# Πρόλογος

Οι πρωτεΐνες αποτελούν βασικές δομικές και λειτουργικές μονάδες των οργανισμών. Οι λειτουργίες που επιτελούν εξαρτώνται κυρίως από την τρισδιάστατη δομή τους, η οποία εξαρτάται, με τη σειρά της, πρωτίστως από την αλληλουχία των αμινοξέων που την αποτελούν και δευτερευόντως από ένα σύνολο άλλων (εξωγενών) παραγόντων. Ένα από τα σημαντικότερα ζητήματα που ενδιαφέρουν το χώρο της βιολογίας (και της βιοπληροφορικής) είναι η πρόβλεψη της τρισδιάστατης δομής των πρωτεϊνών, με μόνο δεδομένο την ακολουθία των αμινοξέων τους. Σύμφωνα, μάλιστα, με το περιοδικό IEEE Computer, τεύχος Ιουλίου 2002, σελ. 27, *το ιερό δισκοπότηρο της υπολογιστικής βιολογίας είναι είτε η πρόβλεψη αλληλουχία-δομή-λειτουργία, είτε ο υπολογισμός της αντιστοίχισης γονότυπου-φαινότυπου* . Αν και υπάρχουν τεχνικές για την εύρεση της ακριβούς δομής μιας πρωτεΐνης στο χώρο, αυτές κοστίζουν πολύ και είναι ιδιαίτερα χρονοβόρες, γι' αυτό και καταφεύγουμε σε υπολογιστικές μεθόδους. Λόγοι που ενδιαφερόμαστε για τέτοιες προβλέψεις είναι η κατανόηση της λειτουργίας μιας πρωτεΐνης, χωρίς να καταφεύγουμε σε πειραματικές μεθόδους, η τεχνητή δημιουργία νέων πρωτεϊνών και ενζύμων (γενετική, φαρμακευτική), καθώς και η εξελικτική πορεία των ειδών.

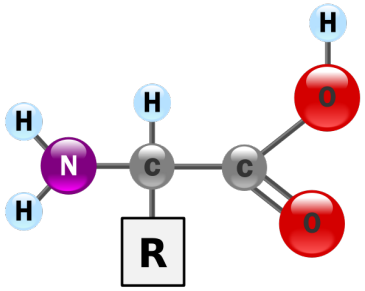
## Βασικά στοιχεία Βιολογίας

Πολλές από τις έννοιες που ακολουθούν προέρχονται από τον χώρο αυτής της επιστήμης. Για το λόγο αυτό, θα αναφέρουμε κάποια εισαγωγικά περί βιολογίας και συγκεκριμένα αναφορικά με το πρόβλημα που εξετάζουμε, ώστε να υπάρχει το υπόβαθρο για την κατανόηση των όσων θα ακολουθήσουν.

## Αμινοξέα (amino acids)

Τα αμινοξέα (amino acids) είναι η βασική, δομική μονάδα από την οποία αποτελούνται οι πρωτεΐνες. Ένα αμινοξύ αποτελείται από:

1. Μια αμινομάδα, $NH_2$
2. Μια καρβοξυλομάδα, COOH και
3. Μια ομάδα R



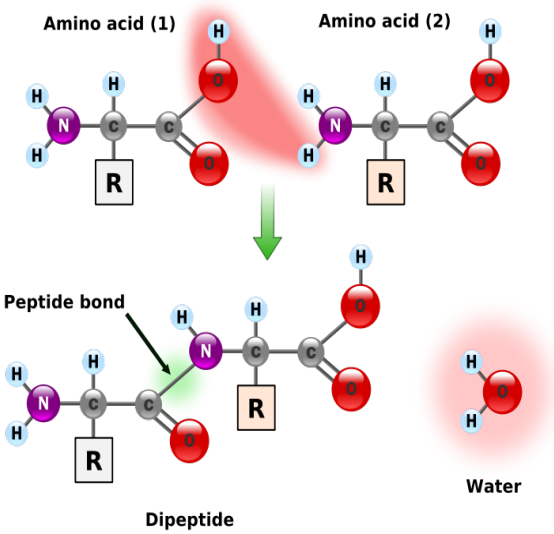
*Εικόνα 1: Αμινοξύ*

Η τυπική μορφή ενός αμινοξέος φαίνεται στην Εικόνα 1. Συνολικά υπάρχουν 20 διαφορετικά αμινοξέα, τα οποία μπορούμε να εντάξουμε σε διάφορες κατηγορίες με βάση κάποιες ιδιότητές τους (βλ. Παράρτημα 1). Το τμήμα εκείνο το οποίο διαφοροποιείται μεταξύ αμινοξέων είναι η ομάδα R, υπάρχουν δηλαδή 20 διαφορετικά R, αλλά σε όλα η σύνδεση της ομάδας R με τα υπόλοιπα μέρη γίνεται με τον ίδιο τρόπο.

## Πολυπεπτιδικές Αλυσίδες

Τα αμινοξέα συνδέονται μεταξύ τους με συγκεκριμένο τρόπο, ο οποίος φαίνεται στην Εικόνα 2. Περισσότερα από δύο αμινοξέα ενώνονται μεταξύ τους με τον ίδιο ακριβώς τρόπο, σχηματίζοντας κατά αυτό τον τρόπο μακροσκελείς αλυσίδες. Οι αλυσίδες αυτές ονομάζονται πολυπεπτιδικές, επειδή οι δεσμοί που σχηματίζονται κατά την ένωση των αμινοξέων

ονομάζονται πεπτιδικοί. Η ένωση δύο (2) αμινοξέων δημιουργεί ένα διπεπτίδιο, τριών (3) τριπεπτίδιο, τεσσάρων (4) τετραπεπτίδιο, μέχρι 20 ολιγοπεπτίδιο, ενώ περισσοτέρων πολυπεπτίδιο.



*Εικόνα 2: Δημιουργία διπεπτίδιου*

Όπως είπαμε, πολλά αμινοξέα ενώνονται μεταξύ τους σχηματίζοντας πολυπεπτιδικές αλυσίδες. Ωστόσο, δεν είναι πρωτεΐνες όλες οι πολυπεπτιδικές αλυσίδες, επειδή δεν εμφανίζουν όλες κάποια χαρακτηριστική λειτουργία. Επίσης, δεν έχουν όλες οι πρωτεΐνες το ίδιο πλήθος αμινοξέων. Έτσι, αν υποθέσουμε ότι έχουμε μια αλληλουχία αμινοξέων μήκους N, το πλήθος των πιθανών συνδυασμών είναι $20^N$ αλλά δεν δίνουν όλοι αυτοί οι συνδυασμοί κάποια πρωτεΐνη. Εμείς, θα αναφερόμαστε από εδώ και πέρα σε πρωτεΐνες μόνο, γιατί αυτές είναι που μας ενδιαφέρουν.

## Ν-ταγής δομή πρωτεΐνης

Η λειτουργία που έχει μια πρωτεΐνη, καθορίζεται (σε τελική ανάλυση) από την αλληλουχία των αμινοξέων που την αποτελούν (το οποίο ενισχύει την άποψη ότι δεν δίνουν όλες οι αλληλουχίες πρωτεΐνες). Η αλληλουχία αυτή των αμινοξέων ονομάζεται **πρωτοταγής** δομή μίας πρωτεΐνης.

Μεταξύ αμινοξέων μιας πρωτεΐνης, που βρίσκονται κοντά στην αλυσίδα (γειτονικά), αναπτύσσονται ασθενείς δεσμοί (ηλεκτροστατικοί, δεσμοί υδρογόνου). Αυτό έχει ως αποτέλεσμα τη δημιουργία αναδιπλώσεων (folds) και περιελίξεων (twists), το οποίο γίνεται για λόγους ευστάθειας της αλυσίδας. Η μορφή που παίρνει η πρωτεΐνη ως αποτέλεσμα αυτών των αναδιπλώσεων ονομάζεται **δευτεροταγής** δομή. Τρεις είναι οι κύριες κατηγορίες στις οποίες εντάσσουμε δευτεροταγείς δομές: *a*-helix, *β*-sheets και *coils*. Η τελευταία περιλαμβάνει ουσιαστικά ό,τι δεν ανήκει στις άλλες δύο.

Πέραν της δευτεροταγούς δομής, η πρωτεΐνη κάνει επιπλέον αναδιπλώσεις στο χώρο, ανεξαρτήτως γειτονικών ή μη αμινοξέων. Οι αναδιπλώσεις αυτές είναι αποτέλεσμα των χαρακτηριστικών που έχουν οι ομάδες R (υδρόφοβες, υδρόφιλες) και του περιβάλλοντα χώρου. Με το σύνολο αυτών των αναδιπλώσεων, η πρωτεΐνη αποκτά την **τριτοταγή** δομή της. Ωστόσο, στη λειτουργία που επιτελεί μια πρωτεΐνη δεν συμμετέχει όλη η πολυπεπτιδική αλυσίδα που την αποτελεί. Το τμήμα που είναι υπεύθυνο για τη λειτουργία μιας πρωτεΐνης καλείται **ενεργό μέρος**.

Τέλος, σε περίπτωση που μια πρωτεΐνη αποτελείται από πολλές πολυπεπτιδικές αλυσίδες, ο τρόπος που οι τριτοταγές δομές τους συνδυάζονται και συνδέονται καλείται **τεταρτοταγής** δομή.

