Βιοπληροφορική Ι

Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

Πανεπιστήμιο Θεσσαλίας
Λαμία 2015

Προγνωστικές μέθοδοι

- Οι μέθοδοι πρόγνωσης έρχονται χρονικά αλλά και λογικά να καλύψουν το κενό που έχουν αφήσει οι μέθοδοι ομοιότητας
- 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

Είναι απαραίτητες;

- τα μοριακά δεδομένα (γονιδιώματα, γονίδια, πρωτεΐνες κ.ο.κ.) συσσωρεύονται με εκθετικούς ρυθμούς
- 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.

Πως λειτουργούν

- 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
- υπάρχει ένα τεράστιο εύρος εφαρμογών τέτοιων μεθόδων με μεγάλη πρακτική χρησιμότητα, αλλά και διαφορετικών μαθηματικών και υπολογιστικών τεχνικών που χρησιμοποιούνται για το σκοπό αυτό

Παραδείγματα

- πρόγνωση της δευτεροταγούς δομής
- πρόγνωση διαμεβρανικών τμημάτων
- πρόβλεψη ιδιαίτερων χαρακτηριστικών των πρωτεϊνικών δομών, όπως οι θέσεις δράσης διαφόρων ενζύμων (μετα-μεταφραστική τροποποίηση, δισουλφιδικοί δεσμοί, σηματοδοτικές αλληλουχίες κλπ)
- 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.

Παραδείγματα (2)

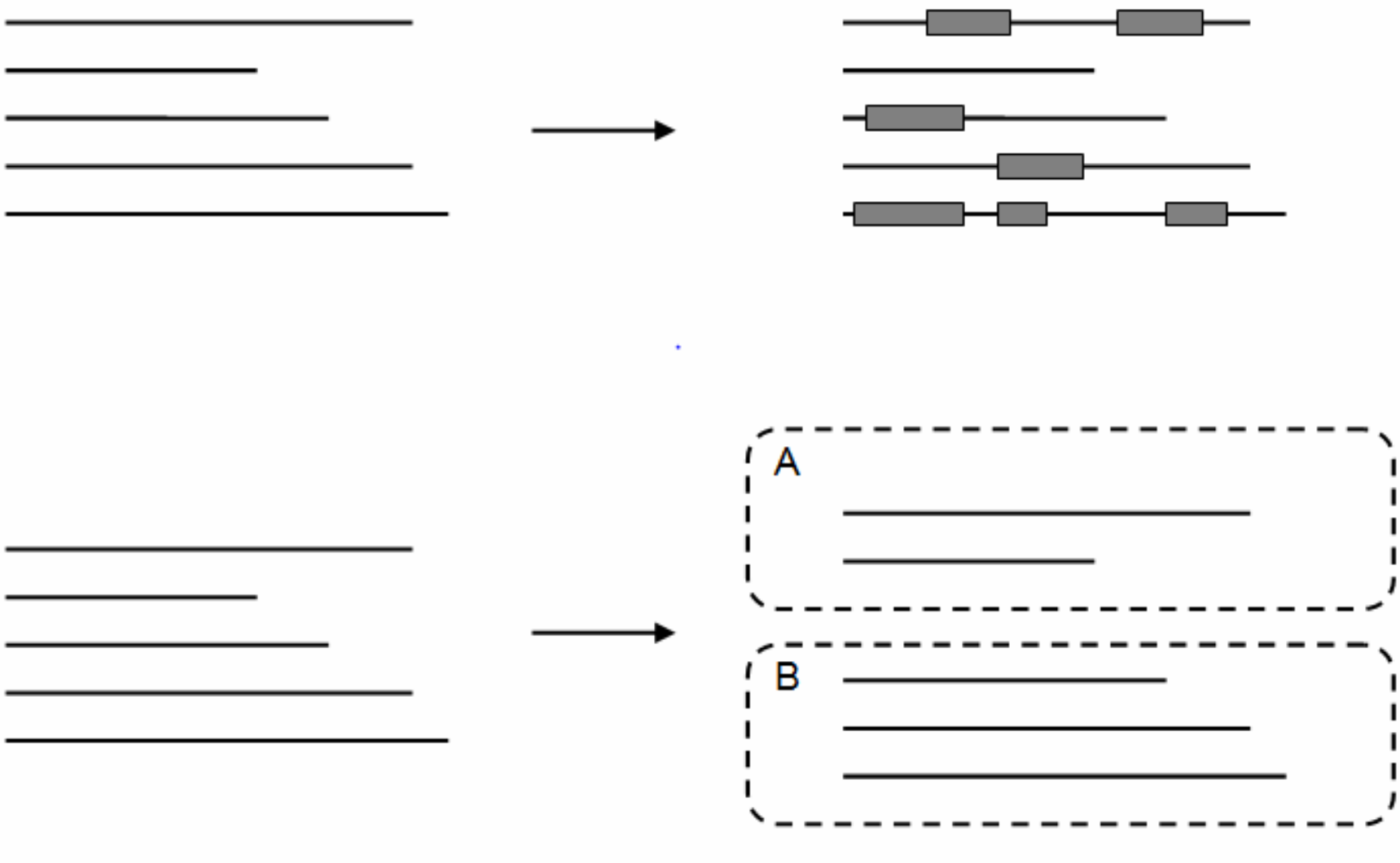
- Στην περίπτωση των αλληλουχιών DNA, το κλασικότερο παράδειγμα είναι η εύρεση γονιδίων (gene finding), πρόβλημα το οποίο είναι σημαντικό τόσο σε ευκαρυωτικούς όσο και προκαρυωτικούς οργανισμούς, και είναι μια μέθοδος που χρησιμοποιείται συνεχώς στον προσδιορισμό νέων γονιδιωμάτων.
- Φυσικά, το πρόβλημα 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

Κωδικοποίηση 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 αλλληλουχίες 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
- Όπως γίνεται ήδη φανερό, μια πρώτη μορφή «μεθόδου πρόγνωσης» είναι δυνατό να κατασκευαστεί με τη χρήση των μεθόδων εύρεσης προτύπων και προφίλ σε αλληλουχίες

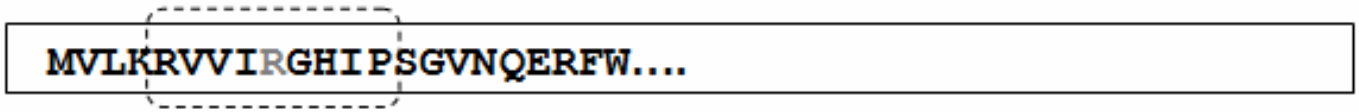
Δύο κατηγορίες μεθόδων

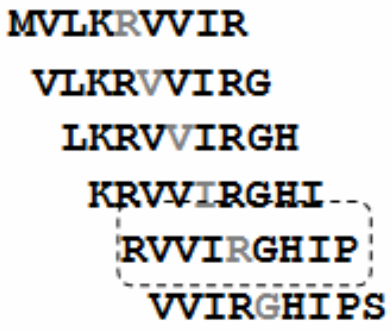
- Γενικά, υπάρχουν δύο κατηγορίες προβλημάτων πρόγνωσης ή πρόβλεψης
- Στην πρώτη περίπτωση ενδιαφερόμαστε για τοπική πρόγνωση κατά μήκος της αλληλουχίας. Ενδιαφερόμαστε δηλαδή να δούμε ποια συγκεκριμένα κατάλοιπα ή νουκλεοτίδια ανήκουν σε μια κατηγορία και ποια σε άλλη
- Στη δεύτερη κατηγορία, ενδιαφερόμαστε να ταξινομήσουμε κάποιες αλληλουχίες σε δύο ή περισσότερες κατηγορίες


A
B

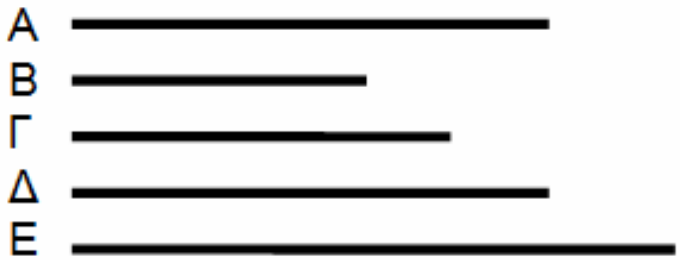
Τρόπος αντιμετώπισης

- οι δύο κατηγορίες μεθόδων απαιτούν και διαφορετικούς τρόπους χειρισμού των δεδομένων αλληλουχιών
- 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
- 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





	X_1	X_2	X_3	X_4	...	X_k
Παράθυρο 1						
Παράθυρο 2						
Παράθυρο 3						
Παράθυρο 4						
Παράθυρο 5						
Παράθυρο 6						



	X_1	X_2	X_3	X_4	...	X_k
Ακολουθία Α						
Ακολουθία Β						
Ακολουθία Γ						
Ακολουθία Δ						
Ακολουθία Ε						

Τοπική κωδικοποίηση

- 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).
- 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.
- 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.

Τοπική κωδικοποίηση

- 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
- 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.

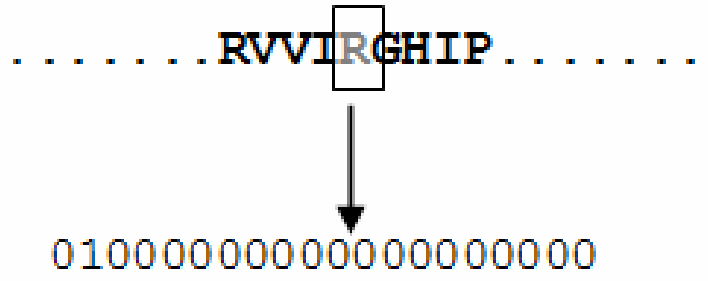
Τοπική κωδικοποίηση

- 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
- 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).

A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

...

...



Ολική κωδικοποίηση

- 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).

Ολική κωδικοποίηση

- 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),
- 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/.

