Ειδικά Θέματα Βιοπληροφορικής

Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

Πανεπιστήμιο Θεσσαλίας
Λαμία, 2015

Δομική Βιοπληροφορική

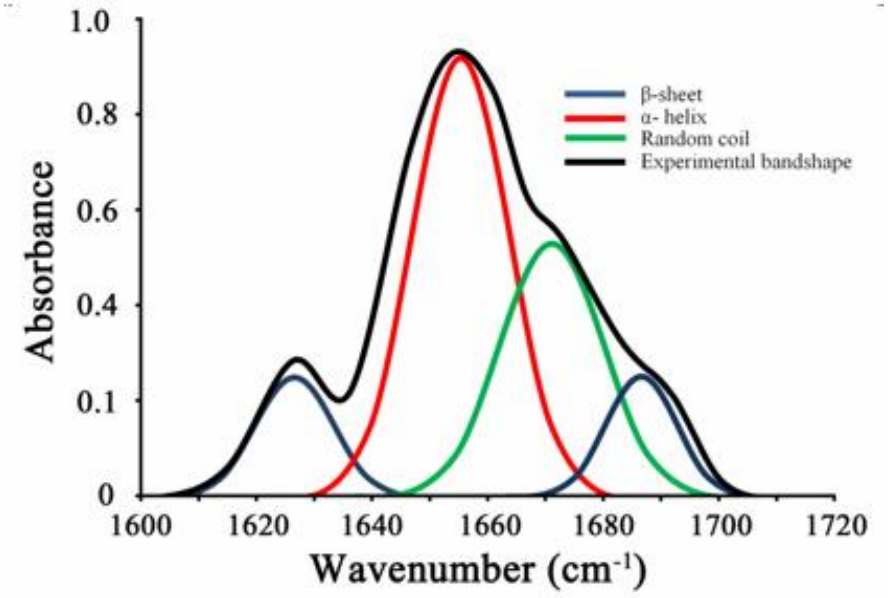
- Το αντικείμενο της μελέτης της δομικής βιοπληροφορικής, είναι οι τρισδιάστατες δομές, δηλαδή οι συντεταγμένες των ατόμων ενός βιολογικού μακρομορίου. Η εύρεση της τρισδιάστατης δομής, είναι από μόνη της μια ιδιαίτερα επίπονη και κοστοβόρα διαδικασία, που αποτελεί τον περιοριστικό παράγοντα στον τομέα. 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.

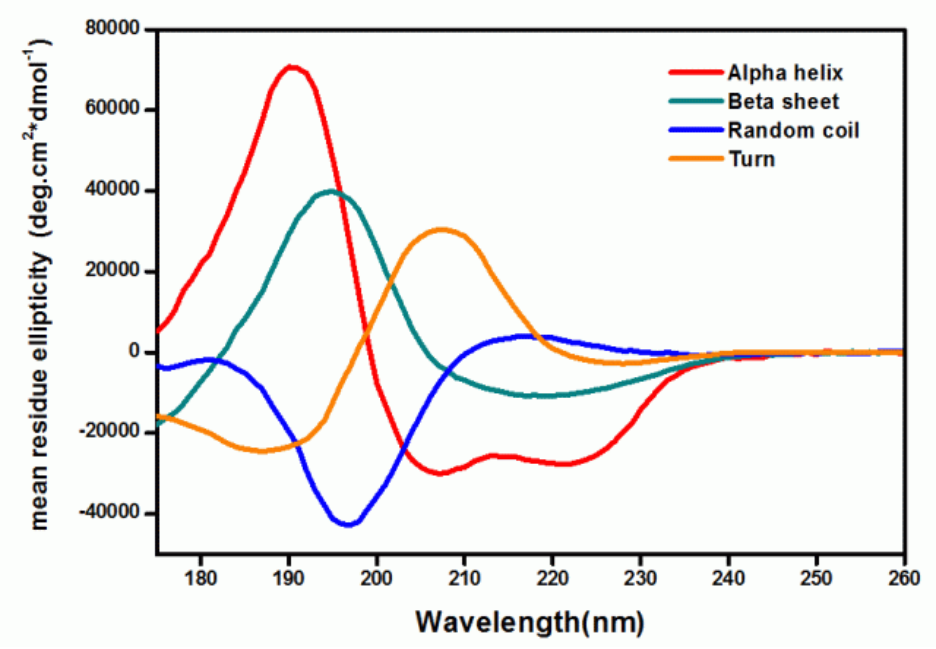
Δομική Βιοπληροφορική

- Προσδιορισμός δομής
- Οπτικοποίηση δομής
- Υπέρθεση και στοίχιση δομών
- Πρόβλεψη δομής
  - Προτυποποίηση με βάση την ομολογία
  - Ύφανση και αναγνώριση διπλώματος
  - Ab initio πρόβλεψη
- Αγκυροβόληση

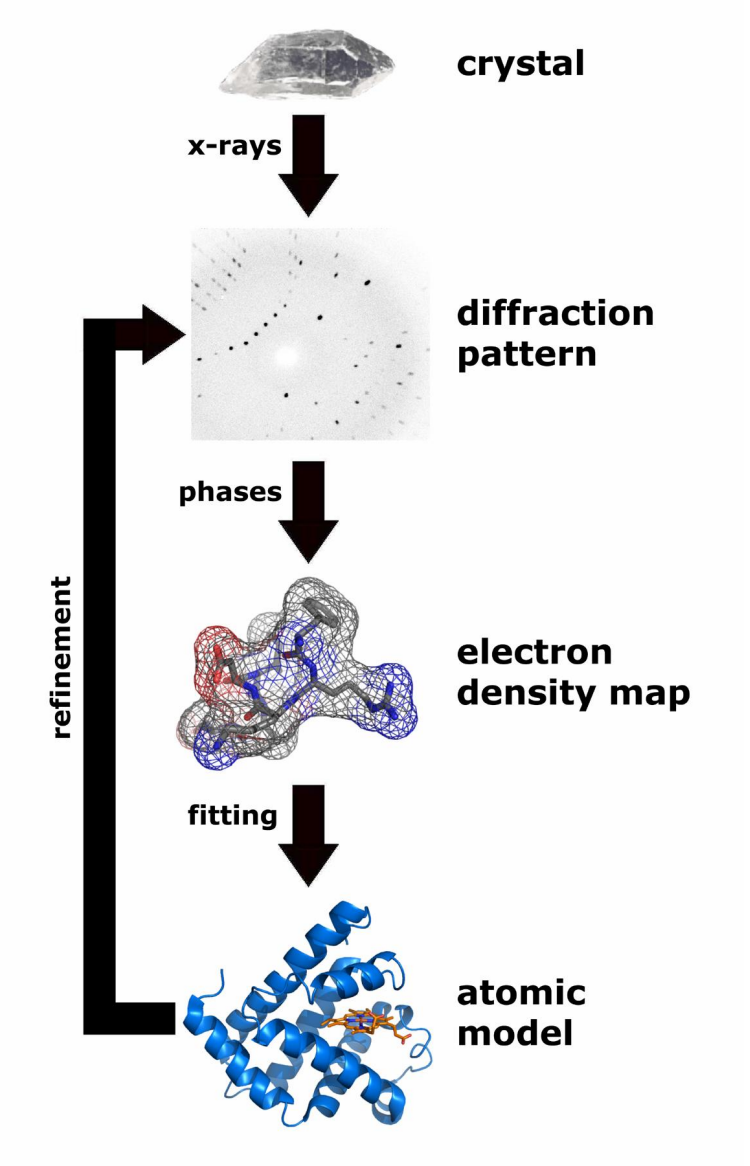
Προσδιορισμός δομής

- IR - CD





Κρυσταλλογραφία ακτίνων Χ



Λογισμικό

- 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 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 παραδοσικά 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)

Λογισμικό ελέγχου

- 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 προγραμμα σε αυτον τον τομεα ηταν το **PROCHECK** (Laskowski, MacArthur, Moss, & Thornton, 1993). Τα προγράμματα που χρησιμοποιουνται πλεον ευρεως για αξιολογηση και επαληθευση 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 επαληθευση των δομων ειναι πλεον απαραιτητη για την δημοσιευση τους και υπαρχουν συγκεριμενες οδηγιες για αυτο τον σκοπο (Read et al., 2011). Μια νεα αντιμετωπιση, ειναι και η λογικη της «ενεργους επαληθευσης» οπου οι υπαρχουσες δομες διορθωνονται με αυτοματοποιημενους αλγοριθμους μοντελοποιησης με βαση τα κρυσταλλογραφικα δεδομενα που κατατίθενται στην PDB (Joosten et al., 2009).

Λογισμικό καθορισμού δομής

- 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 προέκυπταν 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, αλλά και άλλα χαρακτηριστικά όπως την προσβασιμότητα των διαφόρων καταλοίπων (δηλαδή, αν είναι εκτεθιμένα 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.

DSSP

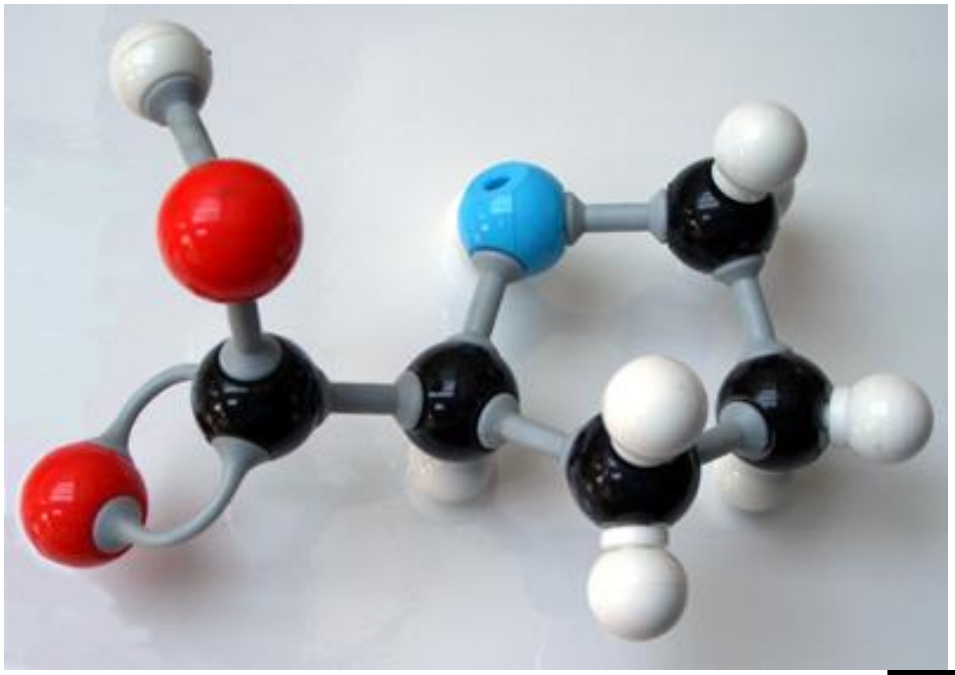
- Το **DSSP** (Define Secondary Structure of Proteins), διαθέσιμο στη διεύθυνση http://swift.cmbi.ru.nl/gv/dssp/ ήταν ο πρώτος αλγόριθμος που προτάθηκε για το σκοπό αυτό, και είναι ακόμα ο ευρύτερα χρησιμοποιούμενος (Kabsch & Sander, 1983). Το DSSP αναγνωρίζει τον κύριο ανθρακικό σκελετό της πρωτεϊνης και εντοπίζει τους δεσμούς υδρογόνου που σχηματίζονται, με βάση έναν καθαρά ηλεκτροστατικό ορισμό. Με βάση τον ενεργειακό υπολογισμό, το DSSP αναγνωρίζει και κατάτασσει τα στοιχεία δευτεροταγούς δομής σε 8 κατηγορίες. Η 310-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 πχ 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).

