ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ ΚΑΙ ΔΙΚΤΥΩΝ

Πρόβλεψη δομικών στοιχείων σφαιρικών πρωτεϊνών με προσομοίωση Monte Carlo σε διακριτό χώρο.

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Χριστίνας Αγγελή

Επιβλέπων καθηγητής:
Ακρίτας Αλκιβιάδης

Δεύτερο μέλος επιτροπής:
Θηραίου Τριάς

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τους καθηγητές μου κύριο Ατλαμάζογλου Βασίλειο και κυρία Θηραίου Τριάς για την πολύτιμη βοήθεια που μου πρόσφεραν κατά την διάρκεια της συγγραφής της διπλωματικής μου εργασίας. Τους ευχαριστώ θερμά για την καθοδήγηση και τον πολύτιμο χρόνο που πάντα ήταν πρόθυμοι να μου διαθέσουν.

Επίσης θα ήθελα να ευχαριστήσω τον καθηγητή μου κύριο Ακρίτα Αλκιβιάδη που με προθυμία δέχτηκε να είναι επιβλέπων της διπλωματικής μου εργασίας καθώς επίσης για τη συνεργασία μας, στο πλαίσιο των μαθημάτων, τα χρόνια αυτά.

Τέλος, οφείλω να πω ένα μεγάλο «ευχαριστώ» στην οικογένεια μου και στους κοντινούς μου ανθρώπους για την αγάπη, την εμψύχωση και την υποστήριξη που πρόσφεραν όλα αυτά τα χρόνια .

Πίνακας περιεχομένων

Πίνακας περιεχομένων σχημάτων

Πίνακας περιεχομένων πινάκων

Κεφάλαιο 1: Εισαγωγή

1.1 Εισαγωγή στις πρωτεΐνες

Οι πρωτεΐνες αποτελούν τα ποιο διαδεδομένα και πολυδιάστατα τόσο στη μορφή όσο και στη λειτουργία τους μακρομόρια. Ακόμη και σε ένα απλό κύτταρο των βακτηρίων εντοπίζονται εκατοντάδες διαφορετικές πρωτεΐνες που κάθε μια εξ αυτών έχει ιδιαίτερο ρόλο. Οι πρωτεΐνες αποτελούν είτε το δομικό συστατικό του κυττάρου είτε συνεργούν σε κάποια συγκεκριμένη λειτουργία.

Οι πρωτεΐνες είναι μεγάλα σύνθετα βιομόρια, με μοριακό βάρος από 10.000 μέχρι πάνω από 1 εκατομμύριο, αποτελούμενα από αμινοξέα, τα οποία ενώνονται μεταξύ τους με πεπτιδικούς δεσμούς σχηματίζοντας μια γραμμική αλυσίδα, καλούμενη αλυσίδα πολυπεπτιδίων. Όλες οι πρωτεΐνες περιέχουν άνθρακα, οξυγόνο και άζωτο και οι περισσότερες εξ αυτών και θείο.

Η ακολουθία αμινοξέων σε μια πρωτεΐνη καθορίζεται από ένα γονίδιο και κωδικοποιείται κατά τον γενετικό κώδικα DNA. Παρόλο που ο γενετικός κώδικας κωδικοποιεί 20 αμινοξέα, τα αμινοξέα που συνιστούν την πρωτεΐνη συχνά υφίστανται χημικές αλλαγές. Περισσότερες από μια πρωτεΐνες συχνά λειτουργούν μαζί για να επιτύχουν κάποια συγκεκριμένη λειτουργία, ή μπορεί ακόμα και να συσσωματωθούν για να διαμορφώσουν τα σταθερά σύμπλοκα.

Οι πρωτεΐνες παράγονται στο κυτόπλασμα και συγκεκριμένα στα ριβοσώματα όπου ξεκινούν ως απλές μη διακλαδωμένες αλληλουχίες αμινοξέων, δηλαδή πεπτιδίων ή πολυπεπτιδίων, σχηματίζοντας την «πρωτοταγή δομή», επί της οποίας και για την οποία καθοριστικοί παράγοντες είναι τα νουκλεϊκά οξέα, τα οποία και φέρονται να ελέγχουν όλες τις λειτουργίες αλλά και τα κληρονομικά γνωρίσματα των οργανισμών.

Στη συνέχεια όλα τα πρωτεϊνικά μόρια υφίστανται μια φυσική αναδιάταξη προκειμένου να δώσουν μια «δευτεροταγή δομή» η οποία προκαλείται από δεσμούς υδρογόνου μεταξύ των καρβοξυλομάδων και των αμινομάδων των αμινοξέων.

Κατά τη δευτεροταγή δομή δε λαμβάνονται υπ' όψιν οι αλληλεπιδράσεις μεταξύ των πλευρικών ομάδων των αμινοξέων. Ο πλέον διαδεδομένος τύπος τέτοιας μορφής είναι η λεγόμενη «α-έλικα», δεξιόστροφη, όπου οι σπείρες διατηρούνται στη θέση τους με δεσμούς υδρογόνου μεταξύ των καρβοξυλομάδων και αμινομάδων. Μια άλλη δευτεροταγής δομή είναι η λεγόμενη «β-πτυχωτή επιφάνεια» όπου στη περίπτωση αυτή διασταυρώνονται παράλληλες αλυσίδες πολυπεπτιδίων που ενώνονται στις διασταυρώσεις με δεσμούς υδρογόνου σχηματίζοντας έτσι μια εξαιρετικά σφιχτή δομή, όπως στο μετάξι. Οι πρωτεΐνες με τέτοιες σχετικά απλές δισδιάστατες δευτερογενείς δομές ονομάζονται γενικά ινώδεις πρωτεΐνες. Παρόλα αυτά οι πρωτεΐνες υφίστανται ακόμα ποιο περίπλοκο δίπλωμα (πτύχωση) το οποίο καλείται «τριτοταγής δομή». Με τον όρο τριτοταγή δομή, εννοούμε το τελικό και λειτουργικό σχήμα που αποκτά η πρωτεΐνη μετά κι από την αλληλεπίδραση των πλευρικών ομάδων των αμινοξέων (π.χ. σχηματισμός δισουλφιδικών δεσμών μεταξύ δύο κυστεϊνικών καταλοίπων). Τέλος, υπάρχουν και πρωτεΐνες που αποτελούνται από

πολλές πολυπεπτιδικές αλυσίδες που είναι χαλαρά ενωμένες και αυτό αποτελεί τη λεγόμενη «τεταρτοταγή δομή». Παράδειγμα είναι η αιμοσφαιρίνη [1]. Περισσότερες πληροφορίες για την δομή των πρωτεϊνών αναφέρονται στη συνέχεια του κεφαλαίου.

Γενικά οι πρωτεΐνες ανάλογα της μορφής τους διακρίνονται σε ινώδεις πρωτεΐνες και σε σφαιρικές πρωτεΐνες. Με κριτήριο τη σύνθεση τους διακρίνονται σε απλές (όταν αποτελούνται μόνο από αμινοξέα) και σε σύνθετες (όταν στο μόριο τους περιλαμβάνονται και μη πρωτεϊνικά τμήματα όπως μέταλλα, σάκχαρα, λίπη κ.λπ.). Επίσης με κριτήριο ακόμη τη λειτουργία τους διακρίνονται σε δομικές (όταν αποτελούν τα δομικά υλικά του κυττάρου), και λειτουργικές (όταν συμβάλλουν σε κάποιες λειτουργίες).

Οι διάφορες λειτουργίες που παρατηρούνται στους οργανισμούς γίνονται χάρη στις πρωτεΐνες. Ο δε βιολογικός τους ρόλος καθορίζεται κάθε φορά από την τρισδιάστατη δομή τους που είναι συνέπεια της αλληλουχίας των αμινοξέων, η οποία και ξεκινά από την πρωτοταγή δομή.

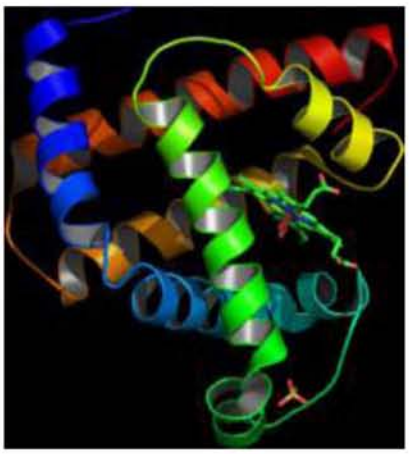
Όπως άλλα βιολογικά μακρομόρια (π.χ. οι πολυσακχαρίτες, τα λιπίδια, και νουκλεϊκά οξέα) έτσι και οι πρωτεΐνες είναι απαραίτητες για όλους τους ζωντανούς οργανισμούς και συμμετέχουν σε κάθε διαδικασία μέσα στα κύτταρα. Πολλές πρωτεΐνες δρουν ως ένζυμα που καταλύουν τις βιοχημικές αντιδράσεις, και είναι ζωτικής σημασίας στο μεταβολισμό. Άλλες πρωτεΐνες έχουν δομικές ή μηχανικές λειτουργίες, όπως οι πρωτεΐνες του κυτταρικού σκελετού, οι οποίες συμβάλλουν στη διατήρηση της μορφής των κυττάρων. Οι πρωτεΐνες είναι επίσης σημαντικές στη διακυτταρική επικοινωνία, τη δράση του ανοσοποιητικού συστήματος, τον σχηματισμό κυτταρικών ιστών, και τον κυτταρικό κύκλο [1].

Όσον αφορά τις *σφαιρικές πρωτεΐνες* πάνω στις οποίες θα γίνει η μελέτη για την πρόβλεψη της δομής τους, αξίζει να αναφέρουμε λίγες πληροφορίες σχετικά με αυτές. Οι σφαιρικές πρωτεΐνες ή αλλιώς σφαιρο – πρωτεΐνες είναι περισσότερο ή λιγότερο διαλυτές σε υδατικά διαλύματα. Αυτό είναι το βασικό χαρακτηριστικό που τις κάνει να ξεχωρίζουν από τις ινώδεις πρωτεΐνες οι οποίες είναι πρακτικά αδιάλυτες. Ο όρος «σφαιρική πρωτεΐνη» είναι σχετικά παλιός (από τον $19^{ο}$ αιώνα) και μπορεί να χαρακτηριστεί ακόμα και ως αρχαϊκός δεδομένου του τεράστιου πλήθους των πρωτεϊνών και του πιο σαφούς και περιγραφικού λεξιλογίου που χρησιμοποιείται σήμερα για τα δομικά μοτίβα των πρωτεϊνών. Αυτή η σφαιρική δομή των πρωτεϊνών αυτών προκαλείται από την τριτοταγή τους δομή. Τα υδρόφοβα αμινοξέα βυθίζονται προς το εσωτερικό του μορίου ενώ τα υδρόφιλα αμινοξέα βρίσκονται προς την εξωτερική πλευρά του μορίου επιτρέποντας αλληλεπιδράσεις με τον διαλύτη, πράγμα που εξηγεί την διαλυτότητα του μορίου.

Σε αντίθεση με τις ινώδεις πρωτεΐνες που έχουν μόνο δομική λειτουργία ,οι σφαιρικές πρωτεΐνες μπορούν να λειτουργήσουν και ως:

1. Ένζυμα, καταλύοντας οργανικές αντιδράσεις.

2. Αγγελιαφόροι, μεταδίδοντας μηνύματα ώστε να ρυθμιστούν διάφορες βιολογικές λειτουργίες.
3. Μεταφορείς άλλων μορίων μέσω μεμβρανών.
4. Παρακαταθήκες αμινοξέων.
5. Ρυθμιστές διαφόρων λειτουργιών στον οργανισμό [2].



Σχήμα 1.1 *Αναπαράσταση της τρισδιάστατης δομής της μυογλοβίνης, που παρουσιάζεται με χρωματισμένες τις άλφα έλικες. Αυτή ήταν η πρώτη πρωτεΐνη, η δομή της οποίας προσδιορίστηκε με κρυσταλλογραφία ακτίνων Χ.*

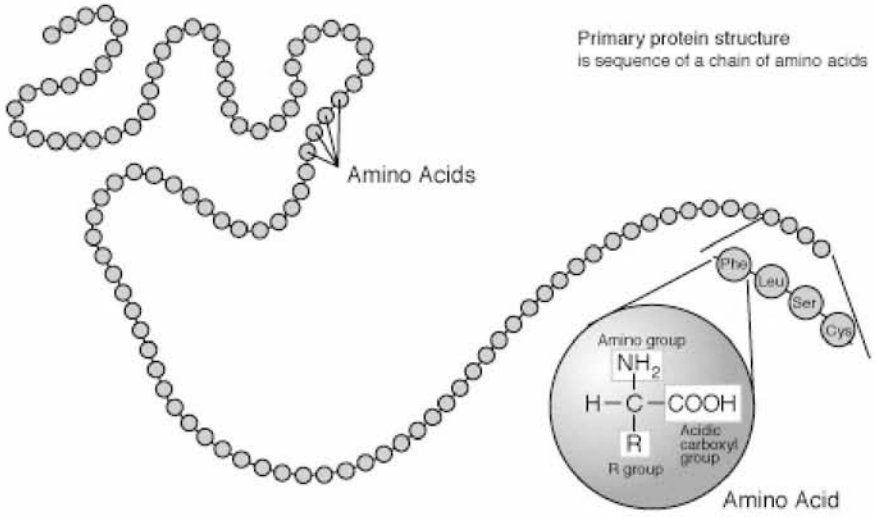
1.2 Δομή των πρωτεϊνών

Η δομή των πρωτεϊνών μπορεί να διακριθεί σε τρία επίπεδα: την πρωτοταγή δομή, την δευτεροταγή και την τριτοταγή. Στην ενότητα αυτή θα αναφέρουμε συνοπτικά κάποιες πληροφορίες σχετικά με το κάθε επίπεδο ώστε να έχουμε μια καλύτερη εικόνα για τη γενική δομή των πρωτεϊνών.

(α) Πρωτοταγής δομή

Στην βιοχημεία όταν αναφερόμαστε στην πρωτοταγή δομή ενός βιολογικού μορίου εννοούμε τον ακριβή προσδιορισμό της ατομικής του σύνθεσης και τους χημικούς δεσμούς που συνδέουν τα άτομα αυτά. Όσον αφορά τις πρωτεΐνες, η πρωτοταγής δομή είναι ισοδύναμη με τον προσδιορισμό των μονομερών υπομονάδων της δηλαδή την πεπτιδική αλυσίδα. Ο όρος «πρωτοταγής δομή» επινοήθηκε το 1951 από τον Linderstrøm-Lang. Από σύμβαση η πρωτοταγής δομή ξεκινά από το αμινο – τερματικό (N) άκρο και φτάνει μέχρι το καρβοξυλο- τερματικό (C) άκρο. Γενικά τα πολυπεπτίδια είναι αδιακλάδωτα πολυμερή και έτσι η πρωτοταγής τους δομή μπορεί να καθοριστεί από την ακολουθία των αμινοξέων κατά μήκος της αλυσίδας.

Η πρωτοταγής δομή των πρωτεϊνών καθορίζει κατά ένα μεγάλο βαθμό την τρισδιάστατη δομή τους δηλαδή την τριτοταγή δομή αλλά τα νουκλεϊκά οξέα και η πρωτεϊνική αναδίπλωση είναι θέματα τόσο σύνθετα που γνωρίζοντας την πρωτοταγή δομή δεν είναι εύκολο να προβλέψουμε το σχήμα ή στοιχεία της δευτεροταγούς δομής των πρωτεϊνών όπως πτυχωτές επιφάνειες ή έλικες. Παρόλα αυτά γνωρίζοντας τη δομή μιας παρόμοιας ομόλογης ακολουθίας (για παράδειγμα ένα μέλος της ίδιας πρωτεϊνικής οικογένειας) μπορούμε να αναγνωρίσουμε την τριτοταγή δομή της δεδομένης πρωτεΐνης [3]. Εμείς στην παρούσα εργασία θα προσπαθήσουμε να προβλέψουμε την τριτοταγή δομή των σφαιρικών πρωτεϊνών χρησιμοποιώντας μια ab initio μέθοδο.



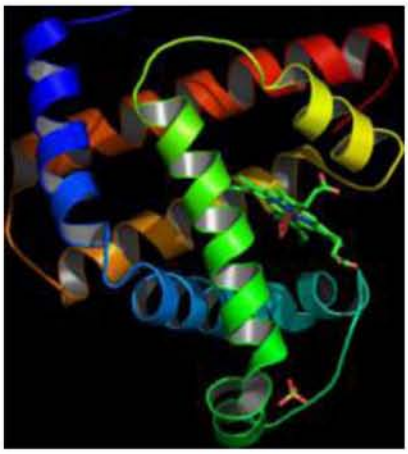
Σχήμα 1.2 *Η πρωτοταγής δομή των πρωτεϊνών είναι μια αλυσίδα αμινοξέων.*

(β) Δευτεροταγής δομή

Η δευτεροταγής δομή γενικά στη βιοχημεία και στη δομική βιολογία είναι η γενική τρισδιάστατη δομή τοπικών τμημάτων βιοπολυμερών όπως πρωτεΐνες και νουκλεϊκά οξέα. Παρόλα αυτά δεν περιγράφουν συγκεκριμένες ατομικές θέσεις στον τρισδιάστατο χώρο όπως η τριτοταγής δομή.

Η δευτεροταγής δομή στις πρωτεΐνες αποτελείται από τοπικές ενδοαμινοξικές αλληλεπιδράσεις στις οποίες ενδεχομένως να παρεμβάλλονται υδρογονικοί δεσμοί ανάμεσα στις αμινομάδες και στις καρβοξυλομάδες. Οι πιο κοινές δευτεροταγείς δομές είναι οι α-έλικες και οι β-πτυχωτές επιφάνειες. Άλλες έλικες όπως οι 3_{10} έλικες και οι π έλικες πολύ σπάνια παρατηρούνται στις φυσικές πρωτεΐνες. Επίσης, άλλες εκτεταμένες δομές όπως οι polyproline έλικες και α-πτυχωτές επιφάνειες παρατηρούνται σπάνια σε πρωτεΐνες που έχουν αναδιπλωθεί στην τελική τους κατάσταση αλλά συχνά υποθέτουμε την ύπαρξη τους στα ενδιάμεσα στάδια της πρωτεϊνικής αναδίπλωσης. Όσον αφορά τα coil αυτά δεν αποτελούν μια πραγματική δευτεροταγή δομή αλλά ουσιαστικά είναι η τάξη των διαφόρων σχηματισμών που υποδεικνύουν την απουσία κανονικής δευτεροταγούς δομής.

Τα αμινοξέα ποικίλλουν ως προς την ικανότητα τους να σχηματίζουν διάφορα στοιχεία δευτεροταγούς δομής. Η προλίνη και η γλυκίνη για παράδειγμα, είναι γνωστές και ως «helix breakers» και αυτό γιατί διαταράσσουν την κανονικότητα του σχηματισμού α-έλικας. Τα αμινοξέα που προτιμούν την υιοθέτηση ελικοειδών σχηματισμών στις πρωτεΐνες είναι η μεθιονίνη, η αλανίνη, η λευκίνη, το γλουταμικό οξύ και η λυσίνη. Αντιθέτως, τα μεγάλα αρωματικά κατάλοιπα (τρυπτοφάνη, τυροσίνη και φαινυλαλανίνη) και αμινοξέα όπως ισολευκίνη, βαλίνη και θρεονίνη προτιμούν την υιοθέτηση σχηματισμών β-πτυχωτών επιφανειών. Παρόλα αυτά οι προτιμήσεις αυτές δεν είναι αρκετά ισχυρές ώστε να μπορούν να παράγουν μια αξιόπιστη μέθοδο πρόβλεψης της δευτεροταγούς δομής έχοντας σαν μόνη πληροφορία την ακολουθία [4].



Σχήμα 1.3 *Η πρωτεΐνη μυογλοβίνη. Οι α-έλικες αναπαριστώνται με χρώμα και το coil σε λευκό ενώ δεν υπάρχουν β-πτυχωτές επιφάνειες.*

(γ) Τριτοταγής δομή

Η τριτοταγής δομή μιας πρωτεΐνης ή κάποιου άλλου μακρομορίου είναι η τρισδιάστατη δομή του όπως καθορίζεται από τις ατομικές συντεταγμένες.

Η τριτοταγής δομή θεωρείται πως καθορίζεται κατά ένα πολύ μεγάλο βαθμό από την πρωτοταγή δομή της πρωτεΐνης δηλαδή την ακολουθία των αμινοξέων από την οποία αποτελείται. Η προσπάθεια να προβλεφθεί η τριτοταγής δομή από την πρωτοταγή είναι γενικά γνωστή ως «πρόβλεψη της δομής της πρωτεΐνης από πρώτες αρχές». Παρόλα αυτά το περιβάλλον μέσα στο οποίο συντίθεται και αναδιπλώνεται μια πρωτεΐνη αποτελεί ένα βασικό παράγοντα που καθορίζει την τελική μορφή της και συνήθως δε λαμβάνεται άμεσα υπόψη από τις τρέχουσες μεθόδους πρόβλεψης.

Στις σφαιρικές πρωτεΐνες οι τριτοταγείς αλληλεπιδράσεις συνήθως σταθεροποιούνται από την απομόνωση των υδρόφοβων καταλοίπων στον πρωτεϊνικό πυρήνα από τον οποίο το νερό αποκλείεται, καθώς και από τον επακόλουθο εμπλουτισμό των υδρόφιλων καταλοίπων στην επιφάνεια της πρωτεΐνης η οποία είναι εκτεθειμένη στο νερό. Οι πρωτεΐνες κατηγοριοποιούνται όσον αφορά τα διπλώματα τους σε βάσεις δεδομένων όπως η SCOP και η CATH.

Ο πιο κοινός σχηματισμός μιας πρωτεΐνης στο κυτταρικό της περιβάλλον γενικά αναφέρεται ως «τελική κατάσταση» ή native state. Θεωρείται πως αυτός ο σχηματισμός είναι ο πιο σταθερός σχηματισμός από θερμοδυναμική άποψη και είναι ουσιαστικά η λειτουργική μορφή της πρωτεΐνης.

Η πλειοψηφία των γνωστών πρωτεϊνικών δομών μέχρι σήμερα έχει επιλυθεί πειραματικά με την τεχνική κρυσταλλογραφίας ακτινών Χ που τυπικά παρέχει δεδομένα υψηλής ανάλυσης αλλά δεν παρέχει πληροφορίες σχετικά με την προσαρμοστικότητα των πρωτεϊνών. Ένας δεύτερος κοινός τρόπος επίλυσης πρωτεϊνικών δομών χρησιμοποιεί το NMR το οποίο γενικά παρέχει δεδομένα χαμηλότερης ανάλυσης και περιορίζεται σε σχετικά μικρές πρωτεΐνες αλλά μπορεί να παρέχει πληροφορίες σχετικά με την κίνηση των πρωτεϊνών. Αξίζει να σημειώσουμε πως διαθέτουμε περισσότερες πληροφορίες σχετικά με την τριτοταγή δομή των σφαιρικών πρωτεϊνών απ ότι των μεμβρανο-πρωτεϊνών γιατί οι τελευταίες είναι πολύ δύσκολο να μελετηθούν χρησιμοποιώντας τις παραπάνω μεθόδους.

Η πρώτη δομή σφαιρικής πρωτεΐνης που προβλέφθηκε ήταν το cyclol μοντέλο της Dorothy Wrinch αλλά γρήγορα βρέθηκε πως ήταν ασυνεπές με τα πειραματικά δεδομένα [5]. Οι σύγχρονες μέθοδοι είναι ορισμένες φορές ικανές να προβλέψουν την τριτοταγή δομή de novo δηλαδή από την ακολουθία και μόνο και αυτό ακριβώς μελετείται και στην παρούσα εργασία.

1.3 Ο μηχανισμός αναδίπλωσης της πρωτεΐνης

Η πρωτεϊνική αναδίπλωση είναι η φυσική διαδικασία κατά την οποία ένα πολυπεπτίδιο αναδιπλώνεται στην χαρακτηριστική και λειτουργική τρισδιάστατη δομή του [6].

Κάθε πρωτεΐνη ξεκινά ως πολυπεπτίδιο. Το πολυπεπτίδιο αυτό δεν έχει ανεπτυγμένη τρισδιάστατη δομή. Παρόλα αυτά κάθε αμινοξύ της αλυσίδας θεωρείται ότι έχει συγκεκριμένα χημικά χαρακτηριστικά. Μπορεί για παράδειγμα να είναι υδρόφοβα, υδρόφιλα ή ηλεκτρικά φορτισμένα. Αυτά αλληλεπιδρούν μεταξύ τους καθώς και με τον περιβάλλοντα χώρο τους ώστε να σχηματιστεί ένα καλώς ορισμένο τρισδιάστατο σχήμα, η αναδιπλωμένη πρωτεΐνη, γνωστή και ως «native state» δηλαδή τελική στερεοδιάταξη. Η τρισδιάστατη αυτή δομή καθορίζεται από την ακολουθία των αμινοξέων όπως αναφέρει και το δόγμα του «Anfinsen»[7]. Ο μηχανισμός αναδίπλωσης της πρωτεΐνης ακόμα και σήμερα δεν είναι πλήρως κατανοητός.

Η πρωτοταγής δομή μιας πρωτεΐνης ουσιαστικά επηρεάζει την τελική στερεοδιάταξη. Η πρωτεΐνη θα αναδιπλωθεί αυθόρμητα κατά την διάρκεια ή μετά την σύνθεση. Ο μηχανισμός αναδίπλωσης εξαρτάται ισοδύναμα από χαρακτηριστικά όπως η φύση του διαλύματος, η συγκέντρωση των αλάτων, η θερμοκρασία και άλλα [8].

Οι περισσότερες αναδιπλωμένες πρωτεΐνες διαθέτουν υδρόφοβο πυρήνα όπου το «πακετάρισμα» των πλευρικών αλυσίδων σταθεροποιεί την όλη αναδιπλωμένη δομή καθώς και φορτισμένες ή πολικές πλευρικές αλυσίδες στην επιφάνεια όπου είναι εκτεθειμένη στο διάλυμα, οι οποίες αλληλεπιδρούν με τα περιβάλλοντα μόρια νερού. Είναι γενικά αποδεκτό ότι η βασική κινητήριος δύναμη πίσω από τον μηχανισμό της αναδίπλωσης είναι η ελαχιστοποίηση του πλήθους των υδρόφοβων πλευρικών αλυσίδων που είναι εκτεθειμένες στο νερό [9]. Παρόλα αυτά προτάθηκε και μια πιο πρόσφατη θεωρία η οποία επανεξετάζει την συνεισφορά των υδρογονικών δεσμών [10]. Η ισχύς των υδρογονικών δεσμών σε μια πρωτεΐνη ποικίλλει καθώς εξαρτάται από το μικροπεριβάλλον τους. Για παράδειγμα οι υδρογονικοί δεσμοί στον υδρόφοβο πυρήνα έχουν μεγαλύτερη ισχύ απ ότι σε υδατικό περιβάλλον [11].

Η διαδικασία της «in vivo» αναδίπλωσης συχνά γίνεται κατά τέτοιο τρόπο ώστε το N-τερματικό της πρωτεΐνης ξεκινά να αναδιπλώνεται ενώ το C-τερματικό ακόμα συντίθεται από το ριβόσωμα. Ειδικές πρωτεΐνες που ονομάζονται chaperones βοηθούν στη διαδικασία της αναδίπλωσης [12].

Το πιο βασικό γεγονός της αναδίπλωσης πάντως παραμένει το ότι η αμινοξική ακολουθία κάθε πρωτεΐνης περιέχει την πληροφορία που καθορίζει την τελική στερεοδιάταξη. Αυτό βέβαια δε σημαίνει ότι πανομοιότυπες αμινοξικές ακολουθίες πάντα αναδιπλώνονται με όμοιο τρόπο [13]. Οι τελικές στερεοδιατάξεις διαφοροποιούνται και από περιβαλλοντολογικούς παράγοντες. Πανομοιότυπες πρωτεΐνες αναδιπλώνονται διαφορετικά ανάλογα με το που βρίσκονται.

Η αναδίπλωση είναι μια αυθόρμητη διαδικασία ανεξάρτητη από ενεργειακές εισόδους. Οι δυνάμεις που οδηγούν στην αναδίπλωση είναι οι μη ομοιοπολικές αλληλεπιδράσεις μεταξύ ατόμων της πολυπεπτιδικής αλυσίδας. Ιδιαίτερη σημασία έχουν οι αλληλεπιδράσεις μεταξύ ατόμων που ανήκουν σε αμινοξέα που βρίσκονται μακριά το ένα από το άλλο κατά μήκος της αλυσίδας. Χάρη σ' αυτές, επιτυγχάνεται το δίπλωμα σε συμπαγή και σταθερή στερεοδομή.

Με τον όρο «μη ομοιοπολικές» εννοούμε τις δυνάμεις Van der Waals, τους δεσμούς υδρογόνου, τις ηλεκτροστατικές δυνάμεις, καθώς και την επίδραση του υδατικού περιβάλλοντος του κυττάρου. Η επίδραση του νερού έχει ως αποτέλεσμα τα περισσότερα υδρόφοβα αμινοξέα να συγκεντρώνονται στο εσωτερικό της πρωτεΐνης (τον πυρήνα της) και τα υδρόφιλα να ευρίσκονται κατά κανόνα στο εξωτερικό της κέλυφος.

Οι παραπάνω δυνάμεις είναι κατά πολύ ασθενέστερες των ομοιοπολικών, όμως αφορούν έναν μεγάλο αριθμό ζευγών ατόμων, γεγονός που οδηγεί στην ενεργειακή σταθερότητα της διπλωμένης πρωτεΐνης. Παρακάτω αναφέρουμε ενδεικτικά λίγες πληροφορίες σχετικά με τις μη ομοιοπολικές δυνάμεις.

(α)Δεσμοί υδρογόνου

Στο δεσμό υδρογόνου δύο αρνητικά ιόντα κρατιούνται ενωμένα μέσω της έλξης και των δύο με το πρωτόνιο του ατόμου του Η. Η ενέργεια ενός υδρογονικού δεσμού (τυπικά 5 ως 30 kJ/mole) είναι συγκρίσιμη με αυτή των αδύναμων ομοιοπολικών δεσμών και ένας τυπικός ομοιοπολικός δεσμός είναι μόλις 20 φορές ισχυρότερος από έναν ενδομοριακό υδρογονικό δεσμό. Αυτοί οι δεσμοί μπορούν να αναπτυχθούν ανάμεσα σε μόρια ή μέσα σε διάφορα μέρη ενός μορίου [14]. Ο δεσμός αυτός είναι ασθενέστερος των ομοιοπολικών, ιονικών και μεταλλικών δεσμών και μπορεί να αναπτυχθεί τόσο σε ανόργανα μόρια όσο και σε οργανικά .

Ο δεσμός υδρογόνου είναι εξαιρετικά ασθενής αλλά με πολύ μεγάλη βιολογική σημασία, επειδή είναι ο δεσμός που συνδέει τμήματα πολύ βασικών βιολογικών μορίων (π.χ. στο DNA τις στροφές της έλικας). Επίσης οι δεσμοί υδρογόνου είναι εν μέρει υπεύθυνοι για την δευτεροταγή, τριτοταγή και τεταρτοταγή δομή των πρωτεϊνών [15].

(β) Δυνάμεις Van der Waals

Οι δυνάμεις Van der Waals ονομάστηκαν από τον Ολλανδό επιστήμονα Johannes Diderik van der Waals και είναι οι ελκτικές ή απωθητικές δυνάμεις μεταξύ των μορίων, διαφορετικές των ομοιοπολικών δεσμών ή της ηλεκτροστατικής αλληλεπίδρασης των ιόντων [16]. Είναι κυρίως δεσμοί μεταξύ μορίων (ή ατόμων που δεν συνδέονται με ιοντικούς ή ομοιοπολικούς δεσμούς) για το σχηματισμό της στερεάς, υγρής ή αέριας ύλης. Οφείλονται στις αδύναμες ηλεκτροστατικές έλξεις ανάμεσα στα ουδέτερα αυτά μόρια. Οι δυνάμεις Van der Waals είναι τριών ειδών:

- *Δύναμη διπόλου-διπόλου*: Αλληλεπίδραση μορίων με μόνιμη διπολική ροπή (λόγω της ανομοιόμορφης κατανομής φορτίου στο κάθε μόριο). Το μόριο αλληλεπιδρά με το ηλεκτρικό πεδίο που δημιουργείται από το άλλο μόριο.

- *Δύναμη διπόλου-διπόλου εξ επαγωγής*: Το ηλεκτρικό πεδίο του ενός μορίου-διπόλου πολώνει το δεύτερο μόριο με αλληλεπίδραση διπόλου-διπόλου εξ επαγωγής.

- *Δύναμη διασποράς*: Δύναμη μεταξύ μη πολικών μορίων. Λόγω της ανομοιόμορφης κατανομής φορτίου στο κάθε μόριο τα φορτία του ενός, όταν τα μόρια έλθουν κοντά, "βλέπουν" τα φορτία του άλλου.

Οι δεσμοί Van der Waals είναι αυτοί που καθορίζουν αν μια ουσία θα είναι υγρή ή αέρια σε θερμοκρασία περιβάλλοντος. Μια ουσία που αποτελείται από μη πολικά μόρια περιμένουμε να είναι αέρια (π.χ. H_2) ενώ μια ουσία που αποτελείται από πολικά μόρια θα είναι υγρή [15].

(γ) Ηλεκτροστατικές δυνάμεις

Το πρωτόνιο είναι το στοιχειώδες θετικό φορτίο του ηλεκτρισμού και βρίσκεται μέσα στον πυρήνα. Στη φυσική του κατάσταση, το άτομο οποιουδήποτε στοιχείου περιέχει ίσο αριθμό ηλεκτρονίων και πρωτονίων. Το αρνητικό φορτίο κάθε ηλεκτρονίου είναι ίσο σε μέγεθος με το θετικό φορτίο κάθε πρωτονίου άρα τα δύο αυτά αντίθετα φορτία αντισταθμίζονται και το άτομο είναι ηλεκτρικά ουδέτερο.

Ένα από τα μυστήρια του ατόμου είναι ότι το ηλεκτρόνιο και ο πυρήνας έλκονται μεταξύ τους. Αυτή η έλξη ονομάζεται ηλεκτροστατική δύναμη και είναι αυτή που κρατά σε τροχιά τα ηλεκτρόνια.

Οι δυνάμεις αυτές λοιπόν είναι συνδυασμός απωστικών και ελκτικών δυνάμεων λόγω άπωσης των ομώνυμων φορτίων και έλξης των ετερώνυμων φορτίων των ατόμων [15,17].

Σε συγκεκριμένα διαλύματα και κάτω από συγκεκριμένες συνθήκες οι πρωτεΐνες δεν δύναται να αναδιπλωθούν στην βιοχημική – λειτουργική δομή τους. Θερμοκρασίες υψηλότερες ή μερικές φορές και χαμηλότερες από αυτές που ζουν τα κύτταρα θα προκαλέσουν το ξεδίπλωμα ή την αποδιάταξη των θερμικά ασταθών πρωτεϊνών. Το ίδιο μπορεί να προκληθεί και από υψηλές συγκεντρώσεις διαλυμάτων, υψηλές τιμές pH, παρουσία μηχανικών δυνάμεων και χημικών αλλοιώσεων. Μια ολικά αποδιατεταγμένη πρωτεΐνη στερείται δευτεροταγούς και τριτοταγούς δομής και υφίσταται ως το λεγόμενο «random coil». Υπό συγκεκριμένες συνθήκες κάποιες πρωτεΐνες μπορούν να διπλωθούν ξανά αλλά σε πολλές περιπτώσεις η αποδιάταξη είναι μη αναστρέψιμη [18].

1.4 Μια παλιότερη προσέγγιση για την πρόβλεψη του τρόπου αναδίπλωσης

Μία συνήθης παραδοχή που υπάρχει σε πολλές μελέτες της αναδίπλωσης είναι η ανάλυση της πρωτεΐνης σε επιμέρους δομικά στοιχεία. Η επιτυχής προσέγγιση της στερεοδομής περνά από την επιτυχή εντοπισμό αυτών των στοιχείων και την συσχέτισή τους με την αμινοξική ακολουθία. Κλασσικό παράδειγμα είναι το λεγόμενο «ιεραρχικό δίπλωμα» που ήταν ιδιαίτερα δημοφιλής προσέγγιση κατά τις δεκαετίες '70 και '80.

Είναι η κλασικότερη προσέγγιση στο πρόβλημα της αναδίπλωσης. Σύμφωνα με αυτή, ο μηχανισμός της αναδίπλωσης ακολουθεί το σχήμα πρωτοταγής - δευτεροταγής - τριτοταγής δομή (Σχήμα 1.4). Με άλλα λόγια, πρώτα σχηματίζονται τα στοιχεία της δευτεροταγούς δομής και μετά "πακετάρονται" στην τριτοταγή δομή. Με αυτή την παραδοχή, η πρόβλεψη της στερεοδομής ανάγεται στην επιτυχή πρόβλεψη των στοιχείων της δευτεροταγούς δομής από την αμινοξική ακολουθία. Στη συνέχεια, το πακετάρισμα στην τριτοταγή δομή είναι εφικτό να προβλεφθεί. Πράγματι, ο αριθμός των στοιχείων δευτεροταγούς δομής είναι μικρός και η εύρεση της ενεργειακά ευνοϊκότερης διάταξής τους είναι υπολογιστικά εφικτή.

Σχήμα 1.4 *Γραφική απεικόνιση του μηχανισμού του «ιεραρχικού διπλώματος».*

Ωστόσο αυτή η προσέγγιση δεν έχει γενική ισχύ: έχει αποδειχθεί πειραματικά ότι ο σχηματισμός της δευτεροταγούς δομής δεν πραγματοποιείται στα πρώτα στάδια του διπλώματος, παρά σε ορισμένες μόνο περιπτώσεις. Πολλές φορές μάλιστα, η δευτεροταγής δομή σχηματίζεται στα τελικά στάδια του διπλώματος. Συγκεκριμένα οι β-πτυχωτές επιφάνειες σταθεροποιούνται από αποστάσεις μακράς εμβέλειας. Άρα η σταθερότητα των στοιχείων της εξαρτάται συχνά σε μεγάλο βαθμό από την μεταξύ

τους αλληλεπίδραση και δεν μπορούν να εξεταστούν ως ανεξάρτητες οντότητες. Γι' αυτό τον λόγο και οι μέθοδοι προσδιορισμού της δευτεροταγούς δομής από την ακολουθία έχουν περιορισμένη επιτυχία στον ακριβή προσδιορισμό της, ιδιαίτερα όσον αφορά τα όρια των στοιχείων της.

Δύο βασικά συμπεράσματα προκύπτουν από την πρόσφατη έρευνα της αναδίπλωσης:

1. Δεν υπάρχει ενιαίος μηχανισμός αναδίπλωσης των πρωτεϊνών. Ακόμα και για την ίδια πρωτεΐνη, μπορούν να υπάρξουν περισσότερα του ενός μονοπάτια σχηματισμού της μοναδικής στερεοδομής της.

2. Η δευτεροταγής δομή δεν μπορεί να αποτελέσει ενιαίο τρόπο περιγραφής των πρωτεϊνικών δομών.

Έτσι αναπόφευκτα, οι πρόσφατες ερευνητικές εργασίες στο χώρο της αναδίπλωσης κινούνται πέρα από την δευτεροταγή δομή.

1.5 Σύγχρονες προσεγγίσεις για την πρόβλεψη του τρόπου αναδίπλωσης

Όπως αναφέραμε και προηγουμένως, η πρόβλεψη της δομής των πρωτεϊνών είναι ένας από τους σημαντικότερους σκοπούς του τομέα της βιοπληροφορικής και της θεωρητικής χημείας. Σκοπός είναι η πρόβλεψη της τρισδιάστατης δομής των πρωτεϊνών από την αμινοξική ακολουθία. Η πρόβλεψη της τρισδιάστατης δομής των πρωτεϊνών είναι υψηλής σημασίας για την φαρμακευτική και την βιοτεχνολογία..

Η πρακτική σημασία της πρόβλεψης της πρωτεϊνικής δομής είναι τώρα πιο σημαντική από ποτέ. Παράγονται συνεχώς μεγάλες ποσότητες δεδομένων πρωτεϊνικών ακολουθιών και ο πειραματικός προσδιορισμός των δομών τους (χρησιμοποιώντας μεθόδους όπως κρυσταλλογραφία ακτινών Χ) είναι τόσο ακριβός όσο και χρονοβόρος. Αυτό συμβαίνει κυρίως για δύο λόγους. Πρώτον, το πλήθος των πιθανών δομών στις οποίες μπορεί να αναδιπλωθεί μια πρωτεΐνη είναι εξαιρετικά μεγάλο και δεύτερον η φυσική βάση πάνω στην οποία σταθεροποιείται η πρωτεϊνική δομή δεν είναι πλήρως κατανοητή. Σαν αποτέλεσμα, κάθε μέθοδος πρόβλεψης της δομής της πρωτεΐνης χρειάζεται έναν τρόπο να εξερευνεί τον χώρο των πιθανών δομών με αποτελεσματικό τρόπο δηλαδή χρειάζεται μια στρατηγική αναζήτησης καθώς και έναν τρόπο ώστε να αναγνωρίζει τις πιο ρεαλιστικές δομές μέσω μιας συνάρτησης ενέργειας.

Στις μεθόδους συγκριτικής πρόβλεψης δομών, ο χώρος αναζήτησης περιορίζεται από την υπόθεση ότι η μελετώμενη πρωτεΐνη υιοθετεί μια δομή που είναι λογικά παρόμοια με τη δομή μιας τουλάχιστον γνωστής πρωτεΐνης. Στις ab initio μεθόδους, στη δικιά μας περίπτωση δηλαδή, καμία τέτοια υπόθεση δε γίνεται, πράγμα που καθιστά το μέγεθος του προβλήματος της αναζήτησης ακόμα μεγαλύτερο και δυσκολότερο. Και στις δύο περιπτώσεις απαιτείται η χρήση μιας συνάρτησης ενέργειας ώστε να αναγνωριστεί η τελική αναδιπλωμένη στερεοδομή αλλά και για την καθοδήγηση της αναζήτησης προς αυτή την στερεοδομή. Δυστυχώς η κατασκευή μιας τέτοιας συνάρτησης ενέργειας είναι ένα ανοιχτό πρόβλημα κατά ένα μεγάλο βαθμό.

Η άμεση προσομοίωση της πρωτεϊνικής αναδίπλωσης σε λεπτομέρεια ατόμων είναι συνήθως δύσκολη λόγω του υψηλού υπολογιστικού κόστους. Έτσι, οι περισσότερες ab initio μέθοδοι πρόβλεψης βασίζονται σε απλοποιημένα μοντέλα της ατομικής δομής των πρωτεϊνών.

Η συγκριτική πρόβλεψη δομών χρησιμοποιεί προηγούμενες πειραματικά λυμένες δομές ως αρχικά σημεία ή πρότυπα. Αυτό είναι αποτελεσματικό καθώς παρότι το πλήθος των πρωτεϊνών είναι μεγάλο, τα μοτίβα τριτοταγής δομής τους είναι πεπερασμένα. Η μέθοδος αυτή μπορεί να διαχωριστεί σε 2 ομάδες [19]:

- Homology Modeling το οποίο βασίζεται στην υπόθεση ότι δύο ομόλογες πρωτεΐνες μοιράζονται πανομοιότυπες δομές

- Protein threading στο οποίο οι αμινοξικές ακολουθίες πρωτεϊνών άγνωστης δομής σαρώνονται έναντι μιας βάσης λυμένων δομών. Σε κάθε περίπτωση μια συνάρτηση σκοραρίσματος αξιολογεί τη συμβατότητα της ακολουθίας με την δομή και έτσι παράγονται τα διάφορα τρισδιάστατα μοντέλα [20].

Στην παρούσα εργασία ακολουθήθηκε μια ab initio μέθοδος πρόβλεψης της πρωτεϊνικής δομής. Πληροφορίες σχετικά με το «πνεύμα» των τεχνικών αυτών ακολουθούν στο παρόν κεφάλαιο.

1.6 Ab initio τεχνική πρόβλεψης της δομής – Η δική μας προσέγγιση στο πρόβλημα.

Οι ab initio ή «από πρώτες αρχές» τεχνικές για την πρόβλεψη της δομής των πρωτεϊνών σχετίζονται αλλά διαφέρουν από τις διάφορες μελέτες για την αναδίπλωση των πρωτεϊνών. Οι τεχνικές αυτές προσπαθούν με μόνη πληροφορία την αμινοξική ακολουθία να προσομοιώσουν τη διαδικασία αναδίπλωσης και να καταλήξουν στην τρισδιάστατη δομή που είναι περισσότερο ενεργειακά πιθανότερο να σχηματιστεί. Έτσι, δε βασίζονται σε προηγούμενες πειραματικά λυμένες δομές αλλά προσπαθούν από το μηδέν να παράγουν την τελική στερεοδιάταξη ακολουθώντας διάφορες φυσικές αρχές που διέπουν την πρωτεϊνική αναδίπλωση. Ένα γενικό παράδειγμα ab initio τεχνικής είναι η δειγματοληψία του χώρου των πιθανών στερεοδιατάξεων, η οποία καθοδηγείται από συναρτήσεις ενέργειας και άλλους παράγοντες που έχουν να κάνουν με την αμινοξική ακολουθία ώστε να παραχθεί ένα μεγάλο σύνολο πιθανών στερεοδιατάξεων. Στη συνέχεια, χρησιμοποιώντας συναρτήσεις «σκοραρίσματος» επιλέγονται οι κατάλληλες στερεοδιατάξεις που ομοιάζουν με την τελική στερεοδιάταξη. Αξίζει να τονίσουμε πως οι τεχνικές αυτές απαιτούν πολύπλοκους αλγορίθμους και πολλούς υπολογιστικούς πόρους και γι αυτό συνήθως εφαρμόζονται σε σχετικά μικρές πρωτεΐνες [21].

Επειδή η προσομοίωση της αναδίπλωσης είναι εξαιρετικά χρονοβόρα (πχ μία τυπική πρωτεΐνη 100 αμινοξέων αποτελείται από πολλές χιλιάδες άτομα και ο αριθμός των δυνατών στερεοδιατάξεων της είναι της τάξης του 3^{100}), χρησιμοποιούνται διάφορες απλουστεύσεις. Γι αυτό, για την αναπαράσταση της πρωτεΐνης χρησιμοποιούνται απλοποιημένα μοντέλα μειωμένης απεικόνισης (με ψευδό – άτομα να αντιπροσωπεύουν ομάδες ατόμων), οι δε διάφορες αλληλεπιδράσεις ανάμεσα στα άτομα αυτά απλοποιούνται και αυτές χρησιμοποιώντας κάποιο στατιστικό πεδίο δυνάμεων. Η χρήση απλοποιημένων μοντέλων έχει ένα κόστος και αυτό είναι πως προβλέπονται τα επιμέρους στοιχείων της πρωτεϊνικής δομής π.χ. εύρεση των αμινοξέων που απαρτίζουν τον υδρόφοβο πυρήνα μίας πρωτεΐνης. Για αυτό το λόγο και η παρούσα εργασία εφαρμόζεται σε σφαιρικές πρωτεΐνες, καθώς παρουσιάζουν εξαιρετικό ενδιαφέρον.

Συγκεκριμένα, στην παρούσα εργασία, το απλοποιημένο μοντέλο αναπαράστασης της πρωτεΐνης που χρησιμοποιήθηκε είναι το μοντέλο 2 – 1 – 0 το οποίο βασίζεται στο απλό κυβικό μοντέλο, είναι δηλαδή μοντέλο διακριτού χώρου. Υπάρχουν φυσικά πολλά μοντέλα αναπαράστασης που θα μπορούσαν να χρησιμοποιηθούν και έχουν με επιτυχία χρησιμοποιηθεί σε άλλες εργασίες όπως το απλό κυβικό μοντέλο που αναφέραμε προηγουμένως ή το μοντέλο diamond. Παρόλα αυτά το συγκεκριμένο μοντέλο έχει το πλεονέκτημα να είναι τόσο απλό όσο και ρεαλιστικό άρα και ιδιαιτέρως αποτελεσματικό.

Στη συνέχεια έπρεπε να καθοριστεί ένας δυναμικός αλγόριθμος που θα προσομοιώνει την αναδίπλωση μέσω μίας σειράς μεταβάσεων σε στερεοδιατάξεις χαμηλότερης ενέργειας, άρα και πιο συμπαγείς. Ο δυναμικός αυτός αλγόριθμος που

χρησιμοποιήθηκε είναι ο δυναμικός αλγόριθμος Monte Carlo ο οποίος μάλιστα χρησιμοποιείται συχνά στα μοντέλα διακριτού χώρου.

Στα απλοποιημένα μοντέλα διακριτού χώρου δεν είναι δυνατή η εφαρμογή των πεδίων της μοριακής μηχανικής. Αντίθετα, χρειαζόμαστε δυνάμεις μεταξύ αμινοξέων. Με άλλα λόγια, χρειαζόμαστε ένα μέσο πεδίο δυνάμεων (η ενεργειών) για κάθε ζεύγος αμινοξέων. Ένα τέτοιο πεδίο θα έχει την μορφή ενός πίνακα 20 X 20 (20 είδη αμινοξέων). Συγκεκριμένα χρησιμοποιήθηκε ο 20×20 πίνακας των Miyazawa και Jernigan ο οποίος προήλθε από στατιστική μελέτη.

Η προσέγγιση της πραγματικής στερεοδομής μίας πρωτεΐνης δεν είναι εφικτή με τόσο απλά μοντέλα σαν το Monte-Carlo σε διακριτό χώρο όπως προείπαμε. Όμως, αυτές οι προσεγγίσεις μας δίνουν πληροφορίες για τα αμινοξέα που καθορίζουν τον σχηματισμό του υδρόφοβου πυρήνα και παίζουν κρίσιμο ρόλο στον σχηματισμό της τελικής δομής, μπορούμε δηλαδή να αντλήσουμε χρήσιμες πληροφορίες για επιμέρους τμήματα της δομής. Προκειμένου να γίνει αυτό χρησιμοποιήθηκε ένας αλγόριθμος που υπολογίζει τα Ισχυρά Αλληλεπιδρώντα Αμινοξέα (Mostly Interacting Residues). Ο αλγόριθμος αυτός υπολογίζει κρίσιμες θέσεις στην αμινοξική ακολουθία οι οποίες σχετίζονται με τον υδρόφοβο πυρήνα και συγκεκριμένα με τα άκρα των κλειστών βρόχων (TEF) και τις τοποϋδρόφοβες θέσεις περισσότερες πληροφορίες για τα οποία θα αναφέρουμε στη συνέχεια του κεφαλαίου. Να σημειώσουμε πως στο κεφάλαιο αυτό έγινε μια πολύ σύντομη αναφορά σχετικά με την ab initio μέθοδο που χρησιμοποιήθηκε στην εργασία. Λεπτομερής ανάλυση σχετικά με το μοντέλο 2-1-0, το στατιστικό πεδίο δυνάμεων, τον δυναμικό αλγόριθμο Monte Carlo αλλά και τον αλγόριθμο υπολογισμού των MIR που χρησιμοποιήθηκαν στην εργασία υπάρχει στο κεφάλαιο «Μέθοδοι».

Αξίζει να τονίσουμε πως εφαλτήριο για τις ab-initio τεχνικές πρόβλεψης της τριτοταγούς δομής των πρωτεϊνών έχει αποτελέσει το δόγμα του Anfinsen ή αλλιώς «η θερμοδυναμική υπόθεση» [22]. Το δόγμα αυτό αναφέρει πως, τουλάχιστο για τις μικρές σφαιρικές πρωτεΐνες, η τελική κατάσταση ή αλλιώς native state καθορίζεται μονοσήμαντα από την ακολουθία των αμινοξέων της πρωτεΐνης. Στις περιβαλλοντολογικές συνθήκες στις οποίες λαμβάνει χώρα η αναδίπλωση (θερμοκρασία, συγκέντρωση και σύνθεση διαλύματος κλπ), η τελική αναδιπλωμένη μορφή είναι ουσιαστικά ένα μοναδικό, σταθερό και κινητικά προσπελάσιμο ελάχιστο της ελεύθερης ενέργειας. Ισχύουν τρεις συνθήκες:

- Μοναδικότητα: Η συνθήκη αυτή απαιτεί ότι η ακολουθία δεν έχει άλλη διάταξη με συγκρίσιμη ελεύθερη ενέργεια. Επομένως το ελάχιστο της ελεύθερης ενέργειας πρέπει να είναι αναντίρρητο.

