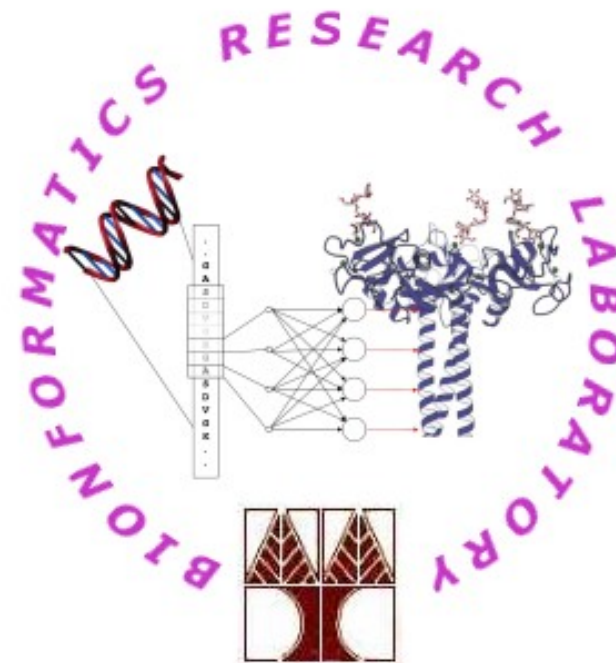


Προγνωστικές μέθοδοι με βάση αμινοξικές αλληλουχίες



Vasilis Promponas
Bioinformatics Research Laboratory
Department of Biological Sciences
University of Cyprus

ΣΥΝΟΨΗ

- Εισαγωγή
- Πρόγνωση της δομής πρωτεϊνών
- Πρόγνωση στοιχείων της πρωτεϊνικής δομής
- Συζήτηση ..

Πίσω στο “Κεντρικό Δόγμα”

Σχεδόν για όλες τις πρωτεΐνες



Το πρόβλημα ...

Ακολουθία

...LHYFRAQTVGKIMVGRRT...

**ΕΧΟΥΜΕ ΠΟΛΛΕΣ ΑΠΟ
ΔΑΥΤΕΣ!!!**

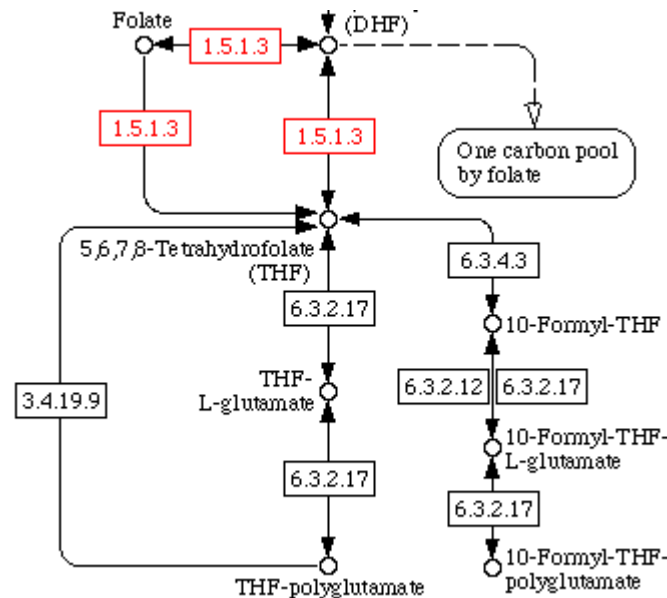
Ακολουθία

...LHYFRAQTVGKIMVVGRRT...

3D-δομή



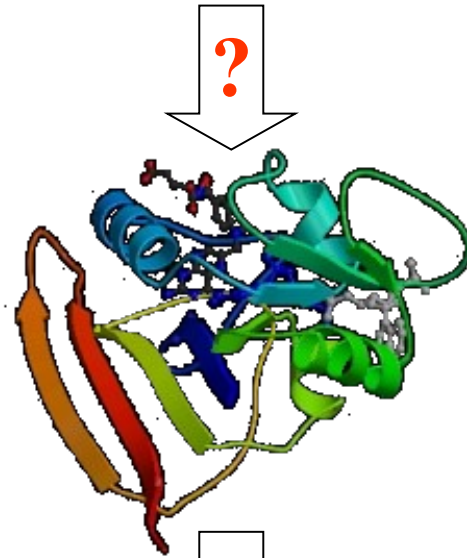
Λειτουργία



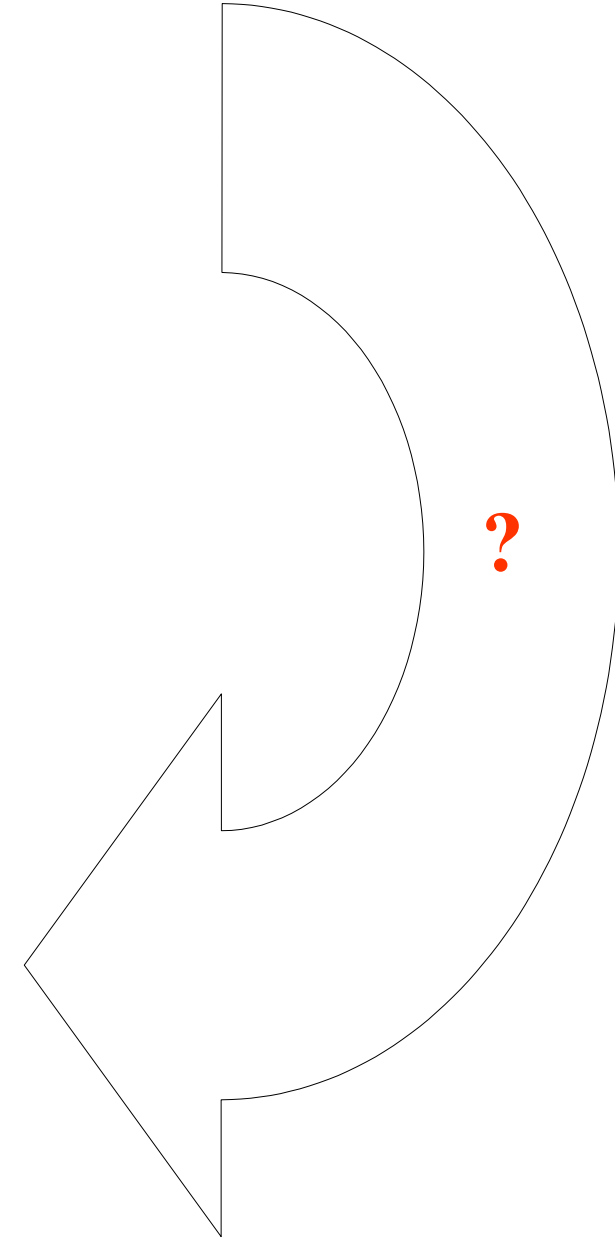
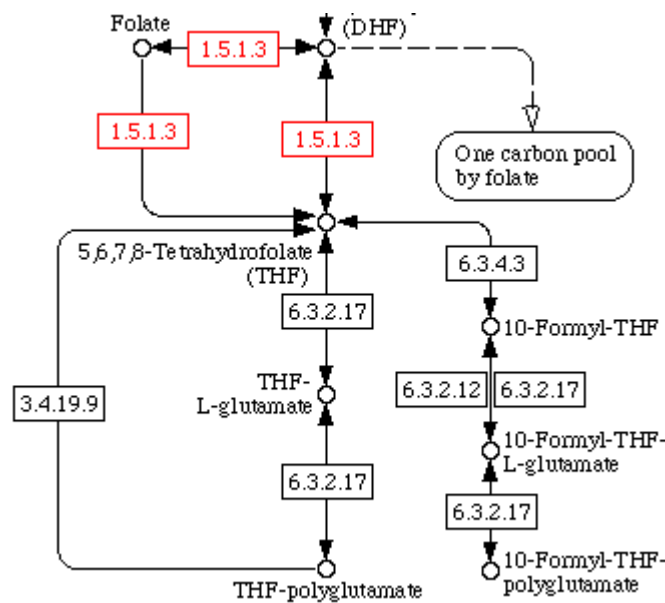
Ακολουθία

...LHYFRAQTNGKIMVGRRT...