# Πρόβλεψη Δομής Πρωτεϊνών

## *Ορισμός του προβλήματος*

Αντικείμενο του συγκεκριμένου προβλήματος είναι η ανακάλυψη της 3D δομής μιας πρωτεΐνης, έχοντας ως μοναδικό δεδομένο την ακολουθία των αμινοξέων της (πρωτοταγής δομή). Αυτό που πρέπει να επισημάνουμε σε αυτό το σημείο, προς αποφυγήν παρεξηγήσεων, είναι πως όταν μιλάμε για 3D δομή μιας πρωτεΐνης δεν αναφερόμαστε αυστηρά στην τριτοταγή δομή της. Αν και οι δύο όροι συχνά χρησιμοποιούνται ως ταυτόσημοι, ακόμα και η δευτεροταγής δομή μιας πρωτεΐνης είναι 3D, δηλαδή στον χώρο. Το να βρεθεί η πρωτοταγής δομή μιας πρωτεΐνης είναι αρκετά εύκολο σήμερα, και σίγουρα κατά πολύ ευκολότερο από τον προσδιορισμό της τριτοταγούς της δομής, η οποία και μας ενδιαφέρει για τον προσδιορισμό της λειτουργίας της. Υπάρχουν φυσικές μέθοδοι για τον προσδιορισμό της τριτοταγούς δομής μιας πρωτεΐνης, όπως π.χ. κρυσταλλογραφία ακτίνων-Χ ή NMR πρισματοσκοπία, αλλά αυτές προϋποθέτουν ότι η πρωτεΐνη είναι σε κρυσταλλική μορφή, μια διαδικασία που είναι ιδιαίτερα χρονοβόρα και κοστοβόρα και η οποία, με τα σημερινά μέσα, δεν μπορεί να αυτοματοποιηθεί. Ενδεικτικό των δυσκολιών που ενέχουν αυτές οι τεχνικές είναι ότι ενώ γνωρίζουμε την πρωτοταγή δομή εκατομμυρίων πρωτεϊνών, μόλις για μερικές χιλιάδες από αυτές γνωρίζουμε την τριτοταγή τους.

## *Γιατί μας ενδιαφέρει*

Η αναζήτηση αυτή έχει προκύψει ως αποτέλεσμα της πεποίθησης ότι *η λειτουργία που επιτελεί μια πρωτεΐνη καθορίζεται από την τριτοταγή δομή της (Anfinsen, 1973)*. Ο λόγος είναι ότι η τριτοταγής δομή μιας πρωτεΐνης καθορίζει και το ενεργό τμήμα της. Γνωρίζοντας, λοιπόν, την τριτοταγή δομή της μπορούμε να εντοπίσουμε το ενεργό της μέρος και να καθορίσουμε, σε δεύτερο στάδιο, τη λειτουργία που επιτελεί (μέσω πχ. προσομοίωσης).

Με βάση αυτή την πληροφορία μπορούμε να κατασκευάζουμε νέες πρωτεΐνες και ένζυμα (φαρμακευτική, γενετική). Επίσης, μπορούμε να εξάγουμε γνώση όσον αφορά την εξελικτική πορεία των ειδών αλλά κυρίως να έχουμε πληροφορία για το τι μπορεί να γίνει σε περίπτωση που συμβεί (εκούσια ή ακούσια) μια μετάλλαξη. Όταν μιλάμε για μετάλλαξη, εννοούμε πως ένα τμήμα της πρωτεΐνης έχει διαφορετική αλληλουχία αμινοξέων απ' ότι η κανονική πρωτεΐνη. Το διαφοροποιημένο μέρος της πρωτεΐνης μπορεί να οδηγεί τελικά σε διαφορετική τριτοταγή δομή ή να επηρεάσει το ενεργό της τμήμα, με φυσικό επακόλουθο τη διαφοροποίηση στη λειτουργία που αυτή επιτελεί. Είναι χαρακτηριστικό ότι γνωστές ασθένειες, όπως π.χ. το Αλτζχάιμερ και το σύνδρομο των τρελών αγελάδων ,ή αναπνευστικά προβλήματα οφείλονται σε κακή αναδίπλωση των πρωτεϊνών στο χώρο.

## *Γιατί είναι δύσκολο πρόβλημα*

Κατά πρώτον, μια βασική δυσκολία που αντιμετωπίζεται και καταφεύγουμε στη μηχανική μάθηση είναι ότι οι σημερινές φυσικές μέθοδοι που έχουμε στη διάθεσή μας για την αναγνώριση της τριτοταγούς δομής μιας πρωτεΐνης είναι ιδιαίτερα ακριβές και χρονοβόρες, όπως ήδη αναφέρθηκε.

Κατά δεύτερον, το να υπολογίσουμε όλες τις πιθανές αναδιπλώσεις

που πραγματοποιεί μια πρωτεΐνη είναι υπολογιστικά ακριβό. Δεδομένου ότι υπάρχουν 3 διαφορετικές κατηγορίες δευτεροταγούς δομής, στις οποίες μπορεί να ανήκει ένα αμινοξύ, για μια πρωτεΐνη που αποτελείται από 100 αμινοξέα, δίνει ως αποτέλεσμα $3^{100}$ διαφορετικές πιθανές καταστάσεις. Ωστόσο, οι πρωτεΐνες αναδιπλώνονται με σκοπό να ελαχιστοποιήσουν τη συνολική ενέργειά τους (ώστε να μη χρειάζεται να ενωθούν με στοιχεία από τον περιβάλλοντα χώρο). Το να αναζητήσουμε το ολικό ελάχιστο είναι ένα πρόβλημα βελτιστοποίησης, το οποίο ανήκει στα *NP-hard*, όπου ακόμα και με χρήση στοχαστικών μοντέλων είναι ιδιαίτερα δύσκολο. Αν και υπάρχει ερευνητική δραστηριότητα στο χώρο αυτό, συνήθως ο προσανατολισμός είναι σε μικρές πρωτεΐνες ώστε να μπορούν να γίνουν οι απαραίτητοι υπολογισμοί.

Άλλοι λόγοι που καθιστούν δύσκολο το συγκεκριμένο πρόβλημα είναι ότι ορισμένες πρωτεΐνες δέχονται βοήθεια και από εξωτερικούς παράγοντες, όπως πχ. άλλες πρωτεΐνες που ονομάζονται *chaperons (Martin and Hartl, 1997)*. Επίσης, όπως ήδη αναφέραμε, σημαντικό ρόλο παίζει και ο περιβάλλων χώρος της πρωτεΐνης (πχ. οξύτητα του διαλύματος στο οποίο βρίσκεται) καθώς και συγκεκριμένες ιδιότητες των ομάδων R που την αποτελούν.

## *Τεχνικές που χρησιμοποιούνται*

Η πρόβλεψη της δομής των πρωτεϊνών είναι ένας τομέας με μεγάλο ερευνητικό ενδιαφέρον, ιδιαίτερα στην μετά-γονιδιακή περίοδο (post-genomic era), δηλαδή μετά την αποκρυπτογράφηση του γονιδιώματος του ανθρώπου (2003). Χαρακτηριστικό είναι ότι αλγοριθμικές προσπάθειες προτάθηκαν από το 1970 [54], ενώ τεχνικές μηχανικής μάθησης χρονολογούνται από το 1992 [4] (ίσως και νωρίτερα).

Υπάρχουν διάφοροι τρόποι που έχουν προταθεί και χρησιμοποιούνται για να αντιμετωπίσουν το συγκεκριμένο πρόβλημα. Κατηγορίες τις οποίες μπορούμε να διακρίνουμε είναι:

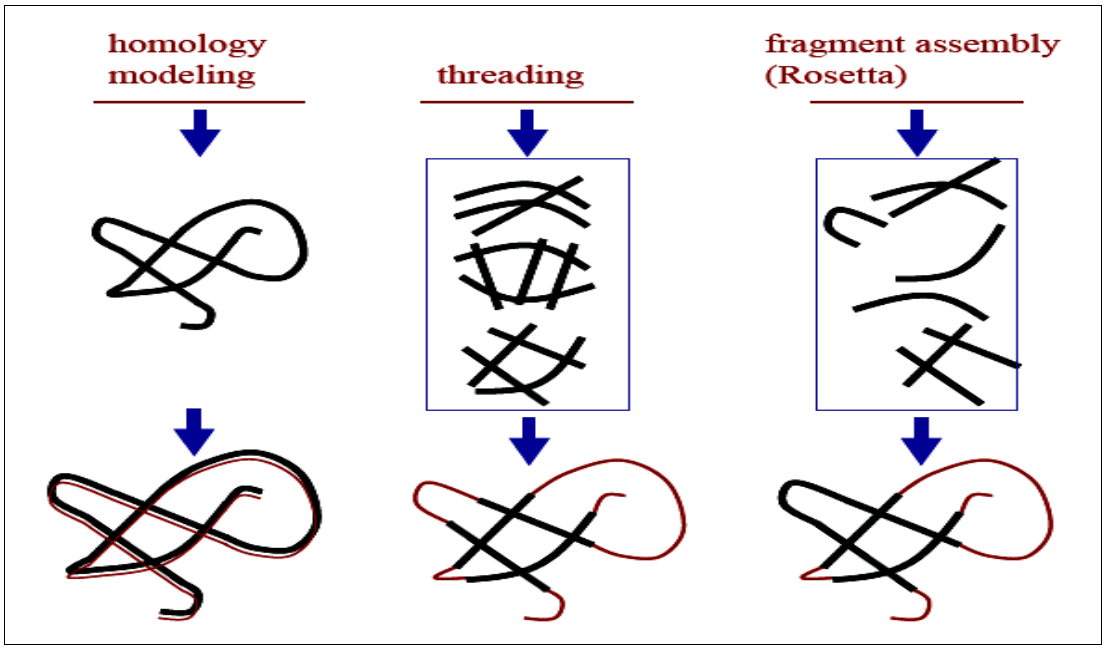
- Μοντελοποίηση ομολόγων (*homology modeling*)
- *Protein threading*[1]
- *Ab initio* ή *de novo* μέθοδος
- *Lego approach*
- Μηχανική Μάθηση (*Machine Learning techniques*)