Μεθοδολογίες για την εκπαίδευση και τον έλεγχο μιας μεθόδου πρόγνωσης

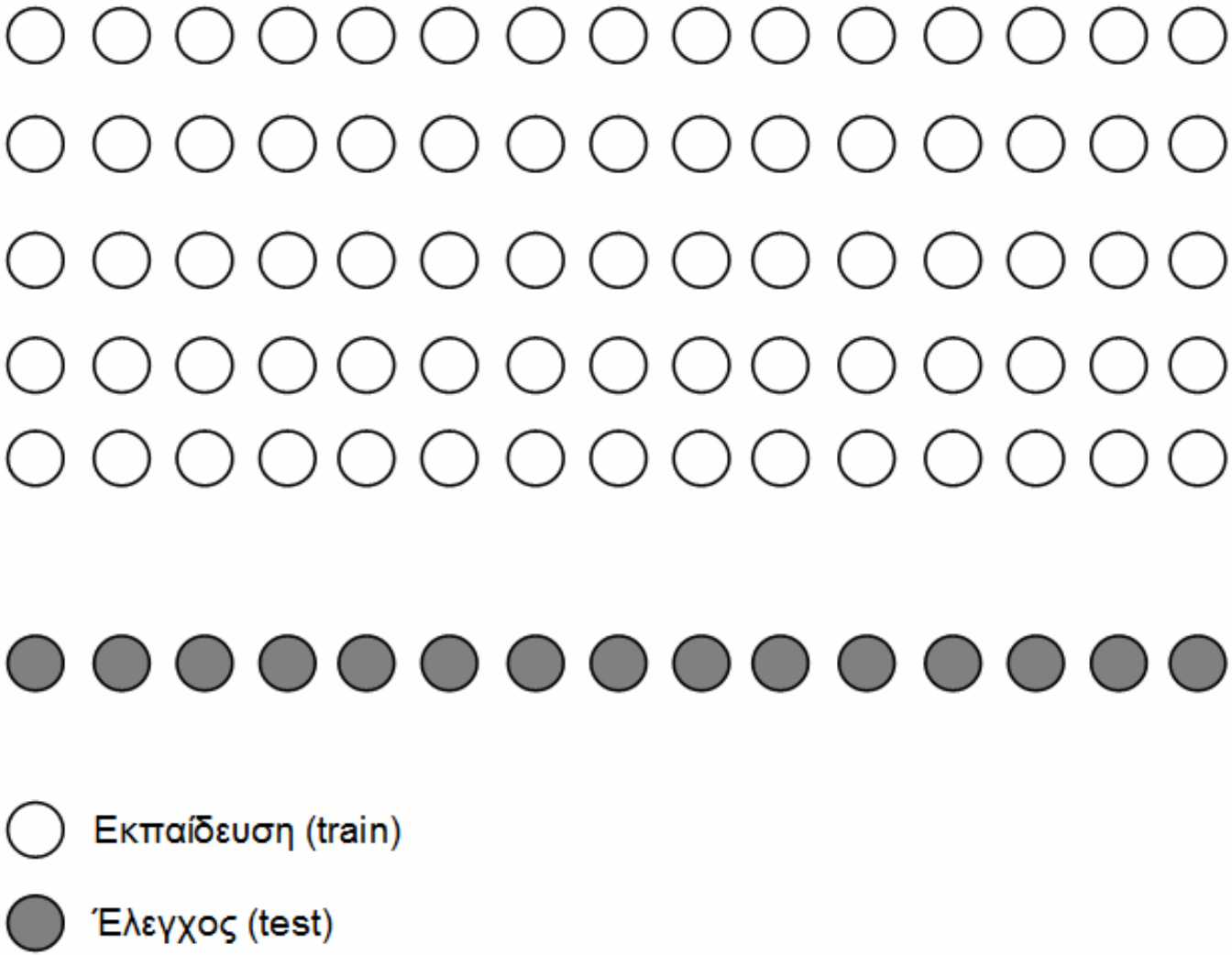
- Αν το πρόβλημα είναι νέο, το σύνολο των δεδομένων εκπαίδευσης θα πρέπει να συλλεχθεί από τις γνωστές βάσεις δεδομένων ή από τη βιβλιογραφία με τις κατάλληλες επερωτήσεις (οι οποίες μπορεί να είναι και ιδιαίτερα δύσκολες).
- Συνηθισμένη είναι και η περίπτωση κάποιος να χρησιμοποιεί κάποιο σύνολο που 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.

Σύνολο εκπαίδευσης

- 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

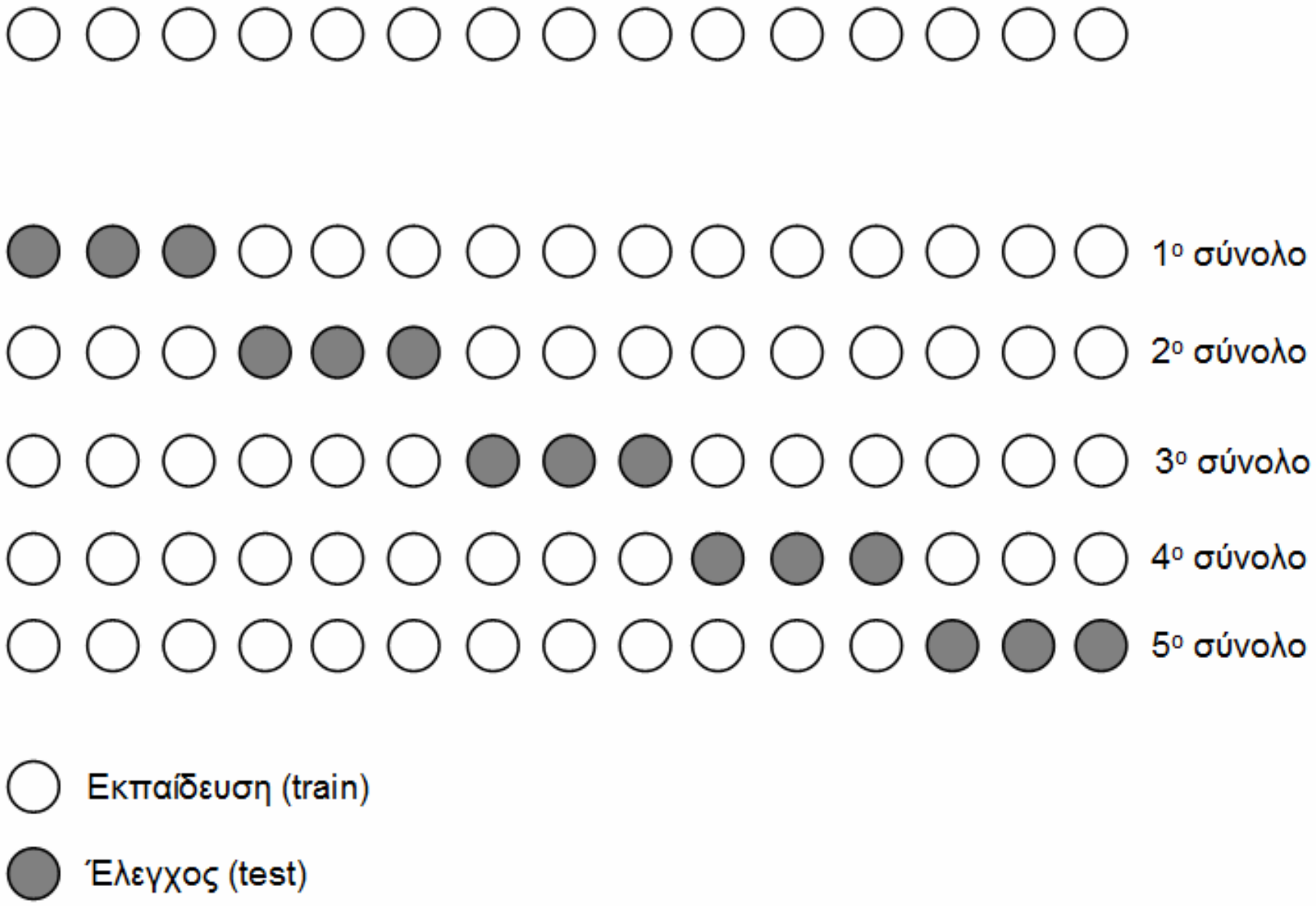
Διαδικασία

- 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 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.
- 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


Εκπαίδευση (train)
Έλεγχος (test)

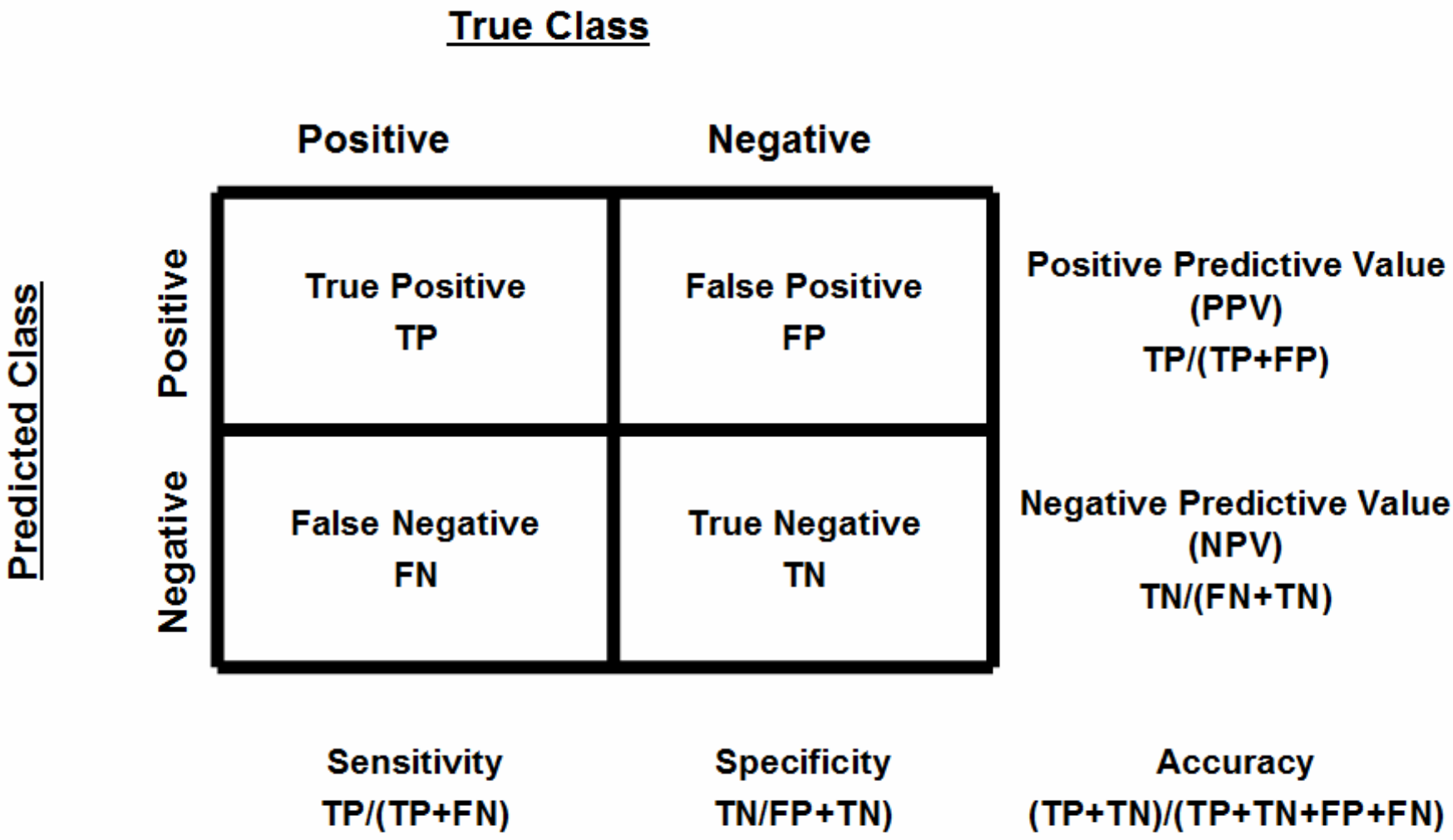
Διαδικασία

- 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
- 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).


1ο σύνολο
2ο σύνολο
3ο σύνολο
4ο σύνολο
5ο σύνολο
Εκπαίδευση (train)
Έλεγχος (test)

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).


True Class
Positive
Negative
Predicted Class
Positive
Negative
True Positive
TP
False Positive
FP
False Negative
FN
True Negative
TN
Positive Predictive Value
(PPV)
TP/(TP+FP)
Negative Predictive Value
(NPV)
TN/(FN+TN)
Sensitivity
TP/(TP+FN)
Specificity
TN/FP+TN)
Accuracy
(TP+TN)/(TP+TN+FP+FN)

