Proteins

Eugene Shakhnovich
Harvard University
Boulder, CO July 25-27 2012

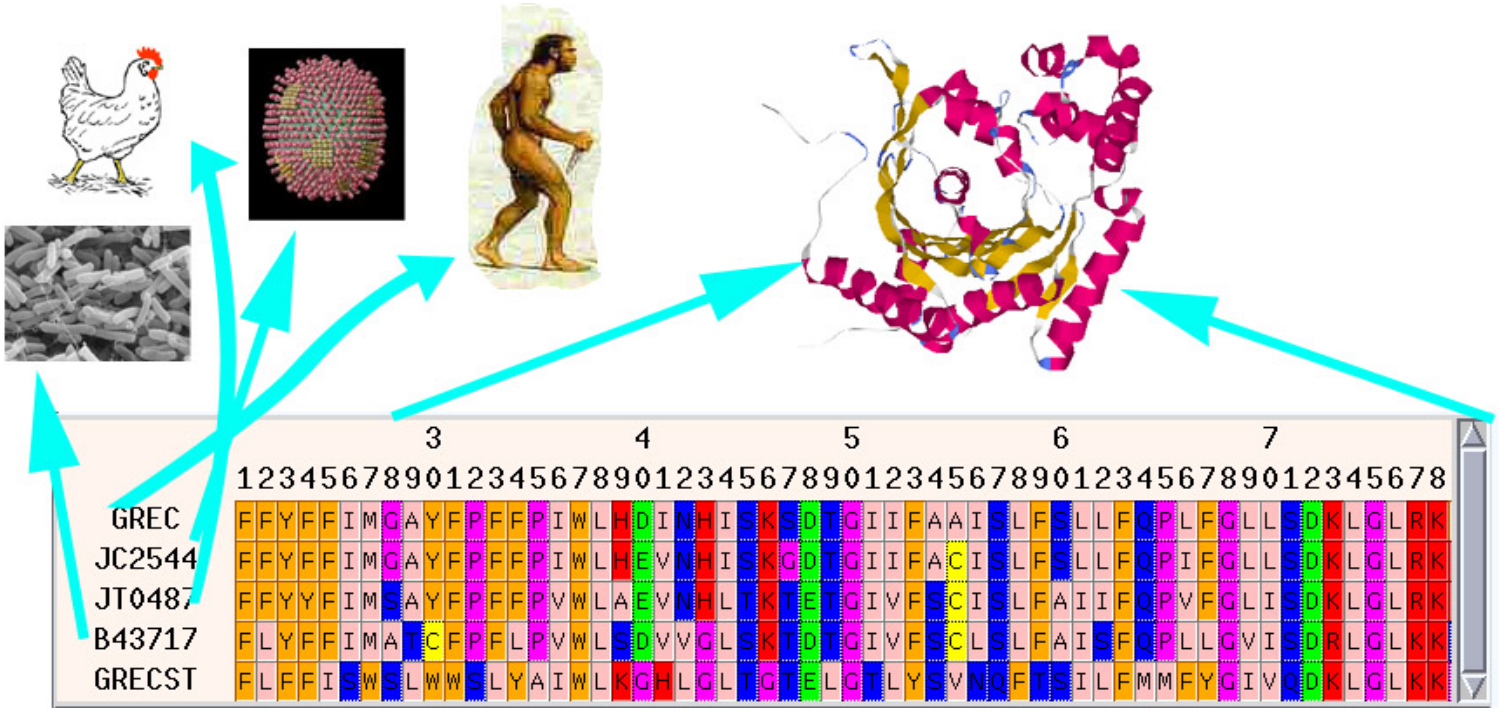
Outline

Lecture 1. Introduction: Principles of structural organization of proteins and Protein Universe.
Concepts of convergent and divergent molecular evolution.

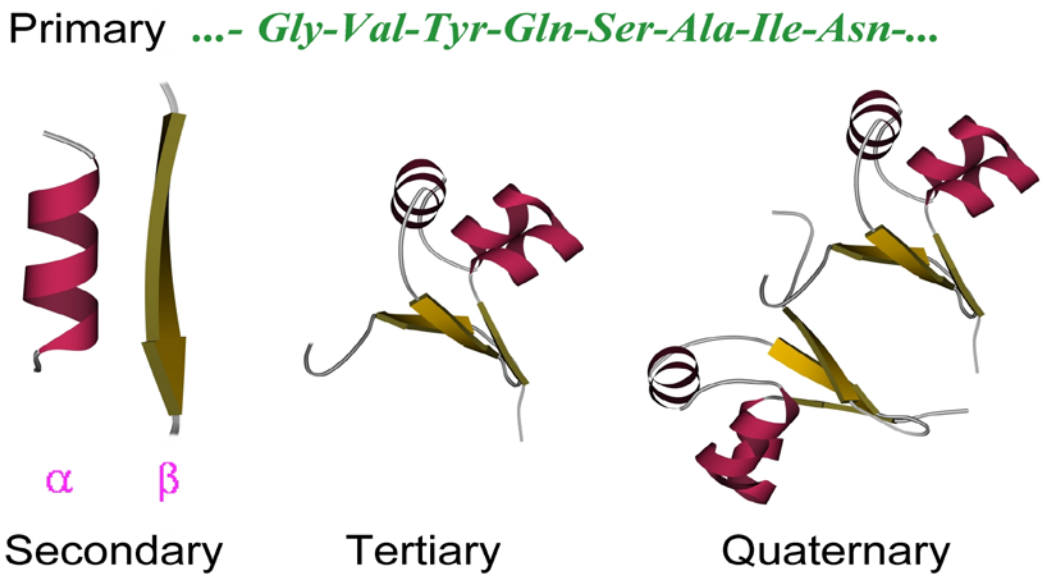
Lecture 2. Protein folding – Experiment and Theory and Simulations

Lecture 3: Bridging scales: From Protein Biophysics to Population Genetics (‘’From Crick to Darwin and back’’)

Marriage made in heaven: genomics and protein folding/evolution

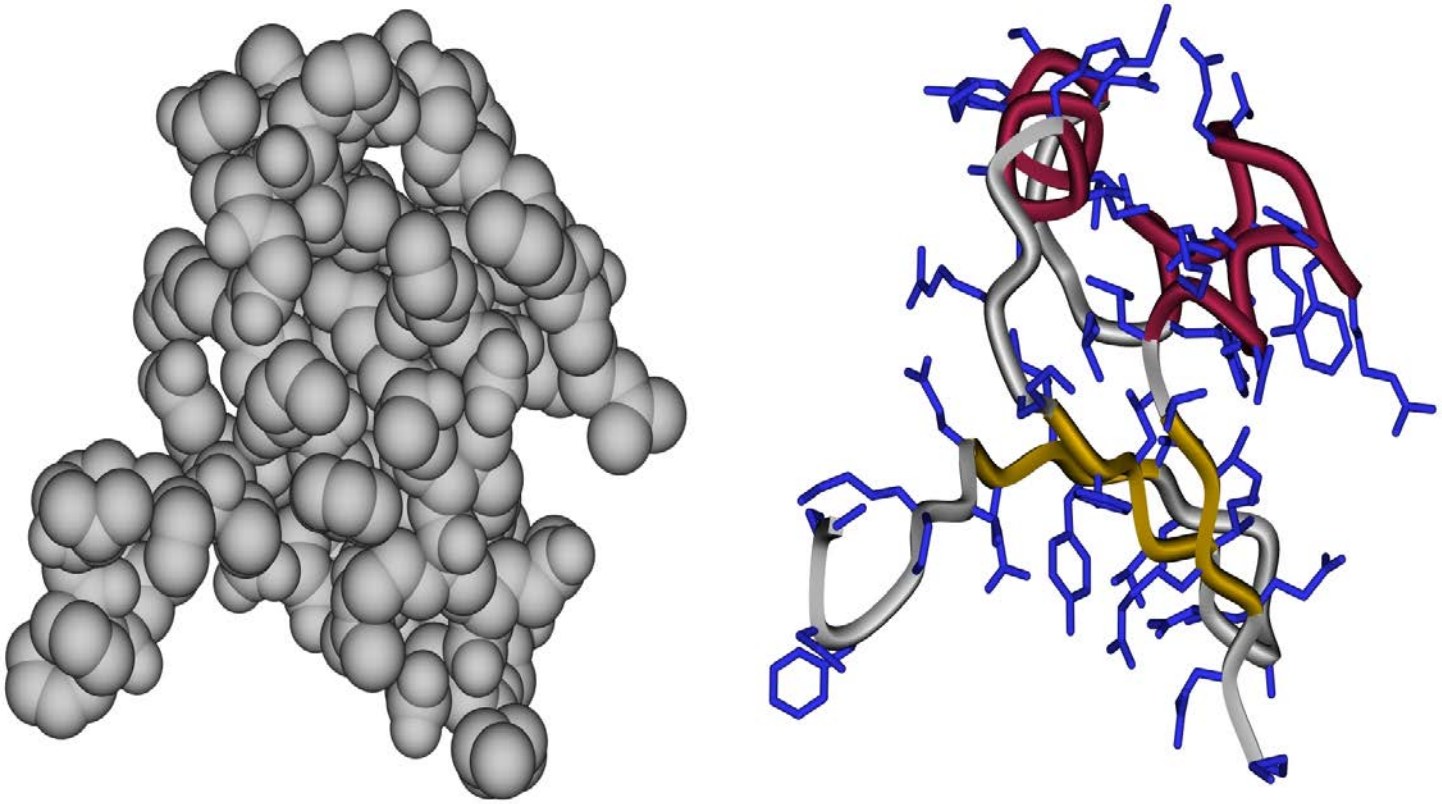


Proteins are folded on various scales

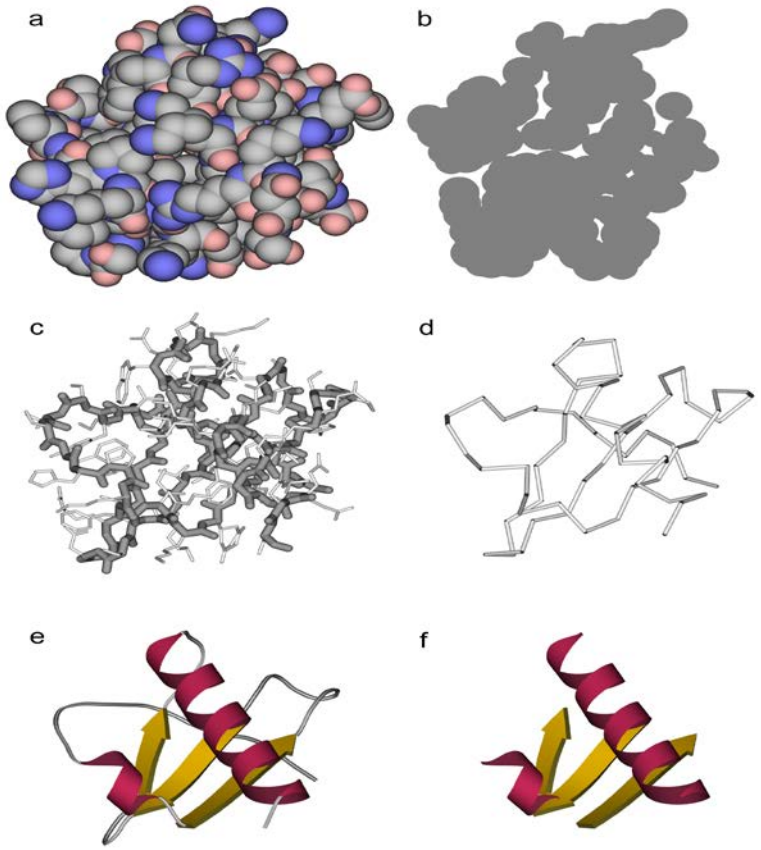


As of now we know hundreds of thousands of sequences (Swissprot) and a few thousand of Structures (protein data bank)

