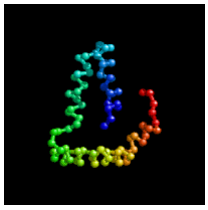
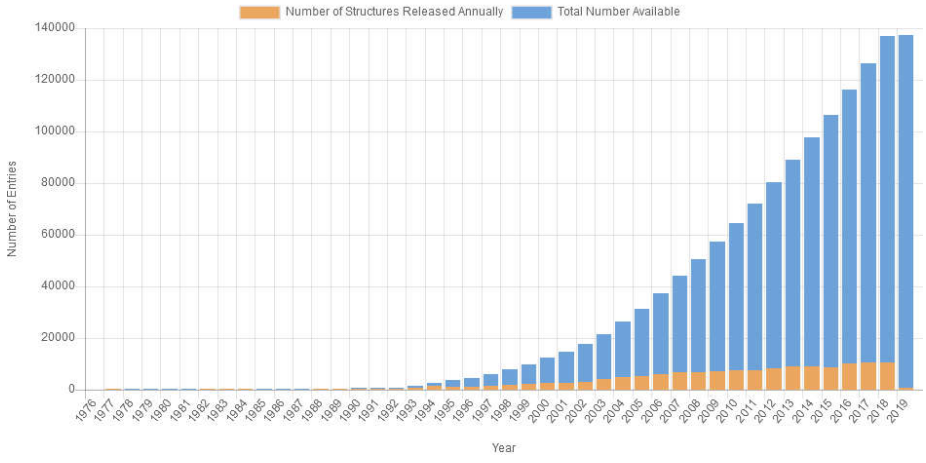


Πρόβλεψη δομής πρωτεϊνών

(Prediction of Protein Structure)

http://lectures.molgen.mpg.de/ProteinStructure/ComparativeModelling/

UniProtKB/TrEMBL PROTEIN DATABASE RELEASE 2019_01 STATISTICS
Release 2019_01 of 16-Jan-2019 of UniProtKB/TrEMBL contains
139694261 sequence entries



- Ενώ οι ακολουθίες εξελίσσονται, οι 3D δομές τείνουν να συντηρούνται
- Η λειτουργία τείνει επίσης να συντηρείται
- Όμως η λειτουργία τείνει να αλλάζει ταχύτερα από ό,τι η δομή

Εξελικτικές οικογένειες πρωτεϊνών

- Στο πλαίσιο μιας εξελικτικής οικογενείας πρωτεϊνών αναμένουμε:
  - μόνο μια βασική 3D δομή
  - ίσως περισσότερες από μια διαφορετικές λειτουργίες
- Οι λειτουργικές διαφορές μπορεί να είναι ελάσσονες ή μείζονες
  - μεταβολή στην ενζυμική εξειδίκευση (ελάσσονες)
  - μεταβολή από ένζυμο σε δομική πρωτεΐνη (μείζονες).

Τί είναι η πρόβλεψη της δομής πρωτεϊνών;

- Στην πιο γενική της μορφή
  - Πρόβλεψη της σχετικής θέσης στον χώρο κάθε ατόμου, προερχόμενη από την γνώση μόνο της πρωτοταγούς δομής (ακολουθία)

Γιατί πρόβλεψη δομής;

- Χάσμα ακολουθίας - δομής
  - 139694261 γνωστές ακολουθίες, ~140000 με γνωστές δομές
- Γνώση της δομής συμβάλει στην
  - κατανόηση του μηχανισμού λειτουργίας
  - πρόβλεψη της λειτουργίας

Γιατί πρόβλεψη δομής;

- Σχεδιασμός φαρμάκων με βάση την δομή
- Κατανόηση των αποτελεσμάτων στην δομή ή στην λειτουργία προερχόμενα από μεταλλάξεις
- Εξαιρετικά ενδιαφέρουσα επιστημονική πρόκληση
  - Παραμένει άλυτο πρόβλημα στην γενική του μορφή μετά από 30ετείς ερευνητικές προσπάθειες

