2015

Βιοπληροφορική

Παντελής Γ. Μπάγκος

Πανεπιστήμιο Θεσσαλίας

Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα και Βοηθήματα
www.kallipos.gr
HEALLINK
Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών
Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης
ΕΣΠΑ 2007-2013

Βιοπληροφορική

Συγγραφή

Παντελής Γ. Μπάγκος
Αναπληρωτής Καθηγητής
Πανεπιστήμιο Θεσσαλίας

Κριτικός αναγνώστης

Δημήτριος Δ. Λεωνίδας
Αναπληρωτής Καθηγητής
Πανεπιστήμιο Θεσσαλίας

Συντελεστές έκδοσης

ΓΛΩΣΣΙΚΗ ΕΠΙΜΕΛΕΙΑ: Ευφροσύνη-Άλκηστη Παρασκευοπούλου-Κόλλια

ΓΡΑΦΙΣΤΙΚΗ ΕΠΙΜΕΛΕΙΑ: Παναγιώτα Κοντού

ΤΕΧΝΙΚΗ ΕΠΕΞΕΡΓΑΣΙΑ: Παναγιώτα Κοντού

ΣΥΝΔΕΣΜΟΣ ΕΛΛΗΝΙΚΩΝ ΑΚΑΔΗΜΑΪΚΩΝ ΒΙΒΛΙΟΘΗΚΩΝ

Εθνικό Μετσόβιο Πολυτεχνείο

Ηρώων Πολυτεχνείου 9, 15780 Ζωγράφου

www.kallipos.gr

ISBN: 978-960-603-329-2

Αφιερωμένο στην Κωνσταντίνα, την Φωτεινή και τον Γιώργο

Πίνακας περιεχομένων

Πρόλογος

Η βιοπληροφορική είναι ένας διεπιστημονικός κλάδος και παρόλο που ένας κοινά αποδεκτός ορισμός δεν υπάρχει, μια προσπάθεια ορισμού της θα ήταν ως ο επιστημονικός χώρος όπου η σύμπραξη της βιολογίας με την πληροφορική, τη στατιστική και τα μαθηματικά εξερευνά νέους τρόπους για την προσέγγιση των βιολογικών προβλημάτων, καθώς και την αντίληψη βασικών αρχών της βιολογίας. Πρόκειται για γνωστικό χώρο με συγκεκριμένο όσο και ευρύ πεδίο εφαρμογών και αλληλεπίδρασης με τη σύγχρονη δομική, μοριακή, πληθυσμιακή και περιβαλλοντική βιολογία αλλά και περιοχές επαφής με την ιατρική πληροφορική και τη βιοστατιστική. Οι πρωτοπόροι του κλάδου, προέρχονταν αρχικά από διαφορετικές ειδικότητες (όπως τα μαθηματικά, τη μοριακή βιολογία, τη φυσική, την πληροφορική κ.ο.κ.), αλλά στην πορεία συνδιαμόρφωσαν ένα ενιαίο πλαίσιο μέσα στο οποίο οι παραπάνω (κυρίως, αλλά και άλλες) ειδικότητες συνυπάρχουν και αλληλεπιδρούν, με στόχο πάντα, την ανάπτυξη μαθηματικών και υπολογιστικών μεθόδων για την ανάλυση της βιολογικής πληροφορίας (και κυρίως, της πληροφορίας των μακρομορίων). Ο κλάδος της βιοπληροφορικής σήμερα θεωρείται, παγκόσμια, ένας από τους πλέον αναπτυσσόμενους, ενώ έχει ήδη επιδείξει σημαντικά επιτεύγματα και έχει συγκεντρώσει ιδιαίτερα σημαντικές επενδύσεις.

Η βιοπληροφορική έχει ήδη φτάσει, από τις προηγούμενες δεκαετίες, σε ένα αρκετά υψηλό επίπεδο ανάπτυξης αλλά και αυτονόμησης ως ξεχωριστή επιστημονική οντότητα, με ίδρυση επιστημονικών περιοδικών, διεξαγωγή συνεδρίων, αλλά και με λειτουργία εξειδικευμένων προγραμμάτων σπουδών. Παρόμοια είναι η κατάσταση και στη χώρα μας, τηρουμένων πάντα, όλων των αναλογιών. Στην Ελλάδα η βιοπληροφορική έχει ήδη να επιδείξει μια ερευνητικά ιδιαίτερα ενεργή, όσο και ποικιλόμορφη κοινότητα, γεγονός που σηματοδοτήθηκε (ή μάλλον, γεγονός που επισφραγίστηκε), από την ίδρυση της Ελληνικής Εταιρίας Υπολογιστικής Βιολογίας και Βιοπληροφορικής (ΕΕΥΒΒ) εδώ και μερικά χρόνια, από τη διεξαγωγή των ετήσιων συνεδρίων της, αλλά και από τις υπόλοιπες δράσεις της. Επιπλέον δε, μαθήματα βιοπληροφορικής διδάσκονται πλέον σε δεκάδες πανεπιστημιακά τμήματα σε όλη την Ελλάδα, τόσο σε τμήματα βιολογίας και βιολογικών επιστημών γενικότερα (τμηματα βιοχημείας και βιοτεχνολογίας, τμήματα μοριακής βιολογίας και γενετικής, κ.ο.κ.), όσο και σε τμήματα πληροφορικής και μηχανικών Η/Υ. Επίσης, εξειδικευμένα μεταπτυχιακά προγράμματα σπουδών, λειτουργούν στη χώρα μας εδώ για πάνω από μία δεκαετία και μάλιστα, σε αρκετά πανεπιστήμια, τόσο κεντρικά, όσο και περιφερειακά.

Παρ' όλα αυτά, ένα εισαγωγικό εγχειρίδιο για την επιστήμη της βιοπληροφορικής, γραμμένο στα ελληνικά, είναι κάτι που έλειπε από την ακαδημαϊκή κοινότητα και τη σχετική ελληνική βιβλιογραφία. Έχουν βέβαια υπάρξει κάποιες αξιόλογες προσπάθειες, τόσο για μετάφραση κάποιων πετυχημένων ξενόγλωσσων βιβλίων, όσο και για συγγραφή από την αρχή ενός αντίστοιχου συγγράμματος. Προσπάθειες όμως, που είτε είναι πλέον λίγο ξεπερασμένες, καθώς οι εξελίξεις στο λογισμικό και τη μεθοδολογία τρέχουν, είτε αρκετά εξειδικευμένες. Το παρόν βιβλίο, έρχεται να καλύψει αυτό το κενό και μάλιστα έχει έναν αρκετά φιλόδοξο στόχο: να μπορέσει να αποτελέσει ένα βασικό εγχειρίδιο το οποίο θα μπορεί να χρησιμοποιηθεί σαν διδακτικό υλικό σε μαθήματα βιοπληροφορικής, τόσο σε τμήματα βιολογίας, όσο και σε τμήματα πληροφορικής και μηχανικών Η/Υ.

Το βιβλίο αυτό, έρχεται σαν αποτέλεσμα της εργασίας μου την τελευταία δεκαετία, κατά την οποία δίδαξα μαθήματα βιοπληροφορικής σε τρία διαφορετικά πανεπιστήμια, τόσο σε προπτυχιακό, όσο και σε μεταπτυχιακό επίπεδο. Διδάσκοντας (αλλά και, ασχολούμενος με την έρευνα σε αυτό) το αντικείμενο, διαμορφώθηκε με τα χρόνια ένας σκελετός της ύλης και μια δομή, την οποία από καιρό σκόπευα να μετουσιώσω σε βιβλίο. Η ύλη αυτή, είχε κάποιες διαφοροποιήσεις σε σχέση με τις κλασικές προσεγγίσεις των εγχειριδίων που απευθύνονται σε βιολόγους, κυρίως στο ότι περιείχε σε κάπως μεγαλύτερο βαθμό λεπτομέρειες των αλγορίθμων και των μαθηματικών μεθόδων. Από την άλλη, σε σχέση με τα κείμενα που απευθύνονται σε μηχανικούς και φοιτητές πληροφορικής, έπρεπε να υπάρχει, και η γενικότερη εποπτική εικόνα, η περιγραφή του βιολογικού προβλήματος, οι πρακτικοί τρόποι αντιμετώπισης, αλλά και η περιγραφή των εργαλείων λογισμικού που είναι διαθέσιμα κάθε φορά με τα αντίστοιχα πλεονεκτήματα και μειονεκτήματα (χαρακτηριστικά δηλαδή που απευθύνονται κυρίως σε βιολόγους). Δεν ξέρω κατά πόσο το τελικό αποτέλεσμα δικαιώνει αυτές μου τις προσδοκίες (οι αναγνώστες θα το κρίνουν αυτό, τελικά), αλλά σίγουρα έκανα ό,τι καλύτερο μπορούσα για να πετύχω και τους δύο στόχους και η προσπάθεια αυτή με έχει ευχαριστήσει. Σίγουρα πάντως, το βιβλίο είναι διεπιστημονικό και δεν νομίζω ότι θα μπορούσε εύκολα κανείς να ισχυριστεί ότι απευθύνεται μόνο σε βιολόγους, ή μόνο σε πληροφορικούς. Αυτό το χαρακτηριστικό, είναι ιδιαίτερα σημαντικό και αντικατοπτρίζει μια ολόκληρη σχολή σκέψης στη σύγχρονη βιοπληροφορική, αυτήν

που καλεί για διεπιστημονική εκπαίδευση (σε ατομικό επίπεδο), και όχι απλά για διεπιστημονικές συνεργασίες (σε επίπεδο ομάδας) ατόμων με διαφορετικές ειδικεύσεις. Περισσότερα για το θέμα, θα συζητηθούν στο κεφάλαιο 1.

Η συγγραφή αυτού το βιβλίου, ήταν ένα ιδιαίτερα κοπιαστικό έργο και έγινε κατά τη διάρκεια μιας χρονιάς ιδιαίτερα βεβαρυμένης, τόσο σε προσωπικό και επαγγελματικό επίπεδο, όσο και αναφορικά με τη γενικότερη κατάσταση στη χώρα αλλά και, ειδικότερα, στα Ελληνικά πανεπιστήμια. Η επιλογή του βιβλίου από τον "*Κάλλιπο*" και από τα "*Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα και Βοηθήματα*", ήταν αφενός μεν, μια τιμητική επιλογή, αφετέρου δε, μια επιπλέον πηγή πίεσης την οποία επέλεξα συνειδητά έτσι ώστε να μπορέσω να ολοκληρώσω έγκαιρα το βιβλίο (το οποίο όπως είπα παραπάνω, ήθελα από καιρό να γράψω, αλλά ποτέ δεν το αποφάσιζα). Το γεγονός, ότι το βιβλίο θα διατίθεται δωρεάν ως ηλεκτρονικό σύγγραμμα με άδεια χρήσης Creative Commons, αποτέλεσε ένα επιπλέον κριτήριο, καθώς πιστεύω ότι τέτοιες προσπάθειες, οι οποίες βοηθούν στην ελεύθερη διακίνηση της γνώσης, πρέπει να τονωθούν και να ενισχυθούν με κάθε τρόπο. Τέλος, το γεγονός ότι η συγγραφή του βιβλίου συνέπεσε με μια αντίστοιχη δράση, τα "*Ανοιχτά Ψηφιακά Μαθήματα*" (open courses), η οποία ήδη ήταν υπό εξέλιξη στο Πανεπιστήμιο Θεσσαλίας, βοήθησε επίσης, καθώς οι νέες διαφάνειες για τις διαλέξεις των μαθημάτων "*Βιοπληροφορική Ι*" και "*Βιοπληροφορική ΙΙ*", οι οποίες είναι πλέον διαθέσιμες στους φοιτητές αλλά και στο ευρύ κοινό, περιέχουν το νέο υλικό που περιγράφεται στο βιβλίο και διαμορφώθηκαν σχεδόν παράλληλα. Θα ήθελα βέβαια να έχω διαθέσιμο και άλλο υλικό, όπως βιντεοσκοπημένες διαλέξεις, αλλά δυστυχώς κάτι τέτοιο δεν έγινε εφικτό λόγω του μεγάλου φόρτου εργασίας και της πίεσης χρόνου. Επιπλέον δε, η συνεχής αλληλεπίδραση και η ανατροφοδότηση με σχόλια των προπτυχιακών και μεταπτυχιακών φοιτητών, αλλά και συναδέλφων, συνεργατών και φίλων, ήταν μια ιδιαίτερα χρήσιμη διαδικασία σε σχέση με τη συγγραφή του βιβλίου, καθώς μου δόθηκε η ευκαιρία να εντοπίσω κάποια από τα σημεία στα οποία υπάρχουν μεγαλύτερες δυσκολίες από πλευράς φοιτητών στην προσπάθεια κατανόησης δύσκολων εννοιών (ή αν προτιμάτε, μεγαλύτερες δικές μας δυσκολίες στο πώς θα παρουσιάσουμε αυτές τις έννοιες με απλό και κατανοητό τρόπο) και να αναπροσαρμόσω έτσι την ύλη και το διδακτικό υλικό με χρήση κατάλληλων παραδειγμάτων.

Το βιβλίο είναι αρκετά μεγάλο σε έκταση (400 σελίδες), αλλά και πάλι αισθάνομαι ότι κάποια πράγματα θα έπρεπε να προστεθούν. Η ύλη, καλύπτει βέβαια σε ικανοποιητική θα έλεγα λεπτομέρεια, μεγάλο μέρος της "κλασικής" βιοπληροφορικής (τουλάχιστον, όπως αυτή διδασκόταν μέχρι σήμερα). Στο κεφάλαιο 1 περιέχεται, μια αρκετά εκτενής αλλά και αρκετά ενδιαφέρουσα, ιστορική αναδρομή στην εξέλιξη της βιοπληροφορικής, με αναφορές στα κύρια προβλήματα που αυτή αντιμετώπισε, στα ορόσημά της, αλλά και στην επιστημολογική της βάση και τις σχέσεις της με άλλες συναφείς επιστήμες, ενώ γίνεται και ιδιαίτερη αναφορά, τόσο στην ερευνητική, όσο και στην εκπαιδευτική κατάσταση, στην Ελλάδα. Τα κεφάλαια 2 έως και το 9 περιγράφουν τις βάσεις βιολογικών δεδομένων, την ομοιότητα αλληλουχιών και τη στοίχιση, την πολλαπλή στοίχιση, τα πρότυπα και τα προφίλ, τη φυλογενετική ανάλυση, τις μεθόδους πρόγνωσης και τη δομική βιοπληροφορική. Τα Hidden Markov Models, μια μεθοδολογία ιδιαίτερα σημαντική, τόσο στις αναζητήσεις ομοιότητας, όσο και στις μεθόδους πρόγνωσης, αναφέρονται ξεχωριστά και σε μεγαλύτερη έκταση, στο κεφάλαιο 8. Τα κεφάλαια 10 και 11 επεκτείνεται σε κάποια θέματα τα οποία θεωρούνται κάπως πιο προχωρημένα (υπολογιστικές γραμματικές και υπολογιστική γονιδιωματική, αντίστοιχα). Τέλος, το κεφάλαιο 12 περιέχει μια εισαγωγή στη γλώσσα Perl, με αρκετά πρακτικά παραδείγματα και ασκήσεις εμπνευσμένα από τις μεθοδολογίες των προηγούμενων κεφαλαίων (αναζήτηση προτύπων, μέθοδοι πρόγνωσης, χειρισμός αλληλουχιών, εύρεση γονιδίων κ.ο.κ.). Προσπάθησα αρκετά, αλλά δεν ξέρω αν τα κατάφερα, να αποφύγω τη συνηθισμένη παγίδα στην οποία πέφτουν πολλοί συγγραφείς, δηλαδή να εστιάζονται περισσότερο σε αντικείμενα που άπτονται των ειδικών ερευνητικών τους ενδιαφερόντων. Ίσως στα Hidden Markov Models (κεφάλαιο 8) να έδωσα λίγο μεγαλύτερη έμφαση στους αλγορίθμους, από όσο θα χρειαζόταν σε ένα εισαγωγικό βιβλίο, ενώ και σε ότι αφορά τις μεθόδους πρόγνωσης (κεφάλαιο 7), αφιερώνω περισσότερη έκταση στις μεθόδους που ασχολούνται με πρωτεΐνες, παρά σε αυτές που ασχολούνται με DNA/RNA. Γενικότερα όμως, πιστεύω ότι το βιβλίο περιέχει τα περισσότερα από όσα θα περίμενε να βρει κανείς σε ένα εισαγωγικό εγχειρίδιο βιοπληροφορικής. Θα ήθελα να συμπεριλάβω στο βιβλίο και ένα εισαγωγικό κεφάλαιο για την ανάλυση δεδομένων γονιδιακής έκφρασης, αλλά δυστυχώς, τόσο οι χρονικοί, όσο και οι περιορισμοί στην έκταση του συγγράμματος, δεν μου το επέτρεψαν.

Με την παραπάνω διάρθρωση της ύλης, το βιβλίο θα μπορούσε να χρησιμοποιηθεί ως διδακτικό σύγγραμμα, τόσο σε ένα εξαμηνιαίο μάθημα, όσο και σε δύο. Στην πρώτη περίπτωση, τα κεφάλαια 2 έως και 9 θα μπορούσαν να αποτελέσουν τον βασικό κορμό (ίσως και με την προσθήκη του κεφαλαίου 12, αν το μάθημα έχει αρκετές διδακτικές μονάδες και προβλέπει διεξαγωγή εργαστηριακών ασκήσεων). Στη δεύτερη

περίπτωση, την οποία εφαρμόζουμε και εμείς στο Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας, στο μάθημα "*Βιοπληροφορική Ι*" θα μπορούσαν να καλύπτονται τα κεφάλαια 2-8 (ίσως σε λίγο μεγαλύτερη ανάλυση σε σχέση με την προηγούμενη περίπτωση), ενώ στο μάθημα "*Βιοπληροφορική ΙΙ*" θα μπορούσαν να καλύπτονται τα κεφάλαια 9-12, με μεγαλύτερη έμφαση στον προγραμματισμό (με παράλληλες εργαστηριακές ασκήσεις) και ίσως, και με την προσθήκη 1-2 ακόμα θεμάτων που δεν καλύπτονται στο παρόν βιβλίο (όπως για παράδειγμα οι αναλύσεις δεδομένων γονιδιακής έκφρασης). Φυσικά, επιλεγμένα κεφάλαια, με την επιπλέον προσθήκη πρωτότυπης βιβλιογραφίας, θα μπορούσαν να αποτελέσουν βοηθητικό υλικό σε διάφορα μαθήματα βιοπληροφορικής σε μεταπτυχιακό επίπεδο (είτε σε εισαγωγικό μάθημα, είτε κυρίως, σε κάποιο μάθημα ανάλυσης αλληλουχιών). Η σειρά με την οποία πρέπει να διδάσκονται τα κεφάλαια, με έχει επίσης προβληματίσει αρκετά, τόσο κατά τα προηγούμενα χρόνια στα οποία ο σκελετός αυτός διαμορφωνόταν, όσο και κατά τη διάρκεια της συγγραφής του βιβλίου. Το πιο σημαντικό ερώτημα, αφορά τη θέση των Hidden Markov Models σε σχέση με τις υπόλοιπες μεθοδολογίες, και κυρίως, αν αυτά τα μοντέλα θα πρέπει να παρουσιαστούν μετά από τις κλασικές μεθόδους πολλαπλής στοίχισης, κάνοντας ταυτόχρονη αναφορά στα μοτίβα αλληλουχιών και στα Position Specific Scoring Matrices, ή, αν θα πρέπει να παρουσιαστούν μετά από τις κλασικές μεθόδους πρόγνωσης. Στο βιβλίο ακολούθησα τη δεύτερη επιλογή, αλλά στην πράξη, η διδασκαλία απαιτεί πολλές φορές ένα "μπρος-πίσω" σε αυτό το σημείο. Επίσης, τα Νευρωνικά Δίκτυα, τα οποία στον αρχικό σχεδιασμό υπήρχε πρόβλεψη να αποτελέσουν ξεχωριστό κεφάλαιο, κρίθηκε σκόπιμο να αναφερθούν μέσα στο κεφάλαιο 7 (το οποίο αφορά τις μεθόδους πρόγνωσης). Ένα άλλο παρόμοιο πρόβλημα, έχει να κάνει με τις μεθόδους πρόγνωσης του RNA. Τα γενικά χαρακτηριστικά τέτοιων μεθόδων, αναφέρονται στο κεφάλαιο 7, έστω και επιγραμματικά. Παρ' όλα αυτά, η κατανόηση των περισσότερων μεθοδολογιών που χρησιμοποιούνται για την πρόγνωση της δομής αυτών των μορίων, απαιτεί την κατανόηση των γραμματικών, οι οποίες παρουσιάζονται στο κεφάλαιο 10. Συνεπώς και σε αυτή την περίπτωση μια τέτοια "διαταραχή" της γραμμικής ροής του βιβλίου, είναι απαραίτητη. Τέλος, το κεφάλαιο 12, αν και βρίσκεται στο τέλος του βιβλίου, θα μπορούσε να αποτελεί και έναν ανεξάρτητο οδηγό για εργαστηριακές ασκήσεις, οι οποίες θα διεξάγονται παράλληλα, κατα τη διάρκεια του εξαμήνου.

Από το βιβλίο απουσιάζει μια εισαγωγή στις βασικές βιολογικές έννοιες (DNA, RNA, αντιγραφή, μεταγραφή, μετάφραση, πρωτεΐνες, μετα-μεταφραστικές τροποποιήσεις, δομή του κυττάρου, οργανίδια, μεταφορά ουσιών, σηματοδότηση, κ.ο.κ.), οι οποίες παρουσιάζονται μέσα από τα ειδικά τους προβλήματα, τα οποία καλείται να επιλύσει η βιοπληροφορική. Είχα σκοπό αρχικά να γράψω μια τέτοια μικρή εισαγωγή, αλλά γρήγορα διαπίστωσα ότι δεν υπήρχε ιδιαίτερος λόγος. Στα τμήματα βιολογικών επιστημών, η γνώση αυτή έχει ήδη διδαχθεί στους φοιτητές που θα παρακολουθήσουν βιοπληροφορική, ενώ στα υπόλοιπα τμήματα (πληροφορικής ή μηχανικών), η σωστή πρακτική επιβάλει να έχει προηγηθεί κάποιο εισαγωγικό μάθημα στις βιολογικές επιστήμες γενικότερα. Όπως όμως είδαμε από την ανάλυση, η οποία παρουσιάζεται στο κεφάλαιο 1, μόνο μια μειοψηφία από τα τμήματα πληροφορικής και μηχανικών Η/Υ των Ελληνικών πανεπιστημίων, προσφέρουν μάθημα βιοπληροφορικής, ενώ σε κάποια από αυτά ήδη προσφέρονται και εισαγωγικά μαθήματα βιολογίας. Κατά συνέπεια, μια εισαγωγή στις βασικές βιολογικές έννοιες, θα αφορούσε λίγα πανεπιστημιακά τμήματα και, στην πράξη, η πληροφορία που αυτή θα περιείχε, θα μπορούσε να βρεθεί σε μια μεγάλη σειρά εισαγωγικών σημειώσεων που κυκλοφορούν ελεύθερα στο διαδίκτυο, ή ακόμα και από τα σχολικά βιβλία βιολογίας (ούτως ή άλλως, οι ειδικότερες βιολογικές έννοιες που αναφέρονται σε διάφορα σημεία του βιβλίου, εξηγούνται εκεί, όσο αυτό είναι δυνατό). Με βάση όλα τα παραπάνω, θεωρώ ότι το βιβλίο μπορεί τελικά να χρησιμοποιηθεί ως διδακτικό σύγγραμμα, τόσο σε τμήματα βιολογικών επιστημών, όσο και σε τμήματα πληροφορικής και μηχανικών Η/Υ, με τις αντίστοιχες κάθε φορά προσαρμογές στον τρόπο διδασκαλίας και στην παρουσίαση της ύλης. Έτσι, σε κάποιο τμήμα πληροφορικής μπορεί να δοθεί μεγαλύτερη έμφαση στους αλγορίθμους και στις υλοποιήσεις τους, ενώ στα τμήματα βιολογίας, η έμφαση θα μπορούσε να δοθεί περισσότερο στη πλεονεκτήματα και τα μειονεκτήματα των εργαλείων λογισμικού.

Το βιβλίο αυτό, ολοκληρώθηκε χάρη στη συμβολή και τις άοκνες προσπάθειες, μιας σειράς ανθρώπων, τους οποίους πρέπει να ευχαριστήσω δημόσια. Καταρχάς, θα πρέπει να ευχαριστήσω τον κριτικό αναγνώστη, Αναπληρωτή Καθηγητή του Τμήματος Βιοχημείας και Βιοτεχνολογίας, του Πανεπιστημίου Θεσσαλίας, Δημήτριο Λεωνίδα, ο οποίος μου συμπαραστάθηκε σε όλη αυτή τη δύσκολη προσπάθεια, τόσο με τα εύστοχα σχόλια και τις παρατηρήσεις του, όσο και με την ενθάρρυνση που μου παρείχε. Ελάχιστες από τις υποδείξεις του, δεν τις ακολούθησα τελικά (για λόγους που δεν μπορούν να αναλυθούν εδώ), αλλά ελπίζω ότι αυτό δεν θα τον κάνει να αλλάξει την καλή γνώμη, που φαίνεται ότι έχει, για το βιβλίο. Η Παναγιώτα Κοντού, υποψήφια διδάκτορας στο Πανεπιστήμιο Θεσσαλίας, κατέβαλε κάθε προσπάθεια, ούτως ώστε το

κείμενο αυτό να έχει άψογη μορφή. Μορφοποίησε το κείμενο, έκανε τη σελιδοποίηση, διόρθωσε λάθη, έφτιαξε εικόνες και πίνακες, έλεγξε τα προγράμματα, ενώ ανέλαβε και τη δημιουργία του αρχείου της μορφής epub. Η δε γνώμη της ήταν γενικότερα πολύτιμη, καθώς τα τελευταία χρόνια δίδασκε τις εργαστηριακές ασκήσεις στα σχετικά μαθήματα, και είχε άμεση γνώμη για τον τρόπο με τον οποίο προσλαμβάνουν οι φοιτητές το γνωστικό αντικείμενο. Σίγουρα το βιβλίο αυτό, δεν θα είχε έρθει σε αυτή τη μορφή, αν δεν ήταν αυτή (και σίγουρα, η αμοιβή της από το πρόγραμμα δεν ανταποκρίνεται στη συνεισφορά της). Η Ευφροσύνη-Άλκηστη Παρασκευοπούλου-Κόλλια, βοήθησε κάνοντας τη γλωσσική επιμέλεια σε όλα τα κεφάλαια, και μάλιστα χωρίς αμοιβή, οπότε πρέπει να την ευχαριστήσω διπλά. Τα υπόλοιπα μέλη της ερευνητικής μας ομάδας στο Εργαστήριο Μοριακής και Υπολογιστικής Βιολογίας και Γενετικής του Τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική, του Πανεπιστημίου Θεσσαλίας, η Νίκη Δήμου, ο Νίκος Νικολακάκης, η Κατερίνα Πανταβού, η Γκρέτα Μπράλιου και η Αθανασία Παυλοπούλου, βοήθησαν, άλλοι λιγότερο και άλλοι περισσότερο, σε διάφορα στάδια, είτε λέγοντας τη γνώμη τους για κάποιο ειδικό θέμα, είτε διαβάζοντας και σχολιάζοντας κάποιο προσχέδιο, είτε συμμετέχοντας στις γενικότερες συζητήσεις και στο καλό κλίμα της ομάδας. Τα προσχέδια των κεφαλαίων είχαν χρησιμοποιηθεί, σαν διδακτικό υλικό, σε διάφορα μαθήματα και πολλοί προπτυχιακοί και μεταπτυχιακοί φοιτητές έκαναν εύστοχες παρατηρήσεις. Ειδική αναφορά όμως, χρειάζεται στη Γεωργία Καπούλα, η οποία εντόπισε πολλά τυπογραφικά και ορθογραφικά λάθη και πρότεινε επιπλέον τρόπους για να γίνει το κείμενο περισσότερο απλό και κατανοητό. Ένα προσχέδιο του κεφαλαίου 9, το είχαν διαβάσει ο Αναστάσης Περράκης και ο Νίκος Παπανδρέου, οι οποίοι και έκαναν ουσιαστικά σχόλια, ενώ συμβουλές για τα εργαλεία λογισμικού του κεφαλαίου 10, έδωσε ο Κώστας Παπαδημητρίου. Δεν έχω βέβαια την ψευδαίσθηση, ότι το βιβλίο δεν θα έχει λάθη, είτε συντακτικά, ορθογραφικά και τυπογραφικά, είτε ακόμα και λάθη ουσίας (αν και ελπίζω τα τελευταία να είναι τα λιγότερα). Φυσικά, όλα αυτά τα εναπομείναντα λάθη, είναι αποκλειστικά δική μου ευθύνη και θα παρακαλούσα θερμά τους αναγνώστες που θα τα εντοπίσουν, να τα αναφέρουν, είτε σαν σχόλιο στη σελίδα: www.compgen.org/books/bioinformatics, είτε με ένα email στη διεύθυνση: books@compgen.org. Φυσικά, αν έκαναν το ίδιο, και όλοι όσοι θέλουν να αφήσουν οποιοδήποτε άλλο σχόλιο στο βιβλίο, ή όσοι έχουν οποιαδήποτε άλλη απορία, θα με χαροποιούσε ιδιαίτερα.

Δεν θα πρέπει, φυσικά, να παραλείψω να ευχαριστήσω, τον πρώην καθηγητή μου, Ομότιμο Καθηγητή του Πανεπιστημίου Αθηνών, Σταύρο Ι. Χαμόδρακα, ο οποίος ήταν και ο άνθρωπος ο οποίος με έφερε για πρώτη φορά σε επαφή με τη βιοπληροφορική σε προπτυχιακό επίπεδο (τη δεκαετία του 1990, όταν ακόμα και ο ίδιος ο όρος "βιοπληροφορική" δεν ήταν ευρέως διαδεδομένος). Στη συνέχεια, επέβλεψε τη διδακτορική μου διατριβή και τη μετα-διδακτορική μου έρευνα, ενώ είχα τη χαρά να διδάξουμε μαζί μεταπτυχιακά μαθήματα, να συνεργαστούμε σε μια σειρά από ερευνητικά θέματα, αλλά και να συνυπηρετήσουμε για χρόνια στο Διοικητικό Συμβούλιο της ΕΕΥΒΒ. Ο ίδιος δεν έχει δει ακόμα δείγματα από το βιβλίο αυτό (καθώς δεν ήθελα να τον κουράσω με προσχέδια), αλλά έχω την αίσθηση ότι στην τελική μορφή, το βιβλίο θα του αρέσει.

Ακόμα και χωρίς τη συγγραφή ενός βιβλίου, η δουλειά ενός πανεπιστημιακού (τουλάχιστον, ενός σοβαρού πανεπιστημιακού), είναι ήδη, παρά τις περί του αντιθέτου φήμες, αρκετά απαιτητική. Απαιτεί ταξίδια, μετακινήσεις, συναντήσεις, διάβασμα και δουλειά στο σπίτι και γενικά, δουλειά σε ώρες και μέρες κατά τις οποίες οι περισσότεροι άνθρωποι ασχολούνται με άλλα, περισσότερο ανέμελα, πράγματα. Το παρόν βιβλίο, όπως ανέφερα ήδη, υλοποιήθηκε σε μια δύσκολη χρονιά, κατά την οποία, εκτός των προπτυχιακών και μεταπτυχιακών μαθημάτων που συνήθως διδάσκω (3 σε κάθε εξάμηνο) και τη συνήθη επίβλεψη αρκετών πτυχιακών εργασιών, είχα την ευθύνη και για δύο απαιτητικά ερευνητικά προγράμματα τα οποία έπρεπε να υλοποιηθούν μέσα στις ίδιες περίπου προθεσμίες. Συνεπώς, το βιβλίο αυτό προσέθεσε όχι μόνο επιπλέον κόπο, αλλά και περισσότερο άγχος. Δεν θα μπορούσα να έχω ανταποκριθεί σε αυτές τις απαιτήσεις, αν δεν είχα την αμέριστη συμπαράσταση και τη στήριξη των δικών μου ανθρώπων, της συζύγου μου Κωνσταντίνας και των παιδιών μου, Φωτεινής και Γιώργου. Εκτός από την ηθική και ψυχολογική στήριξη, έκαναν και αυτοί τις δικές τους μικρές θυσίες, καθώς με ανέχτηκαν να δουλεύω με τις ώρες, να λείπω για μέρες, ενώ ακόμα και στις διακοπές (Χριστούγεννα, Πάσχα, καλοκαίρι), ήμουν αγκαλιά με έναν υπολογιστή. Για όλα τα παραπάνω, αλλά και για τη γενικότερη παρουσία τους και τη χαρά που φέρνουν στη ζωή μου, θα ήθελα να τους ευχαριστήσω και να τους αφιερώσω αυτό το έργο.

Λαμία, 17/11/2015 Παντελής Μπάγκος

Πίνακας συντομεύσεων-ακρωνύμια

BLAST	Basic Local Alignment Search Tool
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
MSA	Multiple Sequence Alignment
HMM	Hidden Markov Model
PSSM	Position-Specific Scoring Matrix
EBI	European Bioinformatics Institute
NCBI	National Center for Biotechnology Information
NLM	National Library of Medicine
NIH	National Institutes of Health
PDB	Protein Data Bank
CABIOS	Computer Applications in the Biosciences
SRS	Sequence Retrieval System
EEYBB	Ελληνική Εταιρία Υπολογιστικής Βιολογίας και Βιοπληροφορικής
ISCB	International Society for Computational Biology
ISMB	Intelligent Systems for Molecular Biology
CAFASP	Critical Assessment of Fully Automated Structure Prediction
CAPRI	Critical Assessment of Prediction of Interactions
CASP	Critical Assessment of protein Structure Prediction

Κεφάλαιο 1: Εισαγωγή στη Βιοπληροφορική

Σύνοψη
Η βιοπληροφορική είναι ένας ταχέα αναπτυσσόμενος διεπιστημονικός κλάδος. Παρόλο που ένας ακριβής ορισμός δεν μπορεί να δοθεί, και υπάρχουν μάλιστα και πολλές διαφωνίες ανάλογα με την οπτική και το υπόβαθρο του καθενός, είναι σαφές ότι πρόκειται για τον επιστημονικό κλάδο που βρίσκεται στην περιοχή επαφής της βιολογίας με τα μαθηματικά και την επιστήμη υπολογιστών. Στο κεφάλαιο αυτό, θα προσπαθήσουμε να εξετάσουμε τέτοια θέματα από όλες τις πλευρές. Θα δούμε το ιστορικό πλαίσιο ανάπτυξης της βιοπληροφορικής (ή καλύτερα, της υπολογιστικής βιολογίας), το διεπιστημονικό χαρακτήρα της, τους μύθους που τη συνοδεύουν, αλλά θα δούμε και τις τελευταίες εξελίξεις στη βιβλιογραφία της βιοπληροφορικής, τόσο διεθνώς όσο και στην Ελλάδα. Με τα περιεχόμενα αυτού το κεφαλαίου, ευελπιστούμε ότι οι αναγνώστες θα μπορέσουν να αποκτήσουν μια εποπτική εικόνα αυτού του σύνθετου ερευνητικού πεδίου η οποία θα τους βοηθήσει στην κατανόηση των επόμενων κεφαλαίων.

Προαπαιτούμενη γνώση
Ο αναγνώστης πρέπει να έχει τις βασικές γνώσεις μοριακής βιολογίας και γενετικής.

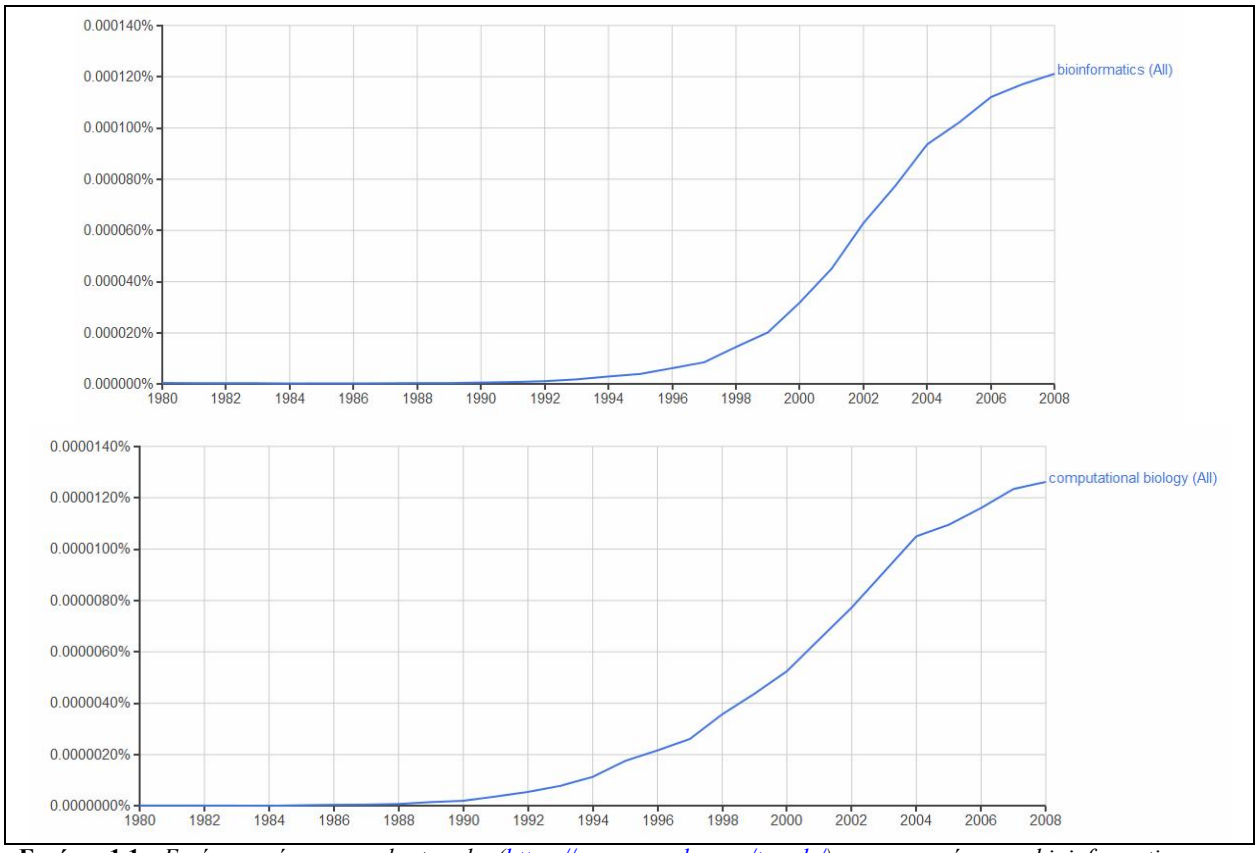
1. Εισαγωγή

Η βιοπληροφορική είναι ένας διεπιστημονικός κλάδος, και παρόλο που ένας κοινώς αποδεκτός ορισμός δεν υπάρχει, μια προσπάθεια ορισμού της θα ήταν ως ο επιστημονικός χώρος όπου η σύμπραξη της βιολογίας με την πληροφορική, τη στατιστική και τα μαθηματικά εξερευνά νέους τρόπους για την προσέγγιση των βιολογικών προβλημάτων, καθώς και την αντίληψη βασικών αρχών της βιολογίας. Πρόκειται για γνωστικό χώρο με συγκεκριμένο όσο και ευρύ πεδίο εφαρμογών και αλληλεπίδρασης με τη σύγχρονη δομική, μοριακή, πληθυσμιακή και περιβαλλοντική βιολογία. Ο κλάδος της βιοπληροφορικής σήμερα θεωρείται, παγκόσμια, ένας από τους πλέον αναπτυσσόμενους, ενώ έχει ήδη επιδείξει σημαντικά επιτεύγματα και έχει συγκεντρώσει ιδιαίτερα σημαντικές επενδύσεις.

Καθώς η βιοπληροφορική είναι διεπιστημονικός κλάδος, υπάρχουν πολλές και αντικρουόμενες απόψεις σχετικά με τον ορισμό της αλλά και σχετικά με το επιστημολογικό της καθεστώς. Η πιο απλοϊκή προσέγγιση λέει ότι η βιοπληροφορική είναι απλώς η εφαρμογή κάποιων τεχνολογιών (μαθηματικών, υπολογιστικών, κ.ο.κ.) σε βιολογικά προβλήματα. Η πιο σύνθετη, στην οποία προσχωρεί και ο συγγραφέας αυτού του βιβλίου, είναι ότι η βιοπληροφορική είναι πλέον μια ξεχωριστή επιστήμη, η οποία ναι μεν χρησιμοποιεί υπολογιστικά και μαθηματικά εργαλεία σε βιολογικά προβλήματα, αλλά κάνει και κάτι άλλο: παράγει (ή τουλάχιστον, προσπαθεί να παράγει) και γενικότερους νόμους που διέπουν αυτά τα βιολογικά συστήματα. Με αυτόν τον τρόπο, μιλάμε για μια βιολογικής κατεύθυνσης επιστήμη ή ειδικότητα, με τη δική της παράδοση και τις δικές της μεθοδολογίες. Πολλές φορές χρησιμοποιείται παράλληλα και ο όρος υπολογιστική βιολογία, ενώ από πολλούς οι δυο αυτοί όροι χρησιμοποιούνται αδιάκριτα μεταξύ τους. Όπως θα δούμε παρακάτω η άποψη αυτή μάλλον δικαιώνεται ιστορικά. Παρ' όλα αυτά, ένας λογικός διαχωρισμός είναι ότι ο όρος βιοπληροφορική αναφέρεται κυρίως στην πρακτική εφαρμογή, δηλαδή στη χρήση αλγόριθμων και υπολογιστικών τεχνικών που επιτρέπουν την απάντηση βιολογικών ερωτημάτων (π.χ. αναζήτηση μιας ακολουθίας σε μια βάση δεδομένων, χειρισμός μεγάλου όγκου ακολουθιών ή δεδομένων γονιδιακής έκφρασης κλπ), ενώ ο όρος υπολογιστική βιολογία είναι κάπως πιο θεωρητικός και αναφέρεται στα θεωρητικά αποτελέσματα στα οποία στηριζόμαστε για να αναπτύξουμε έναν αλγόριθμο, μια μεθοδολογία ή ένα γενικό νόμο.

Η προσωπική άποψη του συγγραφέα, είναι ότι παρ' όλες τις επιμέρους διαφορές που αναλύθηκαν παραπάνω, ο όρος υπολογιστική βιολογία θα έπρεπε να χρησιμοποιείται γενικά αντί της βιοπληροφορικής. Και τούτο, γιατί με αυτόν τον τρόπο θα δίναμε έμφαση στο αντικείμενο που μελετάμε (τα βιολογικά συστήματα) και όχι στον τρόπο με τον οποίο το κάνουμε αυτό (την υπολογιστική μεθοδολογία). Με λίγα λόγια, πρέπει να γίνει κατανοητό ότι η βιοπληροφορική/υπολογιστική βιολογία, είναι πρώτα από όλα βιολογία, μελέτη των ζωντανών οργανισμών με υπολογιστικές μεθοδολογίες. Αυτό δεν σημαίνει όμως ότι πρέπει να υπαχθεί στη συντεχνιακή αντίληψη (κάτι συνηθισμένο στη χώρα μας) ότι αυτή είναι μια δραστηριότητα μόνο για βιολόγους. Το αντίθετο, είναι ένας διεπιστημονικός κλάδος, στον οποίο μπορούν και πρέπει να συνεισφέρουν επιστήμονες εκπαιδευμένοι σε διάφορες ειδικότητες (βιολόγοι, μαθηματικοί, επιστήμονες Η/Υ, μηχανικοί κ.ο.κ.). Αυτό όμως που πρέπει να γίνει, είναι ότι πρέπει επιπλέον, να υπάρξει και

διεπιστημονική εκπαίδευση έτσι ώστε να υπάρξει ένας κοινός τόπος και μια κοινή γλώσσα στην οποία όλοι αυτοί οι ειδικοί θα μπορούν να συνεννοούνται. Και φυσικά, πρέπει να υπάρξει και προσπάθεια δημιουργίας διεπιστημονικών ατόμων, όχι μόνο ομάδων. Αξίζει να αναφερθεί πάντως, ότι παρόλο που υπάρχουν δεκάδες επιστημονικά περιοδικά με σαφή αναφορά στη βιοπληροφορική, οι γενικότερες ταξινομήσεις των επιστημονικών περιοδικών από την ISI, περιλαμβάνουν σαν ξεχωριστή κατηγορία μόνο την γενικότερη περίπτωση (Mathematical and Computational Biology), κατηγορία που περιλαμβάνει και περιοδικά βιοπληροφορικής και υπολογιστικής βιολογίας αλλά και περιοδικά βιοστατιστικής και ιατρικής πληροφορικής. Όπως θα δούμε παρακάτω, τέτοιες συνέργειες με άλλα παρεμφερή επιστημονικά πεδία, είναι αρκετά κοινές στο χώρο.



Εικόνα 1.1: *Εικόνα από το google trends (https://www.google.com/trends/) για τους όρους «bioinformatics» και «computational biology», αντίστοιχα.*

Στις επόμενες παραγράφους, θα προσπαθήσουμε να αποδώσουμε μια σύντομη ιστορική αναδρομή του επιστημονικού κλάδου της βιοπληροφορικής και να ανιχνεύσουμε τις ρίζες του. Θα δούμε τη διεπιστημονικότητά του, αλλά και τις περιοχές επαφής με τις γειτονικές επιστήμες και, τέλος, θα δούμε τις τάσεις στη διεθνή βιβλιογραφία αλλά και αναλυτικά την κατάσταση στην Ελλάδα.

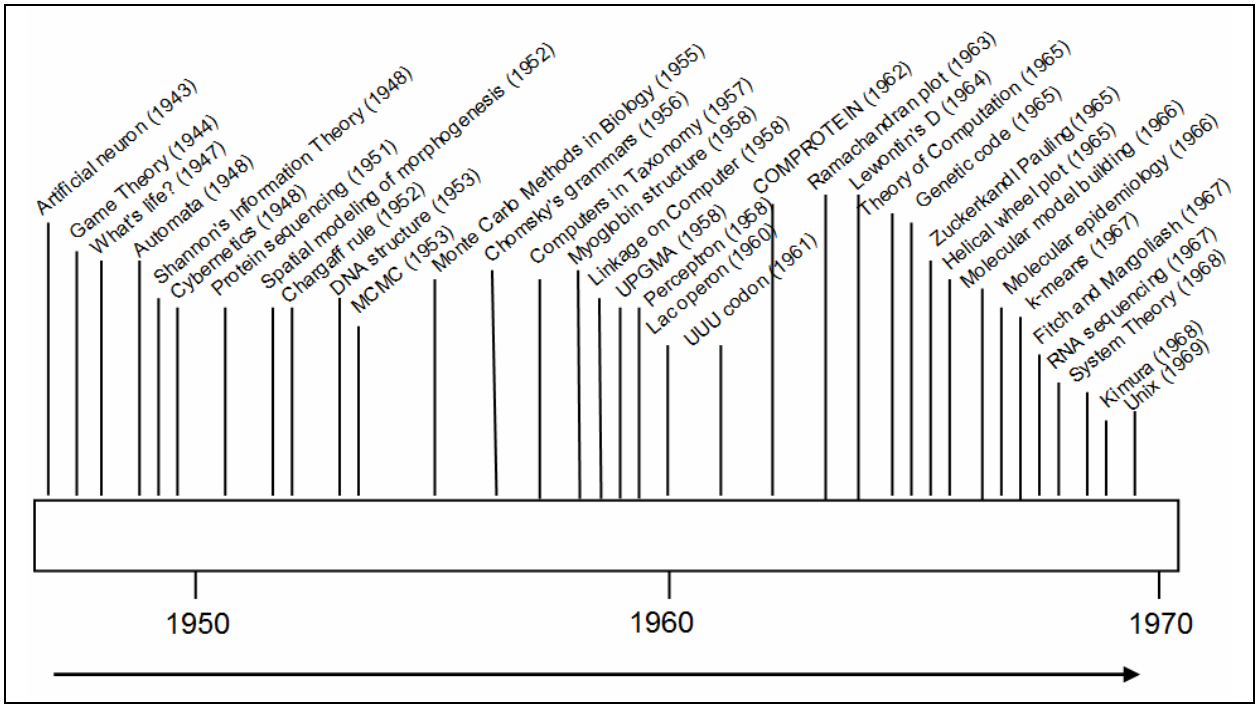
1.1. Η ιστορία της βιοπληροφορικής και της υπολογιστικής βιολογίας

Όπως είδαμε παραπάνω, ο όρος βιοπληροφορική (bioinformatics), είναι ιδιαίτερα πρόσφατος και εμφανίστηκε στη δεκαετία του 1990. Επίσης, είναι ένας ιδιαίτερα επιτυχημένος όρος καθώς έχει επικρατήσει η χρήση του διεθνώς, κάτι μάλλον περίεργο, ιδιαίτερα αν σκεφτούμε ότι ο όρος informatics στα αγγλικα δεν χρησιμοποιείται και πολύ. Όπως θα δούμε παρακάτω όμως, ο όρος αυτός είναι και παραπλανητικός, καθώς συνήθως αναφερόμαστε στην υπολογιστική ανάλυση των βιολογικών συστημάτων (με βασική αναφορά στις βιολογικές αλληλουχίες) και μια τέτοια προσπάθεια δεν ξεκίνησε προφανώς τη δεκαετία του 1990, αλλά πολύ πιο πριν. Αν δούμε τα πράγματα από μια ιστορική σκοπιά, θα δούμε ότι ήδη από τις αρχές του 20ου αιώνα, υπήρχαν πολλές προσπάθειες μαθηματικοποίησης και ποσοτικοποίησης των βιολογικών φαινομένων, αλλά

αυτές οι προσπάθειες συμβάδιζαν πάντα με τις υπολογιστικές μεθοδολογίες της εποχής, όσο και με το είδος των βιολογικών δεδομένων που ήταν κάθε φορά διαθέσιμα. Τα περισσότερα από όσα περιγράφονται παρακάτω, βασίζονται σε γνωστές ιστορικές ανασκοπήσεις (Hagen, 2000; Ouzounis & Valencia, 2003; Roberts, 2000; Searls, 2010; Trifonov, 2000), αλλά και στην προσωπική εμπειρία του συγγραφέα. Προφανώς μια τέτοια θεώρηση, και ειδικά στην έκταση και την ανάλυση ενός διδακτικού εγχειριδίου, θα είναι αναγκαστικά ελλιπής, αλλά ελπίζω ότι η γενική εικόνα που θα μείνει τελικά στον αναγνώστη θα είναι αποκαλυπτική όσο και χρήσιμη.

1.1.1. Οι δεκαετίες του 1950 και 1960

Αν προσπεράσουμε τις προσπάθειες μαθηματικοποίησης της γενετικής, που έδωσαν γένεση στη γενετική πληθυσμών και τη σύγχρονη εξελικτική θεωρία, από την εποχή των Hardy, Weinberg και Fisher, Wright κλπ, θα πρέπει να ανιχνεύσουμε τις απαρχές της σύγχρονης υπολογιστικής βιολογίας, στις απαρχές της ίδιας της σύγχρονης μοριακής βιολογίας τη δεκαετία του 1950 και 1960 (Εικόνα 1.2). Για παράδειγμα, τα πειράματα του Chargaff που έδειξαν ότι το ποσοστό Αδενίνης είναι το ίδιο με το ποσοστό της Θυμίνης και το ποσοστό Γουανίνης ίσο με αυτό της Κυτοσίνης σε κάθε μόριο DNA, ήταν οι πρώτες ενδείξεις για κάποια μορφή ψηφιακής πληροφορίας στις βιολογικές αλληλουχίες. Τα πειράματα αυτά, ως γνωστόν χρησιμοποιήθηκαν από τους Watson και Crick για να μπορέσουν να προσδιορίσουν την τρισδιάστατη δομή του DNA η οποία τους έδωσε και το νόμπελ (χρησιμοποιώντας δεδομένα του Wilkins, ο οποίος βραβεύτηκε μαζί τους αλλά και της Franklin η οποία όμως είχε πεθάνει στο ενδιάμεσο). Τη δεκαετία του 1960 έγιναν επίσης και οι πρωτοποριακές μελέτες των Jacob και Monod στη γονιδιακή ρύθμιση (το οπερόνιο της λακτόζης). Ενώ όσον αφορά τις πρωτεΐνες, μετά τον προσδιορισμό των πρώτων τρισδιάστατων δομών (ινσουλίνη και μυογλοβίνη), και τη βράβευση των Perutz και Kendrew με το νόμπελ το 1962, ακολούθησαν τη δεκαετία του 1960 μια σειρά παρόμοιες σημαντικές δομές (λυσοζύμη, παπαΐνη, ριβονουκλεάση κ.ο.κ.), και άνοιξε ο δρόμος για τη μελέτη της δομής και της λειτουργίας των πρωτεϊνών σε ατομικό επίπεδο. Επίσης, η εύρεση της πρωτοταγούς δομής των πρωτεϊνών έγινε το 1951, και του RNA το 1967. Βλέπουμε λοιπόν ότι πολλά από τα προβλήματα που απασχολούν τη βιοπληροφορική μέχρι σήμερα, έχουν τις ρίζες τους στην έκρηξη που πραγματοποιήθηκε στη μοριακή βιολογία τη δεκαετία του 1960.



Εικόνα 1.2: *Η εξέλιξη των ιδεών με τις μεγαλύτερες ανακαλύψεις σχετικές με τη βιοπληροφορική μέχρι και το τέλος της δεκαετίας του 1960.*

Παράλληλα, από τη δεκαετία του 1950 και του 1960 είχαν τεθεί ήδη και τα θεμέλια της σύγχρονης θεωρητικής πληροφορικής, με τη θεωρία υπολογισμού, τη θεωρία πληροφορίας του Shannon, τη μηχανή του

Turing, τα αυτόματα και τη θεωρία παιγνίων του von Neumann, τη μελέτη των συμβολοσειρών (strings), την θεωρία συστημάτων, την κυβερνητική και τον ορισμό των γραμματικών από τον Chomsky. Έτσι, δεν είναι περίεργο, αν αναλογιστούμε και τα παραπάνω, ότι οι πρώτες προσπάθειες υπολογιστικής αντιμετώπισης βιολογικών προβλημάτων, εμφανίστηκαν τη δεκαετία του 1960 και σε αυτές βρίσκονται τα πρώτα ψήγματα αυτού που σήμερα ονομάζουμε υπολογιστική βιολογία και βιοπληροφορική. Έτσι, η αποκρυπτογράφηση του γενετικού κώδικα, ήταν κομβικό σημείο στην ανάπτυξη της μοριακής βιολογίας και όλων των βιοεπιστημών. Αυτή η ίδια η φύση του γενετικού κώδικα, ο οποίος στην πραγματικότητα είναι μια συνάρτηση, μια διμελής απεικόνιση από το σύνολο των 64 τριπλετών στο σύνολο των 20 αμινοξέων, ήταν ήδη αντικείμενο έντονης θεωρητικής αλλά και υπολογιστικής επεξεργασίας από τη δεκαετία του 1960, ενώ όταν διαλευκάνθηκε πειραματικά έγιναν και πολλές θεωρητικές μελέτες για τις ιδιότητές του και την προέλευση του. Εμφανίστηκαν επίσης οι εφαρμογές των πρώτων υπολογιστών της εποχής στη βιολογία, με τη χρήση τους μεταξύ άλλων στην Ταξινομική και στην κατασκευή μοριακών μοντέλων για την κρυσταλλογραφία. Την ίδια εποχή, είδαμε και την πρώτη προσπάθεια χρήσης βιολογικών αλληλουχιών για εξελικτικές μελέτες από τους Zuckerkandl και Pauling, τη χρήση τους για την κατασκευή φυλογενετικών δέντρων από τους Fitch and Margoliash αλλά και τα πρώτα μαθηματικά μοντέλα της μοριακής εξέλιξης από τους Kimura και Nei. Στο επίπεδο των πρωτεϊνών είδαμε τις πρώτες εργασίες του Ramachandran για τη μελέτη των δομικών ιδιοτήτων και των περιορισμών των αμινοξικών καταλοίπων σε μια πρωτεϊνική δομή, από τις οποίες έχει προκύψει το πασίγνωστο διάγραμμα Ramachandran με τις επιτρεπτές δίεδρες γωνίες που εμφανίζονται σε πρωτεϊνικές δομές, αλλά και τα πρώτα helical wheel plots.

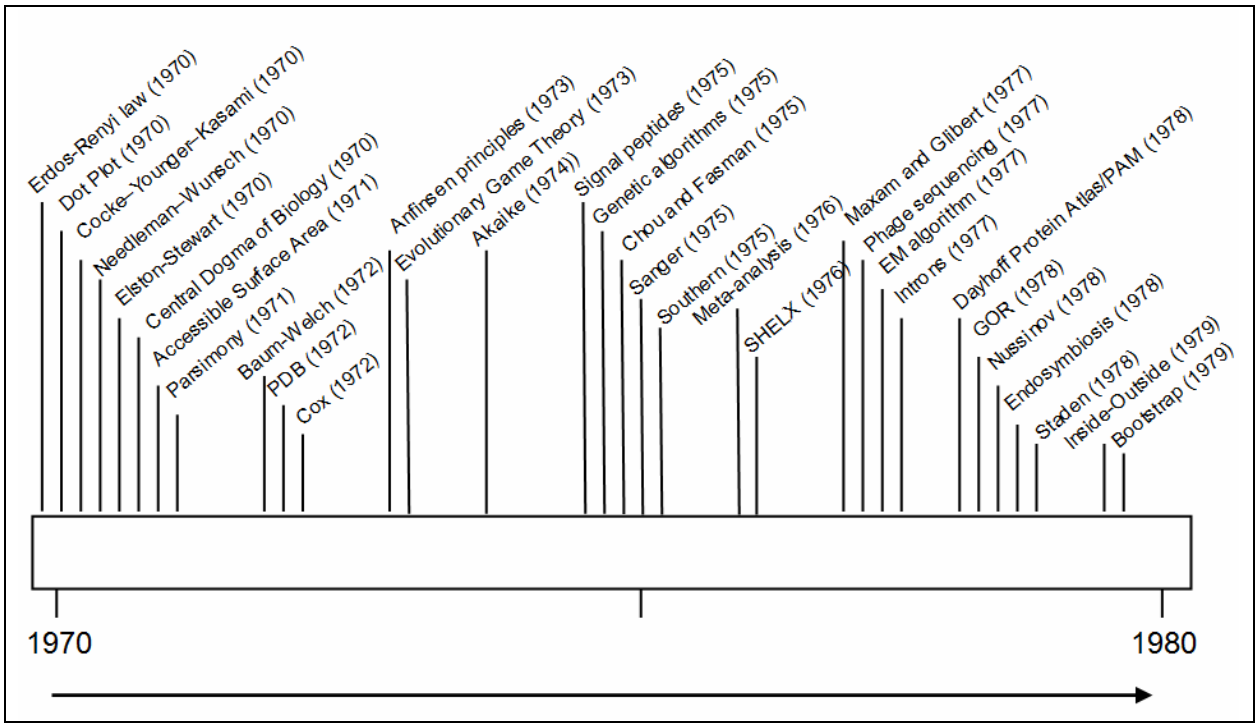
1.1.2. Η δεκαετία του 1970

Την επόμενη δεκαετία, η έρευνα συνεχίστηκε με αυξανόμενο ρυθμό (Εικόνα 1.3). Μια από τις πιο σημαντικές συνεισφορές αυτής της περιόδου, με ευρύτερες συνέπειες για τις βιοεπιστήμες, ήταν η σύγκλιση της κλασικής πληθυσμιακής γενετικής με τη μοριακή εξέλιξη, με αφορμή τις εργασίες του Kimura που είδαμε πριν. Έτσι, φτάσαμε στην εμφάνιση της θεωρίας της ουδέτερης εξέλιξης και στην υπόθεση του σταθερού ρυθμού εξελικτικών αλλαγών, η οποία είναι γνωστή ως το «μοριακό ρολόι». Την ίδια εποχή εμφανίστηκε και ο γνωστός αλγόριθμος του Fitch για την φειδωλή ανακατασκευή φυλογενετικών δέντρων με τη χρήση αλληλουχιών (μέθοδος της μέγιστης φειδωλότητας). Καθώς ο γενετικός κώδικας είχε αποκαλυφθεί, και είχε διαλευκανθεί ο ρόλος των RNA στην μεταγραφή και τη μετάφραση, το κεντρικό δόγμα της βιολογίας διατυπώθηκε από τον Crick το 1970. Την ίδια δεκαετία, εμφανίστηκαν και οι πρώτες μεθοδολογίες αλληλούχισης νουκλεϊκών οξέων από τους Sanger και Maxam-Glbert, μεθοδολογίες που έδωσαν ώθηση στη μελέτη των γονιδιωμάτων και με διάφορες παραλλαγές και τροποποιήσεις έχουν φτάσει μέχρι σήμερα, στις σύγχρονες μεθόδους αλληλούχισης.

Στο επίπεδο των πρωτεϊνών, οι πρωτοποριακές εργασίες του Anfinsen για τις αρχές που καθορίζουν την πρωτεϊνική δομή έδωσαν ώθηση στην έρευνα σε αυτό το πεδίο. Τότε εμφανίστηκαν οι πρώτες μέθοδοι υπολογισμού της προσβασιμότητας στο διαλύτη, όσο και οι πρώτες εργασίες για τις προτιμήσεις των αμινοξέων για τα διάφορα στοιχεία δευτεροταγούς δομής, οι οποίες οδήγησαν στον πρώτο αλγόριθμό πρόγνωσης της δευτεροταγούς δομής πρωτεϊνών από τους Chou και Fasman το 1975 (ενώ φυσικά ακολούθησαν και άλλοι τα επόμενα χρόνια). Επίσης λίγο αργότερα, εμφανίστηκαν και οι πρώτες προσπάθειες πρόγνωσης της δομής των RNA. Τα μοριακά γραφικά, αλλά και οι πρώτες προσπάθειες προσομοίωσης του πρωτεϊνικού διπλώματος με μοριακή δυναμική, εμφανίστηκαν επίσης εκείνη την εποχή. Παρόμοια με τις πρωτεΐνες, εμφανίστηκαν και οι πρώτοι αλγόριθμοι πρόγνωσης της δευτεροταγούς δομής του RNA με τις πρωτοποριακές εργασίες της Nussinov.

Μια από τις πιο σημαντικές αλγοριθμικές συνεισφορές στην υπολογιστική βιολογία που συνέβησαν τη δεκαετία του 1970, ήταν η εμφάνιση των αλγορίθμων δυναμικού προγραμματισμού για τη στοίχιση βιολογικών αλληλουχιών (κυρίως πρωτεϊνών), με πρώτο τον αλγόριθμο για ολική στοίχιση των Needleman και Wunsch που παρουσιάστηκε το 1970. Ακολούθησαν και άλλες προσεγγίσεις και μελέτες στη μεθοδολογία και τα στατιστικά της στοίχισης, ενώ το 1970 έκανε και την εμφάνισή του το διάγραμμα σημείων (dot-plot). Τέλος, αυτή τη δεκαετία εμφανίστηκαν και οι πρώτες βάσεις βιολογικών δεδομένων. Η PDB εμφανίστηκε το 1972 (όταν υπήρχαν μόλις 10 τρισδιάστατες δομές πρωτεϊνών), ενώ η Dayhoff παρουσίασε το 1978 και την πρώτη συλλογή πρωτεϊνικών αλληλουχιών οι οποίες ήταν γνωστές εκείνα τα χρόνια, μια συλλογή που κατά κάποιον τρόπο μπορεί να θεωρηθεί ο πρόδρομος της PIR. Τέλος, τα πρώτα προγράμματα H/Y για απλές αναλύσεις σε βιολογικές αλληλουχίες έκαναν την εμφάνισή τους (μετάφραση μιας κωδικής αλληλουχίας,

εύρεση προτύπων, αναγνώριση υποκινητών και θέσεων δράσης περιοριστικών ενζύμων κ.ο.κ.). Όπως είναι φανερό, ήδη από τη δεκαετία του 1970 είχε ήδη σχηματιστεί μια καθαρή εικόνα του ερευνητικού πεδίου της βιοπληροφορικής. Υπήρχαν οι αλγόριθμοι στοίχισης, η θεωρία της μοριακής εξέλιξης και η ποσοτικοποίηση των εξελικτικών αλλαγών, η κατασκευή φυλογενετικών δέντρων, οι μεθοδολογίες μελέτης και πρόγνωσης της δευτεροταγούς και τριτοταγούς δομής των πρωτεϊνών και οι πρώτες βιολογικές βάσεις δεδομένων.



Εικόνα 1.3: *Η εξέλιξη των ιδεών με τις μεγαλύτερες ανακαλύψεις σχετικές με τη βιοπληροφορική στη δεκαετία του 1970.*

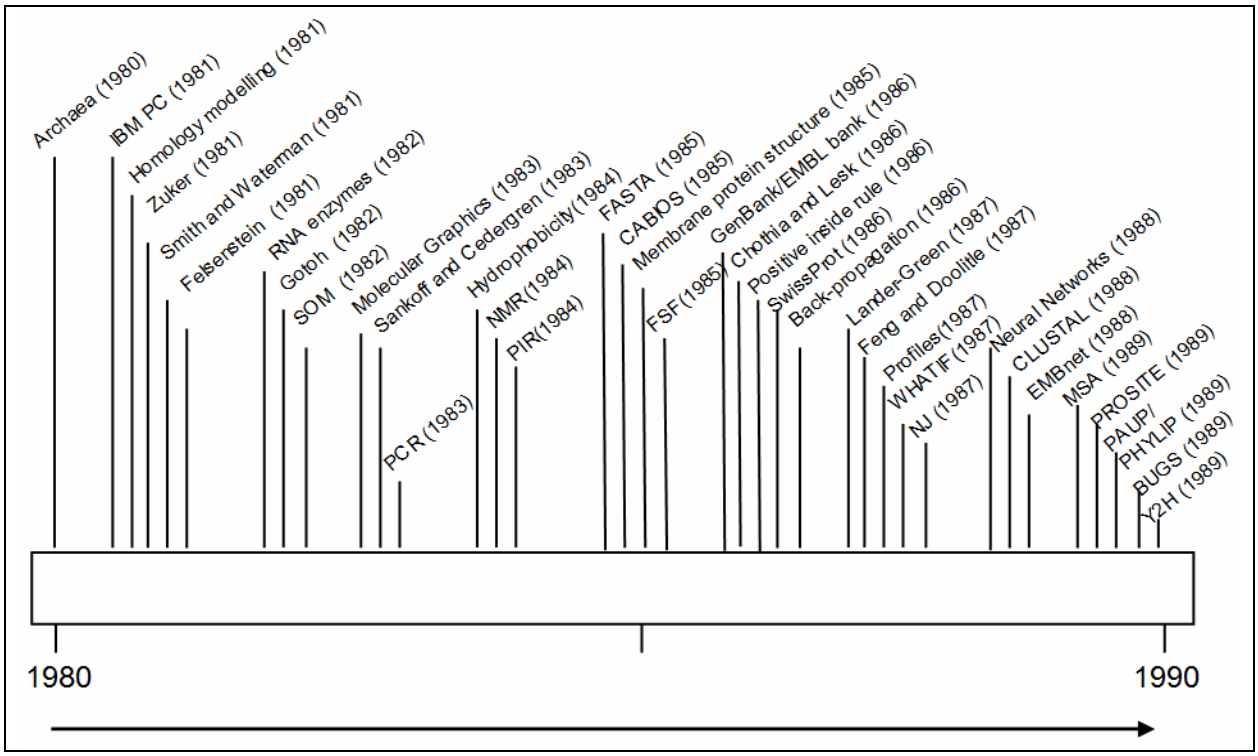
1.1.3. Η δεκαετία του 1980

Η δεκαετία του 1980 ήταν η δεκαετία στην οποία το πεδίο της υπολογιστικής βιολογίας πήρε πλέον μια ξεκάθαρη μορφή, σαν ένας ξεχωριστός κλάδος θέτοντας τα δικά του προβλήματα αλλά και παρουσιάζοντας και τα σημαντικά του επιτεύγματα. Αρχίζουν να κάνουν μαζική εμφάνιση οι αντίστοιχες δημοσιεύσεις στα υψηλού κύρους βιολογικά περιοδικά (Science, Nature, Nucleic Acid Research), ενώ και τα πρώτα εξειδικευμένα περιοδικά κάνουν την εμφάνισή τους (Computer Applications in Biosciences). Φυσικά, πρέπει να έχουμε στο μυαλό μας ότι την εποχή αυτή είχε αρχίσει να γίνεται διαδεδομένη η χρήση υπολογιστικών συστημάτων και έτσι πολλές από τις παρακάτω ανακαλύψεις ακολούθησαν και επωφελήθηκαν από την πρόοδο στον τομέα του υλικού και του λογισμικού (Εικόνα 1.4).

Στο πεδίο της ανάλυσης αλληλουχιών μακρομορίων, η μελέτη πάνω στους αλγόριθμους στοίχισης και στις αποτελεσματικές υλοποιήσεις τους συνεχίστηκε με εντατικό ρυθμό. Βασικό ρόλο έπαιξαν σε αυτή την πρόοδο η ανακάλυψη του αλγορίθμου τοπικής στοίχισης από τους Smith και Waterman το 1981, οι αλγόριθμοι προσεγγιστικού ταιριάσματος συμβολοσειρών, η μελέτη των στατιστικών ιδιοτήτων της στοίχισης από τους Aratia, Waterman και Karlin, αλλά και οι πρώτες αποτελεσματικές υλοποιήσεις για γρήγορη στοίχιση και αναζήτηση ομοιότητας σε μια βάση δεδομένων (FASTA). Παράλληλα έγιναν οι πρώτες θεωρητικές επεξεργασίες της πολλαπλής στοίχισης, επινοήθηκε η ιεραρχική πολλαπλή στοίχιση και παρουσιάστηκε το CLUSTAL. Ιδιαίτερα σημαντική επινόηση αυτής της περιόδου ήταν τα προφίλ αλληλουχιών (sequence profiles) τα οποία αποτέλεσαν πανίσχυρο εργαλείο στη μελέτη των πρωτεϊνικών οικογενειών, εφαρμόστηκαν σε πάρα πολλά παραδείγματα με εντυπωσιακά αποτελέσματα και εξακολουθούν να χρησιμοποιούνται μέχρι σήμερα. Τέλος, πρέπει να σημειώσουμε ότι την εποχή αυτή εμφανίστηκαν και τα πρώτα βιβλία σχετικά με την υπολογιστική ανάλυση αλληλουχιών.

Η πρόοδος στις μεθόδους αλληλούχισης DNA, η εμφάνιση της PCR, αλλά και η ραγδαία βελτίωση στις τεχνικές προσδιορισμού της τρισδιάστατης δομής των μακρομορίων, οδήγησαν την εποχή αυτή στη

ραγδαία αύξηση του όγκου των δεδομένων και στη δημιουργία μεγαλύτερων και πιο οργανωμένων βάσεων βιολογικών δεδομένων (φυσικά, και αυτή η δραστηριότητα αναπτύχθηκε παράλληλα με τις εξελίξεις στα πληροφοριακά συστήματα και τα συστήματα βάσεων δεδομένων). Έτσι, το 1986 έκαναν την εμφάνιση τους οι δύο πιο γνωστές μέχρι σήμερα βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών (GenBank και EMBL Data Library), ενώ η SwissProt, η βάση δεδομένων των πρωτεϊνικών αλληλουχιών εμφανίστηκε το 1987. Την ίδια εποχή έκαναν την εμφάνισή τους προτάσεις για δημιουργία δικτύων που θα διευκόλυναν την υπολογιστική έρευνα στη βιολογία (EMBnet και BIONET), ενώ εμφανίστηκαν και οι πρώτοι κατάλογοι με σχετικό λογισμικό (LiMB). Τέλος, οι ερευνητικοί οργανισμοί όπως το NIH και το EMBL άρχισαν τη δημιουργία εξειδικευμένων τμημάτων αφιερωμένων στην έρευνα στην υπολογιστική βιολογία.



Εικόνα 1.4: *Η εξέλιξη των ιδεών με τις μεγαλύτερες ανακαλύψεις σχετικές με τη βιοπληροφορική στη δεκαετία του 1980.*

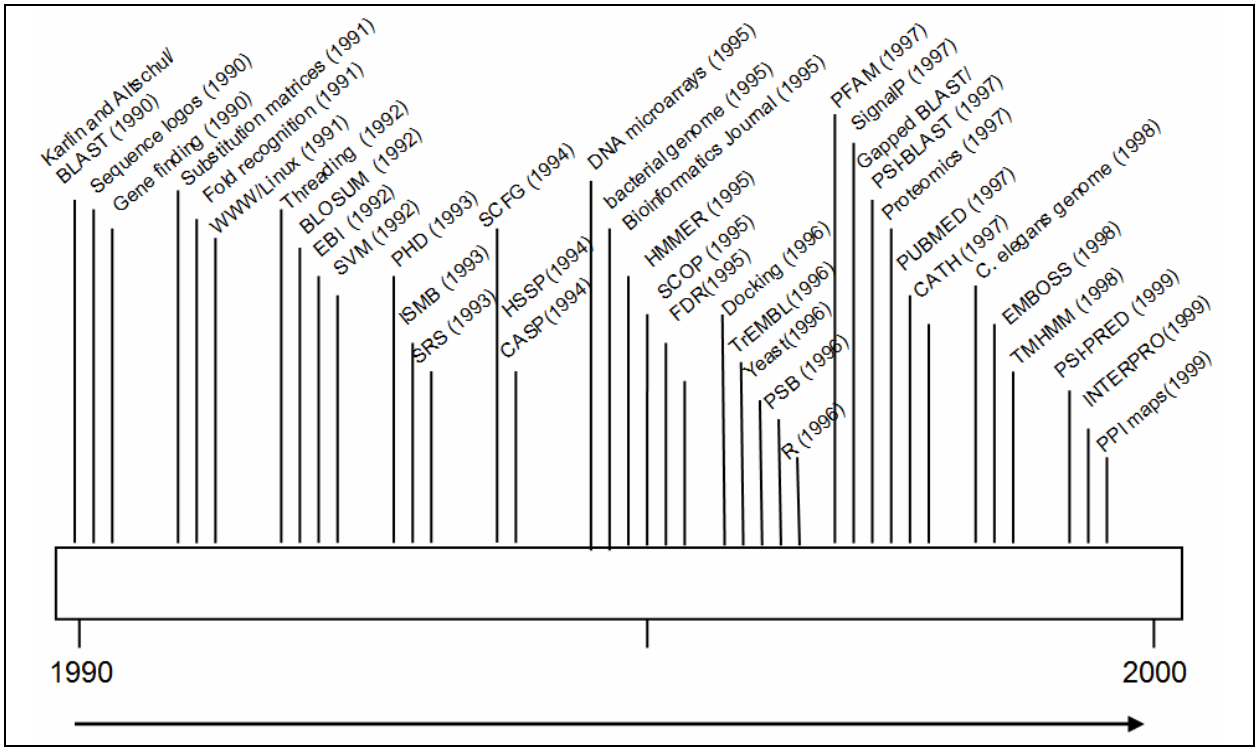
Στον τομέα της ανάλυσης και πρόγνωσης της δομής των πρωτεϊνών, έγιναν επίσης σημαντικές εξελίξεις. Η κρυσταλλογραφία συνέχισε να βελτιώνεται, και εμφανίστηκε και το NMR ενώ οι μεθοδολογίες αναπαράστασης μοριακών τρισδιάστατων δομών εξακολούθησαν να εξελίσσονται παράλληλα με τις εξελίξεις στον τομέα των γραφικών και της υπολογιστικής γεωμετρίας. Με την αύξηση των δομών, αλλά και την εξάπλωση των αλγόριθμων στοίχισης, εμφανίστηκαν και οι πρώτες προσπάθειες αυτόματης προτυποποίησης πρωτεϊνικών δομών με βάση την ομολογία (homology modelling). Η αύξηση των πρωτεϊνικών δομών και η κατάταξή τους σε οικογένειες και διπλώματα (δομικά μοτίβα), οδήγησε και στις πρώτες προσπάθειες μελέτης των κατηγοριών πρωτεϊνικού διπλώματος και προσπάθειες πρόγνωσής του. Επιπλέον, οι μελέτες των δομών κατέληξαν στο πολύ σημαντικό συμπέρασμα ότι οι δομές των πρωτεϊνών συντηρούνται περισσότερο από ότι οι αλληλουχίες τους. Οι αλγόριθμοι πρόγνωσης δευτεροταγούς δομής συνέχισαν να εξελίσσονται, δεχόμενες και τη βοήθεια νέων εξελίξεων στην τεχνητή νοημοσύνη (νευρωνικά δίκτυα), ενώ οι πρώτες κρυσταλλικές δομές μεμβρανικών πρωτεϊνών έδωσαν το έναυσμα για την ανάπτυξη των πρώτων μεθόδων ανάλυσης της αλληλουχίας των πρωτεϊνών αυτών (διαγράμματα υδροφοβικότητας, υδροφοβικές ροπές, positive inside rule) ενώ επίσης έκαναν την εμφάνιση τους και οι πρώτοι αλγόριθμοι πρόγνωσης για τις πρωτεΐνες αυτές. Παρόμοια με τις πρωτεΐνες, εξαπλώθηκαν και οι αλγόριθμοι πρόγνωσης της δευτεροταγούς δομής του RNA.

Στον τομέα της φυλογενετικής ανάλυσης, την εποχή αυτή προτάθηκε ο αλγόριθμος του Fenselstein για την εκτίμηση φυλογενετικών δέντρων μέσω της μέγιστης πιθανοφάνειας, μια πολύ σημαντική ανακάλυψη που έδωσε ώθηση στο αντίστοιχο πεδίο (ενώ παράλληλα αναπτύχθηκαν και πολλά από τα γνωστά μέχρι σήμερα μαθηματικά μοντέλα για την αντικατάσταση βάσεων σε φυλογενετικές μελέτες). Την ίδια εποχή

έγιναν σημαντικές συνεισφορές και στη μελέτη των εξελικτικών σχέσεων των πρωτεϊνών (μιλήσαμε ήδη για την ανακάλυψη ότι οι δομές συντηρούνται περισσότερο από την αλληλουχία). Μελετήθηκαν οι στατιστικές ιδιότητες των πινάκων αντικατάστασης αμινοξέων (PAM), μελετήθηκε έντονα το φαινόμενο της ομολογίας, αλλά και περιπτώσεις ομοιότητας λόγω συγκλίνουσας εξέλιξης, ενώ έγιναν πολλές μελέτες της εξελικτικής ιστορίας συγκεκριμένων πρωτεϊνικών οικογενειών οι οποίες είχαν ευρύτερη σημασία στη βιολογία (π.χ. ανοσοσφαιρίνες, πρωτεάσες, κυτοχρώματα, ριβονουκλεάσες κ.ο.κ.). Τέλος, έγιναν σημαντικά βήματα στην εξελικτική μελέτη των γονιδιωμάτων καθώς μελετήθηκαν οι φυλογενετικοί δείκτες όπως το rRNA, αλλά και η εξελικτική ιστορία των εσωνίων, των εξωνίων και της συρραφής.

1.1.4. Η δεκαετία του 1990

Η δεκαετία του 1990 ήταν η δεκαετία κατά την οποία η έρευνα στην υπολογιστική βιολογία εκτινάχθηκε (θα δούμε παρακάτω και εμπειρικά μετρήσιμα δεδομένα για αυτό). Φυσικά, για άλλη μια φορά δεν πρέπει να αμελήσουμε να αναφέρουμε ότι η δεκαετία αυτή σημαδεύτηκε επίσης από την ανάπτυξη του διαδικτύου και του παγκοσμίου ιστού, αλλά και από την εξάπλωση των προσωπικών Η/Υ (Εικόνα 1.5). Πρέπει να σημειώσουμε επίσης, ότι η ευρεία χρήση του όρου «βιοπληροφορική» συντελέστηκε μέσα στη δεκαετία του 1990. Ενδεικτικά, το πιο γνωστό περιοδικό του χώρου, το Bioinformatics, πήρε το όνομα αυτό το 1995 αλλάζοντας το προηγούμενο όνομα «Computer Applications in the Biosciences» (CABIOS).



Εικόνα 1.5: *Η εξέλιξη των ιδεών με τις μεγαλύτερες ανακαλύψεις σχετικές με τη βιοπληροφορική στη δεκαετία του 1990*

Η δεκαετία αυτή σημαδεύτηκε από μια σειρά μεγάλων ανακαλύψεων. Στον τομέα της στοίχισης αλληλουχιών, η πιο σημαντική ίσως δημοσίευση όλων των εποχών στο χώρο, αφορά το BLAST (Basic Local Alignment Search Tool), από επιστήμονες του NCBI το 1990. Το BLAST «πάτησε» πάνω στις ανακαλύψεις για τη στατιστική κατανομή του σκορ (score) της τοπικής στοίχισης (το γνωστό θεώρημα Karlin-Altschul) και πραγματικά ήταν μια επαναστατική συμβολή στον τρόπο που θα διεξάγεται από κει και πέρα η αναζήτηση ομοιότητας σε βάσεις δεδομένων και η στοίχιση, καθώς ήταν πιο γρήγορο από κάθε άλλο αλγόριθμο επιτρέποντας ταχείες αναζητήσεις, αλλά έδινε και για πρώτη φορά μια εκτίμηση για τη στατιστική σημαντικότητα των στοιχίσεων. Στην πρώτη του έκδοση δεν παρήγαγε στοιχίσεις με κενά, αλλά στη δεύτερη, παρείχε και αυτή τη δυνατότητα, ενώ περιλάμβανε και άλλες παραλλαγές όπως το PSI-BLAST. Επιπλέον, τα

προγράμματα πολλαπλής στοίχισης έκαναν τη δυναμική τους εμφάνιση (CLUSTAL) με εκδόσεις για μαζική χρήση σε Η/Υ, δίνοντας ακόμα και εκδόσεις για παραθυρικό περιβάλλον.

Στον τομέα της ανάλυσης των πρωτεϊνικών δομών και της πρόγνωσης έγιναν επίσης μεγάλες ανακαλύψεις. Εμφανίστηκαν τα πρώτα προγράμματα ευρείας χρήσης για την οπτικοποίηση και την ανάλυση πρωτεϊνικών δομών, όπως το Rasmol και το Kinemage, ενώ έγιναν και οι πρώτες επιτυχημένες προσπάθειες για ύφανση πρωτεϊνών (threading), αλλά και για αγκυροβόληση (docking) πρωτεϊνικών δομών. Στον τομέα της πρόγνωσης της δευτεροταγούς δομής, η χρήση νευρωνικών δικτύων παράλληλα με τη χρήση εξελικτικής πληροφορίας με τη μορφή πολλαπλών στοιχίσεων, έδωσε για πρώτη φορά ποσοστά επιτυχίας πάνω από 70% και άνοιξε ολόκληρες λεωφόρους στη μελέτη των αλγορίθμων πρόγνωσης με εφαρμογές και σε πλήθος άλλων περιπτώσεων (διαμεμβρανικές πρωτεΐνες, προσβασιμότητα του διαλύτη, κ.ο.κ.). Παράλληλα, ξεκίνησαν και οι διαγωνισμοί/συνέδρια του CASP.

Οι πρώτες επιτυχημένες προσπάθειες αλληλούχισης ολόκληρων γονιδιωμάτων, πρώτα βακτηρίων και στη συνέχεια και ευκαρυωτικών οργανισμών, άνοιξαν επίσης νέους δρόμους στη συγκριτική γονιδιωματική, ενώ πυροδότησαν και την ανάπτυξη των πρώτων αλγορίθμων εύρεσης γονιδίων (gene finders). Η δεκαετία αυτή, σηματοδότησε επίσης την εμφάνιση των μικροσυστοιχιών DNA για τη μέτρηση της γονιδιακής έκφρασης, τεχνολογία που είχε, όπως θα δούμε στη συνέχεια, μεγάλη επίδραση τόσο στη βιοπληροφορική όσο και στην ιατρική πληροφορική και τη βιοστατιστική, και σηματοδότησε την απαρχή της λειτουργικής γονιδιωματικής.

Στον τομέα των βάσεων δεδομένων, η εκθετική αύξηση των δεδομένων όλων των κατηγοριών συνεχίστηκε και μια σειρά νέες βάσεις δεδομένων αναπτύχθηκαν. Ανάμεσά τους ήταν βάσεις με δομικές ταξινομήσεις των πρωτεϊνών (όπως η SCOP και η CATH), αλλά και βάσεις με ταξινομήσεις βασισμένες σε χαρακτηριστικά πρότυπα (patterns) της ακολουθίας, όπως η PROSITE, η PFAM και τελικά η INTERPRO. Επίσης, μια πολύ σημαντική εξέλιξη αυτής της περιόδου, ήταν η ίδρυση του EBI (European Bioinformatics Institure), του μεγαλύτερου ινστιτούτου βιοπληροφορικής της Ευρώπης, το οποίο ιδρύθηκε στη Μεγάλη Βρετανία (Hinxton) το 1992 μέσα από μια κοινοπραξία του EMBL και του Welcome Trust. Στο EBI στεγάστηκαν αρχικά οι βάσεις δεδομένων του EMBL, EMBL-Bank και SwissProt-TrEMBL και δημιουργήθηκαν ερευνητικές ομάδες για να συνδράμουν στα διάφορα γονιδιωματικά προγράμματα εκείνης της εποχής, ενώ λίγο αργότερα λειτούργησε και η TrEMBL. Τέλος, το 1993 ξεκίνησαν τα συνέδρια ISMB και λίγα χρόνια αργότερα ιδρύθηκε η ISCB.

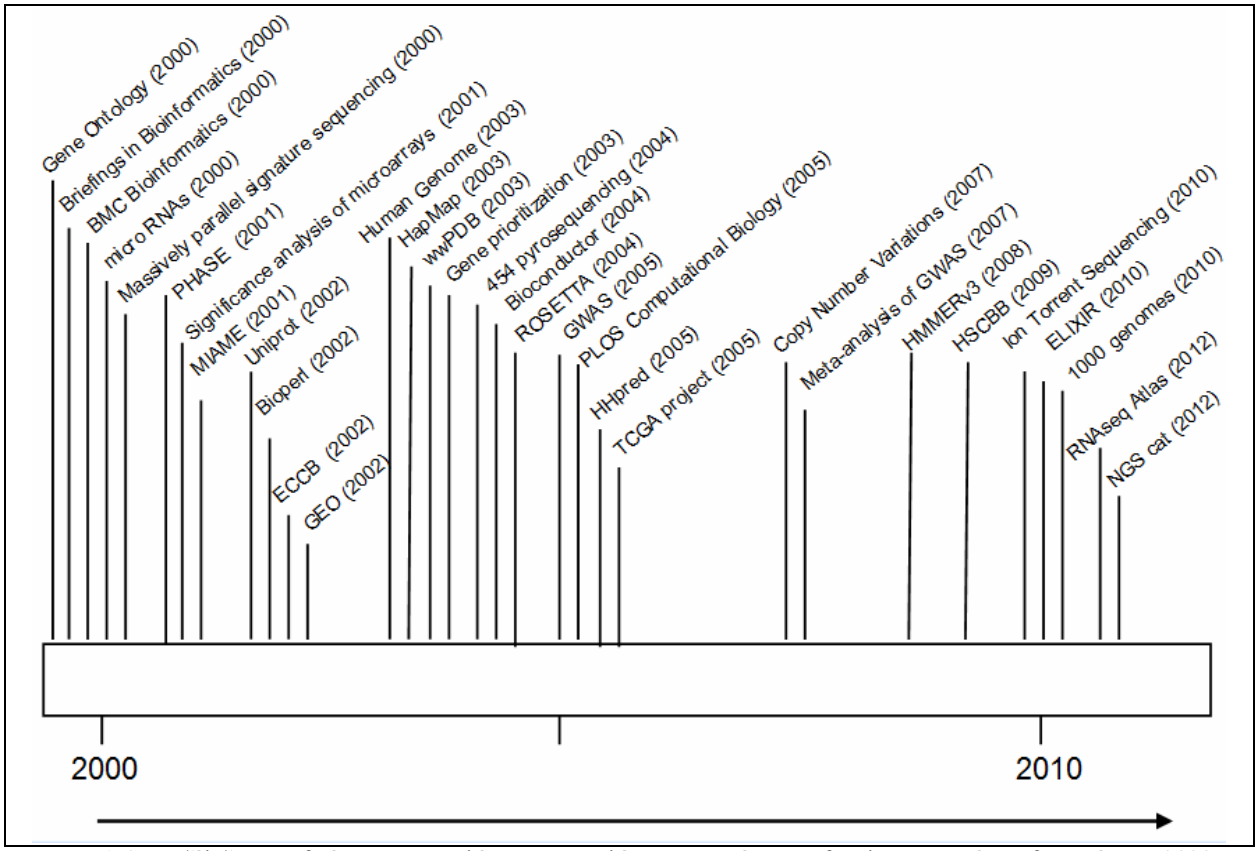
Τέλος, τη δεκαετία αυτή έκανε την εμφάνισή της, μετά τις πρωτοποριακές εργασίες των Krogh, Eddy, Hughey κλπ, και μια μεθοδολογία που θα επικρατούσε τα επόμενα χρόνια στην ανάλυση αλληλουχιών, το Hidden Markov Model (HMM), το οποίο βρήκε εφαρμογές τόσο στην μοντελοποίηση των πολλαπλών στοιχίσεων και την αναζήτηση μακρινών ομοιοτήτων, όσο και στις μεθόδους πρόγνωσης. Το γνωστό πακέτο HMMER για πολλαπλές στοιχίσεις και αναζητήσεις μακρινών ομολόγων με profile HMM, πάνω στο οποίο βασίζεται η βάση δεδομένων πρωτεϊνικών οικογενειών PFAM, έκανε την εμφάνιση του εκείνη την περίοδο, ενώ το ίδιο συνέβη και για δυο από τους πιο επιτυχημένους αλγόριθμους πρόγνωσης, το TMHMM για τις μεμβρανικές πρωτεΐνες, και το SignalP για τις σηματοδοτικές αλληλουχίες. Το HMM αξίζει μια ειδική αναφορά, γιατί παρόλο που σαν μαθηματική μέθοδος ήταν γνωστή από καιρό και είχε χρησιμοποιηθεί στην αναγνώριση ομιλίας, η υιοθέτησή του σε μεθόδους υπολογιστικής βιολογίας, έδωσε νέα πνοή και στην ίδια την αναζήτηση μεθοδολογίας καθώς μια σειρά από αλγόριθμους και τροποποιήσεις του μοντέλου εμφανίστηκαν ειδικά για τα προβλήματα της βιολογίας (το profile HMM, οι αλγόριθμοι για σημασμένες αλληλουχίες, αλλά και μια σειρά αλγόριθμοι εκπαίδευσης και αποκωδικοποίησης). Σήμερα, δεν νοείται κείμενο, ακόμα και εισαγωγικό, στη βιοπληροφορική που να μην περιγράφει το HMM.

1.1.5. Η «σύγχρονη» εποχή

Η εποχή μετά το 2000 σηματοδότησε την ώριμη φάση της υπολογιστικής βιολογίας, καθώς η διδασκαλία της έχει γίνει βασική πλέον και σε προπτυχιακό αλλά και σε μεταπτυχιακό επίπεδο. Παράλληλα, ιδρύονται και αναπτύσσονται επαγγελματικές οργανώσεις, κυκλοφορούν όλο και περισσότερα εξειδικευμένα επιστημονικά περιοδικά κ.ο.κ. Η ολοκλήρωση του προγράμματος προσδιορισμού του ανθρώπινου γονιδιώματος, μπορεί να μη δικαίωσε όλες τις αρχικές προσδοκίες («*θα βρούμε το φάρμακο για κάθε ασθένεια*») αλλά σίγουρα άνοιξε νέους δρόμους σε μια σειρά από κλάδους (Εικόνα 1.6). Για παράδειγμα, ο μεγάλος ρυθμός προσδιορισμού των γονιδιωμάτων, οδήγησε στην αλματώδη ανάπτυξη της γονιδιωματικής στις διάφορες μορφές της. Συγκριτική γονδιωματική για τη σύγκριση γονιδιωμάτων, λειτουργική γονιδιωματική για τις μελέτες

γονιδιακής έκφρασης με μικροσυστοιχίες και δομική γονιδιωματική για τη μαζική παραγωγή πρωτεϊνών για δομικές μελέτες και κρυσταλλογραφία ακτίνων Χ. Παράλληλα, εμφανίστηκαν οι τεχνικές αλληλούχισης νέας γενιάς (Next Generation Sequencing), οι οποίες έδωσαν νέα πνοή σε μελέτες γονιδιακής έκφρασης (RNAseq), έκαναν εύκολο τον εντοπισμό πολυμορφικών θέσεων και αναμένεται να επηρεάσουν και την προσωποποιημένη ιατρική. Με όλα αυτά τα δεδομένα δημιουργήθηκε μια μεγάλη ανάγκη για περισσότερους αλγόριθμους πρόγνωσης έτσι ώστε να τεθεί 'σε τάξη' ο τεράστιος αυτός όγκος δεδομένων, αλλά και μια μεγάλη ανάγκη για δημιουργία εξειδικευμένων βάσεων δεδομένων και οντολογιών που να τις περιγράφουν. Την εποχή αυτή, είδαμε την έκρηξη των διαδικτυακών εφαρμογών (web-servers), αλλά και του ελεύθερου λογισμικού βιοπληροφορικής. Επιπλέον, τα προγράμματα μετα-γονδιωματικής (meta-genomics) έθεσαν νέα αλγοριθμικά προβλήματα στην εύρεση γονιδίων και την ομαδοποίηση δεδομένων.

Ειδικά στις βάσεις δεδομένων, μέσα στη δεκαετία του 2000, εκτός από την ανάπτυξη μικρών εξειδικευμένων βάσεων δεδομένων, είδαμε και τις πρώτες συγχωνεύσεις των μεγάλων βάσεων δεδομένων, καθώς η SwissProt και η PIR ένωσαν τις προσπάθειές τους σε μια προσπάθεια να ανταπεξέλθουν στον τεράστιο όγκο δεδομένων, σχηματίζοντας την Uniprot (η οποία πλέον στεγάζεται στο EBI). Οι προβλέψεις για το μέλλον είναι κάπως δυσοίωνες, καθώς με τη συνεχόμενη εκθετική αύξηση των δεδομένων, σε λίγα χρόνια θα υπάρχει μεγάλο πρόβλημα αποθήκευσης και διαμοιρασμού των δεδομένων, γι' αυτό και έχει ξεκινήσει από το EBI η πρωτοβουλία του ELIXIR να διανεμηθούν, κατά κάποιον τρόπο, οι δημόσια διαθέσιμες βάσεις σε διάφορες χώρες και φορείς.



Εικόνα 1.6: *Η εξέλιξη των ιδεών με τις μεγαλύτερες ανακαλύψεις σχετικές με τη βιοπληροφορική στη δεκαετία του 2000.*

Η γνώση της αλληλουχίας του ανθρώπινου γονιδιώματος έδωσε επίσης μεγάλη ώθηση στη Γενετική Επιδημιολογία, καθώς με τον εντοπισμό εκατομμυρίων πολυμορφισμών (SNPs) και τη χρήση της τεχνολογίας των GWAS (Genome-Wide Association Studies), μας δόθηκε η δυνατότητα να κάνουμε μαζικά μελέτες γενετικής συσχέτισης, μελετώντας ταυτόχρονα εκατομμύρια πολυμορφισμούς σε μαζική κλίμακα, όπως ακριβώς και με τις μικροσυστοιχίες DNA. Παράλληλα, έγιναν μελέτες για την απλοτυπική σύσταση και την προέλευση των ανθρώπινων πληθυσμών, την κατανομή τους, το βαθμό ανασυνδυασμού κλπ (HapMap project), ενώ μεγάλη αποδοχή έχουν λάβει οι μεθοδολογίες μετα-ανάλυσης και ενοποίησης δεδομένων (data

integration). Οι περιοχές αυτές, είναι περιοχές που πλέον η βιοπληροφορική έρχεται σε επαφή με την Γενετική και την Επιδημιολογία και τη βιοστατιστική. Επιπλέον δε, η γνώση της αλληλουχίας του γονιδιώματος, επηρέασε και άλλους κλάδους όπως την πρωτεομική (Proteomics), ενώ η λεπτομερής μελέτη του, έδωσε και νέες βιολογικές ανακαλύψεις, όπως τα μη κωδικά RNA (ncRNA), για τα οποία αναπτύχθηκαν μια σειρά αλγόριθμων και μεθοδολογιών για την πρόγνωση και τη μελέτη τους, αλλά και των μεγάλων επαναληπτικών αλληλουχιών (Copy Number Variations- CNVs). Και οι δύο περιπτώσεις, δεν ήταν προηγουμένως γνωστές και πλέον βρίσκονται στο επίκεντρο της μοριακής έρευνας.

Αλγοριθμικά, εμφανίστηκε δυναμικά η χρήση των Support Vector Machines (SVMs) σε προβλήματα πρόγνωσης, μια μεθοδολογία που αποδείχθηκε πιο αποτελεσματική από τα νευρωνικά δίκτυα σε κάποιες περιπτώσεις. Επίσης, παρουσιάστηκαν ολοκληρωμένοι αλγόριθμοι για σύγκριση HMM-HMM, αλλά και η νέα έκδοση του HMMER η οποία στηρίζεται σε μια σειρά θεωρητικές ανακαλύψεις που βελτιώνουν δραματικά την ταχύτητά του και φιλοδοξούν πλέον να το καταστήσουν αντικαταστάτη του BLAST. Επίσης, έγιναν μεγάλες πρόοδοι στην ab initio πρόγνωση δομής πρωτεϊνών. Τέλος, τη δεκαετία αυτή, ακολουθώντας την τεράστια αύξηση των βάσεων δεδομένων και των οντολογιών, έκανε την εμφάνιση της η βιολογία συστημάτων, η οποία μελετάει πλέον πολύπλοκα δίκτυα με τις αλληλεπιδράσεις των μερών τους, αντί για μεμονωμένες οντότητες, αλλά και οι οντολογίες. Τέτοια δίκτυα είναι τα δίκτυα πρωτεϊνικών αλληλεπιδράσεων, τα ρυθμιστικά δίκτυα, τα τροφικά δίκτυα κλπ. Στην ανάπτυξη αυτή, εκτός από την ποσότητα των δεδομένων και την αυξημένη διαθέσιμη υπολογιστική ισχύ, σημαντικό ρόλο έπαιξαν και οι εξελίξεις στη μαθηματική θεωρία των γράφων και στη θεωρητική πληροφορική.

Στα επόμενα χρόνια, αναμένεται ο ρόλος και η μορφή της Υπολογιστικής Βιολογίας να αλλάξει όλο και περισσότερο, ακολουθώντας τις ραγδαίες εξελίξεις της τεχνολογίας και την ολοένα μεγαλύτερη συσσώρευση μοριακών δεδομένων. Δεν μπορούμε να προβλέψουμε ακριβώς ποια θα είναι η μορφή αυτή, αλλά σίγουρα οι εξελίξεις στην προσωποποιημένη ιατρική, στην αλληλούχιση, στη νανοτεχνολογία, στο βιολογικό υπολογισμό αλλά και στην ίδια την επιστήμη υπολογιστών, θα επηρεάσουν και τον τρόπο που η Υπολογιστική Βιολογία προσεγγίζει τα πράγματα και εξερευνά νέα μονοπάτια (Ouzounis, 2012).

1.2. Η διεπιστημονικότητα της βιοπληροφορικής

Όπως έγινε ελπίζω κατανοητό από τα προηγούμενα η βιοπληροφορική ή, καλύτερα, η υπολογιστική βιολογία είναι διεπιστημονικός κλάδος που έλκει την καταγωγή του από τη μοριακή βιολογία και ιδιαίτερα από τη μελέτη των βιολογικών αλληλουχιών και των δομών. Ο ορισμός που δώσαμε, είναι περισσότερο συμβατός με τον ορισμό που δίνει το NCBI (http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/bioinformatics.html):

«Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information»

ή, τον αντίστοιχο ορισμό του Luscombe (Luscombe, Greenbaum, & Gerstein, 2001):

«Bioinformatics is conceptualizing biology in terms of macromolecules (in the sense of physical-chemistry) and then applying "informatics" techniques (derived from disciplines such as applied maths, computer science, and statistics) to understand and organize the information associated with these molecules, on a large-scale»

και αυτόν του Fredj Tekaia:

«The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information».

αλλά υπάρχουν και ορισμοί που διαφωνούν ριζικά, ιδιαίτερα ορισμοί που κάνουν σαφή διάκριση μεταξύ της βιοπληροφορικής (διαχείριση μεγάλου όγκου δεδομένων και βάσεων δεδομένων) και της υπολογιστικής βιολογίας (ανάπτυξη αλγόριθμων και μεθοδολογιών). Για παράδειγμα ο Richard Durbin λέει ότι:

«I do not think all biological computing is bioinformatics, e.g. mathematical modelling is not bioinformatics, even when connected with biology-related problems. In my opinion, bioinformatics has to do with management and the subsequent use of biological information, particular genetic information».

Όπως και να έχει, η International Society for Computational Biology (ISCB) αποφεύγει τους ορισμούς, και αυτοπροσδιορίζεται, κάπως πιο γενικά, ως:

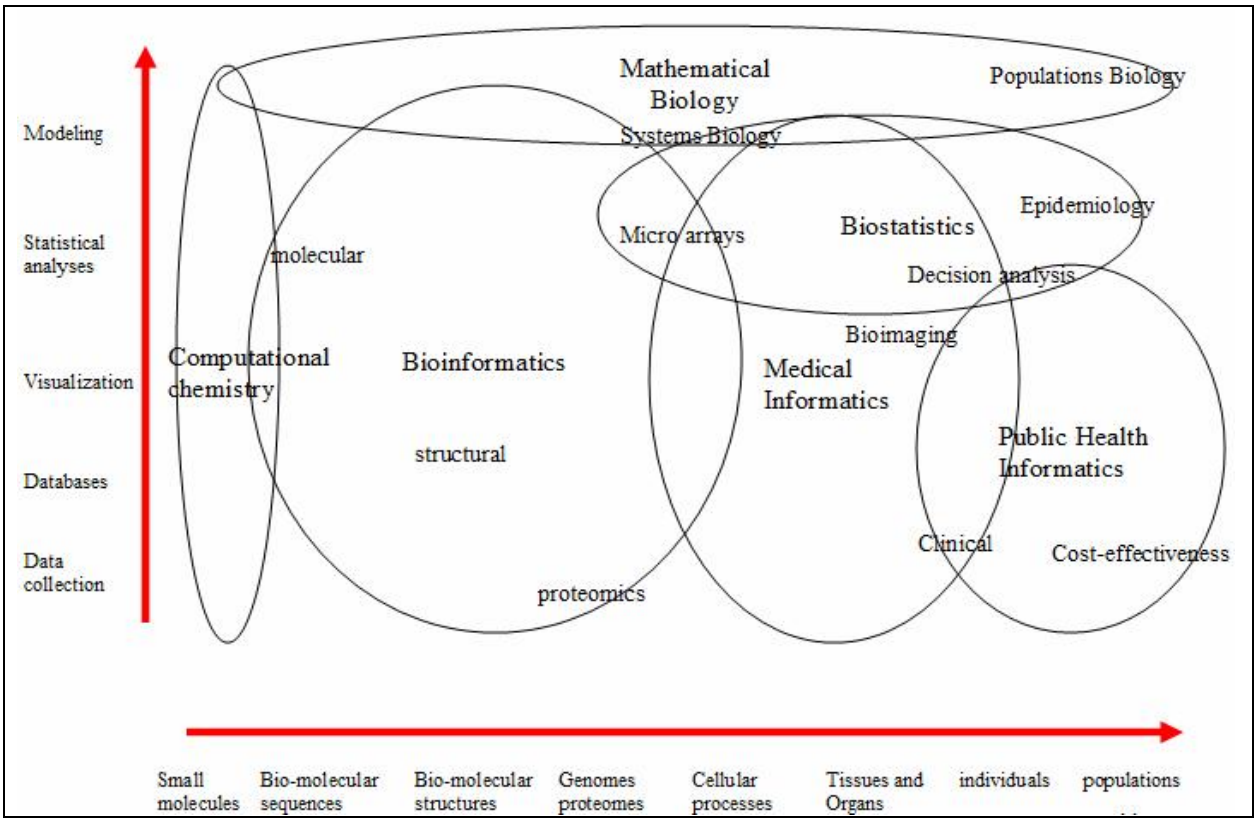
«a scholarly society dedicated to advancing the scientific understanding of living systems through computation».

Μερικοί, πάνε αυτόν τον ορισμό ακόμα πιο μακριά και δέχονται ότι και δραστηριότητες που σήμερα ταξινομούνται στην ιατρική πληροφορική και τη βιοϊατρική τεχνολογία, όπως η ανάλυση ιατρικών εικόνων και η διαχείριση του ιατρικού φακέλου του ασθενούς, ανήκουν στη βιοπληροφορική. Γενικά πάντως, οι περισσότεροι εξακολουθούν να δέχονται ότι κυρίως το είδος των δεδομένων (μοριακά), το πλήθος τους (μεγάλο), αλλά και η μεθοδολογία ανάλυσης, ορίζουν το χώρο της βιοπληροφορικής. Σε μια προσπάθεια να αποδώσουμε γραφικά κάτι τέτοιο, μπορούμε (αν και είναι σίγουρο ότι πολλοί δεν θα συμφωνήσουν) να θεωρήσουμε απλουστευτικά δύο άξονες: τον άξονα που περιέχει το είδος των δεδομένων υπό ανάλυση και τον άξονα που περιέχει το είδος της ανάλυσης και να παραστήσουμε εκεί το χώρο, που περιέχει τις δραστηριότητες της βιοπληροφορικής αλλά και των συναφών επιστημών (Εικόνα 1.7).

Με μια τέτοια προσέγγιση, βλέπουμε τις περιοχές επαφής και αλληλεπικάλυψης, μικρές ή μεγάλες, με τις γειτονικές συναφείς ειδικότητες. Βλέπουμε λοιπόν ότι στην περιοχή των μικρών μορίων (φαρμάκων κλπ), η βιοπληροφορική/υπολογιστική βιολογία εφάπτεται με την υπολογιστική χημεία, ενώ στην περιοχή της μοντελοποίησης, πολλές φορές ταυτίζεται με την μαθηματική βιολογία. Μια μεγάλη περιοχή επικάλυψης υπάρχει με την ιατρική πληροφορική στην περιοχή της μελέτης των κυτταρικών διεργασιών και των δεδομένων γονιδιακής έκφρασης από μικροσυστοιχίες DNA, επικάλυψη που θα δούμε ότι επιβεβαιώνεται και από εμπειρικά δεδομένα της βιβλιογραφίας. Επίσης, η ίδια περιοχή, όπως και η περιοχή της ανάλυσης των γενετικών διαφορών των ατόμων και της συσχέτισης των πολυμορφισμών με ασθένειες (GWAS), αποτελεί περιοχή επικάλυψης αλλά και σύγκλισης της βιοπληροφορικής με τη βιοστατιστική (Molenberghs, 2005). Προφανώς, αν προσθέταμε και άλλους άξονες στο διάγραμμα αυτό, όπως π.χ. το σκοπό της μελέτης, θα μπορούσαμε να «διαχωρίσουμε» καλύτερα τις ειδικότητες. Όπως και να έχει, μιλάμε για μια υπεραπλουστευμένη ανάλυση, η οποία εν τούτοις μας δίνει κάποια χρήσιμα στοιχεία.

Η διεπιστημονικότητα της βιοπληροφορικής, είναι επίσης μια έννοια με μεγάλες συζητήσεις και διχογνωμίες γύρω από αυτήν, καθώς επηρεασμένοι από την ανάγκη να δώσουν αποτελέσματα τα μεγάλα συνεργατικά προγράμματα (όπως το πρόγραμμα προσδιορισμού του ανθρώπινου γονιδιώματος), πολλοί, μεταξύ των οποίων και μεγάλοι ερευνητικοί οργανισμοί, δίνουν έμφαση στο σχηματισμό διεπιστημονικών ομάδων αντί στην εκπαίδευση διεπιστημονικών ατόμων. Όπως αναφέρει και ο Eddy (Eddy, 2005), πολλοί από εμάς που ασχολούμαστε με τη βιοπληροφορική για χρόνια, αντιμετωπίζουμε το ίδιο πρόβλημα που αντιμετώπιζαν και οι πρώτοι μοριακοί βιολόγοι: δεν μπορούμε να ταξινομηθούμε εύκολα στις «παραδοσιακές» ειδικότητες. Αναφορικά με την προσωπική μου διαδρομή στο χώρο και φυσικά χωρίς να επιχειρώ να την παραλληλισω με αυτή του Eddy, θα πρέπει να αναφερω οτι σπούδασα Βιολογία και έκανα μεταπτυχιακό στη βιοστατιστική (σε ένα διεπιστημονικό πρόγραμμα που αποτελούσε συνεργασία του Τμήματος Μαθηματικών και της Ιατρικής Σχολής), ενώ το διδακτορικό μου εκπονήθηκε σε Τμήμα Βιολογίας, αλλά με θέμα που είναι ξεκάθαρα θέμα βιοπληροφορικής («Πρόγνωση δομής και λειτουργίας μεμβρανικών πρωτεϊνών»). Το σύνολο του ερευνητικού μου έργου αφορά σε θέματα πρόγνωσης δομής και λειτουργίας πρωτεϊνών, κατασκευής βιολογικών βάσεων δεδομένων, ανάπτυξη αλγορίθμων για HMM, ανάπτυξη στατιστικής μεθοδολογίας για μετα-ανάλυση γενετικών δεδομένων και εφαρμογές σε σημαντικές ασθένειες. Για όλα τα παραπάνω, χρησιμοποιώ τις βιολογικές γνώσεις μου, τις μαθηματικές μου γνώσεις στο σχεδιασμό αλγορίθμων και στατιστικών μεθόδων, ενώ στο τέλος κάποια από αυτά τα δημιουργήματά μου τα υλοποιώ σε κάποια γλώσσα προγραμματισμού. Παρόλο που σίγουρα υπάρχουν βιολόγοι με πολύ περισσότερες γνώσεις και δεξιότητες από μένα, στατιστικοί με καλύτερη θεωρητική κατάρτιση και

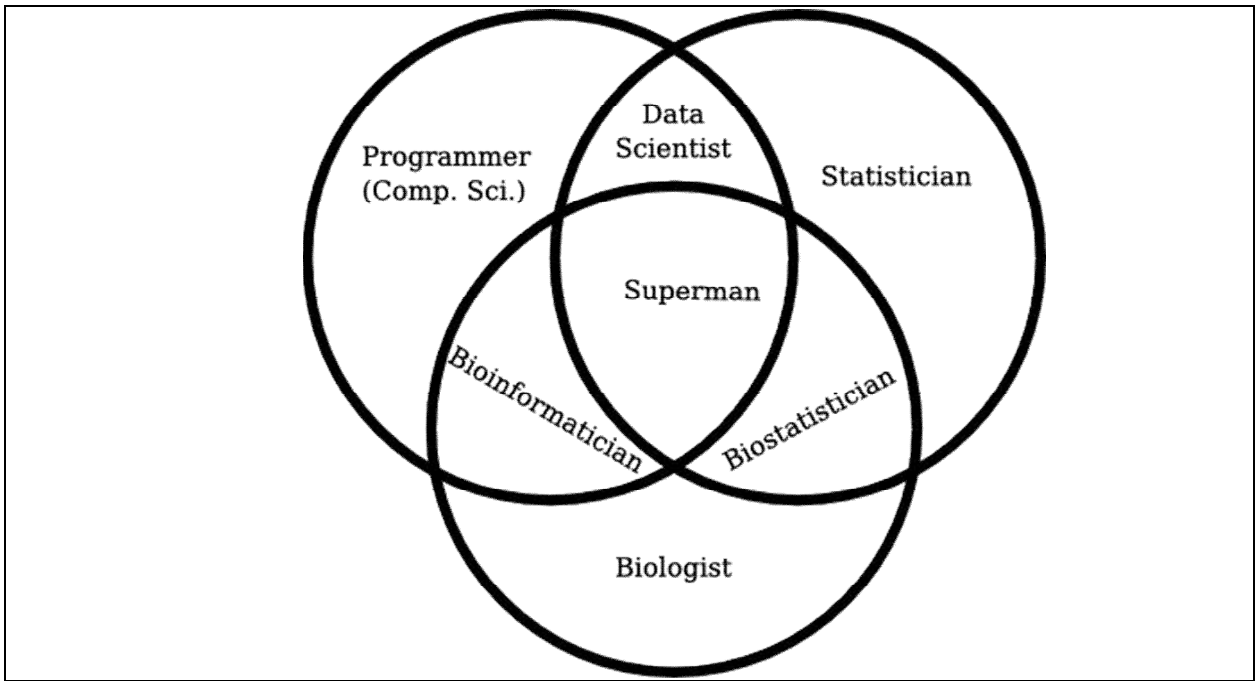
προγραμματιστές πολύ πιο αποδοτικοί, είμαι σύμφωνα με τα περισσότερα κριτήρια ένας σχετικά επιτυχημένος βιοπληροφορικός. Παρ' όλα αυτά, σίγουρα θα βρεθούν βιολόγοι που θα αμφισβητήσουν ότι αυτό που κάνω είναι βιολογία («*πάντα θα χρειάζεσαι ένα πείραμα*», βλ. παρακάτω) ή όπως είπε και ο Eddy «*I'm sure my union card has expired*» (Eddy, 2005). Οι μαθηματικοί από την άλλη θα πουν ότι δεν έχω βασικό πτυχίο στα μαθηματικά ή τη στατιστική και ότι δεν έχω αποδείξει πολλά θεωρήματα, ενώ όσον αφορά την πληροφορική, τα πράγματα είναι χειρότερα: παρόλο που έχω διδάξει προγραμματισμό για χρόνια, έχω δημοσιεύσει αλγόριθμους και το λογισμικό μου χρησιμοποιείται από επιστήμονες σε όλον τον κόσμο, δεν έχω ούτε ένα σχετικό πτυχίο, ούτε καν ECDL (αυτό βέβαια, δεν ξέρω αν λέει περισσότερο κάτι για τη δική μου αξία ή για αυτήν του ECDL).



Εικόνα 1.7: *Μια προσπάθεια απεικόνισης της θέσης της βιοπληροφορικής σε σχέση με τις συγγενικές επιστήμες.*

Το τελικό συμπέρασμα είναι ότι ναι μεν χρειάζεται σίγουρα διεπιστημονική συνεργασία διαφορετικών ειδικοτήτων, ιδιαίτερα στα πολύ μεγάλα και δύσκολα προβλήματα, αλλά σε έναν κλάδο που έχει αναπτύξει ήδη την δική του κουλτούρα, δυναμική και βιβλιογραφία, χρειάζεται πρώτα από όλα η εκπαίδευση διεπιστημονικών ατόμων, ατόμων που θα μπορούν να καταλάβουν τα βασικά από όλες τις «συνιστώσες» της βιοπληροφορικής, αλλά δεν είναι ανάγκη να είναι άριστοι και στις τρεις. Υπάρχουν πολλά ανέκδοτα περιστατικά, στα οποία μια διεπιστημονική ομάδα δεν μπόρεσε καν να συνεννοηθεί στα βασικά (είναι σα να στέλνεις αντιπροσώπους στον ΟΗΕ, διπλωμάτες που δεν μιλάνε ξένη γλώσσα, όπως είπε πάλι πολύ εύστοχα ο Eddy). Μια ιστορία που μου έχουν διηγηθεί αφορούσε μια ομάδα στατιστικών που πήγε να συνεργαστεί σε ένα μεγάλο πρόγραμμα με μοριακούς βιολόγους. Στην πρώτη συνάντηση, οι βιολόγοι μίλαγαν επί μία και πλέον ώρα «*τα κατάλοιπα (σ.σ. residues) αυτό, τα κατάλοιπα το άλλο*» κ.ο.κ. Μετά από πολλή ώρα, κάποιος από τους στατιστικούς ρώτησε «*Ωραία όλα αυτά, αλλά δεν καταλαβαίνω τι εννοείτε. Κάνατε κάποια παλινδρόμηση; Από που προήλθαν αυτά τα κατάλοιπα;*» (σ.σ. residues ή residuals λέγονται τα κατάλοιπα της παλινδρόμησης, οι διαφορές δηλαδή που προκύπτουν αν από τις τιμές που προκύπτουν από το μοντέλο της παλινδρόμησης, αφαιρεθούν οι παρατηρηθείσες τιμές της εξαρτημένης μεταβλητής – οι βιολόγοι από την άλλη εννοούσαν απλά τα αμινοξικά κατάλοιπα της πρωτεΐνης). Καταλαβαίνουμε έτσι, ότι κάποιος που ξεκίνησε βιολόγος, δεν είναι απαραίτητο να είναι και ο καλύτερος προγραμματιστής, ούτε να αποδεικνύει θεωρήματα (αλλά είναι απαραίτητο να μπορεί να καταλάβει τι είναι ένας αλγόριθμος, και να μπορεί να

γράψει 10 γραμμές κώδικα για να κάνει μια απλή ανάλυση). Όμοια, κάποιος που ξεκίνησε από την πληροφορική ή τα μαθηματικά, δεν είναι απαραίτητο να είναι γνώστης όλων των τελευταίων εξελίξεων και τεχνικών στη μοριακή βιολογία (αλλά πρέπει να μπορεί να καταλάβει τι είναι αμινοξύ, τι είναι κωδικόνιο, τι γονίδιο και τι πρωτεΐνη). Όπως αποτύπωσε με ένα διάγραμμα Venn ο Anthony Fejes, δεν είναι αναγκαίο να είναι κανείς υπεράνθρωπος για να είναι καλός βιοπληροφορικός (Εικόνα 1.8).

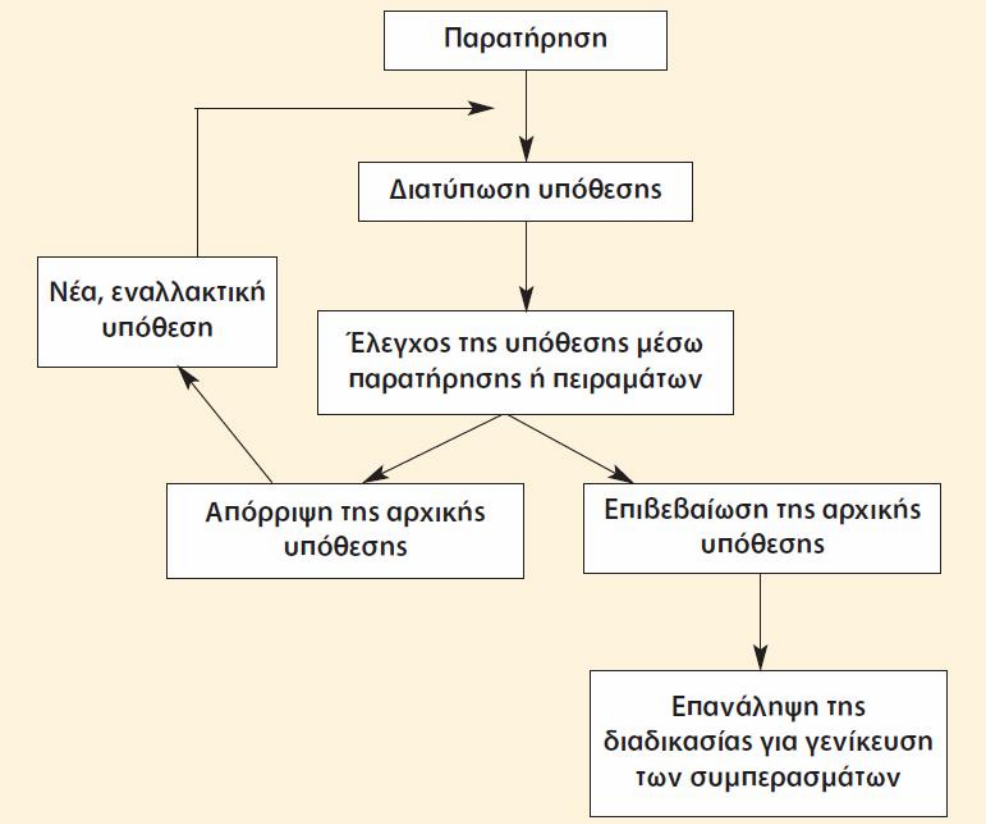


Εικόνα 1.8: *Από το http://blog.fejes.ca/?p=2418*

Ένα σημείο που πρέπει να διευκρινιστεί παράλληλα με τη διεπιστημονικότητα, είναι και το ίδιο το επιστημολογικό καθεστώς της βιοπληροφορικής (Ouzounis, 2002). Το θέμα μπορεί να γίνει κατανοητό, με ένα απλό παράδειγμα, που όμως δεν απέχει και πολύ από την πραγματικότητα. Ας υποθέσουμε λοιπόν ότι έχουμε έναν άψογα εκπαιδευμένο μοριακό βιολόγο και τον στείλουμε να σπουδάσει είτε σε προπτυχιακό είτε σε μεταπτυχιακό επίπεδο, πληροφορική. Αυτό θα τον κάνει αυτομάτως βιοπληροφορικό; Η απάντηση είναι ένα κατηγορηματικό όχι. Φυσικά, το ίδιο ισχύει, και ίσως για την ακρίβεια να είναι και χειρότερη η κατάσταση, για κάποιον μαθηματικό ή πληροφορικό που θα κληθεί να εκπαιδευτεί στη βιολογία. Ακριβώς όπως στην περίπτωση της διεπιστημονικής ομάδας ατόμων που είδαμε παραπάνω, έτσι και εδώ, υπάρχουν πολλά περισσότερα από μια απλή εκπαίδευση στις βασικές ειδικότητες, όσο καλή και επιτυχημένη και αν είναι αυτή. Η βιοπληροφορική, πέρα από το συνδυασμό γνώσεων από τις βασικές «συνιστώσες» της, έχει και το δικό της επιστημονικό βάθος. Έχει τη δική της ορολογία, τα δικά της προβλήματα, τις δικές της μεθοδολογίες, την ξεχωριστή της κουλτούρα, αλλά και τη δική της βιβλιογραφία που όπως είδαμε πηγαίνει 50 χρόνια πίσω. Όλοι οι παραπάνω παράγοντες, έχουν φυσικά τις ρίζες τους στις βασικές «συνιστώσες» (βιολογία, μαθηματικά, πληροφορική), οι οποίες μπορούν να εντοπιστούν ιστορικά, αλλά το γεγονός παραμένει, ότι η βιοπληροφορική διεκδικεί πλέον, και κατά τη γνώμη μου το έχει πετύχει, καθεστώς αυτόνομης επιστημονικής οντότητας, αναγνωρίζοντας φυσικά τις συγγένειες και τις εξαρτήσεις με τα άλλα πεδία. Επιπλέον, γίνεται κατανοητό με το παραπάνω παράδειγμα, ότι η πιο σωστή ονομασία θα ήταν Υπολογιστική Βιολογία, για να τονιστεί ακριβώς η έμφαση στο αντικείμενο της μελέτης (τα βιολογικά συστήματα), κατά ανάλογο τρόπο με τη μοριακή βιολογία. Αντίθετα, το όνομα βιοπληροφορική, παραπέμπει σε άλλους κλάδους όπως π.χ. τη βιοστατιστική, η οποία όμως είναι κατά βάση ειδικότητα της στατιστικής (με τις όποιες ιδιαιτερότητές της), ή τη γεωπληροφορική η οποία είναι απλά η εφαρμογή μεθόδων πληροφορικής σε προβλήματα χωροταξίας και χαρτογράφησης.

Το παραπάνω, είναι ένα κομβικό σημείο στην κατανόηση του επιστημολογικού πλαισίου της βιοπληροφορικής και την αντιμετώπισή της ως υπολογιστική βιολογία, και είναι κάτι που πολλές φορές δεν γίνεται κατανοητό ούτε από τους παραδοσιακούς βιολόγους, οι οποίοι επίσης αντιμετωπίζουν τη

βιοπληροφορική απλά σαν μια «εφαρμογή μεθόδων». Ένα κλασικό παράδειγμα, βρίσκεται σε μια φράση που όλοι όσοι ασχολούμαστε με τη βιοπληροφορική έχουμε λίγο πολύ ακούσει («*εντάξει, καλά είναι αυτά που μας λες, οι προγνώσεις, οι στοιχίσεις και όλα αυτά, αλλά πάντα θα χρειάζεσαι το πείραμα*»). Αυτή η φράση, μπορεί φυσικά να περιέχει αλήθεια σε πολλές περιπτώσεις, δεν μπορεί όμως να έχει καθολική εφαρμογή και δείχνει απλά μια προσκόλληση σε μια υπερβολικά απλουστευμένη, απλοϊκή και τελικά στρεβλή εκδοχή αυτού που ακόμα και στα σχολικά βιβλία βιολογίας αναφέρεται ως «επιστημονική μέθοδος». Το μοντέλο αυτό, καθώς είναι επηρεασμένο από το θετικισμό αλλά και τη διαψευσιοκρατία, θεωρείται από τις σύγχρονες προσεγγίσεις περί φιλοσοφίας της επιστήμης ως μη επαρκές θεωρητικά, αλλά παρ' όλα αυτά έχει δώσει μεγάλες επιτυχίες στη σύγχρονη βιολογία και καλώς χρησιμοποιείται. Τα θεωρητικά επιστημολογικά προβλήματα αυτής της προσέγγισης αφορούν κυρίως την εξάρτηση της παρατήρησης από τη θεωρία, την επισφάλεια όλων των πειραμάτων, αλλά και την τελική αδυναμία να δοθεί μια ξεκάθαρη απάντηση και μια μέθοδος για το πώς παράγεται τελικά μια ολοκληρωμένη θεωρία (Chalmers, 1999).



Εικόνα 1.9: *Σχήμα που απεικονίζει την επιστημονική μέθοδο (Μαυρικάκη, Γκούβρα, & Καμπούρη, 2014)*

Για αρχή, η παρατήρηση δεν είναι κάτι ουδέτερο και εξαρτάται από τη θεωρία και το ολοκληρωμένο σύστημα αξιών μέσα στο οποίο πραγματοποιείται. Λίγες φορές στις σύγχρονες επιστήμες η παρατήρηση είναι οπτική (αλλά ακόμα και τότε είναι επισφαλής), ενώ στις περισσότερες των περιπτώσεων η κατανόηση του τι παρατηρήθηκε απαιτεί την αποδοχή ενός συνόλου κανόνων, τεχνικών, ορισμάτων κ.ο.κ. Η «παρατήρηση» ότι μια συγκεκριμένη πρωτεΐνη έχει μια συγκεκριμένη τρισδιάστατη μορφή, απαιτεί την αποδοχή της τεχνικής της κρυσταλλογραφίας, της περίθλασης ακτίνων Χ, της επίλυσης του προβλήματος φάσης, της ανασύσταση της δομής κ.ο.κ. Το ίδιο φυσικά ισχύει και για άλλες «παρατηρήσεις», οι οποίες στην ουσία είναι οι ίδιες πειράματα (η αλληλούχιση, η PCR, η ηλεκτροφόρηση κ.ο.κ.). Το ένα πρόβλημα με αυτή τη θεώρηση, είναι ότι δεν μπορείς να παρατηρήσεις εύκολα κάτι που δεν ταιριάζει στο δικό σου σύστημα. Υπάρχουν αρκετά τέτοια παραδείγματα στην ιστορία της μοριακής βιολογίας (π.χ. τα δεδομένα κρυσταλλογραφίας ακτίνων Χ με τα οποία οι Watson και Crick προσδιόρισαν τη δομή του DNA ήταν διαθέσιμα από καιρό αλλά δεν μπορούσαν να αξιοποιηθούν). Το άλλο πρόβλημα, είναι ότι τα δεδομένα αυτά

είναι από τη φύση τους επισφαλή. Οι τεχνικές έχουν σφάλματα, υπόκεινται σε πειραματικό λάθος και είναι κάθε άλλο παρά τέλειες. Υπάρχουν επίσης πολλά παραδείγματα όπου ένα πείραμα δεν εκτελέστηκε σωστά (π.χ. έγινε μια επιμόλυνση της καλλιέργειας) ή, ακόμα χειρότερα, ένα μικρό σφάλμα στην όλη διαδικασία (π.χ. ένα μικρό τυπογραφικό λάθος σε ένα από τα πολλά προγράμματα κρυσταλλογραφίας) οδήγησε σε τρισδιάστατες δομές που ήταν τελείως λάθος.

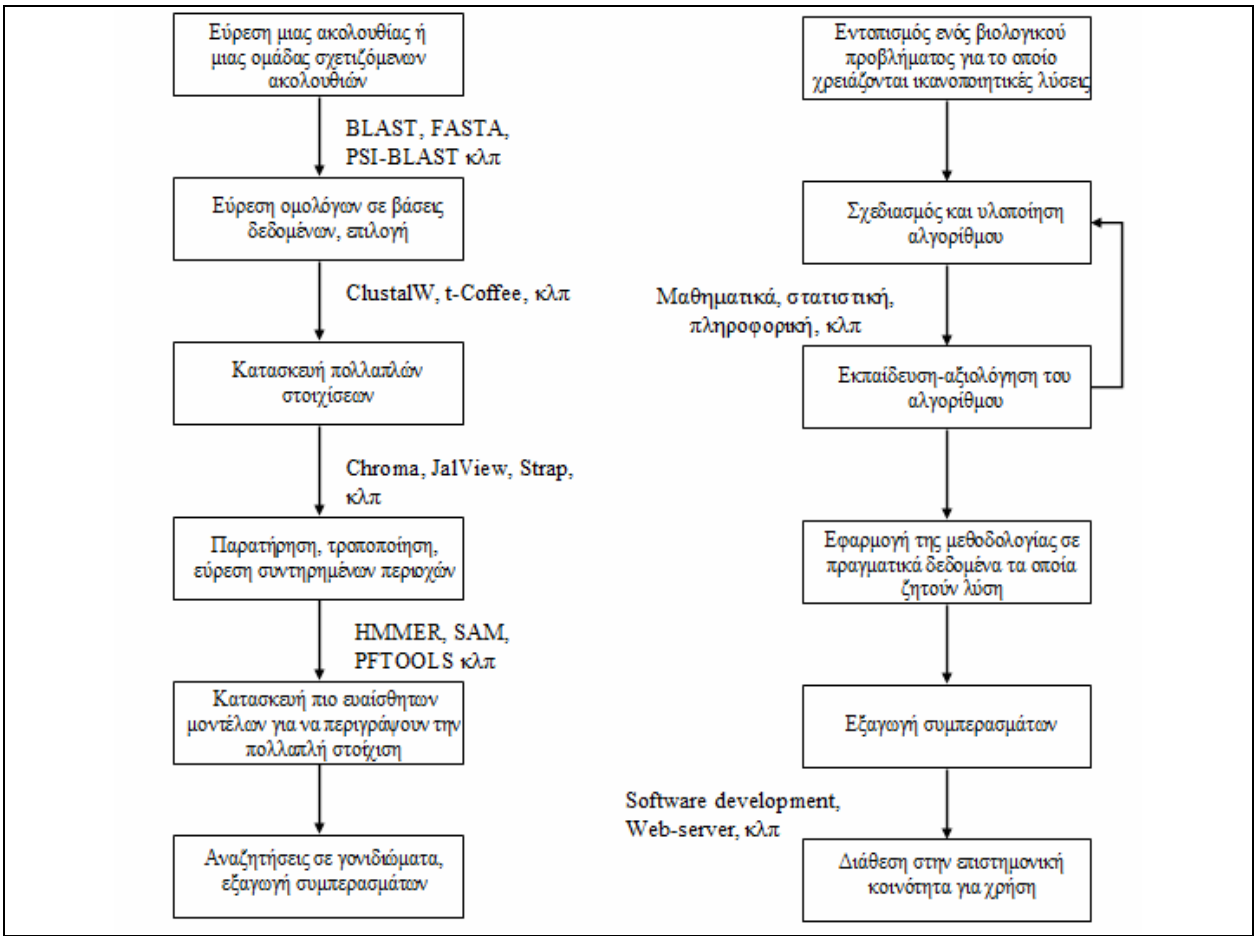
Το βασικό όμως πρόβλημα, για να επιστρέψουμε στο αρχικό μας ερώτημα, δεν είναι όλα τα παραπάνω (καθώς η βιοπληροφορική εντάσσεται ξεκάθαρα στον κορμό των βιολογικών και των άλλων θετικών επιστημών), αλλά η στρεβλή και απλοϊκή αντιμετώπιση του τελευταίου βήματος, του τρόπου δηλαδή παραγωγής της θεωρίας και εξαγωγής των γενικών νόμων. Για παράδειγμα, κάποια φαινόμενα είναι απλά, με την έννοια ότι επιδέχονται μια απλή και ξεκάθαρη απάντηση. Έτσι, όταν ξεκίνησε η διερεύνηση του γενετικού κώδικα, υπήρξε η παρατήρηση ότι τη τριπλέτα UUU κωδικοποιεί το αμινοξύ Φενυλαλανίνη (Phe). Αυτή η παρατήρηση, δεν χώραγε αμφισβήτηση, ενώ με τα αντίστοιχα (και ιδιαίτερα έξυπνα μπορούμε να πούμε) πειράματα για τις υπόλοιπες τριπλέτες αποκωδικοποιήθηκε ολόκληρος ο γενετικός κώδικας με τρόπο αδιαμφισβήτητο. Ο γενετικός κώδικας σύμφωνα με όσα είναι γνωστά από τη δεκαετία του 1960, είναι μια απλή συνάρτηση μίας μεταβλητής, μια απεικόνιση του συνόλου των κωδικονίων στο σύνολο των αμινοξέων, με την οποία κάθε μέλος του πρώτου συνόλου αντιστοιχείται σε ένα μόνο μέλος του δεύτερου συνόλου (μια συνάρτηση όμως που δεν είναι «ένα προς ένα», και κατά συνέπεια δεν είναι αντιστρέψιμη). Τι γίνεται όμως με περιπτώσεις στις οποίες τα πράγματα δεν είναι τόσο ξεκάθαρα; Σε περιπτώσεις που οι αντιστοιχίσεις δεν είναι τόσο απλές; Στην περίπτωση λ.χ. της δομής των πρωτεϊνών, παρ' όλα τα πειράματα και τα θεωρητικά επιχειρήματα που δείχνουν ξεκάθαρα ότι η αλληλουχία καθορίζει τη δομή, δεν υπάρχει κάποιος ξεκάθαρος κώδικας, κάποιος κανόνας που να λέει ότι η αλληλουχία των Χ-Υ-Ζ αμινοξέων θα έχει πάντα τη δομής α-έλικας, ενώ η αλληλουχία των Α-Β-Γ αμινοξέων θα έχει τη δομή β-πτυχωτής επιφάνειας. Εδώ, το πρόβλημα που προκύπτει είναι εγγενώς ασαφές και πολυδιάστατο και όσα δεδομένα και αν συλλέξουμε, όσα πειράματα προσδιορισμού δομών και αν κάνουμε, δεν θα μπορέσουμε ποτέ να καταλήξουμε (τουλάχιστον με τον απλοϊκό τρόπο που είδαμε παραπάνω) σε τόσο απλά διατυπωμένους καθολικούς νόμους.

Σε αυτό το σημείο έρχεται να συμβάλει η βιοπληροφορική, η οποία χρησιμοποιώντας τα υπάρχοντα πειραματικά δεδομένα (τα οποία μπορεί να αντιπροσωπεύουν χιλιάδες ανθρωποώρες επίπονης δουλειάς των εργαστηριακών βιολόγων), χρησιμοποιεί τεχνικές της στατιστικής, των μαθηματικών και της πληροφορικής, με σκοπό να εξάγει ένα γενικό νόμο ή έστω κάποιους κανόνες που να τον προσεγγίζουν. Εισάγει δηλαδή τη μαθηματικοποίηση και την ποσοτικοποίηση των βιολογικών φαινομένων, μια προσέγγιση που όπως είδαμε δεν είναι καθόλου νέα στη βιολογία. Στο παράδειγμα της δομής των πρωτεϊνών, η υπολογιστική μελέτη των χιλιάδων τρισδιάστατων δομών των πρωτεϊνών, έχει καταλήξει σε κάποιους γενικούς νόμους, οι οποίοι δεν αποτελούν -και δεν θα μπορούσαν να αποτελέσουν ποτέ- το αποτέλεσμα κάποιου συγκεκριμένου «πειράματος». Αυτοί οι νόμοι, λένε ότι: α) οι πρωτεϊνικές δομές συντηρούνται περισσότερο από τις πρωτεϊνικές αλληλουχίες και β) οι περισσότερες σημειακές μεταλλάξεις στις πρωτεΐνες συμβαίνουν στην επιφάνειά τους παρά στο εσωτερικό της δομής. Αυτοί οι νόμοι μπορεί να θεωρηθούν ως γενικότεροι νόμοι των βιολογικών επιστημών, καθώς έχουν γενικότερες συνέπειες σε πολλά πεδία και δίνουν άμεσες απαντήσεις σε πολλά πρακτικά ερωτήματα. Για παράδειγμα, αν εντοπίσουμε μια πρωτεΐνη με 99% ομοιότητα με μια πρωτεΐνη γνωστής δομής και λειτουργίας, οι νόμοι αυτοί μας λένε ξεκάθαρα και με μεγάλη βεβαιότητα ότι τα αμινοξικά κατάλοιπα που διαφέρουν θα βρίσκονται στην επιφάνεια του μορίου, θα έχουν ελάχιστη επίδραση στην τρισδιάστατη δομή και κατά κανόνα δεν θα επηρεάζουν σημαντικά τη γενική βιολογική λειτουργία. Επιπλέον δε, οι αλγόριθμοι πρόγνωσης της δομής, ακόμα και αν δεν δίνουν κάποιον ξεκάθαρο κανόνα, μπορούν να κάνουν προβλέψεις για τη δομή μιας πρωτεΐνης, προβλέψεις για τις οποίες ξέρουμε με μεγάλη αξιοπιστία τι ποσοστό επιτυχίας αναμένουμε να έχουν. Τέτοιοι νόμοι και η διαδικασία με την οποία προέρχονται (η βιοπληροφορική δηλαδή) μπορούν από τη μία μεριά να επιταχύνουν τη βιολογική έρευνα οργανώνοντας τον όγκο των πειραματικών δεδομένων, αλλά και αντικαθιστώντας από την άλλη, όπου αυτό είναι δυνατό, τα επιπλέον πειράματα αποκλείοντας μη πιθανές εκδοχές. Εάν λάβουμε υπ' όψη όλα τα παραπάνω, δεν συνεπάγεται ότι πρέπει να μειώσουμε την αξία του πειραματισμού και της πειραματικής βιολογίας (αν μη τι άλλο, αν δεν υπήρχαν τα τεράστια αποθέματα πειραματικών δεδομένων, δεν θα υπήρχε και βιοπληροφορική). Αυτό που πρέπει να γίνει, είναι να αναγνωριστεί η βιοπληροφορική και η υπολογιστική βιολογία σαν ένας αυτόνομος κλάδος των βιολογικών επιστημών με τον ίδιο, ή περίπου τον ίδιο, τρόπο με τον οποίο οι φυσικοί έχουν αποδεχθεί την υπολογιστική φυσική και οι χημικοί την υπολογιστική χημεία.

Η παραπάνω λανθασμένη αντίληψη περί βιοπληροφορικής, αποτελεί έναν από τους «μύθους περί βιοπληροφορικής», που ανέλυσε ο Χρήστος Ουζούνης (Ouzounis, 2000). Ένας άλλος μύθος που σχετίζεται

βέβαια με αυτή τη λάθος θεώρηση, είναι ότι «*η βιοπληροφορική είναι μια νέα τεχνολογία*». Αφενός μεν, όπως είδαμε προηγουμένως, αν και ο όρος είναι νέος, το αντικείμενο της υπολογιστικής βιολογίας και βιοπληροφορικής έχει βάθος δεκαετιών. Αφετέρου δε, με όσα αναφέραμε ήδη, θα πρέπει να γίνεται κατανοητό ότι δεν είναι «τεχνολογία». Τεχνολογία, τουλάχιστον στο βαθμό που μας αφορά σχετικά με τη συνάφειά της με τη βιολογία, είναι για παράδειγμα, οι μικροσυστοιχίες, το pyrosequencing, οι σχεσιακές βάσεις δεδομένων ή μια νέα γλώσσα προγραμματισμού. Αντίθετα, η βιοπληροφορική είναι ένα εκτεταμένο σύνολο εννοιών και μεθοδολογιών, που όπως είδαμε παραπάνω συνιστά ξεχωριστή επιστημονική ειδικότητα. Αυτή η παρανόηση προέρχεται από την αποσπασματική εικόνα που έχουν πολλοί, ειδικά στο χώρο της βιολογίας. Για παράδειγμα, ένας ερευνητής που ασχολείται για χρόνια με τη γονιδιακή έκφραση, έρχεται για πρώτη φορά σε επαφή με την τεχνολογία των μικροσυστοιχιών η οποία απογειώνει τη δουλειά του. Μαζί όμως με τα πανάκριβα μηχανήματα, τα αντιδραστήρια, και τον υπόλοιπο εξοπλισμό, βλέπει και έναν ‘περίεργο τύπο’ να «πατάει τα κουμπιά» και να βγάζει αποτέλεσμα. Είναι λογικό κατά κάποιον τρόπο να υποθέσει ότι αυτό μόνο είναι η βιοπληροφορική, μια τεχνολογική πλατφόρμα για να διευκολύνει τη δουλειά του, όπως οι Η/Υ, τα λειτουργικά συστήματα και ο επεξεργαστής κειμένου στον οποίο θα γράψει την εργασία του. Αγνοεί όμως, αφενός μεν το θεωρητικό υπόβαθρο που υπάρχει πίσω από όλες τις πλατφόρμες λογισμικού που χρησιμοποιεί ο «τεχνικός» και την ειδική γνώση που χρειάζεται για την κατανόηση των αποτελεσμάτων, αφετέρου δε το γεγονός ότι υπάρχει και άλλου είδους βιοπληροφορική. Υπάρχουν αυτοί που κατασκευάζουν τους αλγόριθμους, αυτοί που αποδεικνύουν τα θεωρήματα, αυτοί που σχεδιάζουν το λογισμικό, κ.ο.κ. και όλα αυτά, για ένα μεγάλο εύρος ερευνητικών ερωτημάτων, όχι μόνο για τις μικροσυστοιχίες που είναι (ή ήταν) της μόδας (π.χ. στοίχιση αλληλουχιών, φυλογενετική ανάλυση, πρόγνωση δομής, μοντελοποίηση με βάση την ομολογία, κατασκευή βιολογικών βάσεων δεδομένων κ.ο.κ.). Μια τέτοια αντίληψη υπάρχει δυστυχώς παγκοσμίως σε μερίδα των βιολόγων, και όπως έγραφε με παράπονο ο Edgar Wingender: «*Thus, scientific articles publishing experimental findings which have been evaluated using computational tools, very often give credit to them in the Methods or Results sections with phrases such as "Computer analysis revealed that ...", without any appropriate reference. In contrast, any experimental methodology used is extensively explained in these papers, down to the detailed listing of buffer systems, voltage/current conditions of the electrophoresis systems etc*» (Wingender, 1998).

Η λάθος αυτή εντύπωση που δημιουργείται σε πολλούς, οδηγεί και σε έναν άλλο πολύ διαδεδομένο μύθο, αυτόν που λέει ότι «η βιοπληροφορική είναι εύκολη», με επακόλουθο το ότι «ο καθένας μπορεί να το κάνει» και «οι εργασίες οι δικές σας βγαίνουν εύκολα, πατάτε δυο κουμπιά και βγάζετε δημοσιεύσεις (paper)». Πάλι, μια τέτοια θεώρηση είναι λανθασμένη, αν και πρέπει να αναγνωρίσουμε στους πειραματικούς βιολόγους ότι ειδικά στην Ελλάδα η υψηλού επιπέδου έρευνα στις μοριακές επιστήμες είναι ακριβή, αλλά το κυριότερο, δύσκολη καθώς, εκτός από την εξασφάλιση των κονδυλίων, πρέπει κανείς να αναμετρηθεί και με τη γραφειοκρατία, να ασχοληθεί με διαγωνισμούς και προμήθειες και ακόμα και αν είναι τυχερός με όλα αυτά, μπορεί να περιμένει μήνες για να παραλάβει τα ακριβά του αντιδραστήρια και τα μηχανήματα. Παρ’ όλα αυτά, η κατάσταση για κάποιον που έχει μια εποπτική εικόνα της βιοπληροφορικής, δεν είναι στο σύνολο της, δραματικά καλύτερη. Ναι, υπάρχουν όντως κάποια ερευνητικά πεδία που απαιτούν λιγότερη επένδυση σε υλικό και λογισμικό, αλλά στις περισσότερες περιπτώσεις απαιτούνται ισχυροί υπολογιστές και μεγάλοι αποθηκευτικοί χώροι. Όπως και να έχει όμως, όλες οι εργασίες βιοπληροφορικής απαιτούν εξειδικευμένο προσωπικό υψηλών προσόντων, κάτι το οποίο δεν είναι ούτε εύκολο να βρεθεί, αλλά ούτε και «χαμηλού κόστους». Κατά συνέπεια, το «ο καθένας μπορεί να το κάνει» δεν ευσταθεί, γιατί... αν μπορούσε, θα το είχε κάνει. Δεν μπορεί ο καθένας να φτιάξει έναν επιτυχημένο αλγόριθμο πρόγνωσης, γιατί μια τέτοια διαδικασία απαιτεί ακριβώς τις εξειδικευμένες γνώσεις που απαιτεί η διεπιστημονική προσέγγιση που αναφέραμε, ούτε μπορεί ο καθένας να κάνει μια συγκριτική ανάλυση όλων των γνωστών γονιδιωμάτων, γιατί κάτι τέτοιο προϋποθέτει επιπλέον και μεγάλη υπολογιστική ισχύ και αποθηκευτικό χώρο. Αλλά ούτε και όταν κάποιος έχει τη δυνατότητα να φτιάξει έναν τέτοιον αλγόριθμο ή να κάνει μια τέτοια ανάλυση, αυτό σημαίνει ότι αυτά γίνονται «γρήγορα». Τέτοιες δραστηριότητες, απαιτούν προσεκτικό σχεδιασμό και πειραματισμό, δοκιμή και σφάλμα, διαδικασίες δηλαδή επίπονες και χρονοβόρες. Είναι δηλαδή, από όλες τις απόψεις, πραγματικά πειράματα και σαν τέτοια θα έπρεπε να αντιμετωπίζονται.



Εικόνα 1.10: *Αριστερά, η διεργασία που επιτελεί ένας χρήστης βιοπληροφορικής. Δεξιά, η διεργασία που επιτελεί ένας βιοπληροφορικός*

Σε αυτό το σημείο, πρέπει να κάνουμε όμως και έναν επιπλέον διαχωρισμό των δραστηριοτήτων βιοπληροφορικής. Οι δραστηριότητες που αναφέραμε προηγουμένως, η κατασκευή μεθόδου πρόγνωσης, η μεγάλης κλίμακας υπολογιστικές αναλύσεις, ο σχεδιασμός αλγορίθμων και λογισμικού κ.ο.κ. ανήκουν στην κατηγορία δραστηριοτήτων που όντως απαιτούν μια μεγάλη εξειδίκευση και δεν μπορούν να γίνουν από τον καθένα. Σήμερα όμως, με την πρόοδο που έχει επιτευχθεί σε όλους τους τομείς, υπάρχει και μια μεγάλη ομάδα δραστηριοτήτων που μπορεί να τις επιτελέσει ο καθένας, και μάλιστα θα έλεγα ότι είναι απαραίτητο να μπορεί να τις πραγματοποιεί ο καθένας που ασχολείται με τη βιολογική έρευνα. Η τεράστια ανάπτυξη της υπολογιστικής βιολογίας, όπως είδαμε, έχει φέρει τη χρήση αλγοριθμικών και υπολογιστικών εργαλείων στην καθημερινότητα του βιολόγου (δεν υπάρχει βιολόγος που να μην έχει χρειαστεί να χρησιμοποιήσει το BLAST). Έτσι θα λέγαμε ότι υπάρχουν διάφορες κατατάξεις στον τρόπο που ένας ερευνητής χρησιμοποιεί και εμπλέκεται με δραστηριότητες βιοπληροφορικής. Στην πρώτη κατηγορία, έχουμε τις απλές αναλύσεις που αφορούν τη χρήση λογισμικού (στοίχιση, πολλαπλή στοίχιση, μέθοδος πρόγνωσης, αναζήτηση σε βάσεις δεδομένων κ.ο.κ.). Αυτές τις δραστηριότητες θα μπορούσε και θα έπρεπε να μπορεί να της φέρει σε πέρας ο κάθε βιολόγος ανεξάρτητα της ειδικότητας του και του αντικειμένου της έρευνάς του, αλλά από την εμπειρία μας έχουμε δει ότι η έλλειψη θεωρητικής κατανόησης για το τι ακριβώς κάνει η κάθε μέθοδος, οδηγεί σε πολλά προβλήματα (λάθος χρήση μιας μεθόδου). Στην επόμενη κατηγορία, μπορούμε να κατατάξουμε αυτούς που χρησιμοποιούν κατά κύριο λόγο τέτοια έτοιμα εργαλεία και αλγόριθμους για να πραγματοποιήσουν σύνθετες αναλύσεις και να απαντήσουν σε κάποιο βιολογικό ερώτημα. Τα τελευταία χρόνια, η εμφάνιση των δεδομένων γονιδιακής έκφρασης, τα δεδομένα αλληλούχισης νέας γενιάς κ.ο.κ. έχουν αυξήσει αυτού του είδους τις αναλύσεις και την ανάγκη για άτομα που να μπορούν να τις φέρουν σε πέρας. Το βασικό όμως χαρακτηριστικό αυτής της ομάδας είναι ότι δεν αναπτύσσει αλγόριθμους, ούτε λογισμικό. Τέλος, υπάρχουν και αυτοί που εστιάζοντας σε κάποιο συγκεκριμένο πρόβλημα αναπτύσσουν και αλγόριθμους και λογισμικό.

Οι αλγόριθμοι μπορεί να είναι αλγόριθμοι πρόγνωσης, στοίχισης ή οποιαδήποτε άλλη κατηγορία από αυτές που έχουμε αναλύσει.

Υπάρχει στη βιβλιογραφία μια έντονη διχογνωμία για το πώς πρέπει να ονομαστεί ο επιστήμονας της κάθε κατηγορίας, ειδικά για τις κατηγορίες 2 και 3 (στην 1 είναι οι βιολόγοι όπως είπαμε). Έτσι, έχουν προταθεί οι όροι «bioinformatician» για τους επιστήμονες της κατηγορίας 2 και «bioinformaticist» για αυτούς της κατηγορίας 3, αλλά η διάκριση δεν έχει γίνει αποδεκτή και ο τελευταίος όρος δεν χρησιμοποιείται πολύ. Άλλοι χρησιμοποιούν τον όρο «bioinformatics scientist» και «bioinformatics engineer» για τις κατηγορίες 2 και 3 αντίστοιχα (Welch et al., 2014), αλλά και αυτή η προσέγγιση προσωπικά δεν μου φαίνεται σωστή, γιατί μπορεί να δημιουργήσει την εντύπωση (ειδικά στην Ελλάδα), ότι πρέπει υποχρεωτικά οι επιστήμονες να είναι μηχανικοί, δηλαδή απόφοιτοι πολυτεχνείου. Ίσως ένας περιφραστικός ορισμός να είναι αναγκαίος, πάντα ανάλογα με το περιεχόμενο και την περίσταση. Όπως και να έχει όμως, η άποψή μου είναι ότι μια σωστή διεπιστημονική εκπαίδευση, ακόμα και στο πλαίσιο του βασικού πτυχίου (αλλά σίγουρα και στο πλαίσιο ενός μεταπτυχιακού προγράμματος), θα επέτρεπε στους βιολόγους να μπορούν να αποδίδουν τα μέγιστα, ακόμα και στην περιοχή της εξειδίκευσής τους. Υπάρχουν ένα σωρό παραδείγματα υπολογιστικών αναλύσεων οι οποίες θα μπορούσαν να είχαν γίνει ακόμα και από έναν «παραδοσιακό» βιολόγο (με την έννοια ότι δεν χρειάζεται ειδικές γνώσεις προγραμματισμού), όπως η κατασκευή τρισδιάστατων μοντέλων πρωτεϊνών, ο χαρακτηρισμός μιας πρωτεϊνικής οικογένειας και η εύρεση μακρινών ομολόγων κ.ο.κ., αλλά αυτές αφέθηκαν στους «ειδικούς», τους βιοπληροφορικούς. Επιπλέον, μια μεταπτυχιακή εκπαίδευση στη χρήση των βασικών εργαλείων, ειδικά όσον αφορά τις νέες τεχνολογίες αλληλούχισης και γονιδιακής έκφρασης, θα μπορούσε να αποτελέσει και μια επαγγελματική διέξοδο με περισσότερες προοπτικές, καθώς οι τεχνολογίες αυτές χρησιμοποιούνται πλέον στα περισσότερα εργαστήρια μοριακής βιολογίας αλλά και σε νοσοκομεία και διαγνωστικά κέντρα. Φυσικά, σε αυτή την κατηγορία δεν είναι απαραίτητο να εντάσσονται μόνο βιολόγοι (αν και θα ήταν ίσως πιο εύκολο για αυτούς), αλλά και επιστήμονες άλλων ειδικοτήτων όπως πληροφορικοί, μηχανικοί και στατιστικοί, αφού θα έχουν περάσει πρώτα από κάποιου είδους εκπαίδευση. Οι επιστήμονες της 3^{ης} κατηγορίας, συνιστούν την πιο ετερογενή ομάδα, καθώς σε αυτό το πρότυπο μπορεί να ταιριάζουν επιστήμονες με διαφορετικό προφίλ, από μοριακούς βιολόγους και φυσικούς, μέχρι θεωρητικούς πληροφορικούς και μαθηματικούς που ασχολήθηκαν με ένα συγκεκριμένο πρόβλημα της υπολογιστικής βιολογίας και εστιάζουν στο πώς θα αναπτύξουν αλγόριθμους και λογισμικό για την αντιμετώπισή του.

Αφού είδαμε την ιστορική διαδρομή της βιοπληροφορικής και τις θεωρητικές αναλύσεις που δικαιολογούν τη διεπιστημονικότητα του κλάδου, αναγκαστικά καταλήγουμε στη συζήτηση για την εκπαίδευση των βιοπληροφορικών. Γενικά, υπάρχει μεγάλη συζήτηση στη βιβλιογραφία για το ποια θα πρέπει να είναι η κατάλληλη εκπαίδευση, για το πώς θα πρέπει να δομούνται τα διάφορα προγράμματα σπουδών ειδικά όταν θα απευθύνονται σε διαφορετικό ακροατήριο, για το πώς θα επιτυγχάνεται η διεπιστημονική προσέγγιση, αλλά και για το πώς θα ενσωματωθούν τα μαθήματα βιοπληροφορικής στα βασικά προγράμματα σπουδών των βιοεπιστημών και της ιατρικής (Altman, 1998; Ditty et al., 2010; Floriano, 2008; Honts, 2003; Searls, 2012; Welch, et al., 2014; Yan, Ban, & Tan, 2014). Αυτή η διεπιστημονική εκπαίδευση, η οποία είχε ξεκινήσει ήδη από τη δεκαετία του 1990 στο εξωτερικό, έχει αρχίσει να γίνεται αποδεκτή σταδιακά και στην Ελλάδα. Προγράμματα Μεταπτυχιακών Σπουδών (ΠΜΣ) έχουν ήδη ιδρυθεί και λειτουργούν εδώ και χρόνια (βλ. παρακάτω), τα οποία δέχονται αποφοίτους όλων των παραπάνω κατηγοριών (βιολόγους, γιατρούς, μαθηματικούς, μηχανικούς, στατιστικούς, πληροφορικούς), ενώ το πρόγραμμα σπουδών τους διαιρείται με άξονα τις βασικές αρχές που περιέγραψε πρώτος ο Altman: βασική βιολογία, βιοστατιστική, προγραμματισμός και βασικά στοιχεία επιστήμης υπολογιστών και τέλος, ειδικές γνώσεις της βιοπληροφορικής (ανάλυση ακολουθιών, ανάλυση δομών, βιολογικές βάσεις δεδομένων, κ.ο.κ.). Στο προπτυχιακό επίπεδο, όπως θα δούμε παρακάτω, τα πράγματα είναι πιο σύνθετα, καθώς η βιοπληροφορική πρέπει να ενσωματωθεί σε ένα υπάρχον πρόγραμμα σπουδών. Έτσι, τα περισσότερα τμήματα βιολογίας έχουν ανταποκριθεί στις ανάγκες της εποχής εντάσσοντας στο πρόγραμμά τους κάποιο μάθημα βιοπληροφορικής, αλλά το ίδιο δε συμβαίνει για τα περισσότερα τμήματα πληροφορικής ή μηχανικών Η/Υ. Το Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας, είναι κατά κάποιον τρόπο μοναδικό στον τομέα αυτό, καθώς παρέχει και σε προπτυχιακό επίπεδο τη διεπιστημονικότητα που χρειάζεται, καθώς στον κορμό των μαθημάτων πληροφορικής που είναι κοινός με τα περισσότερα τμήματα πληροφορικής, παρέχει και μια σειρά μαθημάτων στη Βιολογία, Βιοχημεία, Γενετική, Φυσιολογία, και φυσικά Βιοπληροφορική, Βιοστατιστική και Ιατρική Πληροφορική.

1.3. Η κατάσταση στον κόσμο

Την τελευταία δεκαετία έχουν γίνει πολλές επιστημομετρικές μελέτες με σκοπό να μελετήσουν ειδικά την επιστήμη της βιοπληροφορικής και συγκεκριμένα την επιστημονική βιβλιογραφία του χώρου. Άλλες δίνουν έμφαση στις ερευνητικές κατευθύνσεις που ακολουθεί ο κλάδος παγκοσμίως, άλλες επιχειρούν να σκιαγραφήσουν το τοπίο εστιάζοντας στα επιδραστικά περιοδικά, στους συγγραφείς και στα ερευνητικά ιδρύματα που εμπλέκονται στο χώρο, ενώ άλλες επιχειρούν να εστιάσουν στις ομοιότητες και τις διαφορές με άλλους συγγενείς κλάδους, κυρίως με την Ιατρική Πληροφορική. Σε όλες τις περιπτώσεις ένα βασικό πρόβλημα που αντιμετωπίζει μια τέτοια μελέτη είναι στο πώς θα ορίσει το τι είναι βιοπληροφορική. Κάποιοι επιχειρούν να εντοπίσουν τις σχετικές δημοσιεύσεις με χρήση κάποιων καθορισμένων εκ των προτέρων λέξεων-κλειδιών (keywords ή MESH terms), άλλοι επιλέγουν από την αρχή τα επιστημονικά περιοδικά που θεωρούν ότι είναι χαρακτηριστικά του χώρου, ενώ άλλοι ακολουθούν μια μικτή στρατηγική. Σε κάθε περίπτωση, τα αποτελέσματα είναι ιδιαίτερα ενδιαφέροντα και κάποια από αυτά θα προσπαθήσουμε να περιγράψουμε παρακάτω.

Σε μια από τις πρώτες τέτοιες μελέτες, οι Patra και Mishra (Patra & Mishra, 2006) χρησιμοποίησαν κάποια γενικά MESH terms όπως "Bioinformatics" OR "Bioinformatics" OR "Computational Biology" OR "Computational Molecular Biology" OR "Biology Computational" OR "Molecular Biology; Computational" OR "Genomics" για να πραγματοποιήσουν αναζήτηση στην PUBMED και συγκέντρωσαν 16.178 επιστημονικά άρθρα, τα οποία είχαν δημοσιευθεί μέχρι το 2004, προερχόμενα από 1806 διαφορετικά επιστημονικά περιοδικά. Όπως θα ανέμενε κανείς, η αύξηση την περίοδο πριν από το 2000 ήταν εκθετική. Ενδεικτικά, το 1990 είχαν δημοσιευθεί μόνο 12 άρθρα, ενώ το 2000 ο αντίστοιχος αριθμός ξεπέρασε τα 1000. Η μεγάλη πλειοψηφία των εργασιών αυτών ήταν άρθρα σε περιοδικά (98%) και ήταν γραμμένες στα Αγγλικά (97%). Οι επιστήμονες από τις ΗΠΑ είχαν το μεγαλύτερο μερίδιο στους συγγραφείς (42%) ακολουθούμενοι από τους Βρετανούς (10%), τους Γερμανούς (6%) και τους Ιάπωνες (4%).

Bioinformatics
Nucleic Acids Research
Genome Research
Science
Nature
Proceedings of the National Academy of Sciences USA
Proteomics
Genome Biology
Journal of Molecular Biology
Proteins
Nature Biotechnology
BMC Bioinformatics
Pacific Symposium on Biocomputing
Journal of Computational Biology
Tanpakushitsu Kakusan Koso
Journal of Biological Chemistry
Drug Discovery Today
Trends in Biotechnology
Genomics
Briefing in Bioinformatics

Πίνακας 1.1: *Τα 20 κύρια περιοδικά βιοπληροφορικής που εντοπίστηκαν στη μελέτη των Patra και Mishra (Patra & Mishra, 2006)*.

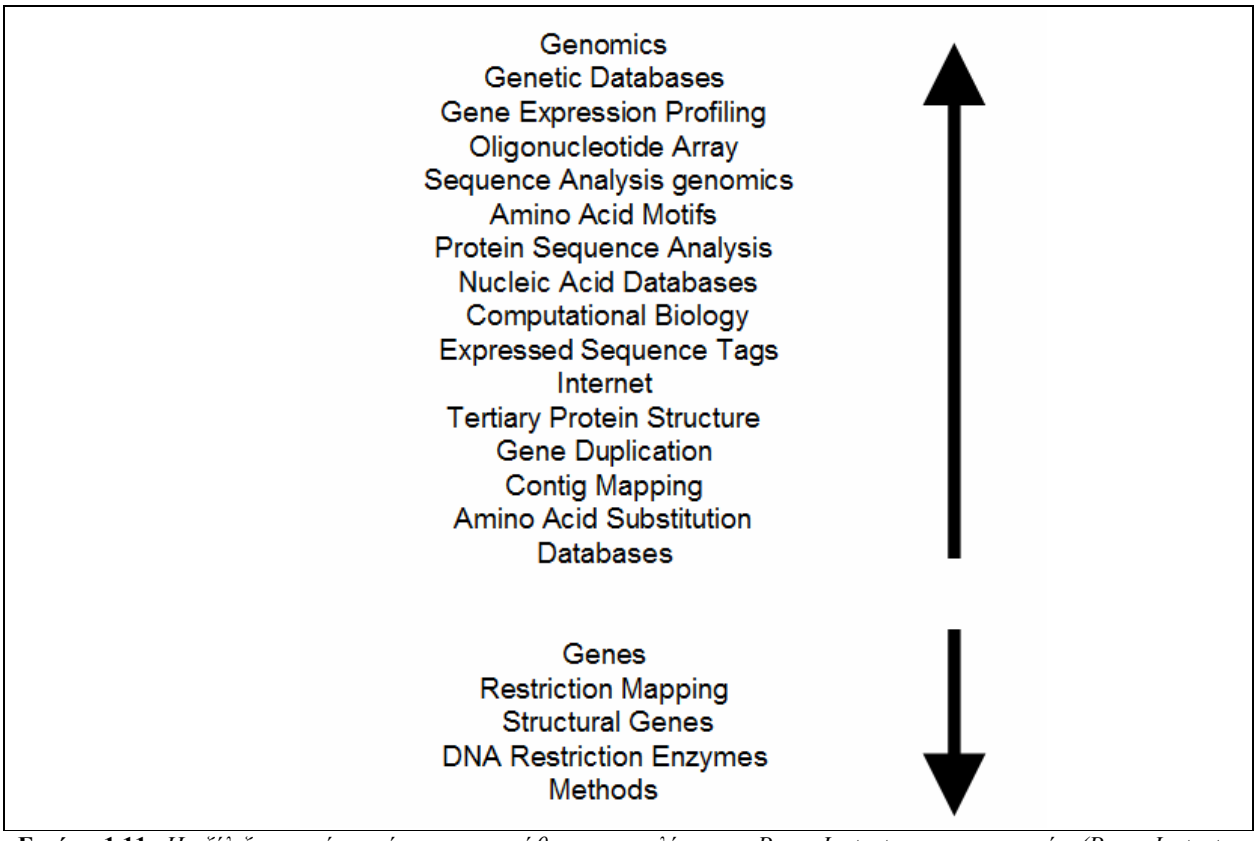
Από την κατανομή των άρθρων ανά περιοδικό, εντόπισαν τα 20 «κύρια» επιστημονικά περιοδικά στα οποία είχαν δημοσιευθεί το 1/3 των εργασιών αυτών. Τα περιοδικά αυτά φαίνονται στον Πίνακα 1.1. Όπως είναι αναμενόμενο, κάποια από αυτά τα περιοδικά είναι αποκλειστικά περιοδικά βιοπληροφορικής (Bioinjormatics, BMC Bioinformatics, Pacific Symposium on Biocomputing, Journal of Computational Biology κ.ο.κ.), αλλά υπάρχουν και πολλά από τα κορυφαία περιοδικά βιολογικού (Nucleic Acids Research, Genome Research, Genome Biology, Journal of Molecular Biology) ή γενικότερου ενδιαφέροντος (Science, Nature, Proceedings of the National Academy of Sciences USA). Παρατηρήθηκε επίσης, ότι τα περισσότερα

από τα περιοδικά αυτά είχαν αυξήσει το δείκτη επιρροής (Impact Factor) τους από το 2001 και μετά, πιθανότατα λόγω αυξημένων αναφορών από τις εργασίες βιοπληροφορικής, ενώ κάποια από αυτά είχαν αλλάξει ακόμα και το όνομά τους για να ανταποκριθούν καλύτερα στις νέες συνθήκες (π.χ. το Computer Application in Biosciences μετονομάστηκε σε Bioinfornatics το 1998, ενώ το PCR Methods and Its Applications μετονομάστηκε σε Genome Research το 1994).

Όσον αφορά τον αριθμό των συγγραφέων σε κάθε άρθρο, η μελέτη βρήκε ότι το 23% των εργασιών είχε γραφτεί από έναν συγγραφέα, ενώ το 23% από δύο. Ο μέγιστος αριθμός συγγραφέων σε ένα άρθρο ήταν 40 και βρέθηκαν 67 τέτοια άρθρα. Τα περισσότερα από αυτά τα άρθρα εμφανίστηκαν μετά το 2000 κάτι που προφανώς έχει σχέση με τα τεράστια συνεργατικά consortia που ασχολήθηκαν με την αλληλούχιση γονιδιωμάτων. Συνολικά, υπήρχαν 39.435 συγγραφείς για τα 16.178 άρθρα (2,43 συγγραφείς/εργασία). Παρ' όλα αυτά, το 73,58% των συγγραφέων είχε συνεισφορά μόνο σε ένα άρθρο, ενώ το 14,34% σε δύο και το 5,30% σε τρία. Τα αποτελέσματα αυτά είναι σύμφωνα με το νόμο του Lotka που λέει ότι ο αριθμός των συγγραφέων που έχει n εργασίες, είναι περίπου ίσος με το $1/n^2$ αυτών που έχουν μόνο μία. Σύμφωνα με αυτόν το νόμο μόνο το 6% των συγγραφέων σε ένα ερευνητικό πεδίο θα έχει πάνω από 10 εργασίες. Τα αποτελέσματα αυτά εξηγούνται αν αναλογιστούμε ότι πολλοί συγγραφείς κάνουν μία ή δύο εργασίες και μετά εγκαταλείπουν την έρευνα σε αυτό το αντικείμενο, είτε γιατί είναι φοιτητές που δεν συνεχίζουν την ερευνητική καριέρα, είτε γιατί αλλάζουν αντικείμενο. Επίσης, η διεπιστημονικότητα της βιοπληροφορικής συντείνει στο να υπάρχουν και αρκετοί καταξιωμένοι επιστήμονες από άλλους χώρους (μαθηματικά, πληροφορική, βιολογία), οι οποίοι μόνο περιστασιακά ενεπλάκησαν στη συγγραφή εργασιών βιοπληροφορικής.

Σε μια άλλη εργασία του 2006, οι Perez-Iratxeta, Andrade-Navarro και Wren (Perez-Iratxeta, Andrade-Navarro, & Wren, 2007) έκαναν μια ανάλυση των λέξεων σε όλα τα άρθρα της PUBMED με σκοπό να εντοπίσουν βιοπληροφορικές εργασίες εστιάζοντας σε 3 διαφορετικούς τομείς (υπολογισμούς, διαδίκτυο και βάσεις δεδομένων). Έκαναν την αναζήτηση μεταξύ των ετών 1996 και 2005 αναζητώντας όρους (MESH terms) που παραπέμπουν σε υπολογιστική ανάλυση βιολογικών δεδομένων όπως: 'comput*', '*informatic*', 'algorithm*', 'software' ή 'database' ενώ επιπλέον αναζήτηση έγινε για λέξεις κλειδιά όπως 'internet', 'online', 'world wide web', 'web-based', 'http:*' και 'ftp:*'.

Η μελέτη αυτή έδειξε ότι το ποσοστό των εργασιών που χρησιμοποιούν υπολογιστικές τεχνικές στη βιοϊατρική έρευνα αυξήθηκε από το 1,6% το 1975, στο 10% το 2005. Η χρήση διαδικτυακών πηγών αυξήθηκε από το 0,05% στο 0.87% την ίδια περίοδο με τη μεγαλύτερη άνοδο να συμβαίνει τη δεκαετία του 1990 με την εξάπλωση του διαδικτύου και των Η/Υ. Επίσης, παρόμοια αύξηση υπάρχει και στην αναφερόμενη χρήση βάσεων δεδομένων, κάτι που απεικονίζει και την τεράστια αύξηση των δεδομένων που βρίσκονται κατατεθειμένες σε αυτές αλλά και την αντίστοιχη τεχνολογική πρόοδο που έκανε εύκολη τη συλλογή και αποθήκευση αυτών των δεδομένων. Γενικά, τα αποτελέσματα έδειξαν ότι η αύξηση στη χρήση υπολογιστικών μεθόδων συνέβη πρώτα, μετά ακολούθησε η εξάπλωση του διαδικτύου και στο τέλος έγινε η εμφάνιση των βάσεων δεδομένων, λόγω την ανάπτυξης των άλλων δύο τεχνικών. Τέλος, η χρονική ανάλυση των MESH terms υπέδειξε μια μεγάλη ομάδα όρων για τις οποίες υπήρξε μια τεράστια αύξηση στην εμφάνιση τους τα χρόνια 2000-2003 σε σχέση με τα προηγούμενα, και μια αντίστοιχη ομάδα όρων με μειωμένη χρήση (Εικόνα 1.11). Οι όροι αυτοί αντικατοπτρίζουν την μετάβαση από τις απλές μοριακές και βιοχημικές τεχνικές (π.χ. τη μελέτη ενός γονιδίου) στις υπολογιστικές αναλύσεις γονιδιωμάτων και μεγάλων συνόλων δεδομένων (κάτι που είναι χαρακτηριστικό της βιοπληροφορικής).



Εικόνα 1.11: *Η εξέλιξη των όρων όπως αποτυπώθηκε στη μελέτη των Perez-Iratxeta και συνεργατών (Perez-Iratxeta, Andrade-Navarro, & Wren, 2007)*

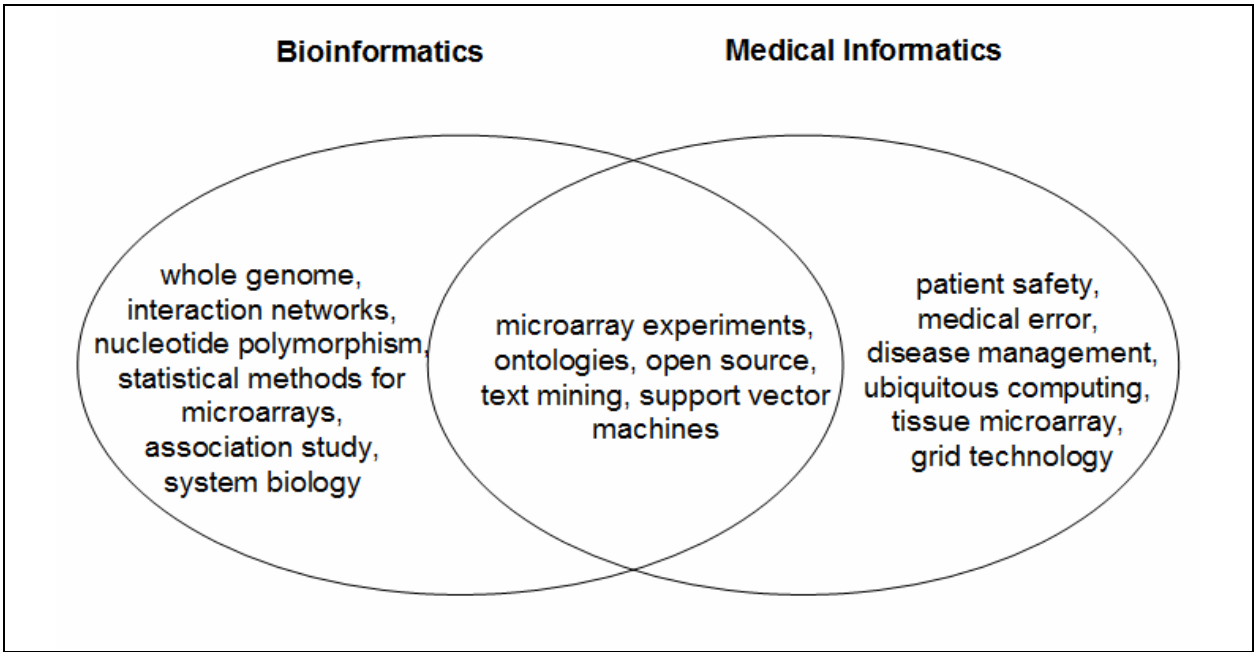
Μια άλλη μελέτη από το 2007, η οποία προέκυψε από μια μεγάλη διεπιστημονική συνεργασία (SYMBIOMATICS), αποσκοπούσε στο να μετρήσει με μια κλασική μέθοδο ανάλυσης κειμένου (bigrams, δηλαδή με τα ζευγάρια λέξεων) τη διαχρονική πορεία της σχετικής βιβλιογραφίας, να συγκρίνει τις «πρόσφατες» (2000-2005) με τις «παλιές» (1990-2000) εργασίες και να εντοπίσει έτσι τις αναδυόμενες τάσεις στο ερευνητικό πεδίο της βιοπληροφορικής, αλλά και επιπλέον, να εντοπίσει τις περιοχές σύγκλισης και διαφοροποίησης της βιοπληροφορικής με την ιατρική πληροφορική (Rebholz-Schuhman et al., 2007). Τα περιοδικά που θεωρήθηκαν από τους συγγραφείς ότι ανήκουν στις δύο κατηγορίες δίνονται στον Πίνακα 1.2.

Βιοπληροφορική	**Ιατρική Πληροφορική**
Bioinformatics	AMIA Annu Symp Proc
Biosystems	Artif Intell Med
BMC Bioinformatics	BMC Med Inform Decis Mak
Brief Bioinform	Int J Med Inform
Comput Methods Programs Biomed	J Am Med Inform Assoc
IEEE Trans Inf Technol Biomed	Medinfo
J Bioinform Comput Biol	Methods Inf Med
J Biomed Inform	Proc AMIA Symp
J Comput Aided Mol Des	
J Comput Biol	
Pac Symp Biocomput	

Πίνακας 1.2: *Τα περιοδικά βιοπληροφορικής και ιατρικής πληροφορικής που χρησιμοποιήθηκαν στη μελέτη SYMBIOMATICS.*

Η ανάλυση έδειξε καταρχάς ότι η μεγάλη αύξηση της βιβλιογραφίας της βιοπληροφορικής έγινε την περίοδο 2000-2005, ενώ η αντίστοιχη αύξηση της βιβλιογραφίας της ιατρικής πληροφορικής είχε συντελεστεί την δεκαετία 1990-2000. Για όλη τη δεκαπενταετία της ανάλυσης (1990-2005) οι πιο συνηθισμένες λέξεις

κλειδιά για τη Βιοπληροφορική ήταν «gene expression», «amino acid», και «protein sequence», ενώ για την ιατρική πληροφορική «information system», «health care» και «decision support». Ιδιαίτερο ενδιαφέρον όμως προκύπτει από τη διαχρονική ανάλυση και τη σύγκριση της περιόδου 1990-2000 με την αντίστοιχη περίοδο 2000-2005 από την οποία προκύπτουν οι αναδυόμενες τάσεις στα δύο πεδία καθώς και οι περιοχές σύγκλισής τους. Όπως φαίνεται στην Εικόνα 1.12 έννοιες που σχετίζονται με τις μικροσυστοιχίες, με τις οντολογίες, το ανοικτό λογισμικό, την ανάλυση κειμένου και τα Support Vector Machines είναι κοινές και στους δύο κλάδους και καταδεικνύουν μια μεγάλη περιοχή διεπαφής και συνέργειας. Οι συνέργιες αυτές των δύο περιοχών υπάρχουν και αυξάνονται διαχρονικά για μια σειρά από λόγους: Πρώτον, και οι δύο επιστημονικές περιοχές επωφελούνται από, και ασχολούνται με, τις νέες τεχνολογίες της βιοϊατρικής (π.χ. μικροσυστοιχίες), έστω και αν τις αντιμετωπίζουν με διαφορετικό τρόπο (ανάπτυξη μεθόδων από τη βιοπληροφορική, εφαρμογή στην κλινική πράξη από την ιατρική πληροφορική). Δεύτερον, και οι δύο επωφελούνται από νέες ανακαλύψεις στα μαθηματικά και την πληροφορική (π.χ. support vector machines). Τέλος, υπάρχει οριζόντια διάχυση καθώς τεχνολογίες και μεθοδολογίες που αναπτύχθηκαν αρχικά στον ένα κλάδο διαχέονται τελικά και στον άλλον και κατόπιν αναπτύσσονται παράλληλα με όφελος και για τους δύο (π.χ. οντολογίες, ανάλυση κειμένου, ανοιχτό λογισμικό κ.ο.κ.).



Εικόνα 1.12: *Τα ερευνητικά πεδία της βιοπληροφορικής και ιατρικής πληροφορικής με τις ομοιότητες και τις διαφορές τους, όπως βρέθηκαν από τη μελέτη SYMBIOMATICS*

Τέλος, το 2014 έγινε η πιο πρόσφατη βιβλιομετρική εργασία ειδικά για το πεδίο της βιοπληροφορικής (Song, Kim, Zhang, Ding, & Chambers, 2014). Οι ερευνητές προσπάθησαν να αντιμετωπίσουν τους περιορισμούς που είχαν οι προηγούμενες αναλύσεις τόσο από πλευράς κάλυψης της βιβλιογραφίας, όσο και από την πλευρά της μεθοδολογίας της ανάλυσης, ενώ φυσικά, το γεγονός ότι μιλάμε για μια σύγχρονη εργασία βοηθάει ιδιαίτερα στο να κατανοήσουμε τις τελευταίες εξελίξεις στον τομέα. Συνδυάζοντας δεδομένα από προηγούμενες μελέτες, οι ερευνητές κατέληξαν σε ένα μεγάλο σύνολο περιοδικών (μεγαλύτερο από τις προηγούμενες μελέτες), από τα οποία μόνο το 73% καλύπτεται από τη βάση δεδομένων του WoS (αλλά όλα είναι καταχωρημένα στην PUBMED). Με την επιλογή να αναλύσουν τα πλήρη κείμενα αντί για τις περιλήψεις, μπόρεσαν να κάνουν πιο ολοκληρωμένη ανάλυση κειμένου, με αντίτιμο το γεγονός ότι τα πλήρη κείμενα ήταν κατατεθειμένα στην PubmedCentral, το υποσύνολο της PUBMED στο οποίο γίνεται κατάθεση των εργασιών που δημοσιεύονται με το πρότυπο ανοιχτής πρόσβασης (open access) ή κατατίθενται από άλλα περιοδικά αλλά αφού έχει περάσει κάποιος χρόνος από την αρχική δημοσίευση. Με αυτόν τον τρόπο, κάποια περιοδικά ή ακόμα και κάποια άρθρα κάποιων περιοδικών δεν συμπεριλήφθηκαν στην ανάλυση. Ένα άλλο πλεονέκτημα αυτής της εργασίας ήταν ότι μπόρεσαν ταυτόχρονα να αναλύσουν και τις

αναφορές των εργασιών, με αποτέλεσμα να εξαχθούν συμπεράσματα για τα πιο επιδραστικά περιοδικά, τους συγγραφείς, τις εργασίες αλλά και τα ιδρύματα.

Τα περιοδικά που χρησιμοποιήθηκαν δίνονται στον Πίνακα 1.3 και η γενικότερη ανάλυση δείχνει ότι η αύξηση της σχετικής βιβλιογραφίας είναι εκθετική, κάτι που επιβεβαιώνει τις προηγούμενες μελέτες. Κατά την περίοδο 2000-2003, το κυρίαρχο αντικείμενο ήταν η μελέτη των πρωτεϊνών και κυρίως η λειτουργική μελέτη τους. Κατά την περίοδο 2004-2007 τα αντικείμενα γίνονται πιο ετερογενή και περιλαμβάνουν τη δομική ανάλυση γονιδίων, τον εγκέφαλο, τον καρκίνο και τους ιούς. Κατά την περίοδο 2008-2011, τα αντικείμενα συνεχίζουν να διαφοροποιούνται αλλά παρουσιάζουν ομοιότητες με την δεύτερη περίοδο. Παρ' όλα αυτά, νέοι όροι εμφανίζονται σε αυτή την περίοδο όπως mutation και RNA. Γενικά, τα πιο συνηθισμένα αντικείμενα όλης της περιόδου περιλαμβάνουν όρους όπως protein binding, algorithm/method, cell/model, network/interaction, genome sequence, immune/virus, gene expression, genetic/evolution, database/software, gene transcription, DNA/chromosome, ontology/mining, gene/genomics και cancer/cell.

Σχετικά με τις χώρες προέλευσης των εργασιών οι ΗΠΑ, η Μεγάλη Βρετανία και η Γερμανία βρίσκονται σταθερά στις πρώτες θέσεις σε όλες τις περιόδους, ακολουθούμενες από τη Γαλλία και τον Καναδά. Σταθερή άνοδο παρουσιάζουν η Κίνα και η Ιαπωνία. Η πρώτη χώρα, από μια θέση μεγαλύτερη της 20ης την περίοδο 2000-2003, φτάνει στην περίοδο 2009-2011 να βρίσκεται στην 6η θέση, ενώ η δεύτερη από την 9η θέση ανεβαίνει σταδιακά στην 7η. Η Ιταλία και η Ισπανία βρίσκονται σταθερά μέσα στην πρώτη δεκάδα ενώ η Ελλάδα όπως ίσως θα αναμέναμε, λόγω μεγέθους αλλά και ΑΕΠ, βρίσκεται σε θέση μεγαλύτερη της 20ης σε όλες τις περιόδους (περισσότερα για την Ελλάδα στην επόμενη ενότητα). Τα πρότυπα της σχετικής συνεισφοράς και της κατάταξης των διαφόρων κρατών στην έρευνα στη βιοπληροφορική, ακολουθούν τα γενικότερα πρότυπα που έχουν βρεθεί για το σύνολο της επιστημονικής παραγωγικότητας, τόσο από βιβλιομετρικές επιστημονικές μελέτες (King, 2004), όσο και από κατατάξεις της παγκόσμιας βιβλιογραφίας (http://www.natureindex.com/).

BMC Bioinformatics	Source Code for Biology and Medicine
BMC Genomics	Advanced Bioinformatics
PLoS Biology	BioData Mining
Genome Biology	Journal of Computational Neuroscience
PLoS Genetics	Journal of Proteome Research
PLoS Computational Biology	Journal of Biomedical Semantics
BMC Research Notes	Journal of Computer-Aided Molecular Design
Bioinformatics	Genome Integration
Molecular Systems Biology	Journal of Molecular Modeling
BMC Systems Biology	Bulletin of Mathematical Biology
Comparative and Functional Genomics	Pharmacogenetics and Genomics
Bioinformation	Statistical Methods in Medical Research
Theoretical Biology and Medical Modeling	Neuroinformatics
Human Molecular Genetics	Genomics
The EMBO Journal	Protein Science
Cancer Informatics	Physiological Genomics
Genome Medicine	Trends in Genetics
Evolutionary Bioinformatics	Journal of Proteomics
Biochemistry	Proteomics
Algorithms for Molecular Biology	Trends in Biochemical Sciences
EURASIP Journal on Bioinformatics and Systems Biology	Journal of Biotechnology
Journal of Molecular Biology	Trends in Biotechnology
Molecular & Cellular Proteomics	Briefings in Functional Genomics & Proteomics
Mammalian Genome	Journal of Theoretical Biology

Πίνακας 1.3: *Τα περιοδικά βιοπληροφορικής που μελετήθηκαν στην εργασία των (Song, Kim, Zhang, Ding, & Chambers, 2014).*

Όσον αφορά την κατάταξη των πανεπιστημίων το Stanford University, το Harvard University και το University of Washington βρίσκονται σταθερά ψηλά στη σχετική λίστα. Το Stanford ήταν 3ο την περίοδο 2000-2003 και 1ο από το 2004 και μετά. Το Harvard ήταν 6ο, 2ο και 3ο αντίστοιχα στις περιόδους 2000-2003, 2004-2007 και 2009-2011. Τέλος, το University of Washington ήταν 5ο στις δύο πρώτες περιόδους και 2ο στην τελευταία. Τρία πανεπιστήμια είχαν σταθερά ανοδική πορεία στις αντίστοιχες περιόδους, το University

of Cambridge (11^{ο}, 8^{ο} και 5^{ο}), το University College London (11^{ο}, 11^{ο} και 10^{ο}), ενώ το University of Oxford δεν ήταν καν στη λίστα με τα κορυφαία ιδρύματα την περίοδο 2000-2003, αλλά ανέβηκε στην 10^{η} θέση την περίοδο 2004-2008 και στην 6^{η} την περίοδο 2009-2011. Από την άλλη μεριά, το Brandeis University που ήταν 1^{ο} την περίοδο 2000-2003, έπεσε στην 12^{η} θέση την περίοδο 2004-2007 και δεν συμπεριλαμβάνεται καν στη λίστα την περίοδο 2008-2011. Παρόμοια πτωτική πορεία είχε το University of California, Berkeley το οποίο έπεσε από τη 2^{η} θέση στην 7^{η} και τελικά στην 14^{η} στις αντίστοιχες περιόδους.

Η ανάλυση έδωσε επίσης δεδομένα για τις πιο επιδραστικές εργασίες στο χώρο καθώς και πληροφορίες για τους συγγραφείς που συμμετείχαν σε αυτές. Η ανάλυση αυτή είναι ιδιαίτερα χρήσιμη γιατί έτσι μπορούν να αναγνωριστούν κάποια από τα ερευνητικά πεδία που κυριάρχησαν στις επόμενες περιόδους. Για την περίοδο 2000-2003 η εργασία με τις περισσότερες αναφορές είχε τίτλο "*Gene ontology: tool for the unification of biology*" και δημοσιεύτηκε στο Nature Genetics με συγγραφείς το Gene Ontology Consortium το οποίο αποτελούταν από 20 ερευνητές. Οκτώ από αυτούς συγκαταλέγονται στους πιο επιδραστικούς συγγραφείς αυτής της χρονικής περιόδου (D. Botstein, G. Rubin, G. Sherlock, M. Ashburner, J. Cherry, C. Ball, J. Matese, H. Butler). Η δεύτερη σε αριθμό αναφορών εργασία, ήταν η δημοσίευση του ανθρώπινου γονιδιώματος ("*Initial sequencing and analysis of the human genome*") στο Nature. Οι συγγραφείς ήταν 249 από 48 διαφορετικά ινστιτούτα. Η 3^{η} πιο σημαντική εργασία αυτής της περιόδου ήταν το "*Significance analysis of microarrays applied to the ionizing radiation response*" από τους V. Tusher, R. Tibshirani, και G. Chu, από το Stanford. Ο R. Tibshirani επίσης εμφανίζεται στη 12^{η} θέση των συγγραφέων την ίδια περίοδο. Κατά την περίοδο 2004-2007, η εργασία με τις περισσότερες αναφορές είχε τίτλο "*Bioconductor: open software development for computational biology and bioinformatics*" και την συνέγραψαν 25 συγγραφείς από 19 ινστιτούτα. Ανάμεσα τους 4 βρίσκονται στη λίστα με τους πιο επιδραστικούς επιστήμονες της ίδιας περιόδου. Η δεύτερη εργασία της περιόδου αυτής ήταν η εργασία που περιέγραφε το στατιστικό πακέτο R, "*R: A language and environment for statistical computing*" και η 3^{η} είχε τίτλο "*Transcriptional regulatory code of a eukaryotic genome*" από 20 συγγραφείς από 4 διαφορετικά ινστιτούτα. Τέλος, κατά την περίοδο 2008-2011, η εργασία με τις περισσότερες αναφορές ήταν η εργασία που παρουσίαζε τη βάση δεδομένων PFAM "*The Pfam protein families database*" με 13 συγγραφείς από 3 διαφορετικά ινστιτούτα (ανάμεσά τους ο A Bateman και ο R. Durbin, οι οποίοι βρίσκονται σταθερά μέσα στη λίστα των πιο επιδραστικών επιστημόνων του χώρου). Η δεύτερη στη σειρά εργασία αφορούσε την περιγραφή της KEGG, "*KEGG for linking genomes to life and the environment*" με 11 συγγραφείς από 3 διαφορετικά ινστιτούτα της Ιαπωνίας, ενώ η τρίτη είχε τίτλο "*Mapping short DNA sequencing reads and calling variants using mapping quality scores*" με συγγραφείς τους H. Li, J Ruan και R. Durbin. Αυτά τα δεδομένα δείχνουν την κατεύθυνση της έρευνας την τελευταία 10ετία και βρίσκονται σε συμφωνία με τα δεδομένα που αναφέρθηκαν στις προηγούμενες μελέτες (ανάπτυξη των μεθόδων αλληλούχισης, ανάπτυξη στατιστικών μεθόδων ελέγχου της γονιδιακής έκφρασης, ανάπτυξη του ανοιχτού λογισμικού βιοπληροφορικής αλλά και των βάσεων δεδομένων και των οντολογιών).

Από την ανάλυση των αναφορών προέκυψε επίσης και η κατάταξη των επιδραστικών περιοδικών του χώρου της βιοπληροφορικής, η οποία βέβαια αναμένουμε να έχει παρόμοια δομή με την αντίστοιχη κατάταξη με βάση το Impact Factor. Έτσι, βλέπουμε ότι σε όλες τις περιόδους τα περιοδικά Proceedings of the National Academy of Sciences, Nucleic Acids Research, Nature, Bioinformatics και Science βρίσκονται σταθερά στην κορυφαία πεντάδα των περιοδικών. Το BMC Bioinformatics το οποίο έκανε την εμφάνιση του σχετικά αργότερα είναι 6^{ο} την περίοδο 2004-2007 και 5^{ο} την περίοδο 2008-2011. Εκτός από το BMC Bioinformatics, την περίοδο 2004-2007 εμφανίζεται και το PLoS Biology, το BMC Genomics, και το Nature Reviews Genetics με τη σειρά κατάταξής τους να αυξάνει συνεχώς. Νέα περιοδικά που έκαναν δυναμική εμφάνιση στα 20 καλύτερα την περίοδο 2008-2009 ήταν το PLoS One, PLoS Genetics, PLoS Computational Biology, Nature Biotechnology και το Nature Methods. Αξίζει να σημειωθεί εδώ ότι τα περισσότερα από τα κορυφαία περιοδικά (με εξαίρεση τα Bioinformatics, BMC Bioinformatics και PLoS Computational Biology) είναι βιολογικά περιοδικά ή περιοδικά γενικότερου ενδιαφέροντος στα οποία δημοσιεύονται και εργασίες βιοπληροφορικής. Αυτό αφενός κάνει δύσκολη την αναζήτηση της σχετικής βιβλιογραφίας για παρόμοιες αναλύσεις, αλλά παράλληλα δείχνει και τη σπουδαιότητα που έχει η βιοπληροφορική στο πλαίσιο της σύγχρονης βιολογικής και βιοϊατρικής έρευνας.

Συνολικά, από τη μελέτη αυτή προέκυψε ότι το πεδίο της βιοπληροφορικής έχει υποστεί σημαντικές αλλαγές κατά την τελευταία 10ετία και εξελίσσεται παράλληλα με άλλες ειδικότητες της βιοϊατρικής, καθώς η εστίαση και το αντικείμενο της μελέτης αλλάζει με την πάροδο του χρόνου (γονιδιακή έκφραση, γονιδιώματα, πολυμορφισμοί, πολύπλοκα συστήματα κ.ο.κ.). Η ανάπτυξη των υπολογιστικών προσεγγίσεων

έχει βοηθήσει την εξάπλωση των βιολογικών βάσεων δεδομένων και των εργαλείων ανάλυσής τους ακόμα και σε άλλες συγγενικές ειδικότητες. Ειδικότερα, οι υπολογιστικές τεχνικές έγιναν αναπόσπαστο τμήμα της βιοϊατρικής έρευνας την περίοδο 2000-2003, ενώ μετά το 2004 αυξήθηκε και η ανάγκη δημιουργίας και διαχείρισης βάσεων βιολογικών δεδομένων. Τέλος, σημαντικά εργαλεία και μεθοδολογίες της βιοπληροφορικής που αναπτύχθηκαν αυτή την περίοδο, όπως οι μικροσυστοιχίες οι οντολογίες και οι μεθοδολογίες ανάλυσης βιολογικών δικτύων έγιναν βασικά κομμάτια της σύγχρονης βιοπληροφορικής έρευνας αλλά όπως είδαμε και προηγουμένως, οι μικροσυστοιχίες και οι οντολογίες έγιναν και βασικά κομμάτια της ιατρικής πληροφορικής.

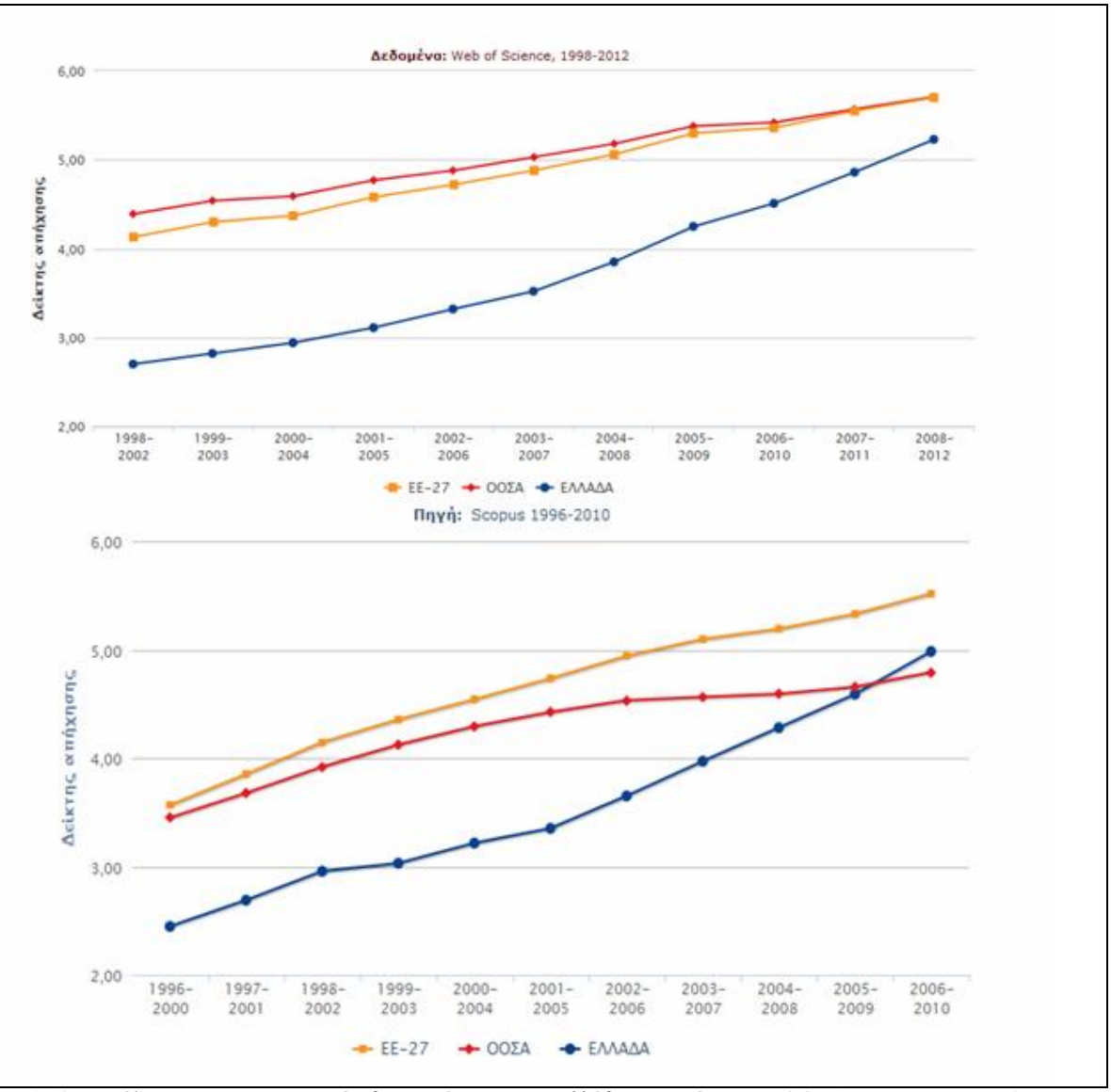
1.4. Η κατάσταση στην Ελλάδα

Στην ενότητα αυτή θα μελετήσουμε την κατάσταση της βιοπληροφορικής στην Ελλάδα. Θα δούμε τις επιστημονικές δημοσιεύσεις που έχουν προέλθει από Ελληνικά ιδρύματα, την κατάσταση της επιστημονικής κοινότητας της βιοπληροφορικής στην Ελλάδα, αλλά και την εκπαίδευση.

1.4.1. Η έρευνα στην Ελλάδα

Γενικά, παρά τις περί του αντιθέτου κραυγές που ακούγονται καθημερινά στα ΜΜΕ, η επιστημονική έρευνα στην Ελλάδα πάει καλά (ειδικά αν αναλογιστούμε τα χρήματα που δαπανώνται σε αυτή, βλ. παρακάτω). Όλες οι αποτιμήσεις της ερευνητικής δραστηριότητας που έχουν γίνει τα τελευταία χρόνια από το Εθνικό Κέντρο Τεκμηρίωσης (ΕΚΤ), δείχνουν ότι τόσο ο αριθμός των δημοσιεύσεων όσο και των αναφορών που παίρνουν εργασίες προερχόμενες από ελληνικά ιδρύματα αυξάνονται συνεχώς, αλλά το πιο σημαντικό είναι ότι αυξάνονται και οι ποιοτικοί δείκτες με αποτέλεσμα η Ελλάδα να συγκλίνει σταδιακά προς το μέσο όρο των χωρών της ΕΕ και του ΟΟΣΑ (Εικόνα 1.13) σε μέτρα που αφορούν την ποιότητα (π.χ. το μέσο αριθμό αναφορών ανά εργασία). Όλες οι μελέτες του ΕΚΤ την τελευταία δεκαετία συγκλίνουν στο αποτέλεσμα αυτό, ανεξαρτήτως της βάση δεδομένων βιβλιογραφίας που χρησιμοποιείται (Σαχίνη, Μάλλιου, & Χούσος, 2012; Σαχίνη, Μάλλιου, Χούσος, & Καραϊσκος, 2013; Σαχίνη, Μάλλιου, Χούσος, & Καραΐσκος, 2014) και το ίδιο φαίνεται να ισχύει ειδικά και για τον τομέα των βιοϊατρικών επιστημών (Σαχίνη, Μάλλιου, Χούσος, & Καραϊσκος, 2012). Οι ίδιες μελέτες δείχνουν καθαρά ότι το μεγαλύτερο σε όγκο μέρος της ερευνητικής δραστηριότητας της χώρας προέρχεται από τα Ελληνικά Πανεπιστήμια, αν και φυσικά υπάρχουν και ιδιαίτερα αξιόλογα Ερευνητικά Κέντρα, αλλά και νησίδες αριστείας στα ΤΕΙ.

Φυσικά, πάντα χρειάζεται προσοχή στο πώς σταθμίζουμε τέτοιες αναλύσεις γιατί θα πρέπει να λαμβάνουμε υπόψη μας και παράγοντες όπως ο πληθυσμός μιας χώρας αλλά και το ΑΕΠ (Ακαθάριστο Εγχώριο Προϊόν) αυτής. Για παράδειγμα, το επιστημονικό περιοδικό Nature, δημοσίευσε πρόσφατα μια μελέτη (http://www.natureindex.com) που απαριθμεί τα άρθρα υψηλής επιστημονικής απήχησης του τελευταίου χρόνου (τις εργασίες που έχουνε δημοσιευθεί σε μια συλλογή από τα σημαντικότερα διεθνή επιστημονικά περιοδικά). Με βάση αυτούς τους πίνακες (Εικόνα 1.14), που μετράνε παραγωγικότητα έρευνας υψηλού επιπέδου χωρίς καμία άλλη στάθμιση (δείκτης WFC), η Ελλάδα, κατατάσσεται στην 32η θέση στον κόσμο, ενώ βρίσκεται στην προτελευταία ανάμεσα στις χώρες της Ευρωπαϊκής Ένωσης (ΕΕ) πριν τη διεύρυνση του 2004 (προσπερνά το Λουξεμβούργο).

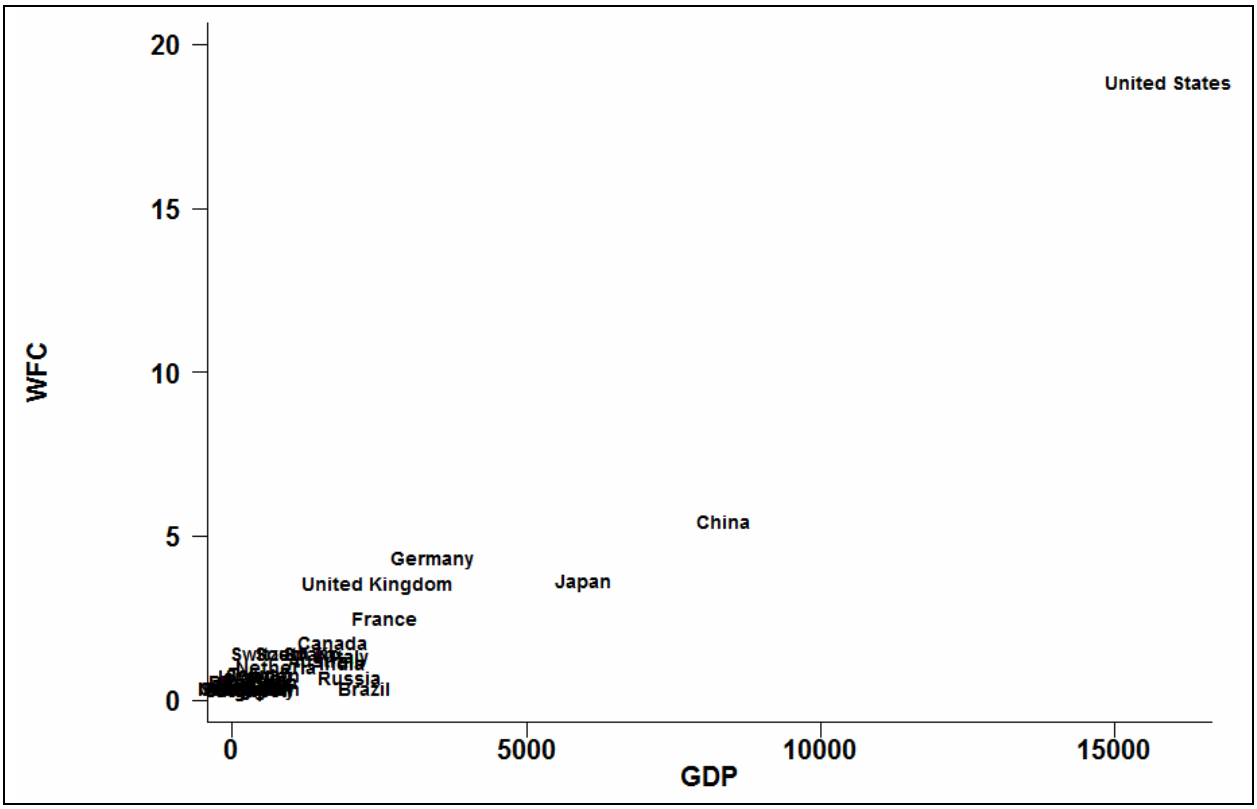


Εικόνα 1.13: *Η αύξηση των επιστημονικών δημοσιεύσεων στην Ελλάδα, στις χώρες του ΟΟΣΑ και στην ΕΕ των 27.*

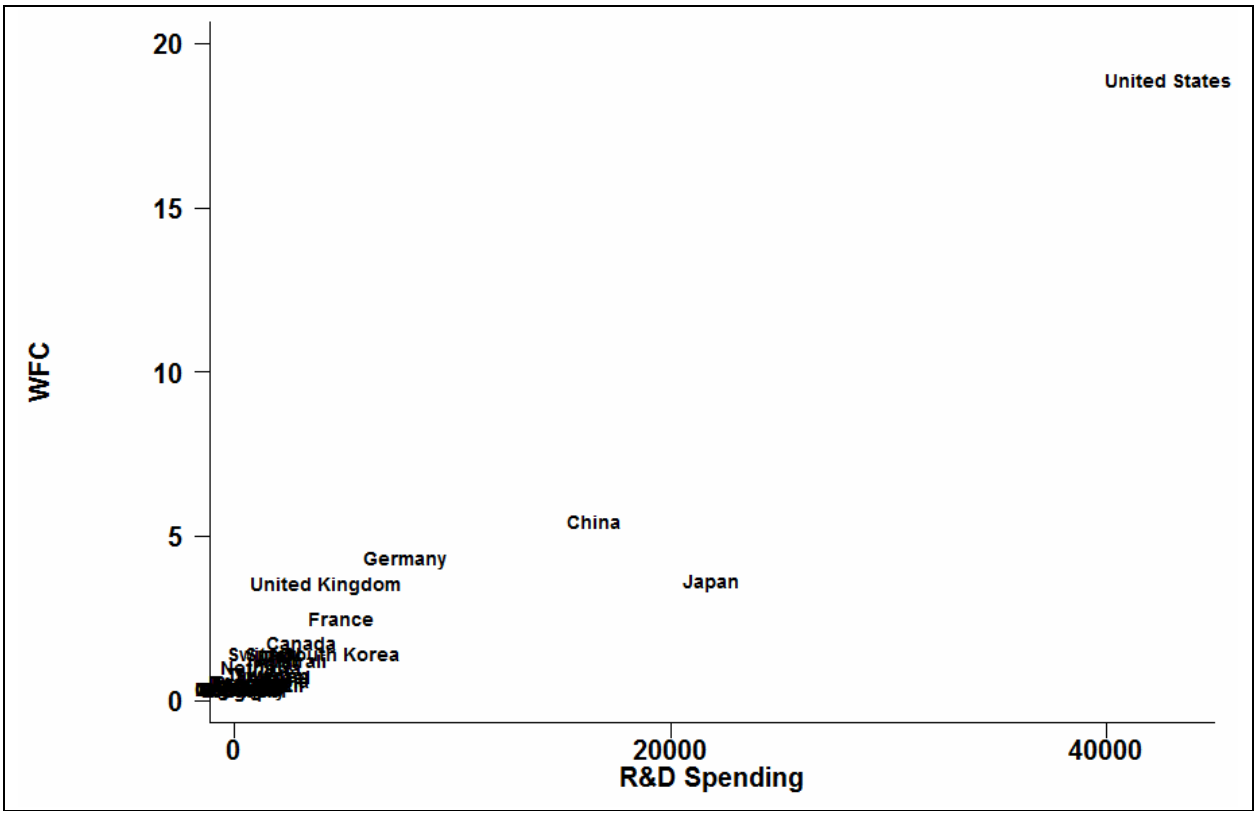
Country	AC	FC	WFC
United States of America (USA)	26590	19225.62	17220.79
China	9766	6998.83	6652.23
Germany	9057	4659.45	4079.92
United Kingdom (UK)	8288	4002.57	3367.43
France	5437	2528.30	2137.20
Japan	4995	3293.35	3073.94
Italy	3392	1455.80	1068.70
Canada	3295	1641.46	1485.02
Spain	3206	1364.05	1093.13
Switzerland	2943	1254.76	1165.33
Australia	2655	1145.23	933.60
Netherlands	2447	891.67	729.63
South Korea	2085	1227.33	1143.33
Sweden	1599	603.14	526.39
India	1563	1030.53	910.47
Russia	1325	466.38	364.36
Israel	1206	590.32	527.75
Denmark	1196	366.32	318.29
Belgium	1159	400.67	336.19
Chile	1014	215.24	99.61
Taiwan	990	456.32	414.47
Poland	943	318.70	234.56
Brazil	942	330.25	236.12
Singapore	923	506.81	505.82
Austria	916	309.33	275.07
Finland	699	219.01	186.95
Czech Republic	563	173.82	136.32
Portugal	552	151.62	128.06
South Africa	546	123.39	76.18
Norway	532	149.48	128.20
Mexico	485	147.64	83.68
Saudi Arabia	470	99.13	97.13
Greece	431	107.70	82.45
Ireland	426	125.55	115.75

Εικόνα 1.14: *Η κατάταξη των χωρών με βάση το www.natureindex.com*

Βρίσκεται λοιπόν η Ελληνική επιστημονική έρευνα σε τόσο μεγάλη απαξίωση; Μια προσεκτικότερη ματιά στους πίνακες δείχνει κάτι απλό: η σειρά των χωρών στη λίστα θυμίζει πάρα πολύ τη σειρά των χωρών με βάση το συνολικό ΑΕΠ τους, έναν δείκτη που αντικατοπτρίζει την οικονομική δραστηριότητα αλλά και το μέγεθος της κάθε χώρας (η Ελλάδα στη σχετική λίστα βρίσκεται, παρά την ύφεση, στην 43η θέση παγκοσμίως, http://data.worldbank.org/, στοιχεία 2012-2013). Κατά συνέπεια, ο συντελεστής συσχέτισης του δείκτη WFC με το ΑΕΠ των χωρών της ΕΕ, αλλά και με το συνολικό ποσό επένδυσης στην έρευνα για κάθε χώρα, είναι περίπου 95%. Όσο περισσότερα λεφτά έχεις, τόσο περισσότερη έρευνα υψηλής ποιότητας παράγεις. Καμία έκπληξη. Αξίζει να σημειωθεί, ότι αν η ανάλυση επαναληφθεί στις πρώτες 33 χώρες στη λίστα του Nature (που περιλαμβάνει πάνω-κάτω τις περισσότερες χώρες του ΟΟΣΑ), τα αποτελέσματα είναι παρόμοια.



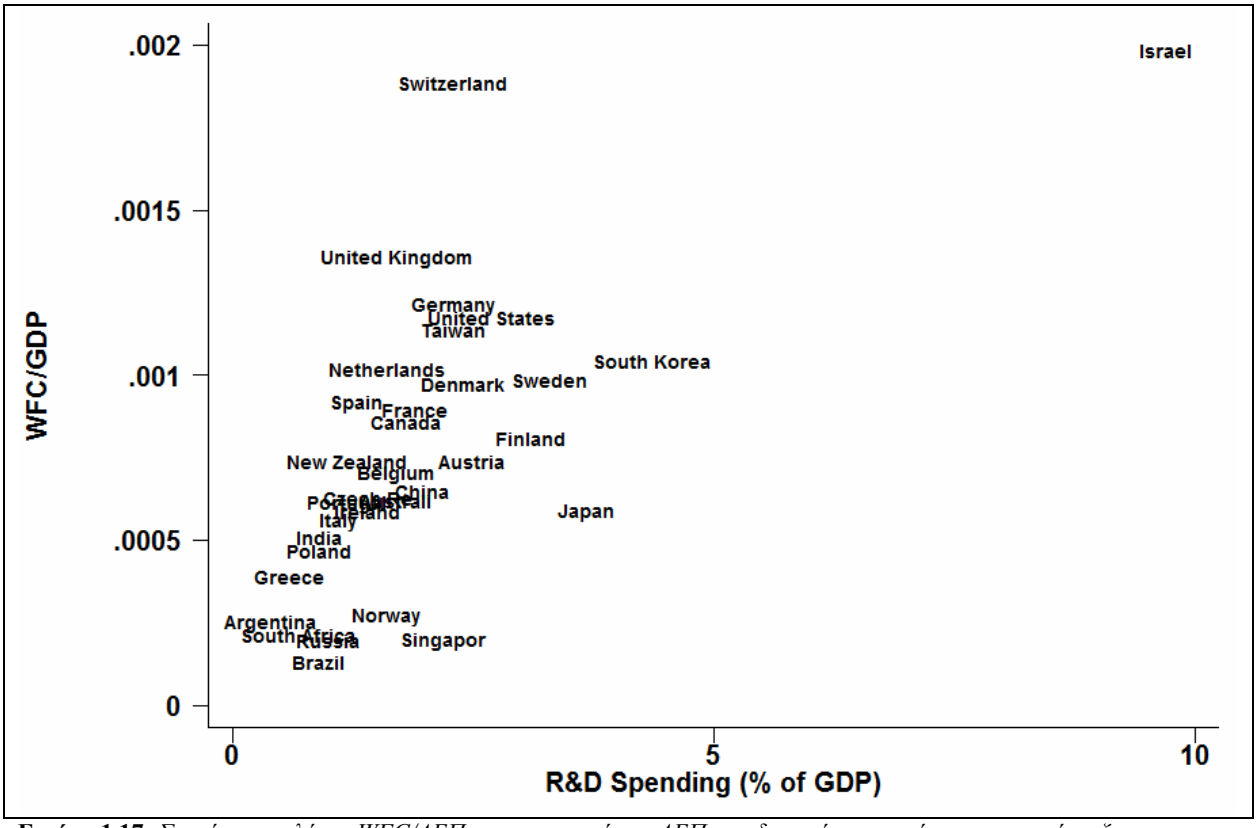
Εικόνα 1.15: *Συσχέτιση του δείκτη WFC με το ΑΕΠ της κάθε χώρας.*



Εικόνα 1.16: *Συσχέτιση του δείκτη WFC με το ποσό που δαπανάται για έρευνα και ανάπτυξη.*

Ένας σταθμισμένος δείκτης που είναι ανεξάρτητος από την ποσότητα της έρευνας και το μέγεθος της χώρας, βασίζεται στις αναφορές των δημοσιευμένων επιστημονικών εργασιών από άλλες επιστημονικές εργασίες. Εκφράζεται ως ο λόγος των αναφορών προς τον αριθμό των εργασιών και κάνει μια εκτίμηση της απήχησης της έρευνας, χωρίς παραδοχές για την ποιότητα του περιοδικού δημοσίευσης. Πρόκειται δηλαδή για το ίδιο κριτήριο με αυτό που χρησιμοποίησε το ΕΚΤ στις αναλύσεις του, που αναφέραμε παραπάνω, αλλά για ευκολία χρησιμοποιήσαμε δεδομένα από άλλη πηγή (http://www.scimagojr.com, στοιχεία 2011-2013). Επειδή αυτός ο δείκτης είναι ήδη σταθμισμένος, δείχνει μια μάλλον αμελητέα συσχέτιση με το ΑΕΠ και το συνολικό ποσό επένδυσης στην έρευνα, αλλά μια αξιοπρόσεκτη συσχέτιση μόνο με το ποσοστό του ΑΕΠ που επενδύεται στην έρευνα, ~70%. Όσο περισσότερα λεφτά βάζεις στην έρευνα αναλογικά με το ΑΕΠ, τόσο πιο μεγάλη απήχηση έχει η έρευνα που παράγεις.

Με βάση τις υψηλές συσχετίσεις αυτών των τριών ζευγαριών αξιολόγησης, μπορέσαμε να φτιάξουμε ένα απλό μοντέλο που προβλέπει το δείκτη WFC με βάση το ΑΕΠ και με βάση το συνολικό ποσό επένδυσης στην έρευνα και το δείκτη ετεροαναφορών ανά εργασία με βάση το ποσοστό του ΑΕΠ που επενδύει κάθε χώρα στην έρευνα. Από αυτό το μοντέλο μπορούμε να υπολογίσουμε την απόκλιση, θετική ή αρνητική, κάθε χώρας από το αναμενόμενο. Σύμφωνα με αυτή την ανάλυση, η Ελλάδα τα πάει κατά 29% καλύτερα από το αναμενόμενο με βάση το ποσοστό του ΑΕΠ που επενδύεται στην έρευνα (και βρίσκεται στην δεύτερη θέση) και κατά 48% καλύτερα από το αναμενόμενο με βάση το συνολικό ποσό επένδυσης στην έρευνα, και βρίσκεται στην τρίτη θέση μαζί με το Ηνωμένο Βασίλειο (48%). Όταν όμως συγκρίνουμε την απόκλιση της Ελλάδας με βάση το συνολικό ΑΕΠ, αυτή βρίσκεται στο -15% από το αναμενόμενο, στη 19η θέση ανάμεσα στις χώρες της ΕΕ.



Εικόνα 1.17: *Συσχέτιση το λόγου WFC/ΑΕΠ με το ποσοστό του ΑΕΠ που δαπανάται στην έρευνα και ανάπτυξη.*

Η ανάλυση αυτή είναι φυσικά περιορισμένη, και η επιλεγμένη μεθοδολογία καθώς και τα πρωτογενή δεδομένα μπορούν να αμφισβητηθούν. Επίσης, θα πρέπει να σημειώσουμε ότι σε όλες αυτές τις αναλύσεις η συσχέτιση δεν σημαίνει υποχρεωτικά αιτιολογική σχέση, αν και ειδικά στην περίπτωση του ζεύγους ΑΕΠ και ερευνητικής παραγωγής, υπάρχουν εμπειρικά και θεωρητικά δεδομένα που να υποστηρίζουν μια τέτοια σχέση. Σε κάθε περίπτωση, η ανάλυση αυτή είναι χρήσιμη και μπορεί να βοηθήσει στην εξαγωγή κάποιων συμπερασμάτων. Το μήνυμα από μια τέτοια ανάλυση είναι μάλλον απλό: καμία μα καμία βελτίωση δεν είναι

δυνατόν να γίνει στην ποιότητα της Ελληνικής έρευνας εάν δεν αυξηθεί το ποσοστό του ΑΕΠ που επενδύεται στην έρευνα. Η Ελλάδα με βάση το απόλυτο ποσό αλλά και το ποσοστό του ΑΕΠ που επενδύεται στην έρευνα, τα πάει εξαιρετικά σε σχέση με τη συντριπτική πλειοψηφία άλλων Ευρωπαϊκών χωρών. Τα πάει όμως μάλλον άσχημα ή μέτρια σε σχέση με το (κουτσουρεμένο λόγω ύφεσης) ΑΕΠ της, για τον απλό λόγο ότι ελάχιστο ποσοστό του ΑΕΠ επενδύεται στην έρευνα. Από τα μικρότερα στην Ευρώπη και λιγότερο από το 1/3 του Ευρωπαϊκού στόχου για 2.1% επί του ΑΕΠ (0.69%) (Μπάγκος & Περράκης, 2014).

1.4.2. Η βιοπληροφορική έρευνα στην Ελλάδα

Ειδικότερα για την κοινότητα της βιοπληροφορικής στην Ελλάδα και την έρευνα που γίνεται στον τομέα αυτό, δεν υπάρχουν πολλά δημοσιευμένα δεδομένα. Τα τελευταία χρόνια όμως, έχουν γίνει σημαντικές προσπάθειες για την οργάνωση και την καταγραφή αυτής της δραστηριότητας στο πλαίσιο της ΕΕΥΒΒ. Η Ελληνική Εταιρεία Υπολογιστικής Βιολογίας και Βιοπληροφορικής (ΕΕΥΒΒ), http://www.hscbb.gr είναι η επιστημονική εταιρεία στην οποία συμμετέχουν δεκάδες επιστήμονες από την Ελλάδα και την Κύπρο, οι οποίοι ασχολούνται με την Υπολογιστική Βιολογία και τη Βιοπληροφορική. Είναι η μοναδική επιστημονική εταιρεία (σωματείο) με το αντικείμενο αυτό στην Ελλάδα, λειτουργεί με τη νομική μορφή σωματείου από το 2009 και είναι συνδεδεμένο μέλος (affiliated) της αντίστοιχης διεθνούς ομοσπονδίας (International Society for Computational Biology, βλ. http://www.iscb.org/iscb-affiliates-europe#hellenic), ενώ συμμετέχει και σε διάφορες άλλες διεθνείς πρωτοβουλίες όπως το GOBLET (http://www.mygoblet.org) και το ELIXIR (https://www.elixir-europe.org/).

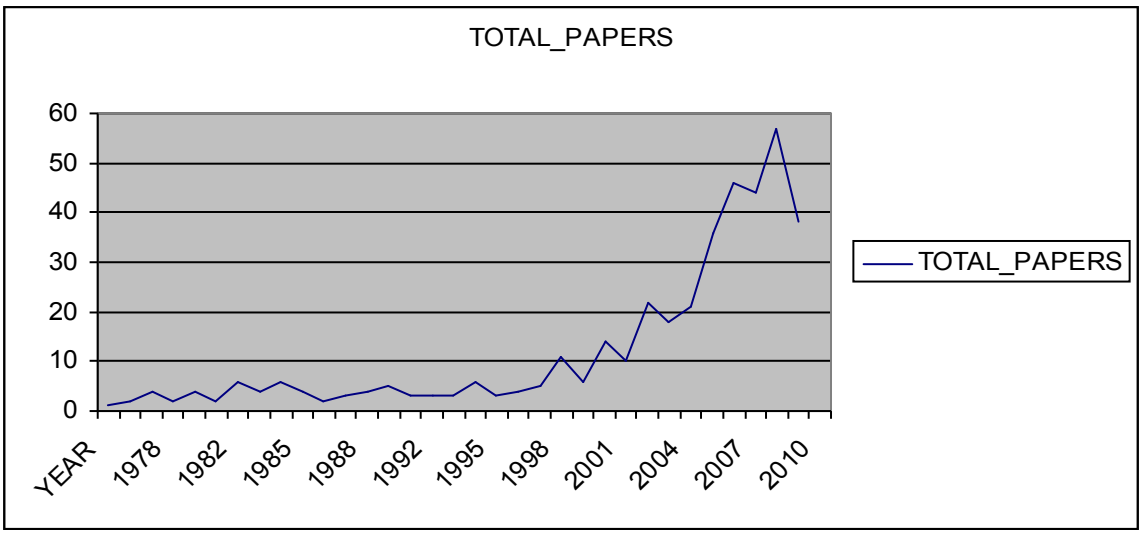
Τα συνέδρια της ΕΕΥΒΒ, διεξάγονται σε ετήσια βάση με μεγάλη επιτυχία και προβολή (π.χ. http://hscbb11.hscbb.gr, http://hscbb12.hscbb.gr κ.ο.κ.) ενώ συμμετέχουν σταθερά ως προσκεκλημένοι σε αυτά ομιλητές, διακεκριμένοι επιστήμονες από το εξωτερικό. Προηγούμενα συνέδρια έχουν γίνει σε πόλεις με σχετικά με το επιστημονικό πεδίο πανεπιστημιακά τμήματα και ερευνητικά ιδρύματα, π.χ. στην Αθήνα, στην Πάτρα, στο Ηράκλειο, στην Αλεξανδρούπολη και στη Λαμία. Αξίζει να σημειωθεί, ότι στα προηγούμενα συνέδρια, οι σύνεδροι ξεπερνούν τους 120 με συνεχώς αυξανόμενους αριθμούς, ενώ μεγάλο μέρος από αυτούς είναι προπτυχιακοί και μεταπτυχιακοί φοιτητές. Τα ενεργά μέλη της εταιρίας (μέλη ΔΕΠ, Ερευνητές, και γενικότερα κάτοχοι διδακτορικού) είναι περίπου 40 από όλη την Ελλάδα, αλλά ο συνολικός αριθμός είναι μεγαλύτερος καθώς κάποιοι συμμετέχουν μόνο περιστασιακά. Γενικά η ενεργή κοινότητα της βιοπληροφορικής στην Ελλάδα αριθμεί πάνω από 30 ερευνητικές ομάδες σε διάφορα πανεπιστήμια και ερευνητικά ιδρύματα, έστω και αν για πολλούς από αυτούς η βιοπληροφορική δεν είναι το μόνο ή το κύριο ερευνητικό αντικείμενο.

Σε μια προσπάθεια να καταγραφεί αναλυτικά η ερευνητική δραστηριότητα στην Ελλάδα, είχε γίνει μια σχετική εργασία που παρουσιάστηκε στο ετήσιο συνέδριο της ΕΕΥΒΒ το 2010 (Bagos, 2010). Σε αυτή την εργασία έγινε συστηματική προσπάθεια να αναλυθεί η βιβλιογραφία της βιοπληροφορικής συλλέγοντας όλες της εργασίες στις οποίες συμμετείχαν συγγραφείς με διεύθυνση εργασίας κάποιο Ελληνικό ίδρυμα και έγινε ανάλυση που αφορά τον αριθμό των εργασιών και των αναφορών τους, τους συγγραφείς, τα ινστιτούτα τους αλλά και τα ερευνητικά ενδιαφέροντα και τις τάσεις τους στην πορεία των χρόνων. Και σε αυτή την περίπτωση τα βασικά προβλήματα αυτών των εργασιών παραμένουν: δηλαδή το πώς θα συλλεχθεί ένα όσο το δυνατό μεγαλύτερο σύνολο από εργασίες από διάφορα περιοδικά (και κυρίως πώς θα διαχωριστούν με ακρίβεια τα περιοδικά βιοπληροφορικής), από ποια βάση δεδομένων θα γίνει η καταγραφή των δεδομένων, και με ποιον τρόπο θα γίνει η ανάλυση του κειμένου. Η επιλογή σε αυτή την περίπτωση ήταν να στηριχθούμε στην γενικότερη κατηγορία του ISI WoS, με τον τίτλο "MATHEMATICAL & COMPUTATIONAL BIOLOGY" και να συλλεχθούν όλα τα περιοδικά αυτής της κατηγορίας. Η κατηγορία αυτή περιέχει τα μεγαλύτερα αμιγώς βιοπληροφορικά περιοδικά, αλλά και κάποια αξιόλογα περιοδικά ιατρικής πληροφορικής και βιοστατιστικής. Σε μια προσπάθεια να διευρυνθεί το σύνολο δεδομένων, επιλέχθηκε και μια επιπλέον λίστα περιοδικών που έχουν ξεκάθαρη αναφορά στον τίτλο τους σε «βιοπληροφορική» ή «υπολογιστική βιολογία» αλλά και τα ειδικά τεύχη του Nucleic Acids Research που είναι αφιερωμένα σε εφαρμογές βιοπληροφορικής (web-server και database issues). Όλα τα περιοδικά αυτά, ήταν ενταγμένα στη βάση δεδομένων PUBMED (ακόμα και αν δεν ήταν στο WoS). Τα περιοδικά και των δύο κατηγοριών που χρησιμοποιήθηκαν στην ανάλυση δίνονται στον Πίνακα 1.4. Προφανώς, ένας σημαντικός περιορισμός του τρόπου αναζήτησης ήταν ότι πολλές εργασίες βιοπληροφορικής, ή εργασίες που έκαναν εκτεταμένη χρήση υπολογιστικών μεθόδων και είχαν δημοσιευτεί σε καθαρά βιολογικά περιοδικά (JMB, Plos Biology, Protein Engineering κ.ο.κ.), αλλά και εργασίες στα κορυφαία περιοδικά γενικού ενδιαφέροντος (Science, Nature), δεν

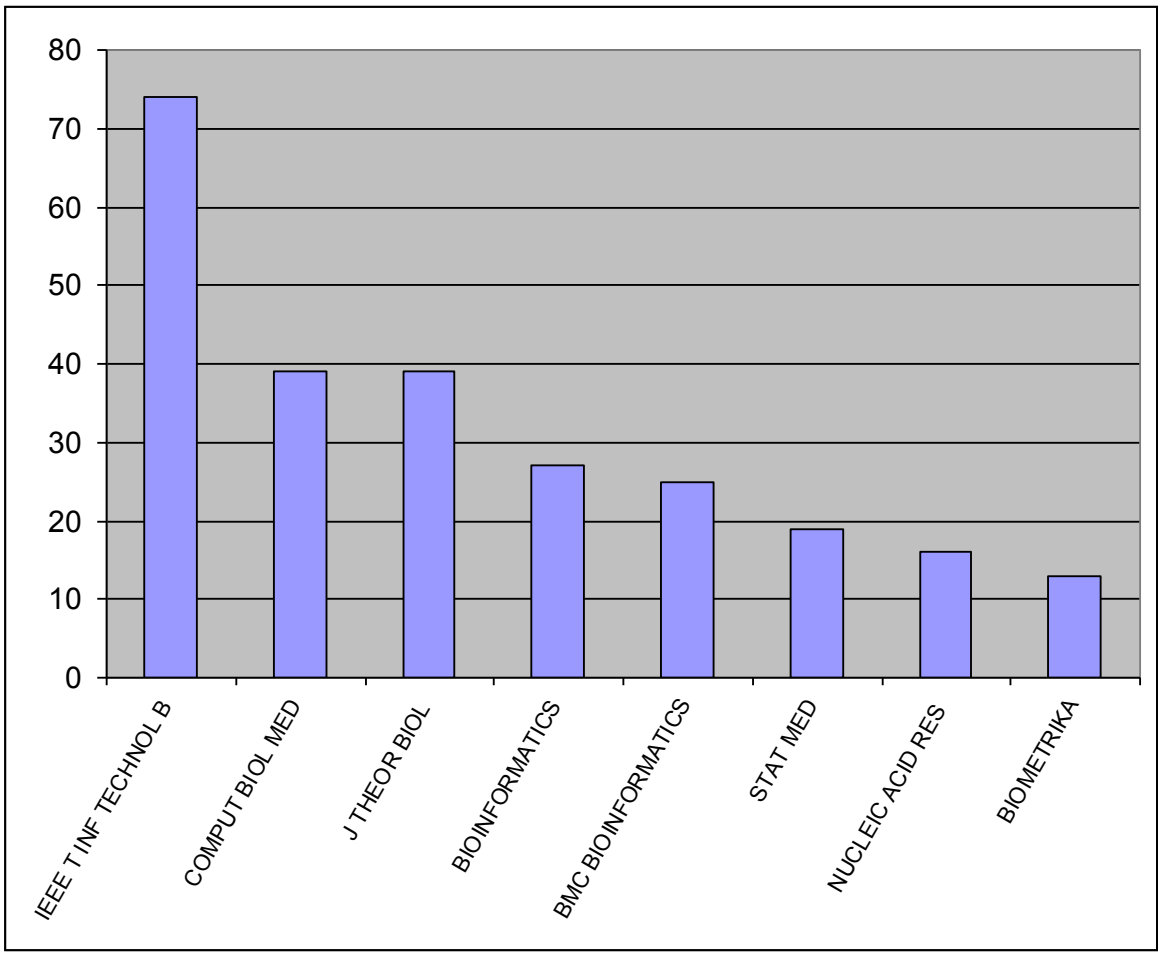
συμπεριλήφθηκαν στη μελέτη. Το ίδιο ισχύει και για αντίστοιχες εργασίες που έχουν δημοσιευθεί σε περιοδικά της πληροφορικής (Machine Learning, Pattern Recognition κ.ο.κ.). Κατά συνέπεια, τα πορίσματα αυτά αποτελούν υπο-εκτίμηση της πραγματικής διάστασης της σχετικής βιβλιογραφίας στην Ελλάδα. Επίσης, για μια σειρά από σημαντικούς επιστήμονες του χώρου, μεγάλο μέρος από τις δημοσιευμένες εργασίες τους είχαν πραγματοποιηθεί όταν αυτοί εργάζονταν σε ιδρύματα του εξωτερικού, κατά συνέπεια δεν έχουν συμπεριληφθεί στην ανάλυση. Οι αναζητήσεις έγιναν με επιπλέον λέξη-κλειδί τη χώρα προέλευσης (GREECE ή CYPRUS), τα δεδομένα αποθηκεύτηκαν σε μια βάση δεδομένων SQL και αναλύθηκαν με στατιστικά εργαλεία και η ανάλυση του κειμένου έγινε με τη σχετική εφαρμογή του Yahoo, το γνωστό «Term Extraction web service» (http://developer.yahoo.com/search/content/V1/termExtraction.html), το οποίο απομονώνει από τις περιλήψεις τις σημαντικές λέξεις-κλειδιά (η επιλογή αυτή έγινε γιατί είδαμε ότι τα KEYWORDS του ίδιου του WoS είναι πολλές φορές αποπροσανατολιστικά ή πολύ γενικά).

PLOS Comput Biol	Journal of Computer-Aided Molecular Design
Bioinformatics	Nucleic Acids Research (web-server and database issues)
BMC Syst Biol	The Open Bioinformatics Journal
BMC Bioinformatics	Statistical Applications in Genetics and Molecular Biology
Biostatistics	Source Code for Biology and Medicine
J Theor Biol	Online Journal of Bioinformatics
Stat Method Med Res	Journal of Integrative Bioinformatics
IET Syst Biol	Journal of Bioinformatics and Computational Biology
J Comput Neurosci	International Journal of Data Mining and Bioinformatics
J Mol Graph Model	International Journal of Computational Biology and Drug Design
Stat Med	International Journal of Bioinformatics Research and Applications
Biometrika	In Silico Biology
Evol Bioinform	Genomics, Proteomics & Bioinformatics
B Math Biol	Genome Informatics
Biometrics	EURASIP Journal on Bioinformatics and Systems Biology
Algorithm Mol Biol	Current Bioinformatics
Med Biol Eng Comput	BioData Mining
J Math Biol	Advances and Applications in Bioinformatics and Chemistry
IEEE T Inf Technol B	Applied Bioinformatics
J Comput Biol	International Journal of Bioinformatics
Curr Bioinform	Bioinformation
SAR QSAR Environ Res	Pac Symp Biocomput
Math Biosci	Database
Comput Biol Med	Genome Res
Biometrical J	BMC Genomics
Math Med Biol	
Int J Data Min Bioin	
J Agr Biol Envir St	
J Biol Syst	

Πίνακας 1.4: *Τα περιοδικά που χρησιμοποιήθηκαν στη δική μας ανάλυση για τη βιοπληροφορική στην Ελλάδα.*



Εικόνα 1.18: *Χρονική εξέλιξη των δημοσιεύσεων βιοπληροφορικής στην Ελλάδα.*



Εικόνα 1.19: *Τα περιοδικά με τις περισσότερες εργασίες που εντοπίστηκαν στη μελέτη για τη βιοπληροφορική στην Ελλάδα.*

Η ανάλυση έδωσε 405 εργασίες οι οποίες είχαν πραγματοποιηθεί από το 1976 μέχρι τις αρχές του 2010. Η αύξηση είναι, όπως και στην περίπτωση της διεθνούς βιβλιογραφίας, εκθετική μετά το 1999, καθώς

μέχρι εκείνη τη χρονιά είχαμε περίπου 5 εργασίες το χρόνο (Εικόνα 1.18). Συνολικά, στις 405 εργασίες είχαν συμμετάσχει 681 διαφορετικοί συγγραφείς (1,68 συγγραφείς ανά εργασία). Ο σχετικά μικρός αριθμός συγγραφέων ανά εργασία (σε σχέση με το αναμενόμενο), δικαιολογείται αν αναλογιστούμε ότι στη μελέτη περιλαμβάνονται και πολλές εργασίες, μαθηματικής βιολογίας και βιοστατιστικής, οι οποίες έχουν λίγους συγγραφείς, πολλές φορές και μόνο έναν. Από τους 681 συγγραφείς, οι 636 είχαν εμπλακεί σε 3 το πολύ εργασίες ενώ οι 45 είχαν συμμετάσχει σε περισσότερες, ενώ μόλις 18 είχαν συμμετάσχει σε περισσότερες από 9 εργασίες. Τα δεδομένα αυτά είναι συμβατά με τα αντίστοιχα στη διεθνή βιβλιογραφία που αναλύσαμε παραπάνω (ειδικά αν αναλογιστούμε ότι κάποιοι συγγραφείς που εμφανίζονται με λίγες εργασίες είχαν δημοσιεύσει περισσότερες όταν εργάζονταν στο εξωτερικό). Συνολικά, εντοπίστηκαν 63 διαφορετικά εκπαιδευτικά και ερευνητικά ιδρύματα, τα περισσότερα εκ των οποίων ήταν πανεπιστήμια.

Από τους συγγραφείς, ο πιο επιδραστικός είναι ο ομότιμος καθηγητής Σταύρος Χαμόδρακας από το ΕΚΠΑ, τόσο σε απόλυτο αριθμό εργασιών, όσο και σε αριθμό αναφορών αλλά και συνολικό Impact Factor. Το γεγονός αυτό είναι κάτι αναμενόμενο, καθώς ο καθηγητής Χαμόδρακας ήταν από τους πρώτους που ασχολήθηκαν με τη βιοπληροφορική στην Ελλάδα και ήταν ο συγγραφέας των 2 από τις 3 εργασίες που είχαν δημοσιευθεί από ερευνητές Ελληνικών ιδρυμάτων στο περιοδικό Computer Applications in Biosciences, τον πρόδρομο του πιο γνωστού περιοδικού στο χώρο, του Bioinformatics. Οι εργασίες αυτές, είχαν τίτλο «*A protein secondary structure prediction scheme for the IBM PC and compatibles*» του 1988 και «*PBM: a software package to create, display and manipulate interactively models of small molecules and proteins on IBM-compatible PCs*» του 1995 (με Perrakis A, Constantinides C, Athanasiades A), εργασίες που δικαίως μπορούν να χαρακτηριστούν τόσο ως οι πρώτες καθαρά βιοπληροφορικές εργασίες στην Ελλάδα, όσο και αντιπροσωπευτικές του ερευνητικού κλίματος της εποχής εκείνης.

Η πιο επιδραστική εργασία πριν το 2000 ήταν η εργασία των Fickett JW, Hatzigeorgiou AC. με τίτλο «*Eukaryotic promoter recognition*» στο Genome Res. 2η ήταν η εργασία των Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C, Ouzounis CA με τίτλο «*CAST: an iterative algorithm for the complexity analysis of sequence tracts*» στο Bioinformatics και 3η η εργασία των Pavlou S, και Kevrekidis IG με τίτλο «*Microbial predation in a periodically operated chemostat- a global study of the interaction between natural and externally imposed frequencies*», η οποία δημοσιεύθηκε στο Math Biosci. Την περίδο 2001-2005, πιο επιδραστική εργασία ήταν των Carninci P, Waki K, Shiraki T, Konno H, Shibata K, Itoh M, Aizawa K, Arakawa T, Ishii Y, Sasaki D et al. (με συμμετοχή του Έλληνα ερευνητή V Aidinis) με τίτλο «*Targeting a complex transcriptome: The construction of the mouse full-length cDNA encyclopedia*» στο περιοδικό Genome Res, 2η ήταν η εργασία των Patrinos GP, Giardine B, Riemer C, Miller W, Chui DHK, Anagnou NP, Wajcman H, Hardison RC με τίτλο «*Improvements in the HbVar database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies*» στο Nucleic Acids Res και 3η η εργασία των Bagos PG, Liakopoulos TD, Spyropoulos IC, Hamodrakas SJ με τίτλο «*PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins*», επίσης στο Nucleic Acids Res. Τέλος, την περίοδο 2006-2010, η πιο επιδραστική εργασία ήταν των Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC με τίτλο «*The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide*» στο Nucleic Acids Res, στη 2η θέση ήταν η εργασία της ίδια ομάδας (Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC) με τίτλο «The *Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata*» επίσης στο Nucleic Acids Res, ενώ στην 3η θέση ήταν η εργασία των Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z et al. με τίτλο «*BioMagResBank*» και αυτή στο ίδιο περιοδικό (Nucleic Acids Res).

Η ανάλυση των 20 εργασιών με τις περισσότερες αναφορές, μας δείχνει ότι ανάμεσά τους περιλαμβάνονται 7 εργασίες που περιγράφουν βιολογικές βάσεις δεδομένων και 4 εργασίες με web-servers ή μεθόδους πρόγνωσης. Όμοια, ανάλυση των 20 εργασιών με τις περισσότερες αναφορές/χρόνο δείχνει ότι ανάμεσά τους βρίσκονται 8 εργασίες που περιγράφουν βιολογικές βάσεις δεδομένων και 9 εργασίες με web-servers ή μεθόδους πρόγνωσης. Μια στατιστική ανάλυση έδειξε ότι οι σημαντικότεροι παράγοντες που «προβλέπουν» τον αριθμό αναφορών που θα πάρει μια εργασία (εκτός από το Impact Factor του περιοδικού, κάτι το οποίο είναι αναμενόμενο), είναι ο αριθμός των συγγραφέων (όσο περισσότεροι, τόσο το καλύτερο), η συμμετοχή συγγραφέων από το εξωτερικό και το αν η εργασία περιγράφει μια βιολογική βάση δεδομένων ή όχι. Ο αριθμός των συγγραφέων και η συμμετοχή ξένων επιστημόνων φαίνεται ότι είναι παράγοντες ενδεικτικοί της ποιότητας της εργασίας, αν και υπάρχουν υπόνοιες για κάποιου είδους συστηματικό σφάλμα (π.χ. οι εργασίες με ξένους επιστήμονες ενδέχεται να θεωρούνται καλύτερες και γι' αυτό να αναφέρονται

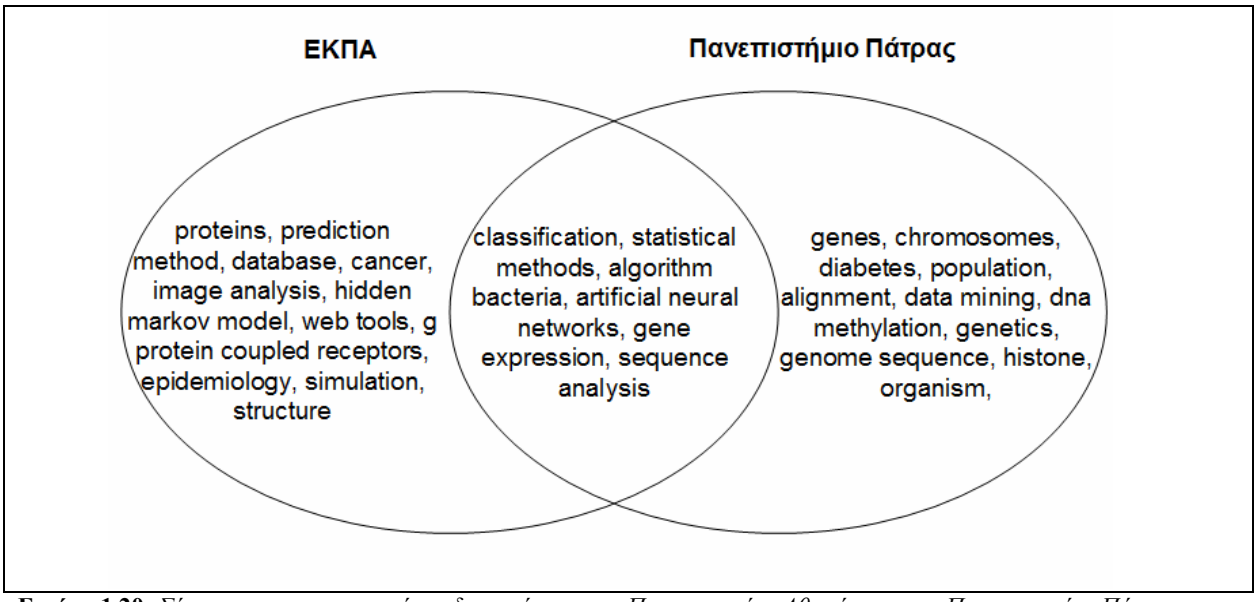
περισσότερο). Η τόσο μεγάλη επιρροή που φαίνεται ότι έχουν ο βιολογικές βάσεις δεδομένων, ειδικά μετά το 2005, είναι σε συμφωνία με τα όσα είδαμε στις προηγούμενες ενότητες σχετικά με τη διεθνή βιβλιογραφία.

Σε σχέση με τα ινστιτούτα από τα οποία προήλθαν οι εργασίες βιοπληροφορικής, στην πρώτη θέση τόσο όσον αφορά τον απόλυτο αριθμό των εργασιών, αλλά και των αναφορών και του Impact Factor, βρίσκεται το ΕΚΠΑ και ακολουθούν το Πανεπιστήμιο Πάτρας, το Πανεπιστήμιο Ιωαννίνων, το ΑΠΘ και το ΕΜΠ (Πίνακας 1.5). Τη δεκάδα συμπληρώνουν το Πανεπιστήμιο Κρήτης και το Πανεπιστήμιο Κύπρου, ενώ οι μόνες παρουσίες ερευνητικών κέντρων στην πρώτη δεκάδα είναι το ΙΤΕ (6η θέση), το ΕΚΕΦΕ Δημόκριτος (8η θέση) και το ΙΙΒΕΑΑ (9η θέση). Αυτή η κατάταξη αφορά διαχρονικά την εποχή στην οποία δημοσιεύτηκε η εργασία. Όταν η ίδια κατάταξη γίνει με βάση την τωρινή θέση που κατέχουν οι επιστήμονες (δηλαδή με βάση τις σημερινές θέσεις εργασίας των 18 επιστημόνων με τις περισσότερες εργασίες), η σειρά αλλάζει λίγο και είναι αυτή που φαίνεται στον Πίνακα 1.5. Παρατηρούμε, ότι η γενική εικόνα δεν έχει αλλάξει πολύ, για παράδειγμα το ΕΚΠΑ εξακολουθεί να είναι πρώτο, τα Πανεπιστήμια Ιωαννίνων και Πάτρας, αλλά και το ΕΜΠ είναι μέσα στην πρώτη πεντάδα κ.ο.κ. Παρ' όλα αυτά, βλέπουμε τη δυναμική εμφάνιση του Πανεπιστημίου Στερεάς Ελλάδας (3η θέση), το οποίο βέβαια καταργήθηκε το 2012 και συγχωνεύτηκε με το Πανεπιστήμιο Θεσσαλίας, αλλά και την είσοδο δύο νέων ιδρυμάτων στην πρώτη δεκάδα, του ΑΛ. ΦΛΕΜΙΝΓΚ (6η θέση) αλλά και του ΤΕΙ Αθήνας (9η θέση).

Ίδρυμα	**Εργασίες**	**Αναφορές**	**Impact Factor**
University of Athens	53	706	159,86
University of Ioannina	41	192	65,53
University of Central Greece	34	498	97,39
Natl Tech Univ Athens	28	139	54,21
University of Patras	21	66	48,17
BSRC Alexander Fleming	18	362	103,11
Aristotle Univ Thessaloniki	16	49	25,66
Natl Ctr Sci Res Demokritos	15	110	35,9
Tech Educ Inst Athens	14	46	28,73
Acad Athens|Biomed Res Fdn	12	13	35,33
Ctr Res & Technol Hellas CERTH	9	116	32,87
University of Thessaly	9	70	17,46
University of Cyprus	8	139	32,53
Tech Educ Inst Lamia	6	42	11,47
Democritus Univ Thrace	6	31	15,44

Πίνακας 1.5: *Τα κυριότερα ιδρύματα που εντοπίστηκαν στη μελέτη μας, με τις εργασίες, τις αναφορές και το συνολικό δείκτη επιρροής.*

Συμπερασματικά, και παρ' όλους τους περιορισμούς της μελέτης αυτής τους οποίους αναλύσαμε παραπάνω, μπορούμε να πούμε ότι η δραστηριότητα στον Ελληνικό χώρο την τελευταία 15ετία είναι ιδιαίτερα αυξημένη και έχει αρχίσει να σχηματίζεται η κρίσιμη μάζα επιστημόνων οι οποίοι θα προωθήσουν το πεδίο. Η διεπιστημονικότητα στην προέλευση αυτών των επιστημόνων είναι εμφανής, τόσο όταν αναλογιστούμε το υπόβαθρο των παλαιότερων ερευνητών στο χώρο, όσο και βλέποντας τα ιδρύματα στα οποία διεξάγεται η έρευνα αυτή. Τα παλαιότερα και μεγαλύτερα πανεπιστήμια, όπως ήταν αναμενόμενο, κυριαρχούν στο χώρο, αλλά τα νεότερα περιφερειακά πανεπιστήμια έχουν έντονη παρουσία τα τελευταία χρόνια.



Εικόνα 1.20: *Σύγκριση των ερευνητικών ενδιαφερόντων του Πανεπιστημίου Αθηνών και του Πανεπιστημίου Πάτρας.*

1.4.3. Η εκπαίδευση στη βιοπληροφορική στην Ελλάδα

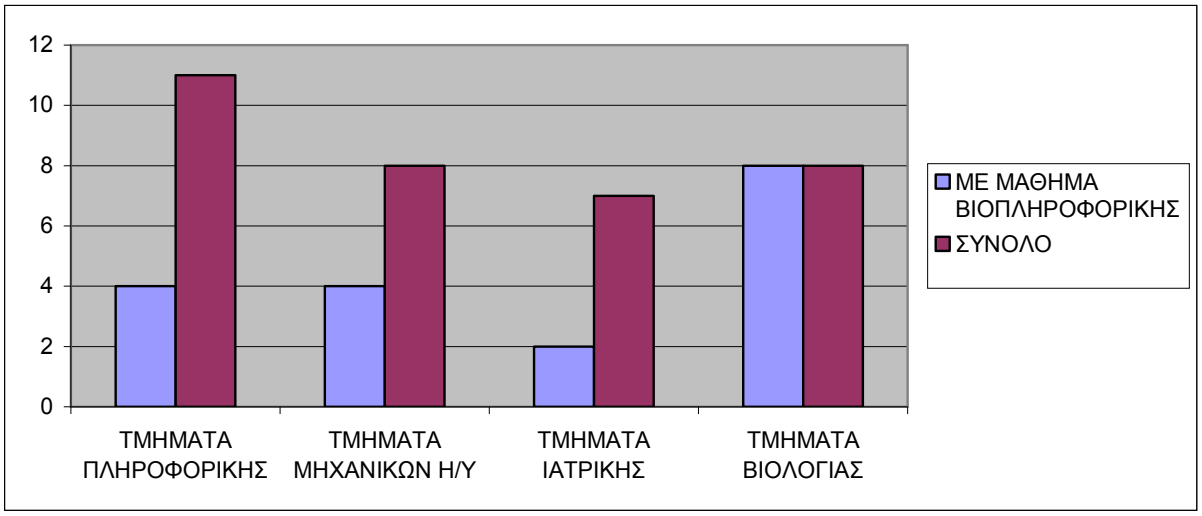
Τέλος, έχει ιδιαίτερη αξία να δούμε εκτός από την ερευνητική και την εκπαιδευτική δραστηριότητα στον Ελληνικό χώρο, τόσο σε προπτυχιακό όσο και σε μεταπτυχιακό επίπεδο (τουλάχιστον στο πλαίσιο κάποιου οργανωμένου προγράμματος σπουδών). Εκτεταμένη αναζήτηση στις ιστοσελίδες και στα προγράμματα σπουδών των Ελληνικών Τμημάτων Βιολογίας (και των υπόλοιπων συναφών τμημάτων βιολογικών επιστημών), Ιατρικής, Πληροφορικής, αλλά και Μηχανικών Η/Υ, δείχνει ότι βιοπληροφορική διδάσκεται σε προπτυχιακό επίπεδο σε 18 τμήματα σε όλη την Ελλάδα (Πίνακας 1.6). Οκτώ από αυτά είναι βιολογικά ή συναφή τμήματα, τέσσερα είναι τμήματα Μηχανικών Η/Υ, τέσσερα είναι τμήματα Πληροφορικής και δύο είναι Ιατρικές σχολές. Βλέπουμε λοιπόν ότι σε όλα σχεδόν τα βιολογικής κατεύθυνσης τμήματα της χώρας, υπάρχει σχετικό μάθημα βιοπληροφορικής στο πρόγραμμα σπουδών. Αντίθετα, το ίδιο συμβαίνει στα λιγότερο από τα μισά τμήματα Πληροφορικής, Μηχανικών Η/Υ αλλά και Ιατρικής. Στα 11 από τα 18 τμήματα, υπηρετεί μέλος ΔΕΠ με γνωστικό αντικείμενο Βιοπληροφορική ή συναφές (π.χ. Υπολογιστική Βιολογία), σε 3 από τα 18 τμήματα υπηρετεί μέλος ΔΕΠ το οποίο έχει τη βιοπληροφορική στα κύρια ερευνητικά του ενδιαφέροντα ενώ στα υπόλοιπα 3 τμήματα, δεν υπάρχει σχετικό μέλος ΔΕΠ (στο Τμήμα Βιολογίας του ΕΚΠΑ ο καθ. Σ. Χαμόδρακας αφυπηρέτησε πρόσφατα, ενώ στο Τμήμα Πληροφορικής του Πανεπιστημίου Πειραιά το μάθημα δεν προσφέρεται καν στους φοιτητές).

Ενδιαφέρουσα περίπτωση είναι το Πανεπιστήμιο Θεσσαλίας, στο οποίο τέσσερα διαφορετικά τμήματα, διαφορετικών σχολών, έχουν την βιοπληροφορική στο πρόγραμμα σπουδών τους, ενώ και στα τέσσερα υπηρετεί μέλος ΔΕΠ με αντίστοιχο γνωστικό αντικείμενο. Στο Τμήμα Πληροφορικής με εφαρμογές στη Βιοϊατρική, διδάσκονται μάλιστα 3 μαθήματα βιοπληροφορικής ενώ υπάρχουν στο πρόγραμμα σπουδών και άλλα συναφή μαθήματα (Βιοστατιστική, Βιολογία, Βιοχημεία, Γενετική, αλλά και αρκετά μαθήματα Ιατρικής Πληροφορικής). Επίσης, το Τμήμα Μοριακής Βιολογίας του Δημοκρίτειου Πανεπιστημίου Θράκης είναι μια ειδική περίπτωση καθώς εκεί διδάσκονται 4 εξαμηνιαία μαθήματα βιοπληροφορικής και υπολογιστικής βιολογίας, τα περισσότερα από κάθε άλλο τμήμα. Ιδιαίτερη έμφαση στη Βιοπληροφορική δίνεται και στο Πανεπιστήμιο Κρήτης, όπου το μάθημα διδάσκεται στα Τμήματα Βιολογίας, Ιατρικής, και Επιστήμης Υπολογιστών, ενώ σε όλες τις περιπτώσεις στα τμήματα αυτά υπηρετεί μέλος ΔΕΠ με συναφές γνωστικό αντικείμενο. Ειδικά στο Τμήμα Επιστήμης Υπολογιστών διδάσκονται δύο σχετικά μαθήματα ενώ στους φοιτητές δίνεται η δυνατότητα να επιλέξουν και μαθήματα βιολογίας από το αντίστοιχο τμήμα. Στο Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών διδάσκεται βιοπληροφορική σαν επιλεγόμενο μάθημα τόσο στο τμήμα Βιολογίας όσο και στο Τμήμα Πληροφορικής και Τηλεπικοινωνιών, ενώ στο Πανεπιστήμιο Πατρών, διδάσκεται στο Τμήμα Βιολογίας και στο Τμήμα Μηχανικών Η/Υ και Πληροφορικής (δεν υπάρχει τμήμα Πληροφορικής). Τέλος, εντύπωση προκαλεί αρχικά η απουσία του ΕΜΠ και των Τμημάτων Πληροφορικής και Μηχανικών του ΑΠΘ από τη σχετική λίστα αλλά τα ευρήματα αυτά γίνονται κατανοητά

αν αναλογιστούμε τη χαμηλή συνεισφορά των ιδρυμάτων αυτών στην έρευνα στη βιοπληροφορική, όπως είδαμε στην προηγούμενη παράγραφο. Στα ιδρύματα αυτά οι περισσότεροι ερευνητές που εμπλέκονται σε θέματα βιοϊατρικής πληροφορικής, ασχολούνται κυρίως με την Ιατρική Πληροφορική και την Πληροφορική Υγείας, τομείς που εφάπτονται μεν, αλλά δεν ταυτίζονται με τη βιοπληροφορική.

Πανεπιστήμιο	**Τμήμα**	**Μαθήματα**
Πανεπιστήμιο Θεσσαλίας	Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών (*)	Βιοπληροφορική
	Τμήμα ιατρικής (*)	Βιοπληροφορική-Βιομετρία
	Τμήμα Βιοχημείας και Βιοτεχνολογίας (*)	Βιοπληροφορική
	Τμήμα Πληροφορικής με εφαρμογές στη Βιοϊατρική (*)	Βιοπληροφορική Ι, Βιοπληροφορική ΙΙ, Ειδικά Θέματα Βιοπληροφορικής και Βιοηθική
Πανεπιστήμιο Κρήτης	Τμήμα Επιστήμης Υπολογιστών (*)	Αλγόριθμοι στη βιοπληροφορική, Εισαγωγή στον προγραμματισμό για Βιοπληροφορική
	Τμήμα Ιατρικής (*)	Εισαγωγή στη Βιοπληροφορική
	Τμήμα Βιολογίας (*)	Υπολογιστική Βιολογία
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών	Τμήμα Βιολογίας	Βιοπληροφορική
	Τμήμα Πληροφορικής και Τηλεπικοινωνιών (**)	Αλγόριθμοι Βιοπληροφορικής
Πανεπιστήμιο Πατρών	Τμήμα Βιολογίας (**)	Βιοπληροφορική
	Τμήμα μηχανικών Η/Υ και Πληροφορικής (**)	Εισαγωγή στη Βιοπληροφορική
Δημοκρίτειο Πανεπιστήμιο Θράκης	Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών	Βιοπληροφορική
	Τμήμα Μοριακής Βιολογίας και Γενετικής (*)	Εισαγωγή στην Υπολογιστική Βιολογία, Βιοπληροφορική, Ειδικά θέματα Βιοπληροφορικής, Ειδικά Θέματα Υπολογιστικής Βιολογίας
Πανεπιστήμιο Ιωαννίνων	Τμήμα Βιολογικών εφαρμογών και Τεχνολογιών (*)	Βιοπληροφορική, Ειδικά θέματα Βιοπληροφορικής
Πανεπιστήμιο Δυτικής Μακεδονίας	Τμήμα Μηχανικών Πληροφορικής και Τηλεπικοινωνιών (*)	Βιοπληροφορική
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης	Τμήμα Βιολογίας (*)	Βιοπληροφορική
Γεωπονικό Πανεπιστήμιο Αθηνών	Τμήμα Βιοτεχνολογίας (*)	Βιοπληροφορική
Πανεπιστήμιο Πειραιά	Τμήμα Πληροφορικής	Βιοπληροφορική

Πίνακας 1.6: *Τα πανεπιστημιακά τμήματα στα οποία διδάσκονται μαθήματα βιοπληροφορικής. Με (*) συμβολίζονται τα τμήματα στα οποία υπηρετεί μέλος ΔΕΠ με συναφές γνωστικό αντικείμενο.*



Εικόνα 1.21: *Τα τμήματα που έχουν μάθημα βιοπληροφορικής σε σχέση με τα υπόλοιπα συναφή τμήματα σε Ελληνικά Πανεπιστήμια.*

Όσον αφορά τη μεταπτυχιακή εκπαίδευση, αυτή τη στιγμή υπάρχουν στη χώρα 6 προγράμματα μεταπτυχιακών σπουδών (ΠΜΣ) που οδηγούν σε μεταπτυχιακό δίπλωμα ειδίκευσης (ΜΔΕ) στης βιοπληροφορική ή σε συναφείς ειδικότητες (δεν αναφέρονται τα προγράμματα βιοστατιστικής και ιατρικής πληροφορικής). Τα προγράμματα αυτά δίνονται στον Πίνακα 1.7.

Πανεπιστήμιο	**Τμήμα**	**ΜΔΕ**
Πανεπιστήμιο Θεσσαλίας	Τμήμα Πληροφορικής με εφαρμογές στη Βιοϊατρική και Τμήμα Πληροφορικής	Πληροφορική και Υπολογιστική Βιοϊατρική (με ροή «Υπολογιστική Ιατρική και Βιολογία»)
	Τμήμα ιατρικής	Μεθοδολογία Βιοϊατρικής Έρευνας, Βιοστατιστική και Κλινική Βιοπληροφορική
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών	Τμήμα Βιολογίας	Βιοπληροφορική
	Τμήμα Πληροφορικής και Τηλεπικοινωνιών, ΤΕΙ Αθήνας	Τεχνολογίες Πληροφορικής στην Ιατρική και τη Βιολογία (με ροή «Βιοπληροφορική»)
Πανεπιστήμιο Πατρών	Τμήματα Ιατρικής, Βιολογίας, Φυσικής, Φαρμακευτικής και Μηχανικών Η/Υ και Πληροφορικής	Πληροφορική Επιστημών Ζωής (με ροή «Βιοπληροφορική»)
Γεωπονικό Πανεπιστήμιο Αθηνών	Βιοτεχνολογίας	Βιολογία Συστημάτων

Πίνακας 1.7: *Τα μεταπτυχιακά συναφή με τη βιοπληροφορική στα Ελληνικά Πανεπιστήμια.*

Από τα ΠΜΣ αυτά, τα παλιότερα είναι το ΠΜΣ «Βιοπληροφορικής» του Τμήματος Βιολογίας του ΕΚΠΑ, και το διατμηματικό ΠΜΣ «Πληροφορική Επιστημών Ζωής» (με ροή «Βιοπληροφορική») το οποίο συνδιοργανώνεται από τα Ιατρικής, Βιολογίας, Φυσικής, Φαρμακευτικής και Μηχανικών Η/Υ και Πληροφορικής του Πανεπιστημίου Πατρών. Τα δύο αυτά μεταπτυχιακά ιδρύθηκαν το 2003, ενώ λίγα χρόνια αργότερα ιδρύθηκε το ΠΜΣ «Τεχνολογίες Πληροφορικής στην Ιατρική και τη Βιολογία» (με ροή «Βιοπληροφορική») από το Τμήμα Πληροφορικής και Τηλεπικοινωνιών του ΕΚΠΑ σε συνεργασία με το ΤΕΙ Αθήνας. Βλέπουμε, ότι η έντονη ερευνητική δραστηριότητα των δύο αυτών ιδρυμάτων οδήγησε και στη δημιουργία των πρώτων ΠΜΣ στην Ελλάδα. Παρ' όλα αυτά, στην περίπτωση της Πάτρας είδαμε μια διατμηματική συνεργασία, ενώ στην περίπτωση του ΕΚΠΑ κάθε τμήμα οργάνωσε το δικό του πρόγραμμα σπουδών. Ιδιαίτερη περίπτωση είναι και πάλι το Πανεπιστήμιο Θεσσαλίας, στο οποίο λειτουργούν εδώ και λίγο καιρό δύο διαφορετικά μεταπτυχιακά προγράμματα σπουδών, το «Μεθοδολογία Βιοϊατρικής Έρευνας, Βιοστατιστική και Κλινική Βιοπληροφορική» από το Τμήμα Ιατρικής στη Λάρισα, και το «Πληροφορική και Υπολογιστική Βιοϊατρική» (με ροή «Υπολογιστική Ιατρική και Βιολογία») από τα Τμήματα Πληροφορικής

με εφαρμογές στη Βιοϊατρική και Πληροφορικής, στη Λαμία. Εδώ, η γεωγραφική απομόνωση και το γεγονός ότι τα τμήματα βρίσκονται σε διαφορετικές πόλεις οδήγησε στη δημιουργία διαφορετικών προγραμμάτων σπουδών. Παρ' όλα αυτά, είναι εμφανές και εδώ ότι η έντονη ερευνητική παρουσία του Πανεπιστημίου Θεσσαλίας αλλά και η ύπαρξη μελών ΔΕΠ με συναφές με τη βιοπληροφορική γνωστικό αντικείμενο έχει παίξει καταλυτικό ρόλο. Τελευταία προσθήκη είναι το ΠΜΣ «Βιολογία Συστημάτων», στο Τμήμα Βιοτεχνολογίας του Γεωπονικού Πανεπιστημίου Αθηνών, μεταπτυχιακό πρόγραμμα που εντάσσεται στα ερευνητικά ενδιαφέροντα του τμήματος και είναι συμβατό με την προπτυχιακή εκπαίδευση στο ίδρυμα αυτό. Εντύπωση προκαλεί η απουσία ΠΜΣ από το Πανεπιστήμιο Ιωαννίνων αλλά ακόμα περισσότερο από το Πανεπιστήμιο Κρήτης, τα οποία όπως είδαμε διαθέτουν και προσωπικό σε συναφή γνωστικά αντικείμενα αλλά και έχουν να επιδείξουν ερευνητική δραστηριότητα στον τομέα. Πιθανότατα, οι ανάγκες μεταπτυχιακής εκπαίδευσης στα ιδρύματα αυτά καλύπτονται από άλλα συναφή ή πιο γενικά προγράμματα σπουδών και οι φοιτητές που επιθυμούν να ασχοληθούν με βιοπληροφορική, βρίσκουν διέξοδο σε επίπεδο εκπόνησης διδακτορικής διατριβής.

Βιβλιογραφία

Altman, R. B. (1998). A curriculum for bioinformatics: the time is ripe. *Bioinformatics, 14*(7), 549-550.

Bagos, P. G. (2010). *Bioinformatics and Computational Biology in Greece: a bibliometric study*. Paper presented at the 5th Conference of HSCBB (HSCBB10), Alexandroupolis.

Chalmers, A. (1999). *What Is This Thing Called Science?* (3rd revised edition ed.). Hackett: University of Queensland Press, Open University press.

Ditty, J. L., Kvaal, C. A., Goodner, B., Freyermuth, S. K., Bailey, C., Britton, R. A., . . . Kerfeld, C. A. (2010). Incorporating genomics and bioinformatics across the life sciences curriculum. *PLoS Biol, 8*(8), e1000448.

Eddy, S. R. (2005). "Antedisciplinary" science. *PLoS Comput Biol, 1*(1), e6.

Floriano, W. B. (2008). A portable bioinformatics course for upper-division undergraduate curriculum in sciences. *Biochem Mol Biol Educ, 36*(5), 325-335.

Hagen, J. B. (2000). The origins of bioinformatics. *Nat Rev Genet, 1*(3), 231-236.

Honts, J. E. (2003). Evolving strategies for the incorporation of bioinformatics within the undergraduate cell biology curriculum. *Cell Biol Educ, 2*(4), 233-247.

King, D. A. (2004). The scientific impact of nations. *Nature, 430*(6997), 311-316.

Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med, 40*(4), 346-358.

Molenberghs, G. (2005). Biometry, biometrics, biostatistics, bioinformatics,..., bio-X. *Biometrics, 61*(1), 1-9.

Ouzounis, C. A. (2000). Two or three myths about bioinformatics. *Bioinformatics, 16*(3), 187-189.

Ouzounis, C. A. (2002). Bioinformatics and the theoretical foundations of molecular biology. *Bioinformatics, 18*(3), 377-378.

Ouzounis, C. A. (2012). Rise and demise of bioinformatics? Promise and progress. *PLoS Comput Biol, 8*(4), e1002487.

Ouzounis, C. A., & Valencia, A. (2003). Early bioinformatics: the birth of a discipline--a personal view. *Bioinformatics, 19*(17), 2176-2190.

Patra, S. K., & Mishra, S. (2006). Bibliometric study of bioinformatics literature. *Scientometrics, 67*(3), 477-489.

Perez-Iratxeta, C., Andrade-Navarro, M. A., & Wren, J. D. (2007). Evolving research trends in bioinformatics. *Brief Bioinform, 8*(2), 88-95.

Rebholz-Schuhman, D., Cameron, G., Clark, D., van Mulligen, E., Coatrieux, J. L., Del Hoyo Barbolla, E., . . . Van der Lei, J. (2007). SYMBIOmatics: synergies in Medical Informatics and Bioinformatics--exploring current scientific literature for emerging topics. *BMC Bioinformatics, 8 Suppl 1*, S18.

Roberts, R. J. (2000). The early days of bioinformatics publishing. *Bioinformatics, 16*(1), 2-4.

Searls, D. B. (2010). The roots of bioinformatics. *PLoS Comput Biol, 6*(6), e1000809.

Searls, D. B. (2012). An online bioinformatics curriculum. *PLoS Comput Biol, 8*(9), e1002632.

Song, M., Kim, S., Zhang, G., Ding, Y., & Chambers, T. (2014). Productivity and influence in bioinformatics: A bibliometric analysis using PubMed central. *Journal of the Association for Information Science and Technology, 65*(2), 352-371.

Trifonov, E. N. (2000). Earliest pages of bioinformatics. *Bioinformatics, 16*(1), 5-9.

Welch, L., Lewitter, F., Schwartz, R., Brooksbank, C., Radivojac, P., Gaeta, B., & Schneider, M. V. (2014). Bioinformatics curriculum guidelines: toward a definition of core competencies. *PLoS Comput Biol, 10*(3), e1003496.

Wingender, E. (1998). ISB: Just Another Journal? *In Silico Biol, 1*(1), 1-4.

Yan, B., Ban, K. H., & Tan, T. W. (2014). Integrating translational bioinformatics into the medical curriculum. *Int J Med Educ, 5*, 132-134.

Μαυρικάκη, Ε., Γκούβρα, Μ., & Καμπούρη, Α. (2014). *Βιολογία Γ Γυμνασίου*. Αθήνα: ΟΕΔΒ

Μπάγκος, Π., & Περράκης, Α. (2014, 25/11/2014). Το Ελληνικό παράδοξο στην επιστημονική έρευνα. *Το ΒΗΜΑ*.

Σαχίνη, Ε., Μάλλιου, Ν., & Χούσος, Ν. (2012). Ελληνικές Επιστημονικές Δημοσιεύσεις 1996-2010: Βιβλιομετρική Ανάλυση Ελληνικών Δημοσιεύσεων σε Διεθνή Επιστημονικά Περιοδικά Retrieved from http://reports.metrics.ekt.gr/

Σαχίνη, Ε., Μάλλιου, Ν., Χούσος, Ν., & Καραϊσκος, Δ. (2012). Ελληνικές Επιστημονικές Δημοσιεύσεις 2000-2010 - Τομέας Βιοεπιστημών Retrieved from http://metrics.ekt.gr/el/node/15

Σαχίνη, Ε., Μάλλιου, Ν., Χούσος, Ν., & Καραϊσκος, Δ. (2013). Ελληνικές Επιστημονικές Δημοσιεύσεις 1996-2010: Βιβλιομετρική Ανάλυση Ελληνικών Δημοσιεύσεων σε Διεθνή Επιστημονικά Περιοδικά - Scopus Retrieved from http://report03.metrics.ekt.gr

Σαχίνη, Ε., Μάλλιου, Ν., Χούσος, Ν., & Καραΐσκος, Δ. (2014). Ελληνικές Επιστημονικές Δημοσιεύσεις 1998-2012: Βιβλιομετρική Ανάλυση Ελληνικών Δημοσιεύσεων σε Διεθνή Επιστημονικά Περιοδικά - Web of Science Retrieved from http://report04.metrics.ekt.gr/

Κεφάλαιο 2: Βιολογικές Βάσεις Δεδομένων

Σύνοψη

Στο κεφάλαιο αυτό, θα γίνει η απαραίτητη εισαγωγή στις βιολογικές βάσεις δεδομένων έτσι ώστε ο αναγνώστης να μπορεί, στα επόμενα κεφάλαια, να ανατρέχει στις πηγές που χρησιμοποιούνται για την ανάλυση των αντίστοιχων κάθε φορά δεδομένων (αλληλουχίες, δομές, οικογένειες πρωτεϊνών, δεδομένα έκφρασης, πολυμορφισμοί κ.ο.κ.). Ανάλογα με το είδος της πληροφορίας που περιέχουν, θα παρουσιαστούν οι κύριες βάσεις κάθε κατηγορίας και θα τονιστούν τα βασικά χαρακτηριστικά τους. Ειδικό κομμάτι στο τέλος του κεφαλαίου, θα αφιερωθεί στις εξειδικευμένες βάσεις (κυριώς πρωτεϊνικών) δεδομένων, οι οποίες καταλαμβάνουν σημαντικό μερίδιο στην έρευνα των μικρών και μεσαίου μεγέθους ερευνητικών εργαστηρίων και αποτελούν σημαντικό εργαλείο στη βιοπληροφορική μελέτη των πρωτεϊνών.

Προαπαιτούμενη γνώση

Προαπαιτούμενο για το κεφάλαιο αυτό, είναι η εξοικείωση με τις βασικές γνώσεις και έννοιες της μοριακής βιολογίας (DNA, RNA, πρωτεϊνες κλπ).

2. Εισαγωγή

Οι βιολογικές βάσεις δεδομένων, αποτελούν βασικό κομμάτι της σύγχρονης βιοπληροφορικής, καθώς αποτελούν τη βασική πηγή δεδομένων από την οποία ένας ερευνητής αντλεί τα δεδομένα στα οποία θα βασίσει την ανάλυση του. Ακόμα και για αυτούς οι οποίοι παράγουν οι ίδιοι πρωτογενή δεδομένα, η ύπαρξη τους είναι σημαντική καθώς τις περισσότερες φορές είναι αναγκασμένοι να καταθέτουν τα δεδομένα τους σε αυτές, έτσι ώστε να είναι διαθέσιμα στην επιστημονική κοινότητα. Όταν δημιουργήθηκαν οι πρώτες βάσεις δεδομένων ο όγκος της πληροφορίας ήταν μικρός, με αποτέλεσμα η συντήρηση και η ανανέωση των βάσεων να απαιτούν μικρό κόστος τόσο σε υποδομές όσο και σε ανθρώπινο δυναμικό. Η πρόσβαση στις εγγραφές γινόταν μέσω επικοινωνίας με τους επιστημονικούς υπευθύνους της βάσης, οι οποίοι συνήθως έστελναν στον ενδιαφερόμενο όλη την βάση αποθηκευμένη σε δισκέτες ή μαγνητοταινία, με συμβατικό ταχυδρομείο.

Η τεχνολογική εξέλιξη όμως οδήγησε στην αύξηση του όγκου των πειραματικών εργασιών και της διεκπεραίωσής τους, που σε συνδυασμό με τον διαρκή προσδιορισμό γονιδιωμάτων διαφόρων οργανισμών, αύξησε σημαντικά τον όγκο της πληροφορίας σε όλα τα επίπεδα και ιδιαίτερα στο επίπεδο της αλληλουχίας. Στις μέρες μας οι βάσεις περιέχουν πολύ μεγάλο όγκο δεδομένων ενώ είναι απαραίτητο να ανανεώνονται καθημερινά. Η συντήρηση μιας βάσης απαιτεί μεγάλο αριθμό εξειδικευμένων επιστημόνων που θα ασχολούνται αποκλειστικά με την επισήμανση ενδεχόμενων λαθών καθώς και με το σχολιασμό (annotation) των νεοεισερχόμενων δεδομένων.

Δυο χαρακτηριστικά παραδείγματα βάσεων αποτελούν η UiprotKB/SWISS-PROT, η κύρια βάση πρωτεϊνικών αλληλουχιών που περιέχει 547.599 αλληλουχίες (Rel. 2015_02 – Φεβρουάριος 2015) και η EMBL Nucleotide Sequence Database που περιέχει νουκλεοτιδικές αλληλουχίες και έχει 510.014.239 εγγραφές (Rel. 122 - Νοέμβριος 2014). Κάθε ερευνητής μπορεί να έχει πρόσβαση στις βάσεις αυτές μέσω της χρήσης διαδικτύου. Αρκεί η επίσκεψη στην ιστοσελίδα της βάσης, η αναζήτηση των δεδομένων ενδιαφέροντος και στη συνέχεια η αποθήκευσή τους στον υπολογιστή. Παράλληλα έχουν δημιουργηθεί βάσεις στις οποίες η πληροφορία στο επίπεδο της αλληλουχίας και της δομής είναι ταξινομημένη με τέτοιο τρόπο ώστε η πληροφορία να είναι οργανωμένη για την εξαγωγή συμπερασμάτων ως προς την βιολογική τους σημασία.

Οι βιολογικές βάσεις δεδομένων, γενικά, μπορούν να διακριθούν σε 2 μεγάλες κατηγορίες, με επιμέρους κατηγοριοποιήσεις, όπως περιγράφονται παρακάτω.

Καταρχάς, υπάρχουν οι πρωτογενείς βάσεις δεδομένων, οι οποίες περιέχουν τα πρωτογενή πειραματικά δεδομένα και οι οποίες αναλύονται κυρίως σε:

- Βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών
- Βάσεις δεδομένων αμινοξικών αλληλουχιών πρωτεϊνών
- Βάσεις δεδομένων τρισδιάστατων βιολογικών δομών
- Βάσεις δεδομένων γονιδιακής έκφρασης
- Βάσεις δεδομένων γενετικής ποικιλομορφίας
- Βάσεις δεδομένων βιβλιογραφίας

Έπειτα, υπάρχουν οι δευτερογενείς βάσεις δεδομένων, στις οποίες υπάρχουν κυρίως ταξινομήσεις των πρωτογενών δεδομένων, χρήσιμες για αναλυτικούς σκοπούς, οι οποίες διακρίνονται σε:

- Βάσεις δεδομένων οικογενειών (κυρίως πρωτεϊνών)
- Εξειδικευμένες βάσεις δεδομένων (όλες οι άλλες κατηγορίες)

2.1. Πρωτογενείς βάσεις δεδομένων

Οι πρωτογενείς βάσεις δεδομένων, είναι οι βάσεις που περιέχουν τα βιολογικά δεδομένα όπως αυτά προσδιορίζονται πειραματικά, και συνήθως περιέχουν επιπλέον ταξινόμηση και σχολιασμό. Γενικά, θα μπορούσαμε να τοποθετήσουμε σε αυτή την κατηγορία τις γενικές βάσεις δεδομένων αλληλουχιών, δομών, δεδομένων έκφρασης, γενετικής ποικιλομορφίας αλλά και για λόγους που θα γίνουν κατανοητοί αργότερα, και τις βάσεις δεδομένων βιβλιογραφίας.

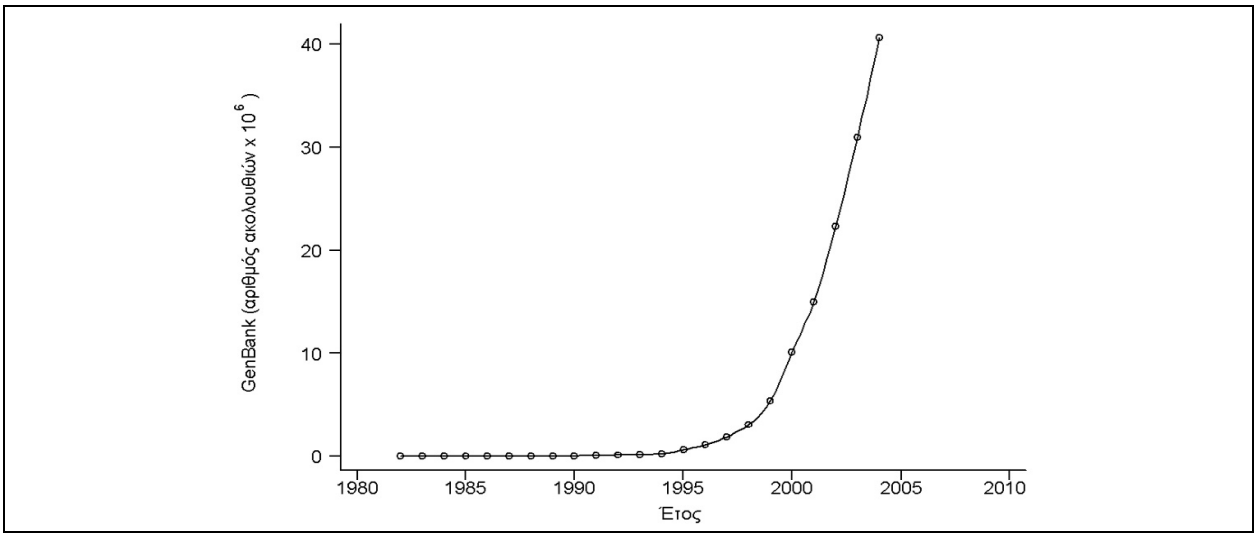
2.1.1 Βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών

Ο όγκος της πληροφορίας που περιέχεται στις βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών καθώς και ο εκθετικός ρυθμός συσσώρευσης των δεδομένων που εμφανίζουν (Εικόνα 2.1), τις έχουν καταστήσει ως τις μεγαλύτερες βάσεις της Βιολογίας. Η εξέλιξη της τεχνολογίας στην εύρεση της αλληλουχίας (sequencing), κυρίως του DNA αλλά και δευτερευόντως του RNA, οδήγησε στον προσδιορισμό της αλληλουχίας ολόκληρων γονιδιωμάτων αρκετών οργανισμών (π.χ. ο άνθρωπος) και στη δημιουργία εξειδικευμένων βάσεων δεδομένων που περιέχουν τις αλληλουχίες για έναν και μόνο από αυτούς.

