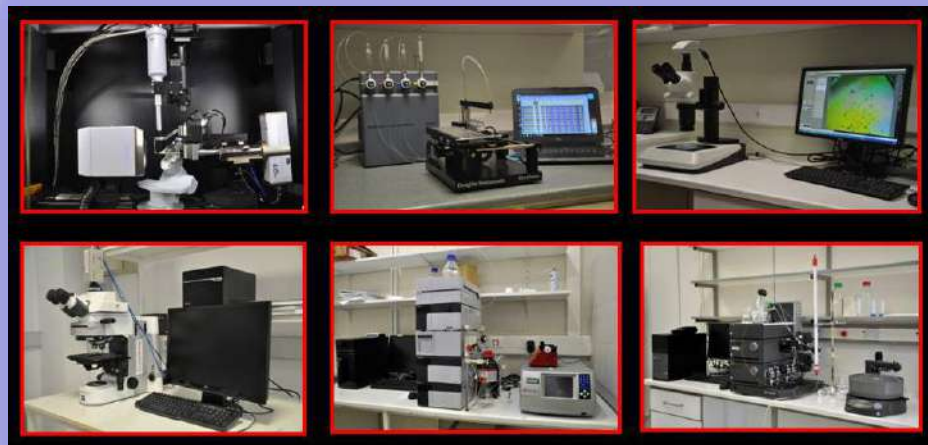


Μ. ΚΟΚΚΙΝΙΔΗΣ

ΔΟΜΙΚΗ ΒΙΟΛΟΓΙΑ

ΠΑΝ/ΜΙΟ ΚΡΗΤΗΣ, Τμ. Βιολογίας & ΙΜΒΒ/ΙΤΕ

30-3-2024 ΑΛΕΞΑΝΔΡΟΥΠΟΛΗ



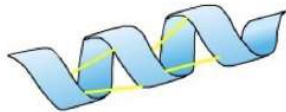
THE PROTEIN FOLDING PROBLEM (ΤΟ ΠΡΟΒΛΗΜΑ ΑΝΑΔΙΠΛΩΣΗΣ ΤΩΝ ΠΡΩΤΕΙΝΩΝ)



(a)

ΜΗΧΑΝΙΣΜΟΙ

Secondary Structure:

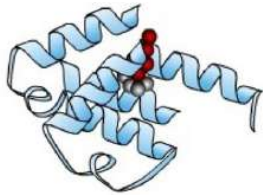


(b)

ΠΡΟΒΛΕΨΗ/ ΥΠΟΛΟΓΙΣΜΟΙ

ΠΡΩΤΕΪΝΙΚΩΝ ΔΟΜΩΝ

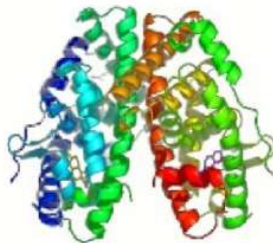
Tertiary Structure:



(c)

ΕΦΑΡΜΟΓΕΣ

Quaternary Structure:



(d)

(Protein structural dynamics)

Hierarchy of protein structure

128

R. JAENICKE

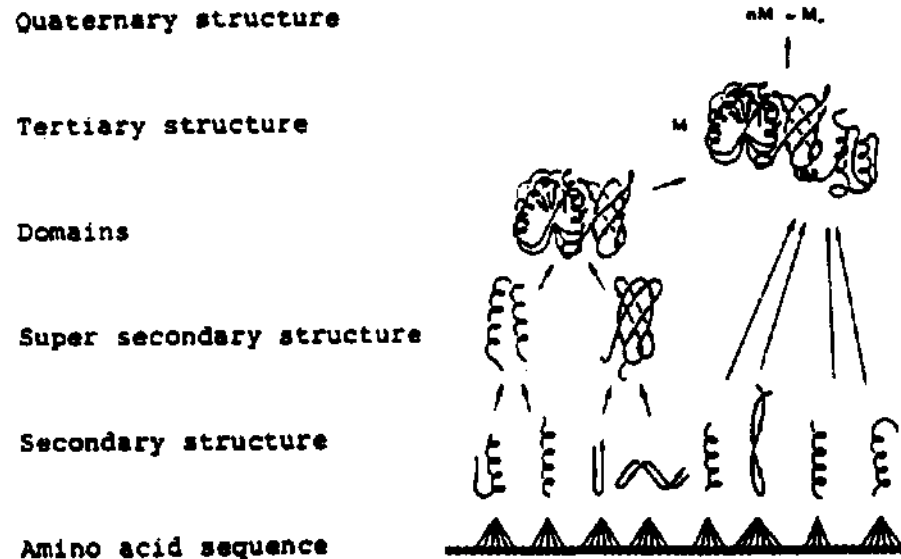
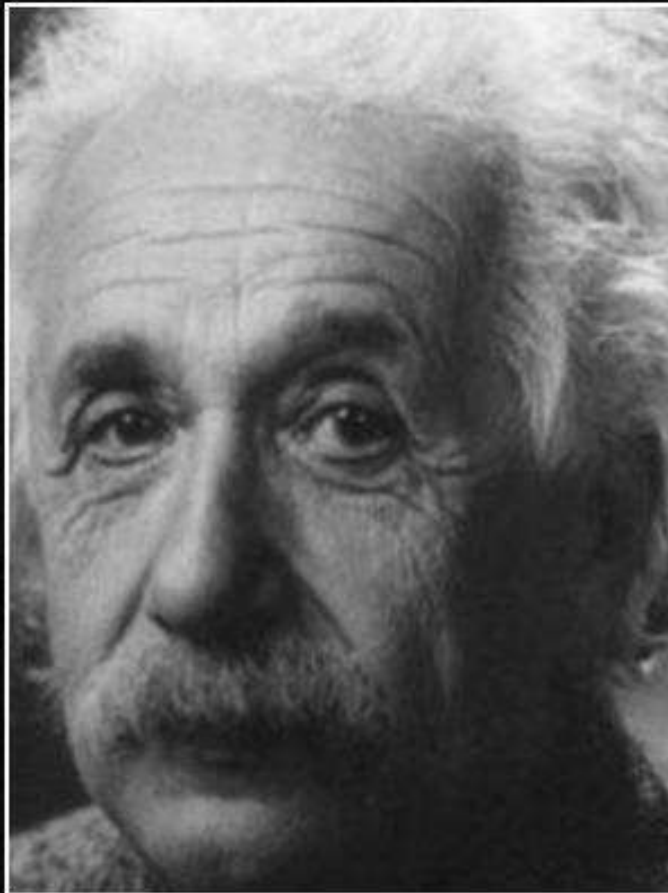


FIG. 3. Hierarchy of protein structure and protein folding. The one-dimensional primary structure (●●●●) determines the folding and association of globular proteins. Next-neighbor (short-range) interactions at the secondary structural level are supplemented by long-range interactions causing docking and merging of structural motifs and domains (supersecondary structure and tertiary structure).



A clever person solves a problem. A
wise person avoids it.

— *Albert Einstein* —

AZ QUOTES

ΑΝΑΔΙΠΛΩΣΗ ΤΩΝ ΠΡΩΤΕΙΝΩΝ & Ο “ΚΩΔΙΚΑΣ” ΤΗΣ ΑΝΑΔΙΠΛΩΣΗΣ

- Anfinsen, δεκαετία 1960's: Η πολύπλοκη **τριδιάστατη δομή** των πρωτεϊνικών μορίων **κωδικοποιείται αποκλειστικά στις αλληλουχίες των αμινοξέων** τους και οι πολυπεπτιδικές αλυσίδες **διπλώνουν** αυτόνομα.
- <https://youtu.be/pZee0XCCqH4>
- Η διαλεύκανση του “**κώδικα**” που ελέγχει την διαδικασία αναδίπλωσης – “**δεύτερο μέρος του γενετικού κώδικα**”-, ήταν και είναι μια από τις μεγαλύτερες προκλήσεις της μοριακής βιολογίας.



ΤΟ ΠΕΔΙΟ ΤΗΣ ΑΝΑΔΙΠΛΩΣΗΣ ΤΩΝ ΠΡΩΤΕΙΝΩΝ ΣΗΜΕΡΑ

- Πολυάριθμες θεωρητικές/υπολογιστικές και πειραματικές μελέτες της αναδίπλωσης πρωτεϊνών έχουν προχωρήσει σημαντικά πιο πέρα από τα ευρήματα του Anfinsen.
- Στο ζωντανό κύτταρο, η αναδίπλωση συμβαίνει σε ένα σύνθετο περιβάλλον, και σε ορισμένες περιπτώσεις μπορεί να αποτύχει οδηγώντας σε λάθος αναδίπλωση, ή σε σχηματισμό συσσωματωμάτων ή αμυλοειδών ινών. Οι περιπτώσεις αυτές σχετίζονται με μια σειρά σοβαρών ασθενειών, πχ αρκετών νευροεκφυλιστικών ασθενειών.
- Η μελέτη της αναδίπλωσης πρωτεϊνών αποτελεί βασικό τομέα της μοριακής βιολογίας, της βιοϊατρικής έρευνας και της βιοτεχνολογίας.

Important issues (1): Folding in vivo vs. folding in vitro

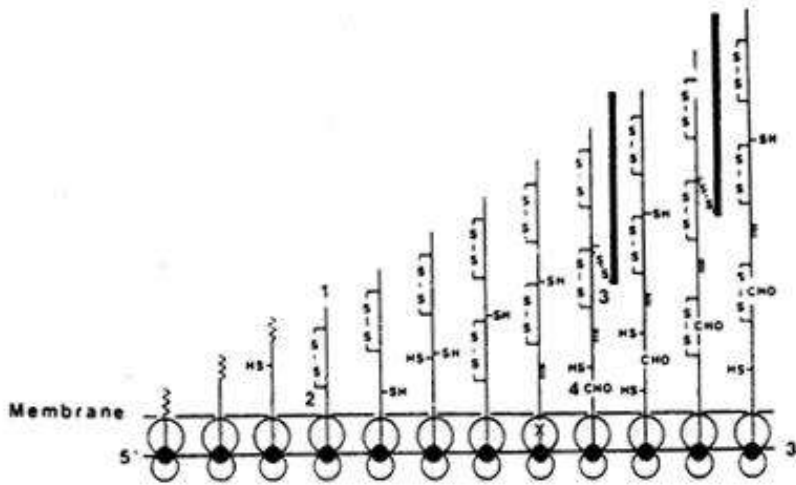


FIG. 2. Co-translational folding of the nascent IgG heavy chains in mouse myeloma cells (MPC 11) indicating co-translational modification events. (1) cleavage of signal peptide (wavy); (2) formation of intrachain disulfide bonds; (3) formation of interchain cysteine bridges with light chain (indicated by thick line); (4) transfer of core oligosaccharide (CHO) to Asn-acceptor (X). About half of the nascent heavy chain forms an interchain disulfide bond with a complete light chain before heavy chain completion and release from the polysome (from Bergman and Kuehl, 1979c).

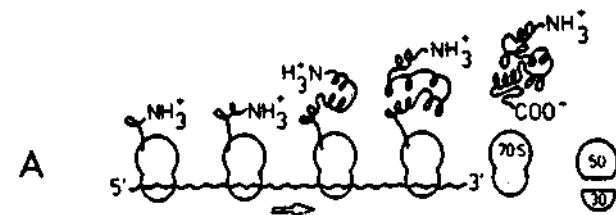


FIG. 1. Folding *in vivo* and folding *in vitro*. A. Ribosomes moving along the mRNA in 5' → 3' direction release the growing polypeptide chain. Folding may occur either co-translationally (i.e. as a "vectorial" process, from the N- to the C-terminus) or post-translationally, as in B. B. Unfolding and refolding of a single-chain, one domain protein demonstrating the non-vectorial character of *in vitro* reconstitution.

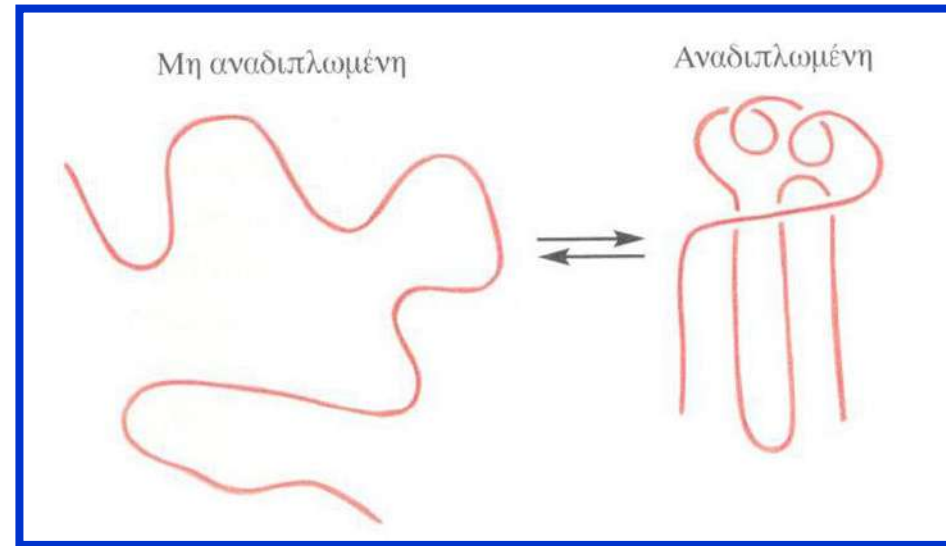
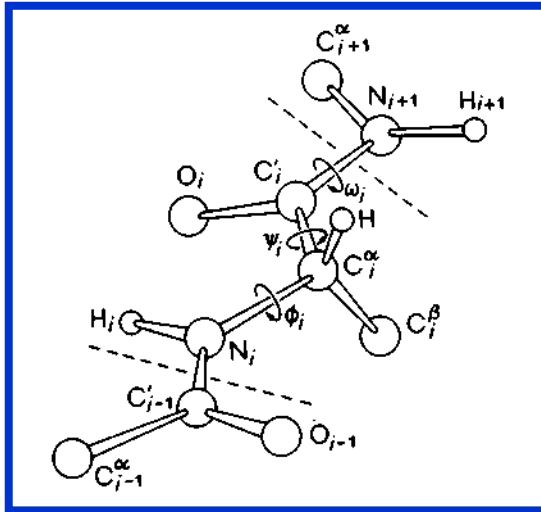
Protein Folding vs. Protein Association

- Folding: a) Spontaneous Acquisition of 3D-structure and the capacity to form higher-order structures b) Spatial arrangement of polypeptide chain backbone
- Association: Formation of stoichiometrically and spatially well defined quaternary structure of oligomeric & multimeric proteins

Important issues (summary)

- Thermodynamic vs. kinetic control of folding
- Unique protein structure vs. a dynamic, fluctuating system (breathing motions etc)
- Role of water & amino acid properties

ΟΙ ΣΦΑΙΡΙΚΕΣ ΠΡΩΤΕΪΝΕΣ ΕΙΝΑΙ ΟΡΙΑΚΑ ΣΤΑΘΕΡΕΣ



Folded & unfolded proteins:

The polypeptide has
considerable conformational
flexibility

Requirements for the folded state

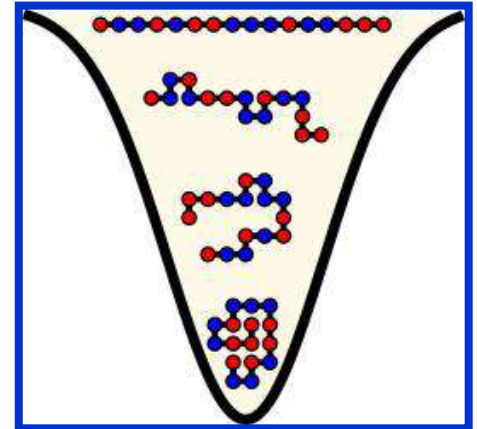
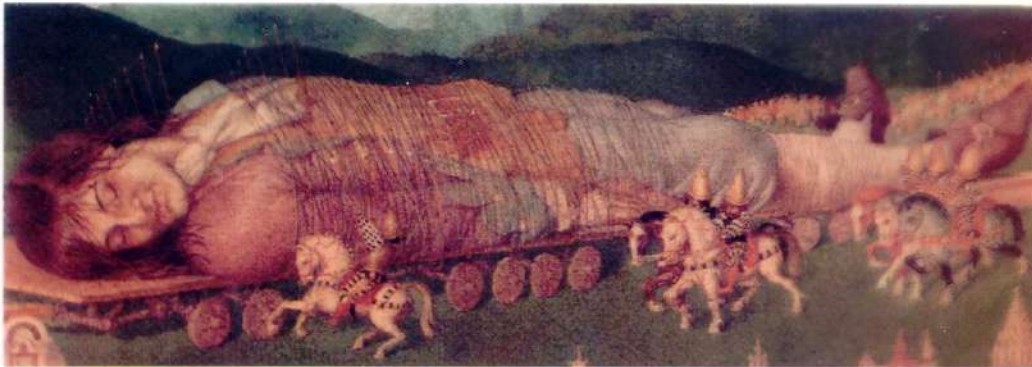
- TO MONTEΛO TΩN 2 KATACTACEΩN (2-STATE-MODEL)

N
(native)



U
(unfolded)

$$\Delta G \cong 15 \text{ kcal/mol}$$
$$E_N \cong E_U \cong 10^7 \text{ kcal/mol}$$

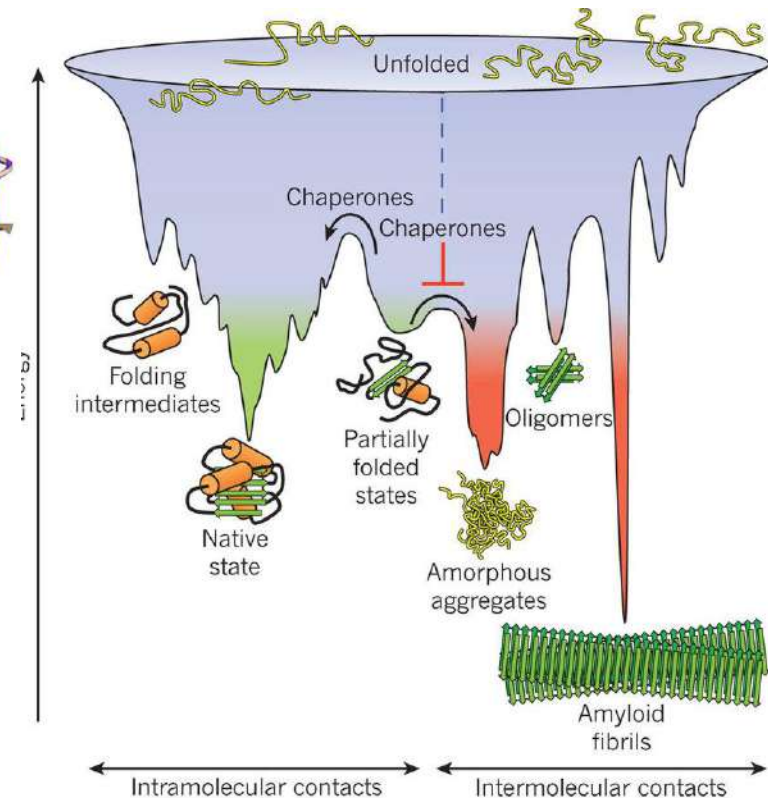
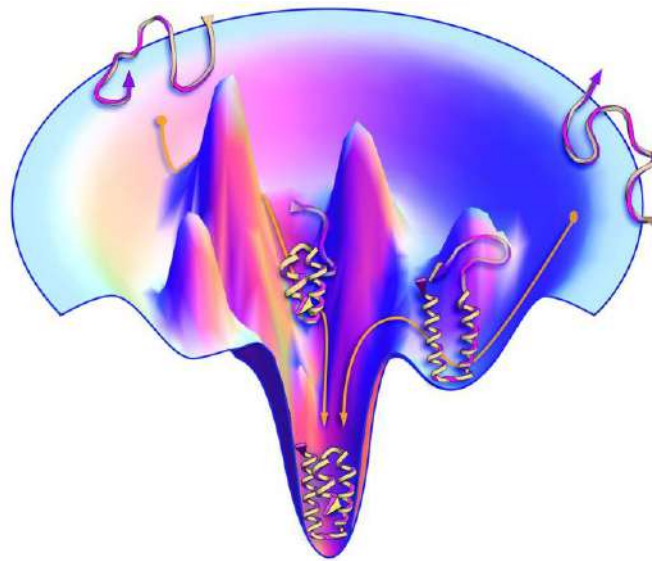


Η κατάσταση U ευνοείται από την τεράστια εντροπία διαμόρφωσης της. Η κατάσταση N ευνοείται από ένα πολύ μεγάλο αριθμό ασθενών αλληλεπιδράσεων που δρουν όμως ταυτόχρονα και συνεργατικά.

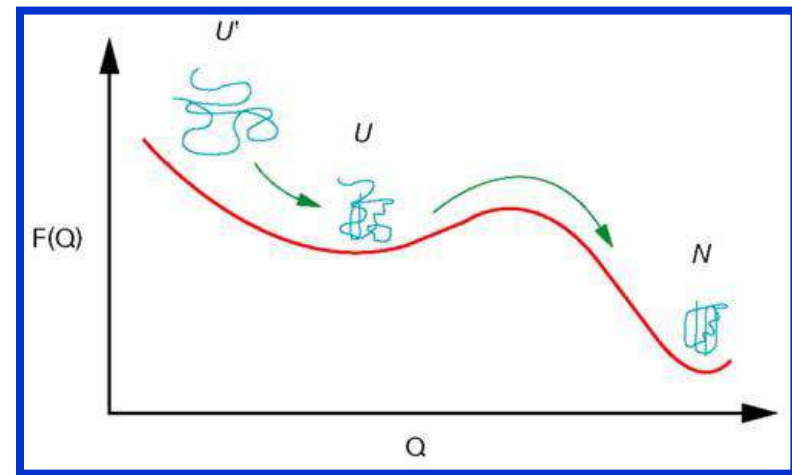
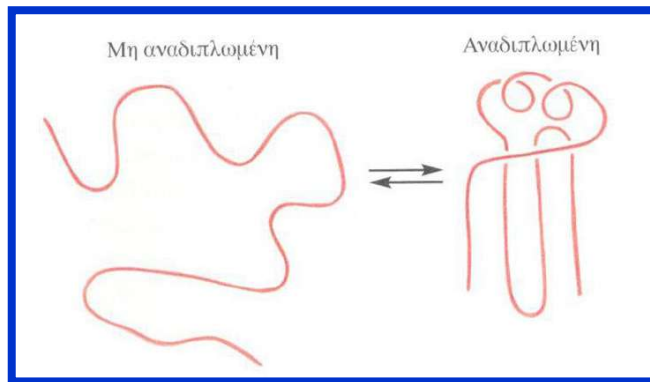
ΕΝΕΡΓΕΙΑΚΟΙ ΠΑΡΑΓΟΝΤΕΣ ΣΤΗΝ ΑΝΑΔΙΠΛΩΣΗ

Η χαμηλότερη
ελεύθερη
ενέργεια
αφορά την
native
διαμόρφωση?

Προκύπτουν
λάθος
συμπεράσματα
από
ενεργειακούς
υπολογισμούς?



