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.

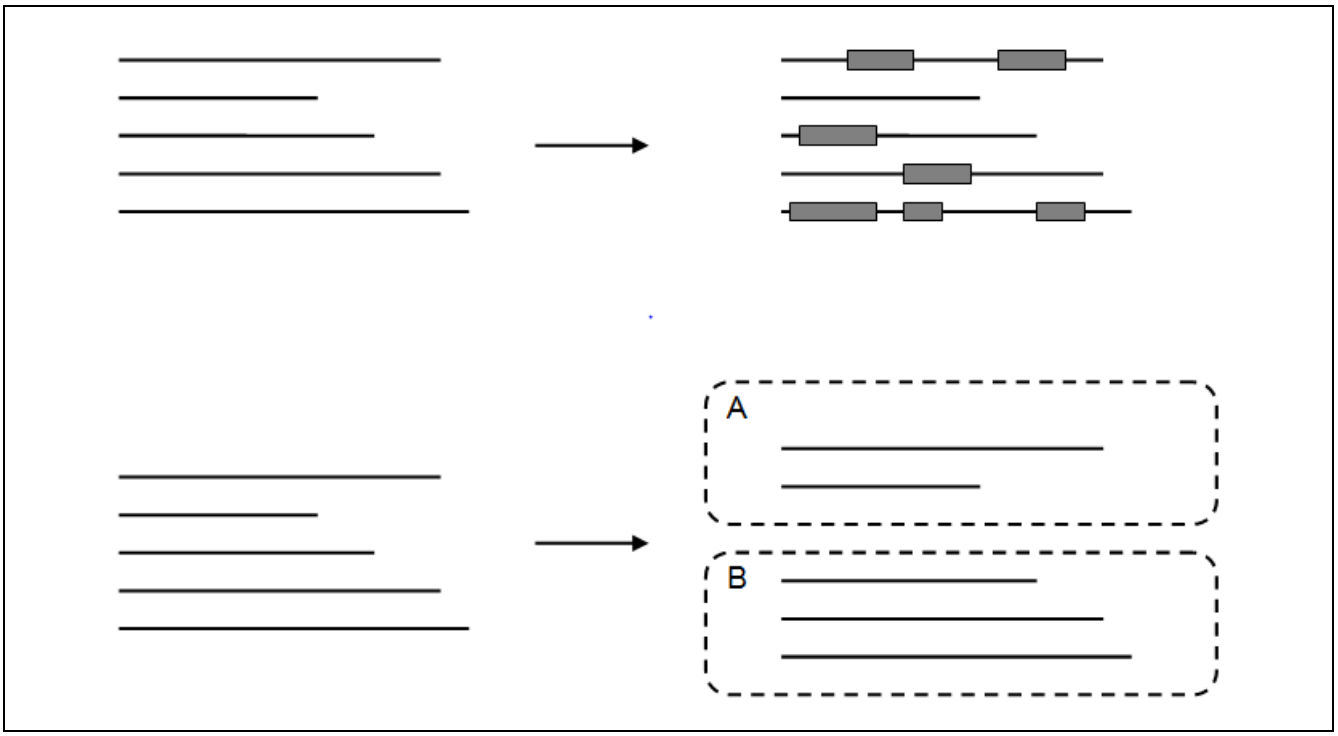
Στο κεφάλαιο αυτό, θα αναλύσουμε τις βασικές μεθοδολογίες που χρησιμοποιούνται στην πρόγνωση λειτουργικών και δομικών χαρακτηριστικών με χρήση αλληλουχιών. Θα αναλύσουμε τις βασικές κατηγορίες τέτοιων μεθόδων, θα δούμε πώς κατασκευάζεται και πώς αξιολογείται μια τέτοια μέθοδος ενώ στο τέλος θα δούμε 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.

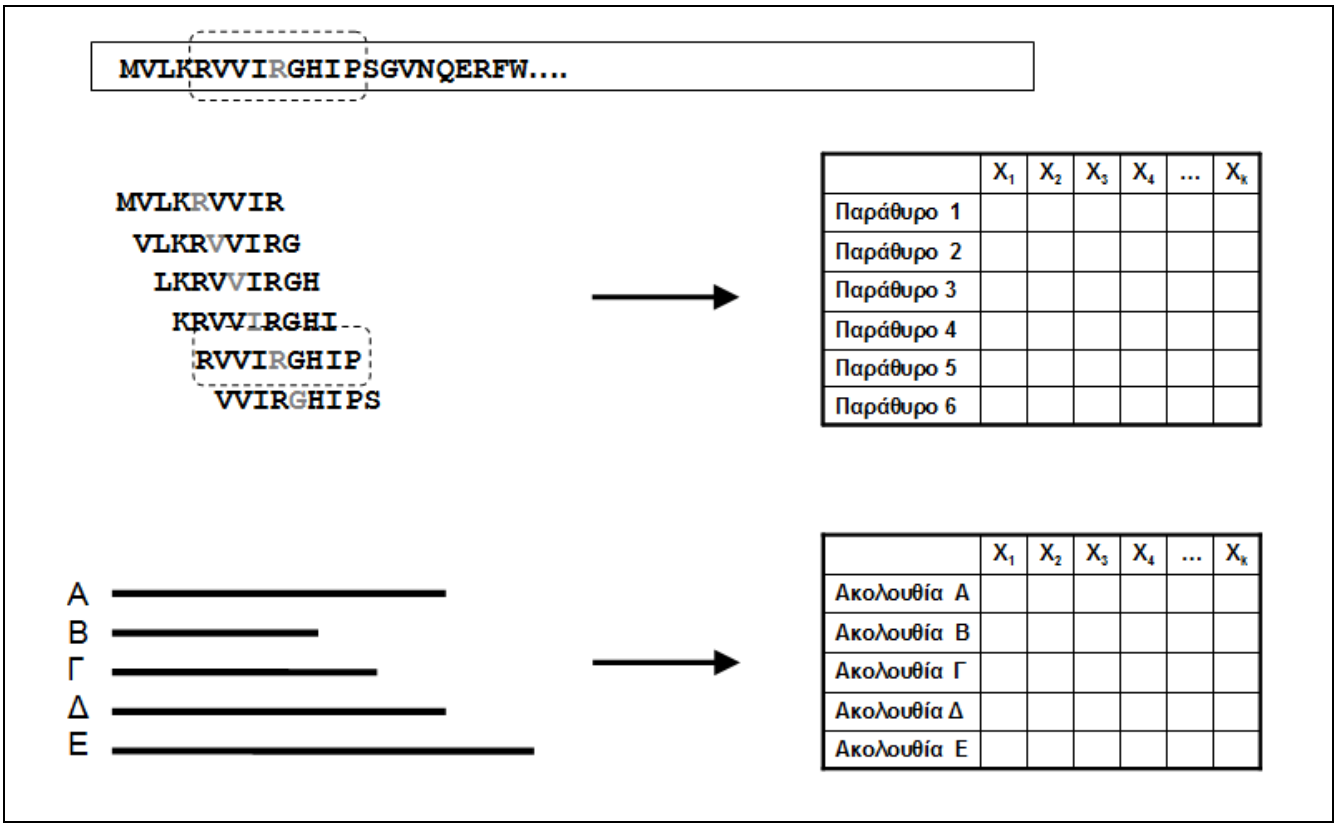


Εικόνα 7.1: *Παραδείγματα μεθόδων πρόγνωσης. Πάνω, δίνεται ένα υποθετικό παράδειγμα πρόγνωσης κάποιου χαρακτηριστικού κατά μήκος της αλληλουχίας. Κάτω, δίνεται ένα υποθετικό παράδειγμα διαχωρισμού των αλληλουχιών σε ομάδες.*

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.*

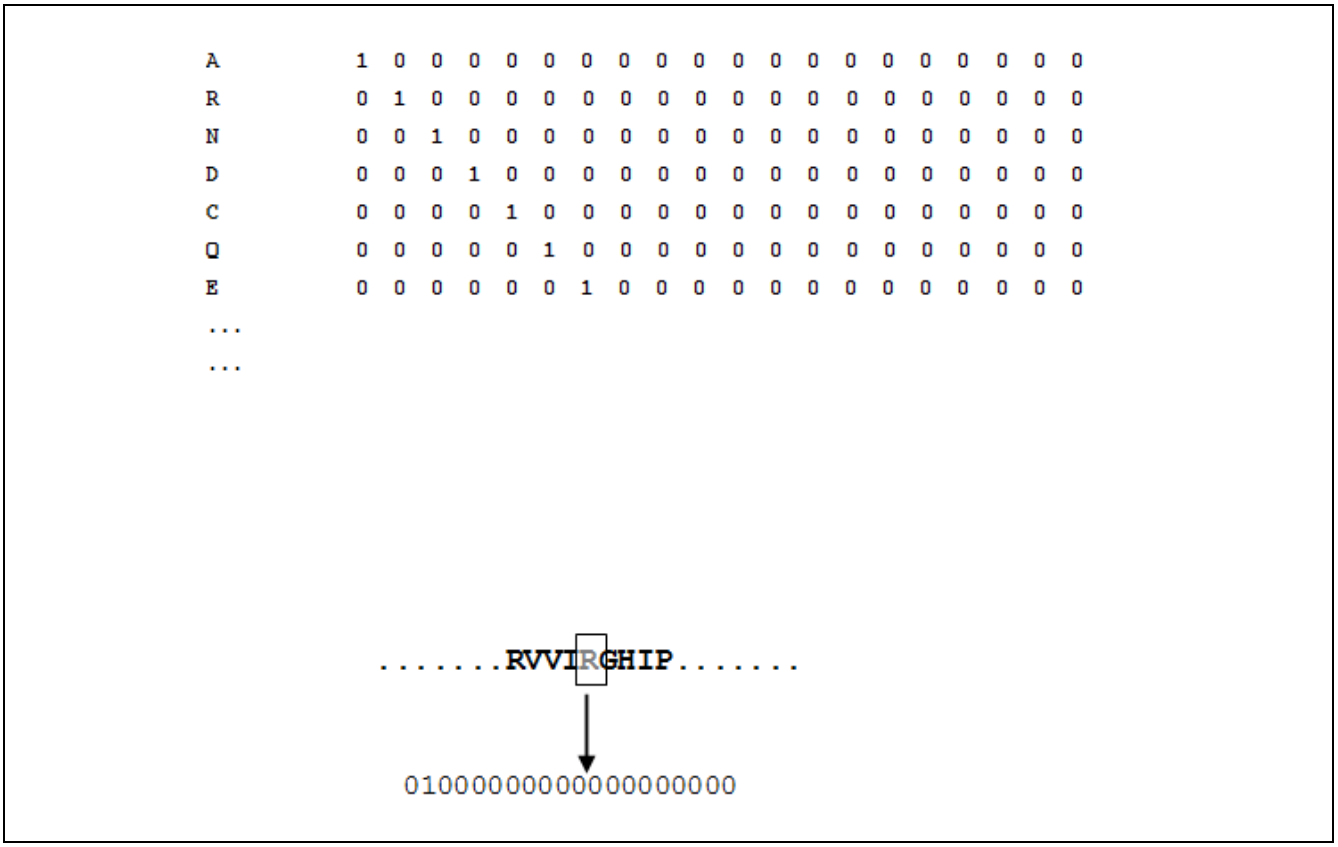
Φυσικά, υπάρχουν πολλά σημεία που απαιτούν διευκρινίσεις και μπορεί να διαφέρουν από μέθοδο σε μέθοδο. Ένα πρώτο θέμα έχει να κάνει με το μήκος του παραθύρου, και εξαρτάται πολύ από το συγκεκριμένο πρόβλημα. Στα περισσότερα προβλήματα πρόβλεψης δομής (δευτεροταγής δομή, διαμεβρανικές έλικες, προσβασιμότητα 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