Οι τρεις μεγαλύτερες βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών που είναι ελεύθερα διαθέσιμες στην ακαδημαϊκή κοινότητα είναι οι GENBANK (NCBI), DNA Data Bank of Japan (DDBJ) και EMBL Nucleotide Sequence Database (EBI). Οι τρεις αυτές βάσεις, βρίσκονται σε συνεργασία, δηλαδή ανταλλάσσουν σε καθημερινή βάση τις εγγραφές που κατατίθενται ανεξάρτητα σε καθεμία, έχοντας θέσει παράλληλα κοινούς κανόνες ταξινόμησης και σχολιασμού δεδομένων. Από αυτήν την συνεργασία έχει δημιουργηθεί η International Nucleotide Sequence Database Collaboration. Παρακάτω παρουσιάζονται τα βασικά χαρακτηριστικά των βάσεων δεδομένων που συμμετέχουν στην International Nucleotide Sequence Database Collaboration :

GENBANK: Η GENBANK (http://www.ncbi.nlm.nih.gov/Genbank/index.html) είναι μια βάση νουκλεοτιδικών αλληλουχιών (Benson et al., 2014), διατίθεται ελεύθερα στην επιστημονική κοινότητα και βρίσκεται και υπό την αιγίδα του Εθνικού Ινστιτούτου Υγείας των Η.Π.Α. (National Institutes of Health). Τα δεδομένα της βάσης προέρχονται από υποβολές δεδομένων διαφόρων ερευνητικών ομάδων όπως αυτά προκύπτουν από πειραματικές διεργασίες. Η διαδικασία υποβολής γίνεται με την συμπλήρωση κατάλληλης φόρμας μέσω διαδικτύου. Τα δεδομένα που υποβάλλονται στην βάση επεξεργάζονται, σχολιάζονται (annotate) από τους υπεύθυνους της βάσης και στη συνέχεια δημοσιοποιούνται σε αυτήν. Σε συχνά χρονικά διαστήματα τα δεδομένα που έχουν καταχωρηθεί στη βάση επανεξετάζονται και διορθώνονται σε περίπτωση που έχουν προκύψει νέα δεδομένα. Ο αριθμός των νουκλεοτιδικών βάσεων που περιέχονται στην GENBANK διπλασιάζεται κάθε 14 μήνες με αποτέλεσμα η τελευταία έκδοση (Rel. 206, Φεβρουάριος 2015) να περιέχει 181.336.445 αλληλουχίες και 187.893.826.750 συνολικό αριθμό βάσεων.

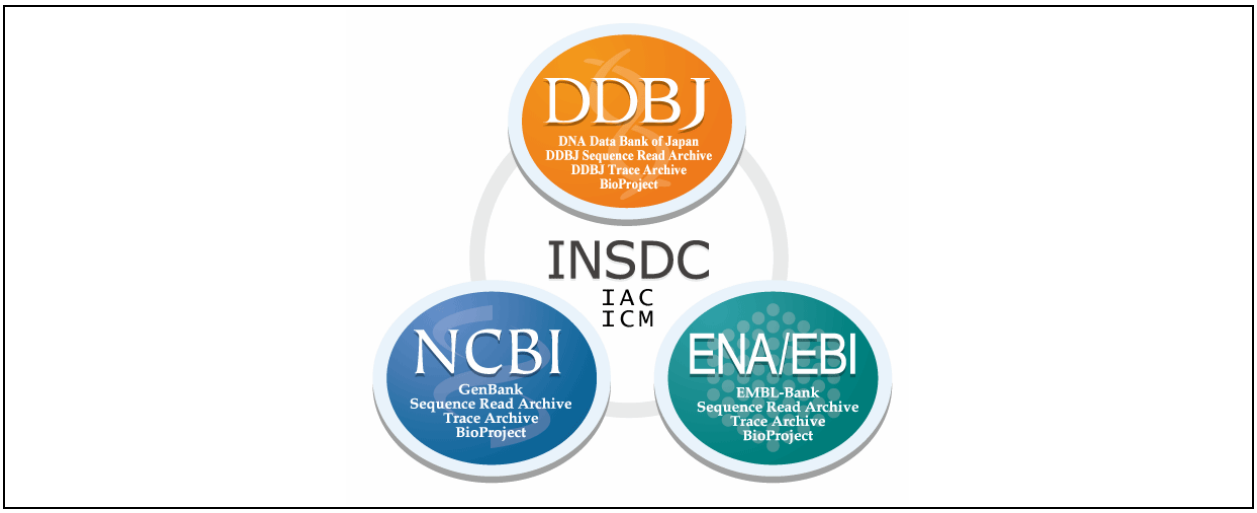
EMBL-Bank: Η EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/) αποτελεί τη μεγαλύτερη βάση νουκλεοτιδικών αλληλουχιών στην Ευρώπη, βρίσκεται υπό την αιγίδα του Ευρωπαϊκού Εργαστηρίου Μοριακής Βιολογίας (EMBL) ενώ εδράζεται και συντηρείται από το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (EBI) στο Cambridge, UK. Οι αλληλουχίες κατατίθενται στην EMBL-Bank μέσω διαδικτύου, ακολουθώντας μία απλή διαδικασία από ανεξάρτητα ερευνητικά εργαστήρια ή ομάδες που ασχολούνται με τον προσδιορισμό των γονιδιωμάτων διαφόρων οργανισμών. Αντίστοιχα με την GENBANK, οι νέες καταχωρήσεις αλληλουχιών επεξεργάζονται, σχολιάζονται από τους υπεύθυνους της βάσης και δημοσιοποιούνται. Παράλληλα διατίθενται διάφορα εργαλεία ανάλυσης των δεδομένων όπως το Fasta και το BLAST. Η παρούσα έκδοση της EMBL-Bank (Rel. 122 - Νοέμβριος 2014) περιέχει 510.014.239 εγγραφές. Ο συνολικός αριθμός νουκλεοτιδίων φτάνει τα 1.094.969.877.589.



Εικόνα 2.1: *Η εκθετική αύξηση των αλληλουχιών οι οποίες είναι κατατεθειμένες στην GenBank, από το 1982 έως το τέλος του 2004.*

DDBJ: Η DNA Databank of Japan (DDBJ - http://www.ddbj.nig.ac.jp/) είναι η μοναδική διεθνώς αναγνωρισμένη βάση νουκλεοτιδικών αλληλουχιών στην Ιαπωνία. Ιδρύθηκε το 1986 στο Εθνικό Ινστιτούτο Γενετικής (NIG) και βρίσκεται υπό την αιγίδα του Υπουργείου Παιδείας, Επιστημών και Αθλητισμού της Ιαπωνίας. Βασική πηγή δεδομένων της βάσης αποτελούν οι εργασίες των Ιαπώνων ερευνητών. Επιπλέον στην DDJB είναι διαθέσιμα διάφορα εργαλεία ανάλυσης νουκλεοτιδικών αλληλουχιών. Η παρούσα έκδοση της DDJB (Rel. 99, Δεκέμβριος 2014) περιέχει 178.825.615 εγγραφές και συνολικά 184.410.381.191 νουκλεοτιδικές βάσεις που περιέχονται στις αλληλουχίες.

Οι κυριότερες βάσεις δεδομένων με αλληλουχίες DNA στον διεθνή χώρο, η Genbank στις ΗΠΑ, η DDBJ στην Ιαπωνία και η EMBL Data Bank στην Ευρώπη, συνεργάζονται μέσω του International Nucleotide Sequence Collaboration, μιας οργάνωσης που οι ίδιοι δημιούργησαν, και έτσι μια αλληλουχία αφού καταχωρηθεί σε μια από αυτές μέσα από μια διαδικασία έγκρισης, καταχωρείται και στις άλλες (Εικόνα 2.2). Πρακτική συνέπεια αυτού του γεγονότος, είναι ότι εκτός ελαχίστων εξαιρέσεων, οι 3 βάσεις περιέχουν τις ίδιες καταχωρήσεις, άρα δεν έχει και πολύ μεγάλη σημασία σε ποια από τις 3 βάσεις δεδομένων θα απευθυνθούμε για μια έρευνα.

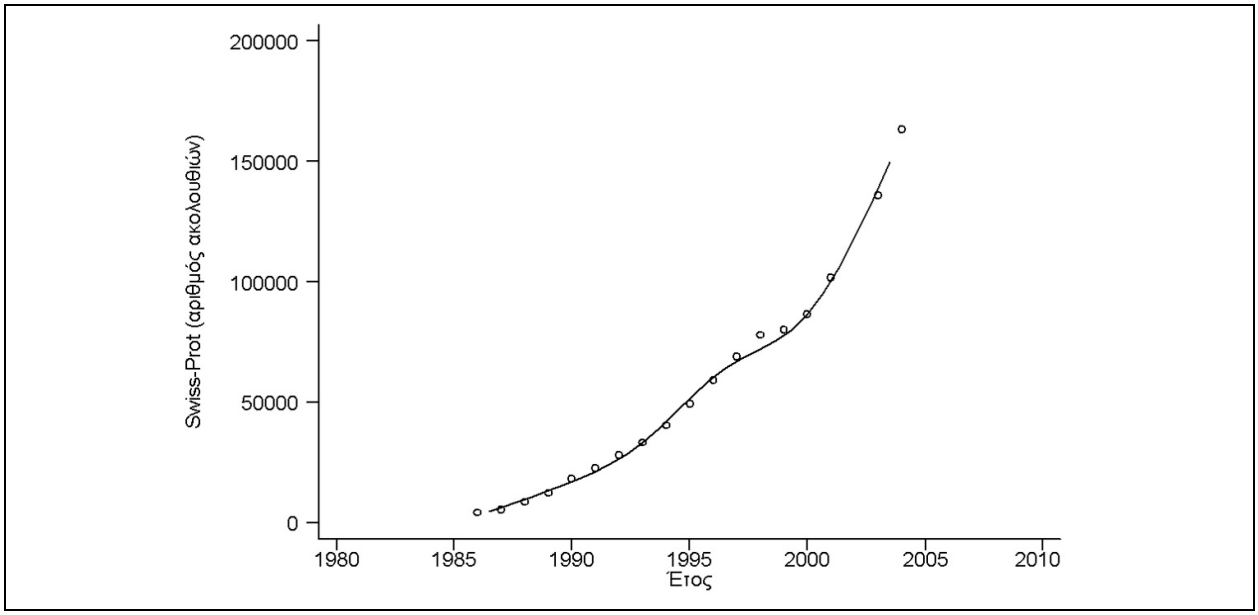


Εικόνα 2.2: *Διάγραμμα που απεικονίζει τη συνεργασία και τη ροή δεδομένων των 3 μεγάλων βάσεων νουκλεοτιδικών αλληλουχιών (INSDC; International Nucleotide Sequence Database Collaboration, http://www.ddbj.nig.ac.jp/insdc/insdc-e.html).*

2.1.2 Βάσεις δεδομένων πρωτεϊνικών αλληλουχιών.

Οι βάσεις δεδομένων πρωτεϊνικών αλληλουχιών, αποτελούν το δεύτερο μεγαλύτερο σε όγκο τμήμα του συνόλου των βιολογικών βάσεων δεδομένων (μετά τις αλληλουχίες DNA), αλλά αποτελούν ίσως το σημαντικότερο τμήμα, καθώς οι αμινοξικές αλληλουχίες πρωτεϊνών παρουσιάζουν μεγάλη ποικιλομορφία τόσο στη δομή όσο και στη λειτουργία τους. Κατά συνέπεια, μεγάλο μέρος της σύγχρονης βιοπληροφορικής ανάλυσης, αναφέρεται στις πρωτεΐνες και υπάρχει τεράστιος όγκος λειτουργικών δεδομένων που παράγονται συνεχώς πειραματικά, τα οποία αποτελούν ή θα έπρεπε να αποτελούν μέρος της πληροφορίας που περιέχεται σε αυτές τις βάσεις.

Η **UniprotKB** (Uniprot Knowledgebase, http://www.uniprot.org/), αποτελεί την κύρια, σε παγκόσμιο επίπεδο βάση δεδομένων πρωτεϊνικών αλληλουχιών (UniProt, 2014). Αποτελείται από δύο υποσύνολα, την Uniprot/SwissProt η οποία περιέχει τις καλά σχολιασμένες πρωτεϊνικές αλληλουχίες και την Uniprot/TrEMBL η οποία περιέχει τις πρωτεϊνικές αλληλουχίες που έχουν προκύψει από αυτόματη (ηλεκτρονική) μετάφραση γονιδιωματικών αλληλουχιών. Η UniprotKB/SwissProt περιέχει 547.599 αλληλουχίες (Rel. 2015_02 – Φεβρουάριος 2015) οι οποίες έχουν περάσει από κάποιου είδους έλεγχο και συνοδεύονται από συμπληρωματικά σχόλια όπως βιβλιογραφικές αναφορές, γενικά στοιχεία δευτεροταγούς δομής, σύνδεσμους σε άλλες βάσεις δεδομένων σχετικές με κάθε εγγραφή, καθώς και σημειώσεις για τη βιολογική λειτουργία (αν είναι γνωστές), καθώς και άλλες χρήσιμες πληροφορίες. Η Uniprot/TrEMBL περιέχει σήμερα (Rel. 2015_02 – Φεβρουάριος 2015) 92.124.243 αλληλουχίες οι οποίες όμως δεν έχουν υποστεί ανθρώπινο σχολιασμό. Περιοδικά, οι σχολιαστές της UniprotKB εντοπίζοντας δεδομένα από τη βιβλιογραφία αλλά και με χρήση αυτοματοποιημένων εργαλείων, αλλάζουν το σχολιασμό των καταχωρήσεων και έτσι μια πρωτεϊνική αλληλουχία ενδέχεται να "περάσει" από την Uniprot/TrEMBL στην Uniprot/SwissProt. Το είδος, το εύρος και η μεγάλη ποικιλομορφία του σχολιασμού που μπορεί να υπάρχει σε επίπεδο πρωτεϊνικής αλληλουχίας είναι τεράστιο (σε ποιο κυτταρικό οργανίδιο υπάρχει, σε ποιον ιστό εκφράζεται, ποια είναι η δευτεροταγής δομή της, ποιος ο βιολογικός της ρόλος, ποια τα μονοπάτια στα οποία εμπλέκεται κ.ο.κ.), και κατά συνέπεια, ο όγκος της πληροφορίας στην Uniprot/SwissProt είναι τεράστιος, όπως επίσης και η πιθανότητα (παρόλες τις προσπάθειες), η πληροφορία αυτή να είναι λαθεμένη ή απλά ελλειπής. Περισσότερα, αναλύονται στην ειδική ενότητα παρακάτω που αφορά στις εξειδικευμένες βάσεις δεδομένων και στα προβλήματα σχολιασμού τους. Ένα τυπικό αρχείο Uniprot με τις επεξηγήσεις των πιο σημαντικών πεδίων, παρουσιάζεται στο παράρτημα.



Εικόνα 2.3: *Η εκθετική αύξηση των αμινοξικών αλληλουχιών πρωτεϊνών οι οποίες είναι κατατεθειμένες στην Swiss-Prot, από το 1986 έως το τέλος του 2004.*

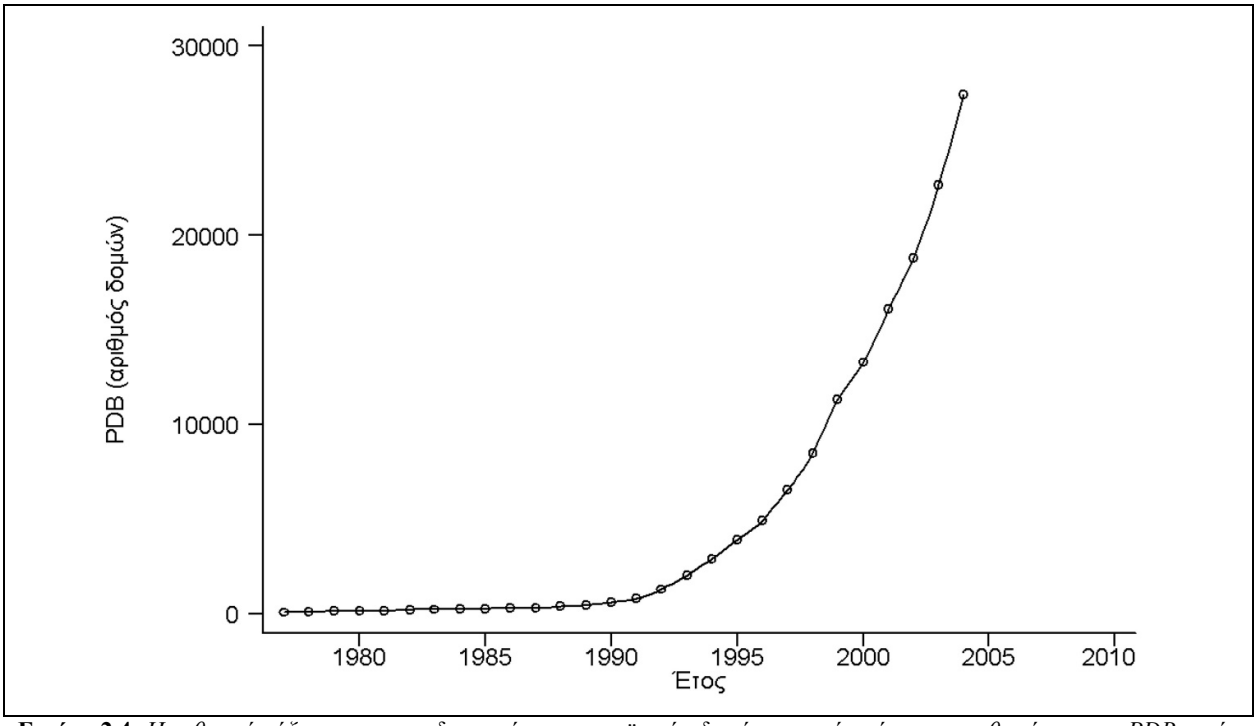
Ιστορικά, αξίζει να αναφερθεί ότι η Uniprot προέκυψε το 2002 από μια συνένωση των δύο μεγαλύτερων τότε βάσεων δεδομένων, της SwissProt και της PIR. Η SwissProt Ιδρύθηκε το 1986 στο Ελβετικό ινστιτούτο Βιοπληροφορικής (Swiss Institute of Bioinformatics) και λειτουργούσε σε συνεργασία με το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (European Bioinformatics Institute). Η **Protein Information Resource** (PIR - http://pir.georgetown.edu/) ήταν η αντίστοιχη Αμερικάνικη βάση δεδομένων. Η έδρα της ήταν στο Πανεπιστήμιο του Georgetown και αποτελούσε τμήμα του Εθνικού Ιδρύματος Βιοϊατρικής Έρευνας (NBRF) των Η.Π.Α. Η κυριότερη βάση που περιέχει είναι η PIR-International Protein Sequence Database (PSD), της οποίας τα δεδομένα προκύπτουν από την συνεργασία της PIR με το Munich Information Center for Protein Sequences (MIPS) και την Japanese International Protein Information Database (JIPID). Το 2002, η PIR σε μια κοινή προσπάθεια με το EBI (European Bioinformatics Institute) και το SIB (Swiss Institute of Bioinformatics) σχημάτισαν το UniProt consortium. Με αυτόν τον τρόπο οι αλληλουχίες της PIR-PSD αλλά και ο σχολιασμός τους, ενσωματώθηκαν στην UniProt Knowledgebase. Προστέθηκαν διασυνδέσεις μεταξύ των καταχωρήσεων της UniProt και της PIR-PSD για να διευκολυνθεί ο εντοπισμός παλαιών καταχωρήσεων της PIR-PSD. Πρωτεΐνες που ήταν μοναδικές στην PIR-PSD όπως και οι αναφορές τους αλλά και τα πειραματικά δεδομένα που υπήρχαν στις σχετικές καταχωρήσεις μπορούν πλέον να βρεθούν στις αντίστοιχες καταχωρήσεις της UniProt.

2.1.3 Βάσεις δεδομένων τρισδιάστατων βιολογικών δομών.

Οι βάσεις αυτές περιέχουν δεδομένα που έχουν να κάνουν με την τρισδιάστατη δομή βιολογικών μακρομορίων. Οι τρισδιάστατες δομές αποτελούν το τελικό στάδιο μιας επίπονης διαδικασίας η οποία μετά τη χρήση μοριακών τεχνικών (κλωνοποίηση, απομόνωση, κρυστάλλωση κ.ο.κ.), οδηγεί τελικά στην υπολογιστική επίλυση της δομής μέσω της διαδικασίας της κρυσταλλογραφίας ακτίνων Χ, ή, σε πιο σπάνιες περιπτώσεις με φασματοσκοπία NMR. Το μεγαλύτερο ενδιαφέρον, βέβαια, έχουν οι δομές πρωτεϊνών, καθώς οι πρωτεΐνες είναι τα μακρομόρια των οποίων η μεγάλη ποικιλομορφία της δομής συνδέεται άμεσα με την βιολογική δράση. Η μοναδική βάση αυτού το είδους παγκοσμίως, είναι η PDB, η οποία και αναλύεται παρακάτω.

Protein Data Bank: Η Protein Data Bank (PDB, www.rcsb.org) είναι παγκοσμίως η μοναδική βάση στην οποία περιέχονται τρισδιάστατες δομές βιολογικών μακρομορίων (Kouranov et al., 2006). Ιδρύθηκε το 1971 στα εργαστήρια Brookhaven National Laboratories (BNL) των ΗΠΑ. Αρχικά αποτελούνταν από 7 δομές μακρομορίων οι οποίες προέκυψαν από κρυσταλλογραφικές μελέτες ενώ είχε μικρό ρυθμό αύξησης εγγραφών μέχρι τα τέλη της δεκαετίας του '70. Την δεκαετία του '80 παρατηρήθηκε σημαντική αύξηση του ρυθμού προσθήκης δεδομένων λόγω της τεχνολογικής εξέλιξης σε κάθε στάδιο του προσδιορισμού των δομών, ενώ πλέον η PDB περιέχει και δομές που έχουν προκύψει με φασματοσκοπία Πυρηνικού Μαγνητικού Συντονισμού (NMR). Σήμερα (Φεβρουάριος 2015) η PDB περιλαμβάνει 106.858 δομές βιομορίων. Οι εγγραφές της PDB περιλαμβάνουν εκτός από τις συντεταγμένες των ατόμων που απαρτίζουν τη δομή και επιπρόσθετα βοηθητικά στοιχεία όπως βιβλιογραφικές αναφορές, λεπτομέρειες για τον προσδιορισμό της δομής καθώς και άλλα στοιχεία που προκύπτουν από τη συγκεκριμένη δομή. Κάθε δομή πριν δημοσιευθεί στην βάση ελέγχεται για την ορθότητα της με τη χρήση ειδικού λογισμικού. Στη συνέχεια εφόσον περάσει τις δοκιμές με επιτυχία αποκτά ένα χαρακτηριστικό κωδικό και προστίθεται στη βάση.

Πρέπει να τονιστεί, ότι η καταχώρηση στην PDB είναι η τρισδιάστατη δομή, και όχι η πρωτεΐνη. Κατά συνέπεια, είναι δυνατόν να υπάρχει μια καταχώρηση της PDB η οποία να περιέχει περισσότερες από μία (ακόμα και μερικές δεκάδες) αμινοξικές αλληλουχίες πρωτεϊνών, όπως για παράδειγμα όταν αναφερόμαστε σε πολυενζυμικά σύμπλοκα τα οποία περιέχουν πολλές υπομονάδες. Επίσης, είναι δυνατόν να υπάρχουν περισσότερες από μία δομές μιας συγκεκριμένης πρωτεΐνης, καθώς είναι δυνατόν να έχουν γίνει διαφορετικά πειράματα είτε σε διαφορετικές συνθήκες, είτε παρουσία άλλων παραγόντων, είτε και απλά με άλλη τεχνική για να επιτευχθεί καλύτερη ευκρίνεια. Φυσικά, όπως είναι αναμενόμενο μόνο ένα μικρό υποσύνολο των γνωστών πρωτεϊνών έχουν γνωστή τρισδιάστατη δομή, καθώς η διαδικασία επίλυσης της δομής είναι χρονοβόρα και δύσκολη. Αυτό φαίνεται ξεκάθαρα αν συγκρίνουμε τον αριθμό των καταχωρήσεων της Uniprot με αυτόν της PDB. Ειδικότερα δε, για κάποιες ειδικές κατηγορίες πρωτεϊνών όπως οι διαμεμβρανικές πρωτεΐνες, τα πράγματα είναι ακόμα πιο δύσκολα από πειραματικής πλευράς και οι τρισδιάστατες δομές τους, είναι ακόμα πιο σπάνιες. Ένα τυπικό αρχείο PDB με τις επεξηγήσεις των πιο σημαντικών πεδίων, παρουσιάζεται στο παράρτημα. Τέλος, αξίζει να αναφερθεί, ότι παρόμοια βάση (MMDB) συντηρείται και στις ΗΠΑ στα πλαίσια του NCBI, με συνεχή όμως επαφή και ενημέρωση από την PDB.



Εικόνα 2.4: *Η εκθετική αύξηση των προσδιορισμένων πρωτεϊνικών δομών οι οποίες είναι κατατεθειμένες στην PDB, από το 1977 έως το τέλος του 2004.*

2.1.4 Βάσεις δεδομένων γονιδιακής έκφρασης

Εκτός από τις βάσεις δεδομένων αλληλουχιών και δομών, σημαντική είναι τα τελευταία χρόνια και η ανάπτυξη των βάσεων δεδομένων γονιδιακής έκφρασης. Με την εξέλιξη της τεχνολογίας και τη δημιουργία νέων οικονομικότερων τσιπ μικροσυστοιχιών, αλλά και με την εμφάνιση των τεχνολογιών Next Generation Sequencing, τα πειράματα ανάλυσης γονιδιακής έκφρασης πραγματοποιούνται με μεγαλύτερο ρυθμό και έτσι υπάρχει ανάγκη αποθήκευσης και ανάλυσης όλων αυτών των δεδομένων. Τη λύση στο παραπάνω πρόβλημα έδωσαν οι βάσεις δεδομένων οι οποίες περιέχουν δεδομένα από χιλιάδες πειράματα μικροσυστοιχιών. Οι βάσεις δεδομένων αυτές επιτρέπουν την καταχώρηση αποτελεσμάτων από πειράματα μικροσυστοιχιων, ενώ κάποιες από αυτές προσφέρουν και επιπλέον εργαλεία ανάλυσης. Επίσης, παρέχουν πληροφορίες σχετικά με το είδος των δεδομένων, την πλατφόρμα μικροσυστοιχίων που χρησιμοποιήθηκε στο πείραμα, τα γονίδια τα οποία μελετώνται καθώς επίσης και πληροφορίες σχετικά με τα είδη των δειγμάτων τα οποία χρησιμοποιήθηκαν. Η βασική δομή αυτών των αρχείων, διαφέρει πολύ από αυτά που αναφέραμε μέχρι τώρα, καθώς έχουμε να κάνουμε με έναν πίνακα, στον οποίο αναγράφονται τιμές "έκφρασης" ενός γονιδίου για κάθε άτομο. Συνήθως τα πειράματα αυτά αφορούν λίγα άτομα, αλλά ανάλογα με την πλατφόρμα μπορούμε να έχουμε δεδομένα έκφρασης για μερικές εκατοντάδες έως μερικές δεκάδες χιλιάδες γονίδια.

Επειδή ο όγκος των δεδομένων γονιδιακής έκφρασης είναι μεγάλος και πολύπλοκος, για να καταχωρηθούν τα δεδομένα των μικροσυστοιχιών στις δημόσιες βάσεις δεδομένων θα πρέπει να ακολουθούν ένα συγκεκριμένο πρωτόκολλο με βάση το οποίο καταχωρείται η ελάχιστη πληροφορία που περιγράφει ένα πείραμα μικροσυστοιχιών (MIAME:Minimun Information About a Microarray Experiment). Τα τελευταία χρόνια, γίνεται μεγάλη προσπάθεια το πρωτόκολλο αυτό να "επιβάλλεται" στους συγγραφείς οι οποίοι πρόκειται να δημοσιεύσουν μια σχετική εργασία. Δηλαδή, πριν η εργασία γίνει αποδεκτή από το επιστημονικό περιοδικό, θα πρέπει οι συγγραφείς να έχουν καταθέσει τα δεδομένα τους σε μια σχετική βάση δεδομένων (κάτι παρόμοιο ισχύει από χρόνια για τις αλληλουχίες και τις δομές μακρομορίων). Οι πιο γνώστες και συχνά χρησιμοποιούμενες βάσεις δεδομένων μικροσυστοιχιών είναι:

GeneExpression Omnibus (GEO): Βάση δεδομένων του NCBI που παρέχει δεδομένα γονιδιακής έκφρασης, τόσο από μικροσυστοιχίες όσο και από αλληλούχιση (next generation sequenicng) (Barrett & Edgar, 2006) Είναι διαθέσιμη στην ιστοσελίδα http://www.ncbi.nlm.nih.gov/geo/, ενώ στην ίδια διεύθυνση

υπάρχουν διαθέσιμα και κάποια διαδικτυακά εργαλεία που επιτρέπουν απλές αναλύσεις των δεδομένων της βάσης. Τα δεδομένα υπάρχουν τόσο σε ακατέργαστη (raw) όσο και σε επεξεργασμένη μορφή (με κανονικοποιήσεις κ.ο.κ.). Η βάση περιέχει (τον Φεβρουάριο του 2015), δεδομένα από 14.031 διαφορετικές πλατφόρμες έκφρασης, προερχόμενα από 1.357.732 "δείγματα", δηλαδή άτομα (στα οποία όμως δεν περιέχονται μόνο άνθρωποι, μπορεί να υπάρχουν δεδομένα από ζώα, φυτά ή ακόμα και μικρο-οργανισμούς), ταξινομημένα 55.725 "σειρές" (series) και 3.848 "σύνολα δεδομένων" (datasets). Το ίδιο δείγμα μπορεί να περιέχεται σε διαφορετικές σειρές και η ίδια σειρά σε ένα ή περισσότερα σύνολα δεδομένων.

Array Express: Δημόσια βάση δεδομένων μικροσυστοιχιών η οποία διατηρείται στο Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής, EBI, διαθέσιμη στην ιστοσελίδα http://www.ebi.ac.uk/arrayexpress/ (Brazma et al., 2003). Είναι της ίδιας λογικής με την GEO, την οποία περιέχει ως υποσύνολο βάσει της συνεργασίας των ιδρυμάτων. Στην ιστοσελίδα υπάρχουν επίσης διαθέσιμα εργαλεία για ανάλυση, οδηγίες για προγραμματιστική πρόσβαση στις υπηρεσίες και tutorials. Τον Φεβρουάριο του 2015, η βάση περιέχει δεδομένα για 57.009 πειράματα (experiments, τα οποία αντιστοιχούν στα series της GEO) και 1.689.237 μετρήσεις (assays, τα οποία περιέχουν ένα ή περισσότερα δείγματα).

Stanford Microarray Database (SMD): Βάση δεδομένων που κατασκευάστηκε αρχικά για να καλύπτει τις ανάγκες διαμοιρασμού αρχείων των ερευνητών του Stanford, αλλά μετεξελίχθηκε σταδιακά σε ένα δημόσιο αποθετήριο δεδομένων για μικροσυστοιχίες, http://smd.stanford.edu/ (Demeter et al., 2007). Περιέχει μικρότερο αριθμό δεδομένων από τις υπόλοιπες βάσεις, καθώς αυτή τη στιγμή έχει δεδομένα για 84.051 πειράματα από 631 δημοσιεύσεις.

2.1.5 Βάσεις δεδομένων γενετικής ποικιλομορφίας

Οι βάσεις αυτές, αν και συνδέονται στενά με τις βάσεις δεδομένων αλληλουχιών DNA, δεν αποτελούν ευθέως παράγωγα τους, αλλά μάλλον ανεξάρτητες οντότητες. Τούτο είναι κατανοητό αν σκεφτούμε ότι σε μια δεδομένη θέση ενός γονιδιώματος ενός είδους (πχ του ανθρώπου), τα διαφορετικά άτομα είναι δυνατόν να έχουν διαφορετική γενετική πληροφορία (πχ Α αντί για Τ, κ.ο.κ.). Η βάση η οποία καταγράφει τους πολυμορφισμούς και τις συχνότητες τους στους διάφορους πληθυσμούς είναι η dbSNP, ενώ η βάση που καταγράφει πρωτογενώς τουλάχιστον τις αλληλοσυσχετίσεις των πολυμορφισμών αυτών, είναι η HapMap.

dbSNP: Η dbSNP είναι η δημόσια βάση για τους νουκλεοτιδικούς πολυμορφισμούς http://www.ncbi.nlm.nih.gov/snp (Sherry et al., 2001). Εκτός από νουκλεοτιδικούς πολυμορφισμούς (single nucleotide polymorphisms - SNPs), περιέχει και δεδομένα για πολυμορφικές θέσεις που αφορούν απαλοιφές ή εισαγωγές βάσεων (deletion insertion polymorphisms -DIPs), καθώς και για ένθετα μεταθετά στοιχεία και μικροδορυφορικές επαναλήψεις (short tandem repeats - STRs). Κάθε καταχώρηση στην dbSNP περιέχει πληροφορίες για το που βρίσκεται ο πολυμορφισμός (δηλαδή την περιβάλλουσα αλληλουχία), τη συχνότητα του πολυμορφισμού σε διάφορους πληθυσμούς, αλλά και για την πειραματική μέθοδο, τα πρωτόκολλα και τις συνθήκες με τις οποίες μετρήθηκε η ποικιλομορφία. Η dbSNP δέχεται επίσης υποβολές για καταχωρήσεις πολυμορφισμών από κάθε είδος, αλλά και από διαφορετικά σημεία του γονιδιώματος. Λεπτομερής περιγραφή της βάσης δεδομένων υπάρχει στο ελεύθερο διαδικτυακό βιβλίο του NCBI στη διεύθυνση http://www.ncbi.nlm.nih.gov/books/NBK3848/. Στην έκδοση 129 (2008) η βάση είχε πάνω από 14 εκατομύρια πολυμορφισμούς, αλλά προφανώς ο αριθμός αυτός αυξάνεται συνεχώς.

HapMap: Το International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) είναι το αποτέλεσμα μιας διεθνούς συνεργασίας σε μια προσπάθεια να εντοπισθούν και να καταγραφούν οι γενετικές διαφορές αλλά και οι ομοιότητες των ανθρώπινων πληθυσμών (HapMap, 2003). Ο σκοπός του προγράμματος είναι να συγκρίνει τις γενετικές αλληλουχίες διαφορετικών ατόμων (από διαφορετικούς πληθυσμούς) και να εντοπίσει με αυτόν τον τρόπο χρωμοσωμικές περιοχές στις οποίες οι γενετικές παραλλαγές (συνήθως, νουκλεοτιδικοί πολυμορφισμοί), κληρονομούνται μαζί. Στην αρχική φάση του προγράμματος, έγινε χρήση γενετικών δεδομένων από 4 πληθυσμούς Αφρικανικής, Ασιατικής και Ευρωπαϊκής καταγωγής. Σε μεταγενέστερες εκδόσεις, προστέθηκαν και άλλοι πληθυσμοί, σε μια προσπάθεια να υπάρχει όσο το δυνατό μεγαλύτερη κάλυψη παγκοσμίως. Τα τελικά δεδομένα που είναι διαθέσιμα από τη βάση αυτή, είναι οι απλότυποι, δηλαδή οι συνδυασμοί πολυμορφισμών που συνκληρονομούνται, και ακριβέστερα οι συντελεστές ανισορροπίας σύνδεσης (Linkage Disequilibrium), των διαφόρων πολυμορφισμών του ιδίου χρωμοσώματος, μεταξύ τους. Με τη χρήση αυτής της πληροφορίας, είναι δυνατόν να σχεδιαστούν μεθόδοι και αλγόριθμοι στατιστικής γενετικής με τους οποίους θα επιχειρείται να απαντηθούν ερωτήματα σχετικά με τη γενετική προδιάθεση σε

ασθένειες και την ανταπόκριση σε φάρμακα. Επιπλέον, τέτοια δεδομένα είναι πολύ χρήσιμα στη μελέτη της γενετικής δομής των ανθρώπινων πληθυσμών.

2.1.6 Βάσεις δεδομένων βιβλιογραφίας

Παρόλο που οι βάσεις αυτές δεν είναι με την στενή έννοια «βιολογικές βάσεις δεδομένων», ιστορικά, αλλά και για λόγους που θα φανούν στην πορεία, είναι καλό να γίνεται αναφορά και σε αυτές. Οι βάσεις αυτές, έχουν σαν «καταχώρηση» τα στοιχεία μιας επιστημονικής δημοσίευσης (συγγραφέας, περιοδικό, περίληψη κ.ο.κ.). Η κυριότερη βάση του είδους, είναι η **PubMed** (http://www.ncbi.nlm.nih.gov/pubmed) η οποία στεγάζεται στο NCBI και περιλαμβάνει περισσότερα από 24 εκατομύρια καταχωρήσεις επιστημονικών άρθρων από τη βιοϊατρική βιβλιογραφία (έχοντας κάλυψη της MEDLINE, άλλων περιοδικών των επιστημών της ζωής αλλά και από κάποια online βιβλία). Οι αναφορές μπορεί να περιέχουν συνδέσμους στο πλήρες κείμενο των εργασιών, είτε μέσω της PubMed Central (το υποσύνολο με τις ελεύθερα διαθέσιμα δημοσιεύσεις πλήρους κειμένου), είτε απευθείας μέσω των ιστοσελίδων των εκδοτικών οίκων. Παρόλο που τα στοιχεία της PubMed είναι δημόσια διαθέσιμα, το να έχει πρόσβαση κανείς στο πλήρες κείμενο μιας εργασίας, εξαρτάται από την πολιτική του εκδοτικού οίκου. Στην ίδια ιστοσελίδα, υπάρχουν διαθέσιμα και tutorials για τη χρήση της υπηρεσίας (http://www.nlm.nih.gov/bsd/disted/pubmed.html).

Άλλες βάσεις δεδομένων, παρόμοιας φύσης, είναι το SCOPUS (http://www.scopus.com/) και το Web of Science (http://webofknowledge.com/). Οι βάσεις αυτές, παρέχουν περισσότερες πληροφορίες, με την κυριότερη να είναι οι βιβλιογραφικές αναφορές (citations) που έχει πάρει κάθε δημοσιευμένη εργασία. Αυτό επιτρέπει την αντίστροφη αναζήτηση (πχ εύρεση του ποια εργασία έχει αναφέρει μια δεδομένη εργασία), αλλά και την αξιολόγηση του συνολικού έργου (ενός συγγραφέα, ενός περιοδικού ή ενός ιδρύματος). Το βασικότερο μειονέκτημα αυτών των βάσεων είναι ότι διατίθενται από ιδιωτικούς οργανισμούς και απαιτούν συνδρομή του χρήστη είτε του ινστιτούτου του.

Η πρόσβαση στη βιβλιογραφία, εκτός του ότι είναι απαραίτητη εργασία στην καθημερινότητα ενός επιστήμονα, αποτελεί επιπλέον, ένα ιδιαίτερα αναπτυσσόμενο κομμάτι της επιστήμης της πληροφορικής (text mining), το οποίο έχει βρει ιδιαίτερες εφαρμογές στη βιοπληροφορική, καθώς η ύπαρξη ενός τεράστιου όγκου δεδομένων από κείμενα (περιλήψεις εργασιών κυρίως), έχει δώσει την αφορμή για μελέτες αυτών των κειμένων με σκοπό την ανακάλυψη συσχετίσεων και την εξαγωγή βιολογικών συμπερασμάτων (Ananiadou, Kell, & Tsujii, 2006; Scherf, Epple, & Werner, 2005).

2.2. Δευτερογενείς βάσεις δεδομένων

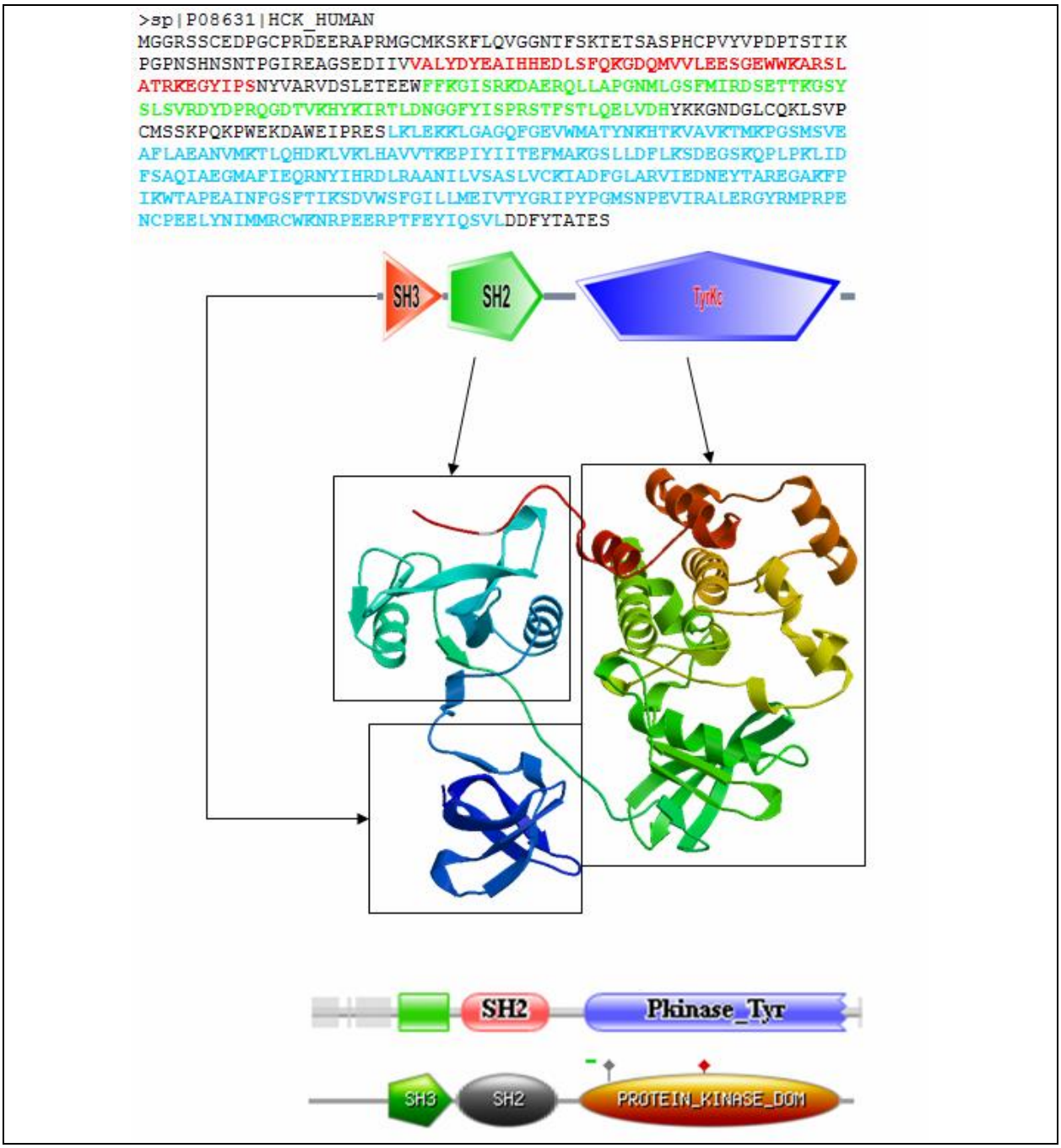
Σε αυτήν την μεγάλη αλλά και ετερογενή κατηγορία περιλαμβάνονται κυρίως βάσεις δεδομένων που περιέχουν διαφόρων ειδών ταξινομήσεις των πρωτογενών δεδομένων, χρήσιμες για αναλυτικούς σκοπούς, και διακρίνονται περαιτέρω σε βάσεις οικογενειών και σε εξειδικευμένες βάσεις δεδομένων.

2.2.1 Βάσεις δεδομένων οικογενειών

Όπως είναι γνωστό, οι πρωτεΐνες γενικά απότελούνται από μία ή περισσότερες διακριτές λειτουργικές περιοχές (domains), οι οποίες πολλές φορές είναι και δομικά αυτοτελείς. Οι περιοχές αυτές, θεωρείται ότι μπορούν να λειτουργήσουν αλλά και και να εξελιχθούν ανεξάρτητα από το υπόλοιπο τμήμα της πρωτεΐνης. Διαφορετικοί συνδυασμοί τέτοιων περιοχών οδηγούν σε μια μεγάλη ποικιλία των πρωτεϊνών στη φύση. Συνεπώς, η ανίχνευση τέτοιων περιοχών είναι σημαντική στην προσπάθεια λειτουργικής ταξινόμησης των πρωτεϊνών. Στο κεφάλαιο της στοίχισης αλληλουχιών θα μιλήσουμε αναλυτικά για το ρόλο που παίζει αυτό το φαινόμενο στην αναζήτηση ομοιότητας αλληλουχιών (τοπική στοίχιση), ενώ στο κεφάλαιο της γονιδιωματικής θα μιλήσουμε για το πως μπορεί η ανίχνευση πρωτεϊνών με διαφορετική σύσταση σε τέτοιες περιοχές να δώσει στοιχεία για τη λειτουργική ή άλλη αλληλεπίδραση μεταξύ πρωτεϊνών μή όμοιων μεταξύ τους.

Οι βάσεις που αναλύονται παρακάτω, επιτελούν πολύ σημαντικό ρόλο στην ταξινόμηση των αμινοξικών αλληλουχιών πρωτεϊνών σε οικογένειες. Επιπλέον δε, πρέπει να έχουμε υπόψη μας ότι καθώς οι δομές είναι περισσότερο συντηρημένες από τις αλληλουχίες, η ύπαρξη αυτών των βάσεων βοηθάει στην εύκολη ταυτοποίηση και κατηγοριοποίηση νέων πρωτεϊνών, και στην εύκολη αναγνώριση ενός νέου πρωτεϊνικού διπλώματος. Οι βάσεις διαφέρουν μεταξύ τους, κυρίως α) στον τρόπο εύρεσης και μαθηματικής

μοντελοποίησης της περιοχής (με τοπική ομοιότητα, με pattern, με HMM κ.ο.κ.), και β) στον τρόπο με τον οποίο έχει καθοριστεί εξαρχής η περιοχή. Οι CATH και SCOP βασίζονται αποκλειστικά σε δομικά κριτήρια, ενώ οι PROSITE, PFAM, INTERPRO λαμβάνουν υπόψη κυρίως την αλληλουχία. Κατά συνέπεια, περιέχουν μεγαλύτερο αριθμό καταχωρήσεων, καθώς οι πρωτεΐνες με γνωστή δομή είναι πολύ λιγότερες. Επιπλέον δε λόγω αυτού του γεγονότος, είναι δυνατόν, σε κάποιες περιπτώσεις οι περιοχές που έχουν οριστεί να διαφέρουν.



Εικόνα 2.5: *Αναπαράσταση της ανθρώπινης κινάσης τυροσίνης HCK (Uniprot: P08631, PDB: 2HCK_A). Φαίνεται η αμινοξική αλληλουχία, και η διάρθρωση των δομικών αυτοτελών περιοχών (domains) στην τρισδιάστατη δομή. Κάτω, η ίδια πρωτεΐνη όπως την αναπαριστούν οι βάσεις PFAM και PROSITE αντίστοιχα. Καθώς οι περιοχές αυτής της πρωτεΐνης είναι δομικά αυτοτελείς, ίδια αναπαράσταση υπάρχει και στην SCOP. Σε άλλες περιπτώσεις, οι περιοχές που αναπαρίστανται στην PFAM και την PROSITE, μπορεί να μην αντιστοιχούν σε δομικά αυτοτελείς περιοχές, οπότε υπάρχει ενδεχόμενο οι βάσεις αυτές να διαφωνούν μεταξύ τους όσον αφορά στα όρια των περιοχών.*

Η **PROSITE** (http://www.expasy.ch/prosite/) αποτελεί μια βάση ταξινόμησης αμινοξικών αλληλουχιών πρωτεϊνών και αυτοτελών περιοχών αλληλουχιών (sequence domains) σε οικογένειες (Sigrist et al., 2010). Η ταξινόμηση σε οικογένειες πραγματοποιείται βάσει των ομοιοτήτων που παρουσιάζουν οι περιοχές των αλληλουχιών μεταξύ τους. Πρωτεΐνες ή περιοχές που ανήκουν στην ίδια οικογένεια έχουν πιθανότατα την ίδια λειτουργία και προέρχονται από κοινό πρόγονο. Υπάρχουν τμήματα των αμινοξικών αλληλουχιών πρωτεϊνών που είναι περισσότερο συντηρημένα στην πορεία της εξέλιξης τους και σχετίζονται άμεσα με τη λειτουργία τους και με τη δομή των πρωτεϊνών στο χώρο. Η ανάλυση αμινοξικών αλληλουχιών πρωτεϊνών που ανήκουν στην ίδια οικογένεια, μέσω μιας πολλαπλής στοίχισης, είναι πιθανό να οδηγήσει σε ένα 'αποτύπωμα' χαρακτηριστικό για κάθε οικογένεια, ικανό να τη διαχωρίζει από τις πρωτεϊνικές αλληλουχίες που δεν ανήκουν σε αυτήν την οικογένεια.

Υπάρχουν γενικά δύο τρόποι για τη δημιουργία των 'αποτυπωμάτων'. Ο ένας βασίζεται στη χρήση μιας γλώσσας παρόμοιας με αυτής των "κανονικών εκφράσεων" (regular expressions) ή μοτίβων, και είναι ο πιο παλιός και εύκολος στη δημιουργία, ενώ ο άλλος βασίζεται στην κατασκευή προφίλ (profiles), πίνακες δηλαδή με ειδικές ανά θέση πιθανότητες εμφάνισης αμινοξέων, μέθοδος η οποία είναι πιο σύνθετη αλλά και πιο ευαίσθητη. Περισσότερα για τις τεχνικές αυτές, θα αναφερθούν σε επόμενο κεφάλαιο. Μέχρι σήμερα η PROSITE περιέχει 'αποτυπώματα' για περίπου 1716 οικογένειες για καθεμία από τις οποίες συμπεριλαμβάνεται λεπτομερής ανάλυση για τη δομή και τη λειτουργία των πρωτεϊνών που την αποτελούν. Συνολικά, υπάρχουν στη βάση 1308 μοτίβα ή πρότυπα (patterns), 1107 προφίλ και 1105 "κανόνες" (τα οποία αφορούν κυρίως πληροφορίες για το που θα πρέπει να βρίσκεται το μοτίβο για να θεωρηθεί έγκυρο αλλά και πληροφορίες για συνδυασμούς από μοτίβα). Προφανώς, υπάρχουν οικογένειες για τις οποίες υπάρχουν διαθέσιμα και μοτίβα και προφίλ (συνήθως, οι παλαιότερες καταχωρήσεις αφορούσαν το μοτίβο). Στην βάση υπάρχουν επίσης αναλύσεις για τις πρωτεΐνες της Uniprot που ανήκουν σε κάθε οικογένεια όσο και για τις πρωτεΐνες στις οποίες εμφανίζεται ένα "αποτύπωμα" (κυρίως όταν έχουμε να κάνουμε με μοτίβο) αλλά είναι γνωστό ότι αυτές δεν ανήκουν λειτουργικά στην οικογένεια αυτή. Τέλος, υπάρχουν εργαλεία για την αναζήτηση των μοτίβων και των προφίλ σε αλληλουχίες, όσο και εργαλεία αναπαράστασης της "σπονδυλωτής" δομής των πρωτεϊνών, δηλαδή της αναπαράστασης των περιοχών αυτών και την αποτύπωση τους πάνω σε μια δεδομένη αλληλουχία.

PFAM*:* Η βάση Pfam (http://pfam.xfam.org/) αποτελεί μια μεγάλη συλλογή πρωτεϊνικών οικογενειών (Finn et al., 2014). Βασίζεται στην ίδια λογική με την PROSITE (ειδικά με το υποσύνολο της που βασίζεται σε προφίλ), αλλά η μεγάλη διαφορά είναι ότι εδώ οι οικογένειες χαρακτηρίζονται από ένα hidden Markov model (HMM), μέθοδος η οποία είναι πιο ευαίσθητη στον εντοπισμό μακρινών ομόλογων, χωρίς όμως να υστερεί σε ταχύτητα και αποτελεσματικότητα. Στην τρέχουσα έκδοση (2013), η βάση περιέχει δεδομένα για 14.831 οικογένειες παρέχοντας κάλυψη για πάνω από το 80% των πρωτεϊνικών καταχωρήσεων της UNIPROT.

Η PFAM αποτελείται από δύο υποσύνολα, την PFAM-A, και την PFAM-B. Η PFAM-A αποτελείται από καταχωρήσεις (οικογένειες) υψηλής «ποιότητας», καθώς έχουν όλες υποστεί σχολιασμό από ειδικούς, ενώ υπάρχουν αναφορές σε άλλες βάσεις δεδομένων και κυρίως σε βιβλιογραφία. Η PFAM-B είναι το υποσύνολο, το οποίο προκύπτει με αυτοματοποιημένο τρόπο εντοπίζοντας τις ομοιότητες ανάμεσα στις πρωτεϊνικές περιοχές που απομένουν όταν αφαιρεθούν οι περιοχές που αντιστοιχούν στις καταχωρήσεις της PFAM-A. Η PFAM-B είναι ιδιαίτερα χρήσιμη, γιατί με στοχευμένη ανάλυση αυτών των «οικογενειών», μπορούν να προκύψουν οικογένειες που μετέπειτα θα «προαχθούν» στην PFAM-A. Το βασικό χαρακτηριστικό της PFAM και αυτό που την κάνει τόσο δημοφιλή, είναι ότι με τη χρήση του HMM (και ειδικά του πακέτου HMMER, βλ. στο αντίστοιχο κεφάλαιο), μπορεί να επιλεγεί για κάθε οικογένεια μία τιμή διαχωριστικού κατωφλίου στο σκορ, και κατά συνέπεια κάθε πρωτεΐνη ταξινομείται μόνο σε μία οικογένεια (σε αυτή που σκοράρει πάνω από το κατώφλι). Παρ' όλα αυτά, χαμηλότερη ομοιότητα μπορεί να υπάρχει μεταξύ πρωτεϊνών που ανήκουν σε διαφορετικές οικογένειες, για το λόγο αυτό η βάση περιέχει και μια ανώτερη κατηγορία οργάνωσης, την υπερ-οικογένεια (clan).

CATH: Η CATH (http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html) είναι μια βάση ιεραρχικής ταξινόμησης πρωτεϊνικών δομών που αποτελούν εγγραφές της PDB με βάση τις αυτοτελείς δομικές περιοχές (domains) που τις απαρτίζουν (Knudsen & Wiuf, 2010). Η CATH περιέχει αποκλειστικά πρωτεϊνικές δομές που είναι προσδιορισμένες σε ευκρρίνεια μεγαλύτερη των 3 Angstroms και χρησιμοποιεί κυρίως αυτοματοποιημένες μεθόδους για την ταξινόμησή τους. Σε ειδικές περιπτώσεις και όταν αυτό κρίνεται απαραίτητο χρησιμοποιούνται και ανθρώπινα κριτήρια. Η ιεραρχία αποτελείται κυρίως από τέσσερα επίπεδα: 1) την Τάξη (Class), 2) την Αρχιτεκτονική (Architecture), 3) την Τοπολογία (Οικογένεια διπλώματος)

(Topology (fold family)) και 4) την Ομόλογη Οικογένεια (Homologous superfamily). Οι πρωτεΐνες που αποτελούνται από περισσότερες από μία αυτοτελείς δομικές περιοχές (domains), αναλύονται στα επιμέρους στοιχεία αυτόματα με βάση ειδικούς αλγόριθμους αναγνώρισης των περιοχών. Η αυτόματη αυτή διαδικασία κατατάσσει το 53% των δομών. Οι υπόλοιπες διαχωρίζονται στις επιμέρους αυτοτελείς δομικές περιοχές με παρατηρήσεις που προκύπτουν είτε από τους αλγόριθμους αυτόματου διαχωρισμού είτε από τη βιβλιογραφία. Η ταξινόμηση πραγματοποιείται μόνο στις αυτοτελείς δομικές περιοχές. Η ανάλυση της ιεραρχίας στην CATH έχει ως εξής:

C - Τάξη (Class): Οι δομές ταξινομούνται σε 4 μεγάλες ομάδες βάσει των στοιχείων δευτεροταγούς δομής των αυτοτελών δομικών περιοχών και είναι οι: 1) mainly-alpha, όπου τα στοιχεία δευτεροταγούς δομής είναι στην συντριπτική τους πλειοψηφία α-έλικες, 2) mainly-beta, όπου τα στοιχεία δευτεροταγούς δομής είναι κυρίως β-εκτεταμένες δομές, 3) alpha-beta, όπου παρατηρούνται εναλλασσόμενες α/β και α+β δομές και 4) δομές με χαμηλό ποσοστό δευτεροταγών δομών. Η διαδικασία της ταξινόμησης γίνεται αυτόματα για το 90% των πρωτεϊνών ενώ για το υπόλοιπο 10% χρησιμοποιούνται κυρίως δεδομένα από τη βιβλιογραφία.

A - Αρχιτεκτονική (Architecture): Η ταξινόμηση πραγματοποιείται βάσει της γενικότερης δομής της αυτοτελούς δομικής περιοχής (domain), λαμβάνοντας υπόψη τον προσανατολισμό των στοιχείων δευτεροταγούς δομής αλλά όχι τον τρόπο διασύνδεσης μεταξύ τους π.χ. βαρέλια (barrels).

T - Τοπολογία (Topology): Οι δομές ομαδοποιούνται με βάση τον προσανατολισμό των στοιχείων δευτεροταγούς δομής καθώς και τον τρόπο σύνδεσής τους.

H - Ομόλογη οικογένεια (Homology superfamily): Σε αυτό το επίπεδο ταξινομούνται τα δομικά στοιχεία που έχουν ομοιότητα 35% στο επίπεδο της αλληλουχίας τους με αποτέλεσμα να θεωρείται ότι προέρχονται από ένα κοινό πρόγονο.

S - Αλληλουχία (Sequence family): Τα μέλη της εμφανίζουν ομοιότητα πάνω από 35% στο επίπεδο της αλληλουχίας με αποτέλεσμα να θεωρούνται ότι έχουν παρόμοια δομή και λειτουργία.

SCOP: Ο βασικός στόχος της βάσης SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/index.html) είναι η ανάλυση των δομικών και εξελικτικών σχέσεων που παρατηρούνται μεταξύ όλων των πρωτεϊνών γνωστής δομής καταχωρημένων στην PDB (Andreeva, et al., 2004). Η ταξινόμηση των πρωτεϊνών πραγματοποιείται βάσει αυτών των δομικών και εξελικτικών σχέσεων. Τα βασικά επίπεδα ταξινόμησης είναι τέσσερα: 1) η οικογένεια (Family), 2) η υπερ-οικογένεια (Superfamily), 3) το δίπλωμα (Fold) και 4) η τάξη (Class).

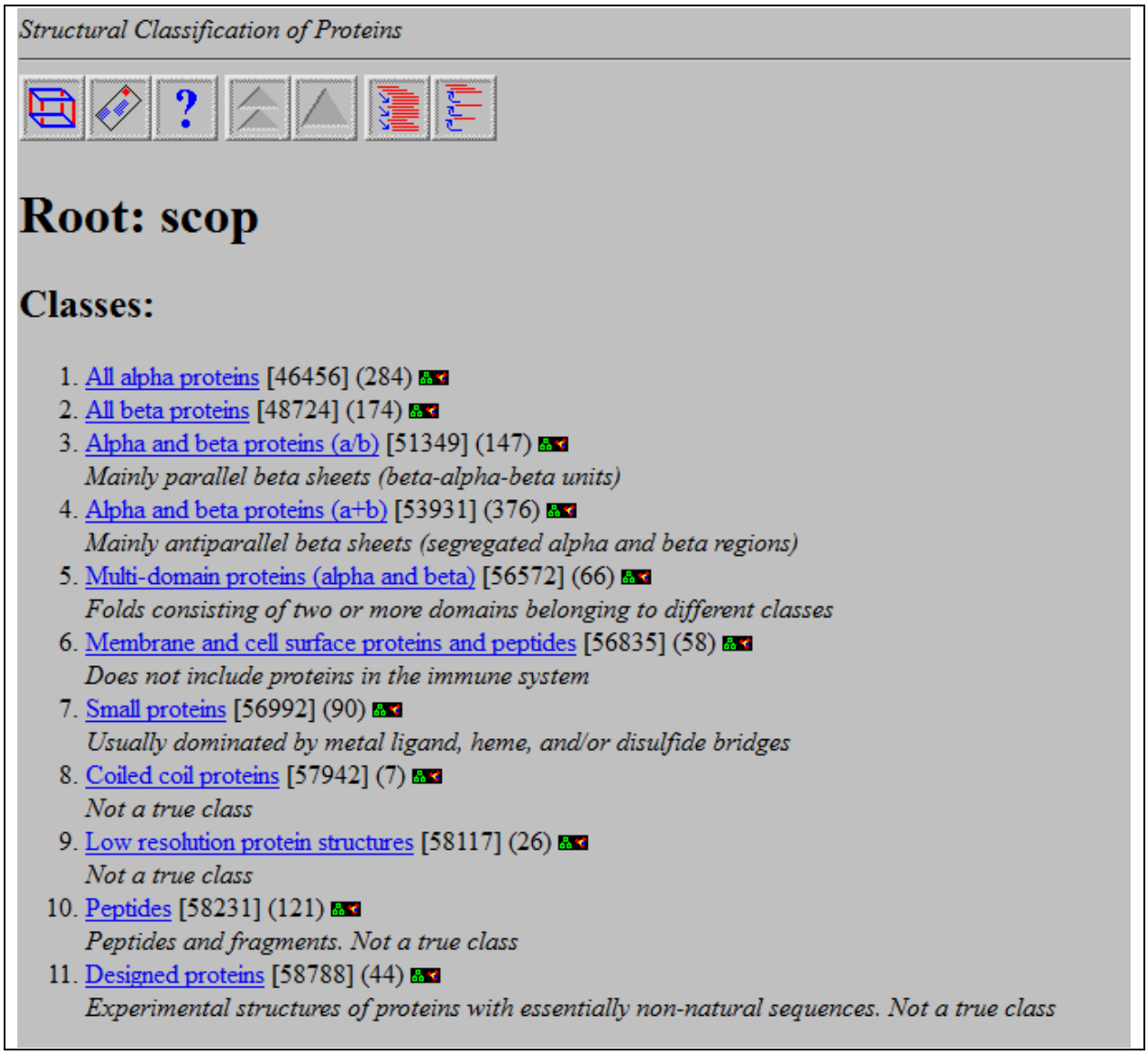
Οικογένεια (Family): Μεταξύ των μελών της οικογένειας παρατηρείται ξεκάθαρη εξελικτική σχέση. Η ομοιότητα σε επίπεδο αλληλουχίας είναι ίση ή μεγαλύτερη του 30%. Παρ' όλα αυτά υπάρχουν περιπτώσεις στις οποίες οι δομές και η λειτουργία είναι παρόμοιες υποδηλώνοντας κοινό πρόγονο ενώ η ομοιότητα σε επίπεδο αλληλουχίας είναι μικρότερη του 30% (σφαιρίνες, 15%).

Υπερ-οικογένεια (Superfamily): Οι πρωτεΐνες που κατατάσσονται στις υπερ-οικογένειες εμφανίζουν πολύ μικρή ομοιότητα στο επίπεδο της αλληλουχίας αλλά τα δομικά τους χαρακτηριστικά και η λειτουργία τους υποδηλώνουν ότι πιθανά έχουν προέλθει από κοινό πρόγονο.

Δίπλωμα (Fold): Σε αυτό το επίπεδο κατατάσσονται πρωτεΐνες που παρουσιάζουν ομοιότητα σε επίπεδο δομής. Οι πρωτεΐνες που εμφανίζουν το ίδιο δίπλωμα έχουν τα ίδια, σε μεγάλο βαθμό, χαρακτηριστικά δευτεροταγούς δομής, με κοινό προσανατολισμό και τις ίδιες τοπολογικές συνδέσεις μεταξύ τους. Πρωτεΐνες που έχουν το ίδιο δίπλωμα αλλά δεν είναι όμοιες από άποψη αμινοξικής αλληλουχίας έχουν ορισμένα περιφερειακά στοιχεία της δευτεροταγούς τους δομής και στροφές ανόμοια και όσον αφορά στο μέγεθος και όσον αφορά στη διαμόρφωση. Πρωτεΐνες που εμφανίζουν κοινό δίπλωμα δεν είναι απαραίτητο να έχουν κοινή εξελικτική προέλευση.

Τάξη (Class): Η ταξινόμηση γίνεται με βάση το δίπλωμα των στοιχείων δευτεροταγούς δομής των πρωτεϊνών σε τέσσερις κύριες δομικές κατηγορίες: 1) την all-α, όπου η δομή σχηματίζεται από α-έλικες, 2) την all-β, όπου η δομή αποτελείται από β-πτυχωτές επιφάνειες, 3) την α/β, όπου στην δομή της πρωτεΐνης εναλλάσσονται α-έλικες και β-πτυχωτές επιφάνειες και 4) την α+β, όπου σε διακριτές περιοχές της δομής βρίσκονται α-έλικες και β-πτυχωτές επιφάνειες.

Η αναγνώριση των σχέσεων καθώς και η ταξινόμηση βάσει των σχέσεων μεταξύ των πρωτεϊνών πραγματοποιείται αποκλειστικά από ειδικούς επιστήμονες μετά από λεπτομερή μελέτη και σύγκριση των πρωτεϊνικών δομών. Αυτοματοποιημένες μέθοδοι χρησιμοποιούνται μόνο για την ομοιογένεια των δεδομένων που περιέχονται στη βάση.



Εικόνα 2.6: *Εικόνα της ιεραρχίας στη βάση SCOP (τάξη, δίπλωμα, υπεροικογένεια, οικογένεια).*

2.2.2 Εξειδικευμένες βάσεις δεδομένων

Εκτός από τις μεγάλες, δημόσια διαθέσιμες και ευρέως χρηματοδοτούμενες βάσεις δεδομένων που αναφέρθηκαν παραπάνω, σημαντικό ρόλο στην πρόοδο της βιοπληροφορικής παίζουν και οι εξειδικευμένες βάσεις δεδομένων. Συνήθως, αλλά όχι πάντα, αφορούν τις αμινοξικές αλληλουχίες πρωτεϊνων (γιατί για αυτές υπάρχει μεγάλη πληθώρα λειτουργικών δεδομένων, σε μεγάλη λεπτομέρεια, που δεν μπορεί να καλυφθεί από τις βάσεις όπως η Uniprot), και τις περισσότερες φορές, συντηρούνται από μικρές ή μεσαίου μεγέθους ερευνητικές ομάδες. Στην ενότητα αυτή θα εξεταστούν προβλήματα που αντιμετωπίζουν οι διαχειριστές αυτών των βάσεων δεδομένων και θα συζητηθούν οι λόγοι που οι επιστήμονες μπορεί να προτιμούν να δημοσιεύουν τα δεδομένα τους σε βάσεις δεδομένων αντί ιστοσελίδες ή παραδοσιακά σε άρθρα επιστημονικών περιοδικών. Τονίζεται η ανάγκη δημιουργίας, πηγών εξειδικευμένων βάσεων δεδομένων, ειδικά όταν τα δεδομένα είναι δύσκολο ή αδύνατο να παρουσιαστούν στις παραδοσιακές πηγές.

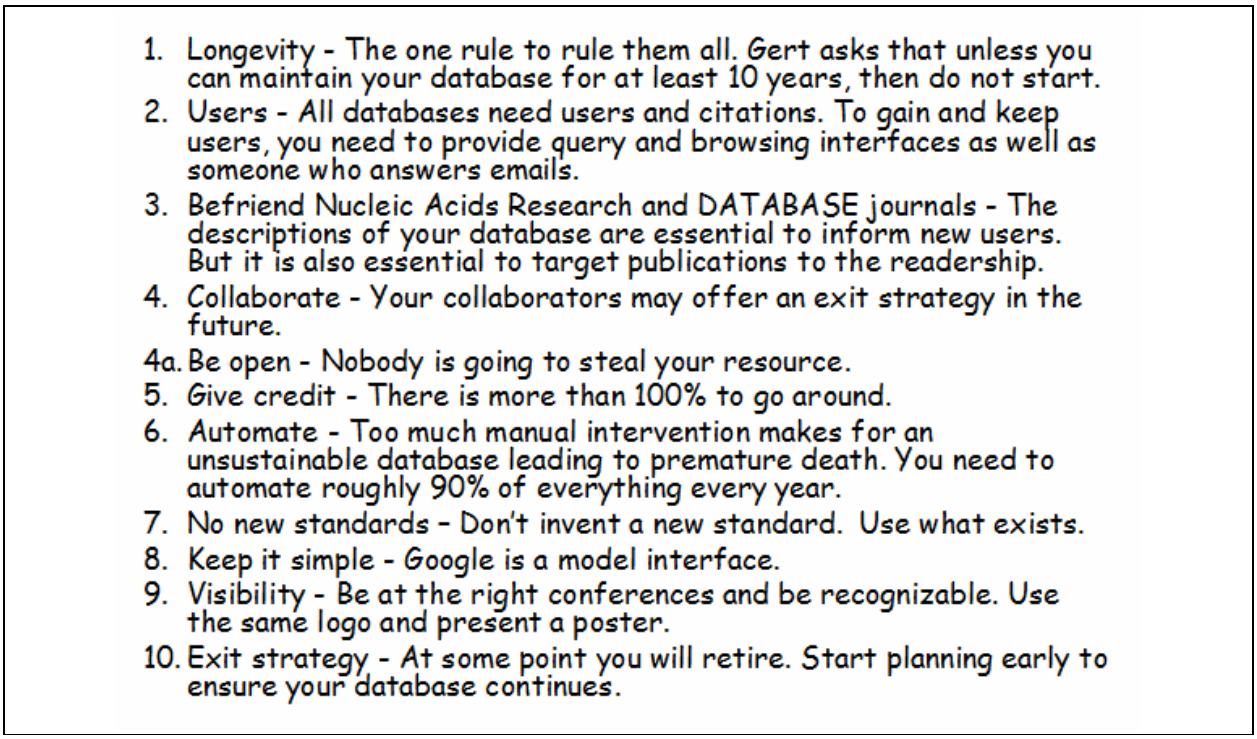
Στις 11-12 Αυγούστου 2014 πραγματοποιήθηκε με την χρηματοδότηση του Wellcome Trust, στο Hinxton της Αγγλίας, μία συνάντηση είκοσι ενός κύριων ερευνητών που ο καθένας διατηρεί μια εξειδικευμένη πρωτεϊνική βάση δεδομένων ή διεξάγει έρευνα σχετικά με την διατήρηση ενός τέτοιου αποθετηρίου (Specialized Protein Resources Network). Το θέμα της συνάντησης ήταν η χάραξη πολιτικής για

την δημιουργία και διατήρηση πρωτεϊνικών βάσεων δεδομένων και αποτελούνταν από πέντε ενότητες: (1) βασικές προκλήσεις, (2) εισαγωγή δεδομένων, (3) βέλτιστες πρακτικές για τη διατήρηση και την επιμέλεια, (4) ροή πληροφοριών προς και από τα μεγάλα κέντρα δεδομένων, και (5) επικοινωνία και χρηματοδότηση. Στο τέλος συνοψίζονται τα συνολικά συμπεράσματα που προέκυψαν από αυτήν την συνάντηση (Holliday et al., 2015).

Στην συνάντηση συμμετείχαν ερευνητές που διατηρούν «εξειδικευμένες» ηλεκτρονικές βάσεις δεδομένων συγκεκριμένων ειδών πρωτεϊνών (όπως αυτές ορίζονται από τα ενζυματικά, λειτουργικά ή δομικά χαρακτηριστικά τους) αλλά και διαχειριστές μεγάλων πρωτεϊνικών αποθετηρίων (συμπεριλαμβανομένων των Pfam, RefSeq, Swiss-Prot, και UniProt). Αυτά τα μεγάλα κέντρα δεδομένων χρησιμοποιούν διάφορες προσεγγίσεις για να συντηρήσουν το περιεχόμενο των δεδομένων τους, όπως η υπολογιστική ανάλυση, η συνεργασία, η ενοποίηση δεδομένων από πολλαπλές πηγές, και η επιμέλεια από ειδικούς σχολιαστές. Όλες οι βάσεις δεδομένων υποστηρίζονται από ειδικό σχολιασμό ώστε να εξασφαλιστεί η ακρίβεια και η πληρότητα των στοιχείων που παρουσιάζονται σε κάθε μικρή ή μεγάλη πρωτεϊνική βάση δεδομένων. Ένα κοινό πρόβλημα όλων των συμμετεχόντων της συνάντησης ήταν η επιμέλεια και η ανανέωση των βάσεων, δεδομένου ότι είναι δύσκολη η ανάκτηση πληροφοριών από δημοσιευμένα άρθρα επειδή συχνά δεν αναφέρουν αναλυτικές συγκεκριμένες πληροφορίες για τον υπό μελέτη οργανισμό (ειδικά για τα strains), ή τις ακριβείς πληροφορίες της αλληλουχίας που αναλύθηκε (πχ ο κωδικός πρόσβασης στη Uniprot ή το gi). Η διεύρυνση των συνεργασιών για τη διόρθωση λαθών στις βάσεις δεδομένων και η διάδοση της γνώσης αναγνωρίστηκαν από όλους ως βασικοί τρόποι δράσης, που θα ωφελήσουν όλες τις πρωτεϊνικές πηγές αλλά και τους χρήστες τους.

Η δημιουργία μίας βάσης δεδομένων θα μπορούσε να θεωρηθεί εύκολη διαδικασία όταν υπάρχουν κάποια στοιχεία διαθέσιμα, στην πραγματικότητα όμως υπάρχουν πολλές προκλήσεις και εμπόδια που πρέπει να αντιμετωπιστούν. Κάθε βάση δεδομένων έχει τις δικές της μοναδικές προκλήσεις και προβλήματα, αλλά κάποια από αυτά είναι κοινά σε όλες τις βάσεις και μπορούν να συνδυαστούν σε ένα βασικό ερώτημα: Τι κάνει μια βάση δεδομένων σημαντική;

Κύρια πρόκληση είναι η αξιοπιστία και όχι η ποσότητα των δεδομένων. Εξαρτάται εξολοκλήρου από το πεδίο εφαρμογής και τη λειτουργία της βάσης δεδομένων. Για παράδειγμα η βάση δεδομένων ESTHER, που εξετάζει μόνο τις εστεράσες και τα άλφα-βήτα ένζυμα υδρολάσης, και η GPCRDB που εξετάζει μόνο τα GPCRs, δεν πρόκειται ποτέ να έχουν τον ίδιο αριθμό καταχωρήσεων με την UniProtKB, η οποία περιλαμβάνει όλες τις αλληλουχίες αμινοξέων που έχουν βρεθεί μέχρι τώρα. Από μία ανάλυση που πραγματοποιήθηκε το 2009 (Schnoes, Brown, Dodevski, & Babbitt, 2009) προκύπτει ότι ορισμένες βάσεις δεδομένων έχουν ποσοστό σφαλμάτων/λάθος σχολιασμών (misannotation) περίπου 80%. Αντίθετα η Swiss-Prot που είναι το τμήμα της UniProtKB στο οποίο τα σχόλια καταχωρούνται χειροκίνητα από τους διαχειριστές, είχε ποσοστό σφάλματος περίπου 0%. Υπάρχουν πολλοί διαφορετικοί τύποι σφαλμάτων που μπορούν να βρεθούν στις πηγές δεδομένων. Κάποια είναι σχετικά εύκολο να εντοπιστούν μέσω αυτοματοποιημένων διαδικασιών, όπως για παράδειγμα τα ορθογραφικά λάθη στο σχολιασμό. Σφάλματα όμως που σχετίζονται με επιστημονικές πληροφορίες είναι πολύ πιο δύσκολο να βρεθούν, ειδικά αφού η γνώση εξελίσσεται πολύ γρήγορα. Χαρακτηριστικό παράδειγμα τέτοιου είδους σφάλματος αποτελεί ο ενζυματικός μηχανισμός δράσης της λυσοζύμης. Για πάνω από 50 χρόνια ο κοινά αποδεκτός μηχανισμός περιελάμβανε ένα ενδιάμεσο ζεύγος ιόντων. Νέα πειράματα όμως έδειξαν ότι περιλαμβάνει το σχηματισμό ενός ομοιοπολικού συμπλόκου γλυκοσυλενζύμου (Kirby, 2001). Οπότε τίθενται διάφορα ερωτήματα όπως για παράδειγμα: Ο αρχικός μηχανισμός ήταν πραγματικά λάθος; Μπορούμε ποτέ να ελπίζουμε ότι θα μπορέσουμε να προσδιορίσουμε τέτοιου είδους πληροφορίες; Έχουν καταχωρηθεί και προωθηθεί οι νέες πληροφορίες σε όλες τις βάσεις δεδομένων; Πιθανόν όχι, αλλά το κλειδί για να διατηρούνται οι βάσεις ενημερωμένες είναι οι διαχειριστές (και/ή χρήστες) να ανατρέχουν συχνά στη βιβλιογραφία ώστε να ενημερώνονται για ό,τι νέο υπάρχει, έχει αλλάξει ή θεωρείται απαρχαιωμένο.



Εικόνα 2.7: *Ο δεκάλογος της "καλής λειτουργίας" μιας βάσης δεδομένων, όπως παρουσιάστηκε από τον Καθ. Gert Vriend*

Υπάρχουν πολλοί ακόμα τύποι σφαλμάτων, για παράδειγμα ένα συχνό σφάλμα στην ανάλυση πρωτεϊνικών αλληλουχιών σχετίζεται με την σπονδυλωτή (modular) δομή πολλών πρωτεϊνών. Συνήθως συναντάται σε ενεργοποιημένα ένζυμα από υδατάνθρακες όπου ένα μέρος της σπονδυλωτής δομής (Module) που προσδένει υδατάνθρακες (carbohydrate-binding module, CBM) βρίσκεται συχνά προσαρτημένο σε καταλυτικές περιοχές που ανήκουν σε διάφορες οικογένειες ή ακόμη και σε δομές άγνωστης λειτουργίας. Μία καλή στοίχιση στο Blast, η οποία στοιχίζει μόνο το CBM, οδηγεί συχνά σε λανθασμένο σχολιασμό των παρακείμενων δομικών περιοχών. Το ίδιο μπορεί να λειτουργήσει και με αντίθετο τρόπο, όπως για παράδειγμα όταν μία πρωτεΐνη με μία μόνο περιοχή (single domain protein) αντιστοιχίζεται σε μια πρωτεΐνη πολλαπλής δομής και ο σχολιασμός μεταφέρεται από την δομή που δεν είναι αντιστοιχισμένη (π.χ. το ένζυμο που σχετίζεται με την αμινοτρανσφεράση (UniProtKB: B8NM72)). Αυτή η πρωτεΐνη, που εμπλέκεται στη βιοσύνθεση ενός δευτερογενούς μεταβολίτη τύπου-πεπτιδίου, στο παρελθόν θεωρούνταν ότι ήταν μια μη-ριβοσωμική πεπτιδική συνθετάση (NRPS), πιθανότατα λόγω της μεταφοράς του αυτόματου σχολιασμού από τα ομόλογά της που έχουν τη δομή και λειτουργία του NRPS. Ωστόσο, με προσεκτικό και χειρωνακτικό σχολιασμό των εμπλεκόμενων πρωτεϊνών (Umemura et al., 2014), διαπιστώθηκε ότι από την πρωτεΐνη αυτή έλειπε η περιοχή NRPS και ότι στην πραγματικότητα ήταν μια ριβοσωμική πρωτεΐνη. Αυτή είναι μια περίπτωση όπου ακόμη και ένα μικρό λάθος μπορεί να οδηγήσει πολλούς ερευνητές σε λανθασμένα συμπεράσματα. Επίσης γίνεται εμφανές, γιατί ο χειρωνακτικός σχολιασμός είναι απαραίτητος στις Εξειδικευμένες Πρωτεϊνικές Βάσεις Δεδομένων (Specialist Protein Resources - SPRs).

Ένας άλλος τύπος σφάλματος προκαλείται από την υπερεκτίμηση που συνάγεται από την "απόδειξη μέσα από την επανάληψη". Αυτό ενισχύεται περαιτέρω από το γεγονός ότι η λειτουργία μιας πρωτεΐνης μπορεί να οριστεί από το μοριακό/χημικό της ρόλο (π.χ. μία κινάση σερίνης) ή από την ευρεία βιολογική διαδικασία στην οποία μεσολαβεί (π.χ. η πρωτεΐνη η οποία μεσολαβεί στην πήξη του αίματος). Γενικά, είναι αρκετά δύσκολο να αποκρυπτογραφηθεί ο βιολογικός ρόλος της πρωτεΐνης στο ενδογενές πλαίσιο χρησιμοποιώντας υπολογιστικές μεθόδους και συνεπώς, τέτοιες προβλέψεις θα πρέπει να χρησιμοποιούνται με προσοχή. Η αναζήτηση με το BLAST στη non-redundant βάση δεδομένων πρωτεϊνών του NCBI, συχνά εντοπίζει ένα μεγάλο αριθμό παρόμοιων πρωτεϊνών που προέρχονται σχεδόν αποκλειστικά από γονιδιωματικές αλληλουχίες. Εξετάζοντας προσεχτικά τα ονόματά τους παρατηρείται ότι είναι ετερογενείς και ότι γίνεται μετάβαση από την μία στην άλλη χωρίς επίβλεψη. Επιπλέον πολλά ομόλογα ένζυμων στερούνται των καθοριστικών αμινοξέων καταλοίπων στο ενεργό κέντρο, καθιστώντας τα μη ενεργά. Από την άλλη πλευρά, σφάλματα συναρμολόγησης γονίδιων (gene assembly errors) προκαλούν υποθετικές

πρωτεΐνες στις οποίες έχει ταυτοποιηθεί λάθος εναρκτήρια μεθιονίνη, ή με εξώνια που έχουν παραλειφθεί, έχοντας σαν πιθανό αποτέλεσμα την παράλειψη ενεργών κατάλοιπων. Παρά το γεγονός ότι αυτά τα λάθη μπορούν στη συνέχεια να διορθωθούν, ο έλεγχος και η διόρθωση των σχολιασμών τους αποτελεί πρόκληση για τους διαχειριστές των SPRs.

Πως μπορούν να διορθωθούν σφάλματα που έχουν εντοπιστεί στις βάσεις δεδομένων; Πολλές πηγές, όπως η UniProtKB, διαθέτουν μηχανισμούς με τους οποίους οι χρήστες μπορούν να αναφέρουν τα πιθανά πιθανά προβλήματα. Άλλες βάσεις, όπως η PDB, δεν επιτρέπουν την διόρθωση των δεδομένων, αν και υπάρχει η PDB_REDO (Joosten, Long, Murshudov, & Perrakis, 2014) η οποία επιτρέπει την διόρθωση των ατομικών συντεταγμένων. Για τα άλλα είδη σφαλμάτων σχολιασμού και ιδιαίτερα εκείνων που σχετίζονται με την αλληλουχία των αμινοξέων, έχουν προταθεί ειδικές μεθοδολογίες για τον εντοπισμό και την διόρθωσή τους (Nagy et al., 2008; Wong, Maurer-Stroh, & Eisenhaber, 2010).

Όταν το σφάλμα διορθωθεί, πως μπορούμε να ενημερώσουμε όλες τις βάσεις δεδομένων που χρησιμοποιούν την αρχική εγγραφή; Η προέλευση των δεδομένων είναι συχνά δύσκολο να εντοπιστεί. Οι διαχειριστές των βάσεων δεδομένων έλαβαν την πληροφορία από την UniProtKB, ή από την πρωτογενή βιβλιογραφία; Ίσως να την πήραν από το SFLD, οπότε τίθεται το ερώτημα: οι διαχειριστές του SFLD από πού την πήραν; Ορισμένες πηγές (π.χ. UniProtKB) έχουν αρχίσει να χρησιμοποιούν το ECO (Evidence Code Ontology) (Chibucos et al., 2014), στο οποίο περιλαμβάνονται και οι πηγές. Η συμπλήρωση όμως πηγών με σχολιασμό τέτοιου είδους είναι συχνά περίπλοκη διαδικασία, καθώς όλα τα δεδομένα πρέπει να διασταυρώνονται και να ελέγχονται. Ένας από τους μελλοντικούς στόχους είναι η δημιουργία κανόνων καταχώρησης σχολιασμού βάσεων όπου θα είναι σημαντική η δυνατότητα γνώσης της πηγής τους με αποτέλεσμα η χρήση του ECO ή κάποιου παρόμοιου κώδικα να είναι απαραίτητη.

Με τον συνεχώς αυξανόμενο όγκο διαθέσιμων δεδομένων, πώς θα μπορούσε να διατηρηθεί ή ακόμα και να βελτιωθεί η αξιοπιστία των πηγών των βάσεων; Ο ειδικός διαχειριστής της βάσης (άτομο που είναι εκπαιδευμένο σε έναν συγκεκριμένο τομέα) παίζει πάντα καθοριστικό ρόλο. Οι βάσεις τις οποίες τις διαχειρίζονται άνθρωποι οι οποίοι πραγματοποιούν διασταύρωση στοιχείων, έχουν μεγαλύτερη αξιοπιστία σε σχέση με τις βάσεις στις οποίες τα δεδομένα απλά καταχωρούνται αυτόματα και συνήθως αναπαράγουν σφάλματα. Ωστόσο, ο ρόλος των χρηστών θα είναι πολύ σημαντικός στο μέλλον. Για παράδειγμα, οι χρήστες όταν εντοπίσουν κάποιο σφάλμα θα μπορούν να επικοινωνούν με τη βάση δεδομένων (δίνοντας αποδείξεις για την ύπαρξη του σφάλματος) ώστε να διορθώνεται η καταχώρηση. Υπάρχει περίπτωση βέβαια ο κατάλογος των σφαλμάτων να ξεπεράσει πολύ γρήγορα την ικανότητα της βάσης δεδομένων να τα διορθώσει. Επιπλέον, οι χρήστες θα μπορούν να προτείνουν νέες καταχωρήσεις ή ακόμα και να καταχωρούν δεδομένα. Το μεγαλύτερο εμπόδιο σε αυτή τη μέθοδο σχολιασμού είναι η εκπαίδευση των χρηστών για τον εντοπισμό των σφαλμάτων.

Ένας τρόπος (όπως εφαρμόστηκε στην διεθνή κοινότητα κρυσταλλογραφίας) είναι η εισαγωγή δεδομένων να αποτελεί προαπαιτούμενο της δημοσίευσης των αποτελεσμάτων. Ωστόσο, χωρίς την υποστήριξη και την επιβολή αυτής της απόφασης από τα περιοδικά, είναι αδύνατη η απόκτηση επαρκούς λειτουργικής πληροφορίας. Τα δεδομένα που καταχωρούνται σε πολλές περιπτώσεις δεν χρειάζεται να είναι ιδιαίτερα λεπτομερειακά. Για παράδειγμα, σημαντική πρόοδος θα μπορούσε να είναι, μαζί με τον αριθμό πρόσβασης αλληλουχιών να συμπεριλαμβάνεται και ο αριθμός Enzyme Commission (EC).

Ένας άλλος τρόπος, θα μπορούσε είναι να μέσω της Βικιπαίδεια (WikiPedia). Αυτή η μέθοδος χρησιμοποιείται ήδη από τη βάση Rfam. Οι συγγραφείς θα μπορούσαν να δημιουργούν μια σελίδα της Wikipedia, η οποία θα χρησιμοποιείται για να συμπληρωθεί η βάση δεδομένων Rfam. Ωστόσο, ποιες πηγές θα είναι οι πρωτογενείς συλλέκτες δεδομένων; Θα είναι η Swiss-Prot η μοναδική πηγή για όλους τους σχολιασμούς των αμινοξέων που θα έχουν ως αναφορά οι άλλες πηγές; Θα συμφωνήσουν όλα τα περιοδικά στην προτεινόμενη διαδικασία; Η διαδικασία σχολιασμού θα είναι αρκετά απλή και ολοκληρωμένη ώστε οι συγγραφείς να την ακολουθήσουν; Δυστυχώς δεν υπάρχει απλή απάντηση σε αυτά τα ερωτήματα, αλλά καθώς αυξάνεται ο όγκος των δεδομένων, οι δημιουργοί των SPRs, οι χρήστες και οι εκδότες θα πρέπει να τα αντιμετωπίσουν.

Για να είναι χρήσιμη μία πηγή, η γλώσσα που χρησιμοποιούν και οι δύο θα πρέπει να είναι τυποποιημένη. Παράδειγμα ενός τέτοιου εγχειρήματος αποτελεί το ερευνητικό πρόγραμμα EMBRACE (Pettifer et al., 2010). Αυτό που εννοεί μία βάση με τον όρο superfamily μπορεί να μην σημαίνει το ίδιο σε μία άλλη βάση. Για παράδειγμα, ο ορισμός SFLD απαιτεί οι πρωτεΐνες να είναι όχι μόνο εξελικτικά σχετικές, αλλά και να έχουν μια συντηρημένη χημεία. Η TIGRFAM, από την άλλη πλευρά, απαιτεί απλώς να υπάρχει εξελικτική συγγένεια. Η SFLD έχει μια ιεραρχία, η οποία αντιστοιχίζεται στο PANTHER, αλλά οι όροι που χρησιμοποιούνται είναι διαφορετικοί (μια υποομάδα SFLD είναι το ισοδύναμο μιας οικογένειας PANTHER).

Επίσης, υπάρχει το θέμα της χημείας. Η πιο κοινή μέθοδος ταξινόμησης της χημείας του ενζύμου είναι ο αριθμός Enzyme Commission (EC). Δημιουργήθηκε για να αποφευχθεί η πληθώρα εσωτερικών ονομασιών (όπως ινβερτάση, σουμπτιλισίνη, κ.λ.π.) και να συνδέσει τα ονόματα με τα μόρια (συνήθως τα υποστρώματα και τους συνολικούς χημικούς μετασχηματισμούς που συμβαίνουν, αλλά όχι την αλληλουχία και την δομή). Ακόμη και μεταξύ των βάσεων MACiE και EzCatDB, οι οποίες ταξινομούν αντιδράσεις ενζύμων, είναι πιθανό να χρειάζεται τυποποίηση της γλώσσας ή του λεξιλογίου, δεδομένου ότι διαχείρισή τους γίνεται με διαφορετικούς τρόπους. Η βάση MACiE κατατάσσει τα στάδια της αντίδρασης, ενώ η EzCatDB ταξινομεί ολόκληρες τις αντιδράσεις που αποτελούνται από ένα ή περισσότερα στάδια. Για αρκετά χρόνια γινόταν μία προσπάθεια βιοχημικού χαρακτηρισμού των ενζύμων, προσδιορίζοντας τον αντίστοιχο αριθμό EC (που χαρακτηρίζει τη γενική χημική αντίδραση που καταλύει το ένζυμο) για το κάθε ένα.

Σημαντικό πρόβλημα αποτελεί το γεγονός ότι ο αριθμός EC ορίστηκε την δεκαετία του ’50. Από τότε έχει βρεθεί ότι πολλά από τα ένζυμα που έχουν αριθμό EC είναι μη ειδικά. Παρ’ όλα αυτά, ο αριθμός EC χρησιμοποιείται μέχρι σήμερα ενώ δημιουργούνται ακόμα και νέοι αριθμοί EC, με αποτέλεσμα να μην μπορούμε να αγνοήσουμε την ύπαρξη του ως ένα χρήσιμο εργαλείο. Ωστόσο, με την αύξηση της δυσκολίας της δημοσίευσης των χαρακτηρισμών των ενζύμων σε περιοδικά υψηλής απήχησης, η σωστή μέθοδος συσχετισμού των πρωτεϊνών με τον αντίστοιχο αριθμό EC έχει εξαλειφθεί, και σήμερα χρησιμοποιείται σχεδόν αποκλειστικά στο πεδίο της βιοπληροφορικής. Επιπλέον, ο αριθμός EC είχε ως στόχο να χαρακτηρίσει την χημική αντίδραση που εκτελείται από ένα ένζυμο, και όχι να αποδίδεται ανάλογα με την ομοιότητα των αλληλουχιών καθώς η ίδια αντίδραση μπορεί να καταλύεται από πολλές μη σχετιζόμενες οικογένειες αλληλουχιών (π.χ. οι β-λακταμάσες) και πολλά ένζυμα ομαδοποιημένα στην ίδια οικογένεια καταλυτικών αντιδράσεων που περιγράφονται από διαφορετικούς αριθμούς EC (π.χ. οι ενδονουκλεάσες). Συμπεραίνουμε επομένως ότι η απόδοση του αριθμού EC είναι πολύ πιο περίπλοκη διαδικασία σε σχέση με την απλή απόδοση του EC βάσει της καλύτερης στοίχισης στο BLAST. Τέτοιου είδους ζητήματα μεταφοράς σχολιασμού αποτελούν πρόκληση όχι μόνο για τους χρήστες των βάσεων δεδομένων, αλλά και για τους ειδικούς. Πώς ξέρουμε πότε θα πρέπει να διαδοθούν και πότε όχι οι λειτουργικές πληροφορίες; Σε μερικές περιπτώσεις, όπως η βάση δεδομένων CAZy, προτιμάται να μην διαδίδονται οι λειτουργικές πληροφορίες και απλά αναφέρονται οι λειτουργίες που έχουν προσδιοριστεί πειραματικά. Άλλες βάσεις, όπως η MACiE και η EzCatDB, απλώς αναφέρουν τους αριθμούς EC των ομολόγων, αλλά περιλαμβάνουν την ταυτότητα των συντηρημένων κατάλοιπων, έτσι ώστε οι χρήστες να μπορούν να βγάλουν τα δικά τους συμπεράσματα ως προς την εγκυρότητα των προβλέψεων.

Παρόλο που δεν είναι αποδεκτή η άποψη ότι όλες οι πηγές θα πρέπει να χρησιμοποιούν την ίδια γλώσσα (η βιολογία είναι πολύπλοκη, οπότε ένας όρος σε ένα πεδίο δεν μπορεί να μεταφραστεί με ακρίβεια σε κάποιο άλλο πεδίο), πιθανότατα θα ήταν χρήσιμο να βρεθεί ένας τρόπος να μεταφράζονται οι έννοιες. Οι οντολογίες είναι ίσως ο πιο κατάλληλος τρόπος. Παρά το γεγονός ότι η οντολογία γονιδίων (GO) είναι ίσως η πιο ευρέως γνωστή οντολογία στον τομέα της βιοπληροφορικής, πρέπει να γνωρίζουμε ότι δεν είναι η μοναδική. Μια αναζήτηση στο PubMed με τον όρο "οντολογία" στον τίτλο των εγγράφων αποδίδει περίπου 1.500 αποτελέσματα. Αν και μπορεί να μην υπάρχουν οντολογίες για κάθε πεδίο της βιοχημείας και της βιολογίας, υπάρχουν σε πολλά, και για μερικές από τις βασικές έννοιες (π.χ. ένα ένζυμο) υπάρχει η δυνατότητα σύνδεσης των δεδομένων σε όλες τις πηγές που έχουν παρόμοια στοιχεία. Είναι θεμιτό να γνωρίζουμε την οντολογία ή το λεξιλόγιο που χρησιμοποιείται, αλλά θα ήταν πιο χρήσιμο για όλους, τους χρήστες και τους διαχειριστές, και τις βάσεις οντολογίας, όπως η BioPortal (Grosjean, Soualmia, Bouarech, Jonquet, & Darmoni, 2014) και η OBO Foundry (Smith et al., 2007), η συλλογή όσο το δυνατόν περισσότερων πληροφοριών σε μια ενιαία βάση.

Εικόνα 2.8: *Φωτογραφία των συμμετεχόντων στο Protein Bioinformatics and Community Resources Retreat. Το όνομα κάθε επιστήμονα ακολουθείται από το όνομα της εξειδικευμένης βάσης δεδομένων την οποία διευθύνει.. Πίσω σειρά: David Landsman (Histone database), Dan Haft (TIGRFAMS), Bernard Henrissat (CAZy), Rob Finn (InterPro and Pfam), David Craik (ConoServer and CyBASE), Arnaud Chatonnet (ESTHER), Neil Rawlings (MEROPS); Μεσαία σειρά: Amos Bairoch (neXtProt), Gerard Manning (Kinase.com), Michael Spedding (IUPHAR), Gert Vriend (GPCRDB), Milton Saier (TCDB), Pantelis Bagos (OMPdb); Εμπρός σειρά: Narayanaswamy Srinivasan (KinG), Ramanathan Sowdhamini (PASS2), Alex Bateman (Pfam & UniProt), Patsy Babbitt (SFLD), Kim Pruitt (RefSeq), Claire O'Donovan (UniProt), Gemma Holliday (MACiE), Nozomi Nagano (EzCatDB).*

Η βάση δεδομένων του περιοδικού Nucleic Acids Research (Fernández-Suárez, Rigden, & Galperin, 2014) περιείχε το 2014 συνολικά 1.552 βάσεις δεδομένων από τις οποίες οι 58 ήταν νέες και οι 123 παλιότερες που ανανεώθηκαν. Η δημιουργία μίας βάσης δεδομένων είναι εύκολη διαδικασία. Για παράδειγμα πολλές βάσεις δεδομένων έχουν δημιουργηθεί στο πλαίσιο διδακτορικών διατριβών ή μεταπτυχιακών προγραμμάτων. Η δυσκολία έγκειται στην διατήρησή της. Μία μελέτη που πραγματοποιήθηκε το 2008 έδειξε ότι περίπου το 40% των διευθύνσεων URL των βάσεων δεδομένων που ήταν δημοσιευμένες σε επιστημονικά περιοδικά πλέον δεν ήταν διαθέσιμες (Wren, 2008). Ωστόσο, η διατήρηση της βάσης δεν αφορά μόνο την ύπαρξη μιας σταθερής διεύθυνσης URL. Το πρώτο πράγμα που χρειάζεται κάθε βάση δεδομένων είναι το προσωπικό που θα την διατηρεί και θα συνεχίσει να την αναπτύσσει. Μερικές από αυτές συντηρούνται από ειδικούς επιστήμονες που εργάζονται μόνοι τους ή/και στον ελεύθερο χρόνο τους, αλλά αυτό είναι δύσκολο να λειτουργήσει μακροπρόθεσμα. Τι συμβαίνει όταν ο επιστήμονας που διατηρεί την βάση πρέπει να προχωρήσει και δεν υπάρχει κανείς να τον αντικταστήσει; Ίσως μια λύση στο πρόβλημα αυτό είναι η ενοποίηση των βάσεων. Παράδειγμα αποτελεί η Interpro, μία πηγή που ενσωματώνει πολλές διαφορετικές πηγές. Οι βάσεις δεδομένων που την αποτελούν εξακολουθούν να διατηρούν τη δική τους ταυτότητα και να έχουν τον δικό τους ρόλο χωρίς να μπορεί να διατηρήσει τις βάσεις η Interpro. Η εξασφάλιση χρηματοδότησης είναι μια συνεχής πρόκληση για μικρές ή/και ανεξάρτητες (από μεγαλύτερες πηγές, όπως η REFSEQ ή UniProtKB) SPRs. Υπάρχουν αρκετοί τρόποι αύξησης των πόρων που διατίθενται για τις SPRs, όπως για παράδειγμα, οι επιχορηγήσεις οργανισμών, (π.χ. η SFLD αυτή τη στιγμή υποστηρίζεται από μια επιχορήγηση του NIH), χρηματοδότηση χρηστών (εμπορική), (π.χ. η KEGG (Kanehisa et al., 2014) στην οποία επιτρέπεται η πρόσβαση σε συνδρομητές), δηλαδή χρηματοδοτείται από τους χρήστες, ενώ έχουν προταθεί και άλλα πιο σύνθετα μοντέλα.

Η συνεχής επικαιροποίηση των δεδομένων των βάσεων αποτελεί ίσως την μεγαλύτερη πρόκληση που έχουμε να αντιμετωπίσουμε σήμερα. Ο όγκος των δεδομένων που διατίθενται είναι τεράστιος ενώ η αύξηση των διαθέσιμων δεδομένων είναι εκθετική. Η UniProtKB τον Νοέμβριο του 2014 είχε πάνω από 86 εκατομμύρια εγγραφές, εκ των οποίων σχολιάστηκαν χειροκίνητα ή αναθεωρήθηκαν περίπου μισό εκατομμύριο. Για κάθε πληροφορία που γνωρίζουμε για μία μόνο πρωτεΐνη, υπάρχουν ακόμη περισσότερες πρωτεΐνες για τις οποίες δεν έχουμε κανένα στοιχείο, εκτός από την πρωταρχική αλληλουχία αμινοξέων. Ο αυτόματος σχολιασμός και οι υποθέσεις είναι κρίσιμης σημασίας για να συνεχιστεί η καταχώριση της πληθώρας των πρωτογενών αλληλουχιών. Διότι, ακόμη και με αναλύσεις υψηλής απόδοσης, όπως αυτές που παρέχονται από το Structural Genomics Consortium, οι πειραματιστές δεν μπορούν να προχωρήσουν σε βιοχημικές ή ακόμα και υπολογιστικές προβλέψεις ενώ το πρωτεϊνικό δίπλωμα δεν παρέχει σχεδόν ποτέ ακριβείς πληροφορίες για την πρωτεϊνική λειτουργία ώστε να είναι απευθείας χρήσιμες σε έναν βιολόγο. Για παράδειγμα, μια πρωτεΐνη της οποίας το όνομα έχει δοθεί από την τάξη διπλώματος, όπως η «putative glycoside hydrolase» ή ένα ομόλογο πεπτιδάσης, δεν παρέχει στον χρήστη καμιά είδους ειδικότητα και άρα επαρκείς πληροφορίες ώστε να γνωρίζει ακριβώς τη λειτουργία της. Για παράδειγμα, οι κυτταρινάσες (τα ένζυμα που διασπούν την κυτταρίνη των φυτών) και η νευραμινιδάση (που επιτρέπει στον ιό της γρίπης να ολοκληρώσει επιτυχώς τον κύκλο μόλυνσής του), είναι και τα δύο γλυκοσιδάσες (glycoside hydrolases). Είναι επίσης σημαντικό να γίνει διάκριση μεταξύ της βιοχημικής λειτουργίας (όπως η χημική διάσπαση της κυτταρίνης) και του βιολογικού ρόλου (όπως παροχή βοήθειας σε ένα ιό για να ολοκληρώσει τον κύκλο μόλυνσης του) καθώς οι βιοχημικές λειτουργίες σχετίζονται περισσότερο με πρωτεϊνικές αλληλουχίες και δομές (Martin et al., 1998) από ότι με τον βιολογικό τους ρόλο. Για παράδειγμα οι πρωτεΐνες που έχουν πάνω από μία λειτουργία (moonlighting proteins) έχουν απολύτως όμοιες αλληλουχίες και δομές, αλλά έχουν διαφορετικούς ρόλους, συχνά ανάλογα με την κυτταρική τους θέση. Ως εκ τούτου, είναι πιο δύσκολο να προσδιοριστεί η λειτουργία μιας πρωτεΐνης από ότι η τρισδιάστατη δομή της και η μεταφορά του σχολιασμού μπορεί τουλάχιστον να βοηθήσει τους χρήστες δίνοντας ένα αρχικό στοιχείο για την τεκμαιρόμενη λειτουργία της. Ωστόσο, οι χρήστες, οι διαχειριστές και οι δημιουργοί των βάσεων, πρέπει όλοι να γνωρίζουν τις διαφορές στον καθορισμό αυτών των λειτουργικών επίπεδων, όπως τη σημασιολογική ακρίβεια που θα βοηθήσει τους χρήστες να βρουν τις πληροφορίες που θέλουν, αλλά και τη αυτόματη μεταφορά του σχολιασμού που εξακολουθεί να απαιτεί όχι μόνο ένα καλό μοντέλο για την πραγματοποίησή του, αλλά και υψηλής ποιότητας και όσο το δυνατόν πληρέστερα δεδομένα. Επομένως, τα μέλη της κοινότητας SPR πρέπει να εργάζονται από κοινού για να ελαχιστοποιηθεί η επικάλυψη των προσπαθειών, έχοντας ως κοινό στόχο την διατήρηση της ποιότητας αλλά και της ποσότητας των δεδομένων ώστε οι χρήστες των βάσεων να έχουν τα καλύτερα δυνατά δεδομένα. Πολύτιμη είναι επίσης και η βοήθεια των χρηστών, χωρίς τους οποίους καμία πηγή δεν μπορεί να αναπτυχθεί και να ευδοκιμήσει.

Παρακάτω, δίνεται μια σύντομη περιγραφή των εξειδικευμένων βάσεων των οποίων οι επιστημονικοί υπεύθυνοι και εκπρόσωποι συμμετείχαν στη συνάντηση του Protein Bioinformatics and Community Resources Retreat (Εικόνα 2.8).

TCDB: Πολλά από τα αποθετήρια πληροφοριών που συνήθως θεωρούνται ιστοσελίδες, στην πραγματικότητα αποτελούν σχεσιακές βάσεις δεδομένων (Stein, 2013) και επιτρέπουν την διάθεση των δεδομένων και την οργάνωση της γνώσης, ταυτόχρονα βάσει πολλαπλών κριτήριων. Οι αλληλουχίες είναι πιθανό να έχουν πολλά ονόματα και να ανήκουν σε πολλές ομάδες, καθεμία από τις οποίες αποθηκεύεται ιεραρχικά. Πλεονεκτήματα των σχεσιακών βάσεων δεδομένων (structured query language - SQL) αποτελούν η οργάνωση και η αναζήτηση των δεδομένων με πολλούς διαφορετικούς τρόπους καθώς και το γεγονός ότι συνδέονται με άλλα συστήματα (Jamison, 2003). Σε αυτή την ενότητα, θα συζητήσουμε σχετικά με το σύστημα διαχείρισης της βάση δεδομένων Transporter Classification Database (TCDB, www.tcdb.org (Saier, Reddy, Tamang, & Västermark, 2014)). Η TCDB ξεκίνησε ως μια απλή ιστοσελίδα σε HTML. Το 1998 μετατράπηκε σε μια σχεσιακή (Oracle MySQL) βάση δεδομένων με διεπαφή PHP και σήμερα στεγάζεται στο San Diego Supercomputer Center (www.sdsc.edu). Πρόκειται για μια βάση δεδομένων των πρωτεϊνικών συστημάτων μεταφοράς από όλους τους ζωντανούς οργανισμούς που κατατάσσονται σύμφωνα με την κατηγορία, υποκατηγορία, οικογένεια, υπο-οικογένεια και το σύστημα. Τα συστήματα μεταφοράς μπορεί να αποτελούνται από απλές ή σύνθετες πρωτεΐνες (multi-component), με μέγιστο έως περίπου 100 πρωτεΐνες ανά σύστημα. Η TCDB χρησιμοποιείται ως κοινό σημείο αναφοράς για τον χαρακτηρισμό άγνωστων συστημάτων. Αυτή τη στιγμή περιλαμβάνει 7 κατηγορίες, 56 υπεροικογένειες, 937 οικογένειες, 9098 συστήματα, 11806 πρωτεΐνες και 12086 βιβλιογραφικές αναφορές. Τα συστήματα έχουν καταχωρηθεί με την ημερομηνία δημοσίευσης, ενώ όλες οι καταχωρήσεις επιμελούνται και σχολιάζονται από ειδικούς. Το

σύστημα TC σχεδιάστηκε με βάση το EC (Enzyme Commission) και είναι παρόμοιο με αυτό (Bairoch, 1999), με την διαφορά ότι βασίζεται τόσο στη λειτουργία (κατηγορία και υποκατηγορία) όσο και τη φυλογένεση (οικογένεια, υπο-οικογένεια και υπερ-οικογένεια). Είναι το μόνο σύστημα εγκεκριμένο από την Διεθνή Ένωση Βιοχημείας και Μοριακής Βιολογίας (International Union of Biochemistry and Molecular Biology, IUBMB) που χρησιμοποιείται σήμερα για τα διαμεμβρανικά μοριακά συστηματα μεταφοράς (Saier, 2000). Το σύστημα διαχείρισης της TCDB έχει πολλά πλεονεκτήματα: Η γνώση είναι ιεραρχικά δομημένη, υπάρχει πρόβλεψη για διαχείριση αρχείων ασφαλείας και ανάκτηση αυτών, ενώ έχουν αναπτυχθεί και ειδικές εφαρμογές βιοπληροφορικής πάνω στο σύστημα ταξινόμησης όπως το TC-BLAST καθώς και λογισμικό για την ανίχνευση μακρινών φυλογενετικών σχέσεων, βάσει της Υπεροικογένειας. Τέλος, γίνονται προσπάθειες για ενοποίηση της πληροφορίας με άλλες βάσεις όπως η PFAM και η OMPdb.

OMPdb: Η βάση δεδομένων OMPdb (Tsirigos, Bagos, & Hamodrakas, 2011), διατίθεται στην ιστοσελίδα http://www.ompdb.org, είναι διαθέσιμη στο κοινό και περιέχει διαμεμβρανικά β-βαρελια της εξωτερικής μεμβράνης των κατά Gram αρνητικών βακτηρίων. Παρουσιάστηκε για πρώτη φορά το 2011 και περιείχε περίπου 70.000 εγγραφές. Μέσα στα επόμενα 3 χρόνια, περιελάμβανε περισσότερες από 500.000 καταχωρήσεις. Όλες οι πρωτεΐνες της OMPdb ταξινομούνται σε 91 οικογένειες, βάσει δομικών και λειτουργικών κριτηρίων. Κάθε οικογένεια χαρακτηρίζεται από διαφορετικό προφίλ Hidden Markov Model (pHMM), που την διαχωρίζει από τις υπόλοιπες. Οι περισσότερες από αυτές τις οικογένειες είχαν ήδη αναφερθεί στη βάση Pfam (Finn, et al., 2014), εκτενής όμως βιβλιογραφική έρευνα επέτρεψε την αναγνώριση όχι μόνο οικογενειών που δεν υπήρχαν στην αντίστοιχη Pfam clan (MBB clan - CL0193), αλλά και κάποιων που αναγνωρίζονταν ως περιοχές άγνωστης λειτουργίας (Domains of Unknown function - DUFs). Επιπλέον, συνολικά 15 οικογένειες, έλειπαν από την Pfam ή είχαν χαρακτηριστεί με αυτόματο τρόπο στην Pfam-B. Για κάθε πρωτεΐνη, ο χρήστης μπορεί να ανακτήσει πληροφορίες σχετικά με την παρουσία των σηματοδοτικών αλληλουχιών και τον σχολιασμό των διαμεμβρανικών τμημάτων. Για κάθε εγγραφή οικογένειας, και εφόσον είναι διαθέσιμη, παρατίθεται λίστα πρωτεϊνών που έχουν κρυσταλλογραφικά προσδιορισμένη δομή. Ο χαρακτηρισμός των πρωτεϊνών βάσει του προφίλ pHMM και η υιοθέτηση του συστήματος ταξινόμησης της Pfam, επιτρέπει στους επιμελητές να ακολουθήσουν ένα ημι-αυτόματο σύστημα ανάκτησης δεδομένων. Αρχικά, μια οικογένεια αναγνωρίζεται μέσω της βιβλιογραφικής αναζήτησης, στη συνέχεια δημιουργούνται μοντέλα pHMM και συγκρίνονται έναντι της βάσης Pfam, και τέλος, προσδιορίζονται τα μέλη της οικογένειας και αποθηκεύονται στη βάση δεδομένων. Το σύστημα αυτό παρουσιάζει έναν πιο πλήρη και ακριβή σχολιασμό των πρωτεϊνών δομής β-βαρελιού, λόγω της πρόσθετης αξίας του χειρωνακτικού σχολιασμού και των λεπτομερών βιβλιογραφικών αναφορών. Από την άλλη πλευρά, η σύγκριση της OMPdb με τις άλλες εξειδικευμένες βάσεων δεδομένων που περιέχουν πρωτεΐνες δομής β-βαρελιού, αποκαλύπτει ότι υπερέχει από όλες τις πλευρές, διότι διαθέτει το μεγαλύτερο αριθμό εγγραφών, πρωτεϊνών και οικογενειών. Διαθέτει τα πιο πλήρη και αποκλειστικά δεδομένα τα διαμεμβρανικά β-βαρελια, και προσφέρει την πιο ολοκληρωμένη διασύνδεση με άλλες δημόσιες βάσεις δεδομένων, βιβλιογραφικές αναφορές, εργαλεία πρόβλεψης και σχολιασμού αλληλουχιών. Η OMPdb συνεργάζεται με τους επιμελητές των βάσεων TCDB και Pfam (και οι δύο βάσεις δεδομένων περιέχουν τις οικογένειες των πρωτεϊνών δομής β-βαρελιού εξωτερικής μεμβράνης των κατά Gram αρνητικών βακτηρίων), προκειμένου να επιτύχει τη ενοποίηση των βάσεων δεδομένων με τη διασύνδεση των οικογενειών και τη διατήρηση των πληροφοριών ενημερωμένων (ανταλλαγή σχολιασμού, αναφορές κ.τ.λ.). Η διαδικτυακή εφαρμογή βασίζεται στο συνδυασμό δύο επιπέδων. Το βασικό επίπεδο είναι ένα σύστημα βάσης δεδομένων MySQL, και το δεύτερο επίπεδο είναι ένας διακομιστής εφαρμογών Apache-PHP που λαμβάνει τις αναζητήσεις των χρηστών. Παρόλο που η ιεραρχία της βάσης δεδομένων είναι μάλλον απλή (δηλαδή υπάρχει μόνο ένα επίπεδο, η οικογένεια), η βάση αποθηκεύεται σε MySQL, ώστε να διευκολυνθεί η διαδικασία των εξειδικευμένων ερωτημάτων και να γίνεται πιο εύκολη η ενημέρωση της βάσης. Η διεπαφή ιστού της OMPdb προσφέρει στο χρήστη τη δυνατότητα όχι μόνο να δει τα διαθέσιμα δεδομένα, αλλά και να υποβάλει εξειδικευμένες αναζητήσεις για την αναζήτηση ανάμεσα στις εγγραφές των πρωτεϊνών της βάσης. Η ύπαρξη ενός τόσο μεγάλου και αξιόπιστου συνόλου δεδομένων διαμεμβρανικών β-βαρελιών μπορούν να χρησιμοποιηθούν για αναλύσεις μεγάλης κλίμακας σχετικά με την ακρίβεια ταξινόμησης των υφιστάμενων προγνωστικών αλγορίθμων, για την δημιουργία νέων μεθόδων πρόβλεψης και για μελέτες μοντελοποίησης. Μακροπρόθεσμο στόχο αποτελεί η διατήρηση της OMPdb όσο το δυνατόν πιο ενημερωμένη, ακολουθώντας τις τακτικές ενημερώσεις της UniProt και κάνοντας ανασκόπηση της βιβλιογραφίας για νέες πειραματικά επαληθευμένες πρωτεΐνες δομής β-βαρελιού, προκειμένου να συμπεριληφθούν στη βάση ή να ενταχθούν σε νέες οικογένειες. Παρόμοια με άλλες βάσεις δεδομένων, η OMPdb βρίσκεται υπό εξέλιξη, και η αλληλεπίδρασή της με την κοινότητα των

χρηστών είναι ζωτικής σημασίας για την ανάπτυξη και την τελειοποίηση της. Εκτός από τη συνεργασία με τις υπόλοιπες σχετικές βάσεις δεδομένων που αναφέρθηκαν παραπάνω (Pfam και TCDB), οι διαχειριστές ενθαρρύνουν τους χρήστες να υποβάλουν στοιχεία, να διορθώσουν πιθανά λάθη, και να διατυπώσουν προτάσεις ώστε η OMPdb να αποκτήσει μεγαλύτερη χρησιμότητα για την επιστημονική κοινότητα.

CAZy: Η βάση δεδομένων CAZy (www.cazy.org) περιγράφει οικογένειες παρόμοιας καταλυτικής δομής και περιοχής πρόσδεσης υδατανθράκων (carbohydrate-binding modules) των ενζύμων, που διασπούν, τροποποιούν, ή δημιουργούν γλυκοσιδικούς δεσμούς (Cantarel et al., 2009; Lombard, Ramulu, Drula, Coutinho, & Henrissat, 2014). Η βάση δεδομένων CAZy δημοσιεύθηκε το 1991, πριν από οποιαδήποτε αλληλούχιση γονιδιώματος (Henrissat, 1991) και περιλαμβάνει την ταξινόμηση της οικογένειας αλληλουχιών των γλυκοζιδικών υδρολασών. Ξεκίνησε στις αρχές της δεκαετίας του '90 και επεκτάθηκε και σε άλλες κατηγορίες ενζύμων ενεργών υδατανθράκων όπως οι γλυκοζυλοτρανσφεράσες (Campbell, Davies, Bulone, & Henrissat, 1997). Έγινε διαθέσιμη αρχικά μέσω μιας απλής ιστοσελίδας τον Σεπτέμβριο του 1998, ενώ μετατράπηκε σε ολοκληρωμένη βάση δεδομένων τύπου SQL (το 1999), ώστε να είναι πιο εύκολη η διαχείριση τους και να βελτιωθεί ο ρυθμός συλλογής τους. Παρά την ταχεία αύξηση των δεδομένων, κάθε αλληλουχία που εμφανίζεται στην CAZy συνεχίζει να ελέγχεται από κάποιον επιμελητή, εκτός εάν η νέα αλληλουχία είναι σε συμφωνία, χωρίς κανένα κενό και με περισσότερο από 50% ταύτιση με μια ήδη ταξινομημένη αλληλουχία. Ο ανθρώπινος παράγοντας στην επιμέλεια της βάσης, που περιλαμβάνει διορθώσεις σφαλμάτων μετά από αίτημα κάποιου χρήστη, καθώς και η απόδοση αριθμών EC αποκλειστικά σε ένζυμα που έχουν χαρακτηριστεί πειραματικά χωρίς μεταφορά σχολιασμού λόγω ομοιότητας αλληλουχίας, καθιέρωσε την CAZy ως πηγή αναφοράς για τις γλυκο-επιστήμες. Ωστόσο, θα ήταν χρήσιμο, μια τέτοια βάση δεδομένων να είναι συμπληρωμένη με μια εγκυκλοπαιδική πηγή που θα είναι σε θέση να παρέχει στους ερευνητές ακριβή επισκόπηση της γνώσης για κάθε οικογένεια. Αυτή η διαπίστωση ήταν το βασικό κίνητρο για την ανάπτυξη της συμπληρωματικής ιστοσελίδας CAZypedia (http://www.cazypedia.org), που αποτελεί την λογική επέκταση της βάσης δεδομένων CAZy. Υπεύθυνος της CAZypedia είναι ο καθηγητής Harry Brumer του πανεπιστημίου British Columbia ενώ υποστηρίζεται από μια επιτροπή έμπειρων επιμελητών από όλο τον κόσμο, οι οποίοι επιζητούν υπεύθυνους επιμελητές και εξειδικευμένους συνεργάτες για να σχολιάζουν τα δεδομένα πετυχαίνοντας έτσι την συμμετοχή όλων των επιστημόνων που έχουν ως πεδίο έρευνας τις γλυκοεπιστήμες. Οι επιστήμονες αυτοί, τηρούν τις συμβάσεις ονοματοδοσίας που διέπουν το σύστημα ταξινόμησης CAZy. Κατά συνέπεια, συχνά αυτοί που ανακαλύπτουν μια νέα οικογένεια CAZymes, πριν από τη δημοσίευση της έρευνάς τους, ζητούν από τη βάση δεδομένων CAZy τον αριθμό της οικογένειας, ώστε να μπορούν να τον χρησιμοποιήσουν στην δημοσίευση. Ομοίως, όταν ανακαλυφθεί μια νέα δραστηριότητα σε μια υπάρχουσα οικογένεια, πολλοί από τους επιστήμονες ενημερώνουν τη βάση, προκειμένου να συμπληρωθεί η λειτουργική αυτή πληροφορία στην CAZy. Παρά τις προσπάθειες που έχουν γίνει μέχρι τώρα, οι πειραματικές πληροφορίες που παρουσιάζονται στην CAZy είναι αναγκαστικά ελλιπής. Οι ερευνητές μπορούν να βοηθήσουν επισημαίνοντας δεδομένα υποστρώματος/προϊόντων που έχουν δημοσιευθεί αλλά δεν έχουν ακόμα καταχωρηθεί στην CAZy. Επειδή η σύγχρονη βιοχημεία σταδιακά δημιουργεί πολύ μεγάλα σύνολα δεδομένων με τις ενεργότητες να αναφέρονται στα δημοσιευμένα άρθρα σε δεκάδες (και σύντομα χιλιάδες) ενζύμων, θα ήταν μεγάλο πλεονέκτημα αν οι ερευνητές κατέθεταν τα δεδομένα που χρησιμοποίησαν ως συμπληρωματικό υλικό σε μορφή πίνακα, που θα περιελάμβανε τη σειρά καταχώρησης στη βάση δεδομένων για κάθε χαρακτηρισμένο ένζυμο, τα υποστρώματα που χρησιμοποιήθηκαν και τα προϊόντα που ανιχνεύθηκαν. Αν τα επιστημονικά περιοδικά κάνουν υποχρεωτικό αυτόν τον απλό τρόπο καταχώρησης των δεδομένων, θα είναι δυνατή, προς όφελος όλων, μια πιο ολοκληρωμένη και αξιόπιστη συλλογή δεδομένων για τη βάση. Η πρακτική αυτή θα διευκόλυνε την λειτουργική συλλογή δεδομένων και σε άλλες βάσεις δεδομένων εκτός της CAZy.

MEROPS: Η βάση Merops (http://merops.sanger.ac.uk) αποτελεί μια βάση ταξινόμησης και ονοματολογίας πρωτεολυτικών ενζύμων και των πρωτεϊνών και μικρών μορίων αναστολέων που επηρεάζουν την ενζυματική τους δράση (Rawlings, Waller, Barrett, & Bateman, 2014). Τα πρωτεολυτικά ένζυμα έχουν πολλές βιολογικές λειτουργίες, που περιλαμβάνουν την πέψη των πρωτεϊνών, την ανακύκλωση των πρωτεϊνών, την επεξεργασία και μετατόπιση των νεοσυντιθέμενων πρωτεϊνών, την αφαίρεση των σηματοδοτικών αλληλουχιών στόχευσης, την ενεργοποίηση (και απενεργοποίηση) των ενζύμων, τις πεπτιδικές ορμόνες, τους υποδοχείς κυτταρικής επιφάνειας και τους νευροδιαβιβαστές, την αναδιαμόρφωση στις εξωκυττάριες μήτρες, την πήξη του αίματος και την ινωδόλυση. Οι πεπτιδάσες εμπλέκονται σε ευρύ φάσμα ασθενειών (ιική λοίμωξη, εισβολή παράσιτων, καρκίνο, διαβήτη τύπου ΙΙ, οστεοαρθρίτιδα και στη νόσο Alzheimer), και

χρησιμοποιούνται συχνά στη βιομηχανία (βιολογικά απορυπαντικά, βυρσοδεψία, παρασκευή τυριών και σάλτσας σόγιας, για εργαστηριακή χρήση στην πρωτεομική φασματοσκοπία μάζας, για προσδιορισμό αλληλουχίας πρωτεϊνών κ.ο.κ.). Η βάση δεδομένων δημιουργήθηκε το 1996 και σήμερα περιλαμβάνει πάνω από 400.000 αλληλουχίες πεπτιδασών. Οι αλληλουχίες που έχουν/μοιράζονται παρόμοια πρωτεϊνική αναδίπλωση οργανώνονται σε μια υπεροικογένεια (clan). Οι αλληλουχίες που έχουν ομοιότητες στην περιοχή της πεπτιδάσης οργανώνονται σε οικογένειες. Συνολικά υπάρχουν 61 υπεροικογένειες, 251 οικογένειες και 4.236 αναγνωριστικά (εκ των οποίων μόνο τα 377 περιλαμβάνονται στο *Enzyme Nomenclature*). Η συλλογή των δεδομένων έχει επεκταθεί για να συμπεριλάβει πάνω από 28.000 αναστολείς πεπτιδάσης που προέρχονται από γονιδιακά προϊόντα, καθώς και πάνω από 1.200 μικρά μόρια αναστολείς. Στη βάση περιλαμβάνονται αναφορές από πάνω από 53.000 δημοσιεύσεις και συνεργάζεται με διάφορες άλλες βάσεις δεδομένων αλληλουχίας πρωτεϊνών, συμπεριλαμβανομένων των UniProt, Pfam και Interpro. Όπως συμβαίνει και σε άλλες εξειδικευμένες βάσεις πρωτεϊνών, υπάρχει επίγνωση των λαθών στα πρωτογενή δεδομένα που διαρρέουν σε άλλες βάσεις. Είναι σχεδόν αδύνατο να διορθωθούν σφάλματα στις βάσεις πρωτογενών αλληλουχιών χωρίς τη συγκατάθεση των ατόμων που τις καταχώρησαν. Οι σχολιαστές των γονιδιωμάτων θα πρέπει να γνωρίζουν ότι τα ένζυμα, και ιδιαίτερα οι πεπτιδάσες, για να μπορούν να χρησιμοποιηθούν, πρέπει να συνοδεύονται από την πλήρη γνώση των κατάλοιπων του ενεργού κέντρου, και ότι αν σε κάποια καταχώρηση αλληλουχίας λείπει οποιοδήποτε από αυτά δεν θα πρέπει να σχολιάζεται ως ενεργό ένζυμο. Ο μεγάλος και διαρκώς αυξανόμενος όγκος δημοσιευμένων δεδομένων, καθιστά απαραίτητη τη συμμετοχή όλων των μελών της επιστημονικής κοινότητας στο σχολιασμό. Τα οφέλη για τον ερευνητή που συμβάλλει σε μία βιολογική βάση δεδομένων είναι: η αναγνώριση της συνεισφοράς του, η προβολή των δημοσιεύσεών του, η βελτίωση των συλλογών δεδομένων και η διόρθωση λαθών, ενώ επίσης βοηθά άλλους ερευνητές που χρησιμοποιούν τα δεδομένα.

neXtProt: Η neXtProt (http://www.nextprot.org/) είναι μια διαδικτυακή βάση δεδομένων πρωτεϊνών του ανθρώπινου οργανισμού (Lane et al., 2012). Ενσωματώνει πληροφορίες που προέρχονται από την UniProtKB/Swiss-Prot με μια πληθώρα άλλων στοιχείων που προέρχονται από τα αποθετήρια και βάσεις δεδομένων που περιέχουν αποτελέσματα πειραμάτων υψηλής απόδοσης στον τομέα της πρωτεομικής, μεταγραφομικής και γονιδιωματικής. Υπάρχουν μια σειρά από δυσκολίες για τη διατήρηση παρόμοιων πηγών με την neXtProt. Η πρώτη έγκειται στην επιλογή και αξιολόγηση της ποιότητας των πληροφοριών που καταχωρούνται στη βάση. Ένας από τους στόχους της ομάδας επιμέλειας της neXtProt είναι η ταξινόμηση των πειραματικών αποτελέσματων σε τρεις κατηγορίες: «χάλκινο» (> 5% ποσοστό σφάλματος), «αργυρό» (1-5% ποσοστό σφάλματος) και «χρυσό» (λιγότερο από 1% ποσοστό σφάλματος). Τα χάλκινα δεδομένα δεν ενσωματώνονται στην neXtProt. Η αξιολόγηση της ποιότητας των πειραματικών αποτελεσμάτων δεν είναι γενικά πολύ εύκολο να επιτευχθεί, δεδομένου ότι συχνά η καταχώρηση μελετών ή δεδομένων στα repositories δεν παρέχουν τα απαραίτητα κριτήρια για να αξιολογηθεί αντικειμενικά η ποιότητα της πειραματικής διάταξης αλλά ούτε και των αποτελεσμάτων. Στην ιδανική περίπτωση, οι εκτιμήσεις αυτές θα πρέπει να επανεξετάζονται σε τακτά χρονικά διαστήματα, όταν οι τεχνικές αλλάζουν και έχουν αντικατασταθεί από καλύτερες και πιο ακριβείς μεθόδους. Μια άλλη σημαντική πρόκληση για πηγές παρόμοιες με την neXtProt που προσπαθούν να ενσωματώσουν μεγάλη ποικιλία πληροφοριών που προέρχονται από πολλές ετερογενείς πηγές είναι η συνεχής ανάγκη τροποποίησης και ενημέρωσης της πληροφορίας που παρέχεται από τη βάση δεδομένων. Η neXtProt προσπαθεί να ακολουθήσει ένα μηνιαίο χρονοδιάγραμμα έκδοσης δημοσιεύσεων αλλά αυτό μερικές φορές διαταράσσεται από αλλαγές σε τουλάχιστον μία από τις ενσωματωμένες πηγές. Σημαντικό πρόβλημα δημιουργούν αλλαγές στην μορφή των δεδομένων. Τέλος, σημαντικό πρόβλημα αποτελεί ότι οι πηγές που ενσωματώνονται στην βάση δεν έχουν όλες την ίδια προτυποποίηση. Η παραγωγή και η διατήρηση πινάκων αντιστοίχισης μεταξύ διαφορετικών οντολογιών ή ελεγχόμενων λεξιλόγιων είναι γενικά απαραίτητη. Αυτό το πρόβλημα είναι ιδιαίτερα οξύ στην περίπτωση των ανθρώπινων ασθενειών καθώς υπάρχουν πάνω από 10 οντολογίες που χρησιμοποιούνται από τις επιστημονικές κοινότητες της ιατρικής και των επιστημών ζωής.

PASS2: Η βάση PASS2 περιέχει στοιχίσεις δομών πρωτεϊνικών αλληλουχιών σε επίπεδο υπερ-οικογενειών (Protein sequence Alignments of Structural Superfamilies). Η πρώτη έκδοση της βάσης αναφέρεται ως «CAMPASS» (Sowdhamini et al., 1998). Η πολλαπλή στοίχιση αλληλουχιών μελών μιας υπερ-οικογένειας πρωτεϊνών που διαφέρουν μεταξύ τους, αποτελεί δύσκολη διαδικασία εξαιτίας την μικρής ομοιότητας των αλληλουχιών, παρόλο που μπορεί να υπάρχουν αναμφισβήτητες εξελικτικές συνδέσεις, λειτουργικές και δομικές ομοιότητες. Οι προηγούμενες εκδόσεις της PASS2 ήταν σε μορφή HTML, ενώ η τρέχουσα έκδοση

της (PASS2.4) (Gandhimathi, Nair, & Sowdhamini, 2012), λειτουργεί σε μια πλατφόρμα MYSQL με διεπαφή σε PHP. Αυτή η έκδοση της βάσης, η οποία είναι σε άμεση αντιστοιχία με την SCOP 1.75 (Murzin, Brenner, Hubbard, & Chothia, 1995) για τον ορισμό των μελών της υπερ-οικογένειας, αυτή τη στιγμή προσφέρει στοιχίσεις αλληλουχιών βάσει της δομής 1961 υπεροικογενειών. Σημαντική πρόκληση για τη διατήρηση και την ενημέρωση της βάσης, λαμβάνοντας υπόψη τη συνεχή συσσώρευση πρόσθετων μελών και υπερ-οικογενειών, είναι η μείωση της χειρωνακτικής παρέμβασης, και η αυτοματοποίηση όσο το δυνατόν περισσότερο της διαδικασίας, διατηρώντας όμως την ποιότητα των δεδομένων σε υψηλό επίπεδο. Αυτό είναι πράγματι δύσκολο, δεδομένου ότι η εξέλιξη φέρνει μαζί της διαφοροποιήσεις, και αυτό σημαίνει ότι θα μπορούσαν να υπάρχουν αρκετά «outliers» (Gandhimathi, et al., 2012), τα οποία είναι δύσκολο να εντοπιστούν κατά τη διάρκεια της αυτόματης στοίχισης των πρωτεϊνικών αυτοτελών δομικών περιοχών των υπερ-οικογενειών. Η μελέτη των λειτουργικών αποκλίσεων των καταλοίπων και της ειδικής κατηγορίας της φύσης των διατηρημένων κατάλοιπων ή μοτίβων από τους πειραματιστές σε ένα ελεγχόμενο λεξιλόγιο στις αναφορές προσδιορισμού της δομής τους, θα μπορούσε να καταστήσει δυνατή την έγκαιρη αναγνώριση των outliers.

KinG: Η βάση δεδομένων Kinases in Genomes (KinG) αποτελεί μία πηγή κινασών Ser/Thr/Tyr που κωδικοποιούνται στα πλήρως αλληλουχημένα γονιδιώματα των προκαρυωτικών, ιικών και ευκαρυωτικών κυττάρων (Krupa, Abhinandan, & Srinivasan, 2004). Το πλήρες ρεπερτόριο των κινασών σε διάφορα πλήρως αλληλουχημένα γονιδιώματα παρουσιάζεται στο δίκτυο Garuda India στην ιστοσελίδα http://megha.garudaindia.in/king/. Το δίκτυο παρέχει λεπτομερή κατάλογο των Ser/Thr/Tyr και άτυπων κινασών πρωτεϊνών διάφορων οργανισμών συνοδευόμενα από χαρακτηριστικά, όπως η ταξινόμηση σε υπο-οικογένειες πρωτεϊνικών κινασών και η οργάνωση των αυτοτελών δομικών περιοχών. Η βάση επιτρέπει επίσης την ανάκτηση των κινασών πρωτεϊνών που ανήκουν σε καθορισμένη υπο-οικογένεια ή σε συγκεκριμένους συνδυασμούς αυτοτελών δομικών περιοχών. Ο χρήστης μπορεί αναζητήσει συγκεκριμένες αλληλουχίες ώστε να προσδιορίσει την καταλυτική περιοχή της κινάσης και τα διάφορα λειτουργικά κατάλοιπα στην καταλυτική περιοχή. Στην πρώτη έκδοση της KinG που δημοσιεύθηκε το 2004 (Krupa, et al., 2004), δημοσιεύθηκαν κινάσες μόνο από 40 οργανισμούς. Η KinG ανανεώνεται κάθε χρόνο. Οι Κινάσες εκφράζονται έντονα, ειδικά στους ευκαρυωτικούς οργανισμούς. Επιπλέον, καθώς ο αριθμός των πλήρως αλληλουχημένων γονιδιωμάτων αυξάνεται με γρήγορο ρυθμό, σε κάθε ανανέωση της βάσης αυξάνεται και ο αριθμός των κινασών που πρέπει να διαχειριστούν. Στην τρέχουσα έκδοση της KinG μελετώνται 12200 ομάδες δεδομένων γονιδιωματικής με αποτέλεσμα τον εντοπισμό και την ταξινόμηση 131.921 κινασών. Εκτός από το ότι πρέπει η βάση να συμβαδίζει με την αύξηση του αριθμού των κινασών, υπάρχουν μερικές επιπλέον ενδιαφέρουσες προκλήσεις που πρέπει να αντιμετωπιστούν. Η ταξινόμηση των κινασών σε υπο-οικογένειες στην βάση KinG πραγματοποιείται σύμφωνα με το σύστημα ταξινόμησης των Hanks και Hunter's (Hanks & Hunter, 1995) προσαρμοσμένο με μια προσέγγιση πολλαπλών ειδικών ανά θέση πινάκων (multiple position-specific scoring matrices - PSSM) (Gowri, Krishnadev, Swamy, & Srinivasan, 2006). Η ομαδοποίηση των αλληλουχιών κινάσης σε αυτές τις υπο-οικογένειες οδηγεί σε αναγνώριση γνήσιων υπο-οικογενειών που δεν περιέχονται στο αρχικό πλαίσιο ταξινόμησης. Ως εκ τούτου, τα συστήματα ταξινόμησης αναδιοργανώνονται, προκειμένου να συμπεριληφθούν όσο το δυνατόν περισσότερες κινάσες. Μια άλλη πρόκληση είναι η ασυμφωνία μεταξύ της ταξινόμησης σε υπο-οικογένειες, η οποία βασίζεται αποκλειστικά στις αλληλουχίες των καταλυτικών περιοχών, και τους συνδυασμούς των περιοχών των κινασών. Πρόσφατες αναλύσεις (Deshmukh, Anamika, & Srinivasan, 2010; Rakshambikai, Gnanavel, & Srinivasan, 2014) επέτρεψαν την αναγνώριση της εμφάνισης υβριδικών (hybrid) κινασών οι οποίες χαρακτηρίζονται από μια υπο-οικογένεια κινάσης, που αναγνωρίζεται με βάση μόνο την αλληλουχία των καταλυτικών περιοχών, και η οποία χαρακτηρίζεται σε μια άλλη υπο-οικογένεια κινασών με βάση την αρχιτεκτονική της καταλυτικής περιοχής. Αυτή η περίπλοκη κατάσταση δεν επιτρέπει την ταξινόμηση της κινάσης σε καμία από τις δύο υπο-οικογένειες, και ως εκ τούτου προτείνεται να ταξινομούνται ως υβριδικές κινάσες με χαρακτηριστικά των δύο διαφορετικών υπο-οικογενειών κινάσης. Επιπλέον, δυσκολία στην ταξινόμηση προκαλείται από την εμφάνιση κινασών που η καταλυτική περιοχή τους συνδέεται με μία συγκεκριμένη οικογένεια κινασών αλλά η συν-ύπαρξη των περιοχών δεν χαρακτηρίζεται από κάποια υπο-οικογένεια. Γι' αυτό τα χαρακτηριστικά τους αποκλίνουν από αυτά των υπόλοιπων μελών της υπο-οικογένειας. Αυτές οι κινάσες ονομάζονται κινάσες rogue (Deshmukh, et al., 2010; Rakshambikai, et al., 2014). Ο σχεδιασμός του προτεινόμενου συστήματος ταξινόμησης των κινασών διευκολύνεται από την ανάλυση δεδομένων υψηλής απόδοσης που προκύπτουν συνεχώς. Τα δεδομένα και το σύστημα ταξινόμησης τροφοδοτούν το ένα το άλλο. Καθώς βελτιώνεται συνεχώς το σύστημα ταξινόμησης (Bhaskara et al., 2014; Gnanavel et al., 2014; Martin, Anamika, &

Srinivasan, 2010), έτσι θα συνεχίζεται και η ενημέρωση της βάσης δεδομένων KinG. Η παρούσα έκδοση της KinG έχει δημιουργηθεί χρησιμοποιώντας το NetBeans IDE σε πυρήνα Java, JSP, Servlets, AJAX, Jquery, XML, HTML και CSS ενώ το περιβάλλον της είναι φιλικό προς το χρήστη, και οι αναζητήσεις πραγματοποιούνται γρήγορα.

EzCatDB: Η βάση EzCatDB (http://ezcatdb.cbrc.jp/EzCatDB/) δημιουργήθηκε το 2004 με στόχο να αποτελέσει έναν οδηγό κατάταξης ενζυμικών αντιδράσεων, των δομών των ενεργών κέντρων των ενζύμων, αλλά και των καταλυτικών τους μηχανισμών. Βασίζεται σε πληροφορίες από την βιβλιογραφία (Nagano, 2005; Nagano et al., 2014) και διαφέρει από την Enzyme Commission (E.C.) (NC-IUBMB; http://www.chem.qmul.ac.uk/iubmb/enzyme/) η οποία ταξινομεί τα ένζυμα με βάση τις χημικές δομές των υποστρωμάτων και των προϊόντων (Fleischmann et al., 2004; McDonald, Boyce, & Tipton, 2009; Tipton, 1994). Αν και η ταξινόμηση της ιεραρχικής αντίδρασης (RLCP) στην EzCatDB αρχικά περιελάμβανε μόνο αντιδράσεις των πυρηνόφιλων υποκαταστάσεων (nucleophilic substitution reactions), όπως υδρολύσεις και αντιδράσεις μεταφοράς, στη συνέχεια επεκτάθηκε και σε άλλες αντιδράσεις όπως προσθήκης, αφαίρεσης, ισομερισμού, μεταφοράς υδριδίου και μεταφοράς ηλεκτρονίων (Nagano, et al., 2014).

Η EzCatDB περιλαμβάνει τις τριτοταγείς δομές των ενζύμων της Protein Data Bank (PDB) (Rose et al., 2013) και τα αντίστοιχα δεδομένα αλληλουχίας αμινοξέων της UniProt, ιδιαίτερα με την αντίστοιχη ταξινόμηση CATH (Cuff et al., 2011). Εκτός από αυτές τις βάσεις, για τα δεδομένα ενώσεων που σχετίζονται με τα ένζυμα λαμβάνεται υπόψη και η βάση KEGG (Kanehisa, et al., 2014). Στην EzCatDB χρησιμοποιείται το σύστημα διαχείρισης βάσεων δεδομένων PostgreSQL. Συνεπώς η αναζήτηση των δεδομένων ενός ενζύμου μπορεί να γίνει με διάφορους τρόπους (Nagano, 2005) όπως για παράδειγμα χρησιμοποιώντας τον αριθμό E.C., τα IDs από άλλες βάσεις δεδομένων, τους τύπους των αμινοξέων που βρίσκονται στο ενεργό κέντρο των ενζύμων και τους τύπους των προσδετών που μπορούν να συνδυαστούν για την αναζήτηση (Nagano, 2005). Επιπλέον, για κάθε εγγραφή είναι δυνατή η δημιουργία ενός πίνακα σχολιασμού των προσδετών για τα δεδομένα της PDB, στον οποίο τα μόρια προσδέτη που συνδέονται με τις δομές του ενζύμου έχουν περιγραφεί χειροκίνητα ως συμπαράγοντες, υποστρώματα, προϊόντα ή ενδιάμεσοι, φυσικοί προσδέτες ή ανάλογοι προσδέτες. Επίσης δημιουργείται ένας πίνακας με πληροφορίες σχετικά με τα αμινοξικά κατάλοιπα που βρίσκονται στο ενεργό κέντρο του ενζύμου (Nagano, 2005). Τα δεδομένα αυτά είναι απαραίτητα για την κατανόηση των καταλυτικών μηχανισμών του ενζύμου.

Όλες οι διαδικασίες που σχετίζονται με τον χειρωνακτικό σχολιασμό είναι χρονοβόρες, λόγω του μεγάλου όγκου των δεδομένων και της δυσκολίας αναζήτησης στη βιβλιογραφία. Συγκεκριμένα, η εξαγωγή και ανάλυση πληροφορίας από την βιβλιογραφία είναι η πιο χρονοβόρα και απαιτεί τοπική αποθήκευση της λίστας των δημοσιεύσεων, παραγγελία του πλήρους κειμένου, αναζήτηση των λέξεων κλειδιών κλπ. Η EzCatDB περιέχει σήμερα 871 εγγραφές ένζυμων, που αφορούν 1.610 αλληλουχίες της UniProtKB και 6.704 εγγραφές της PDB. Είναι επομένως φανερό ότι τα διαθέσιμα δεδομένα ενζύμων είναι περιορισμένα. Επιπλέον βρίσκονται στο στάδιο της επεξεργασίας 300 εγγραφές, κάτι όμως που αποτελεί δύσκολο έργο λόγω των περιορισμών στο ανθρώπινο δυναμικό και στη χρηματοδότηση.

MACiE: Η MACiE (Mechanism, Annotation and Classification in Enzymes, http://www.ebi.ac.uk/thornton-srv/databases/MACiE), είναι μια βάση δεδομένων που περιέχει μηχανισμούς αντίδρασης ενζύμων (Holliday et al., 2012). Στην MACiE συγκεντρώνονται και αποθηκεύονται πληροφορίες σχετικά με τα ένζυμα, τους συνολικούς χημικούς μετασχηματισμούς τους, τους μηχανισμούς αντίδρασης, τους συμπαράγοντες και τα καταλυτικά κατάλοιπα. Κάθε εγγραφή της MACiE αντιστοιχίζεται σε τουλάχιστον μια κρυσταλλική δομή στην PDB και σε έναν καλά καθορισμένο μηχανισμό από την πρωτογενή βιβλιογραφία. Τα δεδομένα της MACiE μπορούν να θεωρηθούν ως μια εννοιολογική ιεράρχηση, η οποία προκύπτει από το γεγονός ότι ένα ένζυμο μπορεί να οριστεί σαφώς από τα στοιχεία του.

Με αυτήν την ιεράρχηση ένα ένζυμο μπορεί να οριστεί κατά την πιο απλή του μορφή, σαν ένα βιοπολυμερές που έχει μια πρωτοταγή αμινοξική αλληλουχία και καταλύει έναν συνολικό χημικό μετασχηματισμό (ο ορισμός αυτός δεν περιλαμβάνει τα ριβοένζυμα). Ένας συνολικός χημικός μετασχηματισμός πρέπει να αποτελείται από τουλάχιστον ένα υπόστρωμα και ένα προϊόν, και να έχει ένα μηχανισμό που όμως είναι πιθανό να μην γνωστός με κάθε λεπτομέρεια. Το γεγονός ότι τα δεδομένα αυτά μπορούν να διαταχθούν ιεραρχικά αναδεικνύει την σχέση που υπάρχει μεταξύ τους καθώς και ότι είναι δυνατή η περιγραφή τους σε μια σχεσιακή βάση δεδομένων. Η MACiE χρησιμοποιεί την ανοιχτή βάση δεδομένων MySQL. Αυτό το επίπεδο των σχεσιακών πληροφοριών επιτρέπει την γρήγορη εκτέλεση

σύνθετων αναζητήσεων. Τέτοιες ιεραρχικές αναζητήσεις έχουν ήδη υλοποιηθεί πολλές φορές στην ιστοσελίδα της MACiE (Holliday et al., 2007).

Με τη χρήση των σχέσεων αυτών είναι εύκολο να περάσουμε από το ένα δεδομένο στο άλλο, με την προϋπόθεση ότι και τα δύο είναι στοιχεία που απαιτούνται για τον καθορισμό του ενζύμου. Η MACiE περιέχει μεγάλο αριθμό μετα-δεδομένων που μπορούν να συνδεθούν με τα ένζυμα και τους μηχανισμούς αντίδρασης τους. Τα βασικά στοιχεία περιλαμβάνουν (1) τις λεπτομερείς λειτουργίες κάθε καταλυτικού αμινοξικού κατάλοιπου στην θέση κατάλυσης, (2) την παρουσία καταλυτικών δυάδων ή τριάδων, (3), τη μηχανιστική περιγραφή του κάθε σταδίου αντίδρασης, (4) τις μεταβολές των δεσμών, (5) τα κέντρα των αντιδράσεων, τους συμπαράγοντες και τις λειτουργίες τους, και (6) τα στοιχεία σύνδεσης με εξωτερικές βάσεις δεδομένων, όπως για παράδειγμα ο αριθμός EC, τα αναγνωριστικά της UniProtKB, CATH, και οι κωδικοί PDB.

Μία τελική βασική συνιστώσα της δομής των δεδομένων της βάσης MACiE είναι η χρήση ενός αυστηρά ελεγχόμενου λεξιλογίου και ο εκτενής έλεγχος των σφαλμάτων κατά την καταχώρηση. Ελέγχεται αν η πρωτεΐνη έχει ήδη καταχωρηθεί στην MACiE και αν τα σχολιασμένα αμινοξικά κατάλοιπα υπάρχουν στην κρυσταλλική δομή. Εξαιρέσεις γίνονται όταν η καταλυτική μονάδα, δηλαδή η μικρότερη μονάδα που απαιτείται για να συμβεί η κατάλυση, δεν αντιστοιχεί στην ασύμμετρη μονάδα στο αρχείο PDB. Επίσης ελέγχεται εάν τα σχόλια ταιριάζουν με την ιεράρχηση, π.χ., όταν ένα κατάλοιπο είναι σχολιασμένο ως αντιδρόν, αλλά δεν έχει σχολιασμένη λειτουργία. Οι έλεγχοι βοηθούν στην ελαχιστοποίηση των ανθρώπινων λαθών, κάτι που είναι πάντα πιθανό να συμβεί με τον χειροκίνητο σχολιασμό.

ESTHER:Η βάση δεδομένων ESTHER (ESTerases and alpha/beta-Hydrolase Enzymes and Relatives) περιέχει την ανάλυση πρωτεϊνών που ανήκουν στην υπερ-οικογένεια των α/β-υδρολάσων (www.bioweb.supagro.inra.fr/esther). Οι α/β υδρολάσες αποτελούν μια από τις μεγαλύτερες και πιο ποικιλόμορφές υπερ-οικογένειες πρωτεϊνών οι οποίες χαρακτηρίζονται ένα μόνο είδος διπλώματος. Μέχρι στιγμής περιλαμβάνει περισσότερες από 800.000 αλληλουχίες (που αντιστοιχοπυν σε 42.000 μη ομόλογες εγγραφές) ομαδοποιημένες σε 175 υπο-οικογένειες (Lenfant, Hotelier, Bourne, Marchot, & Chatonnet, 2013). Κάθε υπο-οικογένεια έχει δημιουργηθεί σύμφωνα με ένα προφίλ HMM (Lenfant et al., 2013). Μέλη της υπερ-οικογένειας έχουν βασικό ρόλο σχεδόν σε όλες τις φυσιολογικές διαδικασίες και αποτελούν στόχους φαρμάκων για την θεραπεία ασθενειών όπως ο διαβήτης, η παχυσαρκία, και οι νευροεκφυλιστικές διαταραχές. Παρά τις προτεινόμενες κοινές ονομασίες τους, πολλές από αυτές τις πρωτεΐνες δεν είναι ένζυμα, καθώς κάποιες από αυτές έχουν χάσει όλα τα αναγκαία κατάλοιπα που δυνητικά μπορούν να αποτελέσουν ένα ενεργό κέντρο (Lenfant, Hotelier, Bourne, Marchot, & Chatonnet, 2014; Marchot & Chatonnet, 2012). Οι λειτουργίες λίγων εκπροσώπων αυτής της τελευταίας ομάδας είναι γνωστές: ενδοκυτταρικοί υποδοχείς μικρών μορίων, πρόδρομα μόρια ορμονών, αλληλεπίδραση με μόρια σε κυτταρικές επιφάνειες κ.ά. Ένας από τους στόχους της βάσης είναι η σύνδεση των βιολογικών δεδομένων με τις διαφορετικές υπο-οικογένειες, προκειμένου να συμβάλλει στον καθορισμό της λειτουργίας τους. Η βάση ESTHER δημιουργήθηκε το 1994 σε εξυπηρετητή Gopher και γρήγορα πέρασε στο WWW (Cousin, Hotelier, Lievin, Toutant, & Chatonnet, 1996). Το σύστημα στο οποίο βασίζεται είναι το ACeDB.

Η βάση δεδομένων περιέχει επίσης μικρά μόρια που αλληλεπιδρούν με εστεράσες ως υποστρώματα, αναστολείς ή ενεργοποιητές και άλλα συναφή κινητικά δεδομένα (Chatonnet, Cousin, & Robinson, 2001). Στην παρούσα φάση γίνεται προσπάθεια επέκτασης αυτής της ενότητας. Έχει ενσωματωθεί το πακέτο R, προκειμένου να επιτραπεί η στατιστική σύγκριση των κινητικών παραμέτρων των διαφόρων ενζύμων ή μεταλλακτών με διάφορα υποστρώματα και / ή των αναστολέων κάτω από διαφορετικές πειραματικές συνθήκες.

ConoServer: Το δηλητήριο του σαλιγκαριού Cone είναι πιθανά μια μεγάλη πηγή αρκετών εκατοντάδων χιλιάδων ενεργών πεπτιδίων εξαιρετικά επιλεκτικών για τους υποδοχείς και μεταφορείς του νευρικού συστήματος με εφαρμογές σε νευρολογικούς ανιχνευτές και φάρμακα (Akondi et al., 2014; Terlau & Olivera, 2004). Η ποικιλομορφία αυτών των δηλητηρίων (Davis, Jones, & Lewis, 2009) έχει αναλυθεί σε μελέτες γενετικής (Biggs et al., 2010; Chang & Duda, 2012; Puillandre, Koua, Favreau, Olivera, & Stocklin, 2012) και οικολογικές (Duda, Chang, Lewis, & Lee, 2009; Duda & Lee, 2009) μελέτες. Τον Δεκέμβριο του 2014, η βάση δεδομένων ConoServer (Kaas, Yu, Jin, Dutertre, & Craik, 2012) περιείχε περισσότερα από 2000 κονο-πεπτίδια, ενώ βοηθά στη συστηματοποίηση των τριών συστημάτων ταξινόμησης που περιγράφουν την εξέλιξη του κονοπεπτιδίου, τρισδιάστατες δομές και μοριακούς στόχους (Kaas, Westermann, & Craik, 2010).

Η βάση ConoServer (www.conoserver.org) δημιουργήθηκε (όπως και ο σχολιασμός) με στόχο την όσο το δυνατόν μικρότερη απαίτηση ανθρώπινων πόρων, δηλαδή ένα άτομο. Ο στόχος αυτός επετεύχθηκε με την εφαρμογή ενός επιπλέον επιπέδου. Επινοήθηκε ένας ψευδο-πίνακας που συνδέει τα δεδομένα που είναι αποθηκευμένα στην σχεσιακή βάση δεδομένων MySQL σε περιβάλλον PHP. Τα πρωτογενή δεδομένα που υποστηρίζουν την ConoServer προέρχονται από την GenBank (Benson, et al., 2014), την UniProt-KB (UniProt, 2014) και την PDB (Berman, Henrick, Nakamura, & Markley, 2007), καθώς και από τη βιβλιογραφία ή των υποβολών από τους συγγραφείς. Οι πρόσφατα ανακτημένες αλληλουχίες, πριν δημοσιευτούν, σχολιάζονται από διάφορα κείμενα που εισάγουν δεδομένα στη βάση MySQL, τα δεδομένα αυτά στη συνέχεια αναθεωρούνται και τροποποιούνται χειροκίνητα μέσω μιας διεπαφής σχολιασμού.

Ένας αριθμός αλληλουχιών κονοπεπτιδίων είναι διαθέσιμος μόνο σε πίνακες, σχήματα ή συμπληρωματικές πληροφορίες από έγκριτα επιστημονικά άρθρα, και έχουν εισαχθεί χειροκίνητα μέσω μιας διεπαφής του διαδικτύου. Οι κονοπεπτιδικές αλληλουχίες των πρωτεϊνών και των προδρόμων αλληλουχιών νουκλεϊκών οξέων σχολιάζονται βάσει των χαρακτηριστικών των αλληλουχιών και ταξινομούνται σύμφωνα με τρία τυποποιημένα συστήματα ταξινόμησης. Το ConoPrec (Kaas, et al., 2012) είναι ένα διαδικτυακό εργαλείο το οποίο επιτρέπει τη χρήση αυτής της διαδικασίας σχολιασμού, βοηθώντας τους χρήστες να αναλύσουν αλληλουχίες χρησιμοποιώντας πρότυπα της ConoServer πριν από τη δημοσίευση τους, απλοποιώντας έτσι αργότερα την είσοδό τους στην ConoServer.

CyBase: Οι ριβοσωμικώς συντιθέμενες κυκλικές πρωτεΐνες, έχουν παρατηρηθεί σε όλα τα βασίλεια της ζωής (Craik, 2006; Kedarisetti, Mizianty, Kaas, Craik, & Kurgan, 2014). Η κυκλοποίηση της κύριας αλυσίδας καθιστά τις πρωτεΐνες αδιαπέραστες στις εξωπρωτεάσες και οδηγεί σε δραματική βελτίωση της σταθερότητάς τους έναντι της ενζυματικής αποικοδόμησης καθώς και της θερμικής ή χαοτρόπου αποδιάταξης (Trabi & Craik, 2002). Η υψηλή σταθερότητα των κυκλικών πεπτιδίων και πρωτεϊνών έχει προσελκύσει έντονα το ενδιαφέρον των σχεδιαστών φαρμάκων για τη σταθεροποίηση βιοδραστικών πεπτιδικών επιτόπων (Poth, Chan, & Craik, 2013). Η βάση δεδομένων CyBase (www.cybase.org.au) είναι μια βάση που παρέχει πρόσβαση σε πληροφορίες σχετικά με την κωδικοποίηση των γονιδίων, την κύρια αλυσίδα και τις κυκλικές πρωτεΐνες (Wang, Kaas, Chiche, & Craik, 2008). Από τον Δεκέμβριο του 2014 η CyBase περιέχει πληροφορίες για περίπου 420 φυσικά δημιουργημένες και περίπου 160 συνθετικές κυκλικές πρωτεΐνες. Αυτές οι πρωτεΐνες έχουν ταξινομηθεί σε εννέα κύριες κατηγορίες, η μεγαλύτερη των οποίων είναι η κατηγορία cyclotide, με 282 εγγραφές. Οι στρατηγικές καταχώρησης και σχολιασμού της CyBase είναι παρόμοιες με αυτές που περιγράφονται στην βάση ConoServer. Στην CyBase πραγματοποιείται αναζήτηση αλγόριθμων που έχουν προσαρμοστεί για τον χειρισμό των κυκλικών πρωτεϊνών, ενώ τα εργαλεία που χρησιμοποιούνται συχνότερα είναι η στοίχιση αλληλουχίας και η φασματομετρία μάζας στις αναζητήσεις αποτυπωμάτων (Wang, et al., 2008). Ένα ιδιαίτερο χαρακτηριστικό της CyBase είναι η σε βάθος περιγραφή του κειμένου βιολογικής ανάλυσης και φυσικοχημικών χαρακτηρισμών της κάθε κυκλικής πρωτεΐνης.

GPCRDB: Οι υποδοχείς που είναι συζευγμένοι με G-πρωτεΐνες (G protein-coupled receptors, GPCRs) αποτελούν τη μεγαλύτερη οικογένεια μεμβρανικών πρωτεϊνών σε ορισμένους ευκαρυωτικούς οργανισμούς. Ρυθμίζουν μια πληθώρα φυσιολογικών διεργασιών που εκτείνονται από το νευρικό και ενδοκρινικό σύστημα, μέχρι και την αίσθηση των οσμών, της γεύση και του φωτός (Bockaert & Pin, 1999; Lagerstrom & Schioth, 2008). Αποτελούν τους στόχους περίπου του 30% των φαρμάκων της αγοράς, αν και μέχρι σήμερα έχουν αξιοποιηθεί θεραπευτικά λίγοι μόνο από τους υποδοχείς (Garland, 2013; Overington, Al-Lazikani, & Hopkins, 2006). Η βάση δεδομένων GPCR, η GPCRDB (http://gpcrdb.org), ξεκίνησε το 1993. Εκείνη την εποχή ταυτοποιήθηκε μέσω κλωνοποίησης γονιδίων ένας μεγάλος αριθμός αλληλουχιών υποδοχέων, και, καθώς δεν είχαν ακόμη έχουν δημιουργηθεί οι περιηγητές του διαδικτύου, η GPCRDB ήταν αρχικά ένα αυτόματο σύστημα απόκρισης ηλεκτρονικών μηνυμάτων το οποίο μπορούσε να στέλνει αλληλουχίες, στοιχίσεις και μοντέλα ομολογίας. Μετά από δύο δεκαετίες, η GPCRDB εξελίχθηκε σε ένα ολοκληρωμένο πληροφοριακό σύστημα (Horn et al., 2003; Horn et al., 1998; Vroling et al., 2011). Το 2013, η GPCRDB μεταφέρθηκε στην ομάδα Gloriam του Πανεπιστημίου της Κοπεγχάγης, η οποία υποστηρίζεται από το EU COST GPCR Action 'GLISTEN'. Σήμερα, η GPCRDB στοχεύει σε ένα διεπιστημονικό κοινό αντί να αποτελεί πηγή κυρίως για βιοπληροφορικούς. Αυτό περιλαμβάνει την δημοσίευση νέων δεδομένων φιλικών προς το χρήστη, διαγράμματα και εργαλεία, καθώς και αναφορές με σημαντικές συμπληρωματικές βάσεις δεδομένων (Isberg et al., 2014).

Η GPCRDB περιέχει τις μεγαλύτερες συλλογές in vitro μεταλλάξεων στα γονίδια των υποδοχέων οι οποίες δημιουργήθηκαν μετά από αρκετά χρόνια επιμέλειας της επιστημονικής βιβλιογραφίας. Αποτελεί μια

ανοιχτή βάση δεδομένων και επιτρέπει τη συνεισφορά δεδομένων μεταλλαξιγένεσης από τους ερευνητές ώστε να αυξηθεί η διάδοση των δεδομένων και να είναι δυνατή η σύγκρισή τους με δεδομένα που έχουν ήδη δημοσιευθεί. Οι μεταλλάξεις συχνά παρατίθενται μέσα στα διαγράμματα κατάλοιπων των υποδοχέων, τα οποία μπορούν να ανακτηθούν και να χρησιμοποιηθούν σε δημοσιεύσεις ή παρουσιάσεις. Η GPCRDB επίσης διατηρεί συλλογή επιμελημένων αναφορών όλων των κρυσταλλικών δομών GPCR, οι οποίες έχουν αυξηθεί εκθετικά σε αριθμό, λόγω των πρόσφατων τεχνολογικών ανακαλύψεων (Katritch, Cherezov, & Stevens, 2013). Οι δομές μπορούν να αναζητηθούν και να φιλτραριστούν από δεδομένα προσδετών και υποδοχέων και από μέτρα ομοιότητας του στόχου-πρότυπου αλληλουχιών. Πρόσθετα εργαλεία του εξυπηρετητή του διαδικτύου επιτρέπουν τη διαχείριση δομών, για παράδειγμα υπέρθεση στη συνολική δομή ή υπο-περιοχές κατάλοιπων που συνιστούν περιοχές πρόσδεσης του προσδέτη.

Η GPCRDB βρίσκεται στη διαδικασία μετάβασης σε πιο σύγχρονες τεχνολογίες διαδικτύου. Η νέα διεπαφή χρησιμοποιεί τεχνολογίες HTML5 (συμπεριλαμβανομένου του CSS3) και JavaScript ώστε να παρέχει στον χρήστη ένα διαδραστικό περιβάλλον. Για τη δημιουργία διαδραστικών διαγράμματων χρησιμοποιούνται Scalable Vector Graphics (SVG), που μπορούν να αποθηκευτούν σε υψηλή ανάλυση και να μπορούν να παρουσιαστούν σε δημοσιεύσεις. Τα δεδομένα αποθηκεύονται χρησιμοποιώντας τη σχεσιακή βάση MySQL, ενώ ο εξυπηρετητής διαδικτύου Apache χρησιμοποιείται για τις ιστοσελίδες των χρηστών. Η GPCRDB προσφέρει υπηρεσίες διαδικτύου SOAP για πρόσβαση μέσω προγραμματισμού, και έχει σαν στόχο να καταστήσει περισσότερη πληροφορία/περιεχόμενο προσβάσιμο μέσω άλλων διαδικτυακών ιστότοπων.

IUPHAR/BPS Guide to PHARMACOLOGY: Η βάση IUPHAR/BPS Guide to PHARMACOLOGY (GtoPdb, (Pawson et al., 2014)) έχει αναπτυχθεί από κοινού από τη Διεθνή Ένωση Βασικής και Κλινικής Φαρμακολογίας (Union of Basic and Clinical Pharmacology, IUPHAR) και τη Βρετανική Φαρμακολογική Εταιρία (BPS) για να παρέχει πρόσβαση σε υψηλής ποιότητας πληροφορίες για φαρμακευτικούς στόχους. Η GtoPdb καταγράφει τιμές συγγένειας με την χαρτογράφηση βιοδραστικών χημικών δομών των πρωτεϊνών. Η GtoPdb (http://www.guidetopharmacology.org/) είναι μια βάση που συγκεντρώνει προηγουμένως διαφορετικές αλλά συμπληρωματικές πληροφορίες που είχαν αρχικά καταχωρηθεί στη βάση IUPHAR (IUPHAR-DB, (Harmar et al., 2009) και στον Οδηγό Υποδοχέων και Καναλιών (Guide to Receptors and Channels, GRAC), που αποτελεί σειρά δημοσιεύσεων στο περιοδικό BPS, British Journal of Pharmacology (π.χ., (Alexander, Mathie, & Peters, 2011)).
Η GtoPdb έχει σαν στόχο:

- Την παροχή πρόσβασης σε δεδομένα σχετικά με όλους τους γνωστούς βιολογικούς στόχους καθώς και με τους υποδοχείς/κανάλια ιόντων
- Να προτείνει προσδέτες για χαρακτηρισμό των εν λόγω στόχων
- Την παροχή ενός σημείου εισόδου στην βιβλιογραφία της φαρμακολογίας
- Την παροχή μιας ολοκληρωμένης πηγής εκπαίδευσης με υψηλή ποιότητα εξάσκησης στις αρχές της βασικής και κλινικής φαρμακολογίας καθώς και τις τεχνικές της
- Την προώθηση καινοτόμων φαρμακευτικών ανακαλύψεων

Μερικά προβλήματα στην τρέχουσα ανακάλυψη φαρμάκων αφορούν τον αριθμό των μεταβλητών που εμπλέκονται στις αλληλεπιδράσεις φαρμάκου-υποδοχέα, τις νέες περιοχές ncRNAs (σε εξέλιξη, με HGNC), στην επιγενετική (Tough, Lewis, Rioja, Lindon, & Prinjha, 2014), την εναλλακτική συρραφή (Bonner, 2014), το allostery (Christopoulos et al., 2014), και τις ανοσολογικές αντιδράσεις (σε εξέλιξη), που συμβάλλουν σημαντικά στις διεργασίες μιας νόσου (Spedding, 2011). Για τον λόγο αυτό η συμβολή εξειδικευμένων επιστημόνων μπορεί να βοηθήσει σημαντικά στην επίλυση δυσκολιών που προκύπτουν.

Η GtoPdb περιλαμβάνει σήμερα πάνω από 2.700 επιβεβαιωμένους ή πιθανούς στόχους φαρμάκων και σχετικών πρωτεϊνών (υποδοχείς συζευγμένοι με G-πρωτεΐνη συμπεριλαμβανομένων των ορφανών GPCRs, κανάλια ιόντων, υποδοχείς πυρηνικών ορμονών, καταλυτικοί υποδοχείς, κινάσες, πρωτεάσες, μεταφορείς κλπ), μαζί με προσδέτες, οι οποίοι είναι είτε φάρμακα που διατίθενται ήδη στην αγορά ή πιθανά φάρμακα για ανάπτυξη, ή τα καλύτερα διαθέσιμα πειραματικά εργαλεία για την αξιολόγηση των εν λόγω στόχων. Οι περαιτέρω στόχοι περιλαμβάνουν τη σύντομη εισαγωγή στην φαρμακολογία, ανάγνωση υποβάθρου, και δημοσιευμένα δεδομένα σχετικά με τη συγγένεια των προσδετών και των στόχων τους. Για τα υποσύνολα πολύ σημαντικών στόχων παρέχονται λεπτομερείς σχολιασμοί όπως λειτουργία, φυσιολογία, και βιολογικές ή κλινικές σχετικές παραλλαγές.

Η GtoPdb μπορεί να μην έχει την έκταση της ChEMBL (Gaulton et al., 2012), ωστόσο συμπληρώνει τις προσεγγίσεις μεγάλης κλίμακας με το να είναι μια εστιασμένη βάση δεδομένων σε προσεκτικά

επιλεγμένους προσδέτες και στόχους και που ο σχολιασμός της γίνεται από εξειδικευμένους επιστήμονες. Παρέχει βασικές πληροφορίες και σχόλια που προστίθενται στο γενικό πλαίσιο. Επιπλέον, παρέχονται σύνδεσμοι σε αντίστοιχες εγγραφές σε άλλες πηγές, π.χ. UniProt, Ensembl, Entrez Gene, KEGG, OMIM. Οι πληροφορίες σχετικά με τους προσδέτες περιλαμβάνουν χημικές δομές, αλληλουχίες πεπτιδίων, κλινικά δεδομένα και ονοματολογία, που συνδέονται με βασικές πηγές, συμπεριλαμβανομένων των PubChem, DrugBank και ChEMBL. Τέλος, υπάρχει ένας δημοσιευμένος οδηγός φαρμακολογίας 'Concise Guide to PHARMACOLOGY', που δημιουργήθηκε στην GtoPdb από περιλήψεις οικογενειών στόχων, και χρησιμεύει ως οδηγός γρήγορης αναφοράς. Δημοσιεύεται ανά διετία στο British Journal of Pharmacology και αντικαθιστά την GRAC (Alexander et al., 2013).

Kinase.com: Η Kinase.com διερευνά τις λειτουργίες και την εξέλιξη των πρωτεϊνικών κινασών, οι οποίες αποτελούν βασικούς ρυθμιστές των περισσότερων βιοχημικών μονοπατιών και είναι ιδιαίτερα σημαντικές για την υγεία και τις ασθένειες (Manning, Whyte, Martinez, Hunter, & Sudarsanam, 2002). Επικεντρώνεται στο "kinome", δηλαδή στο σύνολο των κινασών σε ένα γονιδίωμα. Η ιστοσελίδα της βάσης, KinBase, είναι διαδραστική και περιλαμβάνει πληροφορίες για πάνω από 7.000 γονίδια πρωτεϊνικών κινασών που βρίσκονται στο ανθρώπινο γονιδίωμα, καθώς επίσης 14 επιπλέον γονιδιώματα (Bradham et al., 2006; Caenepeel, Charydczak, Sudarsanam, Hunter, & Manning, 2004; Eisen et al., 2006; Goldberg et al., 2006; Srivastava et al., 2010; Stajich et al., 2010). Οι κινάσες κατατάσσονται ιεραρχικά σε 10 ομάδες, 287 οικογένειες και 356 υποοικογένειες. Η αναζήτηση στην KinBase μπορεί να γίνει βάσει του ονόματος των γονιδίων, των συμπληρωματικών δομικών μοτίβων, ή σύμφωνα με την ταξινόμηση. Επιπλέον, η ιστοσελίδα παρέχει την υπηρεσία BLAST ώστε η αναζήτηση των κινασών να μπορεί να πραγματοποιηθεί με βάση την ομοιότητα αλληλουχίας.

Κάθε κινάση έχει τη δική της σελίδα που περιέχει την ταξινόμηση, την αλληλουχία, τον σχολιασμό της από εξωτερικές πηγές, το γράφημα του συδυασμού των δομικών μοτίβων, και συγκεκριμένα, τη σύγκριση με το αντίστοιχο προφίλ HMM των ομάδων κινάσης, των οικογενειών και των υπεροικογενειών. Κάθε κατηγορία κινάσης (ομάδα, οικογένεια και υποοικογένεια) έχει τη δική της σελίδα που περιέχει την στοίχηση αλληλουχίας, το προφίλ HMM και το φυλογενετικό δέντρο των πρωτεϊνικών αλληλουχιών και των μοτίβων κινασών. Εκτός από την ιστοσελίδα KinBase, η Kinase.com περιέχει ένα δημόσιο σύστημα wiki, το WiKinome, που εστιάζει στην εξέλιξη και τη λειτουργία των κινασών. Ο απώτερος στόχος είναι η δημιουργία μίας σελίδας wiki για κάθε οικογένεια και υπο-οικογένεια κινάσης.

Η βάση Kinase.com δημιουργήθηκε το 1999 για την υποστήριξη της δημοσιευμένης ανάλυσης της εταιρείας σχεδιασμού φαρμάκων Sugen σχετικά με τις κινάσες του Caenorhabditis elegans (Bingham, Plowman, & Sudarsanam, 2000; Manning, 2005; Plowman, Sudarsanam, Bingham, Whyte, & Hunter, 1999) και του Saccharomyces cerevisiae (Hunter & Plowman, 1997). Η βάση δεδομένων KinBase δημιουργήθηκε το 2002 για την υποστήριξη των περαιτέρω εργασιών ανθρώπινων πρωτεϊνικών κινασών (Manning, Whyte, et al., 2002) και των μυγών των φρούτων (Manning, Plowman, Hunter, & Sudarsanam, 2002). Έχει αναπτυχθεί χρησιμοποιώντας βάση δεδομένων MySQL και τη γλώσσα προγραμματισμού Perl. Η ιστοσελίδα της έχει ανανεωθεί πρόσφατα με τη χρήση σύγχρονων τεχνολογιών ανάπτυξης ιστοσελίδων συμπεριλαμβανομένων των Model–view–controller web framework, HTML5, CSS5 και JavaScript.

Βασική δυσκολία αποτελεί η ενημέρωση της βάσης δεδομένων με στοιχεία υψηλής ποιότητας. Παρά το γεγονός ότι έχουν αλληλουχηθεί πάνω από 6.000 ευκαρυωτικά γονιδιώματα, η Kinase.com περιλαμβάνει τα kinomes μόνο των 15 γονιδιωμάτων. Έχει γίνει προσπάθεια αυτόματης εύρεσης και καταχώρησης πρωτεϊνικών κινασών για όλα τα γονιδιωμάτα, αλλά δεν έχει επιτευχθεί η ίδια ποιότητα, από πλευράς μοντέλου γονιδίου ή/και ταξινόμησης, όπως για τα kinomes των 15 γονιδιωμάτων που αναφέρθηκαν. Χωρίς την χειροκίνητη επιμέλεια και άλλων ερευνητών, είναι δύσκολο να πραγματοποιούνται τακτικές και συχνές ενημερώσεις υψηλής ποιότητας των kinomes.

Structure-Function Linkage Database: Η βάση δεδομένων Structure-Function Linkage Database (SFLD; http://sfld.rbvi.ucsf.edu/django/), (Akiva et al., 2014; Pegg et al., 2006), παρέχει την ιεραρχική ταξινόμηση των λειτουργικά διαφορετικών υπερ-οικογενειών των ενζύμων και συνδέει τις αλληλουχίες και τα δομικά χαρακτηριστικά για κάθε ένζυμο. Δημιουργήθηκε για να διευκολύνει την εννοιολογική κατανόηση του τρόπου που εκπροσωπούνται οι διάφορες αντιδράσεις σε υπερ-οικογένειες που εξελίσσονται (Gerlt & Babbitt, 2001). Η SFLD είναι η μοναδική από τις πηγές που σχολιάζουν πρωτεΐνες η οποία χρησιμοποιεί ως βάση για την συσχέτιση των αλληλουχιών, της δομής και των καταλυτικών χαρακτηριστικών το "χημικά - περιορισμένο" μοντέλο (Babbitt & Gerlt, 1997).

Τέτοιες υπερ-οικογένειες συναντώνται στη φύση και εκτιμάται ότι αντιπροσωπεύουν τουλάχιστον το ένα τρίτο του συνόλου των υπερ-οικογενειών των ενζύμων (Almonacid & Babbitt, 2011). Όλα τα μέλη της υπερ-οικογένειας εμφανίζουν συντήρηση των λειτουργικά σημαντικών καταλοίπων του ενεργού κέντρου, ενώ τα υποστρώματα τους, τα προϊόντα και ακόμη και οι συνολικές αντιδράσεις μπορεί να είναι ουσιωδώς διαφορετικά. Στο ανώτερο επίπεδο της ιεραρχίας (επίπεδο υπερ-οικογένειας), η SFLD συνδέει αυτά τα συντηρημένα μοτίβα του ενεργού κέντρου με χημικά χαρακτηριστικά που είναι κοινά σε όλα τα μέλη. Για παράδειγμα, στην περίπτωση της ιεραρχίας της υπερ-οικογένειας ενολάσης, όλα τα ένζυμα που χαρακτηρίζονται μέλη της υπερ-οικογένειας έχουν κοινή μια παρόμοια αρχιτεκτονική ενεργού κέντρου που συνδέεται με μια συγκεκριμένη μερική αντίδραση, την αφαίρεση ενός πρωτονίου προς ένα υπόστρωμα καρβοξυλικού άλατος, και τον σχηματισμό ενός κοινού τύπου ενδιαμέσου ενολικού ανιόντος (Babbitt et al., 1996; Gerlt, Babbitt, & Rayment, 2005).

Η βάση δεδομένων παρέχει υψηλής ποιότητας επιμέλεια των λειτουργικών ιδιοτήτων σε επίπεδο υπερ-οικογένειας, υποομάδας και οικογένειας για ένα μικρό σύνολο μεγάλων και ποικίλων υπερ-οικογενειών (Core SFLD), μαζί με πολλούς τύπους σχετικών μετα-δεδομένων και αποτελεσμάτων ανάλυσης. Τα δεδομένα και οι πληροφορίες για κάθε μία από τις υπερ-οικογένειες, υποομάδες, και τα επίπεδα οικογενειών διατίθενται ελεύθερα μέσω ενός εξελιγμένου γραφικού περιβάλλοντος διεπαφής του χρήστη (user interface). Το διαθέσιμο υλικό περιλαμβάνει αρχείο της υπερ-οικογένειας των αλληλουχιών, των HMMs, των σχολιασμένων πολλαπλών στοιχίσεων, απεικονίσεων των 3D δομών και σχολιασμένων ενεργών κέντρων που μπορεί να επεξεργαστεί χρησιμοποιώντας το ελεύθερα διαθέσιμο λογισμικό Chimera (Pettersen et al., 2004), καθώς και συνδέσμους με πολλές άλλες σχετικές πηγές. Η ενότητα Extended SFLD παρέχει λιγότερο επιμελημένες πληροφορίες για ένα μεγαλύτερο σύνολο λειτουργικά διαφορετικών υπεροικογενειών ενζύμων.

Για τα δεδομένα που διατίθενται στους χρήστες, η SFLD χρησιμοποιεί το Django, το οποίο είναι ένα πλαίσιο υψηλού επιπέδου του Python Web, για να δημιουργήσει το διαδικυακό περιβάλλον. Η χρήση αυτού του πλαισίου διευκολύνει σημαντικά την ανάπτυξη της διεπαφής που αλληλεπιδρούν οι χρήστες, καθώς και την καταχώρηση δεδομένων από τους επιμελητές μέσω ενός εξελιγμένου γραφικού περιβάλλοντος χρήστη (GUI), το οποίο επιτρέπει επίσης τον έλεγχο λαθών κατά την καταχώρηση των δεδομένων.

Καθώς τα δεδομένα πρωτεϊνικών αλληλουχιών συνεχίζουν να αυξάνονται με εκθετικό ρυθμό, έχουν αυξηθεί και τα μέλη των υπερ-οικογενειών που σε ορισμένες περιπτώσεις έχουν ξεπεράσει τις 100.000 αλληλουχίες. Για την αντιμετώπιση αυτής της πρόκλησης και την παροχή υποστήριξης για την εφαρμογή του ιεραρχικού μοντέλου SFLD σε αυτές τις μεγάλες και διαφορετικές υπερ-οικογένειες του Core και Extended SFLD, χρησιμοποιούνται τα δίκτυα ομοιότητας πρωτεϊνών (Atkinson, Morris, Ferrin, & Babbitt, 2009).

Histone Database: Η βάση Histone (http://research.nhgri.nih.gov/histones/) ιδρύθηκε το 1996 (Baxevanis & Landsman, 1996), ως αποτέλεσμα έρευνας αναφορικά με το δίπλωμα των ιστόνων, πρωτεϊνών που προσδένουν το DNA (Baxevanis, Arents, Moudrianakis, & Landsman, 1995). Όλα αυτά τα χρόνια έχουν χρησιμοποιηθεί διάφορα εργαλεία για τον εντοπισμό των ιστόνων στις βάσεις δεδομένων αλληλουχιών, συμπεριλαμβανομένων των πρόσφατων εκδόσεων των PSI-BLAST (Altschul et al., 1997) και HMMER (Eddy, 2009). Μετά την ταυτοποίηση, τα εργαλεία αυτά χρησιμοποιούνται για τον έλεγχο της στοίχισης των αλληλουχιών και των εσφαλμένων καταχωρήσεων στις βάσεις δεδομένων. Τα περισσότερα από αυτά τα λάθη σχετίζονται με την εσφαλμένη τοποθέτηση των κωδικονίων έναρξης. Δεδομένου ότι οι ιστόνες είναι άκρως συντηρημένες και ότι το δίπλωμα των ιστόνων περιγράφεται και σχολιάζεται επαρκώς, είναι αρκετά εύκολο να ταυτοποιηθούν οι ιστόνες που είναι εσφαλμένα σχολιασμένες στις δημόσιες βάσεις δεδομένων. Οι στοιχίσεις της κάθε οικογένειας (H1, H2A, H2B, H3, H4) είναι διαθέσιμες για μεταφόρτωση, ενώ υπάρχει επίσης μία σελίδα αναζήτησης για την εξαγωγή μόνο των αλληλουχιών για τις οποίες ενδιαφέρεται ο χρήστης. Οι στοιχίσεις αυτές έχουν χρησιμοποιηθεί για την ταυτοποίηση τέτοιων πρωτεϊνών (Baxevanis, et al., 1995), καθώς και την πρόσφατη προσθήκη της οικογένειας των ιστονών των Αρχαιοβακτηρίων (Marino-Ramirez et al., 2011). Αυτές οι τελευταίες συλλογές πρωτεϊνών ταξινομούνται στη βάση Histone Database ως ξεχωριστές οικογένειες. Η Histone Database περιέχει επίσης λίστες προσδιορισμένων τρισδιάστατων δομών ιστονών που εξάγονται από την Protein Data Bank (PDB) (Rose et al., 2014), ως επί το πλείστον με τη μορφή του νουκλεοσωματικών δομών που προσδιορίζονται με κρυσταλλογραφία ακτίνων-Χ.

Οι μελλοντικές προκλήσεις περιλαμβάνουν την ενημέρωση της βάσης με νέες αλληλουχίες από τις δημόσιες βάσεις δεδομένων και την υποδιαίρεση κάθε οικογένειας των ιστονών σε διάφορες υπο-οικογένειες ή υπο-τύπους (π.χ.κεντρομερικές ιστόνες (CENP-A και CSE4), ιστόνες H3.3, H2A.B , H2A.Z, H2B.Z, macroH2A (136) και, ενδεχομένως, διάφορους υπο-τύπους ιστόνης H1).

Άλλες εξειδικευμένες βάσεις δεδομένων

Παρόλο που, όπως αναφέραμε ήδη, οι περισσότερες εξειδικευμένες βάσεις αφορούν πρωτεϊνικές αλληλουχίες, υπάρχουν δεκάδες άλλες δημόσια διαθέσιμες βάσεις δεδομένων που είναι δυνατόν να ανήκουν σε οποιαδήποτε από τις κατηγορίες των πρωτογενών βάσεων που αναφέρθηκαν παραπάνω.

Κάποιες από αυτές μπορεί να είναι εξειδικευμένες με την έννοια ότι συλλέγουν όλη τη διαθέσιμη πληροφορία για το γονιδίωμα ενός οργανισμού και τις πρωτεΐνες που αυτό κωδικοποιεί (όμοια με την nextProt που αναφέραμε παραπάνω). Τέτοια παραδείγματα είναι η **SubtiList** (http://genolist.pasteur.fr/SubtiList/) (Moszer, Jones, Moreira, Fabry, & Danchin, 2002) για τον *Bacillus subtilis* και η **EcoCyc** (http://ecocyc.org/) για την *Escherichia coli* K-12 (Karp et al., 2002). Παρόμοιες βάσεις υπάρχουν και για άλλους οργανισμούς, ενώ η **Genome Online Database** (GOLD), περιέχει κατάλογο με όλους τους οργανισμούς με πλήρως προσδιορισμένο γονιδίωμα (https://gold.jgi-psf.org/) (Reddy et al., 2015).

Για τα δεδομένα γονιδιακής έκφρασης, υπάρχουν επίσης αρκετές εξειδικευμένες βάσεις δεδομένων. Για παράδειγμα η **ONCOMINE,** είναι βάση δεδομένων που περιέχει πειράματα μικροσυστοιχιών που αφορούν διαφόρους τύπους καρκίνου. Επίσης παρέχει στο χρήστη εργαλεία διαχείρισης των δεδομένων για την αποδοτικότερη εύρεση των επιθυμητών πειραμάτων και γονιδίων, http://www.oncomine.org/, (Rhodes et al., 2004). Το **RNA-Seq Atlas** (http://medicalgenomics.org/rna_seq_atlas) είναι μια δημόσια βάση δεδομένων για δεδομένα από αλληλούχιση RNA (RNA-Seq). Περιέχει δεδομένα έκφρασης για 11 διαφορετικούς ιστούς από υγιείς ανθρώπους. Η βάση περιέχει επίσης εργαλεία για τη σύγκριση μεταξύ των ιστών καθώς και για την εύρεση γονιδίων με εντοπισμένη έκφραση σε κάποιον ιστό (Krupp et al., 2012). Το **Next Generation Sequencing Catalog (NGS Catalog)** είναι μια δημόσια βάση δεδομένων για τη συλλογή δεδομένων έκφρασης από μελέτες αλληλούχισης νέας γενιάς σε ανθρώπους και βασίζεται σε συλλογή δεδομένων από τη βιβλιογραφία (http://bioinfo.mc.vanderbilt.edu/NGS/index.html). Η βάση περιέχει βιβλιογραφικά δεδομένα, βιολογικές πληροφορίες όπως πληροφορίες για την ασθένεια ή τον πληθυσμό και τεχνικές λεπτομέρειες για τη διαδικασία αλληλούχισης (Xia et al., 2012).

Ειδική αναφορά αξίζει στις όλο και περισσότερο αναπτυσσόμενες τα τελευταία χρόνια, βάσεις δεδομένων γενετικής συσχέτισης. Οι βάσεις αυτές περιέχουν πληροφορίες που εμπλέκουν γονίδια και παραλλαγές των γονιδίων (τους πολυμορφισμούς δηλαδή) με ασθένειες. Παραδοσιακά, υπήρχε η **OMIM (Online Mendelian Inheritance in Man),** η οποία περιέχει κυρίως πληροφορίες για νοσήματα μονογονιδιακής αιτιολογίας (http://www.ncbi.nlm.nih.gov/omim). Τα τελευταία χρόνια όμως, με την ανάπτυξη της γενετικής επιδημιολογίας και των μελετών γενετικής συσχέτισης, έχουν αρχίσει να αναπτύσσονται και οι αντίστοιχες βάσεις δεδομένων, οι οποίες βασίζονται κυρίως σε ανάλυση των δημοσιευμένων εργασιών. Το πρώτο παράδειγμα ήταν η **GAD** (Genetic Association Database, http://geneticassociationdb.nih.gov/) η οποία συνέλλεγε όλες τις σχετικές δημοσιεύσεις από την PubMed αλλά πλέον σταμάτησε τη λειτουργία της (Becker, Barnes, Bright, & Wang, 2004), ενώ το **Catalog of Published Genome-Wide Association Studies** (http://www.genome.gov/gwastudies/) και η **GWASdb** (http://jjwanglab.org/gwasdb) επικεντρώνονται στις ευρυγονιδιωματικές μελέτες (genomewide association studies), οι οποίες στηρίζονται σε μια τεχνολογία υψηλής απόδοσης ανάλογης με αυτήν των μικροσυστοιχιών DNA. Επιπλέον δε, έχουν αναπτυχθεί και μικρότερες βάσεις δεδομένων, οι οποίες συλλέγουν και επεξεργάζονται δεδομένα ειδικά για μια συγκεκριμένη ασθένεια, όπως για παράδειγμα η **Epilepsy Genetic Association Database (epiGAD)** για την επιληψία (Tan & Berkovic, 2010), η **Cancer GAMAdb** (Schully et al., 2011) για τον καρκίνο, ή η **AlzGene** για τη νόσο Alzheimer (Bertram, McQueen, Mullin, Blacker, & Tanzi, 2007).

Τέλος, παρόλο που στην ενότητα για τις εξειδικευμένες βάσεις δεδομένων που συμμετείχαν στο δίκτυο SPRN αναφέρθηκαν και περιγράφηκαν μια σειρά από τέτοιες βάσεις, είναι προφανές πως υπάρχουν δεκάδες άλλες βάσεις που περιέχουν σημαντικά δεδομένα και αξίζουν περιγραφή. Ένα καλό σημείο αναφοράς, είναι η συλλογή Database Collection του περιοδικού Nucleic Acids Research, στο οποίο κάθε χρόνο σε ειδικό τεύχος δημοσιεύονται άρθρα που περιγράφουν βάσεις βιολογικών δεδομένων (http://www.oxfordjournals.org/our_journals/nar/database/cap/). Μερικές τέτοιες βάσεις που αξίζουν ειδικής αναφοράς είναι η **PDBTM** (http://pdbtm.enzim.hu/) που περιέχει τις τοπολογίες από τις τρισδιάστατες δομές διαμεβρανικών πρωτεϊνών (Kozma, Simon, & Tusnady, 2013), η **ExTopoDB** (http://bioinformatics.biol.uoa.gr/ExTopoDB) η οποία περιέχει πειραματικά δεδομένα για την τοπολογία μεμβρανικών πρωτεϊνών με όχι γνωστή τρισδιάστατη δομή (Tsaousis et al., 2010) και η **gpDB**

(http://bioinformatics.biol.uoa.gr/gpDB), η οποία περιέχει δεδομένα για τους GPCRs και τις αλληλεπιδράσεις τους με τις G-πρωτεΐνες (Theodoropoulou, Bagos, Spyropoulos, & Hamodrakas, 2008). Η **DBPTM** (http://dbptm.mbc.nctu.edu.tw/) είναι μια βάση δεδομένων που συλλέγει διαφόρων τύπων δεδομένα για μετα-μεταφραστικές τροποποιήσεις πρωτεϊνών (Lu et al., 2013), ενώ η **DIP** (http://dip.doe-mbi.ucla.edu/dip/Main.cgi) περιέχει δεδομένα, προερχόμενα από διαφορετικές μεθοδολογίες, που αφορούν αλληλεπιδράσεις πρωτεϊνών-πρωτεϊνών (Xenarios et al., 2002). Τέλος, η **bioGrid** (http://thebiogrid.org/) περιέχει γενικά δεδομένα και εργαλεία ανάλυσης για βιολογικές αλληλεπιδράσεις (Stark et al., 2006).

Όσον αφορά τα νουκλεϊκά οξέα (DNA και RNA), τα πράγματα είναι επίσης παρόμοια. Υπάρχουν δεκάδες διαθέσιμες εξειδικευμένες βάσεις δεδομένων και αξίζει εδώ να αναφέρουμε τουλάχιστον αυτές που περιέχουν miRNA και στόχους αυτών, όπως η **MiRBase** (http://www.mirbase.org/) (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006), η **MirTarBase** (http://mirtarbase.mbc.nctu.edu.tw/) (Hsu et al., 2011) και η **TarBase** (http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index/) (Sethupathy, Corda, & Hatzigeorgiou, 2006). Άλλες σημαντικές βάσεις δεδομένων, είναι αυτές που περιέχουν δεδομένα για εσώνια-εξώνια όπως η **EID** (http://bpg.utoledo.edu/~afedorov/lab/eid.html) (Shepelev & Fedorov, 2006), αλλά και αυτές που ασχολούνται με τους υποκινητές των γονιδίων όπως η **EPD** (http://epd.vital-it.ch/) (Dreos, Ambrosini, Cavin Perier, & Bucher, 2013) και η **MMPROMdb** (http://mpromdb.wistar.upenn.edu/) (Sun et al., 2006). Φυσικά, η λίστα δεν τελειώνει εδώ, και για εξειδικευμένες αναζητήσεις, οι χρήστες θα πρέπει να παρακολουθούν τη βιβλιογραφία και να ενημερώνονται για δημοσιεύσεις που περιγράφουν νέες βάσεις δεδομένων.

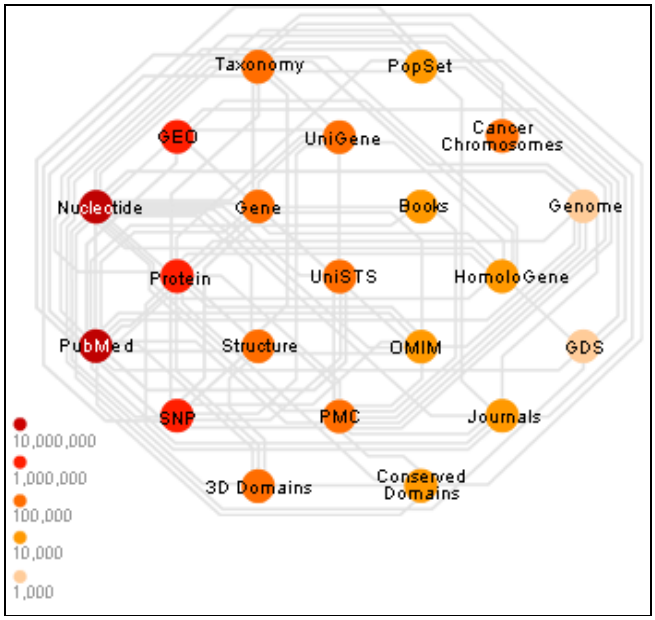
2.3. Ολοκληρωμένα συστήματα ανάκτησης πληροφοριών από βάσεις δεδομένων.

Το **SRS**, είναι ένα ειδικό λογισμικό που διατίθεται από την εταιρία LION Bioscience και αποτελεί ένα ισχυρό και εύχρηστο σύστημα διαχείρισης βιολογικών δεδομένων. Είναι μεν εμπορικό λογισμικό, αλλά διατίθεται δωρεάν για ακαδημαϊκή χρήση. Παρέχει την δυνατότητα αναζήτησης και ανάκτησης δεδομένων σε ένα φιλικό προς τον χρήστη γραφικό περιβάλλον και σε περισσότερες από 400 βάσεις δεδομένων οι οποίες μπορεί να είναι αποθηκευμένες στον ίδιο κεντρικό υπολογιστή. Το βασικό πλεονέκτημα του SRS είναι η δυνατότητα ταυτόχρονης αναζήτησης πληροφοριών σε περισσότερες από μία βάσεις, οι οποίες είναι πιθανό να περιέχουν πληροφορίες διαφορετικού είδους καθώς και η δυνατότητα που δίνει έτσι ώστε η μορφοποίηση των δεδομένων σε καθεμιά από αυτές να είναι διαφορετική. Επιπλέον, λαμβάνοντας υπόψη τον τεράστιο όγκο πληροφορίας και τον μεγάλο αριθμό βάσεων που μπορεί να διαχειρίζεται ταυτόχρονα, σημαντικό πλεονέκτημα αποτελεί η ταχύτητα με την οποία εκτελούνται οι αναζητήσεις. Τέλος δίνεται η δυνατότητα στον κάτοχο του συστήματος να ενσωματώνει σε αυτό και βάσεις που έχει δημιουργήσει ο ίδιος ή προγράμματα για κάθε είδος υπολογιστική ανάλυση χωρίς να επηρεάζεται η απόδοση του συστήματος. Πάνω στο SRS είχαν χτιστεί παλιότερα οι βάσεις του EBI και άλλων μεγάλων ερευνητικών ινστιτούτων. Παρ' όλα αυτά, πλέον θεωρείται παροχημένο και οι σύγχρονες βάσεις όπως η Uniprot χρησιμοποιούν ειδικά κατασκευασμένα συστήματα βάσεων δεδομένων για την αποθήκευση του όλο και μεγαλύτερου όγκου των δεδομένων.

Το **Entrez** αποτελεί ένα σύστημα διαχείρισης δεδομένων για την αναζήτηση και ανάκτηση πληροφοριών όλων των βάσεων δεδομένων που περιέχονται στο NCBI (National Center for Biotechnology Information) των ΗΠΑ. Το Entrez είναι ανάλογο του SRS και παρέχει στον χρήστη τη δυνατότητα αναζήτησης σε βάσεις δεδομένων νουκλεοτιδικών και πρωτεϊνικών αλληλουχιών, δομές βιομορίων και γονιδιωμάτων. Επιπλέον, μέσω του ίδιου γραφικού περιβάλλοντος, παρέχει την δυνατότητα αναζήτησης στη βάση βιβλιογραφίας PUBMED καθώς και πιο πολύπλοκες αναζητήσεις ανάμεσα στα στοιχεία τους. Βασικό μειονέκτημα αποτελεί το γεγονός ότι περιορίζεται μόνο στις βάσεις δεδομένων του NCBI και ότι δεν επιτρέπει ιδιαίτερα πολύπλοκες αναζητήσεις. Παρ' όλα αυτά, αποτελεί για χρόνια τώρα την διεπαφή όλων των βάσεων δεδομένων του NCBI, και επιτρέπει με τον ίδιο απλό τρόπο ο χρήστης να πραγματοποιήσει αναζητήσεις σε τελείως διαφορετικές βάσεις δεδομένων.

Αξίζει να αναφερθεί, ότι μία από τις διαπιστώσεις της συνάντησης του δικτύου SPRN, όσον αφορά τις εξειδικευμένς βάσεις δεδομένων, ήταν ότι στην συντριπτική τους πλειοψηφία, οι βάσεις αυτές στηρίζονται σε κάποιο γενικό σύστημα βάσης δεδομένων όπως η MySQL σε συνδυασμό με PHP. Όπως αναφέρθηκε, παρόλο που στις περισσότερες περιπτώσεις η ιεραρχία ήταν απλή, και θα αρκούσε και μια απλή ιστοσελίδα, το σύστημα διαχείρισης και μόνο (πχ για να γίνονται γρήγορες ανανεώσεις της βάσης ή αντίγραφα ασφαλείας κλπ), ήταν αρκετό για τους ερευνητές για να επιλέξουν αυτόν τον σχεδιασμό. Σε άλλες περιπτώσεις με πιο

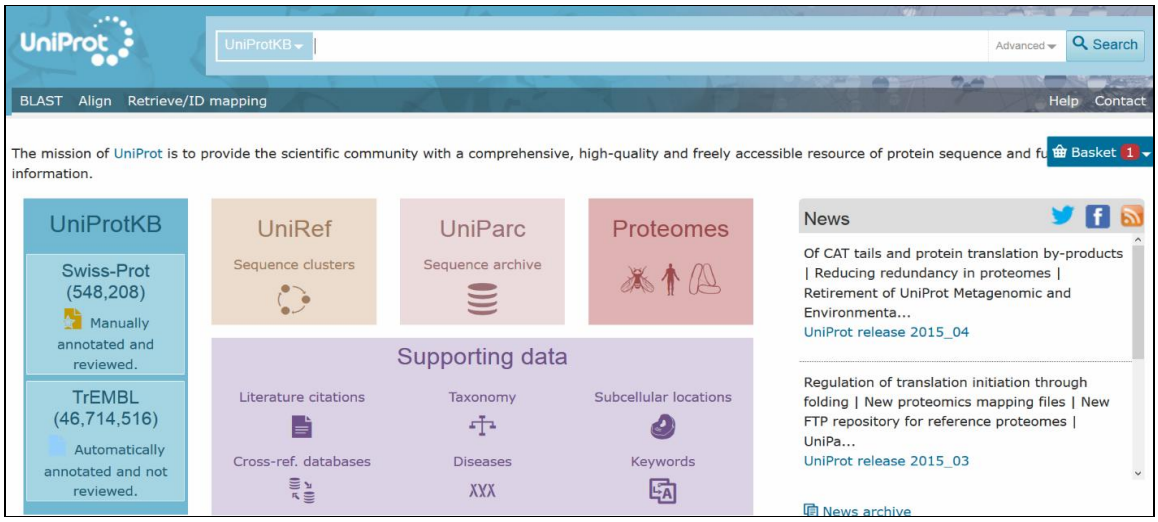
πολύπλοκη ιεραρχία, η SQL προσδίδει επίσης τα απαραίτητα χαρακηριστικά στους διαχειριστές, και έτσι φαίνεται ότι αυτό το μοντέλο είναι αρκετά διαδεδομένο (αν και δεν υπάρχουν πλήρη δεδομένα για όλες τις μικρές βάσεις που έχουν δημοσιευτεί).



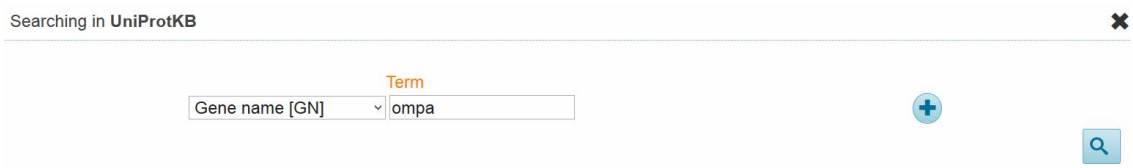
Εικόνα 2.9: *Διαγραμματική απεικόνιση της διασύνδεσης των διαφορετικών βάσεων του NCBI οι οποίες στηρίζονται στο Entrez. Τα διαφορετικά χρώματα, αντιστοιχούν σε διαφορετικό αριθμό καταχωρήσεων. Το NCBI διαθέτει ένα ολοκληρωμένο σύστημα που καλύπτει όλο το εύρος των δημόσιων βάσεων δεδομένων, ακόμα και αυτών που στηρίζονται σε άλλες πηγές. Για παράδειγμα, η Conserved Domains Database είναι αντίστοιχη της PROSITE ενώ η Structure (MMDB) είναι αντίστοιχη της PDB.*

Πρακτικό Μέρος

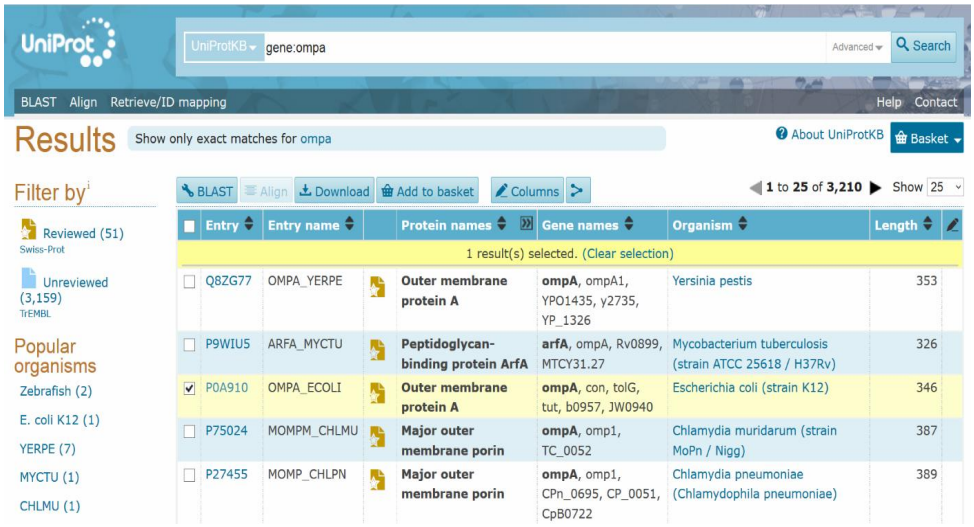
1. Να γίνει αναζήτηση της πρωτεϊνικης αλληλουχίας ompA του βακτηρίου Escherichia coli στη UNIPROT, με βάση το πεδίο Gene Name και να ανακτήσετε την εγγραφή της UNIPROT.



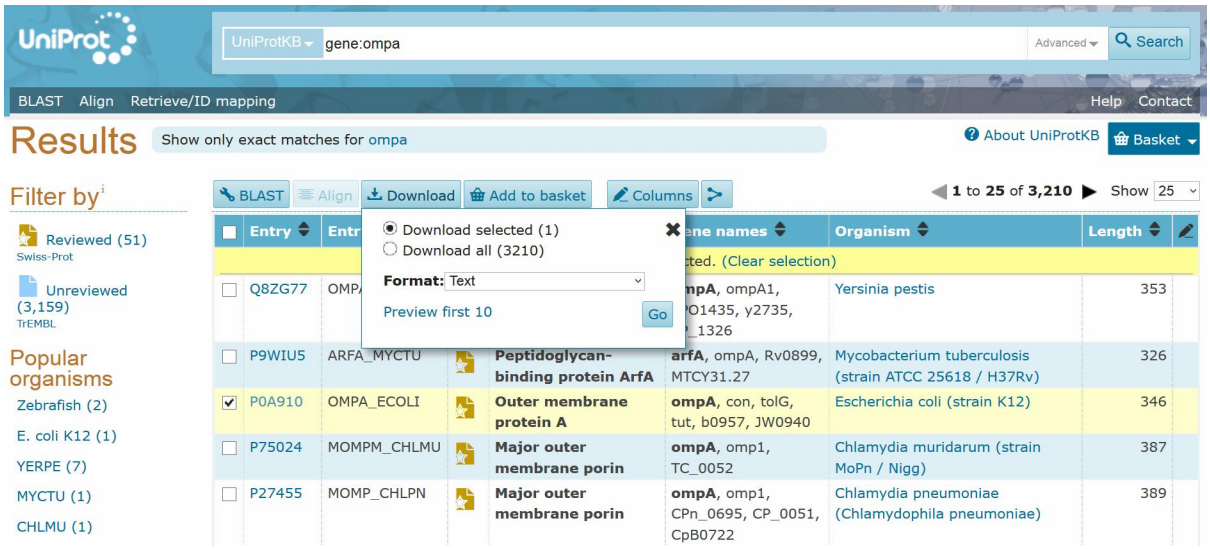
Εικόνα 2.10: *Η αρχική σελίδα της Uniprot*



Εικόνα 2.11: *Επιλέγουμε το όνομα του γονιδίου στο αντίστοιχο πεδίο*

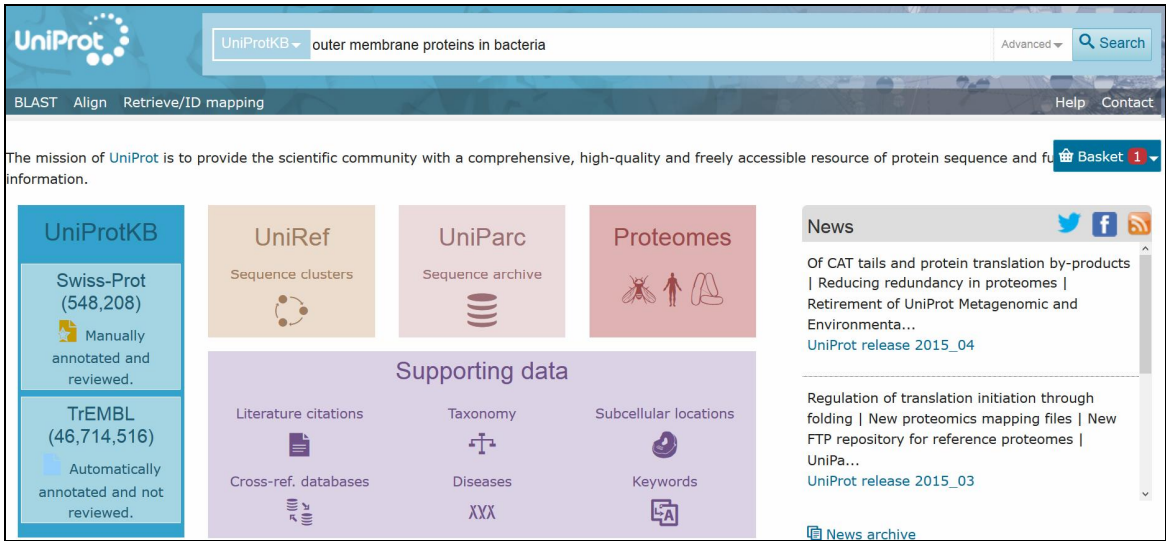


Εικόνα 2.12: *Επιλέγουμε την πρωτεΐνη του οργανισμού που θέλουμε*

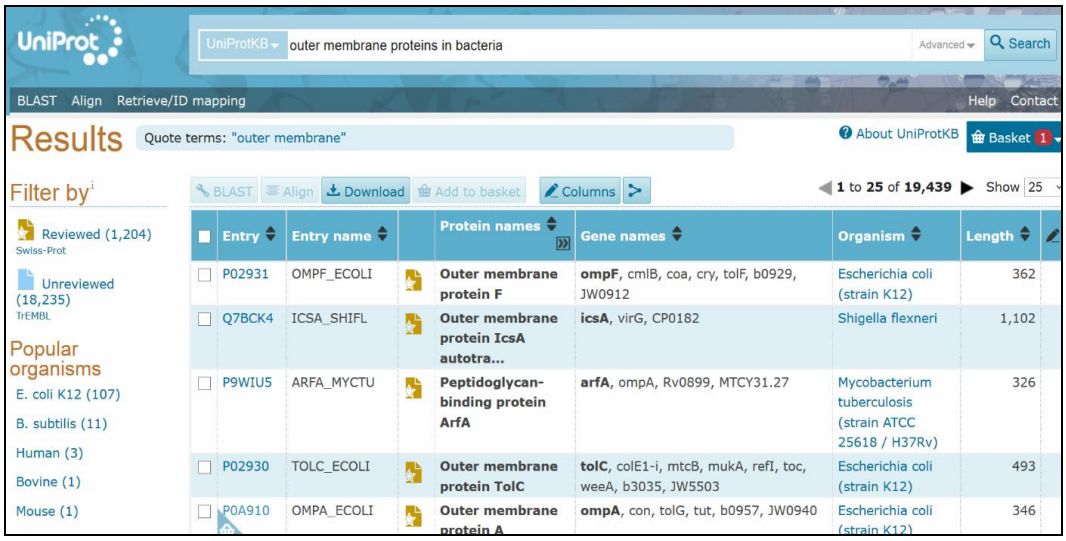


Εικόνα 2.13: *Η επιλογή για να μεταφορτώσουμε όλη την καταχώρηση*

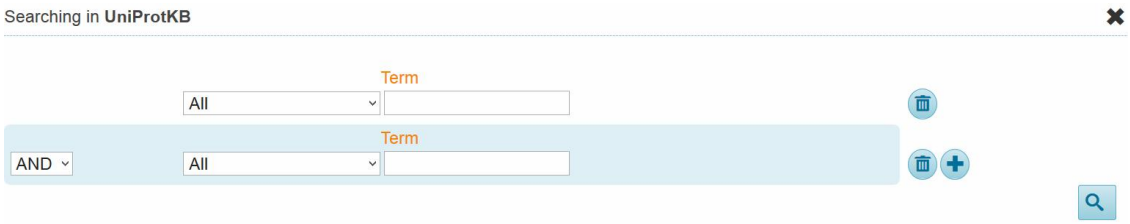
2. Να γίνει αναζήτηση στη βάση δεδομένων Uniprot με σκοπό την ανεύρεση των πρωτεινων της εξωτερικής μεμβράνης (outer membrane) των βακτηρίων με γνωστή προσδιορισμένη δομή.
(η συνολική επερώτηση είναι: taxonomy:"Bacteria [2]" existence:"evidence at protein level" database:(type:pdb) locations:(location:"Cell outer membrane [SL-0040]") keyword:"Cell outer membrane [KW-0998]")



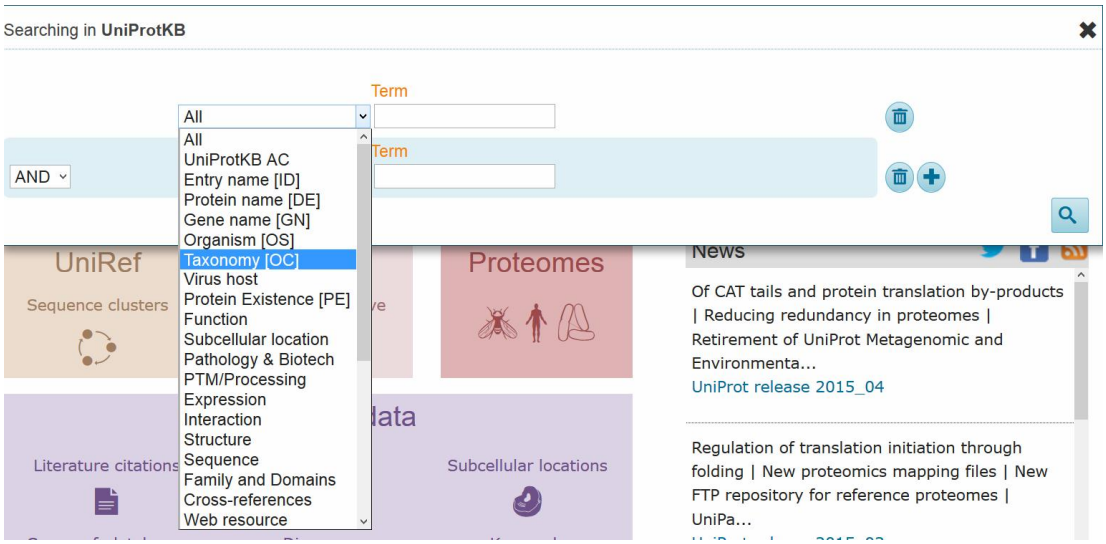
Εικόνα 2.14: *Στην αρχική σελίδα της Uniprot αν κάνουμε μια γενική επερώτηση (σε όλα τα πεδία), θα πάρουμε και πολλές άσχετες απαντήσεις*



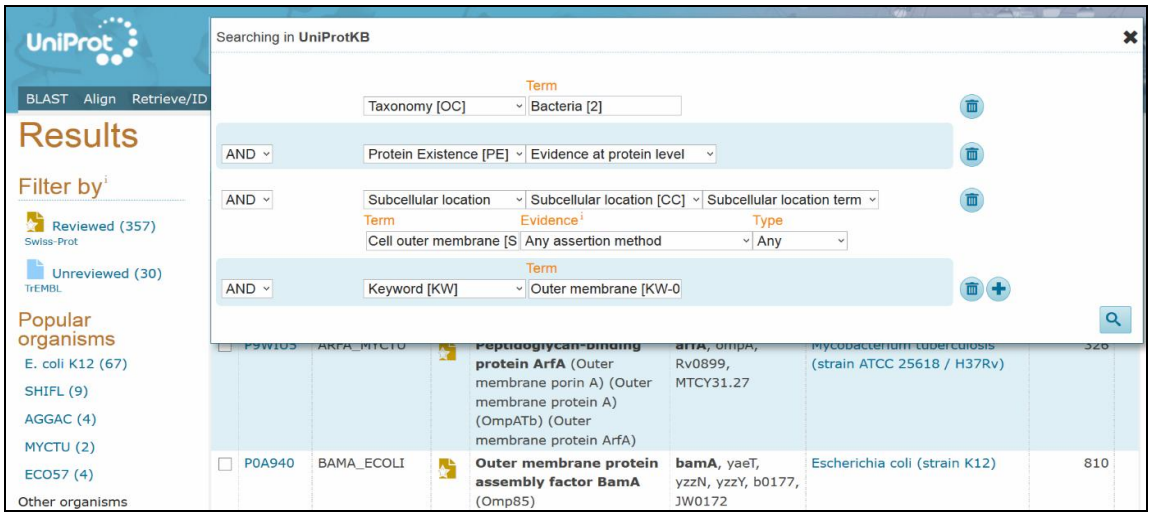
Εικόνα 2.15: *Στην αρχική σελίδα της Uniprot αν κάνουμε μια γενική επερώτηση (σε όλα τα πεδία), θα πάρουμε και πολλές άσχετες απαντήσεις*



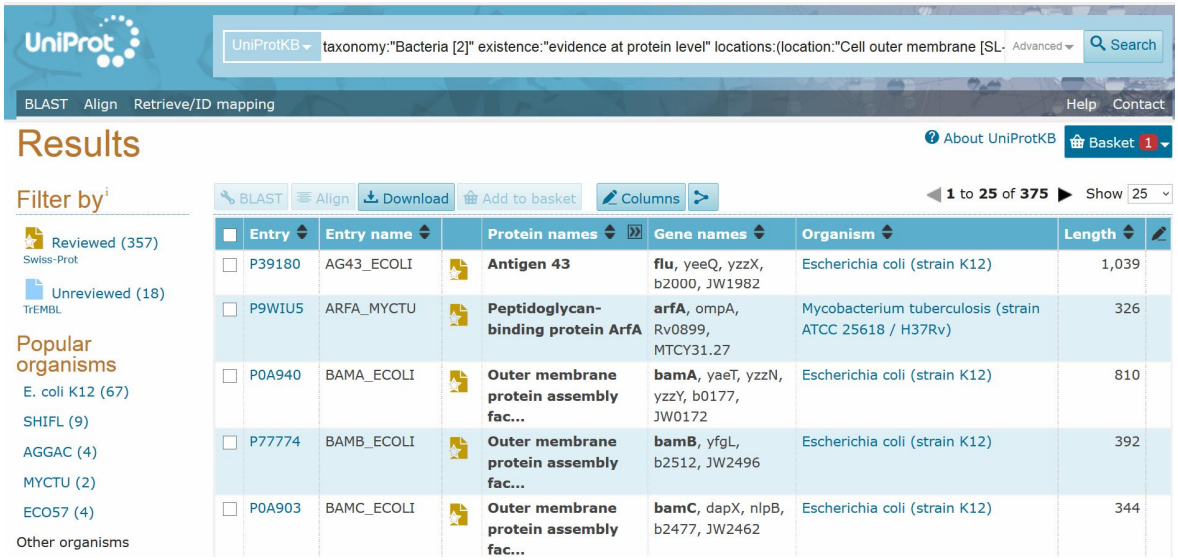
Εικόνα 2.16: *Θα πρέπει να επιλέξουμε να κάνουμε επερωτήσεις για κάθε πεδίο ξεχωριστά*



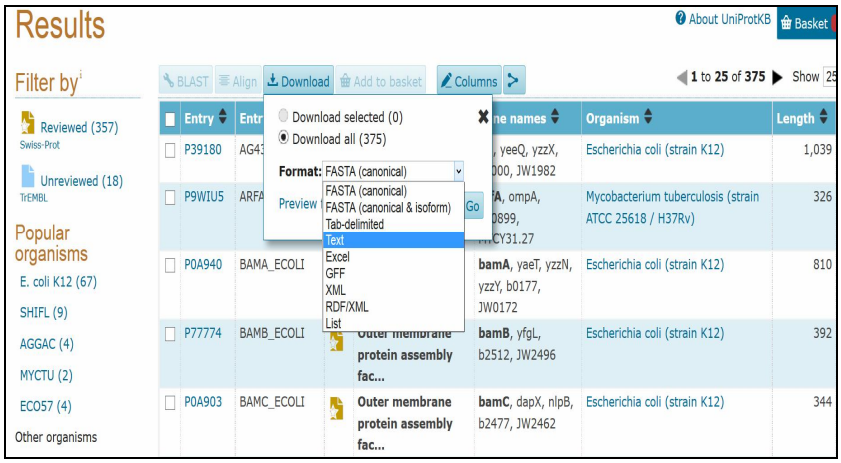
Εικόνα 2.17: *Θα πρέπει να επιλέξουμε να κάνουμε επερωτήσεις για κάθε πεδίο ξεχωριστά. Εδώ, διαλέγουμε την ταξινομική βαθμίδα του οργανισμού*



Εικόνα 2.18: *Θα πρέπει να επιλέξουμε να κάνουμε επερωτήσεις για κάθε πεδίο ξεχωριστά. Εδώ, φαίνονται και τα υπόλοιπα πεδία συμπληρωμένα*



Εικόνα 2.19: *Τα αποτελέσματα της αναζήτησης*



Εικόνα 2.20: *Η επιλογή όλων για μεταφόρτωση σε μορφή κειμένου (υπάρχουν και άλλες επιλογές)*

3. Να γίνει αναζήτηση στη βάση δεδομένων Uniprot με σκοπό την ανεύρεση ανθρώπινων υποδοχέων συζευγμένων με G-πρωτεϊνες οι οποίοι έχουν γνωστή (προσδιορισμένη) τρισδιάστατη δομή:

Η συνολική επερώτηση είναι:
taxonomy:"keyword:"G-protein coupled receptor [KW-0297]" AND organism:"Human [9606]" AND existence:"evidence at protein level" AND database:(type:pdb)

Ποιες από τις πρωτεΐνες έχουν δομή στη βάση δεδομένων PDB; Για τις παραπάνω πρωτεΐνες, να σημειωθεί ποιες αντιστοιχούν στην περιοχή της πρωτεΐνης στην οποία βρίσκεται το σύνολο των διαμεμβρανικών τμημάτων, ποιες αντιστοιχούν σε μέρος της περιοχής των διαμεμβρανικών τμημάτων και ποιες δεν περιλαμβάνουν κανένα τμήμα της αλληλουχίας το οποίο να αντιστοιχεί σε αλληλουχία διαμεμβρανικών τμημάτων.

Παράρτημα (Παραδείγματα από τις βάσεις δεδομένων)

1. Εγγραφή της GENBANK για το γονίδιο της πρωτεΐνης Outer membrane protein A (ompA) από τον οργανισμό Escherichia coli.

```
LOCUS       NC_000913               1041 bp    DNA     linear   CON 16-DEC-2014

DEFINITION  Escherichia coli str. K-12 substr. MG1655, complete genome.

ACCESSION   NC_000913 REGION: complement(1019013..1020053)
VERSION     NC_000913.3  GI:556503834
DBLINK      BioProject: PRJNA57779
            BioSample: SAMN02604091
KEYWORDS    RefSeq.
SOURCE      Escherichia coli str. K-12 substr. MG1655
  ORGANISM  Escherichia coli str. K-12 substr. MG1655
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 1041)
  AUTHORS   Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R.,
            Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T.,
            Mori,H., Perna,N.T., Plunkett,G. III, Rudd,K.E., Serres,M.H.,
            Thomas,G.H., Thomson,N.R., Wishart,D. and Wanner,B.L.
  TITLE     Escherichia coli K-12: a cooperatively developed annotation
            snapshot--2005
  JOURNAL   Nucleic Acids Res. 34 (1), 1-9 (2006)
   PUBMED   16397293
  REMARK    Publication Status: Online-Only
REFERENCE   2  (bases 1 to 1041)
  AUTHORS   Hayashi,K., Morooka,N., Yamamoto,Y., Fujita,K., Isono,K., Choi,S.,
            Ohtsubo,E., Baba,T., Wanner,B.L., Mori,H. and Horiuchi,T.
  TITLE     Highly accurate genome sequences of Escherichia coli K-12 strains
            MG1655 and W3110
  JOURNAL   Mol. Syst. Biol. 2, 2006 (2006)
   PUBMED   16738553
REFERENCE   3  (bases 1 to 1041)
  AUTHORS   Blattner,F.R., Plunkett,G. III, Bloch,C.A., Perna,N.T., Burland,V.,
            Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F.,
            Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J.,
            Mau,B. and Shao,Y.
  TITLE     The complete genome sequence of Escherichia coli K-12
  JOURNAL   Science 277 (5331), 1453-1462 (1997)
   PUBMED   9278503
REFERENCE   4  (bases 1 to 1041)
  AUTHORS   Arnaud,M., Berlyn,M.K.B., Blattner,F.R., Galperin,M.Y.,
            Glasner,J.D., Horiuchi,T., Kosuge,T., Mori,H., Perna,N.T.,
            Plunkett,G. III, Riley,M., Rudd,K.E., Serres,M.H., Thomas,G.H. and
            Wanner,B.L.
  TITLE     Workshop on Annotation of Escherichia coli K-12
  JOURNAL   Unpublished
  REMARK    Woods Hole, Mass., on 14-18 November 2003 (sequence corrections)
REFERENCE   5  (bases 1 to 1041)
  AUTHORS   Glasner,J.D., Perna,N.T., Plunkett,G. III, Anderson,B.D.,
            Bockhorst,J., Hu,J.C., Riley,M., Rudd,K.E. and Serres,M.H.
  TITLE     ASAP: Escherichia coli K-12 strain MG1655 version m56
  JOURNAL   Unpublished
  REMARK    ASAP download 10 June 2004 (annotation updates)
REFERENCE   6  (bases 1 to 1041)
  AUTHORS   Hayashi,K., Morooka,N., Mori,H. and Horiuchi,T.
  TITLE     A more accurate sequence comparison between genomes of Escherichia
            coli K12 W3110 and MG1655 strains
```

```
  JOURNAL   Unpublished
  REMARK    GenBank accessions AG613214 to AG613378 (sequence corrections)
REFERENCE   7  (bases 1 to 1041)
  AUTHORS   Perna,N.T.
  TITLE     Escherichia coli K-12 MG1655 yqiK-rfaE intergenic region, genomic
            sequence correction
  JOURNAL   Unpublished
  REMARK    GenBank accession AY605712 (sequence corrections)
REFERENCE   8  (bases 1 to 1041)
  AUTHORS   Rudd,K.E.
  TITLE     A manual approach to accurate translation start site annotation: an
            E. coli K-12 case study
  JOURNAL   Unpublished
REFERENCE   9  (bases 1 to 1041)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (26-AUG-2014) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
REFERENCE   10 (bases 1 to 1041)
  AUTHORS   Blattner,F.R. and Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (30-JUL-2014) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
  REMARK    Protein update by submitter
REFERENCE   11 (bases 1 to 1041)
  AUTHORS   Blattner,F.R. and Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (15-NOV-2013) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
  REMARK    Protein update by submitter
REFERENCE   12 (bases 1 to 1041)
  AUTHORS   Blattner,F.R. and Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-SEP-2013) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
  REMARK    Sequence update by submitter
REFERENCE   13 (bases 1 to 1041)
  AUTHORS   Rudd,K.E.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-FEB-2013) Department of Biochemistry and Molecular
            Biology, University of Miami Miller School of Medicine, 118 Gautier
            Bldg., Miami, FL 33136, USA
  REMARK    Sequence update by submitter
REFERENCE   14 (bases 1 to 1041)
  AUTHORS   Rudd,K.E.
  TITLE     Direct Submission
  JOURNAL   Submitted (24-APR-2007) Department of Biochemistry and Molecular
            Biology, University of Miami Miller School of Medicine, 118 Gautier
            Bldg., Miami, FL 33136, USA
  REMARK    Annotation update from ecogene.org as a multi-database
            collaboration
REFERENCE   15 (bases 1 to 1041)
  AUTHORS   Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (07-FEB-2006) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
  REMARK    Protein updates by submitter
REFERENCE   16 (bases 1 to 1041)
  AUTHORS   Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (10-JUN-2004) Laboratory of Genetics, University of
```

```
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
  REMARK    Sequence update by submitter
REFERENCE   17 (bases 1 to 1041)
  AUTHORS   Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (13-OCT-1998) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
REFERENCE   18 (bases 1 to 1041)
  AUTHORS   Blattner,F.R. and Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-SEP-1997) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
REFERENCE   19 (bases 1 to 1041)
  AUTHORS   Blattner,F.R. and Plunkett,G. III.
  TITLE     Direct Submission
  JOURNAL   Submitted (16-JAN-1997) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence is identical to U00096.
            On Nov 3, 2013 this sequence version replaced gi:49175990.
            RefSeq Category: Reference Genome
                        FGS: First Genome sequenced
                        MOD: Model Organism
                        PHY: Based on Phylogenetics
                        UPR: UniProt Genome
            Current U00096 annotation updates are derived from EcoGene
            http://ecogene.org. Suggestions for updates can be sent to Dr.
            Kenneth Rudd (krudd@miami.edu). These updates are being generated
            from a collaboration that also includes ASAP/ERIC, the Coli Genetic
            Stock Center, EcoliHub, EcoCyc, RegulonDB and UniProtKB/Swiss-Prot.
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..1041
                     /organism="Escherichia coli str. K-12 substr. MG1655"
                     /mol_type="genomic DNA"
                     /strain="K-12"
                     /sub_strain="MG1655"
                     /db_xref="taxon:511145"
     gene            1..1041
                     /gene="ompA"
                     /locus_tag="b0957"
                     /gene_synonym="con; ECK0948; JW0940; tolG; tut"
                     /db_xref="EcoGene:EG10669"
                     /db_xref="GeneID:945571"
     CDS             1..1041
                     /gene="ompA"
                     /locus_tag="b0957"
                     /gene_synonym="con; ECK0948; JW0940; tolG; tut"
                     /function="membrane; Outer membrane constituents"
                     /GO_component="GO:0009279 - cell outer membrane;
                     GO:0009274 - peptidoglycan-based cell wall"
                     /note="outer membrane protein 3a (II*;G;d)"
                     /codon_start=1
                     /transl_table=11
                     /product="outer membrane protein A (3a;II*;G;d)"
                     /protein_id="NP_415477.1"
                     /db_xref="GI:16128924"
                     /db_xref="ASAP:ABE-0003240"
                     /db_xref="UniProtKB/Swiss-Prot:P0A910"
                     /db_xref="EcoGene:EG10669"
                     /db_xref="GeneID:945571"
```

```
                     /translation="MKKTAIAIAVALAGFATVAQAAPKDNTWYTGAKLGWSQYHDTGF
                     INNNGPTHENQLGAGAFGGYQVNPYVGFEMGYDWLGRMPYKGSVENGAYKAQGVQLTA
                     KLGYPITDDLDIYTRLGGMVWRADTKSNVYGKNHDTGVSPVFAGGVEYAITPEIATRL
                     EYQWTNNIGDAHTIGTRPDNGMLSLGVSYRFGQGEAAPVVAPAPAPAPEVQTKHFTLK
                     SDVLFNFNKATLKPEGQAALDQLYSQLSNLDPKDGSVVVLGYTDRIGSDAYNQGLSER
                     RAQSVVDYLISKGIPADKISARGMGESNPVTGNTCDNVKQRAALIDCLAPDRRVEIEV
                     KGIKDVVTQPQA"
ORIGIN
        1 atgaaaaaga cagctatcgc gattgcagtg gcactggctg gtttcgctac cgtagcgcag
       61 gccgctccga aagataacac ctggtacact ggtgctaaac tgggctggtc ccagtaccat
      121 gacactggtt tcatcaacaa caatggcccg acccatgaaa accaactggg cgctggtgct
      181 tttggtggtt accaggttaa cccgtatgtt ggctttgaaa tgggttacga ctggttaggt
      241 cgtatgccgt acaaaggcag cgttgaaaac ggtgcataca aagctcaggg cgttcaactg
      301 accgctaaac tgggttaccc aatcactgac gacctggaca tctacactcg tctgggtggc
      361 atggtatggc gtgcagacac taaatccaac gtttatggta aaaaccacga caccggcgtt
      421 tctccggtct tcgctggcgg tgttgagtac gcgatcactc ctgaaatcgc tacccgtctg
      481 gaataccagt ggaccaacaa catcggtgac gcacacacca tcggcactcg tccggacaac
      541 ggcatgctga gcctgggtgt ttcctaccgt ttcggtcagg gcgaagcagc tccagtagtt
      601 gctccggctc cagctccggc accggaagta cagaccaagc acttcactct gaagtctgac
      661 gttctgttca acttcaacaa agcaaccctg aaaccggaag gtcaggctgc tctggatcag
      721 ctgtacagcc agctgagcaa cctggatccg aaagacggtt ccgtagttgt tctgggttac
      781 accgaccgca tcggttctga cgcttacaac cagggtctgt ccgagcgccg tgctcagtct
      841 gttgttgatt acctgatctc caaaggtatc ccggcagaca agatctccgc acgtggtatg
      901 ggcgaatcca acccggttac tggcaacacc tgtgacaacg tgaaacagcg tgctgcactg
      961 atcgactgcc tggctccgga tcgtcgcgta gagatcgaag ttaaaggtat caaagacgtt
     1021 gtaactcagc cgcaggctta a
//
```

Επεξηγήσεις των σημαντικότερων πεδίων μιας εγγραφής στην GENBANK

LOCUS: Περιέχει ένα μικρό όνομα για τον χαρακτηρισμό της εγγραφής.

DEFINITION: Μια λεπτομερής περιγραφή της αλληλουχίας.

ACCESSION: Κωδικός που αποκτά μια νεοεισερχόμενη εγγραφή χαρακτηριστικός για την GENBANK. Ο κωδικός παραμένει σταθερός

VERSION: Ειδικός κωδικός που απαρτίζεται από το πρωταρχικό Accession Number, ακολουθεί το σύμβολο της τελείας και στη συνέχεια ένας αριθμός που δηλώνει την έκδοση της παρούσας εγγραφής.

KEYWORDS: Χαρακτηριστικές λέξεις-κλειδιά που σχετίζονται με την νουκλεοτιδική αλληλουχία και τις ιδιότητες των προϊόντων της.

SOURCE: Βιολογική πηγή της αλληλουχίας όπου αναφέρεται ο οργανισμός από τον οποίο έχει απομονωθεί με τα ιδιαίτερα χαρακτηριστικά του (πιθανές μεταλλάξεις, πλασμίδια κ.α.).

ORGANISM: Οργανισμός απ' όπου προήλθε η αλληλουχία. Ακολουθείται η διώνυμη ονομασία κατά Λινναίο. Επίσης παρατίθεται και η συστηματική ταξινόμηση του οργανισμού.

- Τα παρακάτω πεδία σχετίζονται με την δημοσιευμένη εργασία στην οποία αναφέρεται ο προσδιορισμός της παρούσας αλληλουχίας.

REFERENCE: Περιέχει τον αριθμό της αναφοράς καθώς και το μήκος της αλληλουχίας που έχει προσδιοριστεί στην παρούσα εργασία.

AUTHORS: Αναφέρονται οι συμμετέχοντες στην διεξαγωγή της παρούσας εργασίας.

TITLE: Τίτλος της δημοσιευμένης εργασίας.

JOURNAL: Περιέχει λεπτομέρειακα στοιχεία για την αναζήτηση της αναφοράς όπως είναι ο τίτλος του περιοδικού που εκδόθηκε, τεύχος, ημερομηνία έκδοσης και σελίδες που καταλαμβάνει στο συγκεκριμένο τεύχος.

MEDLINE: Κωδικός για την βιβλιογραφική αναφορά στην βάση δεδομένων MEDLINE.

COMMENT: Περιέχει κάποιες γενικές παρατηρήσεις, ή αναφορές και σε άλλες βάσεις.

FEATURES: Πίνακας που περιέχει πληροφορίες σχετικά με τα προϊόντα της αλληλουχίας όπως πολυπεπτιδικές αλυσίδες (από μετάφραση) και RNA (από μεταγραφή) και στοιχεία από πειραματικά δεδομένα που καταδεικνύουν τη βιολογική της σημασία.

BASE COUNT: Αριθμητική ανάλυση της αλληλουχίας στα επιμέρους συστατικά της. Περιέχει το σύνολο καταλοίπων Αδενίνης, Γουανίνης, Κυτοσίνης, Θυμίνης.

ORIGIN: Θέση της πρώτης βάσης της κατατεθειμένης αλληλουχίας σε σχέση με το γονιδίωμα από το οποίο έχει απομονωθεί.

Ακριβώς από κάτω παρατίθεται η αλληλουχία της παρούσας εγγραφής.

Η αναπαράσταση της αλληλουχίας είναι της μορφής:

```
ORIGIN
        1 atgaaaaaga cagctatcgc gattgcagtg gcactggctg gtttcgctac cgtagcgcag
       61 gccgctccga aagataacac ctggtacact ggtgctaaac tgggctggtc ccagtaccat
      121 gacactggtt tcatcaacaa caatggcccg acccatgaaa accaactggg cgctggtgct
      181 tttggtggtt accaggttaa cccgtatgtt ggctttgaaa tgggttacga ctggttaggt
      241 cgtatgccgt acaaaggcag cgttgaaaac ggtgcataca aagctcaggg cgttcaactg
      301 accgctaaac tgggttaccc aatcactgac gacctggaca tctacactcg tctgggtggc
      361 atggtatggc gtgcagacac taaatccaac gtttatggta aaaaccacga caccggcgtt
      421 tctccggtct tcgctggcgg tgttgagtac gcgatcactc ctgaaatcgc tacccgtctg
      481 gaataccagt ggaccaacaa catcggtgac gcacacacca tcggcactcg tccggacaac
      541 ggcatgctga gcctgggtgt ttcctaccgt ttcggtcagg gcgaagcagc tccagtagtt
      601 gctccggctc cagctccggc accggaagta cagaccaagc acttcactct gaagtctgac
      661 gttctgttca acttcaacaa agcaaccctg aaaccggaag gtcaggctgc tctggatcag
      721 ctgtacagcc agctgagcaa cctggatccg aaagacggtt ccgtagttgt tctgggttac
      781 accgaccgca tcggttctga cgcttacaac cagggtctgt ccgagcgccg tgctcagtct
      841 gttgttgatt acctgatctc caaaggtatc ccggcagaca agatctccgc acgtggtatg
      901 ggcgaatcca acccggttac tggcaacacc tgtgacaacg tgaaacagcg tgctgcactg
      961 atcgactgcc tggctccgga tcgtcgcgta gagatcgaag ttaaaggtat caaagacgtt
     1021 gtaactcagc cgcaggctta a
//
```

- Τα νουκλεοτίδια απεικονίζονται με τον κώδικα ενός γράμματος ανάλογα με την αζωτούχο βάση την οποία αποτελούνται.

- Κάθε αλληλουχία αποτελείται από 60 αμινοξικά κατάλοιπα ανά γραμμή, σε ομάδες των δέκα αμινοξικών καταλοίπων, ξεκινώντας πάντα από την θέση 11 της γραμμής. Οι ομάδες των 10 καταλοίπων χωρίζονται μεταξύ τους με κενό διάστημα.

- Από τη θέση 9 της γραμμής και προς τα αριστερά υπάρχει ένας αριθμός που δείχνει την αρίθμηση του πρώτου καταλοίπου κάθε γραμμής.

//: Λήξη της εγγραφής.

2. Εγγραφή της Uniprot για την πρωτεϊνική αλληλουχία της Outer membrane protein A (ompA) από τον οργανισμό Escherichia coli.

```
ID   OMPA_ECOLI              Reviewed;         346 AA.
AC   P0A910; P02934;
DT   20-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT   20-JUL-1986, sequence version 1.
DT   06-JAN-2015, entry version 99.
DE   RecName: Full=Outer membrane protein A;
DE   AltName: Full=Outer membrane protein II*;
DE   Flags: Precursor;
GN   Name=ompA; Synonyms=con, tolG, tut; OrderedLocusNames=b0957, JW0940;
OS   Escherichia coli (strain K12).
OC   Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
OC   Enterobacteriaceae; Escherichia.
OX   NCBI_TaxID=83333;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RC   STRAIN=K12;
RX   PubMed=6253901; DOI=10.1093/nar/8.13.3011;
RA   Beck E., Bremer E.;
RT   "Nucleotide sequence of the gene ompA coding the outer membrane
RT   protein II of Escherichia coli K-12.";
RL   Nucleic Acids Res. 8:3011-3027(1979).
RN   [2]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RC   STRAIN=K12;
RX   PubMed=6260961; DOI=10.1016/0022-2836(80)90193-X;
RA   Movva N.R., Nakamura K., Inouye M.;
RT   "Gene structure of the OmpA protein, a major surface protein of
RT   Escherichia coli required for cell-cell interaction.";
RL   J. Mol. Biol. 143:317-328(1979).
RN   [3]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RC   STRAIN=K12 / W3110 / ATCC 27325 / DSM 5911;
RX   PubMed=8905232; DOI=10.1093/dnares/3.3.137;
RA   Oshima T., Aiba H., Baba T., Fujita K., Hayashi K., Honjo A.,
RA   Ikemoto K., Inada T., Itoh T., Kajihara M., Kanai K., Kashimoto K.,
RA   Kimura S., Kitagawa M., Makino K., Masuda S., Miki T., Mizobuchi K.,
RA   Mori H., Motomura K., Nakamura Y., Nashimoto H., Nishio Y., Saito N.,
RA   Sampei G., Seki Y., Tagami H., Takemoto K., Wada C., Yamamoto Y.,
RA   Yano M., Horiuchi T.;
RT   "A 718-kb DNA sequence of the Escherichia coli K-12 genome
RT   corresponding to the 12.7-28.0 min region on the linkage map.";
RL   DNA Res. 3:137-155(1995).
RN   [4]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RC   STRAIN=K12 / MG1655 / ATCC 47076;
RX   PubMed=9278503; DOI=10.1126/science.277.5331.1453;
RA   Blattner F.R., Plunkett G. III, Bloch C.A., Perna N.T., Burland V.,
RA   Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F.,
RA   Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J.,
RA   Mau B., Shao Y.;
RT   "The complete genome sequence of Escherichia coli K-12.";
RL   Science 277:1453-1462(1996).
RN   [5]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RC   STRAIN=K12 / W3110 / ATCC 27325 / DSM 5911;
RX   PubMed=16738553; DOI=10.1038/msb4100049;
RA   Hayashi K., Morooka N., Yamamoto Y., Fujita K., Isono K., Choi S.,
RA   Ohtsubo E., Baba T., Wanner B.L., Mori H., Horiuchi T.;
RT   "Highly accurate genome sequences of Escherichia coli K-12 strains
RT   MG1655 and W3110.";
RL   Mol. Syst. Biol. 2:E1-E5(2005).
```

RN [6]
RP PROTEIN SEQUENCE OF 22-346.
RC STRAIN=K12;
RX PubMed=7001461; DOI=10.1073/pnas.77.8.4592;
RA Chen R., Schmidmayr W., Kramer C., Chen-Schmeisser U., Henning U.;
RT "Primary structure of major outer membrane protein II (ompA protein)
RT of Escherichia coli K-12.";
RL Proc. Natl. Acad. Sci. U.S.A. 77:4592-4596(1979).
RN [7]
RP PROTEIN SEQUENCE OF 22-34.
RC STRAIN=K12 / EMG2;
RX PubMed=9298646; DOI=10.1002/elps.1150180807;
RA Link A.J., Robison K., Church G.M.;
RT "Comparing the predicted and observed properties of proteins encoded
RT in the genome of Escherichia coli K-12.";
RL Electrophoresis 18:1259-1313(1996).
RN [8]
RP PROTEIN SEQUENCE OF 22-32.
RC STRAIN=K12 / W3110 / ATCC 27325 / DSM 5911;
RA Pasquali C., Sanchez J.-C., Ravier F., Golaz O., Hughes G.J.,
RA Frutiger S., Paquet N., Wilkins M., Appel R.D., Bairoch A.,
RA Hochstrasser D.F.;
RL Submitted (AUG-1994) to UniProtKB.
RN [9]
RP PROTEIN SEQUENCE OF 22-26.
RC STRAIN=K12 / W3110 / ATCC 27325 / DSM 5911;
RX PubMed=9629924; DOI=10.1002/elps.1150190539;
RA Molloy M.P., Herbert B.R., Walsh B.J., Tyler M.I., Traini M.,
RA Sanchez J.-C., Hochstrasser D.F., Williams K.L., Gooley A.A.;
RT "Extraction of membrane proteins by differential solubilization for
RT separation using two-dimensional gel electrophoresis.";
RL Electrophoresis 19:837-844(1997).
RN [10]
RP MUTANTS RESISTANT TO PHAGE ENTRY.
RX PubMed=6086577;
RA Morona R., Klose M., Henning U.;
RT "Escherichia coli K-12 outer membrane protein (OmpA) as a
RT bacteriophage receptor: analysis of mutant genes expressing altered
RT proteins.";
RL J. Bacteriol. 159:570-578(1983).
RN [11]
RP MUTANTS RESISTANT TO PHAGE ENTRY.
RX PubMed=3902787;
RA Morona R., Kramer C., Henning U.;
RT "Bacteriophage receptor area of outer membrane protein OmpA of
RT Escherichia coli K-12.";
RL J. Bacteriol. 164:539-543(1984).
RN [12]
RP PORIN ACTIVITY.
RC STRAIN=K12;
RX PubMed=1370823;
RA Sugawara E., Nikaido H.;
RT "Pore-forming activity of OmpA protein of Escherichia coli.";
RL J. Biol. Chem. 267:2507-2511(1991).
RN [13]
RP SUBCELLULAR LOCATION.
RX PubMed=7813480; DOI=10.1111/j.1432-1033.1994.00891.x;
RA Kuhn A., Kiefer D., Koehne C., Zhu H.-Y., Tschantz W.R., Dalbey R.E.;
RT "Evidence for a loop-like insertion mechanism of pro-Omp A into the
RT inner membrane of Escherichia coli.";
RL Eur. J. Biochem. 226:891-897(1993).

```
RN   [14]
RP   TOPOLOGY.
RX   PubMed=8106193;
RA   Gromiha M.M., Ponnuswamy P.K.;
RT   "Prediction of transmembrane beta-strands from hydrophobic
RT   characteristics of proteins.";
RL   Int. J. Pept. Protein Res. 42:420-431(1992).
RN   [15]
RP   IDENTIFICATION BY 2D-GEL.
RX   PubMed=9298644; DOI=10.1002/elps.1150180805;
RA   VanBogelen R.A., Abshire K.Z., Moldover B., Olson E.R.,
RA   Neidhardt F.C.;
RT   "Escherichia coli proteome analysis using the gene-protein database.";
RL   Electrophoresis 18:1243-1251(1996).
RN   [16]
RP   TOPOLOGY.
RX   PubMed=10368142;
RA   Koebnik R.;
RT   "Structural and functional roles of the surface-exposed loops of the
RT   beta-barrel membrane protein OmpA from Escherichia coli.";
RL   J. Bacteriol. 181:3688-3694(1998).
RN   [17]
RP   DIMERIZATION, AND SUBCELLULAR LOCATION.
RC   STRAIN=BL21-DE3;
RX   PubMed=16079137; DOI=10.1074/jbc.M506479200;
RA   Stenberg F., Chovanec P., Maslen S.L., Robinson C.V., Ilag L.,
RA   von Heijne G., Daley D.O.;
RT   "Protein complexes of the Escherichia coli cell envelope.";
RL   J. Biol. Chem. 280:34409-34419(2004).
RN   [18]
RP   SUBCELLULAR LOCATION.
RC   STRAIN=K12 / MG1655 / ATCC 47076;
RX   PubMed=21778229; DOI=10.1074/jbc.M111.245696;
RA   Fontaine F., Fuchs R.T., Storz G.;
RT   "Membrane localization of small proteins in Escherichia coli.";
RL   J. Biol. Chem. 286:32464-32474(2010).
RN   [19]
RP   X-RAY CRYSTALLOGRAPHY (2.5 ANGSTROMS) OF 22-192.
RX   PubMed=9808047; DOI=10.1038/2983;
RA   Pautsch A., Schulz G.E.;
RT   "Structure of the outer membrane protein A transmembrane domain.";
RL   Nat. Struct. Biol. 5:1013-1017(1997).
RN   [20]
RP   X-RAY CRYSTALLOGRAPHY (1.65 ANGSTROMS).
RX   PubMed=10764596; DOI=10.1006/jmbi.2000.3671;
RA   Pautsch A., Schulz G.E.;
RT   "High-resolution structure of the OmpA membrane domain.";
RL   J. Mol. Biol. 298:273-282(1999).
RN   [21]
RP   STRUCTURE BY NMR OF 22-197.
RX   PubMed=11276254; DOI=10.1038/86214;
RA   Arora A., Abildgaard F., Bushweller J.H., Tamm L.K.;
RT   "Structure of outer membrane protein A transmembrane domain by NMR
RT   spectroscopy.";
RL   Nat. Struct. Biol. 8:334-338(2000).
RN   [22]
RP   MASS SPECTROMETRY.
RX   PubMed=10757971; DOI=10.1021/bi000150m;
RA   le Coutre J., Whitelegge J.P., Gross A., Turk E., Wright E.M.,
RA   Kaback H.R., Faull K.F.;
RT   "Proteomics on full-length membrane proteins using mass
```

```
RT   spectrometry.";
RL   Biochemistry 39:4237-4242(1999).
CC   -!- FUNCTION: Required for the action of colicins K and L and for the
CC       stabilization of mating aggregates in conjugation. Serves as a
CC       receptor for a number of T-even like phages. Also acts as a porin
CC       with low permeability that allows slow penetration of small
CC       solutes.
CC   -!- SUBUNIT: Homodimer.
CC   -!- INTERACTION:
CC       P0C0V0:degP; NbExp=5; IntAct=EBI-371347, EBI-547165;
CC       P0A850:tig; NbExp=3; IntAct=EBI-371347, EBI-544862;
CC   -!- SUBCELLULAR LOCATION: Cell outer membrane
CC       {ECO:0000269|PubMed:16079137, ECO:0000269|PubMed:21778229,
CC       ECO:0000269|PubMed:7813480}; Multi-pass membrane protein
CC       {ECO:0000269|PubMed:16079137, ECO:0000269|PubMed:21778229,
CC       ECO:0000269|PubMed:7813480}.
CC   -!- MASS SPECTROMETRY: Mass=35177; Method=Electrospray; Range=22-346;
CC       Evidence={ECO:0000269|PubMed:10757971;
CC   -!- SIMILARITY: Belongs to the OmpA family. {ECO:0000305}.
CC   -!- SIMILARITY: Contains 1 OmpA-like domain. {ECO:0000255|PROSITE-
CC       ProRule:PRU00473}.
CC   ---------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   ---------------------------------------------------------------------
DR   EMBL; V00307; CAA23588.1; -; Genomic_DNA.
DR   EMBL; U00096; AAC74043.1; -; Genomic_DNA.
DR   EMBL; AP009048; BAA35715.1; -; Genomic_DNA.
DR   PIR; A93707; MMECA.
DR   RefSeq; NP_415477.1; NC_000913.3.
DR   RefSeq; YP_489229.1; NC_007779.1.
DR   PDB; 1BXW; X-ray; 2.50 A; A=21-192.
DR   PDB; 1G90; NMR; -; A=22-197.
DR   PDB; 1QJP; X-ray; 1.65 A; A=22-192.
DR   PDB; 2GE4; NMR; -; A=22-197.
DR   PDB; 2JMM; NMR; -; A=23-197.
DR   PDB; 3NB3; EM; -; A/B/C=1-346.
DR   PDBsum; 1BXW; -.
DR   PDBsum; 1G90; -.
DR   PDBsum; 1QJP; -.
DR   PDBsum; 2GE4; -.
DR   PDBsum; 2JMM; -.
DR   PDBsum; 3NB3; -.
DR   ProteinModelPortal; P0A910; -.
DR   SMR; P0A910; 22-192, 209-346.
DR   DIP; DIP-31879N; -.
DR   IntAct; P0A910; 11.
DR   MINT; MINT-1308131; -.
DR   STRING; 511145.b0957; -.
DR   TCDB; 1.B.6.1.1; the ompa-ompf porin (oop) family.
DR   SWISS-2DPAGE; P0A910; -.
DR   PaxDb; P0A910; -.
DR   PRIDE; P0A910; -.
DR   EnsemblBacteria; AAC74043; AAC74043; b0957.
DR   EnsemblBacteria; BAA35715; BAA35715; BAA35715.
DR   GeneID; 12931038; -.
DR   GeneID; 945571; -.
DR   KEGG; ecj:Y75_p0929; -.
DR   KEGG; eco:b0957; -.
DR   PATRIC; 32117133; VBIEscCol129921_0991.
DR   EchoBASE; EB0663; -.
```

```
DR   EcoGene; EG10669; ompA.
DR   eggNOG; COG2885; -.
DR   HOGENOM; HOG000274199; -.
DR   InParanoid; P0A910; -.
DR   KO; K03286; -.
DR   OMA; EYALTKN; -.
DR   OrthoDB; EOG6PP9QB; -.
DR   BioCyc; EcoCyc:EG10669-MONOMER; -.
DR   BioCyc; ECOL316407:JW0940-MONOMER; -.
DR   EvolutionaryTrace; P0A910; -.
DR   PRO; PR:P0A910; -.
DR   Proteomes; UP000000318; Chromosome.
DR   Proteomes; UP000000625; Chromosome.
DR   Genevestigator; P0A910; -.
DR   GO; GO:0009279; C:cell outer membrane; IDA:EcoliWiki.
DR   GO; GO:0016021; C:integral component of membrane; IDA:EcoliWiki.
DR   GO; GO:0016020; C:membrane; IDA:EcoliWiki.
DR   GO; GO:0019867; C:outer membrane; IDA:EcoliWiki.
DR   GO; GO:0046930; C:pore complex; IEA:UniProtKB-KW.
DR   GO; GO:0015288; F:porin activity; IDA:EcoCyc.
DR   GO; GO:0005198; F:structural molecule activity; IEA:InterPro.
DR   GO; GO:0006974; P:cellular response to DNA damage stimulus; IEP:EcoliWiki.
DR   GO; GO:0000746; P:conjugation; IMP:EcoliWiki.
DR   GO; GO:0009597; P:detection of virus; IMP:EcoliWiki.
DR   GO; GO:0034220; P:ion transmembrane transport; IDA:EcoCyc.
DR   GO; GO:0006811; P:ion transport; IDA:EcoliWiki.
DR   GO; GO:0006810; P:transport; IDA:EcoliWiki.
DR   GO; GO:0046718; P:viral entry into host cell; IMP:EcoliWiki.
DR   Gene3D; 2.40.160.20; -; 1.
DR   Gene3D; 3.30.1330.60; -; 1.
DR   InterPro; IPR011250; OMP/PagP_b-brl.
DR   InterPro; IPR006664; OMP_bac.
DR   InterPro; IPR002368; OmpA.
DR   InterPro; IPR006690; OMPA-like_CS.
DR   InterPro; IPR000498; OmpA-like_TM_dom.
DR   InterPro; IPR006665; OmpA/MotB_C.
DR   Pfam; PF00691; OmpA; 1.
DR   Pfam; PF01389; OmpA_membrane; 1.
DR   PRINTS; PR01021; OMPADOMAIN.
DR   PRINTS; PR01022; OUTRMMBRANEA.
DR   SUPFAM; SSF103088; SSF103088; 1.
DR   SUPFAM; SSF56925; SSF56925; 1.
DR   PROSITE; PS01068; OMPA_1; 1.
DR   PROSITE; PS51123; OMPA_2; 1.
PE   1: Evidence at protein level;
KW   3D-structure; Cell outer membrane; Complete proteome; Conjugation;
KW   Direct protein sequencing; Disulfide bond; Ion transport; Membrane;
KW   Porin; Reference proteome; Repeat; Signal; Transmembrane;
KW   Transmembrane beta strand; Transport.
FT   SIGNAL        1     21       {ECO:0000269|PubMed:7001461,
FT                                ECO:0000269|PubMed:9298646,
FT                                ECO:0000269|PubMed:9629924,
FT                                ECO:0000269|Ref.8}.
FT   CHAIN        22    346       Outer membrane protein A.
FT                                /FTId=PRO_0000020094.
FT   TOPO_DOM     22     26       Periplasmic.
FT   TRANSMEM     27     37       Beta stranded.
FT   TOPO_DOM     38     54       Extracellular.
FT   TRANSMEM     55     66       Beta stranded.
FT   TOPO_DOM     67     69       Periplasmic.
FT   TRANSMEM     70     78       Beta stranded.
```

```
FT   TOPO_DOM     79     95       Extracellular.
FT   TRANSMEM     96    107       Beta stranded.
FT   TOPO_DOM    108    111       Periplasmic.
FT   TRANSMEM    112    124       Beta stranded.
FT   TOPO_DOM    125    137       Extracellular.
FT   TRANSMEM    138    151       Beta stranded.
FT   TOPO_DOM    152    155       Periplasmic.
FT   TRANSMEM    156    163       Beta stranded.
FT   TOPO_DOM    164    181       Extracellular.
FT   TRANSMEM    182    190       Beta stranded.
FT   TOPO_DOM    191    346       Periplasmic.
FT   REPEAT      201    202       1.
FT   REPEAT      203    204       2.
FT   REPEAT      205    206       3.
FT   REPEAT      207    208       4.
FT   DOMAIN      210    338       OmpA-like. {ECO:0000255|PROSITE-
FT                                ProRule:PRU00473}.
FT   REGION      197    208       Hinge-like.
FT   REGION      201    208       4 X 2 AA tandem repeats of A-P.
FT   DISULFID    311    323
FT   STRAND       27     37       {ECO:0000244|PDB:1QJP}.
FT   STRAND       41     43       {ECO:0000244|PDB:1G90}.
FT   STRAND       46     48       {ECO:0000244|PDB:1G90}.
FT   STRAND       50     53       {ECO:0000244|PDB:2GE4}U.
FT   STRAND       55     67       {ECO:0000244|PDB:1QJP}.
FT   STRAND       70     81       {ECO:0000244|PDB:1QJP}.
FT   STRAND       93    128       {ECO:0000244|PDB:1QJP}.
FT   STRAND      130    132       {ECO:0000244|PDB:1QJP}.
FT   STRAND      134    153       {ECO:0000244|PDB:1QJP}.
FT   STRAND      156    165       {ECO:0000244|PDB:1QJP}.
FT   TURN        172    175       {ECO:0000244|PDB:1G90}.
FT   STRAND      182    190       {ECO:0000244|PDB:1QJP}.
SQ   SEQUENCE   346 AA;  37201 MW;  195147734CDF8B04 CRC64;
     MKKTAIAIAV ALAGFATVAQ AAPKDNTWYT GAKLGWSQYH DTGFINNNGP THENQLGAGA
     FGGYQVNPYV GFEMGYDWLG RMPYKGSVEN GAYKAQGVQL TAKLGYPITD DLDIYTRLGG
     MVWRADTKSN VYGKNHDTGV SPVFAGGVEY AITPEIATRL EYQWTNNIGD AHTIGTRPDN
     GMLSLGVSYR FGQGEAAPVV APAPAPAPEV QTKHFTLKSD VLFNFNKATL KPEGQAALDQ
     LYSQLSNLDP KDGSVVVLGY TDRIGSDAYN QGLSERRAQS VVDYLISKGI PADKISARGM
     GESNPVTGNT CDNVKQRAAL IDCLAPDRRV EIEVKGIKDV VTQPQA
//
```

<u>Επεξηγήσεις των σημαντικότερων πεδίων μιας εγγραφής UNIPROT</u>

ID (Identification):

Είναι της μορφής *Entry_name data_class; molecule_type; sequence length*

Entry_name: Το όνομα της αλληλουχίας χαρακτηριστικό για τη βάση UNIPROT.

π.χ. OMPA_ECOLI. Το πρώτο τμήμα υποδηλώνει το όνομα της αλληλουχίας όπως είναι κατατεθειμένο στην βάση. Μπορεί να έχει μήκος μέχρι 4 χαρακτήρες. Το δεύτερο καθορίζει το είδος από το οποίο προέρχεται η αλληλουχία. Μπορεί να έχει μήκος μέχρι 5 χαρακτήρες.

data_class: Δηλώνει αν η εγγραφή έχει σχολιαστεί ή όχι με βάση τα κριτήρια της βάσης UNIPROT.

molecule_type: Δηλώνει σε ποια ομάδα μακρομορίων ανήκει η αλληλουχία. Για τις εγγραφές της UNIPROT είναι PRT (Protein).

sequence length: Το μήκος της αλληλουχίας σε αμινοξικά κατάλοιπα (AA).

AC (Accession number): Είναι ένας χαρακτηριστικός κωδικός που αποκτά μια πολυπεπτιδική αλυσίδα όταν κατατίθεται στην βάση. Χρησιμεύει στην αναγνώριση εγγραφών ανάμεσα στις διαφορετικές εκδόσεις της βάσης όπως αυτή ανανεώνεται ανά τακτά χρονικά διαστήματα.

DT (Date): Αναγραφή ημερομηνίας για τη δημιουργία της παρούσας εγγραφής, τελευταίας τροποποιήσης, προσθήκης σχολίων.

DE (Description): Γενική περιγραφή για την αλληλουχία.

GN (Gene name): Γονίδιο από το οποίο με μετάφραση προέκυψε η αμινοξική αλληλουχία.

OS (Organism Species): Οργανισμός απ' όπου προήλθε η αλληλουχία. Ακολουθείται η διώνυμη ονομασία κατά Λινναίο.

OG (Organelle): Επεξηγεί αν το γονίδιο που κωδικοποιεί την συγκεκριμένη αλληλουχία εδράζεται σε μιτοχόνδρια, χλωροπλάστες ή πλασμίδιο.

OC (Organism Classification): Συστηματική ταξινόμηση του οργανισμού απ'όπου προήλθε η αλληλουχία.

OX (Organism taxonomy cross-reference): Παραπομπή σε βάση δεδομένων συστηματικής ταξινόμησης των οργανισμών.

- **RN, RP, RC, RX, RA, RT, RL :** Τα παρακάτω πεδία σχετίζονται με βιβλιογραφικές αναφορές σχετικές με την παρούσα εγγραφή.

RN (Reference number): Αύξων αριθμός αναφοράς σχετικής με την παρούσα εγγραφή.

RP (Reference Position): Περιέχει λίγες πληροφορίες σχετικές με το τι πραγματεύεται η συγκεκριμένη αναφορά.

RX (Reference cross-reference): Παραπομπές σε βιβλιογραφικές βάσεις δεδομένων π.χ. PUBMED.

RA (Reference author): Λίστα με τους συγγραφείς της παρούσας αναφοράς.

RT (Reference title): Τίτλος της παρούσας εργασίας όπως δημοσιεύτηκε σε επιστημονικά περιοδικά.

RL (Reference Location): Περιοδικό ή βιβλίο όπου δημοσιεύτηκε η παρούσα εργασία.

CC (Comments): Το πεδίο αυτό περιέχει μία σειρά από πληροφορίες πάσης φύσεως σχετικές με την αλληλουχία. Χωρίζεται σε υπο-πεδία όπως:

CATALYTIC ACTIVITY: Περιγραφή της αντίδρασης που καταλύεται αν η αλληλουχία είναι ένζυμο.

ALTERNATIVE PRODUCTS: Αναφέρεται αν υπάρχουν σχετικές με αυτή αλληλουχίες που έχουν προκύψει από εναλλακτικό μάτισμα.

FUNCTION: Σύντομη περιγραφή της λειτουργίας που συμμετέχει η αλληλουχία.

SUBCELLULAR LOCATION: Θέση της αλληλουχίας στο κύτταρο.

SUBUNIT: Το πεδίο εμφανίζεται στην περίπτωση που η αλληλουχία συμμετέχει στην δημιουργία τεταρτοταγούς δομής μιας πρωτεΐνης.

Πρέπει να σημειωθεί πως τα παραπάνω είναι μερικά από τα υπο-πεδία που μπορεί να περιέχονται στο πεδίο CC (Comments).

DR (Database cross-reference): Το πεδίο αυτό δίνει διασυνδέσεις σε άλλες βάσεις δεδομένων που σχετίζονται με την παρούσα εγγραφή όπως η PDB, η EMBL κ.α. με τους αντίστοιχους κωδικούς τους.

KW (Keyword): Το πεδίο αυτό περιέχει ειδικές λέξεις-κλειδιά για τον χαρακτηρισμό της αλληλουχίας όπως αυτές ταξινομούνται με βάση κριτήρια όπως η λειτουργία και η δομή τους.

FT (Feature Table): Το πεδίο αυτό περιέχει στοιχεία χαρακτηριστικά για την αλληλουχία αυτή καθεαυτή και αφορά συγκεκριμένα τμήματά της. Περιλαμβάνει πληροφορίες για:

α. Μεταμεταφραστικές τροποποιήσεις

β. Ποια τμήματα της αλληλουχίας είναι υπεύθυνα για την δέσμευση κάποιου μορίου (π.χ. Receptor-Ligand).

γ. Ποια τμήματα της αλληλουχίας συμμετέχουν για το σχηματισμό του ενεργού κέντρου αν πρόκειται για ένζυμο.

δ. Στοιχεία για τη δευτεροταγή δομή της αλληλουχίας.

ε. Επίσης στο πεδίο αυτό μπορεί και να σημειώνονται και διαφορές στην αλληλουχία εάν έχουν προκύψει και αναφέρονται σε άλλες βιβλιογραφικές αναφορές.

SQ (Sequence): Το πεδίο αυτό περιέχει το μήκος της αλληλουχίας σε αμινοξικά κατάλοιπα (AA), το μοριακό βάρος (MW) σε Daltons.

Ακολουθεί η αναπαράσταση της αλληλουχίας ακολουθώντας τους παρακάτω κανόνες:

- Κάθε αμινοξικό κατάλοιπο απεικονίζεται με τον κώδικα του ενός γράμματος κατά IUPAC.

- Κάθε αλληλουχία αποτελείται από 60 αμινοξικά κατάλοιπα ανά γραμμή, σε ομάδες των δέκα αμινοξικών καταλοίπων, ξεκινώντας πάντα από την θέση 6 της γραμμής. Οι ομάδες των 10 καταλοίπων χωρίζονται μεταξύ τους με κενό διάστημα.

//: Τα σύμβολα αυτά υποδηλώνουν το τέλος της εγγραφής.

Π.χ.

```
SQ   SEQUENCE   346 AA;  37201 MW;  195147734CDF8B04 CRC64;
     MKKTAIAIAV ALAGFATVAQ AAPKDNTWYT GAKLGWSQYH DTGFINNNGP THENQLGAGA
     FGGYQVNPYV GFEMGYDWLG RMPYKGSVEN GAYKAQGVQL TAKLGYPITD DLDIYTRLGG
     MVWRADTKSN VYGKNHDTGV SPVFAGGVEY AITPEIATRL EYQWTNNIGD AHTIGTRPDN
     GMLSLGVSYR FGQGEAAPVV APAPAPAPEV QTKHFTLKSD VLFNFNKATL KPEGQAALDQ
     LYSQLSNLDP KDGSVVVLGY TDRIGSDAYN QGLSERRAQS VVDYLISKGI PADKISARGM
     GESNPVTGNT CDNVKQRAAL IDCLAPDRRV EIEVKGIKDV VTQPQA
//
```

3. Εγγραφή της PROSITE για την πρωτεϊνική αλληλουχία της Outer membrane protein A (ompA).

```
ID   OMPA_1; PATTERN.
```

```
AC   PS01068;
DT   NOV-1995 (CREATED); DEC-2004 (DATA UPDATE); FEB-2015 (INFO UPDATE).
DE   OmpA-like domain.
PA   [LIVMA]-x-[GT]-x-[TA]-[DAN]-x(2,3)-[DG]-[GSTPNKQ]-x(2)-[LFYDEPAVI]-[NQS]-
PA   x(2)-[LI]-[SG]-[QEA]-[KRQENAD]-R-A-x(2)-[LVAIT]-x(3)-[LIVMF]-x(4,5)-
PA   [LIVMF]-x(4)-[LIVM]-x(3)-[SGW]-x-G.
NR   /RELEASE=2015_04,548208;
NR   /TOTAL=55(55); /POSITIVE=55(55); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR   /FALSE_NEG=10; /PARTIAL=2;
CC   /TAXO-RANGE=???P?; /MAX-REPEAT=1;
CC   /VERSION=1;
DR   P65594, ARFA_MYCBO , T; A1KH31, ARFA_MYCBP , T; P9WIU4, ARFA_MYCTO , T;
DR   P9WIU5, ARFA_MYCTU , T; Q9S3P9, MOTY_VIBAN , T; P46233, MOTY_VIBPA , T;
DR   Q8U9L5, OMP16_AGRT5, T; P0A3S9, OMP16_BRUAB, T; P0A3S7, OMP16_BRUME, T;
DR   P0A3S8, OMP16_BRUSU, T; Q98F85, OMP16_RHILO, T; Q926C3, OMP16_RHIME, T;
DR   P07050, OMP3_NEIGO , T; Q9S3R8, OMP40_PORGI, T; Q9S3R9, OMP41_PORGI, T;
DR   P0A0V2, OMP4_NEIMA , T; P0A0V3, OMP4_NEIMB , T; P43840, OMP51_HAEIN, T;
DR   P38368, OMP52_HAEIF, T; P45996, OMP53_HAEIF, T; Q05146, OMPA_BORAV , T;
DR   P57414, OMPA_BUCAI , T; Q8K9L4, OMPA_BUCAP , T; P24016, OMPA_CITFR , T;
DR   P0A911, OMPA_ECO57 , T; P0A910, OMPA_ECOLI , T; P09146, OMPA_ENTAE , T;
DR   B7LNW7, OMPA_ESCF3 , T; P0C8Z2, OMPA_ESCFE , T; P24754, OMPA_ESCHE , T;
DR   P24017, OMPA_KLEPN , T; Q8Z7S0, OMPA_SALTI , T; P02936, OMPA_SALTY , T;
DR   P04845, OMPA_SERMA , T; P24755, OMPA_SEROD , T; I2BAK7, OMPA_SHIBC , T;
DR   P0DJO6, OMPA_SHIBL , T; P02935, OMPA_SHIDY , T; Q8ZG77, OMPA_YERPE , T;
DR   P38399, OMPA_YERPS , T; Q89AJ5, PAL_BUCBP  , T; P0A913, PAL_ECO57  , T;
DR   P0A912, PAL_ECOLI  , T; P10324, PAL_HAEIN  , T; P26493, PAL_LEGPN  , T;
DR   Q51886, PAL_PASMU  , T; Q9I4Z4, PAL_PSEAE  , T; P0A138, PAL_PSEPK  , T;
DR   P0A139, PAL_PSEPU  , T; P0A914, PAL_SHIFL  , T; P13794, PORF_PSEAE , T;
DR   P37726, PORF_PSEFL , T; P22263, PORF_PSESY , T; P38369, TPN50_TREPA, T;
DR   P37665, YIAD_ECOLI , T;
DR   P85410, OMP5_HAEPR , P; P80444, OMPA_ACTLI , P;
DR   D3GSC3, LAFU_ECO44 , N; Q47154, LAFU_ECOLI , N; Q6RYW5, OMP38_ACIBA, N;
DR   A3M8K2, OMP38_ACIBT, N; P84838, OMPC_GLUDA , N; P07021, YFIB_ECOLI , N;
DR   P0C536, YN58_BRUAB , N; Q2YJ83, YP57_BRUA2 , N; Q8YDY8, YU36_BRUME , N;
DR   Q9RPX3, YU58_BRUSU , N;
3D   1OAP; 1R1M; 2AIZ; 2HQS; 2K1S; 2KGW; 2L26; 2LBT; 2LCA; 2W8B;
DO   PDOC00819;
//
```

Επεξηγήσεις των σημαντικότερων πεδίων μιας εγγραφής στην PROSITE

ID (Identification): Είναι της γενικής μορφής

ID ENTRY_NAME; ENTRY_TYPE

Το πρώτο τμήμα είναι η χαρακτηριστική ονομασία που εμφανίζει η εγγραφή χαρακτηριστική για τη βάση PROSITE, ενώ το δεύτερο τμήμα υποδηλώνει τον τύπο της εγγραφής.

AC (ACcession number): Πρόκειται για τον χαρακτηριστικό κωδικό που αποκτά μια νεοεισερχόμενη εγγραφή στην PROSITE και χρησιμεύει στην αναγνώριση της εγγραφής ανάμεσα στις διαφορετικές εκδόσεις της βάσης PROSITE.

DT (DaTe): Το πεδίο αυτό περιέχει τις ημερομηνίες δημιουργίας και τελευταίας ανανέωσης (σχολιασμός) της εγγραφής.

DE (DEscription): Περιέχει μια γενική περιγραφή για την συγκεκριμένη εγγραφή.

PA (PAttern): Στο πεδίο αυτό αναγράφεται το πρότυπο της αλληλουχίας (pattern) που ακολουθούν τα μέλη της συγκεκριμένης εγγραφής.

Οι συμβάσεις που ακολουθούμε για την αναπαράσταση του pattern είναι:

1. Τα αμινοξέα απεικονίζονται με τον κώδικα του ενός γράμματος κατά IUPAC.
2. Το σύμβολο **x** σημαίνει ότι στη θέση αυτή μπορεί να υπάρχει οποιοδήποτε αμινοξύ.
3. [...] Τα αμινοξέα που περιέχονται μέσα στις αγκύλες είναι τα επιτρεπτά για τη συγκεκριμένη θέση. Για παράδειγμα αν περιέχεται στις αγκύλες [ALT] σημαίνει ότι στη συγκεκριμένη θέση επιτρέπεται να βρίσκεται Αλανίνη ή Λευκίνη ή Θρεονίνη.
4. Τα άγκιστρα υποδηλώνουν ότι όσα αμινοξέα περιέχονται σε αυτά δεν επιτρέπεται να βρίσκονται στις συγκεκριμένες θέσεις.
5. Κάθε στοιχείο του μοτίβου χωρίζεται από το γειτονικό του με μια παύλα (-).
6. Αν ένα στοιχείο επαναλαμβάνεται μπορεί να αναπαρασταθεί με ένα αριθμητικό δείκτη σε παρενθέσεις που δηλώνει τον αριθμό των επαναλήψεων π.χ. x(3). Στην περίπτωση που εντός της παρενθέσεως περιέχονται δύο αριθμοί που χωρίζονται μεταξύ τους με κόμμα τούτο σημαίνει ότι ο αριθμός των επαναλήψεων μπορεί να παίρνει ένα εύρος τιμών που καθορίζεται από τις τιμές που περιέχονται στις παρενθέσεις π.χ. (2,4) Ο αριθμός των επαναλήψεων μπορεί να είναι 2 ή 3 ή 4.
7. Αν το μοτίβο περιορίζεται στο αμινοτελικό ή το καρβοξυτελικό άκρο η αναπαράσταση ξεκινά με τα σύμβολα '<' και '>' αντίστοιχα.
8. Η τελεία υποδηλώνει το τέλος του pattern.

NR (Numerical Results): Τα πεδία αυτά περιέχουν στοιχεία που προκύπτουν από την σάρωση (pattern scan) της βάσης SWISS-PROT με το pattern της PROSITE.

Πιο συγκεκριμένα περιλαμβάνουν:

/RELEASE: Η έκδοση της UNIPROT που έχει χρησιμοποιηθεί καθώς και ο αριθμός των εγγραφών που περιέχονται σε αυτή.

/TOTAL: Συνολικός αριθμός εγγραφών της UNIPROT όπου φαίνεται να συναντάται το μοτίβο.

/POSITIVE: Αριθμός των εγγραφών που είναι βέβαιο ότι συναντάται το pattern και ανήκουν σε οικογένεια της PROSITE.

/UNKNOWN: Αριθμός των εγγραφών που πιθανά ανήκει στην οικογένεια της PROSITE.

/FALSE_POS: Εγγραφές της UNIPROT όπου εμφανίζεται το pattern αλλά δεν σχετίζονται με την συγκεκριμένη οικογένεια.

/FALSE_NEG: Αριθμός εγγραφών της UNIPROT που ανήκουν στη συγκεκριμένη οικογένεια αλλά δεν βρέθηκαν κατά τη σάρωση μοτίβου.

/PARTIAL: Αριθμός αλληλουχιών της UNIPROT που δεν είναι πλήρεις (fragments), ανήκουν στην συγκεκριμένη οικογένεια της PROSITE, αλλά δεν ανιχνεύονται από το PROSITE λόγω έλλειψης τμημάτων της αλληλουχίας.

CC (Comments): Στα υπο-πεδία του Comments περιέχονται γενικά σχόλια που σχετίζονται με την PROSITE.

DR (Database Reference): Περιέχει όλες τις εγγραφές της UNIPROT που ακολουθούν το συγκεκριμένο μοτίβο.

3D (3D Structure): Περιέχει όλες τις εγγραφές της Protein Data Bank που περιέχει τις δομές βιομακρομορίων και ακολουθούν το συγκεκριμένο μοτίβο.

DO (Documentation): Σύνδεσμος για εγγραφή που περιέχει αναλυτικά στοιχεία σχετικά με τη βιολογική λειτουργία των αλληλουχιών που περιέχουν το συγκεκριμένο μοτίβο καθώς και βιβλιογραφικές αναφορές.

//: Δηλώνει το τέλος της εγγραφής.