ΚΙΝΗΤΙΚΟΙ ΠΑΡΑΓΟΝΤΕΣ ΕΙΝΑΙ ΣΗΜΑΝΤΙΚΟΙ ΓΙΑ ΤΗΝ ΑΝΑΔΙΠΛΩΣΗ



Cyrus Levinthal estimation: 10^{48} yrs for the folding process to search all possible conformations for a 150 aa protein with 1 ps steps (age of the universe 13.7×10^9 yrs!).

→ **folding pathways** (μονοπάτια αναδίπλωσης) ?

Requirements for folded state

- For a protein to fold, the folded state must be kinetically accessible and have lower free energy than the unfolded state. Free energies are determined by **all the physical interactions** that take place within the molecule and the solvent **plus entropic considerations**.

EXPERIMENTAL APPROACHES TO THE FOLDING AND ASSOCIATION OF PROTEINS

Equilibrium measurements

State of association

Electron microscopy, ultracentrifugation, (elastic and inelastic) light scattering, gel permeation chromatography, SDS-polyacrylamide gel electrophoresis (with and without cross-linking)

Conformation

Spectroscopy (absorption, fluorescence, optical rotatory dispersion, circular dichroism, nuclear magnetic resonance), hydrogen-deuterium-(tritium) exchange, stability towards denaturation or proteolysis, binding of ligands (coenzymes, substrates, etc.),* chemical modification ("group specific labels")

Function (activity)

Enzymatic assays, ligand binding (affinity chromatography)

Kinetic measurements†

Assembly (association)

Turbidity, light scattering, chemical cross-linking, hybridization

Folding

Spectroscopy (absorption, fluorescence, circular dichroism), hydrogen-deuterium exchange, limited proteolysis (fragment analysis using gel electrophoresis), ligand binding (antibodies, allosteric effectors, etc.)*

Function (activity)

Enzymatic assays, ligand binding (coenzymes, substrates)*

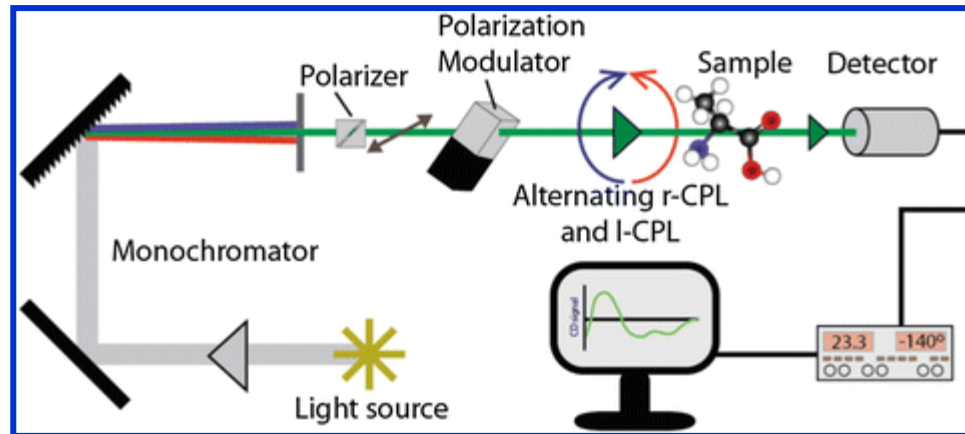
* Ligand binding may cause artifacts by shifting equilibria or stabilizing intermediates.

† Depending on the time range, methods include manual mixing, stopped flow, quench stopped flow (double jump), relaxation techniques (temperature jump, pressure jump, etc).

Stability data and association enthalpies may be deduced directly from calorimetry.

Frequently refolding studies are very informative; agreement of thermodynamic/kinetic data?

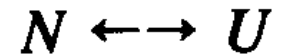
Refolding experiments



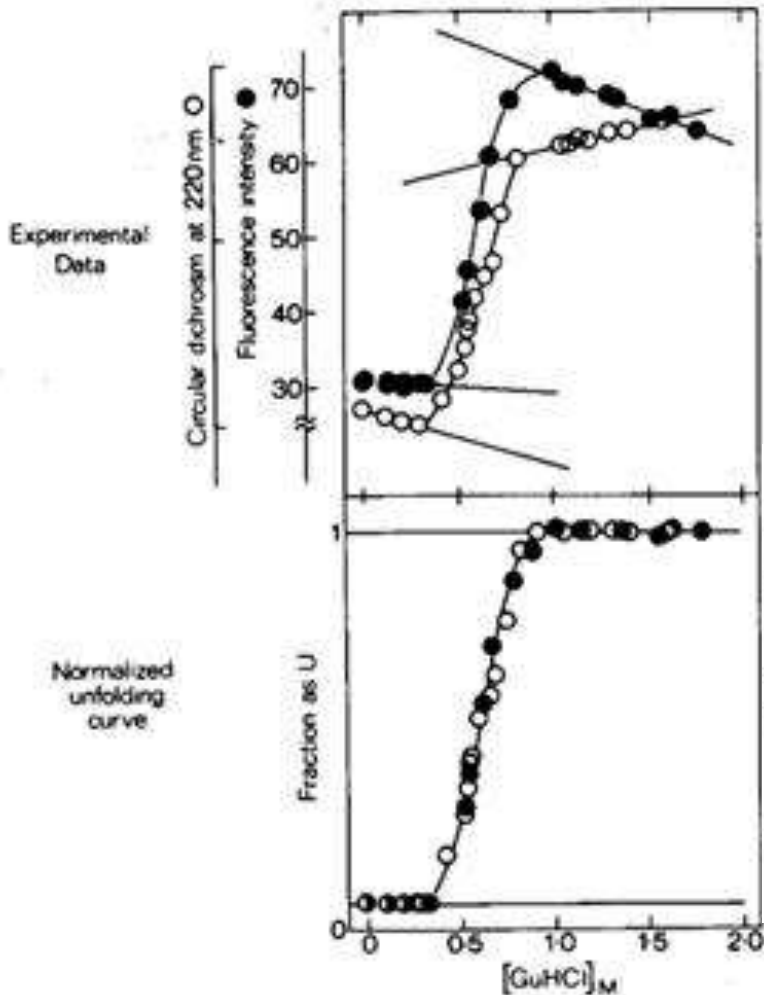
Circular Dichroism

Denaturants & folding by altering the environment

Guanidinium chloride, urea: they solvate almost equally all parts of the protein



$$f_u = \frac{\theta_N - \theta}{\theta_N - \theta_U}$$



The GuHCl-induced unfolding transition of yeast phosphoglycerate kinase detected by the fluorescence intensity of 340 nm (\bullet) and circular dichroism molar ellipticity at 220 nm (\circ). The experimental data are shown at the top; the straight lines show the effects of GuHCl on the spectral properties of the folded state at low concentrations, and on the unfolded state at high concentrations. The same effects are assumed throughout the transition region. Correcting for this, the fraction of unfolding indicated by the two spectral measurements is plotted in the lower half. The two curves coincide and are consistent with a two-state equilibrium unfolding transition. (From Nojima et al. [19].)

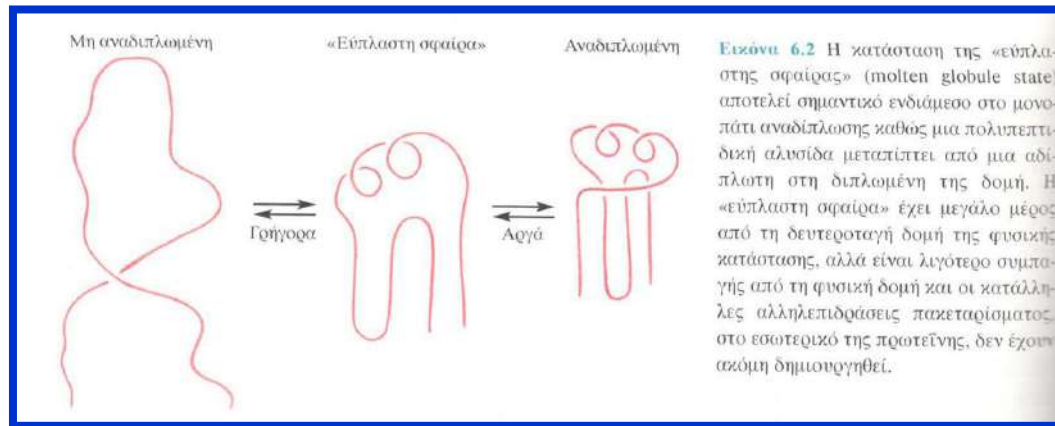
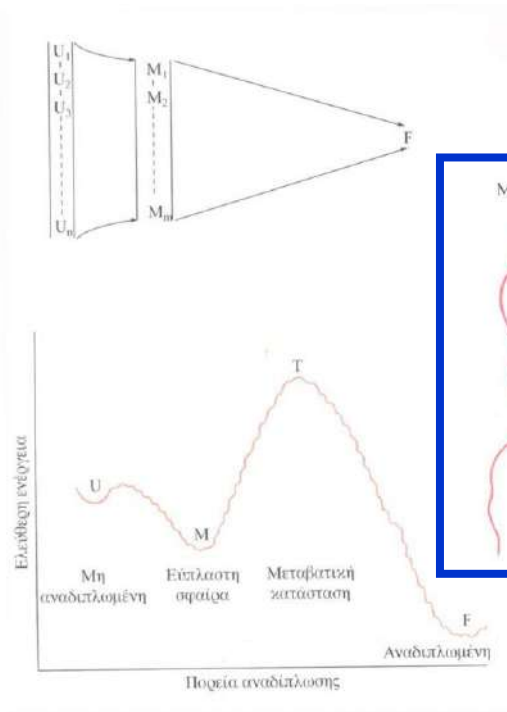
Energetics of folding

$$K_U = \frac{[U]}{[N]} = \frac{f_u}{1-f_u}$$

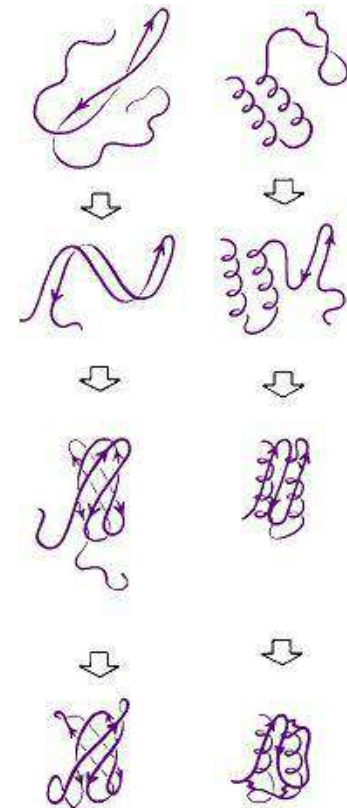
$$\Delta G_f = G_N - G_U = RT \ln K_U$$

Highest stability (lowest ΔG) at about 5°C; pH is important for the stability of the folded state because most proteins are ionized.

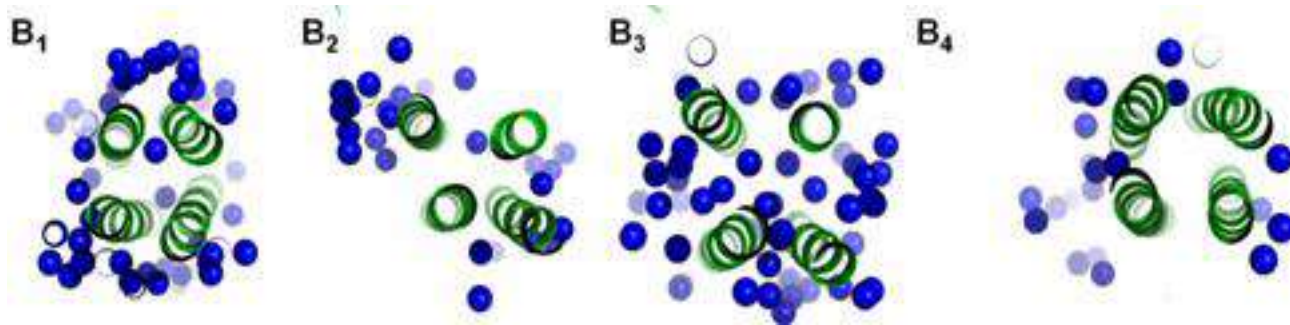
ΟΙ ΕΥΠΛΑΣΤΕΣ ΣΦΑΙΡΕΣ (MOLTEN GLOBULES) ΕΙΝΑΙ ΕΝΔΙΑΜΕΣΑ ΤΗΣ ΠΟΡΕΙΑΣ ΑΝΑΔΙΠΛΩΣΗΣ



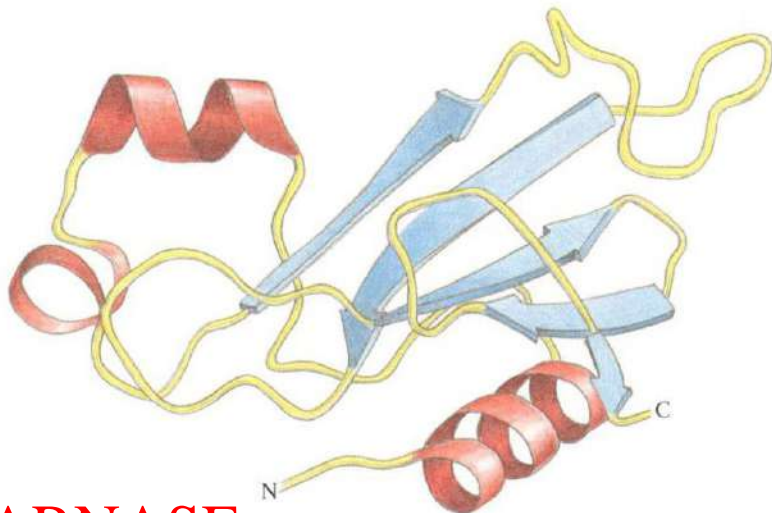
Εικόνα 6.2 Η κατάσταση της «εύπλαστης σφαίρας» (molten globule state) αποτελεί σημαντικό ενδιάμεσο στο μονοπάτι αναδίπλωσης καθώς μια πολυπεπτιδική αλυσίδα μεταπίπτει από μια αδιάπλωτη στη διπλωμένη της δομή. Η «εύπλαστη σφαίρα» έχει μεγάλο μέρος από τη δευτεροταγή δομή της φυσικής κατάστασης, αλλά είναι λιγότερο συμπαγής από τη φυσική δομή και οι κατάλληλες αλληλεπιδράσεις πακεταρίσματος, στο εσωτερικό της πρωτεΐνης, δεν έχουν ακόμη δημιουργηθεί.



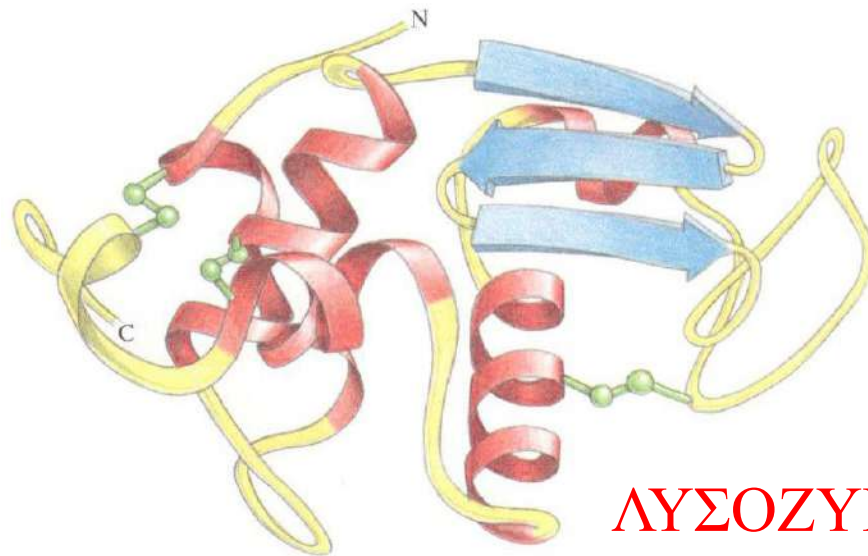
ΤΟ “ΘΑΨΙΜΟ” ΤΩΝ ΥΔΡΟΦΟΒΙΚΩΝ ΟΜΑΔΩΝ ΕΙΝΑΙ ΚΡΙΣΙΜΟ ΒΗΜΑ ΣΤΗΝ ΑΝΑΔΙΠΛΩΣΗ



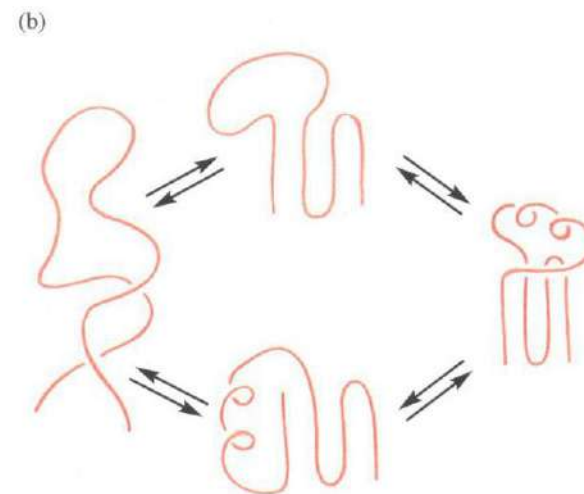
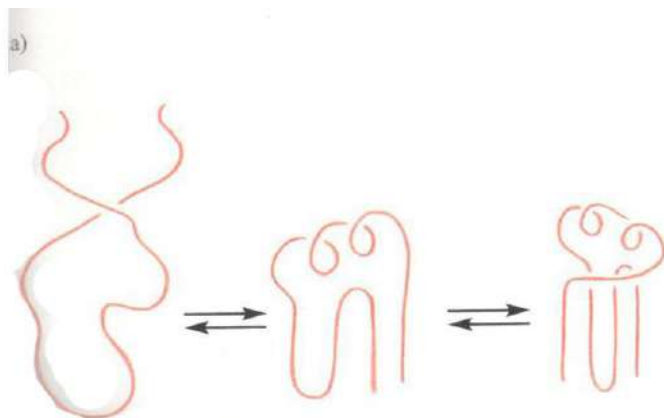
ΑΠΛΑ & ΠΟΛΛΑΠΛΑ ΜΟΝΟΠΑΤΙΑ ΑΝΑΔΙΠΛΩΣΗΣ



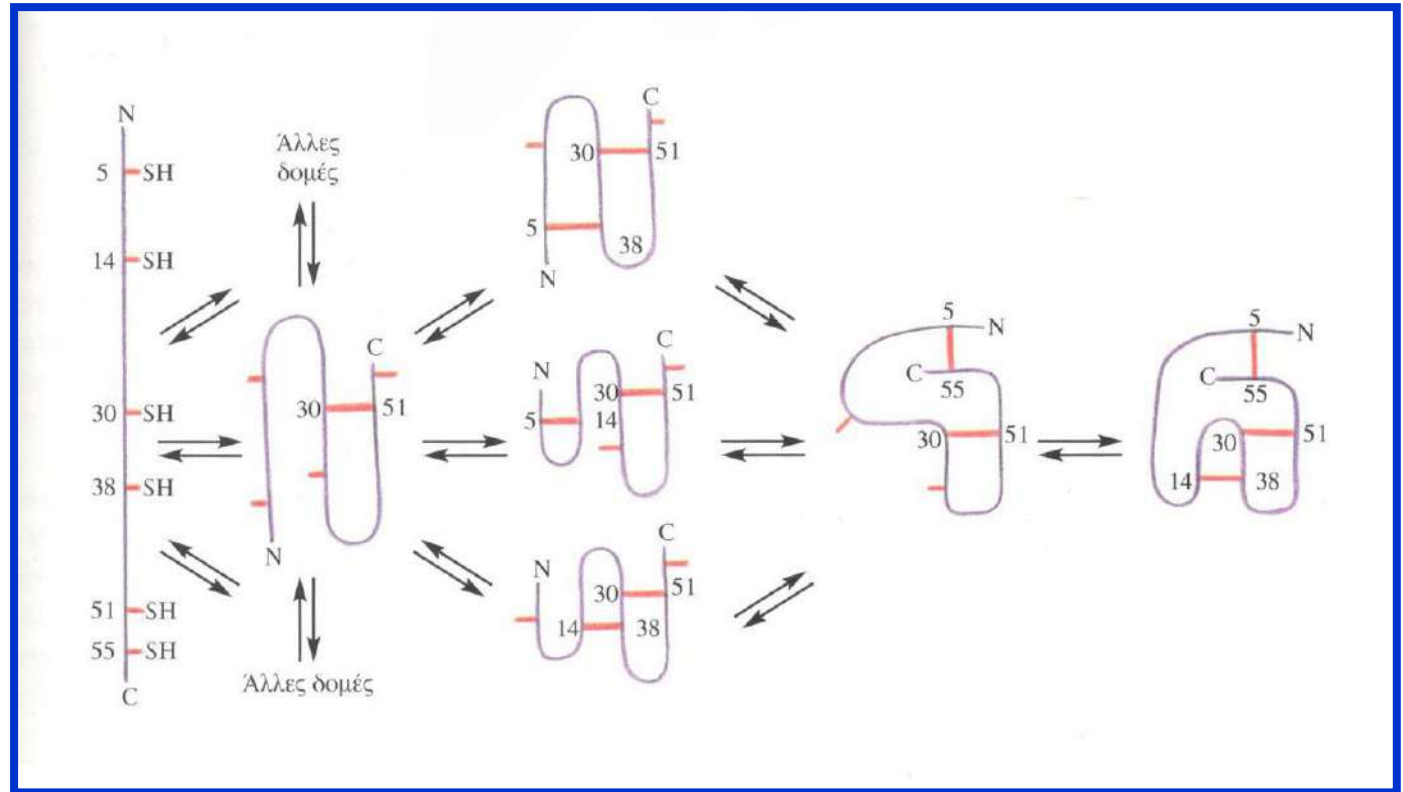
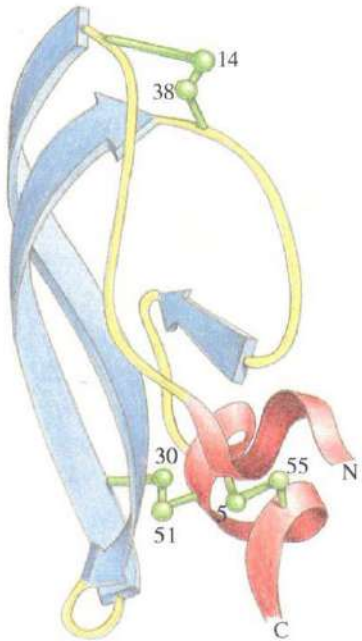
BARNASE