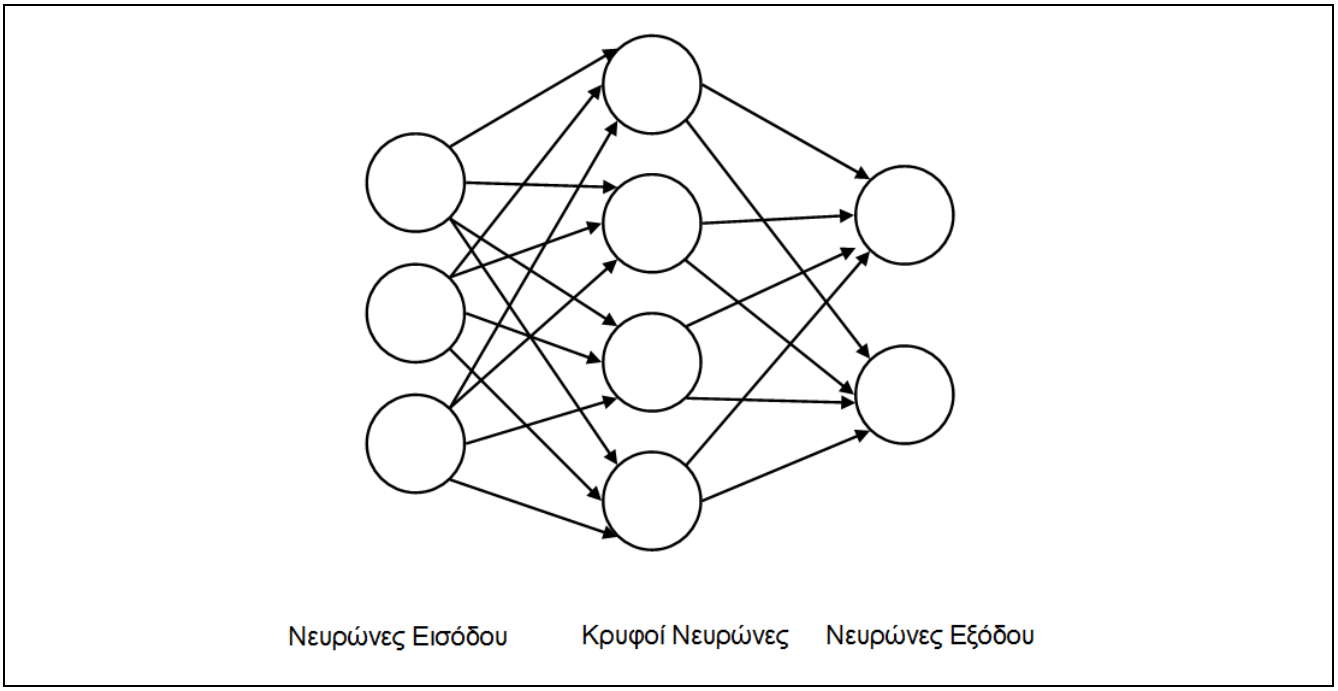
υδροφοβικότητα, η ύπαρξη άλλων χαρακτηριστικών όπως τα διαμεμβρανικά τμήματα, τα πεπτίδια οδηγητές, διάφορα πρότυπα που εμφανίζονται κ.ο.κ. (τα οποία βέβαια με τη σειρά τους προέρχονται από μεθόδους πρόγνωσης!). Μια άλλη εναλλακτική είναι η χρησιμοποίηση μαθηματικών τεχνικών που περιγράφουν την περιοδικότητα 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 (Pierre 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.

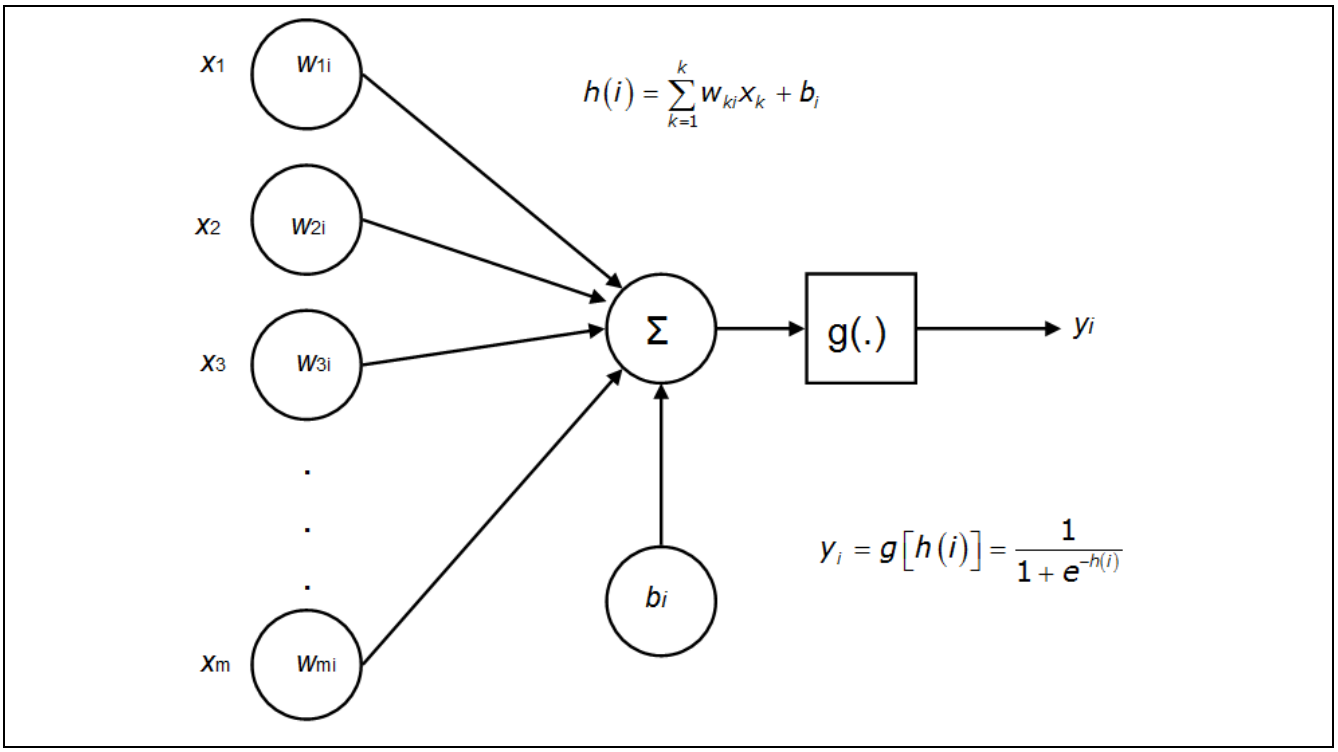
Το είδος της συνάρτησης ενεργοποίησης είναι επίσης κάτι καθοριστικό για τη δομή και τις ιδιότητες του δικτύου (προφανώς, συναρτήσεις ενεργοποίησης έχουν νόημα για τους νευρώνες εξόδου και του κρυφούς νευρώνες). Έτσι, αν το τελικό αποτέλεσμα που θέλουμε να προβλέψουμε είναι δίτιμο (ανήκει σε μια ομάδα/δεν 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:

$$g\left[h(i)\right] = \frac{1}{1+e^{-h(i)}} - \frac{1}{2}$$

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

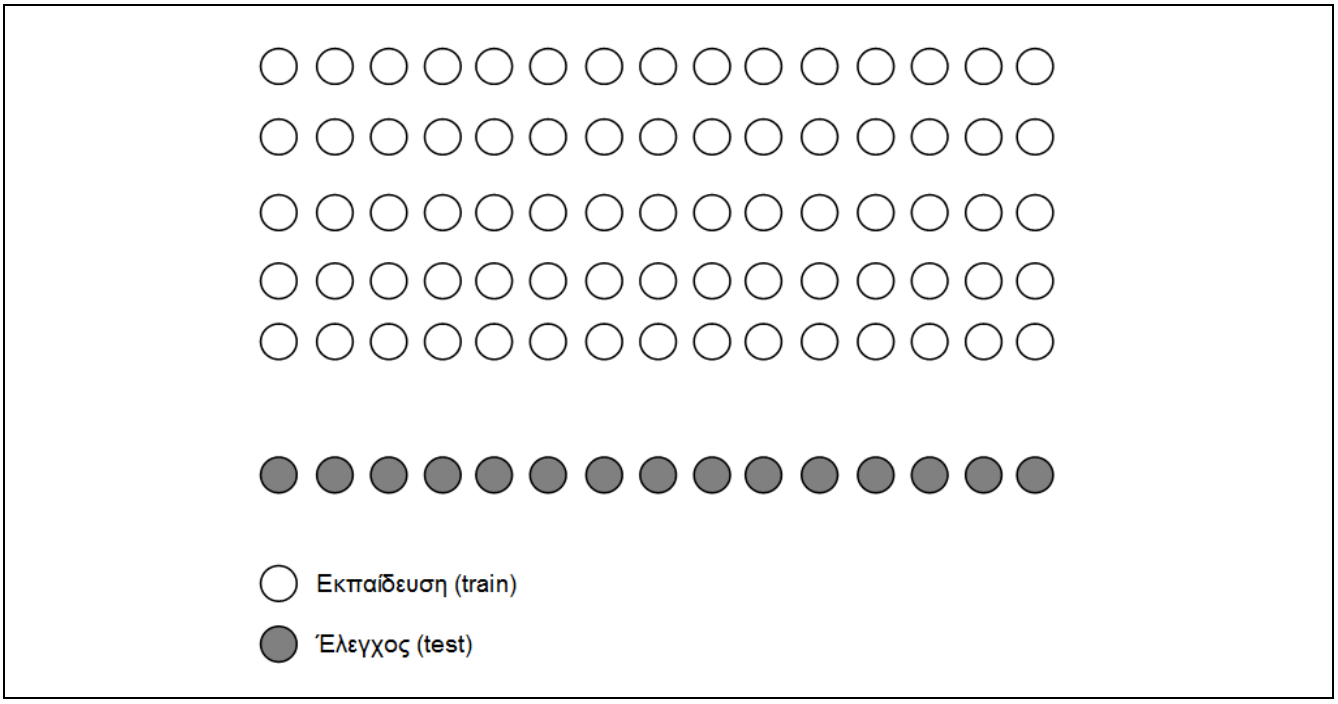
Η μέθοδος αυτή, φυσικά απαιτεί διάφορους υπολογισμούς που παραλείπονται εδώ, αλλά πρέπει να τονιστεί ότι σαν μέθοδος gradient descent, είναι μια ευριστική μέθοδος. Αναμένουμε, αν όλα πάνε καλά, ότι το σφάλμα θα μειώνεται συνεχώς, αλλά δεν υπάρχει μαθηματική εγγύηση. Έχουν αναπτυχθεί επίσης πάρα πολλές 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.