4. Εγγραφή της PDB για την δομή στο χώρο της Outer membrane protein A (ompA) από τον οργανισμό Escherichia coli.

```
HEADER    MEMBRANE PROTEIN                          03-OCT-98   1BXW
TITLE     OUTER MEMBRANE PROTEIN A (OMPA) TRANSMEMBRANE DOMAIN
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: PROTEIN (OUTER MEMBRANE PROTEIN A);
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: TRANSMEMBRANE DOMAIN;
COMPND   5 ENGINEERED: YES;
COMPND   6 MUTATION: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: ESCHERICHIA COLI BL21(DE3);
SOURCE   3 ORGANISM_TAXID: 469008;
SOURCE   4 STRAIN: BL21DE3;
SOURCE   5 GENE: OMPA;
SOURCE   6 EXPRESSION_SYSTEM: ESCHERICHIA COLI BL21(DE3);
SOURCE   7 EXPRESSION_SYSTEM_TAXID: 469008;
SOURCE   8 EXPRESSION_SYSTEM_STRAIN: BL21DE3;
SOURCE   9 EXPRESSION_SYSTEM_PLASMID: PET3B-171
KEYWDS    OUTER MEMBRANE, TRANSMEMBRANE PROTEIN
EXPDTA    X-RAY DIFFRACTION
AUTHOR    G.E.SCHULZ,A.PAUTSCH
REVDAT   3   24-FEB-09 1BXW    1       VERSN
REVDAT   2   22-DEC-99 1BXW    4       HEADER COMPND REMARK JRNL
REVDAT   2 2                    4       ATOM   SOURCE SEQRES
REVDAT   1   14-OCT-98 1BXW    0
JRNL        AUTH   A.PAUTSCH,G.E.SCHULZ
JRNL        TITL   STRUCTURE OF THE OUTER MEMBRANE PROTEIN A
JRNL        TITL 2 TRANSMEMBRANE DOMAIN.
JRNL        REF    NAT.STRUCT.BIOL.              V.   5  1013 1998
JRNL        REFN                   ISSN 1072-8368
JRNL        PMID   9808047
JRNL        DOI    10.1038/2983
REMARK   1
REMARK   2
REMARK   2 RESOLUTION.    2.50 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : REFMAC
REMARK   3   AUTHORS     : MURSHUDOV,VAGIN,DODSON
REMARK   3
REMARK   3  DATA USED IN REFINEMENT.
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 2.50
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 50.00
REMARK   3   DATA CUTOFF            (SIGMA(F)) : 0.000
REMARK   3   COMPLETENESS FOR RANGE        (%) : 89.0
REMARK   3   NUMBER OF REFLECTIONS             : 8328
REMARK   3
REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   CROSS-VALIDATION METHOD          : THROUGHOUT
REMARK   3   FREE R VALUE TEST SET SELECTION  : RANDOM
REMARK   3   R VALUE     (WORKING + TEST SET) : NULL
REMARK   3   R VALUE            (WORKING SET) : 0.189
REMARK   3   FREE R VALUE                     : 0.235
REMARK   3   FREE R VALUE TEST SET SIZE   (%) : 5.000
REMARK   3   FREE R VALUE TEST SET COUNT      : 404
REMARK   3
REMARK   3  NUMBER OF NON-HYDROGEN ATOMS USED IN REFINEMENT.
REMARK   3   PROTEIN ATOMS            : 1330
REMARK   3   NUCLEIC ACID ATOMS       : 0
REMARK   3   HETEROGEN ATOMS          : 21
REMARK   3   SOLVENT ATOMS            : 39
```

```
REMARK   3
REMARK   3  B VALUES.
REMARK   3   FROM WILSON PLOT           (A**2) : 49.20
REMARK   3   MEAN B VALUE      (OVERALL, A**2) : 60.40
REMARK   3   OVERALL ANISOTROPIC B VALUE.
REMARK   3    B11 (A**2) : NULL
REMARK   3    B22 (A**2) : NULL
REMARK   3    B33 (A**2) : NULL
REMARK   3    B12 (A**2) : NULL
REMARK   3    B13 (A**2) : NULL
REMARK   3    B23 (A**2) : NULL
REMARK   3
REMARK   3  ESTIMATED OVERALL COORDINATE ERROR.
REMARK   3   ESU BASED ON R VALUE                            (A): NULL
REMARK   3   ESU BASED ON FREE R VALUE                       (A): NULL
REMARK   3   ESU BASED ON MAXIMUM LIKELIHOOD                 (A): NULL
REMARK   3   ESU FOR B VALUES BASED ON MAXIMUM LIKELIHOOD (A**2): 3.640
REMARK   3
REMARK   3  RMS DEVIATIONS FROM IDEAL VALUES.
REMARK   3   DISTANCE RESTRAINTS.                     RMS    SIGMA
REMARK   3    BOND LENGTH                    (A) : 0.015 ; NULL
REMARK   3    ANGLE DISTANCE                 (A) : 0.030 ; NULL
REMARK   3    INTRAPLANAR 1-4 DISTANCE       (A) : NULL  ; NULL
REMARK   3    H-BOND OR METAL COORDINATION   (A) : NULL  ; NULL
REMARK   3
REMARK   3   PLANE RESTRAINT                 (A) : NULL  ; NULL
REMARK   3   CHIRAL-CENTER RESTRAINT      (A**3) : NULL  ; NULL
REMARK   3
REMARK   3   NON-BONDED CONTACT RESTRAINTS.
REMARK   3    SINGLE TORSION                 (A) : NULL  ; NULL
REMARK   3    MULTIPLE TORSION               (A) : NULL  ; NULL
REMARK   3    H-BOND (X...Y)                 (A) : NULL  ; NULL
REMARK   3    H-BOND (X-H...Y)               (A) : NULL  ; NULL
REMARK   3
REMARK   3   CONFORMATIONAL TORSION ANGLE RESTRAINTS.
REMARK   3    SPECIFIED                (DEGREES) : NULL  ; NULL
REMARK   3    PLANAR                   (DEGREES) : NULL  ; NULL
REMARK   3    STAGGERED                (DEGREES) : NULL  ; NULL
REMARK   3    TRANSVERSE               (DEGREES) : NULL  ; NULL
REMARK   3
REMARK   3  ISOTROPIC THERMAL FACTOR RESTRAINTS.    RMS    SIGMA
REMARK   3   MAIN-CHAIN BOND              (A**2) : NULL  ; NULL
REMARK   3   MAIN-CHAIN ANGLE             (A**2) : NULL  ; NULL
REMARK   3   SIDE-CHAIN BOND              (A**2) : NULL  ; NULL
REMARK   3   SIDE-CHAIN ANGLE             (A**2) : NULL  ; NULL
REMARK   3
REMARK   3  OTHER REFINEMENT REMARKS:  DISORDERED REGIONS ARE FROM GLY22-
REMARK   3  GLY28, GLY65-GLU68 AND ILE147-PRO147 WERE MODELED
REMARK   3  STEREOCHEMICALLY
REMARK   4
REMARK   4 1BXW COMPLIES WITH FORMAT V. 3.15, 01-DEC-08
REMARK 100
REMARK 100 THIS ENTRY HAS BEEN PROCESSED BY RCSB ON 19-AUG-99.
REMARK 100 THE RCSB ID CODE IS RCSB008140.
REMARK 200
REMARK 200 EXPERIMENTAL DETAILS
REMARK 200  EXPERIMENT TYPE                : X-RAY DIFFRACTION
REMARK 200  DATE OF DATA COLLECTION        : 15-JAN-98
REMARK 200  TEMPERATURE           (KELVIN) : 298
REMARK 200  PH                             : 5.0
REMARK 200  NUMBER OF CRYSTALS USED        : 1
```

```
REMARK 200
REMARK 200  SYNCHROTRON              (Y/N) : N
REMARK 200  RADIATION SOURCE               : ROTATING ANODE
REMARK 200  BEAMLINE                       : NULL
REMARK 200  X-RAY GENERATOR MODEL          : RIGAKU RU200
REMARK 200  MONOCHROMATIC OR LAUE    (M/L) : M
REMARK 200  WAVELENGTH OR RANGE        (A) : 1.5418
REMARK 200  MONOCHROMATOR                  : NI FILTER
REMARK 200  OPTICS                         : NULL
REMARK 200
REMARK 200  DETECTOR TYPE                  : AREA DETECTOR
REMARK 200  DETECTOR MANUFACTURER          : SIEMENS
REMARK 200  INTENSITY-INTEGRATION SOFTWARE : XDS
REMARK 200  DATA SCALING SOFTWARE          : CCP4 (SCALA)
REMARK 200
REMARK 200  NUMBER OF UNIQUE REFLECTIONS   : 8328
REMARK 200  RESOLUTION RANGE HIGH      (A) : 2.500
REMARK 200  RESOLUTION RANGE LOW       (A) : 50.000
REMARK 200  REJECTION CRITERIA  (SIGMA(I)) : NULL
REMARK 200
REMARK 200 OVERALL.
REMARK 200  COMPLETENESS FOR RANGE     (%) : 89.0
REMARK 200  DATA REDUNDANCY                : 2.100
REMARK 200  R MERGE                    (I) : NULL
REMARK 200  R SYM                      (I) : 0.02800
REMARK 200  <I/SIGMA(I)> FOR THE DATA SET  : 16.8000
REMARK 200
REMARK 200 IN THE HIGHEST RESOLUTION SHELL.
REMARK 200  HIGHEST RESOLUTION SHELL, RANGE HIGH (A) : 2.50
REMARK 200  HIGHEST RESOLUTION SHELL, RANGE LOW  (A) : 2.64
REMARK 200  COMPLETENESS FOR SHELL     (%) : 53.0
REMARK 200  DATA REDUNDANCY IN SHELL       : 1.20
REMARK 200  R MERGE FOR SHELL          (I) : NULL
REMARK 200  R SYM FOR SHELL            (I) : 0.11000
REMARK 200  <I/SIGMA(I)> FOR SHELL         : 6.600
REMARK 200
REMARK 200 DIFFRACTION PROTOCOL: SINGLE WAVELENGTH
REMARK 200 METHOD USED TO DETERMINE THE STRUCTURE: MIRAS
REMARK 200 SOFTWARE USED: SHARP
REMARK 200 STARTING MODEL: NULL
REMARK 200
REMARK 200 REMARK: NULL
REMARK 280
REMARK 280 CRYSTAL
REMARK 280 SOLVENT CONTENT, VS   (%): 66.70
REMARK 280 MATTHEWS COEFFICIENT, VM (ANGSTROMS**3/DA): 3.70
REMARK 280
REMARK 280 CRYSTALLIZATION CONDITIONS: 10 % PEG-8000 10 % MPD 0.05 M
REMARK 280  POTASSIUM PHOSPHATE PH 5.0
REMARK 290
REMARK 290 CRYSTALLOGRAPHIC SYMMETRY
REMARK 290 SYMMETRY OPERATORS FOR SPACE GROUP: C 1 2 1
REMARK 290
REMARK 290      SYMOP   SYMMETRY
REMARK 290     NNNMMM   OPERATOR
REMARK 290       1555   X,Y,Z
REMARK 290       2555   -X,Y,-Z
REMARK 290       3555   X+1/2,Y+1/2,Z
REMARK 290       4555   -X+1/2,Y+1/2,-Z
REMARK 290
REMARK 290     WHERE NNN -> OPERATOR NUMBER
```

```
REMARK 290           MMM -> TRANSLATION VECTOR
REMARK 290
REMARK 290 CRYSTALLOGRAPHIC SYMMETRY TRANSFORMATIONS
REMARK 290 THE FOLLOWING TRANSFORMATIONS OPERATE ON THE ATOM/HETATM
REMARK 290 RECORDS IN THIS ENTRY TO PRODUCE CRYSTALLOGRAPHICALLY
REMARK 290 RELATED MOLECULES.
REMARK 290   SMTRY1   1  1.000000  0.000000  0.000000        0.00000
REMARK 290   SMTRY2   1  0.000000  1.000000  0.000000        0.00000
REMARK 290   SMTRY3   1  0.000000  0.000000  1.000000        0.00000
REMARK 290   SMTRY1   2 -1.000000  0.000000  0.000000        0.00000
REMARK 290   SMTRY2   2  0.000000  1.000000  0.000000        0.00000
REMARK 290   SMTRY3   2  0.000000  0.000000 -1.000000        0.00000
REMARK 290   SMTRY1   3  1.000000  0.000000  0.000000       34.59000
REMARK 290   SMTRY2   3  0.000000  1.000000  0.000000       38.97500
REMARK 290   SMTRY3   3  0.000000  0.000000  1.000000        0.00000
REMARK 290   SMTRY1   4 -1.000000  0.000000  0.000000       34.59000
REMARK 290   SMTRY2   4  0.000000  1.000000  0.000000       38.97500
REMARK 290   SMTRY3   4  0.000000  0.000000 -1.000000        0.00000
REMARK 290
REMARK 290 REMARK: NULL
REMARK 300
REMARK 300 BIOMOLECULE: 1
REMARK 300 SEE REMARK 350 FOR THE AUTHOR PROVIDED AND/OR PROGRAM
REMARK 300 GENERATED ASSEMBLY INFORMATION FOR THE STRUCTURE IN
REMARK 300 THIS ENTRY. THE REMARK MAY ALSO PROVIDE INFORMATION ON
REMARK 300 BURIED SURFACE AREA.
REMARK 350
REMARK 350 COORDINATES FOR A COMPLETE MULTIMER REPRESENTING THE KNOWN
REMARK 350 BIOLOGICALLY SIGNIFICANT OLIGOMERIZATION STATE OF THE
REMARK 350 MOLECULE CAN BE GENERATED BY APPLYING BIOMT TRANSFORMATIONS
REMARK 350 GIVEN BELOW.  BOTH NON-CRYSTALLOGRAPHIC AND
REMARK 350 CRYSTALLOGRAPHIC OPERATIONS ARE GIVEN.
REMARK 350
REMARK 350 BIOMOLECULE: 1
REMARK 350 AUTHOR DETERMINED BIOLOGICAL UNIT: MONOMERIC
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
REMARK 350   BIOMT3   1  0.000000  0.000000  1.000000        0.00000
REMARK 470
REMARK 470 MISSING ATOM
REMARK 470 THE FOLLOWING RESIDUES HAVE MISSING ATOMS(M=MODEL NUMBER;
REMARK 470 RES=RESIDUE NAME; C=CHAIN IDENTIFIER; SSEQ=SEQUENCE NUMBER;
REMARK 470 I=INSERTION CODE):
REMARK 470   M RES CSSEQI  ATOMS
REMARK 470     HIS A  31    CG   ND1  CD2  CE1  NE2
REMARK 475
REMARK 475 ZERO OCCUPANCY RESIDUES
REMARK 475 THE FOLLOWING RESIDUES WERE MODELED WITH ZERO OCCUPANCY.
REMARK 475 THE LOCATION AND PROPERTIES OF THESE RESIDUES MAY NOT
REMARK 475 BE RELIABLE. (M=MODEL NUMBER; RES=RESIDUE NAME;
REMARK 475 C=CHAIN IDENTIFIER; SSEQ=SEQUENCE NUMBER; I=INSERTION CODE)
REMARK 475   M RES C SSEQI
REMARK 475     GLY A   22
REMARK 475     LEU A   23
REMARK 475     ILE A   24
REMARK 475     ASN A   25
REMARK 475     ASN A   26
REMARK 475     ASN A   27
REMARK 475     GLY A   28
REMARK 475     GLY A   65
```

```
REMARK 475     SER A   66
REMARK 475     VAL A   67
REMARK 475     GLU A   68
REMARK 475     ILE A  147
REMARK 475     GLY A  148
REMARK 475     ASP A  149
REMARK 475     ALA A  150
REMARK 475     HIS A  151
REMARK 475     THR A  152
REMARK 475     ILE A  153
REMARK 475     GLY A  154
REMARK 475     THR A  155
REMARK 475     ARG A  156
REMARK 475     PRO A  157
REMARK 480
REMARK 480 ZERO OCCUPANCY ATOM
REMARK 480 THE FOLLOWING RESIDUES HAVE ATOMS MODELED WITH ZERO
REMARK 480 OCCUPANCY. THE LOCATION AND PROPERTIES OF THESE ATOMS
REMARK 480 MAY NOT BE RELIABLE. (M=MODEL NUMBER; RES=RESIDUE NAME;
REMARK 480 C=CHAIN IDENTIFIER; SSEQ=SEQUENCE NUMBER; I=INSERTION CODE):
REMARK 480   M RES C SSEQI ATOMS
REMARK 480     LYS A   64   CB    CG    CD    CE    NZ
REMARK 500
REMARK 500 GEOMETRY AND STEREOCHEMISTRY
REMARK 500 SUBTOPIC: CLOSE CONTACTS
REMARK 500
REMARK 500 THE FOLLOWING ATOMS THAT ARE RELATED BY CRYSTALLOGRAPHIC
REMARK 500 SYMMETRY ARE IN CLOSE CONTACT.  AN ATOM LOCATED WITHIN 0.15
REMARK 500 ANGSTROMS OF A SYMMETRY RELATED ATOM IS ASSUMED TO BE ON A
REMARK 500 SPECIAL POSITION AND IS, THEREFORE, LISTED IN REMARK 375
REMARK 500 INSTEAD OF REMARK 500.  ATOMS WITH NON-BLANK ALTERNATE
REMARK 500 LOCATION INDICATORS ARE NOT INCLUDED IN THE CALCULATIONS.
REMARK 500
REMARK 500 DISTANCE CUTOFF:
REMARK 500 2.2 ANGSTROMS FOR CONTACTS NOT INVOLVING HYDROGEN ATOMS
REMARK 500 1.6 ANGSTROMS FOR CONTACTS INVOLVING HYDROGEN ATOMS
REMARK 500
REMARK 500  ATM1  RES C  SSEQI   ATM2  RES C  SSEQI  SSYMOP   DISTANCE
REMARK 500   OD1  ASN A    26     CA   PRO A    29     2556     1.44
REMARK 500   OD1  ASN A    26     C    PRO A    29     2556     1.68
REMARK 500   OD1  ASN A    26     N    PRO A    29     2556     1.72
REMARK 500   OD1  ASN A     5     CD1  ILE A   147     2657     2.03
REMARK 500   OD1  ASN A    26     O    PRO A    29     2556     2.08
REMARK 500   CG   ASN A    26     N    PRO A    29     2556     2.11
REMARK 500
REMARK 500 REMARK: NULL
REMARK 500
REMARK 500 GEOMETRY AND STEREOCHEMISTRY
REMARK 500 SUBTOPIC: COVALENT BOND LENGTHS
REMARK 500
REMARK 500 THE STEREOCHEMICAL PARAMETERS OF THE FOLLOWING RESIDUES
REMARK 500 HAVE VALUES WHICH DEVIATE FROM EXPECTED VALUES BY MORE
REMARK 500 THAN 6*RMSD (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 500 IDENTIFIER; SSEQ=SEQUENCE NUMBER; I=INSERTION CODE).
REMARK 500
REMARK 500 STANDARD TABLE:
REMARK 500 FORMAT: (10X,I3,1X,2(A3,1X,A1,I4,A1,1X,A4,3X),1X,F6.3)
REMARK 500
REMARK 500 EXPECTED VALUES PROTEIN: ENGH AND HUBER, 1999
REMARK 500 EXPECTED VALUES NUCLEIC ACID: CLOWNEY ET AL 1996
REMARK 500
```

```
REMARK 500  M RES CSSEQI ATM1   RES CSSEQI ATM2   DEVIATION
REMARK 500    GLY A  28   C     PRO A  29   N       0.125
REMARK 500    GLY A 148   N     GLY A 148   CA      0.090
REMARK 500    ARG A 156   CA    ARG A 156   C       0.206
REMARK 500    PRO A 157   N     PRO A 157   CA     -0.251
REMARK 500    PRO A 157   CD    PRO A 157   N      -0.368
REMARK 500    PRO A 157   CA    PRO A 157   C      -0.164
REMARK 500
REMARK 500 REMARK: NULL
REMARK 500
REMARK 500 GEOMETRY AND STEREOCHEMISTRY
REMARK 500 SUBTOPIC: COVALENT BOND ANGLES
REMARK 500
REMARK 500 THE STEREOCHEMICAL PARAMETERS OF THE FOLLOWING RESIDUES
REMARK 500 HAVE VALUES WHICH DEVIATE FROM EXPECTED VALUES BY MORE
REMARK 500 THAN 6*RMSD (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 500 IDENTIFIER; SSEQ=SEQUENCE NUMBER; I=INSERTION CODE).
REMARK 500
REMARK 500 STANDARD TABLE:
REMARK 500 FORMAT: (10X,I3,1X,A3,1X,A1,I4,A1,3(1X,A4,2X),12X,F5.1)
REMARK 500
REMARK 500 EXPECTED VALUES PROTEIN: ENGH AND HUBER, 1999
REMARK 500 EXPECTED VALUES NUCLEIC ACID: CLOWNEY ET AL 1996
REMARK 500
REMARK 500  M RES CSSEQI ATM1   ATM2   ATM3
REMARK 500    ASP A   4   CA  -  C   -  N   ANGL. DEV. =  18.3 DEGREES
REMARK 500    GLY A  22   O   -  C   -  N   ANGL. DEV. =  11.3 DEGREES
REMARK 500    ASN A  25   C   -  N   -  CA  ANGL. DEV. = -16.5 DEGREES
REMARK 500    ASN A  25   CA  -  C   -  N   ANGL. DEV. = -14.2 DEGREES
REMARK 500    ASN A  25   O   -  C   -  N   ANGL. DEV. =  14.8 DEGREES
REMARK 500    GLY A  28   O   -  C   -  N   ANGL. DEV. = -11.6 DEGREES
REMARK 500    ARG A  60   CD  -  NE  -  CZ  ANGL. DEV. =   9.8 DEGREES
REMARK 500    ARG A  60   NE  -  CZ  -  NH1 ANGL. DEV. =   3.7 DEGREES
REMARK 500    ARG A  60   NE  -  CZ  -  NH2 ANGL. DEV. =  -3.3 DEGREES
REMARK 500    GLU A  68   C   -  N   -  CA  ANGL. DEV. = -16.3 DEGREES
REMARK 500    GLN A  75   CB  -  CA  -  C   ANGL. DEV. =  13.1 DEGREES
REMARK 500    ASP A  90   CB  -  CG  -  OD1 ANGL. DEV. =   5.7 DEGREES
REMARK 500    SER A 120   N   -  CA  -  CB  ANGL. DEV. =  -9.2 DEGREES
REMARK 500    VAL A 122   CB  -  CA  -  C   ANGL. DEV. = -12.2 DEGREES
REMARK 500    ILE A 135   CA  -  CB  -  CG2 ANGL. DEV. =  15.6 DEGREES
REMARK 500    ARG A 138   CA  -  CB  -  CG  ANGL. DEV. =  14.8 DEGREES
REMARK 500    ARG A 138   CD  -  NE  -  CZ  ANGL. DEV. =  10.7 DEGREES
REMARK 500    ARG A 138   NE  -  CZ  -  NH2 ANGL. DEV. =  -3.3 DEGREES
REMARK 500    HIS A 151   CB  -  CA  -  C   ANGL. DEV. = -38.2 DEGREES
REMARK 500    ALA A 150   CA  -  C   -  N   ANGL. DEV. = -16.2 DEGREES
REMARK 500    ALA A 150   O   -  C   -  N   ANGL. DEV. =  17.6 DEGREES
REMARK 500    HIS A 151   CA  -  C   -  N   ANGL. DEV. = -38.2 DEGREES
REMARK 500    HIS A 151   O   -  C   -  N   ANGL. DEV. =  43.4 DEGREES
REMARK 500    THR A 152   C   -  N   -  CA  ANGL. DEV. =  30.0 DEGREES
REMARK 500    ARG A 156   CB  -  CA  -  C   ANGL. DEV. =  12.4 DEGREES
REMARK 500    ARG A 156   N   -  CA  -  CB  ANGL. DEV. = -15.9 DEGREES
REMARK 500    ARG A 156   NH1 -  CZ  -  NH2 ANGL. DEV. =  -6.6 DEGREES
REMARK 500    ARG A 156   NE  -  CZ  -  NH2 ANGL. DEV. =   3.7 DEGREES
REMARK 500    THR A 155   O   -  C   -  N   ANGL. DEV. = -14.6 DEGREES
REMARK 500    ARG A 156   C   -  N   -  CA  ANGL. DEV. = -16.4 DEGREES
REMARK 500    PRO A 157   CA  -  N   -  CD  ANGL. DEV. = -25.7 DEGREES
REMARK 500    PRO A 157   N   -  CA  -  CB  ANGL. DEV. = -25.4 DEGREES
REMARK 500    PRO A 157   CB  -  CG  -  CD  ANGL. DEV. = -24.6 DEGREES
REMARK 500    PRO A 157   N   -  CD  -  CG  ANGL. DEV. = -33.9 DEGREES
REMARK 500    PRO A 157   N   -  CA  -  C   ANGL. DEV. =  19.1 DEGREES
REMARK 500    PRO A 157   CA  -  C   -  O   ANGL. DEV. = -17.2 DEGREES
```

```
REMARK 500    PRO A 157   C   -  N   -  CA  ANGL. DEV. = -16.9 DEGREES
REMARK 500
REMARK 500 REMARK: NULL
REMARK 500
REMARK 500 GEOMETRY AND STEREOCHEMISTRY
REMARK 500 SUBTOPIC: TORSION ANGLES
REMARK 500
REMARK 500 TORSION ANGLES OUTSIDE THE EXPECTED RAMACHANDRAN REGIONS:
REMARK 500 (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN IDENTIFIER;
REMARK 500 SSEQ=SEQUENCE NUMBER; I=INSERTION CODE).
REMARK 500
REMARK 500 STANDARD TABLE:
REMARK 500 FORMAT:(10X,I3,1X,A3,1X,A1,I4,A1,4X,F7.2,3X,F7.2)
REMARK 500
REMARK 500 EXPECTED VALUES: GJ KLEYWEGT AND TA JONES (1996). PHI/PSI-
REMARK 500 CHOLOGY: RAMACHANDRAN REVISITED. STRUCTURE 4, 1395 - 1400
REMARK 500
REMARK 500  M RES CSSEQI        PSI       PHI
REMARK 500    ASN A   5       57.62   -113.00
REMARK 500    TYR A  18      120.18    166.90
REMARK 500    ASP A  20     -140.62   -156.38
REMARK 500    LEU A  23      150.39     68.74
REMARK 500    ASN A  25      -90.55     -9.26
REMARK 500    ASN A  26      121.49    -29.94
REMARK 500    HIS A  31      175.58    173.40
REMARK 500    TYR A  63      102.76   -169.58
REMARK 500    SER A  66       52.30    128.49
REMARK 500    VAL A  67       90.53     49.93
REMARK 500    VAL A 110      -72.06    -67.54
REMARK 500    ALA A 150     -164.72    173.09
REMARK 500    HIS A 151      -97.21     35.47
REMARK 500    THR A 152     -139.61   -128.76
REMARK 500    THR A 155     -137.22   -149.83
REMARK 500    ARG A 156     -162.43   -178.66
REMARK 500
REMARK 500 REMARK: NULL
REMARK 500
REMARK 500 GEOMETRY AND STEREOCHEMISTRY
REMARK 500 SUBTOPIC: NON-CIS, NON-TRANS
REMARK 500
REMARK 500 THE FOLLOWING PEPTIDE BONDS DEVIATE SIGNIFICANTLY FROM BOTH
REMARK 500 CIS AND TRANS CONFORMATION.  CIS BONDS, IF ANY, ARE LISTED
REMARK 500 ON CISPEP RECORDS.  TRANS IS DEFINED AS 180 +/- 30 AND
REMARK 500 CIS IS DEFINED AS 0 +/- 30 DEGREES.
REMARK 500                                    MODEL     OMEGA
REMARK 500 ARG A  156     PRO A  157                 -147.47
REMARK 500
REMARK 500 REMARK: NULL
REMARK 500
REMARK 500 GEOMETRY AND STEREOCHEMISTRY
REMARK 500 SUBTOPIC: MAIN CHAIN PLANARITY
REMARK 500
REMARK 500 THE FOLLOWING RESIDUES HAVE A PSEUDO PLANARITY
REMARK 500 TORSION, C(I) - CA(I) - N(I+1) - O(I), GREATER
REMARK 500 10.0 DEGREES. (M=MODEL NUMBER; RES=RESIDUE NAME;
REMARK 500 C=CHAIN IDENTIFIER; SSEQ=SEQUENCE NUMBER;
REMARK 500 I=INSERTION CODE).
REMARK 500
REMARK 500  M RES CSSEQI        ANGLE
REMARK 500    ARG A 156        -11.76
REMARK 500
```

```
REMARK 500 REMARK: NULL
REMARK 800
REMARK 800 SITE
REMARK 800 SITE_IDENTIFIER: AC1
REMARK 800 EVIDENCE_CODE: SOFTWARE
REMARK 800 SITE_DESCRIPTION: BINDING SITE FOR RESIDUE C8E A 172
DBREF  1BXW A    0   171  UNP    P0A910   OMPA_ECOLI      21    192
SEQADV 1BXW MET A    0  UNP  P0A910    ALA    21 SEE REMARK 999
SEQADV 1BXW LEU A   23  UNP  P0A910    PHE    44 MUTATION
SEQADV 1BXW LYS A   34  UNP  P0A910    GLN    55 MUTATION
SEQADV 1BXW TYR A  107  UNP  P0A910    LYS   128 MUTATION
SEQRES   1 A  172  MET ALA PRO LYS ASP ASN THR TRP TYR THR GLY ALA LYS
SEQRES   2 A  172  LEU GLY TRP SER GLN TYR HIS ASP THR GLY LEU ILE ASN
SEQRES   3 A  172  ASN ASN GLY PRO THR HIS GLU ASN LYS LEU GLY ALA GLY
SEQRES   4 A  172  ALA PHE GLY GLY TYR GLN VAL ASN PRO TYR VAL GLY PHE
SEQRES   5 A  172  GLU MET GLY TYR ASP TRP LEU GLY ARG MET PRO TYR LYS
SEQRES   6 A  172  GLY SER VAL GLU ASN GLY ALA TYR LYS ALA GLN GLY VAL
SEQRES   7 A  172  GLN LEU THR ALA LYS LEU GLY TYR PRO ILE THR ASP ASP
SEQRES   8 A  172  LEU ASP ILE TYR THR ARG LEU GLY GLY MET VAL TRP ARG
SEQRES   9 A  172  ALA ASP THR TYR SER ASN VAL TYR GLY LYS ASN HIS ASP
SEQRES  10 A  172  THR GLY VAL SER PRO VAL PHE ALA GLY GLY VAL GLU TYR
SEQRES  11 A  172  ALA ILE THR PRO GLU ILE ALA THR ARG LEU GLU TYR GLN
SEQRES  12 A  172  TRP THR ASN ASN ILE GLY ASP ALA HIS THR ILE GLY THR
SEQRES  13 A  172  ARG PRO ASP ASN GLY MET LEU SER LEU GLY VAL SER TYR
SEQRES  14 A  172  ARG PHE GLY
HET    C8E  A 172      21
HETNAM     C8E (HYDROXYETHYLOXY)TRI(ETHYLOXY)OCTANE
FORMUL   2  C8E    C16 H34 O5
FORMUL   3  HOH   *39(H2 O)
SHEET    1  S1 1 THR A   6  SER A  16  0
SHEET    1  S2 1 LYS A  34  VAL A  45  0
SHEET    1  S3 1 VAL A  49  ARG A  60  0
SHEET    1  S4 1 TYR A  72  PRO A  86  0
SHEET    1  S5 1 LEU A  91  THR A 106  0
SHEET    1  S6 1 ASN A 114  ALA A 130  0
SHEET    1  S7 1 ILE A 135  TRP A 143  0
SHEET    1  S8 1 MET A 161  PHE A 170  0
LINK         OD2 ASP A 149                 C17 C8E A 172     2657   1555  1.24
LINK         CB  ASP A 149                 C17 C8E A 172     2657   1555  1.88
LINK         OD1 ASP A 149                 C17 C8E A 172     2657   1555  1.59
LINK         OD2 ASP A 149                 O18 C8E A 172     2657   1555  1.96
LINK         CA  ASP A 149                 O18 C8E A 172     2657   1555  1.92
LINK         CB  ASP A 149                 O18 C8E A 172     2657   1555  1.18
LINK         OD1 ASP A 149                 O18 C8E A 172     2657   1555  1.38
LINK         N   ASP A 149                 C19 C8E A 172     2657   1555  1.68
LINK         C   ASP A 149                 C19 C8E A 172     2657   1555  1.87
LINK         CA  ASP A 149                 C20 C8E A 172     2657   1555  1.45
LINK         C   ASP A 149                 C20 C8E A 172     2657   1555  1.22
LINK         CB  ASP A 149                 C20 C8E A 172     2657   1555  2.02
LINK         N   ALA A 150                 O21 C8E A 172     2657   1555  1.70
LINK         N   ALA A 150                 C20 C8E A 172     2657   1555  1.34
SITE     1 AC1  4 TYR A  43  PHE A  51  LEU A  79  GLY A  99
CRYST1   69.180   77.950   50.930  90.00  91.52  90.00 C 1 2 1       4
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.014455  0.000000  0.000383        0.00000
SCALE2      0.000000  0.012829  0.000000        0.00000
SCALE3      0.000000  0.000000  0.019642        0.00000
```

Επεξήγηση πεδίων μιας εγγραφής PDB

HEADER: *Περιέχει ένα τετραψήφιο κωδικό για την αναγνώριση της εγγραφής στην PDB, μια γενική ταξινόμηση του μακρομορίου καθώς και την ημερομηνία κατάθεσης της δομής στην Protein Data Bank.*

TITLE: *Τίτλος που περιλαμβάνει συνήθως τα περιεχόμενα της εγγραφής, τι είδους πειραματική διαδικασία χρησιμοποιήθηκε, ύπαρξη μεταλλάξεων. Επιτρέπει στον ερεύνητη που κατέθεσε τη δομή να καταδείξει τη σημαντικότητα της εργασίας αυτής.*

COMPOUND: *Το πεδίο compound περιέχει πληροφορίες για το μακρομόριο που αναφέρεται στη δομή καθώς και τα άλλα μόρια (μικρές οργανικές ενώσεις, μέταλλα) με τα οποία έχει τυχόν συμπλοκοποιηθεί.*

SOURCE: *Βιολογική προέλευση του μακρομορίου που αναφέρεται στην εγγραφή.*

KEYWDS: *Χαρακτηριστικές λέξεις-κλειδιά για τον χαρακτηρισμό της εγγραφής.*

EXPDTA: *Πειραματική τεχνική για τον προσδιορισμό της δομής (X-Ray Crystallography/NMR/Theoretical Model).*

AUTHOR: *Λίστα με τα ονόματα των ερευνητών που συμμετείχαν στον προσδιορισμό της δομής.*

JRNL: *Πρωταρχική βιβλιογραφική αναφορά η οποία αναφερέται στον προσδιορισμό της δομής που αναφέρεται στην συγκεκριμένη εγγραφή.*

REMARK: *Το πεδίο REMARK περιλαμβάνει μια σειρά από πληροφορίες σχετικές με την κατατεθειμένη δομή.*

Καταρχήν περιέχει βιβλιογραφικές αναφορές που σχετίζονται άμεσα με το προς μελέτη μακρομόριο.

Στο πεδίο REMARK περιλαμβάνονται και στοιχεία σχετικά με την πειραματική διαδικασία που ακολουθήθηκε για την λύση της δομής όπως είναι τα προγράμματα που χρησιμοποιήθηκαν, οι τιμές διαφόρων δεικτών, γενικά πληροφορίες που αποδεικνύουν την ορθότητα της δομής.

SEQRES: *Περιέχει την αλληλουχία του προς μελέτη μακρομορίου. Για τις πρωτεΐνες ακολουθείται ο κώδικας των 3 γραμμάτων.*

HET: *Αναφέρεται στα μόρια (ετεροάτομα) που δεν είναι αμινοξέα ή νουκλεοτίδια. Αυτά μπορεί να είναι προσθετικές ομάδες και ιόντα για τα οποία έχουν προσδιοριστεί οι συντεταγμένες τους. Τα στοιχεία που δίνονται για αυτά είναι ένας κωδικός για να διευκρινίζονται σε σχέση με τα άλλα κατάλοιπα της εγγραφής, η αρίθμηση που έχουν μέσα στο αρχείο των συντεταγμένων και τέλος ο αριθμός των ατόμων από τα οποία αποτελούνται.*

HETNAM: *Ονοματολογία των καταλοίπων που περιέχονται στο πεδίο HET.*

FORMUL: *Μοριακός τύπος των καταλοίπων που αναφέρονται στο πεδίο HET.*

HELIX: *Τμήματα της αλληλουχίας που έχουν ελικοειδή δομή.*

SHEET: *Τμήματα της αλληλουχίας που έχουν εκτεταμένη δομή.*

CRYST1: *Περιέχει τις παραμέτρους μοναδιαίας κυψελλίδας και την ομάδα συμμετρίας χώρου.*

ORIGXn(n=1..3): *Πίνακας Μετατροπής από σύστημα ορθογωνίων συντεταγμένων στις συντεταγμένες που κατατέθηκαν αρχικά στην PDB.*

SCALEn: *Πίνακας Μετατροπής από σύστημα ορθογωνίων συντεταγμένων στις κρυσταλλογραφικές συντεταγμένες.*

ATOM: *Περιέχει τις συντεταγμένες των ατόμων στους άξονες X, Y, Z. Περιλαμβάνει επίσης και άλλα στοιχεία όπως τα άτομα για τα οποία αναφέρονται οι συντεταγμένες και σε ποια κατάλοιπα ανήκουν. Πρέπει να σημειωθεί ότι κάθε είδους δεδομένο που περιέχεται στο πεδίο ATOM είναι τοποθετημένο σε καθορισμένες θέσεις (στήλες) της εγγραφής όπως αυτές παρουσιάζονται παρακάτω:*

ΣΤΗΛΕΣ Περιεχόμενα κάθε στήλης

---

1 - 6 *"ATOM " δηλώνει ότι πρόκειται για το πεδίο ATOM.*

7 - 11 *Αύξων αριθμός του ατόμου.*

13 - 16 *Τύπος ατόμου.*

18 - 20 *Όνομα καταλοίπου. Για τα αμινοξέα ακολουθείται ο κώδικας των 3 γραμμάτων.*

22 *(chainID) Χαρακτήρας που ταυτοποιεί την αλυσίδα, αν περιέχονται περισσότερες από μια στην εγγραφή.*

23 - 26 *Αρίθμηση του καταλοίπου στην αλυσίδα*

31 - 38 *x Συντεταγμένες ατόμου (σε Angstroms) στον άξονα X σε τρισορθογώνιο σύστημα αξόνων.*

39 - 46 *y Συντεταγμένες ατόμου (σε Angstroms) στον άξονα Y σε τρισορθογώνιο σύστημα αξόνων.*

47 - 54 *z Συντεταγμένες ατόμου (σε Angstroms) στον άξονα Z σε τρισορθογώνιο σύστημα αξόνων.*

55 - 60 *Συντελεστής κατάληψης(occupancy)*

61 - 66 *Παράγοντας θερμοκρασίας(Temperature factor)*

77 - 78 *Σύμβολο του ατόμου.*

79 - 80 *Φορτίο του ατόμου (Αν υπάρχει).*

TER: *Το πεδίο TER δηλώνει το τέλος της παράθεσης των ατόμων που απαρτίζουν μια αλυσίδα.*

HETATM: *Συντεταγμένες των ετεροατόμων. Η μορφοποίηση τους ακολουθεί τους ίδιους κανόνες με το πεδίο ATOM.*

CONECT: *Το πεδίο CONECT καθορίζει τα άτομα τα οποία συμμετέχουν στον σχηματισμό δεσμών. Κάθε άτομο συμβολίζεται με την αρίθμηση του όπως είναι καθορισμένη στα πεδία ATOM.*

MASTER: *Αποτελεί ένα πεδίο που χρησιμοποιείται για μια απλή οργάνωση της εγγραφής. Πρόκειται για μια σειρά από αριθμούς που δεν είναι τίποτε άλλο από το άθροισμα των γραμμών για συγκεκριμένα πεδία της εγγραφής.*

END: *Υποδηλώνει τη λήξη της εγγραφής.*

Βιβλιογραφία

Akiva, E., Brown, S., Almonacid, D. E., Barber, A. E., 2nd, Custer, A. F., Hicks, M. A., . . . Babbitt, P. C. (2014). The Structure-Function Linkage Database. *Nucleic acids research, 42*(Database issue), D521-530.

Akondi, K. B., Muttenthaler, M., Dutertre, S., Kaas, Q., Craik, D. J., Lewis, R. J., & Alewood, P. F. (2014). Discovery, synthesis, and structure-activity relationships of conotoxins. *Chemical reviews, 114*(11), 5815-5847.

Alexander, S. P., Benson, H. E., Faccenda, E., Pawson, A. J., Sharman, J. L., McGrath, J. C., . . . Zimmermann, M. (2013). The Concise Guide to PHARMACOLOGY 2013/14: overview. *British journal of pharmacology, 170*(8), 1449-1458.

Alexander, S. P., Mathie, A., & Peters, J. A. (2011). Guide to Receptors and Channels (GRAC), 5th edition. *British journal of pharmacology, 164 Suppl 1*, S1-324.

Almonacid, D. E., & Babbitt, P. C. (2011). Toward mechanistic classification of enzyme functions. *Current opinion in chemical biology, 15*(3), 435-442.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research, 25*(17), 3389-3402.

Ananiadou, S., Kell, D. B., & Tsujii, J. (2006). Text mining and its potential applications in systems biology. *Trends Biotechnol, 24*(12), 571-579.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C., & Murzin, A. G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res, 32*(Database issue), D226-229.

Atkinson, H. J., Morris, J. H., Ferrin, T. E., & Babbitt, P. C. (2009). Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PloS one, 4*(2), e4345.

Babbitt, P. C., & Gerlt, J. A. (1997). Understanding enzyme superfamilies. Chemistry As the fundamental determinant in the evolution of new catalytic activities. *The Journal of biological chemistry, 272*(49), 30591-30594.

Babbitt, P. C., Hasson, M. S., Wedekind, J. E., Palmer, D. R., Barrett, W. C., Reed, G. H., . . . Gerlt, J. A. (1996). The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochemistry, 35*(51), 16489-16501.

Bairoch, A. (1999). The ENZYME data bank in 1999. *Nucleic acids research, 27*(1), 310-311.

Barrett, T., & Edgar, R. (2006). Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*. *Methods Mol Biol, 338*, 175-190.

Baxevanis, A. D., Arents, G., Moudrianakis, E. N., & Landsman, D. (1995). A variety of DNA-binding and multimeric proteins contain the histone fold motif. *Nucleic acids research, 23*(14), 2685-2691.

Baxevanis, A. D., & Landsman, D. (1996). Histone Sequence Database: a compilation of highly-conserved nucleoprotein sequences. *Nucleic acids research, 24*(1), 245-247.

Becker, K. G., Barnes, K. C., Bright, T. J., & Wang, S. A. (2004). The genetic association database. *Nat Genet, 36*(5), 431-432.

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2014). GenBank. *Nucleic acids research*.

Berman, H., Henrick, K., Nakamura, H., & Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic acids research, 35*(Database issue), D301-303.

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., & Tanzi, R. E. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet, 39*(1), 17-23.

Bhaskara, R. M., Mehrotra, P., Rakshambikai, R., Gnanavel, M., Martin, J., & Srinivasan, N. (2014). The relationship between classification of multi-domain proteins using an alignment-free approach and their functions: a case study with immunoglobulins. *Molecular BioSystems, 10*(5), 1082-1093.

Biggs, J. S., Watkins, M., Puillandre, N., Ownby, J. P., Lopez-Vera, E., Christensen, S., . . . Olivera, B. M. (2010). Evolution of Conus peptide toxins: analysis of Conus californicus Reeve, 1844. *Molecular phylogenetics and evolution, 56*(1), 1-12.

Bingham, J., Plowman, G. D., & Sudarsanam, S. (2000). Informatics issues in large-scale sequence analysis: elucidating the protein kinases of C. elegans. *J Cell Biochem, 80*(2), 181-186.

Bockaert, J., & Pin, J. P. (1999). Molecular tinkering of G protein-coupled receptors: an evolutionary success. *Embo J, 18*(7), 1723-1729.

Bonner, T. I. (2014). Should pharmacologists care about alternative splicing? IUPHAR Review 4. *British journal of pharmacology, 171*(5), 1231-1240.

Bradham, C. A., Foltz, K. R., Beane, W. S., Arnone, M. I., Rizzo, F., Coffman, J. A., . . . Manning, G. (2006). The sea urchin kinome: a first look. *Dev Biol, 300*(1), 180-193.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., . . . Sansone, S. A. (2003). ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res, 31*(1), 68-71.

Caenepeel, S., Charydczak, G., Sudarsanam, S., Hunter, T., & Manning, G. (2004). The mouse kinome: discovery and comparative genomics of all mouse protein kinases. *Proc Natl Acad Sci U S A, 101*(32), 11707-11712.

Campbell, J. A., Davies, G. J., Bulone, V., & Henrissat, B. (1997). A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *The Biochemical journal, 326 (Pt 3)*, 929-939.

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., & Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic acids research, 37*(Database issue), D233-238.

Chang, D., & Duda, T. F., Jr. (2012). Extensive and continuous duplication facilitates rapid evolution and diversification of gene families. *Molecular biology and evolution, 29*(8), 2019-2029.

Chatonnet, A., Cousin, X., & Robinson, A. (2001). Links between kinetic data and sequences in the alpha/beta-hydrolases fold database. *Briefings in bioinformatics, 2*(1), 30-37.

Chibucos, M. C., Mungall, C. J., Balakrishnan, R., Christie, K. R., Huntley, R. P., White, O., . . . Giglio, M. (2014). Standardized description of scientific evidence using the Evidence Ontology (ECO). *Database (Oxford), 2014*.

Christopoulos, A., Changeux, J. P., Catterall, W. A., Fabbro, D., Burris, T. P., Cidlowski, J. A., . . . Langmead, C. J. (2014). International union of basic and clinical pharmacology. XC. multisite pharmacology: recommendations for the nomenclature of receptor allosterism and allosteric ligands. *Pharmacological reviews, 66*(4), 918-947.

Cousin, X., Hotelier, T., Lievin, P., Toutant, J. P., & Chatonnet, A. (1996). A cholinesterase genes server (ESTHER): a database of cholinesterase-related sequences for multiple alignments, phylogenetic relationships, mutations and structural data retrieval. *Nucleic acids research, 24*(1), 132-136.

Craik, D. J. (2006). Chemistry. Seamless proteins tie up their loose ends. *science, 311*(5767), 1563-1564.

Cuff, A. L., Sillitoe, I., Lewis, T., Clegg, A. B., Rentzsch, R., Furnham, N., . . . Orengo, C. A. (2011). Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic acids research, 39*(Database issue), D420-426.

Davis, J., Jones, A., & Lewis, R. J. (2009). Remarkable inter- and intra-species complexity of conotoxins revealed by LC/MS. *Peptides, 30*(7), 1222-1227.

Demeter, J., Beauheim, C., Gollub, J., Hernandez-Boussard, T., Jin, H., Maier, D., . . . Ball, C. A. (2007). The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res, 35*(Database issue), D766-770.

Deshmukh, K., Anamika, K., & Srinivasan, N. (2010). Evolution of domain combinations in protein kinases and its implications for functional diversity. *Progress in biophysics and molecular biology, 102*(1), 1-15.

Dreos, R., Ambrosini, G., Cavin Perier, R., & Bucher, P. (2013). EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Res, 41*(Database issue), D157-164.

Duda, T. F., Jr., Chang, D., Lewis, B. D., & Lee, T. (2009). Geographic variation in venom allelic composition and diets of the widespread predatory marine gastropod Conus ebraeus. *PloS one, 4*(7), e6245.

Duda, T. F., Jr., & Lee, T. (2009). Ecological release and venom evolution of a predatory marine snail at Easter Island. *PloS one, 4*(5), e5558.

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome informatics. International Conference on Genome Informatics, 23*(1), 205-211.

Eisen, J. A., Coyne, R. S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J. R., . . . Orias, E. (2006). Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. *PLoS Biol, 4*(9), e286.

Fernández-Suárez, X. M., Rigden, D. J., & Galperin, M. Y. (2014). The 2014 nucleic acids research database issue and an updated NAR online molecular biology database collection. *Nucleic acids research, 42*(D1), D1-D6.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., . . . Punta, M. (2014). Pfam: the protein families database. *Nucleic acids research, 42*(Database issue), D222-D230.

Fleischmann, A., Darsow, M., Degtyarenko, K., Fleischmann, W., Boyce, S., Axelsen, K. B., . . . Apweiler, R. (2004). IntEnz, the integrated relational enzyme database. *Nucleic acids research, 32*(Database issue), D434-437.

Gandhimathi, A., Nair, A. G., & Sowdhamini, R. (2012). PASS2 version 4: an update to the database of structure-based sequence alignments of structural domain superfamilies. *Nucleic acids research, 40*(Database issue), D531-534.

Garland, S. L. (2013). Are GPCRs Still a Source of New Targets? *Journal of Biomolecular Screening, 18*(9), 947-966.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., . . . Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research, 40*(Database issue), D1100-1107.

Gerlt, J. A., & Babbitt, P. C. (2001). Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annual review of biochemistry, 70*, 209-246.

Gerlt, J. A., Babbitt, P. C., & Rayment, I. (2005). Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. *Archives of biochemistry and biophysics, 433*(1), 59-70.

Gnanavel, M., Mehrotra, P., Rakshambikai, R., Martin, J., Srinivasan, N., & Bhaskara, R. M. (2014). CLAP: a web-server for automatic classification of proteins with special reference to multi-domain proteins. *BMC bioinformatics, 15*, 343.

The dictyostelium kinome--analysis of the protein kinases from a simple model organism, 3, 2 Cong. Rec. e38 (2006).

Gowri, V. S., Krishnadev, O., Swamy, C. S., & Srinivasan, N. (2006). MulPSSM: a database of multiple position-specific scoring matrices of protein domain families. *Nucleic acids research, 34*(Database issue), D243-246.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., & Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res, 34*(Database issue), D140-144.

Grosjean, J., Soualmia, L., Bouarech, K., Jonquet, C., & Darmoni, S. (2014). *An Approach to Compare Bio-Ontologies Portals.* Paper presented at the MIE'2014: 26th International Conference of the European Federation for Medical Informatics.

Hanks, S. K., & Hunter, T. (1995). Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology, 9*(8), 576-596.

HapMap Consortium. (2003). The International HapMap Project. *Nature, 426*(6968), 789-796.

Harmar, A. J., Hills, R. A., Rosser, E. M., Jones, M., Buneman, O. P., Dunbar, D. R., . . . Spedding, M. (2009). IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels. *Nucleic acids research, 37*(Database issue), D680-685.

Henrissat, B. (1991). A classification of glycosyl hydrolases based on amino acid sequence similarities. *The Biochemical journal, 280 (Pt 2)*, 309-316.

Holliday, G. L., Almonacid, D. E., Bartlett, G. J., O'Boyle, N. M., Torrance, J. W., Murray-Rust, P., . . . Thornton, J. M. (2007). MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. *Nucleic acids research, 35*(Database issue), D515-520.

Holliday, G. L., Andreini, C., Fischer, J. D., Rahman, S. A., Almonacid, D. E., Williams, S. T., & Pearson, W. R. (2012). MACiE: exploring the diversity of biochemical reactions. *Nucleic acids research, 40*(Database issue), D783-789.

Holliday, G. L., Bairoch, A., Bagos, P. G., Chatonnet, A., Craik, D. J., Flinn, R. D., . . . Bateman, A. (2015). Key challenges for the creation and maintenance of specialist protein resources. *Proteins*.

Horn, F., Bettler, E., Oliveira, L., Campagne, F., Cohen, F. E., & Vriend, G. (2003). GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res, 31*(1), 294-297.

Horn, F., Weare, J., Beukers, M. W., Horsch, S., Bairoch, A., Chen, W., . . . Vriend, G. (1998). GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res, 26*(1), 275-279.

Hsu, S. D., Lin, F. M., Wu, W. Y., Liang, C., Huang, W. C., Chan, W. L., . . . Huang, H. D. (2011). miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res, 39*(Database issue), D163-169.

Hunter, T., & Plowman, G. D. (1997). The protein kinases of budding yeast: six score and more. *Trends Biochem Sci, 22*(1), 18-22.

Isberg, V., Vroling, B., van der Kant, R., Li, K., Vriend, G., & Gloriam, D. (2014). GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res, 42*(1), D422-425.

Jamison, D. C. (2003). Structured Query Language (SQL) fundamentals. *Curr Protoc Bioinformatics, Chapter 9*, Unit9 2.

Joosten, R. P., Long, F., Murshudov, G. N., & Perrakis, A. (2014). The PDB_REDO server for macromolecular structure model optimization. *IUCrJ, 1*(4), 0-0.

Kaas, Q., Westermann, J. C., & Craik, D. J. (2010). Conopeptide characterization and classifications: an analysis using ConoServer. *Toxicon : official journal of the International Society on Toxinology, 55*(8), 1491-1509.

Kaas, Q., Yu, R., Jin, A. H., Dutertre, S., & Craik, D. J. (2012). ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic acids research, 40*(Database issue), D325-330.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research, 42*(D1), D199-D205.

Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M., . . . Gama-Castro, S. (2002). The EcoCyc Database. *Nucleic Acids Res, 30*(1), 56-58.

Katritch, V., Cherezov, V., & Stevens, R. C. (2013). Structure-function of the G protein-coupled receptor superfamily. *Annu Rev Pharmacol Toxicol, 53*, 531-556.

Kedarisetti, P., Mizianty, M. J., Kaas, Q., Craik, D. J., & Kurgan, L. (2014). Prediction and characterization of cyclic proteins from sequences in three domains of life. *Biochimica et biophysica acta, 1844*(1 Pt B), 181-190.

Kirby, A. J. (2001). The lysozyme mechanism sorted — after 50 years. *Nature Structural Biology, 8*, 737-739.

Knudsen, M., & Wiuf, C. (2010). The CATH database. *Hum Genomics, 4*(3), 207-212.

Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P. E., & Berman, H. M. (2006). The RCSB PDB information portal for structural genomics. *Nucleic Acids Res, 34*(Database issue), D302-305.

Kozma, D., Simon, I., & Tusnady, G. E. (2013). PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res, 41*(Database issue), D524-529.

Krupa, A., Abhinandan, K., & Srinivasan, N. (2004). KinG: a database of protein kinases in genomes. *Nucleic acids research, 32*(suppl 1), D153-D155.

Krupp, M., Marquardt, J. U., Sahin, U., Galle, P. R., Castle, J., & Teufel, A. (2012). RNA-Seq Atlas--a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics, 28*(8), 1184-1185.

Lagerstrom, M. C., & Schioth, H. B. (2008). Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat Rev Drug Discov, 7*(4), 339-357.

Lane, L., Argoud-Puy, G., Britan, A., Cusin, I., Duek, P. D., Evalet, O., . . . Bairoch, A. (2012). neXtProt: a knowledge platform for human proteins. *Nucleic acids research, 40*(Database issue), D76-83.

Lenfant, N., Hotelier, T., Bourne, Y., Marchot, P., & Chatonnet, A. (2013). Proteins with an alpha/beta hydrolase fold: Relationships between subfamilies in an ever-growing superfamily. *Chemico-biological interactions, 203*(1), 266-268.

Lenfant, N., Hotelier, T., Bourne, Y., Marchot, P., & Chatonnet, A. (2014). Tracking the origin and divergence of cholinesterases and neuroligins: the evolution of synaptic proteins. *Journal of molecular neuroscience : MN, 53*(3), 362-369.

Lenfant, N., Hotelier, T., Velluet, E., Bourne, Y., Marchot, P., & Chatonnet, A. (2013). ESTHER, the database of the alpha/beta-hydrolase fold superfamily of proteins: tools to explore diversity of functions. *Nucleic acids research, 41*(Database issue), D423-429.

Lombard, V., Ramulu, H. G., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic acids research, 42*(D1), D490-D495.

Lu, C. T., Huang, K. Y., Su, M. G., Lee, T. Y., Bretana, N. A., Chang, W. C., . . . Huang, H. D. (2013). DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic Acids Res, 41*(Database issue), D295-305.

Manning, G. (2005). Genomic overview of protein kinases. *WormBook*, 1-19.

--- Evolution of protein kinase signaling from yeast to man, 10, 27 Cong. Rec. 514-520 (2002).

--- The protein kinase complement of the human genome, 5600, 298 Cong. Rec. 1912-1934 (2002).

Marchot, P., & Chatonnet, A. (2012). Enzymatic activity and protein interactions in alpha/beta hydrolase fold proteins: moonlighting versus promiscuity. *Protein and peptide letters, 19*(2), 132-143.

Marino-Ramirez, L., Levine, K. M., Morales, M., Zhang, S., Moreland, R. T., Baxevanis, A. D., & Landsman, D. (2011). The Histone Database: an integrated resource for histones and histone fold-containing proteins. *Database (Oxford), 2011*, bar048.

Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., . . . Thornton, J. M. (1998). Protein folds and functions. *Structure, 6*(7), 875-884.

Martin, J., Anamika, K., & Srinivasan, N. (2010). Classification of protein kinases on the basis of both kinase and non-kinase regions. *PloS one, 5*(9), e12460.

McDonald, A. G., Boyce, S., & Tipton, K. F. (2009). ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic acids research, 37*(Database issue), D593-597.

Moszer, I., Jones, L. M., Moreira, S., Fabry, C., & Danchin, A. (2002). SubtiList: the reference database for the Bacillus subtilis genome. *Nucleic Acids Res, 30*(1), 62-65.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology, 247*(4), 536-540.

Nagano, N. (2005). EzCatDB: the Enzyme Catalytic-mechanism Database. *Nucleic acids research, 33*(Database issue), D407-412.

Nagano, N., Nakayama, N., Ikeda, K., Fukuie, M., Yokota, K., Doi, T., . . . Tomii, K. (2014). EzCatDB: the enzyme reaction database, 2015 update. *Nucleic acids research*.

Nagy, A., Hegyi, H., Farkas, K., Tordai, H., Kozma, E., Banyai, L., & Patthy, L. (2008). Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. *BMC bioinformatics, 9*, 353.

Overington, J. P., Al-Lazikani, B., & Hopkins, A. L. (2006). How many drug targets are there? *Nat Rev Drug Discov, 5*(12), 993-996.

Pawson, A. J., Sharman, J. L., Benson, H. E., Faccenda, E., Alexander, S. P., Buneman, O. P., . . . Harmar, A. J. (2014). The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic acids research, 42*(Database issue), D1098-1106.

Pegg, S. C., Brown, S. D., Ojha, S., Seffernick, J., Meng, E. C., Morris, J. H., . . . Babbitt, P. C. (2006). Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry, 45*(8), 2545-2555.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry, 25*(13), 1605-1612.

Pettifer, S., Ison, J., Kalas, M., Thorne, D., McDermott, P., Jonassen, I., . . . Vriend, G. (2010). The EMBRACE web service collection. *Nucleic Acids Res, 38*(Web Server issue), W683-688.

Plowman, G. D., Sudarsanam, S., Bingham, J., Whyte, D., & Hunter, T. (1999). The protein kinases of Caenorhabditis elegans: a model for signal transduction in multicellular organisms. *Proc Natl Acad Sci U S A, 96*(24), 13603-13610.

Poth, A. G., Chan, L. Y., & Craik, D. J. (2013). Cyclotides as grafting frameworks for protein engineering and drug design applications. *Biopolymers, 100*(5), 480-491.

Puillandre, N., Koua, D., Favreau, P., Olivera, B. M., & Stocklin, R. (2012). Molecular phylogeny, classification and evolution of conopeptides. *Journal of molecular evolution, 74*(5-6), 297-309.

Rakshambikai, R., Gnanavel, M., & Srinivasan, N. (2014). Hybrid and rogue kinases encoded in the genomes of model eukaryotes. *PloS one, 9*(9), e107956.

Rawlings, N. D., Waller, M., Barrett, A. J., & Bateman, A. (2014). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic acids research, 42*(Database issue), D503-D509.

Reddy, T. B., Thomas, A. D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., . . . Kyrpides, N. C. (2015). The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res, 43*(Database issue), D1099-1106.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., . . . Chinnaiyan, A. M. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia, 6*(1), 1-6.

Rose, P. W., Bi, C., Bluhm, W. F., Christie, C. H., Dimitropoulos, D., Dutta, S., . . . Bourne, P. E. (2013). The RCSB Protein Data Bank: new resources for research and education. *Nucleic acids research, 41*(Database issue), D475-482.

Rose, P. W., Prlic, A., Bi, C., Bluhm, W. F., Christie, C. H., Dutta, S., . . . Burley, S. K. (2014). The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic acids research*.

Saier, M. H., Jr. (2000). A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol Mol Biol Rev, 64*(2), 354-411.

Saier, M. H., Reddy, V. S., Tamang, D. G., & Västermark, Å. (2014). The Transporter Classification Database. . *Nucleic acids research, 42*(Database issue), D251-D258.

Scherf, M., Epple, A., & Werner, T. (2005). The next generation of literature analysis: integration of genomic analysis into text mining. *Brief Bioinform, 6*(3), 287-297.

Schnoes, A. M., Brown, S. D., Dodevski, I., & Babbitt, P. C. (2009). Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLoS computational biology, 5*(12), e1000605.

Schully, S. D., Yu, W., McCallum, V., Benedicto, C. B., Dong, L. M., Wulf, A., . . . Khoury, M. J. (2011). Cancer GAMAdb: database of cancer genetic associations from meta-analyses and genome-wide association studies. *Eur J Hum Genet, 19*(8), 928-930.

Sethupathy, P., Corda, B., & Hatzigeorgiou, A. G. (2006). TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA, 12*(2), 192-197.

Shepelev, V., & Fedorov, A. (2006). Advances in the Exon-Intron Database (EID). *Brief Bioinform, 7*(2), 178-185.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res, 29*(1), 308-311.

Sigrist, C. J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., & Hulo, N. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res, 38*(Database issue), D161-166.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., . . . Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology, 25*(11).

Sowdhamini, R., Burke, D. F., Huang, J. F., Mizuguchi, K., Nagarajaram, H. A., Srinivasan, N., . . . Blundell, T. L. (1998). CAMPASS: a database of structurally aligned protein superfamilies. *Structure, 6*(9), 1087-1094.

Spedding, M. (2011). Resolution of controversies in drug/receptor interactions by protein structure. Limitations and pharmacological solutions. *Neuropharmacology, 60*(1), 3-6.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M. E., Mitros, T., . . . Rokhsar, D. S. (2010). The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature, 466*(7307), 720-726.

Stajich, J. E., Wilke, S. K., Ahren, D., Au, C. H., Birren, B. W., Borodovsky, M., . . . Pukkila, P. J. (2010). Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom Coprinopsis cinerea (Coprinus cinereus). *Proc Natl Acad Sci U S A, 107*(26), 11889-11894.

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res, 34*(Database issue), D535-539.

Stein, L. (2013). Creating databases for biological information: an introduction. *Curr Protoc Bioinformatics, Chapter 9*, Unit9 1.

Sun, H., Palaniswamy, S. K., Pohar, T. T., Jin, V. X., Huang, T. H., & Davuluri, R. V. (2006). MPromDb: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-chip experimental data. *Nucleic Acids Res, 34*(Database issue), D98-103.

Tan, N. C., & Berkovic, S. F. (2010). The Epilepsy Genetic Association Database (epiGAD): analysis of 165 genetic association studies, 1996-2008. *Epilepsia, 51*(4), 686-689.

Terlau, H., & Olivera, B. M. (2004). Conus venoms: a rich source of novel ion channel-targeted peptides. *Physiological reviews, 84*(1), 41-68.

Theodoropoulou, M. C., Bagos, P. G., Spyropoulos, I. C., & Hamodrakas, S. J. (2008). gpDB: a database of GPCRs, G-proteins, effectors and their interactions. *Bioinformatics, 24*(12), 1471-1472.

Tipton, K. F. (1994). Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement: corrections and additions. *European journal of biochemistry / FEBS, 223*(1), 1-5.

Tough, D. F., Lewis, H. D., Rioja, I., Lindon, M. J., & Prinjha, R. K. (2014). Epigenetic pathway targets for the treatment of disease: accelerating progress in the development of pharmacological tools: IUPHAR Review 11. *British journal of pharmacology, 171*(22), 4981-5010.

Trabi, M., & Craik, D. J. (2002). Circular proteins--no end in sight. *Trends Biochem Sci, 27*(3), 132-138.

Tsaousis, G. N., Tsirigos, K. D., Andrianou, X. D., Liakopoulos, T. D., Bagos, P. G., & Hamodrakas, S. J. (2010). ExTopoDB: a database of experimentally derived topological models of transmembrane proteins. *Bioinformatics, 26*(19), 2490-2492.

Tsirigos, K. D., Bagos, P. G., & Hamodrakas, S. J. (2011). OMPdb: a database of {beta}-barrel outer membrane proteins from Gram-negative bacteria. *Nucleic acids research, 39*(Database issue), D324-331.

Umemura, M., Nagano, N., Koike, H., Kawano, J., Ishii, T., Miyamura, Y., . . . Machida, M. (2014). Characterization of the biosynthetic gene cluster for the ribosomally synthesized cyclic peptide ustiloxin B in Aspergillus flavus. *Fungal Genet Biol, 68*, 23-30.

UniProt. (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic acids research, 42*(Database issue), D191-198.

Vroling, B., Sanders, M., Baakman, C., Borrmann, A., Verhoeven, S., Klomp, J., . . . Vriend, G. (2011). GPCRDB: information system for G protein-coupled receptors. *Nucleic acids research, 39*(Database issue), D309-D319.

Wang, C. K., Kaas, Q., Chiche, L., & Craik, D. J. (2008). CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic acids research, 36*(suppl 1), D206-D210.

Wong, W. C., Maurer-Stroh, S., & Eisenhaber, F. (2010). More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology. *PLoS Comput Biol, 6*(7), e1000867.

Wren, J. D. (2008). URL decay in MEDLINE—a 4-year follow-up study. *Bioinformatics, 24*(11), 1381-1385.

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., & Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res, 30*(1), 303-305.

Xia, J., Wang, Q., Jia, P., Wang, B., Pao, W., & Zhao, Z. (2012). NGS catalog: A database of next generation sequencing studies in humans. *Hum Mutat, 33*(6), E2341-2355.

Κεφάλαιο 3: Αλγόριθμοι Στοίχισης Αλληλουχιών

Σύνοψη
Στο κεφάλαιο αυτό θα παρουσιαστούν αρχικά, τα απαραίτητα μαθηματικά μοντέλα που περιγράφουν τις αλληλουχίες μακρομορίων και κάποια βασικά ασυμπτωτικά αποτελέσματα που αναφέρονται σε αυτές. Στη συνέχεια θα παρουσιαστούν τα βασικά θεωρητικά αποτελέσματα που αφορούν στη στοίχιση βιολογικών αλληλουχιών. Θα παρουσιαστούν οι τρόποι βαθμονόμησης της στοίχισης, οι τρόποι εύρεσης της στοίχισης, καθώς και τα διαφορετικά είδη αλγορίθμων στοίχισης, ενώ ιδιαίτερη έμφαση θα δοθεί στην αξιολόγηση της στατιστικής σημαντικότητας μιας στοίχισης. Τέλος, θα παρουσιαστούν οι βασικοί ευριστικοί αλγόριθμοι τοπικής στοίχισης (FASTA, BLAST), οι οποίοι χρησιμοποιούνται καθημερινά στη βιοπληροφορική.

Προαπαιτούμενη γνώση
Προαπαιτούμενη γνώση για το κεφάλαιο αυτό, είναι η γνώση των βασικών νόμων των πιθανοτήτων και στοιχειώδεις γνώσεις σχετικά με τις βιολογικές αλληλουχίες..

3. Εισαγωγή

Η ομοιότητα αλληλουχιών είναι ένα από τα θεμελιώδη ζητήματα στη Βιοπληροφορική, καθώς πλέον αποτελεί αναπόσπαστο τμήμα των αναλύσεων που πραγματοποιεί καθημερινά οποιοσδήποτε ασχολείται με το γνωστικό αυτό αντικείμενο, αλλά, ακόμα περισσότερο, ο καθένας που ασχολείται ερευνητικά με τη μοριακή βιολογία με οποιονδήποτε τρόπο. Η ομοιότητα των βιολογικών αλληλουχιών τις περισσότερες φορές υποδηλώνει ομολογία (δηλαδή, κοινή εξελικτική προέλευση), και κατά συνέπεια (ειδικά για τις πρωτεΐνες), παρόμοια τρισδιάστατη δομή και παρόμοια λειτουργία.

Τα προβλήματα που καλείται κάποιος να λύσει, όταν μελετάει την ομοιότητα αλληλουχιών, είναι πολλαπλά. Με ποιον αλγόριθμο θα πραγματοποιήσει την «στοίχιση» των δύο αλληλουχιών (δηλαδή την εύρεση της καλύτερης περιοχής ομοιότητας τους); Πώς θα ποσοτικοποιήσει αυτή την ομοιότητα; Τι υποθέσεις θα αναγκαστεί να κάνει; Και τέλος, πως θα αξιολογήσει αν μια στοίχιση είναι σημαντική ή όχι; Το τελευταίο, είναι ίσως και το σπουδαιότερο από τα θέματα αυτά, γιατί όλοι καταλαβαίνουν ότι αν δυο πρωτεϊνικές αλληλουχίες είναι ταυτόσημες λ.χ. σε ποσοστό 99%, τότε υπάρχει πολύ μεγάλη πιθανότητα να είναι και όμοιας δομής και παρόμοιας λειτουργίας (εκτός ίσως από τις περιπτώσεις στις οποίες οι λίγες αλλαγές συμβαίνουν στο ενεργό κέντρο ενός ενζύμου και αναστέλλουν τη δράση του). Με 80% ομοιότητα περιμένουμε ότι πάλι οι πρωτεΐνες θα έχουν μεγάλη ομοιότητα στη δομή. Ποιο είναι όμως το όριο όσο κατεβαίνουμε στο επίπεδο ομοιότητας; Παραδοσιακά οι βιολόγοι χρησιμοποιούν τον εμπειρικό κανόνα του «30% ομοιότητα σε μήκος στοίχισης μεγαλύτερο από 80 αμινοξικά κατάλοιπα», κανόνας που σε γενικές γραμμές λειτουργεί σωστά, αλλά χρειαζόμαστε περισσότερη ακρίβεια σε τέτοια ζητήματα, ειδικά όσο οι βάσεις δεδομένων μεγαλώνουν και οι πιθανότητες εμφάνισης μιας τυχαίας ομοιότητας αυξάνονται.

Στο κεφάλαιο αυτό, θα προσπαθήσουμε να παρουσιάσουμε τα βασικά θεωρητικά εργαλεία που θα μας βοηθήσουν να καταλάβουμε τις απαντήσεις των παραπάνω ερωτημάτων, αλλά επίσης και να αναγνωρίσουμε τους περιορισμούς τους. Για το λόγο αυτό, θα ξεκινήσουμε από τη στατιστική μελέτη των βιολογικών αλληλουχιών και θα παρουσιάσουμε το βασικό μοντέλο της ανεξαρτησίας, το οποίο αποτελεί το «θεωρητικό» ή, σε μια πιο στατιστική ορολογία, αποτελεί το μοντέλο από το οποίο προκύπτει η «μηδενική υπόθεση» έναντι της οποίας θα συγκρίνουμε τα ευρήματα μιας αναζήτησης ομοιότητας, έτσι ώστε να μπορέσουμε να καταλάβουμε αν η δεδομένη στοίχιση είναι «σημαντική» ή όχι. Στη συνέχεια θα παρουσιαστούν οι κύριοι αλγόριθμοι εύρεσης ομοιότητας, και θα συζητηθούν πρακτικά θέματα που προκύπτουν, ειδικά σε αναζητήσεις σε βάσεις δεδομένων.

3.1. Η ακολουθία ως σειρά ανεξάρτητων γεγονότων

Το πιο απλό μοντέλο που περιγράφει μια βιολογική αλληλουχία (DNA, RNA ή πρωτεΐνης) είναι το μοντέλο της ανεξαρτησίας, δηλαδή το μοντέλο που θεωρεί ότι η αλληλουχία των γραμμάτων του αλφάβητου Ω, -στην περίπτωση του DNA των τεσσάρων νουκλεοτιδίων-, είναι μια σειρά *n* ανεξάρτητων δοκιμών με τέσσερις διακριτές εκβάσεις. Οι πιθανότητες για τα 4 ενδεχόμενα (A, T, G, C) είναι αντίστοιχα:

$$p_A, p_T, p_G, p_C \quad \text{με } p_k \geq 0 \text{ και } \sum_{k\in\{A,T,G,C\}} p_k = 1.$$

Όμοια, ισχύουν και στην περίπτωση των πρωτεϊνών, μόνο που θα έχουμε 20 διαφορετικά σύμβολα και 20 διαφορετικές πιθανότητες. Μια δεδομένη ακολουθία DNA, $\mathbf{x}=x_1, x_2,...,x_n$ με $x_i \in \{A,T,G,C\}$ έχει συνολική πιθανότητα να παρατηρηθεί κάτω από τις προϋποθέσεις του «τυχαίου» αυτού μοντέλου ίση με:

$$p_{ολ} = P(\mathbf{x}) = \prod_{i=1}^{n} p_{X_i}$$

Προφανώς το άθροισμα των πιθανοτήτων όλων των πιθανών ακολουθιών (που είναι όσες οι δυνατές διατάξεις των 4 στοιχείων ανά *n* με επανάληψη δηλαδή 4^n) είναι ίσο με 1 δηλαδή:

$$\sum_j P(\mathbf{x}_j) = 1.$$

Έστω **x** μια τέτοια τυχαία ακολουθία *n* βάσεων του DNA. Η συχνότητα της εμφάνισης των 4 βάσεων ακολουθεί την πολυωνυμική κατανομή, δηλαδή:

$$P\left(n_A, n_T, n_G, n_C\right) = \frac{n!}{n_A!\, n_T!\, n_G!\, n_C!}\, p_A^{n_A}\, p_T^{n_T}\, p_G^{n_G}\, p_C^{n_C} \tag{3.1}$$

Αν τώρα θεωρήσουμε τις συχνότητες εμφάνισης κάθε μιας από τις βάσεις ξεχωριστά, τότε αυτές ακολουθούν τη διωνυμική κατανομή, δηλαδή :

$$P\left(X = x\right) = \binom{n}{x} p_A^x \left(1 - p_A\right)^{n-x} \tag{3.2}$$

και όμοια για τις άλλες 3 βάσεις (T, G, C). Έτσι η μια ακολουθία των βάσεων του DNA μπορεί να θεωρείται ως μια ακολουθία ανεξάρτητων δοκιμών Bernoulli με $p=p_A$ και $q=1-p_A$. Όμοια θεώρηση μπορεί να γίνει και για τις άλλες 3 βάσεις. Προφανώς η κατανομή των συχνοτήτων εμφάνισης των βάσεων του DNA (ή των αμινοξέων μιας πρωτεΐνης) δεν είναι επαρκής πληροφορία για να περιγράψει τη βιολογική πληροφορία μιας δεδομένης αλληλουχίας. Η βιολογική σημασία μιας αλληλουχίας βάσεων (ή αμινοξέων) έγκειται στην ακριβή αλληλουχία των 4 βάσεων (ή των 20 αμινοξέων), δηλαδή στον τρόπο που το ένα σύμβολο διαδέχεται το άλλο. Εντούτοις, η παραπάνω θεώρηση της τυχαίας και ανεξάρτητης εμφάνισης των συμβόλων μας είναι ιδιαίτερα χρήσιμη καθώς μας προμηθεύει με μια μηδενική υπόθεση (H_0) έναντι της οποίας θα μπορούμε να συγκρίνουμε μια δεδομένη αλληλουχία για να διαπιστώσουμε αν η -συγκεκριμένη αλληλουχία- είναι δυνατόν να έχει προκύψει τυχαία, ή αν, αντίθετα, έχει κάποια βιολογική σημασία (Durbin, Eddy, Krogh, & Mithison, 1998).

Στο μοντέλο αυτό, μια επιπλέον υπόθεση που μπορούμε να κάνουμε (αν δεν έχουμε λόγους να πιστεύουμε το αντίθετο, όπως για παράδειγμα αν έχουμε μια αλληλουχία από ένα γονιδίωμα με γνωστές τις συχνότητες εμφάνισης των βάσεων) είναι ότι τα ενδεχόμενα εμφάνισης των βάσεων εκτός από ανεξάρτητα είναι και ισοπίθανα, δηλαδή:

$$p_A = p_T = p_G = p_C = 1/4$$

Πριν προχωρήσουμε παρακάτω θα πρέπει να κάνουμε μια μικρή παρένθεση για να παραθέσουμε κάποιους ορισμούς δανεισμένους από την Θεωρία Πληροφορίας (Information Theory). Συγκεκριμένα θα αποδώσουμε τον ορισμό της έννοιας της εντροπίας και της πληροφορίας. Μια δεδομένη αλληλουχία DNA, όπως την ορίσαμε παραπάνω, λέμε ότι έχει συνάρτηση εντροπίας κατά Shannon ίση με:

$$H(\mathbf{x}) = -\sum_i P(x_i) \log P(x_i) \tag{3.3}$$

Η εντροπία γίνεται μέγιστη όταν οι βάσεις είναι ισοπίθανες, δηλαδή όταν $p_A=p_G=p_T=p_C=1/4$ οπότε θα έχει τιμή ίση με $H(\mathbf{x})=\sum(1/4)\log(1/4)=\log 4$. Συνήθως σε αυτές τις περιπτώσεις παίρνουμε λογάριθμους με βάση το 2, έτσι ώστε η μονάδα μέτρησης να είναι το bit. Η πληροφορία μιας ακολουθίας ορίζεται ως:

$$I(\mathbf{x}) = H_{\max} - H_{obs} \tag{3.4}$$

άρα αν έχουμε μια αλληλουχία με σύσταση βάσεων διαφορετική από την αναμενόμενη με βάση το τυχαίο μοντέλο η εντροπία της θα είναι μικρότερη από τα 2 bits, και η πληροφορία που φέρει αυτή η αλληλουχία θα είναι μεγαλύτερη από το 0.

Ένα διαφορετικό μέτρο για το πληροφοριακό περιεχόμενο μιας βιολογικής αλληλουχίας, έχει δοθεί από τους (Wootton & Federhen, 1993) και βασίζεται επίσης στη Θεωρία Πληροφορίας. Αυτό το μέτρο, που ονομάστηκε "πολυπλοκότητα της σύνθεσης" (compositional complexity), ορίζεται, για ένα παράθυρο μήκους *k* της ακολουθίας, ως εξής:

$$K = \frac{1}{k}\log_{N_\Omega}\left(\frac{k!}{\prod_{\forall s\in\Omega} n_s!}\right) \tag{3.5}$$

Στην παραπάνω σχέση, το n_s είναι ο αριθμός εμφανίσεων του συμβόλου *s* στο παράθυρο και N_Ω το μέγεθος του αλφάβητου (4 για τα νουκλεοτίδια, 20 για τα αμινοξέα). Διαισθητικά, το μέτρο αυτό δείχνει την ποσότητα της πληροφορίας που απαιτείται σε κάθε θέση της ακολουθίας για να καθορίσει κανείς το σύμβολο (της θέσης), δεδομένης της σύνθεσης όλου του παραθύρου. Για παράδειγμα, ένα παράθυρο 4 νουκλεοτιδίων με σύσταση AAAA, θα έχει πολυπλοκότητα ίση με

$$K = \frac{1}{4}\log_4\left(\frac{4!}{4!0!0!0!}\right) = \frac{1}{4}\log_4(1) = 0\,.$$

Πράγμα που σημαίνει, ότι αν ξέρουμε την ακολουθία του παραθύρου, δεν χρειαζόμαστε καμιά άλλη πληροφορία για να βρούμε ποιο κατάλοιπο βρίσκεται σε μια δεδομένη θέση. Αντίθετα, ένα παράθυρο με σύσταση ATGC θα έχει

$$K = \frac{1}{4}\log_4\left(\frac{4!}{1!1!1!1!}\right) = \frac{1}{4}\log_4(24) = 0.573$$

Οι Wootton και Federhen χρησιμοποίησαν επίσης και την εντροπία, δίνοντας όμως έναν ισοδύναμο ορισμό:

$$H_k = -\sum_{\forall s\in\Omega}\frac{n_s}{k}\left(\log_2\frac{n_s}{k}\right) \tag{3.6}$$

Στην ίδια εργασία, έδειξαν ότι η εντροπία και η πολυπλοκότητα, είναι ασυμπτωτικά ισοδύναμες ποσότητες (δηλαδή, όταν το παράθυρο είναι πολύ μεγάλο θα δίνουν το ίδιο αποτέλεσμα), ενώ έκαναν την παρατήρηση ότι η πολυπλοκότητα όπως την όρισαν, είναι σύμφωνη με τον ορισμό περί εντροπίας του Boltzman (σε αντίθεση με τον κλασικό ορισμό της έννοιας της εντροπίας κατά Shannon).

Η χρήση της εντροπίας, της πολυπλοκότητας και της πληροφορίας, βρίσκουν πολλές εφαρμογές σε προκαταρκτικές περιγραφικές αναλύσεις γονιδιωμάτων, αλλά και σε άλλες πιο εξειδικευμένες αναλύσεις. Οι Wootton και Federhen για παράδειγμα, χρησιμοποίησαν τα μέτρα αυτά για τον εντοπισμό περιοχών χαμηλής πολυπλοκότητας σε αμινοξικές αλληλουχίες πρωτεϊνών ή σε γονιδιώματα. Η ανεύρεση τέτοιων περιοχών είναι σημαντική, γιατί στην περίπτωση αμινοξικών αλληλουχιών πρωτεϊνών η ύπαρξή τους μπορεί να επηρεάσει τα στατιστικά της στοίχισης και τα αποτελέσματα της αναζήτησης ομοιότητας (βλ. παρακάτω), ενώ στην περίπτωση DNA μπορεί να σηματοδοτεί την ύπαρξη ρυθμιστικών περιοχών. Τέλος, όπως θα δούμε στο επόμενο κεφάλαιο, η εντροπία χρησιμεύει στην περιγραφή και στην ποσοτικοποίηση μιας πολλαπλής στοίχισης αλληλουχιών.

Μια άλλη σχετική έννοια, είναι αυτή της σχετικής εντροπίας (Relative Entropy). Η σχετική εντροπία δυο καταστάσεων *P*, *Q* (γνωστή και ως μέτρο της απόστασης των Kullback-Leibler) εκφράζει τη σχετική απόσταση, ή διαφορά, μεταξύ των δυο καταστάσεων και δίνεται από τον τύπο:

$$H(P,Q) = \sum_i P(x_i)\log\frac{P(x_i)}{Q(x_i)} \tag{3.7}$$

Το $P(x_i)$ είναι όπως είδαμε παραπάνω η πιθανότητα εμφάνισης μιας βάσης (A,T,G,C) στην *i* θέση της συγκεκριμένης ακολουθίας, ενώ το $Q(x_i)$ η αντίστοιχη πιθανότητα εμφάνισης μιας βάσης σε μια άλλη ακολουθία. Αυτή η άλλη ακολουθία μπορεί να είναι μια άλλη πραγματική ακολουθία με την οποία θέλουμε να συγκρίνουμε την πρώτη, ή να είναι μια θεωρητική κατανομή, όπως αυτή που υποθέτει ισοπίθανη ή τυχαία εμφάνιση των βάσεων. Προφανώς αν $Q(x_i)=1/4$ (ισοκατανομή των βάσεων) τότε $H(P,Q)=I(P)$

Μια άλλη πολύ σημαντική έννοια που θα ξανασυναντήσουμε και στα επόμενα κεφάλαια είναι αυτή της αμοιβαίας πληροφορίας (Mutual Information). Δυο τ.μ *X,Y* έχουν αμοιβαία πληροφορία που δίνεται από τη σχέση:

$$M(X,Y)=\sum_{i,j}P(x_i,y_j)\log\frac{P(x_i,y_j)}{P(x_i)P(y_j)} \tag{3.8}$$

Σε αυτή την περίπτωση, έχουμε δυο ακολουθίες, **x** και **y**. Η αμοιβαία πληροφορία μετράει πόση διάφορα έχει η από κοινού κατανομή της σ.π. των X και Y που συμβολίζουμε με $P(x_i,y_i)$, με την υποθετική από κοινού κατανομή που θα είχαν αν ήταν ανεξάρτητες με $P(x_i,y_i)=P(x_i)P(y_i)$. Προφανώς $P(x_i)$ και $P(y_i)$ είναι οι περιθώριες σ.π. των X,Y αντίστοιχα. Δηλαδή, η αμοιβαία πληροφορία μετράει το «πόσο ανεξάρτητες» είναι οι δυο κατανομές. Η σχετική εντροπία και η αμοιβαία πληροφορία, βρίσκουν πολλές εφαρμογές όταν μελετάμε ταυτόχρονα πολλές ακολουθίες και σχετικά παραδείγματα θα δούμε στο κεφάλαιο που περιγράφει την πολλαπλή στοίχιση.

3.2. Ροές - Νόμος Erdos και Renyi

Το επόμενο θέμα που θα μας απασχολήσει είναι γνωστό στη βιβλιογραφία ως το πρόβλημα της μέγιστης ροής όμοιων αποτελεσμάτων (longest run of heads). Η πιο απλή του εφαρμογή είναι η απάντηση στο ερώτημα «ποια είναι η αναμενόμενη τιμή για το μέγιστο αριθμό επαναλήψεων-κορώνα ή γράμματα-σε μια διαδοχική σειρά από n διαδοχικά στριψίματα ενός νομίσματος (δίτιμες δοκιμές Bernoulli)». Το θέμα αυτό είναι πολύ σημαντικό, καθώς στη θεωρία των ροών βασίζονται τα στατιστικά της τοπικής στοίχισης ακολουθιών, τα οποία θα μελετήσουμε παρακάτω.

A G G C G A T A A A A A A A A A A A A A A A C G G A T G C A T C G

Εικόνα 3.1: Μια ροή από 16 συνεχόμενες A σε ένα μόριο DNA

Η πρώτη απάντηση που δόθηκε στο ερώτημα αυτό είναι γνωστή ως νόμος του $\log(n)$, ή αλλιώς γνωστός ως νόμος των Erdos και Renyi (Erdos & Renyi, 1970). Το θεώρημα λέει ότι σε μια ακολουθία n ανεξάρτητων δοκιμών Bernoulli με πιθανότητα «επιτυχίας» p, με $0\le p\le 1$, το αναμενόμενο μήκος R_n μέγιστης δυνατής ροής ευνοϊκών αποτελεσμάτων, είναι ίσο κατά προσέγγιση με $\log_{1/p}(n)$ ή αλλιώς:

$$\frac{R_n}{\log_{1/p}(n)}\to 1 \text{ με πιθανότητα 1.} \tag{3.9}$$

Η απόδειξη είναι αρκετά περίπλοκη αλλά μια διαισθητική ερμηνεία του αποτελέσματος μπορεί να γίνει ως εξής (Waterman, 1995): αν το ευνοϊκό αποτέλεσμα έχει πιθανότητα p τότε μια ροή k συνεχών ευνοϊκών αποτελεσμάτων έχει πιθανότητα p^k. Αν έχουμε n επαναλήψεις ($n\to+\infty$) τότε έχουμε περίπου n δυνατές ροές και

$$E(\text{αριθμός ροών μήκους } x)=np^k$$

Αν τώρα, η μέγιστη ροή είναι μοναδική, το μήκος της, R_n, ικανοποιεί τη σχέση $1=np^k$, άρα:

$$R_n=\log_{1/p}(n)$$

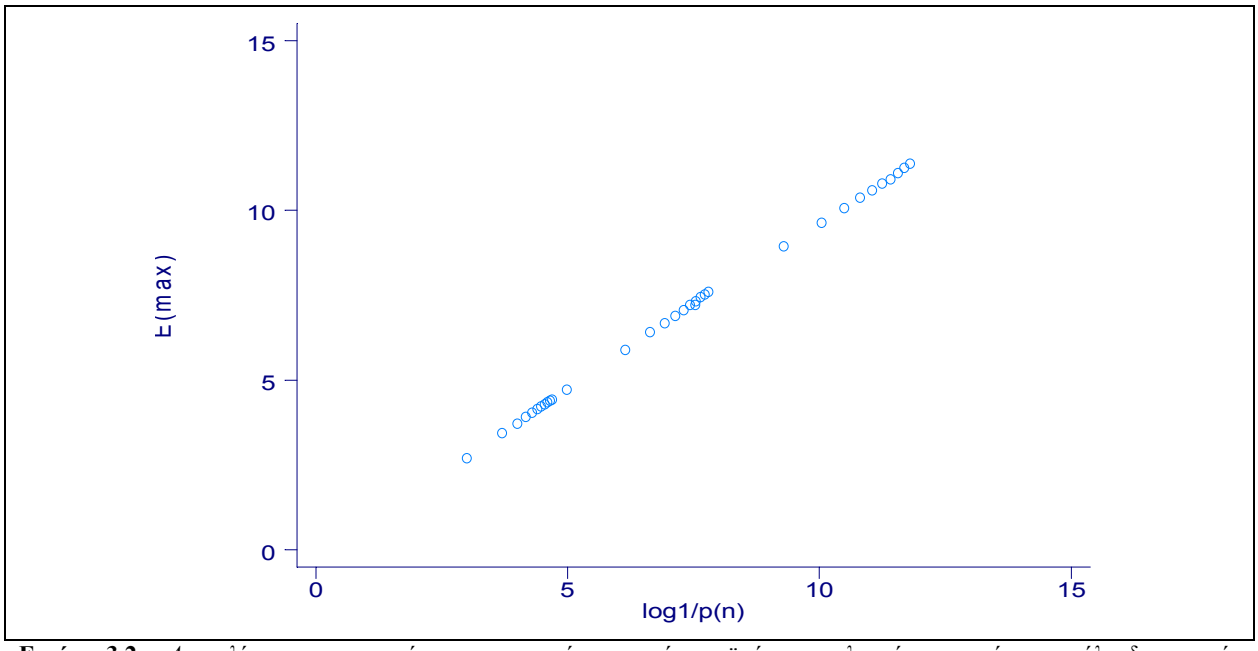
Παρατηρούμε, ότι όσο μεγαλώνει το μήκος της ακολουθίας, τόσο μεγαλώνει και το μήκος της μέγιστης ροής που αναμένουμε να βρούμε λόγω τύχης.

Παράδειγμα 3.2.1

Σε μια ακολουθία n=10000 βάσεων του DNA, θεωρώντας αυτές ισοπίθανες (δηλαδή p_κ=1/4), μας ενδιαφέρει να βρούμε την αναμενόμενη τιμή για τον αριθμό των μέγιστων επαναλήψεων A που μπορεί να έχει συμβεί κατά τύχη. Θεωρώντας ότι η αλληλουχία είναι τυχαία, τότε το αναμενόμενο μήκος της μέγιστης ροής από A θα είναι:

$$R_n=\log_{1/p}(n)\Rightarrow R_n=\log_4(10000)\Rightarrow R_n=\frac{\log_{10}10000}{\log_{10}4}=\frac{4}{0.60205}=6.64$$

Στα παρακάτω διαγράμματα φαίνονται τα αποτελέσματα των προσομοιώσεων για τη μέση τιμή της ροής ευνοϊκών αποτελεσμάτων για n=1000 έως 50000 και για p=0.1, 0.25, 0.4 (1000 επαναλήψεις) οι οποίες επιβεβαιώνουν τη σχέση (3.9).



Εικόνα 3.2: *Αποτελέσματα προσομοίωσης για τη μέγιστη ροή ευνοϊκών αποτελεσμάτων, σε ένα μοντέλο διωνυμικής κατανομής με πιθανότητες p=0.1, 0.25 και 0.4. Το n κυμαίνεται από 1.000 μέχρι 50.000.*

3.3. Επεκτάσεις στον Νόμο Erdos και Renyi

Μέχρι τώρα ασχοληθήκαμε μόνο με τον αριθμό των επαναλήψεων σε μια σειρά ανεξάρτητων δοκιμών. Παρ' όλα αυτά όμως, ξέρουμε ότι υπάρχουν περιπτώσεις στις οποίες μπορεί να μας ενδιαφέρει ο αριθμός των δοκιμών που περιέχουν π.χ. 90% επαναλήψεις από «επιτυχίες». Ένα τέτοιο παράδειγμα, είναι ο εντοπισμός περιοχών με «πολλά» και συνεχόμενα υδρόφοβα κατάλοιπα σε μια πρωτεΐνη. Ξέρουμε ότι οι περιοχές με συνεχόμενα υδρόφοβα κατάλοιπα είναι πιθανό να είναι διαμεμβρανικά τμήματα, αλλά δεν αναμένουμε να συναντήσουμε περιοχές αποκλειστικά με υδρόφοβα αμινοξέα (είναι γνωστό, ότι ακόμα και μέσα σε πραγματικά διαμεμβρανικά τμήματα πρωτεϊνών συναντάμε περιστασιακά 1-2 πολικά κατάλοιπα). Επίσης, ένα άλλο χαρακτηριστικό που μπορεί να μας ενδιαφέρει, είναι το να προσδιορίσουμε τη στατιστική σημαντικότητα μιας τέτοιας παρατήρησης.

Σ' αυτήν την ενότητα θα παρουσιάσουμε κάποια πορίσματα της θεωρίας των μεγάλων αποκλίσεων (Large Deviation Theory) και θα τα χρησιμοποιήσουμε για να επεκτείνουμε τα αποτελέσματα των προηγούμενων παραγράφων. Στην Εικόνα 3.3 φαίνεται γραμμοσκιασμένη μια περιοχή 20 βάσεων η οποία περιέχει 16 Αδενίνες. Σε μια αλληλουχία που θεωρείται τυχαία (και άρα η συχνότητα εμφάνισης των βάσεων δεν έχει λόγο να αποκλίνει από τη συνολική συχνότητα εμφάνισης σε όλη την αλληλουχία), μας ενδιαφέρει το πόσο συχνά μπορεί να εμφανιστεί μια τέτοια περιοχή.

A G G C G A T A A A A A A A A T A A G A C C A A A A A C G G A T G C A T

Εικόνα 3.3: *Μια ροή 20 νουκλεοτιδίων που περιέχει 80% A.*

Η σχετική εντροπία δυο καταστάσεων α, p εκφράζει τη σχετική απόσταση, δηλαδή τη διαφορά μεταξύ των δυο καταστάσεων και, ειδικά για την περίπτωση της διωνυμικής κατανομής δίνεται από τον τύπο:

$$H(\alpha, p) \equiv \alpha \log\left(\frac{\alpha}{p}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{1-p}\right) = \log\frac{a^a(1-a)^{1-a}}{p^a(1-p)^{1-a}} = -\log\left(\frac{p}{a}\right)^a\left(\frac{1-p}{1-a}\right)^{1-a} \qquad (3.10)$$

Η συνάρτηση αυτή μετρά τη διαφορά μεταξύ της κατανομής B(k,p) από την οποία προέρχονται τα δεδομένα μας (η οποία έχει δώσει γένεση σε μια ακολουθία DNA με πιθανότητα εμφάνισης των βάσεων ίση με p), και μιας άλλης, υποθετικής, B(k,α) για την οποία υποπτευόμαστε ότι έχει δώσει γένεση σε μια τοπική υπό-ακολουθία μήκους n στην οποία παρατηρούμε ότι για παράδειγμα η εμφάνιση μιας βάσης, διαφέρει πολύ

από την αναμενόμενη καθώς έχει συχνότητα $\alpha=s/k$. Προφανώς, $0\le p,a\le 1$. Το κλειδί στην κατανόηση των μεγάλων αποκλίσεων, είναι το γεγονός ότι έχουμε να κάνουμε με δυο διαφορετικές πιθανότητες (α,p) στον ίδιο χώρο πιθανών εκβάσεων. Ένα από τα αποτελέσματα της θεωρίας μεγάλων αποκλίσεων, έχει αρκετό ενδιαφέρον και βρίσκει εφαρμογές στον υπολογισμό διωνυμικών πιθανοτήτων. Συγκεκριμένα, αν έχουμε $0\le p\le a\le 1$ και $Y\sim B(k,p)$, τότε μια προσεγγιστική σχέση για την διωνυμική πιθανότητα $P(Y\ge ak)$, δίνεται από τον τύπο:

$$P(Y \ge \alpha k) \approx e^{-kH(a,p)} \tag{3.11}$$

Τώρα που έχουμε δώσει τον ορισμό της έννοιας της σχετικής εντροπίας μπορούμε να προχωρήσουμε και να επεκτείνουμε τη σχέση (3.9). Το αποτέλεσμα αυτό, λέει ότι σε μια ακολουθία n ανεξάρτητων δοκιμών Bernoulli με πιθανότητα «επιτυχίας» p, με $0\le p\le a\le 1$, το πλήθος R_n^a διαδοχικών δοκιμών που περιέχουν $100\alpha\%$ ευνοϊκά αποτελέσματα, ικανοποιεί τη σχέση (Erdos & Renyi, 1970; Erdos & Revesz, 1975):

$$\frac{R_n^a}{\log(n)} \to \frac{1}{H(a,p)} \text{ με πιθανότητα 1} \tag{3.12}$$

Μια διαισθητική ερμηνεία του αποτελέσματος έχει ως εξής: Από τη θεωρία των μεγάλων αποκλίσεων (Large Deviations) της σχέσης (3.11) βρίσκουμε ότι μια περιοχή μήκους k η οποία περιέχει $100\alpha\%$ ευνοϊκά αποτελέσματα, έχει πιθανότητα περίπου ίση με $e^{-kH(a,p)}$. Επειδή τώρα κάθε ροή έχει περίπου $n-k+1\approx n$ δυνατές περιοχές έναρξης έχουμε:

$$1 = ne^{-kH(a,p)} \Rightarrow R_n^a = \frac{\log(n)}{H(a,p)}$$

Παρατηρούμε, με την χρήση του κανόνα De L' Hospital, ότι για $\alpha=1\to H(\alpha,p)=\log(1/p)$ και τα αποτελέσματα συμφωνούν με τη σχέση (3.9).

Παράδειγμα 3.3.1

Σε μια αλληλουχία 10.00.000 βάσεων DNA, η αναμενόμενη τιμή R_n^a για τη μέγιστη περιοχή ροή που θα περιέχει κατ' ελάχιστο 80% βάσεις Αδενίνης (A) είναι :

$$R_n^a = \frac{\log(n)}{H(a,p)} = \frac{\log(1000000)}{0.666} = 20.744$$

(να σημειωθεί εδώ ότι όταν γράφουμε log εννοούμε λογάριθμο με βάση το e)

3.4. Η Κατανομή της Μέγιστης Ροής - Η Κατανομή των Ακραίων Τιμών (EVD)

Επεκτείνοντας τα προηγούμενα, πολλές φορές μπορεί να χρειαστεί να βρούμε την προσεγγιστική κατανομή που ακολουθεί η τυχαία μεταβλητή του μήκους της μέγιστης ροής ενός αποτελέσματος. Η πλήρης κατανομή, μας είναι χρήσιμη, και από θεωρητική σκοπιά, αλλά κυρίως από πρακτική, γιατί με τη γνώση της κατανομής θα μπορούμε να πραγματοποιήσουμε έλεγχο υποθέσεων (χρειαζόμαστε εκτός από τη μέση τιμή, και τη διασπορά της τ.μ.). Όταν αναφερόμαστε σε μέγιστα (ή και σε ελάχιστα) μιας ακολουθίας τυχαίων μεταβλητών καταλήγουμε συνήθως στις κατανομές των ακραίων τιμών (Extreme Value Distributions). Πιο αυστηρά, αν έχουμε ένα δείγμα $X_1, X_2, ..., X_n$, από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές (iid) τότε μας ενδιαφέρει η οριακή κατανομή του:

$$M_n = a_n[\max(X_1, X_2, ..., X_n) - b_n], n \to \infty$$

Όπου αn, bn κατάλληλες σταθερές κανονικοποίησης τέτοιες ώστε να προκύπτει μη τετριμμένη κατανομή. Η απάντηση είναι ότι αν υπάρχει μια μη τετριμμένη και ορισμένη ΑΣΚ (cdf) για κάποιες ακολουθίες α_n, b_n, τότε πρέπει να ανήκει σε μια από τις περιπτώσεις (Davison, 1998):

1. $F(y) = \exp(-e^{-y}), -\infty \le y \le \infty$ (Gumbel)
2. $F(y) = \begin{cases} 0, y \le 0 \\ \exp(-y^{-a}), y \ge 0, a > 0 \end{cases}$ (Frechet)
3. $F(y) = \begin{cases} \exp(-(-y)^{a}), y < 0, a > 0 \\ 1, y \ge 0 \end{cases}$ (Weibull)

Η κατανομή που αφορά την δική μας περίπτωση είναι αυτή του Gumbel, και προκύπτει από την γενικευμένη μορφή της κατανομής των ακραίων τιμών (Generalized Extreme Value Distribution – GEVD):

$$H(y) = \exp\left\{-\left(1+k\left(\frac{y-a}{b}\right)\right)^{-\frac{1}{k}}\right\} \text{ με } -\infty < \alpha, k < \infty, b > 0 \tag{3.13}$$

η οποία ορίζεται όταν $1+k\left(\frac{y-a}{b}\right) > 0$, ως το όριο καθώς $k \to 0$ (οι άλλες δύο μορφές αντιστοιχούν στην περίπτωση που k>0 (Frechet) και k<0 (Weibull) αντίστοιχα). Αν θέσουμε $z = \left(\frac{y-a}{b}\right)$, και $t = -\frac{1}{k}$ θα έχουμε:

$$H(y) = \exp\left\{-\left(1-\frac{z}{t}\right)^{t}\right\}$$

και αν πάρουμε το όριο καθώς το $k \to 0 \Rightarrow t \to \infty$ επειδή είναι γνωστή η σχέση:

$$\lim_{n\to\infty}\left(1-\frac{a}{n}\right)^{n} = e^{-a}$$

θα έχουμε

$$\lim_{t\to\infty} H(y) = \lim_{t\to\infty} \exp\left\{-\left(1-\frac{z}{t}\right)^{t}\right\} = \exp\left\{-e^{-z}\right\} = \exp\left\{-e^{-\left(\frac{y-a}{b}\right)}\right\}$$

Έτσι η κατανομή του $Y_n = \max(X_1, X_2, ..., X_n)$ γίνεται (Gumbel, 1958):

$$F(Y) = \exp(-e^{-\frac{(y-a)}{b}}), -\infty \le y \le \infty \tag{3.14}$$

$$E = a - b\Gamma'(1), \quad V = \frac{b^2\pi^2}{6} \tag{3.15}$$

Αποδεικνύεται (Arratia, Gordon, and Waterman, 1986; Arratia, Gordon, and Waterman, 1990; Waterman, 1995), ότι στην περίπτωση της συνεχούς ροής ενός αποτελέσματος (νόμισμα η βάσεις DNA) για $a_n = \frac{\log(qn)}{\lambda}, b_n = \frac{1}{\lambda}$ όπου $\lambda = \log(1/p)$ ισχύει:

$$\lim_{n\to\infty}\left(R_n < \frac{\log(nq)}{\lambda} + \frac{y}{\lambda}\right) = \exp\left(-e^{-y}\right)$$

Για την ΑΣΚ της τ.μ. R_n θα ισχύει:

$$F(y) = P(R_n \le y) \approx \exp\left(-\exp\left(-\frac{y - \log(nq)/\lambda}{1/\lambda}\right)\right) \tag{3.16}$$

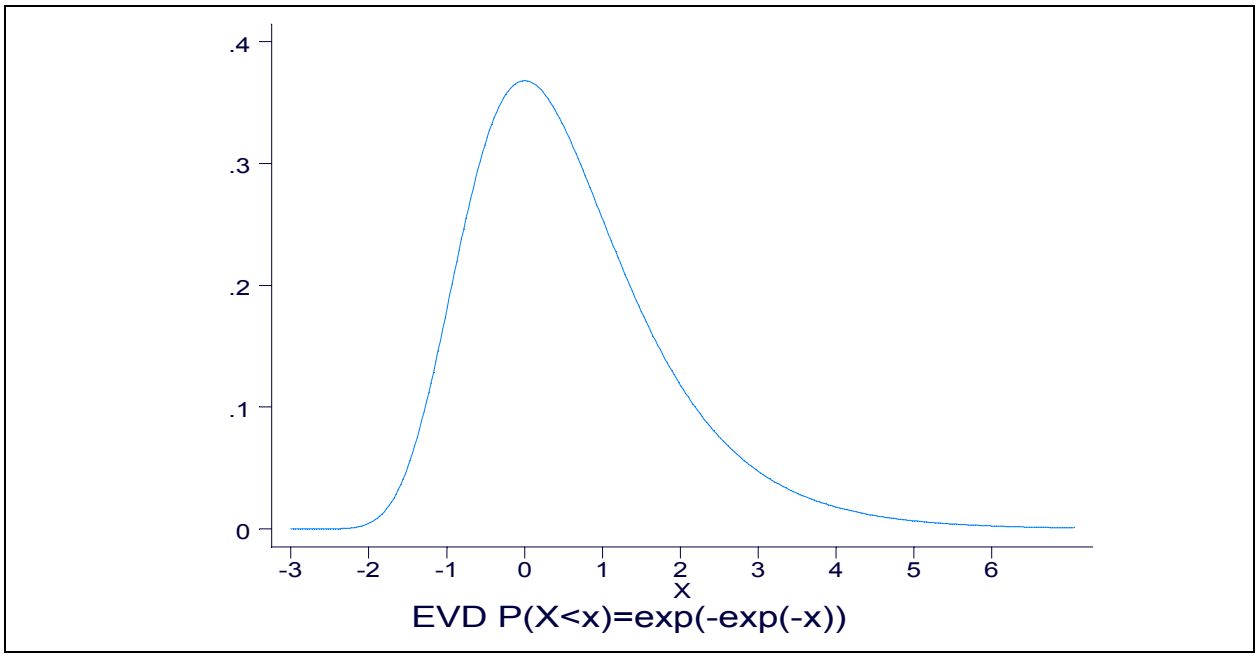
Από τα παραπάνω προκύπτει, ότι η κατανομή της μέγιστης ροής είναι αυτή των ακραίων τιμών του Gumbel. Δηλαδή, οι σχέσεις (3.16) και (3.14) είναι ισοδύναμες. Κατά συνέπεια, θα έχουμε:

$$E(R_n) \approx \frac{\log(n)}{\lambda} + \frac{\log(q)}{\lambda} + \frac{\gamma}{\lambda} - \frac{1}{2} \Rightarrow E(R_n) \approx \log_{1/p}(n) + \log_{1/p}(q) + \gamma/\lambda - 1/2 \tag{3.17}$$

και

$$\text{var}(R_n) \approx \frac{\pi^2}{6\lambda^2} + \frac{1}{12} \tag{3.18}$$

όπου γ = -Γ'(1) = 0.5772… η σταθερά Euler-Mascheroni. Η αφαίρεση από τη μέση τιμή του ½ και η πρόσθεση στη διασπορά 1/12 είναι η διόρθωση συνέχειας του Sheppard, και γίνεται διότι όταν μετατρέπουμε μια συνεχή τ.μ. σε διακριτή αυξάνεται η μέση τιμή της και μειώνεται η διασπορά.



Εικόνα 3.4: *Η γραφική παράσταση της κατανομής του Gumbel*

Παράδειγμα 3.4.1

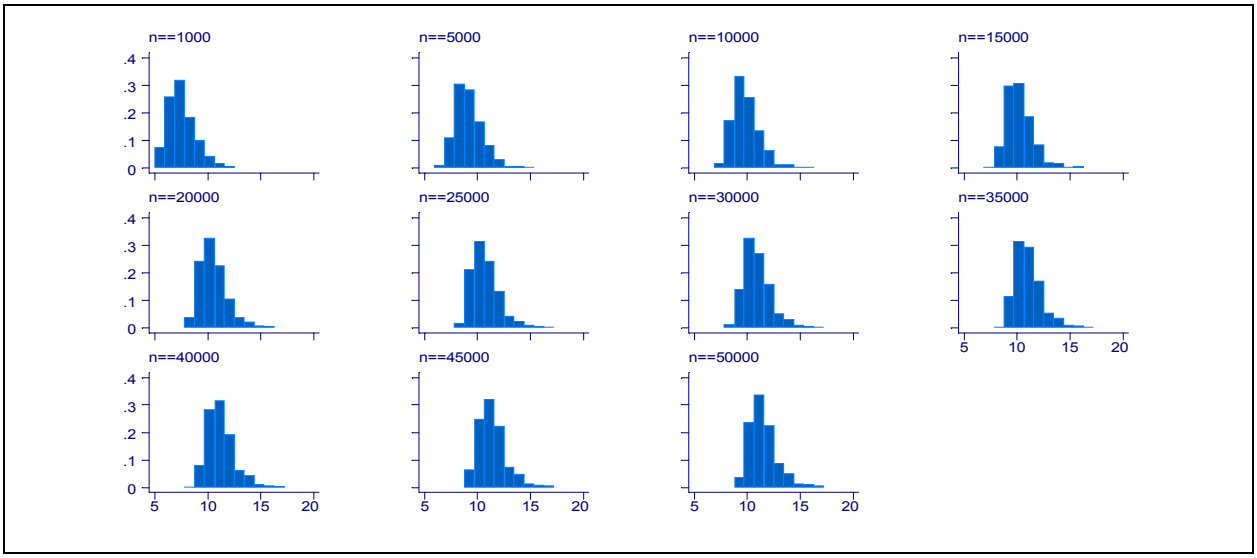
Αν χρησιμοποιήσουμε τις σχέσεις (3.17) και (3.18) στα δεδομένα του παραδείγματος 3.2.1 έχουμε:

$$E(R_n) \approx \log_{1/p}(n) + \log_{1/p}(q) + \gamma/\lambda - 1/2 \Rightarrow$$

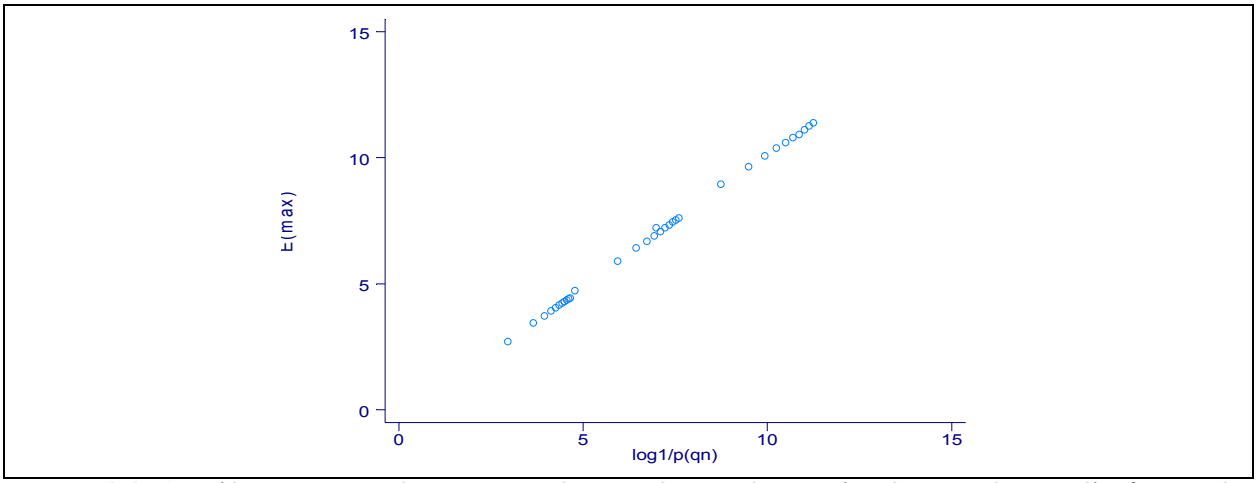
$$E(R_n) \approx \log_4(10000) + \log_4(3/4) + \frac{0.5772}{\log(4)} - 1/2 = 6.3518$$

και $\text{var}(R_n) \approx \frac{\pi^2}{6\lambda^2} + \frac{1}{12} = 0.939$

Παρατηρούμε, ότι παρόλο που η σχέση (3.17), είναι καλύτερη προσέγγιση του αναμενόμενου μήκους της μέγιστης ροής, η πραγματική διαφορά που βρίσκουμε σε σχέση με την πιο απλή εκδοχή, όπως αυτή αποτυπώθηκε στη σχέση (3.9), είναι αρκετά μικρή. Αυτό συμβαίνει, γιατί ο κύριος παράγοντας που καθορίζει την τελική τιμή εξακολουθεί να είναι η τιμή του $\log(n)$, καθώς το $\log(q)$ και το γ/λ είναι σχετικά μικρές ποσότητες. Προφανώς, όσο το n μεγαλώνει, η διαφορά θα γίνεται ακόμα μικρότερη. Από τις προσομοιώσεις, βλέπουμε ξεκάθαρα ότι το μήκος R_n των ροών ακολουθεί την κατανομή του Gumbel (EVD) με μέση τιμή και διασπορά που δίνονται από τις σχέσεις (3.17) και (3.18).



Εικόνα 3.5: *Αποτελέσματα προσομοίωσης για την κατανομή της μέγιστης ροής ευνοϊκών αποτελεσμάτων, σε ένα μοντέλο διωνυμικής κατανομής με πιθανότητες p=0.1, 0.25 και 0.4. Το n κυμαίνεται από 1,000 μέχρι 50,000.*



Εικόνα 3.6: *Αποτελέσματα προσομοίωσης για τη μέγιστη ροή ευνοϊκών αποτελεσμάτων, σε ένα μοντέλο διωνυμικής κατανομής με πιθανότητες p=0.1, 0.25 και 0.4. Το n κυμαίνεται από 1000 μέχρι 50000.*

3.5. Η Κατανομή του Μέγιστου Τμηματικού Σκορ (Maximal Segment Score)

Στη γενικότερη περίπτωση που ενδιαφερόμαστε για την κατανομή που ακολουθεί η τυχαία μεταβλητή του πλήθους R_n^a διαδοχικών δοκιμών που περιέχουν 100α% ευνοϊκά αποτελέσματα, είναι αναγκαίο να ορίσουμε ένα είδος αθροιστικού σκορ (score), που να το περιγράφει. Η προσέγγιση αυτή, έχει όπως θα δούμε πολλά πλεονεκτήματα, καθώς είναι πολύ γενική αλλά περιλαμβάνει και τη ροή ευνοϊκών αποτελεσμάτων σαν ειδική περίπτωση.

Σύμφωνα με τη μέθοδο αυτή, κατά την οποία ενδιαφερόμαστε για την εύρεση μιας περιοχής π.χ. πλούσιας κατά 80% σε A (Karlin & Altschul, 1990; Karlin & Brendel, 1992), πρέπει να ορίσουμε κάποιου είδους σκορ. Τότε, η τυχαία μεταβλητή του πλήθους R_n^a διαδοχικών δοκιμών που περιέχουν 100α% ευνοϊκά αποτελέσματα, μπορεί να περιγράφει με ένα προσθετικό σκορ της μορφής:

$$s_k = \log\left(a_k / p_k\right) \tag{3.19}$$

όπου p_k είναι η πιθανότητα εμφάνισης μιας βάσης σε ολόκληρη την ακολουθία (π.χ. p=1/4) και a_k η πραγματική πιθανότητα εμφάνισης μιας βάσης (target frequency) στο συγκεκριμένο τμήμα της αλληλουχίας

(δηλαδή, σε ένα παράθυρο), το οποίο θέλουμε να ανιχνεύσουμε. Τα p, α είναι τα ίδια τα οποία συναντήσαμε στην θεωρία μεγάλων αποκλίσεων. Αθροίζοντας τα σκορ για τα i κατάλοιπα ενός «παραθύρου» παίρνουμε το τμηματικό σκορ (segment score) και αν το παράθυρο αυτό είναι η μέγιστη περιοχή που περιέχει κατ' ελάχιστο 100α% ευνοϊκά αποτελέσματα, τότε το σκορ ονομάζεται μέγιστο τμηματικό σκορ (maximal segment score) και για μια αλληλουχία μήκους n θα συμβολίζεται ως $M(n)$.

Στον υπολογισμό εργαζόμαστε ως εξής (έστω α=0.8, p=0.25): Από τη σχέση (3.19) έχουμε ότι για κάθε εμφάνιση Α, έχουμε συνεισφορά στο σκορ s_A=log(0.8/0.25)=1.163 και για κάθε εμφάνιση άλλης βάσης θα έχουμε s_N=log(0.2/0.25)=-0.223. Έτσι αν σε ένα τμήμα 20 βάσεων της αλληλουχίας έχουμε 10 Α, τότε το σκορ θα είναι s=16*1.163-4*0.223= 17.716.

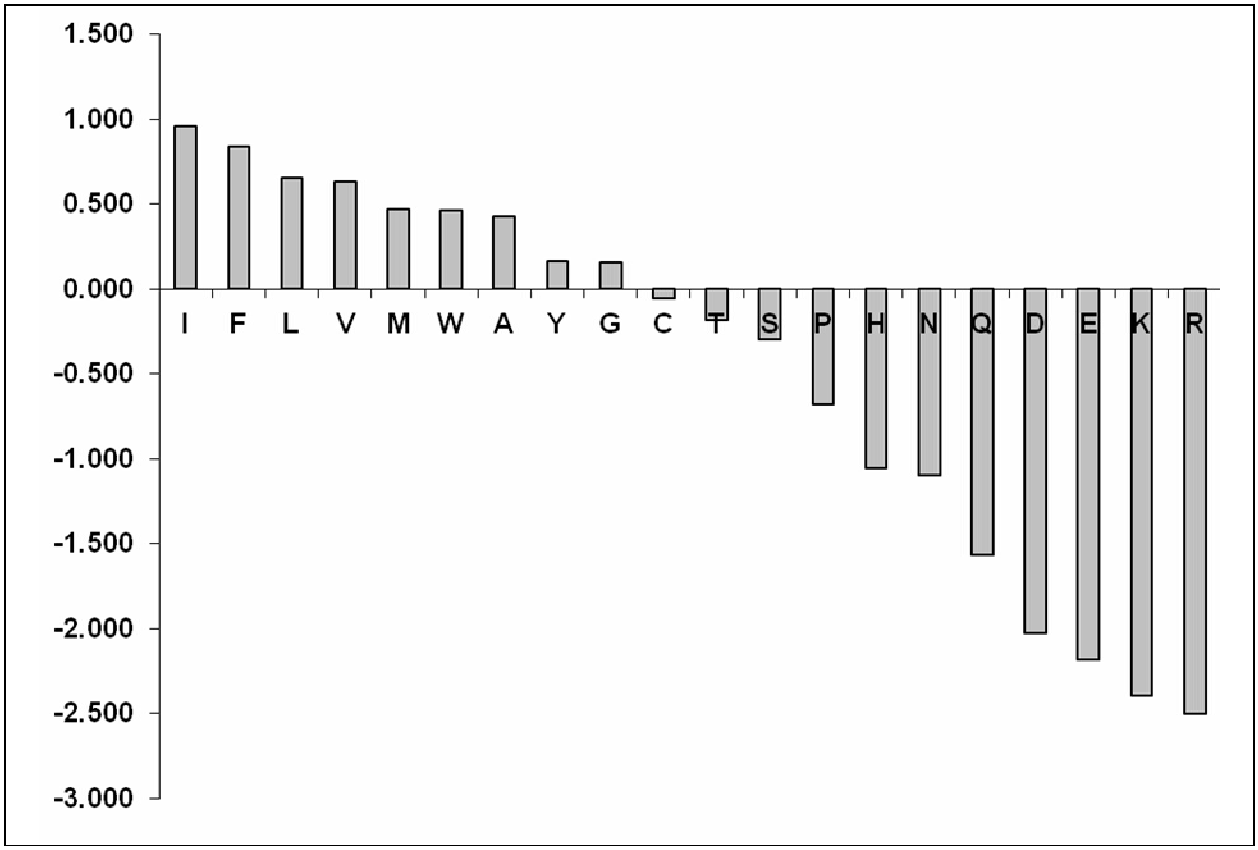
Αμινοξύ (k)	**Πιθανότητα εμφάνισης σε διαμεμβρανικές περιοχές (a_k)**	**Πιθανότητα εμφάνισης σε μη διαμεμβρανικές περιοχές (p_k)**	**Σκορ ($\log(a_k/p_k)$)**
A	0.109	0.071	0.429
C	0.019	0.020	-0.051
D	0.007	0.053	-2.024
E	0.007	0.062	-2.181
F	0.090	0.039	0.836
G	0.082	0.070	0.158
H	0.008	0.023	-1.056
I	0.120	0.046	0.959
K	0.005	0.055	-2.398
L	0.168	0.087	0.658
M	0.040	0.025	0.470
N	0.016	0.048	-1.099
P	0.028	0.055	-0.675
Q	0.009	0.043	-1.564
R	0.005	0.061	-2.501
S	0.053	0.071	-0.292
T	0.050	0.060	-0.182
V	0.115	0.061	0.634
W	0.027	0.017	0.463
Y	0.040	0.034	0.163

Πίνακας 3.1: *Στον πίνακα αυτόν, έχουμε τα στατιστικά στοιχεία από μια ανάλυση 160 διαμεμβρανικών πρωτεϊνών με γνωστή διαμεμβρανική τοπολογία (Krogh, Larsson, von Heijne, & Sonnhammer, 2001). Αν κάποιο αμινοξύ είχε a_k=0 θα έπρεπε να είχαμε κάνει μια μικρή διόρθωση προσθέτοντας μια πολύ μικρή τιμή (πχ 0.0001). Με αυτόν τον τρόπο το σκορ θα έπαιρνε μια πολύ μικρή αρνητική τιμή (πχ -9 ή μικρότερο). Παρατηρήστε ότι η πιθανότητα εμφάνισης των αμινοξέων στις μη-μεμβρανικές περιοχές, είναι πολύ κοντά στη συνολική πιθανότητα εμφάνισης αμινοξέων στη βάση Uniprot (http://web.expasy.org/docs/relnotes/relstat.html).*

Ειδικά στην περίπτωση της ροής R_n^a μιας βάσης, ορίζουμε σκορ =1 (ή κάποιον άλλο θετικό αριθμό) για κάθε εμφάνιση της βάσης αυτής, και -∞ (ή κάποιο άλλο κατ' απόλυτη τιμή πολύ μεγάλο αρνητικό αριθμό) για κάθε άλλη βάση που θα εμφανιστεί. Έτσι μια ροή από π.χ. k=10 βάσεις θα δίνει σκορ=10 ενώ κάθε άλλη περίπτωση θα αποκλείεται (σκορ =-∞). Προφανώς, η προσέγγιση αυτή είναι ισοδύναμη με την πρώτη για την ειδική περίπτωση που α=1, καθώς αν θέλουμε να ανιχνεύσουμε την ροή από Α, θα πρέπει να δίνουμε σκορ για Α, s_A=log(1/0.25)=1.386, και για κάθε άλλη βάση s_N=log(0/0.25)= -∞ (πρακτικά, το κάνουμε θέτοντας την αντίστοιχη τιμή ίση με κάποιον πολύ μικρό αρνητικό αριθμό, πχ -10.000). Σε αυτή την περίπτωση το σκορ για μια καθαρή ροή από 16 Α, θα είναι 16*1.386=22.176.

Η μέθοδος αυτή, είναι όμως πολύ πιο γενική. Στον Πίνακα 3.1 βλέπουμε τα στατιστικά από μια ανάλυση 160 διαμεμβρανικών πρωτεϊνών με γνωστή διαμεμβρανική τοπολογία (Krogh, et al., 2001). Στις πρωτεΐνες αυτές, έχουμε μελετήσει την αμινοξική σύσταση στα διαμεμβρανικά τμήματα αλλά και στις υπόλοιπες (μη μεμβρανικές) περιοχές. Με τον τρόπο αυτόν, μπορούμε με τη χρήση της σχέσης (3.19) να κατασκευάσουμε ένα σκορ που θα μπορούμε να το χρησιμοποιήσουμε για τον εντοπισμό περιοχών με μεγάλη πιθανότητα να είναι διαμεμβρανικά τμήματα. Μεγάλες τιμές του σκορ (όπως για παράδειγμα αυτές που έχουν τα αμινοξέα I, F, L, V, M και W), αντιστοιχούν σε αμινοξέα που έχουν μεγαλύτερη πιθανότητα να εμφανιστούν σε μια διαμεμβρανική περιοχή παρά σε μια μη-μεμβρανική (τα οποία είναι κατά βάση τα

υδρόφοβα αμινοξέα). Αντίθετα, τα πολικά αμινοξέα (Q, D, E, K, και R), εμφανίζουν αρνητικές τιμές στο σκορ. Παρόμοια σκορ, είναι δυνατόν να οριστούν με πολλούς άλλους διαφορετικούς τρόπους. Για την ακρίβεια, τέτοιες προσεγγίσεις αποτελούν τη βάση των προγνωστικών αλγορίθμων, τους οποίους θα συναντήσουμε σε επόμενο κεφάλαιο.



Εικόνα 3. 7: *Γραφική παράσταση με τα σκορ των 20 αμινοξέων διαετεταγμένα σε φθίνουσα σειρά μεγέθους. Τα υδρόφοβα αμινοξέα έχουν θετικές τιμές ενώ τα πολικά και τα φορτισμένα, αρνητικές.*

Για τον υπολογισμό του μέγιστου τμηματικού (τοπικού) σκορ πρέπει να θέσουμε και κάποιους περιορισμούς. Συγκεκριμένα:

- Τουλάχιστον ένα σκορ πρέπει να είναι θετικό
- Η αναμενόμενη τιμή του σκορ για κάθε βάση να είναι αρνητική, δηλαδή

$$E\left(s_k\right) = \sum p_k s_k = \sum p_k \log\left(\frac{a_k}{p_k}\right) < 0 \tag{3.20}$$

Ο πρώτος περιορισμός είναι απαραίτητος για να είμαστε σίγουροι ότι έχουμε τοπικό σκορ και δεν αναφερόμαστε σε ολόκληρη την ακολουθία, ενώ ο δεύτερος ισχύει σχεδόν πάντα καθώς το $E(s_k)$ είναι ίσο με $-H(\alpha,p)$.

Για την κατανομή που ακολουθεί το μέγιστο τμηματικό score M_n στην γενική περίπτωση, είναι γνωστό το επόμενο θεώρημα (Karlin & Altschul, 1990) που λέει ότι η τυχαία μεταβλητή M_n (το μέγιστο τμηματικό score) έχει προσεγγιστική κατανομή την:

$$P\left\{M_n > {\log(n)}/{\lambda} + x\right\} \approx 1 - \exp\left\{-Ke^{-\lambda x}\right\} \tag{3.21}$$

Αυτή είναι η κατανομή των ακραίων τιμών του Gumbel, ενώ K και λ είναι οι σταθερές της και υπολογίζονται με αριθμητικές μεθόδους. Για το λ ειδικά ισχύει το ότι είναι η μοναδική θετική λύση της εξισώσεως:

$$\sum_k p_k \exp\left\{\lambda s_k\right\} = 1 \tag{3.22}$$

Οι ίδιοι συγγραφείς, έδειξαν επίσης ότι καθώς το μήκος n της τυχαίας ακολουθίας τείνει στο άπειρο, η συχνότητα a_k της εμφάνισης κάποιας βάσης σε ένα τμήμα με αρκετά μεγάλο σκορ προσεγγίζει το $p_k \exp\{\lambda s_k\}$ με πιθανότητα 1. Για την ακρίβεια όταν έχουμε το μέγιστο σκορ, τότε:

$$a_k = p_k \exp\{\lambda s_k\} \tag{3.23}$$

Παρατηρούμε επίσης ότι καθώς η ύπαρξη τέτοιων τμημάτων με μεγάλο σκορ (μεγαλύτερο από x) είναι σπάνια γεγονότα (rare events), θα ακολουθούν την κατανομή Poisson με μέση τιμή, E=Knexp(-λx) οπότε η σχέση (3.21) μπορεί να ξαναγραφτεί ως εξής:

$$P(M_n \geq x) \approx 1 - e^{-E} \tag{3.24}$$

Όταν η μέση τιμή-αναμενόμενη τιμή (E-value) είναι πολύ μικρή τότε επειδή ισχύει η προσεγγιστική σχέση:

$$1 - \exp(-\exp(-t)) \approx 1 - (1 - \exp(-t)) = \exp(-t) \tag{3.25}$$

θα έχουμε το P-value περίπου ίσο με το E-value. Επομένως, κάνοντας χρήση της κατανομής Poisson, η πιθανότητα να βρούμε σε μια ακολουθία μήκους n, m τμήματα με σκορ $S_{(m)}$ μεγαλύτερο ή ίσο από το x θα είναι:

$$P(S_{(m)} \geq x) \approx 1 - \exp\left(-Kne^{-\lambda x}\right) \sum_{i=0}^{m-1} \frac{\left(Kne^{-\lambda x}\right)^i}{i!} \tag{3.26}$$

Στην ειδική περίπτωση της ροής R_n, όπως είδαμε παραπάνω, μπορούν να δοθούν κλειστές εκφράσεις για τα K και λ και αυτές είναι :

$$K = 1 - p = q \tag{3.27}$$

και

$$\lambda = \log(1/p) \tag{3.28}$$

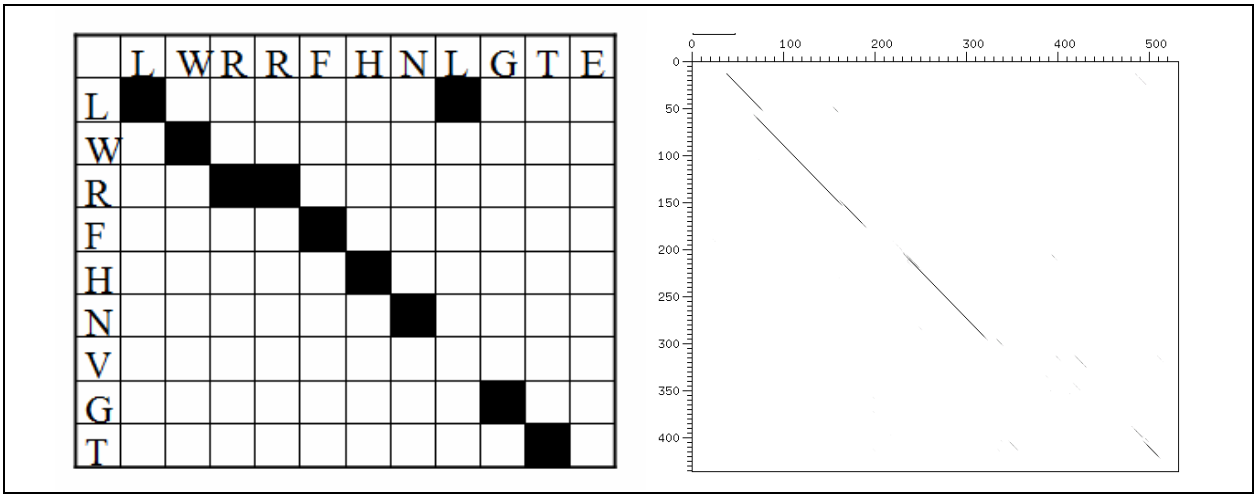
3.6. Στοίχιση αλληλουχιών

Το πρόβλημα της στοίχισης δύο βιολογικών αλληλουχιών, είναι ένα από τα παλιότερα αλλά και πιο σημαντικά θέματα στη βιβλιογραφία της υπολογιστικής βιολογίας. Δύο αλληλουχίες που είναι σε μεγάλο βαθμό «όμοιες», είναι πιθανό να έχουν κοινή εξελικτική προέλευση και, αν μιλάμε για πρωτεΐνες, να έχουν παρόμοια τρισδιάστατη δομή και παρόμοιες λειτουργίες.

Έστω ότι έχουμε δυο βιολογικές αλληλουχίες $\mathbf{x}=x_1,x_2,\ldots,x_n$ και $\mathbf{y}=y_1,y_2,\ldots,y_m$ και θέλουμε να ελέγξουμε κατά πόσο αυτές είναι όμοιες ή όχι. Δημιουργούνται αυτόματα μια σειρά από ερωτήματα:

- Το πρώτο πρόβλημα που προκύπτει είναι με ποιο τρόπο θα μετρήσουμε την ομοιότητα (το πρόβλημα του σκορ)
- Το δεύτερο, αφορά τον τρόπο με τον οποίο θα γίνει η στοίχιση (alignment) των δυο αλληλουχιών (ο αλγόριθμος)
- Το τρίτο, αφορά την επιλογή του είδους της στοίχισης, και τέλος
- Το τελευταίο ερώτημα, αφορά στο πώς θα αποφασίσουμε αν μια δεδομένη στοίχιση είναι σημαντική ή όχι (η στατιστική σημαντικότητα)

Ένας παλιός, αλλά ταυτόχρονα και διαισθητικός τρόπος σύγκρισης δύο αλληλουχιών, είναι το λεγόμενο διάγραμμα σημείων (dot plot). Σύμφωνα με αυτήν την απλοϊκή προσέγγιση, οι δύο αλληλουχίες τοποθετούνται σε ένα δισδιάστατο πίνακα. Σε κάθε κελί του πίνακα, το οποίο αντιστοιχεί σε ένα ζεύγος «συμβόλων» από τις δύο αλληλουχίες (νουκλεοτίδια ή αμινοξέα), βάζουμε μαύρο χρώμα αν τα δύο σύμβολα είναι όμοια, και λευκό, αν είναι ανόμοια. Διαισθητικά, αναμένουμε ότι αν οι δυο αλληλουχίες είναι 100% όμοιες, το σχήμα που θα παρατηρήσουμε θα είναι μια ευθεία γραμμή στη διαγώνιο. Αν οι οι αλληλουχίες δεν έχουν καμία ομοιότητα, θα περιμένουμε μια τυχαία κατανομή των μαύρων (γραμμοσκιασμένων) κελιών. Προφανώς, σε περιπτώσεις μερικής ομοιότητας, θα περιμένουμε να δούμε «κάτι» που να μοιάζει με γραμμή πάνω ή γύρω από τη διαγώνιο. Αν η ομοιότητα εντοπίζεται μόνο σε ένα ορισμένο σημείο, και δεν εκτείνεται σε όλο το μήκος των αλληλουχιών τότε θα περιμένουμε μια διαγώνιο γραμμή να βρίσκεται κάπου μέσα στον πίνακα (και όχι απαραίτητα στην κύρια διαγώνιο).



Εικόνα 3.8: *Δύο παραδείγματα διαγραμμάτων σημείων (dot plot). Στο αριστερό σχήμα, βλέπουμε το διάγραμμα που αντιστοιχεί στη στοίχιση δύο μικρών πρωτεϊνών, στο οποίο μπορούμε να δούμε τα όμοια και ανόμοια αμινοξέα. Οι δύο αλληλουχίες έχουν μεγάλη ομοιότητα, έστω και αν βλέπουμε 1-2 μικροδιαφορές. Στα δεξιά, βλέπουμε τη σύγκριση δύο πραγματικών πρωτεϊνών μεγάλου μήκους. Στην περίπτωση αυτή δεν μπορούμε να δούμε τα αμινοξέα, αλλά η περιοχή ομοιότητας είναι εμφανής (τουλάχιστον μέχρι το κατάλοιπο 300 των δύο αλληλουχιών)*

Στις επόμενες παραγράφους, θα προσπαθήσουμε να αναλύσουμε περισσότερο τα θέματα αυτά, να τα εξειδικεύσουμε και να τα ποσοτικοποιήσουμε. Θα ξεκινήσουμε λοιπόν, με το πρόβλημα της ποσοτικοποίησης της ομοιότητας. Αν θεωρήσουμε ότι οι δυο αλληλουχίες DNA $\mathbf{x}=x_1,x_2,\ldots,x_n$ και $\mathbf{y}=y_1,y_2,\ldots,y_m$ είναι ασυσχέτιστες (τυχαίες), τότε η πιθανότητα να συμπίπτουν σε κάποιο τμήμα τους είναι:

$$P(\mathbf{x},\mathbf{y} \mid R) = \prod_i q_{x_i} \prod_j q_{y_j} \quad \text{με } i = 1,2,...,n \text{ και } j = 1,2,...,m \tag{3.29}$$

Εναλλακτικά, αν θεωρήσουμε ότι οι δυο αλληλουχίες είναι συσχετισμένες τότε η πιθανότητα να συμπίπτουν δίνεται από την από κοινού κατανομή της πιθανότητας:

$$P(\mathbf{x},\mathbf{y} \mid M) = \prod_i p_{x_i y_i} \quad \text{με } i = 1,2,...,n \tag{3.30}$$

όπου για απλότητα (έτσι ώστε να έχει νόημα και η πλήρης ταύτιση-σε όλο το μήκος), μπορούμε να δεχθούμε ότι $n=m$. Ο λόγος των δυο αυτών πιθανοφανειών (likelihood ratio) ονομάζεται και odds ratio και είναι ίσος με:

$$\frac{P(\mathbf{x},\mathbf{y} \mid M)}{P(\mathbf{x},\mathbf{y} \mid R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \prod_j q_{y_j}} = \prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \tag{3.31}$$

Αν πάρουμε τους λογαρίθμους έχουμε:

$$S = \sum_i \log\left(\frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}\right) = \sum_i s(x_i, y_i) \tag{3.32}$$

οπότε ορίζουμε με αυτόν τον τρόπο, το score για την ομοιότητα δυο αλληλουχιών το οποίο πλέον έχει προσθετικές ιδιότητες. Για τα 4 νουκλεοτίδια του DNA μπορούν να φτιαχτούν πίνακες 4x4 που να απεικονίζουν τις παραπάνω συνεισφορές στο score για κάθε μια από τις 16 περιπτώσεις ταύτισης βάσεων σε μια στοίχιση αλλά πολλές φορές μπορούμε απλώς να θέσουμε:

$$s(x_i, y_i) = \begin{cases} 1, & \text{αν } x_i = y_i \\ -1, & \text{αν } x_i \neq y_i \end{cases} \tag{3.33}$$

δηλαδή ορίζουμε συνεισφορά στο score 1 για ταύτιση (match) και –1 για διαφορά (mismatch). Ο πίνακας αυτός ονομάζεται πίνακας ταύτισης (match matrix) ενώ αν έχουμε ποινή για την εμφάνιση διαφοράς $\rightarrow$-∞ (π.χ. –10000) αποκλείεται να εμφανιστεί μια στοίχιση που να περιέχει τίποτα άλλο εκτός από απολύτως όμοια νουκλεοτίδια, και ο πίνακας αυτός ονομάζεται ταυτοτικός πίνακας (identity matrix). Οι ονομασίες για τους παραπάνω πίνακες (π.χ. ταυτοτικός) δεν πρέπει να συγχέονται με την αντίστοιχη ονομασία των πινάκων στη

γραμμική άλγεβρα. Η ερμηνεία του ταυτοτικού πίνακα είναι ότι με μια τόσο μεγάλη ποινή για εμφάνιση διαφοράς (→-∞) αποκλείεται να εμφανιστεί μια στοίχιση που να περιέχει τίποτα άλλο εκτός από απολύτως όμοια νουκλεοτίδια.

Αυτό το απλοϊκό σχήμα του περιγράψαμε, θυμίζει αρκετά τη μέθοδο των διαγραμμάτων σημείων. Ειδικά για τις πρωτεΐνες υπάρχουν πάρα πολλοί τέτοιοι πίνακες υποκατάστασης (substitution matrices) οι οποίοι υπολογίζουν τις αντίστοιχες συνεισφορές στο score για μη ταύτιση των διάφορων αμινοξέων (mismatches) στηριζόμενοι σε παρατηρηθείσες αντικαταστάσεις αμινοξέων αλλά με διαφορετικό τρόπο ο καθένας. Οι πίνακες υποκατάστασης όπως έδειξε ο Altschul (Altschul, 1991), έχουν ξεκάθαρη ερμηνεία υπό το πρίσμα της θεωρίας της πληροφορίας. Τέτοιοι πίνακες είναι οι πίνακες των οικογενειών PAM (Dayhoff, Schwartz, & Orcutt, 1978), BLOSUM (Henikoff & Henikoff, 1992), GONNET (Gonnet, Cohen, & Benner, 1992) οι οποίοι επιπλέον, έχουν μια ξεκάθαρη εξελικτική ερμηνεία όπως θα δούμε παρακάτω.

Τέλος, είναι απαραίτητο να προβλέψουμε και την ύπαρξη κενών στις στοιχισμένες αλληλουχίες. Η ύπαρξη αυτή είναι απαραίτητη καθώς θα δούμε παρακάτω, ότι ένα από τα βασικά χαρακτηριστικά των μεταλλάξεων μέσω των οποίων προχωράει η εξέλιξη είναι η προσθήκη (insertion) και η απαλοιφή (deletion) νουκλεοτιδίων. Όταν στη στοίχιση δυο αλληλουχιών εμφανίζεται (στη μια από τις δυο), το κενό (gap) δεν είναι δυνατό να ξέρουμε αν αυτό προήλθε (εξελικτικά) από απαλοιφή βάσης σ' αυτή την αλληλουχία, ή από προσθήκη στην άλλη αλληλουχία με την οποία συγκρίνεται. Προφανώς η ύπαρξη του κενού πρέπει να «τιμωρείται» από το score γιατί αλλιώς δυο οποιεσδήποτε αλληλουχίες με την προσθήκη «κατάλληλου» αριθμού κενών θα δίνουν μια άριστη στοίχιση. Η συνεισφορά των κενών (η οποια πρέπει να είναι αρνητική) στο score ορίζεται από μια συνάρτηση $\gamma(g)$, όπου g είναι ο αριθμός των κενών, και μπορεί να είναι είτε γραμμική:

$$\gamma(g) = -gd \tag{3.34}$$

είτε πιο σύνθετη:

$$\gamma(g) = -d - (g-1)e \tag{3.35}$$

όπου d είναι η ποινή για την ύπαρξη κενού (gap open penalty) και e η ποινή για την διεύρυνση του κενού (gap extension penalty). Η σωστή επιλογή της συνάρτησης για τα κενά είναι δύσκολη διαδικασία και υπάρχει πλούσια βιβλιογραφία για αυτό το θέμα (Vingron & Waterman, 1994).

3.7. Πίνακες ομοιότητας

Σε περιπτώσεις στοίχισης αλληλουχιών DNA, χρησιμοποιούνται συνήθως απλοί πίνακες ομοιότητας της μορφής της σχέσης (3.33). Στις περισσότερες περιπτώσεις, αν δεν θέλουμε να επιτρέψουμε πολλές ταυτίσεις ανόμοιων βάσεων χρησιμοποιούμε την τιμή 1 για τη ταύτιση και -3 για τη διαφορά, ενώ στις πιο συνηθισμένες περιπτώσεις, χρησιμοποιούμε μια τιμή 5 για την ταύτιση και -4 για τη διαφορά.

Από την άλλη πλευρά, ειδικά για τις πρωτεΐνες υπάρχουν πάρα πολλοί εξειδικευμένοι πίνακες υποκατάστασης (substitution matrices) οι οποίοι υπολογίζουν τις αντίστοιχες συνεισφορές στο score για μη ταύτιση των διάφορων αμινοξέων (mismatches) στηριζόμενοι σε παρατηρηθείσες αντικαταστάσεις αμινοξέων αλλά με διαφορετικό τρόπο ο καθένας. Οι πίνακες υποκατάστασης όπως έδειξε ο Altschul (Altschul, 1991), έχουν ξεκάθαρη ερμηνεία υπό το πρίσμα της θεωρίας της πληροφορίας. Τέτοιοι πίνακες είναι οι πίνακες των οικογενειών PAM (Dayhoff, et al., 1978), BLOSUM (Henikoff & Henikoff, 1992), GONNET (Gonnet, et al., 1992) αλλά και άλλοι.

Οι πίνακες BLOSUM (Henikoff & Henikoff, 1992), έχουν υπολογισθεί από τμήματα από πολλαπλές στοιχίσεις αλληλουχιών για τις οποίες υπάρχουν ξεκάθαρες ενδείξεις ότι έχουν φυλογενετική σχέση. Τα τμήματα αυτά (blocks), επιλέχθηκαν με προσοχή από ένα μεγάλο εύρος πρωτεϊνικών οικογενειών και διατηρήθηκαν τελικά μόνο τα πιο καλά στοιχισμένα τμήματα (αυτά που δεν περιείχαν κενά). Για τον υπολογισμό χρησιμοποιήθηκε ο τύπος:

$$s_{ij} = \frac{1}{\lambda}\log\left(\frac{q_{ij}}{p_i p_j}\right) \tag{3.36}$$

όπου q_{ij}, είναι η πιθανότητα αντικατάστασης του i από το j σε σχετιζόμενες πρωτεΐνες (target frequencies), p_i, p_j είναι οι πιθανότητες εμφάνισης των αμινοξέων σε οποιαδήποτε θέση (background frequencies) και λ είναι μια σταθερά κανονικοποίησης έτσι ώστε οι τιμές να μετατραπούν σε ακέραιους. Η ομοιότητα με τη σχέση

(3.32) είναι εμφανής. Οι πίνακες αυτοί δεν προϋποθέτουν ένα εξελικτικό μοντέλο αλλά το προσεγγίζουν εμπειρικά.

Η άλλη μεγάλη οικογένεια πινάκων αντικατάστασης, είναι οι πίνακες της οικογένειας Point Accepted Mutations (PAM) (Dayhoff, et al., 1978). Οι συγγραφείς, όρισαν ως «Αποδεκτή Σημειακή Μεταλλαγή» (PAM) σε μια πρωτεΐνη την αντικατάσταση ενός αμινοξικού κατάλοιπού της με ένα κατάλοιπο διαφορετικού τύπου, η οποία έχει γίνει αποδεκτή μέσω της διαδικασίας της Φυσικής Επιλογής. Η τιμή PAM1 προέκυψε από πολλαπλή στοίχιση αλληλουχιών με γνωστή εξελικτική σχέση και ομοιότητα μεγαλύτερης του 85%. Μέσω αυτής της τιμής και με χρήση ενός μαρκοβιανού μοντέλου εξέλιξης (θα παρουσιαστούν στο κεφάλαιο της φυλογενετικής ανάλυσης), προέκυψαν οι πίνακες PAM30, PAM250 κ.ο.κ. δεδομένου ότι οι πίνακες αυτοί είναι πολλαπλασιαστικοί καθώς ισχύει PAMN=(PAM1)^N. Η χρήση πινάκων με μικρό N ενδείκνυται όταν οι εξεταζόμενες αλληλουχίες είναι πολύ όμοιες (μικρή εξελικτική απόσταση), ενώ στην περίπτωση περισσότερο απομακρυσμένων ομοιοτήτων χρησιμοποιούμε πίνακες μεγαλύτερου N. Στις περιπτώσεις εκείνες κατά τις οποίες δεν γνωρίζουμε εκ των προτέρων την ομοιότητα των προς σύγκριση αλληλουχιών (π.χ. σε αναζητήσεις έναντι βάσεων δεδομένων) επιλέγουμε έναν ενδιάμεσο πίνακα, όπως τον PAM250, ο οποίος αντιστοιχεί σε συντήρηση της τάξης του 20-25%.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1

Εικόνα 3.9: *Ο Πίνακας BLOSUM62. Παρατηρήστε ότι αμινοξέα τα οποία έχουν παρόμοιες φυσικοχημικές ιδιότητες (π.χ. υδρόφοβα, πολικά, αρωματικά κ.ο.κ.), έχουν γενικά θετικές τιμές για τις μεταξύ τους αντικαταστάσεις*

Παρόλο που οι δύο οικογένειες πινάκων έχουν διαφορές, μπορούμε σε γενικές γραμμές να κάνουμε μια «αντιστοίχιση». Γενικά, μικρές τιμές των πινάκων PAM, και μεγάλες τιμές των πινάκων BLOSUM αντιστοιχούν σε, και κατά συνέπεια πρέπει να χρησιμοποιούνται για, αλληλουχίες με μικρή εξελικτική απόσταση, δηλαδή με μεγάλες ομοιότητες. Αντίθετα, μεγάλες τιμές των πινάκων PAM, και μικρές τιμές των πινάκων BLOSUM αντιστοιχούν σε, και κατά συνέπεια πρέπει να χρησιμοποιούνται για, αλληλουχίες με μεγάλη εξελικτική απόσταση, δηλαδή με μικρότερες ομοιότητες (Πίνακας 3.2).

PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM60
PAM200	BLOSUM52
PAM250	BLOSUM45

Πίνακας 3.2 *Πίνακας κατά προσέγγιση αντιστοίχιση των πινάκων της οικογένειας PAM με αυτούς της οικογένειας BLOSUM*

Τα παραπάνω προφανώς, ισχύουν σε ειδικές περιπτώσεις, όταν ξέρουμε εκ των προτέρων πόσο αναμένουμε να μοιάζουν δύο υπό σύγκριση αλληλουχίες. Στη γενική περίπτωση όμως που πραγματοποιούμε αναζήτηση σε μια βάση δεδομένων, τότε η πιο συνετή επιλογή είναι να χρησιμοποιήσουμε έναν πίνακα

«ενδιάμεσης» ομοιότητας, όπως τον BLOSUM62. Αν το αποτέλεσμα της αναζήτησης μας δώσει πάρα πολλές πρωτεΐνες τις οποίες δυσκολευόμαστε να διαχωρίσουμε (έχουν όλες μεγάλη ομοιότητα), τότε μπορούμε να επαναλάβουμε την αναζήτηση με έναν πίνακα όπως ο BLOSUM90. Αν, από την άλλη μεριά, η αρχική αναζήτηση μας δώσει λίγα αποτελέσματα, τότε θα πρέπει να ξανακάνουμε αναζήτηση επιλέγοντας για παράδειγμα τον BLOSUM45. Αυτό που πρέπει σε κάθε περίπτωση να θυμάται ο αναγνώστης, είναι ότι αλλαγή στον πίνακα υποκατάστασης, σημαίνει και αλλαγή (μικρή ή μεγάλη) στα αποτελέσματα της αναζήτησης αλλά και της προκύπτουσας στοίχισης.

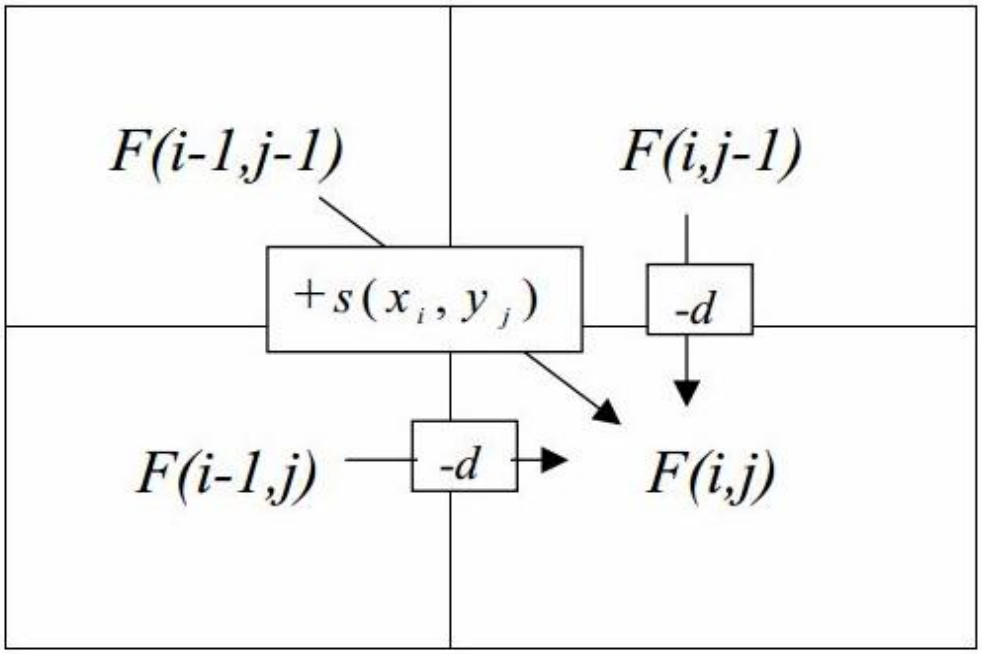
Διαισθητικά, οι πίνακες αυτοί, έχουν την εξής ερμηνεία: αμινοξέα τα οποία έχουν παρόμοιες φυσικοχημικές ιδιότητες (πχ υδρόφοβα, πολικά, αρωματικά κ.ο.κ.), έχουν θετικές τιμές για τις μεταξύ τους αντικαταστάσεις. Αυτό σημαίνει ότι σε γενικές γραμμές, μια αντικατάσταση ενός αμινοξέος με ένα άλλο παρόμοιο, θα είναι «αποδεκτή» διαδικασία για τη δομή και τη λειτουργία της πρωτεΐνης. Αυτό με τη σειρά του, σημαίνει ότι είναι δυνατόν δύο πρωτεΐνες στις οποίες μεγάλο μέρος των αμινοξέων έχουν αντικατασταθεί με «παρόμοια» (και κατά συνέπεια, δεν εμφανίζουν μεγάλη ονομαστική ταύτιση), Παρ' όλα αυτά να θεωρούνται «όμοιες» και να λαμβάνουν μεγάλο score στις στοιχίσεις. Φυσικά, αναμένουμε ότι για κάθε αμινοξύ, τη μεγαλύτερη τιμή για αντικατάσταση θα την έχει ο εαυτός του (οι τιμές στη διαγώνιο) αλλά δεν αναμένουμε όλες οι τιμές της διαγωνίου να είναι ίδιες γιατί οι τιμές αυτές εξαρτώνται και από την πιθανότητα εμφάνισης του κάθε αμινοξέος. Για παράδειγμα στον BLOSUM62, η Κυστεΐνη (C) και η Τρυπτοφάνη (W), οι οποίες είναι τα πιο σπάνια αμινοξέα, έχουν και τις μεγαλύτερες τιμές στη διαγώνιο (9 και 11, αντίστοιχα), ενώ η Αλανίνη (A), η οποία είναι ένα από τα πιο συνηθισμένα, έχει τη μικρότερη τιμή (μόλις 4). Τέλος, πρέπει να τονίσουμε, ότι οι πίνακες που περιγράψαμε είναι φτιαγμένοι για γενική χρήση. Για πιο ειδικά προβλήματα, είναι δυνατόν να κατασκευαστούν ειδικοί πίνακες, όπως για παράδειγμα στην περίπτωση της αναζήτησης για διαμεμβρανικές πρωτεΐνες, ο πίνακας PHAT (Ng, Henikoff, & Henikoff, 2000) και ο SLIM (Muller, Rahmann, & Rehmsmeier, 2001). Ο τελευταίος μάλιστα, είναι και μη-συμμετρικός, ιδιότητα που του επιτρέπει να υπολογίζει καλύτερα την ασυμμετρία που υπάρχει στις κατανομές των αμινοξέων στις μεμβρανικές πρωτεΐνες (σε σχέση με τις γενικές πιθανότητες «υποβάθρου»).

3.8. Αλγόριθμοι δυναμικού προγραμματισμού

Αφού ορίσαμε τον τρόπο ποσοτικοποίησης της ομοιότητας των αλληλουχιών, το επόμενο βήμα είναι να βρούμε τον καλύτερο τρόπο στοίχισης. Για άλλη μια φορά, η αναφορά στο διάγραμμα σημείων είναι χρήσιμη, καθώς είπαμε πριν ότι σε περιπτώσεις μερικής ομοιότητας, θα περιμένουμε να δούμε «κάτι» που να μοιάζει με γραμμή πάνω στη διαγώνιο, ή γύρω από αυτή. Ο σκοπός μιας στοίχισης, είναι να εντοπίσει τη βέλτιστη διαδρομή πάνω σε έναν τέτοιον πίνακα. Όταν η διαδρομή είναι γνωστή, η παράθεση των ζευγών των συμβόλων που αντιστοιχούν στα κελιά του πίνακα με την «καλύτερη» διαδρομή, αντιστοιχεί στην τελική στοίχιση. Προφανώς, «καλύτερη διαδρομή» σημαίνει η διαδρομή η οποία μεγιστοποιεί το σκορ όπως το ορίσαμε λίγο πριν. Οι πιθανές στοιχίσεις όμως, δηλαδή οι πιθανές διαδρομές στον πίνακα, είναι πάρα πολλές. Οι πιθανοί τρόποι παράθεσης των δυο αλληλουχιών η μια κάτω από την άλλη, αν υποθέσουμε ότι μπορεί να υπάρχουν και οσαδήποτε κενά, είναι $\binom{n+m}{n}$ και στην ειδική περίπτωση που n=m, έχουμε από τον τύπο του Stirling (Durbin, et al., 1998):

$$\binom{2n}{n} = \frac{(2n!)}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{2\pi n}} \tag{3.37}$$

Προφανώς ο αριθμός αυτός είναι πολύ μεγάλος και δεν υπάρχει τρόπος να υπολογισθούν όλα τα σκορ που αντιστοιχούν σε αυτούς τους συνδυασμούς. Αντίθετα χρησιμοποιούνται αλγόριθμοι δυναμικού προγραμματισμού που με διαδοχικά βήματα βρίσκουν τον καλύτερο τρόπο στοίχισης. Ο δυναμικός προγραμματισμός, είναι μια τεχνική που βρίσκει εφαρμογές σε πολλά δύσκολα προβλήματα στη βιοπληροφορική. Το βασικό χαρακτηριστικό των αλγορίθμων αυτών, είναι ότι «σπάνε» το μεγάλο πρόβλημα (το οποίο απαιτεί πολλούς υπολογισμούς για να λυθεί), σε μικρότερα προβλήματα τα οποία λύνονται πιο εύκολα. Το βασικό σημείο, είναι κάθε φορά, μια επαγωγική απόδειξη η οποία θα δείχνει ότι το άθροισμα των μικρότερων αυτών προβλημάτων, δίνει και τη λύση του μεγάλου προβλήματος.



Εικόνα 3.10: *Οι αλγόριθμοι δυναμικού προγραμματισμού, υπολογίζουν κάθε φορά το στοιχείο F(i,j) από τα 3 γειτονικά κελιά του F(i-1,j) F(i,j-1) F(i-1,j-1).*

Οι αλγόριθμοι δυναμικού προγραμματισμού στη στοίχιση αλληλουχιών (Gonnet, et al., 1992) εργάζονται σε γενικές γραμμές ως εξής: τοποθετούν τις δυο αλληλουχίες $\mathbf{x}=x_1,x_2,\ldots,x_n$ και $\mathbf{y}=y_1,y_2,\ldots,y_m$ σε ένα nm πίνακα με στοιχεία $F(i,j)$ όπου κάθε στοιχείο αυτού του πίνακα είναι η τιμή του σκορ για την καλύτερη στοίχιση μέχρι το στοιχείο x_i, και το y_j. Στην ουσία, δουλεύουν πάνω στον πίνακα του διαγράμματος σημείων που είδαμε πριν, τοποθετώντας αριθμητικές τιμές στα κελιά του.

Προφανώς, αν γνωρίζουμε την τιμή της συνεισφοράς στο σκορ $s(x_i,y_i)$ για κάθε δυνατό συνδυασμό βάσεων και τη συνάρτηση της ποινής για το κενό, τότε με γνωστά τα στοιχεία $F(i\text{-}1,j)$, $F(i,j\text{-}1)$ και $F(i\text{-}1,j\text{-}1)$ μπορούμε να υπολογίσουμε αναδρομικά το $F(i,j)$ όπως φαίνεται στην Εικόνα 3.10 όπου απεικονίζονται οι τρεις οι πιθανοί τρόποι μετάβασης από το $F(i\text{-}1,j\text{-}1)$ στο $F(i,j)$. Προφανώς κάθε μη διαγώνια μετάβαση σημαίνει την εισαγωγή του κενού σε μια από τις δυο αλληλουχίες. Έχοντας υπολογίσει όλα τα στοιχεία αυτού του πίνακα μπορούμε κινούμενοι προς τα πίσω να βρούμε την καλύτερη δυνατή στοίχιση των δυο αλληλουχιών.

3.9. Ολική στοίχιση - Ο αλγόριθμος των Needleman και Wunsch

Η πρώτη περίπτωση η οποία θα εξετάσουμε είναι η λεγόμενη ολική στοίχιση δυο αλληλουχιών (*global alignment*). Στην περίπτωση αυτή έχουμε δυο αλληλουχίες περίπου ίδιου μήκους και θέλουμε να δούμε ποιος είναι ο καλύτερος δυνατός τρόπος να στοιχηθούν παράλληλα η μια κάτω από την άλλη σε όλο το μήκος τους (π.χ. μπορεί να είναι δυο γονίδια για την ίδια πρωτεΐνη από διαφορετικούς οργανισμούς) ώστε να μπορέσουμε να εξετάσουμε την πιθανή εξελικτική ή λειτουργική σχέση τους.

Ο αλγόριθμος που επιτυγχάνει τα παραπάνω είναι ο αλγόριθμος των *Needleman-Wunsch* (Needleman & Wunsch, 1970). Σύμφωνα με τον αλγόριθμο αυτό το σκορ κάθε κελιού υπολογίζεται με τον αναδρομικό τύπο:

$$F(i,j)=\max\left\{F(i-1,j-1)+s(x_i,y_j),F(i-1,j)-d,F(i,j-1)-d\right\} \tag{3.38}$$

Η τιμή για το κάτω δεξιά στοιχείο του πίνακα είναι εξ' ορισμού το σκορ για την καλύτερη δυνατή στοίχιση, ενώ για την αρχικοποίηση της πρώτης στήλης και της πρώτης γραμμής, έχουμε $F(i,0)=-id$ και $F(0,j)=-jd$. Από το κάτω δεξιά στοιχείο, θα πρέπει να ξεκινήσει μια αναδρομή (recursion) στον πίνακα, η οποία ακολουθώντας κάθε φορά τα μέγιστα θα αποκαλύψει τη βέλτιστη διαδρομή, δηλαδή τη βέλτιστη στοίχιση.

Παράδειγμα 3.9.1 (Waterman, 1995)
Έστω ότι έχουμε τις εξής δυο αλληλουχίες DNA **y**=CAGTATCGCA και **x**=AAGTTAGCAG. Θέλουμε να δούμε ποια είναι η καλύτερη ολική στοίχιση που μπορούν να έχουν με:

$$s\left(x_i, y_i\right) = \begin{cases} 1, & \text{αν } x_i = y_i \\ -1, & \text{αν } x_i \neq y_i \end{cases} \tag{3.39}$$

και d=1, τότε συμπληρώνοντας τον πίνακα έχουμε:

	-	*A*	*A*	*G*	*T*	*T*	*A*	*G*	*C*	*A*	*G*
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
C	-1	-1	-2	-3	-4	-5	-6	-7	-6	-7	-8
A	-2	0	0	-1	-2	-3	-4	-5	-6	-5	-6
G	-3	-1	-1	1	0	-1	-2	-3	-4	-5	-4
T	-4	-2	-2	0	2	1	0	-1	-2	-3	-4
A	-5	-3	-1	-1	1	1	2	1	0	-1	-2
T	-6	-4	-2	-2	0	2	1	1	0	-1	-2
C	-7	-5	-3	-3	-1	1	1	0	2	1	0
G	-8	-6	-4	-2	-2	0	0	2	1	1	2
C	-9	-7	-5	-3	-3	-1	-1	1	3	2	1
A	-10	-8	-6	-4	-4	-2	0	0	2	4	3

οπότε η ολική στοίχιση είναι:

A A G T – T A G C A G
C A G T A T C G C A –

η οποία έχει σκορ ίσο με 3 (η τιμή του κελιού κάτω δεξιά στον πίνακα).

3.10. Προσαρμογή αλληλουχιών

Μια άλλη περίπτωση έχουμε όταν θέλουμε να δούμε την προσαρμογή (fit) μιας μικρής αλληλουχίας σε μια μεγαλύτερη, δηλαδή όταν θέλουμε να ανιχνεύσουμε αν μια μικρή αλληλουχία με βιολογική σημασία υπάρχει σε μια μεγαλύτερη. Ο αλγόριθμος αυτός χρησιμοποιεί τη σχέση (3.38) με κάποιες διαφοροποιήσεις όμως. Πιο συγκεκριμένα (Galas, Eggert, & Waterman, 1985):

$$\begin{aligned} &F\left(i,j\right) = \max\left\{F\left(i-1,j-1\right)+s\left(x_i,y_j\right), F\left(i-1,j\right)-d, F\left(i,j-1\right)-d\right\} \\ &\text{με } F\left(i,0\right) = -id \text{ και } F\left(0,j\right) = 0 \end{aligned} \tag{3.40}$$

Παράδειγμα 3.10.1 (Waterman, 1995)
Έστω ότι θέλουμε να ανιχνεύσουμε αν στην αλληλουχία του γονιδίου lacI της E.coli υπάρχει η γνωστή αλληλουχία του υποκινητή (promoter). Έστω ακόμα ότι το τμήμα του γονιδίου έχει αλληλουχία:

x=TCGCGGTATGGCATGATAGCGCCCGGAA,

και η αλληλουχία του υποκινητή είναι:

y=TATAAT

Αν θέσουμε επίσης $s\left(x_i, y_i\right) = \begin{cases} 1, & \text{αν } x_i = y_i \\ -1, & \text{αν } x_i \neq y_i \end{cases}$ και d=2, τότε ο πίνακας F παίρνει τη μορφή:

	T	C	G	C	G	G	T	A	T	G	G	C	A	T	G	A	T	A	G	C	G	C	C	C	G	G	A	A
T	1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
A	0	0	-2	-2	-2	-2	-1	2	0	0	-2	-2	0	-1	0	0	-1	2	0	-2	-2	-2	-2	-2	-2	-2	0	0
T	1	-1	-1	-3	-3	-3	-1	0	3	1	-1	-3	-2	1	-1	-1	1	0	1	-1	-3	-3	-3	-3	-3	-3	-2	-1
A	-1	0	-2	-2	-4	-4	-3	0	1	2	0	-2	-2	-1	0	0	-1	2	0	0	-2	-4	-4	-4	-4	-4	-2	-1
A	-3	-2	-1	-3	-3	-5	-5	-2	-1	0	1	-1	-1	-3	-2	1	-1	0	1	-1	-1	-3	-5	-5	-5	-5	-3	-1
T	-3	-4	-3	-2	-4	-4	-4	-4	-1	-2	-1	0	-2	0	-2	-1	2	0	-1	0	-2	-2	-4	-6	-6	-6	-5	-3

Παρατηρούμε ότι ο αλγόριθμος εντόπισε μια αλληλουχία πιθανού υποκινητή

C A T G A T

η οποία έχει σκορ ίσο με 2 (επειδή το 2 είναι το μέγιστο στοιχείο στην τελευταία σειρά του πίνακα, και με αναδρομή στη γραμμοσκιασμένη περιοχή βρίσκουμε την παραπάνω αλληλουχία).

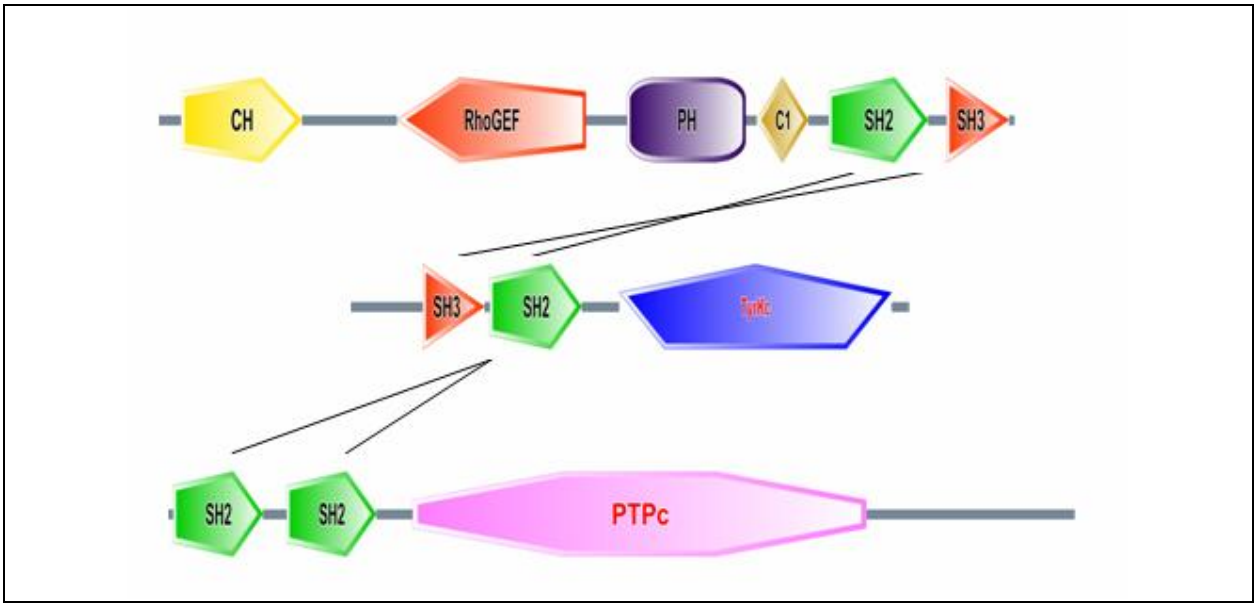
3.11. Τοπική στοίχιση – ο αλγόριθμος Smith και Waterman

Τέλος, μια τρίτη περίπτωση, η οποία όμως παρουσιάζει ιδιαίτερο ενδιαφέρον είναι αυτή που χρησιμοποιείται για τη σύγκριση δυο αλληλουχιών στην περίπτωση που θέλουμε να βρούμε την καλύτερη δυνατή στοίχιση δυο υπό-ακολουθιών τους. Η μέθοδος αυτή ονομάζεται τοπική στοίχιση (local alignment) και δίνει πολλές φορές συνταρακτικά αποτελέσματα ακόμα και σε αλληλουχίες που δεν έχουν καθόλου εμφανή ολική ομοιότητα (ομολογία). Η μέθοδος αυτή είναι η ευρύτερα χρησιμοποιούμενη καθώς μας επιτρέπει και από εξελικτική σκοπιά να διαχωρίζουμε τις αλληλουχίες σε περιοχές που βρίσκονται κάτω από ισχυρή εξελικτική πίεση (και άρα μεταλλάσσονται πολύ αργά) και σε άλλες που μπορεί να διαφέρουν πάρα πολύ (Pearson & Wood, 2001). Όπως επίσης έχουμε αναφέρει, η μέθοδος έχει μεγάλη σημασία στη σύγκριση πρωτεϊνικών αλληλουχιών, καθώς οι πρωτεΐνες αποτελούνται από διαφορετικούς συνδυασμούς περιοχών (domains), και κατά συνέπεια μας ενδιαφέρει πολλές φορές να μπορούμε να εντοπίσουμε τέτοιου είδους ομοιότητες.

Ο αλγόριθμος που επιτυγχάνει τα παραπάνω είναι ο αλγόριθμος των Smith – Waterman (Smith & Waterman, 1981) και χρησιμοποιεί τον εξής αναδρομικό τύπο:

$$F(i,j) = \max\left\{F(i-1,j-1) + s(x_i, y_j), F(i-1,j) - d, F(i,j-1) - d, 0\right\}$$
$$\text{με } F(i,0) = 0 \text{ και } F(0,j) = 0 \tag{3.41}$$

Παρατηρούμε ότι ο αλγόριθμος είναι ίδιος με αυτόν για την ολική στοίχιση με τη διαφορά ότι όποτε μια στοίχιση δίνει αρνητικό *σκορ* αυτή τερματίζεται και αρχίζει μια νέα. Επίσης, και αυτό είναι πολύ σημαντικό, η αρχικοποίηση του πίνακα είναι διαφορετική για να μπορεί να εντοπίσει ομοιότητες σε οποιοδήποτε σημείο εκτός της κύριας διαγωνίου.



Εικόνα 3.11: *Ένα παράδειγμα τοπικής ομοιότητας πρωτεϊνών με διαφορετική σύσταση των περιοχών. Η πρώτη πρωτεΐνη έχει δύο περιοχές που μοιάζουν με περιοχές της δεύτερης πρωτεΐνης (αλλά δεν βρίσκονται στην ίδια θέση στην αλληλουχία). Αντίθετα, η τρίτη πρωτεΐνη διαθέτει μόνο μία από τις περιοχές αυτές, αλλά σε δύο αντίγραφα.*

Παράδειγμα 3.11.1 (Waterman, 1995)

Αν εφαρμόσουμε τη μέθοδο αυτή στις αλληλουχίες του παραδείγματος 3.10.1 έχουμε $s(x_i, y_i) = \begin{cases} 1, & \text{αν } x_i = y_i \\ -1, & \text{αν } x_i \neq y_i \end{cases}$ και d=1 και ο πίνακας F παίρνει τη μορφή:

	-	A	A	G	T	T	A	G	C	A	G
-	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	1	0	0
A	0	1	1	0	0	0	1	0	0	2	1
G	0	0	0	2	1	0	0	2	1	1	3
T	0	0	0	1	3	2	1	1	1	0	2
A	0	1	1	0	2	2	3	2	1	2	1
T	0	0	0	0	1	3	2	2	1	1	1
C	0	0	0	0	0	2	2	1	3	2	1
G	0	0	0	1	0	1	1	3	2	2	3
C	0	0	0	0	0	0	0	2	4	3	2
A	0	1	1	0	0	0	1	1	3	5	4

Επομένως, η καλύτερη τοπική στοίχιση είναι:

A G T A T C G C A
A G T – T A G C A

με σκορ ίσο με 5 (το κελί με τη μεγαλύτερη τιμή στον πίνακα). Πρέπει εδώ να τονίσουμε ότι ο απαιτούμενος χρόνος για να εκτελεστούν οι παραπάνω αλγόριθμοι δυναμικού προγραμματισμού είναι ανάλογος του γινόμενου των μηκών των ακολουθιών και συμβολίζεται $O(mn)$. Το σύμβολο O(mn) (*big-O notation*) δηλώνει ότι ο αριθμός των υπολογισμών που απαιτούνται για να ολοκληρωθεί ο αλγόριθμος, είναι ανάλογος του nm, δηλαδή του πλήθους των κελιών του πίνακα. Κατά συνέπεια, λέμε ότι ο αλγόριθμος είναι γραμμικός ως προς το μήκος των αλληλουχιών. Σε αντιδιαστολή, ο απλοϊκός αλγόριθμος απαρίθμησης όλων των πιθανών στοιχίσεων, είναι εκθετικός ως προς το μήκος των αλληλουχιών (είναι ανάλογος του 2^n).

Πρέπει να τονιστεί τέλος, ότι οι αλγόριθμοι που περιγράφηκαν παραπάνω, αφορούν μόνο την περίπτωση που η ποινή για τα κενά είναι απλή. Σε πραγματικά προβλήματα, απαιτούνται αλγόριθμοι που να υλοποιούν το ρεαλιστικότερο μοντέλο της σύνθετης ποινής για τα κενά. Σε αυτή την περίπτωση, ο αντίστοιχος αλγόριθμος έχει μεγαλύτερη πολυπλοκότητα, της τάξης του $O(nm^2+mn^2)$ γιατί σε κάθε βήμα θα πρέπει να «θυμάται» αν το κενό που έβαλε είναι το πρώτο (open) ή κάποιο από τα επόμενα (extension). Ο αλγόριθμος στην περίπτωση της ολικής στοίχισης γίνεται:

$$F(i,j)=\max\begin{Bmatrix}F(i-1,j-1)+s(x_i,y_j),\\ F(k,j)-\gamma(i-k),k=0,..,i-1\\ F(i,k)-\gamma(j-k),k=0,..,j-1\end{Bmatrix} \tag{3.42}$$

Όμοια τροποποίηση μπορεί να γίνει και στον αλγόριθμο της τοπικής στοίχισης. Έχουν προταθεί, Παρ' όλα αυτά, τροποποιήσεις οι οποίες πραγματοποιούν τον ίδιο υπολογισμό σε χρόνο της τάξης $O(mn)$, με το αντιστάθμισμα, ότι απαιτείται μεγαλύτερη χρήση της μνήμης. Η βασική απαίτηση, είναι ότι η σύνθετη ποινή για τα κενά θα πρέπει να είναι της μορφής της σχέσης (3.35). Ο αλγόριθμος σε αυτή την περίπτωση απαιτεί 3 διαφορετικούς πίνακες, και θα είναι:

$$F(i,j)=\max\begin{Bmatrix}F(i-1,j-1)+s(x_i,y_j)\\ I_x(i-1,j)+s(x_i,y_j)\\ I_y(i,j-1)+s(x_i,y_j)\end{Bmatrix} \tag{3.43}$$

με

$$I_x(i,j)=\max\{F(i-1,j)-d,I_x(i-1,j)-e\} \tag{3.44}$$

και

$$I_y(i,j)=\max\{F(i,j-1)-d,I_y(i,j-1)-e\} \tag{3.45}$$

3.12. Ο νόμος των Erdos και Renyi για τη σύγκριση αλληλουχιών

Είδαμε στις προηγούμενες παραγράφους, με ποιους τρόπους μπορούμε να βρούμε την καλύτερη στοίχιση δύο αλληλουχιών. Το τελευταίο πρόβλημα που μένει, είναι αυτό της εκτίμησης της στατιστικής σημαντικότητας. Συγκεκριμένα μας ενδιαφέρει το πώς μπορούμε να διαχωρίσουμε «τυχαία» ευρήματα από «σημαντικά». Το P-value ενός στατιστικού ελέγχου (γιατί περί τέτοιου πρόκειται) είναι η πιθανότητα, ένα αποτέλεσμα τόσο ακραίο ή και περισσότερο, να έχει προκύψει κατά τύχη, δεδομένου ότι ισχύει η μηδενική υπόθεση (στην περίπτωση μας, ότι οι δύο αλληλουχίες που συγκρίναμε δεν έχουν καμία σχέση μεταξύ τους). Είναι προφανές ότι αναφερόμαστε σε παραμετρικό έλεγχο και χρειάζεται να ξέρουμε την κατανομή που ακολουθεί η τυχαία μεταβλητή που μας ενδιαφέρει, το σκορ δηλαδή. Για να απαντήσουμε σε αυτό το ερώτημα, θα πρέπει να δούμε πάλι το θέμα των ροών των επιτυχιών και τα δεδομένα της κατανομής του Gumbel.

Όπως είδαμε ήδη, η σύγκριση ακολουθιών, μοιάζει με τη μελέτη των ροών σε αλληλουχίες. Η διαφορά είναι ότι το πρόβλημα είναι τώρα δισδιάστατο. Κατά συνέπεια, οι ασυμπτωτικοί νόμοι των Erdos και Renyi που ισχύουν για τις ροές, βρίσκουν εφαρμογή και εδώ (Waterman, 1995). Έστω ότι έχουμε δυο αλληλουχίες $x=x_1,x_2,\ldots,x_n$ και $y=y_1,y_2,\ldots,y_m$. Τότε η μέγιστη περιοχή ταύτισης μεταξύ τους έχει μήκος $M_n \cong \log_{1/p}(mn)$ ή αλλιώς:

$$\frac{M_n}{\log_{1/p}(mn)}\to 1 \text{ με πιθανότητα } 1 \tag{3.46}$$

Προφανώς, η πιθανότητα ταύτισης p είναι ίση με $p=P(x_i=y_j)\Leftrightarrow p=p_A^2+p_T^2+p_G^2+p_C^2$ αν η κατανομή των συμβόλων στις δυο αλληλουχίες είναι ίδια. Η διαισθητική ερμηνεία εδώ είναι εντελώς ανάλογη με αυτή που δώσαμε στην περίπτωση της μελέτης μίας ακολουθίας, με τη μόνη διαφορά ότι τώρα υπάρχουν περίπου mn δυνατά σημεία εμφάνισης της περιοχής ταύτισης. Αντίστοιχο αποτέλεσμα θα έχουμε και για το μήκος της περιοχής μη απόλυτης ταύτισης. Αν για παράδειγμα έχουμε δυο αλληλουχίες $\mathbf{x}=x_1,x_2,\ldots,x_n$ και $\mathbf{y}=y_1,y_2,\ldots,y_m$ με $0\le p<\alpha\le 1$. Τότε για τη μέγιστη περιοχή που περιέχει $100\alpha\%$ όμοια νουκλεοτίδια μεταξύ τους ισχύει:

$$\frac{M_n}{\log_{1/p}(mn)} \to \frac{1}{H(a,p)} \text{ με πιθανότητα 1} \tag{3.47}$$

Η ποσότητα $H(\alpha,p)$ είναι η σχετική εντροπία όπως την ορίσαμε παραπάνω. Όπως θα εξηγήσουμε, οι παραπάνω σχέσεις ισχύουν για τοπικές (local) συγκρίσεις ακολουθιών. Γενικά από εδώ και πέρα θα ασχοληθούμε με την κατανομή του Local Similarity Score αφ' ενός μεν γιατί είναι πιο σημαντικό από πρακτική άποψη αφ' ετέρου δε γιατί τα πιο σημαντικά θεωρητικά αποτελέσματα που έχουν βρεθεί, αφορούν αυτό. Επιπλέον, τα παραπάνω αποτελέσματα έχουν επεκταθεί (Arratia, Gordon and Waterman, 1986; Arratia, Gordon and Waterman, 1990) και έχουν δοθεί ακόμα πιο ακριβείς τύποι για τη μέση τιμή του μήκους της μέγιστης περιοχής ταύτισης. Για την ακρίβεια, ο Arratia (1990) έδωσε μια καλύτερη προσέγγιση για τη μέση τιμή του μήκους της μέγιστης περιοχής ταύτισης μεταξύ δύο ακολουθιών, η οποία είναι:

$$E(M_n) \approx \frac{\log(mn)}{\lambda} + \frac{\log(q)}{\lambda} + \frac{\gamma}{\lambda} - \frac{1}{2} \tag{3.48}$$

όπου q=1-p, και γ=-Γ'(1) = 0.5772… η σταθερά Euler-Mascheroni, και λ=log(1/p). Παράλληλα, έδωσαν και προσεγγιστικό τύπο για την αντίστοιχη διασπορά:

$$Var(M_n) \approx \frac{\pi^2}{6\lambda^2} + \frac{1}{12} \tag{3.49}$$

Οι σχέσεις αυτές, είναι εντελώς ανάλογες με τις αντίστοιχες σχέσεις (3.17) και (3.18) που είδαμε για τη μελέτη μίας αλληλουχίας. Εντελώς ανάλογη είναι και η ερμηνεία για τις διαφορές των σχέσεων (3.48) και (3.47), καθώς ο κύριος παράγοντας που καθορίζει την τελική τιμή εξακολουθεί να είναι η τιμή του log(mn), καθώς το log(q) και το γ/λ είναι σχετικά μικρές ποσότητες. Προφανώς, όσο το m και n μεγαλώνουν, η διαφορά θα γίνεται ακόμα μικρότερη. Οι Arratia και Waterman (Arratia & Waterman, 1989), έδωσαν και ακριβή τύπο για τον υπολογισμό της μέγιστης περιοχής ταύτισης δυο αλληλουχιών, με δεδομένο τον αριθμό k των μη όμοιων νουκλεοτιδίων (mismatches). Σε αυτή την περίπτωση, έχουμε δυο αλληλουχίες $\mathbf{x}=x_1,x_2,\ldots,x_n$ και $\mathbf{y}=y_1,y_2,\ldots,y_m$ και μας ενδιαφέρει η μέση τιμή για το μήκος της μέγιστης περιοχής ταύτισης μεταξύ τους, όταν υπάρχουν k μη κοινά νουκλεοτίδια (k mismatches):

$$E(M_n) \approx \log_{1/p}(qn^2) + k\log_{1/p}\log_{1/p}(qn^2) + k\log_{1/p}(q) - \log_{1/p}(k!) + k + \frac{\gamma}{\lambda} - \frac{1}{2} \tag{3.50}$$

Για την αντίστοιχη διασπορά ισχύει όπως και προηγουμένως:

$$\text{var}(M_n) \approx \frac{\pi^2}{6\lambda^2} + \frac{1}{12} \tag{3.51}$$

Όπως και παραπάνω, q=1-p, και γ=-Γ'(1)=0.5772…η σταθερά Euler-Mascheroni, και λ=log(1/p).

3.13. Η ασυμπτωτική κατανομή του local similarity score

Το επόμενο λογικό βήμα είναι να προσδιορίσουμε την ακριβή κατανομή του Local Similarity Score και κατ' επέκταση του μήκους της μέγιστης κοινής υπό-ακολουθίας, έτσι ώστε να μπορούμε να υπολογίσουμε τη στατιστική σημαντικότητα μιας δεδομένης σύγκρισης δυο αλληλουχιών. Δηλαδή, να μπορέσουμε να προσδιορίσουμε αν το αποτέλεσμα μιας τέτοιας σύγκρισης οφείλεται στην τύχη και μόνο, ή αν υπάρχει μια βιολογική σημαντικότητα στην ομοιότητα αυτή των δυο αλληλουχιών. Παραδοσιακά, οι βιολόγοι είχαν αναπτύξει διάφορους εμπειρικούς κανόνες. Ο πιο ακριβής από αυτούς, λέει ότι δυο πρωτεΐνες είναι «όμοιες» αν έχουν τουλάχιστον 30% ομοιότητα (similarity) σε μήκος στοίχισης τουλάχιστον 80 αμινοξικά κατάλοιπα. Θα δούμε, ότι σε γενικές γραμμές ο κανόνας αυτός αποδίδει, αλλά ο ακριβής υπολογισμός της στατιστικής σημαντικότητας μπορεί να προσδώσει πολλά πλεονεκτήματα, ειδικά στις οριακές καταστάσεις, και ακόμα περισσότερο στις αναζητήσεις σε βάσεις δεδομένων, όπου το πρόβλημα των πολλαπλών ελέγχων είναι υπαρκτό.

Πρέπει να τονίσουμε εδώ ότι ακριβή θεωρητικά αποτελέσματα έχουν δοθεί μόνο για την κατανομή του «Local Similarity Score without gaps», δηλαδή για την περίπτωση κατά την οποία στην στοίχιση δεν υπάρχουν κενά, αν και υπάρχουν ενδείξεις ότι τα ίδια αποτελέσματα γενικεύονται και για την περίπτωση ύπαρξης κενών. Κατ' αρχήν πρέπει να δούμε τις δυο ακραίες περιπτώσεις. Πρώτον, όταν ορίσουμε

$$s(x_i,y_i) = \begin{cases} 1, & \text{αν } x_i = y_i \\ 0, & \text{αν } x_i \neq y_i \end{cases} \text{ και d=0}$$

δηλαδή, όταν δεν υπάρχουν ποινές για κενά και για διαφορές, τότε έχουμε $s(x_i,y_j)\sim cn$ και βρισκόμαστε στη λεγόμενη γραμμική περιοχή για την οποία δεν υπάρχουν προς το παρόν ενδείξεις για την κατανομή του σκορ, αλλά δεν υπάρχει και πρακτικό ενδιαφέρον καθώς με τις παραπάνω ποινές οποιεσδήποτε αλληλουχίες θα μπορούν να στοιχηθούν «καλά». Αντίθετα αν χρησιμοποιήσουμε

$$s\left(x_i, y_i\right) = \begin{cases} 1, \text{ αν } x_i = y_i \\ -\infty, \text{ αν } x_i \neq y_i \end{cases} \text{ και d}=\infty$$

δηλαδή, αν δεν επιτρέπονται καθόλου τα κενά και οι διαφορές, τότε έχουμε $s(x_i,y_j)\sim k\log n$ και βρισκόμαστε στη λογαριθμική περιοχή όπου ισχύει η σχέση (3.46). Προφανώς σ' αυτήν την περιοχή ανήκει και η σχέση (3.47) γιατί και σ' αυτή βλέπουμε ότι το μήκος της περιοχής ταύτισης αυξάνεται με τον λογάριθμο του n. Μειώνοντας σταδιακά τις ποινές για τις διαφορές και τα κενά, μεταπίπτουμε από τη λογαριθμική περιοχή του σκορ, στη γραμμική. Αυτή η μετάπτωση φάσεως (phase transition) έχει περιγραφεί θεωρητικά από τους Arratia, Gordon και Waterman (Arratia & Waterman, 1994; Waterman, 1995; Waterman, Gordon, & Arratia, 1987), αλλά παρ' όλα αυτά δεν υπάρχει αναλυτική έκφραση για τις τιμές των παραμέτρων m (mismatch) και d (gap), στις οποίες συμβαίνει αυτή η μετάπτωση.

Όπως είναι φανερό εμείς ενδιαφερόμαστε για την κατανομή του σκορ στη λογαριθμική περιοχή. Στην περιοχή αυτή η εμφάνιση θετικών σκορ, δηλαδή η ύπαρξη κοινών υπό-ακολουθιών, είναι σπάνια γεγονότα. Επομένως, ασυμπτωτικά θα περιγράφονται από μια κατάλληλη κατανομή Poisson, με μέση τιμή:

$$\mathrm{E}\left(S \geq x\right) = Kmne^{-\lambda x} = Kmnp^{x} \tag{3.52}$$

όπου K είναι μια σταθερά <1 η οποία διορθώνει τον παράγοντα mn, και λ η μοναδική θετική ρίζα της εξίσωσης $\sum q_i q_j e^{\lambda S}$ =1. Στην ιδανική περίπτωση για την οποία δεν επιτρέπονται τα κενά αλλά ούτε και διαφορές, μπορούν να δοθούν κλειστές εκφράσεις για τα K και λ και αυτές είναι :

$$K = 1 - p = q \tag{3.53}$$

και

$$\lambda = \log\left(1/p\right) \tag{3.54}$$

Όταν από την άλλη πλευρά επιτρέπονται διαφορές, τότε τα K, λ υπολογίζονται με αριθμητικές μεθόδους. Πιο συγκεκριμένα, κομβικό ρόλο στη μελέτη των στατιστικών της τοπικής στοίχισης έχει το θεώρημα των Karlin και Altschul (Karlin & Altschul, 1990) το οποίο λέει ότι στη σύγκριση δύο αλληλουχιών $\mathbf{x}=x_1,x_2,\ldots,x_n$ και $\mathbf{y}=y_1,y_2,\ldots,y_m$ με σκορ S όπως το ορίσαμε στη σχέση (3.32), η πιθανότητα να προκύψει ένα σκορ μεγαλύτερο από το x (δηλαδή, το p-value), δίνεται από τη σχέση:

$$P\left(S > x\right) \approx 1 - \exp\left(-Kmne^{-\lambda x}\right) \tag{3.55}$$

Για τον υπολογισμό του μέγιστου τοπικού σκορ (local similarity score) πρέπει να θέσουμε κάποιους περιορισμούς, όμοιους με την περίπτωση του maximal segment score. Πιο συγκεκριμένα:

- Τουλάχιστον ένα σκορ πρέπει να είναι θετικό
- Η αναμενόμενη τιμή του σκορ για μια τυχαία θέση στη στοίχιση να είναι αρνητική, δηλαδή:

$$E\left(s_{ij}\right) = \sum q_i q_j s_{ij} = \sum q_i q_j \log\left(\frac{q_i q_j}{p_{ij}}\right) < 0 \tag{3.56}$$

Το λ είναι όπως είπαμε ήδη, η μοναδική θετική ρίζα της εξίσωσης $\sum q_i q_j e^{\lambda S}$ =1. Προφανώς, οι παραπάνω δυο περιορισμοί είναι απαραίτητοι για να είμαστε σίγουροι ότι το σκορ θα παίρνει τιμές στη λογαριθμική περιοχή, και κατά συνέπεια θα είναι όντως τοπικό. Η σχέση αυτή, γράφεται ισοδύναμα ως:

$$P\left(S \leq x\right) = \exp\left(-Kmne^{-\lambda x}\right) \tag{3.57}$$

Η τελευταία σχέση είναι, με άλλη παραμετροποίηση, η α.σ.κ. της κατανομής των ακραίων τιμών του Gumbel (EVD). Αν κάνουμε μετασχηματισμό, βλέπουμε ότι ισχύει:

$$P\left(S \leq x\right) = \exp\left(-e^{-\frac{(x-a)}{b}}\right), -\infty \leq x \leq \infty \tag{3.58}$$

με

$$E\left(x\right) = a - b\Gamma'\left(1\right),\ V\left(x\right) = \frac{b^2\pi^2}{6} \tag{3.59}$$

Οι παράμετροι a,b είναι προφανώς $a=\frac{\log(kmn)}{\lambda}, b=\frac{1}{\lambda}$ με $\lambda=\log\left(\frac{1}{p}\right)$ και $K=1-p=q$, όταν δεν επιτρέπονται διαφορές. Από τις παραπάνω σχέσεις είναι δυνατόν να υπολογιστεί το p-value για ένα δεδομένο σκορ που προέκυψε από τη σύγκριση δυο αλληλουχιών. Αφού τυποποιήσουμε τη μεταβλητή, έχουμε (Pearson, 1998; Pearson & Wood, 2001):

$$P(Z \geq z) = 1 - \exp\left(-\exp-\left(\left(\frac{\pi}{\sqrt{6}}\right)z - \Gamma'(1)\right)\right) \tag{3.60}$$

Η σχέση αυτή, είναι ισοδύναμη με την (3.55) αλλά δεν περιέχει σταθερές ενώ το σκορ είναι εκφρασμένο σε τυποποιημένες μονάδες (z). Από την παραπάνω σχέση μπορούμε να υπολογίσουμε την πιθανότητα να εμφανισθεί ένα σκορ που να ξεπερνά κάποιες (z) φορές την τυπική απόκλιση της θεωρητικής κατανομής δεδομένου ότι οι αλληλουχίες είναι τυχαίες (ασυσχέτιστες).

Πρέπει σ' αυτό το σημείο να τονίσουμε ότι διαφορετική ερμηνεία έχει ένα p-value που προκύπτει από τη σύγκριση δυο διαφορετικών αλληλουχιών, και άλλη έχει ένα p-value που θα προκύψει από σύγκριση μεταξύ της ίδιας αλληλουχίας με μια μεγάλη βάση δεδομένων με πολλές χιλιάδες αλληλουχίες. Αυτό ισχύει ακόμα και αν η στοίχιση που προέκυψε στις δύο περιπτώσεις είναι πανομοιότυπη. Έτσι ένα p-value που προκύπτει από τη σύγκριση δυο διαφορετικών αλληλουχιών με τιμή 10^{-4}, μπορεί να φαίνεται στατιστικά σημαντικό αλλά αν πρόκειται για σύγκριση μεταξύ μιας ακολουθίας με μια μεγάλη βάση δεδομένων με 100.000 αλληλουχίες τότε λόγω της τύχης και μόνο αναμένεται να εμφανιστεί τουλάχιστον 10 φορές.

Όταν συγκρίνουμε μια αλληλουχία με μια ολόκληρη βάση δεδομένων, η οποία περιέχει D αλληλουχίες, τότε η παρατήρηση ακολουθιών οι οποίες εμφανίζουν μικρό p-value (μεγάλη ομοιότητα- p-match) είναι σπάνιο ενδεχόμενο, και θα περιγράφεται από την κατανομή Poisson. Άρα (Pearson & Wood, 2001): P = Pr(τουλάχιστον 1 σκορ $S \geq x$) = $1-e^{-Dp}$ και αν το Dp είναι πολύ μικρό (<0.01) θα έχουμε: $P \approx Dp$. Στο ίδιο αποτέλεσμα θα καταλήγαμε αν υπολογίζαμε την αναμενόμενη τιμή για τις εμφανίσεις περιοχών με σκορ $S \geq x$, έπειτα από D συγκρίσεις με τις αλληλουχίες της βάσης δεδομένων. Αυτό το E-value (expectation value) είναι ίσο με $E(S \geq x) = DP(S \geq x)$ όπου D είναι ο αριθμός των ανεξάρτητων αλληλουχιών που περιέχει η υπό έλεγχο βάση δεδομένων.

Για να έχουμε περισσότερο ακριβή αποτελέσματα, μια πιο σωστή προσέγγιση θα προέκυπτε αν λαμβάναμε υπόψη το γεγονός ότι όλες οι αλληλουχίες στη βάση δεδομένων δεν έχουν τον ίδιο αριθμό βάσεων. Πρακτικά αυτό σημαίνει ότι θεωρούμε ολόκληρη τη βάση δεδομένων ως μια τεράστια αλληλουχία από N νουκλεοτίδια (ή αμινοξέα) και συγκρίνουμε με αυτήν τη συγκεκριμένη αλληλουχία μας η οποία έχει μήκος n βάσεις (ή αμινοξέα). Κατά μέσο όρο κάθε μια από τις αλληλουχίες της βάσης περιέχει $m=N/D$ βάσεις, οπότε η πιθανότητα να υπάρχει μια περιοχή με σκορ μεγαλύτερο από x, όπως είπαμε παραπάνω είναι $P(S>x) = 1-e^{-E(S)} = 1-\exp(-Kne^{-\lambda x})$ ενώ η αναμενόμενη τιμή (E-value) θα είναι $E(S \geq x) = Kne^{-\lambda x} = DKmne^{-\lambda x}$.

Πολλά προγράμματα όπως το BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990), αντί του p-value, αναφέρουν ως αποτέλεσμα (output) αυτήν την τιμή, επειδή είναι πιο εύκολη η ερμηνεία της από κάποιο μη ειδικό, αλλά όπως είδαμε όταν το E-value είναι πολύ μικρό τότε, επειδή ισχύει η προσεγγιστική σχέση (Waterman, 1995): $1-\exp(-\exp(-t)) \approx 1-(1-\exp(-t)) = \exp(-t)$, το p-value θα είναι περίπου ίσο με το E-value. Είναι φανερό ότι, σήμερα που οι βάσεις δεδομένων αυξάνονται συνεχώς σε μέγεθος, είναι καλύτερο κάθε φορά που γίνονται τέτοιες συγκρίσεις να αναφέρονται τουλάχιστον μαζί το p-value και το E-value.

3.14. Η κατανομή του σκορ όταν υπάρχουν κενά

Όταν στη στοίχιση δυο αλληλουχιών υπάρχουν κενά δεν υπάρχει μαθηματική απόδειξη που να περιγράφει την κατανομή που ακολουθεί το σκορ. Παρ' όλα αυτά πολλοί ερευνητές έχουν προτείνει (Altschul et al., 1997; Clote & Backofen, 2000; Mott, 2000), ότι και σ' αυτήν την περίπτωση η κατανομή του σκορ είναι η κατανομή των ακραίων τιμών του Gumbel:

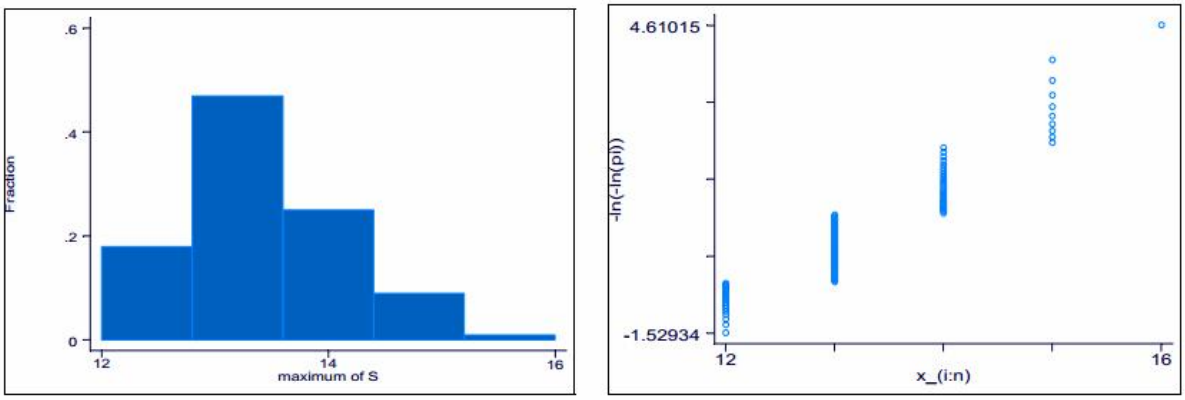
$$P(S \leq x) = \exp\left(-Kmne^{-\lambda x}\right) \tag{3.61}$$

με τη διαφορά ότι οι παράμετροι K,λ είναι διαφορετικές από την περίπτωση της στοίχισης χωρίς κενά.

Η πρώτη προσπάθεια να υπολογιστούν οι παράμετροι της κατανομής του Gumbel όταν υπάρχουν κενά έγινε από τον Mott, το 1992 (Mott, 1992). Συγκεκριμένα θεώρησε μια παραλλαγή της σχέσης (3.58) και

έθεσε $A = a_0 + \frac{a_1}{\lambda} + \frac{\alpha_2 \log(mn)}{\lambda}, B = \frac{b_1}{\lambda}$. Τις σταθερές α_0, α_1, α_2 και b_1 τις εκτίμησε με μέγιστη πιθανοφάνεια. Το λ είναι και πάλι η μοναδική θετική ρίζα της εξίσωσης $\sum q_i q_j e^{\lambda S} = 1$.

Ένας πιο απλός, αλλά και χρονοβόρος τρόπος υπολογισμού των παραμέτρων αυτών είναι η απευθείας εκτίμηση (direct estimation) (Waterman, 1995; Waterman & Vingron, 1994). Στην περίπτωση αυτή απαιτείται ένας μεγάλος αριθμός προσομοιώσεων (συγκρίσεις με τυχαίες αλληλουχίες). Πιο συγκεκριμένα, αφού πραγματοποιήσουμε την τοπική στοίχιση των δυο αλληλουχιών, πραγματοποιούμε πολλές (στη βιβλιογραφία αναφέρεται ότι πρέπει να είναι τουλάχιστον 1000) συγκρίσεις μεταξύ τυχαίων αλληλουχιών με παρόμοια σύσταση βάσεων με τις αρχικές (αυτό ονομάζεται shuffling, ανακάτεμα αλληλουχιών). Κατόπιν, αφού υπολογίσουμε την εμπειρική α.σ.κ. (e.c.d.f.) και εφαρμόσουμε κατάλληλο μετασχηματισμό (log[-log[cdf]]) κάνουμε μια απλή γραμμική παλινδρόμηση του log[-log[cdf]] με το *S*. Η κλίση (slope) της ευθείας ελαχίστων τετραγώνων θα είναι ίση με –λ και η σταθερά της (constant) θα είναι ίση με log(*Kmn*).



Εικόνα 3.12: *Κατανομή της μέγιστης κοινής υπο-ακολουθίας από τις συγκρίσεις αλληλουχιών DNA μήκους 10000 βάσεων. Αριστερά βλέπουμε την κατανομή του σκορ ενώ δεξιά, τον μετασχηματισμό log[-log[cdf]]από τον οποίο θα εκτιμήσουμε τις σταθερές K και λ της κατανομής.*

Εναλλακτικά, αν πραγματοποιείται σύγκριση μιας αλληλουχίας με μια βάση δεδομένων, είναι δυνατόν για την εύρεση της κατανομής, να χρησιμοποιηθούν τα αποτελέσματα από τις συγκρίσεις που έγιναν κατά τη διάρκεια της αναζήτησης. Ιδιαίτερη προσοχή εδώ θέλει το γεγονός, ότι για να πετύχει η μέθοδος πρέπει να έχουν απομακρυνθεί όλες οι αλληλουχίες της βάσης με πολύ μεγάλα (z>7) και πολύ μικρά (z<-3) σκορ, έτσι ώστε να αποφευχθεί μεροληψία στα αποτελέσματα (systematic error – bias) (Pearson, 1998).

Μια άλλη μέθοδος που προτάθηκε από τους Waterman και Vingron (Waterman & Vingron, 1994; Waterman & Vingron, 1994), αναδεικνύει τη δύναμη της προσέγγισης Poisson (Poisson Approximation - (Arratia, Goldstein, & Gordon, 1989; Chen, 1975)) και ονομάζεται "de-clumping estimation". Η μέθοδος αυτή χρησιμοποιεί το διατεταγμένο δείγμα $S_{(1)} \geq S_{(2)} \geq \ldots S_{(k)}$ και όχι μόνο το μέγιστο σκορ. Βασική προϋπόθεση εδώ είναι ότι οι τοπικές περιοχές που θα δώσουν αυτά τα σκορ πρέπει να είναι ανεξάρτητες μεταξύ τους, δηλαδή να μην ταυτίζονται σε κάποιο μικρότερο τμήμα τους. Οι διατεταγμένες παρατηρήσεις $S_{(i)}$ ακολουθούν ασυμπτωτικά την κατανομή Poisson με μέση τιμή, όπως είδαμε:

$$E(S \geq x) = Kmne^{-\lambda x} \tag{3.62}$$

Άρα, η πιθανότητα να υπάρχουν k ανεξάρτητες μεταξύ τους περιοχές, με σκορ μεγαλύτερο από x θα είναι:

$$P(S_{(k)} > x) \approx 1 - \exp\left(-Kmne^{-\lambda x}\right) \sum_{i=0}^{k-1} \frac{\left(Kmne^{-\lambda x}\right)^i}{i!} \tag{3.63}$$

και η μέση τιμή αυτής της κατανομής θα δίνεται από τη σχέση (3.52). Επομένως, παριστάνοντας γραφικά το log[data] (το λογάριθμο του αριθμού τοπικών περιοχών με σκορ πάνω από κάποιο όριο) σε σχέση με το $Kmne^{-\lambda x}$ (τη μέση τιμή του σκορ για τις περιοχές πάνω από το όριο αυτό) παίρνουμε ευθεία γραμμή και μια απλή γραμμική παλινδρόμηση δίνει αμέσως εκτιμήτριες για τα *K*,λ. Ισοδύναμα μια παλινδρόμηση Poisson μεταξύ αριθμού παρατηρήσεων που ξεπερνούν κάποιο σημείο, και της μέσης τιμής του σκορ για παρατηρήσεις πάνω από το κάθε σημείο, δίνει τις εκτιμήτριες για τα *K*,λ.

Οι δυο προηγούμενες μέθοδοι δίνουν ταυτόσημα αποτελέσματα αλλά η δεύτερη είναι πολλές φορές γρηγορότερη καθώς απαιτεί λιγότερες προσομοιώσεις. Τούτο συμβαίνει διότι για κάθε προσομοίωση η μέθοδος του Poisson υπολογίζει μόνο μια φορά τον πίνακα *nm* από τον αλγόριθμο Smith-Waterman, και από αυτόν υπολογίζει τα *k* υποβέλτιστα σκορ (sub-optimal alignments). Οι Waterman και Vingron, αναφέρουν ότι χρειάζονται περίπου 10 προσομοιώσεις, με κατάλληλο αριθμό sub-optimal scores για να πάρουμε καλούς εκτιμητές για τα *K*,*λ*.

Μια άλλη προσέγγιση στην εύρεση της στατιστικής σημαντικότητας όταν επιτρέπεται η ύπαρξη κενών, είναι αυτή που προτάθηκε από τον Pearson (Pearson, 1995, 1998; Pearson & Wood, 2001), και αφορά τη σύγκριση μιας αλληλουχίας με μια βάση δεδομένων. Είναι δηλαδή παραλλαγή της μεθόδου απευθείας εκτίμησης. Κατά τη διαδικασία αυτή, η βάση δεδομένων χωρίζεται σε *k* υποσύνολα σύμφωνα με το μήκος των αλληλουχιών n_1, n_2,…, n_k που περιέχουν, έτσι ώστε τα υποσύνολα αυτά να διαφέρουν το καθένα από το επόμενο στο μέσο μήκος αλληλουχιών που περιέχουν, κατά περίπου 10%. Υπολογίζονται κατόπιν, όλα τα σκορ *S*, για την τοπική ομοιότητα των αλληλουχιών, και στη συνέχεια μια ευθεία σταθμισμένης γραμμικής παλινδρόμησης (weighted linear regression) για τη σχέση:

$$S = a + b\log\left(n_i\right) \tag{3.64}$$

Εδώ, το n_i, είναι το μήκος των αλληλουχιών του *i* υποσυνόλου της βάσης δεδομένων, ενώ το $\log(n_i)$ είναι σταθμισμένο με την αντίστροφη διασπορά (1/var) των σκορ σε αυτό το υποσύνολο, καθώς τμήματα με πολύ μεγάλο σκορ θα έχουν και μεγάλη διασπορά. Υπολογίζεται τέλος η εκτιμήτρια της διασποράς $\hat{\sigma}^2$, των καταλοίπων της παλινδρόμησης (residual variance) η οποία καθορίζει το z-score:

$$z = \frac{S - \left(a + b\log\left(n_i\right)\right)}{\text{var}} \tag{3.65}$$

Οι αλληλουχίες με πολύ μεγάλη διασπορά του σκορ εξαιρούνται, επειδή θεωρούνται ότι είναι αυτές με μεγάλη ομοιότητα, και άρα θα προσδώσουν συστηματικό σφάλμα στις εκτιμήσεις των παραμέτρων. Η όλη διαδικασία επαναλαμβάνεται έως και 5 φορές, για να απομακρυνθούν όλες οι συσχετισμένες (με μεγάλο σκορ) αλληλουχίες. Τελικά υπολογίζονται όλα τα z-scores, για τις αλληλουχίες της βάσης, και από τη σχέση :

$$P\left(Z \geq z\right) = 1 - \exp\left(-\exp-\left(\left(\frac{\pi}{\sqrt{6}}\right)z - \Gamma'(1)\right)\right) \tag{3.66}$$

υπολογίζεται η στατιστική σημαντικότητα (p-value, E-value), για κάθε μια από τις συγκρίσεις που έχουν πραγματοποιηθεί. Η μέθοδος αυτή είναι πολύ χρήσιμη καθώς επιτρέπει «εσωτερική» ρύθμιση για την ακρίβεια των παραμέτρων που εκτιμούμε, και έχει αποδειχθεί ότι δίνει πολύ καλά αποτελέσματα στην πράξη.

Ο Mott (Mott, 2000), πρότεινε τέλος, μια άλλη μέθοδο για τον υπολογισμό των παραμέτρων *K*, *λ* της κατανομής του Gumbel όταν υπάρχουν κενά. Συγκεκριμένα χρησιμοποιώντας μια μικρή τροποποίηση του αλγόριθμου των Smith-Waterman, και ένα συνδυασμό αριθμητικών και στατιστικών μεθόδων, κατόρθωσε να δώσει ακριβείς τύπους για τον υπολογισμό των *K*, *λ* ως συνάρτηση της ποινής για τα κενά. Οι τύποι αυτοί, στην περίπτωση που δεν επιτρέπονται κενά, ανάγονται στις αντίστοιχες παραμέτρους όπως τις προβλέπει η γενική θεωρία. Η παραπάνω μέθοδος δίνει επίσης πολύ καλά αποτελέσματα και επιπλέον λαμβάνει υπόψη τη διαφορετική σύνθεση κάθε αλληλουχίας της βάσης δεδομένων.

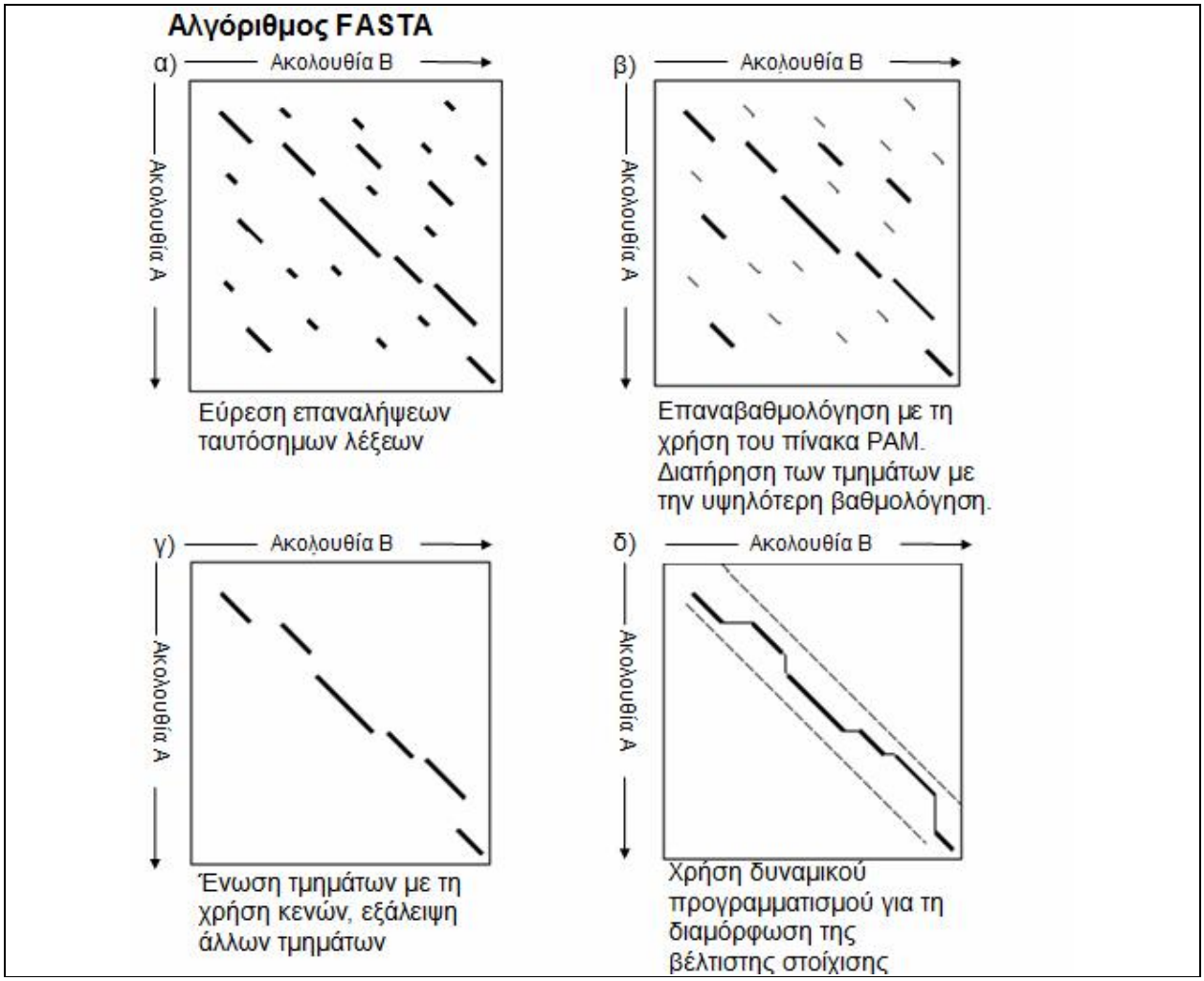
Όπως είναι φανερό, αν και δεν έχουμε το πλήρες θεωρητικό πλαίσιο για να εκτιμήσουμε τη στατιστική σημαντικότητα από συγκρίσεις αλληλουχιών όταν επιτρέπονται τα κενά, έχουμε στη διάθεση μας πληθώρα μεθόδων, που δίνουν πολύ καλά προσεγγιστικά αποτελέσματα. Το ποια μέθοδος θα χρησιμοποιηθεί, πέραν από τη διαθεσιμότητα των προγραμμάτων και τους πρακτικούς περιορισμούς, εξαρτάται κυρίως από το είδος των αλληλουχιών που συγκρίνουμε και το είδος της ομοιότητας που ελπίζουμε να ανακαλύψουμε.

3.15. Ευριστικοί αλγόριθμοι - BLAST και FASTA

Όπως είπαμε ήδη, ο αλγόριθμος Smith και Waterman δίνει σε κάθε περίπτωση την καλύτερη δυνατή στοίχιση μεταξύ δύο αλληλουχιών, λαμβάνοντας υπόψιν τον πίνακα ομοιότητας και τις ποινές για τα κενά. Παρ' όλα αυτά, σε πρακτικές εφαρμογές, είναι δύσχρηστος. Τούτο συμβαίνει, όχι τόσο για την περίπτωση που ενδιαφερόμαστε για τη σύγκριση δύο αλληλουχιών, αλλά περισσότερο για την αναζήτηση ομοιότητας σε μια βάση δεδομένων.

Οι αναζητήσεις στις βάσεις δεδομένων, είναι ένα βασικό εργαλείο στην υπολογιστική ανάλυση αλληλουχιών και είναι στην πραγματικότητα, μέρος της καθημερινής ρουτίνας ακόμα και των εργαστηριακών

μοριακών βιολόγων. Το μεγάλο πρόβλημα προκύπτει, όπως έχουμε δει στο κεφάλαιο 2, από τη συνεχή αύξηση του όγκου των δεδομένων που βρίσκονται κατατεθειμένα στις δημόσιες βάσεις. Είδαμε, ότι ο αριθμός των καταχωρήσεων διπλασιάζεται σε λιγότερο από δύο χρόνια, και ο ρυθμός αυτός είναι ίσως και πιο γρήγορος από την αύξηση της υπολογιστικής ισχύος. Αυτό, το είχαν αντιληφθεί ήδη από τη δεκαετία του 1980, οπότε και ξεκίνησε η έρευνα για τη δημιουργία γρήγορων και αποδοτικών ευριστικών αλγορίθμων, οι οποίοι θα κάνουν την ίδια δουλειά αλλά σε μικρότερο χρόνο. Η βασική απαίτηση από έναν ευριστικό (heuristic) αλγόριθμο, είναι να αποδίδει «σχεδόν» πάντα το ίδιο καλά με τον αυστηρά μαθηματικό αλγόριθμο, αλλά να πραγματοποιεί τις αναλύσεις πολλές φορές πιο γρήγορα. Το «σχεδόν πάντα», δεν μπορεί να αποδειχθεί θεωρητικά αλλά μπορεί να τεκμηριωθεί με εμπειρικές αναλύσεις. Οι δύο πιο σημαντικοί αλγόριθμοι αυτής της κατηγορίας, είναι το BLAST (Altschul, et al., 1990; Altschul, et al., 1997) και το FASTA (Lipman & Pearson, 1985; Wilbur & Lipman, 1983).

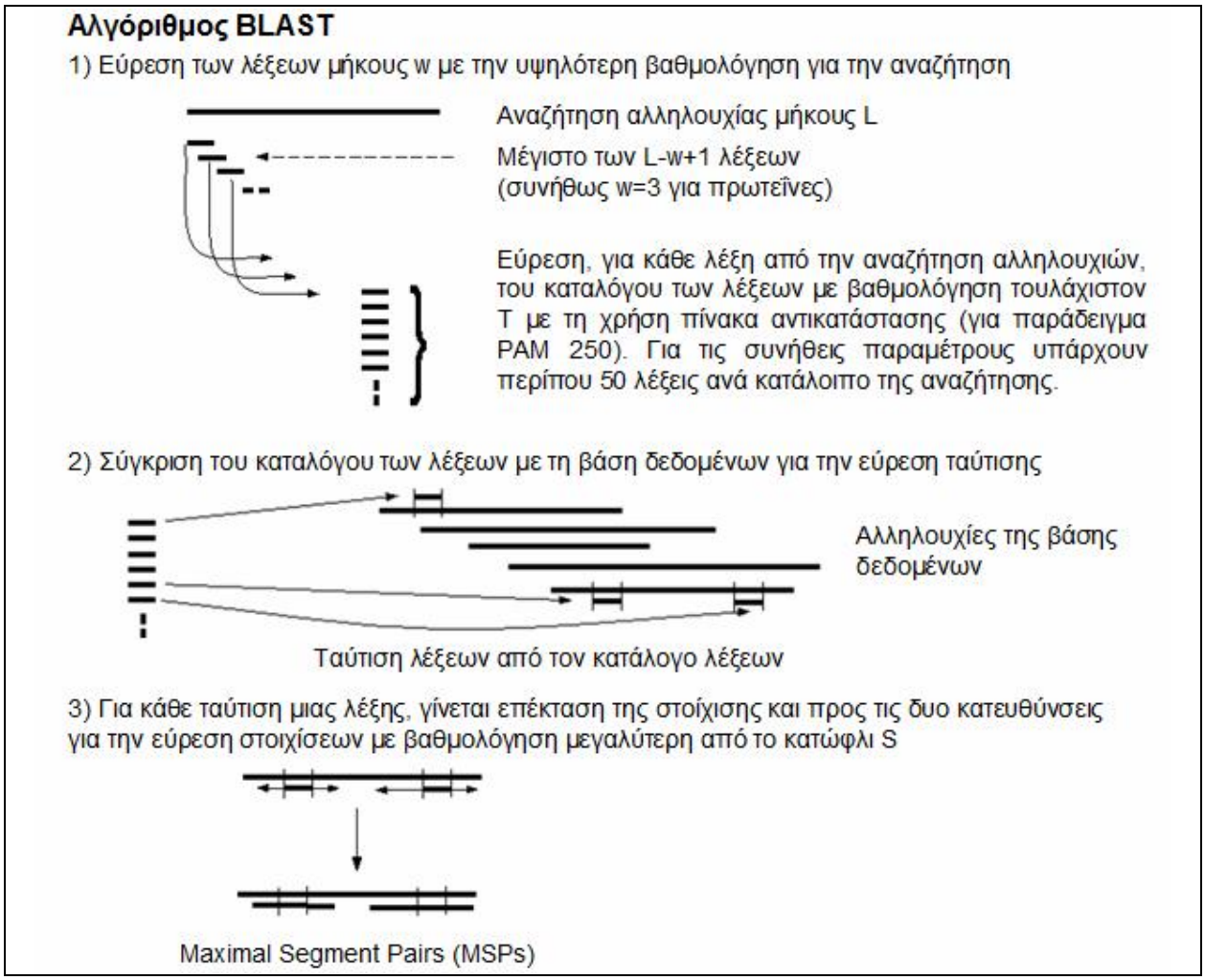


Εικόνα 3.13: *Διαγραμματική απεικόνιση του αλγόριθμου FASTA*

Η βασική ιδέα του **FASTA** (www.ebi.ac.uk/fasta33/), είναι να εντοπίσει κατά προσέγγιση τη διαγώνιο γύρω από την οποία βρίσκεται η στοίχιση, για να περιορίσει έτσι κατά πολύ το εύρος της αναζήτησης. Για το σκοπό αυτό χρησιμοποιεί τα εξής βήματα:

- Στην αρχή δημιουργείται ένα ευρετήριο με τις θέσεις όλων των *k*-tuples (λέξεων με μέγεθος *k*, τυπικό μήκος για αμινοξικές αλληλουχίες είναι το 1 ή 2) που υπάρχουν ταυτόχρονα και στις δύο αλληλουχίες.
- Από τη διαφορά των θέσεων τους στις δύο αλληλουχίες εντοπίζεται η διαγώνιος στην οποία βρίσκονται, οπότε στο επόμενο βήμα εντοπίζονται οι διαγώνιες με τα περισσότερα *k*-tuples.

- Ακολούθως, αυτές οι περιοχές ταύτισης συνενώνονται επιτρέποντας την εισαγωγή κενών με τον υπολογισμό της αντίστοιχης ποινής, και
- Τελικά πραγματοποιείται η διαδικασία πλήρους δυναμικού προγραμματισμού (με τον επιλεγμένο πίνακα αντικατάστασης), περιορισμένου όμως μόνο σε μια ζώνη γύρω από τις συγκεκριμένες διαγωνίους.



Εικόνα 3.14: *Διαγραμματική απεικόνιση του αλγόριθμου BLAST*

Η διαδικασία του **BLAST** (www.ncbi.nlm.nih.gov/BLAST/), μοιάζει στα αρχικά στάδια με αυτή το FASTA, αλλά είναι ακόμα πιο γρήγορη καθώς πολλές παραμέτρους τις έχει προϋπολογισμένες και αποφεύγει τον να στοιχίσει αλληλουχίες της βάσης δεδομένων που ο αλγόριθμος κρίνει ότι δεν έχουν σημαντική ομοιότητα:

- Η διαδικασία της σύγκρισης ξεκινά με την κατασκευή ενός καταλόγου όλων των λέξεων που θα ταίριαζαν με κάποια λέξη της άγνωστης αλληλουχίας και ξεπερνούν την τιμή κατωφλίου (προκαθορισμένη τιμή για πρωτεϊνικές αλληλουχίες T=13).
- Στη συνέχεια, ο αλγόριθμος αναζητά αυτές τις λέξεις στις αλληλουχίες της βάσης δεδομένων και κάθε φορά που εντοπίζει κάποια τέτοια ξεκινάει μια διαδικασία επέκτασης του 'ευρήματος' προς τις δύο κατευθύνσεις, όσο η βαθμολογία συνεχίζει και αυξάνει.
- Οι περιοχές μέγιστης βαθμολογίας που εντοπίζονται σε αυτό το στάδιο είναι οι υποψήφιες περιοχές ομοιότητας (HSPs, high scoring pairs).
- Από όλα τα HSPs αναφέρονται στα αποτελέσματα εκείνες οι περιοχές στις οποίες η βαθμολογία υπερβαίνει μια δεύτερη τιμή κατωφλίου S

- Τελικά, επιλέγονται να αναφερθούν εκείνες μόνο οι τοπικές ομοιότητες οι οποίες εμφανίζουν υψηλή στατιστική σημαντικότητα, ο προσδιορισμός της οποίας βασίζεται στο θεώρημα Karlin και Altschul.

Οι αρχικές εκδόσεις του BLAST, δεν επέτρεπαν την εισαγωγή κενών και έτσι ο αλγόριθμος ήταν ένα απλά εύχρηστο και γρήγορο εργαλείο για τον εντοπισμό όμοιων αλληλουχιών. Από τη $2^{η}$ έκδοση όμως του προγράμματος και μετά, προστέθηκε και η δυνατότητα εισαγωγής κενών με συνέπεια το BLAST να μπορεί να χρησιμοποιηθεί και σαν γενικό πρόγραμμα στοίχισης. Το BLAST γενικά, έχει κερδίσει την αποδοχή της κοινότητας, τόσο γιατί είναι ελεύθερα διαθέσιμο και συνδεδεμένο με τις βάσεις του NCBI, όσο και γιατί είναι ο πιο γρήγορος από τους αλγόριθμους στοίχισης.

Διαφορετικός είναι ο τρόπος υπολογισμού της στατιστικής σημαντικότητας των ευρημάτων. Ενώ το BLAST υπολογίζει τις παραμέτρους της κατανομής (K, λ) από προσομοιώσεις, που έχει πραγματοποιήσει από πριν και έχει αποθηκευμένες τις παραμέτρους, το FASTA τις υπολογίζει από όλες τις άλλες αλληλουχίες της βάσης δεδομένων και για αυτόν τον λόγο είναι και πιο αργό.

Οι νεότερες εκδόσεις του BLAST περιέχουν πολλές τροποποιήσεις που επιτρέπουν πιο ακριβείς υπολογισμούς με χρήση προφίλ και ειδικών ανά θέση πινάκων ομοιότητας (PSI-BLAST) (Altschul, et al., 1997), τεχνικές που θα περιγράψουμε σε επόμενο κεφάλαιο.

Το BLAST, χρησιμοποιεί επίσης και μια σειρά βελτιστοποιήσεων για τον ακριβή υπολογισμό της στατιστικής σημαντικότητας, πέραν της κλασικής θεωρίας των Karlin και Altschul. Έτσι, χρησιμοποιεί επιπλέον και μια διόρθωση στο μήκος των αλληλουχιών για να λάβει υπόψη το «αποτελεσματικό μήκος» (effective length) της αλληλουχίας και της βάσης δεδομένων. Συγκεκριμένα θέτει

$$m' = m - \frac{\log(Kmn)}{H} \text{ και } n' = n - \frac{\log(Kmn)}{H}$$

δηλαδή αναπροσαρμόζει το αποτελεσματικό μήκος της αλληλουχίας και της βάσης δεδομένων για να λάβει υπόψη το γεγονός ότι με αυτά τα μήκη και τον δεδομένο πίνακα (substitution matrix) δεν επιτρέπονται όλες οι στοιχίσεις. Πρακτικά, αυτό δίνει διαφορά όταν οι αλληλουχίες της βάσης είναι μικρές. Θεωρητικά, αν η βάση ήταν ολόκληρη μια τεράστια αλληλουχία, το αποτελεσματικό μήκος της αλληλουχίας και το πραγματικό, θα ήταν τα ίδια. Το H είναι η σχετική εντροπία του πίνακα για τη δεδομένη σύσταση και το μήκος των αλληλουχιών που συγκρίνονται.

$$H = \sum_{i=1}^{20}\sum_{j=1}^{20} q_{ij} s_{ij} \tag{3.67}$$

Η σχετική εντροπία εκφράζει το μέσο ποσό πληροφορίας που είναι διαθέσιμο για κάθε ζεύγος καταλοίπων που στοιχίζεται, και διαχωρίζει την προκύπτουσα στοίχιση από μια τυχαία στοίχιση που οφείλεται απλά στις συχνότητες υποβάθρου. Υψηλότερη τιμή της σχετικής εντροπίας συνεπάγεται εύκολο διαχωρισμό μεταξύ των συχνοτήτων στόχων και υποβάθρου. Η ποσότητα αυτή, εκφράζει την αναμενόμενη τιμή του σκορ της αντικατάστασης (expected substitution score per position), και σε αντίθεση με την παρόμοια ποσότητα στη σχέση (3.56), η οποία εκφράζει την αναμενόμενη τιμή του σκορ για κάθε θέση της στοίχισης (expected per position alignment score), πρέπει να είναι θετική.

Κάτι άλλο που πρέπει να τονιστεί είναι ότι, λόγω του γεγονότος ότι πολλές φορές χρησιμοποιούνται διαφορετικά σχήματα για το σκορ (gap penalties, mismatches), είναι αναγκαίο να αναφέρεται και μια αντικειμενική τιμή για το σκορ. Αυτό μπορεί να επιτευχθεί κανονικοποιώντας το σκορ με βάση το bit (Altschul, et al., 1990; Altschul, et al., 1997):

$$S_{bit} = \frac{\lambda S_{raw} - \log K}{\log 2} \tag{3.68}$$

όπου S_{raw}, είναι το σκορ που υπολογίστηκε με κάποιες συγκεκριμένες τιμές για κενά και διαφορές. Αντικαθιστώντας τώρα στην σχέση (3.47) θα έχουμε:

$$E(S_{bit}) = mn2^{-S_{bit}} \tag{3.69}$$

Η τελευταία σχέση, δίνει ακριβώς ίδιες τιμές με την (3.61) αλλά είναι πιο εύκολη στον υπολογισμό, όταν έχουμε σαν δεδομένο το bit Score.

Το FASTA ενσωματώνει επίσης ακριβέστερους τρόπους υπολογισμού της στατιστικής σημαντικότητας ενός ευρήματος όταν υπάρχουν κενά (Pearson, 1998). Πρέπει όμως να τονιστεί, ότι το BLAST σε αντίθεση με το FASTA δεν μπορεί να χρησιμοποιήσει κάθε μέθοδο, ειδικά αυτές που για τον υπολογισμό της σημαντικότητας χρησιμοποιούν τα αποτελέσματα της αναζήτησης στη βάση. Αυτό συμβαίνει

γιατί το BLAST για τις αλληλουχίες για τις οποίες δεν βρήκε κάποια ομοιότητα, δεν θα έχει υπολογισμένο κάποιο σκορ της στοίχισης.

Τέλος, πρέπει να σημειώσουμε, ότι τα πακέτα αυτά περιέχουν πολλές εκδόσεις που επιτρέπουν τη σύγκριση αλληλουχιών DNA με DNA, πρωτεΐνες με πρωτεΐνες, αλλά και εναλλακτικούς συνδυασμούς, δηλαδή τη σύγκριση ενός γονίδιου (DNA) με μια βάση δεδομένων πρωτεϊνών (μετάφραση του γονιδίου), τη σύγκριση μιας πρωτεΐνης με μια βάση αλληλουχιών DNA, και τέλος τη σύγκριση DNA με DNA αφού πρώτα αυτά μεταφραστούν (δηλαδή σύγκριση DNA-DNA στο πρωτεϊνικό επίπεδο). Σε γενικές γραμμές και το BLAST και το FASTA παρέχουν αποτελέσματα σχεδόν παραπλήσια με τους κλασικούς αλγόριθμους δυναμικού προγραμματισμού και το ποιο πακέτο θα χρησιμοποιηθεί από κάποιον είναι θέμα που εξαρτάται κυρίως από το πού αποσκοπεί η έρευνά του (ακρίβεια), από την ταχύτητα και από τις ανάγκες παραμετροποίησης που έχει (είδος ακολουθίας που συγκρίνεται, πλήθος των πινάκων του σκορ, ποινές για κενά κλπ).

Βιβλιογραφία

Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol, 219*(3), 555-565.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res, 25*(17), 3389-3402.

Arratia, R., Goldstein, L., & Gordon, L. (1989). Two moments suffice for Poisson approximation: The Chen-Stein method. *Ann. Probab., 17*, 9-25.

Arratia, R., Gordon, L. and Waterman, M. S. (1986). An extreme value theory for sequence matching. *Ann. Statist., 14*, 971-993.

Arratia, R., Gordon, L. and Waterman, M. S. (1990). The Erdos-Renyi law in distribution, for coin tossing and sequence matching. *Ann. Statist., 18*, 539-570.

Arratia, R., & Waterman, M. S. (1989). The Erdos-Renyi strong law for pattern matching with a given proportion of mismatches. *Ann. Probab., 17*, 1152-1169.

Arratia, R., & Waterman, M. S. (1994). A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab., 4*, 200-225.

Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Probab., 3*, 534-545.

Clote, P., & Backofen, R. (2000). *Computational Molecular Biology, an Introduction.*: John Wiley and Sons, Ltd. USA.

Davison, A. C. (1998). Extreme Values *Encyclopedia of Biostatistics*: John Wiley & Sons, Ltd.

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). A model of evolutionary change in Proteins. In M. Dayhoff (Ed.), *In Atlas of protein sequence and structure* (Vol. 5, Suppl. 3, pp. 345-352): National biomedical research foundation, Silver Spring, MD.

Durbin, R., Eddy, S., Krogh, A., & Mithison, G. (1998). *Biological sequence analysis, probabilistic models of proteins and nucleic acids.*: Cambridge University Press.

Erdos, P., & Renyi, A. (1970). On a new law of large numbers. *J. Anal. Math., 22*, 103-111.

Erdos, P., & Revesz, P. (1975). On the length of the longest head-run. *Topics in Inforrmation Theory. Colloquia Math. Soc. J. Bolyai, 16*, 219-228.

Galas, D. J., Eggert, M., & Waterman, M. S. (1985). Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from Escherichia coli. *J Mol Biol, 186*(1), 117-128.

Gonnet, G. H., Cohen, M. A., & Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science, 256*(5062), 1443-1445.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitutions matrices from protein blocks. *Proceedings of the National Academy of Sciences (USA), 89*, 10915-10919.

Karlin, S., & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the USA., 87*, 2264-2268.

Karlin, S., & Brendel, V. (1992). Chance and statistical significance in protein and DNA sequence analysis. *Science, 257*(5066), 39-49.

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol, 305*(3), 567-580.

Lipman, D. J., & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science, 227*(4693), 1435-1441.

Mott, R. (1992). Maximum likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bulletin of Mathematical Biology, 54*, 59-75.

Mott, R. (2000). Accurate formula for P-values of gapped local sequence and profile alignments. *J Mol Biol, 300*(3), 649-659.

Muller, T., Rahmann, S., & Rehmsmeier, M. (2001). Non-symmetric score matrices and the detection of homologous transmembrane proteins. *Bioinformatics, 17 Suppl 1*, S182-189.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol, 48*(3), 443-453.

Ng, P. C., Henikoff, J. G., & Henikoff, S. (2000). PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. *Bioinformatics, 16*(9), 760-766.

Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Science, 4*, 1145-1160.

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J Mol Biol, 276*(1), 71-84.

Pearson, W. R., & Wood, T. C. (2001). Statistical significance in biological sequence comparison. In D. J. Balding, M. Bishop & C. Cannings (Eds.), *In handbook of statistical genetics.* (pp. 39-65): John Wiley and Sons, Ltd. England.

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol, 147*(1), 195-197.

Vingron, M., & Waterman, M. S. (1994). Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J Mol Biol, 235*(1), 1-12.

Waterman, M. S. (1995). *Introduction to Computational Biology*: Chapman and Hall, London.

Waterman, M. S., Gordon, L., & Arratia, R. (1987). Phase transitions in sequence matches and nucleic acid structure. *Proceedings of the National Academy of Sciences of the USA., 84*, 1239-1243.

Waterman, M. S., & Vingron, M. (1994). Rapid and accurate estimates of statistical significance for sequence database searches. *Proceedings of the National Academy of Sciences of the USA., 91*, 4625-4628.

Waterman, M. S., & Vingron, M. (1994). Sequence comparison significance and Poisson approximation. *Statistical Science, 2*, 367-381.

Wilbur, W. J., & Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences of the USA., 80*, 726-730.

Wootton, J. C., & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers & chemistry, 17*(2), 149-163.

Ερωτήσεις

1) Στον ορισμό της σχετικής εντροπίας

$$H(\alpha, p) \equiv \alpha \log\left(\frac{\alpha}{p}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{1-p}\right)$$

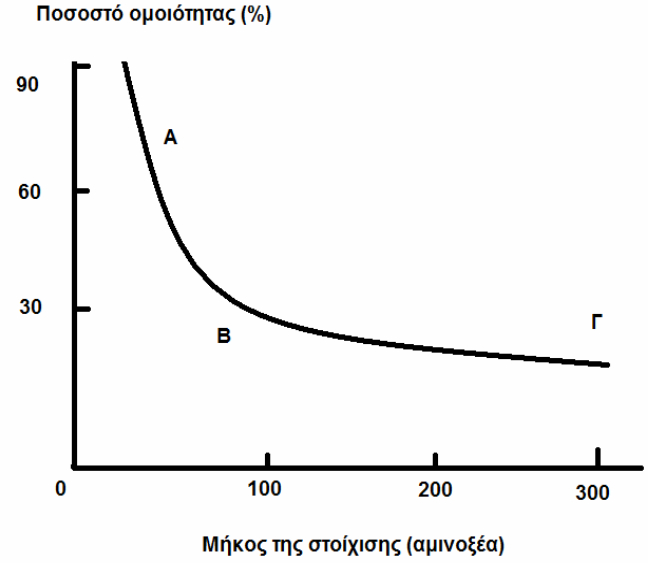
διερευνήστε τι θα συμβεί στην οριακή περίπτωση που το α=1. Τι επιπτώσεις θα έχει αυτή η λύση για τις σχέσεις (3.11) και (3.12);

2) Για τα δεδομένα του πίνακα 3.1, δείξτε ότι ισχύει η ισότητα:

$$E(s_k) = \sum p_k s_k = \sum p_k \log\left(\frac{a_k}{p_k}\right) < 0$$

Τι επιπτώσεις μπορεί να έχει αυτό για την πιθανή χρήση των δεδομένων του πίνακα;

3) Δίνεται η παρακάτω γραφική παράσταση στην οποία αντιστοιχίζονται οι τιμές του ποσοστού ομοιότητας (%) σε συνάρτηση του μήκους μιας κατά ζεύγη στοίχισης δύο πρωτεϊνικών αλληλουχιών.



Α) Τι αναπαριστά η καμπύλη; Ποια είναι η σημασία των δύο περιοχών στις οποίες διαχωρίζει το επίπεδο;
Β) Τι μπορείτε να πείτε για τα σημεία Α, Β και Γ;

4) Δύο αλληλουχίες μήκους 250 αμινοξέων στοιχίζονται με αλγόριθμο τοπικής στοίχισης και τον πίνακα PAM250, και προκύπτει η στοίχιση:

```
F W L E V E G N S M T A P T G
F W L D V Q G D S M T A P A G
```

Υπολογίστε το σκορ της στοίχισης και τη στατιστική σημαντικότητα, αν το K=0.09, και το λ=0.229. Ποιο είναι το bit-score αυτής της στοίχισης;

5) Εκτελούμε τοπική στοίχιση μιας αλληλουχίας Α μήκους 300 αμινοξέων με μια αλληλουχία Β μήκους 550 αμινοξέων. Η στοίχιση που προκύπτει δίνει μια ομοιότητα (similar residues) σε 61 από τα 166 στοιχισμένα κατάλοιπα, ενώ το Bit Score της στοίχισης είναι ίσο με 39.
Α) Ποιο είναι το E-value που προκύπτει από την παραπάνω στοίχιση και πως προκύπτει;
Β) Τι θα συνέβαινε αν το μήκος της στοίχισης ήταν το μισό με αντίστοιχη μείωση του Σκορ; Τι θα συνέβαινε αν το μήκος της στοίχισης ήταν το ίδιο και το Σκορ μειωνόταν στο μισό; Τι θα συνέβαινε αν το μήκος της στοίχισης ήταν το διπλάσιο και το Σκορ παρέμενε ίδιο;

Γ) Ποιο θα ήταν το E-value αν η ίδια στοίχιση είχε προκύψει πραγματοποιώντας αναζήτηση της αλληλουχίας Β έναντι μιας βάσης δεδομένων που περιέχει 500.000 αλληλουχίες με ίδιο μήκος με την Α;

6) Δίνεται τμήμα του αποτελέσματος από την αναζήτηση ομοιότητας με το BLAST μιας πρωτεΐνης έναντι της βάσης δεδομένων NR του NCBI.

Αλληλουχία Α

```
 Score = 34.3 bits (77),  Expect = XXXXX
 Identities = 28/85 (32%), Positives = 44/85 (51%), Gaps = 11/85 (12%)

Query  96   INDWASIYGVVGVGYGKFQTTEYPTY---KHDTSDYGFSYGAGLQ--FNPMENVALDFSY  150
            I++   I+G +G  YG+ +T+  P +     D S +G SYGAG++  FNP     L+  +
Sbjct  118  ISEQFDIFGKLGTTYGRTKTSGNPGFGVATGDDSGFGLSYGAGVRWAFNPQWAAVLE--W  175

Query  151  EQSRIR----SVDVGTWIAGVGYRF  171
            E+ R+       DV     GV YR+
Sbjct  176  ERHRLHFADGKSDVDMTTIGVQYRY  200
```

Αλληλουχία Β

```
 Score = 77.4 bits (189),  Expect = XXXXX
 Identities = 62/201 (30%), Positives = 101/201 (50%), Gaps = 32/201 (15%)

Query  1    MKKIACLSALAAVLAFTAGTSVAAT---STVTGGY--AQSDAQGQMNKMGGFNLKYRYEE  55
            M+K+      AA+    +G    A+   ST++ GY    ++  G  +++ G N+KYRYE
Sbjct  1    MRKLYAAILSAAICLAVSGAPAWASEHQSTLSAGYLHVSTNVPGS-DELNGINVKYRYEF  59

Query  56   DNSPLGVIGSFTY--------TEKSRTASSGDYNKNQYYGITAGPAYRINDWASIYGVVG  107
             ++ LG++ SF+Y        T  S T    D  +N+++ + AGP+ R+N+W S Y + G
Sbjct  60   TDT-LGMVTSFSYAGDKNRQLTHYSDTRWHEDSVRNRWFSVMAGPSVRVNEWFSAYAMAG  118

Query  108  VGYGKFQT--------TEYPTYKHDT---------SDYGFSYGAGLQFNPMENVALDFSY  150
            + Y +  T         T+     HD          S+   ++GAG+Q NP E+VA+D +Y
Sbjct  119  MAYSRVSTFSGDYLRVTDNKGKTHDVLTGSDDGRHSNTSLAWGAGVQVNPTESVAIDIAY  178

Query  151  EQSRIRSVDVGTWIAGVGYRF  171
            E S          +I GVGY+F
Sbjct  179  ECSGSGDWRTDGFIVGVGYKF  199
```

```
Gapped
Lambda     K      H
   0.267   0.0410    0.140
Number of Sequences: 4496249
Length of query: 171
Length of database: 1544746084
Length adjustment: 122
Effective length of query: 49
Effective length of database: 996203706
Effective search space: 48813981594
Effective search space used: 48813981594
```

Α) Υπολογίστε το E-value (Expectation) για τις δυο παραπάνω στοιχίσεις. Ποια συμπεράσματα βγάζετε για τη στατιστική σημαντικότητα των στοιχίσεων αυτών;
Β) Είναι τα συμπεράσματα αυτά σύμφωνα με τους εμπειρικούς κανόνες για την ομοιότητα δυο αλληλουχιών;
Γ) Τι θα συνέβαινε αν οι δύο παραπάνω στοιχίσεις είχαν προκύψει σε μία κατά ζεύγη στοίχιση και όχι σε μια αναζήτηση στη βάση δεδομένων;

Κεφάλαιο 4: Πολλαπλή Στοίχιση Ακολουθιών

Σύνοψη
Η πολλαπλή στοίχιση είναι μια διαδικασία με κεντρική σημασία στη σύγχρονη βιοπληροφορική. Πολλαπλές στοιχίσεις χρησιμοποιούνται για να εντοπιστούν τα συντηρημένα τμήματα σε μια ομάδα πρωτεϊνικών ακολουθιών και για να χαρακτηριστεί η αντίστοιχη οικογένεια, αλλά και για άλλες αναλύσεις, όπως η εκτίμηση φυλογενετικών σχέσεων και η υποβοήθηση της απόδοσης προγνωστικών αλγορίθμων. Το βασικό πρόβλημα της πολλαπλής στοίχισης είναι ότι δεν υπάρχει εύκολος τρόπος να βρεθεί μαθηματικά, η βέλτιστη λύση στο πρόβλημα, όπως έγινε στην περίπτωση της κατά ζεύγη στοίχισης. Στο κεφάλαιο αυτό θα μελετήσουμε τους κύριους αλγόριθμους πολλαπλής στοίχισης και τις αντίστοιχες υλοποιήσεις. Θα δούμε επίσης πώς αξιολογείται μια μέθοδος πολλαπλής στοίχισης, ποια εργαλεία υπάρχουν για την οπτικοποίηση και την επεξεργασία της, και τέλος, θα δούμε πρακτικές συμβουλές για μια καλή πολλαπλή στοίχιση.

Προαπαιτούμενη γνώση
Στο κεφάλαιο αυτό απαραίτητη είναι η γνώση των εννοιών του κεφαλαίου 3 (στοίχιση ακολουθιών).

4. Εισαγωγή

Αφού μελετήσαμε αναλυτικά την περίπτωση της στοίχισης δύο βιολογικών ακολουθιών, είναι εύλογο, ότι το επόμενο βήμα θα είναι η προσπάθεια ταυτόχρονης μελέτης περισσότερων από 2 ακολουθιών. Αυτό είναι το αντικείμενο της πολλαπλής στοίχισης το οποίο θα μελετήσουμε σε αυτό το κεφάλαιο. Το θέμα της πολλαπλής στοίχισης, είναι επίσης πολύ σημαντικό στη σύγχρονη υπολογιστική βιολογία και βιοπληροφορική. Οι χρήσεις μιας πολλαπλής στοίχισης, είναι πολλές, διαπερνούν όλο το φάσμα της υπολογιστικής ανάλυσης βιολογικών ακολουθιών, και μπορούμε να τις διακρίνουμε σε τρεις κατηγορίες.

Η προφανής χρήση μιας πολλαπλής στοίχισης αναφέρεται στην ταυτόχρονη μελέτη μιας ομάδας σχετιζόμενων ακολουθιών (συνήθως πρωτεϊνών) και στην προσπάθεια εύρεσης των κοινών χαρακτηριστικών τους. Αυτό, οδηγεί στο χαρακτηρισμό μιας "οικογένειας" πρωτεϊνών και στην αναγνώριση των περιοχών που είναι συντηρημένες. Με τη σειρά του αυτό, μπορεί να οδηγήσει σε χρήσιμες πληροφορίες για διάφορα δομικά ή λειτουργικά χαρακτηριστικά όλων των πρωτεϊνών της οικογένειας (π.χ. συντηρημένα στοιχεία δευτεροταγούς δομής, συντηρημένα κατάλοιπα τα οποία μπορεί να χαρακτηρίζουν το ενεργό κέντρο ενός ενζύμου). Η λογική συνέχεια όλων αυτών των διεργασιών, είναι να κατασκευαστεί με κάποιον μαθηματικό τρόπο, ένα μοντέλο που θα περιγράφει ολόκληρη την πολλαπλή στοίχιση και θα μπορεί να χρησιμοποιηθεί σε μια αναζήτηση σε βάση δεδομένων για ακολουθίες που ταιριάζουν με το μοντέλο πλέον, και όχι με μια συγκεκριμένη ακολουθία. Τέτοια παραδείγματα, είναι η κατασκευή μοντέλων κανονικών προτύπων ή μοτίβων (patterns), προφίλ αλλά και προφίλ Hidden Markov Models τα οποία θα περιγράψουμε σε επόμενα κεφάλαια. Είδαμε ήδη στο κεφάλαιο 2 ότι υπάρχουν μεγάλες βάσεις δεδομένων οι οποίες περιέχουν κατηγοριοποιήσεις των πρωτεϊνών σε οικογένειες, με τη χρήση τέτοιων μεθόδων (PROSITE, PFAM κ.α.).

Μια δεύτερη, πολύ σημαντική χρήση των πολλαπλών στοιχίσεων προκύπτει στην περίπτωση μελέτης των φυλογενετικών σχέσεων των βιολογικών ακολουθιών, και κατ' επέκταση των οργανισμών προέλευσής τους. Καθώς θεωρούμε ότι οι ακολουθίες έχουν όλες δημιουργηθεί μέσω της διαδικασίας της εξέλιξης από μεταλλάξεις παλαιότερων μορφών, είναι αναμενόμενο ότι οι ομοιότητες και οι διαφορές μιας ομάδας ακολουθιών, μπορούν να ανακατασκευάσουν ένα εξελικτικό δέντρο, το οποίο θα δείχνει τη σειρά με την οποία οι ακολουθίες αυτές εμφάνισαν απόκλιση από τον κοινό πρόγονο. Η διαδικασία αυτή, είναι πολύ σύνθετη και μπορεί να πραγματοποιηθεί με πολλούς διαφορετικούς τρόπους (όπως θα δούμε στο κεφάλαιο 6), αλλά το σημαντικότερο που πρέπει να θυμάται ο αναγνώστης είναι ότι σε κάθε περίπτωση, χρειάζεται μια καλής ποιότητας πολλαπλή στοίχιση σαν σημείο εκκίνησης.

Τέλος, μια πολύ σημαντική χρήση των πολλαπλών στοιχίσεων σχετίζεται με την υποβοήθηση (και μάλιστα σε μεγάλο βαθμό) των αλγορίθμων πρόγνωσης της δομής των πρωτεϊνών. Καθώς είναι γνωστό ότι η δομή συντηρείται περισσότερο από την ακολουθία, μια συντηρημένη περιοχή όπως αποτυπώνεται σε μια πολλαπλή στοίχιση μπορεί να προσφέρει μεγάλη βοήθεια στην προσπάθεια πρόγνωσης. Αυτό διαφέρει από την απλή αναγνώριση μοτίβων και τον εντοπισμό συντηρημένων περιοχών, τα οποία αναφέραμε πριν, αλλά επεκτείνεται και σε αυτοματοποιημένες χρήσεις των πολλαπλών στοιχίσεων, ως δεδομένα εισόδου σε

αλγορίθμους πρόγνωσης της δομής, ή άλλων χαρακτηριστικών των πρωτεϊνών. Το θέμα θα εξεταστεί σε επόμενο κεφάλαιο, καθώς υπάρχουν διαφορετικοί τρόποι με τους οποίους μπορεί να γίνει αυτή η χρήση. Για παράδειγμα, μπορεί η μέθοδος πρόγνωσης να εφαρμόζεται διαδοχικά σε όλες τις πρωτεΐνες της οικογένειας και στο τέλος να γίνεται "προβολή" των αποτελεσμάτων πάνω στην αρχική ακολουθία, ή, εναλλακτικά, οι μέθοδοι πρόγνωσης θα μπορούσαν να τροποποιηθούν έτσι ώστε να χρησιμοποιούν κατευθείαν κάποιο παράγωγο της πολλαπλής στοίχισης, όπως για παράδειγμα ένα προφίλ (profile).

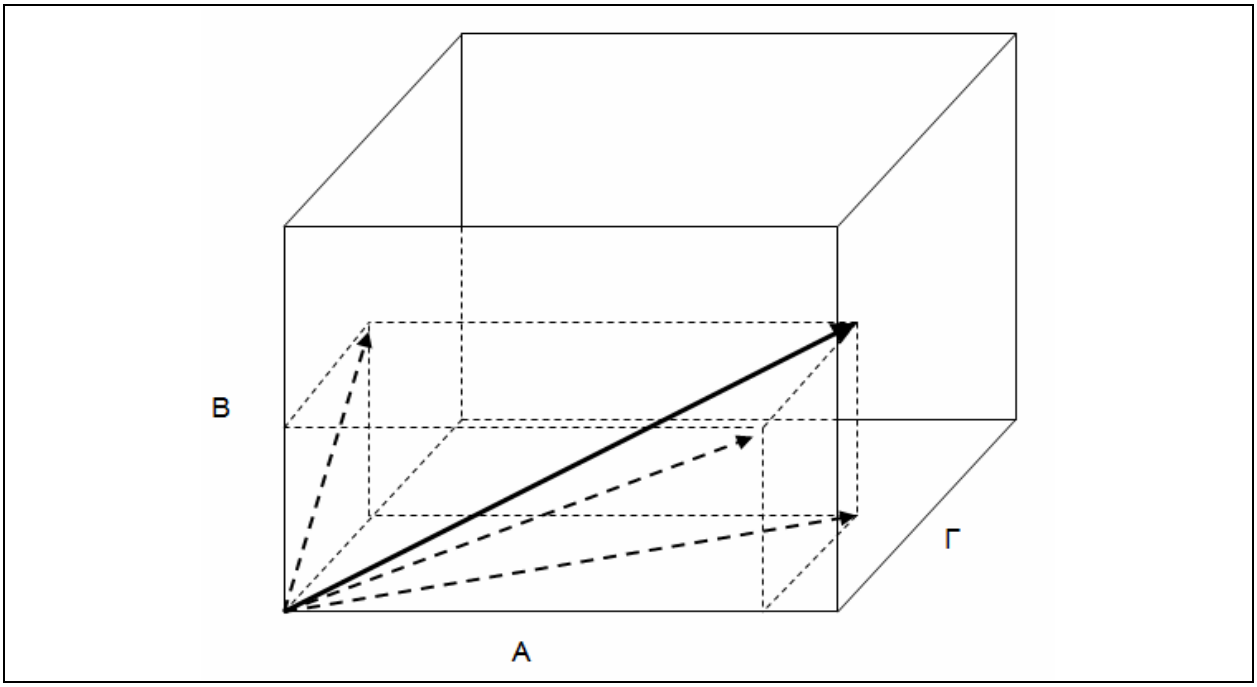
Στις επόμενες παραγράφους, θα παρουσιάσουμε τα βασικά θεωρητικά θέματα που εμπλέκονται στην πολλαπλή στοίχιση και τους βασικούς τύπους αλγορίθμων. Θα παρουσιαστεί επίσης το διαθέσιμο λογισμικό για το σκοπό αυτό, αλλά και οι τρόποι αξιολόγησης και οπτικοποίησης μιας πολλαπλής στοίχισης.

4.1. Πολλαπλή Στοίχιση – Δυναμικός Προγραμματισμός

Για να μελετήσουμε την πολλαπλή στοίχιση, είναι απαραίτητο πλέον να μελετήσουμε ταυτόχρονα περισσότερες από μία ακολουθίες. Έστω ότι έχουμε *r* ακολουθίες:

$$\begin{aligned} \mathrm{x}_1 &= x_{11}x_{12}\ldots x_{1n} \\ \mathrm{x}_2 &= x_{21}x_{22}\ldots x_{2n} \\ &\ldots\ldots\ldots \\ \mathrm{x}_\mathrm{r} &= x_{r1}x_{r2}\ldots x_{rn} \end{aligned} \tag{4.1}$$

ο πιο φυσικός τρόπος που μπορούμε να σκεφτούμε είναι να επεκτείνουμε τους αλγόριθμους δυναμικού προγραμματισμού του κεφαλαίου 3, στις *r* διαστάσεις. Όπως είχαμε δει στο κεφάλαιο 3, το πρόβλημα της στοίχισης δύο ακολουθιών ανάγεται στην εύρεση του βέλτιστου μονοπατιού στον πίνακα που αντιστοιχεί στο διάγραμμα σημείων. Κατ' αναλογία, όταν έχουμε τρεις ακολουθίες, η πολλαπλή στοίχιση αντιστοιχεί στην εύρεση του βέλτιστου μονοπατιού στον τρισδιάστατο πίνακα του οποίου οι έδρες είναι οι πίνακες που αντιστοιχούν στις κατά ζεύγη στοιχίσεις των ακολουθιών.



Εικόνα 4.1: *Σχηματική αναπαράσταση του πίνακα δυναμικού προγραμματισμού, για μια πολλαπλή στοίχιση 3 ακολουθιών Σε περίπτωση περισσότερων ακολουθιών η οπτικοποίηση γίνεται δυσκολότερη καθώς απαιτούνται περισσότερες διαστάσεις.*

Για να ξεκινήσουμε την πολλαπλή στοίχιση είναι αναγκαίο να ορίσουμε μια συνάρτηση για το σκορ:

$$S(m) = G + \sum_i S(m_i) \tag{4.2}$$

όπου m_i είναι η στήλη i της πολλαπλής στοίχισης m, $S(m_i)$ το σκορ της και G είναι μια συνάρτηση (απλή ή σύνθετη) για τα κενά. Για απλότητα, το κενό μπορεί να εισαχθεί και σαν ένα 5^o σύμβολο στις ακολουθίες (-), αν και στην πραγματικότητα αυτό δεν χρησιμοποιείται από τους περισσότερους σύγχρονους αλγόριθμους, γιατί θα ισοδυναμούσε με γραμμική εισαγωγή κενών. Παρ' όλα αυτά, για λόγους απλότητας, στις επόμενες ενότητες θα χρησιμοποιήσουμε αυτόν τον ορισμό, έτσι ώστε να μπορέσουμε να μελετήσουμε πιο εύκολα τους αλγόριθμους. Έτσι, θα έχουμε:

$$S(m) = \sum_i S(m_i) \tag{4.3}$$

Οι πιθανοί τρόποι να ορίσουμε το πολυδιάστατο σκορ, είναι πολλοί. Ο πρώτος τρόπος τον οποίο θα σκεφτόταν κάποιος, αναλογιζόμενος τους αλγορίθμους του προηγούμενου κεφαλαίου είναι να ορίσει ένα log-odds για τις r διαστάσεις:

$$S\left(m\right) = \sum_i S(m_i) = \sum_i \log\left(\frac{p_{x_{1i}x_{2i}\ldots x_{ri}}}{q_{x_{1i}}q_{x_{2i}}\ldots q_{x_{ri}}}\right) = \sum_i s(x_{1i}, x_{2i}, \ldots, x_{ri}) \tag{4.4}$$

Πρακτικά, αυτό είναι πολύ δύσκολο, γιατί θα σήμαινε ότι για παράδειγμα η δουλειά που έγινε για τους πίνακες ομοιότητας (PAM, BLOSUM κλπ), θα έπρεπε να έχει επαναληφθεί για κάθε πιθανό αριθμό ακολουθιών για τις οποίες θα επιχειρήσουμε μια πολλαπλή στοίχιση. Με άλλα λόγια, θα έπρεπε να υπάρχει προϋπολογισμένος ένας πίνακας για τις στοιχίσεις 3 ακολουθιών, άλλος πίνακας για τις στοιχίσεις 4 ακολουθιών, κ.ο.κ., κάτι που είναι πρακτικά αδύνατο.

Ένας άλλος τρόπος, θα ήταν αν κάναμε χρήση της έννοιας της εντροπίας την οποία συναντήσαμε στο προηγούμενο κεφάλαιο. Αν ονομάσουμε m_i^j το σύμβολο στην i στήλη της j ακολουθίας και $n_b(i)$ τον αριθμό των εμφανίσεων του συμβόλου b στη στήλη i, τότε η συνολική πιθανότητα της στήλης αυτής, θα είναι ίση με:

$$P\left(m_i\right) = \prod_{\forall b \in \Omega} p_b\left(i\right)^{n_b(i)} \tag{4.5}$$

όπου $p_s(i)$ θα είναι η πιθανότητα του συμβόλου s στη στήλη i, οποία θα δίνεται από τη σχέση:

$$p_b\left(i\right) = \frac{n_b\left(i\right)}{\sum_{\forall b' \in \Omega} n_{b'}\left(i\right)} \tag{4.6}$$

Τότε, αν πάρουμε το λογάριθμο, θα έχουμε:

$$S\left(m_i\right) = -\sum_{\forall b \in \Omega} n_b\left(i\right) \log p_b\left(i\right) \tag{4.7}$$

Αυτή η σχέση, είναι ξεκάθαρα ένα μέτρο εντροπίας, όπως το ορίσαμε στο προηγούμενο κεφάλαιο, με τη διαφορά ότι τώρα δεν αφορά ένα παράθυρο κατά μήκος της ακολουθίας, αλλά μία στήλη της πολλαπλής στοίχισης. Παρ' όλα αυτά, η ερμηνεία του είναι απλή και διαισθητική: μία στήλη η οποία είναι 100% συντηρημένη, θα έχει εντροπία ίση με το 0, αντίθετα, μια στήλη με τελείως τυχαία κατανομή συμβόλων, θα έχει μέγιστη εντροπία. Κατά συνέπεια, ένα καλό σκορ, θα ήταν αυτό το οποίο θα ελαχιστοποιούσε την εντροπία της πολλαπλής στοίχισης σε όλο το μήκος της (ή, εναλλακτικά, αυτό το οποίο θα μεγιστοποιούσε την πληροφορία). Στα παραπάνω, κάναμε σιωπηλά, δύο σημαντικές παραδοχές, οι οποίες είναι απαραίτητο να γίνουν, αλλά και απαραίτητο να διευκρινιστούν. Πρώτον, θεωρήσαμε τις r ακολουθίες ανεξάρτητες, -πράγμα το οποίο μπορεί να μην ισχύει ειδικά αν βρίσκονται εξελικτικά πολύ κοντά (σε αυτό θα επανέλθουμε). Δεύτερον, για να αθροίσουμε συνεισφορές του σκορ σε όλο το μήκος της στοίχισης, θεωρούμε και πάλι ότι οι στήλες είναι ανεξάρτητες μεταξύ τους. Σε κάθε περίπτωση, οι περισσότεροι αλγόριθμοι δεν χρησιμοποιούν το σύστημα του σκορ που περιγράψαμε παραπάνω. Παρ' όλα αυτά, η εντροπία χρησιμοποιείται για να αξιολογήσει το τελικό αποτέλεσμα μιας πολλαπλής στοίχισης ή για να συγκριθούν μεταξύ τους οι διάφοροι αλγόριθμοι όταν εφαρμοστούν στα ίδια δεδομένα.

Στους περισσότερους αλγόριθμους πολλαπλής στοίχισης, χρησιμοποιείται το λεγόμενο SP (Sum of Pairs) σκορ. Το σκορ αυτό ορίζεται για μία στήλη της στοίχισης ως:

$$SP\left(m_i\right) = \sum_{j<j'} s\left(m_i^j, m_i^{j'}\right) \tag{4.8}$$

Όπου οι τιμές της συνάρτησης s δίνονται από κάποιον από τους γνωστούς από την κατά ζεύγη στοίχιση ακολουθιών, αλγορίθμους. Για τη συνολική στοίχιση, θα μπορούσε να γραφτεί και ως εξής:

$$SP(m) = \sum_i \sum_{j<j'} s\left(m_i^j, m_i^{j'}\right) = \sum_i \left\{ \log\left(\frac{p_{x_{1i}x_{2i}}}{q_{x_i} q_{x_{2i}}}\right) + ... + \log\left(\frac{p_{x_{(r-1)i}x_{ri}}}{q_{x_{(r-1)i}} q_{x_{ri}}}\right) \right\} \quad (4.9)$$

καθώς είναι το άθροισμα των σκορ για όλες τις ανά 2 συγκρίσεις των *r* ακολουθιών. Η Μέθοδος αυτή, είναι πολύ βολική, αλλά έχει το μειονέκτημα ότι δεν έχει καλές μαθηματικές ιδιότητες. Το άθροισμα των ανά δύο σκορ, δεν έχει κάποια φυσική ερμηνεία, και οδηγεί σε κάποια παράδοξα. Για παράδειγμα, υπάρχουν περιπτώσεις, στις οποίες μια εισαγωγή ενός διαφορετικού συμβόλου σε μια κατά τα άλλα τέλεια στοίχιση (π.χ. 10 ή 20 όμοιες ακολουθίες), οδηγεί σε μεγαλύτερη μείωση του σκορ στη στοίχιση με τις περισσότερες, σε σχέση με τη στοίχιση με τις λιγότερες ακολουθίες (Durbin, Eddy, Krogh, & Mithison, 1998). Αυτό είναι αντίθετο με τη διαίσθηση, γιατί θα περιμέναμε η μείωση του σκορ να είναι μικρότερη, λ.χ. στην περίπτωση που έχουμε 19/20 ακολουθίες ίδιες, παρά αν είχαμε 9/10, αλλά εξηγείται αν παρατηρήσουμε ότι η μία εισαγωγή του διαφορετικού συμβόλου θα επηρεάσει περισσότερους όρους στο άθροισμα στην πρώτη περίπτωση.

Παρ' όλα αυτά, αυτή είναι η μέθοδος που χρησιμοποιούν οι περισσότεροι αλγόριθμοι, κυρίως για την υπολογιστική ευκολία που προσφέρει αλλά και λόγω του ότι μπορεί και ενσωματώνει εύκολα την πληροφορία των πινάκων ομοιότητας οι οποίοι είναι ήδη διαθέσιμοι. Δεν πρέπει να ξεχνάμε, ότι στην περίπτωση των πρωτεϊνών, είναι σχεδόν αδύνατο να βρούμε μια στοίχιση πολλών ακολουθιών (π.χ. >50) οι οποίες να έχουν 100% συντηρημένες παρά μόνο λίγες θέσεις. Αυτό συμβαίνει, γιατί πολλές φορές αμινοξέα με παρόμοιες φυσικοχημικές ιδιότητες αντικαθιστούν κάποια άλλα στην εξέλιξη, χωρίς να επηρεάσουν τη δομή και τη λειτουργία της πρωτεΐνης. Ένα μέτρο σαν την εντροπία, θα «έχανε» αυτή την πληροφορία, η οποία όμως εντοπίζεται με χρήση των πινάκων ομοιότητας. Αξίζει να αναφερθεί, ότι το ίδιο ισχύει ακόμα και στην περίπτωση των περιοχών χαμηλής πολυπλοκότητας που είδαμε στο προηγούμενο κεφάλαιο (για παράδειγμα μπορεί να έχουμε επαναλήψεις παρόμοιων αμινοξέων), και έχουν προταθεί και μέτρα πολυπλοκότητας που λαμβάνουν υπόψη τους πίνακες ομοιότητας αμινοξέων.

Αφού έχουμε δει τα βασικά για το σκορ μιας πολλαπλής στοίχισης, ας δούμε πώς διαμορφώνεται ένας αλγόριθμος δυναμικού προγραμματισμού για το σκοπό αυτό. Αν ονομάσουμε $a_{i1,i2,\ ...,iN}$ το μέγιστο σκορ μιας στοίχισης μέχρι και τις υποακολουθίες που τελειώνουν στο $x_{i1}^1, x_{i2}^2, ..., x_{iN}^N$, τότε μια απευθείας επέκταση των αλγορίθμων του κεφαλαίου 3, δίνει:

$$a_{i1}, a_{i2}, ..., a_{in} = \max_{\Delta_1+...+\Delta_n} \begin{cases} a_{i1-1,i2-1,...,in-1} + S\left(x_{i1}^1, x_{i2}^2, ..., x_{in}^n\right) \\ a_{i1,i2-1,...,in-1} + S\left(-, x_{i2}^2, ..., x_{in}^n\right) \\ ... \\ a_{i1-1,i2-1,...,in} + S\left(x_{i1}^1, x_{i2}^2, ..., -\right) \\ a_{i1,i2,i3-1,...,in-1} + S\left(-, -, ..., x_{in}^n\right) \\ ... \end{cases} \quad (4.10)$$

Σε αυτή την περίπτωση το κενό το αντιμετωπίζουμε όπως είπαμε λόγω ευκολίας, σαν ένα πέμπτο σύμβολο (-). Στη σχέση (4.10) στο δεξί σκέλος επιτρέπονται όλοι οι συνδυασμοί των κενών, εκτός από αυτόν στον οποίο όλες οι θέσεις έχουν κενό (2^N-1 συνολικοί συνδυασμοί). Ένας πιο συμπυκνωμένος τρόπος να γραφτεί ο αλγόριθμος, θα είναι (Durbin, et al., 1998; Waterman, 1995):

$$a_{i1}, a_{i2}, ..., a_{in} = \max_{\Delta_1+...+\Delta_n>0} \left\{ a_{i1-\Delta_1, i2-\Delta_2, ..., in-\Delta_n} + S\left(\Delta_1 x_{i1}^1, \Delta_2 x_{i2}^2, ..., \Delta_n x_{in}^n\right) \right\} \quad (4.11)$$

όπου Δ είναι στοιχεία μιας συνάρτησης για την οποία ισχύει:

$$\Delta_i x = \begin{cases} (x), \text{ αν } \Delta_i = 1 \\ (-), \text{ αν } \Delta_i = 0 \end{cases} \quad (4.12)$$

Όπως είναι φανερό, ο αλγόριθμος αυτός, αν έχουμε *r* ακολουθίες με *n* νουκλεοτίδια η κάθε μια (ή, αμινοξικά κατάλοιπα αν μιλάμε για πρωτεΐνες), απαιτεί χρόνο της τάξης του $O(n^r 2^r)$ και χώρο στη μνήμη $O(n^r)$. Πρακτικά λοιπόν, ένας τέτοιος αλγόριθμος θα είχε μεγάλη πολυπλοκότητα και θα ήταν ιδιαίτερα αργός, δηλαδή θα χρειαζόταν απαγορευτικό χρόνο ακόμα και για λίγες σχετικά ακολουθίες και κατά συνέπεια θα πρέπει να αναζητηθούν τρόποι να περιοριστούν οι απαιτήσεις αυτές, περιορίζοντας το εύρος της

αναζήτησης. Έναν τέτοιο αλγόριθμο πρότειναν οι Carrilo και Lipman (Carrillo & Lipman, 1988). Ο αλγόριθμος αυτός, βρίσκει ένα κάτω φράγμα στο σκορ για κάθε ζεύγος στοιχίσεων μεταξύ των ακολουθιών, και στη συνέχεια, ελέγχει στην πολλαπλή στοίχιση μόνο τις περιοχές αυτές που έχουν σκορ μεγαλύτερο από την τιμή αυτή. Οσο πιο υψηλή τιμή έχει αυτό το φράγμα, τόσο πιο γρήγορος θα είναι ο αλγόριθμος. Πρακτικά, για να βρεθεί αυτή η τιμή θα πρέπει να χρησιμοποιηθεί πρώτα ένας γρήγορος ευριστικός αλγόριθμος πολλαπλής στοίχισης (όπως αυτοί της προοδευτικής πολλαπλής στοίχισης που θα περιγραφούν παρακάτω). Ο αλγόριθμος των Carrilo και Lipman έχει υλοποιηθεί στο γνωστό πρόγραμμα **MSA** (Lipman, Altschul, & Kececioglu, 1989), διαθέσιμο στη διεύθυνση http://xylian.igh.cnrs.fr/msa/msa.html το οποίο όμως, ακόμα και έτσι, πρακτικά είναι ικανό να στοιχίσει μόνες μερικές ακολουθίες πρωτεϊνών.

Όπως είναι φανερό, στην περίπτωση της πολλαπλής στοίχισης, η ανάγκη να αναζητήσουμε ευριστικούς αλγόριθμους είναι ακόμα μεγαλύτερη σε σχέση με την περίπτωση της κατά ζεύγη στοίχισης. Στις επόμενες παραγράφους, θα περιγράψουμε τις βασικές κατηγορίες ευριστικών αλγορίθμων για πολλαπλή στοίχιση, και τις παραλλαγές τους όπως αυτές χρησιμοποιούνται στα σύγχρονα εργαλεία λογισμικού.

4.2. Προοδευτική πολλαπλή στοίχιση

Η πιο γνωστή ευριστική (heuristic) μέθοδος που χρησιμοποιείται για πολλαπλή στοίχιση, είναι η λεγόμενη progressive multiple alignment method (προοδευτική πολλαπλή στοίχιση). Κατά τη μέθοδο αυτή η στοίχιση των ακολουθιών γίνεται προοδευτικά ξεκινώντας από δυο ακολουθίες (συνήθως αυτές με την μεγαλύτερη ομοιότητα), και σταδιακά προστίθενται στην στοίχιση μια-μια, οι υπόλοιπες ακολουθίες. Αν και υπάρχουν πολλές παραλλαγές ήδη από τις αρχές της δεκαετίας του 1980, η γενική μέθοδος της προοδευτικής πολλαπλής στοίχισης όπως διατυπώθηκε από τους Feng και Doolittle το 1987 (Feng & Doolittle, 1987) περιλαμβάνει τα παρακάτω κύρια βήματα:

- Αρχικές κατά ζεύγη στοιχίσεις όλων των ακολουθιών
- Με βάση αυτές τις στοιχίσεις, κατασκευή πίνακα αποστάσεων και ενός δέντρου οδηγού (guide tree)
- Προοδευτική στοίχιση των πιο όμοιων ακολουθιών μεταξύ τους, μέχρι τέλους

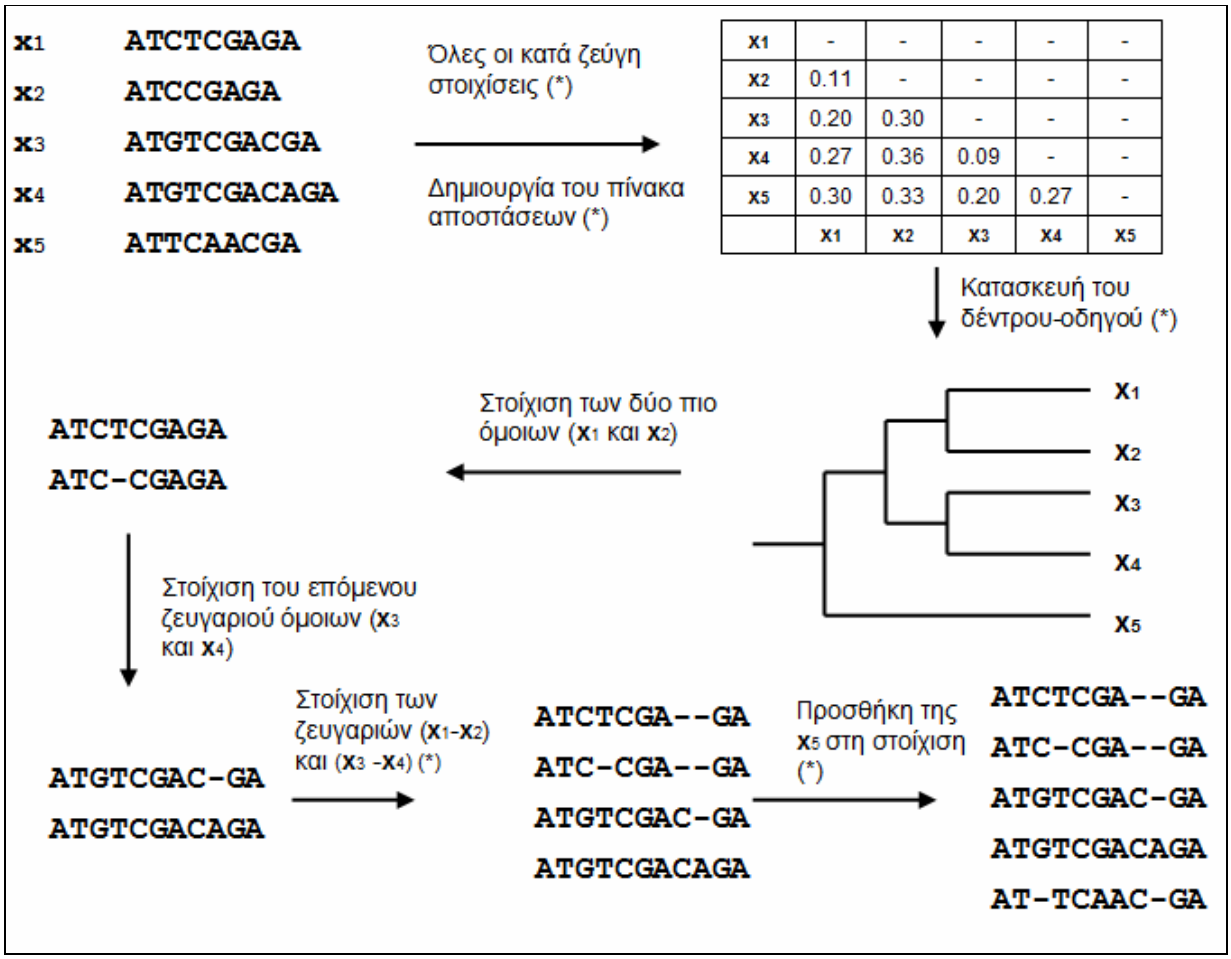
Όπως είναι εμφανές, τα βήματα αυτά, θα μπορούσαν να υλοποιηθούν με διαφορετικούς τρόπους. Για παράδειγμα, οι κατά ζεύγη στοιχίσεις θα μπορούσαν να γίνουν με δυναμικό προγραμματισμό ή με ευριστική μέθοδο (BLAST, FASTA). Ο πίνακας των αποστάσεων θα μπορούσε να οριστεί με τελείως διαφορετικά κριτήρια, ενώ και το δέντρο οδηγός θα μπορούσε να κατασκευαστεί με μια πλειάδα αλγορίθμων ομαδοποίησης (clustering). Τέλος, υπάρχει και το αλγοριθμικό θέμα σχετικά με το πώς θα προχωρήσει η στοίχιση μιας ακολουθίας με μια ήδη υπάρχουσα στοίχιση, ή, μιας στοίχισης με μια άλλη στοίχιση. Στις πιο παλιές μεθόδους, υπήρχαν και άλλες πιο βασικές διαφορές, όπως για παράδειγμα η ίδια η ύπαρξη του δέντρου, αλλά εδώ θα μελετήσουμε κυρίως παραλλαγές πάνω σε αυτή τη μέθοδο. Η μέθοδος των Feng και Doolittle (Feng & Doolittle, 1987), ήταν όπως είπαμε μια από τις πρώτες τέτοιες μεθόδους, και έκανε αρχικά τις στοιχίσεις με αλγόριθμο δυναμικού προγραμματισμού (ολικής στοίχισης). Στη συνέχεια, υπολόγιζε τις αποστάσεις, από τα σκορ των στοιχίσεων χρησιμοποιώντας τον τύπο:

$$D = -\log S = \log \frac{S_{obs} - S_{rand}}{S_{\max} - S_{rand}} \tag{4.13}$$

Το S_{obs} είναι το πραγματικό σκορ για τη στοίχιση των δύο ακολουθιών, όπως προέκυψε από τον αλγόριθμο. Το S_{max} είναι το θεωρητικό μέγιστο που θα μπορούσε να προκύψει από τη στοίχιση, αν στοιχίζαμε οποιαδήποτε από τις δύο ακολουθίες με τον εαυτό της, ενώ το S_{rand} είναι το αναμενόμενο σκορ από μια τέτοια στοίχιση ακολουθιών οι οποίες δεν είχαν καμία σχέση μεταξύ τους. Θα μπορούσε να προκύψει με κάποια προσομοίωση όπως περιγράψαμε στο προηγούμενο κεφάλαιο (με shuffling), αλλά οι Feng και Doolittle έδωσαν έναν προσεγγιστικό υπολογισμό. Όπως είναι φανερό, ο τρόπος υπολογισμού του κλάσματος δίνει περίπου το ποσοστό ομοιότητας των ακολουθιών, οπότε η προσθήκη του –log κάνει το μέτρο περίπου γραμμικό και δίνει μεγαλύτερες τιμές (μεγαλύτερη απόσταση) σε ζευγάρια με μικρή ομοιότητα.

Αφού έχουν υπολογιστεί οι αποστάσεις, ο αλγόριθμος κατασκευάζει ένα δέντρο με τη χρήση του αλγορίθμου των Fitch και Margoliash (Fitch & Margoliash, 1967). Ο αλγόριθμος αυτός είναι από τους πιο γρήγορους αλγόριθμους ομαδοποίησης, και προτάθηκε αρχικά για την κατασκευή φυλογενετικών δέντρων. Γενικά, το δέντρο-οδηγός της προοδευτικής πολλαπλής στοίχισης, έχει πολλά κοινά με τα φυλογενετικά δέντρα τα οποία θα εξετάσουμε στο κεφάλαιο 6, αλλά πρέπει να σημειώσουμε, ότι δεν είναι το ίδιο ένα τέτοιο δέντρο, τουλάχιστον όχι με την αυστηρή έννοια. Ο σκοπός εδώ είναι να παραχθεί γρήγορα μια ομαδοποίηση

η οποία θα κατευθύνει τη στοίχιση, και όχι να βρεθεί ο κοινός πρόγονος των ακολουθιών και ο χρόνος απόκλισης της κάθε μίας.

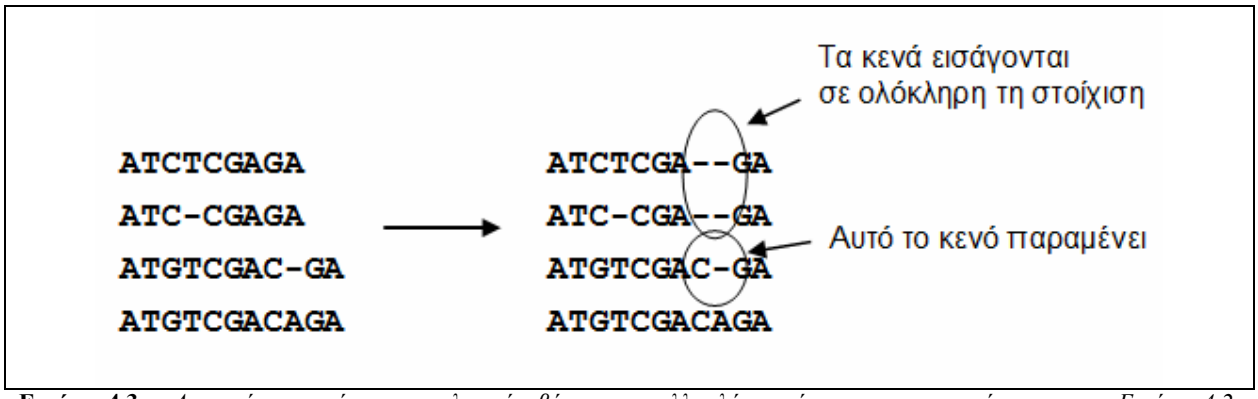


Εικόνα 4.2: *Παράδειγμα προοδευτικής πολλαπλής στοίχισης 5 ακολουθιών (Duret & Abdeddaim, 2000). Με (*) σημειώνονται τα σημεία στα οποία θα μπορούσε να υπάρξει διαφοροποίηση μεταξύ των αλγορίθμων. Όταν σχηματιστεί η στοίχιση των ακολουθιών x_1 και x_2 από τη μια μεριά, και των x_3 και x_4 από την άλλη, στο επόμενο βήμα, το δέντρο υπαγορεύει ότι οι δύο αυτές στοιχίσεις πρέπει να ενωθούν, καθώς οι τέσσερις ακολουθίες που περιέχονται έχουν μεγαλύτερες ομοιότητες μεταξύ τους παρά με την x_5. Στο σημείο αυτό, τη στοίχιση των δύο στοιχίσεων, την κατευθύνει απόλυτα το ζευγάρι με τη μεγαλύτερη ομοιότητα (x_1-x_3). Το ίδιο συμβαίνει και στο επόμενο βήμα, το οποίο καθορίζεται απόλυτα από τη στοίχιση x_3-x_5.*

Τέλος, στο επόμενο βήμα, οι ακολουθίες στοιχίζονται σταδιακά χρησιμοποιώντας την πληροφορία του δέντρου, ξεκινώντας από τις πιο όμοιες. Το βασικό σημείο εδώ, είναι ότι μια ακολουθία (ή μια στοίχιση) προστίθεται σε μια άλλη πολλαπλή στοίχιση, με βάση το ζευγάρι των ακολουθιών που είχε το μεγαλύτερο σκορ (τη μικρότερη απόσταση). Με τον τρόπο αυτό, μια στοίχιση όταν δημιουργηθεί, δεν αλλάζει ξανά, και το μόνο που μπορεί να συμβεί είναι να προστεθούν κενά. Η μέθοδος βέβαια αυτή, έχει και ένα άλλο μειονέκτημα: τη στοίχιση μιας στοίχισης με μια άλλη στοίχιση, την καθοδηγεί απόλυτα το ζευγάρι το οποίο εμφανίζει τη μέγιστη ομοιότητα. Στο παράδειγμα στην Εικόνα 4.2 βλέπουμε πώς λειτουργεί η μέθοδος για μια στοίχιση 5 ακολουθιών.

Η μέθοδος αυτή, δεν είναι πάντα αποτελεσματική, καθώς το ζευγάρι με τη μεγαλύτερη ομοιότητα μπορεί να προσδώσει συστηματικό σφάλμα (bias) στην πολλαπλή στοίχιση, και όπως είπαμε τα λάθη στην προοδευτική πολλαπλή στοίχιση δεν διορθώνονται σε κάποιο επόμενο βήμα. Για το λόγο αυτό, θα ήταν επιθυμητό, όλες οι ακολουθίες της πολλαπλής στοίχισης να παίζουν κάποιο ρόλο στο πώς μια άλλη ακολουθία θα προστεθεί στη στοίχιση. Ένας τρόπος για να γίνει αυτό, θα ήταν να δημιουργηθεί από κάθε μια πολλαπλή στοίχιση, μια συναινετική ακολουθία (consensus), δηλαδή μια «ψεύτικη» ακολουθία στην οποία

κάθε σύμβολο θα ήταν αυτό το οποίο εμφανίζεται με μεγαλύτερο ποσοστό στην πολλαπλή στοίχιση. Αυτή θα ήταν μια εύκολη λύση, αλλά και πάλι αδυνατεί να λάβει υπόψη όλες τις ακολουθίες. Φανταστείτε για παράδειγμα μια στήλη στην οποία υπάρχουν 2 A, 1 T, 1 G, και 1 C. Το A είναι φυσικά, το σύμβολο με τη μεγαλύτερη πιθανότητα, αλλά και πάλι η πληροφορία για το 60% των άλλων συμβόλων της συγκεκριμένης θέσης, δεν χρησιμοποιείται.