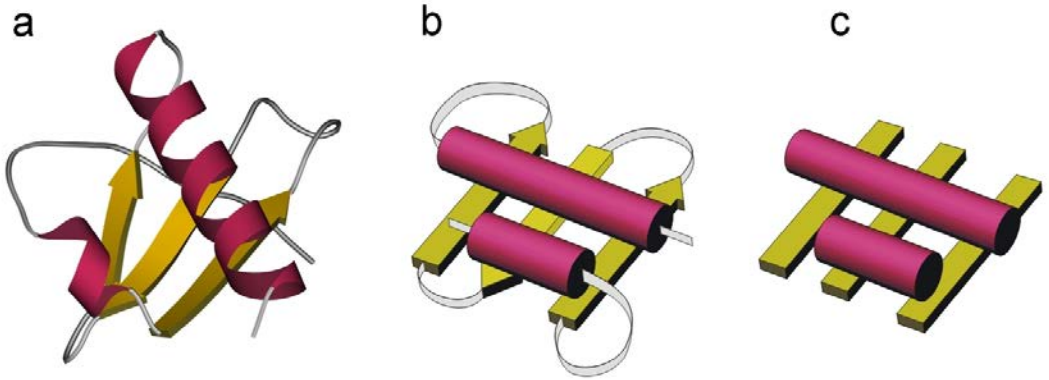
Proteins are tightly packed



Structural organization of globular proteins

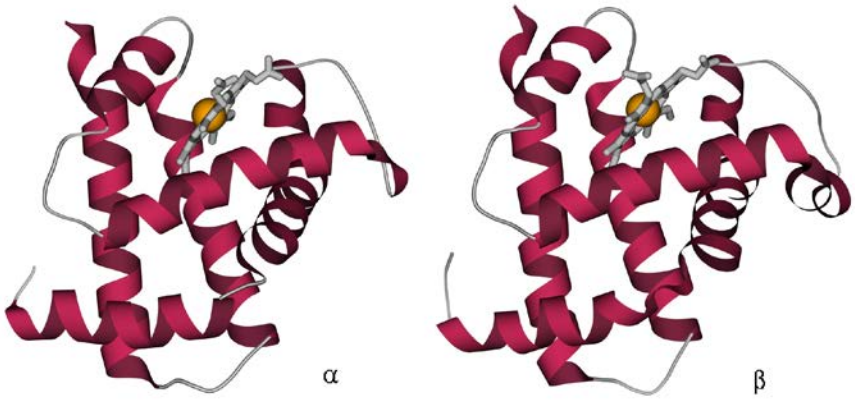


Fold, Folding Pattern and Packing



Simplified models of protein structures. a), A detailed fold describing positions of secondary structures in the protein hain and in space (see also Fig. 13-1e).
(b), The folding pattern of the protein chain with omitted details of loop pathways, the size and exact orientation of α-helices (shown as cylinders) and β-strands (shown as arrows). (c), Packing: a stack of structural segments with no loop shown and omitted details of size, orientation and direction of α-helices and β-strands (which are therefore presented as ribbons rather than as arrows).

Similarities at the fold level, differences at the detailed structural level

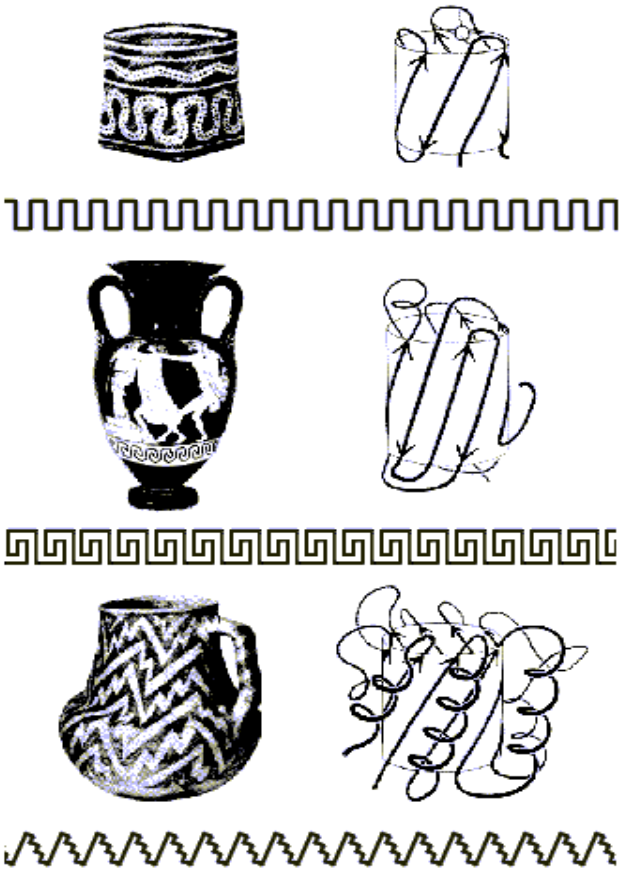


Two close relatives: horse hemoglobin α and horse hemoglobin β (both possessing a heme shown as a wire model with iron in the center). Find similarities and differences. (For tips: they are highly similar although have some differences in details of loop conformations, in details of orientation of some helices, and in one additional helical turn available in the β globin, on the right).

Fold databases group many proteins into fold families

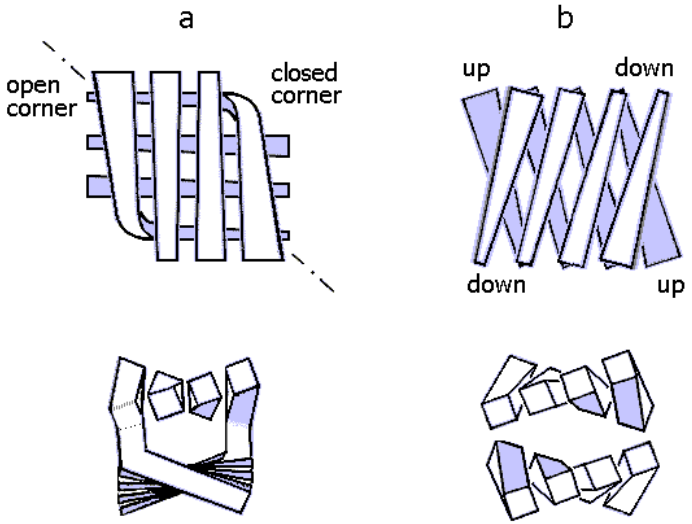
SCOP – Murzin, Chothia et al
CATH – Orengo and Thornton
FSSP – Holm and Sander

Common motives in protein ‘’topologies’’



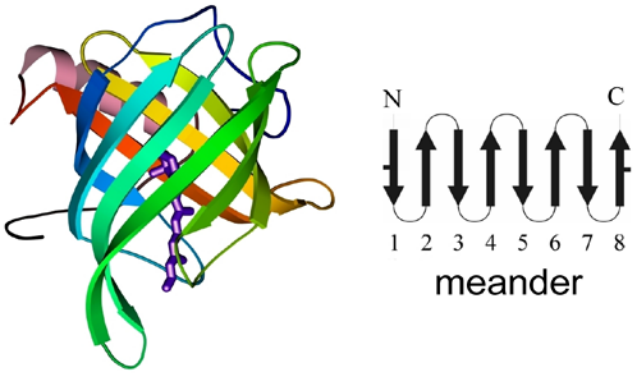
Folding patterns of protein chains and ornaments on American Indian and Greece pottery: two solutions of the problem of enveloping a volume with a non-self-intersecting line. On top, the meander motif; in the center, the Greek key motif; at the bottom, the zigzag “lightning” motif. Reprinted with permission from the cover of Nature, v.268, No.5620, 1977 (© 1977, Macmillan Magazines Limited), where a paper by J. Richardson on folding patterns of protein chains is published.

Strands packing in β-sheet

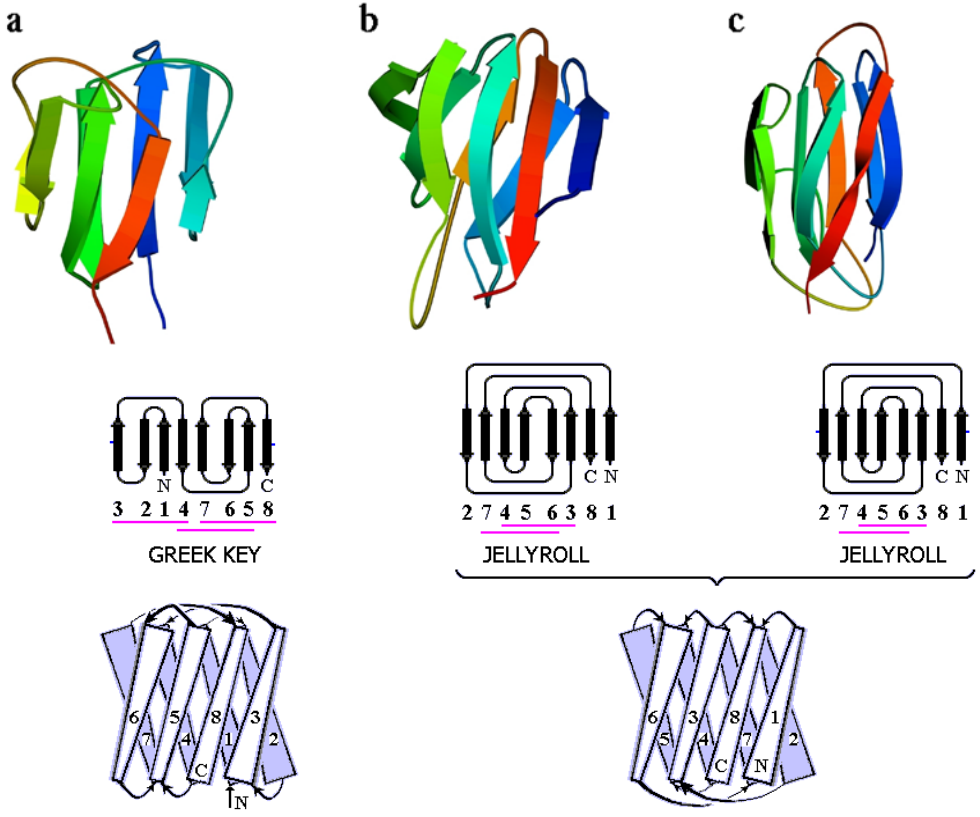


The orthogonal (a) and aligned (b) packing of β-sheets viewed face on (above) and from their lower end. At the face view, β-strands are wider when approaching the reader. The dashed line shows the axis of the orthogonal β-barrel to which both "open" corners belong. Here the two β-sheets are most splayed. At the two "closed" corners the sheets are extremely close to each other; here the chain bends and passes from one layer to the next. In the orthogonal packing the hydrophobic core is almost cylindrical. In contrast, in the aligned packing, the core is flat, the distance between the twisted sheets remains virtually unchanged, and the relative arrangement of the sheets allows hydrophobic faces of twisted β-strands to contact over a great length. Adapted from C.Chothia & A.V.Finkelstein, Annu. Rev. Biochem. (1990) 59:1007-1039.