3D-δομή



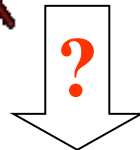
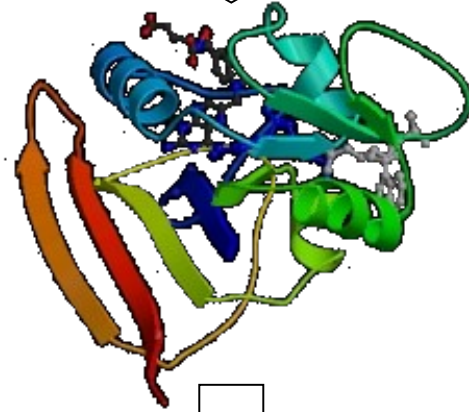
Λειτουργία



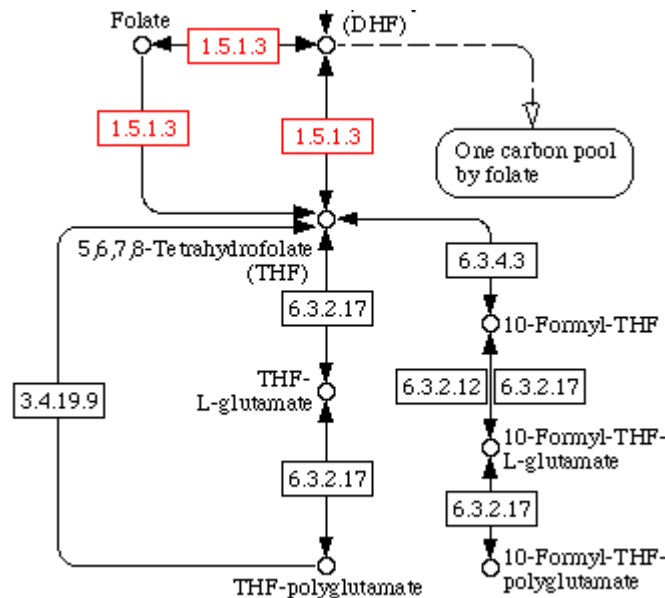
Ακολουθία

...LHYFRAQTNGKIMVGRRT...

3D-δομή



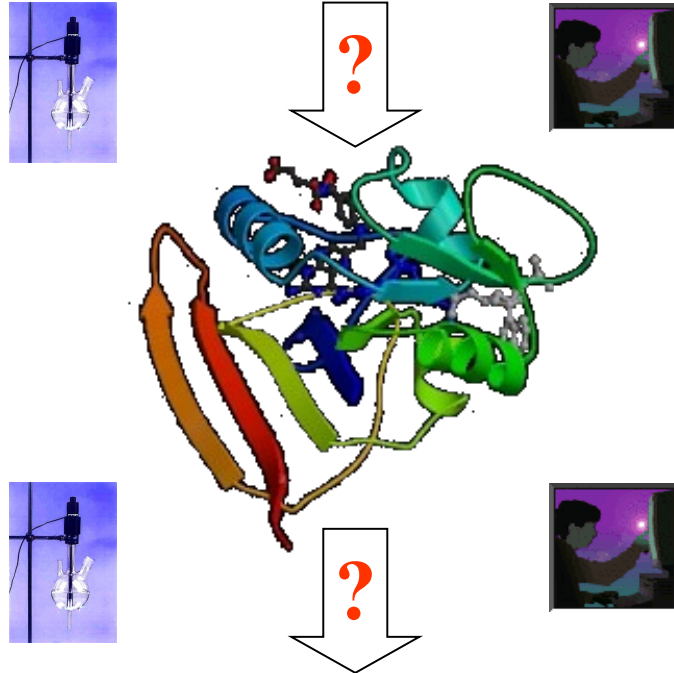
Λειτουργία



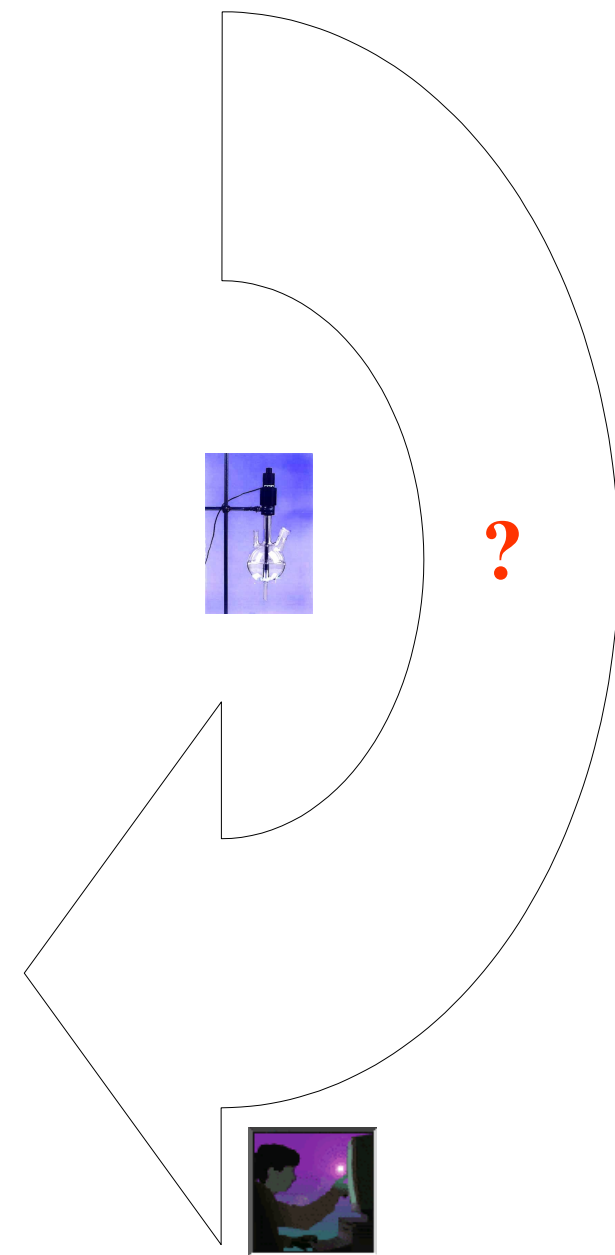
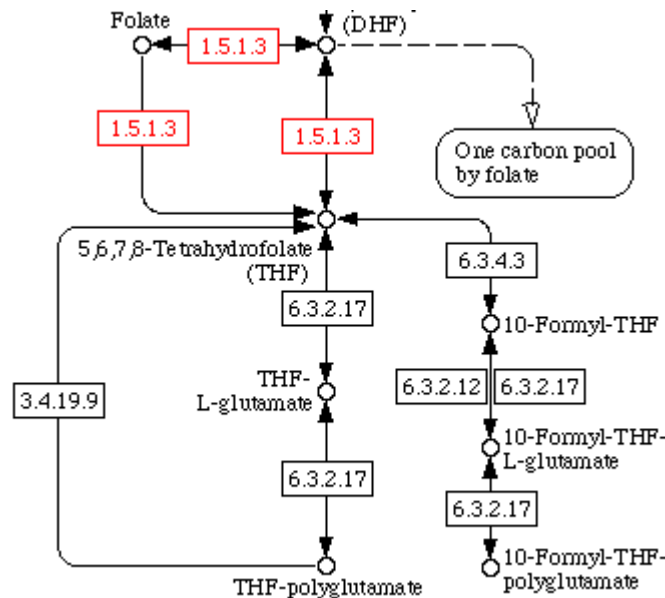
Ακολουθία

...LHYFRAQTNGKIMVGRRT...