Συντελεστής συσχέτισης

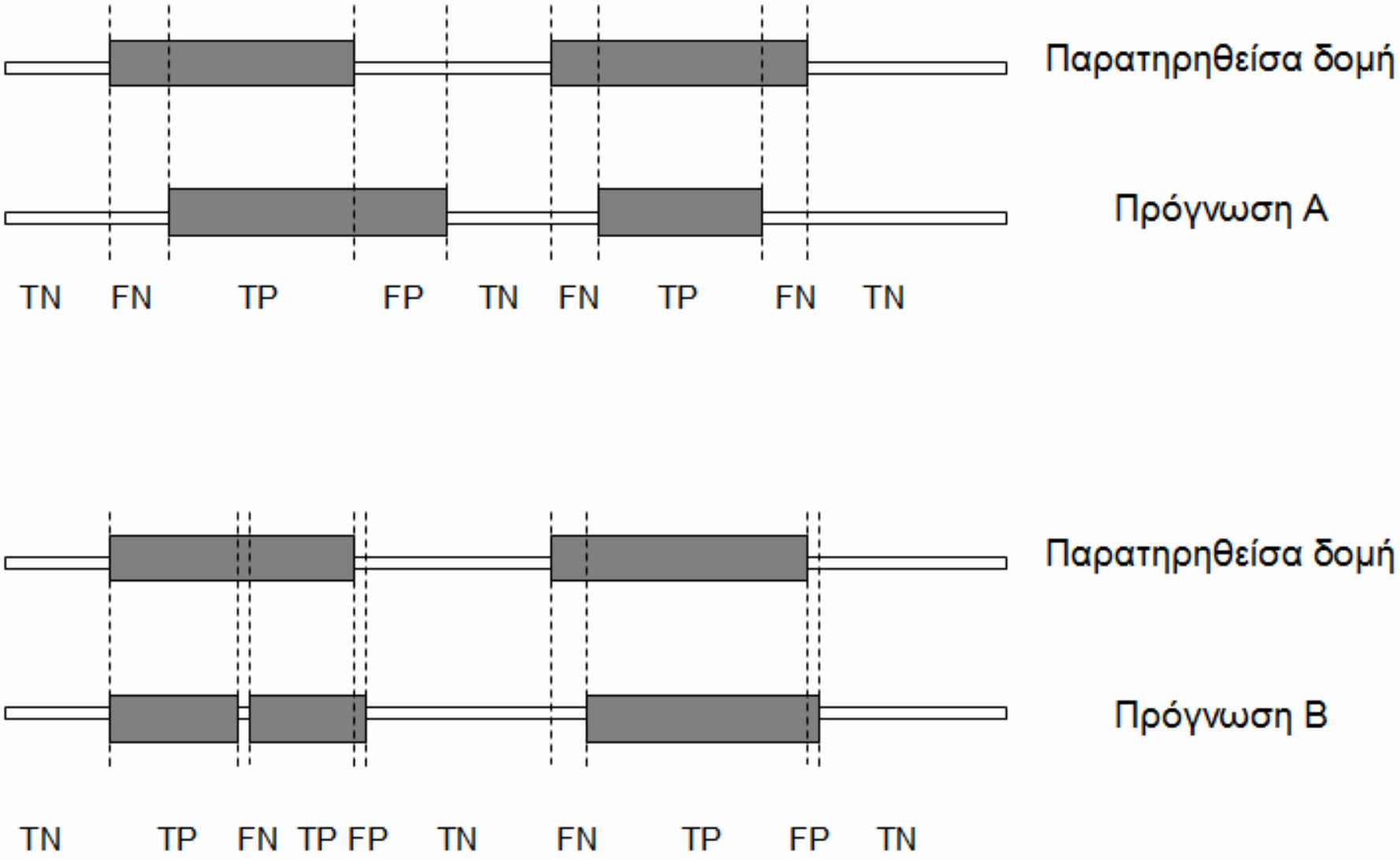
- ς, ένα άλλο μέτρο που χρησιμοποιείται είναι ο γνωστός συντελεστής συσχέτισης του Matthews

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

- 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.

SOV

- 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

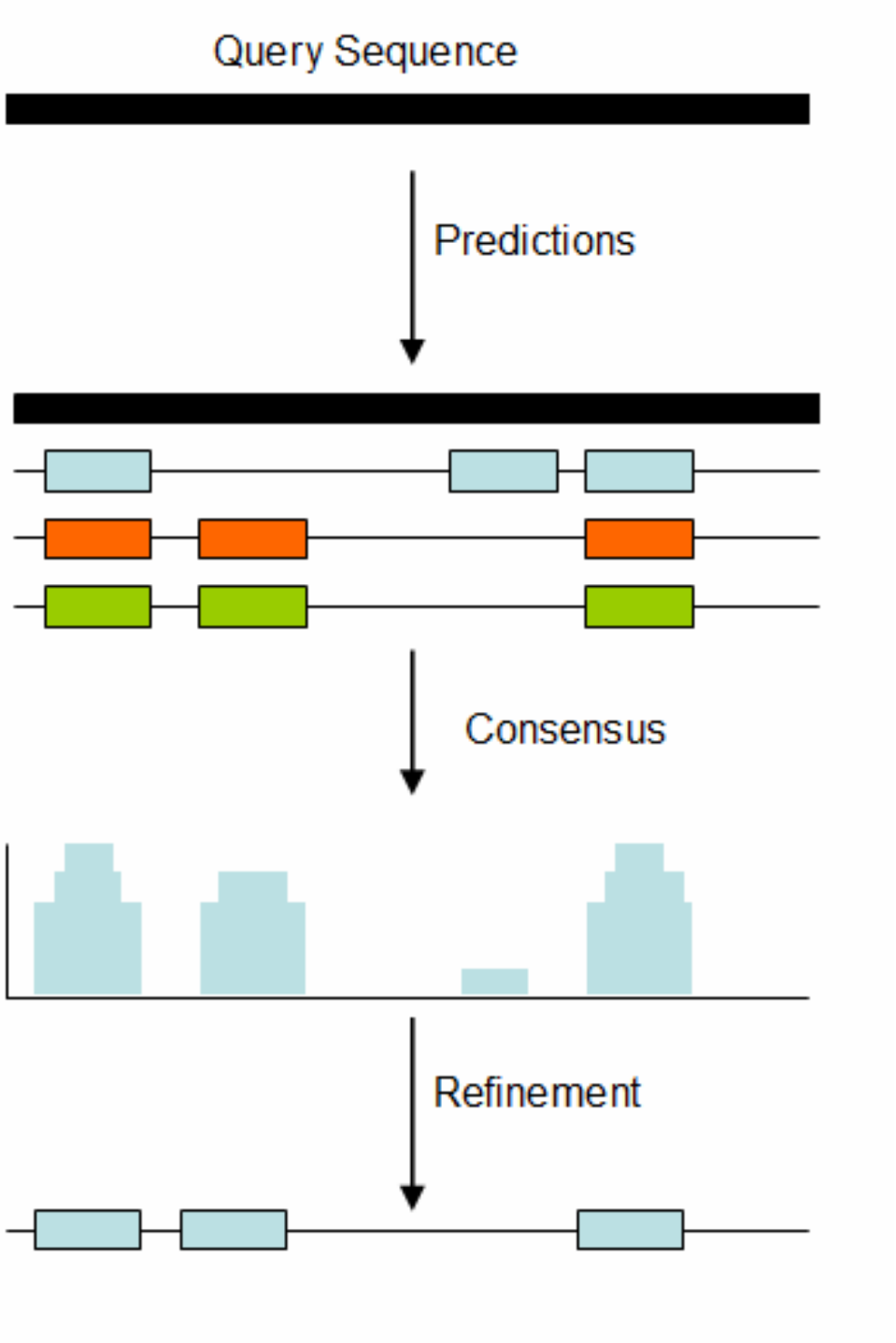

Παρατηρηθείσα δομή
Πρόγνωση Α
TN FN TP FP TN FN TP FN TN
Παρατηρηθείσα δομή
Πρόγνωση Β
TN TP FN TP FP TN FN TP FP TN

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

Συνδυαστικές μέθοδοι

- Μια συνδυαστική/συναινετική μέθοδος πρόγνωσης, βασίζεται στην βασική απλή ιδέα, ότι αν συνδυαστούν ανεξάρτητες μέθοδοι το αποτέλεσμα είναι πάντα καλύτερο.
- Οι μεθοδολογίες αυτές έχουν χρησιμοποιηθεί σε διάφορους τομείς, είτε με απλό τρόπο (majority vote, consensus) είτε με πιο σύνθετους αλγόριθμους μηχανικής μάθησης (ensemble learning, meta-algorithms κ.ο.κ.).
- Η απλή αυτή διαίσθηση («ας ακούσουμε πολλές γνώμες»), έχει επίσης βρει και τη μαθηματική της τεκμηρίωση καθώς υπάρχουν θεωρητικές αποδείξεις ότι ο συνδυασμός «ασθενών ταξινομητών» (weak classifiers), δηλαδή ταξινομητών οι οποίοι αποδίδουν μεν καλύτερα από το τυχαίο (π.χ. συντελεστής συσχέτισης >0, ή Q>0.5), δίνει πάντα έναν ταξινομητή με καλύτερη αποτελεσματικότητα.
- Φυσικά, είναι προφανές ότι αν κάποια από τις μεθόδους είναι ιδιαίτερα καλή (π.χ. συντελεστής συσχέτισης >0.95 ή Q>0.99), τότε η μέθοδος δεν θα δουλέψει καθώς η «ισχυρή» μέθοδος θα υπερισχύει πάντα


Query Sequence
Predictions
Consensus
Refinement

Προβληματισμοί

- 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).

Προβληματισμοί

- 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.

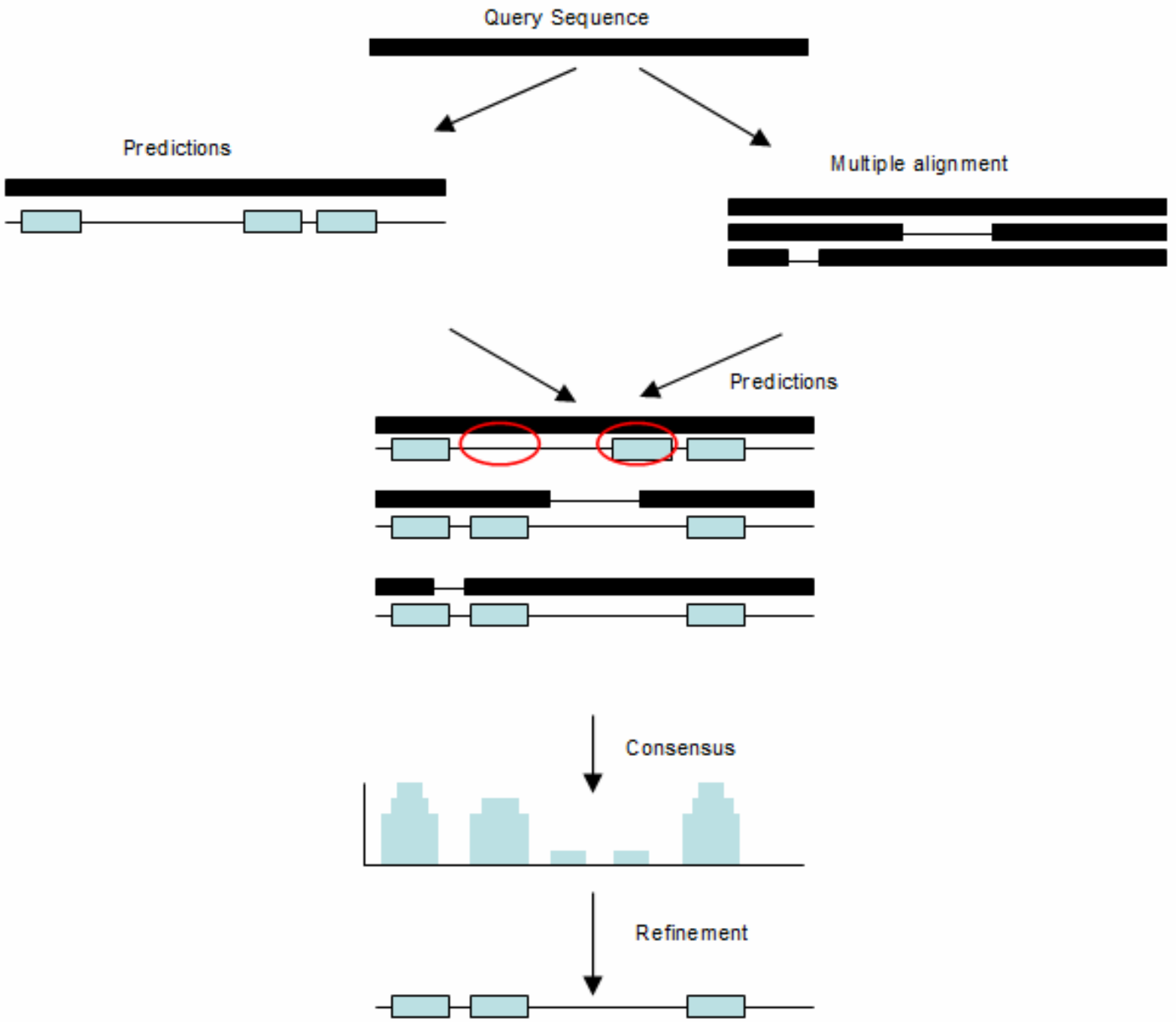
Ενσωμάτωση εξελικτικής πληροφορίας με τη μορφή πολλαπλών στοιχίσεων

- 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).