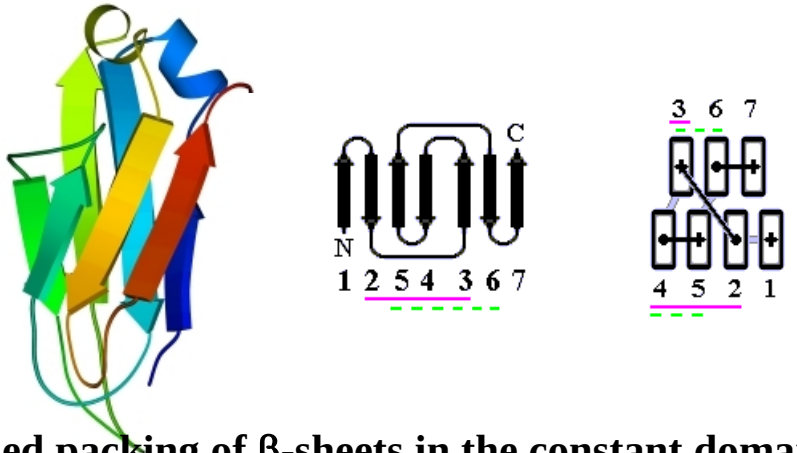
Example of orthogonal packing: Retinol-binding protein



Examples of aligned packing

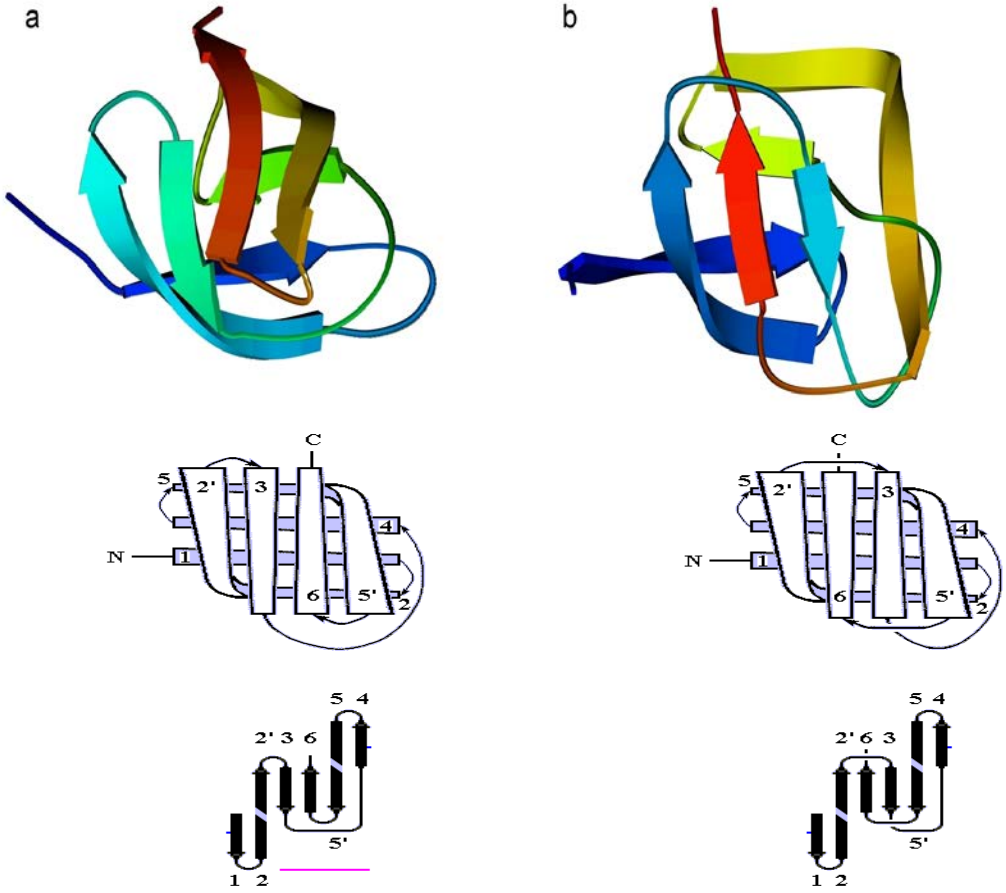


Ig-fold – the most popular β-sandwich

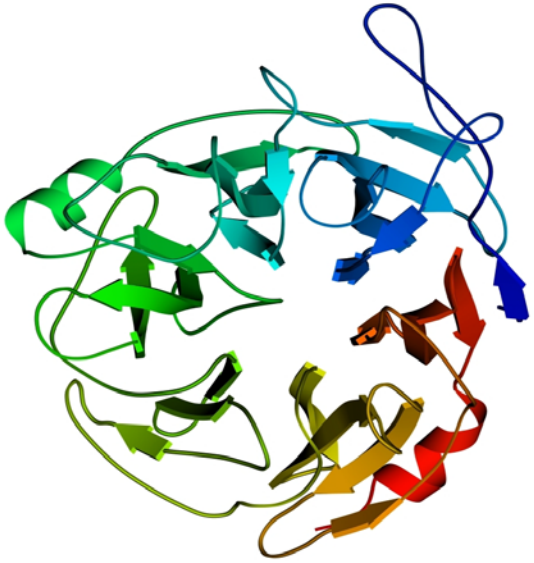


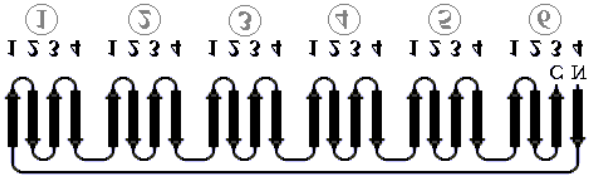
The aligned packing of β-sheets in the constant domain of the light chain of immunoglobulin . On the left, a detailed diagram of the protein is shown; the chain pathway is traced in color (from blue to red) from N- to C-terminus. The topological diagram (in the center) accentuates the "Greek keys". On the right, the protein is shown as viewed from below). The cross corresponds to the strand's N-end (i.e., "the chain runs from the viewer"), and the dot to the C-end (i.e.,"the chain comes to the viewer"). The segment-connecting loops close to the viewer are shown by black lines, and those distant (on the opposite side of the fold) by light lines. Note that such a diagram allows presenting the co-linear packing of these segments (β-strands) in the simplest possible way. Besides, it visualizes the spatial arrangements of "Greek keys" and makes evident that two of them (formed by strands 2, 3, 4, 5 and 3, 4, 5, 6, respectively) differ in their spatial arrangements.

Folding patterns in serine proteases and acid proteases



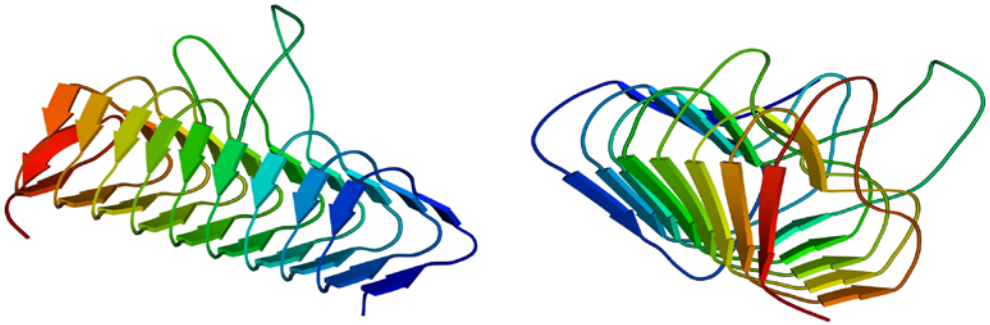
Multiple β-propellers





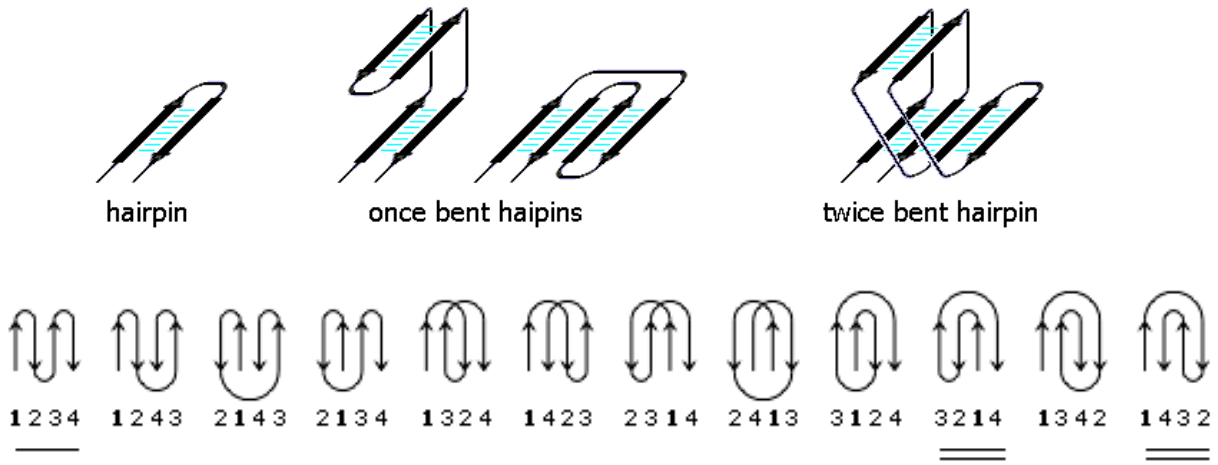
The β-structure in the form of a “six-blade propeller” in neuraminidase, and a topological diagram of this protein composed of six antiparallel β-sheets.

Proteins with all-parallel beta structure are rare but they exist



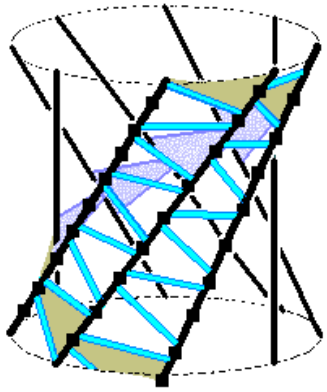
The β-prism in acyl transferase (a) and in pictatylase (b). Pay attention to handedness of the chain's coiling around the axis of the prism: it is unusual, left, in (a) and common, right, in (b). Also note that when the chain's coiling is left-handed, as in scheme (a), the common twist of the β-sheet is absent. This common twist, i.e., the right-handed (if viewed along the β-strands), propeller twist is seen in scheme (b)

Antiparallel architecture of β-proteins is often based on hairpins



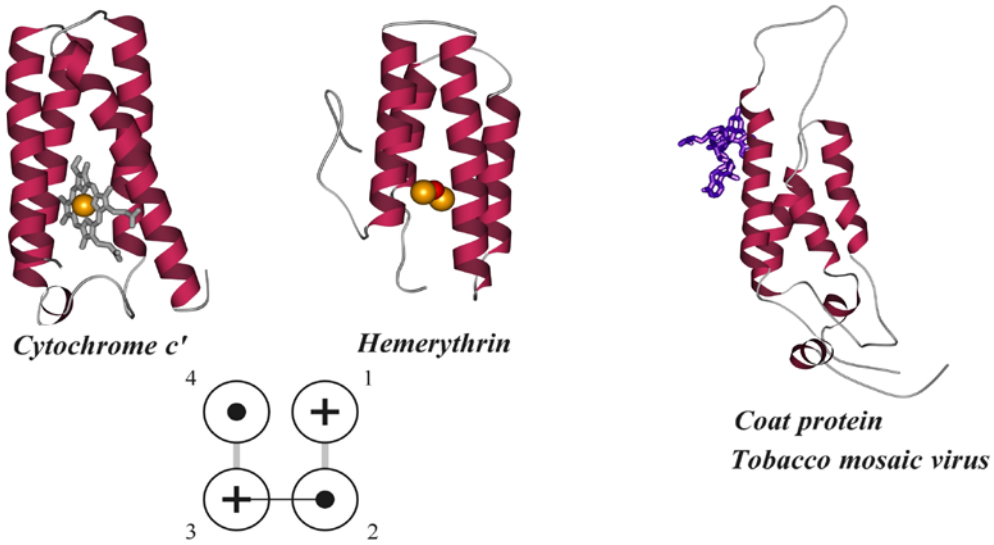
Possible topologies of sheets composed of four β-strands. The scheme includes only the sheets where two adjacent in the chain β-strands are oppositely directed. Among these, the abundant topologies are "meander" (underlined with one line) and two "Greek keys" (underlined with two lines), the latter two being different only in direction of the chain turn from the hairpin consisting of strands 1, 4 to the hairpin consisting of strands 3, 4. The "meander"-containing protein is exemplified by retinol-binding protein ; the examples of "Greek key"-containing proteins are γ crystalline or trypsin

β-cylinders



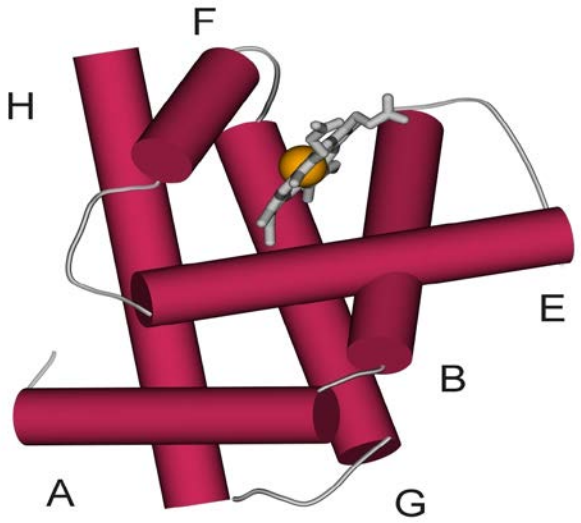
The closed β-cylinder. H-bonds (the blue lines) are shown for one strand only. One line of H-bonding is shown as a gray band. The shear number is equal to 8 in the given case.