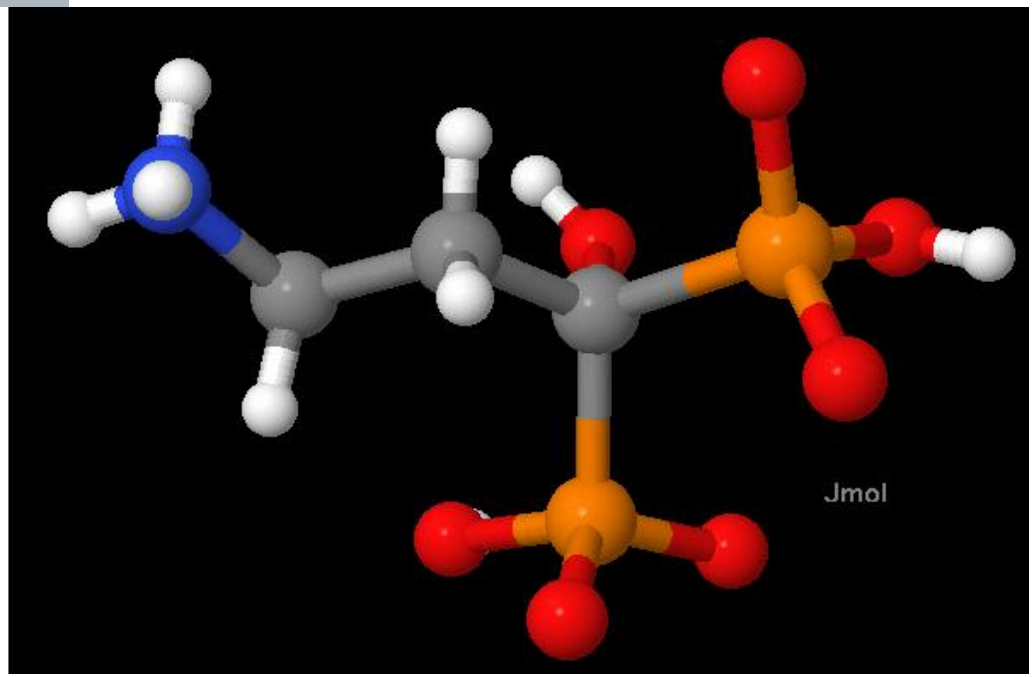
STRIDE

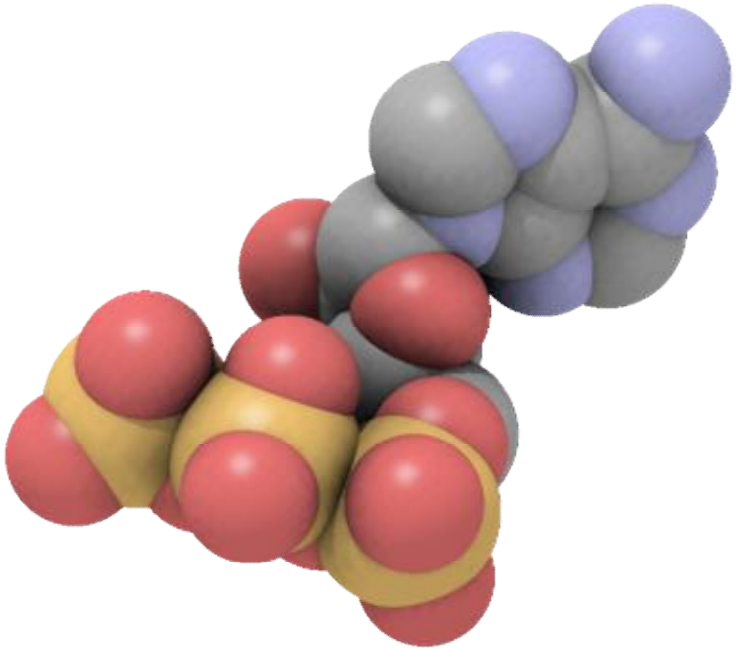
- Το **STRIDE** (STRuctural IDEntification), το οποίο είναι διαθέσιμο στη διεύθυνση http://webclu.bio.wzw.tum.de/stride/ είναι ένας άλλος εναλλακτικός αλγόριθμος για τον προσδιορισμό και την ανάθεση των στοιχείων δευτεροταγούς δομής (Frishman & Argos, 1995). Το STRIDE χρησιμοποιεί μια παρόμοια μέθοδο με το DSSP, καθώς εμαρμόζει 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, οι 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).

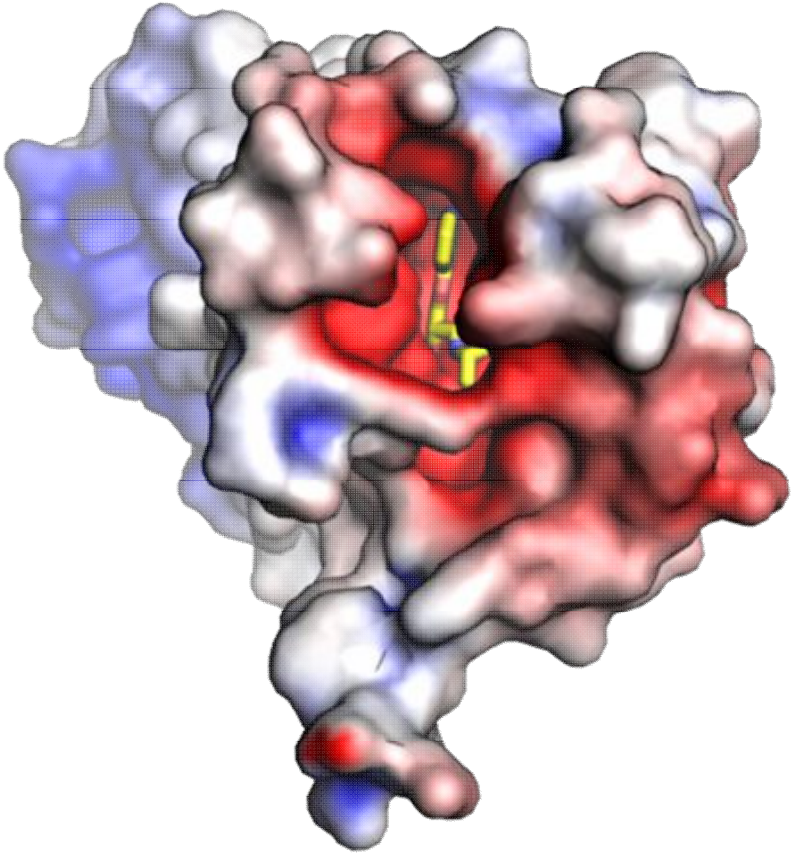
Οπτικοποίηση δομών

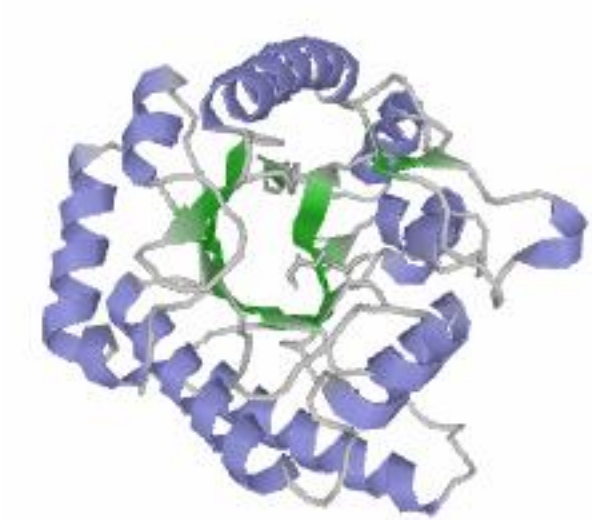
- Ball and stick
- Wireframe
- Space-filling