3D-δομή



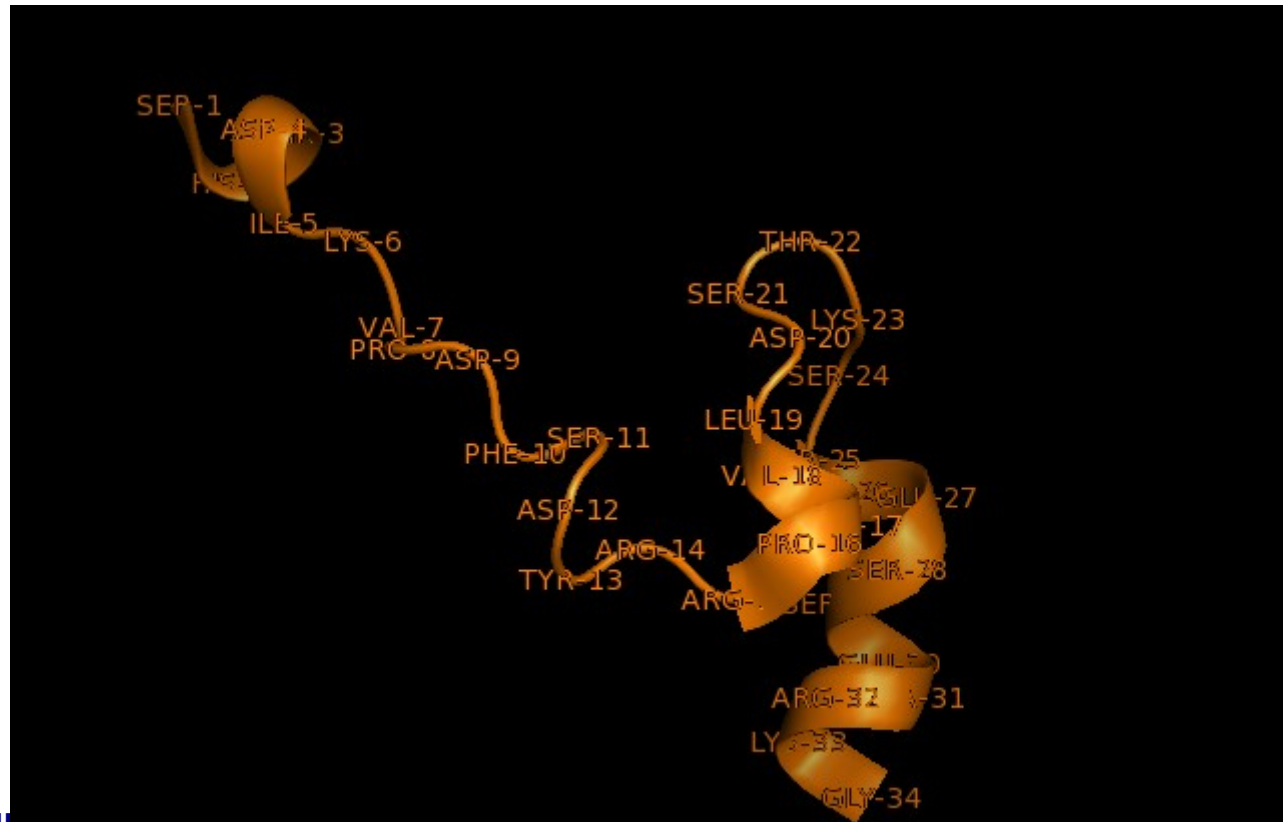
Λειτουργία



Πρόγνωση της τρισδιάστατης πρωτεϊνικής δομής

INPUT

>X-factor hypothetical protein
SHTDIKVPDFSDYRRPEVLD
STKSSKESSEARKGFSY



Πρόγνωση της δομής πρωτεϊνών

- Συγκριτική Προτυποποίηση – Comparative (Homology) Modeling
- Threading (“Αρμάθιασμα”)
- Πρόγνωση από πρώτες αρχές - *ab initio*

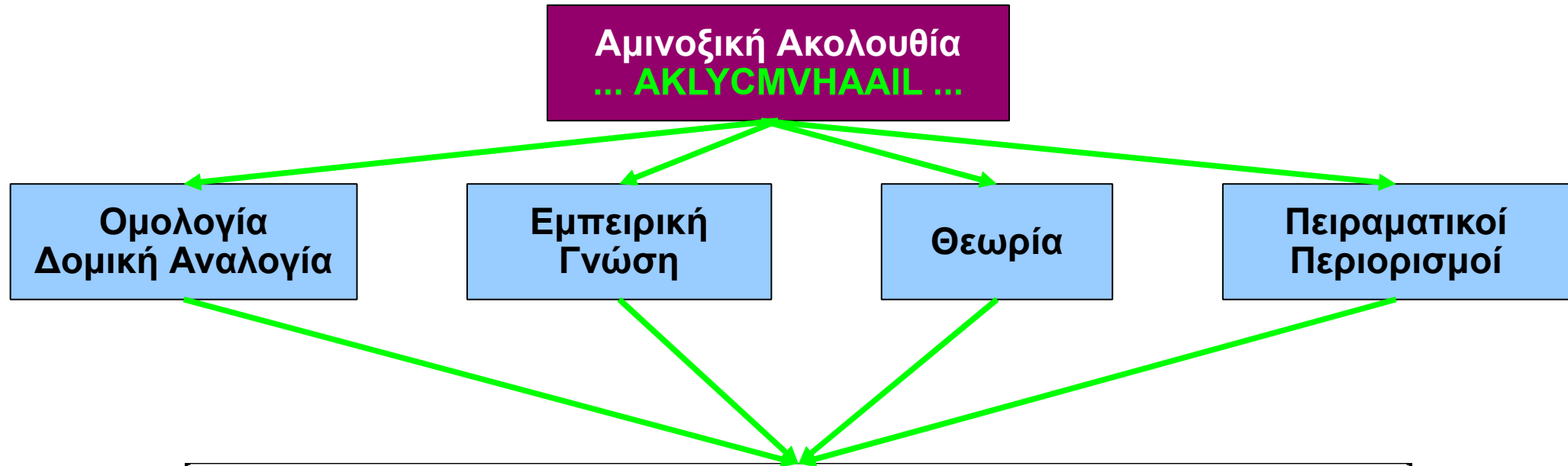
Πρόγνωση της δομής πρωτεϊνών

- ΠΡΟΒΛΗΜΑ:
 - Έχω την ακολουθία μιας νέας πρωτεΐνης
- Ερωτήματα
 - Γνωστό δίπλωμα ?
 - ΝΑΙ: μπορώ να βρώ τις λεπτομέρειες ?
 - ΟΧΙ: ποιο είναι τότε ?
- Γιατί με απασχολεί όμως ??

