Κεφάλαιο 9: Δομική Βιοπληροφορική

Σύνοψη
Δομική Βιοπληροφορική, είναι ο κλάδος της βιοπληροφορικής ο οποίος ασχολείται με την ανάλυση και την πρόγνωση της τρισδιάστατης δομής των βιολογικών μακρομορίων, όπως οι πρωτεΐνες, το DNA, και το RNA. Ασχολείται με όλα τα επίπεδα της ανάλυσης των τρισδιάστατων δομών, από την αναπαράσταση και την οπτικοποίηση, τις συγκρίσεις και τις ομαδοποιήσεις των δομών, τις μελέτες του πρωτεϊνικού διπλώματος, την κατασκευή μοντέλων, τη μελέτη των εξελικτικών σχέσεων έως και τη μελέτη της σχέσης δομής και λειτουργίας. Σαν κλάδος έχει ιδιαίτερες σχέσεις με τη μοριακή βιοφυσική και τη δομική βιολογία. Στο κεφάλαιο αυτό θα ασχοληθούμε με τις βασικές μεθοδολογίες της δομικής βιοπληροφορικής και θα δούμε τα πιο γνωστά πακέτα λογισμικού που χρησιμοποιούνται στον τομέα αυτό.

Προαπαιτούμενη γνώση
Στο κεφάλαιο αυτό απαραίτητη είναι η γνώση των εννοιών των κεφαλαίων που ασχολούνται με τη στοίχιση αλληλουχιών και την πολλαπλή στοίχιση, αλλά και τις βάσεις δεδομένων.

9. Εισαγωγή

Δομική Βιοπληροφορική, είναι ο κλάδος της βιοπληροφορικής ο οποίος ασχολείται με την ανάλυση και την πρόγνωση της τρισδιάστατης δομής των βιολογικών μακρομορίων, όπως οι πρωτεΐνες, το DNA και το RNA. Είναι ένας κλάδος που έχει τις ρίζες του στις πρώτες μεθοδολογίες προσδιορισμού της δομής των βιολογικών μακρομορίων από τις δεκαετίες του 1950 και 1960 και κατά συνέπεια είναι ένας κλάδος που αναπτύχθηκε όλα αυτά τα χρόνια, παράλληλα με την ανάπτυξη της μοριακής βιολογίας, της δομικής βιολογίας και της βιοφυσικής.

Το αντικείμενο της μελέτης της δομικής βιοπληροφορικής, είναι οι τρισδιάστατες δομές, δηλαδή οι συντεταγμένες των ατόμων ενός βιολογικού μακρομορίου. Η εύρεση της τρισδιάστατης δομής, είναι από μόνη της μια ιδιαίτερα επίπονη και κοστοβόρα διαδικασία, που αποτελεί περιοριστικό παράγοντα στον τομέα. Έτσι, είναι γνωστό ότι οι διαθέσιμες τρισδιάστατες δομές είναι μια τάξη μεγέθους λιγότερες από τις διαθέσιμες αλληλουχίες. Από την άλλη, η δομή είναι πολύ σημαντική στην κατανόηση της δράσης των βιολογικών μακρομορίων και ειδικά των πρωτεϊνών.

Γενικά, η δομική βιοπληροφορική ασχολείται με όλα τα επίπεδα της ανάλυσης των τρισδιάστατων δομών. Έτσι, έχουμε σε πρώτο επίπεδο τους αλγόριθμους και το λογισμικό που χρησιμοποιούνται για την αναπαράσταση και την οπτικοποίηση των βιολογικών δομών. Τα εργαλεία αυτά, εκτός από τους ειδικούς, χρησιμοποιούνται πλέον και από τον καθένα που κάνει μια εργασία που αναφέρεται σε βιολογικές δομές. Ένα άλλο επίπεδο είναι οι συγκρίσεις και οι ομαδοποιήσεις των δομών. Εδώ έχουμε το πρόβλημα της στοίχισης και υπέρθεσης δομών, αλλά και το πρόβλημα της αναγνώρισης του πρωτεϊνικού διπλώματος με όλες τις επιπτώσεις του. Όλα αυτά, οδηγούν τελικά στο βασικό πρόβλημα της πρόβλεψης της τρισδιάστατης δομής πρωτεϊνών, της κατασκευής δηλαδή μοντέλων για μια πρωτεΐνη για την οποία δεν υπάρχουν πειραματικά δεδομένα. Σε αυτή την περίπτωση, υπάρχει πληθώρα μεθόδων, από την απλή προτυποποίηση με βάση την ομολογία (homology modelling), μέχρι την ύφανση (threading) και την ab initio πρόγνωση της δομής. Τέλος, είτε με δομές πειραματικά προσδιορισμένες, είτε με δομές που έχουν προκύψει από μοντέλα, αντικείμενο της δομικής βιοπληροφορικής είναι η μελέτη τους με σκοπό την κατανόηση του πρωτεϊνικού διπλώματος και των μηχανισμών του, τη μελέτη των εξελικτικών σχέσεων, αλλά και τη μελέτη για τη σχέση δομής και λειτουργίας, οι οποίες οδηγούν στις μεθοδολογίες αγκυροβόλησης ή ελλιμενισμού (docking) για την κατανόηση της δομής συμπλόκων (της εύρεσης δηλαδή της αλληλεπίδρασης δυο πρωτεϊνών ή πρωτεΐνης/DNA (ή RNA), ή πρωτεΐνης/μικρού μορίου, μελέτες που είναι πολύ σημαντικές στο σχεδιασμό φαρμάκων). Φυσικά, ακόμα και στο αρχικό στάδιο του προσδιορισμού της τρισδιάστατης δομής, η συμβολή των υπολογιστικών μεθόδων είναι σημαντική, καθώς η επεξεργασία των δεδομένων της κρυσταλλογραφίας, η κατασκευή χαρτών ηλεκτρονικής πυκνότητας, η προσαρμογή αλλά και η βελτιστοποίηση του μοντέλου, γίνονται με αλγόριθμους και λογισμικό, αλλά λόγω της φύσης αυτού του εγχειριδίου, δεν θα υπεισέλθουμε σε πολλές λεπτομέρειες.

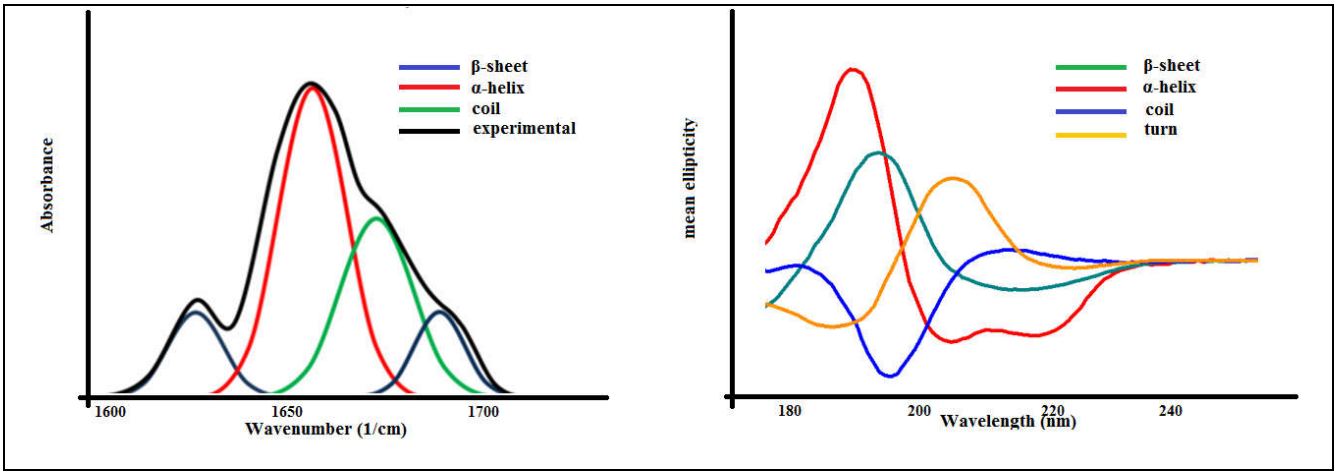
Στις επόμενες ενότητες θα προσπαθήσουμε να διερευνήσουμε τα βασικά μεθοδολογικά θέματα που προκύπτουν στα παραπάνω προβλήματα της δομικής βιοπληροφορικής, να παρουσιάσουμε τα βασικά

εργαλεία λογισμικού που χρησιμοποιούνται για την αντίστοιχη εργασία, αλλά και να προσφέρουμε πρακτικές συμβουλές για την εκτέλεση τέτοιων εργασιών.

9.1. Προσδιορισμός δομής

Το πρώτο βήμα σε κάθε προσπάθεια δομικής βιοπληροφορικής, είναι ο ίδιος ο προσδιορισμός της τρισδιάστατης δομής των μακρομορίων. Αυτές οι μεθοδολογίες ήταν που οδήγησαν τις δεκαετίες του 1950 και του 1960 στη ραγδαία ανάπτυξη της μοριακής βιολογίας, για αυτό και θα κάνουμε μια σύντομη αναφορά, αν και το αντικείμενο αυτό εμπίπτει περισσότερο στον τομέα της δομικής βιολογίας και της βιοφυσικής.

Γενικά, η κατά προσέγγιση σύσταση ενός πολυμερούς, π.χ. μιας πρωτεΐνης σε στοιχεία δευτεροταγούς δομής (π.χ. «η πρωτεΐνη Χ έχει περίπου 40% α-έλικα και 20% β-πτυχωτή επιφάνεια») είναι κάτι που μπορεί να υπολογιστεί με φασματοσκοπία. Για τις πρωτεΐνες, συνηθισμένες μεθοδολογίες φασματοσκοπίας είναι η φασματοσκοπία υπεριώδους (far-UV, 170–250 nm) με κυκλικό διχρωισμό, η φασματοσκοπία υπερύθρου (IR) και η φασματοσκοπία μαγνητικού πυρηνικού συντονισμού (NMR) (Meiler & Baker, 2003; Pelton & McLean, 2000). Σε όλες τις περιπτώσεις, τα διαφορετικά στοιχεία δευτεροταγούς δομής (α-έλικες, β-πτυχωτές επιφάνειες) δίνουν διαφορετικές καμπύλες απορρόφησης, από τις οποίες με υπολογιστική ανάλυση μπορούν να εξαχθούν τα δεδομένα για την εκατοστιαία σύσταση της υπό μελέτης πρωτεΐνης (πάντα με την προϋπόθεση ότι έχουμε καθαρό δείγμα σε κατάλληλο διαλύτη και σε κατάλληλη ποσότητα).



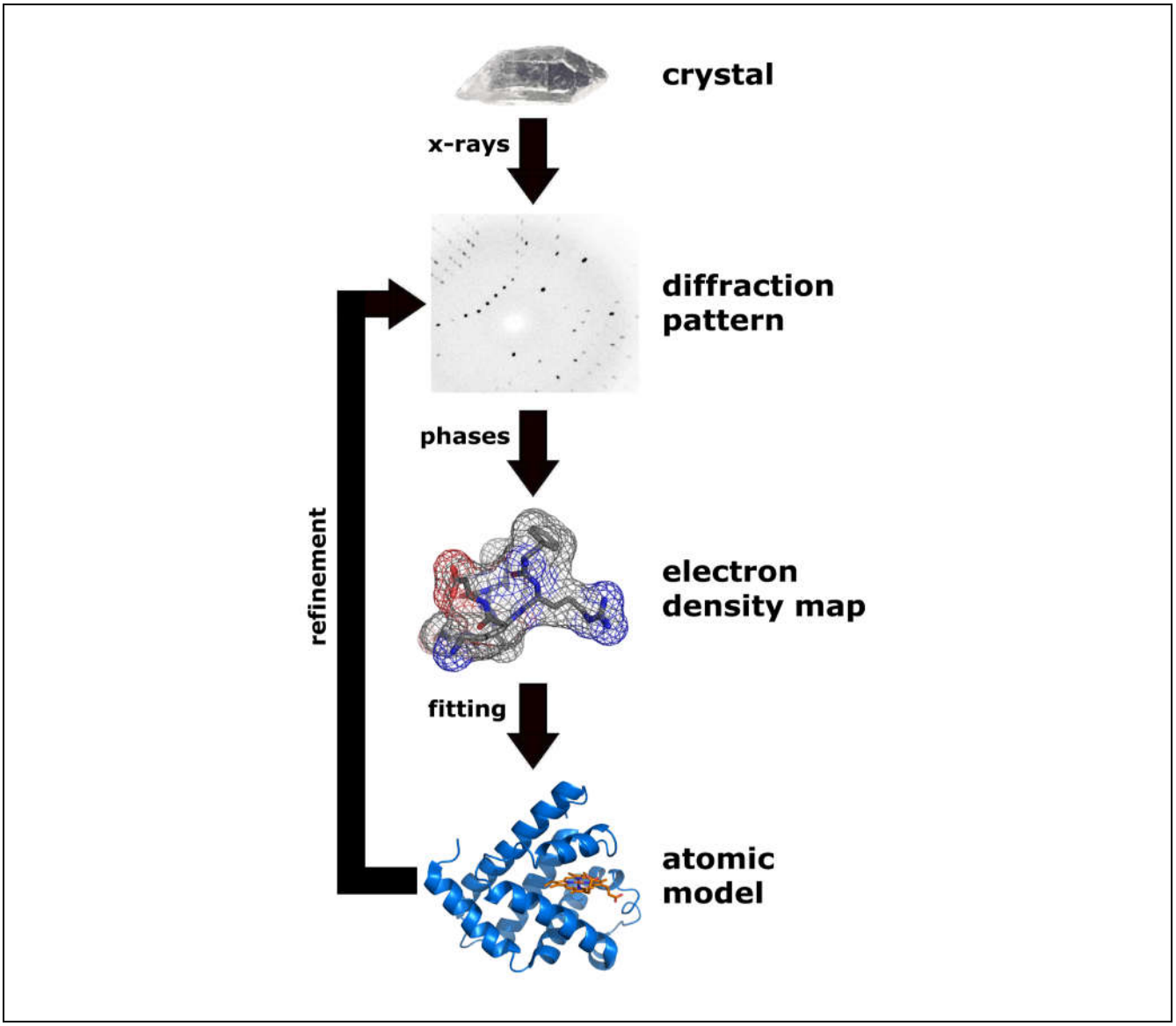
Εικόνα 9.1: *Εικόνα από τη χρήση φασματοσκοπίας για τη διαλεύκανση της δομής πρωτεϊνών (αριστερά φάσμα IR – δεξιά φάσμα CD)*

Η μεθοδολογία της κρυσταλλογραφίας ακτίνων Χ είχε αρχίσει να αναπτύσσεται από τις δεκαετίες πριν τον 2^{ο} Παγκόσμιο Πόλεμο, αλλά οι πρώτες τρισδιάστατες δομές πρωτεϊνών (της μυογλοβίνης και της αιμοσφαιρίνης), επιλύθηκαν προς τα τέλη της δεκαετίας του 1950, λίγο μετά την εύρεση της δομής του DNA. Για τις εργασίες τους αυτές, ο Sir John Kendrew και ο Max Perutz μοιράστηκαν το βραβείο Νόμπελ το 1962, ενώ από τότε αρκετά άλλα βραβεία Νόμπελ έχουν δοθεί σε επιστήμονες που προσδιόρισαν δομές σημαντικών πρωτεϊνών. Η βασική αρχή της κρυσταλλογραφίας ακτίνων Χ βασίζεται στην ίδια αρχή με τους μεγεθυντικούς φακούς και το μικροσκόπιο. Η μεγάλη διαφορά με το οπτικό αλλά και το ηλεκτρονικό μικροσκόπιο, έγκειται στο γεγονός ότι επιθυμούμε να «δούμε» σε ατομική διακριτικότητα, δηλαδή να μπορούμε να ξεχωρίσουμε αντικείμενα με απόσταση λίγα Å. Αυτό σημαίνει ότι η προσπίπτουσα ακτινοβολία πρέπει να έχει ένα μήκος κύματος που την τοποθετεί στο φάσμα των ακτίνων Χ, αλλά οι ακτίνες Χ δεν μπορούν να εστιαστούν με τη χρήση φακών (όπως για παράδειγμα κάνουμε στο οπτικό μικροσκόπιο με τα φωτόνια ή στο ηλεκτρονικό μικροσκόπιο με τα ηλεκτρόνια). Έτσι, πρέπει να μελετήσουμε το πρότυπο περίθλασης των ακτίνων Χ από το δείγμα για να μπορέσουμε να καταλήξουμε σε ένα συμπέρασμα για τη δομή του δείγματος.

Από τη δεκαετία του 1990 και μετά, οι μέθοδοι ετερόλογης έκφραση πρωτεϊνών, η διαθεσιμότητα επιταχυντών ηλεκτρονίων για παραγωγή ακτίνων Χ (συγχροτρονική ακτινοβολία), οι ανιχνευτές περιοχής τύπου CCD αρχικά και απευθείας μέτρησης φωτονίων πιο πρόσφατα, αλλά και η αύξηση της υπολογιστικής ισχύος, οδήγησαν στην εκθετική αύξηση του αριθμού των διαθέσιμων τρισδιάστατων δομών. Σήμερα υπάρχουν πλέον διαθέσιμες πάνω από εκατό χιλιάδες διαθέσιμες δομές στην PDB (βέβαια, παρ' όλα αυτά οι

διαθέσιμες αλληλουχίες, είναι μια τάξης μεγέθους περισσότερες, οπότε το χάσμα ανάμεσα στον αριθμό των δομών και αυτόν των αλληλουχιών συνεχίζει να αυξάνει). Η κρυσταλλογραφία των ακτίνων Χ εξακολουθεί φυσικά να είναι η πιο κοινή μέθοδος εύρεσης τρισδιάστατης δομής, αν και την παρούσα χιλιετία η φασματοσκοπία NMR συμμετέχει σταθερά με περίπου 10% των νέων δομών (κυρίως μικρών πρωτεϊνών) που προσδιορίζονται πειραματικά. Την τελευταία διετία (2013-2015) συντελείται μια επανάσταση στο χώρο της ηλεκτρονικής μικροσκοπίας, που μπορεί πλέον να προσδιορίσει δομές σε ευκρίνεια 2-3Å, με τη χρήση νέων μικροσκοπίων αλλά κυρίως νέων τεχνολογιών ανιχνευτών περιοχής ηλεκτρονίων και μεγάλης υπολογιστικής ισχύος (εκατοντάδες επεξεργαστές, terabyte δεδομένων και δεκάδες gigabyte μνήμης).

Η παλιότερη αλλά και πιο ακριβής μέθοδος κρυσταλλογραφίας ακτίνων Χ είναι αυτή περίθλασης ακτίνων Χ μονοκρυστάλλου (single-crystal X-ray diffraction), στην οποία μία δέσμη από ακτίνες Χ προσκρούει στον κρύσταλλο και παράγει μια σειρά ανακλάσεις οι οποίες καταγράφονται από κάποιον ανιχνευτή. Η ένταση και η γωνία των ανακλάσεων καταγράφεται καθώς ο κρύσταλλος περιστρέφεται. Αν ο κρύσταλλος είναι επαρκούς καθαρότητας και κανονικότητας, τα δεδομένα από την περίθλαση επιτρέπουν τον προσδιορισμό των αποστάσεων των χημικών δεσμών και των γωνιών τους με μεγάλη ακρίβεια (Shi, 2014; Yaffe, 2005).



Εικόνα 9.2:*Σχηματική αναπαράσταση της διαδικασίας προσδιορισμού δομής με κρυσταλλογραφία ακτίνων Χ (https://en.wikipedia.org/wiki/X-ray_crystallography)*

Τα βασικά στάδια της κρυσταλλογραφίας είναι τρία:

- Το πρώτο και συχνά πιο δύσκολο στάδιο είναι η ανάπτυξη ενός κατάλληλου κρυστάλλου για την πρωτεΐνη ή το σύμπλοκο που μελετάται. Ο κρύσταλλος πρέπει να είναι αρκετά μεγάλος (τυπικά μεγαλύτερος από 10-20μm), με καθαρή σύσταση χωρίς προσμίξεις και χωρίς εσωτερικές ατέλειες (σπασίματα, κ.ο.κ.). Κατά συνέπεια, οι πρωτεΐνες πρέπει να απομονωθούν από το δείγμα, να καθαριστούν και να υπάρχουν σε μεγάλη ποσότητα. Μια επιπλέον δυσκολία εντοπίζεται στο γεγονός ότι δεν κρυσταλλώνονται όλες οι πρωτεΐνες στις ίδιες συνθήκες. Έτσι, ακόμα και αν υπάρχει το δείγμα, η κρυστάλλωση είναι μια επίπονη διαδικασία με πολύ πειραματισμό σε διαφορετικές συνθήκες. Υπάρχουν δυο κύριες μεθοδολογίες για την κρυστάλλωση (διάχυση ατμών και διαπίδυση). Νέες τεχνολογίες ελεύθερων επιταχυντών ηλεκτρονίων λέιζερ επιτρέπουν τη χρήση κρυστάλλων μικρότερων από 1 μm.
- Στο δεύτερο στάδιο, ο κρύσταλλος τοποθετείται απέναντι από τη δέσμη των ακτίνων Χ, οι οποίες συνήθως είναι μονοχρωματικές, και κατόπιν συλλέγονται οι ανακλάσεις. Καθώς ο κρύσταλλος περιστρέφεται δημιουργούνται πολλαπλά σύνολα ανακλάσεων καλύπτοντας διαφορετική γωνία έκθεσης στις ακτίνες Χ. Συνολικά έτσι συλλέγονται εκατοντάδες χιλιάδες ανακλάσεις. Πολλές φορές, αν ο κρύσταλλος είναι μικρός ή με μειωμένη κανονικότητα μπορεί να καταστραφεί κατά τη διαδικασία αυτή, προτού συλλεχθούν όλα τα απαραίτητα δεδομένα. Ένας άλλος περιοριστικός παράγοντας σε αυτό το στάδιο είναι η πηγή των ακτίνων Χ, καθώς απαιτείται συνήθως κάποια συσκευή υψηλής ενέργειας, όπως το σύγχροτρο. Κάθε ανάκλαση συλλέγεται αρκετές φορές, και στατιστικές μέθοδοι χρησιμοποιούνται για τη μέτρηση της μέσης τιμής και της τυπικής απόκλισης των μετρήσεων.
- Στο τρίτο στάδιο, τα δεδομένα των ανακλάσεων συνδυάζονται με συμπληρωματικά χημικά δεδομένα για να παραχθεί ένα αρχικό μοντέλο και να βελτιστοποιηθεί στη συνέχεια. Το βασικό πρόβλημα εδώ, είναι ότι από τις ανακλάσεις δεν μπορεί να προσδιοριστεί μονοσήμαντα η θέση των ατόμων και ο χάρτης ηλεκτρονικής πυκνότητας στο δείγμα (το γνωστό πρόβλημα φάσης). Το πρόβλημα αυτό λύνεται με μια σειρά από μεθόδους όπως η εύρεση φάσης εκ του μηδενός (ab initio phasing), η μοριακή αντικατάσταση (molecular replacement), η ανώμαλη σκέδαση ακτίνων Χ (anomalous X-ray scattering) και οι μέθοδοι βαρέων ατόμων (heavy atom methods), έτσι ώστε τελικά να προκύπτει μια αρχική εκτίμηση για τη φάση. Τα επόμενα βήματα, αφορούν την κατασκευή ενός αρχικού μοντέλου και τη βελτιστοποίησή του (refinement).