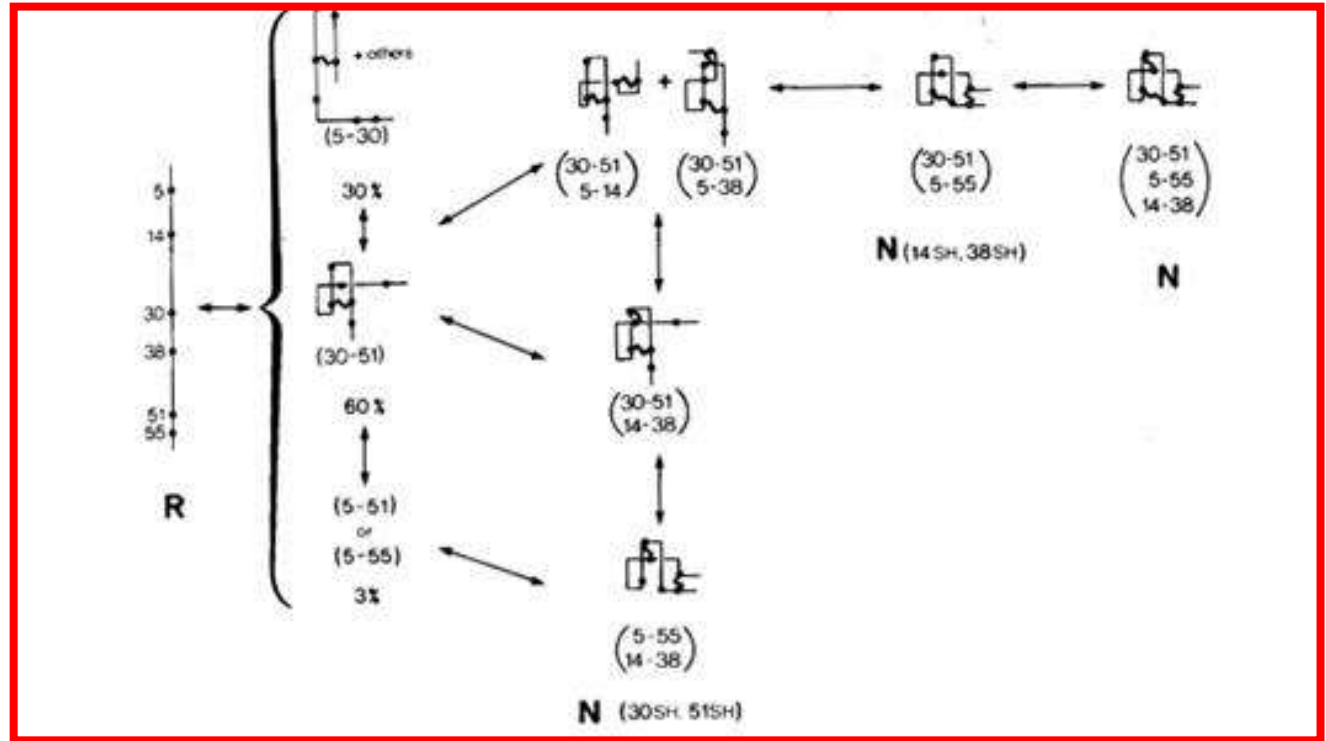
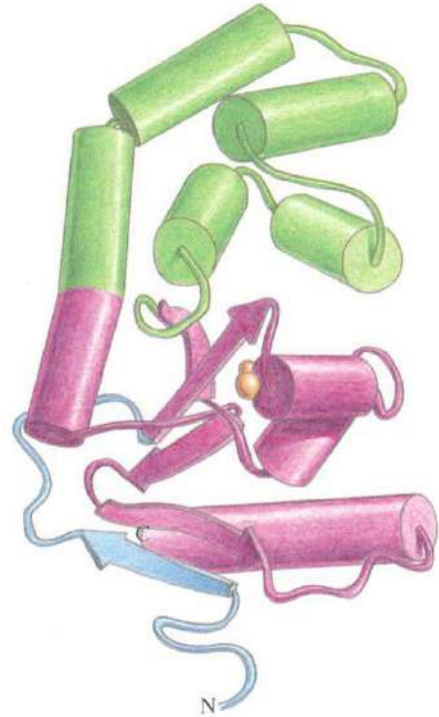
ΛΥΣΟΖΥΜΗ



ΣΧΗΜΑΤΙΣΜΟΣ ΣΩΣΤΩΝ S-S ΔΕΣΜΩΝ ΚΑΤΑ ΤΗΝ ΔΙΑΔΙΚΑΣΙΑ ΑΝΑΔΙΠΛΩΣΗΣ

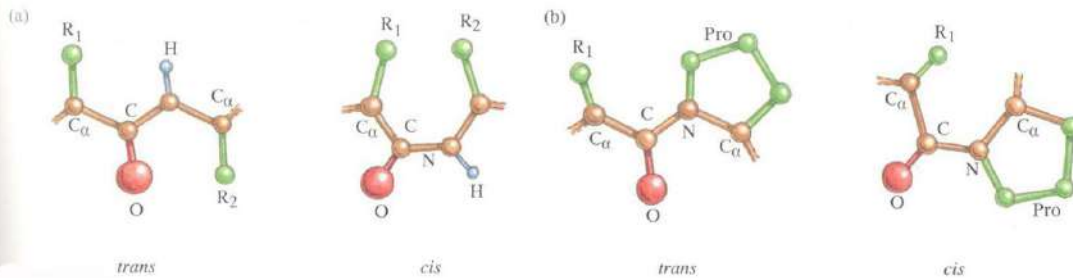


ΟΡΙΣΜΕΝΑ ΕΝΖΥΜΑ ΥΠΟΒΟΗΘΟΥΝ ΤΟΝ ΣΧΗΜΑΤΙΣΜΟ ΣΩΣΤΩΝ S-S ΔΕΣΜΩΝ ΚΑΤΑ ΤΗΝ ΑΝΑΔΙΠΛΩΣΗ



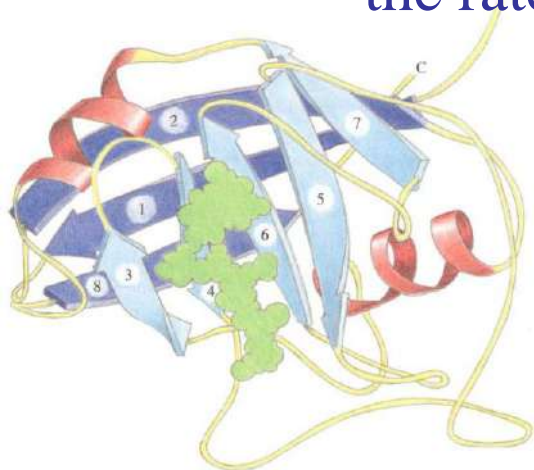
Σχηματικό διάγραμμα του ενζύμου DsbA, το οποίο καταλύει τη δημιουργία και την επαναδιευθέτηση δισουλφιδικών δεσμών. Το ένζυμο αναδιπλώνεται σε δύο επικράτειες, μία επικράτεια που περιλαμβάνει πέντε α-έλικες (πράσινο) και μία δεύτερη που έχει δομή παρόμοια με αυτήν της θειορεδοξίνης (μνεξεδί). Η αμινοτελική επέκταση (μπλε) δεν είναι παρούσα στη θειορεδοξίνη. (Προσαρμοσμένη από τους J. L. Martin et al., *Nature* 365: 464-468, 1993.)

ΙΣΟΜΕΡΙΩΣΗ ΚΑΤΑΛΟΙΠΩΝ ΠΡΟΛΙΝΗΣ ΚΑΙ ΤΑΧΥΤΗΤΑ ΠΡΩΤΕΙΝΙΚΗΣ ΑΝΑΔΙΠΛΩΣΗΣ

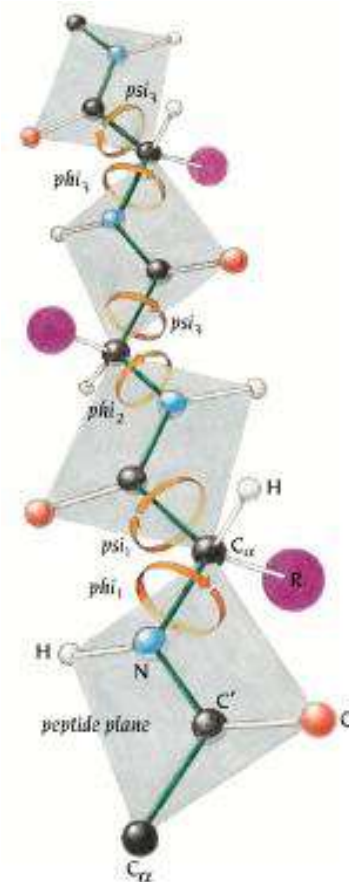


(a) Οι πεπτιδικές ομάδες μπορούν να υιοθετήσουν δύο διαφορετικές στερεοδιατάξεις, την *trans* και τη *cis*. Στην *trans*-μορφή οι ομάδες C=O και N-H διευθετούνται προς αντίθετες κατευθύνσεις, ενώ στη *cis*-μορφή διευθετούνται προς την ίδια κατεύθυνση. Για τα περισσότερα πεπτίδια η *trans*-μορφή είναι περίπου 1.000 φορές πιο σταθερή από τη *cis*-μορφή. (b) Όταν το δεύτερο κατόλοιπο σε ένα πεπτίδιο είναι η προλίνη, η *trans*-μορφή είναι μόνο περίπου τέσσερις φορές πιο σταθερή από τη *cis*-μορφή. Πεπτίδια με *cis*-προλίνη συναντώνται σε πολλές πρωτεΐνες.

Isomerization of proline residues can be the rate-limiting step in protein folding



ΚΥΚΛΟΦΙΛΙΝΗ (ΠΡΟΠΥΛΟ-ΠΕΠΤΙΔΙΚΗ ΙΣΟΜΕΡΑΣΗ)



© 1999 GARLAND PUBLISHING INC.
A member of the Taylor & Francis Group

Pro-isomerization

- The $U \leftrightarrow N$ equilibration fits a 2 state-model on thermodynamic criteria, while the kinetics of the $U \rightarrow N$ transition show higher complexity requiring more than two species

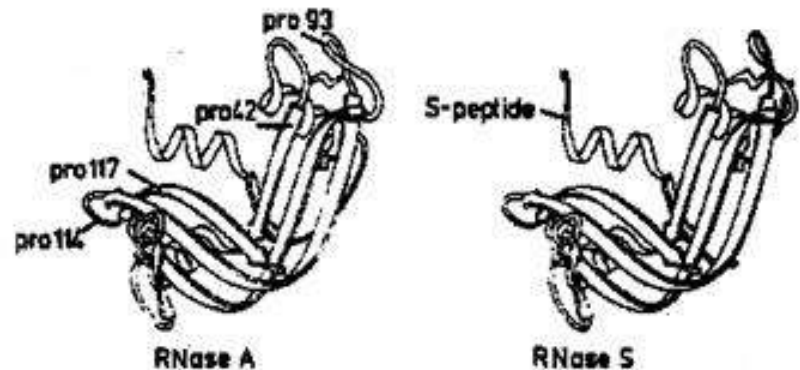
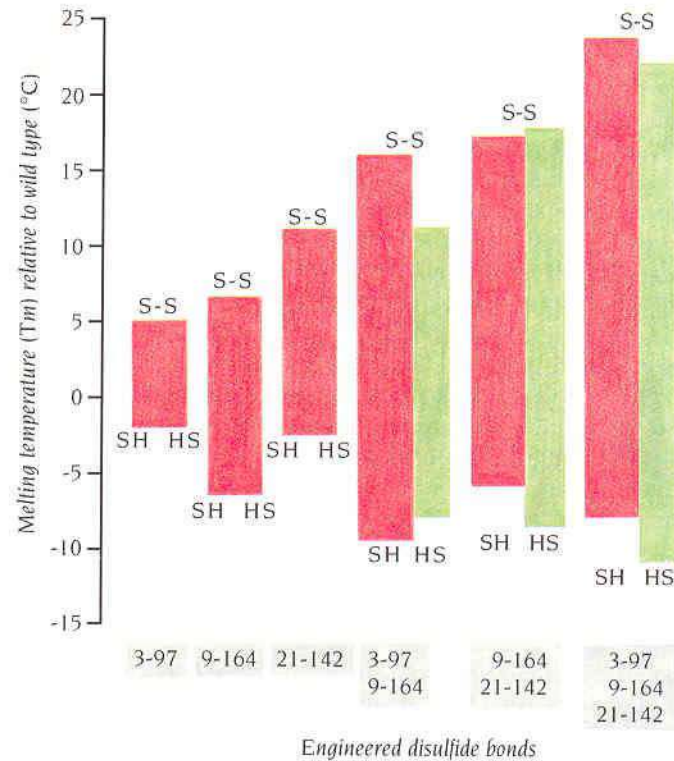


FIG. 17. Schematic drawing of the backbone of ribonuclease A and ribonuclease S, demonstrating the positions of proline residues and the site of subtilisin cleavage. (Adapted from Richardson, 1981).

4 Pro, with Pro93 & Pro114
are *cis* in the native state

Disulfide bonds increase protein stability



Stabilizing the dipoles of α -helices increases stability

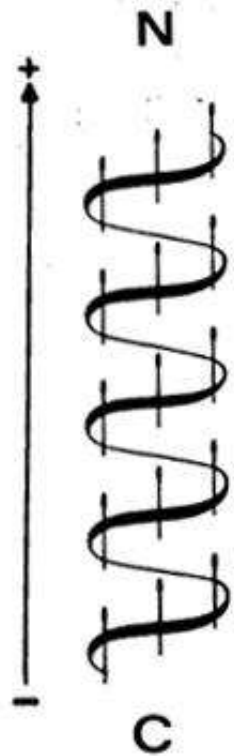
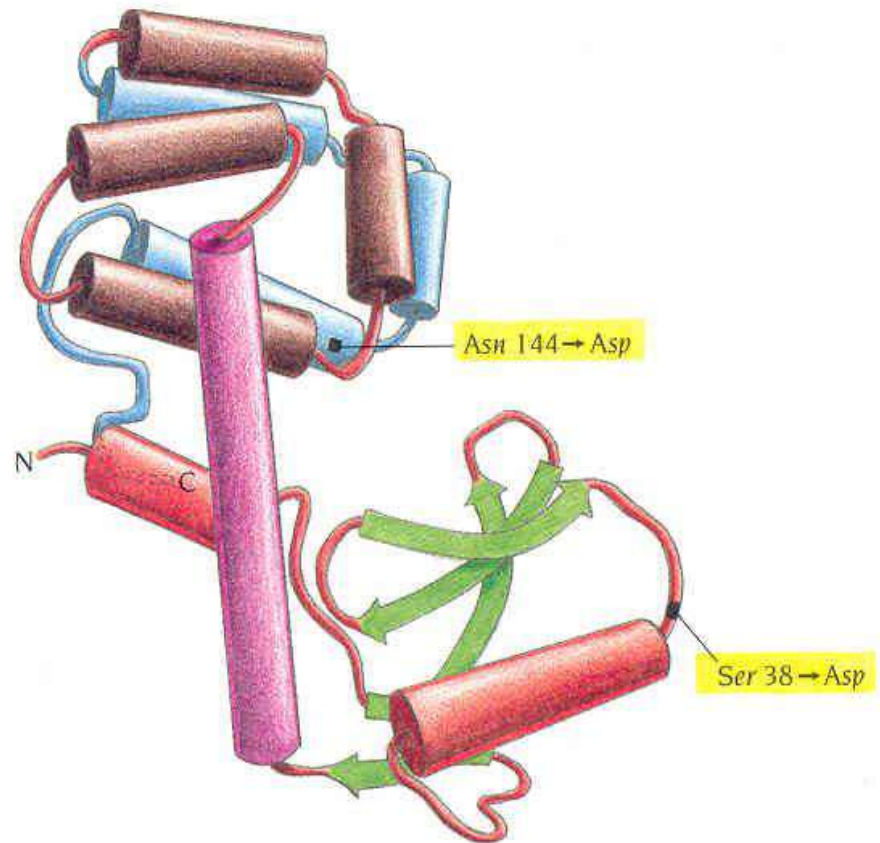
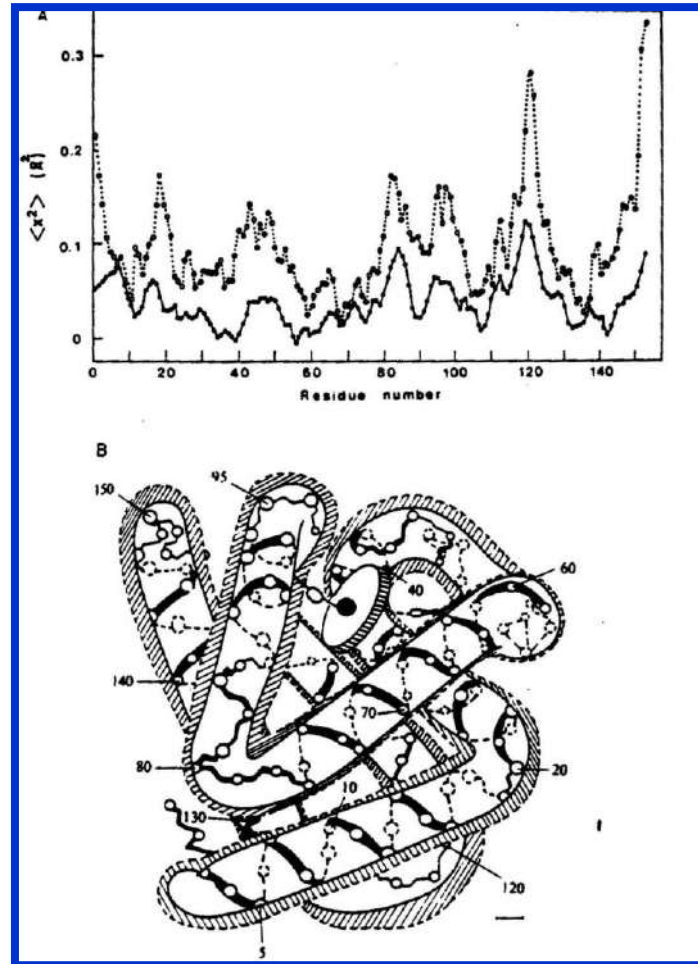


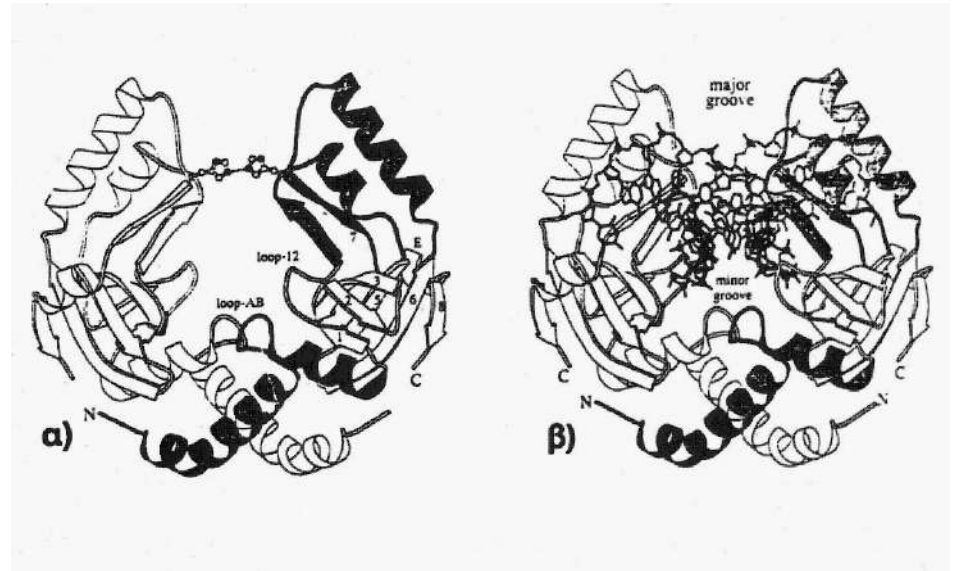
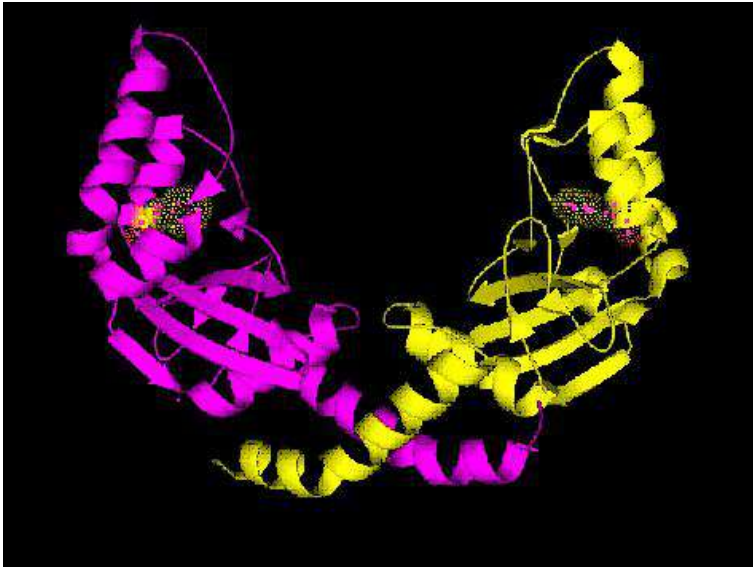
Fig. 6. A schematic drawing of the peptide dipole moments in an α -helix. The entire helix has a dipole moment with the positive pole at the N-terminus and the negative end at the C-terminus. (Adapted from Hol [63].)