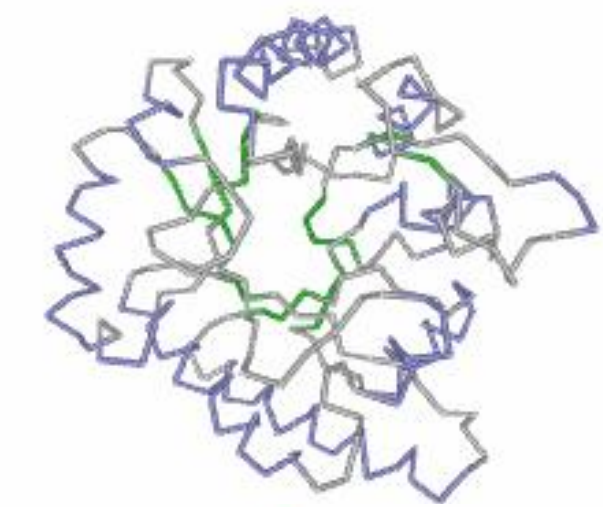


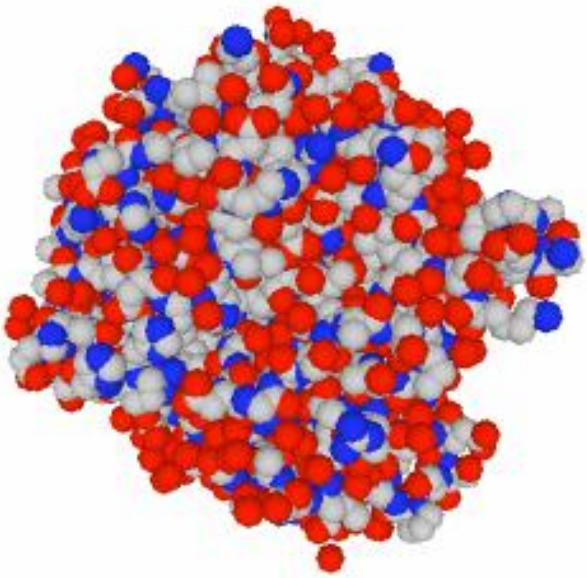

Jmol

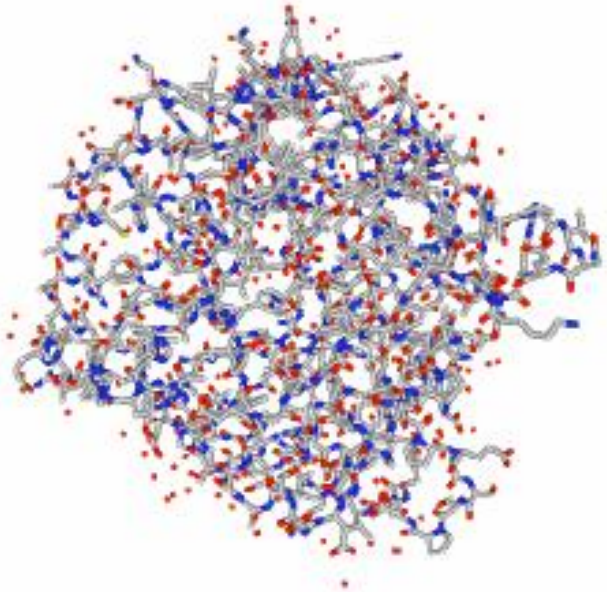












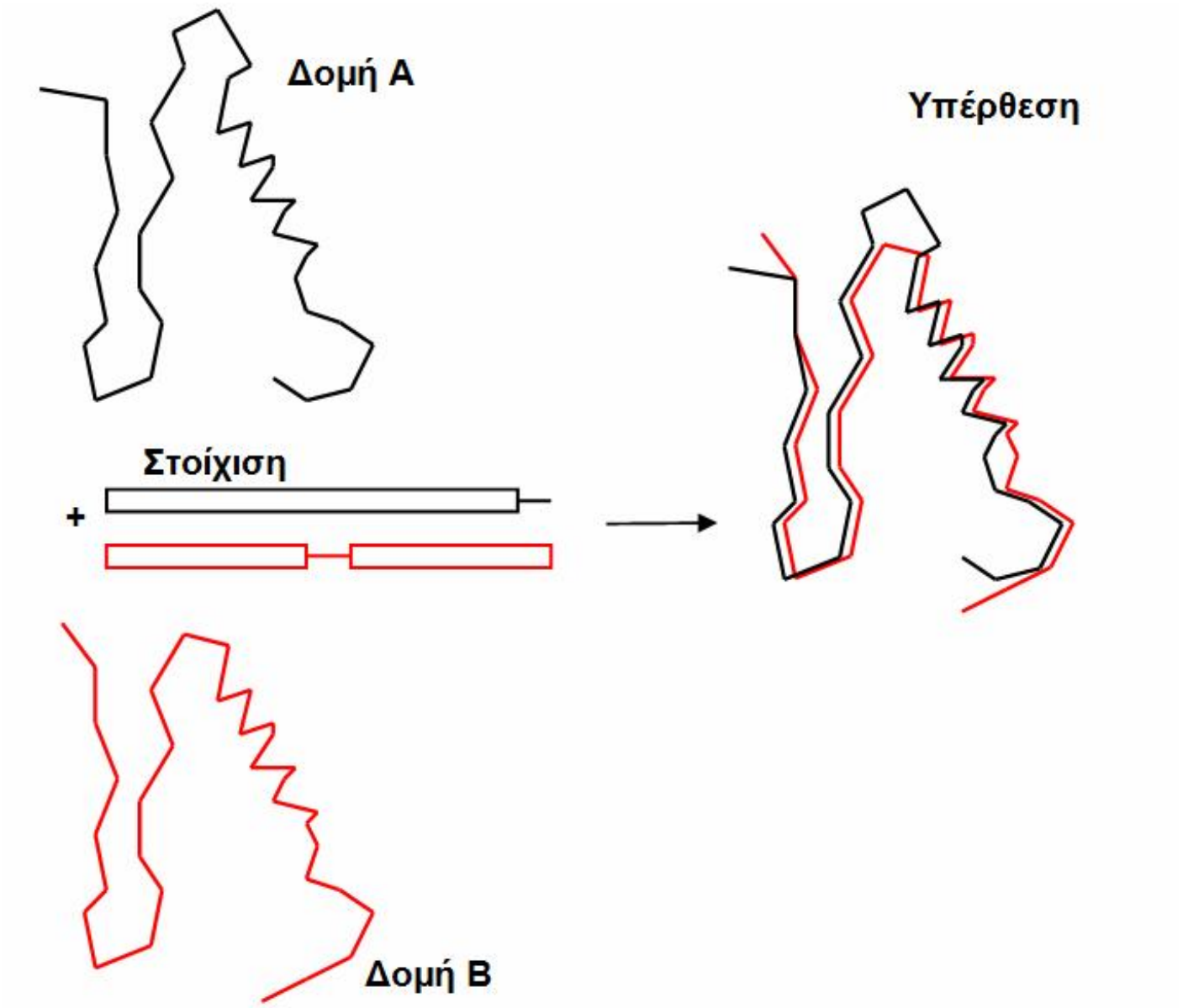
Λογισμικό

- 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 Java Applet και το **Jsmol** που είναι η έκδοση του που 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),
- 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 ξεχάσουμε να κάνουμε αναφορά στο πιο πετυχημένα 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) 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)

Υπέρθεση δομών – Στοίχιση δομών

- Η αναζήτηση ομοιότητας δύο ή περισσότερων τρισδιάστατων δομών.
- Σαν μεθοδολογίες επιφανειακά μοιάζουν αλλά στη βάση τους διαφέρουν αρκετά. Η βασική διαφορά είναι ότι στην υπέρθεση ξέρουμε εκ των προτέρων την στοίχιση των αλληλουχιών, ενώ στη δομική στοίχιση η ταύτιση επιτυγχάνεται μόνο με χρήση της πληροφορίας της δομής γιατί η στοίχιση αλληλουχιών δεν είναι εφικτή (μια στοίχιση των αλληλουχιών όμως προκύπτει από τη στοίχιση δομών).


Δομή Α
Στοίχιση
+
Υπέρθεση
Δομή Β

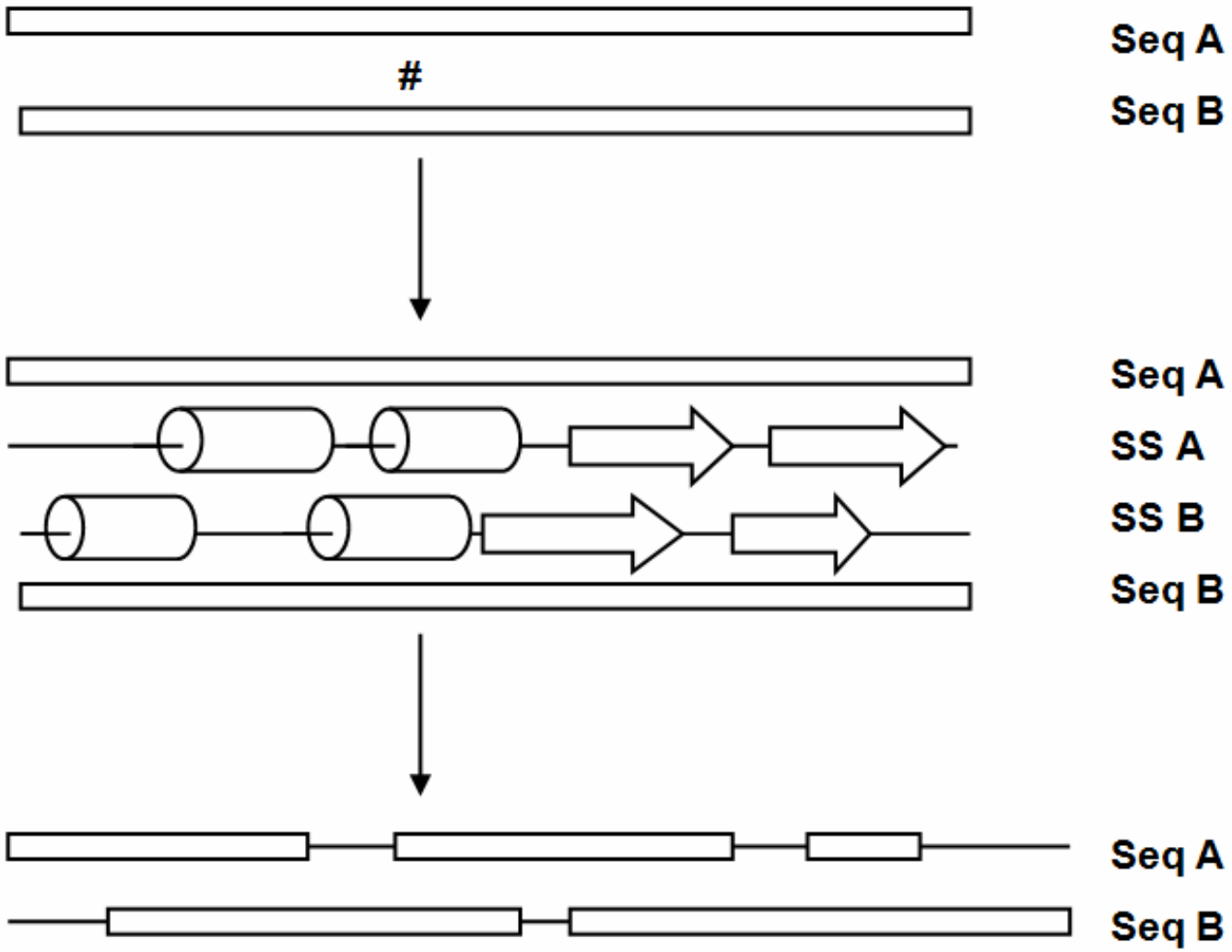
Λογισμικό

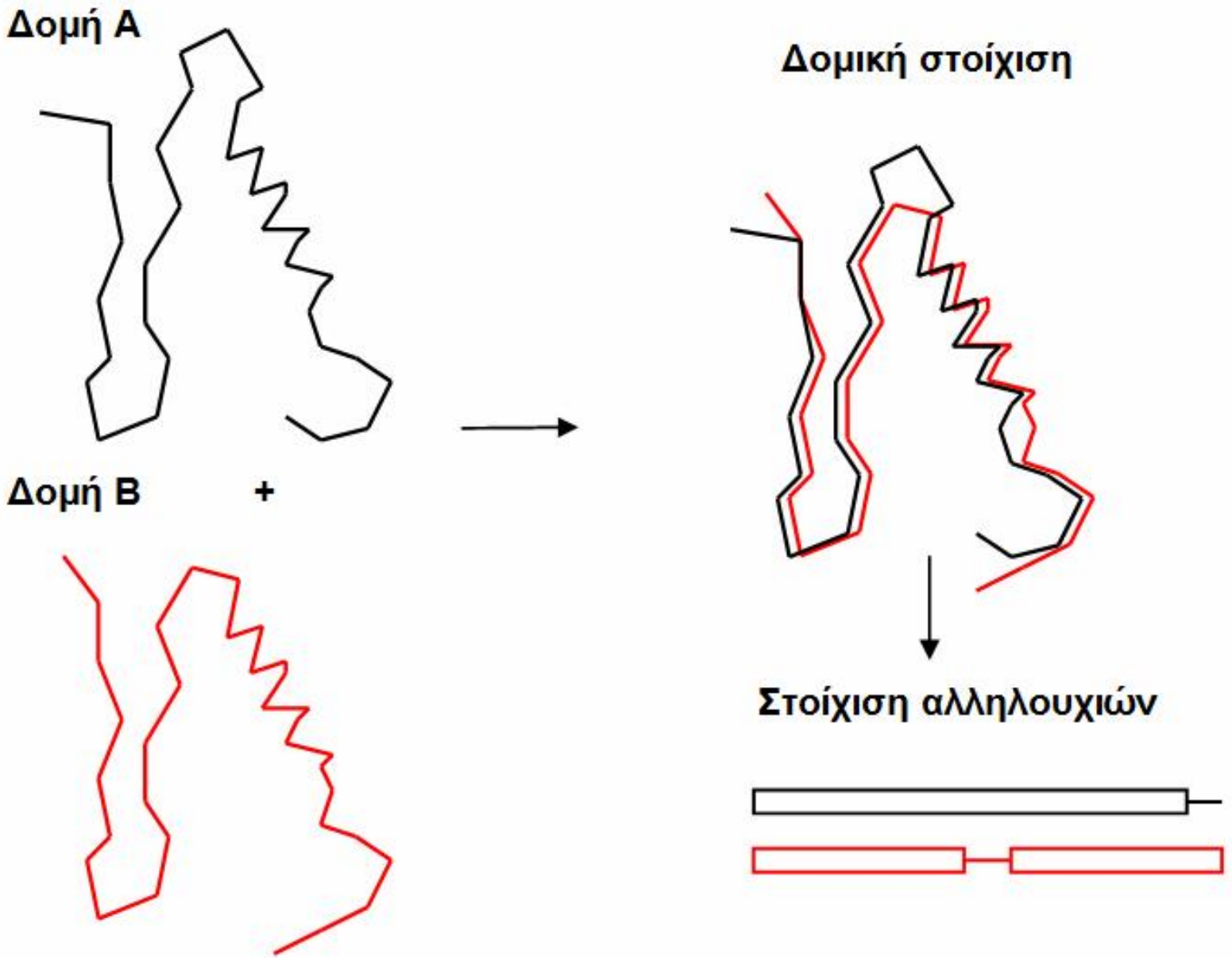
- Στις περισσότερες των περιπτώσεων, μόνο τα άτομα της κύριας ανθρακικής αλυσίδας (Cα) χρησιμοποιούνται για τις συγκρίσεις αυτές καθώς αυτά είναι που θα καθορίσουν το γενικότερο σχήμα και τη δομή της πρωτεΐνης και οι υπολογισμοί είναι ευκολότεροι. Επιπλέον δε, η σύγκριση των ατόμων των πλευρικών αλυσίδων 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.
- Γενικά,
  - 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).