Οι διαδικασίες και των τριών βημάτων απαιτούν ιδιαίτερη χρήση Η/Υ και αλγορίθμων. Στην κρυστάλλωση, οι Η/Υ χρησιμοποιούνται για τον έλεγχο ρομποτικών μονάδων κρυστάλλωσης, την ανάλυση εικόνων από πειράματα κρυστάλλωσης, αλλά και το σχεδιασμό των βέλτιστων συνθηκών για τα πειράματα κρυστάλλωσης. Στο δεύτερο στάδιο χρησιμοποιούνται ειδικά πακέτα ανάλυσης των εικόνων περίθλασης, με έμφαση στον προσδιορισμό των σφαλμάτων μέτρησης, που είναι απαραίτητα ειδικά στην επίλυση του προβλήματος των φάσεων. Υπάρχουν διάφορα πακέτα λογισμικού δομικής βιολογίας που διευκολύνουν τις διαδικασίες αυτές. Στην ιστοσελίδα της International Union of Crystallography αναφέρονται δεκάδες τέτοια πακέτα, από τα οποία κάποια είναι συλλογές με πολλαπλές χρήσεις και άλλα συγκεκριμένες ρουτίνες με εστιασμένο ενδιαφέρον (http://www.iucr.org/resources/other-directories/software). Γενικά πάντως, τα πακέτα με τη μεγαλύτερη αποδοχή, τους περισσότερους χρήστες και τις περισσότερες λειτουργίες, είναι το **CCP4,** το οποίο είναι διαθέσιμο στη διεύθυνση http://www.ccp4.ac.uk/ (Winn et al., 2011), το **PHENIX**, το οποίο είναι διαθέσιμο στη διεύθυνση https://www.phenix-online.org/ (Adams et al., 2010) και το **X-PLOR** (Güntert, 2011), το οποίο είναι ένα από τα παραδοσιακά πακέτα στον τομέα, μαζί με την βελτιωμένη του έκδοση που συντηρείται από τον NIH, το **Xplor-NIH**, το οποίο είναι διαθέσιμο στη διεύθυνση http://nmr.cit.nih.gov/xplor-nih/ (Schwieters, Kuszewski, & Clore, 2006)

Αν όλα τα στάδια στεφθούν με επιτυχία, έχουμε τελικά μια τρισδιάστατη δομή, δηλαδή ένα αρχείο με ατομικές συντεταγμένες που αντιστοιχεί στη δομή της πρωτεΐνης στο δείγμα. Συνήθως τα αρχεία αυτά κατατίθενται σε δημόσιες βάσεις δεδομένων (PDB), καθώς εδώ και χρόνια είναι υποχρεωτική η κατάθεσή τους προκειμένου οι αντίστοιχες εργασίες που τα περιγράφουν να γίνουν δεκτές προς δημοσίευση. Εκεί, πολλές φορές τα δεδομένα αυτά περνάνε και άλλους αυτοματοποιημένους ελέγχους για να αποκλειστεί το ενδεχόμενο σφάλματος και να διασφαλιστεί η ποιότητα. Στην ιστοσελίδα της PDB παρατίθεται μια μεγάλη λίστα με διαθέσιμα προγράμματα για τον έλεγχο, την αξιολόγηση και την επαλήθευση τρισδιάστατων δομών

(http://www.rcsb.org/pdb/static.do?p=software/software_links/analysis_and_verification.html). Το πρωτοπόρο πρόγραμμα σε αυτόν τον τομέα ήταν το **PROCHECK** (Laskowski, MacArthur, Moss, & Thornton, 1993). Τα προγράμματα που χρησιμοποιούνται πλέον ευρέως για αξιολόγηση και επαλήθευση ειναι το **MolProbity** το οποίο είναι διαθέσιμο στη διεύθυνση http://molprobity.biochem.duke.edu/ (Chen et al., 2010) και το **WHATCHECK** το οποίο διατίθεται στη διεύθυνση http://swift.cmbi.ru.nl/gv/whatcheck/ (Hooft, Vriend, Sander, & Abola, 1996). Η επαλήθευση των δομών είναι πλέον απαραίτητη για την δημοσίευσή τους και υπάρχουν συγκριμένες οδηγίες για αυτό τον σκοπό (Read et al., 2011). Μια νέα αντιμετώπιση, είναι και η λογική της «ενεργούς επαλήθευσης» όπου οι υπάρχουσες δομές διορθώνονται με αυτοματοποιημένους αλγόριθμους μοντελοποίησης με βάση τα κρυσταλλογραφικά δεδομένα που κατατίθενται στην PDB (Joosten et al., 2009).

Ένας εναλλακτικός τρόπος προσδιορισμού της δομής, είναι η φασματοσκοπία πυρηνικού μαγνητικού συντονισμού (NMR). Το NMR εκμεταλλεύεται τις μηχανικές ιδιότητες των ατόμων, οι οποίες εξαρτώνται από το περιβάλλον και με αυτόν τον τρόπο παράγει τελικά ένα χάρτη που απεικονίζει τον τρόπο με τον οποίο συνδέονται τα άτομα, πόσο κοντά βρίσκονται στο χώρο, και ποια είναι η σχετική τους κίνηση. Οι ιδιότητες αυτές είναι στην ουσία ίδιες με τη μεθοδολογία του μαγνητικού πυρηνικού συντονισμού (Magnetic Resonance Imaging - MRI), αλλά εδώ εστιαζόμαστε σε αποστάσεις της τάξης του Å, σε αντίθεση με τα mm που αποτελούν το αντικείμενο μελέτης των του MRI. Επίσης, μια άλλη διαφορά είναι ότι εδώ δεν παράγεται απευθείας μια εικόνα, αλλά τα δεδομένα συλλέγονται και με χρήση υπολογιστή κατασκευάζεται ένα τρισδιάστατο μοντέλο της πρωτεΐνης. Στις περισσότερες περιπτώσεις, τα δείγματα βρίσκονται σε υδατικό διάλυμα, αλλά αναπτύσσονται και μεθοδολογίες στερεάς φάσης. Η συλλογή των δεδομένων γίνεται με την τοποθέτηση του δείγματος σε έναν δυνατό μαγνήτη, τη χρήση ραδιοκυμάτων στο δείγμα και τη συλλογή του φάσματος απορρόφησης. Ανάλογα με το περιβάλλον (τόσο του διαλύτη, αλλά και των γειτονικών ατόμων), οι πυρήνες των ατόμων θα απορροφήσουν τα κύματα σε διαφορετικές συχνότητες και οι πληροφορίες αυτές μπορεί να συνδυαστούν με σκοπό να καθοριστεί ένα συνολικό μοντέλο του μορίου. Γενικά, επειδή το δείγμα βρίσκεται σε υδατικό διάλυμα και συνυπολογίζονται ταυτόχρονα οι κινήσεις όλων των ατόμων, η μέθοδος μπορεί να εφαρμοστεί κυρίως σε μικρές πρωτεΐνες (αν και υπάρχουν εξαιρέσεις). Επιπλέον, η τεχνική αυτή λειτουργεί συμπληρωματικά με την κρυσταλλογραφία, καθώς είναι περισσότερο χρήσιμη στη μελέτη της κίνησης και της δυναμικής (ευελιξία, κλπ) των πρωτεϊνικών μορίων, σε αλληλεπιδράσεις μεταξύ πρωτεϊνών με άλλες πρωτεΐνες αλλά και μικρά μόρια (φάρμακα, μεταβολίτες κ.ο.κ.), αλλά και σε περιπτώσεις πρωτεϊνών που δεν μπορούν να κρυσταλλωθούν εύκολα.

Ένα σημαντικό θέμα που πρέπει να αναφερθεί, είναι ο τρόπος καθορισμού της δευτεροταγούς δομής από τα δεδομένα κρυσταλλογραφίας. Παραδοσιακά, οι κρυσταλλογράφοι παρατηρούσαν τις δομές και οπτικά αποφάσιζαν ποιες περιοχές ήταν σε α-έλικα, ποιες σε β-πτυχωτή επιφάνεια κ.ο.κ. Επειδή όμως οι αναθέσεις αυτές, ήταν υποκειμενικές και πολλές φορές προκύπταν διαφωνίες ακόμα και μεταξύ έμπειρων κρυσταλλογράφων, αναπτύχθηκαν αυτοματοποιημένοι αλγόριθμοι οι οποίοι διαβάζουν το αρχείο με τις τρισδιάστατες συντεταγμένες και αποδίδουν όσο πιο αντικειμενικά γίνεται τα στοιχεία δευτεροταγούς δομής, καθώς και άλλα χαρακτηριστικά όπως την προσβασιμότητα των διαφόρων καταλοίπων (δηλαδή, αν είναι εκτεθειμένα ή όχι). Αυτό που πρέπει να τονιστεί, είναι ότι οι αλγόριθμοι αυτοί δεν είναι αλγόριθμοι πρόγνωσης της δευτεροταγούς δομής, δεν κάνουν δηλαδή πρόβλεψη σε κάποια αλληλουχία άγνωστης δομής, απλά εντοπίζουν σε μια προσδιορισμένη τρισδιάστατη δομή το σημείο που βρίσκονται οι α-έλικες και οι β-πτυχωτές επιφάνειες, κάνοντας χρήση αντικειμενικών κριτηρίων.

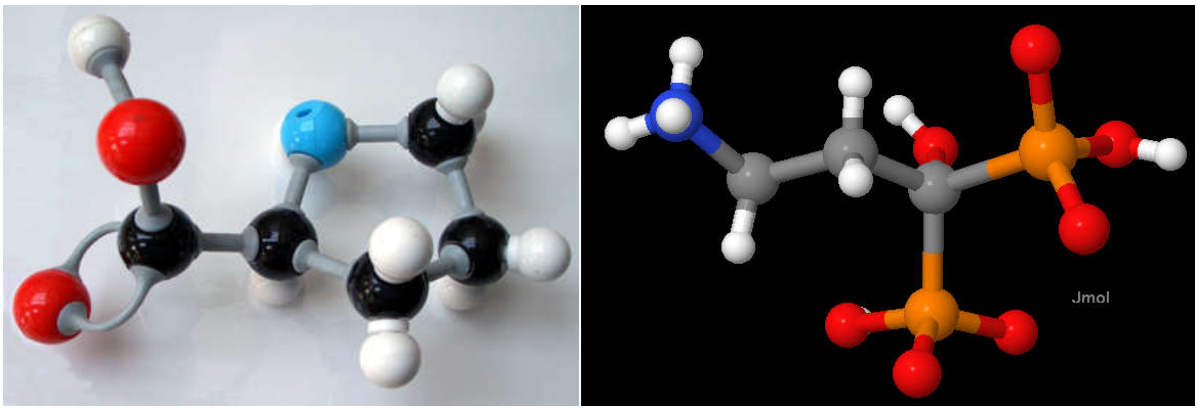
Το **DSSP** (Define Secondary Structure of Proteins), διαθέσιμο στη διεύθυνση http://swift.cmbi.ru.nl/gv/dssp/, ήταν ο πρώτος αλγόριθμος που προτάθηκε για το σκοπό αυτό και είναι ακόμα ο ευρύτερα χρησιμοποιούμενος (Kabsch & Sander, 1983). Το DSSP αναγνωρίζει τον κύριο ανθρακικό σκελετό της πρωτεΐνης και εντοπίζει τους δεσμούς υδρογόνου που σχηματίζονται, με βάση έναν καθαρά ηλεκτροστατικό ορισμό. Με βάση τον ενεργειακό υπολογισμό, το DSSP αναγνωρίζει και κατατάσσει τα στοιχεία δευτεροταγούς δομής σε 8 κατηγορίες. Η 3_{10}-έλικα, η α-έλικα, και η π-έλικα (με σύμβολα G, H και I) αναγνωρίζονται αν υπάρχουν συνεχόμενες επαναλήψεις του δεσμού υδρογόνου με βήμα 3, 4, ή 5 κατάλοιπα αντίστοιχα. Οι β-δομές χωρίζονται σε β-πτυχωτή επιφάνεια (E) και β-γέφυρα (B), το σύμβολο T χρησιμοποιείται για τις στροφές και το S για περιοχές υψηλής καμπυλότητας. Τέλος, περιοχές που δεν ταιριάζουν με κανένα πρότυπο μένουν με το κενό σύμβολο. Συνήθως στις παρακάτω αναλύσεις, όπως π.χ. στην πρόγνωση δευτεροταγούς δομής τα σύμβολα αυτά ομαδοποιούνται και αυτό μπορεί να γίνει με δύο τρόπους. Στην πρώτη περίπτωση α-έλικα μένει το H, β-πτυχωτή επιφάνεια το E και όλα τα άλλα γίνονται τυχαία δομή (coil) με σύμβολο το C. Ο εναλλακτικός τρόπος περιλαμβάνει την ομαδοποίηση στο H και των άλλων ελίκων (G, I), στο E την προσθήκη του B, ενώ τα υπόλοιπα γίνονται C. Το 2002 μια νεότερη έκδοση

του DSSP εμφανίστηκε η οποία πραγματοποιεί ανάθεση με πιο ευέλικτα όρια (continuous DSSP) η οποία φαίνεται να προσφέρει κάποια επιπλέον πλεονεκτήματα (Andersen, Palmer, Brunak, & Rost, 2002).

Το **STRIDE** (STRuctural IDEntification), το οποίο είναι διαθέσιμο στη διεύθυνση http://webclu.bio.wzw.tum.de/stride/ είναι ένας άλλος εναλλακτικός αλγόριθμος για τον προσδιορισμό και την ανάθεση των στοιχείων δευτεροταγούς δομής (Frishman & Argos, 1995). Το STRIDE χρησιμοποιεί μια παρόμοια μέθοδο με το DSSP, καθώς εφαρμόζει μια μέτρηση ενέργειας για τον προσδιορισμό των δεσμών υδρογόνου (ένα δυναμικό Lennard-Jones), αλλά επιπλέον λαμβάνει υπόψη του και τις δίεδρες γωνίες που σχηματίζονται. Στο τέλος, αναθέτει δευτεροταγείς δομές στις ίδιες κατηγορίες που χρησιμοποιεί το DSSP, αλλά επιπλέον δίνει και μια ανά κατάλοιπο τιμή για την αξιοπιστία της ανάθεσης, η οποία έχει προκύψει από εμπειρικές μελέτες. Παρόλο που το DSSP είναι το πιο παλιό και ευρύτερα αποδεκτό πρόγραμμα, το STRIDE πιστεύεται ότι είναι σχετικά καλύτερο και διορθώνει την τάση του DSSP να ορίζει κάπως μικρότερα τμήματα δευτεροταγούς δομής σε σχέση με τους ορισμούς που πραγματοποιούν οι έμπειροι κρυσταλλογράφοι. Τα τελευταία χρόνια, το STRIDE χρησιμοποιείται και στην PDB (παράλληλα με το DSSP), ενώ υπάρχει και διαδικτυακή εφαρμογή διαθέσιμη για άμεση χρήση από το ευρύ κοινό (Heinig & Frishman, 2004).

9.2. Οπτικοποίηση βιολογικών δομών

Δεδομένης της ύπαρξης της τρισδιάστατης δομής μιας πρωτεΐνης, το πρώτο πράγμα που θα ενδιέφερε κάποιον θα ήταν η οπτικοποίηση. Οι μεθοδολογίες απεικόνισης των μακρομορίων, ξεκίνησαν παράλληλα με τις πρώτες επιτυχίες της κρυσταλλογραφίας ακτίνων Χ από την δεκαετία του 1950 και 1960. Αρχικά, για τα μοντέλα αυτά χρησιμοποιήθηκαν ξύλινες σφαίρες για να αναπαραστήσουν τα άτομα και ράβδοι για να αναπαραστήσουν τους δεσμούς. Τέτοια μοντέλα, χρησιμοποιούνται ακόμα και σήμερα για εκπαιδευτικούς λόγους αλλά είναι πλέον πλαστικά και με διαφορετικό χρωματισμό για τα διαφορετικά είδη ατόμων. Αυτή η αναπαράσταση, γίνεται πλέον και σε υπολογιστές και ονομάζεται «ball and stick». Μια παρόμοια αναπαράσταση, είναι και το σκελετικό μοντέλο (wireframe) στο οποίο απεικονίζονται όμως μόνο οι δεσμοί ως σύρματα, ενώ τα άτομα συμπίπτουν με τις κορυφές (γωνίες). Τέτοιες αναπαραστάσεις είναι φυσικά πολύ απλές και μπορούν εύκολα να πραγματοποιηθούν σε υπολογιστή αλλά κάποιες φορές είναι ιδιαίτερα χρήσιμες. Πολλές φορές μάλιστα για λόγους απλότητας αναπαρίστανται μόνο οι Cα (τα ασύμμετρα άτομα άνθρακα, δηλαδή ο κύριος σκελετός της πρωτεΐνης) ενώ άλλες φορές αναπαρίστανται όλα τα άτομα με διαφορετικό χρωματισμό.

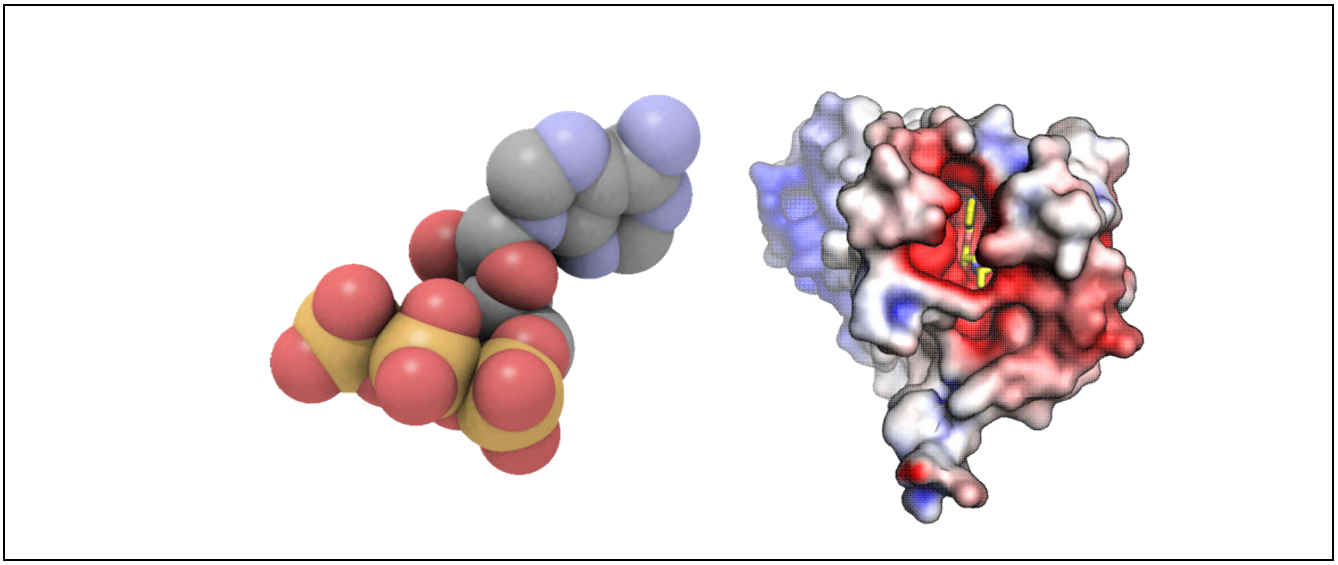


Εικόνα 9.3: *Αριστερά, πλαστικό μοντέλο τύπου «ball-and-stick» (https://en.wikipedia.org/wiki/Molecular_model). Δεξιά, μοντέλο ενός μικρού μορίου ($NH_3CH_2CH_2C(OH)(PO_3H)(PO_3H)$-), όπως παράγεται από το Jmol (https://en.wikipedia.org/wiki/Molecular_graphics)*

Ένας άλλος τρόπος αναπαράστασης, είναι το λεγόμενο χωροπληρωτικό μοντέλο (space-filling model), στο οποίο τα άτομα αναπαρίστανται πάλι με σφαίρες, συνήθως διαφορετικού χρώματος, αλλά με τη σημαντική προσθήκη ότι οι ακτίνες της κάθε σφαίρας είναι ανάλογες με την ακτίνα van der Waals του ατόμου. Τα μοντέλα αυτά ονομάζονται και CPK models από τους Corey, Pauling, και Koltun, οι οποίοι ανέπτυξαν πρώτοι τέτοιες τεχνικές απεικόνισης, ενώ πλαστικά μοντέλα αυτού του είδους χρησιμοποιούνται ακόμα και σήμερα για διδακτικούς σκοπούς. Καθώς οι ακτίνες των ατόμων είναι μικρότερες από την ενδομοριακή απόσταση όταν τα άτομα τα ενώνει ομοιοπολικός δεσμός, οι σφαίρες πρέπει να τέμνονται, και κατά συνέπεια στα πλαστικά μοντέλα αυτού του είδους οι σφαίρες είναι κολοβές καθώς αφαιρείται μια

περιοχή σαν «καπάκι» έτσι ώστε τα άτομα να έρχονται σε επαφή. Τα μοντέλα αυτά είναι πιο ρεαλιστικά, καθώς δίνουν μια απεικόνιση της επιφάνειας του μορίου που βρίσκεται πιο κοντά στην πραγματικότητα. Δεν δίνουν όμως καλή εικόνα της δευτεροταγούς δομής ή της κατεύθυνσης της πολυπεπτιδικής αλυσίδας. Κατά συνέπεια είναι πιο χρήσιμα σε δυναμικές μελέτες, π.χ. για τον υπολογισμό της έκθεσης στο διαλύτη ή για τον υπολογισμό επιφανειών επαφής και μελέτες αγκυροβόλησης. Μια συνηθισμένη παραλλαγή αυτών των μοντέλων είναι αυτή στην οποία η επιφάνεια εμφανίζεται πιο ομαλή και τα άτομα έχουν χρωματιστεί ανάλογα με την ηλεκτραρνητικότητα (κόκκινο) ή την ηλεκτροθετικότητά τους (μπλε).

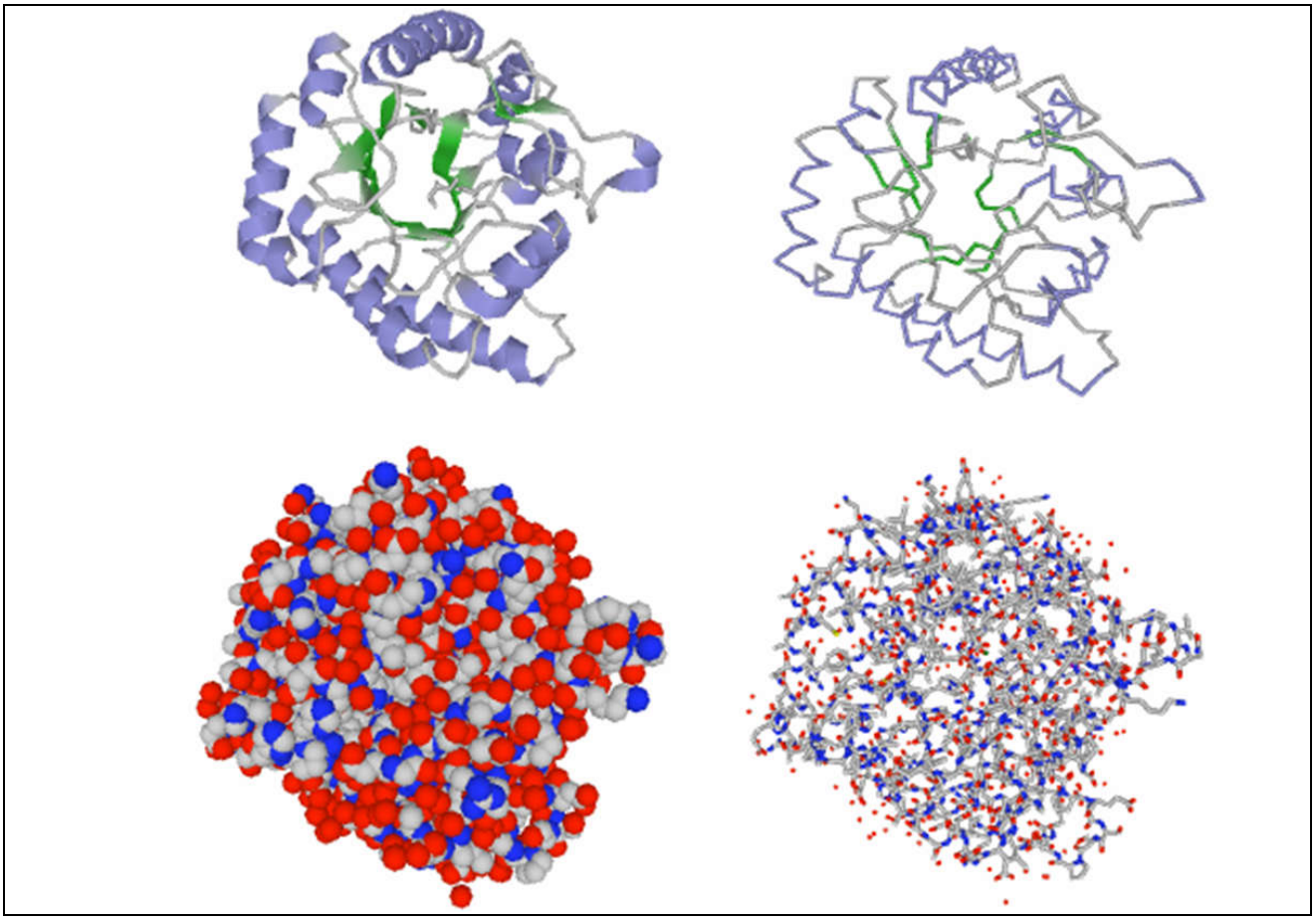
Τέλος, ένας άλλος διαδεδομένος τρόπος αναπαράστασης -που έγινε πολύ διαδεδομένος με τη χρήση H/Y-, είναι το λεγόμενο διάγραμμα κορδέλα (Ribbon diagram) ή καρτούν, το οποίο είναι ιδιαίτερα πληροφοριακό και είναι ίσως από τις δημοφιλέστερες μεθόδους αναπαράστασης. Σε ένα τέτοιο διάγραμμα απεικονίζεται ολόκληρος ο σκελετός της πρωτεΐνης και η κατεύθυνση της πολυπεπτιδικής αλυσίδας και σχεδιάζονται με ειδικό τρόπο τα στοιχεία δευτεροταγούς δομής. Έτσι, οι ά-έλικες αναπαρίστανται σαν κορδέλες (ribbon) που σχηματίζουν ελικοειδή διάταξη ή σαν κύλινδροι, ενώ οι β-πτυχωτές επιφάνειες σαν πεπλατυσμένα βέλη. Τέλος, οι περιοχές με μη κανονική δευτεροταγή δομή, απεικονίζονται σαν καμπυλωτές γραμμές. Επιπλέον δε, τα διαφορετικά στοιχεία δευτεροταγούς δομής χρωματίζονται συνήθως με διαφορετικό τρόπο έτσι ώστε να είναι πιο εύκολη η διάκρισή τους. Η μέθοδος αυτή είναι πολύ πληροφοριακή, γιατί βλέπουμε αμέσως τη διάταξη των στοιχείων δευτεροταγούς δομής και τις μεταξύ τους σχέσεις, αλλά και την κατεύθυνση της αλυσίδας. Για αυτούς τους λόγους τα μοντέλα αυτά χρησιμοποιούνται στις περισσότερες περιπτώσεις στις επιστημονικές δημοσιεύσεις.