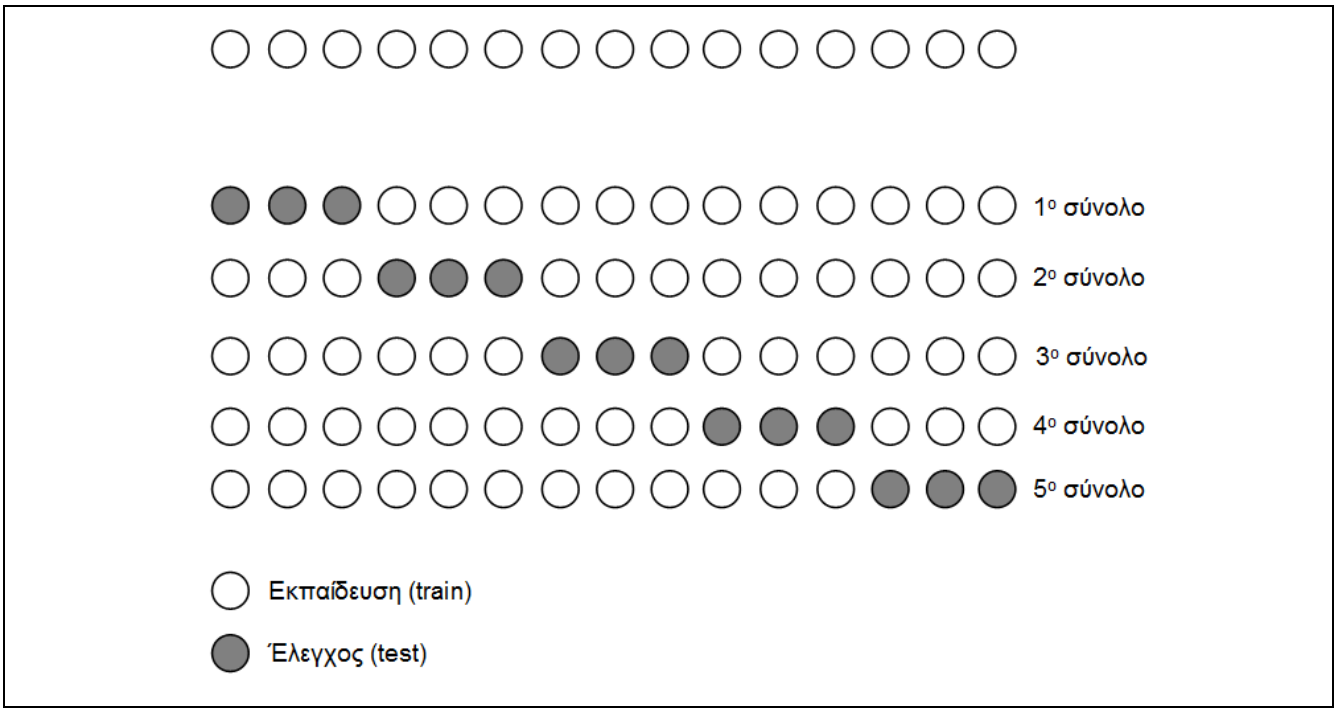


Εικόνα 7.6: *Ένα υποθετικό παράδειγμα με το σύνολο εκπαίδευσης και το ανεξάρτητο σύνολο ελέγχου.*

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).

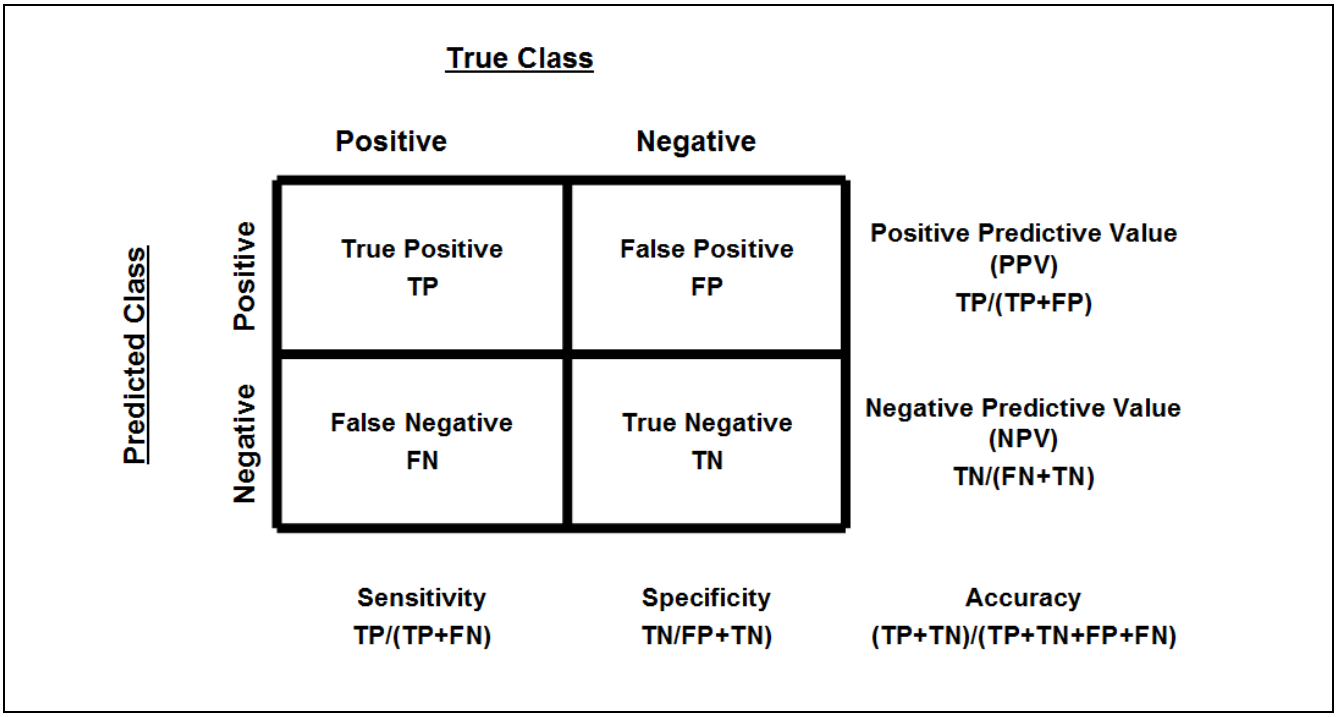
7.4. Μέτρα εκτίμησης της αξιοπιστίας των μεθόδων

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)*

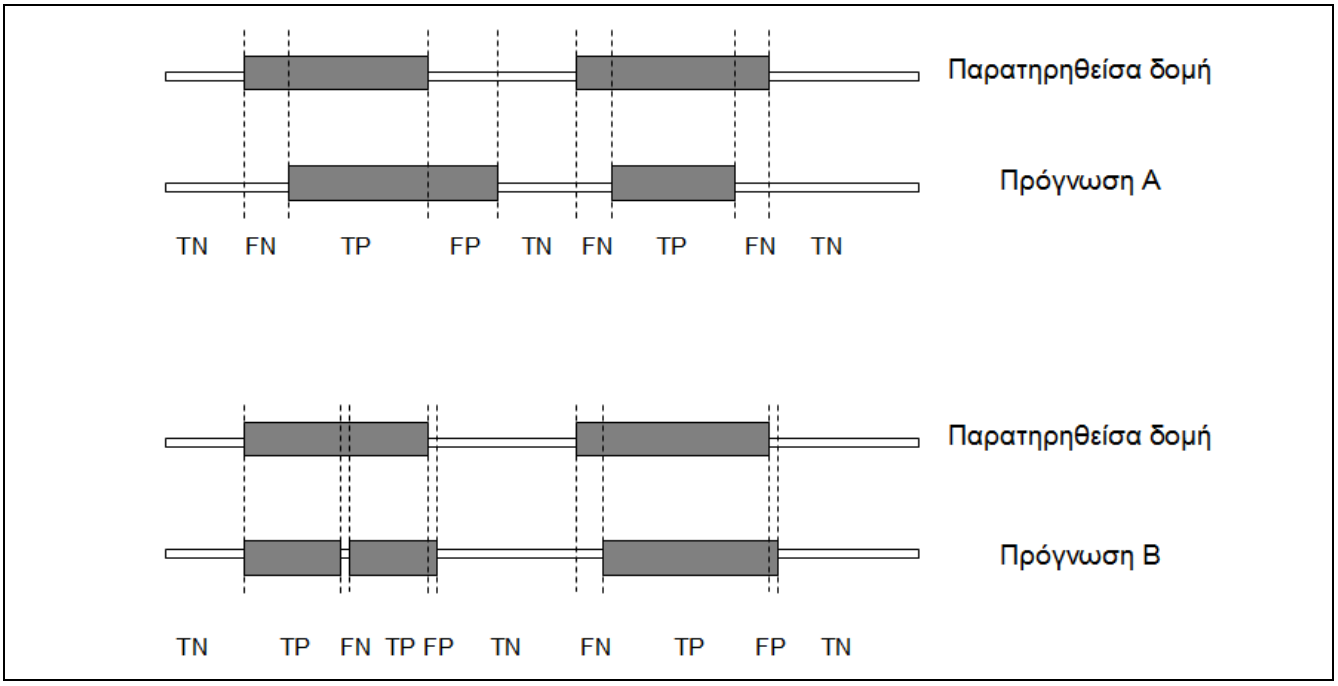
Επίσης, ένα άλλο μέτρο που χρησιμοποιείται είναι ο γνωστός συντελεστής συσχέτισης του Matthews (*C*) (Baldi, Brunak, Chauvin, Andersen, & Nielsen, 2000):

$$C = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

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

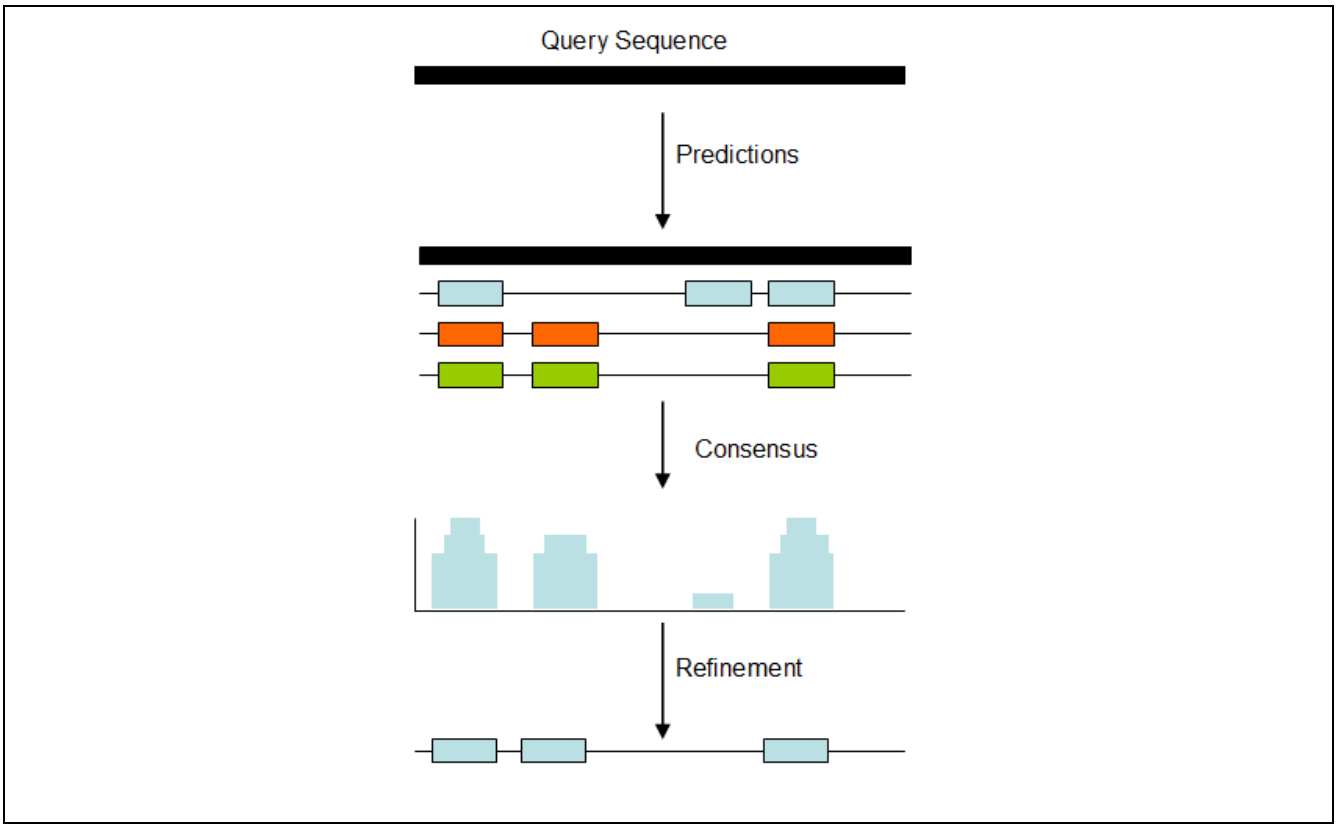
Γενικά, η επιτυχία μιας μεθόδου πρόγνωσης για ένα συγκεκριμένο πάντα πρόβλημα εξαρτάται από το μέγεθος και την ποιότητα του συνόλου εκπαίδευσης και από την επιλογή του αλγορίθμου, δηλαδή της μεθοδολογίας. Το μέγεθος του συνόλου εκπαίδευσης παίζει σίγουρα ένα ρόλο, αλλά η επίδραση δεν είναι γραμμική όπως 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. Συνδυστικές 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.