- 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 αποτελούμενες 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 χρησιμοποιόντας 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 τον χρήση (Maiti, Van Domselaar, Zhang, & Wishart, 2004). Το πρόγραμμα λειτουργεί αυτόματα. Έτσι, όταν οι αλληλουχίες διαφέρουν αλλά μπορούν να στοιχιθούν με κάποιον κλασικό αλγόριθμο ομοιότητας, λειτουργεί με τον κλασικό τρόπο που περιγράψαμε παραπάνω. Όταν όμως οι αλληλουχίες των πρωτεϊνών διαφέρουν πέραν από τα όρια 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.


Seq A
#
Seq B
Seq A
SS A
SS B
Seq B
Seq A
Seq B


Δομή Α
Δομή Β
+
Δομική στοίχιση
Στοίχιση αλληλουχιών

Λογισμικό

- 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 ταύτιση 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.

συνέχεια

- 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).

συνέχεια

- 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 αποστάσεων 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 (W. R. 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.

συνέχεια

- 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.

συνέχεια

- 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 το ότι 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 πρωτεϊνες λόγω συγκλίνουσας εξέλιξης αλλά δεν έχουν ομοιότητα στο δίπλωμα και 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)

συνέχεια

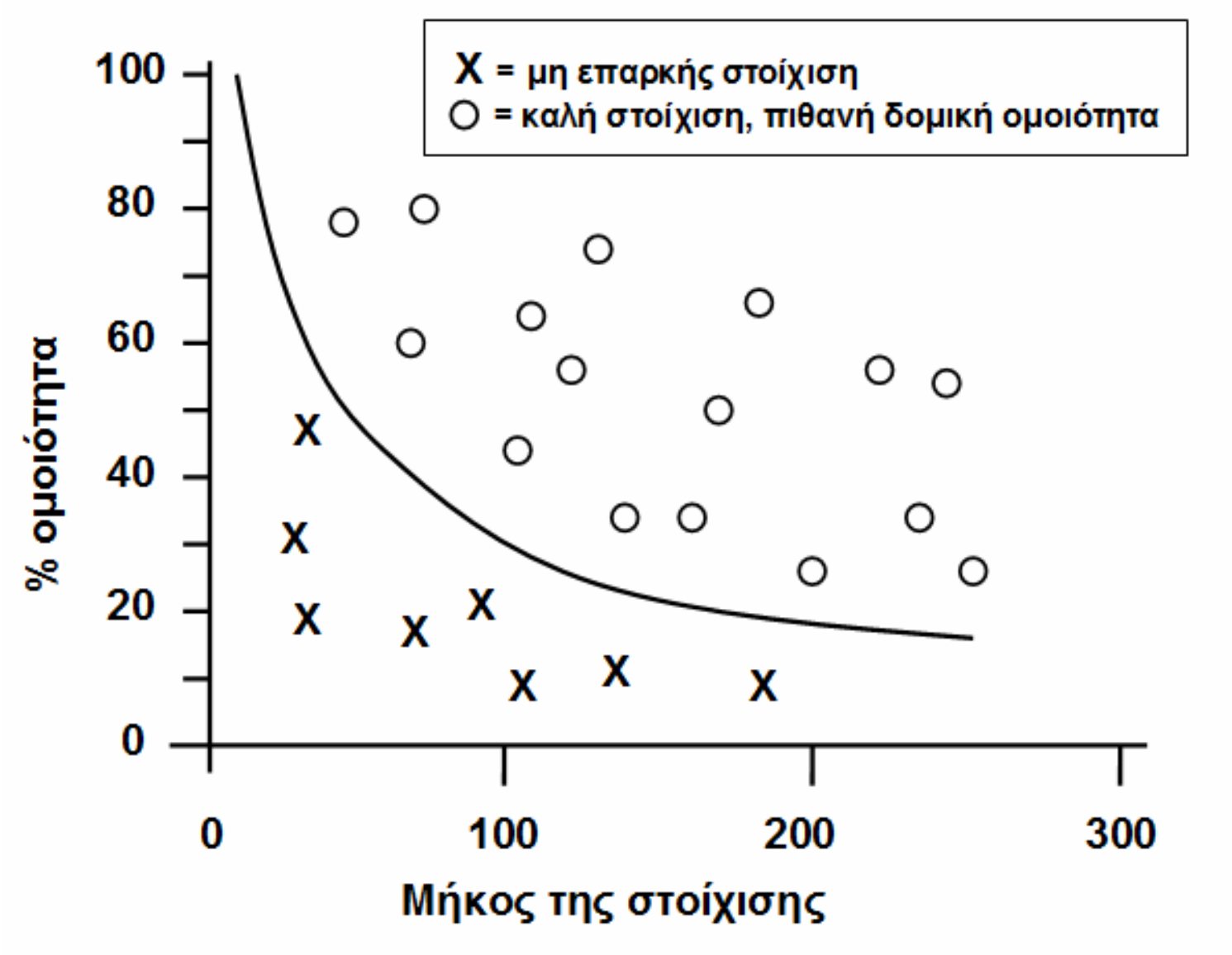
- Το **MUSTANG** (multiple structural alignment algorithm) είναι μια εφαρμογή που σχεδιάστηκε εξαρχής για πολλαπλή δομική στοίχιση (Konagurthu, Whisstock, Stuckey, & Lesk, 2006). Βασίζεται σε ιεραρχική πολλαπλή στοίχιση, με πολλαπλά βήματα που επανυπολογίζονται για βελτιστοποίηση. Στην αρχή υπολογίζει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

Σύγκριση των μεθόδων

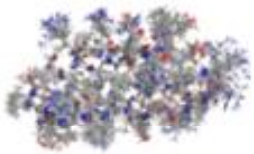
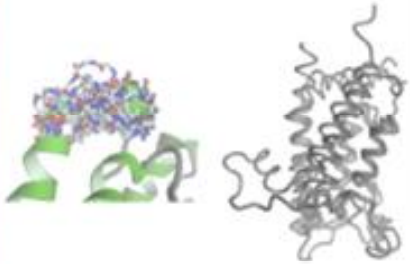
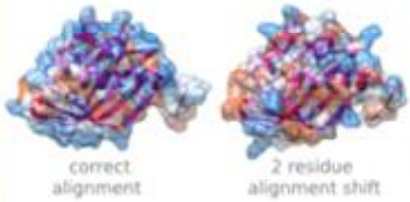
- Όπως αναφέραμε παραπάνω, τα προγράμματα αυτά είναι μόνο ένα μικρό μέρος των προγραμμάτων που είναι διαθέσιμα στην επιστημονική κοινότητα. Στην αντίστοιχη σελίδα της Wikipedia (https://en.wikipedia.org/wiki/Structural_alignment_software) αναφέρονται δεκάδες αντίστοιχα εργαλεία, παρόλα αυτά, κάναμε αναφορά 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, και όταν οι πρωτεϊνες 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. Παρόλα αυτά, υπάρχουν ειδικές περιπτώσεις στις οποίες κάποιο άλλο εργαλείο μπορεί να ενδείκνυται καλύτερα, γιαυτό 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.