Επεξηγήσεις

- 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.

- 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


Query Sequence
Predictions
Multiple alignment
Predictions
Consensus
Refinement

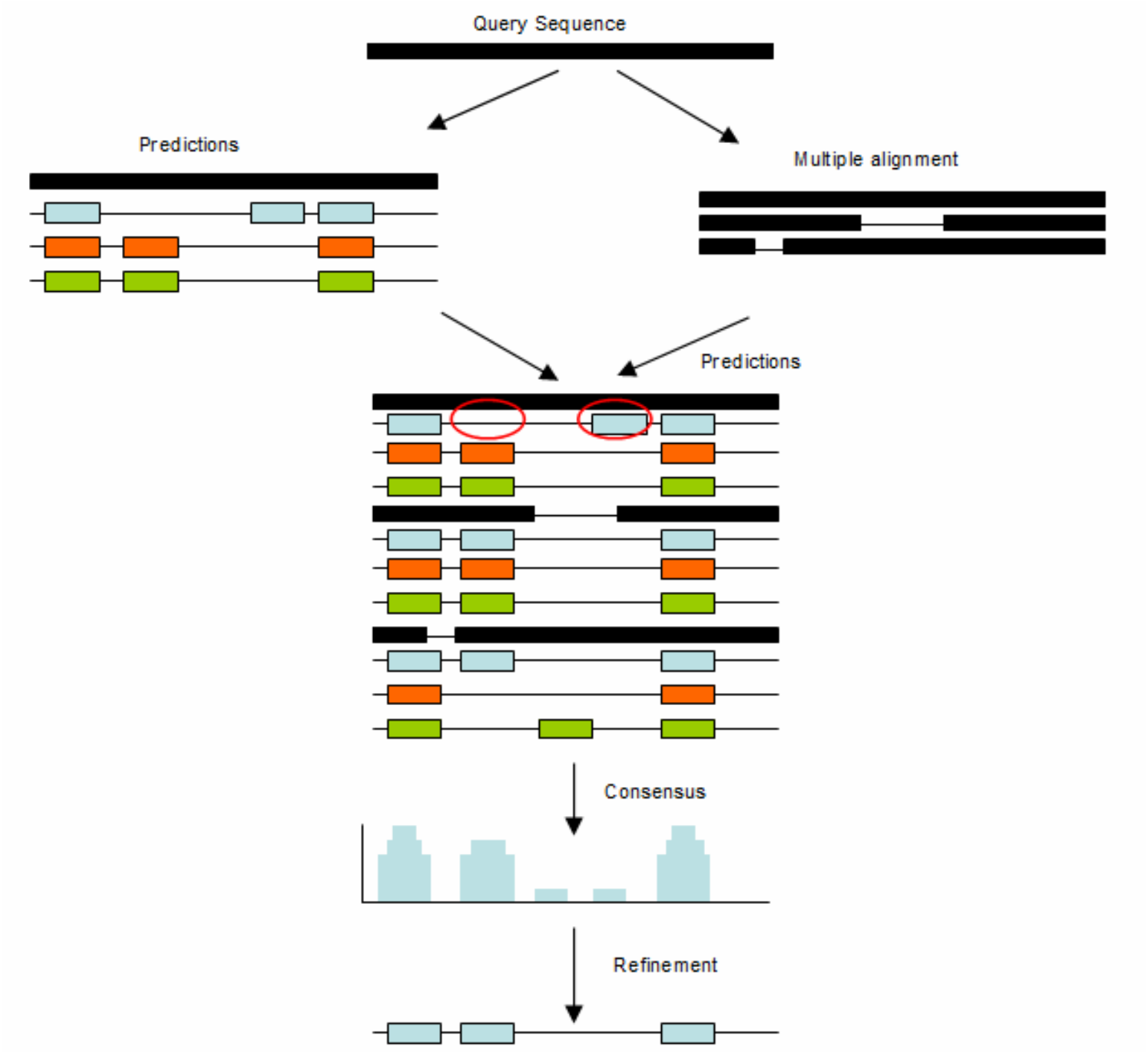
PSI-BLAST

- 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
- 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

		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

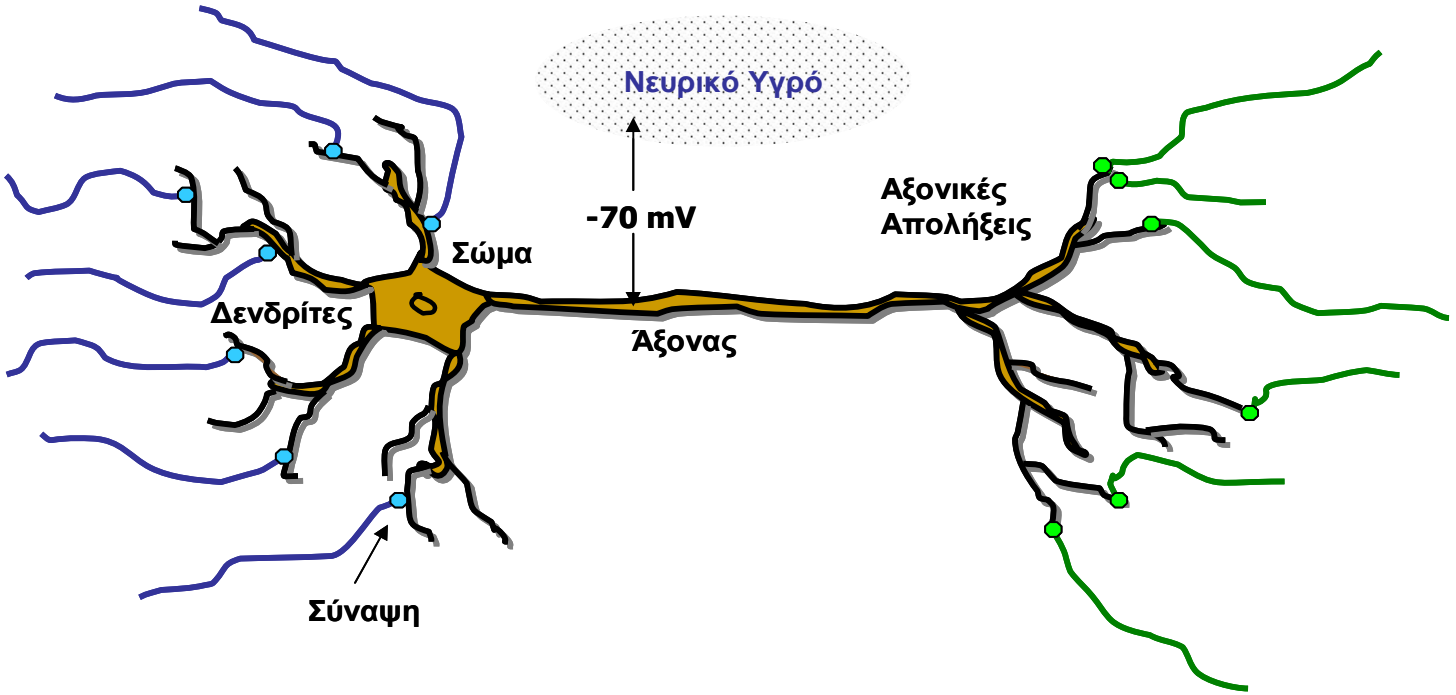
Συνδυασμός

- Τέλος, πρέπει να τονίσουμε ότι οι δύο παραπάνω γενικές μεθοδολογίες (η συνδυαστική πρόγνωση και η χρήση πολλαπλών στοιχίσεων), μπορούν άνετα να συνδυαστούν μεταξύ τους
- 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


Query Sequence
Predictions
Multiple alignment
Predictions
Consensus
Refinement

Νευρωνικά Δίκτυα

- 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),



Είσοδος
Αξονικές απολήξεις
άλλων νευρώνων

Έξοδος
Δενδρίτες
άλλων νευρώνων

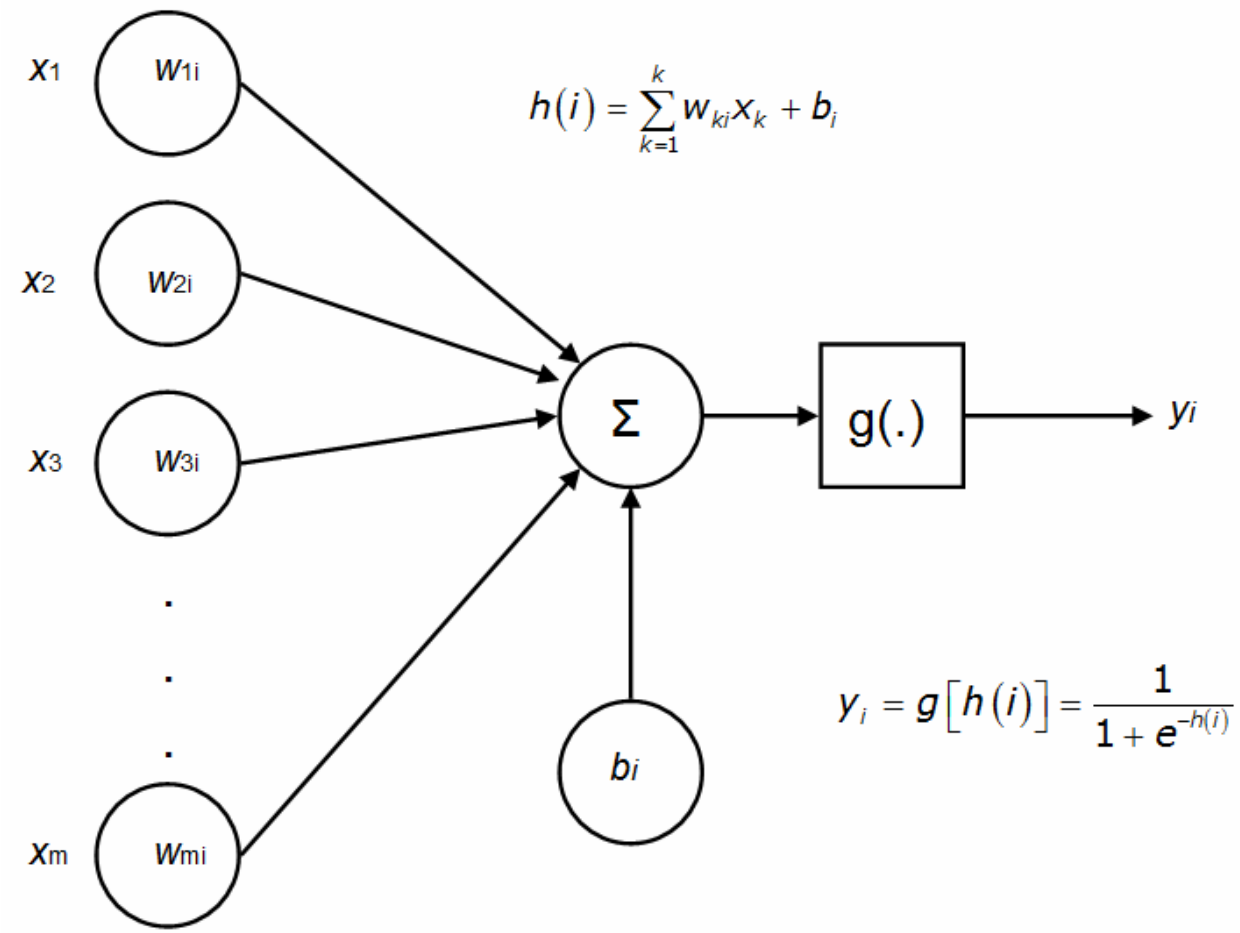
ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΒΙΟΛΟΓΙΚΟΥ ΝΕΥΡΩΝΑ

- Απλή δομική μονάδα
- Ομοιόμορφη αρχιτεκτονική
- Υψηλός παραλληλισμός-Κατανεμημένη επεξεργασία
- Ταχύτητα επεξεργασίας-απόκρισης
- Αξιόπιστη λειτουργία σε «αντίξοες» συνθήκες
- Αποθήκευση-Ανάκληση δεδομένων βάσει εμπειρίας
- Προσαρμογή στην αντιμετώπιση συγκεκριμένων διαδικασιών

ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΣΤΟΙΧΕΙΑ ΤΕΧΝΗΤΟΥ ΝΕΥΡΩΝΑ

- Αριθμητικά «ερεθίσματα»
- Σταθμισμένες Διασυνδέσεις-Συνάψεις
- Αθροιστής
- Συνάρτηση Ενεργοποίησης (μη γραμμική)
- Εξωτερική (σταθερή τιμή) πόλωσης

Ο ΤΕΧΝΗΤΟΣ ΝΕΥΡΩΝΑΣ

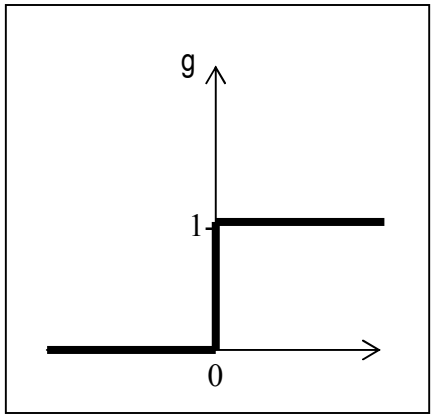


ΣΥΝΑΡΤΗΣΕΙΣ ΕΝΕΡΓΟΠΟΙΗΣΗΣ

1. Συνάρτηση Κατωφλίου (Threshold ή Heaviside function)
2. Μερικώς-Γραμμική Συνάρτηση (Piecewise Linear Function)
3. Σιγμοειδής συνάρτηση (Sigmoid function)

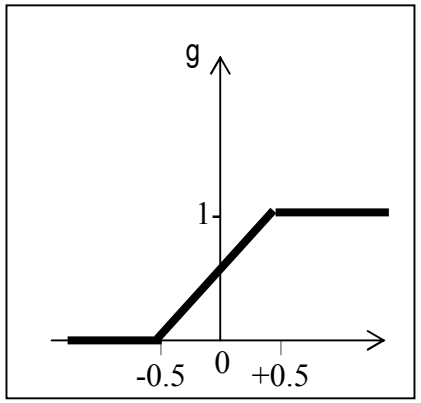
Συνάρτηση Κατωφλίου

$$g(h) = \begin{cases} 1, h \geq 0 \\ 0, h < 0 \end{cases}$$



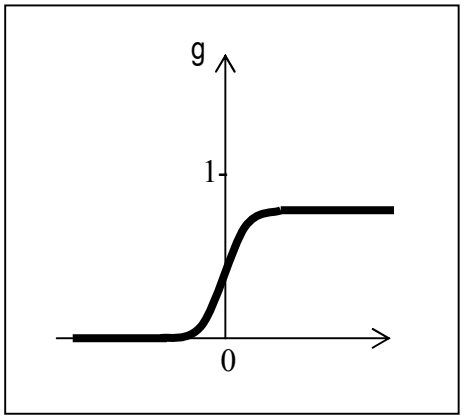
Μερικώς-Γραμμική Συνάρτηση

$$g(h) = \begin{cases} 1 & , h \geq +\frac{1}{2} \\ h + \frac{1}{2} & , -\frac{1}{2} < h < +\frac{1}{2} \\ 0 & , h < -\frac{1}{2} \end{cases}$$



Σιγμοειδείς Συναρτήσεις (Λογιστική Συνάρτηση)

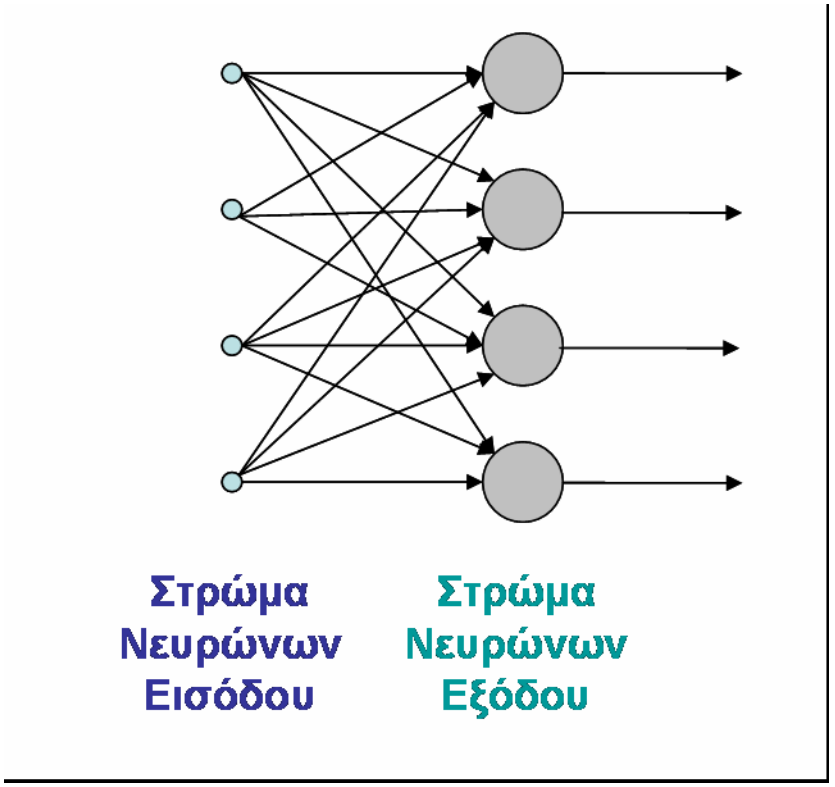
$$g(h) = \frac{1}{1 + e^{-ah}}$$



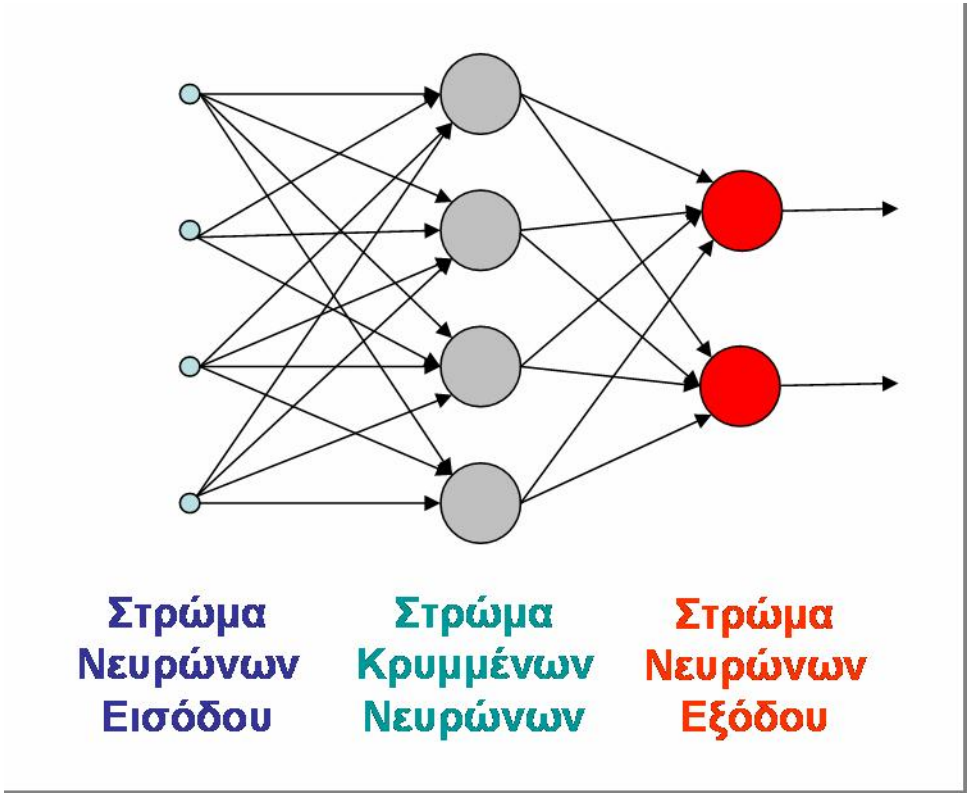
ΑΡΧΙΤΕΚΤΟΝΙΚΗ - ΤΟΠΟΛΟΓΙΑ

- Πλήθος νευρώνων
- Οργάνωση νευρώνων κατά στρώματα
- Διασυνδέσεις μεταξύ νευρώνων
- Πλήθος στρωμάτων (κρυμμένα στρώματα)
- Τρόπος διάδοσης σήματος

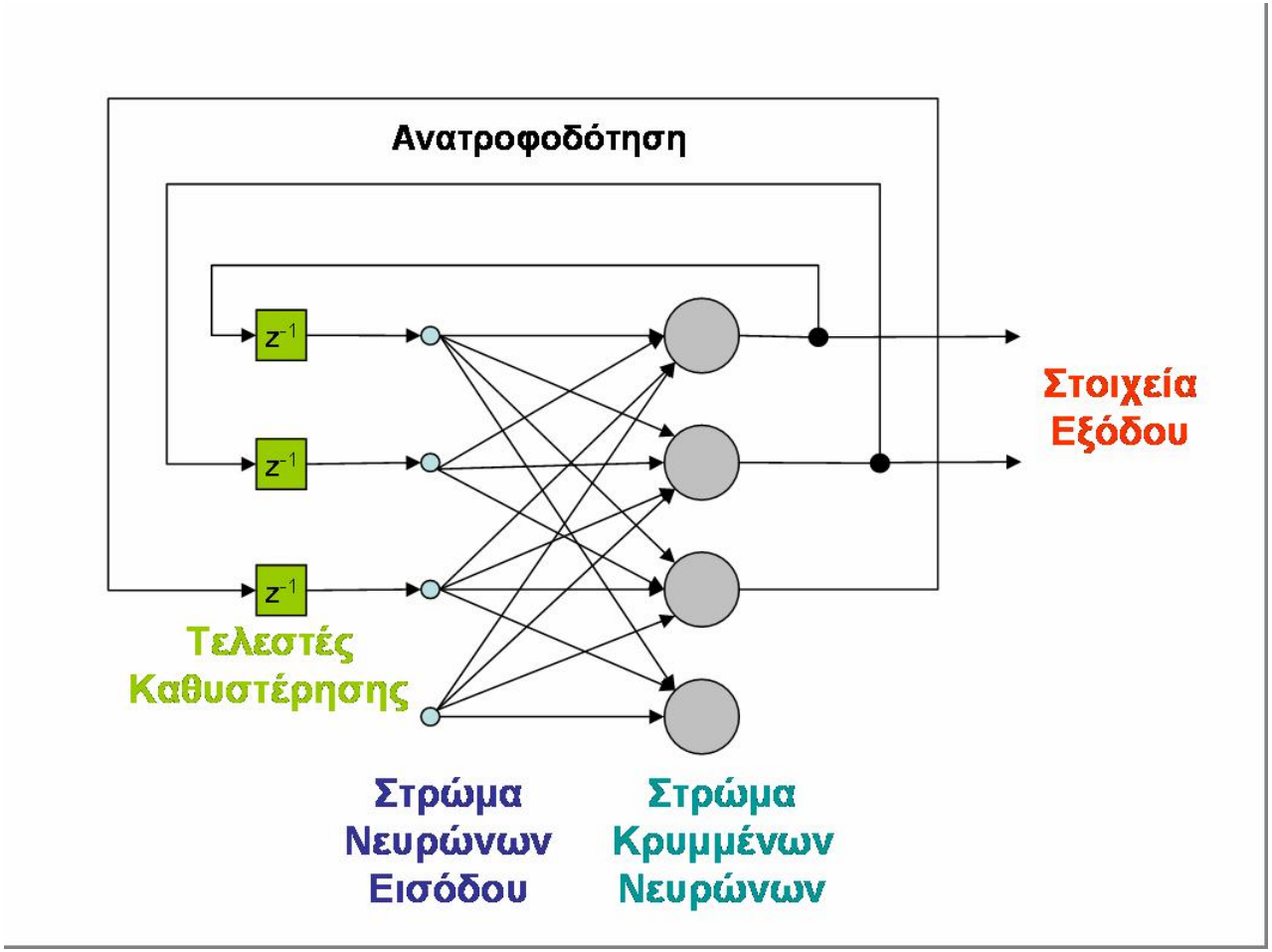
ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ ΕΝΟΣ ΣΤΡΩΜΑΤΟΣ ΝΕΥΡΩΝΩΝ



ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ ΠΟΛΛΑΠΛΩΝ ΣΤΡΩΜΑΤΩΝ ΝΕΥΡΩΝΩΝ



ΑΝΑΔΡΟΜΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ



ΜΑΘΗΣΗ ΑΠΟ ΤΕΧΝΗΤΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

- Μάθηση «υπό εποπτεία»
- Ελεύθερες παράμετροι = Συναπτικά Βάρη
- Δείγμα Δεδομένων Εκπαίδευσης
- Προσαρμογή συναπτικών βαρών (επαναληπτική διαδικασία βελτιστοποίησης για ελαχιστοποίηση του σφάλματος)
- Τερματισμός Διαδικασίας => Έλεγχος Σύγκλισης Βαρών

ΠΡΟΣΔΟΚΙΑ η ΓΕΝΙΚΕΥΣΗ σε άγνωστα παραδείγματα δεδομένων

Ο ΑΛΓΟΡΙΘΜΟΣ ΟΠΙΣΘΙΑΣ ΔΙΑΔΟΣΗΣ ΣΦΑΛΜΑΤΩΝ (error back-propagation)

- Εκπαίδευση συστημάτων Εμπρόσθιας Τροφοδότησης – Πολλαπλών Στρωμάτων
- Αρχικοποίηση βαρών
- Δύο βήματα υπολογισμών
  - Δεδομένα βάρη=> εμπρόσθια διάδοση
  - Έλεγχος σφάλματος=>αναπροσαρμογή βαρών
- Έλεγχος ελαχιστοποίησης σφαλμάτων για τερματισμό διαδικασίας

back-propagation

- 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

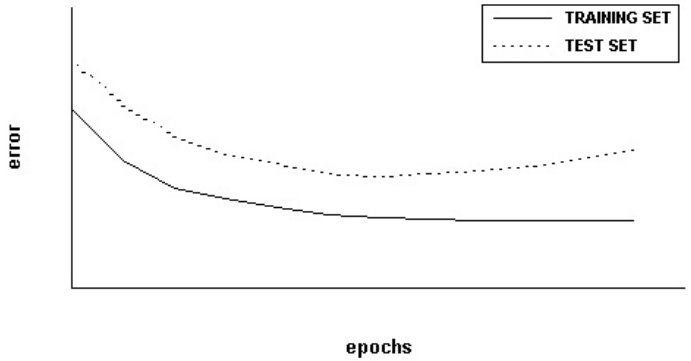
back-propagation

- 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

Λεπτομέρειες

- 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.

ΠΡΟΒΛΗΜΑ: ΥΠΕΡΠΡΟΣΑΡΜΟΓΗ ΣΤΑ ΔΕΔΟΜΕΝΑ (OVERFITTING)



- Αντιστοιχία με «φυσική» μάθηση => «ΠΑΠΑΓΑΛΙΑ»
- Τεχνάσματα για την αποφυγή

Λογισμικό

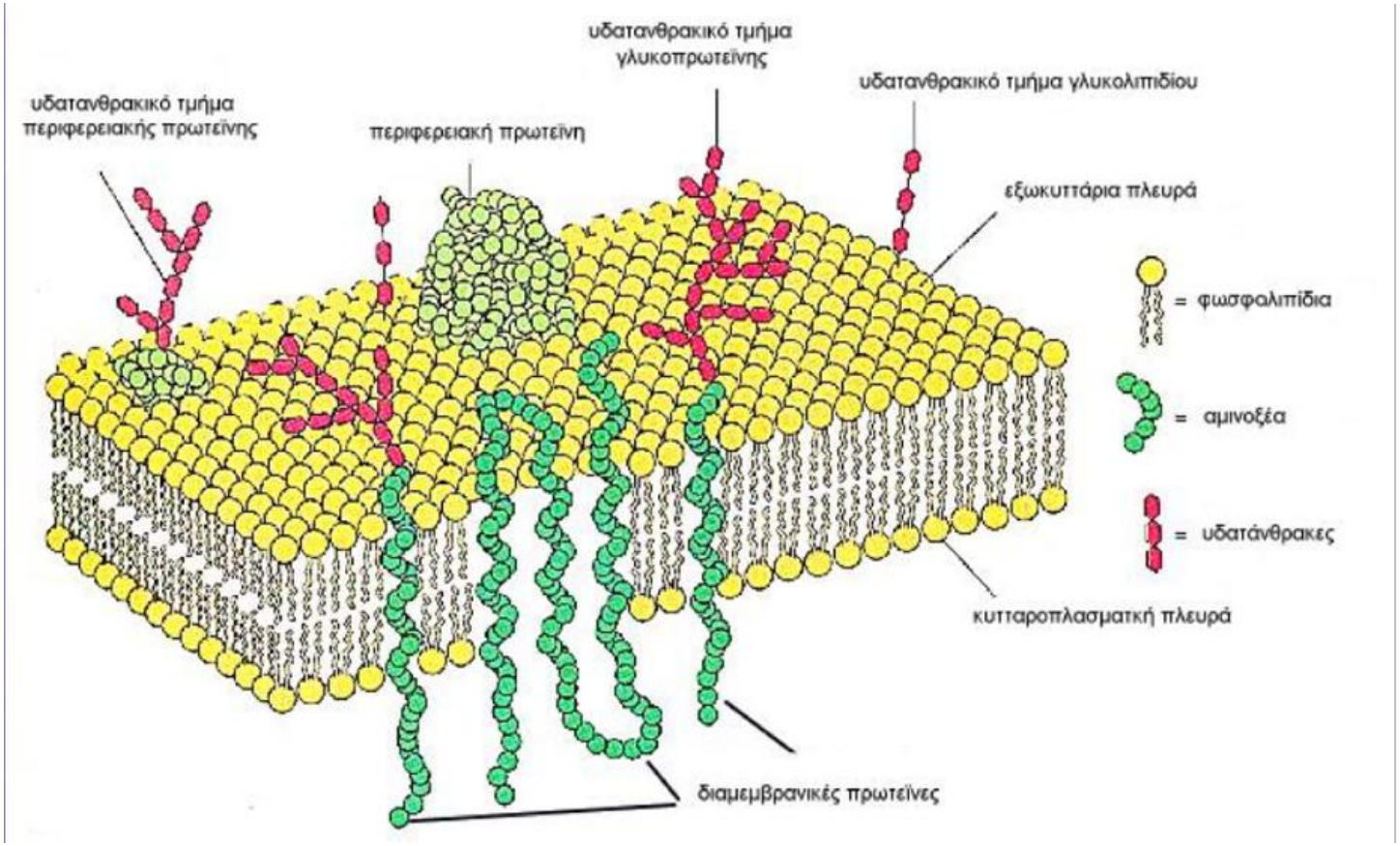
- Για την εφαρμογή νευρωνικών δικτύων σε προβλήματα βιοπληροφορικής, θα πρέπει καταρχάς να γίνουν οι κατάλληλοι μετασχηματισμοί των αλληλουχιών για να έρθουν στη μορφή που περιγράψαμε πριν. Κατόπιν θα πρέπει να χρησιμοποιηθεί κάποιο γενικό πακέτο για νευρωνικά δίκτυα που διαθέτουν τα γνωστά μαθηματικά 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/).

Μέθοδοι πρόγνωσης

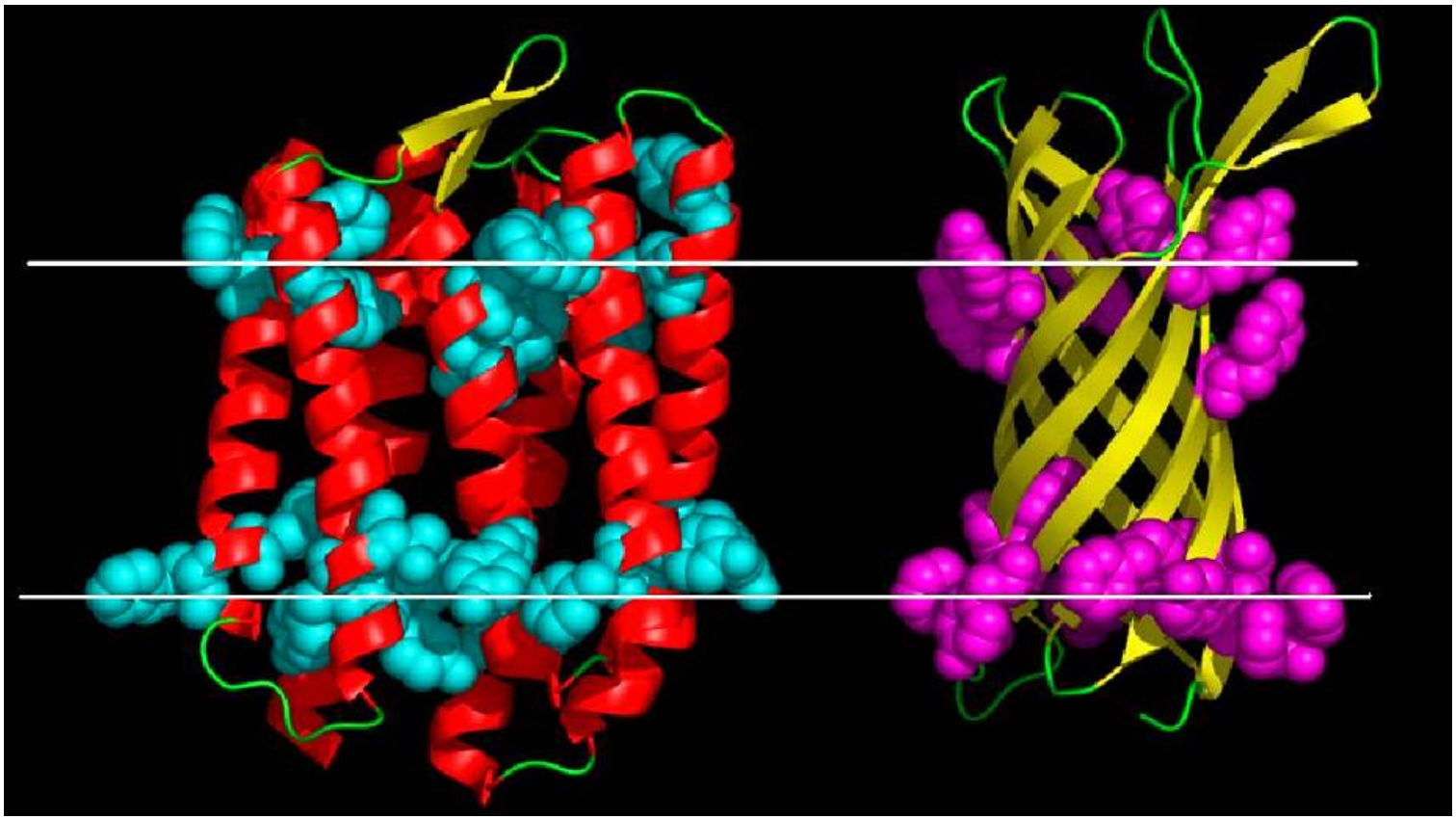
- Μέθοδοι πρόγνωσης για πρωτεΐνες
  - Δευτεροταγής δομή
  - Διαμεμβρανικά τμήματα
  - Σηματοδοτικές αλληλουχίες
  - Στόχευση
  - Μετα-μεταφραστικές τροποποιήσεις
  - Αλληλεπιδράσεις, δομική κατάταξη κλπ
- Μέθοδοι πρόγνωσης DNA/RNA
  - Έύρεση γονιδίων
  - Εύρεση υποκινητών
  - Σημεία συρραφής
  - TIS
  - Poly-A
  - miRNA