Σύμφωνα με την μοντελοποίηση ομολόγων [55], δύο *ομόλογες* πρωτεΐνες θα παρουσιάζουν και παρόμοιες (αν όχι ίδιες) 3D δομές. Ως ομόλογες ορίζονται δύο πρωτεΐνες όταν έχουν υψηλό βαθμό ομοιότητας στην πρωτοταγή δομή τους. Η τεχνική, η οποία χρησιμοποιεί sequence alignemt αλγορίθμους, δίνει πολύ καλά αποτελέσματα όταν η ομοιότητα των δύο πρωτεϊνών είναι ιδιαίτερα μεγάλη, αλλά αντιμετωπίζει προβλήματα όπως: *i*) πρέπει να είναι γνωστή η 3D δομή της πρωτεΐνης ως προς την οποία το ερώτημα έχει υψηλή ομοιότητα, *ii*) δεν ισχύει πάντα η υπόθεση ότι ομόλογες πρωτεΐνες θα έχουν και παρόμοια 3D δομή και *iii*) υπάρχουν περιπτώσεις όπου μη ομόλογες πρωτεΐνες έχουν παρόμοια 3D δομή. Ο λόγος για το τελευταίο είναι ότι μέσω της εξελικτικής διαδικασίας, έχουν γίνει τόσες αλλαγές ώστε οι πρωτεΐνες παύουν να είναι ομόλογες, ωστόσο εξακολουθούν να μοιράζονται κοινή δομή στο χώρο, επειδή είναι (μακρινά) συγγενικές. Η 2η

1 Για τον όρο δεν βρέθηκε κάποια καλή ελληνική μετάφραση, γι' αυτό παρατίθεται αυτούσιος, όπως εμφανίστηκε στη βιβλιογραφία

τεχνική είναι παρόμοια με την 1η, αλλά διαφοροποιείται ως προς τον τρόπο που ορίζει την ομοιότητα μεταξύ δύο πρωτεϊνών. Συγκεκριμένα, χρησιμοποιεί συναρτήσεις βαθμολόγησης, ώστε να αξιολογήσει κατά πόσο μια ακολουθία αμινοξέων μπορεί να πάρει μια συγκεκριμένη 3D δομή, από αυτές που είναι γνωστές.

Η τεχνική *Lego* (ή αλλιώς fragment assembly ) στηρίζεται στην παραδοχή που γίνεται στις δύο προηγούμενες τεχνικές. Η διαφορά της από τις άλλες δύο είναι ότι σπάει τη νέα αλληλουχία σε τμήματα και προσπαθεί να βρει για το κάθε ένα μια ακολουθία που την προσομοιάζει καλύτερα, για την οποία είναι γνωστή η 3D δομή της. Κατόπιν, συνθέτει τις δομές αυτές δίνοντας το τελικό αποτέλεσμα. Στις εικόνες που ακολουθούν, εμφανίζονται οι 3 τεχνικές τις οποίες μόλις αναφέραμε, με τη σειρά: Homology modeling, Protein Threading, Lego.




*Εικόνα 3: Τρεις διαφορετικές προσεγγίσεις για την πρόβλεψη της δομής των πρωτεϊνών*

Η *ab initio* τεχνική στηρίζεται στις φυσικοχημικές ιδιότητες των αμινοξέων και λαμβάνοντας υπόψη πιθανοτικά (στοχαστικά) μοντέλα, προσπαθεί να βρει την κατάσταση στην οποία η πρωτεΐνη παρουσιάζει την ελάχιστη ενέργεια. Προσπαθεί, δηλαδή, να βρει το ολικό ελάχιστο στο πρόβλημα βελτιστοποίησης, που αναφέραμε προηγουμένως. Τέτοια ερευνητικά προγράμματα είναι για παράδειγμα τα *Folding@home*, *Rosetta@home* και πολλά άλλα. Τα συστήματα αυτά είναι πάντα κατανεμημένα και παράλληλα (distributed parallel systems), λόγω των πολλών υπολογισμών, με την έννοια ότι κάθε σύστημα που συμμετέχει αναλαμβάνει ξεχωριστή δουλειά, ώστε να γίνεται ταυτόχρονη αναζήτηση για πολλές πρωτεΐνες. Επίσης, επειδή χρησιμοποιούνται στοχαστικά μοντέλα, γίνεται χρήση πιθανοτικών αλγορίθμων Monte Carlo, το οποίο αυξάνει (προφανώς) τον απαιτούμενο χρόνο (λόγω των πολλαπλών εκτελέσεων).

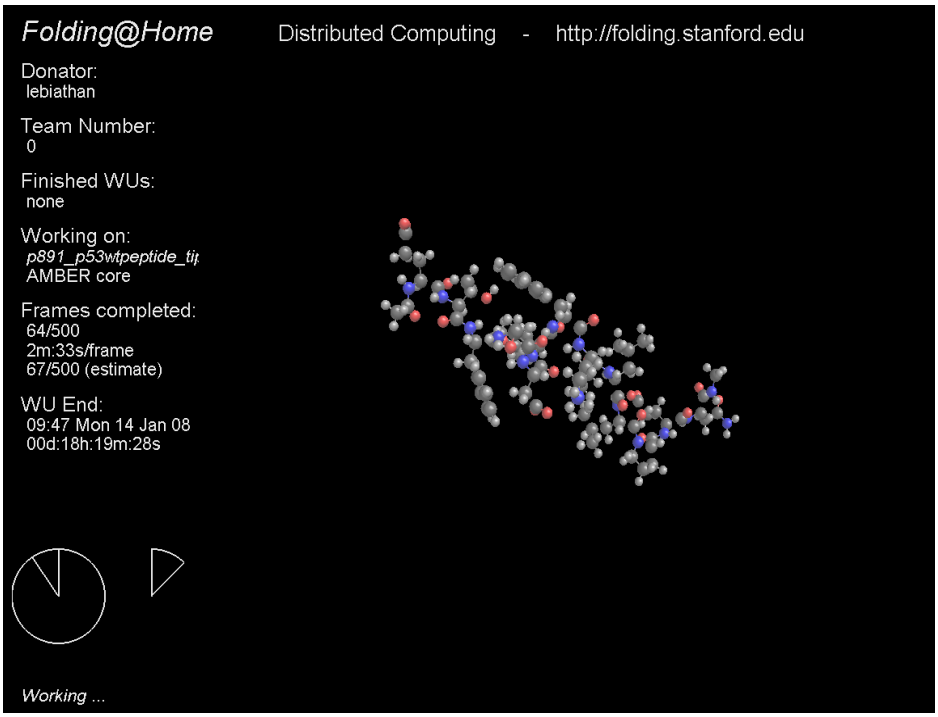
Τέλος, για την αντιμετώπιση του προβλήματος έχουν προταθεί και χρησιμοποιηθεί (και εξακολουθούν ακόμα) τεχνικές μηχανικής μάθησης. Ουσιαστικά, με τη μηχανική μάθηση προσπαθούμε να κατανοήσουμε τις

περιπτώσεις για τις οποίες προκύπτει κάποια αναδίπλωση, έχοντας ως δεδομένα εκπαίδευσης τις πρωτεΐνες για τις οποίες μας είναι ήδη γνωστή η 3D δομή της, με τις τεχνικές που αναφέραμε πιο πάνω. Κατόπιν, χρησιμοποιούμε τη γνώση αυτή, για να προβλέψουμε τη δομή αγνώστων πρωτεΐνών. Τη συγκεκριμένη τεχνική θα εξετάσουμε εδώ σε βάθος.

## *Η συμβολή της Μηχανικής Μάθησης (ΜΜ)*

Η μηχανική μάθηση παρουσιάζει σημαντικά πλεονεκτήματα έναντι των άλλων τεχνικών που χρησιμοποιούνται για την πρόβλεψη της δομής των πρωτεϊνών. Σε αντίθεση με τις πρώτες τρεις (ομόλογες, threading και Lego), δεν απαιτεί να γίνει alignment μιας δεδομένης ακολουθίας ως προς άλλες που υπάρχουν στη βάση. Αν και υπάρχουν σήμερα γρήγοροι αλγόριθμοι για sequence alignment (BLAST), αυτοί εξακολουθούν να είναι πιο αργοί συγκριτικά με την κατηγοριοποίηση βάσει ενός μοντέλου ΜΜ. Ωστόσο, το βασικότερο, ίσως, μειονέκτημα των τεχνικών αυτών είναι ότι θεωρούν πως πράγματι υπάρχει μια *ομόλογη* ακολουθία στη βάση. Σε περίπτωση που κάτι τέτοιο δεν ισχύει οι τεχνικές αυτές αποτυγχάνουν.