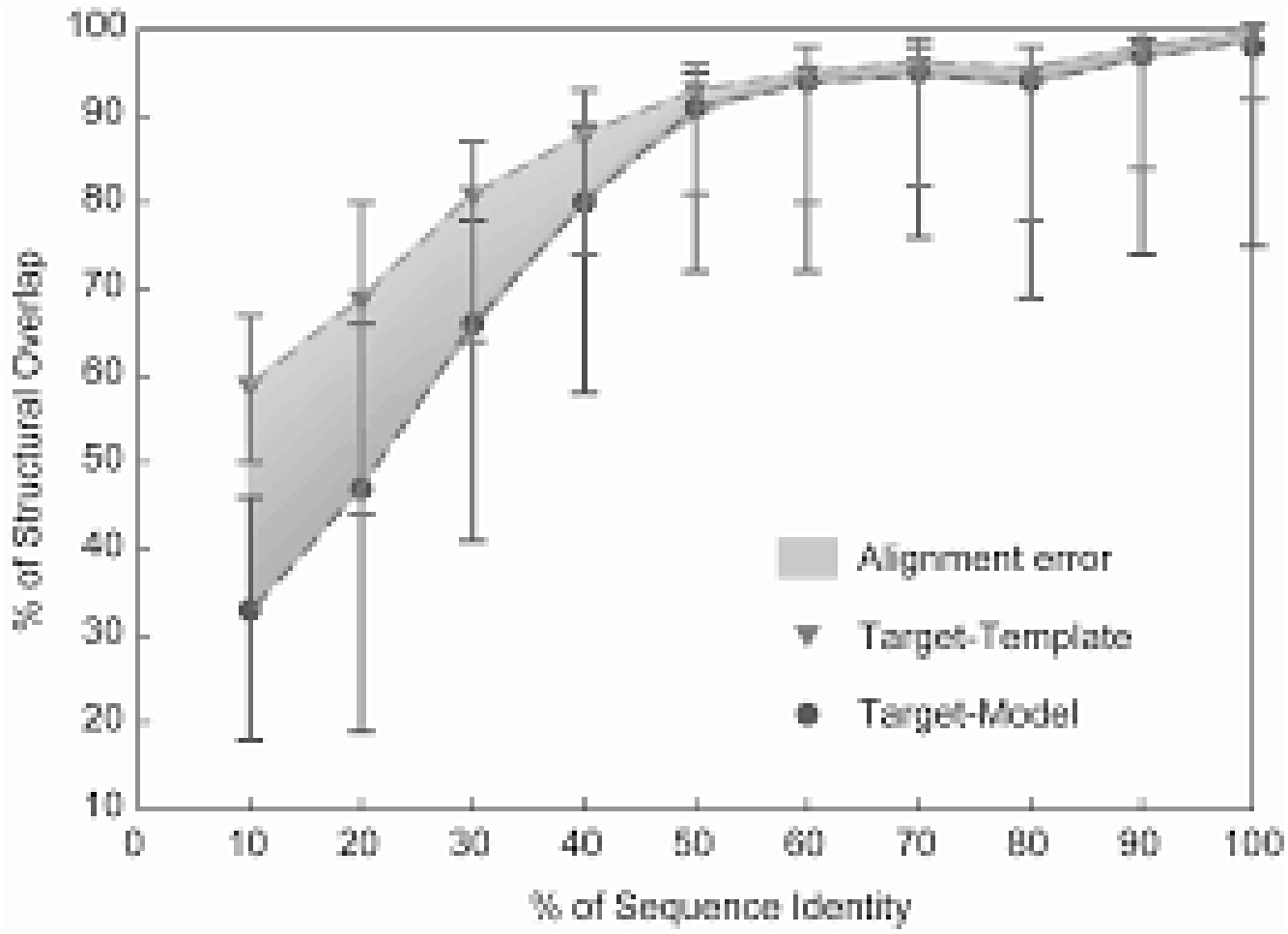
Πρόγνωση τρισδιάστατης δομής πρωτεϊνών

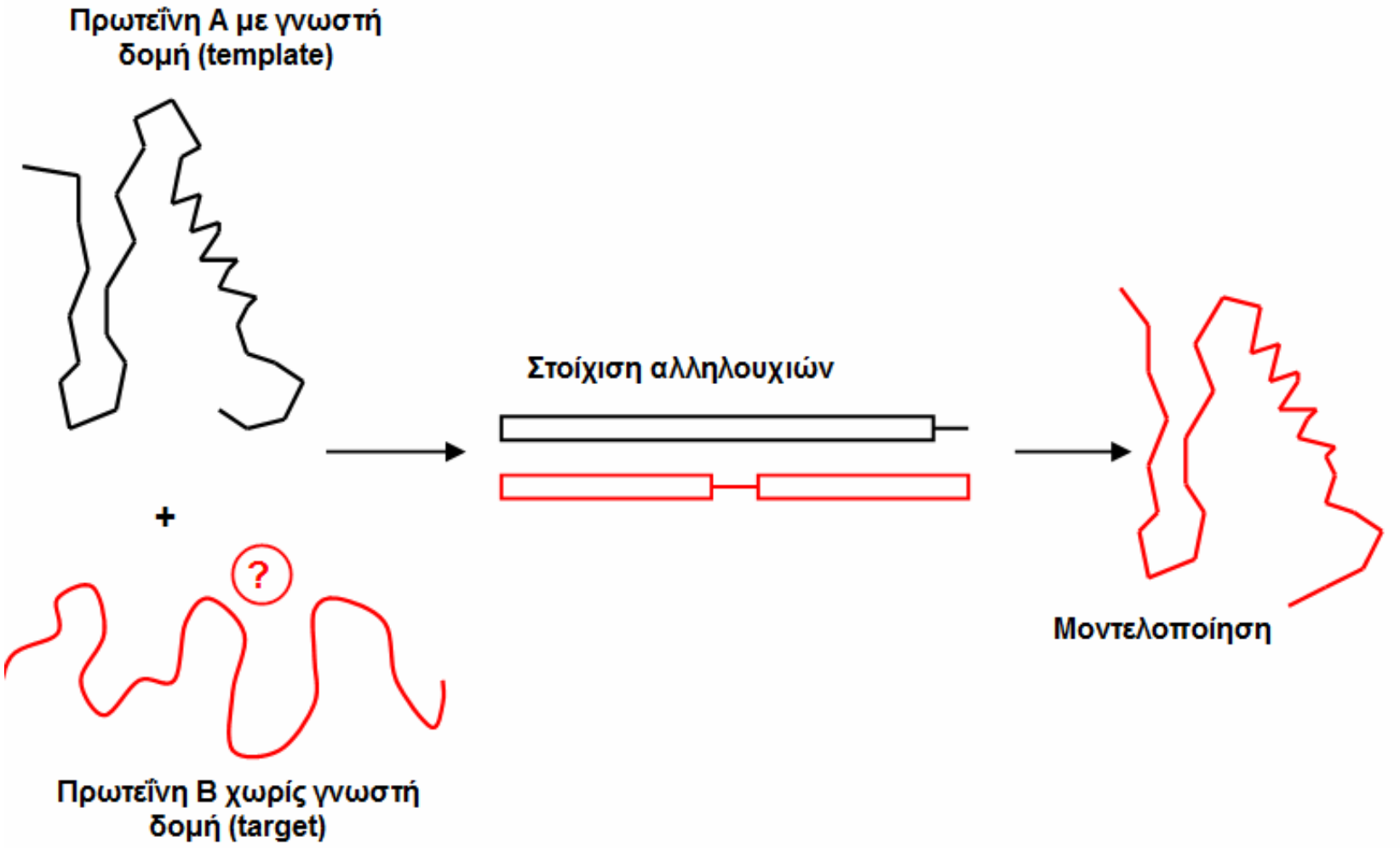
- Ο τελικός σκοπός της υπολογιστικής μελέτης και της μοντελοποίησης των πρωτεϊνών, είναι η πρόγνωση της τρισδιάστατης δομής μιας πρωτεϊνης από την αλληλουχία της. Σε προηγούμενα κεφάλαια, είδαμε την πρόγνωση της δευτεροταγούς δομής, η οποία είναι μόνο ένα σχετικά πιο εύκολο υποκατάστατο για την τελική πρόγνωση της τρισδιάστατης δομής.
- Μια τέτοια πρόβλεψη θα επιτρέψει τη διενέργεια πολλών πειραμάτων in silico (σχεδιασμός φαρμάκων, μελέτη της λειτουργίας της πρωτεϊνης, 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 τις περιπτώσεις αυτές, οι υπολογιστικές μελέτες είναι η μόνη εναλλακτική.
- Οι βασικές αρχές πίσω από τις μελέτες μοντελοποίησης είναι δύο, και είναι γνωστές από χρόνια: α) η αλληλουχία μιας πρωτεϊνης καθορίζει μονοσήμαντα τη δομή μιας πρωτεϊνης, και β) οι πρωτεϊνικές δομές συντηρούνται περισσότερο από τις αλληλουχίες. Μια άμεση συνέπεια των παραπάνω, είναι ότι δυο πρωτεϊνες με μεγάλη ομοιότητα σε επίπεδο αλληλουχίας έχουν κατά βάση παρόμοια δομή, αλλά είναι δυνατό, παρόμοια δομή να έχουν και πρωτεϊνες με μη ανιχνεύσιμη ομοιότητα (στο τελευταίο, σημαντικό ρόλο παίζει και η ύπαρξη περιορισμένου αριθμού πρωτεϊνικών διπλωμάτων).


X = μη επαρκής στοίχιση
○ = καλή στοίχιση, πιθανή δομική ομοιότητα
% ομοιότητα
100
80
60
40
20
0
0
100
200
300
Μήκος 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 ναι μεν δεν διαθέτουν ομόλογη με γνωστή δομή, αλλά είναι δυνατόν να εντοπιστεί με κάποια μέθοδο **αναγνώρισης διπλώματος**, το πρωτεϊνικό δίπλωμα στο οποίο ταιριάζει η συγκεκριμένη αλληλουχία (γιαυτό 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.

Sources of errors			Applications
- experimental errors and uncertainties in X-ray, NMR	1Å 100%		- studying catalytic mechanism / function
- side-chain packing - mis-placed side-chains	1.5Å 95%		- structure-based drug design, ligand docking
- modeling of loop regions (insertions and deletions) - distorsions of aligned regions	60%		- structural support for mutagenesis studies - molecular replacement
- alignment errors	3Å 40%		- integrative modeling - modeling into low-resolution density maps
- sub-optimal template selection - model may even have the wrong fold	>3Å <30%		- domain boundaries - identification of structural motives


100
90
80
70
60
50
40
30
20
10
% of Structural Overlap
0
10
20
30
40
50
60
70
80
90
100
% of Sequence Identity
Alignment error
Target-Template
Target-Model


Πρωτεΐνη Α με γνωστή δομή (template)
+
?
Πρωτεΐνη Β χωρίς γνωστή δομή (target)
Στοίχιση αλληλουχιών
Μοντελοποίηση

προτυποποίηση με βάση την ομολογία

- *Εύρεση του πρότυπου και πραγματοποίηση της στοίχισης*.
- *Κατασκευή του σκελετού της κύριας ανθρακικής αλυσίδας*.
- *Μοντελοποίηση των βρόχων και των πλευρικών αλυσίδων*.
- *Βελτιστοποίηση του μοντέλου*.
- *Έλεγχος ποιότητας του μοντέλου.*

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 (χρησιμοποιόντας 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 χρησιμοποιόντας πολλαπλα 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 και 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 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 επαναλληπτικά 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