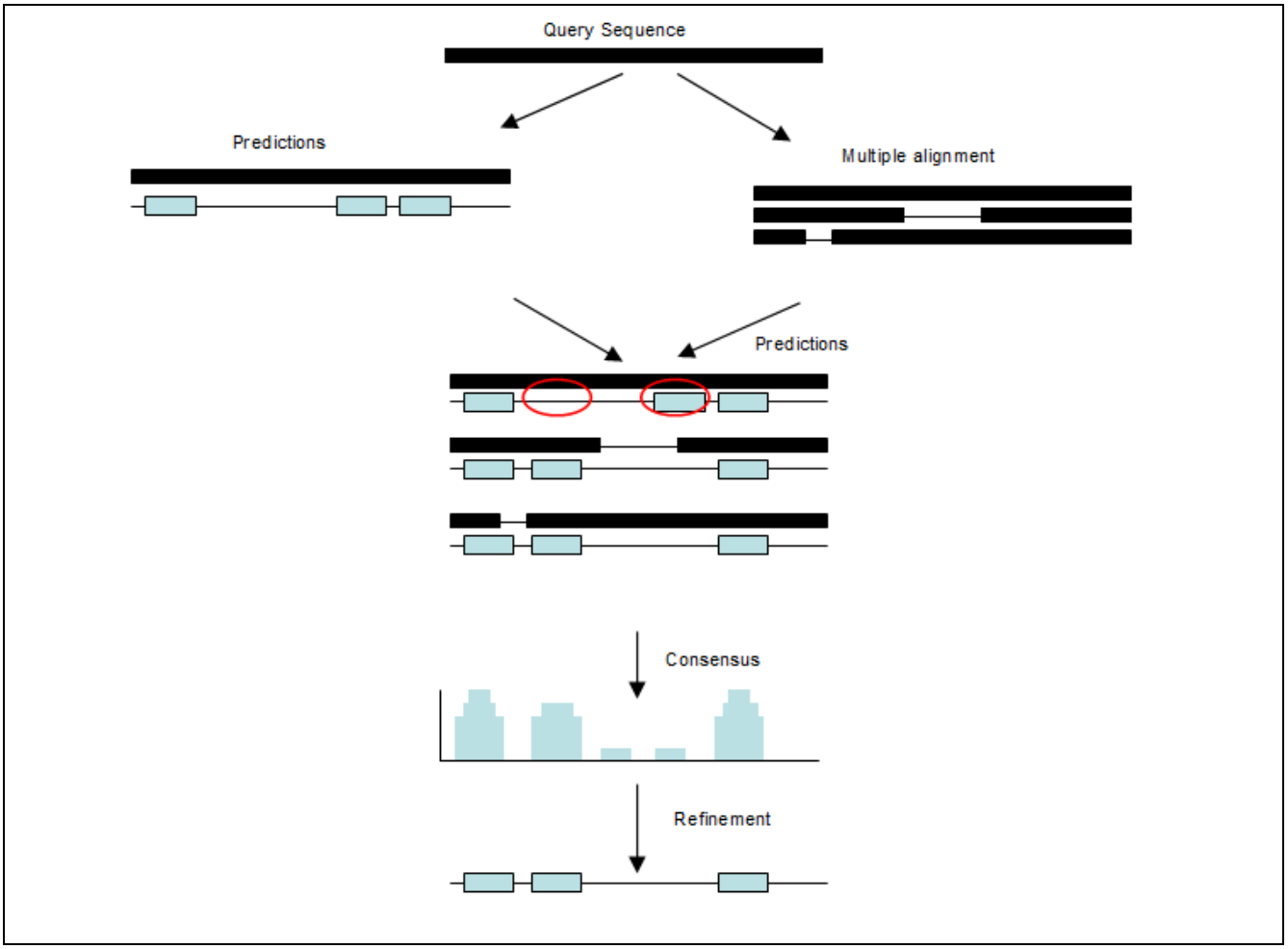
7.5.2. Ενσωμάτωση εξελικτικής πληροφορίας με τη μορφή πολλαπλών στοιχίσεων

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.*

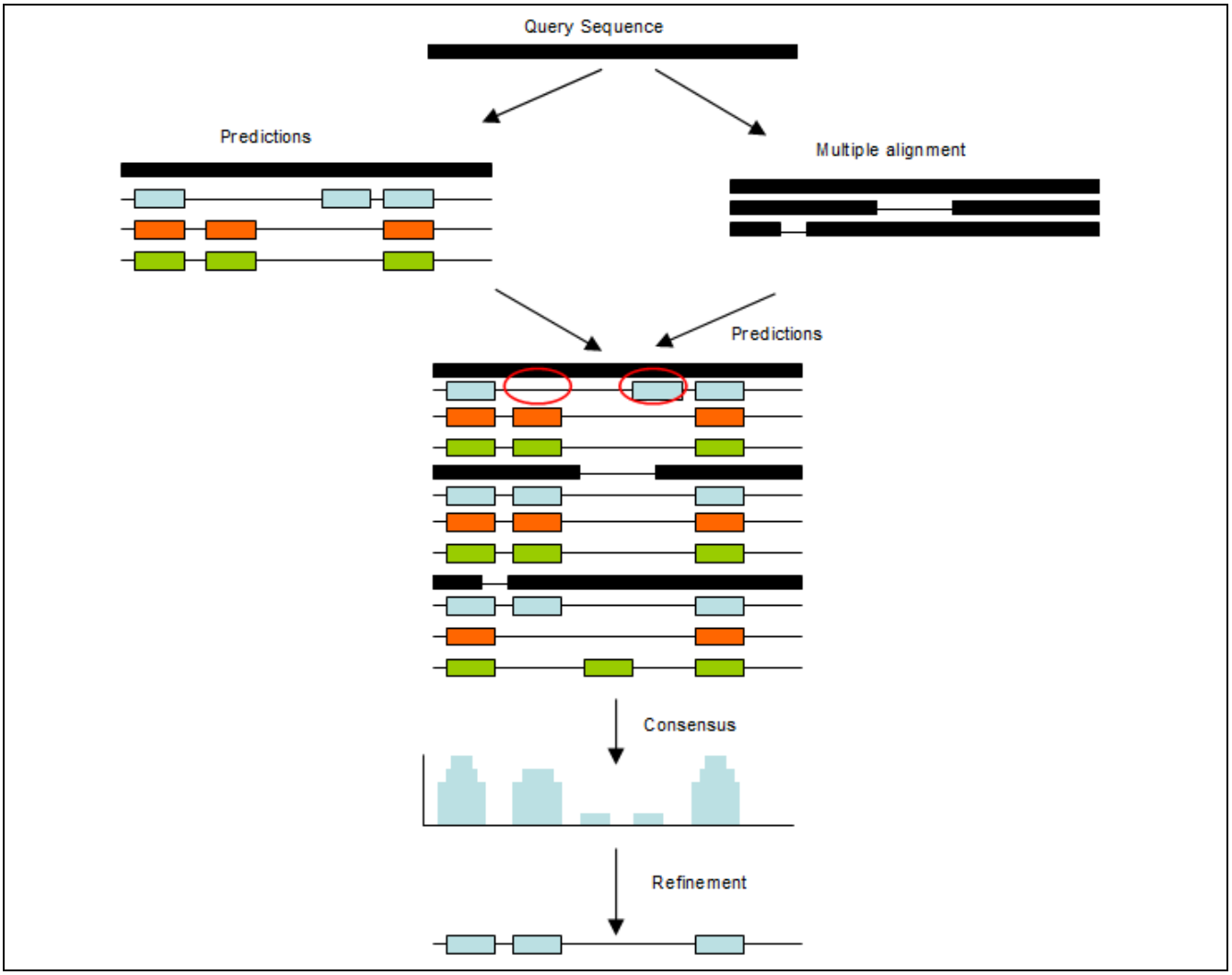
Μια πολύ ενδιαφέρουσα παραλλαγή της μεθόδου αυτής, προέκυψε όταν δημιουργήθηκε το γνωστό PSI-BLAST. Το πρόγραμμα αυτό εντοπίζει, με μια επαναληπτική διαδικασία ομόλογες αλληλουχίες και κατασκευάζει μια ειδικού τύπου πολλαπλή στοίχιση στην οποία δεν περιέχονται κενά στην αλληλουχία επερώτησης, από την 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.

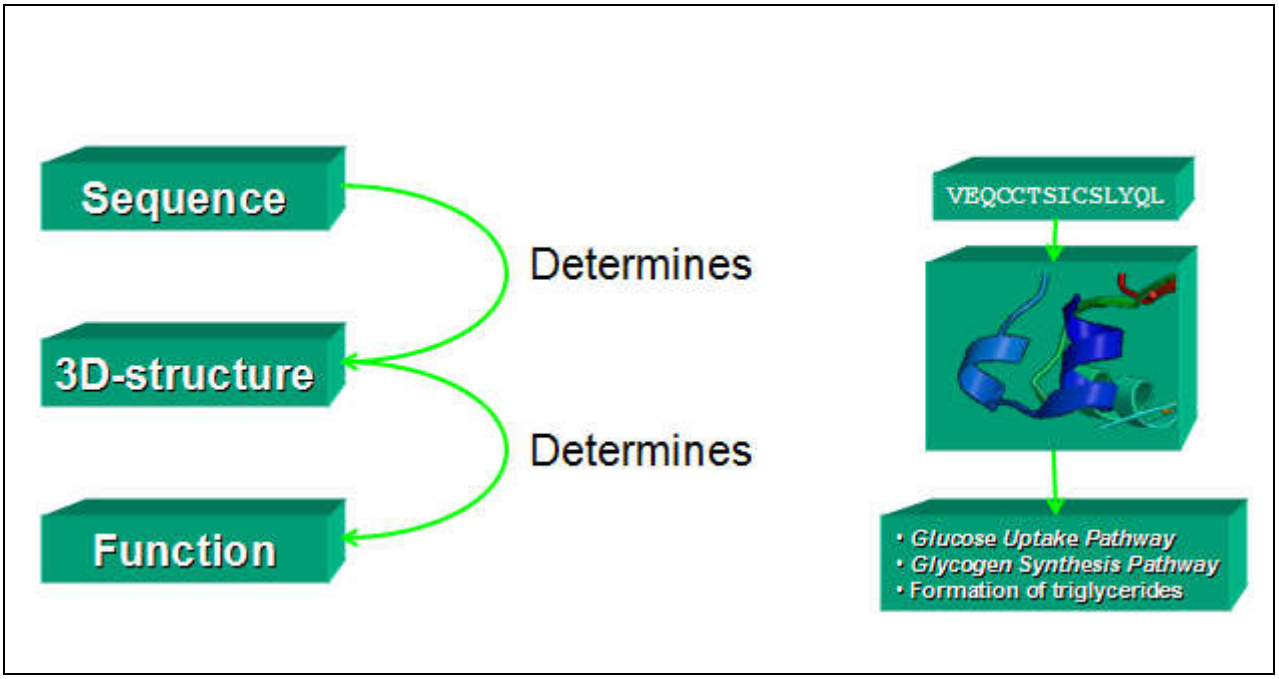


Εικόνα 7.13: *Ένα υποθετικό παράδειγμα συνδυασμού τόσο των συναινετικών μεθόδων αλλά και της χρήσης εξελικτικής πληροφορίας.*