Αν η κατηγοριοποίηση μιας ακολουθίας αμινοξέων με χρήση sequence alignment είναι αργή (συνήθως οι αλγόριθμοι είναι τετραγωνικής πολυπλοκότητας), τότε σίγουρα η χρήση της ab initio μεθόδου χρειάζεται πολύ παραπάνω χρόνο. Προφανώς τέτοιες προσεγγίσεις ενδιαφέρουν σε μεγάλο βαθμό την επιστημονική κοινότητα, αλλά όσον αφορά πρακτικές εφαρμογές (πχ. φαρμακευτική), όπου θέλουμε να ξέρουμε γρήγορα την πρόβλεψη της 3D δομής, δεν μπορούμε να περιμένουμε τόση ώρα[2]. Στη συνέχεια φαίνεται ένα screenshot που τραβήχτηκε κατά την εκτέλεση του προγράμματος Folding@Home.




*Εικόνα 4: Το Folding@Home επί το έργον*

2 Ενδεικτικά αναφέρουμε ότι για τον υπολογισμό ενός τυπικού Working Unit (WU) του Folding@Home, σε έναν Intel Core Duo2,2GHz, 1GB RAM, με πλήρη διαθεσιμότητα του μηχανήματος χρειάστηκαν 20 ώρες. Ένα WU **ΔΕΝ** είναι ολόκληρη πρωτεΐνη.

Επιπλέον, εκτός από τα πλεονεκτήματα αυτά, σύμφωνα με αποτελέσματα που έχουν προκύψει από σχετικές μελέτες, με χρήση της ΜΜ μπορούμε να εξάγουμε συμπεράσματα και για πρωτεΐνες που εμφανίζονται σε διαφορετικά είδη οργανισμών. Τέτοιες πληροφορίες είναι ιδιαίτερα χρήσιμες για την κατανόηση της εξέλιξης των ειδών, εφόσον αυτές αξιολογηθούν και ερμηνευτούν κατάλληλα.

Στη συνέχεια αναφέρουμε διάφορους τρόπους με τους οποίους η μηχανική μάθηση έχει χρησιμοποιηθεί στα πλαίσια της πρόβλεψης 3D δομής πρωτεϊνών. Επειδή το πρόβλημα της πρόβλεψης τριτοταγούς δομής είναι πολύ δύσκολο, η ερευνητική δραστηριότητα από πλευράς ΜΜ επικεντρώνεται στην πρόβλεψη δευτεροταγούς δομής. Η πρόβλεψη της δευτεροταγούς δομής είναι ένα από τα σημαντικότερα βήματα προς την πρόβλεψη και ανακάλυψη της τριτοταγούς δομής (και άρα της πλήρους 3D δομής) μιας πρωτεΐνης. Για παράδειγμα, έχοντας προβλέψει τη δευτεροταγή δομή της πρωτεΐνης, μπορούμε να χρησιμοποιήσουμε ab initio τεχνικές για τον προσδιορισμό της τριτοταγούς δομής (λαμβάνοντας όμως υπόψη μας τις προβλέψεις που αφορούν τη δευτεροταγή δομή, εξοικονομώντας χρόνο).

Στα πλαίσια του προβλήματος που εξετάζουμε, οι συχνότερα χρησιμοποιούμενες τεχνικές της ΜΜ είναι:

1. Τεχνητά Νευρωνικά Δίκτυα
2. Support Vector Machines (SVM's)
3. Hidden Markov Models
4. Decision Trees

Εκτός από τις μεθόδους αυτές, έχουν χρησιμοποιηθεί και κάποιες ακόμα, αλλά και ως συνήθως γίνεται συνδυασμός τους .Στη συνέχεια θα αναφερθούμε σε εφαρμογές τους και στον τρόπο που η κάθε μία το πραγματοποίησε. Ωστόσο, η περιγραφή θα είναι συνοπτική, καθώς πιο πολύ ενδιαφέρει να δούμε τον τρόπο εφαρμογής και τις διαφορετικές προσεγγίσεις που ακολουθήθηκαν παρά τις ακριβείς λεπτομέρειες της κάθε προσέγγισης.

Κάτι το οποίο είναι ανεξάρτητο της τεχνικής με την οποία προσεγγίζεται το πρόβλημα είναι το μέτρο αξιολόγησης του παραγόμενου μοντέλου. Ως μέτρο αξιολόγησης χρησιμοποιείται το ποσοστό των σωστών κατηγοριοποιήσεων. Στα πλαίσια του προβλήματος αυτού, θεωρούμε ότι έχουμε σωστή κατηγοριοποίηση όταν ένα αμινοξύ που κατηγοριοποιείται σε μία από τις 3 ομάδες (α-helix, β-sheet, coil) πράγματι ανήκει σε αυτή την κατηγορία. Το πλήθος των σωστά κατηγοριοποιημένων αμινοξέων προς το πλήθος των συνολικά κατηγοριοποιημένων αμινοξέων δίνει το ποσοστό επιτυχίας του μοντέλου. Επίσης, στη βιβλιογραφία η συνήθης μέθοδος αξιολόγησης είναι το *k*-fold cross validation, αν και το *k* ποικίλει κατά περιπτώσεις. Ο τύπος που δίνει το μέτρο αξιολόγησης δίνεται στη συνέχεια.

$$Q_3 = \frac{N_\alpha + N_\beta + N_c}{N}$$

Στον τύπο αυτό, τα $N_\alpha$, $N_\beta$ και $N_c$ είναι το πλήθος των σωστά κατηγοριοποιημένων αμινοξέων που ανήκουν σε α-helix, β-sheet ή σε άλλο

τύπο αντίστοιχα. Ο παρονομαστής N συμβολίζει το συνολικό πλήθος των κατηγοριοποιημένων αμινοξέων. Ωστόσο, εκτός από το $Q_3$ υπάρχουν και άλλα μέτρα τα οποία χρησιμοποιούνται. Μια λίστα αυτών μπορούν να βρεθούν στο [2], αν και δεν είναι οι μόνες που έχουν προταθεί [4, 3].

Ένα άλλο στοιχείο το οποίο δεν έχει να κάνει με συγκεκριμένο είδος μάθησης αφορά τα δεδομένα εκπαίδευσης. Αυτό το οποίο συνήθως αναφέρεται στη βιβλιογραφία είναι ότι τα δεδομένα εκπαίδευσης και τα δεδομένα για το testing είναι *ανεξάρτητα* [2]. Με την έννοια αυτή εννοούμε ότι δεν υπάρχουν ομόλογες πρωτεΐνες ανάμεσα στα δύο data sets. Ο λόγος που γίνεται αυτό είναι για να μην είναι τετριμμένη η αναζήτηση που πραγματοποιείται.

Τέλος, η συνήθης τακτική που χρησιμοποιείται, επίσης ανεξαρτήτως μεθόδου, είναι ότι για κάθε αμινοξύ εξετάζεται ένα παράθυρο, μεγέθους W γύρω του. Αυτό σημαίνει ότι για κάθε αμινοξύ ελέγχεται μια περιοχή με κέντρο το ίδιο και περιλαμβάνει όλα εκείνα τα αμινοξέα που βρίσκονται στο διάστημα (-W/2, W/2) θέσεις μακριά από αυτό. Ο λόγος για αυτή την προσέγγιση είναι ότι, όπως αναφέρθηκε, οι αναδιπλώσεις στο χώρο που αφορούν στη δευτεροταγή δομή μιας πρωτεΐνης γίνονται σε τοπικό επίπεδο, δηλαδή αφορούν αμινοξέα που βρίσκονται κοντά μεταξύ τους. Αυτή είναι και η βασικότερη προσέγγιση του προβλήματος, όταν μιλάμε για πρόβλεψη της 3D δομής μιας πρωτεΐνης από την πρωτοταγή της ακολουθία. Ο τρόπος που μοντελοποιείται αυτή η πληροφορία αλλάζει με βάση το μοντέλο το οποίο κατασκευάζουμε. Άλλες πληροφορίες οι οποίες συχνά λαμβάνονται υπόψη αφορούν σε: *i*) πολλαπλά sequence alignments, το οποίο συχνά αναφέρεται και ως εξελικτική πορεία , *ii*) φυσικοχημικές ιδιότητες των αμινοξέων (υδροφιλία, υδροφοβία, πολικότητα) , *iii*) ομόλογες πρωτεΐνες, ομάδες και υπερ-ομάδες (families and super-families) στις οποίες εντάσσονται ως προς την δευτεροταγή δομή τους, *iv*) στατικά αντί για ολισθαίνοντα παράθυρα (πιο παλιά τεχνική) και *v*) άλλα εμπειρικά στοιχεία που έχουν προκύψει από επιστήμονες του χώρου της βιολογίας (πχ. προτιμήσεις τάσεις για τρόπους συνδέσεων των αμινοξέων). Παρ' όλα αυτά, η βασικότερη πεποίθηση επικρατεί ότι το σημαντικότερο μέρος της πληροφορίας για τον τρόπο που προκύπτει η 3D δομή μιας πρωτεΐνης βρίσκεται στην ακολουθία των αμινοξέων της.