Εικόνα 9.4: *Αριστερά, ένα παράδειγμα χωροπληρωτικού μοντέλου του ATP. Δεξιά, ένα παράδειγμα χωροπληρωτικού μοντέλου του β2 αδρενεργικού υποδοχέα, (PDB code 2RH1, από https://en.wikipedia.org/wiki/Space-filling_model)*

Σήμερα, υπάρχουν διαθέσιμα δεκάδες προγράμματα για μοριακή απεικόνιση τρισδιάστατων δομών. Τα περισσότερα από αυτά είναι ανοιχτού κώδικα και διανέμονται δωρεάν, άλλα λειτουργούν σαν αυτόνομες εφαρμογές ενώ άλλα λειτουργούν σαν πρόσθετα στον περιηγητή ιστού. Όλα, δέχονται σαν είσοδο ένα αρχείο PDB το οποίο αποτελεί το αποδεκτό πρότυπο για τέτοιου είδους δεδομένα. Τα προγράμματα αυτά, διαθέτουν πλέον πάρα πολλές λειτουργίες και ο κάθε χρήστης μπορεί να βρει κάποιο που να καλύπτει τις ανάγκες του (κάποιος μπορεί να ενδιαφέρεται για την απλότητα, κάποιος για τις υπολογιστικές απαιτήσεις, κάποιος για μια συγκεκριμένη λειτουργία που ένα δεδομένο πρόγραμμα επιτελεί καλύτερα κ.ο.κ.). Τα περισσότερα προγράμματα πάντως, παρέχουν δυνατότητες αναπαράστασης με όλα τα παραπάνω μοντέλα. Δίνουν την επιλογή να επιλέξει ο χρήστης το χρωματισμό που επιθυμεί, ενώ είναι και διαδραστικά καθώς επιτρέπουν στο χρήστη να περιστρέψει το μόριο, να μεγεθύνει σε κάποιο σημείο, να επιλέξει κάποια κατάλοιπα, να τα χρωματίσει διαφορετικά αλλά και να επιτρέψει διαφορετικό τρόπο αναπαράστασης για κάποια επιλεγμένα κατάλοιπα. Το πώς θα τα χρησιμοποιήσει ο κάθε χρήστης, διαφέρει και εξαρτάται από τις ανάγκες του. Για παράδειγμα, κάποιος που ενδιαφέρεται να πάρει μια εικόνα για το δίπλωμα της πρωτεΐνης και το γενικότερο σχήμα της, συνήθως θα επιλέξει ένα διάγραμμα ribbon. Κάποιος που θέλει να δει την τεταρτοταγή δομή, θα επιλέξει διαφορετικό χρωματισμό στις διαφορετικές πολυπεπτιδικές αλυσίδες, ενώ κάποιος που θέλει να μελετήσει τη λειτουργία ενός ενζύμου, θα εστιαστεί στο ενεργό κέντρο και θα χρησιμοποιήσει διαγράμματα

wireframe ή ball and stick και θα χρωματίσει διαφορετικά τα διάφορα άτομα. Τέλος, πολλά από τα προγράμματα αυτά παρέχουν επιπλέον λειτουργίες δομικής βιοπληροφορικής, από υπολογισμό αποστάσεων ατόμων και υπολογισμό φορτίων και επιφανειών, μέχρι και λειτουργίες δομικής στοίχισης (βλ. παρακάτω).



Εικόνα 9.5: *Διαφορετικές αναπαραστάσεις του ίδιου μορίου μπορούν να χρησιμοποιηθούν σε διαφορετικές περιστάσεις και με διαφορετικό σκοπό. Βλέπουμε εδώ τη δομή της Ισομεράσης της Ξυλόζης από τον Planctomyces limnophilus (PDB code 3TVA). Οι εικόνες δημιουργήθηκαν με το PV.*

Στην ιστοσελίδα της PDB παρατίθεται μια μεγάλη λίστα από τέτοια προγράμματα τα οποία καλύπτουν όλες τις ανάγκες (http://www.rcsb.org/pdb/static.do?p=software/software_links/molecular_graphics.html). Η ίδια η PDB έχει ενσωματώσει μια σειρά από τέτοια εργαλεία στη διαδικτυακής της πλατφόρμα με σκοπό ο απλός χρήστης να μπορεί να οπτικοποιήσει αμέσως τις δομές για τις οποίες έχει κάνει αναζήτηση και να δει με διαδραστικό τρόπο τα αποτελέσματα. Τα εργαλεία αυτά ποικίλουν από το απλό **RCSB Simple Viewer (**http://biojava.org/wiki/RCSB_Viewers:About**),** το οποίο βασίζεται στην τεχνολογία Java Web Start και δίνει μια βασική διαδραστικότητα με λειτουργίες του ποντικιού, μέχρι το **Jmol** (http://jmol.sourceforge.net/), το οποίο είναι εφαρμογή Java Applet, και το **Jsmol** το οποίο είναι η ειδική έκδοση του τελευταίου και χρησιμοποιεί JavaScript και HTML5 (http://sourceforge.net/projects/jsmol/). Και τα δύο τελευταία εργαλεία προσφέρουν πολλές λειτουργικότητες και ευκολίες ακόμα και στον πεπειραμένο χρήστη. Υπάρχει ακόμα και το **PV** το οποίο βασίζεται στην τεχνολογία WebGL και παρέχει τις βασικές λειτουργίες με ένα ιδιαίτερα εύχρηστο μενού επιλογών.

Άλλη παρόμοια εφαρμογή που μπορεί να χρησιμοποιήσει κάποιος είναι το **RasMol** (http://www.bernstein-plus-sons.com/software/rasmol/), το οποίο είναι από τις πιο παλιές εφαρμογές για μοριακή απεικόνιση, η οποία εξελίσσεται συνεχώς και έχει αποκτήσει και έκδοση ανοιχτού κώδικα το **OpenRasMol** (http://www.openrasmol.org/). Το πακέτο CCP4 που αναφέραμε παραπάνω, περιέχει και τη δική του αντίστοιχη εφαρμογή, το **CCP4mg** (http://www.ccp4.ac.uk/MG/), ενώ υπάρχουν και άλλες διαδραστικές εφαρμογές που κατασκευάστηκαν ως συμπληρωματικά εργαλεία άλλων διαδικτυακών τόπων και εφαρμογών, όπως για παράδειγμα το **Swiss-PDBviewer** (http://spdbv.vital-it.ch/), το οποίο είναι στενά

συνδεδεμένο με το **SWISS-MODEL** (βλ. παρακάτω) και το **Cn3D** (http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml) το οποίο αποτελεί τμήμα των εφαρμογών του NCBI και είναι στενά συνδεδεμένο με το το Entrez, ενώ παρέχει και δυνατότητες alignment editor. Τέλος, δεν πρέπει να παραλείψουμε να κάνουμε αναφορά στο πιο πετυχημένο ίσως εργαλείο της κατηγορίας αυτής, το **PyMol** (http://www.pymol.org/pymol) το οποίο βασίζεται στη γλώσσα προγραμματισμού Python και κάνει χρήση της τεχνολογίας OpenGL Extension Wrangler Library (GLEW). Το PyMol είναι ίσως η πιο επιτυχημένη εφαρμογή της κατηγορίας, καθώς συνδυάζει άριστη απόδοση γραφικών, πολλές επιλογές για την οπτικοποίηση ακόμα και για τους απαιτητικούς χρήστες και μεγάλη ευκολία στη χρήση ακόμα και για τους αρχάριους. Τέλος, αξίζει μια ειδική αναφορά και στο πακέτο λογισμικού για μοντελοποίηση και επεξεργασία δομών **WHAT IF** (βλ. παρακάτω) που ήταν για πολλά χρόνια το μόνο λογισμικό το οποίο επέτρεπε την 3D αναπαράσταση δομών κάνοντας χρήση των γυαλιών από τα βιντεοπαιχνίδια (σε αντίθεση με τα πολύ πιο ακριβά συστήματα της SGI που ήταν διαθέσιμα για ειδικά συστήματα Unix).

9.3. Στοίχιση και υπέρθεση δομών

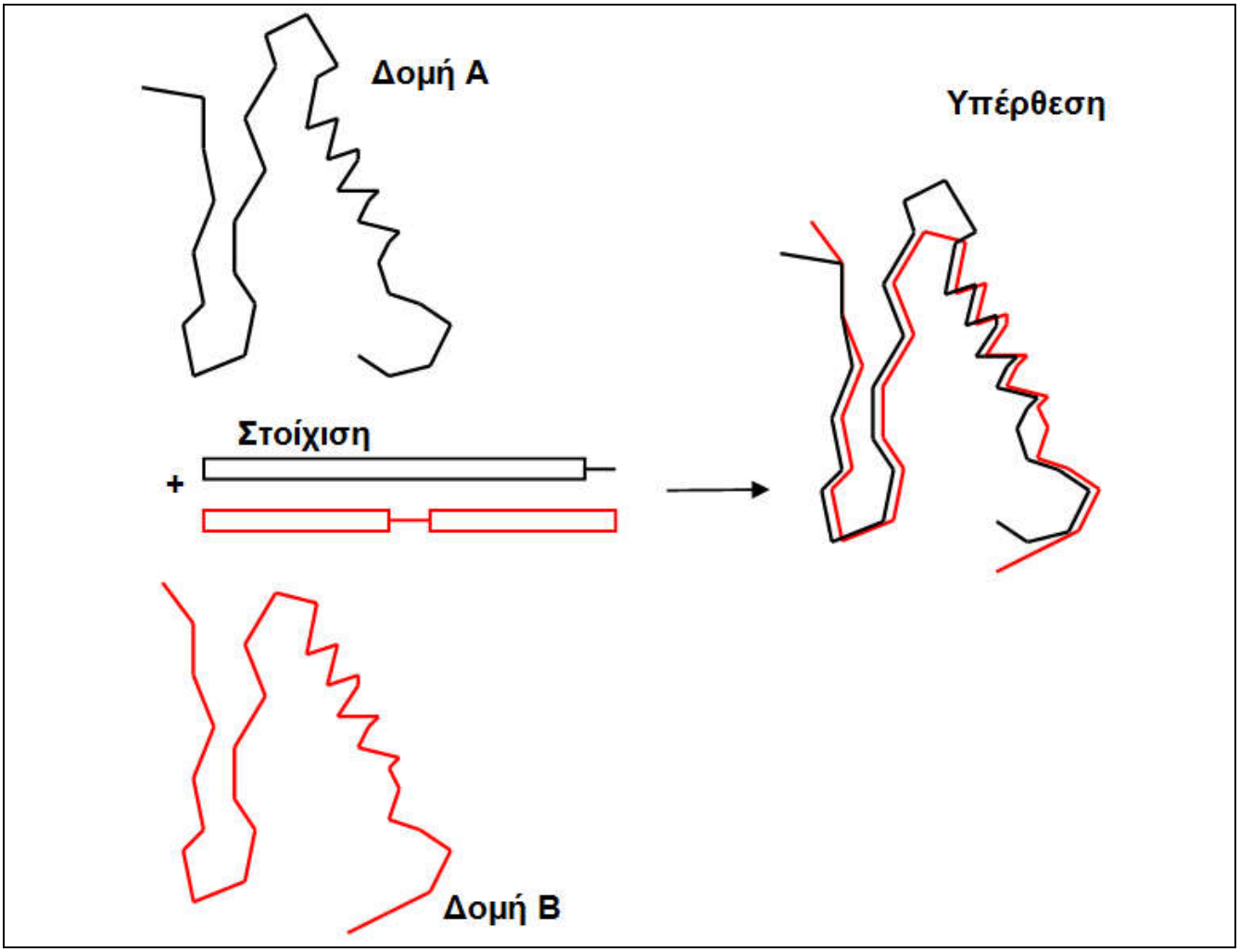
Η δομική στοίχιση (ή στοίχιση δομών) επιχειρεί να εντοπίσει και να τεκμηριώσει την ομολογία δύο πρωτεϊνών μέσω της ομοιότητας των τρισδιάστατων δομών τους (σε αντιδιαστολή με τη στοίχιση αλληλουχιών που επιχειρεί το ίδιο μέσω της ομοιότητας των αλληλουχιών). Γενικά, η διαδικασία και εδώ είναι πιο σύνθετη από τη στοίχιση αλληλουχιών, καθώς θα πρέπει να συγκρίνουμε τρισδιάστατες δομές, δηλαδή συντεταγμένες των ατόμων και όχι απλά δυο μονοδιάστατες αλληλουχίες. Από την άλλη, οι δομικές στοιχίσεις είναι πολύ χρήσιμες, γιατί μπορεί να μας αποκαλύψουν περισσότερα σε σχέση με τις στοιχίσεις αλληλουχιών, καθώς όπως έχουμε αναφέρει η τρισδιάστατη δομή συντηρείται περισσότερο από την αλληλουχία. Κατά συνέπεια, δυο πρωτεΐνες μπορεί να διαφέρουν σε επίπεδο αλληλουχίας περισσότερο από όσο μπορούμε να ανιχνεύσουμε με τις μεθόδους στοίχισης αλληλουχιών, αλλά εντούτοις να εμφανίζουν ξεκάθαρη δομική ομοιότητα. Φυσικά, υπάρχει πάντα ο κίνδυνος να εντοπίσουμε προϊόντα συγκλίνουσας (σε επίπεδο δομής) εξέλιξης και όλα αυτά αποτελούν παράγοντες που πρέπει να λαμβάνονται υπόψη. Οι δομικές στοιχίσεις έχουν και πολλές πρακτικές εφαρμογές, καθώς από αυτές προκύπτουν όπως έχουμε δει στο Κεφάλαιο 4, οι πολλαπλές στοιχίσεις αναφοράς από γνωστές πρωτεϊνικές οικογένειες με τις οποίες αξιολογούμε τις μεθόδους πολλαπλής στοίχισης αλληλουχιών, ενώ με βάση μια δομική στοίχιση μπορούν να γίνουν μια σειρά από δομικές μελέτες για τη σχέση δομής/λειτουργίας μιας δεδομένης πρωτεϊνικής οικογένειας (ή δύο συγκεκριμένων πρωτεϊνών).

Οι περιπτώσεις δομικής στοίχισεις ποικίλουν, ανάλογα με το είδος και τη σχέση των πρωτεϊνών που συγκρίνουμε. Στην πιο απλή περίπτωση, έχουμε την ίδια αλληλουχία με δομή προσδιορισμένη διαφορετικά (με διαφορετική μέθοδο ή σε σύμπλοκο με διαφορετικές ουσίες). Παρόμοια είναι και η περίπτωση δύο πρωτεϊνών με μικρές διαφορές στο επίπεδο της αμινοξικής αλληλουχίας, π.χ. με μία αντικατάσταση σε κάποιο ή κάποια αμινοξέα. Στην περίπτωση αυτή, ξέρουμε εκ των προτέρων ότι τα αμινοξέα της μίας πρωτεΐνης έχουν αντιστοίχιση με τα αμινοξέα της δεύτερης (το $1^{ο}$ με το $1^{ο}$, το $2^{ο}$ με το $2^{ο}$ κ.ο.κ.), και αυτό που έχουμε να κάνουμε είναι μια απλή υπέρθεση δομών (structural superposition). Αν τυχόν οι διαφορές είναι λίγο περισσότερες, αλλά σε κάθε περίπτωση γνωστές εκ των προτέρων, μαζί με τις δύο δομές προμηθεύουμε και μια στοίχιση αλληλουχιών η οποία θα καθοδηγεί το πρόγραμμα όσον αφορά το ποιά ζευγάρια αμινοξέων θα συγκρίνει. Η διαδικασία αυτή ονομάζεται υπέρθεση δομών και είναι η απλούστερη περίπτωση δομικής στοίχισης (έχει όμως αρκετές διαφορές από τη γενικότερη μεθοδολογία που θα δούμε παρακάτω).

Η υπέρθεση των δύο δομών (Εικόνα 9.6), περιλαμβάνει μια διαδικασία σχετικής μετακίνησης της μίας σε σχέση με την άλλη (χωρίς όμως η θέση και η διάταξη των ατόμων της ίδιας πρωτεΐνης να αλλάζει) με σκοπό οι δύο δομές να έρχονται όσο πιο κοντά γίνεται. Η διαδικασία αυτή, αν και φαίνεται απλή σαν ιδέα, έχει αρκετές πρακτικές δυσκολίες. Το μέτρο που αποδίδει αυτή την «ομοιότητα», δηλαδή το πόσο κοντά είναι η μια δομή με την άλλη, είναι το RMSD (Root Mean Square Deviation):

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

όπου N είναι ο αριθμός των ζευγαριών ατόμων που συγκρίνουμε και δ_i η απόσταση στο χώρο του i ζεύγους.

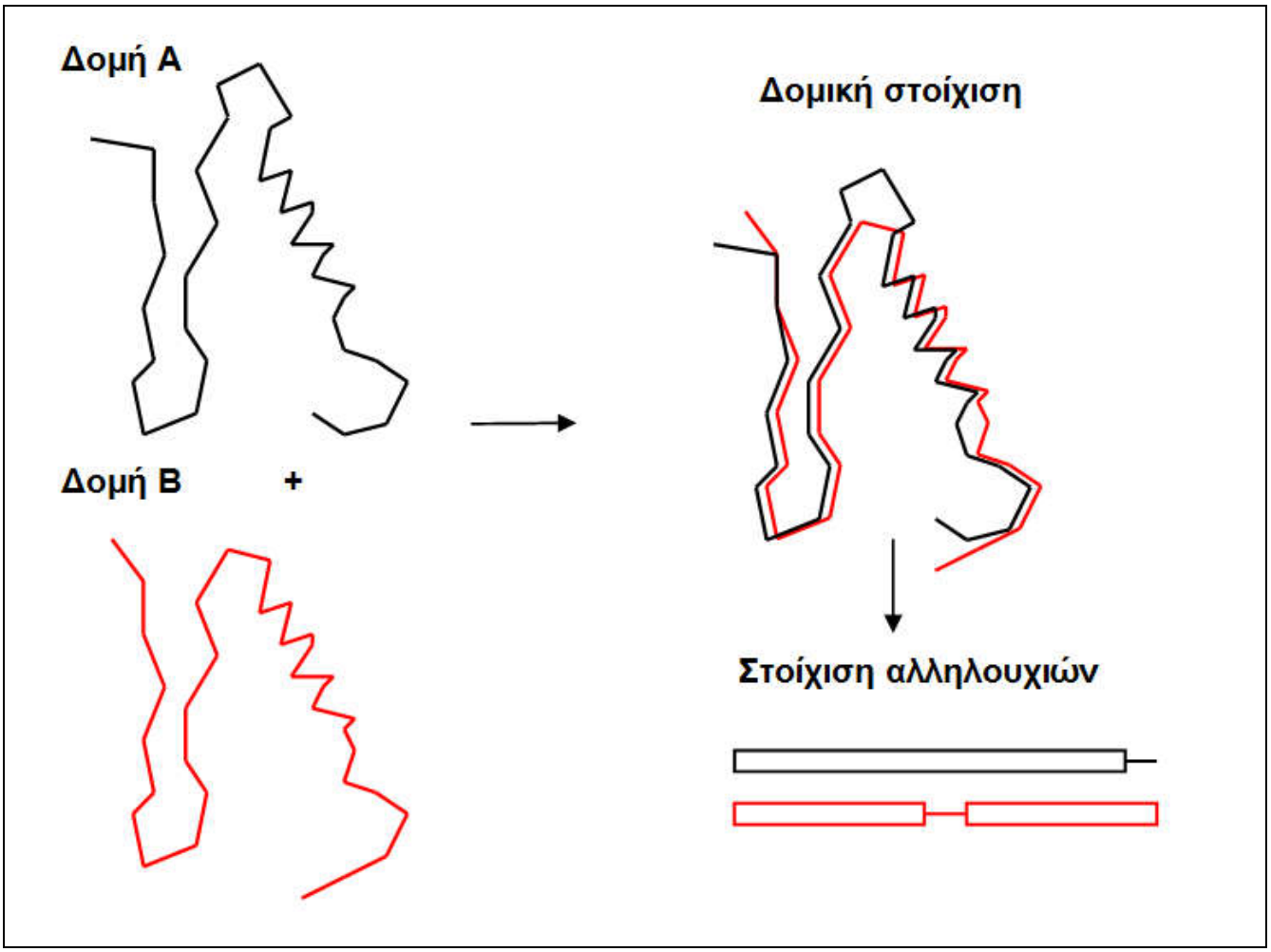


Εικόνα 9.6: *Σχηματική αναπαράσταση της υπέρθεσης δομών.*

Στις περισσότερες των περιπτώσεων, μόνο τα άτομα της κύριας ανθρακικής αλυσίδας (Cα) χρησιμοποιούνται για τις συγκρίσεις αυτές καθώς αυτά είναι που θα καθορίσουν το γενικότερο σχήμα και τη δομή της πρωτεΐνης και οι υπολογισμοί είναι ευκολότεροι. Επιπλέον δε, η σύγκριση των ατόμων των πλευρικών αλυσίδων είναι προβληματική όταν έχουμε να κάνουμε με σύγκριση μη-ταυτόσημων αλληλουχιών. Γενικά, το κριτήριο αυτό χρησιμοποιείται ευρέως, τόσο στην υπέρθεση και τη δομική στοίχιση, αλλά όπως θα δούμε και παρακάτω και σε περιπτώσεις αξιολόγησης θεωρητικών μοντέλων. Η μέθοδος των ελαχίστων τετραγώνων (least squares method) χρησιμοποιείται παραδοσιακά από τους αλγόριθμους υπέρθεσης δομών, αλλά έχουν αναπτυχθεί και μεθοδολογίες που βασίζονται σε αναλύσεις μέγιστης πιθανοφάνειας (maximum likelihood) (Theobald & Wuttke, 2006a, 2006b) αλλά και σταθερών (robust) μεθοδολογιών, όπως η least median squares regression (LMS) (Liu, Fang, & Ramani, 2009). Οι μεθοδολογίες της πρώτης κατηγορίας έχουν υλοποιηθεί στο πρόγραμμα **LSQMAN** (http://xray.bmc.uu.se/usf/lsqman_man.html), της δεύτερης στο **THESEUS** (http://www.theseus3d.org) ενώ της τρίτης στο **LMSfit** (https://engineering.purdue.edu/PRECISE/LMSfit). Το **Profit** (http://www.bioinf.org.uk/software/profit/) είναι μια άλλη γνωστή διαδικτυακή εφαρμογή για υπέρθεση δομών χρησιμοποιώντας τη γρήγορη μέθοδο ελαχίστων τετραγώνων του McLachlan (McLachlan, 1982), ενώ το **3dSS** (http://cluster.physics.iisc.ernet.in/3dss/) είναι μια πιο σύγχρονη εφαρμογή η οποία διασυνδέεται με το RasMol ενώ κάνει και εσωτερικά χρήση του Profit, και επιτρέπει μεταξύ άλλων πολλαπλή υπέρθεση δομών, υπέρθεση υπομονάδων αλλά και άλλες ευκολίες για τον τελικό χρήστη (Sumathi, Ananthalakshmi, Roshan, & Sekar, 2006).