- Σταθερότητα: Μικρές αλλαγές στο εξωτερικό περιβάλλον δε μπορούν να δώσουν αφορμή για αλλαγές στην στερεοδιάταξη ελάχιστης ενέργειας.

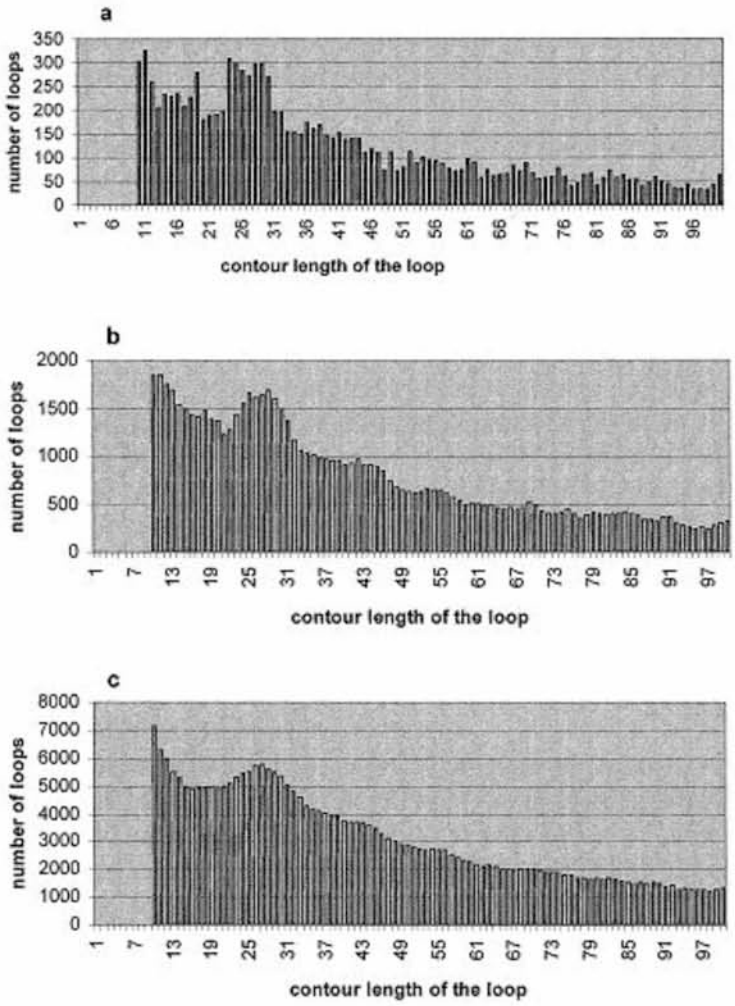
- Κινητική προσιτότητα: Η αναδίπλωση της πρωτεϊνικής αλυσίδας δεν πρέπει να περιλαμβάνει ιδιαίτερα σύνθετες αλλαγές στο σχήμα.

1.7 Τα MIR ως μέσο πρόβλεψης του πρωτεϊνικού πυρήνα.

Η σύγχρονη περιγραφή των δομικών προτύπων στις πρωτεΐνες έχει γίνει πλέον αρκετά εξεζητημένη. Πλέον περιλαμβάνει στοιχεία δευτεροταγούς δομής, βρόχους, τομείς και την ιεραρχία τους. Κάθε δομικό στοιχείο που υπάρχει, βρίσκεται σε συνέπεια με την φυσική, την γεωμετρία και τις ενέργειες των πολυπεπτιδικών αλυσίδων. Παρόλα αυτά, όσο το επιτρέπει βέβαια και η φυσική, η εξέλιξη μπορεί να έχει κάνει κάποιου είδους επιλογή προωθώντας συγκεκριμένα δομικά στοιχεία και καταπιέζοντας άλλα. Γι αυτό προέκυψε το ερώτημα του πόσο πολύ η εξελικτική διαδικασία επηρέασε τη φυσική επιλογή της τελικής στερεοδιάταξης στις πρωτεΐνες. Για το λόγο αυτό έγινε ανάλογη έρευνα. Αυτή η έρευνα αποκάλυψε την ύπαρξη τουλάχιστον ενός βασικού στοιχείου των σφαιρικών πρωτεϊνών: κλειστοί βρόχοι τυπικού μήκους.

Ένα από τα βασικά χαρακτηριστικά τους είναι ο loop factor $P_\delta(l)$. Με ένα συγκεκριμένο όριο μήκους δ, αυτή είναι η πιθανότητα να βρεθούν δύο μονομερή μέσα σε απόσταση δ το ένα από το άλλο εφόσον υπάρχουν l μονομερή ξεχωριστά κατά μήκος της αλυσίδας. Αναλογικά το $P_\delta(l)$ συμπεριφέρεται ως εξής: για πολύ μικρό l είναι και αυτό πολύ μικρό, για μεγάλο l ο loop factor είναι πάλι μικρός αλλά σταθμίζεται για πολύ μεγάλες τιμές του l. Μάλιστα υπάρχει και συγκεκριμένη τιμή για το l με την οποία ο loop factor φτάνει στη μέγιστη τιμή. Πειραματικά αποδείχτηκε πως η κατανομή των βρόχων με κοντινές Ca-Ca επαφές (τυπικά με λιγότερο από δ = 10 Å) έφτασε στη μέγιστη τιμή 25-30 καταλοίπων. Αυτοί οι βρόχοι μπορεί μεν να περιέχουν διάφορα στοιχεία δευτεροταγούς δομής, αλλά έχουν τυπικά το ίδιο περίπου μέγεθος .

Το σχήμα 1.5 απεικονίζει την κατανομή των μηκών των βρόχων που βρέθηκαν με αποστάσεις 5, 7 και 10 Å ανάμεσα στα άκρα τους [23]. Τα ιστογράμματα δείχνουν εξέχοντα μέγιστα στο εύρος 22-32 των μηκών των βρόχων (σε πλήθος αμινοξέων). Με αυτό τον τρόπο φαίνεται πως οι βρόχοι αυτοί προτιμούν συγκεκριμένα μεγέθη. Αξίζει να σημειωθεί πως οι βρόχοι δεν είναι απαραίτητα ανεξάρτητοι και πως ορισμένοι μεγάλοι βρόχοι ενδεχομένως περιέχουν και άλλους μικρότερους.



Σχήμα 1.5 *Ιστογράμματα του πλήθους των βρόχων σε σχέση με το μήκος τους όταν η απόσταση ανάμεσα στα άκρα τους είναι μικρότερη από (a) 5* Å (b) 7 Å (c) 10 Å.

Επίσης βρέθηκε που ακριβώς κατά μήκος της αλυσίδας υπάρχουν αυτοί οι βρόχοι. Πειραματικά δείχθηκε πως το 40% με 85% των αλυσίδων «καλύπτονται» από τους βρόχους αυτούς και ουσιαστικά αποδείχθηκε πόσο σημαντικό δομικό στοιχείο αποτελούν οι βρόχοι αυτοί για τις πρωτεΐνες και ιδιαίτερα τις σφαιρικές και τις μεμβρανο-πρωτεΐνες. Ένα μεγάλο ποσοστό των πρωτεϊνών αυτών αποτελείται από τους βρόχους αυτούς ανεξάρτητα από τον τύπο των διπλωμάτων που προτιμούν οι πρωτεΐνες. Επίσης, οι βρόχοι αυτοί είναι δομικά ετερογενείς δηλαδή μπορεί να αποτελούνται από οποιονδήποτε συνδυασμό ελίκων, β-πτυχωτών επιφανειών και άλλων, διατηρώντας ομοιόμορφο το μέγεθος τους. Αυτή η ομοιομορφία στο μέγεθος τους είναι που ουσιαστικά προσδίδει στην πρωτεΐνη μια τοπολογική ομαλότητα η οποία είναι ανεξάρτητη από την διευθέτηση των στοιχείων δευτεροταγούς δομής. Υπό αυτή την έννοια, τα άκρα των βρόχων οργανώνουν τη συνολική δομή των πρωτεϊνών και είναι πολύ μεγάλης σημασίας [23].

Μια βασική ιδιότητα της πρωτεϊνικής δομής είναι το υδρόφοβο εσωτερικό της και το υδρόφιλο εξωτερικό της. Η πρωτεϊνική αναδίπλωση βρέθηκε ότι περιλαμβάνει πρώτα την δημιουργία υδρόφοβου πυρήνα. Μια από τις πολλές δυνάμεις που σταθεροποιούν τον πυρήνα, είναι οι υδρόφοβες αλληλεπιδράσεις, οι οποίες παίζουν πολύ σημαντικό ρόλο και γι αυτό η κατανομή των υδρόφοβων αμινοξέων τόσο στην ακολουθία των πρωτεϊνών όσο και στην τρισδιάστατη δομή τους είναι μεγάλης σημασίας. Από μελέτες που έγιναν, βρέθηκε πως τα άκρα των κλειστών βρόχων βρίσκονται στον υδρόφοβο πυρήνα των πρωτεϊνών και πως η απόσταση ανάμεσα στα υδρόφοβα αμινοξέα ακολουθεί τον ίδιο κανόνα με τα μεγέθη των βρόχων, απέχουν δηλαδή και αυτά 25-30 αμινοξικές θέσεις. Από αυτό συμπεραίνουμε πως οι υδρόφοβες αλληλεπιδράσεις παίζουν δραστικό ρόλο στο σχηματισμό των βρόχων. Από τη μελέτη αυτή λοιπόν αναδείχθηκε η σημαντικότητα τόσο του υδρόφοβου πυρήνα όσου και των κλειστών βρόχων ως βασικό δομικό στοιχείο των πρωτεϊνών. Επίσης προτάθηκε και ένα νέο σενάριο για την πρωτεϊνική αναδίπλωση που δεν υπήρχε προηγουμένως και το οποίο περιλαμβάνει το σχηματισμό των βρόχων αυτών, με τα άκρα τους στον υδρόφοβο πυρήνα, από τα πρώτα κιόλας στάδια της αναδίπλωσης [24].

Η προσπάθεια ανάλυσης της δομής των σφαιρικών πρωτεϊνών με έναν γενικό τρόπο λοιπόν, οδήγησε στην έννοια των κλειστών βρόχων που αναφέραμε προηγουμένως και που στη βιβλιογραφία αναφέρονται ως closed loops, ή Tightened-End-Fragments (TEF).

Ανακεφαλαιώνοντας, η ανάλυση μεγάλου αριθμού δομών σφαιρικών πρωτεϊνών οδήγησε στο συμπέρασμα πως μία διπλωμένη πολυπεπτιδική αλυσίδα μπορεί να αναλυθεί σε μία αλληλουχία τμημάτων (TEF) με τις εξής ιδιότητες [23,24,25]:

1. Τα τμήματα ξεκινούν και τελειώνουν στον υδρόφοβο πυρήνα της πρωτεΐνης.

2. Έχουν τυπικό μήκος 25-35 αμινοξικά κατάλοιπα.

3. Τα άκρα των τμημάτων βρίσκονται κοντά στον χώρο (τυπικά απέχουν λιγότερο από 10 Å).

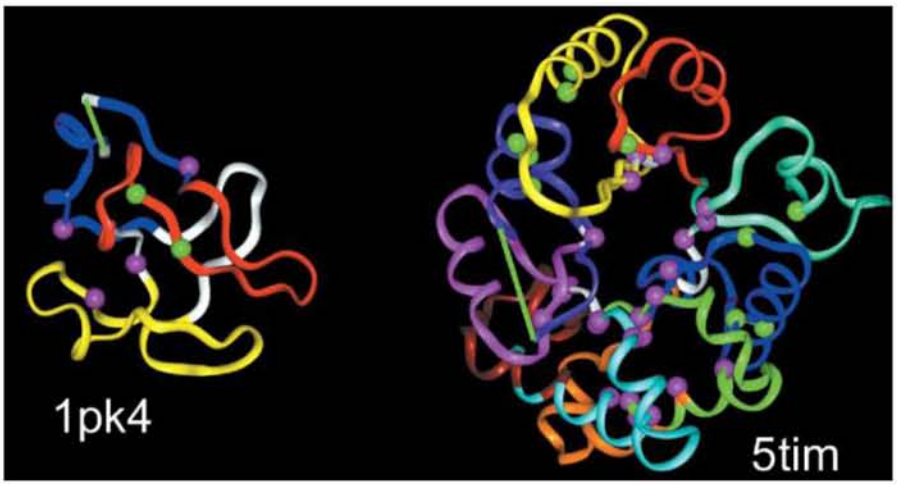
4. Αυτή η ανάλυση αποτελεί τον κανόνα για τις δομές των σφαιρικών πρωτεϊνών.

Επιπλέον, έχει δειχθεί ότι σε μία στερεοδομή, η ενέργεια αλληλεπίδρασης μεταξύ διαδοχικών βρόχων είναι μικρότερη από τις εσωτερικές τους ενέργειες. Αυτό υποδεικνύει ότι οι κλειστοί βρόχοι μπορούν να θεωρηθούν προσεγγιστικά ως ενεργειακά ανεξάρτητες οντότητες.

Σχηματικά, οι δομές των σφαιρικών πρωτεϊνών υπακούουν στην περιγραφή του σχήματος 1.6.

Σχήμα 1.6 *Η πρωτεϊνική δομή ως closed loops.*

Παραδείγματα ανάλυσης πρωτεϊνικών δομών σε κλειστούς βρόχους (σχήμα 1.7)



Σχήμα 1.7 *Η ανάλυση των δομών των πρωτεϊνών 1pk4 και 5tim σε κλειστούς βρόχους όπου κάθε βρόχος αναπαρίσταται με διαφορετικό χρώμα.*

Αξιοσημείωτο είναι πως με την παραδοχή του αρχικού σχηματισμού των TEF λύνεται και το λεγόμενο «παράδοξο του Levinthal» [26]. Το παράδοξο του Levinthal είναι ένα συλλογιστικό πείραμα στον τομέα της πρωτεϊνικής αναδίπλωσης. Το 1969 ο Cyrus Levinthal επισήμανε πως εξαιτίας του πολύ μεγάλου βαθμού ελευθερίας που έχουν οι μη αναδιπλωμένες πολυπεπτιδικές αλυσίδες, το μόριο μπορεί να

αναδιπλωθεί σε αστρονομικά μεγάλο αριθμό πιθανών στερεοδιατάξεων. Ο χρόνος αναδίπλωσης μπορεί να υπολογιστεί από τον τύπο $t = n^L * \tau$. Αν στον τύπο βάλουμε όπου L = 150 αμινοξέα που είναι το τυπικό μήκος μιας πρωτεΐνης, n = 3 στερεοδιατάξεις / αμινοξύ και $\tau = 10^{-12}$ sec όπου είναι ο χρόνος μετάβασης ανάμεσα στις στερεοδιατάξεις τότε ο χρόνος διπλώματος με διερεύνηση όλων των στερεοδιατάξεων είναι $t = 10^{48}$ χρόνια! Φυσικά το παράδοξο είναι ότι πολλές μικρές πρωτεΐνες αναδιπλώνονται αυθόρμητα σε millisecond ή ακόμα και microsecond. Άρα, ουσιαστικά οι πρωτεΐνες δεν εξερευνούν όλο τον χώρο των πιθανών στερεοδιατάξεων και αυτό που ήθελε να επισημάνει ο Levinthal ήταν ότι μια καθαρά τυχαία αναζήτηση στον χώρο των πιθανών στερεοδιατάξεων προκειμένου να προβλεφθεί ο τρόπος αναδίπλωσης των πρωτεϊνών δεν είναι δυνατόν να πετύχει. Αν στον παραπάνω τύπο θέσουμε λοιπόν όπου $t = 10^{-1}$ ως 10^3 sec δηλαδή τον παρατηρούμενο χρόνο διπλώματος, παίρνουμε L = 23-31 αμινοξέα δηλαδή το τυπικό μήκος ενός TEF [27].

Πέρα από τα TEF ανακαλύφθηκε μέσα από μελέτες και κάποιο άλλο πολύ σημαντικό δομικό στοιχείο των πρωτεϊνών που έχει να κάνει με την αμινοξική ακολουθία. Συγκεκριμένα οι έρευνες διενεργήθηκαν στον πυρήνα σφαιρικών πρωτεϊνών και με τη σύγκριση ζευγών από ακολουθίες ομόλογων περιοχών γνωστής τρισδιάστατης δομής, αναδείχθηκαν δυο σημαντικοί πληθυσμοί υδρόφοβων αμινοξέων: αυτά που μοιράζονται τις ίδιες θέσεις στις δυο δομές και αυτά που αντικαθίστανται από ένα μη υδρόφοβο αμινοξύ στην άλλη ακολουθία. Όταν μιλάμε για υδρόφοβα αμινοξέα περιλαμβάνουμε τα V – I – M – W – Y – L – F .Η μέση προσιτότητα διαλύματος για τον πρώτο πληθυσμό είναι μικρή ενώ για τον άλλο είναι σημαντικά μεγαλύτερη.

Στη συνέχεια οι μελέτες προχώρησαν από ζεύγη ακολουθιών σε ολόκληρες πρωτεϊνικές οικογένειες στις οποίες έγινε στοίχιση ακολουθιών και μελετήθηκαν οι ιδιαίτερες ιδιότητες των υδρόφοβων αμινοξέων σε τοπολογικά συντηρημένες περιοχές της δομής [28,29]. Αυτές οι θέσεις οι οποίες ονομάστηκαν τοποϋδρόφοβες και περιλαμβάνουν μόνο υδρόφοβα αμινοξέα, παίζουν πολύ σημαντικό ρόλο στην δομή.

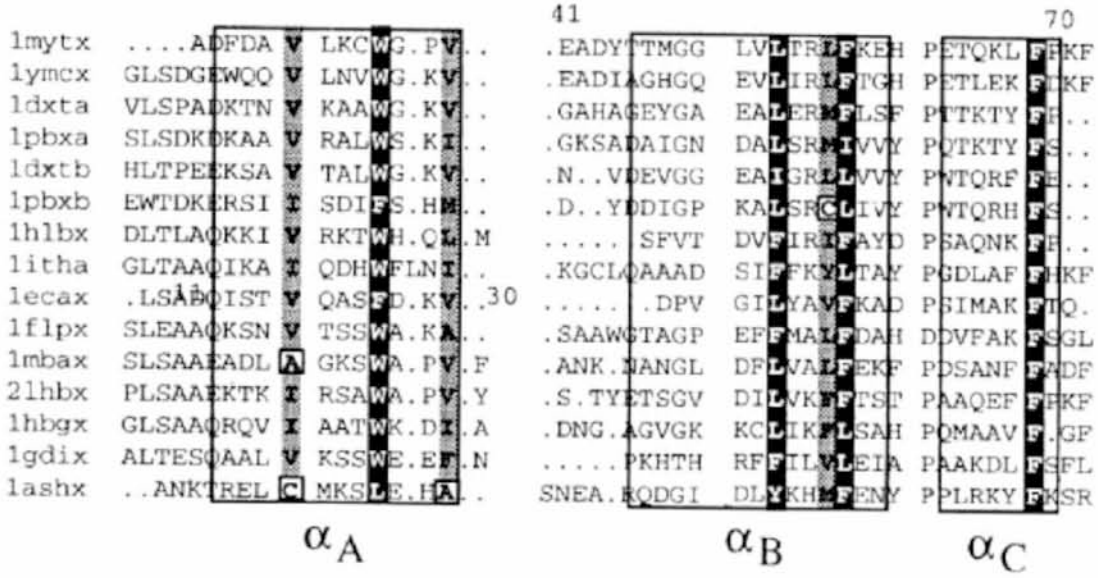
Η εύρεσή τους ακολουθεί την παρακάτω διαδικασία [29]:

1. Υδρόφοβα αμινοξέα εδώ θεωρούνται τα V – I – M – W – Y – L – F.

2. Σε κάθε οικογένεια συλλέγονται οι πρωτεΐνες γνωστής στερεοδομής και βρίσκεται η βέλτιστη υπέρθεση των δομών τους στο χώρο. Με τον τρόπο αυτό προκύπτει μία στοίχιση των ακολουθιών τους. Με βάση τη στοίχιση, από το σύνολο των πρωτεϊνών κρατάμε εκείνες που η ταυτότητα των ακολουθιών τους ανά ζεύγος δεν υπερβαίνει το 30%. Με τον τρόπο αυτό αυξάνει η πιθανότητα ώστε οι ομοιότητες μεταξύ ακολουθιών να αφορούν αμινοξικές θέσεις σημαντικές για την δομή.

3. Πρέπει να βρεθούν τουλάχιστον τέσσερις ακολουθίες που να ικανοποιούν τα παραπάνω και να συνεχιστεί η διαδικασία.

4. Αναζητούνται οι θέσεις της στοίχισης που καταλαμβάνονται τουλάχιστον κατά 75% από υδρόφοβα αμινοξέα, όπως αυτά ορίστηκαν παραπάνω.



Σχήμα 1.8 *Τοποϋδρόφοβες αμινοξικές θέσεις στην οικογένεια της αιμοσφαιρίνης. Στην εικόνα φαίνεται ένα τμήμα της στοίχισης των ακολουθιών της οικογένειας. Αριστερά η στήλη με τους κωδικούς της PDB. Με γκρι οι θέσεις με τουλάχιστον 75% υδρόφοβα και με μαύρο οι 100% υδρόφοβες.*

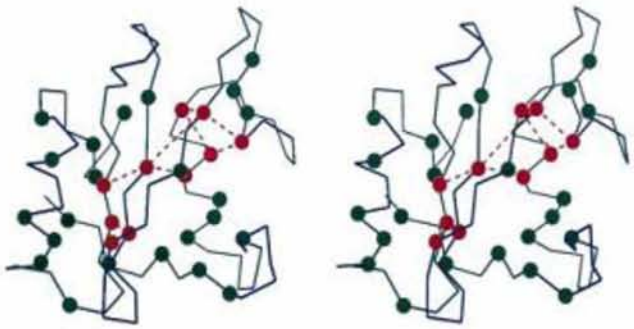
Αφού για τις διάφορες οικογένειες πρωτεϊνών βρέθηκαν οι τοποϋδρόφοβες θέσεις, υπολογίστηκε η μέση προσιτότητα διαλύματος των υδρόφοβων αμινοξέων τόσο για τις τοποϋδρόφοβες όσο και για τις μη τοποϋδρόφοβες θέσεις. Η διαφορά αυτής της μέσης προσιτότητας διαλύματος για το ίδιο αμινοξύ σε τοποϋδρόφοβες και σε μη τοποϋδρόφοβες θέσεις αποδείχτηκε σημαντική.

Για κάθε οικογένεια αυτή η διαφορά ήταν σημαντική στις περισσότερες περιπτώσεις όπου η οικογένεια περιείχε τουλάχιστο 4 πρωτεΐνες. Όταν η οικογένεια περιείχε μόνο 2 ή 3 πρωτεΐνες η διαφορά δεν ήταν και τόσο σημαντική. Παρατηρήθηκαν μόνο ελάχιστες περιπτώσεις όπου η μέση προσιτότητα διαλύματος για ένα αμινοξύ στις τοποϋδρόφοβες θέσεις είναι υψηλότερη απ ότι αυτή του ίδιου αμινοξέως σε μη τοποϋδρόφοβες θέσεις αλλά η διαφορά αυτή δεν ήταν σημαντική.

Από τα παραπάνω προκύπτει το συμπέρασμα πως τα αμινοξέα στις τοποϋδρόφοβες θέσεις είναι πολύ περισσότερο «θαμμένα» απ ότι τα αμινοξέα στις μη τοποϋδρόφοβες θέσεις, δηλαδή οι θέσεις αυτές βρίσκονται συγκεντρωμένες στον πυρήνα των πρωτεϊνών και μάλιστα αποτελούν το βασικό συστατικό στοιχείο του αναδιπλωμένου πυρήνα. Αυτό φάνηκε και από τον εξής υπολογισμό που έγινε στις μελέτες αυτές. Για κάθε πρωτεΐνη μιας οικογενείας υπολογίστηκε η μέση απόσταση ανάμεσα στα βαρυτικά κέντρα όλων των αμινοξέων στις τοποϋδρόφοβες και τις μη τοποϋδρόφοβες θέσεις. Η μέση απόσταση είναι πολύ μικρότερη για την πρώτη περίπτωση και έτσι

προκύπτει το γεγονός πως οι τοποϋδρόφοβες θέσεις βρίσκονται στον πυρήνα της δομής [28,30].

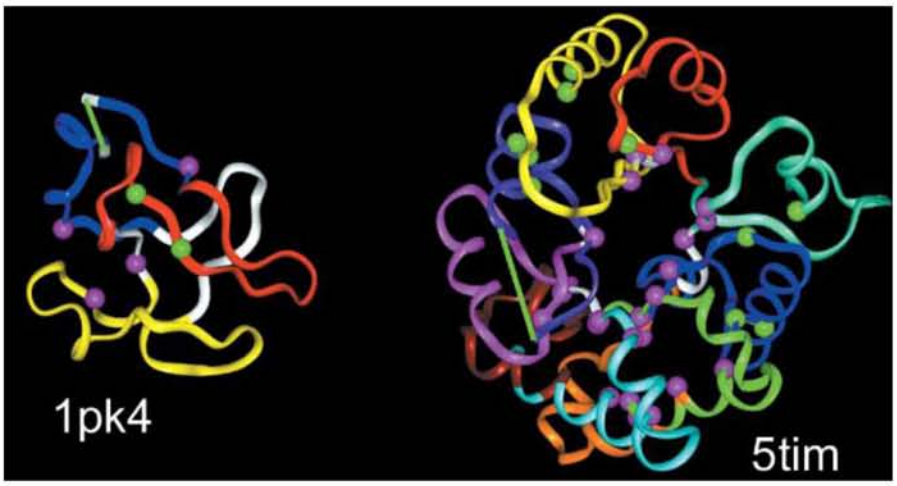
Επίσης, τα αμινοξέα στις τοποϋδρόφοβες θέσεις ενδέχεται να συμμετέχουν σε υδρόφοβες αλληλεπιδράσεις στην πρωτεϊνική δομή και αυτό γιατί βρίσκονται σε σχετικά μικρή απόσταση. Μια άλλη ιδιότητα που έχουν οι τοποϋδρόφοβες θέσεις είναι ότι οι πλευρικές αλυσίδες των αμινοξέων είναι πολύ λιγότερο διασκορπισμένες απ ότι των αμινοξέων στις μη τοποϋδρόφοβες θέσεις. Επίσης, συνθέτουν ένα πλέγμα θέσεων που αλληλεπιδρούν μεταξύ τους όπως φαίνεται στο σχήμα 1.9 [29,30].



Σχήμα 1.9 *Πλέγμα τοποϋδρόφοβων θέσεων. Οι άνθρακες-α των αμινοξέων που βρίσκονται σε τοποϋδρόφοβες θέσεις εμφανίζονται με κόκκινο χρώμα και οι άνθρακες-α των αμινοξέων που βρίσκονται σε μη τοποϋδρόφοβες θέσεις εμφανίζονται με πράσινο χρώμα. Το πλέγμα αναπαρίσταται με κόκκινες διακεκομμένες γραμμές.*

Συνεπώς, η αναγνώριση των τοποϋδρόφοβων θέσεων από την ακολουθία είναι ένας πολύ σημαντικός δομικός περιορισμός ώστε να μειωθεί αισθητά το πλήθος των προτιμητέων στερεοδιατάξεων καθώς το πλέγμα των τοποϋδρόφοβων θέσεων θέτει γεωμετρικούς περιορισμούς [30]. Επίσης, η πλειονότητα των υδρόφοβων αμινοξέων που βρίσκονται στον πυρήνα των σφαιρικών πρωτεϊνών και που παίζουν ένα πολύ σημαντικό ρόλο από τα πρώτα κιόλας στάδια της αναδίπλωσης βρίσκονται σε αυτές τις τοποϋδρόφοβες θέσεις καθώς από μετρήσεις που έγιναν σε προηγούμενες μελέτες αποδείχθηκε ότι τα υδρόφοβα αμινοξέα εμφανίζουν μια ιδιαίτερη προτίμηση στις τοποϋδρόφοβες θέσεις [28,29,30]. Μάλιστα, οι τοποϋδρόφοβες θέσεις εκπροσωπούν περίπου το 30% με 50% του συνολικού πλήθους των υδρόφοβων αμινοξέων μιας περιοχής και συνιστούν μια μικρότερη εσωτερική οντότητα μέσα στον πρωτεϊνικό πυρήνα. Έτσι, η πρόβλεψη της φύσης και των προσεγγιστικών άκρων των στοιχείων δευτεροταγούς δομής σε συνδυασμό με την ανίχνευση των τοποϋδρόφοβων θέσεων που μπορεί να γίνει από την ακολουθία μόνο, δύναται να παρουσιάσει ένα νέο ορίζοντα στις ab-initio μεθόδους πρόβλεψης της τριτοταγούς δομής των σφαιρικών πρωτεϊνών. Στην παρούσα εργασία δεν ασχοληθήκαμε με την δευτεροταγή δομή αλλά χρησιμοποιήσαμε την πληροφορία για τις τοποϋδρόφοβες θέσεις σε συνδυασμό με τους κλειστούς βρόχους (TEF).

Από αυτά που είδαμε και παραπάνω, οι τοποϋδρόφοβες θέσεις (ιδιότητα της ακολουθίας) και οι κλειστοί βρόχοι (ιδιότητα της δομής), έννοιες στενά συνδεδεμένες, αποτελούν κοινά δομικά στοιχεία των σφαιρικών πρωτεϊνών οι οποίες χαρακτηρίζονται κυρίως από την ύπαρξη υδρόφιλης επιφάνειας (τα 2/3 των αμινοξέων μιας περιοχής) και ενός εσωτερικού υδρόφιλου πυρήνα (1/3). Προηγούμενες μελέτες έδειξαν πως η στατιστική κατανομή των μηκών των TEF και αυτή των τοποϋδρόφοβων θέσεων ταιριάζει και πως συνήθως βρίσκονται σε φυσική ταύτιση: τα άκρα των TEF βρίσκονται σε τοποϋδρόφοβες θέσεις ή πολύ κοντά σε αυτές δηλαδή τα TEF ξεκινούν ή καταλήγουν σε θέση που καταλαμβάνεται από ένα ή περισσότερα τοποϋδρόφοβα αμινοξέα (σχήμα 1.10). Επίσης, τα τοποϋδρόφοβα αμινοξέα επιδεικνύουν μια περιοδικότητα στην κατανομή τους στην ακολουθία, παρόμοια με το μήκος των 22-32 καταλοίπων που συνιστούν τα TEF [31]. Το γεγονός αυτό σε συνδυασμό με το ότι τα άκρα των TEF αποτελούνται από υδρόφοβα αμινοξέα που βρίσκονται στον πυρήνα της πρωτεϊνικής δομής μας οδήγησε στο συμπέρασμα πως οι τοποϋδρόφοβες θέσεις δρουν ως άγκυρες για την πρωτεϊνική αναδίπλωση.



Σχήμα 1.10 *Σχηματική απεικόνιση που δείχνει πως τα τοποϋδρόφοβα αμινοξέα (σφαιρίδια στο σχήμα) βρίσκονται εν γένει κοντά στα άκρα των βρόχων.*

Από μια άποψη ο έγκαιρος σχηματισμός ενός πυρήνα που αποτελείται από τοποϋδρόφοβες θέσεις βοηθά στο σχηματισμό των κλειστών βρόχων και επιταχύνει σημαντικά τη διαδικασία της αναδίπλωσης. Έτσι, το ζεύγος εννοιών «τοποϋδρόφοβες θέσεις» και «TEF» παρέχουν ένα γενικό και απλό σενάριο για το μηχανισμό αναδίπλωσης των σφαιρικών πρωτεϊνών καθώς και μια ομάδα κρίσιμων θέσεων στον πρωτεϊνικό πυρήνα. Η ανάλυση της δομής των σφαιρικών πρωτεϊνών σε βρόχους αποτελεί ένα γενικό μοτίβο ανεξάρτητο της δευτεροταγούς δομής και του συγκεκριμένου τρόπου αναδίπλωσης της κάθε πρωτεΐνης [31].

Η εργασία μας βασίζεται σε μια παλιότερη έρευνα η οποία προσπάθησε να προβλέψει αυτές τις κρίσιμες θέσεις από την ακολουθία, πράγμα το οποίο είναι μεγάλης σημασίας για την προσέγγιση της δομής πρωτεϊνών άγνωστου τρόπου αναδίπλωσης. Προκειμένου να «χτιστεί» μια τέτοια δομή, συλλέχτηκαν διάφορα κομμάτια πληροφορίας από τον συνδυασμό διαφόρων μεθόδων.

Η έννοια των τοποϋδρόφοβων θέσεων προτείνει πως οι δυνάμεις οι οποίες «βυθίζουν» αυτά τα κατάλοιπα και οδηγούν σε ένα σταθερό πυρήνα δεν βασίζονται στις λεπτομέρειες της δομής της πλευρικής αλυσίδας των αμινοξέων αλλά σε μια ικανή ακολουθία υδρόφοβων και πολικών αμινοξικών καταλοίπων κατά μήκος της πολυπεπτιδικής αλυσίδας. Έτσι, απλοποιημένα πρωτεϊνικά μοντέλα όπως τα μοντέλα πλέγματος είναι ικανά εργαλεία για τους υπολογισμούς που στοχεύουν στον εντοπισμό των κρίσιμων καταλοίπων [31].

Ξεκινώντας λοιπόν από τυχαίες στερεοδομές, οι προσομοιώσεις Monte Carlo αποκάλυψαν πως μια ομάδα υδρόφοβων καταλοίπων είχε μια ισχυρή τάση να θαφτεί. Αυτά τα κατάλοιπα, ονομάστηκαν «mostly interacting residues» (MIR) ή «ισχυρά αλληλεπιδρώντα αμινοξέα» και στατιστικά βρέθηκε ότι συμπίπτουν με τις τοποϋδρόφοβες θέσεις και τα άκρα των TEF.

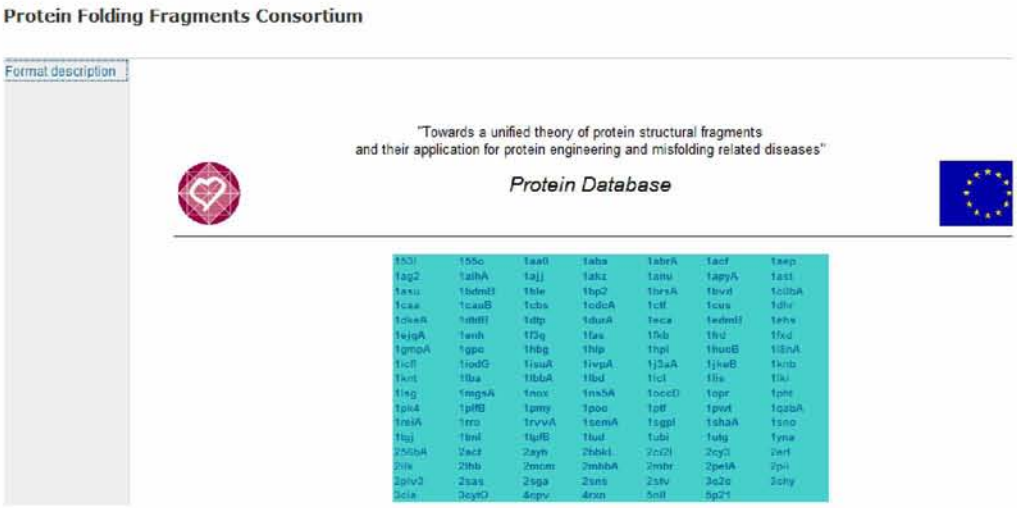
Η παρούσα εργασία προχωρά από τις κρίσιμες θέσεις στα κρίσιμα τμήματα της αμινοξικής ακολουθίας χρησιμοποιώντας την πληροφορία από τα MIR και με την εφαρμογή συγκεκριμένου αλγορίθμου που βασίζεται στη θεωρία των μεταλλάξεων (βλ. ενότητα 2.5), προσπαθεί να προβλέψει τα TEF μέσω των MCF (mutation correlation fragments). Τα αποτελέσματα της εργασίας αυτής και το πόσο αποτελεσματικά τα MCF καταφέρνουν να προβλέψουν τα TEF αναφέρονται στο κεφάλαιο «Αποτελέσματα». Η αξιολόγηση των αποτελεσμάτων έγινε χρησιμοποιώντας διάφορα στατιστικά μέτρα. Τέλος, έγινε και μια οπτικοποίηση των προβλέψεων ώστε να είναι ορατή στον αναγνώστη αυτή η προσπάθεια πρόβλεψης των TEF περιοχών μέσω των MCF.

Κεφάλαιο 2: Μέθοδοι

2.1 Ακολουθίες πρωτεϊνών

Η εργασία έγινε σε ένα dataset 107 σφαιρικών πρωτεϊνών γνωστής δομής οι οποίες αντιπροσωπεύουν 78 διαφορετικά διπλώματα σύμφωνα με την δομική κατηγοριοποίηση των πρωτεϊνών (SCOP classification). Η SCOP είναι μια βάση δεδομένων που βρίσκεται στη διεύθυνση http://scop.mrc-lmb.cam.ac.uk/scop/ .

Οι πρωτεΐνες αυτές βρίσκονται στη βάση δεδομένων PFF από το site http://babylone.ulb.ac.be/LIFE/ (σχήμα 2.1). Η βάση αυτή περιέχει τους κωδικούς PDB για τις μελετώμενες πρωτεΐνες καθώς και πληροφορίες σχετικά με την κατηγορία των πρωτεϊνών και την ταξινόμηση του οργανισμού από τον οποίο προέρχονται (σχήμα 2.2).



Σχήμα 2.1 *Η βάση δεδομένων PFF. Στο πράσινο πλαίσιο βρίσκονται τα links για τις μελετώμενες πρωτεΐνες.*

Συγκεκριμένα, από το σύνολο των 107 πρωτεϊνών οι 24 ανήκουν στην κατηγορία «All Alpha proteins», οι 21 στην κατηγορία «All Beta proteins», οι 26 «Alpha and Beta proteins (a+b)», οι 17 «Alpha and Beta proteins (a/b)», οι 17 «Small proteins», 1 και συγκεκριμένα η 1aa0 στην κατηγορία «Coil» και η 1occD στην κατηγορία «membrane and cell surface proteins and peptides». Από τα νούμερα αυτά γίνεται φανερό ότι στην εργασία μας χρησιμοποιήθηκε ένα ισορροπημένο δείγμα το οποίο αντιπροσωπεύει τα κυριότερα γνωστά διπλώματα και πως δεν υπάρχει κάποια κατηγορία που να υπερέχει πολύ έναντι κάποια άλλης. Αυτό ήταν πολύ σημαντικό ώστε να μην υπάρχει πόλωση των αποτελεσμάτων υπέρ κάποιας συγκεκριμένης κατηγορίας και ενδεχομένως αλλοίωση στην αξιολόγηση των αποτελεσμάτων.

Όσον αφορά την ταξινόμηση τους με βάση τον οργανισμό από τον οποίο προέρχονται, οι 64 πρωτεΐνες προέρχονται από ευκαριωτικούς οργανισμούς, οι 9 από ιούς, οι 32 από βακτήρια, 1 από archaea (αρχαίους οργανισμούς) και η 1abrA η οποία δε κατατάσσεται πουθενά.

Τέλος, να αναφέρουμε ότι τα μήκη των πολυπεπτιδικών αλυσίδων ποικίλουν και κυμαίνονται από 50 ως 250 περίπου κατάλοιπα.

```
==== PROTEIN FRAGMENT CONSORTIUM FILE BASED UPON DSSP FORMAT =====
HEADER     HYDROLASE(O-GLYCOSYL)                   05-MAY-94   153L
COMPND     LYSOZYME (E.C.3.2.1.17)
SOURCE     GOOSE (ANSER ANSER ANSER)
AUTHOR     L.H.WEAVER,M.G.GRUETTER,B.W.MATTHEWS
DESC       Lysozyme from Goose (Anser anser anser)
FOLD       Lysozyme-like
SCOP       a+b
CATH       DOMAIN 153100 CATHCODE 1.10.530.10 CLASS Mainly Alpha ARCH Orthogonal Bundle
CHAIN      -
SWNAME     LYG_ANSAN
SWPAN      P00718
GENACC     ------
SEQPDB     RIDCYGNVNRIDTTGASCKTAKPEGLSYCGVSASKKIAERDLQAMDRYKTIIKKVGEKLC
SEQPDB     VEPAVIAGIISRESHAGKVLKNGWGDRGNGFGLMQVDKRSHKPQGTWNGEVHITQGTTIL
SEQPDB     INFIKTIQKKFPSWTKDQQLKGGISAYNAGAGNVRSYARMDIGTTHDDYANDVVARAQYY
SEQPDB     KQHGY
SEQSW      RTDCYGNVNRIDTTGASCKTAKPEGLSYCGVSASKKIAERDLQAMDRYKTIIKKVGEKLC
SEQSW      VEPAVIAGIISRESHAGKVLKNGWGDRGNGFGLMQVDKRSHKPQGTWNGEVHITQGTTIL
SEQSW      INFIKTIQKKFPSWTKDQQLKGGISAYNAGAGNVRSYARMDIGTTHDDYANDVVARAQYY
SEQSW      KQHGY
NUMS       185  1  2  2  0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS, NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INTERCHAIN)
ACCESS     8330.0   ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)
---------1---------2---------3---------4---------5---------6---------7---------8---------9---------a---------1---------2---------3----
  1   1 - R --- G C    6.3   32.7   50.4  360.0  -73.2  180.0  360.0  360.0 X 137 -- --- - - - - - - - 5 - - --- x -  -0.10   3.42  4.
  2   2 - T --- A C    5.2   32.6   46.8 -124.4    0.2  177.6  360.0   75.4 C  19 -- --- - - - - - - - 5 - - --- x -   0.00  33.70  0.
  3   3 - D --- G C    4.0   36.1   46.3  -98.4   13.5 -177.1   52.4  110.5 C  73 -- --- - - - - - - - 4 - - --- x - ------ ------  1.
  4   4 - C --- G T    7.1   38.2   45.9  -60.4  -25.9 -175.7   87.9   35.1 A  54 -- --- - A - - - - 5 - - --- x -   0.00  29.14 -1.
```

Σχήμα 2.2 *Το αρχείο που περιέχει τις πληροφορίες για τις πρωτεΐνες και το οποίο βρίσκεται στο link RECORD της PFF . Η ακολουθία των αμινοξέων για κάθε πρωτεΐνη βρίσκεται στο πεδίο SEQPDB.*

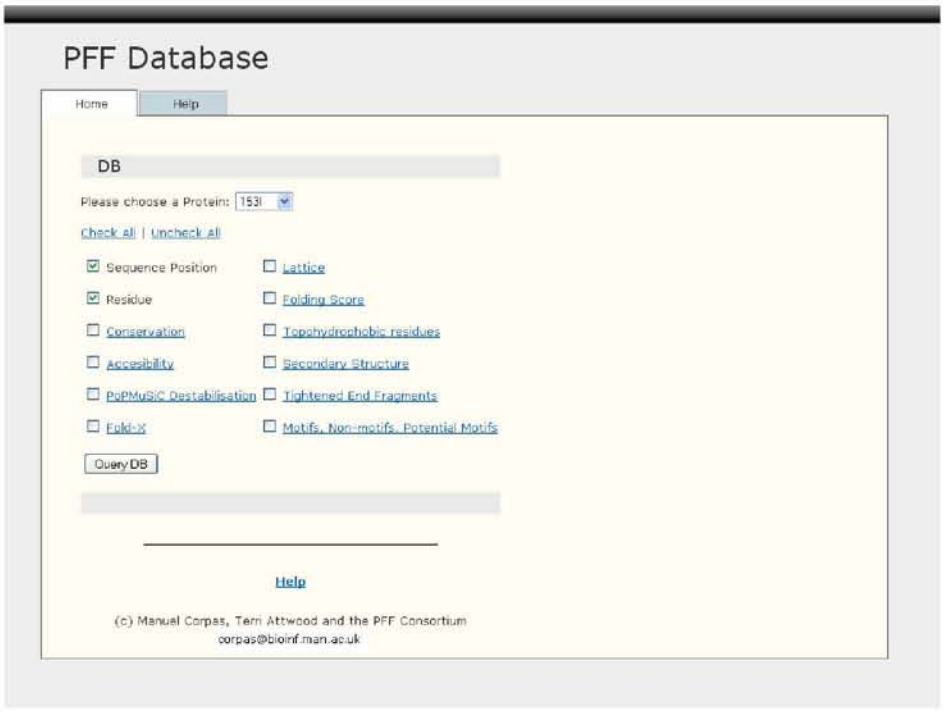
Σ' αυτό λοιπόν το σύνολο των πρωτεϊνών υπολογίστηκαν τα MIR και εν συνεχεία οι περιοχές MCF οι οποίες συγκρίθηκαν με τις γνωστές εκ των προτέρων TEF περιοχές προκειμένου να διαπιστωθεί κατά πόσο η τεχνική που χρησιμοποιείται στην εργασία, μπορεί να προβλέψει με σχετική ακρίβεια την τριτοταγή δομή των σφαιρικών πρωτεϊνών.

2.2 Βάση Δεδομένων PFF (Protein Folding Fragments)

Παρόλη την μεγάλη έρευνα, η κατανόηση του τρόπου με τον οποίο αναδιπλώνονται οι πρωτεΐνες παραμένει ένας τομέας μεγάλου ενδιαφέροντος. Οι «καρποί» αυτής της μαζικής έρευνας είναι: (α) μέθοδοι πρόβλεψης της πιθανούς δομής που θα αποκτήσουν οι αμινοξικές ακολουθίες ή της προσομοίωσης της διαδικασίας αναδίπλωσης και (β) βάσεις δεδομένων που περιέχουν δομική πληροφορία (π.χ. συντεταγμένες στο χώρο, κατηγοριοποιήσεις αναδίπλωσης, δεδομένα για τη δομή κλπ). Στην προσπάθεια αυτή να κατανοηθούν οι βασικές αρχές της πρωτεϊνικής αναδίπλωσης, αναπτύχθηκε μια νέα, ολοκληρωμένη πηγή πληροφοριών σχετικά με τη δομή των πρωτεϊνών η οποία είναι βασισμένη σε ένα μικρό υποσύνολο της βάσης δεδομένων PDB(Protein Data Bank) [32]. Η πηγή αυτή περιέχει πληροφορίες οι οποίες προέρχονται από συνδυασμό εργαλείων ανάλυσης της ακολουθίας, λογισμικού για την ανάλυση της δομής και αλγορίθμων προσομοίωσης της αναδίπλωσης. Προκειμένου τα περιεχόμενα να γίνουν περισσότερο προσβάσιμα στην ευρύτερη κοινότητα αναπτύχθηκε μια εύχρηστη διεπαφή η οποία οπτικοποιεί τα δεδομένα. Το κίνητρο για τον συνδυασμό δεδομένων από τις διάφορες προσεγγίσεις είναι η παροχή πληροφοριών για τον ρόλο που παίζουν οι διάφοροι τύποι αμινοξέων και τμημάτων στην πρωτεϊνική αναδίπλωση ώστε να βελτιωθεί η κατανόηση των παραγόντων που είναι κρίσιμης σημασίας για την διαδικασία της αναδίπλωσης εν γένει.

"Protein Folding Fragments" είναι το όνομα του consortium ευρωπαϊκών εργαστηρίων που μελετά τα πρώτα στάδια της πρωτεϊνικής αναδίπλωσης. Το PFF Project ξεκίνησε το 2003 για να προτείνει μια ενοποιημένη θεωρία των δομικών τμημάτων και να μελετήσει την εφαρμογή τους στον τομέα του protein engineering. Στο project αυτό συμμετέχουν πολλές επιστημονικές ομάδες από την Ευρώπη και το Ισραήλ ώστε να βοηθήσουν να δημιουργηθεί ένα εργαλείο πρόβλεψης της δομής. Αυτή η πρωτοβουλία προσπαθεί να εφαρμόσει την θεωρία των τμημάτων και να αναλύσει τους μηχανισμούς της πρωτεϊνικής αναδίπλωσης με στόχο την πρόβλεψη της δομής, αναγνωρίζοντας τα κρίσιμα τμήματα των πρωτεϊνών. Χρησιμοποιήθηκαν διάφορες επιστημονικές προσεγγίσεις όπως προσεγγίσεις βιοπληροφορικής, τοπολογικές προσεγγίσεις και χρήση συναρτήσεων ενέργειας ώστε να αναγνωριστούν τα δομικά τμήματα πρωτεϊνών γνωστής τριτοταγούς δομής καθώς και οι ξεχωριστές θέσεις αμινοξέων που παίζουν σημαντικό ρόλο στο δίπλωμα. Αυτές οι προσεγγίσεις στοχεύουν στην ab initio πρόβλεψη της δομής [33].

Έτσι λοιπόν δημιουργήθηκε μια βάση δεδομένων που περιέχει πληροφορίες για τη δομή και ταυτόχρονα είναι επαρκώς σχολιασμένη (σχήμα 2.3). Επίσης η version 1.0 της PFF είναι προσπελάσιμη σε DSSP-flat-file format από τη διεύθυνση http://babylone.ulb.ac.be/LIFE



Σχήμα 2.3. *Το site της βάσης PFF που βρίσκεται στη διεύθυνση* [*http://www.bioinf.manchester.ac.uk/corpas/db/index.php*](http://www.bioinf.manchester.ac.uk/corpas/db/index.php)

Κάθε εγγραφή της βάσης διαθέτει πληροφορίες όπως την *αρίθμηση* και το *όνομα* των αμινοξέων, μία *τιμή συντήρησης* για κάθε αμινοξύ της ακολουθίας, *accessibility* σε Angstrom η οποία δείχνει τον βαθμό των εσωτερικών περιορισμών στους οποίους υπόκειται η ευελιξία της δομής των πρωτεϊνών, *τιμές Popmusic* που δείχνουν την μέγιστη τιμή αποσταθεροποίησης για κάθε θέση στην πρωτεϊνική δομή, *σκορ Fold-X* που αντιπροσωπεύει μια ποσοτική εκτίμηση της σημαντικότητας των αλληλεπιδράσεων οι οποίες συνεισφέρουν στην σταθερότητα των πρωτεϊνών και *τιμές προσομοίωσης πλέγματος* που δείχνουν το πλήθος των πλησιέστερων γειτόνων κατά τη διάρκεια του διπλώματος.

Αυτές οι τιμές βασίζονται σε ab initio υπολογισμούς μεταξύ μη γειτονικών καταλοίπων κατά μήκος της πολυπεπτιδικής αλυσίδας. Για τα κατάλοιπα αυτά χρησιμοποιείται μια απλοποιημένη αναπαράσταση Cα με βάση το απλό κυβικό πλέγμα. Ο αλγόριθμος ξεκινά από μια εκτεταμένη αρχική στερεοδιάταξη με την πολυπεπτιδική αλυσίδα να αναδιπλώνεται σύμφωνα με έναν συγκεκριμένο αλγόριθμο Monte – Carlo και οι ενδιάμεσες στερεοδιατάξεις καταγράφονται μέχρι να οδηγηθούμε στην τελική στερεοδιάταξη. Τα αμινοξέα τα οποία κατά τη διάρκεια του αλγορίθμου Monte – Carlo συμμετέχουν σε μεγάλο πλήθος αλληλεπιδράσεων (τυπικά μεγαλύτερο του έξι) ονομάζονται MIR ενώ αυτά που έχουν το μικρότερο πλήθος αλληλεπιδράσεων (μικρότερο του τρία) λέγονται LIR. Η αναγνώριση των LIR και MIR είναι πολύ χρήσιμη για την αναγνώριση των περισσότερο και λιγότερο καθοριστικών αμινοξέων στην διαδικασία της προσομοίωσης του διπλώματος. Τα MIR αναμένεται να είναι περισσότερο συντηρημένα από τον μέσο όρο, σε αντίθεση με τα LIR.