α-helical proteins and their topologies: four helix bundle



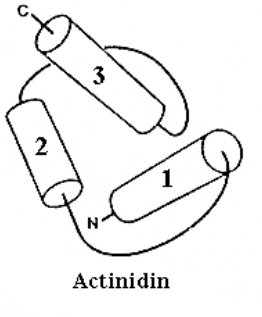
Three similar in architecture ("four-helix bundle") but different in function α-proteins: cytochrome c', mosaic virus coat protein. Both the protein chain and co-factors are shown: wire models represent the heme (in cytochrome) and an RNA fragment (in virus coat protein), orange balls are for iron ions (in the cytochrome heme and in myohemerythrin), and the red ball is for iron-bound oxygen (in myohemerythrin). The overall architecture of such "bundles" reminds the co-linear packing of β-sheets. The topological diagram (below) shows all these proteins as viewed (in the same orientation) from their lower butt-ends. The circles represent the ends of α-helices. The cross corresponds to the N-end of the segment (i.e., the segment goes from the viewer); the dot corresponds to its C-end (i.e., the segment comes to the viewer). The loops connecting the structural segments are shown by the black line (if the loop is close to the viewer) and by the light line (if it is on the opposite side of the fold). The numerals indicate the order of structural elements in the chain (from N- to C-terminus).

Globins: common α-helical proteins

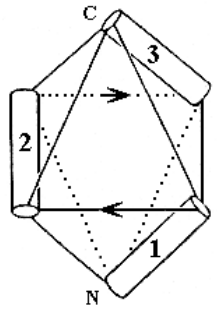


The structure of globin: crossed layers of three α-helices each. The helices A, E, and F (lettered in accordance with their sequence positions) belong to the upper layer, while H, G, and B to the lower layer. The short helices (of 1 – 2 turns each) C and D are not shown since they are not conservative in globins. A crevice in the upper layer houses the heme. Such "crossed layers" resemble the orthogonal packing of β-sheets. [The orthogonal contact of B and E helices is especially close, since both helices have the glycine-formed dents at the contact point.]

Helix packing and polyhedra models of Murzin

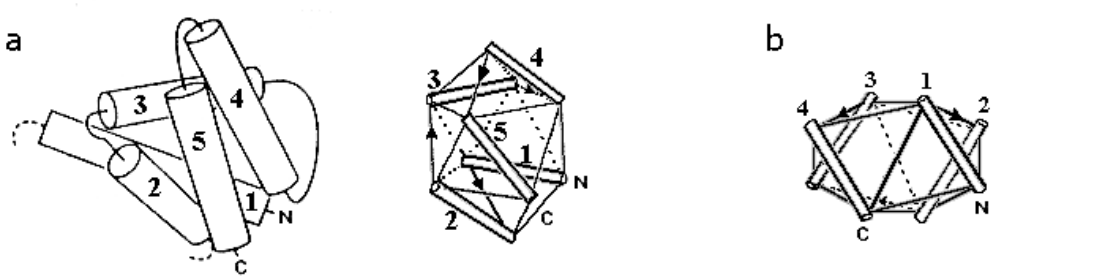


Actinidin



The α-helix positions on the ribs of a quasi-spherical polyhedron that models the N-terminal domain of actinidin. The pathways of helix-connecting loops are shown by arrows

More examples of polyhedra model packing of helixes

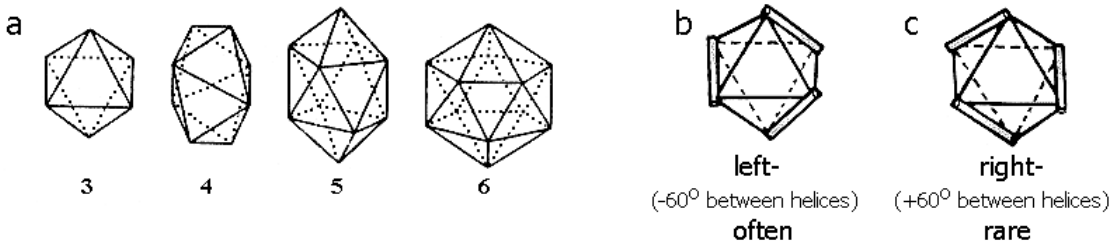


Thermolysin

Cytochrome c', Hemerythrin,...

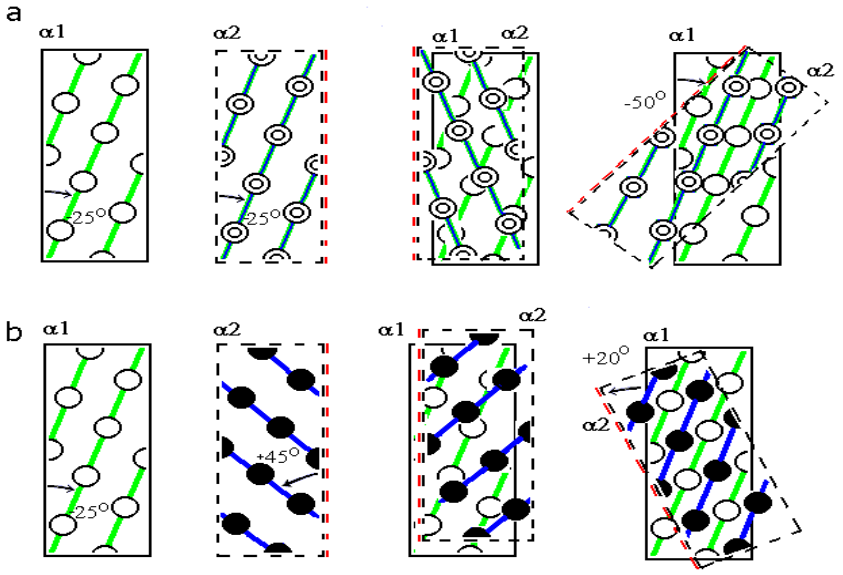
More examples showing how the geometry of helix packings in globular proteins can be described by the quasi-spherical polyhedron model. (a), the C-terminal domain of thermolysin and its model showing the helix positions on the polyhedron ribs; (b), adapted from C.Chothia, Nature (1989) 337:204-205.

The polyhedra model emphasizes the requirement of compact hydrophobic core



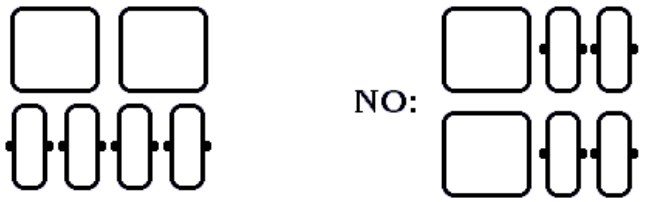
Quasi-spherical polyhedra describe the compact packing of three, four, five, and six helices. Larger assemblies of helices cannot be placed around a spherical core. Each polyhedron describes several packing arrangements, i.e., several types of "stacks" of helices; the stacks differ in helix positioning on the polyhedron ribs. For example, three helices form two different arrangements: (b), a left-handed bundle; (c), a right-handed bundle. Four helices form ten arrangements, five helices form ten arrangements, and six helices form eight arrangements ("stacks" for four-, five- and six-helix globules are not shown, but you can easily construct them by placing the helices on the polyhedral ribs in all possible ways such that each vertex corresponds to one end of a helix). The packings with inter-helical angles favorable for close helix contacts (see Fig. 14-9) are observed in proteins more often than others

Close packing of side chains imposes constraints on mutual orientation of helixes



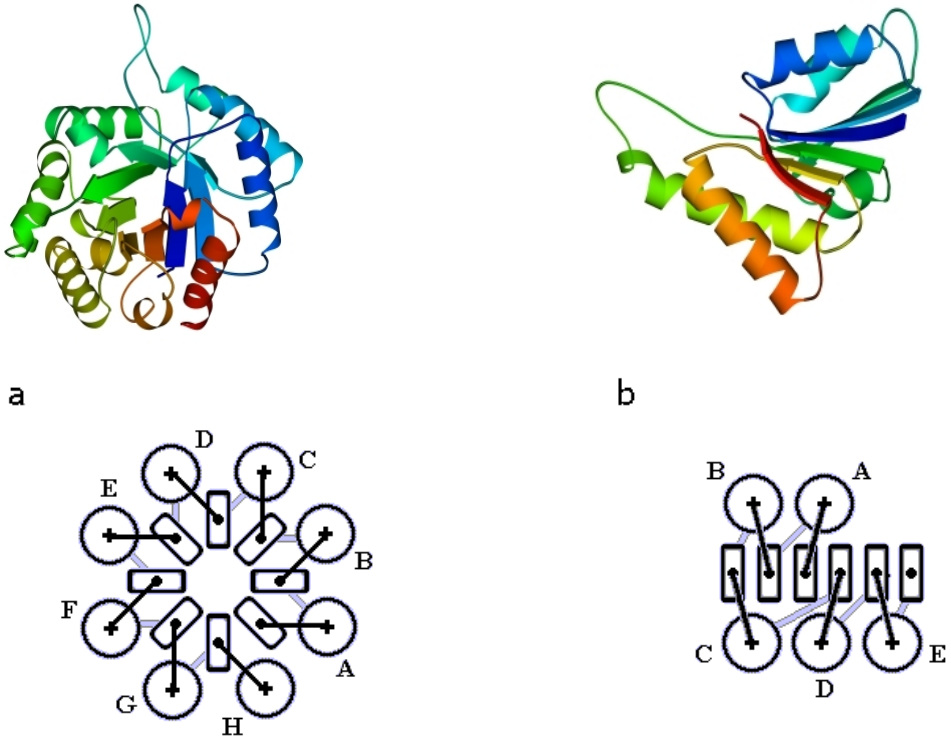
Two basic variants of close packing of side chains: with helix axes inclined at -50° (a) or +20° (b). We look at the contact area through one helix (through α2 turned over through 180° around its axis). The residues of the "lower" (α1) helix are shown as light circles and those of the upper helix (α2) by dark circles.