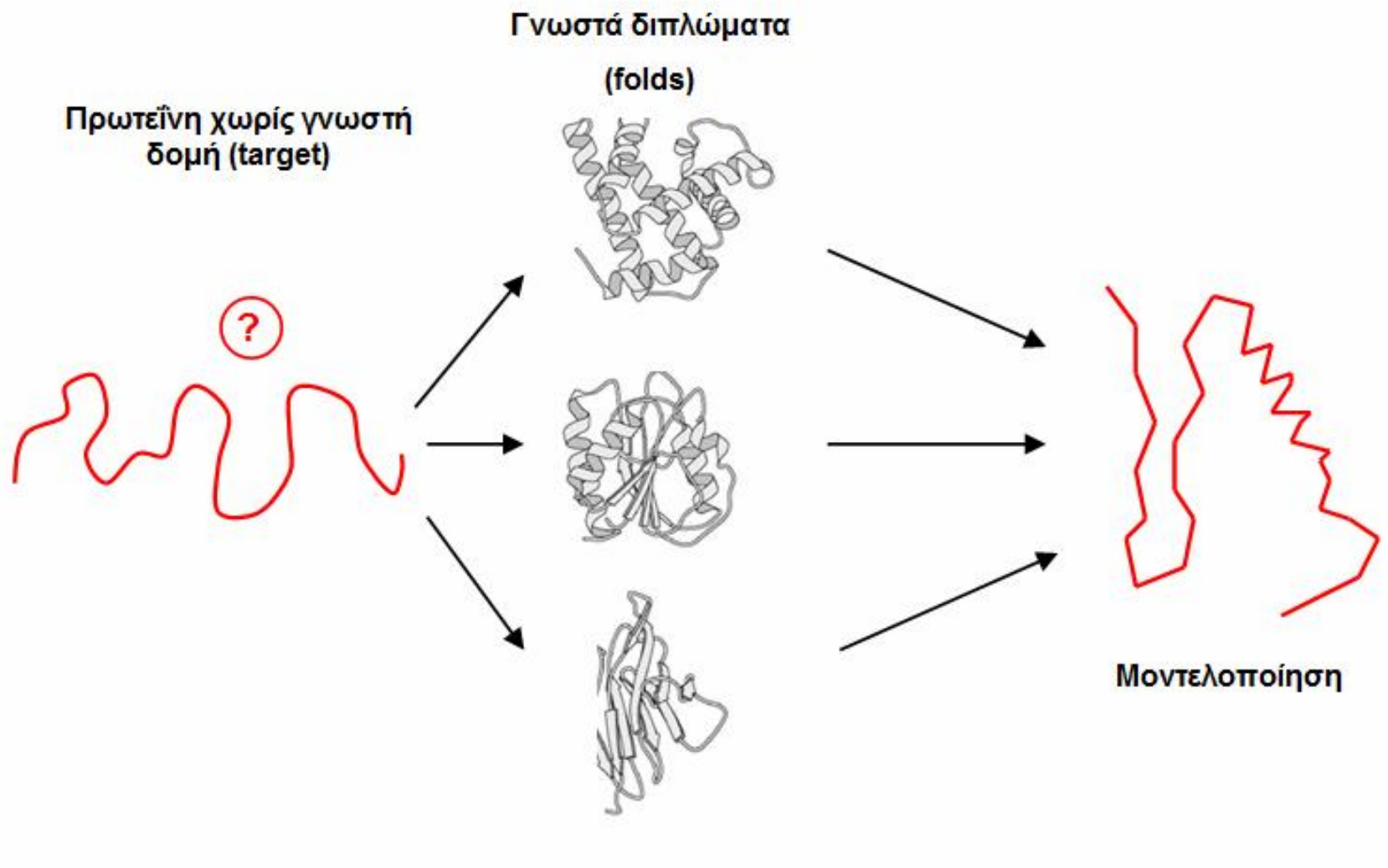
- Αφού το μοντέλο έχει κατασκευαστεί, είναι απαραίτητος ο έλεγχος για την επιβεβαίωσή του. Αυτός μπορεί να γίνει βασικά με δυο τρόπους, είτε με χρήση μοριακής δυναμικής με υπολογισμό της συνολικής ενέργειας το μορίου είτε με εμπειρικές μεθόδους που μετράνε την κανονικότητα διάφορων χαρακτηριστικών (μήκη 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.

Λογισμικό

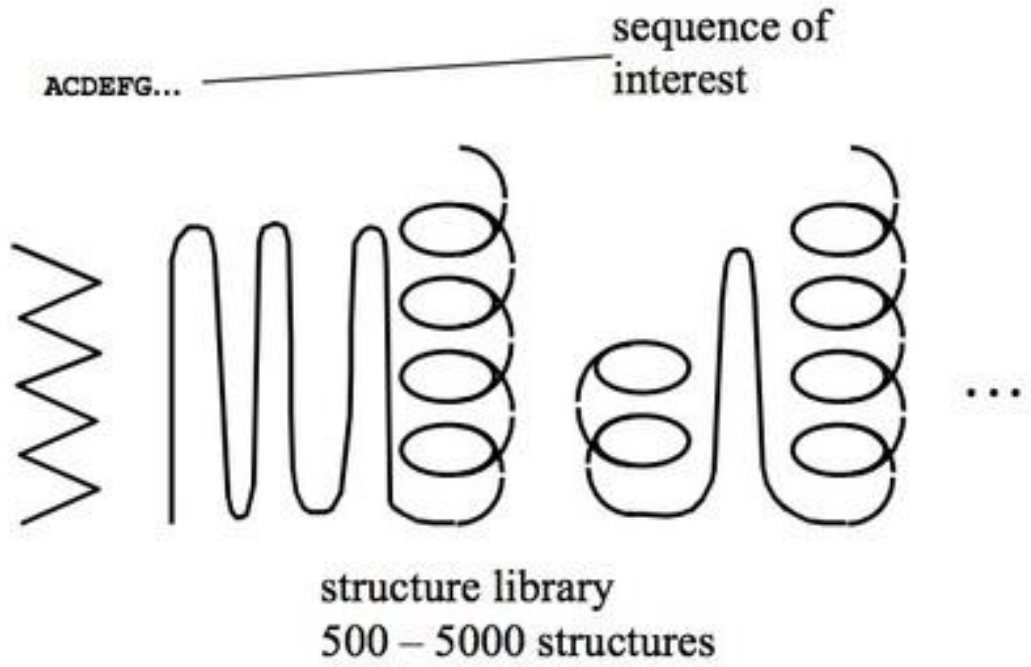
- Το πιο γνωστό αλλά και το πιο παλιό λογισμικό για μοντελοποίηση με βάση την ομολογία, είναι το **WHAT IF** (http://swift.cmbi.ru.nl/whatif/), το οποίο παρουσιάστηκε πρώτη φορά το 1987 από τον Gert Vriend (Vriend, 1990). Από τότε, συνεχίζει να εξελίσσεται και αποτελεί πλέον ένα ολοκληρωμένο περιβάλλον 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 Η/Υ (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, παρέχωντας πλήθος λειτουργιών όπως αυτόματη αναζήτηση στις βάσεις δεδομένων, έλεγχος ποιότητας για την επιλογή του καλύτερου πρότυπου, μοντελοποίηση με πολλαπλά πρότυπα και έλεγχο ποιότητας του προκύπτοντος μοντέλου (Biasini et al., 2014).

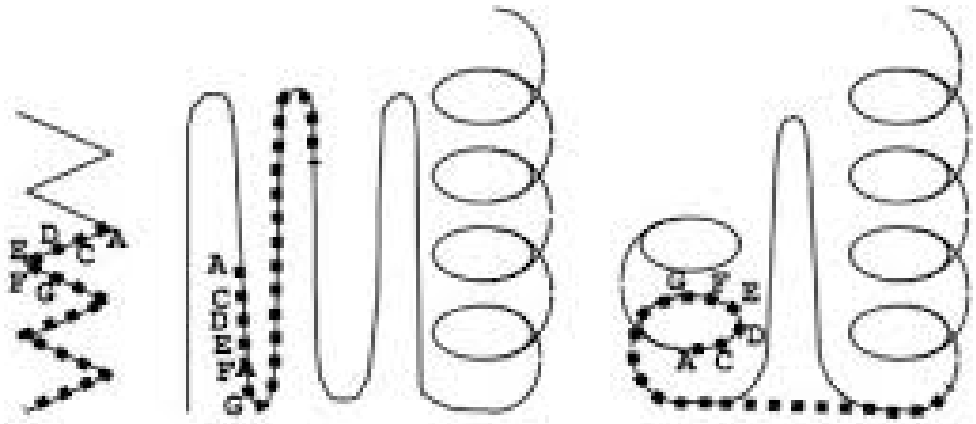
Αναγνώριση διπλώματος και ύφανση

- Όπως είπαμε ήδη, η ύφανση ή αλλιώς αναγνώριση διπλώματος είναι μια τεχνική που χρησιμοποιείται σε περιπτώσεις κατά τις οποίες η πρωτεϊνη στόχος δεν έχει ανιχνεύσιμη ομοιότητα σε επίπεδο αλληλουχίας με κάποια πρωτεϊνη 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 πρωτεϊνες 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 πρωτεϊνη με μη ανιχνεύσιμη ομοιότητα σε επίπεδο αλληλουχίας, είναι παρόλα 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 όντως κάποιες αλγοριθμικές τεχνικές έχουν χρησιμοποιηθεί και στις δύο μεθοδολογίες (πχ είδαμε παραπάνω τη στοίχιση με τη βοήθεια της δευτεροταγούς δομής).

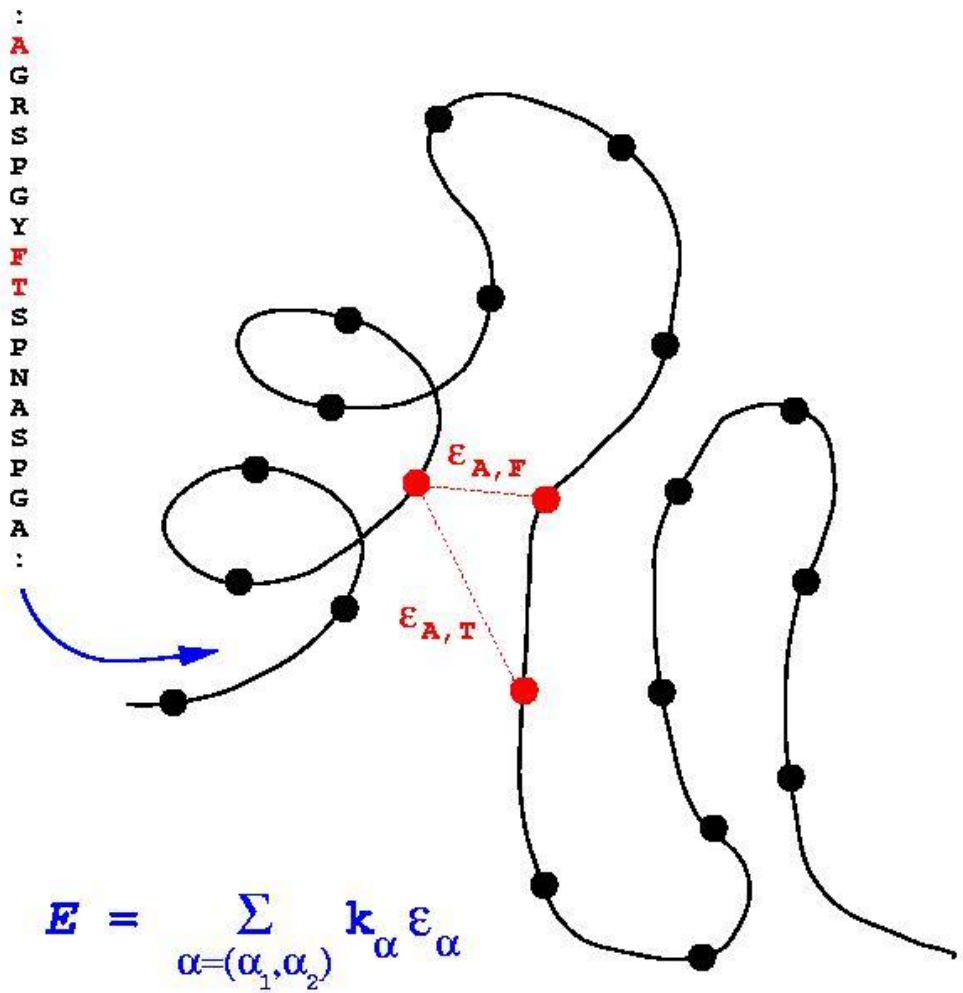

Γνωστά διπλώματα
(folds)
Πρωτεΐνη χωρίς γνωστή
δομή (target)
?
Μοντελοποίηση

- 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 πρωτεϊνη 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 πρωτεϊνης 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 (D. T. Jones, Taylort, & Thornton, 1992) και αρχικά αναφερόταν αποκλειστικά στη χρήση της τρισδιάστατης δομής. Σήμερα παρόλα αυτά, οι όροι ύφανση και αναγνώριση διπλώματος χρησιμοποιούνται συνήθως χωρίς διάκριση.