Όλες οι παραπάνω πληροφορίες που αναφέραμε και που είναι διαθέσιμες για κάθε αμινοξύ στην βάση PFF βρέθηκε ότι συσχετίζονται και μάλιστα συνθέτουν το λεγόμενο "*folding score*". Ο χρήστης έχει πρόσβαση στην τιμή του folding score για κάθε αμινοξύ μέσω της βάσης. Άλλες πληροφορίες που διατίθενται μέσω της βάσης είναι η ύπαρξη η όχι *τοπο-υδρόφοβων αμινοξέων, στοιχεία δευτεροταγούς δομής, ύπαρξη TEF,*καθώς και η *ύπαρξη μοτίβων*. Μια τυπική εγγραφή της βάσης φαίνεται στο σχήμα 2.4 [32].

Position	Residue	Cons.	Access.	PsiMuSiC	Fold-X	Lattice	Folding Score	Topo	2d Str.	TEFs	Mot
1	V	1	149	0.683	-0.430448	5	12.409	-	C	- -	1
2	S	0.807	98	6.4	1.05844	4	11.8949	-	H	- -	1
3	G	0.786	52	1.663	1.77391	4	11.5792	-	H	- -	1
4	L	1	132	19.639	0.6982	5	11.8322	-	H	- -	1
5	N	1	107	3.699	0.481298	4	12.1651	-	H	- -	1
6	N	1	98	3.699	1.2471	4	13.6761	-	H	- -	1
7	A	0.51	44	39.99	-0.0465058	6	13.3884	-	H	- -	1
8	V	1	76	0	0.362154	7	13.6446	-	H	- -	1
9	Q	1	129	8.726	-0.220217	5	13.7502	-	H	- -	1
10	N	1	81	0	0.235971	5	13.4079	-	H	- -	1
11	L	1	109	19.639	0.0978651	7	12.1849	-	H	- -	1
12	Q	1	114	8.726	0.138112	5	12.2421	-	H	- -	1
13	V	1	115	5.312	1.29939	5	13.5265	-	H	- -	1
14	E	1	117	11.406	-0.169979	4	14.0048	-	H	- -	1
15	I	1	88	0	-0.271648	5	13.768	-	H	- -	1
16	G	1	13	34.31	-0.250793	3	12.9455	-	C	- -	1
17	N	1	81	0	1.88791	2	13.0473	-	C	- -	1
18	N	1	116	2.84	1.3438	2	13.1092	-	S	- -	1
19	S	0.807	106	8.288	1.26533	2	12.0859	-	S	- -	1
20	A	0.693	50	0	1.32428	2	12.3005	-	S	- -	1
21	G	1	23	39.376	0.914986	3	12.1491	-	S	- -	1
22	I	1	99	8.332	0.309556	4	12.7169	-	H	- -	1
23	K	1	93	0	0.481263	3	12.1726	-	H	- -	1
24	G	1	11	15.701	0.73183	4	11.6679	-	H	- -	1
25	Q	1	133	8.726	0.245367	4	11.1729	-	H	- -	1

Σχήμα 2.4*. Η εγγραφή της βάσης PFF για την πρωτεΐνη 1aa0.*

Έτσι λοιπόν η βάση PFF συγκέντρωσε παλαιότερες αναλύσεις και πληροφορίες ώστε να είναι δημόσια διαθέσιμες. Ο κύριος στόχος του PFF consortium ήταν να δημιουργηθεί ένα γενικά αποδεκτό εργαλείο «πρόβλεψης» που θα συνδυάζει τα πλεονεκτήματα διαφορετικών μεθόδων. Μάλιστα, αποδείχτηκε ότι η ενσωμάτωση διαφορετικών μεθόδων όντως έχει μεγαλύτερη αξία απ ότι να χρησιμοποιηθεί η κάθε μέθοδος ξεχωριστά. Μάλιστα, η παραγωγή του folding score καταφέρνει να σκιαγραφεί τις περιοχές οι οποίες είναι πιθανό να συμβάλλουν (α) στη σταθερότητα του διπλώματος και (β) στη λειτουργία της πρωτεΐνης [33]. Εμείς στην παρούσα εργασία, χρησιμοποιήσαμε την πληροφορία για την ακολουθία της κάθε μιας από τις 107 πρωτεΐνες, από την version 1.0 της PFF.

2.3 Βάση Δεδομένων PDB (Protein Data Bank)

Στην παρούσα ενότητα θα αναφέρουμε κάποιες πληροφορίες σχετικά με το τι περιέχει και πώς είναι οργανωμένη η βάση δεδομένων PDB. Στοιχεία από αυτή την βάση χρησιμοποιήσαμε τόσο για να πάρουμε πληροφορίες σχετικά με το σε ποιες θέσεις στην αμινοξική ακολουθία βρίσκονται οι TEF περιοχές όσο και για να παράγουμε τα κατάλληλα αρχεία προκειμένου να είναι δυνατή η οπτικοποίηση των αποτελεσμάτων της εργασίας. Πληροφορίες σχετικά με την οπτικοποίηση των προβλέψεων που έγινε, αναφέρονται στην ενότητα 2.8.

Η Protein Data Bank (PDB) (σχήμα 2.5) είναι μια αποθήκη δεδομένων της τρισδιάστατης δομής των πρωτεϊνών και των νουκλεϊκών οξέων. Αυτά τα δεδομένα, που κυρίως προέρχονται από την κρυσταλλογραφία ακτίνων Χ ή την NMR φασματοσκόπηση και στέλνονται στην βάση από βιολόγους και βιοχημικούς από ολόκληρο τον κόσμο, μπορούν να ανακτηθούν από την βάση δωρεάν [34,35].

Σχήμα 2.5 *Η Protein Data Bank το site της οποίας βρίσκεται στη διεύθυνση* [*http://www.rcsb.org/pdb/home/home.do*](http://www.rcsb.org/pdb/home/home.do)

Η PDB δημιουργήθηκε το 1971 στο Brookhaven National Laboratories (BNL) σαν μια συλλογή αρχείων για βιολογικές μακρομοριακές κρυσταλλικές δομές. Αρχικά, αυτή η συλλογή αρχείων περιείχε συνολικά επτά δομές και κάθε χρόνο πρόσθεταν περισσότερες, με αποκορύφωμα την δεκαετία του ‘80 . Μέχρι τις αρχές της δεκαετίας του 1990 η πλειοψηφία των επιστημονικών περιοδικών απαιτούσαν κωδικό πρόσβασης για την βάση PDB, γεγονός που καταμαρτυρεί την αλματώδη επέκταση της βάσης [34,35].

Η αρχική χρήση της βάσης περιορίζονταν σε μια μικρή ομάδα ειδικών που ασχολούνταν με έρευνες σχετικά με την δομή. Σήμερα οι χρήστες της βάσης είναι βιολόγοι, χημικοί, επιστήμονες που ασχολούνται με την πληροφορική, εκπαιδευτικοί και εκπαιδευόμενοι όλων των βαθμίδων. Λόγω της αυξημένης χρήσης της βάσης, προέκυψε η ανάγκη για καινούριους τρόπους συλλογής, οργάνωσης και διανομής των δεδομένων.

Τον Οκτώβριο του 1998 η διαχείριση της PDB ανατέθηκε στο Research Collaboratory for Structural Bioinformatics (RCSB) . Το όραμα του RCSB είναι η δημιουργία μιας πηγής βασισμένης στην πιο μοντέρνα τεχνολογία που να διευκολύνει τη χρήση και την ανάλυση των δομικών δεδομένων ώστε να δημιουργηθεί μια βοηθητική πηγή για βιολογική έρευνα [34]. Στη συνέχεια θα περιγράψουμε την διαδικασία της απόκτησης και επεξεργασίας των δεδομένων.

Ένα βασικό συστατικό για την δημιουργία μιας δημόσιας συλλογής αρχείων πληροφοριών είναι η αποτελεσματική επεξεργασία των δεδομένων. Η επεξεργασία αυτή αποτελείται από την αποστολή, σχολιασμό και επικύρωση των δεδομένων. Στο παρόν σύστημα, τα δεδομένα αποστέλλονται μέσω e-mail ή μέσω του AutoDep Input Tool (ADIT) που αναπτύχθηκε από το RCSB. Αφού η δομή σταλεί μέσω του ADIT, αποστέλλεται αυτόματα ένα αναγνωριστικό PDB στον αποστολέα. Στη συνέχεια, η εγγραφή σχολιάζεται. Αυτή η διαδικασία περιλαμβάνει τη χρήση του ADIT ώστε να διευκολυνθεί η εύρεση σφαλμάτων ή ασυνεπειών στα αρχεία. Αμέσως μετά, στέλνεται πίσω στον αποστολέα η πλήρως σχολιασμένη εγγραφή μαζί με την πληροφορία επικύρωσης. Αφού λοιπόν ξαναδεί ο αποστολέας την επεξεργασμένη εγγραφή, ενδέχεται να αποστείλει αναθεωρήσεις. Αυτή η διαδικασία μπορεί να επαναλαμβάνεται μέχρι την τελική έγκριση του αποστολέα οπότε και η εγγραφή είναι έτοιμη για διανομή.

Όλες αυτές οι διαδικασίες επεξεργασίας των δεδομένων καταγράφονται και αποθηκεύονται στο αρχείο αλληλογραφίας. Με αυτό τον τρόπο το προσωπικό της PDB μπορεί να ανακτά πληροφορίες για κάθε πτυχή της διαδικασίας της αποστολής των δεδομένων αλλά και να παρακολουθεί την αποτελεσματικότητα των εργασιών που γίνονται στη βάση.

Για κάθε εγγραφή αποθηκεύονται στη βάση πληροφορίες σχετικά με την τρέχουσα κατάσταση της, όπως αποστολέας, τίτλος και κατηγορία οι οποίες είναι προσβάσιμες μέσω queries σε WWW interface.

Τα δεδομένα που αποθηκεύονται στη βάση θεωρούνται πρωτεύοντα δεδομένα και περιλαμβάνουν πέρα από τις συντεταγμένες, γενικές πληροφορίες που χρειάζονται για όλες τις αποθηκευμένες δομές καθώς και συγκεκριμένες πληροφορίες σχετικά με την μέθοδο που χρησιμοποιήθηκε προκειμένου να αναγνωριστεί η δομή.

Η επικύρωση των δεδομένων αναφέρεται στη διαδικασία υπολογισμού της ποιότητας των αποθηκευμένων ατομικών μοντέλων καθώς και το πόσο καλά αυτά τα μοντέλα ταιριάζουν με τα πειραματικά δεδομένα. Σε περίπτωση που ανιχνευτούν σοβαρά

λάθη, γίνονται διορθώσεις μέσω της επικοινωνίας με τους αποστολείς. Αυτή η διαδικασία μπορεί να γίνει πριν ακόμα οι δομές αποθηκευτούν στη βάση μέσω ενός server επικύρωσης.

Όσον αφορά την αρχιτεκτονική της βάσης, αυτή αποτελείται από ένα ενιαίο σύστημα ετερογενών βάσεων. Την σχεσιακή βάση που διαχειρίζεται από το Sybase, τα τελικά αρχεία δεδομένων, τις POM – based βάσεις, την Biological Macromolecule Crystallization Database και τον Netscape LDAP server. Η επικοινωνία ανάμεσα σε αυτές τις διαφορετικές βάσεις γίνεται χρησιμοποιώντας το Common Gateway Interface. Ένα ολοκληρωμένο web interface αποστέλλει το ερώτημα στην κατάλληλη βάση η οποία το εκτελεί. Κάθε βάση επιστρέφει τα PDB αναγνωριστικά που ικανοποιούν το ερώτημα και το CGI πρόγραμμα ολοκληρώνει τα αποτελέσματα. Τα ερωτήματα μπορούν να σταλούν μέσω τριών interfaces: του Status Query, του SearchLite και του Search-Fields.

Η βάση διανέμει στους χρήστες της αρχεία συντεταγμένων, αρχεία δομών και αρχεία NMR περιορισμών. Επίσης, παρέχει τεκμηρίωση καθώς και επεξεργασμένα δεδομένα [34].

Μέσω της PDB βρέθηκε ένας τρόπος για τη συλλογή, την επεξεργασία και την διανομή δεδομένων σχετικά με την δομή των μακρομορίων. Μέχρι τις 24 Ιουνίου 2008 η βάση περιείχε 51.491 εγγραφές ατομικών συντεταγμένων, 47.526 από τις οποίες αφορούν πρωτεΐνες, και οι υπόλοιπες εγγραφές είναι σχετικές με νουκλεϊκά οξέα και άλλα μόρια. Αξίζει να αναφέρουμε ότι περίπου 5000 νέες δομές προστίθενται κάθε χρόνο ενώ υπολογίζεται ότι το 2014 το πλήθος των δομών θα φτάσει τις 150000, γεγονός που δείχνει την αλματώδη ανάπτυξη της βάσης [36].

PDB File Format

Τα PDB αρχεία, είναι αρχεία κειμένου που περιγράφουν τις τρισδιάστατες δομές των μορίων και κυρίως πρωτεϊνών και φυσικά αποθηκεύονται στη βάση. Περιέχουν συνήθως εκατοντάδες μέχρι χιλιάδες γραμμές και κάθε γραμμή ξεκινά με μία ετικέτα. Οι γραμμές που ξεκινάνε με ATOM περιγράφουν τις συντεταγμένες των ατόμων που αποτελούν την πρωτεΐνη. Οι γραμμές που ξεκινάνε με HETATM περιγράφουν τις συντεταγμένες ατόμων που δεν αποτελούν τμήμα της πρωτεΐνης. Οι γραμμές που ξεκινάνε με SEQRES δείχνουν τις ακολουθίες των τριών πεπτιδικών αλυσίδων, σε REMARK μπαίνουν σχόλια και γενικές πληροφορίες και τα HEADER, TITLE και AUTHOR παρέχουν πληροφορίες σχετικά με τους ερευνητές που καθόρισαν την δομή [37].

2.4 Αλγόριθμος υπολογισμού των MIR (Mostly Interacting Residues)

Ο κύριος παράγοντας που προδιαγράφει την βιολογική λειτουργία των σφαιρικών πρωτεϊνών είναι η τρισδιάστατη γεωμετρία τους. Αυτή η τρισδιάστατη τελική μορφή δημιουργείται σε χρόνο μεταξύ του 1ms και του 1 λεπτού μετά το σχηματισμό της πεπτιδικής αλυσίδας και εξαρτάται από την αλληλουχία των αμινοξέων που υπάρχουν κατά μήκος της. Οι δυνάμεις που καθορίζουν την πρωτεϊνική γεωμετρία είναι οι μη – ομοιοπολικές αλληλεπιδράσεις μεταξύ μη συνεχόμενων αμινοξέων και μεταξύ των αμινοξέων και του υδατικού περιβάλλοντος της πρωτεΐνης (υδροφοβική δύναμη).

Η διαδικασία αναδίπλωσης είναι υπερβολικά γρήγορη όσον αφορά το αστρονομικό πλήθος των πιθανών χωρικών διαμορφώσεων. Συγκεκριμένα, μία τυπική πρωτεΐνη 100 αμινοξέων αποτελείται από πολλές χιλιάδες άτομα, ο δε αριθμός των δυνατών στερεοδιατάξεων της είναι της τάξης του 3^{100}. Σαν συνέπεια, η εύρεση της στερεοδιάταξης με την ελάχιστη ενέργεια που είναι και το ζητούμενο δεν είναι εφικτή με τα σημερινά υπολογιστικά δεδομένα. Γι αυτό το λόγο, εκτελούμε προσομοιώσεις οι οποίες είναι πολύ χρήσιμες για την πρόβλεψη των κανόνων αναδίπλωσης και καταφεύγουμε σε απλοποιημένα μοντέλα και σε συνδυασμούς διαφόρων μεθόδων, που οδηγούν τουλάχιστον στην πρόβλεψη επιμέρους στοιχείων της πρωτεϊνικής δομής.

Στο εργαστήριο γενετικής του Γεωπονικού Πανεπιστημίου Αθηνών έχει αναπτυχθεί μια τεχνική προσομοίωσης Monte - Carlo που, με μόνη πληροφορία την αμινοξική ακολουθία μίας πρωτεΐνης, μπορεί να προβλέψει τα (υδρόφοβα κατά κανόνα) αμινοξέα που έχουν κρίσιμη σημασία για το σχηματισμό του πυρήνα που σταθεροποιεί την τριτοταγή δομή. Η τεχνική αυτή περιλαμβάνει συν τοις άλλοις έναν αλγόριθμο για την απαρχής πρόβλεψη της πρωτεϊνικής δομής, ο οποίος βασίζεται μόνο στην ακολουθία και στη γνώση των μη ομοιοπολικών δυνάμεων μεταξύ αμινοξέων και έχει την ακόλουθη γενική μορφή:

<u>Είσοδος</u>

1. Αμινοξική ακολουθία.
2. Αλληλεπιδράσεις μεταξύ αμινοξέων.
3. Επίδραση του περιβάλλοντος νερού.

<u>Έξοδος</u>

Πληροφορίες για τη δομή μέσω της αναζήτησης της στερεοδιάταξης ελάχιστης ενέργειας.

Ο αλγόριθμος αυτός, ο οποίος χρησιμοποιήθηκε στην παρούσα εργασία, υπολογίζει τα αμινοξέα που έχουν την τάση να σχηματίσουν τον πυρήνα της πρωτεϊνικής δομής. Τα συγκεκριμένα αμινοξέα ονομάζονται Ισχυρά Αλληλεπιδρώντα Αμινοξέα ή MIR

(Mostly Interacting Residues). Η σύγκριση της προσομοίωσης με αποτελέσματα από την ανάλυση γνωστών δομών 100 περίπου πρωτεϊνών από τη βάση πρωτεϊνικών δομών PDB (Protein Data Bank) έδωσε πολύ ενθαρρυντικά αποτελέσματα. Η γνώση των συγκεκριμένων αμινοξέων είναι ιδιαίτερα χρήσιμη για την πρόβλεψη της δομής.

Τον αλγόριθμο αυτόν τον κατηγοριοποιούμε ως *ab initio*, και αυτό διότι χρησιμοποιεί ως αρχική πληροφορία μόνο την αμινοξική ακολουθία, καθώς και ένα μοντέλο για τις μη ομοιοπολικές αλληλεπιδράσεις μεταξύ αμινοξέων, που είναι υπεύθυνες για την αναδίπλωση και δεν λαμβάνει υπόψη τα στοιχεία της δευτεροταγούς δομής των πρωτεϊνών. Οι αλληλεπιδράσεις αυτές αφορούν σε κατάλοιπα που δεν είναι γειτονικά κατά μήκος της αλυσίδας, αλλά είναι γειτονικά στο χώρο. Παρακάτω θα παραθέσουμε μία λεπτομερή περιγραφή της τεχνικής προσομοίωσης Monte – Carlo η οποία περιλαμβάνει τρία βασικά στοιχεία:

(α) το μοντέλο διακριτού χώρου για την αναπαράσταση της πολυπεπτιδικής αλυσίδας.

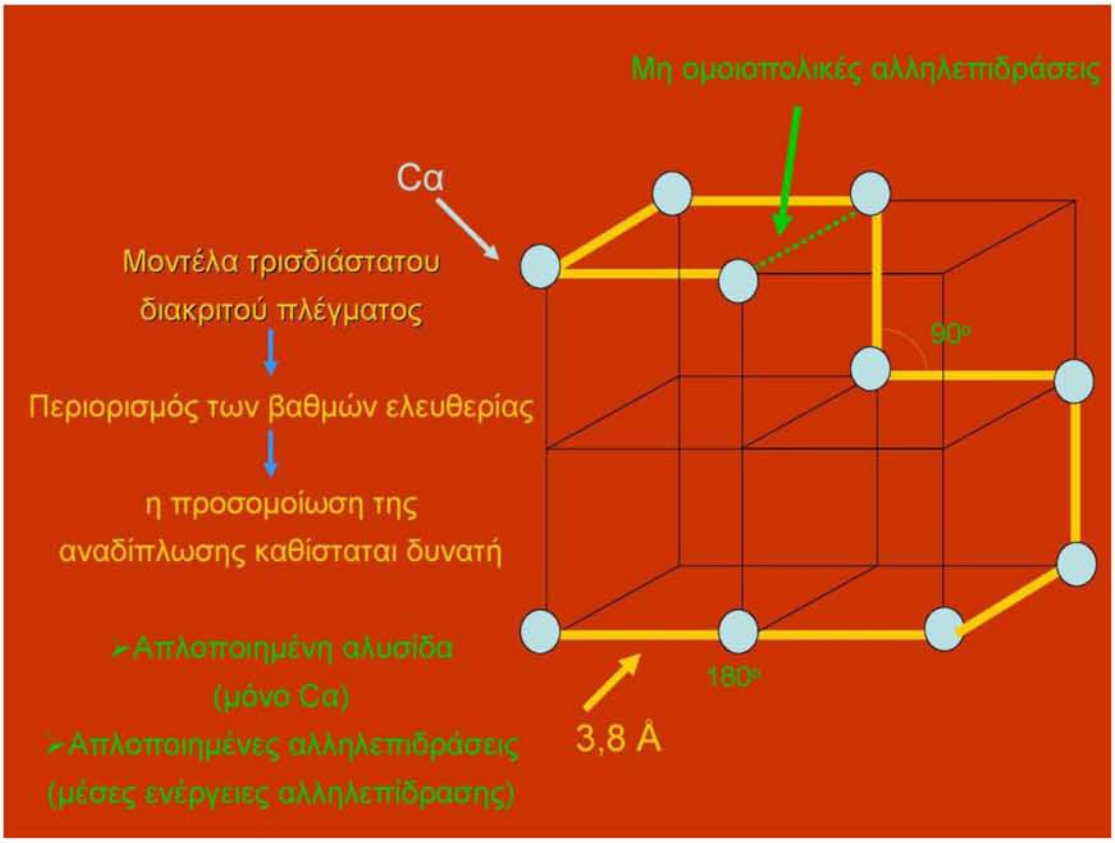
(β) τις αλληλεπιδράσεις ανάμεσα στα άτομα της αλυσίδας.

(γ) τον δυναμικό αλγόριθμο Monte - Carlo.

Με στόχο την ρεαλιστικότερη προσομοίωση της πρωτεϊνικής δομής, κρατώντας όμως την απλότητα της προσέγγισης του διακριτού χώρου, αναπτύχθηκαν διάφορα μοντέλα. Εμείς χρησιμοποιήσαμε ένα από τα πιο πολύπλοκα και συγκεκριμένα το μοντέλο 2 – 1 – 0 το οποίο πρωτοπαρουσιάστηκε από τους ερευνητές το 1991 με πολύ καλά αποτελέσματα [38]. Στο σημείο αυτό θα κάνουμε μία λεπτομερή περιγραφή του μοντέλου αυτού όπως πρωτοπαρουσιάστηκε τότε και στη συνέχεια θα αναφέρουμε κάποιες απλουστεύσεις στις οποίες προχωρήσαμε για την εργασία μας.

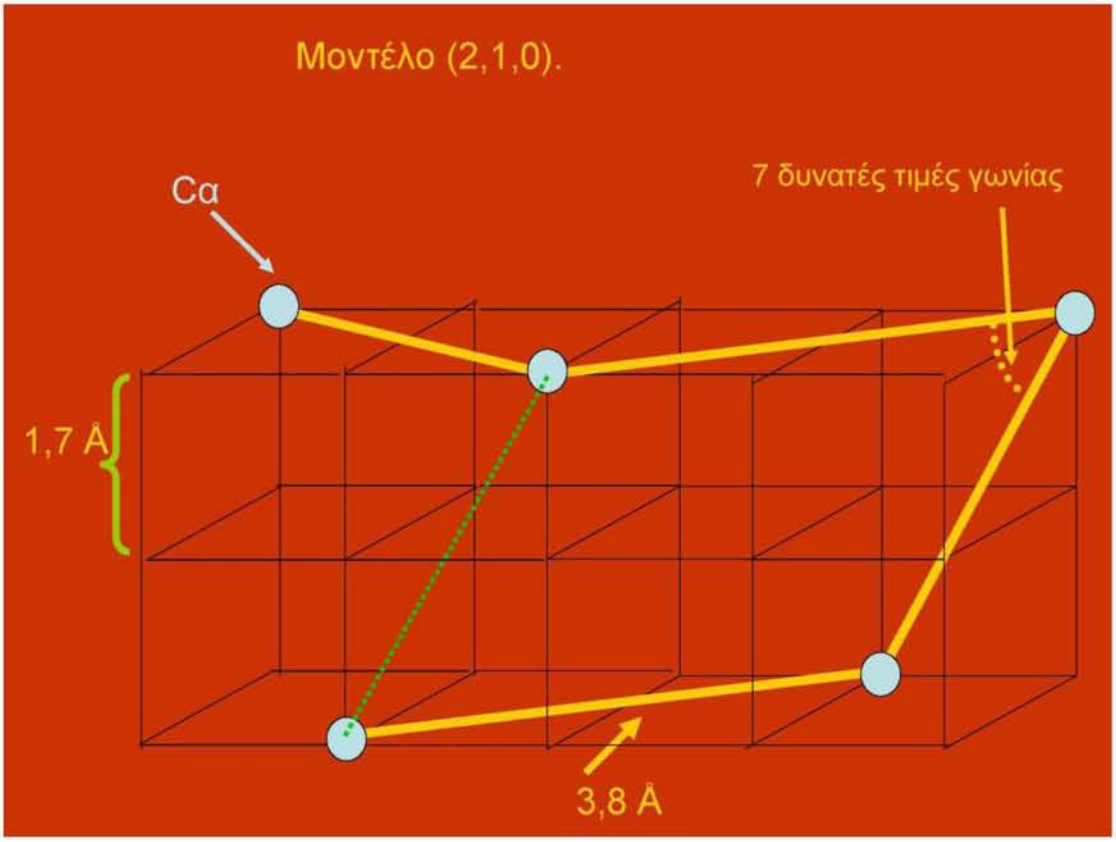
(α) Γεωμετρία της αλυσίδας

Η πρωτεΐνη αποτελείται από αναπαράσταση Cα ατόμων καθώς και Cβ για την πλευρική αλυσίδα για κάθε αμινοξύ. Το όλο σύστημα χρησιμοποιεί σαν βάση το απλό κυβικό πλέγμα το οποίο κατασκευάζεται από διανύσματα της μορφής ($\pm$ 1,0,0) όπως φαίνεται στο σχήμα 2.6 (στο σχήμα φαίνεται μόνο η αναπαράσταση Cα όπου χρησιμοποιήθηκε στην εργασία).



Σχήμα 2.6 *Το απλό κυβικό μοντέλο το οποίο χρησιμοποιεί μόνο αναπαράσταση ανθράκων-α για την πολυπεπτιδική αλυσίδα.*

Στο μοντέλο 2-1-0 δύο διαδοχικά αμινοξέα συνδέονται με διανύσματα της μορφής $(\pm 2,\pm 1,0)$ με όλους τους δυνατούς συνδυασμούς π.χ. (2,-1,0), (-1,-2,0), (1,0,2) κλπ. όπως φαίνεται στο σχήμα 2.7 (στο σχήμα φαίνεται μόνο η αναπαράσταση Cα όπου χρησιμοποιήθηκε στην εργασία). Για κάθε αμινοξύ, υπάρχουν γύρω του 24 τέτοιες θέσεις στον τρισδιάστατο χώρο. Αυτό σημαίνει ότι η απόσταση ανάμεσα σε δύο διαδοχικά σημεία στο κυβικό πλέγμα, δηλαδή το μοναδιαίο διάνυσμα του πλέγματος, είναι 1.7 Å.



Σχήμα 2.7 *Το 2-1-0 μοντέλο το οποίο χρησιμοποιεί μόνο αναπαράσταση ανθράκων-α για την πολυπεπτιδική αλυσίδα.*

Ο σχηματισμός backbone (δηλαδή της κύριας αλυσίδας) του i – οστού ατόμου Cα ορίζεται από το τετράγωνο της απόστασης r^2_θ ανάμεσα στους άνθρακες-α i-1 και i+1 με θ τη γωνία δεσμού ανάμεσα σε τρεις συνεχόμενους άνθρακες-α. Σε μονάδες πλέγματος όπου η απόσταση ανάμεσα σε διαδοχικούς άνθρακες-α ισούται με $(2^2+1^2+0^2)^{1/2} = 5^{1/2}$ (3.785 Å), τιμές με φυσική σημασία για τον όρο r^2_θ είναι οι 6,8,10,12,14,16 και 18 καθώς το r^2_θ εκτείνεται σε μια κλίμακα από το 7 ως το 19 για πραγματικές πρωτεΐνες. Από την κορυφή του κεντρικού άνθρακα-α, η πλευρική αλυσίδα σχηματίζεται από 4 σημεία του πλέγματος. Τρία είναι διανύσματα της μορφής $(\pm1,\pm1,0)$ και το τέταρτο της μορφής $(\pm1,\pm1\pm1)$, διάνυσμα το οποίο είναι το κέντρο της υδρόφοβης ή υδρόφιλης αλληλεπίδρασης. Ο σχηματισμός της πλευρικής αλυσίδας εξαρτάται από το r^2_θ. Για τους υπολογισμούς που γίνονται παρακάτω είτε τα κατάλοιπα είναι γλυκίνη οπότε δεν υπάρχει πλευρική αλυσίδα είτε τα κατάλοιπα έχουν πλευρική αλυσίδα ομοιόμορφου μεγέθους. Παρακάτω θα αναλύσουμε τις αλληλεπιδράσεις μεταξύ των ατόμων που λαμβάνουν χώρα στο μοντέλο αυτό [38].

(β) Αλληλεπιδράσεις

Για μια αλυσίδα που αποτελείται από n κατάλοιπα σε μια αναπαράσταση ανθράκων-α, ο εκάστοτε σχηματισμός περιγράφεται από n-2 γωνίες δεσμού θ και n-3 γωνίες συστροφής (torsional angles) φ. Προκειμένου τα πρώτα και τα τελευταία κατάλοιπα να έχουν έναν ορισμένο σχηματισμό, σε κάθε άκρο της αλυσίδας έχει προστεθεί ένα εικονικό κατάλοιπο. Αυτά τα εικονικά κατάλοιπα είναι τελείως αδρανή και απλά καταλαμβάνουν χώρο. Έτσι, με την προσθήκη των καταλοίπων αυτών έχουμε n γωνίες δεσμού και n-1 γωνίες συστροφής. Για ευκολία η αλυσίδα απαριθμείται από το 1 ως το N.

Η εγγενής προτίμηση σχηματισμού για αυτή την κλάση των μοντέλων αναπαρίσταται από την προτίμηση καθενός καταλοίπου για τις διάφορες τιμές των γωνιών δεσμού. Αυτές περιγράφονται από τον όρο r^2_θ. Αφού για κάθε κατάλοιπο i υπάρχουν επτά διαφορετικές τιμές για το r^2_θ,καθορίστηκε η τοπική ενεργειακή προτίμηση $\varepsilon_\theta(r^2_{\theta i})$ για καθένα από τα r^2_θ. Προκειμένου να διατηρηθεί το πλήθος των διαφόρων παραμέτρων στο ελάχιστο, αν εξαιρέσουμε ορισμένους σχηματισμούς όπου $\varepsilon_\theta(r^2_{\theta i}) = 0$, όλοι οι υπόλοιποι σχηματισμοί έχουν $\varepsilon_\theta > 0$.

Για τα κατάλοιπα 2 ως N-2 καθορίστηκε και ένα δυναμικό δίεδρης γωνίας (dihedral angle potential). Υπάρχουν 324 καταστάσεις περιστροφής για κάθε εσωτερικό δεσμό και σε όλες έχει αντιστοιχηθεί μια σχετική ενέργεια ε_φ. Προκειμένου να γίνει αυτό όμως, πρώτα ορίστηκαν στατιστικά βάρη για κάθε μια από τις καταστάσεις αυτές μέσω μεταγενέστερων εφαρμογών σε πραγματικές πρωτεΐνες. Στην έρευνα του 1991 η πλειοψηφία των σχηματισμών θεωρήθηκε ισοενεργειακή. Η προτίμηση της μικρής και της μεσαίας εμβέλειας εκπροσωπείται από τους όρους $\varepsilon_\theta(r^2_\theta)$ και ε_φ αντίστοιχα.

Αν ορίσουμε ως r_{kl} την απόσταση ανάμεσα στο k-οστό και στο l-οστό ατομικό κέντρο ανθράκων-α που δεν συνδέονται με κάποιο δεσμό, τότε η ενέργεια απώθησης E_{rep} είναι της μορφής

$$E_{rep} = \begin{bmatrix} \infty & (r_{kl})^2 = 0,1 \\ 3\varepsilon_{rep} & (r_{kl})^2 = 3 \\ \varepsilon_{rep} & (r_{kl})^2 = 5 \\ 0 & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{bmatrix} \quad (1)$$

με το ε_{rep} να παίρνει την τιμή 6.

Στη συνέχεια περιγράφουμε τις αλληλεπιδράσεις της τριτοταγής δομής που καθορίστηκαν. Προκειμένου να προσομοιωθεί η επίδραση του υδρογονικού δεσμού και των διπολικών αλληλεπιδράσεων, παρουσιάστηκε μια συνεργατική αλληλεπίδραση για κάθε άνθρακα-α του m-οστού καταλοίπου που βρίσκεται σε απόσταση 3 μονάδων πλέγματος από τον άνθρακα-α του k-οστού καταλοίπου. Έτσι ορίστηκε το

$$E_{ckl} = \varepsilon_c \{\delta_{|bk|:|bm|} + \delta_{|bk+1|:|bm|} + \delta_{|bk|:|bm+1|} + \delta_{|bk+1|:|bm+1|}\} \quad (2)$$

όπου δ_{ij} το λεγόμενο Kronecker delta. Το ε_c εφαρμόστηκε ομοιόμορφα σε όλα τα ζεύγη καταλοίπων ανεξαρτήτως των σχηματισμών τους [38].

Ένα ζεύγος πλευρικών αλυσίδων θεωρούμε ότι βρίσκεται σε αλληλεπίδραση όταν είναι σε απόσταση $\sqrt{2}$. Υπάρχουν 12 τέτοιες περιοχές στο πλέγμα. Για τα αρχικά στάδια των υπολογισμών χρησιμοποιήθηκε μια πλήρης κλίμακα υδροφοβικότητας των Miyazawa & Jernigan και στη συνέχεια προτιμήθηκε ένα απλουστευμένο μοντέλο υδροφοβικότητας. Οι πλευρικές αλυσίδες μπορούν να είναι υδρόφοβες, υδρόφιλες ή αδρανείς. Ζεύγη υδρόφοβων πλευρικών αλυσίδων αλληλεπιδρούν μ' ένα ελκτικό στατιστικό πεδίο δυνάμεων, υδρόφοβα/υδρόφιλα ζεύγη αλληλεπιδρούν μ' ένα απωθητικό στατιστικό πεδίο δυνάμεων και υδρόφιλα ζεύγη μπορεί να είναι ελαφρώς ελκτικά ή απωθητικά χωρίς να παρατηρείται διαφορά στην ποιοτική τους συμπεριφορά.

Το αμινοξύ γλυκίνη δεν έχει πλευρικές αλυσίδες και ανατίθεται δείκτης υδροφοβικότητας h(i) ίσος με το μηδέν. Τα υδρόφοβα αμινοξέα έχουν h(i) < 0 και τα υδρόφιλα h(i) > 0.Για όλες τις πλευρικές αλυσίδες με μήκος μεγαλύτερο των δύο αμινοξέων, τα στοιχεία του πίνακα αλληλεπίδρασης am(i,j) μεταξύ των ζευγών πλευρικών αλυσίδων i και j είναι της μορφής

$$am(i, j) = - h(i) * h(j) * \varepsilon \quad (3)$$

Εδώ $\varepsilon = \varepsilon_{phobe\ phobr}$ (> 0), αν h(i) και h(j) είναι και τα δύο αρνητικά, $\varepsilon = \varepsilon_{phobe\text{-}phil}$ (>0) αν ένα κατάλοιπο είναι υδρόφοβο και το άλλο υδρόφιλο και $\varepsilon = -\varepsilon_{phil\text{-}phil}$ με $\varepsilon_{phil\text{-}phil}$ >0 αν h(i) και h(j) είναι και τα δύο θετικά.

Στην προσομοίωση, οι ενεργειακές παράμετροι ε_φ, ε_θ, ε_{rep}, $\varepsilon_{phobe\ phobr}$, $\varepsilon_{phobe\text{-}phil}$ και $\varepsilon_{phil\text{-}phil}$ είναι διαβαθμισμένες ομοιόμορφα από έναν παράγοντα θερμοκρασίας T^*. Έτσι, έγινε μια απλουστευμένη εκτίμηση ότι οι σχετικοί λόγοι των διάφορων αλληλεπιδράσεων είναι ανεξάρτητοι από τη θερμοκρασία . Ακολούθως θα περιγράψουμε τον δυναμικό αλγόριθμο Monte – Carlo όπως ακριβώς δημοσιεύτηκε από τους συγγραφείς το 1991 [38].

(γ) Δυναμικός αλγόριθμος Monte – Carlo

Για την δυναμική προσομοίωση της αναδίπλωσης, στα μοντέλα διακριτού χώρου συνήθως χρησιμοποιείται ο αλγόριθμος Monte-Carlo (σχήμα 2.8).

Τα στάδιά του σε μια γενική περιγραφή είναι τα εξής:

1. Αρχικά, η πολυπεπτιδική αλυσίδα βρίσκεται σε μία τυχαία, μη αναδιπλωμένη στερεοδομή. Η αναδίπλωση προσομοιώνεται με μία σειρά από κινήσεις των αμινοξέων στους κόμβους του τρισδιάστατου πλέγματος. Στην απλούστερη προσέγγιση, σε κάθε κίνηση συμμετέχει ένα αμινοξύ.

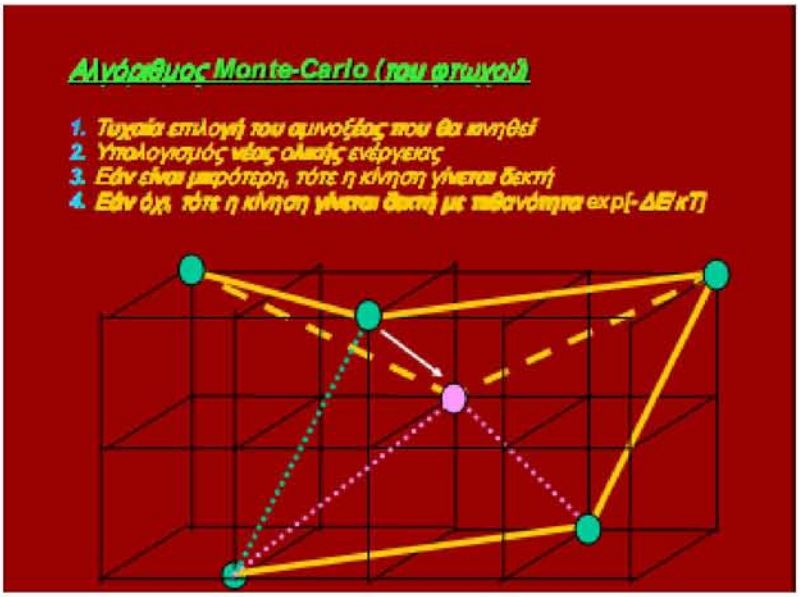
2. Υπολογίζεται η ολική ενέργεια της αρχικής στερεοδομής, σαν το άθροισμα των ενεργειών αλληλεπίδρασης μεταξύ όλων των ζευγών αμινοξέων

3. Επιλέγεται τυχαία το αμινοξύ που θα κινηθεί.

4. Εξετάζεται εάν είναι δυνατό να κινηθεί το αμινοξύ (μπορεί πχ να μην υπάρχουν ελεύθερες γειτονικές θέσεις στο πλέγμα).

5. Εάν όχι, επαναλαμβάνεται το 3ο βήμα. Εάν ναι, υπολογίζεται η ενέργεια της νέας στερεοδιάταξης.

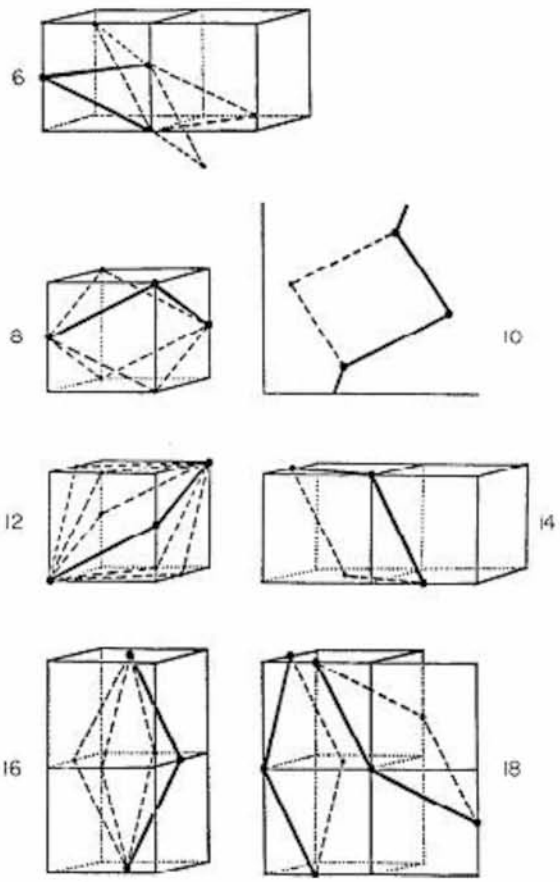
6. Εάν η νέα ενέργεια είναι χαμηλότερη της προηγούμενης, η νέα στερεοδιάταξη γίνεται αποδεκτή και συνεχίζουμε επιστρέφοντας στο βήμα 3 για την επόμενη κίνηση. Σε αντίθετη περίπτωση, υπολογίζεται η πιθανότητα να υπερπηδηθεί ο ενεργειακός φραγμός με τη θερμική κίνηση. Η πιθανότητα αυτή είναι $e^{-\Delta E/kT}$. ΔΕ είναι η διαφορά ενέργειας μεταξύ αρχικής κα νέας στερεοδομής (νέα - αρχική) k η σταθερά του Boltzmann και T η θερμοκρασία. Σαν πιθανότητα, κυμαίνεται μεταξύ 0 και 1 και είναι τόσο μεγαλύτερη, όσο μικρότερη είναι η διαφορά ενέργειας ΔΕ. Η προσομοίωση της θερμικής κίνησης γίνεται με την δημιουργία ενός τυχαίου αριθμού μεταξύ 0 και 1. Εάν αυτός είναι μικρότερος από $e^{-\Delta E/kT}$ τότε η νέα στερεοδομή γίνεται δεκτή και επιστρέφουμε στο 3ο βήμα, εάν όχι δεν γίνεται δεκτή και επιστρέφουμε στο 3ο βήμα με την προηγούμενη στερεοδιάταξη.



Σχήμα 2.8 *Σχηματική απεικόνιση του αλγορίθμου Monte - Carlo*

Με τον τρόπο αυτό η πολυπεπτιδική αλυσίδα μεταβαίνει προοδευτικά σε συμπαγέστερες στερεοδιατάξεις, με τα υδρόφοβα αμινοξέα να έχουν την τάση να σχηματίσουν τον πυρήνα.

Αξίζει να σημειώσουμε πως οι κινήσεις που επιτρέπονται για κάθε αμινοξύ δεν πρέπει να παραβιάζουν διάφορους περιορισμούς δεσμών. Έτσι επιτρέπονται (α) single bead flips ή κινήσεις "spike" όπως φαίνεται δίπλα στο σχήμα 2.9 και (β) δύο bead end flips όπου τα δύο άκρα των δεσμών μετατρέπονται σε ένα καινούριο σετ διανυσμάτων [38].

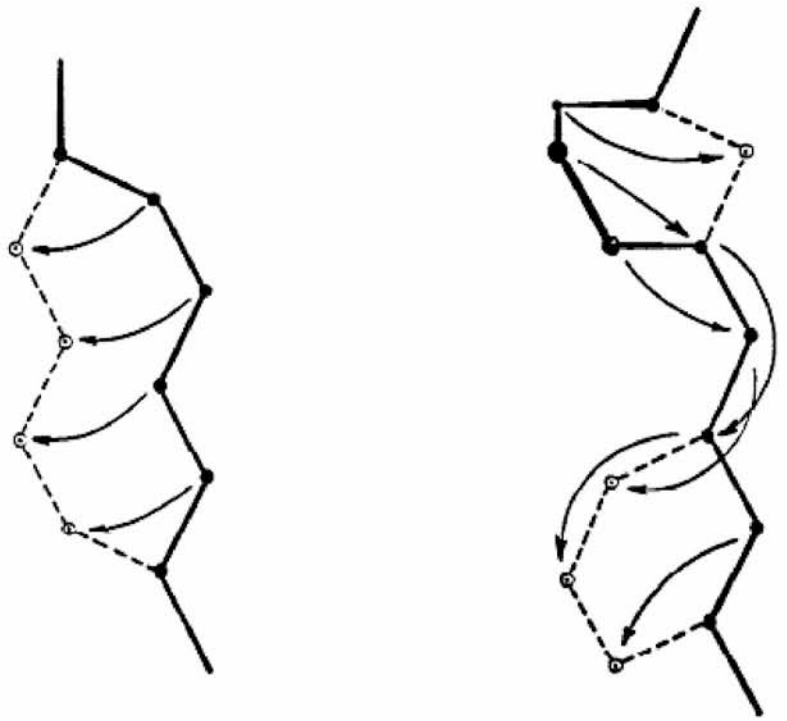


Σχήμα 2.9 *Τοπικές single bead flips που συσχετίζονται με τις διάφορες γωνίες δεσμού που αντιστοιχούν σε* r^2_θ = *6,8,10,12,14,16 και 18. Οι συνεχόμενες γραμμές δείχνουν το αρχικό σχήμα του ζεύγους των δεσμών ενώ οι διακεκομμένες τα πιθανά αποτελέσματα.*

Παρότι αυτές οι κινήσεις μπορούν να εφαρμόσουν την δυναμική του αλγορίθμου Monte-Carlo, παρουσιάζουν το μειονέκτημα ότι δε μπορούν να προσομοιώσουν ακριβώς τις κινήσεις που απαιτούνται για την αναδίπλωση συγκεκριμένων τμημάτων της δευτεροταγούς δομής και συγκεκριμένα των α-ελίκων. Γι αυτό το λόγο, επιτρέπονται και οι εξής κινήσεις:

(γ) Περιστροφές αλυσίδας σε τυχαία επιλεγμένα τμήματα όπως φαίνεται στο σχήμα 2.10α.

(δ) Κινήσεις που μοιάζουν με εσωτερικό κύμα όπως φαίνονται στο σχήμα 2.10β.



(α) (β)

Σχήμα 2.10 *Αναπαράσταση (α) εσωτερικών περιστροφών και (β) κινήσεων που μοιάζουν με κύμα. Οι συνεχόμενες γραμμές δείχνουν τον αρχικό σχηματισμό της αλυσίδας και οι διακεκομμένες τον τελικό σχηματισμό της.*

Μετά την ολοκλήρωση των κινήσεων αυτών υπολογίζεται η ενέργεια του καινούριου σχηματισμού E_{new} και συγκρίνεται με την ενέργεια του παλαιότερου σχηματισμού E_{old} χρησιμοποιώντας ένα ασύμμετρο Metropolis κριτήριο.

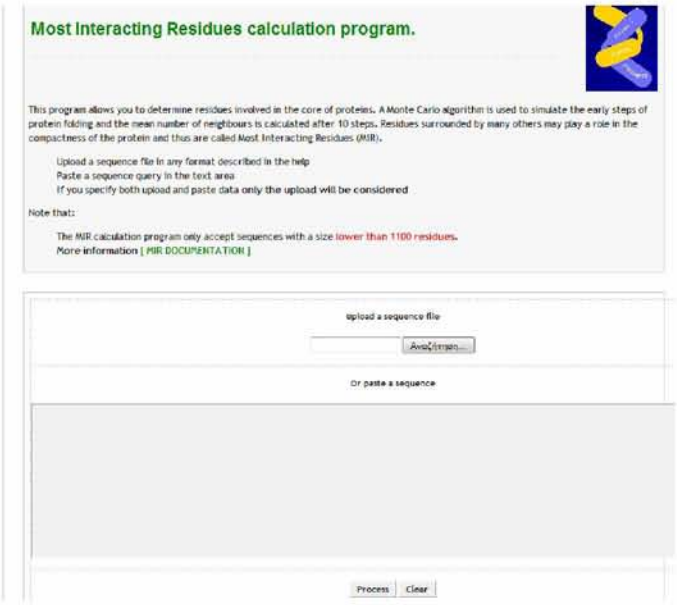
Ένα χρονικό βήμα του αλγορίθμου Monte-Carlo αποτελείται από N προσπάθειες της κίνησης (α), σε κάθε άκρο αλυσίδας επιχειρείται η κίνηση (β) καθώς και μία προσπάθεια των κινήσεων (γ) και (δ).

Το σύστημα εκκινείται σε μια τυχαία κατάσταση υψηλής θερμοκρασίας T^*. Με την πάροδο του χρόνου η θερμοκρασία αυτή μειώνεται και το όλο σύστημα εξισορροπείται μέχρι να μεταβούμε σε έναν αναδιπλωμένο σχηματισμό. Για να συμβεί αυτό απαιτούνται περίπου 1.25×10^6 βήματα του αλγορίθμου Monte-Carlo. Το σύνολο αυτό των στοιχειωδών κινήσεων ικανοποιεί μια βασική εξίσωση στοχαστικών κινήσεων η οποία περιγράφει την χρονική εξέλιξη του συστήματος και προσομοιώνει τη συνολική διαδικασία αναδίπλωσης με αρκετά καλό τρόπο.

Αφού περιγράψαμε αναλυτικά το μοντέλο 2-1-0 που πρωτοπαρουσιάστηκε το 1991 από τους Skolnick και Kolinski [38], στο σημείο αυτό θα αναφέρουμε τις απλουστεύσεις στις οποίες προχωρήσαμε στην εργασία μας καθώς και τη διαδικασία

που ακολουθήθηκε επακριβώς από τον έτοιμο αλγόριθμο που χρησιμοποιήσαμε προκειμένου να υπολογιστούν τα MIR [25,39].

Το πρόγραμμα υπολογισμού των MIR είναι διαθέσιμο on-line στη διεύθυνση http://bioserv.rpbs.jussieu.fr/cgi-bin/MIR όπου εισάγουμε μία ακολουθία και παίρνουμε την πρόβλεψη όπως φαίνεται και στο σχήμα 2.11. Η ακολουθία των αμινοξέων την οποία περνάμε σαν είσοδο υπάρχει στη βάση δεδομένων PFF: http://babylone.ulb.ac.be/LIFE/ και συγκεκριμένα από το πεδίο SEQPDB του link RECORD. Ο χρόνος υπολογισμού κυμαίνεται τυπικά από 5 λεπτά ως 1-2 ώρες ή παραπάνω, ανάλογα με το μήκος της πρωτεΐνης και τον φόρτο του server.



Σχήμα 2.11 *Το πρόγραμμα υπολογισμού των MIR. Στο γκρίζο πλαίσιο εισάγαμε την ακολουθία για την οποία θέλαμε να υπολογιστούν τα MIR και με το κουμπί «Process» ξεκινούσαμε την προσομοίωση.*

Για την προσομοίωση, το πρόγραμμα αυτό χρησιμοποιεί μια απλή αναπαράσταση άνθρακα –α για την πολυπεπτιδική αλυσίδα με βάση το κυβικό πλέγμα ακριβώς όπως φαίνεται στο σχήμα 2.6. Αναπαράσταση Cβ για την πλευρική αλυσίδα ήταν περιττή και αυτό γιατί στόχος ήταν να κρατηθεί σχετικά απλό το μοντέλο το οποίο παρόλα αυτά δίνει μία πολύ ρεαλιστικότερη αναπαράσταση της πολυπεπτιδικής αλυσίδας. Για παράδειγμα, η γωνία που σχηματίζουν τρεις διαδοχικοί άνθρακες-α μπορεί να πάρει 7 διαφορετικές τιμές, μεταξύ 66 και 143 μοιρών (όπως φαίνεται και στο σχήμα 2.7), που είναι τιμές που συναντώνται στις πραγματικές πρωτεΐνες όσον αφορά το εύρος τιμών της ψευδογωνίας τ. Αυτό επιτυγχάνεται με τον περιορισμό της απόστασης μεταξύ των αμινοξέων i και i+2 από τα 4.1 στα 7.2 Å. Επίσης, η απόσταση ανάμεσα σε διαδοχικούς ή μη άνθρακες-α μπορεί να είναι τουλάχιστον 3.8 Å και έτσι, καθίσταται δυνατή και η προβολή μίας πρωτεϊνικής δομής σε διακριτό χώρο [25,39].