α/β and α+β proteins



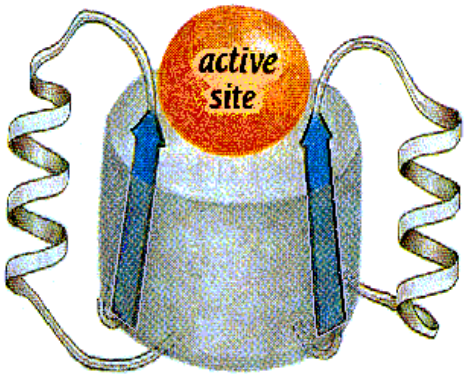
The layered structure of mixed (α/β and α+β) proteins viewed along the α-helices and β-strands to stress their close packing (helix ends are shown as squares and strand ends as rectangles). α-helices and β-strands cannot belong to the same layer because this would cause dehydration of H-bonds at the β-sheet edge (H-bond donors and acceptors in the β-sheet are shown as dots).

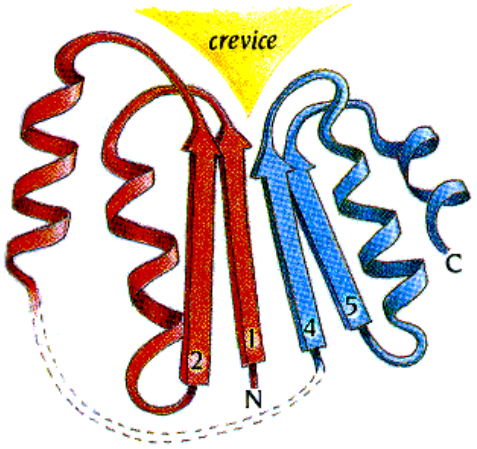
Most popular: TIM Barrel and Rossman folds



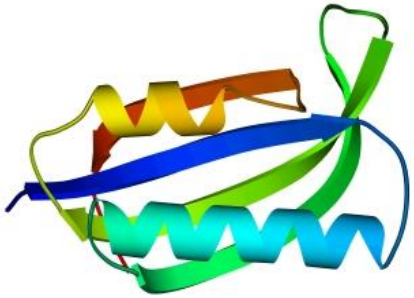
Typical folding patterns of α/β proteins and their simplified models as viewed from the β-layer bottom-end: the "α/β-cylinder" in triose phosphate isomerase (a); the "Rossmann fold" in the NAD-binding domain of malate dehydrogenase (b). The detailed drawing of the former shows a viewer-facing funnel formed by rosette-like loops and directed towards the center of the β-cylinder. The latter has a crevice at its upper side; the crevice is formed by loops going upwards and downwards.

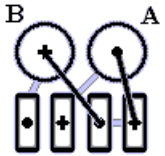
Most TIM-barrels and Rossman fold peptides are enzymes. They feature supersite located in the loop region





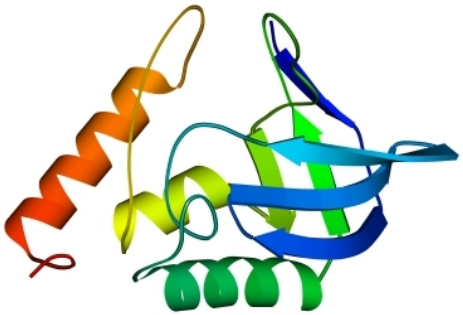
α+β proteins: αβ-plaits

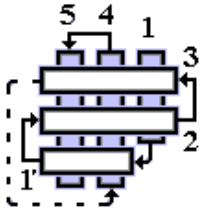




A typical structural motif for α+β proteins: the αβ-plait in the ribosomal protein S6. The αβ-plait is distinct for a more regular alternation of secondary structures in the chain as compared with the other α+β proteins (in this case, the alternation is βαββαβ). S6 represents an example of the so-called "ferredoxin fold". The rainbow coloring (blue-green-yellow-orange-red) traces the pathway of the chain from N- to C-terminus. On the right, a schematic diagram of this protein as viewed along its almost co-linear structural elements. The helices are lettered. An α- or β-region going from the viewer (i.e., viewed from ts N-terminus) is marked with "+", and that approaching the viewer with a dot.

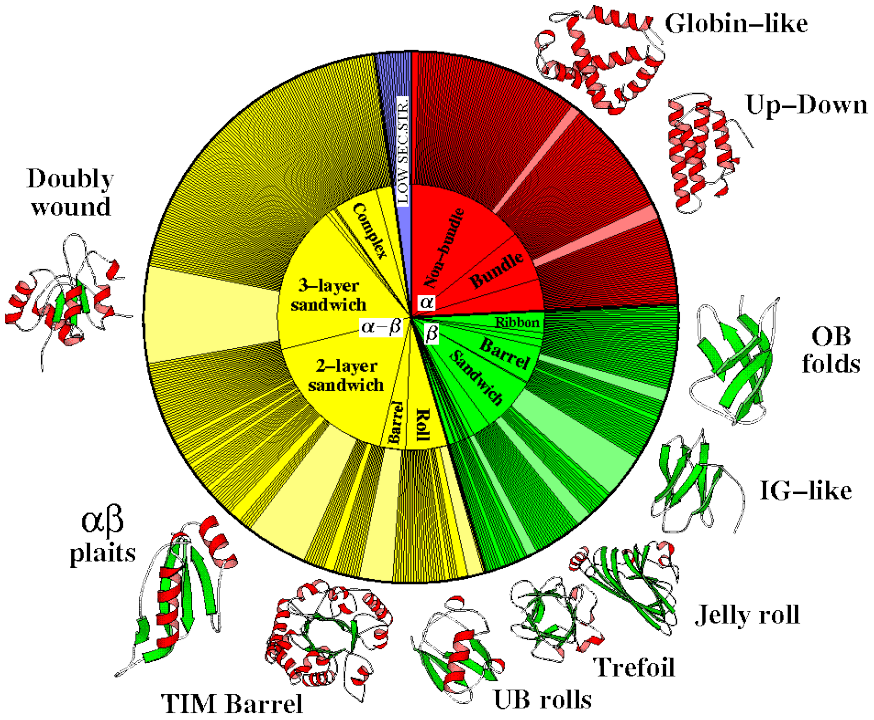
α+β proteins: ‘’Russian doll’’ effect in staphylococcal nuclease





A typical structural motif for α+β proteins: staphylococcus nuclease. This “usual” α+β protein is characterized by a less regular (as compared with α/β proteins or αβ-plaits) alternation of secondary structures in the chain (in this case, ββββαββααα), and these α and β structures are more separated in space. The folding pattern observed in the β-sub-domain of the nuclease is called "OB-fold" (i.e., "Oligonucleotide-Binding fold"). On the right, a schematic diagram of the OB-fold (the orthogonal packing of β-strands is viewed from above) that is abundant in various multi- and mono-domain proteins. The β-strands are marked with numerals. The first strand is bent (actually, it is broken); its two halves are marked as 1 и 1'. Pay attention to the “Russian doll effect”: one characteristic fold (OB-fold) is a part of another characteristic fold (nuclease fold).

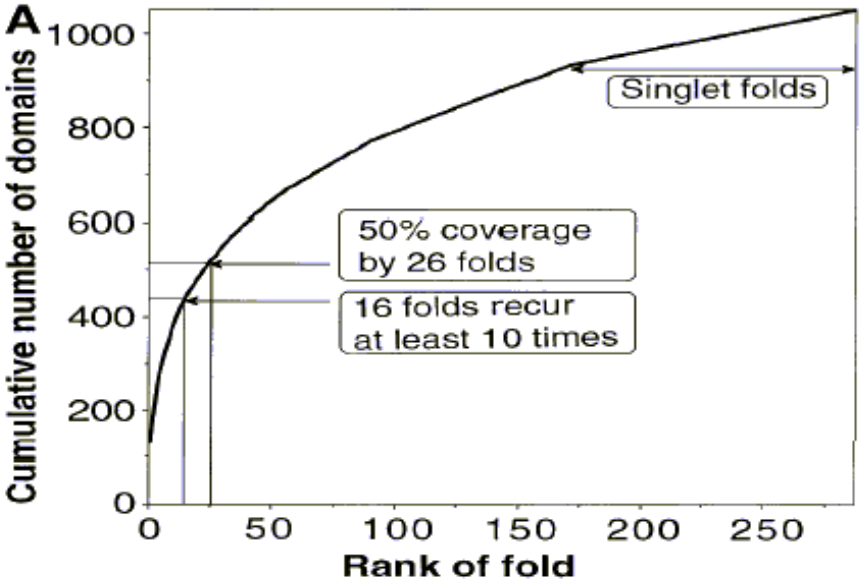
Summary: protein fold universe



Next lecture: Why some folds are more populated than the other?

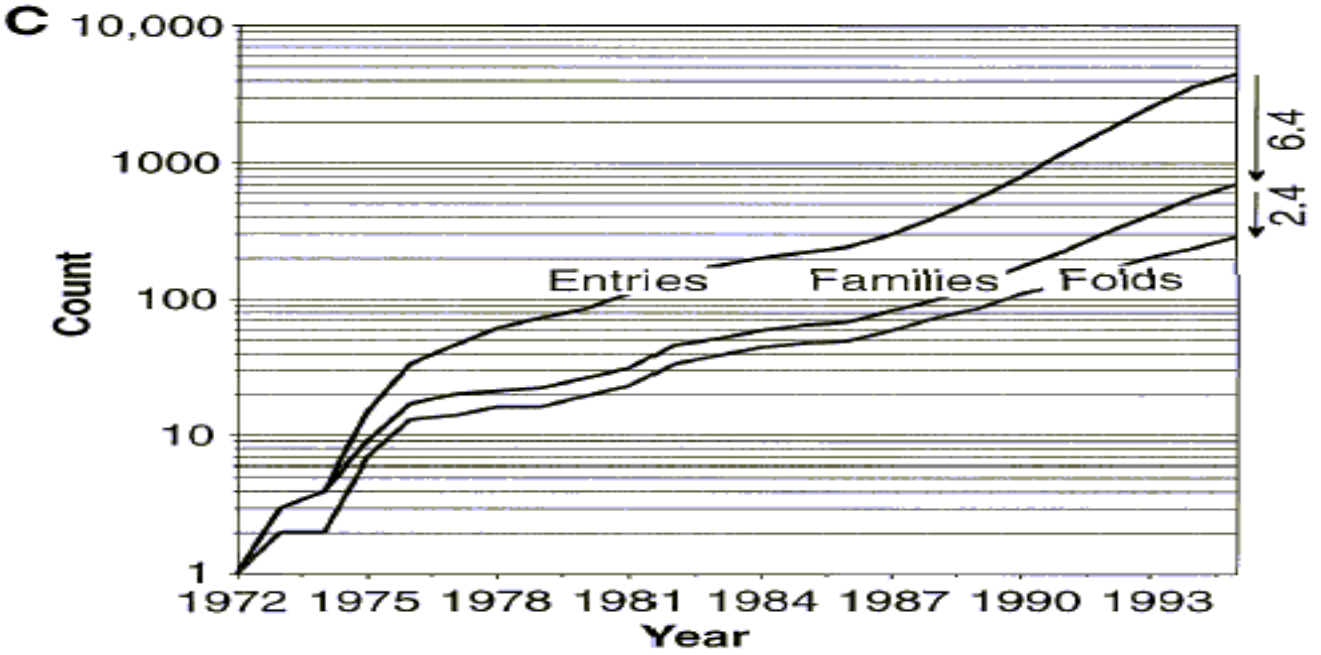
Physics or Biology?