ACDEFG...
sequence of interest
...
structure library
500 – 5000 structures



Inter-residue folding potentials



$$E = \sum_{\alpha=(\alpha_1,\alpha_2)} k_\alpha \varepsilon_\alpha$$

α_1, α_2 – types of amino acids in contact

k_α – number of contacts of type α

Λογισμικό

- 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 (D. T. 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** (D. T. Jones, 1999) το οποίο βασίζεται στην πρόγνωση δευτεροταγούς δομής του PSI-PRED, αλλά και σε επιπλέον βήμα με χρήση νευρωνικού δικτύου για να δώσει μια συνολική τιμή για την αξιοπιστία της μεθόδου και είναι διαθέσιμο μαζί με τις υπόλοιπες προγνώσεις του συκγκεκριμένου 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.

συνέχεια

- 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.

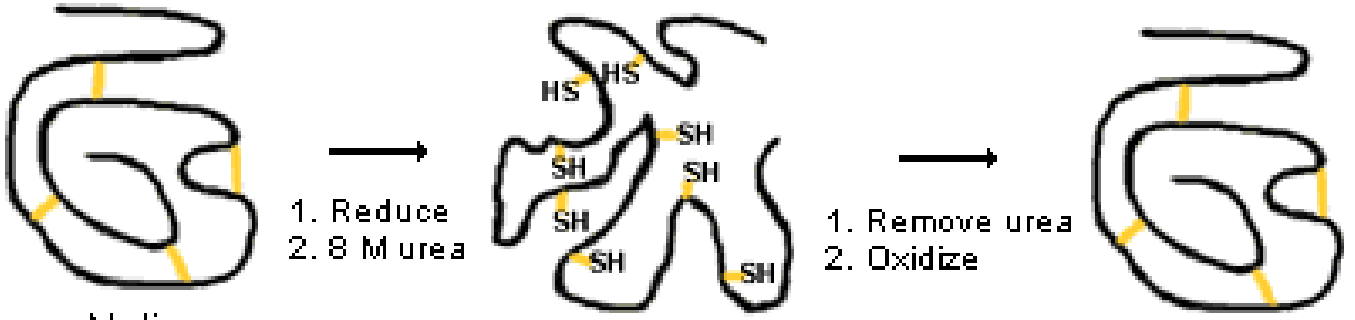
συνέχεια

- 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 δουλευουν 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 χρησιμοποεί 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 είναι σήμερα η 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 μοντελοποιούναι ab initio. Όταν κανένα πρότυπο δεν βρεθεί, τότε το λογισμικό θα κατασκευάσει μοντέλο με μέθοδο ab initio για ολόκληρη την πρωτεϊνη. Στο τελευταίο στάδιο γίνεται βελτιστοποίηση του μοντέλου με προσομοιώσεις Το I-TASSER, ένσωματώνει επίσης μια σειρά βελτιώσεις που επιτρέπουν στο χρήστη να εισάγει δομική πληροφορία με τη μορφή περιορισμών, όπως επαφές των αμινοξέων, δευτεροταγής δομή κ.ο.κ. Οι περιορισμοί αυτοί μπορεί να είναι ιδιαίτερα χρήσιμη σε περίπτωση που τα πρότυπα είναι λίγα ή η ποιότητα της στοίχισης δεν είναι καλή (Roy, Kucukural, & Zhang, 2010).

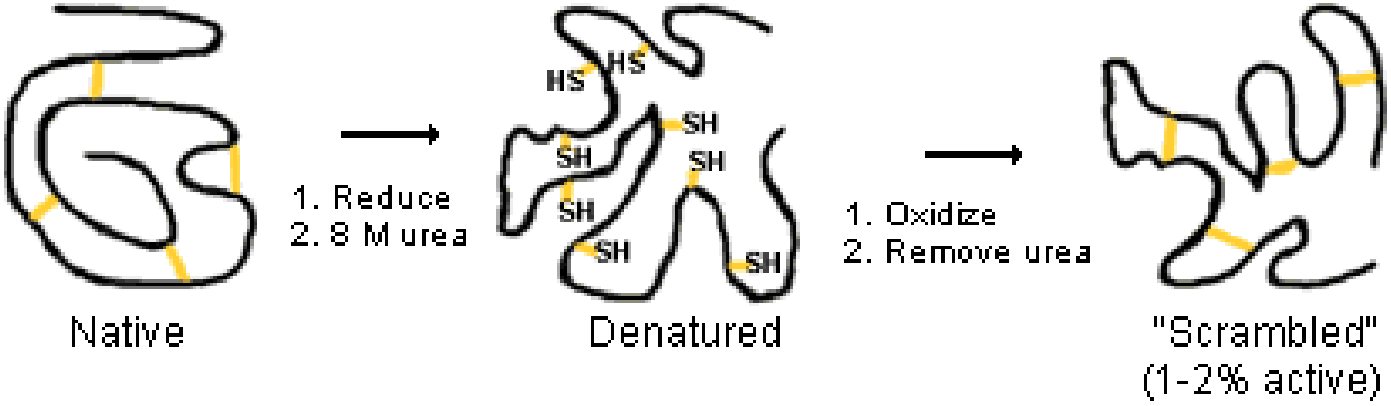
Ab initio και de novo πρόγνωση δομής

- Στην πιο ακραία περίπτωση, η πρωτεϊνή 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 κάνουμε 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 σε ποια ακριβώς κατηγορία κατατάσσεται κάθε μέθοδος, ειδικά εφόσον οι περισσότερες από αυτές χρησιμοποιούν συνδυασμό μεθοδολογιών.

The Observation:

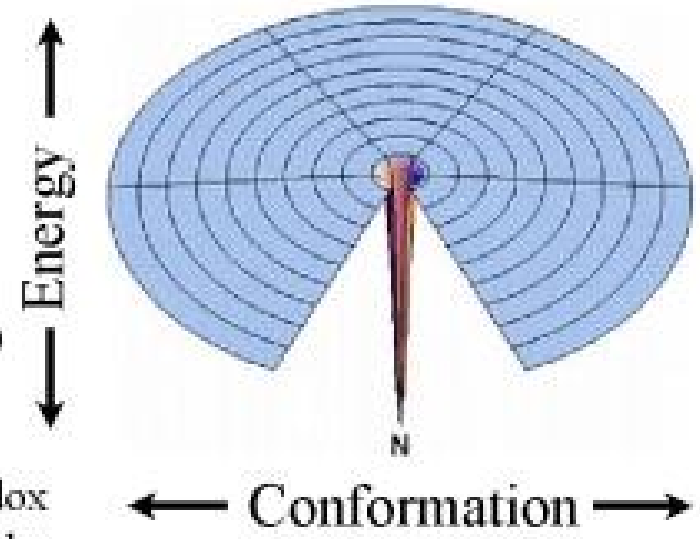


The Control:



The Levinthal Paradox

The Levinthal paradox assumes that all of the possible conformations will be sampled with equal probability until the proper one (N = native) is found. Thus, the funnel surface looks like a hole in a golf course. The paradox states that if a protein samples all 6^M conformations it will take a time longer than the age of the universe to find the native fold (N). Note that the statistical entropy is $S = R\ln\Omega = R\ln(6^M) = M\ln(6)R$ in this example. M = number of residues, 6 = number of conformations per residue



http://www.ncbi.nlm.nih.gov/pubmed/20570731

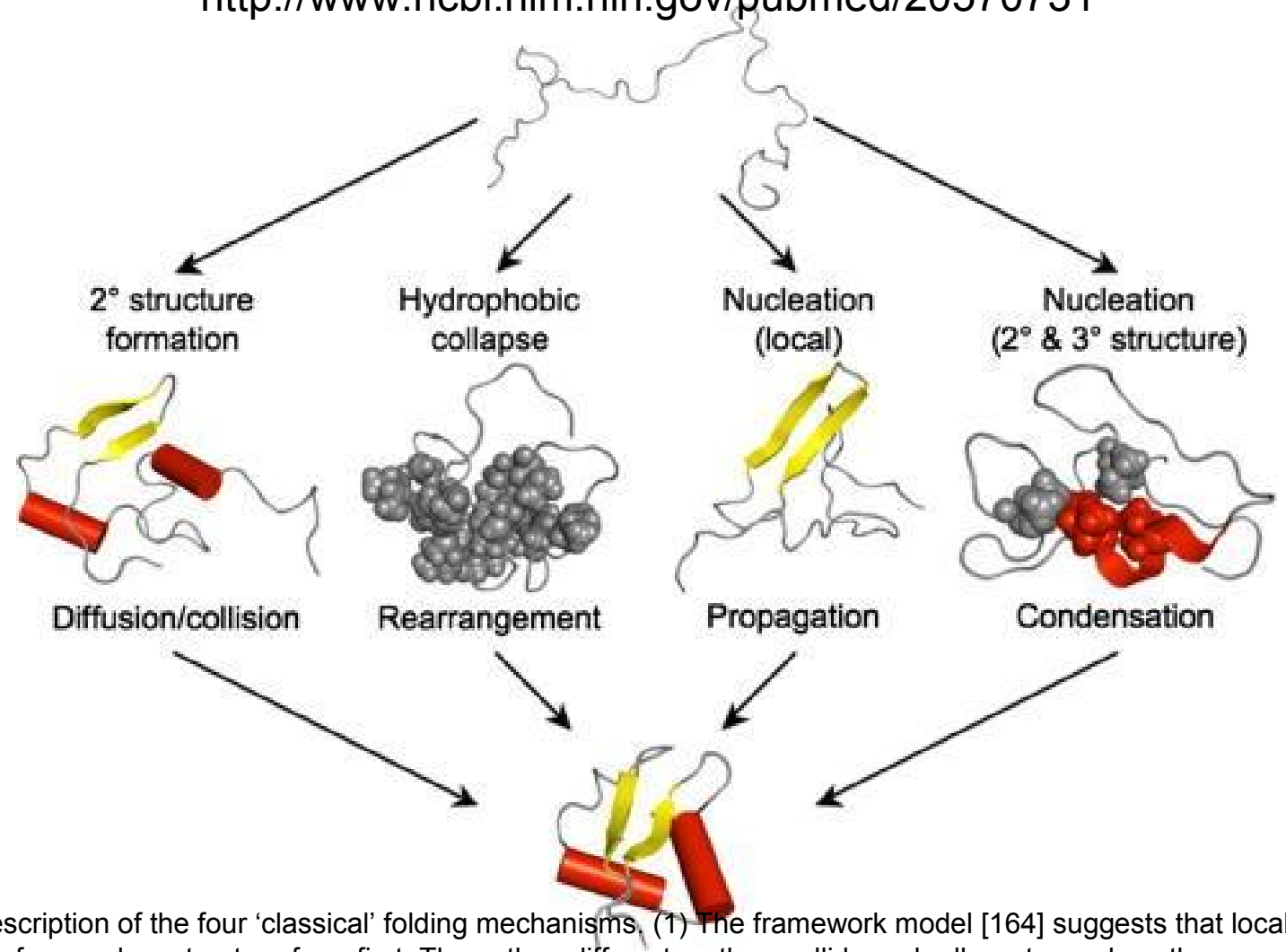
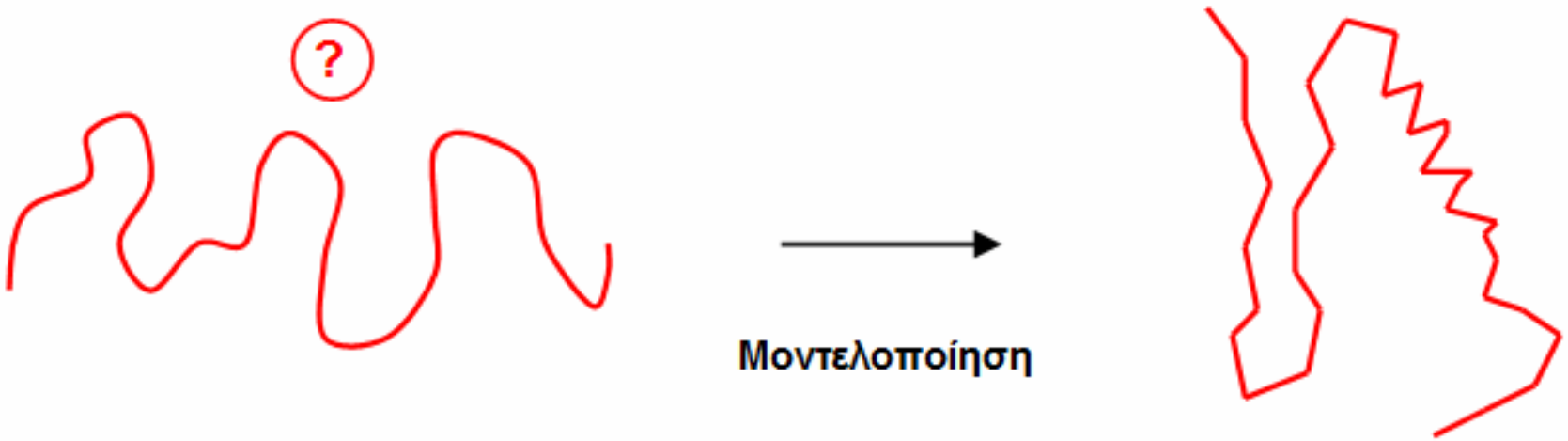