### *Τεχνητά Νευρωνικά Δίκτυα*

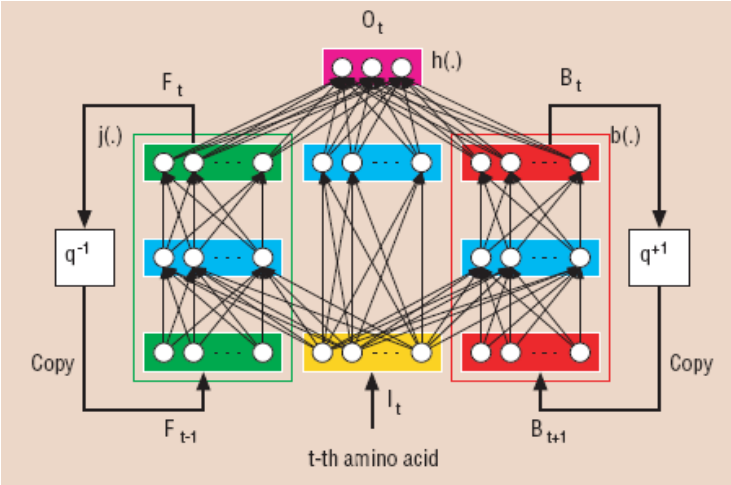
Τα Τεχνητά Νευρωνικά Δίκτυα (*artificial neural networks*), έχουν εφαρμοστεί πολλές φορές για την αντιμετώπιση του προβλήματος της πρόβλεψης της δομής πρωτεϊνών. Ο λόγος για την μεγάλη εφαρμογή που βρήκαν στον τομέα είναι ότι πρόκειται για μηχανισμούς με μεγάλη ικανότητα μάθησης, όπως αντικατοπτρίζεται από τα σχετικά ποσοστά επιτυχίας τους. Επίσης, τα ΤΝΔ είναι γνωστό ότι είναι πολύ καλά στην κατανόηση και εκμάθηση μη γραμμικών δεδομένων, όπως είναι η 3D δομή των πρωτεϊνών, δεδομένης της πρωτοταγούς ακολουθίας τους.

Όσον αφορά τα ΤΝΔ, η πληροφορία που δίνεται ως είσοδος αφορά την πρωτοταγή δομή της πρωτεΐνης, ενώ δίνεται και η αντίστοιχη έξοδος που είναι η κατηγορία (α-helix, β-sheet, coil) στην οποία ανήκει ένα αμινοξύ. Στη συνέχεια, με δεδομένη μια νέα ακολουθία την οποία θέλουμε να προβλέψουμε, η έξοδος είναι συνήθως η ομάδα στην οποία ανήκει κάθε αμινοξύ ή η πιθανότητα να ανήκει στη συγκεκριμένη ομάδα.

Η πιο απλή εφαρμογή ΤΝΔ για την πρόβλεψη της δευτεροταγούς δομής πρωτεϊνών είναι το nnPredict (*http://alexander.compbio.ucsf.edu/~nomi/nnpredict.html*). Πρόκειται στην ουσία για ένα δίκτυο 2 επιπέδων: ένα επίπεδο εισόδου και ένα εξόδου, χωρίς κανένα ενδιάμεσο επίπεδο. Η εφαρμογή που παρατίθεται στο συγκεκριμένο site υλοποιεί το ΤΝΔ με προκαθορισμένη χρήση των βαρών των νευρώνων, σύμφωνα με εξωτερική μάθηση του μοντέλου.

Πιο πολύπλοκα συστήματα που στηρίζονται σε νευρωνικά δίκτυα είναι αυτό των Rost και Sander (1993), το οποίο συνδυάζει εκτός από την απλή πληροφορία της πρωτοταγούς δομής και πολλαπλή στοίχιση ακολουθιών αμινοξέων (multiple sequence alignment), δίνοντας έτσι υψηλότερα ποσοστά επιτυχίας. Πρόκειται δηλαδή για ένα συνδυασμό μηχανικής μάθησης και μοντελοποίησης ομολόγων, δείχνοντας κατ' αυτό τον τρόπο ότι οι διάφορες τεχνικές δεν είναι αποκλειόμενες μεταξύ τους, αλλά περισσότερο συμπληρωματικές. Υπενθυμίζουμε σε αυτό το σημείο ότι με το multiple sequence alignment γίνεται και εύρεση της εξελικτικής πορείας.

Μια ακόμα σημαντική προσέγγιση του προβλήματος με χρήση ΤΝΔ είναι αυτή των Baldi και Pollastri. Το αντίστοιχο σύστημα ονομάζεται BRNN, από τα αρχικά των λέξεων *Bidirectional Recurrent Neural Network*. Όπως αναφέρεται χαρακτηριστικά και στο [1], το πρόβλημα που προσπαθεί να αντιμετωπίσει η προσέγγιση αυτή, είναι ότι τα συνήθη (μέχρι τότε - 2002) ΤΝΔ ήταν feed-forward, με μέγεθος παραθύρων από 9 μέχρι 15 αμινοξέα. Λαμβάνοντας υπόψη, όμως, μόνο τοπικά παράθυρα γύρω από το κάθε αμινοξύ, αγνοείται χρήσιμη πληροφορία σε πιο ευρύ επίπεδο. Αυτό έχει ως αποτέλεσμα τη μείωση της απόδοσης του προσδιορισμού β-sheets σε μία πρωτεΐνη, αφού αυτά δεν είναι πάντα σε τοπικό επίπεδο, και τα σχετικά παράθυρα μπορεί να αποτύχουν στο να τα βρουν. Η λύση που προτάθηκε για το σκοπό αυτό είναι ένα ΤΝΔ με ανατροφοδότηση (η έξοδος χρησιμοποιείται μετά από κάποια βήματα και ως είσοδος), λαμβάνοντας κατ' αυτό τον τρόπο υπόψη και πληροφορία σε πιο ευρύ επίπεδο και όχι μόνο τοπικά. Μια μετέπειτα έκδοση του συγκεκριμένου μοντέλου, κάνει και αυτό χρήση πληροφορίας σχετικά με ομόλογες πρωτεΐνες (PSI-BLAST). Η αρχιτεκτονική του συστήματος BRNN φαίνεται στην Εικόνα 5.



*Εικόνα 5: Αρχιτεκτονική συστήματος BRNN*

Συνδυασμός ΤΝΔ χρησιμοποιείται και από άλλα συστήματα [56], όπου γίνεται αξιοποίηση 2 ΤΝΔ. Στην πρώτη φάση, χρησιμοποιείται ένα ΤΝΔ με ολισθαίνον (sliding) παράθυρο 17 αμινοξέων για την πρόβλεψη των τριών κυρίων κατηγοριών. Σε δεύτερο επίπεδο, αφού έχει προκύψει το αποτέλεσμα από το 1ο ΤΝΔ, χρησιμοποιείται ένα ολισθαίνον παράθυρο 19 αμινοξέων σε δεύτερο ΤΝΔ, για την επιβεβαίωση και εκλέπτυνση των αποτελεσμάτων.

Τέλος, άλλες προσεγγίσεις του προβλήματος χρησιμοποιούν πληθώρα ΤΝΔ, της τάξης πολλών εκατοντάδων κατά περιπτώσεις, τα οποία εκπαιδεύονται σε ανεξάρτητα σύνολα δεδομένων και κατόπιν τα αποτελέσματά τους συνδυάζονται.

*Support Vector Machines*

Παρεμφερείς με την τεχνική των ΤΝΔ είναι εκείνες όπου χρησιμοποιούνται τα SVM's για να προσδιορίσουν τη δομή των πρωτεϊνών. Και στο συγκεκριμένο τομέα έχουν προταθεί ιδιαίτερα πολλές προσεγγίσεις, χαρακτηριστικές από τις οποίες αναλύουμε στη συνέχεια. Επίσης, όσον αφορά τα SVM, οι προσεγγίσεις δεν διαφοροποιούνται μόνο ως προς τις αρχιτεκτονικές των συστημάτων, όπως είδαμε προηγουμένως, αλλά και στα σχετικά kernel functions. Έτσι, προσφέρουν ένα σημαντικό πεδίο έρευνας όσον αφορά τον τομέα αυτό.

Ο κυριότερος λόγος που χρησιμοποιούνται τα SVM's για το συγκεκριμένο πρόβλημα είναι η πολύ καλή απόδοσή τους, καθώς παρουσιάζουν επιδόσεις που ξεπερνούν κατά περιπτώσεις αυτές των ΤΝΔ. Άλλα πλεονεκτήματα είναι οι σχετικά λίγες παράμετροι που απαιτούν για τη ρύθμισή τους, η ευελιξία τους στο να ενσωματώσουν διαφορετικές συναρτήσεις πυρήνα (kernel functions) λαμβάνοντας υπόψη πολλά στοιχεία ιδιότητες, η ανοχή τους σε σφάλματα και η εγγενής τους αποφυγή για υπερ-μοντελοποίηση.

Μία τυπική προσέγγιση του προβλήματος που αναλύουμε, με χρήση SVM είναι αυτή που παρουσιάζεται στο [β edge strands]. Ωστόσο, το πρόβλημα που προσπαθεί να αντιμετωπίσει δεν αφορά τη γενικότητα της πρόβλεψης της δομής πρωτεϊνών, αλλά εστιάζει συγκεκριμένα σε β-sheets. Παρ' όλα αυτά, η μεθοδολογία είναι η ίδια και για το σύνολο του προβλήματος (α-helix, coil). Έτσι, για παράδειγμα, δεν χρησιμοποιείται μόνο η ακολουθία των αμινοξέων, αλλά και κάποιες συγκεκριμένες ιδιότητες τους, καθώς επίσης και παρατηρήσεις που έχουν γίνει, όπως πχ. ότι κάποια αμινοξέα (ή ακολουθίες αμινοξέων) δεν ευνοούν τη δημιουργία β-sheets στις άκρες της πρωτεΐνης. Μια παρόμοια, πιο γενική θεώρηση του προβλήματος μπορεί κανείς να βρει στο [secondary structure prediction with support vector machines.pdf], δείχνοντας μάλιστα ότι τα SVM παράγουν ικανοποιητικά αποτελέσματα συγκριτικά με άλλες τεχνικές και σε αντίθεση με πιο παλιές πεποιθήσεις.