Few folds populate most of the protein universe



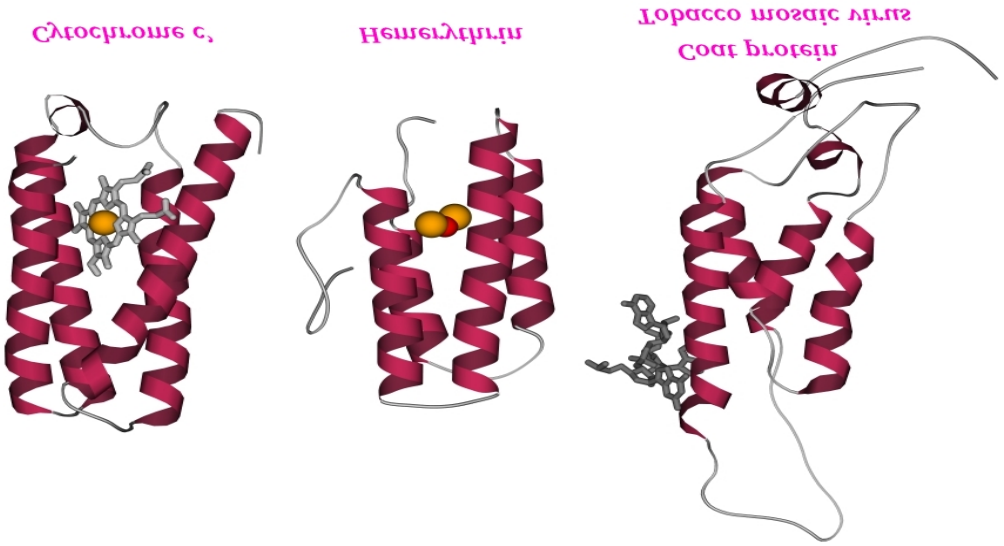
Holm and sander, Science’96

This situation remains the same as new folds get discovered



Holm and Sander’96

Sequence determines structure uniquely but inverse is not true.

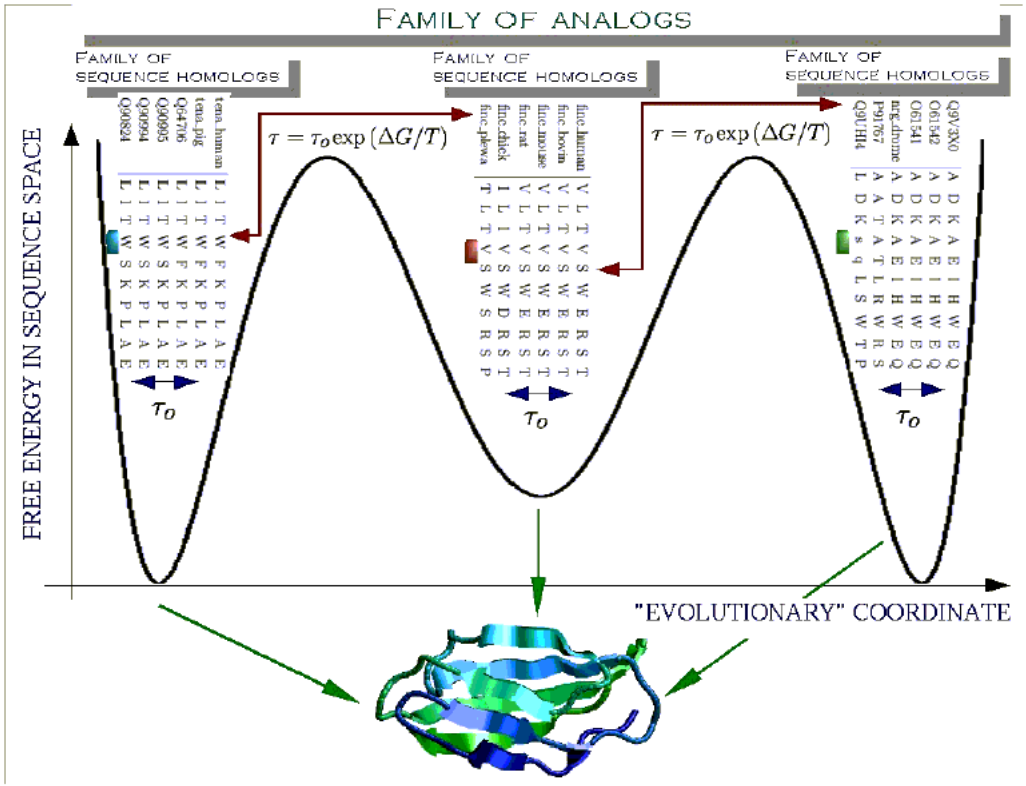


Major mystery in structural genomics:

Why some folds are very populated while others are ‘’orphans’’?

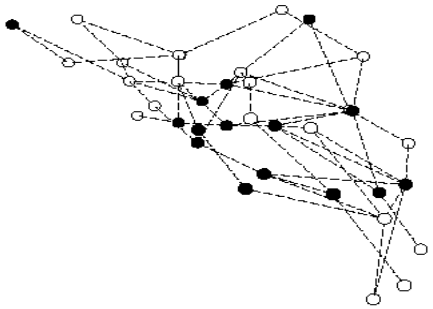
(databases: FSSP, SCOP,CATH, Pfam – all available online)

Cruising sequence space: a landscape view on protein genesis and evolution



Statistical Mechanics of Evolution: Sequence Chance

Statistical Mechanics of Evolution



- Probability to find a sequence $\{\sigma_i\}$

$$P\{\sigma_i\} = \frac{e^{-\frac{E_0\{\sigma_i\})}{T_{evol}}}\delta\Big(\sum_i^N \sigma_i - N_u\Big)}{\tilde{Z}}$$

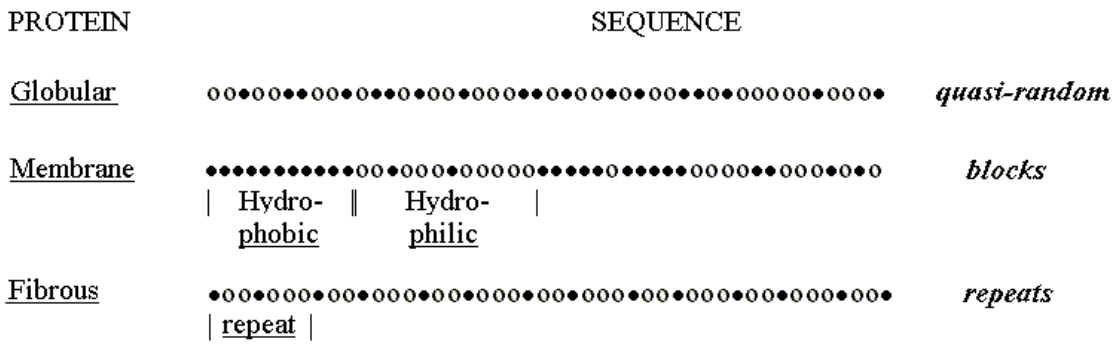
- $\tilde{Z}$ is a partition function in sequence space:

$$\tilde{Z} = \sum_{\{\sigma_i\}} e^{-\frac{E_0\{\sigma_i\})}{T_{evol}}}\delta\Big(\sum_i^N \sigma_i - N_u\Big)$$

- Statistics of sequences is isomorphic to Ising model of ferromagnetism

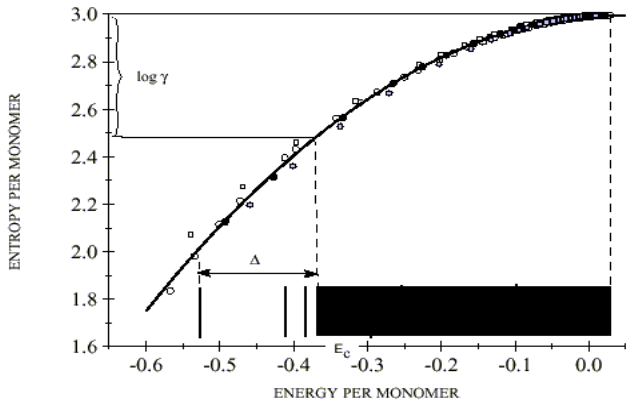
Designability principle: a convergent evolution viewpoint

Simple physical and statistical principles dictate some observed structural features of proteins, e.g. their two-layered structure.



What is the chance to fish out a Myoglobin from a random pool?

WHAT IS THE CHANCE TO "FISH OUT" A MYOGLOBIN SEQUENCE?



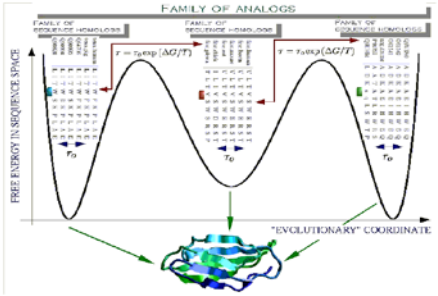
THE NUMBER OF SEQUENCES THAT MAY BE STABLE IN MYOGLOBIN CONFORMATION:

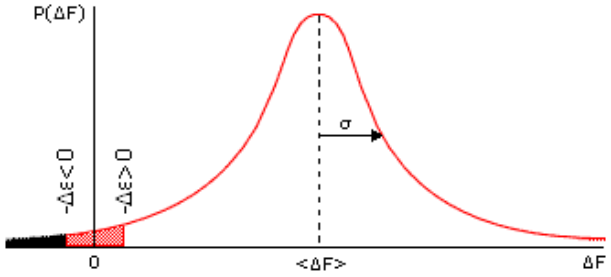
$$e^{2*153}$$

THE CHANCE TO "FISH OUT " A MYOGLOBIN SEQUENCE:

$$e^{2*153}/20^{153}$$

Designability principle: ‘’more stable folds’’ can accommodate more sequences



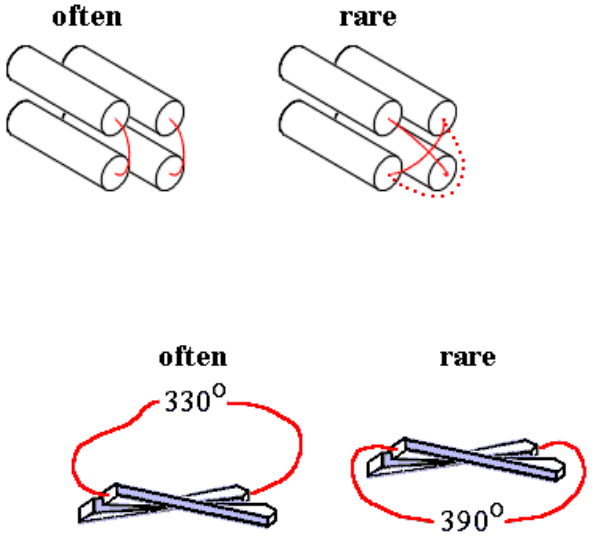


A typical Gaussian curve

$P(\Delta F) \approx \{(2\pi\sigma^2)^{-1/2} \times \exp[- \langle\Delta F\rangle^2/2\sigma^2]\} \times \exp[\Delta F \times (\langle\Delta F\rangle/\sigma^2)]$.

for ΔF distribution among random sequences. ΔF contains the entire free energy difference between the given fold and the unfolded state of the chain, except for the fixed Δε value of the structural element in question. The values of $\Delta F < -\Delta\varepsilon$ (i.e., those satisfying the condition that $\Delta F+\Delta\varepsilon < 0$) meet the requirements of a stable fold. The area shown in black corresponds to $\Delta F < -\Delta\varepsilon$ values at $\Delta\varepsilon > 0$, while the “red+black” area is for $\Delta\varepsilon < 0$. The latter is larger, which means that a greater number of random sequences stabilize the fold when the free energy of the element in question is below zero ($\Delta\varepsilon < 0$) as compared with its being above zero ($\Delta\varepsilon > 0$).

Examples of favorable and unfavorable structural elements



A closer look at designability principle: exact lattice model (ES and A.Gutin’90)

27 MER: THE IDEA.

SINCE THE TOTAL NUMBER OF ALL MAXIMALLY COMPACT CONFORMATIONS IS MODERATE:103346, THEY CAN BE ENUMERATED EXHAUSTIVELY. HENCE THE STATISTICAL MECHANICS OF THIS MODEL CAN BE EVALUATED NUMERICALLY EXACT (ES and AGutin’90).