Εικόνα 4.3: *Λεπτομέρεια από το προτελευταίο βήμα της πολλαπλής στοίχισης που περιγράφεται στην Εικόνα 4.2. Βλέπουμε τη στοίχιση των ακολουθιών x_1 και x_2 από τη μια μεριά, και των x_3 και x_4 από την άλλη. Στο σημείο αυτό, τη στοίχιση των δύο στοιχίσεων, την κατευθύνει απόλυτα το ζευγάρι με τη μεγαλύτερη ομοιότητα (x_1-x_3). Προσέξτε ότι το κενό που υπήρχε στη στοίχιση των x_3 και x_4 παραμένει, ενώ τα κενά που εισάγονται στις x_1 και x_2, εισάγονται ταυτόχρονα και στις δύο.*

Μια καλύτερη λύση, είναι το λεγόμενο *profile alignment* (στοίχιση προφίλ), το οποίο μετράει τη σχετική συνεισφορά όλων των ακολουθιών της κάθε στοίχισης και τελικά πραγματοποιεί την στοίχιση λαμβάνοντας υπόψη όλες τις ακολουθίες. Τα μαθηματικά της μεθόδου είναι πολύπλοκα, αλλά μπορούν να απλοποιηθούν αν θεωρήσουμε, όπως και παραπάνω, το κενό σαν ένα πέμπτο σύμβολο (-), οπότε θα έχουμε και γραμμική ποινή για τα κενά. Τότε, με τη χρήση του SP σκορ, μπορούμε να σκοράρουμε όλες τις ακολουθίες της μιας στοίχισης με όλες τις ακολουθίες της άλλης. Για απλότητα, θεωρούμε επίσης ότι στη μία στοίχιση περιέχονται οι ακολουθίες από 1 έως n, ενώ στην άλλη, οι ακολουθίες από n+1 έως N. Σε αυτή την περίπτωση η σχέση (4.8) γίνεται:

$$\begin{aligned}\sum_i SP\left(m_i\right) &= \sum_i \sum_{j<j'} s\left(m_i^j, m_i^{j'}\right) \\ &= \sum_i \sum_{j<j'\le n} s\left(m_i^j, m_i^{j'}\right) + \sum_i \sum_{n<j<j'\le N} s\left(m_i^j, m_i^{j'}\right) + \sum_i \sum_{j\le n, n<j'\le N} s\left(m_i^j, m_i^{j'}\right)\end{aligned} \tag{4.14}$$

Τα δύο πρώτα αθροίσματα στο δεξί σκέλος της σχέσης (4.14) δεν αλλάζουν καθώς κάθε μια από τις στοιχίσεις παραμένει σταθερή, οπότε αυτό που μένει να βελτιστοποιηθεί είναι το τελευταίο άθροισμα, το οποίο περιέχει τις συνεισφορές από τις χιαστί συγκρίσεις των ακολουθιών των δύο στοιχίσεων. Η βελτιστοποίηση, γίνεται με τον κλασικό πίνακα του δυναμικού προγραμματισμού που συναντήσαμε στο κεφάλαιο 3 (Εικόνα 4.4). Στην πράξη, η μέθοδος αυτή είναι η πιο αποτελεσματική και χρησιμοποιείται από τους περισσότερους σύγχρονους αλγόριθμους. Παρ' όλα αυτά, υπάρχουν πάρα πολλές επί μέρους διαφοροποιήσεις, ανάλογα με το σύστημα του σκορ και το πώς ο κάθε αλγόριθμος χειρίζεται τα κενά (Edgar & Sjolander, 2004; Wang & Dunbrack, 2004).

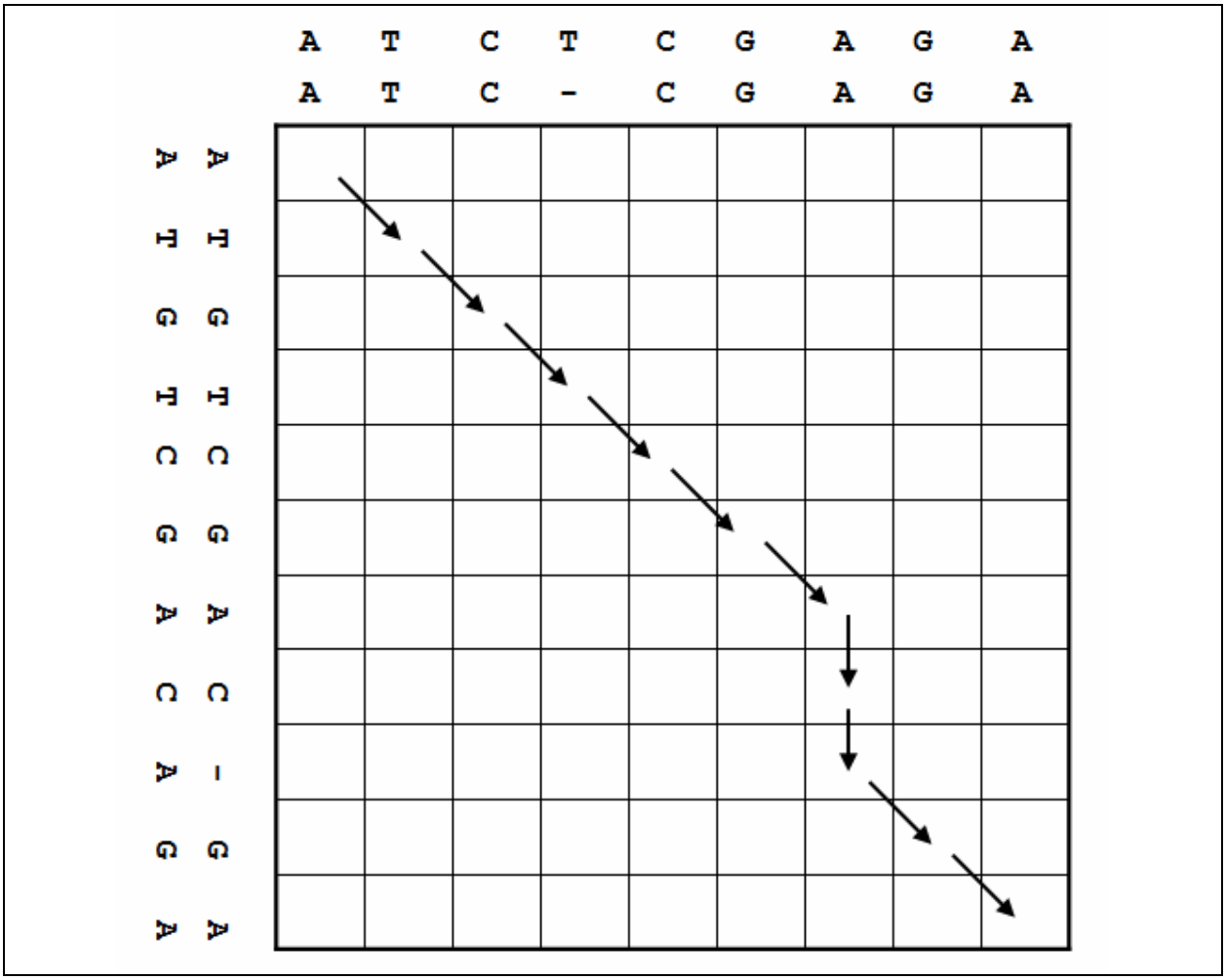
Ίσως το πιο γνωστό και περισσότερο χρησιμοποιημένο, πρόγραμμα πολλαπλής στοίχισης το **CLUSTALW** (Thompson, Higgins, & Gibson, 1994) ήταν το πρώτο που χρησιμοποίησε profile alignment με την προοδευτική πολλαπλή στοίχιση. Το clustal ξεκίνησε από την έκδοση CLUSTALV (Higgins, Bleasby, & Fuchs, 1992) ενώ στην πορεία αναπτύχθηκε η έκδοση CLUSTALW αλλά και η έκδοση που υποστήριζε γραφικά, η CLUSTALX (Thompson, Gibson, & Higgins, 2002). Ο βασικός αλγόριθμος βέβαια, είναι ίδιος και εξελίσσεται με τα χρόνια ενσωματώνοντας πολλές ευριστικές τεχνικές οι οποίες έχουν προκύψει από εμπειρική παρατήρηση και οι οποίες προσδίδουν μεγαλύτερη σταθερότητα και αξιοπιστία στη μέθοδο. Η μέθοδος είναι διαθέσιμη στη διεύθυνση www.ebi.ac.uk/clustalw/.Τα βασικά σημεία της, είναι:

- Στις αρχικές εκδόσεις της μεθόδου, ο αλγόριθμος έκανε τις στοιχίσεις κατά ζεύγη με έναν ευριστικό αλγόριθμο (FASTA), με αποτέλεσμα να είναι ιδιαίτερα γρήγορος. Σε κατοπινές

εκδόσεις δίνει τη δυνατότητα, εναλλακτικά, να χρησιμοποιηθεί ένας αλγόριθμος δυναμικού προγραμματισμού, για καλύτερα αποτελέσματα.

- Οι αποστάσεις υπολογίζονται απευθείας από την επί τοις εκατό ομοιότητα των ακολουθιών x (*D*=1-x/100) ενώ το δέντρο-οδηγός κατασκευάζεται με την ιδιαίτερα αποτελεσματική και σταθερή, μέθοδο Neighbor-Joining (ένωση γειτόνων) (Saitou & Nei, 1987). Η μέθοδος αυτή είναι μια μέθοδος ομαδοποίησης που προτάθηκε αρχικά για χρήση σε φυλογενετικά δέντρα. Θα την αναλύσουμε στο κεφάλαιο 6.
- Για την προσθήκη μιας ακολουθίας (ή μιας πολλαπλής στοίχισης) σε μια υπάρχουσα πολλαπλή στοίχιση, χρησιμοποιεί τη μέθοδο profile alignment.
- Χρησιμοποιεί μια σειρά από πολύ προσεκτικά επιλεγμένες ευριστικές τεχνικές οι οποίες μεγιστοποιούν το αποτέλεσμα. Για παράδειγμα, οι πολύ όμοιες ακολουθίες λαμβάνουν μικρό σχετικό βάρος (weight) έτσι ώστε να μην επηρεάζουν τόσο πολύ και να μην κατευθύνουν την πολλαπλή στοίχιση. Μια άλλη ιδιαιτερότητα είναι ότι ο πίνακας ομοιότητας δεν είναι σταθερός, αλλά επιλέγεται από τον αλγόριθμο ανάλογα με το ποσοστό ομοιότητας που εντοπίζεται στις υπό μελέτη ακολουθίες. Επιπλέον, οι ποινές για τα κενά, δεν είναι σταθερές, αλλά ειδικές ανά θέση (υδρόφοβες περιοχές λαμβάνουν μεγαλύτερη ποινή για τα κενά, με συνέπεια να καθίσταται πιο δύσκολη η εισαγωγή κενών σε αυτές τις περιοχές, αντίθετα, η ποινή μειώνεται αν βρεθούν πάνω από 5 συνεχόμενα υδρόφιλα κατάλοιπα). Τέλος, οι ποινές για τα κενά αυξάνονται αν στην ίδια στήλη της στοίχισης δεν υπάρχουν κενά, αλλά αντίθετα υπάρχει κάπου δίπλα μια περιοχή με πολλά κενά. Αυτό έχει σαν συνέπεια τα κενά να «συσσωρεύονται» σε συγκεκριμένες θέσεις σε μια στοίχιση. Όλες αυτές οι τεχνικές, έχουν βελτιωθεί με τα χρόνια και έχουν κάνει το CLUSTAL να είναι ένα από τα πιο αξιόπιστα εργαλεία πολλαπλής στοίχισης, παρόλο που κατά βάση στηρίζεται σε μια απλή ευριστική μέθοδο.

Ένας άλλος σύγχρονος αλγόριθμος πολλαπλής στοίχισης, ο οποίος βασίζεται στην προοδευτική πολλαπλή στοίχιση, είναι το **Kalign** (Lassmann & Sonnhammer, 2005), (διαθέσιμο στη διεύθυνση http://msa.sbc.su.se/cgi-bin/msa.cgi). Στο Kalign, όλες οι επιλογές της προοδευτικής πολλαπλής στοίχισης είναι βελτιστοποιημένες με σκοπό την ταχύτητα. Βασισμένοι στην παρατήρηση ότι το μεγαλύτερο ποσοστό του υπολογιστικού χρόνου οι αλγόριθμοι το καταναλώνουν στις κατά ζεύγη στοιχίσεις από τις οποίες θα υπολογιστούν οι αποστάσεις, οι Lassmann και Sonnhammer επέλεξαν αντί για έναν αλγόριθμο στοίχισης δυναμικού προγραμματισμού, τον προσεγγιστικό αλγόριθμο ταύτισης συμβολοσειρών, των Wu και Manber, ο οποίος είναι γραμμικός ως προς το μήκος της ακολουθίας (Wu & Manber, 1992). Με αυτόν τον τρόπο το Kalign εκτιμά τις αποστάσεις το ίδιο γρήγορα με τη μέθοδο των k-tuple, αλλά πολύ πιο αποδοτικά. Επιπλέον, το δέντρο-οδηγός κατασκευάζεται με τη μέθοδο UPGMA η οποία είναι ίσως η πιο γρήγορη (αλλά όχι τόσο ακριβής) μέθοδος ομαδοποίησης. Και αυτή τη μέθοδο θα την αναλύσουμε στο κεφάλαιο των φυλογενετικών σχέσεων. Οι στοιχίσεις πραγματοποιούνται με την κλασική μέθοδο profile alignment, με την επιπλέον επιλογή, οι κοινές ακολουθίες που βρέθηκαν στο πρώτο βήμα, να μπορούν να καθοδηγούν τη στοίχιση (αυτή η επιλογή καθυστερεί κάπως τους υπολογισμούς, αλλά είναι πιο ακριβής). Τέλος, μια άλλη ιδιαιτερότητα της μεθόδου, βασίζεται στην παρατήρηση ότι πολύ όμοιες ακολουθίες στοιχίζονται αρκετά καλά ανεξαρτήτως του πίνακα ομοιότητας, αλλά ακολουθίες οι οποίες βρίσκονται εξελικτικά μακριά, απαιτούν τον κατάλληλο πίνακα (BLOSUM50, PAM250 ή GONNET250). Βασισμένοι σε αυτό, οι συγγραφείς επέλεξαν σε όλες τις περιπτώσεις να χρησιμοποιείται ο πίνακας GONNET250 (Gonnet, Cohen, & Benner, 1992), μια επιλογή που διευκολύνει αρκετά τους υπολογισμούς. Με όλες αυτές τις βελτιστοποιήσεις, το Kalign καταφέρνει να αποδίδει ελάχιστα χειρότερα από το CLUSTAL αλλά να πραγματοποιεί τις στοιχίσεις ως και 10 φορές πιο γρήγορα. Όπως θα δούμε παρακάτω, ανάλογα με την εφαρμογή, υπάρχουν περιπτώσεις στις οποίες ο χρόνος είναι πιο καθοριστικός παράγοντας σε σχέση με την ακρίβεια. Περισσότερα για το πως αξιολογούμε την ακρίβεια μιας μεθόδου πολλαπλής στοίχισης, θα δούμε στο τέλος του κεφαλαίου.



Εικόνα 4.4 *Το προτελευταίο βήμα της πολλαπλής στοίχισης που περιγράφεται στην Εικόνα 4.2.όπως θα είχε πραγματοποιηθεί με χρήση profile alignment. Σε έναν κλασικό πίνακα δυναμικού προγραμματισμού, τοποθετούμε τη στοίχιση των ακολουθιών x_1 και x_2 από τη μια μεριά, και των x_3 και x_4 από την άλλη. Οι δύο στοιχίσεις σκοράρονται με τη σχέση (4.14) και η βέλτιστη διαδρομή εντοπίζεται με τον κλασικό τρόπο. Και σε αυτή την περίπτωση, το κενό που υπήρχε στη στοίχιση των x_3 και x_4 παραμένει, ενώ τα κενά που εισάγονται στις x_1 και x_2, εισάγονται ταυτόχρονα και στις δύο ακολουθίες. Στο παράδειγμα αυτό, η τελική στοίχιση είναι ίδια με όλες τις μεθόδους, αλλά αυτό δεν ισχύει γενικά. Σε άλλες περιπτώσεις, η μέθοδος αυτή θα δώσει διαφορετικά αποτελέσματα, τα οποία σε γενικές γραμμές θα είναι και καλύτερα.*

Δεν πρέπει να ξεχνάμε, ότι η προοδευτική πολλαπλή στοίχιση, είναι ευριστική μέθοδος. Δεν βελτιστοποιεί κάποιο ολικό μέτρο «καταλληλότητας» της στοίχισης, και δεν διαχωρίζει τη διαδικασία αξιολόγησης μιας στοίχισης από τον αλγόριθμο βελτιστοποίησης. Το πιο σημαντικό από όλα, είναι το γεγονός ότι με τον τρόπο που δουλεύει η μέθοδος, ένα κενό που εισάγεται νωρίς στη διαδικασία, δεν αναιρείται ποτέ («*once a gap, always a gap*»). Παρ' όλα αυτά, είναι ιδιαίτερα ενδιαφέρον το γεγονός ότι ευριστικοί αλγόριθμοι με προσεκτικά επιλεγμένες επιλογές, καταφέρνουν να αποδίδουν ιδιαίτερα καλά. Στην επόμενη ενότητα, θα δούμε μια άλλη μεγάλη κατηγορία μεθόδων, οι οποίες αν και είναι υπολογιστικά περισσότερο απαιτητικές επιδιώκουν να διορθώσουν τέτοια αρχικά λάθη της στοίχισης.

4.3. Επαναληπτικές μέθοδοι και μέθοδοι που βασίζονται στη συνέπεια

Η βασική ιδέα των επαναληπτικών μεθόδων, είναι να χρησιμοποιηθεί κάποιου είδους προοδευτική πολλαπλή στοίχιση, αλλά αυτή η διαδικασία να γίνει επαναληπτικά έτσι ώστε λάθη που είναι πιθανό να εισχωρήσουν σε αρχικά στάδια της στοίχισης, να μπορούν να αναιρεθούν σε κάποιο μετέπειτα βήμα. Η επαναληπτική διαδικασία, είναι σε γενικές γραμμές μια εύκολα υλοποιήσιμη ιδέα, και εμπειρικές αναλύσεις έχουν δείξει ότι μπορεί να χρησιμοποιηθεί ακόμα και σε ήδη υπάρχοντες αλγόριθμους, αυξάνοντας σημαντικά την απόδοσή

τους. Για παράδειγμα, η ακρίβεια του CLUSTALW αυξάνει κατά 6% με αυτή τη διαδικασία (Wallace, O'Sullivan, & Higgins, 2005).

Μια από τις πρώτες υλοποιήσεις επαναληπτικού αλγόριθμου, ήταν ο αλγόριθμος των Barton και Sternberg (Barton & Sternberg, 1987). Ο αλγόριθμος σε γενικές γραμμές, έκανε τις στοιχίσεις με κλασικό δυναμικό προγραμματισμό και μετά ξεκινούσε την πολλαπλή στοίχιση από τις ακολουθίες με τη μεγαλύτερη ομοιότητα. Στη συνέχεια, πρόσθετε στη στοίχιση την επόμενη πιο όμοια ακολουθία χρησιμοποιώντας profile alignment. Όταν είχε στοιχίσει όλες τις ακολουθίες, τις αφαιρούσε διαδοχικά μία-μία από τη στοίχιση και τις πρόσθετε εκ νέου, έως ότου βρεθεί μια πολλαπλή στοίχιση με καλύτερο σκορ. Παρόμοια στρατηγική είχε και ο αλγόριθμος του Corpet (Corpet, 1988), στον οποίο βασίζεται το πρόγραμμα πολλαπλής στοίχισης **MULTALIN** (http://prodes.toulouse.inra.fr/multalin/multalin.html). Η βασική διαφορά είναι στο πώς γίνεται το επαναληπτικό βήμα. Στο MULTALIN, όταν ολοκληρωθεί η πρώτη πολλαπλή στοίχιση, το νέο δέντρο, το οποίο προκύπτει με ιεραρχική ομαδοποίηση, περιέχει ένα βραχίονα λιγότερο γιατί οι δύο πιο όμοιες ακολουθίες θεωρούνται μία ομάδα και αυτό συνεχίζεται και στις επόμενες επαναλήψεις.

Το **MUSCLE** είναι ένα σύγχρονο πρόγραμμα προοδευτικής στοίχισης το οποίο εργάζεται επαναληπτικά (Edgar, 2004) (είναι διαθέσιμο στη διεύθυνση http://www.drive5.com/muscle). Στον πρώτο κύκλο, το MUSCLE χρησιμοποιεί μια γρήγορη μέθοδο βασισμένη στα *k*-mers (κοινές υπο-ακολουθίες μήκους *k*), για να υπολογίσει αποστάσεις και να κατασκευάσει γρήγορα ένα δέντρο οδηγό με τη μέθοδο UPGMA, από το οποίο θα κατασκευάσει μια πρόχειρη στοίχιση (την ονομάζει MSA1). Από αυτή τη στοίχιση, θα υπολογιστούν αποστάσεις με τη μέθοδο του Kimura (η οποία απαιτεί την ύπαρξη της πολλαπλής στοίχισης και λεπτομέρειες της οποίας θα δούμε στο κεφάλαιο 6), από τις οποίες με προοδευτική πολλαπλή στοίχιση και profile alignment θα κατασκευαστεί η δεύτερη στοίχιση (την οποία ονομάζει MSA2). Στο τελευταίο βήμα (refinement), η μέθοδος διαγράφει διαδοχικά βραχίονες του δέντρου το οποίο έχει προκύψει, και στοιχίζει ξανά τις ακολουθίες αυτού του βραχίονα με τις υπόλοιπες ακολουθίες του δέντρου. Αυτό το βήμα επαναλαμβάνεται μέχρι η μέθοδος να συγκλίνει ή μέχρι να ολοκληρωθεί ένας προκαθορισμένος από το χρήστη αριθμός επαναλήψεων. Το τελικό αποτέλεσμα είναι αυτό που το πρόγραμμα ονομάζει MSA3, αλλά το λογισμικό δίνει επιλογή να σταματάει και στο MSA2 (αυτή είναι η επιλογή MUSCLE-p) σαν μια γρήγορη λύση, καθώς είναι εμφανές ότι το μεγαλύτερο κομμάτι του χρόνου εκτέλεσης αναλώνεται στο τελευταίο επαναληπτικό βήμα του αλγορίθμου. Το MUSCLE-p έχει πολυπλοκότητα υπολογισμών $O(N^2L+NL^2)$ και μνήμης $O(N^2+NL+L^2)$, ενώ το τελευταίο βήμα προσθέτει ένα επιπλέον $O(N^3L)$ στην πολυπλοκότητα των υπολογισμών. Μια άλλη ιδιαιτερότητα του MUSCLE είναι το γεγονός ότι χρησιμοποιεί μια εντελώς διαφορετική μέθοδο για να σκοράρει το profile alignment, τη μέθοδο «log-expectation score». Το MUSCLE θεωρείται ένα από τα καλύτερα σύγχρονα εργαλεία, ενώ είναι ιδιαίτερα δυνατό στη στοίχιση profiles όχι μόνο σαν ενδιάμεσο βήμα στην κατασκευή της πολλαπλής στοίχισης, αλλά και σαν αυτοδύναμη λειτουργία.

Ίσως ένας από τους πιο ενδιαφέροντες επαναληπτικούς αλγόριθμους, είναι ο αλγόριθος του Gotoh (Gotoh, 1996) ο οποίος υλοποιείται στο λογισμικό **PRRP/PRRN** (http://www.genome.ist.i.kyoto-u.ac.jp/~aln_user/prrn/index.html). Ο αλγόριθμος χρησιμοποιεί μια διπλή επαναληπτική στρατηγική με τυχαιοποίηση, η οποία βελτιστοποιεί ένα σταθμισμένο SP σκορ (weighted sums-of-pairs score) με σύνθετη ποινή για τα κενά. Η πρωτοτυπία του, εντοπίζεται στο γεγονός ότι τόσο τα βάρη όσο και η ίδια η στοίχιση βελτιστοποιούνται ταυτόχρονα. Η εσωτερική επαναληπτική διαδικασία βελτιστοποιεί το σταθμισμένο SP σκορ, ενώ η εξωτερική, βελτιστοποιεί τα βάρη τα οποία υπολογίζονται για το φυλογενετικό δέντρο που εκτιμάται από την παρούσα στοίχιση.

Το **PRALINE**, (Simossis & Heringa, 2005) το οποίο είναι διαθέσιμο στη διεύθυνση http://ibivu.cs.vu.nl/programs/pralinewww/, είναι μια επαναληπτική μέθοδος η οποία βασίζεται σε μια διαφορετική επαναληπτική στρατηγική. Οι ακολουθίες αντικαθίστανται από ένα profile το οποίο κατασκευάζεται με PSI-BLAST από μια αρχική πολλαπλή στοίχιση μόνο των πολύ όμοιων ακολουθιών. Αυτή η διαδικασία επαναλαμβάνεται, έως ότου τα profile συγκλίνουν και η συλλογή παραμείνει σταθερή. Κατόπιν, η πολλαπλή στοίχιση πραγματοποιείται με μια κλασική διαδικασία προοδευτικής στοίχισης στην οποία οι ακολουθίες αντικαθίστανται από τα profiles. Καθώς οι πολλαπλές στοιχίσεις παίζουν ρόλο και στους αλγόριθμους πρόγνωσης της δευτεροταγούς δομής, το PRALINE μπορεί να ενσωματώσει και αυτή την πληροφορία. Για τη χρήση των πολλαπλών στοιχίσεων στην πρόγνωση της δομής, θα μιλήσουμε στο αντίστοιχο κεφάλαιο, ενώ η ιδέα να αντικαθίστανται οι ακολουθίες από profile, θα μας απασχολήσει στο επόμενο κεφάλαιο στο οποίο θα μελετήσουμε αναλυτικά τα profiles. Η ιδέα αυτή είναι πολύ ενδιαφέρουσα, γιατί προτείνει μια εναλλαγή ανάμεσα στη διαδικασία πολλαπλής στοίχισης και τη διαδικασία πρόβλεψης της δομής, από την οποία επωφελούνται και οι δύο διαδικασίες. Τέλος, η μέθοδος αυτή είναι πολύ ενδιαφέρουσα και για έναν άλλο λόγο. Με τη μέθοδο αυτή, μπορεί να μετρηθεί η συνέπεια (consistency) ανάμεσα στην

τελική στοίχιση και στη συλλογή των profiles τα οποία χρησιμοποιήθηκαν. Η έννοια της συνέπειας με κάποιο εξωτερικό κριτήριο είναι βασική στην επόμενη μεγάλη ομάδα αλγορίθμων.

Το **Dialign** (Morgenstern, 2014), (διαθέσιμο στη διεύθυνση http://bibiserv.techfak.uni-bielefeld.de/dialign/), είναι μια ιδιαίτερη περίπτωση, καθώς είναι ένας από τους λίγους αλγόριθμους προοδευτικής πολλαπλής στοίχισης, ο οποίος πραγματοποιεί στοιχίσεις και με χαρακτηριστικά τοπικής στοίχισης (αλλά, φυσικά, διαθέτει και χαρακτηριστικά ολικής στοίχισης, όπως όλοι οι αλγόριθμοι πολλαπλής στοίχισης). Αρχικά πραγματοποιούνται όλες οι ανά δυο στοιχίσεις και στη συνέχεια συλλέγονται οι στοιχισμένες περιοχές στις οποίες δεν υπάρχουν κενά. Το όνομα 'Dialign' βγαίνει από αυτές τις διαγώνιες περιοχές (diagonal alignments in a dot plot). Το πρόγραμμα δεν βάζει αρχικά ποινή για τα κενά, και δεν επιχειρεί να στοιχίσει περιοχές που δεν έχουν μεγάλη ομοιότητα. Κατά συνέπεια, για ακολουθίες με μόνο τοπική ομοιότητα, η πολλαπλή στοίχιση περιορίζεται στις περιοχές με ξεκάθαρη ομολογία, αγνοώντας τις μη όμοιες περιοχές. Όταν όμως πρόκειται για ακολουθίες με την ομοιότητα να εκτείνεται σε όλο το μήκος τους, ο αλγόριθμος εντοπίζει τμήματα από τις ακολουθίες, που εκτείνονται σε όλο το μήκος τους, και έτσι μετατρέπεται σε αλγόριθμο ολικής στοίχισης. Σε ενδιάμεσες περιπτώσεις, ο αλγόριθμος δίνει μια μίξη: στοιχίζει τις κοινές περιοχές που έχουν μεγάλη ομοιότητα (ακόμα και σε όλο το μήκος), ενώ τις ακολουθίες στις οποίες δεν υπάρχει μια κοινή περιοχή, τις αφήνει εκτός στοίχισης. Κατά συνέπεια, το Dialign είναι περισσότερο ευέλικτο από τα υπόλοιπα προγράμματα, και μπορεί να εφαρμοστεί σε περισσότερες περιπτώσεις χωρίς παρέμβαση στις ακολουθίες (π.χ. εντοπισμό και αποκοπή των μη όμοιων περιοχών).

Μια άλλη μέθοδος, η οποία χρησιμοποιεί στοιχεία παρόμοια με αυτά τόσο του PRALINE όσο και του Dialign, είναι το **COBALT** (Papadopoulos & Agarwala, 2007), το οποίο αποτελεί τμήμα της σουίτας εργαλείων του NCBI (ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/cobalt). Το COBALT χρησιμοποιεί το BLAST και το RPS-BLAST σαν το εργαλείο ομοιότητας, και επιτελεί προοδευτική στοχίση με ένα δέντρο οδηγό το οποίο παράγεται με τη μέθοδο Neighbour-Joining, αλλά χρησιμοποιεί μια ελαφρώς μετασχηματισμένη απόσταση στην οποία συμμετέχουν τα σκορ των ανά δύο στοιχίσεων (d_{ij}=1-(S_{ij}/2)(1/S_{ii}-1/S_{ij})). Στη συνέχεια, κάνει profile alignment με δυναμικό προγραμματισμό στον οποίο όμως υπάρχουν βασικές τροποποιήσεις, τόσο στο σκοράρισμα όσο και στον τρόπο χειρισμού των κενών. Η βασικότερη όμως ιδιαιτερότητά του, είναι ότι με τη χρήση του BLAST κάνει αναζήτηση τοπικής ομοιότητας, και με τη χρήση του RPS-BLAST, πραγματοποιεί αναζητήσεις των ακολουθιών έναντι της βάσης των συντηρημένων περιοχών του NCBI (CDD). Με αυτόν τον τρόπο καταφέρνει να στοιχίζει καλά τις συντηρημένες περιοχές των ακολουθιών, επιτυγχάνοντας κάτι παρόμοιο με το Dialign.

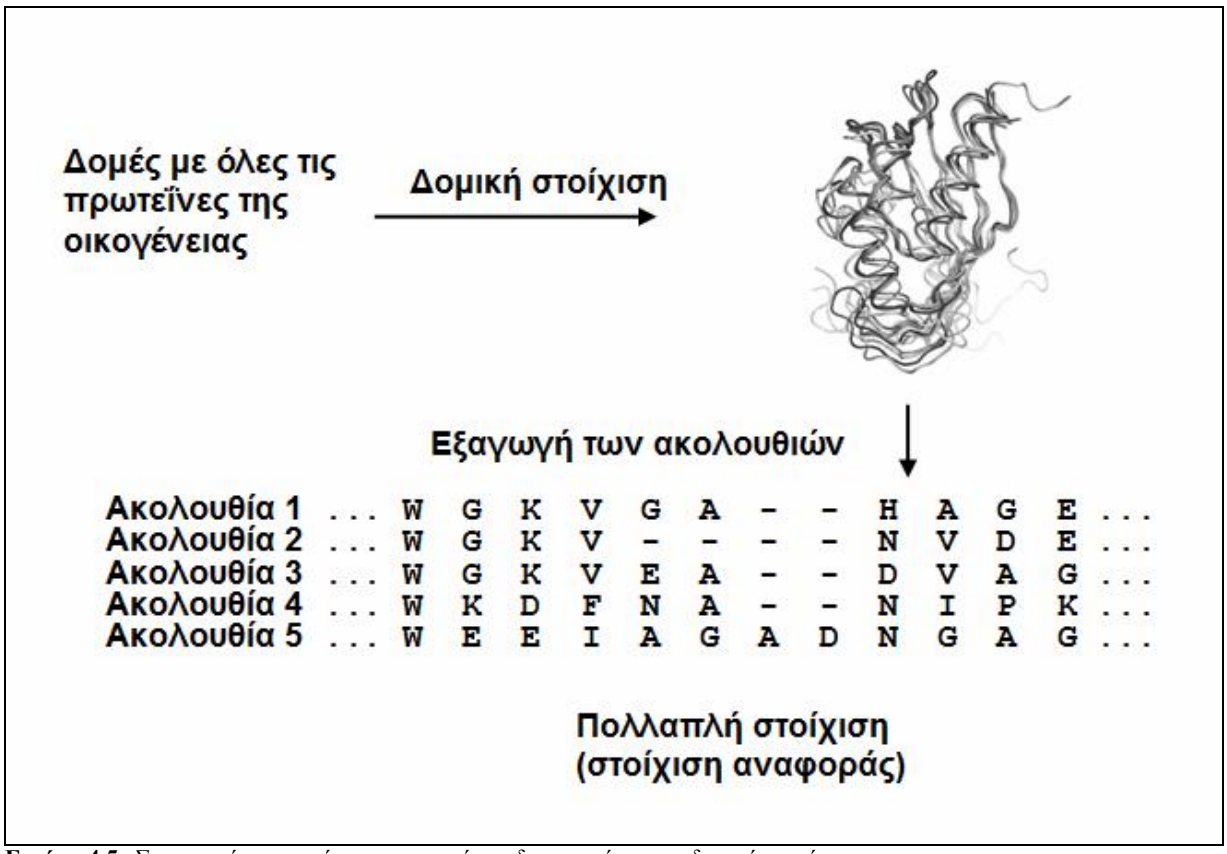
Τέλος, το πιο σημαντικό ίσως από τα εργαλεία που βασίζονται σε προοδευτική πολλαπλή στοίχιση χρησιμοποιώντας την έννοια της συνέπειας, είναι το **T-Coffee** (Magis et al., 2014), (διαθέσιμο στη διεύθυνση http://www.ch.embnet.org/software/TCoffee.html). Το T-Coffee μοιάζει πολύ με το CLUSTALW (τον κώδικα του οποίου μάλιστα, χρησιμοποιεί εσωτερικά): κάνει προοδευτική πολλαπλή στοίχιση, χρησιμοποιεί τον ίδιο αλγόριθμο για την κατασκευή του δέντρου, τον ίδιο τρόπο υπολογισμού των αποστάσεων αλλά και το profile alignment. Η βασική του διαφορά είναι ότι δημιουργεί μια «εκτεταμένη βιβλιοθήκη» όπως την αποκαλεί, στην οποία ένας πίνακας αντικατάστασης ειδικός ανά θέση αντιστοιχίζεται σε κάθε ζεύγος ακολουθιών, και ο οποίος αντικατοπτρίζει τη συμβατότητα της στοίχισης των δύο ακολουθιών με την υπόλοιπη βιβλιοθήκη. Με αυτόν τον τρόπο, οι πίνακες αντικατάστασης, αντικαθίστανται από αυτούς τους ειδικούς ανά θέση πίνακες, οπότε, κάθε πιθανή επέκταση μιας στοίχισης δύο ακολουθιών, ελέγχεται όχι με ένα γενικό πίνακα, αλλά με βάση το αν αυτές ταιριάζουν καλά στις υπόλοιπες ακολουθίες της βιβλιοθήκης. Μια άλλη διαφορά σε σχέση με το CLUSTALW είναι το γεγονός ότι στην αρχική βιβλιοθήκη, αυτή την οποία δημιουργεί από τις κατά ζεύγη στοιχίσεις, περιέχονται τόσο αποτελέσματα ολικής ομοιότητας (το βήμα αυτό επιτελείται με χρήση του CLUSTALW), όσο και αποτελέσματα τοπικής ομοιότητας (με χρήση του LALIGN από το πακέτο FASTA). Με αυτές τις ιδιαιτερότητες, το T-Coffee συνδυάζει τα χαρακτηριστικά από πολλούς από τους αλγορίθμους που αναφέραμε προηγουμένως. Κάνει προοδευτική στοίχιση που είναι γρήγορη, αλλά ελέγχει και τα διάφορα βήματα για τη συνέπειά τους έτσι ώστε να διορθώνονται τα λάθη. Επιπλέον, χρησιμοποιεί και τοπική αλλά και ολική πληροφορία. Με όλα αυτά, ο αλγόριθμος καταφέρνει να πραγματοποιεί πολύ καλές στοιχίσεις και να θεωρείται ίσως ο πιο πετυχημένος αλγόριθμος γενικής χρήσης αυτή τη στιγμή.

Τέλος, αξίζει να αναφερθεί και μια άλλη κατηγορία αλγορίθμων πολλαπλής στοίχισης, οι οποίοι δεν βασίζονται στην προοδευτική πολλαπλή στοίχιση, αλλά βελτιστοποιούν ένα συνολικό κριτήριο πάνω στην πολλαπλή στοίχιση και στηρίζονται σε τεχνικές γνωστές από το χώρο της τεχνητής νοημοσύνης και των στοχαστικών μοντέλων. Μια τέτοια κατηγορία μεθόδων, είναι αυτές που βασίζονται στο λεγόμενο simulated annealing (Kim, Pramanik, & Chung, 1994), αλλά δεν υπάρχουν αυτή τη στιγμή αξιόπιστες σύγχρονες

υλοποιήσεις του. Μια άλλη κατηγορία αποτελούν οι μέθοδοι που βασίζονται στους γενετικούς αλγόριθμους, όπως για παράδειγμα το παλιότερο πρόγραμμα **SAGA** (Notredame & Higgins, 1996), ενώ η πιο μεγάλη κατηγορία είναι οι μέθοδοι που βασίζονται σε πιθανοθεωρητικά μαρκοβιανά μοντέλα (Hidden Markov Models), όπως η μέθοδος **ProbCons** και **ProbAlign** (Roshan, 2014) (διαθέσιμα στο http://probalign.njit.edu/standalone.html). Η τελευταία κατηγορία μεθόδων είναι πολύ σημαντική, γιατί τα μοντέλα αυτά βρίσκουν πολλές εφαρμογές στη βιοπληροφορική, παρέχουν μια πιθανοθεωρητική ερμηνεία των αποτελεσμάτων αποφεύγοντας τις ευριστικές λύσεις, αλλά κυρίως, γιατί καταφέρνουν να επιλύουν παρόμοια προβλήματα με πολύ ικανοποιητικό τρόπο. Σε επόμενο κεφάλαιο, θα αναπτύξουμε κάποια βασικά θέματα που αφορούν τα μοντέλα αυτά.

4.4. Αξιολόγηση των εργαλείων πολλαπλής στοίχισης

Αφού παρουσιάσαμε τους κύριους αλγόριθμους πολλαπλής στοίχισης και τις διάφορες υλοποιήσεις τους, πρέπει να επιστρέψουμε τώρα στο πρόβλημα της αξιολόγησης. Πώς μπορούμε να αξιολογήσουμε αν ένα δεδομένο πρόγραμμα πολλαπλής στοίχισης δουλεύει καλά; Πώς μπορούμε να συγκρίνουμε δύο ή περισσότερα προγράμματα; Όπως είδαμε, υπάρχει ένας τρόπος να αξιολογηθεί μια δεδομένη πολλαπλή στοίχιση, και αυτό μπορεί να γίνει με χρήση κάποιου γενικού στατιστικού μέτρου όπως για παράδειγμα της εντροπίας. Παρ' όλα αυτά, χρειαζόμαστε και κάποιο εξωτερικό κριτήριο αντικειμενικότητας. Σε αυτό, θα πρέπει με κάποιον τρόπο να ενσωματωθεί και η βιολογική όψη του προβλήματος, καθώς για τις πολλαπλές στοιχίσεις, δεν υπάρχει γενικώς αποδεκτός ή εύκολος τρόπος υπολογισμού της στατιστικής σημαντικότητας.



Εικόνα 4.5: *Σχηματική αναπαράσταση του τρόπου δημιουργίας μιας δομικής στοίχισης.*

Γενικά, έχει γίνει αποδεκτό, ότι κριτήριο αναφοράς για μια πολλαπλή στοίχιση, είναι η λεγόμενη «δομική στοίχιση» (structural alignment). Μια δομική στοίχιση, προκύπτει από την υπέρθεση των τρισδιάστατων δομών μια πρωτεϊνικής οικογένειας, για την οποία γνωρίζουμε ότι τα μέλη της έχουν ξεκάθαρη εξελικτική και δομική ομοιότητα. Βασική προϋπόθεση φυσικά, είναι να υπάρχει κάποια ή κάποιες οικογένειες πρωτεϊνών, για τις οποίες υπάρχουν τρισδιάστατες δομές για μεγάλο αριθμό από τα μέλη τους. Η

υπέρθεση των δομών, στην πιο απλή της μορφή, ελαχιστοποιεί τις αποστάσεις των αντίστοιχων ατόμων από τις διαφορετικές πρωτεΐνες και τις φέρνει όσο το δυνατό πιο κοντά στο χώρο. Στην πιο περίπλοκη κατάσταση, κατά την οποία οι πρωτεΐνες δεν έχουν το ίδιο μήκος, επειδή για παράδειγμα σε μερικές λείπουν κάποιες περιοχές, τότε απαιτούνται ειδικοί αλγόριθμοι στοίχισης των τρισδιάστατων δομών. Σε κάθε περίπτωση, από μία καλά κατασκευασμένη δομική στοίχιση, μπορεί να προκύψει μια πολλαπλή στοίχιση ακολουθιών, απλά αγνοώντας τη δομή και κρατώντας την πληροφορία μόνο της ακολουθίας. Με αυτόν τον τρόπο, κάθε στήλη της πολλαπλής στοίχισης αντιστοιχεί στα αντίστοιχα αμινοξέα των περιλαμβανόμενων στη στοίχιση πρωτεϊνών, τα οποία βρίσκονται πιο κοντά στο χώρο. Όπως γίνεται φανερό, μια τέτοια στοίχιση θεωρείται σημείο αναφοράς («gold standard») για τις πρωτεΐνες της οικογένειας, και μια μέθοδος πολλαπλής στοίχισης θα θεωρείται καλή αν καταφέρνει να ανακατασκευάζει αυτή τη στοίχιση ή να την προσεγγίζει. Για το σκοπό αυτό, έχουν αναπτυχθεί μια σειρά από βάσεις δεδομένων οι οποίες περιέχουν τέτοιες δομικές πολλαπλές στοιχίσεις πρωτεϊνικών οικογενειών, με διαφορετικά χαρακτηριστικά. Η πρώτη τέτοια βάση δεδομένων ήταν η **BAliBASE** (Thompson, Plewniak, & Poch, 1999) και για πολλά χρόνια οι περισσότερες αξιολογήσεις γίνονταν πάνω σε αυτήν. Την τελευταία δεκαετία όμως έχουν αναπτυχθεί και άλλες τέτοιες βάσεις δεδομένων/συλλογές πρωτεϊνικών ακολουθιών οι οποίες δίνονται στον Πίνακα 4.1.

Βάση δεδομένων	**Ηλεκτρονική Διεύθυνση**
BAliBASE (Thompson, et al., 1999)	http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE/index.html
OxBench (Raghava, Searle, Audley, Barber, & Barton, 2003)	http://www.compbio.dundee.ac.uk/
SABmark (Van Walle, Lasters, & Wyns, 2005)	http://bioinformatics.vub.ac.be/databases/databases.html
PREFAB (Edgar, 2004)	http://drive5.com/muscle/prefab.htm.

Πίνακας 4.1: *Οι βάσεις δεδομένων με δομικές στοιχίσεις πρωτεϊνικών οικογενειών που χρησιμοποιούνται για την αξιολόγηση των μεθόδων πολλαπλής στοίχισης*

Σε μια αξιολόγηση, συνήθως επιλέγεται ένα μεγάλο και ετερογενές σύνολο από οικογένειες από κάποια βάση, και οι ακολουθίες υποβάλλονται στα προγράμματα για πολλαπλή στοίχιση. Ιδανικά, μια μέθοδος αποδίδει «τέλεια» όταν ανακατασκευάζει στο 100% την αρχική δομική στοίχιση. Προφανώς, όταν οι οικογένειες έχουν πολλά μέλη, αυτό είναι σχετικά δύσκολο να συμβεί και για αυτό το λόγο επιλέγονται διάφορα μέτρα που αξιολογούν, λ.χ. πόσες στήλες της πολλαπλής στοίχισης έχουν ανακατασκευαστεί σωστά από το κάθε πρόγραμμα. Υπάρχουν διάφορα τέτοια μέτρα, με άλλα να υπολογίζουν το ποσοστό των σωστών στηλών επί του συνόλου της στοίχισης, και άλλα να κάνουν τον υπολογισμό με βάση τη στοίχιση αναφοράς (Thompson, Linard, Lecompte, & Poch, 2011), ενώ υπάρχουν ακόμα και μέτρα που αξιολογούν τη σωστή στοίχιση των διαφόρων περιοχών (blocks) (Raghava, et al., 2003). Γενικά στην αξιολόγηση, πρέπει να επιλέγεται ένα μεγάλο δείγμα οικογενειών, αντιπροσωπευτικό του τι αναμένεται να συναντήσει ο ερευνητής σε μια πραγματική κατάσταση. Οι παράγοντες στους οποίους πρέπει να δοθεί βαρύτητα είναι: α) το μέγεθος της οικογένειας και το μέσο μήκος των πρωτεϊνών της οικογένειας (παράγοντες που αναμένεται να επηρεάζουν τόσο τη συνολική απόδοση των μεθόδων σε απόλυτες τιμές, όσο και τον χρόνο εκτέλεσης της στοίχισης), β) η ομοιότητα των μελών της οικογένειας (είναι πολύ σημαντικό να ξέρουμε αν ένα πρόγραμμα δουλεύει καλά τόσο σε κοντινές εξελικτικά πρωτεΐνες, όσο και σε μακρινές), και γ) το αν στην οικογένεια υπάρχουν πρωτεΐνες-μέλη οι οποίες αποτελούν θραύσματα (fragments), καθώς αυτό θα ελέγξει την ικανότητα του προγράμματος στην «τοπική» πολλαπλή στοίχιση και στην αναγνώριση των διακριτών περιοχών.

Υπάρχουν ακόμα και μέτρα που δεν συγκρίνουν τη στοίχιση με κάποια στοίχιση αναφοράς, αλλά κάνουν απευθείας αναγωγή στις τρισδιάστατες δομές για να δώσουν περισσότερο βάρος σε περιοχές με κανονική δευτεροταγή δομή, σε αντίθεση με τις βρόχους στις οποίες τα κενά και τα λάθη είναι περισσότερο κοινά αλλά και λιγότερο σημαντικά (Raghava, et al., 2003). Ένα τέτοιο μέτρο είναι το APDB (O'Sullivan et al., 2003). Το APDB, αξιολογεί μια πολλαπλή στοχίση, με κριτήριο αναφοράς δύο ή περισσότερες δομές από την PDB, χωρίς όμως να απαιτεί μια στοίχιση αναφοράς ή υπέρθεση των δομών. Στη σχετική δημοσίευση, οι συγγραφείς έδειξαν ότι το μέτρο παράγει αξιολογήσεις που σε γενικές γραμμές συμφωνούν με αυτές που προκύπτουν από μια στοίχιση αναφοράς, και κατά συνέπεια το APDB θα μπορούσε να χρησιμοποιηθεί γενικά, ακόμα και σε οικογένειες για τις οποίες δεν υπάρχει μια αξιόπιστη πολλαπλή στοίχιση.

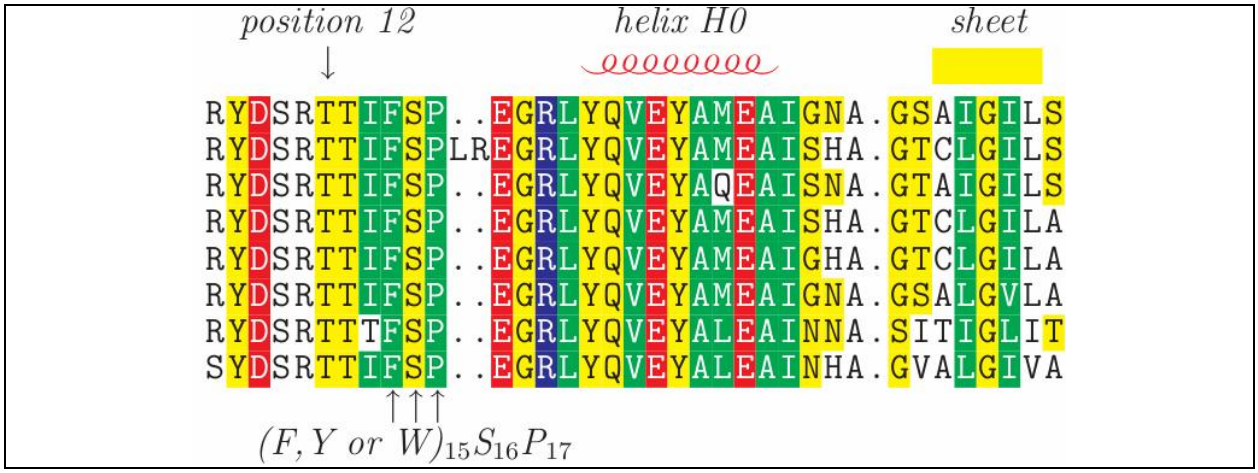
Τα τελευταία χρόνια, έχουν γίνει αρκετές διαφορετικές μελέτες αξιολόγησης των προγραμμάτων πολλαπλής στοίχισης, κάθε φορά με διαφορετικές οικογένειες πρωτεϊνών και πολλές φορές και με διαφορετικά μέτρα αξιολόγησης (Pais, Ruy Pde, Oliveira, & Coimbra, 2014; Thompson, et al., 2011; Thompson, et al., 1999). Επίσης, κάθε νέος αλγόριθμος πολλαπλής στοίχισης, πρέπει πλέον να αξιολογείται με παρόμοια κριτήρια, αν πρόκειται να δημοσιευθεί. Παρόλο που οι επιμέρους μελέτες διαφέρουν πολλές φορές ως προς τη μεθοδολογία, μπορούμε να εξάγουμε κάποια γενικά συμπεράσματα. Για παράδειγμα, τα περισσότερα από τα σύγχρονα εργαλεία που αναφέραμε παραπάνω, σε ένα ευρύ φάσμα συνθηκών αποδίδουν πολύ καλά, πετυχαίνοντας πάνω από 50% επιτυχία στην ανακατασκευή των στοιχίσεων αναφοράς, ακόμα και σε οικογένειες με μέσο ποσοστό ομοιότητας γύρω στο 20%. Το T-Coffee, το ProbCons και το ProbAlign είναι σε γενικές γραμμές οι πιο αποδοτικοί αλγόριθμοι, αλλά είναι και πιο χρονοβόροι και με μεγάλες απαιτήσεις σε μνήμη (ιδιαίτερα τα δύο τελευταία). Το ClustalW και το MUSCLE, ακολουθούν με μικρή διαφορά στην απόδοση, αλλά υπερτερούν σε ταχύτητα εκτέλεσης και σε απαιτήσεις σε μνήμη. Το Prrp/Prrn είναι επίσης καλό, αλλά πιο αργό. Το Kalign, είναι σε γενικές γραμμές ελαφρώς χειρότερο, αλλά είναι έως και 10 φορές γρηγορότερο από το CLUSTALW (πολύ δε περισσότερο από τα υπόλοιπα), και κατά συνέπεια καλύτερο για αναλύσεις μεγάλου όγκου δεδομένων σε καθημερινή βάση. Τέλος, οι αλγόριθμοι που κάνουν ολική στοίχιση, αποδίδουν σε γενικές γραμμές καλύτερα, εκτός αν στις πολλαπλές στοιχίσεις υπάρχουν μεγάλες περιοχές στο αμινοτελικό ή στο καρβοξυτελικό άκρο, οι οποίες δεν ταυτίζονται σε όλα τα μέλη της οικογένειας (δηλαδή, αν υπάρχουν οικογένειες με μέλη τα οποία εμφανίζουν τοπική ομοιότητα). Το T-Coffee γενικά, είναι ένας καλός συμβιβασμός, καθώς τα καταφέρνει σχετικά καλά σε όλες τις περιπτώσεις, ενώ το Dialign αποδεικνύεται καλύτερο μόνο σε κάποια από τα σετ με τέτοιες ακολουθίες (στις πιο ακραίες περιπτώσεις).

Τα δύο τελευταία χαρακτηριστικά, δηλαδή η ταχύτητα και η ικανότητα σωστής στοίχισης σε περιπτώσεις τοπικής ομοιότητας πρέπει να ελέγχονται προσεκτικά και να λαμβάνονται σοβαρά υπόψη στην επιλογή προγράμματος. Η ταχύτητα για παράδειγμα, δεν είναι σημαντική όταν κάνουμε μια μελέτη μιας συγκεκριμένης οικογένειας (θέλουμε να πάρουμε την καλύτερη δυνατή στοίχιση και δεν μας πειράζει να περιμένουμε λίγο). Από την άλλη όμως, είναι ένας σημαντικός παράγοντας αν πρόκειται τις πολλαπλές στοιχίσεις να τις χρησιμοποιούμε λ.χ. για την υποβοήθηση μιας μεθόδου πρόγνωσης, γιατί σε αυτή την περίπτωση θα χρειάζεται να επαναλαμβάνουμε τις στοιχίσεις καθημερινά (για παράδειγμα, αν φτιάχνουμε μια διαδικτυακή εφαρμογή). Κάτι αντίστοιχο ισχύει και για τις τοπικές ομοιότητες των πρωτεϊνών. Αν μελετάμε μια συγκεκριμένη οικογένεια πρωτεϊνών, κατά πάσα πιθανότητα θα ξέρουμε τι είδους στοίχιση να περιμένουμε. Αν όμως πρόκειται η πολλαπλή στοίχιση να χρησιμοποιείται σε μια αυτοματοποιημένη διαδικασία, τότε δεν έχουμε αυτή την πολυτέλεια. Τέλος, ένας άλλος παράγοντας που πρέπει να λαμβάνεται υπόψη είναι και η ευκολία προς τον απλό χρήστη. Τα περισσότερα από τα προγράμματα που αναφέραμε (CLUSTALW, T-Cofffee, Dialign, Kalign, MUSCLE, ProbAlign, Prrp/Prrn), προσφέρονται σαν διαδικτυακές εφαρμογές αλλά και σαν τοπικές εφαρμογές τις οποίες ο χρήστης μπορεί να εγκαταστήσει στον υπολογιστή του. Τα περισσότερα από αυτά, είναι ιδιαίτερα εύκολα στην εγκατάσταση σε όλα τα συστήματα (Windows, Linux, Mac), αλλά το COBALT και το PRALINE, τα οποία απαιτούν χρήση και άλλων προγραμμάτων (PSI-BLAST κλπ), είναι πιο δύσκολα στη ρύθμιση (και για την ακρίβεια, για το PRALINE δεν είμαστε σίγουροι αν υπάρχει και διαθέσιμη εφαρμογή πέραν της διαδικτυακής). Όλα τα παραπάνω είναι παράγοντες που πρέπει να λαμβάνονται σοβαρά υπόψη από τον χρήστη πριν επιλέξει με ποιο πρόγραμμα θα πραγματοποιήσει την ανάλυση του, και σε κάθε περίπτωση, είναι χρήσιμο πάντα κάποιος να δοκιμάζει αρκετές εναλλακτικές προτάσεις.

4.5. Οπτικοποίηση και Επεξεργασία μιας Πολλαπλής Στοίχισης

Το τελευταίο μέρος του κεφαλαίου, είναι αφιερωμένο στο λογισμικό οπτικοποίησης και επεξεργασίας των πολλαπλών στοιχίσεων, όπως και στους τύπους αρχείων πολλαπλής στοίχισης. Τα περισσότερα από τα προγράμματα που αναφέραμε, διαβάζουν ακολουθίες σε απλή μορφή (απλό κείμενο ή FASTA) και παράγουν τις πολλαπλές στοιχίσεις σε κάποια από τις γνωστές μορφές. Υπάρχουν αρκετοί τύποι αρχείων πολλαπλής στοίχισης, αλλά οι βασικοί είναι το Multi-FASTA, το MSF και το CLUSTAL. Το Multi-FASTA, είναι η πιο απλή μορφή και είναι μια γενίκευση του FASTA. Κάθε ακολουθία δίνεται ξεχωριστά, με την πρώτη γραμμή να αποτελεί το όνομα της ή την περιγραφή (με ένα «>» στην αρχή), ενώ οι επόμενες γραμμές περιέχουν την ακολουθία. Για να αναπαρασταθεί η έννοια της στοίχισης, στις ακολουθίες υπάρχουν κενά (-) με συνέπεια το μήκος των ακολουθιών να είναι ίδιο σε κάθε αρχείο. Η μορφή αυτή είναι πολύ απλή, αλλά δεν είναι εύκολα

κατανοητή από το ανθρώπινο μάτι. Η μορφή MSF, λύνει αυτό το πρόβλημα καθώς σε αυτήν, οι ακολουθίες δίνονται σε κομμάτια, στοιχισμένα το ένα κάτω από το άλλο. Αν η στοίχιση έχει για παράδειγμα 3 ακολουθίες, θα υπάρχουν 3 γραμμές με τα πρώτα 50 αμινοξικά κατάλοιπα της κάθε ακολουθίας σε δεκάδες (μαζί με τα κενά της στοίχισης), μετά θα ακολουθούν άλλες 3 γραμμές με τα 50 επόμενα, κ.ο.κ. Προφανώς, σε κάθε γραμμή αναφέρεται το όνομα της πρωτεΐνης για να μπορούμε να την ξεχωρίζουμε. Το αρχείο, στην αρχή, περιέχει σε ξεχωριστή ενότητα (που ξεχωρίζει από τους χαρακτήρες «//») τα ονόματα όλων των ακολουθιών που υπάρχουν στη στοίχιση. Η μορφή CLUSTAL (η οποία προήλθε από το ομώνυμο πρόγραμμα), είναι πιο απλή στην αρχή (δεν περιέχει σε ξεχωριστή ενότητα τα ονόματα των ακολουθιών, και οι ακολουθίες δίνονται συνεχόμενα, χωρίς κενά), αλλά περιέχει μια επιπλέον γραμμή σε κάθε τμήμα το οποίο διαχωρίζει 60 αμινοξικά κατάλοιπα της στοίχισης. Η γραμμή αυτή συμβολίζει τη συνολική «ποιότητα» της στοίχισης και είναι εύκολα κατανοητή από το ανθρώπινο μάτι. Αν σε μια στήλη της πολλαπλής στοίχισης υπάρχει απόλυτη συντήρηση, στη γραμμή αυτή υπάρχει «*». Το «:» και το «.» συμβολίζουν μεγάλη και μικρότερη συντήρηση αντίστοιχα (εξαρτώνται από τον πίνακα ομοιότητας, όχι μόνο από το ποσοστό), ενώ το κενό (« ») συμβολίζει τη διαφορά (μη ταύτιση). Υπάρχουν και άλλες μορφές αρχείων πολλαπλής στοίχισης, όπως το PHYLIP ή το STOCKHOLM, αλλά αυτές που αναφέρθηκαν παραπάνω είναι οι πιο κοινές και διαβάζονται από όλα τα προγράμματα. Εργαλεία, όπως το **READSEQ** (http://www.ebi.ac.uk/Tools/sfc/readseq/), αλλά και αντίστοιχα modules στην BioPerl, BioPython ή BioJava, επιτρέπουν την εύκολη μετατροπή των αρχείων από τη μια μορφή στην άλλη.



Εικόνα 4.6: *Πολλαπλή στοίχιση, όπως αναπαρίσταται από το Strap. Φαίνονται χρωματισμένα τα συντηρημένα κατάλοιπα ανάλογα με τις φυσικοχημικές τους ιδιότητες, και 3 θέσεις οι οποίες έχουν ενδιαφέρον έχουν επισημανθεί. Από πάνω, φαίνονται και οι προγνώσεις δευτεροταγούς δομής με το JNET.*

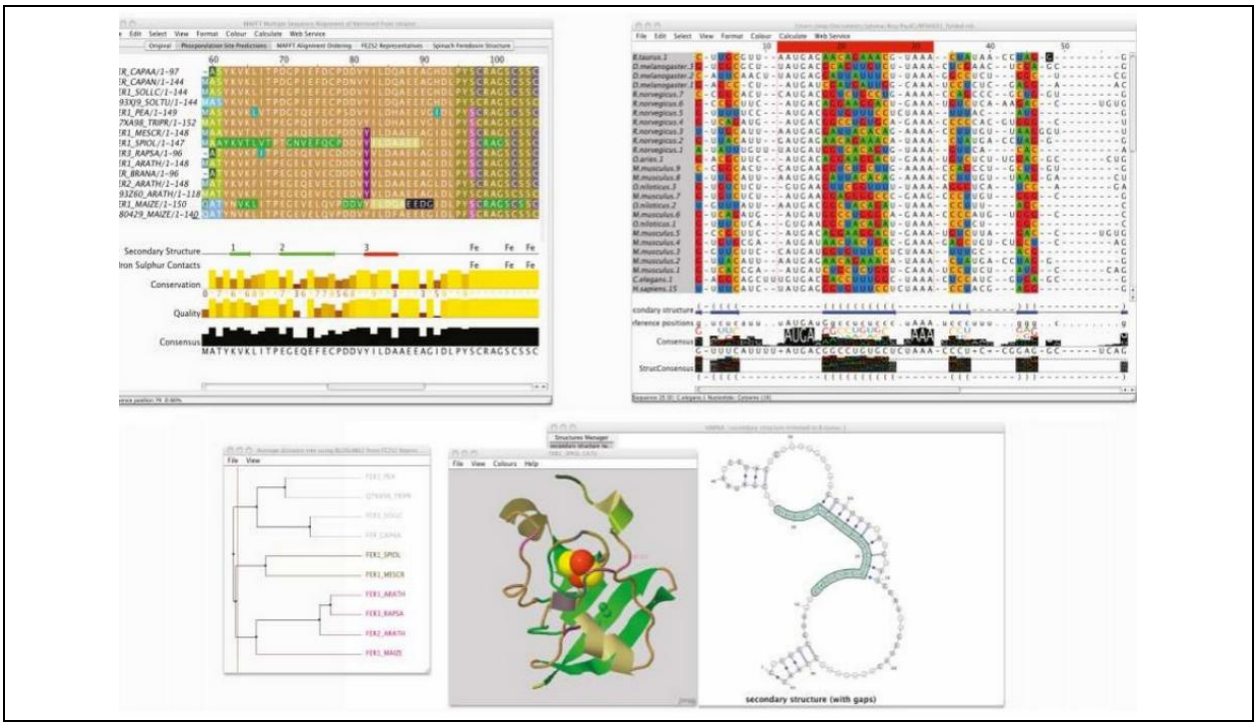
Φυσικά, ένα μεγάλο μέρος της εργασίας που απαιτείται σε μια πολλαπλή στοίχιση, δεν είναι μόνο η ίδια η στοίχιση και η εκτέλεσή της, αλλά και η οπτικοποίηση, η επεξεργασία και η ερμηνεία της. Καταλαβαίνουμε, ότι όσο περιεκτικό και αν είναι από πλευράς πληροφορίας το αρχείο με τα αποτελέσματα μια πολλαπλής στοίχισης, ότι αυτό είναι δύσκολο να μελετηθεί και να αναλυθεί σωστά με το ανθρώπινο μάτι, ειδικά αν μιλάμε για στοιχίσεις μεγάλων πρωτεϊνικών οικογενειών. Για το σκοπό αυτό, έχουν αναπτυχθεί εδώ και χρόνια ειδικά εργαλεία τα οποία οπτικοποιούν τις στοιχίσεις ή τμήματα αυτών, και τις μορφοποιούν σε μορφή κατανοητή και κατάλληλη για παρουσίαση ή δημοσίευση. Τα παλιότερα από αυτά τα εργαλεία, έφτιαχναν απλά στατικές εικόνες βασισμένες σε ένα σύνολο οδηγιών (γραμματοσειρά, χρώμα καταλοίπων κ.ο.κ.). Τα σύγχρονα όμως εργαλεία, προσφέρουν πολλά περισσότερα. Δεν είναι μόνο διαδραστικά εργαλεία (editors), αλλά προσφέρουν και ένα ολοκληρωμένο περιβάλλον εργασίας, με διασυνδέσεις με άλλα εργαλεία (προγράμματα στοίχισης, προγνωστικούς αλγορίθμους κλπ) τόσο τοπικά όσο και στο διαδίκτυο, αλλά και διασυνδέσεις με τις βάσεις δεδομένων (ακολουθιών και δομών). Τα κυριότερα εργαλεία που χρησιμοποιούνται για το σκοπό αυτό είναι τα παρακάτω:

- **Jalview** (http://www.jalview.org/)
- **Strap** (http://www.bioinformatics.org/strap/)
- **Seqpup** (http://iubio.bio.indiana.edu/soft/molbio/seqpup/java/seqpup-doc.html)
- **Seaview** (http://pbil.univ-lyon1.fr/software/seaview.html)

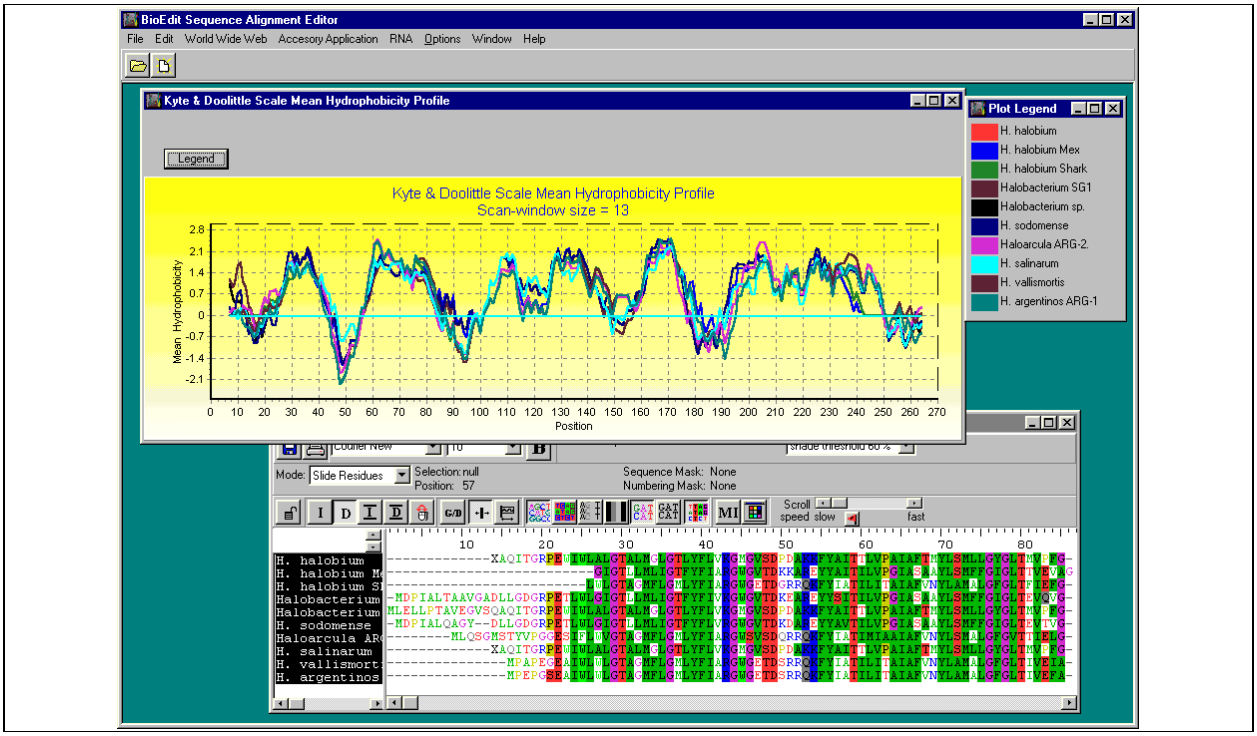
- **Cinema** (http://aig.cs.man.ac.uk/research/utopia/cinema/cinema.php)
- **Boxshade** (http://www.ch.embnet.org/software/BOX_form.html)
- **Bioedit** (http://www.mbio.ncsu.edu/BioEdit/bioedit.html)

Τα εργαλεία αυτά, προσφέρουν μια σειρά από μεγάλες ευκολίες στο χρήστη. Τα περισσότερα είναι εφαρμογές Desktop με ολοκληρωμένο περιβάλλον διαχείρισης αρχείων πολλαπλών στοιχίσεων (κάποια βέβαια, λειτουργούν και διαδικτυακά ως applets). Ο χρήστης μπορεί να φορτώσει μια πολλαπλή στοίχιση και να την επεξεργαστεί. Κάποια μάλιστα, επικοινωνούν και με προγράμματα πολλαπλής στοίχισης, έτσι ώστε η ίδια η πολλαπλή στοίχιση να γίνει μέσω του περιβάλλοντος αυτού. Η βασική εργασία, την οποία επιτελούν όλα τα προγράμματα, είναι να μορφοποιούν την πολλαπλή στοίχιση σε μορφή κατανοητή από το ανθρώπινο μάτι. Για το σκοπό αυτό, χρωματίζουν διαφορετικά τα διάφορα αμινοξικά κατάλοιπα, συνήθως με το όμοιο χρώμα να δίνεται σε αμινοξέα με παρόμοιες φυσικοχημικές ιδιότητες, ενώ τα περισσότερα επιτρέπουν τον καθορισμό του χρωματισμού. Ο χρωματισμός, είναι πολύ βολικός για το χρήστη, γιατί επιτρέπει να εντοπιστούν συντηρημένες περιοχές με μια γρήγορη επισκόπιση της στοίχισης. Το ίδιο έργο επιτελούν και διάφορα στατιστικά, ανά στήλη της πολλαπλής στοίχισης, τα οποία δίνουν το ποσοστό συντήρησης. Τα περισσότερα από τα προγράμματα αυτά, επιτρέπουν την ταυτόχρονη παράθεση της γνωστής ή προβλεφθείσας δευτεροταγούς δομής, παράλληλα με τη στοίχιση. Η δευτεροταγής δομή, παίζει παρόμοιο ρόλο, καθώς είναι γνωστό ότι οι συντηρημένες περιοχές είναι πιθανότερο να έχουν κάποια συγκεκριμένη δομή (α-έλικα ή β-πτυχωτή επιφάνεια), ενώ οι περιοχές με πολλά κενά πιθανότερο είναι να αντιστοιχούν σε βρόχους. Πολλά άλλα παρόμοια στοιχεία είναι επίσης πιθανό να ενσωματώνονται, όπως οι προγνώσεις διαμεμβρανικών τμημάτων, οι θέσεις γλυκοζυλίωσης ή οι δισουλφιδικοί δεσμοί, τα οποία βοηθούν τον ερευνητή να αποκτήσει μια λεπτομερέστερη εικόνα της βιολογίας των υπό μελέτη πρωτεϊνών. Δεδομένα από δημόσιες βάσεις, ειδικά από βάσεις δομών όπως η PDB, είναι επίσης πιθανό να ανασύρονται και να αναπαρίστανται παράλληλα με τη στοίχιση, ενώ πολλά εργαλεία δείχνουν επιπλέον και το φυλογενετικό δέντρο των ακολουθιών της στοίχισης (για ακρίβεια, το δέντρο οδηγό). Τέλος, πρέπει να τονιστεί, ότι κάποια από τα εργαλεία αυτά, είναι παραμετροποιήσιμα, δηλαδή επιτρέπουν στον έμπειρο χρήστη να προσθέσει λειτουργικότητες στο πρόγραμμα, διασυνδέοντάς το με επιπλέον προγράμματα πρόγνωσης, εργαλεία στοίχισης ή τοπικές βάσεις δεδομένων, κάνοντας τα με αυτόν τον τρόπο αναπόσπαστο τμήμα της καθημερινής εργασίας που αφορά την ανάλυση πρωτεϊνικών ακολουθιών.

Έχοντας όλα τα παραπάνω υπόψη μας, μπορούμε τέλος, να δούμε εν συντομία τα βασικά βήματα που πρέπει να κάνει κανείς για να προχωρήσει σε μια σωστή πολλαπλή στοίχιση. Συνήθως, οι ακολουθίες προέρχονται από αναζήτηση ομοιότητας σε κάποια βάση δεδομένων, με βάση 1-2 ακολουθίες αναφοράς που έχει ο ερευνητής και πιθανώς τις έχει μελετήσει πειραματικά. Σε μια τέτοια περίπτωση τα ευρήματα πρέπει να ελέγχονται. Για παράδειγμα, μπορεί οι οργανισμοί από τους οποίους προέρχονται να μην έχουν θεωρητικά τέτοιες πρωτεΐνες ή μπορεί η ομοιότητα που εντοπίσαμε να είναι σε μια μόνο μικρή περιοχή. Η γνώση των περιοχών των πρωτεϊνών είναι πολύ σημαντική. Δεν επιχειρούμε, παρά μόνο σε ειδικές περιπτώσεις και με κατάλληλο λογισμικό, να στοιχίσουμε πρωτεΐνες με μεγάλες αποκλίσεις στο μήκος (και άρα, με μεγάλες αποκλίσεις στη σύσταση των περιοχών τους). Αν έχουμε εντοπίσει την περιοχή που χρειαζόμαστε, καλό είναι να πραγματοποιούμε και τις αναζητήσεις μόνο με αυτήν. Σε κάθε περίπτωση, πληροφορία από αντίστοιχες βάσεις πρωτεϊνικών περιοχών (PFAM, PROSITE κλπ), θα είναι πολύ χρήσιμη, και συνίσταται να γίνεται έλεγχος σε αυτές τις βάσεις (το BLAST παρέχει μια τέτοια επιλογή παράλληλα με την αναζήτηση ομοιότητας).



Εικόνα 4.7: *Παραδείγματα λειτουργίας του Jalview. Πάνω απεικονίζεται μια πολλαπλή στοίχιση πρωτεϊνικών ακολουθιών και μια πολλαπλή στοίχιση RNA. Κάτω απεικονίζεται ένα φυλογενετικό δέντρο, οπτικοποίηση δομών με το Jmol και αναπαράσταση δευτεροταγούς δομής RNA.*



Εικόνα 4.8: *Παραδείγματα λειτουργίας του BioEdit. Το BioEdit εκτός από το ευέυλικτο περιβάλλον που δίνει ο Editor, παρέχει και τα περισσότερα εργαλεία ανάλυσης όπως: αναλύσεις υδοφοβικότητας (στο σχήμα), προγνωστικούς αλγόριθμους, οπτικοποίηση περιοριστικών χαρτών, διασυνδέσεις με πολλές βάσεις δεδομένων, εργαλεία στοίχισης και οπτικοποίησης στοιχίσεων (BLAST, dot plot κλπ), αναζητήσεις προτύπων, αμοιβαία πληροφορία, εργαλεία φυλογενετικής ανάλυσης, αλλά και δυνατότητα προσθήκης και άλλων εργαλείων από το χρήστη.*

Είτε εντοπίσουμε περιοχές ενδιαφέροντος με αυτούς τους τρόπους, είτε όχι, το επόμενο βήμα είναι επίσης σημαντικό. Θα πρέπει να κάνουμε οπωσδήποτε προγνώσεις δευτεροταγούς δομής ή/και άλλων δομικών και λειτουργικών χαρακτηριστικών που ενδέχεται να σχετίζονται με τη συγκεκριμένη πρωτεϊνική οικογένεια. Όπως είδαμε, η δευτεροταγής δομή είναι σημαντική γιατί μπορεί να μας δώσει εικόνα της καταλληλότητας της πολλαπλής στοίχισης, να εντοπίσει ασάφειες κλπ. Τα υπόλοιπα χαρακτηριστικά που πρέπει να προβλέψουμε, μπορεί να ποικίλουν ανάλογα με την περίπτωση. Για παράδειγμα, σε στοιχίσεις εξωκυττάριων πρωτεϊνών, θα ήταν καλό να αφαιρεθεί το πεπτίδιο οδηγητής (signal peptide), και για την ακρίβεια, θα ήταν σημαντικό να ξέρουμε αν έχουν όλες οι πρωτεϊνες μας ένα τέτοιο πεπτίδιο (το ίδιο ισχύει και για όλα τα άλλα προπεπτίδια ή πεπτίδια στόχευσης). Για κάποιες περιπτώσεις, πληροφορίες οι οποίες μπορεί να είναι χρήσιμες στην πολλαπλή στοίχιση μπορεί να μας δώσουν οι δισουλφιδικοί δεσμοί, αλλά και άλλες θέσεις μετα-μεταφραστικής τροποποίησης (γλυκοζυλίωση, φοσφωρυλίωση κ.ο.κ.). Τούτο συμβαίνει, γιατί οι περιοχές αυτές είναι πιθανό να συντηρούνται στην πολλαπλή στοίχιση, αλλά επιπλέον, με αυτόν τον τρόπο μπορούμε να «διορθώσουμε» τη στοίχιση, αν μια τέτοια θέση έχει τοποθετηθεί λάθος. Το τελευταίο βέβαια, είναι αρκετά επικίνδυνο, και πρέπει να γίνεται με μεγάλη προσοχή γιατί απαιτεί εμπειρία (ενώ πρέπει να θυμόμαστε ότι και οι αλγόριθμοι δεν είναι αλάνθαστοι!). Μια εναλλακτική και ίσως πιο συνετή στρατηγική είναι, σε περίπτωση εντοπισμού σφάλματος στη στοίχιση, να αφαιρούμε την ακολουθία που εμφανίζει το πρόβλημα και να επαναλαμβάνουμε τη στοίχιση χωρίς αυτήν. Κατόπιν, μπορούμε να δοκιμάσουμε να την προσθέσουμε εκ των υστέρων στη στοίχιση με χρήση profile alignment, λειτουργία που υποστηρίζουν τα περισσότερα σύγχρονα εργαλεία.

Βιβλιογραφία

Barton, G. J., & Sternberg, M. J. (1987). A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *J Mol Biol, 198*(2), 327-337.

Carrillo, H., & Lipman, D. (1988). The multiple sequence alignment problem in biology. *SIAM Journal on Applied Mathematics 48*(5), 1073-1082.

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res, 16*(22), 10881-10890.

Durbin, R., Eddy, S., Krogh, A., & Mithison, G. (1998). *Biological sequence analysis, probabilistic models of proteins and nucleic acids.*: Cambridge University Press.

Duret, L., & Abdeddaim, S. (2000). Multiple alignment for structural, functional, or phylogenetic analyses of homologous sequences. *Bioinformatics: Sequence, Structure, and Databanks*, 51-76.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics, 5*, 113.

Edgar, R. C., & Sjolander, K. (2004). A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics, 20*(8), 1301-1308.

Feng, D.-F., & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of molecular evolution, 25*(4), 351-360.

Fitch, W. M., & Margoliash, E. (1967). Construction of phylogenetic trees. *science, 155*(3760), 279-284.

Gonnet, G. H., Cohen, M. A., & Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science, 256*(5062), 1443-1445.

Gotoh, O. (1996). Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J Mol Biol, 264*(4), 823-838.

Higgins, D. G., Bleasby, A. J., & Fuchs, R. (1992). CLUSTAL V: improved software for multiple sequence alignment. *Computer applications in the biosciences: CABIOS, 8*(2), 189-191.

Kim, J., Pramanik, S., & Chung, M. J. (1994). Multiple sequence alignment using simulated annealing. *Comput Appl Biosci, 10*(4), 419-426.

Lassmann, T., & Sonnhammer, E. L. (2005). Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics, 6*, 298.

Lipman, D. J., Altschul, S. F., & Kececioglu, J. D. (1989). A tool for multiple sequence alignment. *Proceedings of the National Academy of Sciences, 86*(12), 4412-4415.

Magis, C., Taly, J. F., Bussotti, G., Chang, J. M., Di Tommaso, P., Erb, I., . . . Notredame, C. (2014). T-Coffee: Tree-based consistency objective function for alignment evaluation. *Methods Mol Biol, 1079*, 117-129.

Morgenstern, B. (2014). Multiple sequence alignment with DIALIGN. *Methods Mol Biol, 1079*, 191-202.

Notredame, C., & Higgins, D. G. (1996). SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res, 24*(8), 1515-1524.

O'Sullivan, O., Zehnder, M., Higgins, D., Bucher, P., Grosdidier, A., & Notredame, C. (2003). APDB: a novel measure for benchmarking sequence alignment methods without reference alignments. *Bioinformatics, 19 Suppl 1*, i215-221.

Pais, F. S., Ruy Pde, C., Oliveira, G., & Coimbra, R. S. (2014). Assessing the efficiency of multiple sequence alignment programs. *Algorithms Mol Biol, 9*(1), 4.

Papadopoulos, J. S., & Agarwala, R. (2007). COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics, 23*(9), 1073-1079.

Raghava, G. P., Searle, S. M., Audley, P. C., Barber, J. D., & Barton, G. J. (2003). OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics, 4*, 47.

Roshan, U. (2014). Multiple sequence alignment using Probcons and Probalign. *Methods Mol Biol, 1079*, 147-153.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution, 4*(4), 406-425.

Simossis, V. A., & Heringa, J. (2005). PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res, 33*(Web Server issue), W289-294.

Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2002). Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics, Chapter 2*, Unit 2 3.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research, 22*(22), 4673-4680.

Thompson, J. D., Linard, B., Lecompte, O., & Poch, O. (2011). A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One, 6*(3), e18093.

Thompson, J. D., Plewniak, F., & Poch, O. (1999). A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res, 27*(13), 2682-2690.

Van Walle, I., Lasters, I., & Wyns, L. (2005). SABmark--a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics, 21*(7), 1267-1268.

Wallace, I. M., O'Sullivan, O., & Higgins, D. G. (2005). Evaluation of iterative alignment algorithms for multiple alignment. *Bioinformatics, 21*(8), 1408-1414.

Wang, G., & Dunbrack, R. L., Jr. (2004). Scoring profile-to-profile sequence alignments. *Protein Sci, 13*(6), 1612-1626.

Waterman, M. S. (1995). *Introduction to Computational Biology*: Chapman and Hall, London.

Wu, S., & Manber, U. (1992). Fast text searching allowing errors. *Communications of the ACM 35*(10), 83-91.

Παράρτημα

Οι πιο γνωστές μορφές αρχείων πολλαπλής στοίχισης

Multi-FASTA

```
>sw:CD5R_BOVIN Q28199 Cyclin-dependent kinase 5 activator 1 precursor
MGTVLSLSPSYRKATLFEDGAATVGHYTAVQNSKNAKDKNLKRHSIISVLPWKRIVAVSA
KKKNSKKVQPNSSYQNNITHLNNENLKKSLSCANLSTFAQPPPAQPPAPPASQLSGSQTG
VSSSVKKAPHPAVSSAGTPKRVIVQASTSELLRCLGEFLCRRCYRLKHLSPTDPVLWLRS
VDRSLLLQGWQDQGFITPANVVFLYMLCRDVISSEVGSDHELQAVLLTCLYLSYSYMGNE
ISYPLKPFLVESCKEAFWDRCLSVINLMSSKMLQINADPHYFTQVFSDLKNESGQEDKKR
LLLGLDR
>sw:CD5R_HUMAN Q15078 Cyclin-dependent kinase 5 activator 1 precursor
MGTVLSLSPSYRKATLFEDGAATVGHYTAVQNSKNAKDKNLKRHSIISVLPWKRIVAVSA
KKKNSKKVQPNSSYQNNITHLNNENLKKSLSCANLSTFAQPPPAQPPAPPASQLSGSQTG
GSSSVKKAPHPAVTSAGTPKRVIVQASTSELLRCLGEFLCRRCYRLKHLSPTDPVLWLRS
VDRSLLLQGWQDQGFITPANVVFLYMLCRDVISSEVGSDHELQAVLLTCLYLSYSYMGNE
ISYPLKPFLVESCKEAFWDRCLSVINLMSSKMLQINADPHYFTQVFSDLKNESGQEDKKR
LLLGLDR
>sw:CD5R_MOUSE Q62938 Cyclin-dependent kinase 5 activator 1 precursor
MGTVLSLSPSYRKATLFEDGAATVGHYTAVQNSKNAKDKNLKRHSIISVLPWKRIVAVSA
KKKNSKKAQPNSSYQSNIAHLNNENLKKSLSCANLSTFAQPPPAQPPAPPASQLSGSQTG
VSSSVKKAPHPAITSAGTPKRVIVQASTSELLRCLGEFLCRRCYRLKHLSPTDPVLWLRS
VDRSLLLQGWQDQGFITPANVVFLYMLCRDVISSEVGSDHELQAVLLTCLYLSYSYMGNE
ISYPLKPFLVESCKEAFWDRCLSVINLMSSKMLQINADPHYFTQVFSDLKNESGQEDKKR
LLLGLDR
```

MSF

```
MSF: 307 Type: P Check: 4977 ..
Name: CD5R_BOVIN oo Len: 307 Check: 5281 Weight: 33.3
Name: CD5R_HUMAN oo Len: 307 Check: 5196 Weight: 33.3
Name: CD5R_MOUSE oo Len: 307 Check: 4500 Weight: 33.3
//
CD5R_BOVIN MGTVLSLSPS YRKATLFEDG AATVGHYTAV QNSKNAKDKN LKRHSIISVL
CD5R_HUMAN MGTVLSLSPS YRKATLFEDG AATVGHYTAV QNSKNAKDKN LKRHSIISVL
CD5R_MOUSE MGTVLSLSPS YRKATLFEDG AATVGHYTAV QNSKNAKDKN LKRHSIISVL
CD5R_BOVIN PWKRIVAVSA KKKNSKKVQP NSSYQNNITH LNNENLKKSL SCANLSTFAQ
CD5R_HUMAN PWKRIVAVSA KKKNSKKVQP NSSYQNNITH LNNENLKKSL SCANLSTFAQ
CD5R_MOUSE PWKRIVAVSA KKKNSKKAQP NSSYQSNIAH LNNENLKKSL SCANLSTFAQ
CD5R_BOVIN PPPAQPPAPP ASQLSGSQTG VSSSVKKAPH PAVSSAGTPK RVIVQASTSE
CD5R_HUMAN PPPAQPPAPP ASQLSGSQTG GSSSVKKAPH PAVTSAGTPK RVIVQASTSE
CD5R_MOUSE PPPAQPPAPP ASQLSGSQTG VSSSVKKAPH PAITSAGTPK RVIVQASTSE
CD5R_BOVIN LLRCLGEFLC RRCYRLKHLS PTDPVLWLRS VDRSLLLQGW QDQGFITPAN
CD5R_HUMAN LLRCLGEFLC RRCYRLKHLS PTDPVLWLRS VDRSLLLQGW QDQGFITPAN
CD5R_MOUSE LLRCLGEFLC RRCYRLKHLS PTDPVLWLRS VDRSLLLQGW QDQGFITPAN
CD5R_BOVIN VVFLYMLCRD VISSEVGSDH ELQAVLLTCL YLSYSYMGNE ISYPLKPFLV
CD5R_HUMAN VVFLYMLCRD VISSEVGSDH ELQAVLLTCL YLSYSYMGNE ISYPLKPFLV
CD5R_MOUSE VVFLYMLCRD VISSEVGSDH ELQAVLLTCL YLSYSYMGNE ISYPLKPFLV
CD5R_BOVIN ESCKEAFWDR CLSVINLMSS KMLQINADPH YFTQVFSDLK NESGQEDKKR
CD5R_HUMAN ESCKEAFWDR CLSVINLMSS KMLQINADPH YFTQVFSDLK NESGQEDKKR
CD5R_MOUSE ESCKEAFWDR CLSVINLMSS KMLQINADPH YFTQVFSDLK NESGQEDKKR
CD5R_BOVIN LLLGLDR
CD5R_HUMAN LLLGLDR
CD5R_MOUSE LLLGLDR
```

CLUSTAL

```
CLUSTAL W (1.82) multiple sequence alignment
CD5R_BOVIN MGTVLSLSPSYRKATLFEDGAATVGHYTAVQNSKNAKDKNLKRHSIISVLPWKRIVAVSA
CD5R_HUMAN MGTVLSLSPSYRKATLFEDGAATVGHYTAVQNSKNAKDKNLKRHSIISVLPWKRIVAVSA
CD5R_MOUSE MGTVLSLSPSYRKATLFEDGAATVGHYTAVQNSKNAKDKNLKRHSIISVLPWKRIVAVSA
           ************************************************************
CD5R_BOVIN KKKNSKKVQPNSSYQNNITHLNNENLKKSLSCANLSTFAQPPPAQPPAPPASQLSGSQTG
CD5R_HUMAN KKKNSKKVQPNSSYQNNITHLNNENLKKSLSCANLSTFAQPPPAQPPAPPASQLSGSQTG
CD5R_MOUSE KKKNSKKAQPNSSYQSNIAHLNNENLKKSLSCANLSTFAQPPPAQPPAPPASQLSGSQTG
           *******.*******.**:*****************************************
CD5R_BOVIN VSSSVKKAPHPAVSSAGTPKRVIVQASTSELLRCLGEFLCRRCYRLKHLSPTDPVLWLRS
CD5R_HUMAN GSSSVKKAPHPAVTSAGTPKRVIVQASTSELLRCLGEFLCRRCYRLKHLSPTDPVLWLRS
CD5R_MOUSE VSSSVKKAPHPAITSAGTPKRVIVQASTSELLRCLGEFLCRRCYRLKHLSPTDPVLWLRS
           ***********::*************************************************
CD5R_BOVIN VDRSLLLQGWQDQGFITPANVVFLYMLCRDVISSEVGSDHELQAVLLTCLYLSYSYMGNE
CD5R_HUMAN VDRSLLLQGWQDQGFITPANVVFLYMLCRDVISSEVGSDHELQAVLLTCLYLSYSYMGNE
CD5R_MOUSE VDRSLLLQGWQDQGFITPANVVFLYMLCRDVISSEVGSDHELQAVLLTCLYLSYSYMGNE
           ************************************************************
CD5R_BOVIN ISYPLKPFLVESCKEAFWDRCLSVINLMSSKMLQINADPHYFTQVFSDLKNESGQEDKKR
CD5R_HUMAN ISYPLKPFLVESCKEAFWDRCLSVINLMSSKMLQINADPHYFTQVFSDLKNESGQEDKKR
CD5R_MOUSE ISYPLKPFLVESCKEAFWDRCLSVINLMSSKMLQINADPHYFTQVFSDLKNESGQEDKKR
           ************************************************************
CD5R_BOVIN LLLGLDR
CD5R_HUMAN LLLGLDR
CD5R_MOUSE LLLGLDR
           *******
```

Κεφάλαιο 5: Αναζήτηση Προτύπων σε Αλληλουχίες

Σύνοψη
Στο κεφάλαιο αυτό θα μελετήσουμε τα πρότυπα αλληλουχιών και θα εξετάσουμε τη χρησιμότητά τους. Θα δούμε τον τρόπο ορισμού των προτύπων της PROSITE και τη σχέση τους με τα πρότυπα κανονικών εκφράσεων και θα συζητήσουμε τα πλεονεκτήματα και τα μειονεκτήματά τους. Κατόπιν, θα αξιολογήσουμε πώς κάποια από αυτά τα μειονεκτήματα αντιμετωπίζονται με τους πίνακες του σκορ ειδικούς ανά θέση (PSSMs) και τα προφίλ αλληλουχιών (profiles), τα οποία είναι πιο ευέλικτες στατιστικές περιγραφές των συντηρημένων περιοχών σε μια πολλαπλή στοίχιση. Τέλος, θα μιλήσουμε και για τα πιο γνωστά εργαλεία λογισμικού που χρησιμοποιούνται για την κατασκευή αλλά και για την αναγνώριση τέτοιων προτύπων και προφίλ σε αλληλουχίες.

Προαπαιτούμενη γνώση
Το κεφάλαιο απαιτεί κατανόηση των μεθόδων του κεφαλαίου 3 και του κεφαλαίου 4.

5. Εισαγωγή

Στο κεφάλαιο αυτό, αφού έχουμε ήδη μελετήσει τη στοίχιση και την πολλαπλή στοίχιση αλληλουχιών, θα μελετήσουμε το επόμενο πρόβλημα που προκύπτει: τον τρόπο με τον οποίο θα περιγράψουμε μαθηματικά μια πολλαπλή στοίχιση, με σκοπό να πάρουμε μια πιο συμπυκνωμένη αναπαράσταση της πληροφορίας που περιέχεται στις συντηρημένες περιοχές. Έτσι, θα δούμε στην αρχή τα πρότυπα ακολουθιών (patterns) και τον τρόπο με τον οποίο αυτά περιγράφονται στη λεγόμενη μορφή της PROSITE, ενώ παράλληλα θα δούμε και τις αναλογίες με τις κανονικές εκφράσεις (regular expressions) του UNIX. Αφού μελετήσουμε αναλυτικά τα πρότυπα, θα ασχοληθούμε και με τις αδυναμίες τους, και κατά συνέπεια την ανάγκη για πιο ακριβείς περιγραφές στις οποίες δεν θα υπάρχει απώλεια πληροφορίας. Έτσι, θα μιλήσουμε και για τα προφίλ (profiles) και τους πίνακες σκορ ειδικούς ανά θέση (Position Specific Scoring Matrices). Θα δούμε, ότι με αυτές τις περιγραφές δεν διευκολύνεται μόνο η αναζήτηση συντηρημένων περιοχών σε βάσεις δεδομένων, αλλά επιπλέον, ανοίγει και δρόμος για πιο ευαίσθητες αναζητήσεις και εντοπισμό μακρινών ομολόγων. Για όλα τα παραπάνω θέματα, θα μελετήσουμε επίσης τους αλγόριθμους που χρησιμοποιούνται για την κατασκευή αυτών των αναπαραστάσεων (προτύπων ή προφίλ), αλλά και τα εργαλεία λογισμικού που υπάρχουν διαθέσιμα για το σκοπό αυτό.

5.1. Πρότυπα και μοτίβα αλληλουχιών

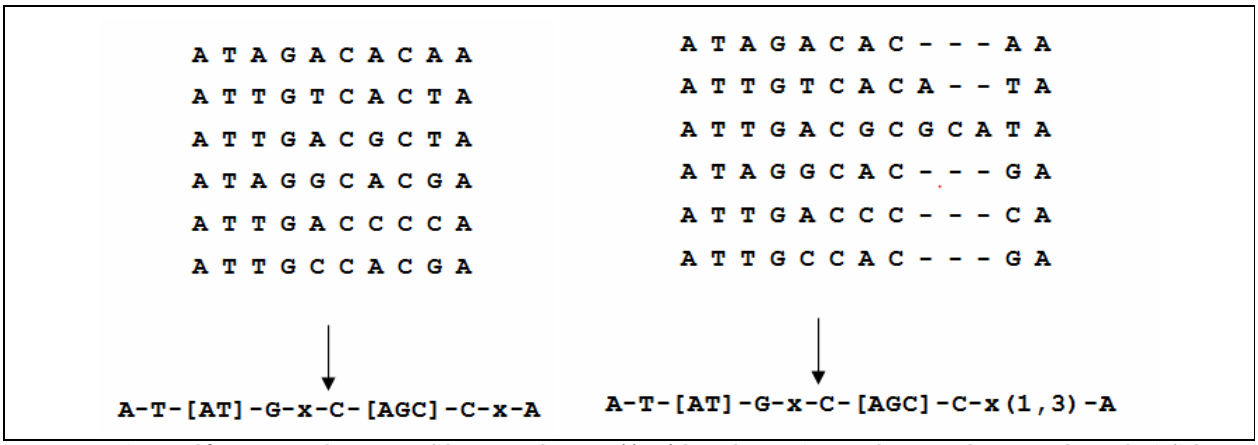
5.1.1 Τι είναι τα πρότυπα

Όταν έχουμε μια καλά προσδιορισμένη πολλαπλή στοίχιση αλληλουχιών (είτε πρωτεϊνών, είτε νουκλεϊκών οξέων), ένα θέμα που μας ενδιαφέρει είναι να μπορούμε να εξάγουμε μια πιο πληροφοριακή περιγραφή της. Είδαμε, για παράδειγμα, ότι μια τέτοια πολλαπλή στοίχιση μπορεί να περιγράφει μια πρωτεϊνική οικογένεια, δηλαδή μια ομάδα πρωτεϊνών με κοινή εξελικτική ιστορία, οι οποίες έχουν κοινά δομικά και πιθανώς και λειτουργικά χαρακτηριστικά. Έτσι, θα μας ενδιέφερε να βρούμε έναν τρόπο να περιγράψουμε τα κοινά χαρακτηριστικά όλων αυτών των αλληλουχιών, και να έχουμε μια εύκολη και κατανοητή περιγραφή, χωρίς να χρειάζεται κάθε φορά να ανατρέχουμε στην ίδια την πολλαπλή στοίχιση η οποία μπορεί να είναι μεγάλη αλλά και δυσνόητη. Επίσης, θα μας ενδιέφερε να βρούμε με βάση αυτή την περιγραφή, έναν εύκολο και γρήγορο τρόπο για να πραγματοποιήσουμε μια αναζήτηση στις βάσεις δεδομένων, για αλληλουχίες που έχουν αυτό το κοινό χαρακτηριστικό, χωρίς όμως να χρειάζεται να πραγματοποιήσουμε εκ νέου στοίχιση.

Οι περιγραφές αυτές, ονομάζονται πρότυπα (patterns) και έχουν μεγάλη ιστορία στην επιστήμη των υπολογιστών για την περιγραφή κειμένων και άλλων ειδών ακολουθιών από χαρακτήρες (είναι τα γνωστά regular expressions). Με τις περιγραφές αυτές, μπορούμε να δούμε σε ποια θέση μιας πολλαπλής στοίχισης υπάρχει μεγάλη ή μικρότερη συντήρηση και έτσι να χαρακτηρίσουμε και να εντοπίσουμε μεταξύ άλλων ενεργά κέντρα, περιοχές δράσης των ενζύμων και θέσεις δισουλφιδικών δεσμών (στις πρωτεΐνες) ή υποκινητές, θέσεις έναρξης γονιδίων και σημεία συρραφής εξωνίων (στα γονίδια). Στη βιοπληροφορική ο παραδοσιακός τρόπος χρήσης τέτοιων εκφράσεων είναι με τα λεγόμενα πρότυπα της PROSITE, τα οποία είναι μεν εντελώς ανάλογα με τις κανονικές εκφράσεις του UNIX αλλά έχουν μια σύνταξη λίγο πιο «εύκολη» και κατανοητή.

Στα πρότυπα αυτά (Εικόνα 5.1), ολόκληρη η πολλαπλή στοίχιση ή για την ακρίβεια, οι στήλες της που εμφανίζουν το μεγαλύτερο ενδιαφέρον, περιγράφονται με μία συμπυκνωμένη έκφραση. Αφενός μεν αυτό προσδίδει μια τεράστια ευκολία καθώς μια πολλαπλή στοίχιση πιθανά εκατοντάδων αλληλουχιών συνοψίζεται σε μία γραμμή, αφετέρου δε, αυτή η διαδικασία αναπόφευκτα οδηγεί σε απώλεια πληροφορίας (θα επανέλθουμε σε αυτό). Τα βασικά χαρακτηριστικά της σύνταξης PROSITE είναι τα παρακάτω:

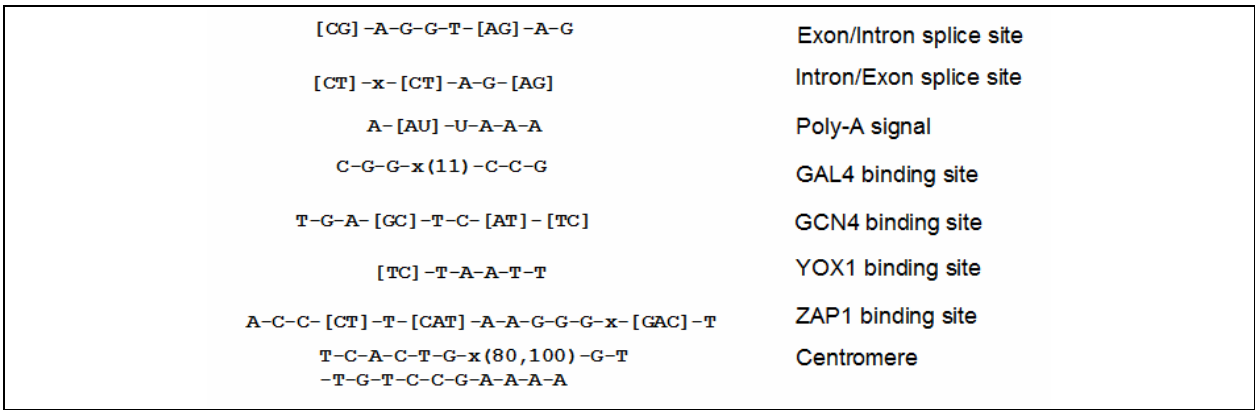
- Τα αμινοξέα ή τα νουκλεοτίδια αναπαρίστανται με τον τυπικό κωδικό του ενός γράμματος της IUPAC.
- Κάθε θέση της πολλαπλής στοίχισης αντιστοιχεί σε μια θέση στο πρότυπο, η οποία διαχωρίζεται από τις υπόλοιπες με μία παύλα (-).
- Οι θέσεις είναι ανεξάρτητες μεταξύ τους.
- Αν σε κάποια θέση εμφανίζεται μόνο ένας χαρακτήρας, τότε στο πρότυπο χρησιμοποιείται αυτούσιος (π.χ. A, T κ.ο.κ.)
- Αν σε κάποια θέση εμφανίζονται δύο ή περισσότεροι χαρακτήρες τότε αυτοί εμφανίζονται μέσα σε άγκιστρο, για παράδειγμα [AT] σημαίνει ότι επιτρέπεται A ή T, ενώ [ACG] σημαίνει ότι επιτρέπεται είτε A, είτε G, είτε C.
- Αν σε κάποια θέση επιτρέπεται να εμφανιστεί οποιοδήποτε σύμβολο, τότε αυτή η θέση συμβολίζεται με x.
- Αν σε κάποια θέση επιτρέπεται να εμφανιστεί οποιοδήποτε σύμβολο εκτός από κάποιο/α, τότε τη θέση τη συμβολίζουμε με {}. Για παράδειγμα, για να πούμε «οποιοδήποτε νουκλεοτίδιο εκτός από A» γράφουμε {A} το οποίο στην περίπτωση του DNA είναι ισοδύναμο με το [CGT]. Προφανώς, αυτός ο κανόνας είναι περισσότερο χρήσιμος στην περίπτωση των πρωτεϊνών με το μεγάλο αλφάβητο.
- Επαναλήψεις συμβολίζονται με παρένθεση () μετά από ένα σύμβολο. Για παράδειγμα το A(3) σημαίνει A-A-A, ενώ το x(3) σημαίνει x-x-x (δηλαδή 3 οποιαδήποτε σύμβολα). Επίσης, μέσα στην παρένθεση μπορεί να μπει και ένα εύρος τιμών. Έτσι, το x(2,4) σημαίνει x-x, ή x-x-x, ή x-x-x-x.
- Η αρχή και το τέλος της αλληλουχίας συμβολίζονται με τα σύμβολα < και > αντίστοιχα. Έτσι, για να πούμε ότι η αλληλουχία αρχίζει με A και μετά ακολουθεί οποιοδήποτε σύμβολο γράφουμε <A-x
- Σε κάποιες ειδικές περιπτώσεις το σύμβολο '>' μπορεί να εμφανιστεί μέσα στα άγκιστρα για να χαρακτηρίσει την πιθανή ύπαρξη καρβοξυτελικού άκρου. Έτσι, το P-R-L-[G>] σημαίνει είτε P-R-L-G ή P-R-L>.



Εικόνα 5.1: *Παράδειγμα προτύπου που εξάγεται από μια πολλαπλή στοίχιση. Αριστερά μια στοίχιση χωρίς κενά. Δεξιά μια στοίχιση με κενά. Στην τελευταία περίπτωση θα πρέπει να αποφασίσουμε ποιες στήλες δεν θα αντιπροσωπευθούν στο πρότυπο.*

Συνήθως στις αλληλουχίες των γονιδίων, τέτοια πρότυπα, δηλαδή πολύ συντηρημένες περιοχές, συναντάμε στις αλληλουχίες των υποκινητών, στις θέσεις αποκοπής των εσωνίων από τα εξώνια κ.ο.κ (Εικόνα 5.2). Στις αλληλουχίες πρωτεϊνών, τέτοιες περιοχές χαρακτηρίζουν τις θέσεις δράσης ενζύμων, τα

ενεργά κέντρα των ενζύμων ή κάποιες πολύ χαρακτηριστικές περιοχές της δευτεροταγούς δομής όπως για παράδειγμα τις θέσεις κυστεϊνών που σχηματίζουν δισουλφιδικούς δεσμούς (Εικόνα 5.3).



Εικόνα 5.2: *Μερικά παραδείγματα γνωστών προτύπων που εμφανίζονται σε αλληλουχίες DNA.*

Όπως είδαμε στο Κεφάλαιο 2, η **PROSITE** (http://www.expasy.ch/prosite/) αποτελεί μια βάση ταξινόμησης πρωτεϊνικών ακολουθιών και αυτοτελών περιοχών ακολουθιών (sequence domains) σε οικογένειες (Sigrist et al., 2010). Ο παραδοσιακός τρόπος καταχώρησης μιας οικογένειας στη βάση αυτή, γίνεται με τους ομώνυμους κανόνες που περιγράψαμε παραπάνω και είναι ο πιο παλιός και εύκολος στη δημιουργία, ενώ ο άλλος βασίζεται στην κατασκευή προφίλ, μέθοδος η οποία είναι πιο σύνθετη αλλά και πιο ευαίσθητη (θα μελετηθεί στη συνέχεια). Μέχρι σήμερα η PROSITE περιέχει καταχωρήσεις για περισσότερες από 1700 οικογένειες. Συνολικά, υπάρχουν στη βάση 1308 πρότυπα, 1107 προφίλ και 1105 "κανόνες" (αφορούν κυρίως πληροφορίες για το πού θα πρέπει να βρίσκεται το πρότυπο για να θεωρηθεί έγκυρο αλλά και πληροφορίες για συνδυασμούς από πρότυπα). Προφανώς, υπάρχουν οικογένειες για τις οποίες υπάρχουν διαθέσιμα και πρότυπα και προφίλ (συνήθως, οι παλαιότερες καταχωρήσεις αφορούσαν το πρότυπα). Στη βάση υπάρχουν επίσης αναλύσεις τόσο για τις πρωτεΐνες της Uniprot που ανήκουν σε κάθε οικογένεια, όσο και για τις πρωτεΐνες στις οποίες εμφανίζεται ένα "αποτύπωμα" (κυρίως όταν έχουμε να κάνουμε με πρότυπα) αλλά είναι γνωστό ότι δεν ανήκουν λειτουργικά στην οικογένεια αυτή. Τέλος, υπάρχουν εργαλεία για την αναζήτηση των προτύπων και των προφίλ σε ακολουθίες, όσο και εργαλεία αναπαράστασης της "σπονδυλωτής" δομής των πρωτεϊνών, δηλαδή της αναπαράστασης των περιοχών αυτών και την αποτύπωση της διάταξής τους πάνω σε μια δεδομένη ακολουθία.

Όπως αναφέραμε ήδη, οι κανονικές εκφράσεις (regular expressions) και οι εκφράσεις της PROSITE είναι ισοδύναμες. Οι διαφορές στη σύνταξη είναι οι εξής:

- Η κάθε θέση αναγράφεται συνεχόμενα χωρίς να μεσολαβεί η παύλα (-).
- Το σύμβολο για «οποιοδήποτε» χαρακτήρα είναι η τελεία (.) αντί για το x
- Το σύμβολο για το «οποιοδήποτε χαρακτήρα εκτός από» είναι το ^ μέσα στην αγκύλη, και όχι το άγκιστρο {}.

Για παράδειγμα, αν θεωρήσουμε το πρότυπο της PROSITE που δίνεται από την έκφραση:

```
[RK]-G-{EDRKHPCG}-[AGSCI]-[FY]-[LIVA]-x-[FYM]
```

τότε η αντίστοιχη κανονική έκφραση θα είναι:

```
[RK]G[^EDRKHPCG][AGSCI][FY][LIVA].[FYM]
```

Κάτι που επίσης πρέπει να τονιστεί, είναι ότι αν και τα πρότυπα αυτά χρησιμοποιούνται εντατικά από τη δεκαετία του 1980 για τον χαρακτηρισμό οικογενειών πρωτεϊνών, και παρά το γεγονός ότι η PROSITE περιέχει πλήθος τέτοιων καταχωρίσεων, η υπόθεση εύρεσης και χαρακτηρισμού προτύπων τα οποία θα μπορούν να χρησιμοποιηθούν για την πρόγνωση δομικών και λειτουργικών χαρακτηριστικών των πρωτεϊνών δεν έχει σταματήσει καθόλου, καθώς τέτοια πρότυπα ανακαλύπτονται συνεχώς. Στην Εικόνα 5.3 βλέπουμε μόνο μερικά από τα εκατοντάδες σχετικά πρότυπα που είναι γνωστά εδώ και πολλά χρόνια. Παρ' όλα αυτά, στην Εικόνα 5.4 βλέπουμε κάποια άλλα πρότυπα, τα οποία έχουν ανακαλυφθεί μέσα στα τελευταία 15 χρόνια.

Για παράδειγμα, τα σήματα πυρηνικού εντοπισμού (nuclear localization signals - NLSs) είναι μικρές αλληλουχίες, γνωστές από παλιά, πλούσιες σε Αργινίνη και Λυσίνη, οι οποίες είναι υπεύθυνες για τη μεταφορά των πρωτεϊνών στον πυρήνα του κυττάρου. Οι Cocol, Nair και Rost, πραγματοποίησαν μια

εκτεταμένη ανάλυση στις γνωστές πυρηνικές πρωτεΐνες και εντόπισαν 214 επιπλέον τέτοια πρότυπα (πέραν των 91 που ήταν ήδη γνωστά) (Cokol, Nair, & Rost, 2000). Ένα άλλο παράδειγμα, αφορά τα σήματα μεταφοράς των πρωτεϊνών στα υπεροξεισώματα. Το πρώτο είδος σήματος στόχευσης των υπεροξεισωμάτων (peroxisomal targeting signal -PTS1) ήταν γνωστό εδώ και χρόνια (το καρβοξυτελικό S-K-L). Το 2004 όμως, ανακαλύφθηκε και ένας δεύτερος μηχανισμός ο οποίος έκανε χρήση ενός αμινοτελικού πεπτιδίου και οι Petriv και συνεργάτες εντόπισαν το πρότυπο που το περιγράφει, το οποίο και ονόμασαν PTS2 (Petriv, Tang, Titorenko, & Rachubinski, 2004).

C-x-C-x(2)-{V}-x(2)-G-{C}-x-C	EGF-like 1 domain
[RK]-x(2,3)-[DE]-x(2,3)-Y	Tyrosine kinase phosphorylation site
N-{P}-[ST]	N-linked glycosylation
[LIVMA]-G-[EQ]-H-G-[DN]-[ST]	L-lactate dehydrogenase active site
P-[LIVM]-C-T-[LIVM]-[KRH]-x-[FT]-P	Ubiquitin-activating enzyme signature
S-K-L>	Peroxisomal Target Sequence 1 (PTS1)
{DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C	Bacterial Lipoprotein signal peptide

Εικόνα 5.3: *Μερικά από τα γνωστά παραδείγματα προτύπων που εμφανίζονται σε πρωτεΐνες.*

Μια άλλη πολύ γνωστή περίπτωση, είναι αυτή των βακτηριακών λιποπρωτεϊνών. Οι πρωτεΐνες αυτές έχουν μια σηματοδοτική αλληλουχία (signal peptide) η οποία μοιάζει αρκετά με αυτή των εκκρινόμενων πρωτεϊνών, αλλά στο καρβοξυτελικό της άκρο φέρει μια χαρακτηριστική αλληλουχία η οποία αναγνωρίζεται από ειδικό ένζυμο, το οποίο αποκόπτει το πεπτίδιο αυτό και ακολούθως η ώριμη πρωτεΐνη προσκολλάται στα λιπίδια της μεμβράνης με ομοιοπολικό δεσμό. Η αλληλουχία που αναγνωρίζει το ένζυμο, έχει μια συντηρημένη κυστεΐνη στο καρβοξυτελικό της άκρο (περίπου στη θέση 17-30 της πρόδρομης πρωτεΐνης, εκεί που γίνεται και η τροποποίηση), ενώ στις προηγούμενες θέσεις υπάρχουν κυρίως Αλανίνες και Βαλίνες. Τέτοια πρότυπα είχαν περιγραφεί από τη δεκαετία του 1980, αλλά το πιο γνωστό είναι το λεγόμενο PS00013, όπως ήταν γνωστό από τον κωδικό της PROSITE (Εικόνα 5.3). Παρ' όλα αυτά, έχουν περιγραφεί και εναλλακτικά πρότυπα πολλές φορές ακόμα και χρόνια αργότερα, όπως το [LVI]-[ASTVI]-[GAS]-C το οποίο χρησιμοποιήθηκε για να κατασκευαστεί η βάση των βακτηριακών λιποπρωτεϊνών (DOLOP). Το 2002 επίσης, οι Sutcliffe και Harrington μελετώντας λιποπρωτεΐνες από βακτήρια θετικά κατά Gram, κατέληξαν σε ένα πιο αυστηρό αλλά ταυτόχρονα και πιο περιεκτικό πρότυπο, το οποίο περιγράφει καλύτερα τις λιποπρωτεΐνες αυτών των βακτηρίων (Sutcliffe & Harrington, 2002). Το πρότυπο αυτό δίνεται (μαζί με άλλα παραδείγματα) στην Εικόνα 5.4.

Μια πιο πρόσφατη εργασία όμως, αφορά τις λιποπρωτεΐνες που εκκρίνονται με το σύστημα TAT (twin-arginine translocation). Το σύστημα αυτό, υπάρχει σε όλα τα βακτήρια αλλά και τους χλωροπλάστες και εκκρίνει πρωτεΐνες με ένα σύστημα διαφορετικό από το γνωστό εκκριτικό μονοπάτι SEC. Για την ακρίβεια, οι πρωτεΐνες εκκρίνονται διπλωμένες στην τρισδιάστατη δομή τους, μέσω ενός διαμεμβρανικού υποδοχέα που λειτουργεί με έναν άγνωστο προς το παρόν μηχανισμό. Από άποψη αλληλουχίας, οι πρωτεΐνες αυτές φέρουν ένα αμινοτελικό πεπτίδιο (σηματοδοτική αλληλουχία), που μοιάζει πάρα πολύ με το κλασικό πεπτίδιο έκκρισης αλλά έχει ένα συντηρημένο πρότυπο που αποτελείται από 2 συνεχόμενες Αργινίνες (R-R-x-[FGAVML]-[LITMVF]). Για τις πρωτεΐνες αυτές έχουν αναπτυχθεί βέβαια πιο ειδικές μέθοδοι πρόγνωσης. Παρ' όλα αυτά, τα τελευταία χρόνια υπήρξαν πειραματικά δεδομένα που έδειχναν ότι υπάρχουν και περίεργες περιπτώσεις, δηλαδή πρωτεΐνες που εκκρίνονται με το TAT αλλά η ώριμη πρωτεΐνη δεν απελευθερώνεται, αντιθέτως προσκολλάται στη μεμβράνη, όπως μια λιποπρωτεϊνη. Με άλλα λόγια, υπάρχουν λιποπρωτεΐνες που χρησιμοποιούν το σύστημα TAT για την έκκριση και όχι το SEC. Έτσι, το 2010 οι Shruthi, Babu και Sankaran, χρησιμοποίησαν αυτές τις απλές παρατηρήσεις και με χρήση μόνο αυτών των δύο απλών προτύπων (R-R-x-[FGAVML]-[LITMVF] για τις πρωτεΐνες TAT και [LVI]-[ASTVI]-[GAS]-C για τις λιποπρωτεΐνες), εντόπισαν όλες τις πιθανές TAT-λιποπρωτεΐνες που υπάρχουν στα βακτηριακά γονιδιώματα και μελέτησαν τις πιθανές λειτουργίες τους (Shruthi, Babu, & Sankaran, 2010).

Τέλος, ένα άλλο σχετικά πρόσφατο παράδειγμα, προέρχεται από την πρόγνωση των διαμεμβρανικών β-βαρελιών των αρνητικών κατά Gram βακτηρίων. Το 2004, οι Berven και συνεργάτες, εντόπισαν ένα συντηρημένο πρότυπο που περιγράφει τις περισσότερες καρβοξυτελικές περιοχές των β-βαρελιών (Berven, Flikka, Jensen, & Eidhammer, 2004). Το πρότυπο αυτό, δεν είναι βέβαια ικανό από μόνο του να διαχωρίσει όλες τις αντίστοιχες πρωτεΐνες, αλλά χρησιμοποιήθηκε σε συνδυασμό με άλλες μεθοδολογίες για την κατασκευή του αλγορίθμου BOMP.

Σε κάθε περίπτωση, αυτό που πρέπει να γίνει κατανοητό από τα παραπάνω, είναι ότι ναι μεν η μεθοδολογία αυτή είναι εξαιρετικά απλή και εύκολη στη χρήση, αλλά ακόμα και σήμερα υπάρχουν περιθώρια για τη χρήση της και πολλά χρήσιμα βιολογικά συμπεράσματα μπορούν να εξαχθούν από αυτή.

[RK]-[LVI]Q]-x(2)-[LVIHQ] -[LSGAK]-x-[HQ]-[LAF]	Peroxisomal Target Sequence 2 (PTS2)
K-R-K-x{11}-K-K-K-S-K-K	Nuclear localization signal (*)
[LVI]-[ASTVI]-[GAS]-C	Alternative bacterial Lipoprotein signal peptide pattern
<[MV]-X(0,13)-[RK]-{DERKQ}(6,20) -[LIVMFESTAG]-[LVIAM]-[IVMSTAFG]-[AG]-C	Pattern specific for lipoproteins of Gram+ bacteria
R-R-x-[FGAVML]-[LITMVF]	Twin-arginine (TAT) signal peptide
x(100,)-{C}-[YFWKLHVITMAD]-{C} -[YFWKLHVITMAD]-{C}-[YFWKLHVITMAD] -{C}-[YFWKLHVITMAD]-{C}-[FYW]	C-terminal beta-strand pattern of bacterial OMPs

Εικόνα 5.4: *Κάποια επιλεγμένα παράδειγμα προτύπων σε πρωτεϊνικές αλληλουχίες, τα οποία ανακαλύφθηκαν μέσα στα τελευταία 15 χρόνια.*

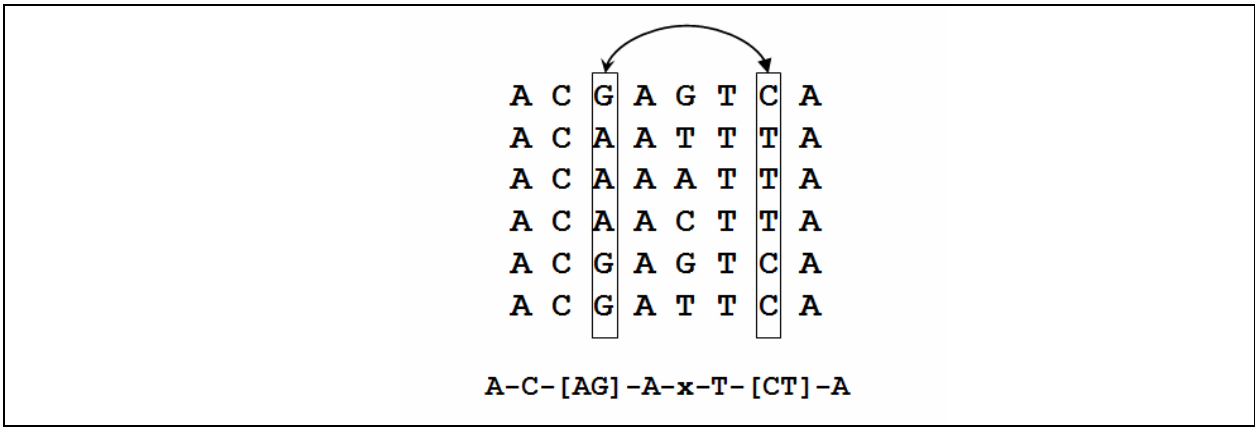
5.1.2 Πλεονεκτήματα και μειονεκτήματα των προτύπων

Τα πρότυπα, έχουν κάποια μοναδικά πλεονεκτήματα. Καταρχάς, είναι κατανοητά στο ανθρώπινο μάτι. Διαβάζοντας μια τέτοια έκφραση, καταλαβαίνουμε αμέσως την πληροφορία που περιέχει. Έτσι, είναι πολύ περιεκτικά και συμπυκνώνουν την πληροφορία μιας πιθανά μεγάλης πολλαπλής στοίχισης, μέσα σε μερικούς μόνο χαρακτήρες. Μας βοηθούν με αυτόν τον τρόπο να ταξινομήσουμε και να κατανοήσουμε φαινόμενα που είναι γενικά δύσκολα. Επίσης, είναι ιδιαίτερα αποδοτικά από υπολογιστικής πλευράς για πρακτικές χρήσης. Τα πρότυπα PROSITE καθώς είναι ισοδύναμα με τις κανονικές εκφράσεις (regular expressions), μπορούν να βασιστούν στις υλοποιήσεις που κάνουν χρήση πεπερασμένων αυτομάτων με συνέπεια να είναι ιδιαίτερα εύκολο και γρήγορο το να αποτελέσουν τμήμα μια υπολογιστικής μεθοδολογίας για ταχείες αναζητήσεις σε μεγάλες βάσεις δεδομένων. Στο κεφάλαιο 12 θα δούμε ότι η υλοποίηση τέτοιων εκφράσεων σε μια γλώσσα προγραμματισμού όπως η Perl είναι κάτι ιδιαίτερα εύκολο, ενώ αντίστοιχες δυνατότητες δίνουν ακόμα και οι βασικές εντολές του UNIX (grep, egrep).

Από την άλλη μεριά όμως, αυτά ακριβώς τα χαρακτηριστικά που κάνουν τα πρότυπα ιδιαίτερα επιτυχημένα, περιέχουν και το σπόρο με τις αδυναμίες τους. Το βασικό μειονέκτημα είναι ότι χάνεται μεγάλο μέρος της πληροφορίας της πολλαπλής στοίχισης. Για παράδειγμα στην Εικόνα 5.1 στην $3^{η}$ στήλη της στοίχισης το πρότυπο προβλέπει [AT], δηλαδή Α ή Τ, αλλά δεν μας δίνει τη σχετική πιθανότητα για το καθένα, παρόλο που από την πολλαπλή στοίχιση βλέπουμε ότι η Θυμίνη (Τ) έχει διπλάσια πιθανότητα από την Αδενίνη (Α). Φανταστείτε ότι έχουμε τώρα την ίδια περίπτωση αλλά σε μια στοίχιση με 100 αλληλουχίες, και εκεί έχουμε 65 Τ και 35 Α. Αν τώρα γίνει γνωστή μια επιπλέον αλληλουχία που ανήκει σίγουρα (με βάση βιολογικά κριτήρια) στη συγκεκριμένη οικογένεια, αλλά στη θέση αυτή έχει G, τι θα πρέπει να γίνει σε αυτή την περίπτωση; Αν το πρότυπο διευρυνθεί για να περιλαμβάνει και τη νέα αλληλουχία (γίνει δηλαδή [AGT]), τότε θα έχουμε χάσει ακόμα μεγαλύτερο μέρος της προβλεπτικής δύναμης. Αν επιλέξουμε να μην κάνουμε αυτή τη διεύρυνση, τότε θα είμαστε αναγκασμένοι να έχουμε ένα πρότυπο το οποίο «χάνει» κάποια από τα πραγματικά μέλη της οικογένειας. Αυτό είναι ένα πραγματικό πρόβλημα, και υπάρχουν και στη βάση

PROSITE πρότυπα τα οποία αδυνατούν να χαρακτηρίσουν το 100% των μελών μιας πρωτεϊνικής οικογένειας. Προφανώς, στην περίπτωση των πρωτεϊνών το πρόβλημα είναι πολύ πιο έντονο καθώς όπως είδαμε στα προηγούμενα κεφάλαια, σε πρωτεϊνικές οικογένειες με πολλά μέλη είναι σχεδόν αδύνατο να βρεις στήλες στην πολλαπλή στοίχιση με απόλυτη ομοφωνία καθώς αυτό που συντηρείται τις περισσότερες φορές είναι οι φυσικοχημικές ιδιότητες (πχ υδρόφοβα αμινοξέα, θετικά φορτισμένα αμινοξέα κ.ο.κ.). Με λίγα λόγια, είναι πολύ συνηθισμένο μια καλή στοίχιση να περιλαμβάνει σε μια στήλη αρκετά, διαφορετικά μεταξύ τους, αμινοξέα. Το πρόβλημα αυτό, θα το λύσουν εν μέρει τα προφίλ αλληλουχιών (sequence profiles) και οι ειδικοί ανά θέση πίνακες σκορ (PSSMs), τους οποίους θα δούμε στην επόμενη ενότητα.

Ένα άλλο πρόβλημα, είναι ότι τα πρότυπα με τον τρόπο που τα ορίσαμε δεν μπορούν να ενσωματώσουν εύκολα τα κενά στην πολλαπλή στοίχιση. Στην Εικόνα 5.1 είδαμε μια πολλαπλή στοίχιση που περιέχει κενά, αλλά όλα προέρχονται από εισαγωγές (τυχαίων) νουκλεοτιδίων σε κάποιες από τις αλληλουχίες της στοίχισης. Έτσι, τα κενά στην $1^{η}$, $4^{η}$, $5^{η}$ και $6^{η}$ αλληλουχία αντιστοιχούν απλά στις εισαγωγές νουκλεοτιδίων στην $2^{η}$ και στην $3^{η}$ αλληλουχία. Τι θα γινόταν όμως αν λ.χ. στην πρώτη αλληλουχία στην $8^{η}$ θέση δεν είχε την Κυτοσίνη (C); Με την υπάρχουσα ορολογία, απλά δεν θα ταίριαζε στο μοντέλο. Το πρόβλημα αυτό το λύνουν εν μέρει τα προφίλ, αντιμετωπίζοντάς το με τον κλασικό τρόπο που είδαμε στη στοίχιση αλληλουχιών (με δυναμικό προγραμματισμό και ποινές για τα κενά), αλλά την πιο ολοκληρωμένη λύση τη δίνουν τα Hidden Markov Models (HMMs) που θα δούμε στο κεφάλαιο 8.



Εικόνα 5.5: *Ένα παράδειγμα πολλαπλής στοίχισης με εξάρτηση μεταξύ 2 γειτονικών θέσεων.*

Τέλος, υπάρχει και ένα μεγαλύτερο πρόβλημα, το οποίο όμως είναι και πιο δύσκολο να εντοπιστεί αλλά και να διορθωθεί. Ο τρόπος που αντιμετωπίζουν τα πρότυπα τις θέσεις της πολλαπλής στοίχισης, είναι σαν να πρόκειται για ανεξάρτητες παρατηρήσεις. Στην Εικόνα 5.5 βλέπουμε μια πολλαπλή στοίχιση με το αντίστοιχο πρότυπο στην οποία υπάρχει ισχυρή συσχέτιση (δηλαδή, αλληλεπίδραση) μεταξύ της στήλης 3 και της στήλης 7. Αν εξετάσουμε κάθε στήλη ξεχωριστά, βλέπουμε ότι στην 3 έχουμε 50% G και 50% A, ενώ στην 7 έχουμε 50% T και 50% C. Το πρότυπο PROSITE θα έδινε για παράδειγμα την ίδια πιθανότητα να εμφανιστεί G ($3^{η}$) και C ($7^{η}$), και να εμφανιστεί G ($3^{η}$) και T ($7^{η}$). Παρατήρηση όμως των συχνοτήτων των δινουκλεοτιδίων, μας δείχνει ότι όταν υπάρχει G ($3^{η}$) υπάρχει πάντα C ($7^{η}$), ενώ όταν υπάρχει A ($3^{η}$) πάντα ακολουθείται από T ($7^{η}$). Αυτή η εξάρτηση, είναι κάτι ιδιαίτερα δύσκολο να μοντελοποιηθεί, καθώς όλες οι μεθοδολογίες που έχουμε δει μέχρι τώρα κάνουν λόγο για ανεξάρτητες θέσεις, ενώ το ίδιο ισχύει και για τα προφίλ που θα δούμε παρακάτω αλλά και για τα HMM που είναι η γενίκευση τους. Μεθοδολογίες που θα μπορούσαν με διαφορετικό τρόπο η καθεμιά να αντιμετωπίσουν αυτό το πρόβλημα, περιλαμβάνουν τα μαρκοβιανά μοντέλα εξάρτησης (Κεφάλαιο 8), τα Νευρωνικά Δίκτυα (Κεφάλαιο 7), αλλά και τις στοχαστικές γραμματικές χωρίς συμφραζόμενα (Κεφάλαιο 10).

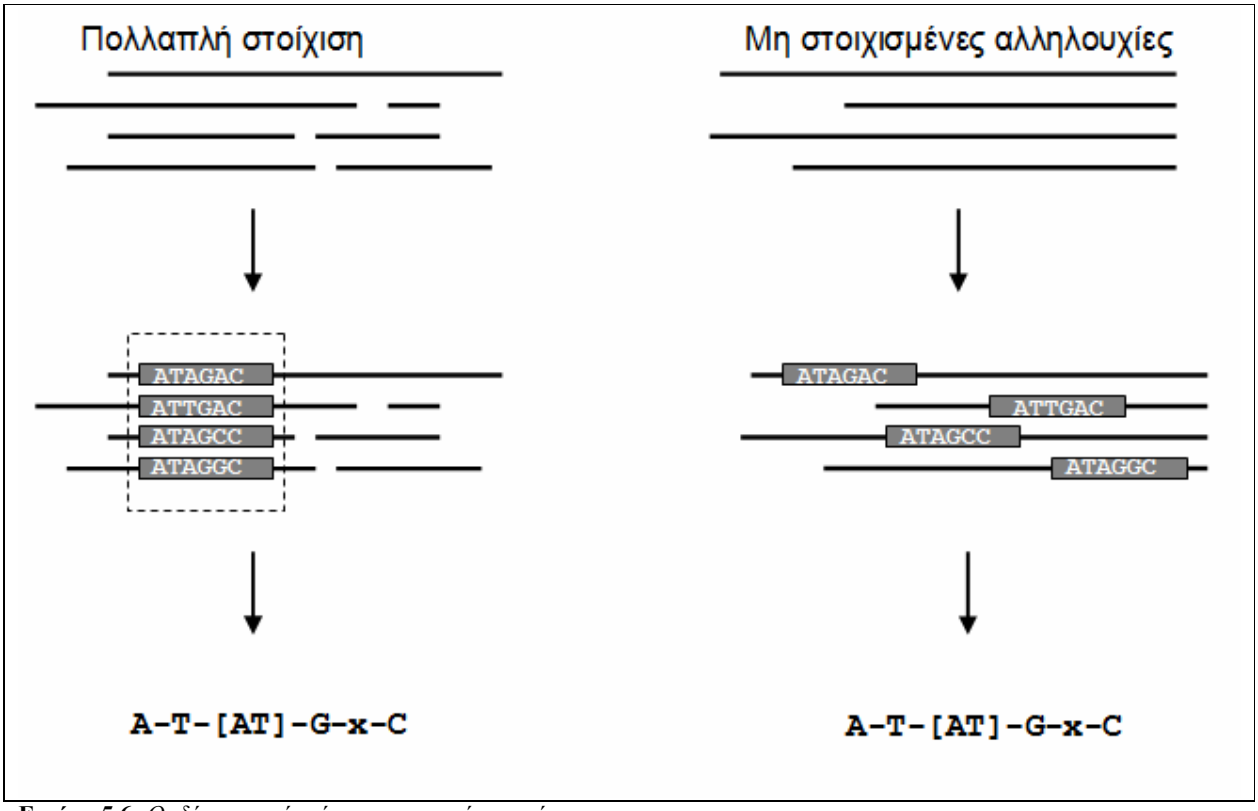
5.1.3 Κατασκευή των προτύπων και λογισμικό

Υπάρχουν δύο γενικοί τρόποι για την κατασκευή προτύπων: είτε ξεκινώντας από στοιχισμένες αλληλουχίες, είτε από μη στοιχισμένες (Εικόνα 5.6). Στην πρώτη περίπτωση, τα πράγματα είναι πιο απλά καθώς έχουμε εντοπίσει τη στοίχιση και η εύρεση των συντηρημένων περιοχών είναι μια τετριμμένη διαδικασία η οποία μπορεί να διεκπεραιωθεί με μια απλή καταμέτρηση. Η δεύτερη περίπτωση όμως, έχει μεγαλύτερο ενδιαφέρον

καθώς αποδεσμεύει το πρόβλημα από τη στοίχιση, αλλά επιπλέον προσφέρει το πλεονέκτημα ότι μπορεί να εντοπίσει πολλαπλές επαναλήψεις του ίδιου προτύπου στην αλληλουχία, αλλά και να εντοπίσει πρότυπα σε μη ομόλογες αλληλουχίες. Το μειονέκτημα βέβαια είναι ότι απαιτείται ειδικός αλγόριθμος.

Γενικά, η εύρεση προτύπων αποτελείται από 3 διακριτά μέρη (Brazma, Jonassen, Eidhammer, & Gilbert, 1998):

- *Επιλογή της γλώσσας*: στο πρώτο στάδιο θα πρέπει να επιλεγεί ο τρόπος περιγραφής των προτύπων. Μπορεί δηλαδή να χρησιμοποιηθεί η σύνταξη της PROSITE, αλλά υπάρχουν και περιπτώσεις στις οποίες επιλέγονται και πιο απλές περιγραφές (π.χ. πρότυπα τα οποία περιέχουν εκφράσεις χωρίς πολλαπλές ταυτίσεις σε κάποια θέση, δηλαδή είτε ένα σύμβολο είτε οποιοδήποτε).
- *Κριτήριο καταλληλότητας*: αυτό είναι το μέτρο με το οποίο θα αξιολογήσουμε ένα πρότυπο ως καλό. Μπορεί να περιλαμβάνει απλές εκφράσεις, όπως τον αριθμό ή το ποσοστό των συντηρημένων θέσεων, μέχρι πιο σύνθετες όπως το συνολικό πληροφοριακό περιεχόμενο ή την πιθανοφάνεια.
- *Αλγόριθμος:* το τελευταίο κομμάτι αφορά τον τρόπο αναζήτησης και είναι περισσότερο σχετικό στην περίπτωση μη στοιχισμένων αλληλουχιών, στις οποίες το πρόβλημα είναι NP-complete, οπότε συνήθως χρησιμοποιούνται ευριστικές τεχνικές (heuristic) ή άπληστοι (greedy) αλγόριθμοι, στους οποίους περιορίζεται το εύρος αναζήτησης (π.χ. αναζήτηση όλων των προτύπων με μέγεθος μέχρι ένα ορισμένο σημείο). Επίσης, χρησιμοποιούνται ευρέως και στατιστικές τεχνικές, όπως ο αλγόριθμος EM (Expectation-Maximization) και ο Gibbs sampler.



Εικόνα 5.6: *Οι δύο γενικοί τρόποι κατασκευής προτύπων.*

Το πιο παλιό και ευρέως χρησιμοποιούμενο εργαλείο για την κατασκευή προτύπων, είναι το **PRATT** (http://web.expasy.org/pratt/). Το PRATT χρησιμοποιεί μια αναπαράσταση γράφων για τα πρότυπα, λειτουργεί με μη στοιχισμένες αλληλουχίες και δέχεται πρότυπα στη μορφή PROSITE (Jonassen, Collins, & Higgins, 1995). Ο χρήστης δίνει σαν δεδομένα εισόδου τις αλληλουχίες και τις γενικές απαιτήσεις των προτύπων, π.χ. το εύρος του μήκους τους, τον αριθμό με τις μη συντηρημένες θέσεις που μπορεί να

περιέχουν, και τον αριθμό των πρωτεϊνών στις οποίες πρέπει να εμφανίζονται. Το PRATT έχει μπορέσει να ανακατασκευάσει αρκετά ήδη γνωστά πρότυπα, ενώ είναι μια ιδιαίτερα εύχρηστη και γρήγορη εφαρμογή που υπάρχει και σε διαδικτυακή έκδοση.

Το **MEME** (http://meme-suite.org/tools/meme) είναι μια επίσης πολύ γνωστή μέθοδος που βασίζεται στον αλγόριθμο EM (Multiple EM For Motif Elicitation). Το MEME διαθέτει πολλές εφαρμογές, κάποιες εκ των οποίων χρησιμοποιούν και προφίλ αλληλουχιών (θα τα εξετάσουμε παρακάτω). Στη γενική περίπτωση, ο αλγόριθμος χρησιμοποιεί μια στατιστική περιγραφή των προτύπων και βασίζεται στο γνωστό πρόβλημα της μίξης των κατανομών (αντιμετωπίζει τις στήλες σαν ανεξάρτητες παρατηρήσεις από πολυωνυμικές κατανομές με διαφορετικές πιθανότητες). Ο αλγόριθμος δέχεται επίσης κάποιες αρχικές παραδοχές για το μήκος του προτύπου και με μια επαναληπτική διαδικασία μέγιστης πιθανοφάνειας εντοπίζει τις βέλτιστες περιοχές πάνω στις αλληλουχίες οι οποίες φέρουν κάποιο χαρακτηριστικό (Bailey & Elkan, 1994).

Μια άλλη παρόμοια εφαρμογή, είναι ο **Gibbs Motif Sampler** ο οποίος όπως λέει το όνομα, βασίζεται στη στατιστική μεθοδολογία του Gibbs sampler (http://ccmbweb.ccv.brown.edu/gibbs/gibbs.html). Η μέθοδος αυτή έχει διάφορες παραλλαγές εστιασμένες σε διαφορετικές απαιτήσεις, όπως για παράδειγμα την εύρεση θέσεων πρόσδεσης μεταγραφικών παραγόντων ή τις επαναληπτικές αλληλουχίες, ενώ είναι διαθέσιμη και ως αυτόνομο λογισμικό (Thompson, Rouchka, & Lawrence, 2003).

Τέλος, ο **TEIRESIAS** ο οποίος αναπτύχθηκε από τον Έλληνα επιστήμονα Ισίδωρο Ριγούτσο όταν αυτός εργαζόταν στην IBM, είναι ίσως ο πιο ενδιαφέρων από τους διαθέσιμους αλγορίθμους (Rigoutsos & Floratos, 1998). Ο αλγόριθμος είναι συνδυαστικός (combinatorial) και εντοπίζει πρότυπα που εμφανίζονται περισσότερες φορές από έναν επιλεγμένο από τον χρήστη αριθμό, αλλά το επιτυγχάνει αυτό χωρίς να απαριθμεί όλα τα ενδεχόμενα. Επιπλέον δε, τα πρότυπα που ανακαλύπτει είναι τα βέλτιστα δυνατά, με την έννοια ότι είναι αδύνατο να γίνουν πιο ειδικά και ταυτόχρονα να εμφανίζονται στις ίδιες ακριβώς θέσεις σε όλες τις αλληλουχίες. Ο TEIRESIAS είναι διαθέσιμος στη διεύθυνση https://cm.jefferson.edu/Teiresias/, ενώ ενδιαφέρον έχει ότι εκτός από τις εφαρμογές του στην ανακάλυψη προτύπων σε αλληλουχίες DNA, έχει χρησιμοποιηθεί και σε άλλου είδους προβλήματα όπως στον εντοπισμό ύποπτων συμπεριφορών στα δίκτυα υπολογιστών.

5.2. Weight Matrices, Profiles και PSSMs

Είδαμε στην προηγούμενη ενότητα τις βασικές αδυναμίες των προτύπων. Η πιο σημαντική από αυτές, είναι ότι σε κάθε θέση «χάνεται» η πληροφορία για τη σχετική αναλογία των συμβόλων του αλφαβήτου, και η αδυναμία να ποσοτικοποιήσει την ταύτιση μιας δεδομένης αλληλουχίας. Τα προβλήματα αυτά, άρχισαν να γίνονται φανερά και πιο έντονα όσο τα δεδομένα συσσωρεύονταν με αποτέλεσμα να εμφανίζονται όλο και περισσότερες περιπτώσεις αλληλουχιών που για μία ή δύο αλλαγές στην αλληλουχία τους, δεν ταίριαζαν στο γνωστό πρότυπο. Τις αδυναμίες αυτές, έρχονται να αντιμετωπίσουν οι σταθμισμένοι πίνακες (weight matrices) και τα προφίλ (profiles). Με τη μεθοδολογία αυτή, κατασκευάζεται ένας πίνακας k x p, όπου k είναι το μέγεθος του αλφαβήτου και p το μέγεθος της περιοχής που μοντελοποιούμε (οι στήλες της πολλαπλής στοίχισης). Έτσι, σε κάθε θέση i της πολλαπλής στοίχισης αντιστοιχίζουμε ένα διάνυσμα με τις πιθανότητες εμφάνισης $p_b(i)$ του κάθε συμβόλου (Εικόνα 5.7). Αν ονομάσουμε $n_b(i)$ τον αριθμό των εμφανίσεων του συμβόλου b στη στήλη i, τότε $p_b(i)$ θα είναι η πιθανότητα του συμβόλου b στη στήλη i, οποία θα δίνεται από τη σχέση:

$$p_b(i) = \frac{n_b(i)}{\sum_{\forall b' \in \Omega} n_{b'}(i)} \tag{5.1}$$

Με αυτόν τον τρόπο μπορούμε αμέσως να αντιμετωπίσουμε και τα δύο προβλήματα που προκύπτουν από την απώλεια πληροφορίας των προτύπων. Μπορούμε να καταλάβουμε ποιο σύμβολο εμφανίζεται με μεγαλύτερη πιθανότητα σε μια θέση, ενώ μπορούμε και να ποσοτικοποιήσουμε την ταύτιση μιας αλληλουχίας με το μοντέλο. Για παράδειγμα στα δεδομένα της Εικόνας 5.1, και αν θυμηθούμε το κεφάλαιο 3, θα δούμε ότι η αλληλουχία ATTGAACTA έχει συνολική πιθανότητα εμφάνισης ίση με:

$$P(\mathbf{x}) = \prod_{i-1}^{p} p_b(i) = P(x_1 = A)P(x_2 = T)P(x_3 = T)...P(x_{10} = A) = 0.074 \tag{5.2}$$

ενώ αντίστοιχα, η πιθανότητα της αλληλουχίας ATAGTTCAA θα είναι ίση με 0.00155 (η δε αλληλουχία AATGAACTA θα έχει πιθανότητα 0, καθώς έχει μια μη αποδεκτή αλλαγή στη θέση 1). Γενικά όμως, η απλή

αυτή μέθοδος δεν είναι πολύ πρακτική κυρίως λόγω των πολύ μικρών πιθανοτήτων που μπορεί να εμφανιστούν. Συνήθως σε τέτοιες περιπτώσεις παίρνουμε το λογάριθμο των πιθανοτήτων, αλλά ακόμα πιο αξιόπιστα αποτελέσματα θα έχουμε αν πάρουμε ένα λογαριθμικό σκορ όπως αυτά που συναντήσαμε στο κεφάλαιο 3. Το προσθετικό αυτό σκορ, θα αντικατοπτρίζει και τη σχετική πιθανότητα ενός συμβόλου και θα είναι της μορφής:

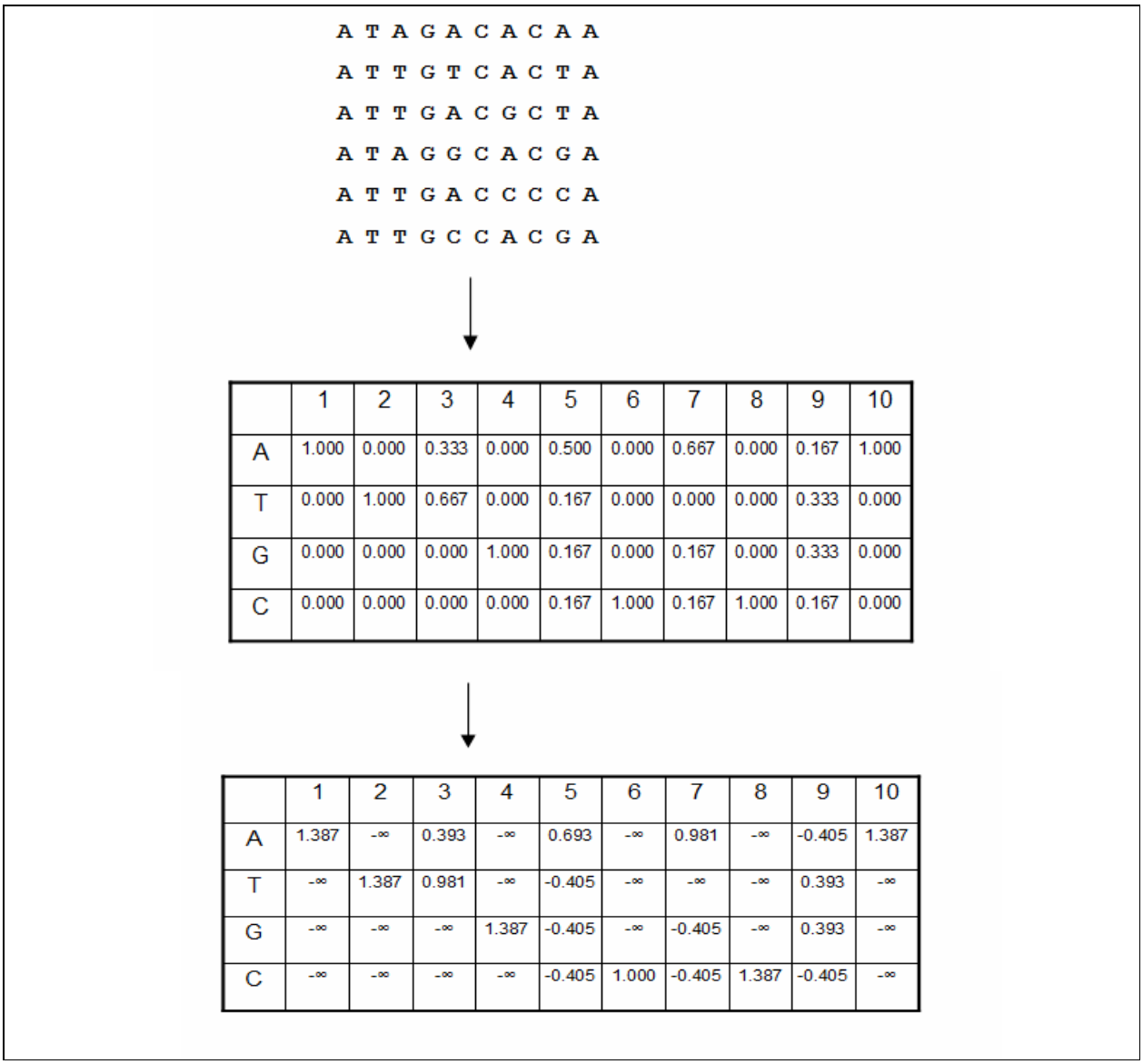
$$s_b(i) = \log\left(p_b(i) / p_b\right) \tag{5.3}$$

όπου p_b είναι η πιθανότητα εμφάνισης ενός συμβόλου (αμινοξέος ή νουκλεοτιδίου) γενικά (στο υπόβαθρο όπως λέμε) και $p_b(i)$ η πραγματική πιθανότητα εμφάνισης του ίδιου συμβόλου στη συγκεκριμένη θέση του πίνακα.

Με τον τρόπο αυτό, έχουμε ένα αθροιστικό σκορ το οποίο λαμβάνει επίσης υπόψη και τις συνολικές πιθανότητες εμφάνισης του κάθε συμβόλου. Ειδικά στις πρωτεΐνες, οι διαφορές μπορεί να είναι μεγάλες καθώς δεν είναι το ίδιο να έχουμε 100% συντήρηση ενός κοινού αμινοξέος με την συντήρηση ενός σπανίου (στη δεύτερη περίπτωση το σκορ θα είναι μεγαλύτερο). Όπως σε όλες τις περιπτώσεις με τα αντίστοιχα σκορ, ένα μικρό πρόβλημα μπορεί να προκύψει στις περιπτώσεις που ένα σύμβολο δεν εμφανίζεται καθόλου σε μια θέση, οπότε η σχέση (3.19) δεν ορίζεται και το σκορ γίνεται -∞. Τότε, υπάρχουν δύο εναλλακτικές. Αν δεν θέλουμε να επιτρέψουμε αυτό το σύμβολο να εμφανιστεί ποτέ, απλά αντικαθιστούμε την τιμή αυτή με έναν ιδιαίτερα μικρό αριθμό (π.χ. -10,000) και όλα λειτουργούν κανονικά, καθώς έστω και μια τέτοια εμφάνιση θα δώσει αρνητικό σκορ. Η άλλη εναλλακτική είναι να προσθέσουμε μικρές ψευδοτιμές, έτσι ώστε να καλύψουμε τη θεωρητική πιθανότητα το σύμβολο αυτό να έχει εμφανιστεί. Έτσι, η σχέση (3.19) θα γίνει:

$$s_b(i) = \log\left(\frac{p_b(i) + z_b}{p_b + \sum_{s=1}^{k} z_s}\right) \tag{5.4}$$

και με αυτόν τον τρόπο θα δοθούν μεγάλες αρνητικές τιμές στα σύμβολα που δεν εμφανίζονται σε κάποια θέση. Εναλλακτικά, οι ψευδοτιμές μπορούν να προστεθούν ευκολότερα στην εξίσωση (5.1) ως ακέραιες τιμές εμφάνισης των συμβόλων. Συνήθως, τέτοιοι πίνακες στρογγυλοποιούνται σε ακέραιες τιμές, για ακόμα μεγαλύτερη ευκολία στους υπολογισμούς. Με αυτόν τον ορισμό, αλληλουχίες με θετικό σκορ έχουν καλή ταύτιση με το μοντέλο, ενώ αλληλουχίες με αρνητικό σκορ θεωρείται ότι δεν έχουν.

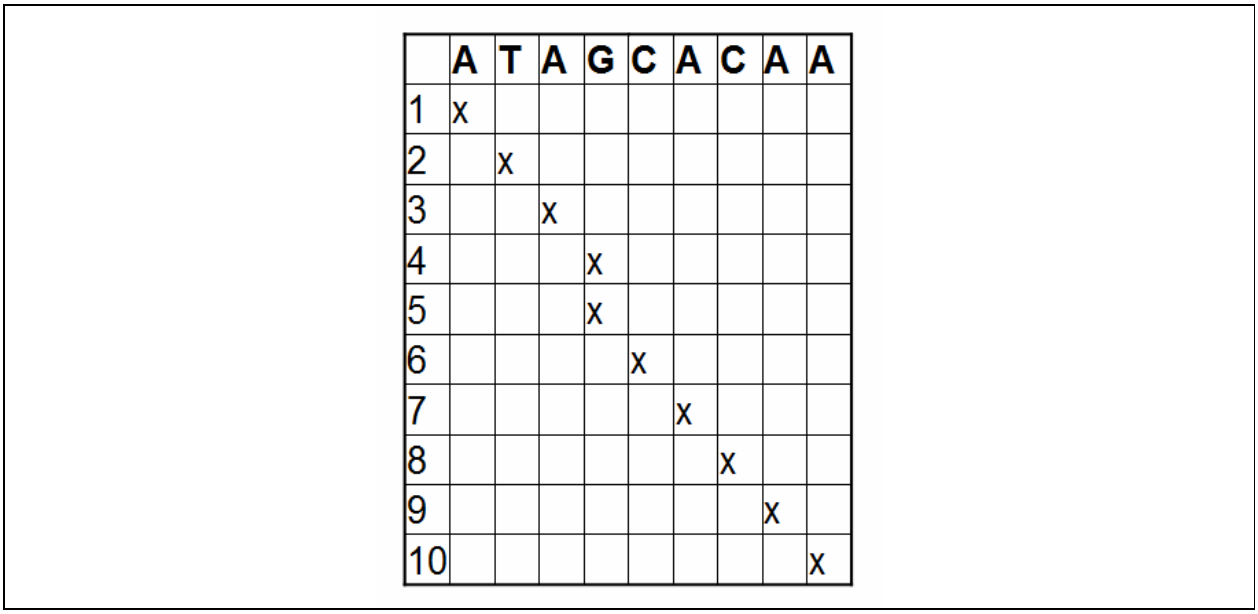


	1	2	3	4	5	6	7	8	9	10
A	1.000	0.000	0.333	0.000	0.500	0.000	0.667	0.000	0.167	1.000
T	0.000	1.000	0.667	0.000	0.167	0.000	0.000	0.000	0.333	0.000
G	0.000	0.000	0.000	1.000	0.167	0.000	0.167	0.000	0.333	0.000
C	0.000	0.000	0.000	0.000	0.167	1.000	0.167	1.000	0.167	0.000

	1	2	3	4	5	6	7	8	9	10
A	1.387	-∞	0.393	-∞	0.693	-∞	0.981	-∞	-0.405	1.387
T	-∞	1.387	0.981	-∞	-0.405	-∞	-∞	-∞	0.393	-∞
G	-∞	-∞	-∞	1.387	-0.405	-∞	-0.405	-∞	0.393	-∞
C	-∞	-∞	-∞	-∞	-0.405	1.000	-0.405	1.387	-0.405	-∞

Εικόνα 5.7: *Ένα παράδειγμα δημιουργίας σταθμισμένου πίνακα (weight matrix) και πίνακα σκορ ειδικού ανά θέση (PSSM), από μια πολλαπλή στοίχιση.*

Οι πίνακες αυτοί, έχουν πάρα πολλές εφαρμογές και σε πολλές περιπτώσεις έχουν αντικαταστήσει τα κλασικά πρότυπα ακριβώς λόγω της ευελιξίας τους. Ανάλογα με το πρόβλημα, μπορεί να υπάρχουν και επιπλέον διαφοροποιήσεις. Για παράδειγμα, η πιο απλή επιλογή είναι να έχουμε κατασκευάσει έναν τέτοιον πίνακα και απλά να κάνουμε μια αναζήτηση ελέγχοντας διαδοχικά τα επικαλυπτόμενα παράθυρα κατά μήκος της αλληλουχίας (Staden, 1990). Αυτό ισοδυναμεί με την υπόθεση ότι το προφίλ που αναζητάμε αναμένουμε να έχει ακριβώς τις ίδιες θέσεις με το αρχικό (μια συνηθισμένη υπόθεση όταν ψάχνουμε για μια καλά χαρακτηρισμένη από λειτουργικής άποψης περιοχή, π.χ. το ενεργό κέντρο ενός ενζύμου ή τη θέση πρόσδεσης ενός μεταγραφικού παράγοντα). Σε άλλες περιπτώσεις μπορεί να ενδιαφερόμαστε για κάτι πιο γενικό, οπότε μπορεί να μας ενδιαφέρει να έχουμε ευελιξία και να επιτρέπουμε κενά (τόσο στην αλληλουχία, όσο και στο προφίλ) (Barton & Sternberg, 1990). Αυτό επιτυγχάνεται με μια μικρή επέκταση των κλασικών αλγορίθμων δυναμικού προγραμματισμού που έχουμε γνωρίσει για την περίπτωση στοίχισης δύο αλληλουχιών. Η διαφορά είναι ότι σε αυτή την εκδοχή αντί να έχουμε στοίχιση αλληλουχίας με αλληλουχία, θα έχουμε τη στοίχιση της αλληλουχίας με το προφίλ. Προφανώς, χρειάζεται και σε αυτή την περίπτωση μια καλά υπολογισμένη, εμπειρικά, ποινή για τα κενά.



	A	T	A	G	C	A	C	A	A
1	x								
2		x							
3			x						
4				x					
5				x					
6					x				
7						x			
8							x		
9								x	
10									x

Εικόνα 5.8: *Στοίχιση μια αλληλουχίας με ένα προφίλ.*

Ειδικά στις πρωτεΐνες, είναι δυνατό να κατασκευαστεί ένα ακόμα πιο ευαίσθητο σύστημα για το σκορ, ικανό να εντοπίζει και μακρινές ομοιότητες. Η μέθοδος αυτή ονομάζεται profile analysis και ήταν μια από τις πρώτες και πολύ ικανοποιητικές προσεγγίσεις στον εντοπισμό μακρινών ομολόγων (Gribskov, McLachlan, & Eisenberg, 1987). Η ιδέα είναι να φτιαχτεί ένας ειδικός ανά θέση πίνακας του σκορ (position specific scoring matric-PSSM), ο οποίος θα μπορεί να χρησιμοποιηθεί αντί των κλασικών πινάκων ομοιότητας (PAM, BLOSUM κλπ) σε μια κλασική μέθοδο στοίχισης. Αρχικά, ξεκινάμε με μια αλληλουχία και εντοπίζουμε τις ομόλογες. Από αυτές, κατασκευάζουμε μια πολλαπλή στοίχιση από την οποία κατασκευάζουμε όμοια με προηγουμένως τον πίνακα με τις πιθανότητες εμφάνισης κάθε καταλοίπου. Βασικό σημείο που χρειάζεται προσοχή εδώ, είναι το γεγονός ότι ο πίνακας έχει τόσες θέσεις, όσο είναι και το μήκος της αρχικής αλληλουχίας. Αυτό συμβαίνει γιατί στήλες στην πολλαπλή στοίχιση που περιέχουν τυχόν κενά στην αρχική αλληλουχία, αγνοούνται. Με άλλα λόγια, η αλληλουχία «μετατρέπεται» σε έναν πίνακα που περιέχει πληροφορίες από όλες τις ομόλογές της και με τον τρόπο αυτόν πετυχαίνουμε μεγαλύτερη ευαισθησία στις αναζητήσεις. Φυσικά, η μέθοδος είναι πιο γενική και μπορεί να χρησιμοποιηθεί και για κατασκευή μοντέλου από μια οποιαδήποτε πολλαπλή στοίχιση, μόνο που τότε θα πρέπει να αποφασιστεί ποιες στήλες δεν θα συμπεριληφθούν στο μοντέλο (αυτές που έχουν κενά περισσότερα από μια προκαθορισμένη τιμή).

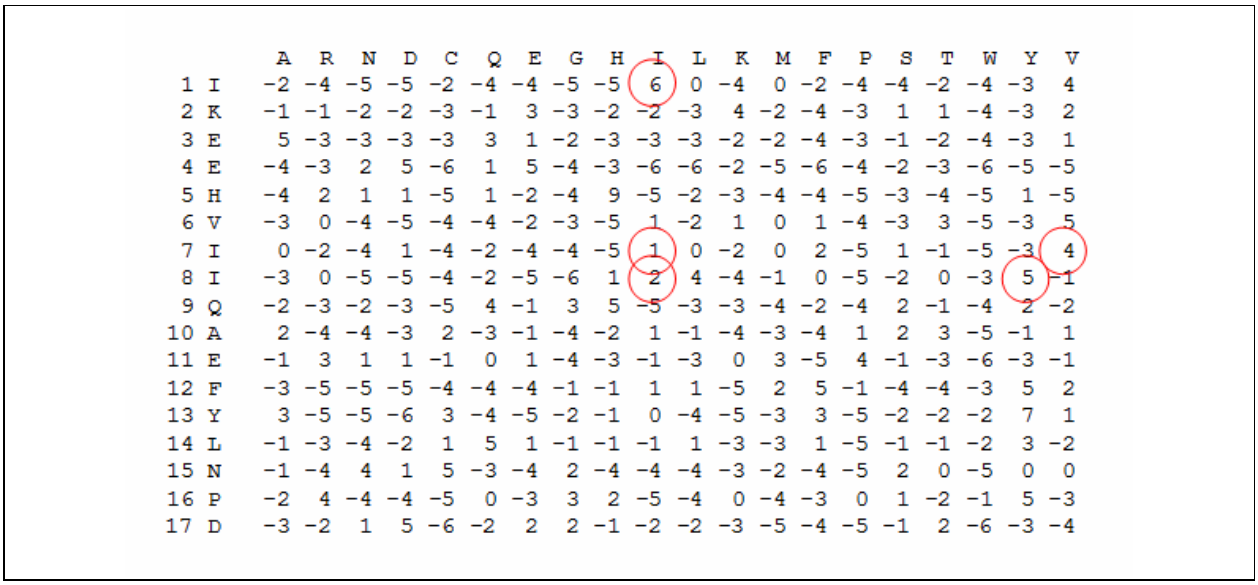
Στον υπολογισμό του σκορ, η βασική διαφορά από την κλασική μέθοδο, έγκειται στο ότι σε κάθε θέση η τιμή του σκορ δίνεται από ένα μέσο όρο όλων των τιμών που προβλέπει ένας κλασικός πίνακας του σκορ για τις συγκρίσεις αλληλουχιών. Έτσι, θα έχουμε:

$$s_b(i) = \sum_{j=1}^{k} p_j(i) S_{bj} \tag{5.5}$$

όπου $p_j(i)$ είναι όμοια με παραπάνω η πιθανότητα εμφάνισης του αμινοξέος j στη θέση i της πολλαπλής στοίχισης (που θα αντιστοιχεί στην ομόλογη θέση της αρχικής αλληλουχίας), ενώ το S_{bj} είναι η τιμή που προβλέπει ο επιλεγμένος πίνακας ομοιότητας (πχ BLOSUM62) για τη σύγκριση των αμινοξέων b και j. Από τον παραπάνω τρόπο υπολογισμού του σκορ, καταλαβαίνουμε ότι ακόμα και αν ένα αμινοξύ δεν εμφανίζεται καθόλου σε μια δεδομένη θέση της πολλαπλής στοίχισης, θα μπορεί να έχει Παρ' όλα αυτά θετική τιμή του σκορ, καθώς θα δεχτεί θετικές συνεισφορές από τα αμινοξέα με τα οποία έχει θετική τιμή στον επιλεγμένο πίνακα ομοιότητας. Στην αρχική εργασία, οι συγγραφείς χρησιμοποίησαν πίνακα ομοιότητας της οικογένειας PAM, αλλά περαιτέρω αναλύσεις έδειξαν ότι ο BLOSUM45 είναι καλύτερος, ενώ επιπλέον η διαφορική στάθμιση των αλληλουχιών, έτσι ώστε οι πολύ όμοιες να συνεισφέρουν λιγότερο στον πίνακα, βελτιώνει τη μεθοδολογία (Lüthy, Xenarios, & Bucher, 1994).

Το σκορ από τις αναζητήσεις με προφίλ, ακολουθεί την κατανομή του Gumbel, όμοια με τη στοίχιση αλληλουχιών. Στις περισσότερες περιπτώσεις, οι παράμετροι της κατανομής υπολογίζονται με

προσομοιώσεις. Σε μερικές απλές περιπτώσεις όμως, όπως για παράδειγμα σε περιπτώσεις χρήσης σταθμισμένου πίνακα και στοίχισης χωρίς κενά, χρησιμοποιούνται και εμπειρικοί κανόνες (π.χ. το σκορ πρέπει να είναι μεγαλύτερο από το 60% της μέγιστης τιμής που προβλέπει ο πίνακας).



		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	I	-2	-4	-5	-5	-2	-4	-4	-5	-5	6	0	-4	0	-2	-4	-4	-2	-4	-3	4
2	K	-1	-1	-2	-2	-3	-1	3	-3	-2	-2	-3	4	-2	-4	-3	1	1	-4	-3	2
3	E	5	-3	-3	-3	-3	3	1	-2	-3	-3	-3	-2	-2	-4	-3	-1	-2	-4	-3	1
4	E	-4	-3	2	5	-6	1	5	-4	-3	-6	-6	-2	-5	-6	-4	-2	-3	-6	-5	-5
5	H	-4	2	1	1	-5	1	-2	-4	9	-5	-2	-3	-4	-4	-5	-3	-4	-5	1	-5
6	V	-3	0	-4	-5	-4	-4	-2	-3	-5	1	-2	1	0	1	-4	-3	3	-5	-3	5
7	I	0	-2	-4	1	-4	-2	-4	-4	-5	1	0	-2	0	2	-5	1	-1	-5	-3	4
8	I	-3	0	-5	-5	-4	-2	-5	-6	1	2	4	-4	-1	0	-5	-2	0	-3	5	-1
9	Q	-2	-3	-2	-3	-5	4	-1	3	5	-5	-3	-3	-4	-2	-4	2	-1	-4	2	-2
10	A	2	-4	-4	-3	2	-3	-1	-4	-2	1	-1	-4	-3	-4	1	2	3	-5	-1	1
11	E	-1	3	1	1	-1	0	1	-4	-3	-1	-3	0	3	-5	4	-1	-3	-6	-3	-1
12	F	-3	-5	-5	-5	-4	-4	-4	-1	-1	1	1	-5	2	5	-1	-4	-4	-3	5	2
13	Y	3	-5	-5	-6	3	-4	-5	-2	-1	0	-4	-5	-3	3	-5	-2	-2	-2	7	1
14	L	-1	-3	-4	-2	1	5	1	-1	-1	-1	1	-3	-3	1	-5	-1	-1	-2	3	-2
15	N	-1	-4	4	1	5	-3	-4	2	-4	-4	-4	-3	-2	-4	-5	2	0	-5	0	0
16	P	-2	4	-4	-4	-5	0	-3	3	2	-5	-4	0	-4	-3	0	1	-2	-1	5	-3
17	D	-3	-2	1	5	-6	-2	2	2	-1	-2	-2	-3	-5	-4	-5	-1	2	-6	-3	-4

Εικόνα 5.9: *Ένα παράδειγμα PSSM. Συνήθως για ευκολία αλλάζουμε τις στήλες με τις γραμμές έτσι ώστε όλοι οι πίνακες να έχουν 20 στήλες αλλά τόσες γραμμές όσα είναι και τα αμινοξέα της πρωτεΐνης. Παρατηρήστε ότι το ίδιο αμινοξύ, π.χ. η Ισολευκίνη μπορεί να έχει σε διαφορετικές θέσεις τελείως διαφορετικό διάνυσμα, καθώς στις αντίστοιχες στήλες της πολλαπλής στοίχισης υπήρχαν διαφορετικά αμινοξέα. Ειδικά στις θέσεις 7 και 8, η πρωτεΐνη μας έχει Ισολευκίνη, αλλά οι περισσότερες πρωτεΐνες της στοίχισης έχουν Βαλίνη και Τυροσίνη, αντίστοιχα.*

5.3. Λογισμικό

Στην ενότητα αυτή, θα παρουσιάσουμε τα πιο γνωστά πακέτα λογισμικού που χρησιμοποιούνται είτε για να κατασκευάζουν PSSMs, είτε για να κάνουν αναζητήσεις. Το πιο γνωστό πρόγραμμα της πρώτης κατηγορίας είναι το **ScanProsite** (http://prosite.expasy.org/scanprosite/). Το ScanProsite είναι κατασκευασμένο για να εντοπίζει πρότυπα και προφίλ της PROSITE, σε οποιαδήποτε αλληλουχία, είτε του χρήστη, είτε κάποια που έχει επιλεγεί από μια βάση δεδομένων. Είναι το εργαλείο που χρησιμοποιείται επίσημα στις αναζητήσεις στην PROSITE και έχει πολλές βελτιστοποιήσεις για να αυξάνεται η ταχύτητα, όπως προϋπολογισμένες ταυτίσεις για τις γνωστές αλληλουχίες κ.ο.κ. (De Castro et al., 2006)

Το **PFTOOLS** (http://web.expasy.org/pftools/) ειναι ένα εργαλείο κατάλληλο τόσο για κατασκευή όσο και για αναζήτηση προφίλ από στοιχισμένες αλληλουχίες (Bucher, Karplus, Moeri, & Hofmann, 1996). Το PFTOOLS είναι πολύ γενικό, και περιλαμβάνει όλες τις περιπτώσεις προφίλ που αναφέραμε στην προηγούμενη ενότητα (πρότυπα, weight matrices, PSSMs), ενώ ενσωματώνει και την πιο γενική περίπτωση στην οποία όλες οι ποινές για τα κενά είναι επίσης ειδικές ανά θέση (generalized profile). Η τελευταία περίπτωση, απέχει ένα μόνο βήμα πριν από το Hidden Markov Model το οποίο θα εξετάσουμε στο κεφάλαιο 8. Το PFTOOLS χρησιμοποιείται κυρίως για την κατασκευή μοντέλων για πρωτεϊνικές οικογένειες, χρησιμοποιώντας μια πολλαπλή στοίχιση των μελών της οικογένειας και διαθέτει διάφορες ρουτίνες, όπως:

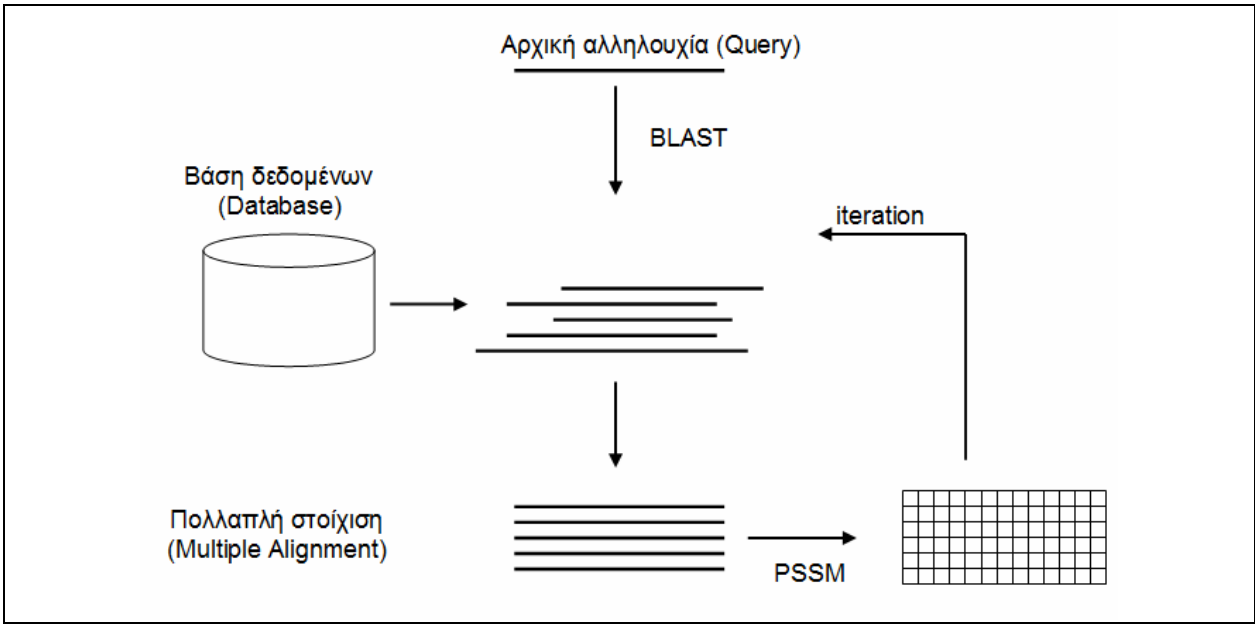
- **pfmake**: κατασκευάζει ένα προφίλ από μια δεδομένη πολλαπλή στοίχιση
- **pfscale**: βρίσκει τις παραμέτρους της κατανομής του Gumbel για να υπολογιστεί η στατιστική σημαντικότητα
- **pfw**: εφαρμόζει τη μέθοδο της διαφορικής στάθμισης των αλληλουχιών για να διορθώσει το συστηματικό σφάλμα από την υπερ-αντιπροσώπευση κάποιων μελών της οικογένειας.
- **pfsearch**: πραγματοποιεί αναζήτηση σε μια βάση δεδομένων αλληλουχιών πρωτεϊνών ή DNA έναντι σε ένα προφίλ.
- **pfscan**: πραγματοποιεί αναζήτηση μιας αλληλουχίας DNA ή πρωτεΐνης έναντι σε μια βιβλιοθήκη με προφίλ.

Επίσης, υπάρχουν μια σειρά από βοηθητικά προγράμματα που μετατρέπουν τα μοντέλα και τις ακολουθίες από και προς διάφορες άλλες γνωστές μορφές, μεταξύ των οποίων συμπεριλαμβάνεται και η μορφή HMMER που θα συναντήσουμε στο κεφάλαιο 8 (**psa2msa**, **gtop**, **htop**, **ptoh**) ή μετατρέπουν τις αλληλουχίες από DNA σε πρωτεϊνικές και το αντίστροφο (**ptof**, **2ft**, **6ft**).

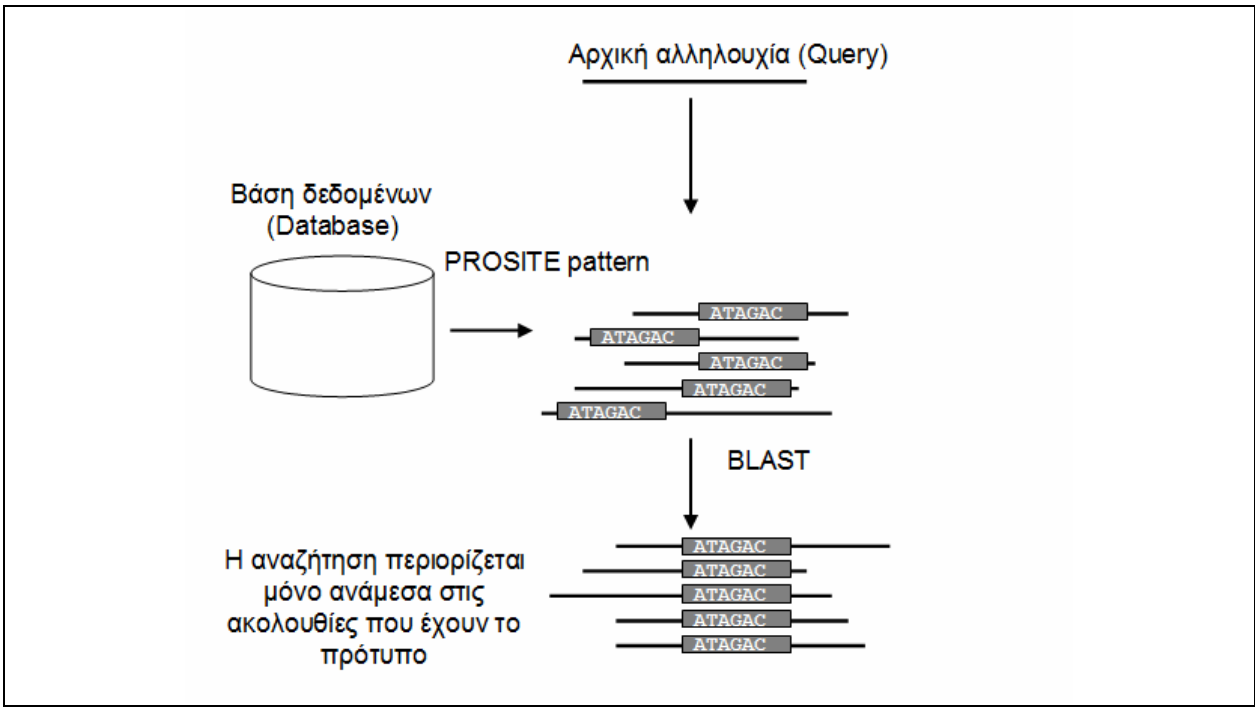
Ίσως η πιο ευρέως χρησιμοποιούμενη εφαρμογή που κάνει χρήση PSSM είναι το **PSI-BLAST** (Position-specific-iterated BLAST) (Altschul et al., 1997). Είναι μια επέκταση του γνωστού αλγορίθμου BLAST και χρησιμοποιείται για την εύρεση μακρινών ομολόγων. Η μέθοδος δουλεύει ως εξής (Εικόνα 5.10): Στην αρχή πραγματοποιείται μια κανονική αναζήτηση με το BLAST και συλλέγονται οι αλληλουχίες με E-value μικρότερο από κάποιο όριο που ορίζεται από τον χρήστη. Αυτές θεωρείται ότι είναι οι «σίγουρες» ομόλογες και χρησιμοποιούνται για να κατασκευαστεί ένας PSSM όπως περιγράψαμε παραπάνω, χωρίς όμως κενά καθώς κάθε στήλη του αντιστοιχεί σε μια θέση της αλληλουχίας της αρχικής πρωτεΐνης. Με αυτόν τον πίνακα, πραγματοποιείται εκ νέου αναζήτηση στη βάση δεδομένων, η οποία πλέον θα δώσει περισσότερες ομόλογες με E-value μικρότερο από το αρχικό όριο. Η διαδικασία αυτή επαναλαμβάνεται αρκετές φορές, είτε μέχρι να σταματήσουν να προστίθενται νέες αλληλουχίες, είτε μέχρι να ξεπεραστεί ένας συγκεκριμένος αριθμός επαναλήψεων (συνήθως 3 ή 4). Η μέθοδος είναι εξαιρετικά αποδοτική και εντοπίζει μεγάλο αριθμό ομόλογων πρωτεϊνών (μακρινών ομολόγων), οι οποίες δεν θα μπορούσαν να εντοπιστούν με μια συμβατική αναζήτηση. Η επαναληπτική αυτή διαδικασία, θυμίζει τον αλγόριθμο EM, και οι μόνες περιπτώσεις στις οποίες μπορεί να αποτύχει είναι είτε όταν δεν βρεθούν καθόλου ομόλογες στην πρώτη αναζήτηση, είτε όταν το όριο είναι αρκετά ψηλά με συνέπεια να συμπεριληφθούν και πρωτεΐνες που δεν έχουν πραγματική ομολογία, οπότε και το προφίλ δεν θα είναι πλέον ειδικό αρκετά (contamination).

Μια ενδιαφέρουσα επέκταση του PSI-BLAST είναι το **DELTA-BLAST** (domain enhanced lookup time accelerated BLAST), το οποίο αντί να κατασκευάσει το PSSM από την αρχή, πραγματοποιεί αναζήτηση σε μια βάση δεδομένων με ήδη χαρακτηρισμένες οικογένειες έτσι ώστε να πετύχει καλύτερη ακρίβεια στην αναγνώριση. Για το σκοπό αυτό, χρησιμοποιεί τη βάση Conserved Domain Database (CDD) του NCBI, και τα αποτελέσματα δείχνουν ότι με τη μέθοδο αυτή, πετυχαίνουμε καλύτερα αποτελέσματα από το PSI-BLAST, καθώς συνδυάζονται τα πλεονεκτήματα της επαναληπτικής διαδικασίας με αυτά της χρήσης της καλά χαρακτηρισμένης βάσης δεδομένων (Boratyn et al., 2012).

Το **PHI-BLAST** (pattern-hit initiated BLAST) είναι άλλη μια παραλλαγή του BLAST, η οποία όμως χρησιμοποιεί πρότυπα κανονικών εκφράσεων (Zhang et al., 1998). Η ιδέα εδώ είναι διαφορετική και συνίσταται στη χρησιμοποίηση γνωστών πρότυπων, τα οποία υπάρχουν στην αλληλουχία επερώτησης και τα καθορίζει ο χρήστης, για να καθοδηγήσουν την αναζήτηση. Με τον τρόπο αυτό, το εύρος της αναζήτησης περιορίζεται και σε πολλές περιπτώσεις εντοπίζονται ομόλογες πρωτεΐνες οι οποίες δεν μπορούσαν να εντοπιστούν με το συμβατικό τρόπο αναζήτησης (Εικόνα 5.11).



Εικόνα 5.10: *Σχηματικό διάγραμμα αναπαράστασης της λειτουργίας του PSI-BLAST.*



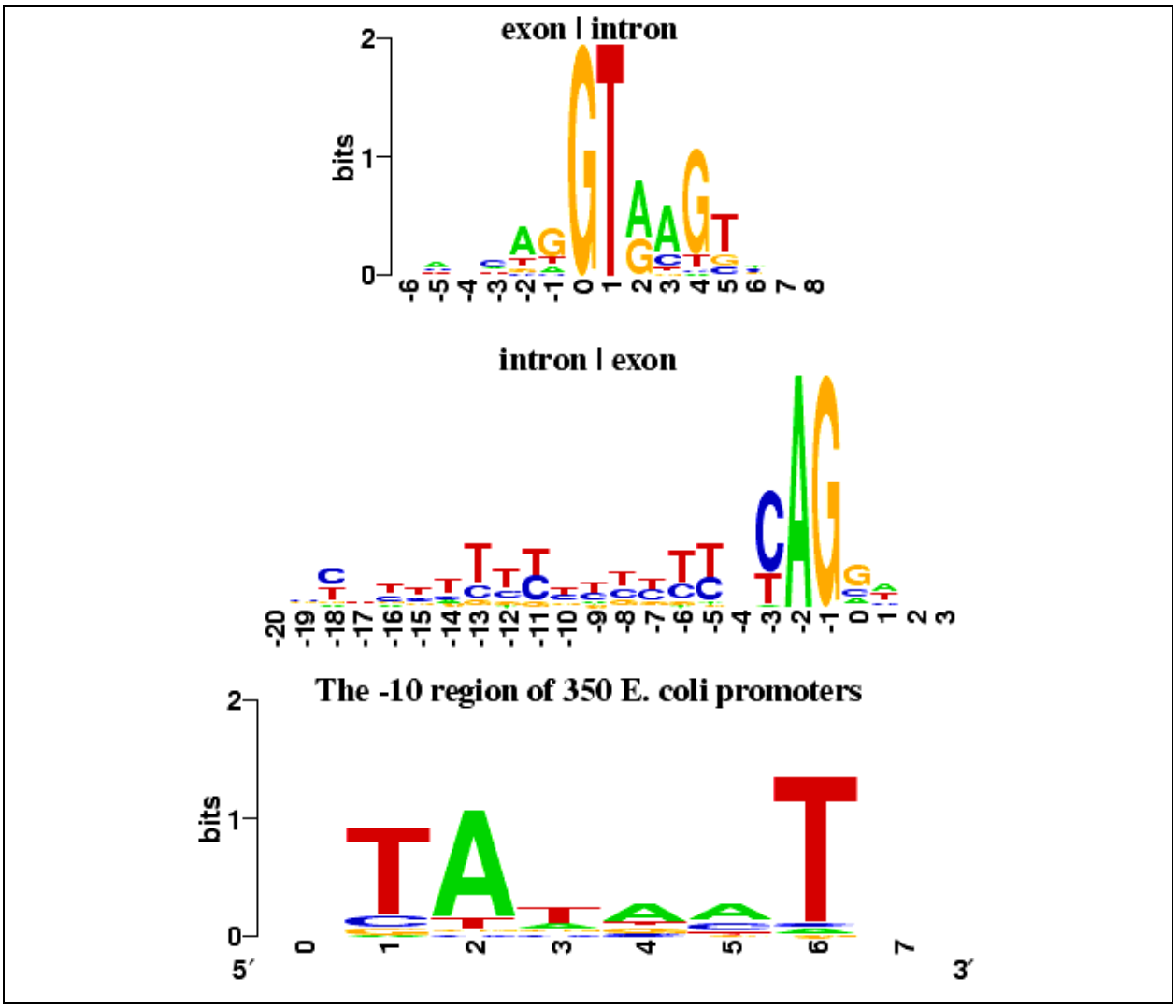
Εικόνα 5.11: *Σχηματικό διάγραμμα αναπαράστασης της λειτουργίας του PHI-BLAST.*

Τέλος, μια πολύ σημαντική εφαρμογή που χρησιμοποιείται για την οπτικοποίηση των περιοχών που απεικονίζονται σε ένα πρότυπο ή προφίλ, είναι το **WebLogo** (http://weblogo.berkeley.edu/) (Crooks, Hon, Chandonia, & Brenner, 2004). Το WebLogo βασίζεται στην απλή ιδέα των Λογότυπων Αλληλουχιών (Sequence Logo) των Schneider και Stephens (Schneider & Stephens, 1990) και απεικονίζει μια πολλαπλή στοίχιση σε μια γραφική αναπαράσταση, με στήλες στις οποίες εμφανίζονται τοποθετημένα κάθετα τα σύμβολα που εμφανίζονται σε αυτή. Το ύψος της στήλης αντιστοιχεί στη συνολική πληροφορία που φέρει η στήλη αυτή, και δίνεται από τον τύπο:

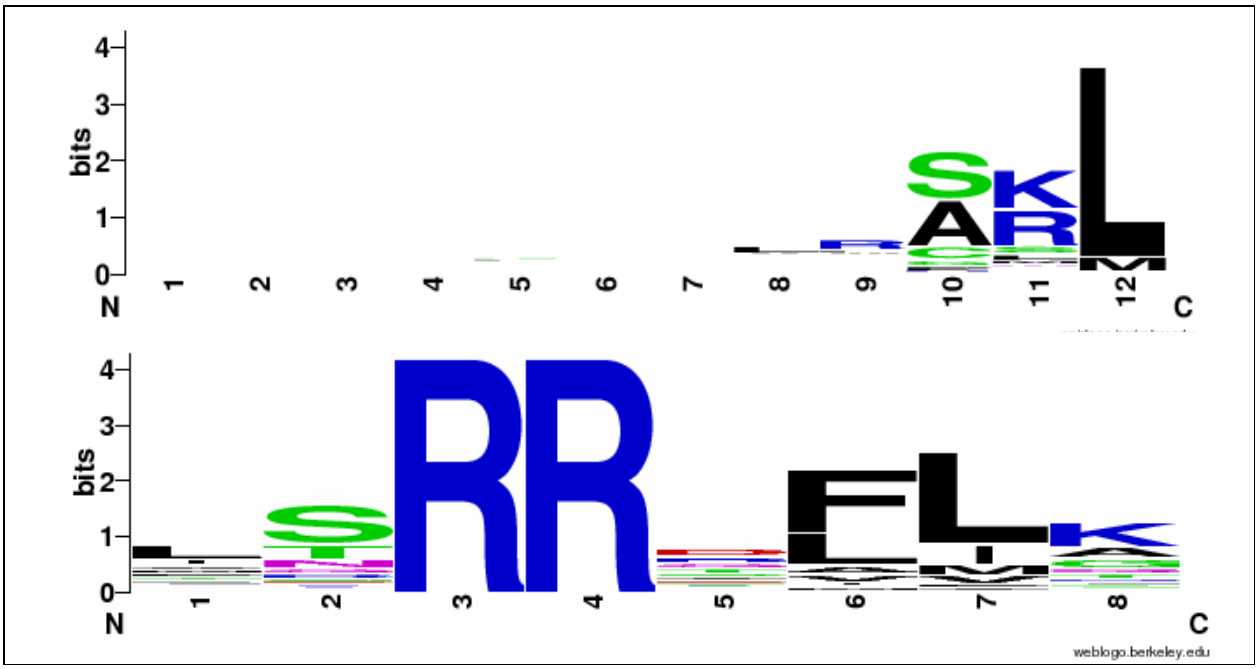
$$R = S_{\max} - S_{obs} = \log_2 k - \left(-\sum_{\forall b \in \Omega} n_b(i) \log p_b(i) \right) \tag{5.6}$$

Στη σχέση αυτή, το S_{max} είναι η μέγιστη εντροπία που μπορεί να έχει η στήλη και S_{obs} η παρατηρηθείσα εντροπία που είδαμε στο κεφάλαιο 3. Το k είναι το μέγεθος του αλφάβητου, το οποίο και καθορίζει τη μέγιστη τιμή (2.2 bits για DNA/RNA και ~4.32 για πρωτεΐνες). Το σχετικό ύψος του κάθε συμβόλου σε κάθε στήλη, δίνεται από τη συχνότητα εμφάνισής του. Το λογισμικό δέχεται σαν είσοδο μια πολλαπλή στοίχιση και παράγει τη γραφική παράσταση, η οποία είναι ιδιαίτερα κατατοπιστική καθώς μας δείχνει με μια γρήγορη ματιά ποιες στήλες είναι συντηρημένες, αλλά και ποια σύμβολα επικρατούν σε κάθε μια από αυτές. Οι στήλες που δεν έχουν ιδιαίτερη συντήρηση, εμφανίζονται σύμφωνα με τη σχέση (5.6) με μικρό ύψος.

Εικόνα 5.12: *Το Sequence Logo της πολλαπλής στοίχισης από την Εικόνα 5.1.*



Εικόνα 5.13: *Παραδείγματα Sequence Logo από αλληλουχίες DNA. Πάνω, απεικονίζονται τα λογότυπα των περιοχών εναλλαγής εσωνίων-εξωνίων, όπως προκύπτουν από τις πειραματικά προσδιορισμένες αλληλουχίες της EID (Exon-Intron database). Κάτω, απεικονίζεται η περιοχή του υποκινητή από 350 γονίδια της E. coli. Παρατηρήστε ότι παρόλο που οι περιοχές αυτές περιγράφονται και από πρότυπα, σε ένα μεγάλο σύνολο δεδομένων, λίγες είναι οι στήλες με απόλυτη συντήρηση.*



Εικόνα 5.14: *Παραδείγματα Sequence Logo από αλληλουχίες πρωτεϊνών. Πάνω απεικονίζονται τα λογότυπα των καρβοξυτελικών περιοχών που περιέχουν το σήμα στόχευσης για το υπεροξείσωμα (PTS1). Κάτω, απεικονίζεται η περιοχή των σηματοδοτικών αλληλουχιών από τις βακτηριακές πρωτεΐνες που εκκρίνονται με το μονοπάτι TAT. Παρατηρήστε ότι παρόλο που οι περιοχές αυτές περιγράφονται και από πρότυπα PROSITE, σε ένα μεγάλο σύνολο δεδομένων λίγες είναι οι στήλες με απόλυτη συντήρηση.*

Βιβλιογραφία

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research, 25*(17), 3389-3402.

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol, 2*, 28-36.

Barton, G. J., & Sternberg, M. J. (1990). Flexible protein sequence patterns: A sensitive method to detect weak structural similarities. *Journal of molecular biology, 212*(2), 389-402.

Berven, F. S., Flikka, K., Jensen, H. B., & Eidhammer, I. (2004). BOMP: a program to predict integral b-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res, 32*(Web Server Issue), W394-W399.

Boratyn, G. M., Schaffer, A. A., Agarwala, R., Altschul, S. F., Lipman, D. J., & Madden, T. L. (2012). Domain enhanced lookup time accelerated BLAST. *Biol Direct, 7*(1), 12.

Brazma, A., Jonassen, I., Eidhammer, I., & Gilbert, D. (1998). Approaches to the automatic discovery of patterns in biosequences. *Journal of computational biology, 5*(2), 279-305.

Bucher, P., Karplus, K., Moeri, N., & Hofmann, K. (1996). A flexible motif search technique based on generalized profiles. *Computers & chemistry, 20*(1), 3-23.

Cokol, M., Nair, R., & Rost, B. (2000). Finding nuclear localization signals. *EMBO reports, 1*(5), 411-415.

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res, 14*(6), 1188-1190.

De Castro, E., Sigrist, C. J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., . . . Hulo, N. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research, 34*(suppl 2), W362-W365.

Gribskov, M., McLachlan, A. D., & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A, 84*(13), 4355-4358.

Jonassen, I., Collins, J. F., & Higgins, D. G. (1995). Finding flexible patterns in unaligned protein sequences. *Protein Science, 4*(8), 1587-1595.

Lüthy, R., Xenarios, I., & Bucher, P. (1994). Improving the sensitivity of the sequence profile method. *Protein Science, 3*(1), 139-146.

Petriv, I., Tang, L., Titorenko, V. I., & Rachubinski, R. A. (2004). A new definition for the consensus sequence of the peroxisome targeting signal type 2. *Journal of molecular biology, 341*(1), 119-134.

Rigoutsos, I., & Floratos, A. (1998). Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics, 14*(1), 55-67.

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res, 18*(20), 6097-6100.

Shruthi, H., Babu, M. M., & Sankaran, K. (2010). TAT-pathway-dependent lipoproteins as a niche-based adaptation in prokaryotes. *Journal of Molecular Evolution, 70*(4), 359-370.

Sigrist, C. J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., & Hulo, N. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res, 38*(Database issue), D161-166.

Staden, R. (1990). Searching for patterns in protein and nucleic acid sequences. *Methods in enzymology, 183*, 193-211.

Sutcliffe, I. C., & Harrington, D. J. (2002). Pattern searches for the identification of putative lipoprotein genes in Gram-positive bacterial genomes. *Microbiology, 148*(Pt 7), 2065-2077.

Thompson, W., Rouchka, E. C., & Lawrence, C. E. (2003). Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Research, 31*(13), 3580-3585.

Zhang, Z., Miller, W., Schäffer, A. A., Madden, T. L., Lipman, D. J., Koonin, E. V., & Altschul, S. F. (1998). Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Research, 26*(17), 3986-3990.

Κεφάλαιο 6: Φυλογενετική Ανάλυση

Σύνοψη
Στο κεφάλαιο αυτό εξετάζονται οι υπολογιστικές όψεις της φυλογενετικής ανάλυσης, δηλαδή, της διαδικασίας εκτίμησης των εξελικτικών σχέσεων των οργανισμών, μέσα από τη μελέτη των αντίστοιχων βιολογικών αλληλουχιών τους. Θα δούμε στην αρχή τους βασικούς ορισμούς για τα φυλογενετικά δέντρα και τα βασικά πιθανοθεωρητικά μοντέλα της εξέλιξης αλληλουχιών. Κατόπιν, θα παρουσιάσουμε τις βασικές κατηγορίες μεθόδων κατασκευής φυλογενετικών δέντρων, και θα σχολιάσουμε τις ομοιότητες και τις διαφορές τους. Τέλος, θα παρουσιάσουμε τα αντίστοιχα πακέτα λογισμικού που υπάρχουν διαθέσιμα για το σκοπό αυτό, θα σχολιάσουμε τα σχετικά πλεονεκτήματα και μειονεκτήματα τους, και θα δώσουμε πρακτικές συμβουλές.

Προαπαιτούμενη γνώση
Βασικές γνώσεις εξελικτικής βιολογίας. Βασικές γνώσεις πιθανοτήτων. Το κεφάλαιο απαιτεί επίσης κατανόηση των μεθόδων του κεφαλαίου 3 και του κεφαλαίου 4.

6. Εισαγωγή

Το θέμα που θα μας απασχολήσει σε αυτό το κεφάλαιο είναι το πρόβλημα του προσδιορισμού των φυλογενετικών σχέσεων, δηλαδή, το πως θα μπορέσουμε από την αμινοξική αλληλουχία κάποιων πρωτεϊνών (ή τις περισσότερες φορές, από την αλληλουχία των αντιστοίχων γονιδίων), οι οποίες προέρχονται από διάφορους οργανισμούς, να προσδιορίσουμε τις εξελικτικές σχέσεις των οργανισμών αυτών.

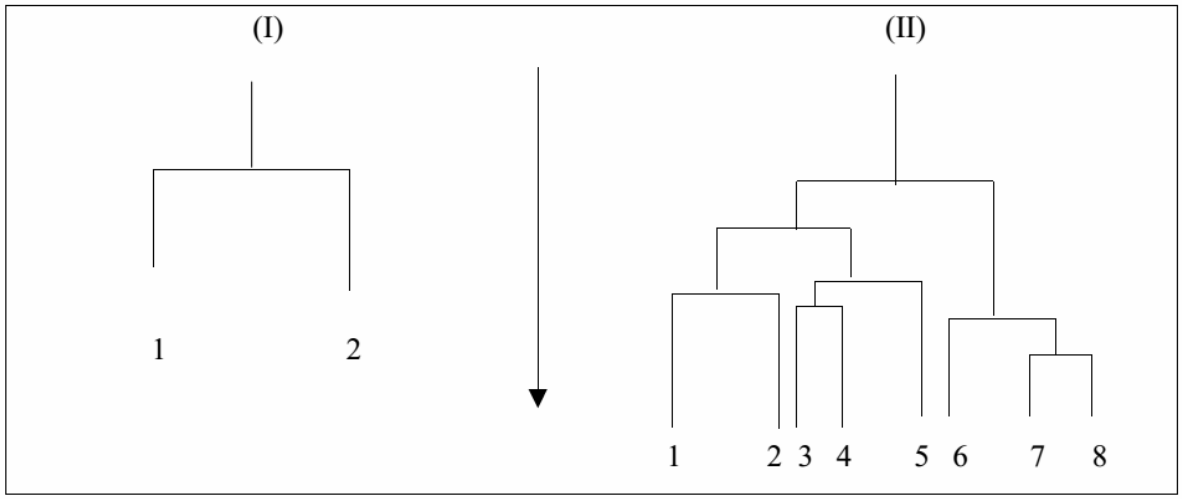
Το θέμα αυτό είναι τεράστιο με πολλές προεκτάσεις (φιλοσοφικές, αλλά και ιδεολογικές) και έχει γίνει από την εποχή του Darwin αντικείμενο για πολλές διαμάχες. Για τους πολέμιους της εξελικτικής θεωρίας (που στη βιολογία βέβαια είναι ελάχιστοι) το θέμα σταματά εδώ καθώς τίποτα από τα παρακάτω δεν έχει νόημα (και ίσως πολλά από όσα προηγήθηκαν), καθώς όπως είπε και ο Thedosius Dobzhansky: «*Nothing in Biology makes sense except in the light of evolution*». Για τους υπόλοιπους, το θέμα αποκτά έντονο ενδιαφέρον καθώς αν δεχθούμε ως βασική αλήθεια την ύπαρξη κοινών προγόνων για όλους τους ζώντες οργανισμούς και την εξέλιξη όλων των σημερινών ειδών από παλαιότερα μέσω της μετάλλαξης και της φυσικής επιλογής, το πρόβλημα της εκτίμησης αυτών των εξελικτικών σχέσεων είναι ένα κατ' εξοχήν μαθηματικό και υπολογιστικό πρόβλημα και κάποιες βασικές όψεις του θα προσπαθήσουμε να παρουσιάσουμε εδώ.

Οι μεθοδολογίες τις οποίες θα πραγματευθούμε σε αυτό το κεφάλαιο, έχουν μακρά ιστορία στο χώρο της βιολογίας. Οι επιστήμονες από τον καιρό του Darwin, προσπαθούσαν να κατασκευάσουν φυλογενετικά δέντρα που να αποδίδουν τις εξελικτικές σχέσεις των οργανισμών και χρησιμοποιούσαν αρχικά για το σκοπό αυτό, τα φαινοτυπικά χαρακτηριστικά. Με την ανάπτυξη όμως της μοριακής βιολογίας, τα μοριακά χαρακτηριστικά (δηλαδή, οι αλληλουχίες των γονιδίων και των πρωτεϊνών), είναι αυτά που κέρδισαν το ενδιαφέρον καθώς αυτά αποτελούν το βασικό υπόστρωμα πάνω στο οποίο δρουν οι εξελικτικές δυνάμεις (η μετάλλαξη και η φυσική επιλογή). Κατά συνέπεια, στο κεφάλαιο αυτό, θα παρουσιάσουμε τους βασικούς τρόπους φυλογενετικής μελέτης βιολογικών αλληλουχιών, θα αναδείξουμε τις ομοιότητες αλλά και τις διαφορές μεταξύ τους, θα εστιάσουμε στα σχετικά πλεονεκτήματα και μειονεκτήματα κάθε μιας, και θα παρουσιάσουμε τα βασικότερα εργαλεία λογισμικού που υπάρχουν διαθέσιμα για το σκοπό αυτό.

6.1. Βασικές Αρχές

Κατ' αρχήν πρέπει να είμαστε σίγουροι για το τι συγκρίνουμε. Αν θέλουμε να εκτιμήσουμε φυλογενετικές σχέσεις από τις αλληλουχίες κάποιων γονιδίων, πρέπει να συγκρίνουμε αντίστοιχα γονίδια, για να εντοπίσουμε την ομολογία τους. *Ομόλογες* πρωτεΐνες (ή γονίδια), λέγονται γενικά οι πρωτεΐνες που έχουν προκύψει μέσω της εξέλιξης από κάποιον κοινό πρόγονο. Συνήθως αυτές επιτελούν παρόμοια λειτουργία (κατ' αντιστοιχία με τα ομόλογα όργανα των οργανισμών), και κατά συνέπεια θα έχουν παρόμοια δομή και αλληλουχία. Τα αντίστοιχα γονίδια σε διαφορετικούς οργανισμούς αναφέρονται και ως *ορθόλογα* (*orthologues*) στη σχετική βιβλιογραφία, και θεωρούμε ότι η όποια διαφοροποίηση τους έχει προκύψει λόγω της ειδογένεσης. Αντίθετα, ομόλογες πρωτεΐνες, ή γονίδια, μέσα στο ίδιο είδος, ονομάζονται *παράομολογα*

(*paralogues*), και θεωρούμε ότι έχουν προκύψει από γονιδιακό διπλασιασμό και ανεξάρτητη εξέλιξη μέσα στο είδος. Παράδειγμα της πρώτης περίπτωσης είναι οι α-αλυσίδες της αιμοσφαιρίνης των θηλαστικών (π.χ. του ανθρώπου, του χιμπαντζή, του σκύλου κ.α.), ενώ για τη δεύτερη περίπτωση θα μπορούσαμε να αναφέρουμε μέσα στο ίδιο είδος (π.χ. τον άνθρωπο), τις α, β, γ, δ, ε, ζ, θ αλυσίδες της αιμοσφαιρίνης αλλά και τη μυοσφαιρίνη. Τέλος, υπάρχουν και τα λεγόμενα *ξενόλογα* ή *ξενοομόλογα* (*xenologues*) γονίδια, τα οποία είναι ομόλογα γονίδια τα οποία έχουν προκύψει από κάποια διαδικασία οριζόντιας γονιδιακής μεταφοράς (συνήθως από προκαρυωτικό οργανισμό).

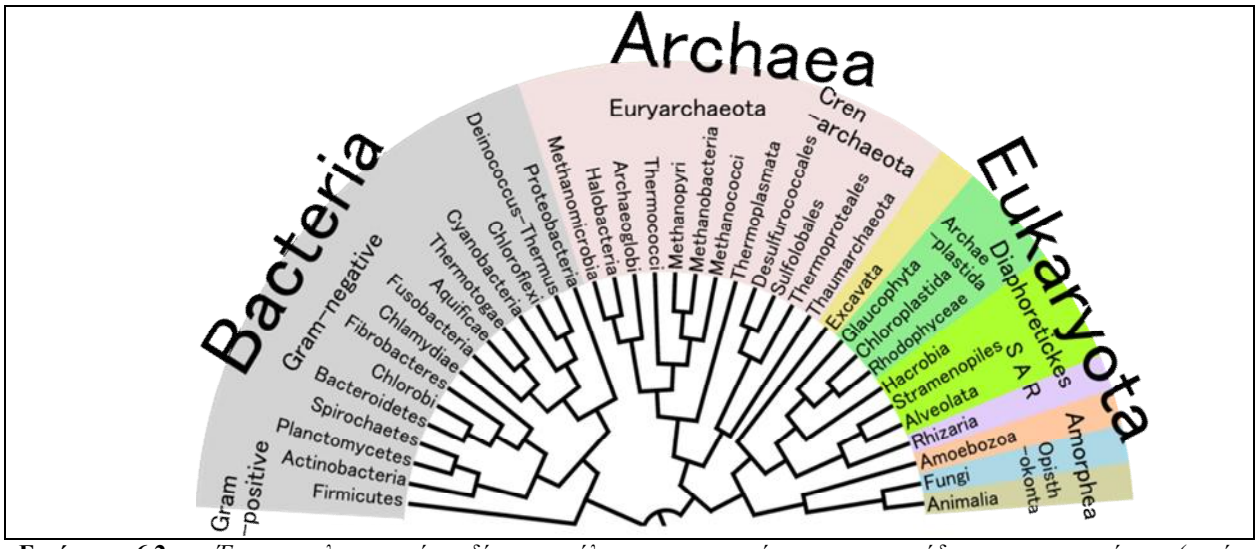


Εικόνα 6.1: *Παράδειγμα ενός δέντρου με δύο κλάδους (I), και ενός άλλου με 8 (II). Και τα δύο δέντρα είναι με ρίζα.*

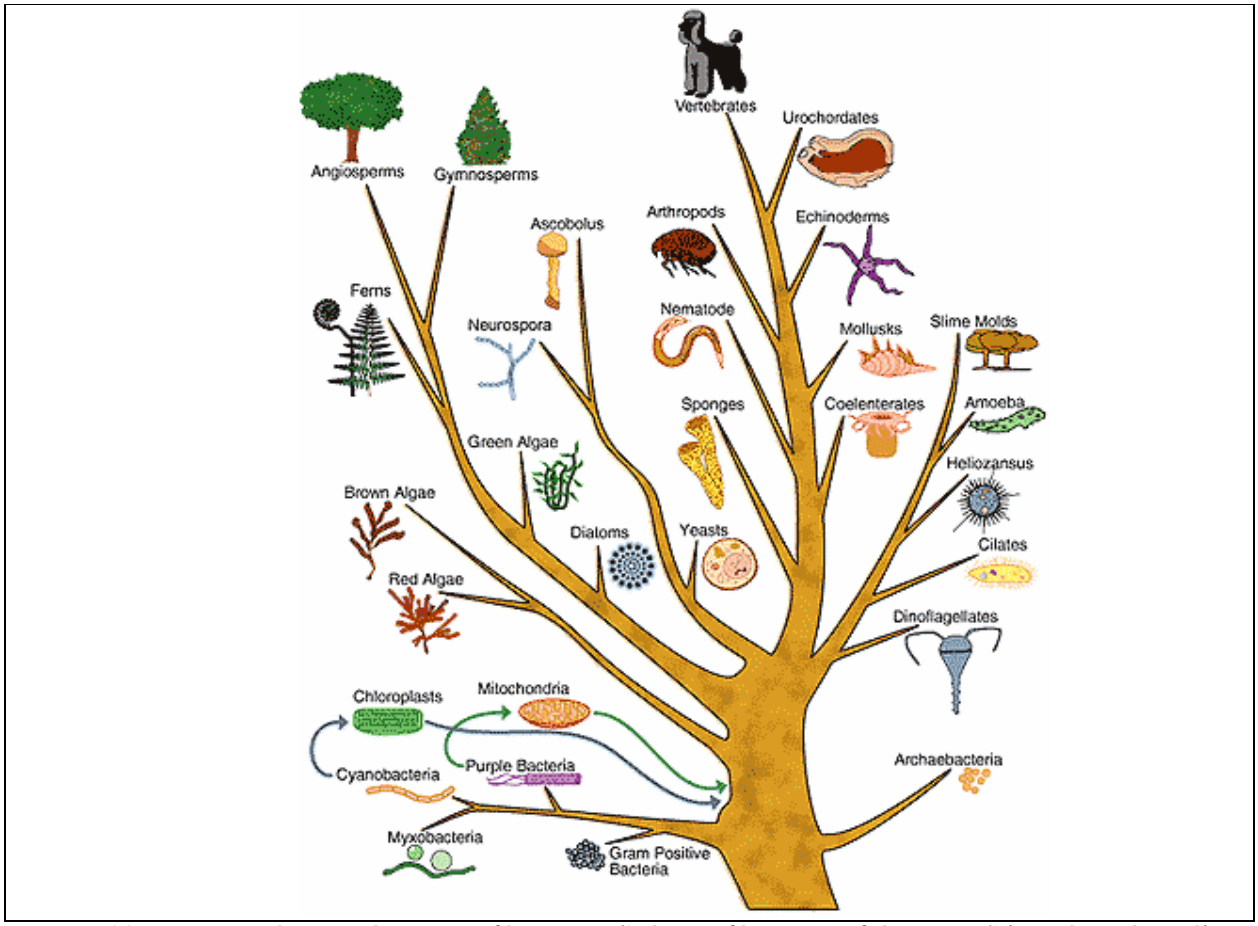
Προτού προχωρήσουμε, είναι απαραίτητο να δώσουμε κάποιους ορισμούς που αφορούν τα φυλογενετικά δέντρα. Ένα φυλογενετικό δέντρο είναι μια αναπαράσταση που συμβολίζει την εξελικτική διαδικασία. Όταν έχουμε κάποιες αλληλουχίες και θέλουμε να εκτιμήσουμε τις φυλογενετικές τους σχέσεις, μια αναπαράσταση σε μορφή δέντρου μας δείχνει πόσο κοντά βρίσκεται η μια αλληλουχία στην άλλη, δηλαδή με ποια σειρά οι αλληλουχίες εξελίχθηκαν η μια από την άλλη, έτσι ώστε, γυρνώντας πίσω στο χρόνο να εντοπίσουμε τελικά τον κοινό τους πρόγονο. Οι ακμές αυτού του δέντρου, είναι οι αλληλουχίες ή γενικότερα, οι ταξινομικές βαθμίδες (taxa) οι οποίες συγκρίνονται. Οι κόμβοι στο δέντρο, δείχνουν τα σημεία διακλάδωσης, δηλαδή το χρονικό σημείο ύπαρξης κοινού προγόνου. Τα μήκη των βραχιόνων, από έναν κόμβο σε μία ακμή, συμβολίζουν τον χρόνο που έχει περάσει. Έναν κλάδο, αποτελούν όλοι οι βραχίονες που ξεκινάνε από έναν κόμβο, και αυτός συμβολίζει μια μονοφυλετική ομάδα (μια ομάδα η οποία περιλαμβάνει όλους τους οργανισμούς που προέρχονται από τον κοινό πρόγονο, χωρίς όμως να περιέχει άλλους οργανισμούς που δεν κατάγονται από αυτόν). Στην Εικόνα 6.1 δίνονται δυο παραδείγματα δέντρων με 2 και 8 αλληλουχίες αντίστοιχα.

Οι βασικές αρχές της φυλογενετικής ανάλυσης, μπορούμε να πούμε ότι στηρίζονται σε μερικές απλές παραδοχές (Brinkman & Leipe, 2001):

- Οποιαδήποτε ομάδα οργανισμών (ή αλληλουχιών) προέρχεται από κάποιον κοινό πρόγονο μέσω της εξέλιξης. Αν οι οργανισμοί (ή οι αλληλουχίες) είναι πολύ διαφορετικοί, ο κοινός πρόγονος υπάρχει αλλά βρίσκεται πολύ πίσω στον εξελικτικό χρόνο.
- Υπάρχει διχαλωτό πρότυπο στην εξέλιξη. Η διαδικασία της εξέλιξης οδηγεί πάντα σε διχοτόμηση ενός taxon ή μιας αλληλουχίας, έτσι ώστε να δημιουργούνται δύο βραχίονες κάτω από έναν κόμβο.
- Αλλαγή στα παρατηρήσιμα χαρακτηριστικά των οργανισμών εμφανίζεται μετά το πέρασμα πολλών γενιών.

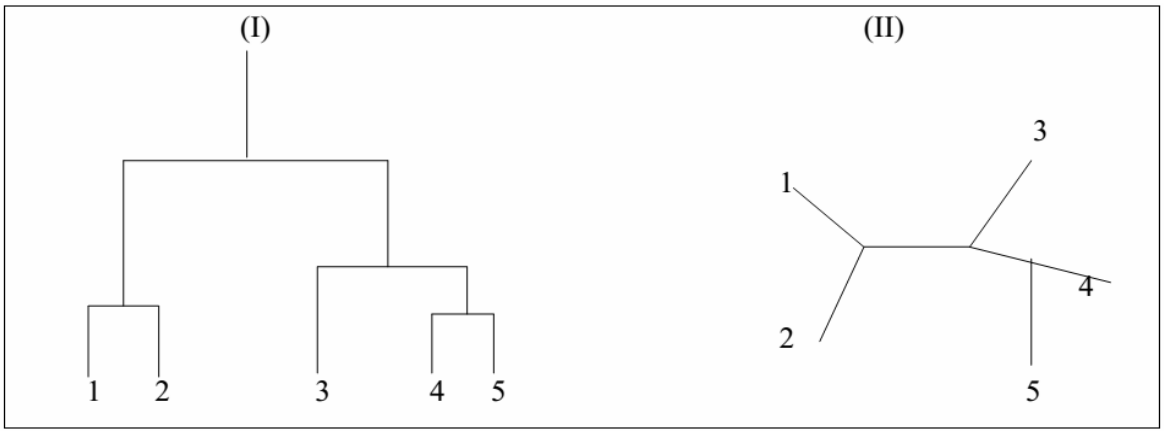


Εικόνα 6.2: *Ένα φυλογενετικό δέντρο όλων των σύγχρονων ομάδων οργανισμών (πηγή: http://commons.wikimedia.org/wiki/File:Phylogenetic_Tree_of_Life.png)*



Εικόνα 6.3: *Μια γραφική αναπαράσταση του δέντρου της ζωής, του δέντρου που δείχνει την εξελικτική συγγένεια όλων των σύγχρονων ομάδων οργανισμών. Η πληροφορία είναι βασικά η ίδια με αυτή της εικόνας 6.2 με τη διαφορά ότι τα μήκη των κλαδιών δεν αντικατοπτρίζουν τους εξελικτικούς χρόνους. Για παράδειγμα, τα βακτήρια και τα αρχαία, φαίνονται σαν δυο μικροί σε μήκος κλάδοι, παρόλο που περιέχουν πολλές και διακριτές μεταξύ τους ομάδες οργανισμών (πηγή: http://creationwiki.org/Macroevolution)*

Τα δέντρα που είδαμε στις προηγούμενες εικόνες είναι δέντρα με ρίζα (rooted). Σε αυτά έχουμε ξεκάθαρη κατεύθυνση του χρόνου, και έτσι μπορούμε να προσδιορίσουμε τον αρχαίο κοινό προγονό (ο οποίος όπως είπαμε, είναι σίγουρο ότι υπάρχει). Εναλλακτικά μπορούμε να έχουμε δέντρα χωρίς ρίζα (unrooted), στα οποία δεν μπορούμε να προσδιορίσουμε την κατεύθυνση κατά την οποία έχει συντελεστεί η εξελικτική διαδικασία. Τέτοια δέντρα δημιουργούνται συνήθως από κάποιους αλγορίθμους (είναι περιορισμός των αλγορίθμων αυτών). Παρακάτω (Εικόνα 6.4) βλέπουμε ένα παράδειγμα για δέντρο με ρίζα (I), και ένα χωρίς ρίζα (II), αμφότερα για 5 αλληλουχίες.



Εικόνα 6.4: *Τυπικά παραδείγματα πιθανών δέντρων για 5 αλληλουχίες. (I) δέντρο με ρίζα, (II) δέντρο χωρίς ρίζα*

Το αν θα έχουμε τελικά δέντρο με ή χωρίς ρίζα είναι αποτέλεσμα της μεθόδου που χρησιμοποιείται καθώς άλλες μέθοδοι παράγουν δέντρα με ρίζα και άλλες χωρίς. Πάντως ακόμα και αν έχουμε δέντρο χωρίς ρίζα είναι δυνατόν να προστεθεί εκ των υστέρων (και μάλιστα, το επιδιώκουμε αυτό), αν συγκρίνουμε όλους τους οργανισμούς με ένα άλλο είδος για το οποίο ξέρουμε ότι απέχει «πολύ» εξελικτικά, από τα υπό μελέτη είδη του δέντρου (το είδος αυτό ονομάζεται εξωομάδα-outgroup). Όσον αφορά τον αριθμό των κλάδων που έχουν τα παραπάνω δέντρα για L ακολουθίες, ξέρουμε από την συνδυαστική ότι για τα δέντρα με ρίζα θα είναι $2L$-1 (1,2,…L για τα τελικά κλαδιά που αντιστοιχούν στις L ακολουθίες και L+1, L+2,…, $2L$-1 για τα εσωτερικά κλαδιά), και για τα δέντρα χωρίς ρίζα $2L$-3. Ο αριθμός N των πιθανών δέντρων που αντιστοιχούν σε L ακολουθίες θα είναι, για τα δέντρα με ρίζα

$$N_{rooted} = \frac{(2L-3)!}{2^{L-2}(L-2)!}$$

ενώ για τα δέντρα χωρίς ρίζα αντίστοιχα, θα έχουμε:

$$N_{unrooted} = \frac{(2L-5)!}{2^{L-3}(L-3)!}$$

Έτσι για παράδειγμα, αν έχουμε L=10 ακολουθίες, τότε μπορεί να κατασκευαστούν $N\approx$35 εκ. δέντρα με ρίζα και $N\approx$2εκ. δεντρα χωρίς ρίζα. Γενικά τα πιθανά δέντρα με ρίζες θα είναι $2L$-3 φορές περισσότερα από τα αντίστοιχα χωρίς ρίζα.

Γενικά η διαδικασία φυλογενετικής ανάλυσης και η κατασκευή φυλογενετικών δέντρων, αποτελείται από τέσσερα διακριτά σημεία:

- Μία πολλαπλή στοίχιση. Από αυτήν ξεκινάνε όλα, και όλα βασίζονται σε αυτή. Αν η αρχική στοίχιση είναι λάθος, όλες οι παρακάτω αναλύσεις θα είναι επισφαλείς. Γι' αυτό, πολλές φορές χρειάζεται εμπειρία και χειροκίνητη επεξεργασία.
- Καθορισμός του μοντέλου αντικατάστασης, δηλαδή του μαθηματικού μοντέλου της εξελικτικής αλλαγής. Αυτή είναι μια απαίτηση των περισσότερων μεθόδων (με εξαίρεση αυτή της μέγιστης φειδωλότητας), και χρειάζεται ιδιαίτερη προσοχή, καθώς ένα απλό μοντέλο μπορεί να κάνει εύκολους τους υπολογισμούς αλλά μπορεί να μην είναι ρεαλιστικό.
- Κατασκευή του δέντρου. Σε αυτό το σημείο, υπάρχουν οι βασικότερες διαφοροποιήσεις των αλγορίθμων. Κάποιες μέθοδοι είναι γρήγορες, άλλες πιο χρονοβόρες, άλλες κάνουν περισσότερες υποθέσεις κ.ο.κ. Γενικά, οι μέθοδοι χωρίζονται σε δύο μεγάλες κατηγορίες,

στις μεθόδους που χρησιμοποιούν απόσταση και στις μεθόδους που χρησιμοποιούν χαρακτήρες.

- Αξιολόγηση του δέντρου. Αφού το δέντρο κατασκευαστεί, πρέπει να υπάρχει και ένας τρόπος να υπολογιστεί η αξιοπιστία του. Ανάλογα με τη μέθοδο κατασκευής, και με το χρησιμοποιούμενο λογισμικό, μπορεί να υπάρχουν και διαφορετικοί τρόποι ελέγχου της αξιοπιστίας του δέντρου.

Θα προσπαθήσουμε να αναλύσουμε αυτά τα θέματα (με την εξαίρεση της πολλαπλής στοίχισης, την οποία μελετήσαμε σε προηγούμενο κεφάλαιο), με τη σειρά που τέθηκαν παραπάνω. Στο τέλος, θα παρουσιάσουμε το διαθέσιμο λογισμικό και θα δώσουμε μερικές πρακτικές συμβουλές.

6.2. Πιθανοθεωρητικά Μοντέλα της Εξέλιξης των Νουκλεοτιδικών Αλληλουχιών

Το πρώτο θέμα που χρειάζεται σχεδόν σε όλες τις φυλογενετικές αναλύσεις, με την εξαίρεση της φειδωλότητας, και αφού θεωρήσουμε δεδομένη την πολλαπλή στοίχιση, είναι ο καθορισμός του μοντέλου με το οποίο θεωρούμε ότι έχει συντελεστεί η εξελικτική διαδικασία. Προχωρώντας στην κατασκευή πιθανοθεωρητικών μοντέλων που περιγράφουν την εξέλιξη των νουκλεοτιδικών αλληλουχιών μέσω της μετάλλαξης θα συναντήσουμε την έννοια της ανέλιξης Markov. Η διαφορά εδώ σε σχέση με τα μοντέλα που θα παρουσιαστούν στο κεφάλαιο 8, είναι ότι δεν ενδιαφερόμαστε για το ποιο νουκλεοτίδιο ακολουθεί κάποιο άλλο στην ίδια αλυσίδα αλλά ποιο νουκλεοτίδιο θα αντικαταστήσει ένα συγκεκριμένο, σε κάποια μελλοντική χρονική στιγμή. Έτσι, οι πιθανότητες μεταβάσεως θα είναι (Durbin, Eddy, Krogh, & Mitchison, 1998):

$$p_{abt} = P(x_i = b \mid x_i = a, t) \tag{6.1}$$

δηλαδή η πιθανότητα το νουκλεοτίδιο *b* να αντικαταστήσει το *α* στην θέση *i* της αλληλουχίας *x* έπειτα από χρόνο *t*. Όπως είναι φανερό, εδώ έχουμε μια ανέλιξη Markov διακριτών καταστάσεων σε συνεχή χρόνο. Αν τώρα έχουμε δυο αλληλουχίες DNA $\mathbf{x} = x_1, x_2, ..., x_n$ και $\mathbf{y} = y_1, y_2, ..., y_n$, η πιθανότητα η **x** να έχει προκύψει από την **y** σε χρόνο *t* είναι:

$$P(\mathbf{x} \mid \mathbf{y}, t) = \prod_{i=1}^{n} P(x_i \mid y_i, t) \tag{6.2}$$

Ορίζουμε στη συνέχεια, έναν 4x4 πίνακα πιθανοτήτων μεταβάσεως ή υποκαταστάσεως ο οποίος εξαρτάται από το *t*,

$$S(t) = \begin{bmatrix} P(A \mid A,t) & P(T \mid A,t) & P(G \mid A,t) & P(C \mid A,t) \\ P(A \mid T,t) & P(T \mid T,t) & P(G \mid T,t) & P(C \mid T,t) \\ P(A \mid G,t) & P(T \mid G,t) & P(G \mid G,t) & P(C \mid G,t) \\ P(A \mid C,t) & P(T \mid C,t) & P(G \mid C,t) & P(C \mid C,t) \end{bmatrix} \tag{6.3}$$

Στον πίνακα αυτό πρέπει να ισχύουν

$$\begin{aligned} &p_{i,j} \geq 0 \text{ με } i, j = 1,2,3,4 \text{ και} \\ &\sum_{j=1}^{4} p_{i,j} = 1 \text{ για κάθε } i \end{aligned} \tag{6.4}$$

δηλαδή ο πίνακας αυτός είναι στοχαστικός.

Πρέπει να τονίσουμε εδώ ότι μια ανέλιξη Markov, σαν αυτές που περιγράφουμε εδώ, μπορεί να έχει τρεις βασικές ιδιότητες (Lio & Goldman, 1998). Πρώτον, η αλυσίδα Markov μπορεί να είναι ομογενής χρονικά (homogeneity), δηλαδή οι πιθανότητες μεταβάσεως να μην εξαρτώνται από το χρόνο. Σε μια ομογενή αλυσίδα οδηγούμαστε τελικά σε μια κατάσταση ισορροπίας (equilibrium). Δεύτερον, η αλυσίδα Markov να είναι στάσιμη (stationary), δηλαδή σε κάθε χρονική στιγμή η κατανομή των βάσεων είναι αυτή της κατάστασης ισορροπίας. Και τρίτον, είναι δυνατόν να ισχύει η αντιστρεπτότητα (reversibility) των πιθανοτήτων μεταβάσεως, δηλαδή οι πιθανότητες να είναι ίδιες και για τις αντίστροφες μεταβάσεις. Στα μοντέλα φυλογενετικής εξέλιξης συνήθως υποθέτουμε ότι πληρούν και τις 3 παραπάνω προϋποθέσεις, για λόγους υπολογιστικής απλότητας.

Έτσι, αν ισχύουν τα παραπάνω τότε οι εξισώσεις Chapman-Kolmogorov γίνονται:

$$S(t)S(s) = S(t+s) \tag{6.5}$$

Τέλος, είναι αναγκαίο να ορίσουμε έναν πίνακα των ρυθμών υποκαταστάσεως ή αντικαταστάσεως (Substitution Rate Matrix) *R* έτσι ώστε:

$$R = \begin{bmatrix} \delta & \alpha & \beta & \gamma \\ \alpha & \delta & \gamma & \beta \\ \beta & \gamma & \delta & \alpha \\ \gamma & \beta & \alpha & \delta \end{bmatrix} \tag{6.6}$$

στον οποίο για να πληρούνται και οι 3 παραπάνω προϋποθέσεις, πρέπει να ισχύει:

$$\delta = -(a+\beta+\gamma)$$

δηλαδή οι γραμμές και οι στήλες του να αθροίζουν στο 0. Επειδή:

$$S(t) = \exp(Rt) \cong I + Rt + \frac{(Rt)^2}{2!} + \frac{(Rt)^3}{3!} + \ldots$$

αν προχωρήσουμε σε φασματική αποικοδόμηση (spectral decomposition), ισχύει επιπλέον:

$$S(t) = U diag\left\{e^{\lambda_1 t}, \ldots, e^{\lambda_n t}\right\} U^{-1}$$

όπου λ_i οι ιδιοτιμές (eigenvalues) του *R*, και *U* το αντίστοιχο ιδιοδιάνυσμα. Ο πίνακας υποκαταστάσεως για ένα «μικρό» χρονικό διάστημα ε γίνεται:

$$S(\varepsilon) = I + R\varepsilon \Rightarrow S(t+\varepsilon) = S(t)S(\varepsilon) = S(t)(I + R\varepsilon)$$
$$\Rightarrow \frac{S(t+\varepsilon) - S(t)}{\varepsilon} \approx S(t)(I + R\varepsilon)$$

και παίρνοντας το όριο καθώς $\varepsilon \to 0$ θα έχουμε:

$$S'(t) = S(t)R \tag{6.7}$$

Λύνοντας αυτές τις εξισώσεις μπορούμε να πάρουμε τις τιμές για τις πιθανότητες μεταβάσεως. Στην περίπτωση που στη σχέση (6.6) έχουμε $a=\beta=\gamma$ τότε:

$$R = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix} \qquad S(t) = \begin{bmatrix} r_t & s_t & s_t & s_t \\ s_t & r_t & s_t & s_t \\ s_t & s_t & r_t & s_t \\ s_t & s_t & s_t & r_t \end{bmatrix}$$

και προκύπτουν οι λύσεις:

$$\begin{aligned} r_t &= \tfrac{1}{4}\left(1 + 3e^{-4at}\right) \\ s_t &= \tfrac{1}{4}\left(1 - e^{-4at}\right) \end{aligned} \tag{6.8}$$

Το μοντέλο αυτό ήταν το πρώτο σχετικό μοντέλο που προτάθηκε από τους Jukes και Cantor (Jukes & Cantor, 1969) και χαρακτηρίζεται ως ένα απλό μοντέλο ανέλιξης Poisson (για συντομία, ονομάζεται JC69). Είναι το πιο απλό ανάμεσα στα σχετικά μοντέλα, αλλά χάνει με αυτόν τον τρόπο κάποια σημαντικά χαρακτηριστικά της εξελικτικής διαδικασίας. Για παράδειγμα δεν αποδίδει σωστά το γεγονός ότι οι μεταπτώσεις (πουρίνη σε πουρίνη) δεν έχουν τον ίδιο ρυθμό με τις μεταστροφές (πουρίνη σε πυριμιδίνη και αντίστροφα). Η επόμενη παραλλαγή, είναι το μοντέλο του διάσημου εξελικτικού βιολόγου Kimura (Kimura, 1980) το οποίο συμβολίζεται ως K2P και προβλέπει:

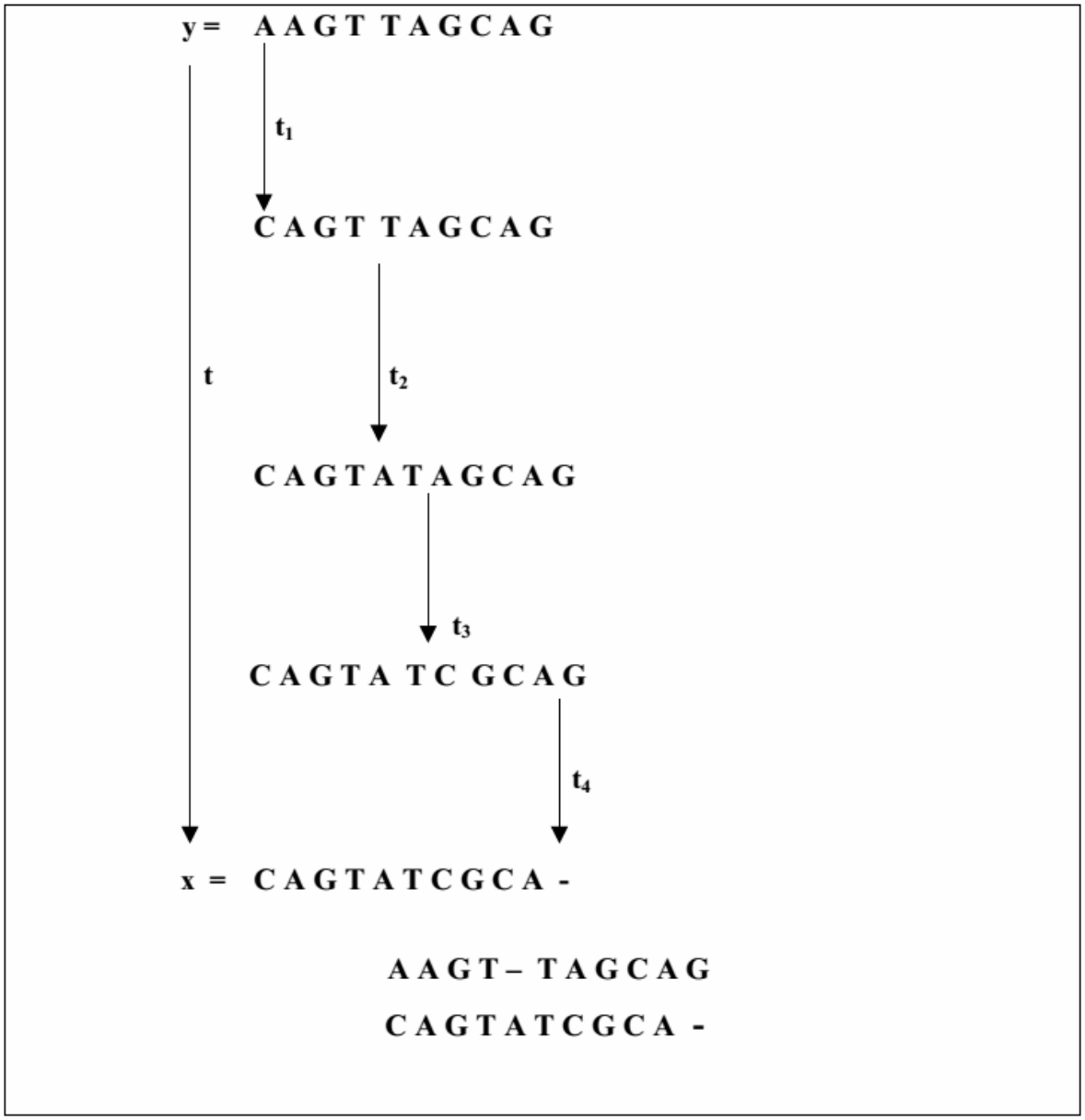
$$R = \begin{bmatrix} -2\beta - \alpha & \beta & \alpha & \beta \\ \beta & -2\beta - \alpha & \beta & \alpha \\ \alpha & \beta & -2\beta - \alpha & \beta \\ \beta & \alpha & \beta & -2\beta - \alpha \end{bmatrix} \qquad S(t) = \begin{bmatrix} r_t & s_t & u_t & s_t \\ s_t & r_t & s_t & u_t \\ u_t & s_t & r_t & s_t \\ s_t & u_t & s_t & r_t \end{bmatrix}$$

με λύσεις

$$s_t = \tfrac{1}{4}\left(1 - e^{-4\beta t}\right)$$
$$u_t = \tfrac{1}{4}\left(1 + e^{-4\beta \mathbf{t}} - 2e^{-2(\alpha+\beta)\mathbf{t}}\right)$$

και

$$r_t = 1 - 2s_t - u_t \tag{6.9}$$



Εικόνα 6.5: *Σχηματική αναπαράσταση της εξελικτικής διαδικασίας μέσα από την οποία μια αλληλουχία y, μέσα από διαδοχικές μεταλλάξεις (αντικαταστάσεις, απαλοιφές, εισαγωγές), μετά από την πάροδο μεγάλου χρονικού διαστήματος t, οδηγεί σε μια αλληλουχία x.*

Όπως είναι φανερό το μοντέλο αυτό είναι δι-παραμετρικό καθώς προβλέπει άλλες πιθανότητες υποκαταστάσεως για μεταπτώσεις (π.χ. A↔G, T↔C) και άλλες για μεταστροφές (π.χ. A↔T, G↔C). Πρέπει να υπενθυμίσουμε εδώ ότι βασική ιδιότητα των δυο αυτών μοντέλων (JC69, K2P) είναι ότι καθώς $t \to \infty$ ισχύει $q_A = q_T = q_G = q_C = 1/4$, δηλαδή στην κατάσταση ισορροπίας μετά την παρέλευση «άπειρου» χρόνου θα έχουμε μια ισοκατανομή των βάσεων του DNA. Αυτό όμως ξέρουμε ότι δεν είναι τόσο ρεαλιστικό καθώς οι οργανισμοί παρουσιάζουν μεγάλη μεταβλητότητα στο λόγο των βάσεων (A+T)/(G+C) και για να αντιμετωπισθεί αυτό έχουν προταθεί άλλα πιο σύνθετα μοντέλα (Felsenstein, 1981; Lio & Goldman, 1998; Penny & Hendy, 2001). Σ 'αυτά, ο πίνακας των ρυθμών υποκαταστάσεως δεν έχει την ιδιότητα (6.6) αλλά στηρίζεται σε παρατηρηθείσες συχνότητες νουκλεοτιδίων στις υπό μελέτη ακολουθίες (π_A, π_G, π_C, π_T), και

κατά συνέπεια επιτρέπει στην κατάσταση ισορροπίας να έχουμε διαφορετικές κατανομές των νουκλεοτιδίων, μια προσέγγιση η οποία είναι πιο ρεαλιστική. Για παράδειγμα, το μοντέλο F81 του Felsenstein (Felsenstein, 1981), μοιάζει αρκετά με το κλασικό μοντέλο JC69, στο ότι χρησιμοποιεί μόνο μία παράμετρο (μ) για τις μεταπτώσεις και τις μεταστροφές, αλλά μοντελοποιεί τα τέσσερα νουκλεοτίδια με διαφορετικές πιθανότητες εμφάνισης (π_A, π_G, π_C, π_T):

$$R = \begin{bmatrix} -\mu(\pi_C + \pi_G + \pi_T) & \mu\pi_C & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\pi_A + \pi_G + \pi_T) & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & -\mu(\pi_C + \pi_A + \pi_T) & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & \mu\pi_G & -\mu(\pi_C + \pi_G + \pi_A) \end{bmatrix} \quad (6.10)$$

Μια λογική επέκταση αυτού του μοντέλου, είναι εφικτή αν θεωρήσουμε ότι οι μεταπτώσεις έχουν διαφορετικό ρυθμό από τις μεταστροφές, χρησιμοποιώντας δύο παραμέτρους (κ και μ):

$$R = \begin{bmatrix} -\mu(\pi_C + \kappa\pi_G + \pi_T) & \mu\pi_C & \mu\kappa\pi_G & \mu\kappa\pi_T \\ \mu\pi_A & -\mu(\pi_A + \pi_G + \kappa\pi_T) & \mu\kappa\pi_G & \mu\kappa\pi_T \\ \mu\kappa\pi_A & \mu\pi_C & -\mu(\pi_C + \kappa\pi_A + \pi_T) & \mu\kappa\pi_T \\ \mu\pi_A & \mu\kappa\pi_C & \mu\pi_G & -\mu(\kappa\pi_C + \pi_G + \pi_A) \end{bmatrix} \quad (6.11)$$

Αυτό είναι το μοντέλο των Hasegawa-Kishino-Yano (Hasegawa, Kishino, & Yano, 1985) (HKY85), ενώ παρόμοιο είναι και το μεταγενέστερο μοντέλο του Felsenstein, το λεγόμενο F84, το οποίο μοντελοποιεί το ίδιο φαινόμενο αλλά διαφέρει στην παραμετροποίηση.

Τέλος, το πιο γενικό μοντέλο αυτής της κατηγορίας, είναι το λεγόμενο γενικό χρονικά αντιστρεπτό μοντέλο (general time reversible model), το οποίο συμβολίζεται ως GTR (Tavare, 1986) και περιγράφεται από τη σχέση:

$$R = \begin{bmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu a\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu b\pi_A & \mu d\pi_C & -\mu(d\pi_C + b\pi_A + f\pi_T) & \mu\kappa\pi_T \\ \mu c\pi_A & \mu e\pi_C & \mu f\pi_G & -\mu(e\pi_C + f\pi_G + c\pi_A) \end{bmatrix} \quad (6.12)$$

Όπως είναι φανερό, το μοντέλο αυτό είναι το πιο γενικό και μπορεί να συμπεριλάβει όλα τα προηγούμενα ως ειδικές περιπτώσεις. Αντίστοιχα, το HKY85 περιλαμβάνει σαν ειδικές περιπτώσεις τα F81, K2P και JC69, ενώ το F81 και το K2P περιλαμβάνουν το καθένα σαν (διαφορετική) ειδική περίπτωση το JC69. Γενικά, όπως και σε κάθε διαδικασία μοντελοποίησης το πιο σύνθετο μοντέλο είναι και το καλύτερο, αλλά, από την άλλη απαιτεί περισσότερα δεδομένα και αλλά και υπολογιστική ισχύ, καθώς χρειάζεται η εκτίμηση μεγαλύτερου αριθμού παραμέτρων. Πάντως, με τους σύγχρονους υπολογιστές και τις μεθόδους εκτίμησης, ο αριθμός των παραμέτρων και η πολυπλοκότητα του μοντέλου δεν αποτελεί πρακτικό πρόβλημα, γι' αυτό και στις περισσότερες πρακτικές αναλύσεις τα σύγχρονα λογισμικά χρησιμοποιούν πλέον κατά βάση το πιο γενικό μοντέλο, το GTR.

Άλλα μοντέλα, ακόμα πιο σύνθετα, έχουν προταθεί κατά καιρούς, τα οποία όμως δεν ασχολούνται με τα ίδια τα νουκλεοτίδια αλλά με τα κωδικόνια επιτρέποντας έτσι απευθείας υπολογισμούς που ανάγονται στο πρωτεϊνικό επίπεδο (Yang, 1994). Έχουν προταθεί και άλλα μοντέλα τα οποία επιτρέπουν διαφορετικούς ρυθμούς αντικατάστασης θεωρώντας ότι αυτοί προέρχονται από έναν πληθυσμό που ακολουθεί την κατανομή Γάμμα (το λεγόμενο random effects model) (Yang, 1993). Πολλές φορές, το μοντέλο της ετερογένειας των ρυθμών εξέλιξης για τις διάφορες θέσεις, μπορεί να συνδυαστεί με κάποιο από τα μοντέλα που παρουσιάσαμε πριν, οπότε μιλάμε για το μοντέλο JC69+Γ, ή GTR+Γ, κ.ο.κ. Η ύπαρξη σταθερών ρυθμών αντικατάστασης είναι γνωστή ως η υπόθεση του «μοριακού ρολογιού» (molecular clock).

Παρόλο που δεν είναι τόσο συνηθισμένο, υπάρχουν και μοντέλα τα οποία δουλεύουν κατευθείαν πάνω σε αμινοξικές αλληλουχίες πρωτεϊνών. Η πιο φυσική επιλογή σε μια τέτοια περίπτωση, είναι οι πίνακες PAM (Dayhoff, Schwartz, & Orcutt, 1978) οι οποίοι έχουν προκύψει από ένα ξεκάθαρο (και μαρκοβιανό) μοντέλο εξελικτικής αλλαγής, ανάλογο με αυτό της σχέσης (6.3). Οι πίνακες αυτής της οικογένειας έχουν ακριβώς τις ίδιες ιδιότητες (χρονική ομογένεια και αντιστρεπτότητα) και σε κατάσταση ισορροπίας οδηγούν μια κατανομή των αμινοξέων ίδια με αυτή που είχε η βάση δεδομένων από την οποία προήλθαν. Πρέπει να τονιστεί εδώ η ολοφάνερη εξελικτική έννοια που παίρνουν οι πίνακες αντικατάστασης αμινοξέων και νουκλεοτιδίων που συζητήσαμε στις τεχνικές στοίχισης αλληλουχιών στο 3^{ο} κεφαλαίο, καθώς μπορούν να

ιδωθούν (Lio & Goldman, 1998) ως πίνακες μεταβάσεως της στοχαστικής ανελίξεως της μετάλλαξης και για την ακρίβεια, σαν το όριο τους καθώς $t\to\infty$ (δεχόμενοι δηλαδή ότι έχει περάσει «άπειρος» χρόνος και έχουμε φτάσει σε μια κατάσταση ισορροπίας, όχι όμως με ισοκατανομή των νουκλεοτιδίων ή των αμινοξέων). Μια τελική παρατήρηση αφορά στη χρονική συμμετρία όλων των σχετικών μοντέλων Markov που αναφέραμε, δηλαδή στην αδυναμία τους να διαχωρίσουν ποια από τις ακολουθίες προέκυψε από την άλλη, με άλλα λόγια δεν μπορούν αν χρησιμοποιηθούν σε μια ανάλυση μέγιστης πιθανοφάνειας να παράγουν φυλογενετικό δέντρο με ρίζα. Παρ' όλα αυτά έχουν προταθεί και μοντέλα που δεν χαρακτηρίζονται από χρονική συμμετρία (Lio & Goldman, 1998).

6.3. Μέθοδοι βασισμένες στην απόσταση

Η μία μεγάλη κατηγορία μεθόδων, είναι οι λεγόμενες *μέθοδοι που χρησιμοποιούν τις αποστάσεις* (*distance-based methods*). Οι μέθοδοι αυτές, ξεκινούν από μία πολλαπλή στοίχιση, υπολογίζουν με κάποιον τρόπο έναν πίνακα αποστάσεων για όλα τα ζευγάρια αλληλουχιών και μετά με βάση αυτόν τον πίνακα κατασκευάζουν το φυλογενετικό δέντρο (Durbin, et al., 1998). Η απόσταση (d_{ij}) μεταξύ των αλληλουχιών i,j , έχει συνήθως τις εξής ιδιότητες:

$$\begin{aligned} d_{ii} &= 0 \\ d_{ij} &= d_{ji} > 0,\ i \neq j \\ d_{ij} &\leq d_{ik} + d_{kj} \end{aligned} \tag{6.13}$$

Η μεγαλύτερη ομάδα από τις προαναφερόμενες μεθόδους, είναι στην ουσία κλασικές τεχνικές της *στατιστικής ομαδοποίησης* (*clustering*), και συγκεκριμένα, αυτές οι μέθοδοι που ονομάζονται «ιεραρχικές», οι οποίες παράγουν ένα είδους δέντρο το οποίο δηλώνει την ομαδοποίηση των δεδομένων. Τέτοια παραδείγματα, είδαμε ήδη στο κεφάλαιο 4 με εφαρμογές στην προοδευτική πολλαπλή στοίχιση.

Ο πίνακας των αποστάσεων, d_{ij}, δυο αλληλουχιών i και j, θα μπορούσε γενικά να προκύψει με πολλούς τρόπους. Ένας εύκολος τρόπος θα ήταν μετρώντας απλά το ποσοστό f από τις θέσεις u, στις οποίες τα κατάλοιπα x_u^i και x_u^j, διαφέρουν. Αυτό είναι ένα λογικό μέτρο, αλλά δεν αποδίδει καλά για ασυσχέτιστες ακολουθίες, καθώς θέλουμε σε αυτή την περίπτωση η απόσταση να αυξάνει. Μια καλύτερη λύση, προκύπτει από την αρχική πολλαπλή στοίχιση με χρήση κάποιου από τα πιθανοθεωρητικά μοντέλα της εξέλιξης που παρουσιάσαμε στην προηγούμενη ενότητα. Για παράδειγμα, το μοντέλο JC69 δίνει την απόσταση:

$$d_{ij} = -\frac{3}{4}\log\left(1-\frac{4}{3}f\right) \tag{6.14}$$

Για το K2P, αντίστοιχα θα έχουμε:

$$d_{ij} = -\frac{1}{2}\log\left(1-2f-g\right) - \frac{1}{4}\log\left(1-2g\right) \tag{6.15}$$

όπου f είναι το ποσοστό των αλλαγών που οφείλονται σε μεταπτώσεις και g το ποσοστό των αλλαγών που οφείλονται σε μεταστροφές. Παρόμοιοι υπολογισμοί, μπορούν να γίνουν και για τα υπόλοιπα μοντέλα, μόνο που οι εκφράσεις είναι πιο σύνθετες καθώς περιέχουν διαφορετικές συχνότητες για τις τέσσερις βάσεις.

Μόλις οι αποστάσεις υπολογιστούν, η πολλαπλή στοίχιση δεν χρησιμοποιείται ξανά, και όλοι οι υπολογισμοί γίνονται με χρήση του πίνακα. Η πρώτη τέτοια μέθοδος, προτάθηκε από τους Socal και Michener, (Sokal & Michener, 1958) και είναι η γνωστή *UPGMA* (*Unweighted Pair Group using Arithmetic Mean*), η οποία ορίζει την απόσταση μεταξύ δυο ομάδων (Clusters) να είναι η μέση απόσταση μεταξύ ζευγών αλληλουχιών από τις δύο ομάδες:

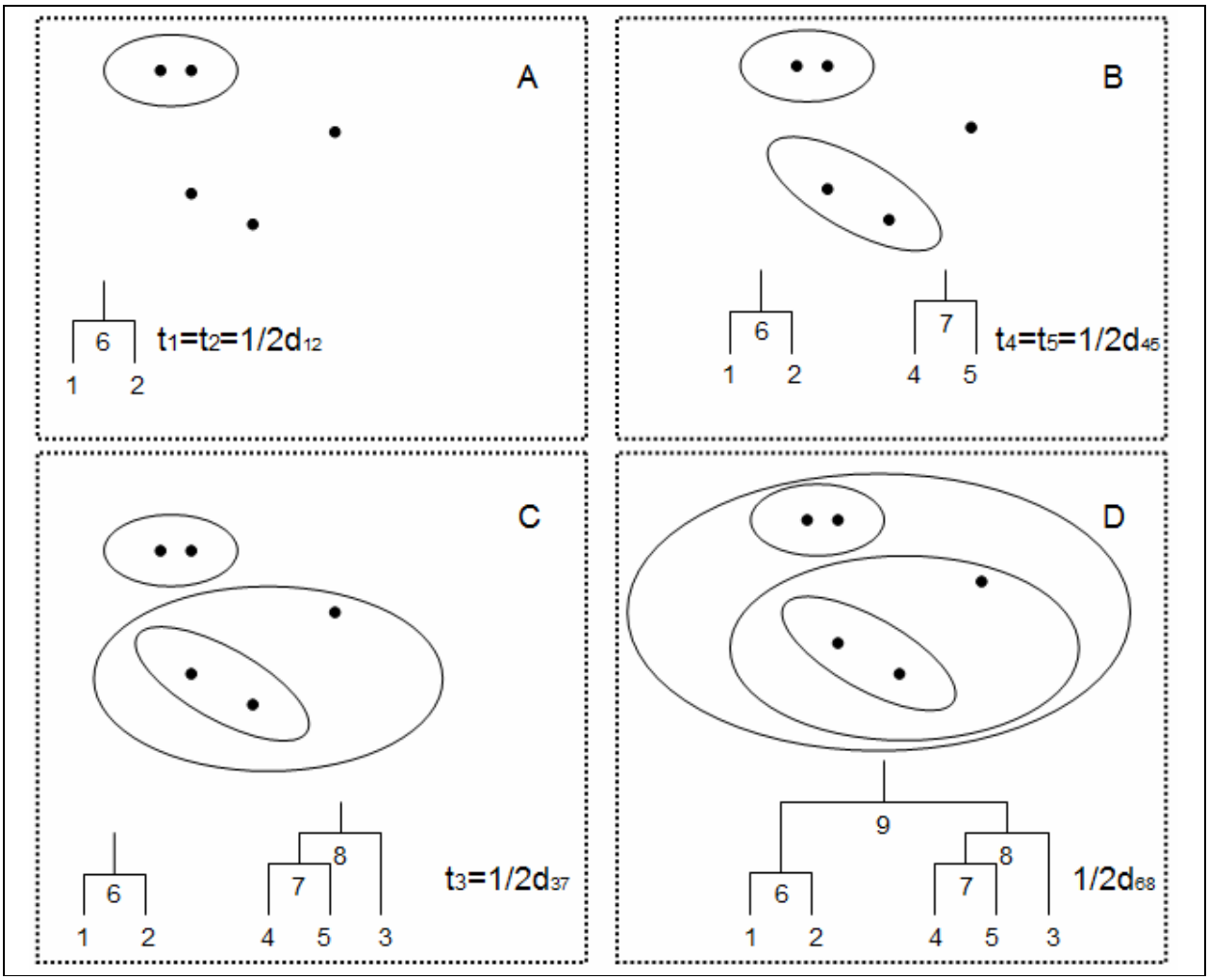
$$d_{ij} = \frac{1}{|C_i||C_j|}\sum_{p\in C_i, q\in C_j} d_{pq} \tag{6.16}$$

Σε αυτή τη σχέση, $|C_i|$ και $|C_j|$ είναι οι αριθμοί των αλληλουχιών στις ομάδες i και j. Η απόσταση της ομάδας k η οποία αποτελεί την ένωση των ομάδων i και j με μια άλλη ομάδα l θα δίνεται από τη σχέση:

$$d_{kl} = \frac{d_{il}|C_i| + d_{jl}|C_j|}{|C_i| + |C_j|} \tag{6.17}$$

Ο αλγόριθμος, δουλεύει όπως και οι αλγόριθμοι ιεραρχικής ομαδοποίησης που περιγράψαμε στην πολλαπλή στοίχιση: στην αρχή ξεκινάει τοποθετώντας κάθε μια ακολουθία στη δική της ομάδα, και

προχωράει ιεραρχικά, τοποθετώντας στην ίδια ομάδα τις ακολουθίες με τη μικρότερη απόσταση. Ο κόμβος σε κάθε βραχίονα τοποθετείται στο ύψος $d_{ij}/2$ (Εικόνα 6.6). Ο αλγόριθμος ανακατασκευάζει το δέντρο σε χρόνο της τάξης του $O(n^2)$. Η μέθοδος αυτή είναι απλή, διαισθητικά σωστή, εύκολα ερμηνεύσιμη και παράγει φυλογενετικά δέντρα με ρίζα. Παρ' όλα αυτά πολλές φορές μπορεί να δώσει λάθος αποτελέσματα, όταν δεν ικανοποιούνται κάποιες προϋποθέσεις, η βασικότερη από τις οποίες είναι ο σταθερός ρυθμός εξελικτικής διαδικασίας σε όλες τις αλληλουχίες, δηλαδή, το «μοριακό ρολόι». Αξίζει να αναφερθεί, ότι στη στατιστική ορολογία, η μέθοδος ονομάζεται «average linkage». Μέθοδοι που βασίζονται και στις υπόλοιπες κατηγορίες linkage (complete linkage, simple linkage κλπ), έχουν επίσης προταθεί για φυλογενετική ανάλυση, αλλά εμπειρικές αναλύσεις έδειξαν ότι δεν αποδίδουν τόσο καλά όσο η μέθοδος UPGMA.



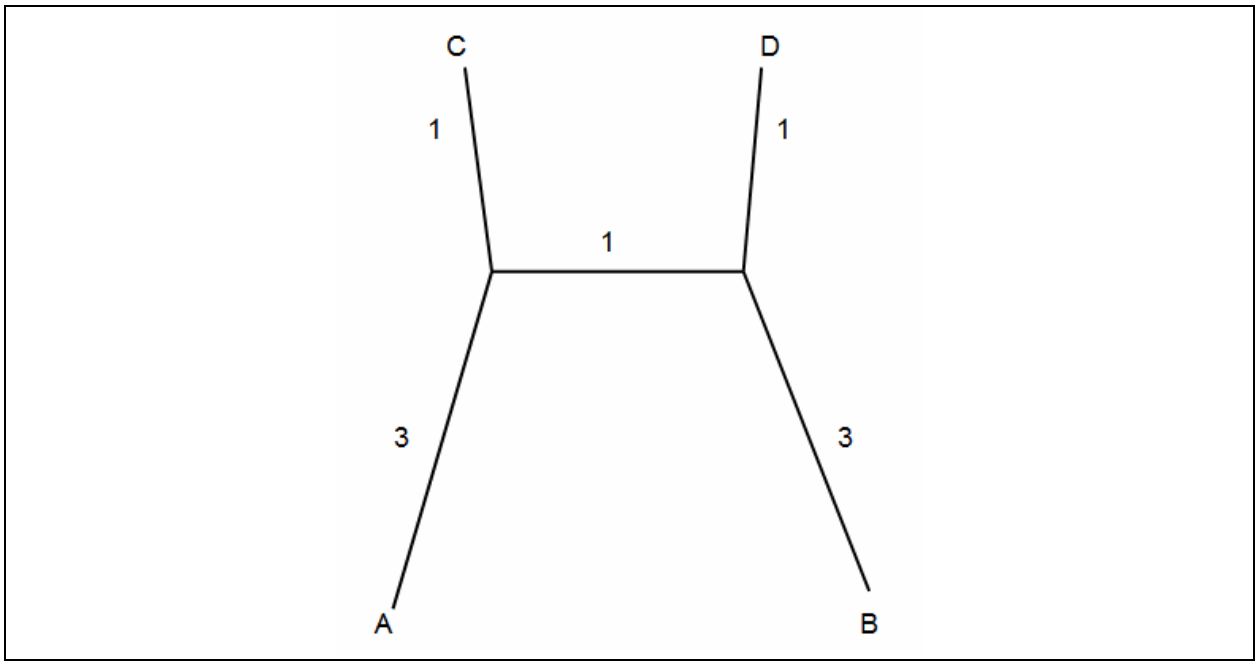
Εικόνα 6.6: *Υποθετικό παράδειγμα της μεθόδου UPGMA. Καταχρηστικά, οι πέντε αλληλουχίες αναπαρίστανται σαν να αντιστοιχούν σε σημεία στο επίπεδο (δεν ισχύει πάντα λόγω των διαφορετικών τρόπων ορισμού της απόστασης). Ο αλγόριθμος προχωράει σε βήματα (A,B,C,D) κατά τα οποία ανακατασκευάζει το δέντρο, ομαδοποιώντας σταδιακά τις πιο όμοιες αλληλουχίες.*

Η μέθοδος UPGMA, εκτός από την υπόθεση του μοριακού ρολογιού, έχει και μια άλλη ιδιότητα, παράγει εξ'ορισμού δέντρα στα οποία ισχύει η προσθετική ιδιότητα. Με αυτό, εννοούμε δέντρα στα οποία η απόσταση δύο οποιονδήποτε άκρων, είναι ίση με το άθροισμα των μηκών των ακμών που τα συνδέουν. Παρ' όλα αυτά, είναι δυνατόν να υπάρχουν περιπτώσεις στις οποίες δεν ισχύει η υπόθεση του μοριακού ρολογιού, αλλά η προσθετική ιδιότητα να εξακολουθεί να ισχύει. Σε αυτή την περίπτωση, πρέπει να αναζητηθεί κάποιος άλλος κατάλληλος αλγόριθμος. Η βασική ιδέα, είναι να έχουμε ένα δέντρο με αθροιστικό μήκος κλαδιών. Τότε, για δύο γειτονικά κλαδιά i, j τα οποία έχουν κοινό κόμβο τον k, θα πρέπει να αφαιρέσουμε τα κλαδιά

αυτά από το δέντρο, να προσθέσουμε το *k* σαν κλαδί του δέντρου και να ορίσουμε την απόστασή του από ένα άλλο κλαδί, *m*, ως:

$$d_{km} = \frac{1}{2}\left(d_{im} + d_{jm} - d_{ij}\right) \tag{6.18}$$

Επειδή ισχύει η προσθετική ιδιότητα, οι αρχικές αποστάσεις του δέντρου, διατηρούνται. Άρα, μπορούμε να προχωρήσουμε βήμα-βήμα και να αφαιρούμε κάθε φορά και από ένα κλαδί του δέντρου, μέχρι να ομαδοποιήσουμε όλες τις παρατηρήσεις. Το βασικό πρόβλημα, είναι να μπορέσουμε να εντοπίσουμε τα γειτονικά κλαδιά, χρησιμοποιώντας τις αποστάσεις και μόνο. Αυτό δεν είναι πάντα απλό, όπως φαίνεται στην Εικόνα 6.7.



Εικόνα 6.7: *Ένα παράδειγμα υποθετικού δέντρου στο οποίο φαίνεται ότι δύο γειτονικοί βραχίονες, είναι δυνατόν να μην είναι οι πιο κοντινοί. Οι αποστάσεις των γειτόνων (A-C και B-D) είναι ίσες με 4, αλλά οι αποστάσεις των μη γειτόνων (A-B και C-D) είναι μικρότερες (ίσες με 3). Το πρόβλημα αυτό, το λύνει η μετασχηματισμένη απόσταση που χρησιμοποιεί η μέθοδος Neighbour-Joining.*

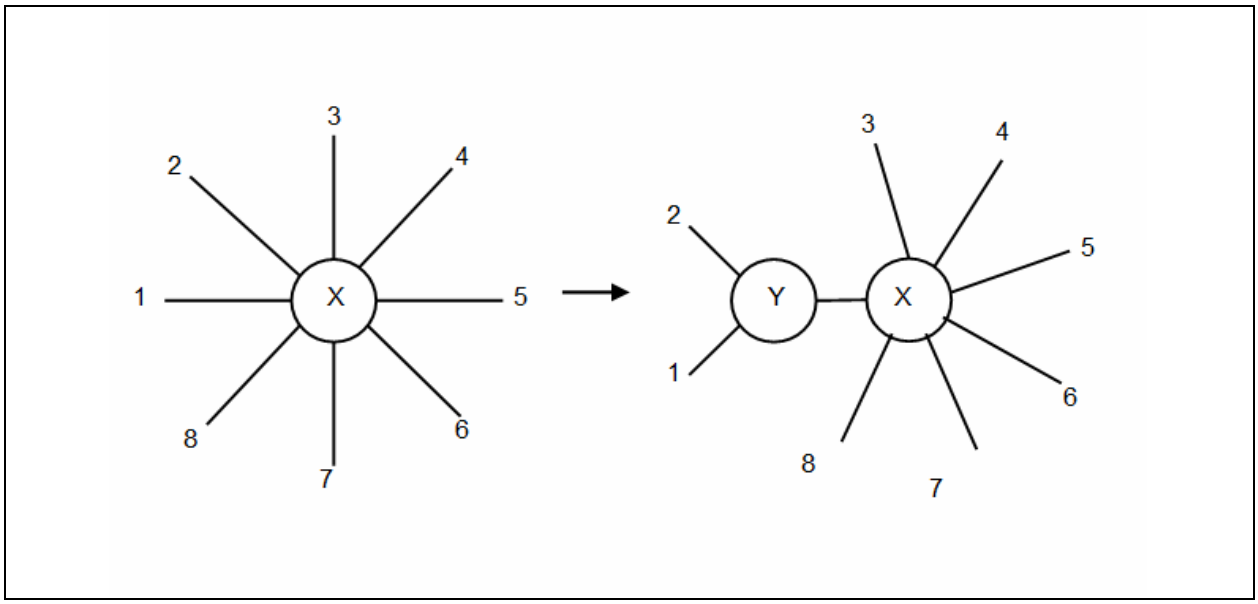
Η μέθοδος της *ένωσης γειτόνων* (*neighbour joining*, NJ) (Saitou & Nei, 1987), είναι ίσως μια από τις πιο συζητημένες και ευρέως χρησιμοποιούμενες μεθόδους αποστάσεων, η οποία ορίζει μια μετασχηματισμένη απόσταση:

$$D_{ij} = d_{ij} - \frac{1}{L-2}\sum_{\forall k}\left(d_{ik} + d_{jk}\right) \tag{6.19}$$

όπου *L* ο αριθμός των κλαδιών του δέντρου, και κατατάσσει τις αλληλουχίες σε ζευγάρια για να βρει τελικά τις πιο «γειτονικές». Με αυτόν τον τρόπο, λύνει το πρόβλημα των αποστάσεων που αναφέραμε στην Εικόνα 6.7 και εντοπίζει τα γειτονικά κλαδιά του δέντρου (δηλαδή, αυτά στα οποία το D_{ij} είναι ελάχιστο). Στο επόμενο βήμα, η μέθοδος θα ενώσει τα δύο επόμενα κλαδιά, κ.ο.κ (Εικόνα 6.8). Ο αλγόριθμος αυτός, για ένα δέντρο με *L* κλαδιά, απαιτεί *L*-3 επαναλήψεις και σε κάθε μία από αυτές, υπολογίζεται ο πίνακας D_{ij} ο οποίος έχει διαστάσεις *L*x*L*. Κατά συνέπεια, ο αλγόριθμος έχει πολυπλοκότητα της τάξης του $O(L^3)$, αν και υπάρχουν τροποποιήσεις που επιτυγχάνουν καλύτερες επιδόσεις.

Η μέθοδος NJ, όπως είπαμε, έχει το σημαντικό πλεονέκτημα ότι δεν απαιτεί την ύπαρξη του μοριακού ρολογιού για να δώσει σωστά αποτελέσματα. Επίσης, αν οι αποστάσεις ακολουθούν ή προσεγγίζουν την προσθετική ιδιότητα, τότε το δέντρο που ανακατασκευάζεται θα είναι πάντα το σωστό. Τέλος, είναι πολύ γρήγορη, και αυτό την κάνει ελκυστική, ειδικά για αναλύσεις μεγάλων συνόλων δεδομένων ή για εφαρμογή στατιστικών τεχνικών όπως το bootsrap. Σε σύγκριση με την UPGMA, είναι πιο αργή, αλλά αυτό αντισταθμίζεται από την χαλάρωση της απαίτησης του μοριακού ρολογιού. Από την άλλη, παράγει

φυλογενετικά δέντρα χωρίς ρίζα και για την εύρεση αυτής μπορούμε να χρησιμοποιήσουμε σαν σημείο αναφοράς ένα «μακρινό» είδος (εξωομάδα) το οποίο ξέρουμε ότι βρίσκεται πολύ μακριά εξελικτικά από όσα εξετάσαμε. Η μέθοδος, προτάθηκε ειδικά για φυλογενετικές αναλύσεις, αλλά είναι στην ουσία μια μέθοδος ομαδοποίησης, η οποία μπορεί να βρει εφαρμογή και σε άλλα πεδία, φτάνει να οριστεί κατάλληλα η απόσταση. Ένα τέτοιο παράδειγμα, είδαμε στην προοδευτική πολλαπλή στοίχιση, καθώς το γνωστό πρόγραμμα CLUSTAL, χρησιμοποιεί αυτόν τον αλγόριθμο για να κατασκευάσει το δέντρο-οδηγό.



Εικόνα 6.8: *Ένα υποθετικό παράδειγμα της λειτουργίας της μεθόδου Neighbour-Joining. Οι αλληλουχίες 1 και 2 έχουν τη μικρότερη (μετασχηματισμένη) απόσταση, και κατά συνέπεια ενώνονται για να δώσουν ένα νέο κόμβο (Υ). Στη συνέχεια, ο αλγόριθμος θα προχωρήσει ενώνοντας διαδοχικά κάθε φορά τους βραχίονες που οδηγούν στο επόμενο πιο κοντινό ζευγάρι γειτόνων.*

Τέλος, πρέπει να αναφέρουμε και μια άλλη μέθοδο που χρησιμοποιεί αποστάσεις. Αυτή, είναι η μέθοδος των *Fitch-Margoliash* (Fitch & Margoliash, 1967), η οποία βασίζεται στη στατιστική τεχνική της γραμμικής παλινδρόμησης, δηλαδή, της ευθείας ελαχίστων τετραγώνων. Η βασική ιδέα της μεθόδου, είναι να ελαχιστοποιήσει το άθροισμα των τετραγώνων των αποκλίσεων που έχουν οι παρατηρηθείσες αποστάσεις σε ένα δέντρο (d_{ij}), από τις θεωρητικές ($\hat{d}_{ij}$). Συνεπώς, η ποσότητα που ελαχιστοποιείται, θα δίνεται από τον τύπο:

$$Q = \sum_{i=1}^{L} \sum_{j=1}^{L} w_{ij} \left(\hat{d}_{ij} - d_{ij} \right)^2 \qquad (6.20)$$

Αυτός είναι ακριβώς το κριτήριο που ελαχιστοποιείται και στην κλασική ευθεία ελαχίστων τετραγώνων ($y=a+bx$). Τα βάρη w_{ij} παίζουν εδώ ακριβώς τον ίδιο ρόλο που παίζουν και στην γραμμική παλινδρόμηση. Η πρώτη εμφάνιση της μεθόδου, όπως προτάθηκε από τους (Cavalli-Sforza & Edwards, 1967) είχε βάρη ίσα με $w_{ij}=1$ (απλή γραμμική παλινδρόμηση), ενώ η πιο προχωρημένη μέθοδος των (Fitch & Margoliash, 1967), χρησιμοποίησε βάρη αντίστροφα του τετραγώνου της απόστασης ($wi_j = 1/\hat{d}_{ij}^2$, σταθμισμένη γραμμική παλινδρόμηση). Σε κάθε περίπτωση, με τη μέθοδο αυτή αναζητούμε το δέντρο, τα μήκη των βραχιόνων του οποίου θα έχουν τη μικρότερη τετραγωνική απόκλιση των αποστάσεων, σε σχέση με όλα τα πιθανά μήκη μονοπατιών. Η μέθοδος είναι απλή, κατανοητή και βασίζεται σε ένα στέρεο στατιστικό υπόβαθρο, αλλά έχει το βασικό μειονέκτημα ότι δεν έχει μηχανισμό για να εντοπίζει τα δέντρα. Αντίθετα, πρέπει η τετραγωνική απόκλιση (Q) να μετρηθεί για κάθε πιθανό δέντρο. Για αυτό το σκοπό, έχουν αναπτυχθεί μια σειρά αλγορίθμων οι οποίοι υπολογίζουν σε εύλογο χρονικό διάστημα τα πιθανά δέντρα τα οποία θα ελέγξει η μέθοδος. Παρ' όλα αυτά, σήμερα δεν χρησιμοποιείται πολύ, γιατί δεν εμφανίζει πολλά πλεονεκτήματα και έχει ξεπεραστεί από την πιο σύγχρονη και πιο ευέλικτη μέθοδο της μέγιστης πιθανοφάνειας.

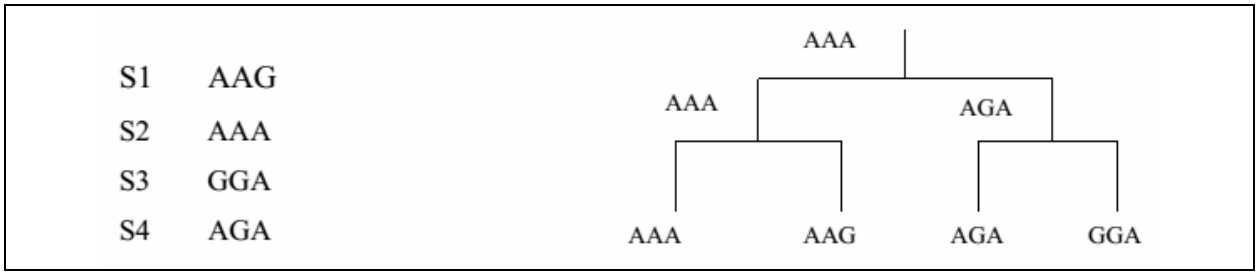
6.4. Μέθοδοι βασισμένες στους χαρακτήρες

Οι *μέθοδοι που βασίζονται στους χαρακτήρες* (*character-based methods*), σε αντίθεση με τις μεθόδους αποστάσεων, δεν μετασχηματίζουν τις αλληλουχίες, αλλά τις χρησιμοποιούν σε όλη τη διαδικασία της εκτίμησης, αντιμετωπίζοντας αυτές, όπως ακριβώς είναι: ακολουθίες διακριτών συμβόλων από ένα πεπερασμένο αλφάβητο (Durbin, et al., 1998). Διακρίνονται σε δύο μεγάλες ομάδες: στις μεθόδους φειδωλότητας, οι οποίες δεν κάνουν καμιά υπόθεση για τον τρόπο με τον οποίο συντελέστηκε η εξελικτική διαδικασία, και στις μεθόδους μέγιστης πιθανοφάνειας, οι οποίες χρησιμοποιούν (ή καλύτερα, απαιτούν) ένα ξεκάθαρο μαθηματικό μοντέλο για την εξέλιξη. Αυτές οι δύο μέθοδοι θα παρουσιαστούν στις επόμενες υποενότητες.

6.4.1 Μέθοδος Φειδωλότητας

Οι μέθοδοι που στηρίζονται στη *μέγιστη φειδωλότητα* (*maximum parsimony*) διαφέρουν ριζικά από τις προηγούμενες μεθόδους των αποστάσεων, στο ότι κάνουν διάκριση μεταξύ πληροφοριακών και μη-πληροφοριακών θέσεων στις αλληλουχίες, με τις πληροφοριακές θέσεις να είναι αυτές που παρουσιάζουν πολυμορφισμό (ύπαρξη πάνω από δυο ειδών νουκλεοτιδίων) τουλάχιστον δυο φορές. Η μέθοδος αυτή εφαρμόζεται στην εξελικτική βιολογία προτού να εμφανιστεί η μοριακή φυλογένεση (εφαρμοζόταν για παράδειγμα σε διάφορα φαινοτυπικά χαρακτηριστικά) και έχει σκοπό να εξηγήσει τις εξελικτικές διάφορες με το μικρότερο δυνατό αριθμό αλλαγών. Είναι δηλαδή, κατά μία έννοια, το φυλογενετικό ανάλογο της φιλοσοφικής μεθόδου του «ξυραφιού του Οκάμ» (Okham Razor), η οποία με απλά λόγια δηλώνει ότι η απλούστερη εξήγηση είναι και η προτιμότερη. Συνήθως η έκφραση αποδίδεται στα Λατινικά ως «*Pluralitas non est ponenda sine necessitate*», η οποία σε ελεύθερη απόδοση σημαίνει «*Όταν δύο θεωρίες παρέχουν εξίσου ακριβείς προβλέψεις, πάντα επιλέγουμε την απλούστερη*».

Ένα σημαντικό χαρακτηριστικό της μεθόδου, είναι ότι απλά αποδίδει κόστος σε μια δεδομένη τοπολογία ενός δέντρου, οπότε πρέπει να έχουμε στο μυαλό μας ότι απαιτείται ειδικός αλγόριθμος για να εντοπίσει το δέντρο του οποίου το κόστος θα υπολογίσουμε. Η πλέον χρησιμοποιούμενη μέθοδος είναι αυτή που χρησιμοποιεί τον αλγόριθμο του Fitch (Fitch, 1971), στην οποία κάθε διαφορά σε μία θέση «σκοράρει», δηλαδή αποδίδει κόστος ίσο με +1 σε όλες τις αλλαγές, αλλά έχουν προταθεί και παραλλαγές οι οποίες σταθμίζουν με διαφορετικό τρόπο τις διαφορές (weighted parsimony), οπότε ο σκοπός της μεθόδου είναι να ελαχιστοποιηθεί το κόστος αυτό. Στην Εικόνα 6.9, απεικονίζονται 4 αλληλουχίες οι οποίες είναι ήδη στοιχισμένες, και θέλουμε να βρούμε ένα φυλογενετικό δέντρο με τη χρήση της μεθόδου της φειδωλότητας. Το δέντρο που δίνεται είναι αυτό το οποίο εξηγεί τις νουκλεοτιδικές αλλαγές με τον μικρότερο αριθμό αντικαταστάσεων (3 συνολικά) από όλα τα άλλα δέντρα με ρίζα (συνολικά υπάρχουν 15 τέτοια δέντρα).



Εικόνα 6.9: *Ένα παράδειγμα δέντρου που εκτιμήηκε με τη μέθοδο της φειδωλότητας. Το δέντρο που δίνεται είναι αυτό το οποίο εξηγεί τις διαφορές των αλληλουχιών με τον μικρότερο αριθμό αντικαταστάσεων (3 συνολικά) από όλα τα άλλα πιθανά δέντρα με ρίζα (συνολικά υπάρχουν 15 τέτοια δέντρα).*

Γενικά, παρόλο που η μέθοδος είναι ιδιαίτερα γρήγορη, η αναζήτηση ανάμεσα σε όλα τα πιθανά δέντρα, γίνεται απαγορευτική όταν οι υπό σύγκριση αλληλουχίες είναι υπερβολικά πολλές. Έχουν προταθεί για αυτό το σκοπό διάφοροι αλγόριθμοι, εκ των οποίων ο λεγόμενος *branch and bound*, είναι αυτός που δίνει εγγυήσεις ότι θα βρει το καλύτερο δέντρο χωρίς να ανατρέξει σε όλες τις πιθανές τοπολογίες. Η βασική του ιδέα είναι να ξεκινάει από τυχαία δέντρα και να προσθέτει βραχίονες στην τύχη. Εκμεταλλεύεται έτσι, το

γεγονός (το οποίο είναι χαρακτηριστικό της μεθόδου της φειδωλότητας), ότι οι αλλαγές στο δέντρο μπορούν να συμβούν μόνο όταν προστεθεί ένας βραχίονας. Αν τώρα, ένας δεδομένος βραχίονας αυξήσει το συνολικό ελάχιστο αριθμό αντικαταστάσεων που έχει παρατηρηθεί μέχρι εκείνη τη στιγμή, τότε η αναζήτηση σε αυτή την κατεύθυνση εγκαταλείπεται και ο βραχίονας διαγράφεται όπως επίσης και όλες οι πιθανές αλλαγές από εκείνο το σημείο και κάτω.

Τα βασικά πλεονεκτήματα της μεθόδου είναι αφενός μεν η ταχύτητά της, που την καθιστά ικανή για αναλύσεις πολλών αλληλουχιών, αφετέρου δε η απλότητα της, καθώς δεν προϋποθέτει κανένα μοντέλο για την εξέλιξη των αλληλουχιών. Πρέπει να τονιστεί βέβαια, ότι η μέθοδος της φειδωλότητας είναι αντικείμενο πολλών αντιπαραθέσεων στην εξελικτική βιολογία, καθώς πολλοί ερευνητές δε δέχονται ότι διαθέτει στατιστική τεκμηρίωση ενώ μερικοί αμφισβητούν ακόμα και τη σχέση της με την παραπάνω αναφερθείσα φιλοσοφική μέθοδο της φειδωλότητας (Yang, 1996). Αξίζει να σημειωθεί τέλος ότι η φειδωλότητα προτάθηκε αρχικά (Edwards & Cavalli-Sforza, 1963) ως μέθοδος υπολογιστικής προσέγγισης στη μέγιστη πιθανοφάνεια την οποία θα αναλύσουμε στην επόμενη ενότητα.

6.4.2 Η Μέθοδος της Μέγιστης Πιθανοφάνειας

Η πιο προφανής από άποψη στατιστικής, μέθοδος εκτίμησης που θα μπορούσε να χρησιμοποιηθεί είναι αυτή της *Μέγιστης Πιθανοφάνειας* (*Maximum Likelihood*). Σε γενικές γραμμές η μέθοδος αυτή αντιμετωπίζει τις ακολουθίες ως ένα σετ *L* μεταβλητών με *n* παρατηρήσεις η κάθε μια. Έτσι στις ακολουθίες:

$$\mathbf{x_1} = x_{11}x_{12}...x_{1n}$$

$$\mathbf{x_2} = x_{21}x_{22}...x_{2n}$$

.....

$$\mathbf{x_L} = x_{L1}x_{L2}...x_{Ln}$$

τα αντίστοιχα διανύσματα των παρατηρήσεων τα οποία αντιστοιχούν στις θέσεις της πολλαπλής στοίχισης, θα είναι:

$$X_1 = \left(x_{11}, x_{21}, ..., x_{L1}\right), X_2 = \left(x_{12}, x_{22}, ..., x_{L2}\right), ..., X_L = \left(x_{1n}, x_{2n}, ..., x_{Ln}\right)$$

Πρέπει εδώ να τονιστούν κάποιες θεμελιώδεις διαφορές ανάμεσα στις γνωστές μεθόδους Μέγιστης Πιθανοφάνειας (Maximum Likelihood) που χρησιμοποιούνται για την εκτίμηση π.χ. παραμέτρων σε ένα Γενικευμένο Γραμμικό Μοντέλο και στην εκδοχή της μεθόδου που χρησιμοποιείται για την εκτίμηση των φυλογενετικών σχέσεων.

- Για να προχωρήσουμε στην ανάλυση, είναι αναγκαίο να έχει γίνει πρώτα μια πολλαπλή στοίχιση των αλληλουχιών (και άρα το αποτέλεσμα μας θα είναι δεσμευμένο στη στοίχιση αυτή). Παρ' όλα αυτά έχουν προταθεί και κάποιες μέθοδοι ταυτόχρονης στοίχισης και φυλογενετικής ανάλυσης.
- Για να εφαρμοστούν οι μέθοδοι αυτές πρέπει να ορίσουμε εξ' αρχής ένα πιθανοθεωρητικό μοντέλο το οποίο να περιγράφει την εξέλιξη των αλληλουχιών (π.χ μοντέλο JC69 ή κάποιο άλλο από αυτά που μελετήσαμε στην αντίστοιχη παράγραφο).
- Η πιθανοφάνεια κάθε φορά υπολογίζεται ως συνάρτηση της τοπολογίας του δέντρου και του μήκους των βραχιόνων του.

Αποτέλεσμα των παραπάνω είναι το γεγονός ότι η συνολική πιθανοφάνεια δεν μπορεί να υπολογιστεί αναλυτικά με κάποιον απλό τρόπο, αλλά απαιτεί υπολογισμούς που ανάγονται στο άθροισμα όλων των πιθανοφανειών για όλα τα πιθανά δέντρα. Στην περίπτωση που έχουμε δύο ακολουθίες:

$$\mathbf{x_1} = x_{11}, x_{12}, ..., x_{1n}$$

$$\mathbf{x_2} = x_{21}, x_{22}, ..., x_{2n}$$

η πιθανότητα το *i* νουκλεοτίδιο στις δυο ακολουθίες να έχει προκύψει από κάποιο *α* αρχικό είναι:

$$P\left(x_{1i}, x_{2i}, a \mid T, t_1, t_2\right) = q_\alpha P\left(x_{i1} \mid a, t_1\right) P\left(x_{i2} \mid a, t_2\right) \tag{6.21}$$

όπου *α* το άγνωστο αρχικό νουκλεοτίδιο, *T* το υποτιθέμενο δέντρο, και t_1,t_2 τα μήκη των βραχιόνων του δέντρου (χρόνος κατά τον οποίο έχουν αποκλίνει εξελικτικά). Επειδή δεν γνωρίζουμε το *α* πρέπει να αθροίσουμε όλες τις εναλλακτικές, άρα:

$$P\left(x_{1i},x_{2i}\mid T,t_1,t_2\right)=\sum_a q_\alpha P\left(x_{i1}\mid a,t_1\right)P\left(x_{i2}\mid a,t_2\right) \tag{6.22}$$

και κατόπιν μπορούμε να υπολογίσουμε τη συνολική πιθανότητα για τις n θέσεις των δυο ακολουθιών ως εξής:

$$P\left(\mathbf{x}_1,\mathbf{x}_2\mid T,t_1,t_2\right)=\prod_{i=1}^{n}P\left(x_{1i},x_{2i}\mid T,t_1,t_2\right) \tag{6.23}$$

Αυτή είναι η πιθανοφάνεια των δυο ακολουθιών (likelihood). Για αριθμητική ευκολία εργαζόμαστε συνήθως με το λογάριθμό της (log-likelihood) ο οποίος είναι:

$$\log P\left(\mathbf{x}_1,\mathbf{x}_2\mid T,t_1,t_2\right)=\sum_{i=1}^{n}\log P\left(x_{1i},x_{2i}\mid T,t_1,t_2\right) \tag{6.24}$$

Στο δεξί μέρος της σχέσης (6.24), θα πρέπει να χρησιμοποιηθούν οι αντίστοιχες εκφράσεις που απορρέουν από το εκάστοτε μοντέλο της εξέλιξης, όπως για παράδειγμα το JC69 ή το GTR, για να καταλήξουμε τελικά σε μια αναλυτική έκφραση για τη συνάρτηση πιθανοφάνειας. Στη γενικότερη περίπτωση που χρησιμοποιούμε L ακολουθίες, θα έχουμε:

$$P\left(x_{1i},x_{2i},\ldots,x_{Li}\mid T,t^{\bullet}\right)=\sum_{a^{L+1},\ldots,a^{2L-1}}q_{a^{2L-1}}\prod_{k=L+1}^{2L-2}P\left(a^k\mid a^{a(k)},t_k\right)\prod_{k=1}^{L}P\left(x_{ki}\mid a^{a(k)},t_k\right) \tag{6.25}$$

Οι επιπλέον συμβολισμοί που εισάγουμε εδώ είναι το $a(k)$ που συμβολίζει το αρχικό νουκλεοτίδιο για κάθε ένα από τα παρακλάδια του δέντρου. Από την συνδυαστική βρίσκουμε ότι για L ακολουθίες έχουμε $2L$-1 παρακλάδια και $2L$-2 σημεία διασταύρωσης των κλαδιών, όταν το δέντρο έχει ρίζα. Από αυτά τα πρώτα L αντιστοιχούν στα παρακλάδια (βραχίονες) που οδηγούν σε μια ακολουθία ενώ τα υπόλοιπα, από L+1εως $2L$-2, αντιστοιχούν στους κλάδους που ομαδοποιούν τις ακολουθίες. Έτσι δικαιολογούνται όλοι οι δυνατοί συνδυασμοί και τα γινόμενα στην παραπάνω εξίσωση, και το τελικό άθροισμα είναι για τα κλαδιά του δέντρου από το L+1 έως το $2L$-1 (είναι ένα παραπάνω γιατί πρέπει να μετρήσουμε και το α στη ρίζα του δέντρου). Τελικά η συνολική πιθανοφάνεια για τις r ακολουθίες θα προκύψει αφού αθροίσουμε τη συνεισφορά όλων των n θέσεων:

$$P\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_L\mid T,t_\circ\right)=\prod_{i=1}^{n}P\left(x_{1i},x_{2i},\ldots,x_{Li}\mid T,t_\circ\right) \tag{6.26}$$

και δουλεύοντας ως συνήθως με το λογάριθμο της (log-likelihood), θα έχουμε:

$$\log P\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_L\mid T,t_\circ\right)=\sum_{i=1}^{n}\log P\left(x_{1i},x_{2i},\ldots,x_{Li}\mid T,t_\circ\right) \tag{6.27}$$

Το παραπάνω μοντέλο, θεωρεί όλες τις θέσεις ανεξάρτητες και υποθέτει ότι ο ρυθμός της εξέλιξης είναι σταθερός για όλα τα παρακλάδια του δέντρου. Όπως είδαμε, έχουν προταθεί και άλλα μοντέλα τα οποία επιτρέπουν διαφορετικούς ρυθμούς αντικατάστασης (Yang, 1993). Η ύπαρξη σταθερών ρυθμών αντικατάστασης είναι γνωστή ως η υπόθεση του «*μοριακού ρολογιού*» (*molecular clock*) και είναι απαραίτητη προϋπόθεση και για την εφαρμογή της μεθόδου UPGMA, την οποία είδαμε σε προηγούμενη παράγραφο (για την ακρίβεια, είναι απαραίτητη προϋπόθεση για να δουλέψει σωστά η μέθοδος). Μια άλλη παρατήρηση που αφορά τη μέγιστη πιθανοφάνεια, η οποία δεν είναι αμέσως εμφανής, είναι ότι η μέθοδος δεν είναι ικανή να δώσει τη θέση της ρίζας του δέντρου, καθώς τα περισσότερα από τα στοχαστικά μοντέλα της εξέλιξης, έχουν την ιδιότητα να είναι χρονικώς αντιστρεπτά (time reversibility). Κατά συνέπεια, επειδή ο πίνακας των αντικαταστάσεων δεν μπορεί να διακρίνει μια αλλαγή A σε T, από μια αλλαγή T σε A, η πιθανοφάνεια του δέντρου είναι ίδια και για τις δύο εναλλακτικές υποθέσεις, οπότε τελικά, η πιθανοφάνεια δεν θα εξαρτάται από τη θέση της ρίζας.