The folded state has frequently a flexible structure



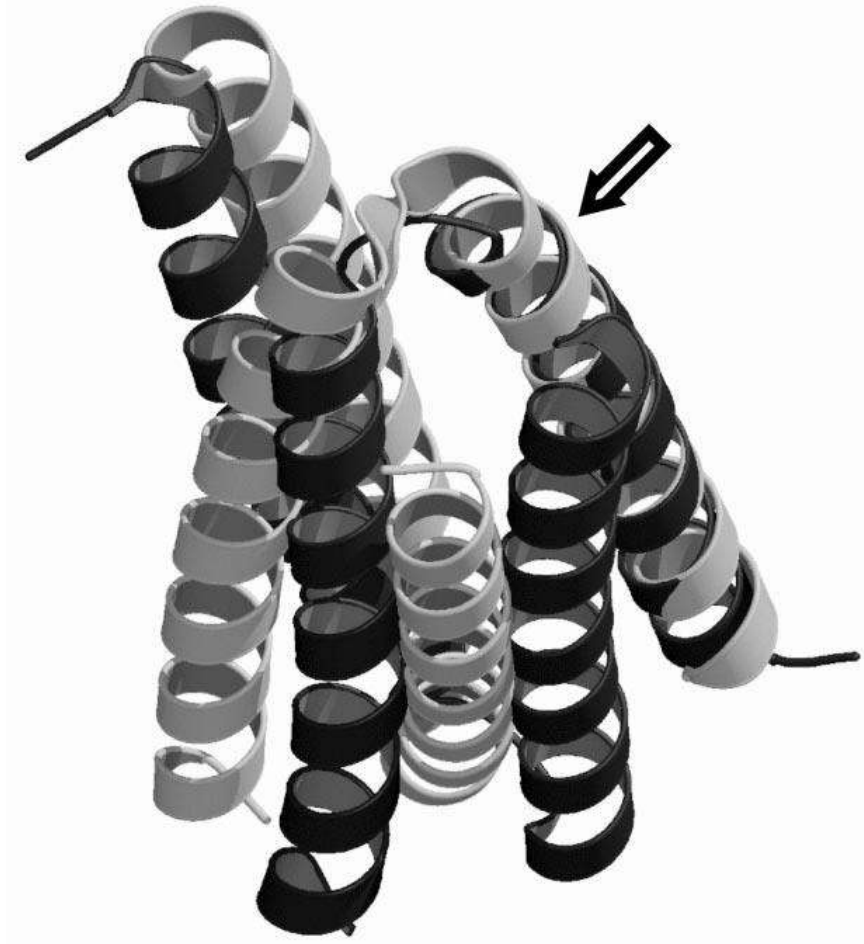
Flexibility of the folded state is essential for function



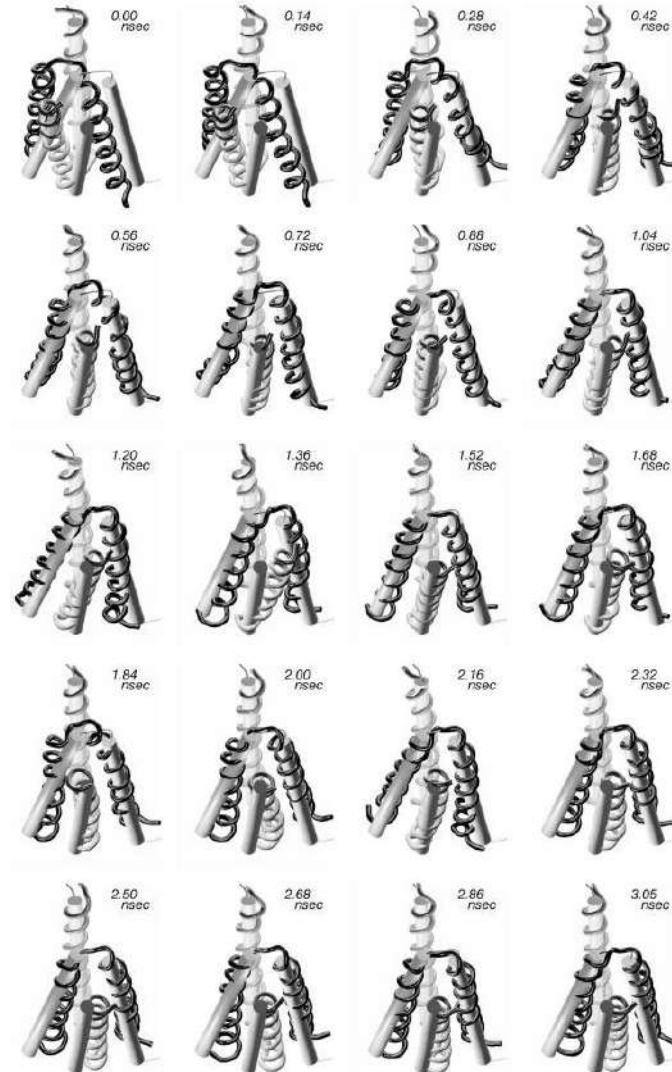
Structural Dynamics

- Biological life depends on motion, and this manifests itself in proteins that display motion over a formidable range of time scales (femtoseconds to micro- or milliseconds).
- Outstanding challenge: a quantitative understanding of the linkages among protein structure, dynamics, and function.
- These linkages are becoming increasingly explorable due to conceptual and methodological advances. BUT: the research questions in the field are becoming increasingly complex (e.g. the mechanistic understanding of high-order interaction networks in allosteric signal propagation through a protein matrix). In analogy to the “protein folding problem”, the way forward lies in the successful integration of experiment and computation, while utilizing the rapid expansion of sequence and structure space.
- Looking forward, the future is bright, and we are in a period where we are on the doorstep to, at least in part, comprehend the importance of dynamics for biological function.

Structural Dynamics



Structural Dynamics



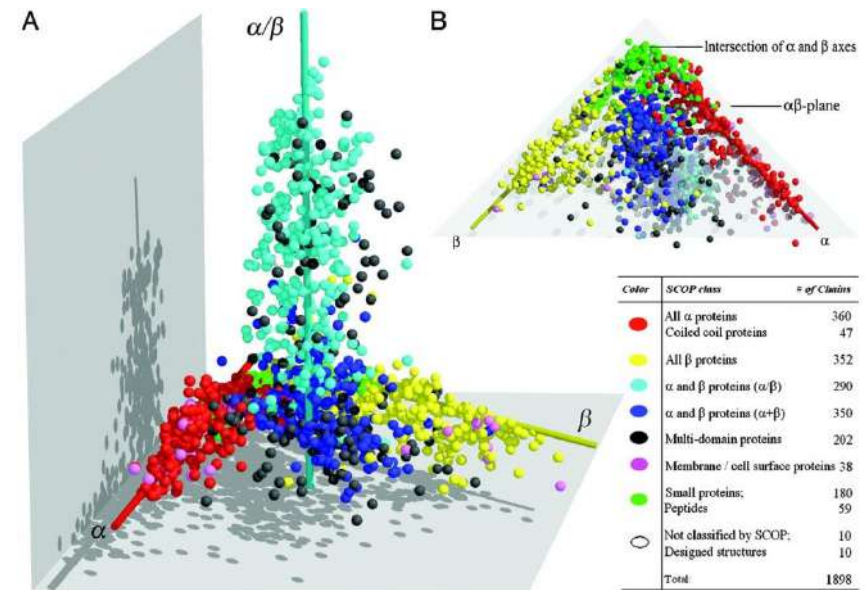
Πρωτεϊνικές αλληλουχίες

Σε μια περίοδο πάνω από 3 δισεκατομμύρια χρόνια μια μεγάλη ποικιλία από μόρια πρωτεϊνών έχει εξελιχθεί για την εκτέλεση των πολύπλοκων λειτουργιών των κυττάρων και οργανισμών. Πιστεύεται ότι αυτά τα μόρια έχουν εξελιχθεί από τυχαία μετάλλαξη των γονιδίων και φυσική επιλογή εκείνων των προϊόντων τους που έχουν αποκτήσει κάποια λειτουργική υπεροχή σε σχέση με την επιβίωση των οργανισμών. Με την έλευση της μοριακής γενετικής και των τεχνικών κλωνοποίησης και εισαγωγής γονιδίων, μπαίνουμε τώρα σε μία εποχή γενετικής εκμετάλλευσης της πληροφορίας που περιέχεται στις αλληλουχίες άλλων οργανισμών, σε βαθμό που ήταν αδιανόητο μόλις 50 χρόνια πριν.

Exploring the sequence/structure space & the protein folding problem



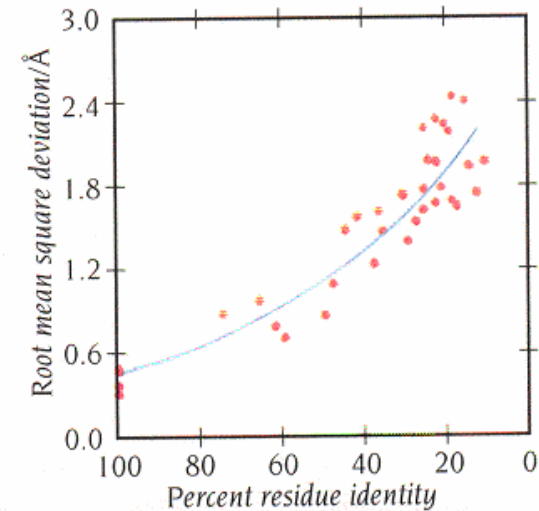
The size of the protein sequence space is astronomical



The known protein structure space is limited

ΟΜΟΛΟΓΕΣ ΠΡΩΤΕΪΝΕΣ ΕΧΟΥΝ ΣΥΝΤΗΡΗΜΕΝΟΥΣ ΔΟΜΙΚΟΥΣ ΠΥΡΗΝΕΣ ΚΑΙ ΜΕΤΑΒΛΗΤΕΣ ΠΕΡΙΟΧΕΣ ΣΤΡΟΦΩΝ

Figure 17.1 The relation between the divergence of amino acid sequence and three-dimensional structure of the core region of homologous proteins. Known structures of 32 pairs of homologous proteins such as globins, serine proteinases, and immunoglobulin domains have been compared. The root mean square deviation of the main-chain atoms of the core regions is plotted as a function of amino acid homology (red dots). The curve represents the best fit of the dots to an exponential function. Pairs with high sequence homology are almost identical in three-dimensional structure, whereas deviations in atomic positions for pairs of low homology are of the order of 2 Å. (From C. Chothia and A. Lesk, *EMBO J.* 5: 823–826, 1986.)



Η ομολογία αλληλουχιών αμινοξέων υποδηλώνει ομοιότητα στην δομή και στις λειτουργίες

```

Human-zCr      MATGQKLMRAVRVFEFGGPEVLKLRSDIAPVPIPKDHOVLKIKVHACGVNPNVETYIRSGTYS
Ecoli-QOR      -----MATRIEFHKHGGPEVLQA-VEFTPADPAENEIQVENKAIGINFIDTYIRSGLYP
                *****
Human-zCr      RKPLLPYTPGSDVAGVIEAVGDNASAFKKGDRVFTSSTISGGYAEYALAADHTVYKLPEK
Ecoli-QOR      -PPSLPSGLGTEAAGIVSKVSGVKHKIKAGDRVVYAQSALGAYSSVHNIADKAAILPAA
                ***

Human-zCr      LDPKQGAAGIGIPYFTAYRALIHSACVKAGESVLVHGASGGVGLAACQIARAYGLKILGTA
Ecoli-QOR      ISPEQAAASFLKGLTVYLLRKYEIKPDEQFLPHAAAGGVGLIACQWAKALGAKLIGTV
                *****

Human-zCr      GTEEGQKIVLQNGAHEVPNHREVNYIDKIKKYVGEKGIDIIIEMLANVNSKDLSSLSHG
Ecoli-QOR      GTAQKAQSALKAGAWQVINYREEDLVERLKEITGGKKRVVVYDSVGRDWTWERSLDCLQRR
                *****

Human-zCr      GRVIVVG-SRGTIEINPRDTMAKES---SIIGVTLFSSSTKEEFQYAAALQAGMEIGWL
Ecoli-QOR      GLMVSFGNSSGAVTVGNLILNQKGSGLYVTRPSLQGYITTREELTEASNELFSLIASGVI
                *****

Human-zCr      KPVIGSQ--YPLEKVAEAHENIHGSGATGKMILLL
Ecoli-QOR      KVDVABQKQVPLKDAQRAHE-ILESRATQGSSLLIP
                *
    
```

Figure 7.2 Optimal global sequence alignment. Alignment of the amino acid sequences of human ζ-crystallin (Swiss-Prot Q08257) and *E. coli* quinone oxidoreductase (Swiss-Prot P28304). It is an optimal global alignment produced by the CLUSTAL W program (Higgins et al., 1996). Identical residues are marked by asterisks below the alignment, while dots indicate conserved residues.

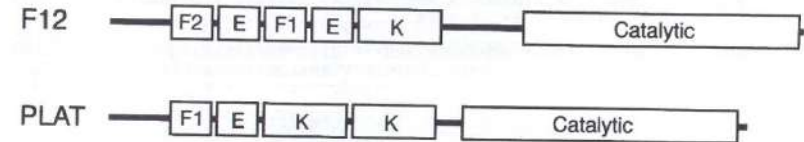


Figure 7.3 Modular structure of two proteins involved in blood clotting. Schematic representation of the modular structure of human tissue plasminogen activator and coagulation factor XII. The module labeled Catalytic are shared by several proteins involved in blood clotting. F1 and F2 are frequently repeated units that were first seen in fibronectin. E is a module resembling epidermal growth factor. A module known as a "kringle domain" is denoted K.

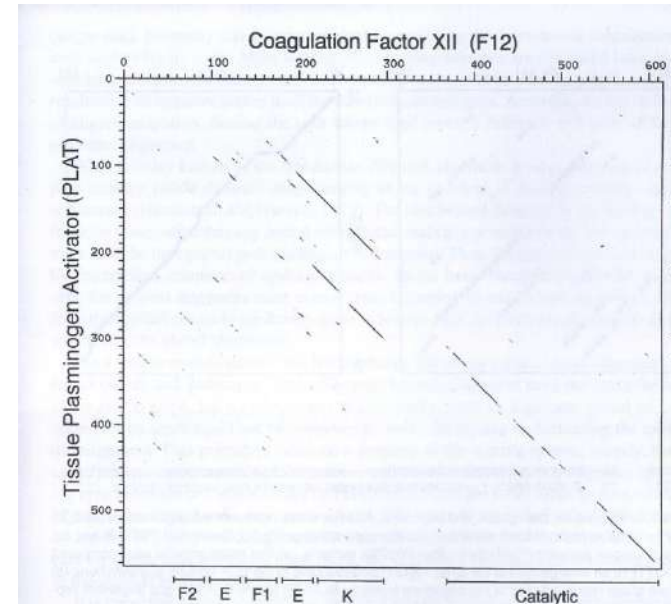
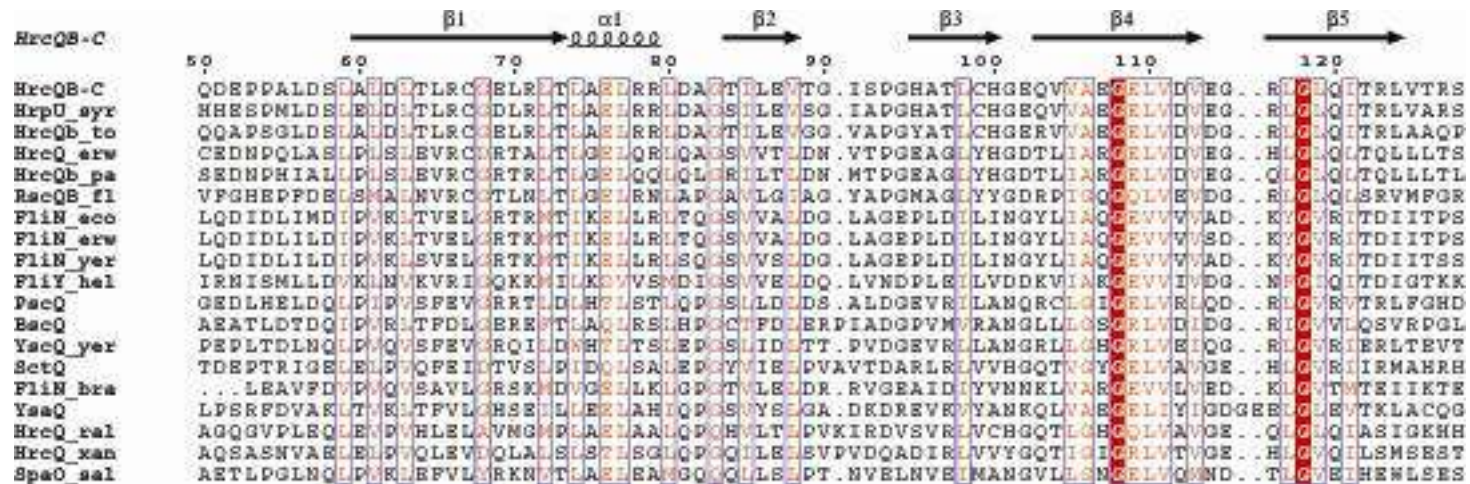
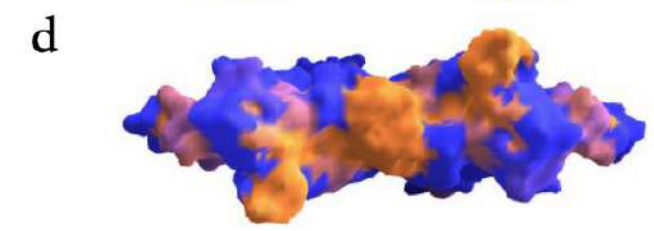
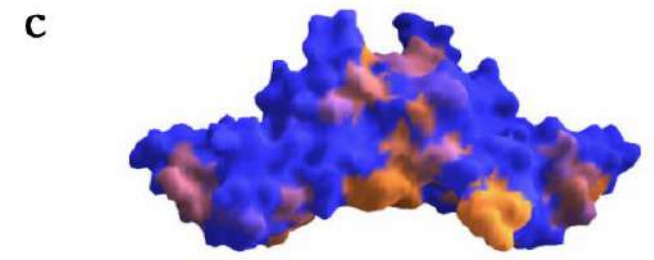
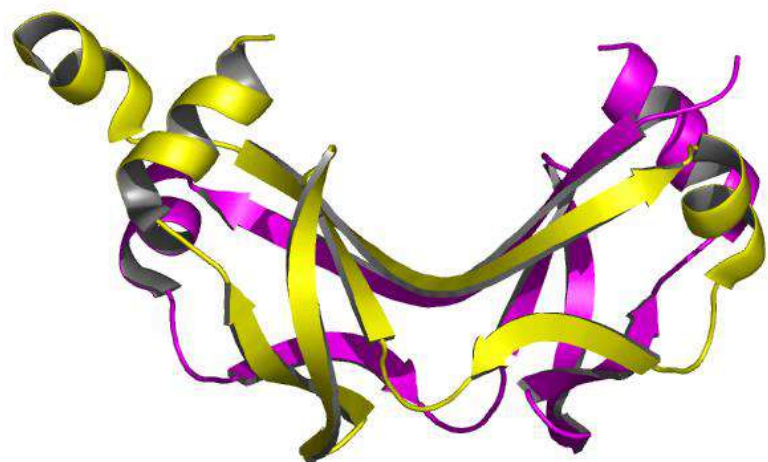
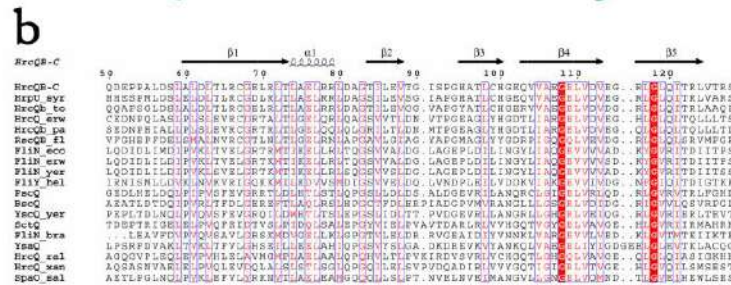
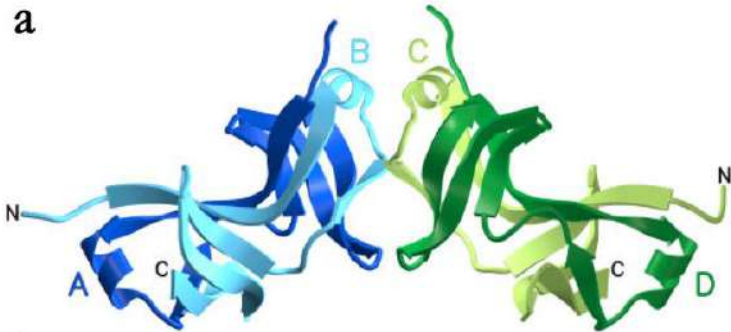


Figure 7.4 Dot matrix sequence comparison. Dot matrix comparison of the amino acid sequences of human coagulation factor XII (F12; Swiss-Prot P00748) and tissue plasminogen activator (PLAT; Swiss-Prot P00750) and proteins. The figure was generated using the DOTTER program (Sonnhammer & Durbin, 1996).

Συντηρημένα πρότυπα στις πρωτεϊνικές αλληλουχίες: ομοιότητα στην δομή και στις λειτουργίες

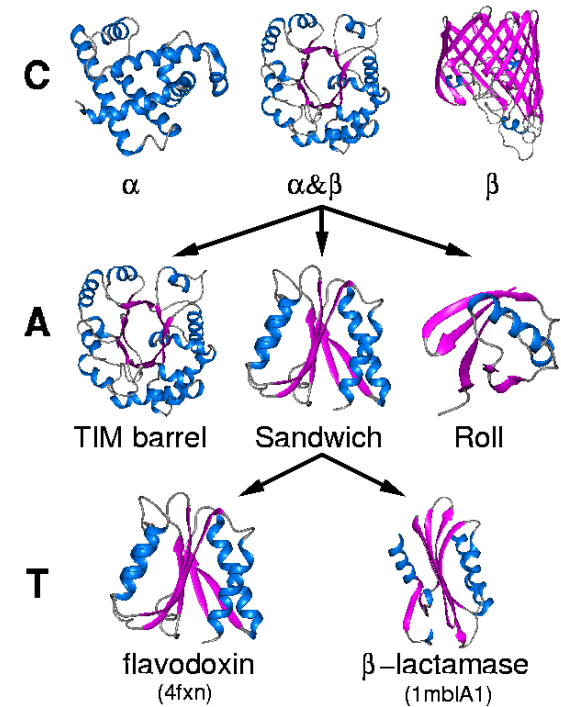
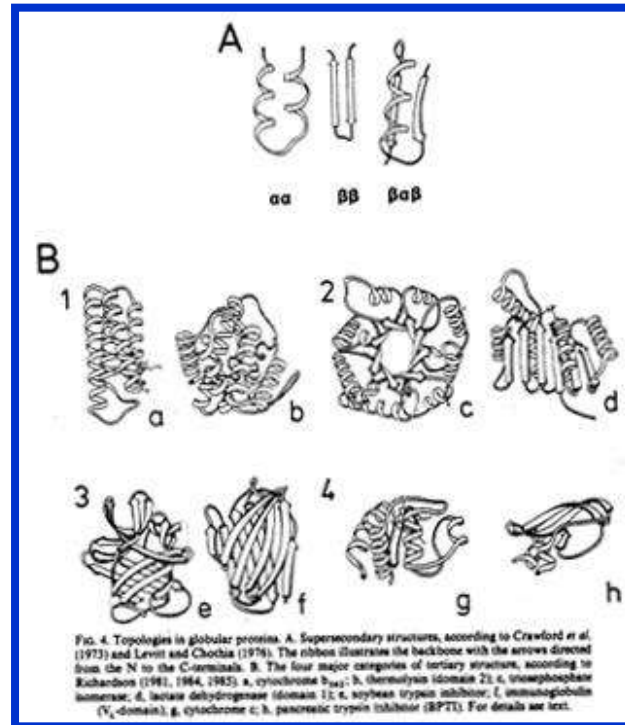


Η πρωτεΐνη HrcQb από το Type III εκκριτικό σύστημα των φυτοπαθογόνων (*P. syringae*)



Many different amino acid sequences give similar 3D-structures

For a 150 aa domain, there are 20^{150} or roughly 10^{200} possible sequences, of which 10^{38} members can be extracted that have less than 20% aa sequence homology. Assuming that 1 out of a billion sequences folds, we are left with 10^{29} possible proteins. However there are only ~1000 topologically different domain structures, which means that there are 10^{26} side-chain arrangements with less than 20% homology that give similar polypeptide folds.



Protein Folding, PSSPs & Protein Design

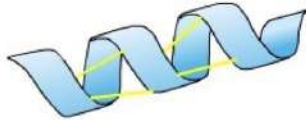
Η γνώση μιας τριτοταγούς δομής πρωτεΐνης είναι αναγκαία προϋπόθεση για τον ανασχεδιασμό της λειτουργίας της για διάφορες βιοτεχνολογικές εφαρμογές.