Στα απλοποιημένα μοντέλα διακριτού χώρου δεν είναι δυνατή η εφαρμογή των πεδίων της Μοριακής Μηχανικής. Πράγματι, τα τελευταία κινούνται σε ατομικό επίπεδο (περιγράφουν δηλαδή δυνάμεις μεταξύ ατόμων). Όμως, στα μοντέλα πλέγματος, χρειαζόμαστε δυνάμεις μεταξύ αμινοξέων. Με άλλα λόγια, χρειαζόμαστε ένα Μέσο Πεδίο Δυνάμεων (η Ενεργειών) για κάθε ζεύγος αμινοξέων Στο πεδίο δυνάμεων λαμβάνεται υπόψη η διαφορετική φύση των αμινοξέων (μιας και το είδος των αμινοξέων δεν φαίνεται από την γεωμετρία της αλυσίδας αφού χρησιμοποιείται αναπαράσταση άνθρακα-α) και με αυτό τον τρόπο αποδίδεται διαφορετική τιμή ενέργειας για κάθε σχηματισμό αλυσίδας. Συγκεκριμένα χρησιμοποιήθηκε ο 20×20 πίνακας των Miyazawa και Jernigan ο οποίος περιέχει τις τιμές των ενεργειών για κάθε ζεύγος καταλοίπων. Ο συγκεκριμένος πίνακας προέκυψε από στατιστική ανάλυση των γνωστών πρωτεϊνικών δομών. Συγκεκριμένα, εξετάστηκε πόσο συχνά συμβαίνει να βρίσκονται κοντά στο χώρο δύο αμινοξέα συγκεκριμένων τύπων. Αν πχ ένα ζεύγος αμινοξέων (πχ ala-phe) βρίσκονται συχνά σε γειτονικές θέσεις, τότε αυτό σημαίνει πως αλληλεπιδρούν ελκτικά και τους αντιστοιχίζεται μία ισχυρή ενέργεια ελκτικής μη ομοιοπολικής αλληλεπίδρασης. Αντίστοιχοι ενεργειακοί όροι υπολογίζονται για όλα τα δυνατά ζεύγη αμινοξέων, ανάλογα με την συχνότητα εμφάνισής τους σε κοντινές θέσεις στις πρωτεϊνικές δομές. Αξίζει να σημειωθεί πως ο όρος «γειτονικές θέσεις» αναφέρεται σε γειτονικές θέσεις στο χώρο και όχι στην πολυπεπτιδική αλυσίδα. Έτσι λοιπόν, αν δύο μη διαδοχικά κατάλοιπα i και j βρίσκονται σε απόσταση μικρότερη ή ίση των 5.88 Å, ένας όρος E_{ij} προστίθεται στη συνολική ενέργεια, ανάλογα πάντα με την φύση των αμινοξέων. Αυτό το μέγιστο όριο απόστασης αλληλεπίδρασης των 5.88 Å αντιστοιχεί σε $\sqrt{12}$ μονάδες πλέγματος και φαίνεται σαν ένα λογικό όριο για την μέση μη ομοιοπολική αλληλεπίδραση μεταξύ αμινοξέων. Πέρα από αυτό τα όριο, η ενέργεια αλληλεπίδρασης είναι μηδέν [25,39].

Το πρόγραμμα αυτό χρησιμοποιεί, πέρα από το μοντέλο διακριτού χώρου 2-1-0 και το στατιστικό πεδίο δυνάμεων των Miyazawa και Jernigan, δύο ακόμα αλγορίθμους: τον αλγόριθμο Monte – Carlo όπως περιγράφηκε προηγουμένως καθώς και έναν αλγόριθμο υπολογισμού των MIR (ο οποίος ουσιαστικά ενσωματώνεται στον αλγόριθμο Monte – Carlo), λεπτομέρειες για τον οποίο θα αναφέρουμε παρακάτω. Προς το παρόν οφείλουμε να σημειώσουμε τα εξής σχετικά με τον αλγόριθμο Monte – Carlo:

Για κάθε πρωτεΐνη παρήχθησαν τυχαία 100 διαφορετικές μη αναδιπλωμένες στερεοδομές, οι οποίες χρησιμοποιήθηκαν ως σημεία εκκίνησης σε 100 διαφορετικές προσομοιώσεις ώστε να αποφευχθεί τυχόν εξάρτηση από την αρχική κατάσταση. Ο μόνος περιορισμός που τέθηκε για αυτές τις αρχικές καταστάσεις ήταν ότι αμινοξέα τα οποία βρίσκονταν μακριά το ένα από το άλλο στην πολυπεπτιδική αλυσίδα δεν επιτρέπονταν να βρίσκονται κοντά στον χώρο προκειμένου να αποτραπεί συσταδοποίηση εξαιτίας συγκεκριμένων αρχικών σχηματισμών. Αυτός ο περιορισμός εισάγει και μία ελάχιστη χωρική απόσταση dmin σύμφωνα με το διαχωρισμό Delta = $|i-j|$ ανάμεσα στα κατάλοιπα i και j: (1) Delta = $6 \div 10$, dmin = 7 Å (2) Delta =

$11 \div 15$, dmin = 11 Å (3) Delta = $16 \div 20$, dmin = 19 Å (4) Delta πάνω από 20, dmin = 27 Å [29].

Οι επιτρεπόμενες κινήσεις για κάθε αμινοξύ για το βήμα 3 του αλγορίθμου Monte – Carlo που αναφέραμε και παραπάνω ήταν δύο ειδών. Κινήσεις end flip για τα τερματικά αμινοξέα N και C και κινήσεις γωνίας (corner movements) όπως ορίζονται από τους Skolnick και Kolinski και περιγράψαμε και εμείς προηγουμένως (σχήμα 2.9 και 2.10) για τα υπόλοιπα [38,39]. Η επιλογή του σετ κινήσεων είναι περισσότερο ή λιγότερο αφηρημένη καθώς οι στοιχειώδεις αυτές κινήσεις παρότι μετακινούν ένα κατάλοιπο τη φορά είναι ικανές ώστε να φέρουν την πρωτεΐνη σε μια αναδιπλωμένη κατάσταση. Με αυτό τον τρόπο αυτός ο περιορισμός σε στοιχειώδεις κινήσεις μονάχα, πέρα από την απλότητα και πρακτικότητα που μας παρέχει, επιτρέπει και μια σειριακή ανάλυση της τάσης της αλυσίδας να σχηματίζει συμπαγή τμήματα γύρω από συγκεκριμένα αμινοξέα από την αρχή της προσομοίωσης [25]. Ο λόγος που επιτρέψαμε κινήσεις ενός αμινοξέως την φορά και όχι κινήσεις πολλαπλών αμινοξέων είναι ότι στην εργασία αυτή προσπαθούμε να αποκαλύψουμε τον ρόλο των τοπικών (υπό την έννοια της ακολουθίας) αλληλεπιδράσεων στα πρώτα στάδια του διπλώματος [39]. Πράγματι αυτές οι στοιχειώδεις κινήσεις καταφέρνουν να οδηγήσουν την πρωτεΐνη σε μια τελική αναδιπλωμένη στερεοδομή [25].

Μετά από κάθε κίνηση αμινοξέως η υπολογιζόμενη ενέργεια της στερεοδιάταξης υποβαλλόταν σε ένα πρότυπο κριτήριο Metropolis σε σταθερή θερμοκρασία που υπολογίζεται επαγωγικά όπως ακριβώς περιγράψαμε προηγουμένως στο βήμα 6 του αλγορίθμου Monte – Carlo. Αν η θερμοκρασία είναι πολύ χαμηλή η πρωτεΐνη παγώνει σε μη αναδιπλωμένες στερεοδιατάξεις καθώς η θερμική ενέργεια είναι πολύ μικρή ώστε να υπερπηδηθούν οι ενεργειακοί φραγμοί που υπάρχουν ανάμεσα στις διάφορες στερεοδιατάξεις, Από την άλλη αν η θερμοκρασία είναι πολύ υψηλή η θερμική κίνηση κάνει την αναδιπλωμένη στερεοδιάταξη ασταθή. Μια καλή τιμή για το T εμπειρικά είναι T = 1.1 [39].

Επειδή ο στόχος ήταν να αναλύσουμε την τάση των αμινοξέων να βυθίζονται από την αρχή του διπλώματος, εξασφαλίσαμε ότι το μέγιστο πλήθος των βημάτων του αλγορίθμου Monte – Carlo ήταν αρκετό ώστε να επιτραπεί ο σχηματισμός συμπαγών τμημάτων αλυσίδας. Εξαιτίας της σειριακής φύσης του αλγορίθμου, αυτό το χρονικό όριο του μέγιστου πλήθους των βημάτων σχετίζεται με το μήκος L της πρωτεϊνικής αλυσίδας. Εμπειρικά διαπιστώθηκε ότι για μικρές πρωτεΐνες μήκους 50 αμινοξέων, η τιμή t_{max} ισούται περίπου με 10^6 βήματα Monte – Carlo. Έτσι η ακόλουθη γραμμική σχέση υιοθετήθηκε ώστε να γενικεύσει τον όρο t_{max} για πρωτεΐνες οποιουδήποτε μήκους L: t_{max} = INT (10^6 L/50) όπου το INT είναι το ακέραιο μέρος μιας και ο όρος t_{max} ορίζεται ως ακέραιος αφού δείχνει τα βήματα του αλγορίθμου Monte – Carlo.

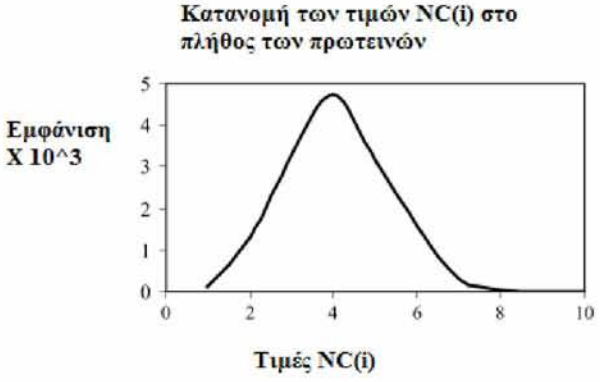
Για κάθε προσομοίωση παράγονται περίπου 10^4 εγγραφές ενδιάμεσων σχηματισμών ανά τακτά χρονικά διαστήματα. Καθώς το πλήθος των προσομοιώσεων ανά πρωτεΐνη είναι 100 (μία προσομοίωση για κάθε αρχική κατάσταση) το τελικό αποτέλεσμα είναι ένα πλήθος 10^6 αρχείων ανά πρωτεΐνη [25,39].

Για κάθε καταγραμμένη στερεοδιάταξη καθώς και για κάθε αμινοξύ υπολογίζεται το πλήθος των αμινοξέων με τα οποία βρίσκεται σε μη ομοιοπολική αλληλεπίδραση. Μη ομοιοπολική αλληλεπίδραση σε χωρικούς όρους έχουμε όταν δύο ή περισσότερα μη γειτονικά αμινοξέα βρίσκονται σε απόσταση το πολύ 5.88 Å. Για μια δεδομένη πρωτεΐνη και για το αμινοξύ i, κατά την r-οστή εγγραφή, το πλήθος των γειτόνων με τους οποίους βρίσκεται σε μη ομοιοπολική αλληλεπίδραση ορίζεται ως nc(i,r). Ο χρονικός μέσος όρος αυτής της ποσότητας είναι [25]

$$\mathrm{NC(i)} = \frac{1}{10^6}\sum_{r=1}^{10^6} nc(i,r)$$

όπου οι τιμές NC(i) στρογγυλοποιούνται στον πλησιέστερο ακέραιο. Αυτό το μέσο πλήθος των πλησιέστερων γειτόνων που βρίσκονται σε μη ομοιοπολική αλληλεπίδραση για το κάθε αμινοξύ, είναι ουσιαστικά ένα ποσοτικό μέτρο που μας δείχνει τη τάση του εκάστοτε αμινοξέως να βυθίζεται και να σχηματίζει τον πυρήνα της πρωτεΐνης. Όσο μεγαλύτερο το NC(i) τόσο μεγαλύτερη και αυτή η τάση.

Αν $\overline{NC}$ είναι η μέση τιμή του όρου NC(i) για μια συγκεκριμένη πρωτεΐνη, τα αμινοξέα για τα οποία το NC(i) είναι σημαντικά μεγαλύτερο από το $\overline{NC}$, παρουσιάζουν ιδιαίτερο ενδιαφέρον και ονομάζονται ισχυρά αλληλεπιδρώντα αμινοξέα (MIR). Προκειμένου να γίνει η επιλογή των MIR πρέπει πρώτα να οριστεί ένα κατώφλι για την τιμή $\overline{NC}$. Βρέθηκε ότι οι τιμές NC(i) κυμαίνονται μεταξύ του 1 και του 8 και ότι $\overline{NC}$ = 4 για όλες τις μελετώμενες πρωτεΐνες. Στο σχήμα 2.12 φαίνεται η κατανομή των διάφορων τιμών του NC(i) για τα κατάλοιπα των 111 πρωτεϊνών. Η πιο συχνά παρατηρούμενη τιμή απ ότι φαίνεται και στο σχήμα είναι η τιμή 4 η οποία συμπίπτει και με την τιμή $\overline{NC}$. Από την κατανομή αυτή προκύπτει ότι το 13% των αμινοξέων έχουν πλήθος τουλάχιστον 6 πλησιέστερων γειτόνων με τους οποίους βρίσκονται σε μη ομοιοπολική αλληλεπίδραση. Άρα για να θεωρείται ένα αμινοξύ i ως MIR πρέπει να έχει NC(i) τουλάχιστον 6 [25].



Σχήμα 2.12 *Η κατανομή των διάφορων τιμών του NC(i) για τα κατάλοιπα των 111 πρωτεϊνών.*

Συνοψίζοντας λοιπόν, ο αλγόριθμος υπολογισμού των MIR ακολουθεί τα παρακάτω βήματα:

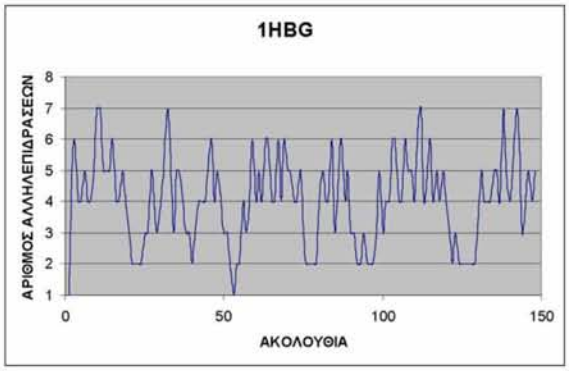
1. Ο αλγόριθμος "τρέχει" τόσο χρόνο ώστε να προσομοιώσει τα αρχικά στάδια της αναδίπλωσης (τυπικά 10^6 βήματα).

2. Σε τακτικά διαστήματα (κάθε 10 βήματα πχ) και για κάθε αμινοξύ i, υπολογίζεται ο αριθμός των αμινοξέων που αλληλεπιδρούν με αυτό δηλαδή τα αμινοξέα που βρίσκονται σε μη ομοιοπολική αλληλεπίδραση εντός εμβέλειας 5.88 Å. Μια και μιλάμε για μη ομοιοπολική αλληλεπίδραση, δεν περιλαμβάνουμε τα γειτονικά αμινοξέα i+1 και i-1.

3. Για κάθε αμινοξύ υπολογίζουμε τον μέσο αριθμό αυτών των αλληλεπιδράσεων κατά την διάρκεια της προσομοίωσης.

4. Παράδειγμα: Για μία προσομοίωση με το μοντέλο 2-1-0 στην αιμοσφαιρίνη (1HBG), η κατανομή του αριθμού των αλληλεπιδράσεων ως προς την αμινοξική ακολουθία δίνεται από το σχήμα αυτό. (σχήμα 2.13)

5. Τα αμινοξέα με αριθμό αλληλεπιδράσεων μεγαλύτερο η ίσο του 6 λέγονται Ισχυρά Αλληλεπιδρώντα Αμινοξέα ή MIR (= Mostly Interacting Residues). Σχεδόν στο σύνολό τους τα MIR είναι υδρόφοβα.



Σχήμα 2.13 *Η κατανομή του αριθμού των αλληλεπιδράσεων ως προς την αμινοξική ακολουθία της αιμοσφαιρίνης.*

Θα πρέπει να σημειώσουμε πως τα MIR υπολογίστηκαν τόσο για την αρχική ακολουθία κάθε πρωτεΐνης όσο και για μεταλλαγμένες ακολουθίες της. Συγκεκριμένα, σε κάθε θέση MIR της αρχικής ακολουθίας εισάγαμε κάποιες μεταλλάξεις, αντικαταστήσαμε δηλαδή το κάθε αμινοξύ με αλανίνη και αν ήταν εξαρχής αλανίνη με το αμινοξύ γλυκίνη, και στη νέα μεταλλαγμένη ακολουθία υπολογίσαμε ξανά τα MIR. Οι μεταλλάξεις δεν έγιναν όλες μαζί αλλά μία τη φορά, δηλαδή κάθε μεταλλαγμένη ακολουθία διέφερε από την αρχική κατά ένα αμινοξύ μονάχα και συγκεκριμένα σε θέση MIR. Αφού υπολογίστηκαν τα MIR και γι αυτές τις μεταλλαγμένες ακολουθίες, χρησιμοποιήσαμε την πληροφορία αυτή και μέσω του αλγορίθμου που παρουσιάζεται στην ενότητα 2.5, παρήχθησαν οι MCF περιοχές που αποτελούν και την πρόβλεψη μας για τις TEF περιοχές. Περισσότερα περί της

θεωρίας των μεταλλάξεων που χρησιμοποιήθηκε στην παρούσα εργασία αναφέρονται στην ενότητα 4.1

2.5 Αλγόριθμος παραγωγής των MCF (Mutation Correlation Fragments) περιοχών - Λογικό Διάγραμμα.

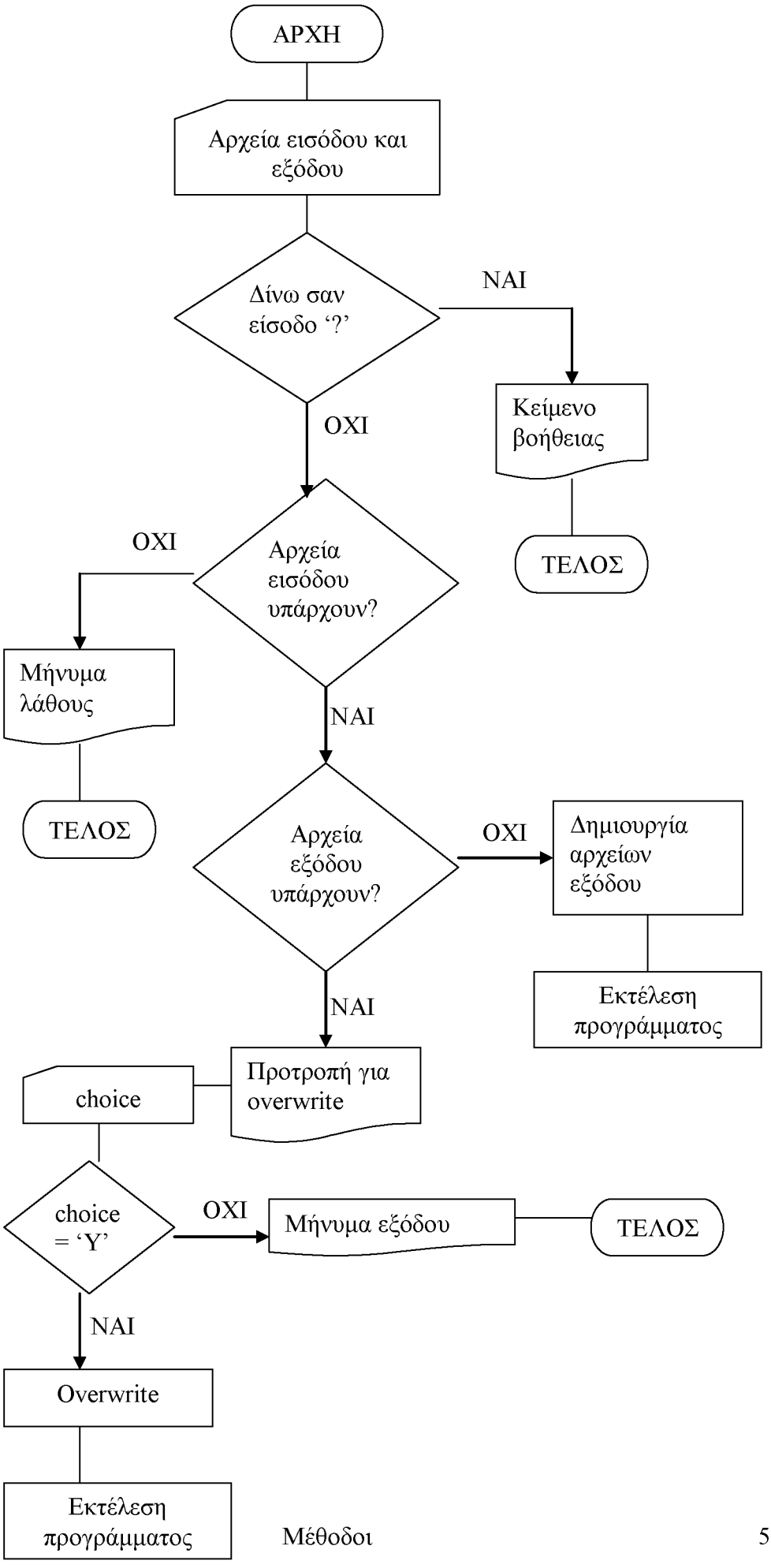
Στην ενότητα αυτή θα παραθέσουμε το λογικό διάγραμμα του αλγορίθμου που χρησιμοποίησε την πληροφορία από τα MIR και αναπτύχθηκε προκειμένου να παραχθούν οι MCF περιοχές. Συγκεκριμένα, ο αλγόριθμος αυτός χρησιμοποιώντας τις θέσεις MIR της αρχικής ακολουθίας καθώς και τις πληροφορίες που προκύπτουν από τις μεταλλάξεις που εισάγαμε σε κάθε θέση MIR καθώς και τα νέα MIR που προκύπτουν στις μεταλλαγμένες ακολουθίες, υπολογίζει τις αρχικές περιοχές MCF. Αυτές στη συνέχεια ελέγχονται για τυχόν επικαλύψεις και με βάση ορισμένα κριτήρια κάποιες ενδεχομένως να απορρίπτονται και κάποιες άλλες να υποσημειώνονται ως «υπό εξέταση». Έτσι, πέρα από το σύνολο των αρχικών περιοχών MCF, παράγονται και κάποια υποσύνολα αυτού, το σύνολο των απορριπτέων, των «υπό εξέταση» και των υπόλοιπων περιοχών. Τα δύο τελευταία σύνολα αποτελούν και την πρόβλεψη μας για τις περιοχές TEF.

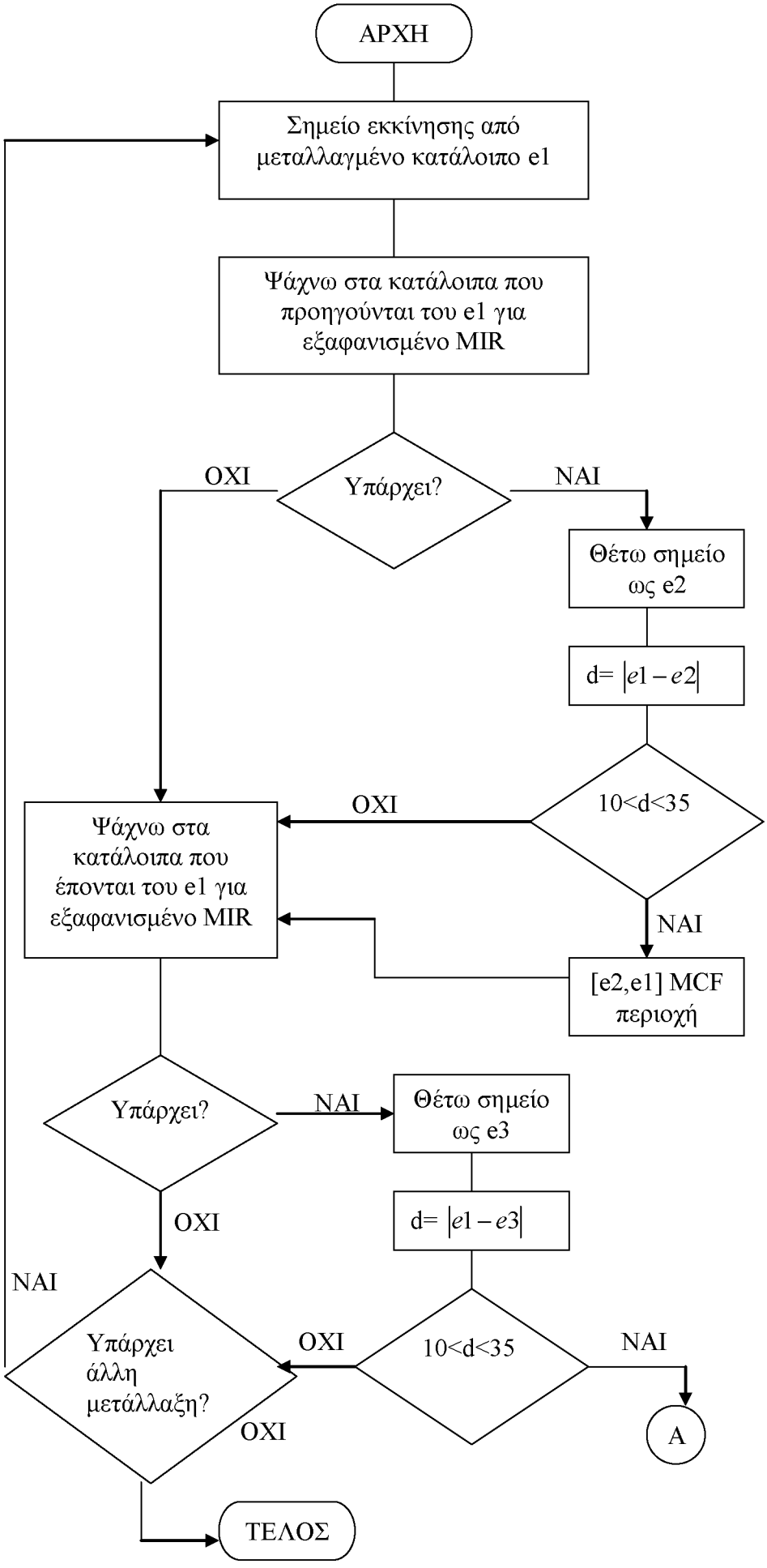
Ο αλγόριθμος υπολογισμού των αρχικών, τελικών καθώς και των υπό εξέταση MCF περιοχών αναπτύχθηκε σε γλώσσα C. Το πρόγραμμα που χρησιμοποιήθηκε ήταν το Dev C++ έκδοση 4.9.9.2

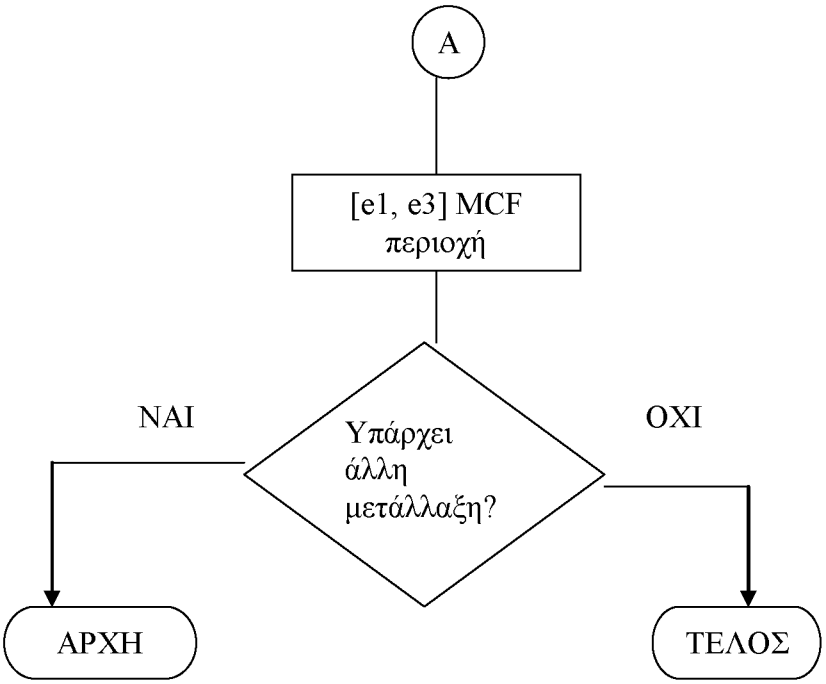
Παρακάτω σχεδιάστηκαν τα λογικά διαγράμματα για τον αλγόριθμο αυτόν. Το πρώτο λογικό διάγραμμα δείχνει την αλληλεπίδραση του προγράμματος με τον χρήστη. Το πρόγραμμα διαβάζει ένα csv αρχείο και παράγει κάποια txt αρχεία. Ο αλγόριθμος σχεδιάστηκε έτσι ώστε να εκτελείται αποκλειστικά σε περιβάλλον windows. Τα csv αρχεία πρέπει να έχουν σαν *character set* το western Europe iso-8859-1 και σαν *text delimeter* το κενό. Το δεύτερο διάγραμμα δείχνει τον αλγόριθμο με τον οποίο παράγονται οι αρχικές MCF περιοχές. Το τρίτο διάγραμμα δείχνει τον αλγόριθμο με τον οποίο παράγονται οι τελικές MCF περιοχές και το τέταρτο δείχνει πώς προκύπτει το σύνολο των υπό εξέταση MCF περιοχών.

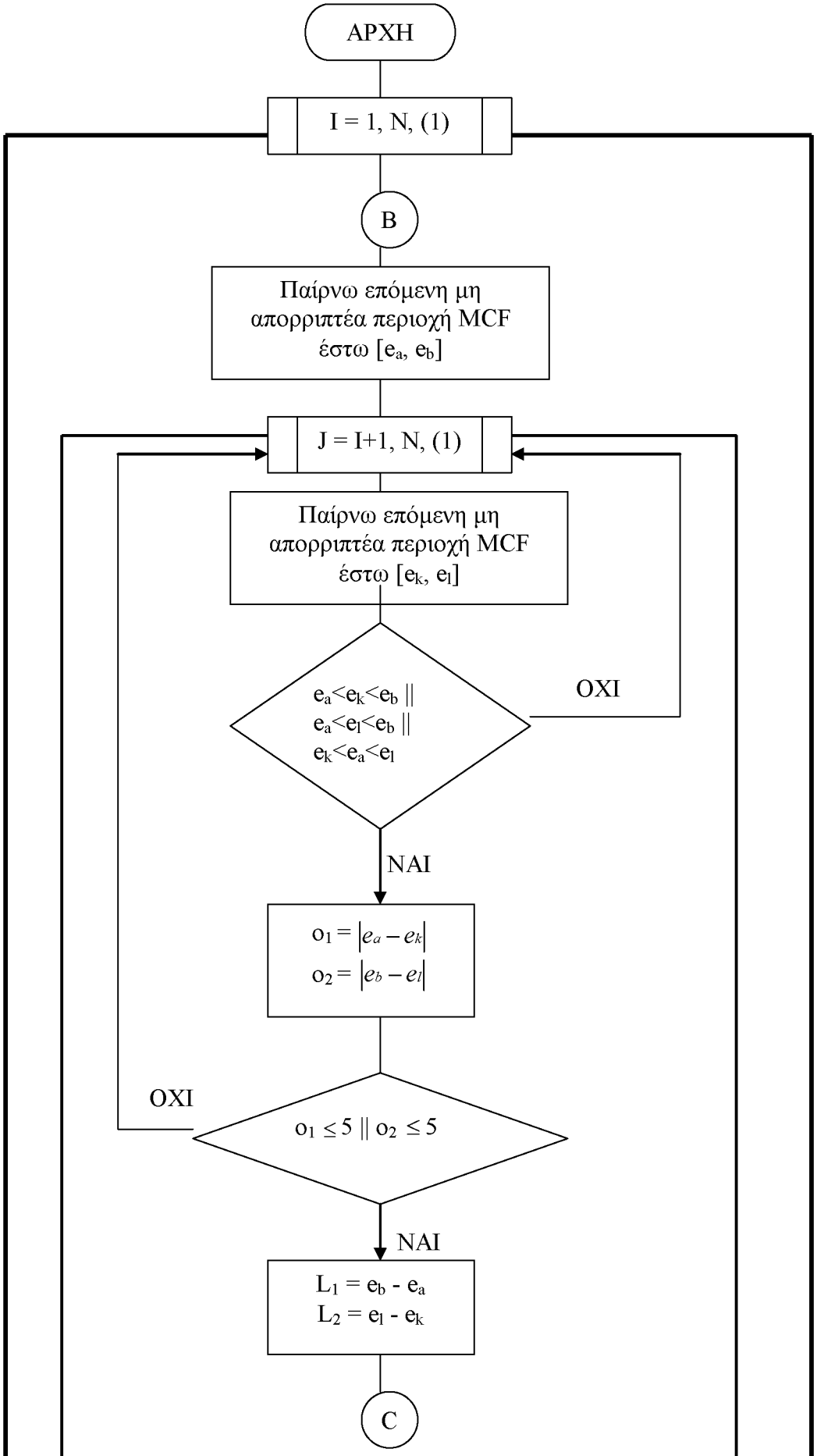
Στο σημείο αυτό θα πρέπει να γίνουν κάποιες απαραίτητες διευκρινήσεις όσον αφορά τα λογικά διαγράμματα ώστε να γίνει πιο εύκολη η ανάγνωση τους. Το δεύτερο λογικό διάγραμμα απαρτίζεται από δύο σελίδες. Η ετικέτα A δείχνει πώς ο αλγόριθμος συνεχίζει στην επόμενη σελίδα. Το τρίτο λογικό διάγραμμα επίσης συνεχίζεται σε δύο σελίδες μέσω της ετικέτας C. Η ετικέτα B στην δεύτερη σελίδα δείχνει πως βγαίνουμε από την εσωτερική επανάληψη και συνεχίζουμε με την εξωτερική αυξάνοντας τον μετρητή. Η διαφορά στο πάχος ορισμένων γραμμών είναι καθαρά για λόγους ευκολότερης ανάγνωσης. Τα ίδια ισχύουν και για το τελευταίο διάγραμμα.

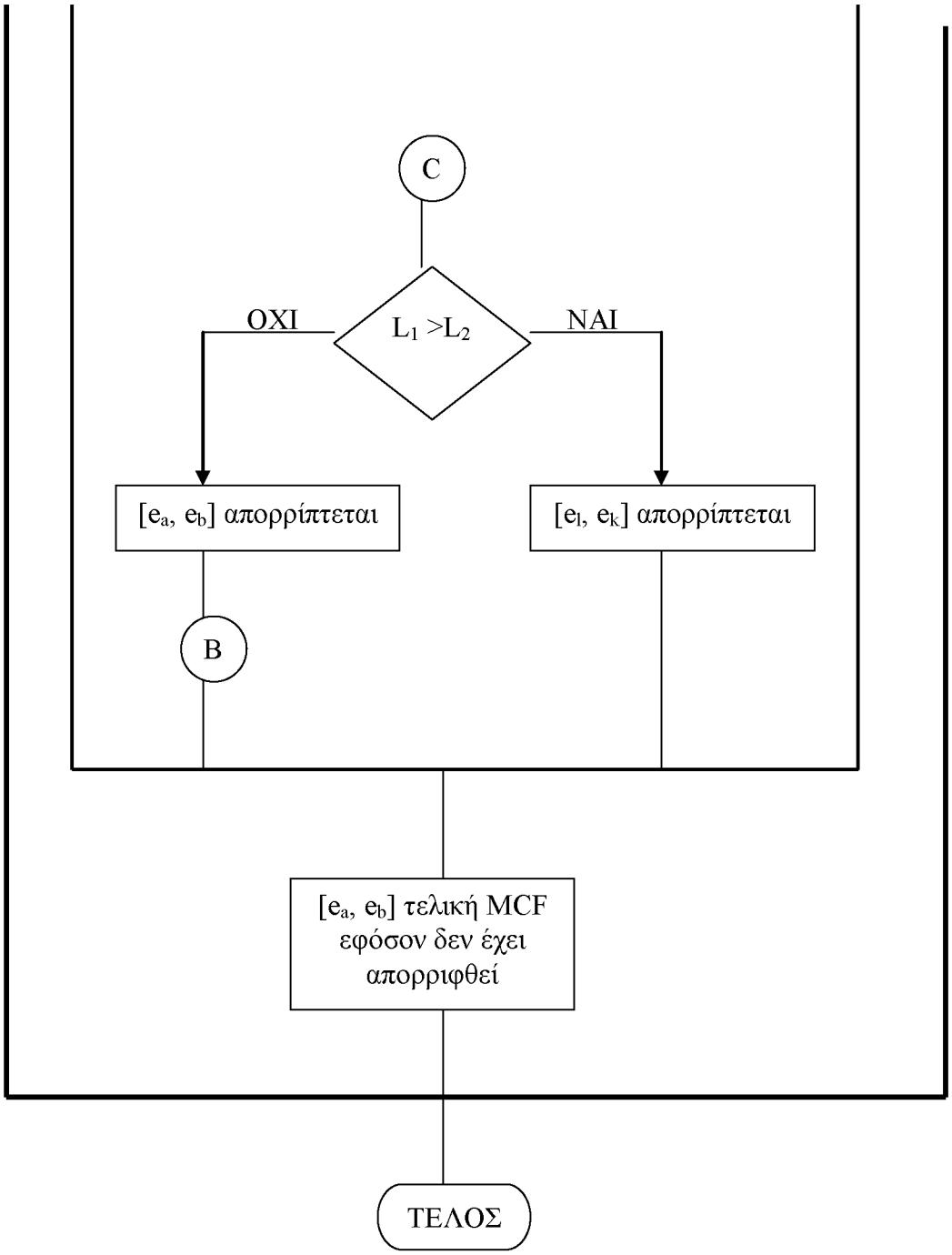
Τέλος, στο τρίτο διάγραμμα όπου N είναι το πλήθος των μεταλλάξεων ενώ στο τελευταίο όπου N έχουμε το πλήθος των αρχικών MCF περιοχών. Κάθε διάγραμμα ξεκινά από το αντίστοιχο σύμβολο start και τερματίζεται μέσω του συμβόλου end.

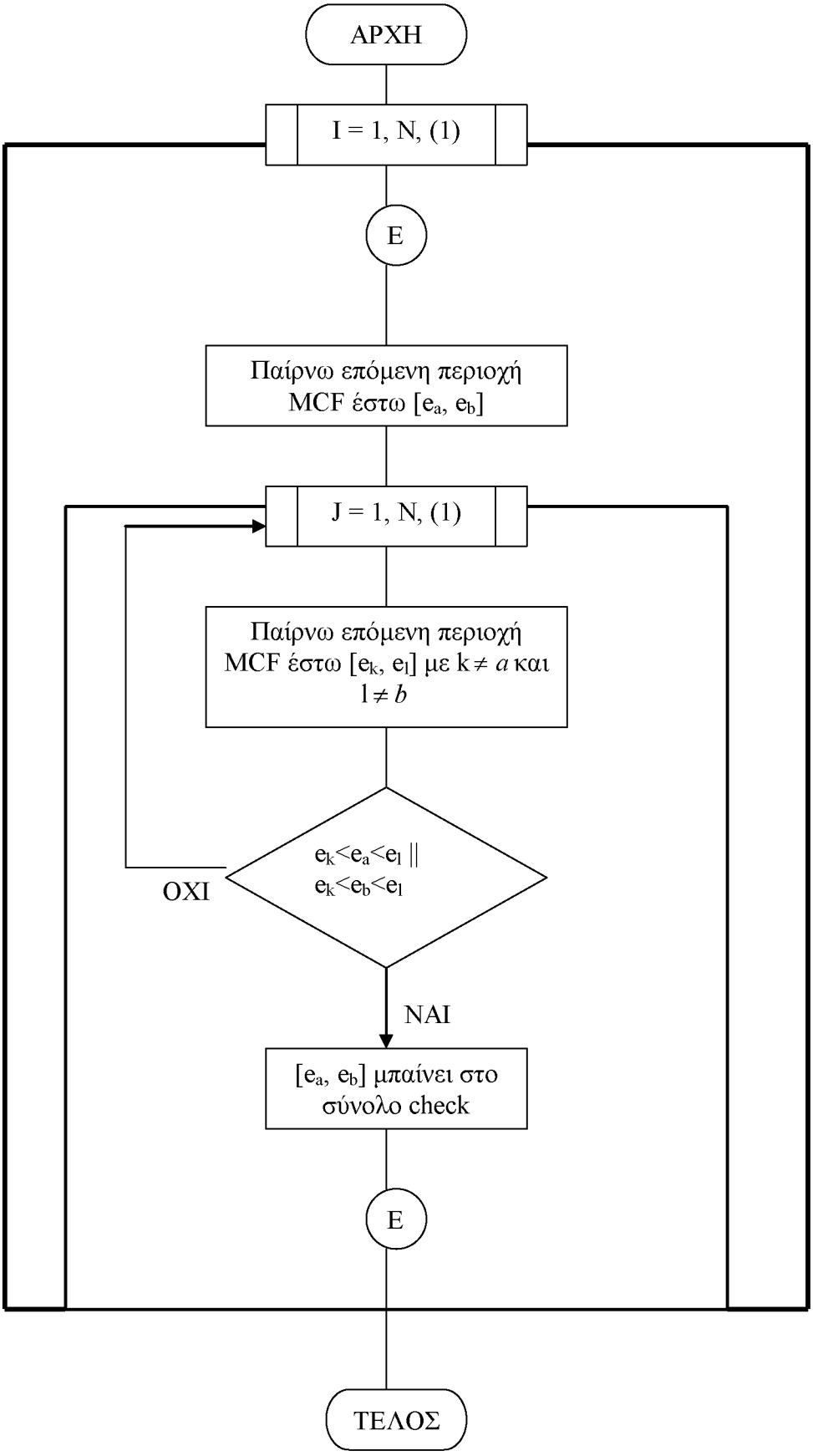

ΑΡΧΗ
Αρχεία εισόδου και εξόδου
Δίνω σαν είσοδο '?'
ΝΑΙ
Κείμενο βοήθειας
ΤΕΛΟΣ
ΟΧΙ
Αρχεία εισόδου υπάρχουν?
ΟΧΙ
Μήνυμα λάθους
ΤΕΛΟΣ
ΝΑΙ
Αρχεία εξόδου υπάρχουν?
ΟΧΙ
Δημιουργία αρχείων εξόδου
Εκτέλεση προγράμματος
ΝΑΙ
choice
Προτροπή για overwrite
choice = 'Y'
ΟΧΙ
Μήνυμα εξόδου
ΤΕΛΟΣ
ΝΑΙ
Overwrite
Εκτέλεση προγράμματος


ΑΡΧΗ
Σημείο εκκίνησης από μεταλλαγμένο κατάλοιπο e1
Ψάχνω στα κατάλοιπα που προηγούνται του e1 για εξαφανισμένο MIR
Υπάρχει?
ΟΧΙ
ΝΑΙ
Θέτω σημείο ως e2
d= |e1 – e2|
10<d<35
ΟΧΙ
ΝΑΙ
Ψάχνω στα κατάλοιπα που έπονται του e1 για εξαφανισμένο MIR
[e2,e1] MCF περιοχή
Υπάρχει?
ΝΑΙ
Θέτω σημείο ως e3
d= |e1 – e3|
ΟΧΙ
ΝΑΙ
Υπάρχει άλλη μετάλλαξη?
ΟΧΙ
10<d<35
ΟΧΙ
ΝΑΙ
Α
ΤΕΛΟΣ


A
[e1, e3] MCF περιοχή
Υπάρχει άλλη μετάλλαξη?
ΝΑΙ
ΟΧΙ
ΑΡΧΗ
ΤΕΛΟΣ

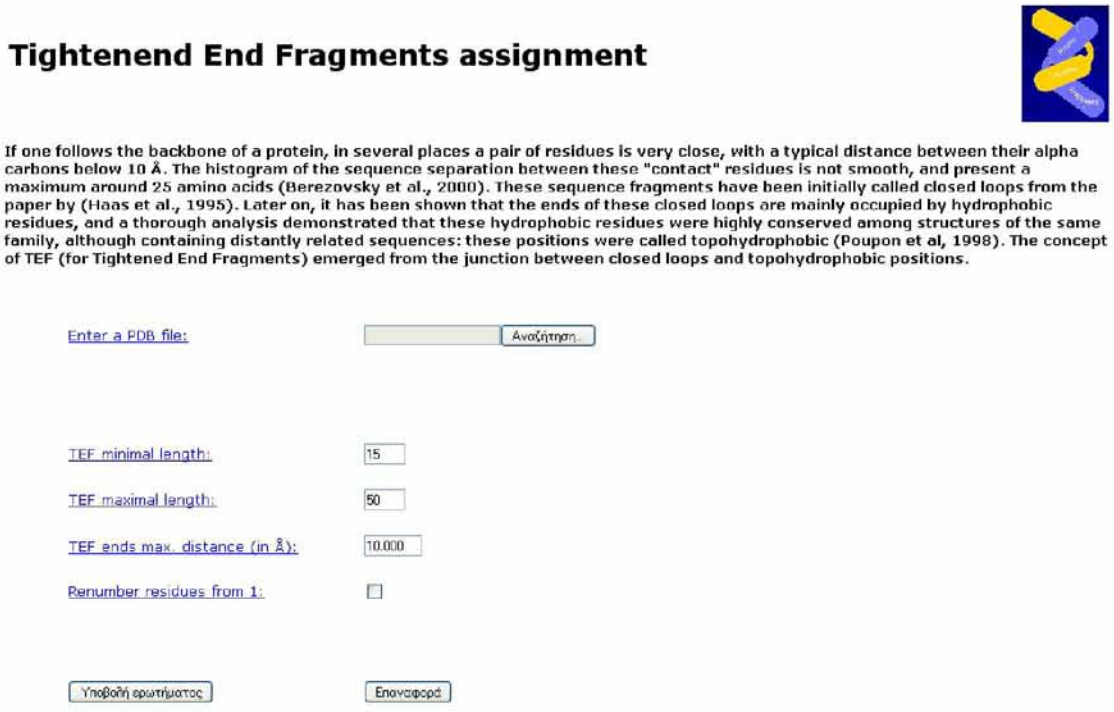

ΑΡΧΗ
I = 1, N, (1)
B
Παίρνω επόμενη μη απορριπτέα περιοχή MCF έστω [ea, eb]
J = I+1, N, (1)
Παίρνω επόμενη μη απορριπτέα περιοχή MCF έστω [ek, el]
ea<ek<eb || ea<el<eb || ek<ea<el
ΟΧΙ
ΝΑΙ
o1 = |ea – ek|
o2 = |eb – el|
ΟΧΙ
o1 ≤ 5 || o2 ≤ 5
ΝΑΙ
L1 = eb - ea
L2 = el - ek
C


C
ΟΧΙ
$L_1 > L_2$
ΝΑΙ
$[e_a, e_b]$ απορρίπτεται
$[e_l, e_k]$ απορρίπτεται
B
$[e_a, e_b]$ τελική MCF εφόσον δεν έχει απορριφθεί
ΤΕΛΟΣ


ΑΡΧΗ
I = 1, N, (1)
E
Παίρνω επόμενη περιοχή MCF έστω [ea, eb]
J = 1, N, (1)
Παίρνω επόμενη περιοχή MCF έστω [ek, el] με k ≠ a και l ≠ b
ek<ea<el || ek<eb<el
ΟΧΙ
ΝΑΙ
[ea, eb] μπαίνει στο σύνολο check
E
ΤΕΛΟΣ

2.6 Tightened End Fragments (TEF)

Σκοπός της εργασίας είναι η ανάπτυξη μιας τεχνικής η οποία θα προσπαθεί να προβλέπει με όσο το δυνατόν μεγαλύτερη ακρίβεια την τριτοταγή δομή των σφαιρικών πρωτεϊνών. Συγκεκριμένα, προσπαθούμε να προβλέψουμε τις περιοχές TEF μέσω των MCF. Προκειμένου να παράγουμε τα αποτελέσματα για την ακριβή θέση των TEF περιοχών των πρωτεϊνών μας, χρησιμοποιήσαμε τον server που βρίσκεται στην εξής διεύθυνση : http://bioserv.rpbs.jussieu.fr/TEF/ (σχήμα 2.14).



Σχήμα 2.14 *Η αρχική σελίδα του ιστότοπου που χρησιμοποιήθηκε για την παραγωγή των δεδομένων που αφορούν την εύρεση των TEF περιοχών των πρωτεϊνών. Στο σχήμα επίσης φαίνονται και οι παράμετροι που χρησιμοποιήσαμε.*

Η διαδικασία που ακολουθήθηκε ήταν η παρακάτω.

Για κάθε μία από τις πρωτεΐνες χρησιμοποιήσαμε τα αντίστοιχα pdb αρχεία από την βάση δεδομένων Protein Data Bank σαν αρχεία εισόδου μέσω της επιλογής «Αναζήτηση» που υπάρχει. Μάλιστα αυτού του είδους τα αρχεία είναι και τα μοναδικά που δέχεται σαν είσοδο το πρόγραμμα. Τονίζεται πως κάποιες ασυνήθιστες περιπτώσεις ενδεχομένως να μην μπορεί να τις επεξεργαστεί σωστά το πρόγραμμα και έτσι είναι σημαντικό να διαθέτει ο χρήστης το πιο ορθό pdb αρχείο, δηλαδή αυτό να μην περιέχει κενά και να διαθέτει καλά ευρετήρια. Παρόλα αυτά αν χρησιμοποιήσουμε την επιλογή «Renumber residues from 1» υπάρχει η πιθανότητα να ξεπεραστούν τα όποια προβλήματα και η επεξεργασία των δεδομένων να γίνει κανονικά.

Αφού λοιπόν εισάγουμε στο πρόγραμμα το αντίστοιχο pdb αρχείο για κάθε πρωτεΐνη, θα πρέπει να θέσουμε τις παραμέτρους του προγράμματος ως εξής. Με την παράμετρο «TEF minimal length» ορίζουμε το ελάχιστο μήκος του TEF προκειμένου να αποφύγουμε την παραγωγή υπερβολικά μικρών τμημάτων. Την παράμετρο αυτή την θέσαμε στα 15 αμινοξέα που είναι και η προεπιλεγμένη τιμή. Αντίστοιχα, με την παράμετρο «TEF maximal length» ορίζουμε το μέγιστο μήκος των TEF. Έχει δειχθεί πως το μήκος των TEF κινείται συνήθως στο εύρος 22-32 αμινοξέα [14,15,21]. Κατά αυτό τον τρόπο το μέγιστο μήκος τους δε μπορεί να υπερβαίνει την διπλάσια τιμή, επομένως η προεπιλεγμένη τιμή είναι 50, τιμή που επίσης χρησιμοποιήσαμε και εμείς. Η παράμετρος «TEF ends max distance» δείχνει την μέγιστη απόσταση που υπάρχει ανάμεσα στα άκρα TEF και η οποία μπορεί να τεθεί σε διάφορες τιμές. Εμείς την θέσαμε ξανά στην προεπιλεγμένη τιμή που δεν είναι άλλη από τα 10 Å. Τέλος, τσεκάραμε και την επιλογή «Renumber residues from 1» για τον λόγο που εξηγήσαμε παραπάνω. Αξίζει να αναφέρουμε πως είναι πολύ σύνηθες στα αρχεία pdb, η αρίθμηση των αμινοξέων να μην ξεκινά από το 1 [40].