Όπως ήδη είπαμε, η συνολική πιθανοφάνεια δεν μπορεί να υπολογιστεί αναλυτικά, αλλά πρέπει να αθροιστούν οι συνεισφορές όλων των πιθανών δέντρων. Ακόμα και όταν αναφερόμαστε σε ένα δεδομένο δέντρο (ανάμεσα στα πολλά πιθανά), η αναλυτική έκφραση για τη σχέση (6.27) είναι ιδιαίτερα πολύπλοκη, και απαιτεί για την επίλυσή της, κάποιου είδους επαναληπτική διαδικασία, κλασική σε προβλήματα μέγιστης πιθανοφάνειας, όπως για παράδειγμα κάποια παραλλαγή της μεθόδου Gradient Descent, Newton-Raphson (Ypma, 1995), ή του αλγορίθμου EM (Dempster, Laird, & Rubin, 1977). Κατά συνέπεια, η μέθοδος έχει το ίδιο πρόβλημα όπως και η μέγιστη φειδωλότητα: χρειάζεται κάποιος τρόπος για να υπολογιστούν γρήγορα όλα τα πιθανά δέντρα, ή τουλάχιστον, τα πιο πιθανά. Ο υπολογισμός της πιθανοφάνειας γίνεται με κάποιον

αλγόριθμο ο οποίος αθροίζει τις συνεισφορές για όλα τα πιθανά δέντρα, και ο πιο γνωστός αλγόριθμος που έχει προταθεί για αυτό το σκοπό, είναι αυτός του Felsenstein (Felsenstein, 1981). Αν και έχουν προταθεί πολλές παραλλαγές, σε γενικές γραμμές η διαδικασία που ακολουθείται για να υπολογιστεί η πιθανοφάνεια είναι η εξής: Ο αλγόριθμος εντοπίζει ένα πιθανό δέντρο και οι παράμετροι για αυτό το δέντρο αλλάζουν λίγο-λίγο με κάποια από τις επαναληπτικές διαδικασίες που αναφέραμε παραπάνω έως ότου βρεθεί το βέλτιστο δέντρο. Η διαδικασία αυτή επαναλαμβάνεται για όλα τα πιθανά δέντρα και αυτό που δίνει τη μέγιστη πιθανοφάνεια, από όλα τα δέντρα, επιλέγεται τελικά ως το δέντρο μέγιστης πιθανοφάνειας.

Η μέθοδος της μέγιστης πιθανοφάνειας έχει μερικά ξεκάθαρα πλεονεκτήματα σε σχέση με τις άλλες μεθόδους (Yang & Rannala, 2012). Καταρχάς, όλες οι προϋποθέσεις των μοντέλων δηλώνονται ξεκάθαρα και μπορούν να αξιολογηθούν εκ των υστέρων. Διαθέτει επίσης μια πληθώρα πιθανοθεωρητικών μοντέλων για την εξέλιξη των αλληλουχιών, τα οποία μπορούν να υλοποιηθούν, να εφαρμοστούν και να ελεγχθούν. Το στέρεο μαθηματικό της υπόβαθρο, το οποίο χρησιμοποιείται και σε πολλές άλλες εφαρμογές στη βιοπληροφορική, επιτρέπει τη χρήση εργαλείων όπως ο έλεγχος του πηλίκου πιθανοφάνειας (likelihood ratio test), με σκοπό τον έλεγχο καλής προσαρμογής και τη σύγκριση ανταγωνιστικών μοντέλων. Συνέπεια όλων αυτών, είναι να αποτελεί πανίσχυρο εργαλείο όχι μόνο στην απλή ανακατασκευή φυλογενετικών δέντρων, αλλά και στη διερεύνηση των ίδιων των μηχανισμών της εξελικτικής διαδικασίας, όπως για παράδειγμα στον έλεγχο της υπόθεσης του μοριακού ρολογιού, ή του τρόπου με τον οποίο επηρεάζει η δαρβινική επιλογή την εξέλιξη των πρωτεϊνών. Ένα βασικό μειονέκτημα της μεθόδου, είναι ότι είναι υπολογιστικά εντατική, με το πιο δύσκολο κομμάτι να αποτελεί η αναζήτηση των δέντρων κάτω από το κριτήριο της μέγιστης πιθανοφάνειας. Παρ' όλα αυτά, η εξέλιξη των υπολογιστών, η αύξηση της υπολογιστικής ισχύος αλλά και μια σειρά βελτιώσεις σε αλγοριθμικό επίπεδο, έχουν κάνει τη μέθοδο να είναι η πλέον αποδεκτή και η περισσότερο χρησιμοποιούμενη τα τελευταία χρόνια, καθώς συνεχώς παρουσιάζονται παράλληλες υλοποιήσεις αλγορίθμων, υλοποιήσεις σε GPU αλλά και υλοποιήσεις σε FPGA.

Μια διαφορετική οπτική στη μέθοδο μέγιστης πιθανοφάνειας, χρησιμοποιούν οι λεγόμενες *Μπεϋζιανές μέθοδοι* (*Bayesian methods*) (Huelsenbeck, Ronquist, Nielsen, & Bollback, 2001). Με την Μπευζιανή στατιστική ανάλυση, ενσωματώνεται στο μοντέλο με «φυσικό» τρόπο, η αβεβαιότητα στην εκτίμηση των παραμέτρων και απαντάται με άμεσο τρόπο το βιολογικό ερώτημα. Η κλασική μέθοδος της μέγιστης πιθανοφάνειας, χρησιμοποιεί την πιθανότητα των παρατηρήσεων, δεδομένου του δέντρου και των παραμέτρων, $P(\mathbf{x}|T,\theta)$, δηλαδή τη σχέση (6.26). Σε αντίθεση (Bland & Altman, 1998), οι μπεϋζιανές μέθοδοι, βασιζόμενες στο θεώρημα του Bayes, αντιστρέφουν το πρόβλημα και αντιμετωπίζουν τις παραμέτρους σαν τυχαίες μεταβλητές χρησιμοποιώντας την εκ των υστέρων κατανομή (posterior distribution):

$$P(T,\theta \mid \mathbf{x}) = \frac{P(T,\theta)P(\mathbf{x} \mid T,\theta)}{P(\mathbf{x})} \tag{6.28}$$

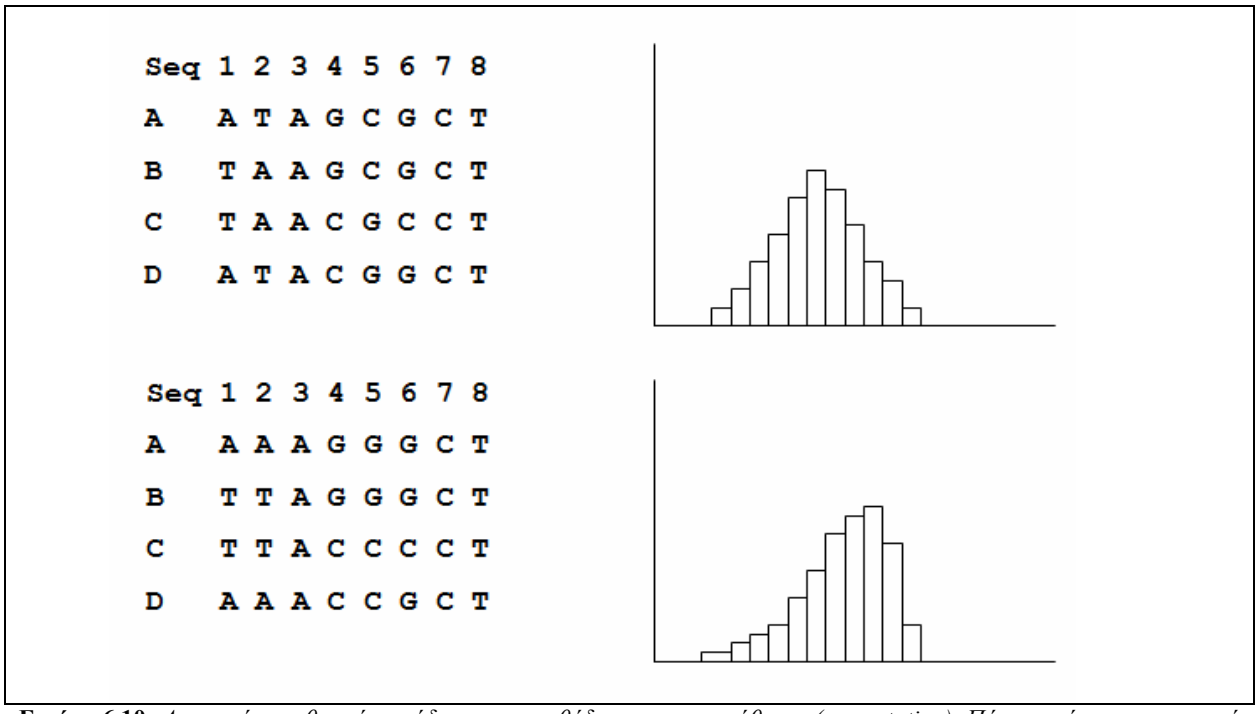
Στη σχέση (6.28), το $P(T,\theta \mid \mathbf{x})$ είναι η εκ των υστέρων κατανομή, $P(\mathbf{x}|T,\theta)$ η πιθανοφάνεια, ενώ $P(T,\theta)$ είναι η εκ των προτέρων κατανομή του δέντρου και των παραμέτρων. Ο παρονομαστής, $P(\mathbf{x})$, είναι μια σταθερά κανονικοποίησης η οποία χρησιμοποιείται έτσι ώστε η εκ των υστέρων κατανομή να είναι όντως πιθανότητα (δηλαδή, να αθροίζει στο ένα). Σε γενικές γραμμές, ακόμα και για τα απλά προβλήματα κλασικής στατιστικής, η εκ των υστέρων κατανομή δεν μπορεί να υπολογιστεί αναλυτικά, καθώς περιέχει δύσκολα ολοκληρώματα με πολλές διαστάσεις. Παρ' όλα αυτά, τέτοιες τεχνικές, χρησιμοποιούνται ευρέως τα τελευταία χρόνια στη βιοστατιστική και τη βιοπληροφορική και πραγματοποιούν τις εκτιμήσεις των παραμέτρων κάνοντας δειγματοληψία από την εκ των υστέρων κατανομή που προκύπτει από μια προσομοίωση με τη χρήση του MCMC (Markov Chain Monte Carlo) (Gilks, Richardson, & Spiegelhalter, 1996). Για το λόγο αυτό, είναι και πιο απαιτητικές από πλευράς υπολογιστικής ισχύος.

Το βασικό πλεονέκτημα της μεθόδου αυτής είναι ότι απαντάει με άμεσο και φυσικό τρόπο, μέσω της εκ των υστέρων κατανομής, στο βιολογικό ερώτημα («ποια είναι η πιθανότητα το δέντρο T να είναι σωστό, δεδομένων των παρατηρήσεων μου και του μοντέλου;»). Σε αντίθεση, η χρήση των τεχνικών της πιθανοφάνειας και των ελέγχων υποθέσεων γενικά, έχει μια δυσκολία στην κατανόηση από τους μη ειδικούς (Bland & Altman, 1998). Ένα άλλο πλεονέκτημα της μεθόδου είναι ότι επιτρέπει την ενσωμάτωση της εκ των προτέρων πληροφορίας η οποία μπορεί να προέρχεται από εξειδικευμένη γνώση. Πρακτικά όμως, τέτοιες περιπτώσεις είναι σπάνιες και μη-πληροφοριακές εκ των προτέρων κατανομές χρησιμοποιούνται στις περισσότερες περιπτώσεις. Το βασικό μειονέκτημα της μεθόδου, είναι ότι είναι ιδιαίτερα απαιτητική υπολογιστικά (περισσότερο και από τη μέθοδο μέγιστης πιθανοφάνειας), ενώ η εκ των υστέρων κατανομή είναι ίσως περισσότερο ευαίσθητη σε περιπτώσεις λάθος ορισμού του μοντέλου. Παρ' όλα αυτά, είναι μια

υποσχόμενη μέθοδος, η οποία κερδίζει συνεχώς έδαφος τα τελευταία χρόνια σε πολλούς τομείς της βιοπληροφορικής και της βιοστατιστικής.

6.5. Αξιολόγηση των δέντρων

Τελευταία, και ίσως πιο δύσκολη διαδικασία σε μια φυλογενετική ανάλυση, είναι το να εκτιμήσουμε πόσο καλό είναι το δέντρο που κατασκευάσαμε και αν ικανοποιούνται οι αρχικές προϋποθέσεις της ανάλυσης. Συνήθως δύο κατηγορίες μεθόδων είναι αυτές που χρησιμοποιούνται, οι εμπειρικές, δηλαδή αυτές που βασίζονται σε κάποια προσομοίωση ή μέθοδο τυχαίας δειγματοληψίας, και οι μέθοδοι που βασίζονται στις μαθηματικές ιδιότητες του ίδιου του μοντέλου.

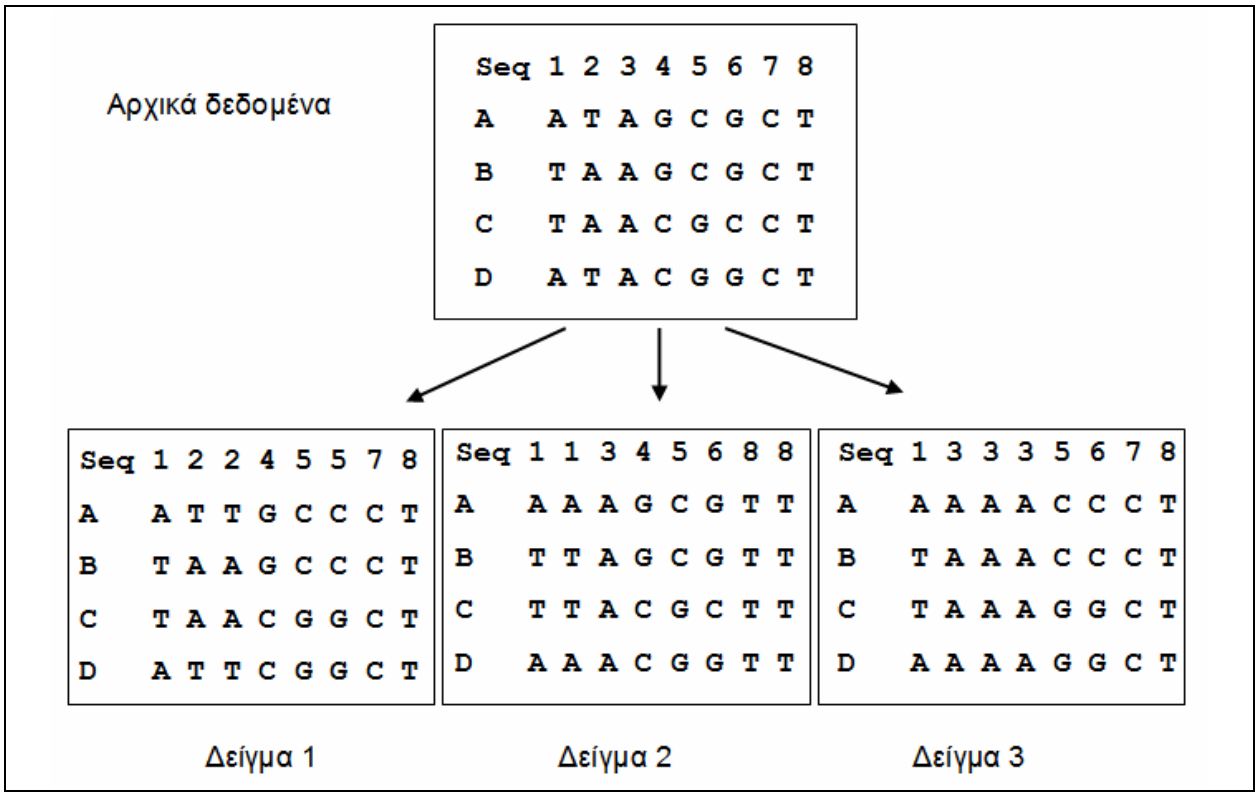


Εικόνα 6.10: *Αριστερά: υποθετικό παράδειγμα της μεθόδου της αντιμετάθεσης (permutation). Πάνω, φαίνεται μια αρχική πολλαπλή στοίχιση στην οποία υπάρχει συσχέτιση μεταξύ των στηλών (στήλες 1 με 2 και 5 με 6), ενώ από κάτω μια τυχαία αντιμετάθεση στην οποία η συσχέτιση έχει εν πολλοίς αφαιρεθεί. Η ανάλυση των μηκών των βραχιόνων τέτοιων δέντρων που προκύπτουν από διαδοχικές αντιμεταθέσεις, θα πρέπει να δείξει μια κατανομή στην οποία, το παρατηρηθέν δέντρο θα βρίσκεται στο αριστερό άκρο (στο ελάχιστο). Δεξιά: η κατανομή των μηκών των βραχιόνων από πολλά τελείως τυχαιοποημένα (randomised) δέντρα. Πάνω, φαίνεται η συμμετρική κατανομή των μηκών από τυχαιοποιημένα δέντρα στα οποία δεν υπάρχει φυλογενετικό σήμα. Κάτω, φαίνεται η λοξή κατανομή που προκύπτει από ένα σύνολο δεδομένων στο οποίο υπάρχει τέτοιο σήμα.*

Στην τελευταία κατηγορία, ανήκει ο γνωστός έλεγχος του πηλίκου πιθανοφάνειας, ο οποίος προφανώς εφαρμόζεται μόνο στην περίπτωση ανάλυσης μέγιστης πιθανοφάνειας. Σε αυτή την περίπτωση, δύο ανταγωνιστικά «μοντέλα» (συνήθως το ένα να αποτελεί ειδική περίπτωση του άλλου) ελέγχονται μέσω του πηλίκου της πιθανοφάνειάς τους και η διαφορά συγκρίνεται με μια θεωρητική κατανομή χ^2. Γενικά η μέθοδος δεν έχει καλές ιδιότητες όταν πρόκειται να συγκριθούν ανταγωνιστικά φυλογενετικά δέντρα για να βρεθεί η σωστή τοπολογία, αλλά δουλεύει σωστά όταν πρόκειται να συγκρίνει εξελικτικά μοντέλα αντικατάστασης, ειδικά σε συνδυασμό με την παραμετρική bootsrap που θα δούμε παρακάτω (Goldman, 1993; Posada & Crandall, 1998). Ένα πλεονέκτημα των μπεϋζιανών μεθόδων, είναι όπως είπαμε, το ότι η εκ των υστέρων κατανομή απαντά άμεσα και απλά στο πρόβλημα της πιθανότητας του δέντρου (αν και η μέθοδος έχει κατηγορηθεί ότι παράγει κάπως υπερβολικά αισιόδοξα ποσοστά).

Μια μέθοδος που χρησιμοποιείται συνήθως με την ανάλυση μέγιστης φειδωλότητας, είναι να ελέγχεται η κατανομή που δίνουν τα μήκη των τυχαιοποιημένων δέντρων και να συγκρίνεται, κυρίως με βάση τη λοξότητα της κατανομής, με αυτή που έχουν δέντρα τελείως τυχαία (χωρίς καμία εξελικτική πληροφορία).

Στα λεγόμενα permutation tests, στήλες από την πολλαπλή στοίχιση αντιμετατίθενται, δηλαδή ανακατεύεται η σειρά εμφάνισης των χαρακτήρων τους, έτσι ώστε να πάρουμε ένα νέο σύνολο δεδομενων, δηλαδή μια στοίχιση, στην οποία οι παρατηρηθήσες συχνότητες σε κάθε στήλη να είναι ίδιες αλλά να έχει αφαιρεθεί η επίδραση της συσχέτισης μεταξύ θέσεων της ίδιας ακολουθίας. Η μέθοδος αυτή εφαρμόζεται συνήθως σε αναλύσεις φειδωλότητας για να δείξει κατά πόσο είναι πιθανό ένα δεδομένο δέντρο να έχει προέλθει κατά τύχη αλλά δεν μπορεί να εντοπίσει αν το δέντρο είναι σωστό ή όχι.



Εικόνα 6.11: *Υποθετικό Παράδειγμα της μεθόδου bootstrap. Πάνω φαίνεται μια αρχική πολλαπλή στοίχιση, ενώ κάτω μια σειρά από τυχαίες δειγματοληψίες με επανάθεση ανάμεσα στις στήλες. Κάθε δείγμα θα αναλυθεί με την ίδια μέθοδο για να ελεγχθεί η σταθερότητα της. Προσέξτε ότι λόγω της δειγματοληψίας κάποιες στήλες δεν θα επιλεγούν σε κάποιο δείγμα, ενώ άλλες μπορεί να επιλεγούν περισσότερες από μία φορά.*

Ίσως η πιο γενική και ισχυρή μέθοδος, είναι αυτή του bootstrap. Είναι μια γνωστή μέθοδος στη στατιστική, και χρησιμοποιείται για ελέγχους σημαντικότητας σε δύσκολες περιπτώσεις (Efron & Tibshirani, 1993). Εν συντομία, η μέθοδος λειτουργεί σε δύο βήματα: 1) δημιουργεί πολλά «τεχνητά» σύνολα δεδομένων κάνοντας *δειγματοληψία με επανάθεση* από τις παρατηρήσεις (δηλαδή τις στήλες της στοίχισης) του αρχικού συνόλου δεδομένων, και 2) επαναλαμβάνει την ανάλυση για κάθε ένα από τα νέα αυτά σύνολα δεδομένων. Αν και φαίνεται εκ πρώτης όψεως παράδοξο, η μέθοδος έχει άριστες μαθηματικές ιδιότητες και οδηγεί σε καλό υπολογισμό των p-values και διαστημάτων εμπιστοσύνης. Στην περίπτωση των φυλογενετικών δέντρων (Hillis & Bull, 1993; Solitis & Solitis, 2003), η ερμηνεία είναι λίγο περίεργη, καθώς αν και αρχικά η μέθοδος προτάθηκε για να μετρήσει την επαναληψιμότητα μιας ανάλυσης, στην πράξη χρησιμοποιήθηκε από πολλούς για να δώσει μια εκτίμηση της πιθανότητας ότι το δέντρο είναι σωστό. Σε ανάλυση μέγιστης πιθανοφάνειας κάτω από προϋποθέσεις (να ισχύουν συγκεκριμένα μοντέλα της εξέλιξης και τα δεδομένα να είναι πολλά), έχει δειχθεί ότι η μέθοδος δίνει όντως μια κατανομή που προσεγγίζει την εκ των υστέρων πιθανότητα του δέντρου (Durbin, et al., 1998). Στην πραγματικότητα όμως, και κάτω από συνθήκες στις οποίες το μοντέλο που χρησιμοποιείται δεν είναι το σωστό, η σωστή ερμηνεία είναι ότι αυτό που μετράει η bootstrap είναι η αξιοπιστία ενός συγκεκριμένου κλάδου του δέντρου (δηλαδή, το πόσο συχνά αυτός ο κλάδος ανακατασκευάζεται σωστά από να «τεχνητά» δεδομένα). Μια άλλη, αλλά όχι τόσο σωστή χρήση της μεθόδου είναι να κατασκευάζει ένα συναινετικό δέντρο (consensus tree) από τα αποτελέσματα των επαναλήψεων και να περιλαμβάνει σε αυτό κλάδους που εμφανίζονται στην πλειοψηφία των επαναλήψεων. Η γενική μέθοδος

που περιγράψαμε, ονομάζεται μη-παραμετρική bootstrap και είναι δυνατό να εφαρμοστεί με κάθε μέθοδο κατασκευής δέντρου.

Μια παραλλαγή της μεθόδου, η λεγόμενη παραμετρική bootstrap, παράγει τεχνητά δεδομένα ίδιου μεγέθους με το αρχικό σύνολο πραγματοποιώντας την προσομοίωση κάτω από τις προϋποθέσεις του ίδιου εξελικτικού μοντέλου με το οποίο έγινε και η ανάλυση (Goldman, 1993; Wollenberg & Atchley, 2000). Κάθε σύνολο δεδομένων αναλύεται με τον ίδιο τρόπο, και τα αποτελέσματα ερμηνεύονται περίπου με τον ίδιο τρόπο όπως και για την μη-παραμετρική bootstrap. Όπως είναι προφανές, η μέθοδος αυτή έχει ιδιαίτερη αξία αν χρησιμοποιηθεί με κάποια μέθοδο κατασκευής δέντρων η οποία υποθέτει ένα συγκεκριμένο μοντέλο της εξελικτικής διαδικασίας γιατί με τον τρόπο αυτό μπορούν να ελεγχθούν καλύτερα τόσο η ικανότητα ανακατασκευής του σωστού μοντέλου, όσο και η σταθερότητα της μεθόδου έναντι σε λάθος καθορισμό του μοντέλου (model misspecification). Οι μέθοδοι αυτές, λόγω της επαναληπτικής τους φύσης, είναι υπολογιστικά εντατικές καθώς απαιτούνται εκατοντάδες επαναλήψεις, στις οποίες ο αλγόριθμος κατασκευής δέντρων θα πρέπει επίσης να εφαρμοστεί επαναλληπτικά.

Γενικά, το πρόβλημα της εκτίμησης της πιθανότητας του δέντρου και της σύγκρισης ανταγωνιστικών δέντρων, είναι σύνθετο και με μεγάλη βιβλιογραφία. Όπως είπαμε παραπάνω, ούτε οι έλεγχοι πηλίκου πιθανοφάνειας, ούτε οι τιμές bootstrap από μόνες τους, προσφέρουν καλή εκτίμηση της πιθανότητας του δέντρου να έχει κατασκευαστεί σωστά. Έχουν αναπτυχθεί Παρ' όλα αυτά, μια σειρά σύνθετες μέθοδοι οι οποίες χρησιμοποιούν τα αποτελέσματα της πιθανοφάνειας από τις διαδοχικές επαναλήψεις της bootstrap και κατασκευάζουν έναν έλεγχο για την ορθότητα του δέντρου, με αρκετά καλές στατιστικές ιδιότητες. Οι πιο γνωστές από αυτές τις τεχνικές, είναι ο έλεγχος των Kishino-Hasegawa (KH), ο έλεγχος των Shimodaira-Hasegawa (SH), ο σταθμισμένος (weighted) έλεγχος των Shimodaira-Hasegawa (WSH), και ο λεγόμενος Approximately Unbiased (AU) έλεγχος του Shimodaira, ο οποίος θεωρείται και ο καλύτερος. Το λογισμικό **CONSEL**, υλοποιεί τους παραπάνω ελέγχους, δεχόμενο σαν είσοδο τα αποτελέσματα από τις πιθανοφάνειες των ανταγωνιστικών δέντρων και τις αντίστοιχες τιμές από τις διαδοχικές επαναλήψεις bootsrap. Επίσης, έχει το πλεονέκτημα ότι μπορεί να χρησιμοποιηθεί με δεδομένα που προέρχονται από διαφορετικούς αλγόριθμους και διαφορετικά λογισμικά (Shimodaira & Hasegawa, 2001).

6.6. Η διαμάχη για την Εγκυρότητα των Μεθόδων-Πρακτικές Συμβουλές

Όπως ήδη αναφέραμε, η διαμάχη για το ποια μέθοδος (Μέγιστη Πιθανοφάνεια ή Μέγιστη Φειδωλότητα) είναι προτιμότερη για την εκτίμηση των φυλογενετικών σχέσεων, είναι παλιά και συνεχίζεται ακόμα, αν και μπορούμε να πούμε ότι τα τελευταία χρόνια, έχει κοπάσει κάπως. Πολλές μελέτες έχουν γίνει με χρήση προσομοιώσεων, οι οποίες έδειξαν αντικρουόμενα αποτελέσματα ως προς την προτίμηση στις δυο ανταγωνιστικές μεθόδους όσον αφορά τη ορθή ανακατασκευή των δέντρων, ανάλογα με το ποιες ήταν οι συνθήκες (το αρχικό μοντέλο) που παρήγαγαν τα δεδομένα.

Αξίζει να αναφερθεί, ότι ενώ αρχικά η μέθοδος της φειδωλότητας είχε προταθεί ως υπολογιστική απλοποίηση για την εύρεση της συνάρτησης πιθανοφάνειας, αμφισβητήθηκε έντονα στη συνέχεια, με κύριο επιχείρημα το ότι δεν έχει στατιστική ερμηνεία και δεν κάνει καθόλου υποθέσεις για τον τρόπο με τον οποίο έγινε η εξέλιξη. Παρ' όλα αυτά ο Felsenstein (Felsenstein, 1973, 1996) ισχυρίστηκε ότι όταν έχει συντελεστεί «λίγη» εξελικτική διαδικασία και ο ρυθμός της είναι περίπου σταθερός, τότε η παραπάνω προσέγγιση δίνει έγκυρα αποτελέσματα. Και αυτό όμως έχει αμφισβητηθεί από πολλούς που στηρίχθηκαν σε προσομοιώσεις χρησιμοποιώντας κάποιες ακραίες συνθήκες. Όταν οι ρυθμοί της εξέλιξης δεν είναι ίδιοι σε όλες τις εξελικτικές γραμμές (κάτι που συμβαίνει σχετικά συχνά), η μέθοδος της φειδωλότητας θα ανακατασκευάζει λανθασμένα δέντρα με μεγάλη πιθανότητα, η οποία θα μεγαλώνει καθώς το μέγεθος και ο αριθμός των ακολουθιών θα μεγαλώνει (Yang, 1996).

Συμπερασματικά, δεν μπορούμε να αποφανθούμε με 100% σιγουριά για το ποια μέθοδος είναι καλύτερη κάτω από όλες τις περιστάσεις, και έτσι χρειάζεται προσοχή όταν έχουμε να εκτιμήσουμε ένα φυλογενετικό δέντρο. Σε γενικές γραμμές, η μέγιστη πιθανοφάνεια, φαίνεται να έχει κερδίσει στη σχετική διαμάχη, κυρίως λόγω του στέρεου μαθηματικού υποβάθρου της, της δυνατότητας χρήσης πολλών εξελικτικών μοντέλων, αλλά και της ευκολίας την οποία προσδίδουν οι σύγχρονοι υπολογιστές και η αυξημένη υπολογιστική ισχύς. Επιπλέον δε, φαίνεται να αποδίδει καλύτερα την ανακατασκευή δέντρων κάτω από τα περισσότερα σενάρια προσομοιώσεων. Παρ' όλα αυτά, η μέθοδος NJ και η φειδωλότητα εξακολουθούν να είναι δημοφιλείς ειδικά για γρήγορες αναλύσεις μεγάλου όγκου δεδομένων. Γενικά επειδή η διαδικασία κατασκευής ενός δέντρου περιλαμβάνει 3 διακριτές λειτουργίες (Penny & Hendy, 2001; Steel &

Penny, 2000), δηλαδή: 1) το κριτήριο καταλληλότητας για το πόσο καλά «προσαρμόζονται» τα δεδομένα στο δέντρο, 2) τη στρατηγική αναζήτησης για να βρούμε το καλύτερο δέντρο, και τέλος 3) τον έλεγχο των προϋποθέσεων κάτω από τις οποίες έχει συντελεστεί η εξέλιξη, είναι δυνατόν να έχουμε συνδυασμό πολλών μεθόδων, πράγμα που εκμεταλλεύονται αρκετά από τα σύγχρονα λογισιμικά τα οποία παρουσιάζουμε στην επόμενη ενότητα. Για παράδειγμα, η μη παραμετρική bootstrap μπορεί να χρησιμοποιηθεί σαν μέθοδος αξιολόγησης με κάθε μέθοδο κατασκευής δέντρων, ενώ τα κλασικά μοντέλα της εξέλιξης (πχ JC69, K2P κλπ), μπορούν να χρησιμοποιηθούν τόσο με τη NJ (και την UPGMA), όσο και με τη μέγιστη πιθανοφάνεια (αλλά προσοχή, όχι με τη φειδωλότητα!). Ο Felsenstein έδειξε επιπλέον, ότι χρησιμοποιώντας οποιοδήποτε από τα γνωστά μοντέλα της εξελικτικής διαδικασίας, μπορούν να οριστούν «αποστάσεις μέγιστης πιθανοφάνειας» (maximum likelihood distance), οι οποίες έχουν την προσθετική ιδιότητα (Felsenstein, 1996). Αυτές οι αποστάσεις, μπορούν να χρησιμοποιηθούν με οποιαδήποτε μέθοδο αποστάσεων (NJ, UPGMA) για να δώσουν μια μέθοδο η οποία θα δίνει καλύτερα αποτελέσματα. Μια άλλη υβριδική μέθοδος είναι η NJML, η οποία αποτελεί συνδυασμό των Neighbour Joining και Maximum Likelihood. Στο πρώτο βήμα κατασκευάζει ένα δέντρο με NJ και η αναζήτηση των πιθανών δέντρων με τη μέθοδο μέγιστης πιθανοφάνειας γίνεται μόνο στα κλαδιά με μεγάλη τιμή bootstrap. Η NJML έδειξε ότι πετυχαίνει καλύτερα αποτελέσματα από την κλασική NJ αλλά σε χρόνο που είναι σημαντικά μικρότερος σε σχέση με τις ιδιαίτερα απαιτητικές μεθόδους πιθανοφάνειας (Ota & Li, 2000).

Συμπερασματικά, ο αναγνώστης θα πρέπει να έχει στο μυαλό του τις ακόλουθες συμβουλές πριν από κάθε ανάλυση (Brinkman & Leipe, 2001):

- Να γίνεται προσεκτικός έλεγχος των δεδομένων εισόδου. Αυτό είναι κάτι που οι περισσότεροι το ξεχνάνε, αλλά όπως είπαμε, όλες οι αναλύσεις στηρίζονται στην αρχική πολλαπλή στοίχιση και αν αυτή είναι λάθος, όλες οι μετέπειτα αναλύσεις είναι επισφαλείς (είναι αυτό που λένε: «*garbage in, garbage out*»)
- Ίσως το πιο σωστό είναι να χρησιμοποιούμε όσο το δυνατόν περισσότερες μεθόδους και να συγκρίνουμε τα αποτελέσματα, προσπαθώντας παράλληλα να ελέγξουμε τις προϋποθέσεις κάτω από τις οποίες ισχύει η κάθε μια (κάτι που δεν είναι και τόσο εύκολο). Γενικά, αν υπάρχει κάτι σημαντικό στα δεδομένα, τις περισσότερες φορές αν οι μέθοδοι εφαρμοστούν σωστά, θα δείξουν το ίδιο ή περίπου το ίδιο.
- Να γίνει έλεγχος της σειράς των αλληλουχιών. Όσο και αν φαίνεται παράξενο, κάποιες μέθοδοι (ειδικά οι πιο παλιές) παράγουν διαφορετικά αποτελέσματα ανάλογα με τη σειρά εισόδου των αλληλουχιών. Αν δεν είμαστε τελείως σίγουροι, καλό είναι να τοποθετούμε τις «περίεργες» αλληλουχίες προς το τέλος, ή αν είναι δυνατόν, να επαναλαμβάνουμε τις αναλύσεις αλλάζοντας με το χέρι τη σειρά των αλληλουχιών.
- Να γίνει προσεκτική επιλογή της εξωομάδας στις περιπτώσεις που η μέθοδος που θα χρησιμοποιήσουμε παράγει δέντρο χωρίς ρίζα. Εκτός από τις κλασικές παραμέτρους που πρέπει να προσέξουμε (να ανήκει σε οργανισμό με μεγάλη εξελικτική απόσταση από τους υπό μελέτη οργανισμούς), θα πρέπει να έχουμε υπόψη μας ότι μπορεί η επιλεγμένη ακολουθία να διαθέτει κάποια «ειδικά» χαρακτηριστικά που να την κάνουν να μοιάζει περισσότερο από το αναμενόμενο σε κάποιες από τις αλληλουχίες υπό σύγκριση. Τέτοια χαρακτηριστικά, είναι η σύσταση σε GC% και ο ρυθμοί εξελικτικής αλλαγής, οπότε πρέπει να είμαστε έτοιμοι για εναλλακτικές στρατηγικές (πχ να υπάρχει και δεύτερη εξωομάδα διαθέσιμη).

6.7. Λογισμικό

Οι περισσότερες από τις μεθόδους που αναφέραμε στις προηγούμενες ενότητες, υπάρχουν διαθέσιμες σε υλοποιήσεις λογισμικού, το οποίο διατίθεται ελεύθερα στον τελικό χρήστη. Τα πιο παλιά από τα προγράμματα αυτά, είναι το PAUP και το PHYLIP τα οποία εξελίσσονται συνεχώς, αλλά τα τελευταία χρόνια υπάρχουν νέες προσθήκες με πακέτα λογισμικού που προσφέρουν μεγάλη ευκολία στο χρήστη αλλά και μεγάλες ικανότητες ανάλυσης κάτω από διαφορετικά μοντέλα και προϋποθέσεις (πχ MEGA, RAxML κλπ). Στην ενότητα αυτή, παρουσιάζεται μια μικρή, αλλά ελπίζουμε κατατοπιστική περιγραφή των βασικότερων αλγοριθμικών υλοποιήσεων και πακέτων λογισμικού για φυλογενετική ανάλυση.

Το **PAUP** (Phylogenetic analysis using parsimony* and other methods), είναι ένα από τα πιο παλιά και γνωστά πακέτα φυλογενετικής ανάλυσης (Wilgenbusch & Swofford, 2003). Όπως φανερώνει και το

όνομά του, αρχικά ξεκίνησε υλοποιώντας μεθόδους φειδολώτητας αλλά σταδιακά εμπλουτίστηκε και πλέον προσφέρει και μεθόδους αποστάσεων αλλά και μέγιστης πιθανοφάνειας με πολλές επιλογές. Το μειονέκτημα του είναι ότι διατίθεται με εμπορική άδεια χρήσης (http://www.sinauer.com/detail.php?id=8060). Το **PHYLIP** (PHYLogeny Inference Package) είναι επίσης ένα από τα πιο παλιά και αξιόπιστα πακέτα το οποία αναπτύχθηκε αρχικά από τον Joe Felsenstein (Retief, 2000). Στις σύγχρονες εκδόσεις υλοποιεί πληθώρα μεθόδων τόσο για μεθόδους αποστάσεων όσο και για μέγιστη πιθανοφάνεια και φειδωλότητα, ενώ διανέμεται πλέον κάτω από άδεια ανοικτού κώδικα (http://evolution.gs.washington.edu/phylip.html). Το **MEGA** (Molecular evolutionary genetic analysis) είναι ίσως το πιο χρησιμοποιημένο, τα τελευταία χρόνια, πακέτο φυλογενετικής ανάλυσης (Kumar, Nei, Dudley, & Tamura, 2008). Ενσωματώνει μεθόδους αποστάσεων, μέγιστης πιθανοφάνειας και φειδωλότητας, και το δυνατό του σημείο είναι ότι είναι ιδιαίτερα εύχρηστο και κατάλληλο για τον απλό χρήστη καθώς τρέχει σε περιβάλλον Windows με παραθυρική διεπαφή (http://www.megasoftware.net).

Η αύξηση της υπολογιστικής ισχύος, έχει δώσει όπως είναι εμφανές, ένα μεγάλο προβάδισμα στις μεθόδους μέγιστης πιθανοφάνειας, καθώς αναλύσεις οι οποίες μέχρι πριν μία-δύο δεκαετίες δεν μπορούσαν να γίνουν παρά μόνο από υπερ-υπολογιστές, πλέον μπορούν να πραγματοποιηθούν από τον μέσο χρήστη στον προσωπικό του Η/Υ. Τέτοιες μέθοδοι, οι οποίες εστιάζονται αποκλειστικά στην χρήση πιθανοφάνειας, είναι το HYPHY, το PAML, το PhyML και το RAxML. Το **HYPHY** (Hypothesis testing using phylogenies), είναι ένα πρόγραμμα το οποίο χρησιμοποιεί μέγιστη πιθανοφάνεια για φυλογενετικές αναλύσεις. Η ιδιαιτερότητα του είναι ότι υλοποιεί μια υψηλού επιπέδου γλώσσα στην οποία ο χρήστης μπορεί να ορίσει το μοντέλο και να πραγματοποιήσει εύκολα ελέγχους πηλίκου πιθανοφάνειας για τη σύγκριση των ανταγωνιστικών μοντέλων (http://www.hyphy.org). Το **PAML** (Phylogenetic analysis by maximum likelihood), ήταν από τα πρώτα πακέτα που εστίαζαν αποκλειστικά στη μέγιστη πιθανοφάνεια. Αναπτύχθηκε από τον Ziheng Yang (Yang, 2007) και η μεγάλη του δύναμη βρίσκεται στους ελέγχους για θετική επιλογή, στην ανακατασκευή προγονικών αλληλουχιών, στη χρονολόγηση μέσω του μοριακού ρολογιού και στην υλοποίηση πολλών διαφορετικών πιθανοθεωρητικών μοντέλων, παρά στις αναζητήσεις και στις συγκρίσεις φυλογενετικών δέντρων (http://abacus.gene.ucl.ac.uk/software/paml.html). Το **PhyML** είναι ένα γρήγορο πρόγραμμα για αναζητήσεις δέντρων κάτω από τη μέγιστη πιθανοφάνεια, το οποίο μπορεί να χρησιμοποιήσει τόσο αλληλουχίες DNA όσο και πρωτεϊνών http://www.atgc-montpellier.fr/phyml/binaries.php, (Bazinet, Zwickl, & Cummings, 2014). Τέλος, το **RAxML**, το οποίο έχει αναπτυχθεί από τον Έλληνα επιστήμονα Αλέξανδρο Σταματάκη (Stamatakis, 2014), είναι ένα γρήγορο και αποτελεσματικό πρόγραμμα για αναλύσεις μέγιστης πιθανοφάνειας με το γενικό μοντέλο (GTR), κάνοντας χρήση τόσο αμινοξικών όσο και νουκλεοτιδικών αλληλουχιών. Το δυνατό του σημείο, είναι οι παράλληλες υλοποιήσεις των αλγορίθμων που επιτρέπουν την ανακατασκευή τεράστιων φυλογενετικών δέντρων http://scoh-its.org/exelixis/software.html.

Μεθόδους μέγιστης πιθανοφάνειας, χρησιμοποιούν και άλλα πακέτα, αλλά με ελαφρώς διαφορετικό τρόπο. Με την Μπευζιανή στατιστική ανάλυση, ενσωματώνεται στο μοντέλο με «φυσικό» τρόπο η αβεβαιότητα στην εκτίμηση των παραμέτρων. Τέτοιες τεχνικές, χρησιμοποιούνται ευρέως τα τελευταία χρόνια στη βιοστατιστική και τη βιοπληροφορική και πραγματοποιούν τις εκτιμήσεις των παραμέτρων κάνοντας δειγματοληψία από την εκ των υστέρων κατανομή που προκύπτει από μια προσομοίωση με τη χρήση του MCMC (Markov Chain Monte Carlo). Για το λόγο αυτό, είναι και πιο απαιτητικές από πλευρά υπολογιστικής ισχύος. Το **MrBayes** είναι ένα από τα πιο γνωστά και παλιά τέτοια εργαλεία, και υλοποιεί φυλογενετική ανάλυση με χρήση MCMC (Huelsenbeck & Ronquist, 2001). Περιέχει επιλογές για όλα τα γνωστά πιθανοθεωρητικά μοντέλα αντικατάστασης των νουκλεοτιδίων αλλά και μοντέλα για αμινοξέα και κωδικόνια (http://mrbayes.net). Το **BEAST** (Bayesian evolutionary analysis sampling tree), είναι ένα άλλο πρόγραμμα Μπεϋζιανής ανάλυσης με χρήση MCMC (Drummond, Suchard, Xie, & Rambaut, 2012). Παράγει δέντρα με ρίζα κάτω από τις προϋποθέσεις του μοριακού ρολογιού, αλλά υποστηρίζει και μια σειρά από μοντέλα που χαλαρώνουν αυτές τις προϋποθέσεις. Μπορεί να χρησιμοποιηθεί με αμινοξικές ή νουκλεοτιδικές αλληλουχίες, αλλά και με άλλου είδους δεδομένα (πχ μορφολογικά). Περιέχει επίσης ρουτίνες, όπως το **Tracer** και το **FigTree**, οι οποίες χρησιμεύουν στα διαγνωστικά και στην οπτικοποίηση των αποτελεσμάτων (http://beast.bio.ed.ac.uk).

Το **GARLI** (Genetic Algorithm for Rapid Likelihood Inference), ακολουθεί μια διαφορετική προσέγγιση στη φυλογενετική ανάλυση με χρήση πιθανοφάνειας (Bazinet, et al., 2014). Χρησιμοποιεί Γενετικούς Αλγόριθμους (μια τεχνική της τεχνητής νοημοσύνης) στην αναζήτηση του δέντρου μέγιστης πιθανοφάνειας. Περιέχει τόσο το γενικό μοντέλο GTR και τις ειδικές περιπτώσεις του, όσο και το μοντέλο της κατανομής Γάμμα, ενώ μπορεί να αναλύσει νουκλεοτιδικές και αμινοξικές αλληλουχίες αλλά και κωδικόνια. Ένα μεγάλο πλεονέκτημα είναι ότι διαθέτει και παράλληλες εκδόσεις των αλγορίθμων

http://code.google.com/p/garli. Το **TNT** (Tree analysis using new technology) (Goloboff, Farris, & Nixon, 2008) είναι ένα πολύ γρήγορο πρόγραμμα για φυλογενετική ανάλυση με χρήση της φειδωλότητας το οποίο είναι ιδιαίτερα ικανό για αναλύσεις μεγάλων δέντρων http://www.lillo.org.ar/phylogeny/tnt/. Τέλος, αξίζει να αναφερθούμε και στο **TreeView** (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html) το οποίο είναι ένα ελεύθερα διαθέσιμο πρόγραμμα οπτικοποίησης φυλογενετικών δέντρων. Το λογισμικό διαβάζει τους περισσότερους τύπους αρχείων που χρησιμοποιούν τα σχετικά προγράμματα (NEXUS, PHYLIP, Hennig86, NONA, MEGA, και ClustalW/X) και μπορεί να υποστηρίξει γραμματοσειρές TrueType and Postscript αλλά και γραφικά PICT (Macintosh) και Windows metafile (Windows) τα οποία επιτρέπουν εύκολη μεταφορά και επεξεργασία. Είναι διαθέσιμο για όλες τις γνωστές πλατφόρμες (Windows, Unix/Linux, και Macintosh), και διαθέτει και editor για επεξεργασία των δέντρων.

Βιβλιογραφία

Bazinet, A. L., Zwickl, D. J., & Cummings, M. P. (2014). A gateway for phylogenetic analysis powered by grid computing featuring GARLI 2.0. *Syst Biol, 63*(5), 812-818.

Bland, J. M., & Altman, D. G. (1998). Bayesians and frequentists. *BMJ, 317*(7166), 1151-1160.

Brinkman, F. S., & Leipe, D. D. (2001). Phylogenetic Analysis. In A. D. Baxevanis & B. F. Ouellette (Eds.), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* (pp. 323-358): John Wiley & Sons, Inc.

Cavalli-Sforza, L. L., & Edwards, A. W. (1967). Phylogenetic analysis. Models and estimation procedures. *Am J Hum Genet, 19*(3 Pt 1), 233-257.

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). A model of evolutionary change in Proteins. In M. Dayhoff (Ed.), *In Atlas of protein sequence and structure* (Vol. 5, Suppl. 3, pp. 345-352): National biomedical research foundation, Silver Spring, MD.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc B, 39*, 1-38.

Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution, 29*(8), 1969-1973.

Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. J. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.*

Edwards, A. W., & Cavalli-Sforza, L. L. (1963). The reconstruction of evolution. . *Annals of Human Genetics, 27*, 105.

Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *American journal of human genetics, 25*(5), 471.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution, 17*(6), 368-376.

Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol, 266*, 418-427.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology, 20*(4), 406-416.

Fitch, W. M., & Margoliash, E. (1967). Construction of phylogenetic trees. *science, 155*(3760), 279-284.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (Eds.). (1996). *Markov Chain Monte Carlo in Practice* Chapman & Hall/CRC.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution, 36*(2), 182-198.

Goloboff, P., A.,, Farris, J. S., & Nixon, K. C. (2008). TNT, a free program for phylogenetic analysis. *Cladistics, 24*(5), 774–786.

Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution, 22*(2), 160-174.

Hillis, D. M., & Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic biology, 42*(2), 182-192.

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics, 17*(8), 754-755.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science, 294*(5550), 2310-2314.

Jukes, T., & Cantor, C. (1969). Evolution of protein molecules Pp. 21–132 in HN Munro, ed. Mammalian protein metabolism: Academic Press, New York.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution, 16*(2), 111-120.

Kumar, S., Nei, M., Dudley, J., & Tamura, K. (2008). MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform, 9*(4), 299-306.

Lio, P., & Goldman, N. (1998). Models of molecular evolution and phylogeny. *Genome research, 8*(12), 1233-1244.

Ota, S., & Li, W.-H. (2000). NJML: a hybrid algorithm for the neighbor-joining and maximum-likelihood methods. *Molecular Biology and Evolution, 17*(9), 1401-1409.

Penny, D., & Hendy, M. (2001). Phylogenetics: parsimony and distance methods. In D. J. Balding, M. Bishop & C. Cannings (Eds.), *Handbook of Statistical Genetics* (pp. 445-484): John Wiley and Sons, Ltd.

Posada, D., & Crandall, K. A. (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics, 14*(9), 817-818.

Retief, J. D. (2000). Phylogenetic analysis using PHYLIP. *Methods Mol Biol, 132*, 243-258.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution, 4*(4), 406-425.

Shimodaira, H., & Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics, 17*(12), 1246-1247.

Sokal, R. R., & Michener, C. D. (1958). A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Science Bulletin, 38*, 1409-1438.

Solitis, P. S., & Solitis, D. E. (2003). Applying the Bootstrap in Phylogeny Reconstruction. *Stat Sci, 18*(2), 256-267.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313.

Steel, M., & Penny, D. (2000). Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular Biology and evolution, 17*(6), 839-850.

Tavare, S. (1986). Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences (American Mathematical Society) 17*, 57–86.

Wilgenbusch, J. C., & Swofford, D. (2003). Inferring evolutionary trees with PAUP*. *Curr Protoc Bioinformatics, Chapter 6*, Unit 6 4.

Wollenberg, K. R., & Atchley, W. R. (2000). Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proceedings of the National Academy of Sciences, 97*(7), 3288-3291.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution, 10*(6), 1396-1401.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution, 39*(1), 105-111.

Yang, Z. (1996). Phylogenetic analysis using parsimony and likelihood methods. *Journal of Molecular Evolution, 42*(2), 294-307.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution, 24*(8), 1586-1591.

Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nat Rev Genet, 13*(5), 303-314.

Ypma, T. J. (1995). Historical development of the Newton-Raphson method. *SIAM Review, 37*(4), 531-551.

Κεφάλαιο 7: Μέθοδοι Πρόγνωσης

Σύνοψη

Στο κεφάλαιο αυτό θα ασχοληθούμε με τις μεθόδους πρόγνωσης δομής και λειτουργίας μακρομορίων, τόσο των πρωτεϊνών όσο και του DNA και RNA. Οι μέθοδοι αυτές είναι ιδιαίτερα σημαντικές καθώς έρχονται να καλύψουν το κενό που προκύπτει σε περιπτώσεις που μια νεοανακαλυφθείσα αλληλουχία δεν εμφανίζει σημαντική ομοιότητα με κάποια άλλη γνωστής δομής ή λειτουργίας. Θα παρουσιάσουμε τις βασικές αρχές με τις οποίες μπορεί να κατασκευαστεί μια προγνωστική μέθοδος, καθώς και τα πιο σημαντικά παραδείγματα τέτοιων μεθόδων τα οποία παρουσιάζουν μεγάλο θεωρητικό και πρακτικό ενδιαφέρον. Έτσι, θα δούμε την πρόγνωση της δευτεροταγούς δομής πρωτεϊνών, την πρόγνωση των διαμεμβρανικών τμημάτων, την πρόγνωση των σηματοδοτικών αλληλουχιών αλλά και παραδείγματα πρόγνωσης μετα-μεταφραστικών τροποποιήσεων. Στην περίπτωση του DNA θα δούμε τις μεθόδους εύρεσης γονιδίων, αλλά και άλλα σχετιζόμενα προβλήματα (εύρεση σημείων αποκοπής εξωνίων/εσωνίων, πρόγνωση πολυαδενυλίωσης κ.ο.κ.), ενώ για RNA θα εστιάσουμε στις μεθόδους πρόγνωσης των micro RNA και των στόχων τους.

Προαπαιτούμενη γνώση

Στο κεφάλαιο αυτό θεωρείται απαραίτητη η γνώση των κεφαλαίων 2, 3, 4 και 5.

7. Εισαγωγή

Στο κεφάλαιο αυτό θα ασχοληθούμε με τις μεθόδους πρόγνωσης που κάνουν χρήση αλληλουχιών πρωτεϊνών ή DNA/RNA. Οι μέθοδοι πρόγνωσης καλύπτουν ένα πολύ σημαντικό κομμάτι της σύγχρονης βιοπληροφορικής έρευνας και στην πράξη χρησιμοποιούνται καθημερινά τόσο από ειδικούς της βιοπληροφορικής (όταν κάνουν αναλύσεις γονιδιωμάτων ή όταν μελετούν μια νέα πρωτεϊνική οικογένεια κ.ο.κ.), όσο και από μοριακούς βιολόγους όταν μελετούν μια συγκεκριμένη πρωτεΐνη ή ένα νέο γονίδιο που έχει εντοπιστεί ή σε πολλές άλλες αντίστοιχες περιπτώσεις. Οι μέθοδοι πρόγνωσης έρχονται χρονικά αλλά και λογικά να καλύψουν το κενό που έχουν αφήσει οι μέθοδοι ομοιότητας των προηγούμενων κεφαλαίων. Όταν έχουμε στα χέρια μας μια άγνωστη αλληλουχία γονιδίου ή πρωτεΐνης, το πρώτο πράγμα που πρέπει να κάνουμε, είναι να ελέγξουμε με τις μεθόδους αναζήτησης ομοιότητας αν μοιάζει σε σημαντικό βαθμό με κάποια αλληλουχία με γνωστά χαρακτηριστικά (δομής ή/και λειτουργίας) και αν έχει αρκετές ομόλογες αλληλουχίες να ελέγξουμε την πολλαπλή τους στοίχιση και τα κοινά μοτίβα που μπορεί να εμφανίζονται. Όταν αναφερόμαστε σε δομικά χαρακτηριστικά (αλλά το ίδιο ισχύει και για τα περισσότερα λειτουργικά χαρακτηριστικά), αν μια αλληλουχία μοιάζει σε μεγάλο βαθμό με μια άλλη γνωστής δομής και λειτουργίας, τότε τα περισσότερα προβλήματα έχουν λυθεί: μπορούμε να κατασκευάσουμε εύκολα ένα τρισδιάστατο μοντέλο της δομής της με προτυποποίηση με βάση την ομολογία (αν μιλάμε για πρωτεΐνη) και να κάνουμε μια πολύ καλή εκτίμηση για την πιθανή λειτουργία της. Προφανώς, όσο μεγαλύτερη είναι η ομοιότητα, τόσο πιο εύκολη είναι αυτή η διαδικασία και τόσο πιο μεγάλη ακρίβεια μας δίνει.

Τα προβλήματα αρχίζουν όταν η αλληλουχία μας δεν έχει σημαντική ομοιότητα με καμία άλλη αλληλουχία από αυτές που βρίσκονται κατατεθειμένες στις βάσεις δεδομένων ή, όταν έχει μεν μεγάλη ομοιότητα αλλά μόνο με άλλες αλληλουχίες, επίσης άγνωστης δομής και λειτουργίας. Εκτιμάται, ότι σε κάθε νεοπροσδιορισθέν γονιδίωμα, περίπου το 20-30% των γονιδίων αντιστοιχούν σε πρωτεϊνικές αλληλουχίες για τις οποίες δεν μπορούν να εξαχθούν σίγουρα συμπεράσματα από μια αναζήτηση ομοιότητας και μόνο. Προφανώς, με τη συνεχή συσσώρευση γονιδιωμάτων και αλληλουχιών, το ποσοστό των «εντελώς νέων» πρωτεϊνών θα μειώνεται συνεχώς, αλλά αυτές που μοιάζουν με κάποιες άλλες άγνωστης όμως δομής και λειτουργίας θα εξακολουθούν να υπάρχουν. Για να λυθεί αυτό το πρόβλημα, έχουν αναπτυχθεί και διάφορες μεθοδολογίες όπως αυτές της αναζήτησης μακρινών ομοιοτήτων (remote homology) ή τεχνικές ύφανσης (threading), αλλά και πάλι το πρόβλημα παραμένει σε μεγάλο βαθμό. Αυτό το κενό έρχονται να καλύψουν οι μέθοδοι πρόγνωσης, των οποίων ο σκοπός είναι να προβλέπουν δομικά ή λειτουργικά χαρακτηριστικά για μία αλληλουχία πρωτεΐνης ή DNA, χρησιμοποιώντας μόνο την ακολουθία της.

Η χρησιμότητα των μεθόδων πρόγνωσης λοιπόν φαίνεται από το γεγονός ότι είναι απαραίτητες για ένα μεγάλο υποσύνολο των πρωτεϊνών από τα νεοανακαλυφθέντα γονιδιώματα και από το ότι προσφέρουν αρκετές πληροφορίες για τις αλληλουχίες αυτές. Αν αναλογιστεί κανείς ότι τα μοριακά δεδομένα (γονιδιώματα, γονίδια, πρωτεΐνες κ.ο.κ.) συσσωρεύονται με εκθετικούς ρυθμούς, τότε γίνεται εύκολα

αντιληπτό ότι ο πειραματικός έλεγχος όλων αυτών είναι πρακτικά αδύνατος. Για παράδειγμα, ενώ πλέον οι αλληλουχίες προσδιορίζονται με διαδικασίες ρουτίνας, οι τρισδιάστατες δομές απαιτούν εντατική ενασχόληση ενώ για κάποιες ειδικές κατηγορίες πρωτεϊνών τα πράγματα είναι πολύ πιο δύσκολα (όπως για παράδειγμα οι μεμβρανικές πρωτεΐνες). Κατά συνέπεια, το κενό ανάμεσα στον αριθμό αλληλουχιών και αυτών των δομών δεν αναμένεται να καλυφθεί ποτέ. Παρόμοια είναι και η κατάσταση στη διερεύνηση της λειτουργίας μιας πρωτεΐνης. Καταλαβαίνουμε λοιπόν ότι οι μέθοδοι πρόγνωσης είναι ένα απαραίτητο κομμάτι της βιοπληροφορικής και έρχονται να καλύψουν το κενό αυτό, «συλλέγοντας» πληροφορίες για τις άγνωστες αλληλουχίες. Φυσικά, δεν υπάρχει μέθοδος που να προβλέπει τέλεια τη δομή ή κάποιο χαρακτηριστικό μιας πρωτεΐνης, ούτε και μέθοδος για κάθε πιθανή λειτουργία, αλλά η εφαρμογή μεθόδων πρόγνωσης σε νεοπροσδιορισμένες αλληλουχίες μπορεί να μειώσει δραστικά τον αριθμό των πειραμάτων που απαιτούνται για την πειραματική αξιολόγηση, καθοδηγώντας κατά κάποιον τρόπο τα επόμενα βήματα. Για παράδειγμα, με τις μεθόδους πρόγνωσης μπορούμε να πάρουμε μια εικόνα για την πιθανή δευτεροταγή δομή της πρωτεΐνης και τη δομική της ταξινόμηση, να δούμε αν είναι διαμεμβρανική πρωτεΐνη ή όχι, να δούμε αν έχει θέσεις δράσης γλυκοζυλίωσης ή άλλων μετα-μεταφραστικών τροποποιήσεων, να δούμε αν είναι εκκρινόμενη πρωτεΐνη κ.ο.κ. Με όλες αυτές τις μεθόδους, μπορούμε να πάρουμε μια φευγαλέα μεν αλλά αρκετά περιεκτική εικόνα για το πώς περίπου είναι και το τι περίπου κάνει αυτή η πρωτεΐνη, με συνέπεια να μπορούμε να σχεδιάσουμε στοχευμένα πειράματα για να απαντήσουμε σε εξειδικευμένα ερωτήματα.

Ο βασικός τρόπος με τον οποίο λειτουργούν αυτές οι μέθοδοι είναι με την «εκπαίδευση» σε κάποια γνωστά παραδείγματα. Κατόπιν, και αν η διαδικασία έχει γίνει σωστά, υπάρχει η ελπίδα ότι η μέθοδος θα προβλέπει σωστά τα αντίστοιχα χαρακτηριστικά ακόμα και σε εντελώς διαφορετικές αλληλουχίες. Όπως θα δούμε, υπάρχει ένα τεράστιο εύρος εφαρμογών τέτοιων μεθόδων με μεγάλη πρακτική χρησιμότητα, αλλά και διαφορετικών μαθηματικών και υπολογιστικών τεχνικών που χρησιμοποιούνται για το σκοπό αυτό. Για παράδειγμα στις πρωτεΐνες, η πρόγνωση της δευτεροταγούς δομής είναι μια από τις παλαιότερες ενασχολήσεις των βιοπληροφορικών (ήδη από τη δεκαετία του 1970) και εξακολουθεί σε διάφορες παραλλαγές να είναι ενεργός κλάδος μέχρι και σήμερα (πρόγνωση διαμεβρανικών τμημάτων κλπ). Επίσης, η πρόβλεψη ιδιαίτερων χαρακτηριστικών των πρωτεϊνικών δομών, όπως οι θέσεις δράσης διαφόρων ενζύμων (μετα-μεταφραστική τροποποίηση, δισουλφιδικοί δεσμοί, σηματοδοτικές αλληλουχίες κλπ) είναι ιδιαίτερα σημαντικός κλάδος. Η λειτουργική πρόβλεψη επίσης, είναι ιδιαίτερα σημαντική, καθώς έχουν αναπτυχθεί μέθοδοι που προβλέπουν λ.χ. το αν μια πρωτεΐνη είναι ένζυμο και τι είδους αντίδραση καταλύει, το αν δεσμεύει DNA ή όχι, κ.ο.κ. Στην περίπτωση των αλληλουχιών DNA, το κλασικότερο παράδειγμα είναι η εύρεση γονιδίων (gene finding), πρόβλημα το οποίο είναι σημαντικό τόσο σε ευκαρυωτικούς όσο και προκαρυωτικούς οργανισμούς, και είναι μια μέθοδος που χρησιμοποιείται συνεχώς στον προσδιορισμό νέων γονιδιωμάτων. Φυσικά, το πρόβλημα αυτό είναι τεράστιο, γι' αυτό και έχουν αναπτυχθεί και μέθοδοι για ειδικές περιπτώσεις όπως η αναγνώριση υποκινητών, η αναγνώριση εσωνίων-εξωνίων, η πρόβλεψη της πολυαδενυλίωσης του RNA κ.ο.κ. Επίσης, ιδιαίτερα τα τελευταία χρόνια έχει δοθεί μεγάλη έμφαση στην πρόβλεψη των microRNA αλλά και των στόχων τους.

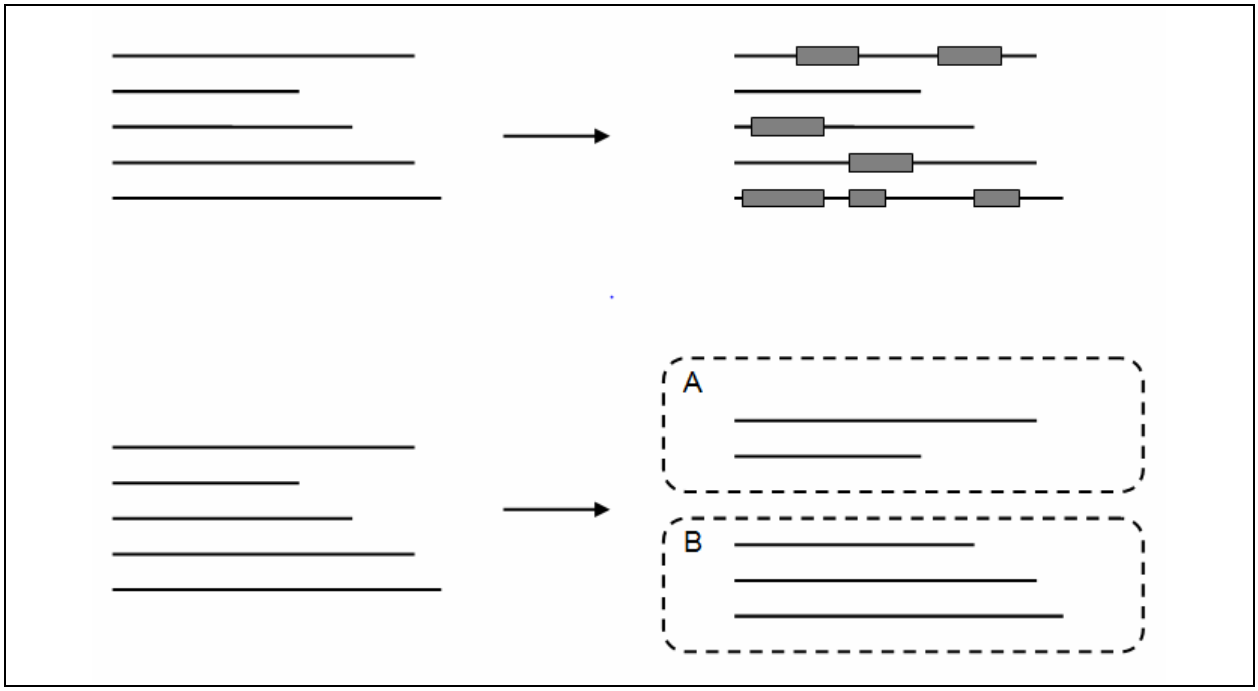
Στο κεφάλαιο αυτό, θα αναλύσουμε τις βασικές μεθοδολογίες που χρησιμοποιούνται στην πρόγνωση λειτουργικών και δομικών χαρακτηριστικών με χρήση αλληλουχιών. Θα αναλύσουμε τις βασικές κατηγορίες τέτοιων μεθόδων, θα δούμε πώς κατασκευάζεται και πώς αξιολογείται μια τέτοια μέθοδος ενώ στο τέλος θα δούμε λεπτομέρειες για τις κυριότερες τέτοιες μεθόδους που υπάρχουν διαθέσιμες σήμερα.

7.1. Κωδικοποίηση των αλληλουχιών

Οι βασικές αρχές όλων των μεθόδων πρόγνωσης στηρίζονται αρχικά σε κάποιες στατιστικές παρατηρήσεις. Για παράδειγμα η Αλανίνη, το Γλουταμικό και η Λευκίνη έχουν ισχυρή προτίμηση να βρίσκονται σε α-έλικα ενώ η Προλίνη, η Γλυκίνη και η Σερίνη όχι, τα υδρόφοβα αμινοξέα έχουν ισχυρή προτίμηση να βρίσκονται σε διαμεμβρανικές περιοχές, ενώ τα υδρόφιλα και τα πολικά, όχι, οι σηματοδοτικές αλλλληλουχίες για τις εκκρινόμενες πρωτεΐνες έχουν συνήθως στο σημείο αποκοπής την αλληλουχία A-X-A, ενώ η γλυκοζυλίωση των πρωτεϊνών στο σύστημα Golgi γίνεται σε αλληλουχίες N-X-[ST]. Στο DNA η έναρξη όλων των γονιδίων κωδικοποιείται από το κωδικόνιο A-U-G, ενώ στο σημείο αποκοπής εξωνίου-εσωνίου, τα νουκλεοτίδια που βρίσκονται συνήθως είναι A-G και G-T αντίστοιχα, κ.ο.κ. Όπως γίνεται ήδη φανερό, μια πρώτη μορφή «μεθόδου πρόγνωσης» είναι δυνατό να κατασκευαστεί με τη χρήση των μεθόδων εύρεσης προτύπων και προφίλ σε αλληλουχίες. Πραγματικά, για πολλές από τις περιπτώσεις πρόγνωσης, τα πρώτα χρόνια χρησιμοποιήθηκαν εντατικά τα πρότυπα της PROSITE. Φυσικά, τα απλά πρότυπα έχουν το πρόβλημα ότι δεν

είναι δυνατό να αποδώσουν σύνθετες δομές, αλλά ακόμα και σήμερα για αρκετές κατηγορίες, τέτοιες μέθοδοι ή επεκτάσεις τους, τα προφίλ, τα HMM και τα προφίλ HMM (τα οποία θα αναφερθούν στο επόμενο κεφάλαιο), θεωρούνται οι καλύτερες εναλλακτικές. Ένα σημαντικό πλεονέκτημα των μεθόδων αυτών, είναι ότι αντιμετωπίζουν εγγενώς (λόγω της γραμματικής που περιέχουν) την αλληλουχία και τα σύμβολά της, όπως πραγματικά είναι, δηλαδή ως διακριτά σύμβολα σε σειρά.

Στα γενικότερα όμως προβλήματα, όπως π.χ. στην πρόγνωση δευτεροταγούς δομής, η εγγενής «ασάφεια» των κανόνων της πρωτεϊνικής αναδίπλωσης, η συμμετοχή αλληλεπιδράσεων μεγάλης απόστασης κατά μήκος της αλληλουχίας και οι μη γραμμικές συσχετίσεις, έχουν κάνει απαραίτητη τη χρήση γενικότερων τεχνικών που χρησιμοποιούνται στη στατιστική και στη μηχανική μάθηση. Το βασικό πρόβλημα που προκύπτει σε τέτοιες περιπτώσεις, είναι η ανάγκη η αλληλουχία συμβόλων να μετατραπεί με κάποιον τρόπο σε αριθμητικά δεδομένα για να μπορέσουν να εφαρμοστούν οι μέθοδοι αυτές.

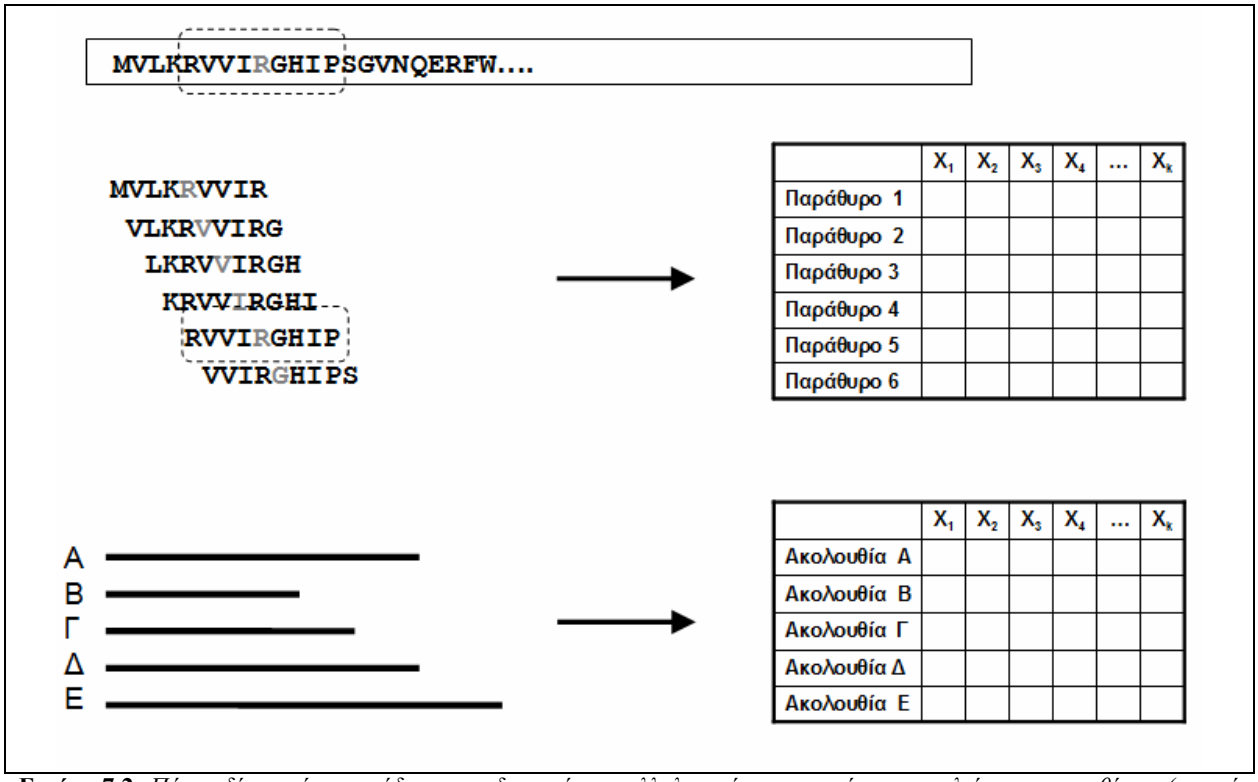


Εικόνα 7.1: *Παραδείγματα μεθόδων πρόγνωσης. Πάνω, δίνεται ένα υποθετικό παράδειγμα πρόγνωσης κάποιου χαρακτηριστικού κατά μήκος της αλληλουχίας. Κάτω, δίνεται ένα υποθετικό παράδειγμα διαχωρισμού των αλληλουχιών σε ομάδες.*

Γενικά, υπάρχουν δύο κατηγορίες προβλημάτων πρόγνωσης ή πρόβλεψης (Εικόνα 7.1). Στην πρώτη περίπτωση ενδιαφερόμαστε για τοπική πρόγνωση κατά μήκος της αλληλουχίας. Ενδιαφερόμαστε δηλαδή να δούμε ποια συγκεκριμένα κατάλοιπα ή νουκλεοτίδια ανήκουν σε μια κατηγορία και ποια σε άλλη. Τέτοια παραδείγματα είναι πολύ συνηθισμένα, καθώς σε αυτήν την κατηγορία ανήκουν όλες οι περιπτώσεις που περιγράψαμε παραπάνω (δευτεροταγής δομή, διαμεμβρανικά τμήματα, εσώνια/εξώνια, θέσεις γλυκοζυλίωσης κ.ο.κ.). Οι περιοχές που προσπαθούμε να εντοπίσουμε, μπορεί να είναι αρκετά συνηθισμένες (όπως στην περίπτωση της δευτεροταγούς δομής), αλλά και πολύ σπάνιες (όπως στην περίπτωση των θέσεων γλυκοζυλίωσης ή των σημάτων πυρηνικού εντοπισμού). Στη δεύτερη κατηγορία, ενδιαφερόμαστε να ταξινομήσουμε κάποιες αλληλουχίες σε δύο ή περισσότερες κατηγορίες. Έτσι, μπορεί να θέλουμε να διαχωρίσουμε τις πρωτεΐνες σε διαμεβρανικές και μη, να κατατάξουμε τους υποδοχείς GPCR σε διάφορες λειτουργικές ομάδες, να ταξινομήσουμε τις πρωτεΐνες στις δομικές τους κατηγορίες (π.χ. α, β, α/β κ.ο.κ.), να προβλέψουμε την ενζυμική λειτουργία μιας πρωτεΐνης ή ακόμα και να διαχωρίσουμε ολόκληρα γονιδιώματα. Σε αυτό το σημείο πρέπει να έχουμε στο μυαλό μας, ότι σε κάποια προβλήματα οι κατηγορίες είναι σχεδόν ισοδύναμες αριθμητικά (π.χ. οι διαμεμβρανικές πρωτεΐνες και οι σφαιρικές υδατοδιαλυτές), ενώ σε άλλα προβλήματα μπορεί η μία από τις κατηγορίες να είναι ιδιαίτερα σπάνια (π.χ. τα διαμεμβρανικά β-βαρέλια), ενώ υπάρχουν και περιπτώσεις στις οποίες οι κατηγορίες που επιθυμούμε να κατατάξουμε τις πρωτεΐνες είναι πολλές. Τέλος, κάτι που χρειάζεται μεγάλη προσοχή είναι το γεγονός ότι πολλές φορές τα προβλήματα είναι

αλληλοσυνδεόμενα, αλλά η αντιμετώπιση τελείως διαφορετική. Για παράδειγμα, μια μέθοδος πρόγνωσης διαμεμβρανικών τμημάτων μπορεί να απαντήσει και στο ερώτημα αν μια πρωτεΐνη είναι μεμβρανική ή όχι. Παρ' όλα αυτά, χρειάζεται μεγάλη προσοχή γιατί υπάρχουν μέθοδοι που αποδίδουν πολύ καλά και σε μη μεμβρανικές πρωτεΐνες (με την έννοια ότι δεν προβλέπουν λάθος διαμεμβρανικά), ενώ άλλες δουλεύουν καλά μόνο σε διαμεμβρανικές (με την έννοια ότι προβλέπουν καλά την ύπαρξη διαμεμβρανικών τμημάτων όταν αυτά υπάρχουν).

Όπως είναι φανερό από τα παραπάνω, οι δύο κατηγορίες μεθόδων απαιτούν και διαφορετικούς τρόπους χειρισμού των δεδομένων αλληλουχιών (Εικόνα 7.2). Στην πρώτη περίπτωση, στην περίπτωση τοπικής πρόβλεψης (*τοπική κωδικοποίηση*), αναγκαστικά θα καταφύγουμε σε μια αναπαράσταση της αλληλουχίας με τη χρήση της τεχνικής του κινούμενου παραθύρου. Με αυτή την τεχνική, ένα κινούμενο παράθυρο ολισθαίνει κατά μήκος της ακολουθίας και κάθε φορά το παράθυρο αυτό «καθορίζει» τη φύση ενός καταλοίπου (συνήθως του κεντρικού). Η ιδέα βασίζεται στη στατιστική ομαλοποίηση (smoothing), και σύμφωνα με αυτή οι ιδιότητες όλου του παραθύρου καθορίζουν τη φύση του εκάστοτε καταλοίπου. Στην περίπτωση των διαμεμβρανικών τμημάτων των πρωτεϊνών, καταλαβαίνουμε εύκολα τη διαίσθηση πίσω από τη μέθοδο (αν βρεις 15 υδρόφοβα κατάλοιπα στη σειρά, είναι πολύ πιο πιθανό να έχεις εντοπίσει μια διαμεμβρανική περιοχή). Το ίδιο ισχύει και στην περίπτωση προβλημάτων που αντιμετωπίζονται με απλά πρότυπα (αναμένεις να βρεις κάποια συγκεκριμένα κατάλοιπα σε κάθε θέση του προτύπου). Σε άλλες περιπτώσεις τα πράγματα είναι πιο ασαφή, όπως π.χ. στην περίπτωση της δευτεροταγούς δομής, στην οποία τα πράγματα είναι πιο σύνθετα αλλά και πάλι οι ίδιοι κανόνες ισχύουν και εδώ (και για την ακρίβεια, αυτό ήταν το πρώτο πρόβλημα από το οποίο ξεκίνησε η ανάπτυξη των μεθόδων αυτών).



Εικόνα 7.2: *Πάνω, δίνεται ένα παράδειγμα κωδικοποίησης αλληλουχιών με τη χρήση του κυλιόμενου παραθύρου (τοπική κωδικοποίηση). Κάτω, δίνεται ένα παράδειγμα ολικής κωδικοποίησης στην οποία κάθε αλληλουχία ανεξαρτήτως μήκους μετασχηματίζεται σε ένα διάνυσμα με συγκεκριμένο αριθμό παραμέτρων.*

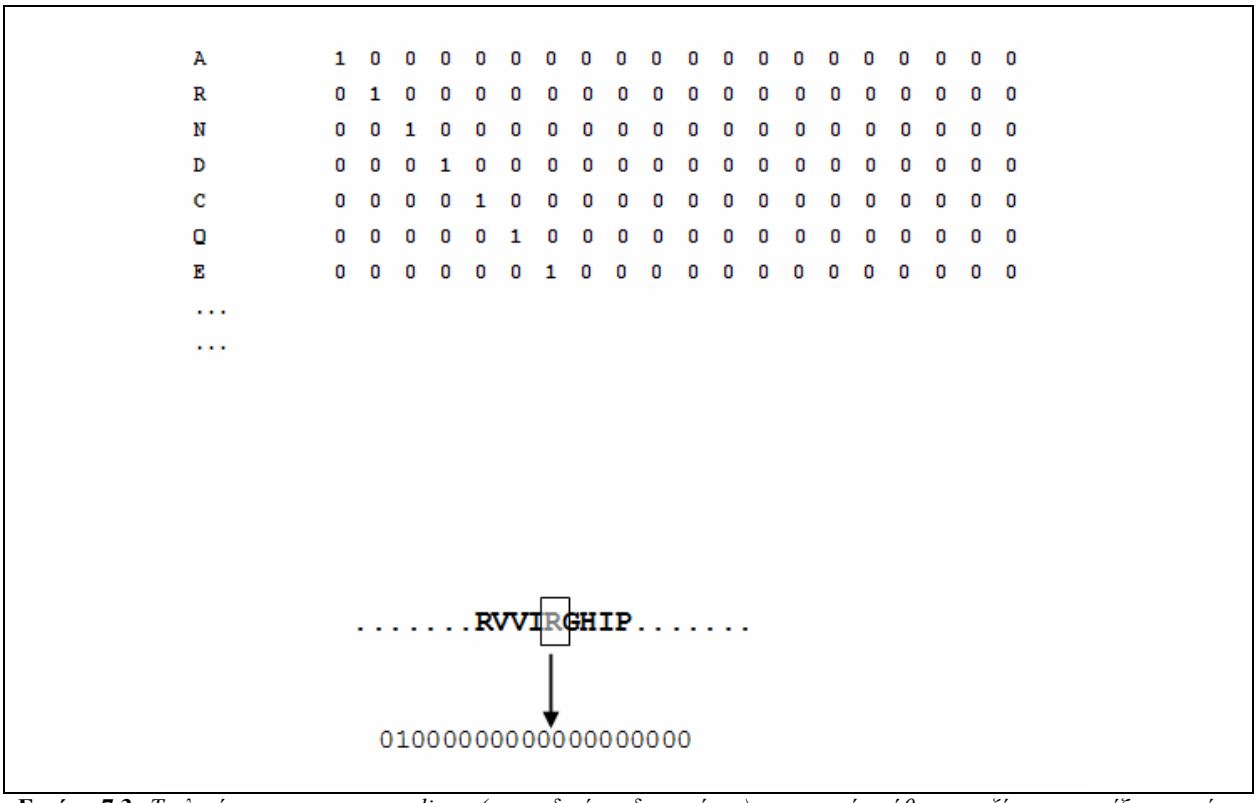
Φυσικά, υπάρχουν πολλά σημεία που απαιτούν διευκρινίσεις και μπορεί να διαφέρουν από μέθοδο σε μέθοδο. Ένα πρώτο θέμα έχει να κάνει με το μήκος του παραθύρου, και εξαρτάται πολύ από το συγκεκριμένο πρόβλημα. Στα περισσότερα προβλήματα πρόβλεψης δομής (δευτεροταγής δομή, διαμεβρανικές έλικες, προσβασιμότητα του διαλύτη κλπ) τα παράθυρα είναι της τάξης των 10-20 αμινοξέων, αν και όπως θα ανέμενε κανείς οι πρώιμες μέθοδοι είχαν χρησιμοποιήσει μικρότερα. Σε άλλα προβλήματα που ανάγονται σε

εύρεση συγκεκριμένων προτύπων, όπως π.χ. οι θέσεις γλυκοζυλίωσης, τα παράθυρα μπορεί να είναι μικρότερα. Γενικά, όσο μεγαλύτερο είναι ένα παράθυρο τόσο περισσότερη πληροφορία γύρω από το κατάλοιπο του ενδιαφέροντος μπορεί να χρησιμοποιηθεί, αλλά αυτό αυξάνει τον αριθμό των παραμέτρων του μοντέλου. Από την άλλη μεριά, από ένα σημείο και μετά η επιπλέον αύξηση του μεγέθους του παραθύρου εισάγει θόρυβο οπότε στα περισσότερα προβλήματα δεν θα δούμε παράθυρα με μέγεθος μεγαλύτερο από τα 30 αμινοξικά κατάλοιπα. Η συμμετρία του παραθύρου είναι ένα άλλο θέμα. Συνήθως στα περισσότερα προβλήματα τα παράθυρα είναι συμμετρικά με μήκος που αντιστοιχεί σε περιττό αριθμό (π.χ. ένα συμμετρικό παράθυρο με μήκος 9 αντιστοιχεί σε ±5 αμινοξικά κατάλοιπα εκατέρωθεν του κεντρικού, κ.ο.κ.). Σε κάποιες ειδικές περιπτώσεις όμως, όπως π.χ. όταν η περιοχή που θέλουμε να εντοπίσουμε βρίσκεται στην αρχή ή στο τέλος της αλληλουχίας (όπως για παράδειγμα στις σηματοδοτικές αλληλουχίες έκκρισης), το παράθυρο δουλεύει καλύτερα όταν είναι μη συμμετρικό.

Τέλος, το πιο σημαντικό θέμα έχει να κάνει με το πώς κωδικοποιείται η πληροφορία της αλληλουχίας του παραθύρου και με το πώς συνδυάζεται για να δώσει μια τελική πρόβλεψη για το κεντρικό κατάλοιπο του παραθύρου. Μια πρώτη προσέγγιση θα μπορούσε να γίνει, με βάση όσα έχουμε δει μέχρι τώρα, με τη χρήση ενός προσθετικού σκορ όπως αυτά που είδαμε στο Κεφάλαιο 3. Αυτή η μέθοδος είναι στατιστικά ορθή, εύκολα κατανοητή και εισηγείται αυτόματα και τον τρόπο με τον οποίο η πληροφορία του κάθε καταλοίπου θα συνδυαστεί (το σκορ το οποίο είναι ήδη σε λογαριθμική κλίματα, θα προστεθεί για όλο το παράθυρο). Όταν επιθυμούμε να χρησιμοποιήσουμε μια κωδικοποίηση που βασίζεται σε κάποιο είδος πρότερης γνώσης σχετικά με τις φυσικοχημικές ιδιότητες των αμινοξέων, υπάρχουν δεκάδες επιλογές. Στην ιστοσελίδα http://web.expasy.org/protscale/ υπάρχουν διαθέσιμες πάρα πολλές επιλογές κωδικοποίησης βασισμένες σε πειραματικές μετρήσεις για την υδροφοβικότητα, την πολικότητα, την ευελιξία, τον όγκο, το μοριακό βάρος ή την προτίμηση για κάποια συγκεκριμένη δευτεροταγή δομή. Με αυτόν τον τρόπο μπορούν να επιλεχθούν (πάντα βέβαια, σε συνάρτηση με το πρόβλημα που θέλουμε να λύσουμε) μία ή περισσότερες από αυτές τις παραμέτρους και να προχωρήσουμε στην κωδικοποίηση. Αν έχουμε λοιπόν παράθυρα με μέγεθος k, τότε επιλέγοντας p από αυτές τις μεταβλητές, σε κάθε παράθυρο θα έχουμε pk ψηφία, ενώ μια αλληλουχία με L αμινοξέα, θα έχει $(L-k+1)$ παράθυρα και συνολικά θα πρέπει να κωδικοποιηθεί με $pk(L-k+1)$ μεταβλητές. Φυσικά, υπάρχουν και άλλες παραπλήσιες εναλλακτικές, π.χ. με χρήση λόγων πιθανοτήτων που θα συνδυαστούν πολλαπλασιαστικά ή με πίνακα σκορ ειδικό ανά θέση όπως στην περίπτωση των weight matrices (το οποίο αναμένεται να είναι καλύτερο αλλά αυξάνει και άλλο τον αριθμό των παραμέτρων). Γενικά, όλες οι μεθοδολογίες που συναντήσαμε στα κεφάλαια 3 (προσθετικά σκορ), 5 (μοτίβα, πίνακες κ.ο.κ.) αλλά και αυτές που θα συναντήσουμε στο κεφάλαιο 8 (HMM), υπάγονται σε αυτήν την κατηγορία. Στην πιο ακραία περίπτωση η πληροφορία θα συνδυαστεί με κάποια μέθοδο τεχνητής νοημοσύνης όπως τα νευρωνικά δίκτυα, η οποία εκτός από το πρόβλημα της ειδικής αντιμετώπισης της κάθε θέσης θα λύσει και το πρόβλημα των συσχετίσεων.

Σε γενικότερα προβλήματα που λύνονται με τέτοιου είδους μεθόδους, η απευθείας κωδικοποίηση της ίδιας της αλληλουχίας και όχι η χρήση κάποιου σκορ είναι προτιμότερη, αλλά όπως θα δούμε αυξάνει εκθετικά τον αριθμό των παραμέτρων του μοντέλου. Ο πιο συχνά χρησιμοποιούμενος, αλλά και ο πιο μαθηματικά σωστός τρόπος, για την κωδικοποίηση των αλληλουχιών σε ένα παράθυρο κατά μήκος της αλληλουχίας, είναι με το λεγόμενο sparse encoding (η σποραδική κωδικοποίηση) στον οποίο κάθε αμινοξύ ή νουκλεοτίδιο αναπαρίσταται με ένα διάνυσμα 20 ή 4 ψηφίων από τα οποία ένα μόνο κάθε φορά θα είναι 1 και τα υπόλοιπα 0 (Εικόνα 7.3). Ο τρόπος αυτός, ο οποίος στη στατιστική ονομάζεται «dummy variables», είναι μαθηματικά σωστός γιατί κάθε σύμβολο αντιμετωπίζεται σαν ξεχωριστός χαρακτήρας και αποφεύγεται η εισαγωγή τεχνητών συσχετίσεων (η οποία θα μπορούσε να προκύψει αν είχαμε χρησιμοποιήσει μια κωδικοποίηση με λιγότερα ψηφία). Παρ' όλα αυτά, είναι φανερό ότι οδηγεί σε μεγάλη υπολογιστική σπατάλη καθώς κάθε σύμβολο (στην περίπτωση των πρωτεϊνών) θα χρησιμοποιεί 20 ψηφία. Αν έχουμε λοιπόν παράθυρα με μέγεθος k τότε σε κάθε παράθυρο θα έχουμε 20k ψηφία, ενώ μια αλληλουχία με L αμινοξέα, θα έχει $(L-k+1)$ παράθυρα και συνολικά θα πρέπει να κωδικοποιηθεί με $20k(L-k+1)$ ψηφία (δηλαδή μεταβλητές). Έχουν προταθεί και άλλες μορφές κωδικοποίησης είτε προσαρμοστικές, δηλαδή με αλγορίθμους που να προσαρμόζονται στο εκάστοτε πρόβλημα, είτε γενικές κατά τις οποίες επιλέγεται κάποια πιο γενική μορφή που να προσδίδει κάποια πλεονεκτήματα. Για παράδειγμα, μια κωδικοποίηση που βασίζεται στην ταξινόμηση των αμινοξέων με βάση τις φυσικοχημικές τους ιδιότητες (υδρόφοβα, πολικά, αρωματικά, θετικά φορτισμένα κ.ο.κ.) μπορεί να μειώσει των αριθμό των ψηφίων στα 7 έως 9, ενώ παράλληλα αντιμετωπίζει τις συσχετίσεις που θα προκύπτουν με αποφασιστικό τρόπο. Μια άλλη περίπτωση, θα ήταν να χρησιμοποιηθεί απευθείας η κωδικοποίηση από τον πίνακα BLOSUM62 (ή κάποιον παρόμοιο), μια προσέγγιση που δεν θα μείωνε τον

αριθμό των παραμέτρων αλλά θα εισήγαγε την επιπλέον πληροφορία για τις σχέσεις των αμινοξέων μεταξύ τους. Τέλος, στην πιο ακραία περίπτωση, θα μπορούσαμε να έχουμε την κωδικοποίηση από ένα PSSM, η οποία θα έδινε και τις επιπλέον πληροφορίες για τις ανά θέση προτιμήσεις των αμινοξέων και θα βελτίωνε κατά πολύ την απόδοση της μεθόδου. Στην πράξη, αυτή η εναλλακτική χρησιμοποιείται στους περισσότερους αλγόριθμους πρόγνωσης (δευτεροταγούς δομής, διαμεμβρανικών τμημάτων κ.ο.κ.).



Εικόνα 7.3: *Το λεγόμενο «sparse encoding» (σποραδική κωδικοποίηση) στην οποία κάθε αμινοξύ αντιστοιχίζεται σε ένα διάνυσμα 20 ψηφίων εκ των οποίων το ένα μόνο είναι «1» ενώ τα υπόλοιπα 19 είναι «0».*

Στη δεύτερη περίπτωση, σε προβλήματα στα οποία ενδιαφερόμαστε να κατατάξουμε μια αλληλουχία σε δύο ή περισσότερες κατηγορίες, χρειαζόμαστε μια μέθοδο *ολικής κωδικοποίησης* της αλληλουχίας. Σε αυτές τις μεθόδους, η αλληλουχία ανεξαρτήτως του μήκους της αναπαρίσταται από ένα διάνυσμα σταθερού μήκους. Ένα κλασικό παράδειγμα αυτής της κατηγορίας αποτελούν τα ποσοστά εμφάνισης των αμινοξέων, μέθοδος με την οποία μπορούμε να κωδικοποιήσουμε οποιαδήποτε πρωτεΐνη σε ένα διάνυσμα 20 μεταβλητών. Με αυτόν τον τρόπο και τη χρήση νευρωνικών δικτύων οι (Reinhardt & Hubbard, 1998) είχαν πετύχει, σε μια από τις πρώτες προσπάθειες του είδους, την πρόγνωση της κυτταρικής στόχευσης (τοποθεσίας) των πρωτεϊνών, τόσο στο βακτηριακό όσο και στο ευκαρυωτικό κύτταρο. Αυτό που γίνεται εμφανές βέβαια, είναι ότι με τη μεθοδολογία αυτή, διευκολύνονται μεν οι υπολογιστικές μεθοδολογίες, αλλά από την άλλη χάνεται ένα σημαντικό μέρος της πληροφορίας που περιέχεται στις αλληλουχίες καθώς πολλές (πρακτικά άπειρες) αλληλουχίες θα αναπαρίστανται με το ίδιο ακριβώς διάνυσμα, ακόμα και αν έχουν τελείως διαφορετικά χαρακτηριστικά (π.χ. η αλληλουχία AAAATTTT και η αλληλουχία ATATATAT θα έχουν ακριβώς την ίδια κωδικοποίηση). Επιπλέον δε, για να δουλέψει στην πράξη η μέθοδος αυτή θα πρέπει οι υπό σύγκριση ομάδες να έχουν σημαντικές διαφορές στις μεταβλητές που χρησιμοποιούνται (στην περίπτωση της κυτταρικής τοποθεσίας, υπάρχουν όντως ενδείξεις ότι πρωτεΐνες που βρίσκονται σε διαφορετικά οργανίδια, έχουν σημαντικές διαφορές στην αμινοξική σύσταση).

Τα προβλήματα αυτής της προσέγγισης, μπορούν να αντιμετωπιστούν μόνο μερικώς καθώς δεν είναι δυνατό να λυθεί τελείως το τελευταίο πρόβλημα, αυτό της απώλειας πληροφορίας. Έτσι, κάποιοι έχουν προτείνει τη χρήση δι- και τρι-πεπτιδίων, μια προσέγγιση που αυξάνει όμως αρκετά τον αριθμό των παραμέτρων του μοντέλου (400 και 8000 αντίστοιχα). Μια άλλη προσέγγιση, θα ήταν να χρησιμοποιηθούν άλλου είδους πληροφορίες συνοπτικής φύσης, όπως το μοριακό βάρος της πρωτεΐνης, η συνολική

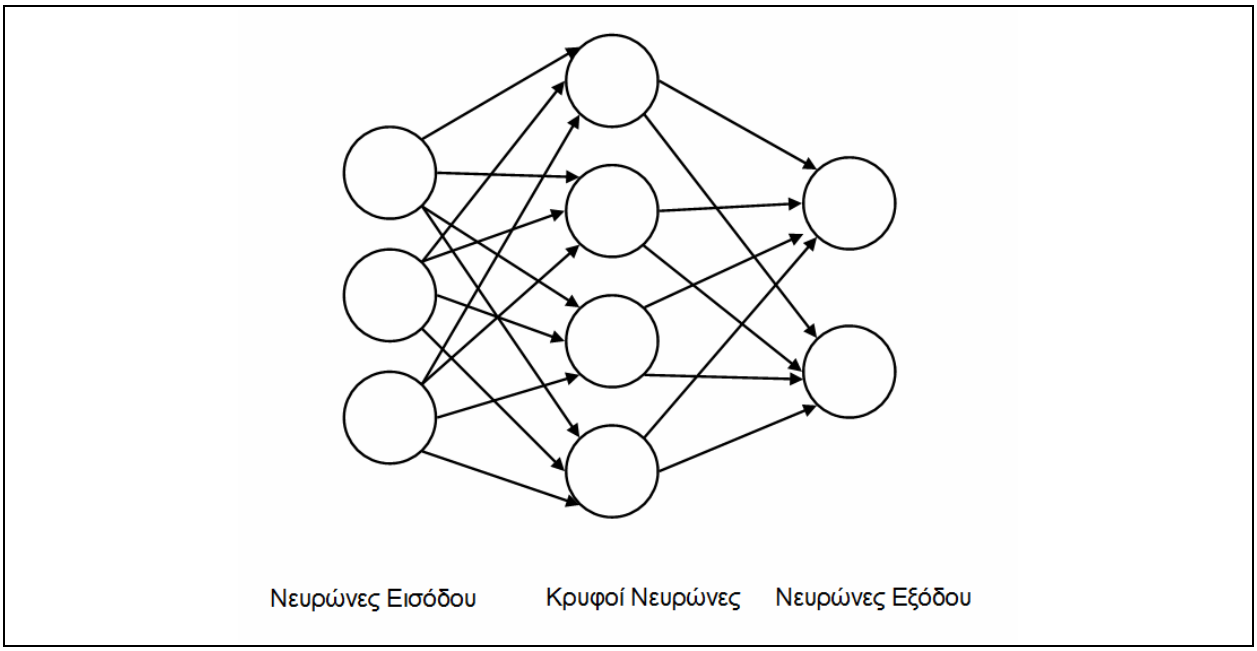
υδροφοβικότητα, η ύπαρξη άλλων χαρακτηριστικών όπως τα διαμεμβρανικά τμήματα, τα πεπτίδια οδηγητές, διάφορα πρότυπα που εμφανίζονται κ.ο.κ. (τα οποία βέβαια με τη σειρά τους προέρχονται από μεθόδους πρόγνωσης!). Μια άλλη εναλλακτική είναι η χρησιμοποίηση μαθηματικών τεχνικών που περιγράφουν την περιοδικότητα που μπορεί να εμφανίζεται σε μια αλληλουχία (π.χ. με μετασχηματισμό Fourier), ενώ η πιο γενικευμένη προσέγγιση είναι η λεγόμενη ψευδοσύσταση σε αμινοξέα (pseudo aminoacid composition) του Chou, η οποία μετράει εκτός από τα αμινοξέα και τις (μέχρι ένα βαθμό) συσχετίσεις τους που εμφανίζονται κατά μήκος της αλληλουχίας. Για παράδειγμα, υπολογίζει (σε μια μεθοδολογία που μοιάζει με τις μαρκοβιανές αλυσίδες), τις συσχετίσεις των αμινοξέων με το επόμενό τους (το *i* με το *i*+1), ή με το μεθεπόμενο (το *i* με το *i*+2), αλλά και παραπάνω (*i*+3). Προφανώς όμως, η παραπάνω αύξηση οδηγεί σε μεγάλη αύξηση του αριθμού των παραμέτρων. Μια διαδικτυακή εφαρμογή που εφαρμόζει τέτοιους μετασχηματισμούς, βρίσκεται διαθέσιμη στη διεύθυνση http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/. Γενικά, το πρόβλημα που κάθε φορά καλούμαστε να λύσουμε μπορεί να υπαγορεύει και την κατάλληλη επιλογή των παραμέτρων, γι' αυτό και χρειάζεται ιδιαίτερα καλή γνώση του εκάστοτε βιολογικού προβλήματος, αλλά και πειραματισμός για την εύρεση του καλύτερου τρόπου κωδικοποίησης.

7.2. Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα, είναι μια μαθηματική τεχνική της τεχνητής νοημοσύνης, με πολλές εφαρμογές στη βιοπληροφορική όπως θα δούμε και παρακάτω (Baldi & Brunak, 2001). Στην ενότητα αυτή θα προσπαθήσουμε να αποδώσουμε τα βασικά στοιχεία της λειτουργίας τους γιατί τα νευρωνικά δίκτυα αποτελούν μια σημαντική μέθοδο που θα συναντήσουμε στην πρόγνωση δευτεροταγούς δομής αλλά και σε άλλες εφαρμογές. Τα νευρωνικά δίκτυα (ή καλύτερα, τα τεχνητά νευρωνικά δίκτυα) είναι υπολογιστικές μηχανές που σκοπό είχαν αρχικά να μιμηθούν τις ικανότητες του ανθρώπινου εγκεφάλου στην αναγνώριση προτύπων (Bishop, 1998). Ο κάθε νευρώνας είναι απλά μια συνάρτηση που δέχεται ερεθίσματα από άλλους νευρώνες και δίνει τελικά ερέθισμα (με βάση τη συνάρτηση αυτή) σε άλλους νευρώνες. Συνήθως, τα δίκτυα τα αναπαριστούμε με ένα γράφο, με τα βέλη να αντιστοιχούν στις συνδέσεις (συνάψεις) μεταξύ των νευρώνων. Πρακτικά, οι νευρώνες διαφοροποιούνται σε νευρώνες εισόδου στους οποίους κωδικοποιούνται οι μεταβλητές εισόδου, σε κρυφούς νευρώνες οι οποίοι δέχονται τα ερεθίσματα από τους νευρώνες εισόδου και στους νευρώνες εξόδου οι οποίοι δέχονται τα ερεθίσματα από τους κρυφούς νευρώνες και τελικά παράγουν το αποτέλεσμα του δικτύου. Καταλαβαίνουμε δηλαδή, πως το συνολικό δίκτυο δεν είναι παρά μια περίπλοκη συνάρτηση που επεξεργάζεται τα δεδομένα εισόδου και παράγει κάποιο τελικό αποτέλεσμα. Φυσικά, με όσα είπαμε παραπάνω, είναι κατανοητό ότι σαν νευρώνες εισόδου μπορούν να χρησιμοποιηθούν μεταβλητές που έχουν προκύψει από μια κατάλληλη κωδικοποίηση μιας βιολογικής αλληλουχίας (είτε με τοπική είτε με ολική κωδικοποίηση), αλλά σε επόμενα κεφάλαια θα δούμε ότι μπορεί να χρησιμοποιηθούν και άλλου είδους δεδομένα, όπως δεδομένα γονιδιακής έκφρασης.

Υπάρχουν πολλών ειδών νευρωνικά δίκτυα, αλλά για λόγους απλότητας θα ασχοληθούμε με τη σημαντικότερη κατηγορία, τα δίκτυα εμπρόσθιας τροφοδότησης (feed forward) στα οποία ένας νευρώνας επικοινωνεί πάντα μόνο με νευρώνες που βρίσκονται σε στρώμα που βρίσκεται παρακάτω, η πληροφορία δηλαδή διαδίδεται πάντα προς τα εμπρός (Εικόνα 7.4). Αν και η συνδεσμολογία μπορεί να σχεδιαστεί, συνήθως για λόγους απλότητας όλοι οι νευρώνες ενός στρώματος επικοινωνούν με όλους τους νευρώνες του επόμενου (πλήρως συνδεδεμένη αρχιτεκτονική). Είναι δυνατό να υπάρχουν δίκτυα με παραπάνω από ένα στρώματα κρυφούς νευρώνες, αλλά και δίκτυα χωρίς κρυφούς νευρώνες. Για την ακρίβεια, πραγματικά «νευρωνικά δίκτυα» θεωρούνται μόνο αυτά που περιέχουν τουλάχιστον ένα στρώμα κρυφών νευρώνων. Τα δίκτυα που δεν διαθέτουν κρυφούς νευρώνες είναι μαθηματικά ισοδύναμα με γραμμικά ή γενικευμένα γραμμικά μοντέλα γνωστά από τη στατιστική (ανάλογα με τον αριθμό των νευρώνων εξόδου και της συνάρτησης ενεργοποίησης είναι δυνατόν να κατασκευαστούν δίκτυα ανάλογα με τη γραμμική παλινδρόμηση, τη λογιστική παλινδρόμηση, τη διαχωριστική ανάλυση, την πολυμεταβλητή γραμμική παλινδρόμηση κ.ο.κ.). Η μεγάλη δύναμη των νευρωνικών δικτύων (με κρυφούς νευρώνες) βρίσκεται στο γεγονός ότι η παρουσία των κρυφών νευρώνων μπορεί να οδηγήσει σε σύνθετες μη-γραμμικές αναπαραστάσεις των δεδομένων εισόδου και με αυτόν τον τρόπο μπορούν να λυθούν προβλήματα που είναι γραμμικά μη-διαχωρίσιμα. Ένα απλό παράδειγμα τέτοιου προβλήματος είναι το πρόβλημα XOR, ενώ ένα αντίστοιχο βιολογικό, είναι η ίδια η ύπαρξη του γενετικού κώδικα, της συνάρτησης δηλαδή που αντιστοιχεί τα κωδικόνια στα αμινοξέα. Ο αριθμός των κρυφών νευρώνων καθορίζει το πόσο «λεπτομερής» θα είναι μια τέτοια συνάρτηση. Για παράδειγμα, ένα νευρωνικό δίκτυο με μεγάλο αριθμό νευρώνων μπορεί να

προσεγγίσει απείρως καλά μια πολυωνυμική συνάρτηση οποιουδήποτε βαθμού (όσο περισσότεροι νευρώνες, τόσο καλύτερη η προσέγγιση). Στη στατιστική ορολογία, οι κρυφοί νευρώνες αντιστοιχούν στις αλληλεπιδράσεις (interaction) μεταξύ των μεταβλητών, μόνο που στην περίπτωση των νευρωνικών δικτύων κατασκευάζουμε μαζικά αλληλεπιδράσεις όλων των πιθανών μεταβλητών. Αυτό όπως θα φανεί στη συνέχεια έχει σαν αρνητικό επακόλουθο την αύξηση του αριθμού των παραμέτρων του μοντέλου, γεγονός που χρειάζεται μεγάλη προσοχή.



Εικόνα 7.4: *Παράδειγμα ενός νευρωνικού δικτύου με 3 νευρώνες εισόδου, 4 κρυφούς νευρώνες (σε ένα στρώμα) και 2 νευρώνες εξόδου. Το δίκτυο είναι εμπρόσθιας τροφοδότησης με πλήρως συνδεδεμένη αρχιτεκτονική.*

Όπως είπαμε, κάθε νευρώνας δέχεται ερεθίσματα από άλλους (Εικόνα 7.5). Έτσι πρέπει να υπάρχει τρόπος να κωδικοποιηθεί αριθμητικά τόσο το ερέθισμα (η τιμή της μεταβλητής) όσο και η σχετική συνεισφορά της στον συγκεκριμένο νευρώνα. Το ρόλο αυτό, παίζουν τα συναπτικά βάρη (weights) με τα οποία μια μεταβλητή συνδέεται με τον νευρώνα. Τα βάρη, είναι στην ουσία οι παράμετροι του μοντέλου και οι τιμές τους πρέπει να βρεθούν με εκπαίδευση όπως θα δούμε παρακάτω. Το κάθε συναπτικό βάρος πολλαπλασιάζεται με την τιμή του νευρώνα εισόδου και οι συνεισφορές όλων των νευρώνων αθροίζονται και περνάνε μέσα από μια συνάρτηση ενεργοποίησης (activation function). Το ανάλογο των συναπτικών βαρών στη στατιστική είναι οι συντελεστές της παλινδρόμησης (συνήθως συμβολίζονται με β). Προσοχή χρειάζεται στο γεγονός ότι όπως ακριβώς και στη στατιστική, έτσι και εδώ χρειάζεται ένας συντελεστής που να δίνει την τιμή του νευρώνα όταν όλα τα δεδομένα εισόδου έχουν τιμή 0. Στα νευρωνικά δίκτυα αυτός ο συντελεστής ονομάζεται πόλωση (bias) και μπορεί να θεωρηθεί ως η τιμή του συναπτικού βάρους για έναν υποθετικό νευρώνα ο οποίος έχει πάντα τιμή +1.

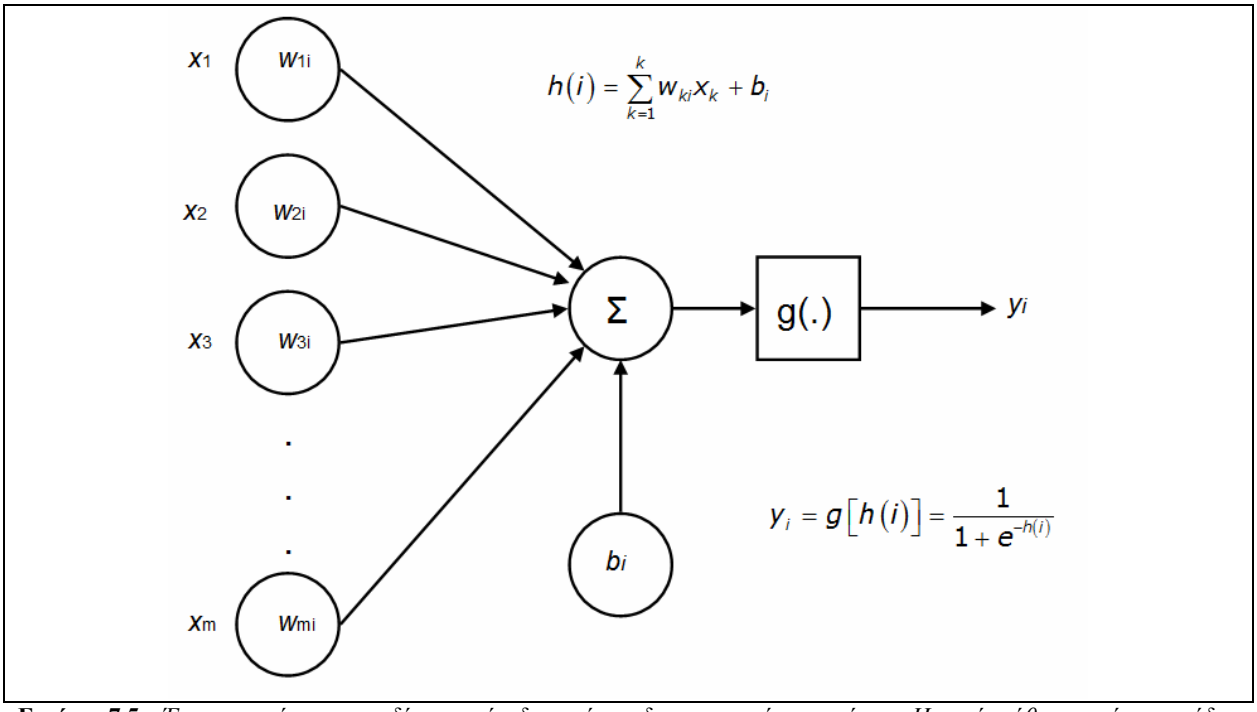
Το είδος της συνάρτησης ενεργοποίησης είναι επίσης κάτι καθοριστικό για τη δομή και τις ιδιότητες του δικτύου (προφανώς, συναρτήσεις ενεργοποίησης έχουν νόημα για τους νευρώνες εξόδου και του κρυφούς νευρώνες). Έτσι, αν το τελικό αποτέλεσμα που θέλουμε να προβλέψουμε είναι δίτιμο (ανήκει σε μια ομάδα/δεν ανήκει), τότε η συνάρτηση ενεργοποίησης του νευρώνα *i* πρέπει να είναι η σιγμοειδής συνάρτηση:

$$g\left[h(i)\right]=\frac{1}{1+e^{-h(i)}}$$

Η συνάρτηση αυτή μοντελοποιεί το αποτέλεσμα του νευρώνα, έτσι ώστε να είναι πάντα μεταξύ 0 και 1, και κατά συνέπεια μπορεί να θεωρηθεί σαν πιθανότητα το δεδομένο παράδειγμα να ανήκει στην συγκεκριμένη κατηγορία. Το ακριβώς αντίστοιχο στη στατιστική είναι η λογιστική παλινδρόμηση (logistic regression). Εκεί, η λογιστική συνάρτηση η οποία είναι η αντίστροφη της σιγμοειδούς, εφαρμόζεται στο αποτέλεσμα και παίρνουμε ακριβώς το ίδιο αποτέλεσμα. Στις περισσότερες περιπτώσεις στη βιοπληροφορική

θα έχουμε τέτοιου είδους προβλήματα και, κατά συνέπεια, τέτοιου είδους συναρτήσεις. Αν σε κάποιο πρόβλημα έχουμε *c* πολλαπλές αμοιβαία αποκλειόμενες ομάδες για να κάνουμε την ταξινόμηση (π.χ. α-έλικα, β-πτυχωτή επιφάνεια, τυχαία δομή), θα πρέπει να ορίσουμε κατάλληλα τους νευρώνες και τότε θα πρέπει να χρησιμοποιηθεί η λεγόμενη συνάρτηση softmax:

$$g\left[h(i)\right]=\frac{e^{-h(i)}}{\sum_{j=1}^{c} e^{-h(j)}}$$



Εικόνα 7.5: *Ένας νευρώνας που δέχεται είσοδο από m διαφορετικούς νευρώνες. Η τιμή κάθε νευρώνα εισόδου πολλαπλασιάζεται με το αντίστοιχο συναπτικό βάρος και αθροίζεται (μαζί με την τιμή του bias) πριν περάσει από τη συνάρτηση ενεργοποίησης η οποία θα δώσει το τελικό αποτέλεσμα. Στο συγκεκριμένο παράδειγμα η συνάρτηση ενεργοποίησης είναι η μη συμμετρική σιγμοειδής.*

Με τη συνάρτηση αυτή, όλα τα αποτελέσματα των c νευρώνων εξόδου, είναι πάντα πιθανότητες μεταξύ 0 και 1 αλλά επιπλέον, αθροίζουν και στη μονάδα. Υπάρχουν βέβαια και περιπτώσεις στις οποίες θα μπορούμε να έχουμε πολλές διαφορετικές κατηγορίες στις οποίες δεν είναι απαραίτητο κάποιο παράδειγμα να ανήκει μόνο σε μία. Ένα τέτοιο φαινόμενο θα δούμε παρακάτω στη σύζευξη των GPCR με τις G-πρωτεΐνες, όπου ένας δεδομένος υποδοχέας μπορεί να κάνει σύζευξη με περισσότερες από μια πρωτεΐνες. Σε αυτή την περίπτωση θα έχουμε ανεξάρτητες μεταξύ τους εξόδους, κάθε μία με τη σιγμοειδή συνάρτηση. Υπάρχουν και άλλες περιπτώσεις συναρτήσεων ενεργοποίησης (συνάρτηση κατωφλίου, ταυτοτική κ.ο.κ.) αλλά δεν έχουν πολλές εφαρμογές στα δικά μας παραδείγματα.

Ειδική μνεία απαιτείται στις συναρτήσεις ενεργοποίησης των κρυφών νευρώνων. Οι κρυφοί νευρώνες, καθώς είναι αυθαίρετα δημιουργήματα μπορούν να έχουν πολλές διαφορετικές συναρτήσεις ενεργοποίησης (ακόμα και ταυτοτικές), αλλά εμπειρικές μελέτες λένε ότι η καλύτερη επιλογή είναι η χρήση μιας συμμετρικής σιγμοειδούς συνάρτησης. Για το σκοπό αυτό μπορεί να χρησιμοποιηθεί μια μικρή τροποποίηση της σιγμοειδούς που να τη «μεταφέρει» σε συμμετρικές τιμές αλλά η καλύτερη και μαθηματικά πιο κομψή επιλογή, είναι η συνάρτηση αντίστροφη-εφαπτομένη (tanh) η οποία δίνεται από τη σχέση:

$$g\left[h(i)\right]=\frac{1-e^{-h(i)}}{1+e^{-h(i)}}$$

η οποία περιορίζει την τιμή της εξόδου μεταξύ -1 και +1.

Το τελευταίο θέμα που χρήζει αναφοράς, είναι το ζήτημα της εκτίμησης παραμέτρων, δηλαδή της εκπαίδευσης του δικτύου. Τα νευρωνικά δίκτυα του είδους που παρουσιάσαμε, είναι μέθοδοι επιβλεπόμενης μάθησης. Χρειάζονται κάποιες παρατηρήσεις με γνωστές (προφανώς) τις τιμές εισόδου, αλλά γνωστές και τις τιμές των μεταβλητών εξόδου, και απαιτείται μια διαδικασία μάθησης. Ο αλγόριθμος αυτός, είναι ο γνωστός αλγόριθμος back-propagation (Rumelhart, Hinton, & Williams, 1988). Ο αλγόριθμος είναι μια ειδική έκδοση του γνωστού αλγόριθμου gradient descent που βασίζεται στη μερική παράγωγο της συνάρτησης σφάλματος σε σχέση με τις παραμέτρους του μοντέλου. Ανάλογα με το είδος των νευρώνων εξόδου, θα πρέπει να ορίσουμε μια συνάρτηση σφάλματος. Αν οι νευρώνες εξόδου είναι γραμμικοί, η συνάρτηση είναι το μέσο τετραγωνικό σφάλμα, ενώ στην πιο συνηθισμένη περίπτωση των δίτιμων μεταβλητών, η τυπική συνάρτηση είναι η σχετική εντροπία που είδαμε στο κεφάλαιο 2 (στην πραγματικότητα είναι ισοδύναμη με την πιθανοφάνεια της διωνυμικής κατανομής). Όταν τα έχουμε ορίσει όλα αυτά, ο αλγόριθμος λειτουργεί με τα εξής βήματα:

- στην αρχή γίνεται μια αρχικοποίηση των βαρών με τυχαίες τιμές (συνήθως μέσα σε κάποιο εύρος τιμών που καθορίζεται από τον αριθμό των νευρώνων)
- με βάση τα αρχικά αυτά βάρη, υπολογίζεται το αποτέλεσμα του δικτύου για όλες τις παρατηρήσεις.
- το αποτέλεσμα χρησιμοποιείται για να υπολογιστεί το σφάλμα, η «απόσταση» δηλαδή από τις παρατηρηθείσες τιμές
- αυτό το σφάλμα, είναι η «πιθανοφάνεια» με την στατιστική έννοια, και είναι μια συνάρτηση των βαρών. Άρα τα νέα βάρη θα βρεθούν με τη μέθοδο gradient descent υπολογίζοντας την παράγωγο αυτής της συνάρτησης και κάνοντας τις κατάλληλες τροποποιήσεις
- το «σήμα» αυτό, προωθείται προς τα πίσω στο δίκτυο, τροποποιώντας διαδοχικά τις τιμές όλων των συναπτικών βαρών. Σε κάθε βήμα προς τα πίσω, οι υπολογισμοί καθορίζονται από τις αντίστοιχες συναρτήσεις ενεργοποίησης, ενώ απαιτούνται και αθροίσματα για όλους τους νευρώνες που δίνουν σήμα σε κάποιον άλλον νευρώνα
- όταν το «σήμα» φτάσει ξανά στους νευρώνες εισόδου, ένας κύκλος έχει ολοκληρωθεί, και πλέον όλα τα βάρη του δικτύου έχουν αλλάξει σε μια κατεύθυνση που να μειώνει το συνολικό σφάλμα. Η διαδικασία επαναλαμβάνεται πλέον με τα νέα βάρη (υπολογίζεται νέο σφάλμα κ.ο.κ.) μέχρι το συνολικό σφάλμα να σταματήσει να μειώνεται ή μέχρι να ολοκληρωθεί ένας προκαθορισμένος αριθμός επαναλήψεων

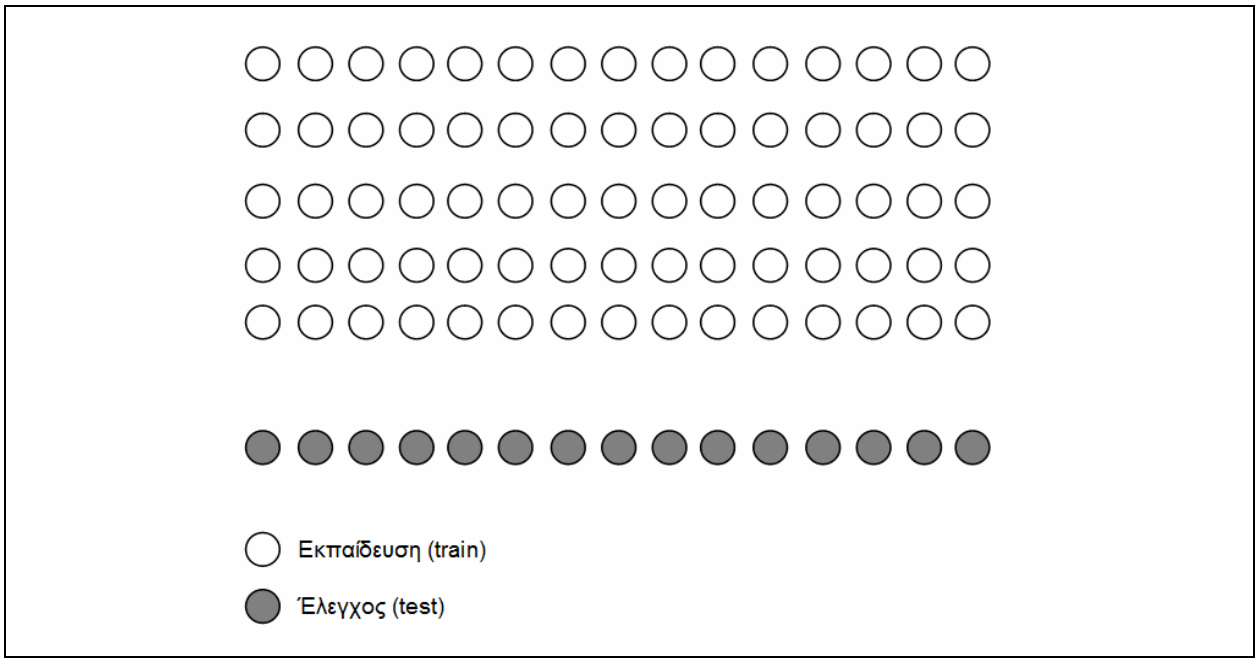
Η μέθοδος αυτή, φυσικά απαιτεί διάφορους υπολογισμούς που παραλείπονται εδώ, αλλά πρέπει να τονιστεί ότι σαν μέθοδος gradient descent, είναι μια ευριστική μέθοδος. Αναμένουμε, αν όλα πάνε καλά, ότι το σφάλμα θα μειώνεται συνεχώς, αλλά δεν υπάρχει μαθηματική εγγύηση. Έχουν αναπτυχθεί επίσης πάρα πολλές παραλλαγές της για να αυξήσουν την πιθανότητα σύγκλισης του αλγορίθμου, αλλά και την ταχύτητα αυτής (π.χ. μέθοδοι που βασίζονται στη δεύτερη παράγωγο της συνάρτησης σφάλματος, κ.ο.κ.). Γενικά, η εκπαίδευση των νευρωνικών δικτύων είναι μια σύνθετη διαδικασία που απαιτεί παρακολούθηση. Ένα μεγάλο πρόβλημα που προκύπτει αφορά κυρίως το μεγάλο αριθμό παραμέτρων (πολλά συναπτικά βάρη) που προκύπτουν τόσο από την κωδικοποίηση των αλληλουχιών όσο και από την αθρόα εισαγωγή μεγάλου αριθμού κρυφών νευρώνων. Για να αντιμετωπιστούν τέτοιου είδους προβλήματα, έχουν προταθεί διάφορες τεχνικές cross-validation (βλ. παρακάτω), ενώ πολλές φορές, λόγω της τυχαιότητας στον αρχικό υπολογισμό των βαρών, αρκετοί ερευνητές προτείνουν τη δημιουργία ικανού αριθμού δικτύων με βάρη που να έχουν ξεκινήσει από διαφορετικές αρχικές τιμές και το τελικό δίκτυο να είναι ένας μέσος όρος των δικτύων αυτών.

Για την εφαρμογή νευρωνικών δικτύων σε προβλήματα βιοπληροφορικής, θα πρέπει καταρχάς να γίνουν οι κατάλληλοι μετασχηματισμοί των αλληλουχιών για να έρθουν στη μορφή που περιγράψαμε πριν. Κατόπιν θα πρέπει να χρησιμοποιηθεί κάποιο γενικό πακέτο για νευρωνικά δίκτυα που διαθέτουν τα γνωστά μαθηματικά πακέτα όπως το **MATLAB** (http://www.mathworks.com/products/neural-network/) ή το **R** (https://cran.r-project.org/web/packages/neuralnet/index.html). Παρ' όλα αυτά, επειδή συνήθως οι εφαρμογές βιοπληροφορικής πρέπει να είναι ανεξάρτητες από την πλατφόρμα, οι περισσότεροι χρησιμοποιούν βιβλιοθήκες για κάποια γενική γλώσσα προγραμματισμού, όπως το **FANN** το οποίο είναι γραμμένο σε C (http://leenissen.dk/fann/wp/) και το **JOONE,** που είναι γραμμένο σε JAVA (http://sourceforge.net/projects/joone/). Επίσης, ιδιαίτερα εύχρηστοι είναι διάφοροι προσομοιωτές (simulators), δηλαδή προγράμματα που υλοποιούν νευρωνικά δίκτυα πολύπλοκης μορφής χωρίς να απαιτείται

από τον χρήστη η ικανότητα προγραμματισμού. Τέτοιες (παλιότερες) προσπάθειες είναι το **BILLNET** (http://www.nongnu.org/billnet/) και το **NevProp** (http://www.cse.unr.edu/brain/nevprop), ενώ το **SNNS** (http://www.ra.cs.uni-tuebingen.de/SNNS/) είναι ίσως το πιο πλήρες πακέτο για το σκοπό αυτό. Τέλος, δεν πρέπει να ξεχνάμε και τις δυνατότητες που δίνουν για χρήση νευρωνικών δικτύων και τα γενικά εργαλεία εξόρυξης γνώσης και μηχανικής μάθησης όπως το **Weka** (http://www.cs.waikato.ac.nz/ml/weka/).

7.3. Μεθοδολογίες για την εκπαίδευση και τον έλεγχο μιας μεθόδου πρόγνωσης

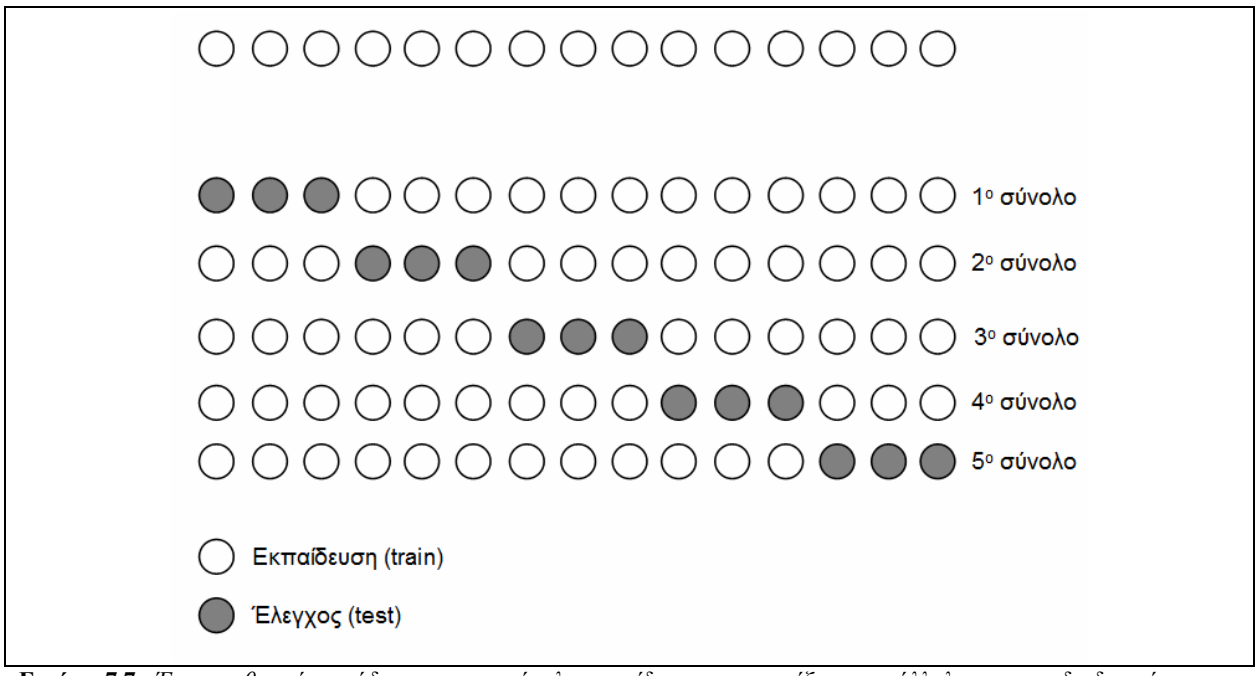
Αρχικά πρέπει να γίνει η συγκέντρωση των δεδομένων εκπαίδευσης και να γίνει αξιολόγησή τους. Αν το πρόβλημα είναι νέο, το σύνολο των δεδομένων εκπαίδευσης θα πρέπει να συλλεχθεί από τις γνωστές βάσεις δεδομένων ή από τη βιβλιογραφία με τις κατάλληλες επερωτήσεις (οι οποίες μπορεί να είναι και ιδιαίτερα δύσκολες). Συνηθισμένη είναι και η περίπτωση κάποιος να χρησιμοποιεί κάποιο σύνολο που είχε χρησιμοποιηθεί παλιότερα. Αυτό έχει νόημα όταν ενδιαφερόμαστε να συγκρίνουμε αμιγώς την επίδραση του νέου αλγορίθμου και να τη διαχωρίσουμε από την επίδραση του συνόλου εκπαίδευσης. Σε όλες τις περιπτώσεις πάντως, πρέπει να έχουμε στο μυαλό μας ότι ακόμα και οι βάσεις δεδομένων περιέχουν λάθη στο σχολιασμό και πολλές φορές τέτοια λάθη μπορεί να έχουν σημαντική επίπτωση στην απόδοση των αλγορίθμων. Επίσης, για διάφορα εξειδικευμένα δομικά και λειτουργικά χαρακτηριστικά, είναι δυνατόν οι βάσεις δεδομένων να μην έχουν την πληροφορία που απαιτείται, οπότε να χρειάζεται αναζήτηση στη βιβλιογραφία. Γι' αυτό το λόγο, πολλές φορές εξειδικευμένες βάσεις δεδομένων κατασκευάζονται από επιστήμονες που ασχολούνται με προγνωστικές μεθόδους. Τα δεδομένα δηλαδή προκύπτουν από τέτοιες αναζητήσεις για τις ανάγκες της μεθόδου πρόγνωσης και κατόπιν δημιουργείται η βάση δεδομένων για να μπορέσει να χρησιμοποιηθεί και από άλλους.



Εικόνα 7.6: *Ένα υποθετικό παράδειγμα με το σύνολο εκπαίδευσης και το ανεξάρτητο σύνολο ελέγχου.*

Γενικά, το σύνολο εκπαίδευσης πρέπει να είναι όσο το δυνατόν πιο αντιπροσωπευτικό γίνεται, αλλά δεν υπάρχουν ξεκάθαροι κανόνες. Επίσης, θα πρέπει να υπάρχουν και κανόνες όσον αφορά το πόσο όμοιες αλληλουχίες περιέχει (να είναι όπως λέμε non-redundant set). Το ποσοστό ομοιότητας όμως που θεωρείται αποδεκτό εξαρτάται από τη φύση του επιμέρους προβλήματος (π.χ. στα προβλήματα δευτεροταγούς δομής η αποδεκτή ομοιότητα είναι στο 30%, ενώ σε άλλες περιπτώσεις όπως στις σηματοδοτικές αλληλουχίες, υπάρχουν άλλα κριτήρια). Τέλος, υπάρχει και περίπτωση το σύνολο εκπαίδευσης να περιέχει αρκετές ομόλογες πρωτεΐνες, αλλά τότε απαιτείται η αξιολόγηση της αποδοτικότητας του αλγορίθμου να γίνει σε ανεξάρτητο σύνολο δεδομένων, οι πρωτεΐνες του οποίου δεν θα έχουν ομοιότητα με αυτές του συνόλου εκπαίδευσης (βλ. παρακάτω).

Το επόμενο βήμα και ίσως το πιο δύσκολο, είναι ο σχεδιασμός του αλγόριθμου. Στο στάδιο αυτό απαιτούνται τόσο ειδικές γνώσεις για τη βιολογική φύση του προβλήματος (τι χαρακτηριστικό είναι αυτό που ψάχνουμε πάνω στις αλληλουχίες), όσο και για τις υπολογιστικές και μαθηματικές τεχνικές που θα χρησιμοποιηθούν για την επίλυσή του. Τα παραπάνω ισχύουν προφανώς τόσο για τις μεθόδους τοπικής πρόβλεψης (η επιλογή του παραθύρου, η κωδικοποίηση των αλληλουχιών, ο αλγόριθμος που θα χρησιμοποιηθεί) όσο και για τις μεθόδους ολικής ταξινόμησης των πρωτεϊνών (όπου πρέπει να γίνει επιλογή των χαρακτηριστικών με τα οποία θα κωδικοποιηθούν οι αλληλουχίες, αλλά και του αλγόριθμου ταξινόμησης). Το στάδιο αυτό, απαιτεί εκτός από εξειδικευμένες γνώσεις και αρκετή φαντασία, καθώς η διαδικασία μοντελοποίησης (γιατί περί αυτού πρόκειται) είναι μια αρκετά δύσκολη διαδικασία χωρίς ξεκάθαρους κανόνες. Φυσικά, η ίδια η επιλογή της μεθοδολογίας και η υλοποίηση του αλγόριθμου που θα χρησιμοποιηθεί απαιτεί πολλές φορές εξειδικευμένες γνώσεις μαθηματικών, στατιστικής, μηχανικής μάθησης, τεχνολογίας λογισμικού, τεχνολογίας διαδικτύου (όταν πρόκειται να κατασκευαστεί διαδικτυακή εφαρμογή), κ.ο.κ.



Εικόνα 7.7: *Ένα υποθετικό παράδειγμα με το σύνολο εκπαίδευσης να χωρίζεται κατάλληλα για μια διαδικασία cross-validation.*

Τέλος, ένα πολύ κρίσιμο σημείο στην όλη διαδικασία κατασκευής μιας μεθόδου πρόγνωσης είναι η σωστή αξιολόγησή της. Ανάλογα με τη μέθοδο, θα επιλέξουμε και τα κατάλληλα στατιστικά μέτρα (βλ. παρακάτω) αλλά αυτό δεν αρκεί. Μια οποιαδήποτε μέθοδος είναι δυνατόν αν εφαρμοστεί στα ίδια τα δεδομένα με τα οποία έχει εκπαιδευτεί, να δώσει υπερβολικά καλά αποτελέσματα. Αυτός ο έλεγχος ονομάζεται έλεγχος αυτο-συνέπειας (self-consistency) αλλά είναι πολύ πιθανό να δώσει μεροληπτικά αποτελέσματα καθώς υπάρχει ο κίνδυνος υπερ-προσαρμογής (over-fitting). Με τον τελευταίο όρο εννοούμε ότι μπορεί η μέθοδος να έχει «εκπαιδευτεί» παραπάνω από όσο χρειάζεται, με αποτέλεσμα να αποδίδει πολύ καλά στο σύνολο εκπαίδευσης αλλά να αποτυγχάνει σε νέα παραδείγματα. Αυτό το φαινόμενο είναι πιο πιθανό να συμβεί όσο πιο εξελιγμένη είναι μια μέθοδος αλγοριθμικά, καθώς οι μεθοδολογίες μηχανικής μάθησης (όπως π.χ. τα νευρωνικά δίκτυα) έχουν μεγάλο αριθμό παραμέτρων. Σε κάθε περίπτωση παντως, ιδανικά μια μέθοδος πρέπει να αποδειχτεί ότι αποδίδει αρκετά καλά σε ένα ανεξάρτητο έλεγχο (independent test) για να έχουμε όσο το δυνατό πιο αμερόληπτα αποτελέσματα (Εικόνα 7.6). Η κατασκευή του ανεξάρτητου συνόλου ελέγχου είναι μια επίσης δύσκολη διαδικασία, αφενός μεν γιατί υπάρχει περίπτωση τα δεδομένα να μην είναι επαρκή, αφετέρου δε γιατί πρέπει οπωσδήποτε οι πρωτεΐνες του ανεξάρτητου συνόλου να είναι όντως «ανεξάρτητες», διαφορετικές δηλαδή από αυτές του συνόλου εκπαίδευσης. Το τι εννοούμε «διαφορετικές» βέβαια, εξαρτάται πάρα πολύ από το πρόβλημα, αλλά μια καλή αρχή στις περισσότερες περιπτώσεις είναι να στηριζόμαστε στα αποδεκτά επίπεδα ομοιότητας σε επίπεδο αλληλουχιών (π.χ. στα

περισσότερα προβλήματα δομής, ακολουθούμε τον κανόνα του 30% ομοιότητα). Φυσικά, για επιμέρους ειδικά προβλήματα τα κριτήρια μπορεί να είναι λιγότερο ή περισσότερο αυστηρά.

Ένας άλλος έλεγχος, ο οποίος είτε γίνεται λόγω ανάγκης εξαιτίας της έλλειψης ανεξάρτητου συνόλου, είτε γίνεται ως ένας επιπλέον έλεγχος λόγω του ότι το ανεξάρτητο σύνολο είναι μικρό, είναι ο λεγόμενος έλεγχος cross-validation (Εικόνα 7.7). Με τη διαδικασία αυτή, το σύνολο εκπαίδευσης χωρίζεται σε *k* υποσύνολα (*k*-fold cross-validation). Έπειτα, ένα υποσύνολο κάθε φορά αφαιρείται από το σύνολο εκπαίδευσης, η εκπαίδευση πραγματοποιείται με τα εναπομείναντα υποσύνολα και κατόπιν η μέθοδος δοκιμάζεται στις ακολουθίες του υποσυνόλου το οποίο έχει αφαιρεθεί. Η διαδικασία επαναλαμβάνεται *k* φορές και το τελικό αποτέλεσμα προσφέρει μια αμερόληπτη (unbiased) εκτίμηση για την πραγματική επιτυχία της μεθόδου, καθώς τα αποτελέσματα έχουν προκύψει χωρίς καμία αλληλουχία να έχει χρησιμοποιηθεί στην κατασκευή της μεθόδου με την οποία έγινε η πρόβλεψη πάνω της. Φυσικά, αυτό εισάγει τον επιπλέον περιορισμό ότι μεταξύ των πρωτεϊνών του συνόλου εκπαίδευσης δεν υπάρχουν ανιχνεύσιμες ομοιότητες (με όποιο κριτήριο και αν έχουμε επιλέξει) ή τουλάχιστον δεν υπάρχουν τέτοιες ομοιότητες μεταξύ των *k* υποσυνόλων. Μια παραλλαγή αυτής της μεθόδου, η οποία είναι πιο αξιόπιστη στατιστικά αλλά απαιτεί πολλούς περισσότερους υπολογισμούς, είναι η λεγόμενη Jackknife κατά την οποία το *k* επιλέγεται να είναι ίσο με το μέγεθος του συνόλου εκπαίδευσης, με συνέπεια το κάθε υποσύνολο να έχει μέγεθος ίσο με ένα. Γενικά, σε σύνολα με μέτριο μέγεθος ή για μεθόδους που είναι γρήγορες, το Jackknife είναι προτιμότερο, γιατί κάθε φορά το σύνολο εκπαίδευσης είναι όσο μεγαλύτερο γίνεται. Αν όμως η μέθοδος είναι αργή ή αν το σύνολο είναι πολύ μεγάλο ή αν δεν υπάρχει εύκολος τρόπος να εξασφαλιστούν οι συνθήκες ομοιότητας, τότε η μέθοδος δεν μπορεί να εφαρμοστεί (και αν εφαρμοστεί θα δώσει επίσης μεροληπτικά αποτελέσματα).

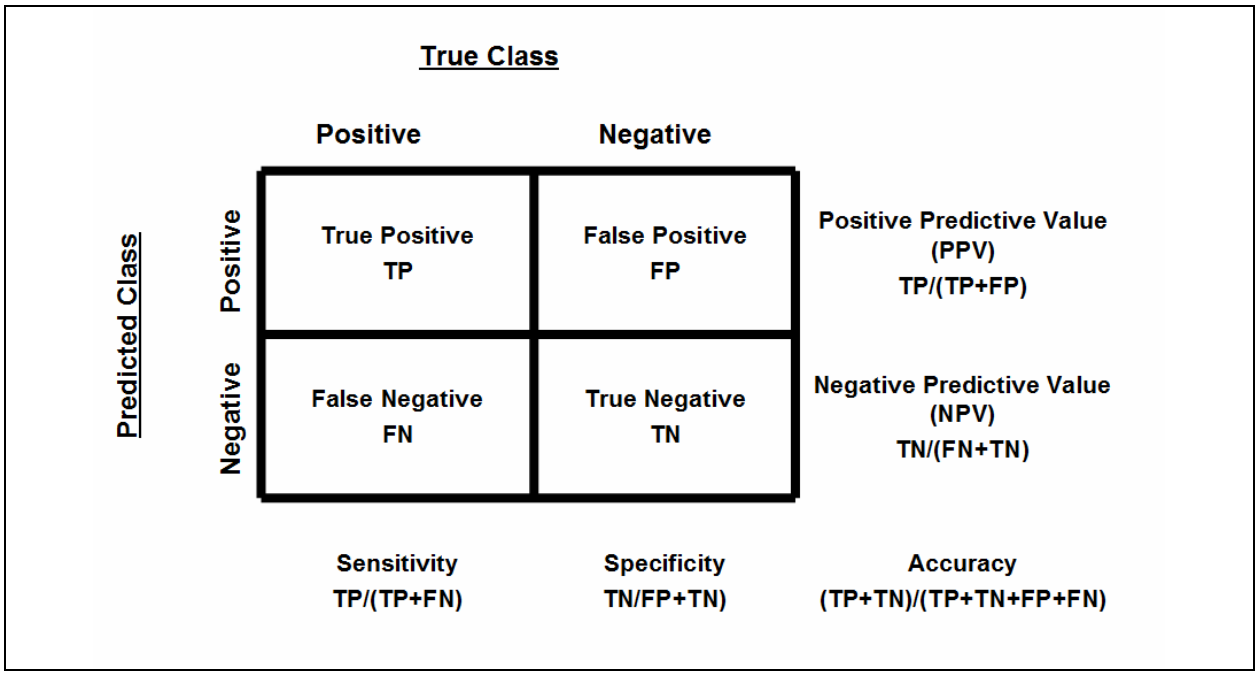
7.4. Μέτρα εκτίμησης της αξιοπιστίας των μεθόδων

Για να μπορέσουμε να μετρήσουμε την επιτυχία και την αξιοπιστία των προγνώσεων που προέρχονται από μια μέθοδο, έχουν προταθεί διάφορα μέτρα. Τα περισσότερα από αυτά, ισχύουν τόσο για την περίπτωση της τοπικής πρόγνωσης (per-residue prediction) όσο και για την κατάταξη αλληλουχιών σε κατηγορίες (per-protein classification). Αν θεωρήσουμε μια πρόγνωση για δύο κατηγορίες, τότε τα δεδομένα μπορούν να αναπαρασταθούν σε έναν πίνακα συνάφειας 2x2 (Εικόνα). Έτσι, συμβολίζουμε με TP (True Positives) τον αριθμό των ορθώς θετικά προσδιορισμένων καταλοίπων, TN (True Negatives) τον αριθμό των ορθώς αρνητικά προσδιορισμένων καταλοίπων, FN (False Negatives) τον αριθμό των εσφαλμένων αρνητικά προσδιορισμένων καταλοίπων και FP (False Positives) τον αριθμό των εσφαλμένων θετικά προσδιορισμένων καταλοίπων. Προφανώς, όταν μιλάμε για κατάταξη αλληλουχιών, οι παρατηρήσεις πλέον δεν είναι τα κατάλοιπα αλλά ολόκληρες οι πρωτεΐνες.

Από τον πίνακα αυτό, το πιο προφανές μέτρο αξιολόγησης είναι το συνολικό ποσοστό των καταλοίπων που έχουν προβλεφθεί σωστά (*Q*), με την πρόγνωση να έχει αναχθεί σε δυο κατηγορίες:

$$Q = \frac{TP + TN}{TP + TN + FP + FN} 100\%$$

Ανάλογα με την περίπτωση, είναι δυνατό να μας απασχολεί περισσότερο η ευαισθησία (sensitivity) της μεθόδου, η οποία μετράει το ποσοστό των σωστών θετικών προβλέψεων (δηλαδή πόσες παρατηρήσεις που άνηκαν στην ομάδα ενδιαφέροντος προβλέφθηκαν σωστά), αλλά και η ειδικότητα (specificity) που συνοψίζει το ποσοστό των σωστών αρνητικών προβλέψεων (δηλαδή το πόσες παρατηρήσεις που δεν άνηκαν στην ομάδα προβλέφθηκαν σωστά). Μπορούμε εύκολα να φανταστούμε περιπτώσεις μεθόδων με καλή ευαισθησία αλλά όχι καλή ειδικότητα, και αντίστροφα, ενώ σε κάποια προβλήματα μπορεί να μας ενδιαφέρει εξαρχής η καλή ευαισθησία και σε άλλα η καλή ειδικότητα. Επιπλέον δε, ανάλογα με το πόσο σπάνια είναι η μία από τις δύο ομάδες, μπορεί να υπάρχουν περιπτώσεις μεθόδων οι οποίες να έχουν ονομαστικά καλή ειδικότητα και ευαισθησία, αλλά να μην αποδίδουν καλά στην πράξη. Αυτό συμβαίνει, γιατί αν για παράδειγμα η μία ομάδα είναι πολύ σπάνια (πχ 5%), τότε ακόμα και μια ευαισθησία και ειδικότητα της τάξης του 95%, θα δώσει πολύ χαμηλή θετική και αρνητική προγνωστική αξία. Τα μέτρα αυτά εκφράζουν την πιθανότητα, μια θετική ή μια αρνητική πρόγνωση αντίστοιχα, να είναι σωστές, και πολλές φορές σε πραγματικά προβλήματα είναι και αυτά παράγοντες που πρέπει να λαμβάνουμε υπόψη μας.



Εικόνα 7.8: *Τα μέτρα που προκύπτουν από έναν δίπτυχο πίνακα ταξινόμησης. Τα μέτρα αυτά μπορεί να εφαρμοστούν τόσο σε επίπεδο αμινοξικών καταλοίπων (ή βάσεων), αλλά και σε επίπεδο αλληλουχιών (Vihinen, 2012)*

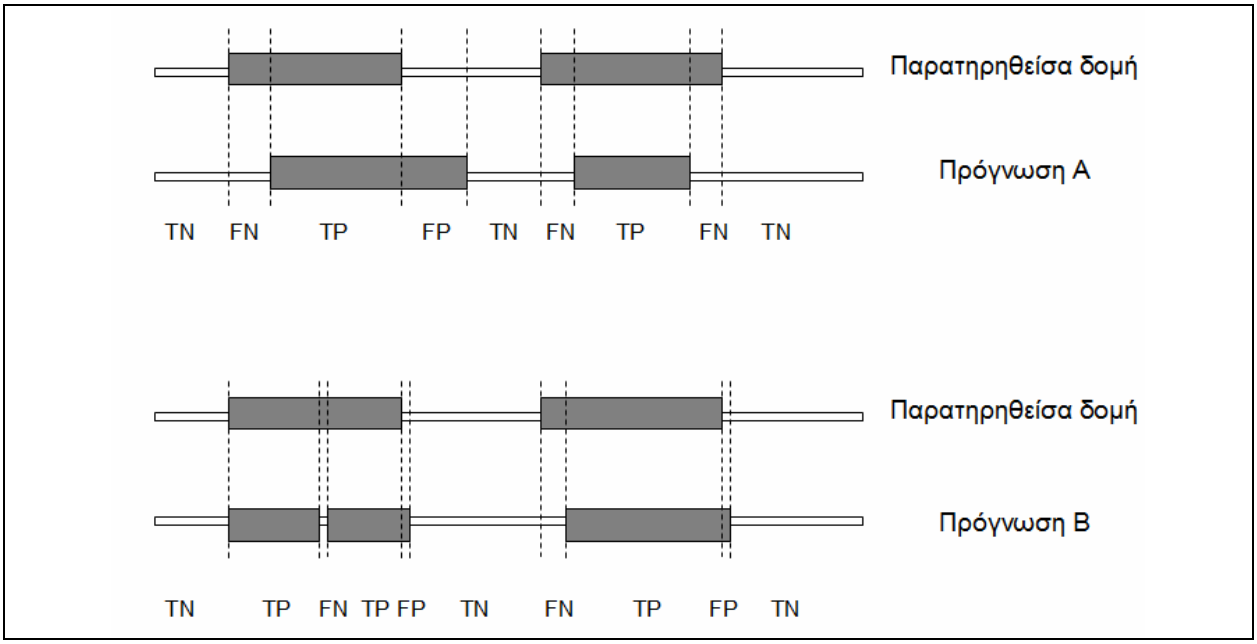
Επίσης, ένα άλλο μέτρο που χρησιμοποιείται είναι ο γνωστός συντελεστής συσχέτισης του Matthews (*C*) (Baldi, Brunak, Chauvin, Andersen, & Nielsen, 2000):

$$C = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

Ο συντελεστής αυτός, είναι ισοδύναμος του γνωστού συντελεστή συσχέτισης του Pearson όταν εφαρμοστεί σε δίτιμα δεδομένα και παίρνει τιμές από το -1 (τελείως αντίθετη πρόγνωση), έως το +1 (τέλεια πρόγνωση), με το 0 να αντιστοιχεί στην τελείως τυχαία πρόγνωση. Το μεγάλο πλεονέκτημα του συντελεστή συσχέτισης είναι ότι συνδυάζει όλες τις τιμές του πίνακα σε μία αριθμητική τιμή.

Βέβαια, όπως αναφέραμε ήδη, τα μέτρα αυτά είναι κατάλληλα για διαχωρισμούς σε δύο κλάσεις. Όταν το πρόβλημα με το οποίο ασχολούμαστε είναι πρόβλημα πολλών κλάσεων (*k*), συνήθως εφαρμόζουμε το *Q* στον αντίστοιχο *k*x*k* πίνακα αλλά το *C* θα πρέπει να υπολογιστεί ξεχωριστά για κάθε ομάδα, αγνοώντας τις υπόλοιπες. Για παράδειγμα στην περίπτωση πρόγνωσης δευτεροταγούς δομής (όπου οι κλάσεις είναι H, E και C), μπορούμε να υπολογίσουμε ένα συνολικό *Q* (όπως φυσικά και τα αντίστοιχα *Qa*, *Qb* κλπ) αλλά για το *C* θα πρέπει να υπολογίσουμε τις επιμέρους τιμές αγνοώντας τις άλλες ομάδες (*Ca*, *Cb*).

Σε περιπτώσεις τοπικών προγνώσεων, όπου και ενδιαφερόμαστε για την πρόβλεψη συγκεκριμένων περιοχών κατά μήκος της αλληλουχίας, είναι δυνατόν τα παραπάνω μέτρα να είναι παραπλανητικά. Για παράδειγμα, στα προβλήματα πρόβλεψης δευτεροταγούς δομής ή διαμεμβρανικών τμημάτων, είναι δυνατόν να έχεις μια μέθοδο με καλύτερα ανά κατάλοιπο μέτρα (TP, TN, Q, C) σε σχέση με μια άλλη μέθοδο, αλλά η δεύτερη μέθοδος να είναι καλύτερη. Αυτό μπορεί να συμβεί αν εμφανίζονται κατακερματισμένες προγνώσεις, π.χ. μια ξεχωριστή περιοχή να προβλέπεται ως δυο διαφορετικές περιοχές ή δυο γειτονικές περιοχές να προβλέπονται ως μία. Για όλα τα παραπάνω, έχει προταθεί σαν πιο αξιόπιστη λύση, η χρήση του μέτρου επικάλυψης των τμημάτων (measure of the segment's overlap-*SOV*), το οποίο θεωρείται ο πιο αξιόπιστος δείκτης της προγνωστικής ικανότητας των αλγορίθμων πρόγνωσης δευτεροταγούς δομής, και παίρνει συνεχείς τιμές στο διάστημα 0-1 (Zemla, Venclovas, Fidelis, & Rost, 1999).



Εικόνα 7.9: *Ένα υποθετικό παράδειγμα της σημασίας του μέτρου SOV. Κάτω, βλέπουμε μια περίπτωση στην οποία, η πρόγνωση δεν είναι καλή, γιατί η πρώτη περιοχή έχει προβλεφθεί σαν δύο διαφορετικές, παρ' όλα αυτά τα μέτρα που εστιάζουν στα κατάλοιπα δίνουν πολύ καλές τιμές. Αντίθετα, στην πάνω εικόνα, παρόλο που τα μέτρα για τα κατάλοιπα είναι χειρότερα, η πρόγνωση γενικά είναι καλύτερη και αυτό απεικονίζεται και στο SOV.*

7.5. Τρόποι βελτίωσης της απόδοσης των μεθόδων πρόγνωσης

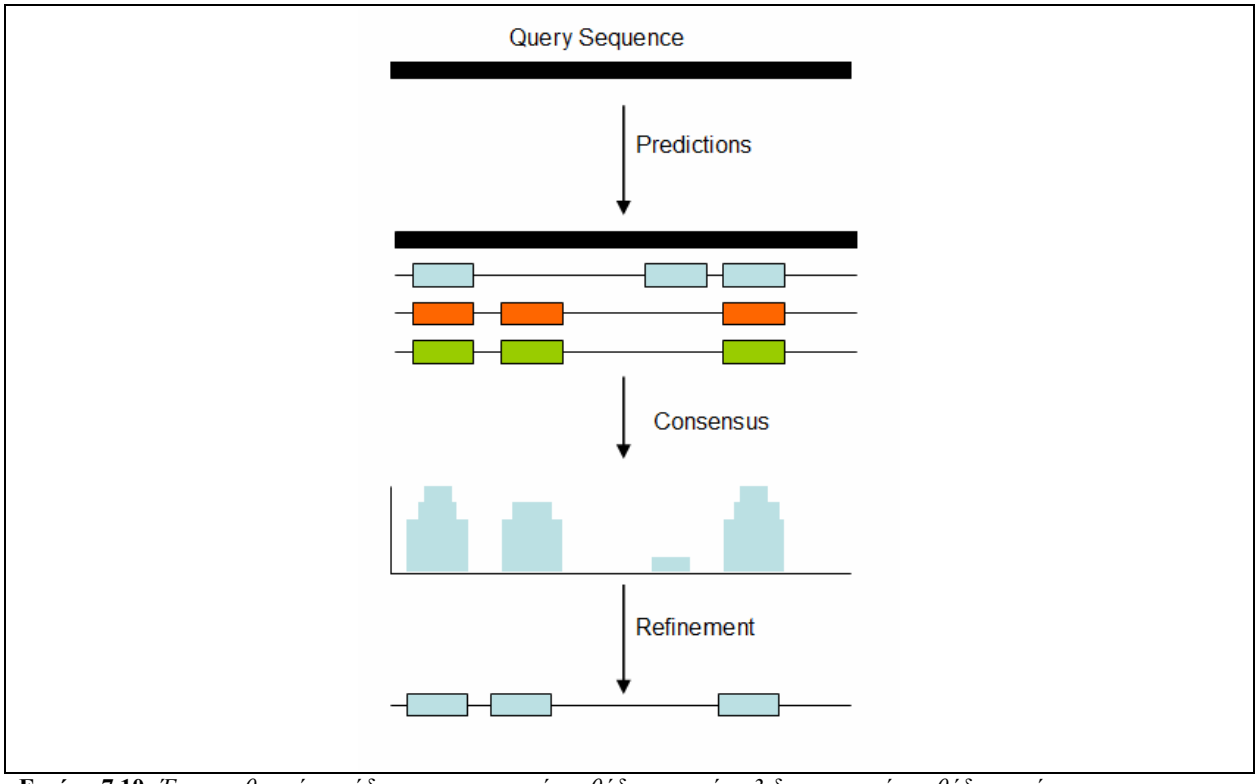
Γενικά, η επιτυχία μιας μεθόδου πρόγνωσης για ένα συγκεκριμένο πάντα πρόβλημα εξαρτάται από το μέγεθος και την ποιότητα του συνόλου εκπαίδευσης και από την επιλογή του αλγορίθμου, δηλαδή της μεθοδολογίας. Το μέγεθος του συνόλου εκπαίδευσης παίζει σίγουρα ένα ρόλο, αλλά η επίδραση δεν είναι γραμμική όπως έχει φανεί από εμπειρικές μελέτες καθώς ενώ υπάρχει γενικά μια αυξητική τάση, από ένα σημείο και μετά δεν μπορούμε να πετύχουμε περαιτέρω αύξηση της απόδοσης. Επίσης, το είδος του αλγορίθμου παίζει ρόλο και στο πώς επηρεάζει το μέγεθος του συνόλου εκπαίδευσης την απόδοση, καθώς οι απλές μέθοδοι έχουν μικρό αριθμό παραμέτρων με συνέπεια να φτάνουν γρήγορα στο σημείο κορεσμού (πλατό), ενώ οι πιο σύνθετες μέθοδοι οι οποίες έχουν μεγαλύτερο αριθμό παραμέτρων απαιτούν και περισσότερα δεδομένα.

Εκτός από αυτά πάντως, υπάρχουν δύο γενικές μεθοδολογίες οι οποίες μπορούν να αυξήσουν σημαντικά την απόδοση οποιασδήποτε μεθόδου πρόγνωσης, και αξίζει να αναφερθούν. Η πρώτη μεθοδολογία είναι οι συναινετικές ή συνδυαστικές μέθοδοι, ενώ η δεύτερη η χρήση εξελικτικής πληροφορίας με τη μορφή πολλαπλών στοιχίσεων.

7.5.1. Συνδυαστικές μέθοδοι

Μια συνδυαστική/συναινετική μέθοδος πρόγνωσης, βασίζεται στην βασική απλή ιδέα, ότι αν συνδυαστούν ανεξάρτητες μέθοδοι το αποτέλεσμα είναι πάντα καλύτερο. Οι μεθοδολογίες αυτές έχουν χρησιμοποιηθεί σε διάφορους τομείς, είτε με απλό τρόπο (majority vote, consensus) είτε με πιο σύνθετους αλγόριθμους μηχανικής μάθησης (ensemble learning, meta-algorithms κ.ο.κ.). Η απλή αυτή διαίσθηση («ας ακούσουμε πολλές γνώμες»), έχει επίσης βρει και τη μαθηματική της τεκμηρίωση καθώς υπάρχουν θεωρητικές αποδείξεις ότι ο συνδυασμός «ασθενών ταξινομητών» (weak classifiers), δηλαδή ταξινομητών οι οποίοι αποδίδουν μεν καλύτερα από το τυχαίο (π.χ. συντελεστής συσχέτισης >0, ή Q>0.5), δίνει πάντα έναν ταξινομητή με καλύτερη αποτελεσματικότητα. Φυσικά, είναι προφανές ότι αν κάποια από τις μεθόδους είναι ιδιαίτερα καλή (π.χ. συντελεστής συσχέτισης >0.95 ή Q>0.99), τότε η μέθοδος δεν θα δουλέψει καθώς η «ισχυρή» μέθοδος θα υπερισχύει πάντα.

Παρακάτω, θα περιγράψουμε πώς λειτουργεί μια τέτοια μέθοδος στα προβλήματα τοπικής πρόγνωσης με δύο κατηγορίες, αλλά φυσικά με τον ίδιο (αν και πιο απλό) τρόπο δουλεύει και για τα προβλήματα ολικής ταξινόμησης. Επίσης, η βασική ιδέα είναι η ίδια και όταν υπάρχουν επιπλέον κατηγορίες (όπως στην περίπτωση της δευτεροταγούς δομής), με τη μόνη διαφορά ότι τότε η ίδια διαδικασία θα πρέπει να επαναληφθεί για την κάθε κατηγορία. Η βασική ιδέα, φαίνεται διαγραμματικά στην Εικόνα 7.10. Έχουμε κάποιες μεθόδους πρόγνωσης, τις οποίες προς το παρόν αντιμετωπίζουμε ως «μαύρα κουτιά», δεν μας ενδιαφέρει δηλαδή πώς λειτουργούν και με ποιον τρόπο. Απλά δίνουμε μια ακολουθία ως δεδομένο εισόδου και παίρνουμε μια πρόγνωση σαν αποτέλεσμα. Αυτή η θεώρηση, είναι όπως θα δούμε αρκετά βολική γιατί μας επιτρέπει να χρησιμοποιήσουμε τη μέθοδο με οποιεσδήποτε μεθόδους πρόγνωσης, χωρίς να έχουμε γνώση του τρόπου με τον οποίον λειτουργούν. Εφαρμόζουμε, στη συνέχεια, την κάθε μέθοδο ξεχωριστά στην ίδια ακολουθία εισόδου. Και για κάθε θέση πάνω στην αλληλουχία, δημιουργούμε ένα σκορ καταμετρώντας πόσες από τις μεθόδους προβλέπουν τη μία κατηγορία και πόσες την άλλη. Το σκορ αυτό συνήθως είναι κανονικοποιημένο για τον αριθμό των μεθόδων, και έτσι παίρνουμε τελικά μια τιμή από το 0 μέχρι το 1 (αλλά αυτό δεν είναι και απαραίτητο). Κατόπιν, μπορούμε επιλέγοντας ένα κατώφλι αξιοπιστίας, να θεωρήσουμε ότι η συνδυαστική μέθοδος αποδίδει μια πρόγνωση όταν η τιμή σε μία θέση είναι μεγαλύτερη από μια τιμή c ($0<c<1$). Η τιμή αυτή, αντιστοιχεί στο αποδεκτό επίπεδο «πλειοψηφίας», εξαρτάται από το είδος του προβλήματος και τη φύση των μεθόδων που χρησιμοποιούνται, και ως εκ τούτου η εύρεσή της αποτελεί αντικείμενο εμπειρικής αξιολόγησης.



Εικόνα 7.10: *Ένα υποθετικό παράδειγμα συναινετικής μεθόδου με χρήση 3 διαφορετικών μεθόδων πρόγνωσης.*

Φυσικά, με τον τρόπο που περιγράφηκε παραπάνω, η μέθοδος είναι αρκετά απλή και υπάρχουν διάφορες επιπλέον παραλλαγές οι οποίες μπορούν να βελτιώσουν την απόδοση. Για παράδειγμα, είναι δυνατό η κάθε μέθοδος να μη συνεισφέρει το ίδιο στο σκορ αλλά να εισαχθούν βάρη που να αντιστοιχούν στην αξιοπιστία της κάθε μεθόδου. Επίσης, είναι δυνατόν διαφορετικοί συνδυασμοί των μεθόδων να δίνουν διαφορετικό αποτέλεσμα (π.χ. όταν η μέθοδος Α και η μέθοδος Β συμφωνούν, τότε αυτό σημαίνει ότι η πρόγνωση είναι σωστή ανεξαρτήτως του τι λένε οι άλλες μέθοδοι). Τέτοιες μεθοδολογίες μπορούν να υλοποιηθούν με τις μεθόδους ensemble learning, και μπορεί να βελτιώσουν θεαματικά την απόδοση. Το μεγάλο μειονέκτημα βέβαια, είναι ότι καθώς απαιτείται εκπαίδευση και έλεγχος για την εύρεση της βέλτιστης τιμής των παραμέτρων, απαιτείται ξεχωριστό σύνολο εκπαίδευσης και ελέγχου για τη νέα συνδυαστική

μέθοδο. Αντίθετα, η απλή συναινετική μέθοδος όπως περιγράφηκε στην προηγούμενη παράγραφο, μπορεί να λειτουργήσει χωρίς αυτή τη διαδικασία, καθώς απαιτείται μόνο η τιμή του κατωφλίου c, το οποίο μπορεί να τεθεί σε μια λογικοφανή τιμή (πχ 0.8).

Τέλος, ένα επιπλέον πρόβλημα μπορεί να προκύψει όταν η τελική πρόγνωση απαιτεί βελτιστοποίηση (refinement). Σε κάποιες περιπτώσεις αυτό δεν απαιτείται, αλλά στα περισσότερα προβλήματα αυτό είναι απαραίτητο είτε λόγω της ύπαρξης πολλών κατηγοριών, είτε κυρίως λόγω της ανάγκης η τελική πρόγνωση να υπακούει σε κάποιους κανόνες (π.χ. το μέγεθος των περιοχών να είναι μέσα σε κάποια όρια όσον αφορά το μήκος). Όπως είναι φανερό, ακόμα και αν οι επιμέρους μέθοδοι που χρησιμοποιούνται παράγουν αποτελέσματα με όρια περιοχών «τυποποιημένα» (δηλαδή, μέσα στα εκάστοτε αποδεκτά όρια), η συνδυαστική μέθοδος εκ των πραγμάτων δεν θα δεσμεύεται από αυτές τις ρυθμίσεις. Σε αυτές τις περιπτώσεις, χρειάζεται ένα επιπλέον βήμα για την τυποποίηση και τον περιορισμό των προβλέψεων. Αυτό μπορεί να γίνει είτε με εισαγωγή *ad-hoc* κανόνων ή (κατά προτίμηση) με την εφαρμογή ενός επιπλέον φίλτρου με κάποιον αλγόριθμο δυναμικού προγραμματισμού για να επιβάλει τους περιορισμούς. Η πρακτική αυτή μπορεί να έχει το μειονέκτημα των επιπλέον υπολογιστικών απαιτήσεων, αλλά στις περισσότερες περιπτώσεις αυξάνει την απόδοση της συνδυαστικής μεθόδου θεαματικά.

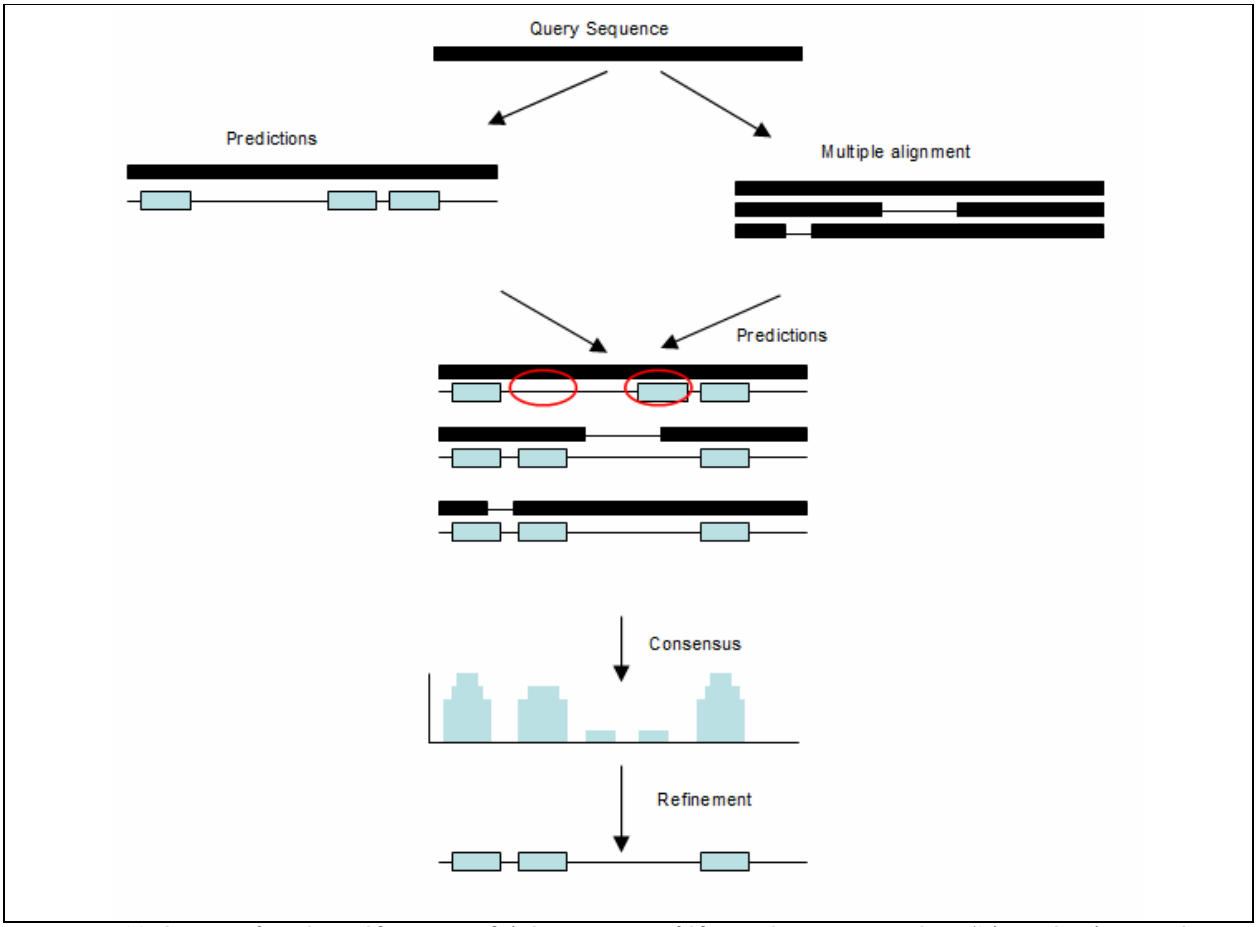
7.5.2. Ενσωμάτωση εξελικτικής πληροφορίας με τη μορφή πολλαπλών στοιχίσεων

Η μέθοδος αυτή βασίζεται στην εξής απλή και γνωστή παρατήρηση, ότι οι πρωτεϊνικές δομές είναι πιο συντηρημένες από τις αλληλουχίες. Με άλλα λόγια, σε μία πολλαπλή στοίχιση ομόλογων πρωτεϊνών αναμένουμε ότι η τρισδιάστατη δομή θα είναι παρόμοια, ακόμα και αν οι επιμέρους αλληλουχίες διαφέρουν. Η μέθοδος της ενσωμάτωσης εξελικτικής πληροφορίας με τη μορφή πολλαπλών στοιχίσεων εκμεταλλεύεται ακριβώς αυτό. Στην πιο απλή της μορφή, η μέθοδος συνίσταται στην εύρεση των ομόλογων πρωτεϊνών της υπό μελέτης αλληλουχίας και την κατασκευή της πολλαπλής στοίχισης. Κατόπιν, με την ίδια μέθοδο πραγματοποιούνται προγνώσεις σε όλες τις αλληλουχίες της πρωτεϊνικής οικογένειας που έχουν εντοπιστεί και οι προγνώσεις αυτές «προβάλλονται» πάνω στην πολλαπλή στοίχιση και κατ' επέκταση στην αρχική αλληλουχία επερώτησης (δηλαδή, σε αυτή στην οποία ενδιαφερόμαστε να πραγματοποιήσουμε την πρόγνωση). Η διαγραμματική αναπαράσταση της μεθόδου, φαίνεται στην Εικόνα 7.11.

Το κλειδί στην κατανόηση της μεθόδου αυτής, βρίσκεται στο γεγονός ότι είναι δυνατό σε μια συγκεκριμένη αλληλουχία, σε ένα δεδομένο σημείο, λόγω της μεταβλητότητας των αμινοξικών αλληλουχιών να υπάρχουν αμινοξέα που «ευνοούν» μια λάθος πρόγνωση. Αφού όμως αναμένουμε ότι τα υπόλοιπα μέλη της οικογένειας μοιράζονται παρόμοια δομή, είναι λογικό να υποθέσουμε, ότι στη δεδομένη θέση της πολλαπλής στοίχισης, η μέθοδος πρόγνωσης θα έχει δώσει διαφορετικό αποτέλεσμα για την πλειοψηφία των αλληλουχιών. Με άλλα λόγια, αντί να στηρίξουμε την πρόγνωση μας σε μια δεδομένη αλληλουχία, η οποία μπορεί να είναι και ειδική περίπτωση, είναι καλύτερο να χρησιμοποιήσουμε για την πρόγνωση την πληροφορία από ολόκληρη την πολλαπλή στοίχιση της οικογένειας.

Υπάρχουν πολλές παραλλαγές αυτής της μεθόδου, που κυρίως έχουν να κάνουν με την επιλογή αλγόριθμου για την εύρεση των ομόλογων αλλά και για την κατασκευή της πολλαπλής στοίχισης. Γενικά, όλες οι επιλογές είναι θεμιτές αλλά μια εύκολη και πρακτική λύση είναι ο συνδυασμός BLAST και CLUSTAL, ενώ σε περιπτώσεις διαδικτυακών εφαρμογών που απαιτούν πολλές στοιχίσεις ίσως η επιλογή του KALIGN να είναι πιο συμφέρουσα. Επίσης, τα τελευταία χρόνια με την εμφάνιση του HMMER 3.0, η εφαρμογή των προφίλ HMM γίνεται μια ελκυστική εναλλακτική.

Η μέθοδος αυτή, είναι πολύ απλή, διαισθητικά σωστή και αποτελεσματική καθώς έχει δειχθεί ότι σε γενικά προβλήματα πρόγνωσης δομής είναι δυνατό να αυξήσει την αποτελεσματικότητα μιας οποιασδήποτε μεθόδου πρόγνωσης κατά περίπου 6-8%. Το βασικό πλεονέκτημά της είναι ότι καθώς αντιμετωπίζει τη μέθοδο πρόγνωσης ως «μαύρο κουτί», είναι δυνατό να εφαρμοστεί με οποιαδήποτε μέθοδο πρόγνωσης ανεξαρτήτως του πώς λειτουργεί. Επίσης, απαιτεί μόνο τη χρήση γνωστών εργαλείων (αναζήτησης ομοιότητας και πολλαπλών στοιχίσεων). Ένα βασικό μειονέκτημα είναι το γεγονός ότι έχει αυξημένες υπολογιστικές απαιτήσεις, κυρίως γιατί απαιτεί την εφαρμογή της μεθόδου πρόγνωσης σε όλες τις πρωτεΐνες της πολλαπλής στοίχισης. Τέλος, ένα άλλο μειονέκτημα είναι κοινό με τη μέθοδο συναινετικής πρόγνωσης. Συγκεκριμένα, ανεξάρτητα με το αν η μέθοδος πρόγνωσης θέτει όρια και περιορισμούς στις περιοχές που προβλέπει, η πρόγνωση που θα προκύπτει από την πολλαπλή στοίχιση δεν είναι σίγουρο ότι θα ακολουθεί τους ίδιους κανόνες. Κατά συνέπεια, χρειάζεται και εδώ το επιπλέον βήμα για το φιλτράρισμα και την εκ των υστέρων επεξεργασία των προγνώσεων.



Εικόνα 7.11: *Ένα υποθετικό παράδειγμα της βελτίωσης μιας μεθόδου πρόγνωσης με χρήση εξελικτικής πληροφορίας σε μορφή πολλαπλών στοιχίσεων.*

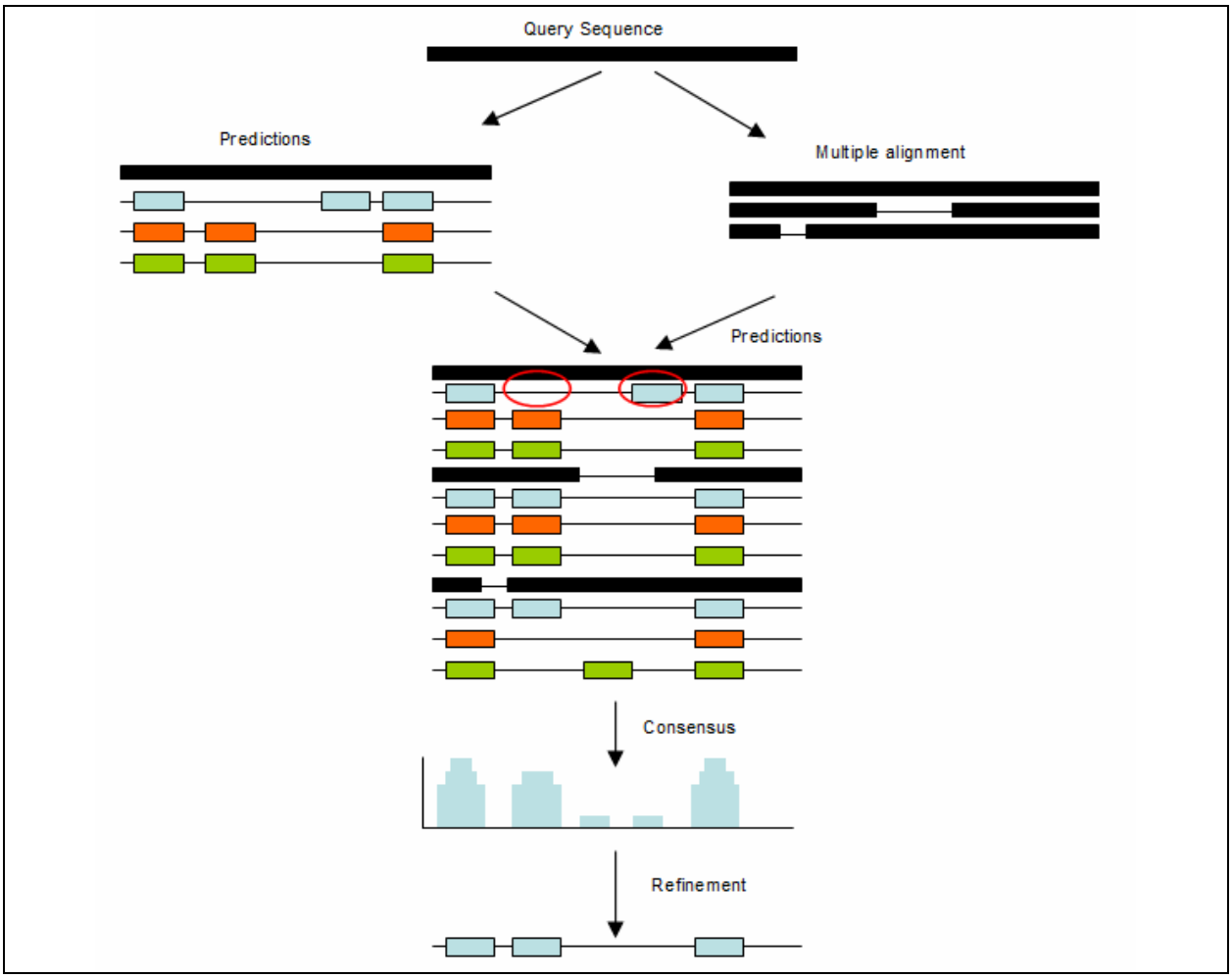
Μια πολύ ενδιαφέρουσα παραλλαγή της μεθόδου αυτής, προέκυψε όταν δημιουργήθηκε το γνωστό PSI-BLAST. Το πρόγραμμα αυτό εντοπίζει, με μια επαναληπτική διαδικασία ομόλογες αλληλουχίες και κατασκευάζει μια ειδικού τύπου πολλαπλή στοίχιση στην οποία δεν περιέχονται κενά στην αλληλουχία επερώτησης, από την οποία προκύπτει τελικά ένας πίνακας σκορ ειδικός ανά θέση (PSSM). Ο πίνακας αυτός συνοψίζει σε μια πολύ βολική μορφή ολόκληρη την πολλαπλή στοίχιση, ανεξάρτητα αν αυτή αποτελείτο από 5 ή 5000 αλληλουχίες (Εικόνα 7.12). Εκτός του ότι το PSI-BLAST είναι πολύ αποδοτικό στον εντοπισμό και τη στοίχιση μακρινών ομολόγων (πράγμα που ενισχύει από μόνο του την απόδοση της μεθόδου), η ύπαρξη του πίνακα κάνει δυνατή την κατασκευή άλλων μεθόδων που θα χρησιμοποιούν κατευθείαν τα δεδομένα του ίδιου του πίνακα και όχι τις αρχικές αλληλουχίες. Τέτοιου είδους αναπαράσταση είναι ιδανική για χρήση νευρωνικών δικτύων, αλλά και άλλες παραλλαγές έχουν προταθεί όπως στην περίπτωση των HMM. Το μεγάλο πλεονέκτημα αυτής της παραλλαγής είναι το ότι με τη συμπυκνωμένη μορφή αποφεύγεται η ανάγκη για πολλαπλή εφαρμογή του αλγορίθμου πρόγνωσης, αλλά από την άλλη, αυτό ακριβώς είναι και αδυναμία της, καθώς έτσι γίνεται απαραίτητη η δημιουργία και εκπαίδευση νέων μεθόδων πρόγνωσης με χρήση του πίνακα. Οι περισσότερες σύγχρονες μέθοδοι πρόγνωσης, κυρίως όσες βασίζονται σε μεθόδους μηχανικής μάθησης όπως τα νευρωνικά δίκτυα, χρησιμοποιούν αποκλειστικά αυτή τη μεθοδολογία καθώς τα νευρωνικά δίκτυα είναι ιδιαίτερα εύκολο να χρησιμοποιηθούν με τέτοιου είδους δεδομένα. Μια παραλλαγή αυτής της μεθόδου, έχει προταθεί κυρίως για αναγνώριση μακρινών ομολόγων. Συγκεκριμένα η μέθοδος αυτή συνίσταται στην εύρεση του προφίλ από το PSI-BLAST και μετέπειτα στην «αντικατάσταση» των αμινοξέων της υπό μελέτη πρωτεΐνης με τα πιο «κοινά» αμινοξικά κατάλοιπα σε κάθε θέση. Με τη μέθοδο αυτή, χάνεται μεν αρκετή πληροφορία (καθώς δεν έχουμε πλέον την πληροφορία για τη σχετική συντήρηση σε κάθε θέση της πολλαπλής στοίχισης), αλλά από την άλλη, με το σχηματισμό αυτής της «ψευτο-ακολουθίας», το

πρόβλημα ανάγεται πάλι στην απλή περίπτωση μίας και μόνο αλληλουχίας πρωτεΐνης, με συνέπεια την εύκολη εφαρμογή μεθόδων που είναι σχεδιασμένες για απλές αλληλουχίες (Przybylski & Rost, 2007).

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	I	-2	-4	-5	-5	-2	-4	-4	-5	-5	6	0	-4	0	-2	-4	-4	-2	-4	-3	4
2	K	-1	-1	-2	-2	-3	-1	3	-3	-2	-2	-3	4	-2	-4	-3	1	1	-4	-3	2
3	E	5	-3	-3	-3	-3	3	1	-2	-3	-3	-3	-2	-2	-4	-3	-1	-2	-4	-3	1
4	E	-4	-3	2	5	-6	1	5	-4	-3	-6	-6	-2	-5	-6	-4	-2	-3	-6	-5	-5
5	H	-4	2	1	1	-5	1	-2	-4	9	-5	-2	-3	-4	-4	-5	-3	-4	-5	1	-5
6	V	-3	0	-4	-5	-4	-4	-2	-3	-5	1	-2	1	0	1	-4	-3	3	-5	-3	5
7	I	0	-2	-4	1	-4	-2	-4	-4	-5	1	0	-2	0	2	-5	1	-1	-5	-3	4
8	I	-3	0	-5	-5	-4	-2	-5	-6	1	2	4	-4	-1	0	-5	-2	0	-3	5	-1
9	Q	-2	-3	-2	-3	-5	4	-1	3	5	-5	-3	-3	-4	-2	-4	2	-1	-4	2	-2
10	A	2	-4	-4	-3	2	-3	-1	-4	-2	1	-1	-4	-3	-4	1	2	3	-5	-1	1
11	E	-1	3	1	1	-1	0	1	-4	-3	-1	-3	0	3	-5	4	-1	-3	-6	-3	-1
12	F	-3	-5	-5	-5	-4	-4	-4	-1	-1	1	1	-5	2	5	-1	-4	-4	-3	5	2
13	Y	3	-5	-5	-6	3	-4	-5	-2	-1	0	-4	-5	-3	3	-5	-2	-2	-2	7	1
14	L	-1	-3	-4	-2	1	5	1	-1	-1	-1	1	-3	-3	1	-5	-1	-1	-2	3	-2
15	N	-1	-4	4	1	5	-3	-4	2	-4	-4	-4	-3	-2	-4	-5	2	0	-5	0	0
16	P	-2	4	-4	-4	-5	0	-3	3	2	-5	-4	0	-4	-3	0	1	-2	-1	5	-3
17	D	-3	-2	1	5	-6	-2	2	2	-1	-2	-2	-3	-5	-4	-5	-1	2	-6	-3	-4

Εικόνα 7.12: *Ένα παράδειγμα PSSM. Παρατηρήστε ότι οι Ισολευκίνες στις θέσεις 1, 7 και 8, έχουν διαφορετική κωδικοποίηση που αντανακλά τις διαφορετικές συχνότητες αμινοξέων στην αντίστοιχη στήλη της πολλαπλής στοίχισης.*

Τέλος, πρέπει να τονίσουμε ότι οι δύο παραπάνω γενικές μεθοδολογίες (η συνδυαστική πρόγνωση και η χρήση πολλαπλών στοιχίσεων), μπορούν άνετα να συνδυαστούν μεταξύ τους (Εικόνα 7.13). Φυσικά, όταν έχεις μια σειρά μεθόδων που η κάθε μία χρησιμοποιεί εξελικτική πληροφορία, τότε όπως είπαμε, αυτές εύκολα συνδυάζονται σε μια συναινετική πρόγνωση. Επιπλέον όμως, ακόμα και αν είχαμε μεθόδους που βασίζονται μόνο σε απλές αλληλουχίες, πάλι θα μπορούσαμε να εφαρμόσουμε πρώτα τη χρήση πολλαπλών στοιχίσεων και μετά τον συνδυασμό των μεθόδων. Πάλι είναι δυνατόν να υπάρξουν πολλές παραλλαγές όσον αφορά τον τρόπο σταθμίσματος της συνεισφοράς κάθε μεθόδου ή όσον αφορά τη βελτιστοποίηση και το φιλτράρισμα των τελικών προβλέψεων, αλλά γενικά η μεθοδολογία είναι εύκολη και κατανοητή και (το πιο σημαντικό) αυξάνει την αποτελεσματικότητα των απλών μεθόδων.

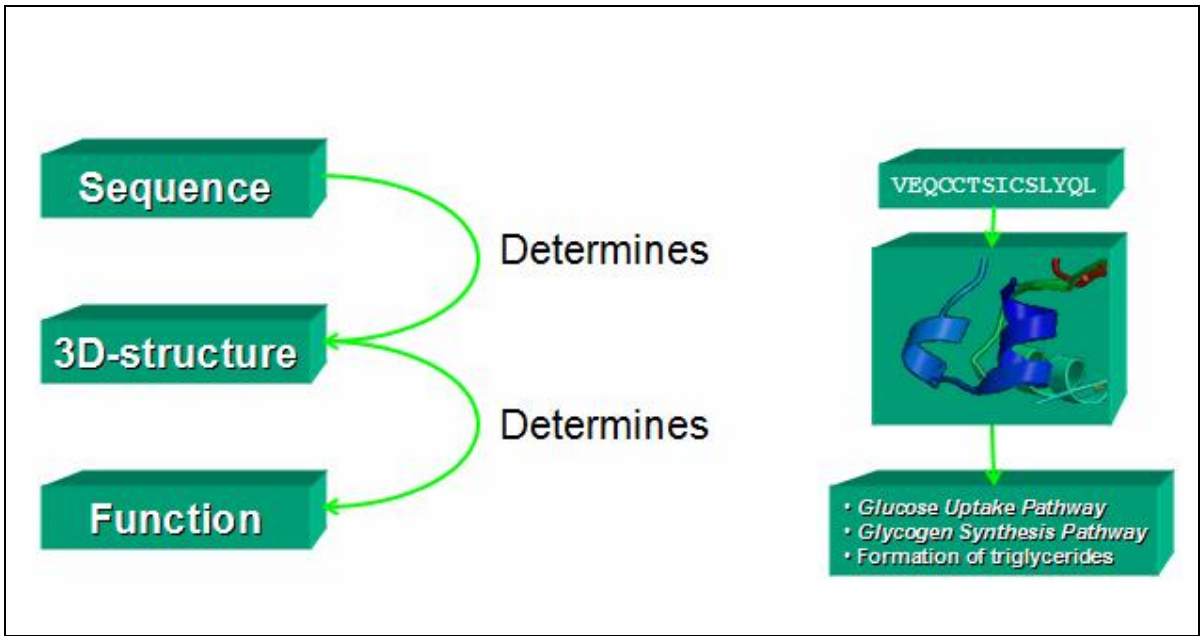


Εικόνα 7.13: *Ένα υποθετικό παράδειγμα συνδυασμού τόσο των συναινετικών μεθόδων αλλά και της χρήσης εξελικτικής πληροφορίας.*

7.6. Μέθοδοι πρόγνωσης για αλληλουχίες πρωτεϊνών

7.6.1 Δευτεροταγής δομή

Η πρόγνωση δευτεροταγούς δομής, είναι ίσως το αρχέτυπο των μεθόδων πρόγνωσης και μαζί με τη στοίχιση αλληλουχιών ένα από τα πιο παλιά προβλήματα, ήδη από τη δεκαετία του 1970, όταν δεν υπήρχε καν ο όρος βιοπληροφορική. Η μεγάλη σημασία των μεθόδων πρόγνωσης δευτεροταγούς δομής έγκειται στη γενικότητά τους, καθώς η δευτεροταγής δομή επηρεάζει πολλά άλλα δομικά χαρακτηριστικά, αλλά και στο αδιαμφισβήτητο γεγονός ότι τα περισσότερα λειτουργικά χαρακτηριστικά εξαρτώνται, λίγο ή πολύ, από την δομή της πρωτεΐνης. Προφανώς, η ακριβής τρισδιάστατη δομή είναι πιο δύσκολο να προβλεφθεί, αλλά η δευτεροταγής δομή, η οποία αφορά την τοπική μόνο διαμόρφωση της πολυπεπτιδικής αλυσίδας σε 3 κατηγορίες α-έλικα (H), β-πτυχωτή επιφάνεια (E) και τυχαία δομή (C), είναι αρκετά πιο εύκολο να αποτελέσει αντικείμενο πρόγνωσης.



Εικόνα 7.14: *Ο «γενετικός κώδικας» της βιολογίας των πρωτεϊνών. Η αλληλουχία καθορίζει τη δομή και η δομή καθορίζει τη λειτουργία.*

Οι πρώτες μέθοδοι που προτάθηκαν στηρίζονταν στη βασική αρχή ότι στις διάφορες κατηγορίες δευτεροταγούς δομής υπάρχουν διαφορετικές προτιμήσεις για την εμφάνιση διαφόρων αμινοξέων. Για παράδειγμα η Αλανίνη, το Γλουταμικό και η Λευκίνη έχουν ισχυρή προτίμηση να βρίσκονται σε α-έλικα ενώ η Προλίνη, η Γλυκίνη και η Σερίνη, όχι. Στην πρώτη και πιο δημοφιλή τέτοια περίπτωση μεθόδου, οι **Chou και Fasman** (Chou & Fasman, 1978) στηριζόμενοι σε ένα (μάλλον μικρό) σύνολο από 29 πρωτεΐνες με γνωστή τρισδιάστατη δομή που ήταν διαθέσιμες τότε, υπολόγισαν τις συχνότητες εμφάνισης αμινοξέων στις 3 κατηγορίες (H, E, C) και με βάση αυτές, υπολόγισαν τις λεγόμενες στερεοδιαταξικές παραμέτρους (P). Η μεθοδολογία ήταν σε γενικές γραμμές η εξής: Ξεκινάμε ορίζοντας ως $f^j(i)$ = τη συχνότητα εμφάνισης του αμινοξέος i στην κατάσταση j (helix, sheet, turn). Στη συνέχεια υπολογίζουμε τη μέση συχνότητα $\langle f^j \rangle$ ως τη μέση τιμή όλων των f για όλα τα αμινοξέα της κατηγορίας j. Τέλος, υπολογίζουμε τη στερεοδιαταξική παράμετρο $P^j(i)$ για κάθε αμινοξύ i και κατάσταση j ως $P^j(i) = f^j(i)/\langle f^j \rangle$. Οι τιμές των παραμέτρων αυτών όπως υπολογίστηκαν από τους Chou και Fasman δίνονται στον Πίνακα 7.1. Για παράδειγμα, στο σύνολο εκπαίδευσης υπήρχαν 228 Αλανίνες (119 σε α-έλικα, 38 σε β-πτυχωτή επιφάνεια και 71 σε τυχαία δομή). Άρα, οι παράμετροι θα είναι $f^H(A) = 0.522$, $f^E(A) = 0.167$ και $f^C(A)$ $f_c = 0.311$. Για την α-έλικα οι μέσες τιμές είναι $\langle f^H \rangle$=890/2473 = 0.359, για τη β-πτυχωτή επιφάνεια $\langle f^E \rangle$=424/2473 = 0.171 και για την τυχαία δομή, $\langle f^C \rangle$=1159/2473 = 0.469. Κατά συνέπεια, οι στερεοδιαταξικές παράμετροι για την Αλανίνη θα είναι P^H(A)= 0.522/0.359 = 1.45, P^E(A)= 0.167/0.171 = 0.97 και P^C(A)=0.311/0.469 = 0.63.

Τιμές με $P^j(i)$>1.0 δηλώνουν μεγάλη προτίμηση του αμινοξέος να βρίσκεται στο δεδομένο στοιχείο δευτεροταγούς δομής. Αφού έχουν υπολογιστεί οι παράμετροι, η μέθοδος απαιτεί την εφαρμογή μιας σειράς κανόνων. Για παράδειγμα, στην αρχή απαιτείται ο εντοπισμός «πυρήνων» δευτεροταγούς δομής, δηλαδή 4 συνεχόμενα κατάλοιπα με $P^H(i)$>1 ή 3 από τα 5 συνεχόμενα κατάλοιπα με $P^E(i)$>1. Όταν εντοπιστούν οι πυρήνες, οι περιοχές επεκτείνονται προς τις δύο κατευθύνσεις μέχρι να εντοπιστούν 4 συνεχόμενα κατάλοιπα με $P^j(i)$<1. Επιπλέον κανόνες, αφορούν τη μη ύπαρξη Προλίνης στις α-έλικες και Γλουταμικού και Προλίνης στις β-πτυχωτές επιφάνειες, οι προτιμήσεις για τα αμινοτελικά και τα καρβοξυτελικά άκρα των ελίκων (Προλίνη, Ασπαρτικό, Γλουταμικό και Ιστιδίνη, Λυσίνη και Αργινίνη αντίστοιχα). Τέλος, ειδικά για τις β-πτυχωτές επιφάνειες (οι οποίες είναι παραδοσιακά οι πιο δύσκολες περιοχές για πρόβλεψη), απαιτείται η παρουσία τουλάχιστον 5 συνεχόμενων καταλοίπων με $P^E(i)$>1.05, και $P^E(i)$> $P^H(i)$ για την ίδια περιοχή. Αξίζει να σημειωθεί ακόμα, ότι στην αρχική εργασία είχαν υπολογιστεί ειδικές παράμετροι για τις στροφές (T, turn), αλλά πλέον δεν χρησιμοποιούνται καθώς οι περισσότεροι προβλέπουν την κατηγορία C (coil, τυχαία δομή) ενώ για τις στροφές υπάρχουν εξειδικευμένες μέθοδοι.

Βλέπουμε, πως η μέθοδος αυτή μοιάζει πολύ με τη γενική μέθοδο του log-odds score και τη χρήση του κινούμενου παραθύρου που έχουμε περιγράψει σε προηγούμενα κεφάλαια. Μία διαφορά είναι ότι με το

log-odds score, η σύγκριση γίνεται απευθείας ανάμεσα στη συχνότητα εμφάνισης του αμινοξέος στην περιοχή, σε σχέση με το σύνολο, ενώ στη μέθοδο Chou-Fasman οι παράμετροι κανονικοποιούνται πρώτα για την περιοχή και μετά για το σύνολο. Επίσης, το log-odds score είναι σε λογαριθμική κλίμακα και κατά συνέπεια λειτουργεί αθροιστικά, ενώ οι στερεοδιαταξικές παράμετροι της μεθόδου Chou-Fasman λειτουργούν πολλαπλασιαστικά. Κατά τα άλλα πάντως, σαν μεθοδολογίες είναι εντελώς συγκρίσιμες από στατιστική άποψη. Η μέθοδος αυτή έδινε υψηλά ποσοστά σωστών προβλέψεων για τα δεδομένα της εποχής (~60%) αλλά μετέπειτα αμερόληπτες μελέτες έριξαν το ποσοστό αυτό στο 55%. Ένα άλλο σημείο κριτικής αφορούσε το γεγονός ότι οι παράμετροι είχαν υπολογιστεί από μικρό αριθμό πρωτεϊνών (και πιθανώς, από μη αντιπροσωπευτικό δείγμα). Παρ' όλα αυτά, μετέπειτα υπολογισμοί σε μεγαλύτερα σύνολα δεδομένων έδωσαν παρόμοια αποτελέσματα. Παρόλο που η μέθοδος αυτή δεν χρησιμοποιείται πλέον, μια υλοποίησή της κυρίως για ιστορικούς λόγους υπάρχει στη διεύθυνση http://cho-fas.sourceforge.net/

aminoacid	P(helix)	P(sheet)	P(coil)
A (Ala)	1.420	0.830	0.660
R (Arg)	0.980	0.930	0.950
N (Asn)	0.670	0.890	1.560
D (Asp)	1.010	0.540	1.460
C (Cys)	0.700	1.190	1.190
Q (Gln)	1.110	1.100	0.980
E (Glu)	1.510	0.370	0.740
G (Gly)	0.570	0.750	1.560
H (His)	1.000	0.870	0.950
I (Ile)	1.080	1.600	0.470
L (Leu)	1.210	1.300	0.590
K (Lys)	1.160	0.740	1.010
M (Met)	1.450	1.050	0.600
F (Phe)	1.130	1.380	0.600
P (Pro)	0.570	0.550	1.520
S (Ser)	0.770	0.750	1.430
T (Thr)	0.830	1.190	0.960
W (Trp)	1.080	1.370	0.960
Y (Tyr)	0.690	1.470	1.140
V (Val)	1.060	1.700	0.500

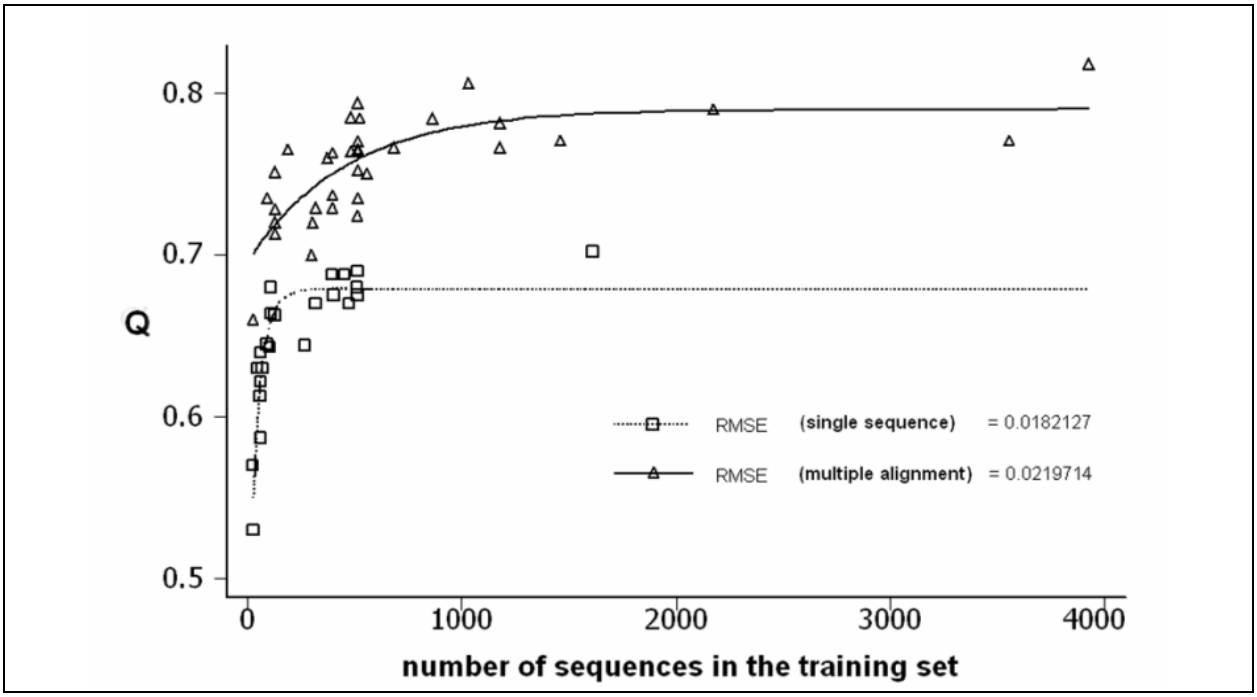
Πίνακας 7.1: *Οι τιμές των παραμέτρων(P) όπως υπολογίστηκαν από τους Chou και Fasman*

Μια αδυναμία της μεθόδου, ήταν το γεγονός ότι αντιμετώπιζε τις διάφορες θέσεις σε ένα δεδομένο παράθυρο ανεξάρτητα. Θεωρητικά, αναμένουμε ότι διαφορετικοί συνδυασμοί των ίδιων αμινοξέων θα δίνουν διαφορετικές προτιμήσεις ανάλογα με τη συγκεκριμένη αλληλουχία (αυτή είναι η ουσία της αλληλεπίδρασης). Αυτό το πρόβλημα ήρθε να λύσει η μέθοδος **GOR** (Garnier-Osguthorpe-Robson). Στην αρχική της μορφή χρησιμοποίησε τη μαθηματικά πιο «σωστή» τεχνική του log-odds score σε συνδυασμό με ένα πίνακα ειδικό ανά θέση με μήκος 17 κατάλοιπα (Garnier, Osguthorpe, & Robson, 1978). Επειδή τα δεδομένα της εποχής ήταν λίγα και δεν επέτρεπαν τον υπολογισμό όλων των πιθανών συσχετίσεων των αμινοξέων, οι συγγραφείς χρησιμοποίησαν στατιστική μεθοδολογία για να βρουν τις αναμενόμενες τιμές για τις δεσμευμένες πιθανότητες χρησιμοποιώντας μόνο τις κατά ζεύγη συσχετίσεις των αμινοξέων (τις προτιμήσεις τους ανά δύο). Η μέθοδος αυτή βασίζεται, όπως είδαμε, σε πιο στέρεες μαθηματικές βάσεις, καθώς χρησιμοποιεί αποτελέσματα της θεωρίας πληροφορίας και μπεϋζιανή μεθοδολογία. Επιπλέον δε, έχει βελτιωθεί με τα χρόνια, με την τελευταία έκδοση, την **GOR IV** (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html), η οποία κάνει χρήση μόνο της αμινοξικής αλληλουχίας, να φτάνει ένα ποσοστό σωστών προγνώσεων της τάξης του 64%, ενώ με την **GOR V** (http://gor.bb.iastate.edu/), η οποία χρησιμοποιεί πολλαπλές στοιχίσεις με τη μορφή προφίλ του PSI-BLAST, φτάνει πλέον σε ένα ποσοστό ακρίβειας της τάξης του 74%.

Το επόμενο μεγάλο βήμα στην πρόγνωση δευτεροταγούς δομής των πρωτεϊνών έγινε το 1987 όταν οι Qian και Sejnowski χρησιμοποίησαν για πρώτη φορά νευρωνικό δίκτυο και η ακρίβεια της μεθόδου ανέβηκε

και άλλο, περίπου στο 68% (Qian & Sejnowski, 1988). Αλλά, η πρώτη φορά που μια μέθοδος πέρασε το όριο του 70% ήταν το 1992 όταν οι Rost και Sander παρουσίασαν την πρώτη έκδοση του **PHD** (Rost & Sander, 1993). Η μέθοδος αυτή ήταν πρωτοποριακή γιατί χρησιμοποίησε ένα συνδυασμό ανεξάρτητων νευρωνικών δικτύων ("jury of networks" το ονόμασαν), μεγάλο σύνολο εκπαίδευσης (130 μη ομόλογες πρωτεΐνες), ένα δεύτερο δίκτυο για το φιλτράρισμα των αποτελεσμάτων (structure-to-structure network) αλλά και για πρώτη φορά έκανε χρήση εξελικτικής πληροφορίας με τη μορφή πολλαπλών στοιχίσεων. Η κίνηση αυτή έδωσε μια αύξηση της ακρίβειας της μεθόδου της τάξης του 6-8% και από τότε έχει γίνει αποδεκτό ότι ακρίβειες μεγαλύτερες από 70% μπορούν να επιτευχθούν μόνο με χρήση πολλαπλών στοιχίσεων. Το **PSI-PRED** (http://bioinf.cs.ucl.ac.uk/psipred/) ήταν η πρώτη μέθοδος πρόγνωσης που χρησιμοποίησε τα profiles του PSI-BLAST (Jones, 1999) και έφτασε μεγαλύτερες τιμές ακρίβειας (της τάξης του 76%). Τα επιπλέον χαρακτηριστικά του PSI-PRED ήταν η χρήση δύο διαδοχικών δικτύων αλλά και η χρήση ενός ακόμα μεγαλύτερου συνόλου εκπαίδευσης (513 μη ομόλογες πρωτεΐνες από 187 διαφορετικά πρωτεϊνικά διπλώματα). Λίγα χρόνια αργότερα, εμφανίστηκε και η νέα έκδοση του PHD, το **PROFphd** το οποίο επίσης χρησιμοποιεί προφίλ από το PSI-BLAST και έφτασε σε παρόμοια επίπεδα επιτυχίας (~75-76%). Παρόμοια μεθοδολογία αλλά και ποσοστά επιτυχίας, εμφανίζει και το επίσης γνωστό **JNET** (Cuff & Barton, 2000). Έκτοτε, έχουν αναπτυχθεί πολλές άλλες μέθοδοι, η πλειοψηφία τους όμως χρησιμοποιεί τα προφίλ του PSI-BLAST, έστω και αν σαν βασική μεθοδολογία τους χρησιμοποιούν διαφορετικές τεχνικές όπως τα Support Vector Machines (SVM) ή τα Recurrent Neural Networks (RNN).

Γενικά, η επιτυχία μιας μεθόδου πρόγνωσης εξαρτάται από το είδους του αλγορίθμου (τα νευρωνικά δίκτυα και οι άλλες τεχνικές μηχανικής μάθησης αποδίδουν καλύτερα από τις απλές στατιστικές τεχνικές), από το μέγεθος του συνόλου εκπαίδευσης (μέθοδοι που χρησιμοποίησαν μεγαλύτερα σύνολα αποδίδουν καλύτερα) και από το αν χρησιμοποιεί πολλαπλές στοιχίσεις (οι μέθοδοι που χρησιμοποιούν πολλαπλές στοιχίσεις αποδίδουν πάντα καλύτερα). Αναδρομικές μελέτες βασισμένες στα δημοσιευμένα αποτελέσματα μεθόδων πρόγνωσης (όταν τα αποτελέσματα προέκυψαν από ανεξάρτητο σύνολο ελέγχου ή cross-validation) έχουν όμως δείξει ότι με τις παρούσες μεθοδολογίες, οι μέθοδοι πρόγνωσης έχουν ένα ανώτατο όριο στην αναμενόμενη ακρίβεια και, μάλιστα, από ένα σημείο και μετά η απόδοση δεν αυξάνει (Bagos, Tsaousis, & Hamodrakas, 2009).



Εικόνα 7.15: *Η αύξηση της απόδοσης των αλγορίθμων δευτεροταγούς δομής σε συνάρτηση με το μέγεθος του συνόλου εκπαίδευσης. Με διαφορετικά σύμβολα απεικονίζονται οι μέθοδοι που βασίζονται μόνο στην αλληλουχία και αυτές που χρησιμοποιούν πολλαπλές στοιχίσεις (Bagos, Tsaousis, et al., 2009).*

Για παράδειγμα οι μέθοδοι που χρησιμοποιούν απλές ακολουθίες, φτάνουν σε ένα ανώτατο όριο γύρω στο 70% και μάλιστα (καθώς συνήθως έχουν λιγότερες παραμέτρους) αυτό το όριο έρχεται σχετικά γρήγορα (όταν το σύνολο εκπαίδευσης είναι περίπου στις 500 πρωτεΐνες). Αντίθετα, οι μέθοδοι που χρησιμοποιούν εξελικτική πληροφορία φτάνουν σε υψηλότερα επίπεδα αλλά και πάλι δεν μπορούν να ξεπεράσουν το 80% όσο και αν αυξηθεί το σύνολο εκπαίδευσης (υπήρχαν και μέθοδοι που εκπαιδεύτηκαν με πάνω από 2000 πρωτεΐνες). Κατά συνέπεια, αν θέλουμε να περάσουμε αυτό το όριο του 80% θα πρέπει να δοθεί έμφαση στην ανάπτυξη νέων μεθοδολογιών και, μάλιστα, σε μεθοδολογίες που θα χρησιμοποιούν τις μακρινές αλληλεπιδράσεις κατά μήκος της αλληλουχίας.

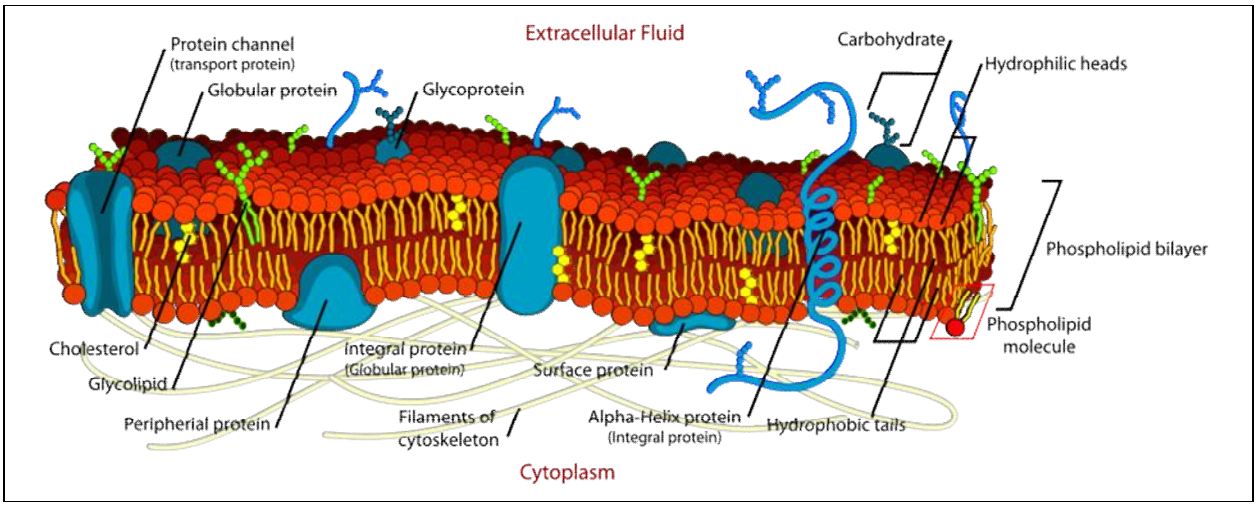
Οι συνδυαστικές/συναινετικές προγνώσεις είχαν επίσης δείξει από παλιά ότι μπορούν να αυξήσουν την επιτυχία των μεθόδων πρόγνωσης. Μια από τις πρώτες προσπάθειες είχε γίνει το 1988 όταν ο Hamodrakas (Hamodrakas, 1988) δημοσίευσε ένα συνδυαστικό αλγόριθμο που έκανε χρήση των τότε διαθέσιμων μεθόδων (Chou-Fasman, GOR, Lim, Dufton-Hider, Burgess, Nagano). Η μέθοδος αυτή έδειξε μια βελτίωση της τάξης του 2-3% και μετέπειτα έγινε και διαθέσιμη σαν διαδικτυακή εφαρμογή με το όνομα **SecStr** (http://athina.biol.uoa.gr/SecStr/). Βέβαια, γίνεται αντιληπτό ότι καθώς οι μέθοδοι που χρησιμοποιεί το SecStr είναι παλιές και κάνουν χρήση μόνο της αλληλουχίας, τα αναμενόμενα ποσοστά επιτυχίας θα είναι περιορισμένα κάτω από το 70%. Το **JPRED** (http://www.compbio.dundee.ac.uk/jpred/) ήταν ίσως η πρώτη μέθοδος που χρησιμοποίησε συνδυασμό μεθόδων και ταυτόχρονα έκανε χρήση εξελικτικής πληροφορίας το 1998 (Cuff, Clamp, Siddiqui, Finlay, & Barton, 1998). Στην πρώτη έκδοση έκανε χρήση του JNET και μιας σειράς άλλων αλγορίθμων της εποχής (NNSSP, DSC, PREDATOR, MULPRED, PHD, ZPRED) και ανέφερε σημαντικά βελτιωμένη απόδοση. Σήμερα, η μέθοδος έχει φτάσει στην έκδοση 4 (JPRED4) και συγκαταλέγεται ανάμεσα στις καλύτερες μεθόδους, έχοντας αυτοματοποιημένη πρόσβαση μέσω διαδικτυακής εφαρμογής και πολλές επιλογές, όπως γραφικές παραστάσεις των αποτελεσμάτων ή τη δυνατότητα ο χρήστης να δώσει τη δική του πολλαπλή στοίχιση. Μια άλλη γνωστή από παλιά συνδυαστική μέθοδος είναι η **NPS@** (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html) η οποία κάνει συνδυαστική πρόγνωση με χρήση των μεθόδων SOPM, SOPMA, HNN, MLRC, DPM, DSC, GOR I, GOR III, GOR IV, PHD, PREDATOR, SIMPA96 ενώ δίνει στο χρήστη τη δυνατότητα να επιλέξει ποιες από αυτές θα χρησιμοποιηθούν. Άλλες πιο πρόσφατες συνδυαστικές μέθοδοι είναι το **CONCORD** (http://helios.princeton.edu/CONCORD/) το οποίο χρησιμοποιεί τα PSIPRED, DSC, GOR IV, Predator, Prof, PROFphd, και SSpro, και το **SYMPRED** (http://www.ibi.vu.nl/programs/sympredwww/) το οποίο κάνει χρήση των PHDpsi, PROFsec, SSPro, Predator, YASPIN, JNet και PSIPRED.

Πρέπει να τονίσουμε σε αυτό το σημείο, ότι η σύγχρονη τάση των μεγάλων εργαστηρίων είναι να διαθέτουν σε μια διαδικτυακή εφαρμογή όλες τις σχετικές μεθόδους πρόγνωσης (δευτεροταγούς δομής, προσβασιμότητας του διαλύτη, διαμεμβρανικών τμημάτων κ.ο.κ.). Έτσι, οι μέθοδοι του B. Rost βρίσκονται όλες μαζί στην ιστοσελίδα **PREDICTPROTEIN** (www.predictprotein.org/), στην ιστοσελίδα του **PSI-PRED** (http://bioinf.cs.ucl.ac.uk/psipred/) διατίθενται εκτός από την ομώνυμη εφαρμογή και άλλες μέθοδοι πρόγνωσης πρωτεϊνών του εργαστηρίου, ενώ αντίστοιχες μέθοδοι διατίθενται στο **SCRATCH** (http://scratch.proteomics.ics.uci.edu/index.html).

Τέλος, πρέπει να αναφερθεί ο τρόπος με τον οποίο αξιολογούνται οι μέθοδοι. Όταν κάποιος δημιουργήσει μια νέα μέθοδο πρόγνωσης είναι λογικό να ελέγξει την αποτελεσματικότητά της σε ένα ανεξάρτητο σύνολο που δεν έχει ομοιότητα με το σύνολο εκπαίδευσης. Με αυτόν τον τρόπο όμως, δεν έχουμε πάντα αξιόπιστα νούμερα για τη σύγκριση καθώς οι διάφορες μέθοδοι δεν έχουν δοκιμαστεί στα ίδια παραδείγματα. Έτσι, από τη δεκαετία του 1990 οι επιστήμονες δημιούργησαν το συνέδριο **CASP** (Critical Assessment of Structure Predictions http://predictioncenter.org/). Σε αυτή την προσπάθεια, εντοπίζονται μετά από επικοινωνία με τους κρυσταλλογράφους οι αλληλουχίες των πρωτεϊνών που είναι «έτοιμες» να προσδιοριστούν πειραματικά. Αφού ελεγχθεί ότι οι αλληλουχίες αυτές δεν εμφανίζουν ομοιότητα με καμία άλλη πρωτεΐνη γνωστής δομής, οι αλληλουχίες ανακοινώνονται και οι διάφοροι αλγόριθμοι δοκιμάζονται. Όταν φτάσει ο καιρός του συνεδρίου τα αποτελέσματα των αλγορίθμων ανακοινώνονται και συγκρίνονται με τις πραγματικές δομές που στο μεταξύ έχουν προσδιοριστεί αλλά παραμένουν μυστικές. Μια άλλη προσπάθεια για συνεχή παραγωγή τέτοιων ανεξάρτητων συνόλων, είχε δημιουργήσει ο Rost. Το πρόγραμμα ονομάζεται **EVA** (Koh et al., 2003) και πραγματοποιούσε κάθε μήνα αναζήτηση στην PDB για νέες δομές και πραγμάτωνε τη σύγκριση με τα γνωστά σύνολα εκπαίδευσης όλων ή των περισσότερων, γνωστών μεθόδων. Έτσι, υπάρχει ένα συνεχώς ανανεωνόμενο σύνολο ανεξάρτητου ελέγχου για κάθε μέθοδο, οπότε με σύγκριση των συνόλων αυτών θα μπορεί ανά πάσα στιγμή να κατασκευαστεί ένα σύνολο που να είναι κατάλληλο για τη σύγκριση δύο ή περισσότερων αλγορίθμων.

7.6.2. Διαμεμβρανικές πρωτεΐνες

Οι βιολογικές μεμβράνες, είναι υπερμοριακοί σχηματισμοί οι οποίοι μπορούν να ειδωθούν τόσο σαν μηχανισμοί απομόνωσης, προστασίας και διαμερισματοποίησης του κυττάρου, όσο και σαν εξειδικευμένα όργανα επικοινωνίας και αλληλεπίδρασης του κυττάρου με το περιβάλλον του. Οι βιολογικές μεμβράνες, σύμφωνα με τις ισχύουσες απόψεις θεωρούμε ότι δομούνται με το μοντέλο του «ρευστού μωσαϊκού» (Singer & Nicolson, 1972) και αποτελούνται από μια διπλοστιβάδα λιπιδίων μέσα στην οποία ή και γύρω από αυτή, βρίσκονται σε διαρκή αλληλεπίδραση διαφόρων ειδών πρωτεΐνες (Εικόνα 7.16).



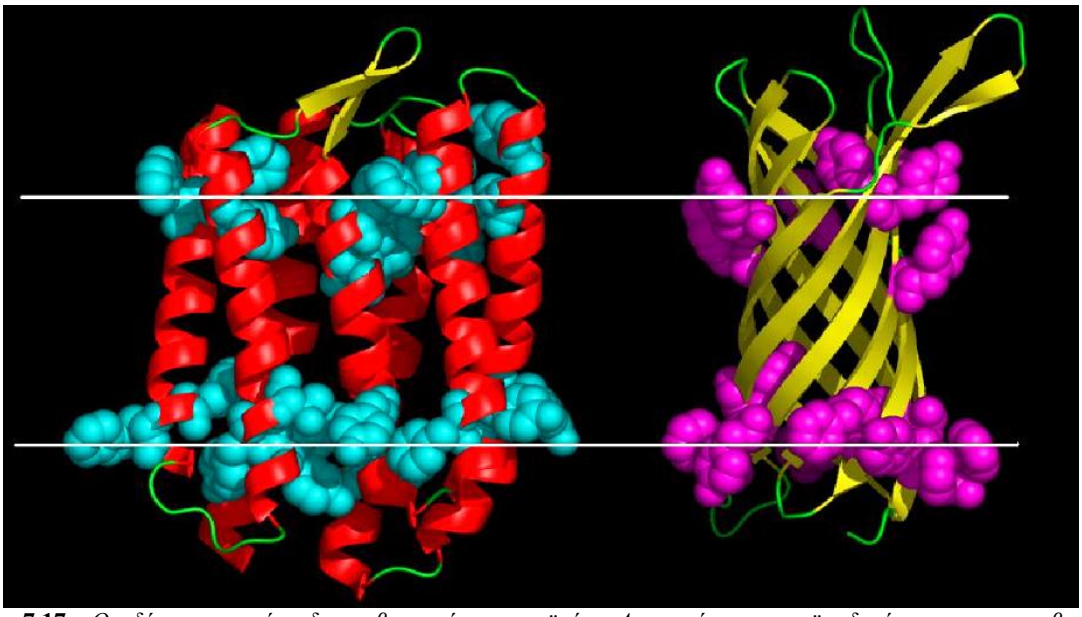
Εικόνα 7.16: *Απεικόνιση μιας τυπικής λιπιδικής διπλοστιβάδας, στην οποία φαίνονται οι διαμεμβρανικές και οι περιφερειακές πρωτεΐνες αλλά και τα υπόλοιπα συστατικά (https://en.wikipedia.org/wiki/Cell_membrane)*

Τα λιπίδια είναι διαφόρων ειδών (φωσφολιπίδια, γλυκολιπίδια, σφιγγολιπίδια, χοληστερόλη κλπ) και το κοινό τους γενικό χαρακτηριστικό είναι ότι διατάσσονται στη διπλοστιβάδα με τις πολικές κεφαλές τους να βρίσκονται προς την εξωτερική πλευρά (εκατέρωθεν της μεμβράνης), ενώ οι υδρόφοβες ουρές τους συσσωρεύονται στον εσωτερικό χώρο, αλληλεπιδρώντας μεταξύ τους και δημιουργώντας έτσι ένα ιδιαίτερα υδρόφοβο περιβάλλον. Συνέπεια αυτού, είναι η μεμβράνη να καθίσταται αδιαπέραστη από τα περισσότερα πολικά μόρια αλλά και από διάφορα μεγαλομόρια όπως π.χ. τις πρωτεΐνες. Τα ιδιαίτερα χαρακτηριστικά κάθε βιολογικής μεμβράνης (πάχος, διαπερατότητα, υδροφοβικότητα, ρευστότητα κλπ) καθορίζονται από την ιδιαίτερη σύστασή της σε λιπίδια αλλά και το είδος και την ποσότητα των πρωτεϊνών που απαντώνται σε αυτή.

Οι μεμβρανικές πρωτεΐνες επιτελούν μια σειρά από πολύ σημαντικές λειτουργίες, απαραίτητες για την ζωή του κυττάρου. Οι λειτουργίες αυτές μπορεί να ποικίλουν από την κυτταρική αναγνώριση, τη λειτουργία τους ως μοριακοί υποδοχείς, τη μεταφορά (παθητική ή ενεργητική) ουσιών διαμέσου της μεμβράνης, την έκκριση ουσιών, ως και την εξειδικευμένη ενζυμική δραστηριότητα (Alberts et al., 1994). Όπως είναι φανερό οι λειτουργίες αυτές είναι πολύ σημαντικές για την επιβίωση των οργανισμών καθώς πιθανή αλλοίωση τέτοιων πρωτεϊνών μπορεί να οδηγήσει σε διαφόρων ειδών ασθένειες. Από την άλλη, οι πρωτεΐνες είναι δυνατόν να αποτελέσουν και οι ίδιες στόχο ουσιών-φαρμάκων, προκειμένου να ανασταλεί ή να ενισχυθεί η λειτουργία τους κατά περίπτωση. Γενικά, οι μεμβρανικές πρωτεΐνες είναι δυνατόν να ταξινομηθούν σε δυο μεγάλες ομάδες, τις διαμεμβρανικές οι οποίες διαπερνούν με την πολυπεπτιδική τους αλυσίδα τη λιπιδική διπλοστιβάδα, και τις περιφερειακές και αγκυροβολημένες πρωτεΐνες οι οποίες βρίσκονται προσκολλημένες στην επιφάνεια της μεμβράνης με ασθενείς αλληλεπιδράσεις (περιφερειακές πρωτεΐνες) ή ομοιοπολικούς δεσμούς με τα λιπίδια (αγκυροβολημένες στη μεμβράνη πρωτεΐνες). Οι διαμεμβρανικές πρωτεΐνες, με τις οποίες θα ασχοληθούμε διεξοδικά παρακάτω, διαθέτουν ειδικά χαρακτηριστικά γνωρίσματα στην αμινοξική σύστασή τους κατά μήκος της ακολουθίας, μέσω των οποίων επιτυγχάνεται αλλά και εξηγείται η ενσωμάτωσή τους στη λιπιδική διπλοστιβάδα. Αντίθετα, οι αγκυροβολημένες με ομοιοπολικό τρόπο στα λιπίδια πρωτεΐνες, επιτυγχάνουν αυτήν την πρόσδεση μέσω αναγνώρισης από ειδικά ένζυμα μιας συγκεκριμένης αλληλουχίας στην αμινοξική τους ακολουθία, ενώ οι περιφερειακές πρωτεΐνες, προσκολλώνται με ασθενείς αλληλεπιδράσεις σε άλλες διαμεμβρανικές πρωτεΐνες

με τρόπο που δεν διαφέρει από τον γενικότερο τρόπο πρωτεϊνικών αλληλεπιδράσεων που συναντάμε στις σφαιρικές υδατοδιαλυτές πρωτεΐνες (Marsh, Horvath, Swamy, Mantripragada, & Kleinschmidt, 2002). Πολλές φορές τέλος, ιδιαίτερα στα ευκαρυωτικά κύτταρα, τα τμήματα των διαμεμβρανικών πρωτεϊνών, που προεξέχουν στον εξωκυττάριο χώρο υφίστανται μετα-μεταφραστικές τροποποιήσεις (π.χ. γλυκοζυλίωση), έτσι ώστε να τροποποιηθεί η λειτουργία τους.

Το γενικότερο σχήμα για την λιπο-πρωτεϊνική φύση των βιολογικών μεμβρανών, που είδαμε παραπάνω, ισχύει, με επιμέρους κατά περίπτωση τροποποιήσεις, και στις μεμβράνες των οργανιδίων που απαντώνται στο εσωτερικό των ευκαρυωτικών κυττάρων (μιτοχόνδρια, χλωροπλάστες, λυσοσώματα, Golgi κλπ). Πολλά δε είδη κυττάρων, διαθέτουν στην εξωτερική πλευρά (πέραν της μεμβράνης), ένα επιπλέον προστατευτικό στρώμα πολυσακχαριτικής προέλευσης, το λεγόμενο κυτταρικό τοίχωμα. Το κυτταρικό τοίχωμα, ποικίλει από κύτταρο σε κύτταρο ως προς τα ιδιαίτερα δομικά και λειτουργικά χαρακτηριστικά του. Έτσι, στα φυτικά κύτταρα το τοίχωμα αποτελείται από τον πολυσακχαρίτη κυτταρίνη, τα τοιχώματα των μυκήτων από χιτίνη, ενώ στα βακτήρια συναντάμε μουρεΐνη. Βασικός ρόλος του κυτταρικού τοιχώματος σε όλες πάντως τις περιπτώσεις, είναι να παρέχει ένα επιπλέον, σταθερό προστατευτικό στρώμα έναντι των επιδράσεων του περιβάλλοντος.

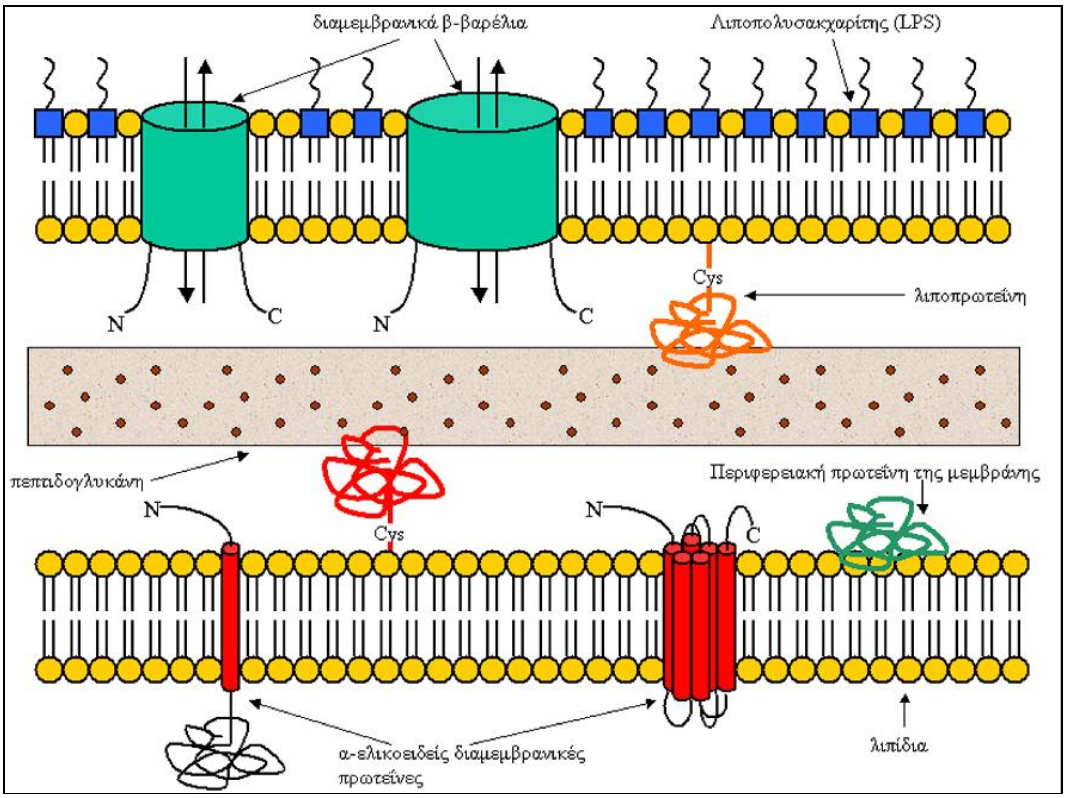


Εικόνα 7.17: *Οι δύο κατηγορίες διαμεμβρανικών πρωτεϊνών. Αριστερά, ο φωτοϋποδοχέας του αρχαιοβακτηρίου Natronomonas pharaonis. Δεξιά, η NspA, του βακτηρίου Neisseria meningitidis. Με τη χωροπληρωτική αναπαράσταση, διακρίνονται τα αρωματικά κατάλοιπα που συνιστούν την αρωματική ζώνη στα όρια της μεμβράνης.*

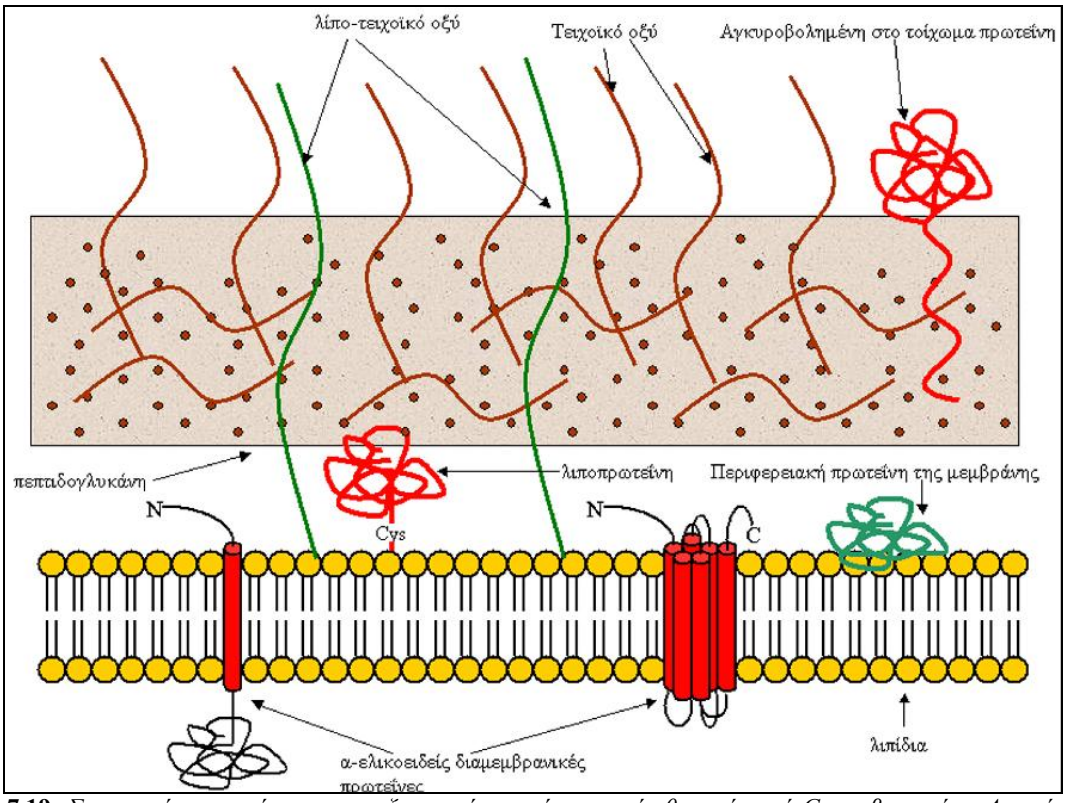
Οι διαμεμβρανικές πρωτεΐνες, σε γενικές γραμμές μπορούν να ταξινομηθούν ανάλογα με την δευτεροταγή δομή που υιοθετούν τα τμήματά τους που διαπερνούν τη λιπιδική διπλοστιβάδα. Έτσι υπάρχουν οι πρωτεΐνες που διαπερνούν τη μεμβράνη σε μορφή α-ελίκων (απομονωμένες ή σε μορφή δεματίου) και οι πρωτεΐνες των οποίων τα διαμεμβρανικά τμήματα αποτελούνται από β-πτυχωτές επιφάνειες σε μορφή αντιπαράλληλων κλειστών βαρελιών (Εικόνα 7.17). Οι πρωτεΐνες κάθε κατηγορίας, διαθέτουν διακριτά χαρακτηριστικά, προφανώς σχετιζόμενα με την τρισδιάστατη δομή των διαμεμβρανικών τμημάτων και την αντίστοιχη διαδικασία διπλώματος που έχει ακολουθηθεί σε κάθε περίπτωση. Κάποια από αυτά τα χαρακτηριστικά αντικατοπτρίζουν τη βιογένεση των μεμβρανικών πρωτεϊνών και των αντίστοιχων μεμβρανών, καθώς επίσης και τα χαρακτηριστικά των μηχανισμών κυτταρικής μεταφοράς αλλά και των περιβαλλοντικών περιορισμών που επιβάλλονται από τις φυσικοχημικές ιδιότητες των διαφόρων τύπων λιπιδικών διπλοστιβάδων.

Οι α-ελικοειδείς διαμεμβρανικές πρωτεΐνες εμφανίζονται σε μεγάλη αφθονία σε όλες σχεδόν τις κυτταρικές μεμβράνες (von Heijne, 1999), σε αντίθεση με τις διαμεμβρανικές πρωτεΐνες με μορφή β-βαρελιού οι οποίες έχουν παρατηρηθεί έως τώρα πειραματικά στην εξωτερική μεμβράνη των αρνητικών κατά Gram βακτηρίων αλλά και στις εξωτερικές μεμβράνες των μιτοχονδρίων και των χλωροπλαστών (Schulz, 2003). Στην πραγματικότητα, όλες οι διαμεμβρανικές πρωτεΐνες της εξωτερικής μεμβράνης των βακτηρίων

που έχουν εντοπισθεί έως σήμερα, πιστεύεται ότι ανήκουν σε αυτήν την κατηγορία αποτελώντας ένα σημαντικό τμήμα της συνολικής μάζας της εξωτερικής μεμβράνης (Εικόνα 7.18). Στα θετικά κατά Gram βακτήρια, γενικά, δεν απαντάται εξωτερική μεμβράνη (Εικόνα 7.19), αλλά το κυτταρικό τοίχωμα εμφανίζεται ιδιαίτερα παχύ. Ιδιαίτερες περιπτώσεις, αποτελούν κάποια οξεότροφα βακτήρια (π.χ. *Mycobacterium*), τα οποία εμφανίζουν ένα είδος εξωτερικής μεμβράνης, που αποτελείται από μυκολικό οξύ το οποίο τους προσδίδει (λόγω του πάχους της μεμβράνης) ιδιαίτερη αντοχή σε αντιβιοτικά.



Εικόνα 7.18: *Σχηματική αναπαράσταση της εξωτερικής επιφάνειας ενός αρνητικού κατά Gram βακτηρίου. Διακρίνουμε τις διαμεμβρανικές πρωτεΐνες (στην εξωτερική αλλά και στην εσωτερική μεμβράνη), τις περιφερειακές πρωτεΐνες, αλλά και τις αγκυροβολημένες στη μεμβράνη πρωτεΐνες. Το στρώμα της πεπτιδογλυκάνης είναι πιο λεπτό από το αντίστοιχο στα θετικά κατά Gram βακτήρια, και παρατηρούμε, επίσης, την ιδιαίτερη σύσταση της εξωτερικής μεμβράνης σε λιπίδια. Ο χώρος ανάμεσα στην εξωτερική και την εσωτερική μεμβράνη, ονομάζεται περιπλασμικός χώρος.*



Εικόνα 7.19: *Σχηματική αναπαράσταση της εξωτερικής επιφάνειας ενός θετικού κατά Gram βακτηρίου. Διακρίνουμε τις διαμεμβρανικές πρωτεΐνες, τις περιφερειακές πρωτεΐνες, αλλά και τις αγκυροβολημένες, είτε στο τοίχωμα είτε στη μεμβράνη, πρωτεΐνες.*

Οι α-ελικοειδείς διαμεμβρανικές πρωτεΐνες, σε πρώτη φάση διαχωρίζονται με βάση τον αριθμό και τον προσανατολισμό των διαμεμβρανικών τους ελίκων. Έτσι, οι Τύπου Ι διαμεμβρανικές πρωτεΐνες είναι αυτές οι οποίες διαθέτουν μια διαμεμβρανική α-έλικα, και των οποίων το αμινοτελικό άκρο βρίσκεται στον εξωκυττάριο χώρο, ενώ οι Τύπου ΙΙ διαμεμβρανικές πρωτεΐνες είναι αυτές οι οποίες έχουν επίσης ένα διαμεμβρανικό τμήμα αλλά το αμινοτελικό άκρο τους βρίσκεται στον ενδοκυττάριο χώρο (Alberts et al., 1994). Οι υπόλοιπες πρωτεΐνες, οι οποίες διαθέτουν περισσότερα του ενός διαμεμβρανικά τμήματα κατατάσσονται στις λεγόμενες *multi-spanning* μεμβρανικές πρωτεΐνες, οι οποίες με τη σειρά τους διαιρούνται σε περαιτέρω κατηγορίες οι οποίες συνήθως αντικατοπτρίζουν και λειτουργικές ομοιότητες. Για παράδειγμα, οι υποδοχείς, οι συζευγμένοι με G πρωτεΐνες (G-Protein Coupled Receptors-GPCRs), απαρτίζουν μια ετερογενή ομάδα υποδοχέων (Kristiansen, 2004) τα μέλη της οποίας όμως, εμφανίζουν δομικές (αριθμός διαμεμβρανικών τμημάτων και τοπολογία) αλλά και λειτουργικές ομοιότητες (μεταγωγή σήματος μέσω ετεροτριμερών G πρωτεϊνών). Οι διαμεμβρανικές πρωτεΐνες με δομή β-βαρελιού, από την άλλη πλευρά, διαιρούνται περαιτέρω όπως θα δούμε, σε διάφορες ομάδες κυρίως με βάση τη δομική τους ομοιότητα, η οποία αφορά των αριθμό των διαμεμβρανικών β-κλώνων και την κλίση τους ως προς το επίπεδο της μεμβράνης, η οποία τις περισσότερες φορές αντικατοπτρίζει επίσης λειτουργικές ομοιότητες.

Ενώ η πρόγνωση των διαμεμβρανικών τμημάτων των α-ελικοειδών διαμεμβρανικών πρωτεϊνών έχει επιχειρηθεί και μάλιστα με αρκετά μεγάλη επιτυχία εδώ και 20 τουλάχιστον χρόνια (Eisenberg, Weiss, & Terwilliger, 1984; Kyte & Doolittle, 1982), η αντίστοιχη διαδικασία για τα διαμεμβρανικά β-βαρέλια είναι πιο δύσκολη, για λόγους τους οποίους θα αναλύσουμε διεξοδικά παρακάτω. Στην περίπτωση των α-ελικοειδών διαμεμβρανικών πρωτεϊνών, ο εντοπισμός περιοχών 15-25 ιδιαίτερα υδρόφοβων καταλοίπων, είναι τις πιο πολλές φορές αρκετός για να μας δώσει μια επιτυχημένη πρόγνωση των πιθανών διαμεμβρανικών τμημάτων. Αν αυτό συνδυαστεί με την εφαρμογή του λεγόμενου «*positive inside rule*», δηλαδή της διαπίστωσης ότι οι περιοχές που βρίσκονται στην κυτοπλασμική πλευρά διαθέτουν πολύ περισσότερα θετικά φορτισμένα κατάλοιπα (von Heijne, 1992), τότε μια μέθοδος πρόγνωσης έχει ήδη κατασκευαστεί. Οι παραπάνω κανόνες (υδροφοβικότητα, *positive inside rule*), ισχύουν για όλες σχεδόν τις βιολογικές μεμβράνες στις οποίες απαντώνται α-ελικοειδείς μεμβρανικές πρωτεΐνες, συμπεριλαμβανομένων

των εσωτερικών μεμβρανών των μιτοχονδρίων (Rojo, Guiard, Neupert, & Stuart, 1999) και των χλωροπλαστών (Houben, de Gier, & van Wijk, 1999). Αλγόριθμοι που βασίζονται σε αυτούς, έχουν αναπτυχθεί εδώ και χρόνια βασισμένοι σε διαφόρων τύπων αλγοριθμικές τεχνικές, από εμπειρικούς αλγόριθμους με κυλιόμενα παράθυρα κατά μήκος της ακολουθίας (Claros & von Heijne, 1994), στατιστικές τεχνικές βασιζόμενες στις προτιμήσεις των αμινοξέων (Pasquier, Promponas, Palaios, Hamodrakas, & Hamodrakas, 1999), έως και σύγχρονες τεχνικές μηχανικής μάθησης όπως τα Hidden Markov Models (Krogh, Larsson, von Heijne, & Sonnhammer, 2001; Tusnady & Simon, 1998) και τα Νευρωνικά Δίκτυα (Pasquier & Hamodrakas, 1999; Rost, Casadio, Fariselli, & Sander, 1995).

Η γνώση της δομής μιας πρωτεΐνης σε ατομική διακριτικότητα, είναι ένα αποφασιστικό βήμα στην προσπάθεια κατανόησης της βιολογικής της λειτουργίας. Υψηλής διακριτικότητας τρισδιάστατες δομές είναι διαθέσιμες για μια μεγάλη ποικιλία σφαιρικών υδατοδιαλυτών πρωτεϊνών, σε αντίθεση με τον αριθμό των μοναδικών τρισδιάστατων δομών για διαμεμβρανικές πρωτεΐνες ο οποίος είναι αναλογικά πολύ μικρός. Παρ' όλη την εκπληκτική πρόοδο που έχει συντελεστεί τα τελευταία χρόνια στην κατευθυνόμενη γονιδιακή έκφραση, στο βιοχημικό καθαρισμό και προσδιορισμό και στις τεχνικές κρυστάλλωσης των πρωτεϊνών, αναμένεται ότι η αποσαφήνιση της μοριακής δομής των διαμεμβρανικών πρωτεϊνών σε ατομική διακριτικότητα θα παραμείνει δύσκολη πρόκληση για τη δομική μοριακή βιολογία (Kyogoku et al., 2003; Loll, 2003; Walian, Cross, & Jap, 2004). Γενικά, ενώ είναι αποδεκτό πλέον ότι οι διαμεμβρανικές πρωτεΐνες αποτελούν περίπου το 25-30% του γονιδιώματος των οργανισμών όλων των εξελικτικών βαθμίδων (Chen & Rost, 2002; Pasquier, Promponas, & Hamodrakas, 2001), οι διαθέσιμες μοναδικές δομές των διαμεμβρανικών πρωτεϊνών είναι περίπου 500, αποτελώντας έτσι ένα πολύ μικρό μόνο ποσοστό (<1%) των πρωτεϊνών με κρυσταλλογραφικά λυμένη δομή (Tusnady, Dosztanyi, & Simon, 2004).

Τα βασικά προβλήματα, που απαντώνται στην προσπάθεια επίλυσης της δομής μιας διαμεμβρανικής πρωτεΐνης, είναι συνυφασμένα με τον κατά βάση υδρόφοβο χαρακτήρα αυτής. Έτσι προσπάθειες αποδιάταξης της μεμβράνης με απορρυπαντικά, έχουν ως συνέπεια την αδυναμία περαιτέρω διαλυτοποίησης της πρωτεΐνης, με τελικό αποτέλεσμα να μην είναι δυνατή η κρυστάλλωση της. Πρόσφατες έρευνες, έδειξαν ότι η πρόοδος στην επίλυση των δομών των διαμεμβρανικών πρωτεϊνών ακολουθεί εκθετική αύξηση, παρόμοια με την αύξηση που παρατηρείται εδώ και 40 χρόνια από τότε που προσδιορίστηκε η πρώτη δομή μιας σφαιρικής υδατοδιαλυτής πρωτεΐνης (White, 2004). Αναμένουμε, λοιπόν, μεγάλη αύξηση του αριθμού των διαμεμβρανικών πρωτεϊνών με γνωστή δομή μέσα στα επόμενα χρόνια, αλλά λόγω του ότι η καθυστέρηση στον προσδιορισμό της πρώτης δομής μεμβρανικής πρωτεΐνης ήταν περίπου 20 χρόνια σε σχέση με τις σφαιρικές υδατοδιαλυτές, το χάσμα ανάμεσα στις γνωστές δομές των πρωτεϊνών των δυο κατηγοριών ίσως να μην καλυφθεί ποτέ. Με βάση τα παραπάνω, γίνεται εμφανές πόσο σημαντική είναι η ανάγκη ύπαρξης αυτοματοποιημένων αλγορίθμων, μέσω των οποίων θα μπορούμε εύκολα και με μεγάλη ακρίβεια να προσδιορίζουμε την πιθανή δομή μιας διαμεμβρανικής πρωτεΐνης.

Οι πρώτοι αλγόριθμοι πρόγνωσης της τοπολογίας των α-ελικοειδών μεμβρανικών πρωτεϊνών βασίστηκαν σε κινούμενα παράθυρα κατά μήκος της αμινοξικής αλληλουχίας. Αρχικά γινόταν χρήση παραθύρων σε συνδυασμό με κάποια κλίμακα υδροφοβικότητας αλλά και με τον κανόνα positive-inside. Έτσι, ένας από τους πρώτους αλγόριθμους πρόγνωσης ήταν το **TopPred** (Claros & von Heijne, 1994)(διαθέσιμο στη διεύθυνση http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::toppred). Το **TMpred** (http://www.ch.embnet.org/software/TMPRED_form.html) ήταν επίσης ένας από τους αρχικούς αλγόριθμους που βασιζόταν σε στατιστικές προτιμήσεις για την εμφάνιση των αμινοξέων. Την ίδια εποχή, εμφανίστηκε και το **MEMSAT** (το οποίο βέβαια έχει εξελιχθεί από τότε), που στηριζόταν σε ένα log-odds score βασισμένο σε στατιστικές προτιμήσεις αμινοξέων και βελτιστοποιούσε τα αποτελέσματα με χρήση δυναμικού προγραμματισμού (http://bioinf.cs.ucl.ac.uk/?id=756). Το **PRED-TMR** ήταν επίσης μια παρόμοια μέθοδος που αναπτύχθηκε λίγο αργότερα, από Έλληνες επιστήμονες (Pasquier et al., 1999)(διαθέσιμο στη διεύθυνση http://athina.biol.uoa.gr/PRED-TMR/) ενώ, καθώς προέβλεπε μόνο την παρουσία των διαμεμβρανικών ελίκων, έπρεπε να συνδυαστεί με έναν άλλον αλγόριθμο, το **orienTM** (http://athina.biol.uoa.gr/orienTM/), το οποίο βασιζόταν επίσης σε στατιστικές προτιμήσεις των αμινοξέων για να προβλέψει τη διευθέτηση των ήδη προβλεφθέντων διαμεμβρανικών περιοχών (Liakopoulos, Pasquier, & Hamodrakas, 2001). Η πρώτη προσπάθεια εφαρμογής Νευρωνικών Δικτύων, συνδυασμένη με πληροφορία από πολλαπλές στοιχίσεις, έγινε το 1996 με το **PHDtm** (www.predictprotein.org), ενώ από τότε έχουν εμφανιστεί πολλοί παρόμοιοι αλγόριθμοι. Ο πρώτος αλγόριθμος βασισμένος σε HMM εμφανίστηκε το 1998 (Sonnhammer, von Heijne, & Krogh, 1998), είναι το **TMHMM** (http://www.cbs.dtu.dk/services/TMHMM/) και θεωρείται ακόμα και σήμερα, ένας από τους καλύτερους αλγορίθμους της κατηγορίας (τουλάχιστον όσον αφορά τους αλγορίθμους

που βασίζονται μόνο στην αμινοξική αλληλουχία). Παρόμοιος αλγόριθμος, αν και κάπως διαφορετικός στην υλοποίηση του μοντέλου είναι το **HMMTOP** (http://www.enzim.hu/hmmtop/) (Tusnady & Simon, 2001). Ένας από τους πρώτους αλγόοριθμους, που χρησιμοποίησαν συνδυαστική πρόγνωση, ήταν το **CoPreThi** (http://athina.biol.uoa.gr/CoPreTHi/) που αναπτύχθηκε στην Ελλάδα και βασιζόταν στους διαθέσιμους εκείνη την εποχή αλγόριθμους SOSUI, Tmpred, ISREC, DAS, TopPred, PHDtm και PRED-TMR (Promponas, Palaios, Pasquier, Hamodrakas, & Hamodrakas, 1999).

Μια άλλη μεγάλη κατηγορία μεθόδων που έκαναν την εμφάνισή τους, ειδικά βασισμένοι σε χρήση των HMM, ήταν οι μέθοδοι που έκαναν ταυτόχρονη πρόγνωση των διαμεμβρανικών τμημάτων και των πεπτιδίων οδηγητών. Η βάση αυτής της μεθοδολογίας βρισκόταν στην παρατήρηση ότι τα αμινοτελικά πεπτίδια οδηγητές (βλ. επόμενη ενότητα) έχουν μια μεγάλη υδρόφοβη περιοχή που μοιάζει με διαμεμβρανική α-έλικα, και κατά συνέπεια πολλοί αλγόριθμοι πρόγνωσης των διαμεμβρανικών τμημάτων τα μπερδεύουν με διαμεμβρανικές περιοχές. Η πρώτη μέθοδος που έκανε αυτή την επέκταση ήταν το **Phobius** (Kall, Krogh, & Sonnhammer, 2004) (διαθέσιμο στη διεύθυνση http://phobius.sbc.su.se/), ενώ αργότερα εμφανίστηκε και το **SPOCTOPUS** (http://octopus.cbr.su.se/index.php?about=SPOCTOPUS). Μια άλλη παρόμοιας φύσεως επέκταση, έχει να κάνει με την ταυτόχρονη πρόγνωση τόσο των διαμεμβρανικών περιοχών όσο και των θέσεων μετα-μεταφραστικών τροποποιήσεων. Οι τροποποιήσεις αυτές, έχουν ειδική στόχευση στην αλληλουχία, αλλά συμβαίνουν και σε διακριτά τμήματα του κυττάρου. Έτσι, μια πρόγνωση για γλυκοζυλίωση μπορεί να βοηθήσει και την πρόγνωση των διαμεμβρανικών τμημάτων καθώς η γλυκοζυλίωση γίνεται σε περιοχές της πρωτεΐνης που βρίσκονται εκτεθειμένες στον εξωκυττάριο χώρο. Αντίθετα, οι θέσεις φωσφορυλίωσης βρίσκονται πάντα στην πλευρά που βρίσκεται στο κυτταρόπλασμα. Η μόνη μέθοδος που προσφέρει μέχρι στιγμής αυτή τη δυνατότητα, είναι το **HMMpTM** (http://bioinformatics.biol.uoa.gr/HMMpTM), το οποίο με αυτόν τον τρόπο πετυχαίνει βελτιωμένη πρόγνωση τόσο στην περίπτωση της διαμεμβρανικής τοπολογίας, όσο και στην περίπτωση των θέσεων γλυκοζυλίωσης και φωσφορυλίωσης (Tsaousis, Bagos, & Hamodrakas, 2014).

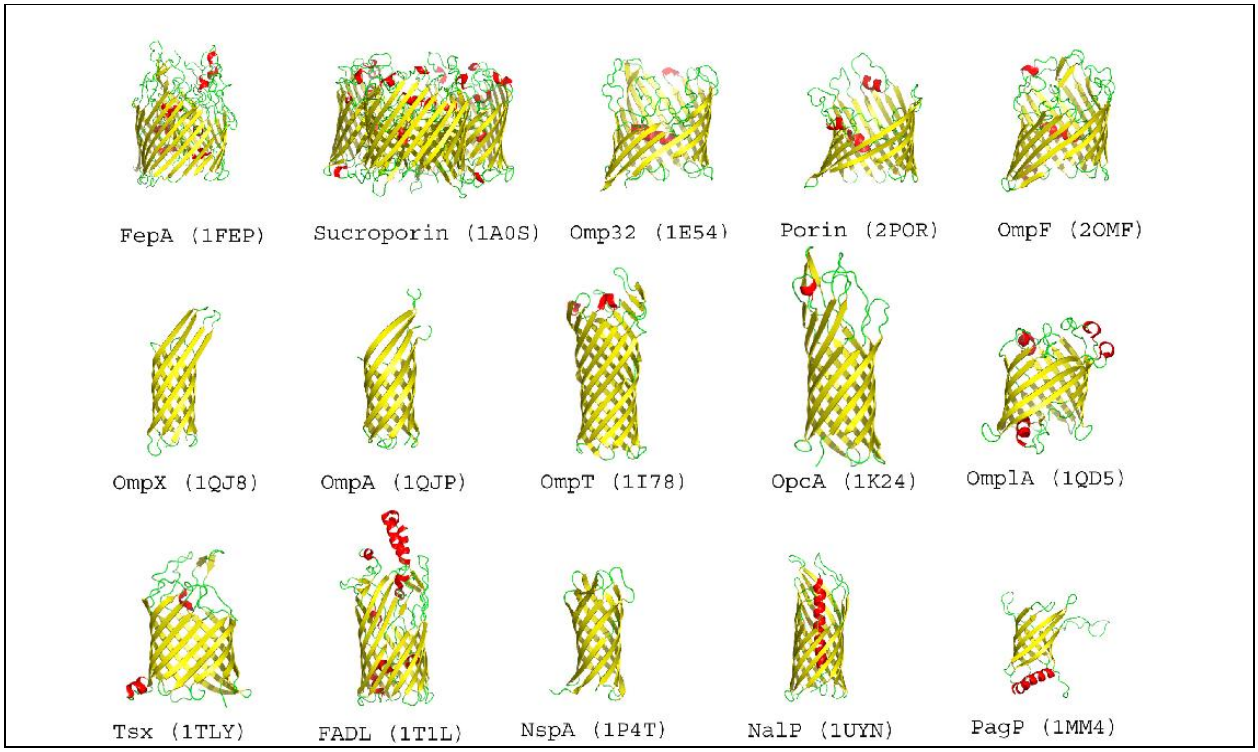
Μια άλλη μεγάλη πρόοδος που έγινε στην περίπτωση πρόγνωσης των διαμεμβρανικών α-ελίκων, έχει να κάνει με την ενίσχυση της απόδοσης της πρόγνωσης όταν γίνει ενσωμάτωση πειραματικής πληροφορίας. Στην περίπτωση διαμεμβρανικών πρωτεϊνών, είναι γνωστό ότι η ενσωμάτωση μιας, ακόμα και περιορισμένης πειραματικά, προσδιορισμένης πληροφορίας σχετικά με την τοπολογία θα βελτίωνε κατά ένα μεγάλο μέρος την απόδοση ακόμα και των καλύτερων μεθόδων. Με την ανάπτυξη εύκολων και γρήγορων πειραματικών τεχνικών βασισμένων σε συντήξεις γονιδίων (gene fusions), με τις οποίες καθορίζεται η θέση του αμινοτελικού άκρου μιας πρωτεΐνης, προτάθηκε ότι (αυτές οι τεχνικές) συνδυαζόμενες θα βελτιώσουν κατά ένα μεγάλο μέρος την απόδοση των προγνωστικών μεθόδων και την εφαρμογή τους σε πλήρως προσδιορισμένα γονιδιώματα (Drew et al., 2002; Melen, Krogh, & von Heijne, 2003). Υπάρχουν αρκετά δεδομένα στην βιβλιογραφία τα οποία δείχνουν και άλλους εναλλακτικούς τρόπους προσδιορισμού της θέσης διαφόρων τμημάτων της ακολουθίας (αντισώματα, πρωτεόλυση κλπ), αλλά οι πιο ολοκληρωμένες πειραματικές αποδείξεις σε μεγάλη κλίμακα γι' αυτήν την βελτίωση, ήρθαν από μελέτες που αφορούν πρωτεΐνες της *E. coli* (Rapp et al., 2004) και του *S. cerevisiae* (Kim, Melen, & von Heijne, 2003).

Από τις ήδη διαθέσιμες προγνωστικές μεθόδους, το TMHMM και το HMMTOP (Tusnady & Simon, 2001), προσφέρουν στο χρήστη την επιλογή να ενσωματώσει στην πρόγνωσή του, πειραματικά προσδιορισμένη πληροφορία για την τοπολογία. Παρόμοια επιλογή, προσφέρεται και από την συνδυασμένη πρόγνωση διαμεμβρανικών α-ελίκων και πεπτιδίων οδηγητών, με τη μέθοδο **Phobius** (Kall et al., 2004). Το **HMM-TM** το οποίο αναπτύχθηκε από την ομάδα μας (http://bioinformatics.biol.uoa.gr/HMM-TM/), ήταν η πρώτη μέθοδος που ενσωμάτωνε τέτοιου είδους πληροφορία σε κάθε αλγόριθμο αποκωδικοποίησης των HMM, ενώ παράλληλα έδινε και τη θεωρητική τεκμηρίωση για αυτήν την τροποποίηση.

Εκτεταμένες εμπειρικές αναλύσεις έχουν δείξει ότι οι μέθοδοι που βασίζονται σε κάποια γραμματική δομή, όπως τα HMM, είναι κατά κανόνα καλύτερες για την πρόγνωση των διαμεμβρανικών α-ελίκων σε σχέση με τις πιο απλές στατιστικές μεθόδους, αλλά και σε σχέση με τα Νευρωνικά Δίκτυα. Επίσης, τόσο ο συνδυασμός πολλών μεθόδων, όσο και η χρήση εξελικτικής πληροφορίας με τη μορφή πολλαπλών στοιχίσεων, είναι παράγοντες που αυξάνουν σημαντικά την απόδοση των μεθόδων αυτών. Έτσι, ακόμα και οι αρχικά ιδιαίτερα επιτυχημένοι αλγόριθμοι, όπως το TMHMM, φάνηκε ότι βελτιώνονται με την προσθήκη πολλαπλών στοιχίσεων σε διάφορες μορφές (**PRO-TMHMM**, **PRODIV-TMHMM**, **S-TMHMM**). Παρόμοιες προσπάθειες, έγιναν και με το Phobius και οδήγησαν στην εμφάνιση του **PolyPhobius** (http://phobius.sbc.su.se/poly.html). Με βάση τα παραπάνω, η καλύτερη σύγχρονη προσέγγιση θα ήταν η χρησιμοποίηση κάποιου αλγόριθμου που συνδυάζει επιτυχημένους αλγορίθμους και ταυτόχρονα κάνει χρήση

εξελικτικής πληροφορίας από πολλαπλές στοιχίσεις. Το πιο πρόσφατο τέτοιο παράδειγμα, είναι το **TOPCONS** (http://topcons.net), το οποίο κάνει χρήση των αλγορίθμων **PolyPhobius**, **OCTOPUS**, **SPOCTOPUS** και **SCAMPI** (οι οποίοι, όλοι κάνουν χρήση εξελικτικής πληροφορίας), αλλά και το **Philius** το οποίο κάνει χρήση μόνο της αλληλουχίας, αλλά στη συνδυαστική πρόγνωση χρησιμοποιείται με τον τρόπο που περιγράψαμε προηγουμένως καθώς οι ομόλογες εντοπίζονται και ενσωματώνονται από τη συνδυαστική μέθοδο. Έτσι, το TOPCONS πετυχαίνει σήμερα, ίσως τις καλύτερες επιδόσεις σε σχέση με τον ανταγωνισμό.

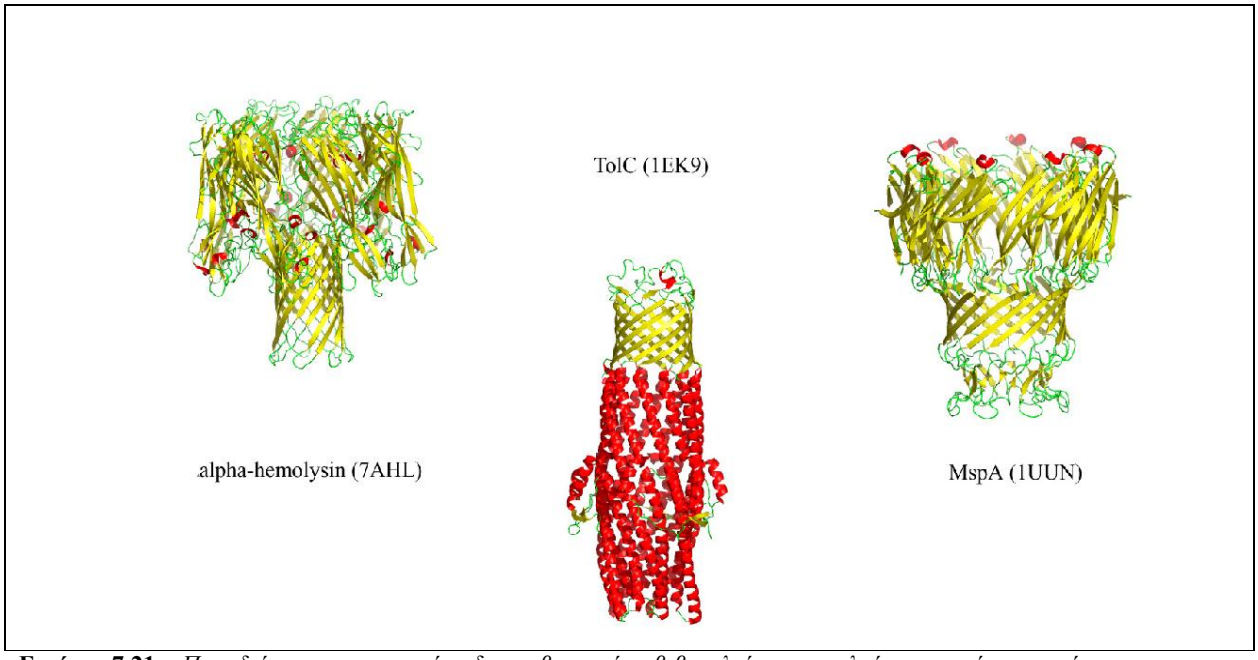
Το β-βαρέλι γενικά, είναι μια β-πτυχωτή επιφάνεια (πτυχωτό φύλλο) που περιελίσσεται και αναδιπλώνεται σχηματίζοντας μια κλειστή δομή σε σχήμα βαρελιού, η οποία σταθεροποιείται από δεσμούς υδρογόνου που σχηματίζονται από την κύρια αλυσίδα. Τα γνωστά παραδείγματα διαμεμβρανικών β-βαρελιών δείχνουν προτίμηση στην ευθυγράμμιση του άξονα του βαρελιού με το κάθετο, στην μεμβράνη, επίπεδο. Επιπλέον, όλες οι γνωστές δομές φαίνεται να απαρτίζονται από γειτονικούς αντιπαράλληλους β-κλώνους που σχηματίζουν μαίανδρο, γεγονός που υποδηλώνει ότι η επαναλαμβανόμενη δομική μονάδα είναι η β-φουρκέτα (β-hairpin). Σήμερα, οι διαθέσιμες δομές υψηλής ανάλυσης διαμεμβρανικών β-βαρελιών περιέχουν βαρέλια διαφόρων μεγεθών και χαρακτηριστικών, με το n να παίρνει τιμές από $8 \leq n \leq 26$ και το S, από $8 \leq S \leq 24$ (Schulz, 2003). Στην περιοχή αυτή των τιμών, αναμένουμε να βρούμε βαρέλια των οποίων οι κλώνοι έχουν μια κλίση σε σχέση με το κατακόρυφο επίπεδο, της τάξης των 30^{o}-60^{o}. Είναι επίσης αξιοσημείωτο, όπως αναφέραμε παραπάνω, το γεγονός ότι σε όλες τις γνωστές δομές, τα διαμεμβρανικά β-βαρέλια αποτελούνται από άρτιο αριθμό β-κλώνων με την εξαίρεση της μοναδικής διαθέσιμης δομής β-βαρελιού από μιτοχόνδρια ευκαρυωτικών οργανισμών, η οποία διαθέτει 19 β-κλώνους.



Εικόνα 7.20: *Μερικά τυπικά παραδείγματα διαμεμβρανικών β-βαρελιών της εξωτερικής μεμβράνης. Με κίτρινο συμβολίζονται οι β-κλώνοι και με κόκκινο οι α-έλικες.*

Σημαντική πρόοδος έχει επιτευχθεί έως σήμερα, στην προσπάθεια κατανόησης της δομής και λειτουργίας των βακτηριακών διαμεμβρανικών β-βαρελιών (Εικόνα 7.20). Παρόλο που αρχικά υπήρχε η εντύπωση ότι οι πρωτεΐνες αυτές ήταν μόνο πόροι (κανάλια) στη μεμβράνη, τα νεότερα δεδομένα δείχνουν ότι εμπλέκονται σχεδόν σε όλες τις διαδικασίες που έχουν εμπλακεί και οι α-ελικοειδείς διαμεμβρανικές πρωτεΐνες. Οι λειτουργικοί τους ρόλοι και οι βιολογικές διεργασίες στις οποίες εμπλέκονται είναι ποικίλοι και ενδέχεται να διαφέρουν από οργανισμό σε οργανισμό. Μεγάλες ευκίνητες στροφές (θηλιές) ανθεκτικές σε πρωτεολυτικά ένζυμα, όπως στην περίπτωση της OmpA (Morona, Kramer, & Henning, 1985) ή σταθερές προεκτάσεις των β-κλώνων που σχηματίζουν το βαρέλι, όπως στην περίπτωση της OmpX (Vogt & Schulz,

1999) όλες στον εξωκυττάριο χώρο, είναι γνωστό ότι παρέχουν θέσεις μοριακής αναγνώρισης. Οι δυο αυτές πρωτεΐνες (OmpA, OmpX), σχηματίζουν βαρέλια με 8 διαμεμβρανικούς κλώνους, αριθμός που θεωρείται ως η ελάχιστη απαίτηση για να μπορέσει να σχηματιστεί βαρέλι από μια και μόνο πολυπεπτιδική αλυσίδα. Ο ακριβής ρόλος της OmpX πιστεύεται ότι είναι η δράση της ως συγκολλητική πρωτεΐνη, ενώ για την OmpA, έχει αναφερθεί ότι με τη μεγάλη καρβοξυτελική περιοχή της συμβάλλει στη σταθερότητα της εξωτερικής μεμβράνης (Ringler & Schulz, 2002), συμμετέχοντας έτσι ως δομική πρωτεΐνη ενώ επιπλέον εμφανίζεται να κατέχει και μικρή ενεργότητα καναλιού (Sugawara & Nikaido, 1992, 1994). Παρόμοιο δίπλωμα εμφανίζουν η NspA (Vandeputte-Rutten, Bos, Tommassen, & Gros, 2003) με 8 διαμεμβρανικά τμήματα, και η OpcA (Prince, Achtman, & Derrick, 2002) με 10 διαμεμβρανικά τμήματα, οι οποίες εμπλέκονται κυρίως σε μολυσματικές διεργασίες μέσω της συγκόλλησης στα κύτταρα του ξενιστή.



Εικόνα 7.21: *Παραδείγματα μη-τυπικών διαμεμβρανικών β-βαρελιών, αποτελούμενα από περισσότερες της μιας πολυπεπτιδικές αλυσίδες. Αριστερά η α-αιμολυσίνη, στο κέντρο η TolC, και δεξιά η MspA.*

Ο διαχωρισμός και η πρόγνωση των διαμεμβρανικών β-βαρελιών, είναι σε γενικές γραμμές πιο δύσκολες διαδικασίες σε σχέση με την πρόγνωση των α-διαμεμβρανικών πρωτεϊνών. Παρ' όλο που οι διαμεμβρανικοί β-κλώνοι σε όλες τις διαθέσιμες δομές τοποθετούνται με σχετικά μεγάλες γωνίες ως προς την λιπιδική διπλοστιβάδα, είναι σημαντικά μικρότεροι από τις διαμεμβρανικές α-έλικες, σε αριθμό αμινοξέων που περιέχουν λόγω της εκτεταμένης διαμόρφωσης, με μήκος που κυμαίνεται από 6 έως 22 αμινοξικά κατάλοιπα. Επιπλέον, είναι αποδεκτό ότι κλώνοι μήκους 7 έως 9 κατάλοιπα είναι ικανοί να διαπεράσουν την λιπιδική διπλοστιβάδα, και καθώς οι β-κλώνοι έρχονται σε επαφή με διαφορετικά μικροπεριβάλλοντα (το υδροφοβικό περιβάλλον της εξωτερικής επιφάνειας του βαρελιού σε αντίθεση με το υδρόφιλο περιβάλλον του υδάτινου πόρου στο εσωτερικό) συχνά συναντάμε εναλλαγές υδρόφοβων-υδρόφιλων καταλοίπων. Αυτή η εναλλαγή δεν είναι πάντα απόλυτη, καθώς τα κατάλοιπα στην εξωτερική επιφάνεια του βαρελιού είναι σχεδόν πάντα υδρόφοβα, αλλά τα κατάλοιπα που αντικρίζουν το εσωτερικό του πόρου μπορεί να μην είναι πάντα πολικά, αλλά να ανήκουν σε άλλες κατηγορίες (π.χ. μπορεί να είναι μικρά ή ουδέτερα).

Παρ' όλο που οι κορυφές στις γραφικές παραστάσεις υδροφοβικότητας συμπίπτουν με τις προγνώσεις των β-πτυχωτών επιφανειών, και συσχετίζονται με την τοποθεσία των διαμεμβρανικών β-πτυχωτών επιφανειών (Zhai & Saier, 2002), η μέση υδροφοβικότητα των τμημάτων αυτών είναι σημαντικά χαμηλότερη από την αντίστοιχη των διαμεμβρανικών α-ελίκων. Το γεγονός αυτό, πρέπει να συνδέεται με τον αντίστοιχο μηχανισμό μετακίνησης, καθώς σε αντίθετη περίπτωση (αν αυτές οι περιοχές ήταν ιδιαίτερα υδρόφοβες), οι πρωτεΐνες της εξωτερικής μεμβράνης, υπήρχε κίνδυνος, να παγιδευτούν στην εσωτερική μεμβράνη κατά τη διάρκεια της μετακίνησης. Επιπλέον, ο ολιγομερισμός των β-βαρελιών, πιθανόν να

εξασθενίζει την ανάγκη για υψηλή υδροφοβικότητα στο εξωτερικό του βαρελιού, καθώς πολικές πλευρικές ομάδες είναι δυνατό να σχηματίζουν ενεργειακά ευνοημένες αλληλεπιδράσεις στην επιφάνεια επαφής.

Ανακεφαλαιώνοντας τα παραπάνω, το σήμα σε επίπεδο ακολουθίας, είναι μάλλον ασθενές για να ανιχνευθεί με απλές στατιστικές αναλύσεις. Επιπλέον, η ύπαρξη κοινών δομικών χαρακτηριστικών με σφαιρικές-υδατοδιαλυτές πρωτεΐνες, οι οποίες έχουν στην τρισδιάστατη δομή τους σχήμα β-βαρελιού, μπορεί να οδηγήσει (την προσπάθεια πρόγνωσης) σε μεγάλο αριθμό ψευδώς θετικών αποτελεσμάτων. Παρ' όλα αυτά προσεκτική παρατήρηση της αμινοξικής ακολουθίας τέτοιων πρωτεϊνών, σε συνδυασμό με τη γνώση της τρισδιάστατης δομής, μπορεί να οδηγήσει σε εξαγωγή κάποιων γενικών κανόνων οι οποίοι θα μπορούν να χρησιμεύσουν σε μια προγνωστική μέθοδο (Schulz, 2002, 2003).

Τέτοια γενικά χαρακτηριστικά είναι:

(1) Οι διαμεμβρανικοί β-κλώνοι είναι κατά βάση αμφιπαθικοί, καθώς εμφανίζουν εναλλαγή υδρόφοβων-πολικών καταλοίπων. Τα υδρόφοβα κατάλοιπα αλληλεπιδρούν με τις υδρόφοβες ουρές των λιπιδίων της μεμβράνης, ενώ τα πολικά στρέφονται προς το εσωτερικό του βαρελιού και άρα αλληλεπιδρούν με το υδάτινο περιβάλλον του πόρου.

(2) Τα αρωματικά κατάλοιπα έχουν την τάση να εμφανίζονται με μεγαλύτερη συχνότητα στις επιφάνειες επαφής με τις πολικές κεφαλές των λιπιδίων, σχηματίζοντας έτσι τις λεγόμενες «αρωματικές ζώνες» στην περιφέρεια του βαρελιού.

(3) Και το αμινοτελικό και το καρβοξυτελικό άκρο των πρωτεϊνών αυτών, είναι τοποθετημένα στον περιπλασμικό χώρο (εσωτερικά σε σχέση με την εξωτερική μεμβράνη). Σε κάποιες περιπτώσεις, μεγάλες αμινοτελικές και καρβοξυτελικές δομικές περιοχές, με μήκος μεγαλύτερο των 100 καταλοίπων, είναι δυνατόν να σχηματίζονται.

(4) Τα τμήματα της ακολουθίας, τα οποία συνδέουν τους διαμεμβρανικούς κλώνους, και τα οποία βρίσκονται στον περιπλασμικό χώρο (εσωτερικές στροφές) είναι γενικά μικρότερου μήκους από τα τμήματα τα οποία βρίσκονται στον εξωκυττάριο χώρο (εξωτερικές θηλιές). Οι στροφές του περιπλασμικού χώρου, σε όλες σχεδόν τις γνωστές δομές, έχουν μήκος 12 ή και λιγότερα κατάλοιπα ενώ αυτές του εξωκυττάριου χώρου μπορεί να έχουν μήκος και πάνω από 30 κατάλοιπα. Αυτό είναι επιτρεπτό λόγω της διαμόρφωσης του μαιάνδρου που υιοθετείται από το β-βαρέλι.

(5) Το μήκος των διαμεμβρανικών β-κλώνων ποικίλει ανάλογα με την κλίση του κλώνου σε σχέση με τον άξονα του βαρελιού και παίρνει τιμές από 6 έως και 22 κατάλοιπα. Παρ' όλα αυτά, σε αρκετές περιπτώσεις, μόνο ένα μικρό τμήμα του κλώνου είναι βυθισμένο στην λιπιδική διπλοστιβάδα, και το υπόλοιπο προεξέχει μακριά από το επίπεδο της μεμβράνης προς τον εξωκυττάριο χώρο, σχηματίζοντας εύκαμπτες φουρκέτες.

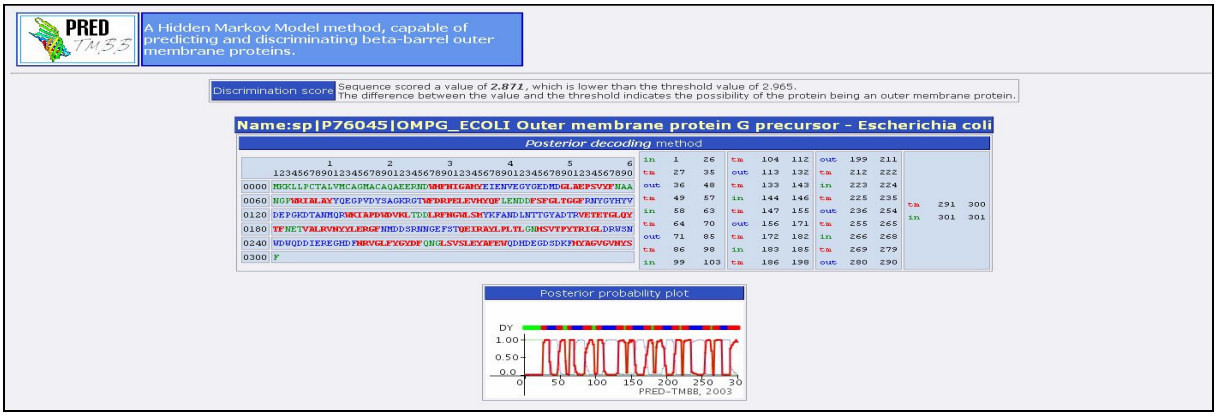
(6) Οι διαμεμβρανικές πρωτεΐνες με μορφή β-βαρελιού εμφανίζουν μικρότερη συντηρητικότητα στις ακολουθίες τους, σε σχέση με τις σφαιρικές-υδατοδιαλυτές πρωτεΐνες. Ακόμα μικρότερη είναι η συντηρητικότητα στις εξωκυττάριες στροφές, οι οποίες δρουν συχνά σαν αντιγονικοί καθοριστές. Το γεγονός αυτό συνεπάγεται, ότι πρωτεΐνες με πολύ μικρή ομοιότητα σε επίπεδο ακολουθίας είναι δυνατόν να διπλώνονται με απολύτως όμοιο τρόπο, αλλά παρ' όλα αυτά οι μέθοδοι αναζήτησης με βάση την ομοιότητα στην ακολουθία να μην μπορούν να τις ανιχνεύσουν.

(7) Οι γειτονικοί β-κλώνοι συνδέονται με ένα δίκτυο δεσμών υδρογόνου, το οποίο σταθεροποιεί τη δομή του βαρελιού.