Το πρόβλημα της επιτυχούς πρόβλεψης/ μοντελοποίησης της αναδίπλωσης των πρωτεϊνών με δεδομένα από την αλληλουχία των αμινοξέων της, αποτελεί διαχρονικά κεντρικό θέμα για την ταχεία πρόοδο της πρωτεϊνικής μηχανικής και για τον σχεδιασμό πρωτεϊνών.



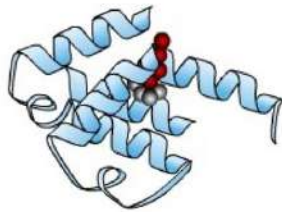
(a)

Secondary Structure:



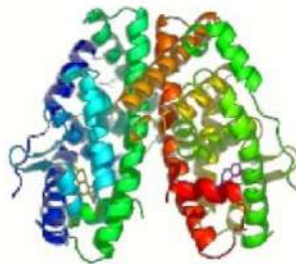
(b)

Tertiary Structure:



(c)

Quaternary Structure:

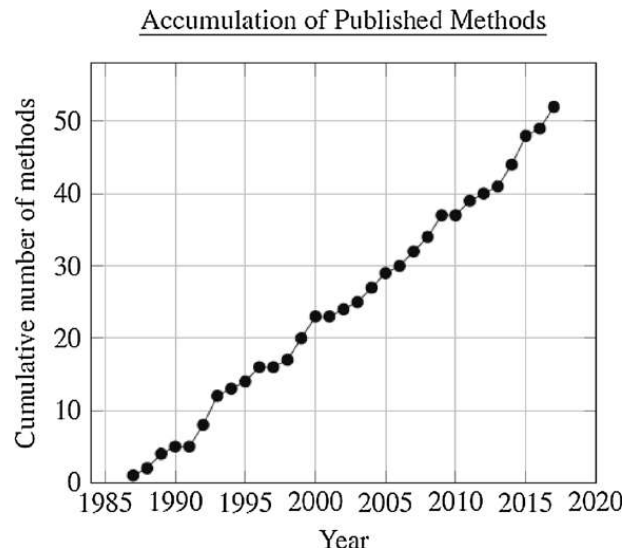


(d)

In vitro methods of obtaining the detailed structure of proteins include X-ray crystallography, nuclear magnetic resonance spectroscopy and electron micrography. Although these methods are accurate, they are time-consuming and costly. Due to these disadvantages, innovative approaches to predict protein structures, such as machine learning, have become the panacea. In the early years, Lim (1974) proposed a method that utilized the physicochemical characteristics of amino acids to predict protein structure. Later, a similar approach was also proposed in Ptitsyn and Finkelstein (1983). Additionally, prediction attempts using sequence patterns and statistical analysis have also been thoroughly investigated in the early years of PSSP.

Protein Folding & Protein Structure Predictions(PSSPs)

- Ever since Kendrew et al. (1958, 1960) and Perutz et al. (1960) determined the first structures of proteins using x-ray crystallography around 1960 (for which they received the shared Nobel Prize (1962)), researchers have been attempting to understand the protein folding problem. By 1988, it was realized that the PSSP problem would require researchers to move away from traditional computing onto newer ways of computation (Rooman and Wodak, 1988; Kneller et al., 1990).
- Hence, machine learning techniques such as ANN were explored. Fig. 4 is a graph that represents the accumulation of the efforts made in improving PSSP with NN over the past 3 decades.



Prediction of protein structure from sequence may exploit our knowledge of molecular forces and evolution


Computational methods proceed along two complementary paths that focus either on the **physical interactions** in the protein structure or the **evolutionary history**.

Physical interactions heavily integrate our understanding of molecular driving forces (thermodynamic/ kinetic simulations or statistical approximations).

Evolutionary considerations provide constraints on protein structures derived from the bioinformatics analysis of the evolution history of proteins, homology to known structures and pairwise evolutionary correlations.

Prediction of protein structure from sequence is a major scientific problem

(Physical interactions, Evolutionary history)

- Complex task 
- Frequently enormous computing time required
- Inverse protein folding problem (which sequence patterns are compatible with a specific fold?)/ Threading techniques

Knowledge of **secondary structure** is frequently necessary for the prediction of tertiary structure

- Modelling tertiary structure from the amino acid sequence alone is for many aa sequences an unsolved problem
- Global tertiary structure imposes frequently local secondary structure

Secondary structure prediction methods benefit from multiple sequence alignments of homologous proteins

Goal: classify each residue as alpha, beta or coil.

Assumption: Secondary structure of a residue is determined by the amino acid at the given position and amino acids at the neighboring ones.

Several prediction methods for secondary structure of proteins have been proposed (most frequently used: Chou & Fasman, GOR, Lim)

All 3 methods assign 1 out of 3 states to each residue

Applied to a set of homologous proteins, the predictive power is higher (underlying assumptions: scaffold more conserved than a.a. sequence)
(best prediction accuracies based on homologies ~72% on the average)

Chow-Fasman algorithm

Chow, P.Y. and Fasman, G.D. Biochemistry (1974)

Statistical approach based on calculation of statistical propensities of each residuum to form an α -helix or β -strand

Low accuracy ($\sim 50\%$) (accuracy of current methods $>75\%$).

GOR

Consider window of 17 positions and see how the conformation of the central residuum depends on this residuum and its 18 neighbors (8 in each direction).

Ideally one would consider all possible combinations of these neighbors. This is impossible: would require collecting statistics for 20^{17} sequences.

Instead assume the central residue depends on its neighbors but the neighbors are independent on each other

Implementation :Statistical information derived from proteins of known structure is stored in three (17X20)matrices, one each for α , β , coil

Structure Predictions

- Model building by homology
- Prediction of loop regions (main chain conformations cluster in sets of similar structures)



Data base of loops

Secondary structure prediction methods

How good are the methods?

Single sequence, single residuum methods

Chou & Fasman 50%

Single sequence, multiple residues methods

GOR IV 65%

Multiple sequence methods

NNSSP 71%

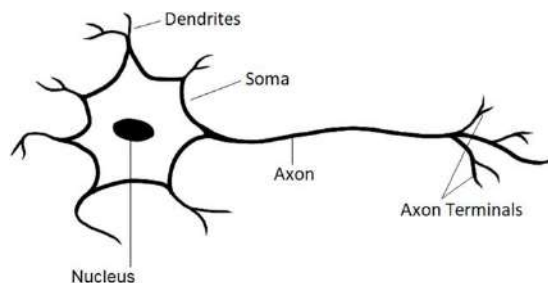
PHD 71%

Taking a weighted consensus of many methods moderate improvement.

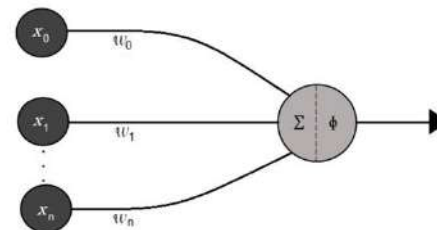
Artificial Neural Network (ANN) methods

Wardah et al., <https://doi.org/10.1016/j.compbiolchem.2019.107093>

- Traditional computing involves human-written instructions in a computer program. On the contrary, artificial intelligence allows a system to modify or write new instructions for itself. One approach of this latter style is through the use of ANNs. This concept is derived from the working patterns of the biological neurons in the brain. Just as the millions of neurons in the brain collectively execute the cognitive processes, ANNs are fashioned in a similar way to carry out intelligent computation.



Biological neuron

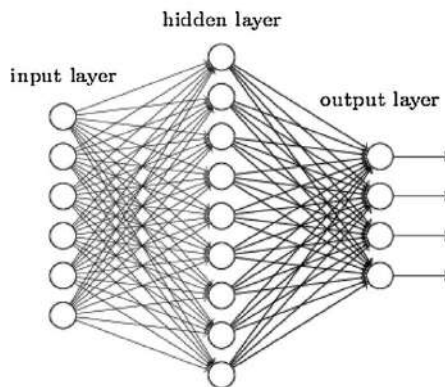


Artificial neuron

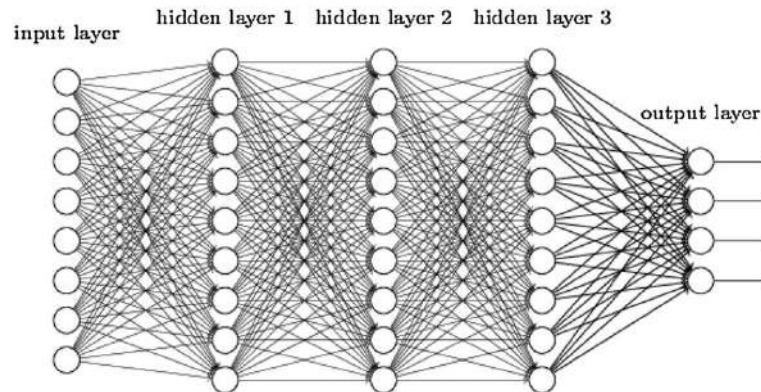
Artificial Neural Network (ANN) methods

- An ANN is a network created by at least 2 layers of neuron-like processing units. The initial layer is called the input layer as it introduces input variables into the network. The final layer is the output layer, which may contain units for carrying out output classification. For networks that contain more than 2 layers, the remaining inner layers are called the hidden layers. A shallow network is one that ideally contains none or one hidden layer. On the other hand, deep network refers to a network of artificial neurons comprising many hidden layers. Evidently, deep NNs have been highly successful in solving complex problems (Bianchini and Scarselli, 2014).

Shallow Neural Network



Deep Neural Network



Artificial Neural Network (ANN) methods

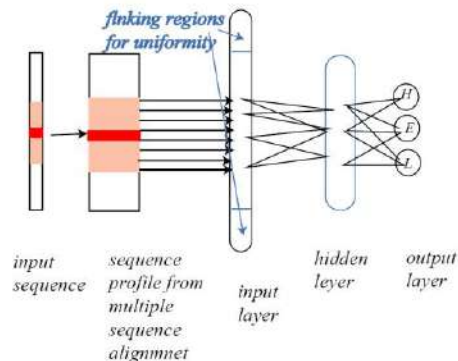
- Inside an ANN, complex matrix computations take place throughout the inner layers. A standard ANN generally accepts a set of input values in the form of vectors containing feature-values (example $x_0, x_1, x_2, \dots, x_n$). Each unit (neuron) that is part of the following layer assigns a designated weight (and other parameters such as bias) to the input, which produces some output. In supervised learning, the real corresponding output is also supplied to the algorithm during training. If the produced output does not match the real output for that particular input, the weights get adjusted automatically through an algorithm of choice. In large networks with high dimensionality like those for the protein structure prediction problems, backpropagation is often used for adjusting weights. Backpropagation refers to the method of revisiting the previous layers and adjusting weights so that the calculated output is closer to the actual expected output

Neural network methods

Inside an ANN, complex matrix computations take place throughout the inner layers. A standard ANN generally accepts a set of input values in the form of vectors containing feature-values (example $x_0, x_1, x_2, \dots, x_n$). Each unit (neuron) that is part of the following layer assigns a designated weight (and other parameters such as bias) to the input, which produces some output. In supervised learning, the real corresponding output is also supplied to the algorithm during training. If the produced output does not match the real output for that particular input, the weights get adjusted automatically through an algorithm of choice. In large networks with high dimensionality like those for the protein structure prediction problems, backpropagation is often used for adjusting weights. Backpropagation refers to the method of revisiting the previous layers and adjusting weights so that the calculated output is closer to the actual expected output

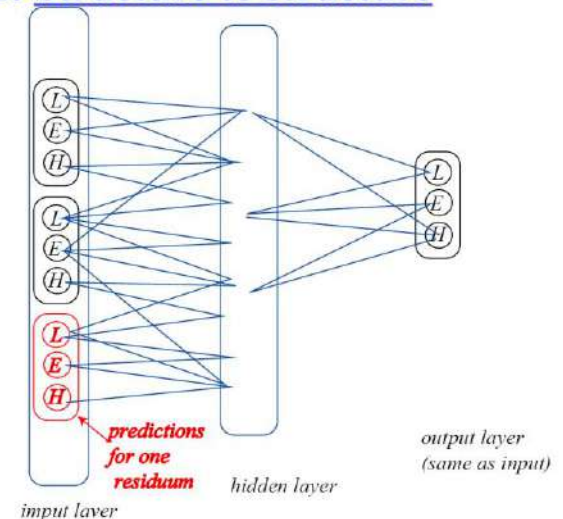
Level 1: sequence to structure

Take window of 13 adjacent residues is (6 before and 6 after the residuum for which we predict secondary structure at the given step). In the output layer, for each residuum we have scores for helix, strand, loop



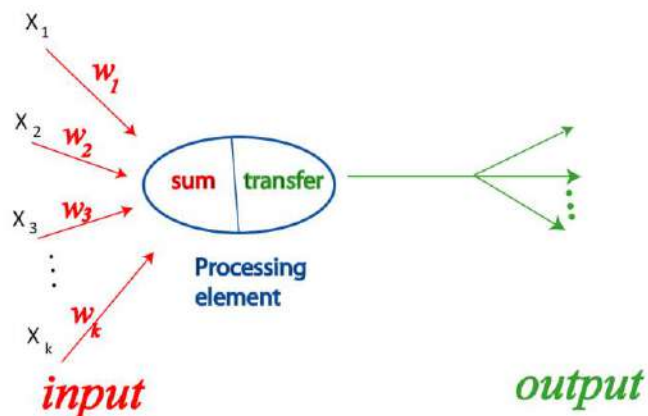
Level 2: Structure to structure

The role for the second level is to include dependence on the conformation predicted for a residuum and conformation of its neighbors



Artificial Neural Network methods

Artificial neuron:



Accumulation of Published Methods

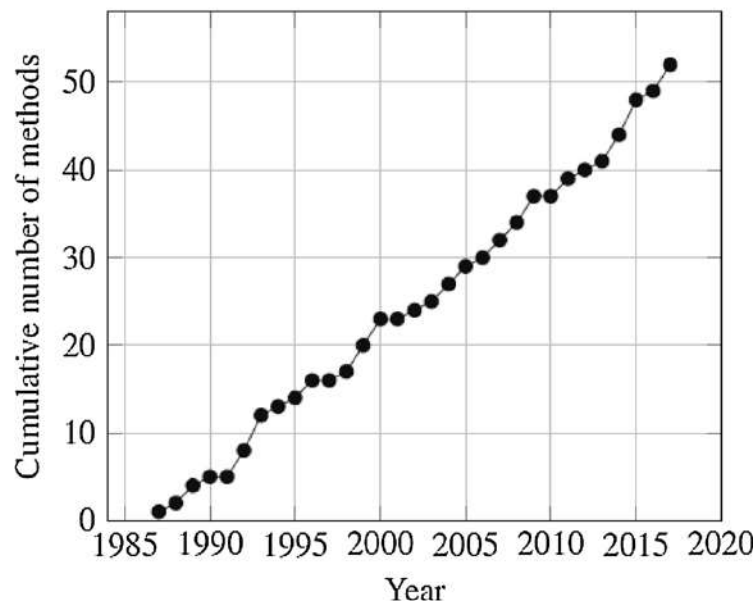


Table 2

The major periodically relevant state-of-the-art methods are shown along with the types of feature values they employed in their networks.

Neural network method	Accuracy (Q3)	Seq info	Evo info	Physico chem info
Qian & Sejnowski 1988 (Qian and Sejnowski, 1988)	64.3%	✓		
PHD 1994 (Rost et al., 1994)	71.4%	✓	✓	
PSIPRED 1997 (Jones, 1999)	76.5%	✓	✓	
JPRED3 2008 (Cole et al., 2007)	81.5%	✓	✓	
SPIDER3 2017 (Heffernan et al., 2017)	84%	✓	✓	✓

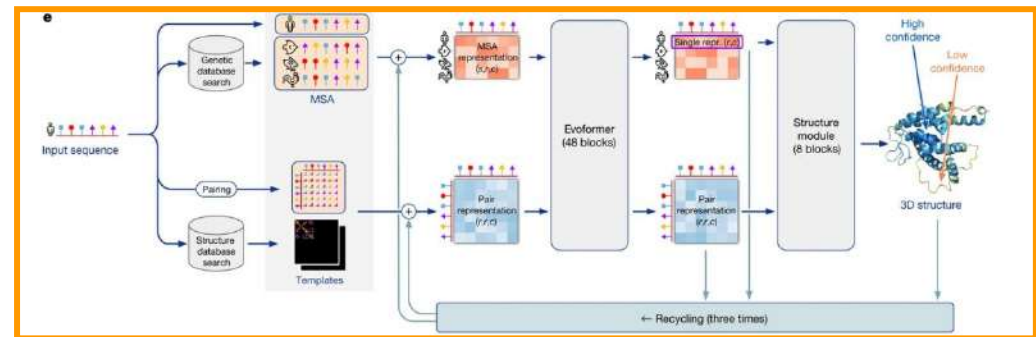
A Structure-Prediction-Miracle?

(...AlphaFold 2 means that predicting a protein structure from sequence will be, for all practical purposes, a solved problem...)



Protein structure predictions focus on either the physical interactions or the evolutionary history of a protein

The Delphi oracle in Greece



A Machine Learning Approach: AlphaFold

Jumper *et al.*, **Highly accurate protein structure prediction with AlphaFold**, *Nature* | Vol 596 | 26 August 2021 | 583

AlphaFold

Article

Improved protein structure prediction using potentials from deep learning

<https://doi.org/10.1038/s41586-019-1923-7>

Received: 2 April 2019

Accepted: 10 December 2019

Published online: 15 January 2020

Andrew W. Senior^{1,4*}, Richard Evans^{1,4}, John Jumper^{1,4}, James Kirkpatrick^{1,4}, Laurent Sifre^{1,4}, Tim Green¹, Chongli Qin¹, Augustin Židek¹, Alexander W. R. Nelson¹, Alex Bridgland¹, Hugo Penedones¹, Stig Petersen¹, Karen Simonyan¹, Steve Crossan¹, Pushmeet Kohli¹, David T. Jones^{2,3}, David Silver¹, Koray Kavukcuoglu¹ & Demis Hassabis¹

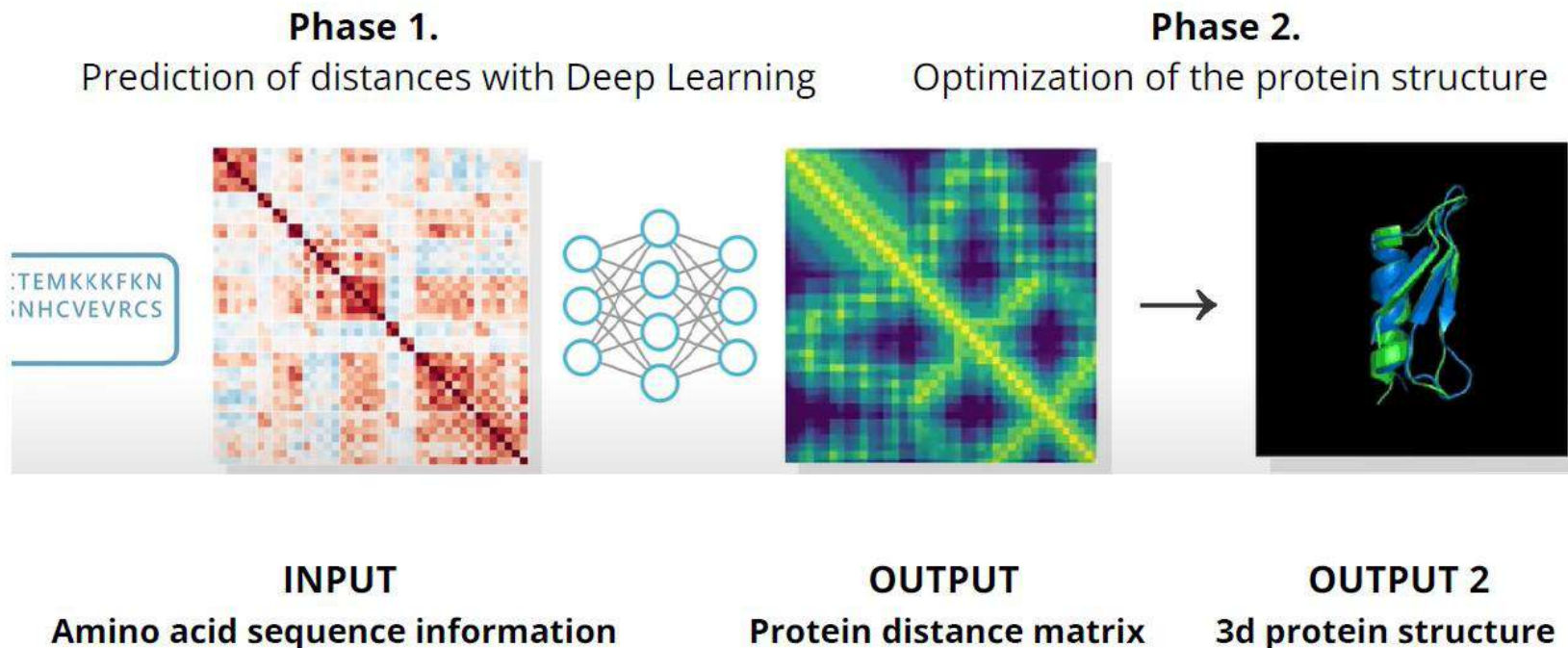
Protein structure prediction can be used to determine the three-dimensional shape of a protein from its amino acid sequence¹. This problem is of fundamental importance as the structure of a protein largely determines its function²; however, protein structures can be difficult to determine experimentally. Considerable progress has recently been made by leveraging genetic information. It is possible to infer which amino acid residues are in contact by analysing covariation in homologous sequences, which aids in the prediction of protein structures³. Here we show that we

Reduce P
image an

AlphaFold

AlphaFold takes the amino acid sequence and convert it into an image for an artificial intelligence algorithm to translate it into another image representing the structure of the protein.