Αφού θέσαμε τις παραμέτρους στις προαναφερθείσες τιμές, πατώντας το κουμπί «Υποβολή Ερωτήματος», ο υπολογισμός γίνεται άμεσα. Τα αποτελέσματα δίνονται στο κάτω μέρος της σελίδας τόσο σε αριθμητική τιμή όσο και γραφικά, ενώ στην αρχή της σελίδας των αποτελεσμάτων εμφανίζεται η λίστα με τα TEF με την εξής μορφή. Η πρώτη στήλη δείχνει το πρώτο κατάλοιπο του TEF, η δεύτερη στήλη το τελευταίο κατάλοιπο, η τρίτη στήλη την απόσταση σε Å ανάμεσα στα πρώτα και τα τελευταία κατάλοιπα του TEF και η τελευταία στήλη το μέγεθος του σε πλήθος αμινοξέων. Για τη καταγραφή των αποτελεσμάτων χρησιμοποιήθηκε το αριθμητικό αποτέλεσμα και όχι το γράφημα και αυτό διότι υπήρχε ένα μικρό bug που δεν επέτρεπε την εξαγωγή συμπερασμάτων από τη γραφική αναπαράσταση.

Τέλος, να τονίσουμε πως στις περιπτώσεις πρωτεϊνών που έχουν περισσότερες της μία αλυσίδες καταγράψαμε τα TEF για την συγκεκριμένη αλυσίδα που μας ενδιαφέρει και μόνο. Το αν σε μια πρωτεΐνη μας ενδιαφέρει μια συγκεκριμένη αλυσίδα, θα επισημαίνεται στον pdb κωδικό της πρωτεΐνης, δηλαδή αν μια πρωτεΐνη περιλαμβάνει πάνω από 4 γράμματα/ψηφία στον κωδικό της, τότε το τελευταίο γράμμα/ψηφίο δηλώνει την αλυσίδα που θέλουμε.

Με τον τρόπο που περιγράψαμε παραπάνω αντλήσαμε την πληροφορία σχετικά με τις TEF περιοχές των σφαιρικών πρωτεϊνών πάνω στις οποίες έγινε η έρευνα, περιοχές τις οποίες επιδιώκουμε να προβλέψουμε μέσω των MCF με την τεχνική που παρουσιάζεται στην παρούσα εργασία.

2.7 Στατιστικά μέτρα αξιολόγησης των προβλέψεων.

2.7.1 Sensitivity, Specificity

Η ευαισθησία (sensitivity) και η εξειδίκευση (specificity) είναι στατιστικά μέτρα αξιολόγησης προβλέψεων σε προβλήματα δυαδικής κατηγοριοποίησης. Στην εργασία μας οι κατηγορίες είναι δύο και συγκεκριμένα TEF και όχι – TEF. Στα προβλήματα αυτά, τα δεδομένα χωρίζονται σε 4 κατηγορίες. True positive όπου εκεί ανήκουν τα δεδομένα όπου ο αλγόριθμος προέβλεψε ότι είναι TEF και στην πραγματικότητα είναι όντως TEF, true negative όπου εκεί ανήκουν τα δεδομένα όπου ο αλγόριθμος προέβλεψε ότι δεν είναι TEF και στην πραγματικότητα όντως δεν είναι TEF, false positive όπου περιλαμβάνονται τα δεδομένα τα οποία δεν είναι TEF αλλά ο αλγόριθμος προέβλεψε ότι είναι TEF και τέλος false negative με τα TEF τα οποία ο αλγόριθμος πρόβλεψης τα κατηγοριοποίησε ως μη TEF.

Η ευαισθησία μετρά το ποσοστό των positives που έχουν προβλεφθεί σωστά και στην περίπτωση μας συγκεκριμένα, το ποσοστό των γνωστών TEF που προβλέπονται σωστά από τον αλγόριθμο προσομοίωσης. Ο μαθηματικός τύπος για την ευαισθησία είναι ο εξής [41]:

$$\mathrm{Sn} = \frac{TruePositive}{AllTrue} = \frac{TruePositive}{TruePositive + FalseNegative}$$

Αν έχουμε λοιπόν ευαισθησία ίση με 100% αυτό σημαίνει πως όλα τα TEF προβλέπονται σωστά. Η ευαισθησία σαν μέτρο από μόνη της δε μας υποδεικνύει πόσο καλά προβλέπονται οι άλλες κατηγορίες. Προκειμένου να το μάθουμε αυτό για τις δυαδικές κατηγοριοποιήσεις πρέπει να μετρήσουμε την ευαισθησία και για την άλλη κατηγορία.

Από την άλλη η εξειδίκευση μετρά το ποσοστό των θετικών προβλέψεων που προβλέπονται σωστά ήτοι στην δικιά μας περίπτωση το ποσοστό των προβλεπόμενων MCF που αντιστοιχούν σε TEF στην πραγματικότητα και ο μαθηματικός της τύπος είναι ο παρακάτω [41]:

$$\mathrm{Sp} = \frac{TruePositive}{AllPositive} = \frac{TruePositive}{TruePositive + FalsePositive}$$

Αν έχουμε λοιπόν εξειδίκευση ίση με 100% αυτό σημαίνει πως όλα τα MCF αντιστοιχούν όντως σε TEF. Πολλές φορές η εξειδίκευση συγχέεται με την ακρίβεια (precision ή την positive predictive value PPV) καθώς όλες αυτές οι έννοιες αναφέρονται στο ποσοστό των positives που είναι true positives. Η διάκριση αυτών των δύο είναι σημαντική όταν οι κατηγορίες είναι πολύ διαφορετικού μεγέθους. Για παράδειγμα, ένας κατηγοριοποιητής υψηλής εξειδίκευσης μπορεί να έχει πολύ χαμηλή ακρίβεια αν το πλήθος των true negatives είναι πολύ μεγαλύτερο από το πλήθος των true positives και αντίστροφα [41].

Στην παρούσα εργασία η ευαισθησία και η εξειδίκευση μετρήθηκαν τόσο για τα TEF limits όσο και για τα TEF σαν τμήματα ολόκληρα. Να σημειωθεί πως στην πρώτη περίπτωση χρησιμοποιήθηκε «παράθυρο» ± 5 θέσεων. Αυτό σημαίνει πως σαν ευαισθησία ορίσαμε το πλήθος των TEF limits που βρίσκονται σε απόσταση μικρότερη κατ απόλυτη τιμή των 5 θέσεων από MCF limits προς το σύνολο των TEF limits και σαν εξειδίκευση το πλήθος των MCF limits που βρίσκονται σε απόσταση μικρότερη κατ απόλυτη τιμή των 5 θέσεων από TEF limits προς το σύνολο των MCF limits. Ο λόγος που έγινε αυτό αναφέρεται στη αρχή της ενότητας 3.1

2.7.2 Accuracy

Η ακρίβεια (accuracy) χρησιμοποιείται ως στατιστικό μέτρο για να δείξει πόσο καλά ένας κατηγοριοποιητής δυαδικής κατηγοριοποίησης αναγνωρίζει ή αποκλείει μια συγκεκριμένη συνθήκη. Η ακρίβεια είναι ένα μέτρο της αναλογίας των σωστών αποτελεσμάτων (true positives και true negatives) στον πληθυσμό του δείγματος. Ο μαθηματικός τύπος για την ακρίβεια είναι ο εξής [42]:

$$\text{Accuracy} = \frac{TruePositive + TrueNegative}{TruePositive + TrueNegative + FalsePositive + FalseNegative}$$

Αν έχουμε λοιπόν ακρίβεια ίση με 100% αυτό σημαίνει πως όλες οι TEF και μη-TEF περιοχές προβλέπονται σωστά από τον αλγόριθμο μας. Φυσικά το μέτρο αυτό όπως είναι λογικό επηρεάζεται σε μεγάλο βαθμό από τα μεγέθη των κατηγοριών. Στην παρούσα εργασία η ακρίβεια μετρήθηκε τόσο για τα TEF limits όσο και για τα TEF σαν τμήματα ολόκληρα, χωρίς να χρησιμοποιηθεί «παράθυρο» ± 5 θέσεων για την πρώτη περίπτωση όπως έγινε με την ευαισθησία και την εξειδίκευση.

2.7.3 Fractional Segment Overlap (SOV)

Ένα από τα στατιστικά μέτρα που χρησιμοποιήσαμε για να αξιολογήσουμε την ποιότητα της πρόβλεψης ήταν και το Sov [43] και συγκεκριμένα ένας τροποποιημένος ορισμός του Sov94 [44], μέτρο το οποίο είναι segment – based δηλαδή συγκρίνει ολόκληρα τμήματα και όχι τόσο θέσεις αμινοξέων. Ορίστηκε για πρώτη φορά το 1994 κυρίως για εκτίμηση της πρόβλεψης της δευτεροταγούς δομής των πρωτεϊνών, αλλά εμείς χρησιμοποιήσαμε τον βελτιωμένο ορισμό του που πρωτοπαρουσιάστηκε το 1999 και ο οποίος βρίσκει εφαρμογή και για την πρόβλεψη της τριτοταγούς δομής των πρωτεϊνών. Παρακάτω θα αναφέρουμε λίγες πληροφορίες τόσο για το αρχικό μέτρο Sov94 όσο και για το τροποποιημένο Sov, μέτρο που χρησιμοποιήθηκε στην παρούσα εργασία.

Παρότι η δευτεροταγής δομή των πρωτεϊνών είναι καταμερισμένη στη φύση, στην πράξη έχει προβλεφθεί και αναλυθεί σε μία «ανά - κατάλοιπο» βάση. Συχνά όμως έτσι δεν μπορούμε να εκμεταλλευθούμε σε βάθος την «χρησιμότητα» των προβλέψεων της δευτεροταγούς δομής [43].

Ένα άλλο θέμα που επηρεάζει την εκτίμηση της πρόβλεψης της δομής σχετίζεται με την ποικιλία που παρατηρείται στα άκρα των τμημάτων της δευτεροταγούς δομής. Ακόμα και για ζευγάρια ομόλογων πρωτεϊνών με παρόμοιες ακολουθίες αμινοξέων, τα στοιχεία της δευτεροταγούς δομής συχνά διαφέρουν ως προς την ακριβή θέση των άκρων τους. Για αυτό το λόγο ενδέχεται να μην είναι τόσο κρίσιμο να προβλέψουμε επακριβώς τα άκρα των τμημάτων. Εφόσον η συνολική τριτοταγής δομή παρέχει τέτοια περιορισμένη ποικιλία, είναι λογικό να «επιτρέψουμε» το ίδιο και στο επίπεδο της εκτίμησης της δευτεροταγούς δομής.

Για όλους αυτούς τους λόγους, ένα πλήθος συγγραφέων τόνισε την ανάγκη να δημιουργηθεί ένα μέτρο της εκτίμησης της πρόβλεψης της δευτεροταγούς δομής που θα είχε μεγαλύτερο «δομικό» νόημα. Αυτό το «δομικά» προερχόμενο μέτρο θα πρέπει να αιτιολογεί και να λαμβάνει υπόψη τα ακόλουθα [43].

- Τον τύπο και την θέση των τμημάτων της δευτεροταγούς δομής παρά μια ανά κατάλοιπο ανάθεση της τελικής στερεοδομής.
- Την φυσική διακύμανση των ορίων των τμημάτων ανάμεσα σε οικογένειες ομόλογων πρωτεϊνών.
- Ενδεχόμενη αμφιβολία στη θέση των άκρων των τμημάτων εξαιτίας των διαφορών στη προσέγγιση της κατηγοριοποίησης της δευτεροταγούς δομής.

Το μέτρο αυτό είναι το Segment Overlap Sov. Αυτό το μέτρο καταφέρνει να τονίζει τη σημασία δομικά σημαντικών χαρακτηριστικών (τμήματα δευτεροταγούς ή και τριτοταγούς δομής) και να ελαττώνει τη σημασία αυτών που είναι δομικά λιγότερο σημαντικά (μικρές διαφοροποιήσεις στη θέση και το μήκος των τμημάτων). Έτσι το Sov έχει επιλεχθεί ως ένα από τα κριτήρια αξιολόγησης της πρόβλεψης για το CASP2 (Critical Assessment of Techniques for Protein Structure Prediction). Παρακάτω αναφέρουμε λίγα λόγια για το αρχικό μέτρο Sov94 και στη συνέχεια θα παρουσιάσουμε πληροφορίες για το νέο βελτιωμένο μέτρο Sov που χρησιμοποιήθηκε στην εργασία.

Το Sov94 συμβιβάζει την ανεκτικότητα σε σφάλματα και την ακρίβεια. Βασίζεται στις εξής ιδέες [44]:

1. Επιτρέπει κάποια διακύμανση στα άκρα των τμημάτων.
2. Παρέχει μία κλιμακούμενη βαθμίδα επικάλυψης τμημάτων η οποία μας παρέχει διαισθητικά αναμενόμενες τιμές για ακραίες περιπτώσεις.
3. Δίνει χαμηλές τιμές για τυχαίες εκτιμήσεις .

Το μέτρο αυτό απλά μετράει το κλασματικό εύρος για το οποίο δύο τμήματα επικαλύπτονται, με κάποια ανοχή για τα κατάλοιπα των άκρων που δεν ταιριάζουν. Υπάρχει μόνο μία ρυθμιζόμενη παράμετρος, το μέγεθος της ανοχής.

Συγκεκριμένα, δεδομένων δύο συνόλων συμβόλων δευτεροταγούς δομής ορίζουμε ως Sov :

$$\text{Sov} = \frac{1}{N} \sum_{s} \frac{\min ov(s1 : s2) + \delta}{\max ov(s1 : s2)} \times \text{len(s1)}$$

όπου N είναι το συνολικό πλήθος των καταλοίπων. Το σύνολο s1 συνήθως περιέχει τα τμήματα που θέλουμε να συγκρίνουμε (observed) και το s2 τα τμήματα που προβλέπονται (predicted) .Το άθροισμα λαμβάνει υπόψη όλα τα ζευγάρια τμημάτων s = {s1,s2} όπου s1 και s2 είναι δύο τμήματα που εμφανίζουν επικάλυψη για τουλάχιστον μία θέση καταλοίπου για την ίδια δομή. Η ασυμμετρία ανάμεσα στα δύο τμήματα εισάγεται μέσω του βάρους len(s1) το οποίο είναι το μήκος του s1 και όπως είπαμε είναι συνήθως το τμήμα της πειραματικά λυμένης δομής. Την πραγματική επικάλυψη ανάμεσα στα δύο τμήματα δείχνει το minov ενώ το maxov δείχνει το συνολικό εύρος των δύο τμημάτων. Η επιτρεπόμενη διακύμανση δ εξασφαλίζει ένα ποσοστό της τάξης του 1 μόνο όταν υπάρχουν μικρές αποκλίσεις στα άκρα των τμημάτων όπως συχνά παρατηρείται σε δομικά ομόλογες πρωτεΐνες. Η τιμή δ επιλέγεται κατά τέτοιο τρόπο ώστε να είναι μικρότερη από το minov και μικρότερη από το μισό του μήκους του τμήματος s1 (δ = 1,2,3 για μικρά, μεσαία, μεγάλα τμήματα). Η αναλογία minov/maxov περιορίζεται σε μια μέγιστη τιμή της τάξης του 1 που σημαίνει ότι αυτός ο περιορισμός δε μπορεί να οδηγήσει σε μία ποσοστιαία τιμή επικάλυψης που να ξεπερνά το τέλειο [44].

Όσον αφορά την ποιότητα αυτού του μέτρου πρέπει να τονίσουμε ότι σε συγκριτικά τεστ που έγιναν αποδείχθηκε ότι το Sov94 δίνει πολύ περισσότερες πληροφορίες σχετικά με την τρισδιάστατη δομή απ ότι τα απλά στατιστικά μέτρα που δε λαμβάνουν υπόψη ολόκληρα τμήματα παρά μόνο ξεχωριστά κατάλοιπα όπως για παράδειγμα το μέτρο Q3.

Παρόλα αυτά δημιουργήθηκε η ανάγκη να βελτιωθεί ο αρχικός ορισμός. Το μέτρο αυτό αρχικά δεν είχε ένα σαφώς ορισμένο άνω όριο, επομένως δεν ήταν δυνατή η άμεση σύγκριση με άλλα στατιστικά μέτρα αξιολόγησης της πρόβλεψης [44]. Παρακάτω θα παρουσιάσουμε τις αλλαγές που έγιναν οι οποίες βελτίωσαν την αρχική αδυναμία [43].

Όπως και πριν, έτσι και τώρα ως s1 ορίζουμε το σύνολο των τμημάτων που προκύπτουν από τις παρατηρούμενες γνωστές δομές και ως s2 το σύνολο των τμημάτων που προβλέπονται. Το (s1,s2) αντιπροσωπεύει ένα ζευγάρι επικαλυπτόμενων τμημάτων, S(i) το σύνολο όλων των επικαλυπτόμενων ζευγών των τμημάτων (s1, s2) που βρίσκονται στην κατάσταση i δηλαδή

$S(i) = \{(s1, s2): s1 \cap s2 \neq \varnothing, \text{ s1 και s2 ανήκουν στην ίδια κατάσταση i}\}$

$S'(i)$ – το σύνολο των τμημάτων s1 τα οποία δεν επικαλύπτονται με κανένα τμήμα από το σύνολο s2 για την ίδια κατάσταση i δηλαδή

S'(i) = {s1 : $\forall$ s2 , s1 $\cap$ s2 = $\varnothing$, s1 και s2 ανήκουν στην ίδια κατάσταση i}

Για δεδομένη κατάσταση i το Sov(i) ορίζεται ως

$$\text{Sov(i)} = 100 \times \frac{1}{N(i)} \sum_{S(i)} \left[\frac{\min ov(s1,s2) + \delta(s1,s2)}{\max ov(s1,s2)} \times \text{len(s1)} \right] \quad (1)$$

με την τιμή κανονικοποίησης N(i) να ορίζεται ως

$$\text{N(i)} = \sum_{S(i)} len(s1) + \sum_{S'(i)} len(s1) \quad (2)$$

Το άθροισμα στην εξίσωση 1 και το πρώτο άθροισμα στην εξίσωση 2 λαμβάνουν υπόψη όλα τα ζεύγη τμημάτων που βρίσκονται στην κατάσταση i και που υπάρχει επικάλυψη μεταξύ τους κατά τουλάχιστον ένα κατάλοιπο, το δεύτερο άθροισμα στην εξίσωση 2 λαμβάνει υπόψη όλα τα υπόλοιπα τμήματα της κατάστασης i, len(s1) είναι το πλήθος των καταλοίπων του συνόλου s1, το minov(s1,s2) είναι το μήκος της πραγματικής επικάλυψης ανάμεσα στα s1 και s2 δηλαδή όπου και τα δύο τμήματα έχουν κατάλοιπα που επικαλύπτονται στην κατάσταση i, maxov(s1,s2) είναι το συνολικό εύρος για το οποίο είτε το σύνολο s1 είτε το σύνολο s2 έχουν ένα κατάλοιπο στην κατάσταση i, και το δ(s1,s2) ορίζεται ως

$$\delta(s1,s2) = \min\{(\text{maxov}(s1,s2) - \text{minov}(s1,s2)) ; \text{minov}(s1,s2) ; \text{int}(\text{len}(s1)/2) ; \text{int}(\text{len}(s2)/2)\} \quad (3)$$

όπου το min{x1;x2;x3;…;xn} είναι το ελάχιστο των n ακεραίων.

Το μέτρο αυτό έτσι όπως ορίστηκε στις εξισώσεις 1 – 3 μπορεί να επεκταθεί ώστε να αξιολογεί πολλαπλές καταστάσεις δευτεροταγούς και τριτοταγούς δομής. Για την δευτεροταγή δομή συγκεκριμένα για τις τρεις καταστάσεις έλικας (H), πτυχωτή επιφάνεια (E) και coil (C) ορίζουμε

$$\text{Sov} = 100 \times \left[\frac{1}{N} \sum_{i \in \{H,C,E\}} \sum_{S(i)} \frac{\min ov(s1,s2) + \delta(s1,s2)}{\max ov(s1,s2)} \times len(s1) \right] \quad (4)$$

όπου η τιμή κανονικοποίησης N ορίζεται ως το άθροισμα των N(i) και για τις τρεις καταστάσεις

$$\text{N} = \sum_{i \in (H,C,E)} N(i) \quad (5)$$

Την ποιότητα του ταιριάσματος για κάθε ζεύγος τμημάτων την παίρνουμε σαν ένα κλάσμα της επικάλυψης των δύο τμημάτων minov(s1,s2) με το συνολικό εύρος του ζεύγους maxov(s1,s2). Ο ορισμός αυτός επιτρέπει την βελτίωση του κλάσματος αυτού με την επέκταση της επικάλυψης κατά την τιμή δ(s1,s2). Η διαδικασία κανονικοποίησης εξασφαλίζει ότι οι τιμές Sov κινούνται στο εύρος 0 – 100 επομένως μπορούν να χρησιμοποιηθούν σε ποσοστιαία κλίμακα ούτος ώστε να είναι εφικτή η άμεση σύγκριση με άλλα στατιστικά μέτρα αξιολόγησης των προβλέψεων, όπως για παράδειγμα το Q3.

Σε σύγκριση με το Sov94, ο νέος ορισμός του μέτρου εισήγαγε δύο νέες αλλαγές όσον αφορά δύο βασικά θέματα: τη διαδικασία κανονικοποίησης και τον ορισμό του δ, δηλαδή τον βαθμό της ελευθερίας που επιτρέπεται στα άκρα των τμημάτων. Ο ορισμός του δ στην εξίσωση 3 γίνεται συμμετρικός όσον αφορά τα παρατηρούμενα και τα προβλεπόμενα τμήματα και ο δε παράγοντας κανονικοποίησης N που ήταν ίσος με το συνολικό πλήθος των καταλοίπων πλέον αντικαθίσταται από τις εξισώσεις 2 και 5.

Η νέα τιμή κανονικοποίησης υπολογίζεται όσον αφορά τα τμήματα της παρατηρούμενης δομής, με το κάθε τμήμα να λαμβάνεται υπόψη τουλάχιστον μία φορά. Αν για κάθε τμήμα που παρατηρούμε πρέπει να λάβουμε υπόψη πάνω από ένα προβλεπόμενο τμήμα, το άθροισμα επεκτείνεται κατάλληλα, δηλαδή το παρατηρούμενο τμήμα αθροίζεται για κάθε επικαλύπτον ζεύγος. Συνεπώς η διαδικασία κανονικοποίησης «χαμηλώνει» το σκορ της πρόβλεψης για τα λάθος προβλεπόμενα τμήματα και αντανακλά την ζευγαρωτή φύση της σύγκρισης των τμημάτων αυτών. Ως εκ τούτου το εύρος τιμών για το Sov κυμαίνεται πλέον από 0 – 100 % ανάλογα με τη ακρίβεια της πρόβλεψης.

Στον ορισμό του Sov94 το δ είχε σχεδιαστεί ώστε να επιτρέπει μια περιορισμένη διακύμανση στα άκρα των τμημάτων της δευτεροταγούς δομής ώστε να δίνεται έμφαση στη σωστή πρόβλεψη των τμημάτων παρά στη ανάθεση μιας συγκεκριμένης κατάστασης σε μεμονωμένα κατάλοιπα. Στον νέο ορισμό η επιτρεπόμενη αυτή διακύμανση περιορίζεται ακόμα περισσότερο, τόσο ώστε για ένα ζευγάρι παρατηρούμενων και προβλεπόμενων τμημάτων το δ δε μπορεί να υπερβεί σαν τιμή το μισό του μήκους του μικρότερου τμήματος. Πρακτικά αυτό σημαίνει πώς αν τα τμήματα έχουν μια αρκετά διευρυμένη επικάλυψη με μόνο μικρές διαφοροποιήσεις στα άκρα τους, θεωρούνται πανομοιότυπα., δηλαδή η πρόβλεψη ήταν τέλεια. Από την άλλη πλευρά αν η επικάλυψη ήταν μικρή, τότε αυτή δε μπορεί να «διευρυνθεί» κατά ένα αξιόλογο ποσό ώστε να παράγει ένα τεχνητά βελτιωμένο σκορ. Παρότι ακόμα και αν δεν λάβουμε υπόψη την τιμή δ και τη θεωρήσουμε ως μηδενική μπορούμε ακόμα και τότε να βγάλουμε ασφαλή συμπεράσματα κατά πόσο η πρόβλεψη μας ήταν καλή ή κακή, θα έχουμε όμως τότε παντελώς αγνοήσει τη φυσική διακύμανση στα άκρα των τμημάτων μας [43].

Στην παρούσα εργασία θεωρήσαμε μόνο μία κατάσταση, τα TEF, άρα δεν ασχοληθήκαμε καθόλου με τις εξισώσεις 4 και 5. Επίσης, αφού ουσιαστικά έχουμε

μία μόνο κατάσταση, το i στις εξισώσεις 1 και 2 είναι πλεονασμός. Το σύνολο s1 αποτελούσαν οι παρατηρούμενες περιοχές δηλαδή τα TEF και το σύνολο s2 οι προβλεπόμενες δηλαδή τα MCF.

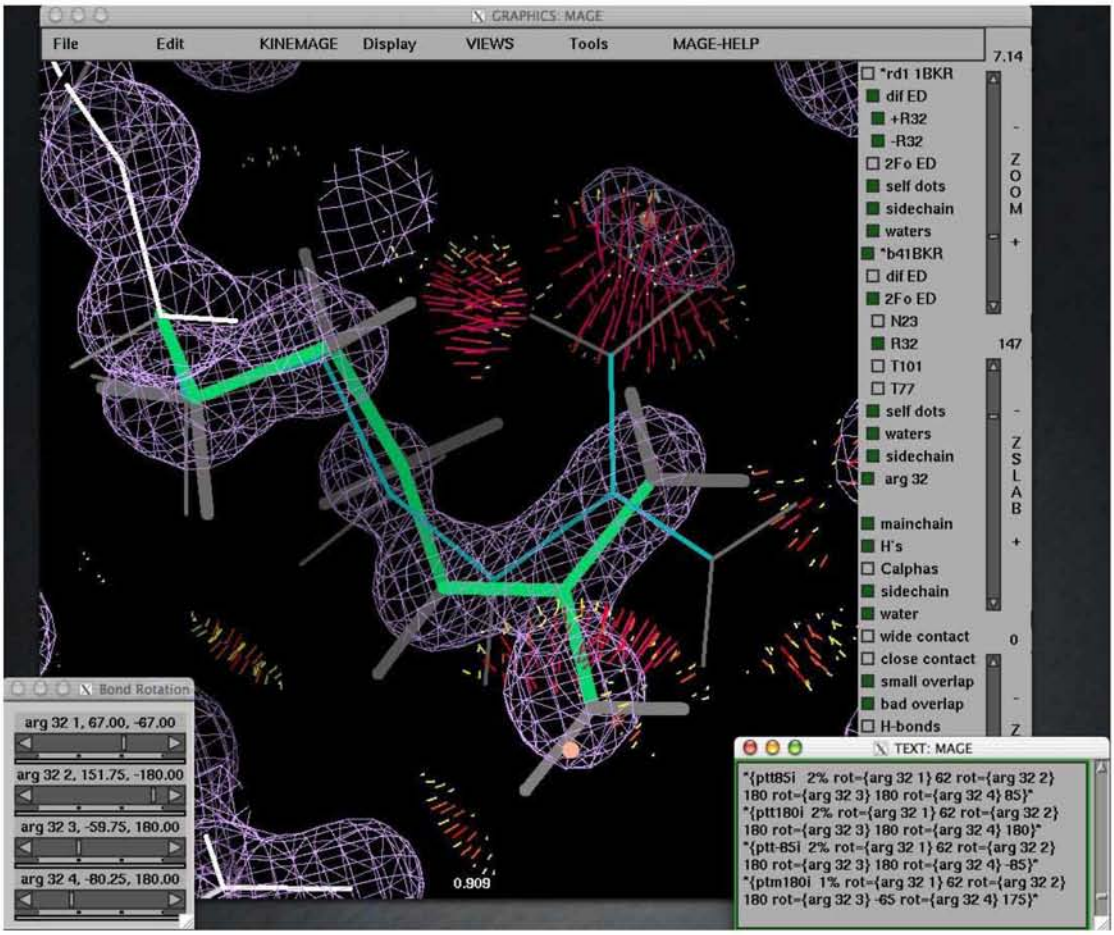
Μέχρι στιγμής, το Sov έχει οριστεί ώστε να εκτιμά την ορθότητα της πρόβλεψης της δομής όσον αφορά την πειραματικά λυμένη δομή ($Sov^{observed}$). Εκτός από αυτό το βασικό μέτρο είναι δυνατό να υπολογίσουμε και μια εναλλακτική εκδοχή ($Sov^{predicted}$) η οποία παρέχει μια τιμή που δείχνει τι ποσοστό των προβλεπόμενων τμημάτων είναι σωστό, τιμή την οποία χρησιμοποιήσαμε και στην εργασία μας για να έχουμε μια καλύτερη εικόνα για τις προβλέψεις. Αυτό το μέτρο υπολογίζεται με το s1 να αντιπροσωπεύει τα προβλεπόμενα τμήματα και στην περίπτωση μας τα MCF και το s2 τα παρατηρούμενα δηλαδή τα TEF. Το $Sov^{predicted}$ είναι ιδιαίτερα χρήσιμο σε μεθόδους ανάπτυξης, για παράδειγμα σε περιπτώσεις όπου η πρόβλεψη για μια συγκεκριμένη κατάσταση υποψιαζόμαστε ότι δεν ανταποκρίνεται στην πραγματικότητα. Αν εξαιρέσουμε την περίπτωση όπου η δομή που προβλέπεται είναι ακριβώς η ίδια με τη δομή που παρατηρείται, τα δυο αυτά μέτρα δεν παράγουν το ίδιο αποτέλεσμα απαραίτητα [43]. Αξίζει να παρατηρήσουμε πως τόσο για ολόκληρα τμήματα όσο και για μεμονωμένα κατάλοιπα, η καλύτερη αξιολόγηση της πρόβλεψης γίνεται όταν τα δύο αυτά μέτρα, δηλαδή το $Sov^{observed}$ και το $Sov^{predicted}$ συνδυαστούν, όπως έγινε δηλαδή και στην παρούσα εργασία..

Το Sov όχι μόνο «τιμωρεί» τις λάθος προβλέψεις αλλά επίσης εκτελεί μια αποτελεσματική αξιολόγηση. Η πρόβλεψη της τριτοταγούς δομής ανάγεται σε ένα σημαντικό θέμα, και η σωστή εκτίμηση της πρόβλεψης είναι κρίσιμης σημασίας για την εξαγωγή συμπερασμάτων [43]. Από αυτή την οπτική σίγουρα το μέτρο Sov παρέχει μια πολύ ποιοτικότερη εκτίμηση για την πρόβλεψη μας απ ότι τα στατιστικά μέτρα που βασίζονται στην ανάθεση ανά κατάλοιπο και γι αυτό το λόγο χρησιμοποιήθηκε και στην εργασία μας.

2.8 Οπτικοποίηση αποτελεσμάτων

Για την οπτικοποίηση των αποτελεσμάτων μας χρησιμοποιήσαμε το πρόγραμμα KiNG (Kinemage, Next Generation). Πρόκειται για ένα διαδραστικό σύστημα τρισδιάστατων γραφικών που υποστηρίζει ένα πλήθος γραφικών δομών ώστε να μπορούν να σχεδιαστούν διάφοροι τύποι γράφων, γραφικών παραστάσεων και απεικονίσεων, παρότι αρχικά είχε χρησιμοποιηθεί για την αναπαράσταση μακρομοριακών δομών για βιοφυσική έρευνα. Βασίζεται στην ιδέα του Mage, Javamage και Kinemage (kinetic image ήτοι κινητική εικόνα) και πρόκειται για μια εφαρμογή που διαθέτει όλα τα χαρακτηριστικά της Java και επιπλέον παρέχει μια εύχρηστη διεπαφή. Επίσης, μπορεί να λειτουργήσει και ως Java applet ώστε να είναι εύκολη η πρόσβαση στις kinemages και από ένα πρόγραμμα περιηγητή [45].

Η βασική ιδιότητα του λογισμικού αυτού αλλά και το μεγάλο του πλεονέκτημα είναι ότι μπορεί να δημιουργήσει κινητικές εικόνες, τις λεγόμενες kinemages (σχήμα 2.15).



Σχήμα 2.15 *Παράδειγμα κινητικής εικόνας (kinemage).*

Η kinemage είναι μια διαδραστική τρισδιάστατη απεικόνιση και διαθέτει απλά γεωμετρικά σχήματα. Με αυτό τον τρόπο μπορεί να απεικονίσει μεγάλη ποικιλία αντικειμένων από απλά σκίτσα μέχρι περίπλοκα γραφήματα και λεπτομερειακές αναπαραστάσεις τρισδιάστατων αντικειμένων. Το βασικό της προτέρημα είναι ότι ο

χρήστης μπορεί να επιλέξει ακριβώς ποια πληροφορία θα παρουσιαστεί στον παρατηρητή και με ποιόν τρόπο, επιλέγοντας τον τρόπο παρουσίασης. Μπορεί δηλαδή να εκτελέσει διάφορες λειτουργίες στην kinemage οι οποίες έχουν άμεσο αποτέλεσμα, μπορεί για παράδειγμα να περιστρέψει την εικόνα σε πραγματικό χρόνο, να ενεργοποιήσει / απενεργοποιήσει τμήματα της απεικόνισης, να αναγνωρίσει συγκεκριμένα σημεία της εικόνας επιλέγοντας τα καθώς και να απεικονίσει με τη μορφή animation τις αλλαγές ανάμεσα σε διάφορες μορφές εικόνων. Οι kinemages βρίσκουν μεγαλύτερη εφαρμογή όταν θέλουμε να οπτικοποιήσουμε τρισδιάστατα δεδομένα, γι αυτό το λόγο χρησιμοποιήσαμε το εν λόγο πρόγραμμα [45,46].

Με το KiNG μπορούμε να δημιουργήσουμε, επεξεργαστούμε και να δούμε kinemages ενώ παρέχει ένα πλήθος από plug-ins και άλλα εργαλεία για την καλύτερη επεξεργασία και παρουσίαση τους. Συνοπτικά θα αναφέρουμε τις βασικές δυνατότητες που μας παρέχει το πρόγραμμα αυτό και στη συνέχεια θα περιγράψουμε ακριβώς τη διαδικασία που ακολουθήσαμε στην εργασία μας για την οπτικοποίηση των αποτελεσμάτων μας.

Η βασική αλληλεπίδραση με τις kinemages περιλαμβάνει την μετακίνηση τους προς οποιαδήποτε κατεύθυνση, την αναγνώριση συγκεκριμένων σημείων τους με ένα απλό κλικ, τη δημιουργία σημείων επισήμανσης, την τοποθέτηση όποιου σημείου επιθυμούμε ως το επίκεντρο της kinemage ώστε η υπόλοιπη γραφική περιοχή να περιστρέφεται γύρω από αυτό το σημείο, zoom – in και zoom – out ώστε να καθορίζουμε εμείς το επίπεδο λεπτομερειών που θα εμφανίζονται και την ενεργοποίηση / απενεργοποίηση συστατικών στοιχείων της kinemage [45]. Οι περισσότερες kinemages αποτελούνται από διάφορα δεδομένα τα οποία ομαδοποιούνται με λογικούς τρόπους. Μια πρωτεΐνη για παράδειγμα μπορεί να έχει ένα group για κάθε υπομονάδα, άλλο για βασική αλυσίδα, άλλο για πλευρική και άλλο για τα υδρογόνα. Στα group καθώς και στα subgroup αυτά δίνονται συγκεκριμένα ονόματα και μπορούν να γίνονται ορατά ή μη από τον χρήστη ανάλογα στο τι θέλει να επικεντρώσει την προσοχή του ή να παρουσιάσει. Επίσης, το πρόγραμμα παρέχει τη δυνατότητα να επιλέξουμε από ένα ευρύ φάσμα προεπιλεγμένων οπτικών γωνιών ώστε να έχουμε καλύτερη εικόνα για το τι δείχνει η kinemage. Στις πιο εξεζητημένες λειτουργίες του KiNG περιλαμβάνονται η εύρεση σημείων στην kinemage (καθώς κάθε σημείο διαθέτει μια αναγνωριστική ετικέτα), η παρακολούθηση animation και η επιλογή τρόπου παρουσίασης των γραφικών ανάμεσα από τις επιλογές perspective και stereo view. Τέλος, το πρόγραμμα αυτό μας επιτρέπει να προσαρμόσουμε με απλό τρόπο τα περισσότερα μέρη μιας kinemage. Η δομή της κινητικής εικόνας μπορεί να αναδιαρθρωθεί αποκόπτοντας, αντιγράφοντας και επικολλώντας ήδη υπάρχοντα στοιχεία, δημιουργώντας καινούρια και διαγράφοντας τα αχρείαστα καθώς και αλλάζοντας τη σειρά τους. Είναι δυνατή ακόμα και η μεταφορά στοιχείων από μια kinemage σε μια άλλη [45].

Με το KiNG μπορούμε να ανοίξουμε όχι μόνο μια kinemage αλλά να δουλέψουμε με περισσότερες και να συνδυάσουμε τα περιεχόμενα τους επισυνάπτοντας τα σε μια τελική kinemage την οποία μπορούμε είτε να αποθηκεύσουμε είτε να εκτυπώσουμε. Επίσης, δίνεται η δυνατότητα να εξάγουμε τις kinemages σε διάφορους τύπους αρχείων όπως jpeg, png, pdf, PostScript και άλλους [45].

Μερικές από τις λειτουργίες που παρέχουν τα εργαλεία και τα plug – in του KiNG περιλαμβάνουν επιπλέον εξειδικευμένες μετρήσεις όπως η γωνία που σχηματίζεται ανάμεσα σε τρία σημεία και η δίεδρη γωνία ανάμεσα σε τέσσερα, η εμφάνιση των XYZ συντεταγμένων του επιθυμητού σημείου, η περιστροφή της εικόνας γύρω από τον Z άξονα και η τροποποίηση των υπαρχόντων σημείων της kinemage μέσα από μια λίστα με διαθέσιμες ενέργειες μερικές από τις οποίες είναι η μετακίνηση σημείων, ο σχεδιασμός διακεκομμένων γραμμών, τριγώνων, σφαιρών και άλλων σχημάτων [45].

Ένα σημαντικό εργαλείο του KiNG είναι και το BACKRUB tool το οποίο χρησιμοποιείται ώστε να προσαρμόζουμε τα μικρά τμήματα της κύριας αλυσίδας των πρωτεϊνών (protein backbone) χωρίς να διαταράσσεται η περιβάλλουσα δομή. Μπορούμε δηλαδή να επιλέξουμε ένα κατάλοιπο και να το περιστρέψουμε γύρω από τον άξονα του ανάμεσα από τους γειτονικούς άνθρακες – α. Το εργαλείο αυτό επίσης παρέχει πληροφορίες σχετικά με την γεωμετρική ποιότητα του τρέχοντος μοντέλου δείχνοντας τα ονόματα των καταλοίπων, την απόκλιση της γωνίας τ σε σχέση με την ιδανική περίπτωση και την θέση του καταλοίπου στη γραφική παράσταση Ramachandran. Η γενικευμένη εκδοχή του BACKRUB είναι το Hinges tool το οποίο μας επιτρέπει να επιλέξουμε οποιαδήποτε συνεχόμενη περιοχή της κύριας αλυσίδας της πρωτεΐνης και η οποία ενώνει δύο άνθρακες – α και στη συνέχεια να περιστρέψουμε την περιοχή αυτή ανάμεσα από τους άνθρακες αυτούς. Το Sidechain Rotator χρησιμοποιείται σε συνδυασμό με το Hinges tool ώστε να μετασκευάσουμε με διαδραστικό τρόπο το πρωτεϊνικό μοντέλο ενώ το Sidechain Mutator χρησιμοποιείται για τον επανασχεδιασμό των πρωτεϊνικών μοντέλων, πάντα σε συνδυασμό με τα προηγούμενα εργαλεία. Τα εργαλεία αυτά δεν μπορούν να λειτουργήσουν χωρίς το Model Manager το οποίο επιτελεί μοριακή προτυποποίηση βασισμένη σε ένα αρχικό μοντέλο το οποίο είναι συνήθως ένα αρχείο PDB και είναι υπεύθυνο για το άνοιγμα και την αποθήκευση αυτών των αρχείων [47].

Άλλα χρήσιμα εργαλεία του KiNG είναι το Dock 3 – on – 3 το οποίο χρησιμοποιείται για την κατασκευή γεωμετρικών σχημάτων και μορφών από ήδη υπάρχοντα τμήματα μορφής kinemage, το Least – Squares Docking το οποίο συνταιριάζει δύο αντικείμενα τα οποία είναι παρόμοια αλλά όχι ίδια και το 90 Degree Rotation το οποίο παρέχει ένα εύκολο τρόπο ώστε να περιστρέψουμε την εικόνα μας κατά 90 μοίρες σε οποιονδήποτε από τους τρεις καρτεσιανούς άξονες. Επίσης, υπάρχει μια άλλη ομάδα εργαλείων τα οποία έχουν σχεδιαστεί ώστε η μετατροπή των kinemages να γίνεται εύκολα χωρίς να χρειάζεται να αλλάξουμε «με το χέρι» το αρχικό PDB αρχείο. Σ αυτή την ομάδα αναφέρουμε ενδεικτικά το Recoloring με το οποίο μπορούμε να αλλάξουμε τα χρώματα σε μια kinemage, το KinFudger το οποίο χρησιμεύει στο να

προσαρμόζουμε εύκολα τις αποστάσεις και τις γωνίες ανάμεσα στα σημεία και το Loop tool με το οποίο απομακρύνουμε τα ανεπιθύμητα κατάλοιπα από τις μακρομοριακές δομές [48].

Τέλος, κάποια πολύ χρήσιμα plug – in που χρησιμοποιούνται για εφαρμογές δομικής βιολογίας είναι το Analyze Geometry το οποίο αναλύει αυτόματα την γεωμετρία της κύριας αλυσίδας και το Fill – Gap plug – in το οποίο γεμίζει αυτόματα τα κενά που υπάρχουν σε πρωτεϊνικά μοντέλα μέσω κάποιου αλγορίθμου [48].

Το KiNG διατίθεται ελεύθερα από τη διεύθυνση http://kinemage.biochem.duke.edu και για να λειτουργήσει θα πρέπει να έχουμε εγκατεστημένη απαραιτήτως την Java καθώς το πρόγραμμα είναι γραμμένο σε αυτή τη γλώσσα και απαιτεί την ύπαρξη των Java βιβλιοθηκών στο σύστημα μας. Για την εργασία μας χρησιμοποιήσαμε την έκδοση 2.13.

Επιλέξαμε 3 χαρακτηριστικές κατά την άποψη μας πρωτεΐνες για να τις οπτικοποιήσουμε. Συγκεκριμένα πρόκειται για τις πρωτεΐνες με κωδικούς pdb 1ag2, 1fkb και 4rxn και στην οπτικοποίηση εκτός από την τρισδιάστατη δομή της πρωτεΐνης θα φαίνονται με χαρακτηριστικά χρώματα οι περιοχές TEF και MCF καθώς και οι κοινές τους περιοχές (δηλαδή αυτές που είναι τόσο MCF όσο και TEF, δηλαδή προβλέφθηκαν σωστά) καθώς επίσης και τα αμινοξέα που σημειώθηκαν ως MIR από τον αλγόριθμο για την αρχική μη μεταλλαγμένη ακολουθία. Να σημειώσουμε πως οι περιοχές MCF που χρησιμοποιήθηκαν για την οπτικοποίηση είναι ελαφρώς διαφορετικές από αυτές που βρήκαμε εμείς μέσω του αλγορίθμου μας στην εργασία και αυτό διότι έχουν ληφθεί και κάποιοι άλλοι παράγοντες υπόψη. Οι προαναφερθείσες πρωτεΐνες παρουσιάζουν μεγάλο πειραματικό ενδιαφέρον και μπορούν να αποδώσουν με αρκετά ευκρινή τρόπο τις περιοχές ενδιαφέροντος καθώς και την πρόβλεψη που έγινε από τον αλγόριθμο.

Σαν είσοδο στο KiNG χρησιμοποιήσαμε τα αρχεία pdb για κάθε πρωτεΐνη τα οποία βρήκαμε στη διεύθυνση http://www.rcsb.org/pdb/download/download.do και τα οποία περιέχουν τις συντεταγμένες για την τρισδιάστατη δομή των πρωτεϊνών την οποία απεικονίζουμε με το πρόγραμμα.. Στη συνέχεια, φτιάξαμε μέσω text editor και τα pdb αρχεία για τα MIR, καθώς και για τις TEF, MCF και κοινές τους περιοχές. Για την δημιουργία του pdb αρχείου για τα MIR επιλέξαμε μόνο τα άτομα Ca των αμινοξέων που σημειώθηκαν ως MIR σύμφωνα με την αρίθμηση που έχουν στην ακολουθία μας και την αντιστοιχία στα pdb αρχεία. Για την δημιουργία των pdb αρχείων για τις TEF, MCF και κοινές τους περιοχές, διαλέξαμε όλα τα άτομα των αμινοξέων που βρίσκονται μεταξύ της αρχής και του τέλους κάθε τμήματος. Έτσι, για την οπτικοποίηση κάθε πρωτεΐνης χρησιμοποιήσαμε το pdb αρχείο της τρισδιάστατης δομής της καθώς και ένα pdb αρχείο που περιέχει τις συντεταγμένες των σημειωμένων ως MIR αμινοξέων, ένα pdb αρχείο για κάθε TEF, ένα για κάθε MCF και ένα για κάθε κοινή τους περιοχή. Η διαδικασία ήταν η εξής :

Αρχικά εισάγαμε το pdb αρχείο με την τρισδιάστατη δομή της πρωτεΐνης και στις επιλογές εμφάνισης διαλέξαμε «protein», «backbone» και «sidechain». Στη συνέχεια εισάγαμε το pdb αρχείο που φτιάξαμε για τα MIR και στις επιλογές εμφάνισης διαλέξαμε «protein», «backbone», «balls on N, O, P etc» και «balls on C atoms too». Με τον ίδιο τρόπο εισάγαμε τα pdb αρχεία που είχαμε φτιάξει για κάθε TEF, MCF και κοινή περιοχή και επιλέξαμε «protein» και «backbone». Με τον τρόπο αυτό δημιουργήσαμε το kin αρχείο για την οπτικοποίηση της κάθε πρωτεΐνης.

Προκειμένου οι παραγόμενες kinemages να είναι πιο ευδιάκριτες ώστε να ξεχωρίζουν οι περιοχές ενδιαφέροντος μεταξύ τους, προχωρήσαμε σε κάποιες αλλαγές. Η μορφή των αρχείων kinemage είναι plain – text και μπορεί να διαβαστεί και να διορθωθεί από τον άνθρωπο με εύκολο τρόπο. Η συνολική δομή των αρχείων αυτών περιλαμβάνει ένα προαιρετικό @text μπλοκ που περιγράφει τα περιεχόμενα του αρχείου και στη συνέχεια ακολουθούν μία ή περισσότερες kinemages. Οι kinemages ξεκινούν με το μπλοκ @kinemage και περιέχουν μία σειρά από δηλώσεις group, subgroup, list και point οι οποίες περιλαμβάνουν με τη σειρά τους μια ιεραρχική οργάνωση τρισδιάστατων βασικών γραφικών στοιχείων όπως γραμμές, σφαίρες και τρίγωνα [49]. Έτσι, ανοίξαμε το kin αρχείο με κάποιο text editor και το τροποποιήσαμε κατάλληλα ώστε τα MIR να φαίνονται ως κόκκινα, στις MCF περιοχές να εναλλάσσονται δύο αποχρώσεις του πράσινου, στις TEF περιοχές να εναλλάσσονται δύο αποχρώσεις του κίτρινου και οι κοινές περιοχές να φαίνονται με καφέ χρώμα. Αυτό το κάναμε αλλάζοντας το πεδίο color στην αντίστοιχη γραμμή. Για παράδειγμα, για την πρωτεΐνη 1enh και την MCF περιοχή 3 – 18 πήγαμε στο κομμάτι

```
@group {1ENH_MCF3_18.pdb A} dominant

@subgroup {(implied)} nobutton

@vectorlist {protein bb} color= white master= {protein} master=
{backbone}
```

και αλλάξαμε το color= white σε color= green. Με τον τρόπο αυτό αλλάξαμε τα χρώματα και για τις υπόλοιπες περιοχές ενδιαφέροντος και έτσι προέκυψαν οι τελικές kinemages για την οπτικοποίηση των αποτελεσμάτων της εργασίας μας.

Κεφάλαιο 3: Αποτελέσματα

3.1 Παράθεση και ανάλυση αποτελεσμάτων

Στο παρούσα ενότητα θα παραθέσουμε τα αποτελέσματα της εργασίας τόσο συγκεντρωτικά για το σύνολο των πρωτεϊνών όσο και κατηγοριοποιημένα ανά ταξινόμηση οργανισμού, SCOP κατηγορίες (στοιχεία δευτεροταγούς δομής) και μήκος ακολουθίας, σε μορφή πινάκων. Για λόγους πληρότητας παραθέτουμε στο Παράρτημα το οποίο βρίσκεται στο τέλος της αναφοράς, τρεις πίνακες (Α, Β, Γ) με όλες τις πρωτεΐνες, τα χαρακτηριστικά τους και τις τιμές που έχουν για τα διάφορα μέτρα αξιολόγησης που χρησιμοποιήσαμε καθώς και πίνακες που περιέχουν τον μέσο όρο, την τυπική απόκλιση καθώς και την μέγιστη και ελάχιστη τιμή που παρουσιάζει η κάθε κατηγορία για τα διάφορα μέτρα αξιολόγησης (Α1-Α9, Β1-Β9, Γ1-Γ6). Οι πίνακες αφορούν τα αποτελέσματα της εργασίας τόσο για την πρόβλεψη των άκρων TEF όσο και την πρόβλεψη των καθαυτών TEF περιοχών και οι τιμές είναι σε μοναδιαία κλίμακα.

Στο σημείο αυτό θα πρέπει να διευκρινίσουμε πως συγκεκριμένα για τα μέτρα sensitivity και specificity και μόνον για την πρόβλεψη των άκρων TEF, χρησιμοποιήθηκε «παράθυρο» ± 5 θέσεων δηλαδή η πρόβλεψη θεωρείται σωστή όχι μόνο στην περίπτωση που ο αλγόριθμος προέβλεψε ακριβώς την θέση του άκρου TEF αλλά και όταν η προβλεπόμενη θέση απέχει το πολύ 5 θέσεις (και προς τις δυο κατευθύνσεις) από την θέση του άκρου TEF. Αυτό έγινε διότι σε προηγούμενη έρευνα που διενεργήθηκε και στην οποία βασίστηκε η παρούσα εργασία και αφορούσε την πρόβλεψη των άκρων των TEF, ελήφθη υπόψη το συγκεκριμένο «παράθυρο» και το χρησιμοποιήσαμε και εμείς ώστε να υπάρχει ομοιομορφία και αναλογία με την προηγούμενη έρευνα [25] και να είναι άμεσα συγκρίσιμα τα αποτελέσματα των δυο εργασιών. Στα υπόλοιπα μέτρα δεν χρησιμοποιήθηκε το συγκεκριμένο «παράθυρο» διότι είναι τέτοια η φύση των μέτρων αυτών που με το «παράθυρο» γίνονται αρκετά περίπλοκοι πλέον οι υπολογισμοί.

Αρχικά παρατίθενται δυο πίνακες, ένας που αφορά την πρόβλεψη των άκρων των TEF και ένας για την πρόβλεψη των TEF περιοχών και οι οποίοι περιέχουν συγκεντρωτικά τα αποτελέσματα για το σύνολο των πρωτεϊνών. Συγκεκριμένα στον πίνακα 3.1 παρατίθεται ο μέσος όρος που παρουσιάζουν τα μέτρα αξιολόγησης sensitivity, specificity και accuracy για την πρόβλεψη των άκρων TEF και στον πίνακα 3.2 ο μέσος όρος που παρουσιάζουν τα μέτρα αξιολόγησης sensitivity, specificity accuracy, SOV observed και SOV predicted για την πρόβλεψη των TEF.