Όπως είπαμε παραπάνω, για την υπέρθεση δομών εκτός από την τετριμένη περίπτωση κατά την οποία έχουμε δύο δομές της ίδιας ακριβώς πρωτεΐνης, είναι απαραίτητη μια στοίχιση των αλληλουχιών. Σε περιπτώσεις πρωτεϊνών με μερικές μόνο μικρές διαφορές στην αμινοξική αλληλουχία, αυτό είναι κάτι εύκολο και κατανοητό. Τί γίνεται όμως όσο οι διαφορές μεταξύ των πρωτεϊνών που επιθυμούμε να συγκρίνουμε

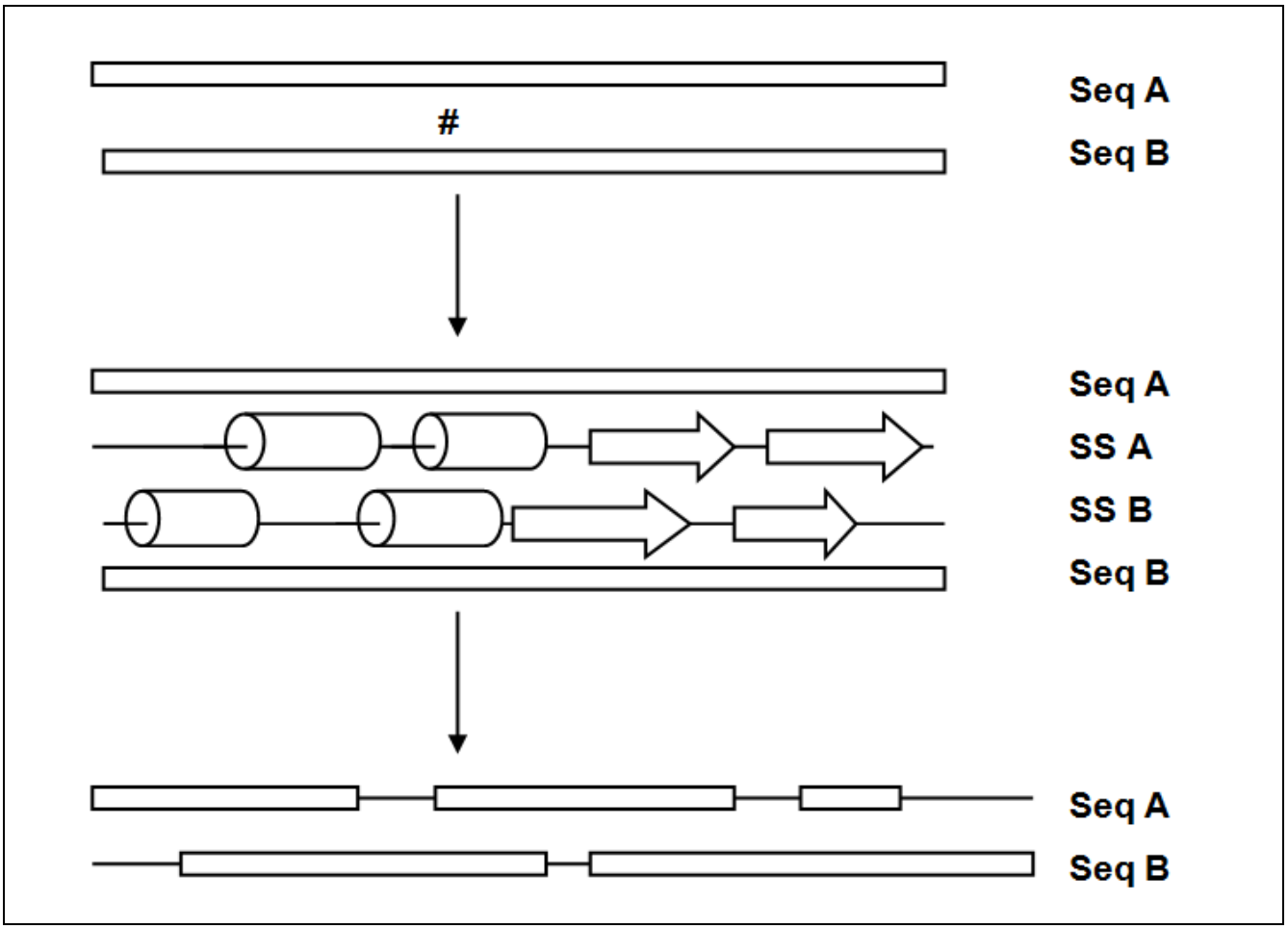
μεγαλώνουν; Καθώς είναι γνωστό ότι η δομή συντηρείται περισσότερο από την αλληλουχία, είναι αναμενόμενο ότι θα υπάρχουν περιπτώσεις πρωτεϊνών με παρόμοια δομή αλλά μικρή μόνο και πιθανώς μη ανιχνεύσιμη ομοιότητα σε επίπεδο αλληλουχίας. Για την ακρίβεια, αυτός ακριβώς είναι και ο λόγος για τον οποίο επιθυμούμε να κάνουμε σύγκριση δομών, για να μπορέσουμε δηλαδή να καταλήξουμε τελικά σε μια στοίχιση αλληλουχιών και να μελετήσουμε την απόκληση δομής/αλληλουχίας (Εικόνα 9.7).



Εικόνα 9.7: *Σχηματική αναπαράσταση της δομικής στοίχισης*.

Ένας απλός τρόπος, για να πραγματοποιήσουμε τη στοίχιση, όταν δεν υπάρχει μεγάλη ομοιότητα, ο οποίος χρησιμοποιείται από διάφορα προγράμματα, είναι να στηριχθούμε στη δευτεροταγή δομή. Η ιδέα είναι απλή και στηρίζεται στο γεγονός ότι η δευτεροταγής δομή μπορεί να κατευθύνει τη στοίχιση. Ένας απλός τρόπος να το επιτύχουμε αυτό, θα ήταν να κάνουμε στοίχιση των ακολουθιών των δευτεροταγών δομών (δηλαδή δυο ακολουθιών που αποτελούνται από τρία σύμβολα: H, E, και C), ενώ ένας λίγο πιο σύνθετος θα ήταν με κάποιον τροποποιημένο αλγόριθμο στοίχισης, στον οποίο η συνεισφορά στο σκορ για δυο αμινοξικά κατάλοιπα θα αυξάνεται αν τα δύο κατάλοιπα έχουν την ίδια δευτεροταγή δομή. Παρόμοιες τεχνικές θα δούμε και στην περίπτωση της ύφανσης (threading) παρακάτω. Μόλις η στοίχιση κατασκευαστεί, τότε είναι εύκολο πλέον να πραγματοποιηθεί η υπέρθεση δομών όπως περιγράφεται παραπάνω χρησιμοποιώντας τη στοίχιση αυτή σαν οδηγό. Το **SuperPose** (http://wishart.biology.ualberta.ca/SuperPose/) είναι μία πολύ εύχρηστη διαδικτυακή εφαρμογή για υπέρθεση δομών η οποία απαιτεί ελάχιστη παρέμβαση από το χρήση (Maiti, Van Domselaar, Zhang, & Wishart, 2004). Το πρόγραμμα λειτουργεί αυτόματα. Έτσι, όταν οι αλληλουχίες διαφέρουν αλλά μπορούν να στοιχηθούν με κάποιον κλασικό αλγόριθμο ομοιότητας, λειτουργεί με τον κλασικό τρόπο που περιγράψαμε παραπάνω. Όταν όμως οι αλληλουχίες των πρωτεϊνών διαφέρουν πέραν των ορίων ανίχνευσης των αλγορίθμων στοίχισης, το SuperPose χρησιμοποιεί την αναφερθείσα τεχνική με τη βοήθεια της δευτεροταγούς δομής για να μπορέσει να κάνει την υπέρθεση των δομών και να δώσει κάτι που μοιάζει με δομική στοίχιση. Πρέπει να τονιστεί βέβαια, ότι παρόλο που αυτό μοιάζει αρκετά με δομική στοίχιση, και σε πολλές περιπτώσεις λειτουργεί και παράγει παρόμοια αποτελέσματα, με βάση τον

ορισμό η μέθοδος αυτή δεν θεωρείται τυπική περίπτωση δομικής στοίχισης, γιατί η στοίχιση δεν πραγματοποιείται με χρήση της δομικής πληροφορίας αλλά με χρήση της αλληλουχίας, ενώ τα όποια κενά εισάγονται μόνο με τη βοήθεια της στοίχισης αλληλουχιών και παραμένουν σταθερά κατά την προσαρμογή των δομών.



Εικόνα 9.8: *Στοίχιση αλληλουχιών με τη βοήθεια της δευτεροταγούς δομής (παρατηρηθείσας ή προβλεφθείσας). Η μέθοδος μπορεί να χρησιμοποιηθεί τόσο στην υπέρθεση δομών όσο και στην ύφανση.*

Προχωρώντας στις πιο κλασικές μεθόδους δομικής στοίχισης, σε αυτές δηλαδή που εφαρμόζονται κάνοντας χρήση της δομής των πρωτεϊνών και είναι ιδανικές για περιπτώσεις στις οποίες δεν υπάρχει ανιχνεύσιμη ομοιότητα των αλληλουχιών, θα πρέπει να κάνουμε κάποιες επισημάνσεις. Καταρχήν, το πρόβλημα της βέλτιστης ύφανσης, δηλαδή της στοίχισης μιας αλληλουχίας με μια δομή, έχει αποδειχτεί NP-complete (Lathrop, 1994), αλλά το πρόβλημα της βέλτιστης στοίχισης δύο δομών, δηλαδή της βέλτιστης προσαρμογής με βελτιστοποίηση κάποιου προκαθορισμένου κριτηρίου, δεν είναι NP (Poleksic, 2009). Έτσι, βέλτιστη λύση μπορεί να βρεθεί (με την προϋπόθεση πάντα ότι μιλάμε για κάποιο δεδομένο κριτήριο ομοιότητας), αλλά η πολυπλοκότητα του προβλήματος είναι μεγάλη και καθιστά την ακριβή λύση απαγορευτική για πρακτικές εφαρμογές. Κατά συνέπεια, οι αλγόριθμοι που χρησιμοποιούνται στην πράξη, βασίζονται σε ευριστικές τεχνικές, μερικές από τις οποίες θα περιγράψουμε παρακάτω.

Ίσως η πιο γνωστή και πετυχημένη σύγχρονη μέθοδος, είναι το **DALI**, (distance alignment matrix method), το οποίο είναι διαθέσιμο στη διεύθυνση http://ekhidna.biocenter.helsinki.fi/dali_server/start σαν υπηρεσία, αλλά και σαν αυτόνομη έκδοση (**DALIlite**). Η μέθοδος είναι από τις πιο παλιές (1993), αλλά έχει εμπλουτιστεί με νέα στοιχεία και πλέον λειτουργεί και με πολλαπλές δομές (πολλαπλή δομική στοίχιση). Η βασική ιδέα της μεθόδου είναι το «σπάσιμο» της δομής σε διαδοχικά εξαπεπτίδια και ο υπολογισμός ενός πίνακα αποστάσεων από τα πρότυπα ενδομοριακών αλληλεπιδράσεων (contacts) που εμφανίζουν τα διαδοχικά εξαπεπτίδια. Τα στοιχεία δευτεροταγούς δομής στα οποία εμπλέκονται συνεχόμενα κατάλοιπα εμφανίζονται στην κύρια διαγώνιο. Όταν στους πίνακες αποστάσεων δύο πρωτεϊνών εμφανίζονται παρόμοια χαρακτηριστικά στην ίδια θέση, οι πρωτεϊνες θα έχουν το ίδιο δίπλωμα. Στο επόμενο βήμα πραγματοποιείται

σύγκριση των επικαλυπτόμενων πινάκων 6x6 και ταύτισή τους με κάποιον αλγόριθμο βελτιστοποίησης (Holm & Rosenström, 2010). Το DALI έχει χρησιμοποιηθεί για την κατασκευή της βάσης FSSP (Families of Structurally Similar Proteins) στην οποία όλες οι γνωστές δομές έχουν στοιχιθεί για να δώσουν μια κατηγοριοποίηση των πρωτεϊνικών διπλωμάτων.

Ένα άλλο ιδιαίτερα πετυχημένο πρόγραμμα είναι το **CE** (Combinatorial extension), το οποίο είναι διαθέσιμο στη διεύθυνση http://source.rcsb.org/jfatcatserver/ceHome.jsp και χρησιμοποιείται από την PDB. Είναι και αυτό μια σχετικά παλιά μέθοδος η οποία εξελίσσεται, ενώ έχει αναπτυχθεί και εφαρμογή για πολλαπλές αλληλουχίες (**CE-MC**). Το CE μοιάζει στο DALI στο γεγονός ότι σπάει τη δομή σε μικρότερα κομμάτια, και μετά επιχειρεί να τα συναρμολογήσει για να κατασκευάσει τη στοίχιση. Συγκρίσεις των θραυσμάτων αυτών (aligned fragment pairs –AFPs) χρησιμοποιούνται για την κατασκευή του πίνακα ομοιότητας στον οποίο γίνεται τελικά η εύρεση του καλύτερου μονοπατιού με δυναμικό προγραμματισμό που καλείται να ενώσει με βέλτιστο τρόπο τα διαδοχικά AFP. Το μέτρο ομοιότητας ήταν αρχικά βασισμένο μόνο στην απόσταση, αλλά στις μετέπειτα εκδόσεις τροποποιήθηκε για να περιλαμβάνει πληροφορία για τη δευτεροταγή δομή, τους δεσμούς υδρογόνου, τις δίεδρες γωνίες κ.ο.κ. (Shindyalov & Bourne, 1998).

Το **SSAP** (Sequential Structure Alignment Program) είναι ίσως η πιο παλιά μέθοδος, η οποία χρησιμοποιεί διπλό δυναμικό προγραμματισμό για να στοιχίσει τις δομές (http://www.biochem.ucl.ac.uk/~orengo/ssap.html). Σε αντίθεση με τις άλλες μεθόδους, χρησιμοποποεί τον Cβ για τους υπολογισμούς, έτσι ώστε να λάβει υπόψη όχι μόνο τη θέση αλλά και τη δευτεροταγή δομή των αμινοξικών καταλοίπων. Στην αρχή η μέθοδος κατασκευάζει μια σειρά διανύσματα αποστάσεων μεταξύ των καταλοίπων και των γειτόνων τους που δεν είναι συνεχόμενα. Έπειτα, κατασκευάζει μια σειρά από πίνακες που περιέχουν τα διανύσματα των διαφορών των απστάσεων μεταξύ γειτόνων. Ο δυναμικός προγραμματισμός στη συνέχεια εφαρμόζεται σε κάθε πίνακα για να δώσει τις τοπικές στοιχίσεις, οι οποίες αθροίζονται ξανά σε ένα συνολικό πίνακα όπου και εφαρμόζεται ξανά δυναμικός προγραμματισμός για να δώσει την τελική στοίχιση (Taylor & Orengo, 1989). Όμοια με τις άλλες μεθόδους, έχει τροποποιηθεί για να δίνει και πολλαπλές στοιχίσεις ενώ χρησιμοποιείται για την ταξινόμηση των πρωτεϊνών στη βάση CATH.

Το **SSM** (http://www.ebi.ac.uk/msd-srv/ssm/), είναι ένας αλγόριθμος που αναπτύχθηκε στο EBI για να καλύψει τις ανάγκες της PDB. Έχει την ιδιαιτερότητα ότι βασίζεται σε μια εντελώς διαφορετική μέθοδο, αυτήν της ταύτισης των στοιχείων δευτεροταγούς δομής και όχι των ατομικών συντεταγμένων (Krissinel & Henrick, 2004). Η μέθοδος αυτή, διαισθητικά θυμίζει αυτό που περιγράψαμε παραπάνω, αλλά το μαθηματικοποιεί περισσότερο και πραγματοποιεί τη μοντελοποίηση σε επίπεδο δομής. Στην αρχή το πρόγραμμα εντοπίζει τα στοιχεία δευτεροταγούς δομής, και δημιουργεί μια γραφοθεωρητική αναπαράσταση της δομής με βάση αυτά. Κατόπιν, κάνει χρήση ενός γρήγορου αλγόριθμου για εύρεση ισομορφισμού γράφων για να συγκρίνει τις δύο αναπαραστάσεις των δομών και επιστρέφει τελικά στις ατομικές συντεταγμένες για να δώσει την τελική στοίχιση.

Το **MASS** (http://bioinfo3d.cs.tau.ac.il/MASS/), είναι επίσης μια μέθοδος πολλαπλής δομικής στοίχισης που βασίζεται στη στοίχιση των στοιχείων δευτεροταγούς δομής (Dror, Benyamini, Nussinov, & Wolfson, 2003). Δυο σημαντικά χαρακτηριστικά του MASS, είναι ότι πεώτον έχει την επιλογή να αγνοεί τη σειρά (είτε των στοιχείων δευτεροταγούς δομής στο πρώτο στάδιο, είτε των καταλοίπων στη συνέχεια), με συνέπεια να μπορεί να ανιχνεύσει κοινά δομικά στοιχεία που έχουν εμφανιστεί σε πρωτεΐνες λόγω συγκλίνουσας εξέλιξης αλλά δεν έχουν ομοιότητα στο δίπλωμα, και δεύτερον, ότι έχει τη δυνατότητα να ανιχνεύσει δομικά μοτίβα που εμφανίζονται μόνο σε ένα υποσύνολο των δομών.

Το **MAMMOTH** είναι μια άλλη πετυχημένη μέθοδος, η οποία έχει επίσης επεκταθεί και για πολλαπλές στοιχίσεις (http://ub.cbm.uam.es/software/online/mamothmult.php). Το MAMMOTH σπάει τη δομή σε επταπεπτίδια, και εφαρμόζει εκεί ένα διαφορετικό μέτρο απόστασης, το unit-vector root mean square (URMS), το οποίο έχει αρκετές επιθυμητές ιδιότητες, το μετατρέπει σε σκορ και μετά χρησιμοποιεί έναν αλγόριθμο δυναμικού προγραμματισμού για να βρει τη βέλτιστη στοίχιση των τμημάτων και μετά βρίσκει τη συνολική δομή που ικανοποιεί κάποιες προϋποθέσεις απόστασης. Μια ιδιαιτερότητα της μεθόδου είναι ότι υπολογίζει και στατιστική σημαντικότητα (Ortiz, Strauss, & Olmea, 2002).

Το **MUSTANG** (multiple structural alignment algorithm) είναι μια εφαρμογή που σχεδιάστηκε εξαρχής για πολλαπλή δομική στοίχιση (Konagurthu, Whisstock, Stuckey, & Lesk, 2006). Βασίζεται σε ιεραρχική πολλαπλή στοίχιση, με πολλαπλά βήματα που επανυπολογίζονται για βελτιστοποίηση. Στην αρχή υπολογίζει τις περιοχές ομοιότητας και τις σκοράρει χρησιμοποιώντας μια «πρόχειρη» στοίχιση αλληλουχιών, την οποία βελτιστοποιεί στη συνέχεια. Κατόπιν, πραγματοποιεί τις κατά ζεύγη δομικές στοιχίσεις με χρήση του RMSD των Cα, και με βάση αυτές διορθώνει τα σκορ από τις συγκρίσεις

αλληλλουχιών και στη συνέχεια προχωρά σε ιεραρχική στοίχιση των δομών (http://www.csse.monash.edu.au/~karun/Site/mustang.html).

Τέλος, θα πρέπει να αναφέρουμε και το **TM-align** (http://zhanglab.ccmb.med.umich.edu/TM-align/) το οποίο είναι μια από τις σχετικά πρόσφατες μεθόδους και έχει κερδίσει μεγάλη δημοφιλία τα τελευταία χρόνια (Zhang & Skolnick, 2005). Στο αρχικό του βήμα χρησιμοποιεί δυναμικό προγραμματισμό για να στοιχίσει τις ακολουθίες της δευτεροταγούς δομής, ενώ κατόπιν στοιχίζει τους άνθρακες της κύριας αλυσίδας (Cα) χρησιμοποιώντας έναν επαναληπτικό ευριστικό αλγόριθμο. Το ιδιαίτερο χαρακτηριστικό του, είναι ότι είναι εξαιρετικά γρήγορο (4 φορές πιο γρήγορο από το CE και 20 φορές πιο γρήγορο από το DALI), χωρίς όμως να υστερεί σε ποιότητα και αξιοπιστία.

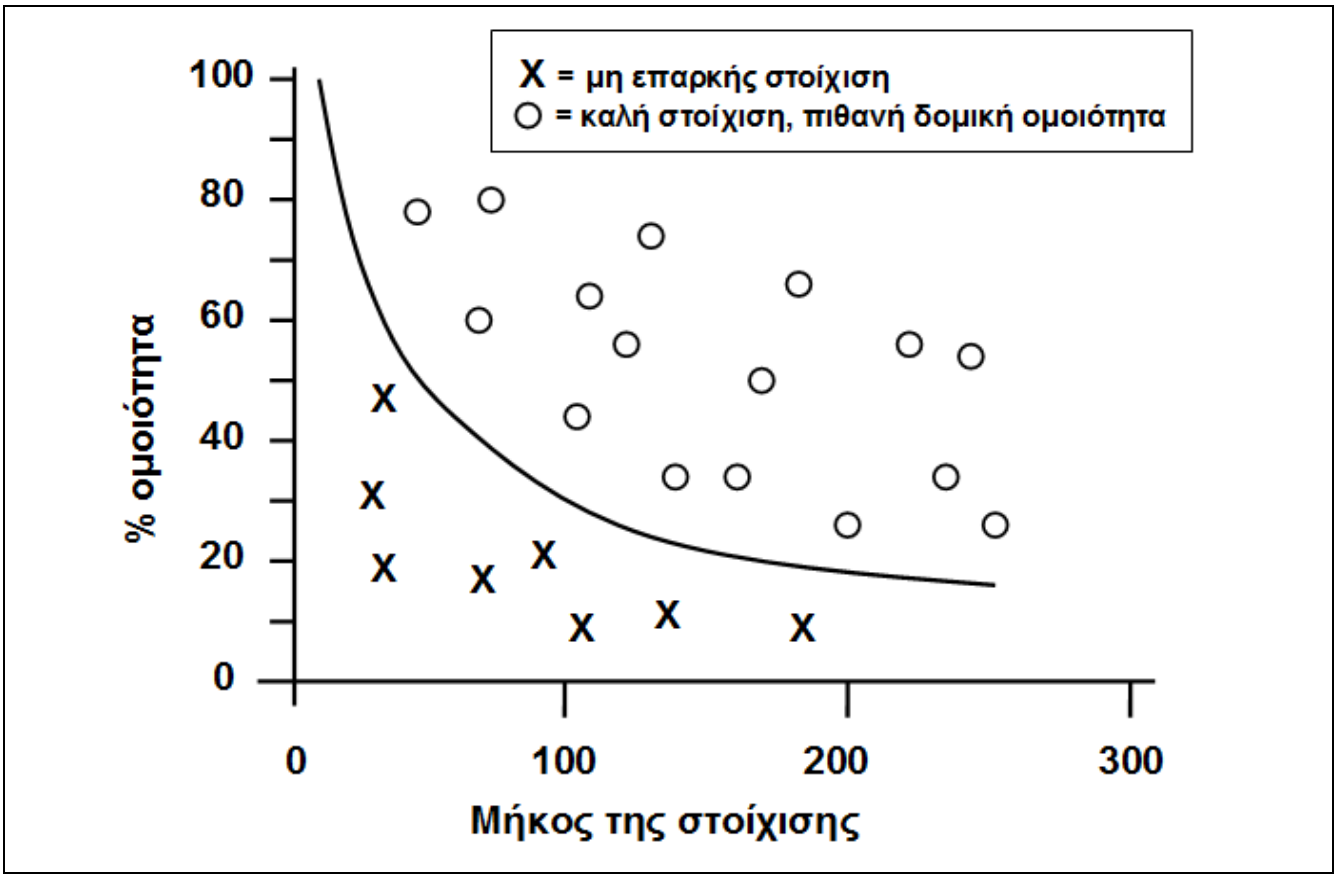
Όπως αναφέραμε παραπάνω, τα προγράμματα αυτά είναι μόνο ένα μικρό μέρος των προγραμμάτων που είναι διαθέσιμα στην επιστημονική κοινότητα. Στην αντίστοιχη σελίδα της Wikipedia (https://en.wikipedia.org/wiki/Structural_alignment_software), αναφέρονται δεκάδες αντίστοιχα εργαλεία, παρ' όλα αυτά, εδώ έγινε αναφορά σε αυτά που θεωρούνται πιο αξιόπιστα και χρησιμοποιούνται από τους περισσότερους ερευνητές. Στη βιβλιογραφία έχουν αναφερθεί μερικές μόνο συγκριτικές μελέτες, οι οποίες όμως έχουν το μειονέκτημα ότι κάθε φορά συγκρίνουν λίγα μόνο από τα διαθέσιμα εργαλεία ενώ χρησιμοποιούν και διαφορετικά σύνολα πρωτεϊνών και διαφορετικά κριτήρια αξιολόγησης (Kolodny, Koehl, & Levitt, 2005; Mayr, Domingues, & Lackner, 2007; Singh & Brutlag, 2000). Σε αυτό που συμφωνούν όλοι, είναι ότι τα περισσότερα από τα εργαλεία που αναφέραμε παραπάνω, αποδίδουν αρκετά καλά στις περισσότερες συνθήκες, ενώ όταν οι πρωτεΐνες έχουν μια στοιχειώδη ομοιότητα, οι στοιχίσεις τους είναι παρόμοιες. Γενικά, το DALI, το CE και το TM-align φαίνεται να είναι τα καλύτερα και τα πιο εύχρηστα, ενώ το τελευταίο είναι και ιδιαίτερα γρήγορο. Παρ' όλα αυτά, υπάρχουν ειδικές περιπτώσεις στις οποίες κάποιο άλλο εργαλείο μπορεί να ενδείκνυται καλύτερα, γι' αυτό και ο χρήστης θα πρέπει να έχει καλή γνώση του βιολογικού προβλήματος και να είναι ενήμερος έτσι ώστε να μπορεί να χρησιμοποιήσει και εναλλακτικές μεθόδους.