Αναγκαιότητα

- Δυσκολία πειραματικού προσδιορισμού
 - Επίπονη – Χρονοβόρα – Ακριβή – Εγγυημένη(?)
 - Τεράστιο πλήθος πρωτεϊνών (π.χ. Genome Projects)
 - Δύσκολη αυτοματοποίηση
- Γνώση 3D-δομής
 - Λειτουργία
 - Ορθολογικός σχεδιασμός φαρμάκων
 - Σχεδιασμός & Μηχανική Πρωτεϊνών
 - Πρόβλεψη πρωτεϊνικών αλληλεπιδράσεων
 - ...

Αφηρημένη γενική προσέγγιση



1H7W.pdb (~:/Documents/COURSES/BIO650) - gedit

Atom	ID	Element	Chain	Res	Chain	X	Y	Z	Occupancy	B-factor	Element
ATOM	1	N	ALA	A	2	121.541	61.214	47.411	1.00	38.90	N
ATOM	2	CA	ALA	A	2	120.185	61.701	47.814	1.00	37.73	C
ATOM	3	C	ALA	A	2	119.796	61.155	49.192	1.00	36.28	C
ATOM	4	O	ALA	A	2	120.015	59.983	49.472	1.00	37.48	O
ATOM	5	CB	ALA	A	2	119.162	61.259	46.773	1.00	39.01	C
ATOM	6	N	PRO	A	3	119.183	61.988	50.057	1.00	33.46	N
ATOM	7	CA	PRO	A	3	118.775	61.558	51.405	1.00	30.18	C
ATOM	8	C	PRO	A	3	117.989	60.246	51.436	1.00	25.97	C
ATOM	9	O	PRO	A	3	117.571	59.733	50.399	1.00	25.25	O
ATOM	10	CB	PRO	A	3	117.918	62.724	51.904	1.00	31.77	C
ATOM	11	CG	PRO	A	3	118.461	63.897	51.167	1.00	33.31	C

Ln 1363, Col 1 INS

... είναι εφικτό ??

- Τα καλά νέα:

- “Η θερμοδυναμική υπόθεση” (C. Anfinsen)

This hypothesis states that the three-dimensional structure of a native protein in its normal physiological milieu (solvent, pH, ionic strength, presence of other components such as metal ions or prosthetic groups, temperature, etc.) is the one in which the Gibbs free energy of the *whole system* is lowest; that is, that the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a *given environment*.

**STUDIES ON THE PRINCIPLES THAT GOVERN
THE FOLDING OF PROTEIN CHAINS**

Nobel Lecture, December 11, 1972
by
CHRISTIAN B. ANFINSEN

- Τα άσχημα νέα:

- Ο τρόπος δεν είναι πάντα προφανής

- Δεν υπάρχει αναλυτική σχέση ακολουθίας-δομής

Συγκριτική προτυποποίηση

(aka Comparative or Homology Modeling)

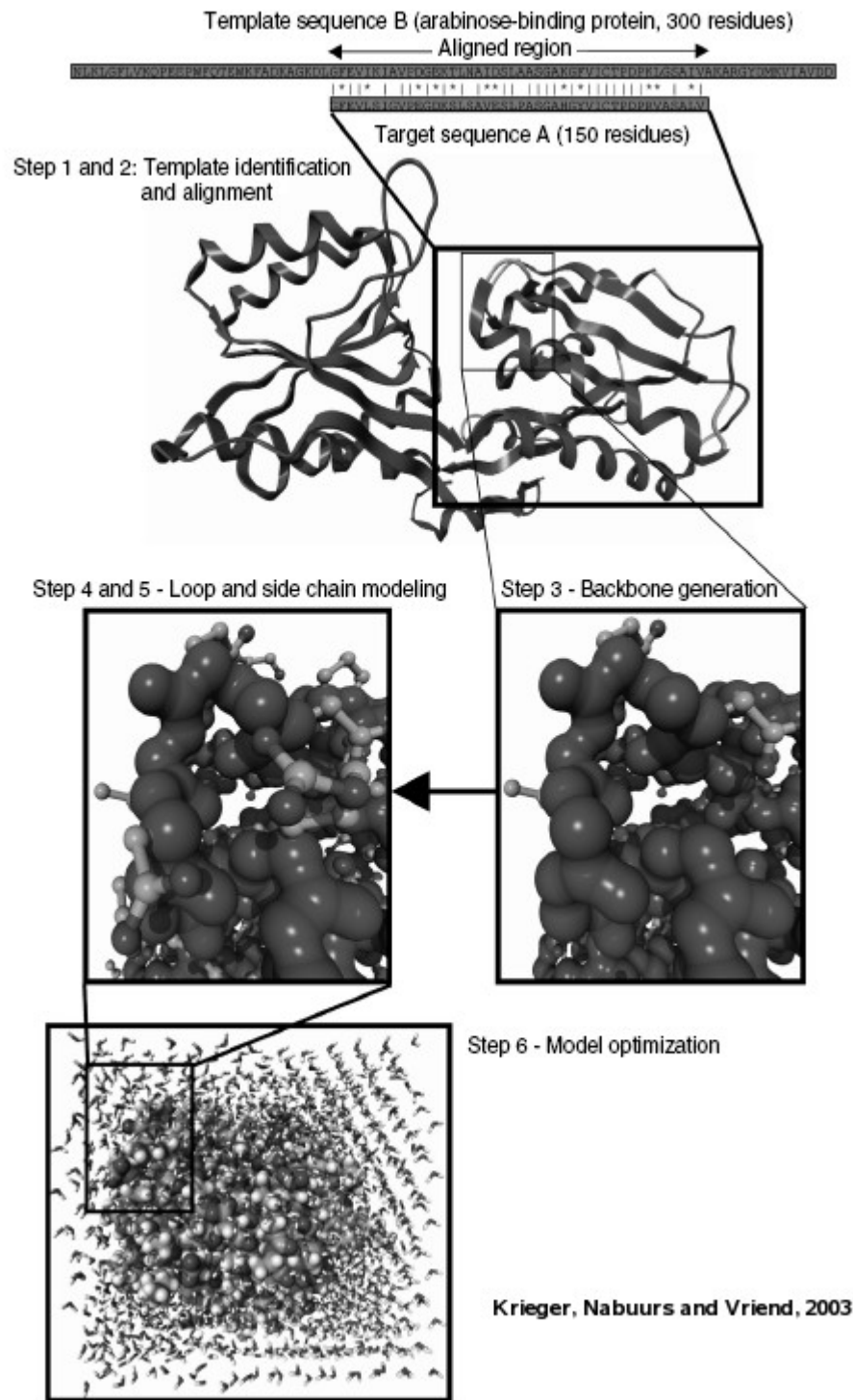
- Ξεκάθαρη σχέση με προσδιορισμένη δομή
 - Ομοιότητα στο επίπεδο της αλληλουχίας
- Βασικές Αρχές
 - Φυσικά η Θερμοδυναμική Υπόθεση
 - (Πολύ) όμοιες ακολουθίες αναμένουμε να διπλώνουν με τον ίδιο τρόπο
 - Η δομή περισσότερο συντηρημένη από ακολουθία

Συγκριτική προτυποποίηση

(aka Comparative or Homology Modeling)