Οι μέθοδοι πρόγνωσης των διαμεμβρανικών β-βαρελιών, επίσης, διακρίνονται σε μεθόδους που βασίζονται στην υδροφοβικότητα, σε στατιστικές τεχνικές και σε μεθόδους μηχανικής μάθησης. Αξίζει να σημειωθεί, ότι είναι άλλο το πρόβλημα της πρόγνωσης της διαμεμβρανικής τοπολογίας των β-βαρειλιών και άλλο το πρόβλημα του εντοπισμού τους. Κατά συνέπεια, έχουν αναπτυχθεί και διαφορετικές μεθοδολογίες για τις παραπάνω περιπτώσεις, αν και κάποιοι από τους αλγόριθμους αυτούς επιτυγχάνουν και τις δύο λειτουργίες. Η πρώτη προσπάθεια εφαρμογής μεθόδων μηχανικής μάθησης για την πρόγνωση της τοπολογίας των διαμεμβρανικών β-βαρελιών, πραγματοποιήθηκε από τον Diederichs και του συνεργάτες του (Diederichs, Freigang, Umhau, Zeth, & Breed, 1998) αλλά πλέον η μέθοδος αυτή δεν είναι διαθέσιμη. Το **B2TMPRED** που αναπτύχθηκε λίγο αργότερα χρησιμοποίησε Νευρωνικά Δίκτυα με ταυτόχρονη χρήση εξελικτικής πληροφορίας αλλά και επιπλέον φιλτράρισμα των αποτελεσμάτων με αλγόριθμο δυναμικού προγραμματισμού (Jacoboni, Martelli, Fariselli, De Pinto, & Casadio, 2001) και είναι διαθέσιμο στη διεύθυνση http://gpcr.biocomp.unibo.it/cgi/predictors/outer/pred_outercgi.cgi. Σε Νευρωνικά Δίκτυα βασίζονται επίσης και το **TBBpred** (http://www.imtech.res.in/raghava/tbbpred/) και το **TMBETA-NET** (http://psfs.cbrc.jp/tmbeta-net/) τα οποία χρησιμοποιούν μόνο την αμινοξική αλληλουχία, αλλά και το

TMBETAPRED-RBF (http://rbf.bioinfo.tw/~sachen/BARRELpredict/TMBETAPRED-RBF.php) και το **TMBpro** (http://tmbpro.ics.uci.edu/) τα οποία χρησιμοποιούν εξελικτική πληροφορία με τη μορφή πολλαπλών στοιχίσεων.

Οι πρώτες μέθοδοι βασισμένες σε Hidden Markov Model (HMM) εμφανίστηκαν επίσης στις αρχές της δεκαετίας του 2000, και από τότε η μεθοδολογία αυτή έχει κυριαρχήσει (Bagos, Liakopoulos, Spyropoulos, & Hamodrakas, 2004a, 2004b; Bigelow, Petrey, Liu, Przybylski, & Rost, 2004; Hayat & Elofsson, 2012; Liu, Zhu, Wang, & Li, 2003; Martelli, Fariselli, Krogh, & Casadio, 2002; Savojardo, Fariselli, & Casadio, 2013; Singh, Goodman, Walter, Helms, & Hayat, 2011). Η πρώτη μέθοδος ήταν το **HMM-B2TMR**, το οποίο χρησιμοποιούσε πολλαπλές στοιχίσεις αλλά έγινε δημόσια διαθέσιμο αργότερα (http://gpcr.biocomp.unibo.it/predictors/), ενώ πλέον έχει εμφανιστεί και μια συνδυαστική μέθοδος από την ίδια ομάδα, το **BetAware** (http://www.biocomp.unibo.it/~savojard/betawarecl). Το **PRED-TMBB** (http://bioinformatics.biol.uoa.gr/PRED-TMBB/) παρουσιάστηκε λίγο αργότερα από εμάς, και ήταν ιδιαίτερα πετυχημένο, καθώς παρ' όλο που χρησιμοποιούσε μόνο πληροφορία από την αμινοξική αλληλουχία, χρησιμοποίησε ένα διαφορετικό κριτήριο για την εκτίμηση των παραμέτρων του μοντέλου, αλλά και διαφορετικούς αλγόριθμους για την εκπαίδευση και την αποκωδικοποίησή του. Ταυτόχρονα είχε εμφανιστεί το **PROFtmb** (https://www.predictprotein.org/) το οποίο έκανε χρήση εξελικτικής πληροφορίας ενώ αργότερα εμφανίστηκαν και άλλες μέθοδοι, όπως το **TMBHMM** και το **TMBhunt**. Η τελευταία και πιο αξιόπιστη μέθοδος, είναι το **BOCTOPUS** (http://boctopus.cbr.su.se/), το οποίο χρησιμοποιεί ένα συνδυασμό Support Vector Machines και HMMs ενώ κάνει και χρήση εξελικτικής πληροφορίας.

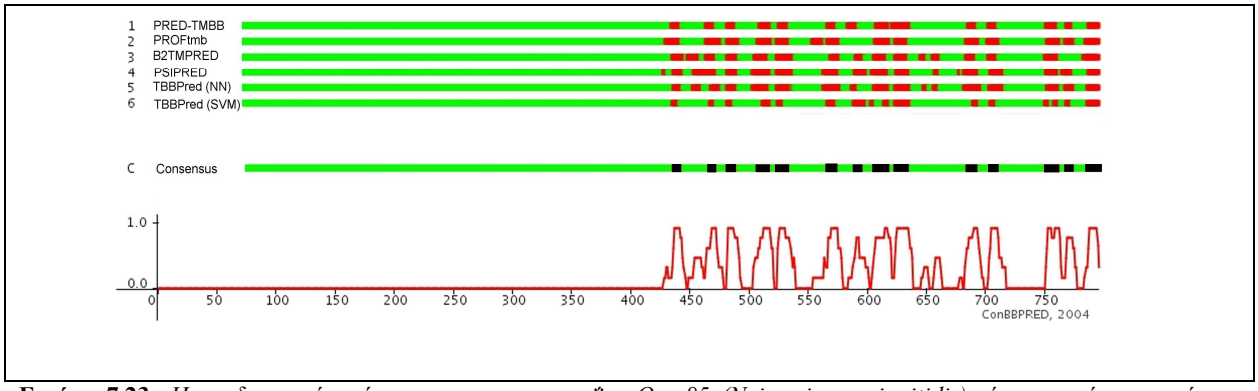


Εικόνα 7.22: *Το αποτέλεσμα που επιστρέφει ο εξυπηρετητής δικτύου του PRED-TMBB για την ακολουθία OMPG_ECOLI. Στο κέντρο φαίνονται με πράσινο χρώμα τα κατάλοιπα του περιπλασμικού χώρου, με κόκκινο τα κατάλοιπα των διαμεμβρανικών β-κλώνων και με μπλε αυτά που προβλέπονται ως εξωκυττάρια.*

Όμοια με τις α-ελικοειδείς μεμβρανικές πρωτεΐνες, εκτεταμένες εμπειρικές αναλύσεις έχουν δείξει ότι οι μέθοδοι που βασίζονται σε κάποια γραμματική δομή όπως τα HMM, είναι κατά κανόνα καλύτερες για την πρόγνωση των διαμεμβρανικών β-βαρελιών σε σχέση με τις πιο απλές στατιστικές μεθόδους, αλλά και σε σχέση με τα Νευρωνικά Δίκτυα. Επίσης, τόσο ο συνδυασμός πολλών μεθόδων όσο και η χρήση εξελικτικής πληροφορίας με τη μορφή πολλαπλών στοιχίσεων είναι παράγοντες που αυξάνουν σημαντικά την απόδοση των μεθόδων αυτών. Με βάση τα παραπάνω, το 2005, παρουσιάσαμε τον μοναδικό μέχρι στιγμής συνδυαστικό αλγόριθμο πρόγνωσης των β-βαρελιών, το **ConBBPRED** (http://bioinformatics.biol.uoa.gr/ConBBPRED/). Το ConBBPRED δίνει τη δυνατότητα στο χρήστη να επιλέξει ποιες μεθόδους θα συμπεριλάβει στη συνδυαστική πρόγνωση ενώ επιπλέον βελτιστοποιεί την τελική πρόγνωση με έναν αλγόριθμο δυναμικού προγραμματισμού. Με τον τρόπο αυτό, η μέθοδος ξεπερνάει σε επιτυχία όλες τις επιμέρους μεθόδους που χρησιμοποιούνται στην πρόγνωση. Το μειονέκτημα της μεθόδου είναι το γεγονός ότι ο χρήστης πρέπει να έχει λάβει μόνος του τα αποτελέσματα από τις επιμέρους μεθόδους και να τα επικολλήσει στην αντίστοιχη φόρμα της διαδικτυακής εφαρμογής (Bagos, Liakopoulos, & Hamodrakas, 2005).

Παρ' όλο που είδαμε ότι ακόμα και για τα β-βαρέλια η αύξηση του μεγέθους του συνόλου εκπαίδευσης δεν οδηγεί σε γραμμική αύξηση της απόδοσης, το μέγεθος παίζει κάποιο ρόλο, ειδικά αν αναλογιστούμε ότι οι πρώτες μέθοδοι ήταν εκπαιδευμένες σε μόλις 10-20 τέτοιες πρωτεΐνες. Έτσι, είναι κατανοητό ότι οι πιο σύγχρονες μέθοδοι όπως το BOCTOPUS, που έχουν εκπαιδευθεί σε μερικές δεκάδες

αλληλουχίες, θα είναι πιο αποδοτικές. Παρ' όλα αυτά, οι αλγοριθμικές επιλογές αλλά και ο σωστός σχεδιασμός του μοντέλου καθιστούν ακόμα και σήμερα το PRED-TMBB μια ιδιαίτερα ανταγωνιστική μέθοδο. Εκτός από το PRED-TMBB και το BOCTOPUS, οι πιο αξιόπιστες μέθοδοι σύμφωνα με τα τελευταία δεδομένα είναι το PROFtmb, το BetAware και το HMM-B2TMR. Μια προσπάθεια να επανεκπαιδευθεί το PRED-TMBB σε νέα δεδομένα αλλά και να χρησιμοποιήσει εξελικτική πληροφορία, έχει δώσει εξαιρετικά μέχρι στιγμής αποτελέσματα και αναμένουμε να δημοσιευτεί σύντομα. Η μέθοδος αυτή, το **PRED-TMBB2** (www.compgen.org/tools/PRED-TMBB2), φαίνεται ότι είναι πλέον η πιο αξιόπιστη μέθοδος, ενώ μια νέα εφαρμογή για συνδυαστική πρόγνωση βρίσκεται υπό κατασκευή.



Εικόνα 7.23: *Η συνδυαστική πρόγνωση για την πρωτεΐνη Omp85 (Neisseria meningitidis), όπως προέκυψε από τον αλγόριθμο ConBBPRED (http://bioinformatics.biol.uoa.gr/ConBBPRED). Πάνω: τα αποτελέσματα των έξι διαφορετικών προγνωστικών μεθόδων που χρησιμοποιήθηκαν. Κάτω: το ιστόγραμμα της τελικής πιθανότητας για την ύπαρξη διαμεμβρανικών β-κλώνων κατά μήκος της ακολουθίας (0-1). Κέντρο: η τελική συνδυαστική πρόγνωση (με μαύρο χρώμα), η οποία βελτιστοποιείται με έναν αλγόριθμο δυναμικού προγραμματισμού.*

Από τις μεθόδους πρόγνωσης της τοπολογίας, κάποιες όπως το PRED-TMBB, το BetAware και το TMBETA-NET προσφέρουν και την επιλογή να κάνουν ταυτόχρονα διαχωρισμό και ταξινόμηση, δηλαδή να προβλέψουν αν μια δοθείσα πρωτεΐνη είναι διαμεμβρανικό β-βαρέλι ή όχι. Όπως είπαμε, αυτή η πρόβλεψη δεν είναι κάτι απλό, γιατί πιθανοί διαμεμβρανικοί β-κλώνοι μπορεί να προβλεφθούν και σε μη μεμβρανικές πρωτεΐνες. Επίσης, η μεγάλη δυσκολία του εγχειρήματος έχει να κάνει και με το γεγονός ότι τα β-βαρέλια είναι σπάνιες πρωτεΐνες (~2% στα γονιδιώματα) και κατά συνέπεια ακόμα και μια μέθοδος με ειδικότητα >90% θα δώσει εκ των πραγμάτων πολλά αρνητικά αποτελέσματα. Έτσι, έχουν αναπτυχθεί και ειδικοί αλγόριθμοι (συνήθως βασισμένοι σε ολική κωδικοποίηση της αλληλουχίας), οι οποίοι έχουν βασικό στόχο το διαχωρισμό αυτών των πρωτεϊνών. Η μια μεγάλη ομάδα τέτοιων αλγορίθμων προέρχεται από την ομάδα του Michael Gromiha και βασίζεται σε ολική πληροφορία με χαρακτηριστικό παράδειγμα το **TMBETADISC-RBF** (http://rbf.bioinfo.tw/~sachen/OMPpredict/TMBETADISC-RBF.php), το οποίο έχει μεγάλη ειδικότητα (94%), αλλά χάνει σε ευαισθησία (85%). Το **BOMP** (http://services.cbu.uib.no/tools/bomp) είναι ένας αρκετά παλιός αλλά πετυχημένος αλγόριθμος που κάνει χρήση κανονικών εκφράσεων και υδροφοβικότητας για την πρόγνωση και έχει μεγάλη ειδικότητα (99%), αλλά χάνει σε ευαισθησία (~68%). Το **PSORTb** (http://www.psort.org/psortb/) είναι ένα γενικότερο εργαλείο πρόγνωσης της υποκυτταρικής θέσης των πρωτεϊνών στα βακτήρια, που μεταξύ άλλων προβλέπει και σαν θέση την εξωτερική μεμβράνη. Κάνει χρήση πολλών εργαλείων πρόγνωσης και συνδυάζει αποτελέσματα από πρότυπα κανονικών εκφράσεων, εμφανίσεις διπεπτιδίων κλπ ενώ η τελική απόφαση βγαίνει από έναν αλγόριθμο μηχανικής μάθησης. Παρ' όλα αυτά, είναι ιδιαίτερα ειδικός (~99.5%) αλλά χάνει σε ευαισθησία (~50%). Το **β-barrel analyzer** (http://beta-barrel.tulane.edu/FW_analysis.php) των Freeman-Wimley, είναι ένα πρόσφατο και ιδιαίτερα καλό εργαλείο που πετυχαίνει καλό συνδυασμό ευαισθησίας (86%) και ειδικότητας (95%). Τέλος, το **HHomp** (http://toolkit.tuebingen.mpg.de/hhomp) είναι ίσως ο καλύτερος αλγόριθμος, καθώς στηρίζεται σε σύγκριση HMM-HMM κατασκευασμένων από τις πρωτεΐνες με γνωστή δομή (στην ουσία κάνει αναγνώριση μακρινών ομολόγων με έναν παρόμοιο τρόπο που θα ξανασυναντήσουμε στο κομμάτι της ύφανσης). Το μεγάλο του μειονέκτημα, που το καθιστά δύσχρηστο σε πραγματικά προβλήματα και αναλύσεις γονιδιωμάτων, είναι ότι είναι ιδιαίτερα αργό λόγω της μεθοδολογίας που χρησιμοποιεί.

7.6.3. Σηματοδοτικές αλληλουχίες και κυτταρική στόχευση

Ένα άλλο πολύ σημαντικό θέμα είναι η πρόβλεψη της στόχευσης των πρωτεϊνών, δηλαδή του προορισμού τους μέσα στο κύτταρο (υποκυτταρική τοποθεσία ή στόχευση). Είναι γνωστό εδώ και δεκαετίες, ότι η πληροφορία για τη στόχευση αυτή βρίσκεται κωδικοποιημένη στην ίδια την αλληλουχία των πρωτεϊνών, τις περισσότερες φορές με τη μορφή μιας αμινοτελικής ή καρβοξυτελικής αλληλουχίας. Σε όλους τους οργανισμούς (Βακτήρια, Αρχαία και Ευκαρυωτικούς), η πλειοψηφία των εκκρινόμενων πρωτεϊνών συντίθεται σαν ένα πρόδρομο μόριο το οποίο φέρει μια αμινοτελική αλληλουχία, η οποία κατευθύνει την έκκριση και μετά αποκόπτεται (αυτή η αλληλουχία ονομάζεται σηματοδοτική αλληλουχία, ή πεπτίδιο οδηγητής). Αυτό το πεπτίδιο, διαθέτει μια σπονδυλωτή δομή με φορτισμένα αμινοξέα στο αμινοτελικό άκρο (n-region), μια υδρόφοβη περιοχή (h-region) η οποία διαπερνά τη μεμβράνη και μια άλλη περιοχή (c-region) η οποία αποτελείται κυρίως από μικρά και μη φορτισμένα κατάλοιπα, η οποία τελειώνει σε μια χαρακτηριστική αλληλουχία αποκοπής (συνήθως με το πρότυπο A-X-A), που αναγνωρίζεται από ειδικό ένζυμο, την πεπτιδάση το σήματος (von Heijne, 1990). Ο μηχανισμός ο οποίος είναι απαραίτητος για τη στόχευση των πρωτεϊνών στο μεμβρανικό σύστημα έκκρισης, είναι παρόμοιος, τόσο στα Βακτήρια (Driessen & Nouwen, 2007), όσο και στους Ευκαρυωτικούς οργανισμούς (Rapoport, Matlack, Plath, Misselwitz, & Staeck, 1999), αλλά και στα Αρχαία (Pohlschroder, Gimenez, & Jarrell, 2005). Μετά τη μεταφορά κατά μήκος της μεμβράνης, το πεπτίδιο οδηγητής αποκόπτεται από την πρόδρομη πρωτεΐνη, με τη χρήση μιας προσδεδεμένης στη μεμβράνη πεπτιδάσης του σήματος (Tuteja, 2005; van Roosmalen et al., 2004). Στους Ευκαρυωτικούς οργανισμούς, οι περισσότερες πρωτεΐνες που κατευθύνονται στα μιτοχόνδρια και τους χλωροπλάστες (αλλά όχι όλες), περιέχουν επίσης αμινοτελικές σηματοδοτικές αλληλουχίες, οι οποίες αποκόπτονται μετά τη μεταφορά, αν και τα γενικά χαρακτηριστικά τους είναι αρκετά διαφορετικά, τόσο όσον αφορά στο μήκος αλλά και όσον αφορά τη σύσταση και την υδροφοβικότητα τους (Habib, Neupert, & Rapaport, 2007; G. von Heijne, Steppuhn, & Herrmann, 1989). Άλλες περιπτώσεις στόχευσης, όπως των πρωτεϊνών του πυρήνα και των υπεροξεισωμάτων, ελέγχονται με διαφορετικό τρόπο. Οι πρωτεΐνες που εισάγονται στον πυρήνα περιέχουν εσωτερικά σήματα αποτελούμενα από μικρές αλληλουχίες πλούσιες σε Αργινίνη και Λυσίνη, ενώ για τα υπεροξεισώματα έχουν βρεθεί δύο μηχανισμοί, ένας που μεσολαβείται με τη δράση καρβοξυτελικών αλληλουχιών (PTS1), και ένας τελείως διαφορετικός, ο οποίος λειτουργεί μέσω αμινοτελικών αλληλουχιών (PTS2).

Τα Βακτήρια, τα Αρχαία και οι χλωροπλάστες διαθέτουν, εκτός από το γενικό εκκριτικό μηχανισμό που περιγράψαμε παραπάνω (Sec), και ένα άλλο σύστημα βασισμένο στο μεταφορέα των δίδυμων αργινινών (Twin-Arginine translocase - Tat). Το σύστημα Tat αναγνωρίζει ελαφρώς μεγαλύτερα και λιγότερο υδρόφοβα πεπτίδια οδηγητές, τα οποία φέρουν μια χαρακτηριστική αλληλουχία από δυο συνεχόμενες αργινίνες (RR) στο n-region (Berks, Palmer, & Sargent, 2005; Lee, Tullman-Ercek, & Georgiou, 2006; Teter & Klionsky, 1999). Μια βασική διαφορά μεταξύ των μονοπατιών Sec και Tat, βρίσκεται στο γεγονός ότι το πρώτο μεταφέρει τις πρωτεΐνες μη διπλωμένες κατά μήκος του καναλιού της μεμβράνης, ενώ στο δεύτερο σύστημα οι πρωτεΐνες μεταφέρονται με έναν άγνωστο προς το παρόν μηχανισμό, αφού έχουν διπλωθεί στην τελική τρισδιάστατη δομή τους (Teter & Klionsky, 1999). Στις περισσότερες εκκρινόμενες πρωτεΐνες (είτε αυτές εκκρίνονται με Sec, είτε με Tat), τα πεπτίδια οδηγητές αποκόπτονται από την πεπτιδάση του σήματος I (Spase I), η οποία αναγνωρίζει πεπτίδια που μοιάζουν αρκετά με αυτά των Ευκαρυωτικών οργανισμών. Επιπλέον όμως, τα Βακτήρια και τα Αρχαία, εκτός από τους δύο παραπάνω μηχανισμούς μεταφοράς, έχουν και έναν δεύτερο μηχανισμό για την αποκοπή των πεπτιδίων οδηγητών. Συγκεκριμένα, υπάρχει η πεπτιδάση του σήματος II (Spase II or Lsp), η οποία είναι ειδική για τις προσδεδεμένες στη μεμβράνη λιποπρωτεΐνες. Το πεπτίδιο οδηγητής των λιποπρωτεϊνών, έχει ακριβώς τα ίδια χαρακτηριστικά με το εκκριτικό πεπτίδιο οδηγητή, με την κύρια διαφορά να εντοπίζεται στην c-region (lipobox), η οποία χαρακτηρίζεται από μια συντηρημένη C, η οποία είναι και απαραίτητη για τη χημική τροποποίηση που θα οδηγήσει στην ομοιοπολική πρόσδεση στα λιπίδια της μεμβράνης. Το πρότυπο που εμφανίζεται σε αυτή την περιοχή μπορεί να χαρακτηριστεί από το μοτίβο [LVI]-[AST]-[GA]-C, αλλά και άλλα παρόμοια μοτίβα που έχουν περιγραφεί κατά καιρούς. Επιπλέον δε, τα τελευταία χρόνια έχουν παρατηρηθεί και λιποπρωτεΐνες που εκκρίνονται με το σύστημα Tat. Καταλαβαίνουμε δηλαδή, ότι το σύστημα είναι τελείως σπονδυλωτό, καθώς οι διαφορετικές περιοχές των πεπτιδίων οδηγητών μπορούν να συνδυαστούν με διαφορετικό τρόπο.

Η υπολογιστική πρόγνωση των πεπτιδίων οδηγητών, αλλά και των άλλων σηματοδοτικών αλληλουχιών, ήταν ένα σημαντικό πρόβλημα, ήδη από τη δεκαετία του 1980. Αρχικά χρησιμοποιήθηκαν weight matrices βασισμένοι στην ανάλυση του Gunnar von Heijne (von Heijne, 1986), και ο πιο γνωστός

αλγόριθμος που βασίζεται σε αυτή τη μέθοδο, είναι το **SigCleave,** το οποίο υπάρχει διαθέσιμο σε πολλές εκδόσεις (http://emboss.sourceforge.net/apps/release/6.6/emboss/apps/sigcleave.html). Μια πιο σύγχρονη μέθοδος βασισμένη σε weight matrices, η οποία έχει εκπαιδευθεί σε περισσότερα και καλύτερης ποιότητας δεδομένα, είναι το **PrediSi** (http://www.predisi.de/). Η μέθοδος αυτή, όπως και οι περισσότερες σύγχρονες μέθοδοι, έχει διαφορετικές εκδόσεις για τις τρεις μεγάλες κατηγορίες οργανισμών (Ευκαρυωτικοί, αρνητικά κατά Gram Βακτήρια, θετικά κατά Gram Βακτήρια). Οι πιο αποδοτικές όμως σύγχρονες μεθοδολογίες, βασίζονται σε μεθόδους μηχανικής μάθησης όπως τα νευρωνικά δίκτυα και τα HMM. Η πιο καλή και η πιο γνωστή από τις σύγχρονες μεθόδους, είναι το **SignalP** (http://www.cbs.dtu.dk/services/SignalP), το οποίο έχει φτάσει ήδη την έκδοση 4.1, και εκτός του ότι διαθέτει ξεχωριστά εργαλεία για την κάθε ομάδα οργανισμών και δυο διαφορετικές μεθόδους (νευρωνικά δίκτυα και HMM), ενώ βασίζεται στην εξαιρετική βιβλιογραφική αναζήτηση για την κατάρτιση του συνόλου εκπαίδευσης, περιλαμβάνοντας έτσι πολλές πρωτεΐνες, αλλά και απομακρύνοντας λάθος καταχωρίσεις (Bendtsen, Nielsen, von Heijne, & Brunak, 2004). Όπως ήδη αναφέραμε, κάποιες μέθοδοι πρόγνωσης διαμεμβρανικών πρωτεϊνών διαθέτουν επιπλέον την ικανότητα να προβλέπουν τα πεπτίδια οδηγητές. Οι μέθοδοι αυτές είναι το **Phobius**, διαθέσιμο στη διεύθυνση http://phobius.sbc.su.se/ (Kall et al., 2004; Kall, Krogh, & Sonnhammer, 2007) και το **Philius** (Reynolds, Kall, Riffle, Bilmes, & Noble, 2008), το οποίο είναι διαθέσιμο στη διεύθυνση http://noble.gs.washington.edu/proj/philius/, οι οποίες χρησιμοποιούν γραφικά μοντέλα (HMM και Bayesian network, αντίστοιχα), ενώ αργότερα εμφανίστηκε και το **SPOCTOPUS** (http://octopus.cbr.su.se/index.php?about=SPOCTOPUS).

Ειδικά για τα πεπτίδια οδηγητές των Αρχαίων, υπήρχε για χρόνια διαμάχη, σχετικά με το ερώτημα αν τα πεπτίδια αυτής της ομάδας μοιάζουν με τα πεπτίδια κάποιας άλλης ομάδας, και με ποιάς ομάδας μοιάζουν περισσότερο. Το βασικό πρόβλημα, ήταν ότι δεν υπήρχαν πολλά παραδείγματα καλά χαρακτηρισμένων τέτοιων πρωτεϊνών και οι περισσότεροι πρότειναν απλά τη χρήση όλων των διαθέσιμων εργαλείων (αυτά που έχουν αναπτυχθεί για τις άλλες ομάδες οργανισμών). Παρ' όλα αυτά, μια εκτεταμένη αναζήτηση στη βιβλιογραφία, μας οδήγησε σε ένα μεγάλο αριθμό τέτοιων πρωτεϊνών, οι οποίες αν και καλά χαρακτηρισμένες στη βιβλιογραφία, δεν είχαν αντίστοιχη πληροφορία στην Uniprot (Bagos, Tsirigos, Plessas, Liakopoulos, & Hamodrakas, 2009). Έτσι, η ανάλυσή μας έδειξε ότι τα πεπτίδια οδηγητές των Αρχαίων μοιάζουν περισσότερο με τα αντίστοιχα των θετικών κατά Gram βακτηρίων, ενώ με το νέο σύνολο εκπαίδευσης κατασκευάσαμε τη μοναδική μέχρι στιγμής διαθέσιμη μέθοδο για τα Αρχαία, το **PRED-SIGNAL** (http://www.compgen.org/tools/PRED-SIGNAL).

Οι βακτηριακές λιποπρωτεΐνες για πολλά χρόνια αναγνωρίζονταν με χρήση κανονικών εκφράσεων της PROSITE, όπως αυτές που αναφέραμε στο κεφάλαιο 5 (π.χ. το PS00013). Παρ' όλα αυτά, τα τελευταία χρόνια αναπτύχθηκαν και για αυτές τις πρωτεΐνες πιο σύγχρονες μέθοδοι. Αρχικά αναπτύχθηκε το **LipoP** (http://www.cbs.dtu.dk/services/LipoP), το οποίο βασίστηκε σε HMM και είχε εκπαιδευθεί να αναγνωρίζει λιποπρωτεΐνες από αρνητικά κατά Gram βακτήρια (Juncker et al., 2003). Το LipoP έχει επιπλέον την ειδική ικανότητα να προβλέπει εξίσου καλά και πεπτίδια οδηγητές εκκρινόμενων πρωτεϊνών, αλλά και διαμεμβρανικές έλικες στο αμινοτελικό άκρο και έχει μια επιτυχία της τάξης του 97% στη σωστή ταξινόμηση στις λιποπρωτεΐνες από αρνητικά κατά Gram βακτήρια, ενώ δίνει λάθος προβλέψεις (δηλαδή, σε μη εκκρινόμενες πρωτεΐνες), της τάξης του 0.3%. Παρ' όλα αυτά, όταν χρησιμοποιηθεί σε λιποπρωτεΐνες από θετικά κατά Gram βακτήρια, η ακρίβειά του πέφτει περίπου στο 90-92%. Έτσι, σε μια παλιότερη εργασία μας, αφού πραγματοποιήσαμε εκτεταμένη αναζήτηση στη βιβλιογραφία για την εύρεση πειραματικά προσδιορισμένων λιποπρωτεϊνών από θετικά κατά Gram βακτήρια, κατασκευάσαμε το **PRED-LIPO** (http://www.compgen.org/tools/PRED-LIPO), το οποίο αποδίδει καλύτερα σε αυτή την κατηγορία βακτηρίων, ενώ παράλληλα προβλέπει με αρκετά μεγάλη ακρίβεια και τα πεπτίδια των εκκρινόμενων πρωτεϊνών, αλλά και τις διαμεμβρανικές έλικες (Bagos, Tsirigos, Liakopoulos, & Hamodrakas, 2008).

Στους ευκαρυωτικούς οργανισμούς, δεν υπάρχουν τέτοιου είδους λιποπρωτεΐνες, αλλά υπάρχει ένας παρόμοιος μηχανισμός για την πρόσδεση πρωτεϊνών στα λιπίδια (GPI-anchor). Οι πρωτεΐνες αυτές κατευθύνονται εκεί, από μια σηματοδοτική αλληλουχία στο καρβοξυτελικό άκρο, η οποία είναι βασικά υδρόφοβη και περιέχει μια ειδική περιοχή αναγνώρισης. Μέθοδοι πρόγνωσης για τις πρωτεΐνες αυτής της κατηγορίας, είναι το **PredGPI** (http://gpcr.biocomp.unibo.it/predgpi/pred.htm), το **big-PI** (http://mendel.imp.ac.at/gpi/gpi_server.html), το **FragAnchor** (http://navet.ics.hawaii.edu/~fraganchor/NNHMM/NNHMM.html) και το **GPI-SOM** (http://gpi.unibe.ch/). Αντίστοιχα στα βακτήρια, υπάρχει ένα σύστημα που αναγνωρίζει μια σηματοδοτική αλληλουχία στο καρβοξυτελικό άκρο και προσδένει την πρωτεΐνη στο κυτταρικό τοίχωμα. Οι πρωτεΐνες αυτές ονομάζονται

συνήθως LPXTG (από το αντίστοιχο πρότυπο πάνω στη σηματοδοτική αλληλουχία που αναγνωρίζει το ειδικό ένζυμο, η σορτάση), και προς το παρόν είναι χαρακτηρισμένες μόνο στα θετικά κατά Gram βακτήρια (αν και υπάρχουν ενδείξεις ότι παρόμοια συστήματα υπάρχουν και στα αρνητικά κατά Gram βακτήρια αλλά και στα αρχαία). Αυτή τη στιγμή, η καλύτερη μέθοδος για την κατηγορία αυτή, το **CW-PRED** (http://bioinformatics.biol.uoa.gr/CW-PRED/), έχει αναπτυχθεί από εμάς, και εκτός από την πρόγνωση κάνει και διαχωρισμό των διαφορετικών ενζύμων (σορτάσες) που αναγνωρίζουν τα υποστρώμματα αυτά.

Παρ' όλο που πολλές από τις μεθόδους που αναφέραμε, μπορούν να προβλέψουν (μέχρι κάποιο βαθμό) και τα πεπτίδια που οδηγούνται μέσω του συστήματος Tat (χωρίς όμως να μπορούν να τα διαχωρίσουν), έχουν αναπτυχθεί τα τελευταία χρόνια και ειδικές μεθοδολογίες που δουλεύουν καλύτερα στις πρωτεΐνες αυτής της κατηγορίας. Η πρώτη τέτοια μέθοδος ήταν το **TATFIND** (http://signalfind.org/tatfind.html), το οποίο βασιζόταν σε ανάλυση υδροφοβικότητας και σε κανονικές εκφράσεις (Rose, Bruser, Kissinger, & Pohlschroder, 2002). Λίγα χρόνια αργότερα εμφανίστηκε το **TatP** (http://www.cbs.dtu.dk/services/TatP/), το οποίο χρησιμοποιεί νευρωνικά δίκτυα αλλά και κανονικές εκφράσεις για να διακρίνει την περιοχή RR (Bendtsen, Nielsen, Widdick, Palmer, & Brunak, 2005). Το TatP είναι γενικά αξιόπιστο, αλλά όχι στα επίπεδα του SignalP, ενώ το TATFIND αναγνωρίζει μόνο την ύπαρξη του σήματος RR, αλλά όχι και το σημείο αποκοπής. Σε μια προσπάθεια να επιλύσουμε όλα αυτά τα προβλήματα, παρουσιάσαμε πρόσφατα το **PRED-TAT** (http://www.compgen.org/tools/PRED-TAT/), μια μέθοδο βασισμένη στα HMMs, η οποία μπορεί αφενός μεν να διαχωρίσει τα πεπτίδια οδηγητές (Sec και Tat), αφετέρου δε, να προβλέψει και τις θέσεις αποκοπής στις δύο κατηγορίες. Η μέθοδος αυτή, είναι αυτή τη στιγμή, η κορυφαία για τα Tat πεπτίδια οδηγητές, αλλά ταυτόχρονα προβλέπει και τα κλασικά πεπτίδια (Sec) σε ικανοποιητικό βαθμό, ενώ υστερεί ελάχιστα σε αυτή την κατηγορία σε σχέση με το SignalP (Bagos et al., 2010).

Σχετικά με τις σηματοδοτικές αλληλουχίες που κατευθύνουν τις πρωτεΐνες στα μιτοχόνδρια και τους χλωροπλάστες, έχουν επίσης αναπτυχθεί εξειδικευμένοι αλγόριθμοι. Για τους χλωροπλάστες, ο πιο γνωστός είναι το **ChloroP** (http://www.cbs.dtu.dk/services/ChloroP), ενώ το **TargetP** (http://www.cbs.dtu.dk/services/TargetP), είναι ένα ολοκληρωμένο σύστημα που προβλέπει τόσο τις εκκριτικές πρωτεΐνες, όσο και αυτές των μιτοχονδρίων και των χλωροπλαστών. Παρόμοιας αρχιτεκτονικής και φιλοσοφίας είναι το κάπως παλιότερο **iPSORT** (http://ipsort.hgc.jp/how.html). Άλλα εργαλεία που προβλέπουν τις μιτοχονδριακές σηματοδοτικές αλληλουχίες, είναι το **MitoProt** (https://ihg.gsf.de/ihg/mitoprot.html), το **Predotar** (http://urgi.versailles.inra.fr/predotar/predotar.html), και το **Tppred2** (http://tppred2.biocomp.unibo.it). Για τις πρωτεΐνες των υπεροξεισωμάτων, υπάρχει το **PTS1 predictor** (http://mendel.imp.ac.at/mendeljsp/sat/pts1/PTS1predictor.jsp), ενώ για τις πρωτεΐνες που κατευθύνονται στον πυρήνα έχει αναπτυχθεί το **cNLS Mapper** (http://nls-mapper.iab.keio.ac.jp/cgi-bin/NLS_Mapper_form.cgi), το **NLStradamus** (http://www.moseslab.csb.utoronto.ca/NLStradamus/), το **NucPred** (http://www.sbc.su.se/~maccallr/nucpred/) και το **PredictNLS** (https://rostlab.org/owiki/index.php/PredictNLS).

Τέλος, το γενικότερο πρόβλημα της υποκυτταρικής στόχευσης των πρωτεϊνών, αντιμετωπίζεται και με μεθόδους που δεν βασίζονται στις σηματοδοτικές αλληλουχίες. Για παράδειγμα, για πολλές κατηγορίες πρωτεϊνών, τέτοιες αλληλουχίες δεν έχουν εντοπιστεί ακόμα (π.χ. πρωτεΐνες της μεμβράνης των μιτοχονδρίων), ενώ για άλλες κατηγορίες δεν υπάρχουν καθόλου (π.χ. πρωτεΐνες των λυσοσωμάτων, του Golgi κ.ο.κ.). Επιπλέον δε, πολλές φορές οι αμινοτελικές αλληλουχίες μπορεί να περιέχουν και σφάλματα λόγω λαθών στην αλληλούχιση. Έτσι, μια σειρά μεθόδων έχουν αναπτυχθεί εδώ και αρκετά χρόνια, οι οποίες βασίζονται σε κάποιας μορφής ολική κωδικοποίηση των αλληλουχιών, κάνοντας χρήση μεγάλου εύρους διαθέσιμων μεθοδολογιώμ (πρότυπα και περιοχές, αμινοξική σύσταση, διπεπτίδια κ.ο.κ.). Η πιο αξιόπιστη και σύγχρονη από αυτές τις μεθόδους, είναι το **WoLF PSORT** (http://wolfpsort.org/), το οποίο αφού βασίζεται στα πιο αξιόπιστα σύγχρονα δεδομένα και κάνει ταξινόμηση σε πολλές κυτταρικές τοποθεσίες των ευκαρυωτικών οργανισμών (αποτελεί τη νεότερη έκδοση του **PSORT** και **PSORT II**). Το αντίστοιχο λογισμικό για τα βακτήρια και τα αρχαία, είναι το **PSORTb** (http://www.psort.org/psortb/index.html). Παρόμοια εργαλεία για τους ευκαρυωτικούς οργανισμούς, είναι το **LOCtree** (http://cubic.bioc.columbia.edu/cgi-bin/var/nair/loctree/query), το **ESLPred2** (http://www.imtech.res.in/raghava/eslpred2/), το **LOCSVMPSI** (http://bioinformatics.ustc.edu.cn/locsvmpsi/locsvmpsi.php), το **CELLO** (http://cello.life.nctu.edu.tw/), το **BaCELLO** (http://gpcr.biocomp.unibo.it/bacello/), το **Protein Prowler** (http://pprowler.imb.uq.edu.au/), το **Hum-Ploc2** (http://www.csbio.sjtu.edu.cn/bioinf/hum-multi-2/), το **AAIndexLoc** (http://aaindexloc.bii.a-

star.edu.sg/) και το **SecretP** (http://cic.scu.edu.cn/bioinformatics/secretp/index.htm). Ειδικά για τους προκαρυωτικούς οργανισμούς, αντίστοιχες μέθοδοι (εκτός από το PSORTb), είναι το **iLoc-Gneg** (http://icpr.jci.edu.cn/bioinfo/iLoc-Gneg), το **Gpos-mPloc** (http://www.csbio.sjtu.edu.cn/bioinf/Gpos-multi/) και το **Gneg-mPloc** (http://www.csbio.sjtu.edu.cn/bioinf/Gneg-multi/) για θετικά και αρνητικά κατά Gram βακτήρια αντίστοιχα, το **SOSUI-GramN** (http://bp.nuap.nagoya-u.ac.jp/sosui/sosuigramn/sosuigramn_submit.html) για αρνητικά κατά Gram βακτήρια, το **PSLPred** (http://www.imtech.res.in/raghava/pslpred/), το **Augur** (http://bioinfo.mikrobio.med.uni-giessen.de/augur), και το **SubLoc** (http://www.bioinfo.tsinghua.edu.cn/SubLoc/).

Τέλος, παρ' όλο που πολλές από τις παραπάνω μεθόδους προβλέπουν εκ των πραγμάτων τη θέση πρωτεϊνών που εκκρίνονται με μη-κλασικά μονοπάτια έκκρισης, έχουν αναπτυχθεί και ειδικές μεθοδολογίες βασισμένες στην ολική σύσταση, που προβλέπουν ειδικά αυτές τις πρωτεΐνες. Έτσι, για τους ευκαρυωτικούς οργανισμούς υπάρχει το **SecretomeP** (http://www.cbs.dtu.dk/services/SecretomeP), ενώ για τα βακτήρια υπάρχει το **NclassG+** (http://www.biolisi.unal.edu.co/web-servers/nclassgpositive/).

7.6.4. Άλλα παραδείγματα μεθόδων πρόγνωσης

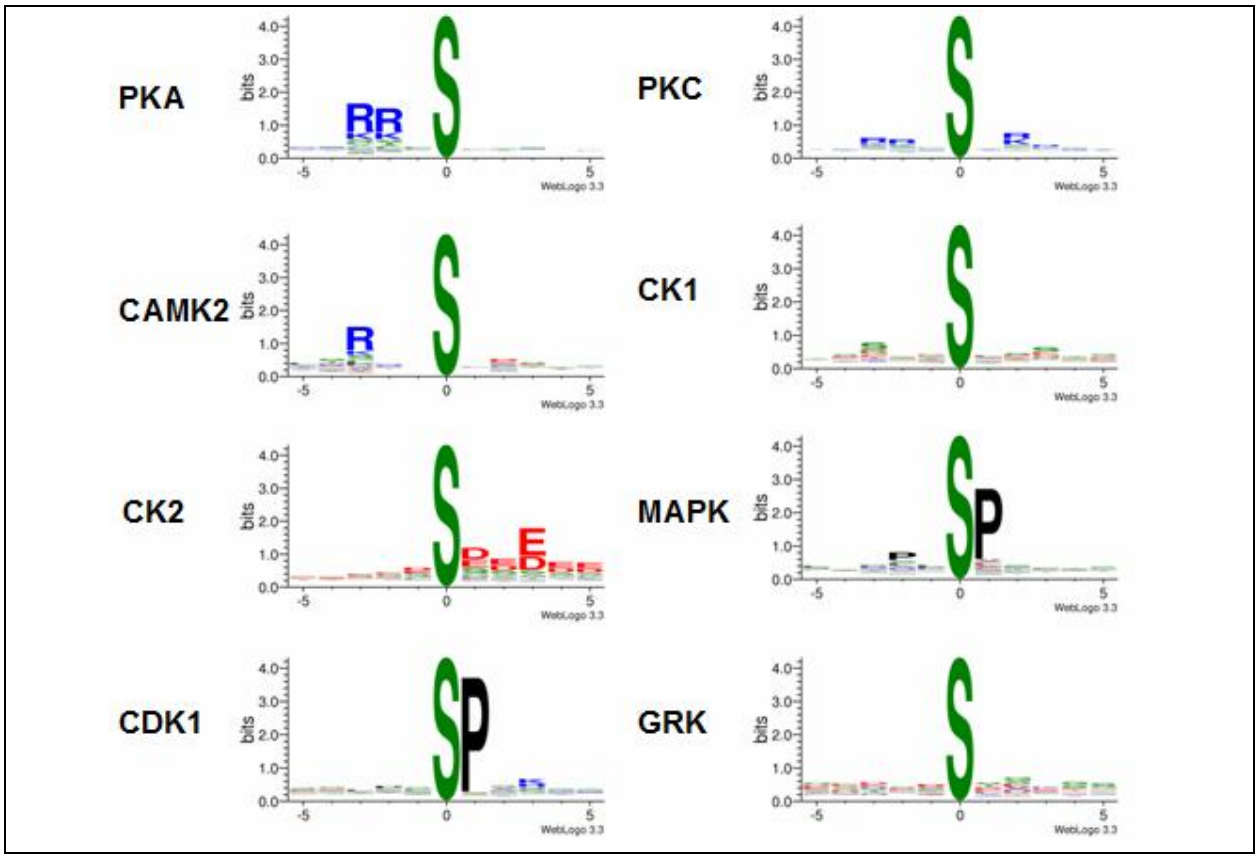
Εκτός από τις παραπάνω περιπτώσεις, υπάρχουν φυσικά και πολλά άλλα παραδείγματα προγνώσεων που μπορεί να γίνουν σε μια πρωτεϊνική αλληλουχία, με σκοπό να αποκαλύψουν διάφορα δομικά ή λειτουργικά χαρακτηριστικά της. Ένα πολύ σημαντικό στοιχείο, που έχει σχέση και με τη δευτεροταγή δομή μιας πρωτεΐνης, αλλά μπορεί να αποκαλύψει και στοιχεία για την τρισδιάστατη δομή της και την δομική της ταξινόμηση, είναι η ύπαρξη υπερελίκων (coiled coil). Το πιο γνωστό από παλιά πρόγραμμα για το σκοπό αυτό, είναι το **COILS** (http://www.ch.embnet.org/software/COILS_form.html), ενώ έχουν προταθεί και νεότερες εκδόσεις όπως το **PAIRCOIL** (http://paircoil2.csail.mit.edu/), το οποίο προβλέπει παράλληλες υπερέλικες, το **MULTICOIL** (http://multicoil2.csail.mit.edu/cgi-bin/multicoil2.cgi), το οποίο προβλέπει και τον ολιγομερισμό, αλλά και το **CCHMM** (http://gpcr.biocomp.unibo.it/cgi/predictors/cc/pred_cchmm.cgi) και το **MARCOIL** (http://bcf.isb-sib.ch/webmarcoil/webmarcoilINFOC1.html), τα οποία βασίζονται σε πιο σύγχρονα μαρκοβιανά μοντέλα.

Ένα άλλο πολύ σημαντικό χαρακτηριστικό, είναι ο εντοπισμός περιοχών με μη σταθερή δομή. Τέτοια χαρακτηριστικά γίνονται ολοένα και πιο σημαντικά τα τελευταία χρόνια, γιατί πολλές πρωτεΐνες εμφανίζονται με μη σταθερή δευτεροταγή δομή, με συνέπεια να μην μπορούν να κρυσταλλωθούν αλλά και να εμπλέκονται λόγω αυτού του χαρακτηριστικού σε πολλές παθολογικές καταστάσεις. Τέτοιοι αλγόριθμοι είναι το **DisEMBL** (http://dis.embl.de/), το **PrDOS** (http://prdos.hgc.jp/cgi-bin/top.cgi), το **DISpro** (http://www.ics.uci.edu/~baldig/dispro.html), το **DISOPRED** (http://bioinf.cs.ucl.ac.uk/psipred/?disopred=1), αλλά και συνδυαστικές μέθοδοι όπως το **MeDor** (http://www.vazymolo.org/MeDor/index.html), το **MetaDisorder** (http://genesilico.pl/metadisorder/) και το **DisProt** (http://www.disprot.org/pondr-fit.php).

Ένα άλλο πολύ σημαντικό δομικό χαρακτηριστικό των πρωτεϊνών, το οποίο μπορεί να δώσει σημαντικά στοιχεία για την τρισδιάστατη δομή, είναι η σύνδεση των κυστεϊνών της ίδιας αλληλουχίας και ο σχηματισμός δισουλφιδικών δεσμών. Οι περισσότεροι αλγόριθμοι αυτής της κατηγορίας, χρησιμοποιούν κάποια τεχνική μηχανικής μάθησης, όπως τα νευρωνικά δίκτυα. Τέτοιοι αλγόριθμοι είναι το **DIpro** (http://download.igb.uci.edu/bridge.html), το **EDBCP** (http://biomedical.ctust.edu.tw/edbcp/), το **CYSPRED** (http://gpcr.biocomp.unibo.it/cgi/predictors/cyspred/pred_cyspredcgi.cgi), το **DiANNA** (http://clavius.bc.edu/~clotelab/DiANNA/), το **Dinosolve** (http://hpcr.cs.odu.edu/dinosolve/), το **DISULFIND** (http://disulfind.dsi.unifi.it/) και το **CysCON** (http://www.csbio.sjtu.edu.cn/bioinf/Cyscon/).

Μια άλλη πολύ μεγάλη κατηγορία μεθόδων πρόγνωσης, είναι οι μέθοδοι που προβλέπουν τις μετα-μεταφραστικές τροποποιήσεις των πρωτεϊνών. Μετα-μεταφραστική τροποποίηση είναι κάθε μεταβολή στη χημική σύσταση της πρωτεΐνης, η οποία πραγματοποιείται αφού έχει γίνει η σύνθεσή της πρωτεΐνης στα ριβωσόματα. Ειδικά στους Ευκαρυωτικούς οργανισμούς, οι μετα-μεταφραστικές τροποποιήσεις αποτελούν πολύ σημαντικούς μηχανισμούς που ελέγχουν και ρυθμίζουν τη δράση των πρωτεϊνών (γι' αυτό και πολλές φορές χαρακτηρίζονται ως «μοριακοί διακόπτες»). Φυσικά, και η αποκοπή των σηματοδοτικών αλληλουχιών που είδαμε πριν, είναι μια μορφή μετα-μεταφραστικής τροποποίησης, όπως είναι και η πρόσδεση σε λιπίδια της μεμβράνης. Αλλά παρ' όλα αυτά, ο όρος συνήθως χρησιμοποιείται για άλλου είδους τροποποιήσεις, κυρίως για την (συνήθως, αλλά όχι πάντα) αντιστρεπτή προσθήκη πλευρικών ομάδων στα αμινοξέα μιας πρωτεΐνης. Τέτοιες τροποποιήσεις, είναι η φωσφορυλίωση, η γλυκοζυλίωση, η μεθυλίωση, η ακετυλίωση, κ.ο.κ.

Η γλυκοζυλίωση είναι η προσθήκη σακχάρων που συμβαίνει συνήθως στο ενδοπλασματικό δίκτυο και το σύμπλεγμα Golgi. Διακρίνεται σε O-γλυκοζυλίωση (γλυκοζυλιώνεται η Ασπαραγίνη), N-γλυκοζυλίωση (γλυκοζυλιώνονται η Σερίνη και η Θρεονίνη) και C-γλυκοζυλίωση (γλυκοζυλιώνεται η Τρυπτοφάνη). Η πιο διαδεδομένη μέθοδος για πρόγνωση N-γλυκοζυλίωσης είναι το **NetNGlyc** (http://www.cbs.dtu.dk/services/NetNGlyc/), ενώ αντίστοιχα για την O-γλυκοζυλίωση έχει αναπτυχθεί το **NetOGlyc** (http://www.cbs.dtu.dk/services/NetOGlyc/) και για την C-γλυκοζυλίωση το **NetCGlyc** (http://www.cbs.dtu.dk/services/NetCGlyc/), ενώ το **YinOYang** (http://www.cbs.dtu.dk/services/YinOYang/) προβλέπει ταυτόχρονη γλυκοζυλίωση και φωσφορυλίωση του ίδιου καταλοίπου σερίνης. Το **GlycoEP** (http://www.imtech.res.in/raghava/glycoep/submit.html), είναι μιά άλλη σύγχρονη μέθοδος που προβλέπει και τις τρεις κατηγορίες γλυκοζυλίωσης, όπως και το **GPP** (http://comp.chem.nottingham.ac.uk/glyco/). Άλλες εφαρμογές, περιλαμβάνουν το **Oglyc** (http://www.biosino.org/Oglyc/), το **ISOGlyP** (http://isoglyp.utep.edu/) και το **CKSSAP_OGlySite** (http://bioinformatics.cau.edu.cn/zzd_lab/CKSAAP_OGlySite/).

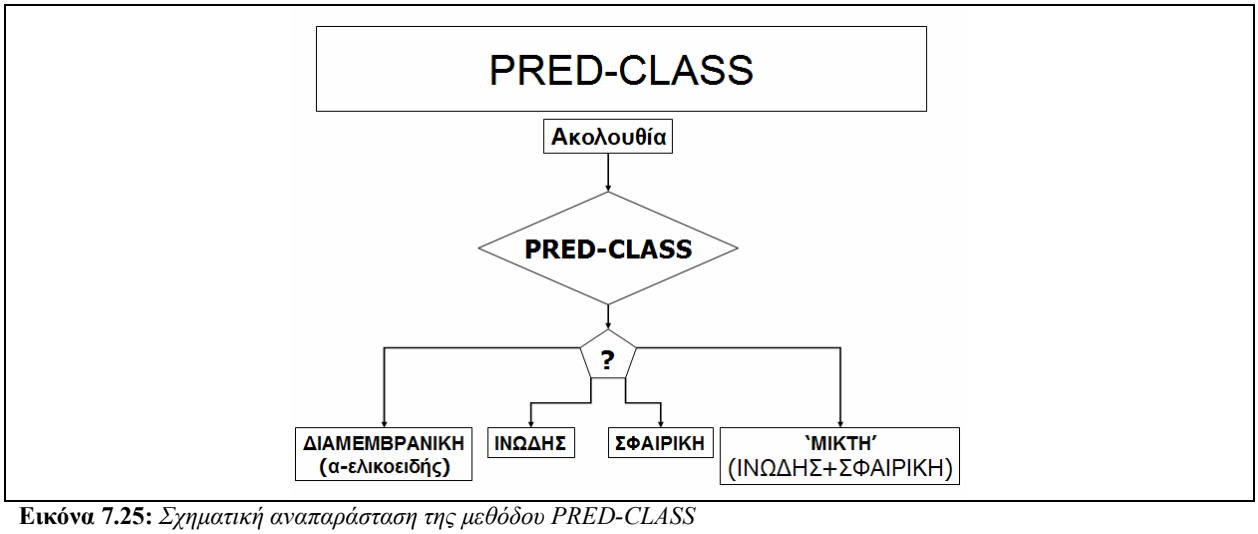


Εικόνα 7.24: *Λογότυπα αλληλουχιών από τις θέσεις δράσης διαφόρων κινασών. Παρόμοια εικόνα δίνουν και οι θέσεις δράσης με Θρεονίνη αντί Σερίνης.*

Η φωσφορυλίωση, είναι επίσης μια πολύ σημαντική κατηγορία τροποποιήσεων που συνίσταται στην προσθήκη φωσφορικής ομάδας, συνήθως στην πλευρική ομάδα της Σερίνης, της Θρεονίνης ή της Τυροσίνης. Τα ένζυμα που πραγματοποιούν αυτές τις αντιδράσεις ονομάζονται κινάσες και η διαδικασία αυτή χρησιμεύει σαν αντιστρεπτός μηχανισμός σηματοδότησης και ενεργοποίησης διαφόρων μηχανισμών. Η πιο γνωστή μέθοδος πρόγνωσης είναι το **NetPhos** (http://www.cbs.dtu.dk/services/NetPhos/) που βασίζεται σε νευρωνικά δίκτυα, ενώ η πιο εξελιγμένη έκδοση **NetPhosK** (http://www.cbs.dtu.dk/services/NetPhosK/) προβλέπει και το είδος της κινάσης που πραγματοποιεί την κάθε αντίδραση. Το **GPS** (http://gps.biocuckoo.org/) είναι ένα άλλο εργαλείο για πρόγνωση της φωσφορυλίωσης (περιέχει και μεθόδους πρόγνωσης και για άλλες μετα-μεταφραστικές τροποποιήσεις). Το **KinasePhos2** (http://kinasephos2.mbc.nctu.edu.tw/) είναι μια ακόμα γνωστή εφαρμογή για πρόγνωση των θέσεων φωσφορυλίωσης που προβλέπει και το είδος της κινάσης και βασίζεται σε HMM. Άλλες μέθοδοι είναι το **PhosphoSVM** (http://sysbio.unl.edu/PhosphoSVM/), το **DISPHOS** (http://www.dabi.temple.edu/disphos/), το **pkaPS** (http://mendel.imp.ac.at/sat/pkaPS/) και το

Predikin (http://predikin.biosci.uq.edu.au/). Εμπειρικές μελέτες έχουν δείξει ότι οι υπάρχουσες μέθοδοι πρόγνωσης έχουν σχετικά μικρή ακρίβεια και πολλές φορές δρουν συμπληρωματικά (άλλες έχουν μεγάλη ευαισθησία, άλλες μεγάλη ειδικότητα), κατά συνέπεια, μια συνδυαστική μέθοδος μπορεί να αποδώσει καλύτερα. Η μόνη προς το παρόν τέτοια μέθοδος είναι το **MetaPredPS** (http://c1.accurascience.com/MetaPred/MetaPredPS_091201/). Πολλές φορές επίσης, σε ειδικές κατηγορίες οργανισμών, οι γενικές μέθοδοι δεν αποδίδουν καλά, οπότε υπάρχει και η ανάγκη για εξειδικευμένες μεθόδους όπως το **NetPhosYeast** (http://www.cbs.dtu.dk/services/NetPhosYeast/) και το **NetPhosBac** (http://www.cbs.dtu.dk/services/NetPhosBac-1.0/).

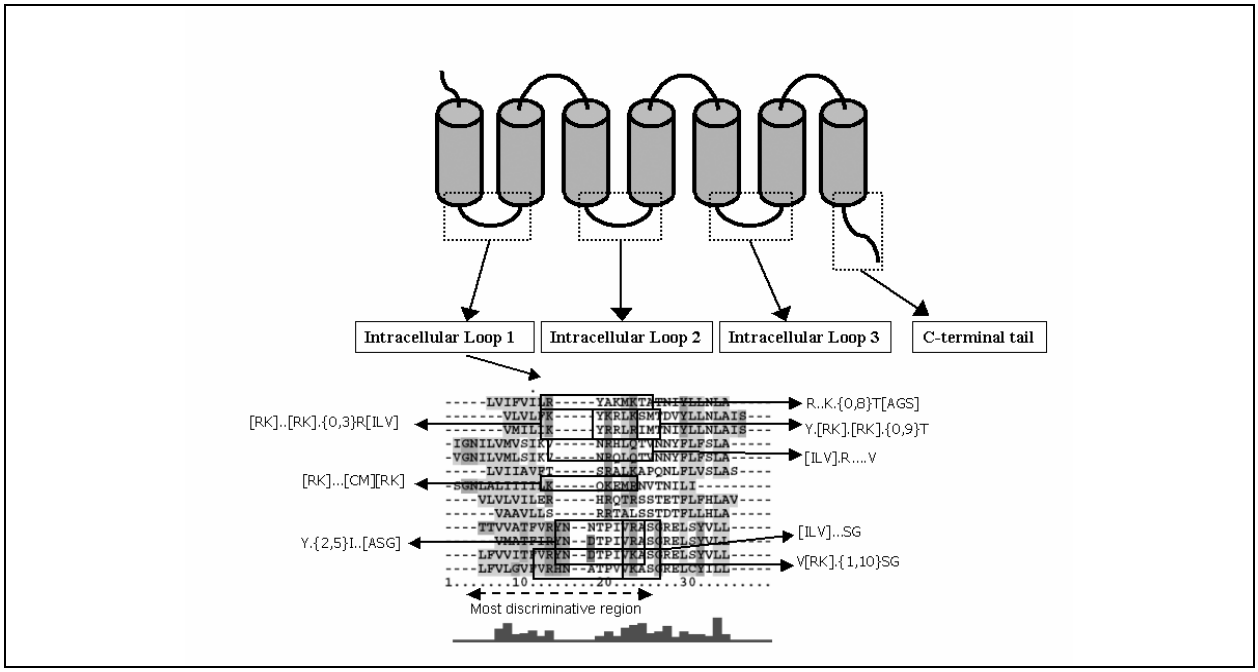
Μια άλλη ομάδα μετα-μεταφραστικών τροποποιήσεων είναι οι τροποποιήσεις που πραγματοποιούνται στο αμινοτελικό άκρο και σχετίζονται με τη σταθερότητα και το χρόνο ημιζωής της πρωτεΐνης. Το **Myristoylator** (http://web.expasy.org/myristoylator/) και το **NMT** (http://mendel.imp.ac.at/myristate/SUPLpredictor.htm) προβλέπουν την προσθήκη ενός λιπιδίου, του μυριστικού οξέως στο αμινοτελικό άκρο, το **NetAcet** (http://www.cbs.dtu.dk/services/NetAcet/) προβλέπει την πιθανή ακετυλίωση του αμινοτελικού άκρου ενώ το **TermiNator** (http://www.isv.cnrs-gif.fr/terminator3/index.html) είναι πιο γενικό και προβλέπει ακετυλίωση, μυριστοϋλίωση ή παλμιτοϋλίωση. Εκτός βέβαια από το αμινοτελικό άκρο, παρόμοιες τροποποιήσεις, ειδικά ακετυλίωση και σουλφυλίωση (προσθήκη ομάδας θειικού οξέος), συμβαίνουν και σε εσωτερικά κατάλοιπα των πρωτεϊνών. Έτσι, το **CSS-Palm** (http://csspalm.biocuckoo.org/) προβλέπει προσθήκη παλμιτικού οξέως σε εσωτερικές θέσεις, το **GPS-TSP** (http://tsp.biocuckoo.org/) και το **Sulfinator** (http://web.expasy.org/sulfinator/) προβλέπουν σουλφιλίωση των τυροσινών, ενώ το **PAIL** (http://bdmpail.biocuckoo.org/) και το **KAT** (http://bioinfo.bjmu.edu.cn/huac/) προβλέπουν εσωτερική ακετυλίωση των λυσινών. Τέλος, μια άλλη πολύ σημαντική κατηγορία τροποποιήσεων είναι η προσθήκη ολόκληρων πρωτεϊνών σαν προσθετικές ομάδες. Με μια διαδικασία σαν αυτή ρυθμίζονται σειρά άλλων διεργασιών όπως η πρωτεϊνική σταθερότητα, η μεταγραφική ρύθμιση της απόπτωσης και οι διεργασίες του κυτταρικού κύκλου. Η Ουμπικουϊτίνη (Ubiquitin) ήταν η πρώτη τέτοια πρωτεΐνη που ανακαλύφθηκε, η οποία ρυθμίζει την αποικοδόμηση των πρωτεϊνών από το πρωτεάσωμα, ενώ ακολούθησαν και άλλες που συνολικά ονομάστηκαν πρωτεΐνες SUMO (Small Ubiquitin-like Modifie). Την προσθήκη της ουμπικουϊτίνης την προβλέπει η μέθοδος **UbPred** (http://www.ubpred.org/), η **BDM-PUB** (http://bdmpub.biocuckoo.org/), η **CKSAAP_UbSite** (http://protein.cau.edu.cn/cksaap_ubsite/), η **iUbiq-Lys** (http://www.jci-bioinfo.cn/iUbiq-Lys) και η **UbiProber** (http://bioinfo.ncu.edu.cn/UbiProber.aspx). Γενικότερα, την προσθήκη SUMO την προβλέπει το **SUMOplot** (http://www.abgent.com/sumoplot) και το **GPS-SUMP** (http://sumosp.biocuckoo.org/).



Εικόνα 7.25: *Σχηματική αναπαράσταση της μεθόδου PRED-CLASS*

Τέλος, πρέπει να κάνουμε και μια αναφορά σε μεθόδους που προβλέπουν γενικότερα δομικά ή λειτουργικά χαρακτηριστικά των πρωτεϊνών. Ένα κλασικό παράδειγμα είναι οι μέθοδοι που προβλέπουν την δομική ταξινόμηση μιας πρωτεΐνης. Το πιο αντιπροσωπευτικό παράδειγμα, είναι η μέθοδος **PRED-CLASS** (http://athina.biol.uoa.gr/PRED-CLASS/) η οποία προβλέπει αν μια πρωτεΐνη είναι μεμβρανική, ινώδης,

μικτή ή σφαιρική υδατοδιαλυτή (οι μικτές είναι αυτές που έχουν και σφαιρικές αλλά και ινώδεις περιοχές, π.χ. μυοσίνη) (Pasquier et al., 2001). Η πρόγνωση βασίζεται σε 3 διαφορετικά επάλληλα νευρωνικά δίκτυα, το καθένα από τα οποία διαχωρίζει μια ομάδα από τις υπόλοιπες. Το πρώτο δίκτυο, είναι βασισμένο σε τοπική πληροφορία και δουλεύει σε διαδοχικά (επικαλυπτόμενα) 'παράθυρα' εύρους 30 καταλοίπων. Για την κωδικοποίηση χρησιμοποιείται μια κωδικοποίηση με παράμετρους υδροφοβικότητας (μία τιμή για κάθε αμινοξύ), ενώ το δίκτυο έχει μόνο δύο κρυφούς νευρώνες, με αποτέλεσμα να είναι πολύ γρήγορο. Ο σκοπός του πρώτου δικτύου είναι να εντοπίσει τις διαμεμβρανικές πρωτεΐνες, αλλά όχι και να προβλέψει όλα τα διαμεμβρανικά τμήματα. Έτσι, όταν εντοπίσει έστω και ένα παράθυρο με μεγάλη τιμή υδροφοβικότητας, το πρόγραμμα σταματάει και η πρωτεΐνη καταχωρείται ως «διαμεμβρανική». Οι υπόλοιπες τρεις κατηγορίες, διαχωρίζονται με τη χρήση δύο άλλων δικτύων, τα οποία όμως βασίζονται σε ολική πληροφορία με παρόμοια κωδικοποίηση. Σαν παράμετροι, χρησιμοποιούνται οι συχνότητες εμφάνισης των αμινοξικών καταλοίπων (20) και κάποιων ομάδων τους (10 παράμετροι, που αφορούν κυρίως τις ομαδοποιήσεις σχετικά με την υδροφοβικότητα, το φορτίο αλλά και τις προτιμήσεις για διάφορα στοιχεία δευτεροταγούς δομής). Επιπλέον, χρησιμοποιούνται και αντίστοιχα δεδομένα (άλλες 30 παράμετροι δηλαδή), από τον αντίστοιχο μετασχηματισμό Fourier (FFT), μια επιλογή που έγινε για να μπορέσει να αποτυπώσει τις πληροφορίες για την περιοδικότητα των αμινοξέων. Το πρώτο από αυτά τα δίκτυα διαχωρίζει τις ινώδεις πρωτεΐνες από τις υπόλοιπες, ενώ το δεύτερο, τις μικτές από τις σφαιρικές (ο πιο δύσκολος ίσως διαχωρισμός). Η μέθοδος αποδίδει πολύ καλά (>95% σωστή ταξινόμηση) και ήταν από τις πρώτες μεθόδους που εφαρμόστηκαν σε πλήρη γονιδιώματα και έδειξαν ότι οι μεμβρανικές πρωτεΐνες αποτελούν περίπου το 30% των πρωτεϊνών που κωδικοποιούνται από αυτά. Παρόμοιας φύσης μέθοδοι, έχουν παρουσιαστεί και για άλλα αντίστοιχα προβλήματα, όπως π.χ. για το διαχωρισμό των δομικών κατηγοριών των σφαιρικών πρωτεϊνών ή για την πρόγνωση της ενζυμικής ενεργότητας των ενζύμων, αλλά δεν υπάρχουν πολλοί αξιόπιστοι και διαθέσιμοι στο ευρύ κοινό αλγόριθμοι.

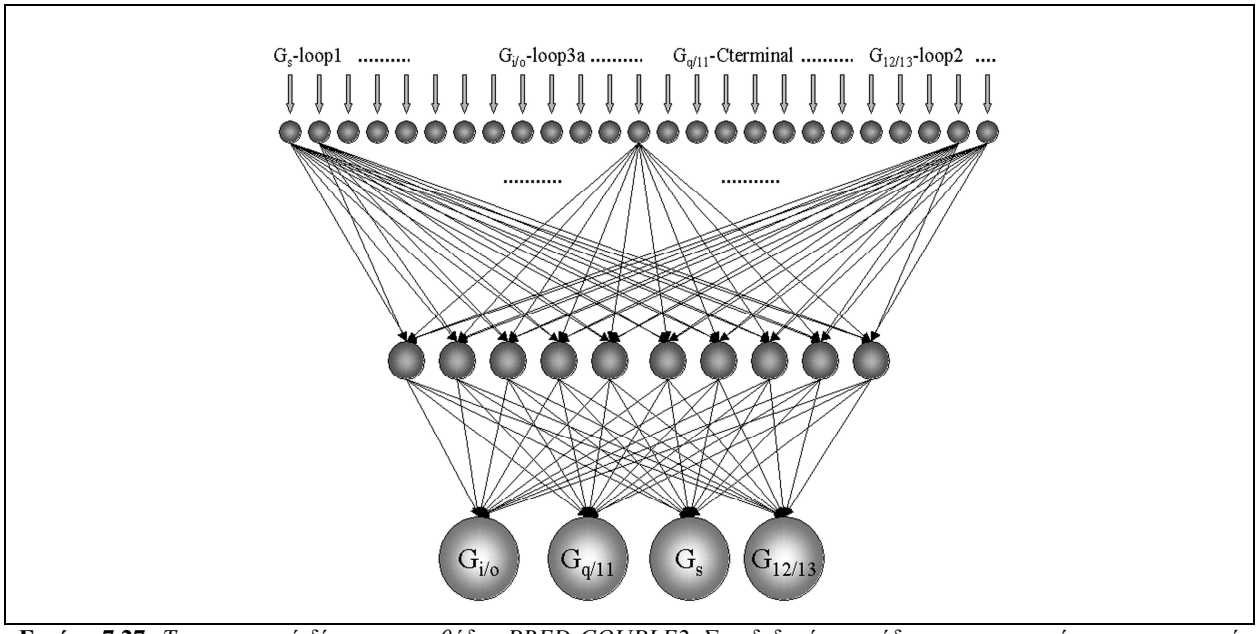


Εικόνα 7.26: *Σχηματική αναπαράσταση της μεθόδου PRED-COUPLE. Από τις στοιχίσεις των υποδοχέων με την ίδια σύζευξη, προέκυψε με εντατική αναζήτηση μια περιοχή που μπορεί να χρησιμοποιηθεί για διαχωρισμό μεταξύ των κατηγοριών και από την οποία κατασκευάζεται το pHMM. Τα πρότυπα κανονικών εκφράσεων που δίνονται, είναι ίδια με αυτά που είχαν εντοπιστεί σε παλιότερη μελέτη.*

Τέλος, ένα άλλο παράδειγμα αφορά τη λειτουργική πρόβλεψη για το διαχωρισμό των υποδοχέων GPCR ανάλογα με την ικανότητά τους να αλληλεπιδρούν με συγκεκριμένη ομάδα G-πρωτεϊνών. Οι Συζευγμένοι με G-πρωτεΐνες Υποδοχείς (G protein-coupled receptors-GPCRs), σχηματίζουν μια από τις μεγαλύτερες ομάδες διαμεμβρανικών υποδοχέων στους ευκαρυωτικούς οργανισμούς. Διαθέτουν επτά διαμεμβρανικές α-έλικες, γεγονός που επιβεβαιώθηκε πειραματικά με την πρόσφατη ανάλυση της

κρυσταλλικής δομής της ροδοψίνης αλλά και τις αναλύσεις των υπολοίπων δομών που έχουν προκύψει από τότε. Όσον αφορά, τη λειτουργική ταξινόμηση των GPCRs, έως πρόσφατα, λίγες ερευνητικές ομάδες είχαν αναπτύξει υπολογιστικούς αλγόριθμους ικανούς να προβλέπουν την ειδικότητά τους σχετικά με τη σύζευξη με G-πρωτεΐνες, αλλά όχι πάντα με τα αναμενόμενα αποτελέσματα. Η μέθοδος **PRED-COUPLE** (http://athina.biol.uoa.gr/bioinformatics/PRED-COUPLE/), η οποία βασίζεται σε εύρεση χαρακτηριστικών περιοχών και χρήση profile Hidden Markov Models, ήταν η πρώτη δημόσια διαθέσιμη στο διαδίκτυο μέθοδος πρόγνωσης της εξειδίκευσης των GPCR σε G-πρωτεΐνες (Sgourakis, Bagos, Papasaikas, & Hamodrakas, 2005). Η βασική της αρχή στηρίζεται στην παραδοχή ότι οι ενδοκυττάριοι βρόχοι, περιέχουν την απαραίτητη πληροφορία σε επίπεδο ακολουθίας, η οποία καθορίζει το δυναμικό της σύζευξης ενός υποδοχέα με μια G-πρωτεΐνη. Η μέθοδος, ταξινομεί τους GPCRs σε τρεις κατηγορίες εξειδίκευσης (Gi/o,Gs και Gq/11) και όταν ελεγχθεί η αποτελεσματικότητά της με μια διαδικασία cross-validation (το σύνολο εκπαίδευσης χωρισμένο σε 5 υποσύνολα), αποδίδει σωστά αποτελέσματα σε ποσοστό 89.7%. Σε ένα ανεξάρτητο σύνολο 30 υποδοχέων με καμία ομοιότητα με αυτούς που χρησιμοποιήθηκαν για εκπαίδευση, προβλέπει σωστά την εξειδίκευση των 25 από αυτούς (83.3%).

Στη δεύτερη έκδοση της μεθόδου, το **PRED-COUPLE2** (http://athina.biol.uoa.gr/bioinformatics/PRED-COUPLE2/), οι ίδιες τεχνικές συνδυάστηκαν με ένα νευρωνικό δίκτυο, το οποίο όμως θα αποφάσιζε αν ένας δεδομένος υποδοχέας κάνει σύζευξη με μία, δύο, τρεις ή και τις τέσσερις από τις κατηγορίες των G-πρωτεϊνών (Sgourakis, Bagos, & Hamodrakas, 2005). Σαν δεδομένα στο νευρωνικό δίκτυο, δίνονται πλέον τα σκορ από τα pHMM που έχουν κατασκευαστεί για τις διάφορες κατηγορίες σύζευξης. Με αυτόν τον τρόπο, η μέθοδος όχι μόνο προβλέπει σε μεγάλο ποσοστό (~95%) τις αλληλεπιδράσεις που είναι «ένα-προς-ένα», αλλά καταφέρνει να προβλέψει και πολλές από τις περιπτώσεις υποδοχέων με μη αποκλειστική σύζευξη. Η μέθοδος είναι η μοναδική που καταφέρνει τέτοιου είδους προγνώσεις. Είχαν αναπτυχθεί και άλλες παρόμοιες μεθοδολογίες, αλλά δεν υπάρχουν αυτή τη στιγμή διαθέσιμες στο κοινό διαδικτυακές εφαρμογές.

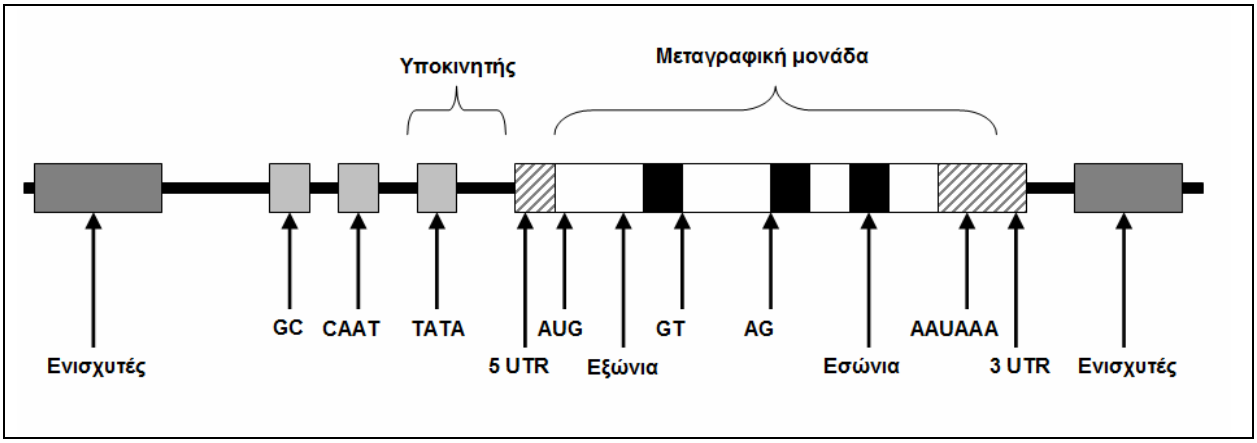


Εικόνα 7.27: *Το νευρωνικό δίκτυο της μεθόδου PRED-COUPLE2. Σαν δεδομένα εισόδου χρησιμοποιούνται τα σκορ από τα pHMM που είχαν κατασκευαστεί με μια διαδικασία όμοια με το PRED-COUPLE.*

7.7. Μέθοδοι πρόγνωσης για αλληλουχίες DNA/RNA

Οι μέθοδοι πρόγνωσης, φυσικά, δεν περιορίζονται μόνο στις περιπτώσεις πρωτεϊνών. Υπάρχουν πολλές και ιδιαίτερα σημαντικές περιπτώσεις κατά τις οποίες χρειαζόμαστε μια μέθοδο πρόγνωσης σχεδιασμένη για αλληλουχίες DNA και RNA. Το πιο βασικό πρόβλημα στην περίπτωση αλληλουχιών DNA είναι αυτό της εύρεσης γονιδίων (gene finding), αλλά και αυτό μπορεί να αντιμετωπιστεί με πολλούς τρόπους ενώ μπορεί

και να χωριστεί σε μικρότερα «υπο-προβλήματα» (Mathé, Sagot, Schiex, & Rouze, 2002). Η εύρεση των πραγματικών γονιδίων που κωδικοποιούνται σε ένα γονιδίωμα, είναι τεράστιας σημασίας πρόβλημα, γιατί όπως έχουμε πει, η αλληλούχιση ενός γονιδιώματος είναι μεν μια δουλειά ρουτίνας, αλλά αυτό δεν σημαίνει ότι και αυτόματα θα έχουμε γνώση των πρωτεϊνών που κωδικοποιεί αυτό το γονιδίωμα. Η εύρεση απλά των ανοιχτών πλαισίων ανάγνωσης, είναι μια σχετικά απλή διαδικασία (ειδικά στους προκαρυωτικούς οργανισμούς), αλλά ακόμα και έτσι υπάρχουν πάρα πολλά ψευδογονίδια ή περιοχές που απλά έτυχε να έχουν το κωδικόνιο έναρξης και λήξης σε διαφορά φάσης (σε απόσταση νουκελοτιδίων που είναι πολλαπλάσιο του 3). Έτσι, η εύρεση των κατάλληλων ρυθμιστικών περιοχών (υποκινητές) που καθορίζουν την έκφραση του γονιδίου, είναι μια πολύ σημαντική διαδικασία. Στους δε ευκαρυωτικούς οργανισμούς, στους οποίους τα γονίδια είναι διακοπτόμενα από εσώνια και εξώνια, επιφέρει μια επιπλέον πολυπλοκότητα στους υπολογισμούς καθώς οι ρυθμιστικές αυτές περιοχές πρέπει να αναγνωριστούν πριν καν εντοπιστούν τα ανοιχτά πλαίσια ανάγνωσης. Επιπλέον δε, στους ευκαρυωτικούς οργανισμούς υπάρχουν και άλλες ρυθμιστικές αλληλουχίες πιο μακριά από τον υποκινητή, οι οποίες πρέπει να εντοπιστούν.



Εικόνα 7.28: *Η τυπική δομή ενός ευκαρυωτικού γονιδίου*

Έτσι καταλαβαίνουμε ότι μπορεί να υπάρξουν μια σειρά μικρότερα από «προβλήματα» προς επίλυση: μπορεί να υπάρχουν μέθοδοι εύρεσης των σημείων αποκοπής και συρραφής των εξωνίων (exon/intron splice site), μέθοδοι αναγνώρισης του υποκινητή (promoter recognition), μέθοδοι αναγνώρισης του σημείου έναρξης της μεταγραφής (translation initiation site prediction) (Saeys, Abeel, Degroeve, & Van de Peer, 2007), μέθοδοι εύρεσης του σημείου πολυαδενυλίωσης στο mRNA (polyadenylation prediction) (Chang et al., 2011), αλλά και, φυσικά, μέθοδοι που προβλέπουν ολόκληρη τη δομή του γονιδίου. Τέλος, οι μέθοδοι έχουν και διαφορετικές στατιστικές ιδιότητες. Ανάλογα με την ευαισθησία και την ειδικότητα που μπορεί να έχει η κάθε μία, είναι δυνατόν να αποδίδουν καλύτερα είτε σε απομονωμένες περιοχές DNA, είτε σε πλήρη γονιδιώματα (Saeys et al., 2007). Ένα άλλο σημείο που χρειάζεται προσοχή, είναι η ειδικότητα ανά οργανισμό ή ομάδα οργανισμών, καθώς οι στατιστικές ιδιότητες των νουκλεοτιδίων (ακόμα και στο πλαίσιο των αποδεκτών κωδικωνίων), διαφέρουν ανάμεσα στις μεγάλες ομάδες. Έτσι, υπάρχουν εξειδικευμένα εργαλεία για ειδικές περιπτώσεις ή εργαλεία που λαμβάνουν υπόψη τους τη φυλογενετική προέλευση του οργανισμού. Γενικά, υπάρχει μια πληθώρα μεθόδων καθώς η σχετική βιβλιογραφία είχε ξεκινήσει από τη δεκαετία του 1980, ενώ τα πρώτα ολοκληρωμένα προγράμματα εμφανίστηκαν τη δεκαετία του 1990 παράλληλα με τις προσπάθειες αλληλούχισης. Οι μεθοδολογίες που έχουν χρησιμοποιηθεί για τα προβλήματα αυτά, καλύπτουν ένα μεγάλο εύρος: από στατιστικές τεχνικές, weight matrices και προφιλ, νευρωνικά δίκτυα, μαρκοβιανές αλυσίδες μέχρι και Hidden Markov Models. Οι μεθοδολογίες που βασίζονται καθαρά σε εκπαίδευση για να κάνουν την πρόγνωση αναφέρονται και ως *ab initio gene finders*, ενώ οι μεθοδολογίες στις οποίες χρησιμοποιείται και πληροφορία από τις ήδη υπάρχουσες γνωστές πρωτεΐνες με σκοπό να «καθοδηγηθεί» η πρόγνωση από τα γνωστά παραδείγματα ονομάζονται *homology-based gene finders*.

Για τους προκαρυωτικούς οργανισμούς, τα πιο γνωστά και πετυχημένα εργαλεία περιλαμβάνουν τα:

- **FrameD** (http://tata.toulouse.inra.fr/apps/FrameD/FD)
- **GeneMark** (http://exon.gatech.edu/GeneMark/gmchoice.html)
- **Glimmer** (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi)
- **EasyGene** (http://www.cbs.dtu.dk/services/EasyGene/)

- **FGENESB** (http://linux1.softberry.com/berry.phtml?topic=fgenesb&group=help&subgroup=gfindb)
- **Prodigal** (http://prodigal.ornl.gov/)

Αντίστοιχα, για τους ευκαρυωτικούς οργανισμούς, τα πιο πετυχημένα αντίστοιχα εργαλεία είναι:

- **FGENESH** (http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind)
- **GlimmerHMM** (https://ccb.jhu.edu/software/glimmerhmm/)
- **HMMgene** (http://www.cbs.dtu.dk/services/HMMgene/)
- **GeneMark.hmm** (http://exon.gatech.edu/GeneMark/hmmchoice.html)
- **GeneID** (http://genome.crg.es/software/geneid/geneid.html)
- **GeneScan** (http://genes.mit.edu/GENSCAN.html)
- **mGene** (http://raetschlab.org/suppl/mgene)
- **Grail** (http://compbio.ornl.gov/grailexp/)

Ειδικά εργαλεία για την έναρξη της μεταγραφής (translation initiation) είναι:

- **ATGpr** (http://atgpr.dbcls.jp/)
- **NetStart** (http://www.cbs.dtu.dk/services/NetStart/)
- **TIS Miner** (http://dnafsminer.bic.nus.edu.sg/Tis.html)
- **StartScan** (http://bioinformatics.psb.ugent.be/webtools/startscan/)

Για την πολυαδενυλίωση του mRNA τα διαθέσιμα εργαλεία αυτή τη στιγμή είναι:

- **Poly(A) Signal Miner** (http://dnafsminer.bic.nus.edu.sg/)
- **PolyAPred** (http://www.imtech.res.in/raghava/polyapred/help.html)
- **POLYAH** (http://www.softberry.com/berry.phtml?topic=polyah&group=programs&subgroup=promoter)
- **PolyApredict** (http://cub.comsats.edu.pk/polyapredict.htm)

Τέλος, μέθοδοι που εστιάζονται στην εύρεση των σημείων αποκοπής και συρραφής εσωνίων/εξωνίων σε ευκαρυωτικά γονιδιώματα, είναι:

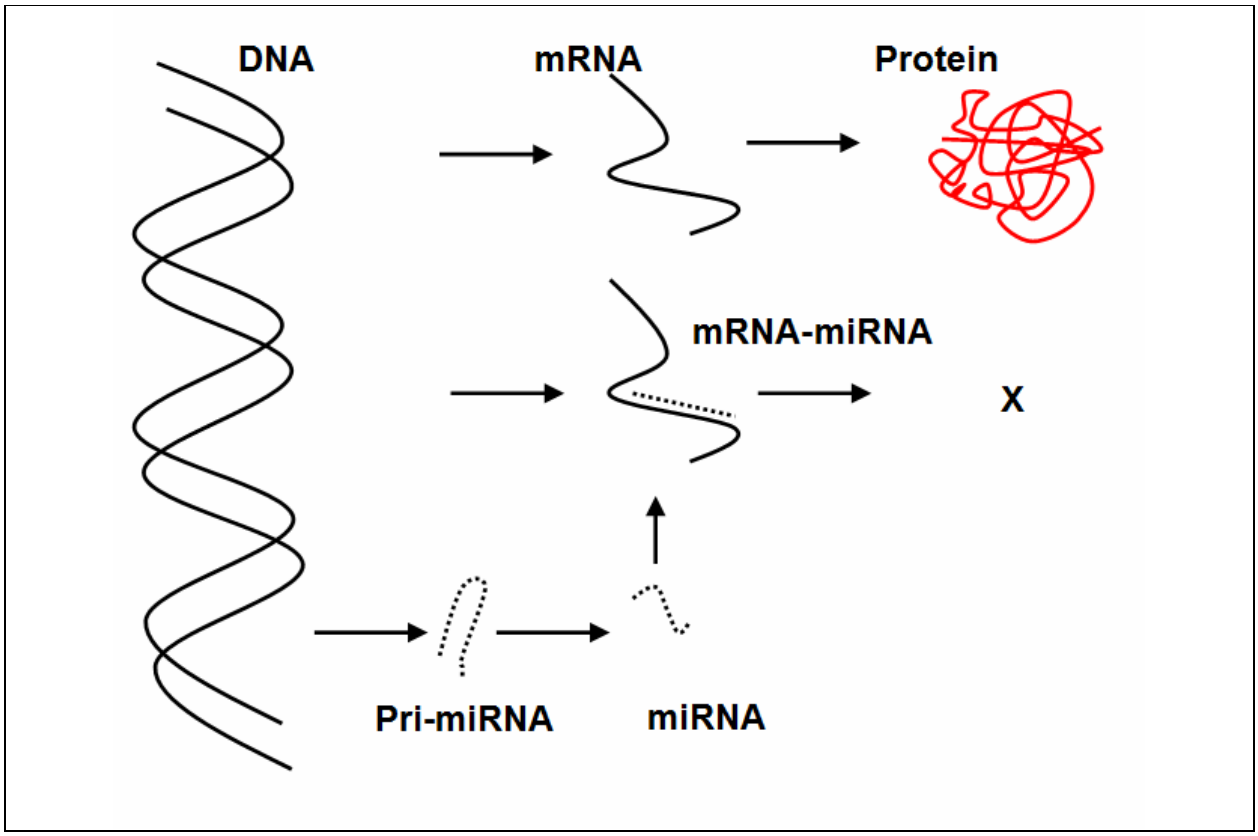
- **Human Splice Finder** (http://www.umd.be/HSF3/)
- **NetGene** (http://www.cbs.dtu.dk/services/NetGene2/)
- **NetPlant** (http://www.cbs.dtu.dk/services/NetPGene/)
- **GeneSplicer** (https://ccb.jhu.edu/software/genesplicer/)
- **SpliceView** (http://bioinfo4.itb.cnr.it/~webgene/wwwspliceview_ex.html)
- **SplicePredictor** (http://bioservices.usd.edu/splicepredictor/)

Φυσικά, υπάρχουν και άλλες μέθοδοι πρόγνωσης για αλληλουχίες DNA που δεν αφορούν μόνο την εύρεση γονιδίων αλλά και μια σειρά άλλων λειτουργικών ή δομικών χαρακτηριστικών. Έτσι, υπάρχουν μέθοδοι πρόγνωσης των θέσεων μεθυλίωσης όπως το **Methylator** (http://bio.dfci.harvard.edu/Methylator/) και γενικότερα των επιγενετικών τροποποιήσεων όπως το **epigram** (http://wanglab.ucsd.edu/star/epigram/), μέθοδοι πρόγνωσης της θέσης των νουκλεοσωμάτων όπως το **NuPoP** (http://nucleosome.stats.northwestern.edu/) και η μέθοδος του Segal (http://genie.weizmann.ac.il/software/nucleo_prediction.html), μέθοδοι πρόγνωσης του σημείου ζέσεως των μορίων DNA όπως το **uMELT** (https://www.dna.utah.edu/umelt/umelt.html), μέθοδοι πρόγνωσης των δομικών χαρακτηριστικών του μορίου του DNA όπως το **DNAshape** (http://rohslab.cmb.usc.edu/DNAshape/) και το **DNAtools** (http://hydra.icgeb.trieste.it/dna/), αλλά και μέθοδοι πρόγνωσης του λειτουργικού αποτελέσματος των νουκελοτιδικών πολυμορφισμών (SNPs), όπως το **SNAP** (https://rostlab.org/services/snap/), το **FuncPred** (http://snpinfo.niehs.nih.gov/snpinfo/snpfunc.htm) και το **PredictSNP** (http://loschmidt.chemi.muni.cz/predictsnp/).

Όσον αφορά τα μόρια RNA, καθώς τα μόρια αυτά παρουσιάζουν ομοιότητες στη δομική ποικιλομορφία με τις πρωτεΐνες, οι αλγόριθμοι πρόγνωσης έχουν να κάνουν περισσότερο με τη δομή. Το βασικό ερώτημα που ενδιαφέρει σε αυτή την περίπτωση, αφορά την πιθανή δευτεροταγή και τριτοταγή δομή ενός μορίου RNA, δεδομένης της αλληλουχίας του. Το ειδικό θέμα που προκύπτει, είναι ότι στα μόρια RNA η συμπληρωματικότητα των βάσεων οδηγεί σε ζευγάρωμα μέσα στο ίδιο μόριο. Αυτού του είδους οι συσχετίσεις χρειάζονται διαφορετικά εργαλεία για να μοντελοποιηθούν, γι' αυτό και τους διάφορους

αλγόριθμους και τις μεθόδους πρόγνωσης για τη δομή των RNA θα τα συζητήσουμε αναλυτικά στο κεφάλαιο 10.

Μια ειδική όμως κατηγορία μορίων RNA, έχει αποκτήσει μεγάλο ενδιαφέρον τα τελευταία χρόνια και έχουν αναπτυχθεί πολλοί αλγόριθμοι πρόγνωσης για τον εντοπισμό τους. Πρόκειται για τα micro RNA (miRNA) τα οποία είναι μικρά μη-κωδικά μόρια RNA (αποτελούμενα συνήθως από 21-22 νουκλεοτίδια, προερχόμενα από ένα μεγαλύτερο πρόδρομο μόριο που σχηματίζει βρόχο, το pri-miRNA) τα οποία βρίσκονται σχεδόν σε όλους τους οργανισμούς και η λειτουργία τους συνίσταται στο να αποσιωπούν τα mRNA και να ρυθμίζουν με αυτόν τον τρόπο μετα-μεταγραφικά τη λειτουργία των γονιδίων (Cai, Yu, Hu, & Yu, 2009). Η λειτουργία αυτή επιτυγχάνεται μέσω ζευγαρώματος με συμπληρωματικές περιοχές που βρίσκονται σε μόρια mRNA. Έτσι, τα mRNA παύουν να λειτουργούν είτε γιατί αποσυντίθενται, είτε γιατί αποσταθεροποιούνται λόγω αλλοίωσης της πολυαδενυλικής ουράς, είτε γιατί δεν μεταφράζονται το ίδιο γρήγορα στα ριβοσώματα. Τα miRNAs μοιάζουν δηλαδή με τα small interfering RNA (siRNA) με τη διαφορά ότι τα miRNAs προέρχονται από μετάγραφα RNA που αναδιπλώνονται για να σχηματίσουν βρόχους ενώ τα siRNAs προέρχονται από μεγαλύτερα δίκλωνα μόρια RNA. Στο ανθρώπινο γονιδίωμα υπάρχουν περίπου 1000 miRNA τα οποία εμφανίζονται σε πολλούς τύπους κυττάρων και φαίνεται ότι στοχεύουν περίπου το 60% των υπόλοιπων γονιδίων, ενώ έχουν και εμπλοκή σε πολλές ασθένειες.



Εικόνα 7.29: *Απλοποιημένη αναπαράσταση του τρόπου λειτουργίας των miRNA. Πολλές λεπτομέρειες της βιοσύνθεσης και της ωρίμανσης παραλείπονται.*

Τα miRNA και ο μηχανισμός τους είναι συντηρημένα στα θηλαστικά και τα φυτά, και πιστεύεται ότι είναι κατάλοιπα μιας παλιάς διαδικασίας ρύθμισης της γονιδιακής έκφρασης. Παρ' όλα αυτά υπάρχουν αρκετά μεγάλες διαφορές τόσο στη βιοσύνθεση όσο και στη λειτουργία ανάμεσα σε φυτά και ζώα. Τα φυτικά miRNA εμφανίζουν συνήθως μια σχεδόν τέλεια συμπληρωματικότητα με τα mRNA στόχους, και κατά συνέπεια επάγουν την αποσιώπηση κυρίως με απευθείας διάσπαση του mRNA. Αντίθετα, τα ζωικά miRNA αναγνωρίζουν το στόχο mRNA χρησιμοποιώντας τη συμπληρωματικότητα μόνο 6–8 νουκελοτιδίων που βρίσκονται στην 5' περιοχή του miRNA. Με αυτόν τον τρόπο δεν είναι ικανά να επάγουν διάσπαση του mRNA αλλά λειτουργούν με τους υπόλοιπους μηχανισμούς που αναφέραμε παραπάνω. Όπως είναι προφανές,

ένα δεδομένο miRNA, ειδικά στα θηλαστικά, έχει πολλά mRNA σαν στόχους, ενώ ένα δεδομένο mRNA είναι πιθανό να ελέγχεται από περισσότερα του ενός miRNA.

Τα υπολογιστικά προβλήματα που προκύπτουν σχετικά με τα miRNA είναι δύο: αφενός μεν ο ίδιος ο εντοπισμός τους στα γονιδιώματα, αφετέρου δε η πρόγνωση των στόχων τους. Και τα δύο αντιμετωπίζονται με συνδυασμούς μεθόδων, όπως νευρωνικά δίκτυα, υπολογιστικές γραμματικές, HMM, τεχνικές μηχανικής μάθησης, αλλά και λαμβάνοντας υπόψη τη συμπληρωματικότητα των βάσεων και την πιθανή δευτεροταγή δομή του RNA. Οι βασικότερες μέθοδοι πρόγνωσης που είναι διαθέσιμες για τον εντοπισμό των miRNA αναφέρονται παρακάτω:

- **CID miRNA** (http://melb.agrf.org.au:8888/cidmirna/)
- **MiRPara** (https://code.google.com/p/mirpara/)
- **HeteroMirPred** (http://ncrna-pred.com/premiRNA.html)
- **HHMMiR** (http://biodev.hgen.pitt.edu/kadriAPBC2009.html)
- **HuntMi** (http://adaa.polsl.pl/agudys/huntmi/huntmi.htm)
- **MaturePred** (http://nclab.hit.edu.cn/maturepred/)
- **microPred** (http://www.cs.ox.ac.uk/people/manohara.rukshan.batuwita/microPred.htm)
- **MiPred** (http://www.bioinf.seu.edu.cn/miRNA/)
- **miRAbela** (http://www.mirz.unibas.ch/cgi/pred_miRNA_genes.cgi)
- **MiRAlign** (http://bioinfo.au.tsinghua.edu.cn/miralign/)
- **miRBoost** (http://evryrna.ibisc.univ-evry.fr/miRBoost/index.html)
- **mirnaDetect** (http://datamining.xmu.edu.cn/main/~leyiwei/mirnaDetect.html)
- **miRNAFold** (http://evryrna.ibisc.univ-evry.fr/miRNAFold/)
- **MiRscan** (http://genes.mit.edu/mirscan/)
- **novoMIR** (http://www.biophys.uni-duesseldorf.de/novomir/)
- **ProMiR** (http://bi.snu.ac.kr/Research/ProMiR/ProMiR.html)
- **RNAmicro** (http://www.tbi.univie.ac.at/~jana/software/RNAmicro.html)
- **tripletSVM** (http://bioinfo.au.tsinghua.edu.cn/mirnasvm/)
- **SplamiR** (http://www.uni-jena.de/SplamiR.html)
- **SSCprofiler** (http://mirna.imbb.forth.gr/SSCprofiler.html)
- **EumiR** (http://miracle.igib.res.in/eumir/)

Αντίστοιχα, οι μέθοδοι που είναι διαθέσιμες για την πρόγνωση των στόχων των miRNA, δίνονται παρακάτω:

- **Diana Micro-T** (http://diana.cslab.ece.ntua.gr/microT/)
- **PicTar** (http://pictar.mdc-berlin.de/)
- **TargetScan** (http://www.targetscan.org/)
- **miRTar** (http://mirtar.mbc.nctu.edu.tw/human/)
- **miRanda** (http://www.microrna.org/microrna/home.do)
- **MaMi** (http://mami.med.harvard.edu/)
- **ComiR** (http://www.benoslab.pitt.edu/comir/) (συνδυαστική μέθοδος)
- **PITA** (http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html)
- **MirMap** (http://mirmap.ezlab.org/)
- **STarMir** (http://sfold.wadsworth.org/starmir.html)

Βιβλιογραφία

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., & Watson, J. (1994). *Molecular Biology of the Cell* (3rd ed.): Garland Publishing, Inc.

Bagos, P. G., Liakopoulos, T. D., & Hamodrakas, S. J. (2005). Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. *BMC Bioinformatics, 6*, 7. doi: 1471-2105-6-7 [pii] 10.1186/1471-2105-6-7

Bagos, P. G., Liakopoulos, T. D., Spyropoulos, I. C., & Hamodrakas, S. J. (2004a). A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC Bioinformatics, 5*, 29. doi: 10.1186/1471-2105-5-29 1471-2105-5-29 [pii]

Bagos, P. G., Liakopoulos, T. D., Spyropoulos, I. C., & Hamodrakas, S. J. (2004b). PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins. *Nucleic Acids Res, 32*(Web Server issue), W400-404. doi: 10.1093/nar/gkh41732/suppl_2/W400 [pii]

Bagos, P. G., Tsaousis, G. N., & Hamodrakas, S. J. (2009). How many 3D structures do we need to train a predictor? *Genomics Proteomics Bioinformatics, 7*(3), 128-137. doi: 10.1016/S1672-0229(08)60041-8 S1672-0229(08)60041-8 [pii]

Bagos, P. G., Tsirigos, K. D., Liakopoulos, T. D., & Hamodrakas, S. J. (2008). Prediction of lipoprotein signal peptides in Gram-positive bacteria with a Hidden Markov Model. *J Proteome Res, 7*(12), 5082-5093.

Bagos, P. G., Tsirigos, K. D., Plessas, S. K., Liakopoulos, T. D., & Hamodrakas, S. J. (2009). Prediction of signal peptides in archaea. *Protein Eng Des Sel, 22*(1), 27-35. doi: gzn064 [pii] 10.1093/protein/gzn064

Bagos, P. G., Nikolaou, E. P., Liakopoulos, T. D., & Tsirigos, K. D. (2010). Combined prediction of Tat and Sec signal peptides with hidden Markov models. *Bioinformatics*, *26*(22), 2811-2817. doi: 10.1093/bioinformatics/btq530

Baldi, P., & Brunak, S. (2001). *Bioinformatics: the machine learning approach*: MIT press.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics, 16*(5), 412-424.

Bendtsen, J. D., Nielsen, H., von Heijne, G., & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol, 340*(4), 783-795. doi: 10.1016/j.jmb.2004.05.028S0022283604005972 [pii]

Bendtsen, J. D., Nielsen, H., Widdick, D., Palmer, T., & Brunak, S. (2005). Prediction of twin-arginine signal peptides. *BMC Bioinformatics, 6*, 167. doi: 1471-2105-6-167 [pii]10.1186/1471-2105-6-167

Berks, B. C., Palmer, T., & Sargent, F. (2005). Protein targeting by the bacterial twin-arginine translocation (Tat) pathway. *Curr Opin Microbiol, 8*(2), 174-181. doi: S1369-5274(05)00021-4 [pii]10.1016/j.mib.2005.02.010

Bigelow, H. R., Petrey, D. S., Liu, J., Przybylski, D., & Rost, B. (2004). Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res, 32*(8), 2566-2577.

Bishop, C. M. (1998). *Neural Networks for Pattern Recognition*: Oxford University Press.

Cai, Y., Yu, X., Hu, S., & Yu, J. (2009). A brief review on the mechanisms of miRNA regulation. *Genomics, proteomics & bioinformatics, 7*(4), 147-154.

Chang, T.-H., Wu, L.-C., Chen, Y.-T., Huang, H.-D., Liu, B.-J., Cheng, K.-F., & Horng, J.-T. (2011). Characterization and prediction of mRNA polyadenylation sites in human genes. *Medical & biological engineering & computing, 49*(4), 463-472.

Chen, C. P., & Rost, B. (2002). State-of-the-art in membrane protein prediction. *Appl Bioinformatics, 1*(1), 21-35.

Chou, P. Y., & Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol, 47*, 45-148.

Claros, M. G., & von Heijne, G. (1994). TopPred II: an improved software for membrane protein structure predictions. *Comput Appl Biosci, 10*(6), 685-686.

Cuff, J. A., & Barton, G. J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins, 40*(3), 502-511.

Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M., & Barton, G. J. (1998). JPred: a consensus secondary structure prediction server. *Bioinformatics, 14*(10), 892-893. doi: btb130 [pii]

Diederichs, K., Freigang, J., Umhau, S., Zeth, K., & Breed, J. (1998). Prediction by a neural network of outer membrane beta-strand protein topology. *Protein Sci, 7*(11), 2413-2420.

Drew, D., Sjostrand, D., Nilsson, J., Urbig, T., Chin, C. N., de Gier, J. W., & von Heijne, G. (2002). Rapid topology mapping of Escherichia coli inner-membrane proteins by prediction and PhoA/GFP fusion analysis. *Proc Natl Acad Sci U S A, 99*(5), 2690-2695.

Driessen, A. J., & Nouwen, N. (2007). Protein Translocation Across the Bacterial Cytoplasmic Membrane. *Annu Rev Biochem*. doi: 10.1146/annurev.biochem.77.061606.160747

Eisenberg, D., Weiss, R. M., & Terwilliger, T. C. (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc Natl Acad Sci U S A, 81*(1), 140-144.

Garnier, J., Osguthorpe, D. J., & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol, 120*(1), 97-120.

Habib, S. J., Neupert, W., & Rapaport, D. (2007). Analysis and prediction of mitochondrial targeting signals. *Methods Cell Biol, 80*, 761-781. doi: S0091-679X(06)80035-X [pii]10.1016/S0091-679X(06)80035-X

Hamodrakas, S. J. (1988). A protein secondary structure prediction scheme for the IBM PC and compatibles. *Comput Appl Biosci, 4*(4), 473-477.

Hayat, S., & Elofsson, A. (2012). BOCTOPUS: improved topology prediction of transmembrane beta barrel proteins. *Bioinformatics, 28*(4), 516-522. doi: 10.1093/bioinformatics/btr710

Houben, E., de Gier, J. W., & van Wijk, K. J. (1999). Insertion of leader peptidase into the thylakoid membrane during synthesis in a chloroplast translation system. *Plant Cell, 11*(8), 1553-1564.

Jacoboni, I., Martelli, P. L., Fariselli, P., De Pinto, V., & Casadio, R. (2001). Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor. *Protein Sci, 10*(4), 779-787.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol, 292*(2), 195-202.

Juncker, A. S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H., & Krogh, A. (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci, 12*(8), 1652-1662.

Kall, L., Krogh, A., & Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *J Mol Biol, 338*(5), 1027-1036. doi: 10.1016/j.jmb.2004.03.016 S0022283604002943 [pii]

Kall, L., Krogh, A., & Sonnhammer, E. L. (2007). Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res, 35*(Web Server issue), W429-432. doi: gkm256 [pii]10.1093/nar/gkm256

Kim, H., Melen, K., & von Heijne, G. (2003). Topology models for 37 Saccharomyces cerevisiae membrane proteins based on C-terminal reporter fusions and predictions. *J Biol Chem, 278*(12), 10208-10213.

Koh, I. Y., Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Eswar, N., . . . Sali, A. (2003). EVA: evaluation of protein structure prediction servers. *Nucleic Acids Research, 31*(13), 3311-3315.

Kristiansen, K. (2004). Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function. *Pharmacol Ther, 103*(1), 21-80.

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol, 305*(3), 567-580.

Kyogoku, Y., Fujiyoshi, Y., Shimada, I., Nakamura, H., Tsukihara, T., Akutsu, H., . . . Nomura, N. (2003). Structural genomics of membrane proteins. *Acc Chem Res, 36*(3), 199-206.

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol, 157*(1), 105-132.

Lee, P. A., Tullman-Ercek, D., & Georgiou, G. (2006). The bacterial twin-arginine translocation pathway. *Annu Rev Microbiol, 60*, 373-395. doi: 10.1146/annurev.micro.60.080805.142212

Liakopoulos, T. D., Pasquier, C., & Hamodrakas, S. J. (2001). A novel tool for the prediction of transmembrane protein topology based on a statistical analysis of the SwissProt database: the OrienTM algorithm. *Protein Eng, 14*(6), 387-390.

Liu, Q., Zhu, Y. S., Wang, B. H., & Li, Y. X. (2003). A HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins. *Comput Biol Chem, 27*(1), 69-76.

Loll, P. J. (2003). Membrane protein structural biology: the high throughput challenge. *J Struct Biol, 142*(1), 144-153.

Marsh, D., Horvath, L. I., Swamy, M. J., Mantripragada, S., & Kleinschmidt, J. H. (2002). Interaction of membrane-spanning proteins with peripheral and lipid-anchored membrane proteins: perspectives from protein-lipid interactions (Review). *Mol Membr Biol, 19*(4), 247-255.

Martelli, P. L., Fariselli, P., Krogh, A., & Casadio, R. (2002). A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics, 18 Suppl 1*, S46-53.

Mathé, C., Sagot, M. F., Schiex, T., & Rouze, P. (2002). Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research, 30*(19), 4103-4117.

Melen, K., Krogh, A., & von Heijne, G. (2003). Reliability measures for membrane protein topology prediction algorithms. *J Mol Biol, 327*(3), 735-744. doi: S0022283603001827 [pii]

Morona, R., Kramer, C., & Henning, U. (1985). Bacteriophage receptor area of outer membrane protein OmpA of Escherichia coli K-12. *J Bacteriol, 164*(2), 539-543.

Pasquier, C., & Hamodrakas, S. J. (1999). An hierarchical artificial neural network system for the classification of transmembrane proteins. *Protein Eng, 12*(8), 631-634.

Pasquier, C., Promponas, V. J., & Hamodrakas, S. J. (2001). PRED-CLASS: cascading neural networks for generalized protein classification and genome-wide applications. *Proteins, 44*(3), 361-369.

Pasquier, C., Promponas, V. J., Palaios, G. A., Hamodrakas, J. S., & Hamodrakas, S. J. (1999). A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm. *Protein Eng, 12*(5), 381-385.

Pohlschroder, M., Gimenez, M. I., & Jarrell, K. F. (2005). Protein transport in Archaea: Sec and twin arginine translocation pathways. *Curr Opin Microbiol, 8*(6), 713-719. doi: S1369-5274(05)00162-1 [pii] 10.1016/j.mib.2005.10.006

Prince, S. M., Achtman, M., & Derrick, J. P. (2002). Crystal structure of the OpcA integral membrane adhesin from Neisseria meningitidis. *Proc Natl Acad Sci U S A, 99*(6), 3417-3421.

Promponas, V. J., Palaios, G. A., Pasquier, C. M., Hamodrakas, J. S., & Hamodrakas, S. J. (1999). CoPreTHi: a Web tool which combines transmembrane protein segment prediction methods. *In Silico Biol, 1*(3), 159-162. doi: 1998010014 [pii]

Przybylski, D., & Rost, B. (2007). Consensus sequences improve PSI-BLAST through mimicking profile–profile alignments. *Nucleic Acids Research, 35*(7), 2238-2246.

Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol, 202*(4), 865-884.

Rapoport, T. A., Matlack, K. E., Plath, K., Misselwitz, B., & Staeck, O. (1999). Posttranslational protein translocation across the membrane of the endoplasmic reticulum. *Biol Chem, 380*(10), 1143-1150.

Rapp, M., Drew, D., Daley, D. O., Nilsson, J., Carvalho, T., Melen, K., Von Heijne, G. (2004). Experimentally based topology models for E. coli inner membrane proteins. *Protein Sci, 13*(4), 937-945.

Reinhardt, A., & Hubbard, T. (1998). Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res, 26*(9), 2230-2236.

Reynolds, S. M., Kall, L., Riffle, M. E., Bilmes, J. A., & Noble, W. S. (2008). Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput Biol, 4*(11), e1000213. doi: 10.1371/journal.pcbi.1000213

Ringler, P., & Schulz, G. E. (2002). OmpA membrane domain as a tight-binding anchor for lipid bilayers. *Chembiochem, 3*(5), 463-466.

Rojo, E. E., Guiard, B., Neupert, W., & Stuart, R. A. (1999). N-terminal tail export from the mitochondrial matrix. Adherence to the prokaryotic "positive-inside" rule of membrane protein topology. *J Biol Chem, 274*(28), 19617-19622.

Rose, R. W., Bruser, T., Kissinger, J. C., & Pohlschroder, M. (2002). Adaptation of protein secretion to extremely high-salt conditions by extensive use of the twin-arginine translocation pathway. *Mol Microbiol, 45*(4), 943-950. doi: 3090 [pii]

Rost, B., Casadio, R., Fariselli, P., & Sander, C. (1995). Transmembrane helices predicted at 95% accuracy. *Protein Sci, 4*(3), 521-533.

Rost, B., & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol, 232*(2), 584-599.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive modeling, 5*, 3.

Saeys, Y., Abeel, T., Degroeve, S., & Van de Peer, Y. (2007). Translation initiation site prediction on a genomic scale: beauty in simplicity. *Bioinformatics, 23*(13), i418-i423.

Savojardo, C., Fariselli, P., & Casadio, R. (2013). BETAWARE: a machine-learning tool to detect and predict transmembrane beta-barrel proteins in prokaryotes. *Bioinformatics, 29*(4), 504-505. doi: 10.1093/bioinformatics/bts728

Schulz, G. E. (2002). The structure of bacterial outer membrane proteins. *Biochim Biophys Acta, 1565*(2), 308-317.

Schulz, G. E. (2003). Transmembrane beta-barrel proteins. *Adv Protein Chem, 63*, 47-70.

Sgourakis, N. G., Bagos, P. G., & Hamodrakas, S. J. (2005). Prediction of the coupling specificity of GPCRs to four families of G-proteins using hidden Markov models and artificial neural networks. *Bioinformatics, 21*(22), 4101-4106. doi: bti679 [pii] 10.1093/bioinformatics/bti679

Sgourakis, N. G., Bagos, P. G., Papasaikas, P. K., & Hamodrakas, S. J. (2005). A method for the prediction of GPCRs coupling specificity to G-proteins using refined profile Hidden Markov Models. *BMC Bioinformatics, 6*, 104. doi: 1471-2105-6-104 [pii]10.1186/1471-2105-6-104

Singer, S. J., & Nicolson, G. L. (1972). The fluid mosaic model of the structure of cell membranes. *Science, 175*(23), 720-731.

Singh, N. K., Goodman, A., Walter, P., Helms, V., & Hayat, S. (2011). TMBHMM: a frequency profile based HMM for predicting the topology of transmembrane beta barrel proteins and the exposure status of transmembrane residues. *Biochim Biophys Acta, 1814*(5), 664-670. doi:10.1016/j.bbapap.2011.03.004

Sonnhammer, E. L., von Heijne, G., & Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol, 6*, 175-182.

Sugawara, E., & Nikaido, H. (1992). Pore-forming activity of OmpA protein of Escherichia coli. *J Biol Chem, 267*(4), 2507-2511.

Sugawara, E., & Nikaido, H. (1994). OmpA protein of Escherichia coli outer membrane occurs in open and closed channel forms. *J Biol Chem, 269*(27), 17981-17987.

Teter, S. A., & Klionsky, D. J. (1999). How to get a folded protein across a membrane. *Trends Cell Biol, 9*(11), 428-431. doi: S0962-8924(99)01652-9 [pii]

Tsaousis, G. N., Bagos, P. G., & Hamodrakas, S. J. (2014). HMMpTM: improving transmembrane protein topology prediction using phosphorylation and glycosylation site prediction. *Biochim Biophys Acta, 1844*(2), 316-322. doi: 10.1016/j.bbapap.2013.11.001S1570-9639(13)00376-2 [pii]

Tusnady, G. E., Dosztanyi, Z., & Simon, I. (2004). Transmembrane proteins in protein data bank: identification and classification. *Bioinformatics*.

Tusnady, G. E., & Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol, 283*(2), 489-506.

Tusnady, G. E., & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics, 17*(9), 849-850.

Tuteja, R. (2005). Type I signal peptidase: an overview. *Arch Biochem Biophys, 441*(2), 107-111. doi: S0003-9861(05)00305-X [pii]10.1016/j.abb.2005.07.013

van Roosmalen, M. L., Geukens, N., Jongbloed, J. D., Tjalsma, H., Dubois, J. Y., Bron, S., . . . Anne, J. (2004). Type I signal peptidases of Gram-positive bacteria. *Biochim Biophys Acta, 1694*(1-3), 279-297. doi: S0167488904001235 [pii]10.1016/j.bbamcr.2004.05.006

Vandeputte-Rutten, L., Bos, M. P., Tommassen, J., & Gros, P. (2003). Crystal structure of Neisserial surface protein A (NspA), a conserved outer membrane protein with vaccine potential. *J Biol Chem, 278*(27), 24825-24830.

Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics, 13*(Suppl 4), S2.

Vogt, J., & Schulz, G. E. (1999). The structure of the outer membrane protein OmpX from Escherichia coli reveals possible mechanisms of virulence. *Structure Fold Des, 7*(10), 1301-1309.

von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res, 14*(11), 4683-4690.

von Heijne, G. (1990). The signal peptide. *J Membr Biol, 115*(3), 195-201.

von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol, 225*(2), 487-494.

von Heijne, G. (1999). Recent advances in the understanding of membrane protein assembly and function. *Quart Rev Biophys, 32*(4), 285-307.

von Heijne, G., Steppuhn, J., & Herrmann, R. G. (1989). Domain structure of mitochondrial and chloroplast targeting peptides. *Eur J Biochem, 180*(3), 535-545.

Walian, P., Cross, T. A., & Jap, B. K. (2004). Structural genomics of membrane proteins. *Genome Biol, 5*(4), 215.

White, S. H. (2004). The progress of membrane protein structure determination. *Protein Sci, 13*(7), 1948-1949.

Zemla, A., Venclovas, C., Fidelis, K., & Rost, B. (1999). A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins, 34*(2), 220-223.

Zhai, Y., & Saier, M. H., Jr. (2002). The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes. *Protein Sci, 11*(9), 2196-2207.

Κεφάλαιο 8: Μαρκοβιανά Μοντέλα

Σύνοψη
Στο κεφάλαιο αυτό, θα γίνει η απαραίτητη εισαγωγή στα μαρκοβιανά μοντέλα εξάρτησης και κατόπιν, παρουσίαση των κρυπτομαρκοβιανών μοντέλων (Hidden Markov Models) τα οποία αποτελούν ένα σημαντικό εργαλείο στη σύγχρονη βιοπληροφορική. Θα αναφερθούμε στα βασικά χαρακτηριστικά των μοντέλων αυτών και στη μαθηματική τους θεμελίωση, ενώ θα παρουσιαστούν σε βάθος οι διάφοροι αλγόριθμοι που χρησιμοποιούνται για τον υπολογισμό της πιθανοφάνειας, για την αποκωδικοποίηση και για την εκτίμηση παραμέτρων στα μοντέλα αυτά. Θα παρουσιαστούν επίσης, τα μοντέλα για σημασμένες αλληλουχίες, τα οποία αποτελούν μια επέκταση του βασικού HMM, η οποία βρίσκει πολλές εφαρμογές στην ανάλυση βιολογικών αλληλουχιών (πρόγνωση διαμεμβρανικών πρωτεϊνών, εύρεση γονιδίων κ.ο.κ.). Τέλος, θα γίνει ειδική αναφορά στο profile HMM το οποίο είναι άλλη μια παραλλαγή του βασικού μοντέλου, η οποία βρίσκει εφαρμογές στη μοντελοποίηση πρωτεϊνικών οικογενειών, στην εύρεση μακρινών ομολόγων και στην πολλαπλή στοίχιση.

Προαπαιτούμενη γνώση
Βασικές γνώσεις πιθανοτήτων. Κατανόηση των εννοιών της στοίχισης και πολλαπλής στοίχισης αλληλουχιών που μελετήθηκαν στα κεφάλαια 3 και 4.

8. Εισαγωγή

Στο κεφάλαιο αυτό θα μελετήσουμε μαθηματικά μοντέλα τα οποία ανήκουν σε μια μεγάλη οικογένεια στοχαστικών-πιθανοθεωρητικών μοντέλων, τα οποία ονομάζονται μοντέλα εξάρτησης του Markov ή αλλιώς Μαρκοβιανά μοντέλα. Θα εισαγάγουμε αρχικά την έννοια της αλυσίδας Markov (Markov Chain), η οποία βρίσκει σημαντικές εφαρμογές στη δημιουργία μοντέλων που περιγράφουν αλληλουχίες DNA ή και πρωτεϊνών. Η θεώρηση μιας ακολουθίας ενδεχομένων ως αλυσίδα Markov στηρίζεται, πολύ απλά, στην ιδέα ότι κάθε ένα από τα ενδεχόμενα εξαρτάται μόνο από το αμέσως προηγούμενό του ή αλλιώς το κάθε ενδεχόμενο καθορίζει με κάποια πιθανότητα το αμέσως επόμενό του. Αν αυτή η εξάρτηση επεκταθεί και σε *2,3,...,k* προηγούμενα ενδεχόμενα τότε μιλάμε για αλυσίδες Markov $2^{ης}$,$3^{ης}$, ...,$k^{ης}$ τάξης.

Πρέπει να τονιστεί εδώ, ότι το μοντέλο Markov θεωρείται από πολλούς ερευνητές ως το πιο φυσικό για να περιγράψει αλληλουχίες μεγαλομορίων όπως του DNA αλλά και των πρωτεϊνών, και αυτό φαίνεται διαισθητικά φυσικό καθώς αυτή η εξάρτηση φαίνεται να προσεγγίζει την έννοια της πληροφορίας που εμπεριέχεται σε μια αλληλουχία. Ήδη από τη δεκαετία του 1970 τα μοντέλα αυτά χρησιμοποιούνταν και χρησιμοποιούνται ακόμα με σκοπό την αναγνώριση και επεξεργασία εικόνας, ήχου κ.α. και υπάρχει πλούσια βιβλιογραφία πάνω στα θέματα αυτά. Η πιο απλή εξήγηση για τα παραπάνω είναι το γεγονός ότι σε οποιοδήποτε κωδικοποιημένο σύστημα επικοινωνίας όπως στις φυσικές γλώσσες, υπάρχει μια εσωτερική δομή που καθορίζει κάποιο είδος εξάρτησης των συμβόλων. Για παράδειγμα, στην αγγλική γλώσσα το γράμμα Q ακολουθείται σχεδόν πάντοτε από το U, άρα η πιθανότητα να εμφανιστεί το U σε μια θέση δεν είναι πάντα ίδια αλλά εξαρτάται από το αν προηγήθηκε το Q. Για την ακρίβεια, ο ίδιος ο Ρώσος Μαθηματικός Andrey Markov (1856-1922) οδηγήθηκε στην σύλληψη της έννοιας των ομώνυμων αλυσίδων, μελετώντας τις εναλλαγές φωνηέντων και συμφώνων σε κάποιο ποίημα του Pushkin (Markov, 1913). Θα προχωρήσουμε στην συνέχεια στον τυπικό ορισμό του μοντέλου Markov (Markov Model-MM) αλλά και του «κρυμμένου» μοντέλου Markov (Hidden Markov Model-HMM) και θα εξετάσουμε τις κυριότερες εφαρμογές τους.

8.1. Αλυσίδες Markov

8.1.1. Ορισμοί

Μια αλυσίδα Markov $1^{ης}$ τάξης ορίζεται ως μια στοχαστική ανέλιξη διακριτών καταστάσεων σε διακριτό χρόνο. Στην περίπτωση των βιολογικών αλληλουχιών, ως καταστάσεις ορίζονται τα σύμβολα της ακολουθίας τα οποία ανήκουν σε ένα πεπερασμένο αλφάβητο, Ω (τα τέσσερα νουκλεοτίδια στην περίπτωση του DNA ή

τα 20 αμινοξέα στην περίπτωση των πρωτεϊνών). Αν θεωρήσουμε μια πρωτεϊνική αλληλουχία μήκους L καταλοίπων, και την ονομάσουμε **x**, έτσι ώστε:

$$\mathbf{x} = x_1, x_2, ..., x_{L-1}, x_L$$

και θεωρήσουμε την κατανομή των αμινοξέων σε κάθε θέση i κατά μήκος της αλληλουχίας ως τυχαία μεταβλητή, τότε μπορούμε να ορίσουμε την αλυσίδα Markov, ως μια στοχαστική ανέλιξη η οποία διαθέτει τη λεγόμενη «Μαρκοβιανή Ιδιότητα». Στη διακριτή περίπτωση (όπως η συγκεκριμένη), η ανέλιξη αποτελείται από την ακολουθία των τυχαίων μεταβλητών **x**, η οποία παίρνει τιμές σε ένα «χώρο καταστάσεων» οριζόμενο από το συγκεκριμένο αλφάβητο. Όπως είδαμε, οι τιμές των x_i συμβολίζουν την «κατάσταση στην οποία βρίσκεται το σύστημα την χρονική στιγμή *i*». Η Μαρκοβιανή ιδιότητα (σε διακριτό χρόνο) ορίζει ότι η δεσμευμένη κατανομή των «μελλοντικών» παρατηρήσεων x_{i+1}, x_{i+2}, x_{i+3} δεδομένου του «παρελθόντος» x_1, x_2, ...x_{i-1}, x_i, εξαρτάται από το παρελθόν μόνο μέσω του x_i. Με άλλα λόγια, η γνώση της πιο πρόσφατης κατάστασης του συστήματος καθιστά τη λιγότερο πρόσφατη ιστορία άχρηστη. Αυτό τυπικά διατυπώνεται ως εξής:

$$P(x_i \mid x_{i-1}, ..., x_1) = P(x_i \mid x_{i-1}) \tag{8.1}$$

Μια συγκεκριμένη Αλυσίδα Markov χαρακτηρίζεται από τον πίνακα των «πιθανοτήτων μετάβασης» (transition probabilities), ο οποίος πιο απλά ονομάζεται πίνακας μεταβάσεων. Τα στοιχεία αυτού του πίνακα, δίνονται από την παρακάτω σχέση:

$$a_{st} = P(x_i = t \mid x_{i-1} = s) = \alpha_{x_{i-1}x_i} \tag{8.2}$$

η οποία δηλώνει, την πιθανότητα το κατάλοιπο t να εμφανιστεί στη θέση i της αλληλουχίας, δεδομένου ότι το προηγούμενο κατάλοιπο (i-1) είναι s. Αν αναλογιστούμε ότι μπορούμε να γενικεύσουμε την εξάρτηση στα k προηγούμενα κατάλοιπα, είναι φυσικό η δεδομένη αλυσίδα να ονομάζεται Αλυσίδα Markov $1^{ης}$ τάξεως. Η συνολική πιθανότητα μιας αλληλουχίας υπολογίζεται ως εξής:

$$P(\mathbf{x}) = P(x_1, x_2, ..., x_{L-1}, x_L) = P(x_L \mid x_{L-1}, ..., x_1) P(x_{L-1} \mid x_{L-2}, ..., x_1) ... P(x_1)$$

και από τη σχέση (8.1), έχουμε:

$$P(\mathbf{x}) = P(x_L \mid x_{L-1}) P(x_{L-1} \mid x_{L-2}) ... P(x_2 \mid x_1) P(x_1) = P(x_1) \prod_{i=2}^{L} P(x_i \mid x_{i-1}) = P(x_1) \prod_{i=2}^{L} \alpha_{x_{i-1}x_i} \tag{8.3}$$

όπου $P(x_1)$ είναι η πιθανότητα για την εμφάνιση του πρώτου συμβόλου. Σύμφωνα με τον ορισμό αυτό, βλέπουμε ότι οι πιθανότητες μεταβάσεως είναι ίδιες, ανεξαρτήτως της θέσης τους στην αλυσίδα δηλαδή:

$p_{ab}(n-1, n) = P(x_i = b \mid x_{i-1} = a) = p_{ab}$ για κάθε n=1,2,..L.

Η αλυσίδα αυτή λεμε ότι έχει στάσιμες πιθανότητες μεταβάσεως, ή, ισοδύναμα, ότι η αλυσίδα αυτή είναι *ομογενής χρονικά*. Ο περιορισμός αυτός χρησιμοποιείται πολλές φορές στις περιπτώσεις μακρομορίων (αν και υπάρχουν εξαιρέσεις, όπως θα αναφέρουμε), κυρίως μεν γιατί προσφέρει υπολογιστική απλότητα αλλά και γιατί δεν έχουμε, στις περισσότερες περιπτώσεις, καμία ένδειξη που να υποστηρίζει μια τέτοια εξάρτηση από τη θέση στην αλυσίδα. Ο πίνακας ο οποίος περιέχει τις πιθανότητες μεταβάσεως, όπως είδαμε, λέγεται πίνακας *πιθανοτήτων μεταβάσεως* ή *πίνακας μεταβάσεως* $1^{ης}$ *τάξης* και πρέπει για ένα αλφάβητο με πλήθος k να ικανοποιεί τα παρακάτω:

$$p_{a,b} \geq 0 \text{ για } a, b = 1,2,..., k$$

και $$\sum_{b=1}^{k} p_{a,b} = 1 \text{ για κάθε } b=1,2,...,k$$

Γενικότερα κάθε τετραγωνικός πίνακας που ικανοποιεί τις δυο αυτές σχέσεις, λέγεται *στοχαστικός*. Όπως είδαμε από τις παραπάνω, σχέσεις ορίζεται πλήρως μια αλυσίδα Markov, αρκεί να ορίσουμε επιπλέον μια πιθανότητα για την κατάσταση της έναρξης της αλυσίδας (B=Begin). Η πιθανότητα αυτή ονομάζεται αρχική πιθανότητα και ορίζεται ως:

$$P(x_1 = a) = p_{Ba} \tag{8.4}$$

Όμοια μπορούμε να ορίσουμε (χωρίς όμως και να είναι απαραίτητο) μια άλλη τελική κατάσταση (E=End) για τον τερματισμό της αλυσίδας με πιθανότητα:

$$P(E \mid x_n = b) = p_{bE} \tag{8.5}$$

Έτσι πλέον μια πλήρης σχηματική αναπαράσταση του μοντέλου Markov φαίνεται στην Εικόνα 8.1 παρακάτω. Παραδοσιακά η λήξη της αλληλουχίας δεν συμπεριλαμβάνεται στο μοντέλο, θεωρούμε δηλαδή

ότι η αλυσίδα μπορεί να τελειώνει οπουδήποτε. Το πλεονέκτημα του να συμπεριληφθεί αυτή η κατάσταση στο μοντέλο, είναι όταν θέλουμε να μελετήσουμε την κατανομή του μήκους της αλυσίδας. Έτσι αν

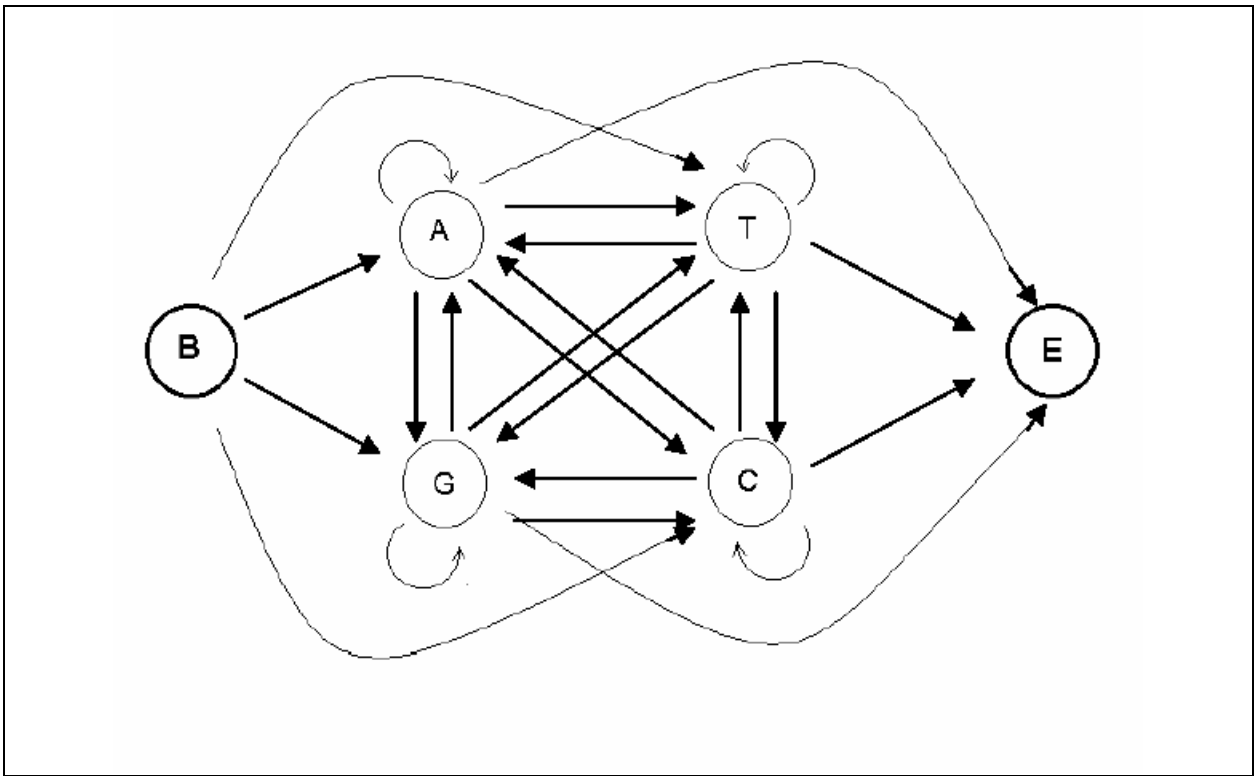
$$P(E \mid x_n = b) = p_{bE} = q$$

τότε η κατανομή του αθροίσματος των πιθανοτήτων της σχέσης (8.3) για μια αλληλουχία μήκους L είναι:

$$p_{o\lambda} = q(1-q)^{L-1} \tag{8.6}$$

δηλαδή η κατανομή του αθροίσματος των πιθανοτήτων για όλες τις αλληλουχίες μήκους L ακολουθεί γεωμετρική κατανομή. Αντίστοιχα το άθροισμα των πιθανοτήτων όλων των πιθανών ακολουθιών είναι (Durbin, Eddy, Krogh, & Mithison, 1998):

$$p_{o\lambda} = \sum_{\{X\}} P(\mathrm{x}) = \sum \sum \ldots \sum P(x_1) \prod_{i=2}^{n} P(x_i \mid x_{i-1}) = 1 \tag{8.7}$$



Εικόνα 8.1: *Ένα τυπικό μοντέλο αλυσίδας Markov, με καταστάσεις τις 4 βάσεις του DNA. Τα βέλη συμβολίζουν τις επιτρεπτές μεταβάσεις.Με B και E, συμβολίζονται οι καταστάσεις έναρξης και τερματισμού του μοντέλου, αντίστοιχα.*

8.1.2. Εκτίμηση Παραμέτρων

Οι εκτιμητές μέγιστης πιθανοφάνειας (Maximum Likelihood Estimates-MLEs) των πιθανοτήτων μεταβάσεως, υπολογίζονται σύμφωνα με τη σχέση:

$$\hat{\alpha}_{x_{i-1}x_i} = \frac{n_{st}}{\sum_{t'} n_{st'}} \tag{8.8}$$

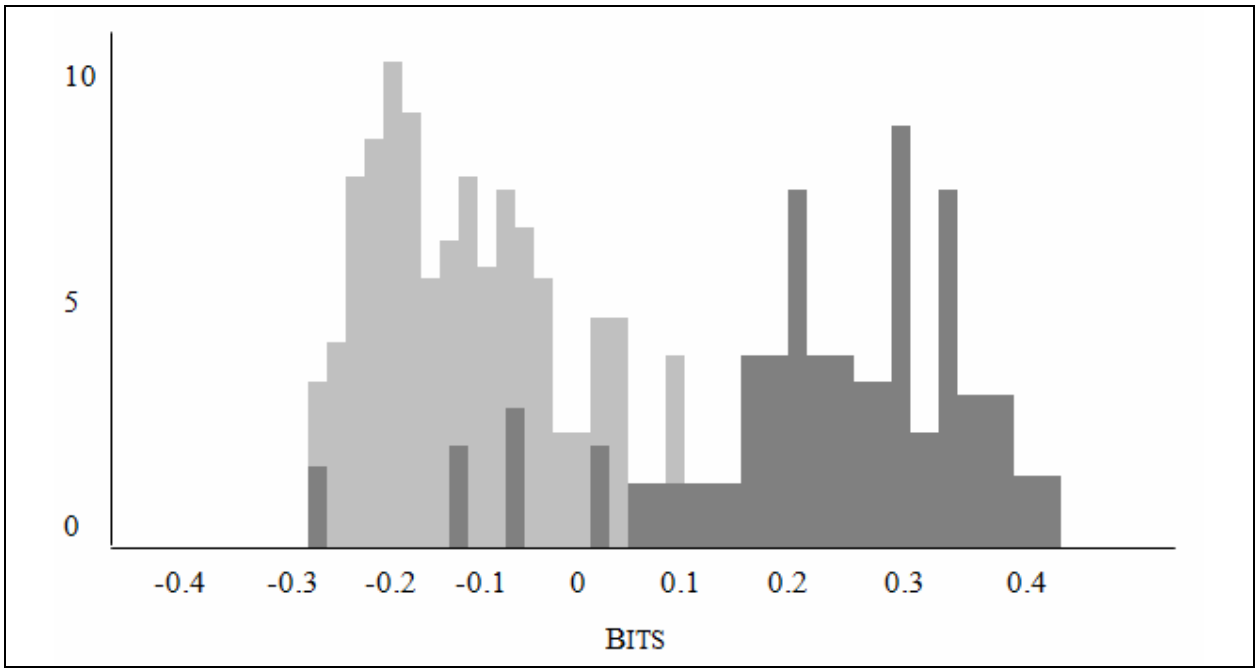
όπου n_{st} είναι οι παρατηρούμενες εμφανίσεις του καταλοίπου s ακολουθούμενο από το κατάλοιπο t στις αλληλουχίες εκπαίδευσης, με το άθροισμα στον παρονομαστή να εκτείνεται σε όλο το αλφάβητο των 20 αμινοξέων (ή των νουκλεοτιδίων αν μιλάμε για DNA). Θεωρώντας δυο διαφορετικά μοντέλα, με τη χρήση δυο πινάκων μεταβάσεων (για παράδειγμα, ένα μοντέλο + για τα λεγόμενα θετικά παραδείγματα και ένα μοντέλο – για τα λεγόμενα αρνητικά), μπορούμε να ορίσουμε ένα *log-odds score,* $S(\mathbf{x})$ για ολόκληρη την αλληλουχία, το οποίο είναι χρήσιμο για διαχωριστικούς σκοπούς:

$$S(\mathbf{x}) = \log \frac{P(\mathbf{x}|+)}{P(\mathbf{x}|-)} = \sum_{i=1}^{L} \log\left(\frac{\alpha^{+}_{x_{i-1}x_i}}{\alpha^{-}_{x_{i-1}x_i}}\right) = \sum_{i=1}^{L} \beta_{x_{i-1}x_i} \tag{8.9}$$

όπου $\beta_{x_{i-1}x_i}$, είναι το log-odds για την πιθανότητα μετάβασης από το κατάλοιπο x_{i-1} στο x_i, και είναι ένα σχετικό μέτρο της τάσης των πιθανοτήτων μετάβασης να εμφανίζονται πιο συχνά στο ένα ή το άλλο μοντέλο. Το σκορ αυτό, είναι εντελώς ανάλογο με τα αντίστοιχα που είδαμε στο κεφάλαιο 2, όπου και μελετούσαμε τις αλληλουχίες, κάτω από τις προϋποθέσεις του μοντέλου της ανεξαρτησίας. Τιμές των $\beta_{x_{i-1}x_i}$ μεγαλύτερες από το 0, υποδηλώνουν προτιμήσεις των συγκεκριμένων μεταβάσεων για το μοντέλο (+), ενώ τιμές μικρότερες από το 0 προτίμηση για το μοντέλο (-). Για να εκμηδενίσουμε την επιρροή του μήκους των ακολουθιών στο συνολικό σκορ, κανονικοποιούμε περαιτέρω τις τιμές, διαιρώντας με το μήκος L της αλληλουχίας έτσι ώστε να πάρουμε ένα log-odds score ανά κατάλοιπο.

$$S^{norm}(\mathbf{x}) = \frac{S(\mathbf{x})}{L} = \frac{\sum_{i=1}^{L} \beta_{x_{i-1}x_i}}{L} \tag{8.10}$$

Χαρακτηριστικό παράδειγμα, 1ης τάξης μοντέλου με την παραπάνω διατύπωση, αναφέρεται στην εύρεση νησίδων CG στα ευκαρυωτικά γονιδιώματα (Durbin, et al., 1998).



Εικόνα 8.2: *Ένα παράδειγμα μαρκοβιανής αλυσίδας για την εύρεση νησίδων CG στα ευκαρυωτικά γονιδιώματα. Ένας αριθμός πραγματικών γονιδίων και ένας αριθμός μη-γονιδιών, αναλύθηκαν με τη βοήθεια των σχέσεων* (8.8) *και* (8.9) *και τα αποτελέσματα για το Σκορ παρατίθενται σε ένα απλό ιστόγραμμα συχνοτήτων. Παρατηρούμε ότι οι δύο κατανομές διαχωρίζονται ικανοποιητικά. Τα bits στις τιμές του σκορ, αναφέρονται σε λογάριθμο με βάση το 2.*

8.1.3. Αλυσίδες ανώτερης τάξεως

Μια $k^{ης}$ τάξεως αλυσίδα Markov, μπορεί να προκύψει αυτόματα από γενίκευση της Μαρκοβιανής ιδιότητας της εξισώσεως (8.1). Συγκεκριμένα, η σχέση αυτή τροποποιείται έτσι ώστε να συμπεριλάβει εξάρτηση στις k προηγούμενες παρατηρήσεις:

$$P(x_i \mid x_{i-1}, \ldots, x_1) = P(x_i \mid x_{i-1}, x_{i-2}, \ldots, x_{i-k}) = \alpha_{x_k \ldots x_{i-1} x_i} \tag{8.11}$$

Δεδομένου ότι ισχύει:

$$P(x_i \mid x_{i-1}, x_{i-2}, \ldots, x_{i-k}) = P(x_i, x_{i-1}, \ldots, x_{i-k+1} \mid x_{i-1}, x_{i-2}, \ldots, x_{i-k}) \tag{8.12}$$

η $k^{ης}$ τάξεως αλυσίδα Markov, είναι ισοδύναμη με μια αλυσίδα $1^{ης}$ τάξεως, αλλά με ένα αλφάβητο της τάξης του 20^k. Κατά συνέπεια, απαιτεί τον υπολογισμό πινάκων μεταβάσεων μεγέθους 20^kx20^k. Άρα, στην

περίπτωση των πρωτεϊνών, ενώ για μια αλυσίδα $1^{ης}$ τάξεως χρειαζόμαστε να υπολογίσουμε 20^2=400 πιθανότητες για κάθε μοντέλο, για ένα μοντέλο $2^{ης}$ τάξεως χρειαζόμαστε 20^3=8000 παραμέτρους, αριθμός υπερβολικά μεγάλος ο οποίος στην περίπτωση αλληλουχιών πρωτεϊνών θα απαιτούσε υπερβολικά μεγάλο αριθμό ακολουθιών για να χρησιμοποιηθούν ως παραδείγματα για την εκπαίδευση των μοντέλων. Περιπτώσεις αλυσίδων ανώτερης τάξης είναι δυνατόν να εφαρμοστούν πιο εύκολα σε αλληλουχίες νουκλεοτιδίων, όπου το αλφάβητο είναι μικρότερο και οι αλληλουχίες πολύ μεγαλύτερες (Ellrott, Yang, Sladek, & Jiang, 2002; Phillips, Arnold, & Ivarie, 1987). Σε μια ενδιαφέρουσα εργασία, οι Audic και Claverie (Audic & Claverie, 1998), χρησιμοποίησαν αλυσίδες ανώτερης τάξης σε συνδυασμό με μια υπολογιστικά εντατική μεθοδολογία έτσι ώστε να ανιχνεύσουν διαφορετικής σύστασης περιοχές σε βακτηριακά γονιδιώματα. Με αυτό, τον τρόπο, διαχώρισαν χωρίς την ανάγκη συνόλου εκπαίδευσης, περιοχές οι οποίες κωδικοποιούν πρωτεΐνες, περιοχές που δεν κωδικοποιούν τίποτα και περιοχές που κωδικοποιούν πρωτεΐνες αλλά στη συμπληρωματική τους αλυσίδα, σε ποσοστό που έφτανε και το 90% (Audic & Claverie, 1998).

Σε περιπτώσεις πρωτεϊνών, έχει προταθεί η προσέγγιση των πιθανοτήτων μετάβασης μεγαλύτερης τάξης. Συγκεκριμένα, η πιθανότητα μετάβασης για μια $k^{ης}$ τάξης αλυσίδα θα μπορούσε να προσεγγισθεί (Yuan, 1999), από τη σχέση:

$$P\left(x_i \mid x_{i-1}, x_{i-2}, ..., x_{i-k}\right) \approx \prod_{j=1}^{k} P\left(x_i \mid x_{i-j}\right) \tag{8.13}$$

Η σχέση (8.13) χρησιμοποιήθηκε (Yuan, 1999) στην προσπάθεια να προβλεφθεί η υποκυτταρική τοποθεσία των βακτηριακών πρωτεϊνών, με αρκετή επιτυχία. Σε γενικές γραμμές, αναμένουμε ότι με μεγαλύτερης τάξεως αλυσίδες θα έχουμε και καλύτερη διαχωριστική ικανότητα των μοντέλων, γεγονός που επιβεβαιώνεται και από τη μελέτη αυτή (Yuan, 1999). Από την άλλη μεριά, μεγαλώνοντας πάρα πολύ την τάξη (>6), ακόμα και για νουκλεοτιδικές αλληλουχίες, πέραν του προβλήματος υπερ-προσαρμογής (over-fitting) και της έλλειψης δεδομένων, ανακύπτει και το πρόβλημα της εισαγωγής θορύβου, από μη-σημαντικές μακρινές αλληλεπιδράσεις (Ellrott, et al., 2002; Phillips, et al., 1987; Yuan, 1999). Άλλη μέθοδος, χρήσιμη κυρίως σε αλληλουχίες νουκλεοτιδίων, είναι αυτή της χρήσης μη-ομογενών αλυσίδων (non-homogenous Markov chains), με την οποία χρησιμοποιούνται διαφορετικοί πίνακες μεταβάσεων, έτσι ώστε να εντοπιστούν καλύτερα οι στατιστικές προτιμήσεις στις διάφορες θέσεις της τριπλέτας βάσεων μιας κωδικής περιοχής (Borodovsky & Peresetsky, 1994).

Μεγάλο ενδιαφέρον, τόσο πρακτικό όσο και θεωρητικό, παρουσιάζουν τα μοντέλα Markov μεταβλητού μήκος (Variable length Markov Models-VMM), τα οποία όπως διατυπώθηκαν από τον Bejerano (Bejerano, 2004) είναι μια επέκταση της διατύπωσης των Στοχαστικών Πεπερασμένων Αυτομάτων (Probabilistic Finite Automata-PFA), από τον Ron και τους συνεργάτες του (Ron, Singer, & Tishby, 1996). Το μοντέλο αυτό, αντί να υπολογίζει όλα τα C πλαίσια παραθύρων μήκους n, τα οποία θα καθορίσουν τις παραμέτρους της αλυσίδας $k^{ης}$ τάξης, υπολογίζει παραμέτρους μόνο για ένα υποσύνολο $C^* \subset C$, το οποίο προσδιορίζεται με εκπαίδευση από τα δεδομένα και προβλέπει μεγαλύτερες εξαρτήσεις στα προηγηθέντα κατάλοιπα, όταν είναι απαραίτητο, ενώ μικρότερες όταν δεν είναι. Έτσι, η προσεγγιστική σχέση που χρησιμοποιείται, είναι η εξής:

$$P\left(x_i \mid x_{i-1}, x_{i-2}, ..., x_{i-k}\right) \approx P\left(x_i \mid \max_{k_i \geq 0}\left\{x_{i-k_i}, ..., x_{i-1} \in C\right\}\right) \tag{8.14}$$

Με τη χρήση αυτής της σχέσης (και μιας πολύπλοκης διαδικασίας εκπαίδευσης που δεν θα αναφερθεί εδώ), ο Bejerano και οι συνεργάτες του (Bejerano, Seldin, Margalit, & Tishby, 2001; Bejerano & Yona, 2001), κατάφεραν να κατασκευάσουν μοντέλα τα οποία διακρίνουν με αρκετά μεγάλη ακρίβεια σχεδόν όλες τις οικογένειες πρωτεϊνών που είναι κατατεθειμένες στην βάση δεδομένων PFAM (Bateman et al., 2004). Η προσέγγιση αυτή, έχει ενδιαφέρον γιατί έδειξε ότι απλούστερα αλλά καλής προγνωστικής αξίας μοντέλα, μπορούν να κατασκευαστούν, και να συναγωνίζονται σε επιτυχία τα πιο πολύπλοκα Hidden Markov Models (βλ. παρακάτω).

Μια άλλη προσέγγιση, είναι το λεγόμενο Mixture Transition Distribution (MTD) model το οποίο προτάθηκε αρχικά από τον (Raftery, 1985a), και στο οποίο οι πιθανότητες μετάβασης της σχέσης (8.11) προσεγγίζονται από τη σχέση:

$$a_{s_k \ldots s_1 s_0} = P\left(x_i \mid x_{i-1}, x_{i-2}, ..., x_{i-k}\right) = \sum_{j=1}^{k} \lambda_j \alpha_{s_j s_0} \tag{8.15}$$

Έτσι, η επίδραση κάθε παλιάς παρατήρησης (j=1,2…k) λαμβάνεται υπόψη ξεχωριστά και τελικά η ανώτερης τάξης πιθανότητα μετάβασης υπολογίζεται σαν ένας γραμμικός συνδυασμός πιθανοτήτων

μετάβασης πρώτης τάξης. Στην πιο γενική μορφή του μοντέλου (MTDg), η οποία προτάθηκε αργότερα (Raftery, 1985b), κάθε παλιά θέση j συνοδεύεται από διαφορετικό πίνακα μεταβάσεων, α^j και έτσι έχουμε:

$$a_{s_k \ldots s_1 s_0} = P\left(x_i \mid x_{i-1}, x_{i-2}, \ldots, x_{i-k}\right) = \sum_{j=1}^{k} \lambda_j \alpha^j_{s_j s_0} \tag{8.16}$$

Προφανώς, για να διατηρεί το μοντέλο την πιθανοθεωρητική του ερμηνεία θα πρέπει να ισχύει:

$$0 \le \sum_{j=1}^{k} \lambda_j \alpha^j_{s_j s_0} \le 1 \tag{8.17}$$

και $\sum_{s_k,\ldots,s_1,s_0,\forall s_0 \in Q} \left(\sum_{j=1}^{k} \lambda_j \alpha^j_{s_j s_0} \right) = 1$, και κατά συνέπεια οι παρακάτω περιορισμοί θα πρέπει να ισχύουν:

$$\sum_{j=1}^{k} \lambda_j = 1, \lambda_j \ge 0 \ \forall \ j = 1, 2, \ldots, k \tag{8.18}$$

Ο Raftery, μελέτησε τις ασυμπτωτικές ιδιότητες του μοντέλου αυτού και έδειξε ότι προσεγγίζει την ανώτερης τάξης αλυσίδα Markov. Παρ' όλα αυτά, αναλυτικές εκφράσεις για τους εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων δεν μπορούν να βρεθούν, και κατά συνέπεια χρειάζεται κάποιου είδους επαναληπτική διαδικασία. Ο Raftery, στην αρχική εργασία (Raftery, 1985a), βελτιστοποίησε την πιθανοφάνεια με μια ρουτίνα γραμμικής βελτιστοποίησης με περιορισμούς (NAG). Ο Bercthold, πρότεινε μια μορφή ευριστικού αλγορίθμου που χρησιμοποιεί το gradient (Berchtold, 2001). Τέλος, μια προσέγγιση που βασίζεται στον αλγόριθμο Expectation-Maximization έγινε πρόσφατα από τους (Lebre & Bourguignon, 2008). Το μοντέλο αυτό είναι ιδιαίτερα υποσχόμενο γιατί έχει μια σειρά από συγκριτικά πλεονεκτήματα (απλότητα, ευκολία στην ερμηνεία των παραμέτρων κ.ο.κ.), αλλά Παρ' όλα αυτά, δεν έχει χρησιμοποιηθεί ακόμα αρκετά σε εφαρμογές στη βιοπληροφορική.

Τυπικά παραδείγματα εφαρμογής των μαρκοβιανών αλυσίδων αφορούν στην εύρεση γονιδίων (gene finding), είτε σε επιβλεπόμενη (Borodovsky & McIninch, 1993; Borodovsky & Peresetsky, 1994) είτε σε μη-επιβλεπόμενη διαδικασία (Audic & Claverie, 1998). Επεκτάσεις του βασικού μοντέλου, όπως οι λεγόμενες interpolated Markov chains ή οι αλυσίδες μεταβλητού μήκους, έχουν επίσης χρησιμοποιηθεί για τον ίδιο σκοπό σε Βακτήρια (Salzberg, Delcher, Kasif, & White, 1998) και σε Ευκαρυωτικούς οργανισμούς (Ohler, Harbeck, Niemann, Noth, & Reese, 1999; Salzberg, Pertea, Delcher, Gardner, & Tettelin, 1999), για την εύρεση μοτίβων σε βιολογικές αλληλουχίες (Barash, Elidan, Friedman, & Kaplan, 2003), για τον εντοπισμό οριζόντιας γονιδιακής μεταφοράς (Dalevi, Dubhashi, & Hermansson, 2006), πρόγνωση της κυτταρικής θέσης των πρωτεϊνών (Yuan, 1999), για ταξινόμηση πρωτεϊνικών αλληλουχιών (Bejerano, et al., 2001), για την ανακατασκευή απλοτύπων στη γενετική (Eronen, Geerts, & Toivonen, 2004) και για τη λεγόμενη ανάλυση πολλών σημείων σε μελέτες γενετικής συσχέτισης (Browning, 2006).

8.2. Hidden Markov Models

8.2.1. Ορισμοί

Ένα Hidden Markov Model (HMM), αποτελείται από ένα σύνολο κρυφών καταστάσεων, ένα σύνολο παρατηρούμενων συμβόλων και δυο σύνολα πιθανοτήτων, τις πιθανότητες μετάβασης και τις πιθανότητες εκπομπής ή εμφάνισης συμβόλων (emissions). Θεωρώντας μια πρωτεϊνική αλληλουχία **x** μήκους L καταλοίπων:

$$\mathbf{x} = x_1, x_2, \ldots, x_{L-1}, x_L \tag{8.19}$$

όπου με x_i συμβολίζουμε τις παρατηρήσεις αποτελούμενες από ένα εκ των 20 αμινοξέων (ή γενικότερα, ενός διακριτού αλφάβητου Ω), στο HMM οι παρατηρήσεις πλέον αποδεσμεύονται από τις καταστάσεις. Συνηθίζεται να συμβολίζουμε την αλληλουχία των καταστάσεων έως μια συγκεκριμένη θέση i στην αλληλουχία, με π_i, και να την ονομάζουμε «μονοπάτι» (path). Έτσι, δυο καταστάσεις k, l συνδέονται μέσω των πιθανοτήτων μετάβασης a_{kl}, σχηματίζοντας μια αλυσίδα Markov 1ης τάξης. Ο τυπικός ορισμός αυτών των πιθανοτήτων μεταβάσεως, έχει ως εξής:

$$a_{kl} = P(\pi_i = l \mid \pi_{i-1} = k) \tag{8.20}$$

και συμβολίζει πολύ απλά την πιθανότητα, η κατάσταση k να δώσει μετάβαση (να προηγείται δηλαδή) προς την κατάσταση l. Αντίστοιχα με το απλό μοντέλο Markov, ορίζονται και εδώ ειδικές καταστάσεις ενάρξεως και τερματισμού της αλληλουχίας, οι οποίες για συντομία ονομάζονται B (Begin)

$$a_{Bk} = P(\pi_1 = k \mid B) \tag{8.21}$$

και E (End) αντίστοιχα

$$a_{kE} = P(E \mid \pi_i = k) \tag{8.22}$$

Η σύνδεση των παρατηρηθέντων συμβόλων με τις καταστάσεις, γίνεται μέσω των πιθανοτήτων εμφάνισης συμβόλων:

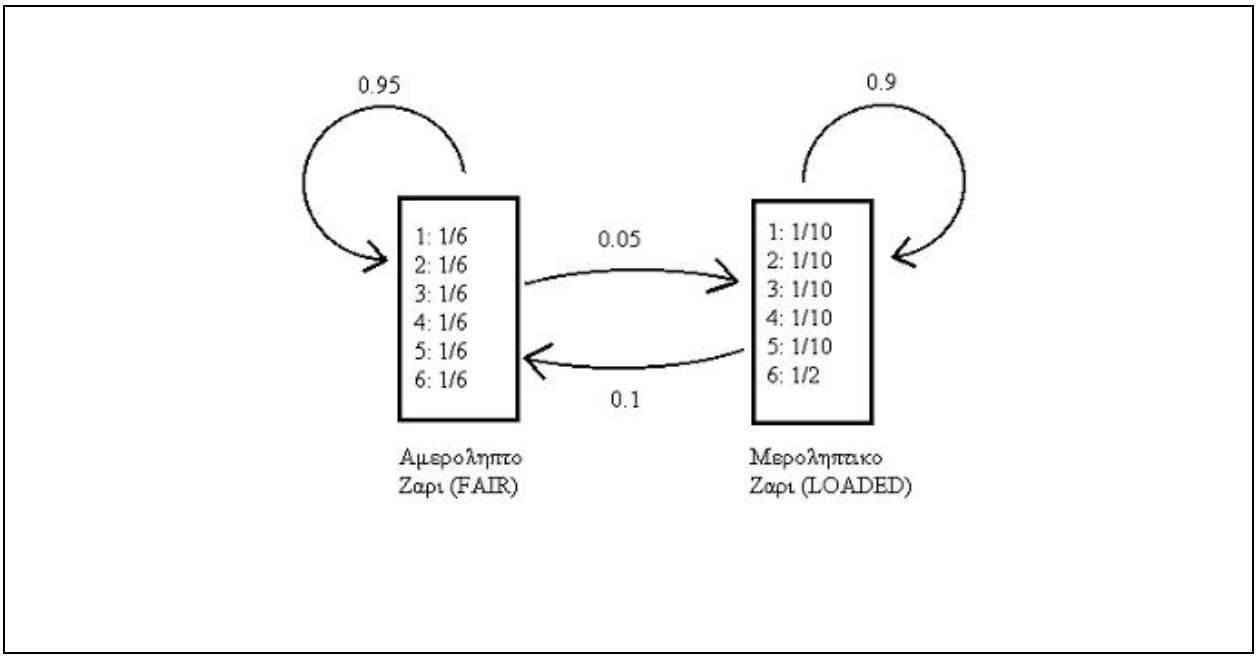
$$e_k(b) = P(x_i = b \mid \pi_i = k) \tag{8.23}$$

οι οποίες, δηλώνουν την πιθανότητα εμφάνισης στη θέση *i* της αλληλουχίας, ενός συγκεκριμένου συμβόλου *b*, δεδομένου ότι το σύστημα βρίσκεται στην κατάσταση *k*. Η από κοινού πιθανότητα μιας αλληλουχίας **x** και του μονοπατιού π, υπολογίζεται ως εξής:

$$P(\mathbf{x}, \pi) = P(x_L, x_{L-1}, ..., x_1, \pi) = a_{B\pi_1} \prod_{i=1}^{L} e_{\pi i}(x_i) a_{\pi_i \pi_{i+1}} \tag{8.24}$$

Ένα χαρακτηριστικό παράδειγμα HMM, το οποίο αναφέρεται στη βιβλιογραφία (Durbin, et al., 1998) είναι αυτό του «ανέντιμου καζίνο» (dishonest casino). Στο παράδειγμα αυτό, το καζίνο χρησιμοποιεί κατά περίσταση «κανονικά» αμερόληπτα ζάρια, αλλά έχει τη δυνατότητα να τα αλλάζει (π.χ. με πιθανότητα 0.05 για κάθε ζαριά) και να χρησιμοποιεί κάποια άλλα μεροληπτικά, δηλαδή ζάρια τα οποία ευνοούν κάποιο συγκεκριμένο αποτέλεσμα. Από την μεριά του ο παίκτης, το μόνο που μπορεί να δει είναι τα αποτελέσματα του ζαριού, αλλά δεν μπορεί να ξέρει ποιο ζάρι χρησιμοποιείται κάθε φορά.

Όπως βλέπουμε (Εικόνα 8.3) η πιθανότητα με την οποία αλλάζει το ζάρι από αμερόληπτο σε μεροληπτικό είναι 0.05 (επιλέχθηκε αυθαίρετα σε αυτό το παράδειγμα) και από μεροληπτικό πίσω σε αμερόληπτο, 0.1 (επίσης αυθαίρετη επιλογή). Το μοντέλο αυτό δίκαια ονομάζεται «κρυμμένο» (hidden) γιατί η πραγματική κατάσταση στην οποία βρίσκεται το ζάρι είναι κρυμμένη από τον παίκτη. Προφανώς η μετάβαση από τη μια κατάσταση του ζαριού (μεροληπτικό) στην άλλη (αμερόληπτο) και πάλι πίσω, είναι μια ανέλιξη Markov. Η σημαντική διαφορά του μοντέλου αυτού (HMM) από το απλό Μοντέλο Markov (MM) είναι το ότι σε αυτή την περίπτωση δεν υπάρχει μια προς μια αντιστοίχιση ανάμεσα στα σύμβολα και στις καταστάσεις του μοντέλου. Δηλαδή βλέποντας ένα σύμβολο (πχ. το ζάρι να έχει φέρει αποτέλεσμα 4), δεν μπορούμε να πούμε από ποια κατάσταση έχει διέλθει το μοντέλο για να δώσει το αποτέλεσμα αυτό.

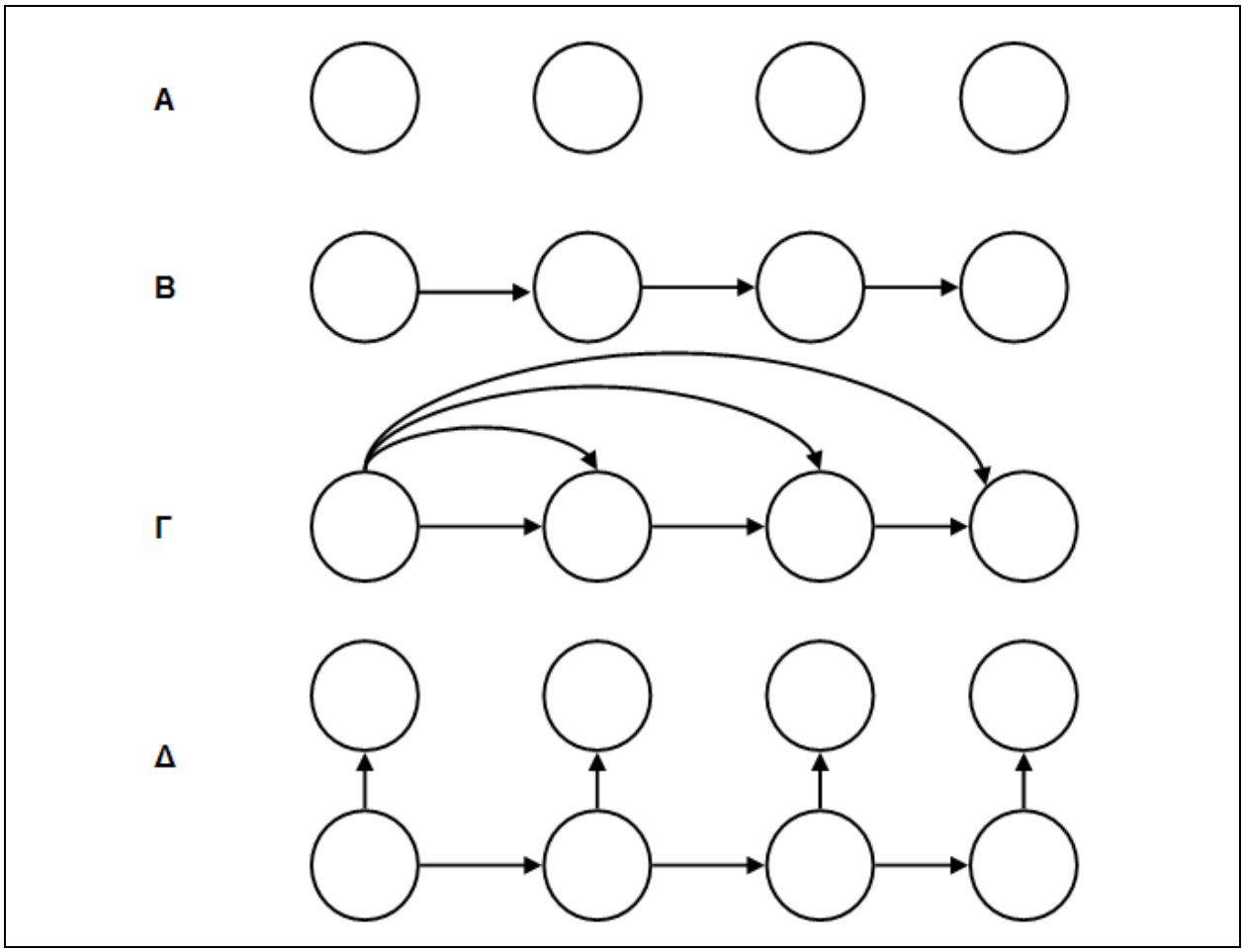


Εικόνα 8.3: *Το παράδειγμα του 'ανέντιμου καζίνο'. Τα δυο παραλληλόγραμμα συμβολίζουν τις δυο καταστάσεις του ζαριού (αμερόληπτο-μεροληπτικό), και τα βέλη τις επιτρεπτές μεταβάσεις. Μέσα σε κάθε κατάσταση, αναγράφονται οι πιθανότητες εμφανίσεως των συμβόλων.*

8.2.2. Τα 3 βασικά ερωτήματα σε ένα HMM

Στην ενότητα αυτή αναπτύσσονται τα 3 βασικά ερωτήματα, τα οποία μπορούν να τεθούν σε ένα HMM όπως διατυπώνονται από τον Rabiner (Rabiner, 1989).

- Δεδομένου ενός μοντέλου θ, πώς μπορούμε να υπολογίσουμε τη συνολική πιθανότητα μια αλληλουχία **x** να έχει εμφανιστεί από αυτό το μοντέλο; Δηλαδή, πώς μπορούμε να υπολογίσουμε την ποσότητα $P(\mathbf{x}|\theta)$;
- Δεδομένου ενός μοντέλου θ και μιας αλληλουχίας **x**, πώς μπορούμε να υπολογίσουμε το μονοπάτι, δηλαδή την αλληλουχία καταστάσεων, με την καλύτερη πιθανότητα; Με άλλα λόγια, πώς μπορούμε να υπολογίσουμε το μονοπάτι π, έτσι ώστε: $\pi^{\max} = \arg\max_{\pi} P(\mathbf{x}, \pi)$;
- Πώς μπορούμε να τροποποιήσουμε τις παραμέτρους του μοντέλου θ, υπό το φως νέων δεδομένων, ώστε να έχουμε καλύτερα μοντέλα; Το πρόβλημα αυτό, ανάγεται στην εκτίμηση παραμέτρων με τη μέθοδο της μέγιστης πιθανοφάνειας. Συγκεκριμένα, ζητάμε τον υπολογισμό των παραμέτρων θ, έτσι ώστε $\theta^{ML} = \arg\max_{\theta} P(\mathbf{x} \mid \theta)$.



Εικόνα 8.4: *Γραφική αναπαράσταση των πιθανοθεωρητικών μοντέλων που έχουμε συναντήσει έως τώρα. Α. Το βασικό μοντέλο της ανεξαρτησίας. Β. Το μαρκοβιανό μοντέλο 1ης τάξης. Γ. Ένα μαρκοβιανό μοντέλο 3ης τάξης. Δ. Το Hidden Markov Model. Στα μοντέλα Α-Γ, οι καταστάσεις αντιστοιχούν σε παρατηρήσιμα σύμβολα. Στο Δ οι καταστάσεις (στην κάτω γραμμή) ακολουθούν μια μαρκοβιανή αλυσίδα 1ης τάξης, κάθε κατάσταση της οποίας «παράγει» με διαφορετική πιθανότητα τα παρατηρήσιμα σύμβολα.*

8.2.3. Υπολογισμός Πιθανοφάνειας

Η σχέση (8.24), είναι όπως είδαμε, η από κοινού πιθανότητα μιας αλληλουχίας **x** και του μονοπατιού π, και δεν μας είναι ιδιαίτερα χρήσιμη γιατί δεν είναι δυνατόν να γνωρίζουμε ποια αλληλουχία καταστάσεων έδωσε γέννηση στην αλληλουχία των παρατηρήσεων. Για να υπολογίσουμε τη συνολική πιθανότητα μιας αλληλουχίας **x** δεδομένου του μοντέλου, θα πρέπει να υπολογίσουμε το άθροισμα τις για όλες τις πιθανές αλληλουχίες καταστάσεων, δηλαδή να αθροίσουμε τη συνεισφορά στη συνολική πιθανότητα, όλων των πιθανών μονοπατιών π.

$$P(\mathbf{x} \mid \theta) = \sum_{\pi} P(\mathbf{x}, \pi \mid \theta) = \sum_{\pi} a_{B\pi_1} \prod_{i=1}^{L} e_{\pi i}(x_i) a_{\pi_i \pi_{i+1}} \tag{8.25}$$

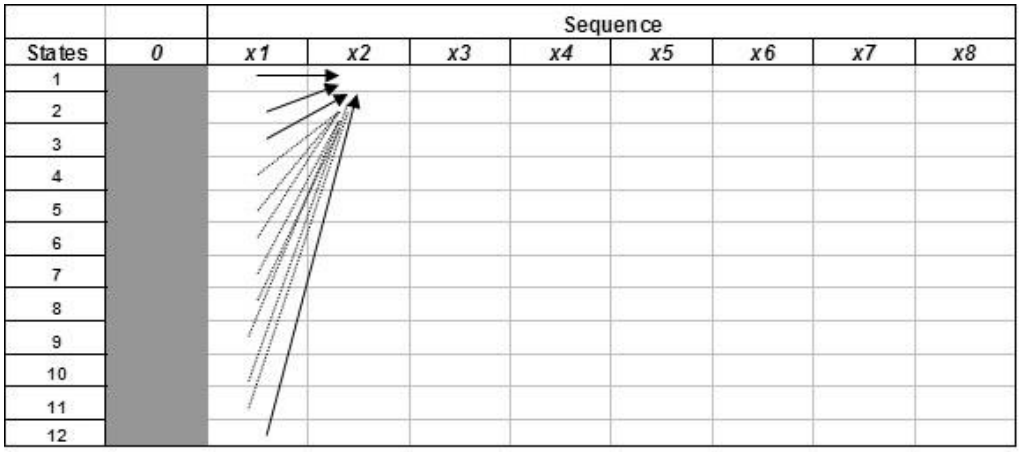
Η ποσότητα αυτή, πρακτικά δεν μπορεί να υπολογιστεί, γιατί μια απλή απαρίθμηση του πλήθους των πιθανών μονοπατιών αυξάνει εκθετικά καθώς αυξάνει το μήκος της αλληλουχίας. Έτσι, π.χ. αν έχουμε ένα μοντέλο με 50 καταστάσεις και μια αλληλουχία μήκους 300 καταλοίπων, τα πιθανά μονοπάτια είναι 50^{300}, αριθμός αστρονομικά μεγάλος. Η συνηθισμένη τακτική σε τέτοιου είδους υπολογιστικά προβλήματα, όπως είδαμε και στο κεφάλαιο 3, είναι ο δυναμικός προγραμματισμός (dynamic programming). Με τη μέθοδο αυτή, το μεγάλο πρόβλημα σπάει σε αρκετά μικρότερα, οι λύσεις των οποίων υπολογίζονται πολύ πιο εύκολα. Ο πιο γνωστός αλγόριθμος δυναμικού προγραμματισμού που έχει προταθεί για το παραπάνω πρόβλημα, είναι ο αλγόριθμος Forward (Εικόνα 8.4), ο οποίος σκιαγραφείται παρακάτω (Durbin, et al., 1998; Rabiner, 1989).

Αλγόριθμος Forward

$$\begin{aligned}
&\forall k \neq B,\ i = 0 : f_B(0) = 1,\ f_k(0) = 0 \\
&\forall 1 \leq i \leq L : f_l(i) = e_l(x_i) \sum_k f_k(i-1) a_{kl} \\
&P(\mathbf{x}|\theta) = \sum_k f_k(L) a_{kE}
\end{aligned} \tag{8.26}$$

Ο αλγόριθμος αυτός, κατασκευάζει έναν πίνακα με διαστάσεις *N(L+1)*, όπου *N* ο αριθμός των καταστάσεων και *L* το μήκος της αλληλουχίας, και θεωρεί μια ενδιάμεση μεταβλητή $f_k(i)$ για κάθε θέση *i* και κατάσταση *k* της αλληλουχίας. Η ποσότητα αυτή, αντιστοιχεί στην από κοινού πιθανότητα της αλληλουχίας έως το κατάλοιπο *i*, και του μονοπατιού που αντιστοιχεί στην κατάσταση *k*. Δηλαδή:

$$f_k(i) = P(x_1, x_2, ..., x_i, \pi_i = k) \tag{8.27}$$



Εικόνα 8.5: *Διαγραμματική απεικόνιση του πίνακα Forward, για ένα υποθετικό μοντέλο με 12 καταστάσεις (states) και μια αλληλουχία από 8 κατάλοιπα. Για τον υπολογισμό της τιμής ενός κελιού (π.χ. του $f_1(2)$), υπολογίζονται οι συνεισφορές όλων των προηγούμενων κελιών στη θέση 1 της αλληλουχίας (βέλη).*

Στο πρώτο βήμα, ο πίνακας των $f_k(i)$ αρχικοποιείται, στο δεύτερο συμπληρώνονται οι τιμές του διαδοχικά από την αρχή ως το τέλος της αλληλουχίας, και στο τελευταίο βήμα αθροίζονται για να προκύψει η τελική πιθανοφάνεια. Αν το μοντέλο δεν έχει κατάσταση λήξεως, τότε στο τελευταίο βήμα οι αντίστοιχες πιθανότητες απλώς απαλείφονται. Ο αλγόριθμος αυτός απαιτεί *NL* υπολογισμούς, γι' αυτό και λέμε ότι είναι της τάξης O(*NL*).

Εντελώς ανάλογος είναι ο αλγόριθμος Backward (Durbin, et al., 1998; Rabiner, 1989), ο οποίος διαφέρει μόνο ως προς την κατεύθυνση προς την οποία διατρέχει την αλληλουχία. Η ενδιάμεση μεταβλητή που χρησιμοποιείται, ονομάζεται πλέον $b_k(i)$, και ορίζεται για κάθε i ως η πιθανότητα της αλληλουχίας από τη θέση i+1 έως το τέλος, δεδομένου ότι στη θέση i συναντάμε την κατάσταση k. Δηλαδή:

$$b_k(i) = P(x_{i+1},...,x_L \mid \pi_i = k) \tag{8.28}$$

Άρα ο αλγόριθμος, διατυπώνεται ως εξής:

Αλγόριθμος Backward

$$\begin{aligned} &\forall k,\ i = L:\ b_k(L) = a_{kE} \\ &\forall\ 1 \le i < L:\ b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i+1) \\ &P(\mathbf{x}|\theta) = \sum_l a_{Bl} e_l(x_1) b_l(1) \end{aligned} \tag{8.29}$$

Όμοια, αν δεν υπάρχουν καταστάσεις λήξεως, στην αρχικοποίηση, οι αντίστοιχες πιθανότητες τίθενται ίσες με 1. Το τελικό αποτέλεσμα του αλγορίθμου, είναι ακριβώς όμοιο με αυτό του Forward.

8.2.4. Αποκωδικοποίηση

Στο δεύτερο ερώτημα, θέλουμε να βρούμε ποια είναι η πιο πιθανή αλληλουχία καταστάσεων από την οποία προέκυψε η αλληλουχία των παρατηρήσεων. Αυτό, αναφέρεται στην ουσία, στην αποκωδικοποίηση (decoding) ενός μοντέλου. Ένας παρόμοιος αλγόριθμος με αυτούς που είδαμε παραπάνω, είναι ο αλγόριθμος του Viterbi (Durbin, et al., 1998; Rabiner, 1989).

Αλγόριθμος Viterbi

$$\begin{aligned} &\forall k \ne B,\ i = 0:\ u_B(0) = 1,\ u_k(0) = 0 \\ &\forall\ 1 \le i \le L:\ u_l(i) = e_l(x_i) \max_k \{u_k(i-1) a_{kl}\} \\ &P(\mathbf{x}, \pi^{\max} \mid \theta) = \max_k \{u_k(L) a_{kE}\} \end{aligned} \tag{8.30}$$

Ο αλγόριθμος του Viterbi, είναι στην ουσία όμοιος με τον Forward, με τη μόνη διαφορά να εντοπίζεται στο ότι τα διαδοχικά αθροίσματα αντικαθίστανται από μεγιστοποιήσεις. Σε αυτή την περίπτωση με π^{max}, συμβολίζουμε το μονοπάτι με τη μεγαλύτερη πιθανότητα και η πιθανότητα αυτή συμβολίζεται με $P(\mathbf{x}, \pi^{max}|\theta)$. Προφανώς, ισχύει ότι $P(\mathbf{x}, \pi^{max}|\theta) \le P(\mathbf{x}|\theta)$. Ένα επιπλέον χαρακτηριστικό του αλγόριθμου αυτού, στο οποίο μοιάζει με τους αλγόριθμους στοίχισης αλληλουχιών, είναι το ότι απαιτεί την ύπαρξη ενός ξεχωριστού πίνακα στον οποίο θα κρατούνται δείκτες (pointers), για την καλύτερη (πιθανότερη) κατάσταση σε κάθε θέση της αλληλουχίας. Με αναδρομή (back-tracking), σε αυτόν τον πίνακα, ανακτά κανείς στο τέλος, το ίδιο το πιθανότερο μονοπάτι.

Εκ των υστέρων αποκωδικοποίηση

Πολλές φορές μπορεί να χρειαστούμε κάτι παραπάνω από τον απλό υπολογισμό του πιο πιθανού μονοπατιού. Μπορεί για παράδειγμα, να θέλουμε να υπολογίσουμε την πιο πιθανή κατάσταση για μια συγκεκριμένη παρατήρηση x_i, ή πιο γενικά μπορεί να θέλουμε να βρούμε την πιθανότητα η παρατήρηση x_i να προέρχεται από μια κατάσταση k, δεδομένης ολόκληρης της αλληλουχίας **x**. Αναζητούμε δηλαδή, την ποσότητα $P(\pi_i=k|\mathbf{x})$. Στην αρχή υπολογίζουμε την από κοινού πιθανότητα της αλληλουχίας **x** και του ενδεχομένου η i παρατήρηση να προέρχεται από την κατάσταση k. Άρα:

$$P(\mathbf{x}, \pi_i = k) = P(x_1, x_2, ..., x_i, \pi_i = k) P(x_{i+1}, ..., x_n \mid x_1, ..., x_i, \pi_i = k)$$
$$= P(x_1, x_2, ..., x_i, \pi_i = k) P(x_{i+1}, ..., x_n \mid \pi_i = k)$$

Ο πρώτος όρος του τελευταίου γινομένου, προκύπτει από τη σχέση (8.27), ενώ ο τελευταίος από τη σχέση (8.28). Άρα, προκύπτει ότι:

$$P(\mathbf{x}, \pi_i = k) = f_k(i) b_k(i)$$

Τέλος, σύμφωνα με το θεώρημα Bayes θα έχουμε:

$$P(\pi_i = k \mid \mathbf{x}) = \frac{f_k(i) b_k(i)}{P(\mathbf{x})} \tag{8.31}$$

Με τον τύπο αυτό, μπορούμε να υπολογίσουμε την πιθανότητα μια παρατήρηση να προέρχεται από μια συγκεκριμένη κατάσταση. Μπορούμε επίσης να ορίσουμε μια άλλη αλληλουχία καταστάσεων, για την οποία ισχύει:

$$\hat{\pi}_i = \arg\max_k P(\pi_i = k \mid \mathbf{x}) \tag{8.32}$$

Η σχέση αυτή μπορεί να είναι πιο χρήσιμη όταν ενδιαφερόμαστε περισσότερο για τον καθορισμό της κατάστασης μιας συγκεκριμένης παρατήρησης και όχι για ολόκληρο το μονοπάτι. Με τον τρόπο αυτό, ορίζουμε το μονοπάτι που μεγιστοποιεί την εκ των υστέρων πιθανότητα.

Μπορούμε επίσης, να ομαδοποιήσουμε κατά κάποιο τρόπο τις καταστάσεις των οποίων η ύπαρξη έχει την ίδια βιολογική σημασία. Με αυτόν το δεύτερο τρόπο, σε γενικές γραμμές, δεν μας ενδιαφέρει η ίδια η αλληλουχία καταστάσεων αλλά κάποια άλλη ιδιότητα που προκύπτει από αυτήν. Για παράδειγμα, στην περίπτωση που έχουμε ένα μοντέλο, στο οποίο οι καταστάσεις, ομαδοποιούνται σε δυο κατηγορίες (π.χ. διαμεμβρανικές-μη διαμεμβρανικές), αν έχουμε μια συνάρτηση $g(k)$ που να ορίζεται πάνω στις ίδιες τις καταστάσεις, με

$$g(k) = \begin{cases} 1, \textbf{αν } k \in C^{TM} \\ 0, \textbf{αν } k \in C^{NTM} \end{cases}$$

τότε

$$G(i \mid \mathrm{x}) = \sum_k P(\pi_i = k \mid x) g(k) \tag{8.33}$$

και αυτή είναι ακριβώς η εκ των υστερών πιθανότητα το αμινοξύ i να ανήκει σε μια διαμεμβρανική περιοχή σύμφωνα με το μοντέλο.

Η βασική αδυναμία των παραπάνω τεχνικών αποκωδικοποίησης, οι οποίες ονομάζονται τεχνικές «εκ των υστέρων αποκωδικοποίησης» (posterior decoding), είναι ότι είναι δυνατόν να δώσουν πρόγνωση ασύμβατη με το μοντέλο. Με άλλα λόγια, μπορεί να προβλέψουν ως την πιο πιθανή αλληλουχία καταστάσεων, μια αλληλουχία η οποία δεν θα μπορούσε να προκύψει μέσω του μοντέλου (μη επιτρεπτές μεταβάσεις). Η σοβαρή αυτή αδυναμία, η οποία μπορεί πρακτικά να ακυρώσει τα πλεονεκτήματα του HMM, μπορεί να αντιμετωπιστεί μέσω μιας διαδικασίας «φιλτραρίσματος» και επεξεργασίας των εκ των υστέρων πιθανοτήτων, με έναν αλγόριθμο δυναμικού προγραμματισμού.

Παράδειγμα 8.2.4.1

Έστω ότι έχουμε ένα υποθετικό μοντέλο το οποίο να περιγράφει τις αλληλουχίες DNA, και το οποίο περιγράφεται παρακάτω (είναι ανάλογο με το παράδειγμα με το ζάρι): υπάρχουν 2 διακριτές περιοχές στις αλληλουχίες οι οποίες υποθέτουμε ότι έχουν κάποια λειτουργική σημασία, και οι οποίες διαφέρουν στις πιθανότητες εμφάνισης των 4 βάσεων (emission probabilities). Έτσι στην περιοχή «1» ισχύουν :

$$e_1(A) = P(x_i = A \mid \pi_i = 1) = 0.7$$
$$e_1(T) = P(x_i = T \mid \pi_i = 1) = 0.1$$
$$e_1(G) = P(x_i = G \mid \pi_i = 1) = 0.1$$
$$e_1(C) = P(x_i = C \mid \pi_i = 1) = 0.1$$

ενώ στην περιοχή «0» ισχύουν αντίστοιχα:

$$e_0(A) = P(x_i = A \mid \pi_i = 0) = 0.25$$
$$e_0(T) = P(x_i = T \mid \pi_i = 0) = 0.25$$

$$e_0(G) = P(x_i = G \mid \pi_i = 0) = 0.25$$

$$e_0(C) = P(x_i = C \mid \pi_i = 0) = 0.25$$

Οι πιθανότητες μεταβάσεως (από τη μία περιοχή στην άλλη) είναι:

$$a_{11} = P(\pi_i = 1 \mid \pi_{i-1} = 1) = 0.9$$

$$a_{10} = P(\pi_i = 0 \mid \pi_{i-1} = 1) = 0.1$$

$$a_{00} = P(\pi_i = 0 \mid \pi_{i-1} = 0) = 0.9$$

$$a_{01} = P(\pi_i = 1 \mid \pi_{i-1} = 0) = 0.1$$

Παρακάτω φαίνεται μια αλληλουχία DNA μήκους 200 βάσεων η οποία προήλθε από το παραπάνω μοντέλο. Στην πρώτη σειρά φαίνεται η αλληλουχία του DNA, ενώ από κάτω φαίνεται η αλληλουχία των καταστάσεων (1/0). Στην Εικόνα 8.7 φαίνεται η διαδικασία αποκωδικοποίησης και οι εκ των υστέρων πιθανότητες για αυτήν την αλληλουχία.

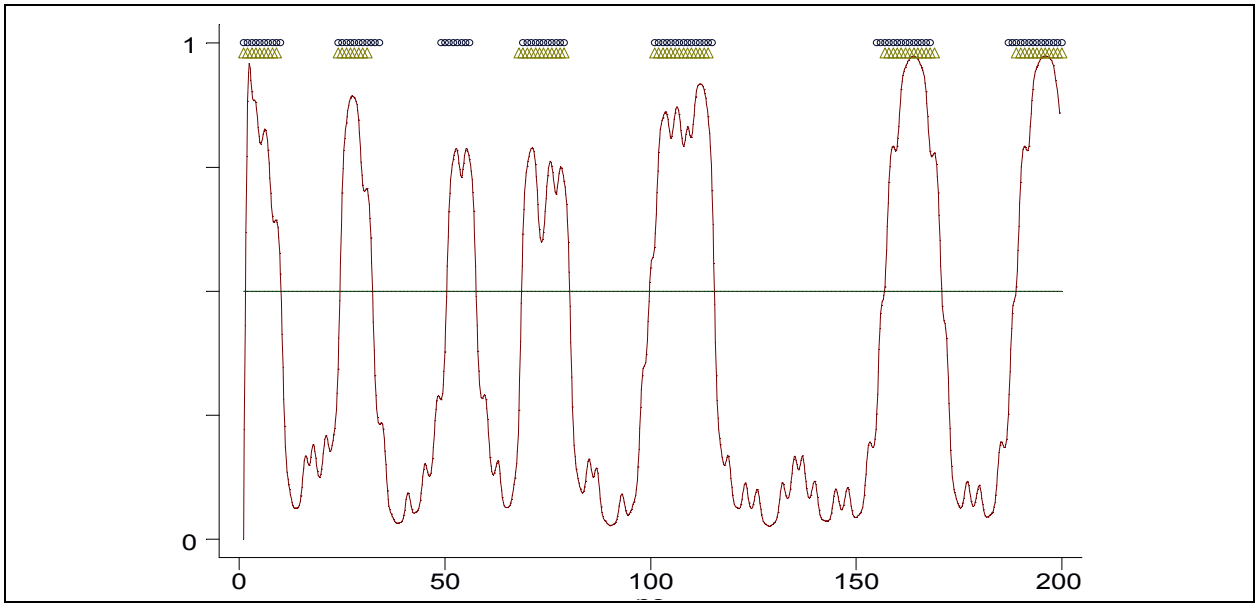
```
AAACAAGAATGCGCACACTACGCAAAAACAATTAGTCGCACTCACGATGAAACAAATTACCACGGTGAA
1111111111000000000000011111111111000000000000001111111100000000000001

AACGAATAAACCTCAGAGGCCCAGCGTATATAAACAAGATAAAAACCTAGTCAGCACTCTGACCAGACG
1111111111000000000000000000000011111111111111100000000000000000000000

AGCTCACGACTTGAGGATAAGAAAAAAACAACAGCTCACGACTTGAGGATAAGAAAAAAACA
0000000000000000111111111111111000000000000000000011111111111111
```

Εικόνα 8.6: *Μια τυχαία αλληλουχία DNA από το παραπάνω μοντέλο. Κάτω από την αλληλουχία δίνεται και το αντίστοιχο μονοπάτι (0/1)*

Στην παρακάτω εικόνα (Εικόνα 8.7) φαίνεται η αποκωδικοποίηση με τους δυο δυνατούς τρόπους (αποκωδικοποίηση Viterbi και εκ των υστέρων αποκωδικοποίηση) που αναφέρθηκαν παραπάνω.



Εικόνα 8.7: *Οι εκ των υστέρων πιθανότητες και η αποκωδικοποίηση Viterbi για την αλληλουχία που δόθηκε στην Εικόνα 8.6. Όπως φαίνεται και οι δυο μέθοδοι δουλεύουν καλά, προβλέποντας την κατάσταση στην οποία βρίσκονται τα νουκλεοτίδια του DNA. Παρατηρούμε ότι η «εκ των υστέρων αποκωδικοποίηση», είναι λίγο πιο αποτελεσματική καθώς έχει προβλέψει σωστά και τις 7 περιοχές τύπου «1», ενώ η «αποκωδικοποίηση Viterbi» έχει αποτύχει να εντοπίσει μία από αυτές.*

Αν σε μια παραλλαγή του υποθετικού αυτού μοντέλου, αλλάξουν οι πιθανότητες μεταβάσεως ως εξής:

$a_{11} = P(\pi_i = 1 \mid \pi_{i-1} = 1) = 0.98$

$a_{10} = P(\pi_i = 0 \mid \pi_{i-1} = 1) = 0.02$

$a_{00} = P(\pi_i = 0 \mid \pi_{i-1} = 0) = 0.97$

$a_{01} = P(\pi_i = 1 \mid \pi_{i-1} = 0) = 0.03$

Τότε, όταν στην παρακάτω αλληλουχία που προέρχεται από το μοντέλο αυτό, εφαρμόσουμε τις δύο παραπάνω μεθόδους αποκωδικοποίησης :

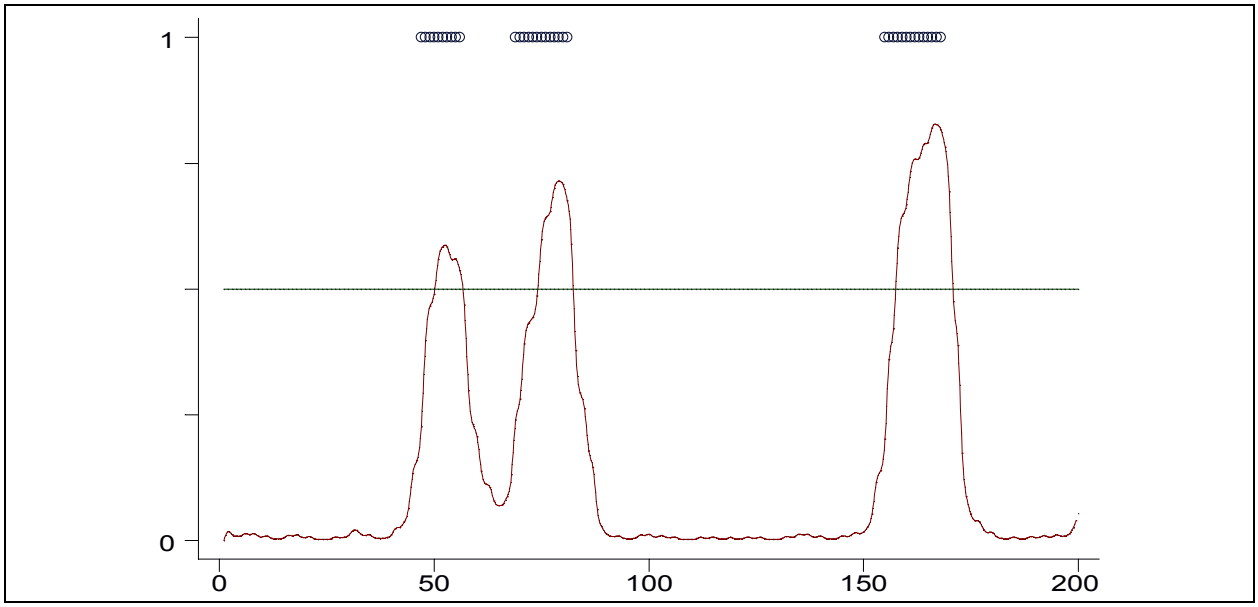
```
GCGCACACTAGCGCACACTACGCCTACGCAATTAGTCGCACTCACGAAGAAACAAATTACCACGGTGAG
000000000000000000000000000000000000000000000001111111110000000000001

AACGAATAAAAATCAGAGGCCCAGCGTATATCAGCACTCTGACCACCTAGTCAGCACTCTGACCAGACG
111111111111000000000000000000000000000000000000000000000000000000000

AGCTCACGACTTGAGGATAAGAATAGAAAAACAGCTCACGACTTGAGGCACGACTAGCTCAG
000000000000000011111111111111000000000000000000000000000000000
```

Εικόνα 8.8: *Μια τυχαία αλληλουχία DNA από το μοντέλο, όταν οι πιθανότητες αλλάξουν. Κάτω από την αλληλουχία δίνεται και το αντίστοιχο μονοπάτι (0/1)*

Θα έχουμε το επόμενο διάγραμμα:



Εικόνα 8.9: *Οι εκ των υστέρων πιθανότητες και η αποκωδικοποίηση Viterbi για την αλληλουχία από το αλλαγμένο μοντέλο. Όπως είναι φανερό σ' αυτή την περίπτωση ο αλγόριθμος του Viterbi, δεν προβλέπει καμία περιοχή να είναι τύπου «1», καθώς το πιο πιθανό μονοπάτι καταστάσεων δεν επισκέπτεται καμία φορά την κατάσταση «1». Αντίθετα η «εκ των υστέρων αποκωδικοποίηση» δίνει μια αρκετά πιο καλή προσέγγιση της αληθινής κατάστασης. Η «εκ των υστέρων αποκωδικοποίηση» είναι σε γενικές γραμμές καλύτερη μέθοδος, ιδιαίτερα σε περιπτώσεις όπως αυτές (που είδαμε παραπάνω) στις οποίες έχουμε πολύ μικρές (ή μεγάλες) πιθανότητες μεταβάσεως. Πρέπει να τονίσουμε εδώ, ότι στο συγκεκριμένο παράδειγμα, δεν έγινε ανάλυση για να υπολογιστεί το κατώφλι (cut-off value) της πιθανότητας πάνω από την οποία θα θεωρούμε ότι η μέθοδος προβλέπει περιοχές τύπου «1». Το κατώφλι που δεχθήκαμε εδώ (αυθαίρετα) είναι το 0.5 (πράσινη γραμμή), και αν ακολουθήσει ανάλυση ευαισθησίας, θα επιτύχουμε καλύτερη πρόβλεψη.*

Ακόμα πιο δύσκολες καταστάσεις αναμένεται να συναντήσουμε σε πραγματικά προβλήματα, όταν το μοντέλο περιέχει αρκετές καταστάσεις που συνδέονται μεταξύ τους με μια σειρά που έχει κάποιο βιολογικό

νόημα. Για παράδειγμα, στην εύρεση γονιδίων, μπορεί η κατάσταση 1 να αντιστοιχεί σε εξώνιο, η κατάσταση 2 στο σημείο αποκοπής, η κατάσταση 3 σε εσώνιο, ενώ στις διαμεμβρανικές πρωτεΐνες οι καταστάσεις μπορεί να συμβολίζουν αντίστοιχα τις εξωκυττάριες, τις διαμεμβρανικές και τις κυτοπλασμικές περιοχές. Σε όλες αυτές τις περιπτώσεις, η εκ των υστέρων αποκωδικοποίηση είναι δυνατόν, υπό συνθήκες, να δώσει ένα πιθανό μονοπάτι του τύπου 1-3-2, κάτι που όμως είναι αδύνατο από βιολογική σκοπιά. Το όλο επιχείρημα πίσω από τη χρήση κρυπτομαρκοβιανών μοντέλων σε αυτά τα προβλήματα, είναι ότι με τα μοντέλα αυτά μπορούμε να μοντελοποιήσουμε καλύτερα το υπό μελέτη βιολογικό σύστημα. Κατά συνέπεια, πρέπει να αναζητηθούν πιο αποδοτικοί αλγόριθμοι αποκωδικοποίησης οι οποίοι (όπως και ο Viterbi) να διαφυλάσσουν τη γραμματική του μοντέλου.

Εκ των υστέρων αποκωδικοποίηση με χρήση δυναμικού προγραμματισμού

Με τη μέθοδο αυτή, οι εκ των υστέρων πιθανότητες φιλτράρονται μέσα από έναν αλγόριθμο δυναμικού προγραμματισμού, ο οποίος περιορίζει τις πιθανές λύσεις μέσα σε κάποια προαποφασισμένα όρια (π.χ. τα ελάχιστα και μέγιστα μήκη των διαμεμβρανικών τμημάτων) και βρίσκει την ολική καλύτερη τοπολογία για την πρωτεΐνη. Η μέθοδος αυτή προτάθηκε αρχικά από τον Jones και τους συνεργάτες του (Jones, Taylor, & Thornton, 1994).

Το συνολικό πρόβλημα του εντοπισμού της βέλτιστης θέσεως και του μήκους των n διαμεμβρανικών τμημάτων σε μια αλληλουχία m καταλοίπων, υποδιαιρείται σε n μικρότερα προβλήματα αντιμετωπίζοντας κάθε διαμεμβρανική περιοχή ξεχωριστά. Έτσι, με s^{il} συμβολίζουμε το συνολικό σκορ (άθροισμα των εκ των υστέρων πιθανοτήτων που αντιστοιχούν σε διαμεμβρανική περιοχή), για το διαμεμβρανικό τμήμα με μήκος l στη θέση i μιας αλληλουχίας. Τότε, το συνολικό σκορ $S_j^i\left(i:1,2,...,n;\, j:1,2,...,m\right)$, θα υπολογίζεται από την αναδρομική σχέση:

$$S_j^i = \max_{l=l_{\min}\to l_{\max}} \left\{ s_j^{il} + \max_{k=1+l+A\to n} \left\{ S_{j-1}^k \right\} \right\} \tag{8.34}$$

όπου j είναι ο συνολικός αριθμός διαμεμβρανικών τμημάτων, l_{min} και l_{max} τα ελάχιστα και μέγιστα επιτρεπόμενα μήκη των διαμεμβρανικών τμημάτων, και A το ελάχιστο επιτρεπόμενο μήκος στροφής. Μια πιο γενική μορφή αυτού του αλγόριθμου προτάθηκε αργότερα (Fariselli et al., 2003), σαν μια γενικότερη λύση του προβλήματος εντοπισμού υπο-περιοχών με συγκεκριμένα χαρακτηριστικά.

Posterior-Viterbi

Πολύ πρόσφατα, ο Fariselli και οι συνεργάτες του, πρότειναν έναν αλγόριθμο αποκωδικοποίησης ο οποίος συνδυάζει χαρακτηριστικά του αλγορίθμου Viterbi και της εκ των υστέρων αποκωδικοποίησης (Fariselli, Martelli, & Casadio, 2005). Ο αλγόριθμος, βρίσκει το μονοπάτι π^{PV}:

$$\pi^{PV} = \arg\max_{\pi\in\Pi_p} \prod_{i=1}^{L} P\left(\pi_i \mid \mathbf{x}\right)$$

όπου Π_p είναι το σύνολο των επιτρεπτών από το μοντέλο μονοπατιών, και $P(\pi_i=k|\mathbf{x})$ η εκ των υστέρων πιθανότητα για μια κατάσταση, όπως ορίστηκε στη σχέση (8.31). Για να οριστούν τα επιτρεπτά μονοπάτια, χρειαζόμαστε μια δίτιμη συνάρτηση η οποία να παίρνει τιμή 1 για μια επιτρεπτή μετάβαση και 0 για μια μη-επιτρεπτή. Έτσι:

$$\delta\left(k,l\right) = \begin{cases} 1, & \text{if } a_{kl} > 0 \\ 0, & \text{otherwise} \end{cases}$$

Τελικά, το βέλτιστο επιτρεπτό εκ των υστέρων μονοπάτι π^{PV}, δίνεται από τη σχέση:

$$\pi^{PV} = \arg\max_{\pi} \prod_{i=1}^{L} \delta\left(\pi_i, \pi_{i+1}\right) P\left(\pi_i \mid \mathbf{x}\right)$$

Ο συνολικός αλγόριθμος, ο οποίος παρουσιάζεται παρακάτω, είναι στην ουσία μια παραλλαγή του αλγορίθμου Viterbi, στην οποία οι πιθανότητες γεννήσεως αντικαθίστανται από τις εκ των υστέρων πιθανότητες και οι πιθανότητες μετάβασης από την δίτιμη συνάρτηση που είδαμε παραπάνω.

Αλγόριθμος Posterior-Viterbi

$$\forall k \neq B,\ i = 0:\ u_B(0) = 1,\ u_k(0) = 0$$

$$\forall\ 1 \le i \le L:\ u_l(i) = P\left(\pi_i = l \mid \mathrm{x}\right) \max_k \left\{u_k(i-1)\delta\left(k,l\right)\right\}$$

$$P(\mathbf{x},\pi^{PV} \mid \theta) = \max_k \left\{u_k(L)\delta\left(k,E\right)\right\}$$

Αποκωδικοποίηση Forward

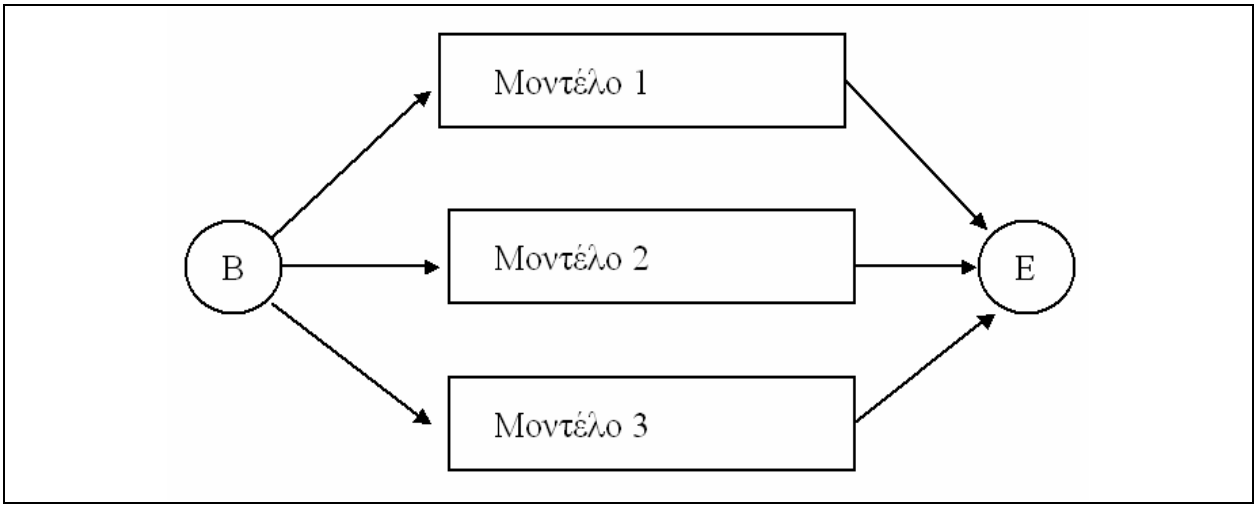
Ένα άλλο είδος αποκωδικοποίησης, είναι επίσης δυνατόν να πραγματοποιηθεί με τη χρήση του αλγόριθμου Forward. Η μέθοδος αυτή, η οποία ονομάζεται πολλές φορές 'Forward Decoding method', είναι χρήσιμη όταν για παράδειγμα έχουμε ένα μοντέλο το οποίο περιέχει πολλαπλούς κλάδους (Εικόνα 8.10). Σε αυτή την περίπτωση, ο κλάδος που δίνει την μεγαλύτερη πιθανότητα, θα είναι ο προτιμώμενος αλλά δεν θα μπορούμε με αυτό τον τρόπο να βρούμε την ακριβή αλληλουχία των καταστάσεων (μπορεί και να μην μας ενδιαφέρει).

Η χρήση της ολικής πιθανότητας από τη σχέση (8.26), δεν αρκεί γιατί μεγαλύτερες αλληλουχίες, κατά κανόνα θα δίνουν και μεγαλύτερη πιθανότητα. Έτσι χρησιμοποιήσαμε μια κανονικοποιημένη σχέση ανάλογη με αυτή της σχέσης (8.10). Συγκεκριμένα, θα έχουμε μια τιμή σκορ:

$$S(\mathbf{x}|\theta) = -\frac{\log P\left(\mathbf{x}|\theta\right)}{L} \tag{8.35}$$

όπου L, θα είναι το μήκος της πρωτεΐνης. Γενικά, μεγάλες τιμές αυτής της συνάρτησης θα είναι ένδειξη ότι η πρωτεΐνη δεν παράγεται από το μοντέλο, ενώ μικρές (που αντιστοιχούν σε μεγάλη πιθανότητα), ένδειξη ότι η πρωτεΐνη ταιριάζει με το μοντέλο. Εναλλακτικές σχέσεις, έχουν προταθεί (Eddy, 1998), οι οποίες χρησιμοποιούν έναν λόγο πηλίκου πιθανοφανειών, συγκρίνοντας για παράδειγμα την πιθανότητα της αλληλουχίας δεδομένου του μοντέλου, με την πιθανότητα της αλληλουχίας δεδομένου ενός 'τυχαίου' μοντέλου θ_0, ενός μοντέλου δηλαδή, το οποίο προϋποθέτει 'τυχαίες' κατανομές των αμινοξέων (μηδενικό – null μοντέλο).

$$S(\mathbf{x}|\theta) = -\frac{\log P\left(\mathbf{x}|\theta\right)}{\log P\left(\mathbf{x}|\theta_0\right)} \tag{8.36}$$



Εικόνα 8.10: *Υποθετική περίπτωση ενός μοντέλου με τρεις κλάδους. Με εφαρμογή του αλγορίθμου Forward, μπορούμε να βρούμε τον κλάδο με τη μεγαλύτερη πιθανότητα.*

Οι πιθανότητες για το τυχαίο αυτό μοντέλο, προκύπτουν από κάποια βάση δεδομένων, και για κάθε κατάλοιπο στην αλληλουχία αντιστοιχούν στα ποσοστά εμφάνισης του καταλοίπου στη βάση. Αθροίζοντας τις πιθανότητες για κάθε κατάλοιπο της αλληλουχίας, έχουμε τελικά την πιθανότητα της αλληλουχίας δεδομένου του μηδενικού μοντέλου.

8.2.5. Εκτίμηση Παραμέτρων στα HMM

Μέγιστη Πιθανοφάνεια

Η εκτίμηση των παραμέτρων ενός στατιστικού-πιθανοθεωρητικού μοντέλου, συνηθίζεται να πραγματοποιείται με τη διαδικασία της Μέγιστης Πιθανοφάνειας (Maximum Likelihood). Οι Εκτιμητές Μέγιστης Πιθανοφάνειας (ΕΜΠ), ορίζονται ως οι τιμές των παραμέτρων θ^{ML} του μοντέλου, οι οποίες μεγιστοποιούν τη συνάρτηση πιθανοφάνειας. Η τελευταία, είναι απλά η από κοινού συνάρτηση κατανομής όλων των παρατηρήσεων, δεδομένων των παραμέτρων του μοντέλου αν θεωρήσουμε τις παραμέτρους σαν τυχαίες μεταβλητές. Άρα:

$$\theta^{ML} = \arg\max_{\theta} P(\mathbf{x} \mid \theta) \tag{8.37}$$

Για λόγους υπολογιστικής απλότητας, συνήθως δουλεύουμε με το λογάριθμο της πιθανοφάνειας $l(\mathbf{x}|\theta)$, ο οποίος μεγιστοποιείται στα ίδια σημεία με αυτή.

$$l(\mathbf{x} \mid \theta) = \log P(\mathbf{x} \mid \theta)$$

Αν εργαζόμαστε με τον λογάριθμο της πιθανοφάνειας ο σκοπός είναι να μεγιστοποιήσουμε την τιμή του, ενώ αν εργαζόμαστε με το αντίθετό του (αρνητική λογαριθμική πιθανοφάνεια) ο σκοπός είναι να ελαχιστοποιήσουμε την αντίστοιχη τιμή. Πρέπει να τονιστεί εδώ, ότι οι ατομικές παρατηρήσεις είναι τα αμινοξικά κατάλοιπα. Κατά συνέπεια, αν διαθέτουμε για εκπαίδευση, περισσότερες αλληλουχίες, τις θεωρούμε ανεξάρτητες και η συνολική πιθανοφάνεια είναι το γινόμενο των πιθανοφανειών τους. Άρα, ο λογάριθμός της, είναι το άθροισμα των λογαρίθμων των πιθανοφανειών κάθε αλληλουχίας. Σε όλα τα παρακάτω, θα θεωρούμε ως δεδομένα του συνόλου εκπαίδευσης μια αλληλουχία **x**, με τη γενίκευση σε περίπτωση πολλαπλών ακολουθιών, να είναι τετριμμένη περίπτωση.

Στην ιδανική (όσο και ανέφικτη) περίπτωση, κατά την οποία γνωρίζουμε τα ακριβή μονοπάτια για τις αλληλουχίες εκπαίδευσης, ο υπολογισμός των ΕΜΠ είναι αρκετά απλός. Συγκεκριμένα, δεν έχουμε παρά να καταμετρήσουμε πόσες φορές παρατηρήθηκε μια συγκεκριμένη μετάβαση από κάθε κατάσταση, και πόσες φορές ένα αμινοξύ εμφανίστηκε σε κάθε κατάσταση. Άρα οι ΕΜΠ, για τις πιθανότητες μετάβασης θα είναι:

$$\hat{a}_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \tag{8.38}$$

και για τις πιθανότητες εμφάνισης συμβόλων,

$$\hat{e}_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b)} \tag{8.39}$$

όπου τα αθροίσματα στους παρονομαστές, εκτείνονται σε όλο το εύρος των παραμέτρων. Πρόβλημα με αυτή την προσέγγιση, μπορεί να υπάρξει, αν κάποια παράμετρος δεν παρατηρηθεί ούτε μια φορά στο σύνολο εκπαίδευσης, με αποτέλεσμα να εμφανιστούν μηδενικές πιθανότητες. Στην περίπτωση αυτή, μπορούμε να προσθέσουμε κάποιες ψευδοτιμές (pseudo-counts), ως εξής:

$$\hat{a}_{kl} = \frac{A_{kl} + r_{kl}}{\sum_{l'} A_{kl'} + \sum_{l'} r_{kl'}} \tag{8.40}$$

$$\hat{e}_k(b) = \frac{E_k(b) + r_k(b)}{\sum_{b'} E_k(b') + \sum_{b'} r_k(b')} \tag{8.41}$$

Αν κάποιος υιοθετήσει μια Μπεϋζιανή προσέγγιση στην ανάλυση των δεδομένων, οι ποσότητες αυτές r_{kl}, $r_k(b)$ αντιστοιχούν στις παραμέτρους μιας κατανομής Dirichlet. Η κατανομή αυτή χρησιμοποιείται ως η εκ των προτέρων (prior) κατανομή, λόγω του ότι είναι συζυγής με την πολυωνυμική κατανομή που ακολουθούν οι πιθανότητες μεταβάσεως και γεννήσεως.

Ο αλγόριθμος Baum-Welch

Στη γενικότερη και πιο συνηθισμένη περίπτωση, κατά την οποία δεν γνωρίζουμε την αλληλουχία των καταστάσεων, το πρόβλημα είναι πιο σύνθετο γιατί πρέπει να εκτιμηθούν οι παράμετροι ταυτόχρονα με τα μονοπάτια. Η λύση, προτάθηκε κατά τη δεκαετία του 1970 από τον Baum και την ερευνητική του ομάδα, και

έγινε γνωστή ως ο αλγόριθμος Baum-Welch (Baum, 1972). Στην πράξη, έχει αποδειχθεί, ότι ο αλγόριθμος αυτός, είναι μια ειδικότερη περίπτωση του αλγορίθμου Expectation-Maximisation (EM) (Dempster, Laird, & Rubin, 1977), ο οποίος προτάθηκε σαν μια γενική προσέγγιση για την εκτίμηση παραμέτρων από δεδομένα με ελλείπουσες τιμές (missing values). Είναι ενδιαφέρον ότι, ο Baum και οι συνεργάτες του, πρότειναν τον αλγόριθμο, χωρίς να γνωρίζουν την γενικευμένη προσέγγιση η οποία προτάθηκε αργότερα (Dempster, et al., 1977). Στα παρακάτω, για λόγους στατιστικής συνέπειας θα παραθέσουμε την παρουσίαση του αλγορίθμου υπό το πρίσμα του αλγορίθμου EM.

Γενικά ο αλγόριθμος EM χρησιμοποιείται για εκτιμήσεις μέγιστης πιθανοφάνειας όταν υπάρχουν ελλείπουσες τιμές (missing values). Εδώ οι ελλείπουσες τιμές είναι οι άγνωστες καταστάσεις π. Σε ολόκληρη την παράγραφο αυτή θ, θ^t,θ^{t+1} κλπ, εννοούμε το σετ των παραμέτρων του μοντέλου σε κάθε επανάληψη (iteration) t και $\mathbf{x}$ γενικά τα δεδομένα μας, δηλαδή τις αλληλουχίες. Ο λογάριθμος της πιθανοφάνειας θα είναι:

$$l(\mathbf{x},\theta) = \log P(\mathbf{x},\theta) = \sum_{\pi} \log P(\mathbf{x},\pi \mid \theta)$$

Επειδή από το θεώρημα του Bayes είναι γνωστό ότι:

$$P(\pi \mid \mathbf{x},\theta) = \frac{P(\mathbf{x},\pi \mid \theta)}{P(\mathbf{x} \mid \theta)}$$

θα έχουμε:

$$\log P(\mathbf{x} \mid \theta) = \log P(\mathbf{x},\pi \mid \theta) - \log P(\pi \mid \mathbf{x},\theta)$$

Τότε αν πολλαπλασιάσουμε με $P(\pi \mid \mathbf{x},\theta^{\mathbf{t}})$, και αθροίσουμε για όλα τα πιθανά μονοπάτια π, θα έχουμε:

$$\log P(\mathbf{x} \mid \theta) = \sum_{\pi} P(\pi \mid \mathbf{x},\theta^{\mathbf{t}}) \log P(\mathbf{x},\pi \mid \theta) - \sum_{\pi} P(\pi \mid \mathbf{x},\theta^{t}) \log P(\pi \mid \mathbf{x},\theta)$$

Τον πρώτο όρο του παραπάνω αθροίσματος τον ονομάζουμε:

$$Q(\theta \mid \theta^t) = \sum_{\pi} P(\pi|\mathrm{x},\theta^{\mathrm{t}}) \log P(\mathrm{x},\pi \mid \theta) \tag{8.42}$$

Για να μεγιστοποιηθεί η πιθανοφάνεια, θέλουμε:

$$\log P(\mathbf{x} \mid \theta) \geq \log P(\mathbf{x} \mid \theta^t)$$

για κάθε σετ παραμέτρων θ, άρα:

$$\log P(\mathbf{x} \mid \theta) - \log P(\mathbf{x} \mid \theta^t) = Q(\theta \mid \theta^t) - Q(\theta^t \mid \theta^t) + \sum_{\pi} P(\pi \mid \mathbf{x},\theta^t) \log \frac{P(\pi \mid \mathbf{x},\theta^t)}{P(\pi \mid \mathbf{x},\theta)}$$

και επειδή ο τελευταίος όρος είναι η σχετική εντροπία και είναι πάντα θετικός εκτός αν $\theta=\theta^t$ θα έχουμε:

$$\log P(\mathbf{x} \mid \theta) - \log P(\mathbf{x} \mid \theta^t) \geq Q(\theta \mid \theta^t) - Q(\theta^t \mid \theta^t)$$

Τότε αν διαλέξουμε το σύνολο των παραμέτρων που μεγιστοποιεί τη συνάρτηση Q, δηλαδή:

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta \mid \theta^t)$$

η πιθανοφάνεια του νέου μοντέλου θα είναι πάντα μεγαλύτερη. Ας δούμε αναλυτικά τώρα τον αλγόριθμο:
Η σχέση (8.24) γίνεται:

$$P(\mathbf{x},\pi \mid \theta) = \prod_{k=1} \prod_{b} [e_k(b)]^{E_k(b,\pi)} \prod_{k=0} \prod_{l=1} a_{kl}^{A_{kl}(\pi)} \tag{8.43}$$

όπου, $E_k(b,\pi)$, και, $A_{kl}(\pi)$ είναι οι συνολικές εμφανίσεις του συμβόλου b, και των μεταβάσεων στην κατάσταση l αντίστοιχα, από την κατάσταση k, σε ένα μονοπάτι π. Αντικαθιστώντας τώρα στην (8.42) έχουμε:

$$Q(\theta \mid \theta^t) = \sum_{\pi} P(\pi|\mathrm{x},\theta^{\mathrm{t}}) \left[\sum_{k=1} \sum_{b} E_k(b,\pi) \log e_k(b) + \sum_{k=0} \sum_{l=1} A_{kl}(\pi) \log \alpha_{\mathrm{kl}} \right] \tag{8.44}$$

Θα δείξουμε παρακάτω, ότι οι αναμενόμενες τιμές $E_k(b,\pi)$ και $A_{kl}(\pi)$ των παραμέτρων, αθροιζόμενες για όλα τα μονοπάτια, μπορούν να εκφραστούν σαν συνάρτηση των μεταβλητών $f_k(i)$, $b_k(i)$, ποσότητες που υπολογίζονται από τους αλγορίθμους forward και backward που είδαμε παραπάνω. Πράγματι,

$$P(\mathbf{x},\pi_i = k,\pi_{i+1} = l \mid \theta) = P(x_1,x_2,...,x_L,\pi_i = k,\pi_{i+1} = l \mid \theta) =$$
$$= P(x_1,x_2,...,x_i,\pi_i = k \mid \theta) P(x_{i+1},x_{i+2},...,x_L,\pi_{i+1} = l \mid x_1,x_2,...,x_i,\pi_i = k,\theta)$$

και επειδή δεν υπάρχει εξάρτηση ούτε των παρατηρήσεων ούτε των καταστάσεων από προηγούμενες παρατηρήσεις, καταλήγουμε:

$$P(\mathbf{x}, \pi_i = k, \pi_{i+1} = l \mid \theta) =$$
$$P(x_1, x_2, ..., x_i, \pi_i = k \mid \theta) P(x_{i+1}, x_{i+2}, ..., x_L, \pi_{i+1} = l \mid \pi_i = k, \theta)$$

Από τη σχέση (8.27), βλέπουμε ότι:

$$f_k(i) = P(x_1, x_2, ..., x_i, \pi_i = k)$$

Επιπλέον,

$$P(x_{i+1}, x_{i+2}, ..., x_L, \pi_{i+1} = l \mid \pi_i = k, \theta) =$$
$$P(x_{i+1}, \pi_{i+1} = l \mid \pi_i = k, \theta) P(x_{i+2}, ..., x_L \mid x_{i+1}, \pi_{i+1} = l, \pi_i = k, \theta) \quad (8.45)$$

Ο πρώτος όρος του γινομένου, γίνεται:

$$P(x_{i+1}, \pi_{i+1} = l \mid \pi_i = k, \theta) =$$
$$= P(\pi_{i+1} = l \mid \pi_i = k) P(x_{i+1} \mid \pi_{i+1} = l) = \quad (8.46)$$
$$= a_{kl} e_l(x_{i+1})$$

ενώ ο δεύτερος, με τη βοήθεια της σχέσης (24):

$$P(x_{i+2}, ..., x_L \mid x_{i+1}, \pi_{i+1} = l, \pi_i = k, \theta) =$$
$$= P(x_{i+2}, ..., x_L \mid \pi_{i+1} = l) = b_l(i+1) \quad (8.47)$$

Αντικαθιστώντας τις σχέσεις (8.46) και (8.47) στην (8.45), έχουμε:

$$P(x_{i+1}, x_{i+2}, ..., x_L, \pi_{i+1} = l \mid \pi_i = k, \theta) = a_{kl} e_l(x_{i+1}) b_l(i+1)$$

και τελικά:

$$P(\mathbf{x}, \pi_i = k, \pi_{i+1} = l \mid \theta) = f_k(i) a_{kl} e_l(x_{i+1}) b_l(i+1) \quad (8.48)$$

απ' όπου με χρήση του θεωρήματος του Bayes:

$$P(\pi_i = k, \pi_{i+1} = l \mid \mathrm{x}, \theta) = \frac{f_k(i) a_{kl} e_l(x_{i+1}) b_l(i+1)}{P(\mathrm{x})} \quad (8.49)$$

όπου $f_k(i)$, $b_k(i)$, είναι οι ποσότητες που υπολογίζονται από τους αλγορίθμους forward και backward. Με όμοιο τρόπο, είδαμε στη σχέση (8.31) ότι:

$$P(\pi_i = k \mid \mathbf{x}) = \frac{f_k(i) b_k(i)}{P(\mathbf{x})}$$

Τότε, από τον ορισμό της αναμενόμενης τιμής, έχουμε για τις μεταβάσεις:

$$A_{kl} = \sum_{\pi} P(\pi \mid \mathbf{x}, \theta^t) A_{kl}(\pi) = \frac{1}{P(\mathbf{x})} \sum_i f_k(i) a_{kl} e_l(x_{i+1}) b_l(i+1) \quad (8.50)$$

και αντίστοιχα για τις πιθανότητες γεννήσεως:

$$E_k(b) = \sum_{\pi} P(\pi \mid \mathbf{x}, \theta^t) E_k(b, \pi) = \frac{1}{P(\mathbf{x})} \sum_{\{i \mid x_i = b\}} f_k^j(i) b_k^j(i) \quad (8.51)$$

και αντικαθιστώντας στη σχέση (8.41) έχουμε:

$$Q(\theta \mid \theta^t) = \sum_{k=1} \sum_b E_k(b) \log e_k(b) + \sum_{k=0} \sum_{l=1} A_{kl} \log \alpha_{kl} \quad (8.52)$$

Και αυτή η συνάρτηση τελικά μεγιστοποιείται από τους εκτιμητές των εξισώσεων (8.38) και (8.39). Έτσι σε γενικές γραμμές ο αλγόριθμος των Baum-Welch αποτελείται από το *E*-βήμα (Expectation) στο οποίο υπολογίζονται οι ποσότητες $f_k(i)$, $b_k(i)$, από τους αλγορίθμους forward και backward αντίστοιχα και κατόπιν υπολογίζονται οι αναμενόμενες τιμές για τις πιθανότητες A_{kl}, και $E_k(b)$ από τις σχέσεις (8.50) και (8.51). Έτσι καθορίζεται μονοσήμαντα η συνάρτηση *Q*. Το *M*-βήμα (Maximization) περιορίζεται στο να τεθούν οι παραπάνω τιμές των A_{kl} και $E_k(b)$ στις σχέσεις (8.38) και (8.39), να υπολογιστούν ξανά οι Ε.Μ.Π. και η πιθανοφάνεια του μοντέλου. Ο αλγόριθμος τερματίζεται απλά όταν οι μεταβολές στην πιθανοφάνεια (log-likelihood) και αντίστοιχα στην συνάρτηση *Q* μετά από κάποια βήματα είναι μικρότερές από μια προκαθορισμένη τιμή (threshold).

Μέθοδοι Gradient-Descent

Ο αλγόριθμος Baum-Welch, παρουσιάζει όπως είδαμε μια σειρά από θαυμαστά προτερήματα. Το βασικό, είναι ότι είναι μαθηματικά εγγυημένος ότι θα συγκλίνει, ενώ δευτερευόντως, αποδεικνύεται και αρκετά γρήγορος. Το βασικό του μειονέκτημα, είναι ότι απαιτεί η ανανέωση (update) των παραμέτρων να γίνεται, αφού όλο το σύνολο εκπαίδευσης έχει παρουσιαστεί (batch mode of learning). Επιπλέον, είναι «απορροφητικός», υπό την έννοια ότι αν μια παράμετρος μηδενιστεί, δεν υπάρχει περίπτωση να αποκτήσει πλέον διαφορετική τιμή. Αυτά τα δυο μειονεκτήματα, προσπάθησαν να αντιμετωπίσουν οι Baldi και Chauvin (Baldi & Chauvin, 1994).

Η μέθοδος Gradient-Descent, είναι μια γενική ευριστική μέθοδος ελαχιστοποίησης ενέργειας. Αν θεωρήσουμε μια συνάρτηση, *f* με *n* μεταβλητές:

$$f\left(\mathbf{x}\right)=f\left(x_1,x_2,...,x_n\right)$$

η οποία είναι παραγωγίσιμη, τότε ένα τοπικό της ελάχιστο, στο πολυδιάστατο σημείο

$$\left(\mathbf{x^0}\right)=\left(x_1{}^0,x_2{}^0,...,x_n{}^0\right)$$

μπορεί να προσδιοριστεί, προσεγγίζοντας διαδοχικά το σημείο μέσω της σχέσης:

$$f(\mathbf{x}^{t+1})=f(\mathbf{x}^t)-\eta\Delta f(\mathbf{x})$$

όπου, Δ είναι το διάνυσμα των μερικών παραγώγων της συνάρτησης και η ένας αρκετά μικρός ρυθμός μάθησης (learning rate). Στην περίπτωση μας, ως «ενέργεια» μπορεί να οριστεί το αντίθετο του λογαρίθμου της πιθανοφάνειας (negative log-likelihood), ενώ οι παράμετροι είναι φυσικά το σύνολο των πιθανοτήτων μεταβάσεως και γεννήσεως. Για κάθε παράμετρο ω του μοντέλου, η ανανέωση επιτυγχάνεται θέτοντας:

$$\omega^{t+1}=\omega^t-\eta\frac{\partial\ell\left(\mathbf{x}\mid\theta\right)}{\partial\omega} \tag{8.53}$$

Άρα το πρόβλημα έγκειται στον υπολογισμό των μερικών παραγώγων του λογαρίθμου της πιθανοφάνειας, ως προς τις παραμέτρους του μοντέλου. Αφού ορίσουμε

$$\ell=-\log P\left(\mathbf{x}\mid\theta\right)$$

για τον υπολογισμό των μερικών παραγώγων κινούμαστε ως εξής:

$$\begin{aligned}\frac{\partial\log P\left(\mathbf{x}\mid\theta\right)}{\partial\omega}&=\frac{1}{P\left(\mathbf{x}\mid\theta\right)}\frac{\partial P\left(\mathbf{x}\mid\theta\right)}{\partial\omega}\\&=\frac{1}{P\left(\mathbf{x}\mid\theta\right)}\frac{\partial P\left(\mathbf{x},\pi\mid\theta\right)}{\partial\omega}\\&=\frac{1}{P\left(\mathbf{x}\mid\theta\right)}\sum_\pi P\left(\mathbf{x},\pi\mid\theta\right)\frac{\partial\log P\left(\mathbf{x},\pi\mid\theta\right)}{\partial\omega}\\&=\sum_\pi P\left(\pi\mid\mathbf{x},\theta\right)\frac{\partial\log P\left(\mathbf{x},\pi\mid\theta\right)}{\partial\omega}\end{aligned} \tag{8.54}$$

Με χρήση της σχέσης (8.43), και αφού λογαριθμοποιήσουμε, έχουμε για τη μερική παράγωγο ως προς τις πιθανότητες μεταβάσεως:

$$\begin{aligned}\frac{\partial\log P\left(\mathbf{x},\pi\mid\theta\right)}{\partial a_{kl}}&=\frac{\partial\left(\sum_{k=0}\sum_{l=1}A_{kl}\left(\pi\right)\log a_{kl}\right)}{\partial a_{kl}}\\&=A_{kl}\left(\pi\right)\frac{\partial\left(\sum_{k=0}\sum_{l=1}\log a_{kl}\right)}{\partial a_{kl}}\\&=\frac{A_{kl}\left(\pi\right)}{a_{kl}}\end{aligned} \tag{8.55}$$

Άρα, η σχέση (8.54) τελικά γίνεται:

$$\frac{\partial \log P(\mathbf{x} \mid \theta)}{\partial a_{kl}} = \sum_{\pi} P(\pi \mid \mathbf{x}, \theta) \frac{A_{kl}(\pi)}{a_{kl}} = \frac{A_{kl}}{a_{kl}} \tag{8.56}$$

Με εντελώς όμοιο συλλογισμό, υπολογίζουμε και την αντίστοιχη μερική παράγωγο για μια πιθανότητα γεννήσεως:

$$\frac{\partial \log P(\mathbf{x} \mid \theta)}{\partial e_k(b)} = \sum_{\pi} P(\pi \mid \mathbf{x}, \theta) \frac{E_k(\pi, b)}{e_k(b)} = \frac{E_k(b)}{e_k(b)} \tag{8.57}$$

Παρατηρούμε, ότι η μερική παράγωγος του λογαρίθμου της πιθανοφάνειας ως προς τις παραμέτρους του μοντέλου, είναι ίση με την αντίστοιχη μερική παράγωγο της βοηθητικής συνάρτησης Q από τη σχέση (8.52). Μπορούμε δηλαδή, να συμπεράνουμε ότι και οι δυο μέθοδοι κινούνται προς την ίδια κατεύθυνση στην ελαχιστοποίηση ενέργειας, και οδηγούν τελικά στο ίδιο αποτέλεσμα καθώς στο τοπικό ελάχιστο και οι δυο μηδενίζονται (Baldi & Chauvin, 1994).

Για να ολοκληρωθεί ο αλγόριθμος, χρειάζεται ένα ακόμη απαραίτητο βήμα. Πρέπει με κάποιο τρόπο, να περιορίσουμε τον αλγόριθμο, έτσι ώστε οι τιμές των παραμέτρων να είναι πραγματικές πιθανότητες. Αν εφαρμόσουμε τη σχέση (8.53), πραγματοποιώντας Gradient Descent, πάνω στις πραγματικές τιμές των παραμέτρων, είναι πιθανόν να πάρουμε ακόμα και αρνητικούς εκτιμητές για τις πιθανότητες αυτές. Κατά συνέπεια, είναι απαραίτητο να ορίσουμε κάποιες βοηθητικές μεταβλητές, οι οποίες να παίρνουν πάντα τιμές μεταξύ 0 και 1, να πραγματοποιήσουμε την ελαχιστοποίηση και κατόπιν να ανακτήσουμε τις τιμές των πιθανοτήτων. Στη συγκεκριμένη περίπτωση, χρησιμοποιήσαμε τη μέθοδο των Krogh και Riss (Krogh & Riis, 1999), η οποία θεωρεί τον λεγόμενο «soft-max» μετασχηματισμό. Για την ακρίβεια, για τις πιθανότητες μετάβασης a_{kl}, ορίζονται μια σειρά από βοηθητικές παραμέτρους $z_{kl,}$ έτσι ώστε:

$$a_{kl} = \frac{\exp(z_{kl})}{\sum_{l'} \exp(z_{kl'})} \tag{8.58}$$

Πραγματοποιώντας τώρα, την ελαχιστοποίηση με τη μέθοδο Gradient Decent, όχι στα a_{kl}, αλλά στα z_{kl}:

$$z_{kl}^{t+1} = z_{kl}^{t} - \eta \frac{\partial \ell^t}{\partial z_{kl}} \tag{8.59}$$

παίρνουμε τις ανανεωμένες παραμέτρους για τις πιθανότητες μετάβασης:

$$a_{kl}^{(t+1)} = \frac{z_{kl}^{(t)} \exp\left(-\eta \frac{\partial \ell^{(t)}}{\partial z_{kl}} \right)}{\sum_{l'} z_{kl'}^{(t)} \exp\left(-\eta \frac{\partial \ell^{(t)}}{\partial z_{kl'}} \right)} \tag{8.60}$$

Με αλλαγή μεταβλητής στη σχέση (8.56), μπορούμε να υπολογίσουμε επίσης τις μερικές παραγώγους του αντίθετου του λογαρίθμου της πιθανοφάνειας ως προς τις βοηθητικές παραμέτρους z_{kl}:

$$\frac{\partial \ell}{\partial z_{kl}} = -\left[A_{kl} - a_{kl} \sum_{l'} A_{kl'} \right] \tag{8.61}$$

Αντικαθιστώντας, τη σχέση (8.61) στη σχέση (8.60), παίρνουμε μια έκφραση η οποία εξαρτάται πλέον μόνο από τις τιμές των παραμέτρων στην προηγούμενη επανάληψη και από τις αναμενόμενες τιμές τους:

$$\alpha_{kl}^{(t+1)} = \frac{\alpha_{kl}^{(t)} \exp\left(-\eta \left[A_{kl} - a_{kl} \sum_{l'} A_{kl'} \right] \right)}{\sum_{l'} \alpha_{kl'}^{(t)} \exp\left(-\eta \left[A_{kl} - a_{kl} \sum_{l'} A_{kl'} \right] \right)} \tag{8.62}$$

Με αυτόν τον τρόπο, η ανανέωση των παραμέτρων επιτυγχάνεται χωρίς να χρειαστεί να υπολογιστούν σε κάποιο ενδιάμεσο βήμα οι βοηθητικές μεταβλητές. Προφανώς, με όμοιο τρόπο εργαζόμαστε και για τις πιθανότητες γέννησης. Με τη μέθοδο αυτή, μπορούμε πλέον να πραγματοποιήσουμε «ομαλή» (smooth) εκπαίδευση, χωρίς τον κίνδυνο μηδενισμού κάποιων παραμέτρων ο οποίος ενυπάρχει στον αλγόριθμο Baum-Welch, αλλά και να πραγματοποιήσουμε τη λεγόμενη διαδικασία «online training», κατά την οποία οι παράμετροι μπορούν να ανανεώνονται κατάλοιπο-κατάλοιπο, και όχι με την παρουσίαση

ολόκληρου του συνόλου εκπαίδευσης. Η μέθοδος αυτή, θα φανεί χρήσιμη παρακάτω, καθώς μόνο με αυτή μπορεί να πραγματοποιηθεί εκπαίδευση Δεσμευμένης Μέγιστης Πιθανοφάνειας (Conditional Maximum Likelihood-CML). Παρ' όλα αυτά το βασικό της μειονέκτημα, καθώς πρόκειται περί ευριστικής μεθόδου, είναι η αυθαίρετη επιλογή της παραμέτρου η (ρυθμός μάθησης), η οποία οδηγεί σε αστάθεια στη διαδικασία εκπαίδευσης και η μικρή της ταχύτητα σύγκλισης στον πολυδιάστατο χώρο των παραμέτρων.

Viterbi training

Τέλος, πρέπει να αναφέρουμε και έναν άλλο, αρκετά απλό τρόπο εκτίμησης παραμέτρων, ο οποίος μάλιστα παρουσιάστηκε σχετικά αργά στη βιβλιογραφία. Ο αλγόριθμος αυτός ονομάζεται Viterbi training ή αλλιώς Segmental *k*-means algorithm και παρουσιάστηκε από τους Juang και Rabiner το 1990 (Juang & Rabiner, 1990). Ο αλγόριθμος στηρίζεται στην απλή ιδέα, να εναλλάσσονται διαδοχικά δύο βήματα: α) εύρεση του καλύτερου μονοπατιού με τον αλγόριθμο Viterbi, και β) θεώρηση του μονοπατιού αυτού ως πραγματικό και χρήση των τύπων (8.34) και (8.35) για την εύρεση των παραμέτρων του μοντέλου.

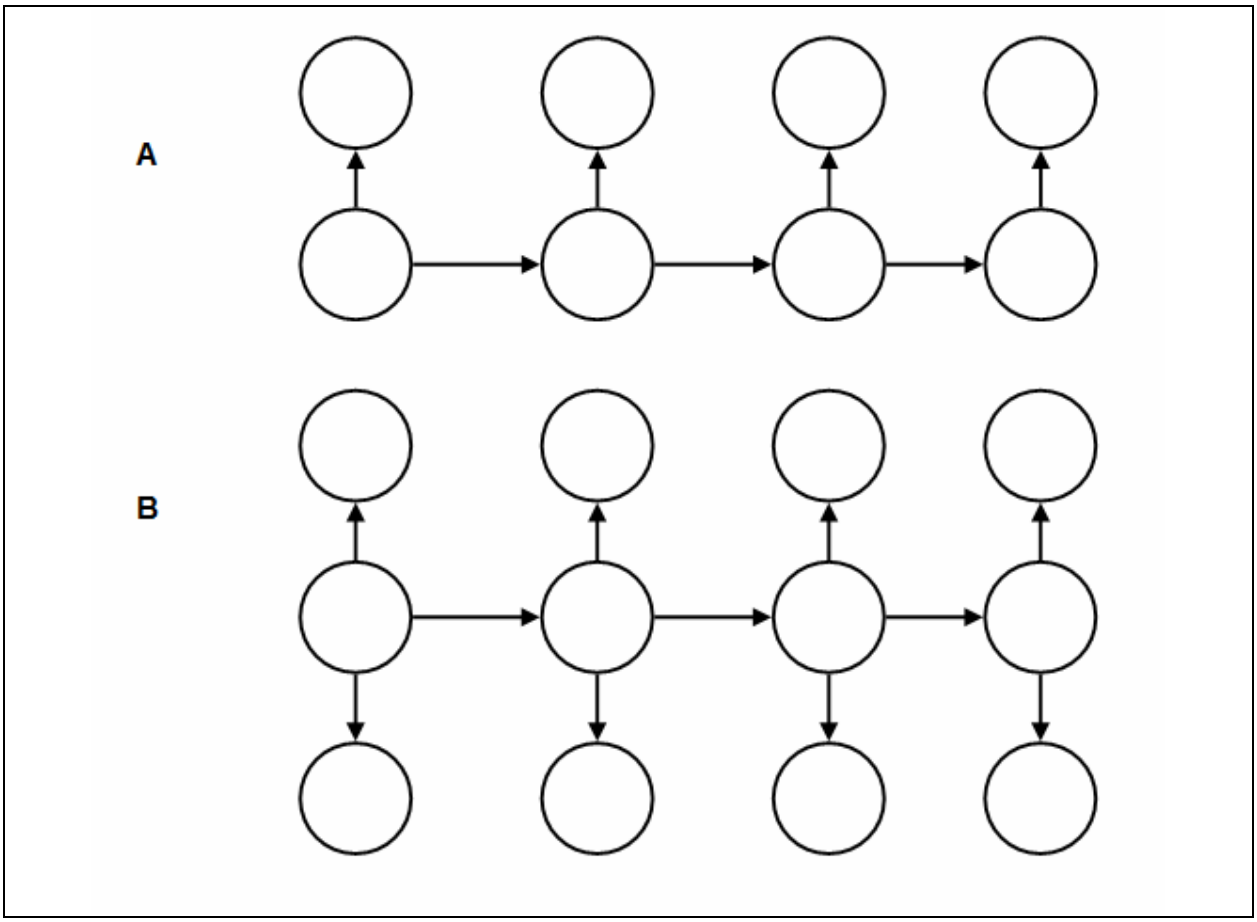
Η ιδέα πίσω από τον αλγόριθμο, είναι αρκετά απλή και έλκει την καταγωγή της από τη βιβλιογραφία της στατιστικής ομαδοποίησης (clustering), όπου και ο αλγόριθμος προτάθηκε για πρώτη φορά το 1968 με το όνομα k-means algorithm (MacQueen, 1967). Με την εύρεση του βέλτιστου μονοπατιού και την παραδοχή ότι αυτό είναι «παρατηρηθέν», η από κοινού πιθανοφάνεια μιας αλληλουχίας και του μονοπατιού αυτού, δίνεται από ένα απλό γινόμενο (8.20), οπότε είναι ανοικτός ο δρόμος για την απλή εφαρμογή των τύπων που δίνουν τους εκτιμητές μέγιστης πιθανοφάνειας. Οι Juang και Rabiner έδειξαν ότι η διαδοχική εφαρμογή των δύο αυτών βημάτων, οδηγεί σε μονοτονική αύξηση της τιμής της πιθανοφάνειας, την οποία και ονόμασαν πιθανοφάνεια βελτιστοποιημένη για τις καταστάσεις (state-optimized likelihood). Ένα επιπλέον ενδιαφέρον σημείο αυτού του αλγορίθμου, είναι το γεγονός ότι, επειδή τα πιθανά μονοπάτια είναι πεπερασμένα, ο αλγόριθμος συγκλίνει σε ένα τοπικό μέγιστο της πιθανοφάνειας σε πεπερασμένο αριθμό βημάτων.

Φυσικά, όπως και όλοι οι αλγόριθμοι βελτιστοποίησης που αναφέρουμε εδώ, δεν υπάρχει εγγύηση ότι ο αλγόριθμος θα εντοπίσει ένα ολικό ελάχιστο της πιθανοφάνειας. Έτσι, εξαρτάται από τις αρχικές τιμές αν το τοπικό αυτό ελάχιστο θα είναι κάποιο το οποίο θα δίνει μοντέλα με σημαντική προγνωστική αξία (το ίδιο φυσικά ισχύει και για τον αλγόριθμο Baum-Welch αλλά και για τους αλγόριθμους gradient). Διαισθητικά, ο αλγόριθμος αυτός είναι μια διακριτή εκδοχή του αλγορίθμου Baum-Welch. Εκεί που ο τελευταίος βρίσκει την αναμενόμενη τιμή αθροίζοντας τη συνεισφορά όλων των πιθανών μονοπατιών, ο πρώτος επιλέγει να κρατήσει τη συνεισφορά μόνο του μονοπατιού με την καλύτερη πιθανότητα. Αυτό έχει σαν αποτέλεσμα, να δίνει, θεωρητικά πάντα, λίγο χειρότερα αποτελέσματα, αλλά από την άλλη έχει το μεγάλο πλεονέκτημα ότι απαιτεί μόνο ένα πέρασμα του αλγορίθμου Viterbi εκεί που ο αλγόριθμος Baum-Welch απαιτεί ένα πέρασμα του Forward και ένα του Backward (κατά συνέπεια, απαιτεί το μισό χρόνο υπολογισμού, πράγμα σημαντικό σε περίπτωση μεγάλων μοντέλων ή/και μεγάλου όγκου δεδομένων).

8.3. Class Hidden Markov Model

8.3.1. Ορισμοί

Το βασικό χαρακτηριστικό των παραπάνω κλασικών προσεγγίσεων στην εκπαίδευση ενός μοντέλου, αποτελεί το γεγονός ότι είναι «μέθοδοι χωρίς επίβλεψη» (unsupervised methods). Το μόνο που έχει να κάνει ο χρήστης, είναι να προσφέρει κάποιες αλληλουχίες-παραδείγματα, και οι αλγόριθμοι θα βρουν την βέλτιστη κατανομή των παραμέτρων για να περιγράψουν τα δεδομένα αυτά. Ένα βασικό πρόβλημα στη βιολογία, ανακύπτει στην περίπτωση κατά την οποία θέλουμε να εκπαιδεύσουμε ένα μεγάλο μοντέλο, το οποίο να περιγράφει με μεγάλη ακρίβεια (χρησιμοποιώντας διαφορετικές καταστάσεις) διαφορετικές περιοχές μέσα σε μια πρωτεϊνική αλυσίδα.



Εικόνα 8.11: *Γραφική αναπαράσταση του κλασικού HMM και του CHMM. Α. Το βασικό HMM όπως το έχουμε περιγράψει έως τώρα. Οι καταστάσεις (στην κάτω γραμμή) ακολουθούν μια μαρκοβιανή αλυσίδα $1^{ης}$ τάξης, κάθε κατάσταση της οποίας «παράγει» με διαφορετική πιθανότητα τα παρατηρήσιμα σύμβολα Β. Το CHMM στην πιο γενική του μορφή, μπορεί να θεωρηθεί ως ένα μοντέλο που σε κάθε κατάσταση «παράγει» ταυτόχρονα δύο σειρές από παρατηρήσιμα σύμβολα. Η μία είναι τα σύμβολα της αλληλουχίας (όπως ακριβώς το HMM), ενώ η άλλη είναι η αλληλουχία των σημάνσεων. Η αλληλουχία των καταστάσεων του μοντέλου, εξακολουθεί να είναι μαρκοβιανή $1^{ης}$ τάξης.*

Χαρακτηριστικότατο παράδειγμα, είναι αυτό της πρόγνωσης των διαμεμβρανικών περιοχών. Είναι φυσικό, να θέλουμε να κατασκευάσουμε ένα πολύπλοκο μοντέλο, το οποίο να περιέχει αρκετές και διαφορετικές καταστάσεις για την διαμεμβρανική περιοχή, άλλες για την εξωτερική περιοχή της μεμβράνης και άλλες για την εσωτερική, έτσι ώστε να μπορούμε να μοντελοποιήσουμε καλύτερα το πρόβλημα και να αποτυπώσουμε πιο αποτελεσματικά την πρότερη βιολογική γνώση. Στην περίπτωση αυτή, θα έπρεπε να χωρίσουμε τις περιοχές, να εκπαιδεύσουμε 3 διαφορετικά μοντέλα και κατόπιν να τα ενώσουμε αυθαίρετα με κάποιες πιθανότητες μετάβασης. Το πρόβλημα αυτό παρακάμπτεται, αν χρησιμοποιήσουμε μια μέθοδο εκμάθησης «με επίβλεψη» (supervised learning). Η μέθοδος αυτή, η οποία βασίζεται στις σημασμένες αλληλουχίες ή αλληλουχίες με ετικέτες (labeled sequences), προτάθηκε από τον Krogh και αντιστοιχεί στο λεγόμενο Class Hidden Markov Model (Anders. Krogh, 1994). Συγκεκριμένα, με την τεχνική αυτή, κάθε αλληλουχία συμβόλων

$$\mathbf{x} = x_1, x_2, ..., x_{L-1}, x_L ,$$

συνοδεύεται, και από μια αλληλουχία σημάνσεων ή ετικετών (labels)

$$\mathbf{y} = y_1, y_2, ..., y_{L-1}, y_L$$

Στη συγκεκριμένη περίπτωση της πρόγνωσης των διαμεμβρανικών τμημάτων, οι σημάνσεις είναι 3: μια για τα διαμεμβρανικά τμήματα (M), μια για την εσωτερική περιοχή (I) και μια για την εξωτερική (O). Επιπλέον, είναι αναγκαίο πλέον να ορίσουμε μια κατανομή για την πιθανότητα ταύτισης μιας κατάστασης με

μια δεδομένη σήμανση. Στην πράξη, ομαδοποιούμε τις καταστάσεις σε ομάδες οι οποίες έχουν μια βιολογική σημασία, δηλαδή ομαδοποιούμε τις καταστάσεις που αντιστοιχούν σε διαμεμβρανικά τμήματα κ.ο.κ. Χρειαζόμαστε έτσι, μια μεταβλητή $\delta_k(c)$ που δηλώνει την πιθανότητα η κατάσταση k να έχει σήμανση c. Η κατανομή που ακολουθεί αυτή η μεταβλητή, είναι προφανώς διωνυμική, αλλά σε όλες τις εφαρμογές που θα χρησιμοποιήσουμε, είναι απλώς μια δίτιμη συνάρτηση (delta function) που παίρνει απλώς την τιμή 1 αν η κατάσταση συμφωνεί με τη σήμανση και 0 σε αντίθετη περίπτωση. Δηλαδή, δεν επιτρέπουμε σε μια κατάσταση να συμπίπτει με περισσότερες από μια σημάνσεις.

8.3.2. Πιθανοφάνεια

Όπως γίνεται πλέον φανερό, με την εισαγωγή των σημάνσεων, ένας τρόπος να επιτύχουμε «επιβλεπόμενη μάθηση», είναι να θεωρήσουμε ως αντικειμενική συνάρτηση την από κοινού πιθανότητα $P(\mathbf{x},\mathbf{y}|\theta)$ των ακολουθιών $\mathbf{x}$ με τις σημάνσεις $\mathbf{y}$, δεδομένου του μοντέλου:

$$P(\mathbf{x},\mathbf{y}\mid\theta)=\sum_{\pi}P(\mathbf{x},\mathbf{y},\pi\mid\theta)=\sum_{\pi\in\Pi_y}P(\mathbf{x},\pi\mid\theta)=\sum_{\pi\in\Pi_y}a_{B\pi_1}\prod_{i=1}^{L}e_{\pi i}(x_i)a_{\pi_i\pi_{i+1}}$$

Η μοναδική διαφορά της σχέσης αυτής, με τη σχέση (8.25), είναι ότι η άθροιση πρέπει να γίνει μόνο για αυτά τα μονοπάτια Π_y, τα οποία διέρχονται μέσω καταστάσεων οι οποίες είναι συμβατές με τη σήμανση. Η ολική από κοινού πιθανότητα των ακολουθιών και των σημάνσεων δεδομένου του μοντέλου, μπορεί να υπολογιστεί με κάποιες τετριμμένες τροποποιήσεις των γνωστών αλγορίθμων Forward και Backward, που συναντήσαμε παραπάνω. Η μόνη διαφορά, έγκειται στον πολλαπλασιασμό των ενδιάμεσων μεταβλητών, με τη δίτιμη συνάρτηση (0,1), η οποία δείχνει την συμφωνία καταστάσεων και σημάνσεων. Κατά συνέπεια, ο τροποποιημένος αλγόριθμος Forward, είναι:

Τροποποιημένος αλγόριθμος Forward

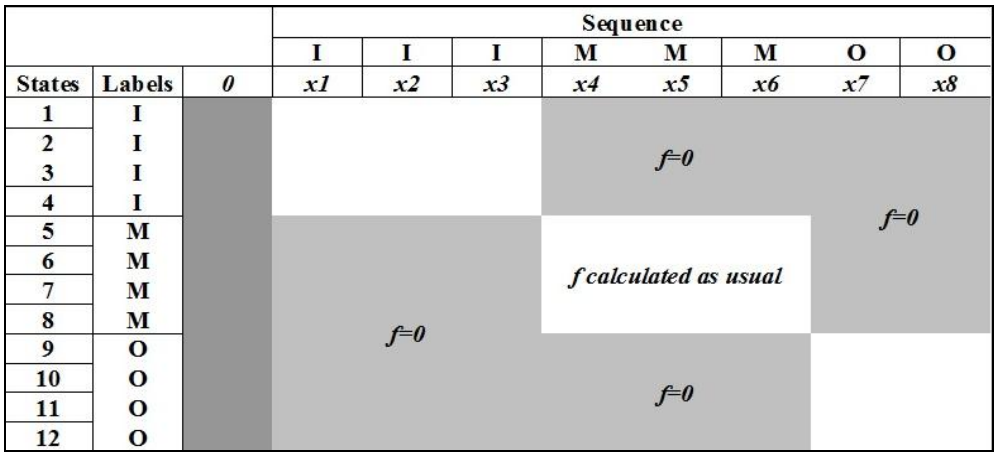
$$\forall k\neq B,\ i=0:\ f_B(0)=1,\ f_k(0)=0$$
$$\forall 1\le i\le L:\ f_l(i)=e_l(x_i)\delta_l(y_i)\sum_k f_k(i-1)a_{kl} \qquad (8.63)$$
$$P(\mathbf{x},\mathbf{y}|\theta)=\sum_k f_k(L)a_{kE}$$

Εντελώς όμοια, ο τροποποιημένος αλγόριθμος Backward, θα είναι:

Τροποποιημένος αλγόριθμος Backward

$$\forall k,\ i=L:\ b_k(L)=a_{kE}$$
$$\forall\ 1\le i<L:\ b_k(i)=\sum_l a_{kl}e_l(x_{i+1})\delta_l(y_{i+1})b_l(i+1) \qquad (8.64)$$
$$P(\mathbf{x},\mathbf{y}|\theta)=\sum_l a_{Bl}e_l(x_1)b_l(1)$$

Διαισθητικά, οι αλγόριθμοι αυτοί απλώς μηδενίζουν τις περιοχές των πινάκων Forward και Backward, στις οποίες δεν υπάρχει συμφωνία καταστάσεων και σημάνσεων (Εικόνα 8.12). Λόγω του ότι το σύνολο των επιτρεπτών μονοπατιών Πy, είναι υποσύνολο του συνόλου όλων των πιθανών μονοπατιών Π, αντιλαμβανόμαστε ότι $P(\mathbf{x},\mathbf{y}|\theta)\le P(\mathbf{x}|\theta)$.



Εικόνα 8.12: *Διαγραμματική απεικόνιση του πίνακα Forward για σημασμένες αλληλουχίες. Έχουμε ένα μοντέλο με 12 υποθετικές καταστάσεις, και μια αλληλουχία με 8 κατάλοιπα για τα οποία είναι γνωστές οι σημάνσεις (labels). Είναι φανερό ότι για τα κατάλοιπα και τις καταστάσεις που δεν συμφωνούν με την σήμανση, οι τιμές του πίνακα απλά μηδενίζονται.*

8.3.3. Εκτίμηση παραμέτρων

Μέγιστη Πιθανοφάνεια

Με την εισαγωγή των αλγορίθμων για σημασμένες αλληλουχίες, είναι εφικτό πλέον να πραγματοποιήσουμε εκτίμηση μέγιστης πιθανοφάνειας:

$$\theta^{ML} = \arg\max_{\theta} P\left(\mathbf{x},\mathbf{y} \mid \theta\right)$$

Όλοι οι αλγόριθμοι που είδαμε ότι ισχύουν για τις μη σημασμένες αλληλουχίες, ισχύουν με μικρές παραλλαγές και εδώ. Λόγω του ότι, οι αλληλουχίες και οι σημάνσεις είναι ανεξάρτητες, οι σχέσεις (8.65) και (8.66) γίνονται αντίστοιχα:

$$A_{kl} = \frac{1}{P\left(\mathbf{x},\mathbf{y} \mid \theta\right)} \sum_i f_k\left(i\right) a_{kl} e_l\left(x_{i+1}\right) \delta_l\left(y_{i+1}\right) b_l\left(i+1\right) \tag{8.67}$$

$$E_k\left(b\right) = \frac{1}{P\left(\mathbf{x},\mathbf{y} \mid \theta\right)} \sum_{\left\{i \mid x_i^j = b\right\}} f_k\left(i\right) b_k\left(i\right) \tag{8.68}$$

όπου οι ποσότητες $P(\mathbf{x},\mathbf{y}|\theta)$, $f_i(i)$ και $b_k(i)$, υπολογίζονται πλέον από τους τροποποιημένους αλγόριθμους που είδαμε παραπάνω. Με όλα τα παραπάνω, μπορούμε άνετα να πραγματοποιήσουμε εκπαίδευση μέγιστης πιθανοφάνειας, είτε με τη μέθοδο των Baum-Welch είτε με τη μέθοδο Gradient Descent.

Δεσμευμένη Μέγιστη Πιθανοφάνεια

Με την εισαγωγή της έννοιας των σημασμένων ακολουθιών, ο Krogh (Anders. Krogh, 1994) πρότεινε και μια νέα μέθοδο εκπαίδευσης, αυτήν της Δεσμευμένης Μέγιστης Πιθανοφάνειας (Conditional Maximum Likelihood). Με το κριτήριο αυτό, αναζητούμε πλέον να μεγιστοποιήσουμε την πιθανότητα των σημάνσεων, δεδομένων των ακολουθιών και του μοντέλου:

$$\theta^{CML} = \arg\max_{\theta} P\left(\mathbf{y} \mid \mathbf{x},\theta\right) = \arg\max_{\theta} \frac{P\left(\mathbf{x},\mathbf{y} \mid \theta\right)}{P\left(\mathbf{x} \mid \theta\right)}$$

Ο Krogh, έδειξε επίσης (Anders. Krogh, 1994), ότι η προσέγγιση αυτή αποτελεί γενική περίπτωση της από παλιότερα γνωστής διαδικασίας εκπαίδευσης με το κριτήριο της Μέγιστης Αμοιβαίας Πληροφορίας (Maximum Mutual Information) όπως αναφέρεται στον (Rabiner, 1989). Ο αρνητικός λογάριθμος αυτής της δεσμευμένης πιθανοφάνειας, μπορεί να εκφραστεί ως η διαφορά:

$$\ell = -\log P\left(\mathbf{y} \mid \mathbf{x},\theta\right) = \ell_c - \ell_f$$

όπου:

$$\ell_c = -\log P\left(\mathbf{x},\mathbf{y} \mid \theta\right)$$

$$\ell_f = -\log P\left(\mathbf{x} \mid \theta\right)$$

Με τους δείκτες *c* και *f*, ονομάζουμε αντίστοιχα την πιθανοφάνεια που υπολογίζεται στη φάση όπου οι σημάνσεις λαμβάνονται υπόψη (clamped phase), και αυτή στην οποία οι σημάνσεις δεν υπολογίζονται (free-running phase). Από τις σχέσεις (8.56) και (8.57), μπορούμε εύκολα να υπολογίσουμε τις αναμενόμενες τιμές και τις μερικές παραγώγους των παραμέτρων του μοντέλου:

$$\frac{\partial \ell}{\partial a_{kl}} = \frac{\partial \ell_c}{\partial a_{kl}} - \frac{\partial \ell_f}{\partial a_{kl}} = -\frac{A_{kl}^c - A_{kl}^f}{a_{kl}} \tag{8.69}$$

$$\frac{\partial \ell}{\partial e_k\left(b\right)} = \frac{\partial \ell_c}{\partial e_k\left(b\right)} - \frac{\partial \ell_f}{\partial e_k\left(b\right)} = -\frac{E_k^c(b) - E_k^f(b)}{e_k(b)} \tag{8.70}$$

Όπως είναι φανερό, ο αλγόριθμος Baum-Welch δεν μπορεί να χρησιμοποιηθεί, λόγω του ότι η διαφορά των αναμενόμενων τιμών θα δώσει αρνητικές εκτιμήσεις για τις παραμέτρους. Παρ' όλα αυτά, με τη χρήση των μερικών παραγώγων μπορούμε να προχωρήσουμε σε εκπαίδευση με τη μέθοδο Gradient Descent. Δουλεύοντας με τον ίδιο μετασχηματισμό όπως και στις προηγούμενες παραγράφους για τις πιθανότητες μεταβάσεως, η μερική παράγωγος της πιθανοφάνειας ως προς τις βοηθητικές μεταβλητές, θα είναι:

$$\frac{\partial \ell}{\partial z_{kl}} = -\left(A_{kl}^c - A_{kl}^f - a_{kl} \sum_{l'} \left(A_{kl'}^c - A_{kl'}^f \right) \right) \tag{8.71}$$

και τελικά η σχέση η οποία θα δώσει τις ανανεωμένες τιμές των παραμέτρων στην επανάληψη *t*, θα είναι, εντελώς ανάλογα:

$$\alpha_{kl}^{(t+1)} = \frac{\alpha_{kl}^{(t)} \exp\left(-\eta \left[A_{kl}^c - A_{kl}^f - a_{kl} \sum_{l'} \left(A_{kl'}^c - A_{kl'}^f \right) \right] \right)}{\sum_{l'} \alpha_{kl'}^{(t)} \exp\left(-\eta \left[A_{kl}^c - A_{kl}^f - a_{kl} \sum_{l'} \left(A_{kl'}^c - A_{kl'}^f \right) \right] \right)} \tag{8.72}$$

Μια βασική αδυναμία αυτής της μεθόδου, η οποία θεωρείται και «μέθοδος εκπαίδευσης προσανατολισμένη στο διαχωρισμό» (Discriminative Training), είναι ότι απαιτεί διπλάσιο υπολογιστικό χρόνο και μνήμη στον υπολογιστή, καθώς χρειάζονται δυο «περάσματα» των αλγορίθμων για τις δυο διαφορετικές πιθανοφάνειες που υπολογίζονται. Παρ' όλα αυτά, τα πλεονεκτήματα τα οποία προσδίδει μια τέτοια διαδικασία υπερτερούν, κυρίως λόγω της καλύτερης ικανότητας πρόγνωσης που προσφέρει. Το μεγάλο μειονέκτημα του αλγορίθμου, είναι η μικρή του ταχύτητα και η ανάγκη εύρεσης μιας βέλτιστης τιμής για την επιλογή της παραμέτρου η (ρυθμός μάθησης). Το πρόβλημα αυτό, λύθηκε σε μεγάλο βαθμό με τη δημιουργία ενός αλγορίθμου ο οποίος χρησιμοποιεί διαφορετικούς ρυθμούς μάθησης για κάθε παράμετρο, αλλά έχει και την ιδιότητα να τους αναπροσαρμόζει κατά τη διάρκεια της διαδικασίας εκπαίδευσης (Bagos, Liakopoulos, & Hamodrakas, 2004).

8.3.4. Αποκωδικοποίηση

1-best Decoding

Καθώς είδαμε ότι με την προσέγγιση των σημασμένων ακολουθιών, αποκτάμε μια αντιστοίχιση των καταστάσεων με τις σημάνσεις, μια λογική απορία θα ήταν αν θα μπορεί να έχει κανείς μια μέθοδο αποκωδικοποίησης η οποία να βρίσκει την πιο πιθανή αλληλουχία (μονοπάτι) των σημάνσεων και όχι των καταστάσεων. Μια αλληλουχία καταστάσεων σύμφωνα με τον ορισμό που δώσαμε, αντιστοιχεί μονοσήμαντα σε μια αλληλουχία σημάνσεων αλλά το αντίθετο δεν ισχύει, καθώς είναι δυνατόν να έχουμε περισσότερες από μια αλληλουχίες καταστάσεων οι οποίες δίνουν την ίδια σήμανση κατά μήκος της αλληλουχίας συμβόλων. Ακριβής αλγόριθμος, ο οποίος να υπολογίζει την *ολικά καλύτερη* αλληλουχία σημάνσεων δεν υπάρχει, αλλά έχουν προταθεί προσεγγιστικές λύσεις.

Ο αλγόριθμος 1-best (Krogh, 1997), είναι μια τροποποίηση του αλγορίθμου N-best, ο οποίος είχε προταθεί παλαιότερα για αναγνώριση ομιλίας (Schwartz & Chow, 1990). Στην ουσία, πρόκειται για έναν ευριστικό αλγόριθμο δυναμικού προγραμματισμού, ο οποίος αναζητά την εύρεση της πιο πιθανής αλληλουχίας σημάνσεων ymax αντί αυτή της πιο πιθανής αλληλουχίας καταστάσεων. Ο αλγόριθμος, για κάθε

θέση i της αλληλουχίας, αποθηκεύει όλες τις πιθανές «ενεργές υποθέσεις» hi-1 για τη σήμανση, οι οποίες αποτελούνται από όλες τις πιθανές αλληλουχίες σημάνσεων μέχρι εκείνο το σημείο. Κατόπιν, για κάθε κατάσταση l «προωθεί» τις υποθέσεις προσθέτοντας στο τέλος κάθε μια από τις πιθανές σημάνσεις yi και διαλέγει την καλύτερη. Η όλη διαδικασία επαναλαμβάνεται ως το τέλος της αλληλουχίας. Σε αντίθεση με τον αλγόριθμο του Viterbi, ο αλγόριθμος 1-best δεν χρειάζεται αναδρομή αλλά έχει και μεγαλύτερες υπολογιστικές απαιτήσεις τόσο σε μνήμη όσο και σε πραγματοποιούμενες πράξεις.

Αλγόριθμος 1-best

$$
\begin{aligned}
& i = 1: \ \gamma_l(h_1) = a_{Bl} e_l(x_1) \\
& \forall 1 < i \le L: \ \gamma_l(h_i \mathrm{y}_i) = e_l(x_i) \sum_k \gamma_k(h_{i-1}) a_{kl} \\
& P\left(\mathbf{x}, \mathbf{y}^{\max} \mid \theta\right) = \sum_k \gamma_k(h_L) a_{kE}
\end{aligned}
\tag{8.73}
$$

Η πιθανότητα της βέλτιστης αυτή σήμανσης, είναι πάντα μεγαλύτερη ή ίση από την πιθανότητα του βέλτιστου μονοπατιού καταστάσεων, καθώς πολλά διαφορετικά μονοπάτια καταστάσεων συνεισφέρουν σε αυτή, άρα:

$$P\left(\mathbf{x}, \pi^{\max} \mid \theta\right) \le P\left(\mathbf{x}, \mathbf{y}^{\max} \mid \theta\right) \le P\left(\mathbf{x} \mid \theta\right)$$

Optimal Accuracy Posterior Decoder

Πριν από μερικά χρόνια, ο Kall και συνεργάτες παρουσίασαν έναν εναλλακτικό αλγόριθμο, τον λεγόμενο Optimal Accuracy Posterior Decoder (Kall, Krogh, & Sonnhammer, 2005). Ο αλγόριθμος αυτός μοιάζει πάρα πολύ με τον Posterior-Viterbi, αλλά διαθέτει κάποιες διαφοροποιήσεις οι οποίες εμφανίζονται ειδικά στην περίπτωση του CHMM (αν και γενικά, οι δύο αλγόριθμοι αποδίδουν σχεδόν ταυτόσημα στα περισσότερα προβλήματα που έχουν εφαρμοστεί). Ο αλγόριθμος, λειτουργεί ως εξής: για κάθε θέση στην αλληλουχία, αθροίζει τις εκ των υστέρων πιθανότητες της κάθε σήμανσης (posterior label probabilities -PLPs), χρησιμοποιώντας τη σχέση (8.29), και μετά υπολογίζοντας μόνο τις επιτρεπτές μεταβάσεις, υπολογίζει με έναν αλγόριθμο τύπου Viterbi τη βέλτιστη αλληλουχία των σημάνσεων που μεγιστοποιεί την ποσότητα:

$$\pi^{OAPD} = \arg\max_{\pi} \sum_{i=1}^{L} \left\{ \delta\left(\pi_i, \pi_{i+1}\right) \left(\sum_k P\left(\pi_i \mid \mathbf{x}\right) \lambda_k\left(c\right) \right) \right\} \tag{8.74}$$

Αλγόριθμος Optimal Accuracy Posterior Decoder

$$
\begin{aligned}
& \forall k \ne B, \ i = 0: \ A_B(0) = 0, \ A_k(0) = -\infty \\
& \forall \ 1 \le i \le L: \ A_l(i) = P\left(y_i = c^l \mid \mathbf{x}, \theta\right) + \max_k \left\{ A_k(i-1) \delta\left(k, l\right) \right\} \\
& P(\mathbf{x}, \pi^{OAPD} \mid \theta) = \max_k \left\{ A_k(L) \delta\left(k, E\right) \right\}
\end{aligned}
\tag{8.75}
$$

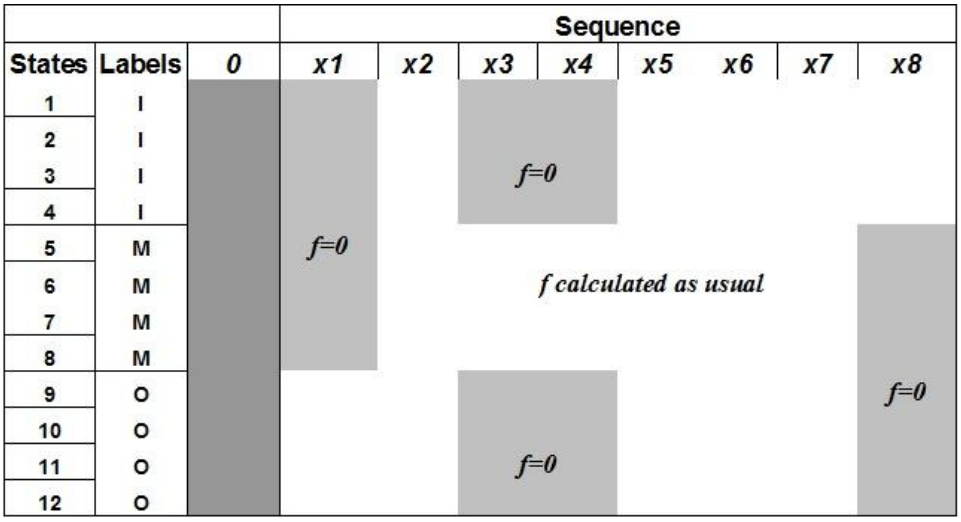
Όπως είπαμε, ο αλγόριθμος αυτός μοιάζει πολύ με τον Posterior-Viterbi και οι μόνες διαφορές τους είναι ότι, α) χρησιμοποιεί τις εκ των υστέρων πιθανότητες των σημάνσεων και όχι των καταστάσεων, και β) αντί για το γινόμενο των πιθανοτήτων αυτών, μεγιστοποιεί το άθροισμά τους. Αυτό, είναι αναγκαίο γιατί ο Posterior-Viterbi υπολογίζει τελικά ένα μονοπάτι καταστάσεων, ενώ ο Optimal Accuracy Posterior Decoder μια αλληλουχία από σημάνσεις που είναι όμως συμβατές με το μοντέλο, αλλά ενδέχεται να περιέχουν πολλά εναλλακτικά μονοπάτια. Για το λόγο αυτό, οι τελική πιθανότητα που αποδίδει ο αλγόριθμος αυτός, δεν είναι συγκρίσιμη σε απόλυτες τιμές με τις πιθανότητες των άλλων αλγορίθμων.

8.3.5. Δεσμευμένη πρόγνωση και αλγόριθμοι για ενσωμάτωση πειραματικής πληροφορίας

Σε διάφορα βιολογικά προβλήματα, όπως για παράδειγμα στην περίπτωση της πρόγνωσης των διαμεμβρανικών πρωτεϊνών, είναι γνωστό ότι η ενσωμάτωση μιας ακόμα και περιορισμένης πειραματικά

προσδιορισμένης πληροφορίας σχετικά με την τοπολογία θα βελτίωνε κατά ένα μεγάλο μέρος την απόδοση ακόμα και των καλύτερων μεθόδων. Με την ανάπτυξη εύκολων και γρήγορων πειραματικών τεχνικών βασισμένων σε συντήξεις γονιδίων (gene fusions), με την οποία καθορίζεται η θέση του αμινοτελικού άκρου μιας πρωτεΐνης, προτάθηκε ότι αυτές οι τεχνικές συνδυαζόμενες μαζί θα βελτιώσουν κατά ένα μεγάλο μέρος την απόδοση των προγνωστικών μεθόδων και την εφαρμογή τους σε πλήρως προσδιορισμένα γονιδιώματα (Drew et al., 2002; Melen, Krogh, & von Heijne, 2003). Δεδομένα στην βιβλιογραφία υπάρχουν αρκετά τα οποία δείχνουν και άλλους εναλλακτικούς τρόπους προσδιορισμού της θέσης διαφόρων τμημάτων της αλληλουχίας (αντισώματα, πρωτεόλυση κλπ) αλλά οι πιο ολοκληρωμένες πειραματικές αποδείξεις σε μεγάλη κλίμακα γι' αυτήν την βελτίωση, ήρθαν από μελέτες που αφορούν πρωτεΐνες της *E. coli* (Rapp et al., 2004) και του *S. cerevisiae* (Kim, Melen, & von Heijne, 2003).

Από τις ήδη διαθέσιμες προγνωστικές μεθόδους, το TMHMM και το HMMTOP (Tusnady & Simon, 2001), προσφέρουν επιλογή στο χρήστη έτσι ώστε να ενσωματώσει στην πρόγνωσή του πειραματικά προσδιορισμένη πληροφορία για την τοπολογία, όπως επίσης τέτοια επιλογή, προσφέρεται και από την συνδυασμένη πρόγνωση διαμεμβρανικών ελίκων και πεπτιδίων οδηγητών από την μέθοδο Phobius (Kall, Krogh, & Sonnhammer, 2004). Η πρώτη όμως προσπάθεια να αναλυθούν αυτοί οι αλγόριθμοι, και να ενσωματωθούν με γενική μορφή σε κάθε αλγόριθμο αποκωδικοποίησης, έγινε από τους (Bagos, Liakopoulos, & Hamodrakas, 2006) και η εφαρμογή τους στον αλγόριθμο HMMTM.



Εικόνα 8.13: *Διαγραμματική απεικόνιση του πίνακα Forward, όταν γίνεται η ενσωμάτωση της εκ των προτέρων πληροφορίας. Έχουμε ένα (υποθετικό) μοντέλο από 12 καταστάσεις, και μια αλληλουχία* ***x*** *από 8 κατάλοιπα. Στον υπολογισμό της πιθανοφάνειας της αλληλουχίας* ***x,*** *ενσωματώνεται η πληροφορία ότι τα κατάλοιπα 3,4 είναι διαμεμβρανικά, ότι το κατάλοιπο 1 βρίσκεται στην εξωκυττάρια πλευρά και ότι το κατάλοιπο 8 βρίσκεται στο κυτταρόπλασμα.*

Οι τροποποιήσεις αυτές, είναι εντελώς ανάλογες με τις τροποποιήσεις που επιτρέπουν την εκπαίδευση με σημασμένες αλληλουχίες, με τη διαφορά ότι εδώ χρησιμοποιούνται στο πλαίσιο της αποκωδικοποίησης. Οι συγγραφείς έδειξαν, ότι οι πιθανοφάνειες που προκύπτουν με αυτό τον τρόπο, μπορούν να εκφραστούν σαν εκ των υστέρων πιθανότητες των πειραματικών πληροφοριών δεδομένης της αλληλουχίας και του μοντέλου, διατηρώντας έτσι την πιθανοθεωρητική ερμηνεία των αποτελεσμάτων. Παρόμοιας φύσεως τροποποιήσεις εισάγονται σε όλους τους αλγόριθμους αποκωδικοποίησης τους οποίους αναφέραμε ήδη, και με αυτόν τον τρόπο, είμαστε σε θέση να πραγματοποιήσουμε δεσμευμένη πρόγνωση για οποιασδήποτε μορφής πειραματική πληροφορία, και με οποιαδήποτε μέθοδο αποκωδικοποίησης.

Κατ' αρχάς ορίζουμε την έννοια της *Πληροφορίας (Information)* **ω,** η οποία αποτελείται από $1 \le r \le L$, αμοιβαίως αποκλειόμενα, μη-μηδενικού μήκους τμήματα στην αλληλουχία, τα οποία συμβολίζονται με ω_1, ω_2, ...ω_r, και για τα οποία γνωρίζουμε την ακριβή πειραματικώς προσδιορισμένη τοπολογία και κατά συνέπεια τη σήμανση. Όμοια με την περίπτωση της εκπαίδευσης με σημασμένες αλληλουχίες, μπορούμε να ορίσουμε μια δίτιμη συνάρτηση που να δείχνει τη συμφωνία της σήμανσης με την κατάσταση:

$$d_k(i) = \begin{cases} 0, \text{ if } \lambda_k(\omega_i) \neq 0 \text{ and } i \in \omega^r \\ 1, \text{ otherwise} \end{cases}$$

Οι απλές τροποποιήσεις οι οποίες προτάθηκαν για την περίπτωση του αλγόριθμου forward, συνίστανται στο να θέσουμε τη forward μεταβλητή f ίση με το μηδέν για κάθε θέση i και κατάσταση k η οποία δε συμφωνεί με την πληροφορία (Εικόνα 8.13). Αυτό είναι ακριβώς όμοιο με τη διαδικασία εκπαίδευσης με σημασμένες αλληλουχίες, όπου επιτρέπουμε μόνο τα μονοπάτια Π_y τα οποία είναι σε συμφωνία με τη σήμανση **y**, να συνεισφέρουν στην ολική πιθανοφάνεια. Στη συγκεκριμένη περίπτωση, επιτρέπουμε απλά, τη συνεισφορά μόνο των μονοπατιών Π_ω τα οποία είναι σε συμφωνία με την εκ των προτέρων πληροφορία **ω**. Εκτός από τους αλγόριθμους forward και backward που υπολογίζουν την πιθανοφάνεια, με τον ίδιο ακριβώς τρόπο τροποποιούνται και όλοι οι αλγόριθμοι αποκωδικοποίησης που είδαμε σε προηγούμενες ενότητες. Με τους αλγόριθμους αυτούς, μπορούμε να κάνουμε δεσμευμένες προγνώσεις ενσωματώνοντας όλη τη διαθέσιμη εκ των προτέρων πληροφορία για μια αλληλουχία. Τέτοια παραδείγματα, που αφορούν τις διαμεμβρανικές πρωτεΐνες, θα δούμε σε επόμενο κεφάλαιο. Λεπτομέρειες για τους τροποποιημένους αλγόριθμους, μπορούν να βρεθούν στην αντίστοιχη δημοσίευση (Bagos, et al., 2006).

8.3.6. Λεπτομέρειες της αλγοριθμικής υλοποίησης

Ένα σημαντικό πρόβλημα στην υλοποίηση των αλγορίθμων, προκύπτει καθώς οι διαδοχικοί πολλαπλασιασμοί μικρών πιθανοτήτων οδηγούν με μαθηματική ακρίβεια σε τελικό μηδενισμό των πιθανοτήτων, λόγω της μαθηματικής ακρίβειας των υπολογισμών (underflow error). Η χρησιμοποίηση του λογαρίθμου της πιθανοφάνειας λύνει εν μέρει αυτό το πρόβλημα, καθώς τα γινόμενα των πιθανοτήτων εύκολα μετατρέπονται σε αθροίσματα λογαριθμικών πιθανοτήτων.

Παρ' όλα αυτά, χρειάζεται μια επιπλέον τροποποίηση στην περίπτωση κατά την οποία πρέπει να υπολογιστεί ο λογάριθμος ενός αθροίσματος από τους λογαρίθμους των προσθετέων. Έτσι, λειτουργούμε ως εξής:

$$\log(a+b) = \log\left(a\left(1+{}^{a}\!/_{b}\right)\right) = \log(a) + \log\left(1+{}^{a}\!/_{b}\right)$$
$$= \log(a) + \log\left(1+\exp\left(\log(a)-\log(b)\right)\right) \quad (8.76)$$

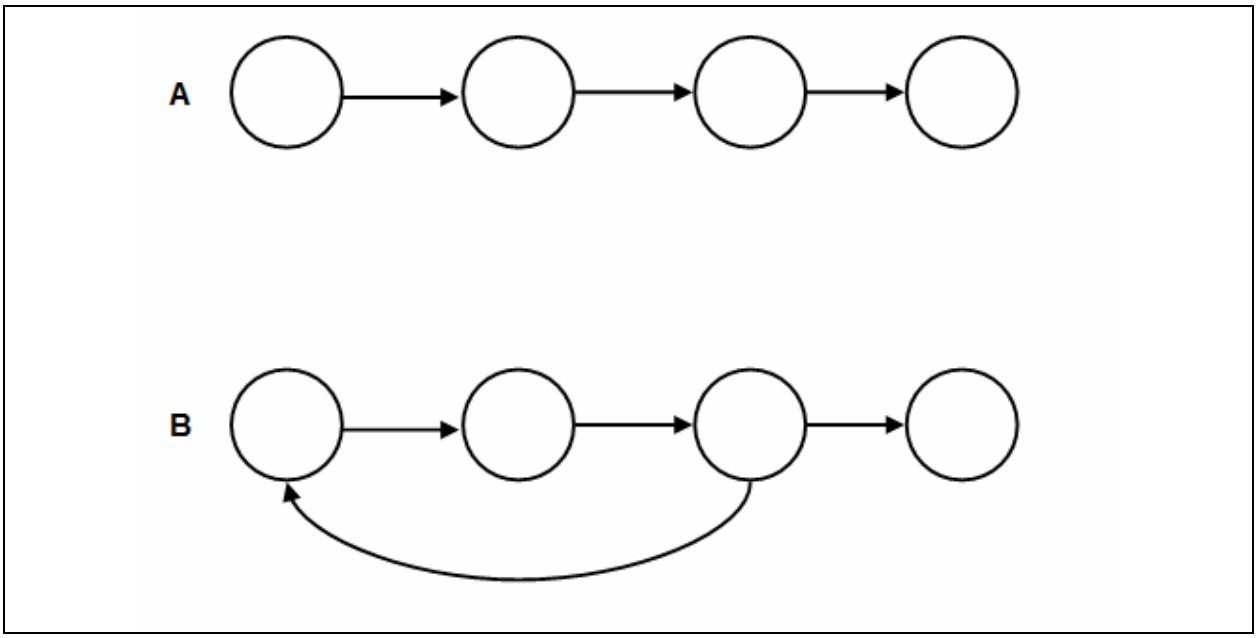
Όταν η διαφορά $|\log(a)-\log(b)|$ είναι μικρότερη από -37, το οποίο είναι το όριο για τους αριθμούς διπλής ακρίβειας, τότε ο δεύτερος προσθετέος της παραπάνω σχέσης θα είναι περίπου ίσος με 0, αλλιώς χρησιμοποιείται αυτούσια η σχέση (8.76).

8.4. Σχεδιασμός της δομής των μοντέλων

Ίσως το πιο ενδιαφέρον αλλά και πιο επίπονο στάδιο στην κατασκευή ενός προγνωστικού αλγορίθμου βασισμένο σε HMM, είναι η διαδικασία σχεδιασμού του κατάλληλου μοντέλου. Παρ΄ όλο που διάφορες μέθοδοι έχουν προταθεί για την εύρεση της Βέλτιστης δομής ενός μοντέλου είτε με γενικότερες τεχνικές Μέγιστης Πιθανοφάνειας (Ostendorf & Singer, 1997; Vasko, El-Jaroudi, & Boston, 1996), είτε με χρήση Γενετικών αλγορίθμων (Won, Prugel-Bennett, & Krogh, 2004; Yada, Ishikawa, H., & Asai, 1994), γενικώς, σε πολύπλοκα μοντέλα αυτές δεν αποδίδουν τόσο καλά και το καλύτερο μοντέλο προκύπτει πάντα από ανθρώπινο χέρι. Η διαδικασία, απαιτεί άριστη γνώση των αλγορίθμων που χρησιμοποιούνται όσο και βαθιά κατανόηση του βιολογικού προβλήματος το οποίο καλούμαστε να αντιμετωπίσουμε. Η μεγάλη δύναμη του HMM είναι ότι μπορεί να μοντελοποιήσει πολλά βιολογικά προβλήματα, τουλάχιστον όταν αυτά αφορούν αλληλουχίες στις οποίες υπάρχει «διαμερισματοποίηση», δηλαδή περιοχές με ξεκάθαρες διαφορές στη σύσταση των αμινοξέων. Όταν πηγαίνουμε στα πιο σύνθετα CHMM, όταν δηλαδή υπάρχουν σημάνσεις, κάθε σήμανση αντιστοιχεί σε ένα επιπλέον υπο-μοντέλο.