Similar idea for 2d models: K.Dill and coworkers’89

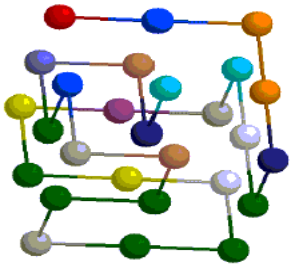
Applications to thermodynamics:

Probability that a random sequence has a gap (ES and AG’90),

Test of analytical theory and the REM analogy (ES and AG’90,Dinner et al’94).

’’Designability’’ concept and its evolutionary implications: Goldstein and coworkers’95-98, Tang and Coworkers’96

Derivation of parameters: rigorous tests (Finkelstein et al’93,Thomas and Dill’96, Mirny and ES’96)



103346 compact conformations

Lattice model analysis of designability principle (Li et al, Science, 96)

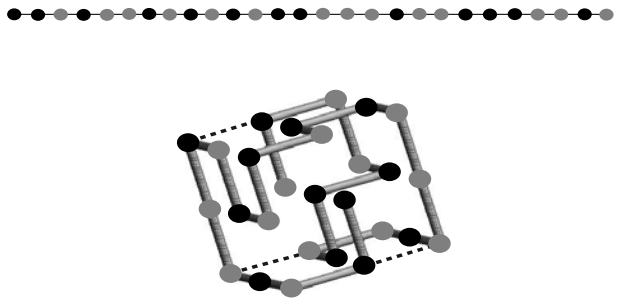
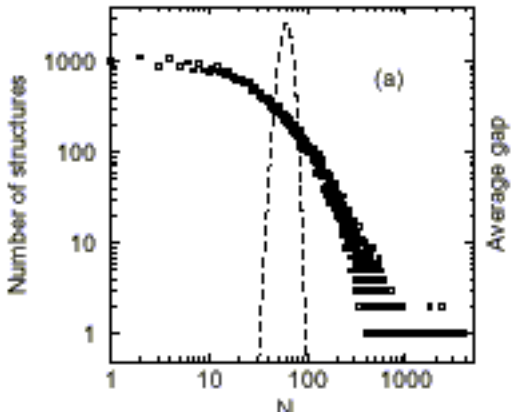


Fig. 5. A 3D lattice HP model. A sequence of H (dark disc) and P (light disc) (a) is folded into a 3D structure (b).

In a tour de force calculation these authors enumerated all 103346 compact Conformations for all 2^{26} HP sequences and determined ground state(s) for each Sequence. Some sequences featured unique ground state some (96% majority) did not.

Lattice designability calculations: results and caveats



Li et al observed that unique ground state conformations were more Symmetric. However subsequent analytical solution for this model by E.Kussell and ES (PRL' 99) showed that this conclusion is extremely model-dependent

Analytical model of protein designability allows a closed-form solution for a class of contact potentials (England and ES, PRL, 2003)

$$H = \frac{1}{2}\sum_{i,j}^{N,N} C_{i,j}\vec{s}^{(i)} \cdot (B\vec{s}^{(j)})$$

$$B = \begin{bmatrix} V_{1,1} & -V_{1,1} & V_{1,2} & -V_{1,2} & \cdots \\ -V_{1,1} & V_{1,1} & -V_{1,2} & V_{1,2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix}$$

$$\mathbf{U}_{M(i-1)+k,M(j-1)+l} = C_{i,j}V_{k,l}$$

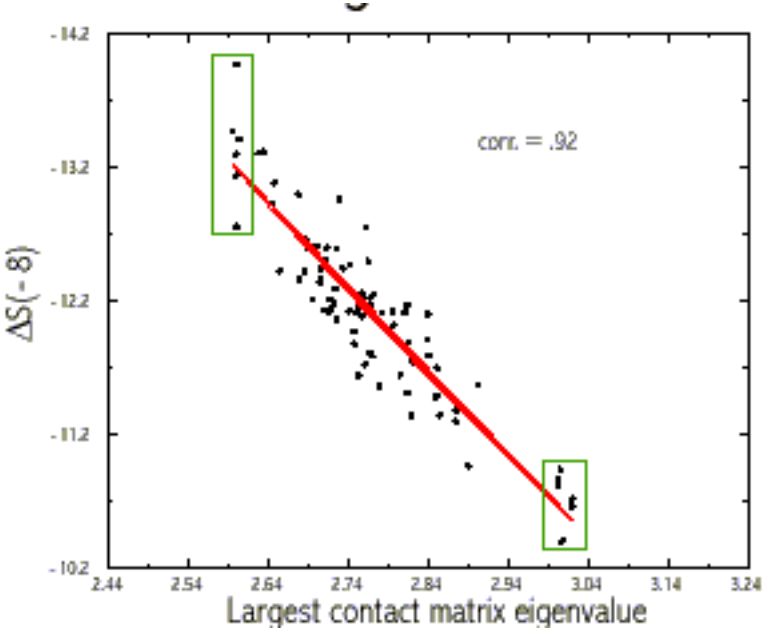
$$z(\beta) \equiv \frac{Z(\beta)}{Z(0)} = \sqrt{\frac{\det[\mathbf{M}]}{\det[\mathbf{M}+\beta\mathbf{U}]}} = \frac{1}{\sqrt{\mathbf{I}+\beta\mathbf{u}}}$$

$$F = -\frac{\beta}{4}(\mathrm{Tr}\, v^2)(\mathrm{Tr}\, C^2) + \frac{\beta^2}{6}(\mathrm{Tr}\, v^3)(\mathrm{Tr}\, C^3) - \frac{\beta^3}{8}(\mathrm{Tr}\, v^4)(\mathrm{Tr}\, C^4) + O(\beta^4)$$

Main result of the analytical calculation:

The ‘’Contact Trace’’
(well approximated by largest eigenvalue of the contact matrix)
is a structural determinant of protein designability

Correlation between designability and largest CM eigenvalue: 200 randomly selected 27-mer structures



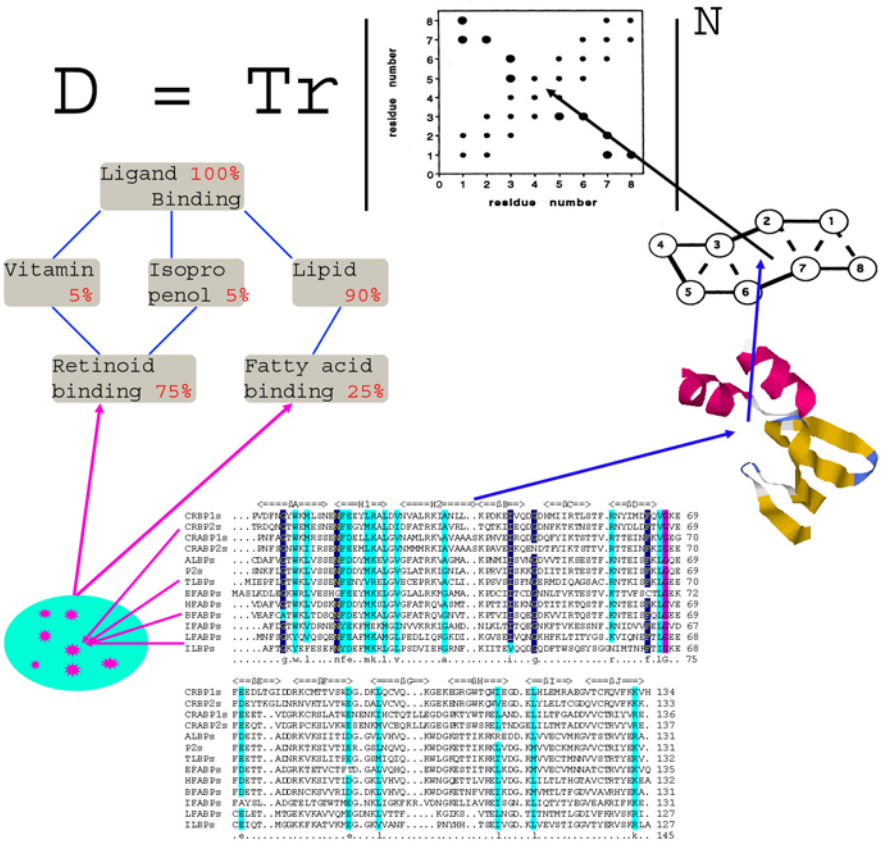
Proteomic Implications

More designable folds are:

- more plastic,
- more functionally diverse,
- have more paralogous genes.

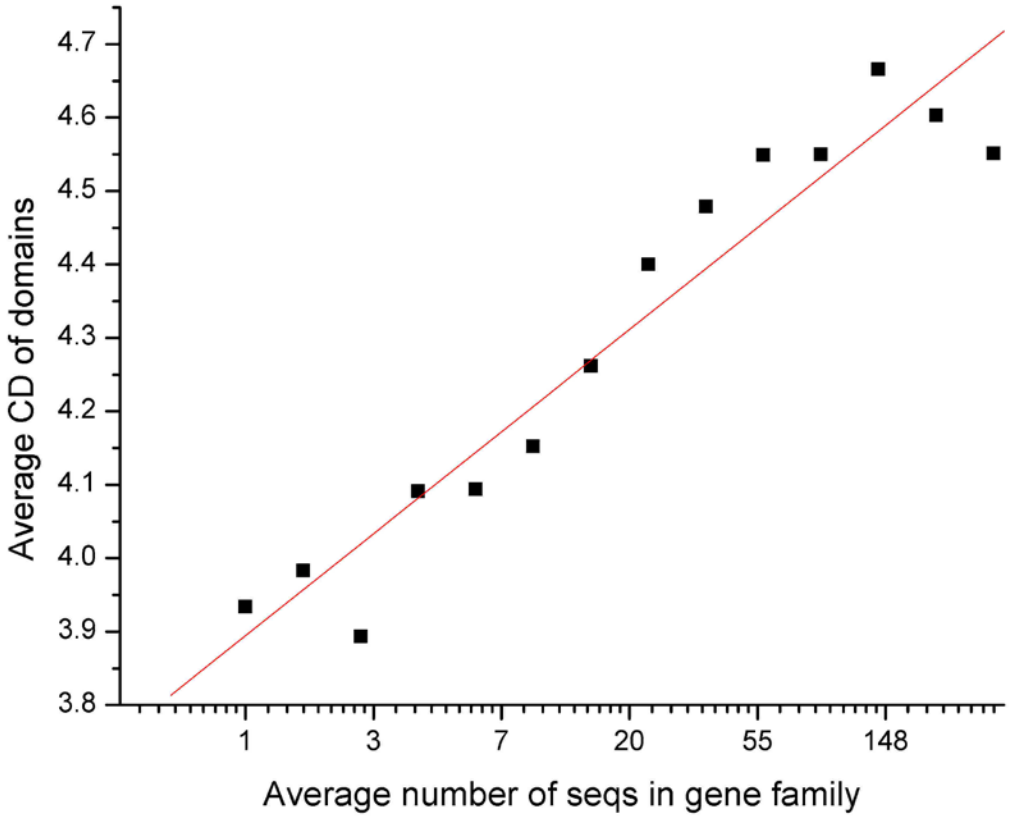
If CT is a proxy for designability,
there should be a correlation
between CT and functional diversity of a fold

Quantification of family size/functional diversity using GO and InterPro



Is CD a predictor of Gene and Fold Family Sizes?