Μέθοδοι πρόβλεψης δομής

- Συγκριτικός σχεδιασμός (Comparative modelling)
- Πρόβλεψη δευτεροταγούς δομής
- Αναγνώριση αναδίπλωσης (Fold recognition/threading)
- *Ab initio* προσεγγίσεις

Συγκριτικός σχεδιασμός (Comparative modelling)

- Προβλέπει τριτοταγή δομή επί τη βάση
  - ***γνωστών*** δομών πρωτεϊνών με όμοια ακολουθία προς την πρωτεΐνη «στόχο» (target), οι οποίες καλούνται «πρότυπα» (templates)
  - στοίχισης μεταξύ των προτύπων και του στόχου
- Υπενθύμιση: ~25% seq ID σημαίνει ότι δύο πρωτεΐνες έχουν την ίδια βασική δομή

Συγκριτικός σχεδιασμός (Comparative modelling)

- Επιλογή κατάλληλης ακολουθίας προτύπου

PQFHLWKRPVVTAHIEGQPVEVLLDTGAEDSIVTAIEIGVHYTPKIVGGIGGFINTKEYK
NVEVEVLGKRIKGTIMTGNTPMNIFGRNLLTALGMSLNF

- Προσαρμογή του «στόχου» στο «πρότυπο»
- Κατασκευή του σκελετού
- Κατασκευή βρόγχων και πλευρικών αλυσίδων

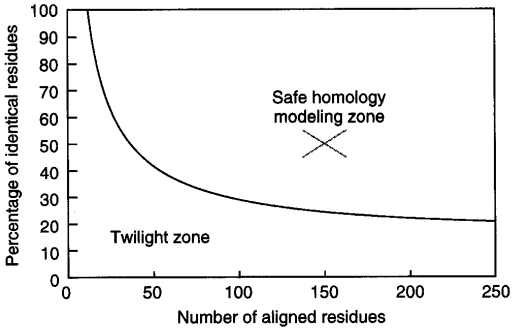
Εύλογα αποτελέσματα >50% Seq ID

Συγκριτικός σχεδιασμός (Comparative modelling)

- Ανεύρεση καταλλήλου προτύπου
  - Blast-search στην PDB

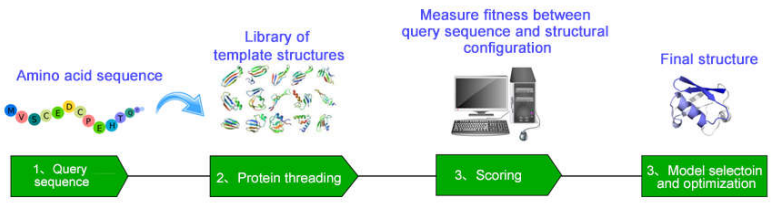
(https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins, https://web.expasy.org/blast/)

  - Ποσοστό ταυτότητας και συντηρημένες περιοχές



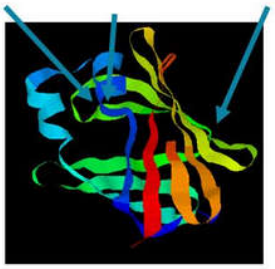
Αναγνώριση αναδίπλωσης (Fold recognition/Threading)

- Αναγνώριση αναδίπλωσης (fold recognition/ threading)
- Εξετάζεται αν η ακολουθία στόχος είναι συμβατή με μια γνωστή αναδίπλωση, ακόμα και αν δεν έχει σημαντική ομοιότητα με την ακολουθία της.



MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE

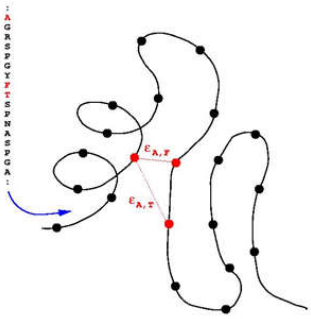
What is "probability" that two specific residues are in contact?



How well does a specific residue fit structural environment?

Alignment gap penalty?