9.4. Πρόγνωση τρισδιάστατης δομής πρωτεϊνών

Ο τελικός σκοπός της υπολογιστικής μελέτης και της μοντελοποίησης των πρωτεϊνών, είναι η πρόγνωση της τρισδιάστατης δομής μιας πρωτεΐνης από την αλληλουχία της. Σε προηγούμενα κεφάλαια, είδαμε την πρόγνωση της δευτεροταγούς δομής, η οποία είναι ένα σχετικά εύκολο υποκατάστατο για την τελική πρόγνωση της τρισδιάστατης δομής. Μια τέτοια πρόβλεψη θα επιτρέψει τη διενέργεια πολλών πειραμάτων in silico (σχεδιασμός φαρμάκων, μελέτη της λειτουργίας της πρωτεΐνης, μελέτη αλληλεπιδράσεων κ.ο.κ.), για τα οποία σήμερα είναι απαραίτητη η διεξαγωγή των επίπονων και κοστοβόρων πειραμάτων προσδιορισμού της δομής. Επιπλέον δε, υπάρχουν και περιπτώσεις πρωτεϊνών που αποδεικνύονται δύσκολες στις μελέτες αυτές (δυσκολίες στην κρυστάλλωση κ.ο.κ.) και για τέτοιες περιπτώσεις, οι υπολογιστικές μελέτες είναι η μόνη διαθέσιμη εναλλακτική. Οι βασικές αρχές πίσω από τις μελέτες μοντελοποίησης είναι δύο, και είναι γνωστές από χρόνια: α) η αλληλουχία μιας πρωτεΐνης καθορίζει μονοσήμαντα τη δομή μιας πρωτεΐνης, και β) οι πρωτεϊνικές δομές συντηρούνται περισσότερο από τις αλληλουχίες. Μια άμεση συνέπεια των παραπάνω, είναι ότι δυο πρωτεΐνες με μεγάλη ομοιότητα σε επίπεδο αλληλουχίας έχουν κατά βάση παρόμοια δομή, αλλά είναι δυνατό, παρόμοια δομή να έχουν και πρωτεΐνες με μη ανιχνεύσιμη ομοιότητα (στο τελευταίο, σημαντικό ρόλο παίζει και η ύπαρξη περιορισμένου αριθμού πρωτεϊνικών διπλωμάτων).

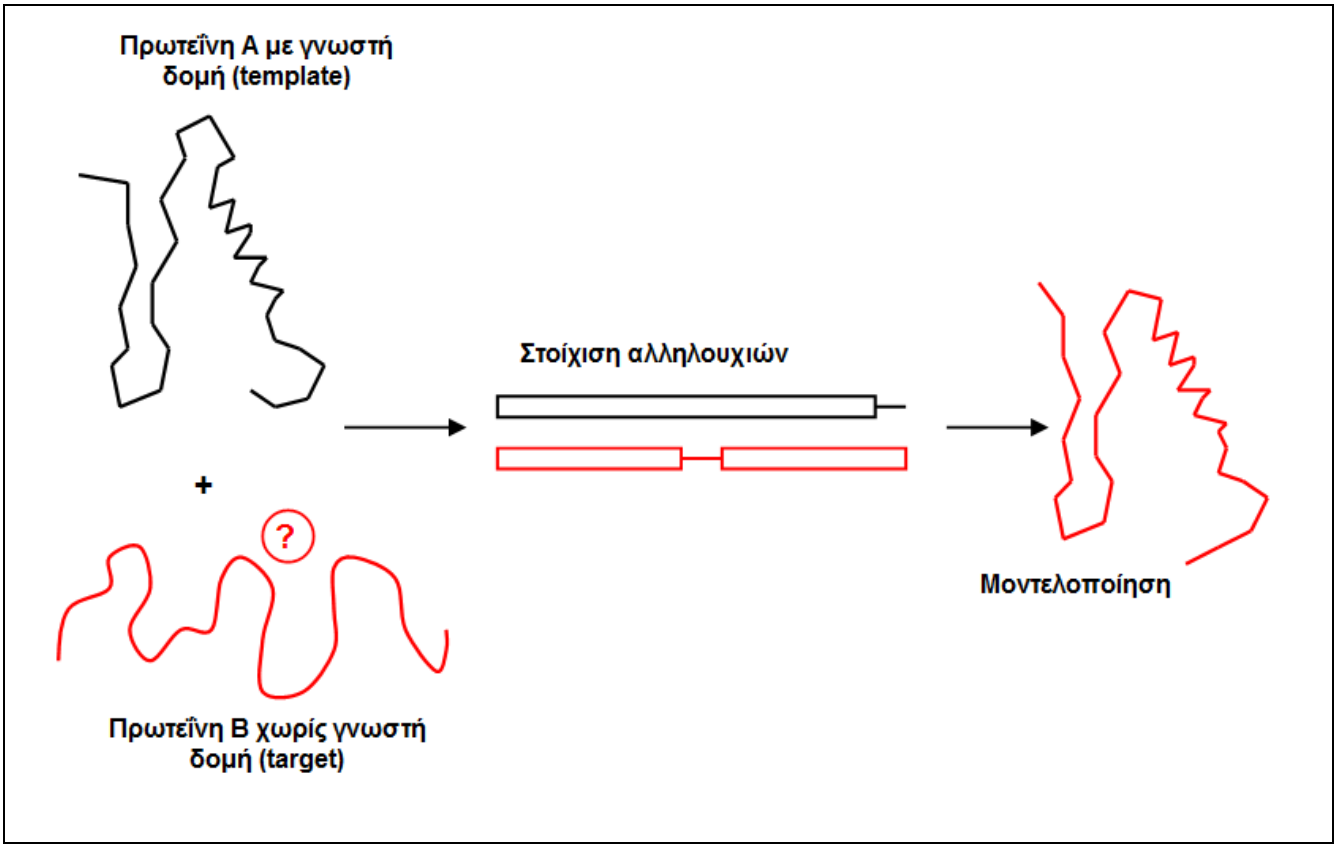
Με όλα τα παραπάνω, μπορούμε να φανταστούμε ένα σενάριο στο οποίο έχουμε μια πρωτεΐνη με άγνωστη δομή, την οποία θέλουμε να προβλέψουμε, και έχουμε εντοπίσει μια ομόλογή της (μια πρωτεΐνη δηλαδή με μεγάλη ομοιότητα σε επίπεδο αλληλουχίας) η οποία διαθέτει γνωστή τρισδιάστατη δομή. Σε μια αντίστροφη πορεία από αυτήν που είχαμε ακολουθήσει στη δομική στοίχιση και την υπέρθεση, μπορούμε να φανταστούμε μια σειρά περιπτώσεων, στις οποίες η πρωτεΐνη με την άγνωστη δομή (target) μοντελοποιείται κάνοντας χρήση της δομής της ομόλογής της σαν καλούπι ή πρότυπο (template). Στην πιο απλή περίπτωση, αν λ.χ. διαφοροποιείται ένα αμινοξικό κατάλοιπο, είναι λογικό να υποθέσουμε ότι η υπόλοιπη δομή θα είναι ίδια και η μόνη διαφορά θα υφίσταται στο συγκεκριμένο κατάλοιπο και σε όσα βρίσκονται σε άμεση επαφή με αυτό (τέτοιες περιπτώσεις αν και τετριμμένες, μπορούν να έχουν σημασία όταν για παράδειγμα μελετάμε μια συγκεκριμένη αλλαγή στο ενεργό κέντρο ενός ενζύμου). Όσο πέφτει το επίπεδο ομοιότητας, αναμένουμε να υπάρχουν περισσότερες αλλαγές στο πρωτεϊνικό μόριο (διαφορετικές πλευρικές ομάδες, διαφορετικές αλληλεπιδράσεις κ.ο.κ.), αλλά η γενικότερη δομή θα είναι περίπου ίδια (Εικόνα 9.9). Το ερώτημα είναι όμως, πού ακριβώς τίθεται το όριο κάτω από το οποίο δεν μπορούμε πλέον να υποθέσουμε ότι οι δυο πρωτεΐνες έχουν την ίδια δομή με σιγουριά; Όπως έχουμε δει στο κεφάλαιο της στοίχισης αλληλουχιών, η απάντηση στο

πρόβλημα αυτό (η οποία στην ουσία απαντά στο πρόβλημα της στατιστικής σημαντικότητας μιας στοίχισης), εξαρτάται από δύο παράγοντες: από την ομοιότητα σε αμινοξέα της στοίχισης, και από το μήκος της στοίχισης, τα οποία μπορούν να παρασταθούν γραφικά και να σχηματιστεί εκεί μια γραμμή η οποία θα χωρίσει το επίπεδο των πιθανών στοιχίσεων σε αποδεκτές και μη αποδεκτές. Για μεγάλα ποσοστά ομοιότητας, το απαραίτητο μήκος της στοίχισης είναι μικρό. Για μικρότερα ποσοστά ομοιότητας όμως απαιτείται μεγαλύτερη στοίχιση. Γενικά, για ομοιότητες κάτω από το 30% απαιτούνται μεγάλες στοιχίσεις, ενώ για ομοιότητα κάτω από 20% δεν υπάρχει απλή μέθοδος στοίχισης για να δείξει καν αυτή την ομοιότητα (η περιοχή ονομάζεται και twilight zone).



Εικόνα 9.9: *Η σχέση της % ομοιότητας σε μια στοίχιση με το μήκος της στοίχισης καθορίζει την ποιότητα της στοίχισης.*

Μπορούμε, όπως είπαμε παραπάνω, να φανταστούμε ένα ολόκληρο φάσμα περιπτώσεων πρωτεϊνών που πιθανώς να συναντήσουμε σε μια προσπάθεια μοντελοποίησης της δομής. Κάποιες βρίσκονται στην «καλή» περιοχή, δηλαδή έχουν μια ξεκάθαρη ομοιότητα για μεγάλο μήκος της αλληλουχίας τους με πρωτεΐνες γνωστής δομής (σε διάφορα επίπεδα, 80%, 50%, 40% κ.ο.κ.), ενώ κάποιες εμφανίζουν πολύ μικρές ομοιότητες (<30%) για μικρά τμήματα τους ή δεν θα εμφανίζουν καμία ομοιότητα. Αυτές τις περιπτώσεις έρχονται να αντιμετωπίσουν οι διαφορετικές τεχνικές μοντελοποίησης της δομής, τις οποίες και αυτές πρέπει να τις αντιμετωπίσουμε σε ένα «συνεχές» φάσμα. Έτσι, για τις πρωτεΐνες της πρώτης κατηγορίας, υπάρχει η τεχνική που με απλά λόγια περιγράψαμε παραπάνω, η λεγόμενη *προτυποποίηση με βάση την ομολογία* (homology modelling). Για τις περιπτώσεις της δεύτερης κατηγορίας, υπάρχει η τεχνική της **ύφανσης** (threading), αλλά και η τεχνική της *προτυποποίηση εκ του μηδενός* (ab initio modelling), οι οποίες είναι τελείως διαφορετικές μεταξύ τους και θα παρουσιαστούν ξεχωριστά. Γενικά η ύφανση εφαρμόζεται σε πρωτεΐνες στόχους, που αφενός μεν δεν διαθέτουν ομόλογη πρωτεΐνη με γνωστή δομή, αφετέρου δε είναι δυνατόν να εντοπιστεί, με κάποια μέθοδο *αναγνώρισης διπλώματος*, το πρωτεϊνικό δίπλωμα στο οποίο ταιριάζουν (γι' αυτό και πολλές φορές οι όροι «αναγνώριση διπλώματος» και «ύφανση», χρησιμοποιούνται χωρίς διάκριση μεταξύ τους). Οι μέθοδοι ab initio πρόγνωσης, μπορούν φυσικά να εφαρμοστούν σε όλες τις περιπτώσεις, αλλά επειδή είναι και οι πιο υπολογιστικά απαιτητικές, αλλά και αυτές με τη μεγαλύτερη επισφάλεια ως προς το αποτέλεσμα, χρησιμοποιούνται περισσότερο για τις πρωτεΐνες για τις οποίες ούτε καν κάποιο πιθανό δίπλωμα δεν μπορεί να αναγνωριστεί.



Εικόνα 9.9: *Σχηματική αναπαράσταση της προτυποποίησης με βάση την ομολογία.*

Προφανώς, οι 3 μεθοδολογίες παράγουν και τρισδιάστατα μοντέλα με διαφορετική αξιοπιστία και κατά συνέπεια κατάλληλα για διαφορετικές χρήσεις. Για παράδειγμα, η μοντελοποίηση με βάση την ομολογία, ειδικά όταν το πρότυπο έχει μεγάλη ομοιότητα με το στόχο (>70%) μπορεί να δώσει μοντέλα πολύ κοντά στην πραγματική δομή (RMSD<1 Å), μοντέλα δηλαδή που μπορούν να χρησιμοποιηθούν για λεπτομερείς δομικές μελέτες (για μελέτη ενζυμικών μηχανισμών κλπ). Όταν η ομολογία είναι μικρότερη, τα μοντέλα έχουν μεγαλύτερη απόκλιση από την πραγματική δομή και στην περίπτωση της ύφανσης μιλάμε πλέον για απόκλιση της τάξης των 2-4 Å. Τέλος, στην περίπτωση της ab initio πρόγνωσης, στην καλύτερη των περιπτώσεων δίνουν μοντέλα με RMSD της τάξης των 4-10 Å, τιμές που αρκούν για να δώσουν πληροφορίες μόνο για το γενικότερο σχήμα της πρωτεΐνης και όχι για λεπτομερείς αλληλεπιδράσεις της.

9.4.1. Μοντελοποίηση με βάση την ομολογία

Όπως είπαμε, η μοντελοποίηση (ή προτυποποίηση) με βάση την ομολογία, είναι η ενδεικνυόμενη μέθοδος για τις περιπτώσεις στις οποίες μια ομόλογη πρωτεΐνη με γνωστή δομή μπορεί να αναγνωριστεί εύκολα με μεθόδους στοίχισης αλληλουχιών (Εικόνα 9.10). Στη μέθοδο αυτή, η οποία μπορεί και να χαρακτηριστεί διαισθητικά και ως το αντίστροφο της υπέρθεσης δομών, η στοίχιση αλληλουχιών και μόνο αυτή είναι που κατευθύνει τη δημιουργία του μοντέλου. Τα βασικά βήματα της μοντελοποίησης με βάση την ομολογία είναι τα εξής:

- *Εύρεση του πρότυπου και πραγματοποίηση της στοίχισης*. Συνήθως χρησιμοποιούνται μέθοδοι όπως το BLAST και το FASTA, αν και κάποιες φορές η ολική στοίχιση είναι προτιμότερη, ειδικά αν υπάρχει ξεκάθαρη ομοιότητα. Επίσης, είναι πιθανό να αναγνωριστούν πολλά πρότυπα οπότε μπορούν να κατασκευαστούν πολλά εναλλακτικά μοντέλα. Πολλές φορές μια διόρθωση είναι απαραίτητη, ειδικά σε περιοχές με μικρή ομοιότητα. Η διόρθωση μπορεί να γίνει είτε με χρήση πρότερης γνώσης είτε με τη χρήση αλγορίθμων πολλαπλής στοίχισης (χρησιμοποιώντας δηλαδή την πληροφορία και από άλλες ομόλογες).
- *Κατασκευή του σκελετού της κύριας ανθρακικής αλυσίδας*. Στη φάση αυτή «χτίζεται» η νέα δομή ακολουθώντας το πρότυπο και τη στοίχιση. Σε περιοχές που τα κατάλοιπα είναι ίδια, η

κατάσταση είναι απλή. Εκεί που υπάρχουν διαφορετικά κατάλοιπα τοποθετούνται μόνο τα άτομα του σκελετού (C, Cα, N, και O). Ένα πρόβλημα μπορεί να υπάρξει σε περιοχές της δομής του προτύπου που δεν έχουν προσδιοριστεί καλά, και πολλά προγράμματα το διορθώνουν χρησιμοποιώντας πολλαπλά πρότυπα.

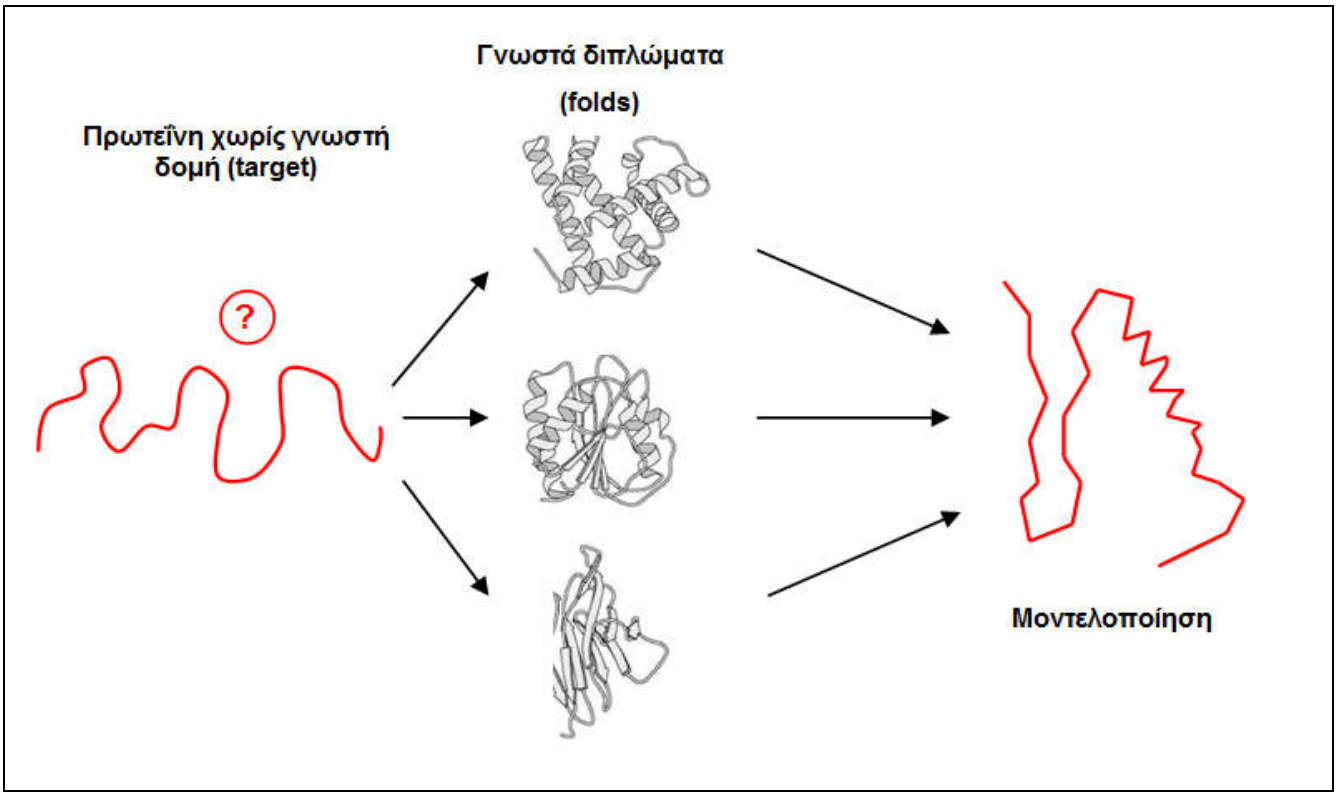
- *Μοντελοποίηση των βρόχων και των πλευρικών αλυσίδων*. Στις περισσότερες περιπτώσεις στις στοιχίσεις θα υπάρχουν κενά. Όταν τα κενά βρίσκονται στην αλληλουχία του στόχου, θα πρέπει τα κατάλοιπα πριν και μετά το κενό να μετακινηθούν στην τελική δομή. Όταν όμως το κενό βρίσκεται στην αλληλουχία του προτύπου, δηλαδή έχει γίνει εισαγωγή στην αλληλουχία στόχο, τότε τα επιπλέον κατάλοιπα θα πρέπει να σχηματίσουν ένα βρόχο (loop), τη δομή του οποίου θα πρέπει να υπολογίσουμε. Επιπλέον δε, οι βρόχοι ούτως ή άλλως είναι ευκίνητες περιοχές οι οποίες είναι πολύ πιθανό να διαφέρουν αρκετά, ακόμα και σε πολύ όμοιες αλληλουχίες. Για να μοντελοποιηθεί σωστά ένας βρόχος, υπάρχουν δύο βασικές στρατηγικές, η πρώτη που μοιάζει περισσότερο με ύφανση και τη χρησιμοποιούν τα περισσότερα προγράμματα, στην οποία το πρόγραμμα ψάχνει στην PDB για περιοχές με παρόμοια κατάλοιπα, ενώ στη δεύτερη που είναι στην ουσία ab initio μέθοδος, γίνεται ελαχιστοποίηση ενέργειας για τον υπολογισμό της βέλτιστης δομής. Στην περίπτωση των πλευρικών ομάδων, το ζήτημα αφορά την ελεύθερη περιστροφή γύρω από το δεσμό Cα-Cβ. Κάποιες προσεγγίσεις στηρίζονται στην απλή μεταφορά αυτής της δομικής πληροφορίας από το πρότυπο, αλλά αυτό είναι επιτυχημένο μόνο για μεγάλη ομοιότητα (>35%) σε επίπεδο αλληλουχίας. Παράλληλα υπάρχουν και άλλες προσεγγίσεις που βασίζονται σε ανίχνευση όμοιων περιοχών στην PDB αλλά και σε ενεργειακούς υπολογισμούς.
- *Βελτιστοποίηση του μοντέλου*. Στο βήμα αυτό γίνεται βελτιστοποίηση όλης της δομής ταυτόχρονα, έτσι ώστε να ληφθούν υπόψη παράλληλα και ο προσανατολισμός των Cα αλλά και των βρόχων και των πλευρικών αλυσίδων (γιατί το ένα μπορεί να επηρεάζει το άλλο). Συνήθως το βήμα αυτό γίνεται επαναληπτικά και απαιτεί πιο προσεκτικά σχεδιασμένη συνάρτηση ενέργειας (σε σχέση με το προηγούμενο βήμα), ενώ η πιο απλή περίπτωση είναι να χρησιμοποιηθεί προσομοίωση μοριακής δυναμικής (molecular dynamics). Ανάλογα με τη συνάρτηση ενέργειας που μπορεί να χρησιμοποιηθεί, το βήμα αυτό μπορεί να είναι υπολογιστικά απαιτητικό.
- *Έλεγχος ποιότητας του μοντέλου*. Αφού το μοντέλο έχει κατασκευαστεί, είναι απαραίτητος ο έλεγχος για την επιβεβαίωσή του. Αυτός μπορεί να γίνει βασικά με δυο τρόπους, είτε με χρήση μοριακής δυναμικής με υπολογισμό της συνολικής ενέργειας το μορίου είτε με εμπειρικές μεθόδους που μετράνε την κανονικότητα διάφορων χαρακτηριστικών (μήκη δεσμών, αποστάσεις, γωνίες κ.ο.κ.). Η δεύτερη μέθοδος είναι πιο εύχρηστη καθώς επιτρέπει τον εντοπισμό των λαθών σε συγκεκριμένα σημεία κατά μήκος της αλληλουχίας.