INPUT AND OUTPUT: Multiple Sequence Alignments and DISTOGRAMS



Article

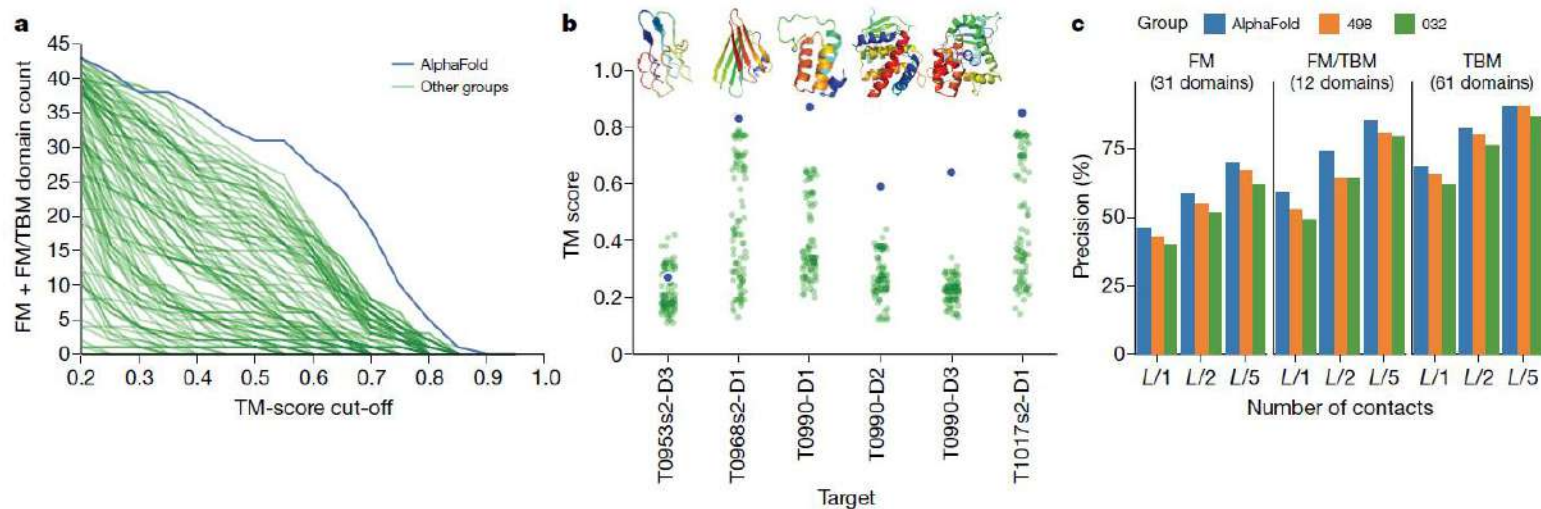


Fig. 1 | The performance of AlphaFold in the CASP13 assessment. **a**, Number of FM (FM + FM/TBM) domains predicted for a given TM-score threshold for AlphaFold and the other 97 groups. **b**, For the six new folds identified by the CASP13 assessors, the TM score of AlphaFold was compared with the other groups, together with the native structures. The structure of T1017s2-D1 is not available for publication. **c**, Precisions for long-range contact prediction in

CASP13 for the most probable L , $L/2$ or $L/5$ contacts, where L is the length of the domain. The distance distributions used by AlphaFold in CASP13, thresholded to contact predictions, are compared with the submissions by the two best-ranked contact prediction methods in CASP13: 498 (RaptorX-Contact²⁶) and 032 (TripletRes³²) on 'all groups' targets, with updated domain definitions for T0953s2.

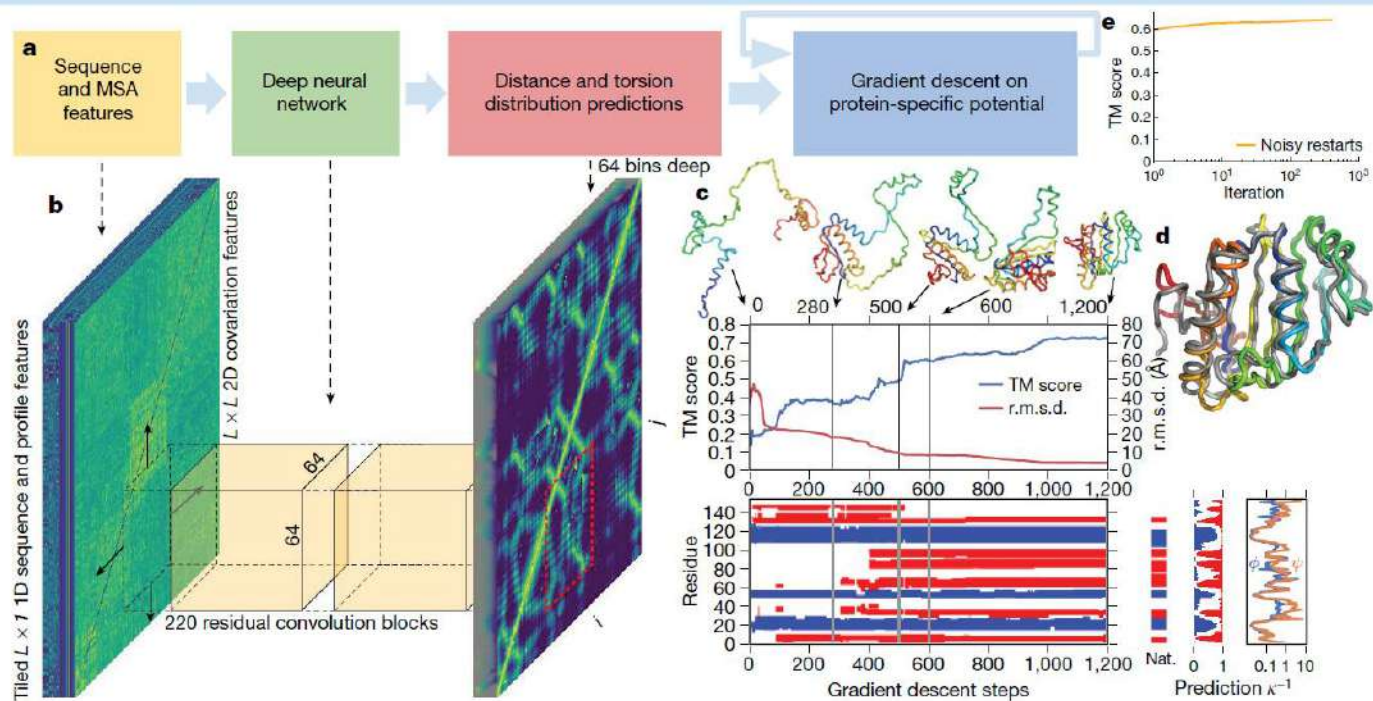


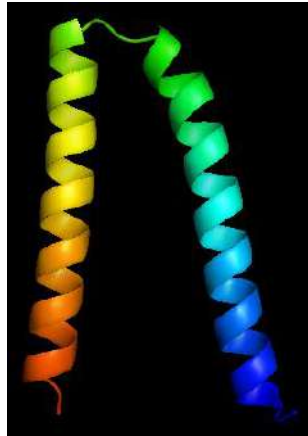
Fig. 2 | The folding process illustrated for CASP13 target T0986s2. CASP target T0986s2, $L = 155$, PDB: 6N9V. **a**, Steps of structure prediction. **b**, The neural network predicts the entire $L \times L$ distogram based on MSA features, accumulating separate predictions for 64×64 -residue regions. **c**, One iteration of gradient descent (1,200 steps) is shown, with the TM score and root mean square deviation (r.m.s.d.) plotted against step number with five snapshots of the structure. The secondary structure (from SSI³³) is also shown (helix in blue, strand in red) along with the native secondary structure (Nat.), the secondary

structure prediction probabilities of the network and the uncertainty in torsion angle predictions (as κ^{-1} of the von Mises distributions fitted to the predictions for φ and ψ). While each step of gradient descent greedily lowers the potential, large global conformation changes are effected, resulting in a well-packed chain. **d**, The final first submission overlaid on the native structure (in grey). **e**, The average (across the test set, $n = 377$) TM score of the lowest-potential structure against the number of repeats of gradient descent per target (log scale).

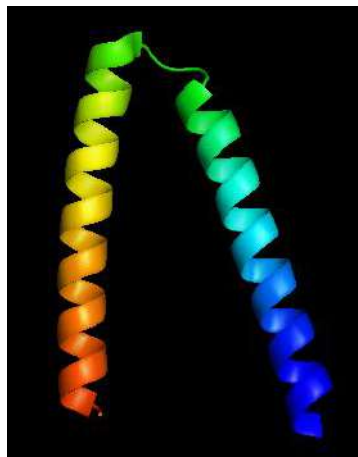
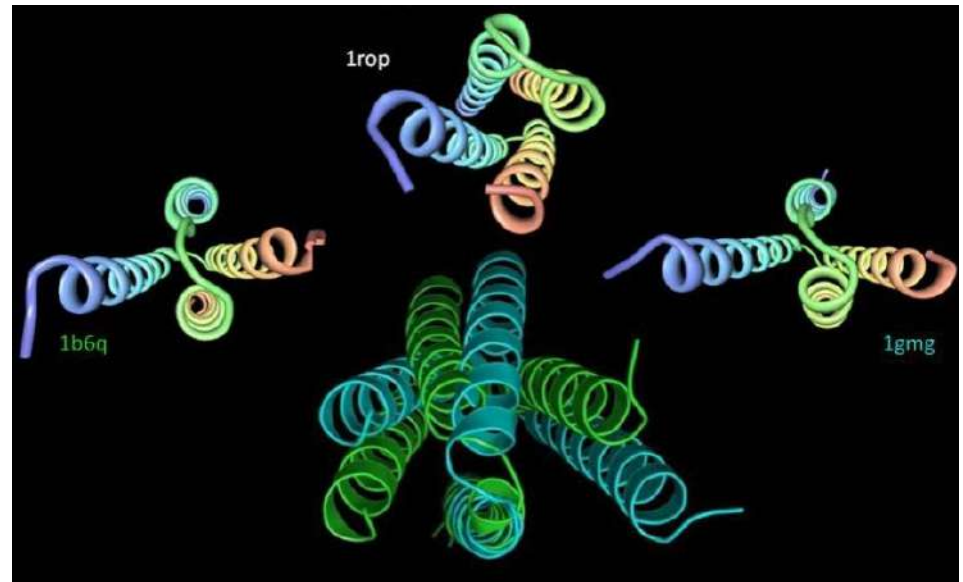
Machine Learning vs Experiment



AlphaFold



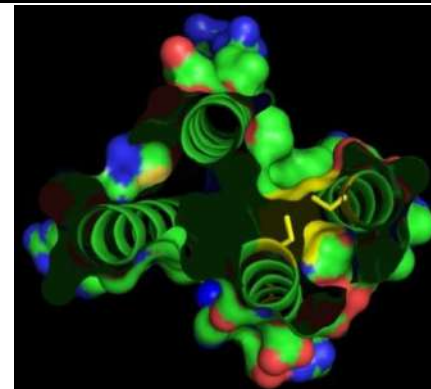
1b6q



1gmg



1rop



1gmg: Formation of transient S-S bridges upon folding

APPLICATIONS

Mutants with cavities



Figure 1. The wild-type ROP dimer with the sites of mutation (Leu41) and the internal water molecules.

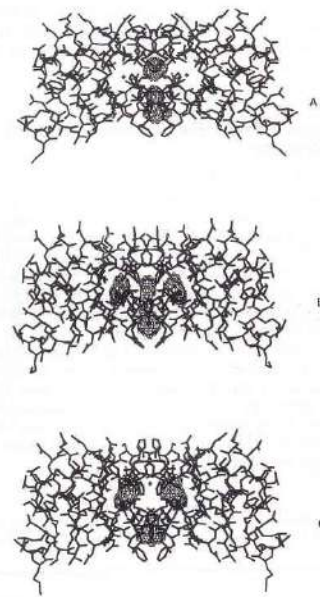


Figure 2. Cavities in wild-type ROP (a), L41V (b) and L41A (c).

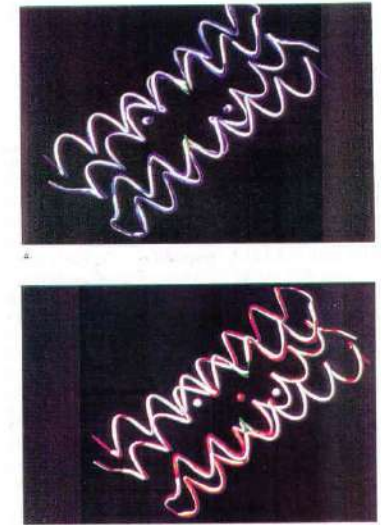


Figure 3. Superposition of (a) wild-type ROP (white) and L41V (blue), and (b) wild-type ROP (white) and L41A (red). The site of mutation is highlighted (green). Internal water molecules are shown as spheres.

Mutants with cavities

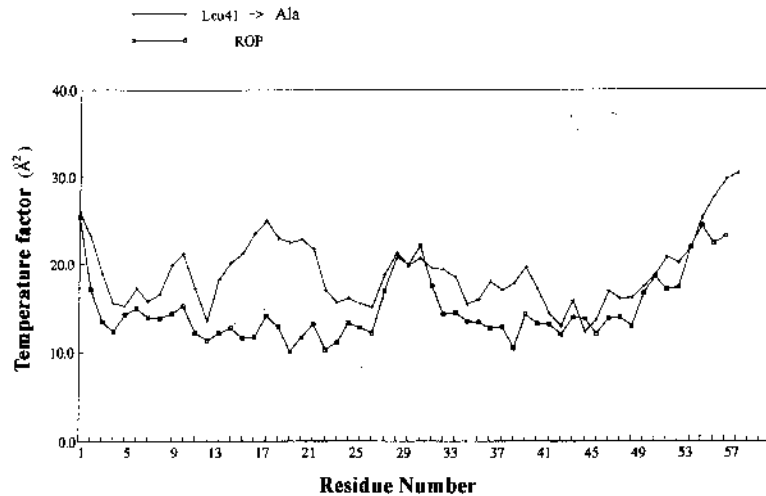


Figure 4. Variation of the average isotropic temperature factors (\AA^2) of ROP and L41A.

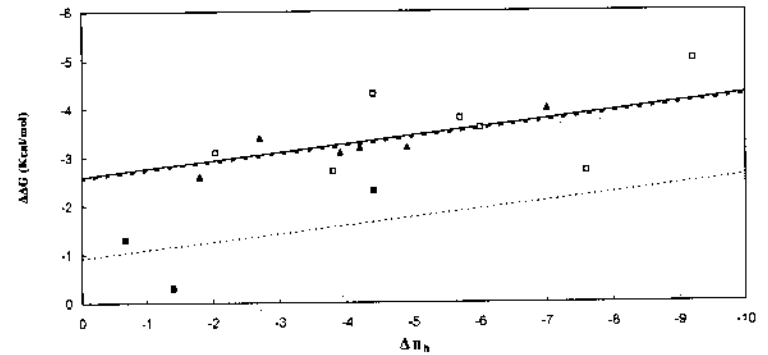


Figure 6. The destabilization, $\Delta\Delta G$, of ROP, barnase and T4 lysozyme mutants presented as a function of $\Delta\pi_0$ (Table 1). The least-squares line for Leu \rightarrow Ala mutants was used to determine the slope as $0.17 \text{ kcal mol}^{-1} \text{ contact}^{-1}$. The lines for the other two sets of mutants are drawn parallel. The extrapolated intercepts with the $\Delta\Delta G$ axis are given in Table 2. The symbols are the same as described in the legend to Figure 5.

Mutants with cavities

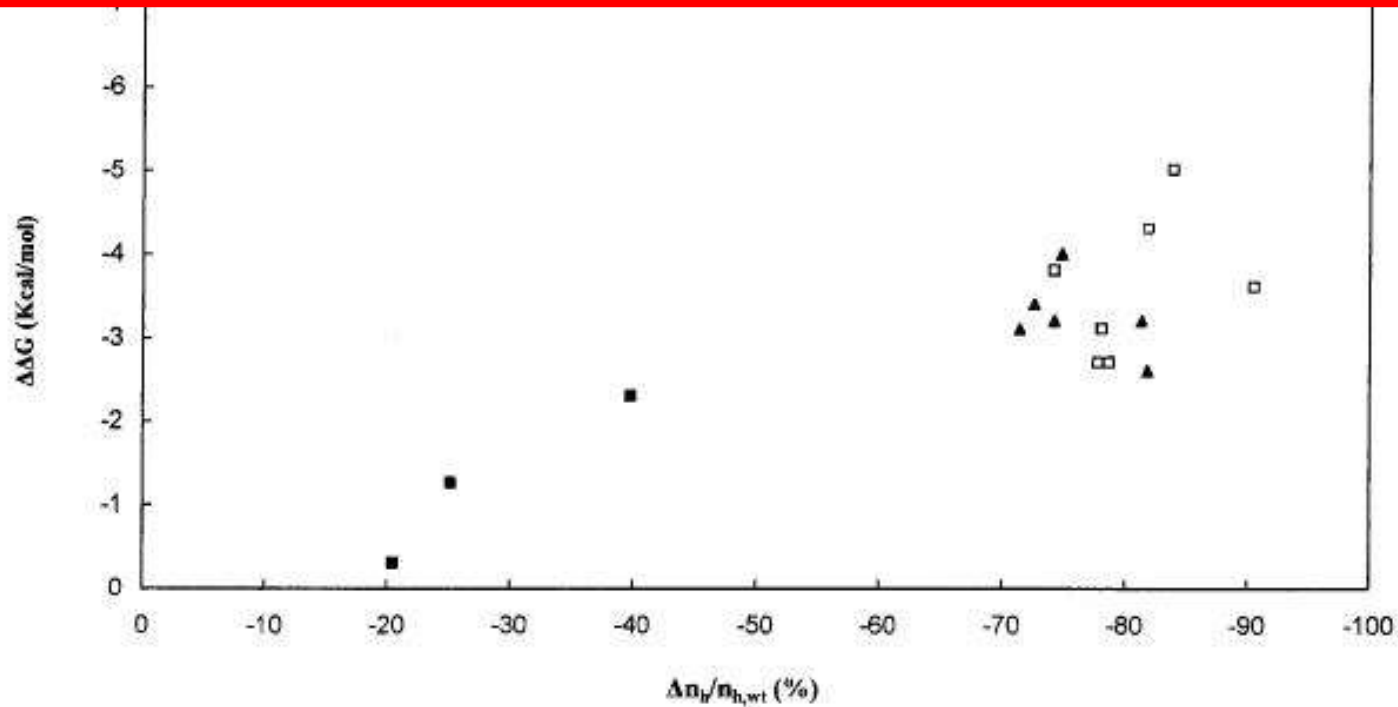
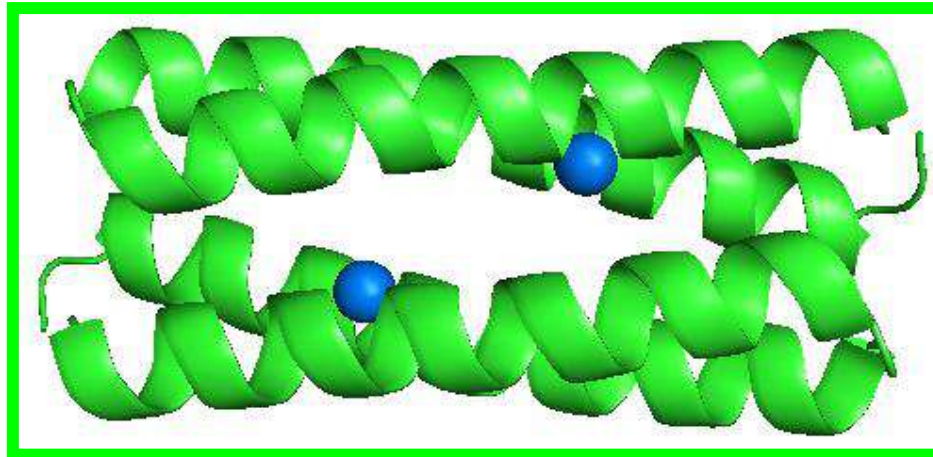
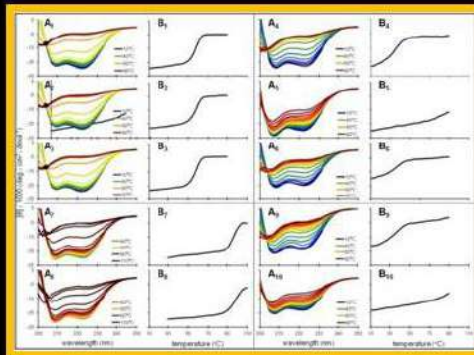


Figure 7. The destabilization, $\Delta\Delta G$, of ROP, bamase and T4 lysozyme mutants (Table 1), presented as a function of the per cent change of contacts relative to the wild-type protein (calculated as $\Delta n_H/n_{H,wt}$). The symbols are the same as described in Figure 5.

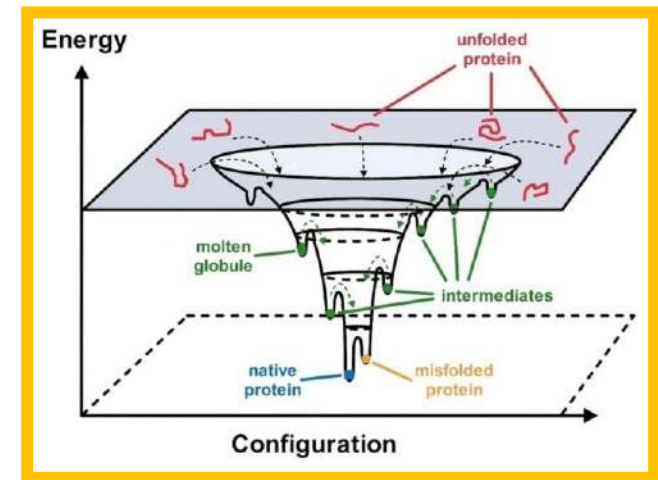
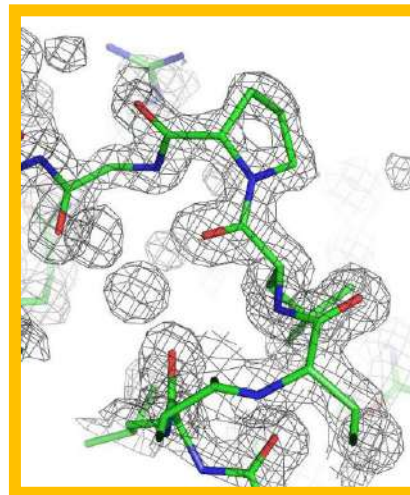
Protein folding studies of recurrent tertiary motifs



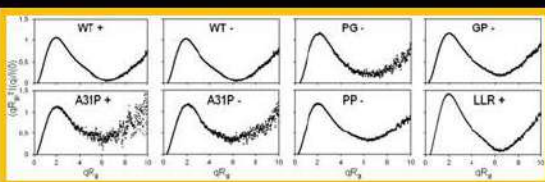
α -helical bundles



Thermal unfolding for Rop and its variants recorded by CD

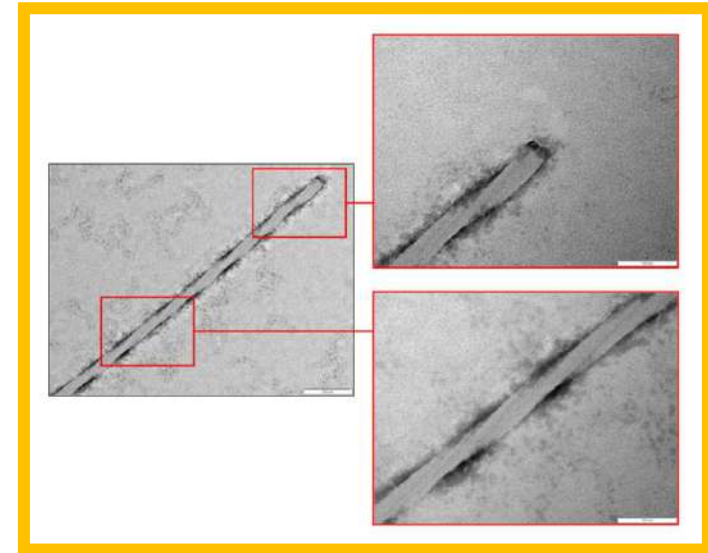
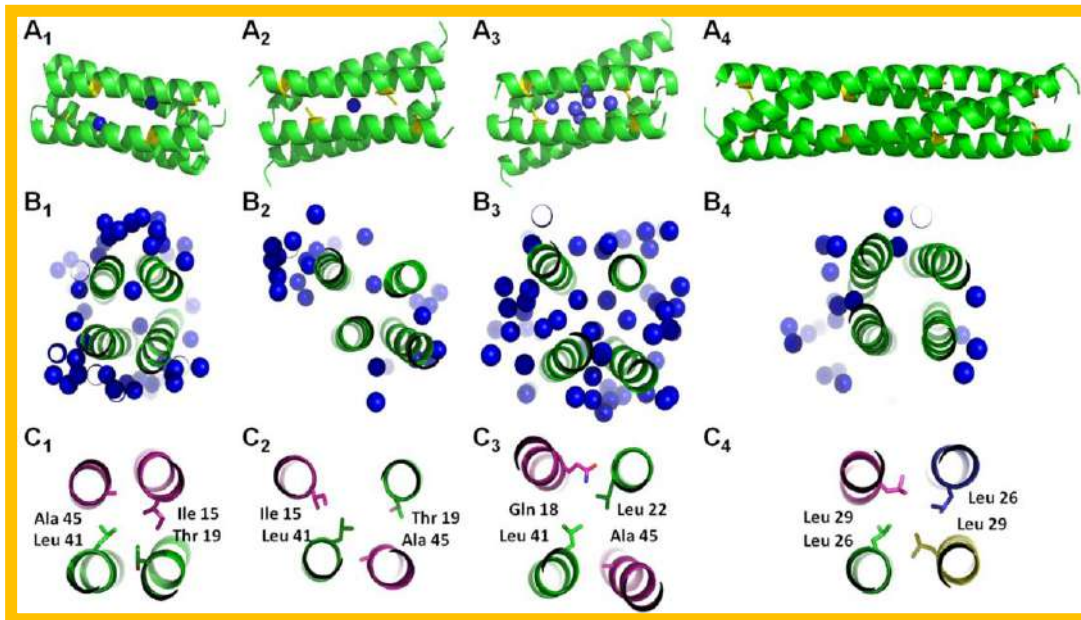


Amprazi *et al.*, PNAS (2014)
Kefala *et al.*, IJMS (2021)



Dimensionless Kratky plots

Novel protein folds have been engineered



Arnittali et al, Int. J. Mol. Sci.