Inter-residue folding potentials



Total energy: E_p + E_s + E_g

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

α_1, α_2 - types of amino acids in contact

k_α - number of contacts of type α

Επιλογή των μεθόδων πρόβλεψης

- Διαθέσιμες όμοιες ακολουθίες γνωστής δομής → ο *συγκριτικός σχεδιασμός* είναι ο καλύτερος τρόπος
  - όλες οι άλλες μέθοδοι είναι λιγότερο αξιόπιστες
- Όμοιες ακολουθίες γνωστής δομής δεν είναι πάντα διαθέσιμες

Αν δεν είναι δυνατός ο συγκριτικός σχεδιασμός;

- Πρώτο βήμα είναι η πρόβλεψη δευτεροταγούς δομής (1D)
  - Προβλέπει για κάθε αμινοξύ αν ανήκει σε έλικα (H), β-κλώνο (E) ή C (coil/loop)
  - Σε αντίθεση με την πρόβλεψη της τριτοταγούς δομής στην πρόβλεψη της δευτεροταγούς δομής έχουν γίνει σημαντικές πρόοδοι.

Ab initio αναδίπλωση πρωτεΐνης

- Στοχεύει στην πρόβλεψη της τριτοταγούς δομής με βάση φυσικοχημικές αρχές
  - Δεν βασίζεται στην διαπίστωση ομοιότητας με ακολουθίες γνωστής δομής
- Ενδιαφέρον επιστημονικό ερώτημα
- Προς το παρόν αναξιόπιστη μέθοδος για πρακτική χρήση

Ευστοχία της πρόβλεψης

- Συγκριτικός σχεδιασμός
  - Υψηλός βαθμός ευστοχίας όταν οι ακολουθίες πρότυπο και στόχος έχουν μεγάλη ομοιότητα
  - Μερικές φορές RMSD < 1.0 Angstrom (τετραγωνική ρίζα της μέσης τετραγωνικής απόκλισης μεταξύ των θέσεων κάθε ατόμου στην προβλεπόμενη και στην πραγματική δομή)

$$RMSD = \sqrt{\frac{\sum_{1}^{N} (r_i^A - r_i^B)^2}{N}}$$

Παράγοντες που επηρεάζουν τον βαθμό ευστοχίας

- Ποιότητα της στοίχισης μεταξύ της ακολουθίας «στόχος» και «πρότυπο»

  - Η προσαρμογή είναι ευκολότερη όταν οι ακολουθίες είναι πολύ όμοιες (seq ID > 80%).

  Οι καλύτερες μέθοδοι δίνουν μέση ευστοχία μόλις ~ 73% (% αμινοξέων που έχουν προβλεφθεί σωστά)

Μέθοδοι πρόβλεψης δευτεροταγούς δομής

Βασική ιδέα: Αποσπάσματα ακολουθιών από διαδοχικά αμινοξέα δείχνουν να προτιμούν συγκεκριμένες δευτεροταγείς δομές.

Φυσικοχημικές αρχές, συστήματα εμπειρογνωμόνων, θεωρία γραφημάτων, γραμμική και πολυγραμμική στατιστική, αλγόριθμοι εγγύτερων γειτόνων, μοριακή δυναμική και νευρωνικά δίκτυα → 60% ευστοχία

Αξιοποίηση μόνο τοπικής πληροφορίας, ο σχηματισμός β-ελασμάτων προέρχεται από μη τοπικές αλληλεπιδράσεις

Μέθοδοι πρόβλεψης δευτεροταγούς δομής

Διαπίστωση: 20% των ορθά προβλεφθέντων αμινοξέων ήταν σε κλώνους (strands), 30% σε έλικες και 50% σε μη κανονικές δομές.

Η διαπίστωση αυτή άλλαξε τις παραμέτρους εκπαίδευσης νευρωνικών δικτύων και βελτίωσε στο 60% τα αμινοξέα που ανήκουν σε strands και προβλέπονται σωστά.

Μέθοδοι πρόβλεψης δευτεροταγούς δομής

Διαπίστωση: 67% των αμινοξέων μπορούν να ανταλλαχθούν σε μια πρωτεΐνη χωρίς μεταβολή της δομής.