Το πιο γνωστό αλλά και το πιο παλιό λογισμικό για μοντελοποίηση με βάση την ομολογία, είναι το **WHAT IF** (http://swift.cmbi.ru.nl/whatif/), το οποίο παρουσιάστηκε πρώτη φορά το 1987 από τον Gert Vriend (Vriend, 1990). Από τότε, συνεχίζει να εξελίσσεται και αποτελεί πλέον ένα ολοκληρωμένο περιβάλλον για τη μελέτη των πρωτεϊνικών δομών ενσωματώνοντας συνεχώς νέες λειτουργίες (οπτικοποίηση, υπέρθεση, 3D γραφικά, έλεγχο εγκυρότητας δομών, μοριακή δυναμική, υπολογισμούς φορτίων κ.ο.κ.), ενώ είναι διαθέσιμο ελεύθερα στην επιστημονική κοινότητα για διάφορες πλατφόρμες, αλλά και ως διαδικτυακή εφαρμογή. Το **MODELLER** (https://salilab.org/modeller/) είναι επίσης ένα κλασικό πακέτο λογισμικού για μοντελοποίηση με βάση την ομολογία (Eswar et al., 2006). Το MODELLER είναι ιδιαίτερα εύχρηστο καθώς στην πιο απλή εκδοχή ο χρήστης προμηθεύει ο ίδιος μια στοίχιση του στόχου με το πρότυπο (αυτό είναι ιδιαίτερα σημαντικό, όπως θα δούμε παρακάτω, καθώς μπορεί να κάνει χρήση και τεχνικών ύφανσης) και το λογισμικό υπολογίζει αυτόματα την τρισδιάστατη δομή. Το MODELLER χρησιμοποιεί την τεχνική των Sali και Blundell (Šali & Blundell, 1993), αλλά ενσωματώνει πολλές άλλες λειτουργίες όπως de novo μοντελοποίηση των βρόχων, βελτιστοποίηση του μοντέλου, πολλαπλή στοίχιση αλληλουχιών και δομών, ομαδοποίηση, αναζήτηση σε βάσεις δεδομένων, σύγκριση δομών κ.ο.κ. Παράλληλα, είναι και ελεύθερα διαθέσιμο για τις περισσότερες πλατφόρμες H/Y (Unix/Linux, Windows, και Mac) ενώ έχει αναπτυχθεί και ένα παραθυρικό περιβάλλον για τη λειτουργία του, το **EasyModeller** (http://modellergui.blogspot.gr/). Τέλος, το **SWISS-MODEL** (http://swissmodel.expasy.org/) αποτελεί ίσως την πιο εύχρηστη εναλλακτική για μοντελοποίηση με βάση την ομολογία. Το εργαλείο λειτουργεί σαν μια αυτοματοποιημένη διαδικτυακή εφαρμογή, παρέχοντας πλήθος λειτουργιών όπως αυτόματη αναζήτηση στις

βάσεις δεδομένων, έλεγχο ποιότητας για την επιλογή του καλύτερου πρότυπου, μοντελοποίηση με πολλαπλά πρότυπα και έλεγχο ποιότητας του προκύπτοντος μοντέλου (Biasini et al., 2014).

9.4.2. Αναγνώριση διπλώματος και ύφανση

Όπως είπαμε ήδη, η ύφανση ή αλλιώς αναγνώριση διπλώματος είναι μια τεχνική που χρησιμοποιείται σε περιπτώσεις κατά τις οποίες η πρωτεΐνη στόχος δεν έχει ανιχνεύσιμη ομοιότητα σε επίπεδο αλληλουχίας με κάποια πρωτεΐνη γνωστής δομής, αλλά μοιράζεται το ίδιο δίπλωμα με αυτές. Με τη διαδικασία αυτή, γίνεται έλεγχος αν η αλληλουχία μπορεί να ταιριάξει με κάποιο από τα γνωστά διπλώματα και μετά κατασκευάζεται η στοίχιση με το δίπλωμα αυτό (με τη δομή δηλαδή). Η βασική διαφορά από την μοντελοποίηση με βάση την ομολογία, στην οποία το πρότυπο το χειριζόμαστε ως αλληλουχία, είναι ότι στην ύφανση το πρότυπο χρησιμοποιείται σαν δομή. Οι μέθοδοι αναγνώρισης διπλώματος έχουν αποκτήσει μεγάλη δημοφιλία, λόγω της γνωστής αρχής ότι η δομή συντηρείται περισσότερο από την αλληλουχία και, κατά συνέπεια, από την παρατήρηση ότι ακόμα και διαφορετικές πρωτεΐνες μπορεί να έχουν παρόμοια δομή (ίδιο δίπλωμα). Επιπλέον δε, πιστεύεται γενικά ότι ο αριθμός των διπλωμάτων είναι πεπερασμένος και βρίσκεται κάπου ανάμεσα στο 1000-2000 (σήμερα πιστεύεται ότι έχουν εντοπιστεί 1300 διαφορετικά διπλώματα). Συνεπώς, μια πρωτεΐνη με μη ανιχνεύσιμη ομοιότητα σε επίπεδο αλληλουχίας, είναι παρ' όλα αυτά πολύ πιθανό να μπορεί να ταυτιστεί με κάποιο από τα ήδη γνωστά διπλώματα. Μια άλλη ενδιαφέρουσα παρατήρηση που πρέπει να γίνει, είναι ότι η αναγνώριση διπλώματος μοιάζει σε κάποιο βαθμό με τη δομική στοίχιση, και πράγματι κάποιες αλγοριθμικές τεχνικές έχουν χρησιμοποιηθεί και στις δύο μεθοδολογίες (π.χ. είδαμε παραπάνω τη στοίχιση με τη βοήθεια της δευτεροταγούς δομής).



Εικόνα 9.10: *Σχηματική αναπαράσταση της ύφανσης.*

Οι μεθοδολογίες που χρησιμοποιούνται στην αναγνώριση διπλώματος, εμφανίζουν τεράστια ετερογένεια αλλά χωρίζονται γενικά σε δύο μεγάλες κατηγορίες. Στην πρώτη κατηγορία ανήκουν οι μέθοδοι που μετατρέπουν τις τρισδιάστατες δομές σε μια μονοδιάστατη αλληλουχία (1D), σε ένα είδος προφίλ, και μετά στοιχίζουν την πρωτεΐνη στόχο με αυτό το προφίλ συνήθως με χρήση κλασικού δυναμικού προγραμματισμού. Σαν προφίλ μπορεί να χρησιμοποιηθεί πληροφορία από τη δευτεροταγή δομή, την προσβασιμότητα στο διαλύτη κ.ο.κ., ενώ για να εφαρμοστεί η στοίχιση απαιτείται και η κατασκευή κάποιου είδους πίνακα για το σκορ, που να συνδέει τα γράμματα του νέου «αλφαβήτου» στο οποίο έχει μεταφραστεί η

δομή, με τις αλληλουχίες αμινοξέων οι οποίες θα χρησιμοποιηθούν σαν στόχοι. Στη δεύτερη κατηγορία, χρησιμοποιείται κατευθείαν η τρισδιάστατη δομή (3D) και η ομοιότητα αξιολογείται με σύγκριση των ατομικών αποστάσεων. Συνήθως σε αυτή την περίπτωση, κατασκευάζεται ένα είδος σκορ που να μετράει τις πιθανές αλληλεπιδράσεις των ατόμων της πρωτεΐνης στην πιθανή δομή (δίπλωμα), ενώ ο δυναμικός προγραμματισμός έχει μεγαλύτερη πολυπλοκότητα. Όπως είναι φανερό, οι μέθοδοι της δεύτερης κατηγορίας χρησιμοποιούν περισσότερη πληροφορία, αλλά είναι οι πιο πολύπλοκες και κοστοβόρες από άποψη χρόνου. Η μέθοδος με τα προφίλ προτάθηκε πρώτη φορά από τους Bowie, Lüthy και Eisenberg το 1991 (Bowie, Luthy, & Eisenberg, 1991) ενώ ο ίδιος ο όρος ύφανση (threading) χρησιμοποιήθηκε για πρώτη φορά από τους Jones, Taylor και Thornton το 1992 (Jones, Taylort, & Thornton, 1992) και αρχικά αναφερόταν αποκλειστικά στη χρήση της τρισδιάστατης δομής. Σήμερα παρ' όλα αυτά, οι όροι ύφανση και αναγνώριση διπλώματος χρησιμοποιούνται συνήθως χωρίς διάκριση.

Ένα από τα πρώτα δημόσια διαθέσιμα εργαλεία για ύφανση, ήταν το **THREADER** του David Jones (διαθέσιμο στην ιστοσελίδα http://bioinf.cs.ucl.ac.uk/?id=747), που υλοποιούσε τον αλγόριθμο του διπλού δυναμικού προγραμματισμού του 1992 και πλέον βρίσκεται μετά από διάφορες προσθήκες στην έκδοση 3.5 (Jones, 1998). Ένα από τα πρώτα εργαλεία που εφάρμοζαν τη μέθοδο με τη μετατροπή της δομής σε ένα μονοδιάστατο προφίλ, ήταν το **PHDthreader** του Burkhart Rost (Rost, Schneider, & Sander, 1997), το οποίο αποτελεί τμήμα των εφαρμογών που καλύπτονται από τον server Predict Protein (www.predictprotein.org). Το PHDthreader κάνει χρήση του PHD για την πρόγνωση της δευτεροταγούς δομής και μετά στοιχίζει τις δομές (την παρατηρηθείσα για το πρότυπο, με την προβλεφθείσα για το στόχο). Παρόμοια στρατηγική χρησιμοποιεί και το **genTHREADER** (Jones, 1999), το οποίο βασίζεται στην πρόγνωση δευτεροταγούς δομής του PSI-PRED, αλλά και σε ένα επιπλέον βήμα με χρήση νευρωνικού δικτύου για να δώσει μια συνολική τιμή για την αξιοπιστία της μεθόδου, ενώ είναι διαθέσιμο μαζί με τις υπόλοιπες προγνώσεις του συγκεκριμένου server (http://bioinf.cs.ucl.ac.uk/psipred/). Το genTHREADER συνδυάζεται εύκολα με το MODELLER που είδαμε παραπάνω, για να δώσει τρισδιάστατα μοντέλα σε περίπτωση μη ικανοποιητικής ομοιότητας.

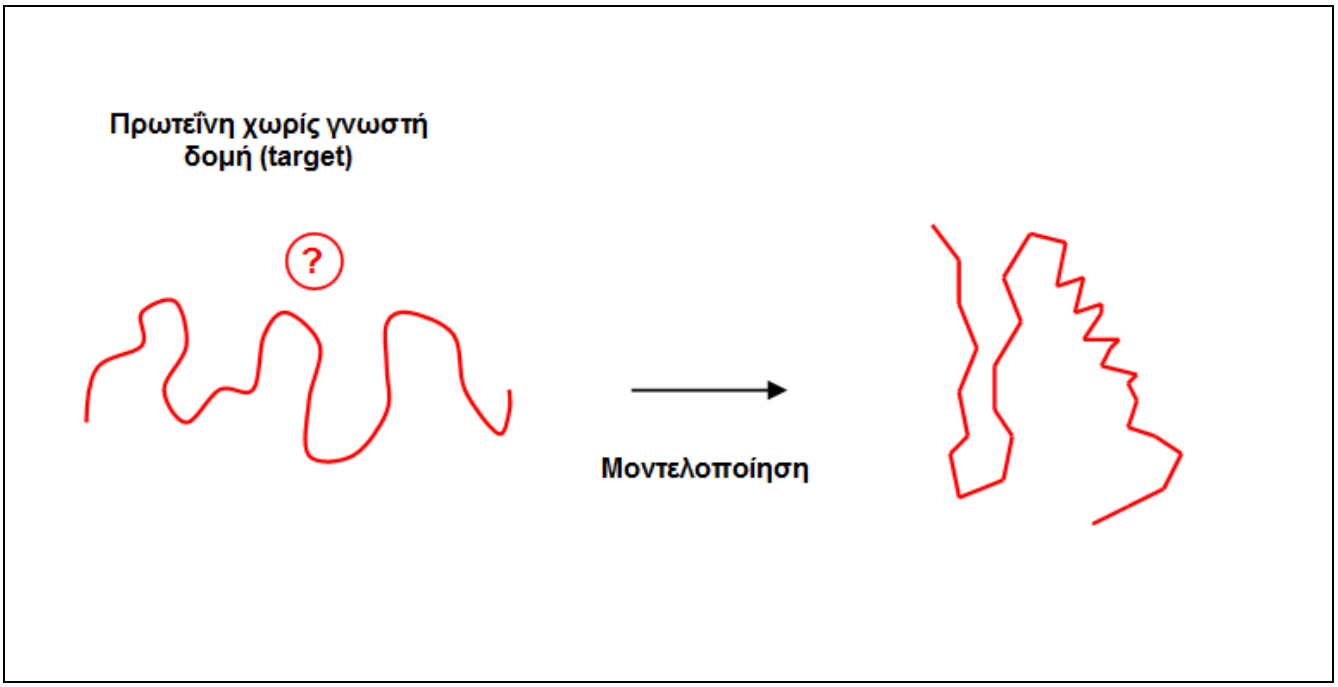
Μια σύγχρονη και ιδιαίτερα ικανοποιητική μέθοδος, είναι το **HHpred** (Söding, Biegert, & Lupas, 2005). Το HHpred βασίζεται σε μια ιδιαίτερα αποδοτική μέθοδο για στοίχιση και σύγκριση μεταξύ profile HMM (το HHsearch), κάτι που επιτρέπει ιδιαίτερα ευαίσθητες αναζητήσεις και εντοπισμό μακρινών ομολόγων (Söding, 2005). Η διαδικτυακή εφαρμογή δέχεται είσοδο είτε ακολουθία είτε μια πολλαπλή στοίχιση και επιτρέπει αναζήτηση σε διάφορες βάσεις (PDB, SCOP, PFAM, SMART κ.ο.κ.), τα αποτελέσματα επιστρέφονται πολύ γρήγορα σε κατανοητή μορφή, ενώ υπάρχει και διασύνδεση με το MODELLER για την παραγωγή του τρισδιάστατου μοντέλου (http://toolkit.tuebingen.mpg.de/hhpred). Το **Phyre2** είναι ένα άλλο παρόμοιο εργαλείο για αναγνώριση διπλώματος (http://www.sbg.bio.ic.ac.uk/phyre2). Η αρχική έκδοση, χρησιμοποιούσε έναν αλγόριθμο για στοίχιση profile-profile, βασισμένο σε PSSM, αλλά η νεότερη έκδοση χρησιμοποιεί και αυτή το HHsearch (Kelley, Mezulis, Yates, Wass, & Sternberg, 2015). Το Phyre2 ενσωματώνει διάφορες λειτουργίες όπως πρόγνωση δευτεροταγούς δομής με το PSI-RPED, πρόγνωση διαμεμβρανικών τμημάτων με το MEMSAT, πρόγνωση μη-κανονικών περιοχών με το DISOPRED, ενώ επιτρέπει πολλαπλές αναλύσεις όπως μελέτες προσδετών, μελέτες μη συνώνυμων πολυμορφισμών αλλά και ab initio προγνώσεις. Γενικά, αυτή η στρατηγική, να χρησιμοποιούνται σε ένα μόνο περιβάλλον, με απλό τρόπο χρήσης, όλες οι διαθέσιμες τεχνικές (πρόγνωση δευτεροταγούς δομής, μοντελοποίηση με βάση την ομολογία, αναγνώριση διπλώματος και ab initio προβλέψεις), αντιπροσωπεύει την κυρίαρχη τάση στις μεθόδους όπως θα δούμε και στις επόμενες παραγράφους.

Τελος, το **RaptorX** (http://raptorx.uchicago.edu/) και το **MUSTER** (http://zhang.bioinformatics.ku.edu/MUSTER) είναι δυο από τους πιο επιτυχημένους αλγόριθμους για ύφανση, καθώς δουλεύουν ικανοποιητικά ακόμα και σε περιπτώσεις κατά τις οποίες η ύπαρξη ομολόγων είναι περιορισμένη. Το RaptorX βασίζεται σε πιθανοθεωρητικά γραφικά μοντέλα και χρησιμοποιεί παράλληλα και την πληροφορία από τις δομές αλλά και από τις αλληλουχίες, ενώ χρησιμοποιεί και πληροφορία από όλα τα πιθανά πρότυπα για να χτίσει καλύτερα το μοντέλο (multiple template threading) (Peng & Xu, 2011). Το MUSTER χρησιμοποιεί δυναμικό προγραμματισμό, αλλά ενσωματώνει επίσης πολλαπλές πηγές πληροφορίας (δευτεροταγής δομή, προσβασιμότητα του διαλύτη, υδροφοβικότητα, πιθανές δίεδρες γωνίες κ.ο.κ.), ενώ κατασκευάζει το μοντέλο χρησιμοποιώντας διαφορετικό πρότυπο για κάθε πρωτεϊνική περιοχή του στόχου (Wu & Zhang, 2008). Τέλος, υπάρχει και στην περίπτωση της ύφανσης η περίπτωση της συνδυαστική πρόγνωσης με το **LOMETS** (http://zhanglab.ccmb.med.umich.edu/LOMETS/) το οποίο είναι ένας meta-server που χρησιμοποιεί 9 διαφορετικά εργαλεία (FFAS-3D, HHsearch, MUSTER, pGenTHREADER, PPAS, PRC, PROSPECT2, SP3, και SPARKS-X) για να παράγει έτσι μοντέλα

μεγαλύτερης πιστότητας (Wu & Zhang, 2007). Το LOMETS, χρησιμοποιείται από το **I-TASSER** (http://zhanglab.ccmb.med.umich.edu/I-TASSER/about.html), το οποίο αποτελεί σήμερα την καλύτερη και πιο ολοκληρωμένη λύση στην πρόγνωση τριτοταγούς δομής, πετυχαίνοντας την πρώτη θέση στους τελευταίους διαγωνισμούς του CASP. Το I-TASSER αναγνωρίζει τα πρότυπα και με τα διάφορα τμήματα κατασκευάζει ένα μοντέλο με μια τεχνική που ονομάζεται replica exchange Monte Carlo simulations και οι βρόχοι μοντελοποιούνται ab initio. Όταν κανένα πρότυπο δεν βρεθεί, τότε το λογισμικό θα κατασκευάσει μοντέλο με μέθοδο ab initio για ολόκληρη την πρωτεΐνη. Στο τελευταίο στάδιο γίνεται βελτιστοποίηση του μοντέλου με προσομοιώσεις Το I-TASSER, ενσωματώνει επίσης μια σειρά βελτιώσεις που επιτρέπουν στο χρήστη να εισάγει δομική πληροφορία με τη μορφή περιορισμών, όπως τις επαφές των αμινοξέων, τη δευτεροταγή δομή κ.ο.κ. Οι περιορισμοί αυτοί μπορεί να είναι ιδιαίτερα χρήσιμοι σε περίπτωση που τα πρότυπα είναι λίγα ή η ποιότητα της στοίχισης δεν είναι καλή (Roy, Kucukural, & Zhang, 2010).

9.4.3. Ab initio και de novo πρόγνωση δομής

Στην πιο ακραία περίπτωση, η πρωτεΐνη στόχος δεν μπορεί να ταυτοποιηθεί ούτε με βάση την ομολογία αλλά ούτε και με βάση το δίπλωμα. Το πρόβλημα σε αυτή την περίπτωση, καταλήγει στο πασίγνωστο πρόβλημα του πρωτεϊνικού διπλώματος, (protein folding problem) της πρόγνωσης δηλαδή της τρισδιάστατης δομής απευθείας από την αμινοξική αλληλουχία. Το πρόβλημα αυτό, είναι στην ουσία ένα από τα μεγαλύτερα προβλήματα της σύγχρονης βιολογίας και δεκάδες ερευνητές έχουν ασχοληθεί (προφανώς, είναι ένα δύσκολο πρόβλημα καθώς έχει αποδειχτεί ότι είναι NP-complete). Γενικά, υπάρχουν δύο όροι για να περιγράψουν τις μεθόδους αυτές, και αν και πολλές φορές χρησιμοποιούνται αδιάκριτα μεταξύ τους, είναι καλό να πραγματοποιούμε το διαχωρισμό. Έτσι, με τον όρο ab initio πρόγνωση, παραδοσιακά αναφερόμαστε στην πρόγνωση με χρήση μόνο των βασικών αρχών της φυσικής (αλληλεπιδράσεις ατόμων και υπολογισμοί ενέργειας). Από την άλλη, ο όρος de Novo πρόγνωση, είναι κάπως πιο γενικός και αναφέρεται σε όλες τις μεθόδους που επιχειρούν πρόγνωση χωρίς τη χρήση προτύπου με γνωστή δομή. Γενικά πάντως δεν υπάρχει απόλυτη συμφωνία ως προς το σε ποια ακριβώς κατηγορία κατατάσσεται κάθε μέθοδος, ειδικά εφόσον οι περισσότερες από αυτές χρησιμοποιούν συνδυασμό μεθοδολογιών.



Εικόνα 9.11: *Σχηματική αναπαράσταση της ab initio πρόγνωσης της δομής*

.

Γενικά το θέμα της πρόγνωσης της τρισδιάστατης δομής των πρωτεϊνών έχει απασχολήσει τους επιστήμονες για δεκαετίες. Οι εργασίες του Anfinsen έδειξαν μεν ότι οι αλληλουχίες των πρωτεϊνών καθορίζουν μονοσήμαντα την τρισδιάστατη δομή οδηγώντας στη δομή με την ελάχιστη ενέργεια, αλλά το παράδοξο του Levinthal έδειξε με ξεκάθαρο τρόπο ότι οι πιθανές διαμορφώσεις μιας πρωτεΐνης δεν είναι δυνατό να δοκιμαστούν όλες. Για παράδειγμα, αν μια πρωτεΐνη έχει 100 αμινοξέα (ένας κάπως μικρός

αριθμός), υπάρχουν 99 πεπτιδικοί δεσμοί και κατά συνέπεια 198 διαφορετικές γωνίες φ και ψ οι οποίες μπορούν να περιστραφούν ελεύθερα. Αν υποθέσουμε ότι κάθε γωνία έχει μόνο 3 πιθανές τιμές (πάλι ένας μετριοπαθής υπολογισμός), τότε οι πιθανές διαμορφώσεις ολόκληρου του πρωτεϊνικού μορίου είναι 3^{198}, ένας αριθμός εξωπραγματικός. Αν η πρωτεΐνη έπρεπε να δοκιμάσει με τυχαίες κινήσεις όλες τις πιθανές διαμορφώσεις, τότε δεν θα προλάβαινε να διπλωθεί σωστά ακόμα και αν περιμέναμε ως το... τέλος του σύμπαντος προσπαθώντας. Στην πράξη βέβαια, οι πρωτεΐνες διπλώνονται σε χρόνους της τάξης των microsecond ή millisecond, κάτι που σημαίνει ότι λειτουργεί κάποιος άλλος μηχανισμός (έχουν προταθεί διάφοροι τρόποι, όπως το μοντέλο της υδροφοβικής κατάρρευσης, το μοντέλο του σχηματισμού πυρήνων δευτεροταγούς δομής κ.ο.κ.). Σε κάθε περίπτωση, όλη αυτή η 'συζήτηση', εκτός από θεωρητική σημασία, έως και σήμερα βασικό ρόλο στις προσπάθειες ab initio ή de novo πρόγνωσης της δομής.