1. Εντοπισμός κατάλληλης δομής-μήτρας (template) και αρχική στοίχιση (seq) [αν δεν βρω?]
2. Βελτίωση της στοίχισης (seq)
3. Κατασκευή κύριας αλυσίδας
4. Προτυποποίηση θηλιών
5. Προτυποποίηση πλευρικών αλυσίδων
6. Βελτιστοποίηση μοντέλου
7. Επικύρωση μοντέλου

Σε κάθε βήμα υπάρχουν εναλλακτικοί τρόποι προσέγγισης



1. Εντοπισμός κατάλληλης δομής-μήτρας (template) - στοίχιση

- Βασιζόμαστε στις υπάρχουσες μεθοδολογίες (DP, BLASTP, FASTA – μη βιάζεστε, στα επόμενα ...)
- Αναζήτηση έναντι των ακολουθιών εγγραφών της PDB
- Δυνητικά σε 2 βήματα
- Κρίσιμα σημεία
 - Επιλογή 1(??) δομής-μήτρας [κριτήρια??]
 - Στοίχιση στο επίπεδο ακολουθίας

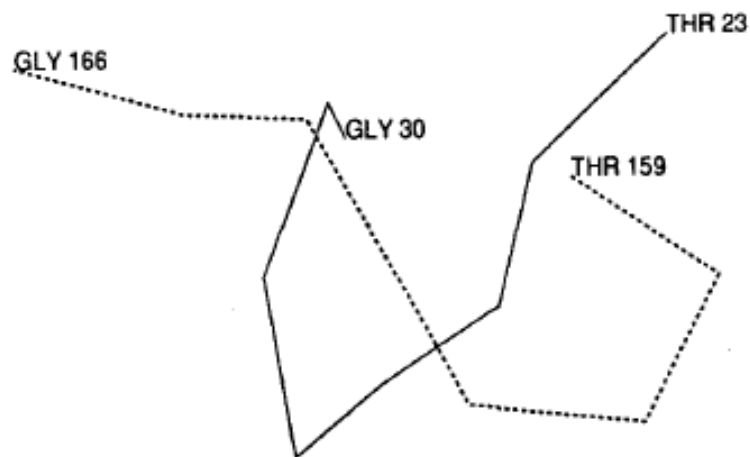


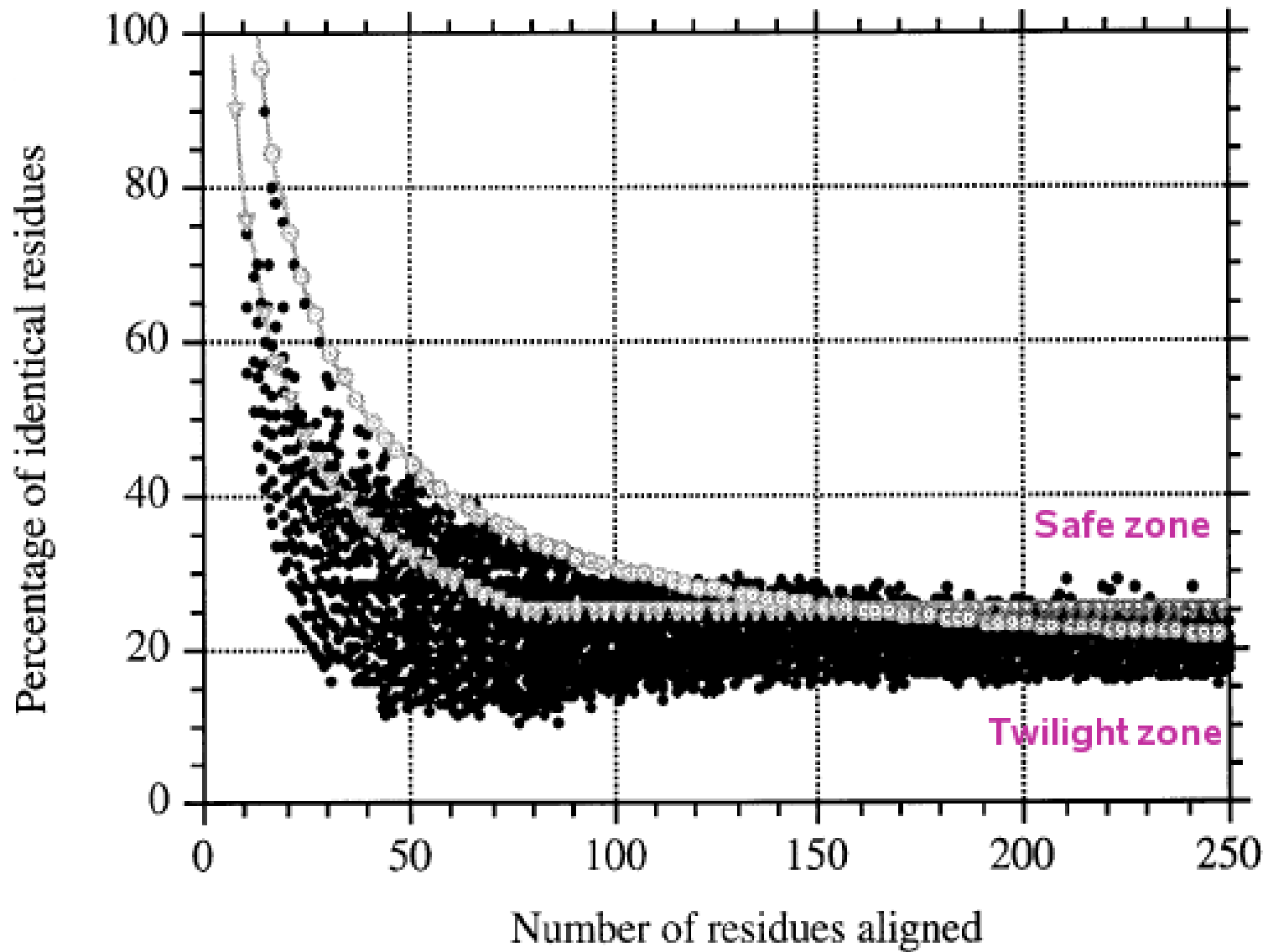
Fig. 1. The structural meaning of sequence similarity depends strongly on the length of alignments. In this extreme example, two short peptides have sequence similarity normally sufficient for structural homology (75% identical residues), yet their structures are very different. Residues 159–166 of a subtilisin protease (dashed line, data set 2SBT⁷) and residues 23–30 of an immu-

TABLE I. Homology Threshold for Different Alignment Lengths*

Alignment length L (number of residues)	Homology threshold t (% residue identity)
<10	–
10	79.6
12	71.9
14	65.9
16	61.2
18	57.2
20	53.9
22	51.1
24	48.7
26	46.6
28	44.7
30	43.0
35	39.4
40	36.6
45	34.2
50	32.3
55	30.6
60	29.1
65	27.8
70	26.7
80	24.8
>80	24.8

noglobulin (solid line, data set 3FAB⁸) have 6 out of 8 identical residues (TGSSSTVG/TGSSSNIG), but differ by 4.7 Å rms deviation in C(α) positions. Secondary structures are also very different (LTTSLLLL/ELLTTSST where T, H-bonded turn; S, geometrical turn; E, part of beta strand; L, extended loop). Protein fragments as stereo C(α) traces.