Διαπίστωση: Ανταλλαγή πολύ συγκεκριμένων αμινοξέων μπορεί να αποσταθεροποιήσει την δομή.

Εξελικτική πληροφορία: Πολλαπλή στοίχιση σε οικογένεια πρωτεϊνών δίνει πρότυπα ανταλλαγής αμινοξέων ενδεικτικά της δομής. Ένα προφίλ διαδοχικών αμινοξέων μιας στοίχισης περιέχει μη τοπική πληροφορία, αφού η εξέλιξη δουλεύει σε αντικείμενο 3D και όχι σε ακολουθία.

Πρόβλεψη δευτεροταγούς δομής

Διαπίστωση: 20% των ορθά προβλεφθέντων αμινοξέων ήταν σε β-κλώνους, 30% σε έλικες και 50% σε μη κανονικές δομές.

Η διαπίστωση αυτή άλλαξε τις παραμέτρους εκπαίδευσης νευρωνικών δικτύων και βελτίωσε στο 60% τα αμινοξέα που ανήκουν σε β-κλώνους και προβλέπονται σωστά.

Amino acid	Abbreviations		Hydropathy index (KD)	Hydropathy index
Alanine	Ala	A	1.8	1.6
Arginine	Arg	R	-4.5	-12.3
Asparagine	Asn	N	-3.5	-4.8
Aspartic acid	Asp	D	-3.5	-9.2
Cysteine	Cys	C	2.5	2
Glutamine	Gln	Q	-3.5	-4.1
Glutamic	Glu	E	-3.5	-8.2
Glycine	Gly	G	-0.4	1
Histidine	His	H	-3.2	-3
Isoleucine	Ile	I	4.5	3.1
Leucine	Leu	L	3.8	2.8
Lysine	Lys	K	-3.9	-8.8
Methionine	Met	M	1.9	3.4
Phenylalanine	Phe	F	2.8	3.7
Proline	Pro	P	-1.6	-0.2
Serine	Ser	S	-0.8	0.6
Threonine	Thr	T	-0.7	1.2
Tryptophan	Trp	W	-0.9	1.2
Tyrosine	Tyr	Y	-1.3	-0.7
Valine	Val	V	4.2	2.6

Table 4.2 Propensities of Amino Acids to Form α-Helices $\langle P_\alpha \rangle$ and β-Sheets $\langle P_\beta \rangle$

α-Residues	$\langle P_\alpha \rangle$	α-Assignment	β-Residues	$\langle P_\beta \rangle$	β-Assignment
Glu	1.44 ± 0.06	H_α	Val	1.64 ± 0.07	H_β
Ala	1.39 ± 0.05	H_α	Ile	1.57 ± 0.08	H_β
Met	1.32 ± 0.11	H_α	Thr	1.33 ± 0.07	h_β
Leu	1.30 ± 0.05	H_α	Tyr	1.31 ± 0.09	h_β
Lys	1.21 ± 0.05	h_α	Trp	1.24 ± 0.14	h_β
His	1.12 ± 0.08	h_α	Phe	1.23 ± 0.09	h_β
Gln	1.12 ± 0.07	h_α	Leu	1.17 ± 0.06	h_β
Phe	1.11 ± 0.07	h_α	Cys	1.07 ± 0.12	h_β
Asp	1.06 ± 0.06	h_α	Met	1.01 ± 0.13	I_β
Trp	1.03 ± 0.10	I_α	Gln	1.00 ± 0.09	I_β
Arg	1.00 ± 0.07	I_α	Ser	0.94 ± 0.06	i_β
Ile	0.99 ± 0.06	i_α	Arg	0.94 ± 0.09	i_β
Val	0.97 ± 0.05	i_α	Gly	0.87 ± 0.05	i_β
Cys	0.95 ± 0.09	i_α	His	0.83 ± 0.09	i_β
Thr	0.78 ± 0.05	i_α	Ala	0.79 ± 0.05	i_β
Asn	0.78 ± 0.06	i_α	Lys	0.73 ± 0.06	b_β
Tyr	0.73 ± 0.06	b_α	Asp	0.66 ± 0.06	b_β
Ser	0.72 ± 0.04	b_α	Asn	0.66 ± 0.06	b_β
Gly	0.63 ± 0.04	B_α	Pro	0.62 ± 0.07	B_β
Pro	0.55 ± 0.05	B_α	Glu	0.51 ± 0.06	B_β

Listed are values compiled from the crystal structures of 64 proteins, and the assignments as former (H and h), indifferent (I and i) and breakers (b and B) for each type of structure.