Γενικά με βάση τα παραπάνω, οι προσπάθειες ab initio πρόγνωσης, είναι (παρόλες τις αλγοριθμικές επινοήσεις), ιδιαίτερα απαιτητικές υπολογιστικά και για χρόνια ήταν περιορισμένες σε μικρές πρωτεΐνες (μέχρι 50 αμινοξέα μήκος), και ακόμα και σε αυτές τις περιπτώσεις τα αποτελέσματα χρειάζονταν πολύ χρόνο. Τα τελευταία χρόνια τόσο η αύξηση της υπολογιστικής ισχύος, αλλά και νέες αλγοριθμικές τεχνικές επέτρεψαν την πρόγνωση με σχετική ακρίβεια, για μεγαλύτερες πρωτεΐνες, και σε εύλογο χρονικό διάστημα (μερικές ώρες ή μέρες). Τα γενικά θέματα που έχει να αντιμετωπίσει μια τέτοια μέθοδος είναι τρία:

- *Η αναπαράσταση της πρωτεϊνικής δομής*. Στην ιδανική περίπτωση θα έπρεπε στους υπολογισμούς να λαμβάνουν μέρος όλα τα άτομα της πρωτεΐνης, αλλά κάτι τέτοιο είναι απαγορευτικό από πλευράς υπολογιστικής ισχύος. Έτσι, έχουν χρησιμοποιηθεί διαφορετικές προσεγγίσεις: από την απλή χρήση μόνο του Cα, την προσθήκη του Cβ, μέχρι και σύνθετες μετρήσεις στις οποίες ολόκληρη η πλευρική ομάδα αντικαθίσταται από ένα σημείο με τη συνολική μάζα στο κέντρο βάρους. Οι επιτρεπτές γωνίες είναι επίσης ένας σημαντικός παράγοντας σε αυτό το σημείο. Έτσι, κάποιες μέθοδοι επιτρέπουν μόνο προκαθορισμένες γωνίες φ και ψ, ενώ άλλες υπολογίζουν τη δομή κομματιών μήκους 6-7 αμινοξέων για να ελαττώσουν ακόμα περισσότερο το χρόνο.
- *Ο υπολογισμός της ενέργειας*. Το κομμάτι αυτό αφορά το πώς θα αξιολογηθεί μια δομή ως «καλή». Θα πρέπει δηλαδή να υπάρχει ένα κριτήριο που να ξεχωρίζει τις δομές ελάχιστης ενέργειας. Η πιο προφανής λύση εδώ, είναι η χρήση καθαρά φυσικοχημικών τεχνικών κατά τις οποίες υπολογίζονται οι ελκτικές και απωστικές δυνάμεις μεταξύ όλων των ατόμων, σε μια προσπάθεια να μιμηθούμε το δίπλωμα των πρωτεϊνών στη φύση. Οι μεθοδολογίες αυτής της κατηγορίας περιλαμβάνουν τα πεδία AMBER, CHARMM, UNRES και ASTRO-FOLD. Η άλλη εναλλακτική είναι να χρησιμοποιηθεί μια εμπειρική συνάρτηση η οποία θα έχει προκύψει από στατιστικές μετρήσεις (τέτοιες συναρτήσεις χρησιμοποιούνται από το ROSSETA και το TASSER/I-TASSER).
- *Η στρατηγική αναζήτησης*. Αυτό το σημείο αναφέρεται στο πώς θα γίνει η αναζήτηση στο χώρο των πιθανών διαμορφώσεων για την εύρεση της δομής με την ελάχιστη ενέργεια. Η πιο συνηθισμένη μέθοδος εδώ, είναι η προσομοίωση Monte Carlo, αλλά έχουν χρησιμοποιηθεί και άλλες στατιστικές τεχνικές όπως το Simulated Annealing (προσομοίωση ανώπτησης), αλλά και τεχνικές της τεχνητής νοημοσύνης όπως οι γενετικοί αλγόριθμοί. Μια άλλη μεγάλη κατηγορία μεθόδων είναι η Μοριακή Δυναμική (Molecular Dynamics), κατά την οποία επιλύονται οι εξισώσεις κίνησης του Νεύτωνα και προσομοιώνεται η κίνηση των ατόμων στο χρόνο. Η τεχνική αυτή είναι η πιο αξιόπιστη, αλλά καθώς συνήθως συνδυάζεται με συνάρτηση ενέργειας φυσικοχημικού τύπου, απαιτεί πάρα πολλούς υπολογισμούς. Κατά συνέπεια, είναι εφαρμόσιμη περισσότερο σε περιπτώσεις που μας ενδιαφέρει η διαδικασία διπλώματος μιας πρωτεΐνης. Μοριακή δυναμική επίσης χρησιμοποιείται γενικά για τη μοντελοποίηση των βρόχων και για τη βελτιστοποίηση ενός ήδη κατασκευασμένου μοντέλου.

Τα πιο γνωστά και παλιά προγράμματα μοριακής δυναμικής, είναι το **CHARMM** (http://www.charmm.org/) και το **AMBER** (http://ambermd.org/) τα οποία συμβαίνει να αντιστοιχούν και στις δύο πιο γνωστές κατηγορίες δυναμικών πεδίων για υπολογισμούς μοριακής δυναμικής. Το CHARMM (Chemistry at HARvard Macromolecular Mechanics) αποτελεί ένα μεγάλο συνεργατικό πρόγραμμα με πολλούς ερευνητές, υπό την καθοδήγηση του Martin Karplus στο Harvard (Brooks et al., 2009). Είναι το παλιότερο πρόγραμμα μοριακής δυναμικής και έχει εξελιχθεί με τα χρόνια έτσι ώστε να παρέχει πολλές διαφορετικές λειτουργίες και επιλογές (προσομοίωση, ελαχιστοποίηση ενέργειας μιας δεδομένης δομής, και

υπολογισμοί μοριακής δυναμικής). Το AMBER (Assisted Model Building with Energy Refinement) ξεκίνησε αρχικά σαν μια άλλη οικογένεια δυναμικών πεδίων που αναπτύχθηκαν από τον Peter Kollman στο University of California και το ομώνυμο πακέτο αναπτύχθηκε σαν υλοποίηση αυτών των υπολογισμών (Case et al., 2005). Ένα μεγάλο μειονέκτημα και των δύο πακέτων είναι ότι δεν είναι ελεύθερα διαθέσιμα. Μια επιλογή για αντίστοιχο λογισμικό ανοιχτού κώδικα, είναι το **GROMACS** (http://www.gromacs.org). Το GROMACS παρέχει πολλές δυνατότητες για μοντελοποίηση πολλών κατηγοριών βιομορίων με χρήση διαφορετικών πεδίων, ενώ παρέχει και δυνατότητες παράλληλης επεξεργασίας ακόμα και σε συστήματα Windows (Pronk et al., 2013). Αυτό που πρέπει να τονιστεί βέβαια, είναι ότι τα προγράμματα αυτά δεν μπορούν να χρησιμοποιηθούν σε πρακτικές εφαρμογές για ab initio πρόγνωση της δομής πρωτεϊνών, αλλά κυρίως για βελτιστοποίηση μιας υπάρχουσας δομής, για τη μελέτη της διαδικασίας του διπλώματος και για μελέτη των αλληλεπιδράσεων με άλλα μόρια.

Το πιο γνωστό από τα προγράμματα για ab initio/de novo πρόγνωση τρισδιάστατης δομής, είναι το **ROSETTA** (http://robetta.bakerlab.org/). Το ROSETTA αρχικά αναγνωρίζει τις πρωτεϊνικές περιοχές, και στη συνέχεια τις μοντελοποιεί με μια γρήγορη ab initio μεθοδολογία που χρησιμοποιεί τα μικρότερα τμήματα (κυρίως 9μερή) από γνωστές δομές της PDB. Η μεθοδολογία αυτή βασίζεται σε μια ιδέα των Bowie και Eisenberg από το 1994. Το αρχικό μοντέλο κατασκευάζεται με χρήση μόνο της κύριας ανθρακικής αλυσίδας και των Cβ ενώ στη συνέχεια κάποια από τα καλύτερα (από άποψη ενέργειας) μοντέλα υφίστανται βελτιστοποίηση με όλα τα άτομα παρόντα, κάνοντας χρήση προσομοίωσης Monte Carlo και μιας στατιστικής φύσεως συνάρτησης ενέργειας (Rohl, Strauss, Misura, & Baker, 2004).

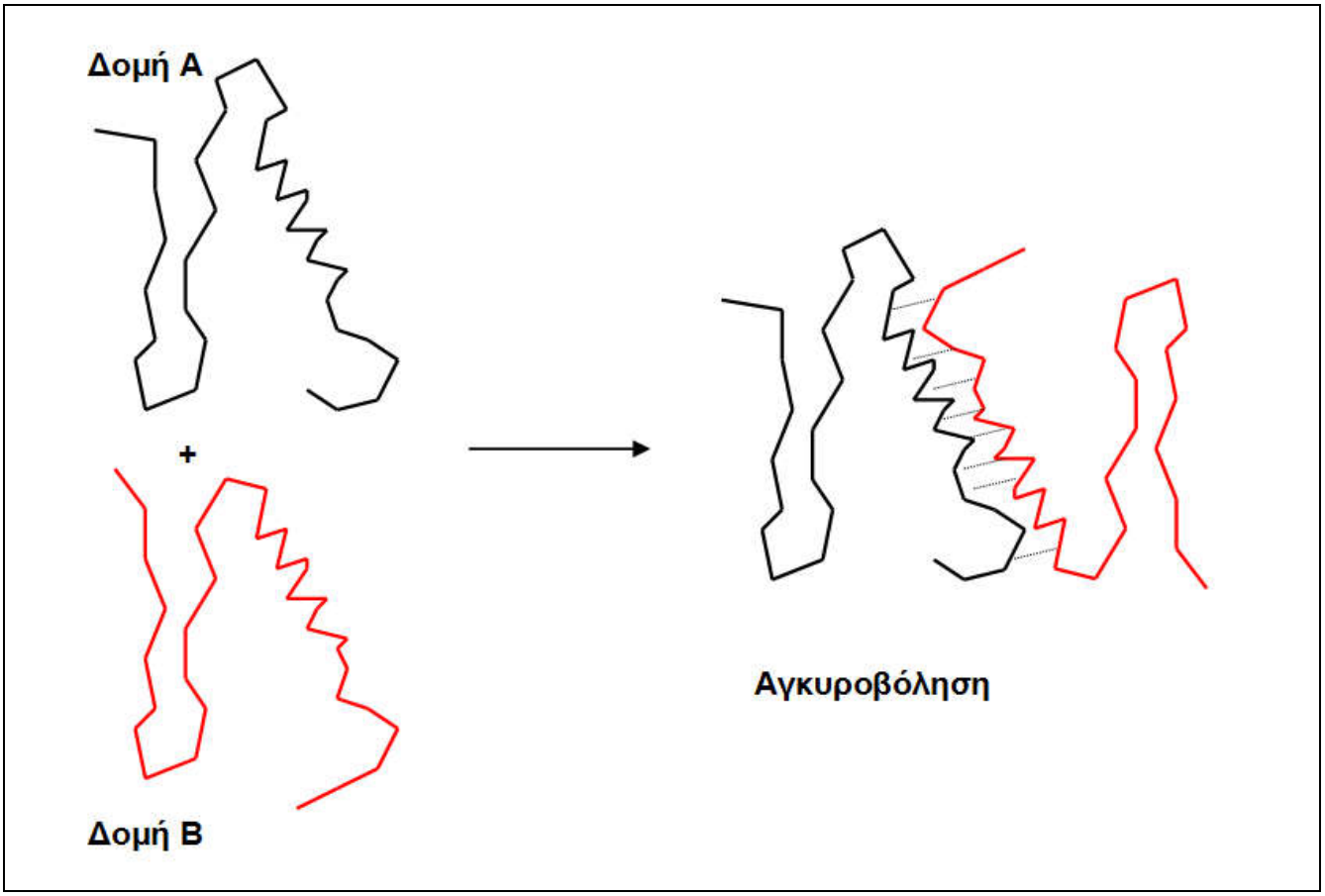
Όπως αναφέραμε ήδη, το **I-TASSER** (http://zhanglab.ccmb.med.umich.edu/I-TASSER/about.html) είναι μια εφαρμογή που πραγματοποιεί και ύφανση αλλά και ab initio μοντελοποίηση όταν δεν μπορεί να εντοπίσει δομές με παρόμοιο δίπλωμα. Το I-TASSER, είναι σήμερα η καλύτερη και πιο ολοκληρωμένη λύση στην πρόγνωση τριτοταγούς δομής, όπως πιστοποιείται από την πρώτη θέση που καταλαμβάνει στους τελευταίους διαγωνισμούς του CASP αλλά και σε εμπειρικές μελέτες αξιολόγησης (Helles, 2008). Το μεγάλο του πλεονέκτημα, είναι εκτός από την ακρίβεια στην πρόβλεψη, η μεγάλη ταχύτητα στους υπολογισμούς. Και το I-TASSER και το Rosetta χρησιμοποιούν προσομοίωση Monte Carlo (αν και με διαφορετικές παραλλαγές), συναρμογή τμημάτων και στατιστικής φύσεως συναρτήσεις ενέργειας, αλλά διαφέρουν στην αναπαράσταση της δομής και στις αποδεκτές δίεδρες γωνίες. Άλλες δημόσια διαθέσιμες μέθοδοι λιγότερο γνωστές είναι το **ePROPAINOR** (http://www.math.iitb.ac.in/epropainor) και το **PROTinfo** (http://ram.org/compbio/protinfo/), οι οποίες όμως δεν είναι τόσο επιτυχημένες (πάντα σε σχέση με το I-TASSER και το ROSETTA). Επίσης, υπάρχουν μια σειρά από μέθοδοι όπως το **QUARK** (http://zhanglab.ccmb.med.umich.edu/QUARK/), το **CABSfold** (http://biocomp.chem.uw.edu.pl/CABSfold/), το **PEP-FOLD** (http://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD/) και το **BHAGEERATH** (http://www.scfbio-iitd.res.in/bhageerath/index.jsp), οι οποίες όμως ενδείκνυνται περισσότερο για πεπτίδια και μικρές πρωτεΐνες (<100 αμινοξέα), καθώς ο υπολογιστικός χρόνος για μεγαλύτερους υπολογισμούς είναι απαγορευτικός.

Τέλος, αξίζει μια ειδική αναφορά στα κατανεμημένα (distributed) συστήματα ab initio πρόγνωσης. Τέτοιου είδους εφαρμογές, ξεκίνησαν με το **ROSETTA@home** (http://boinc.bakerlab.org/rosetta/) και το **Folding@home** (http://folding.stanford.edu/). Με τις μεθοδολογίες αυτές, ο χρήστης που έχει εγκαταστήσει την ειδική εφαρμογή «δανείζει» υπολογιστικό χρόνο από τον υπολογιστή του όταν αυτός δεν λειτουργεί, με σκοπό να βοηθήσει στην επίλυση του προβλήματος του διπλώματος «δύσκολων» πρωτεϊνών. Πολλοί από τους επιστήμονες που είχαν εμπλοκή στο σχέδιο του ROSETTA@home, αποφάσισαν αργότερα να εμπλέξουν ακόμα περισσότερους χρήστες και να αναπτύξουν ένα παιχνίδι που θα προσομοιώνει το δίπλωμα των πρωτεϊνών. Η ιδέα ήταν να χρησιμοποιηθούν οι ικανότητες αναγνώρισης προτύπων που διαθέτει ο ανθρώπινος εγκέφαλος, και να εφαρμοστούν σε παρόμοιες δύσκολες περιπτώσεις. Έτσι αναπτύχθηκε το **FOLDit** (http://fold.it/portal/) στο οποίο οι χρήστες σε ένα είδος παιχνιδιού στον Η/Υ κατασκευάζουν μοντέλα τρισδιάστατης δομής γνωστών πρωτεϊνών προτείνοντας τη δομή με το κατάλληλο δίπλωμα (δηλαδή, με τη μικρότερη ενέργεια). Η ιδέα είναι ότι το σύστημα μπορεί να «εκπαιδευτεί» με τις λύσεις που προτείνει ο ανθρώπινος εγκέφαλος, έτσι ώστε ένα αυτοματοποιημένο παρόμοιο σύστημα να μπορέσει να υλοποιηθεί αργότερα.

9.5. Αγκυροβόληση

Με τον όρο αγκυροβόληση ή ελλιμενισμό (docking) εννοούμε τη διαδικασία με την οποία υπολογίζουμε ή προβλέπουμε τον προτιμώμενο προσανατολισμό ενός μορίου σε σχέση με ένα άλλο όταν σχηματίζουν ένα

σταθερό σύμπλοκο. Στη διαδικασία αυτή, γίνεται η υπόθεση ότι το σταθερό αυτό σύμπλοκο βρίσκεται σε μια διαμόρφωση ελάχιστης ενέργειας. Το σύμπλοκο το οποίο θα επιχειρήσουμε να μοντελοποιήσουμε με τη διαδικασία της αγκυροβόλησης μπορεί να είναι μεταξύ δύο πρωτεϊνών (Bonvin, 2006; Gray, 2006; Sternberg, Gabb, & Jackson, 1998), αλλά και μεταξύ μιας πρωτεΐνης και ενός μικρού μορίου το οποίο μπορεί να είναι ορμόνη, φάρμακο, αναστολέας, βιταμίνη κ.ό.κ. (Taylor, Jewsbury, & Essex, 2002). Φυσικά, υπάρχουν και περιπτώσεις αλληλεπιδράσεων DNA-πρωτεϊνών αλλά και DNA-μικρών μορίων. Η γνώση αυτή, μπορεί να είναι χρήσιμη στο να κατανοήσουμε το βιολογικό μηχανισμό της λειτουργίας της πρωτεΐνης, την ένταση και την ισχύ της δέσμευσης του μικρού μορίου, το μηχανισμό λειτουργίας, αλλά και τον μηχανισμό με τον οποίο αλληλεπιδρούν δυο πρωτεΐνες είτε σαν ένζυμο-υπόστρωμα, είτε σαν υποδοχέας-προσδέτης, αλλά και γενικότερα στη μελέτη των πρωτεϊνικών αλληλεπιδράσεων και της τεταρτοταγούς δομής. Ειδικά στην περίπτωση των μικρών μορίων, η αγκυροβόληση βρίσκει πολλές εφαρμογές στο σχεδιασμό νέων φαρμάκων, και λόγω της σημασίας αυτής της διαδικασίας, στη φαρμακευτική βιομηχανία, έχει δοθεί μεγάλη ώθηση στο πεδίο από τέτοιες αντίστοιχες μελέτες (Alvarez, 2004).



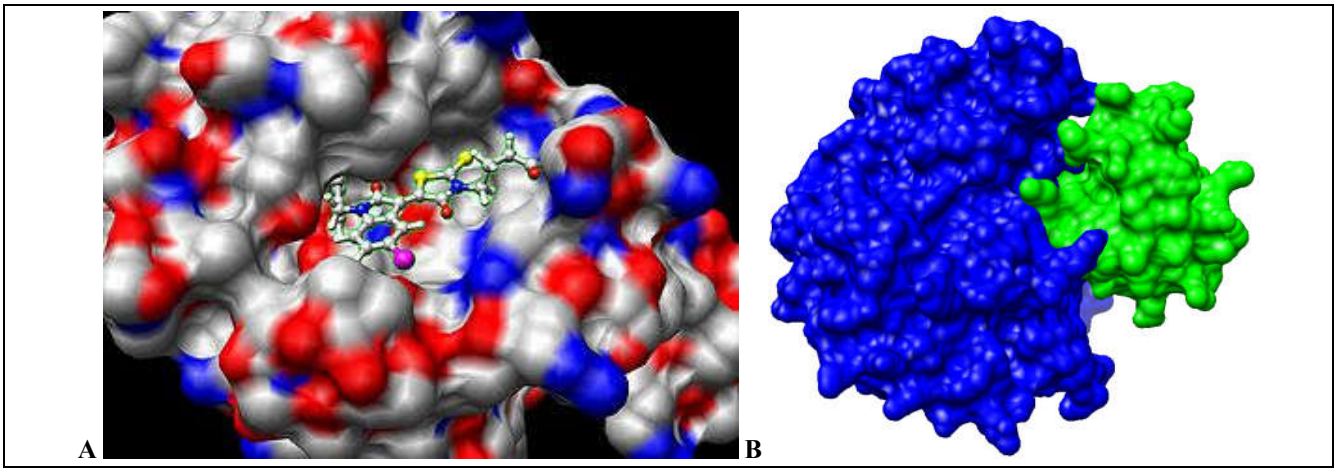
Εικόνα 9.12: *Σχηματική αναπαράσταση της αγκυροβόλησης δύο πρωτεϊνικών δομών.*

Το πρόβλημα της αγκυροβόλησης μπορούμε να το δούμε ανατρέχοντας στις γνωστές θεωρίες για τη δράση των ενζύμων και των πρωτεϊνών γενικότερα. Έτσι, η πιο απλή προσέγγιση κάνει λόγο για το μοντέλο «κλειδιού-κλειδαριάς», σύμφωνα με το οποίο οι επιφάνειες των δύο πρωτεϊνών είναι συμπληρωματικές ή το ενεργό κέντρο του ενζύμου είναι συμπληρωματικό σαν γεωμετρικό σχήμα με το υπόστρωμα (ή, του υποδοχέα με τον προσδέτη κ.ο.κ.). Σύμφωνα με αυτή τη θεωρία, αναπτύχθηκαν οι πρώτες μέθοδοι αγκυροβόλησης, οι λεγόμενες μέθοδοι «αγκυροβόλησης σταθερού σώματος» (rigid docking), σύμφωνα με τις οποίες οι τρισδιάστατες δομές του κάθε μορίου δεν αλλάζουν (δηλαδή τα άτομά τους δεν αλλάζουν καθόλου τη σχετική τους θέση), αλλά απλά μετακινούνται για να βρεθεί η επιφάνεια επαφής. Παρ' όλα αυτά, ξέρουμε ότι το μοντέλο αυτό δεν είναι επαρκές καθώς σε πολλές περιπτώσεις η πρόσδεση επηρεάζει (σε μικρότερο ή μεγαλύτερο βαθμό) τη διαμόρφωση του κάθε μορίου (το μοντέλο της «επαγόμενης προσαρμογής»). Αυτό οδήγησε σε πιο σύνθετες τεχνικές αγκυροβόλησης, τις λεγόμενες μεθοδολογίες «ευέλικτης αγκυροβόλησης»

(flexible docking) στις οποίες η τρισδιάστατη δομή των μορίων αλλάζει (έστω και ελάχιστα) για να επιτευχθεί η καλύτερη δυνατή αναγνώριση.

Από άποψη υπολογιστικής μεθοδολογίας, και σύμφωνα με τα παραπάνω, μπορούμε να διακρίνουμε, δύο κατηγορίες προσεγγίσεων στην αγκυροβόληση. Στην πρώτη περίπτωση, έχουμε τις προσεγγίσεις που βασίζονται στη συμπληρωματικότητα του σχήματος. Στις μεθοδολογίες αυτής της κατηγορίας, τα εμπλεκόμενα βιομόρια αντιμετωπίζονται ως τρισδιάστατα σχήματα και η συμπληρωματικότητα επιτυγχάνεται με μετακίνηση των δομών με τρόπο που να τις κάνει να συμπίπτουν όσο το δυνατό καλύτερα. Οι μεθοδολογίες αυτής της κατηγορίας είναι γρήγορες και σταθερές, αλλά με τις απλουστεύσεις που κάνουν δεν μπορούν να δώσουν τα βέλτιστα αποτελέσματα. Έτσι, χρησιμοποιούνται συνήθως στα αρχικά στάδια των μελετών για πιθανούς στόχους για φάρμακα, ούτως ώστε να γίνει μια γρήγορη διαλογή των πιθανών στόχων. Λόγω του ότι βασίζονται κυρίως στη γεωμετρική αναπαράσταση των δομών, στις μεθοδολογίες αυτές χρησιμοποιείται ιδιαίτερα η προσέγγιση των «φαρμακοφόρων».