Sander and Schneider, 1993

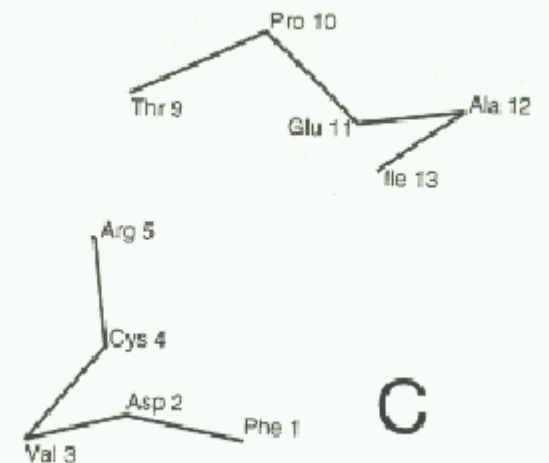
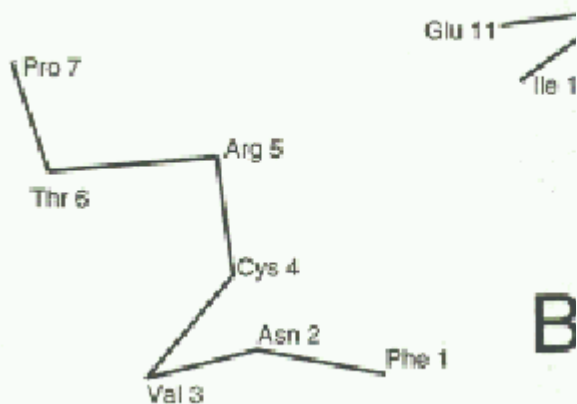
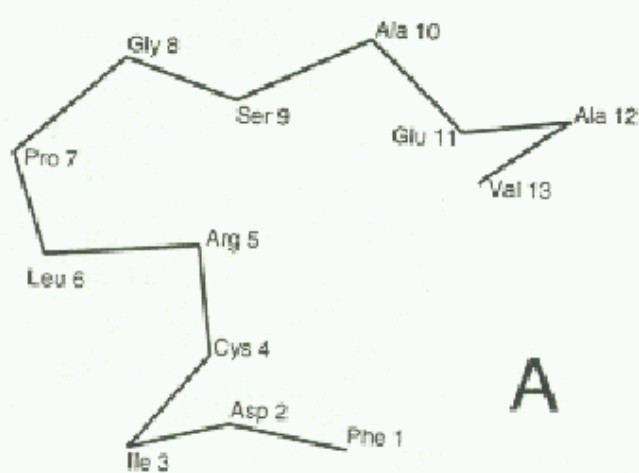
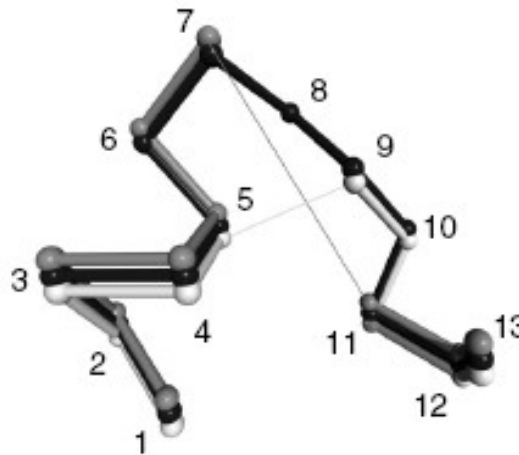


Rost, 1999

2. Βελτίωση της στοίχισης

- Δυναμικός προγραμματισμός (δείτε επόμενα)
 - Μαθηματικά βέλτιστη στοίχιση
 - Βιολογικά/Δομικά??
- Πιθανή χρήση (επιπλέον) ομόλογων
 - profile alignment
 - multiple sequence alignment
- Αξιοποίηση δομικών περιορισμών
 - Κανονικές δευτεροταγείς δομές
 - Θηλιές

	1	2	3	4	5	6	7	8	9	10	11	12	13
Template	PHE	ASP	ILE	CYS	ARG	LEU	PRO	GLY	SER	ALA	GLU	ALA	VAL
Model (bad) 1	PHE	ASN	VAL	CYS	ARG	ALA	PRO	---	---	---	GLU	ALA	ILE
Model (good) 2	PHE	ASN	VAL	CYS	ARG	---	---	---	ALA	PRO	GLU	ALA	ILE



3. Κατασκευή κύριας αλυσίδας

- “Αντιγραφή” από τη δομή-μήτρα
- Αποφυγή λαθών εγγραφών PDB
 - Εντοπισμός
 - πχ PDBREPORT: <http://swift.cmbi.ru.nl/gv/pdbreport/>
 - επιλογή μήτρας με τα λιγότερα σφάλματα
- Χρήση πολλαπλών μητρών
 - Πλεονεκτικό: κάλυψη μεγαλύτερου τμήματος της άγνωστης πρωτεΐνης
 - Δεν είναι πάντα απλό ...

(Εντοπισμός πιθανών σφαλμάτων)

- Μικρές αποστάσεις ατόμων
- Ασυνήθιστα μήκη/γωνίες δεσμών
- Ασυνήθιστα διαστροφόμερη
- Missing atoms
- ...

4. Προτυποποίηση θηλιών

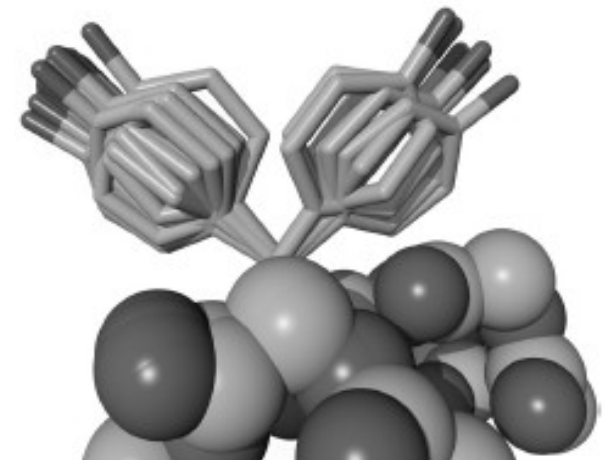
- Στοιχίσεις με κενά
 - Insertions => δεν υπάρχει δομή για τη θηλιά
 - Deletions => πρέπει να ενωθούν μη διαδοχικά κατάλοιπα
- Απαιτούνται στεροδιαταξικές αλλαγές στην κύρια αλυσίδα της μήτρας
 - όχι σε κανονικές δευτεροταγείς δομές
 - δεν είναι εύκολο να προβλεφθούν
 - υπάρχει εγγενές πρόβλημα ακόμη και χωρίς κενά
 - επιφανειακές θηλιές => κρυσταλλογραφικές επαφές
 - μέγεθος πλευρικών αλυσίδων

4. Προτυποποίηση θηλιών (II)