From P. Y. Chou (1989), in *Prediction of Protein Structure and the Principles of Protein Conformation*, ed. G. D. Fasman, 549–586, Plenum Press, New York.

Μέθοδοι πρόβλεψης δευτεροταγούς δομής

- PHD - Rost and Sander (artificial neural network)
- DSC - King and Sternberg (linear discriminant analysis)
- NNSSP -Salomov and Solevyev (nearest neighbour algorithm)
- PREDATOR - Frishman and Argos
  (Αναγνώριση πιθανών ζευγών καταλοίπων που συνδέονται με δεσμούς υδρογόνου)
- JPred2
- SSpro2

> 70% ευστοχία

Πόροι πρόβλεψης δομής

- Πρόβλεψη δευτεροταγούς δομής
  - Jpred (http://www.compbio.dundee.ac.uk/Software/JPred/jpred.html)
  - PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/
  - Και αρκετοί άλλοι στο WWW
- Συγκριτικός σχεδιασμός
  - SWISSMODEL (http://www.espasy.ch/swissmod/SWISS-MODEL.html)

Περιορισμός

- Οι μέθοδοι που αναφέρθηκαν αφορούν σε υδατοδιαλυτές πρωτεΐνες
  - Δεν γνωρίζουμε πολλές 3D δομές διαμεμβρανικών πρωτεϊνών

Συγκριτικός σχεδιασμός

Εύρεση κατάλληλου προτύπου (template)

- Blast την ακολουθία στόχο (target) στην PDB
- Επιλογή στοίχισης με την καλύτερη βαθμολογία
- Επιλογή στοίχισης με λιγότερα χάσματα
- Επιλογή δομών με καλύτερη διακριτική ικανότητα (resolution)
- Πολλαπλή στοίχισης συγγενών δομών
- Σύγκριση στοίχισης με πρόβλεψη δευτεροταγούς δομής

PQFHLWKRPVVTAHIEGQPVEVLLDTGAEDSIVTAIEIGVHYTPKIVGGIGGFINTKEYK
NVEVEVLGKRIKGTIMTGNTPMNIFGRNLLTALGMSLNF

https://blast.ncbi.nlm.nih.gov/Blast.cgi

Συγκριτικός σχεδιασμός

Προσαρμογή στόχου στο πρότυπο

- Ενδεχομένως διόρθωση στοίχισης
- Χρήση προγραμμάτων για την προσαρμογή (fit) των μερών με καλή στοίχιση στα αντίστοιχα πρότυπα.