Άκρα TEF

Πίνακας 3.1

Πίνακας 3.1 *Ο μέσος όρος που παρουσιάζουν τα μέτρα αξιολόγησης sensitivity, specificity και accuracy για την πρόβλεψη των άκρων των TEF.*

Στατιστικό Μέτρο	**Μέσος Όρος**
Sensitivity	0.42
Specificity	0.48
Accuracy	0.873

TEF Περιοχές

Πίνακας 3.2

Πίνακας 3.2 *Ο μέσος όρος που παρουσιάζουν τα μέτρα αξιολόγησης sensitivity, specificity, accuracy, SOV observed και SOV predicted για την πρόβλεψη των περιοχών TEF.*

Στατιστικό Μέτρο	**Μέσος Όρος**
Sensitivity	0.61
Specificity	0.63
Accuracy	0.572
SOV observed	0.50
SOV predicted	0.58

Σε γενικές γραμμές παρατηρούμε πως τα διάφορα στατιστικά μέτρα εμφανίζουν μέσους όρους που κυμαίνονται γύρω από την τιμή 0.5, είτε λίγο χαμηλότερη (όπως στην περίπτωση των sensitivity και specificity των άκρων TEF, είτε λίγο υψηλότερη όπως στην περίπτωση της πρόβλεψης των TEF περιοχών).

Συγκεκριμένα, για την πρόβλεψη των άκρων TEF, οι μέσοι όροι για τα μέτρα sensitivity και specificity κυμαίνονται στα ίδια πλαίσια και λίγο χαμηλότερα από την τιμή 0.5. Ο μέσος όρος του μέτρου accuracy κυμαίνεται σε υψηλότερα επίπεδα γιατί επηρεάζεται περισσότερο από την ομάδα των true negatives που είναι αρκετά μεγάλη σε πληθυσμό. Γι' αυτό τον λόγο θα λέγαμε πως τα πιο κατάλληλα μέτρα για να περιγράψουν την ποιότητα της πρόβλεψης των άκρων TEF είναι τα sensitivity και specificity και ακριβώς σε αυτά θα επικεντρωθούμε και θα παρουσιάσουμε και στην συνέχεια του κεφαλαίου και από την στιγμή που χρησιμοποιήσαμε το «παράθυρο» των ±5 θέσεων, οι ορισμοί των sensitivity και specificity διαφοροποιούνται ως εξής. Το στατιστικό μέτρο sensitivity, ουσιαστικά μας δείχνει το ποσοστό των άκρων TEF που βρίσκονται σε απόσταση ±5 θέσεων από άκρο MCF ενώ το στατιστικό μέτρο specificity, ουσιαστικά μας δείχνει το ποσοστό των προβλεπόμενων άκρων MCF που βρίσκονται σε απόσταση ±5 θέσεων από άκρο TEF.

Έτσι, παρατηρώντας τον πίνακα 3.1 βλέπουμε πως το 42% των άκρων TEF βρίσκονται σε απόσταση ±5 θέσεων από άκρο MCF, ενώ το 48% των άκρων MCF βρίσκονται σε απόσταση ±5 θέσεων από άκρο TEF. Σε σύγκριση με παλιότερη έρευνα που έχει διενεργηθεί και πάνω στην οποία βασίστηκε η παρούσα εργασία [25], παρατηρούμε το εξής. Χωρίς την εισαγωγή των μεταλλάξεων, η πρόβλεψη για τα άκρα των TEF γινόταν μέσω των MIR. Από τα αποτελέσματα της παλιότερης αυτής έρευνας, βρέθηκε πως το ποσοστό των MIR που βρίσκονται σε απόσταση ±5 θέσεων από άκρο TEF ήταν 57% ή στην μοναδιαία κλίμακα 0.57. Στην παρούσα έρευνα, έγινε μια απόπειρα να εισάγουμε ένα «φίλτρο» επιλογής των πιο «κατάλληλων» MIR μέσω του αλγόριθμου παραγωγής των MCF περιοχών (βλ. ενότητα 2.5) και πλέον η πρόβλεψη για τα άκρα των TEF γίνεται μέσω των άκρων MCF. Παρότι τα νούμερα σε πολύ γενικές γραμμές κρίνονται σχετικά καλά, παρόλα αυτά παρατηρούμε πως το ποσοστό της πρόβλεψης μειώθηκε από 0.57 σε 0.48.

Όσον αφορά την πρόβλεψη των TEF περιοχών, παρατηρούμε πως οι μέσοι όροι για τα μέτρα sensitivity και specificity έχουν σχεδόν την ίδια τιμή. Ο μέσος όρος του accuracy είναι αρκετά μικρότερος από τον αντίστοιχο της πρόβλεψης των άκρων TEF και αυτό γιατί η ομάδα των true negatives η οποία επηρεάζει αρκετά την τιμή του μέτρου αυτού, είναι πλέον μικρότερη σε πληθυσμό στην περίπτωση της πρόβλεψης των TEF περιοχών. Τέλος, οι μέσοι όροι των μέτρων SOV observed και SOV predicted κυμαίνονται λίγο πάνω από το 0.50. Εφόσον όμως πρόκειται για πρόβλεψη τμημάτων και όχι θέσεων, ως περισσότερο κατάλληλα στατιστικά μέτρα, λόγω της φύσης τους, κρίνονται τα SOV observed και SOV predicted, με τα οποία και θα ασχοληθούμε περισσότερο και στη συνέχεια του κεφαλαίου. Αυτό συμβαίνει γιατί είναι έτσι ορισμένα ώστε να δίνουν περισσότερη έμφαση στην επικάλυψη των διαφόρων τμημάτων και να αφήνουν ένα ποσοστό ανοχής για τα άκρα τους.

Το SOV observed δείχνει τι ποσοστό των περιοχών TEF προβλέφθηκε σωστά από τις MCF περιοχές, ενώ το SOV predicted δείχνει τι ποσοστό των περιοχών MCF ανήκει σε περιοχές TEF. Ουσιαστικά τα SOV είναι το αντίστοιχο των στατιστικών μέτρων sensitivity και specificity για ολόκληρα τμήματα. Να υπενθυμίσουμε πως τα μέτρα SOV δεν υπολογίζουν πλήθος περιοχών που προβλέφθηκαν σωστά αλλά τον «βαθμό επικάλυψης» ανάμεσα στις περιοχές αυτές και χρησιμοποιούνται μαζί ώστε να έχουμε περισσότερο ολοκληρωμένη άποψη για τις προβλέψεις μας. Παρατηρώντας λοιπόν τον πίνακα 3.2 και αν συνυπολογίσουμε τις τιμές των SOV, βλέπουμε πως αυτές κινούνται λίγο πάνω από το 0.50.

Αφού λοιπόν είχαμε μια πρώτη συνολική εικόνα της απόδοσης του αλγορίθμου μέσω των πινάκων 3.1 και 3.2, στη συνέχεια της ενότητας θα προχωρήσουμε στην «συμπεριφορά» ανά κατηγορία. Για το λόγο αυτό θα παραθέσουμε πίνακες οι οποίοι θα περιέχουν για κάθε κατηγορία και για τις τρεις κατηγοριοποιήσεις που έγιναν, τον μέσο όρο των μέτρων sensitivity και specificity για την πρόβλεψη των άκρων TEF και τον μέσο όρο των μέτρων SOV observed και SOV predicted για την πρόβλεψη των TEF περιοχών. Για τα υπόλοιπα μέτρα καθώς και για πρόσθετες πληροφορίες

(τυπική απόκλιση, μέγιστες και ελάχιστες τιμές) σχετικά με τα μέτρα αυτά, ο αναγνώστης μπορεί να ανατρέξει στο Παράρτημα.

Στη πρώτη στήλη κάθε πίνακα αναφέρεται η ιδιότητα των πρωτεϊνών με βάση την οποία γίνεται η εκάστοτε κατηγοριοποίηση. Η πρώτη κατηγοριοποίηση έγινε με βάση την ταξινόμηση του οργανισμού από τον οποίον προέρχεται η κάθε πρωτεΐνη (βακτήρια, ευκαριωτικά, ιοί), η δεύτερη κατηγοριοποίηση έγινε με βάση τις SCOP κατηγορίες (all a, all b, a+b, a/b, small) και η τελευταία με βάση το μήκος της αλυσίδας (0-100, πάνω από 100). Στη δεύτερη στήλη με τίτλο «αριθμός πρωτεϊνών» αναφέρεται το πλήθος των πρωτεϊνών που ανήκουν στην κάθε κατηγορία ώστε να υπάρχει μια εποπτική εικόνα σχετικά με την κατανομή των πληθυσμών. Η τελευταία στήλη αναφέρεται σε καθένα από τα στατιστικά μέτρα (sensitivity, specificity για τα άκρα TEF και SOV observed και SOV predicted για τις TEF περιοχές) και για καθένα από αυτά έχει υπολογιστεί η μέση τιμή για κάθε κατηγορία.

Η πρώτη κατηγοριοποίηση έγινε με βάση την ταξινόμηση των οργανισμών από τον οποίον προέρχονται οι πρωτεΐνες. Έτσι, οι πρωτεΐνες ουσιαστικά χωρίζονται σε τρεις κατηγορίες: βακτήρια, ευκαριωτικά και ιοί με τις δύο πρώτες ομάδες να είναι μεγαλύτερες σε πληθυσμό. Δυστυχώς το δείγμα για την κατηγορία των ιών είναι σχετικά μικρό με αποτέλεσμα η εκπροσώπηση της κατηγορίας αυτής να είναι ισχνή. Σίγουρα, αν ο πληθυσμός της κατηγορίας αυτής ήταν μεγαλύτερος, τα αποτελέσματα μας θα είχαν και ακόμα μεγαλύτερη εγκυρότητα και θα μας οδηγούσαν σε πιο ασφαλή συμπεράσματα. Εκτός από αυτές υπάρχει και άλλη μια ομάδα πρωτεϊνών, τα αρχαία, με μόλις μία πρωτεΐνη να ανήκει στην ομάδα αυτή και συγκεκριμένα την πρωτεΐνη με pdb κωδικό 1caa καθώς και άλλη μια πρωτεΐνη με κωδικό 1abrA η οποία δεν κατατάσσεται πουθενά. Οι συγκεκριμένες πρωτεΐνες δεν εμφανίζονται πουθενά στους αντίστοιχους πίνακες και αυτό διότι δεν χρησιμεύουν στο να εξαχθεί κάποιο συμπέρασμα. Οι παρακάτω αναλύσεις της συγκεκριμένης κατηγοριοποίησης δίνουν βαρύτητα στις τρεις πρώτες κατηγορίες.

Άκρα TEF

Πίνακας 3.3

Πίνακας 3.3 *Ο μέσος όρος που παρουσιάζουν τα μέτρα αξιολόγησης sensitivity, specificity για κάθε μια από τις τρεις κατηγορίες που αφορούν την ταξινόμηση του οργανισμού από τον οποίο προέρχονται (βακτήρια, ευκαριωτικά, ιοί) όσον αφορά την πρόβλεψη των άκρων TEF.*

Ταξινόμηση	αριθμός πρωτεϊνών	Sensitivity	Specificity
		Μέση τιμή	Μέση τιμή
Βακτήρια	32	0.39	0.46
Ευκαριωτικά	64	0.44	0.49
Ιοί	9	0.43	0.53

TEF Περιοχές

Πίνακας 3.4

Πίνακας 3.4 *Ο μέσος όρος που παρουσιάζουν τα μέτρα αξιολόγησης SOV observed, SOV predicted για κάθε μια από τις τρεις κατηγορίες που αφορούν την ταξινόμηση του οργανισμού από τον οποίο προέρχονται (βακτήρια, ευκαριωτικά, ιοί) όσον αφορά την πρόβλεψη των περιοχών TEF.*

Ταξινόμηση	αριθμός πρωτεϊνών	SOV observed	SOV predicted
		Μέση τιμή	Μέση τιμή
Βακτήρια	32	0.50	0.65
Ευκαριωτικά	64	0.49	0.55
Ιοί	9	0.55	0.56

Από τον πίνακα 3.3 παρατηρούμε πως τα ζεύγη των μέσων όρων sensitivity – specificity είναι μεγαλύτερα για την κατηγορία των ιών, ακολουθεί η κατηγορία των ευκαριωτικών ενώ μικρότερους μέσους όρους παρουσιάζει η κατηγορία των βακτηρίων. Βέβαια πρέπει να τονίσουμε πως οι διαφορές γενικά είναι συγκριτικά πολύ μικρές μεταξύ τους. Επίσης, σε σύγκριση με την παλιότερη έρευνα [25], μπορούμε να παρατηρήσουμε πως ο μέσος όρος του specificity για την κατηγορία των ιών πλησιάζει το 0.57 που είχε υπολογιστεί στην έρευνα αυτή, ενώ οι μέσοι όροι του μέτρου αυτού για τις υπόλοιπες κατηγορίες είναι πιο κοντά στον μέσο όρο που υπολογίστηκε στην παρούσα εργασία (0.48). Από τον πίνακα 3.4 μπορούμε να παρατηρήσουμε πως τα ζεύγη των μέσων όρων SOV observed και SOV predicted παρουσιάζονται ελαφρώς πιο χαμηλά για την κατηγορία των ευκαριωτικών ενώ οι υπόλοιπες δύο κατηγορίες εμφανίζουν ζεύγη μέσων όρων περίπου στα ίδια πλαίσια και γύρω από τον μέσο όρο που υπολογίστηκε στον πίνακα 3.2.

Αν σε όλα αυτά συμπεριλάβουμε και τους πίνακες που υπάρχουν στο παράρτημα τότε μπορούμε να συμπεράνουμε πως η κατηγορία που εμφανίζει συχνότερα τους υψηλότερους μέσους όρους στα συγκεκριμένα μέτρα αξιολόγησης σε σχέση με τις άλλες κατηγορίες είναι η κατηγορία των ιών. Συγκεκριμένα, εμφανίζει τους υψηλότερους μέσους όρους, σε σχέση με τις υπόλοιπες κατηγορίες, για την πρόβλεψη των άκρων TEF στα μέτρα specificity (πίνακας 3.3) και accuracy (παράρτημα: πίνακας Α3) και στην πρόβλεψη των TEF στα μέτρα sensitivity (παράρτημα: πίνακας Β1), accuracy (παράρτημα: πίνακας Β3) καθώς και στο μέτρο SOV observed (πίνακας 3.4), στις 5 δηλαδή από τις 8 διαφορετικές περιπτώσεις που έχουμε για τους συνδυασμούς πρόβλεψη – στατιστικό μέτρο (πίνακες Α1 ως Α3, Β1 ως Β3, Γ1 και Γ2 του παραρτήματος). Τους δε χαμηλότερους μέσους όρους, σε σχέση με τις υπόλοιπες κατηγορίες, παρουσιάζει η κατηγορία των ευκαριωτικών σε 4 από τις 8 περιπτώσεις και συγκεκριμένα στα μέτρα accuracy (παράρτημα: πίνακας Α3) της πρόβλεψης των άκρων TEF και specificity (παράρτημα: πίνακας Β2), SOV observed και SOV predicted (πίνακας 3.4) της πρόβλεψης των TEF με τις πρωτεΐνες που προέρχονται

από τα βακτήρια να έχουν τους μικρότερους μέσους όρους στις υπόλοιπες περιπτώσεις. Εξαίρεση αποτελεί το μέτρο accuracy της πρόβλεψης των TEF (παράρτημα: πίνακας Β3) όπου τόσο η κατηγορία των βακτηρίων όσο και η κατηγορία των ευκαριωτικών παρουσιάζουν από κοινού τους χαμηλότερους μέσους όρους.

Από τα παραπάνω συμπεραίνουμε πως με βάση τους μέσους όρους των στατιστικών μέτρων, καλύτερα προβλέπονται (τόσο για τα TEF όσο και για τα άκρα τους) οι πρωτεΐνες που προέρχονται από ιούς, ενώ με βάση τους πίνακες 3.3 και 3.4 θα λέγαμε πως χειρότερα προβλέπονται τα άκρα TEF για την κατηγορία των βακτηρίων ενώ χειρότερα προβλέπονται οι περιοχές TEF για την κατηγορία των ευκαριωτικών. Θα μπορούσαμε να ισχυριστούμε πως οι δυο αυτές κατηγορίες δε φαίνεται να έχουν ιδιαίτερα μεγάλες διαφορές μεταξύ τους. Αυτή που φαίνεται να ξεχωρίζει κάπως είναι η κατηγορία των ιών, που όπως προείπαμε όμως, έχει μικρή εκπροσώπηση στο στατιστικό μας δείγμα με συνέπεια να μην έχουμε την μέγιστη δυνατή εγκυρότητα για τα αποτελέσματα μας. Τέλος, περισσότερο συγκεντρωμένες στην μέση τιμή παρουσιάζονται οι τιμές των στατιστικών μέτρων για την κατηγορία των ιών. Οι τιμές της τυπικής απόκλισης για την κατηγοριοποίηση αυτή και για κάθε στατιστικό μέτρο φαίνονται στους πίνακες Α1-Α3, Β1-Β3, Γ1 και Γ2 του παραρτήματος.

Η δεύτερη κατηγοριοποίηση έγινε με βάση τις scop κατηγορίες στις οποίες εντάσσονται οι πρωτεΐνες, δηλαδή τα στοιχεία δευτεροταγούς δομής που διαθέτουν. Έτσι, οι πρωτεΐνες ουσιαστικά χωρίζονται σε πέντε μεγάλες και σχεδόν ίσες σε πληθυσμό κατηγορίες: all a, all b, a+b, a/b και small. Εκτός από αυτές, υπάρχουν επίσης και μια πρωτεΐνη coil και μια membrane με κωδικούς 1aa0 και 1occD αντίστοιχα. Όπως είναι φυσικό, η ανάλυση των αποτελεσμάτων δίνει βαρύτητα στις 5 προαναφερθείσες κατηγορίες και είναι αυτές οι 5 που αναφέρονται και στους πίνακες.

Άκρα TEF

Πίνακας 3.5

Πίνακας 3.5 *Ο μέσος όρος που παρουσιάζουν τα μέτρα αξιολόγησης sensitivity, specificity για κάθε μια από τις scop κατηγορίες (all a, all b, a+b, a/b, small) όσον αφορά την πρόβλεψη των άκρων TEF.*

Scop κατηγορίες	αριθμός πρωτεϊνών	Sensitivity	Specificity
		Μέση τιμή	Μέση τιμή
all a	24	0.75	0.49
all b	21	0.30	0.45
a+b	26	0.39	0.48
a/b	17	0.52	0.56
small	17	0.43	0.43

TEF Περιοχές

Πίνακας 3.6

Πίνακας 3.6 *Ο μέσος όρος που παρουσιάζουν τα μέτρα αξιολόγησης SOV observed, SOV predicted για κάθε μια από τις scop κατηγορίες (all a, all b, a+b, a/b, small) όσον αφορά την πρόβλεψη των περιοχών TEF.*

Scop κατηγορίες	αριθμός πρωτεϊνών	SOV observed	SOV predicted
		Μέση τιμή	Μέση τιμή
all a	24	0.46	0.58
all b	21	0.45	0.58
a+b	26	0.53	0.60
a/b	17	0.54	0.62
small	17	0.55	0.54

Από τον πίνακα 3.5 παρατηρούμε πως τα ζεύγη των μέσων όρων sensitivity – specificity παρουσιάζονται συγκριτικά μεγαλύτερα στις κατηγορίες «all a» και «a/b». Μάλιστα, στην περίπτωση της κατηγορίας «a/b», ο μέσος όρος του specificity είναι σχεδόν ίσος με το 0.57 που είχε υπολογιστεί στην προηγούμενη έρευνα ενώ οι μέσοι όροι του μέτρου αυτού για τις υπόλοιπες κατηγορίες είναι πιο κοντά στον μέσο όρο που υπολογίστηκε στην παρούσα εργασία και που φαίνεται στον πίνακα 3.2. Χαμηλότερους συγκριτικά μέσους όρους φαίνεται να έχει η κατηγορία «all b». Από τον πίνακα 3.6 μπορούμε να παρατηρήσουμε πως τα ζεύγη των μέσων όρων SOV observed και SOV predicted παρουσιάζονται περίπου να κυμαίνονται στα ίδια πλαίσια, χωρίς ιδιαίτερες διαφορές μεταξύ τους. Ελαφρώς υψηλότεροι πάντως, φαίνονται οι μέσοι όροι για τις κατηγορίες «a/b» και «a+b», με την κατηγορία «all b» ξανά να είναι στις συγκριτικά λιγότερο καλά προβλεπόμενες πρωτεΐνες.

Από τους παραπάνω πίνακες καθώς και από τους πίνακες του παραρτήματος μπορούμε να ισχυριστούμε πως η κατηγορία που εμφανίζει συχνότερα τους υψηλότερους μέσους όρους στα συγκεκριμένα μέτρα αξιολόγησης σε σχέση με τις άλλες κατηγορίες είναι η κατηγορία «a/b». Συγκεκριμένα, εμφανίζει τους υψηλότερους μέσους όρους για την πρόβλεψη των άκρων TEF στο μέτρο specificity (πίνακας 3.5) και στην πρόβλεψη των TEF στα μέτρα sensitivity (παράρτημα: πίνακας Β4), specificity (παράρτημα: πίνακας Β5), accuracy (παράρτημα: πίνακας Β6) και SOV predicted (πίνακας 3.6), στις 5 δηλαδή από τις 8 διαφορετικές περιπτώσεις που έχουμε για τους συνδυασμούς πρόβλεψη – στατιστικό μέτρο (πίνακες Α4 ως Α6, Β4 ως Β6, Γ3 και Γ4 του παραρτήματος). Τους δε χαμηλότερους μέσους όρους παρουσιάζει η κατηγορία «all b» σε 4 από τις 8 περιπτώσεις και συγκεκριμένα στα μέτρα sensitivity (πίνακας 3.5) της πρόβλεψης των άκρων TEF και sensitivity (παράρτημα: πίνακας Β4), accuracy (παράρτημα: πίνακας Β6) και SOV observed (πίνακας 3.6) της πρόβλεψης των TEF.

Συνοψίζοντας, με βάση τους μέσους όρους των συγκεκριμένων στατιστικών μέτρων υπάρχει μια τάση να προβλέπεται κάπως καλύτερα η κατηγορία «a/b», και οι λιγότερο καλά προβλεπόμενη κατηγορία φαίνεται από τους πίνακες η «all b» η οποία σημειωτέον, σε καμία περίπτωση συνδυασμού πρόβλεψη – στατιστικό μέτρο δεν εμφανίζει υψηλότερο μέσο όρο από τις υπόλοιπες κατηγορίες. Τέλος, περισσότερο συγκεντρωμένες στην μέση τιμή είναι οι τιμές των στατιστικών μέτρων για την κατηγορία «a/b». Οι τιμές της τυπικής απόκλισης για την κατηγοριοποίηση αυτή και για κάθε στατιστικό μέτρο φαίνονται στους πίνακες Α4-Α6, Β4-Β6, Γ3 και Γ4 του παραρτήματος.

Η τελευταία κατηγοριοποίηση έγινε με βάση το μήκος της ακολουθίας των πρωτεϊνών. Έτσι, οι πρωτεΐνες χωρίζονται σε δύο κατηγορίες: η πρώτη κατηγορία περιλαμβάνει αυτές που το μήκος τους είναι από 0-100 κατάλοιπα και η δεύτερη αυτές που το μήκος τους είναι μεγαλύτερο των 100 καταλοίπων. Η δεύτερη κατηγορία είναι κάτι παραπάνω από διπλάσια σε πληθυσμό σε σχέση με την πρώτη.

Άκρα TEF

Πίνακας 3.7

Πίνακας 3.7 *Ο μέσος όρος που παρουσιάζουν τα μέτρα αξιολόγησης sensitivity, specificity για κάθε μια από τις κατηγορίες που αφορούν το μήκος της αλυσίδας (0 – 100, >100) όσον αφορά την πρόβλεψη των άκρων TEF.*

Μήκος Ακολουθίας	αριθμός πρωτεϊνών	Sensitivity	Specificity
		Μέση τιμή	Μέση τιμή
0 – 100	33	0.37	0.36
>100	74	0.44	0.54

TEF Περιοχές

Πίνακας 3.8

Πίνακας 3.8 *Ο μέσος όρος που παρουσιάζουν τα μέτρα αξιολόγησης SOV observed, SOV predicted για κάθε μια από τις κατηγορίες που αφορούν το μήκος της αλυσίδας (0 – 100, >100) όσον αφορά την πρόβλεψη των περιοχών TEF.*

Μήκος Ακολουθίας	αριθμός πρωτεϊνών	SOV observed	SOV predicted
		Μέση τιμή	Μέση τιμή
0 – 100	33	0.55	0.57
>100	74	0.48	0.59

Από τον πίνακα 3.7 παρατηρούμε πως τα ζεύγη των μέσων όρων sensitivity – specificity παρουσιάζονται συγκριτικά μεγαλύτερα στην κατηγορία «μήκος μεγαλύτερο του 100». Μάλιστα, στην περίπτωση της κατηγορίας αυτής, ο μέσος όρος

του specificity είναι σχεδόν ίσος με το 0.57 που είχε υπολογιστεί στην προηγούμενη έρευνα ενώ ο μέσος όρος του μέτρου αυτού για την άλλη κατηγορία είναι συγκριτικά αρκετά μικρότερος από τον μέσο όρο που υπολογίστηκε στην παρούσα εργασία και που φαίνεται στον πίνακα 3.2. Από τον πίνακα 3.8 μπορούμε να παρατηρήσουμε πως τα ζεύγη των μέσων όρων SOV observed και SOV predicted παρουσιάζονται περίπου να κυμαίνονται στα ίδια πλαίσια, χωρίς ιδιαίτερες διαφορές μεταξύ τους. Ελαφρώς υψηλότερος πάντως, φαίνεται ο μέσος όρος του SOV observed για την κατηγορία «0-100» χωρίς όμως σημαντικές διαφορές, ενώ ο μέσος όρος του SOV predicted είναι σχεδόν ίδιος και για τις δύο κατηγορίες.

Από τους παραπάνω πίνακες αλλά και από τους πίνακες του παραρτήματος συμπεραίνουμε πως η κατηγορία που εμφανίζει συχνότερα τους υψηλότερους μέσους όρους στα συγκεκριμένα μέτρα αξιολόγησης σε σχέση με τις άλλες κατηγορίες είναι η κατηγορία «μήκος μεγαλύτερο του 100». Συγκεκριμένα, εμφανίζει τους υψηλότερους μέσους όρους για την πρόβλεψη των άκρων TEF στα μέτρα sensitivity, specificity (πίνακας 3.7) και accuracy (παράρτημα: πίνακας Α9) και στην πρόβλεψη των TEF στα μέτρα specificity (παράρτημα: πίνακας Β8) και accuracy (παράρτημα: πίνακας Β9) καθώς και στο μέτρο SOV predicted (πίνακας 3.8), στις 6 δηλαδή από τις 8 διαφορετικές περιπτώσεις που έχουμε για τους συνδυασμούς πρόβλεψη – στατιστικό μέτρο (πίνακες Α7 ως Α9, Β7 ως Β9, Γ5 και Γ6 του παραρτήματος). Άρα, η κατηγορία «μήκος μεγαλύτερο του 100» προβλέπεται καλύτερα από την κατηγορία «μήκος 0-100» κυρίως όσον αφορά την πρόβλεψη των άκρων των TEF, ενώ όσον αφορά την πρόβλεψη των TEF περιοχών, θα μπορούσαμε να ισχυριστούμε κυρίως βάσει των πινάκων 3.7 και 3.8 ότι ο αλγόριθμος δεν παρουσιάζει κάποια έντονη συμπεριφορά υπέρ κάποιας κατηγορίας. Τέλος, περισσότερο συγκεντρωμένες στην μέση τιμή είναι οι τιμές των στατιστικών μέτρων για την κατηγορία «μήκος μεγαλύτερο του 100». Οι τιμές της τυπικής απόκλισης για την κατηγοριοποίηση αυτή και για κάθε στατιστικό μέτρο φαίνονται στους πίνακες Α7-Α9, Β7-Β9, Γ5 και Γ6 του παραρτήματος.

Τέλος, να επισημάνουμε πως για τις πρωτεΐνες με κωδικό 1bvd, 1edmB, 1ejgA, 1knt, 1pwt, ο αλγόριθμος που χρησιμοποιήθηκε δεν παρήγαγε ως έξοδο κάποια MCF περιοχή και γι αυτό οι πρωτεΐνες αυτές παρουσιάζουν μηδενικές τιμές για τα στατιστικά μέτρα.

Κλείνοντας την ενότητα αυτή θα λέγαμε πως οι μέσοι όροι κρίνονται σχετικά καλοί σε πολύ γενικά πλαίσια, παρόλα αυτά θα πρέπει να τονίσουμε πως πλέον το ποσοστό των MIR (στην παρούσα εργασία τα άκρα MCF) που βρίσκονται σε απόσταση ±5 θέσεων από άκρο TEF έπεσε από το 57% στο 48%. Αυτό σε πρώτη ανάγνωση σημαίνει πως η παρούσα εργασία δε δείχνει να βελτιώνει τις προβλέψεις σε σχέση με την αρχική έρευνα [25]. Από την άλλη, τα ποσοστά επικάλυψης των περιοχών TEF με τις προβλεπόμενες περιοχές (MCF) είναι πολύ λίγο μεγαλύτερα από την τάξη του 50%, ένα σχετικά ικανοποιητικό ποσοστό.

Όσον αφορά τη «συμπεριφορά» του αλγορίθμου, θα λέγαμε πως γενικά δεν παρουσιάζει κάποια πολύ ισχυρή «προτίμηση» υπέρ ή κατά κάποιας κατηγορίας στις διάφορες κατηγοριοποιήσεις που έγιναν. Η πρόβλεψη φαίνεται να είναι καλύτερη για τις πρωτεΐνες που προέρχονται από ιούς, όπως και για πρωτεΐνες των οποίων η SCOP κατηγοριοποίηση είναι «a/b». Αντίθετα, λιγότερο καλή φαίνεται η πρόβλεψη για τις πρωτεΐνες που ανήκουν στην κατηγορία «all b». Τέλος, όσον αφορά το μήκος της ακολουθίας, επίσης φαίνεται ότι ο αλγόριθμος δεν «προτιμά» αυστηρά κάποια κατηγορία έναντι άλλης. Μια πολύ ελαφρά «προτίμηση» παρατηρείται υπέρ της κατηγορίας «μήκος μεγαλύτερο του 100» τουλάχιστον όσον αφορά την πρόβλεψη των άκρων TEF, ενώ οι δυο αυτές κατηγορίες παρουσιάζονται να προβλέπονται εξίσου καλά στην περίπτωση της πρόβλεψης των περιοχών TEF.

3.2 Οπτικοποίηση προβλέψεων

Στην ενότητα αυτή θα παραθέσουμε τρισδιάστατα γραφήματα, τις λεγόμενες kinemages, ήτοι κινητικές εικόνες που δημιουργήσαμε με τη βοήθεια του προγράμματος KiNG προκειμένου να έχει ο αναγνώστης μια πιο πλήρη εικόνα των προβλέψεων.

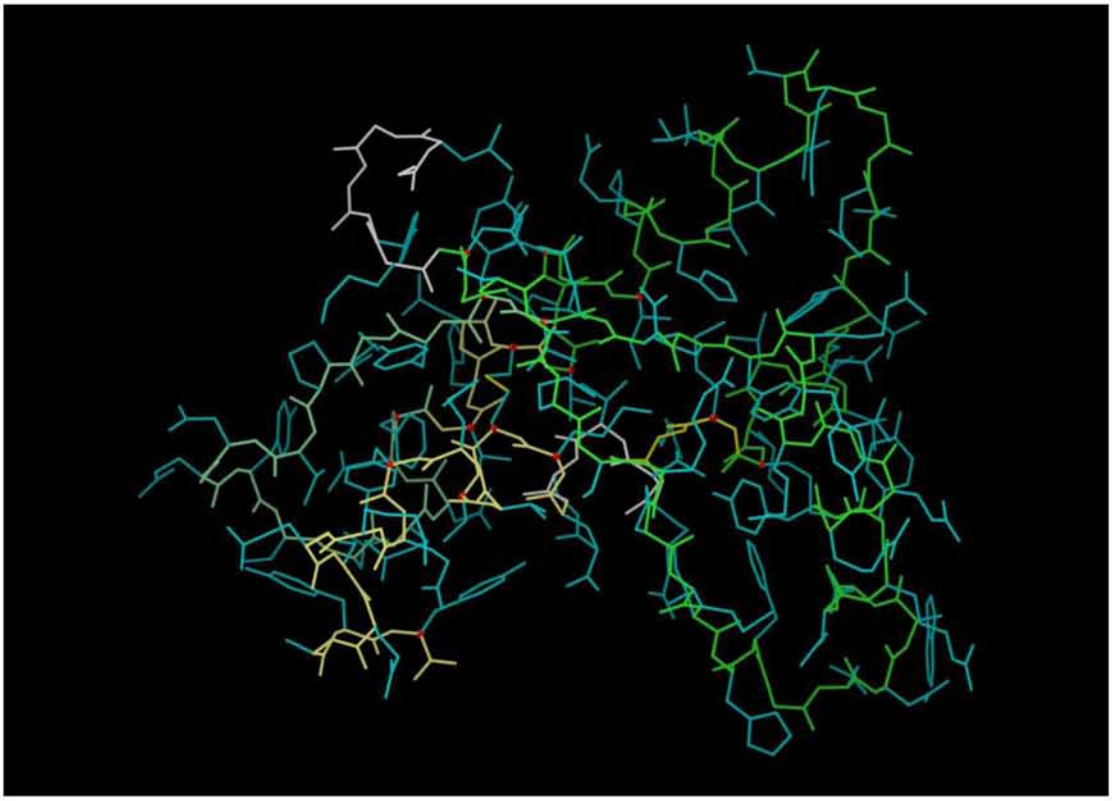
Από το σύνολο των 107 πρωτεϊνών, επιλέξαμε να παραθέσουμε τα γραφήματα για 3 χαρακτηριστικές πρωτεΐνες οι οποίες παρουσιάζουν μεγαλύτερο πειραματικό ενδιαφέρον, είναι ευκρινείς και ανήκουν στις καλύτερα προβλεπόμενες πρωτεΐνες με βάση τους πίνακες του Παραρτήματος.

Για κάθε πρωτεΐνη παραθέτουμε 4 γραφήματα, το πρώτο το οποίο δείχνει το σύνολο της πρωτεΐνης περιλαμβάνοντας τόσο την κύρια όσο και τις πλευρικές αλυσίδες, το δεύτερο το οποίο περιλαμβάνει μόνο την κύρια αλυσίδα περισσότερο για λόγους ευκρίνειας και μόνο τις TEF περιοχές, το τρίτο το οποίο περιλαμβάνει μόνο την κύρια αλυσίδα και μόνο τις MCF περιοχές και το τέταρτο το οποίο περιλαμβάνει μόνο την κύρια αλυσίδα και το σύνολο των TEF, MCF και σωστά προβλεπόμενων περιοχών δηλαδή τις περιοχές TEF οι οποίες προβλέφθηκαν σωστά από τον αλγόριθμο ως MCF, προκειμένου να έχουμε μια καλύτερη εικόνα για τις προβλέψεις. Στα γραφήματα τα οποία θα παρουσιάσουμε, τα ακόλουθα στοιχεία των πρωτεϊνών φαίνονται με τα εξής χρώματα:

- Πλευρικές αλυσίδες: χρώμα κυανό.
- Περιοχές TEF: εναλλασσόμενες αποχρώσεις κίτρινου χρώματος.
- Περιοχές MCF: εναλλασσόμενες αποχρώσεις πράσινου χρώματος.
- Περιοχές TEF που προβλέφθηκαν σωστά ως MCF (κοινές περιοχές): χρώμα καφέ.
- MIR: μικρές σφαίρες κόκκινου χρώματος.
- Περιοχές της πρωτεΐνης που ανήκουν στην κύρια αλυσίδα αλλά δεν είναι ούτε MCF ούτε TEF: αποχρώσεις του γκρι.

Στο σημείο αυτό να τονίσουμε πως οι MCF περιοχές που χρησιμοποιήθηκαν στα γραφήματα αυτά είναι ελαφρώς διαφορετικές από αυτές που παρήχθησαν σαν έξοδο από τον αλγόριθμο μας και αυτό επειδή ελήφθησαν και κάποιοι άλλοι παράγοντες υπόψη. Η λογική και το πνεύμα της πρόβλεψης φυσικά παραμένουν τα ίδια. Στη συνέχεια, παραθέτουμε τα γραφήματα για κάθε μία από τις πρωτεΐνες που επιλέξαμε.

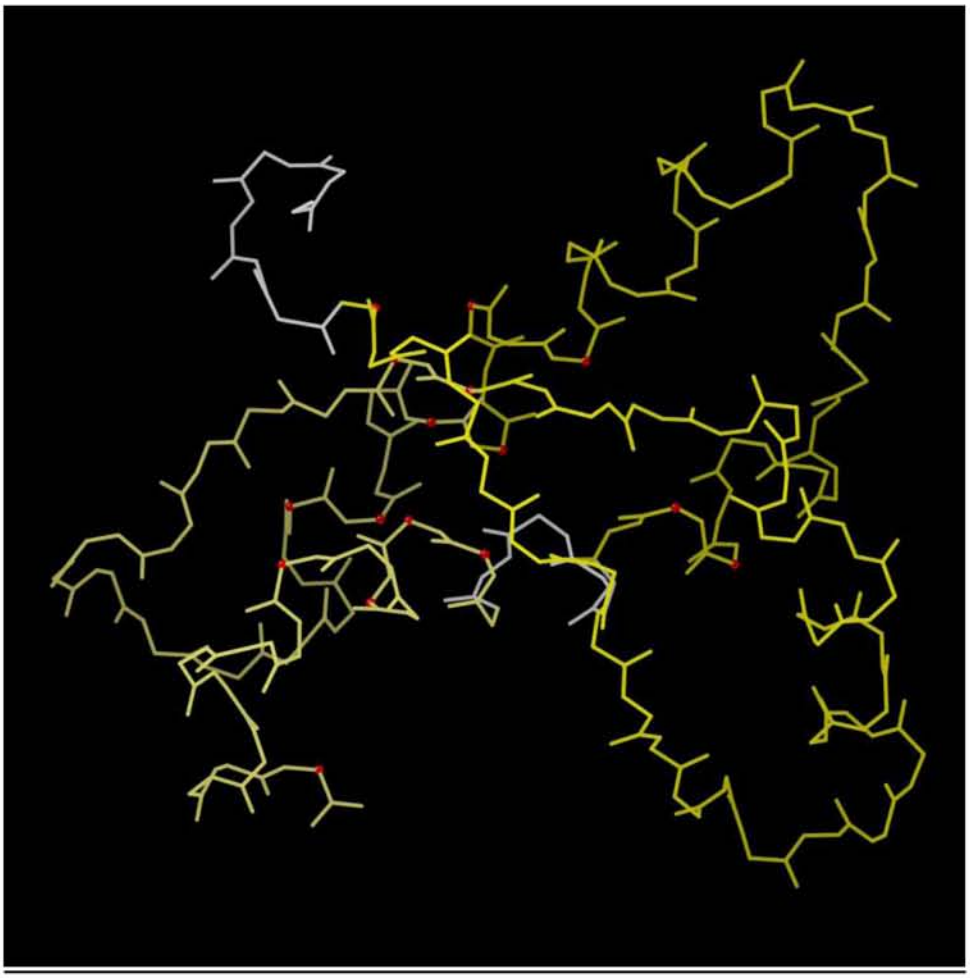
Πρωτεΐνη 1ag2



Σχήμα 3.1 *Κύρια και πλευρικές αλυσίδες της πρωτεΐνης 1ag2. Με εναλλαγές του πράσινου χρώματος φαίνονται οι προβλεπόμενες MCF περιοχές, με εναλλαγές του κίτρινου φαίνονται οι TEF περιοχές, με κόκκινο χρώμα τα MIR, με κυανό οι πλευρικές αλυσίδες και με αποχρώσεις του γκρι οι υπόλοιπες περιοχές της κύριας αλυσίδας.*

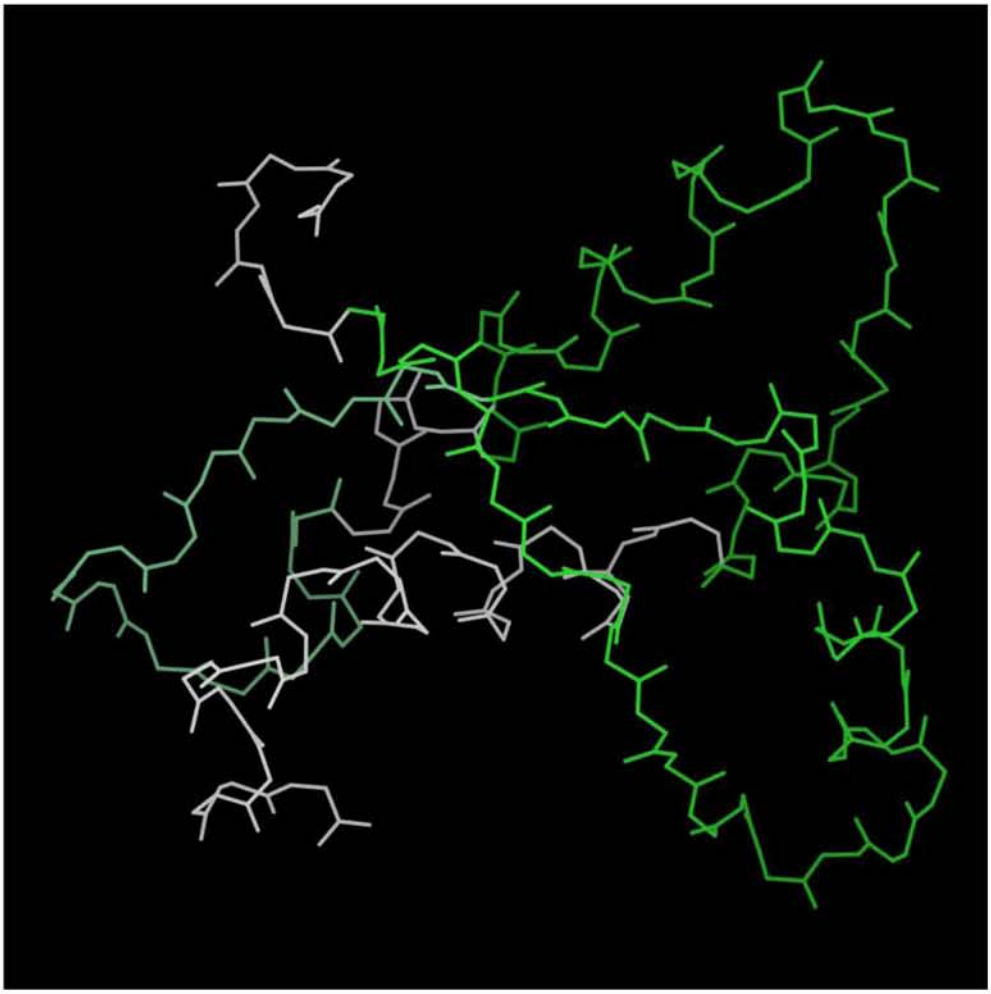
.

Πρωτεΐνη 1ag2 – TEF



Σχήμα 3.2 *Κύρια αλυσίδα της πρωτεΐνης 1ag2. Με εναλλαγές του κίτρινου φαίνονται οι TEF περιοχές, με κόκκινο χρώμα τα MIR και με αποχρώσεις του γκρι οι υπόλοιπες περιοχές της κύριας αλυσίδας.*

Πρωτεΐνη 1ag2 – MCF

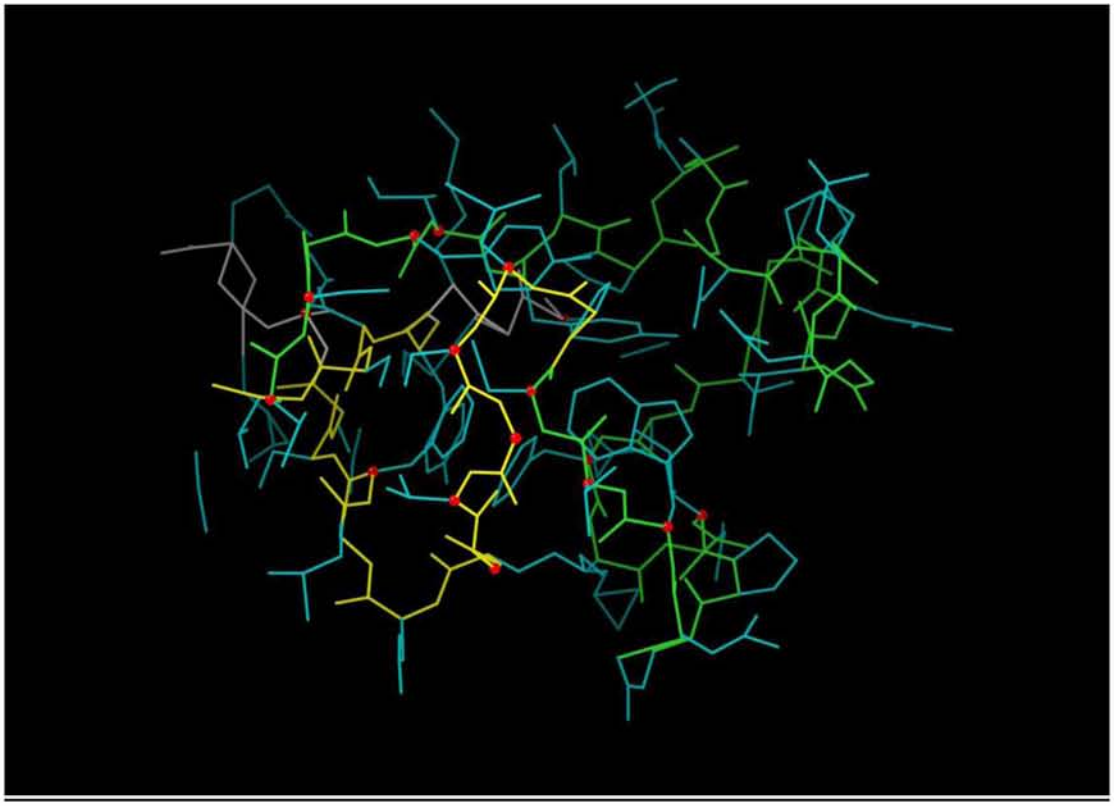


Σχήμα 3.3 *Κύρια αλυσίδα της πρωτεΐνης 1ag2. Με εναλλαγές του πράσινου φαίνονται οι MCF περιοχές, με κόκκινο χρώμα τα MIR και με αποχρώσεις του γκρι οι υπόλοιπες περιοχές της κύριας αλυσίδας.*

Πρωτεΐνη 1ag2 – TEF – MCF – Κοινές Περιοχές

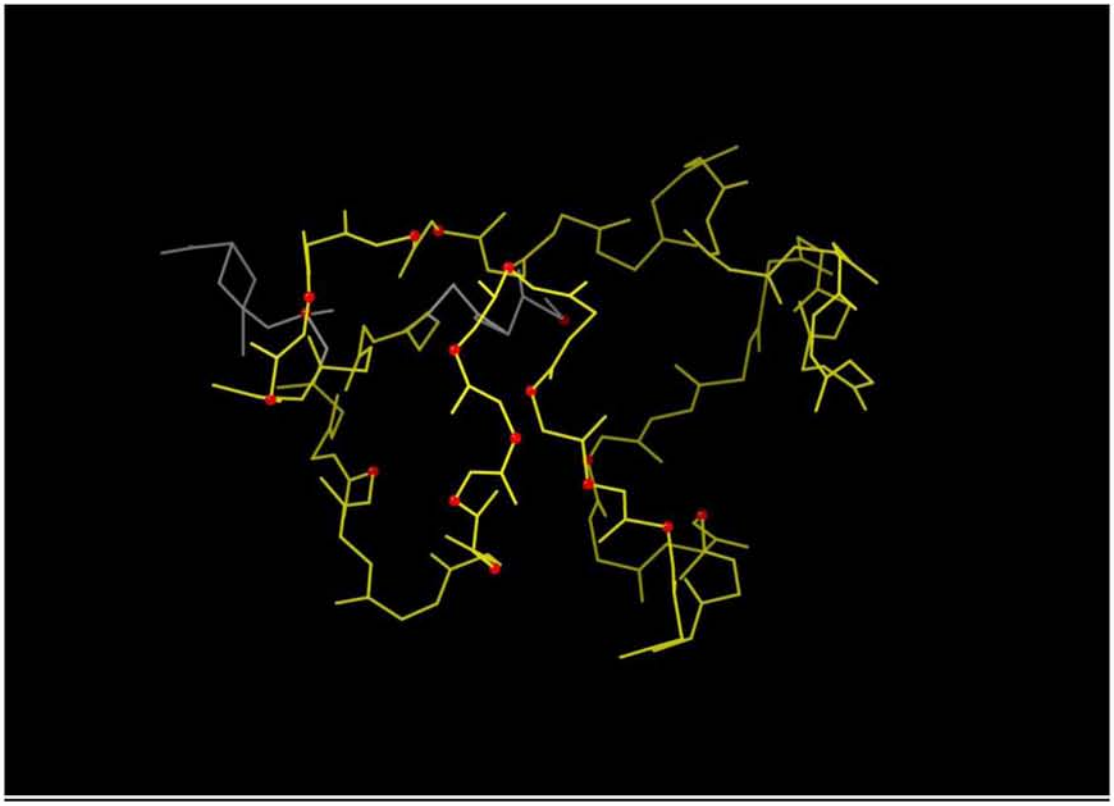
Σχήμα 3.4 *Κύρια αλυσίδα της πρωτεΐνης 1ag2. Με εναλλαγές του κίτρινου φαίνονται οι TEF περιοχές, με κόκκινο χρώμα τα MIR, με καφέ οι περιοχές που είναι τόσο TEF όσο και MCF (δηλαδή προβλέφθηκαν σωστά) και με αποχρώσεις του γκρι οι υπόλοιπες περιοχές της κύριας αλυσίδας. Στην συγκεκριμένη πρωτεΐνη δεν έχουμε πράσινο χρώμα γιατί όλες οι MCF περιοχές ήταν και TEF.*

Πρωτεΐνη 4rxn



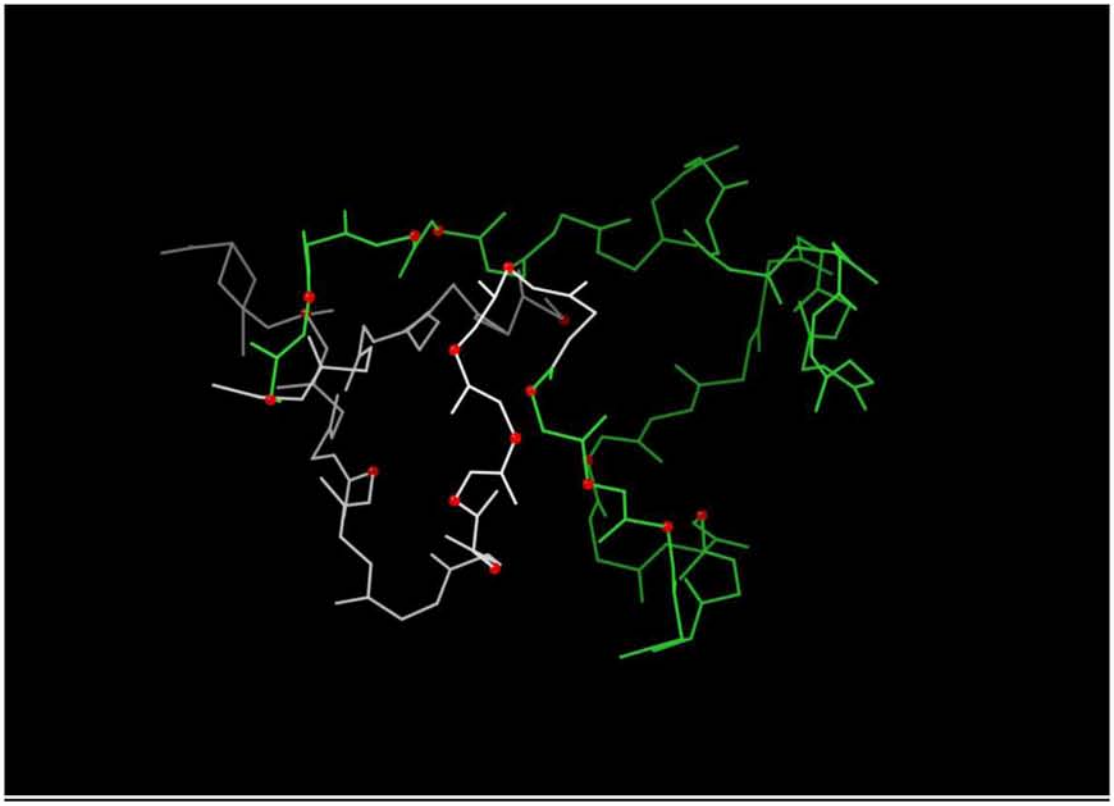
Σχήμα 3.5 *Κύρια και πλευρικές αλυσίδες της πρωτεΐνης 4rxn. Με εναλλαγές του πράσινου χρώματος φαίνονται οι προβλεπόμενες MCF περιοχές, με εναλλαγές του κίτρινου φαίνονται οι TEF περιοχές, με κόκκινο χρώμα τα MIR, με κυανό οι πλευρικές αλυσίδες και με αποχρώσεις του γκρι οι υπόλοιπες περιοχές της κύριας αλυσίδας.*