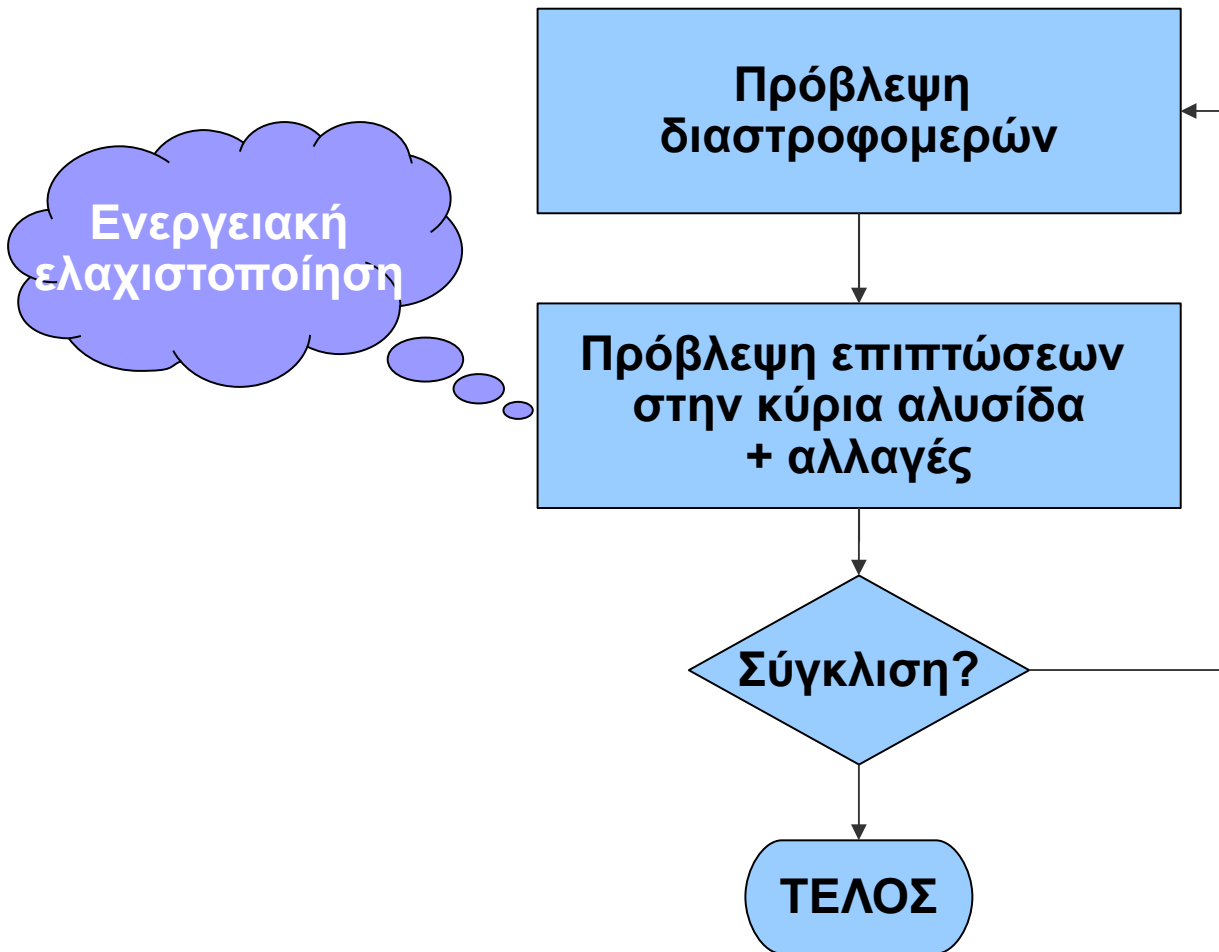
- Δύο κύριες προσεγγίσεις
 - Knowledge based
 - Αναζήτηση στην PDB για θηλιές με τα ίδια άκρα
 - Αντιγραφή της στερεοδιάταξης
 - Energy based
 - Χρησιμοποίηση *ab initio* τεχνικών (δείτε παρακάτω ??)
 - Ελαχιστοποίηση ενέργειας (Monte Carlo, Simulated annealing)
 - Αποδοτικές μόνο για μικρές θηλιές (<10aa)

5. Προτυποποίηση πλευρικών αλυσίδων

- Ύπαρξη διαστροφομερών (rotamer)
 - Διαφορετικές στερεοδιατάξεις σχετικά με την κύρια αλυσίδα
 - Συχνά ομόλογες δομές έχουν παρόμοιες χ^1
- Προτιμήσεις => rotamer libraries
 - Επιρροή από γειτονικές πλευρικές αλυσίδες
 - Επιρροή στην κύρια αλυσίδα
 - Επιρροή από την κύρια αλυσίδα
- Απόδοση
 - Καλή στον υδρόφοβο πυρήνα
 - Μειωμένη σε επιφανειακά κατάλοιπα



6. Βελτιστοποίηση μοντέλου



6. Βελτιστοποίηση μοντέλου (II)

- Ελαχιστοποίηση ενέργειας
 - Στο σύνολο της δομής!!!
 - Απαιτείται ακρίβεια στη συνάρτηση ενέργειας
 - Αποφυγή “ΧΟΝΤΡΩΝ” λαθών
 - Εισαγωγή μικρών (συχνά ανεπιθύμητων) λαθών
- Χρήση δομικών περιορισμών
- Χρήση με φειδώ ...
- Χρήση τεχνικών μοριακής δυναμικής

7. Επικύρωση μοντέλου

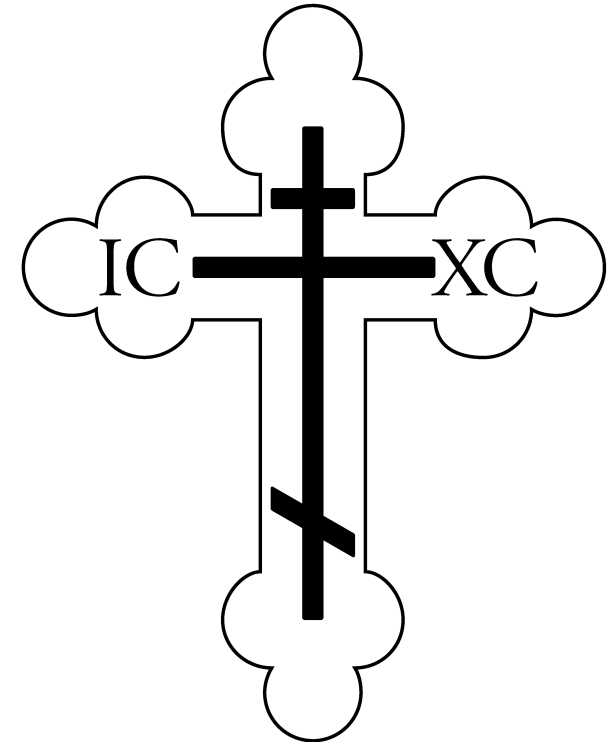
- Εντοπισμός πιθανών λαθών
 - Χρήση ενεργειακών υπολογισμών
 - Έλεγχος “κανονικότητας”
 - μήκη δεσμών
 - (επίπεδες) γωνίες δεσμών
 - δίδρες γωνίες
 - 3D κατανομή πολικών/υδρόφοβων καταλοίπων
 - γεωμετρία ατομικών επαφών
- Το πλήθος (και η σοβαρότητα) τους εξαρτάται
 - ομοιότητα σε επίπεδο ακολουθίας
 - σφάλματα στη δομή-μήτρα

Συγκριτική προτυποποίηση: Επιδόσεις

- Πλεονεκτήματα
 - Ταχύτητα
 - Ακρίβεια (RMSD 1-3Å)
- Μειονεκτήματα
 - Ακρίβεια εξαρτώμενη της περιοχής (πχ θηλιές)
 - Εξαρτάται από την ποιότητα στοίχισης, μέγεθος PDB, αντιπροσωπευτικού δείγματος δομών
 - Προβληματική στο 'Twilight Zone', ORFans
 - Δε δίνει πληροφορία για το μηχανισμό διπλώματος
 - Όχι και τόσο απλή διαδικασία τελικά :-)

Εάν δε βρω κατάλληλη δομή?