2021



Applications: Scaffolds for protein engineering and new biomaterials

Accessing remote regions of protein sequence space: Backwards reading the sequences of α -Helical Bundles

	f	g	a	b	c	d	e
1	(Met)	Thr					
3	Lys	Gln	Glu	Lys	Thr	Ala	Leu
10	Asn	Met	Ala	Arg	Phe	Ile	Arg
17	Ser	Gln	Thr	Leu	Thr	Leu	Leu
24	Glu	Lys	Leu	Asn	Glu	Leu	Asp
31	← Ala →			Asp	Glu	Gln	Ala
36	Asp	Ile	Cys	Glu	Ser	Leu	His
43	Asp	His	Ala	Asp	Glu	Leu	Tyr
50	Arg	Ser	Cys	Leu	Ala	(Arg	Phe)
57	(Gly	Asp	Asp	Gly	Glu	Asn	Leu)

	f	g	a	b	c	d	e
1		(Leu	Asn	Glu	Gly	Asp	Asp
7	Gly)	Phe	Arg	Ala	Leu	Cys	Ser
14	Arg	Tyr	Leu	Glu	Asp	Ala	His
21	Asp	His	Leu	Ser	Glu	Cys	Ile
28	Asp	Ala	Gln	Glu	Asp	←	Ala
34	→	Asp	Leu	Glu	Asn	Leu	Lys
40	Glu	Leu	Leu	Thr	Leu	Thr	Gln
47	Ser	Arg	Ile	Phe	Arg	Ala	Met
54	Asn	Leu	Ala	Thr	Lys	Glu	Gln
61	Lys	(Thr	Met)				

Parent
protein

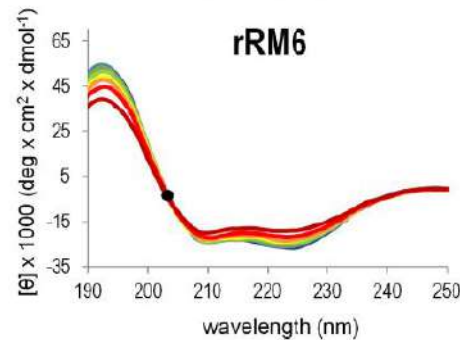
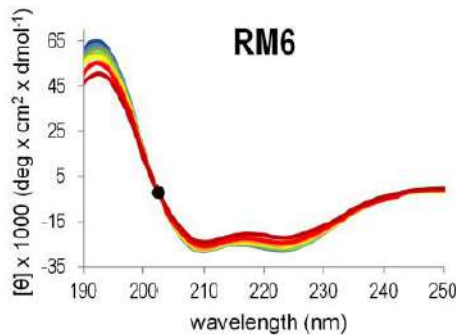
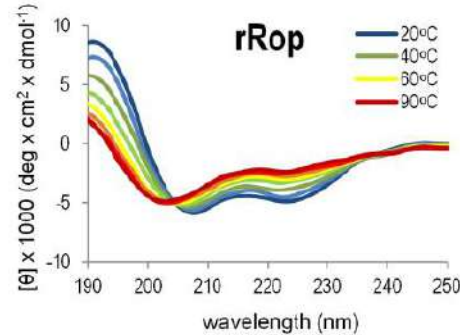
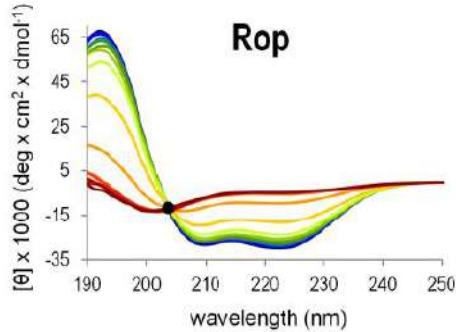
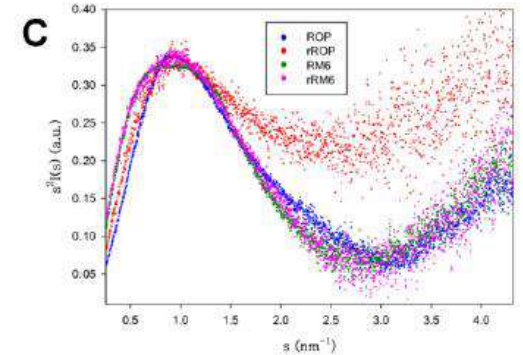
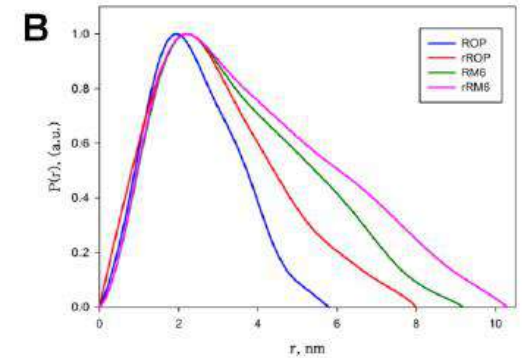
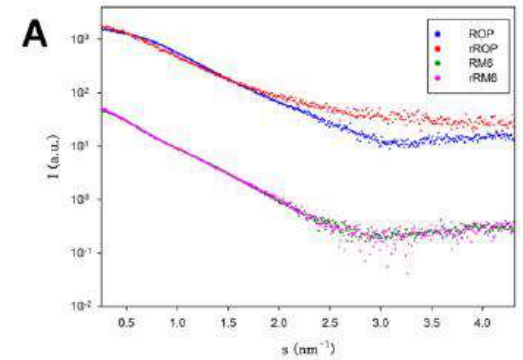
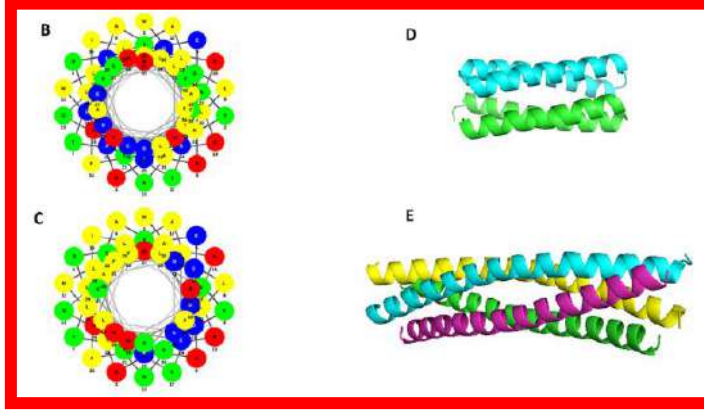
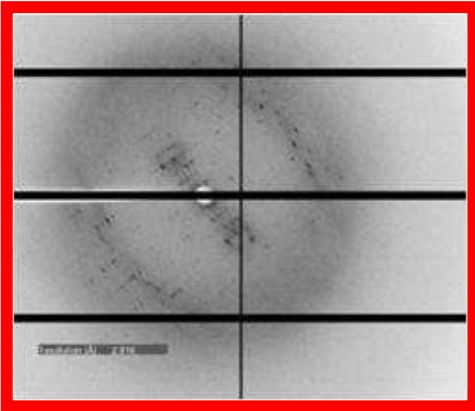
Retro
protein

```
rRM6
gi|503929373|re|MLNEGDDGFRALCSRYLEDAHDHLSSECIDALENLKELLTQTQSRIFRAMNLATKEQKTM
gi|931383334|gb|VLYKGDSDMIKPCAVYFRDYYDHTVEIIDIVETYREMAS
gi|931366270|gb|-----YLRDYYDHTIOVIDTIVETRYDMLM
gi|517782630|re|-----FLRDYYDHTIOVIDTIESLRDMLM
gi|501781185|re|LMREGTALFDAGTMPYLRDYYDHTVHIMDLLESYREMAS
LMREGTTLFDAGTLPYLRDYYDHTVHIMDLLESYREMAS
cons      ::*::**  :* :*  :::
```

```
rRop
g1|769156590|re|MLNEGDDGFRALCSRYLEDAHDHLSSECIDAQEDALENLKELLTQSRIFRAMNLATKEQKTM
g1|291075667|gb|LLVEHNKIYRALCSRHVVEEAKOAMREHIDNOEITVLKNIKE
g1|524623934|em|LLVEHNKIYRALCSRHVVEEAKOAMREHIDNOEITVLKNIKE
g1|569410540|gb|-----RHVSKVKDHSVKIMAVWEDVLEGTRHLDLSILLEAVFRAFNLK
g1|295092621|em|LLVEHNKIYKALCSRHVVEEAKLAMREHIDNOEITVLKNIKE
cons      *:::..  : : : *  * : . :
```



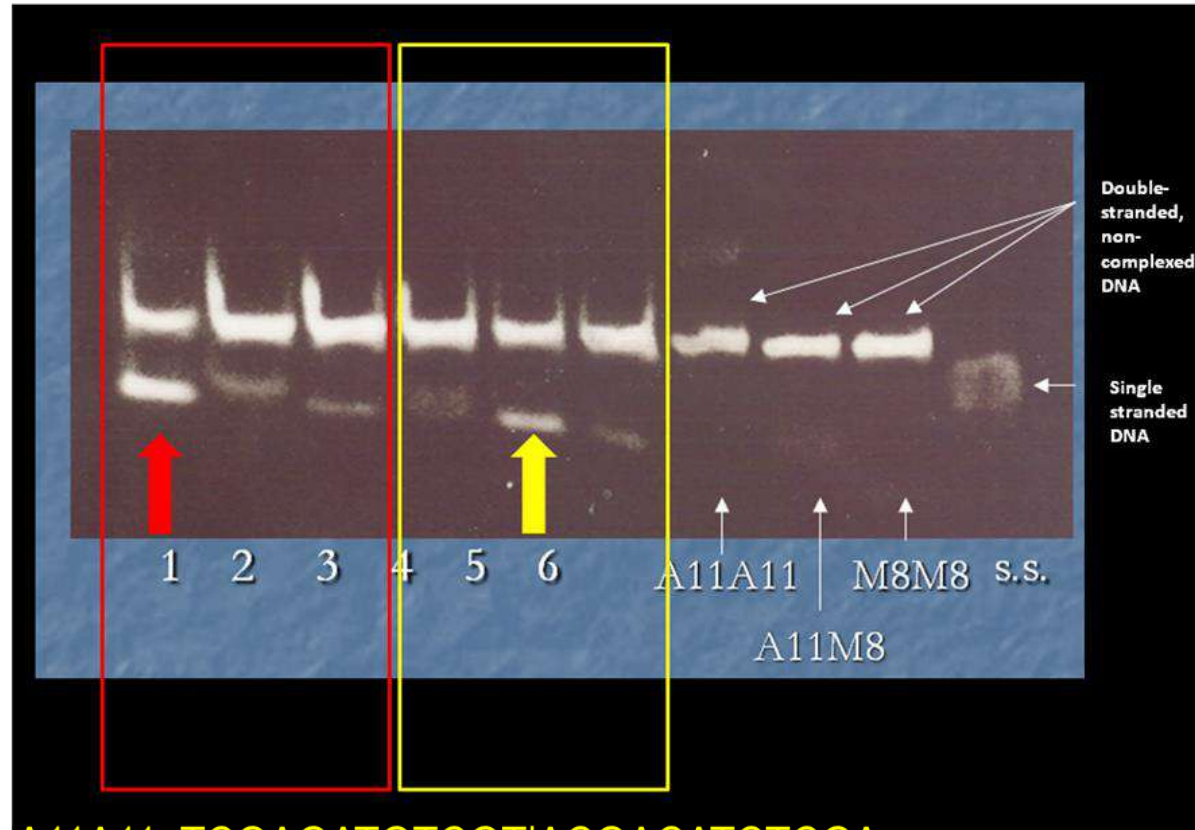
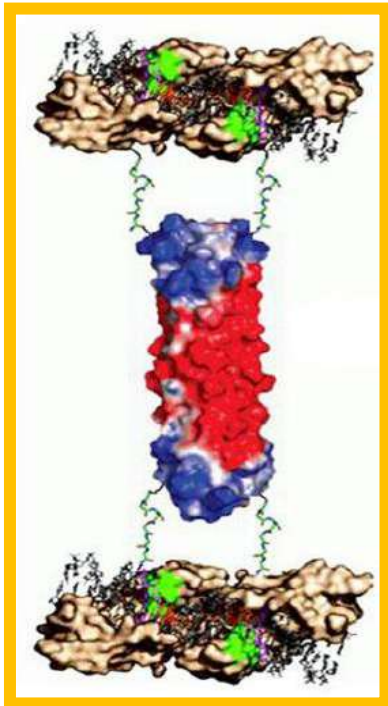
Probing Protein Folding with Sequence-Reversed α -Helical Bundles



Πρωτεϊνικές αλληλουχίες & Βιοτεχνολογία

Στην Βιοτεχνολογία διακρίνουμε δύο βασικές κατευθύνσεις που αξιοποιούν τις σχέσεις μεταξύ δομής-αμινοξικής αλληλουχίας-δομής στις πρωτεΐνες: την **πρωτεϊνική μηχανική**, δηλαδή την μεταλλαγή του γονιδίου μιας υπάρχουσας πρωτεΐνης σε μια προσπάθεια να τροποποιήσουμε την λειτουργία της με ένα προβλέψιμο τρόπο, και τον **σχεδιασμό πρωτεϊνών**, που έχει τον πιο φιλόδοξο σκοπό να σχεδιάσει de novo μια πρωτεΐνη που να εκτελεί μια επιθυμητή λειτουργία.



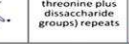










Programmable meganucleases via helical scaffolds



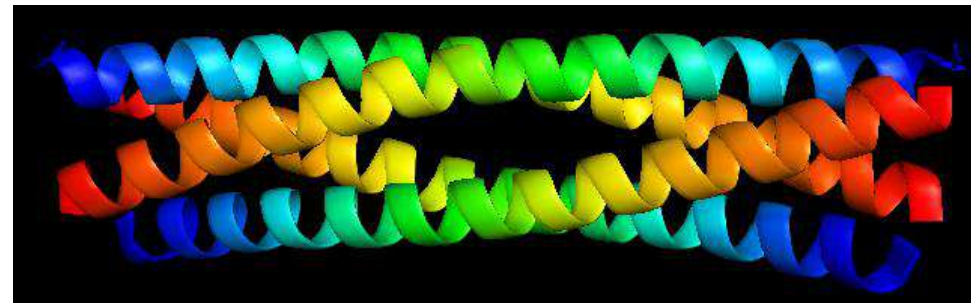
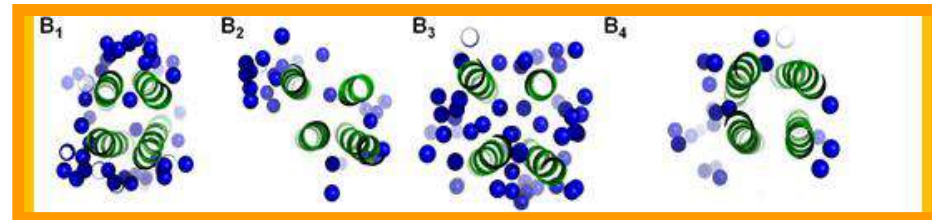
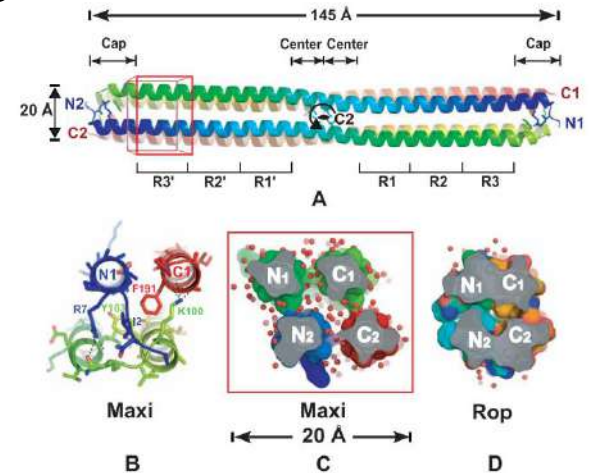
A11A11 TCGAGATGTCGT|ACGACATCTCGA
TCGAGATGTCAC ACAGAGGTACGA
M8M8 TCGTACCTCTGT|ACAGAGGTACGA

A heart of ice: Engineering of antifreeze proteins based on helical bundles

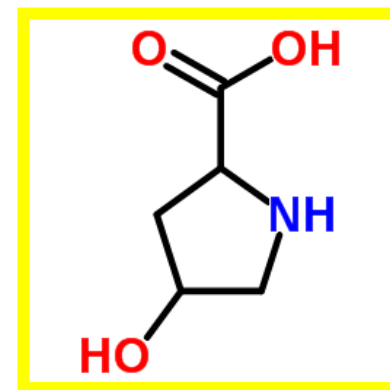
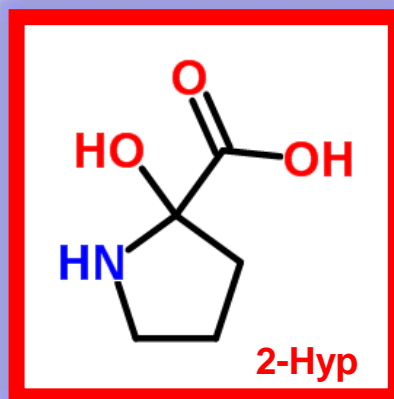
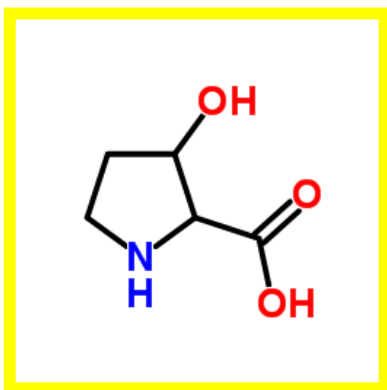
Antifreeze proteins are a class of proteins that adsorb to the surface of ice crystals to prevent their growth

AFP (Type I)	AFP (Type II)	AFP (Type III)	AFGP	Hyperactive
Alanine-rich α -helical	Cysteine-rich globular	Globular	(Alanine-alanine-threonine plus disaccharide groups) repeats	β -helical
				
Crystal planes of hexagonal ice (Ih)				
				
Binding plane				
Binding Plane				
	AFGP	Type I	Type II	Type III
				Insect AFP

Inspired from Maxi protein which prevents the blood of winter flounder (*Pseudopleuronectes americanus*) from freezing, an effect of its water-filled core. This water structure sticks outside the protein, where it appears to bind to ice. Science 343, 795–798 (2014)



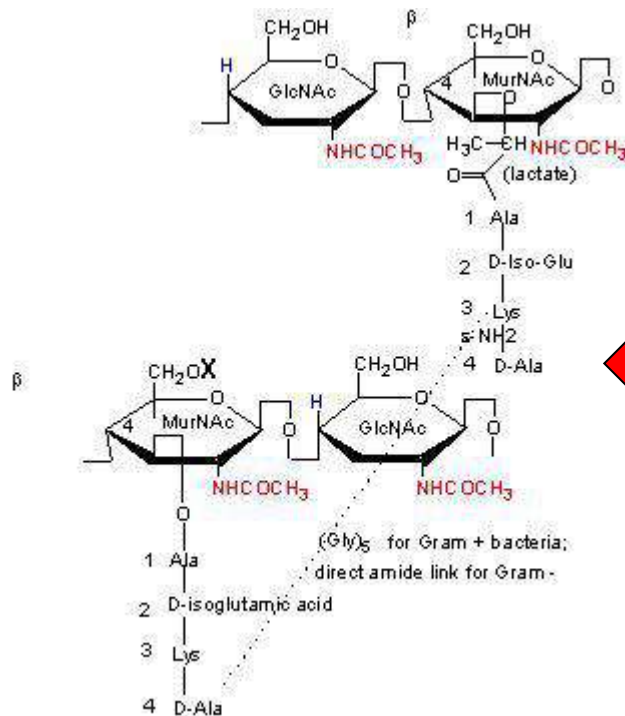
B) The autocatalytic hydroxylation of the C_α atom