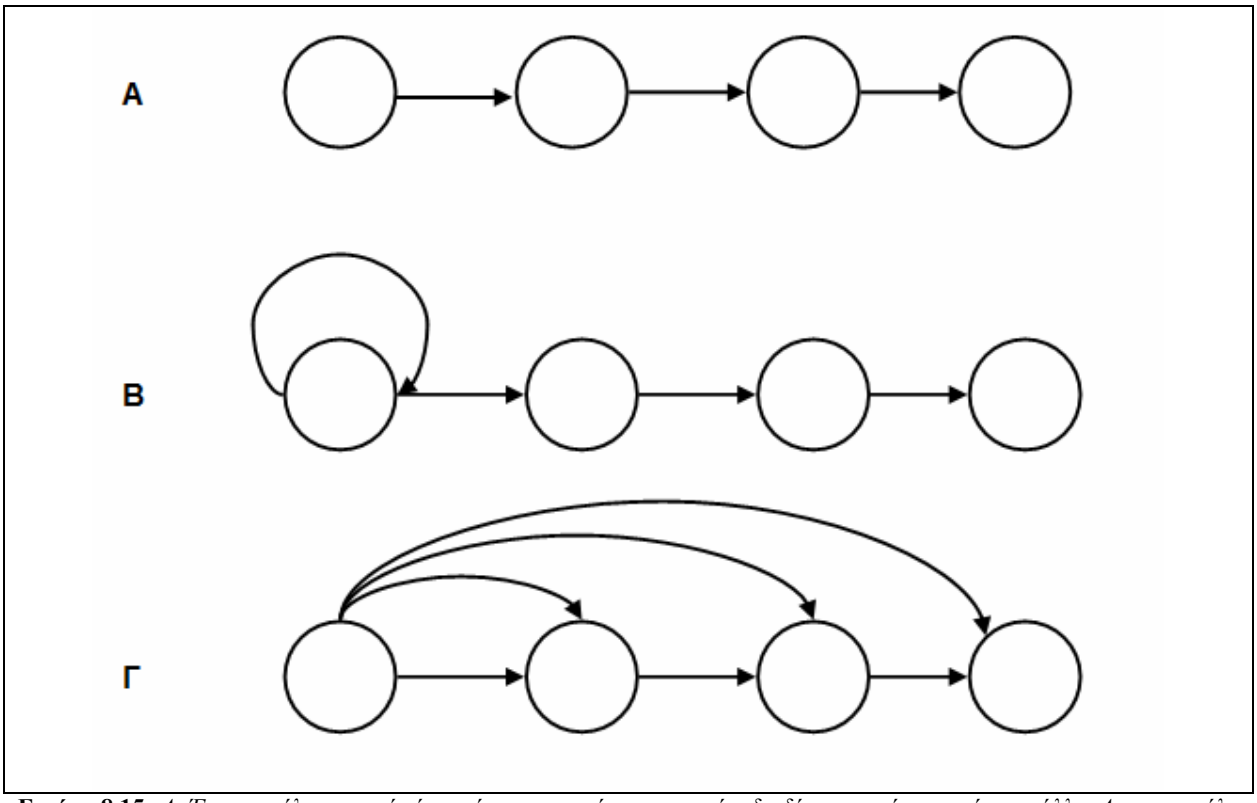
Γενικά, για να μοντελοποιήσουμε σύνθετες περιοχές (πχ δευτεροταγή δομή, διαμεμβρανικά τμήματα, εσώνια-εξώνια κ.ο.κ.) πρέπει να λάβουμε υπόψη μας δύο πράγματα: τις πιθανότητες εμφάνισης συμβόλων και τις μεταβάσεις μεταξύ των καταστάσεων. Πολλές φορές, ειδικά όταν έχουμε περιοχές με μεγάλο μήκος, χρειάζεται να χρησιμοποιήσουμε «παρόμοιες» καταστάσεις. Δηλαδή, καταστάσεις που είναι μεν διαφορετικές αλλά περιμένουμε να έχουν τις ίδιες ιδιότητες. Για παράδειγμα, στα διαμεμβρανικά τμήματα πρέπει να έχουμε μεγάλο αριθμό καταστάσεων για να μοντελοποιήσουμε τα διαμεμβρανικά τμήματα τα οποία έχουν

μήκος 15 έως 35 αμινοξικά κατάλοιπα. Σε αυτές τις περιπτώσεις, μια καλή πρακτική είναι να κάνουμε το λεγόμενο «parameter tying» ή αλλιώς «sharing». Πρακτικά, αυτό σημαίνει ότι θα έχουμε περισσότερες από μία καταστάσεις οι οποίες όμως θα έχουν τις ίδιες πιθανότητες εμφάνισης συμβόλων. Αυτό επιτυγχάνεται με το να αθροίζονται οι αναμενόμενες τιμές *E* στους αλγόριθμους forward και backward. Το βασικό πλεονέκτημα μια τέτοιας στρατηγικής, είναι ότι ελαττώνονται κατά πολύ οι παράμετροι του μοντέλου (αν ενώσουμε 10 καταστάσεις, θα έχουμε τελικά μόνο 20 πιθανότητες εμφάνισης συμβόλων αντί για 10*20).



Εικόνα 8.14: *Α. Τυπική εικόνα ενός γραμμικού μοντέλου (left to right). Το μοντέλο αν αφήσει μια δεδομένη κατάσταση, δεν επιστρέφει ποτέ. Β. Ένα τυπικό παράδειγμα κυκλικού μοντέλου. Το μοντέλο μπορεί να επιστρέψει (απεριόριστες φορές) σε μια δεδομένη κατάσταση.*

Γενικά, τα μοντέλα μπορεί να είναι γραμμικά ή κυκλικά, ανάλογα με το τι μονοπάτια επιτρέπονται κάθε φορά. Στο γραμμικό μοντέλο αν αφήσουμε μια δεδομένη κατάσταση, δεν επιστρέφουμε ποτέ σε αυτήν, ενώ σε ένα κυκλικό μοντέλο μπορούμε να επιστρέψουμε απεριόριστες φορές (Εικόνα 8.14). Το άλλο μεγάλο πρόβλημα, είναι τι είδους πιθανότητες μετάβασης θα επιτρέπουμε. Όπως είδαμε, αν και υπάρχουν προτάσεις για αυτόματη επιλογή του βέλτιστου μοντέλου, συνήθως στις περισσότερες εφαρμογές χρειάζεται ανθρώπινη παρέμβαση. Γενικά, εδώ χρειάζεται και κάποια φαντασία και γνώση κάποιων βασικών κανόνων. Το βασικό που χρειάζεται να λάβουμε υπόψη μας, είναι το μήκος των περιοχών και το πόσο περιοριστικό θέλουμε να είναι το μοντέλο. Ένα απλό μοντέλο με λίγες καταστάσεις, μπορεί να ταιριάζει σε κάθε είδους αλληλουχία που μπορεί να συναντήσει, αλλά μάλλον δεν θα έχει μεγάλη προβλεπτική αξία. Από την άλλη, ένα μοντέλο με πολλούς περιορισμούς, ενδέχεται να αποτύχει σε κάποιες αλληλουχίες που δεν ταιριάζουν σε αυτό. Συνήθως υπάρχουν 3 γενικοί τρόποι για να μοντελοποιηθεί μια περιοχή από *k* καταστάσεις (Εικόνα 8.15). Η πιο απλή περίπτωση είναι όταν η αλληλουχία είναι τελείως γραμμική και η μία κατάσταση ακολουθεί υποχρεωτικά την άλλη. Με αυτόν τον τρόπο, ο οποίος οδηγεί σε μοντέλα ανάλογα με τα profile τα οποία μελετήσαμε σε προηγούμενο κεφάλαιο, αναγκάζουμε το μοντέλο να περάσει ακριβώς μία φορά (ούτε περισσότερες, ούτε λιγότερες) από τις καταστάσεις της περιοχής αυτής (από το υπο-μοντέλο αυτό). Μια άλλη περίπτωση, συναντάμε όταν η περιοχή που θέλουμε να μοντελοποιήσουμε έχει ένα ελάχιστο μήκος, αλλά δεν είναι εύκολο να υπολογίσουμε κάποιο μέγιστο. Σε αυτή την περίπτωση, εισάγουμε μια κατάσταση η οποία έχει μετάβαση προς τον εαυτό της, με συνέπεια να μπορεί να επαναληφθεί επ' άπειρον. Η κατανομή που μπορεί να δώσει μια τέτοια τοπολογία, είναι η γεωμετρική κατανομή. Φυσικά, με κατάλληλη αλλαγή των πιθανοτήτων μετάβασης μπορεί το αναμενόμενο μήκος μιας τέτοιας αλληλουχίας καταστάσεων που θα παράξει αυτό το μοντέλο, να αλλάξει. Τέλος, σε περιπτώσεις που η περιοχή που θέλουμε να μοντελοποιήσουμε έχει ξεκάθαρα ελάχιστα και μέγιστα όρια, μπορούμε να επιλέξουμε μια τοπολογία κατά την οποία από μια δεδομένη κατάσταση επιτρέπονται μεταβάσεις προς όλες τις επόμενές της. Τέτοια περίπτωση θα δούμε ότι προκύπτει στην πρόγνωση των διαμεμβρανικών τμημάτων.



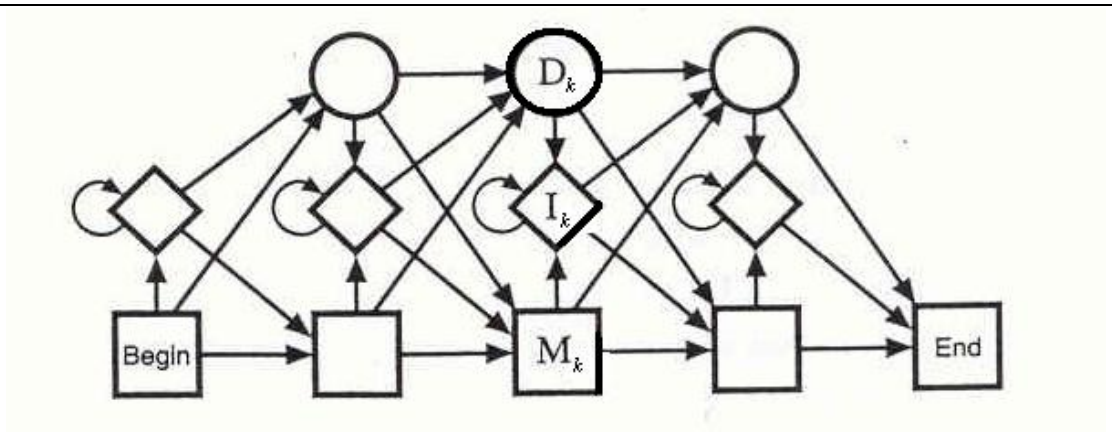
Εικόνα 8.15: *Α. Ένα μοντέλο το οποίο έχει πάντα καταστάσεις οι οποίες διαδέχονται πάντα η μία την άλλη. Αν το μοντέλο φτάσει στην κατάσταση 1, θα περάσει αναγκαστικά και από τις 2,3 και 4. Β. Ένα μοντέλο με 4 καταστάσεις, από τις οποίες όμως η μία επαναλαμβάνεται με μία πιθανότητα. Αυτό το μοντέλο μπορεί να μοντελοποιήσει περιοχές με μήκος πάνω από 4 κατάλοιπα (δεν υπάρχει όμως ανώτατο όριο). Γ. Ένα μοντέλο με ενδιάμεσες μεταβάσεις. Το μοντέλο αυτό μπορεί να περιγράψει περιοχές με μήκος από 1 έως 4 κατάλοιπα.*

Ένα άλλο θέμα που συχνά προκύπτει στον καθορισμό της δομής του μοντέλου είναι ο καθορισμός κάποιων «σιωπηλών καταστάσεων» (silent states). Σιωπηλές καταστάσεις ονομάζονται οι καταστάσεις που δεν παράγουν κάποιο σύμβολο, όπως για παράδειγμα οι καταστάσεις που συνδέονται με την έναρξη και τον τερματισμό του μοντέλου. Μια πιθανή χρησιμότητα τέτοιων καταστάσεων είναι όταν θέλουμε να επιτρέπουμε σε κάθε κατάσταση να συνδέεται με κάποιες από τις επόμενές της. Αν αυτό γίνει χωρίς τη χρήση silent states τότε αυξάνεται εκθετικά ο αριθμός των παραμέτρων (πιθανότητες μεταβάσεως) του μοντέλου, ενώ αν χρησιμοποιηθούν, το πιθανό μειονέκτημα είναι η αύξηση της πολυπλοκότητας των αλγορίθμων οι οποίοι χρειάζονται τροποποίηση.

8.5. Profile Hidden Markov Models

Μια πολύ ειδική κατηγορία Hidden Markov Models, είναι τα λεγόμενα προφίλ (profile) Hidden Markov Models (Eddy, 1998), τα οποία επέκτειναν την έννοια του προφίλ αλληλουχιών (Gribskov, Luthy, & Eisenberg, 1990; Gribskov, McLachlan, & Eisenberg, 1987) τα οποία συναντήσαμε σε προηγούμενα κεφάλαια και την επένδυσαν με πιθανοθεωρητικό χαρακτήρα. Ένα Profile Hidden Markov Model (pHMM), είναι στην ουσία ένα HMM το οποίο περιγράφει με ακρίβεια μια πολλαπλή στοίχιση αλληλουχιών. Η βασική διαφορά του profile από τα κλασικά μοντέλα που αναφέραμε παραπάνω, είναι ότι σε αυτό κάθε κατάσταση περιγράφει μια συγκεκριμένη θέση (στήλη) στην πολλαπλή στοίχιση. Κατά συνέπεια, το μοντέλο έχει ειδικές παραμέτρους ανά θέση (position specific) και ως εκ τούτου η κατεύθυνση των μεταβάσεων θα πρέπει να είναι πάντα μονόδρομη. Γι' αυτόν ακριβώς το λόγο, τα μοντέλα αυτά ονομάζονται μοντέλα *left-to-right* σε αντιδιαστολή με τα κυκλικά που είδαμε παραπάνω τα οποία επιτρέπουν στο μοντέλο να επισκεφθεί μια κατάσταση περισσότερες από μια φορές. Από την πλευρά των κλασικών προφίλ αλληλουχιών που είδαμε σε

προηγούμενα κεφάλαια, το προφίλ HMM είναι μια γενίκευση, στην οποία δεν μοντελοποιούνται μόνο οι πιθανότητες εμφάνισης συμβόλων σε κάθε θέση της πολλαπλής στοίχισης, αλλά μοντελοποιούνται με πιθανοθεωρητικό τρόπο και οι πιθανότητες εισαγωγής κενού και απαλοιφής (insert/deletion). Αυτό είναι πολύ σημαντικό, καθώς το πρόβλημα της επιλογής της ποινής για τα κενά, αποτελούσε μέχρι τώρα ένα πρόβλημα στο οποίο η απάντηση δινόταν με ξεκάθαρα εμπειρικό τρόπο, χωρίς καμία θεωρητική τεκμηρίωση.



Εικόνα 8.16: *Σχηματική αναπαράσταση ενός τυπικού profile Hidden Markov Model*

Μια άλλη σημαντική διαφορά των μοντέλων αυτών, είναι η ύπαρξη ειδικών καταστάσεων οι οποίες δεν εκπέμπουν κανένα σύμβολο, οι οποίες ονομάζονται σιωπηρές καταστάσεις (silent states). Οι καταστάσεις αυτές, χρησιμοποιούνται για να πραγματοποιήσουν μεταβάσεις από ένα κατάλοιπο σε κάποιο άλλο αρκετά κατάλοιπα μακριά του χωρίς να υπάρχει ανάγκη για πολλαπλές μεταβάσεις από την κατάσταση αυτή. Η ύπαρξη τέτοιων καταστάσεων, είναι αναγκαία για να μειώσει τον αριθμό των παραμέτρων του μοντέλου οι οποίες μεγαλώνουν υπερβολικά καθώς κάθε στήλη στην πολλαπλή στοίχιση αντιστοιχεί πλέον σε μια κατάσταση. Όπως είδαμε στην προηγούμενη παράγραφο, μια τοπολογία στην οποία μια κατάσταση επιτρέπεται να κάνει μετάβαση σε *k* επόμενες, θα μπορούσε να χρησιμοποιηθεί για να μοντελοποιήσει τα κενά σε μια πολλαπλή στοίχιση. Παρ' όλα αυτά, μια τέτοια στρατηγική θα αύξανε πάρα πολύ τον αριθμό των παραμέτρων του μοντέλου και θα δημιουργούσε δυσκολία στην εκπαίδευση. Ένα τυπικό pHMM, φαίνεται στην Εικόνα 8.16.

Οι καταστάσεις που παρατηρούνται σε ένα τέτοιο μοντέλο (εκτός αυτών της εκκίνησης και του τερματισμού) χωρίζονται σε 3 κατηγορίες:

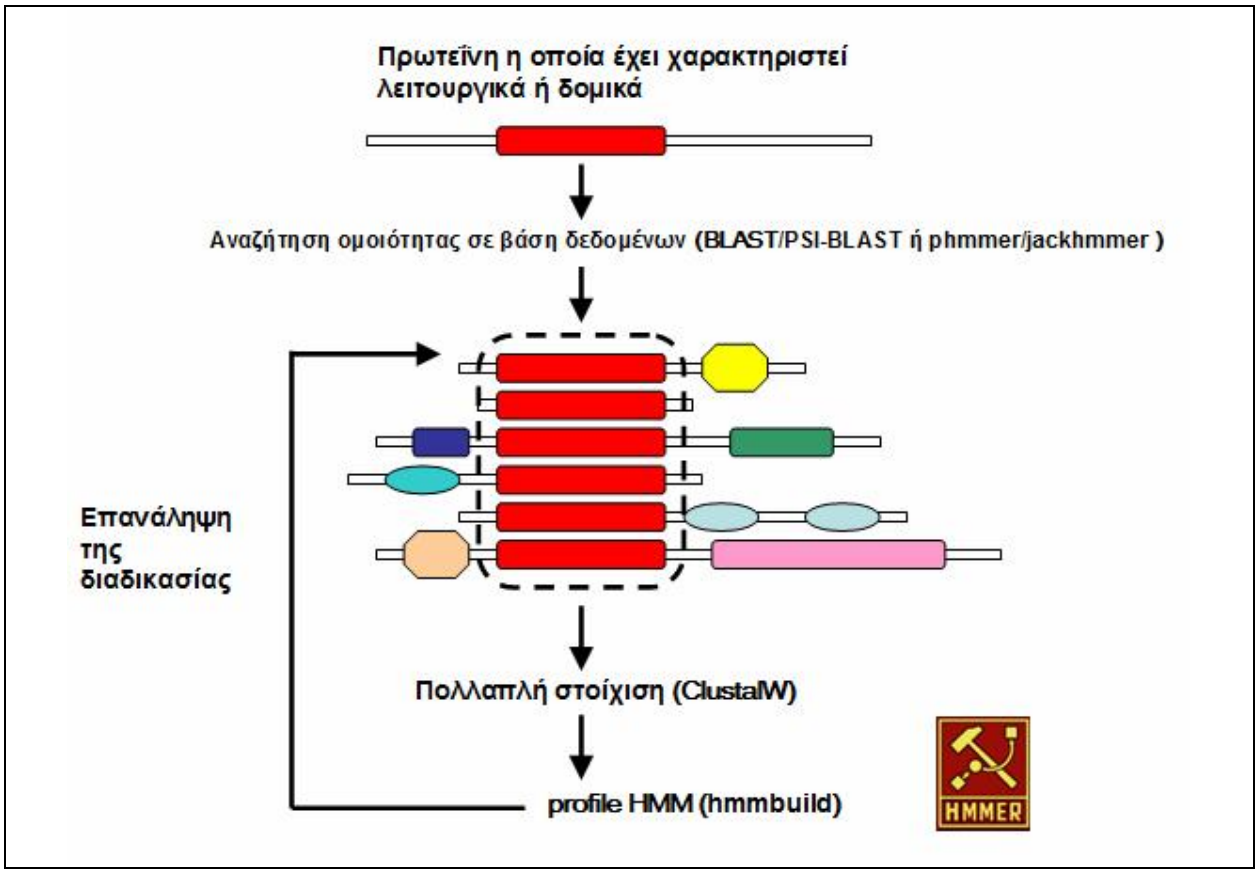
- Καταστάσεις Ταύτισης (Match states) M_k τετράγωνα
- Καταστάσεις Εισαγωγής (Insertion states) I_k ρόμβοι
- Καταστάσεις Απαλοιφής (Deletion states) D_k κύκλοι

και συνδέονται με τις αντίστοιχες πιθανότητες μεταβάσεως, που συμβολίζονται με βέλη. Αντίστοιχα ορίζονται οι πιθανότητες εμφάνισης συμβόλων οι οποίες γεννούν τα σύμβολα σε κάθε κατάσταση. Έτσι υπάρχει και εδώ μια αλληλουχία καταστάσεων η οποία είναι κρυφή και μια αλληλουχία συμβόλων που είναι φανερή, και θεωρούμε ότι παράγεται από την αλληλουχία των καταστάσεων. Οι καταστάσεις ταύτισης και εισαγωγής, είναι κανονικές καταστάσεις οι οποίες συνδέονται μέσω των πιθανοτήτων γέννησης με την εμφάνιση συμβόλων. Η διαφορά τους είναι η εξής: οι μεν καταστάσεις ταύτισης αντιστοιχούν σε στήλες της πολλαπλής στοίχισης οι οποίες στοιχίζονται καλά και άρα αντιστοιχούν σε περιοχή με ομοιότητα, ενώ οι καταστάσεις εισαγωγής, αντιστοιχούν σε περιοχές στις οποίες έχουμε εισαγωγή χαρακτήρων που δεν στοιχίζονται καλά. Οι περιοχές αυτές, οι οποίες δεν υπάρχουν στις υπόλοιπες αλληλουχίες, εμφανίζονται ως κενά τα οποία μοντελοποιούνται μέσω των σιωπηρών καταστάσεων απαλοιφής.

Στο αλγοριθμικό κομμάτι, μετατροπές χρειάζονται για την ενσωμάτωση των σιωπηρών καταστάσεων, καθώς και για να μην καταμετρώνται μεταβάσεις σε καταστάσεις που προηγούνται, σε όλους τους παραπάνω αλγορίθμους. Έτσι είναι δυνατόν με χρήση όλων των βασικών αλγόριθμων δυναμικού προγραμματισμού (Viterbi, backward, forward, Baum-Welch κλπ) που αναφέραμε πριν, να υπολογίσουμε τις παραμέτρους του HMM που περιγράφει μια πολλαπλή στοίχιση.

8.6. Εφαρμογές των profile HMM

Με την εισαγωγή των διαφορετικών καταστάσεων ταύτισης και εισαγωγής, γίνεται μια σημαντική τομή με τις κλασικές μεθόδους στοίχισης, οι οποίες καθώς δεν προϋποθέτουν ένα μοντέλο δεν διαχωρίζουν τις πληροφοριακές θέσεις στη στοίχιση από τις απλές τυχαίες εισαγωγές. Επιπλέον, για πρώτη φορά οι ποινές για την εισαγωγή κενών (gap penalties), δεν τίθενται εκ των προτέρων αλλά εκτιμώνται από τα δεδομένα και αναπαρίστανται με καθαρά πιθανοθεωρητικό τρόπο, αποκλείοντας την υποκειμενική παρέμβαση. Έτσι, με τέτοια μοντέλα, είμαστε σε θέση να πραγματοποιήσουμε ιδιαίτερα ευαίσθητες αναζητήσεις και να εντοπίσουμε απομακρυσμένες ομολογίες (remote homologies), τις οποίες οι παραδοσιακοί αλγόριθμοι στοίχισης δεν θα μπορούσαν να εντοπίσουν.

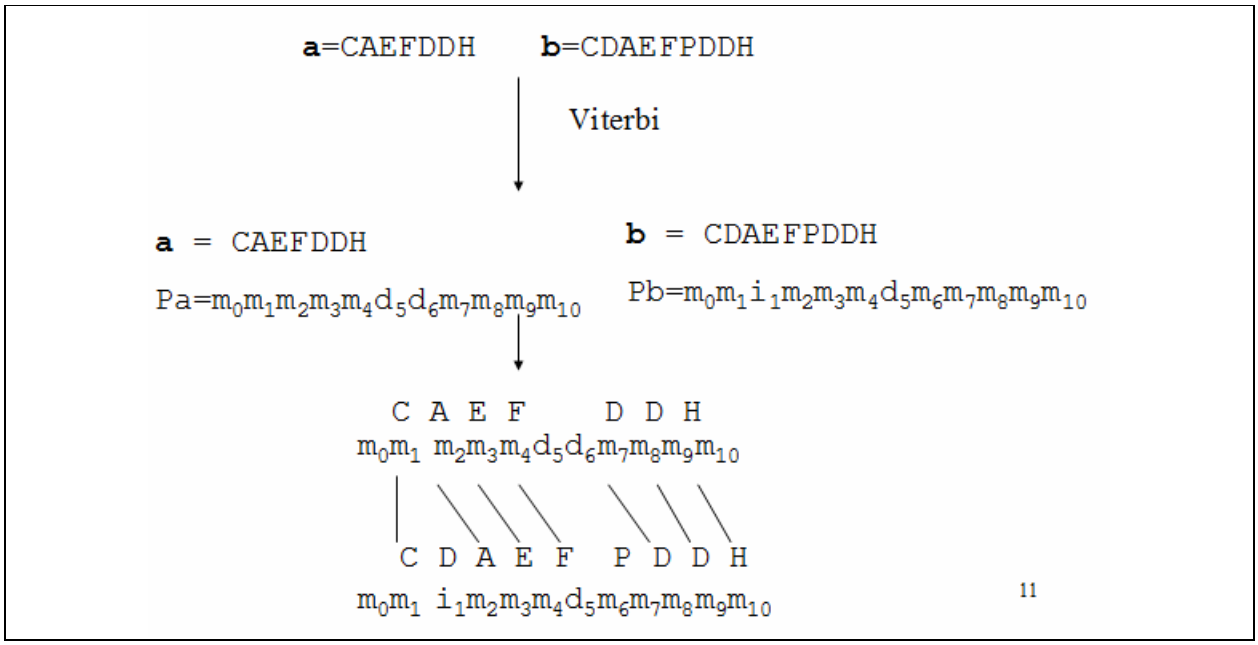


Εικόνα 8.17: *Σχηματική αναπαράσταση της διαδικασίας χαρακτηρισμού μιας νέας πρωτεϊνικής οικογένειας. Για λεπτομέρειες, δείτε το κείμενο.*

Η χρησιμότητα των pHMM, ξεκινάει από δημιουργία πολλαπλών στοιχίσεων, οι οποίες πολλές φορές είναι εφάμιλλες αντίστοιχων δομικών πολλαπλών στοιχίσεων (Eddy, 1995) και φτάνει στη δημιουργία μοντέλων μέσω των οποίων μπορεί να γίνει ευαίσθητη αναζήτηση απομακρυσμένων ομολογιών (Krogh, Brown, Mian, Sjolander, & Haussler, 1994). Ιδιαίτερα με την τελευταία μέθοδο και τη δημιουργία όλο και περισσότερων μοντέλων για την ταξινόμηση πρωτεϊνικών οικογενειών, έχουν κατασκευαστεί ειδικές βάσεις δεδομένων όπως η PFAM (Bateman, et al., 2004), μέσω της οποίας ταξινομείται μεγάλο μέρος των άγνωστης λειτουργίας πρωτεϊνών που προσδιορίζονται καθημερινά, π.χ. με την αποκωδικοποίηση των γονιδιωμάτων. Παρόλο που υπάρχουν και άλλες αντίστοιχες βάσεις, όπως για παράδειγμα η TIGRFAM, η PFAM θεωρείται σήμερα η κορυφαία βάση δεδομένων για πρωτεϊνικές οικογένειες. Αυτό οφείλεται, τόσο στο σχεδιασμό της, βάσει του οποίου μία πρωτεϊνική περιοχή μπορεί να ανήκει μόνο σε μία οικογένεια, όσο και στο λογισμικό το οποίο χρησιμοποιεί, το οποίο είναι ιδιαίτερα αποδοτικό και θα περιγραφεί στην επόμενη ενότητα. Τέλος, πρέπει να σημειωθεί ότι μια σειρά θεωρητικές και αλγοριθμικές βελτιώσεις που έχουν προκύψει τα τελευταία

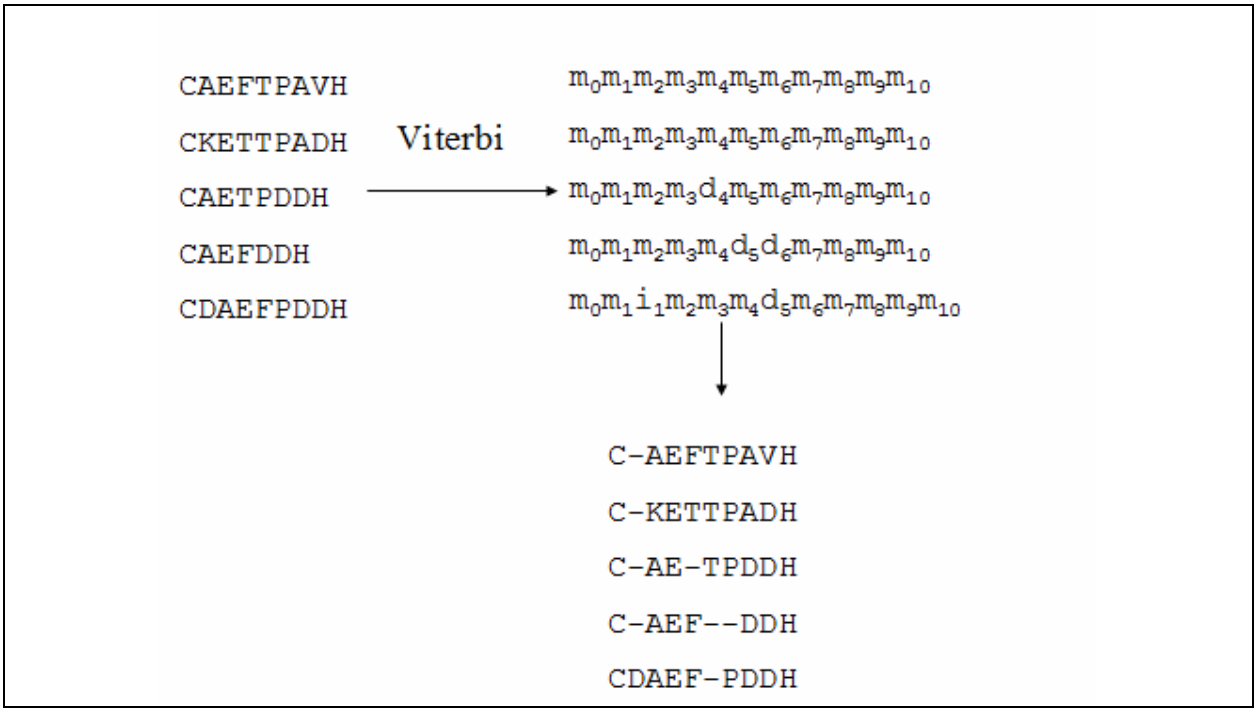
χρόνια, έχουν επιτρέψει την υλοποίηση αλγορίθμων που επιτελούν ακόμα και κλασικές αναζητήσεις ομοιότητας σε μια βάση δεδομένων, με ένα profile HMM (Eddy, 2011). Γενικά μια διαδικασία χαρακτηρισμού μιας πρωτεϊνικής οικογένειας έχει τα εξής βήματα:

- Στην αρχή, ξεκινάμε με μια αλληλουχία για την οποία υπάρχουν πειραματικές ενδείξεις για τη λειτουργία ή τη δομή της
- Γίνεται αναζήτηση σε βάσεις δεδομένων (BLAST, PSI-BLAST ή πλέον, με το HMMER)
- Συλλογή ομολόγων, επιλογή και ξεσκαρτάρισμα
- Γίνεται μια πολλαπλή στοίχιση (μπορεί και τροποποίηση αυτής με το χέρι)
- Ανάλογα με την περίπτωση, πραγματοποιούνται προγνώσεις (δευτεροταγούς δομής, διαμεμβρανικών τμημάτων ή οποιουδήποτε άλλου χρήσιμου χαρακτηριστικού)
- Γίνεται κατασκευή HMM και αξιολόγηση του (HMMER)
- Αναζήτηση εκ νέου σε βάσεις δεδομένων, μέχρι να μην προκύπτουν νέα μέλη της οικογένειας.



Εικόνα 8.18: *Στοίχιση δύο αλληλουχιών με ένα profile HMM.*

Ειδικά η περίπτωση της πολλαπλής στοίχισης αλληλουχιών με ένα profile HMM, είναι μια σημαντική διαδικασία, καθώς όπως είδαμε στο αντίστοιχο κεφάλαιο, αλγόριθμοι δυναμικού προγραμματισμού είναι δύσκολο να υλοποιηθούν, και στην πράξη όλα τα αντίστοιχα προγράμματα ακόμα και τα πιο πετυχημένα, βασίζονται σε ευριστικούς αλγόριθμους. Κατά συνέπεια, το profile HMM είναι μια ιδιαίτερα αξιόπιστη εναλλακτική, η οποία μάλιστα έχει αποδειχθεί ότι λειτουργεί εξαιρετικά καλά (Eddy, 1995). Η βασική αρχή της μεθόδου, φαίνεται στις Εικόνες 8.18 και 8.19. Το βασικό χαρακτηριστικό της μεθόδου, είναι ότι κάθε αλληλουχία στοιχίζεται με το μοντέλο, ανεξάρτητα από τις υπόλοιπες, και κατά συνέπεια, υπάρχει μόνο γραμμική εξάρτηση από τον αριθμό των αλληλουχιών. Για κάθε αλληλουχία, ο αλγόριθμος Viterbi βρίσκει το μονοπάτι των καταστάσεων από τις οποίες έχει περάσει. Αυτές οι καταστάσεις, όπως είδαμε είναι για κάθε θέση *i* στην αλληλουχία, τριών ειδών (*m*, *d*, *i*), ταύτιση, απαλοιφή και εισαγωγή. Επειδή όμως οι καταστάσεις αντιστοιχούν στη θέση *i* με αυτόν τον τρόπο μπορούμε άμεσα να στοιχίσουμε τις αλληλουχίες, αντιστοιχίζοντας απλά τις ταυτίσεις σε κάθε θέση (m_i).



Εικόνα 8.19: *Πολλαπλή στοίχιση αλληλουχιών με ένα profile HMM.*

8.7. Το πακέτο λογισμικού HMMER

Το πιο γνωστό πακέτο λογισμικού για κατασκευή Profile Hidden Markov Models, είναι το **HMMER** (Eddy, 2000). Το πακέτο αυτό, είναι μια συλλογή προγραμμάτων, διανεμόμενων ελεύθερα με την άδεια 'ανοικτού κώδικα' GPL (GNU Public License), η οποία επιτρέπει ελεύθερη πρόσβαση στον πηγαίο κώδικα, και το οποίο έχει αποδειχθεί το καλύτερο ίσως πακέτο τέτοιου είδους με πολλές εφαρμογές σε μεγάλο εύρος βιολογικών δεδομένων.

Το HMMER, στην έκδοση 3, περιέχει μεταξύ άλλων τα παρακάτω προγράμματα:

- **hmmbuild:** Πρόγραμμα με χρήση του οποίου, ξεκινώντας από μια αρχική πολλαπλή στοίχιση, κατασκευάζεται ένα μοντέλο HMM το οποίο να την περιγράφει.
- **hmmalign:** Πρόγραμμα με το οποίο μια σειρά αλληλουχιών οι οποίες προέρχονται από ένα HMM, στοιχίζονται σε μια πολλαπλή στοίχιση. Η πολλαπλή στοίχιση, επιτυγχάνεται μέσω διαδοχικών στοιχίσεων των αλληλουχιών με το μοντέλο.
- **hmmsearch:** Πρόγραμμα το οποίο, πραγματοποιεί αναζητήσεις ενός μοντέλου HMM έναντι μιας βάσης αλληλουχιών πρωτεϊνών.
- **phmmer:** Πρόγραμμα το οποίο πραγματοποιεί αναζήτηση μια πρωτεϊνικής αλληλουχίας έναντι μιας βάσης δεδομένων πρωτεϊνών (ανάλογο με το BLASTP)
- **jackhmmer:** Πρόγραμμα το οποίο πραγματοποιεί επαναληπτικές αναζητήσεις μια πρωτεϊνικής αλληλουχίας έναντι μιας βάσης δεδομένων πρωτεϊνών (ανάλογο με το PSI-BLAST)
- **hmmscan:** Πρόγραμμα με το οποίο πραγματοποιούνται αναζητήσεις μιας η περισσότερων αλληλουχιών έναντι μιας βάσης δεδομένων από μοντέλα HMM. Πρέπει να τονιστεί εδώ, ότι αν έχουμε μια αλληλουχία και ένα HMM, τα δυο παραπάνω προγράμματα επιστρέφουν ακριβώς το ίδιο αποτέλεσμα. Αν διαφέρουν, είτε οι αλληλουχίες είτε τα μοντέλα, τότε δίνουν άλλο αποτέλεσμα, λόγω του διαφορετικού τρόπου υπολογισμού της στατιστικής σημαντικότητας.

- **nhmmer:** Πρόγραμμα που πραγματοποιεί αναζήτηση μιας αλληλουχίας DNA, μιας στοίχισης ή ενός pHMM, έναντι μιας βάσης αλληλουχιών DNA. (ανάλογο με το BLASTN)
- **nhmmscan:** Πρόγραμμα που πραγματοποιεί αναζήτηση μιας αλληλουχίας DNA έναντι μιας βάσης δεδομένων από DNA profile HMM.
- **hmmconvert:** Πρόγραμμα που μετατρέπει μοντέλα HMM από και προς τη μορφή του HMMER3.
- **hmmemit:** Πρόγραμμα, με το οποίο 'εκπέμπεται' η καλύτερη (ανάλογα με τον ορισμό) αλληλουχία η οποία θα μπορούσε να παραχθεί από το μοντέλο.
- **hmmpress:** Μετατρέπει μια βάση δεδομένων HMM σε δυαδικό κώδικα για το hmmscan.
- **hmmstat:** Δείχνει συνοπτικά στατιστικά για μια βάση δεδομένων HMM.

Τα παραπάνω προγράμματα, περιέχουν μια σειρά από βελτιστοποιήσεις με σκοπό την επιτάχυνση των διαδικασιών. Για παράδειγμα, υπάρχουν βελτιστοποιήσεις για την ταχύτητα κατά την εκτέλεση των αλγορίθμων με τον μη υπολογισμό των προηγούμενων καταστάσεων, βελτιστοποιήσεις στον υπολογισμό των κατανομών συχνοτήτων των αμινοξέων του μηδενικού (null) μοντέλου με την εισαγωγή μίξεων από εκ των προτέρων κατανομές, διαφορικό στάθμισμα των αλληλουχιών με διαφορετικό βαθμό ομοιότητας, και μια σειρά από βελτιστοποιήσεις στη δομή του μοντέλου. Το τελευταίο, είναι πολύ σημαντικό, καθώς με αυτή τη διαδικασία, ο χρήστης δεν ασχολείται καθόλου με τη δομή και το μέγεθος του μοντέλου. Με επαναλήψεις ενός βασικού μοτίβου το οποίο αποτελεί παραλλαγή του κλασικού μοντέλου, επιτυγχάνεται, αναλόγως και του μήκους της πολλαπλής στοίχισης, η τελική διαμόρφωση του μοντέλου. Το παραπάνω μοντέλο, διαφέρει από την τυπική έκδοση του Profile Hidden Markov Model, που είδαμε παραπάνω, στο ότι δεν επιτρέπει μεταβάσεις από μια κατάσταση εισαγωγής κενού (I) σε κατάσταση απαλοιφής (D) και το αντίστροφο. Το πρόγραμμα, είναι διαθέσιμο στην ηλεκτρονική διεύθυνση: http://hmmer.janelia.org/.

Ιστορικά, αξίζει να αναφερθεί ότι υπάρχουν μεγάλες διαφορές μεταξύ των εκδόσεων του HMMER. Οι εκδόσεις μέχρι τη 1.8 επέτρεπαν τη δημιουργία μοντέλου ακόμα και από μη στοιχισμένες αλληλουχίες. Από την έκδοση 2.0 και μετά, η ύπαρξη μιας πολλαπλής στοίχισης έγινε απαραίτητη, καθώς το λογισμικό εξειδικεύτηκε σε αναλύσεις πρωτεϊνών. Παρ' όλα αυτά, το λογισμικό ήταν εξαιρετικά πετυχημένο και χρησιμοποιήθηκε για χιλιάδες αναλύσεις. Σε μερικές μάλιστα περιπτώσεις, χρησιμοποιήθηκε και για αναλύσεις που φαινομενικά δεν προϋπέθεταν κάποια «ομοιότητα» στις υπό μελέτη αλληλουχίες (Zhang & Wood, 2003). Εκτός αυτού, υπήρξαν και βελτιώσεις του, από τρίτους επιστήμονες, έτσι ώστε να μπορεί να πραγματοποιεί «διαχωριστική εκπαίδευση» (discriminative training) (Srivastava, Desai, Nandi, & Lynn, 2007). Από την έκδοση 3.0 και μετά, το λογισμικό, βασισμένο σε μια σειρά αλγοριθμικές και θεωρητικές βελτιώσεις, έγινε πολύ καλύτερο (Eddy, 2011). Καταρχάς, έγινε πολύ πιο γρήγορο. Επίσης, δεν χρειάζεται πλέον να πραγματοποιούνται προσομοιώσεις για να υπολογιστούν οι παράμετροι της κατανομής του Gumbel, αλλά αυτές προκύπτουν θεωρητικά. Τέλος, βασισμένο εν μέρει και στα προηγούμενα, το λογισμικό μπορεί πλέον να επιτελέσει και αναζητήσεις μιας αλληλουχίας έναντι μιας βάσης δεδομένων αλληλουχιών, -όπως ακριβώς το BLAST και το PSI-BLAST-, τις οποίες πραγματοποιεί καλύτερα και μάλιστα σε συγκρίσιμο χρόνο. Με αυτόν τον τρόπο, φαίνεται πως αργά αλλά σταθερά οδηγούμαστε σε καθολική αποδοχή των πιθανοθεωρητικών μοντέλων, τα οποία θα αντικαταστήσουν τις ευριστικές τεχνικές ομοιότητας.

Βιβλιογραφία

Audic, S., & Claverie, J. M. (1998). Self-identification of protein-coding regions in microbial genomes. *Proc Natl Acad Sci U S A, 95*(17), 10026-10031.

Bagos, P. G., Liakopoulos, T. D., & Hamodrakas, S. J. (2004). Faster Gradient Descent Conditional Maximum Likelihood Training of Hidden Markov Models, Using Individual Learning Rate Adaptation. In G. Paliouras & Y. Sakakibara (Eds.), *Grammatical Inference: Algorithms and Applications* (Vol. 3264, pp. 40-52): Spinger Berlin/Heidelberg.

Bagos, P. G., Liakopoulos, T. D., & Hamodrakas, S. J. (2006). Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins. *BMC Bioinformatics, 7*, 189.

Baldi, P., & Chauvin, Y. (1994). Smooth On-Line Learning Algorithms for Hidden Markov Models. *Neural Comput, 6*(2), 305-316.

Barash, Y., Elidan, G., Friedman, N., & Kaplan, T. (2003). *Modeling dependencies in protein-DNA binding sites*. Paper presented at the Proceedings of the seventh annual international conference on Computational molecular biology. RECOMB '03., New York, NY, USA.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., . . . Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Res, 32*(Database issue), D138-141.

Baum, L. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities, 3*, 1-8.

Bejerano, G. (2004). Algorithms for variable length Markov chain modeling. *Bioinformatics, 20*(5), 788-789.

Bejerano, G., Seldin, Y., Margalit, H., & Tishby, N. (2001). Markovian domain fingerprinting: statistical segmentation of protein sequences. *Bioinformatics, 17*(10), 927-934.

Bejerano, G., & Yona, G. (2001). Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics, 17*(1), 23-43.

Berchtold, A. (2001). Estimation in the Mixture Transition Distribution Model. *Journal of Time Series Analysis, 22*(4), 379-397.

Borodovsky, M., & McIninch, J. (1993). GeneMark: parallel gene recognition for both DNA strands. *Comput Chem, 17*(19), 123-133.

Borodovsky, M., & Peresetsky, A. (1994). Deriving non-homogeneous DNA Markov chain models by cluster analysis algorithm minimizing multiple alignment entropy. *Comput Chem, 18*(3), 259-267.

Browning, S. R. (2006). Multilocus association mapping using variable-length Markov chains. *Am J Hum Genet, 78*(6), 903-913.

Dalevi, D., Dubhashi, D., & Hermansson, M. (2006). Bayesian classifiers for detecting HGT using fixed and variable order markov models of genomic signatures. *Bioinformatics, 22*(5), 517-522.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc B, 39*, 1-38.

Drew, D., Sjostrand, D., Nilsson, J., Urbig, T., Chin, C. N., de Gier, J. W., & von Heijne, G. (2002). Rapid topology mapping of Escherichia coli inner-membrane proteins by prediction and PhoA/GFP fusion analysis. *Proc Natl Acad Sci U S A, 99*(5), 2690-2695.

Durbin, R., Eddy, S. R., Krogh, A., & Mithison, G. (1998). *Biological sequence analysis, probabilistic models of proteins and nucleic acids.*: Cambridge University Press.

Eddy, S. R. (1995). Multiple alignment using hidden Markov models. *Proc Int Conf Intell Syst Mol Biol, 3*, 114-120.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics, 14*(9), 755-763.

Eddy, S. R. (2000). *HMMER: profile hidden Markov models for biological sequence analysis*. St Louis, MO: Washington University school of medicine.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol, 7*(10), e1002195.

Ellrott, K., Yang, C., Sladek, F. M., & Jiang, T. (2002). Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics, 18 Suppl 2*, S100-S109.

Eronen, L., Geerts, F., & Toivonen, H. (2004). A Markov chain approach to reconstruction of long haplotypes. *Pac Symp Biocomput*, 104-115.

Fariselli, P., Finelli, M., Marchignoli, D., Martelli, P. L., Rossi, I., & Casadio, R. (2003). MaxSubSeq: an algorithm for segment-length optimization. The case study of the transmembrane spanning segments. *Bioinformatics, 19*(4), 500-505.

Fariselli, P., Martelli, P. L., & Casadio, R. (2005). A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. *BMC Bioinformatics, 6 Suppl 4*, S12.

Gribskov, M., Luthy, R., & Eisenberg, D. (1990). Profile analysis. *Methods Enzymol, 183*, 146-159.

Gribskov, M., McLachlan, A. D., & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A, 84*(13), 4355-4358.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry, 33*(10), 3038-3049.

Juang, B. H., & Rabiner, L. R. (1990). The Segmental K-Means Algorithm for Estimating Parameters of Hidden Markov Models. *IEEE Transactions on Acoustics, Speech and Signal Processing, 38*(9), 1639-1641.

Kall, L., Krogh, A., & Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *J Mol Biol, 338*(5), 1027-1036.

Kall, L., Krogh, A., & Sonnhammer, E. L. (2005). An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics, 21 Suppl 1*, i251-i257.

Kim, H., Melen, K., & von Heijne, G. (2003). Topology models for 37 Saccharomyces cerevisiae membrane proteins based on C-terminal reporter fusions and predictions. *J Biol Chem, 278*(12), 10208-10213.

Krogh, A. (1994). Hidden Markov models for labelled sequences. *Proceedings of the12th IAPR International Conference on Pattern Recognition*, 140-144.

Krogh, A. (1997). Two methods for improving performance of an HMM and their application for gene finding. *Proc Int Conf Intell Syst Mol Biol, 5*, 179-186.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol, 235*(5), 1501-1531.

Krogh, A., & Riis, S. K. (1999). Hidden neural networks. *Neural Comput, 11*(2), 541-563.

Lebre, S., & Bourguignon, P. Y. (2008). An EM algorithm for estimation in the mixture transition distribution model. *Journal of Statistical Computation and Simulation, 78*(8), 713-729.

MacQueen, B. (1967). *Some Methods for classification and Analysis of Multivariate Observations*. Paper presented at the Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability.

Markov, A. A. (1913). An example of statistical study on text of Eugeny Onegin illustrating the linking of events to a chain. *Izvestija Imp. Akad. nauk, 6*(3), 153-162.

Melen, K., Krogh, A., & von Heijne, G. (2003). Reliability measures for membrane protein topology prediction algorithms. *J Mol Biol, 327*(3), 735-744.

Ohler, U., Harbeck, S., Niemann, H., Noth, E., & Reese, M. G. (1999). Interpolated markov chains for eukaryotic promoter recognition. *Bioinformatics, 15*(5), 362-369.

Ostendorf, M., & Singer, H. (1997). HMM topology design using maximum likelihood successive state splitting. *Computer Speech & Language, 11*(1), 17-41.

Phillips, G. J., Arnold, J., & Ivarie, R. (1987). Mono- through hexanucleotide composition of the Escherichia coli genome: a Markov chain analysis. *Nucleic Acids Res, 15*(6), 2611-2626.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE, 77*(2), 257-286.

Raftery, A. E. (1985a). A model for high-order Markov chains. *Journal of the Royal Statistical Society. Series B (Methodological), 47*(3), 528-539.

Raftery, A. E. (1985b). A new model for discrete-valued time series: Autocorrelations and extensions. *Rassegna di Metodi Statistici ed Applicazioni, 3-4* 149-162.

Rapp, M., Drew, D., Daley, D. O., Nilsson, J., Carvalho, T., Melen, K., . . . Von Heijne, G. (2004). Experimentally based topology models for E. coli inner membrane proteins. *Protein Sci, 13*(4), 937-945.

Ron, D., Singer, Y., & Tishby, N. (1996). The power of amnesia: learning probabilistic automata with variable memory length. *Machine Learning, 25*, 117-149.

Salzberg, S. L., Delcher, A. L., Kasif, S., & White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res, 26*(2), 544-548.

Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J., & Tettelin, H. (1999). Interpolated Markov models for eukaryotic gene finding. *Genomics, 59*(1), 24-31.

Schwartz, R., & Chow, Y. L. (1990). The N-Best Algorithm: An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses. *Proc IEEE Int Conf Acoust, Speech, Sig Proc, 1*, 81-84.

Srivastava, P. K., Desai, D. K., Nandi, S., & Lynn, A. M. (2007). HMM-ModE--improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences. *BMC Bioinformatics, 8*, 104.

Tusnady, G. E., & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics, 17*(9), 849-850.

Vasko, R. C. J., El-Jaroudi, A., & Boston, J. R. (1996). *An algorithm to determine hidden Markov model topology.* Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96.

Won, K. J., Prugel-Bennett, A., & Krogh, A. (2004). Training HMM structure with genetic algorithm for biological sequence analysis. *Bioinformatics, 20*(18), 3613-3619.

Yada, T., Ishikawa, M., H., T., & Asai, K. (1994). DNA Sequence Analysis using Hidden Markov Model and Genetic Algorithm. *Genome Informatics, 5*, 178-179.

Yuan, Z. (1999). Prediction of protein subcellular locations using Markov chain models. *FEBS Lett, 451*(1), 23-26.

Zhang, Z., & Wood, W. I. (2003). A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics, 19*(2), 307-308.

Ερωτήσεις

1) Στο μοντέλο με το «μεροληπτικό ζάρι» που αναφέρεται στην παράγραφο 8.3.1, δίνεται η ακολουθία συμβόλων **x**= `21452643663656166623214 5` και το μονοπάτι π= `--------++++++++++++++-------`. Εκτιμήστε τις παραμέτρους του μοντέλου με βάση τις σχέσεις (8.34) και (8.35).

2) Σε ένα Hidden Markov Model (HMM) δίνονται ο πίνακας των πιθανοτήτων μετάβασης (*a*, transitions), και αυτός των πιθανοτήτων εμφάνισης των συμβόλων (*e*, emissions), αντίστοιχα:

$$a=\begin{bmatrix} 0.7 & 0.3 & 0 & x_1 \\ 0 & 0 & 0.8 & x_2 \\ 0 & 0 & 0.9 & 0.1 \\ x_3 & 0 & 0 & 0 \end{bmatrix},\ e=\begin{bmatrix} 0.2 & 0.1 & x_4 & 0.5 \\ 0.5 & x_5 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.1 & 0.6 \\ 0.4 & 0.4 & 0.1 & 0.1 \end{bmatrix}$$

Α) Από πόσες καταστάσεις αποτελείται αυτό το μοντέλο και πόσα είναι τα γράμματα του αλφαβήτου του; Εξηγήστε.
Β) Υπολογίστε τα x_1, x_2, x_3, x_4 και x_5. Εξηγείστε.
Γ) Απεικονίστε γραφικά το παραπάνω μοντέλο. Το μοντέλο αυτό είναι γραμμικό ή κυκλικό;
Δ) Πως θα τροποποιηθεί το μοντέλο αν προστεθούν καταστάσεις έναρξης (Begin) και τερματισμού (End);

3) Η κανονική έκφραση (regular expression) που περιγράφει την περιοχή συρραφής (splicing) σε ευκαρυωτικά γονίδια είναι η ακόλουθη:
[AC]AGGT[AG]AGT
1 2 3 4 5 6 7 8 9,
όπου οι θέσεις αριθμούνται με 1-9 και η αποκοπή γίνεται μεταξύ των θέσεων 3 και 4.
Α) Κατασκευάστε (και σχεδιάστε) ένα Hidden Markov Model, το οποίο να είναι *εντελώς ανάλογο* με την παραπάνω κανονική έκφραση.
Β) Αν σας δοθεί η επιπλέον πληροφορία ότι στη θέση 1, η Αδενίνη (A) εμφανίζεται με ποσοστό 10%, ενώ στη θέση 6 η πιθανότητα εμφάνισης της Γουανίνης (G) είναι τριπλάσια από αυτή της Αδενίνης (A), τροποποιήστε κατάλληλα το μοντέλο.
Γ) Δίνονται δύο αλληλουχίες DNA
x: AAACAGGTGAGTAAA
y: TTTAAGGTAAGTGGG
Ποια από τις δυο έχει μεγαλύτερες πιθανότητες να εμφανιστεί κάτω από τις προϋποθέσεις του μοντέλου, σύμφωνα με τον αλγόριθμο Forward? Εξηγήστε *ποιοτικά* τα αποτελέσματα.
Δ) Ποια είναι τα πλεονεκτήματα του HMM που κατασκευάσατε στο ερώτημα (Β), και κάποιων ακόμα καλύτερων που μπορεί να κατασκευαστούν υπό το φως περισσότερων δεδομένων (αλληλουχιών), σε σχέση με το απλό regular expression?

Σημείωση: για όλα τα παραπάνω θεωρήστε ότι σε τυχαίες περιοχές (εκτός της περιοχής συρραφής) οι πιθανότητες εμφάνισης κάθε βάσης (A, C, T, G) είναι ίσες.

4) Δίνεται η παρακάτω πρωτεϊνική αλληλουχία και η δευτεροταγής δομή της:

```
1     GSAPSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPTAYI
      CCCCCCEEEEEEEECCCCCHHHHHHHHHHHHCCCCCCEEEEEEECCCCCH
51    DFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSER
      HHHHHHCCCCEEEEEEEECCCCCCCCCCCHHHHHHHCCCCEEEEECHHHH
101   RHVFGESDELIGQKVAHALAEGLG
      HCCCCCCHHHHHHHHHHHHCCCC
```

Α) Σε ποια κατηγορία πρωτεϊνικού διπλώματος πιστεύετε ότι κατατάσσεται η εν λόγω πρωτεΐνη στη βάση δεδομένων SCOP και γιατί?
Β) Κατασκευάστε ένα όσο το δυνατό πιο απλό Hidden Markov Model το οποίο να προβλέπει τη

δευτεροταγή δομή πρωτεϊνών σε 3 καταστάσεις (H, E, C) και εξηγήστε. Πώς θα μπορούσε το μοντέλο αυτό να γίνει καλύτερο?
Γ) Αν θεωρήσετε σαν δεδομένα εκπαίδευσης την παραπάνω αλληλουχία και τη δομή της, να υπολογίσετε τις παραμέτρους του εν λόγω μοντέλου (πιθανότητες μετάβασης και πιθανότητες εμφάνισης συμβόλων).

Σημείωση: για λόγους απλότητας θεωρήστε ότι στο μοντέλο δεν υπάρχουν καταστάσεις έναρξης και τερματισμού

5) Δίνεται το παρακάτω τμήμα μιας αλληλουχίας DNA η οποία προέρχεται από ένα γονίδιο με τον αντίστοιχο σχολιασμό (E=εξόνιο, I=εσώνιο, N=μη-κωδική περιοχή εκτός γονιδίου, 5=5'-άκρο συρραφής, 3=3'-άκρο συρραφής)

```
...AGCATATGCGTAGTCTGACTGCCAAGATATAAAGTTATAATCTATATACGATCGCTGTCAATGCT...
...NNNNNEEEEEEEEEEEEEEEEEEE5IIIIIIIIIIIIIIIIIIII3EEEEEEEEEEEEEE...
```

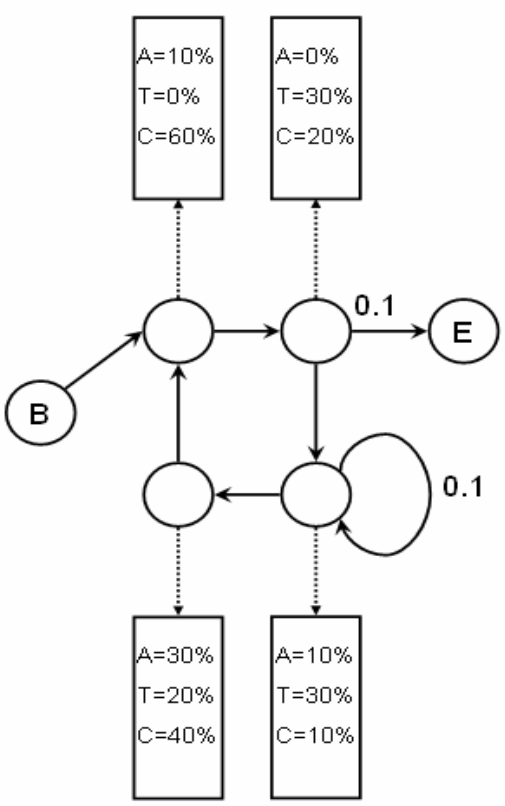
Α) Από τι είδους οργανισμό (ευκαρυωτικό ή προκαρυωτικό) προέρχεται το παραπάνω γονίδιο; Εξηγήστε την απάντησή σας.
Β) Κατασκευάστε ένα όσο το δυνατό πιο απλό Hidden Markov Model το οποίο να προβλέπει την ύπαρξη γονιδίων σε 5 καταστάσεις (E, I, 5, 3, N) και εξηγήστε. Με ποιους τρόπους θα μπορούσε το μοντέλο αυτό να γίνει καλύτερο;
Γ) Αν θεωρήσετε σαν δεδομένα εκπαίδευσης την παραπάνω αλληλουχία, να υπολογίσετε τις παραμέτρους του εν λόγω μοντέλου (πιθανότητες μετάβασης και πιθανότητες εμφάνισης συμβόλων).

Σημείωση: Για λόγους απλότητας θεωρήστε ότι στο μοντέλο δεν υπάρχουν καταστάσεις έναρξης και τερματισμού. Για όσες παραμέτρους δεν έχετε στοιχεία, βάλτε κάποια υποθετικά νούμερα.

6) Δίνεται το παρακάτω σχήμα στο οποίο απεικονίζεται η γραφική αναπαράσταση ενός HMM.
Α) Περιγράψτε το μοντέλο αναλυτικά (Πίνακες μεταβάσεων και εμφανίσεων συμβόλων, αλφάβητο, καταστάσεις).
Β) Δίνονται οι παρακάτω αλληλουχίες:

$\mathbf{x}_1$=CGGCCG
$\mathbf{x}_2$=AT
$\mathbf{x}_3$=ACCGAT
$\mathbf{x}_4$=TACGAT
$\mathbf{x}_5$=GGCCG

Υπολογίστε τη συνολική πιθανότητα κάθε μιας με βάση το παρακάτω μοντέλο. Ποιες από τις αλληλουχίες δεν μπορούν να παραχθούν από το μοντέλο αυτό?
Γ) Πως θα τροποποιούσαμε το μοντέλο έτσι ώστε να μπορεί να παράγει και τις αλληλουχίες του ερωτήματος Β?



7) Θεωρήστε ένα όσο το δυνατόν πιο απλό HMM, το οποίο να περιέχει 2 καταστάσεις και 2 σύμβολα. Σχεδιάστε το μοντέλο και δώστε μια γενική έκφραση για τους πίνακες *a* και *e*. Δείξτε ότι το ίδιο μοντέλο, μπορεί να αναπαρασταθεί με ένα κλασικό μαρκοβιανό μοντέλο, του οποίου τις παραμέτρους θα ορίσετε. Δοκιμάστε να αυξήσετε την πολυπλοκότητα του HMM, αυξάνοντας για παράδειγμα των αριθμό των καταστάσεων σε 3, 4, κ.ο.κ. Τι παρατηρείτε για το αντίστοιχο μαρκοβιανό μοντέλο;

8) Αποδείξτε ότι σε ένα HMM, ισχύει :

$$P(\pi_i = k \mid \mathbf{x}, \theta) = \frac{f_k(i) b_k(i)}{P(\mathbf{x}|\theta)}$$

Σημείωση: Ξεκινήστε από την πιθανότητα $P(\mathbf{x}, \pi_i = k \mid \theta)$ και θυμηθείτε ότι

$f_k(i) = P(x_1, x_2, ..., x_i, \pi_i = k \mid \theta)$

$b_k(i) = P(x_{i+1}, ..., x_L \mid \pi_i = k, \theta)$

9) Πώς πραγματοποιείται μια πολλαπλή στοίχιση με χρήση profile HMM?
Δίνονται οι αλληλουχίες:

x_a=WAYDDR, και

x_b=WDAYPDDR

και τα αντίστοιχα μονοπάτια (paths) από τον αλγόριθμο του Viterbi:

$p_a = m_0 m_1 m_2 m_3 d_4 d_5 m_6 m_7 m_8 m_9$, και $p_b = m_0 m_1 i_1 m_2 m_3 d_4 m_5 m_6 m_7 m_8 m_9$

Δώστε την στοίχιση των δυο αλληλουχιών και εξηγείστε.

Κεφάλαιο 9: Δομική Βιοπληροφορική

Σύνοψη
Δομική Βιοπληροφορική, είναι ο κλάδος της βιοπληροφορικής ο οποίος ασχολείται με την ανάλυση και την πρόγνωση της τρισδιάστατης δομής των βιολογικών μακρομορίων, όπως οι πρωτεΐνες, το DNA, και το RNA. Ασχολείται με όλα τα επίπεδα της ανάλυσης των τρισδιάστατων δομών, από την αναπαράσταση και την οπτικοποίηση, τις συγκρίσεις και τις ομαδοποιήσεις των δομών, τις μελέτες του πρωτεϊνικού διπλώματος, την κατασκευή μοντέλων, τη μελέτη των εξελικτικών σχέσεων έως και τη μελέτη της σχέσης δομής και λειτουργίας. Σαν κλάδος έχει ιδιαίτερες σχέσεις με τη μοριακή βιοφυσική και τη δομική βιολογία. Στο κεφάλαιο αυτό θα ασχοληθούμε με τις βασικές μεθοδολογίες της δομικής βιοπληροφορικής και θα δούμε τα πιο γνωστά πακέτα λογισμικού που χρησιμοποιούνται στον τομέα αυτό.

Προαπαιτούμενη γνώση
Στο κεφάλαιο αυτό απαραίτητη είναι η γνώση των εννοιών των κεφαλαίων που ασχολούνται με τη στοίχιση αλληλουχιών και την πολλαπλή στοίχιση, αλλά και τις βάσεις δεδομένων.

9. Εισαγωγή

Δομική Βιοπληροφορική, είναι ο κλάδος της βιοπληροφορικής ο οποίος ασχολείται με την ανάλυση και την πρόγνωση της τρισδιάστατης δομής των βιολογικών μακρομορίων, όπως οι πρωτεΐνες, το DNA και το RNA. Είναι ένας κλάδος που έχει τις ρίζες του στις πρώτες μεθοδολογίες προσδιορισμού της δομής των βιολογικών μακρομορίων από τις δεκαετίες του 1950 και 1960 και κατά συνέπεια είναι ένας κλάδος που αναπτύχθηκε όλα αυτά τα χρόνια, παράλληλα με την ανάπτυξη της μοριακής βιολογίας, της δομικής βιολογίας και της βιοφυσικής.

Το αντικείμενο της μελέτης της δομικής βιοπληροφορικής, είναι οι τρισδιάστατες δομές, δηλαδή οι συντεταγμένες των ατόμων ενός βιολογικού μακρομορίου. Η εύρεση της τρισδιάστατης δομής, είναι από μόνη της μια ιδιαίτερα επίπονη και κοστοβόρα διαδικασία, που αποτελεί περιοριστικό παράγοντα στον τομέα. Έτσι, είναι γνωστό ότι οι διαθέσιμες τρισδιάστατες δομές είναι μια τάξη μεγέθους λιγότερες από τις διαθέσιμες αλληλουχίες. Από την άλλη, η δομή είναι πολύ σημαντική στην κατανόηση της δράσης των βιολογικών μακρομορίων και ειδικά των πρωτεϊνών.

Γενικά, η δομική βιοπληροφορική ασχολείται με όλα τα επίπεδα της ανάλυσης των τρισδιάστατων δομών. Έτσι, έχουμε σε πρώτο επίπεδο τους αλγόριθμους και το λογισμικό που χρησιμοποιούνται για την αναπαράσταση και την οπτικοποίηση των βιολογικών δομών. Τα εργαλεία αυτά, εκτός από τους ειδικούς, χρησιμοποιούνται πλέον και από τον καθένα που κάνει μια εργασία που αναφέρεται σε βιολογικές δομές. Ένα άλλο επίπεδο είναι οι συγκρίσεις και οι ομαδοποιήσεις των δομών. Εδώ έχουμε το πρόβλημα της στοίχισης και υπέρθεσης δομών, αλλά και το πρόβλημα της αναγνώρισης του πρωτεϊνικού διπλώματος με όλες τις επιπτώσεις του. Όλα αυτά, οδηγούν τελικά στο βασικό πρόβλημα της πρόβλεψης της τρισδιάστατης δομής πρωτεϊνών, της κατασκευής δηλαδή μοντέλων για μια πρωτεΐνη για την οποία δεν υπάρχουν πειραματικά δεδομένα. Σε αυτή την περίπτωση, υπάρχει πληθώρα μεθόδων, από την απλή προτυποποίηση με βάση την ομολογία (homology modelling), μέχρι την ύφανση (threading) και την ab initio πρόγνωση της δομής. Τέλος, είτε με δομές πειραματικά προσδιορισμένες, είτε με δομές που έχουν προκύψει από μοντέλα, αντικείμενο της δομικής βιοπληροφορικής είναι η μελέτη τους με σκοπό την κατανόηση του πρωτεϊνικού διπλώματος και των μηχανισμών του, τη μελέτη των εξελικτικών σχέσεων, αλλά και τη μελέτη για τη σχέση δομής και λειτουργίας, οι οποίες οδηγούν στις μεθοδολογίες αγκυροβόλησης ή ελλιμενισμού (docking) για την κατανόηση της δομής συμπλόκων (της εύρεσης δηλαδή της αλληλεπίδρασης δυο πρωτεϊνών ή πρωτεΐνης/DNA (ή RNA), ή πρωτεΐνης/μικρού μορίου, μελέτες που είναι πολύ σημαντικές στο σχεδιασμό φαρμάκων). Φυσικά, ακόμα και στο αρχικό στάδιο του προσδιορισμού της τρισδιάστατης δομής, η συμβολή των υπολογιστικών μεθόδων είναι σημαντική, καθώς η επεξεργασία των δεδομένων της κρυσταλλογραφίας, η κατασκευή χαρτών ηλεκτρονικής πυκνότητας, η προσαρμογή αλλά και η βελτιστοποίηση του μοντέλου, γίνονται με αλγόριθμους και λογισμικό, αλλά λόγω της φύσης αυτού του εγχειριδίου, δεν θα υπεισέλθουμε σε πολλές λεπτομέρειες.

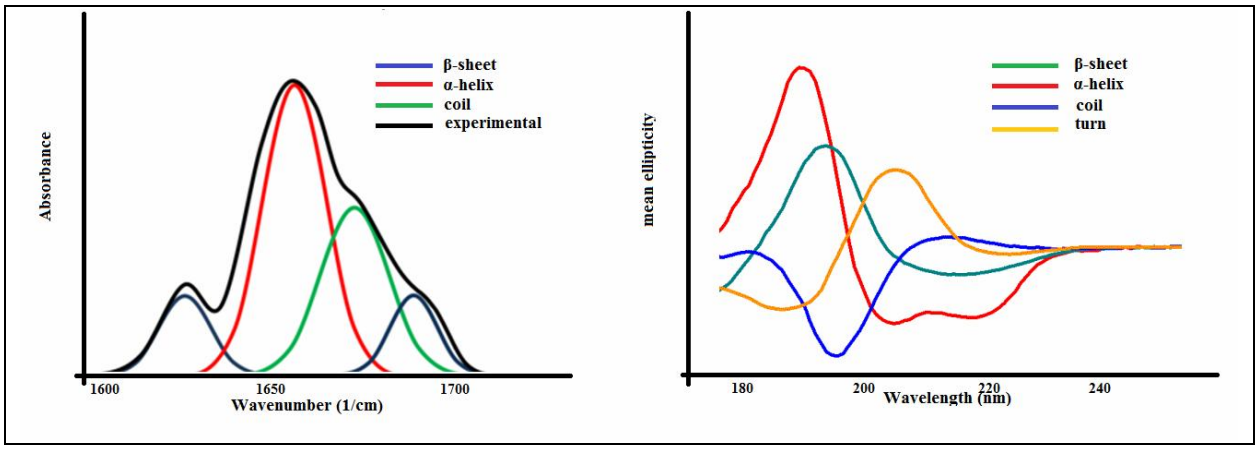
Στις επόμενες ενότητες θα προσπαθήσουμε να διερευνήσουμε τα βασικά μεθοδολογικά θέματα που προκύπτουν στα παραπάνω προβλήματα της δομικής βιοπληροφορικής, να παρουσιάσουμε τα βασικά

εργαλεία λογισμικού που χρησιμοποιούνται για την αντίστοιχη εργασία, αλλά και να προσφέρουμε πρακτικές συμβουλές για την εκτέλεση τέτοιων εργασιών.

9.1. Προσδιορισμός δομής

Το πρώτο βήμα σε κάθε προσπάθεια δομικής βιοπληροφορικής, είναι ο ίδιος ο προσδιορισμός της τρισδιάστατης δομής των μακρομορίων. Αυτές οι μεθοδολογίες ήταν που οδήγησαν τις δεκαετίες του 1950 και του 1960 στη ραγδαία ανάπτυξη της μοριακής βιολογίας, για αυτό και θα κάνουμε μια σύντομη αναφορά, αν και το αντικείμενο αυτό εμπίπτει περισσότερο στον τομέα της δομικής βιολογίας και της βιοφυσικής.

Γενικά, η κατά προσέγγιση σύσταση ενός πολυμερούς, π.χ. μιας πρωτεΐνης σε στοιχεία δευτεροταγούς δομής (π.χ. «η πρωτεΐνη Χ έχει περίπου 40% α-έλικα και 20% β-πτυχωτή επιφάνεια») είναι κάτι που μπορεί να υπολογιστεί με φασματοσκοπία. Για τις πρωτεΐνες, συνηθισμένες μεθοδολογίες φασματοσκοπίας είναι η φασματοσκοπία υπεριώδους (far-UV, 170–250 nm) με κυκλικό διχρωισμό, η φασματοσκοπία υπερύθρου (IR) και η φασματοσκοπία μαγνητικού πυρηνικού συντονισμού (NMR) (Meiler & Baker, 2003; Pelton & McLean, 2000). Σε όλες τις περιπτώσεις, τα διαφορετικά στοιχεία δευτεροταγούς δομής (α-έλικες, β-πτυχωτές επιφάνειες) δίνουν διαφορετικές καμπύλες απορρόφησης, από τις οποίες με υπολογιστική ανάλυση μπορούν να εξαχθούν τα δεδομένα για την εκατοστιαία σύσταση της υπό μελέτης πρωτεΐνης (πάντα με την προϋπόθεση ότι έχουμε καθαρό δείγμα σε κατάλληλο διαλύτη και σε κατάλληλη ποσότητα).



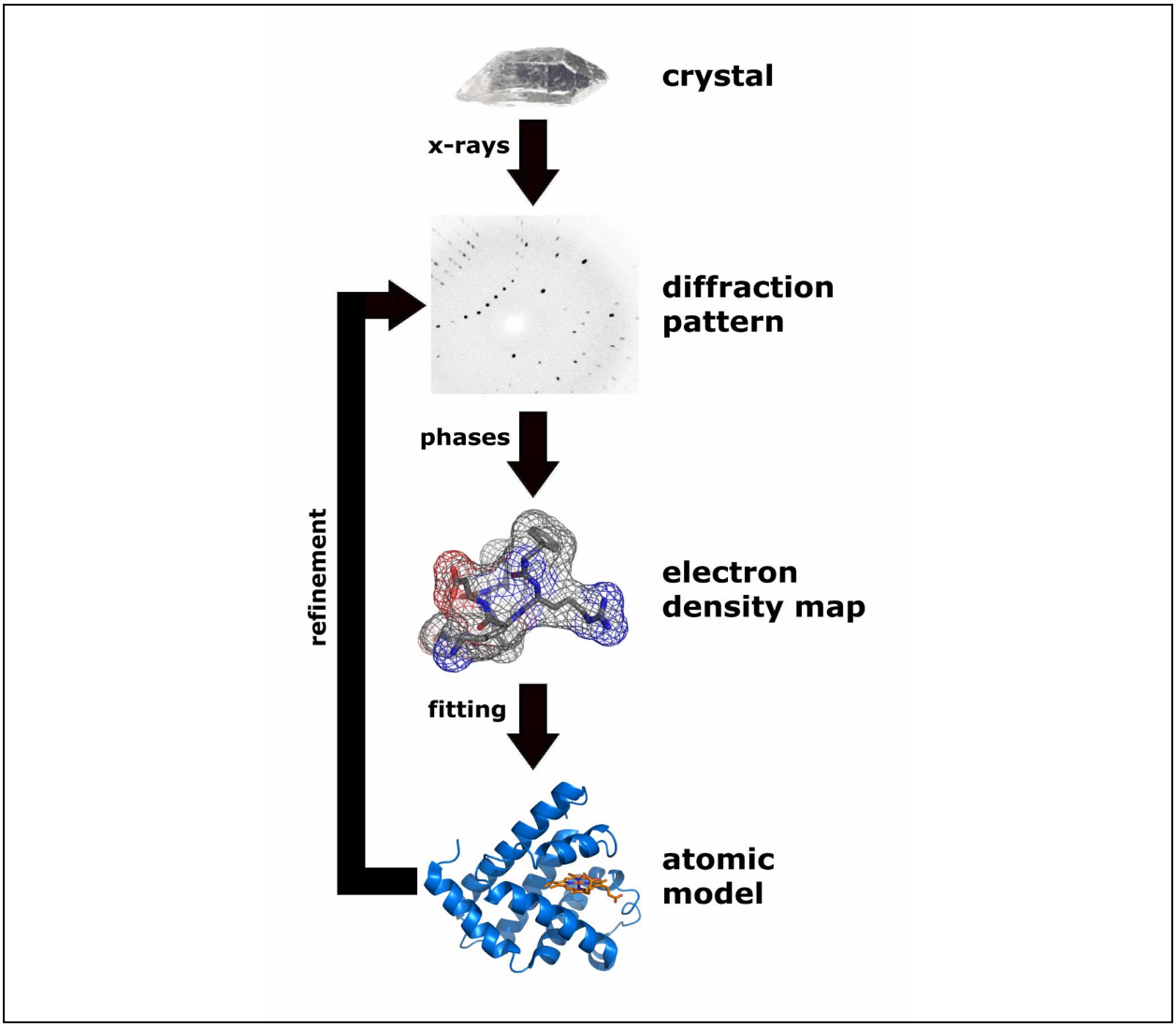
Εικόνα 9.1: *Εικόνα από τη χρήση φασματοσκοπίας για τη διαλεύκανση της δομής πρωτεϊνών (αριστερά φάσμα IR – δεξιά φάσμα CD)*

Η μεθοδολογία της κρυσταλλογραφίας ακτίνων Χ είχε αρχίσει να αναπτύσσεται από τις δεκαετίες πριν τον 2ο Παγκόσμιο Πόλεμο, αλλά οι πρώτες τρισδιάστατες δομές πρωτεϊνών (της μυογλοβίνης και της αιμοσφαιρίνης), επιλύθηκαν προς τα τέλη της δεκαετίας του 1950, λίγο μετά την εύρεση της δομής του DNA. Για τις εργασίες τους αυτές, ο Sir John Kendrew και ο Max Perutz μοιράστηκαν το βραβείο Νόμπελ το 1962, ενώ από τότε αρκετά άλλα βραβεία Νόμπελ έχουν δοθεί σε επιστήμονες που προσδιόρισαν δομές σημαντικών πρωτεϊνών. Η βασική αρχή της κρυσταλλογραφίας ακτίνων Χ βασίζεται στην ίδια αρχή με τους μεγεθυντικούς φακούς και το μικροσκόπιο. Η μεγάλη διαφορά με το οπτικό αλλά και το ηλεκτρονικό μικροσκόπιο, έγκειται στο γεγονός ότι επιθυμούμε να «δούμε» σε ατομική διακριτικότητα, δηλαδή να μπορούμε να ξεχωρίσουμε αντικείμενα με απόσταση λίγα Å. Αυτό σημαίνει ότι η προσπίπτουσα ακτινοβολία πρέπει να έχει ένα μήκος κύματος που την τοποθετεί στο φάσμα των ακτίνων Χ, αλλά οι ακτίνες Χ δεν μπορούν να εστιαστούν με τη χρήση φακών (όπως για παράδειγμα κάνουμε στο οπτικό μικροσκόπιο με τα φωτόνια ή στο ηλεκτρονικό μικροσκόπιο με τα ηλεκτρόνια). Έτσι, πρέπει να μελετήσουμε το πρότυπο περίθλασης των ακτίνων Χ από το δείγμα για να μπορέσουμε να καταλήξουμε σε ένα συμπέρασμα για τη δομή του δείγματος.

Από τη δεκαετία του 1990 και μετά, οι μέθοδοι ετερόλογης έκφραση πρωτεϊνών, η διαθεσιμότητα επιταχυντών ηλεκτρονίων για παραγωγή ακτίνων Χ (συγχροτρονική ακτινοβολία), οι ανιχνευτές περιοχής τύπου CCD αρχικά και απευθείας μέτρησης φωτονίων πιο πρόσφατα, αλλά και η αύξηση της υπολογιστικής ισχύος, οδήγησαν στην εκθετική αύξηση του αριθμού των διαθέσιμων τρισδιάστατων δομών. Σήμερα υπάρχουν πλέον διαθέσιμες πάνω από εκατό χιλιάδες διαθέσιμες δομές στην PDB (βέβαια, παρ' όλα αυτά οι

διαθέσιμες αλληλουχίες, είναι μια τάξης μεγέθους περισσότερες, οπότε το χάσμα ανάμεσα στον αριθμό των δομών και αυτόν των αλληλουχιών συνεχίζει να αυξάνει). Η κρυσταλλογραφία των ακτίνων Χ εξακολουθεί φυσικά να είναι η πιο κοινή μέθοδος εύρεσης τρισδιάστατης δομής, αν και την παρούσα χιλιετία η φασματοσκοπία NMR συμμετέχει σταθερά με περίπου 10% των νέων δομών (κυρίως μικρών πρωτεϊνών) που προσδιορίζονται πειραματικά. Την τελευταία διετία (2013-2015) συντελείται μια επανάσταση στο χώρο της ηλεκτρονικής μικροσκοπίας, που μπορεί πλέον να προσδιορίσει δομές σε ευκρίνεια 2-3Å, με τη χρήση νέων μικροσκοπίων αλλά κυρίως νέων τεχνολογιών ανιχνευτών περιοχής ηλεκτρονίων και μεγάλης υπολογιστικής ισχύος (εκατοντάδες επεξεργαστές, terabyte δεδομένων και δεκάδες gigabyte μνήμης).

Η παλιότερη αλλά και πιο ακριβής μέθοδος κρυσταλλογραφίας ακτίνων Χ είναι αυτή περίθλασης ακτίνων Χ μονοκρυστάλλου (single-crystal X-ray diffraction), στην οποία μία δέσμη από ακτίνες Χ προσκρούει στον κρύσταλλο και παράγει μια σειρά ανακλάσεις οι οποίες καταγράφονται από κάποιον ανιχνευτή. Η ένταση και η γωνία των ανακλάσεων καταγράφεται καθώς ο κρύσταλλος περιστρέφεται. Αν ο κρύσταλλος είναι επαρκούς καθαρότητας και κανονικότητας, τα δεδομένα από την περίθλαση επιτρέπουν τον προσδιορισμό των αποστάσεων των χημικών δεσμών και των γωνιών τους με μεγάλη ακρίβεια (Shi, 2014; Yaffe, 2005).



Εικόνα 9.2: *Σχηματική αναπαράσταση της διαδικασίας προσδιορισμού δομής με κρυσταλλογραφία ακτίνων Χ (https://en.wikipedia.org/wiki/X-ray_crystallography)*

Τα βασικά στάδια της κρυσταλλογραφίας είναι τρία:

- Το πρώτο και συχνά πιο δύσκολο στάδιο είναι η ανάπτυξη ενός κατάλληλου κρυστάλλου για την πρωτεΐνη ή το σύμπλοκο που μελετάται. Ο κρύσταλλος πρέπει να είναι αρκετά μεγάλος (τυπικά μεγαλύτερος από 10-20μm), με καθαρή σύσταση χωρίς προσμίξεις και χωρίς εσωτερικές ατέλειες (σπασίματα, κ.ο.κ.). Κατά συνέπεια, οι πρωτεΐνες πρέπει να απομονωθούν από το δείγμα, να καθαριστούν και να υπάρχουν σε μεγάλη ποσότητα. Μια επιπλέον δυσκολία εντοπίζεται στο γεγονός ότι δεν κρυσταλλώνονται όλες οι πρωτεΐνες στις ίδιες συνθήκες. Έτσι, ακόμα και αν υπάρχει το δείγμα, η κρυστάλλωση είναι μια επίπονη διαδικασία με πολύ πειραματισμό σε διαφορετικές συνθήκες. Υπάρχουν δυο κύριες μεθοδολογίες για την κρυστάλλωση (διάχυση ατμών και διαπίδυση). Νέες τεχνολογίες ελεύθερων επιταχυντών ηλεκτρονίων λέιζερ επιτρέπουν τη χρήση κρυστάλλων μικρότερων από 1 μm.
- Στο δεύτερο στάδιο, ο κρύσταλλος τοποθετείται απέναντι από τη δέσμη των ακτίνων Χ, οι οποίες συνήθως είναι μονοχρωματικές, και κατόπιν συλλέγονται οι ανακλάσεις. Καθώς ο κρύσταλλος περιστρέφεται δημιουργούνται πολλαπλά σύνολα ανακλάσεων καλύπτοντας διαφορετική γωνία έκθεσης στις ακτίνες Χ. Συνολικά έτσι συλλέγονται εκατοντάδες χιλιάδες ανακλάσεις. Πολλές φορές, αν ο κρύσταλλος είναι μικρός ή με μειωμένη κανονικότητα μπορεί να καταστραφεί κατά τη διαδικασία αυτή, προτού συλλεχθούν όλα τα απαραίτητα δεδομένα. Ένας άλλος περιοριστικός παράγοντας σε αυτό το στάδιο είναι η πηγή των ακτίνων Χ, καθώς απαιτείται συνήθως κάποια συσκευή υψηλής ενέργειας, όπως το σύγχροτρο. Κάθε ανάκλαση συλλέγεται αρκετές φορές, και στατιστικές μέθοδοι χρησιμοποιούνται για τη μέτρηση της μέσης τιμής και της τυπικής απόκλισης των μετρήσεων.
- Στο τρίτο στάδιο, τα δεδομένα των ανακλάσεων συνδυάζονται με συμπληρωματικά χημικά δεδομένα για να παραχθεί ένα αρχικό μοντέλο και να βελτιστοποιηθεί στη συνέχεια. Το βασικό πρόβλημα εδώ, είναι ότι από τις ανακλάσεις δεν μπορεί να προσδιοριστεί μονοσήμαντα η θέση των ατόμων και ο χάρτης ηλεκτρονικής πυκνότητας στο δείγμα (το γνωστό πρόβλημα φάσης). Το πρόβλημα αυτό λύνεται με μια σειρά από μεθόδους όπως η εύρεση φάσης εκ του μηδενός (ab initio phasing), η μοριακή αντικατάσταση (molecular replacement), η ανώμαλη σκέδαση ακτίνων Χ (anomalous X-ray scattering) και οι μέθοδοι βαρέων ατόμων (heavy atom methods), έτσι ώστε τελικά να προκύπτει μια αρχική εκτίμηση για τη φάση. Τα επόμενα βήματα, αφορούν την κατασκευή ενός αρχικού μοντέλου και τη βελτιστοποίησή του (refinement).

Οι διαδικασίες και των τριών βημάτων απαιτούν ιδιαίτερη χρήση Η/Υ και αλγορίθμων. Στην κρυστάλλωση, οι Η/Υ χρησιμοποιούνται για τον έλεγχο ρομποτικών μονάδων κρυστάλλωσης, την ανάλυση εικόνων από πειράματα κρυστάλλωσης, αλλά και το σχεδιασμό των βέλτιστων συνθηκών για τα πειράματα κρυστάλλωσης. Στο δεύτερο στάδιο χρησιμοποιούνται ειδικά πακέτα ανάλυσης των εικόνων περίθλασης, με έμφαση στον προσδιορισμό των σφαλμάτων μέτρησης, που είναι απαραίτητα ειδικά στην επίλυση του προβλήματος των φάσεων. Υπάρχουν διάφορα πακέτα λογισμικού δομικής βιολογίας που διευκολύνουν τις διαδικασίες αυτές. Στην ιστοσελίδα της International Union of Crystallography αναφέρονται δεκάδες τέτοια πακέτα, από τα οποία κάποια είναι συλλογές με πολλαπλές χρήσεις και άλλα συγκεκριμένες ρουτίνες με εστιασμένο ενδιαφέρον (http://www.iucr.org/resources/other-directories/software). Γενικά πάντως, τα πακέτα με τη μεγαλύτερη αποδοχή, τους περισσότερους χρήστες και τις περισσότερες λειτουργίες, είναι το **CCP4,** το οποίο είναι διαθέσιμο στη διεύθυνση http://www.ccp4.ac.uk/ (Winn et al., 2011), το **PHENIX**, το οποίο είναι διαθέσιμο στη διεύθυνση https://www.phenix-online.org/ (Adams et al., 2010) και το **X-PLOR** (Güntert, 2011), το οποίο είναι ένα από τα παραδοσιακά πακέτα στον τομέα, μαζί με την βελτιωμένη του έκδοση που συντηρείται από τον NIH, το **Xplor-NIH**, το οποίο είναι διαθέσιμο στη διεύθυνση http://nmr.cit.nih.gov/xplor-nih/ (Schwieters, Kuszewski, & Clore, 2006)

Αν όλα τα στάδια στεφθούν με επιτυχία, έχουμε τελικά μια τρισδιάστατη δομή, δηλαδή ένα αρχείο με ατομικές συντεταγμένες που αντιστοιχεί στη δομή της πρωτεΐνης στο δείγμα. Συνήθως τα αρχεία αυτά κατατίθενται σε δημόσιες βάσεις δεδομένων (PDB), καθώς εδώ και χρόνια είναι υποχρεωτική η κατάθεσή τους προκειμένου οι αντίστοιχες εργασίες που τα περιγράφουν να γίνουν δεκτές προς δημοσίευση. Εκεί, πολλές φορές τα δεδομένα αυτά περνάνε και άλλους αυτοματοποιημένους ελέγχους για να αποκλειστεί το ενδεχόμενο σφάλματος και να διασφαλιστεί η ποιότητα. Στην ιστοσελίδα της PDB παρατίθεται μια μεγάλη λίστα με διαθέσιμα προγράμματα για τον έλεγχο, την αξιολόγηση και την επαλήθευση τρισδιάστατων δομών

(http://www.rcsb.org/pdb/static.do?p=software/software_links/analysis_and_verification.html). Το πρωτοπόρο πρόγραμμα σε αυτόν τον τομέα ήταν το **PROCHECK** (Laskowski, MacArthur, Moss, & Thornton, 1993). Τα προγράμματα που χρησιμοποιούνται πλέον ευρέως για αξιολόγηση και επαλήθευση ειναι το **MolProbity** το οποίο είναι διαθέσιμο στη διεύθυνση http://molprobity.biochem.duke.edu/ (Chen et al., 2010) και το **WHATCHECK** το οποίο διατίθεται στη διεύθυνση http://swift.cmbi.ru.nl/gv/whatcheck/ (Hooft, Vriend, Sander, & Abola, 1996). Η επαλήθευση των δομών είναι πλέον απαραίτητη για την δημοσίευσή τους και υπάρχουν συγκριμένες οδηγίες για αυτό τον σκοπό (Read et al., 2011). Μια νέα αντιμετώπιση, είναι και η λογική της «ενεργούς επαλήθευσης» όπου οι υπάρχουσες δομές διορθώνονται με αυτοματοποιημένους αλγόριθμους μοντελοποίησης με βάση τα κρυσταλλογραφικά δεδομένα που κατατίθενται στην PDB (Joosten et al., 2009).

Ένας εναλλακτικός τρόπος προσδιορισμού της δομής, είναι η φασματοσκοπία πυρηνικού μαγνητικού συντονισμού (NMR). Το NMR εκμεταλλεύεται τις μηχανικές ιδιότητες των ατόμων, οι οποίες εξαρτώνται από το περιβάλλον και με αυτόν τον τρόπο παράγει τελικά ένα χάρτη που απεικονίζει τον τρόπο με τον οποίο συνδέονται τα άτομα, πόσο κοντά βρίσκονται στο χώρο, και ποια είναι η σχετική τους κίνηση. Οι ιδιότητες αυτές είναι στην ουσία ίδιες με τη μεθοδολογία του μαγνητικού πυρηνικού συντονισμού (Magnetic Resonance Imaging - MRI), αλλά εδώ εστιαζόμαστε σε αποστάσεις της τάξης του Å, σε αντίθεση με τα mm που αποτελούν το αντικείμενο μελέτης των του MRI. Επίσης, μια άλλη διαφορά είναι ότι εδώ δεν παράγεται απευθείας μια εικόνα, αλλά τα δεδομένα συλλέγονται και με χρήση υπολογιστή κατασκευάζεται ένα τρισδιάστατο μοντέλο της πρωτεΐνης. Στις περισσότερες περιπτώσεις, τα δείγματα βρίσκονται σε υδατικό διάλυμα, αλλά αναπτύσσονται και μεθοδολογίες στερεάς φάσης. Η συλλογή των δεδομένων γίνεται με την τοποθέτηση του δείγματος σε έναν δυνατό μαγνήτη, τη χρήση ραδιοκυμάτων στο δείγμα και τη συλλογή του φάσματος απορρόφησης. Ανάλογα με το περιβάλλον (τόσο του διαλύτη, αλλά και των γειτονικών ατόμων), οι πυρήνες των ατόμων θα απορροφήσουν τα κύματα σε διαφορετικές συχνότητες και οι πληροφορίες αυτές μπορεί να συνδυαστούν με σκοπό να καθοριστεί ένα συνολικό μοντέλο του μορίου. Γενικά, επειδή το δείγμα βρίσκεται σε υδατικό διάλυμα και συνυπολογίζονται ταυτόχρονα οι κινήσεις όλων των ατόμων, η μέθοδος μπορεί να εφαρμοστεί κυρίως σε μικρές πρωτεΐνες (αν και υπάρχουν εξαιρέσεις). Επιπλέον, η τεχνική αυτή λειτουργεί συμπληρωματικά με την κρυσταλλογραφία, καθώς είναι περισσότερο χρήσιμη στη μελέτη της κίνησης και της δυναμικής (ευελιξία, κλπ) των πρωτεϊνικών μορίων, σε αλληλεπιδράσεις μεταξύ πρωτεϊνών με άλλες πρωτεΐνες αλλά και μικρά μόρια (φάρμακα, μεταβολίτες κ.ο.κ.), αλλά και σε περιπτώσεις πρωτεϊνών που δεν μπορούν να κρυσταλλωθούν εύκολα.

Ένα σημαντικό θέμα που πρέπει να αναφερθεί, είναι ο τρόπος καθορισμού της δευτεροταγούς δομής από τα δεδομένα κρυσταλλογραφίας. Παραδοσιακά, οι κρυσταλλογράφοι παρατηρούσαν τις δομές και οπτικά αποφάσιζαν ποιες περιοχές ήταν σε α-έλικα, ποιες σε β-πτυχωτή επιφάνεια κ.ο.κ. Επειδή όμως οι αναθέσεις αυτές, ήταν υποκειμενικές και πολλές φορές προκύπταν διαφωνίες ακόμα και μεταξύ έμπειρων κρυσταλλογράφων, αναπτύχθηκαν αυτοματοποιημένοι αλγόριθμοι οι οποίοι διαβάζουν το αρχείο με τις τρισδιάστατες συντεταγμένες και αποδίδουν όσο πιο αντικειμενικά γίνεται τα στοιχεία δευτεροταγούς δομής, καθώς και άλλα χαρακτηριστικά όπως την προσβασιμότητα των διαφόρων καταλοίπων (δηλαδή, αν είναι εκτεθειμένα ή όχι). Αυτό που πρέπει να τονιστεί, είναι ότι οι αλγόριθμοι αυτοί δεν είναι αλγόριθμοι πρόγνωσης της δευτεροταγούς δομής, δεν κάνουν δηλαδή πρόβλεψη σε κάποια αλληλουχία άγνωστης δομής, απλά εντοπίζουν σε μια προσδιορισμένη τρισδιάστατη δομή το σημείο που βρίσκονται οι α-έλικες και οι β-πτυχωτές επιφάνειες, κάνοντας χρήση αντικειμενικών κριτηρίων.

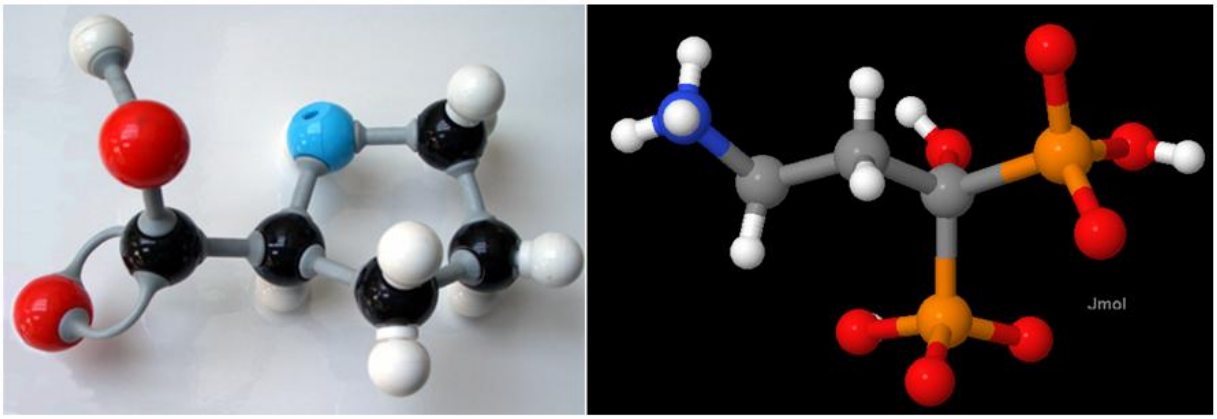
Το **DSSP** (Define Secondary Structure of Proteins), διαθέσιμο στη διεύθυνση http://swift.cmbi.ru.nl/gv/dssp/, ήταν ο πρώτος αλγόριθμος που προτάθηκε για το σκοπό αυτό και είναι ακόμα ο ευρύτερα χρησιμοποιούμενος (Kabsch & Sander, 1983). Το DSSP αναγνωρίζει τον κύριο ανθρακικό σκελετό της πρωτεΐνης και εντοπίζει τους δεσμούς υδρογόνου που σχηματίζονται, με βάση έναν καθαρά ηλεκτροστατικό ορισμό. Με βάση τον ενεργειακό υπολογισμό, το DSSP αναγνωρίζει και κατατάσσει τα στοιχεία δευτεροταγούς δομής σε 8 κατηγορίες. Η 3_{10}-έλικα, η α-έλικα, και η π-έλικα (με σύμβολα G, H και I) αναγνωρίζονται αν υπάρχουν συνεχόμενες επαναλήψεις του δεσμού υδρογόνου με βήμα 3, 4, ή 5 κατάλοιπα αντίστοιχα. Οι β-δομές χωρίζονται σε β-πτυχωτή επιφάνεια (E) και β-γέφυρα (B), το σύμβολο T χρησιμοποιείται για τις στροφές και το S για περιοχές υψηλής καμπυλότητας. Τέλος, περιοχές που δεν ταιριάζουν με κανένα πρότυπο μένουν με το κενό σύμβολο. Συνήθως στις παρακάτω αναλύσεις, όπως π.χ. στην πρόγνωση δευτεροταγούς δομής τα σύμβολα αυτά ομαδοποιούνται και αυτό μπορεί να γίνει με δύο τρόπους. Στην πρώτη περίπτωση α-έλικα μένει το H, β-πτυχωτή επιφάνεια το E και όλα τα άλλα γίνονται τυχαία δομή (coil) με σύμβολο το C. Ο εναλλακτικός τρόπος περιλαμβάνει την ομαδοποίηση στο H και των άλλων ελίκων (G, I), στο E την προσθήκη του B, ενώ τα υπόλοιπα γίνονται C. Το 2002 μια νεότερη έκδοση

του DSSP εμφανίστηκε η οποία πραγματοποιεί ανάθεση με πιο ευέλικτα όρια (continuous DSSP) η οποία φαίνεται να προσφέρει κάποια επιπλέον πλεονεκτήματα (Andersen, Palmer, Brunak, & Rost, 2002).

Το **STRIDE** (STRuctural IDEntification), το οποίο είναι διαθέσιμο στη διεύθυνση http://webclu.bio.wzw.tum.de/stride/ είναι ένας άλλος εναλλακτικός αλγόριθμος για τον προσδιορισμό και την ανάθεση των στοιχείων δευτεροταγούς δομής (Frishman & Argos, 1995). Το STRIDE χρησιμοποιεί μια παρόμοια μέθοδο με το DSSP, καθώς εφαρμόζει μια μέτρηση ενέργειας για τον προσδιορισμό των δεσμών υδρογόνου (ένα δυναμικό Lennard-Jones), αλλά επιπλέον λαμβάνει υπόψη του και τις δίεδρες γωνίες που σχηματίζονται. Στο τέλος, αναθέτει δευτεροταγείς δομές στις ίδιες κατηγορίες που χρησιμοποιεί το DSSP, αλλά επιπλέον δίνει και μια ανά κατάλοιπο τιμή για την αξιοπιστία της ανάθεσης, η οποία έχει προκύψει από εμπειρικές μελέτες. Παρόλο που το DSSP είναι το πιο παλιό και ευρύτερα αποδεκτό πρόγραμμα, το STRIDE πιστεύεται ότι είναι σχετικά καλύτερο και διορθώνει την τάση του DSSP να ορίζει κάπως μικρότερα τμήματα δευτεροταγούς δομής σε σχέση με τους ορισμούς που πραγματοποιούν οι έμπειροι κρυσταλλογράφοι. Τα τελευταία χρόνια, το STRIDE χρησιμοποιείται και στην PDB (παράλληλα με το DSSP), ενώ υπάρχει και διαδικτυακή εφαρμογή διαθέσιμη για άμεση χρήση από το ευρύ κοινό (Heinig & Frishman, 2004).

9.2. Οπτικοποίηση βιολογικών δομών

Δεδομένης της ύπαρξης της τρισδιάστατης δομής μιας πρωτεΐνης, το πρώτο πράγμα που θα ενδιέφερε κάποιον θα ήταν η οπτικοποίηση. Οι μεθοδολογίες απεικόνισης των μακρομορίων, ξεκίνησαν παράλληλα με τις πρώτες επιτυχίες της κρυσταλλογραφίας ακτίνων Χ από την δεκαετία του 1950 και 1960. Αρχικά, για τα μοντέλα αυτά χρησιμοποιήθηκαν ξύλινες σφαίρες για να αναπαραστήσουν τα άτομα και ράβδοι για να αναπαραστήσουν τους δεσμούς. Τέτοια μοντέλα, χρησιμοποιούνται ακόμα και σήμερα για εκπαιδευτικούς λόγους αλλά είναι πλέον πλαστικά και με διαφορετικό χρωματισμό για τα διαφορετικά είδη ατόμων. Αυτή η αναπαράσταση, γίνεται πλέον και σε υπολογιστές και ονομάζεται «ball and stick». Μια παρόμοια αναπαράσταση, είναι και το σκελετικό μοντέλο (wireframe) στο οποίο απεικονίζονται όμως μόνο οι δεσμοί ως σύρματα, ενώ τα άτομα συμπίπτουν με τις κορυφές (γωνίες). Τέτοιες αναπαραστάσεις είναι φυσικά πολύ απλές και μπορούν εύκολα να πραγματοποιηθούν σε υπολογιστή αλλά κάποιες φορές είναι ιδιαίτερα χρήσιμες. Πολλές φορές μάλιστα για λόγους απλότητας αναπαρίστανται μόνο οι Cα (τα ασύμμετρα άτομα άνθρακα, δηλαδή ο κύριος σκελετός της πρωτεΐνης) ενώ άλλες φορές αναπαρίστανται όλα τα άτομα με διαφορετικό χρωματισμό.

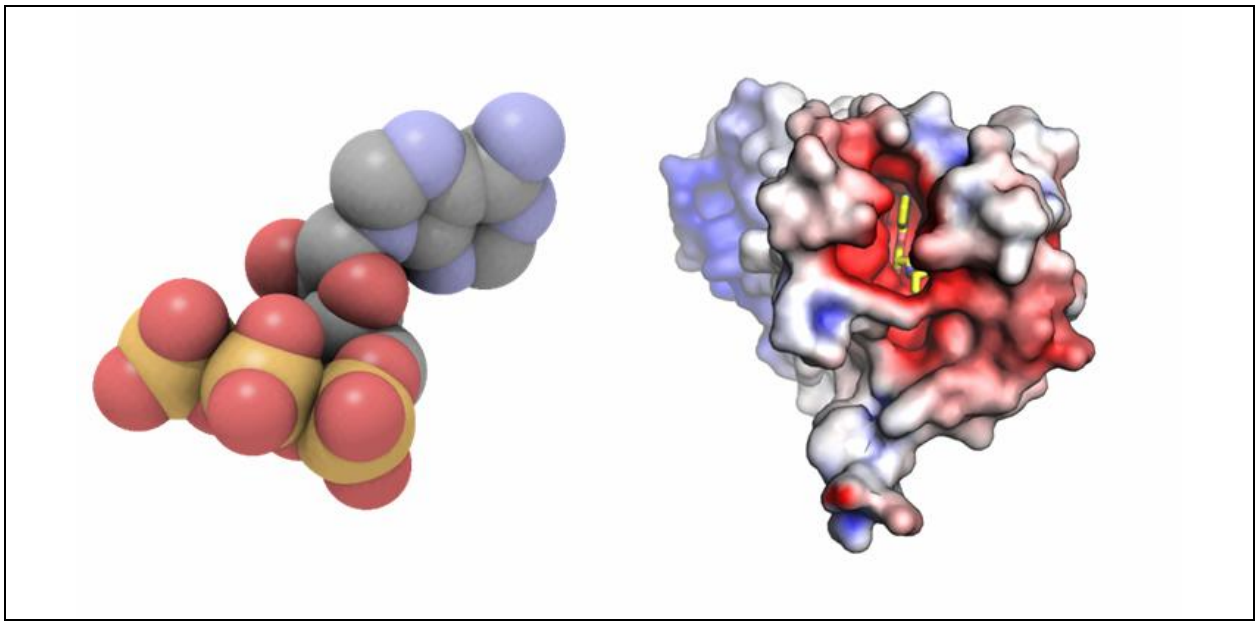


Εικόνα 9.3: *Αριστερά, πλαστικό μοντέλο τύπου «ball-and-stick» (https://en.wikipedia.org/wiki/Molecular_model). Δεξιά, μοντέλο ενός μικρού μορίου ($NH_3CH_2CH_2C(OH)(PO_3H)(PO_3H)$-), όπως παράγεται από το Jmol (https://en.wikipedia.org/wiki/Molecular_graphics)*

Ένας άλλος τρόπος αναπαράστασης, είναι το λεγόμενο χωροπληρωτικό μοντέλο (space-filling model), στο οποίο τα άτομα αναπαρίστανται πάλι με σφαίρες, συνήθως διαφορετικού χρώματος, αλλά με τη σημαντική προσθήκη ότι οι ακτίνες της κάθε σφαίρας είναι ανάλογες με την ακτίνα van der Waals του ατόμου. Τα μοντέλα αυτά ονομάζονται και CPK models από τους Corey, Pauling, και Koltun, οι οποίοι ανέπτυξαν πρώτοι τέτοιες τεχνικές απεικόνισης, ενώ πλαστικά μοντέλα αυτού του είδους χρησιμοποιούνται ακόμα και σήμερα για διδακτικούς σκοπούς. Καθώς οι ακτίνες των ατόμων είναι μικρότερες από την ενδομοριακή απόσταση όταν τα άτομα τα ενώνει ομοιοπολικός δεσμός, οι σφαίρες πρέπει να τέμνονται, και

κατά συνέπεια στα πλαστικά μοντέλα αυτού του είδους οι σφαίρες είναι κολοβές καθώς αφαιρείται μια περιοχή σαν «καπάκι» έτσι ώστε τα άτομα να έρχονται σε επαφή. Τα μοντέλα αυτά είναι πιο ρεαλιστικά, καθώς δίνουν μια απεικόνιση της επιφάνειας του μορίου που βρίσκεται πιο κοντά στην πραγματικότητα. Δεν δίνουν όμως καλή εικόνα της δευτεροταγούς δομής ή της κατεύθυνσης της πολυπεπτιδικής αλυσίδας. Κατά συνέπεια είναι πιο χρήσιμα σε δυναμικές μελέτες, π.χ. για τον υπολογισμό της έκθεσης στο διαλύτη ή για τον υπολογισμό επιφανειών επαφής και μελέτες αγκυροβόλησης. Μια συνηθισμένη παραλλαγή αυτών των μοντέλων είναι αυτή στην οποία η επιφάνεια εμφανίζεται πιο ομαλή και τα άτομα έχουν χρωματιστεί ανάλογα με την ηλεκτραρνητικότητα (κόκκινο) ή την ηλεκτροθετικότητά τους (μπλε).

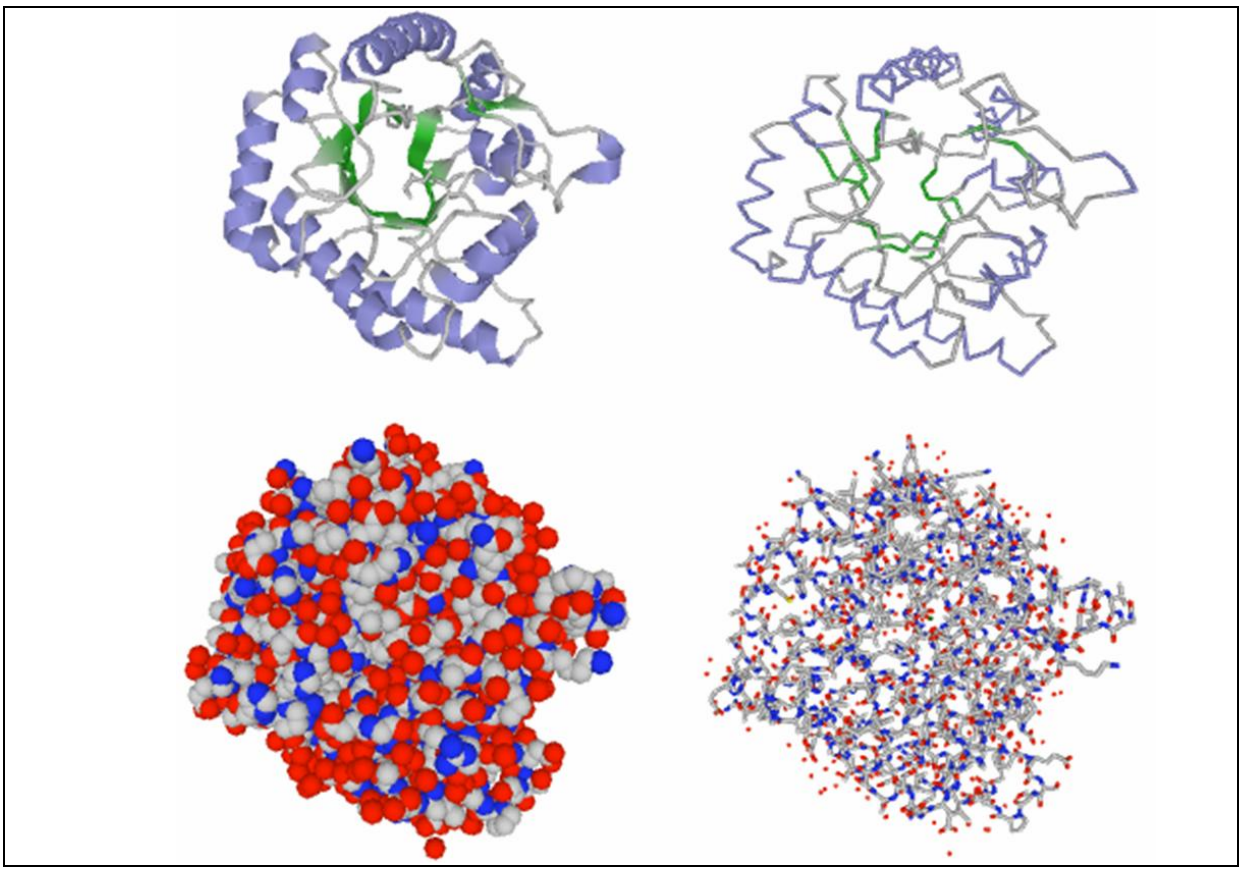
Τέλος, ένας άλλος διαδεδομένος τρόπος αναπαράστασης -που έγινε πολύ διαδεδομένος με τη χρήση H/Y-, είναι το λεγόμενο διάγραμμα κορδέλα (Ribbon diagram) ή καρτούν, το οποίο είναι ιδιαίτερα πληροφοριακό και είναι ίσως από τις δημοφιλέστερες μεθόδους αναπαράστασης. Σε ένα τέτοιο διάγραμμα απεικονίζεται ολόκληρος ο σκελετός της πρωτεΐνης και η κατεύθυνση της πολυπεπτιδικής αλυσίδας και σχεδιάζονται με ειδικό τρόπο τα στοιχεία δευτεροταγούς δομής. Έτσι, οι ά-έλικες αναπαρίστανται σαν κορδέλες (ribbon) που σχηματίζουν ελικοειδή διάταξη ή σαν κύλινδροι, ενώ οι β-πτυχωτές επιφάνειες σαν πεπλατυσμένα βέλη. Τέλος, οι περιοχές με μη κανονική δευτεροταγή δομή, απεικονίζονται σαν καμπυλωτές γραμμές. Επιπλέον δε, τα διαφορετικά στοιχεία δευτεροταγούς δομής χρωματίζονται συνήθως με διαφορετικό τρόπο έτσι ώστε να είναι πιο εύκολη η διάκρισή τους. Η μέθοδος αυτή είναι πολύ πληροφοριακή, γιατί βλέπουμε αμέσως τη διάταξη των στοιχείων δευτεροταγούς δομής και τις μεταξύ τους σχέσεις, αλλά και την κατεύθυνση της αλυσίδας. Για αυτούς τους λόγους τα μοντέλα αυτά χρησιμοποιούνται στις περισσότερες περιπτώσεις στις επιστημονικές δημοσιεύσεις.



Εικόνα 9.4: *Αριστερά, ένα παράδειγμα χωροπληρωτικού μοντέλου του ATP. Δεξιά, ένα παράδειγμα χωροπληρωτικού μοντέλου του β2 αδρενεργικού υποδοχέα, (PDB code 2RH1, από https://en.wikipedia.org/wiki/Space-filling_model)*

Σήμερα, υπάρχουν διαθέσιμα δεκάδες προγράμματα για μοριακή απεικόνιση τρισδιάστατων δομών. Τα περισσότερα από αυτά είναι ανοιχτού κώδικα και διανέμονται δωρεάν, άλλα λειτουργούν σαν αυτόνομες εφαρμογές ενώ άλλα λειτουργούν σαν πρόσθετα στον περιηγητή ιστού. Όλα, δέχονται σαν είσοδο ένα αρχείο PDB το οποίο αποτελεί το αποδεκτό πρότυπο για τέτοιου είδους δεδομένα. Τα προγράμματα αυτά, διαθέτουν πλέον πάρα πολλές λειτουργίες και ο κάθε χρήστης μπορεί να βρει κάποιο που να καλύπτει τις ανάγκες του (κάποιος μπορεί να ενδιαφέρεται για την απλότητα, κάποιος για τις υπολογιστικές απαιτήσεις, κάποιος για μια συγκεκριμένη λειτουργία που ένα δεδομένο πρόγραμμα επιτελεί καλύτερα κ.ο.κ.). Τα περισσότερα προγράμματα πάντως, παρέχουν δυνατότητες αναπαράστασης με όλα τα παραπάνω μοντέλα. Δίνουν την επιλογή να επιλέξει ο χρήστης το χρωματισμό που επιθυμεί, ενώ είναι και διαδραστικά καθώς επιτρέπουν στο χρήστη να περιστρέψει το μόριο, να μεγεθύνει σε κάποιο σημείο, να επιλέξει κάποια κατάλοιπα, να τα χρωματίσει διαφορετικά αλλά και να επιτρέψει διαφορετικό τρόπο αναπαράστασης για κάποια επιλεγμένα κατάλοιπα. Το πώς θα τα χρησιμοποιήσει ο κάθε χρήστης, διαφέρει και εξαρτάται από τις ανάγκες του. Για

παράδειγμα, κάποιος που ενδιαφέρεται να πάρει μια εικόνα για το δίπλωμα της πρωτεΐνης και το γενικότερο σχήμα της, συνήθως θα επιλέξει ένα διάγραμμα ribbon. Κάποιος που θέλει να δει την τεταρτοταγή δομή, θα επιλέξει διαφορετικό χρωματισμό στις διαφορετικές πολυπεπτιδικές αλυσίδες, ενώ κάποιος που θέλει να μελετήσει τη λειτουργία ενός ενζύμου, θα εστιαστεί στο ενεργό κέντρο και θα χρησιμοποιήσει διαγράμματα wireframe ή ball and stick και θα χρωματίσει διαφορετικά τα διάφορα άτομα. Τέλος, πολλά από τα προγράμματα αυτά παρέχουν επιπλέον λειτουργίες δομικής βιοπληροφορικής, από υπολογισμό αποστάσεων ατόμων και υπολογισμό φορτίων και επιφανειών, μέχρι και λειτουργίες δομικής στοίχισης (βλ. παρακάτω).



Εικόνα 9.5: *Διαφορετικές αναπαραστάσεις του ίδιου μορίου μπορούν να χρησιμοποιηθούν σε διαφορετικές περιστάσεις και με διαφορετικό σκοπό. Βλέπουμε εδώ τη δομή της Ισομεράσης της Ξυλόζης από τον Planctomyces limnophilus (PDB code 3TVA). Οι εικόνες δημιουργήθηκαν με το PV.*

Στην ιστοσελίδα της PDB παρατίθεται μια μεγάλη λίστα από τέτοια προγράμματα τα οποία καλύπτουν όλες τις ανάγκες (http://www.rcsb.org/pdb/static.do?p=software/software_links/molecular_graphics.html). Η ίδια η PDB έχει ενσωματώσει μια σειρά από τέτοια εργαλεία στη διαδικτυακής της πλατφόρμα με σκοπό ο απλός χρήστης να μπορεί να οπτικοποιήσει αμέσως τις δομές για τις οποίες έχει κάνει αναζήτηση και να δει με διαδραστικό τρόπο τα αποτελέσματα. Τα εργαλεία αυτά ποικίλουν από το απλό **RCSB Simple Viewer (**http://biojava.org/wiki/RCSB_Viewers:About**),** το οποίο βασίζεται στην τεχνολογία Java Web Start και δίνει μια βασική διαδραστικότητα με λειτουργίες του ποντικιού, μέχρι το **Jmol** (http://jmol.sourceforge.net/), το οποίο είναι εφαρμογή Java Applet, και το **Jsmol** το οποίο είναι η ειδική έκδοση του τελευταίου και χρησιμοποιεί JavaScript και HTML5 (http://sourceforge.net/projects/jsmol/). Και τα δύο τελευταία εργαλεία προσφέρουν πολλές λειτουργικότητες και ευκολίες ακόμα και στον πεπειραμένο χρήστη. Υπάρχει ακόμα και το **PV** το οποίο βασίζεται στην τεχνολογία WebGL και παρέχει τις βασικές λειτουργίες με ένα ιδιαίτερα εύχρηστο μενού επιλογών.

Άλλη παρόμοια εφαρμογή που μπορεί να χρησιμοποιήσει κάποιος είναι το **RasMol** (http://www.bernstein-plus-sons.com/software/rasmol/), το οποίο είναι από τις πιο παλιές εφαρμογές για μοριακή απεικόνιση, η οποία εξελίσσεται συνεχώς και έχει αποκτήσει και έκδοση ανοιχτού κώδικα το

OpenRasMol (http://www.openrasmol.org/). Το πακέτο CCP4 που αναφέραμε παραπάνω, περιέχει και τη δική του αντίστοιχη εφαρμογή, το **CCP4mg** (http://www.ccp4.ac.uk/MG/), ενώ υπάρχουν και άλλες διαδραστικές εφαρμογές που κατασκευάστηκαν ως συμπληρωματικά εργαλεία άλλων διαδικτυακών τόπων και εφαρμογών, όπως για παράδειγμα το **Swiss-PDBviewer** (http://spdbv.vital-it.ch/), το οποίο είναι στενά συνδεδεμένο με το **SWISS-MODEL** (βλ. παρακάτω) και το **Cn3D** (http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml) το οποίο αποτελεί τμήμα των εφαρμογών του NCBI και είναι στενά συνδεδεμένο με το το Entrez, ενώ παρέχει και δυνατότητες alignment editor. Τέλος, δεν πρέπει να παραλείψουμε να κάνουμε αναφορά στο πιο πετυχημένο ίσως εργαλείο της κατηγορίας αυτής, το **PyMol** (http://www.pymol.org/pymol) το οποίο βασίζεται στη γλώσσα προγραμματισμού Python και κάνει χρήση της τεχνολογίας OpenGL Extension Wrangler Library (GLEW). Το PyMol είναι ίσως η πιο επιτυχημένη εφαρμογή της κατηγορίας, καθώς συνδυάζει άριστη απόδοση γραφικών, πολλές επιλογές για την οπτικοποίηση ακόμα και για τους απαιτητικούς χρήστες και μεγάλη ευκολία στη χρήση ακόμα και για τους αρχάριους. Τέλος, αξίζει μια ειδική αναφορά και στο πακέτο λογισμικού για μοντελοποίηση και επεξεργασία δομών **WHAT IF** (βλ. παρακάτω) που ήταν για πολλά χρόνια το μόνο λογισμικό το οποίο επέτρεπε την 3D αναπαράσταση δομών κάνοντας χρήση των γυαλιών από τα βιντεοπαιχνίδια (σε αντίθεση με τα πολύ πιο ακριβά συστήματα της SGI που ήταν διαθέσιμα για ειδικά συστήματα Unix).

9.3. Στοίχιση και υπέρθεση δομών

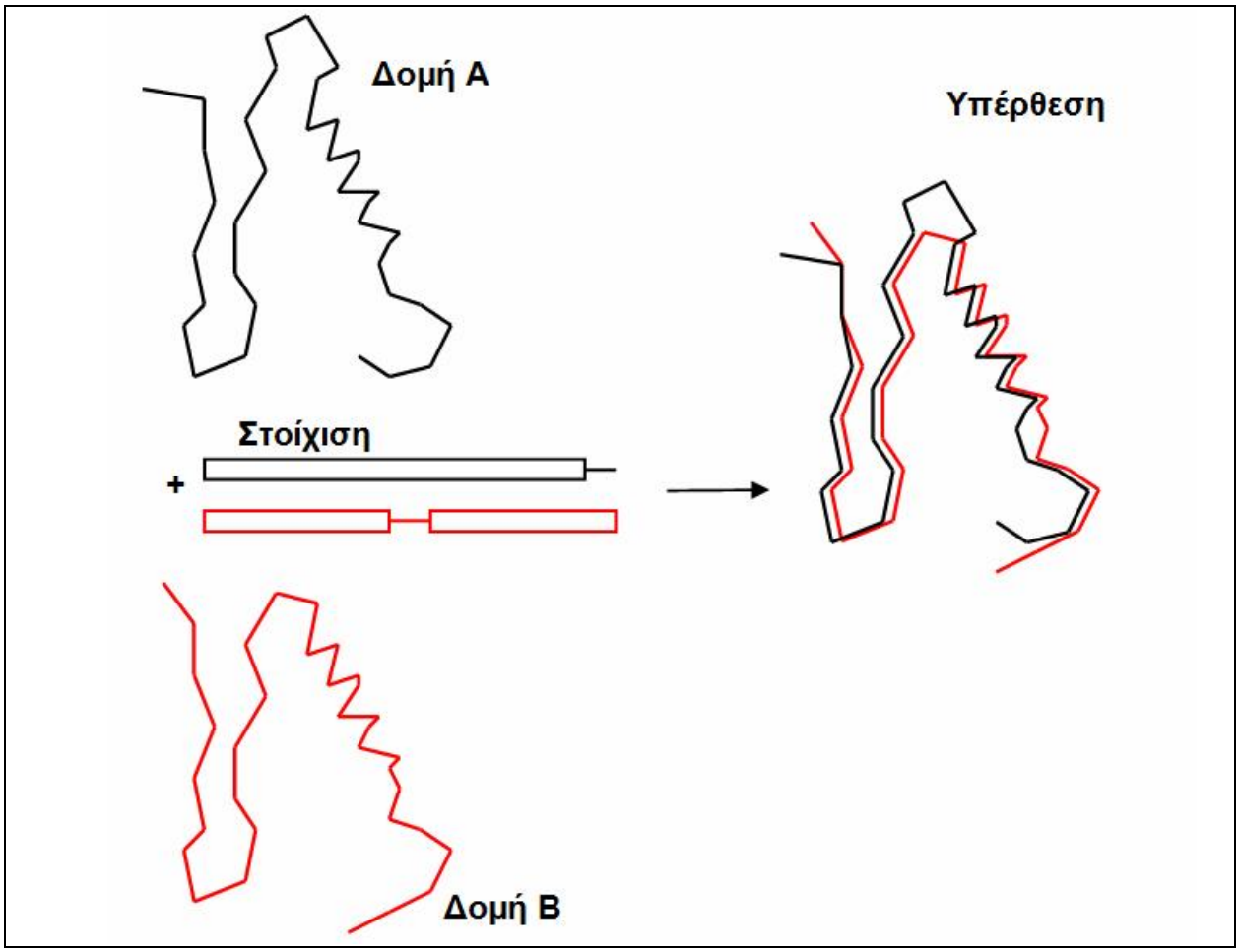
Η δομική στοίχιση (ή στοίχιση δομών) επιχειρεί να εντοπίσει και να τεκμηριώσει την ομολογία δύο πρωτεϊνών μέσω της ομοιότητας των τρισδιάστατων δομών τους (σε αντιδιαστολή με τη στοίχιση αλληλουχιών που επιχειρεί το ίδιο μέσω της ομοιότητας των αλληλουχιών). Γενικά, η διαδικασία και εδώ είναι πιο σύνθετη από τη στοίχιση αλληλουχιών, καθώς θα πρέπει να συγκρίνουμε τρισδιάστατες δομές, δηλαδή συντεταγμένες των ατόμων και όχι απλά δυο μονοδιάστατες αλληλουχίες. Από την άλλη, οι δομικές στοιχίσεις είναι πολύ χρήσιμες, γιατί μπορεί να μας αποκαλύψουν περισσότερα σε σχέση με τις στοιχίσεις αλληλουχιών, καθώς όπως έχουμε αναφέρει η τρισδιάστατη δομή συντηρείται περισσότερο από την αλληλουχία. Κατά συνέπεια, δυο πρωτεΐνες μπορεί να διαφέρουν σε επίπεδο αλληλουχίας περισσότερο από όσο μπορούμε να ανιχνεύσουμε με τις μεθόδους στοίχισης αλληλουχιών, αλλά εντούτοις να εμφανίζουν ξεκάθαρη δομική ομοιότητα. Φυσικά, υπάρχει πάντα ο κίνδυνος να εντοπίσουμε προϊόντα συγκλίνουσας (σε επίπεδο δομής) εξέλιξης και όλα αυτά αποτελούν παράγοντες που πρέπει να λαμβάνονται υπόψη. Οι δομικές στοιχίσεις έχουν και πολλές πρακτικές εφαρμογές, καθώς από αυτές προκύπτουν όπως έχουμε δει στο Κεφάλαιο 4, οι πολλαπλές στοιχίσεις αναφοράς από γνωστές πρωτεϊνικές οικογένειες με τις οποίες αξιολογούμε τις μεθόδους πολλαπλής στοίχισης αλληλουχιών, ενώ με βάση μια δομική στοίχιση μπορούν να γίνουν μια σειρά από δομικές μελέτες για τη σχέση δομής/λειτουργίας μιας δεδομένης πρωτεϊνικής οικογένειας (ή δύο συγκεκριμένων πρωτεϊνών).

Οι περιπτώσεις δομικής στοίχισεις ποικίλουν, ανάλογα με το είδος και τη σχέση των πρωτεϊνών που συγκρίνουμε. Στην πιο απλή περίπτωση, έχουμε την ίδια αλληλουχία με δομή προσδιορισμένη διαφορετικά (με διαφορετική μέθοδο ή σε σύμπλοκο με διαφορετικές ουσίες). Παρόμοια είναι και η περίπτωση δύο πρωτεϊνών με μικρές διαφορές στο επίπεδο της αμινοξικής αλληλουχίας, π.χ. με μία αντικατάσταση σε κάποιο ή κάποια αμινοξέα. Στην περίπτωση αυτή, ξέρουμε εκ των προτέρων ότι τα αμινοξέα της μίας πρωτεΐνης έχουν αντιστοίχιση με τα αμινοξέα της δεύτερης (το $1^{ο}$ με το $1^{ο}$, το $2^{ο}$ με το $2^{ο}$ κ.ο.κ.), και αυτό που έχουμε να κάνουμε είναι μια απλή υπέρθεση δομών (structural superposition). Αν τυχόν οι διαφορές είναι λίγο περισσότερες, αλλά σε κάθε περίπτωση γνωστές εκ των προτέρων, μαζί με τις δύο δομές προμηθεύουμε και μια στοίχιση αλληλουχιών η οποία θα καθοδηγεί το πρόγραμμα όσον αφορά το ποιά ζευγάρια αμινοξέων θα συγκρίνει. Η διαδικασία αυτή ονομάζεται υπέρθεση δομών και είναι η απλούστερη περίπτωση δομικής στοίχισης (έχει όμως αρκετές διαφορές από τη γενικότερη μεθοδολογία που θα δούμε παρακάτω).

Η υπέρθεση των δύο δομών (Εικόνα 9.6), περιλαμβάνει μια διαδικασία σχετικής μετακίνησης της μίας σε σχέση με την άλλη (χωρίς όμως η θέση και η διάταξη των ατόμων της ίδιας πρωτεΐνης να αλλάζει) με σκοπό οι δύο δομές να έρχονται όσο πιο κοντά γίνεται. Η διαδικασία αυτή, αν και φαίνεται απλή σαν ιδέα, έχει αρκετές πρακτικές δυσκολίες. Το μέτρο που αποδίδει αυτή την «ομοιότητα», δηλαδή το πόσο κοντά είναι η μια δομή με την άλλη, είναι το RMSD (Root Mean Square Deviation):

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

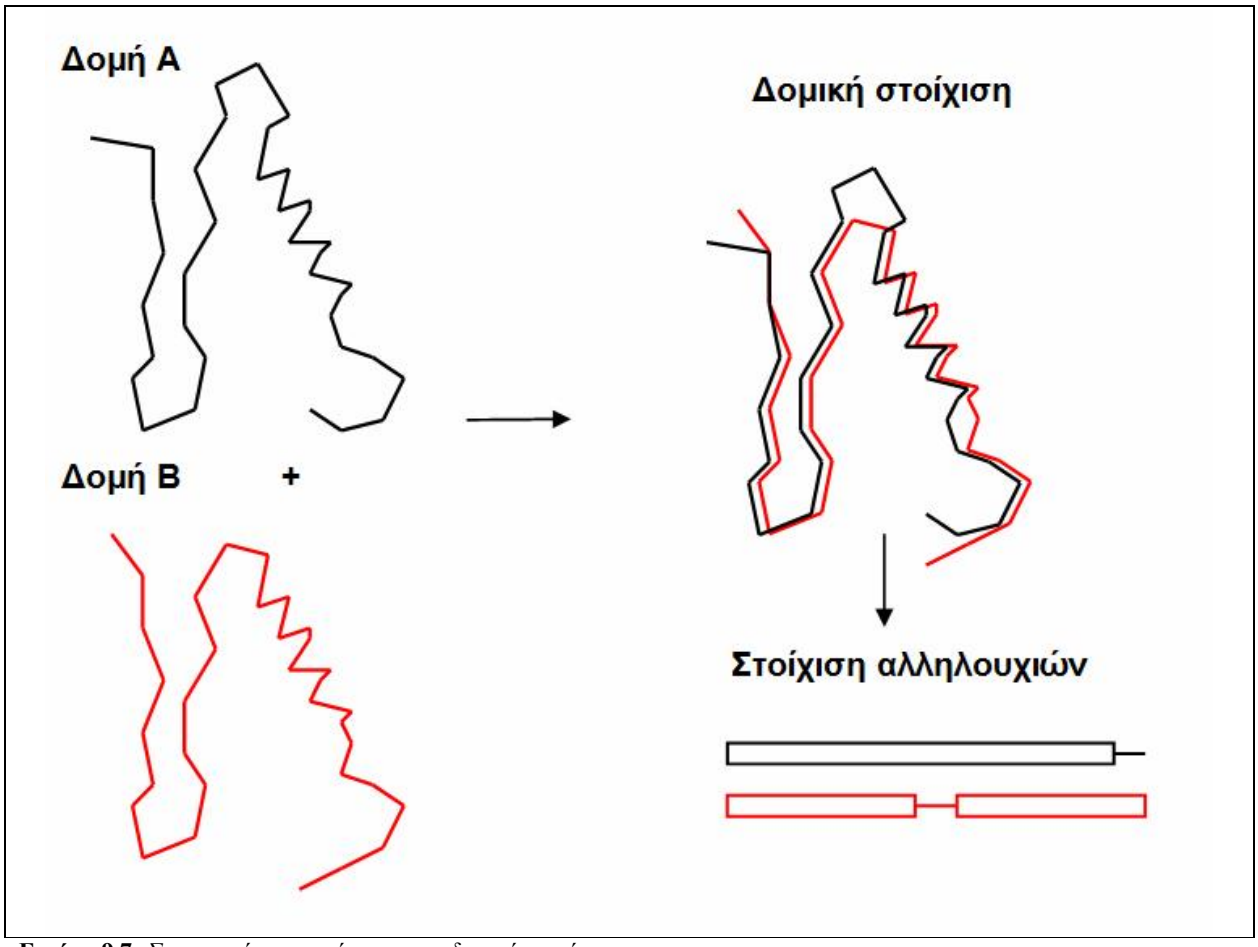
όπου N είναι ο αριθμός των ζευγαριών ατόμων που συγκρίνουμε και δ_i η απόσταση στο χώρο του i ζεύγους.



Εικόνα 9.6: *Σχηματική αναπαράσταση της υπέρθεσης δομών.*

Στις περισσότερες των περιπτώσεων, μόνο τα άτομα της κύριας ανθρακικής αλυσίδας (Cα) χρησιμοποιούνται για τις συγκρίσεις αυτές καθώς αυτά είναι που θα καθορίσουν το γενικότερο σχήμα και τη δομή της πρωτεΐνης και οι υπολογισμοί είναι ευκολότεροι. Επιπλέον δε, η σύγκριση των ατόμων των πλευρικών αλυσίδων είναι προβληματική όταν έχουμε να κάνουμε με σύγκριση μη-ταυτόσημων αλληλουχιών. Γενικά, το κριτήριο αυτό χρησιμοποιείται ευρέως, τόσο στην υπέρθεση και τη δομική στοίχιση, αλλά όπως θα δούμε και παρακάτω και σε περιπτώσεις αξιολόγησης θεωρητικών μοντέλων. Η μέθοδος των ελαχίστων τετραγώνων (least squares method) χρησιμοποιείται παραδοσιακά από τους αλγόριθμους υπέρθεσης δομών, αλλά έχουν αναπτυχθεί και μεθοδολογίες που βασίζονται σε αναλύσεις μέγιστης πιθανοφάνειας (maximum likelihood) (Theobald & Wuttke, 2006a, 2006b) αλλά και σταθερών (robust) μεθοδολογιών, όπως η least median squares regression (LMS) (Liu, Fang, & Ramani, 2009). Οι μεθοδολογίες της πρώτης κατηγορίας έχουν υλοποιηθεί στο πρόγραμμα **LSQMAN** (http://xray.bmc.uu.se/usf/lsqman_man.html), της δεύτερης στο **THESEUS** (http://www.theseus3d.org) ενώ της τρίτης στο **LMSfit** (https://engineering.purdue.edu/PRECISE/LMSfit). Το **Profit** (http://www.bioinf.org.uk/software/profit/) είναι μια άλλη γνωστή διαδικτυακή εφαρμογή για υπέρθεση δομών χρησιμοποιώντας τη γρήγορη μέθοδο ελαχίστων τετραγώνων του McLachlan (McLachlan, 1982), ενώ το **3dSS** (http://cluster.physics.iisc.ernet.in/3dss/) είναι μια πιο σύγχρονη εφαρμογή η οποία διασυνδέεται με το RasMol ενώ κάνει και εσωτερικά χρήση του Profit, και επιτρέπει μεταξύ άλλων πολλαπλή υπέρθεση δομών, υπέρθεση υπομονάδων αλλά και άλλες ευκολίες για τον τελικό χρήστη (Sumathi, Ananthalakshmi, Roshan, & Sekar, 2006).

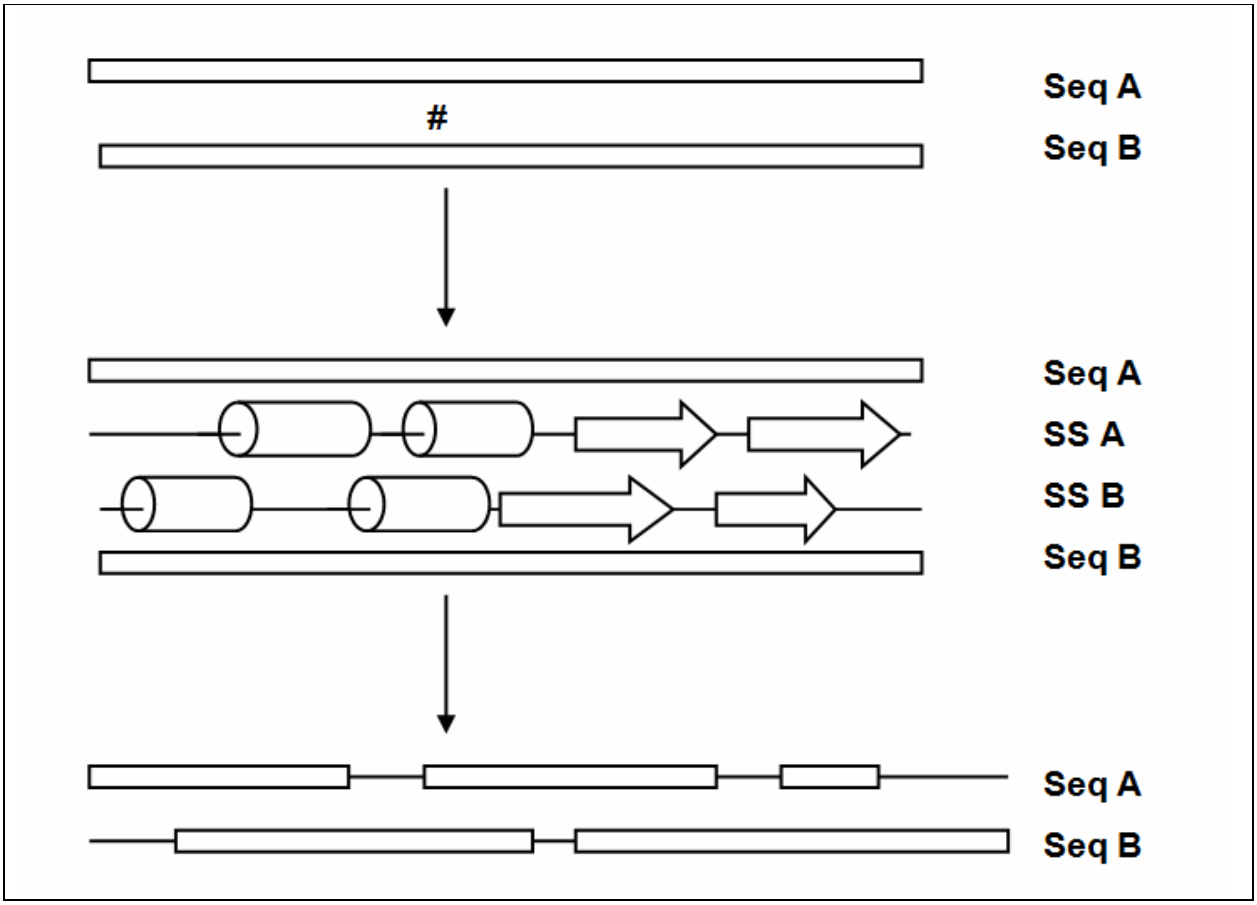
Όπως είπαμε παραπάνω, για την υπέρθεση δομών εκτός από την τετριμένη περίπτωση κατά την οποία έχουμε δύο δομές της ίδιας ακριβώς πρωτεΐνης, είναι απαραίτητη μια στοίχιση των αλληλουχιών. Σε περιπτώσεις πρωτεϊνών με μερικές μόνο μικρές διαφορές στην αμινοξική αλληλουχία, αυτό είναι κάτι εύκολο και κατανοητό. Τί γίνεται όμως όσο οι διαφορές μεταξύ των πρωτεϊνών που επιθυμούμε να συγκρίνουμε μεγαλώνουν; Καθώς είναι γνωστό ότι η δομή συντηρείται περισσότερο από την αλληλουχία, είναι αναμενόμενο ότι θα υπάρχουν περιπτώσεις πρωτεϊνών με παρόμοια δομή αλλά μικρή μόνο και πιθανώς μη ανιχνεύσιμη ομοιότητα σε επίπεδο αλληλουχίας. Για την ακρίβεια, αυτός ακριβώς είναι και ο λόγος για τον οποίο επιθυμούμε να κάνουμε σύγκριση δομών, για να μπορέσουμε δηλαδή να καταλήξουμε τελικά σε μια στοίχιση αλληλουχιών και να μελετήσουμε την απόκλιση δομής/αλληλουχίας (Εικόνα 9.7).



Εικόνα 9.7: *Σχηματική αναπαράσταση της δομικής στοίχισης*.

Ένας απλός τρόπος, για να πραγματοποιήσουμε τη στοίχιση, όταν δεν υπάρχει μεγάλη ομοιότητα, ο οποίος χρησιμοποιείται από διάφορα προγράμματα, είναι να στηριχθούμε στη δευτεροταγή δομή. Η ιδέα είναι απλή και στηρίζεται στο γεγονός ότι η δευτεροταγής δομή μπορεί να κατευθύνει τη στοίχιση. Ένας απλός τρόπος να το επιτύχουμε αυτό, θα ήταν να κάνουμε στοίχιση των ακολουθιών των δευτεροταγών δομών (δηλαδή δυο ακολουθιών που αποτελούνται από τρία σύμβολα: H, E, και C), ενώ ένας λίγο πιο σύνθετος θα ήταν με κάποιον τροποποιημένο αλγόριθμο στοίχισης, στον οποίο η συνεισφορά στο σκορ για δυο αμινοξικά κατάλοιπα θα αυξάνεται αν τα δύο κατάλοιπα έχουν την ίδια δευτεροταγή δομή. Παρόμοιες τεχνικές θα δούμε και στην περίπτωση της ύφανσης (threading) παρακάτω. Μόλις η στοίχιση κατασκευαστεί, τότε είναι εύκολο πλέον να πραγματοποιηθεί η υπέρθεση δομών όπως περιγράφεται παραπάνω χρησιμοποιώντας τη στοίχιση αυτή σαν οδηγό. Το **SuperPose** (http://wishart.biology.ualberta.ca/SuperPose/) είναι μία πολύ εύχρηστη διαδικτυακή εφαρμογή για υπέρθεση δομών η οποία απαιτεί ελάχιστη παρέμβαση από το χρήση (Maiti, Van Domselaar, Zhang, & Wishart, 2004). Το πρόγραμμα λειτουργεί αυτόματα. Έτσι, όταν οι αλληλουχίες διαφέρουν αλλά μπορούν να στοιχηθούν με κάποιον κλασικό αλγόριθμο ομοιότητας, λειτουργεί με τον κλασικό τρόπο που περιγράψαμε παραπάνω. Όταν όμως οι αλληλουχίες των πρωτεϊνών διαφέρουν

πέραν των ορίων ανίχνευσης των αλγορίθμων στοίχισης, το SuperPose χρησιμοποιεί την αναφερθείσα τεχνική με τη βοήθεια της δευτεροταγούς δομής για να μπορέσει να κάνει την υπέρθεση των δομών και να δώσει κάτι που μοιάζει με δομική στοίχιση. Πρέπει να τονιστεί βέβαια, ότι παρόλο που αυτό μοιάζει αρκετά με δομική στοίχιση, και σε πολλές περιπτώσεις λειτουργεί και παράγει παρόμοια αποτελέσματα, με βάση τον ορισμό η μέθοδος αυτή δεν θεωρείται τυπική περίπτωση δομικής στοίχισης, γιατί η στοίχιση δεν πραγματοποιείται με χρήση της δομικής πληροφορίας αλλά με χρήση της αλληλουχίας, ενώ τα όποια κενά εισάγονται μόνο με τη βοήθεια της στοίχισης αλληλουχιών και παραμένουν σταθερά κατά την προσαρμογή των δομών.



Εικόνα 9.8: *Στοίχιση αλληλουχιών με τη βοήθεια της δευτεροταγούς δομής (παρατηρηθείσας ή προβλεφθείσας). Η μέθοδος μπορεί να χρησιμοποιηθεί τόσο στην υπέρθεση δομών όσο και στην ύφανση.*

Προχωρώντας στις πιο κλασικές μεθόδους δομικής στοίχισης, σε αυτές δηλαδή που εφαρμόζονται κάνοντας χρήση της δομής των πρωτεϊνών και είναι ιδανικές για περιπτώσεις στις οποίες δεν υπάρχει ανιχνεύσιμη ομοιότητα των αλληλουχιών, θα πρέπει να κάνουμε κάποιες επισημάνσεις. Καταρχήν, το πρόβλημα της βέλτιστης ύφανσης, δηλαδή της στοίχισης μιας αλληλουχίας με μια δομή, έχει αποδειχτεί NP-complete (Lathrop, 1994), αλλά το πρόβλημα της βέλτιστης στοίχισης δύο δομών, δηλαδή της βέλτιστης προσαρμογής με βελτιστοποίηση κάποιου προκαθορισμένου κριτηρίου, δεν είναι NP (Poleksic, 2009). Έτσι, βέλτιστη λύση μπορεί να βρεθεί (με την προϋπόθεση πάντα ότι μιλάμε για κάποιο δεδομένο κριτήριο ομοιότητας), αλλά η πολυπλοκότητα του προβλήματος είναι μεγάλη και καθιστά την ακριβή λύση απαγορευτική για πρακτικές εφαρμογές. Κατά συνέπεια, οι αλγόριθμοι που χρησιμοποιούνται στην πράξη, βασίζονται σε ευριστικές τεχνικές, μερικές από τις οποίες θα περιγράψουμε παρακάτω.

Ίσως η πιο γνωστή και πετυχημένη σύγχρονη μέθοδος, είναι το **DALI**, (distance alignment matrix method), το οποίο είναι διαθέσιμο στη διεύθυνση http://ekhidna.biocenter.helsinki.fi/dali_server/start σαν υπηρεσία, αλλά και σαν αυτόνομη έκδοση (**DALIlite**). Η μέθοδος είναι από τις πιο παλιές (1993), αλλά έχει εμπλουτιστεί με νέα στοιχεία και πλέον λειτουργεί και με πολλαπλές δομές (πολλαπλή δομική στοίχιση). Η βασική ιδέα της μεθόδου είναι το «σπάσιμο» της δομής σε διαδοχικά εξαπεπτίδια και ο υπολογισμός ενός

πίνακα αποστάσεων από τα πρότυπα ενδομοριακών αλληλεπιδράσεων (contacts) που εμφανίζουν τα διαδοχικά εξαπεπτίδια. Τα στοιχεία δευτεροταγούς δομής στα οποία εμπλέκονται συνεχόμενα κατάλοιπα εμφανίζονται στην κύρια διαγώνιο. Όταν στους πίνακες αποστάσεων δύο πρωτεϊνών εμφανίζονται παρόμοια χαρακτηριστικά στην ίδια θέση, οι πρωτεΐνες θα έχουν το ίδιο δίπλωμα. Στο επόμενο βήμα πραγματοποιείται σύγκριση των επικαλυπτόμενων πινάκων 6x6 και ταύτισή τους με κάποιον αλγόριθμο βελτιστοποίησης (Holm & Rosenström, 2010). Το DALI έχει χρησιμοποιηθεί για την κατασκευή της βάσης FSSP (Families of Structurally Similar Proteins) στην οποία όλες οι γνωστές δομές έχουν στοιχιθεί για να δώσουν μια κατηγοριοποίηση των πρωτεϊνικών διπλωμάτων.

Ένα άλλο ιδιαίτερα πετυχημένο πρόγραμμα είναι το **CE** (Combinatorial extension), το οποίο είναι διαθέσιμο στη διεύθυνση http://source.rcsb.org/jfatcatserver/ceHome.jsp και χρησιμοποιείται από την PDB. Είναι και αυτό μια σχετικά παλιά μέθοδος η οποία εξελίσσεται, ενώ έχει αναπτυχθεί και εφαρμογή για πολλαπλές αλληλουχίες (**CE-MC**). Το CE μοιάζει στο DALI στο γεγονός ότι σπάει τη δομή σε μικρότερα κομμάτια, και μετά επιχειρεί να τα συναρμολογήσει για να κατασκευάσει τη στοίχιση. Συγκρίσεις των θραυσμάτων αυτών (aligned fragment pairs –AFPs) χρησιμοποιούνται για την κατασκευή του πίνακα ομοιότητας στον οποίο γίνεται τελικά η εύρεση του καλύτερου μονοπατιού με δυναμικό προγραμματισμό που καλείται να ενώσει με βέλτιστο τρόπο τα διαδοχικά AFP. Το μέτρο ομοιότητας ήταν αρχικά βασισμένο μόνο στην απόσταση, αλλά στις μετέπειτα εκδόσεις τροποποιήθηκε για να περιλαμβάνει πληροφορία για τη δευτεροταγή δομή, τους δεσμούς υδρογόνου, τις δίεδρες γωνίες κ.ο.κ. (Shindyalov & Bourne, 1998).

Το **SSAP** (Sequential Structure Alignment Program) είναι ίσως η πιο παλιά μέθοδος, η οποία χρησιμοποιεί διπλό δυναμικό προγραμματισμό για να στοιχίσει τις δομές (http://www.biochem.ucl.ac.uk/~orengo/ssap.html). Σε αντίθεση με τις άλλες μεθόδους, χρησιμοποποεί τον Cβ για τους υπολογισμούς, έτσι ώστε να λάβει υπόψη όχι μόνο τη θέση αλλά και τη δευτεροταγή δομή των αμινοξικών καταλοίπων. Στην αρχή η μέθοδος κατασκευάζει μια σειρά διανύσματα αποστάσεων μεταξύ των καταλοίπων και των γειτόνων τους που δεν είναι συνεχόμενα. Έπειτα, κατασκευάζει μια σειρά από πίνακες που περιέχουν τα διανύσματα των διαφορών των απστάσεων μεταξύ γειτόνων. Ο δυναμικός προγραμματισμός στη συνέχεια εφαρμόζεται σε κάθε πίνακα για να δώσει τις τοπικές στοιχίσεις, οι οποίες αθροίζονται ξανά σε ένα συνολικό πίνακα όπου και εφαρμόζεται ξανά δυναμικός προγραμματισμός για να δώσει την τελική στοίχιση (Taylor & Orengo, 1989). Όμοια με τις άλλες μεθόδους, έχει τροποποιηθεί για να δίνει και πολλαπλές στοιχίσεις ενώ χρησιμοποιείται για την ταξινόμηση των πρωτεϊνών στη βάση CATH.

Το **SSM** (http://www.ebi.ac.uk/msd-srv/ssm/), είναι ένας αλγόριθμος που αναπτύχθηκε στο EBI για να καλύψει τις ανάγκες της PDB. Έχει την ιδιαιτερότητα ότι βασίζεται σε μια εντελώς διαφορετική μέθοδο, αυτήν της ταύτισης των στοιχείων δευτεροταγούς δομής και όχι των ατομικών συντεταγμένων (Krissinel & Henrick, 2004). Η μέθοδος αυτή, διαισθητικά θυμίζει αυτό που περιγράψαμε παραπάνω, αλλά το μαθηματικοποιεί περισσότερο και πραγματοποιεί τη μοντελοποίηση σε επίπεδο δομής. Στην αρχή το πρόγραμμα εντοπίζει τα στοιχεία δευτεροταγούς δομής, και δημιουργεί μια γραφοθεωρητική αναπαράσταση της δομής με βάση αυτά. Κατόπιν, κάνει χρήση ενός γρήγορου αλγόριθμου για εύρεση ισομορφισμού γράφων για να συγκρίνει τις δύο αναπαραστάσεις των δομών και επιστρέφει τελικά στις ατομικές συντεταγμένες για να δώσει την τελική στοίχιση.

Το **MASS** (http://bioinfo3d.cs.tau.ac.il/MASS/), είναι επίσης μια μέθοδος πολλαπλής δομικής στοίχισης που βασίζεται στη στοίχιση των στοιχείων δευτεροταγούς δομής (Dror, Benyamini, Nussinov, & Wolfson, 2003). Δυο σημαντικά χαρακτηριστικά του MASS, είναι ότι πεώτον έχει την επιλογή να αγνοεί τη σειρά (είτε των στοιχείων δευτεροταγούς δομής στο πρώτο στάδιο, είτε των καταλοίπων στη συνέχεια), με συνέπεια να μπορεί να ανιχνεύσει κοινά δομικά στοιχεία που έχουν εμφανιστεί σε πρωτεΐνες λόγω συγκλίνουσας εξέλιξης αλλά δεν έχουν ομοιότητα στο δίπλωμα, και δεύτερον, ότι έχει τη δυνατότητα να ανιχνεύσει δομικά μοτίβα που εμφανίζονται μόνο σε ένα υποσύνολο των δομών.

Το **MAMMOTH** είναι μια άλλη πετυχημένη μέθοδος, η οποία έχει επίσης επεκταθεί και για πολλαπλές στοιχίσεις (http://ub.cbm.uam.es/software/online/mamothmult.php). Το MAMMOTH σπάει τη δομή σε επταπεπτίδια, και εφαρμόζει εκεί ένα διαφορετικό μέτρο απόστασης, το unit-vector root mean square (URMS), το οποίο έχει αρκετές επιθυμητές ιδιότητες, το μετατρέπει σε σκορ και μετά χρησιμοποιεί έναν αλγόριθμο δυναμικού προγραμματισμού για να βρει τη βέλτιστη στοίχιση των τμημάτων και μετά βρίσκει τη συνολική δομή που ικανοποιεί κάποιες προϋποθέσεις απόστασης. Μια ιδιαιτερότητα της μεθόδου είναι ότι υπολογίζει και στατιστική σημαντικότητα (Ortiz, Strauss, & Olmea, 2002).

Το **MUSTANG** (multiple structural alignment algorithm) είναι μια εφαρμογή που σχεδιάστηκε εξαρχής για πολλαπλή δομική στοίχιση (Konagurthu, Whisstock, Stuckey, & Lesk, 2006). Βασίζεται σε ιεραρχική πολλαπλή στοίχιση, με πολλαπλά βήματα που επανυπολογίζονται για βελτιστοποίηση. Στην αρχή

υπολογίζει τις περιοχές ομοιότητας και τις σκοράρει χρησιμοποιώντας μια «πρόχειρη» στοίχιση αλληλουχιών, την οποία βελτιστοποιεί στη συνέχεια. Κατόπιν, πραγματοποιεί τις κατά ζεύγη δομικές στοιχίσεις με χρήση του RMSD των Cα, και με βάση αυτές διορθώνει τα σκορ από τις συγκρίσεις αλληλλουχιών και στη συνέχεια προχωρά σε ιεραρχική στοίχιση των δομών (http://www.csse.monash.edu.au/~karun/Site/mustang.html).

Τέλος, θα πρέπει να αναφέρουμε και το **TM-align** (http://zhanglab.ccmb.med.umich.edu/TM-align/) το οποίο είναι μια από τις σχετικά πρόσφατες μεθόδους και έχει κερδίσει μεγάλη δημοφιλία τα τελευταία χρόνια (Zhang & Skolnick, 2005). Στο αρχικό του βήμα χρησιμοποιεί δυναμικό προγραμματισμό για να στοιχίσει τις ακολουθίες της δευτεροταγούς δομής, ενώ κατόπιν στοιχίζει τους άνθρακες της κύριας αλυσίδας (Cα) χρησιμοποιώντας έναν επαναληπτικό ευριστικό αλγόριθμο. Το ιδιαίτερο χαρακτηριστικό του, είναι ότι είναι εξαιρετικά γρήγορο (4 φορές πιο γρήγορο από το CE και 20 φορές πιο γρήγορο από το DALI), χωρίς όμως να υστερεί σε ποιότητα και αξιοπιστία.

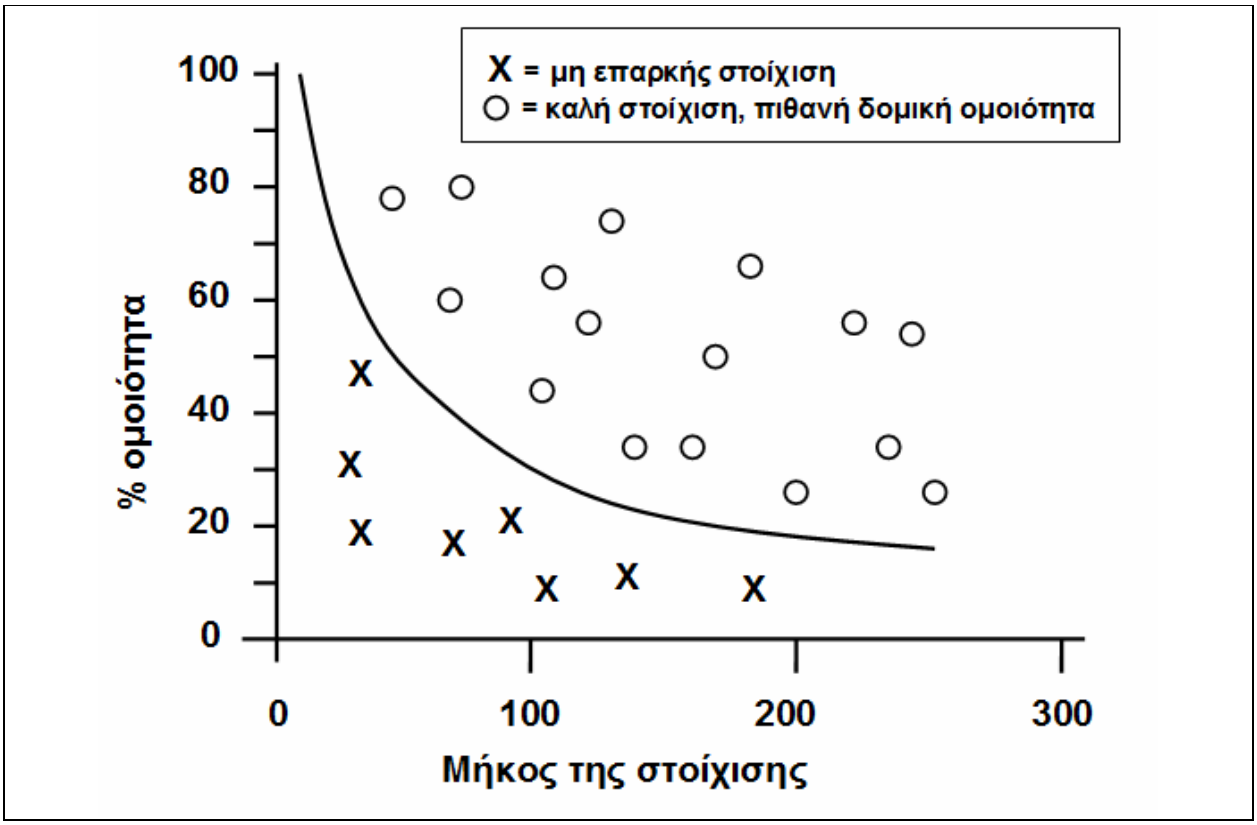
Όπως αναφέραμε παραπάνω, τα προγράμματα αυτά είναι μόνο ένα μικρό μέρος των προγραμμάτων που είναι διαθέσιμα στην επιστημονική κοινότητα. Στην αντίστοιχη σελίδα της Wikipedia (https://en.wikipedia.org/wiki/Structural_alignment_software), αναφέρονται δεκάδες αντίστοιχα εργαλεία, παρ' όλα αυτά, εδώ έγινε αναφορά σε αυτά που θεωρούνται πιο αξιόπιστα και χρησιμοποιούνται από τους περισσότερους ερευνητές. Στη βιβλιογραφία έχουν αναφερθεί μερικές μόνο συγκριτικές μελέτες, οι οποίες όμως έχουν το μειονέκτημα ότι κάθε φορά συγκρίνουν λίγα μόνο από τα διαθέσιμα εργαλεία ενώ χρησιμοποιούν και διαφορετικά σύνολα πρωτεϊνών και διαφορετικά κριτήρια αξιολόγησης (Kolodny, Koehl, & Levitt, 2005; Mayr, Domingues, & Lackner, 2007; Singh & Brutlag, 2000). Σε αυτό που συμφωνούν όλοι, είναι ότι τα περισσότερα από τα εργαλεία που αναφέραμε παραπάνω, αποδίδουν αρκετά καλά στις περισσότερες συνθήκες, ενώ όταν οι πρωτεΐνες έχουν μια στοιχειώδη ομοιότητα, οι στοιχίσεις τους είναι παρόμοιες. Γενικά, το DALI, το CE και το TM-align φαίνεται να είναι τα καλύτερα και τα πιο εύχρηστα, ενώ το τελευταίο είναι και ιδιαίτερα γρήγορο. Παρ' όλα αυτά, υπάρχουν ειδικές περιπτώσεις στις οποίες κάποιο άλλο εργαλείο μπορεί να ενδείκνυται καλύτερα, γι' αυτό και ο χρήστης θα πρέπει να έχει καλή γνώση του βιολογικού προβλήματος και να είναι ενήμερος έτσι ώστε να μπορεί να χρησιμοποιήσει και εναλλακτικές μεθόδους.

9.4. Πρόγνωση τρισδιάστατης δομής πρωτεϊνών

Ο τελικός σκοπός της υπολογιστικής μελέτης και της μοντελοποίησης των πρωτεϊνών, είναι η πρόγνωση της τρισδιάστατης δομής μιας πρωτεΐνης από την αλληλουχία της. Σε προηγούμενα κεφάλαια, είδαμε την πρόγνωση της δευτεροταγούς δομής, η οποία είναι ένα σχετικά εύκολο υποκατάστατο για την τελική πρόγνωση της τρισδιάστατης δομής. Μια τέτοια πρόβλεψη θα επιτρέψει τη διενέργεια πολλών πειραμάτων in silico (σχεδιασμός φαρμάκων, μελέτη της λειτουργίας της πρωτεΐνης, μελέτη αλληλεπιδράσεων κ.ο.κ.), για τα οποία σήμερα είναι απαραίτητη η διεξαγωγή των επίπονων και κοστοβόρων πειραμάτων προσδιορισμού της δομής. Επιπλέον δε, υπάρχουν και περιπτώσεις πρωτεϊνών που αποδεικνύονται δύσκολες στις μελέτες αυτές (δυσκολίες στην κρυστάλλωση κ.ο.κ.) και για τέτοιες περιπτώσεις, οι υπολογιστικές μελέτες είναι η μόνη διαθέσιμη εναλλακτική. Οι βασικές αρχές πίσω από τις μελέτες μοντελοποίησης είναι δύο, και είναι γνωστές από χρόνια: α) η αλληλουχία μιας πρωτεΐνης καθορίζει μονοσήμαντα τη δομή μιας πρωτεΐνης, και β) οι πρωτεϊνικές δομές συντηρούνται περισσότερο από τις αλληλουχίες. Μια άμεση συνέπεια των παραπάνω, είναι ότι δυο πρωτεΐνες με μεγάλη ομοιότητα σε επίπεδο αλληλουχίας έχουν κατά βάση παρόμοια δομή, αλλά είναι δυνατό, παρόμοια δομή να έχουν και πρωτεΐνες με μη ανιχνεύσιμη ομοιότητα (στο τελευταίο, σημαντικό ρόλο παίζει και η ύπαρξη περιορισμένου αριθμού πρωτεϊνικών διπλωμάτων).

Με όλα τα παραπάνω, μπορούμε να φανταστούμε ένα σενάριο στο οποίο έχουμε μια πρωτεΐνη με άγνωστη δομή, την οποία θέλουμε να προβλέψουμε, και έχουμε εντοπίσει μια ομόλογή της (μια πρωτεΐνη δηλαδή με μεγάλη ομοιότητα σε επίπεδο αλληλουχίας) η οποία διαθέτει γνωστή τρισδιάστατη δομή. Σε μια αντίστροφη πορεία από αυτήν που είχαμε ακολουθήσει στη δομική στοίχιση και την υπέρθεση, μπορούμε να φανταστούμε μια σειρά περιπτώσεων, στις οποίες η πρωτεΐνη με την άγνωστη δομή (target) μοντελοποιείται κάνοντας χρήση της δομής της ομόλογής της σαν καλούπι ή πρότυπο (template). Στην πιο απλή περίπτωση, αν λ.χ. διαφοροποιείται ένα αμινοξικό κατάλοιπο, είναι λογικό να υποθέσουμε ότι η υπόλοιπη δομή θα είναι ίδια και η μόνη διαφορά θα υφίσταται στο συγκεκριμένο κατάλοιπο και σε όσα βρίσκονται σε άμεση επαφή με αυτό (τέτοιες περιπτώσεις αν και τετριμμένες, μπορούν να έχουν σημασία όταν για παράδειγμα μελετάμε μια συγκεκριμένη αλλαγή στο ενεργό κέντρο ενός ενζύμου). Όσο πέφτει το επίπεδο ομοιότητας, αναμένουμε να υπάρχουν περισσότερες αλλαγές στο πρωτεϊνικό μόριο (διαφορετικές πλευρικές ομάδες, διαφορετικές

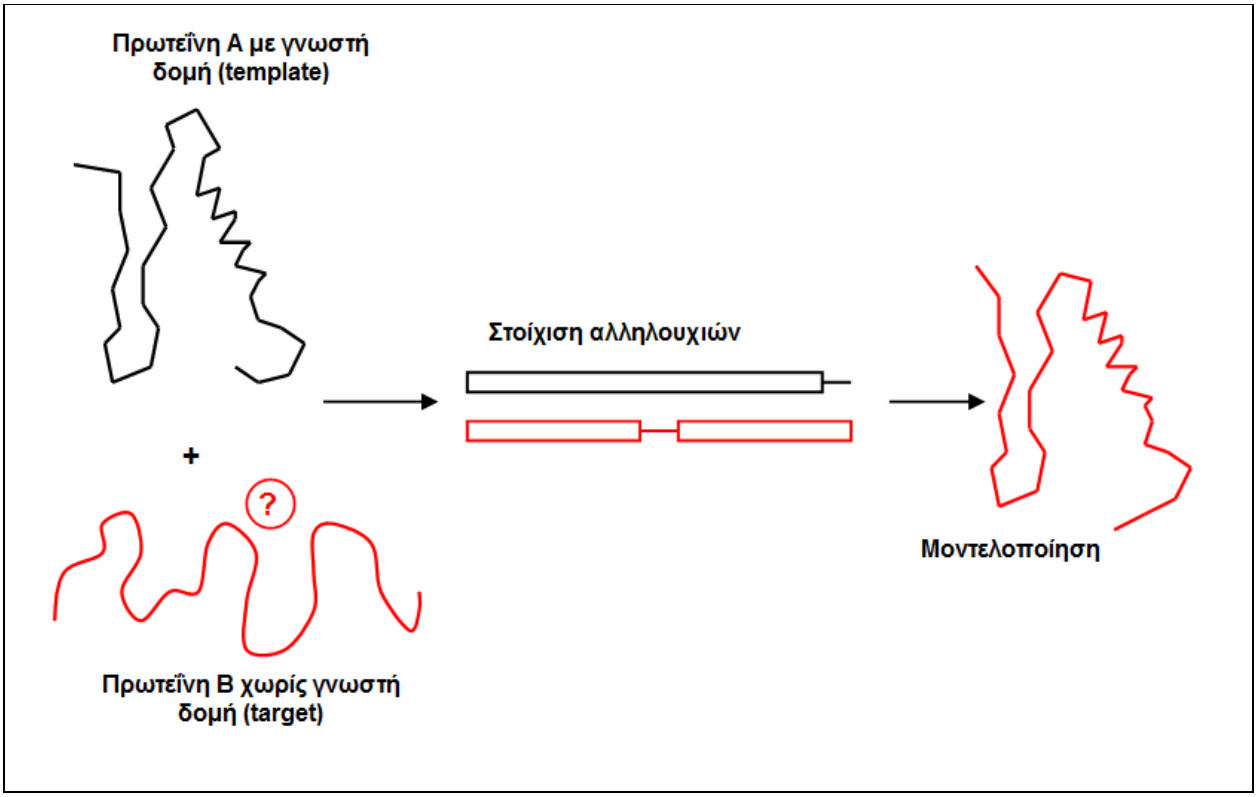
αλληλεπιδράσεις κ.ο.κ.), αλλά η γενικότερη δομή θα είναι περίπου ίδια (Εικόνα 9.9). Το ερώτημα είναι όμως, πού ακριβώς τίθεται το όριο κάτω από το οποίο δεν μπορούμε πλέον να υποθέσουμε ότι οι δυο πρωτεΐνες έχουν την ίδια δομή με σιγουριά; Όπως έχουμε δει στο κεφάλαιο της στοίχισης αλληλουχιών, η απάντηση στο πρόβλημα αυτό (η οποία στην ουσία απαντά στο πρόβλημα της στατιστικής σημαντικότητας μιας στοίχισης), εξαρτάται από δύο παράγοντες: από την ομοιότητα σε αμινοξέα της στοίχισης, και από το μήκος της στοίχισης, τα οποία μπορούν να παρασταθούν γραφικά και να σχηματιστεί εκεί μια γραμμή η οποία θα χωρίσει το επίπεδο των πιθανών στοιχίσεων σε αποδεκτές και μη αποδεκτές. Για μεγάλα ποσοστά ομοιότητας, το απαραίτητο μήκος της στοίχισης είναι μικρό. Για μικρότερα ποσοστά ομοιότητας όμως απαιτείται μεγαλύτερη στοίχιση. Γενικά, για ομοιότητες κάτω από το 30% απαιτούνται μεγάλες στοιχίσεις, ενώ για ομοιότητα κάτω από 20% δεν υπάρχει απλή μέθοδος στοίχισης για να δείξει καν αυτή την ομοιότητα (η περιοχή ονομάζεται και twilight zone).



Εικόνα 9.9: *Η σχέση της % ομοιότητας σε μια στοίχιση με το μήκος της στοίχισης καθορίζει την ποιότητα της στοίχισης.*

Μπορούμε, όπως είπαμε παραπάνω, να φανταστούμε ένα ολόκληρο φάσμα περιπτώσεων πρωτεϊνών που πιθανώς να συναντήσουμε σε μια προσπάθεια μοντελοποίησης της δομής. Κάποιες βρίσκονται στην «καλή» περιοχή, δηλαδή έχουν μια ξεκάθαρη ομοιότητα για μεγάλο μήκος της αλληλουχίας τους με πρωτεΐνες γνωστής δομής (σε διάφορα επίπεδα, 80%, 50%, 40% κ.ο.κ.), ενώ κάποιες εμφανίζουν πολύ μικρές ομοιότητες (<30%) για μικρά τμήματα τους ή δεν θα εμφανίζουν καμία ομοιότητα. Αυτές τις περιπτώσεις έρχονται να αντιμετωπίσουν οι διαφορετικές τεχνικές μοντελοποίησης της δομής, τις οποίες και αυτές πρέπει να τις αντιμετωπίσουμε σε ένα «συνεχές» φάσμα. Έτσι, για τις πρωτεΐνες της πρώτης κατηγορίας, υπάρχει η τεχνική που με απλά λόγια περιγράψαμε παραπάνω, η λεγόμενη *προτυποποίηση με βάση την ομολογία* (homology modelling). Για τις περιπτώσεις της δεύτερης κατηγορίας, υπάρχει η τεχνική της **ύφανσης** (threading), αλλά και η τεχνική της *προτυποποίηση εκ του μηδενός* (ab initio modelling), οι οποίες είναι τελείως διαφορετικές μεταξύ τους και θα παρουσιαστούν ξεχωριστά. Γενικά η ύφανση εφαρμόζεται σε πρωτεΐνες στόχους, που αφενός μεν δεν διαθέτουν ομόλογη πρωτεΐνη με γνωστή δομή, αφετέρου δε είναι δυνατόν να εντοπιστεί, με κάποια μέθοδο *αναγνώρισης διπλώματος*, το πρωτεϊνικό δίπλωμα στο οποίο ταιριάζουν (γι' αυτό και πολλές φορές οι όροι «αναγνώριση διπλώματος» και «ύφανση», χρησιμοποιούνται χωρίς διάκριση μεταξύ τους). Οι μέθοδοι ab initio πρόγνωσης, μπορούν φυσικά να εφαρμοστούν σε όλες τις

περιπτώσεις, αλλά επειδή είναι και οι πιο υπολογιστικά απαιτητικές, αλλά και αυτές με τη μεγαλύτερη επισφάλεια ως προς το αποτέλεσμα, χρησιμοποιούνται περισσότερο για τις πρωτεΐνες για τις οποίες ούτε καν κάποιο πιθανό δίπλωμα δεν μπορεί να αναγνωριστεί.



Εικόνα 9.10: *Σχηματική αναπαράσταση της προτυποποίησης με βάση την ομολογία.*

Προφανώς, οι 3 μεθοδολογίες παράγουν και τρισδιάστατα μοντέλα με διαφορετική αξιοπιστία και κατά συνέπεια κατάλληλα για διαφορετικές χρήσεις. Για παράδειγμα, η μοντελοποίηση με βάση την ομολογία, ειδικά όταν το πρότυπο έχει μεγάλη ομοιότητα με το στόχο (>70%) μπορεί να δώσει μοντέλα πολύ κοντά στην πραγματική δομή (RMSD<1 Å), μοντέλα δηλαδή που μπορούν να χρησιμοποιηθούν για λεπτομερείς δομικές μελέτες (για μελέτη ενζυμικών μηχανισμών κλπ). Όταν η ομολογία είναι μικρότερη, τα μοντέλα έχουν μεγαλύτερη απόκλιση από την πραγματική δομή και στην περίπτωση της ύφανσης μιλάμε πλέον για απόκλιση της τάξης των 2-4 Å. Τέλος, στην περίπτωση της ab initio πρόγνωσης, στην καλύτερη των περιπτώσεων δίνουν μοντέλα με RMSD της τάξης των 4-10 Å, τιμές που αρκούν για να δώσουν πληροφορίες μόνο για το γενικότερο σχήμα της πρωτεΐνης και όχι για λεπτομερείς αλληλεπιδράσεις της.

9.4.1. Μοντελοποίηση με βάση την ομολογία

Όπως είπαμε, η μοντελοποίηση (ή προτυποποίηση) με βάση την ομολογία, είναι η ενδεικνυόμενη μέθοδος για τις περιπτώσεις στις οποίες μια ομόλογη πρωτεΐνη με γνωστή δομή μπορεί να αναγνωριστεί εύκολα με μεθόδους στοίχισης αλληλουχιών (Εικόνα 9.10). Στη μέθοδο αυτή, η οποία μπορεί και να χαρακτηριστεί διαισθητικά και ως το αντίστροφο της υπέρθεσης δομών, η στοίχιση αλληλουχιών και μόνο αυτή είναι που κατευθύνει τη δημιουργία του μοντέλου. Τα βασικά βήματα της μοντελοποίησης με βάση την ομολογία είναι τα εξής:

- *Εύρεση του πρότυπου και πραγματοποίηση της στοίχισης*. Συνήθως χρησιμοποιούνται μέθοδοι όπως το BLAST και το FASTA, αν και κάποιες φορές η ολική στοίχιση είναι προτιμότερη, ειδικά αν υπάρχει ξεκάθαρη ομοιότητα. Επίσης, είναι πιθανό να αναγνωριστούν πολλά πρότυπα οπότε μπορούν να κατασκευαστούν πολλά εναλλακτικά μοντέλα. Πολλές φορές μια διόρθωση είναι απαραίτητη, ειδικά σε περιοχές με μικρή ομοιότητα. Η διόρθωση μπορεί να

γίνει είτε με χρήση πρότερης γνώσης είτε με τη χρήση αλγορίθμων πολλαπλής στοίχισης (χρησιμοποιώντας δηλαδή την πληροφορία και από άλλες ομόλογες).

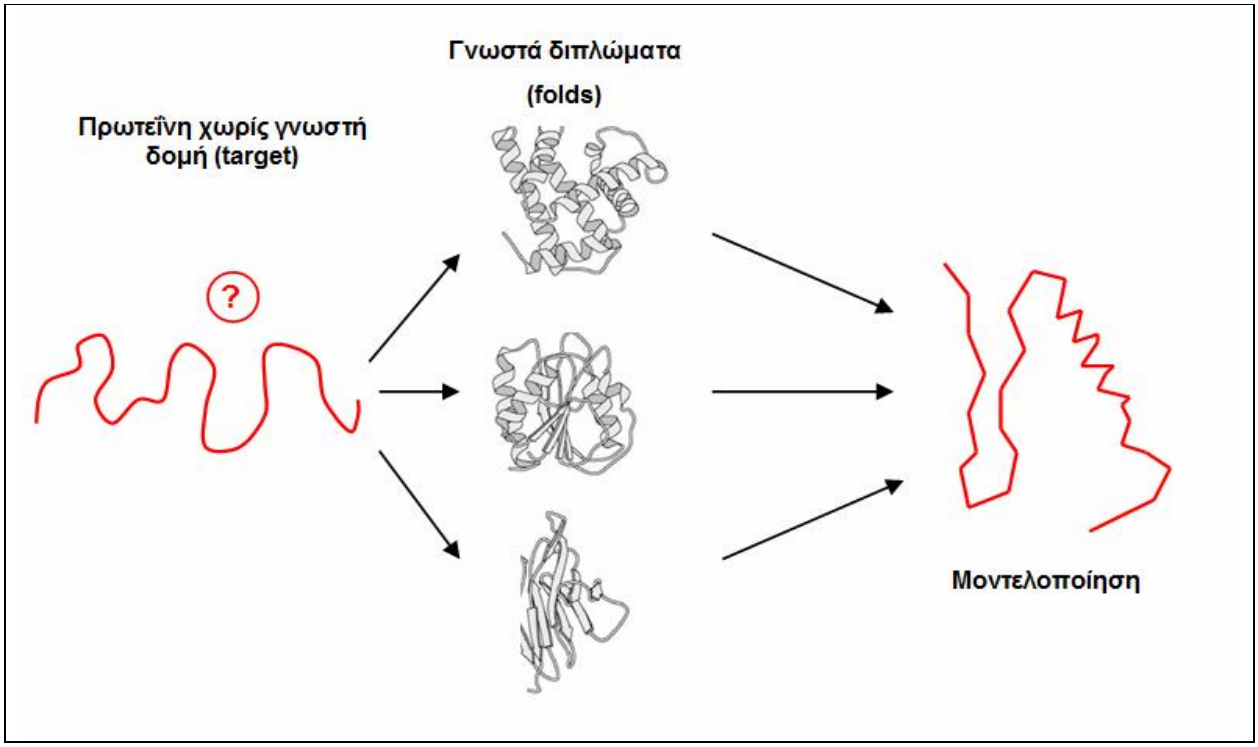
- *Κατασκευή του σκελετού της κύριας ανθρακικής αλυσίδας*. Στη φάση αυτή «χτίζεται» η νέα δομή ακολουθώντας το πρότυπο και τη στοίχιση. Σε περιοχές που τα κατάλοιπα είναι ίδια, η κατάσταση είναι απλή. Εκεί που υπάρχουν διαφορετικά κατάλοιπα τοποθετούνται μόνο τα άτομα του σκελετού (C, Cα, N, και O). Ένα πρόβλημα μπορεί να υπάρξει σε περιοχές της δομής του προτύπου που δεν έχουν προσδιοριστεί καλά, και πολλά προγράμματα το διορθώνουν χρησιμοποιώντας πολλαπλά πρότυπα.
- *Μοντελοποίηση των βρόχων και των πλευρικών αλυσίδων*. Στις περισσότερες περιπτώσεις στις στοιχίσεις θα υπάρχουν κενά. Όταν τα κενά βρίσκονται στην αλληλουχία του στόχου, θα πρέπει τα κατάλοιπα πριν και μετά το κενό να μετακινηθούν στην τελική δομή. Όταν όμως το κενό βρίσκεται στην αλληλουχία του προτύπου, δηλαδή έχει γίνει εισαγωγή στην αλληλουχία στόχο, τότε τα επιπλέον κατάλοιπα θα πρέπει να σχηματίσουν ένα βρόχο (loop), τη δομή του οποίου θα πρέπει να υπολογίσουμε. Επιπλέον δε, οι βρόχοι ούτως ή άλλως είναι ευκίνητες περιοχές οι οποίες είναι πολύ πιθανό να διαφέρουν αρκετά, ακόμα και σε πολύ όμοιες αλληλουχίες. Για να μοντελοποιηθεί σωστά ένας βρόχος, υπάρχουν δύο βασικές στρατηγικές, η πρώτη που μοιάζει περισσότερο με ύφανση και τη χρησιμοποιούν τα περισσότερα προγράμματα, στην οποία το πρόγραμμα ψάχνει στην PDB για περιοχές με παρόμοια κατάλοιπα, ενώ στη δεύτερη που είναι στην ουσία ab initio μέθοδος, γίνεται ελαχιστοποίηση ενέργειας για τον υπολογισμό της βέλτιστης δομής. Στην περίπτωση των πλευρικών ομάδων, το ζήτημα αφορά την ελεύθερη περιστροφή γύρω από το δεσμό Cα-Cβ. Κάποιες προσεγγίσεις στηρίζονται στην απλή μεταφορά αυτής της δομικής πληροφορίας από το πρότυπο, αλλά αυτό είναι επιτυχημένο μόνο για μεγάλη ομοιότητα (>35%) σε επίπεδο αλληλουχίας. Παράλληλα υπάρχουν και άλλες προσεγγίσεις που βασίζονται σε ανίχνευση όμοιων περιοχών στην PDB αλλά και σε ενεργειακούς υπολογισμούς.
- *Βελτιστοποίηση του μοντέλου*. Στο βήμα αυτό γίνεται βελτιστοποίηση όλης της δομής ταυτόχρονα, έτσι ώστε να ληφθούν υπόψη παράλληλα και ο προσανατολισμός των Cα αλλά και των βρόχων και των πλευρικών αλυσίδων (γιατί το ένα μπορεί να επηρεάζει το άλλο). Συνήθως το βήμα αυτό γίνεται επαναληπτικά και απαιτεί πιο προσεκτικά σχεδιασμένη συνάρτηση ενέργειας (σε σχέση με το προηγούμενο βήμα), ενώ η πιο απλή περίπτωση είναι να χρησιμοποιηθεί προσομοίωση μοριακής δυναμικής (molecular dynamics). Ανάλογα με τη συνάρτηση ενέργειας που μπορεί να χρησιμοποιηθεί, το βήμα αυτό μπορεί να είναι υπολογιστικά απαιτητικό.
- *Έλεγχος ποιότητας του μοντέλου.* Αφού το μοντέλο έχει κατασκευαστεί, είναι απαραίτητος ο έλεγχος για την επιβεβαίωσή του. Αυτός μπορεί να γίνει βασικά με δυο τρόπους, είτε με χρήση μοριακής δυναμικής με υπολογισμό της συνολικής ενέργειας το μορίου είτε με εμπειρικές μεθόδους που μετράνε την κανονικότητα διάφορων χαρακτηριστικών (μήκη δεσμών, αποστάσεις, γωνίες κ.ο.κ.). Η δεύτερη μέθοδος είναι πιο εύχρηστη καθώς επιτρέπει τον εντοπισμό των λαθών σε συγκεκριμένα σημεία κατά μήκος της αλληλουχίας.

Το πιο γνωστό αλλά και το πιο παλιό λογισμικό για μοντελοποίηση με βάση την ομολογία, είναι το **WHAT IF** (http://swift.cmbi.ru.nl/whatif/), το οποίο παρουσιάστηκε πρώτη φορά το 1987 από τον Gert Vriend (Vriend, 1990). Από τότε, συνεχίζει να εξελίσσεται και αποτελεί πλέον ένα ολοκληρωμένο περιβάλλον για τη μελέτη των πρωτεϊνικών δομών ενσωματώνοντας συνεχώς νέες λειτουργίες (οπτικοποίηση, υπέρθεση, 3D γραφικά, έλεγχο εγκυρότητας δομών, μοριακή δυναμική, υπολογισμούς φορτίων κ.ο.κ.), ενώ είναι διαθέσιμο ελεύθερα στην επιστημονική κοινότητα για διάφορες πλατφόρμες, αλλά και ως διαδικτυακή εφαρμογή. Το **MODELLER** (https://salilab.org/modeller/) είναι επίσης ένα κλασικό πακέτο λογισμικού για μοντελοποίηση με βάση την ομολογία (Eswar et al., 2006). Το MODELLER είναι ιδιαίτερα εύχρηστο καθώς στην πιο απλή εκδοχή ο χρήστης προμηθεύει ο ίδιος μια στοίχιση του στόχου με το πρότυπο (αυτό είναι ιδιαίτερα σημαντικό, όπως θα δούμε παρακάτω, καθώς μπορεί να κάνει χρήση και τεχνικών ύφανσης) και το λογισμικό υπολογίζει αυτόματα την τρισδιάστατη δομή. Το MODELLER χρησιμοποιεί την τεχνική των Sali και Blundell (Šali & Blundell, 1993), αλλά ενσωματώνει πολλές άλλες λειτουργίες όπως de novo μοντελοποίηση των βρόχων, βελτιστοποίηση του μοντέλου, πολλαπλή στοίχιση αλληλουχιών και δομών, ομαδοποίηση, αναζήτηση σε βάσεις δεδομένων, σύγκριση δομών κ.ο.κ. Παράλληλα, είναι και ελεύθερα διαθέσιμο για τις περισσότερες πλατφόρμες Η/Υ (Unix/Linux, Windows, και Mac) ενώ

έχει αναπτυχθεί και ένα παραθυρικό περιβάλλον για τη λειτουργία του, το **EasyModeller** (http://modellergui.blogspot.gr/). Τέλος, το **SWISS-MODEL** (http://swissmodel.expasy.org/) αποτελεί ίσως την πιο εύχρηστη εναλλακτική για μοντελοποίηση με βάση την ομολογία. Το εργαλείο λειτουργεί σαν μια αυτοματοποιημένη διαδικτυακή εφαρμογή, παρέχοντας πλήθος λειτουργιών όπως αυτόματη αναζήτηση στις βάσεις δεδομένων, έλεγχο ποιότητας για την επιλογή του καλύτερου πρότυπου, μοντελοποίηση με πολλαπλά πρότυπα και έλεγχο ποιότητας του προκύπτοντος μοντέλου (Biasini et al., 2014).

9.4.2. Αναγνώριση διπλώματος και ύφανση

Όπως είπαμε ήδη, η ύφανση ή αλλιώς αναγνώριση διπλώματος είναι μια τεχνική που χρησιμοποιείται σε περιπτώσεις κατά τις οποίες η πρωτεΐνη στόχος δεν έχει ανιχνεύσιμη ομοιότητα σε επίπεδο αλληλουχίας με κάποια πρωτεΐνη γνωστής δομής, αλλά μοιράζεται το ίδιο δίπλωμα με αυτές. Με τη διαδικασία αυτή, γίνεται έλεγχος αν η αλληλουχία μπορεί να ταιριάξει με κάποιο από τα γνωστά διπλώματα και μετά κατασκευάζεται η στοίχιση με το δίπλωμα αυτό (με τη δομή δηλαδή). Η βασική διαφορά από την μοντελοποίηση με βάση την ομολογία, στην οποία το πρότυπο το χειριζόμαστε ως αλληλουχία, είναι ότι στην ύφανση το πρότυπο χρησιμοποιείται σαν δομή. Οι μέθοδοι αναγνώρισης διπλώματος έχουν αποκτήσει μεγάλη δημοφιλία, λόγω της γνωστής αρχής ότι η δομή συντηρείται περισσότερο από την αλληλουχία και, κατά συνέπεια, από την παρατήρηση ότι ακόμα και διαφορετικές πρωτεΐνες μπορεί να έχουν παρόμοια δομή (ίδιο δίπλωμα). Επιπλέον δε, πιστεύεται γενικά ότι ο αριθμός των διπλωμάτων είναι πεπερασμένος και βρίσκεται κάπου ανάμεσα στο 1000-2000 (σήμερα πιστεύεται ότι έχουν εντοπιστεί 1300 διαφορετικά διπλώματα). Συνεπώς, μια πρωτεΐνη με μη ανιχνεύσιμη ομοιότητα σε επίπεδο αλληλουχίας, είναι παρ' όλα αυτά πολύ πιθανό να μπορεί να ταυτιστεί με κάποιο από τα ήδη γνωστά διπλώματα. Μια άλλη ενδιαφέρουσα παρατήρηση που πρέπει να γίνει, είναι ότι η αναγνώριση διπλώματος μοιάζει σε κάποιο βαθμό με τη δομική στοίχιση, και πράγματι κάποιες αλγοριθμικές τεχνικές έχουν χρησιμοποιηθεί και στις δύο μεθοδολογίες (π.χ. είδαμε παραπάνω τη στοίχιση με τη βοήθεια της δευτεροταγούς δομής).



Εικόνα 9.11: *Σχηματική αναπαράσταση της ύφανσης.*

Οι μεθοδολογίες που χρησιμοποιούνται στην αναγνώριση διπλώματος, εμφανίζουν τεράστια ετερογένεια αλλά χωρίζονται γενικά σε δύο μεγάλες κατηγορίες. Στην πρώτη κατηγορία ανήκουν οι μέθοδοι που μετατρέπουν τις τρισδιάστατες δομές σε μια μονοδιάστατη αλληλουχία (1D), σε ένα είδος προφίλ, και

μετά στοιχίζουν την πρωτεΐνη στόχο με αυτό το προφίλ συνήθως με χρήση κλασικού δυναμικού προγραμματισμού. Σαν προφίλ μπορεί να χρησιμοποιηθεί πληροφορία από τη δευτεροταγή δομή, την προσβασιμότητα στο διαλύτη κ.ο.κ., ενώ για να εφαρμοστεί η στοίχιση απαιτείται και η κατασκευή κάποιου είδους πίνακα για το σκορ, που να συνδέει τα γράμματα του νέου «αλφαβήτου» στο οποίο έχει μεταφραστεί η δομή, με τις αλληλουχίες αμινοξέων οι οποίες θα χρησιμοποιηθούν σαν στόχοι. Στη δεύτερη κατηγορία, χρησιμοποιείται κατευθείαν η τρισδιάστατη δομή (3D) και η ομοιότητα αξιολογείται με σύγκριση των ατομικών αποστάσεων. Συνήθως σε αυτή την περίπτωση, κατασκευάζεται ένα είδος σκορ που να μετράει τις πιθανές αλληλεπιδράσεις των ατόμων της πρωτεΐνης στην πιθανή δομή (δίπλωμα), ενώ ο δυναμικός προγραμματισμός έχει μεγαλύτερη πολυπλοκότητα. Όπως είναι φανερό, οι μέθοδοι της δεύτερης κατηγορίας χρησιμοποιούν περισσότερη πληροφορία, αλλά είναι οι πιο πολύπλοκες και κοστοβόρες από άποψη χρόνου. Η μέθοδος με τα προφίλ προτάθηκε πρώτη φορά από τους Bowie, Lüthy και Eisenberg το 1991 (Bowie, Luthy, & Eisenberg, 1991) ενώ ο ίδιος ο όρος ύφανση (threading) χρησιμοποιήθηκε για πρώτη φορά από τους Jones, Taylor και Thornton το 1992 (Jones, Taylort, & Thornton, 1992) και αρχικά αναφερόταν αποκλειστικά στη χρήση της τρισδιάστατης δομής. Σήμερα παρ' όλα αυτά, οι όροι ύφανση και αναγνώριση διπλώματος χρησιμοποιούνται συνήθως χωρίς διάκριση.

Ένα από τα πρώτα δημόσια διαθέσιμα εργαλεία για ύφανση, ήταν το **THREADER** του David Jones (διαθέσιμο στην ιστοσελίδα http://bioinf.cs.ucl.ac.uk/?id=747), που υλοποιούσε τον αλγόριθμο του διπλού δυναμικού προγραμματισμού του 1992 και πλέον βρίσκεται μετά από διάφορες προσθήκες στην έκδοση 3.5 (Jones, 1998). Ένα από τα πρώτα εργαλεία που εφάρμοζαν τη μέθοδο με τη μετατροπή της δομής σε ένα μονοδιάστατο προφίλ, ήταν το **PHDthreader** του Burkhart Rost (Rost, Schneider, & Sander, 1997), το οποίο αποτελεί τμήμα των εφαρμογών που καλύπτονται από τον server Predict Protein (www.predictprotein.org). Το PHDthreader κάνει χρήση του PHD για την πρόγνωση της δευτεροταγούς δομής και μετά στοιχίζει τις δομές (την παρατηρηθείσα για το πρότυπο, με την προβλεφθείσα για το στόχο). Παρόμοια στρατηγική χρησιμοποιεί και το **genTHREADER** (Jones, 1999), το οποίο βασίζεται στην πρόγνωση δευτεροταγούς δομής του PSI-PRED, αλλά και σε ένα επιπλέον βήμα με χρήση νευρωνικού δικτύου για να δώσει μια συνολική τιμή για την αξιοπιστία της μεθόδου, ενώ είναι διαθέσιμο μαζί με τις υπόλοιπες προγνώσεις του συγκεκριμένου server (http://bioinf.cs.ucl.ac.uk/psipred/). Το genTHREADER συνδυάζεται εύκολα με το MODELLER που είδαμε παραπάνω, για να δώσει τρισδιάστατα μοντέλα σε περίπτωση μη ικανοποιητικής ομοιότητας.

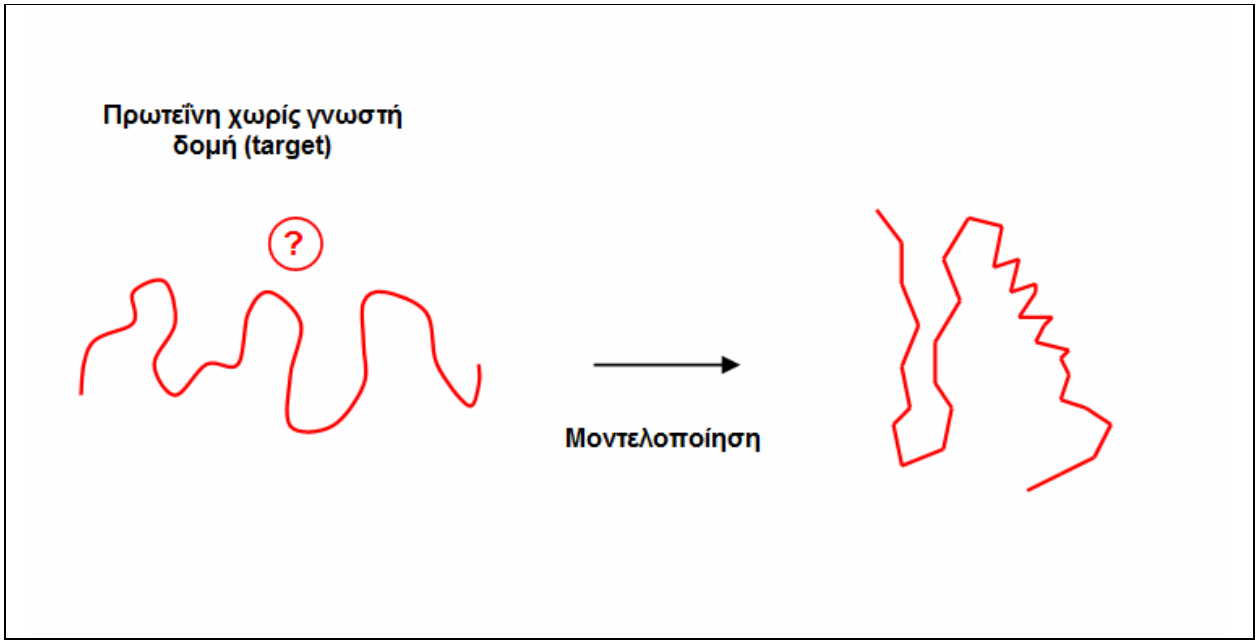
Μια σύγχρονη και ιδιαίτερα ικανοποιητική μέθοδος, είναι το **HHpred** (Söding, Biegert, & Lupas, 2005). Το HHpred βασίζεται σε μια ιδιαίτερα αποδοτική μέθοδο για στοίχιση και σύγκριση μεταξύ profile HMM (το HHsearch), κάτι που επιτρέπει ιδιαίτερα ευαίσθητες αναζητήσεις και εντοπισμό μακρινών ομολόγων (Söding, 2005). Η διαδικτυακή εφαρμογή δέχεται είσοδο είτε ακολουθία είτε μια πολλαπλή στοίχιση και επιτρέπει αναζήτηση σε διάφορες βάσεις (PDB, SCOP, PFAM, SMART κ.ο.κ.), τα αποτελέσματα επιστρέφονται πολύ γρήγορα σε κατανοητή μορφή, ενώ υπάρχει και διασύνδεση με το MODELLER για την παραγωγή του τρισδιάστατου μοντέλου (http://toolkit.tuebingen.mpg.de/hhpred). Το **Phyre2** είναι ένα άλλο παρόμοιο εργαλείο για αναγνώριση διπλώματος (http://www.sbg.bio.ic.ac.uk/phyre2). Η αρχική έκδοση, χρησιμοποιούσε έναν αλγόριθμο για στοίχιση profile-profile, βασισμένο σε PSSM, αλλά η νεότερη έκδοση χρησιμοποιεί και αυτή το HHsearch (Kelley, Mezulis, Yates, Wass, & Sternberg, 2015). Το Phyre2 ενσωματώνει διάφορες λειτουργίες όπως πρόγνωση δευτεροταγούς δομής με το PSI-RPED, πρόγνωση διαμεμβρανικών τμημάτων με το MEMSAT, πρόγνωση μη-κανονικών περιοχών με το DISOPRED, ενώ επιτρέπει πολλαπλές αναλύσεις όπως μελέτες προσδετών, μελέτες μη συνώνυμων πολυμορφισμών αλλά και ab initio προγνώσεις. Γενικά, αυτή η στρατηγική, να χρησιμοποιούνται σε ένα μόνο περιβάλλον, με απλό τρόπο χρήσης, όλες οι διαθέσιμες τεχνικές (πρόγνωση δευτεροταγούς δομής, μοντελοποίηση με βάση την ομολογία, αναγνώριση διπλώματος και ab initio προβλέψεις), αντιπροσωπεύει την κυρίαρχη τάση στις μεθόδους όπως θα δούμε και στις επόμενες παραγράφους.

Τελος, το **RaptorX** (http://raptorx.uchicago.edu/) και το **MUSTER** (http://zhang.bioinformatics.ku.edu/MUSTER) είναι δυο από τους πιο επιτυχημένους αλγόριθμους για ύφανση, καθώς δουλεύουν ικανοποιητικά ακόμα και σε περιπτώσεις κατά τις οποίες η ύπαρξη ομολόγων είναι περιορισμένη. Το RaptorX βασίζεται σε πιθανοθεωρητικά γραφικά μοντέλα και χρησιμοποιεί παράλληλα και την πληροφορία από τις δομές αλλά και από τις αλληλουχίες, ενώ χρησιμοποιεί και πληροφορία από όλα τα πιθανά πρότυπα για να χτίσει καλύτερα το μοντέλο (multiple template threading) (Peng & Xu, 2011). Το MUSTER χρησιμοποιεί δυναμικό προγραμματισμό, αλλά ενσωματώνει επίσης πολλαπλές πηγές πληροφορίας (δευτεροταγής δομή, προσβασιμότητα του διαλύτη, υδροφοβικότητα, πιθανές δίεδρες γωνίες κ.ο.κ.), ενώ κατασκευάζει το μοντέλο χρησιμοποιώντας διαφορετικό πρότυπο για κάθε

πρωτεϊνική περιοχή του στόχου (Wu & Zhang, 2008). Τέλος, υπάρχει και στην περίπτωση της ύφανσης η περίπτωση της συνδυαστικής πρόγνωσης με το **LOMETS** (http://zhanglab.ccmb.med.umich.edu/LOMETS/) το οποίο είναι ένας meta-server που χρησιμοποιεί 9 διαφορετικά εργαλεία (FFAS-3D, HHsearch, MUSTER, pGenTHREADER, PPAS, PRC, PROSPECT2, SP3, και SPARKS-X) για να παράγει έτσι μοντέλα μεγαλύτερης πιστότητας (Wu & Zhang, 2007). Το LOMETS, χρησιμοποιείται από το **I-TASSER** (http://zhanglab.ccmb.med.umich.edu/I-TASSER/about.html), το οποίο αποτελεί σήμερα την καλύτερη και πιο ολοκληρωμένη λύση στην πρόγνωση τριτοταγούς δομής, πετυχαίνοντας την πρώτη θέση στους τελευταίους διαγωνισμούς του CASP. Το I-TASSER αναγνωρίζει τα πρότυπα και με τα διάφορα τμήματα κατασκευάζει ένα μοντέλο με μια τεχνική που ονομάζεται replica exchange Monte Carlo simulations και οι βρόχοι μοντελοποιούνται ab initio. Όταν κανένα πρότυπο δεν βρεθεί, τότε το λογισμικό θα κατασκευάσει μοντέλο με μέθοδο ab initio για ολόκληρη την πρωτεΐνη. Στο τελευταίο στάδιο γίνεται βελτιστοποίηση του μοντέλου με προσομοιώσεις Το I-TASSER, ενσωματώνει επίσης μια σειρά βελτιώσεις που επιτρέπουν στο χρήστη να εισάγει δομική πληροφορία με τη μορφή περιορισμών, όπως τις επαφές των αμινοξέων, τη δευτεροταγή δομή κ.ο.κ. Οι περιορισμοί αυτοί μπορεί να είναι ιδιαίτερα χρήσιμοι σε περίπτωση που τα πρότυπα είναι λίγα ή η ποιότητα της στοίχισης δεν είναι καλή (Roy, Kucukural, & Zhang, 2010).

9.4.3. Ab initio και de novo πρόγνωση δομής

Στην πιο ακραία περίπτωση, η πρωτεΐνη στόχος δεν μπορεί να ταυτοποιηθεί ούτε με βάση την ομολογία αλλά ούτε και με βάση το δίπλωμα. Το πρόβλημα σε αυτή την περίπτωση, καταλήγει στο πασίγνωστο πρόβλημα του πρωτεϊνικού διπλώματος, (protein folding problem) της πρόγνωσης δηλαδή της τρισδιάστατης δομής απευθείας από την αμινοξική αλληλουχία. Το πρόβλημα αυτό, είναι στην ουσία ένα από τα μεγαλύτερα προβλήματα της σύγχρονης βιολογίας και δεκάδες ερευνητές έχουν ασχοληθεί (προφανώς, είναι ένα δύσκολο πρόβλημα καθώς έχει αποδειχτεί ότι είναι NP-complete). Γενικά, υπάρχουν δύο όροι για να περιγράψουν τις μεθόδους αυτές, και αν και πολλές φορές χρησιμοποιούνται αδιάκριτα μεταξύ τους, είναι καλό να πραγματοποιούμε το διαχωρισμό. Έτσι, με τον όρο ab initio πρόγνωση, παραδοσιακά αναφερόμαστε στην πρόγνωση με χρήση μόνο των βασικών αρχών της φυσικής (αλληλεπιδράσεις ατόμων και υπολογισμοί ενέργειας). Από την άλλη, ο όρος de Novo πρόγνωση, είναι κάπως πιο γενικός και αναφέρεται σε όλες τις μεθόδους που επιχειρούν πρόγνωση χωρίς τη χρήση προτύπου με γνωστή δομή. Γενικά πάντως δεν υπάρχει απόλυτη συμφωνία ως προς το σε ποια ακριβώς κατηγορία κατατάσσεται κάθε μέθοδος, ειδικά εφόσον οι περισσότερες από αυτές χρησιμοποιούν συνδυασμό μεθοδολογιών.



Εικόνα 9.12: *Σχηματική αναπαράσταση της ab initio πρόγνωσης της δομής*

.

Γενικά το θέμα της πρόγνωσης της τρισδιάστατης δομής των πρωτεϊνών έχει απασχολήσει τους επιστήμονες για δεκαετίες. Οι εργασίες του Anfinsen έδειξαν μεν ότι οι αλληλουχίες των πρωτεϊνών καθορίζουν μονοσήμαντα την τρισδιάστατη δομή οδηγώντας στη δομή με την ελάχιστη ενέργεια, αλλά το παράδοξο του Levinthal έδειξε με ξεκάθαρο τρόπο ότι οι πιθανές διαμορφώσεις μιας πρωτεΐνης δεν είναι δυνατό να δοκιμαστούν όλες. Για παράδειγμα, αν μια πρωτεΐνη έχει 100 αμινοξέα (ένας κάπως μικρός αριθμός), υπάρχουν 99 πεπτιδικοί δεσμοί και κατά συνέπεια 198 διαφορετικές γωνίες φ και ψ οι οποίες μπορούν να περιστραφούν ελεύθερα. Αν υποθέσουμε ότι κάθε γωνία έχει μόνο 3 πιθανές τιμές (πάλι ένας μετριοπαθής υπολογισμός), τότε οι πιθανές διαμορφώσεις ολόκληρου του πρωτεϊνικού μορίου είναι 3^{198}, ένας αριθμός εξωπραγματικός. Αν η πρωτεΐνη έπρεπε να δοκιμάσει με τυχαίες κινήσεις όλες τις πιθανές διαμορφώσεις, τότε δεν θα προλάβαινε να διπλωθεί σωστά ακόμα και αν περιμέναμε ως το... τέλος του σύμπαντος προσπαθώντας. Στην πράξη βέβαια, οι πρωτεΐνες διπλώνονται σε χρόνους της τάξης των microsecond ή millisecond, κάτι που σημαίνει ότι λειτουργεί κάποιος άλλος μηχανισμός (έχουν προταθεί διάφοροι τρόποι, όπως το μοντέλο της υδροφοβικής κατάρρευσης, το μοντέλο του σχηματισμού πυρήνων δευτεροταγούς δομής κ.ο.κ.). Σε κάθε περίπτωση, όλη αυτή η 'συζήτηση', εκτός από θεωρητική σημασία, έως και σήμερα βασικό ρόλο στις προσπάθειες ab initio ή de novo πρόγνωσης της δομής.

Γενικά με βάση τα παραπάνω, οι προσπάθειες ab initio πρόγνωσης, είναι (παρόλες τις αλγοριθμικές επινοήσεις), ιδιαίτερα απαιτητικές υπολογιστικά και για χρόνια ήταν περιορισμένες σε μικρές πρωτεΐνες (μέχρι 50 αμινοξέα μήκος), και ακόμα και σε αυτές τις περιπτώσεις τα αποτελέσματα χρειάζονταν πολύ χρόνο. Τα τελευταία χρόνια τόσο η αύξηση της υπολογιστικής ισχύος, αλλά και νέες αλγοριθμικές τεχνικές επέτρεψαν την πρόγνωση με σχετική ακρίβεια, για μεγαλύτερες πρωτεΐνες, και σε εύλογο χρονικό διάστημα (μερικές ώρες ή μέρες). Τα γενικά θέματα που έχει να αντιμετωπίσει μια τέτοια μέθοδος είναι τρία:

- *Η αναπαράσταση της πρωτεϊνικής δομής*. Στην ιδανική περίπτωση θα έπρεπε στους υπολογισμούς να λαμβάνουν μέρος όλα τα άτομα της πρωτεΐνης, αλλά κάτι τέτοιο είναι απαγορευτικό από πλευράς υπολογιστικής ισχύος. Έτσι, έχουν χρησιμοποιηθεί διαφορετικές προσεγγίσεις: από την απλή χρήση μόνο του Cα, την προσθήκη του Cβ, μέχρι και σύνθετες μετρήσεις στις οποίες ολόκληρη η πλευρική ομάδα αντικαθίσταται από ένα σημείο με τη συνολική μάζα στο κέντρο βάρους. Οι επιτρεπτές γωνίες είναι επίσης ένας σημαντικός παράγοντας σε αυτό το σημείο. Έτσι, κάποιες μέθοδοι επιτρέπουν μόνο προκαθορισμένες γωνίες φ και ψ, ενώ άλλες υπολογίζουν τη δομή κομματιών μήκους 6-7 αμινοξέων για να ελαττώσουν ακόμα περισσότερο το χρόνο.
- *Ο υπολογισμός της ενέργειας*. Το κομμάτι αυτό αφορά το πώς θα αξιολογηθεί μια δομή ως «καλή». Θα πρέπει δηλαδή να υπάρχει ένα κριτήριο που να ξεχωρίζει τις δομές ελάχιστης ενέργειας. Η πιο προφανής λύση εδώ, είναι η χρήση καθαρά φυσικοχημικών τεχνικών κατά τις οποίες υπολογίζονται οι ελκτικές και απωστικές δυνάμεις μεταξύ όλων των ατόμων, σε μια προσπάθεια να μιμηθούμε το δίπλωμα των πρωτεϊνών στη φύση. Οι μεθοδολογίες αυτής της κατηγορίας περιλαμβάνουν τα πεδία AMBER, CHARMM, UNRES και ASTRO-FOLD. Η άλλη εναλλακτική είναι να χρησιμοποιηθεί μια εμπειρική συνάρτηση η οποία θα έχει προκύψει από στατιστικές μετρήσεις (τέτοιες συναρτήσεις χρησιμοποιούνται από το ROSSETA και το TASSER/I-TASSER).
- *Η στρατηγική αναζήτησης*. Αυτό το σημείο αναφέρεται στο πώς θα γίνει η αναζήτηση στο χώρο των πιθανών διαμορφώσεων για την εύρεση της δομής με την ελάχιστη ενέργεια. Η πιο συνηθισμένη μέθοδος εδώ, είναι η προσομοίωση Monte Carlo, αλλά έχουν χρησιμοποιηθεί και άλλες στατιστικές τεχνικές όπως το Simulated Annealing (προσομοίωση ανώπτησης), αλλά και τεχνικές της τεχνητής νοημοσύνης όπως οι γενετικοί αλγόριθμοί. Μια άλλη μεγάλη κατηγορία μεθόδων είναι η Μοριακή Δυναμική (Molecular Dynamics), κατά την οποία επιλύονται οι εξισώσεις κίνησης του Νεύτωνα και προσομοιώνεται η κίνηση των ατόμων στο χρόνο. Η τεχνική αυτή είναι η πιο αξιόπιστη, αλλά καθώς συνήθως συνδυάζεται με συνάρτηση ενέργειας φυσικοχημικού τύπου, απαιτεί πάρα πολλούς υπολογισμούς. Κατά συνέπεια, είναι εφαρμόσιμη περισσότερο σε περιπτώσεις που μας ενδιαφέρει η διαδικασία διπλώματος μιας πρωτεΐνης. Μοριακή δυναμική επίσης χρησιμοποιείται γενικά για τη μοντελοποίηση των βρόχων και για τη βελτιστοποίηση ενός ήδη κατασκευασμένου μοντέλου.

Τα πιο γνωστά και παλιά προγράμματα μοριακής δυναμικής, είναι το **CHARMM** (http://www.charmm.org/) και το **AMBER** (http://ambermd.org/) τα οποία συμβαίνει να αντιστοιχούν και

στις δύο πιο γνωστές κατηγορίες δυναμικών πεδίων για υπολογισμούς μοριακής δυναμικής. Το CHARMM (Chemistry at HARvard Macromolecular Mechanics) αποτελεί ένα μεγάλο συνεργατικό πρόγραμμα με πολλούς ερευνητές, υπό την καθοδήγηση του Martin Karplus στο Harvard (Brooks et al., 2009). Είναι το παλιότερο πρόγραμμα μοριακής δυναμικής και έχει εξελιχθεί με τα χρόνια έτσι ώστε να παρέχει πολλές διαφορετικές λειτουργίες και επιλογές (προσομοίωση, ελαχιστοποίηση ενέργειας μιας δεδομένης δομής, και υπολογισμοί μοριακής δυναμικής). Το AMBER (Assisted Model Building with Energy Refinement) ξεκίνησε αρχικά σαν μια άλλη οικογένεια δυναμικών πεδίων που αναπτύχθηκαν από τον Peter Kollman στο University of California και το ομώνυμο πακέτο αναπτύχθηκε σαν υλοποίηση αυτών των υπολογισμών (Case et al., 2005). Ένα μεγάλο μειονέκτημα και των δύο πακέτων είναι ότι δεν είναι ελεύθερα διαθέσιμα. Μια επιλογή για αντίστοιχο λογισμικό ανοιχτού κώδικα, είναι το **GROMACS** (http://www.gromacs.org). Το GROMACS παρέχει πολλές δυνατότητες για μοντελοποίηση πολλών κατηγοριών βιομορίων με χρήση διαφορετικών πεδίων, ενώ παρέχει και δυνατότητες παράλληλης επεξεργασίας ακόμα και σε συστήματα Windows (Pronk et al., 2013). Αυτό που πρέπει να τονιστεί βέβαια, είναι ότι τα προγράμματα αυτά δεν μπορούν να χρησιμοποιηθούν σε πρακτικές εφαρμογές για ab initio πρόγνωση της δομής πρωτεϊνών, αλλά κυρίως για βελτιστοποίηση μιας υπάρχουσας δομής, για τη μελέτη της διαδικασίας του διπλώματος και για μελέτη των αλληλεπιδράσεων με άλλα μόρια.

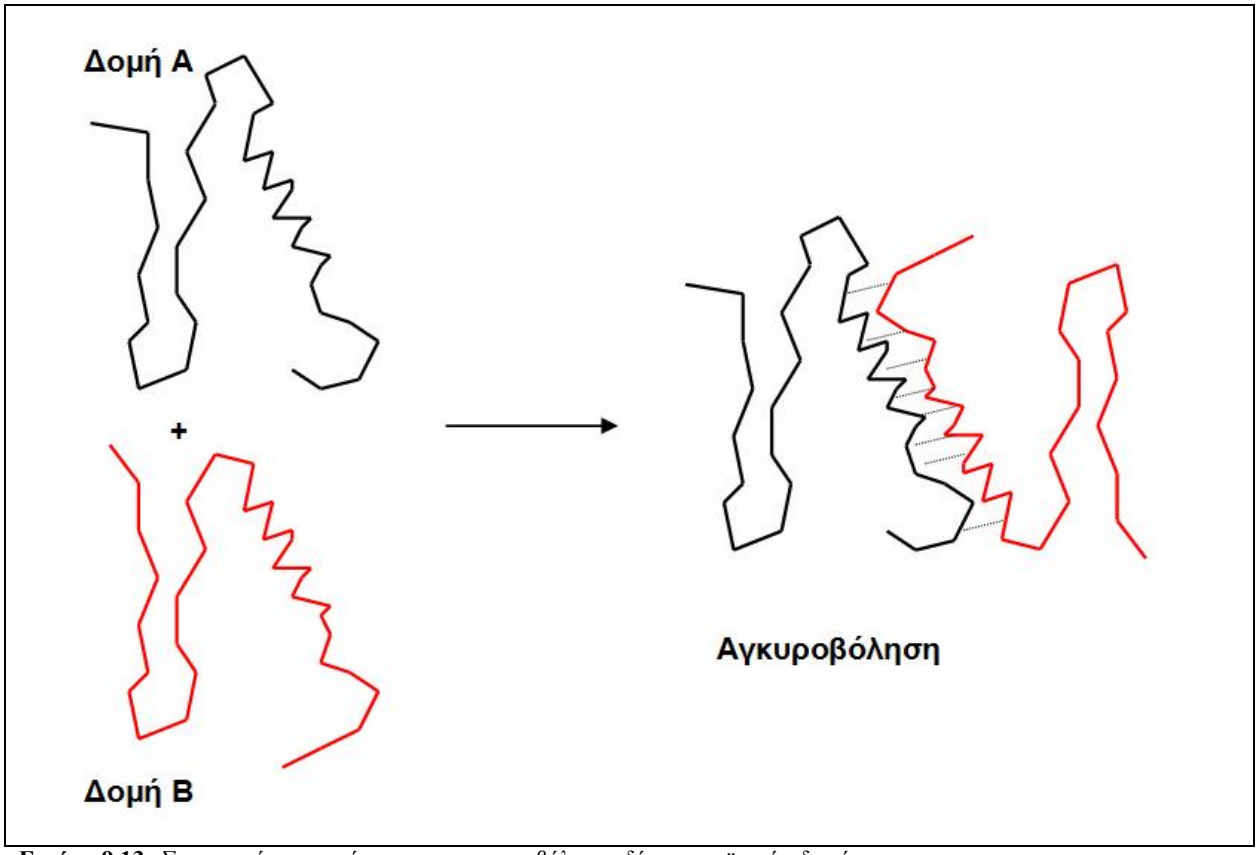
Το πιο γνωστό από τα προγράμματα για ab initio/de novo πρόγνωση τρισδιάστατης δομής, είναι το **ROSETTA** (http://robetta.bakerlab.org/). Το ROSETTA αρχικά αναγνωρίζει τις πρωτεϊνικές περιοχές, και στη συνέχεια τις μοντελοποιεί με μια γρήγορη ab initio μεθοδολογία που χρησιμοποιεί τα μικρότερα τμήματα (κυρίως 9μερή) από γνωστές δομές της PDB. Η μεθοδολογία αυτή βασίζεται σε μια ιδέα των Bowie και Eisenberg από το 1994. Το αρχικό μοντέλο κατασκευάζεται με χρήση μόνο της κύριας ανθρακικής αλυσίδας και των Cβ ενώ στη συνέχεια κάποια από τα καλύτερα (από άποψη ενέργειας) μοντέλα υφίστανται βελτιστοποίηση με όλα τα άτομα παρόντα, κάνοντας χρήση προσομοίωσης Monte Carlo και μιας στατιστικής φύσεως συνάρτησης ενέργειας (Rohl, Strauss, Misura, & Baker, 2004).

Όπως αναφέραμε ήδη, το **I-TASSER** (http://zhanglab.ccmb.med.umich.edu/I-TASSER/about.html) είναι μια εφαρμογή που πραγματοποιεί και ύφανση αλλά και ab initio μοντελοποίηση όταν δεν μπορεί να εντοπίσει δομές με παρόμοιο δίπλωμα. Το I-TASSER, είναι σήμερα η καλύτερη και πιο ολοκληρωμένη λύση στην πρόγνωση τριτοταγούς δομής, όπως πιστοποιείται από την πρώτη θέση που καταλαμβάνει στους τελευταίους διαγωνισμούς του CASP αλλά και σε εμπειρικές μελέτες αξιολόγησης (Helles, 2008). Το μεγάλο του πλεονέκτημα, είναι εκτός από την ακρίβεια στην πρόβλεψη, η μεγάλη ταχύτητα στους υπολογισμούς. Και το I-TASSER και το Rosetta χρησιμοποιούν προσομοίωση Monte Carlo (αν και με διαφορετικές παραλλαγές), συναρμογή τμημάτων και στατιστικής φύσεως συναρτήσεις ενέργειας, αλλά διαφέρουν στην αναπαράσταση της δομής και στις αποδεκτές δίεδρες γωνίες. Άλλες δημόσια διαθέσιμες μέθοδοι λιγότερο γνωστές είναι το **ePROPAINOR** (http://www.math.iitb.ac.in/epropainor) και το **PROTinfo** (http://ram.org/compbio/protinfo/), οι οποίες όμως δεν είναι τόσο επιτυχημένες (πάντα σε σχέση με το I-TASSER και το ROSETTA). Επίσης, υπάρχουν μια σειρά από μέθοδοι όπως το **QUARK** (http://zhanglab.ccmb.med.umich.edu/QUARK/), το **CABSfold** (http://biocomp.chem.uw.edu.pl/CABSfold/), το **PEP-FOLD** (http://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD/) και το **BHAGEERATH** (http://www.scfbio-iitd.res.in/bhageerath/index.jsp), οι οποίες όμως ενδείκνυνται περισσότερο για πεπτίδια και μικρές πρωτεΐνες (<100 αμινοξέα), καθώς ο υπολογιστικός χρόνος για μεγαλύτερους υπολογισμούς είναι απαγορευτικός.

Τέλος, αξίζει μια ειδική αναφορά στα κατανεμημένα (distributed) συστήματα ab initio πρόγνωσης. Τέτοιου είδους εφαρμογές, ξεκίνησαν με το **ROSETTA@home** (http://boinc.bakerlab.org/rosetta/) και το **Folding@home** (http://folding.stanford.edu/). Με τις μεθοδολογίες αυτές, ο χρήστης που έχει εγκαταστήσει την ειδική εφαρμογή «δανείζει» υπολογιστικό χρόνο από τον υπολογιστή του όταν αυτός δεν λειτουργεί, με σκοπό να βοηθήσει στην επίλυση του προβλήματος του διπλώματος «δύσκολων» πρωτεϊνών. Πολλοί από τους επιστήμονες που είχαν εμπλοκή στο σχέδιο του ROSETTA@home, αποφάσισαν αργότερα να εμπλέξουν ακόμα περισσότερους χρήστες και να αναπτύξουν ένα παιχνίδι που θα προσομοιώνει το δίπλωμα των πρωτεϊνών. Η ιδέα ήταν να χρησιμοποιηθούν οι ικανότητες αναγνώρισης προτύπων που διαθέτει ο ανθρώπινος εγκέφαλος, και να εφαρμοστούν σε παρόμοιες δύσκολες περιπτώσεις. Έτσι αναπτύχθηκε το **FOLDit** (http://fold.it/portal/) στο οποίο οι χρήστες σε ένα είδος παιχνιδιού στον Η/Υ κατασκευάζουν μοντέλα τρισδιάστατης δομής γνωστών πρωτεϊνών προτείνοντας τη δομή με το κατάλληλο δίπλωμα (δηλαδή, με τη μικρότερη ενέργεια). Η ιδέα είναι ότι το σύστημα μπορεί να «εκπαιδευτεί» με τις λύσεις που προτείνει ο ανθρώπινος εγκέφαλος, έτσι ώστε ένα αυτοματοποιημένο παρόμοιο σύστημα να μπορέσει να υλοποιηθεί αργότερα.

9.5. Αγκυροβόληση

Με τον όρο αγκυροβόληση ή ελλιμενισμό (docking) εννοούμε τη διαδικασία με την οποία υπολογίζουμε ή προβλέπουμε τον προτιμώμενο προσανατολισμό ενός μορίου σε σχέση με ένα άλλο όταν σχηματίζουν ένα σταθερό σύμπλοκο. Στη διαδικασία αυτή, γίνεται η υπόθεση ότι το σταθερό αυτό σύμπλοκο βρίσκεται σε μια διαμόρφωση ελάχιστης ενέργειας. Το σύμπλοκο το οποίο θα επιχειρήσουμε να μοντελοποιήσουμε με τη διαδικασία της αγκυροβόλησης μπορεί να είναι μεταξύ δύο πρωτεϊνών (Bonvin, 2006; Gray, 2006; Sternberg, Gabb, & Jackson, 1998), αλλά και μεταξύ μιας πρωτεΐνης και ενός μικρού μορίου το οποίο μπορεί να είναι ορμόνη, φάρμακο, αναστολέας, βιταμίνη κ.ό.κ. (Taylor, Jewsbury, & Essex, 2002). Φυσικά, υπάρχουν και περιπτώσεις αλληλεπιδράσεων DNA-πρωτεϊνών αλλά και DNA-μικρών μορίων. Η γνώση αυτή, μπορεί να είναι χρήσιμη στο να κατανοήσουμε το βιολογικό μηχανισμό της λειτουργίας της πρωτεΐνης, την ένταση και την ισχύ της δέσμευσης του μικρού μορίου, το μηχανισμό λειτουργίας, αλλά και τον μηχανισμό με τον οποίο αλληλεπιδρούν δυο πρωτεΐνες είτε σαν ένζυμο-υπόστρωμα, είτε σαν υποδοχέας-προσδέτης, αλλά και γενικότερα στη μελέτη των πρωτεϊνικών αλληλεπιδράσεων και της τεταρτοταγούς δομής. Ειδικά στην περίπτωση των μικρών μορίων, η αγκυροβόληση βρίσκει πολλές εφαρμογές στο σχεδιασμό νέων φαρμάκων, και λόγω της σημασίας αυτής της διαδικασίας, στη φαρμακευτική βιομηχανία, έχει δοθεί μεγάλη ώθηση στο πεδίο από τέτοιες αντίστοιχες μελέτες (Alvarez, 2004).



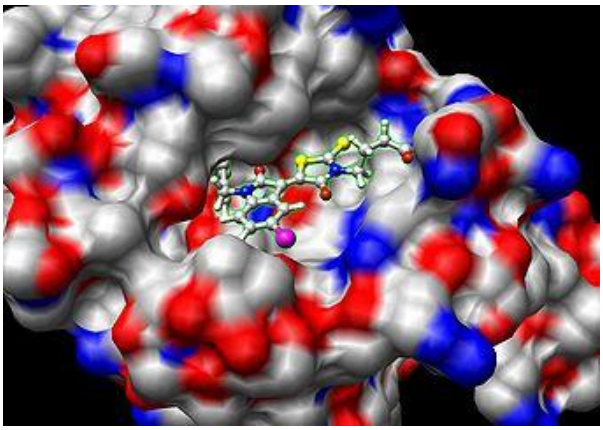
Εικόνα 9.13: *Σχηματική αναπαράσταση της αγκυροβόλησης δύο πρωτεϊνικών δομών.*

Το πρόβλημα της αγκυροβόλησης μπορούμε να το δούμε ανατρέχοντας στις γνωστές θεωρίες για τη δράση των ενζύμων και των πρωτεϊνών γενικότερα. Έτσι, η πιο απλή προσέγγιση κάνει λόγο για το μοντέλο «κλειδιού-κλειδαριάς», σύμφωνα με το οποίο οι επιφάνειες των δύο πρωτεϊνών είναι συμπληρωματικές ή το ενεργό κέντρο του ενζύμου είναι συμπληρωματικό σαν γεωμετρικό σχήμα με το υπόστρωμα (ή, του υποδοχέα με τον προσδέτη κ.ο.κ.). Σύμφωνα με αυτή τη θεωρία, αναπτύχθηκαν οι πρώτες μέθοδοι αγκυροβόλησης, οι λεγόμενες μέθοδοι «αγκυροβόλησης σταθερού σώματος» (rigid docking), σύμφωνα με τις οποίες οι τρισδιάστατες δομές του κάθε μορίου δεν αλλάζουν (δηλαδή τα άτομά τους δεν αλλάζουν καθόλου τη

σχετική τους θέση), αλλά απλά μετακινούνται για να βρεθεί η επιφάνεια επαφής. Παρ' όλα αυτά, ξέρουμε ότι το μοντέλο αυτό δεν είναι επαρκές καθώς σε πολλές περιπτώσεις η πρόσδεση επηρεάζει (σε μικρότερο ή μεγαλύτερο βαθμό) τη διαμόρφωση του κάθε μορίου (το μοντέλο της «επαγόμενης προσαρμογής»). Αυτό οδήγησε σε πιο σύνθετες τεχνικές αγκυροβόλησης, τις λεγόμενες μεθοδολογίες «ευέλικτης αγκυροβόλησης» (flexible docking) στις οποίες η τρισδιάστατη δομή των μορίων αλλάζει (έστω και ελάχιστα) για να επιτευχθεί η καλύτερη δυνατή αναγνώριση.

Από άποψη υπολογιστικής μεθοδολογίας, και σύμφωνα με τα παραπάνω, μπορούμε να διακρίνουμε, δύο κατηγορίες προσεγγίσεων στην αγκυροβόληση. Στην πρώτη περίπτωση, έχουμε τις προσεγγίσεις που βασίζονται στη συμπληρωματικότητα του σχήματος. Στις μεθοδολογίες αυτής της κατηγορίας, τα εμπλεκόμενα βιομόρια αντιμετωπίζονται ως τρισδιάστατα σχήματα και η συμπληρωματικότητα επιτυγχάνεται με μετακίνηση των δομών με τρόπο που να τις κάνει να συμπίπτουν όσο το δυνατό καλύτερα. Οι μεθοδολογίες αυτής της κατηγορίας είναι γρήγορες και σταθερές, αλλά με τις απλουστεύσεις που κάνουν δεν μπορούν να δώσουν τα βέλτιστα αποτελέσματα. Έτσι, χρησιμοποιούνται συνήθως στα αρχικά στάδια των μελετών για πιθανούς στόχους για φάρμακα, ούτως ώστε να γίνει μια γρήγορη διαλογή των πιθανών στόχων. Λόγω του ότι βασίζονται κυρίως στη γεωμετρική αναπαράσταση των δομών, στις μεθοδολογίες αυτές χρησιμοποιείται ιδιαίτερα η προσέγγιση των «φαρμακοφόρων».

Στη δεύτερη κατηγορία μεθόδων, ανήκουν οι μέθοδοι που βασίζονται στην προσομοίωση. Οι μεθοδολογίες αυτές είναι πιο σύνθετες και πιο απαιτητικές και μοιάζουν αρκετά με τις αντίστοιχες μεθοδολογίες της ab initio πρόγνωσης που είδαμε στην προηγούμενη ενότητα. Με λίγα λόγια, τα μόρια του ζευγαριού πρωτεΐνη-πρωτεΐνη ή πρωτεΐνη-μικρό μόριο, αφήνονται σε μια κάποια απόσταση και μέσω της προσομοίωσης επιχειρείται μέσα από διαδοχικές «κινήσεις» να βρεθεί η καλύτερη, από άποψη ελεύθερης ενέργειας, αλληλεπίδραση μεταξύ τους. Οι κινήσεις μπορεί να αφορούν τόσο μετακινήσεις ολόκληρου του μορίου αλλά και σχετικές μεταβολές στη στερεοδιάταξή του έτσι ώστε να βρεθεί η καλύτερη πιθανή διαμόρφωση. Όπως είναι φανερό, οι μεθοδολογίες αυτές είναι περισσότερο ρεαλιστικές, αλλά ιδιαίτερα χρονοβόρες και όπως και στην περίπτωση της ab initio πρόγνωσης μόνο τα τελευταία χρόνια, με τη ανάπτυξη ισχυρών υπολογιστών και την έμφαση στην παράλληλη επεξεργασία, τέτοιες μέθοδοι απέκτησαν ευρεία χρήση.



Εικόνα 9.14: *Αγκυροβόληση πρωτεΐνης με μικρό μόριο (από https://en.wikipedia.org/wiki/Docking_%28molecular%29).*

Οι μεθοδολογίες αυτές, έχουν πολλά κοινά με τις αντίστοιχες που χρησιμοποιούνται στην ab initio πρόγνωση δομής και ειδικά στο κομμάτι της βελτιστοποίησης, καθώς στην αγκυροβόληση ξεκινάμε σχεδόν πάντα από βιομόρια γνωστής ή σχεδόν γνωστής δομής. Έτσι, δύο είναι τα σημαντικότερα προβλήματα στην αγκυροβόληση: ο υπολογισμός της ενέργειας και η στρατηγική αναζήτησης (Halperin, Ma, Wolfson, & Nussinov, 2002; Moreira, Fernandes, & Ramos, 2010). Αντιθέτως, η αναπαράσταση της δομής συνήθως δεν είναι, γιατί εδώ ενδιαφερόμαστε για μελέτη όλων των ατόμων του μορίου. Επίσης, μια άλλη διαφορά είναι ότι επιχειρούμε μοντελοποίηση και των διαμοριακών αλληλεπιδράσεων και όχι μόνο των ενδομοριακών.

Πακέτα λογισμικού κατάλληλα για αγκυροβόληση, υπάρχουν δεκάδες, τόσο σε αυτόνομες εφαρμογές όσο και σε διαδικτυακές. Μια ιδιαιτερότητα σε σχέση με άλλες κατηγορίες λογισμικού Βιοπληροφορικής είναι το γεγονός ότι καθώς η αγκυροβόληση βρίσκει πολλές εφαρμογές στο σχεδιασμό φαρμάκων (computer aided drug discovery), υπάρχουν και πολλές εφαρμογές που είναι εμπορικές. Στην ιστοσελίδα του Swiss Institude for Bioinformatics υπάρχει αναλυτική λίστα με όλα τα λογισμικά για τα διάφορα στάδια στην

ανακάλυψη φαρμάκων και στην αντίστοιχη κατηγορία για την αγκυροβόληση αναφέρονται δεκάδες πακέτα λογισμικού (http://www.click2drug.org/directory_Docking.html). Παρακάτω θα προσπαθήσουμε να κάνουμε μια σύντομη αναφορά στα πιο σημαντικά από αυτά τα πακέτα, παρουσιάζοντας τα βασικά πλεονεκτήματα του καθενός (Rodrigues & Bonvin, 2014). Γενικά, οι παράγοντες που παίζουν ρόλο στην αποτελεσματικότητα ενός τέτοιου λογισμικού είναι, α) η ταχύτητα, β) η σωστή εύρεση της επιφάνειας επαφής, γ) η δυνατότητα να χειριστεί αγκυροβόληση πρωτεΐνης-πρωτεΐνης, δ) η δυνατότητα να πραγματοποιήσει ευέλικτη αγκυροβόληση και δ) η δυνατότητα να ορίζει ο χρήστης τις πιθανές επιφάνειες επαφής κάνοντας χρήση εξωτερικής πληροφορίας.

Ένα από τα πιο γνωστά και ευρέως χρησιμοποιούμενα προγράμματα για αγκυροβόληση είναι το **GRAMM** (http://vakser.bioinformatics.ku.edu/resources/gramm/gramm1/). Το GRAMM (από τα αρχικά Global RAnge Molecular Matching) χρησιμοποιεί εμπειρική συνάρτηση ενέργειας και εκτελεί εκτεταμένες περιστροφές και μετακινήσεις των μορίων για να εντοπίσει την πιθανή θέση πρόσδεσης και μπορεί να χρησιμοποιηθεί σε ευρύ φάσμα συνθηκών, τόσο για αγκυροβόληση μικρών μορίων, όσο και για αγκυροβόληση πρωτεϊνών αλλά και πρωτεϊνικών περιοχών. Επίσης, μπορεί να χρησιμοποιηθεί τόσο για δομές υψηλής ανάλυσης όσο και για δομές πιο χαμηλής ανάλυσης. Η ποιότητα της πρόβλεψης εξαρτάται όμως από την ακρίβεια των δομών. Έτσι, μια αγκυροβόληση σε δομή μεγάλης διακριτικότητας με μικρές αλλαγές στη στερεοδιάταξη, θα δώσει πιο αξιόπιστες προβλέψεις σε σχέση με μια περίπτωση λ.χ. με δομές χαμηλής διακριτικότητας, όπου και θα πάρουμε μόνο τα γενικά χαρακτηριστικά του συμπλόκου. Υπάρχει επίσης και μια άλλη έκδοση με βελτιωμένους αλγόριθμους για αγκυροβόληση πρωτεϊνών, το **GRAMM-X** (http://vakser.compbio.ku.edu/resources/gramm/grammx/), το οποίο είναι ιδιαίτερα γρήγορο αλλά δεν μπορεί να χειριστεί ευέλικτα σύμπλοκα.

Το **AutoDock** (http://autodock.scripps.edu/) είναι ένα ολόκληρο πακέτο με εργαλεία αγκυροβόλησης. Χρησιμοποιείται κυρίως για την αγκυροβόληση μικρών μορίων και αυτή τη στιγμή υπάρχουν δύο εκδόσεις του πακέτου: το AutoDock 4 και το AutoDock Vina. Το πρώτο επιτρέπει περισσότερες παρεμβάσεις του χρήστη στην οπτικοποίηση του πλέγματος στο οποίο θα γίνει η αγκυροβόληση, κάτι που μπορεί να βοηθήσει τους χημικούς στη σύνθεση μικρών μορίων. Το δεύτερο κάνει αυτούς τους υπολογισμούς εσωτερικά και είναι πιο αυτοματοποιημένο. Επίσης υπάρχει και μια γραφική διεπαφή, το AutoDockTools, εργαλείο το οποίο βοηθάει το χρήστη να επιλέξει τους δεσμούς που θα περιστρέφονται στον προσδέτη και στην ανάλυση την αγκυροβόλησης.

Το **HADDOCK** (High Ambiguity Driven protein-protein DOCKing) είναι μια ιδιαίτερα δημοφιλής εφαρμογή για αγκυροβόληση η οποία χρησιμοποιείται κυρίως για αλληλεπιδράσεις πρωτεϊνών. Το HADDOCK διακρίνεται από τις υπόλοιπες ab initio προσεγγίσεις στο ότι δέχεται εξωτερική πληροφορία για τις πιθανές περιοχές επαφής (http://haddocking.org/). Ο χρήστης δίνει τα δύο μόρια και μια λίστα πιθανών (γνωστών ή προβλεφθέντων) καταλοίπων της επιφάνειας επαφής για να κατευθύνει με αυτόν τον τρόπο τη διαδικασία της αγκυροβόλησης. Η διαδικτυακή εφαρμογή είναι ιδιαίτερα εύχρηστη για την πραγματοποίηση της ανάλυσης, ενώ υπάρχουν και επιπλέον επιλογές για την πλήρη εκμετάλλευση των δυνατοτήτων του HADDOCK και για την εξατομίκευση της διαδικασίας.

Το **FTDock** (http://www.sbg.bio.ic.ac.uk/docking/ftdock.html) είναι μια ιδιαίτερα γρήγορη εφαρμογή αγκυροβόλησης η οποία βασίζεται στη συμπληρωματικότητα των σχημάτων. Ο αλγόριθμος επεξεργάζεται το σχήμα των μορίων χρησιμοποιώντας μετασχηματισμούς Fourier και προαιρετικά εφαρμόζει και ένα ηλεκτροστατικό φίλτρο.

Το **DOT** (http://www.sdsc.edu/CCMS/DOT/) είναι μια εφαρμογή για αγκυροβόληση που μπορεί να δεχτεί σαν δεδομένα εισόδου τόσο ζευγάρια πρωτεϊνών-πρωτεϊνών όσο και άλλες κατηγορίες μορίων. Το DOT εργάζεται με αλγόριθμο σταθερής αγκυροβόλησης που ψάχνει αναλυτικά όλες τις πιθανές διευθετήσεις του ενός μορίου σε σχέση με το άλλο. Στον υπολογισμό της ενέργειας υπολογίζονται τα ηλεκτροστατικά δυναμικά αλλά και οι αλληλεπιδράσεις van der Waals, ενώ κάνει και χρήση μετασχηματισμών Fourier.

Το **ZDOCK** (http://www.umassmed.edu/zlab/) είναι ένα άλλο πετυχημένο εργαλείο για γρήγορη αγκυροβόληση και εύρεση των αλληλεπιδράσεων μεταξύ δύο πρωτεϊνών. Βασίζεται σε μια μεθοδολογία «στέρεας» αγκυροβόλησης με συμπληρωματικότητα σχημάτων, με ειδικές συναρτήσεις για υπολογισμό των ηλεκτροστατικών αλληλεπιδράσεων. Το ZDOCK είναι ιδιαίτερα γρήγορο αλλά δεν μπορεί να χειριστεί ευέλικτα σύμπλοκα.

Το **ClusPro** (http://cluspro.bu.edu/), είναι ένας άλλος αλγόριθμος που έχει δώσει πολύ καλά αποτελέσματα σε αξιολογήσεις. Στηρίζεται σε μια γρήγορη αναζήτηση με βάση τη συμπληρωματικότητα των σχημάτων με χρήση μετασχηματισμών Fourier. Στο δεύτερο στάδιο πραγματοποιεί ομαδοποίηση με βάση το

RMSD και στο τέλος βελτιστοποιεί τις επιλεγμένες δομές με το CHARMM. Ένα μειονέκτημά του είναι ότι δεν κάνει ευέλικτη αγκυροβόληση.

Το **SwissDock** (http://www.swissdock.ch/) είναι μια διαδικτυακή εφαρμογή που επιτρέπει με εύκολο και γρήγορο τρόπο την πρόβλεψη των αλληλεπιδράσεων μιας πρωτεΐνης με ένα μικρό μόριο. Το SwissDock βασίζεται στο λογισμικό EADock DSS, ο αλγόριθμος του οποίου περιλαμβάνει αρχικά την κατασκευή πολλών μοντέλων είτε σε μια εντοπισμένη περιοχή (local docking) ή γύρω από όλες τις πιθανές κοιλότητες του πρωτεϊνικού μορίου (blind docking). Παράλληλα, το CHARMM χρησιμοποιείται για τον υπολογισμό ενέργειας και στο τέλος τα μοντέλα με τις καλύτερες τιμές ενέργειας επιλέγονται και ομαδοποιούνται.

Το **rDock** (http://rdock.sourceforge.net/) είναι επίσης μια εφαρμογή για την αγκυροβόληση μικρών μορίων σε πρωτεΐνες εστιασμένη στην ταχύτητα και την ευελιξία. Είναι λογισμικό ανοιχτού κώδικα και είναι σχεδιασμένο ειδικά για τις λεγόμενες διαδικασίες High Throughput Virtual Screening (HTVS). Είναι επίσης ιδιαίτερα ελαφρύ σαν λογισμικό, και μπορεί να εγκατασταθεί σε όλα τα συστήματα Linux, ενώ με την ευέλικτη αρχιτεκτονική του μπορεί να εγκατασταθεί σε cluster και να χρησιμοποιήσει απεριόριστο αριθμό CPUs.

Τέλος, το **RosettaDock** (http://rosie.rosettacommons.org/docking2) είναι η παραλλαγή του γνωστού αλγόριθμου Rosetta, στην αγκυροβόληση. Βασίζεται σε μια μέθοδο προσομοίωσης Monte Carlo (MC) και εργάζεται σε δύο βήματα: στο πρώτο γίνεται μια ελαχιστοποίηση χαμηλής διακριτικότητας για τη διευθέτηση της κύριας αλυσίδας (ξεκινώντας είτε από τυχαίες θέσεις είτε από μια θέση επιλεγμένη από το χρήστη), ενώ στο δεύτερο βήμα γίνεται μια ελαχιστοποίηση ενέργειας για τη βελτιστοποίηση της διευθέτησης των πλευρικών ομάδων. Στα διαφορετικά στάδια, χρησιμοποιούνται επίσης και εμπειρικές συναρτήσεις ενέργειας με διαφορετικά χαρακτηριστικά.

Γενικά η αξιολόγηση των τόσων πολλών και διαφορετικών μεταξύ τους μεθόδων και λογισμικών είναι μια δύσκολη διαδικασία (Rodrigues & Bonvin, 2014; R. D. Taylor, et al., 2002). Κατ' αναλογία με τους διαγωνισμούς CASP και CAFASP για την πρόγνωση της δομής των πρωτεϊνών, υπάρχει, ειδικά για τον εντοπισμό των αλληλεπιδράσεων μεταξύ πρωτεϊνών, ο διαγωνισμός **CAPRI** (Critical Assessment of PRediction of Interactions). Το CAPRI είναι μια συνεχής διαδικασία κατά την οποία οι ερευνητές εφαρμόζουν τις μεθοδολογίες τους για αγκυροβόληση πρωτεϊνών στο ίδιο σύνολο δεδομένων, που αποτελείται από πρωτεΐνες των οποίων οι δομές έχουν πρόσφατα προσδιοριστεί πειραματικά, αλλά παραμένουν κρυφές με τη συναίνεση των ερευνητών που έκαναν τον προσδιορισμό. Το όλο πείραμα είναι διπλότυφλο, με την έννοια ότι ούτε οι επιστήμονες που κάνουν την πρόγνωση ξέρουν τη δομή, αλλά ούτε και οι αξιολογητές ξέρουν τον δημιουργό της κάθε πρόγνωσης (http://www.ebi.ac.uk/msd-srv/capri/).

Παρόλο που οι μέθοδοι αγκυροβόλησης απέχουν αρκετά από το να χαρακτηριστούν τέλειες, εξελίσσονται συνεχώς και αναμένεται στα επόμενα χρόνια να υπάρξει πρόοδος στον τομέα. Έτσι, αναμένεται πρόοδος στον τομέα της ενσωμάτωσης πειραματικής πληροφορίας, με σκοπό την παραγωγή όλο και πιο ρεαλιστικών μοντέλων. Στην περίπτωση της αγκυροβόλησης μικρών μορίων, η χρησιμότητα στην ανακάλυψη φαρμάκων είναι προφανής. Παρ' όλα αυτά, και η αγκυροβόληση πρωτεΐνης-πρωτεΐνης είναι ένας ιδιαίτερα σημαντικός τομέας, καθώς μπορεί να δώσει απαντήσεις σε πολλά προβλήματα που αφορούν τη σχέση δομής και λειτουργίας των πρωτεϊνών (τεταρτοταγής δομή, πρωτεϊνικές αλληλεπιδράσεις, μηχανισμοί δράσης ενζύμων κ.ο.κ.).

Βιβλιογραφία

Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., . . . Zwart, P. H. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr, 66*(Pt 2), 213-221.

Alvarez, J. C. (2004). High-throughput docking as a source of novel drug leads. *Current opinion in chemical biology, 8*(4), 365-370.

Andersen, C. A., Palmer, A. G., Brunak, S., & Rost, B. (2002). Continuum secondary structure captures protein flexibility. *Structure, 10*(2), 175-184.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., . . . Bordoli, L. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, gku340.

Bonvin, A. M. (2006). Flexible protein–protein docking. *Current Opinion in Structural Biology, 16*(2), 194-200.

Bowie, J. U., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science, 253*(5016), 164-170.

Brooks, B. R., Brooks, C. L., MacKerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., . . . Boresch, S. (2009). CHARMM: the biomolecular simulation program. *Journal of computational chemistry, 30*(10), 1545-1614.

Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., . . . Woods, R. J. (2005). The Amber biomolecular simulation programs. *Journal of computational chemistry, 26*(16), 1668-1688.

Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., . . . Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr, 66*(Pt 1), 12-21.

Dror, O., Benyamini, H., Nussinov, R., & Wolfson, H. J. (2003). Multiple structural alignment by secondary structures: algorithm and applications. *Protein Science, 12*(11), 2492-2507.

Eswar, N., Marti-Renom, M. A., Webb, B., Madhusudhan, M. S., Eramian, D., Shen, M., . . . Sali, A. (2006). Comparative Protein Structure Modeling With MODELLERCurrent Protocols in Bioinformatics (Vol. 5.6.1-5.6.30): John Wiley & Sons, Inc.

Frishman, D., & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins, 23*(4), 566-579.

Güntert, P. (2011). Automated protein structure determination from NMR data. In A. J. Dingley & S. M. Pascal (Eds.), *Biomolecular NMR spectroscopy* (pp. 341). Amsterdam: IOS Press.

Gray, J. J. (2006). High-resolution protein–protein docking. *Current Opinion in Structural Biology, 16*(2), 183-193.

Halperin, I., Ma, B., Wolfson, H., & Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics, 47*(4), 409-443.

Heinig, M., & Frishman, D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res, 32*(Web Server issue), W500-502.

Helles, G. (2008). A comparative study of the reported performance of ab initio protein structure prediction algorithms. *Journal of the Royal Society Interface, 5*(21), 387-396.

Holm, L., & Rosenström, P. (2010). Dali server: conservation mapping in 3D. *Nucleic Acids Research, 38*(suppl 2), W545-W549.

Hooft, R. W., Vriend, G., Sander, C., & Abola, E. E. (1996). Errors in protein structures. *Nature, 381*(6580), 272.

Jones, D. T. (1998). THREADER : Protein Sequence Threading by Double Dynamic Programming. In S. Salzberg, D. Searls & S. Kasif (Eds.), Computational Methods in Molecular Biology: Elsevier Science.

Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of molecular biology, 287*(4), 797-815.

Jones, D. T., Taylort, W. R., & Thornton, J. M. (1992). A new approach to protein fold recognition.

Joosten, R. P., Salzemann, J., Bloch, V., Stockinger, H., Berglund, A. C., Blanchet, C., . . . Vriend, G. (2009). PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *J Appl Crystallogr, 42*(Pt 3), 376-384.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers, 22*(12), 2577-2637.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protocols, 10*(6), 845-858.

Kolodny, R., Koehl, P., & Levitt, M. (2005). Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *Journal of molecular biology, 346*(4), 1173-1188.

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., & Lesk, A. M. (2006). MUSTANG: a multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics, 64*(3), 559-574.

Krissinel, E., & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography, 60*(12), 2256-2268.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK - a program to check the stereochemical quality of protein structures. *J. App. Cryst, 26*, 283-291.

Lathrop, R. H. (1994). The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein engineering, 7*(9), 1059-1068.

Liu, Y.-S., Fang, Y., & Ramani, K. (2009). Using least median of squares for structural superposition of flexible proteins. *BMC Bioinformatics, 10*(1), 29.

Maiti, R., Van Domselaar, G. H., Zhang, H., & Wishart, D. S. (2004). SuperPose: a simple server for sophisticated structural superposition. *Nucleic Acids Research, 32*(suppl 2), W590-W594.

Mayr, G., Domingues, F. S., & Lackner, P. (2007). Comparative analysis of protein structure alignments. *BMC Structural Biology, 7*(1), 50.

McLachlan, A. D. (1982). Rapid comparison of protein structures. *Acta Crystallogr D Biol Crystallogr, A38*, 871-873

Meiler, J., & Baker, D. (2003). Rapid protein fold determination using unassigned NMR data. *Proc Natl Acad Sci U S A, 100*(26), 15404-15409.

Moreira, I. S., Fernandes, P. A., & Ramos, M. J. (2010). Protein–protein docking dealing with the unknown. *Journal of computational chemistry, 31*(2), 317-342.

Ortiz, A. R., Strauss, C. E., & Olmea, O. (2002). MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Science, 11*(11), 2606-2621.

Pelton, J. T., & McLean, L. R. (2000). Spectroscopic methods for analysis of protein secondary structure. *Anal Biochem, 277*(2), 167-176.

Peng, J., & Xu, J. (2011). RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function, and Bioinformatics, 79*(S10), 161-171.

Poleksic, A. (2009). Algorithms for optimal protein structure alignment. *Bioinformatics, 25*(21), 2751-2756.

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., . . . van der Spoel, D. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, btt055.

Read, R. J., Adams, P. D., Arendall, W. B., 3rd, Brunger, A. T., Emsley, P., Joosten, R. P., . . . Zwart, P. H. (2011). A new generation of crystallographic validation tools for the protein data bank. *Structure, 19*(10), 1395-1412.

Rodrigues, J. P., & Bonvin, A. M. (2014). Integrative computational modeling of protein interactions. *FEBS Journal, 281*(8), 1988-2003.

Rohl, C. A., Strauss, C. E., Misura, K. M., & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods in enzymology, 383*, 66-93.

Rost, B., Schneider, R., & Sander, C. (1997). Protein fold recognition by prediction-based threading. *Journal of molecular biology, 270*(3), 471-480.

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols, 5*(4), 725-738.

Šali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology, 234*(3), 779-815.

Schwieters, C. D., Kuszewski, J. J., & Clore, G. M. (2006). Using Xplor-NIH for NMR molecular structure determination. *Progr. NMR Spectroscopy 48*, 47-62

Shi, Y. (2014). A glimpse of structural biology through X-ray crystallography. *Cell, 159*(5), 995-1014.

Shindyalov, I. N., & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein engineering, 11*(9), 739-747.

Singh, A. P., & Brutlag, D. L. (2000). Protein Structure Alignment: A comparison of methods. *Bioinformatics*.

Söding, J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics, 21*(7), 951-960.

Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research, 33*(suppl 2), W244-W248.

Sternberg, M. J., Gabb, H. A., & Jackson, R. M. (1998). Predictive docking of protein—protein and protein—DNA complexes. *Current Opinion in Structural Biology, 8*(2), 250-256.

Sumathi, K., Ananthalakshmi, P., Roshan, M. M., & Sekar, K. (2006). 3dSS: 3D structural superposition. *Nucleic Acids Research, 34*(suppl 2), W128-W132.

Taylor, R. D., Jewsbury, P. J., & Essex, J. W. (2002). A review of protein-small molecule docking methods. *Journal of computer-aided molecular design, 16*(3), 151-166.

Taylor, W. R., & Orengo, C. A. (1989). Protein structure alignment. *Journal of molecular biology, 208*(1), 1-22.

Theobald, D. L., & Wuttke, D. S. (2006a). Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem. *Proceedings of the National Academy of Sciences, 103*(49), 18521-18527.

Theobald, D. L., & Wuttke, D. S. (2006b). THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics, 22*(17), 2171-2172.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *Journal of molecular graphics, 8*(1), 52-56.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., . . . Wilson, K. S. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr, 67*(Pt 4), 235-242.

Wu, S., & Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Research, 35*(10), 3375-3382.

Wu, S., & Zhang, Y. (2008). MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information. *Proteins: Structure, Function, and Bioinformatics, 72*(2), 547-556.

Yaffe, M. B. (2005). X-ray crystallography and structural biology. *Crit Care Med, 33*(12 Suppl), S435-440.

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research, 33*(7), 2302-2309.

Κεφάλαιο 10: Υπολογιστικές Γραμματικές

Σύνοψη
Στο κεφάλαιο αυτό θα μελετήσουμε μια γενικότερη κατηγορία μεθόδων, τις υπολογιστικές γραμματικές, οι οποίες περιλαμβάνουν σαν ειδικές περιπτώσεις μοντέλα που είδαμε σε προηγούμενα κεφάλαια (πρότυπα, HMM), αλλά και πιο σύνθετες δομές οι οποίες μπορούν να μοντελοποιήσουν καλύτερα μια σειρά από βιολογικά προβλήματα. Θα δούμε τις ιστορικές καταβολές αυτών των μεθόδων και θα μελετήσουμε τις κατηγορίες εκείνες που μπορούν να χρησιμοποιηθούν στην πρόγνωση της δευτεροταγούς δομής του RNA καλύπτοντας και το ζευγάρωμα των βάσεων. Θα μιλήσουμε για τις γνωστές εφαρμογές και το λογισμικό που βασίζεται σε τέτοιες μεθοδολογίες, και στο τέλος θα δούμε και κάποιες εφαρμογές στην πρόγνωση δευτεροταγούς δομής πρωτεϊνών, εφαρμογές που κάνουν ένα πρώτο βήμα στην πρόγνωση των μακρινών αλληλεπιδράσεων.

Προαπαιτούμενη γνώση
Στο κεφάλαιο αυτό απαραίτητη είναι η γνώση των εννοιών των κεφαλαίων που ασχολούνται με τα πρότυπα (κεφάλαιο 4), τα HMM (κεφάλαιο 8), αλλά και τις μεθόδους πρόγνωσης (κεφάλαιο 7).

10. Εισαγωγή

Στο κεφάλαιο αυτό, αφού έχουμε δει λεπτομερώς τις κανονικές εκφράσεις (regular expressions), τα πρότυπα (patterns) στις αλληλουχίες, τα προφίλ (profiles) αλλά και τα Hidden Markov Models (HMMs), θα προχωρήσουμε ένα βήμα παραπέρα. Θα δούμε πώς εντάσσονται τα παραπάνω συστήματα στη μεγάλη κατηγορία των τυπικών γλωσσών που χρησιμοποιούνται στην υπολογιστική γλωσσολογία. Στη συνέχεια, αφού έχουμε μελετήσει τις περιπτώσεις στις οποίες τα απλά αυτά μοντέλα δεν επαρκούν, θα δούμε και παραδείγματα εφαρμογών πιο σύνθετων μοντέλων.

Η θεωρία των τυπικών γλωσσών (formal language theory), ορίζει μια «γλώσσα» ως ένα σύνολο συμβόλων από κάποιο αλφάβητο. Η γραμματική, είναι μια προσέγγιση για τον ορισμό της γλώσσας, η οποία βασίζεται σε ένα σύνολο από κανόνες. Οι κανόνες αυτοί (rewriting rules) παίρνουν τη μορφή όπως $A \rightarrow xB$, όπου τα κεφαλαία γράμματα συμβολίζουν τα προσωρινά, αφηρημένα, μη-τερματικά σύμβολα (nonterminal symbols), τα οποία δεν εμφανίζονται στο αλφάβητο, ενώ τα πεζά γράμματα συμβολίζουν τα παρατηρήσιμα, τερματικά (terminal symbols) σύμβολα, τα οποία υπάρχουν στο αλφάβητο. Ο παραπάνω κανόνας καθορίζει ότι κάθε εμφάνιση του μη-τερματικού συμβόλου A μπορεί να αντικατασταθεί από το τερματικό x και το μη-τερματικό B. Γενικά, ξεκινώντας από ένα μη-τερματικό σύμβολο S, η παραγωγή της γραμματικής συνίσταται σε μια σειρά από τέτοια βήματα αντικατάστασης, τα οποία τερματίζονται όταν το τελευταίο μη-τερματικό σύμβολο έχει εξαφανιστεί.

Η ταξινόμηση και η υπολογιστική μελέτη των τυπικών γραμματικών, οφείλει την ύπαρξή της στον μεγάλο γλωσσολόγο του MIT και διάσημο αναρχικό φιλόσοφο και στοχαστή, Noam Chomsky. Η συνεισφορά του αυτή ήταν ορόσημο για την υπολογιστική γλωσσολογία και οι μεθοδολογίες αυτές χρησιμοποιούνται μέχρι σήμερα, τόσο στη μελέτη των φυσικών γλωσσών, όσο και στη θεωρητική πληροφορική (γλώσσες προγραμματισμού κλπ), αλλά και στη Βιοπληροφορική, όπως θα δούμε στο κεφάλαιο αυτό. Τα περισσότερα περιεχόμενα αυτού του κεφαλαίου, ακολουθούν την ορολογία και τη δομή του αντίστοιχου κεφαλαίου των (Durbin, Eddy, Krogh, & Mithison, 1998) ενώ αναφορές γίνονται και σε αντίστοιχα άρθρα ανασκόπησης, π.χ. (Searls, 2002).

10.1. Η ιεραρχία των γραμματικών του Chomsky

Το 1956 ο Noam Chomsky ταξινόμησε τις τυπικές γραμματικές σε ιεραρχία με κριτήριο τους τύπους των κανόνων παραγωγής τους (Chomsky, 1956). Σύμφωνα με αυτήν την ταξινόμηση μια τυπική γλώσσα G αποτελείται από:

- Ένα πεπερασμένο σύνολο V από μη τερματικά σύμβολα
- Ένα πεπερασμένο σύνολο T από τερματικά σύμβολα
- Ένα πεπερασμένο σύνολο P από κανόνες παραγωγής
- Ένα αρχικό σύμβολο S

Έτσι, μια τυπική γραμματική συμβολίζεται ως $G(V, T, P, S)$. Η ιεραρχία, περιλαμβάνει σε αυξημένη σειρά πολυπλοκότητας, τις κανονικές γραμματικές, τις γραμματικές χωρίς συμφραζόμενα, τις γραμματικές με συμφραζόμενα και τέλος, τις γενικές γραμματικές.

Στις **κανονικές γραμματικές** (regular grammars), οι οποίες ονομάζονται και γραμματικές τύπου 3, η μορφή των κανόνων παραγωγής τους είναι δεξιογραμμικές (right-linear) ή αριστερογραμμικές (left-linear). Αν είναι δεξιογραμμικές, τότε:

$$W_1 \rightarrow aW_2 \text{ ή } W \rightarrow a$$

ενώ, αν είναι αριστερογραμμικές:

$$W_1 \rightarrow W_2a \text{ ή } W \rightarrow a$$

Στις κανονικές γραμματικές, το πρώτο μέλος του κανόνα παραγωγής αποτελείται μόνο από ένα μη τερματικό σύμβολο, ενώ το δεύτερο μέλος περιέχει μια ακολουθία τερματικών συμβόλων και ένα μη τερματικό σύμβολο στα αριστερά ή στα δεξιά, ανάλογα αν η γλώσσα είναι δεξιογραμμική ή αριστερογραμμική, αντίστοιχα. Τις κανονικές γραμματικές αναγνωρίζουν τα Πεπερασμένα Αυτόματα (Finite State Automata). Αυτή η κατηγορία γλωσσών αντιστοιχεί, όπως θα δούμε, στις κανονικές εκφράσεις (regular expressions), οι οποίες έχουν πολλές εφαρμογές τόσο στη Βιοπληροφορική όσο και στην ανάλυση κειμένου. Κανονικές γλώσσες χρησιμοποιούνται, επίσης, για να οριστεί η λεξικογραφική δομή των γλωσσών προγραμματισμού.

Στις **γραμματικές χωρίς συμφραζόμενα** (context free grammar), οι οποίες ονομάζονται και γραμματικές τύπου 2, η μορφή των κανόνων παραγωγής τους είναι:

$$W \rightarrow \beta$$

Εδώ, το β είναι συμβολοσειρά (string) αποτελούμενη από οποιαδήποτε τερματικά ή μη-τερματικά σύμβολα (χωρίς όμως να συμπεριλαμβάνεται η κενή συμβολοσειρά). Τα αυτόματα που αναγνωρίζουν γραμματικές χωρίς συμφραζόμενα είναι τα Αυτόματα Στοίβας (Push Down Automata). Γλώσσες χωρίς συμφραζόμενα, αποτελούν τη θεωρητική βάση για τη δομή των φράσεων των περισσότερων γλωσσών προγραμματισμού παρόλο που το συντακτικό τους περιλαμβάνει και άλλα χαρακτηριστικά.

Στις **γραμματικές με συμφραζόμενα** (context sensitive grammar), οι οποίες ονομάζονται και γραμματικές τύπου 1, ανήκουν οι μονοτονικές γραμματικές (monotonic grammar). Η μορφή των κανόνων παραγωγής είναι:

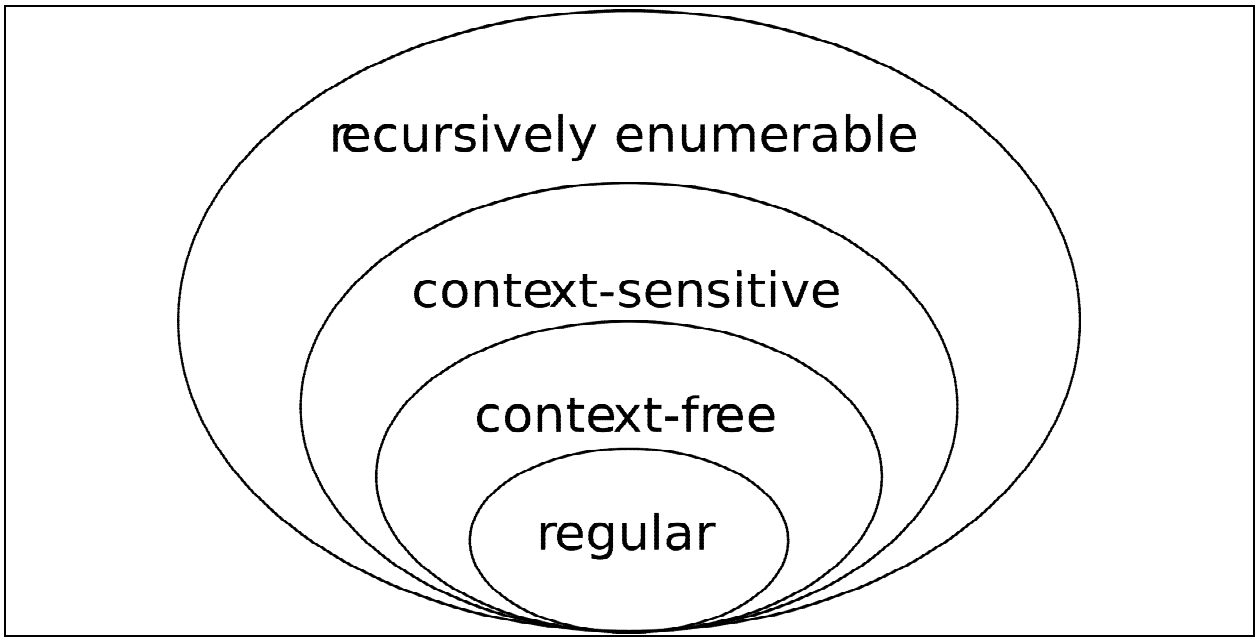
$$\alpha_1 W \alpha_2 \rightarrow a_1 \beta a_2$$

Εδώ, το a είναι ένα οποιοδήποτε τερματικό σύμβολο, το α οποιοσδήποτε συνδυασμός τερματικών και μη-τερματικών συμβόλων που περιλαμβάνει και την κενή συμβολοσειρά, ενώ το β οποιοσδήποτε συνδυασμός τερματικών και μη-τερματικών συμβόλων που δεν περιλαμβάνει και την κενή συμβολοσειρά. Παράγονται, έτσι, συμβολοσειρές μικρότερου μήκους από αυτό της αρχικής συμβολοσειράς. Γι' αυτό άλλωστε οι γλώσσες αυτές ονομάζονται μονοτονικές. Τα αυτόματα που αναγνωρίζουν γραμματικές χωρίς συμφραζόμενα είναι τα Γραμμικά Περιορισμένα Αυτόματα (Linearly Bounded Automata).

Τέλος, στις **γενικές γραμματικές** (unrestricted grammars), οι οποίες ονομάζονται και γραμματικές τύπου 0, η μορφή των κανόνων παραγωγής είναι:

$$\alpha_1 W \alpha_2 \rightarrow \beta$$

όπου β είναι οποιοσδήποτε συνδυασμός τερματικών και μη-τερματικών συμβόλων που περιλαμβάνει και την κενή συμβολοσειρά. Σε αυτήν την περίπτωση, οι συμβολοσειρές των κανόνων παραγωγής μπορούν να αποτελούνται από οποιαδήποτε σύμβολα της αλφαβήτου της γλώσσας. Από μια οποιαδήποτε συμβολοσειρά (εκτός της κενής) μπορεί να παραχθεί οποιαδήποτε άλλη (ή και η ίδια) συμβολοσειρά. Οι γενικές γραμματικές είναι γραμματικές με μόνο περιορισμό ότι από το κενό σύμβολο δεν παράγεται συμβολοσειρά. Επειδή δεν υπάρχουν άλλοι περιορισμοί, το σύνολο των γλωσσών που ανήκουν στις γενικές γραμματικές είναι το πιο ευρύ (συγκριτικά με τις υπόλοιπες γραμματικές της Ιεραρχίας του Τσόμσκι) και μέσα σε αυτό εμπεριέχονται τα σύνολα των γλωσσών που ανήκουν στις γραμματικές χαμηλότερης ιεραρχίας. Αυτές οι γλώσσες ονομάζονται και Αναδρομικώς Απαριθμήσιμες Γλώσσες (recursively enumerable languages). Οι γενικές γραμματικές αναγνωρίζονται από τις Μηχανές Τούρινγκ (Turing Machines).



Εικόνα 10.1: *Διαγραμματική απεικόνιση της ιεραρχίας των γραμματικών του Τσόμσκι. Κάθε ανώτερη γραμματική περιλαμβάνει σαν ειδική περίπτωση αυτές που βρίσκονται σε κατώτερο επίπεδο (από https://en.wikipedia.org/wiki/Chomsky_hierarchy).*

10.2. Κανονικές γραμματικές

Οι κανονικές γραμματικές, είδαμε ότι είναι γραμμικές, είτε δεξιογραμμικές είτε αριστερογραμμικές. Ας θεωρήσουμε μια απλή δεξιογραμμική γραμματική με αλφάβητο τα σύμβολα x και y, και κανόνες $S \Rightarrow xS$ και $S \Rightarrow y$. Αυτή η γραμματική μπορεί να παράξει όλες τις συμβολοσειρές που ξεκινάνε με αυθαίρετο αριθμό x και τελειώνουν με ένα μόνο y. Μπορεί να παράγει για παράδειγμα μια ακολουθία $S \rightarrow xS \rightarrow xxS \rightarrow xxxS \rightarrow xxxy$, όπου το βέλος συμβολίζει την εφαρμογή του κανόνα (μερικοί συγγραφείς συμβολίζουν την εφαρμογή του κανόνα παραγωγής με το διπλό βέλος ⇒). Σε αυτή την περίπτωση έχουμε τρεις διαδοχικές εφαρμογές του πρώτου κανόνα και μια εφαρμογή του δεύτερου, έτσι ώστε να παραχθεί τελικά η συγκεκριμένη συμβολοσειρά, μία από τις άπειρες συμβολοσειρές που μπορούν να παραχθούν από αυτή τη γλώσσα.

Όπως είπαμε, τις κανονικές γραμματικές τις αναγνωρίζουν, δηλαδή τις διαβάζουν (parsing) τα Πεπερασμένα Αυτόματα (Finite State Automata). Αυτή η κατηγορία γλωσσών αντιστοιχεί όπως θα δούμε στις κανονικές εκφράσεις (regular expressions) οι οποίες έχουν πολλές εφαρμογές τόσο στη Βιοπληροφορική όσο και στην ανάλυση κειμένου. Ας θεωρήσουμε μια πολλαπλή στοίχιση, όπως αυτή της Εικόνας 10.2, και μια κανονική έκφραση ή ένα πρότυπο της PROSITE που να την περιγράφουν (όπως είδαμε, οι κανονικές εκφράσεις και οι εκφράσεις της PROSITE είναι ισοδύναμες).

```
RU1A_HUMAN    SRSLKMRGQAFVIFKEVSSAT
SXLF_DROME    KLTGRPRGVAFVRYNKREEAQ
ROC_HUMAN     VGCSVHKGFAFVQYVNERNAR
ELAV_DROME    GNDTQTKGVGFIRFDKREEAT
```

Εικόνα 10.2: *Ένα παράδειγμα πολλαπλής στοίχισης.*

Στη συγκεκριμένη περίπτωση, το πρότυπο της PROSITE θα δίνεται από την έκφραση:

```
[RK]-G-{EDRKHPCG}-[AGSCI]-[FY]-[LIVA]-x-[FYM]
```

ενώ η αντίστοιχη κανονική έκφραση θα ήταν:

```
[RK]G[^EDRKHPCG][AGSCI][FY][LIVA].[FYM]
```

Η μετατροπή αυτών των εκφράσεων σε μια κανονική γραμματική, γίνεται αν γράψουμε έναν ξεχωριστό κανόνα για κάθε θέση της παραπάνω έκφρασης. Έτσι, στην πρώτη θέση θα έχουμε έναν κανόνα

που θα λέει ότι μετά την έναρξη (S) το πρώτο σύμβολο θα είναι R, και έναν άλλον κανόνα που θα λέει ότι το πρώτο σύμβολο μπορεί να είναι G. Αυτοί οι δύο κανόνες, συμπτύσσονται σε έναν που λέει ότι το πρώτο σύμβολο θα είναι R ή G. Στη συνέχεια προχωράμε στον επόμενο κανόνα για τη δεύτερη θέση, κ.ο.κ. Σε αυτό το παράδειγμα, αλλά και σε όλα τα παρακάτω, για να είμαστε σύμφωνοι με τους κανόνες των γραμματικών που λένε ότι για τα τερματικά σύμβολα χρησιμοποιούνται πεζά γράμματα, θα χρησιμοποιούμε r αντί για R, κ.ο.κ. Τελικά, το σύνολο των κανόνων που περιγράφει αυτή την κανονική έκφραση, θα είναι:

$$S \rightarrow rW_1|kW_1$$
$$W_1 \rightarrow gW_2$$
$$W_2 \rightarrow [afilmnqrstvwy]\ W_3$$
$$W_3 \rightarrow [agsci]\ W_4$$
$$W_4 \rightarrow fW_5|yW_5$$
$$W_5 \rightarrow lW_6|iW_6|vW_6|aW_6$$
$$W_6 \rightarrow [acdefghiklmnpqrstvwy]\ W_7$$
$$W_7 \rightarrow f|y|m$$

Μια αλληλουχία αμινοξέων που συμφωνεί με αυτή τη γραμματική, δηλαδή συμφωνεί με την παραπάνω κανονική έκφραση, θα παραχθεί με διαδοχική εφαρμογή των κανόνων:

$$S \rightarrow rW_1$$
$$\rightarrow rgW_2$$
$$\rightarrow rgaW_3$$
$$\rightarrow rgacW_4$$
$$\rightarrow rgacfW_5$$
$$\rightarrow rgacfvW_6$$
$$\rightarrow rgacfvkW_7$$
$$\rightarrow rgacfvky$$

Όλες οι κανονικές εκφράσεις, και κατά συνέπεια όλα τα πρότυπα της PROSITE μπορούν να περιγραφούν με όρους μιας τέτοιας κανονικής γραμματικής. Για την ακρίβεια, οι αλγόριθμοι που κάνουν αναζήτηση τέτοιων εκφράσεων βασίζονται στη θεωρία των κανονικών γραμματικών και των πεπερασμένων αυτομάτων. Όπως θα παρατηρήσατε ήδη από το πρώτο βήμα του παραδείγματος, εμφανίζεται πάλι το δίλημμα «ποιο είναι πιο πιθανό; Το *r* ή το *g*;». Προφανώς, όπως είδαμε και στην περίπτωση των κανονικών εκφράσεων, μια τέτοια διάκριση δεν μπορεί να γίνει και όλες οι (πεπερασμένες) αλληλουχίες που παράγονται από αυτή τη γλώσσα είναι το ίδιο πιθανές. Μια αλληλουχία, όπως η *rgacfvky* ή η *kgacfvky*, απλά ταιριάζει στο μοντέλο, ενώ μια άλλη όπως η *agacfvky* απλά δεν ταιριάζει.

Για να μπορέσουμε να κάνουμε τη διάκριση ανάμεσα στα τερματικά σύμβολα με τη μεγαλύτερη πιθανότητα, από αυτά με τη μικρότερη θα πρέπει να κάνουμε, όμοια με την περίπτωση των προτύπων, την εισαγωγή των στοχαστικών γραμματικών. Όπως είδαμε στην περίπτωση των προτύπων, μια πρώτη γενίκευση είναι η περίπτωση των προφίλ, ενώ η πιο γενική μορφή είναι τα HMM. Ας θεωρήσουμε το πιο απλό HMM που είχαμε δει στο κεφάλαιο 8, ένα μοντέλο με δύο μόνο καταστάσεις (M + και M-) και 4 σύμβολα (αναφερόμαστε στο DNA). Το μοντέλο αυτό απεικονίζεται στην Εικόνα 10.3 και όπως θα δούμε, μπορεί με τους κατάλληλους ορισμούς να μετασχηματιστεί σε μία εντελώς ισοδύναμη στοχαστική κανονική γραμματική, απλώς με την προσθήκη των κατάλληλων πιθανοτήτων. Δηλαδή, για μια στοχαστική κανονική γραμματική, χρειαζόμαστε τα τερματικά και τα μη-τερματικά σύμβολα, τους κανόνες παραγωγής (όπως στις κανονικές γραμματικές), αλλά και τις αντίστοιχες πιθανότητες.

Στην αρχή, θα πρέπει να μοντελοποιήσουμε τους κανόνες που αναφέρονται στις μεταβάσεις μεταξύ των καταστάσεων, δηλαδή μεταξύ των μη-τερματικών συμβόλων (έχουμε εδώ και καταστάσεις έναρξης και τερματισμού):

$$B \rightarrow M+|M-|E$$
$$M+ \rightarrow M+|M-|E$$
$$M- \rightarrow M+|M-|E$$

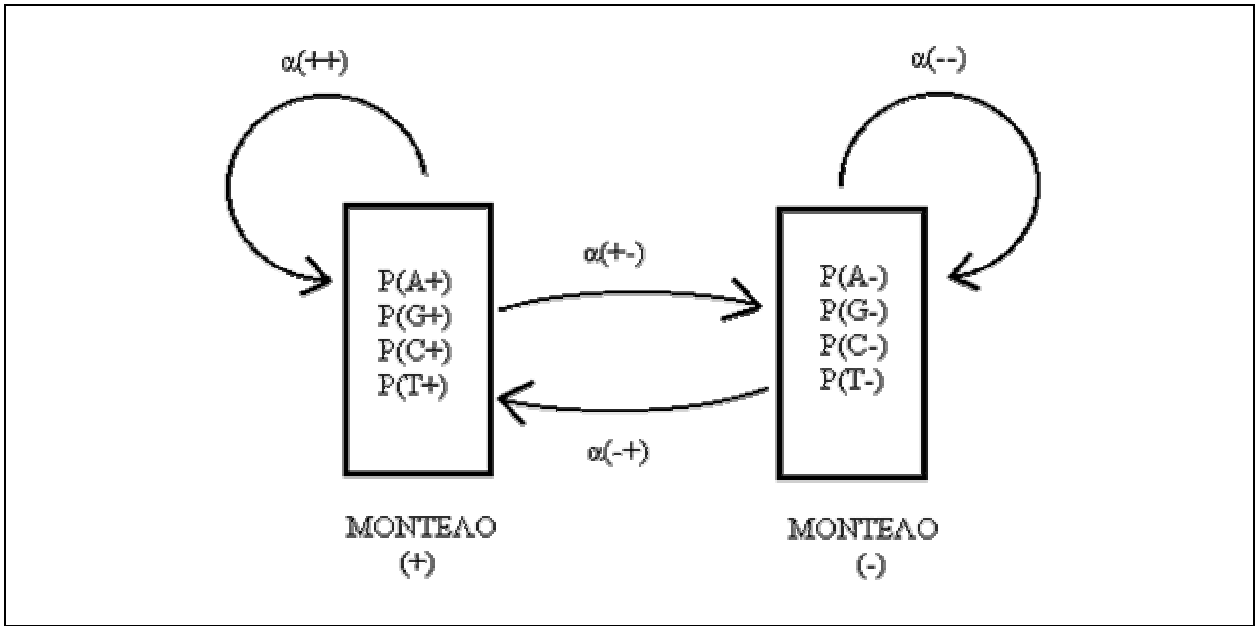
Επίσης, πρέπει να ορίσουμε τις πιθανές περιπτώσεις εμφάνισης συμβόλων από κάθε κατάσταση, χωρίς ακόμα να ορίσουμε την αντίστοιχη πιθανότητα:

$$M+: a|c|g|t$$
$$M-: a|c|g|t$$

Σύμφωνα με την ορολογία των γραμματικών, αυτά είναι τα τερματικά σύμβολα. Έτσι για να ολοκληρωθεί το μοντέλο, πρέπει να συνδυαστούν τα παραπάνω υπολογίζοντας όλες τις πιθανές περιπτώσεις:

$$B \rightarrow aM+|cM+|gM+|tM+|aM-|cM-|gM-|tM-|E$$
$$M+ \rightarrow aM+|cM+|gM+|tM+|aM-|cM-|gM-|tM-|E$$
$$M- \rightarrow aM+|cM+|gM+|tM+|aM-|cM-|gM-|tM-|E$$

Και τέλος, σε όλους αυτούς τους κανόνες, θα πρέπει να αντιστοιχήσουμε μια κατάλληλα υπολογισμένη πιθανότητα. Για παράδειγμα, για τον κανόνα $B \rightarrow aM+$ πρέπει να ορίσουμε την αντίστοιχη πιθανότητα ως $P(B \rightarrow aM+) = P(M+|B)P(a|M+)$, για τον κανόνα $M+ \rightarrow aM-$ την πιθανότητα $P(M+ \rightarrow aM-) = P(M-|M+)P(a|M-)$, κ.ο.κ.



Εικόνα 10.3: Ένα HMM με δύο καταστάσεις που έχουμε ήδη συναντήσει στο κεφάλαιο 8.

Ένα παράδειγμα παραγωγής (από τα άπειρα που μπορούν να υπάρξουν) από την παραπάνω γραμματική, είναι το:

$$B \rightarrow aM-$$
$$\rightarrow aaM+$$
$$\rightarrow aacM+$$
$$\rightarrow aactM+$$
$$\rightarrow aactgM-$$
$$\rightarrow aactgcM-$$
$$\rightarrow aactgcaE$$

Τα πεπερασμένα αυτόματα (Finite State Automata) που διαβάζουν τέτοιες γραμματικές είναι σε γενικές γραμμές οι μηχανές του Meale και οι μηχανές του Moore, οι οποίες αν και ορίζονται με διαφορετικό τρόπο, είναι σχεδόν ισοδύναμες μεταξύ τους (κάθε μηχανή του Moore μπορεί να μετατραπεί σε μια ισοδύναμη μηχανή του Meale, αλλά οι μηχανές του Meale δεν μπορούν όλες να μετατραπούν σε εντελώς ισοδύναμη μορφή). Γενικά πάντως, τέτοια αυτόματα παρόλο που χρησιμοποιούνται για κάποιες εφαρμογές στη μοντελοποίηση κυκλωμάτων, στη Βιοπληροφορική δεν έχουν εφαρμογές καθώς χρησιμοποιούνται οι πιο εύχρηστες δομές των κανονικών εκφράσεων και των HMM.

10.3. Γραμματικές χωρίς συμφραζόμενα και η πρόγνωση του RNA

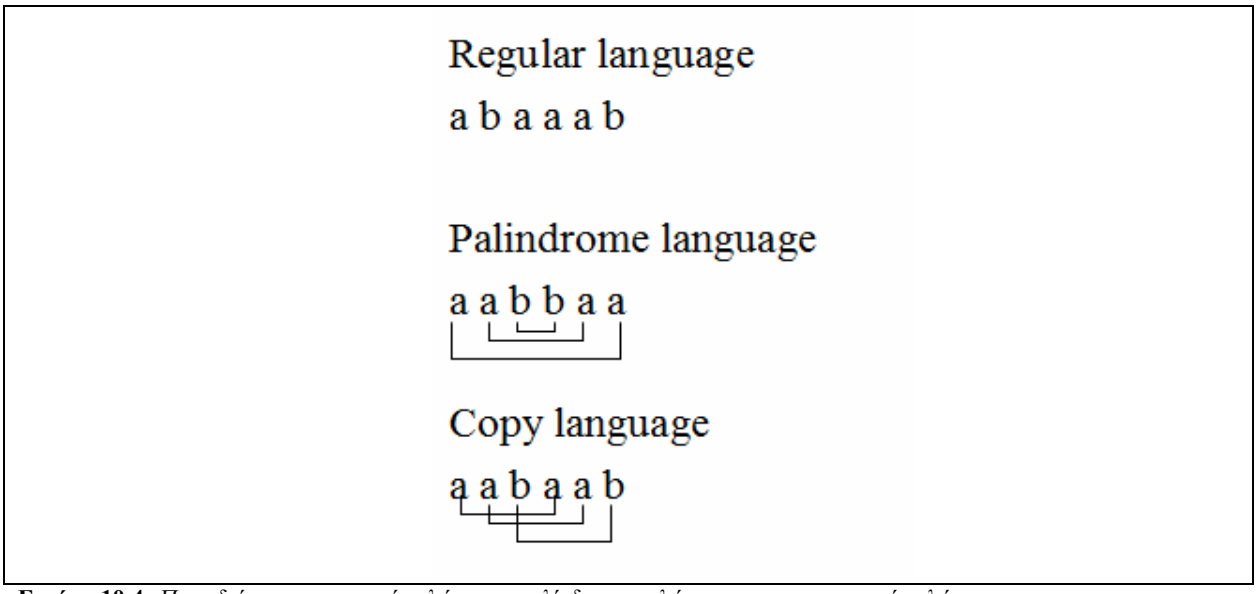
Όπως είδαμε, οι κανονικές γραμματικές έχουν μια συγκεκριμένη κατεύθυνση στον τρόπο παραγωγής (αριστερά ή δεξιά). Κατά συνέπεια, κάποιες πιο σύνθετες δομές που απαντώνται στις βιολογικές αλληλουχίες (αλλά και σε άλλου είδους γλώσσες) δεν μπορούν να μοντελοποιηθούν με επάρκεια. Γραμματικές που έχουν τη δυνατότητα να παραθέσουν οποιονδήποτε συνδυασμό τερματικών και μη-τερματικών συμβόλων στο δεξί μέρος του κανόνα παραγωγής, έχουν μεγαλύτερη εκφραστική δύναμη και αποτελούν το επόμενο επίπεδο στην ιεραρχία του Chomsky. Οι γραμματικές αυτές όπως είδαμε, ονομάζονται γραμματικές χωρίς συμφραζόμενα (context-free grammars) και περιλαμβάνουν, για παράδειγμα, περιπτώσεις κατά τις οποίες μπορεί να παραχθεί ένας αριθμός συμβόλων ενός είδους ακολουθούμενος από ίσο αριθμό συμβόλων άλλου είδους (π.χ. η ακολουθία *xxxxyyyy*). Τέτοιες ακολουθίες δεν μπορούν να παραχθούν αμφιμονοσήμαντα (άρα, και να αναγνωριστούν στη συνέχεια) από μια κανονική γραμματική, γιατί τα πεπερασμένα αυτόματα δεν έχουν τρόπο να «θυμούνται» πόσες φορές έχει προηγηθεί ένα σύμβολο έτσι ώστε να παραχθούν άλλες τόσες φορές αντίγραφα από το άλλο σύμβολο. Στα ΗΜΜ είδαμε ότι υπάρχουν τρόποι να μοντελοποιηθούν κάποιες ειδικές περιπτώσεις με τη χρήση πολλών διαφορετικών καταστάσεων, αλλά το γενικό πρόβλημα παραμένει.