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

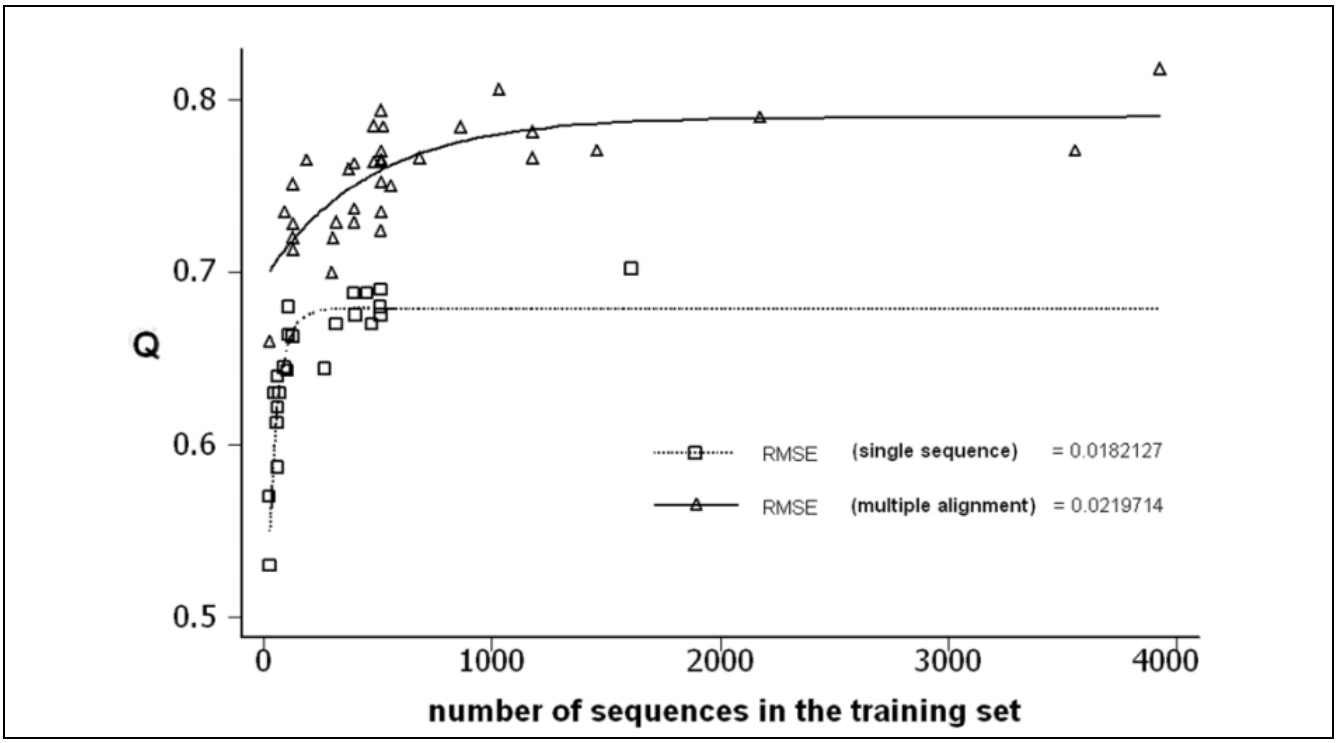
Πίνακας 7.1: *Οι τιμές των παραμέτρων(P) όπως υπολογίστηκαν από τους Chou και Fasman*

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.

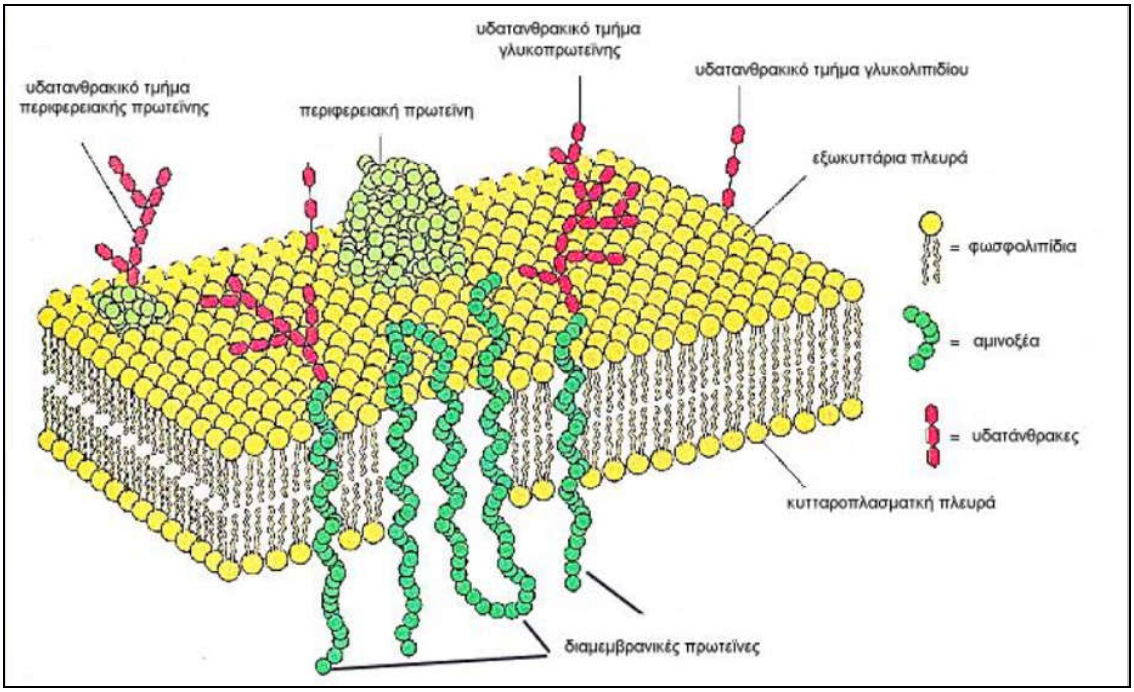
Οι συνδυαστικές/συναινετικές προγνώσεις είχαν επίσης δείξει από παλιά ότι μπορούν να αυξήσουν την επιτυχία των μεθόδων πρόγνωσης. Μια από τις πρώτες προσπάθειες είχε γίνει το 1988 όταν ο Hamodrakas (Hamodrakas, 1988) δημοσίευσε ένα συνδυαστικό αλγόριθμο που έκανε χρήση των τότε διαθέσιμων μεθόδων (Chou-Fasman, 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).



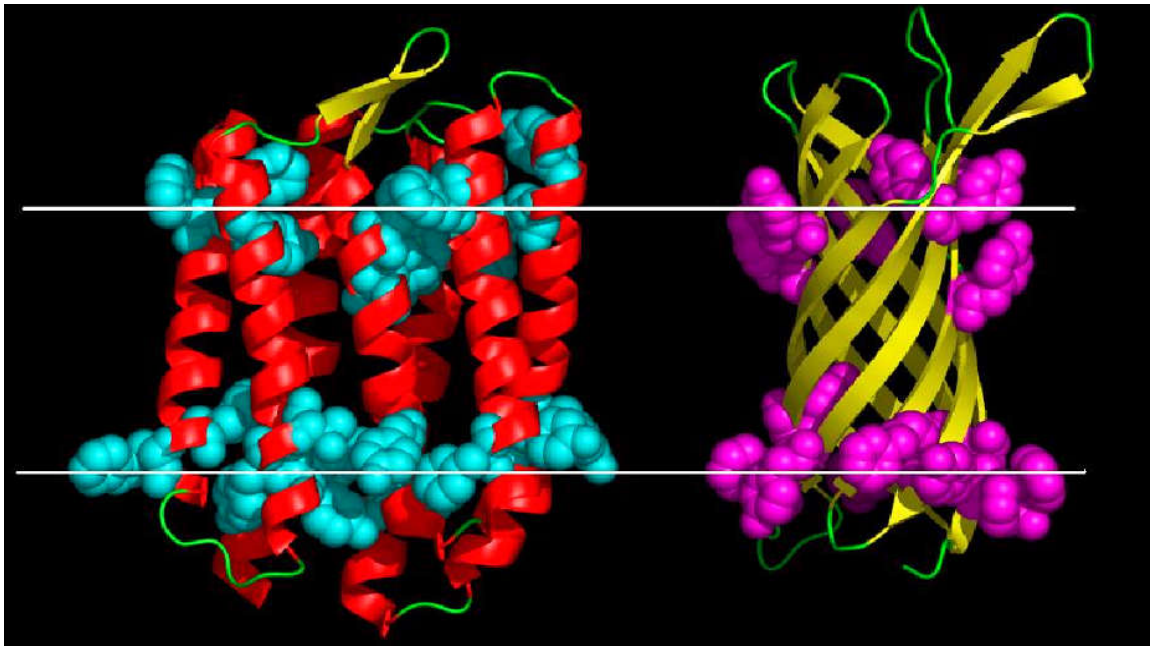
Εικόνα 7.16: *Απεικόνιση μιας τυπικής λιπιδικής διπλοστιβάδας, στην οποία φαίνονται οι διαμεμβρανικές και οι περιφερειακές πρωτεΐνες.*

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.

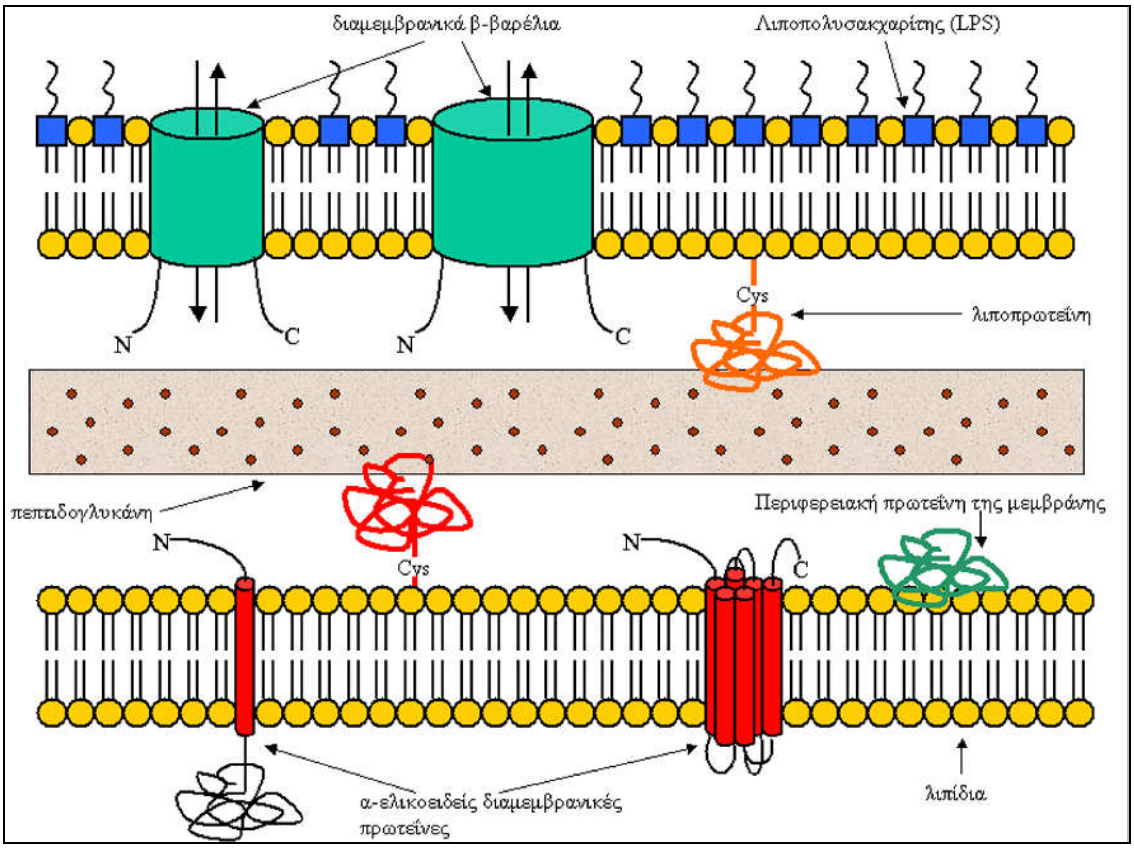
Το γενικότερο σχήμα για την λιπο-πρωτεϊνική φύση των βιολογικών μεμβρανών, που είδαμε παραπάνω, ισχύει, με επιμέρους κατά περίπτωση τροποποιήσεις, και στις μεμβράνες των οργανιδίων που απαντώνται στο εσωτερικό των ευκαρυωτικών κυττάρων (μιτοχόνδρια, χλωροπλάστες, λυσοσώματα, Golgi κλπ). Πολλά δε είδη 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.