Οι τρόποι με τους οποίους έχει χρησιμοποιηθεί αυτή η μέθοδος ποικίλουν. Πολλές φορές κατασκευάζονται SVM's τύπου one-versus-one (ένας-προς-ένας) όσον αφορά τις κλάσεις κατηγοριοποίησης, ενώ άλλες φορές είναι της μορφής one-versus-rest (ένας προς όλους). Στην πρώτη κατηγορία κατασκευάζονται όλες οι πιθανές δυάδες των κατηγοριοποιητών που μπορούν να προκύψουν, δηλαδή H/C, C/E, E/H. Στη δεύτερη ο κατηγοριοποιητής μαθαίνει αν κάτι ανήκει στην κατηγορία ή όχι, δηλαδή H/~H, E/~E, C/~C, όπου ~ είναι η άρνηση της αντίστοιχης κατηγορίας. Ο 2ος τρόπος δεν θεωρείται τόσο καλός, καθώς τα δεδομένα δεν έχουν πάντα την ίδια συχνότητα εμφανίσεων (κατανομή), με αποτέλεσμα ο κατηγοριοποιητής να μεροληπτεί (σε κάποιες περιπτώσεις).

Ανάλογα με τα SVM που κατασκευάζονται για την αντιμετώπιση του προβλήματος, έχουν προταθεί διάφορες τεχνικές για τον προσδιορισμό τελικά της κλάσης στην οποία ανήκει ένα αμινοξύ μιας πρωτεΐνης. Για παράδειγμα έχει προταθεί η χρήση ψηφοφορίας, λαμβάνοντας υπόψη την

πλειοψηφία που προκύπτει από τα αποτελέσματα όλων των δυαδικών κατηγοριοποιητών. Σε περίπτωση που προκύπτει ισοπαλία ,το αμινοξύ θεωρείται ότι ανήκει στην κλάση C (coil). Άλλη τεχνική είναι η SVM_MAX_D, όπου χρησιμοποιούνται οι ένα-προς-άλλους κατηγοριοποιητές και τελικά επιλέγεται εκείνη η κλάση η οποία έχει τη μέγιστη θετική τιμή, ενώ οι αρνητικές τιμές δε δίνουν καμία πληροφορία (και συνεπώς δε λαμβάνονται υπόψη). Υπάρχει επίσης η SVM_Representative, που είναι ίδια με την SVM_MAX_D, μόνο που λαμβάνονται υπόψη και αρνητικές τιμές. Συγκεκριμένα, επιλέγεται ο κατηγοριοποιητής που παρουσιάζει τη μέγιστη *απόλυτη* τιμή από το υπερ-επίπεδο που επιστρέφεται. Κατόπιν, με βάση αν η τιμή είναι θετική η αρνητική, επιλέγεται και η αντίστοιχη κλάση. Έτσι, μια θετική τιμή αντιπροσωπεύει την πρώτη κλάση του κατηγοριοποιητή, ενώ μια αρνητική τιμή τη δεύτερη [3].

Απόπειρες για τη βελτίωση των kernel functions που χρησιμοποιούνται έχουν γίνει πάρα πολλές κατά καιρούς. Σκοπός των προσεγγίσεων αυτών είναι να βρεθούν οι συναρτήσεις εκείνες για τις οποίες ο διαχωρισμός θα γίνεται πιο εύκολα από τα SVM's. Έτσι, η προσπάθεια σε αυτό το επίπεδο επικεντρώνεται στην ενσωμάτωση περισσότερης πληροφορίας σχετικά με την πρωτοταγή ακολουθία (και άλλα στοιχεία της πρωτεΐνης) στις συναρτήσεις πυρήνα [23, 3]. Τέτοιες πληροφορίες αφορούν τις φυσικοχημικές ιδιότητες των αμινοξέων (υδρόφιλα υδρόφοβα ,πολικά μη πολικά, κ.ά.). Σχετικοί πίνακες με τέτοια πληροφορία υπάρχουν πάρα πολλοί (πχ. BLOSUM62, MAMMOTH)

Εναλλακτικά, έχουν προταθεί συνδυασμοί από SVM, με την έννοια όμως της συνεργασίας. Συγκεκριμένα, όπως φαίνεται και στο [nguyen.pdf], υλοποιείται ένας κατηγοριοποιητής που αποτελείται από δύο επίπεδα SVM, όπου η έξοδος του πρώτου είναι η είσοδος στο δεύτερο. Η τεχνική αυτή είναι παρόμοια με την τεχνική που χρησιμοποιείται στα νευρωνικά από το [56], όπου στο πρώτο επίπεδο λαμβάνονται υπόψη τα γειτονικά αμινοξέα, ενώ στο δεύτερο λαμβάνεται υπόψη και η δομή της πρωτεΐνης στο σύνολο.

*Hidden Markov Model*

Μια άλλη τεχνική που χρησιμοποιείται είναι τα Hidden Markov Models. Ένα HMM μπορεί να θεωρηθεί ως το πιο απλό δυναμικό Bayesian μοντέλο. Συγκεκριμένα, πρόκειται για ένα στατιστικό μοντέλο όπου οι παράμετροι, όμως, είναι άγνωστες προς τον χρήστη. Τα HMM βασίζονται στην ιδιότητα ότι η πιθανότητα μιας κατάστασης είναι ανεξάρτητη απ' το τι έχει συμβεί στο παρελθόν, πρόκειται δηλαδή για συστήματα χωρίς μνήμη .Με δεδομένη αυτή την παράμετρο, η δυσκολία των HMM έγκειται στο να βρεθούν οι παράμετροι που είναι κρυφές στον χρήστη. Το θετικό των μοντέλων αυτών είναι ότι μπορούν να αναπαρασταθούν γραφικά, το οποίο βοηθάει την κατανοησιμότητά τους σε σημαντικό βαθμό.

Μια χαρακτηριστική δουλειά που χρησιμοποιεί τα HMM για την αντιμετώπιση του προβλήματος της πρόβλεψης μπορεί να βρει κανείς στο. Οι παράμετροι του μοντέλου στην περίπτωση αυτή δεν καθορίζονται με κάποιο συγκεκριμένο τρόπο, αλλά προκύπτουν μέσα από μια διαδικασία *trial and error* . Επίσης, μια web εφαρμογή που πραγματοποιεί πρόβλεψη δομής πρωτεϊνών με HMM μπορεί να βρεθεί στο [53]. Για τον προσδιορισμό των πιθανοτήτων στα κρυφά επίπεδα, κατασκευάστηκαν πολλά διαφορετικά

μοντέλα και λαμβάνοντας υπόψη συγκεκριμένα κριτήρια, όπως η απόδοση στο Q3, το Bayesian Information Criterion και η στατιστική απόσταση μεταξύ των μοντέλων. Το μοντέλο εκπαιδεύεται με βάση ένα σύνολο δεδομένων και επαληθεύεται. Επίσης, η συγκεκριμένη ερευνητική ομάδα χρησιμοποίησε και ένα σύνολο δεδομένων για τον έλεγχο του overfitting. Συγκεκριμένα, κράτησε ένα μέρος του συνόλου εκπαίδευσης εκτός, και αφού κατασκευάστηκε το μοντέλο ερευνήθηκε η δυνατότητα πρόβλεψης στα δεδομένα τα οποία δεν είχε δει. Παρ' όλα αυτά, η πρόβλεψη των α-helix είναι ιδιαίτερα καλή, ενώ αντιθέτως των β-sheets και των coils όχι.

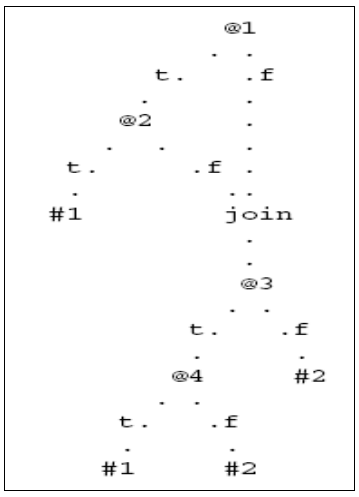
Επίσης, επειδή τα HMM στηρίζονται σε θεωρία πιθανοτήτων κάποιος μπορεί να χρησιμοποιήσει εκ των προτέρων γνώση (κάτι το οποίο δεν έγινε στη συγκεκριμένη περίπτωση), όσον αφορά τα δεδομένα που έχει στη διάθεσή του, με σκοπό τη βελτίωση των αποτελεσμάτων. Ωστόσο, σε τέτοιες περιπτώσεις πρέπει να δίνεται μεγαλύτερη προσοχή σε θέματα overfitting.

*Δέντρα Απόφασης*

Όπως είναι γνωστό, τα δέντρα απόφασης είναι ιδιαίτερα κατανοητά από τον άνθρωπο. Μπορούν πολύ εύκολα να μετατραπούν σε κανόνες, οι οποίοι είναι ακόμα πιο κατανοητοί και μπορούν να οδηγήσουν σε χρήσιμα συμπεράσματα. Τέτοιοι κανόνες στα πλαίσια της βιοπληροφορικής βοηθούν στην κατανόηση τόσο του λόγου για τον οποίο λήφθηκε μία απόφαση (πχ. γιατί ένα αμινοξύ ανήκει σε α-helix) όσο και σε πιο ευρύ επίπεδο, για την κατανόηση της εξελικτικής πορείας.