Σχεδιασμός των αναστροφών

Αξιολόγηση του μοντέλου

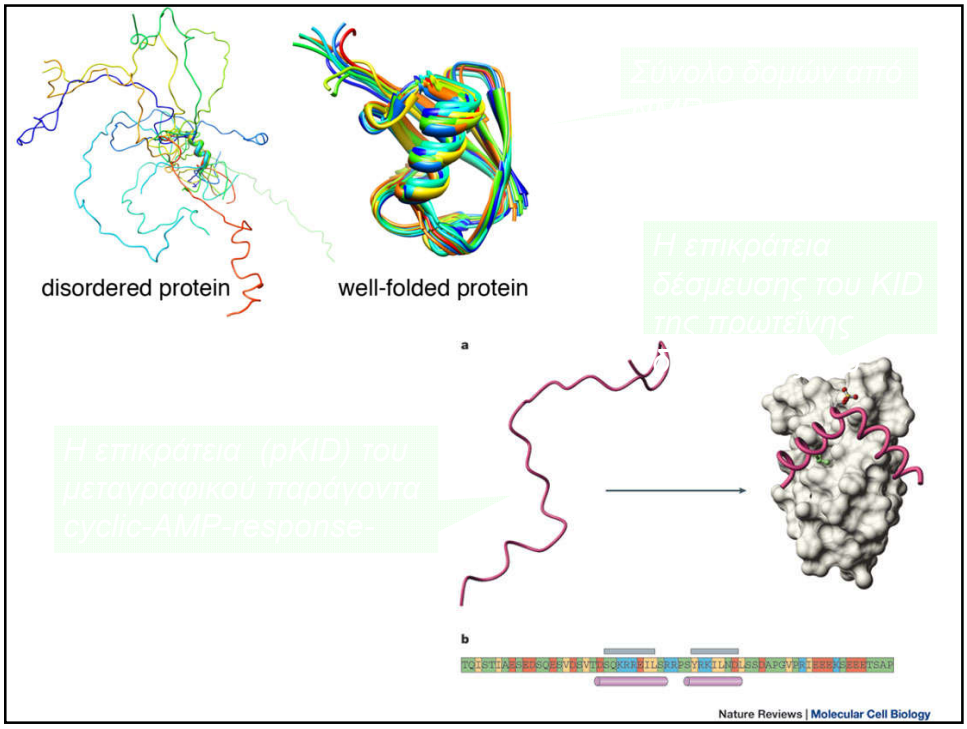
- Ramachandran plot: Ένα καλό μοντέλο έχει > 90% των καταλοίπων του στα πλαίσια των στενά επιτρεπτών περιοχών του Ramachandran plot και >98% στις ευρύτερα επιτρεπτές.
- ProQ: http://www.sbc.su.se/~bjornw/ProQ/ProQ.cgi

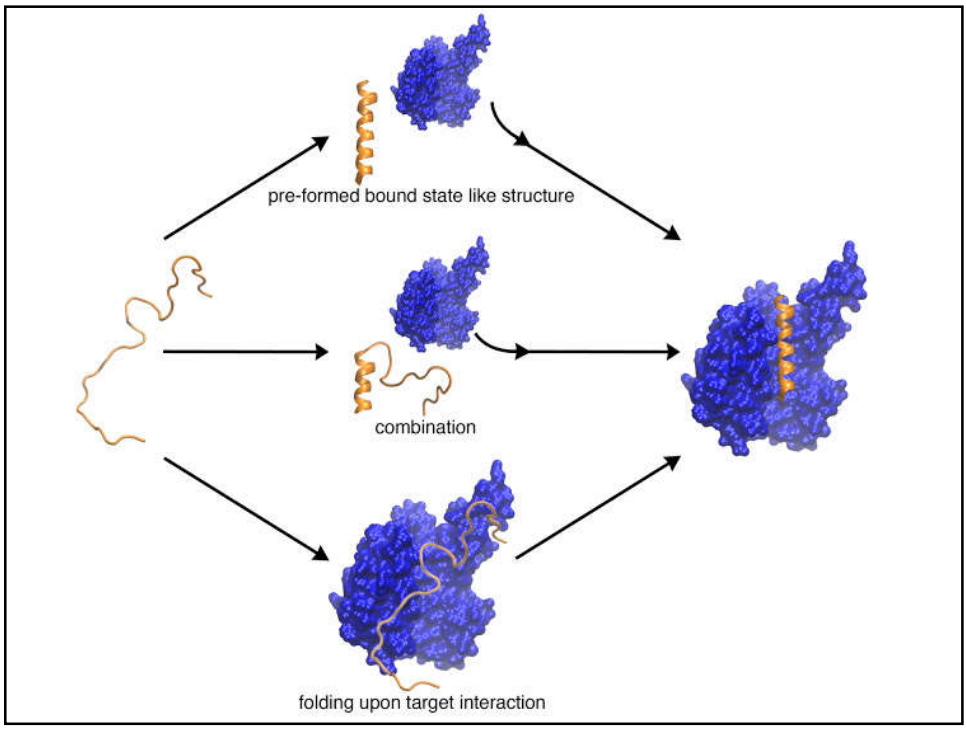
Ενδογενώς αδόμητες πρωτεΐνες

(intrinsically unstructured proteins ή naturally unfolded proteins or disordered proteins)