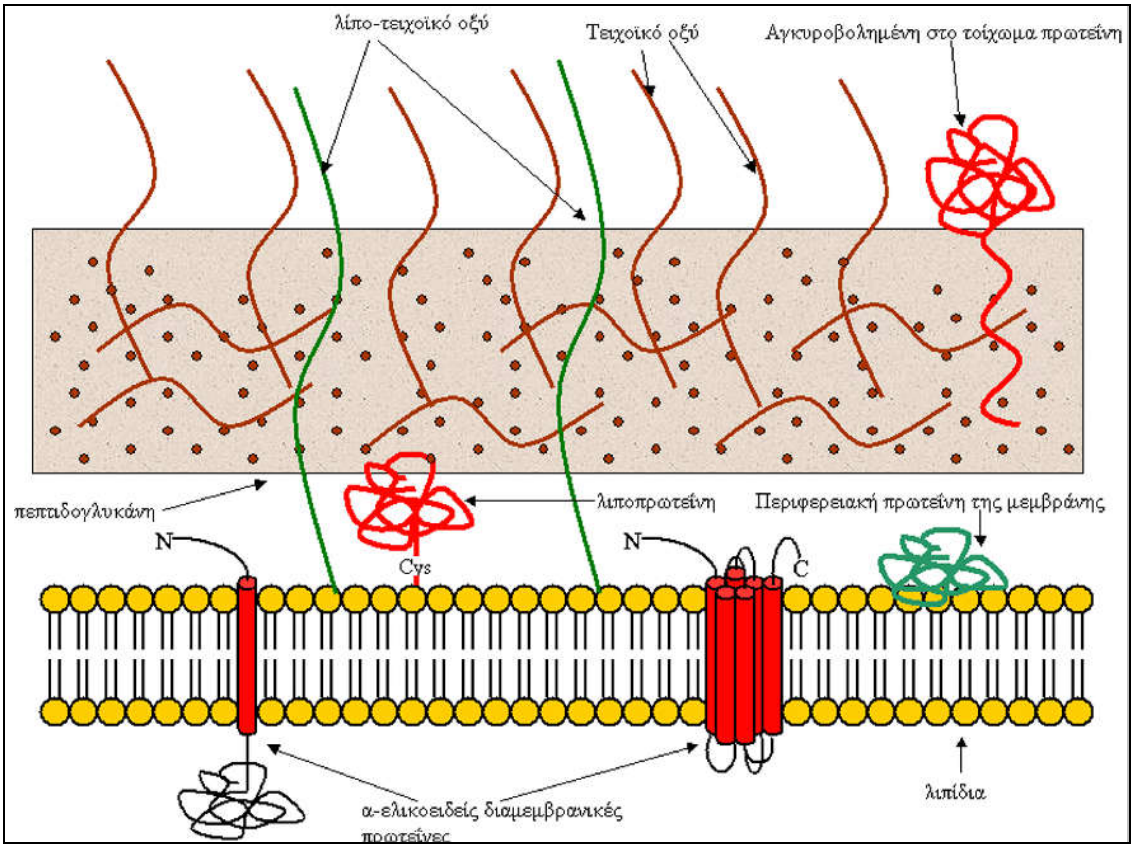
Εικόνα 7.17: *Οι δύο κατηγορίες διαμεμβρανικών πρωτεϊνών. Αριστερά, ο φωτοϋποδοχέας του αρχαιοβακτηρίου Natronomonas pharaonis. Δεξιά, η NspA, του βακτηρίου Neisseria meningitidis. Με τη χωροπληρωτική αναπαράσταση, διακρίνονται τα αρωματικά κατάλοιπα που συνιστούν την αρωματική ζώνη στα όρια της μεμβράνης.*

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).

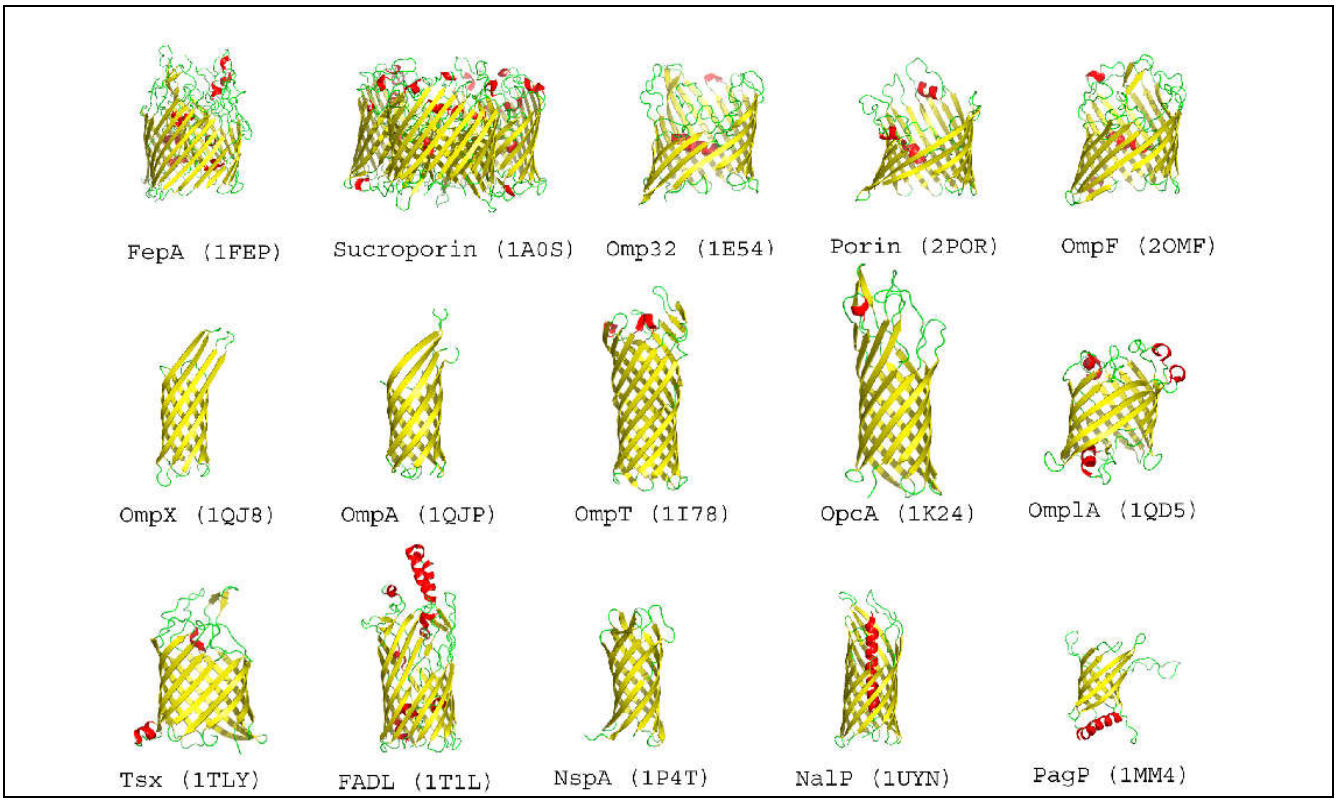
Μια άλλη μεγάλη πρόοδος που έγινε στην περίπτωση πρόγνωσης των διαμεμβρανικών α-ελίκων, έχει να κάνει με την ενίσχυση της απόδοσης της πρόγνωσης όταν γίνει ενσωμάτωση πειραματικής πληροφορίας. Στην περίπτωση διαμεμβρανικών πρωτεϊνών, είναι γνωστό ότι η ενσωμάτωση μιας, ακόμα και περιορισμένης πειραματικά, 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.

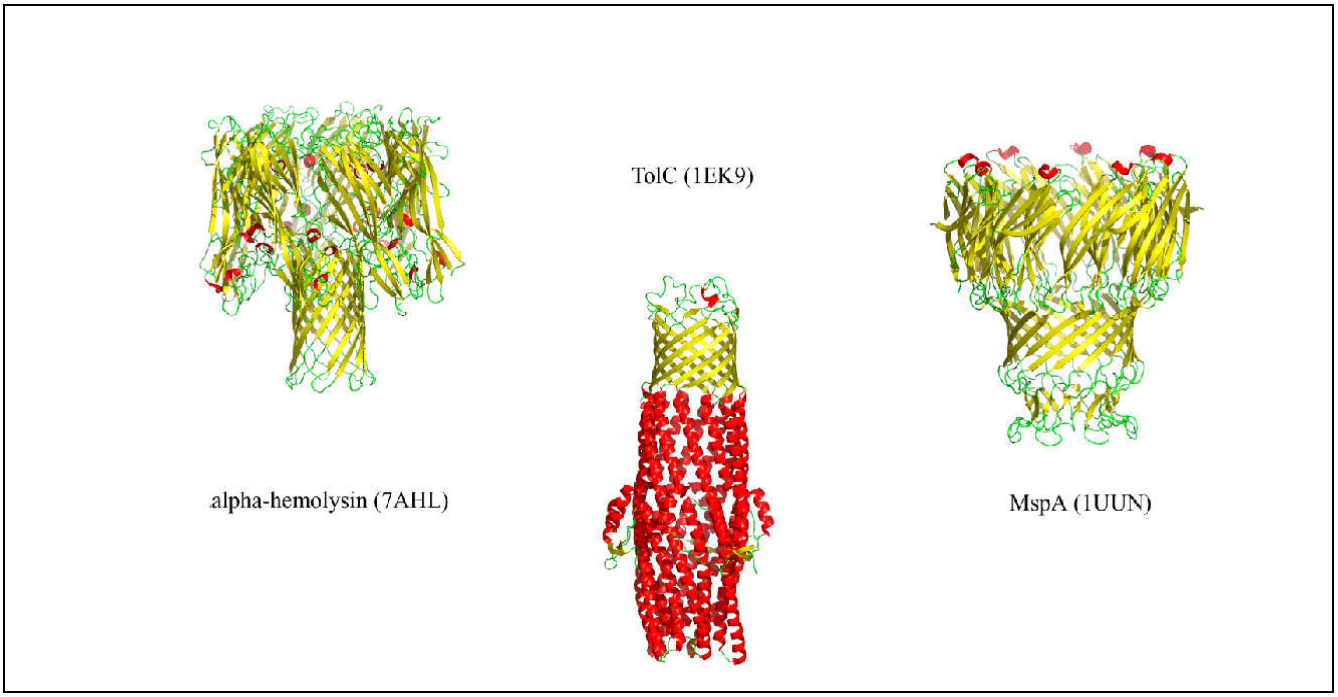
Το β-βαρέλι γενικά, είναι μια β-πτυχωτή επιφάνεια (πτυχωτό φύλλο) που περιελίσσεται και αναδιπλώνεται σχηματίζοντας μια κλειστή δομή σε σχήμα βαρελιού, η οποία σταθεροποιείται από δεσμούς υδρογόνου που σχηματίζονται από την κύρια αλυσίδα. Τα γνωστά παραδείγματα διαμεμβρανικών β-βαρελιών δείχνουν προτίμηση 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.