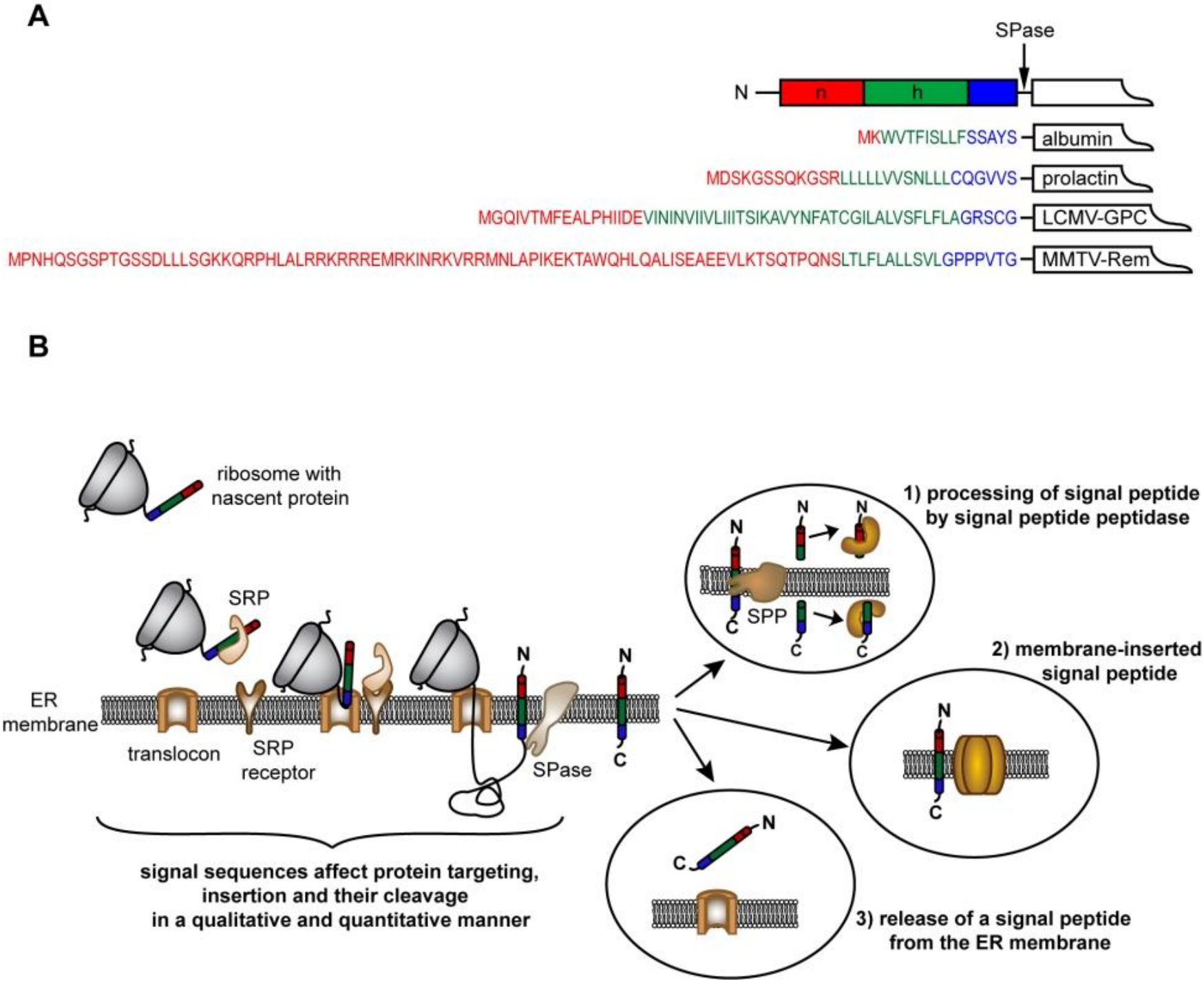
Δευτεροταγής Δομή

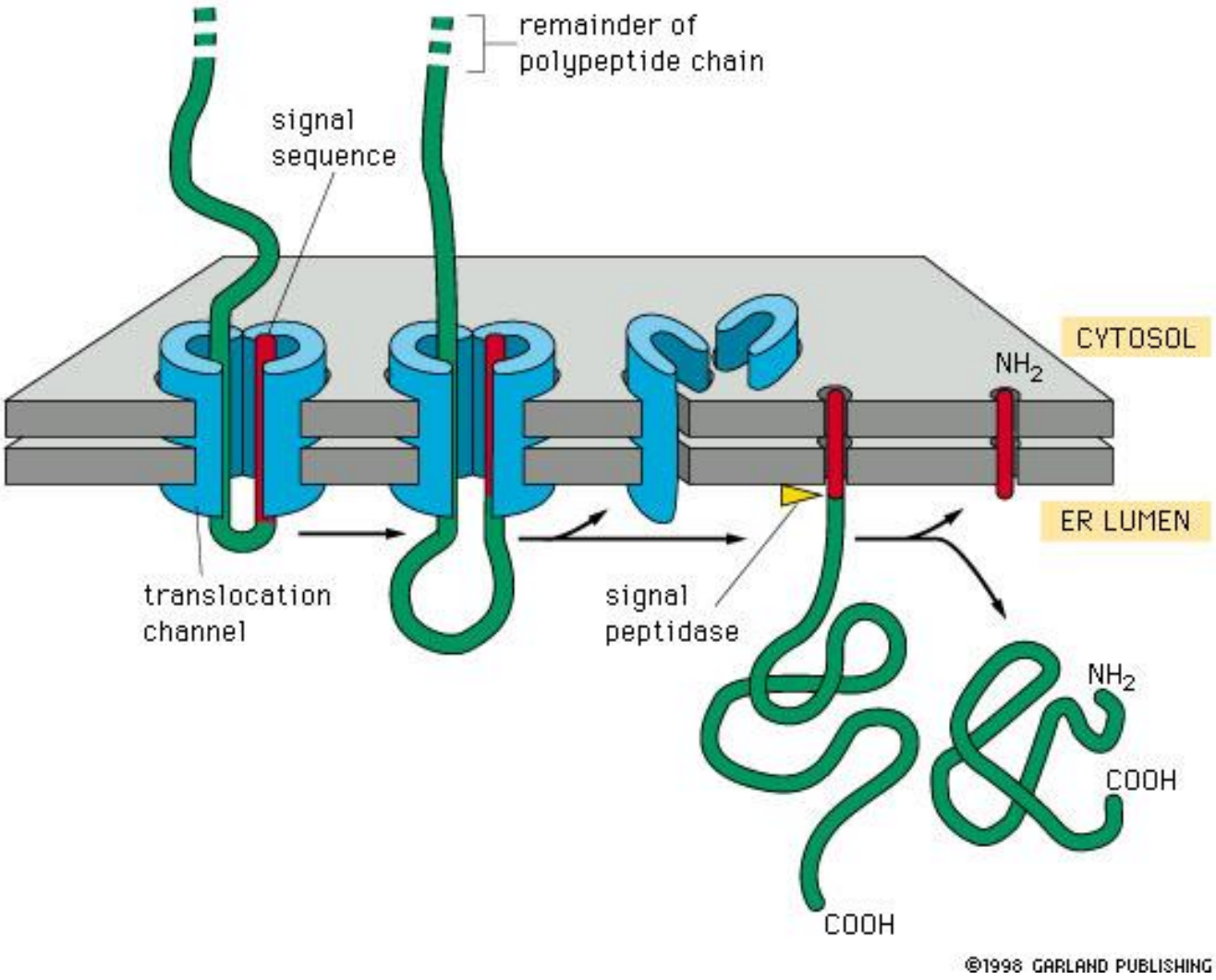
Διαμεμβρανικά τμήματα



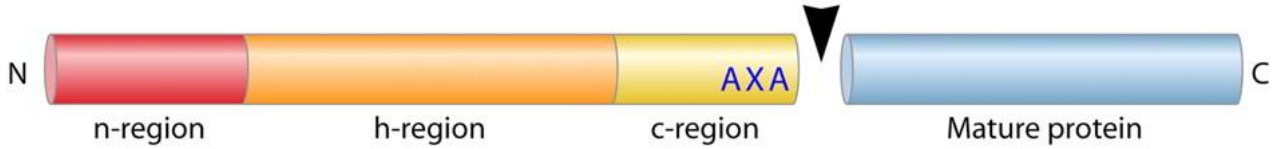
Διαμεμβρανικές πρωτεΐνες




A
SPase
N
n
h
MKWVTFISLLFSSAYS
albumin
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVS
prolactin
MGQIVTMFEALPHIIDEVININVIIVLIIITSIKAVYNFATCGILALVSFLFLAGRSCG
LCMV-GPC
MPNHQSGSPTGSSDLLLSGKKQRPHLALRRKRRREMRKINRKVRRMNLAPIKEKTAWQHLQALISEAEEVLKTSQTPQNSLTLFLALLSVLGPPPVTG
MMTV-Rem
B
ribosome with
nascent protein
SRP
ER
membrane
translocon
SRP
receptor
SPase
N
C
signal sequences affect protein targeting,
insertion and their cleavage
in a qualitative and quantitative manner
1) processing of signal peptide
by signal peptide peptidase
SPP
2) membrane-inserted
signal peptide
3) release of a signal peptide
from the ER membrane

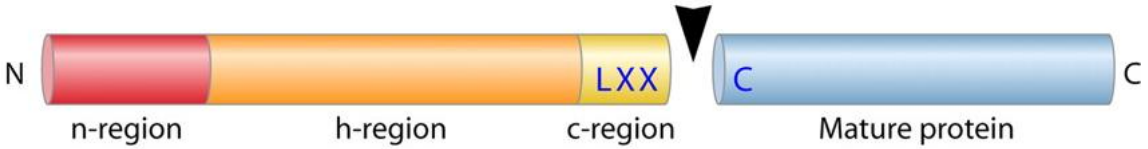


Bacterial signal peptide (SPI)

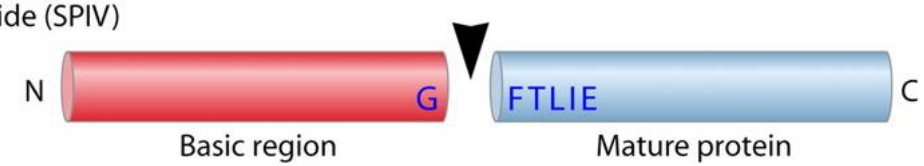


Tat signal peptide (SPI)

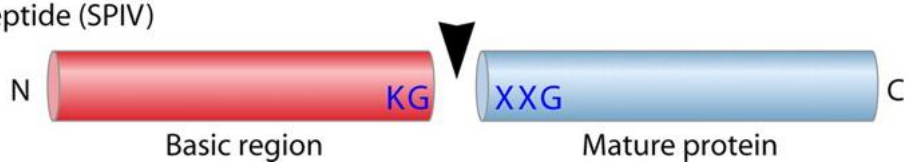
Lipoprotein signal peptide (SPII)



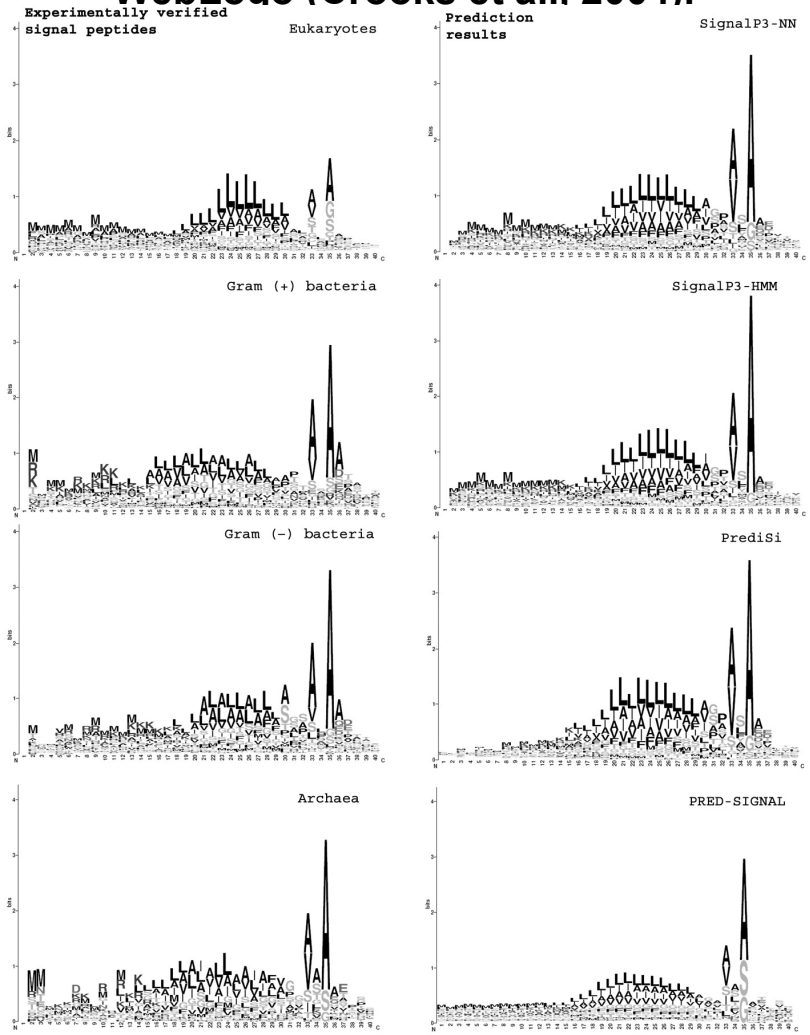
Bacterial prepilin signal peptide (SPIV)