fig2: A description of the four 'classical' folding mechanisms. (1) The framework model [164] suggests that local elements of secondary structure form first. These then diffuse together, collide and adhere to produce the correct tertiary structure in the rate determining step. (2) The hydrophobic collapse model [165] implies that a protein collapses rapidly around its hydrophobic side-chains, and then rearranges from the restricted conformation of this 'molten-globule' intermediate. (3) The nucleation propagation model [166] states that local interactions form a small amount of native secondary structure, which acts as a nucleus for the outward propagation of further native structure. (4) The nucleation condensation model [167] suggests the presence of a metastable nucleus that is unable to trigger folding until a sufficient number of stabilising long-range interactions have built up. Once this occurs, the native structure condenses so rapidly that the nucleus is not yet fully formed in the transition state.

Πρωτεΐνη χωρίς γνωστή δομή (target)



Τα γενικά θέματα που έχει να αντιμετωπίσει μια τέτοια μέθοδος

- *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 πως 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 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 πως 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 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.

Λογισμικό

- 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, μια μεθοδολογία που 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).

συνέχεια

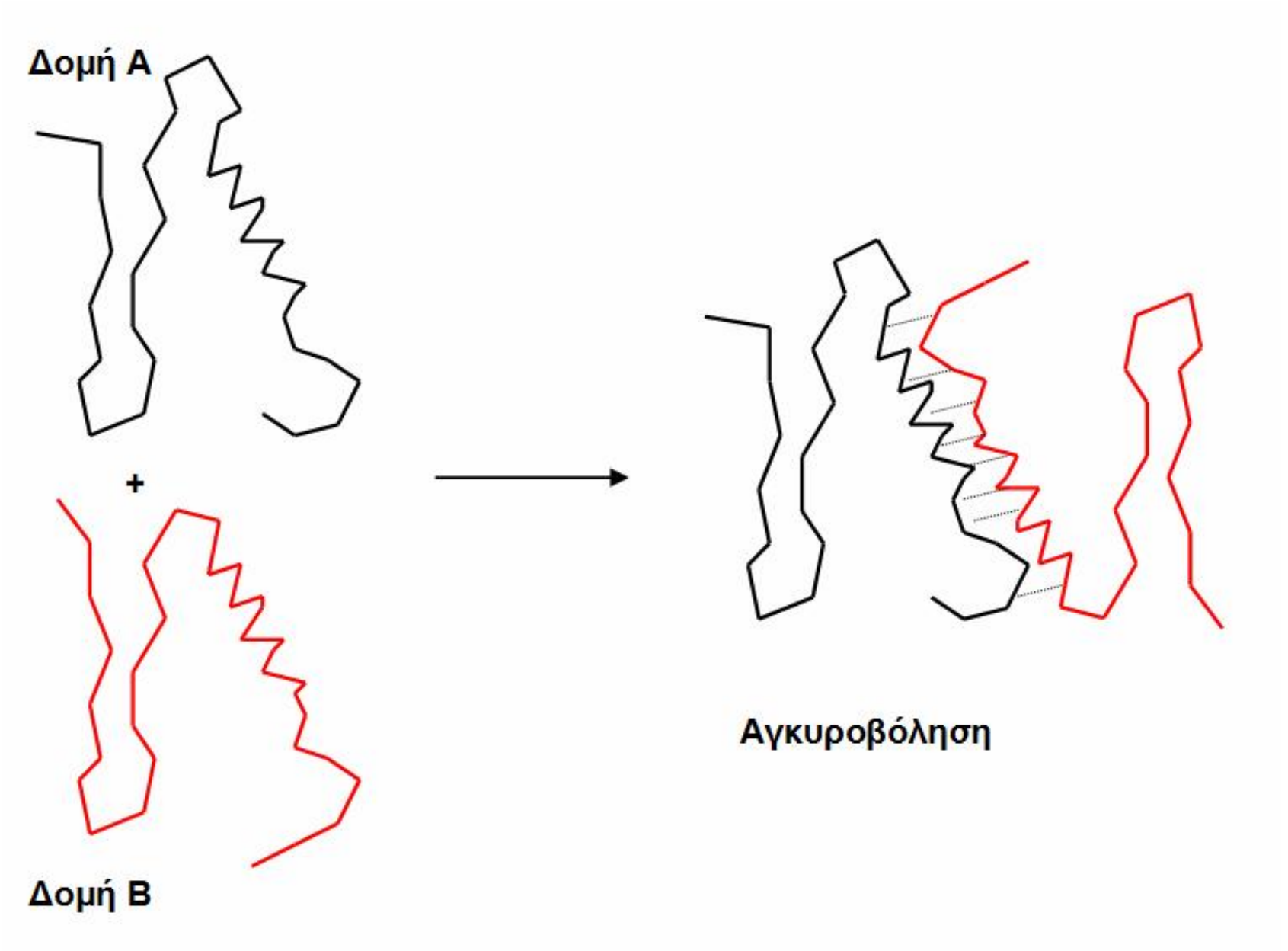
- Το **I-TASSER** (http://zhanglab.ccmb.med.umich.edu/I-TASSER/about.html) είναι μια εφαρμογή που πραγματοποιεί και ύφανση αλλά και ab initio μοντελοποίηση όταν δεν μπορεί να εντοπίσει δομές με παρόμοιο δίπλωμα. Το I-TASSER, είναι σήμερα η καλύτερη και πιο ολοκληρωμένη λύση στην πρόγνωση τριτοταγούς 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 είναι 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), τα οποία όμως ενδείκνυνται περισσότερο για πεπτίδια και μικρές πρωτεΐνες (<100 αμινοξέα), καθώς ο υπολογιστικός χρόνος για μεγαλύτερους υπολογισμούς είναι απαγορευτικός.

συνέχεια

- 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 δε 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.

Αγκυροβόληση

- Με τον όρο αγκυροβόληση ή ελλιμενισμό (docking) εννοούμε τη διαδικασία με την οποία υπολογίζουμε ή προβλέπουμε τον προτιμόμενο 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 πρωτεϊνων (Bonvin, 2006; Gray, 2006; Sternberg, Gabb, & Jackson, 1998), αλλά και μεταξύ μιας πρωτεϊνης και ενός μικρού μορίου το οποίο μπορεί να είναι ορμόνη, φάρμακο, αναστολέας, βιταμίνη κ.ό.κ. (R. D. 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 πρωτεϊνης, την ένταση και την ισχύ της δέσμευσης του μικρού μορίου, το μηχανισμό λειτουργίας, αλλά και τον μηχανισμό με τον οποίο αλληλεπιδρούν δυο πρωτεϊνες 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, μεγάλη ώθηση στο πεδίο έχει δοθεί από τέτοιες μελέτες (Alvarez, 2004).


Δομή Α
+
Δομή Β
Αγκυροβόληση

Παραλλαγές

- 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 αλλάζει (δηλαδή τα άτομα τους δεν αλλάζουν καθόλου τη σχετική τους θέση), αλλά απλά μετακινούνται για να βρεθεί η επιφάνεια επαφής. Παρόλα αυτά, ξέρουμε ότι το μοντέλο αυτό δεν είνα εκαρκές καθώς σε πολλές περιπτώσεις η πρόσδεση επηρρεάζει (σε μικρότερο ή μεγαλύτερο βαθμό) τη διαμόρφωση του κάθε μορίου (το μοντέλο της «επαγώμενης προσαρμογής»). Αυτό οδήγησε σε πιο σύνθετες τεχνικές αγκυροβόλησης, τις λεγέμενες μεθοδολογίες «ευέλικτης αγκυροβόλησης» (flexible docking) στις οποίες η τρισδιάστατη δομή των μορίων αλλάζει (έστω και ελάχιστα) για να επιτευχθεί η καλύτερη δυνατή αναγνώριση.

Κατηγορίες μεθόδων

- 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.
  - 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.

Μέθοδοι συμπληρωματικότητας σχήματος

- 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 μεθοδολογές 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, έτσι 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».

Μέθοδοι προσομοίωσης

- 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 πρωτεϊνη-πρωτεϊνη ή πρωτεϊνη-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.

Λογισμικό

- 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.

Λογισμικό

- 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 αλγορίθμου 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.

Λογισμικό

- 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, για αγκυροβόληση πρωτεϊνών αλλά και πρωτεϊνικών περιοχών. Επίσης, μπορεί να χρησιμοποιηθεί τόσο για δομές υψηλής ανάλυσης όσο και για πιο δομές 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 βλετιωμένους αλγόριθμους για αγκυροβόληση πρωτεϊνών, το **GRAMM-X** (http://vakser.compbio.ku.edu/resources/gramm/grammx/), το οποίο είναι ιδιαίτερα γρήγορο αλλά δεν μπορεί να κάνει 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 αυτοματοποιημενο. Επίσης υπάρχει και μια γραφική διεπαφή, το AutoDockTools, 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 έυχρηστη 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 χρησιμοποιόντας 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 αγκυροβοληση 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 κάνει 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 μειονέκτημα 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 is 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.

Σύγκριση

- 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 ολο 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/).