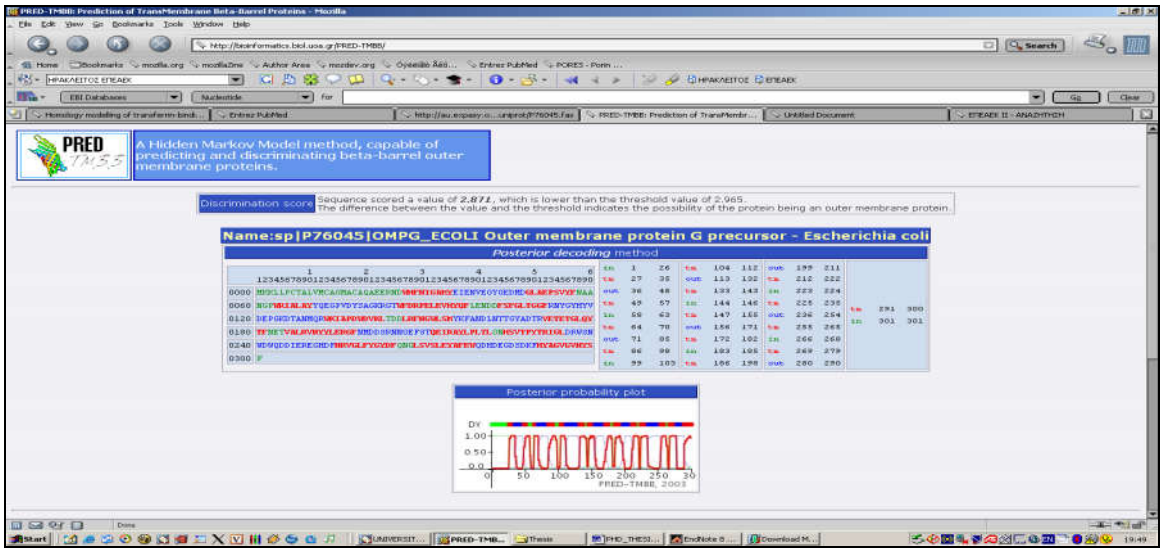
(6) Οι διαμεμβρανικές πρωτεΐνες με μορφή β-βαρελιού εμφανίζουν μικρότερη συντηρητικότητα στις ακολουθίες τους, σε σχέση με τις σφαιρικές-υδατοδιαλυτές πρωτεΐνες. Ακόμα μικρότερη είναι η συντηρητικότητα στις εξωκυττάριες στροφές, οι οποίες δρουν συχνά σαν αντιγονικοί καθοριστές. Το γεγονός αυτό συνεπάγεται, 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.

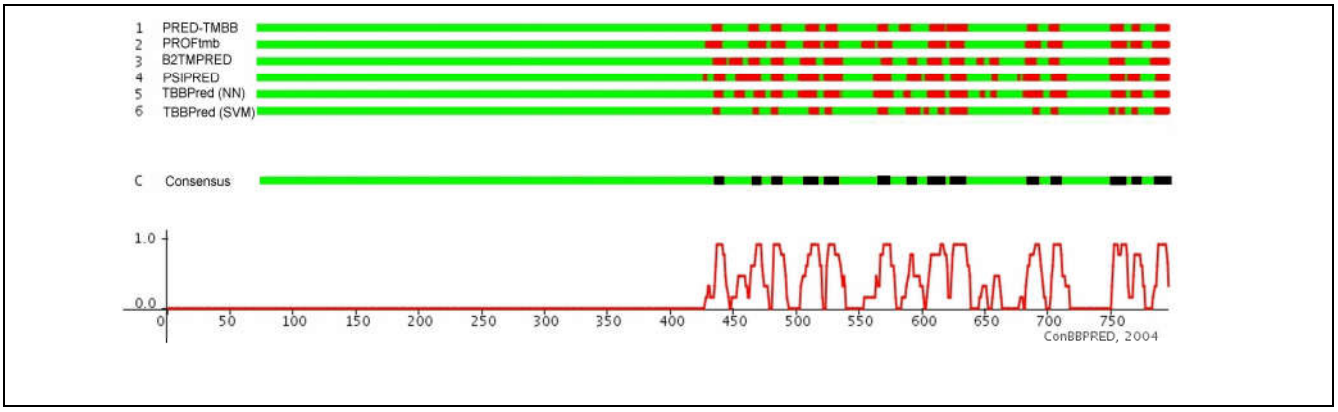
Οι πρώτες μέθοδοι βασισμένες σε Hidden Markov Model (HMM) εμφανίστηκαν επίσης στις αρχές της δεκαετίας του 2000, και από τότε η μεθοδολογία αυτή έχει κυριαρχήσει (Bagos, Liakopoulos, Spyropoulos, & Hamodrakas, 2004a, 2004b; Bigelow, Petrey, Liu, Przybylski, & Rost, 2004; Hayat & Elofsson, 2012; Liu, 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.

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.

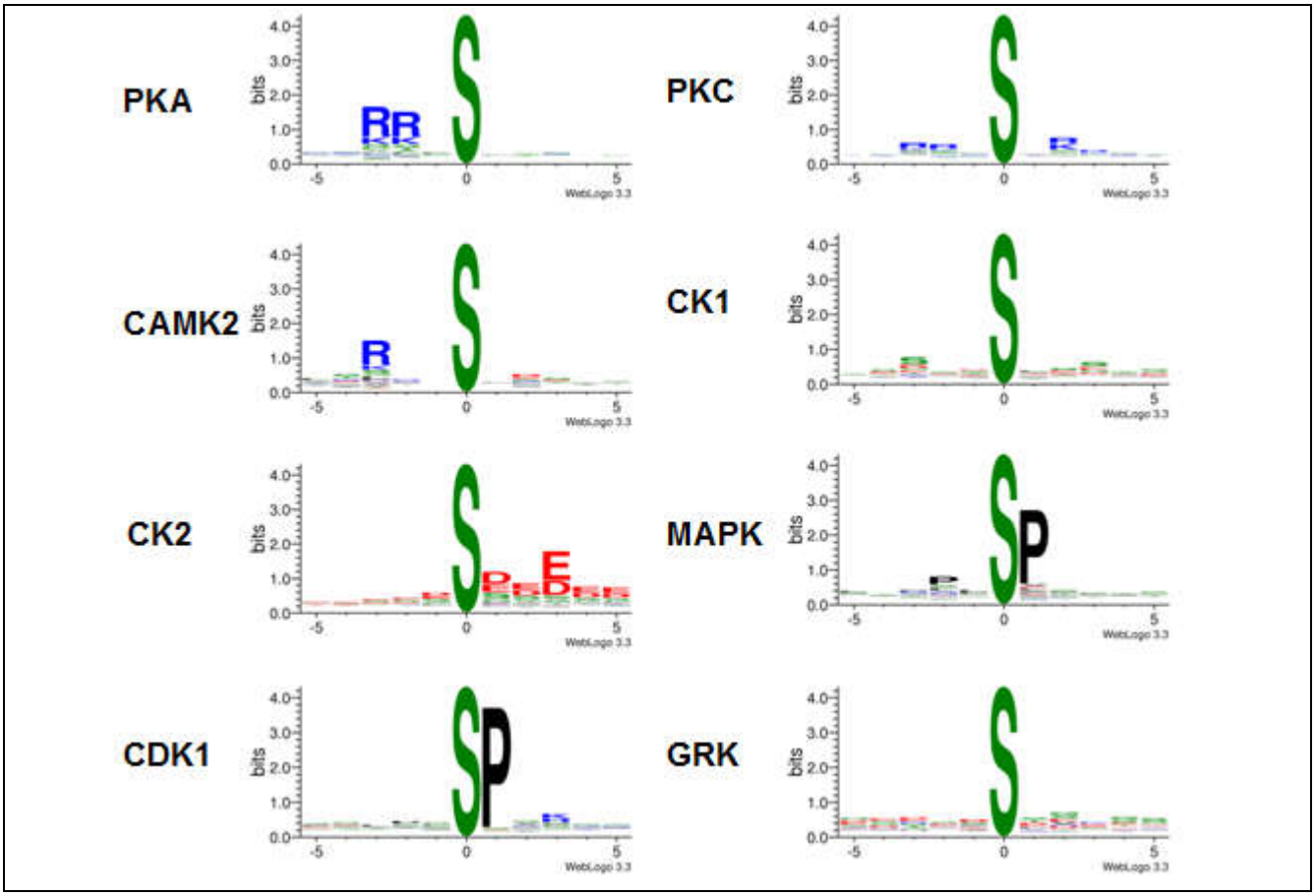
Ένα άλλο πολύ σημαντικό χαρακτηριστικό, είναι ο εντοπισμός περιοχών με μη σταθερή δομή. Τέτοια χαρακτηριστικά γίνονται ολοένα και πιο σημαντικά τα τελευταία χρόνια, γιατί πολλές πρωτεΐνες εμφανίζονται με μη σταθερή δευτεροταγή δομή, με συνέπεια να μην μπορούν να κρυσταλλωθούν αλλά και να εμπλέκονται 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/).

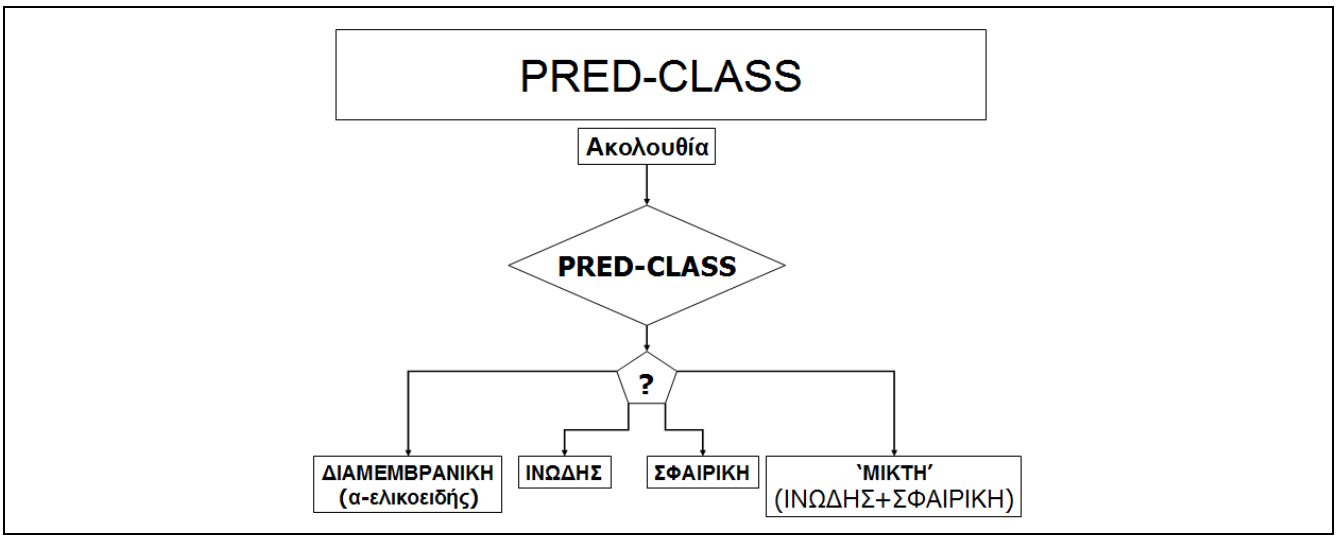


Εικόνα 7.24: *Λογότυπα αλληλουχιών από τις θέσεις δράσης διαφόρων κινασών. Παρόμοια εικόνα δίνουν και οι θέσεις δράσης με Θρεονίνη αντί Σερίνης.*

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/).