Όπως και πιο πάνω, ένας ενδεικτικός τρόπος χρήσης των δέντρων απόφασης φαίνεται στο [β-edge strands]. Και πάλι, αυτή η περίπτωση έχει να κάνει συγκεκριμένα με την πρόβλεψη των β-sheets, αλλά η διαδικασία είναι η ίδια και γενικεύεται και για τις άλλες κλάσεις.

Είναι χαρακτηριστικό ότι τα Decision Trees χρησιμοποιήθηκαν από πολύ νωρίς για την αντιμετώπιση του μοντέλου. Συγκεκριμένα, χρησιμοποιήθηκαν οι γράφοι απόφασης (Decision Graphs), το 1992, όπως μπορεί κανείς να δει στο [4]. Ο λόγος ήταν η κατανοησιμότητα που προσφέρουν έναντι των άλλων τεχνικών. Οι γράφοι απόφασης είναι μια γενίκευση των δέντρων απόφασης, όπου επιτρέπονται και συζεύξεις και διαζεύξεις στον ίδιο κόμβο. Ένα παράδειγμα ενός δέντρου απόφασης φαίνεται στο σχήμα δίπλα. Η τεχνική αυτή, όμως, είναι ιδιαίτερα παλιά και τα αποτελέσματά της δεν μπορούν να είναι συγκρίσιμα με αυτά των σημερινών τεχνικών (ενδεικτικά αναφέρουμε ότι είχε μέσο ποσοστό επιτυχίας 54%), κυρίως επειδή τότε δεν υπήρχαν και πολλά δεδομένα εκπαίδευσης. Επίσης, στην περίπτωση είχαν χρησιμοποιηθεί και επιπλέον τεχνικές Ωστόσο, μας δείχνει ότι πρόκειται για ένα πρόβλημα που απασχολεί την επιστημονική κοινότητα ακόμα, 16 χρόνια αργότερα! Σκοπός της χρησιμότητας αυτής της μεθόδου ήταν κυρίως να ληφθούν κανόνες τους οποίους θα μπορούσε να κατανοήσει και να ερμηνεύσει ο άνθρωπος.




Είναι σημαντικό να αναφέρουμε ότι όσον αφορά τους παραγόμενους κανόνες, όπως προκύπτει και από τα αποτελέσματα της 1ης περίπτωσης,

αυτοί είναι αρκετά κοντά στην γνώση που έχει μέχρι σήμερα η επιστημονική κοινότητα, το οποίο αυξάνει την εμπιστοσύνη μας στη χρήση της μηχανικής μάθησης.

*Υβριδικά Μοντέλα*

Όπως πάντα, υπάρχουν και τα υβριδικά μοντέλα στις περιπτώσεις της μηχανικής μάθησης, όπου συνδυάζονται διαφορετικές μέθοδοι ΜΜ. Χαρακτηριστικό είναι ότι σε συγκεκριμένες περιπτώσεις [rule generation for] συνδυάζονται τα SVM με τα δέντρα απόφασης. Ο λόγος είναι ότι τα πρώτα παρουσιάζουν υψηλά ποσοστά επιτυχίας, ενώ τα δεύτερα παρουσιάζουν ευνόητους και αντιληπτούς κανόνες. Έτσι, χρησιμοποιούνται τα SVM για την δημιουργία προβλέψεων και κατόπιν, με κατάλληλη επιλογή στιγμιοτύπων από τα δεδομένα εκπαίδευσης (τα οποία κατηγοριοποιήθηκαν σωστά), εκτελούνται τα δέντρα απόφασης πάνω τους, ώστε να προκύψουν οι κανόνες. Αν και η απόδοση των δεύτερων (decision trees) στην περίπτωση αυτή δεν είναι τόσο καλή όσο των SVM, παρ' όλα αυτά παράγουν ικανοποιητικούς κανόνες, οι οποίοι είναι και το ζητούμενο.

Άλλες υβριδικές προσεγγίσεις, όπως αυτή στο [bagging like], αφορούν σε περιπτώσεις όπου υπάρχουν πάρα πολλά δεδομένα τα οποία δε μπορούν να είναι αυτομάτως όλα στη μνήμη. Ο τρόπος που προτείνεται για την αντιμετώπιση του συγκεκριμένου προβλήματος είναι η προσομοίωση της τεχνικής του Bagging ( bootstrap aggregation ). Κατά τη διαδικασία του bagging αυτό το οποίο γίνεται είναι να παράγονται διαφορετικά dataset, και να δημιουργούνται χωριστοί κατηγοριοποιητές, οι οποίοι τελικά συνδυάζονται με κάποιο τρόπο (κάποιοι από αυτούς ήδη αναφέρθηκαν). Έτσι, η προσομοίωση που προτείνεται έχει να κάνει με τη δημιουργία N διαφορετικών κατηγοριοποιητών, όπου ο κάθε ένας έχει μάθει ένα 1/N πλήθος των αρχικών δεδομένων. Κατόπιν, η τελική απόφαση προκύπτει με συνδυασμό των κατηγοριοποιητών.

Άλλες υβριδικές μέθοδοι αφορούν στη γενίκευση ήδη υπαρχόντων μεθόδων. Η αντιμετώπιση τεχνικών που παρουσιάζουν παρόμοια χαρακτηριστικά υπό ένα κοινό πρίσμα, παρέχει τη δυνατότητα να χρησιμοποιούνται εν δυνάμει όλα τα μοντέλα τα οποία πληρούν τα χαρακτηριστικά αυτά. Έτσι, για παράδειγμα, έχουν γίνει γενίκευσεις των Bayesian δικτύων, τα οποία μπορούν, εν δυνάμει να προσομοιαστούν και από τα ΤΝΔ και από τα Hidden Markov Models (ειδικά με δεδομένο ότι και τα δεύτερα στηρίζονται σε πιθανότητες). Το γεγονός ότι τα ΗΜΜ έχουν και κρυφά επίπεδα για τα οποία ο χρήστης δεν γνωρίζει τις τιμές του, προσομοιάζουν σε σημαντικό βαθμό τα ΤΝΔ, για τα οποία ο χρήστης βλέπει ένα μαύρο κουτί χωρίς να γνωρίζει τι γίνεται στο εσωτερικό του .

*Συμπεράσματα*

Συμπερασματικά, αναφέρουμε ότι η ανακάλυψη της 3D δομής των πρωτεϊνών από την ακολουθία των αμινοξέων της είναι ένα πολύ σημαντικό και δύσκολο ζήτημα, το οποίο η επιστημονική κοινότητα προσπαθεί για πολλά χρόνια να λύσει. Διαφορετικές προσεγγίσεις έχουν προταθεί και χρησιμοποιούνται για την αντιμετώπισή του, οι οποίες είναι συμπληρωματικές και όχι αντικρουόμενες. Η Μηχανική Μάθηση βοηθάει στην

αντιμετώπιση του προβλήματος με πληθώρα αλγορίθμων, οι οποίοι αυξάνουν διαρκώς την απόδοση πρόβλεψης. Αναφέρουμε, ενδεικτικά, ότι για το $Q_3$ μέτρο, το βέλτιστο ποσοστό μέχρι στιγμής είναι ~80% (από τα αποτελέσματα που βρέθηκαν).

*Μηχανική Μάθηση και πρωτεΐνες: επιπλέον θέματα*

Επιγραμματικά να αναφέρουμε ότι η Μηχανική Μάθηση, δε βοηθάει μόνο στην πρόβλεψη της 3D δομής πρωτεϊνών, αλλά χρησιμοποιείται και σε θέματα ταξινόμησης (clustering) όσον αφορά τις πρωτεΐνες.

Συγκεκριμένα, υπάρχει το πρόγραμμα SCOP, που σκοπό έχει να κατηγοριοποιήσει τις πρωτεΐνες σε ομάδες / κατηγορίες (ονομάζονται οικογένειες και υπεροικογένειες ως μετάφραση των όρων families και super-families ), με βάση κοινά τους χαρακτηριστικά ως προς την *δευτεροταγή* και *τριτοταγή* δομή τους. Σχετικές δουλειές μπορεί να βρει κάποιος στα [32, 33].

Τέλος, η Μηχανική Μάθηση έχει συμβάλει και με άλλους τρόπους, όπως μπορεί να δει κανείς στο [31], όπου χρησιμοποιείται για να κάνει επιλογή χαρακτηριστικών και να μειώσει το χώρο του προβλήματος αναζήτησης που πραγματοποιείται από το πρόγραμμα Rosetta (*ab initio* μέθοδος).