Εκτιμώνται ≈30% των ευκαρυωτικών πρωτεϊνών

- Σε αδέσμευτη κατάσταση και σε φυσιολογικές συνθήκες δεν διαθέτουν σταθερή τριτοταγή διαμόρφωση.
- Μερικές αποκτούν σταθερή δομή μετά από πρόσδεση σε άλλο μακρομόριο.
- Πολλές ρυθμίζουν την συγγένεια δέσμευσης με μετα-μεταφραστικές τροποποιήσεις.
- Συχνή εμφάνιση σε διακυτταρική επικοινωνία, μεταγραφή και σε λειτουργίες ανασχεδιασμού της χρωματίνης


disordered protein
well-folded protein
a
b
Nature Reviews | Molecular Cell Biology


pre-formed bound state like structure
combination
folding upon target interaction

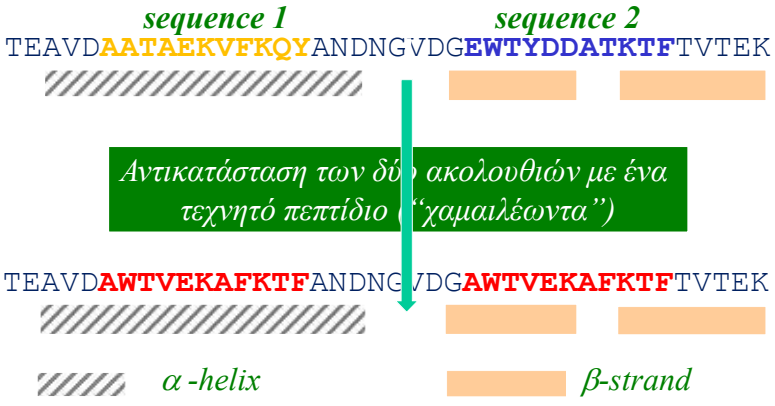
Πειραματική ταυτοποίηση

- *Οι αναδιπλωμένες πρωτεΐνες έχουν υψηλή πυκνότητα (ειδικό όγκο = Όγκος / MB , 0.72-0.74 mL/g) και ανάλογα μικρή γυροσκοπική ακτίνα* $R_g^2 = \frac{\sum_i m_i (r_i - r_c)^2}{M}$ *Άρα ανιχνεύονται με μεθόδους ευαίσθητες στο μοριακό μέγεθος, στην πυκνότητα ή στην υδροδυναμική αντίσταση, όπως **size exclusion chromatography, analytical ultracentrifugation, Small angle X-ray scattering (SAXS)** καθώς και μετρήσεις του συντελεστή διάχυσης.*
- *Επειδή χαρακτηρίζονται από έλλειψη δευτεροταγούς δομής μπορούν να ανιχνευθούν με **CD** στο άπω υπεριώδες (170-250 nm) ή με φασματοσκοπία **IR**.*
- *Εμφανίζουν εκτεθειμένες στον διαλύτη πεπτιδικές ομάδες του σκελετού και έτσι προσβάλλονται εύκολα από πρωτεάσες.*
- *Επιτρέπουν γρήγορη ανταλλαγή H→D.*
- *Στο **NMR** εμφανίζουν μικρή διασπορά (<1 ppm) στην χημική μετατόπιση των αμιδικών πρωτονίων, ενώ οι αναδιπλωμένες πρωτεΐνες μέχρι και 5 ppm. Το NMR αποτελεί την κύρια μέθοδο άντλησης πληροφοριών για αδόμητες περιοχές μιας πρωτεΐνης.*
- *Η απουσία ηλεκτρονιακής πυκνότητας στην **κρυσταλλογραφία ακτίνων X** μπορεί επίσης να είναι σημάδι απουσίας δομής.*

ΠΡΟΒΛΗΜΑ

Η ακολουθία "Χαμαιλέων"

The "Chameleon" sequence



Πηγή: *Minor and Kim. 1996. Nature 380:730-734*