Πρωτεΐνη 4rxn – TEF



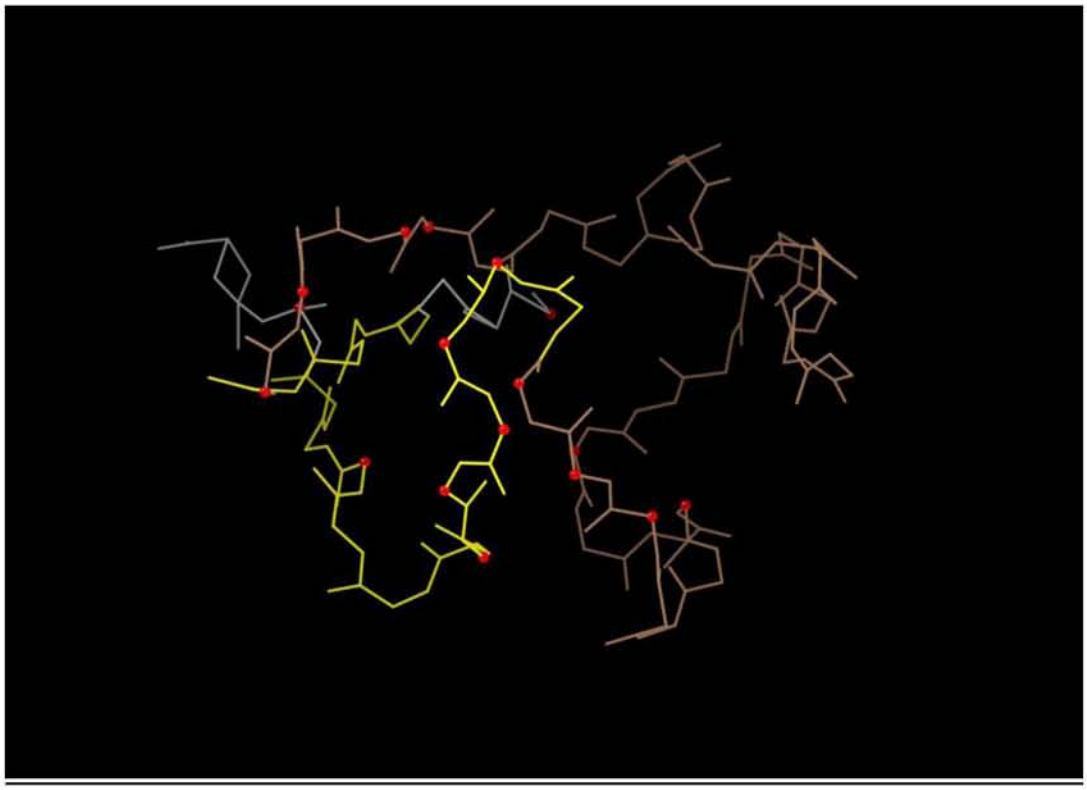
Σχήμα 3.6 *Κύρια αλυσίδα της πρωτεΐνης 4rxn. Με εναλλαγές του κίτρινου φαίνονται οι TEF περιοχές, με κόκκινο χρώμα τα MIR και με αποχρώσεις του γκρι οι υπόλοιπες περιοχές της κύριας αλυσίδας.*

Πρωτεΐνη 4rxn – MCF



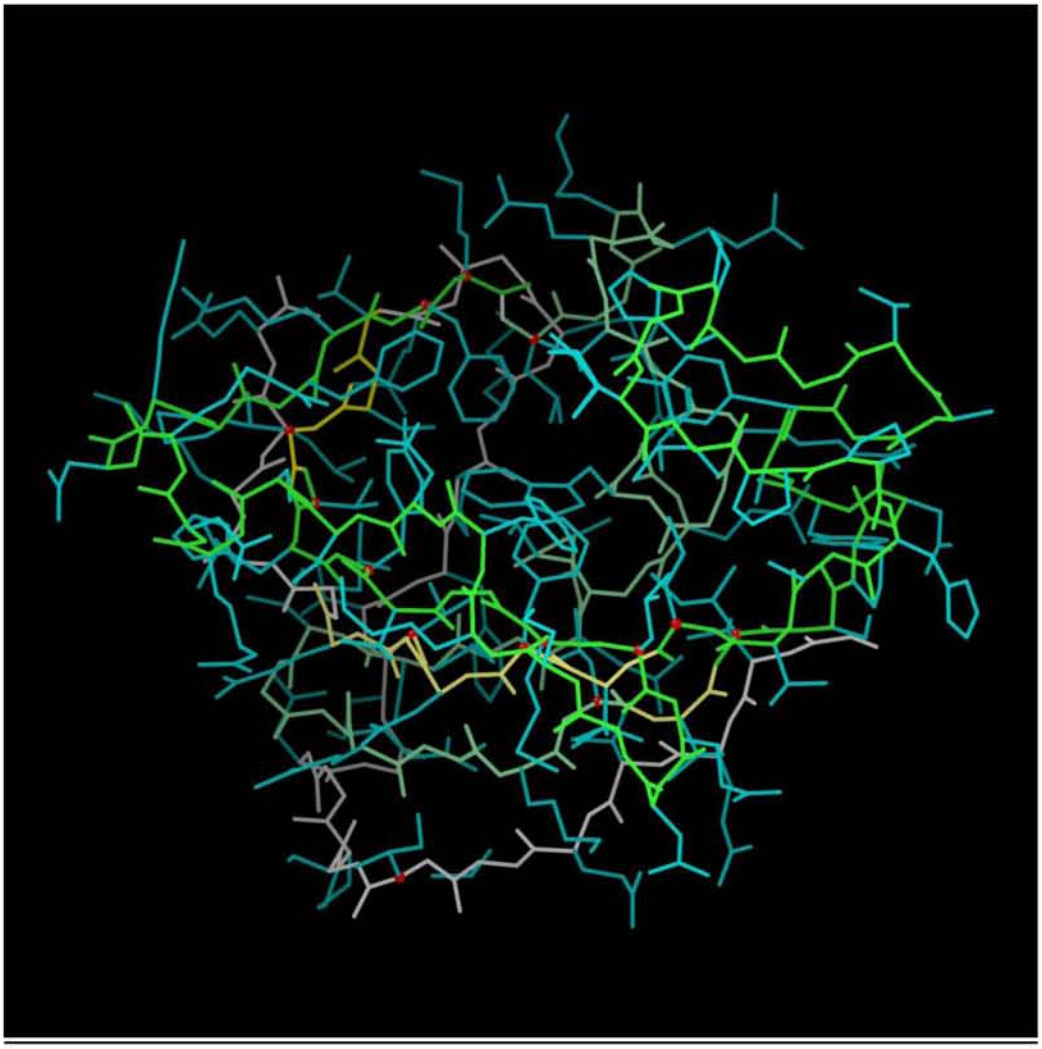
Σχήμα 3.7 *Κύρια αλυσίδα της πρωτεΐνης 4rxn. Με εναλλαγές του πράσινου φαίνονται οι MCF περιοχές, με κόκκινο χρώμα τα MIR και με αποχρώσεις του γκρι οι υπόλοιπες περιοχές της κύριας αλυσίδας.*

Πρωτεΐνη 4rxn – TEF – MCF – Κοινές Περιοχές



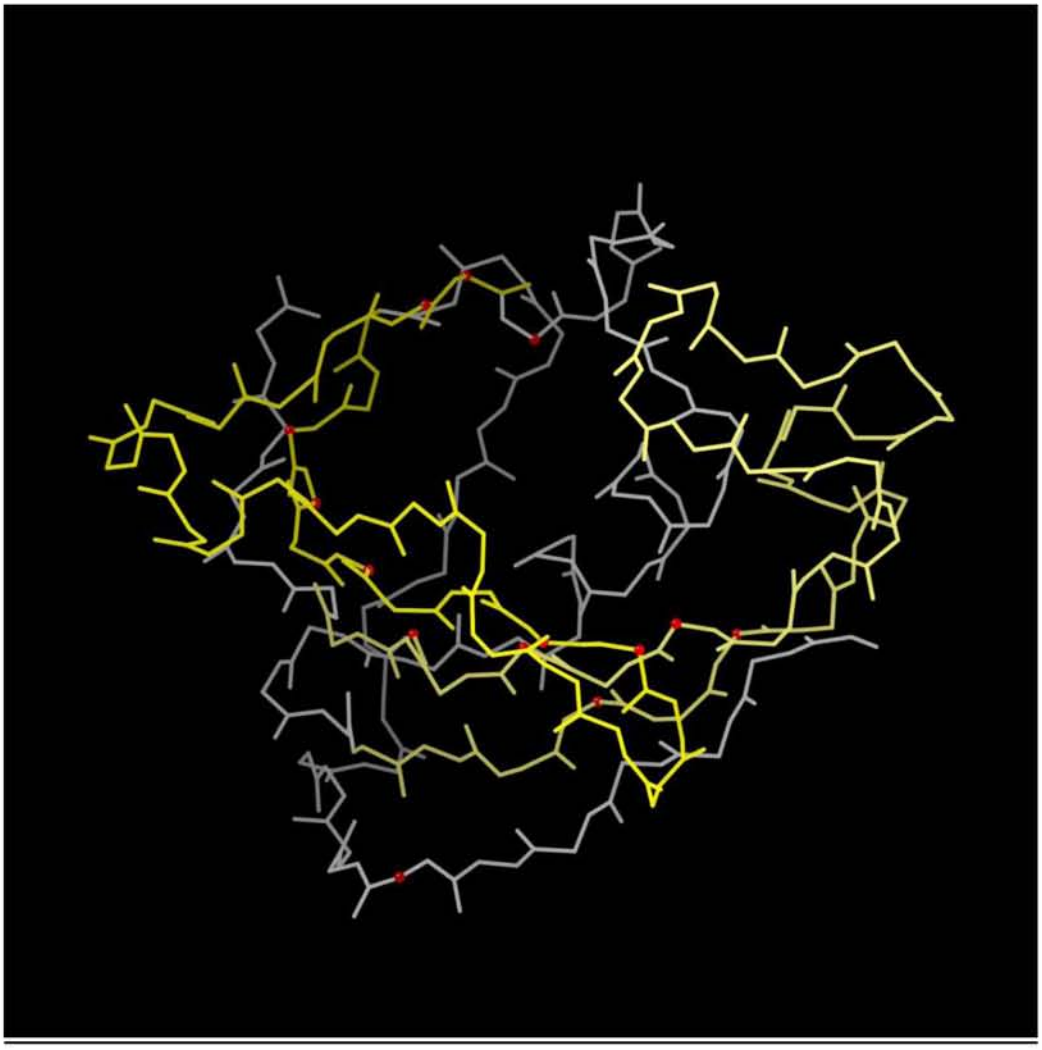
Σχήμα 3.8 *Κύρια αλυσίδα της πρωτεΐνης 4rxn. Με εναλλαγές του κίτρινου φαίνονται οι TEF περιοχές, με κόκκινο χρώμα τα MIR, με καφέ οι περιοχές που είναι τόσο TEF όσο και MCF (δηλαδή προβλέφθηκαν σωστά) και με αποχρώσεις του γκρι οι υπόλοιπες περιοχές της κύριας αλυσίδας. Στην συγκεκριμένη πρωτεΐνη δεν έχουμε πράσινο χρώμα γιατί όλες οι MCF περιοχές ήταν και TEF.*

Πρωτεΐνη 1fkb



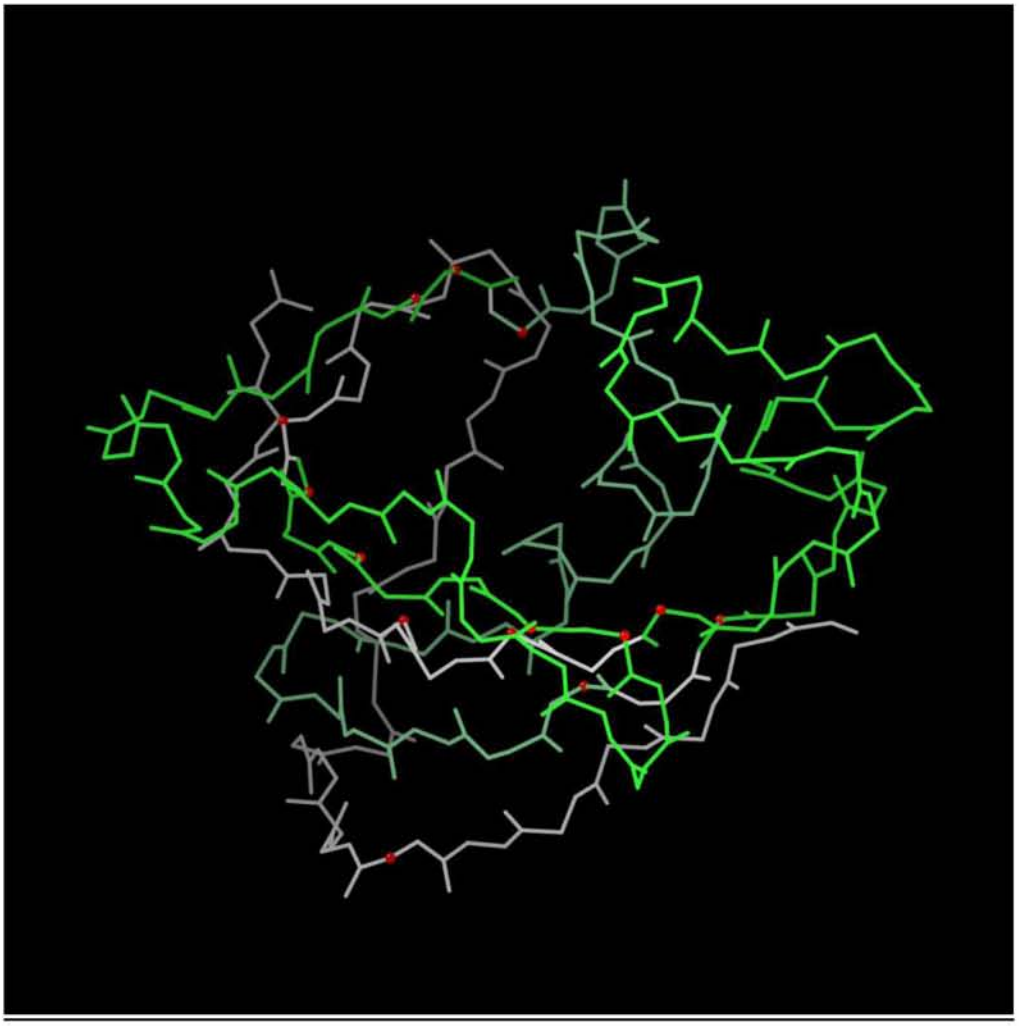
Σχήμα 3.9 *Κύρια και πλευρικές αλυσίδες της πρωτεΐνης 1fkb. Με εναλλαγές του πράσινου χρώματος φαίνονται οι προβλεπόμενες MCF περιοχές, με εναλλαγές του κίτρινου φαίνονται οι TEF περιοχές, με κόκκινο χρώμα τα MIR, με κυανό οι πλευρικές αλυσίδες και με αποχρώσεις του γκρι οι υπόλοιπες περιοχές της κύριας αλυσίδας.*

Πρωτεΐνη 1fkb – TEF



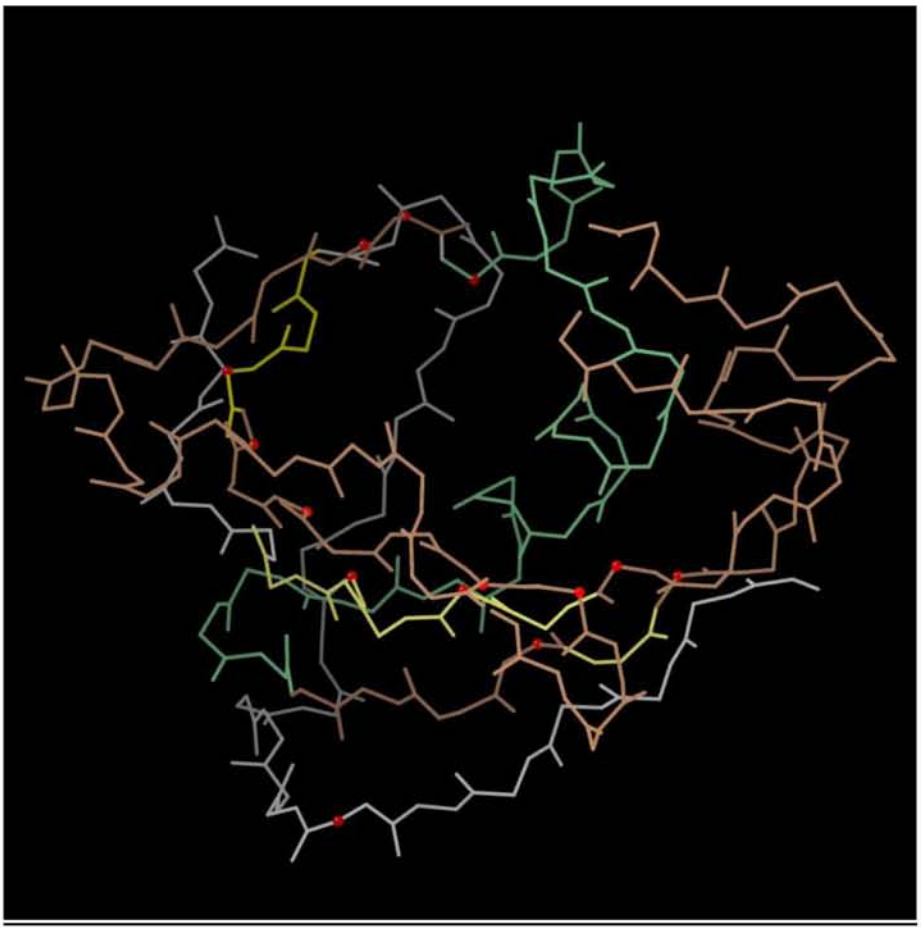
Σχήμα 3.10 *Κύρια αλυσίδα της πρωτεΐνης 1fkb. Με εναλλαγές του κίτρινου φαίνονται οι TEF περιοχές, με κόκκινο χρώμα τα MIR και με αποχρώσεις του γκρι οι υπόλοιπες περιοχές της κύριας αλυσίδας.*

Πρωτεΐνη 1fkb – MCF



Σχήμα 3.11 *Κύρια αλυσίδα της πρωτεΐνης 1fkb. Με εναλλαγές του πράσινου φαίνονται οι MCF περιοχές, με κόκκινο χρώμα τα MIR και με αποχρώσεις του γκρι οι υπόλοιπες περιοχές της κύριας αλυσίδας.*

Πρωτεΐνη 1fkb – TEF – MCF – Κοινές Περιοχές



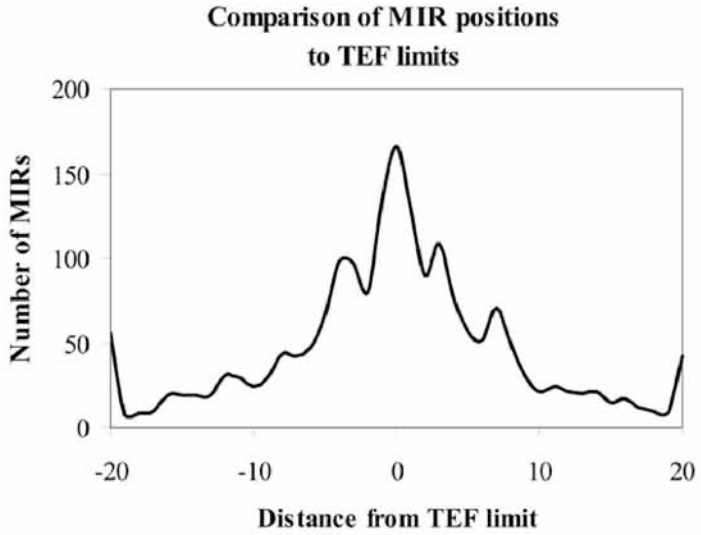
Σχήμα 3.12 *Κύρια αλυσίδα της πρωτεΐνης 1fkb. Με εναλλαγές του κίτρινου φαίνονται οι TEF περιοχές, με εναλλαγές του πράσινου οι MCF, με κόκκινο χρώμα τα MIR, με καφέ οι περιοχές που είναι τόσο TEF όσο και MCF (δηλαδή προβλέφθηκαν σωστά) και με αποχρώσεις του γκρι οι υπόλοιπες περιοχές της κύριας αλυσίδας.*

Κεφάλαιο 4: Συζήτηση

4.1 Σκοπός της εργασίας.

Σκοπός της παρούσας εργασίας είναι να βρούμε τμήματα (fragments) της ακολουθίας, που είναι σημαντικά για το δίπλωμα της πρωτεΐνης. Η εργασία αυτή βασίστηκε σε μια προηγούμενη έρευνα και συγκεκριμένα σε μια τεχνική προσομοίωσης Monte-Carlo που, με μόνη πληροφορία την αμινοξική ακολουθία μίας πρωτεΐνης, μπορεί να προβλέψει τα (υδρόφοβα κατά κανόνα) αμινοξέα που έχουν κρίσιμη σημασία για το σχηματισμό του πυρήνα που σταθεροποιεί την τριτοταγή δομή. Ο αλγόριθμος αυτός, που αναπτύχθηκε στο Εργαστήριο Γενετικής του Γεωπονικού Πανεπιστημίου Αθηνών υπολογίζει τα αμινοξέα που έχουν την τάση να σχηματίσουν τον πυρήνα της πρωτεϊνικής δομής, τα λεγόμενα MIR.

Η σύγκριση της προσομοίωσης με αποτελέσματα από την ανάλυση γνωστών δομών 100 περίπου πρωτεϊνών από τη βάση πρωτεϊνικών δομών PDB (Protein Data Bank) έδωσε πολύ ενθαρρυντικά αποτελέσματα. Η στατιστική ανάλυση της σχετικής θέσης των MIR ως προς τα άκρα των βρόχων και τις τοποϋδρόφοβες θέσεις σε ένα σύνολο 100 πρωτεϊνών διαφόρων οικογενειών έφερε τα ακόλουθα αποτελέσματα [25]: τα 2/3 των MIR βρίσκονται σε απόσταση το πολύ ±5 θέσεων από το πλησιέστερο άκρο βρόχου ή αντίστοιχα από την πλησιέστερη τοποϋδρόφοβη θέση. Συγκεκριμένα, το 63% των MIR βρέθηκε σε απόσταση το πολύ ±5 θέσεων από μια τοποϋδρόφοβη θέση, ενώ το 57% των MIR βρέθηκε σε απόσταση το πολύ ±5 θέσεων από άκρο TEF. Επίσης, από το διάγραμμα φαίνονται δύο δευτερεύοντα μέγιστα στις θέσεις ±3. Το αποτέλεσμα αυτό επιτρέπει μία πρώτη προσέγγιση στην πρόβλεψη των κρίσιμων αμινοξέων του πρωτεϊνικού πυρήνα από την αμινοξική ακολουθία.



Σχήμα 4.1 *Η σχέση μεταξύ των MIR και των άκρων TEF. Στον οριζόντιο άξονα συμβολίζεται η απόσταση από άκρο TEF και στον κάθετο το πλήθος των MIR. Όπως βλέπουμε, η πλειονότητα των MIR βρίσκεται σε απόσταση* το *πολύ ±5 θέσεων από το πλησιέστερο άκρο βρόχου.*

Η γνώση των συγκεκριμένων αμινοξέων είναι ιδιαίτερα χρήσιμη για την πρόβλεψη της δομής. Ο αλλεπάλληλος σχηματισμός του πρωτεϊνικού σφαιρικού πυρήνα, ο οποίος κυρίως αποτελείται από αυτά τα κατάλοιπα, οδηγεί σε μια εκπληκτική βελτιστοποίηση της διαδικασίας αναδίπλωσης, μειώνοντας το χώρο των πιθανών στερεοδιατάξεων που πρέπει να εξερευνηθεί. Άρα, με την ανάδειξη αυτών των «κρίσιμων» αμινοξέων έγινε ένα πρώτο βήμα για την πρόβλεψη της τελικής τρισδιάστατης αναδιπλωμένης στερεοδιάταξης των σφαιρικών πρωτεϊνών. Η υπόθεση που έγινε στην έρευνα ήταν ότι προκειμένου να επιτευχθεί γρήγορη αναδίπλωση, τα «κρίσιμα» αυτά αμινοξέα θα πρέπει να έχουν την τάση να αλληλεπιδρούν το ένα με το άλλο και έτσι ουσιαστικά δημιουργούν τις απαρχές του υδρόφοβου πυρήνα.. Τα αποτελέσματα που παραθέσαμε παραπάνω καθώς και το σχήμα 4.1 επιβεβαίωσαν την υπόθεση αυτή καθώς τα «βυθισμένα» MIR χρησιμεύουν ως «άγκυρες» για τον σχηματισμό των κλειστών βρόχων TEF [25].

Η παρούσα εργασία χρησιμοποιεί όλη την πληροφορία από την έρευνα αυτή και επιδιώκει να κάνει ουσιαστικά το επόμενο βήμα για την πρόβλεψη των δομικών στοιχείων των σφαιρικών πρωτεϊνών, προχωράει δηλαδή την προηγούμενη έρευνα στο επόμενο επίπεδο. Με άλλα λόγια, περνάμε από τις «κρίσιμες» θέσεις στα «κρίσιμα» τμήματα της αμινοξικής ακολουθίας. Αυτό γίνεται ως εξής: Ο σκοπός και η λογική της εργασίας είναι να παρέχει ένα «φίλτρο» για τα αρχικά MIR ώστε να βρεθούν τα πιο «κρίσιμα» και αυτό γιατί το πλήθος των MIR είναι πάρα πολύ μεγαλύτερο από το πλήθος των άκρων TEF και συνεπώς σκοπός είναι να βρεθούν τα πιο «σημαντικά» MIR. Αυτό γίνεται με τη βοήθεια της θεωρίας των μεταλλάξεων την οποία «δανειστήκαμε» από την πειραματική βιολογία και την χρησιμοποιούμε στις προσομοιώσεις.

Τα MIR είναι εν γένει σαφώς υδρόφοβα [25]. Στην εργασία αυτή έγιναν μεταλλάξεις στις θέσεις των MIR, δηλαδή σε κάθε πρωτεΐνη, παρήχθησαν τόσες μεταλλαγμένες ακολουθίες, όσες και τα MIR. Κάθε μία διέφερε από την αρχική και σε μία θέση MIR, όπου το εκάστοτε κατάλοιπο είχε αντικατασταθεί από μία αλανίνη (Ala, A). Αν κάποιο MIR ήταν A (Ala), τότε αντικαθίστατο από την γλυκίνη G (Gly). Σε αυτές τις μεταλλαγμένες ακολουθίες υπολογίζονταν ξανά τα MIR.

Αυτές οι μεταλλάξεις βασίζονται στην υπόθεση πως αν κάποιο MIR είναι σημαντικό και το αντικαταστήσουμε με την αλανίνη ή την γλυκίνη που είναι τα πιο απλά αμινοξέα που υπάρχουν στη φύση, τότε θα «χαλάσει» το δίπλωμα της πρωτεΐνης. Αν δε «χαλάσει» το δίπλωμα, τότε σημαίνει πως το MIR τελικά δεν ήταν όντως τόσο σημαντικό για τον σχηματισμό της δομής της δεδομένης πρωτεΐνης και ενδεχομένως να είναι και εκτός πυρήνα. Το αν «χαλάει» ή όχι το δίπλωμα φαίνεται από το αν εξαφανίζονται ή όχι τα γειτονικά MIR στο χώρο. Η μετάλλαξη ενός MIR σε Ala μειώνει εν γένει την αλληλεπίδρασή του με τα άλλα αμινοξέα, άρα και με τα άλλα MIR. Άρα, ένα MIR μεταλλαγμένο σε Ala, αναμένουμε να συσχετίζεται με τα εξαφανιζόμενα MIR. Συγκεκριμένα, για κάθε μεταλλαγμένη ακολουθία εξετάσαμε αν υπάρχουν εξαφανιζόμενα MIR τα οποία βρίσκονται σε απόσταση μεγαλύτερη των 10 και μικρότερη των 35 θέσεων από την αμινοξική θέση στην οποία έγινε η μετάλλαξη.

Αν ναι τότε το MIR στο οποίο έγινε η μετάλλαξη θεωρείται κατά την υπόθεση μας «κρίσιμο» και πως συσχετίζεται με τα οριζόμενα ως άνωθεν, γειτονικά MIR. Με τον τρόπο αυτό παρέχεται ένα «φίλτρο» το οποίο καθορίζει ποιά από τα αρχικά MIR (της μη μεταλλαγμένης ακολουθίας δηλαδή) είναι όντως σημαντικά για την δομή. Εκτός αυτού, με τον τρόπο που περιγράψαμε παραπάνω, βρίσκουμε και τα ζεύγη τα οποία συσχετίζονται σύμφωνα με την υπόθεση μας και δεν είναι άλλα από το μεταλλαγμένο MIR και το εξαφανιζόμενο MIR που βρίσκεται μέσα στην εμβέλεια 10 – 35. Έτσι, προκύπτουν ζεύγη που συσχετίζονται και τα οποία με τον τρόπο αυτό οριοθετούν την περιοχή που βρίσκεται ανάμεσα τους και η οποία δεν είναι άλλη από την MCF, την πρόβλεψη μας δηλαδή για τους κλειστούς βρόχους TEF που απ ότι είδαμε από την προηγούμενη έρευνα, παίζουν σημαντικό ρόλο στην αναδίπλωση και τη δομή των σφαιρικών πρωτεϊνών.

Πέρα όμως από την πρόβλεψη των βασικότερων δομικών στοιχείων των σφαιρικών πρωτεϊνών η έρευνα αυτή μπορεί επίσης να αξιοποιηθεί για την πρόβλεψη της επίδρασης μεταλλάξεων στο σχηματισμό του δομικού πυρήνα των πρωτεϊνών, συμβάλλοντας στη μελέτη και κατανόηση των μηχανισμών της αναδίπλωσης. Έχει βρεθεί πειραματικά πως υπάρχουν περιπτώσεις συγκεκριμένων μεταλλάξεων που για κάποιες πρωτεΐνες είτε καταστρέφουν την αναδίπλωση (περίπτωση της Drosophila melanogaster Engrailed Homeodomain που θα δούμε παρακάτω) είτε σχηματίζουν αμυλοειδή που σχετίζονται με τον διαβήτη τύπου 2 (περίπτωση των transthyretin και a-synuclein). Οι συνέπειες των μεταλλάξεων στις πρωτεΐνες είναι πολύ δύσκολο να προβλεφθούν. Πολυάριθμες μεταλλάξεις δεν έχουν καμία επίδραση στη βιολογική δραστηριότητα, ενώ άλλες μεταλλάξεις που επηρεάζουν αμινοξέα μακριά από τις ενεργές περιοχές της πρωτεΐνης μπορούν να καταργήσουν εντελώς την λειτουργικότητα τους. Ένα παρόμοιο πρόβλημα προκύπτει με μεταλλαγμένες πρωτεΐνες που δε μπορούν να αναδιπλωθούν [30]. Η γνώση λοιπόν της επίδρασης των μεταλλάξεων μας βοηθά στην καλύτερη κατανόηση του μηχανισμού αναδίπλωσης.

Πέραν αυτού όμως, η μελέτη της διαδικασίας αναδίπλωσης είναι εξαιρετικής σημασίας για την κατανόηση της προέλευσης ασθενειών όπως η νόσος Alzheimer, η κυστική ίνωση ,η νόσος των τρελών αγελάδων, η νόσος Creutzfeldt-Jakob και η νόσος Parkinson που σχετίζονται με μη σωστή πρωτεϊνική αναδίπλωση. Αυτές οι εκφυλιστικές ασθένειες συνδέονται με τον πολυμερισμό των μη σωστά αναδιπλωμένων πρωτεϊνών σε αδιάλυτα, εξωκυτταρικά ενιαία τμήματα και/ή ενδοκυτταρικά τμήματα που περιλαμβάνουν β – πτυχωτές επιφάνειες και αμυλοειδείς ίνες. Δεν είναι ξεκάθαρο αν τα ενιαία αυτά τμήματα είναι η αιτία ή απλά μια «αντανάκλαση» της απώλειας της πρωτεϊνικής ομοιόστασης. Η μη σωστή πρωτεϊνική αναδίπλωση μπορεί να οδηγήσει σε ένα πλήθος πρωτοπαθών ασθενειών όπως το εμφύσημα που συνδέεται με την αντιθρυψίνη ,και την κυστική ίνωση όπου η απώλεια της λειτουργίας της πρωτεΐνης είναι η αιτία της δυσλειτουργίας .

4.2 Η περίπτωση της πρωτεΐνης En – HD.

Η πρωτεΐνη Drosophila melanogaster Engrailed Homeodomain (En - HD) είναι μια πρωτεΐνη 60 καταλοίπων με τρία ελικοειδή τμήματα : 10 – 22 (H1), 28 – 37 (H2) και 42 – 56(H3). Αναδιπλώνεται σε χρονική κλίμακα μικροδευτερολέπτων μέσω αναδιπλούμενων ενδιάμεσων καταστάσεων και το μονοπάτι που ακολουθεί για την αναδίπλωση έχει μελετηθεί καλά μέσω πειραμάτων και προσομοιώσεων. Κατασκευάστηκε μια αποδιατεταγμένη (denatured) μορφή υπό συνθήκες που ευνοούν την αναδίπλωση, εισάγοντας την μετάλλαξη L16A. Το L16 είναι ένα υψηλά συντηρημένο, πλήρως βυθισμένο κατάλοιπο που βρίσκεται στο μέσον του τμήματος H1. Σε γενικές γραμμές, η μετάλλαξη L16A, απομακρύνει ένα πλήθος τοπικών αλληλεπιδράσεων καθώς και πολυάριθμες μακρινές αλληλεπιδράσεις με κατάλοιπα που βρίσκονται στα τμήματα H2 και κυρίως H3. Διάφορες αναλύσεις αλλά και η προσομοίωση που έγιναν σε παλιότερη έρευνα, έδειξαν πως ο πυρήνας διαταράσσεται σε αυτή την περιοχή αλλά η ελικοειδής δευτεροταγής δομή των H1, H2 και H3 παραμένει άθικτη. Η μετάλλαξη δρα ώστε να ανατρέψει την ενεργειακή ισορροπία της πρωτεΐνης ώστε η L16A να είναι αποδιατεταγμένη υπό φυσιολογική ιονική ισχύ [50].

Στη συνέχεια θα αναφέρουμε λίγα λόγια για τα χαρακτηριστικά της μετάλλαξης L16A για την προκειμένη περίπτωση. Η μετάλλαξη αυτή αποσταθεροποιεί τις πρωτεΐνες κατά 4–5 kcal mol^{-1}. Για μια πρωτεΐνη όπως η En-HD που έχει ελεύθερη ενέργεια αναδίπλωσης μόλις 2.5 kcal mol^{-1} μια τέτοια αλλαγή της τάξης του 4–5 kcal mol^{-1} αλλάζει έτσι την ισορροπία ώστε η πρωτεΐνη να αποδιατάσσεται υπό φυσιολογικές συνθήκες.

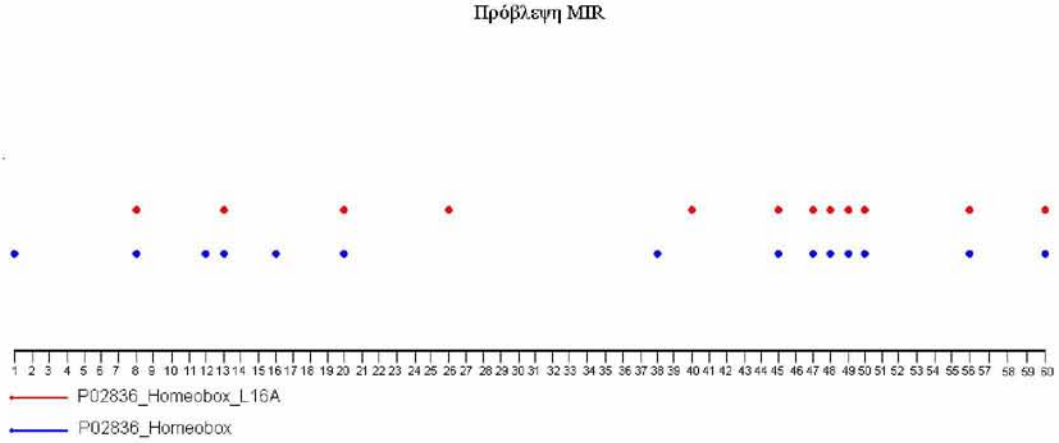
Η μετάλλαξη της λευκίνης στη θέση 16 σε αλανίνη αποσταθεροποίησε την En-HD σε σημείο που η μεταλλαγμένη πρωτεΐνη ήταν αποδιατεταγμένη υπό όλες τις συνθήκες. Η μεταλλαγμένη πρωτεΐνη ήταν ένα καλό μοντέλο για την αποδιατεταγμένη κατάσταση σε πλήρως φυσιολογικές συνθήκες. Τα κατάλοιπα 50 – 55 εμφανίστηκαν πιο ευπροσάρμοστα απ ότι το υπόλοιπο της αλυσίδας στην L16A αλλά οι δευτεροταγείς χημικές μεταβολές τους ήταν σημαντικές, υποδεικνύοντας κάποιο ελικοειδές περιεχόμενο. Αυτό σε συνδυασμό με άλλες παρατηρήσεις απέδειξε ότι οι ενδιάμεσες καταστάσεις αναδίπλωσης της αρχικής πρωτεΐνης δεν ήταν μόνιμα ελικοειδείς σε αυτές τις θέσεις. Αυτό το γεγονός ενίσχυσε την εγκυρότητα της L16A ως μοντέλο για την αποδιατεταγμένη ή ενδιάμεση κατάσταση της αρχικής πρωτεΐνης [51].

Πρώιμες μελέτες σχετικά με την πρωτεϊνική αναδίπλωση ισχυρίζονταν ότι οι ενδιάμεσες καταστάσεις ενδεχομένως να ήταν απαραίτητες για το παραγωγικό δίπλωμα. Η ύπαρξη καλά ορισμένων μονοπατιών με υποχρεωτικές ενδιάμεσες καταστάσεις είχε προταθεί για την επίλυση του παραδόξου του Levinthal. Αργότερα όμως διαπιστώθηκε πως μικρές πρωτεΐνες μπορούσαν να αναδιπλωθούν χωρίς να περάσουν από σταθερές ενδιάμεσες καταστάσεις και μάλιστα σε εξαιρετικά σύντομο χρονικό διάστημα [51].

Το σημείο εκκίνησης για την αναδίπλωση, μόλις η πολυπεπτιδική αλυσίδα ισορροπήσει σε νέες συνθήκες που προκαλούν αποδιάταξη, είναι άγνωστο. Αποδιατεταγμένες καταστάσεις που παράγονται μέσω μεταλλάξεων, όπως η περίπτωση της L16A, μπορεί να είναι η καλύτερη προσέγγιση για να χαρακτηρισθεί, η πολυπεπτιδική αλυσίδα υπό φυσιολογικές συνθήκες και εκεί έγκειται και η χρησιμότητα της μετάλλαξης [51]. Στην κατανόηση δηλαδή του μηχανισμού της αναδίπλωσης.

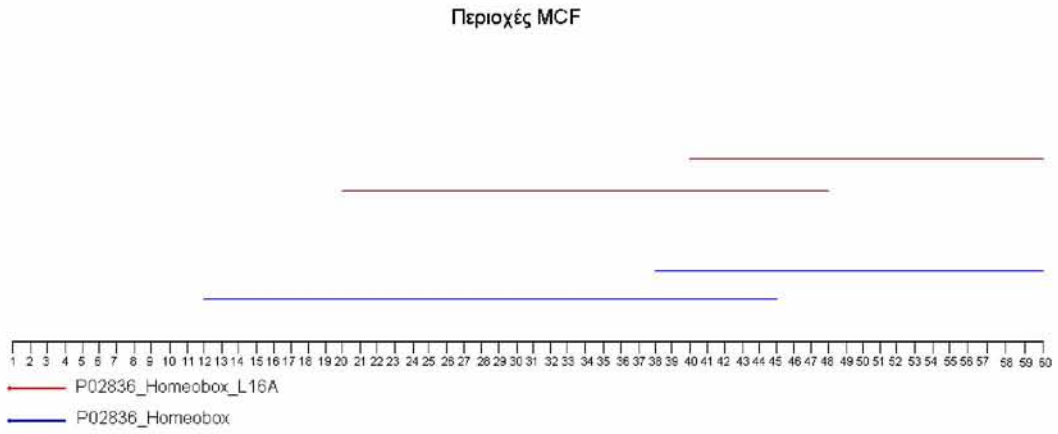
Αυτό που κάναμε στη δική μας εργασία ήταν να κάνουμε μια μικρή εφαρμογή των MIR και MCF σε περιπτώσεις πρωτεϊνών που έχει δειχθεί ότι συγκεκριμένες μεταλλάξεις καταστρέφουν το δίπλωμα, όπως συμβαίνει με την En – HD. Συγκεκριμένα, θέλουμε να δούμε αν η μετάλλαξη βρίσκεται κοντά σε MIR, MCF, MCF limits και γενικά κατά πόσο αλλάζουν τα MIR και MCF μετά την εισαγωγή της μετάλλαξης L16A.

Στο σημείο αυτό, να τονίσουμε πως ο κωδικός pdb της En – HD είναι ο 1enh και ο κωδικός swiss-prot είναι PO2836_Homeobox ενώ ο swiss-prot κωδικός της μεταλλαγμένης πρωτεΐνης είναι PO2836_Homeobox_L16A. Εμείς χρησιμοποιήσαμε την ακολουθία της πρωτεΐνης PO2836_Homeobox και μέσω της προσομοίωσης Monte-Carlo, υπολογίσαμε τα MIR τόσο για την αρχική όσο και για τις μεταλλαγμένες ακολουθίες. Στη συνέχεια, μέσω του αλγορίθμου παραγωγής των MCF περιοχών, υπολογίσαμε τις περιοχές MCF. Το ίδιο κάναμε και για την μεταλλαγμένη πρωτεΐνη PO2836_Homeobox_L16A η οποία όπως είναι λογικό, έχει την ίδια ακολουθία με την PO2836_Homeobox μόνο που στην θέση του καταλοίπου λευκίνη που βρίσκεται στη θέση 16, εισάγαμε την αλανίνη. Στα διαγράμματα που ακολουθούν, μπορούμε να δούμε εποπτικά τις αλλαγές ανάμεσα στις δυο πρωτεΐνες, την αρχική δηλαδή και την μεταλλαγμένη, όσον αφορά τα MIR και τα MCF.



Σχήμα 4.2 *Οι θέσεις των MIR για τις πρωτεΐνες PO2836_Homeobox και PO2836_Homeobox_L16A. Στον οριζόντιο άξονα φαίνεται η αρίθμηση για τις 60 θέσεις των αμινοξέων, με μπλε κουκίδες τα MIR για την PO2836_Homeobox και με κόκκινες κουκίδες τα MIR για την PO2836_Homeobox_L16A.*

Η αρχική πρωτεΐνη εμφανίζει 14 MIR σε αντίθεση με την μεταλλαγμένη που εμφανίζει 12. Παρατηρούμε ότι το MIR της θέσης 1 εξαφανίζεται τελείως στην μεταλλαγμένη πρωτεΐνη όπως και τα MIR των θέσεων 12 και 16. Το MIR της θέσης 38 εμφανίζεται δύο θέσεις μακρύτερα στην μεταλλαγμένη πρωτεΐνη και συγκεκριμένα στη θέση 40. Δύο νέα MIR εμφανίζονται στην μεταλλαγμένη πρωτεΐνη σε θέσεις που δεν υπήρχαν στην αρχική πρωτεΐνη: στις θέσεις 26 και 40. Όλα τα υπόλοιπα MIR, 10 στο σύνολο, εμφανίζονται στις ίδιες θέσεις τόσο στην αρχική όσο και στη μεταλλαγμένη πρωτεΐνη.



Σχήμα 4.3 *Οι περιοχές MCF για τις πρωτεΐνες PO2836_Homeobox και PO2836_Homeobox_L16A. Στον οριζόντιο άξονα φαίνεται η αρίθμηση για τις 60 θέσεις των αμινοξέων, με μπλε ευθύγραμμα τμήματα οι MCF περιοχές για την PO2836_Homeobox και με κόκκινα ευθύγραμμα τμήματα οι MCF περιοχές για την PO2836_Homeobox_L16A.*

Και στις δυο πρωτεΐνες εμφανίζονται δυο MCF περιοχές όπου εκτείνονται για την αρχική πρωτεΐνη στις θέσεις από 12-45 και 38-60 και για την μεταλλαγμένη στις θέσεις 20-48 και 40-60. Για το πρώτο MCF η διαφορά στις θέσεις των άκρων τους ανάμεσα στις δυο πρωτεΐνες είναι 8 θέσεις για το ένα άκρο και 3 για το άλλο, ενώ για το δεύτερο MCF η διαφορά είναι 2 θέσεις για το ένα άκρο ενώ το δεύτερο άκρο συμπίπτει και στις δυο περιπτώσεις. Επίσης, τα MCF εμφανίζουν διαφορετικό μήκος στις δύο πρωτεΐνες. Το πρώτο και το δεύτερο MCF της αρχικής πρωτεΐνης εμφανίζουν μήκος 33 και 22 κατάλοιπα αντίστοιχα, ενώ στην μεταλλαγμένη πρωτεΐνη εμφανίζουν μικρότερο μήκος, 28 και 20 αντίστοιχα.

Παρατηρούμε πως εισάγοντας μια μόλις μετάλλαξη στην ακολουθία της En-HD αλλάζουν τα MIR, MCF και MCF limits στην πρωτεΐνη. Οι διαφορές στη δομή λόγω της μετάλλαξης αυτής εξάλλου αποδείχθηκε από προηγούμενες έρευνες [50,51] και αναφέρθηκε στην αρχή της παρούσας ενότητας.

4.3 Επίλογος

Τα αποτελέσματα της παρούσας εργασίας όσο και της εργασίας στην οποία βασίστηκε, είναι σχετικά καλά σε πολύ γενικές γραμμές όπως μπορεί να φανεί τόσο από τους πίνακες της ενότητας 3.1 όσο και από τους συγκεντρωτικούς πίνακες του παραρτήματος. Πρέπει να σημειωθεί όμως πως πλέον το ποσοστό των MIR (στην παρούσα εργασία τα άκρα MCF) που βρίσκονται σε απόσταση ±5 θέσεων από άκρο TEF έπεσε από το 57% της προηγούμενης εργασίας [25] στο 48%. Άρα, μπορούμε να ισχυριστούμε πως η παρούσα εργασία δε δείχνει να βελτιώνει τις προβλέψεις συγκριτικά με την προηγούμενη έρευνα.

Δοκιμάστηκε κάτι καινούριο και συγκεκριμένα η εισαγωγή της θεωρίας των μεταλλάξεων ως «φίλτρο» » για τα αρχικά MIR και ασφαλώς χρειάζεται και απαιτείται περεταίρω μελέτη ώστε η συγκεκριμένη μεθοδολογία να βελτιώσει τις προβλέψεις και γενικά να αποκτήσει μεγάλη αξία ως μέθοδο πρόβλεψης καθώς το θέμα το οποίο πραγματεύεται είναι πειραματικό και διαρκώς εξελισσόμενο σε διεθνή κλίμακα.

Η παρούσα εργασία σίγουρα μπορεί να βελτιωθεί περισσότερο προς αυτή την κατεύθυνση. Για παράδειγμα, μπορούν να βελτιωθούν οι αλγόριθμοι που χρησιμοποιούνται. Είτε ο αλγόριθμος υπολογισμού των MIR είτε ο αλγόριθμος παραγωγής των περιοχών MCF μπορούν να βελτιστοποιηθούν και ενδεχομένως να κάνουν χρήση επιπρόσθετων παραμέτρων και πληροφοριών ώστε να εμπλουτιστεί η μέθοδος πρόβλεψης. Επίσης, προκειμένου να γίνει καλύτερη αξιολόγηση της μεθόδου αυτής, προτείνεται να χρησιμοποιηθεί ακόμα μεγαλύτερο στατιστικό δείγμα, εφαρμόζοντας την τεχνική αυτή σε ακόμα μεγαλύτερο πλήθος πρωτεϊνών με ενίσχυση των κατηγοριών που δεν είχαν αρκετά μεγάλη εκπροσώπηση στην παρούσα εργασία (πχ πρωτεΐνες προερχόμενες από ιούς) και χρησιμοποιώντας ενδεχομένως περισσότερα και πιο βελτιωμένα στατιστικά μέτρα αξιολόγησης, κατάλληλα για την φύση του προβλήματος. Με τον τρόπο αυτό η επιστημονική κοινότητα θα έχει ακόμα περισσότερες πληροφορίες διαθέσιμες και θα μπορεί με μεγαλύτερη βεβαιότητα να ελέγξει την εγκυρότητα της συγκεκριμένης μεθόδου πρόβλεψης.

Η μελέτη του μηχανισμό της αναδίπλωσης και γενικά η κατανόηση της προέλευσης ασθενειών που σχετίζονται με την μη σωστή αναδίπλωση των πρωτεϊνών στο χώρο – με τα οποία ασχολείται η παρούσα εργασία – είναι σε γενικές γραμμές ένα ανοιχτό ερευνητικό θέμα. Η εργασία αυτή προσπάθησε να προχωρήσει από τις κρίσιμες θέσεις στα κρίσιμα τμήματα της πολυπεπτιδικής αλυσίδας των πρωτεϊνών, προσπάθησε να προχωρήσει δηλαδή ένα βήμα παρακάτω, παλαιότερες αντίστοιχες έρευνες. Παρόλα αυτά όμως, η προσπάθεια δε σταματάει εδώ αλλά αντιθέτως προτείνεται η συνεχής βελτίωση της μεθόδου αυτής ώστε να βελτιώσει τα αποτελέσματα της και να αποκτήσει κύρος και αξία ως μια αξιόπιστη μέθοδος πρόβλεψης. Σίγουρα απαιτείται λοιπόν, περισσότερη μελέτη και έρευνα προκειμένου να βελτιώσει τις προβλέψεις και ώστε να συντελεστεί ουσιαστική πρόοδος στον τομέα της κατανόησης του μηχανισμού της πρωτεϊνικής αναδίπλωσης εν γένει.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Βικιπαίδια Η Ελεύθερη Εγκυκλοπαίδεια http://el.wikipedia.org/wiki τίτλος άρθρου *«Πρωτεΐνη»*.

2. Wikipedia The Free Encyclopedia www.wikipedia.org τίτλος άρθρου *«Globular Protein»*.

3. Wikipedia The Free Encyclopedia www.wikipedia.org τίτλος άρθρου *«Primary Structure»*.

4. Wikipedia The Free Encyclopedia www.wikipedia.org τίτλος άρθρου *«Secondary Structure»*.

5. Wikipedia The Free Encyclopedia www.wikipedia.org τίτλος άρθρου *«Tertiary Structure»*.

6. Alberts, Bruce; Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walters (2002) *"The Shape and Structure of Proteins"* Molecular Biology of the Cell; Fourth Edition. New York and London: Garland Science.

7. Anfinsen C (1972) *"The formation and stabilization of protein structure"*. Biochem. J. **128** (4): 737–49.

8. van den Berg B, Wain R, Dobson CM, Ellis RJ (August 2000) *"Macromolecular crowding perturbs protein refolding kinetics: implications for folding inside the cell"*. Embo J. **19** (15): 3870–5.