Στη δεύτερη κατηγορία μεθόδων, ανήκουν οι μέθοδοι που βασίζονται στην προσομοίωση. Οι μεθοδολογίες αυτές είναι πιο σύνθετες και πιο απαιτητικές και μοιάζουν αρκετά με τις αντίστοιχες μεθοδολογίες της ab initio πρόγνωσης που είδαμε στην προηγούμενη ενότητα. Με λίγα λόγια, τα μόρια του ζευγαριού πρωτεΐνη-πρωτεΐνη ή πρωτεΐνη-μικρό μόριο, αφήνονται σε μια κάποια απόσταση και μέσω της προσομοίωσης επιχειρείται μέσα από διαδοχικές «κινήσεις» να βρεθεί η καλύτερη, από άποψη ελεύθερης ενέργειας, αλληλεπίδραση μεταξύ τους. Οι κινήσεις μπορεί να αφορούν τόσο μετακινήσεις ολόκληρου του μορίου αλλά και σχετικές μεταβολές στη στερεοδιάταξή του έτσι ώστε να βρεθεί η καλύτερη πιθανή διαμόρφωση. Όπως είναι φανερό, οι μεθοδολογίες αυτές είναι περισσότερο ρεαλιστικές, αλλά ιδιαίτερα χρονοβόρες και όπως και στην περίπτωση της ab initio πρόγνωσης μόνο τα τελευταία χρόνια, με τη ανάπτυξη ισχυρών υπολογιστών και την έμφαση στην παράλληλη επεξεργασία, τέτοιες μέθοδοι απέκτησαν ευρεία χρήση.



Εικόνα 9.13: *Αριστερά, αγκυροβόληση πρωτεΐνης με μικρό μόριο (από https://en.wikipedia.org/wiki/Docking_%28molecular%29). Δεξιά, αγκυροβόληση δύο πρωτεϊνών, της HDAC3 με την NCOR2 (από http://www.zbi.uni-saarland.de/de/%C3%BCber-bioinformatik/docking.html)*

Οι μεθοδολογίες αυτές, έχουν πολλά κοινά με τις αντίστοιχες που χρησιμοποιούνται στην ab initio πρόγνωση δομής και ειδικά στο κομμάτι της βελτιστοποίησης, καθώς στην αγκυροβόληση ξεκινάμε σχεδόν πάντα από βιομόρια γνωστής ή σχεδόν γνωστής δομής. Έτσι, δύο είναι τα σημαντικότερα προβλήματα στην αγκυροβόληση: ο υπολογισμός της ενέργειας και η στρατηγική αναζήτησης (Halperin, Ma, Wolfson, & Nussinov, 2002; Moreira, Fernandes, & Ramos, 2010). Αντιθέτως, η αναπαράσταση της δομής συνήθως δεν είναι, γιατί εδώ ενδιαφερόμαστε για μελέτη όλων των ατόμων του μορίου. Επίσης, μια άλλη διαφορά είναι ότι επιχειρούμε μοντελοποίηση και των διαμοριακών αλληλεπιδράσεων και όχι μόνο των ενδομοριακών.

Πακέτα λογισμικού κατάλληλα για αγκυροβόληση, υπάρχουν δεκάδες, τόσο σε αυτόνομες εφαρμογές όσο και σε διαδικτυακές. Μια ιδιαιτερότητα σε σχέση με άλλες κατηγορίες λογισμικού Βιοπληροφορικής είναι το γεγονός ότι καθώς η αγκυροβόληση βρίσκει πολλές εφαρμογές στο σχεδιασμό φαρμάκων (computer aided drug discovery), υπάρχουν και πολλές εφαρμογές που είναι εμπορικές. Στην ιστοσελίδα του Swiss Institude for Bioinformatics υπάρχει αναλυτική λίστα με όλα τα λογισμικά για τα διάφορα στάδια στην ανακάλυψη φαρμάκων και στην αντίστοιχη κατηγορία για την αγκυροβόληση αναφέρονται δεκάδες πακέτα

λογισμικού (http://www.click2drug.org/directory_Docking.html). Παρακάτω θα προσπαθήσουμε να κάνουμε μια σύντομη αναφορά στα πιο σημαντικά από αυτά τα πακέτα, παρουσιάζοντας τα βασικά πλεονεκτήματα του καθενός (Rodrigues & Bonvin, 2014). Γενικά, οι παράγοντες που παίζουν ρόλο στην αποτελεσματικότητα ενός τέτοιου λογισμικού είναι, α) η ταχύτητα, β) η σωστή εύρεση της επιφάνειας επαφής, γ) η δυνατότητα να χειριστεί αγκυροβόληση πρωτεΐνης-πρωτεΐνης, δ) η δυνατότητα να πραγματοποιήσει ευέλικτη αγκυροβόληση και δ) η δυνατότητα να ορίζει ο χρήστης τις πιθανές επιφάνειες επαφής κάνοντας χρήση εξωτερικής πληροφορίας.

Ένα από τα πιο γνωστά και ευρέως χρησιμοποιούμενα προγράμματα για αγκυροβόληση είναι το **GRAMM** (http://vakser.bioinformatics.ku.edu/resources/gramm/gramm1/). Το GRAMM (από τα αρχικά Global RAnge Molecular Matching) χρησιμοποιεί εμπειρική συνάρτηση ενέργειας και εκτελεί εκτεταμένες περιστροφές και μετακινήσεις των μορίων για να εντοπίσει την πιθανή θέση πρόσδεσης και μπορεί να χρησιμοποιηθεί σε ευρύ φάσμα συνθηκών, τόσο για αγκυροβόληση μικρών μορίων, όσο και για αγκυροβόληση πρωτεϊνών αλλά και πρωτεϊνικών περιοχών. Επίσης, μπορεί να χρησιμοποιηθεί τόσο για δομές υψηλής ανάλυσης όσο και για δομές πιο χαμηλής ανάλυσης. Η ποιότητα της πρόβλεψης εξαρτάται όμως από την ακρίβεια των δομών. Έτσι, μια αγκυροβόληση σε δομή μεγάλης διακριτικότητας με μικρές αλλαγές στη στερεοδιάταξη, θα δώσει πιο αξιόπιστες προβλέψεις σε σχέση με μια περίπτωση λ.χ. με δομές χαμηλής διακριτικότητας, όπου και θα πάρουμε μόνο τα γενικά χαρακτηριστικά του συμπλόκου. Υπάρχει επίσης και μια άλλη έκδοση με βελτιωμένους αλγόριθμους για αγκυροβόληση πρωτεϊνών, το **GRAMM-X** (http://vakser.compbio.ku.edu/resources/gramm/grammx/), το οποίο είναι ιδιαίτερα γρήγορο αλλά δεν μπορεί να χειριστεί ευέλικτα σύμπλοκα.

Το **AutoDock** (http://autodock.scripps.edu/) είναι ένα ολόκληρο πακέτο με εργαλεία αγκυροβόλησης. Χρησιμοποιείται κυρίως για την αγκυροβόληση μικρών μορίων και αυτή τη στιγμή υπάρχουν δύο εκδόσεις του πακέτου: το AutoDock 4 και το AutoDock Vina. Το πρώτο επιτρέπει περισσότερες παρεμβάσεις του χρήστη στην οπτικοποίηση του πλέγματος στο οποίο θα γίνει η αγκυροβόληση, κάτι που μπορεί να βοηθήσει τους χημικούς στη σύνθεση μικρών μορίων. Το δεύτερο κάνει αυτούς τους υπολογισμούς εσωτερικά και είναι πιο αυτοματοποιημένο. Επίσης υπάρχει και μια γραφική διεπαφή, το AutoDockTools, εργαλείο το οποίο βοηθάει το χρήστη να επιλέξει τους δεσμούς που θα περιστρέφονται στον προσδέτη και στην ανάλυση την αγκυροβόλησης.

Το **HADDOCK** (High Ambiguity Driven protein-protein DOCKing) είναι μια ιδιαίτερα δημοφιλής εφαρμογή για αγκυροβόληση η οποία χρησιμοποιείται κυρίως για αλληλεπιδράσεις πρωτεϊνών. Το HADDOCK διακρίνεται από τις υπόλοιπες ab initio προσεγγίσεις στο ότι δέχεται εξωτερική πληροφορία για τις πιθανές περιοχές επαφής (http://haddocking.org/). Ο χρήστης δίνει τα δύο μόρια και μια λίστα πιθανών (γνωστών ή προβλεφθέντων) καταλοίπων της επιφάνειας επαφής για να κατευθύνει με αυτόν τον τρόπο τη διαδικασία της αγκυροβόλησης. Η διαδικτυακή εφαρμογή είναι ιδιαίτερα εύχρηστη για την πραγματοποίηση της ανάλυσης, ενώ υπάρχουν και επιπλέον επιλογές για την πλήρη εκμετάλλευση των δυνατοτήτων του HADDOCK και για την εξατομίκευση της διαδικασίας.

Το **FTDock** (http://www.sbg.bio.ic.ac.uk/docking/ftdock.html) είναι μια ιδιαίτερα γρήγορη εφαρμογή αγκυροβόλησης η οποία βασίζεται στη συμπληρωματικότητα των σχημάτων. Ο αλγόριθμος επεξεργάζεται το σχήμα των μορίων χρησιμοποιώντας μετασχηματισμούς Fourier και προαιρετικά εφαρμόζει και ένα ηλεκτροστατικό φίλτρο.

Το **DOT** (http://www.sdsc.edu/CCMS/DOT/) είναι μια εφαρμογή για αγκυροβόληση που μπορεί να δεχτεί σαν δεδομένα εισόδου τόσο ζευγάρια πρωτεϊνών-πρωτεϊνών όσο και άλλες κατηγορίες μορίων. Το DOT εργάζεται με αλγόριθμο σταθερής αγκυροβόλησης που ψάχνει αναλυτικά όλες τις πιθανές διευθετήσεις του ενός μορίου σε σχέση με το άλλο. Στον υπολογισμό της ενέργειας υπολογίζονται τα ηλεκτροστατικά δυναμικά αλλά και οι αλληλεπιδράσεις van der Waals, ενώ κάνει και χρήση μετασχηματισμών Fourier.

Το **ZDOCK** (http://www.umassmed.edu/zlab/) είναι ένα άλλο πετυχημένο εργαλείο για γρήγορη αγκυροβόληση και εύρεση των αλληλεπιδράσεων μεταξύ δύο πρωτεϊνών. Βασίζεται σε μια μεθοδολογία «στέρεας» αγκυροβόλησης με συμπληρωματικότητα σχημάτων, με ειδικές συναρτήσεις για υπολογισμό των ηλεκτροστατικών αλληλεπιδράσεων. Το ZDOCK είναι ιδιαίτερα γρήγορο αλλά δεν μπορεί να χειριστεί ευέλικτα σύμπλοκα.

Το **ClusPro** (http://cluspro.bu.edu/), είναι ένας άλλος αλγόριθμος που έχει δώσει πολύ καλά αποτελέσματα σε αξιολογήσεις. Στηρίζεται σε μια γρήγορη αναζήτηση με βάση τη συμπληρωματικότητα των σχημάτων με χρήση μετασχηματισμών Fourier. Στο δεύτερο στάδιο πραγματοποιεί ομαδοποίηση με βάση το RMSD και στο τέλος βελτιστοποιεί τις επιλεγμένες δομές με το CHARMM. Ένα μειονέκτημά του είναι ότι δεν κάνει ευέλικτη αγκυροβόληση.

Το **SwissDock** (http://www.swissdock.ch/) είναι μια διαδικτυακή εφαρμογή που επιτρέπει με εύκολο και γρήγορο τρόπο την πρόβλεψη των αλληλεπιδράσεων μιας πρωτεΐνης με ένα μικρό μόριο. Το SwissDock βασίζεται στο λογισμικό EADock DSS, ο αλγόριθμος του οποίου περιλαμβάνει αρχικά την κατασκευή πολλών μοντέλων είτε σε μια εντοπισμένη περιοχή (local docking) ή γύρω από όλες τις πιθανές κοιλότητες του πρωτεϊνικού μορίου (blind docking). Παράλληλα, το CHARMM χρησιμοποιείται για τον υπολογισμό ενέργειας και στο τέλος τα μοντέλα με τις καλύτερες τιμές ενέργειας επιλέγονται και ομαδοποιούνται.

Το **rDock** (http://rdock.sourceforge.net/) είναι επίσης μια εφαρμογή για την αγκυροβόληση μικρών μορίων σε πρωτεΐνες εστιασμένη στην ταχύτητα και την ευελιξία. Είναι λογισμικό ανοιχτού κώδικα και είναι σχεδιασμένο ειδικά για τις λεγόμενες διαδικασίες High Throughput Virtual Screening (HTVS). Είναι επίσης ιδιαίτερα ελαφρύ σαν λογισμικό, και μπορεί να εγκατασταθεί σε όλα τα συστήματα Linux, ενώ με την ευέλικτη αρχιτεκτονική του μπορεί να εγκατασταθεί σε cluster και να χρησιμοποιήσει απεριόριστο αριθμό CPUs.

Τέλος, το **RosettaDock** (http://rosie.rosettacommons.org/docking2) είναι η παραλλαγή του γνωστού αλγόριθμου Rosetta, στην αγκυροβόληση. Βασίζεται σε μια μέθοδο προσομοίωσης Monte Carlo (MC) και εργάζεται σε δύο βήματα: στο πρώτο γίνεται μια ελαχιστοποίηση χαμηλής διακριτικότητας για τη διευθέτηση της κύριας αλυσίδας (ξεκινώντας είτε από τυχαίες θέσεις είτε από μια θέση επιλεγμένη από το χρήστη), ενώ στο δεύτερο βήμα γίνεται μια ελαχιστοποίηση ενέργειας για τη βελτιστοποίηση της διευθέτησης των πλευρικών ομάδων. Στα διαφορετικά στάδια, χρησιμοποιούνται επίσης και εμπειρικές συναρτήσεις ενέργειας με διαφορετικά χαρακτηριστικά.

Γενικά η αξιολόγηση των τόσων πολλών και διαφορετικών μεταξύ τους μεθόδων και λογισμικών είναι μια δύσκολη διαδικασία (Rodrigues & Bonvin, 2014; R. D. Taylor, et al., 2002). Κατ' αναλογία με τους διαγωνισμούς CASP και CAFASP για την πρόγνωση της δομής των πρωτεϊνών, υπάρχει, ειδικά για τον εντοπισμό των αλληλεπιδράσεων μεταξύ πρωτεϊνών, ο διαγωνισμός **CAPRI** (Critical Assessment of PRediction of Interactions). Το CAPRI είναι μια συνεχής διαδικασία κατά την οποία οι ερευνητές εφαρμόζουν τις μεθοδολογίες τους για αγκυροβόληση πρωτεϊνών στο ίδιο σύνολο δεδομένων, που αποτελείται από πρωτεΐνες των οποίων οι δομές έχουν πρόσφατα προσδιοριστεί πειραματικά, αλλά παραμένουν κρυφές με τη συναίνεση των ερευνητών που έκαναν τον προσδιορισμό. Το όλο πείραμα είναι διπλότυφλο, με την έννοια ότι ούτε οι επιστήμονες που κάνουν την πρόγνωση ξέρουν τη δομή, αλλά ούτε και οι αξιολογητές ξέρουν τον δημιουργό της κάθε πρόγνωσης (http://www.ebi.ac.uk/msd-srv/capri/).

Παρόλο που οι μέθοδοι αγκυροβόλησης απέχουν αρκετά από το να χαρακτηριστούν τέλειες, εξελίσσονται συνεχώς και αναμένεται στα επόμενα χρόνια να υπάρξει πρόοδος στον τομέα. Έτσι, αναμένεται πρόοδος στον τομέα της ενσωμάτωσης πειραματικής πληροφορίας, με σκοπό την παραγωγή όλο και πιο ρεαλιστικών μοντέλων. Στην περίπτωση της αγκυροβόλησης μικρών μορίων, η χρησιμότητα στην ανακάλυψη φαρμάκων είναι προφανής. Παρ' όλα αυτά, και η αγκυροβόληση πρωτεΐνης-πρωτεΐνης είναι ένας ιδιαίτερα σημαντικός τομέας, καθώς μπορεί να δώσει απαντήσεις σε πολλά προβλήματα που αφορούν τη σχέση δομής και λειτουργίας των πρωτεϊνών (τεταρτοταγής δομή, πρωτεϊνικές αλληλεπιδράσεις, μηχανισμοί δράσης ενζύμων κ.ο.κ.).

Βιβλιογραφία

Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., . . . Zwart, P. H. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr, 66*(Pt 2), 213-221.

Alvarez, J. C. (2004). High-throughput docking as a source of novel drug leads. *Current opinion in chemical biology, 8*(4), 365-370.

Andersen, C. A., Palmer, A. G., Brunak, S., & Rost, B. (2002). Continuum secondary structure captures protein flexibility. *Structure, 10*(2), 175-184.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., . . . Bordoli, L. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, gku340.

Bonvin, A. M. (2006). Flexible protein–protein docking. *Current Opinion in Structural Biology, 16*(2), 194-200.

Bowie, J. U., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science, 253*(5016), 164-170.

Brooks, B. R., Brooks, C. L., MacKerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., . . . Boresch, S. (2009). CHARMM: the biomolecular simulation program. *Journal of computational chemistry, 30*(10), 1545-1614.

Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., . . . Woods, R. J. (2005). The Amber biomolecular simulation programs. *Journal of computational chemistry, 26*(16), 1668-1688.

Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., . . . Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr, 66*(Pt 1), 12-21.

Dror, O., Benyamini, H., Nussinov, R., & Wolfson, H. J. (2003). Multiple structural alignment by secondary structures: algorithm and applications. *Protein Science, 12*(11), 2492-2507.

Eswar, N., Marti-Renom, M. A., Webb, B., Madhusudhan, M. S., Eramian, D., Shen, M., . . . Sali, A. (2006). Comparative Protein Structure Modeling With MODELLERCurrent Protocols in Bioinformatics (Vol. 5.6.1-5.6.30): John Wiley & Sons, Inc.

Frishman, D., & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins, 23*(4), 566-579.

Güntert, P. (2011). Automated protein structure determination from NMR data. In A. J. Dingley & S. M. Pascal (Eds.), *Biomolecular NMR spectroscopy* (pp. 341). Amsterdam: IOS Press.

Gray, J. J. (2006). High-resolution protein–protein docking. *Current Opinion in Structural Biology, 16*(2), 183-193.

Halperin, I., Ma, B., Wolfson, H., & Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics, 47*(4), 409-443.

Heinig, M., & Frishman, D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res, 32*(Web Server issue), W500-502.

Helles, G. (2008). A comparative study of the reported performance of ab initio protein structure prediction algorithms. *Journal of the Royal Society Interface, 5*(21), 387-396.

Holm, L., & Rosenström, P. (2010). Dali server: conservation mapping in 3D. *Nucleic Acids Research, 38*(suppl 2), W545-W549.

Hooft, R. W., Vriend, G., Sander, C., & Abola, E. E. (1996). Errors in protein structures. *Nature, 381*(6580), 272.

Jones, D. T. (1998). THREADER : Protein Sequence Threading by Double Dynamic Programming. In S. Salzberg, D. Searls & S. Kasif (Eds.), Computational Methods in Molecular Biology: Elsevier Science.

Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of molecular biology, 287*(4), 797-815.

Jones, D. T., Taylort, W. R., & Thornton, J. M. (1992). A new approach to protein fold recognition.

Joosten, R. P., Salzemann, J., Bloch, V., Stockinger, H., Berglund, A. C., Blanchet, C., . . . Vriend, G. (2009). PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *J Appl Crystallogr, 42*(Pt 3), 376-384.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers, 22*(12), 2577-2637.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protocols, 10*(6), 845-858.

Kolodny, R., Koehl, P., & Levitt, M. (2005). Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *Journal of molecular biology, 346*(4), 1173-1188.

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., & Lesk, A. M. (2006). MUSTANG: a multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics, 64*(3), 559-574.

Krissinel, E., & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography, 60*(12), 2256-2268.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK - a program to check the stereochemical quality of protein structures. *J. App. Cryst, 26*, 283-291.

Lathrop, R. H. (1994). The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein engineering, 7*(9), 1059-1068.

Liu, Y.-S., Fang, Y., & Ramani, K. (2009). Using least median of squares for structural superposition of flexible proteins. *BMC Bioinformatics, 10*(1), 29.

Maiti, R., Van Domselaar, G. H., Zhang, H., & Wishart, D. S. (2004). SuperPose: a simple server for sophisticated structural superposition. *Nucleic Acids Research, 32*(suppl 2), W590-W594.

Mayr, G., Domingues, F. S., & Lackner, P. (2007). Comparative analysis of protein structure alignments. *BMC Structural Biology, 7*(1), 50.

McLachlan, A. D. (1982). Rapid comparison of protein structures. *Acta Crystallogr D Biol Crystallogr, A38*, 871-873

Meiler, J., & Baker, D. (2003). Rapid protein fold determination using unassigned NMR data. *Proc Natl Acad Sci U S A, 100*(26), 15404-15409.

Moreira, I. S., Fernandes, P. A., & Ramos, M. J. (2010). Protein–protein docking dealing with the unknown. *Journal of computational chemistry, 31*(2), 317-342.

Ortiz, A. R., Strauss, C. E., & Olmea, O. (2002). MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Science, 11*(11), 2606-2621.

Pelton, J. T., & McLean, L. R. (2000). Spectroscopic methods for analysis of protein secondary structure. *Anal Biochem, 277*(2), 167-176.

Peng, J., & Xu, J. (2011). RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function, and Bioinformatics, 79*(S10), 161-171.

Poleksic, A. (2009). Algorithms for optimal protein structure alignment. *Bioinformatics, 25*(21), 2751-2756.

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., . . . van der Spoel, D. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, btt055.

Read, R. J., Adams, P. D., Arendall, W. B., 3rd, Brunger, A. T., Emsley, P., Joosten, R. P., . . . Zwart, P. H. (2011). A new generation of crystallographic validation tools for the protein data bank. *Structure, 19*(10), 1395-1412.

Rodrigues, J. P., & Bonvin, A. M. (2014). Integrative computational modeling of protein interactions. *FEBS Journal, 281*(8), 1988-2003.

Rohl, C. A., Strauss, C. E., Misura, K. M., & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods in enzymology, 383*, 66-93.

Rost, B., Schneider, R., & Sander, C. (1997). Protein fold recognition by prediction-based threading. *Journal of molecular biology, 270*(3), 471-480.

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols, 5*(4), 725-738.

Šali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology, 234*(3), 779-815.

Schwieters, C. D., Kuszewski, J. J., & Clore, G. M. (2006). Using Xplor-NIH for NMR molecular structure determination. *Progr. NMR Spectroscopy 48*, 47-62

Shi, Y. (2014). A glimpse of structural biology through X-ray crystallography. *Cell, 159*(5), 995-1014.

Shindyalov, I. N., & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein engineering, 11*(9), 739-747.

Singh, A. P., & Brutlag, D. L. (2000). Protein Structure Alignment: A comparison of methods. *Bioinformatics*.

Söding, J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics, 21*(7), 951-960.

Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research, 33*(suppl 2), W244-W248.

Sternberg, M. J., Gabb, H. A., & Jackson, R. M. (1998). Predictive docking of protein—protein and protein—DNA complexes. *Current Opinion in Structural Biology, 8*(2), 250-256.

Sumathi, K., Ananthalakshmi, P., Roshan, M. M., & Sekar, K. (2006). 3dSS: 3D structural superposition. *Nucleic Acids Research, 34*(suppl 2), W128-W132.

Taylor, R. D., Jewsbury, P. J., & Essex, J. W. (2002). A review of protein-small molecule docking methods. *Journal of computer-aided molecular design, 16*(3), 151-166.

Taylor, W. R., & Orengo, C. A. (1989). Protein structure alignment. *Journal of molecular biology, 208*(1), 1-22.

Theobald, D. L., & Wuttke, D. S. (2006a). Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem. *Proceedings of the National Academy of Sciences, 103*(49), 18521-18527.

Theobald, D. L., & Wuttke, D. S. (2006b). THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics, 22*(17), 2171-2172.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *Journal of molecular graphics, 8*(1), 52-56.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., . . . Wilson, K. S. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr, 67*(Pt 4), 235-242.

Wu, S., & Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Research, 35*(10), 3375-3382.

Wu, S., & Zhang, Y. (2008). MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information. *Proteins: Structure, Function, and Bioinformatics, 72*(2), 547-556.

Yaffe, M. B. (2005). X-ray crystallography and structural biology. *Crit Care Med, 33*(12 Suppl), S435-440.

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research, 33*(7), 2302-2309.