- ΑΤΥΧΗΣΕΣ!!!
- *Ab initio* και Fold recognition
- Εναλλακτικά ελάτε στο BIO650 :-)

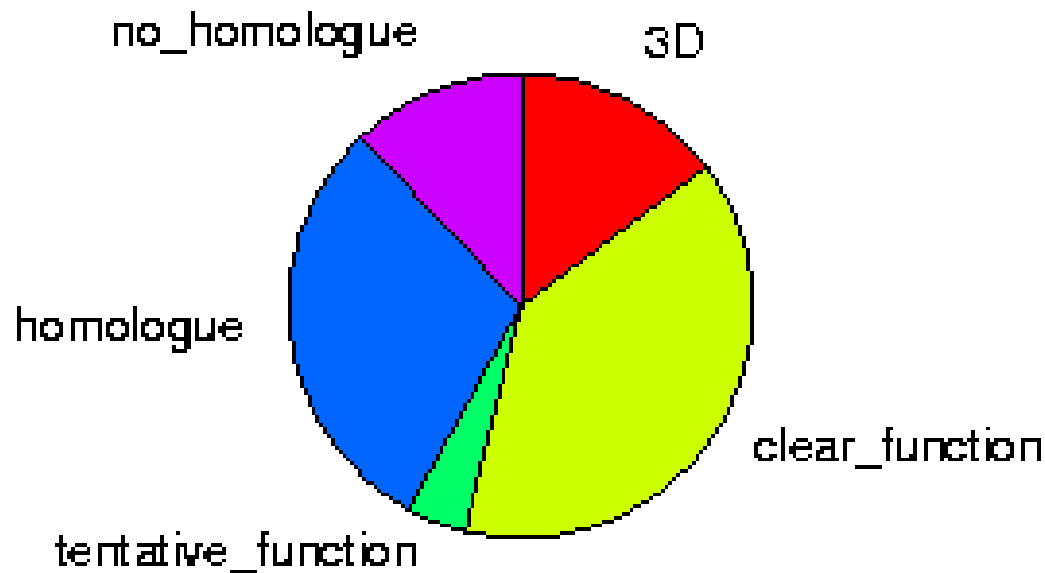


Υπολογιστική Ανάλυση Αμινοξικών Ακολουθιών

- Μέθοδοι Βασισμένοι στην Ανίχνευση Ομοιότητας (σχετικές διαλέξεις ..)
- Εμπειρικές Μέθοδοι
- Τεχνικές Μηχανικής Μάθησης
- Αυτοματοποιημένα ή «με το χέρι??»

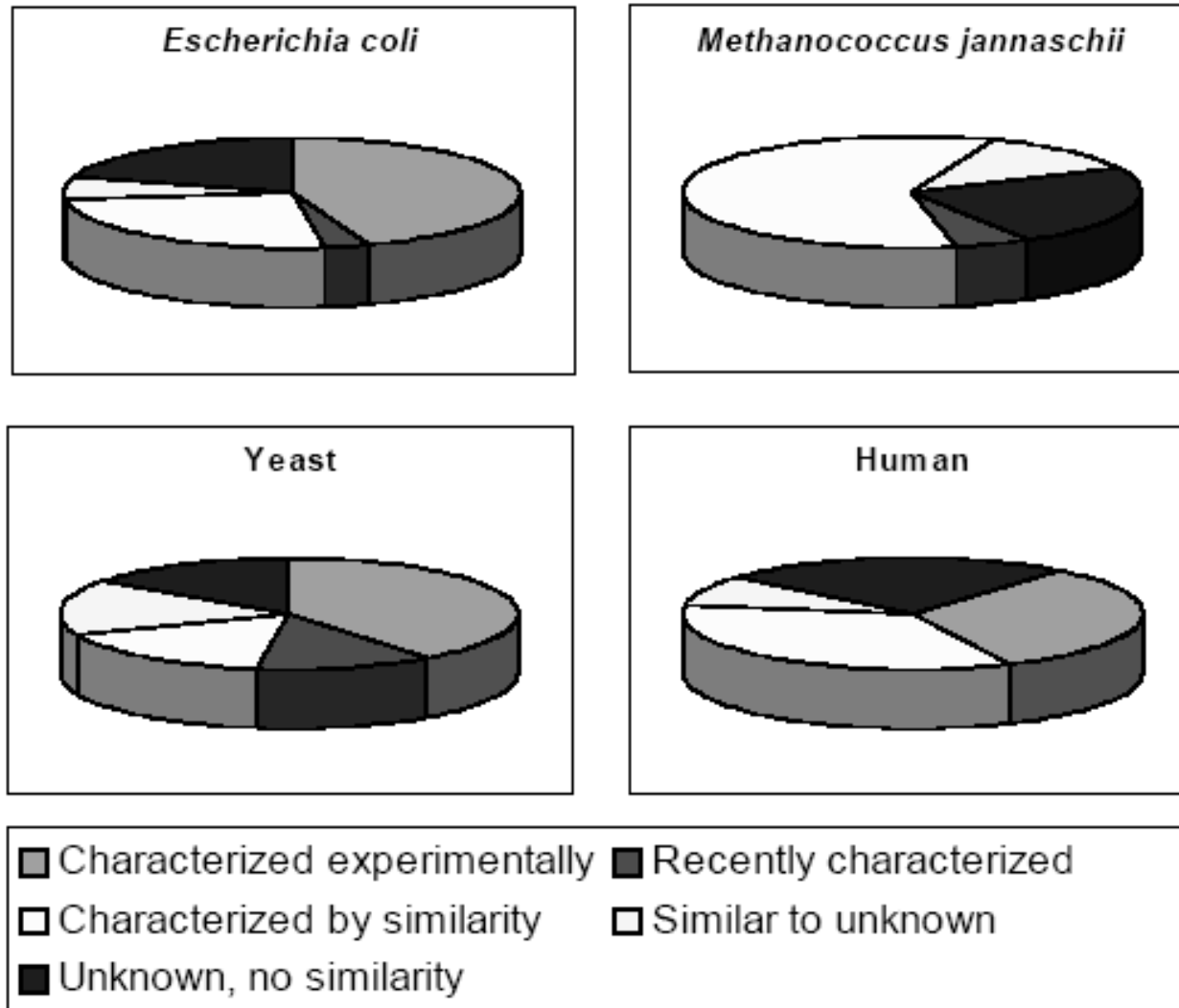
Αυτοματοποιημένα ??

Αυτοματοποιημένος Σχολιασμός βασισμένος σε ομοιότητες (σύστημα GeneQuiz, Μάϊος 2000) για τα ORFs του γονιδιώματος του Αρχαίου *Methanococcus jannaschii*.



<http://jura.ebi.ac.uk:8765/ext-genequiz//genomes/mj0005/index.html>

Αυτοματοποιημένα ?? (2)



Από Koonin EV and Galperin M, (2003) "Sequence – Evolution – Function: Computational Approaches in Comparative Genomics"

Πρόγνωση στοιχείων της τρισδιάστατης δομής

- Πρόγνωση δευτεροταγούς δομής σφαιρικών υδατοδιαλυτών πρωτεϊνών
- Πρόγνωση της τοπολογίας διαμεμβρανικών πρωτεϊνών
 - α-ελικοειδείς διαμεμβρανικές πρωτεΐνες
 - διαμεμβρανικά β-βαρέλια
- Πρόγνωση πεπτιδίων-οδηγών
- Πρόγνωση μετα-μεταφραστικών τροποποιήσεων

Συζήτηση ...

Διδακτικό υλικό:

<http://troodos.biol.ucy.ac.cy/BRL/courses/BIO331/index.html>