# Παράρτημα 1 *Πίνακας Αμινοξέων*

| Amino Acid | 3-Letter | 1-Letter | Side chain polarity | Side chain acidity or basicity | Hydropathy index |
|---|---|---|---|---|---|
| Alanine | Ala | A | nonpolar | neutral | 1.8 |
| Arginine | Arg | R | polar | basic (strongly) | -4.5 |
| Asparagine | Asn | N | polar | neutral | -3.5 |
| Aspartic acid | Asp | D | polar | acidic | -3.5 |
| Cysteine | Cys | C | polar | neutral | 2.5 |
| Glutamic acid | Glu | E | polar | acidic | -3.5 |
| Glutamine | Gln | Q | polar | neutral | -3.5 |
| Glycine | Gly | G | nonpolar | neutral | -0.4 |
| Histidine | His | H | polar | basic (weakly) | -3.2 |
| Isoleucine | Ile | I | nonpolar | neutral | 4.5 |
| Leucine | Leu | L | nonpolar | neutral | 3.8 |
| Lysine | Lys | K | polar | basic | -3.9 |
| Methionine | Met | M | nonpolar | neutral | 1.9 |
| Phenylalanine | Phe | F | nonpolar | neutral | 2.8 |
| Proline | Pro | P | nonpolar | neutral | -1.6 |
| Serine | Ser | S | polar | neutral | -0.8 |
| Threonine | Thr | T | polar | neutral | -0.7 |
| Tryptophan | Trp | W | nonpolar | neutral | -0.9 |
| Tyrosine | Tyr | Y | nonpolar | neutral | -1.3 |
| Valine | Val | V | nonpolar | neutral | 4.2 |

*Πίνακας 1: Τα 20 αμινοξέα και ιδιότητες (Πηγή: wikipedia)*

# Βιβλιογραφία

## Δημοσιεύσεις Βιβλία


1. ***A Machine Learning Strategy for Protein Analysis*** ,P. Baldi, G. Pollastri, in *IEEE Intelligent Systems in Biology 2002*
2. **Review: Protein Secondary Structure Prediction Continues to Rise** , Burkhard Rost, in *Journal of Structural Biology 134, 204 218 (2001)*
3. **Improved Protein Secondary Structure Prediction Using Support Vector Machine With a New Encoding Scheme and an Advanced Tertiary Classifier** ,Hae-Jin Hu, Yi Pan, Robert Harrison and Phang C. Tai, in *IEEE Transactions on Nanobioscience, Vol. 3, No. 4, December 2004*
4. **A Decision Graph Explanation of Protein Secondary Structure Prediction** ,David L. Dowe, Jonathan Oliver, Lloyd Allison, Christopher S. Wallace & Trevor I. Dix, *Technical Report 92/163 June 1992*
5. **Rule Generation for Protein Secondary Structure Prediction With Support Vector Machines and Decision Tree** , Jieyue He, Hae-Jin Hu, Robert Harrison, Phang C. Tai and Yi Pan, in *IEEE Transactions on Nanobioscience, Vol. 5, No. 1, March 2006*
6. **Analysis of an optimal hidden Markov model for secondary structure prediction** ,Juliette Martin, Jean-François Gibrat and François Rodolphe, *BMC Structural Biology 2006, 6:25, December 2006*
7. **Edge strands in protein structure prediction and aggregation** , Jennifer A. Siepen, Sheena E. Radford, and David R. Westhead, *Protein Science (2003), 12:2348 2359*
8. **Predicting protein function by machine learning on amino acid sequences a critical evaluation** ,Ali Al-Shahib, Rainer Breitling and David R Gilbert, in *BMC Genomics 2007, 8:78 March 2007*
9. **Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information** , Gianluca Pollastri, Alberto JM Martin, Catherine Mooney and Alessandro Vullo, *BMC Bioinformatics 2007, 8:201 June 2007*
10. **A machine learning information retrieval approach to protein fold recognition** ,Jianlin Cheng and Pierre Baldi, *Vol. 22 no. 12 2006, pages 1456 1463*
11. **New Machine Learning Methods for the Prediction of Protein Topologies** ,P.Baldi, Gianluca Pollastri, P.Frasconi and A.Vullo, in *Artificial Intelligence and Heuristic Methods in Bioinformatics, 2003*
12. **Machine Learning and Bioinformatics: An Introduction** ,Byoung-Tak Zhang, *cbit.snu.ac.kr/tutorial-2002/ ppt/CBIT_ML_Tutorial_bw.pdf*
13. **BaggingLike Effects for Decision Trees and Neural Nets in Protein Secondary Structure Prediction** ,Nitesh Chawla, Thomas E. Moore Jr., Kevin W. Bowyer, Lawrence O. Hall, Clayton Springer and Philip Kegelmeyer, *Workshop in Data Mining in Bioinformatics, 2001*
14. **Introduction to Protein Structure Prediction** , Mark Craven, http://www.biostat.wisc.edu/bmi576/lectures/protein-structure.pdf
15. **Proteomics: Prediction of protein structures** ,C. Stan Tsai, *An Introduction to Computational Biochemistry, 2002, ISBN: 0-471-40120-X*
16. **Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Proles** ,Gianluca Pollastri, Darisz Przybylski, Burkhard Rost, Pierre Baldi, in *Proteins*

*(Proteins) 2002, vol. 47, no2, pp. 228-235*

17. **Coupled prediction of protein secondary and tertiary structure** , Jens Meiler and David Baker, http://www.pnas.org/cgi/reprint/100/21/1205.pdf
18. **Secondary structure prediction with support vector machines** J. J. Ward, L. J. McGuffin, B. F. Buxton and D. T. Jones, *Vol. 19 no. 13 2003, pages 1650 1655*
19. **Two-Stage Multi-Class Support Vector Machines to Protein Secondary Structure Prediction** M.N. Nguyen and J.C. Rajapakse Pacific, *Symposium on Biocomputing 10:346-357(2005)*
20. **Local protein structure prediction using discriminative models** , Oliver Sander, Ingolf Sommer and Tomas Lengauer, in *BMC Bioinformatics 2006, 7:14*
21. **Ab initio methods for protein structure prediction** , Christoph Best, *Praktikum Genomorientierte Bioinformatik WS, November 2003*
22. **Protein Structure Prediction and Modelling** ,H.A.Nagarajaram, *Laboratory of Computational Biology*
23. **Hybrid SVM kernels for protein secondary structure prediction** , by Gulsah Altun, Hae-Jin Hu, Dumitru Brinza, Robert W. Harrison, Alex Zelikovsky, Yi Pan, *IEEE 2006*
24. **Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure** , Darrin P. Lewis1, Tony Jebara1 and William Stafford Noble, *Vol. 22 no. 22 2006, pages 2753 2760*
25. **SCRATCH: a protein structure and structural feature prediction server** , J. Cheng, A. Z. Randall, M. J. Sweredoski and P. Baldi, *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
26. **Machine Learning Algorithms for Protein Structure Prediction ,** *DISSERTATION, by J. Cheng*
27. **The simulated annealing method applied to protein structure prediction** , Wei-Feng Chen' ,Kai-Yang Li', JuanLiu, *in Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, 26-29 August 2004*
28. **Protein Secondary Structure Prediction Using Dynamic Programming** , Jing ZHAO1,2,3, Pei-Ming SONG1, Qing FANG1, and Jian-Hua LUO, *Acta Biochimica et Biophysica Sinica 2005, 37(3): 167 172*
29. **PREDICTION OF PROTEIN SECONDARY STRUCTURE USING GENETIC PROGRAMMING** , by Varun Aggarwal, *Summer Internship Project Report During June-July 2003*
30. **Prediction of Protein Secondary Structure from Amino Acid Sequence** , by Jen Tsi Yang 1, *Journal of Protein Chemistry, Vol. 15, No. 2, 1996*
31. **Feature selection methods for improving protein structure prediction with Rosetta** , B. Blum, M.I. Jordan, D. Kim, R. Das, P. Bradley and D. Baker, in *Advances in Neural Information Processing Systems (NIPS) 21, 2007*
32. **Improved K-Means Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property** , by Wei Zhong, Gulsah Altun, Robert Harrison, Phang C. Tai, and Yi Pan, in *IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL. 4, NO. 3, SEPTEMBER 2005 255*
33. **Decision Tree Based Information Integraion for Automated Protein Classification** ,ORHAN CAMOGLU, TOLGA CAN, AMBUJ K. SINGH, YUAN-FANG WANG, *Journal of Bioinformatics and Computational Biology*
34. **Genomic Biology and** Bioinformatics ,The Bio Team, http://www.bioteam.net/resources/guides.html

35. **Introduction to protein secondary structure Second Edition** , by Carl Branden and John Tooze, Garland Publishing
36. **Bioinformatics Computing** ,By Bryan Bergeron Publisher: Prentice Hall PTR Pub Date: November 19, 2002 ISBN: 0-13-100825-0
37. **Prediction of protein secondary structure at 80% accuracy** , by T. N. Pedersen, C. Lundegaard, M. Nielsen, H. Bohr, J. Bohr, S. Brunak, G. P. Gippert, and O. Lund, *in Proteins, 41(1):17{20, 2000*
38. **PROTEIN SECONDARY STRUCTURE PREDICTION USING SUPPORT VECTOR MACHINES, NUERAL NETWORKS AND GENETIC ALGORITHMS** , ANJUM B REYAZ-AHMED, *A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of M.Sc.*

Ιστοσελίδες

39. http://en.wikipedia.org/wiki/Amino_acid
40. http://en.wikipedia.org/wiki/Secondary_structure
41. http://en.wikipedia.org/wiki/Tertiary_structure
42. http://en.wikipedia.org/wiki/Protein_structure_prediction
43. http://folding.stanford.edu/
44. http://www.dsi.unifi.it/neural/ProteinStructure/proteomics.html
45. http://cubic.bioc.columbia.edu/papers/1998_encyclopedia/abstract.html
46. http://speedy.embl-heidelberg.de/gtsp/
47. http://boinc.bakerlab.org/rosetta/
48. http://alexander.compbio.ucsf.edu/~nomi/nnpredict.html
49. http://www.ncbi.nlm.nih.gov/blast/
50. http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/psi1.html
51. http://www.predictprotein.org/
52. http://en.wikipedia.org/wiki/Hidden_Markov_model
53. http://www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php
54. http://en.wikipedia.org/wiki/Chou-Fasman_method
55. http://en.wikipedia.org/wiki/Homology_modeling
56. http://www.compbio.dundee.ac.uk/~www-jpred/