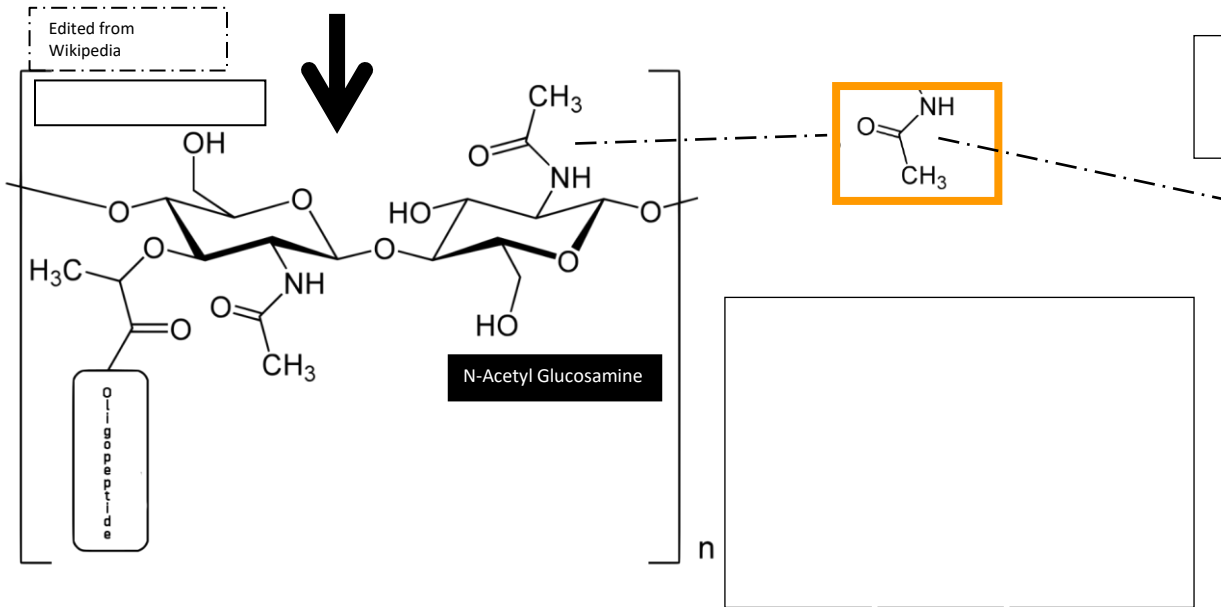
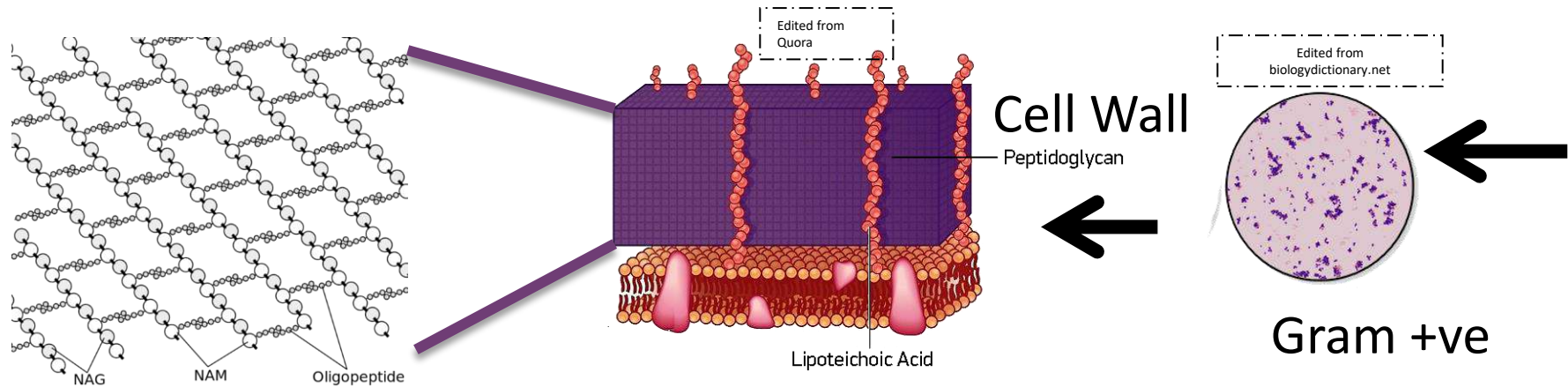
Polysaccharide deacetylases (PDAs) from *B.cereus* and *B.anthraxis*

<i>B. cereus</i> ATCC 14579	<i>B. anthracis</i> st. Ames	Possible function	Identity	Similarity
NP_831730 (275) (BC1960)	NP_844369 (275) (BA1961)	Peptidoglycan GlcNAc deacetylase	94	97
NP_833348 (213) (BC3618)	NP_845942 (213) (BA3679)	Peptidoglycan GlcNAc deacetylase	97	100
NP_832677 (275) (BC2929)	NP_845280 (275) (BA2944)	Peptidoglycan GlcNAc deacetylase	94	97
NP_834868 (245) (BC5204)	NP_847604 (245) (BA5436)	Peptidoglycan GlcNAc deacetylase	93	96
NP_831744 (273) (BC1974)	NP_844383 (273) (BA1977)	Peptidoglycan GlcNAc deacetylase	98	99
NP_830306 (260) (BC0467)	NP_842967 (273) (BA0424)	Peptidoglycan MurNAc deacetylase	98	99
NP_830050 (254) (BC0171)	NP_842717 (254) (BA0150)	Chitooligosaccharide deacetylase	95	99
NP_831543 (234) (BC1768)	NP_844255 (234) (BA1836)	Chitooligosaccharide deacetylase	92	96
NP_833526 (299) (BC3804)	NP_846187 (299) (BA3943)	Chitooligosaccharide deacetylase	95	97
NP_830200 (360) (BC0361)	NP_842877 (360) (BA0330)	PDA	91	94
NP_830200 (360) (BC0361)	NP_842878 (367) (BA0331)	PDA	53	69

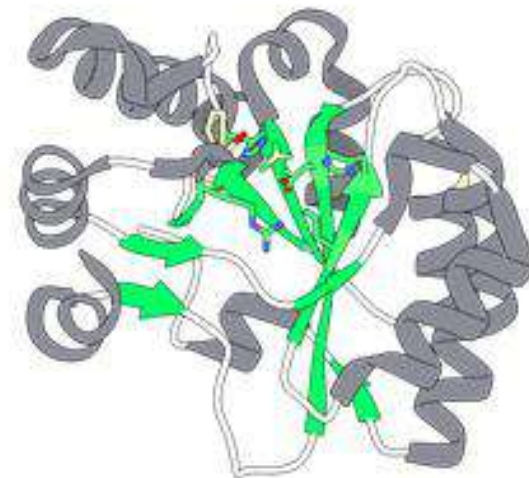
- All these enzymes share a **universal conserved region** called **polysaccharide deacetylase domain** (according to the Henrissat classification). All five members of this family catalyze the **hydrolysis** of either ***N*-linked acetyl group** from ***N*-acetylglucosamine residues** (chitin deacetylases, NodB and peptidoglycan *N*-acetylglucosamine deacetylases), or ***O*-linked acetyl groups** from ***O*-acetylxylose residues** (acetyl xylan esterases, xylanases).
- **Peptidoglycan modification**, specifically *N*-deacetylation, is a **highly efficient strategy** used by **pathogenic bacteria to evade innate host defenses**. For example, de-*N*-acetylation of peptidoglycan GlcNAc confers resistance to lysozyme, an exogenous muramidase, upon several bacterial species, such as *S. pneumoniae*, *Bacillus cereus*, *L. monocytogenes*, *Lactococcus lactis* and *Helicobacter pylori*.



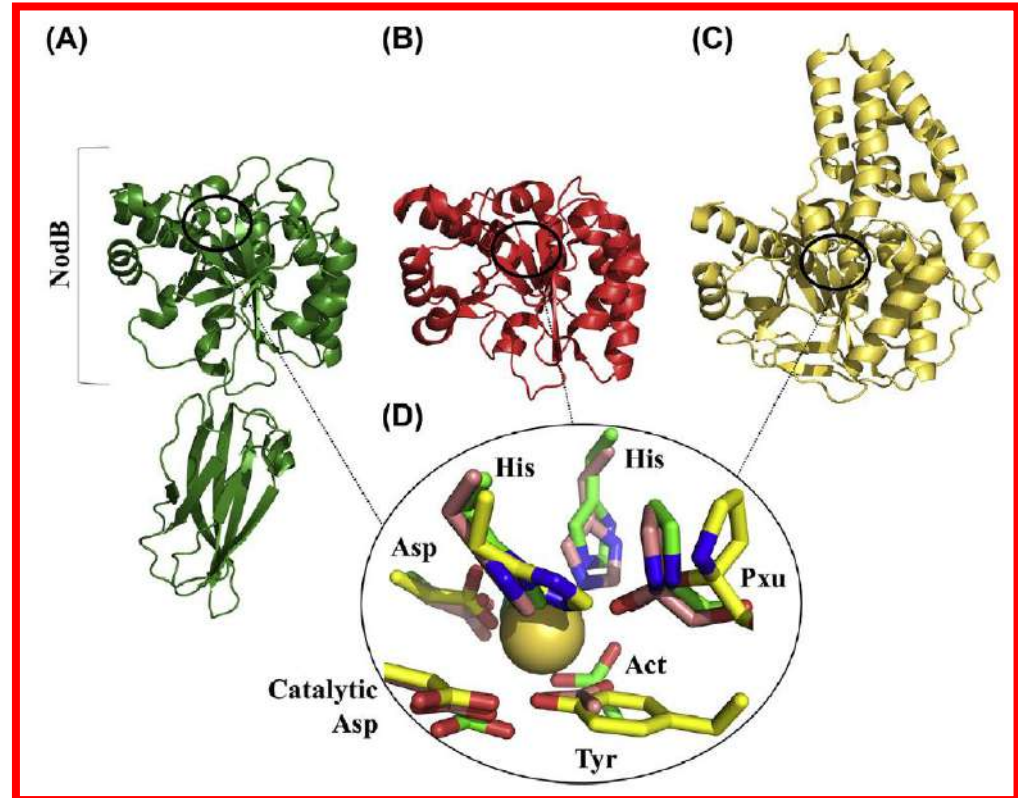
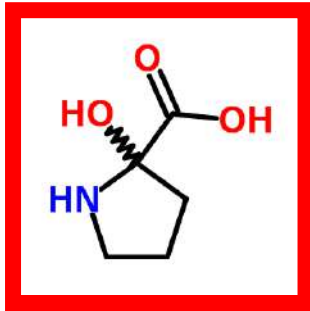
Peptidoglycan N-deacetylation occurs at N-linked acetyl groups of GlcNAc or MurNAc (coloured in red)



Make bacteria invisible to the immune system



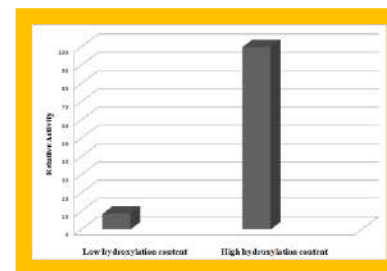
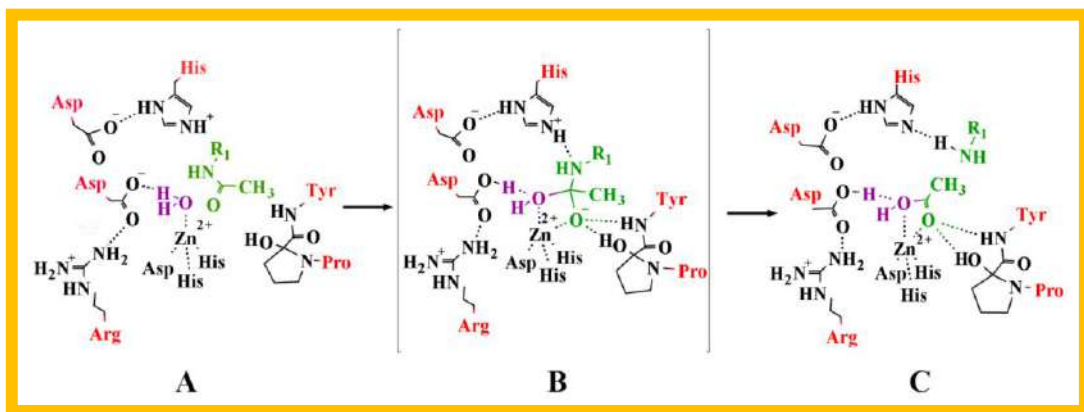
A new form of hydroxyproline (2-Hyp) is a frequent occurrence in active sites of PDAs and is associated with conserved sequence motifs



Kokkinidis *et al*, Adv. Prot.
Chem. Struct. Biol., **2020**

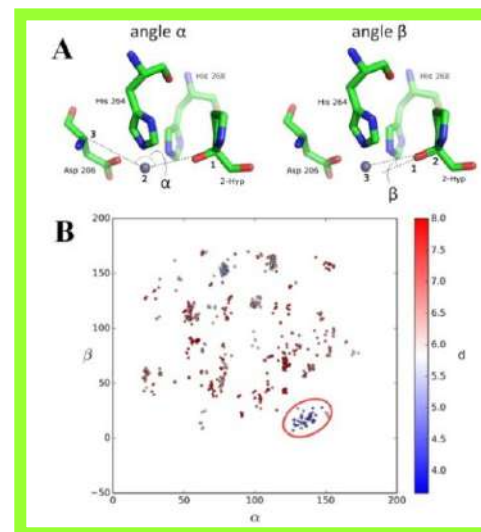
Pro C_α-hydroxylation is an active site maturation process

- The origin of the –OH group of 2-Hyp is molecular oxygen.
- The Pro→2-Hyp conversion is autocatalytic, highly specific, and occurs partially.
- Pro C_α hydroxylation is functionally intertwined with the deacetylation reaction, the two processes share the same active site and one key/catalytic (Asp) residue.
- The introduction of the additional –OH group in the active site via the Pro→2-Hyp conversion represents an active site maturation event which enhances 10x the deacetylation activity.



Fadoulglou *et al*, JACS, 2017

Extent of the autocatalytic C_α-hydroxylation in other systems? →



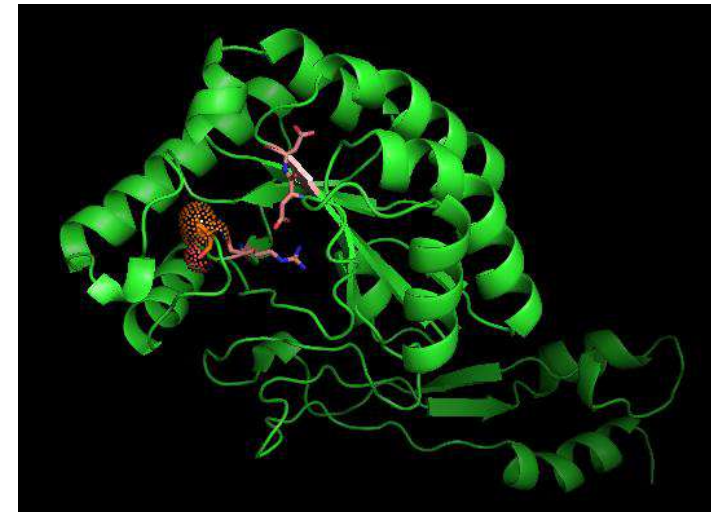
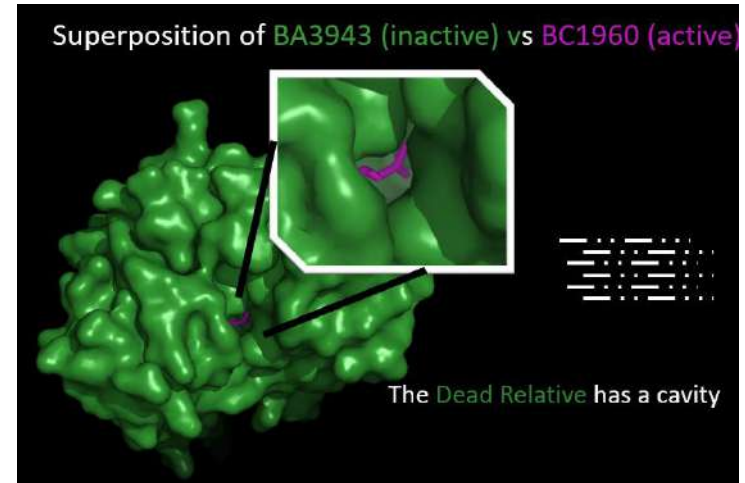
Pseudoenzymes from *B.cereus* & *B.anthraxis*

(Two deaths followed by resurrection)

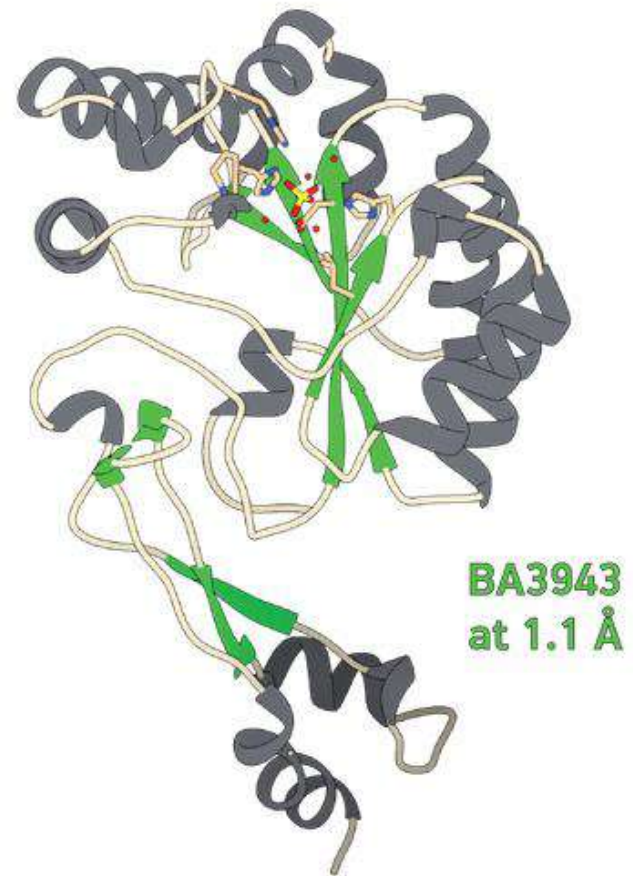
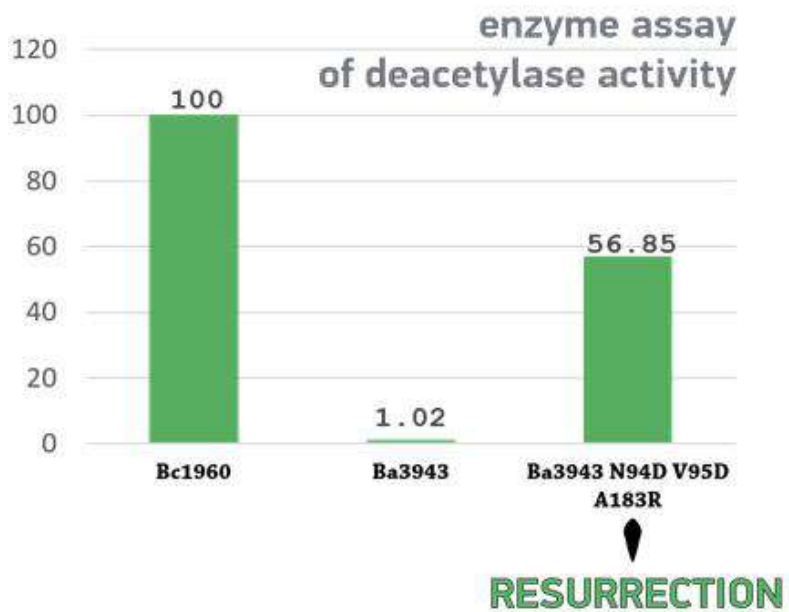
BC1960	77	LTFDDG	128	IGNHTYSHP	165	PKFIRPXYG
BC1974	73	LTFDDG	123	VGMHSMTHN	160	PKLTRPPYG
BA0330	202	VTFDDG	261	MQSHTATHA	296	VIAVAYXFG
BA0330 D205A	202	VTFADG	261	MQSHTATHA	296	VIAVAYXFG
BA3943	91	LTINVA	141	VGNHSYTHP	177	VRWFAPPSG
BA3943 N94D	91	LTIDVA	141	VGNHSYTHP	177	VRWFAPPSG
BA3943 N94D V95D	91	LTIDDA	141	VGNHSYTHP	177	VRWFAPPSG
		Motif 1		Motif 2		Motif 3

Ba3943: Motif 1 is corrupted and a sizeable cavity is located in the interior of the protein. No deacetylation/ C_α hydroxylation activities

Rolling back evolution (?): A triple mutant (motif 1 restored, cavity “filled” with Arg), is “resurrected” (as PDA) with self-hydroxylation & deacetylation activity



The Dead RELATIVE



- Thank you

ΕΡΓΑΣΙΑ

Knowledge about protein structure incorporated in AlphaFold

*Jumper et al., Nature | Vol 596 | 26 August 2021
/ 583*

- **3 pages max**

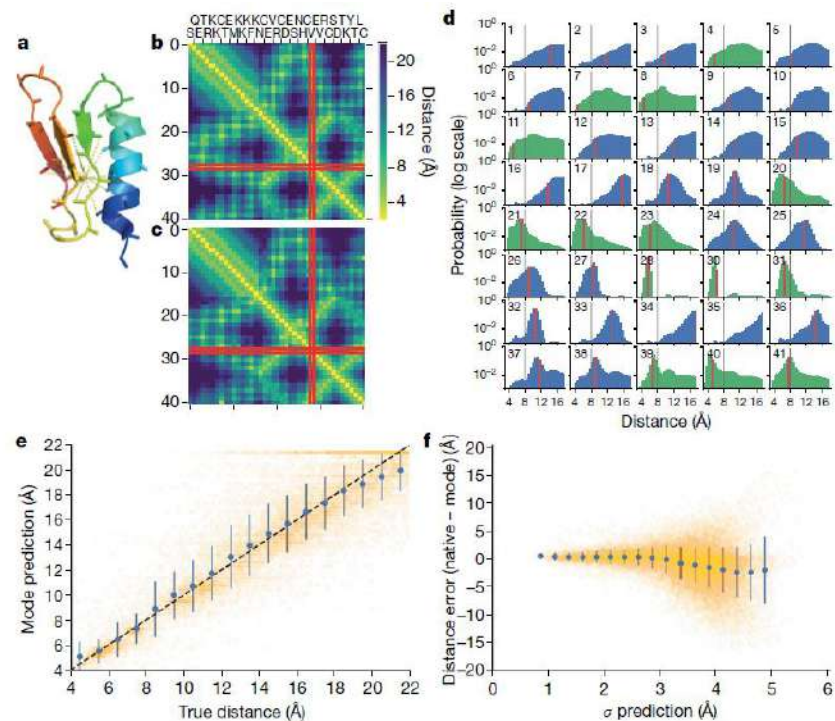


Fig. 3 | Predicted distance distributions compared with true distances.

a–d, CASP target T0955, $L=41$, PDB 5W9F. a, Native structure showing distances under 8 Å from the C_{α} of residue 29. b, c, Native inter-residue distances (b) and the mode of the distance predictions (c), highlighting residue 29. d, The predicted probability distributions for distances of residue 29 to all other residues. The bin corresponding to the native distance is highlighted in red, 8 Å is drawn in black. The distributions of the true contacts are plotted in green, non-contacts in blue. e, f, CASP target T0990, $L=552$, PDB 6N9V.

e, The mode of the predicted distance plotted against the true distance for all residue pairs with distances ≤ 22 Å, excluding distributions with s.d. > 3.5 Å ($n=28,678$). Data are mean \pm s.d. calculated for 1 Å bins. f, The error of the mode distance prediction versus the s.d. of the distance distributions, excluding pairs with native distances > 22 Å ($n=61,872$). Data are mean \pm s.d. are shown for 0.25 Å bins. The true distance matrix and histogram for T0990 are shown in Extended Data Fig. 2b, c.