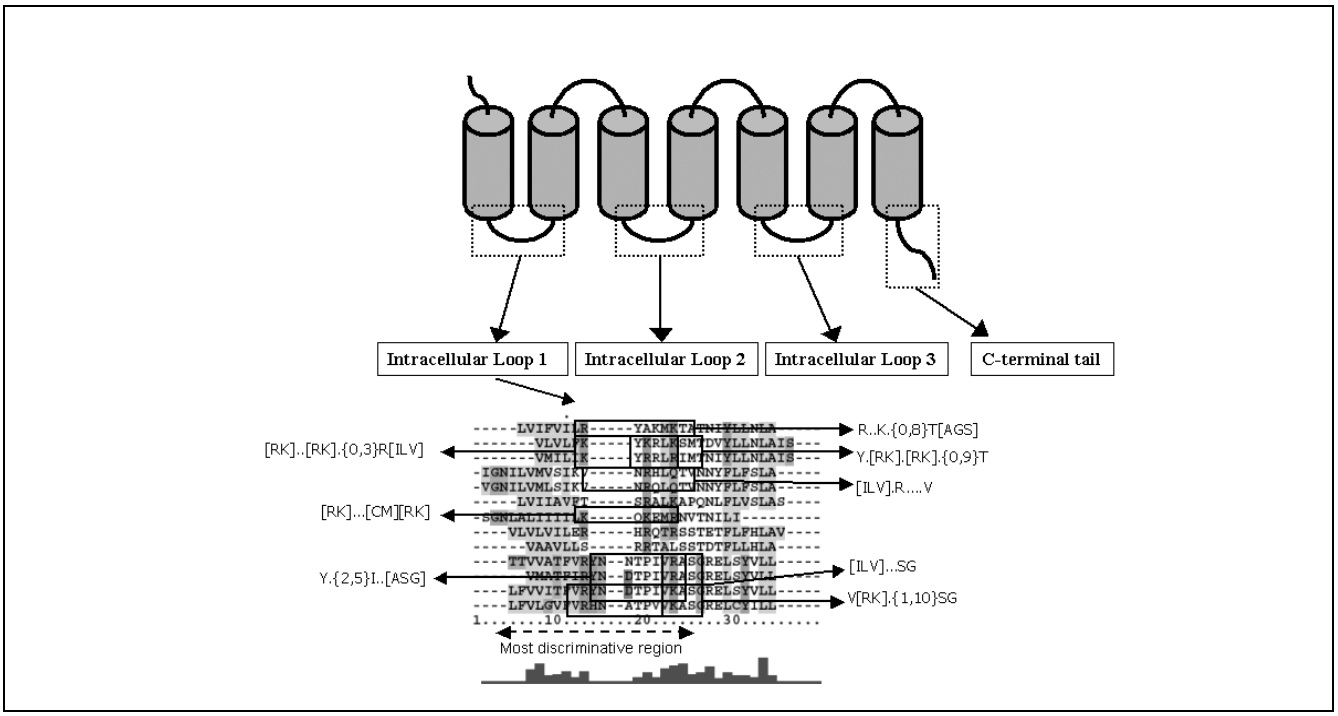
Μια άλλη ομάδα μετα-μεταφραστικών τροποποιήσεων είναι οι τροποποιήσεις που πραγματοποιούνται στο αμινοτελικό άκρο και σχετίζονται με τη σταθερότητα και το χρόνο ημιζωής της πρωτεΐνης. Το **Myristoylator** (http://web.expasy.org/myristoylator/) και το **NMT** (http://mendel.imp.ac.at/myristate/SUPLpredictor.htm) 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.

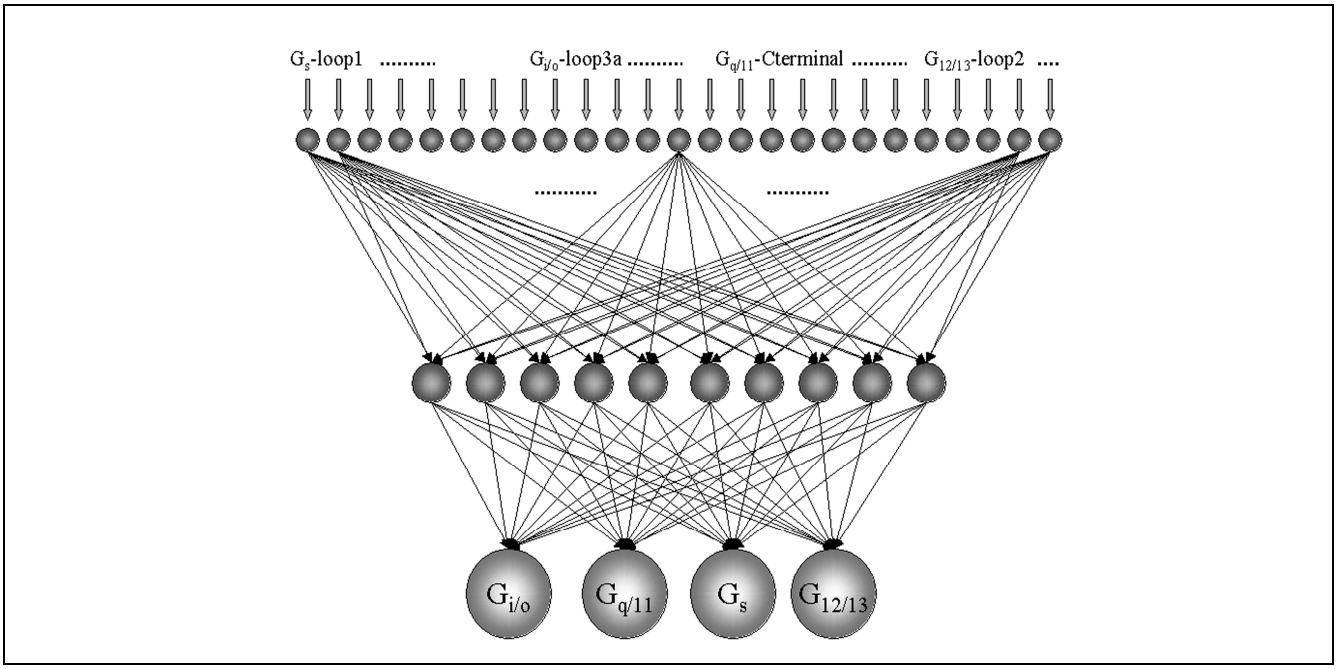


Εικόνα 7.26: *Σχηματική αναπαράσταση της μεθόδου PRED-COUPLE. Από τις στοιχίσεις των υποδοχέων με την ίδια σύζευξη, προέκυψε με εντατική αναζήτηση μια περιοχή που μπορεί να χρησιμοποιηθεί για διαχωρισμό μεταξύ των κατηγοριών και από την οποία κατασκευάζεται το pHMM. Τα πρότυπα κανονικών εκφράσεων που 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.

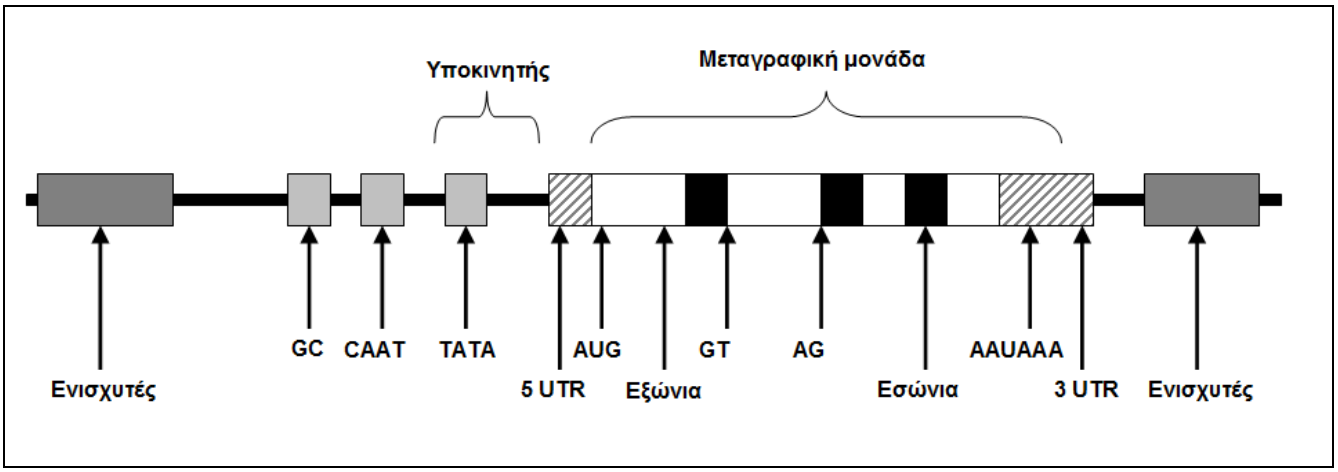


Εικόνα 7.27: *Το νευρωνικό δίκτυο της μεθόδου PRED-COUPLE2. Σαν δεδομένα εισόδου χρησιμοποιούνται τα σκορ από τα pHMM που είχαν κατασκευαστεί με μια διαδικασία όμοια με το PRED-COUPLE.*

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:

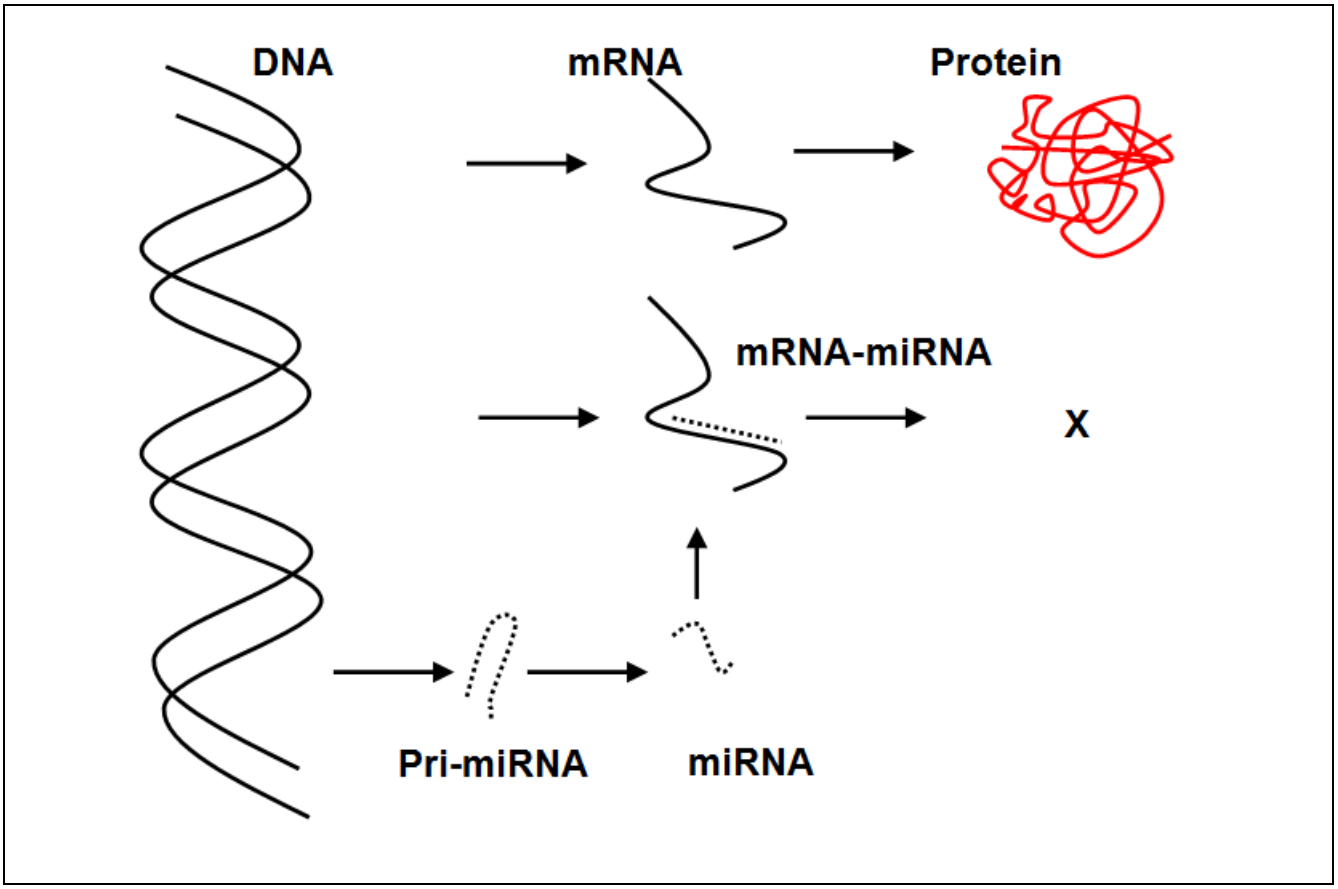
- **Human Splice Finder** (http://www.umd.be/HSF3/)
- **NetGene** (http://www.cbs.dtu.dk/services/NetGene2/)
- **NetPlant** (http://www.cbs.dtu.dk/services/NetPGene/)
- **GeneSplicer** (https://ccb.jhu.edu/software/genesplicer/)
- **SpliceView** (http://bioinfo4.itb.cnr.it/~webgene/wwwspliceview_ex.html)
- **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

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.