Archaeal preflagellin signal peptide (SPIV)



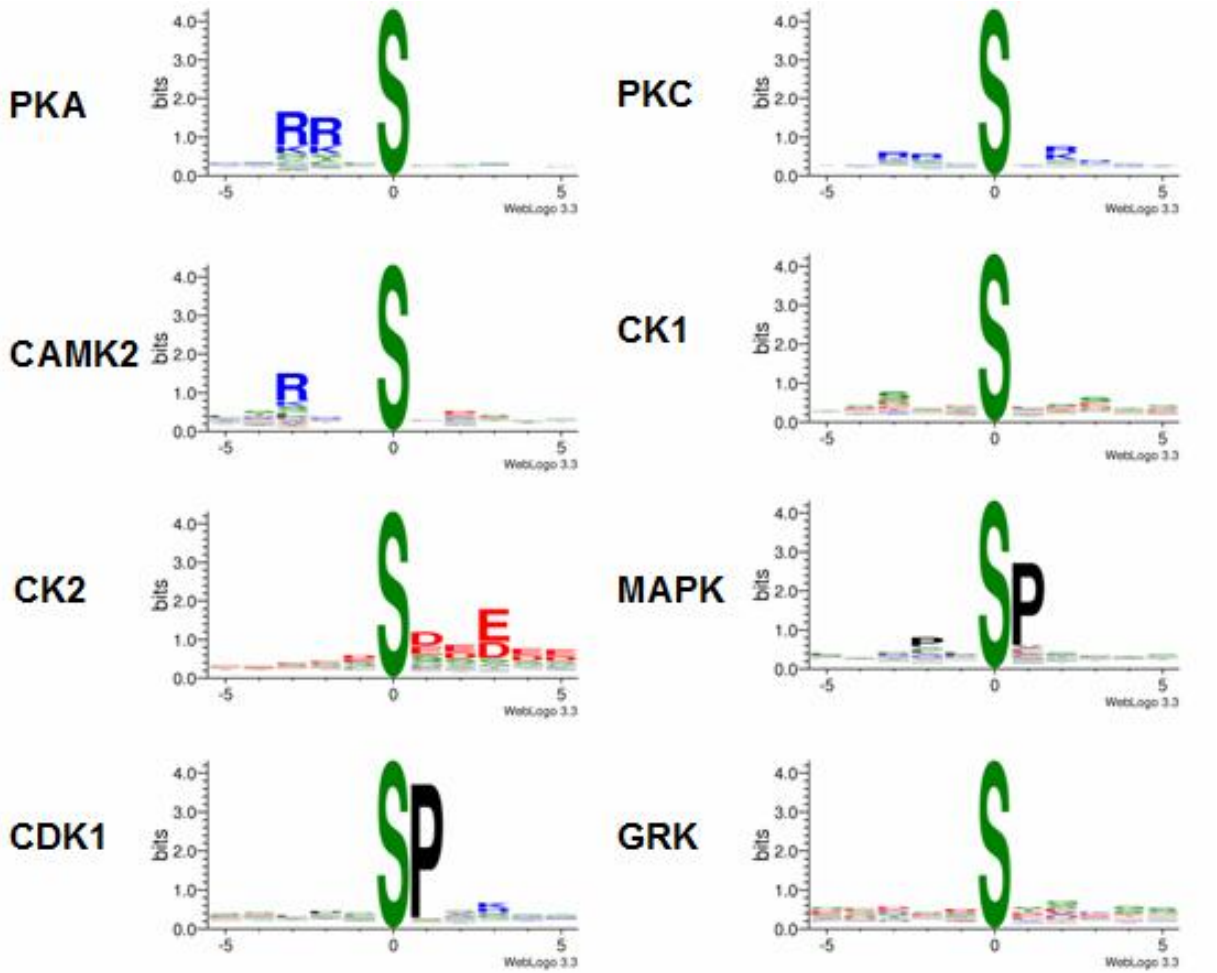
Left panel (from top to bottom): the sequence logos of experimentally verified eukaryal, gram-positive, gram-negative and archaeal signal peptides (SPs), respectively, produced by WebLogo (Crooks et al., 2004).



P.G. Bagos et al. Protein Engineering, Design and Selection 2009;22:27-35

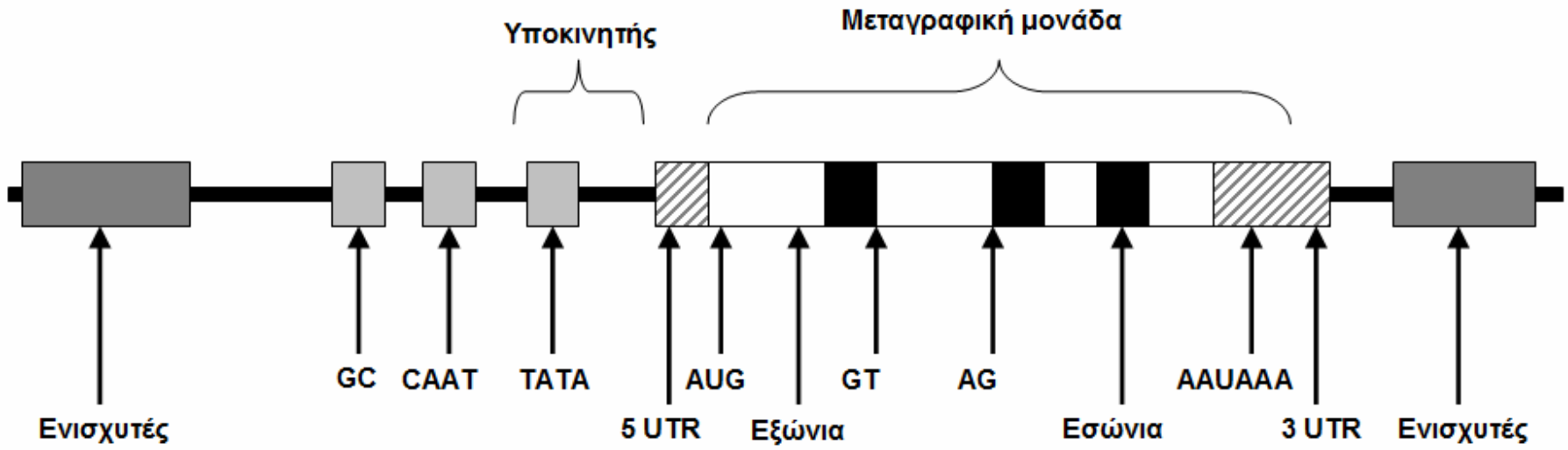
peds
protein engineering design & selection

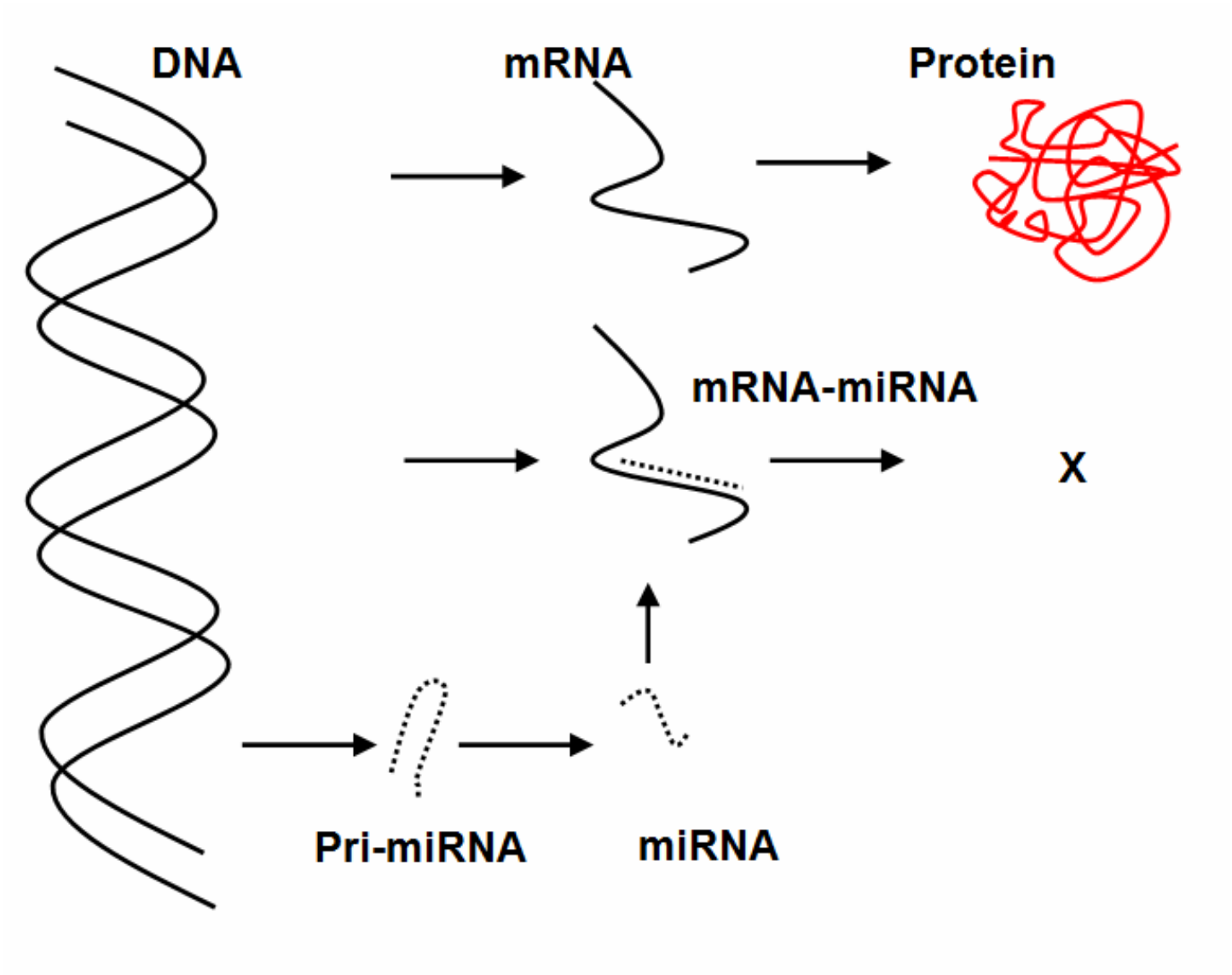
Μετα-μεταφραστικές τροποποιήσεις



DNA/RNA

- Έύρεση γονιδίων
- Εύρεση υποκινητών
- Σημεία συρραφής
- TIS
- Poly-A
- miRNA


Υποκινητής
Μεταγραφική μονάδα
GC
CAAT
TATA
AUG
GT
AG
AAUAAA
Ενισχυτές
5 UTR
Εξώνια
Εσώνια
3 UTR
Ενισχυτές


DNA
mRNA
Protein
mRNA-miRNA
X
Pri-miRNA
miRNA