Το πρόβλημα αυτό το αντιμετωπίζουν πολύ εύκολα οι γραμματικές χωρίς συμφραζόμενα με το να επιτρέπουν κανόνες παραγωγής του είδους $S \Rightarrow xSy$ οι οποίοι εξασφαλίζουν πάντα την ταυτόχρονη παραγωγή (αριστερά και δεξιά) ενός x και ενός y. Τα αυτόματα στοίβας που χαρακτηρίζουν αυτές τις γραμματικές, μπορούν να κρατήσουν τη «μνήμη» τέτοιων διαδοχικών καταστάσεων και έτσι η παραγωγή αλληλουχιών (συμβολοσειρών γενικότερα) που διαθέτουν τέτοιες εξαρτήσεις των τερματικών συμβόλων, είναι εφικτές, με την προϋπόθεση ότι οι εξαρτήσεις αυτές είναι είτε ανεξάρτητες μεταξύ τους (δηλαδή μεταξύ των ζευγαριών), είτε φωλιασμένες, αλλά ποτέ διασταυρούμενες. Το πιο χαρακτηριστικό παράδειγμα γλώσσας με την παραπάνω δομή, είναι η παλίνδρομη γλώσσα (Εικόνα 10.4). Τέτοια κείμενα, είναι γνωστά σαν καρκινικές επιγραφές και ονομάζονται οι συμμετρικές φράσεις οι οποίες μπορούν να διαβαστούν είτε από την αρχή είτε από το τέλος. Το κλασικότερο παράδειγμα από την Ελληνική ιστορία, είναι το «*ΝΙΨΟΝ ΑΝΟΜΗΜΑΤΑ ΜΗ ΜΟΝΑΝ ΟΨΙΝ*» η οποία χαρασσόταν συχνά σε πηγές και σε ελεύθερη μετάφραση στα νέα ελληνικά σημαίνει: «*πλύνε τις αμαρτίες, όχι μόνο το πρόσωπο*». Το πιο γνωστό σύγχρονο παράδειγμα τέτοιας τεχνητά κατασκευασμένης φράσης, αναφέρεται στην ερευνητική ομάδα του Bletchley Park, στην οποία συμμετείχε ο Alan Turing και είχε σκοπό (τον οποίο και πέτυχε τελικά) να σπάσει τον κώδικα του ENIGMA, της μηχανής κρυπτογράφησης που χρησιμοποιούσαν οι Γερμανοί στον Β' παγκόσμιο πόλεμο για να κωδικοποιούν τα μηνύματά τους. Οι κρυπτογράφοι αυτοί, είχαν σαν παιχνίδι να φτιάχνουν τέτοιες φράσεις και, όπως αναφέρουν, η πιο ωραία παλίνδρομη έκφραση που κατασκευάστηκε ποτέ, αποδίδεται στον Peter Hilton: "*Doc, note. I dissent. A fast never prevents fatness. I diet on cod*." (η φράση αυτή δεν είναι μόνο όμορφη γραμματικά, αλλά δίνει και σωστές...διατροφικές συμβουλές!).

Για παράδειγμα, η παλίνδρομη φράση aabaabaa μπορεί να παραχθεί από μια γραμματική με τους κανόνες

$$S \rightarrow aSa|bSb|aa|bb$$

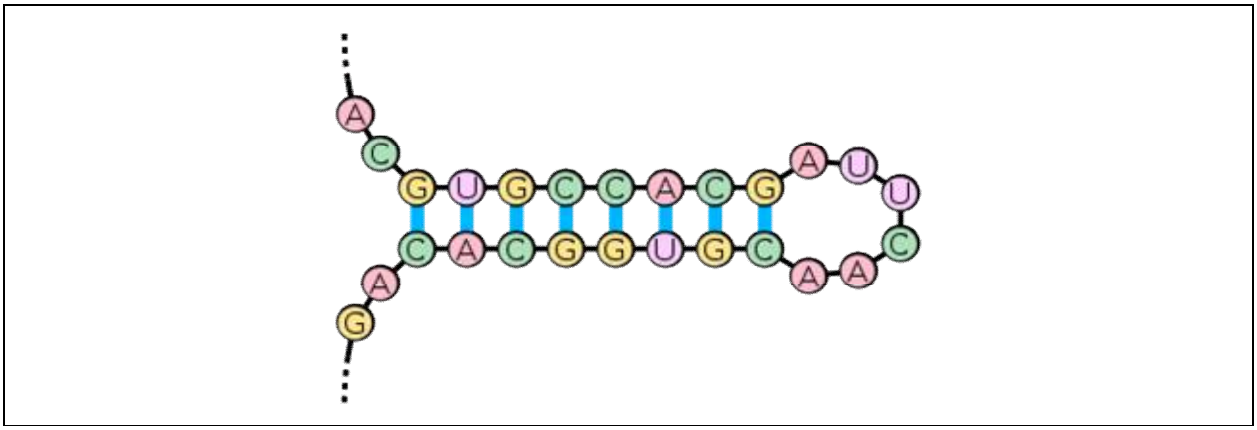
ως εξής: $S \Rightarrow aSa \Rightarrow aaSaa \Rightarrow aabSbaa \Rightarrow aabaabaa$. Βλέπουμε, ότι για κάθε παραγωγή a υπάρχει και ένα συμμετρικό a (και όμοια για τα b). Το μήκος που θα έχει η φράση καθορίζεται από το αν θα υπάρξουν πολλές επαναλήψεις του κανόνα ο οποίος περιέχει το μη-τερματικό σύμβολο (αν εμφανιστεί στο δεξί μέλος ο κανόνας που περιέχει μόνο τερματικά σύμβολα, η αλληλουχία τερματίζεται).



Εικόνα 10.4: *Παραδείγματα κανονικής γλώσσας, παλίνδρομης γλώσσας και αντιγραφικής γλώσσας.*

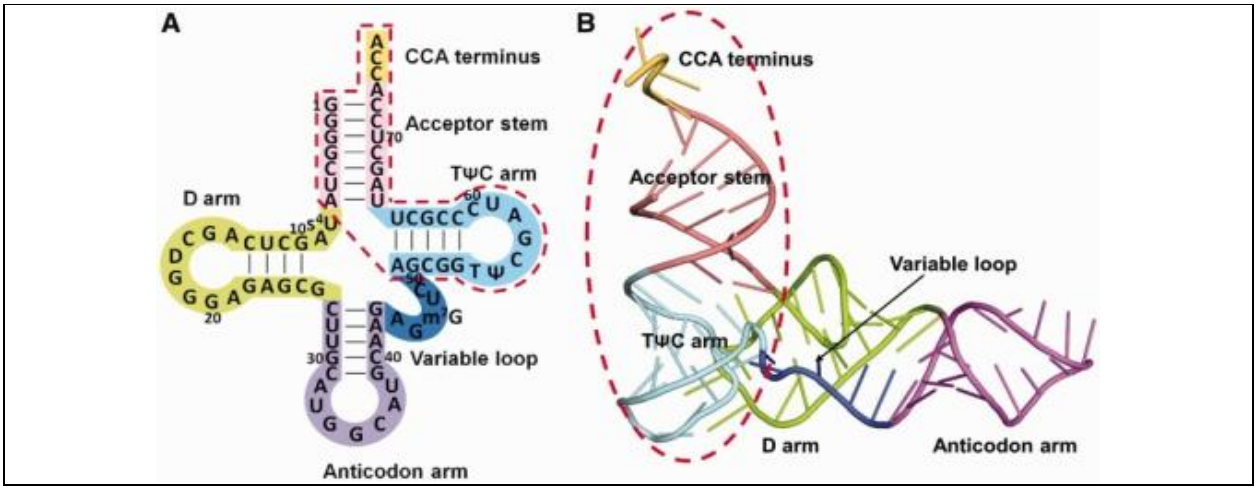
Στη βιολογία, ένα παράδειγμα χαρακτηριστικό της παλίνδρομης γλώσσας, είναι η δευτεροταγής δομή του RNA. Τα μόρια αυτά, σε αντίθεση με το DNA, αναδιπλώνονται και σχηματίζουν πολύπλοκες τρισδιάστατες δομές (παρόμοιες με των πρωτεϊνών), με σχηματισμό δεσμών υδρογόνου μεταξύ των συμπληρωματικών βάσεων (A-U, G-C). Έτσι, υπάρχει μια ξεκάθαρη «εξάρτηση» μεταξύ τμημάτων της αλληλουχίας που βρίσκονται μακριά το ένα από το άλλο. Στην πιο απλή μορφή, η εξάρτηση αυτή παίρνει τη μορφή φουρκέτας στην οποία τα απέναντι τοποθετημένα νουκλεοτίδια σχηματίζουν δεσμούς υδρογόνου. Καταλαβαίνουμε λοιπόν, ότι αν μπορούσαμε να μοντελοποιήσουμε κατάλληλα ένα τέτοιο σύστημα, θα μπορούσαμε να προβλέψουμε τη δευτεροταγή δομή του RNA και με τη χρήση αυτών των εξαρτήσεων να έχουμε μια πολύ καλή προσέγγιση για την τρισδιάστατη δομή του. Στην Εικόνα 10.5 δίνεται ένα παράδειγμα μια τέτοιας λούπας (loop) στην οποία φαίνονται και οι δεσμοί υδρογόνου μεταξύ των συμπληρωματικών βάσεων. Η δευτεροταγής δομή που προκύπτει για την περιγραφή ενός τέτοιου δομικού μοτίβου, περιγράφεται ως:

```
acgugccacgauucaacguggcacag
..(((((((((.....)))))))))..
```



Εικόνα 10.5: *Παράδειγμα ενός τυπικού βρόχου σε μόριο RNA (από https://en.wikipedia.org/wiki/Stem-loop).*

Στη δομή αυτή, οι εξαρτήσεις συμβολίζονται με τις παρενθέσεις, και βλέπουμε έτσι ότι η Γουανίνη στη θέση 3 κάνει δεσμό υδρογόνου με την Κυτοσίνη στη θέση 25, η Ουρακίλη στη θέση 3 με την Αδενίνη στη θέση 24 κ.ο.κ. Οι βάσεις οι οποίες δεν εμπλέκονται σε τέτοιες αλληλεπιδράσεις, συμβολίζονται με την τελεία (.).



Εικόνα 10.6: *Δευτεροταγής (A) και τριτοταγής δομή (B) του tRNAPhe (PDB code 6TNA) της E. coli,. (Ito et al., 2012)*

Όπως γίνεται φανερό, για να πετύχουμε την παραπάνω αναπαράσταση με μια γραμματική με συμφραζόμενα, αρκεί να ορίσουμε ρητά στους κανόνες ότι θα εξασφαλίζεται η συμμετρική κατανομή και εμφάνιση των τερματικών με τρόπο που να υπακούν στον κανόνα της συμπληρωματικότητας:

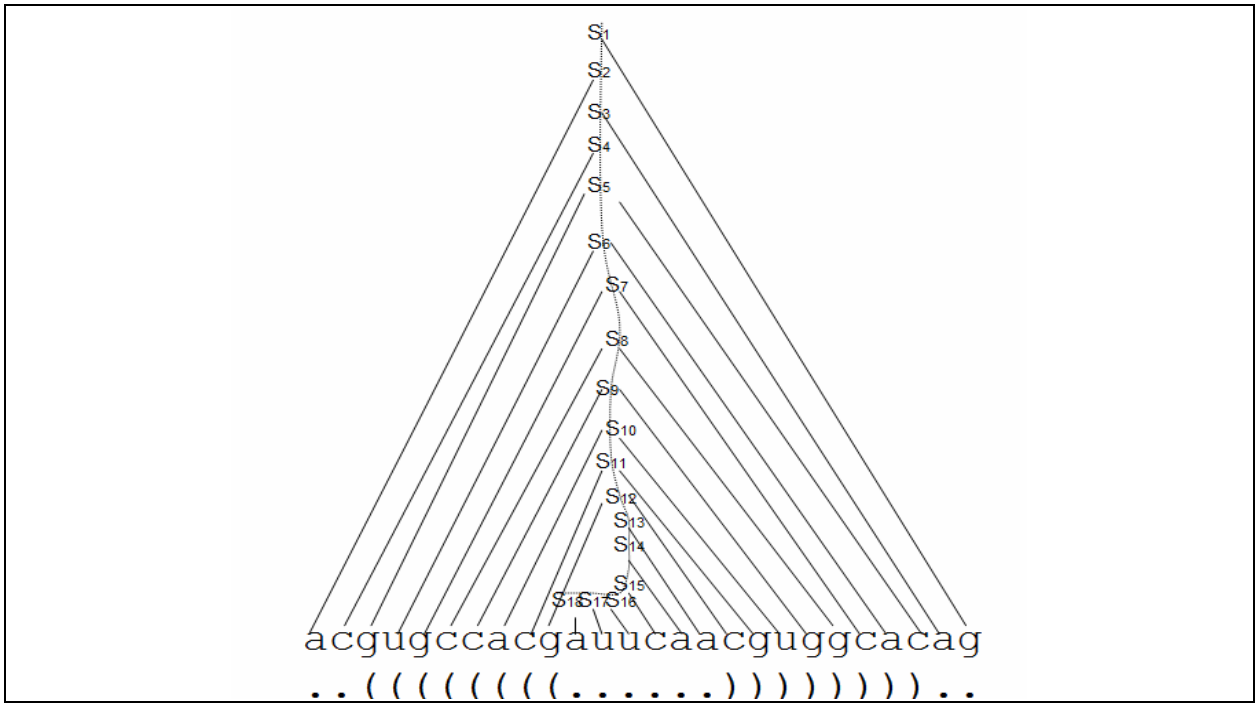
$$S \rightarrow gSc|cSg|\ aSu|uSa$$

Προφανώς, για να μπορούμε να μοντελοποιήσουμε μια πραγματικά πολύπλοκη δομή, οι κανόνες πρέπει να είναι περισσότεροι και να διαδέχονται ο ένας τον άλλον, ενώ θα πρέπει να εξασφαλίσουμε και το τι θα γίνεται στις περιπτώσεις που έχουμε βάσεις που δε συμμετέχουν σε δεσμούς υδρογόνου (π.χ. θα πρέπει να υπάρχουν και κανόνες του τύπου $S \rightarrow gS|\ Sg$), τι θα γίνεται στην περίπτωση περισσότερων βρόχων (θα πρέπει να υπάρχουν διακλαδώσεις που εξασφαλίζονται από κανόνες του τύπου $S \rightarrow S_1S_2$) και πώς θα τερματίζεται η αλληλουχία (θα πρέπει να υπάρχουν κανόνες του τύπου $S \rightarrow g$).

$$\begin{array}{ll} S_1 \rightarrow S_2g & S_{10} \rightarrow aS_{11}u \\ S_2 \rightarrow aS_3 & S_{11} \rightarrow cS_{12}g \\ S_3 \rightarrow S_4a & S_{12} \rightarrow gS_{13}c \\ S_4 \rightarrow aS_5 & S_{13} \rightarrow aS_{14} \\ S_5 \rightarrow gS_6c & S_{14} \rightarrow uS_{15} \\ S_6 \rightarrow uS_7a & S_{15} \rightarrow uS_{16} \\ S_7 \rightarrow gS_8c & S_{16} \rightarrow cS_{17} \\ S_8 \rightarrow cS_9g & S_{17} \rightarrow aS_{18} \\ S_9 \rightarrow cS_{10}g & S_{18} \rightarrow a \end{array}$$

$$\begin{array}{c} S_1 \rightarrow S_2g \\ \rightarrow aS_3g \\ \rightarrow aS_4ag \\ \rightarrow aaS_5ag \\ \rightarrow aagS_6cag \\ \rightarrow aaguS_7acag \\ \rightarrow aagugS_8cacag \\ \rightarrow aagugcS_9gcacag \\ \rightarrow aagugccS_{10}ggcacag \\ \rightarrow aagugccaS_{11}uggcacag \\ \rightarrow aagugccacS_{12}guggcacag \\ \rightarrow aagugccacgS_{13}cguggcacag \\ \rightarrow aagugccacgaS_{14}cguggcacag \\ \rightarrow aagugccacgauS_{15}cguggcacag \end{array}$$

$\rightarrow aagugccacgauuS_{16}cguggcacag$
$\rightarrow aagugccacgauucS_{17}cguggcacag$
$\rightarrow aagugccacgauucaS_{18}cguggcacag$
$\rightarrow aagugccacgauucaacguggcacag$



Εικόνα 10.7: *Η παραγωγή της δευτεροταγούς δομής του RNA σε αναπαράσταση δέντρου.*

Αυτό που γίνεται βέβαια κατανοητό για τη γλώσσα του παραπάνω παραδείγματος είναι ότι, είναι πολύ ειδική, μπορεί να παράγει δηλαδή μόνο τη συγκεκριμένη αλληλουχία RNA. Σε μια πραγματική περίσταση θα έπρεπε οι κανόνες να είναι πολύ περισσότεροι και πιο γενικοί, έτσι ώστε να μπορούν να παραχθούν από αυτούς, αλλά και το κυριότερο, να μπορούν να αναγνωριστούν, περισσότερα μόρια μιας συγκεκριμένης κατηγορίας (πχ tRNA).

Το επόμενο λογικό βήμα, θα είναι οι παραπάνω γραμματικές, να γίνουν στοχαστικές. Αυτό επιτυγχάνεται, όπως και στην περίπτωση των κανονικών γραμματικών, με την προσθήκη μιας κατάλληλης πιθανότητας σε κάθε κανόνα. Η διαδικασία αυτή οδηγεί στις πολύ γνωστές «στοχαστικές γραμματικές χωρίς συμφραζόμενα» (stochastic context-free grammars). Το βασικό πλεονέκτημα που έχουν αυτά τα μοντέλα, είναι το αντίστοιχο που είχαν και οι στοχαστικές κανονικές γραμματικές έναντι των κανονικών γραμματικών, ή τα HMM έναντι των κανονικών εκφράσεων: η επέκταση και εκλέπτυνση των αποτελεσμάτων και η ενσωμάτωση των περιπτώσεων με μικρή πιθανότητα εμφάνισης (με αντίστοιχη ποσοτικοποίηση). Ένα κλασικό παράδειγμα στην περίπτωση του RNA είναι το ότι μπορεί πλέον με τη χρήση (μικρών) πιθανοτήτων να επιτρέπουμε το «λαθεμένο» ζευγάρωμα βάσεων, G-U, C-A, κάτι που πιθανώς να δώσει πιο ρεαλιστικές προβλέψεις.

Κατ' αντιστοιχία με τα HMM, στις στοχαστικές γραμματικές χωρίς συμφραζόμενα, έχουν εφαρμογή τα τρία κλασικά ερωτήματα:

- Πώς θα επιτύχουμε την καλύτερη στοίχιση μιας ακολουθίας με μια γραμματική (alignment-parsing problem)
- Πώς θα υπολογίσουμε την πιθανότητα μιας ακολουθίας δεδομένης μιας γραμματικής (scoring problem)
- Πώς θα γίνει η εύρεση των βέλτιστων παραμέτρων μιας γραμματικής αν υπάρχουν γνωστά παραδείγματα (training problem)

Παρόλο που λόγω πολυπλοκότητας και χώρου δεν θα μπούμε σε λεπτομέρειες, οι απαντήσεις στα προβλήματα αυτά ακολουθούν επίσης μια πορεία ανάλογη με αυτήν των HMM. Το πρώτο ερώτημα, αφορά την εφαρμογή του αλγόριθμου του Viterbi στις γραμματικές. Ο αλγόριθμος αυτός ονομάζεται αλγόριθμος των Cocke-Younger-Kasami (CYK algorithm) και πρώτη φορά προτάθηκε από τον Youmger το 1967 (Younger, 1967). Στο δεύτερο ερώτημα, ο αντίστοιχος αλγόριθμος είναι ο αλγόριθμος Inside (outside algorithm) που είναι αντίστοιχος του Forward (ενώ ο outside είναι ο αντίστοιχος του Backward). Τέλος, το συνολικό πρόβλημα της εκπαίδευσης, απαντάται από τον αλγόριθμο Inside-Outside ο οποίος είναι αντίστοιχος του αλγορίθμου Baum-Welch (Forward-Backward) και προτάθηκε το 1979 (Baker, 1979). Όπως είναι φανερό, οι αλγόριθμοι αυτοί, σε αντίθεση με τους αλγόριθμους των HMM οι οποίοι αντιμετώπιζαν την αλληλουχία σειριακά, θα πρέπει να αντιμετωπίσουν την αλληλουχία κάνοντας χρήση του δέντρου, γι' αυτό προκύπτει και το όνομα (inside/outside). Επόμενο είναι λοιπόν, όλα τα παραπάνω να αποτυπώνονται και στην αλγοριθμική πολυπλοκότητα και στις απαιτήσεις σε μνήμη αυτών των αλγορίθμων (Πίνακας 10.1).

Μια άλλη σημαντική αναφορά, πρέπει να γίνει στη λεγόμενη «*κανονική μορφή του Chomsky*» (Chomsky Normal Form). Ο Chomsky πρότεινε το 1959 (Chomsky, 1959) ότι κάθε γραμματική χωρίς συμφραζόμενα, μπορεί να γραφτεί κάνοντας χρήση μόνο κανόνων όπως:

$$W_1 \rightarrow W_2 W_3 \text{ ή}$$
$$W_1 \rightarrow a$$

Επίσης, ισχύει και το αντίστροφο, δηλαδή κάθε γραμματική που παίρνει αυτή τη μορφή, είναι υποχρεωτικά γραμματική χωρίς συμφραζόμενα. Όταν μετατρέπουμε μια γραμματική χωρίς συμφραζόμενα στην κανονική μορφή Chomsky το μέγεθος της νέας γραμματικής αναγκαστικά θα μεγαλώνει, αλλά δεν μπορεί να είναι μεγαλύτερο από το τετράγωνο του μεγέθους της αρχικής γραμματικής (όπου «μέγεθος» εννοούμε τον αριθμό των κανόνων). Προφανώς, η μετατροπή πολύπλοκων γραμματικών δεν είναι απλή υπόθεση και έχουν προταθεί αλγόριθμοι με προκαθορισμένα βήματα για το σκοπό αυτό (Lange & Leiß, 2009). Το μεγάλο πλεονέκτημα από τη χρήση της κανονικής μορφής, από όπου προκύπτει και η σημασία της, εντοπίζεται στη χρησιμότητά της στους υπολογισμούς και στους αλγόριθμους, καθώς με τη μορφή αυτή διευκολύνονται πολύ οι υπολογισμοί και η υλοποίηση.

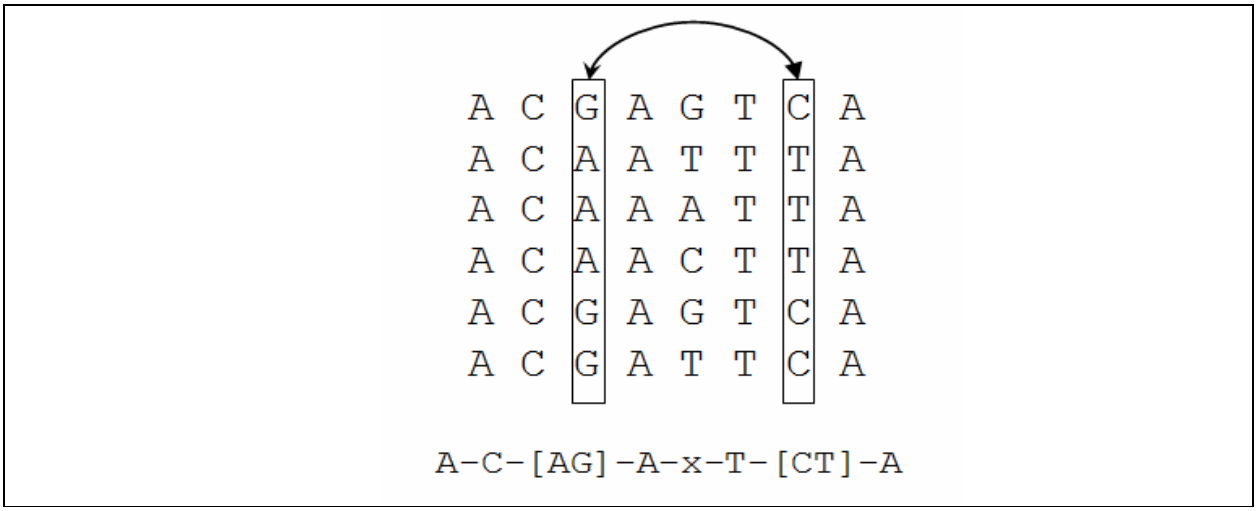
Στόχος	**HMM**	**SCFG**
Βέλτιστη στοίχιση	Viterbi	CYK
$P(\mathbf{x}\|\theta)$	Forward	Inside
EM algorithm	Baum-Welch	Inside-Outside
Απαιτήσεις σε μνήμη	$O(LM)$	$O(L^2M)$
Πολυπλοκότητα	$O(LM^2)$	$O(L^3M^3)$

Πίνακας 10.1: *Αντιστοίχιση των εννοιών μεταξύ HMM και SCFG.*

Οι πρώτες εφαρμογές των στοχαστικών γραμματικών με συμφραζόμενα στη μελέτη του RNA έγιναν τη δεκαετία του 1990 από τον Sakakibara (Sakakibara et al., 1994). Αξίζει να σημειωθεί, ότι μέχρι τότε οι πιο επιτυχημένες προσπάθειες πρόγνωσης των RNA βασίζονταν στις εργασίες της Nussinov και του Zuker. Η Nussinov είχε παρουσιάσει πρώτη το 1978 έναν κομψό αλγόριθμο δυναμικού προγραμματισμού ο οποίος μεγιστοποιούσε το σύνολο των ζευγαριών βάσεων που βρίσκονταν σε δίκλωνη μορφή (Nussinov, Pieczenik, Griggs, & Kleitman, 1978). Ο Zuker παρουσίασε λίγα χρόνια αργότερα έναν αλγόριθμο βασισμένο στη θερμοδυναμική, ο οποίος μεγιστοποιούσε μια συνάρτηση ελεύθερης ενέργειας (ΔG). Με τη μέθοδο αυτή λαμβάνεται υπόψη η συνολική ενεργειακή κατάσταση του μορίου, και πιθανώς να επιτρέπεται και το «λαθεμένο» ζευγάρωμα βάσεων, G-U, C-A και κατά συνέπεια, η μέθοδος αυτή αποδίδει καλύτερα (Zuker & Stiegler, 1981). Και οι δυο αλγόριθμοι, βρέθηκε αργότερα ότι μπορούν να γραφούν σε μια ισοδύναμη μορφή SCFG, αλλά σε γενικές γραμμές οι μεθοδολογίες που βασίζονται σε θερμοδυναμικούς υπολογισμούς ελεύθερης ενέργειας εξακολουθούν να είναι ιδιαίτερα ακριβείς, κυρίως λόγω των ευριστικών τεχνικών που ενσωματώνουν. Μια απλή εξήγηση των αλγορίθμων δυναμικού προγραμματισμού, παραθέτει ο Eddy (Sean R Eddy, 2004). Στη μέθοδο του Zuker βασίζεται η πολύ γνωστή μέθοδος **MFOLD** (http://unafold.rna.albany.edu/?q=mfold), η οποία είναι ίσως και μια από τις παλιότερες διαδικτυακές εφαρμογές στη Βιοπληροφορική. Το **RNAfold** (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) είναι επίσης μια πολύ γνωστή εφαρμογή, που χρησιμοποιεί μεταξύ άλλων, και τον αλγόριθμο του Zuker για την πρόγνωση των RNA (Lorenz et al., 2011). Το **PFOLD** (http://www.daimi.au.dk/~compbio/pfold) είναι ίσως η πιο επιτυχημένη εφαρμογή για πρόγνωση δομής RNA και βασίζεται σε γραμματικές χωρίς συμφραζόμενα (Knudsen & Hein, 2003). Οι Dowell και Eddy (Dowell & Eddy, 2004) πραγματοποίησαν μια μεγάλη

συγκριτική μελέτη στην οποία υλοποίησαν μια σειρά από διαφορετικές γραμματικές χωρίς συμφραζόμενα, ειδικά για την περίπτωση της πρόγνωσης της δομής του RNA. Μελέτησαν τις διαφορές των διαφόρων γραμματικών και πραγματοποίησαν συγκρίσεις έναντι των κλασικών αλγορίθμων ελαχιστοποίησης ενέργειας. Τα αποτελέσματα έδειξαν ότι κάποιες από τις γραμματικές αυτές, μπορούσαν να δώσουν αποτελέσματα συγκρίσιμα με τους κλασικούς αλγορίθμους, ενώ ο αλγόριθμος του PFOLD ήταν και ιδιαίτερα φειδωλός (και άρα και γρήγορος). Η μελέτη αυτή έχει και μια ιδιαίτερη σημασία καθώς ο κώδικας των γραμματικών αυτών, το λογισμικό **CONUS**, είναι διαθέσιμος, για μελλοντική χρήση και πειραματισμούς (http://selab.janelia.org/software/conus/). Παρόμοιο αποτελέσματα έδειξε και μια μεταγενέστερη μελέτη με χρήση του λογισμικού **TORNADO** το οποίο δίνει περισσότερες δυνατότητες μοντελοποίησης και εφαρμογής σε άλλες περιπτώσεις (http://selab.janelia.org/software/tornado/tornado.tar.gz) (Rivas, Lang, & Eddy, 2012).

Μια άλλη πολύ σημαντική εφαρμογή των γραμματικών αυτών στη μελέτη των RNA, ξεκίνησε από την δουλειά των Eddy και Durbin, σε μια κατηγορία μοντέλων που ονομάζονται Covariance Models (μοντέλα συνδυακύμανσης) (Eddy & Durbin, 1994). Τα μοντέλα αυτά, είναι μια ειδική περίπτωση γραμματικών χωρίς συμφραζόμενα, κατάλληλα φτιαγμένα για να περιγράφουν μια πολλαπλή στοίχιση RNA. Τα μοντέλα αυτά ειναι για τα SCFG, ό,τι είναι τα profile HMM για τα HMM. Η διαφορά τους από τα γενικά μοντέλα, είναι ότι είναι φτιαγμένα για να λειτουργούν πάνω στην πολλαπλή στοίχιση. Αυτό τους δίνει μεγαλύτερη ευελιξία από άποψη υπολογιστικής πολυπλοκότητας, αλλά καθιστά πιο δύσκολη την εφαρμογή τους σε περιπτώσεις που μια πολλαπλή στοίχιση μιας οικογένειας δεν είναι διαθέσιμη. Το όνομα βγαίνει από τη συνδιακύμανση των διαδοχικών στηλών μιας πολλαπλής στοίχισης. Η δομή της γραμματικής, επιτρέπει τη μοντελοποίηση μιας τέτοιας αλληλεπίδρασης, η οποία δεν ήταν δυνατή με τα πρότυπα κανονικών εκφράσεων, αλλά ούτε και με τα HMM.



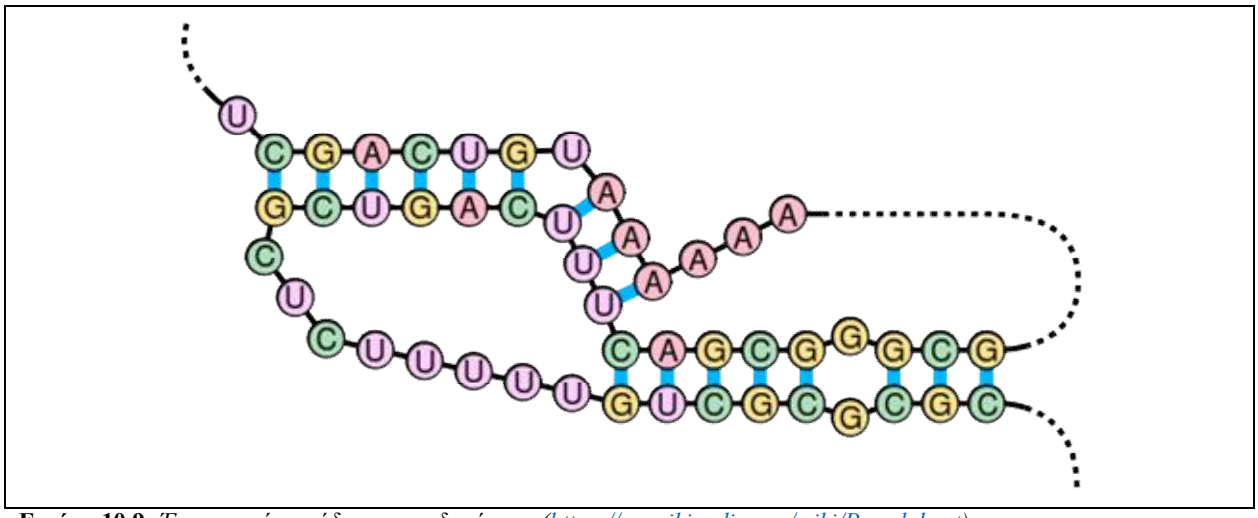
Εικόνα 10.8: *Παράδειγμα πολλαπλής στοίχισης στην οποία υπάρχει ισχυρή συσχέτιση (συνδιακύμανση) μεταξύ δύο στηλών.*

Μια πολλαπλή στοίχιση στην οποία υπάρχει ισχυρή συνδιακύμανση (δηλαδή, αλληλεπίδραση) μεταξύ της στήλης 3 και της στήλης 7, φαίνεται στην Εικόνα 10.9. Αν εξετάσουμε κάθε στήλη ξεχωριστά, βλέπουμε ότι στην 3 έχουμε 50% G και 50% A, ενώ στην 7 έχουμε 50% T και 50% C. Το απλό HMM ή ένα πρότυπο PROSITE (το οποίο θα ήταν A-C-[AG]-A-x-T-[CT]-A), θα έδινε για παράδειγμα την ίδια πιθανότητα να εμφανιστεί G ($3^{η}$) και C($7^{η}$), και στο να εμφανιστεί G($3^{η}$) και T ($7^{η}$). Παρατήρηση όμως των συχνοτήτων των δινουκλεοτιδίων, μας δείχνει ότι όταν υπάρχει G ($3^{η}$) υπάρχει πάντα C($7^{η}$), ενώ όταν υπάρχει A($3^{η}$) πάντα ακολουθείται από T ($7^{η}$). Αυτή η εξάρτηση, μπορεί να αποτυπωθεί στο covariance model, και στο συγκεκριμένο παράδειγμα το μοντέλο αυτό θα έδινε πολύ μεγάλο σκορ σε αυτή την πολλαπλή στοίχιση και θα ταξινομούσε σε αυτή την κατηγορία μια αλληλουχία όπως την ACGATTCA, αλλά θα απέρριπτε την ακολουθία ACGATTTA. Και οι δύο όμως αλληλουχίες θα ταίριαζαν (λαθεμένα) με το παραπάνω πρότυπο PROSITE.

Στα μοντέλα συνδιακύμανσης, βασίζεται το γνωστό πακέτο λογισμικού **INFERNAL**, http://infernal.wustl.edu/ το οποίο έχει υλοποιήσει και συντηρεί ο Sean Eddy (Nawrocki, Kolbe, & Eddy, 2009), και παρουσιάζει πολλές ομοιότητες με το ήδη γνωστό πακέτο HMMER για τα HMM. Για την

ακρίβεια, το INFERNAL δεν προβλέπει δευτεροταγή δομή, αλλά βρίσκει αν ένα RNA ανήκει σε μια γνωστή οικογένεια, αν ταιριάζει σε μια δεδομένη πολλαπλή στοίχιση. Αν τώρα κάποιο μέλος της οικογένειας διαθέτει δομή, η πρόγνωση γίνεται έμμεσα. Φυσικά, ένα μεγάλο πλεονέκτημα του λογισμικού είναι η ευκολία στη χρήση και η δυνατότητα, ο χρήστης να κατασκευάσει μοντέλα για τις δικές του οικογένειες RNA. Κατ' αναλογία με τη βάση PFAM, η οποία περιέχει στοιχίσεις οικογενειών πρωτεϊνών, στο INFERNAL βασίζεται η βάση δεδομένων RFAM, η οποία περιέχει οικογένειες RNA, http://rfam.xfam.org/ (Gardner et al., 2011). Το **EvoFold**, http://users.soe.ucsc.edu/~jsp/EvoFold/, χρησιμοποιεί μια παρόμοια αλλά κάπως πιο προχωρημένη τεχνική για να περιγράψει τις πολλαπλές στοιχίσεις, η οποία βασίζεται σε φυλογενετική ανάλυση και εξελικτική πληροφορία (phylo-SCFG). Το πλεονέκτημα της μεθόδου είναι ότι μπορεί να χρησιμοποιηθεί και για άλλες κατηγορίες RNA όπως microRNA (Pedersen et al., 2006). Το **RNAz**, http://www.tbi.univie.ac.at/~wash/RNAz/ βασίζεται σε ένα συνδυασμό θερμοδυναμικών παραμέτρων και πολλαπλών στοιχίσεων που δείχνουν την εξελικτική πληροφορία (Washietl, Hofacker, & Stadler, 2005). Τέλος, το **CONTRAfold**, το οποίο είναι διαθέσιμο στη διεύθυνση http://contra.stanford.edu/contrafold, βασίζεται σε ένα κάπως διαφορετικό στοχαστικό μοντέλο το οποίο αποτελεί γενίκευση των SCFG και ανήκει στην κατηγορία των «διαχωριστικών» μοντέλων, και ονομάζεται conditional log-linear model (CLLM). Το CONTRAfold είναι από τις λίγες καθαρά πιθανοθεωρητικές μεθόδους που προσεγγίζει την ακρίβεια πρόγνωσης των θερμοδυναμικών μεθόδων (Do, Woods, & Batzoglou, 2006).

Πέραν όσων αναφέραμε παραπάνω, υπάρχουν περιπτώσεις ακόμα και στη δομή του RNA που ακόμα και οι γραμματικές χωρίς συμφραζόμενα δεν επαρκούν. Μια τέτοια επιπλοκή στην πρόγνωση δευτεροταγούς δομής του RNA έφερε η ανακάλυψη των ψευδοκόμπων (pseudoknots). Ψευδοκόμπος (Εικόνα 10.10) είναι μια δομή των νουκλεϊκών οξέων στην οποία οι βρόχοι διασταυρώνονται και διακλαδώνονται με συνέπεια η μία πλευρά (το στέλεχος) του ενός να κάνει δεσμούς υδρογόνου με τη μία πλευρά του άλλου. Όπως φαίνεται και από την Εικόνα 10.10, ο ψευδοκόμπος δεν μπορεί να αναπαρασταθεί από μια παλίνδρομη γλώσσα γιατί θα απαιτείται δέντρο με διασταυρούμενες συσχετίσεις, με άλλα λόγια οι κανόνες που παράγουν τα ζευγάρια βάσεων δεν μπορούν να είναι φωλιασμένοι (ο ένας μέσα στον άλλον). Για παράδειγμα, η αλληλουχία AAUUCCGG μπορεί να αναπαρασταθεί σαν δυο φωλιασμένες (nested) παλίνδρομες αλληλουχίες, αλλά η αλληλουχία AACCUUGG απαιτεί να υπάρχει διασταύρωση (crossing) των παλίνδρομων, πράγμα που δεν επιτρέπεται. Οι ψευδοκόμποι εισάγουν πολλά προβλήματα στους αλγορίθμους πρόγνωσης που αναφέραμε παραπάνω, τόσο στους κλασικούς αλγόριθμους δυναμικού προγραμματισμού, όσο και στις αντίστοιχες γραμματικές. Έχουν προταθεί ειδικοί αλγόριθμοι δυναμικού προγραμματισμού για να υπολογίζουν τη δομή σε μόρια με ψευδοκόμπους, αλλά υπάρχουν κάποια προβλήματα. Ένας ακριβής αλγόριθμος υπάρχει, αλλά μόνο για την περίπτωση που μεγιστοποιούμε απλά το σύνολο των ζευγαριών βάσεων (όπως ο αλγόριθμος της Nussinov), αλλά ακόμα και τότε η πολυπλοκότητά του είναι μεγάλη με συνέπεια να είναι αργός. Αν πάμε σε θερμοδυναμικούς υπολογισμούς, τότε έχει αποδειχθεί ότι το πρόβλημα είναι NP-complete (Lyngsø & Pedersen, 2000). Οι πιο συνηθισμένες περιπτώσεις, αφορούν αλγόριθμους δυναμικού προγραμματισμού που αντιμετωπίζουν κάποιες μόνο ειδικές περιπτώσεις ψευδοκόμπων, τις οποίες είναι ίσως πιθανό να συναντήσουμε στην πράξη. Έτσι, μια από τις πρώτες υλοποιήσεις αποτελεί ο αλγόριθμος **PKNOTS** http://selab.janelia.org/software/pknots/pknots.tar.gz (Rivas & Eddy, 1999). Το **CYLOFOLD** είναι ένας άλλος πιο σύγχρονος τέτοιος αλγόριθμος http://cylofold.abcc.ncifcrf.gov/ (Bindewald, Kluth, & Shapiro, 2010), όπως επίσης και το **KineFOLD** http://kinefold.curie.fr/cgi-bin/form.pl (Isambert, 2009), αλλά και το **IPknot** https://github.com/satoken/ipknot, (Sato, Kato, Hamada, Akutsu, & Asai, 2011). Τέλος, το **SimulFold**, http://www.cs.ubc.ca/~irmtraud/simulfold/, επιτυγχάνει κάτι παρόμοιο αλλά χρησιμοποιεί επιπλέον και πολλαπλές στοιχίσεις (Meyer & Miklós, 2007).



Εικόνα 10.9: *Ένα τυπικό παράδειγμα ψευδοκόμπου (https://en.wikipedia.org/wiki/Pseudoknot).*

Εκτός από τις διασταυρούμενες παλίνδρομες αλληλουχίες υπάρχουν και άλλες περιπτώσεις στις οποίες οι γραμματικές χωρίς συμφραζόμενα δεν επαρκούν. Ένα τέτοιο παράδειγμα είναι οι επαναληπτικές αλληλουχίες, όπως η *xxxyyyzzz*, η οποία προκύπτει από την επανάληψη *x*, *y* και *z* σε ίσους αριθμούς. Τέτοιες αλληλεπιδράσεις διασταυρώνονται αναγκαστικά και για να αποτυπωθούν τέτοιοι κανόνες, χρειάζεται μια γραμματική με συμφραζόμενα (context-sensitive grammar) ικανή να εκφράσει αυτή τη γλώσσα η οποία ονομάζεται αντιγραφική γλώσσα (copy language). Οι γλώσσες αυτές, και οι αντίστοιχες γραμματικές που τις περιγράφουν, έχουν μεγάλο θεωρητικό ενδιαφέρον κυρίως λόγω της μεγάλης τους περιγραφικής δύναμης. Παρ' όλα αυτά, οι υπερβολικές υπολογιστικές απαιτήσεις τέτοιων εφαρμογών έχουν μέχρι στιγμής λειτουργήσει αποτρεπτικά στην εφαρμογή τους σε βιολογικά προβλήματα.

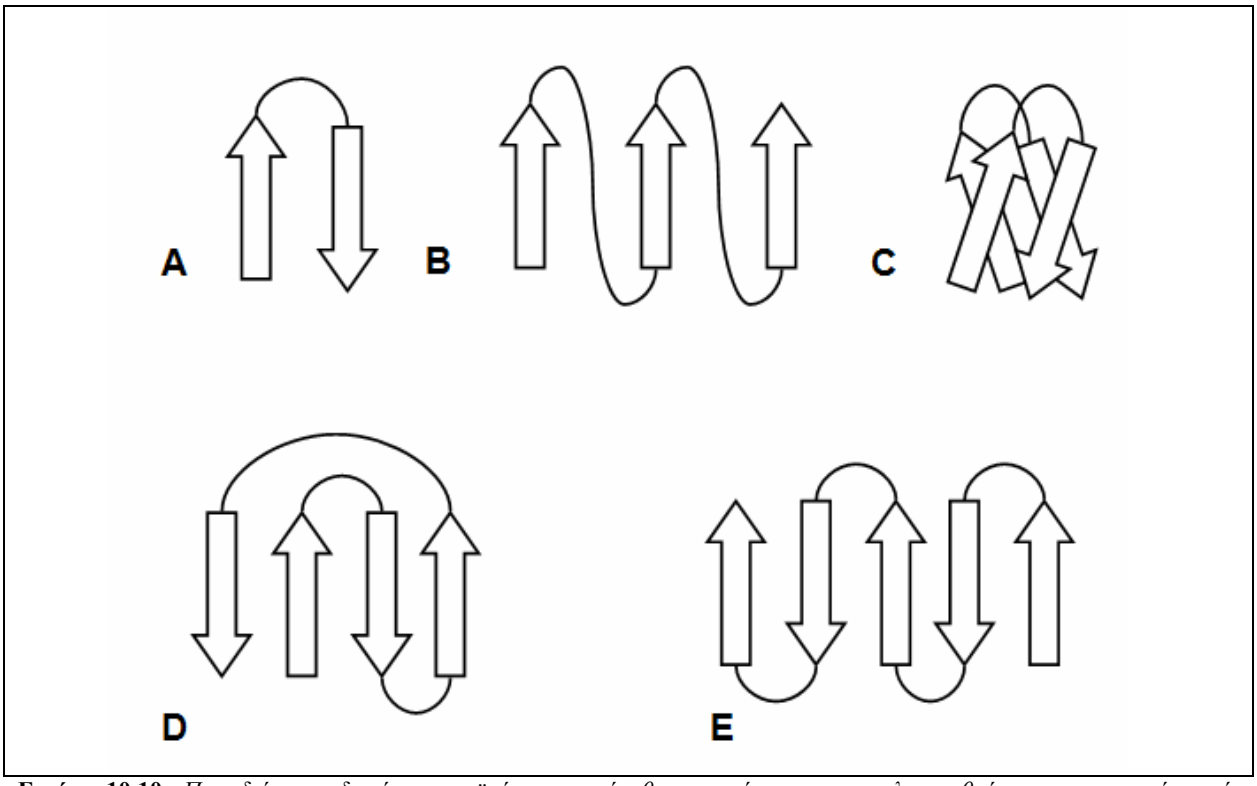
10.4. Εφαρμογές στην περίπτωση των πρωτεϊνών

Μέχρι τώρα ασχοληθήκαμε αποκλειστικά με την περίπτωση των RNA, τα οποία προσέφεραν το κλασικό παράδειγμα για την εφαρμογή των γραμματικών χωρίς συμφραζόμενα (παλίνδρομες γλώσσες και μακρινές αλληλεπιδράσεις). Παρ' όλα αυτά, μακρινές αλληλεπιδράσεις και μάλιστα ιδιαίτερα πολύπλοκες, υπάρχουν και στην περίπτωση των πρωτεϊνών και μάλιστα είναι υπεύθυνες σε μεγάλο βαθμό για την πολυπλοκότητα των τρισδιάστατων δομών τους.

Η περίπτωση των πρωτεϊνών είναι πιο περίπλοκη, για μια σειρά από λόγους:

- Οι πρωτεΐνες έχουν μεγαλύτερο αλφάβητο (20 αμινοξέα αντί για 4 νουκλεοτίδια)
- Οι αλληλεπιδράσεις που σταθεροποιούν τη δομή είναι διαφόρων ειδών (δεσμοί υδρογόνου, υδρόφοβες αλληλεπιδράσεις, δεσμοί άλατος, αλληλεπιδράσεις Van der Waals)
- Δεν υπάρχει ξεκάθαρος κανόνας για τη συμπληρωματικότητα των αμινοξέων που συμμετέχουν σε αυτές, αν και φυσικά υπάρχουν προτιμήσεις

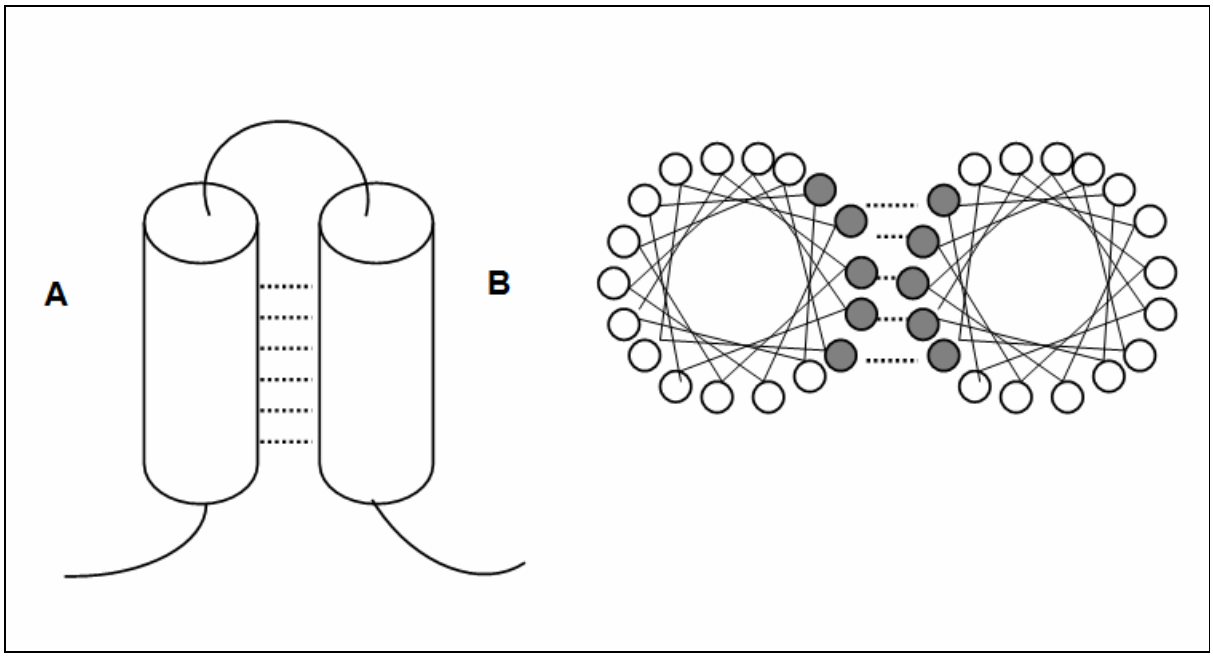
Ένα κλασικό παράδειγμα που προσεγγίζει όσο περισσότερο γίνεται την περίπτωση των αλληλεπιδράσεων των RNA, είναι η β-πτυχωτή επιφάνεια (Εικόνα 10.11). Υπάρχουν πολλές περιπτώσεις β-πτυχωτών επιφανειών, από την απλή φουρκέτα και τις παράλληλες και αντιπαράλληλες β-πτυχωτές επιφάνειες, μέχρι πιο σύνθετες δομές όπως το Greek-key motif αλλά και συνδυασμοί τέτοιων δομών για να δώσουν υπερ-δευτεροταγείς δομές, όπως το β-σάντουιτς, η β-προπέλα, το β-βαρέλι και η β-έλικα. Σε όλες τις περιπτώσεις, το χαρακτηριστικό γνώρισμα είναι ο σχηματισμός δεσμών υδρογόνου μεταξύ των N-H και C=O των πεπτιδικών δεσμών των αμινοξέων που βρίσκονται «απέναντι» στην αλυσίδα. Παρ' όλα αυτά, υπάρχει η βασική διαφορά ότι το ζευγάρωμα αυτό δεν είναι αποκλειστικό, αν και φυσικά υπάρχουν προτιμήσεις στα ζευγάρια αμινοξέων. Επίσης, το ζευγάρωμα δεν είναι αποκλειστικό και με την έννοια ότι το ίδιο αμινοξύ εμπλέκεται σε ένα δεσμό υδρογόνου προς τη μία κατεύθυνση (με την ομάδα N-H), αλλά και προς την άλλη (με την ομάδα C=O).



Εικόνα 10.10: *Παραδείγματα δομών πρωτεϊνών τα οποία θα μπορούσαν να μοντελοποιηθούν με γραμματική χωρίς συμφραζόμενα.*

Όλα τα παραπάνω, καθιστούν την απλή εφαρμογή των μεθόδων που περιγράψαμε ήδη, ιδιαίτερα προβληματική. Παρ' όλα αυτά, η μελέτη τέτοιων περιπτώσεων είναι ιδιαίτερα σημαντική καθώς πιθανώς να ανοίξει το δρόμο προς την πρόγνωση της δομής πρωτεϊνών η οποία να προβλέπει και τις μακρινές αλληλεπιδράσεις, γιατί όπως είδαμε σε προηγούμενα κεφάλαια οι μέθοδοι πρόγνωσης έχουν ένα ανώτατο όριο επιτυχίας, καθώς αδυνατούν να λάβουν υπόψη τους τις μακρινές αλληλεπιδράσεις. Η πρώτη προσπάθεια εφαρμογής γραμματικών χωρίς συμφραζόμενα στην περίπτωση της πρόγνωσης δομής των πρωτεϊνών έγινε το 1994 από τους Mamitsuka και Abe (Mamitsuka & Abe, 1994). Οι συγγραφείς, επιχείρησαν να μοντελοποιήσουν τις β-πτυχωτές επιφάνειες, με όλους τους περιορισμούς που αναφέραμε παραπάνω, και κατέληξαν στη χρήση μιας ειδικής κατηγορίας γραμματικών, γνωστών και ως *Stochastic Ranked Node Rewriting Grammars* (SRNRG). Με τις γραμματικές αυτές και μια σειρά από τροποποιήσεις (μείωση αλφαβήτου αμινοξέων, χρήση μιας πιο γρήγορης εκδοχής του αλγορίθμου inside-outside, αλλά και παραλληλοποίηση των υπολογισμών), κατάφεραν να προβλέψουν με επιτυχία τις β-πτυχωτές επιφάνειες σε ένα σύνολο δεδομένων από τη βάση HSSP με λιγότερο από 25% ομοιότητα με το σύνολο εκπαίδευσης. Τα αποτελέσματα αυτά ήταν σημαντικά, γιατί όχι μόνο προβλέφθηκαν οι θέσεις των β-κλώνων αλλά το ίδιο έγινε και για τους δεσμούς υδρογόνου που σταθεροποιούν την επιφάνεια.

Βέβαια, οι υπολογιστικές απαιτήσεις τέτοιων εγχειρημάτων καθυστέρησαν αρκετά την εφαρμογή και εξάπλωση τέτοιων μεθόδων για περίπου μια δεκαετία. Το 2006, ο Searls με τους συνεργάτες του παρουσίασαν ένα γενικό πλαίσιο για την εφαρμογή γραμματικών κανόνων στην περιγραφή βιομοριακών αλληλουχιών, και ειδικά, στην περιγραφή των πρωτεϊνών (Chiang, Joshi, & Searls, 2006). Το 2005 ο Waldispuhl παρουσίασε μια ενδιαφέρουσα εφαρμογή των γραμματικών στην πρόγνωση των διαμεμβρανικών α-ελίκων (Waldispühl & Steyaert, 2005). Χρησιμοποίησε το λεγόμενο *multi-tape S-attributed grammar* το οποίο είναι για τις γραμματικές το αντίστοιχο του class HMM, καθώς αναθέτει στα μη-τερματικά σύμβολα μια «ιδιότητα», η οποία στη συγκεκριμένη περίπτωση συμβόλιζε την αλληλεπίδραση με τις γειτονικές έλικες. Με τη μέθοδο αυτή (**TMMTSAG**), πέτυχαν συγκρίσιμα αποτελέσματα στην πρόβλεψη των διαμεμβρανικών α-ελίκων, σε σχέση με τις υπάρχουσες μεθόδους, αλλά το σημαντικότερο ήταν ότι προέβλεψαν και την αλληλεπίδραση των αμινοξέων με τις άλλες έλικες (ή την έκθεση προς τα λιπίδια της μεμβράνης). Δυστυχώς η μέθοδος αυτή δεν είναι δημόσια διαθέσιμη.



Εικόνα 10.11: *Παράδειγμα αλληλεπίδρασης μεταξύ δύο α-ελίκων (διαμεμβρανικές ή μη). Α. Η αναπαράσταση των δύο ελίκων με το κλασικό μοντέλο. Β. Αναπαράσταση με το helical wheel plot, το οποίο δείχνει τα αμινοξέα από τον κάθετο άξονα της έλικας. Με γκρι φαίνονται τα αμινοξέα τα οποία βρίσκονται σε επαφή μεταξύ τους, ενώ με άσπρο τα αμινοξέα που βρίσκονται σε επαφή με τον περιβάλλοντα χώρο. Λόγω της περιοδικότητας της έλικας, ένα κάθε τέσσερα ή πέντε αμινοξέα αναμένουμε να είναι «γκρι», αλλά στο χώρο αυτά συσσωρεύονται σε μια επιφάνεια επαφής.*

Η ίδια ομάδα, επιχείρησε να εφαρμόσει τις ίδιες μεθοδολογίες και στην περίπτωση των διαμεμβρανικών β-βαρελιών. Έτσι, δημιουργήθηκε η μέθοδος **transFold** (http://bioinformatics.bc.edu/clotelab/transFold) η οποία προβλέπει με αρκετά ικανοποιητικό τρόπο, τουλάχιστον σε σύγκριση με τους υπόλοιπους αλγόριθμους του είδους, τους διαμεμβρανικούς β-κλώνους αλλά και το δίκτυο των δεσμών υδρογόνου που σταθεροποιούν το β-βαρέλι (Waldispuhl, Berger, Clote, & Steyaert, 2006). Ένα μειονέκτημα της μεθόδου είναι το γεγονός ότι είναι αργή, αλλά και ότι απαιτεί διάφορες παραμέτρους από τον χρήστη (τι είδους βαρέλι περιμένουμε να βρούμε κ.ο.κ.). Μια επέκταση της μεθοδολογίας αυτής, παράλληλα με χρήση ενός επιπλέον βήματος για την ελαχιστοποίηση ενέργειας, αποτελεί η μέθοδος **Partifold** η οποία είναι διαθέσιμη στη διεύθυνση http://partiFold.csail.mit.edu/ (Waldispuhl, O'Donnell, Devadas, Clote, & Berger, 2008).

Τέλος, αξίζει να αναφερθούν και οι εργασίες των Dyrka και Nebel, οι οποίοι ανέπτυξαν ένα ολόκληρο πλαίσιο για την εφαρμογή γραμματικών χωρίς συμφραζόμενα σε παρόμοια προβλήματα πρωτεϊνών, ακολουθώντας μια διαφορετική στρατηγική. Το βασικό χαρακτηριστικό της μεθόδου αυτής, ήταν η χρήση ενός γενετικού αλγόριθμου για να μπορέσει να εξάγει τους κανόνες της γραμματικής, αλλά και η χρήση μιας διαφορετικής αναπαράστασης βασισμένης στις ιδιότητες των αμινοξέων προκειμένου να μειωθεί το αλφάβητο (Dyrka & Nebel, 2009). Η μεθοδολογία αυτή, εφαρμόστηκε σε μια σειρά από εργασίες, τόσο για την πρόγνωση των αλληλεπιδράσεων των διαμεμβρανικών ελίκων (Dyrka, Nebel, & Kotulska, 2013), όσο και για την πρόγνωση θέσεων πρόσδεσης άλλων μορίων πάνω στις πρωτεΐνες (Dyrka & Nebel, 2007). Όλα τα παραπάνω, δείχνουν ότι οι μεθοδολογίες αυτές αν και δεν έχουν εφαρμοστεί ιδιαίτερα λόγω των εγγενών δυσκολιών στις πρωτεΐνες, εν τούτοις είναι πολύ υποσχόμενες και καθώς η υπολογιστική ισχύς αυξάνεται, αλλά και η έρευνα στους αλγόριθμους συνεχίζεται, αναμένεται στο μέλλον να παίξουν σημαντικό ρόλο στην επίλυση τέτοιων προβλημάτων.

Βιβλιογραφία

Baker, J. K. (1979). Trainable grammars for speech recognition. The Journal of the Acoustical Society of America, 65(S1), S132-S132.

Bindewald, E., Kluth, T., & Shapiro, B. A. (2010). CyloFold: secondary structure prediction including pseudoknots. Nucleic Acids Research, 38(suppl 2), W368-W372.

Chiang, D., Joshi, A. K., & Searls, D. B. (2006). Grammatical representations of macromolecular structure. Journal of computational biology, 13(5), 1077-1100.

Chomsky, N. (1956). Three models for the description of language. IRE Transactions on Information Theory, 2(3), 113-124.

Chomsky, N. (1959). On Certain Formal Properties of Grammars. Information and Control 2(2), 137-167.

Do, C. B., Woods, D. A., & Batzoglou, S. (2006). CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics, 22(14), e90-e98.

Dowell, R. D., & Eddy, S. R. (2004). Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinformatics, 5(1), 71.

Durbin, R., Eddy, S. R., Krogh, A., & Mithison, G. (1998). Biological sequence analysis, probabilistic models of proteins and nucleic acids: Cambridge University Press.

Dyrka, W., & Nebel, J.-C. (2007). A probabilistic context-free grammar for the detection of binding sites from a protein sequence. Bmc Systems Biology, 1(Suppl 1), P78.

Dyrka, W., & Nebel, J.-C. (2009). A stochastic context free grammar based framework for analysis of protein sequences. BMC Bioinformatics, 10(1), 323.

Dyrka, W., Nebel, J.-C., & Kotulska, M. (2013). Probabilistic grammatical model for helix-helix contact site classification. Algorithms for Molecular Biology, 8(1), 31.

Eddy, S. R. (2004). How do RNA folding algorithms work? Nature biotechnology, 22(11), 1457-1458.

Eddy, S. R., & Durbin, R. (1994). RNA sequence analysis using covariance models. Nucleic Acids Res, 22(11), 2079-2088.

Gardner, P. P., Daub, J., Tate, J., Moore, B. L., Osuch, I. H., Griffiths-Jones, S., . . . Bateman, A. (2011). Rfam: Wikipedia, clans and the "decimal" release. Nucleic Acids Res, 39(Database issue), D141-145.

Isambert, H. (2009). The jerky and knotty dynamics of RNA. Methods, 49(2), 189-196.

Ito, K., Murakami, R., Mochizuki, M., Qi, H., Shimizu, Y., Miura, K.-i., . . . Uchiumi, T. (2012). Structural basis for the substrate recognition and catalysis of peptidyl-tRNA hydrolase. Nucleic Acids Research, gks790.

Knudsen, B., & Hein, J. (2003). Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Research, 31(13), 3423-3428.

Lange, M., & Leiß, H. (2009). To CNF or not to CNF? An efficient yet presentable version of the CYK algorithm. Informatica Didactica, 8, 2008-2010.

Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. Algorithms for Molecular Biology, 6(1), 26.

Lyngsø, R. B., & Pedersen, C. N. (2000). RNA pseudoknot prediction in energy-based models. Journal of computational biology, 7(3-4), 409-427.

Mamitsuka, H., & Abe, N. (1994). Predicting location and structure of beta-sheet regions using stochastic tree grammars. Proc Int Conf Intell Syst Mol Biol, 2, 276-284.

Meyer, I. M., & Miklós, I. (2007). SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework.

Nawrocki, E. P., Kolbe, D. L., & Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. Bioinformatics, 25(10), 1335-1337.

Nussinov, R., Pieczenik, G., Griggs, J. R., & Kleitman, D. J. (1978). Algorithms for Loop Matchings. SIAM Journal on Applied Mathematics 35(1), 68-82.

Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., . . . Haussler, D. (2006). Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol, 2(4), e33.

Rivas, E., & Eddy, S. R. (1999). A dynamic programming algorithm for RNA structure prediction including pseudoknots. Journal of molecular biology, 285(5), 2053-2068.

Rivas, E., Lang, R., & Eddy, S. R. (2012). A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. RNA, 18(2), 193-212.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjolander, K., Underwood, R. C., & Haussler, D. (1994). Stochastic context-free grammars for tRNA modeling. Nucleic Acids Res, 22(23), 5112-5120.

Sato, K., Kato, Y., Hamada, M., Akutsu, T., & Asai, K. (2011). IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinformatics, 27(13), i85-i93.

Searls, D. B. (2002). The language of genes. Nature, 420(6912), 211-217.

Waldispuhl, J., Berger, B., Clote, P., & Steyaert, J. M. (2006). transFold: a web server for predicting the structure and residue contacts of transmembrane beta-barrels. Nucleic Acids Res, 34(Web Server issue), W189-193.

Waldispuhl, J., O'Donnell, C. W., Devadas, S., Clote, P., & Berger, B. (2008). Modeling ensembles of transmembrane beta-barrel proteins. Proteins, 71(3), 1097-1112.

Washietl, S., Hofacker, I. L., & Stadler, P. F. (2005). Fast and reliable prediction of noncoding RNAs. Proceedings of the National Academy of Sciences of the United States of America, 102(7), 2454-2459.

Younger, D. H. (1967). Recognition and parsing of context-free languages in time n 3. Information and Control, 10(2), 189-208.

Zuker, M., & Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res, 9(1), 133-148.

Κεφάλαιο 11: Υπολογιστική Γονιδιωματική

Σύνοψη
Στο κεφάλαιο αυτό εξετάζονται οι υπολογιστικές τεχνικές που χρησιμοποιούνται στη μελέτη ολόκληρων γονιδιωμάτων. Εκτός από το ξεκάθαρο ενδιαφέρον που έχουν αυτές οι τεχνικές στη φυλογενετική ανάλυση και στην εξέλιξη, υπάρχει και άλλος πιο πρακτικός λόγος για τη χρησιμότητά τους. Παρόλο που τεχνικές που αναπτύχθηκαν σε προηγούμενα κεφάλαια όπως η στοίχιση και οι προγνώσεις είναι εύκολο να χρησιμοποιηθούν σε ολόκληρα γονιδιώματα, η ταυτόχρονη αξιολόγηση της θέσης του κάθε γονιδίου στα γονδιώματα συγγενικών οργανισμών (συγκριτική γονιδιωματική) μπορεί να δώσει πολλές επιπλέον πληροφορίες για μια σειρά από λειτουργικές ιδιότητες οι οποίες δεν θα μπορούσαν να είχαν προβλεφθεί με άλλον τρόπο.

Προαπαιτούμενη γνώση
Το κεφάλαιο απαιτεί κατανόηση των μεθόδων των κεφαλαίων 3, 4, 6 και 7.

11. Εισαγωγή

Γονιδιωματική, ονομάζουμε τον επιστημονικό κλάδο ο οποίος χρησιμοποιεί διαφορετικές τεχνικές της γενετικής, της μοριακής βιολογίας και της βιοπληροφορικής με σκοπό να βρει την αλληλουχία, να κάνει την συναρμολόγηση και να αναλύσει τη δομή και τη λειτουργία των γονιδιωμάτων, δηλαδή, ολόκληρης της γενετικής πληροφορίας που περιέχεται σε ένα κύτταρο ενός οργανισμού. Υπάρχουν πολλές υποδιαιρέσεις της γονιδιωματικής, κυρίως όσον αφορά τις διαφορετικές τεχνικές που είναι δυνατό να χρησιμοποιηθούν κάθε φορά. Για παράδειγμα, η δομική γονιδιωματική ασχολείται με το μαζικό προσδιορισμό τρισδιάστατων δομών πρωτεϊνών από ολόκληρα γονιδιώματα, ενώ η λειτουργική γονιδιωματική ασχολείται κυρίως με τη μελέτη των λειτουργικών περιοχών στα γονιδιώματα (υποκινητές, μικρά RNA κλπ).

Στο παρόν κεφάλαιο, θα εστιάσουμε στις υπολογιστικές τεχνικές που χρησιμοποιούνται στην ανάλυση γονιδωμάτων. Σε πρώτο επίπεδο, και με βάση τον γενικότερο ορισμό, υπολογιστική γονιδιωματική είναι και κάθε προσπάθεια ανάλυσης του γονιδιώματος ενός και μόνο οργανισμού, δηλαδή οι τεχνικές αλληλούχισης και συναρμολόγησης του γονιδιώματος (Zerbino & Birney, 2008), η εύρεση γονιδίων (Picardi & Pesole, 2010), η εύρεση ρυθμιστικών περιοχών (Harbison et al., 2004), η εύρεση μικρών RNA (Rigoutsos, 2010; Vlachos & Hatzigeorgiou, 2013) ή η εύρεση περιοχών οριζόντιας γονιδιακής μεταφοράς (Soucy, Huang, & Gogarten, 2015) και η εύρεση του τρόπου γονιδιακής ρύθμισης. Σε ένα επόμενο επίπεδο, οι τεχνικές που χρησιμοποιούνται είναι απλά εφαρμογές σε ολόκληρα γονιδιώματα, γνωστών μεθόδων και αλγορίθμων που σχεδιάστηκαν για αλληλουχίες (π.χ. μέθοδοι πρόγνωσης), και στη συνέχεια, στατιστική ανάλυση των αποτελεσμάτων με σκοπό την εξαγωγή γενικότερων κανόνων και συμπερασμάτων. Θα παρουσιάσουμε κάποια τέτοια παραδείγματα με σκοπό να εξοικειωθεί ο αναγνώστης με τη μεθοδολογία. Στο επόμενο στάδιο όμως, θα παρουσιαστούν οι πιο ενδιαφέρουσες τεχνικές της συγκριτικής γονιδιωματικής, οι οποίες προσφέρουν κάτι επιπλέον: αξιοποιώντας την πληροφορία για την ύπαρξη, τη θέση και την εσωτερική δομή των γονιδίων στα γονιδιώματα διαφόρων υπό σύγκριση οργανισμών, μπορούν να μας δώσουν επιπλέον πληροφορίες, πληροφορίες που από μια απλή ανάλυση ενός οργανισμού (και του γονιδιώματός του) δεν θα μπορούσαν να εξαχθούν.

Στο τέλος, θα παρουσιαστούν κάποια γνωστά παραδείγματα εφαρμογής των μεθόδων αυτών, αλλά και τα βασικά εργαλεία λογισμικού που χρησιμοποιούνται σε τέτοιου είδους αναλύσεις.

11.1. Υπολογιστική ανάλυση γονιδιωμάτων

Όπως είδαμε, ο όρος «υπολογιστική γονιδιωματική» είναι αρκετά γενικός, και πολλών ειδών υπολογιστικές αναλύσεις μπορούν να θεωρηθούν ως τέτοιες. Για παράδειγμα, κάποιος επιστήμονας μπορεί να μελετήσει ένα γονιδίωμα για να βρει πόσα γονίδια αυτό περιέχει ή ποιες είναι οι αποστάσεις μεταξύ τους ή ποια είναι η κατανομή κάποιου άλλου ειδικού χαρακτηριστικού (π.χ. ποια γονίδια ελέγχονται από κάποιο συγκεκριμένο μεταγραφικό παράγοντα, ποια γονίδια κωδικοποιούν μεμβρανικές πρωτεΐνες κ.ο.κ.). Μπορεί επίσης να ενδιαφέρει η εύρεση μικρών RNA ή η εύρεση περιοχών οριζόντιας γονιδιακής μεταφοράς και η εύρεση του τρόπου γονιδιακής ρύθμισης. Επιπλέον, πολλές από τις αναλύσεις τις γενετικής όπως η εύρεση

πολυμορφισμών ή η εύρεση επαναληπτικών αλληλουχιών μπορεί να εμπίπτει στον ορισμό της γονιδιωματικής.

Μία πιο μεγάλης κλίμακας ανάλυση θα λάβει χώρα όταν αναλυθούν παράλληλα πολλά γονιδιώματα για κάποια από τα παραπάνω χαρακτηριστικά (π.χ. για τη σύσταση GC ή για τον αριθμό των γονιδίων που κωδικοποιούν μεμβρανικές πρωτεΐνες κ.ο.κ.). Σε αυτή την περίπτωση, οδηγούμαστε τελικά σε μια ανάλυση στην οποία κάθε «γραμμή» στο αρχείο μας (δηλαδή, κάθε παρατήρηση όπως λέμε στη στατιστική) αντιστοιχεί σε ένα γονιδίωμα, ενώ κάθε στήλη (μεταβλητή) αποτελεί το χαρακτηριστικό που μελετάμε. Συνήθως τέτοιες αναλύσεις συνδυάζονται, έμμεσα ή άμεσα, με φυλογενετικά δεδομένα με σκοπό να δείξουν την κατανομή του υπό μελέτη χαρακτηριστικού στις διάφορες ταξινομικές βαθμίδες των οργανισμών υπό μελέτη. Ο σκοπός τέτοιων αναλύσεων, είναι η εξαγωγή συνολικών συμπερασμάτων και κανόνων από την ταυτόχρονη μελέτη πολλών διαφορετικών γονιδιωμάτων.

Μία πολύ απλή τέτοια γονιδιωματική ανάλυση, αλλά με τεράστια σημασία, αφορά τις αναλύσεις που έδειξαν ότι σε όλους τους οργανισμούς, οι α-ελικοειδείς διαμεβρανικές πρωτεΐνες αντιστοιχούν σε περίπου 20-30% των πρωτεϊνών που κωδικοποιούνται από τα γονιδιώματα αυτά, ενώ τα διαμεβρανικά β-βαρέλια αντιστοιχούν σε περίπου 1-2% των βακτηριακών γονιδιωμάτων. Άλλες τέτοιες αναλύσεις αφορούν τους GPCRs (οι οποίοι αποτελούν τη μεγαλύτερη οικογένεια διαμεμβρανικών υποδοχέων στα θηλαστικά, με περίπου 2% του γονιδιώματος) ή τις εκκρινόμενες πρωτεΐνες που αντιστοιχούν σε περίπου 15% των πρωτεϊνών που κωδικοποιούνται από τα γονιδιώματα όλων των οργανισμών. Επίσης, σημαντικές γονιδιωματικές αναλύσεις, για την ιδιαίτερα σημαντική αυτή ομάδα των πρωτεϊνών (διαμεμβρανικές πρωτεΐνες), έχουν γίνει για να απαντήσουν το ερώτημα του κατά πόσο στην εξέλιξή τους έχει γίνει εκτεταμένη χρήση του φαινομένου του εσωτερικού γονιδιακού διπλασιασμού. Στην αρχική ανάλυση, βρέθηκε από συσχετίσεις του μήκους των πρωτεϊνών με τον αριθμό των διαμεμβρανικών τμημάτων ότι κάτι τέτοιο είναι πιθανό (Arai, Ikeda, & Shimizu, 2003). Κατόπιν, με στοιχίσεις του πρώτου «μισού» των πρωτεϊνών αυτών με το δεύτερο μισό (δηλαδή, της μισής ακολουθίας προς το αμινοτελικό άκρο με την ακολουθία προς το καρβοξυτελικό άκρο), βρέθηκε ότι ανάμεσα σε 38,174 διαμεμβρανικές πρωτεΐνες από 87 γονιδιώματα, 377 ήταν δυνατό να έχουν παραχθεί από ένα μηχανισμό εσωτερικού διπλασιασμού και αφορούσαν κυρίως περιπτώσεις με 8, 10 και 12 διαμεμβρανικά τμήματα (Shimizu, Mitsuke, Noto, & Arai, 2004).

Φυσικά, σε κάθε ανάλυση γονιδιωμάτων είναι απαραίτητη και μια ανάλυση των πρωτεϊνικών οικογενειών με βάση τα δεδομένα κάποιων από τις βάσεις πρωτεϊνικών δεδομένων που είδαμε στο κεφάλαιο 2. Οι βάσεις αυτές μπορεί να είναι οι βάσεις γενική χρήσης όπως PFAM ή πιο εξειδικευμένες όπως η TCDB, η CAZy κ.ο.κ.

	X_1	X_2	X_3	X_4	…	X_k
Γονιδίωμα Α						
Γονιδίωμα Β						
Γονιδίωμα Γ						
Γονιδίωμα Δ						
…						

Εικόνα 11.1: *Παράδειγμα κωδικοποίησης πληροφορίας από γονιδιώματα.*

Σε άλλες περιπτώσεις, χρειάζεται να συμπτύξουμε την πληροφορία των γονιδιωμάτων και να την περιορίσουμε με χρήση μερικών μόνο παραμέτρων. Το ποσοστό των βάσεων GC είναι μία από τις πιο ευρέως χρησιμοποιούμενες παραμέτρους όταν επιθυμούμε να συγκρίνουμε διαφορετικούς οργανισμούς με βάση το

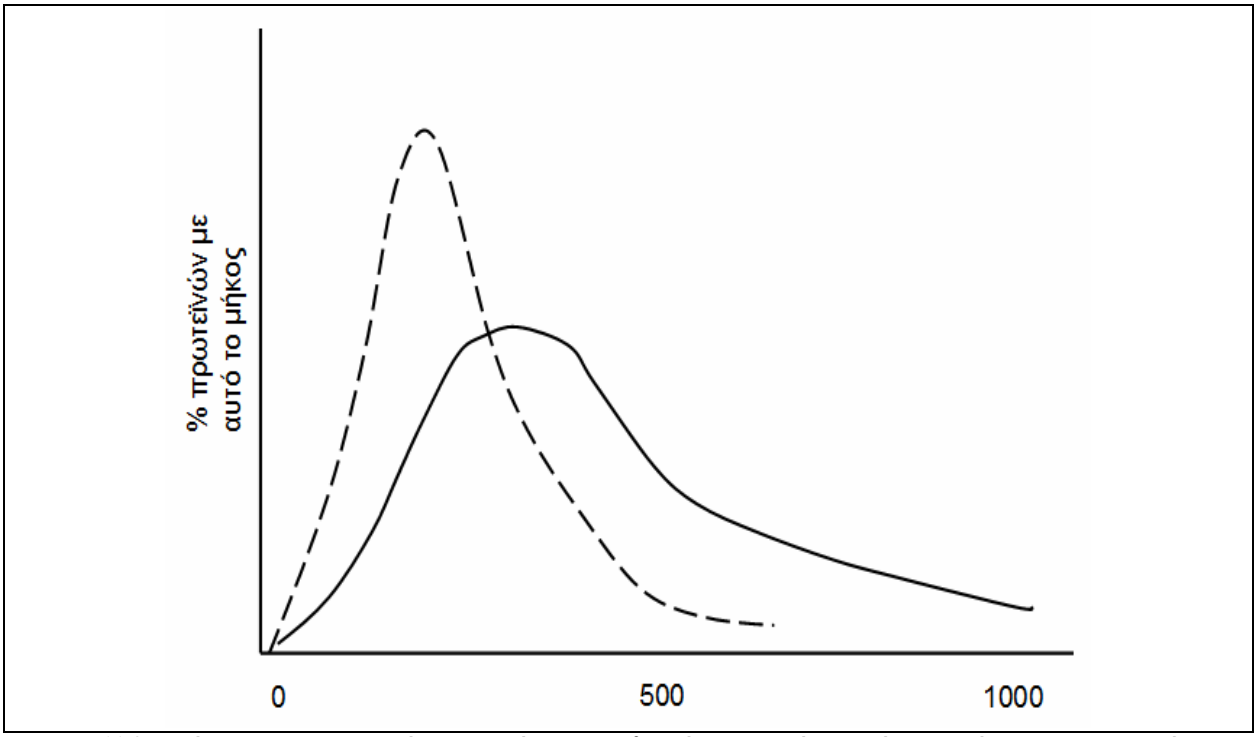
γονιδίωμα τους (Li, 2011). Στα γονιδιώματα διαφόρων οργανισμών παρατηρούνται διαφορετικά ποσοστά εμφάνισης GC (το λεγόμενο GC% content). Οι συνέπειές του είναι πολλές, κυρίως γιατί τα διαφορετικά κωδικόνια για τα ίδια αμινοξέα εμφανίζονται με διαφορετικές συχνότητες και έτσι, τελικά γονιδιώματα με διαφορετικό ποσοστό GC καταλήγουν να κωδικοποιούν πρωτεΐνες με διαφορετική περιεκτικότητα σε αμινοξέα.

Το φαινόμενο αυτό, ονομάζεται «codon bias» (πολωμένη σύσταση των κωδικονίων) και παρατηρείται σε όλους τους οργανισμούς, τόσο στο γονιδιωματικό επίπεδο όσο και μεταξύ λειτουργικώς συνδεδεμένων γονιδίων (π.χ. οπερόνια), αλλά και σε μεμονωμένα γονίδια. Άλλες παραλλαγές του φαινομένου περιλαμβάνουν τα πολωμένα ζευγάρια κωδικονίων και την πολωμένη συν-εμφάνιση κωδικονίων. Παρόλο που είναι γενικά αποδεκτό ότι η έναρξη της μετάφρασης είναι το βασικό σημείο στην πρωτεϊνοσύνθεση, είναι επίσης αναγνωρισμένο ότι το codon bias παίζει ρόλο συνεισφέροντας στην αποδοτικότητα της μετάφρασης ρυθμίζοντας τη φάση της επιμήκυνσης. Επιπλέον, παίζει σημαντικό ρόλο στον έλεγχο πολλών άλλων κυτταρικών διεργασιών οι οποίες ποικίλουν, από τη διαφορική σύνθεση πρωτεϊνών, μέχρι το πρωτεϊνικό δίπλωμα (Quax, Claassens, Soll, & van der Oost, 2015).

Σε μια από τις πρώτες, αρκετά απλές αλλά ιδιαίτερα πληροφοριακές τέτοιες μελέτες, ο Ouzounis και ο Kreil, ανέλυσαν την αμινοξική σύσταση των πρωτεϊνών που κωδικοποιούν τα γονιδιώματα 6 θερμόφιλων αρχαιοβακτηρίων (αρχαίων), 2 θερμόφιλων βακτηρίων, 17 μεσόφιλων βακτηρίων και 2 ευκαρυωτικών οργανισμών. Στην ανάλυση χρησιμοποίησαν την αμινοξική σύσταση και το ποσοστό GC και πραγματοποίησαν ιεραρχική ομαδοποίηση και ανάλυση κύριων συνιστωσών (principal components analysis). Παρόλο που το ποσοστό GC είχε μια ξεκάθαρη επιρροή, τα θερμόφιλα είδη μπορούν να αναγνωριστούν με μόνη χρήση της ολικής αμινοξικής σύστασης (Kreil & Ouzounis, 2001). Αναλύοντας τα αποτελέσματα, φάνηκε ότι τα θερμόφιλα είδη έχουν λιγότερη Γλουταμίνη (Gln) και περισσότερο Γλουταμικό (Glu) σε σχέση με τα μεσόφιλα. Τα θερμόφιλα, έχουν επίσης περισσότερη Βαλίνη (Val) και λιγότερη Θρεονίνη (Thr) σε σχέση με τα μεσόφιλα. Για τα αμινοξέα Ιστιδίνη (His), Σερίνη (Ser) και Ασπαραγίνη (Asn) υπήρχαν επίσης ενδείξεις αλλά με μικρότερο στατιστικό βάρος.

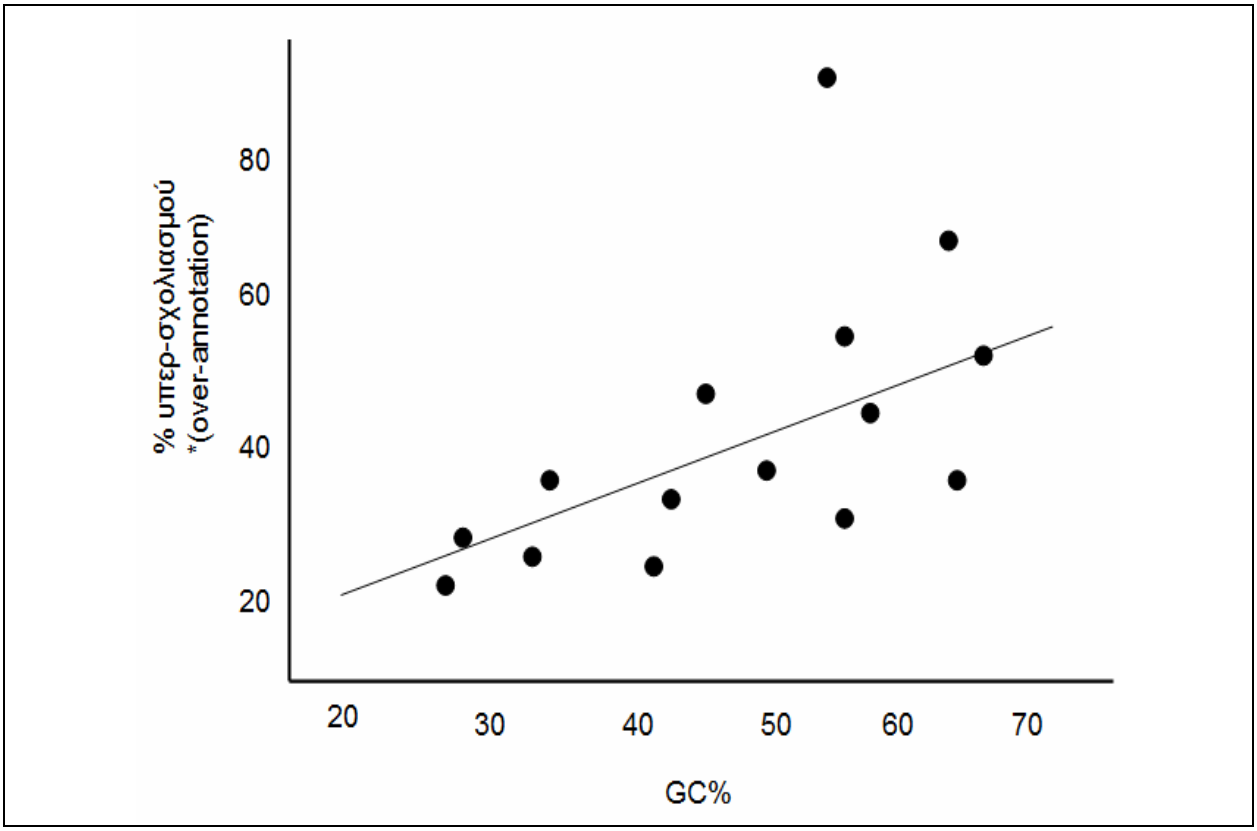
Μια άλλη ιδιαίτερα ενδιαφέρουσα εργασία, χρησιμοποίησε το ποσοστό GC με σκοπό να προσδιορίσει το μήκος των πραγματικών πρωτεϊνών στα γνωστά γονιδιώματα. Όπως είναι γνωστό, οι περισσότερες πρωτεΐνες είναι γνωστές όχι με πειραματικό τρόπο, αλλά έμμεσα κάνοντας χρήση της γονιδιωματικής πληροφορίας (conceptual translation). Έτσι, δημιουργείται το αναπόφευκτο ερώτημα αν όντως όλες αυτές οι «υποθετικές πρωτεΐνες» είναι πραγματικές ή αν αποτελούν τεχνητά προϊόντα, λάθη δηλαδή που προέκυψαν από τα προγράμματα εύρεσης γονιδίων. Η βασική σκέψη πίσω από την ανάλυση αυτή, είναι η εξής: Οι τριπλέτες είναι 64 (61 για αμινοξέα και 3 για λήξη). Αν τα νουκλεοτίδια θεωρηθούν τυχαία κατανεμημένα και ισοπίθανα (όπως απλουστευτικά είχαμε δει στο Κεφάλαιο 2), τότε θα έχουμε μια τριπλέτα λήξης περίπου μετά από 21 αμινοξέα. Όπως είδαμε, αυτό δίνει μια κατανομή για το μήκος των «ανοιχτών πλαισίων ανάγνωσης», η οποία αντιστοιχεί στη γεωμετρική κατανομή. Παρατηρούμε όμως ότι οι τριπλέτες λήξης είναι πλούσιες σε AT (TAA, TGA, TAG). Άρα, σε γονιδιώματα με μεγάλο λόγο AT οι τριπλέτες αυτές θα είναι πιο συχνές, ενώ στα γονιδιώματα με υψηλό λόγο GC, αυτές θα είναι πιο σπάνιες. Κατά συνέπεια το μήκος των «τυχαίων» ανοιχτών πλαισίων ανάγνωσης θα αυξάνει στα πλούσια σε GC γωνιδιώματα.

Οι ερευνητές λοιπόν, προχώρησαν βρίσκοντας όλα τα ORF από τα γνωστά βακτηριακά γονιδιώματα (34 εκείνη την εποχή). Κατόπιν, αφαίρεσαν τις πολύ ομόλογες πρωτεΐνες (Redundancy Reduction) και στη συνέχεια πραγματοποίησαν μια απλή σύγκριση με αναζήτηση ομοιότητας έναντι των πραγματικών (non-hypothetical) πρωτεϊνών της SwissProt (E-value<10^{-6}). Τα αποτελέσματα αναλύθηκαν με γραφικές παραστάσεις και στατιστικές μεθοδολογίες, ειδικά για να μπορέσει να εκτιμηθεί το ποσοστό του «υπερ-σχολιασμού» (over-annotation), δηλαδή των επιπλέον πρωτεϊνών που είχαν προβλεφθεί για το κάθε γονιδίωμα (Skovgaard, Jensen, Brunak, Ussery, & Krogh, 2001).



Εικόνα 11.2: *Ιστόγραμμα της κατανομής των μηκών των υποθετικών πρωτεϊνών που είχαν ομοιότητα με πραγματική πρωτεΐνη της SwissProt (συνεχής γραμμή), και αυτών που δεν είχαν ομοιότητα. (διακεκομμένη γραμμή).*

Τα αποτελέσματα της ανάλυσης ήταν εντυπωσιακά. Οι πρωτεΐνες των γονιδιωμάτων που είχαν ξεκάθαρη ομοιότητα με κάποια «σίγουρη» πρωτεΐνη της Swissprot, είχαν διαφορετική κατανομή του μήκους τους από αυτές οι οποίες δεν εμφάνισαν τέτοια ομοιότητα. Για την ακρίβεια, η δεύτερη ομάδα, αυτές που πιθανότατα ήταν αποτελέσματα ψευδών προβλέψεων των προγραμμάτων εύρεσης γονιδίων, ήταν μικρότερες κατά μέσο όρο και με μια κατανομή που προσέγγιζε τη γεωμετρική (Εικόνα 11.2). Το πρόβλημα αυτό μας θυμίζει αρκετά το πρόβλημα της «μίξης κατανομών» (mixture of distributions) στη στατιστική και στην ομαδοποίηση. Παρόλο που η διαφορά ήταν εμφανής οπτικά, δεν είναι και τόσο εύκολο να προβλεφθεί η ταυτότητα μιας συγκεκριμένης πρωτεΐνης, γιατί οι κατανομές δεν διαχωρίζονται επαρκώς. Για παράδειγμα, για μια πολύ μικρή (π.χ. 100 αμινοξέα) ή για μια αρκετά μεγάλη πρωτεΐνη (π.χ. 500 αμινοξέα), είναι αρκετά εύκολο να κάνουμε μια πρόβλεψη, αλλά για τις περισσότερες πρωτεΐνες που έχουν μήκος στην περιοχή 200 με 300 αμινοξέα, αυτό δεν είναι εύκολο. Παρ' όλα αυτά, είναι εύκολο αλλά και σημαντικό να προβλεφθεί με μεγάλη ακρίβεια ο συνολικός αριθμός των πρωτεϊνών που θα ανήκουν σε κάθε ομάδα. Με βάση αυτή την εκτίμηση, οι ερευνητές προχώρησαν σε μια απλή γραφική παράσταση του ποσοστού GC με το ποσοστό υπερ-σχολιασμού (δηλαδή, το ποσοστό των «ψεύτικων» πρωτεϊνών που αναμένουμε να υπάρχουν στο γονιδίωμα).



Εικόνα 11.3: *Η σχέση του ποσοστού υπερ-σχολιασμού (δηλαδή των επιπλέον αλληλουχιών που έχουν προσδιοριστεί λανθασμένα ως πραγματικές πρωτεΐνες) με το ποσοστό GC των γονιδιωμάτων.*

Τα αποτελέσματα (Εικόνα 11.3), επιβεβαίωσαν πλήρως το θεωρητικό μοντέλο, καθώς τα γονιδιώματα με υψηλό GC ήταν και αυτά με το μεγαλύτερο ποσοστό «ψεύτικων» πρωτεϊνών. Το ένα γονιδίωμα που ξεφεύγει από το διάγραμμα, καθώς εμφανίζει ένα ιδιαίτερα υψηλό ποσοστό «ψεύτικων» πρωτεϊνών, περίπου μία στις δύο πρωτεΐνες (τέτοιες παρατηρήσεις ονομάζονται outlier στη στατιστική), βρέθηκε μετά από αναζήτηση στη βιβλιογραφία ότι ανήκε στο βακτήριο *A. pernix*, στον προσδιορισμό του γονιδιώματος του οποίου, οι ερευνητές δεν χρησιμοποίησαν καν κάποιο πρόγραμμα εύρεσης γονιδίων, αλλά απλά ονόμασαν «πρωτεΐνη» κάθε ανοιχτό πλαίσιο ανάγνωσης. Φυσικά, δεν πρέπει να ξεχνάμε ότι η απλή αυτή γραμμική σχέση δεν εξηγεί 100% την μεταβλητότητα του δείγματος (με άλλα λόγια, τα σημεία είναι διασκορπισμένα αρκετά εκατέρωθεν της ευθείας γραμμής). Ο λόγος είναι προφανής: μιλάμε για διαφορετικά γονιδιώματα, τα οποία προσδιορίστηκαν και αναλύθηκαν από διαφορετικές ερευνητικές ομάδες, σε διαφορετικές εποχές, και με χρήση διαφορετικών εργαλείων. Αν μια παρόμοια ανάλυση γινόταν, για παράδειγμα, στα γονιδιώματα που προσδιορίστηκαν πρόσφατα, θα περιμέναμε ότι το ποσοστό των «ψεύτικων» πρωτεϊνών θα ήταν μειωμένο, γιατί τα σύγχρονα εργαλεία εύρεσης γονιδίων λειτουργούν καλύτερα.

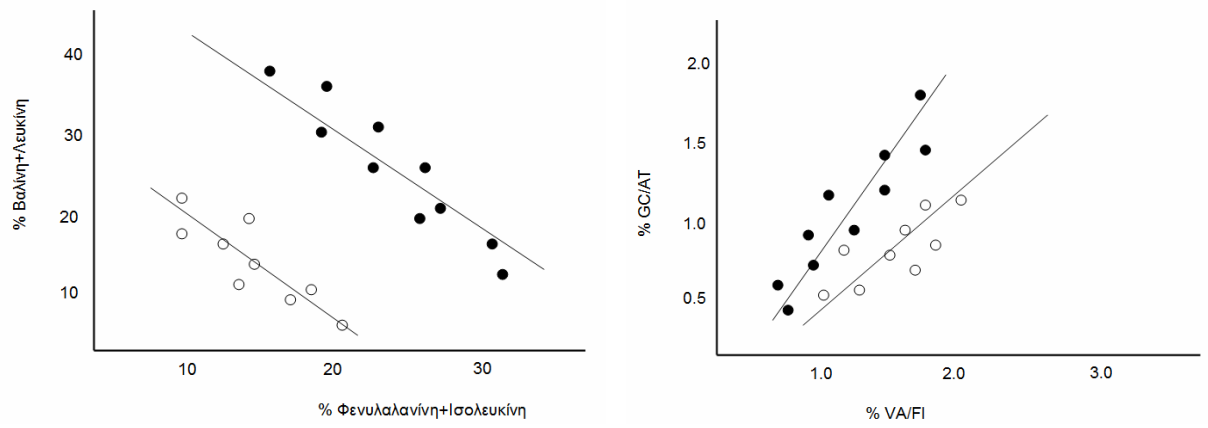
Ένα άλλο παράδειγμα γονιδιωματικής ανάλυσης που δίνει πολύ ενδιαφέροντα συμπεράσματα που σχετίζονται με το ποσοστό GC, αφορά τις διαμεμβρανικές πρωτεΐνες. Η εύρεση των α-ελικοειδών διαμεμβρανικών πρωτεϊνών στηρίζεται στη, με διάφορους τρόπους, αναζήτηση περιοχών πλούσιων σε υδρόφοβα κατάλοιπα. Τα γονιδιώματα όμως διαφέρουν στο ποσοστό GC όπως είδαμε πριν. Επιπλέον, τα κωδικόνια των υδρόφοβων αμινοξέων περιέχουν GC σε διαφορετικό βαθμό. Για την ακρίβεια, όπως φαίνεται στην Εικόνα 11.4, η Αλανίνη και η Γλυκίνη έχουν περισσότερα GC στα κωδικόνια τους, η Βαλίνη και η Λευκίνη έχουν ίδιο αριθμό GC και AT, ενώ η Ισολευκίνη και η Φενυλαλανίνη έχουν περισσότερο AT. Επιπλέον δε, η Αλανίνη και η Γλυκίνη, αν και υδρόφοβα αμινοξέα είναι «λιγότερο» υδρόφοβα σύμφωνα με τις περισσότερες κλίμακες υδροφοβικότητας. Κατά συνέπεια, ένας «γενικής χρήσης» αλγόριθμος πρόγνωσης μπορεί να υπερ- ή υπό-εκτιμά την πρόγνωση διαμεμβρανικών τμημάτων όταν εφαρμόζεται σε πρωτεΐνες από οργανισμούς με διαφορετικό GC. Θεωρητικά, αναμένουμε ότι σε γονιδιώματα πλούσια σε GC, η Αλανίνη και

η Γλυκίνη θα βρίσκονται σε μεγαλύτερη συχνότητα και κατά συνέπεια οι προγνώσεις για διαμεμβρανικά τμήματα θα είναι πιο «σπάνιες» (ή, πιο δύσκολες).

Αμινοξύ	**Κωδικόνιο**	**min(A+T)**	**min(G+C)**
Αλανίνη (Ala)	G-C-X	0	2
Γλυκίνη (Gly)	G-G-X	0	2
Βαλίνη (Val)	G-T-X	1	1
Λευκίνη (Leu)	C-T-X, T-T-[AG]	1	1
Ισολευκίνη (Ile)	A-T-[ACT]	2	0
Φενυλαλανίνη (Phe)	T-T-[CT]	2	0

Εικόνα 11.4: *Κατανομή των κωδικωνίων για τα υδρόφοβα αμινοξέα.*

Στην ανάλυση των τότε γνωστών γονιδιωμάτων, οι ερευνητές χρησιμοποίησαν ένα σχετικά απλό τρόπο εύρεσης των διαμεμβρανικών τμημάτων (έναν αλγόριθμο κυλιόμενου παραθύρου με χρήση κλίμακας υδροφοβικότητας), και συσχέτισαν τα ποσοστά Βαλίνης και Λευκίνης (VL) με αυτά της Φενυλαλανίνης και της Ισολευκίνης (FI). Το ίδιο έγινε όχι μόνο για τις ακολουθίες των διαμεμβρανικών τμημάτων, αλλά και για το σύνολο του πρωτεόματος. Τα αποτελέσματα έδειξαν μια πολύ ισχυρή συσχέτιση, όπως αναμενόταν. Επιπλέον δε, ο λόγος αυτός (VL/FI) εμφανίζει μια ξεκάθαρη συσχέτιση με το λόγο GC/AT. Με άλλα λόγια, επαληθεύεται η αρχική υπόθεση ότι τα γονιδιώματα που είναι πλούσια σε GC, έχουν συγκριτικά περισσότερες Βαλίνες και Αλανίνες σε σύγκριση με Ισολευκίνες και Φενυλαλανίνες. Αυτό έχει σαν συνέπεια τα διαμεμβρανικά τμήματα των διαμεμβρανικών πρωτεϊνών που κωδικοποιούνται σε αυτά τα γονιδιώματα, να είναι λιγότερο υδρόφοβα σε σχέση με αυτά των πρωτεϊνών που προέρχονται από οργανισμούς φτωχούς σε GC. Όλα τα παραπάνω, σημαίνουν ότι σε ακραίες περιπτώσεις, αυτές οι διαφορές θα πρέπει να λαμβάνονται υπόψη και (αν είναι δυνατόν) η πληροφορία αυτή να ενσωματωθεί ακόμα και στους αλγόριθμους πρόγνωσης διαμεμβρανικών τμημάτων.



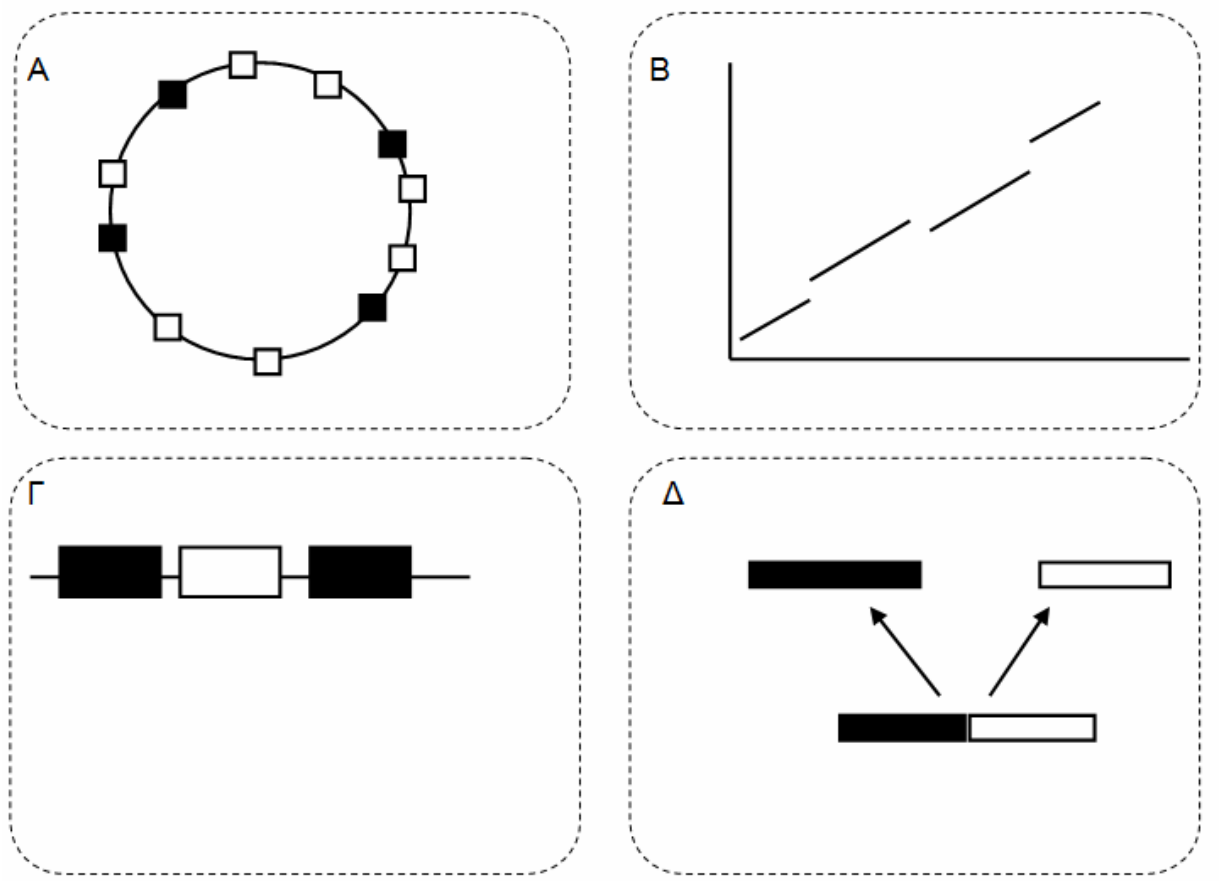
Εικόνα 11.5: *Αριστερά, η συσχέτιση του ποσοστού Βαλίνης καιΛευκίνης με το ποσοστό Φενυλαλανίνης και Ισολευκίνης. Δεξιά, η συσχέτιση του λόγου αυτών των δύο ποσοστών με τον λόγο GC/AT. Τα μαύρα σημεία αντιστοιχούν στις διαμεμβρανικές πρωτεΐνες, ενώ τα λευκά στο σύνολο του γονιδιώματος.*

11.2. Συγκριτική Γονιδιωματική

Η συγκριτική γονιδιωματική, πάει ένα βήμα παραπέρα την υπολογιστική ανάλυση των γονιδιωμάτων. Αντί να εστιάζει μόνο στα συνολικά στατιστικά μέτρα από κάθε γονιδίωμα, όπως π.χ. το ποσοστό GC ή κάποιο άλλο μέτρο, επιχειρεί να χρησιμοποιήσει τη βασική αρχή της φυλογενετικής ανάλυσης, ότι δηλαδή τα γονιδιώματα όλων των οργανισμών προέρχονται από προγονικές μορφές και έχουν διαμορφωθεί έτσι όπως είναι σήμερα μετά από αλλεπάλληλες αλλαγές που έγιναν μέσα σε εκατομμύρια χρόνια. Οι αλλαγές αυτές, αφορούν τόσο τα αντίστοιχα ορθόλογα γονίδια και τις αλληλουχίες τους, όσο και το ίδιο το γονιδίωμα, τη δομή του, και τη διάταξή των γονιδίων πάνω σε αυτό.

Βασικά, η συγκριτική γονιδιωματική κάνει χρήση των κλασικών αλγορίθμων στοίχισης και εύρεσης ομοιότητας μεταξύ γονιδίων ή/και πρωτεϊνών, αλλά συνδυάζοντας αυτή την πληροφορία με τη δομή του

γονιδιώματος και τη διάταξη των γονιδίων πάνω σε αυτό, καταφέρνει να εξάγει πολύ σημαντικά συμπεράσματα, που δεν θα μπορούσαν να έχουν εξαχθεί με άλλον τρόπο (ούτε καν με πρόγνωση). Οι βασικές τεχνικές που χρησιμοποιούνται στη συγκριτική γονιδιωματική είναι τέσσερις (Tsoka & Ouzounis, 2000)



Εικόνα 11.6: *Οι τέσσερις κλασικές μέθοδοι συγκριτικής γονιδιωματικής. Η μέθοδος «αφαίρεσης» γονιδίων (Α), η μέθοδος στοίχισης ολόκληρων γονιδιωμάτων(Β), η μέθοδος σύγκρισης της σειράς των γονιδίων (Γ) και η μέθοδος εντοπισμού προϊόντων γονιδιακής σύντηξης (Δ).*

Η μέθοδος «αφαίρεσης» γονιδίων, στην οποία συγκρίνονται σε μια σειρά οργανισμούς τα κοινά γονίδια και εντοπίζονται τα μοναδικά γονίδια.

Η μέθοδος σύγκρισης της σειράς των γονιδίων, σύμφωνα με την οποία εντοπίζονται γονίδια που έχουν την τάση να βρίσκονται κοντά σε όλα τα υπό μελέτη γονιδιώματα,

Η μέθοδος στοίχισης ολόκληρων γονιδιωμάτων, σύμφωνα με την οποία στοιχίζονται ολόκληρα γονιδιώματα και εντοπίζονται οι περιοχές στις οποίες έχουν μεγάλη ομοιότητα, και τέλος

Η μέθοδος εντοπισμού προϊόντων γονιδιακής σύντηξης, στην οποία εντοπίζονται με υπολογιστικό τρόπο γονίδια τα οποία σε κάποιον άλλον οργανισμό βρίσκονται ενωμένα (σύντηξη), λειτουργούν δηλαδή σαν ανεξάρτητες πρωτεϊνικές περιοχές (domains).

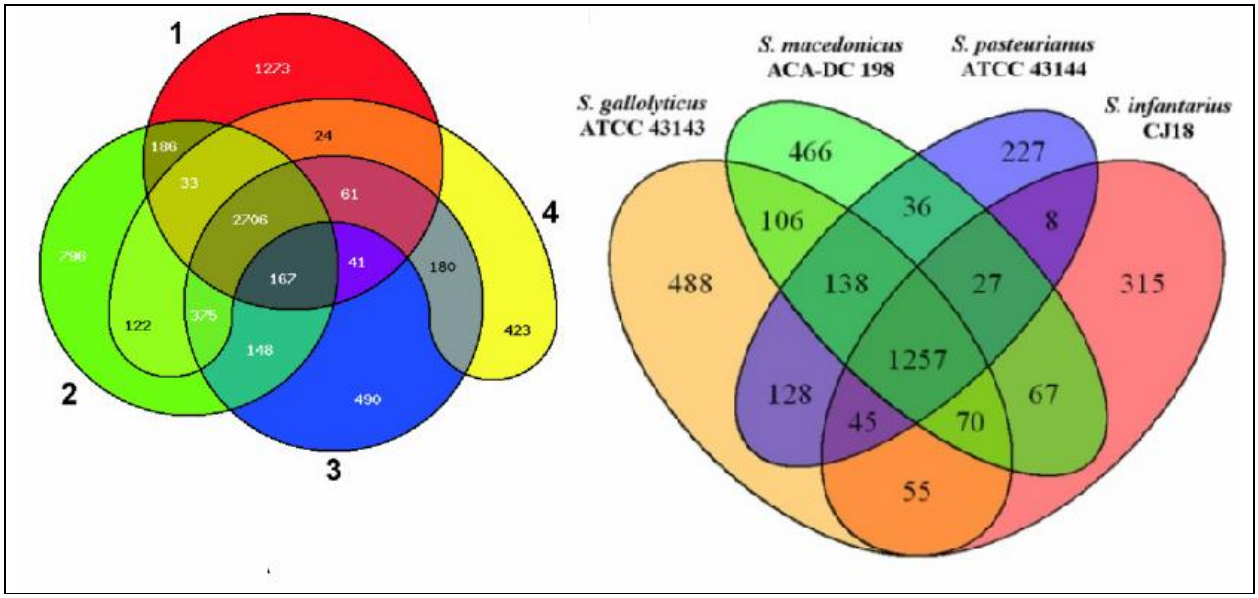
Όλες οι παραπάνω μεθοδολογίες λειτουργούν με χρήση της ομοιότητας των γονιδίων και των πρωτεϊνικών προϊόντων τους και κάνουν χρήση της πληροφορίας από τη σχετική θέση των γονιδίων (ή και την ίδια την ύπαρξή τους) σε διαφορετικούς οργανισμούς. Παρ' όλα αυτά, οι μεθοδολογίες αυτές εντοπίζουν διαφορετικού είδους λειτουργικές συσχετίσεις μεταξύ των γονιδίων. Προσφέρουν δηλαδή διαφορετικά αποτελέσματα, γι' αυτό και στη μεγάλη τους πλειοψηφία δρουν συμπληρωματικά, όπως θα δούμε παρακάτω.

11.2.1 Η μέθοδος «αφαίρεσης» γονιδίων

Η μέθοδος αυτή, βασίζεται στην εύρεση κοινών, ομόλογων δηλαδή, γονιδίων σε μια σειρά υπό σύγκριση οργανισμών. Η βασική αρχή, είναι η γνωστή από παλιά αρχή στη φυλογενετική, ότι οι πιο συγγενικοί

οργανισμοί θα έχουν και περισσότερα κοινά χαρακτηριστικά (δηλαδή, γονίδια στην περίπτωση μας). Με την ενσωμάτωση της γνώσης για τη μοριακή λειτουργία αυτών των γονιδίων, μπορούμε να εντοπίσουμε ποια γονίδια είναι χαρακτηριστικά για μια ομάδα οργανισμών και να εξάγουμε χρήσιμα συμπεράσματα για τη φυλογένεση (λειτουργούν δηλαδή ως απομορφικοί χαρακτήρες). Εξετάζοντας τα γονίδια που είναι μοναδικά σε κάποιον οργανισμό (ή σε κάποιους οργανισμούς) μπορούμε επίσης να εντοπίσουμε ειδικές λειτουργίες που επιτελεί αυτός ο οργανισμός για να επιβιώσει (π.χ. τα μεθανότροφα βακτήρια έχουν ειδικά μεταβολικά μονοπάτια για να αποικοδομούν το μεθάνιο που βρίσκεται σε περίσσεια στο περιβάλλον τους).

Οι μέθοδοι αναπαράστασης τέτοιων αναλύσεων, ξεκινούν από απλά διαγράμματα Venn και φτάνουν μέχρι περίπλοκες αναπαραστάσεις μεταβολικών δρόμων στις οποίες ο κάθε οργανισμός απεικονίζεται με κάποιο χρώμα, έτσι ώστε να φανεί ποιοι οργανισμοί έχουν κάποια συγκεκριμένα ένζυμα ή άλλα μοριακά συστήματα.



Εικόνα 11.7: *Παραδείγματα διαγραμμάτων Venn. Αριστερά: Σύγκριση στελεχών του Xanthomonas Oryzae με το EDGAR (Blom et al., 2009) Δεξιά: Σύγκριση διαφορετικών ειδών Streptococcous με το R (Papadimitriou et al., 2014)*.

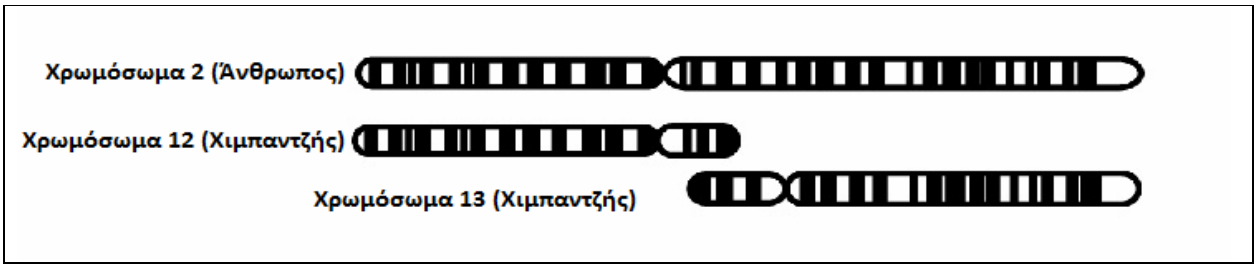
Τέτοιου είδους αναλύσεις, έχουν χρησιμοποιηθεί για να διαλευκανθεί το ερώτημα που αφορά τον τελευταίο κοινό πρόγονο όλων των σύγχρονων οργανισμών (Last Universal Common Ancestor-LUCA). Οι αναλύσεις ξεκίνησαν με τη μελέτη του οργανισμού με το μικρότερο γονιδίωμα, του βακτηρίου *Mycoplasma genitalium* το οποίο είναι υποχρεωτικό ενδοκυτταρικό παράσιτο και κωδικοποιεί μόλις 468 γονίδια που παράγουν πρωτεΐνες. Ακόμα και σε σύγκριση με κάποιο άλλο βακτήριο, π.χ. με το *Haemophilus influenzae* (1703 γονίδια) γίνεται εμφανές ότι μόνο 240 γονίδια του *M. genitalium* έχουν ορθόλογα γονίδια στον *H. influenzae*. Το ερώτημα λοιπόν ήταν αν ο LUCA ήταν ένας οργανισμός με λίγα γονίδια (όπως π.χ. το *Mycoplasma*) ή αν, αντίθετα, ήταν οργανισμός με περισσότερα γονίδια (όπως τα περισσότερα βακτήρια) και τελικά η εξέλιξη οδήγησε κάποιους οργανισμούς να χάσουν τα γονίδια αυτά και άλλους να αποκτήσουν κάποια νέα. Συγκριτικές αναλύσεις, με κάποιες παραδοχές (όπως π.χ. ότι δεν αναμένουμε σε όλους τους οργανισμούς να είναι συντηρημένα όλα τα γονίδια), έδειξε ότι μάλλον η δεύτερη εκδοχή είναι η σωστή. Για παράδειγμα, όταν στην ανάλυση συμπεριλήφθηκαν μόνο προκαρυώτες, βρέθηκε ότι ο κοινός πρόγονος όλων των οργανισμών πρέπει να είχε γονίδια μεταξύ 1006 και 1189, ενώ όταν συμπεριλήφθηκαν και οι ευκαρυώτες, ο αριθμός ανέβηκε στο 1344 με 1529, -αριθμοί που είναι πιο κοντά στο μέσο όρο των σημερινών βακτηρίων παρά στο ελάχιστο (δηλαδή στο *Mycoplasma*) (Ouzounis, Kunin, Darzentas, & Goldovsky, 2006).

Παρόμοιες αναλύσεις, έχουν μεγάλο ενδιαφέρον και στη λεγόμενη «εξωβιολογία», τον κλάδο δηλαδή που μελετάει θεωρητικά το πώς αναμένουμε να είναι οι οργανισμοί που ενδεχομένως βρεθούν σε άλλους πλανήτες, αλλά και στη συνθετική βιολογία και τη γενετική μηχανική. Για παράδειγμα, τέτοιου είδους αναλύσεις, έκαναν δυνατό τον υπολογισμό των απαραίτητων γονιδίων που απαιτούνται για να συντηρήσουν τη ζωή σε ένα βακτήριο και εφαρμόστηκαν πρόσφατα όταν επιστήμονες συνέθεσαν εξ' ολοκλήρου ένα

βακτηριακό γονιδίωμα 1Mbp και το ενσωμάτωσαν σε ένα βακτηριακό κύτταρο από το οποίο είχαν αφαιρέσει το γονιδίωμα. Το «νέο» βακτήριο, το οποίο χρησιμοποιεί αποκλειστικά το συνθετικό DNA (*Mycoplasma mycoides* JCVI-syn1.0), είχε τις αναμενόμενες φαινοτυπικές λειτουργίες και ήταν ικανό να αναπαράγεται (Gibson et al., 2010).

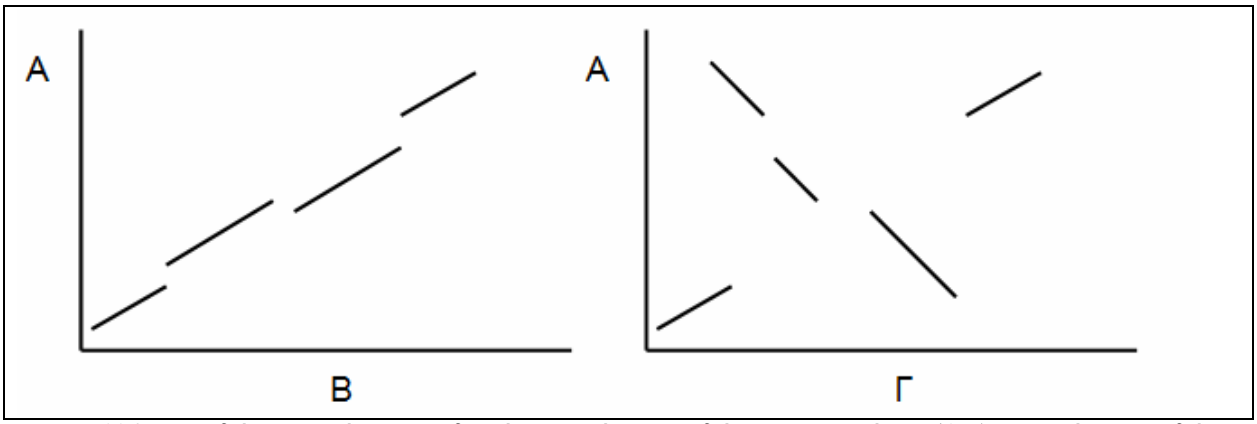
11.2.2 Η μέθοδος στοίχισης ολόκληρων γονιδιωμάτων

Η μέθοδος αυτή βασίζεται στην ίδια αρχή με τις στοιχίσεις αλληλουχιών (οι συγγενικοί οργανισμοί είναι πιο πιθανό να έχουν μεγάλες ομοιότητες στο γονιδίωμα τους). Με τη μέθοδο αυτή στοιχίζονται ολόκληρα γονιδιώματα και εντοπίζονται οι περιοχές στις οποίες έχουν μεγάλη ομοιότητα. Τέτοιες τεχνικές σε πιο πρώιμη μορφή ήταν γνωστές από παλιά, π.χ. από παρατηρήσεις ότι το ανθρώπινο DNA υβριδοποιείται με το αντίστοιχο του χιμπατζή, είχε γίνει γνωστό ότι τα γονιδιώματα του ανθρώπου και των άλλων μεγάλων πιθήκων έχουν μεγάλη ομοιότητα. Παρόμοιες ανακαλύψεις είχαν γίνει και με τη χρήση καρυότυπου, όταν για παράδειγμα έγινε γνωστό ότι το χρωμόσωμα 2 του ανθρώπου εμφανίζει μερική ομοιότητα με το χρωμόσωμα 12 και 13 του χιμπατζή, και έγινε κατανοητό ότι στο απώτατο παρελθόν είχε προκύψει από σύντηξη τελομερών.



Εικόνα 11.8: *Σύγκριση των χρωμοσωμάτων του ανθρώπου και του χιμπατζή.*

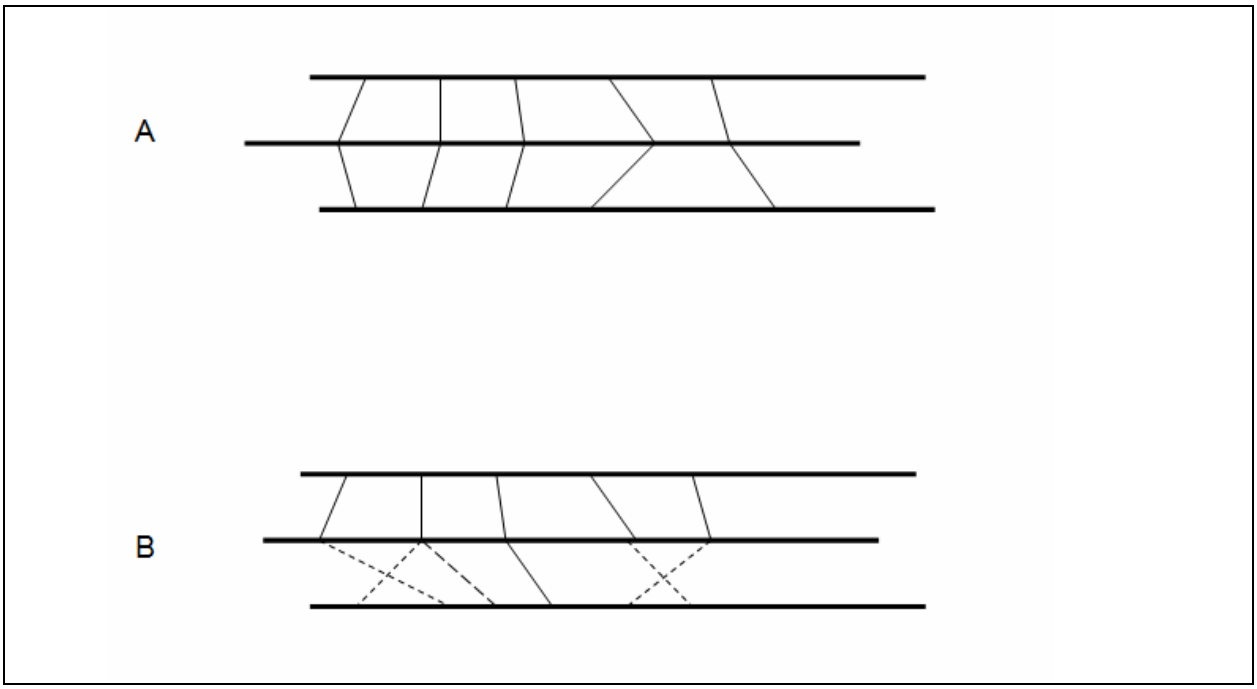
Στοίχιση ολόκληρων γονιδιωμάτων, μπορεί να γίνει με διαφορετικούς τρόπους. Ο πιο απλός είναι με μια παραλλαγή του γνωστού διαγράμματος σημείων (dot-plot) η οποία επεκτείνεται σε όλο το γονιδίωμα ή με κάποια επέκταση κάποιου γνωστού αλγόριθμου στοίχισης (όπως το BLAST) η οποία να επιτρέπει χρήση μεγάλων ακολουθιών. Οι πιο σύγχρονες μεθοδολογίες, συνδυάζουν τους αλγορίθμους τοπικής ή ολικής στοίχισης (για κάθε ζευγάρι ομόλογων γονιδίων) με τη θέση των γονιδίων αυτών στο γονιδίωμα, δείχνοντας π.χ. με διαφορετικό χρωματισμό τα ζευγάρια, ενώ κάποιες από τις τεχνικές αυτές επιτρέπουν και πολλαπλή στοίχιση. Όπως γίνεται εύκολα αντιληπτό, οι τεχνικές αυτές είναι πολύ πιο εύκολο να εφαρμοστούν σε βακτηριακά ή ϊικά γονδιώματα, τόσο γιατί είναι πιο μικρά όσο και γιατί είναι ενιαία, καθώς τα πολλαπλά χρωμοσώματα των ευκαρυωτικών οργανισμών απαιτούν σύγκριση ένα με ένα.



Εικόνα 11.9: *Παραδείγματα στοίχισης γονιδιωμάτων. Στοίχιση που δείχνει συνταινικότητα (Α-Β), και στοίχιση που δείχνει αναστροφή (Α-Γ).*

Οι μεθοδολογίες ολικής στοίχισης γονιδιωμάτων είναι δυνατό να δώσουν πολλές πληροφορίες για τις αλλαγές που έχουν συμβεί στα γονιδιώματα στο πέρασμα του εξελικτικού χρόνου. Για παράδειγμα, μια

στοίχιση και ένα διάγραμμα σημείων περίπου στο ύψος της διαγωνίου δείχνει την κοινή προέλευση και τη στενή σχέση των δύο οργανισμών. Επιπλέον, αλλαγές μεγάλης κλίμακας όπως αναστροφές και διπλασιασμοί είναι ιδιαίτερα εύκολο να εντοπιστούν. Τέλος, περιοχές μη ομοιότητας ανάμεσα σε 2 κατά κανόνα «όμοια» γονιδιώματα είναι δυνατό να δείξουν πρόσφατη απόκτηση γενετικού υλικού (είτε με οριζόντια μεταφορά είτε με κάποιον άλλο τρόπο ενσωμάτωσης DNA).

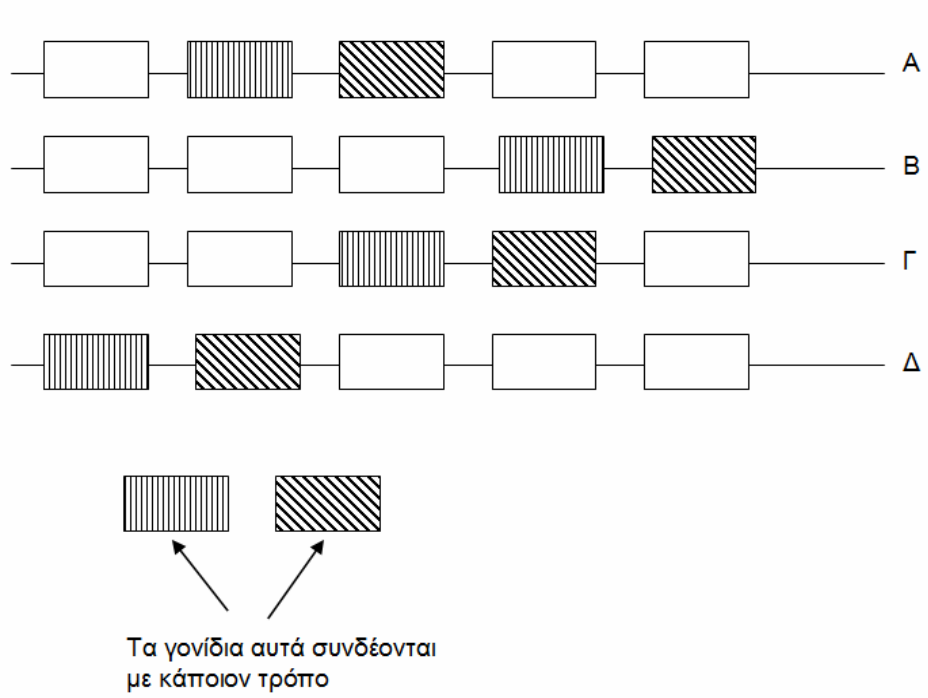


Εικόνα 11.10: *Τρόπος αναπαράστασης της πολλαπλής στοίχισης γονιδιωμάτων. Οι συνεχείς γραμμές δείχνουν ζεύγη γονιδίων που είναι σε συνταινία, ενώ οι διακεκομμένες γραμμές δείχνουν ζεύγη που έχουν υποστεί αναστροφή.*

11.2.3 Η μέθοδος σύγκρισης της σειράς των γονιδίων

Σύμφωνα με μέθοδο αυτή εντοπίζονται γονίδια που έχουν την τάση να βρίσκονται κοντά σε όλα ή στα περισσότερα τα υπό μελέτη γονιδιώματα. Η βασική αρχή της μεθόδου μοιάζει διαισθητικά με την αρχή της σύνδεσης στη γενετική, μόνο που εδώ χρησιμοποιείται σε μεγαλύτερη κλίμακα χρόνου. Η ιδέα είναι ότι γονίδια που βρίσκονται σε πολλούς οργανισμούς δίπλα-δίπλα, το κάνουν για κάποιο λόγο (π.χ. εκφράζονται μαζί ή συμμετέχουν σε κάποιο κοινό μεταβολικό μονοπάτι). Ειδικά στα βακτήρια, είναι γνωστό ότι ομάδες γονιδίων που συμμετέχουν στο ίδιο μονοπάτι, βρίσκονται οργανωμένα σε ομάδες που ονομάζονται οπερόνια, ομάδες οι οποίες εκφράζονται και ελέγχονται ταυτόχρονα.

Με τη μέθοδο αυτή είναι δυνατό να εντοπιστούν συσχετίσεις μεταξύ γονιδίων που κωδικοποιούν τελείως διαφορετικές πρωτεΐνες. Για παράδειγμα, αν υποθέσουμε ότι στο οπερόνιο της λακτόζης, ξέραμε τη λειτουργία της γαλακτοσιδάσης (lacZ) αλλά όχι αυτή της περμεάσης (lacY), με την παρατήρηση ότι σε μια σειρά από οργανισμούς τα δύο γονίδια βρίσκονται πάντα μαζί, θα μπορούσαμε να συμπεράνουμε ότι αποτελούν και τα δύο τμήμα κάποιου οπερονίου. Δεν θα ξέραμε φυσικά ακριβώς τη λειτουργία του νέου γονιδίου, αλλά συνδυάζοντας κάποιες απλές μεθόδους πρόγνωσης, όπως για παράδειγμα την πρόγνωση διαμεμβρανικών τμημάτων, θα βλέπαμε ότι πρόκειται για διαμεμβρανική πρωτεΐνη με 12 πιθανά διαμεμβρανικά τμήματα και αμέσως θα υποθέταμε ότι πρόκειται για κάποιον διαμεμβρανικό υποδοχέα που πιθανότατα εμπλέκεται στο μεταβολισμό της λακτόζης. Μια τόσο λεπτομερής πρόβλεψη για τη λειτουργία μιας πρωτεΐνης δεν θα μπορούσε με κανέναν τρόπο να γίνει δυνατή με χρήση μόνο της ακολουθίας της, αλλά βλέπουμε ότι αυτό συμβαίνει όταν χρησιμοποιήσουμε την πληροφορία από τη σειρά των γονιδίων και τη συντήρησή της στα γονδιώματα.

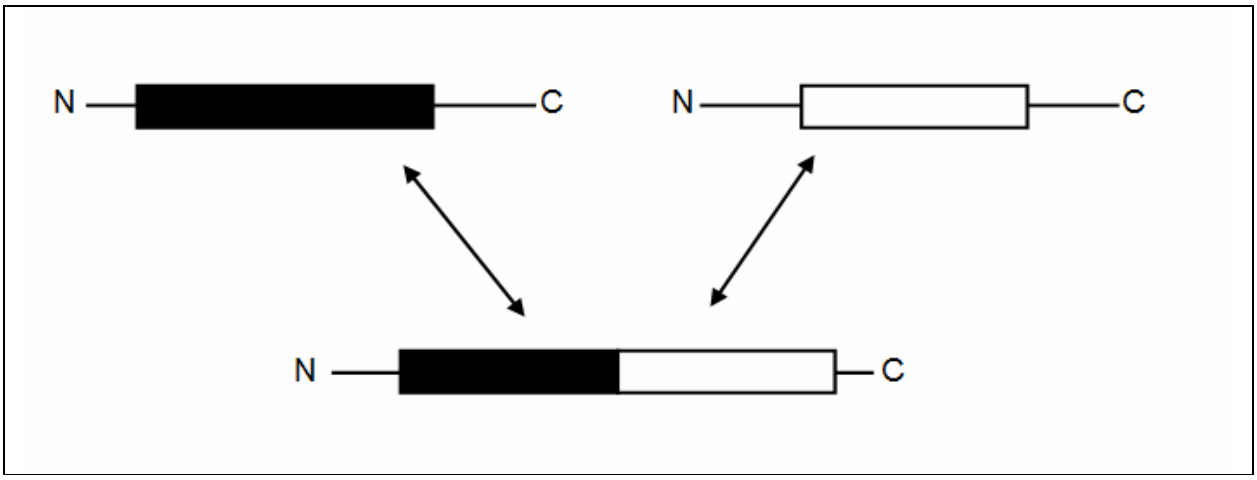


Εικόνα 11.11: *Η μέθοδος της σύγκρισης της σειράς των γονιδίων.*

11.2.4 Η μέθοδος εντοπισμού προϊόντων γονιδιακής σύντηξης

Η βασική αρχή αυτής της μεθόδου βασίζεται στη σπονδυλωτή φύση των πρωτεϊνών, δηλαδή, στην ύπαρξη ανεξάρτητων δομικών και λειτουργικών περιοχών (domains). Έτσι, με τη μέθοδο αυτή εντοπίζονται με υπολογιστικό τρόπο γονίδια ενός οργανισμού Α τα οποία σε κάποιον άλλον οργανισμό Β βρίσκονται ενωμένα, λειτουργούν δηλαδή σαν ανεξάρτητες περιοχές της ίδιας πρωτεΐνης. Η εξήγηση είναι ότι σε κάποια προγονική μορφή, είτε τα γονίδια βρίσκονταν ανεξάρτητα και συνενώθηκαν (σύντηξη γονιδίων) με το πέρασμα του χρόνου στον οργανισμό Β, είτε ότι σε κάποια προγονική μορφή τα γονίδια βρίσκονταν ενωμένα, ήταν δηλαδή πρωτεϊνικές περιοχές και κατόπιν στην πορεία της εξέλιξης αυτή η σχέση διακόπηκε στον οργανισμό Α (Enright, Iliopoulos, Kyrpides, & Ouzounis, 1999). Με τη μέθοδο αυτή, δεν μπορούμε να διακρίνουμε ποια από τις δύο εναλλακτικές όντως συνέβη, αλλά αυτό δεν αποτελεί πρόβλημα σε αυτές τις αναλύσεις, γιατί μπορούμε να εξάγουμε ούτως ή άλλως σημαντικά συμπεράσματα για πρωτεΐνες που ούτε ομοιότητα έχουν, αλλά και ούτε βρίσκονται κοντά στο γονιδίωμα.

Συνήθως τέτοιες περιπτώσεις γονιδίων αφορούν ένζυμα τα οποία εμπλέκονται στον ίδιο μεταβολικό δρόμο, πιθανότατα το προϊόν του ενός να είναι αντιδρόν στο άλλο και με αυτόν τον τρόπο διευκολύνονται οι μεταβολικές οδοί. Ένα κλασικό παράδειγμα, είναι η διυδροφολική αναγωγάση (DHFR) η οποία στους ευκαρυωτικούς οργανισμούς αποτελεί μια πρωτεΐνη με μια μοναδική πρωτεϊνική περιοχή, αλλά στα βακτήρια στο ίδιο μόριο συνυπάρχει και η λειτουργική περιοχή της θυμιδιλικής συνθετάσης (TS) η οποία συμμετέχει στο ίδιο μονοπάτι (σύνθεση νουκλεοτιδίων) και η οποία στους ευκαρυωτικούς οργανισμούς βρίσκεται σε διαφορετικό γονίδιο.

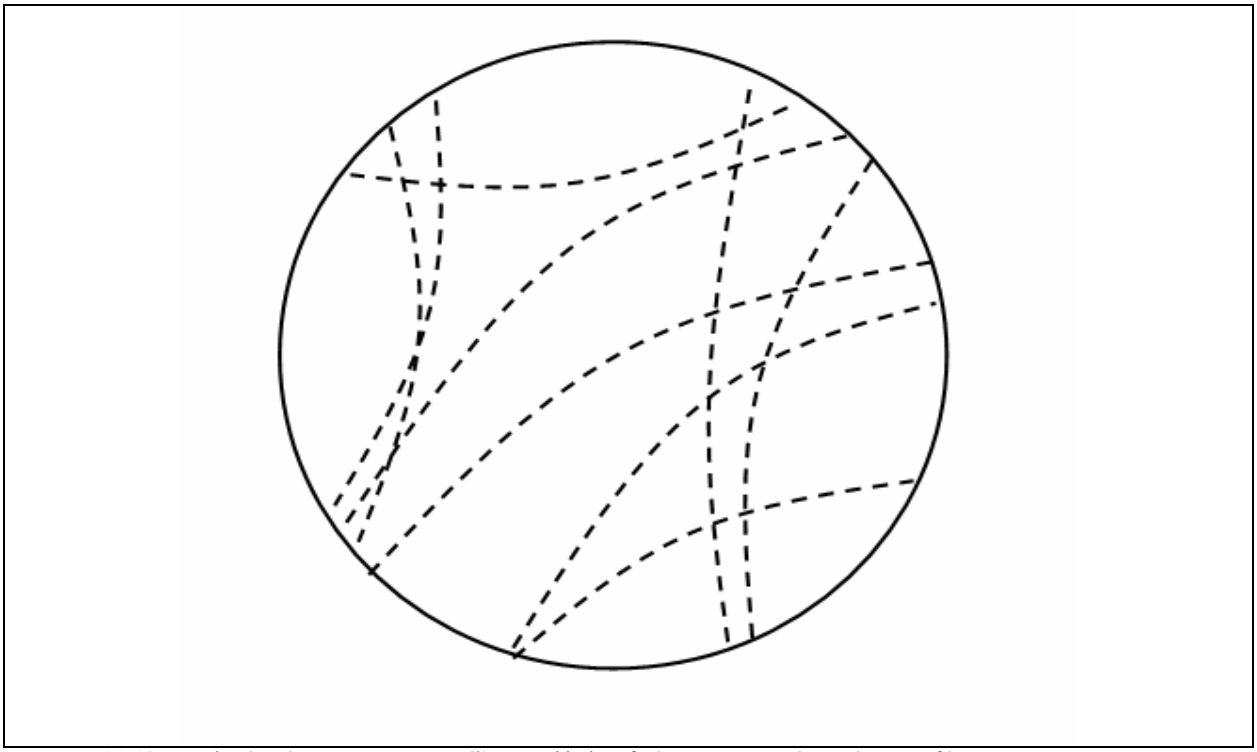


Εικόνα 11.12: *Η διαγραμματική απεικόνιση της μεθόδου σύντηξης γονιδίων.*

Η μέθοδος αυτή, είναι υπολογιστικά απαιτητική καθώς απαιτεί μία προς μία στοιχίσεις όλων των πρωτεϊνών του ενός οργανισμού, με όλες τις πρωτεΐνες του άλλου οργανισμού, ενώ απαιτείται και επιπλέον επεξεργασία για να διασφαλιστεί ότι οι δύο υποψήφιες πρωτεΐνες μοιάζουν μεν με μια άλλη πρωτεΐνη του άλλου οργανισμού αλλά σε διαφορετική περιοχή (δηλαδή, ότι δεν μοιάζουν μεταξύ τους). Από την άλλη μεριά, ένα σημαντικό πλεονέκτημα της μεθόδου σε σχέση με τις υπόλοιπες μεθόδους που αναλύθηκαν παραπάνω, είναι ότι καθώς δεν χρησιμοποιεί τη σειρά των γονιδίων, μπορεί να εφαρμοσθεί με ακριβώς τον ίδιο τρόπο σε κάθε είδους ζευγάρια ή ομάδες οργανισμών, ανεξάρτητα τόσο της εξελικτικής τους απόστασης όσο και του αριθμού χρωμοσωμάτων τους. Μπορεί με άλλα λόγια, να χρησιμοποιηθεί για τη σύγκριση του ανθρώπου με ένα βακτήριο και να δώσει χρήσιμα συμπεράσματα σε αντίθεση με τις προηγούμενες μεθόδους οι οποίες αποδίδουν καλύτερα και πρέπει να χρησιμοποιούνται κυρίως σε συγγενικούς οργανισμούς (και κατά βάση, σε βακτήρια).

Προφανώς το ποιος αλγόριθμος στοίχισης θα χρησιμοποιηθεί είναι ένα ανοιχτό ζήτημα (στην αρχική εργασία χρησιμοποιήθηκε το BLAST και έγινε εκ των υστέρων επεξεργασία με τον αλγόριθμο Smith-Waterman), όπως επίσης και το ποιες θα είναι οι παράμετροι (ποιο E-value θα χρησιμοποιηθεί σαν όριο για την εύρεση της ομοιότητας κ.ο.κ.). Κάτι ακόμα που πρέπει να γίνει σαφές, είναι ότι όσο περισσότεροι οργανισμοί χρησιμοποιούνται στην ανάλυση, τόσο περισσότερες πρωτεϊνικές αλληλεπιδράσεις θα εντοπιστούν στο δεδομένο γονιδίωμα επερώτησης. Αυτό γίνεται, γιατί έστω και σε έναν από τους οργανισμούς αυτούς να βρεθεί μια πρωτεΐνη με τις δύο περιοχές ενωμένες, τότε σε όλους τους υπόλοιπους θα αναγνωριστεί αυτή η «αλληλεπίδραση».

Προσοχή βέβαια χρειάζεται στη χρήση της έννοιας αυτής της «αλληλεπίδρασης» (interaction το ονομάζουν οι συγγραφείς), καθώς μπορεί να γίνει σύγχυση με τη φυσική αλληλεπίδραση των δύο πρωτεϊνών (τη φυσική επαφή). Παρόλο που κάτι τέτοιο είναι φυσικά πολύ πιθανό να συμβαίνει, η μέθοδος δεν προβλέπει απευθείας αυτό, αλλά μόνο μια λειτουργική αλληλεπίδραση, όμοια με αυτές που προβλέπει η προηγούμενη μέθοδος της «σειράς των γονιδίων». Βλέπουμε επομένως ότι οι μέθοδοι αυτές, λειτουργούν περισσότερο συμπληρωματικά παρά ανταγωνιστικά και αυτό είναι κάτι που πρέπει να το έχουμε πάντα στο μυαλό μας. Για παράδειγμα, η μέθοδος της σειράς των γονιδίων εντοπίζει λειτουργικά συνδεδεμένες πρωτεΐνες των οποίων τα γονίδια βρίσκονται πάντα μαζί, ενώ η μέθοδος σύντηξης γονιδίων εντοπίζει παρόμοιες συνδέσεις μεταξύ γονιδίων που βρίσκονται σε διαφορετικά μέρη στο γονιδίωμα. Επιπλέον δε, μπορεί με τη δημιουργία ενός κυκλικού χάρτη των αλληλεπιδράσεων να βρεθούν θερμές περιοχές (hot-spots), περιοχές δηλαδή με μεγάλη πυκνότητα τέτοιων αλληλεπιδράσεων και αυτές οι περιοχές να συσχετιστούν με πιθανά σημεία γονιδιωματικών ανακατατάξεων τα οποία θα εντοπιστούν με τη μέθοδο στοίχισης γονιδιωμάτων.



Εικόνα 11.13: *Κυκλικός χάρτης που απεικονίζει τις αλληλεπιδράσεις πρωτεϊνών σε ένα γονιδίωμα.*

11.3. Λογισμικό

Με τη ραγδαία ανάπτυξη της αλληλούχισης και την πρόοδο της γονιδιωματικής συνολικά, έχουν υλοποιηθεί τα τελευταία χρόνια εκατοντάδες εργαλεία συγκριτικής γονιδιωματικής που υλοποιούν κάποιους από τους αλγόριθμους και τις μεθόδους που αναλύσαμε παραπάνω, με έναν τρόπο εύκολο και βολικό για τον τελικό χρήστη (Edwards & Holt, 2013). Τα πιο πετυχημένα από αυτά τα εργαλεία συνδυάζουν την απλότητα με την ευελιξία καθώς προσφέρουν ένα ολοκληρωμένο περιβάλλον για να διευκολύνει πολλαπλές αναλύσεις, ενώ συνήθως προσφέρονται σαν διαδικτυακές εφαρμογές. Παρακάτω, αναλύουμε τα πιο γνωστά από αυτά τα εργαλεία.

Το **ACT** (https://www.sanger.ac.uk/resources/software/act/) είναι ένα εργαλείο βασισμένο στη Java το οποίο επιτρέπει την οπτικοποίηση γονιδιωμάτων και τη σύγκρισή τους. Για τη στοίχιση των αλληλουχιών χρησιμοποιεί το BLAST. Κατόπιν τα δύο γονιδιώματα και το αποτέλεσμα από την αναζήτηση του BLAST εισάγονται στο ACT για οπιτικοποίηση της σύγκρισης. Επιπλέον, το εργαλείο μπορεί να οπτικοποιήσει ταυτόχρονα περισσότερες από μία συγκρίσεις γονδιωμάτων. Οι ομόλογες περιοχές οι οποίες βρίσκονται στην ίδια κατεύθυνση στο γονιδίωμα χρωματίζονται με κόκκινο ενώ αυτές που βρίσκονται σε αντίθετες κατευθύνσεις, με μπλε. Η ένταση του χρωματισμού αντικατοπτρίζει το επίπεδο ομοιότητας. Τα πλεονεκτήματα του ACT περιλαμβάνουν τη δυνατότητα να απεικονίζει τη στοίχιση σε διαφορετικές μεγεθύνσεις (zoom in – zoom out) έτσι ώστε να μπορεί να απεικονίσει είτε τη στοίχιση ολόκληρου του γονιδιώματος, είτε να εστιάσει σε συγκεκριμένα γονίδια ενδιαφέροντος, αλλά και τη δυνατότητα που προσφέρει στο χρήστη να προσθέσει δικό του σχολιασμό για τα γονιδιώματα που αναλύονται (Carver et al., 2005).

Το **MAUVE** (http://darlinglab.org/mauve/mauve.html) είναι επίσης ένα εργαλείο βασισμένο στη Java κατάλληλο για συγκρίσεις γονιδιωμάτων. Διαθέτει ενσωματωμένο σύστημα απεικόνισης αλλά και τη δυνατότητα να εξάγει την πληροφορία από τη σύγκριση των γονιδιωμάτων σε διάφορες μορφές. Το MAUVE μπορεί να εργαστεί με δεδομένα αλληλούχισης νέας γενιάς, και έτσι παρέχει τη δυνατότητα να τοποθετήσει και να διατάξει μια σειρά από contigs απέναντι σε ένα ολόκληρο γονιδίωμα. Το εργαλείο δέχεται σαν είσοδο τις τελικές μορφές των γονιδιωμάτων και δημιουργεί μια στοίχιση αυτών. Αναγνωρίζει περιοχές με μεγάλη ομολογία και αναθέτει ένα ξεχωριστό χρώμα σε κάθε μία. Κατόπιν, κάθε γονιδίωμα απεικονίζεται σαν μια

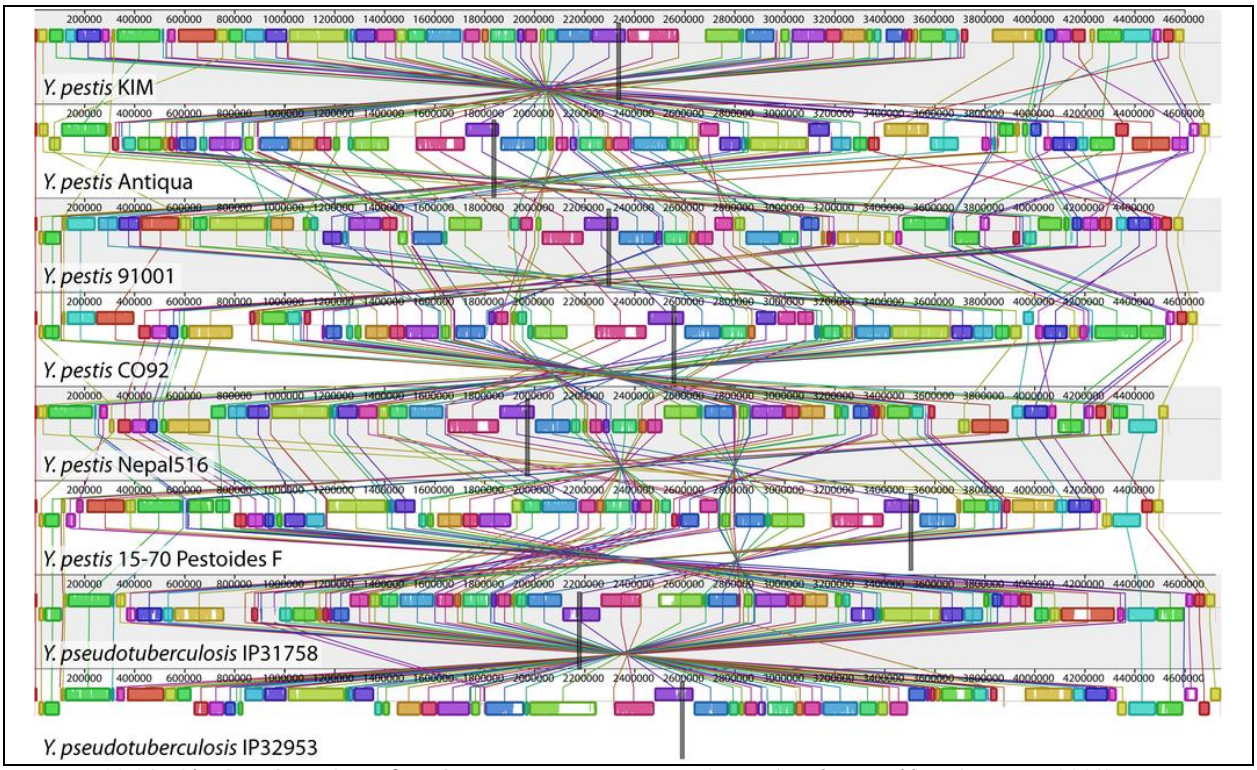
ακολουθία τέτοιων χρωματιστών περιοχών. Με τον τρόπο αυτό, γίνεται εύκολος ο εντοπισμός περιοχών με μοναδικά γονίδια. Επίσης, το MAUVE μπορεί χρησιμοποιηθεί (καθώς δουλεύει όπως αναφέραμε και με δεδομένα αλληλούχισης νέας γενιάς) και για τον εντοπισμό νουκλεοτιδικών πολυμορφισμών (SNPs) οι οποίοι μπορούν να χρησιμοποιηθούν παρακάτω για φυλογενετικές, εξελικτικές ή ιατρικές αναλύσεις (Darling, Mau, & Perna, 2010).

Το **EDGAR** (http://edgar.cebitec.uni-bielefeld.de) είναι ένα ακόμα σύγχρονο διαδικτυακό εργαλείο συγκριτικής γονιδωματικής το οποίο μπορεί να δεχτεί και δεδομένα αλληλούχισης. Το EDGAR είναι σχεδιασμένο έτσι ώστε να διευκολύνει το χρήστη και να απλοποιεί τις διαδικασίες. Ενσωματώνει τις βάσεις δεδομένων του NCBI και έχει στη βάση δεδομένων του όλα τα αποτελέσματα γνωστών γονιδιωμάτων προ-υπολογισμένα, ενώ έχει και τη δυνατότητα να απεικονίσει εξελικτικές και φυλογενετικές σχέσεις οι οποίες πολλές φορές διαλευκάνουν υποθέσεις σύγκρισης γονιδιωμάτων. Επίσης, υποστηρίζει μια σειρά από τρόπους απεικόνισης των αποτελεσμάτων όπως τα διαγράμματα στοίχισης γονιδιωμάτων (synteny plots) και διαγράμματα Venn για τα κοινά γονίδια (Blom, et al., 2009).

Το **CGAT** (http://mbgd.genome.ad.jp/CGAT/) είναι ένα ακόμα παρόμοιο εργαλείο που δημιουργήθηκε για να διευκολύνει τις συγκρίσεις συγγενικών βακτηριακών γονιδιωμάτων. Το CGAT λειτουργεί με αρχιτεκτονική client-server, στην οποία ο client AlignmentViewer (μια εφαρμογή Java) συνεργάζεται με τον DataServer (προγράμματα Perl). Το εργαλείο οπτικοποιεί στοιχίσεις γονιδιωμάτων τόσο στη μορφή των διαγραμμάτων σημείων όσο και στη μορφή των στοιχίσεων. Ο χρήστης μπορεί να προσθέσει πληροφορία στη σύγκριση, όπως για παράδειγμα την ύπαρξη επαναληπτικών αλληλουχιών και αλλαγές στη συχνότητα κωδικονίων έτσι ώστε να διευκολυνθεί στην εξαγωγή συμπερασμάτων. Εκτός από την οπτικοποίηση, ένα πλεονέκτημα του CGAT είναι η ευελιξία του καθώς επιτρέπει τη χρήση πολλών διαφορετικών αλγόριθμων στοίχισης γονιδιωμάτων (Uchiyama, Higuchi, & Kobayashi, 2006).

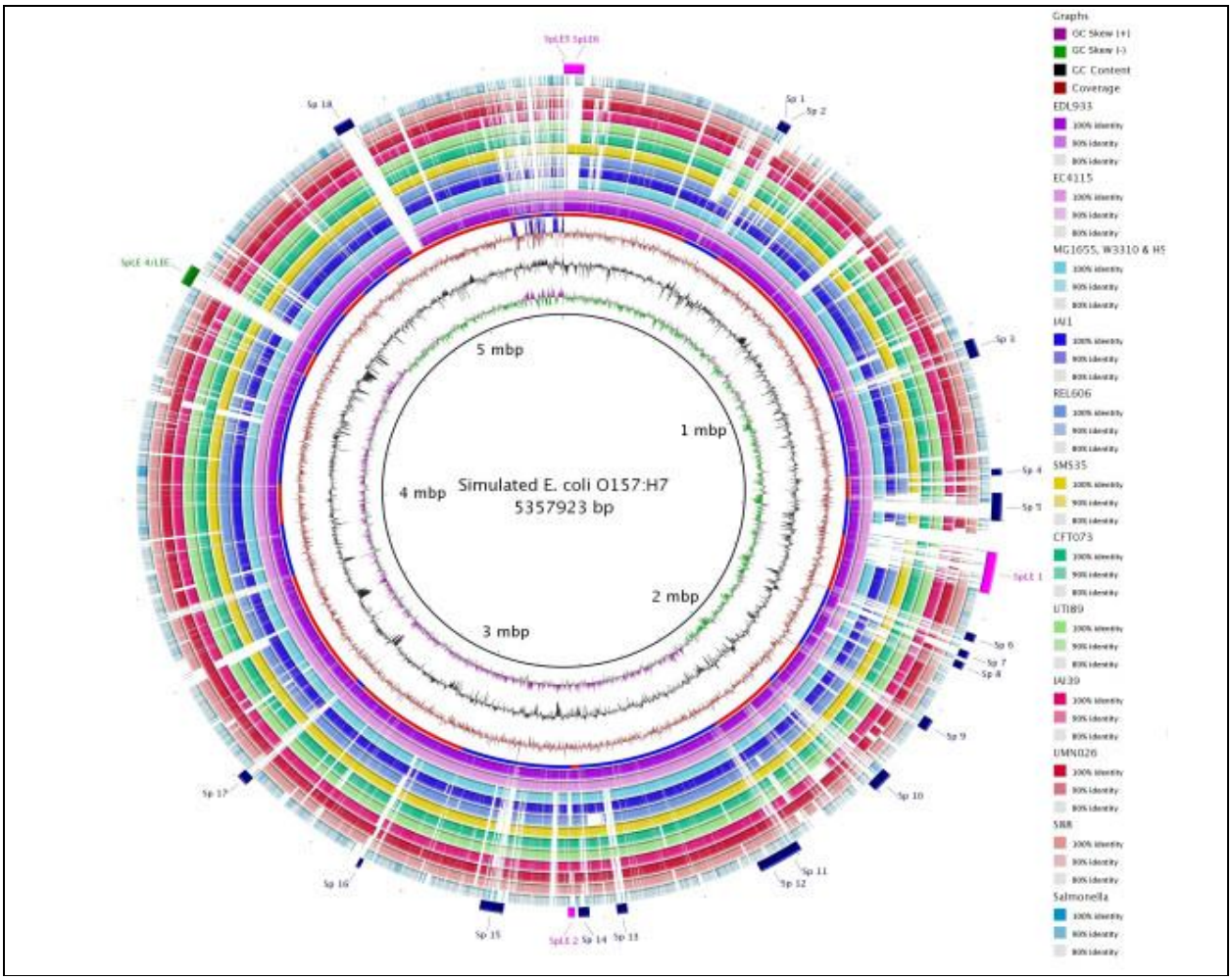
Το **BRIG** (BLAST Ring Image Generator, http://sourceforge.net/projects/brig/) είναι ένα άλλο εργαλείο βασισμένο στη Java, το οποίο οπτικοποιεί τη σύγκριση ενός γονιδιώματος αναφοράς με μία ή περισσότερες άλλες αλληλουχίες. Χρησιμοποιεί έναν ιδιαίτερο τρόπο οπτικοποίησης, σύμφωνα με τον οποίο τα γονιδιώματα αναπαρίστανται ως σειρές από επάλληλους κύκλους (δαχτυλίδια), με ειδικό χρωματισμό, για να δηλώνει την παρουσία μιας περιοχής ή ενός γονιδίου στο γονιδίωμα αναφοράς. Το BRIG είναι αρκετά ευέλικτο και μπορεί να χρησιμοποιηθεί για να απαντήσει πλήθος ερωτημάτων, ανάλογα με την επιλογή των γονιδιωμάτων υπό σύγκριση. Αυτό που πρέπει να τονιστεί είναι το γεγονός ότι η αναπαράσταση είναι εξαρτώμενη από το γονιδίωμα αναφοράς. Με άλλα λόγια, ενώ το εργαλείο απεικονίζει ποιες περιοχές είναι παρούσες ή απούσες από τα γονιδιώματα σύγκρισης, δεν μπορεί να δείξει περιοχές των γονιδιωμάτων αυτών που λείπουν από το γονιδίωμα αναφοράς. Γι' αυτό το λόγο η επιλογή του γονιδιώματος αναφοράς είναι ιδιαίτερα σημαντική (Alikhan, Petty, Ben Zakour, & Beatson, 2011).

Το **VISTA** (http://genome.lbl.gov/vista/index.shtml) ήταν ένα από τα πρώτα εργαλεία οπτικοποίησης στοιχίσεων γονιδιωμάτων και είχε παρουσιαστεί το 2000. Σήμερα, έχει εξελιχθεί σε μια ολοκληρωμένη σουίτα προγραμμάτων τα οποία καλύπτουν κάθε ανάγκη συγκριτικής ανάλυσης γονιδιωμάτων. Διαθέτει ειδικά εργαλεία για διάφορες συγκριτικές αναλύσεις γονιδιωμάτων, διασύνδεση με τις βάσεις δεδομένων γονιδιωμάτων, ενώ διαθέτει και αποθηκευμένα προ-υπολογισμένα αποτελέσματα για τα γνωστά γονιδιώματα (ακόμα και των σπονδυλοτών). Διαθέτει ειδικό σύστημα οπτικοποίησης (VISTA Browser) το οποίο επιτρέπει στο χρήστη να υποβάλει και το δικό του γονιδίωμα για ανάλυση στους διάφορους εξυπηρετητές (VISTA servers, rVista, mVISTA, phyloVISTA, gVISTA κ.ο.κ.) στους οποίους ο χρήστης μπορεί να επιτελέσει στοιχίσεις με διαφορετικούς αλγόριθμους, οπτικοποίηση με διαφορετικούς τρόπους, αλλά και ενσωμάτωση διαφορετικών ειδών πληροφορίας όπως φυλογενετικές σχέσεις, ρυθμιστικές περιοχές κ.ο.κ. (Frazer, Pachter, Poliakov, Rubin, & Dubchak, 2004). Μια επιπλέον δυνατότητα του VISTA είναι ότι διαθέτει και μια ανεξάρτητη (standalone) εφαρμογή με σχεδόν τις ίδιες δυνατότητες, το GenomeVISTA, το οποίο μπορεί να εγκατασταθεί ελεύθερα στον υπολογιστή του χρήστη και να εκτελέσει εκεί τις ίδιες λειτουργίες με τη διαδικτυακή εκδοχή, προσφέροντας μεγαλύτερη ασφάλεια των δεδομένων και ίσως και ταχύτητα (Poliakov, Foong, Brudno, & Dubchak, 2014).

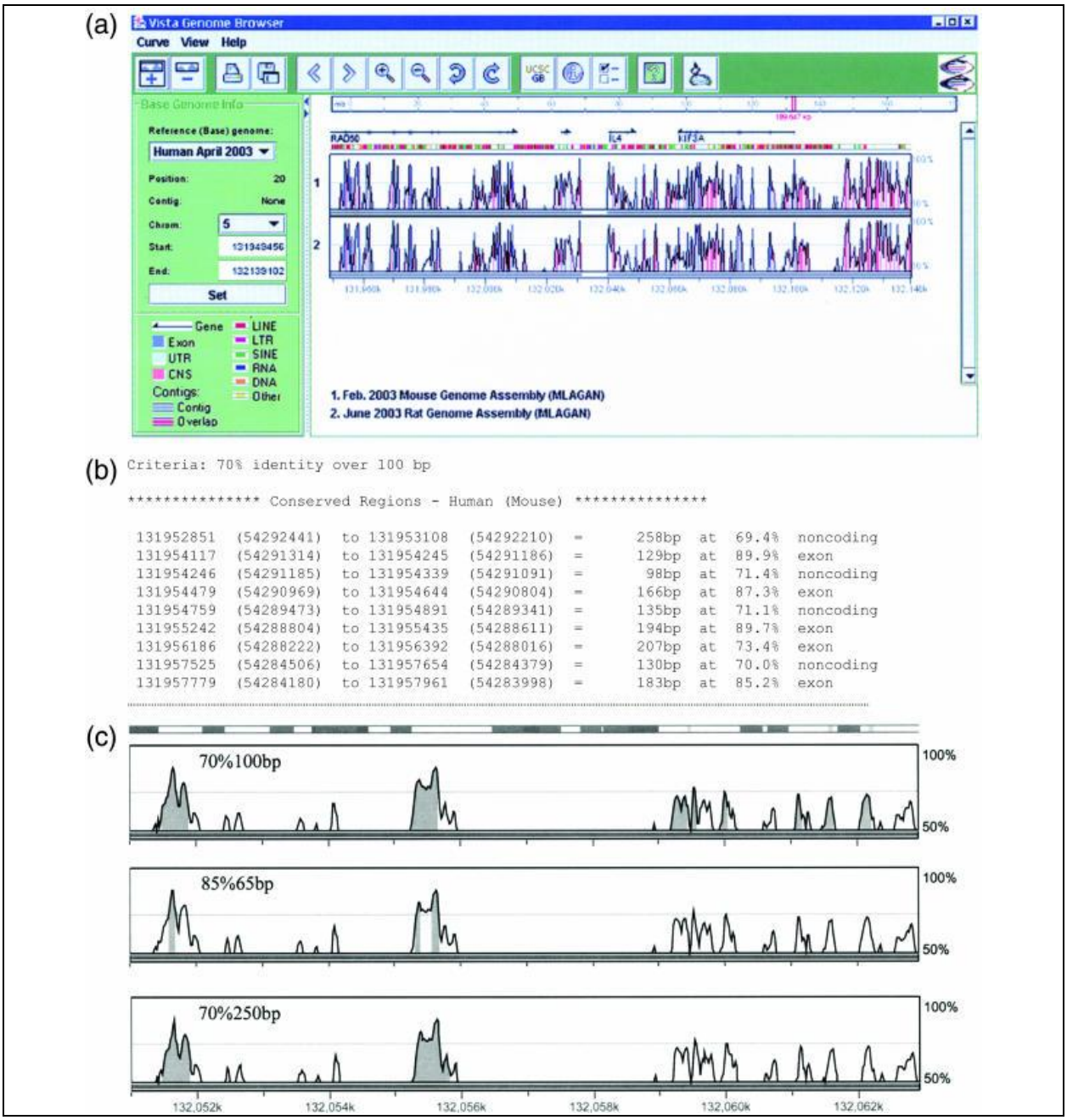


Εικόνα 11.14: *Ολική στοίχιση 8 γονιδιωμάτων της Yersinia με το MAUVE (Darling, Miklos, & Ragan, 2008).*

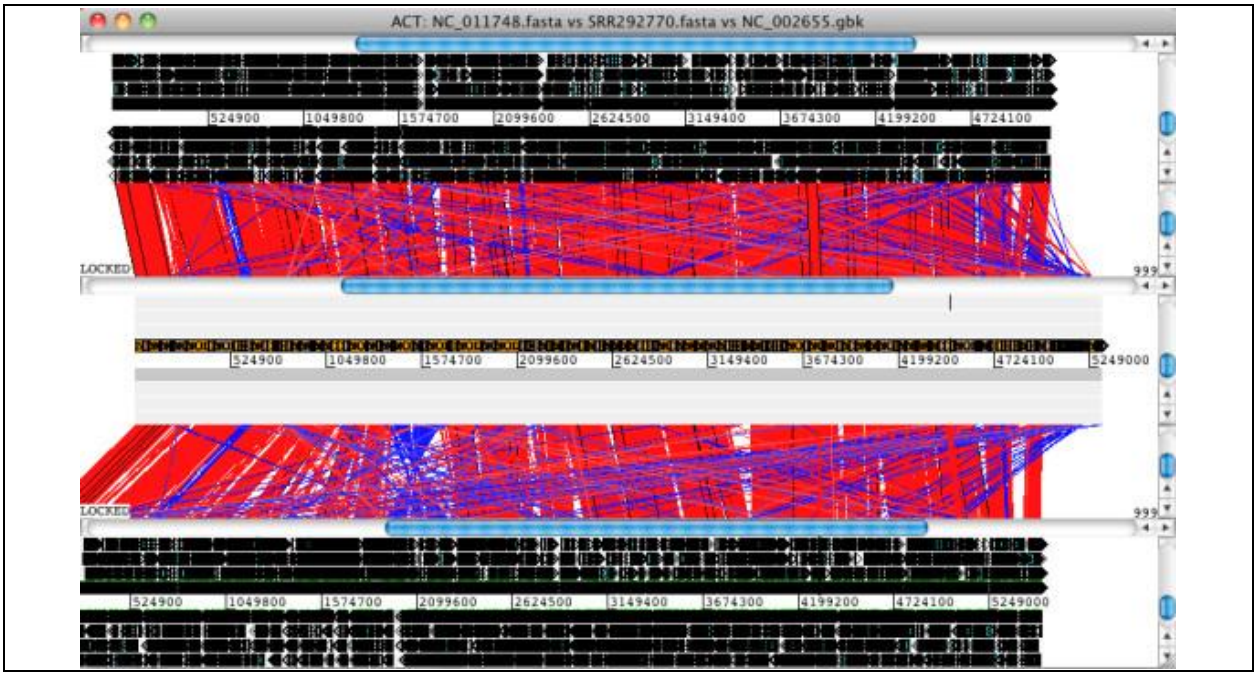
Όπως είδαμε, τα περισσότερα από τα προαναφερθέντα πακέτα λογισμικού παρέχουν τη δυνατότητα χρήσης διαφορετικών αλγορίθμων στοίχισης γονιδιωμάτων. Κάποια, διαθέτουν και δικούς τους αλγόριθμους στοίχισης αλλά τα περισσότερα δίνουν τα δυνατότητα ενσωμάτωσης και άλλων εξειδικευμένων αλγορίθμων. Οι πιο γνωστοί από αυτούς είναι το **MUMMER** (http://mummer.sourceforge.net/), το **MEGA-BLAST** (http://www.ncbi.nlm.nih.gov/BLAST/), το **LAGAN** (http://bioperl.org/wiki/LAGAN) και το **MGA** (http://bibiserv.techfak.uni-bielefeld.de/mga/). Όσον αφορά τους αλγόριθμους εύρεσης σύντηξης γονιδίων, η οποία σαν μέθοδος είναι και η πιο «απόμακρη» (ή ξεχωριστή) από τις υπόλοιπες, υπάρχουν επίσης διαθέσιμες μια σειρά από επιλογές, οι οποίες έχουν πολλαπλασιαστεί ιδιαίτερα τα τελευταία χρόνια με την έλευση της αλληλούχισης νέας γενιάς με τη χρήση τέτοιων τεχνικών σε διάφορες άλλες εφαρμογές, ακόμα και ιατρικές (Carrara et al., 2013). Ενδεικτικά, αναφέρουμε τον αρχικό αλγόριθμο των Ouzounis και συνεργατών, το **GeneRAGE** (Enright & Ouzounis, 2000), αλλά και μερικές νεότερες εφαρμογές όπως το **FusionMap** (http://www.omicsoft.com/fusionmap) (Ge et al., 2011) και το **MosaicFinder** (http://sourceforge.net/projects/mosaicfinder) (Jachiet, Pogorelcnik, Berry, Lopez, & Bapteste, 2013).



Εικόνα 11.15: *Κυκλική αναπαράσταση της στοίχισης του γονιδιώματος της* E. coli *O157:H7 str. Sakai και η σύγκριση με 27 άλλα προκαρυωτικά γονιδιώματα με του BRIG.*



Εικόνα 11.16: ***(a)*** *Διάγραμμα μιας χρωμοσωμικής περιοχής του ανθρώπινου γονιδιώματος που περιέχει το γονίδιο KIF3A (chr5:131949456–132139102) με το VISTA Η σύγκριση δείχνει συντηρημένες περιοχές μεταξύ ανθρώπου και ποντικού και μεταξύ ανθρώπου και αρουραίου.* ***(b)*** *Το VISTA παράγει μια λίστα με τα συντηρημένα στοιχεία μεταξύ ανθρώπου και ποντικού στην περιοχή του KIF3A* ***(c)*** *Η γονιδιωματική περιοχή πριν από το γονίδιο KIF3A, στην οποία εμφανίζονται συντηρημένες μη κωδικές περιοχές (Frazer, et al., 2004).*



Εικόνα 11.17: *Στοίχιση γονιδιωμάτων με το ACT, Το γονιδίωμα της E. coli O104:H4 είναι στη μεσαία σειρά, αυτό της E. coli Ec55989 φαίνεται πάνω, ενώ το γονιδίωμα της E. coli EDL933 είναι κάτω (Edwards & Holt, 2013).*

Βιβλιογραφία

Alikhan, N. F., Petty, N. K., Ben Zakour, N. L., & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics, 12*, 402. doi: 10.1186/1471-2164-12-402

Arai, M., Ikeda, M., & Shimizu, T. (2003). Comprehensive analysis of transmembrane topologies in prokaryotic genomes. [Research Support, Non-U.S. Gov't]. *Gene, 304*, 77-86.

Blom, J., Albaum, S. P., Doppmeier, D., Pühler, A., Vorhölter, F.-J., Zakrzewski, M., & Goesmann, A. (2009). EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics, 10*(1), 154.

Carrara, M., Beccuti, M., Lazzarato, F., Cavallo, F., Cordero, F., Donatelli, S., & Calogero, R. A. (2013). State-of-the-art fusion-finder algorithms sensitivity and specificity. *Biomed Res Int, 2013*, 340620. doi: 10.1155/2013/340620

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G., & Parkhill, J. (2005). ACT: the Artemis Comparison Tool. *Bioinformatics, 21*(16), 3422-3423. doi: 10.1093/bioinformatics/bti553

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One, 5*(6), e11147. doi: 10.1371/journal.pone.0011147

Darling, A. E., Miklos, I., & Ragan, M. A. (2008). Dynamics of genome rearrangement in bacterial populations. *PLoS Genet, 4*(7), e1000128. doi: 10.1371/journal.pgen.1000128

Edwards, D. J., & Holt, K. E. (2013). Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microb Inform Exp, 3*(1), 2. doi: 10.1186/2042-5783-3-2

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature, 402*(6757), 86-90. doi: 10.1038/47056

Enright, A. J., & Ouzounis, C. A. (2000). GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics, 16*(5), 451-457.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., & Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res, 32*(Web Server issue), W273-279. doi: 10.1093/nar/gkh458

Ge, H., Liu, K., Juan, T., Fang, F., Newman, M., & Hoeck, W. (2011). FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics, 27*(14), 1922-1928. doi: 10.1093/bioinformatics/btr310

Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R. Y., Algire, M. A., . . . Venter, J. C. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science, 329*(5987), 52-56. doi: 10.1126/science.1190719

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., . . . Young, R. A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature, 431*(7004), 99-104. doi: 10.1038/nature02800

Jachiet, P. A., Pogorelcnik, R., Berry, A., Lopez, P., & Bapteste, E. (2013). MosaicFinder: identification of fused gene families in sequence similarity networks. *Bioinformatics, 29*(7), 837-844. doi: 10.1093/bioinformatics/btt049

Kreil, D. P., & Ouzounis, C. A. (2001). Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res, 29*(7), 1608-1615.

Li, W. (2011). On parameters of the human genome. [Review]. *J Theor Biol, 288*, 92-104. doi: 10.1016/j.jtbi.2011.07.021

Ouzounis, C. A., Kunin, V., Darzentas, N., & Goldovsky, L. (2006). A minimal estimate for the gene content of the last universal common ancestor--exobiology from a terrestrial perspective. *Res Microbiol, 157*(1), 57-68. doi: 10.1016/j.resmic.2005.06.015

Papadimitriou, K., Anastasiou, R., Mavrogonatou, E., Blom, J., Papandreou, N. C., Hamodrakas, S. J., . . . Pot, B. (2014). Comparative genomics of the dairy isolate Streptococcus macedonicus ACA-DC 198 against related members of the Streptococcus bovis/Streptococcus equinus complex. *BMC Genomics, 15*(1), 272.

Picardi, E., & Pesole, G. (2010). Computational methods for ab initio and comparative gene finding. *Methods Mol Biol, 609*, 269-284. doi: 10.1007/978-1-60327-241-4_16

Poliakov, A., Foong, J., Brudno, M., & Dubchak, I. (2014). GenomeVISTA--an integrated software package for whole-genome alignment and visualization. *Bioinformatics, 30*(18), 2654-2655. doi: 10.1093/bioinformatics/btu355

Quax, T. E., Claassens, N. J., Soll, D., & van der Oost, J. (2015). Codon Bias as a Means to Fine-Tune Gene Expression. *Mol Cell, 59*(2), 149-161. doi: 10.1016/j.molcel.2015.05.035

Rigoutsos, I. (2010). Short RNAs: how big is this iceberg? *Curr Biol, 20*(3), R110-113. doi: 10.1016/j.cub.2009.12.036

Shimizu, T., Mitsuke, H., Noto, K., & Arai, M. (2004). Internal gene duplication in the evolution of prokaryotic transmembrane proteins. [Comparative Study

Research Support, Non-U.S. Gov't]. *J Mol Biol, 339*(1), 1-15. doi: 10.1016/j.jmb.2004.03.048

Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D., & Krogh, A. (2001). On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet, 17*(8), 425-428. doi: S0168-9525(01)02372-1 [pii]

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nat Rev Genet, 16*(8), 472-482. doi: 10.1038/nrg3962

Tsoka, S., & Ouzounis, C. A. (2000). Recent developments and future directions in computational genomics. *FEBS Lett, 480*(1), 42-48.

Uchiyama, I., Higuchi, T., & Kobayashi, I. (2006). CGAT: a comparative genome analysis tool for visualizing alignments in the analysis of complex evolutionary changes between closely related genomes. *BMC Bioinformatics, 7*, 472. doi: 10.1186/1471-2105-7-472

Vlachos, I. S., & Hatzigeorgiou, A. G. (2013). Online resources for miRNA analysis. *Clin Biochem, 46*(10-11), 879-900. doi: 10.1016/j.clinbiochem.2013.03.006

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res, 18*(5), 821-829. doi: 10.1101/gr.074492.107

Κεφάλαιο 12: Η Γλώσσα Προγραμματισμού Perl

Σύνοψη
Η γλώσσα προγραμματισμού Perl είναι μια ιδιαίτερα εύχρηστη γλώσσα με πολλές εφαρμογές στη βιοπληροφορική. Η Perl είναι γλώσσα scripting και κατασκευάστηκε αρχικά για να διευκολύνει τους διαχειριστές συστημάτων UNIX στην καθημερινή τους δουλειά, δηλαδή στην επεξεργασία μεγάλων αρχείων, σε μαζικές αποστολές μηνυμάτων κ.ο.κ., γι' αυτό και βρήκε πολλές εφαρμογές στη βιοπληροφορική, όπου ο χειρισμός μεγάλων αρχείων και η αναζήτηση προτύπων, είναι μέρος της καθημερινής σχεδόν ενασχόλησης. Στο κεφάλαιο αυτό θα παρουσιάσουμε τα βασικά στοιχεία της γλώσσας και θα δείξουμε τη χρησιμότητά της σε πλήθος πρακτικών προβλημάτων που προκύπτουν στην ανάλυση βιολογικών (κυρίως μοριακών) δεδομένων.

Προαπαιτούμενη γνώση
Το κεφάλαιο αυτό δεν έχει ιδιαίτερες απαιτήσεις και προαπαιτούμενα πέρα από τις βασικές γνώσεις της βιολογίας και μια εξοικείωση με τους Η/Υ.

12. Εισαγωγή

Η Perl είναι μια γλώσσα προγραμματισμού ιδιαίτερα γνωστή στο χώρο του UNIX και του Linux. Το όνομα προέρχεται από τα αρχικά των όρων «*Practical Extraction and Report Language*» το οποίο σε ελεύθερη μετάφραση σημαίνει «*Γλώσσα για την Πρακτική Αναφορά και Εξαγωγή (δεδομένων)*». Ορισμένοι βέβαια, χιουμοριστικά, υποστηρίζουν ότι το όνομα Perl προέρχεται από τα αρχικά των λέξεων «*Pathologiacally Eclectic Rubbish Lister*» δηλαδή «*Παθολογικά Επιλεκτική Παρουσίαση Σκουπιδιών*». Και οι δύο εκδοχές όμως φαίνεται ότι γίνονται αποδεκτές από τον δημιουργό της γλώσσας, τον Larry Wall. Η Perl μπορεί να βοηθήσει τον προγραμματιστή στη δημιουργία και εκτέλεση προγραμμάτων τα οποία σε άλλες γλώσσες προγραμματισμού θα ήταν πιο πολύπλοκα και θα χρειαζόταν σημαντικά περισσότερος χρόνος για να υλοποιηθούν. Η Perl θεωρείται παραδοσιακά γλώσσα scripting. Scripts ονομάζονται γενικά, τα μικρά προγραμματάκια που υλοποιούνται για πολύ συγκεκριμένες δουλειές και μπορεί να είναι συχνά μίας χρήσης. Το χαρακτηριστικό τους είναι ότι διερμηνεύονται (interpret) και δεν μεταγλωττίζονται (compile), δηλαδή δεν μετατρέπονται σε γλώσσα μηχανής. Η γλώσσα κατασκευάστηκε αρχικά για να διευκολύνει τους διαχειριστές συστημάτων UNIX στην καθημερινή τους δουλειά που απαιτούσε «scripting», δηλαδή στην επεξεργασία μεγάλων αρχείων, σε μαζικές αποστολές μηνυμάτων κ.ο.κ. Γι' αυτό το λόγο βρήκε και πολλές εφαρμογές στη βιοπληροφορική, όπου ο χειρισμός μεγάλων αρχείων και η αναζήτηση προτύπων είναι μέρος της καθημερινής σχεδόν ενασχόλησης. Η Perl στηρίχθηκε σε ό,τι καλύτερο υπήρχε εκείνη την εποχή διαθέσιμο για τέτοιου είδους δουλειές και το ενσωμάτωσε. Έτσι, στην Perl θα βρούμε στοιχεία του awk, του sed, του ίδιου του bash shell, αλλά και της γλώσσας C. Προφανώς, όποιος γνωρίζει κάποιο από τα παραπάνω εργαλεία, θα βρει την Perl αρκετά εύκολη και μάλλον ενδιαφέρουσα.

Ένα βασικό χαρακτηριστικό της Perl, είναι η απλότητα της σε πολλά επίπεδα που θα εξηγηθούν παρακάτω, η οποία οδηγεί σε μια μάλλον «απότομη» καμπύλη μάθησης. Αυτό σημαίνει, ότι κάποιος που δεν γνωρίζει τη γλώσσα, μπορεί σε σύντομο χρονικό διάστημα να τη χρησιμοποιήσει για να πραγματοποιήσει κάποιες βασικές λειτουργίες (στον ίδιο χρόνο δεν θα μπορούσε να πετύχει το ίδιο σε κάποιες άλλες γλώσσες όπως η C, C++ ή Java). Το μειονέκτημα σε αυτό, είναι ότι με το "χαλαρό" τρόπο της, η Perl μπορεί να οδηγήσει κάποιες φορές τον προγραμματιστή σε λάθος. Το άλλο χαρακτηριστικό της, αποτυπώνεται στο σύνθημα «*TIMTOWTDI*», που προέρχεται από τα αρχικά της φράσης «*There Is More Than One Way To Do It*» (και προφέρεται «*Tim Toady*»). Η γλώσσα σχεδιάστηκε με αυτό σαν στόχο, υπηρετώντας δηλαδή τη βασική ιδέα ότι δεν πρέπει η γλώσσα να επιβάλλει στον προγραμματιστή το πως θα γράψει ένα πρόγραμμα. Αυτό σημαίνει ότι υπάρχουν πολλοί (και μάλλον αρκετά διαφορετικοί) τρόποι να γραφτεί ένα συγκεκριμένο πρόγραμμα, κάνοντας κάθε φορά χρήση συναρτήσεων διαφορετικής προέλευσης και λειτουργίας. Το μειονέκτημα αυτού βέβαια, είναι ότι πολλές φορές τα προγράμματα που έχει φτιάξει κάποιος, δεν είναι εύκολο να διαβαστούν και να γίνουν κατανοητά από κάποιον άλλον, ειδικά αν ο προγραμματιστής δεν επιλέξει να βάλει τα κατάλληλα σχόλια στον κώδικα. Πολλές φορές οι χρήστες της Perl κάνουν και άτυπους (ή και λιγότερο άτυπους), διαγωνισμούς για το ποιος μπορεί να γράψει το πιο μικρό ή το πιο γρήγορο πρόγραμμα για να κάνει μια συγκεκριμένη δουλειά (τα γνωστά oneliners), προγράμματα τα οποία πολλές

φορές απαιτούν...μελέτη για να καταλάβει κανείς τί ακριβώς κάνουν. Στον αντίποδα όλων αυτών, υπάρχει η γνωστή γλώσσα Python, ο μεγαλύτερος αντίπαλος της Perl, το σύνθημα της οποίας αποτυπώνεται στη φράση «*There should be one-and preferably only one-obvious way to do it*» (αλλά καλύτερα, ας αφήσουμε τις συγκρίσεις μεταξύ των γλωσσών, θα οδηγηθούμε σε άσχημα μονοπάτια!).

Τα λειτουργικά συστήματα Unix/Linux και MacOS X έχουν εγκατεστημένη την Perl από «το κουτί», ενώ είναι διαθέσιμη δωρεάν για χρήση και για συστήματα Windows. Αν χρησιμοποιούμε ένα τέτοιο λειτουργικό σύστημα, για να εγκαταστήσουμε την Perl στο σύστημά μας θα πρέπει να επισκεφτούμε τον ιστότοπο http://www.perl.org/get.html και να επιλέξουμε την έκδοση που επιθυμούμε να εγκαταστήσουμε, ενώ στη συνέχεια πρέπει να ακολουθήσουμε τα βήματα που υποδεικνύει το πρόγραμμα εγκατάστασης. Επειδή η Perl δουλεύει με τον παλιό καλό τρόπο, δηλαδή από τη γραμμή εντολών, ιδιαίτερα στα Windows πολλοί επιλέγουν να τη χρησιμοποιούν μέσα από κάποιο είδος κέλυφους (shell) που να θυμίζει και να δίνει τις αντίστοιχες ευκολίες με το περιβάλλον του UNIX. Μια πολύ καλή τέτοια λύση είναι ο εξομοιωτής **Cygwin** (www.cygwin.com), αλλά και διάφορες «native» (φυσικές) όπως λέγονται μεταφορές του περιβάλλοντος του bash στα Windows, όπως τα **UnixUtils** (http://unxutils.sourceforge.net/) και το **MinGW** (http://www.mingw.org/). Μαζί με τη γλώσσα και τον μεταγλωττιστή (compiler), εγκαθίσταται και το σύστημα τεκμηρίωσης της Perl, παραδείγματα και άλλα βοηθήματα.

12.1. Τα βασικά της Perl

Για την δημιουργία ενός αρχείου που περιέχει κώδικα Perl αρκεί να χρησιμοποιήσουμε έναν απλό κειμενογράφο. Συνιστάται βέβαια, να χρησιμοποιούνται κειμενογράφοι που υποστηρίζουν την γλώσσα αυτή ώστε να χρωματίζουν κατάλληλα τις δεσμευμένες λέξεις, τις μεταβλητές και τα διάφορα κομμάτια (blocks) του κώδικα. Αφού συντάξουμε τον κώδικά μας πρέπει να αποθηκεύσουμε το αρχείο, και συνίσταται να το αποθηκεύουμε με κατάληξη .pl ώστε να μπορεί να επεξεργασθεί από τον compiler στην περίπτωση των Windows (αλλά και να αναγνωρίζεται από τους κειμενογράφους για να έχουμε το επιθυμητό αποτέλεσμα στην οπτικοποίηση του κώδικα). Από την στιγμή που έχουμε το αρχείο, μπορούμε να το εκτελέσουμε μέσα από τη γραμμή εντολών, καλώντας την Perl με την εντολή:

```
perl file_name.pl
```

Στην περίπτωση που το αρχείο που θα εκτελέσουμε απαιτεί την εισαγωγή δεδομένων από τον χρήστη για να εκτελεστεί, τα δεδομένα αυτά θα πρέπει να εισαχθούν στην παραπάνω γραμμή μετά την ονομασία του αρχείου χωρισμένα με κενό (αυτά είναι τα ορίσματα σε ένα πρόγραμμα). Εφόσον υπάρχουν λάθη σύνταξης ο μεταγλωττιστής θα μας ειδοποιήσει στη συνέχεια με μήνυμα στην ίδια γραμμή εντολών. Εάν δεν υπάρχουν λάθη και ο κώδικάς μας είναι σωστός θα εκτελεστεί το πρόγραμμα και στην περίπτωση που δεν υπάρχουν ούτε λογικά λάθη θα δούμε το αποτέλεσμά του είτε και πάλι στην ίδια γραμμή εντολών, είτε σε κάποιο άλλο αρχείο που έχουμε ορίσει να δημιουργεί ή να τροποποιεί το πρόγραμμα. Για την εισαγωγή σχολίων στον κώδικα της Perl χρησιμοποιούμε το σύμβολο #. Παρακάτω παρουσιάζεται το πρώτο πρόγραμμα Perl:

```
#!/usr/bin/perl
#Πρώτο Πρόγραμμα
print "Hello world \n";
```

Η συνάρτηση print έχει ως όρισμα μια συμβολοσειρά και το χαρακτήρα \n ο οποίος αντιπροσωπεύει την αλλαγή γραμμής. Η εντολή print, όπως και κάθε εντολή στη γλώσσα Perl αλλά και σε άλλες γλώσσες προγραμματισμού, ολοκληρώνεται με το ελληνικό ερωτηματικό (;). Ειδικά η πρώτη γραμμή αυτού του προγράμματος, είναι προαιρετική και χρησιμεύει στο να δείξει τη θέση του μεταγλωττιστή της Perl. Με αυτόν τον τρόπο, σε συστήματα Linux/UNIX αλλά και σε εξομοιωτές, το παραπάνω πρόγραμμα μπορεί να εκτελεστεί και σαν εκτελέσιμο shell script, δηλαδή απλά με την εντολή:

```
./program.pl
```

Το ./ στην αρχή, το χρησιμοποιούμε όταν το πρόγραμμα βρίσκεται στον τρέχοντα κατάλογο (directory) του δίσκου. Φυσικά, αν το πρόγραμμα μας θέλουμε να χρησιμοποιείται γενικά, μπορούμε να το τοποθετήσουμε

σε κάποιον κατάλογο που βρίσκεται στο μονοπάτι του συστήματος (Path), και έτσι να είναι προσβάσιμο από παντού.

Κάτι άλλο που πρέπει να έχουμε υπόψη μας, είναι ότι η Perl δεν κατασκευάζει κώδικα σε γλώσσα μηχανής (όπως για παράδειγμα η C). Αυτό σημαίνει ότι τα προγράμματα, με ελάχιστες μόνο εξαιρέσεις που αφορούν μονοπάτια αρχείων στο δίσκο κλπ, θα μπορούν να χρησιμοποιηθούν αυτούσια σε όλα τα συστήματα. Το μειονέκτημα σε αυτό, είναι ότι τα προγράμματα είναι σχετικά πιο αργά στην εκτέλεση σε σχέση με τα αντίστοιχα που θα μπορούσαν να είχαν γραφτεί για παράδειγμα στη C (πάλι όμως, υπάρχουν εξαιρέσεις, ειδικά γιατί κάποιες ρουτίνες που είναι εξειδικευμένες στην επεξεργασία αρχείων και κειμένου είναι ιδιαίτερα βελτιστοποιημένες στην Perl). Εν τούτοις, η Perl δεν είναι κλασική γλώσσα με διερμηνέα (interpreter), καθώς το κάθε πρόγραμμα δεν εκτελείται γραμμή-γραμμή, αλλά διαβάζεται συνολικά, ελέγχεται για λάθη και μετασχηματίζεται σε κάποια εσωτερική ενδιάμεση μορφή πριν την εκτέλεση. Γι' αυτούς τους λόγους πολλές φορές θα δούμε ότι αναφέρεται η ύπαρξη του «μεταγλωττιστή» της Perl. Η Python λειτουργεί επίσης με ένα παρόμοιο σύστημα, ενώ στην περίπτωση της Java, η ενδιάμεση μορφή αποθηκεύεται ως bytecode το οποίο μπορεί να χρησιμοποιηθεί μετά σε κάθε σύστημα.

12.2. Μεταβλητές

Στην Perl, τα ονόματα των βαθμωτών μεταβλητών (scalars) ξεκινούν πάντα με το σύμβολο του δολαρίου ($) ακολουθούμενο από οποιοδήποτε όνομα. Κάποια ειδικά ονόματα, είναι δεσμευμένα για ειδικές λειτουργίες της γλώσσας που θα συναντήσουμε παρακάτω ($1, $2, $_, $/ κ.ο.κ.). Μια μεγάλη διαφορά από άλλες γλώσσες προγραμματισμού, βρίσκεται στο γεγονός ότι οι μεταβλητές μπορεί να περιέχουν αριθμούς, μαθηματικές εκφράσεις, σμβολοσειρές ή ακόμα και αναφορά σε άλλες μεταβλητές, ενώ το πιο σημαντικό από όλα είναι ότι δεν χρειάζεται να ορίσουμε από την αρχή τί είδους μεταβλητή θα κατασκευάσουμε. Αυτό αποτελεί μια μεγάλη ευκολία, αλλά χρειάζεται ιδιαίτερη προσοχή καθώς το πώς θα χειριστούμε τη μεταβλητή αυτή, εξαρτάται από μας γιατί μπορούμε σε μια αριθμητική μεταβλητή να εφαρμόσουμε πράξεις για αριθμούς αλλά και για συμβολοσειρές. Για την δημιουργία οποιουδήποτε τύπου μεταβλητής αρκεί να πληκτρολογήσουμε:

```
$var = ___;
```

Με τον παραπάνω τρόπο δημιουργούμε μια μεταβλητή με όνομα $var και τύπο που εξαρτάται από το τι θα βρίσκεται δεξιά του ίσον (κάνουμε δηλαδή, ανάθεση). Δεξιά από την ανάθεση, μπορεί να βρίσκεται μια βαθμωτή τιμή (scalar) ή μια έκφραση. Όταν θέλουμε σε μια μεταβλητή να αναθέσουμε μια συμβολοσειρά μπορούμε να χρησιμοποιήσουμε τα διπλά λατινικά εισαγωγικά ("") για να εισάγουμε την μεταβλητή. Επίσης με τα διπλά εισαγωγικά χρησιμοποιούμε την παρεμβολή μεταβλητών (variable interpolation) και έτσι μπορούμε να περάσουμε το περιεχόμενο άλλων μεταβλητών σε μια συμβολοσειρά, ή να χρησιμοποιήσουμε χαρακτήρες διαφυγής (π.χ. \n). Οι χαρακτήρες αυτοί και ο συμβολισμός με την ανάποδη κάθετο (\) χρησιμοποιούνται για να ορίσουμε ιδιαίτερες λειτουργίες (το \n σημαίνει «αλλαγή γραμμής», το \s σημαίνει «κενό», το \t «στηλοθέτης» κ.ο.κ.). Έτσι, με τους χαρακτήρες διαφυγής μπορούμε να εισάγουμε τέτοιες ειδικές σημασίες των χαρακτήρων, αλλά και να αποφύγουμε την ερμηνεία ενός χαρακτήρα όταν δεν το θέλουμε. Για παράδειγμα, αν θέλουμε να περάσουμε τον χαρακτήρα $ χωρίς η Perl να νομίζει ότι ορίζουμε μεταβλητή, θα πρέπει να χρησιμοποιήσουμε το \$ (και όμοια γίνεται και για άλλους τέτοιους χαρακτήρες).

Αν χρησιμοποιήσουμε τα μονά λατινικά εισαγωγικά ('') οι χαρακτήρες αναγνωρίζονται σαν απλό κείμενο όπως ακριβώς το γράφουμε. Προσοχή χρειάζονται τα ανάποδα εισαγωγικά (``), με τα οποία η Perl ανοίγει εσωτερικά το shell, εκτελείται ο κώδικας που βρίσκεται μέσα στα εισαγωγικά σαν εντολή του συστήματος, και επιστρέφει στο πρόγραμμα το αποτέλεσμα.

Εάν μέσα στο πρόγραμμά μας υπάρχει ο παρακάτω κώδικας:

```
…
$var=3;
....
$var="hello";
….
```

η μεταβλητή $var αλλάζει τύπο και περιεχόμενο. Στο παραπάνω παράδειγμα η μεταβλητή $var περιείχε αρχικά έναν ακέραιο αριθμό και έπειτα μια συμβολοσειρά. Όπως είπαμε, αντίθετα με άλλες γλώσσες η Perl δεν απαιτεί από το χρήστη να δηλώσει τον τύπο των μεταβλητών πριν τις χρησιμοποιήσει. Στη συνέχεια δίνονται κάποια παραδείγματα ορισμού ή τροποποίησης μεταβλητών στην Perl.

```
$number=5;
$name="George";
$exp=3*$number+($number+1);
$a+=5;
$b*=3;
++$a; ή $a++;
$a--;
```

12.2.1. Τελεστές

Τελεστές (Operators) όπως και στις υπόλοιπες γλώσσες, ορίζουμε τα σύμβολα ή τις δεσμευμένες εκφράσεις μια γλώσσας που πραγματοποιούν πράξεις ή συγκρίσεις μεταξύ μεταβλητών. Οι βασικοί τελεστές που χρησιμοποιούνται για αριθμητικές πράξεις, φαίνονται στον Πίνακα 12.1 ενώ οι τελεστές που χρησιμοποιούνται για τις πράξεις και τις συγκρίσεις συμβολοσειρών (string), δίνονται στον Πίνακα 12.2.

Αριθμητικοί Τελεστές	**Περιγραφή**
+	Πρόσθεση
-	Αφαίρεση
*	Πολλαπλασιασμός
/	Διαίρεση
%	Υπόλοιπο Διαίρεσης
**	Δύναμη

Πίνακας 12.1: *Αριθμητικοί τελεστές της Perl*

Αριθμητικοί Τελεστές	**Αλφαριθμητικοί Τελεστές**	**Σύγκριση**
+	.	*πρόσθεση*
*	x	*πολλαπλασιασμός*
==	eq	*Ίσο*
!=	ne	*Άνισο*
>	gt	*Μεγαλύτερο από*
<	lt	*Μικρότερο από*
>=	ge	*Μεγαλύτερο ή ίσο*
<=	le	*Μικρότερο ή ίσο*

Πίνακας 12.2: *Αριθμητικοί τελεστές πράξεων και σύγκρισης και οι αντίστοιχοι τελεστές συμβολοσειρών στην Perl. Ο πολλαπλασιασμός στην περίπτωση των συμβολοσειρών αφορά τον πολλαπλασιασμό αριθμού με συμβολοσειρά.*

12.2.2. Συναρτήσεις της Perl

Κάποιες συνήθεις συναρτήσεις που χρησιμοποιούνται για την διαχείριση μεταβλητών που περιέχουν συμβολοσειρές δίνονται στον Πίνακα 12.3:

chomp	Κόβει το τελικό \n από την μεταβλητή	
chop	Κόβει τον τελευταίο χαρακτήρα από τη μεταβλητή	
substr	$x = substr($name, 0, 1, "L");	Αντικαθιστά στην ακολουθία $name 1 χαρακτήρα ξεκινώντας από τη θέση την 0 και αποθηκεύει την νέα στην μεταβλητή $x Η μεταβλητή $name δεν αλλάζει.
	$x = substr($name, 0, 1);	Αντιγράφει ακολουθία μήκους 1 με αρχική θέση την 0 από την ακολουθία $name στην μεταβλητή $x .
index	Index($name, "k");	Δίνει την θέση του k στην ακολουθία $name
rindex	Επιστρέφει ό,τι και η index μετρώντας όμως από το τέλος της ακολουθίας.	

Πίνακας 12.3: *Συναρτήσεις διαχείρισης συμβολοσειρών στην Perl*

Για παράδειγμα, οι εντολές:

```
$name="Takis";
$x=substr($name,0,1);
```

Θα ορίσουν αρχικά τη μεταβλητή $name και θα της δώσουν περιεχόμενο, ενώ στη συνέχεια θα δημιουργήσουν μια βαθμωτή μεταβλητή που περιέχει τον πρώτο χαρακτήρα του $name (δηλαδή «Τ»).

12.2.3. Κανονική είσοδος (<STDIN>)

Χρησιμοποιώντας τη λειτουργία <STDIN> σε ένα πρόγραμμα, η Perl διαβάζει τη γραμμή από την κανονική είσοδο (δηλαδή, από το πληκτρολόγιο) μέχρι την πρώτη αλλαγή γραμμής και τη χρησιμοποιεί ως τιμή της ειδικής μεταβλητής <STDIN>. Μια συνηθισμένη διαδικασία εισόδου είναι η παρακάτω:

```
$a=<STDIN>; #Αποθήκευση του κειμένου που πληκτρολογεί ο χρήστης στη
μεταβλητή $a
print $a;
```

Το ίδιο αποτέλεσμα θα μπορούσε να προκύψει και με μία μόνο εντολή:

```
chomp ($a=<STDIN>);
```

12.3. Πίνακες και λίστες

Λίστα (list) είναι μια ταξινομημένη σειρά από βαθμωτές τιμές, μεταβλητές ή εκφράσεις. Είναι στην ουσία ένα διάνυσμα. Κάποια παραδείγματα από λίστες, φαίνονται παρακάτω:

```
(1, 2, 3)
("perl", 3, 15)
($x, 3, $x+2, "$y$x")
(1..10)
($a..$b)
```

Η λίστα κατασκευάζεται δυναμικά και χρησιμοποιείται από διάφορες εντολές και συναρτήσεις. Με την χρήση λίστας για παράδειγμα, μπορούμε να κάνουμε πολύ εύκολα αλλαγή των τιμών μεταξύ δύο μεταβλητών (swap):

```
($a,$b)=($b,$a);
```

ή να αναθέσουμε μαζικά τιμές σε βαθμωτές μεταβλητές

```
($a, $b, $c)=(1,2,3);
```

Όταν όμως θέλουμε να αποθηκεύσουμε μια λίστα και να την προσπελάσουμε χρειαζόμαστε μια νέα δομή. Αυτή είναι ο πίνακας (Array). Ένας πίνακας είναι στην ουσία μια μεταβλητή που περιέχει τα περιεχόμενα μιας λίστας. Τα στοιχεία του πίνακα έχουν μια θέση το ένα μετά το άλλο, ανάλογα με την θέση που έχουν δηλωθεί κατά την δημιουργία του πίνακα. Η πιο απλή σύνταξη για την δημιουργία ενός πίνακα είναι η ανάθεση τιμών από μια λίστα:

```
@array = (     );
```

Το όνομα κάθε πίνακα, πρέπει να περιέχει υποχρεωτικά το @ σαν πρώτο χαρακτήρα. Προσοχή χρειάζεται επίσης, στο γεγονός ότι μπορεί να υπάρχει πίνακας @name και ταυτόχρονα βαθμωτή μεταβλητή $name. Επίσης, όμοια με τις βαθμωτές μεταβλητές, κάποια ονόματα είναι δεσμευμένα (@_, @ARGV κ.ο.κ.). Ανάλογα με το τι περιέχεται ανάμεσα στις παρενθέσεις ο πίνακας έχει και τον αντίστοιχο τύπο δηλαδή ο πίνακας μπορεί να περιέχει αριθμούς, συμβολοσειρές ή συνδυασμό τους:

```
@name=("John", "George", "Mike");
@name=(1..10);
@table=1;
@table=(1, 2, @name, 7);
@table=(1,2,3);
```

Η αρίθμηση των στοιχείων, όπως και στις περισσότερες άλλες γλώσσες, ξεκινάει από το 0, αλλά σε αντίθεση με τις άλλες γλώσσες το πλήθος των στοιχείων δεν είναι απαραίτητο να δηλωθεί όταν ορίζεται ο πίνακας. Μπορούμε να ορίσουμε έναν πίνακα και με τον παρακάτω τρόπο χρησιμοποιώντας τον τρόπο δήλωσης των στοιχείων του με τη σειρά:

```
$table [0] = 1;
$table [1] = 2;
$table [2] = 3;
```

Τα στοιχεία του πίνακα, είναι κανονικές βαθμωτές μεταβλητές τόσο στην ονομασία (έχουν πάντα το $ στην αρχή του ονόματος) όσο και στη λειτουργία τους. Αυτό σημαίνει ότι μπορούμε να τα χρησιμοποιήσουμε σε κάθε είδους διεργασία στην οποία θα χρησιμοποιούσαμε μια βαθμωτή μεταβλητή. Μπορούμε δηλαδή να κάνουμε πράξεις με αυτά, να πραγματοποιήσουμε ελέγχους (βλ. παρακάτω) ή να τα βάλουμε σε μια λίστα και να χρησιμοποιήσουμε τη λίστα. Για παράδειγμα θα μπορούσαμε να έχουμε τις παρακάτω περιπτώσεις:

```
$x=$table[0];
$table[1]++;
($table[0], $table[1])= ($table[1], $table[0]);
@table[0,1,2]=@table[1,1,1];
```

Αυτό πρακτικά σημαίνει, ότι μπορούμε και να ορίσουμε έναν πίνακα δίνοντας ως στοιχείο του έναν άλλον πίνακα (με αυτόν τον τρόπο η αρίθμηση των στοιχείων θα μεταφέρεται αυτόματα καθώς τα νέα στοιχεία παρεμβάλλονται στα παλιά):

```
@table=(1, 2, @name, 7)
```

Κάποιες συνήθεις συναρτήσεις που χρησιμοποιούνται για την διαχείριση πινάκων δίνονται στον Πίνακα 12.4. Για παράδειγμα, η push συντάσσεται με τον εξής τρόπο:

```
push @table,$scalar;
```

το οποίο είναι ισοδύναμο, με το να ορίσεις τον πίνακα ως εξής:

```
@table=(@table, $scalar);
```

Αντίστοιχα, η unshift συντάσσεται έτσι:

```
unshift(@table, $scalar);
```

το οποίο είναι ισοδύναμο με την εξής εντολή:

```
@table=($scalar, @table );
```

push	Εισαγωγή στοιχείου στο τέλος του πίνακα.
pop	Διαγραφή στοιχείου από το τέλος του πίνακα.
shift	Προσθέτει στοιχείο στην πρώτη θέση του πίνακα.
unshift	Διαγραφή του πρώτου στοιχείου του πίνακα
reverse	Αντιστρέφει τη σειρά των στοιχείων του πίνακα.
sort	Ταξινομεί τον πίνακα με αύξουσα σειρά. Εάν ο πίνακας είναι αλφαριθμητικός τον ταξινομεί κατά ASCII.
splice	splice(@table,2,1); Από τον πίνακα @table αφαιρεί 1 στοιχείο ξεκινώντας από την θέση 2.

Πίνακας 12.4: *Συναρτήσεις διαχείρισης πινάκων στην Perl*

Τέλος, μια πολύ χρήσιμη λειτουργία των πινάκων, είναι η μεταβλητή $#. Αν υπάρχει ένας πίνακας @name, τότε η βαθμωτή μεταβλητή $#name περιέχει αυτόματα (αποτελεί δηλαδή μια δεσμευμένη ονομασία και η γλώσσα την κατασκευάζει αυτόματα) την τιμή του μεγαλύτερου δείκτη του πίνακα αυτού. Έτσι, αν ορίσουμε π.χ. τον πίνακα @table=(1,2,3), τότε το $#table, θα έχει την τιμή 2. Το $#table είναι μια κανονική βαθμωτή μεταβλητή, άρα μπορούμε να κάνουμε χρήση του, π.χ. να ζητήσουμε το στοιχείο $table[$#table] το οποίο στη συγκεκριμένη περίπτωση θα είναι ίσο με 3. Προσοχή χρειάζεται στο εξής: λόγω του τρόπου ορισμού των πινάκων, αν ο πίνακας έχει οριστεί με «άναρχο» τρόπο, τότε είναι δυνατό να περιέχει κενά στοιχεία. Αν για παράδειγμα στον παραπάνω πίνακα δώσουμε την εντολή $table[5]=12, τότε «παρακάμπτουμε» τα στοιχεία με δείκτες (index) 3 και 4, και τοποθετούμε το στοιχείο 12 στη θέση με δείκτη 5. Αν τώρα ζητήσουμε το $#table θα πάρουμε την τιμή 5, και μπορεί λανθασμένα να θεωρήσουμε ότι ο πίνακας έχει 6 στοιχεία.

12.4. Ευρετήρια

Ευρετήριο (Hash), ή αλλιώς κατακερματισμός ή συσχετιστικός πίνακας, είναι ένα σύνολο από μεταβλητές οι οποίες επιλέγονται με βάση την τιμή κάποιου δείκτη. Αυτοί οι δείκτες, που ονομάζονται κλειδιά (keys), μπορεί να είναι βαθμωτές τιμές ή συμβολοσειρές, και χρησιμοποιούνται για την ανάκτηση των τιμών του ευρετηρίου (values). Σε ένα ευρετήριο, όμοια με το τι συμβαίνει σε έναν πίνακα, δεν χρειάζεται να ορίσουμε μέγεθος, απλώς αναθέτουμε τα ζεύγη τιμών. Τα ευρετήρια όμως σε αντίθεση με τους πίνακες, δεν έχουν σειρά στην αρίθμηση. Επίσης, τα ονόματα των ευρετηρίων ξεκινάνε πάντα με το σύμβολο %. Ουσιαστικά θα μπορούσαμε να παρομοιάσουμε ένα ευρετήριο με ένα σύνολο από δεδομένα ίδιας μορφής τα οποία πέραν της τιμής τους, έχουν και ένα μοναδικό αναγνωριστικό ώστε να ξεχωρίζονται από τα υπόλοιπα του συνόλου.

Ένα ευρετήριο αποτελείται (και κατά συνέπεια, μπορεί να οριστεί) από μία λίστα με ζεύγη τιμών. Τα ζεύγη αυτά αποτελούνται από το κλειδί και την τιμή. Το κλειδί περιγράφει την τιμή και με αυτό την προσπελαύνουμε και την χρησιμοποιούμε. Κατά την δημιουργία ενός ευρετηρίου πρέπει να δοθούν και τα δύο, και το κλειδί και η τιμή ενός ζευγαριού χωρίς αυτό να σημάνει πως δεν είναι δυνατή η μετέπειτα τροποποίηση τους. Για να ορίσουμε ένα ευρετήριο, ένας εύκολος τρόπος είναι να ακολουθούμε την εξής σύνταξη:

```
%day = ("Sun", "Sunday", "Mon", "Monday", "Tue", "Tuesday", "Wed",
"Wednesday", "Thu", "Thursday", "Fri", "Friday", "Sat", "Saturday");
```

Όπως παρατηρούμε ο τρόπος μοιάζει με τον τρόπο ορισμού ενός πίνακα, παρ' όλα αυτά κάθε ζεύγος (σύμφωνα με την θέση τους κατά τον ορισμό) αποτελεί ένα ζεύγος κλειδιού-τιμής. Ένας άλλος τρόπος για να ορίσουμε ένα ευρετήριο έτσι ώστε να μην υπάρξει περίπτωση σύγχυσης μεταξύ των ζευγών είναι:

```
%day = ( "Sun" => "Sunday", "Mon" => "Monday", "Tue" => "Tuesday", "Wed" =>
"Wednesday", "Thu" => "Thursday", "Fri" => "Friday", "Sat" => "Saturday" );
```

Πρέπει να γνωρίζουμε πως εάν έχουμε μεταβλητή τύπου πίνακα μπορούμε να την μετατρέψουμε σε ευρετήριο, για παράδειγμα:

```
%table=@table;
```

Αυτό όμως, είναι δυνατό μόνο με την βασική προϋπόθεση ότι ο αριθμός των στοιχείων του πίνακα είναι άρτιος και ότι τα περιττά στοιχεία (1, 3, 5 κ.ο.κ.) τα οποία θα αντιστοιχιθούν με τα κλειδιά, θα είναι μοναδικά (δεν είναι δυνατόν να υπάρχει δυο φορές το ίδιο κλειδί!). Το ανάποδο, δηλαδή η κατασκευή ενός πίνακα που να περιέχει τα στοιχεία ενός ευρετηρίου, μπορεί να συμβεί σε κάθε περίπτωση:

```
@table=%table;
```

Πρέπει να διευκρινίσουμε εδώ, πως οι μεταβλητές που είναι χρησιμοποιήσιμες είναι οι τιμές (values) από κάθε ζεύγος, ενώ τα κλειδιά (keys) υπάρχουν για να προσφέρουν πρόσβαση στις τιμές αυτές. Για να το κάνουμε αυτό, πρέπει να χρησιμοποιήσουμε το όνομα του κλειδιού ανάμεσα σε αγκύλες ({}) ενώ το όνομα του ευρετηρίου προηγείται με το σύμβολο $ μπροστά του. Για παράδειγμα:

```
$hash{"key"}="value";
```

Με τον παραπάνω τρόπο, μπορούμε να κάνουμε τόσο ανάθεση κλειδιών-τιμών σε ένα ευρετήριο, όσο και προσπέλαση των τιμών αν ξέρουμε το κλειδί. Προφανώς, η έκφραση $hash{"key"} είναι μια κανονική βαθμωτή μεταβλητή, την οποία μπορούμε να χρησιμοποιήσουμε όπως ακριβώς κάναμε και στην περίπτωση των στοιχείων του πίνακα. Με βάση τα όσα είδαμε, καταλαβαίνουμε πως αν θέλαμε να το περιγράψουμε με κάποιο μαθηματικό ανάλογο, το ευρετήριο είναι μια «συνάρτηση», δηλαδή μια αντιστοίχιση των στοιχείων ενός συνόλου Α (κλειδιά) στα μέλη ενός συνόλου Β (τιμές), στην οποία σε κάθε στοιχείο του Α θα πρέπει να αντιστοιχεί ένα μόνο στοιχείο του Β (έτσι καταλαβαίνουμε και το γιατί τα κλειδιά πρέπει να είναι μοναδικά). Αν τώρα, τύχει και τα στοιχεία του Β να είναι μοναδικά, τότε η συνάρτηση θα είναι αντιστρέψιμη (και όντως, υπάρχει συνάρτηση που κάνει αυτή τη δουλειά στα ευρετήρια, βλ. Πίνακα 12.5)

each	Επιστρέφει, ένα προς ένα, τα ζεύγη από τα κλειδιά και τις τιμές του ευρετηρίου (τα επιστρέφει σε λίστα)
delete	Διαγραφή του κλειδιού που δίνεται ως όρισμα. Η αντίστοιχη τιμή διαγράφεται επίσης.
reverse	Αντιστρέφει τις τιμές και τα κλειδιά (ισχύει όμως μόνο όταν και οι τιμές ειναι μοναδικές)

Πίνακας 12.5: *Συναρτήσεις διαχείρισης ευρετηρίων στην Perl*

12.5. Δομές Ελέγχου

Όπως και σε όλες τις γνωστές γλώσσες προγραμματισμού, έτσι και στην Perl υπάρχουν δομές ελέγχου και επανάληψης. Παρακάτω φαίνονται οι κυριότερες από αυτές και η λειτουργία τους.

12.5.1 If/else/elsif

Με την δομή if/else ελέγχουμε μια συνθήκη. Στο σκέλος της if τοποθετούμε την μία περίπτωση και στο σκέλος της else ελέγχεται αυτόματα η εναλλακτικής της χωρίς να ορίσουμε εμείς κάτι στη συνθήκη. Παρ'

όλα αυτά, έχουμε τη δυνατότητα να περιορίσουμε και την αντίθετη συνθήκη που ελέγχεται ορίζοντας διαφορετικές εναλλακτικές. Αυτή την λειτουργία εξυπηρετεί η ύπαρξη του σετ εντολών if/elsif/else. Στην if, όπως και προηγουμένως, βάζουμε την πρώτη μας συνθήκη, στα επόμενα μπλοκ elsif τοποθετούμε τις ενναλακτικές συνθήκες ελέγχου (οι οποίες όμως αφορούν το αρχικό σύνολο μεταβλητών) και στην περίπτωση που επιθυμούμε την τελική συνθήκη, την τοποθετούμε σε ένα μπλοκ else χωρίς να ορίσουμε περιορισμό. Παράδειγμα:

```
if (condition)
{
      …
}
elsif
{
      …
}
else
{
      …
}
```

Ένα σημείο που χρειάζεται προσοχή, είναι ότι το elsif σε άλλες γλώσσες συντάσσεται ως «elseif» ή «else if». Επίσης, στην Perl κάθε κομμάτι κώδικα (block) ή συνθήκη ελέγχου, θα πρέπει να κλείνεται σε αγκύλες ({}) ακόμα και αν περιέχει μόνο μία γραμμή. Γενικά στην Perl, οι συνθήκες ελέγχου εκτός από τις προφανείς περιπτώσεις, επιστρέφουν ψευδή τιμή (false), όταν η τιμή που πρέπει να ελεγχθεί περιέχει το 0 αλλά και όταν περιέχει την κενή συμβολοσειρά.

12.5.2. While/until

Η Perl μπορεί να επαναλάβει την εκτέλεση μιας ομάδας εντολών με τη δομή while ή until. Η συνθήκη ελέγχεται στην αρχή της δομής ελέγχου στην πρώτη επανάληψη οπότε στην περίπτωση που αυτή δεν ικανοποιείται, παραλείπεται χωρίς να εκτελεστεί ποτέ. Για παράδειγμα:

```
while (condition)
{
      …
}

until (condition)
{
      …
}
```

12.5.3. Do –while/until

Στην Perl συναντάμε και τη δομή επανάληψης do/while. Σε αυτή την δομή τοποθετούμε τη δομή ελέγχου στο τέλος, χρησιμοποιείται δηλαδή όταν θέλουμε να εκτελεσθεί τουλάχιστον μια φορά το σύνολο των εντολών πριν πραγματοποιηθεί ο έλεγχος. Στην Perl μπορούμε να αντικαταστήσουμε την εντολή while με την εντολή until τροποποιώντας την συνθήκη μας ανάλογα. Παράδειγμα:

```
do
{
      …
} while (condition)


do
{
      …
```

```
} until (condition)
```

12.5.4. For

Η εντολή for λειτουργεί όπως και στις υπόλοιπες γνωστές γλώσσες. Η συνθήκη της αποτελείται από τρία μέρη. Το πρώτο μέρος ορίζει την αρχική τιμή της μεταβλητής της συνθήκης, στο δεύτερο ορίζουμε το βήμα με το οποίο θα γίνει η μεταβολή και στο τρίτο την τελική τιμή στην οποία θα πρέπει να τερματίσει η επανάληψή μας. Παράδειγμα:

```
for ($i = 1; $i <= 10; $i ++)
{
      print "$i\n";
}
```

Η εντολή αυτή, είναι η πιο ισχυρή, καθώς μπορούμε να την παραμετροποιήσουμε με διαφορετικούς τρόπους (π.χ. στη συνθήκη μπορούμε να κάνουμε χρήση εκφράσεων όπως $x+$y<10 κ.ο.κ.), αλλα πολλές φορές για πρακτικούς λόγους, η ίδια (απλή) δουλειά γίνεται πιο εύκολα με κάποια άλλη εντολή (όπως η foreach που θα δούμε στην επόμενη ενότητα).

12.5.5. Foreach

Η συγκεκριμένη εντολή εκτελεί πανομοιότυπη λειτουργία με την for με την μόνη διαφορά πως δέχεται ένα μόνο όρισμα (μια λίστα, στα στοιχεία της οποίας ανατρέχει η επαναληπτική διαδικασία). Εκτελεί τον κώδικα που έχουμε ορίσει στο εσωτερικό, για κάθε εγγραφή της δομής μας χωρίς εμείς να ξέρουμε καν το πλήθος. Παράδειγμα:

```
@a = (1,2,3,4,5);
foreash $i (@a)
{
      print "$i\n";
}
```

Ένα ιδιαίτερο χαρακτηριστικό που χρειάζεται προσοχή, είναι ότι αν μέσα στην επανάληψη, αλλάξουμε την τιμή της προσωρινής μεταβλητής (στο παράδειγμα, της $i), τότε αλλάζει και η ίδια η τιμή του πίνακα που έχουμε δώσει σαν όρισμα (αν δώσαμε πίνακα φυσικά).

12.5.6. Last/next/redo

Τα παραπάνω αποτελούν εντολές που προσφέρουν λειτουργικότητες επανάληψης ή εξόδου σε κάποιες από τις προαναφερθείσες δομές επανάληψης.
Last: Η εντολή last χρησιμοποιείται με ανάλογο τόπο με την εντολή break την γλώσσας C, δηλαδή για να σταματήσει την επανάληψη και να βγει από το βρόχο όταν ικανοποιείται κάποια συνθήκη. Έχει εφαρμογή στις δομές ελέγχου:

- for
- foreach
- while
- until

Παράδειγμα:

```
while (condition 1)
{
      …
      if(condition 2)
      {
            …
```

```
        last;
        }
}
```

Next: Η εντολή αυτή έχει την ίδια λειτουργία που έχει η εντολή continue στην C. Τοποθετείται στο τέλος ενός μπλοκ εντολών ώστε να σηματοδοτήσει την μετάβαση, είτε στις επόμενες ελεύθερες εντολές, είτε στο επόμενο μπλοκ εντολών. Παράδειγμα:

```
while (condition 1)
{
        …
        if(condition 2)
        {
                …
        next;
}
```

Redo: Με την τοποθέτηση της εντολής redo στο τέλος ενός μπλοκ εντολών πραγματοποιείται η επανάληψη ολόκληρου του μπλοκ ακόμα και αν θεωρητικά έχει τελειώσει ο κύκλος ικανοποίησης της αρχικής συνθήκης. Παράδειγμα:

```
while (condition 1)
{
        #
        …
        if(condition 2)
        {
                …
                redo;
        }
}
```

12.6. Διαχειριστές Αρχείων (Filehandles) – Είσοδος/Εξοδος

Ένα από τα ισχυρά χαρακτηριστικά της Perl, είναι ο τρόπος με τον οποίο επεξεργάζεται αρχεία. Στην περίπτωση που τα δεδομένα εισόδου ενός προγράμματος της Perl προέρχονται από ένα εξωτερικό αρχείο, ή όταν η έξοδος αποστέλλεται σε ένα άλλο αρχείο, είναι απαραίτητη η χρήση των διαχειριστών αρχείων (filehandles). Ο διαχειριστής αρχείων είναι μια σύνδεση μεταξύ ενός προγράμματος και ενός αρχείου. Ένα πρόγραμμα μπορεί να διαβάζει από ένα διαχειριστή αρχείων ανακτώντας από το περιεχόμενό του μια γραμμή κάθε φορά (αν και όπως θα δούμε, και αυτό μπορεί να αλλάξει), και να τυπώνει σε ένα διαχειριστή προσθέτοντας με αυτόν τον τρόπο δεδομένα σε ένα αρχείο. Ένας διαχειριστής αρχείων μπορεί να έχει ένα οποιοδήποτε όνομα με κεφαλαία, και ορίζεται με τον παρακάτω τρόπο:

```
open MYFILE, 'file.txt';
```

Η συνάρτηση open δέχεται ακριβώς δύο ορίσματα. Το πρώτο είναι το όνομα του διαχειριστή αρχείων και το δεύτερο όρισμα είναι το όνομα του αρχείου που θα ανοιχθεί. Το όνομα του διαχειριστή αρχείων μπορεί να είναι ένα οποιοδήποτε όνομα, αρκεί όλα τα γράμματα του να είναι κεφαλαία. Ιδανικά κάθε αρχείο που ανοίγεται πρέπει να κλείνεται στην συνέχεια. Όταν δημιουργούμε ένα διαχειριστή αρχείων και ανοίγουμε ένα αρχείο, μπορούμε να ορίσουμε και τον τρόπο με τον οποίο θα επεξεργαστούμε το αρχείο αυτό, μέσω αυτού του διαχειριστή. Οι λειτουργίες αυτές και η σύνταξή τους, όσον αφορά τα αρχεία, φαίνονται στον Πίνακα 12.6:

open (IN, "filename")	Ανοίγει ένα υπάρχον αρχείο με το ονομα filename και ονομαζει το διαχειριστή, IN
open (OUT, "> filename")	Δημιουργεί ένα νέο αρχείο με το όνομα filename και γράφει σε αυτό χρησιμοποιώντας το όνομα διαχειριστή OUT
open (IN, ">> filename")	Προσθέτει στο τέλος ενός αρχείου με το όνομα filename και χρησιμοποιεί το όνομα διαχειριστή OUT

Πίνακας 12.6: *Τρόποι με τους οποίους συντάσσεται η εντολή open*

Παραδείγματα διαχειριστών αρχείων για άνοιγμα αρχείου και εκτύπωση σε άλλο αρχείο, φαίνονται στο παρακάτω πρόγραμμα:

```
open IN, "/etc/passwd";
$x=<IN>;
print $x;
close IN;
open OUT, ">tempfile";
print OUT "bla bla bla\n";
```

Το πρόγραμμα αυτό, ανοίγει το αρχείο /etc/passwd (αρχείο του συστήματος στο Linux), μέσω του διαχειριστή IN, διαβάζει μια γραμμή, την τυπώνει, κλείνει το διαχειριστή, ανοίγει το διαχειριστή για το tempfile και τυπώνει σε αυτό την πρόταση «bla bla bla». Η μεταβλητή <IN> είναι ο ειδικός τρόπος που έχει η γλώσσα για να συμβολίζει το τι διαβάζει από το αρχείο μέσω του διαχειριστή (προσέξτε την ομοιότητα με το STDIN). Προσοχή χρειάζεται στο γεγονός ότι αν η μεταβλητή χρησιμοποιηθεί δύο ή περισσότερες φορές, θα προχωράει κάθε φορά και ένα «βήμα», διαβάζονται κάθε φορά την επόμενη γραμμή του αρχείου. Ένας εναλλακτικός τρόπος για να προσπελάσουμε αρκετά πιο εύκολα ένα αρχείο είναι και ο εξής:

```
while(<>)
{
      print $_;
}
```

Με αυτό το απλό πρόγραμμα, που κάνει τη δουλειά της εντολής cat του Unix (και μάλιστα, ίσως την κάνει και πιο γρήγορα), προσπελαύνουμε όλο το αρχείο μέχρι να συναντήσουμε End Of File (EOF) αποθηκεύοντας σε κάθε επανάληψη μια γραμμή στην μεταβλητή $_. Η συνθήκη στο while, ικανοποιείται όσο η τιμή του <> είναι μη κενή, δηλαδή όσο διαβάζει διαδοχικές γραμμές από το αρχείο. Το <> (το «διαμάντι») είναι η πιο απλή περίπτωση διαχειριστή που χρησιμοποιεί η Perl αυτόματα. Για να πάρει τιμές, αρκεί απλά να εισάγουμε ένα αρχείο ως όρισμα εισόδου κατά την κλήση του προγράμματος από την γραμμή εντολών. Για να το κάνουμε αυτό αρκεί να γράψουμε στην γραμμή εντολών μας την κλήση ως εξής:

```
perl programme.pl file.txt
```

Με αυτόν τον τρόπο, η Perl θα ανοίξει αυτόματα το αρχείο file.txt, και θα αρχίσει να στέλνει τα περιεχόμενά του στο <>, γραμμή-γραμμή (αν και αυτό ακόμα μπορεί να αλλάξει όπως θα δούμε). Το $_ είναι το ειδικό όνομα που δίνει η Perl στη μεταβλητή που περιέχει κάθε φορά τα στοιχεία αυτά, και είναι μια δεσμευμένη ονομασία, όπως είπαμε στην αρχή. Ο χρήστης δεν μπορεί να τη δημιουργήσει αλλά ούτε και να την τροποποιήσει. Το παραπάνω πρόγραμμα και ο συγκεκριμένος τρόπος, είναι ιδιαίτερα εύχρηστος και προτιμάται για απλές δουλειές. Όταν όμως θέλουμε να κάνουμε πιο σύνθετες εργασίες (όπως π.χ. να επεξεργαστούμε παράλληλα πολλά αρχεία), ή να κάνουμε το πρόγραμμα πιο αυστηρό (π.χ. να βγάζει μήνυμα σφάλματος αν το αρχείο που δώσαμε σαν όρισμα δεν υπάρχει), τότε πρέπει να χρησιμοποιήσουμε τον κλασικό τρόπο με τους διαχειριστές αρχείων. Στην περίπτωση αυτή, τα πολλαπλά ορίσματα που θα περάσουμε στο πρόγραμμα, αποθηκεύονται αυτόματα στον πίνακα @ARGV. Έτσι, τα ονόματα των αρχείων είναι $ARGV[0], $ARGV[1] κ.ο.κ. τα οποία μπορούμε να τα χρησιμοποιήσουμε στην open. Φυσικά, με τον

ίδιο τρόπο μπορούμε να περάσουμε ως όρισμα, οποιαδήποτε άλλη βαθμωτή τιμή που θα χρησιμοποιηθεί εσωτερικά στο πρόγραμμα (και όχι μόνο ονόματα αρχείων).

12.7. Κανονικές Εκφράσεις

Με τις κανονικές εκφράσεις (regular expressions) μπορούμε να ελέγξουμε εάν μια συμβολοσειρά είναι ίδια με κάποια άλλη, ή περιέχει ένα συγκεκριμένο μοτίβο. Η λειτουργία αυτή είναι ιδιαίτερα χρήσιμη καθώς όπως είδαμε στο κεφάλαιο 4, τα μοτίβα είναι ιδιαίτερα διαδεδομένα στην ανάλυση αλληλουχιών στη βιοπληροφορική. Οι κανονικές εκφράσεις οροθετούνται από τις καθέτους (/) και περιέχουν μια ακολουθία από χαρακτήρες που πρέπει να ταιριάξουν με τους χαρακτήρες μέσα στο σώμα μιας συμβολοσειράς. Για παράδειγμα η εντολή:

```
$dna=~/GAATTC/;
```

ελέγχει αν στη συμβολοσειρά $dna περιέχεται η αλληλουχία GAATTC η οποία αντιστοιχεί σε θέση δράσης μιας περιοριστικής ενδονουκλεάσης. Όταν σε μια συνθήκη βρούμε κάτι σαν /GAATTC/, έχουμε μια απλή συνθήκη ταιριάσματος συμβολοσειρών. Επίσης είναι δυνατό μέσω των κανονικών εκφράσεων όταν βρεθεί μια ακολουθία μέσα σε μια συμβολοσειρά να αντικατασταθεί με κάτι άλλο, με την απλή εντολή /GAATTC/CTTAAG/ (σε αυτή την περίπτωση αντικαθιστούμε το GAATTC με CTTAAG).

Στην Perl οι κανονικές εκφράσεις είναι τόσο ευρέως χρησιμοποιούμενες που οι συναρτήσεις και ο τρόπος επεξεργασίας τους έχει απλοποιηθεί στο ελάχιστο διότι ενισχύονται με τη χρήση μεταχαρακτήρων και ποσοδεικτών. Ένας μεταχαρακτήρας εκπροσωπεί μια ολόκληρη κλάση χαρακτήρων. Παραδείγματος χάριν, η τελεία (.) ταιριάζει με οποιοδήποτε χαρακτήρα εκτός από την αλλαγή γραμμής, ενώ το \d δηλώνει οποιοδήποτε ψηφίο. Στον Πίνακα 12.7 παρουσιάζονται οι πιο συχνά χρησιμοποιούμενοι μεταχαρακτήρες.

Μεταχαρακτήρας	**Περιγραφή**
.	Οποιοσδήποτε χαρακτήρας εκτός από την αλλαγή γραμμής
^	Αρχή μιας γραμμής
$	Τέλος μιας γραμμής
\w	Οποιοσδήποτε χαρακτήρας λέξης
\W	Οποιοσδήποτε χαρακτήρας εκτός από χαρακτήρα λέξης
\s	Χαρακτήρας διαστήματος
\S	Οποιοσδήποτε χαρακτήρας εκτός από χαρακτήρα διαστήματος
\d	Οποιοδήποτε ψηφίο
\D	Οποιοσδήποτε χαρακτήρας εκτός από ψηφίο

Πίνακας 12.7: *Μεταχαρακτήρες κανονικών εκφράσεων*

Εξ' ορισμού, οποιοσδήποτε χαρακτήρας ή μεταχαρακτήρας σε μια κανονική έκφραση ταιριάζει ακριβώς μια φορά. Μια αναζήτηση με κανονική έκφραση μπορεί να προσπαθεί να ταιριάξει όσο το δυνατόν περισσότερες φορές μέσα στο «κείμενο» την ακολουθία μας, ή όσο το δυνατόν λιγότερες αντίστοιχα. Με την τοποθέτηση ενός ποσοδείκτη (Quantifier) μετά το χαρακτήρα, η Perl μπορεί να ταιριάξει το χαρακτήρα αυτό για συγκεκριμένο αριθμό επαναλήψεων, ή και διάστημα επαναλήψεων. Ο απλούστερος ποσοδείκτης είναι ο {n}, ο οποίος προσπαθεί να ταιριάζει το πρότυπο ακριβώς n φορές. Για παράδειγμα η εντολή:

```
$sequence=~/AAAT{5}CCG/;
```

ελέγχει αν η συμβολοσειρά $sequence ταιριάζει στην αλληλουχία AAATTTTTCCG (δηλαδή, ελέγχει το πρότυπο PROSITE A-A-A-T(5)-C-C-G). Προσέξτε, ότι αυτή είναι μια έκφραση που μπορεί να έχει αληθείς (true) ή ψευδείς (false) τιμες, και κατά συνέπεια θα πρέπει να ελεγχθεί, συνήθως σε κάποια δομή if. Φυσικά, όπως είδαμε στο κεφάλαιο 4, όλες οι εκφράσεις PROSITE αντιστοιχούν σε μια κανονική έκφραση, οπότε, μπορούμε να χρησιμοποιήσουμε αυτούσιους τους κανόνες που είδαμε εκεί για να πραγματοποιήσουμε αναζητήσεις. Στον Πίνακα 12.8 παρουσιάζονται οι ποσοδείκτες κανονικών εκφράσεων.

Ποσοδείκτης	Περιγραφή
?	0 ή 1 εμφάνιση (είναι σημαντικός, γιατί αναγκάζει το πρότυπο να μην είναι «άπληστο»)
+	1 ή περισσότερες εμφανίσεις
.	0 ή περισσότερες εμφανίσεις
{n,m}	Μεταξύ n και m εμφανίσεων
{n, }	Τουλάχιστον n εμφανίσεις
{ ,m}	Όχι περισσότερες από m εμφανίσεις

Πίνακας 12.8: *Ποσοδείκτες κανονικών εκφράσεων*

12.8. Εφαρμογές της Perl στη Βιοπληροφορική

Στην ενότητα αυτή, θα παρουσιάσουμε βήμα-βήμα, κάποια βασικά προγράμματα σε Perl, τα οποία είναι μεν είναι χρήσιμα σε διάφορα στάδια της ανάλυσης βιολογικών αλληλουχιών, αλλά ταυτόχρονα αποτελούν και ιδανικά παραδείγματα για να παρουσιαστεί και ο τρόπος χρήσης των βασικών λειτουργιών της γλώσσας.

12.8.1 Μετατροπή αρχείου Uniprot σε μορφή Fasta

Μια διαδικασία ρουτίνας στις αναλύσεις αλληλουχιών, είναι η μετατροπή μιας μορφής αρχείων σε μια άλλη. Η πιο συνηθισμένη από τις περιπτώσεις, είναι να θέλουμε να μετατρέψουμε την πληροφορία από ένα αρχείο Uniprot σε μια μορφή που θα μπορεί να χρησιμοποιηθεί από τα περισσότερα προγράμματα ανάλυσης αλληλουχιών, δηλαδή στη μορφή fasta. Για να γράψουμε ένα τέτοιο πρόγραμμα, θα πρέπει να παρατηρήσουμε το αρχείο της Uniprot και να εντοπίσουμε τη δομή του. Για το fasta θα χρειαστούμε και ένα όνομα για την αλληλουχία, οπότε η καλύτερη επιλογή είναι να χρησιμοποιήσουμε το AC ή το ID. Όλες οι γραμμές ενός αρχείου Uniprot ξεκινάνε με δυο χαρακτήρες που καθορίζουν το πεδίο (π.χ. το AC), ενώ μόνο οι γραμμές που περιέχουν την αλληλουχία ξεκινάνε με δύο κενά. Έπειτα, όλες οι γραμμές έχουν 3 κενά και μετά αρχίζει η πληροφορία. Το πρόγραμμα αυτό (uniprot2fasta.pl), είναι πολύ απλό και δίνεται παρακάτω:

```
while (<>)
{
      if ($_=~/^AC\s{3}(.*?)\;/)
      {
          print ">$1\n";
      }
      if ($_=~/^\s{5}(.*)/)
      {
          $sequence=$1;
          $sequence=~s/\s//g;
          print "$sequence\n";
      }
}
```

Για ευκολία, χρησιμοποιούμε το <>. Το πρόγραμμα διαβάζει γραμμή-γραμμή και πραγματοποιεί έλεγχο στη μεταβλητή $_. Στο πρώτο if, γίνεται ο έλεγχος για το AC. Ψάχνουμε για μία γραμμή που να ξεκινάει με AC, να ακολουθούν 3 κενά και μετά οσεσδήποτε επαναλήψεις οποιουδήποτε χαρακτήρα, μέχρι να εμφανιστεί το (;) το οποίο διαχωρίζει τα πολλαπλά AC (θέλουμε να κρατήσουμε μόνο το πρώτο από αυτά). Το ? χρησιμοποιείται εδώ για να κάνει το πρότυπο όχι άπληστο (non-greedy), δηλαδή για να το αναγκάσει να σταματήσει στη μικρότερη επανάληψη του μοτίβου που θα βρει. Το $1 στη συνέχεια, είναι μια άλλη δεσμευμένη μεταβλητή στην οποία αποθηκεύεται το περιεχόμενο της παρένθεσης που έχουμε ορίσει μέσα στο πρότυπο. Με αυτόν τον τρόπο μπορούμε να απομονώσουμε από ένα μεγάλο πρότυπο, μόνο το κομμάτι που χρειαζόμαστε (εδώ, το AC). Προφανώς, αν είχαμε περισσότερες παρενθέσεις, θα είχαμε και $2, $3 κ.ο.κ.

Στον δεύτερο έλεγχο, επιχειρούμε να εντοπίσουμε τις γραμμές που έχουν στην αρχή 5 κενούς χαρακτήρες (δηλαδή, τις γραμμές με την αλληλουχία) και να κρατήσουμε όλους τους υπόλοιπους (δηλαδή την ίδια την αλληλουχία). Επειδή στη μορφή Uniprot η αλληλουχία δίνεται σε διαδοχικές δεκάδες με κενά

ανάμεσα, το περιεχόμενο της μεταβλητής $sequence θα πρέπει σε κάθε βήμα να το αλλάζουμε. Έτσι, χρησιμοποιούμε το ~s/\s//g το οποίο αλλάζει τα κενά με τον κενό χαρακτήρα (το g στο τέλος, συμβολίζει ότι πρέπει να γίνει σε όλη τη συμβολοσειρά, globally- αν δεν το είχαμε βάλει, θα άλλαζε μόνο το πρώτο κενό). Σε κάθε επανάληψη, το πρόγραμμα τυπώνει και μία γραμμή της αλληλουχίας, μέχρι να σταματήσει να βρίσκει αντίστοιχες γραμμές. Προσέξτε ότι τα δύο μπλοκ με τα if δεν είναι απαραίτητο να βρίσκονται καν στη συγκεκριμένη σειρά. Το πρόγραμμα διαβάζει μια γραμμή και κάνει όλους τους ελέγχους ανεξάρτητα, άρα, θα μπορούσε η σειρά να είναι και η αντίστροφη.

Σε κάποιες περιπτώσεις, ιδιαίτερα αν τις αλληλουχίες πρόκειται να τις επεξεργαστούμε εμείς με κάποιο άλλο πρόγραμμα μας, ίσως είναι προτιμότερο να κατασκευάσουμε μια ειδική μορφή fasta, στην οποία η αλληλουχία θα δίνεται σε μόνο μία γραμμή. Με τον τρόπο αυτό θα μπορεί να διαβαστεί πολύ πιο εύκολα. Το παρακάτω πρόγραμμα (uniprot2line.pl) κάνει αυτήν ακριβώς τη δουλειά.

```
$/="\/\/\n";
while (<>)
{
    if ($_=~/^AC\s{3}(.*?)\;/m)
    {
        print ">$1\n";
    }
    while ($_=~/^\s{5}(.*)/mg)
    {
        $sequence=$1;
        $sequence=~s/\s//g;
        print "$sequence";
    }
print "\n";
}
```

Το πρόγραμμα αυτό, μοιάζει αρκετά με το προηγούμενο αλλά έχει κάποιες σημαντικές διαφοροποιήσεις. Στην αρχή η εντολή $/="\/\/\n" λέει στο πρόγραμμα ότι θα πρέπει να αλλάξει ο προκαθορισμένος τρόπος ανάγνωσης από αρχείο (γραμμή-γραμμή) και το πρόγραμμα να διαβάζει πλέον μέχρι να συναντήσει το «//\n». Με τον τρόπο αυτό, το πρόγραμμα σε κάθε επανάληψη θα διαβάζει μια ολόκληρη εγγραφή της Uniprot και θα την αποθηκεύει στο $_. Η άλλη σημαντική διαφορά, βρίσκεται στον χαρακτήρα m στον έλεγχο των κανονικών εκφράσεων. Ο χαρακτήρας αυτός προειδοποιεί την κανονική έκφραση ότι η συμβολοσειρά ($_) περιέχει πολλαπλές γραμμές (multilne). Τέλος, το if έχει αντικατασταθεί από το while, όπως επίσης έχει προστεθεί και το g (global) στην ταύτιση των προτύπων, για να μπορέσει να ανταποκριθεί το πρότυπο και να γίνει η σωστή ταύτιση του σε πολλαπλές γραμμές.

Τέλος, θα μπορούσε να υπάρξει και η περίπτωση στην οποία θα θέλαμε να μετατρέψουμε το αρχείο από fasta με πολλές γραμμές σε fasta με μία. Αυτή τη δουλειά κάνει το παρακάτω πρόγραμμα (fasta2line.pl):

```
$/=">";

while (<>)
{
      $entry=$_;
      chop $entry;
      $entry= ">"."$entry";
      $entry=~/>(.+?)\n(\C*)/g;
      $name=$1;
      $sequence=$2;
      $sequence=~s/\n//g;

      if ($name ne "")
      {
            print ">$name\n$sequence\n";
      }
}
```

```
$/="\n";
```

Όμοια με παραπάνω, το αρχείο διαβάζεται καταγραφή-καταγραφή (μέχρι να συναντηθεί το «>»). Στη συνέχεια, διαχωρίζει το όνομα της πρωτεΐνης από την αλληλουχία, αφαιρεί τις αλλαγές γραμμής που υπάρχουν και τυπώνει την αλληλουχία συνεχόμενη.

12.8.2. Προσομοίωση με τυχαίες αλληλουχίες

Οι προσομοιώσεις για την παραγωγή τυχαίων αλληλουχιών αποτελούν μια συνηθισμένη πρακτική σε πολλές διαδικασίες στη βιοπληροφορική, κυρίως για να εντοπιστούν οι στατιστικές ιδιότητες κάποιου φαινομένου (ροές, στατιστική σημαντικότητα στοίχισης κ.ο.κ.). Το παρακάτω πρόγραμμα φτιάχνει 500 τυχαίες ακολουθίες αποτελούμενες από 200 αμινοξέα η καθεμιά:

```
@aa = (A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y);
for ( $i=0; $i<500; $i++ ) {
  print '>Random', "$i\n";
      for( $j=0; $j<200; $j++ ) {
            $r = $aa[ int (rand 20)];
            print $r;
            print "\n" if ($j+1)%60 == 0 and $j;
       }
  print "\n";
}
```

Το πρόγραμμα στην αρχή ορίζει έναν πίνακα που περιέχει τα αμινοξέα (η σειρά τους δεν παίζει ρόλο). Μετά, εκτελεί μια επανάληψη στην οποίο χρησιμοποιείται η συνάρτηση rand για να παράγει τυχαίους αριθμούς στο διάστημα 0-20, οι οποίοι στρογγυλοποιούνται και χρησιμοποιούνται σαν δείκτες (index) στον πίνακα για να ανακτηθεί το τυχαίο αμινοξύ που παράγεται. Πριν κλείσει ο βρόχος γίνεται και ένας (προαιρετικός) έλεγχος για το αν έχει συμπληρωθεί ο αριθμός 60 που χαρακτηρίζει τις περισσότερες μορφές του αρχείο fasta. Διάφορες τροποποιήσεις θα μπορούσαν να γίνουν σε αυτό το πρόγραμμα, όπως π.χ. να δέχεται το μήκος των αλληλουχιών και τον αριθμό τους ως όρισμα, να τα παράγει από μια συγκεκριμένη κατανομή. ή να παράγει αμινοξέα όχι ισοπίθανα αλλά με βάση κάποιες προκαθορισμένες συχνότητες. Οι τροποποιήσεις αυτές αφήνονται σαν άσκηση.

12.8.3. Υπολογισμός συχνοτήτων αμινοξέων

Μια πολύ συνηθισμένη ανάλυση, είναι να υπολογίσουμε το ποσοστό των αμινοξέων σε μια αλληλουχία πρωτεϊνών, ή το ποσοστό των νουκλεοτιδίων σε μια αλληλουχία DNA/RNA. Όπως είδαμε σε προηγούμενα κεφάλαια, τέτοιες αναλύσεις μπορούν να χρησιμοποιηθούν σε προγνωστικές μεθόδους ή στην υπολογιστική γονιδιωματική. Το παρακάτω πρόγραμμα δέχεται ως είσοδο ένα αρχείο με πρωτεΐνες σε μορφή FASTA με μία γραμμή και υπολογίζει το ποσοστό των αμινοξέων κάθε πρωτεΐνης.

```
@aa = (A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y);
while (<>)
{
      if ($_=~/^>/)
      {
            $id=$_;
            chomp $id;
            print $id."\t";
            $seq=<>;
            chomp $seq;
      }
      $length=length($seq)+1;
      foreach $z(@aa)
      {
            $count = $seq =~s/$z//g;
```

```
            $diairesi = $count/$length;
            $pososto=sprintf( "%.3f", $diairesi );
            print $z."\t".$count.'/'.$length."\t[".$pososto."]\n";
      }
      print "\n";
}
```

Το πρόγραμμα στην αρχή ορίζει τον πίνακα με τα 20 αμινοξέα και στη συνέχεια διαβάζει το αρχείο εισόδου γραμμή-γραμμή. Σε κάθε γραμμή ελέγχει την ύπαρξη του χαρακτήρα > που σηματοδοτεί το header και τότε κρατάει το όνομα της πρωτεΐνης και διαβάζει την επόμενη γραμμή που είναι η αλληλουχία η ίδια (προσέξτε τη χρήση του @seq=<>). Μετά, αφού μετρήσει το μήκος της κάθε αλληλουχίας, το πρόγραμμα εκτελεί ένα βρόχο foreach στον οποίο μετράει ένα-ένα τα αμινοξέα του κάθε τύπου. Αυτό γίνεται πολύ εύκολα με την εντολή $count = $seq =~s/$z//g. Στο δεξί σκέλος, κάνει μια αντικατάσταση όλων των άλλων αμινοξέων (εκτός από αυτό που μετράει κάθε φορά) με τον κενό χαρακτήρα, και τελικά μετράει το πόσες φορές το βρήκε στην αλληλουχία (η εντολή αυτή θα μπορούσε να γίνει και με τη συνάρτηση length αλλά και με άλλους τρόπους - πχ $seq =~s/$z//g; $count=length($seq);). Τέλος, τυπώνει τα αποτελέσματα σε μία γραμμή για κάθε πρωτεΐνη χρησιμοποιώντας την sprintf για καλύτερη μορφοποίηση.

12.8.4. Εύρεση ανοιχτών πλαισίων ανάγνωσης σε αλληλουχίες DNA

Ένα πρώτο βήμα στην εύρεση γονιδίων και στη γονιδιωματική ανάλυση είναι η αναγνώριση των ανοιχτών πλαισίων ανάγνωσης (open reading frames) σε αλληλουχίες DNA. Σε αυτή την ενότητα θα παρουσιάσουμε ένα απλό τέτοιο πρόγραμμα το οποίο θα πραγματοποιεί αυτού του είδους την ανάλυση. Το πρόγραμμα αυτό δέχεται ως είσοδο μια ακολουθία DNA και αρχικά βρίσκει τη συμπληρωματική της. Στη συνέχεια, μεταφράζει και τις 2 ακολουθίες χρησιμοποιώντας τα 6 πιθανά πλαίσια ανάγνωσης και τυπώνει τις αμινοξικές αλληλουχίες των υποθετικών πρωτεϊνών που προκύπτουν από αυτό.

```
%genetic_code = (
'GCA'=>'A', #Alanine
'GCC'=>'A', #Alanine
'GCG'=>'A', #Alanine
'GCT'=>'A', #Alanine
'AGA'=>'R', #Arginine
'AGG'=>'R', #Arginine
'CGA'=>'R', #Arginine
'CGC'=>'R', #Arginine
'CGG'=>'R', #Arginine
'CGT'=>'R', #Arginine
'AAC'=>'N', #Asparagine
'AAT'=>'N', #Asparagine
'GAC'=>'D', #Aspartic acid
'GAT'=>'D', #Aspartic acid
'TGC'=>'C', #Cysteine
'TGT'=>'C', #Cysteine
'GAA'=>'E', #Glutamic acid
'GAG'=>'E', #Glutamic acid
'CAA'=>'Q', #Glutamine
'CAG'=>'Q', #Glutamine
'GGA'=>'G', #Glycine
'GGC'=>'G', #Glycine
'GGG'=>'G', #Glycine
'GGT'=>'G', #Glycine
'CAC'=>'H', #Histidine
'CAT'=>'H', #Histidine
'ATA'=>'I', #Isoleucine
'ATC'=>'I', #Isoleucine
'ATT'=>'I', #Isoleucine
'TTA'=>'L', #Leucine
```

```
'TTG'=>'L', #Leucine
'CTA'=>'L', #Leucine
'CTC'=>'L', #Leucine
'CTG'=>'L', #Leucine
'CTT'=>'L', #Leucine
'AAA'=>'K', #Lysine
'AAG'=>'K', #Lysine
'ATG'=>'M', #Methionine
'TTC'=>'F', #Phenylalanine
'TTT'=>'F', #Phenylalanine
'CCA'=>'P', #Proline
'CCC'=>'P', #Proline
'CCG'=>'P', #Proline
'CCT'=>'P', #Proline
'AGC'=>'S', #Serine
'AGT'=>'S', #Serine
'TCA'=>'S', #Serine
'TCC'=>'S', #Serine
'TCG'=>'S', #Serine
'TCT'=>'S', #Serine
'ACA'=>'T', #Threonine
'ACC'=>'T', #Threonine
'ACG'=>'T', #Threonine
'ACT'=>'T', #Threonine
'TGG'=>'W', #Tryptophan
'TAC'=>'Y', #Tyrosine
'TAT'=>'Y', #Tyrosine
'GTA'=>'V', #Valine
'GTC'=>'V', #Valine
'GTG'=>'V', #Valine
'GTT'=>'V', #Valine
'TAA'=>'-', #STOP
'TAG'=>'-', #STOP
'TGA'=>'-', #STOP
);

$seq="AAAAAAATTAATAGATGAACATATATATAGATTTTCTATATAGACCCTCTACCCGATAAGGCTAC";
$seq2=$seq;
$seq2=~tr/ATCG/TAGC/;
$seq2=reverse($seq2);

print "Forward Strand\n";
Translate($seq);

print "Reverse Strand\n";
Translate($seq2);

sub Translate{
$sub_seq=$_[0];
      for($i=0;$i<=length($sub_seq)-3;$i++)
      {
            $x=substr($sub_seq,$i,3);
            if ($x eq 'ATG')
            {
                  $pos=$i+1;
                  print "position $pos\n";
                  for($j=$i;$j<=length($sub_seq)-3;$j=$j+3)
                  {
                        $y=substr($sub_seq,$j,3);
                        $k=$genetic_code{$y};
                        if($k eq '-')
```

```
                        {
                              print"\n";
                              last;
                        }
                        print "$k";
                  }
            }
      }
}
```

Για ευκολία, στο πρόγραμμα αυτό θα έχουμε την αλληλουχία αποθηκευμένη εσωτερικά σε μια μεταβλητή. Φυσικά, μπορεί να γίνει τροποποίηση και το πρόγραμμα να την διαβάζει από αρχείο. Μεγάλο μέρος του κώδικα αναγκαστικά χρησιμοποιείται για να οριστεί ένα ευρετήριο για τη συσχέτιση των κωδικονίων με τα αμινοξέα (αλλά φυσικά, και αυτό θα μπορούσε να γίνει μέσα από εισαγωγή αρχείου). Στη συνέχεια το tr βρίσκει με μια απλή εντολή τη συμπληρωματική αλυσίδα, ενώ η reverse, αντιστρέφει τη σειρά έτσι ώστε να διαβαστεί από την ορθή κατεύθυνση (υποθέτουμε ότι το 5' άκρο είναι στα αριστερά).

Το επόμενο σημαντικό χαρακτηριστικό αυτού του προγράμματος, είναι ο ορισμός μια υπορουτίνας (sub-routine). Τέτοιες περιπτώσεις συναντάμε όταν έχουμε να πραγματοποιήσουμε μια δεδομένη ενέργεια πολλές φορές (εδώ, έχουμε τη μετάφραση των δύο αλυσίδων) και δεν θέλουμε να έχουμε διπλό κώδικα μέσα στο πρόγραμμα (τόσο για αποφυγή λαθών, όσο και για να είναι πιο ευανάγνωστο). Οι υπορουτίνες είναι πολύ βολικές γιατί κρατάνε τον κώδικα ευανάγνωστο, ενώ μπορούν να τοποθετηθούν σε οποιοδήποτε σημείο του προγράμματος και να κληθούν από οποιοδήποτε σημείο (λόγω του μεταγλωττιστή). Ο ορισμός μιας υπορουτίνας αλλά και η κλήση της, είναι απλή υπόθεση όπως φαίνεται στο παραπάνω πρόγραμμα. Το μόνο σημείο που χρειάζεται διευκρίνιση είναι στο πως η Perl περνάει τα ορίσματα της υπορουτίνας. Αυτό γίνεται με τον ειδικού σκοπού πίνακα @_ στον οποίο αυτά αποθηκεύονται. Έτσι, εσωτερικά στην υπορουτίνα το πρώτο όρισμα είναι το $_[0], το δεύτερο (αν υπάρχει) είναι το $_[1] κ.ο.κ. Προσοχή χρειάζεται στο γεγονός ότι οι μεταβλητές στην Perl είναι εξ'ορισμού "παγκόσμιες" (global), δηλαδή είναι προσβάσιμες από κάθε σημείο του κώδικα. Έτσι, για να αποφύγουμε λάθη καλό είναι να δώσουμε διαφορετικό όνομα στην ακολουθία μέσα στην υπορουτίνα. Αν θέλαμε να έχουμε ιδιωτικές μεταβλητές (private), δηλαδή μεταβλητές που θα είναι προσβάσιμες μόνο στο συγκεκριμένο σημείο (βρόχο ή υπορουτίνα), θα έπρεπε να τις ορίσουμε με το my (πχ my $seq= κ.ο.κ.).

Επί της ουσίας, η υπορουτίνα πραγματοποιεί με απλό τρόπο την μετάφραση του γενετικού υλικού. Στην αρχή διαβάζει επικαλυπτόμενα παράθυρα μήκους 3 (δηλαδή, τριπλέτες) στην αλληλουχία, για να εντοπίσει το κωδικόνιο έναρξης (ATG). Για να το κάνει αυτό, χρησιμοποιεί το substr έτσι ώστε να εξάγει την τριπλέτα και μετά την περνάει στο ευρετήριο σαν κλειδί για να βρει την αντίστοιχη τιμή. Μετά, καθώς έχει ξεκινήσει η μετάφραση εξακολουθεί να διαβάζει τριπλέτες, αλλά πλέον, μη επικαλυπτόμενες, και τυπώνει τα διαδοχικά αμινοξέα έως ότου συναντήσει κωδικόνιο λήξης (προσέξτε εδώ τη χρήση του last). Όταν τελειώσει αυτός ο κύκλος, το πρόγραμμα επιστρέφει και διαβάζει ξανά επικαλυπτόμενες τριπλέτες από το σημείο που είχε σταματήσει. Αυτό είναι πολύ σημαντικό γιατί τα ανοιχτά πλαίσια ανάγνωσης μπορεί να είναι και επικαλυπτόμενα.

12.8.5. Εύρεση βακτηριακών λιποπρωτεϊνών

Όπως είδαμε στο κεφάλαιο 5, οι βακτηριακές λιποπρωτεΐνες έχουν μια σηματοδοτική αλληλουχία (signal peptide) η οποία μοιάζει αρκετά με αυτή των εκκρινόμενων πρωτεϊνών, αλλά στο καρβοξυτελικό της άκρο φέρει μια χαρακτηριστική αλληλουχία η οποία αναγνωρίζεται από ειδικό ένζυμο, το οποίο αποκόπτει το πεπτίδιο αυτό και ακολούθως η ώριμη πρωτεΐνη προσκολλάται στα λιπίδια της μεμβράνης με ομοιοπολικό δεσμό. Η αλληλουχία που αναγνωρίζει το ένζυμο, έχει μια συντηρημένη κυστεΐνη στο καρβοξυτελικό της άκρο (περίπου στη θέση 17-30 της πρόδρομης πρωτεΐνης, εκεί που γίνεται και η τροποποίηση), ενώ στις προηγούμενες θέσεις υπάρχουν κυρίως Αλανίνες και Βαλίνες. Διάφορα πρότυπα είχαν περιγραφεί από τη δεκαετία του 1980, αλλά το πιο γνωστό είναι το λεγόμενο PS00013, όπως είναι γνωστό από τον κωδικό της PROSITE. Παρ' όλα αυτά, έχουν περιγραφεί και εναλλακτικά πρότυπα πολλές φορές ακόμα και χρόνια αργότερα, όπως το [LVI]-[ASTVI]-[GAS]-C το οποίο χρησιμοποιήθηκε για να κατασκευαστεί η βάση των βακτηριακών λιποπρωτεϊνών (DOLOP). Το 2002, οι Sutcliffe και Harrington μελετώντας λιποπρωτεΐνες από

βακτήρια θετικά κατά Gram, κατέληξαν σε ένα πιο αυστηρό αλλά ταυτόχρονα και πιο περιεκτικό πρότυπο, το οποίο περιγράφει καλύτερα τις λιποπρωτεΐνες αυτών των βακτηρίων (Sutcliffe & Harrington, 2002).

Θα χρησιμοποιήσουμε ένα αρχείο με 63 αμινοξικές αλληλουχίες πρωτεϊνών όπως αναφέρθηκαν στην εργασία των (Juncker et al., 2003). Το πρόγραμμα αυτό δέχεται ως είσοδο ένα αρχείο με πρωτεΐνες σε μορφή FASTA με μία γραμμή, ελέγχει την ύπαρξη του συγκεκριμένου κάθε φορά μοτίβου και τυπώνει τον αριθμό των πρωτεϊνών στις οποίες αυτό βρέθηκε (προσέξτε τη διαφορά των κανονικών εκφράσεων, από τα πρότυπα PROSITE).

```
while (<>){
      if    ($_=~/^>(.*)/)
      {
            $name=$1;
            $seq=<>;
            if($seq=~/(.*LA[GA]C)/)
            {
                  $x=length($1);
                  $a=$a+1;
            }
      }
}
print "$a LIPOPROTEINS FOUND";
```

Για τα άλλα μοτίβα το περιεχόμενο της δομή ελέγχου if είναι το μόνο που αλλάζει στον κώδικα της Perl:

```
if($seq=~/(.*[LVI][ASTG][GA]C)/)
```

```
if($seq=~/(.*[^DERK]{6}[LIVMFWSTAG]{2}[LIVMFYSTAGCQ][AGS]C)/)
```

```
if($seq=~/^([MV].{0,13}[RK][^DERK]{6,20}[LIVMFESTAG][LVIAM][IVMSTAFG][AG]C)/)
```

Θα μπορούσε βέβαια να φτιαχτεί και μια υπορουτίνα για να κάνει το ίδιο και να ελέγχει όλα τα πιθανά πρότυπα (αφήνεται ως άσκηση). Τέλος, για να δείξουμε με γραφικό τρόπο την ύπαρξη των παραπάνω μοτίβων θα πρέπει να κάνουμε μια ειδικής μορφής πολλαπλή στοίχιση, στην οποία όλες οι αλληλουχίες που έχουν το συγκεκριμένο μοτίβο θα έχουν στοιχιθεί στο καρβοξυτελικό άκρο του πεπτιδίου οδηγητή, έτσι ώστε να μπορέσουμε να οπτικοποιήσουμε τη συντήρηση στην περιοχή αυτή. Το παρακάτω πρόγραμμα, για το μοτίβο το οποίο βρέθηκε στις περισσότερες πρωτεΐνες, τυπώνει την αλληλουχία των πεπτιδίων οδογητών και προσθέτει μπροστά όσους χαρακτήρες κενού (δηλαδή "–") είναι απαραίτητοι έτσι ώστε όλα τα πεπτίδια να στοιχιθούν στο καρβοξυτελικό άκρο, δηλαδή στην κυστεϊνη. Το αποτέλεσμα του προγράμματος τυπώνεται στην οθόνη και θα χρησιμοποιηθεί ως είσοδος στο πρόγραμμα **WebLogo** (http://weblogo.berkeley.edu/) (Crooks, Hon, Chandonia, & Brenner, 2004) για την οπτικοποίηση της στοίχισης.

```
while (<>)
{
if ($_=~/^>(.*)/)
      {
            $name=$1;
            $seq=<>;
            if ($seq=~/(.*[^DERK]{6}[LIVMFWSTAG]{2}[LIVMFYSTAGCQ][AGS]C)/)
            {
                  $x=length($1);
                  push @table,$1;
            }
      }
      if($x>$max)
      {
```

```
            $max=$x;
      }
}
foreach $signal(@table)
{
      $i="-" x ($max-length($signal));
      $signal=$i.$signal;
      print "$signal\n";
}
```

Το πρόγραμμα διαβάζει το αρχείο εισόδου γραμμή-γραμμή. Σε κάθε γραμμή ελέγχει την ύπαρξη του χαρακτήρα > που σηματοδοτεί το header και τότε κρατάει το όνομα της πρωτεΐνης και διαβάζει την επόμενη γραμμή που είναι η αλληλουχία η ίδια (προσέξτε τη χρήση του @seq=<>). Εντοπίζει το πρότυπο της λιποπρωτεΐνης, αποκόπτει το πεπτίδιο οδηγητή και το αποθηκεύει σε έναν πίνακα (όμοια έχει κάνει και για το αντίστοιχο όνομα της πρωτεΐνης). Στην επόμενη φάση, αφού έχει εντοπίσει την αλληλουχία με το μέγιστο μήκος, ανατρέχει στις αλληλουχίες του πίνακα και τυπώνει στην αρχή τους τόσα "-" όσα απαιτούνται για να επιτευχθεί η στοίχιση. Η στοίχιση των πεπτιδίων οδηγητών όπως προέκυψε από την εκτέλεση του προγράμματος αυτού, δίνεται παρακάτω.

```
      ------------------MRRCMPLVAASVAALMLAGC
      -----------------MKLKQLFAITAIASALVLTGC
      ------------------MKLLSKIMIIALAASMLQAC
      -------------MNKNRGFTPLAVVLMLSGSLALTGC
      --------------------MKRQALAAMIASLFALAAC
      --------------------MRLLPLVAAATAAFLVVAC
      ----------------------MRIVIFILGILLTSC
      ---------------------MFKRFIFITLSLLVFAC
      ----------------------MLKKVYYFLIFLFIVAC
      ---------------------MKKILLTVSLGLALSAC
      -----------------MVKKAIVTAMAVISLFTLMGC
      -----------------MKQLIVNSVATVALASLVAGC
      --------------------MKLKTLALSLLAAGVLAGC
      --------------------MKAYLALISAAVIGLAAC
      --------------------MKLKATLTLAAATLVLAAC
      ------------------MQKTPKKLTALCHQQSTASC
      ----------------MPLPDFRLIRLLPLAALVLTAC
      ----------------MKNQVKKILGMSVVAAMVIVGC
      ----------------------MKKFLPLSISITVLAAC
      -----------------MKRLFLSFVALALLAGSIAAC
      --------------------MCGKILLILFFIMTLSAC
      ----------------------MSKRLLSLASLALLFGC
      --------------------MFKRRYVTLLPLFVLLAAC
      ------------------MKKIIKLSLLSLSIAGLASC
      -----------------MGRSKIVLGAVVLASALLAGC
      ------------------MKAKIVLGAVILASGLLAGC
      ------------------MNNVLKFSALALAAVLATGC
      --------------MKLTTHHLRTGAALLLAGILLAGC
      -------------MAYSVQKSRLAKVAGVSLVLLLAAC
      ------------MSAGSPKFTVRRIAALSLVSLWLAGC
      MDKGEGLRLAATLRQWTRLYGGCHLLLGAVVCSLLAAC
      --------------------MKPFLRWCFVATALTLAGC
      ------------------MNIATKLMASLVASVVLTAC
      ----------------MQNAKLMLTCLAFAGLAALAGC
      ----------------------MKKYLLGIGLILALIAC
      ----------------------MRLLIGFALALALIGC
      --------------MFVTSKKMTAAVLAITLAMSLSAC
```

```
----------------MNKNMAGILSAAAVLTMLAGC
----------------MHVSSLKVVLFGVCCLSLAAC
---------------MYKNGFFKNYLSLFLIFLVIAC
-----------------MNKFVKSLLVAGSVAALAAC
-----------------MKKTNMALALLVAFSVTGC
---------------MSLTHYSGLAAAVSMSLILTAC
-----------------MLRYTRNALVLGSLVLLSGC
--------------------MRNFILFPMMAVVLLSGC
--------------------MRKQWLGICIAAGMLAAC
-------------------MRYLATLLLSLAVLITAGC
-------------------MNMTKGALILSLSFLLAAC
---------------MNKKIFTLFLVVAASAIFAVSC
--------------------MVKRGRFALCLAVLLGAC
-----------------MKVKYALLSAGALQLLVVGC
-----------------MNNPLVNQAAMVLPVFLLSAC
---------------MNAHTLVYSGVALACAAMLGSC
---------------MKLKSLVFSLSALFLVLGFTGC
----------------MREKWVRAFAGVFCAMLLIGC
--------------MKHNVKLMAMTAVLSSVLVLSGC
--------------------MKLRLSALALGTTLLVGC
----------MRKRISAIINKLNISIIIMTVVLMIGC
-----------------MRKRISAIIMTLFMVLVSC
----------MRKRISAIINKLNISIMMMIVVLMIGC
```

Το αποτέλεσμα αυτό, τυπώνεται στην οθόνη, αλλά θα μπορούσαμε πολύ εύκολα να έχουμε τροποποιήσει το πρόγραμμα έτσι ώστε να γράφει σε αρχείο (με την εντολή open). Εναλλακτικά, μπορούμε να εκμεταλλευτούμε τις δυνατότητες του shell (ακόμα και των Windows) και να ανακατευθύνουμε το αποτέλεσμα σε ένα αρχείο της επιλογής μας. Για παράδειγμα, μπορούμε να εκτελέσουμε το πρόγραμμα με την εντολή:

```
perl program.pl input.file > output.file
```

Γενικότερα στο Linux (και στο UNIX) το shell επιτρέπει πιο σύνθετους τρόπους διασύνδεσης μεταξύ προγραμμάτων, τις λεγόμενες "σωληνώσεις" (pipes). Με τις σωληνώσεις, επιτρέπουμε σε μια διεργασία να επικοινωνεί με μια άλλη και το αποτέλεσμα της μίας να αποτελεί είσοδο στην επόμενη. Για παράδειγμα η εντολή:

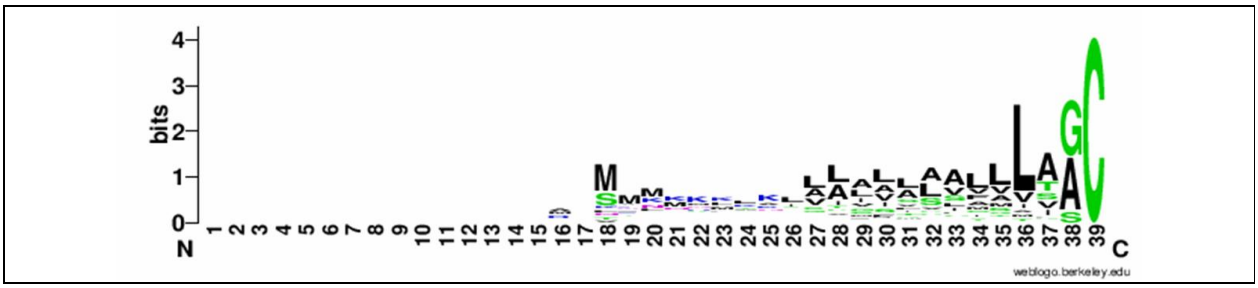
```
ls | wc
```

αποτελεί μια σωλήνωση που επιτρέπει στο αποτέλεσμα της ls να είναι είσοδος στην εντολή wc. Γενικά η Perl έχει πολλούς τρόπους να ελέγχει τις σωληνώσεις, αλλά ο πιο απλός είναι μέσω του shell. Για παράδειγμα μπορούμε να δώσουμε την εντολή:

```
./program.pl input.file | ./program2.pl
```

Με τον παραπάνω τρόπο, τα προγράμματα εκτελούνται σαν εντολές του συστήματος και το ένα δίνει το αποτέλεσμα στο άλλο. Κάτι που πρέπει να προσέξουμε σε αυτόν τον τρόπο, είναι ότι το δεύτερο πρόγραμμα θα πρέπει να επιτρέπει την εισαγωγή δεδομένων μέσα από το Standard Input (δηλαδή με χρήση του <STDIN>). Φυσικά, υπάρχουν και άλλοι τρόποι να χειριστούμε το θέμα, είτε μέσα από τη γλώσσα προγραμματισμού, όσο και από το ίδιο το περιβάλλον του shell αλλά η παρουσίαση τους ξεφεύγει από τους στόχους του παρόντος εισαγωγικού κειμένου.

Σε κάθε περίπτωση, το αποτέλεσμα του παραπάνω προγράμματος (η ιδιότυπη αυτή πολλαπλή στοίχιση), αν δοθεί σαν είσοδος στο WebLogo, θα δώσει το αποτέλεσμα που φαίνεται στην Εικόνα 12.1.



Εικόνα 12.1: *Το λογότυπο αλληλουχιών για την πολλαπλή στοίχιση των σηματοδοτικών αλληλουχιών των βακτηριακών λιποπρωτεϊνών όπως προέκυψε από το WebLogo (http://weblogo.berkeley.edu/)*

12.9. Περαιτέρω μελέτη

Όπως ήδη είπαμε, αυτό το κεφάλαιο προσφέρει μια απλή και όσο το δυνατό πιο κατανοητή εισαγωγή στην Perl, με ταυτόχρονη επίδειξη των ιδιοτήτων της γλώσσας στην επίλυση μερικών απλών προβλημάτων, τα οποία είναι εμπνευσμένα από πρακτικά προβλήματα της βιοπληροφορικής. Κάποιος που επιθυμεί να εμβαθύνει περισσότερο στην κατανόηση της γλώσσας θα πρέπει να ανατρέξει σε περισσότερο κατατοπιστικά και αναλυτικά βιβλία, όπως το *Programming Perl*, το οποίο θεωρείται το εγχειρίδιο αναφοράς της γλώσσας (Wall & Schwartz, 1991), ή το *Learning Perl* το οποίο αποτελεί μια πιο βατή εισαγωγή, καλύτερη για αρχάριους προγραμματιστές (Schwartz & Phoenix, 2001), το οποίο υπάρχει και στα Ελληνικά. Φυσικά, υπάρχουν πάρα πολλά εξειδικευμένα βιβλία, άλλα ειδικά εστιασμένα σε εφαρμογές βιοπληροφορικής (Moorhouse & Barry, 2005; Tisdall, 2001, 2003), άλλα σε αλγόριθμους (Orwant, Hietaniemi, & Macdonald, 1999) και άλλα και σε εφαρμογές διαδικτύου (Castro, 2001). Φυσικά, υπάρχουν άπειρα tutorials αλλά και online ebooks που διατίθενται δωρεάν όπως το *Picking Up Perl*, το οποίο είναι διαθέσιμο στη διεύθυνση http://www.ebb.org/PickingUpPerl/ (Kuhn, 2002), αλλά και πολλά ακόμα που μπορούν να βρεθούν στη διεύθυνση https://www.perl.org/books/library.html και http://www.perlmonks.org/index.pl/Tutorials.

Δεν πρέπει ακόμα να παραλείψουμε να κάνουμε αναφορά στο πρόγραμμα της **BioPerl** (http://www.bioperl.org/wiki/Main_Page), το οποίο είναι ένα μεγάλο έργο που σκοπό έχει να προσφέρει ειδικά προγράμματα για αναλύσεις βιοπληροφορικής σε μορφή βιβλιοθηκών (modules) της Perl. Η BioPerl είναι ένα τεράστιο έργο, που έχει χρησιμοποιηθεί από εκατοντάδες χρήστες και περιέχει εντολές για σχεδόν κάθε είδους ανάλυση (Stajich et al., 2002). Η ιδιαιτερότητα της, είναι ότι χρησιμοποιεί αντικειμενοστραφή προγραμματισμό (object-oriented Perl), κάτι που ίσως δυσκολέψει τον αρχάριο χρήστη. Μια εισαγωγή στην BioPerl, με κάποια απλά παραδείγματα όμοια με αυτά που αντιμετωπίσαμε σε αυτό το κεφάλαιο, υπάρχει στη διεύθυνση: http://www.bioperl.org/wiki/HOWTO:Beginners.

Βιβλιογραφία

Castro, Elizabeth. (2001). Perl and CGI for the world wide web: Visual quickstart guide: Peachpit Press.

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. Genome Res, 14(6), 1188-1190. doi: 10.1101/gr.84900414/6/1188 [pii]

Juncker, A. S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H., & Krogh, A. (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci, 12(8), 1652-1662.

Kuhn, Bradley M. (2002). Picking Up Perl: B. Kuhn.

Moorhouse, Michael, & Barry, Paul. (2005). Bioinformatics biocomputing and Perl: an introduction to bioinformatics computing skills and practice: John Wiley & Sons.

Orwant, Jon, Hietaniemi, Jarkko, & Macdonald, John. (1999). Mastering algorithms with Perl: " O'Reilly Media, Inc.".

Schwartz, Randal L, & Phoenix, Tom. (2001). Learning perl: O'Reilly & Associates, Inc.

Stajich, Jason E, Block, David, Boulez, Kris, Brenner, Steven E, Chervitz, Stephen A, Dagdigian, Chris, . . . Lapp, Hilmar. (2002). The Bioperl toolkit: Perl modules for the life sciences. Genome research, 12(10), 1611-1618.

Sutcliffe, I. C., & Harrington, D. J. (2002). Pattern searches for the identification of putative lipoprotein genes in Gram-positive bacterial genomes. Microbiology, 148(Pt 7), 2065-2077.

Tisdall, James. (2001). Beginning Perl for bioinformatics: " O'Reilly Media, Inc.".

Tisdall, James. (2003). Mastering Perl for bioinformatics: " O'Reilly Media, Inc.".

Wall, Larry, & Schwartz, Randal L. (1991). Programming perl: O'Reilly & Associates Sebastopol, CA.

Ασκήσεις

1) Τροποποιήστε το πρόγραμμα της προσομοίωσης έτσι ώστε να παράγει αμινοξέα ή νουκλεοτίδια σύμφωνα με κάποια προκαθορισμένη σύσταση (δηλαδή, να μην είναι ισοπίθανα). Για παράδειγμα, στην περίπτωση των πρωτεϊνών, μπορείτε να χρησιμοποιήσετε τη σύσταση ολόκληρης της Uniprot για να πάρετε πιο ρεαλιστικά αποτελέσματα (θεωρούμε εδώ ότι τα αμινοξέα δίνονται με τη σειρά που είδαμε ήδη, A, C, κ.ο.κ.):

```
@counts = ('0.077', '0.016', '0.053', '0.065', '0.041', '0.069', '0.023', '0.059',
'0.059', '0.095', '0.024', '0.043', '0.049', '0.039', '0.052', '0.070', '0.055',
'0.066', '0.012', '0.031');
```

2) Χρησιμοποιήστε το πρόγραμμα της προσομοίωσης τυχαίων αλληλουχιών DNA για να ελέγξετε τα αποτελέσματα του κεφαλαίου 3 που αφορούν τις ροές. Κατασκευάστε μια μεγάλη σειρά (π.χ. 1000) τυχαίων αλληλουχιών με γνωστό μήκος, και εντοπίστε τη μέγιστη ροή από A που εμφανίζεται σε κάθε μία από αυτές. Έπειτα, κατασκευάστε το ιστόγραμμα συχνοτήτων για τις 1000 αλληλουχίες και επιβεβαιώστε το νόμο των Erdos και Renyi και την κατανομή των ακραίων τιμών. Επαναλάβετε το πείραμα για διαφορετικά μήκη αλληλουχιών (π.χ. 1000, 5000, 10000 κ.ο.κ.).

3) Κατασκευάστε ένα πρόγραμμα που θα πραγματοποιεί πρόγνωση της θέσης μεμβρανικών τμημάτων σε διαμεμβρανικές πρωτεΐνες και ελέξγτε την αποτελεσματικότητα του σε μια ομάδα πρωτεϊνών από την Uniprot. Οι γνωστής δομής μεμβρανικές πρωτεΐνες, έχουν στο πεδίο FT το χαρακτηριστικό TRANSMEM το οποίο δείχνει την θέση των διαμεμβρανικών περιοχών. Στο πρώτο στάδιο, το πρόγραμμα θα πρέπει να διαβάζει την εγγραφή και θα πρέπει να τυπώνει μια ιδιαίτερη μορφή fasta oneline, στην οποία θα υπάρχει και σήμανση. Για παράδειγμα το αρχείο θα είναι κάπως έτσι:

```
ID   140U_DROME              Reviewed;         261 AA.
AC   P81928; Q9VFM8;
FT   CHAIN         1    261       RPII140-upstream gene protein.
FT                                /FTId=PRO_0000064352.
FT   TRANSMEM     67     87       Potential.
FT   TRANSMEM    131    151       Potential.
FT   TRANSMEM    183    203       Potential.
FT   CONFLICT     64     64       S -> F (in Ref. 1).
SQ   SEQUENCE   261 AA;  29182 MW;  5DB78CF6CFC4435A CRC64;
     MNFLWKGRRF LIAGILPTFE GAADEIVDKE NKTYKAFLAS KPPEETGLER LKQMFTIDEF
     GSISSELNSV YQAGFLGFLI GAIYGGVTQS RVAYMNFMEN NQATAFKSHF DAKKKLQDQF
     TVNFAKGGFK WGWRVGLFTT SYFGIITCMS VYRGKSSIYE YLAAGSITGS LYKVSLGLRG
     MAAGGIIGGF LGGVAGVTSL LLMKASGTSM EEVRYWQYKW RLDRDENIQQ AFKKLTEDEN
     PELFKAHDEK TSEHVSLDTI K
```

Το πρόγραμμα θα πρέπει να μετατρέπει τις αλληλουχίες στη μορφή:

```
>P81928
MNFLWKGRRFLIAGILPTFEGAADEIVDKENKTYKAFLASKPPEETGLERLKQMFTIDEFGSISSELNSVYQAGFLGFL
------------------------------------------------------------------MMMMMMMMMMMM
```

Για την πρόγνωση, θα στηριχθείτε στη μέθοδο των παραθύρων, με χρήση κάποιας κλίμακας υδροφοβικότητας, όπως των Kyte-Doolitle που δίνεται παρακάτω:

```
%hyd =('A' => 0.100,
       'C' => -1.420,
       'D' => 0.780,
       'E' => 0.830,
       'F' => -2.120,
       'G' => 0.330,
       'H' => -0.500,
       'I' => -1.130,
       'K' => 1.400,
       'L' => -1.180,
```

```
            'M' => -1.590,
            'N' => 0.480,
            'P' => 0.730,
            'Q' => 0.950,
            'R' => 1.910,
            'S' => 0.520,
            'T' => 0.070,
            'V' => -1.270,
            'W' => -0.510,
            'Y' => -0.210
            );
```

Θα πρέπει να πειραματιστείτε με τις πιθανές τιμές του παραθύρου που δίνουν το καλύτερο αποτέλεσμα, αλλά και να προσπαθήσετε να αναπαραστήσετε τα αποτελέσματα γραφικά (με χρήση χαρακτήρων όπως το *). Μπορείτε επίσης να χρησιμοποιήσετε αντί της κλίμακας υδροφοβικότητας, και έναν εμπειρικό πίνακα του σκορ, με τιμές από συχνότητες αμινοξέων, όπως λ.χ. ο Πίνακας 3.1 στο κεφάλαιο 3, και να συγκρίνετε τα αποτελέσματα.

4) Κατασκευάστε ένα πρόγραμμα που θα παράγει τυχαίες αλληλουχίες DNA και θα ελέγχει το μήκος των πρωτεϊνών που παράγονται από τα τυχαία ανοιχτά πλαίσια ανάγνωσης που θα εντοπιστούν. Θα πρέπει να συνδυάσετε το πρόγραμμα της προσομοίωσης και το πρόγραμμα εύρεσης γονιδίων. Θα πρέπει να δημιουργήσετε ένα μεγάλο αριθμό τυχαίων γονιδιωμάτων (π.χ. 1000), με μεγάλο μήκος όμως (>10.000), και να αποθηκεύσετε το μήκος των παραγόμενων πρωτεϊνών για να γίνει ένα ιστόγραμμα που θα μας δείξει την κατανομή τους. Θα πρέπει να πειραματιστείτε τόσο με το μήκος του γονιδιώματος, όσο και με τη σύσταση σε νουκλεοτίδια (βλ. και άσκηση 1).

5) Κατασκευάστε ένα πρόγραμμα το οποίο θα διαβάζει τις αλληλουχίες πρωτεϊνών από ένα αρχείο FASTA (με μία γραμμή), και θα τις κωδικοποιεί σε παράθυρα με μήκος που θα ορίζει ο χρήστης, χρησιμοποιώντας το sparse encoding.

Ευρετήριο Όρων