9. Pace C, Shirley B, McNutt M, Gajiwala K (1996) *"Forces contributing to the conformational stability of proteins"*. Faseb J. **10** (1): 75–83.

10. Rose G, Fleming P, Banavar J, Maritan A (2006) *"A backbone-based theory of protein folding"*. Proc. Natl. Acad. Sci. U.S.A **103** (45): 16623–33.

11. Deechongkit S, Nguyen H, Dawson PE, Gruebele M, Kelly JW (2004) *"Context Dependent Contributions of Backbone H-Bonding to β-Sheet Folding Energetics"*. Nature **403** (45): 101–105.

12. Lee S, Tsai F (2005) *"Molecular chaperones in protein quality control"*. J. Biochem. Mol. Biol. **38** (3): 259–65.

13. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. (2007) *"The design and characterization of two proteins with 88% sequence identity but different structure and function"*. Proc Natl Acad Sci U S A. **104** (29): 11963–8.

14. International Union of Pure and Applied Chemistry *"hydrogen bond"* Compendium of Chemical Terminology Internet edition.

15. Κώστας Σούκουλης, Μαρία Καφεσάκη (2008) *Εισαγωγή στη σύγχρονη φυσική.*

16. International Union of Pure and Applied Chemistry (1994) *"van der Waals forces"* Compendium of Chemical Terminology Internet edition.

17. Integrated Publishing http://www.tpub.com

18. Shortle D (1996) *"The denatured state (the other half of the folding equation) and its role in protein stability".* Faseb J **10** (1): 27–34.

19. Zhang Y (2008) *"Progress and challenges in protein structure prediction".* Curr Opin Struct Biol **18** (3): 342-348.

20. Bowie JU, Luthy R, Eisenberg D (1991) *"A method to identify protein sequences that fold into a known three-dimensional structure".* Science **253** (5016): 164–170.

21. Wikipedia The Free Encyclopedia www.wikipedia.org τίτλος άρθρου *«De novo Protein Structure Prediction».*

22. Anfinsen CB (1973) *"Principles that govern the folding of protein chains".* Science **181** (96): 223-230.

23. Igor N. Berezovsky, Alexander Y. Grosberg, Edward N. Trifonov (2000) *Closed loops of nearly standard size: common basic elements of protein structure.* FEBS Lett. **466**, 283-286.

24. Igor N. Berezovsky, Valery M. Kirzhner, Alla Kirzhner, Edward N. Trifonov (2001) *Protein Folding: Looping From Hydrophobic Nuclei.* Proteins **45**, 346-350.

25. Nikolaos Papandreou, Igor N. Berezovsky, Anne Lopes, Elias Eliopoulos, Jacques Chomilier (2004) *Universal positions in globular proteins.* European Journal of Biochemistry **271**, 4762-4768.

26. C. Levinthal (1968) *"Are there pathways for protein folding?"* Journal de Chimie Physique et de Physico-Chimie Biologique **65**: 44–45.

27. Igor N. Berezovsky, Edward N. Trifonov (2002) *Loop fold structure of proteins: resolution of Levinthal's paradox.* J. Biomol. Struct. Dynamics **20**, 5-6.

28. Anne Poupon, Jean – Paul Mornon (1998) *Populations of hydrophobic amino acids within protein globular domains; identification of conserved "topohydrophobic" positions.* Proteins **33**, 329-342.

29. Anne Poupon, Jean – Paul Mornon (1999) *"Topohydrophobic positions" as key markers of globular protein folds.* Theoret. Chem. Accounts **101**, 2-8.

30. Anne Poupon, Jean – Paul Mornon (1999) *Predicting the protein folding nucleus from sequence.* FEBS Lett. **452**, 283-289.

31. Marc Lamarine, Jean – Paul Mornon, Igor N. Berezovsky, Jacques Chomilier (2001) *Distribution of tightened end fragments of globular proteins statistically match that of topohydrophobic positions: towards an efficient punctuation of protein folding?* Cell. Moll. Life Sci. **58**, 492-498.

32. PFF Database http://www.bioinf.manchester.ac.uk/corpas/db/help.html

33. Manuel Corpas, James Sinnott, Dave Thorne, Steve Pettifer, Terri Attwood and the PFF consortium (2007) *PFF – an integrated database of residues and fragments critical for protein folding.* BMC Systems Biology **1**, 48.

34. Helen M.Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov and Philip E. Bourne (2000) *The Protein Data Bank.* Nucleic Acids Res. **28**, 235-242.

35. Berman, H. M. (January 2008) *"The Protein Data Bank: a historical perspective".* Acta Crystallographica Section A: Foundations of Crystallography **A64** (1): 88-95.

36. *PDB Archive Contains More Than 50000 Structures* http://www.rcsb.org

37. Westbrook, J.; et. al. (2005) *"PDBML: the representation of archival macromolecular structure data in XML".* Bioinformatics **21** (7): 988–992.

38. Jeffrey Skolnick, Andrzej Kolinski (1991) *Dynamic Monte Carlo Simulations of a New Lattice Model of Globular Protein Folding, Structure and Dynamics.* J. Mol. Biol. **256**, 623-644.

39. Jacques Chomilier, Marc Lamarine, Jean-Paul Mornon, Jorge Hernandez Torres, Elias Eliopoulos, Nikolaos Papandreou (2004) *Analysis of fragments induced by simulated lattice protein folding.* C. R. Biologies **327**, 431-443.

40. Tightened End Fragments Assignment http://bioserv.rpbs.jussieu.fr/TEF/ *«Help».*

41. Altman DG, Bland JM (1994) *"Diagnostic tests. 1: Sensitivity and specificity".* BMJ **308** (6943).

42. Wikipedia The Free Encyclopedia www.wikipedia.org τίτλος άρθρου *«Accuracy and Precision».*

43. Adam Zelma, Ceslovas Venclovas, Krzysztof Fidelis, Burkhard Rost (1999) *A Modified Definition of Sov, a Segment – Based Measure for Protein Secondary Structure Prediction Assessment.* PROTEINS: Structure, Function and Genetics **34,** 220-223.

44. Burkhard Rost, Chris Sander, Reinhard Schneider (1994) *Redefining the Goals of Protein Secondary Structure Prediction.* J. Mol. Biol. **235,** 13-26.

45. Ian W. Davis, Vincent B. Chen (2007) *The KiNG manual.*

46. Macromolecule Analysis and Kinemage Home Page http://kinemage.biochem.duke.edu/index.php

47. Ian W. Davis (2005) *The Chiropraxis tools manual.*

48. Ian W. Davis, Vincent B. Chen (2007) *The Extratools manual.*

49. Ian W. Davis, Jane S. Richardson, David C. Richardson (2006) *The Kinemage File Format, v1.0*

50. T. L Religa, J. S Markson, U. Mayor, S. M. V. Freund & A. R. Fersht (2005) *Solution structure of a protein denatured state and folding intermediate.* NATURE **Vol 473,** 1053 – 1056.

51. Ugo Mayor, ,J.Gunter Grossmann, , Nicholas W. Foster, Stefan M. V. Freund and Alan R. Fersht (2003) *The Denatured State of Engrailed Homeodomain under Denaturing and Native Conditions* J. Mol. Biol. **333**, 977–991.

ΠΑΡΑΡΤΗΜΑ

Πίνακας Α. *Συγκεντρωτικός πίνακας στοιχείων των 107 πρωτεϊνών για τις οποίες κάναμε την πρόβλεψη. Στην πρώτη στήλη με τον τίτλο «Κωδικός» φαίνεται ο κωδικός pdb για την κάθε πρωτεΐνη. Στην δεύτερη στήλη με τον τίτλο «Μήκος» φαίνεται το μήκος της κάθε πρωτεΐνης σε πλήθος αμινοξέων. Στην τρίτη στήλη με τον τίτλο «Στοιχεία β ταγούς» φαίνεται το είδος της κάθε πρωτεΐνης όσον αφορά τα στοιχεία της δευτεροταγούς δομής. Στην τέταρτη στήλη με τον τίτλο «Ταξινόμηση» φαίνεται η ταξινόμηση της κάθε πρωτεΐνης με βάση τον οργανισμό από τον οποίο προέρχεται. Οι τελευταίες 3 στήλες περιέχουν τις τιμές που έχουν οι πρωτεΐνες για συγκεκριμένα στατιστικά μέτρα αξιολόγησης της πρόβλεψης των άκρων TEF. Η πέμπτη στήλη αφορά το sensitivity, η έκτη το specificity και η τελευταία το accuracy. Για τα μέτρα Sn και Sp έχει χρησιμοποιηθεί «παράθυρο» ± 5 θέσεων.*

Πίνακας Α

Κωδικός	**Μήκος**	**Στοιχεία βταγούς**	**Ταξινόμηση**	**Sn**	**Sp**	**Acc**
1aa0	113	coil	Virus	0.50	0.29	0.92
1aba	87	a/b	Virus	0.50	0.40	0.90
1abrA	251	a+b	-	0.19	0.30	0.90
1acf	125	a+b	Eukaryotic	0.40	0.80	0.88
1aep	153	a	Eukaryotic	0.50	1	0.93
1ag2	103	a+b	Eukaryotic	0.75	1	0.92
1aihA	170	a+b	Virus	0.30	0.60	0.92
1ajj	37	small	Eukaryotic	0.50	0.50	0.89
1akz	223	a/b	Eukaryotic	0.50	0.62	0.88
1anu	138	b	Bacteria	0.33	0.67	0.93
1apyA	161	a+b	Eukaryotic	0.20	0.63	0.89
1ast	200	a+b	Eukaryotic	0.20	0.50	0.93
1asu	162	a/b	Virus	0.38	0.50	0.91
1ble	161	a/b	Bacteria	0.50	0.50	0.93
1bp2	123	a	Eukaryotic	0.83	0.80	0.91
1bvd	153	a	Eukaryotic	0	0	0
1c0bA	124	a+b	Eukaryotic	0	0	0.94
1caa	53	small	Archaea	0.50	0.25	0.89
1cauB	184	b	Eukaryotic	0.13	0.20	0.93
1cbs	137	b	Eukaryotic	0.50	0.67	0.87
1cdcA	96	b	Eukaryotic	0.25	0.20	0.91
1ctf	68	a+b	Bacteria	0.25	0.50	0.91
1cus	197	a/b	Eukaryotic	0.70	0.67	0.90
1dhr	236	a/b	Eukaryotic	0.83	0.57	0.91
1dkeA	141	a	Eukaryotic	0.17	0.25	0.93
1dtdB	61	small	Eukaryotic	1	0.50	0.92
1dtp	190	a+b	Virus	0.40	0.45	0.92

1durA	55	a+b	Bacteria	0	0	0.93
1eca	136	a	Eukaryotic	0.67	0.50	0.91
1edmB	39	small	Eukaryotic	0	0	0
1ehs	48	small	Bacteria	0.50	0.50	0.92
1ejgA	46	small	Eukaryotic	0	0	0
1enh	54	a	Eukaryotic	1	0.50	0.96
1f3g	150	b	Bacteria	0.17	0.20	0.93
1fas	61	small	Eukaryotic	0.25	0.34	0.92
1fkb	107	a+b	Eukaryotic	1	1	0.93
1frd	98	a+b	Bacteria	0.17	0.50	0.92
1fxd	58	a+b	Bacteria	0	0	0.93
1gmpA	96	a+b	Bacteria	0.38	0.50	0.88
1gpc	218	b	Virus	0.25	0.60	0.95
1hbg	147	a	Eukaryotic	0.67	0.50	0.92
1hip	85	small	Bacteria	0.50	0.75	0.91
1hpi	71	small	Bacteria	0.50	0.50	0.94
1hucb	205	a+b	Eukaryotic	0.40	0.57	0.93
1i8nA	89	small	Eukaryotic	0	0	0.92
1icfI	65	small	Eukaryotic	0.50	0.67	0.92
1jkeB	145	a/b	Bacteria	0.84	0.56	0.91
1knb	186	b	Virus	0.50	0.56	0.90
1knt	55	small	Eukaryotic	0	0	0
1lbbA	258	a/b	Eukaryotic	0.34	0.40	0.92
1lbd	238	a	Eukaryotic	0.30	0.60	0.95
1lcl	141	b	Eukaryotic	0.34	0.67	0.94
1lis	131	a	Eukaryotic	0.25	0.20	0.93
1lki	172	a	Eukaryotic	0.50	0.57	0.92
1lsg	144	a+b	Eukaryotic	0.50	0.63	0.90
1mgsA	73	a+b	Eukaryotic	0.75	0.40	0.90
1nox	200	a+b	Bacteria	0	0	0.96
1ns5A	152	a/b	Bacteria	0.75	0.55	0.88
1occD	144	membrane	Eukaryotic	0.50	1	0.96
1opr	213	a/b	Bacteria	0.17	0.50	0.82
1pht	83	b	Eukaryotic	0	0	0.95
1pk4	79	small	Eukaryotic	0.75	1	0.91
1plfB	65	a+b	Eukaryotic	0	0	0.91
1pmy	123	b	Bacteria	0.20	0.50	0.93
1poc	134	a	Eukaryotic	0.50	0.50	0.93
1ptf	87	a+b	Bacteria	0.50	0.40	0.90
1pwt	61	b	Eukaryotic	0	0	0
1qabA	124	b	Eukaryotic	0.67	0.60	0.91
1reiA	107	b	Eukaryotic	0	0	0.91
1rro	108	a	Eukaryotic	0.75	0.67	0.91
1rvvA	154	a/b	Bacteria	0.67	0.80	0.94
1semA	58	b	Eukaryotic	0.50	0.75	0.86
1sgpI	51	small	Eukaryotic	0.50	0.50	0.92
1shaA	103	a+b	Virus	0.63	1	0.93
1sno	136	b	Bacteria	0.67	0.67	0.91

1tgj	112	small	Eukaryotic	0.50	1	0.88
1tml	286	a/b	Bacteria	0.36	0.29	0.89
1tpfB	249	a/b	Eukaryotic	0.58	0.70	0.91
1ubi	76	a+b	Eukaryotic	0.50	0.50	0.89
1utg	70	a	Eukaryotic	0	0	0.94
1yna	194	b	Eukaryotic	0.50	0.83	0.91
2act	218	a+b	Eukaryotic	0.60	0.67	0.91
2ayh	214	b	Bacteria	0.30	0.34	0.91
2bbkL	125	small	Bacteria	0.75	0.34	0.91
2ci2I	65	a+b	Eukaryotic	1	0.67	0.92
2cy3	118	a	Bacteria	0.84	0.60	0.88
2erl	40	a	Eukaryotic	0.50	0.67	0.93
2ilk	155	a	Eukaryotic	0.50	0.50	0.95
2lhb	149	a	Eukaryotic	0.34	0.17	0.93
2mcm	112	b	Bacteria	0.30	1	0.89
2mhbA	141	a	Eukaryotic	0	0	0.94
2mhr	118	a	Eukaryotic	0.38	1	0.93
2pelA	232	b	Eukaryotic	0.40	0.67	0.94
2pii	112	a+b	Bacteria	0.25	0.25	0.93
2sas	185	a	Eukaryotic	0.50	0.43	0.94
2sns	141	b	Bacteria	0	0	0.95
2stv	184	b	Virus	0.38	0.38	0.91
3c2c	112	a	Bacteria	0.67	0.57	0.88
3chy	128	a/b	Bacteria	0.50	0.60	0.91
3cla	213	a/b	Bacteria	0.36	0.67	0.92
3cytO	103	a	Eukaryotic	0.50	0.40	0.93
4cpv	108	a	Eukaryotic	0.75	0.60	0.92
4rxn	54	small	Bacteria	0.50	0.50	0.93
5nll	138	a/b	Bacteria	0.34	0.40	0.93
5p21	166	a/b	Eukaryotic	0.60	0.71	0.90
153l	185	a+b	Eukaryotic	0.75	0.60	0.91
256bA	106	a	Bacteria	0.17	0.50	0.92

Πίνακας Β. *Συγκεντρωτικός πίνακας στοιχείων των 107 πρωτεϊνών για τις οποίες κάναμε την πρόβλεψη. Στην πρώτη στήλη με τον τίτλο «Κωδικός» φαίνεται ο κωδικός pdb για την κάθε πρωτεΐνη. Στην δεύτερη στήλη με τον τίτλο «Μήκος» φαίνεται το μήκος της κάθε πρωτεΐνης σε πλήθος αμινοξέων. Στην τρίτη στήλη με τον τίτλο «Στοιχεία β ταγούς» φαίνεται το είδος της κάθε πρωτεΐνης όσον αφορά τα στοιχεία της δευτεροταγούς δομής. Στην τέταρτη στήλη με τον τίτλο «Ταξινόμηση» φαίνεται η ταξινόμηση της κάθε πρωτεΐνης με βάση τον οργανισμό από τον οποίο προέρχεται. Οι τελευταίες 3 στήλες περιέχουν τις τιμές που έχουν οι πρωτεΐνες για συγκεκριμένα στατιστικά μέτρα αξιολόγησης της πρόβλεψης των περιοχών TEF. Η πέμπτη στήλη αφορά το sensitivity, η έκτη το specificity και η τελευταία το accuracy.*

Πίνακας Β

Κωδικός	**Μήκος**	**Στοιχεία βταγούς**	**Ταξινόμηση**	**Sn**	**Sp**	**Acc**
1aa0	113	coil	Virus	1	0.33	0.44
1aba	87	a/b	Virus	0.90	0.59	0.57
1abrA	251	a+b	-	0.37	0.63	0.41
1acf	125	a+b	Eukaryotic	0.69	0.81	0.68
1aep	153	a	Eukaryotic	0.23	0.71	0.52
1ag2	103	a+b	Eukaryotic	0.82	1	0.83
1aihA	170	a+b	Virus	0.51	0.63	0.59
1ajj	37	small	Eukaryotic	0.25	0.22	0.30
1akz	223	a/b	Eukaryotic	0.78	0.82	0.71
1anu	138	b	Bacteria	0.48	0.90	0.60
1apyA	161	a+b	Eukaryotic	0.76	0.70	0.68
1ast	200	a+b	Eukaryotic	0.25	0.62	0.40
1asu	162	a/b	Virus	0.58	0.68	0.59
1ble	161	a/b	Bacteria	0.88	0.93	0.86
1bp2	123	a	Eukaryotic	0.57	0.57	0.63
1bvd	153	a	Eukaryotic	0	0	0
1c0ba	124	a+b	Eukaryotic	0.11	0.36	0.23
1caa	53	small	Archaea	0.88	0.95	0.85
1cauB	184	b	Eukaryotic	0.59	0.80	0.66
1cbs	137	b	Eukaryotic	0.77	0.90	0.74
1cdcA	96	b	Eukaryotic	0.72	0.29	0.31
1ctf	68	a+b	Bacteria	0.51	1	0.57
1cus	197	a/b	Eukaryotic	0.74	0.65	0.60
1dhr	236	a/b	Eukaryotic	0.93	0.81	0.79
1dkeA	141	a	Eukaryotic	0.46	0.61	0.54
1dtdB	61	small	Eukaryotic	1	0.95	0.97
1dtp	190	a+b	Virus	0.74	0.72	0.62
1durA	55	a+b	Bacteria	0.48	1	0.55
1eca	136	a	Eukaryotic	0.91	0.76	0.71
1edmB	39	small	Eukaryotic	0	0	0

1ehs	48	small	Bacteria	1	0.58	0.77
1ejgA	46	small	Eukaryotic	0	0	0
1enh	54	a	Eukaryotic	1	0.60	0.67
1f3g	150	b	Bacteria	0.58	0.65	0.57
1fas	61	small	Eukaryotic	1	0.83	0.85
1fkb	107	a+b	Eukaryotic	0.77	0.73	0.77
1frd	98	a+b	Bacteria	0.42	0.82	0.56
1fxd	58	a+b	Bacteria	0.76	0.74	0.71
1gmpA	96	a+b	Bacteria	0.60	0.70	0.60
1gpc	218	b	Virus	0.58	0.59	0.67
1hbg	147	a	Eukaryotic	0.84	0.68	0.65
1hip	85	small	Bacteria	0.65	0.58	0.49
1hpi	71	small	Bacteria	0	0	0.04
1hucb	205	a+b	Eukaryotic	0.46	0.59	0.45
1i8nA	89	small	Eukaryotic	0.83	0.34	0.87
1icfI	65	small	Eukaryotic	1	0.66	0.68
1jkeB	145	a/b	Bacteria	1	0.87	0.89
1knb	186	b	Virus	0.77	0.57	0.52
1knt	55	small	Eukaryotic	0	0	0
1lbbA	258	a/b	Eukaryotic	0.63	0.56	0.45
1lbd	238	a	Eukaryotic	0.34	0.65	0.51
1lcl	141	b	Eukaryotic	0.39	0.34	0.52
1lis	131	a	Eukaryotic	0.58	0.35	0.44
1lki	172	a	Eukaryotic	0.55	0.35	0.45
1lsg	144	a+b	Eukaryotic	1	0.69	0.69
1mgsA	73	a+b	Eukaryotic	1	0.52	0.60
1nox	200	a+b	Bacteria	0.15	0.35	0.37
1ns5A	152	a/b	Bacteria	0.88	0.76	0.72
1occD	144	membrane	Eukaryotic	0.33	1	0.65
1opr	213	a/b	Bacteria	0.13	0.51	0.33
1pht	83	b	Eukaryotic	0.98	0.72	0.63
1pk4	79	small	Eukaryotic	0.98	0.72	0.80
1plfB	65	a+b	Eukaryotic	0.22	0.30	0.37
1pmy	123	b	Bacteria	0.25	0.77	0.37
1poc	134	a	Eukaryotic	0.48	0.09	0.56
1ptf	87	a+b	Bacteria	0.93	0.90	0.86
1pwt	61	b	Eukaryotic	0	0	0
1qabA	124	b	Eukaryotic	0.62	0.62	0.50
1reiA	107	b	Eukaryotic	0.61	1	0.70
1rro	108	a	Eukaryotic	0.80	0.74	0.70
1rvvA	154	a/b	Bacteria	0.64	0.78	0.65
1semA	58	b	Eukaryotic	1	0.83	0.86
1sgpI	51	small	Eukaryotic	0.84	0.81	0.78
1shaA	103	a+b	Virus	0.99	0.99	0.98
1sno	136	b	Bacteria	0.81	0.60	0.57
1tgj	112	small	Eukaryotic	0.53	0.77	0.50
1tml	286	a/b	Bacteria	0.84	0.81	0.72
1tpfB	249	a/b	Eukaryotic	0.62	0.95	0.67

1ubi	76	a+b	Eukaryotic	1	0.73	0.76
1utg	70	a	Eukaryotic	0.15	0.26	0.27
1yna	194	b	Eukaryotic	0.49	0.61	0.57
2act	218	a+b	Eukaryotic	0.65	0.70	0.65
2ayh	214	b	Bacteria	0.60	0.70	0.55
2bbkL	125	small	Bacteria	0.49	0.35	0.37
2ci2I	65	a+b	Eukaryotic	1	0.92	0.94
2cy3	118	a	Bacteria	0.74	0.49	0.47
2erl	40	a	Eukaryotic	1	0.71	0.73
2ilk	155	a	Eukaryotic	0.44	0.35	0.66
2lhb	149	a	Eukaryotic	0.56	0.80	0.61
2mcm	112	b	Bacteria	0.18	0.84	0.33
2mhbA	141	a	Eukaryotic	0.15	1	0.40
2mhr	118	a	Eukaryotic	0.36	0.89	0.50
2pelA	232	b	Eukaryotic	0.36	0.46	0.53
2pii	112	a+b	Bacteria	0.77	0.91	0.79
2sas	185	a	Eukaryotic	0.41	0.40	0.38
2sns	141	b	Bacteria	0.23	0.49	0.43
2stv	184	b	Virus	0.43	0.40	0.29
3c2c	112	a	Bacteria	1	0.59	0.62
3chy	128	a/b	Bacteria	0.42	0.43	0.43
3cla	213	a/b	Bacteria	0.36	0.68	0.49
3cytO	103	a	Eukaryotic	0.94	0.41	0.51
4cpv	108	a	Eukaryotic	0.85	0.97	0.89
4rxn	54	small	Bacteria	0.73	1	0.78
5nll	138	a/b	Bacteria	0.65	0.72	0.67
5p21	166	a/b	Eukaryotic	0.88	0.88	0.81
153l	185	a+b	Eukaryotic	1	0.57	0.68
256bA	106	a	Bacteria	0.29	0.47	0.46

Πίνακας Γ. *Συγκεντρωτικός πίνακας στοιχείων των 107 πρωτεϊνών για τις οποίες κάναμε την πρόβλεψη. Στην πρώτη στήλη με τον τίτλο «Κωδικός» φαίνεται ο κωδικός pdb για την κάθε πρωτεΐνη. Στην δεύτερη στήλη με τον τίτλο «Μήκος» φαίνεται το μήκος της κάθε πρωτεΐνης σε πλήθος αμινοξέων. Στην τρίτη στήλη με τον τίτλο «Στοιχεία β ταγούς» φαίνεται το είδος της κάθε πρωτεΐνης όσον αφορά τα στοιχεία της δευτεροταγούς δομής. Στην τέταρτη στήλη με τον τίτλο «Ταξινόμηση» φαίνεται η ταξινόμηση της κάθε πρωτεΐνης με βάση τον οργανισμό από τον οποίο προέρχεται. Οι τελευταίες 2 στήλες περιέχουν τις τιμές που έχουν οι πρωτεΐνες για συγκεκριμένα στατιστικά μέτρα αξιολόγησης της πρόβλεψης των περιοχών TEF. Η πέμπτη στήλη αφορά το SOV observed και η τελευταία το SOV predicted.*

Πίνακας Γ

Κωδικός	Μήκος	Στοιχεία βταγούς	Ταξινόμηση	SOV ob	SOV pre
1aa0	113	coil	Virus	0.79	0.45
1aba	87	a/b	Virus	0.77	0.74
1abrA	251	a+b	-	0.36	0.55
1acf	125	a+b	Eukaryotic	0.51	0.53
1aep	153	a	Eukaryotic	0.23	1
1ag2	103	a+b	Eukaryotic	0.53	0.53
1aihA	170	a+b	Virus	0.24	0.41
1ajj	37	small	Eukaryotic	0.27	0.27
1akz	223	a/b	Eukaryotic	0.61	0.66
1anu	138	b	Bacteria	0.49	0.75
1apyA	161	a+b	Eukaryotic	0.61	0.70
1ast	200	a+b	Eukaryotic	0.27	0.49
1asu	162	a/b	Virus	0.70	0.73
1ble	161	a/b	Bacteria	0.53	0.60
1bp2	123	a	Eukaryotic	0.63	0.69
1bvd	153	a	Eukaryotic	0	0
1c0ba	124	a+b	Eukaryotic	0.15	0.36
1caa	53	small	Archaea	0.96	0.96
1cauB	184	b	Eukaryotic	0.46	0.55
1cbs	137	b	Eukaryotic	0.73	0.67
1cdcA	96	b	Eukaryotic	0.55	0.39
1ctf	68	a+b	Bacteria	0.71	1
1cus	197	a/b	Eukaryotic	0.51	0.49
1dhr	236	a/b	Eukaryotic	0.62	0.55
1dkeA	141	a	Eukaryotic	0.42	0.49
1dtdB	61	small	Eukaryotic	0.84	0.89
1dtp	190	a+b	Virus	0.51	0.51
1durA	55	a+b	Bacteria	0.71	0.71
1eca	136	a	Eukaryotic	0.58	0.57
1edmB	39	small	Eukaryotic	0	0
1ehs	48	small	Bacteria	0.85	0.85

1ejgA	46	small	Eukaryotic	0	0
1enh	54	a	Eukaryotic	0.52	0.66
1f3g	150	b	Bacteria	0.57	0.65
1fas	61	small	Eukaryotic	0.67	0.67
1fkb	107	a+b	Eukaryotic	0.71	0.73
1frd	98	a+b	Bacteria	0.37	0.69
1fxd	58	a+b	Bacteria	0.98	0.98
1gmpA	96	a+b	Bacteria	0.45	0.54
1gpc	218	b	Virus	0.48	0.60
1hbg	147	a	Eukaryotic	0.60	0.63
1hip	85	small	Bacteria	0.48	0.42
1hpi	71	small	Bacteria	0	0
1hucb	205	a+b	Eukaryotic	0.41	0.53
1i8nA	89	small	Eukaryotic	0.76	0.78
1icfI	65	small	Eukaryotic	0.63	0.64
1jkeB	145	a/b	Bacteria	0.56	0.62
1knb	186	b	Virus	0.51	0.51
1knt	55	small	Eukaryotic	0	0
1lbbA	258	a/b	Eukaryotic	0.45	0.54
1lbd	238	a	Eukaryotic	0.38	0.55
1lcl	141	b	Eukaryotic	0.39	0.40
1lis	131	a	Eukaryotic	0.46	0.46
1lki	172	a	Eukaryotic	0.50	0.51
1lsg	144	a+b	Eukaryotic	0.61	0.56
1mgsA	73	a+b	Eukaryotic	0.52	0.46
1nox	200	a+b	Bacteria	0.23	0.27
1ns5A	152	a/b	Bacteria	0.64	0.73
1occD	144	Membrane	Eukaryotic	0.39	0.55
1opr	213	a/b	Bacteria	0.15	0.68
1pht	83	b	Eukaryotic	0.23	0.23
1pk4	79	small	Eukaryotic	0.91	0.87
1plfB	65	a+b	Eukaryotic	0.23	0.37
1pmy	123	b	Bacteria	0.29	1
1poc	134	a	Eukaryotic	0.49	0.47
1ptf	87	a+b	Bacteria	0.62	0.69
1pwt	61	b	Eukaryotic	0	0
1qabA	124	b	Eukaryotic	0.71	0.53
1reiA	107	b	Eukaryotic	0.60	0.73
1rro	108	a	Eukaryotic	0.96	0.94
1rvvA	154	a/b	Bacteria	0.44	0.62
1semA	58	b	Eukaryotic	0.78	0.84
1sgpI	51	small	Eukaryotic	1	1
1shaA	103	a+b	Virus	0.57	0.56
1sno	136	b	Bacteria	0.60	0.69
1tgj	112	small	Eukaryotic	0.43	0.53
1tml	286	a/b	Bacteria	0.59	0.59
1tpfB	249	a/b	Eukaryotic	0.61	0.62
1ubi	76	a+b	Eukaryotic	0.67	0.59

1utg	70	a	Eukaryotic	0.21	0.21
1yna	194	b	Eukaryotic	0.36	0.45
2act	218	a+b	Eukaryotic	0.59	0.57
2ayh	214	b	Bacteria	0.47	0.66
2bbkL	125	small	Bacteria	0.50	0.25
2ci2I	65	a+b	Eukaryotic	0.76	0.76
2cy3	118	a	Bacteria	0.45	0.54
2erl	40	a	Eukaryotic	0.62	0.73
2ilk	155	a	Eukaryotic	0.42	0.48
2lhb	149	a	Eukaryotic	0.59	0.52
2mcm	112	b	Bacteria	0.18	1
2mhbA	141	a	Eukaryotic	0.22	0.65
2mhr	118	a	Eukaryotic	0.30	0.92
2pelA	232	b	Eukaryotic	0.29	0.39
2pii	112	a+b	Bacteria	0.65	0.69
2sas	185	a	Eukaryotic	0.36	0.42
2sns	141	b	Bacteria	0.43	0.56
2stv	184	b	Virus	0.37	0.51
3c2c	112	a	Bacteria	0.55	0.61
3chy	128	a/b	Bacteria	0.45	0.45
3cla	213	a/b	Bacteria	0.38	0.58
3cytO	103	a	Eukaryotic	0.61	0.44
4cpv	108	a	Eukaryotic	0.74	0.82
4rxn	54	small	Bacteria	1	1
5nll	138	a/b	Bacteria	0.65	0.58
5p21	166	a/b	Eukaryotic	0.61	0.73
153l	185	a+b	Eukaryotic	0.72	0.71
256bA	106	a	Bacteria	0.21	0.71

Πίνακας Α1 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου sensitivity για την πρόβλεψη των άκρων TEF ανά κατηγορίες οργανισμών από τους οποίους προέρχονται οι πρωτεΐνες. Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (βακτήρια, ευκαριωτικά ή ιοί), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου sensitivity. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Α1

Ταξινόμηση	**αριθμός πρωτεϊνών**	**Sensitivity**			
		Μέση τιμή	**τυπική απόκλιση**	**μέγιστη τιμή**	**ελάχιστη τιμή**
Βακτήρια	32	0.39	0.25	0.84 (2cy3) (1jkeB)	0 (1durA) (1fxd) (1nox) (2sns)
Ευκαριωτικά	64	0.44	0.29	1 (1dtdB) (1enh) (1fkb) (2ci2I)	0 (1bvd) (1c0bA) (1edmB) (1ejgA) (1i8nA) (1knt) (1pht) (1plfB) (1pwt) (1reiA) (1utg) (2mhbA)
Ιοί	9	0.43	0.12	0.63 (1shaA)	0.25 (1gpc)

Πίνακας Α2 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου specificity για την πρόβλεψη των άκρων TEF ανά κατηγορίες οργανισμών από τους οποίους προέρχονται οι πρωτεΐνες. Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (βακτήρια, ευκαριωτικά ή ιοί), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου specificity. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Α2

Ταξινόμηση	**αριθμός πρωτεϊνών**	**Specificity**			
		Μέση τιμή	**τυπική απόκλιση**	**μέγιστη τιμή**	**ελάχιστη τιμή**
Βακτήρια	32	0.46	0.24	1 (2mcm)	0 (1durA) (1fxd) (1nox) (2sns)
Ευκαριωτικά	64	0.49	0.31	1 (1aep) (1ag2) (1fkb) (1occD) (1pk4) (1tgj) (2mhr)	0 (1bvd) (1c0bA) (1edmB) (1ejgA) (1i8nA) (1knt) (1pht) (1plfB) (1pwt) (1reiA) (1utg) (2mhbA
Ιοί	9	0.53	0.20	1 (1shaA)	0.29 (1aa0)

Πίνακας Α3 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου accuracy για την πρόβλεψη των άκρων TEF ανά κατηγορίες οργανισμών από τους οποίους προέρχονται οι πρωτεΐνες. Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (βακτήρια, ευκαριωτικά ή ιοί), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου accuracy. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Α3

Ταξινόμηση	**αριθμός πρωτεϊνών**	**Accuracy**			
		Μέση Τιμή	**τυπική απόκλιση**	**μέγιστη τιμή**	**ελάχιστη τιμή**
Βακτήρια	32	0.91	0.03	0.96 (1nox)	0.82 (1opr)
Ευκαριωτικά	64	0.85	0.25	0.96 (1enh) (1occD)	0 (1bvd) (1edmB) (1ejgA) (1knt) (1pwt)
Ιοί	9	0.92	0.02	0.95 (1gpc)	0.90 (1knb) (1aba)

Πίνακας Β1 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου sensitivity για την πρόβλεψη των περιοχών TEF ανά κατηγορίες οργανισμών από τους οποίους προέρχονται οι πρωτεΐνες. Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (βακτήρια, ευκαριωτικά ή ιοί), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου sensitivity. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Β1

Ταξινόμηση	**αριθμός πρωτεϊνών**	**Sensitivity**			
		Μέση τιμή	**τυπική απόκλιση**	**μέγιστη τιμή**	**ελάχιστη τιμή**
Βακτήρια	32	0.58	0.28	1 (1ehs) (1jkeB) (3c2c)	0 (1hpi)
Ευκαριωτικά	64	0.61	0.32	1 (1dtp) (1enh) (1fas) (1icfl) (1lsg) (1mgsA) (1semA) (1ubi) (2ci2l) (2erl) (153l)	0 (1bvd) (1edmB) (1ejgA) (1knt) (1pwt)
Ιοί	9	0.72	0.21	1 (1aa0)	0.43 (2stv)

Πίνακας Β2 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου specificity για την πρόβλεψη των περιοχών TEF ανά κατηγορίες οργανισμών από τους οποίους προέρχονται οι πρωτεΐνες. Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (βακτήρια, ευκαριωτικά ή ιοί), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου specificity. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Β2

Ταξινόμηση	**αριθμός πρωτεϊνών**	**Specificity**			
		Μέση Τιμή	**τυπική απόκλιση**	**μέγιστη τιμή**	**ελάχιστη τιμή**
Βακτήρια	32	0.69	0.23	1 (1ctf) (1durA) (4rxn)	0 (1hpi)
Ευκαριωτικά	64	0.60	0.28	1 (1ag2) (1occD) (1reiA) (2mhbA)	0 (1bvd) (1edmB) (1ejgA) (1knt) (1pwt
Ιοί	9	0.61	0.19	0.99 (1shaA)	0.33 (1aa0)

Πίνακας Β3 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου accuracy για την πρόβλεψη των περιοχών TEF ανά κατηγορίες οργανισμών από τους οποίους προέρχονται οι πρωτεΐνες. Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (βακτήρια, ευκαριωτικά ή ιοί), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου accuracy. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Β3

Ταξινόμηση	αριθμός πρωτεϊνών	Accuracy			
		Μέση Τιμή	τυπική απόκλιση	μέγιστη τιμή	Ελάχιστη Τιμή
Βακτήρια	32	0.57	0.19	0.89 (1jkeB)	0.04 (1hpi)
Ευκαριωτικά	64	0.57	0.23	0.97 (1dtdB)	0 (1bvd) (1edmB) (1ejgA) (1knt) (1pwt
Ιοί	9	0.59	0.19	0.98 (1shaA)	0.29 (2stv)

Πίνακας Γ1 *Μέση τιμή ,τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου SOV observed για την πρόβλεψη των περιοχών TEF ανά κατηγορίες οργανισμών από τους οποίους προέρχονται οι πρωτεΐνες. Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (βακτήρια, ευκαριωτικά ή ιοί), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου SOV observed. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Γ1

Ταξινόμηση	αριθμός πρωτεϊνών	SOV observed			
		Μέση Τιμή	τυπική απόκλιση	μέγιστη τιμή	Ελάχιστη Τιμή
Βακτήρια	32	0.50	0.22	1 (4rxn)	0 (1hpi)
Ευκαριωτικά	64	0.49	0.24	1 (1sgpl)	0 (1bvd) (1edmB) (1ejgA) (1knt) (1pwt
Ιοί	9	0.55	0.18	0.79 (1aa0)	0.24 (1aihA)

Πίνακας Γ2 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου SOV predicted για την πρόβλεψη των περιοχών TEF ανά κατηγορίες οργανισμών από τους οποίους προέρχονται οι πρωτεΐνες. Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (βακτήρια, ευκαριωτικά ή ιοί), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου SOV predicted. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Γ2

Ταξινόμηση	**αριθμός πρωτεϊνών**	**SOV predicted**			
		Μέση Τιμή	**τυπική απόκλιση**	**μέγιστη τιμή**	**Ελάχιστη Τιμή**
Βακτήρια	32	0.64	0.22	1 (1ctf) (2mcm) (4rxn)	0 (1hpi)
Ευκαριωτικά	64	0.55	0.24	1 (1aep)	0 (1bvd) (1edmB) (1ejgA) (1knt) (1pwt
Ιοί	9	0.56	0.11	0.74 (1aba)	0.41 (1aihA)

Πίνακας Α4 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου sensitivity για την πρόβλεψη των άκρων TEF ανά scop κατηγορίες Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (all a, all b, a+b, a/b, small), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου sensitivity. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Α4

scop κατηγορίες	**αριθμός πρωτεϊνών**	**Sensitivity**			
		μέση τιμή	**τυπική απόκλιση**	**μέγιστη τιμή**	**ελάχιστη τιμή**
all a	24	0.75	0.40	1 (1enh)	0 (1bvd) (1utg) (2mhbA)
all b	21	0.30	0.21	0.67 (1qabA) (1sno)	0 (1pht) (1pwt) (1reiA) (2sns)
a+b	26	0.39	0.30	1 (1fkb) (2ci2l)	0 (1c0bA) (1durA) (1fxd) (1nox) (1plfB)
a/b	17	0.52	0.19	0.84 (1jkeB)	0.17 (1opr)
small	17	0.43	0.29	1 (1dtdB)	0 (1edmB) (1ejgA) (1knt)

Πίνακας Α5 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου specificity για την πρόβλεψη των άκρων TEF ανά scop κατηγορίες Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (all a, all b, a+b, a/b, small), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου specificity. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Α5

scop κατηγορίες	**αριθμός πρωτεϊνών**	**Specificity**			
		μέση τιμή	**τυπική απόκλιση**	**μέγιστη τιμή**	**ελάχιστη τιμή**
all a	24	0.49	0.27	1 (1aep) (2mhr)	0 (1bvd) (1utg) (2mhbA)
all b	21	0.45	0.30	1 (2mcm)	0 (1pht) (1pwt) (1reiA) (2sns)
a+b	26	0.48	0.30	1 (1fkb) (1ag2)	0 (1c0bA) (1durA) (1fxd) (1nox) (1plfB)
a/b	17	0.56	0.13	0.8 (1rvvA)	0.29 (1tml)
small	17	0.43	0.32	1 (1pk4) (1tgj)	0 (1edmB) (1ejgA) (1knt)

Πίνακας Α6 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου accuracy για την πρόβλεψη των άκρων TEF ανά scop κατηγορίες Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (all a, all b, a+b, a/b, small), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου accuracy. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Α6

scop κατηγορίες	**αριθμός πρωτεϊνών**	**Accuracy**			
		μέση τιμή	**τυπική απόκλιση**	**μέγιστη τιμή**	**ελάχιστη τιμή**
all a	24	0.89	0.19	0.96 (1enh)	0 (1bvd)
all b	21	0.87	0.20	0.95 (1pht)	0 (1pwt)
a+b	26	0.91	0.02	0.94 (1c0bA)	0.88 (1gmpA) (1acf)
a/b	17	0.90	0.03	0.94 (1rvvA)	0.82 (1opr)
Small	17	0.75	0.36	0.94 (1hpi)	0 (1edmB) (1ejgA) (1knt)

Πίνακας Β4 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου sensitivity για την πρόβλεψη των περιοχών TEF ανά scop κατηγορίες Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (all a, all b, a+b, a/b, small), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου sensitivity. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Β4

scop κατηγορίες	**αριθμός πρωτεϊνών**	**Sensitivity**			
		μέση τιμή	**τυπική απόκλιση**	**μέγιστη τιμή**	**ελάχιστη τιμή**
all a	24	0.57	0.30	1 (1enh) (2erl) (3c2c)	0 (1bvd)
all b	21	0.54	0.26	1 (1semA)	0 (1pwt)
a+b	26	0.65	0.28	1 (1lsg) (1mgsA) (1ubi) (2ci2l) (153l)	0.11 (1c0bA)
a/b	17	0.70	0.23	1 (1jkeB)	0.13 (1opr)
Small	17	0.60	0.40	1 (1dtdB) (1icfl) (1ehs) (1fas)	0 (1edmB) (1ejgA) (1hpi) (1knt)

Πίνακας Β5 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου specificity για την πρόβλεψη των περιοχών TEF ανά scop κατηγορίες Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (all a, all b, a+b, a/b, small), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου specificity. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Β5

scop κατηγορίες	**αριθμός πρωτεϊνών**	**Specificity**			
		μέση τιμή	**τυπική απόκλιση**	**μέγιστη τιμή**	**ελάχιστη τιμή**
all a	24	0.53	0.26	1 (2mhbA)	0 (1bvd)
all b	21	0.62	0.24	1 (1reiA)	0 (1pwt)
a+b	26	0.72	0.20	1 (1ctf) (1durA) (1ag2)	0.3 (1plfB)
a/b	17	0.73	0.15	0.95 (1tpfB)	0.43 (3chy)
Small	17	0.52	0.37	1 (4rxn)	0 (1ejgA) (1hpi) (1knt)

Πίνακας Β6 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου accuracy για την πρόβλεψη των περιοχών TEF ανά scop κατηγορίες Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (all a, all b, a+b, a/b, small), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου accuracy. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Β6

scop κατηγορίες	**αριθμός πρωτεϊνών**	**Accuracy**			
		μέση τιμή	**τυπική απόκλιση**	**μέγιστη τιμή**	**ελάχιστη τιμή**
all a	24	0.54	0.18	0.89 (4cpv)	0 (1bvd)
all b	21	0.52	0.19	0.86 (1semA)	0 (1pwt)
a+b	26	0.63	0.18	0.98 (1shaA)	0.23 (1c0bA)
a/b	17	0.64	0.16	0.89 (1jkeB)	0.33 (1opr)
Small	17	0.53	0.35	0.97 (1dtdB)	0 (1edmB) (1ejgA) (1knt)

Πίνακας Γ3 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου SOV observed για την πρόβλεψη των περιοχών TEF ανά scop κατηγορίες Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (all a, all b, a+b, a/b, small), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου SOV observed. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Γ3

Scop κατηγορίες	**αριθμός πρωτεϊνών**	**SOV observed**			
		μέση τιμή	**τυπική απόκλιση**	**μέγιστη τιμή**	**ελάχιστη τιμή**
all a	24	0.46	0.20	0.96 (1rro)	0 (1bvd)
all b	21	0.45	0.19	0.78 (1semA)	0 (1pwt)
a+b	26	0.53	0.20	0.98 (1fxd)	0.15 (1c0bA)
a/b	17	0.54	0.14	0.77 (1aba)	0.15 (1opr)
Small	17	0.55	0.38	1 (1sgpl)	0 (1edmB) (1ejgA) (1hpi) (1knt)

Πίνακας Γ4 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου SOV predicted για την πρόβλεψη των περιοχών TEF ανά scop κατηγορίες Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (all a, all b, a+b, a/b, small), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου SOV predicted. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Γ4

Scop κατηγορίες	**αριθμός πρωτεϊνών**	**SOV predicted**			
		μέση τιμή	**τυπική απόκλιση**	**μέγιστη τιμή**	**ελάχιστη τιμή**
all a	24	0.58	0.22	1 (1aep)	0 (1bvd)
all b	21	0.58	0.24	1 (1pmy) (2mcm)	0 (1pwt)
a+b	26	0.60	0.17	1 (1ctf)	0.27 (1nox)
a/b	17	0.62	0.08	0.74 (1aba)	0.45 (3chy)
Small	17	0.54	0.38	1 (1sgpl) (4rxn)	0 (1edmB) (1ejgA) (1hpi) (1knt)

Πίνακας Α7 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου sensitivity για την πρόβλεψη των άκρων TEF ανά κατηγορία μήκους ακολουθίας Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (0 – 100, >100), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου sensitivity. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Α7

μήκος ακολουθίας	**αριθμός πρωτεϊνών**	**Sensitivity**			
		μέση τιμή	**τυπική απόκλιση**	**μέγιστη τιμή**	**ελάχιστη τιμή**
0 -100	33	0.37	0.31	1 (1dtdB) (1enh) (2ci2I)	0 (1durA) (1edmB) (1ejgA) (1fxd) (1i8nA) (1knt) (1pht) (1plfB) (1pwt) (1utg)
>100	74	0.44	0.24	1 (1fkb)	0 (1bvd) (1c0bA) (1nox) (1reiA) (2mhbA) (2sns)

Πίνακας Α8 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου specificity για την πρόβλεψη των άκρων TEF ανά κατηγορία μήκους ακολουθίας Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (0 – 100, >100), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου specificity. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Α8

μήκος ακολουθίας	αριθμός πρωτεϊνών	Specificity			
		μέση τιμή	τυπική απόκλιση	μέγιστη τιμή	ελάχιστη τιμή
0 -100	33	0.36	0.28	1 (1pk4)	0 (1durA) (1edmB) (1ejgA) (1fxd) (1i8nA) (1knt) (1pht) (1plfB) (1pwt) (1utg)
>100	74	0.54	0.26	1 (1aep) (1ag2) (1fkb) (1occD) (1shaA) (1tgj) (2mcm) (2mhr)	0 (1bvd) (1c0bA) (1nox) (1reiA) (2mhbA) (2sns)

Πίνακας Α9 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου accuracy για την πρόβλεψη των άκρων TEF ανά κατηγορία μήκους ακολουθίας Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (0 – 100, >100), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου accuracy. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Α9

μήκος ακολουθίας	αριθμός πρωτεϊνών	Accuracy			
		μέση τιμή	τυπική απόκλιση	μέγιστη τιμή	ελάχιστη τιμή
0 -100	33	0.80	0.30	0.96 (1enh)	0 (1edmB) (1ejgA) (1knt) (1pwt)
>100	74	0.90	0.11	0.96 (1occD) (1nox)	0 (1bvd)

Πίνακας Β7 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου sensitivity για την πρόβλεψη των περιοχών TEF ανά κατηγορία μήκους ακολουθίας Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (0 – 100, >100), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου sensitivity. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Β7

μήκος ακολουθίας	αριθμός πρωτεϊνών	Sensitivity			
		μέση τιμή	τυπική απόκλιση	μέγιστη τιμή	ελάχιστη τιμή
0 -100	33	0.66	0.38	1 (1jkeB) (1lsg) (3c2c) (153l)	0 (1edmB) (1ejgA) (1hpi) (1knt) (1pwt)
>100	74	0.59	0.25	1 (1aa0) (1jkeB) (1lsg) (3c2c) (153l)	0 (1bvd)

Πίνακας Β8 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου specificity για την πρόβλεψη των περιοχών TEF ανά κατηγορία μήκους ακολουθίας Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (0 – 100, >100), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου specificity. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Β8

μήκος ακολουθίας	αριθμός πρωτεϊνών	Specificity			
		μέση τιμή	τυπική απόκλιση	μέγιστη τιμή	ελάχιστη τιμή
0 -100	33	0.58	0.33	1 (1ctf) (1durA) (4rxn)	0 (1edmB) (1ejgA) (1hpi) (1knt) (1pwt)
>100	74	0.65	0.22	1 (1ag2) (1occD) (1reiA) (2mhbA)	0 (1bvd)

Πίνακας Β9 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου accuracy για την πρόβλεψη των περιοχών TEF ανά κατηγορία μήκους ακολουθίας Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (0 – 100, >100), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου accuracy. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Β9

μήκος ακολουθίας	αριθμός πρωτεϊνών	Accuracy			
		μέση τιμή	τυπική απόκλιση	μέγιστη τιμή	ελάχιστη τιμή
0 -100	33	0.57	0.30	0.97 (1dtdB)	0 (1edmB) (1ejgA) (1knt) (1pwt)
>100	74	0.57	0.17	0.98 (1shaA)	0 (1bvd)

Πίνακας Γ5 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου SOV observed για την πρόβλεψη των περιοχών TEF ανά κατηγορία μήκους ακολουθίας Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (0 – 100, >100),*

στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου SOV observed. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.

Πίνακας Γ5

μήκος ακολουθίας	αριθμός πρωτεϊνών	SOV observed			
		μέση τιμή	τυπική απόκλιση	μέγιστη τιμή	ελάχιστη τιμή
0 -100	33	0.55	0.32	1 (1sgpl) (4rxn)	0 (1edmB) (1ejgA) (1hpi) (1knt) (1pwt)
>100	74	0.48	0.17	0.96 (1rro)	0 (1bvd)

Πίνακας Γ6 *Μέση τιμή, τυπική απόκλιση και μέγιστη/ελάχιστη τιμή του μέτρου SOV predicted για την πρόβλεψη των περιοχών TEF ανά κατηγορία μήκους ακολουθίας Στην πρώτη στήλη αναγράφεται το όνομα της εκάστοτε κατηγορίας (0 – 100, >100), στη δεύτερη το μέγεθος της σε πλήθος πρωτεϊνών, στην τρίτη η μέση τιμή για το μέτρο, στην τέταρτη η τυπική απόκλιση και στις δύο τελευταίες στήλες η μέγιστη και η ελάχιστη τιμή του μέτρου SOV predicted. Τέλος, μέσα σε παρενθέσεις αναφέρεται και ο pdb κωδικός των πρωτεϊνών οι οποίες παρουσιάζουν αυτές τις ακραίες τιμές.*

Πίνακας Γ6

μήκος ακολουθίας	αριθμός πρωτεϊνών	SOV predicted			
		μέση τιμή	τυπική απόκλιση	μέγιστη τιμή	ελάχιστη τιμή
0 -100	33	0.57	0.33	1 (1ctf) (1sgpl) (4rxn)	0 (1edmB) (1ejgA) (1hpi) (1knt) (1pwt)
>100	74	0.59	0.17	1 (1aep) (1pmy) (2mcm)	0 (1bvd)