In one word the answer is "Yes"



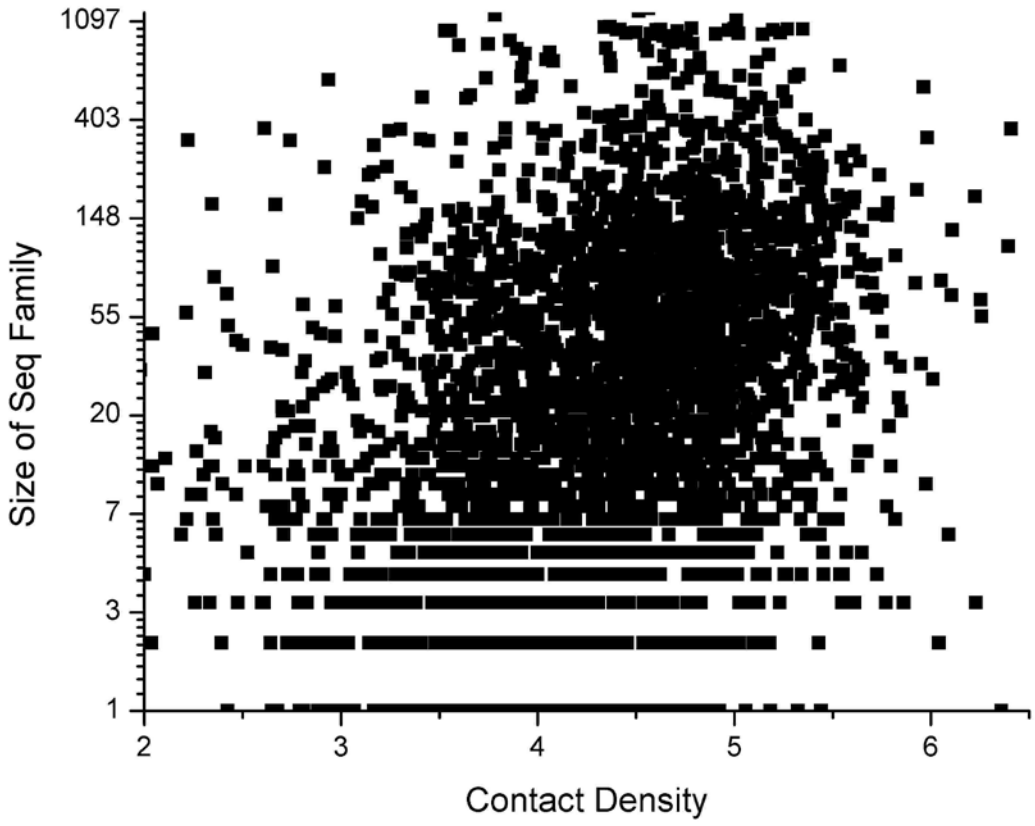
There is a strong statistical correlation (R^2=0.92) between CD of a domain and the number of sequences that encode this domain.

In two words the answer is: ”Not Really”.
Looking at Individual, *not binned* data

Real size depends on “evolutionary history”

CD is a proxy for designability

Designability is potential for acceptance of mutations not the determinant thereof.



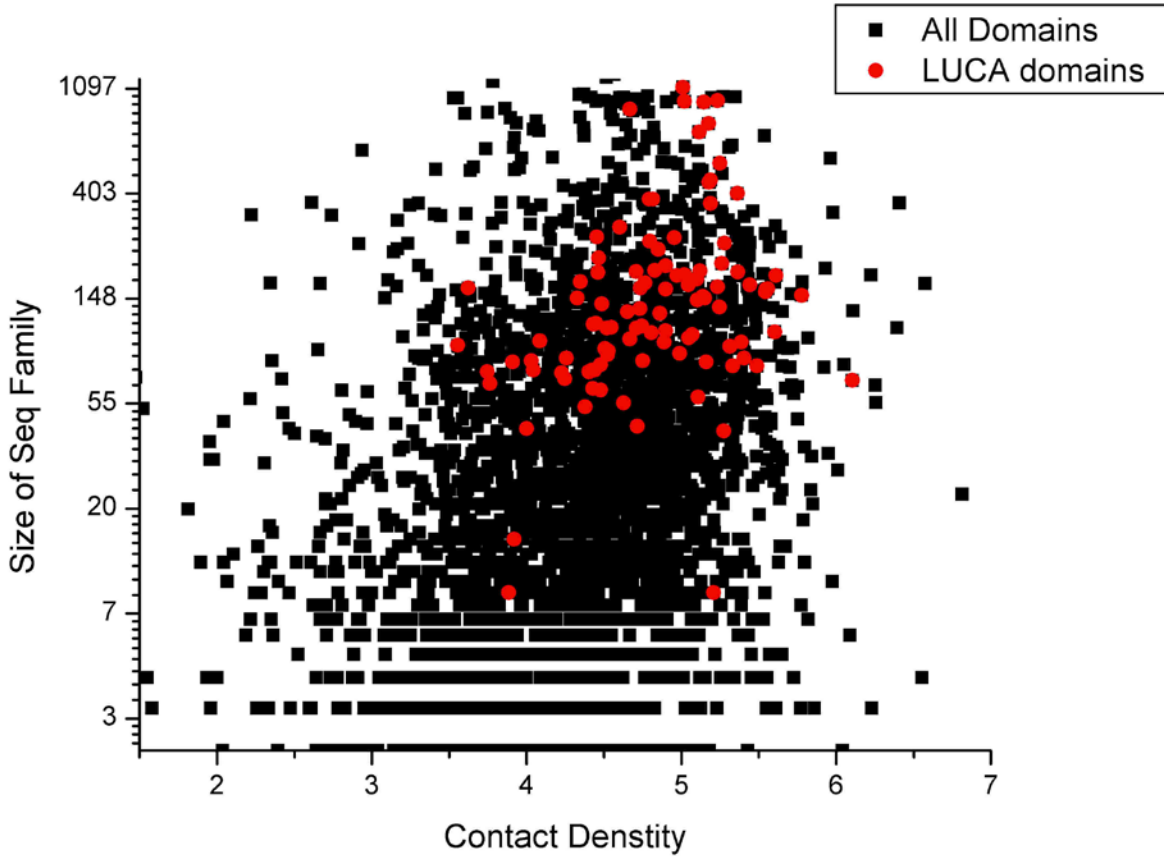
Historically speaking

LUCA: Last Univeral Common Ancestor..

Poor man's LUCA: Domains that are present in all organisms i.e. archea, prokaryotes and eukaryotes **(cf Koonin and coworkers for better LUC'ism)**

Domains Started from higher CD and evolved to lower CD

Size of sequence family is a function of time as well as fitness…



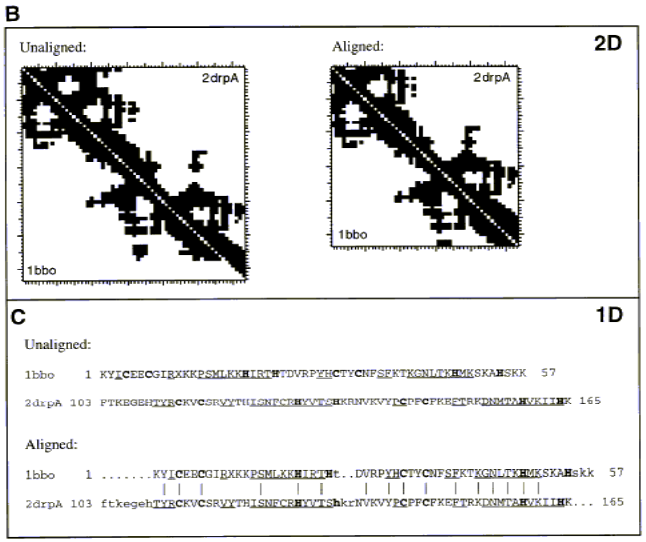
Designability principle:critiques

1) **Assumes convergent evolution based on equilibrium in sequence space.**

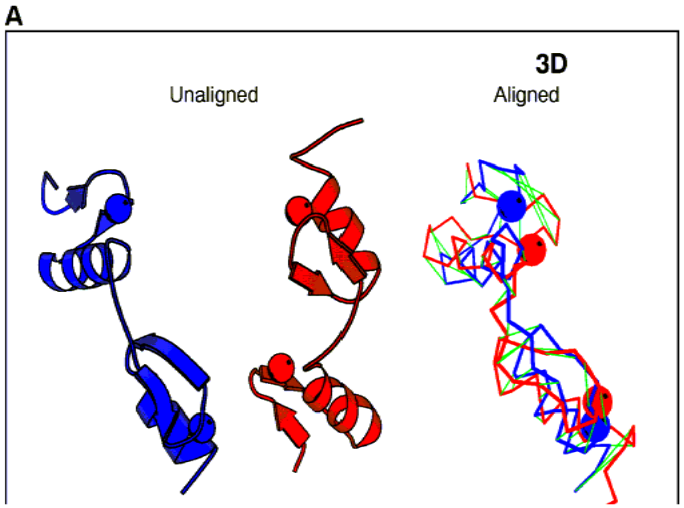
2) **Fails to identify which protein folds would be more populated.**

3) **Is based on very simplified assumptions such as HP-sequences; real sequences feature 20 aminoacids and in this case the difference in designabilities for different protein folds disappears (ES, Folding and Design' 98, see later)**

A quantitative approach to the analysis of protein universe: DALI by Holm and Sander



More on DALI

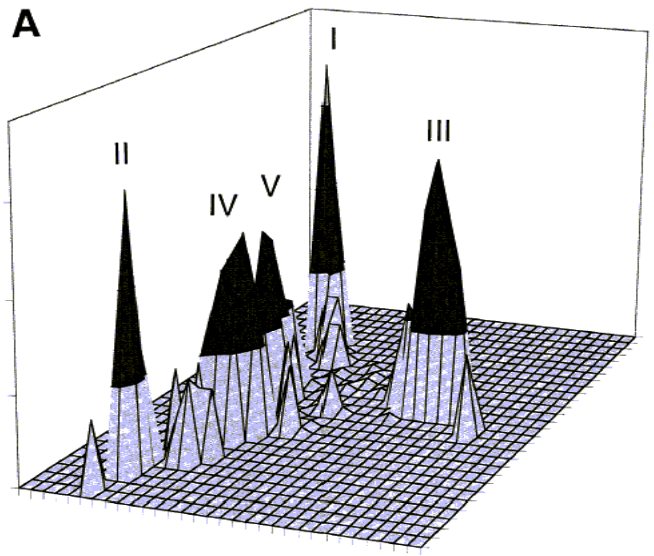


Monte-Carlo-based routine that aligns contact matrices for two Protein structures and applies DALI score that favors good alignment.

Z-score – parameter that shows how does DALI score deviates from such Of two random aligned structures (number of standard deviations from average)

Z-score is a quantitative measure of structure proximity

Clustering protein structures using DALI:Holm and Sander



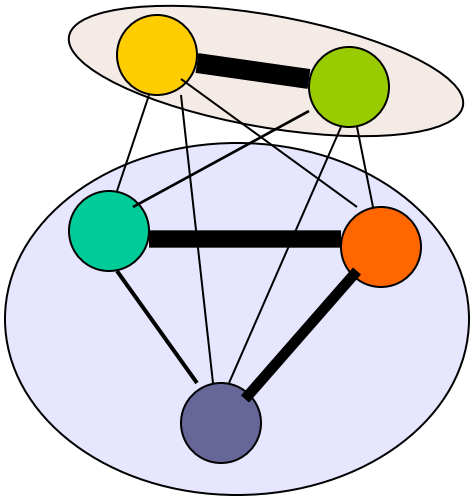
Exploring protein universe using graph theory: Dokholyan and ES'02

Consider all known protein domains as nodes of a graph.
Make all-against all structural comparisons using DALI.
For each pair evaluate their DALI Z-score. If $Z>Z_{min}$ connect nodes with an edge.
Split into disjoint clusters.
That creates Protein Domain Universe Graph (PDUG)

PROTEIN STRUCTURAL SPACE: RECONSTRUCTION

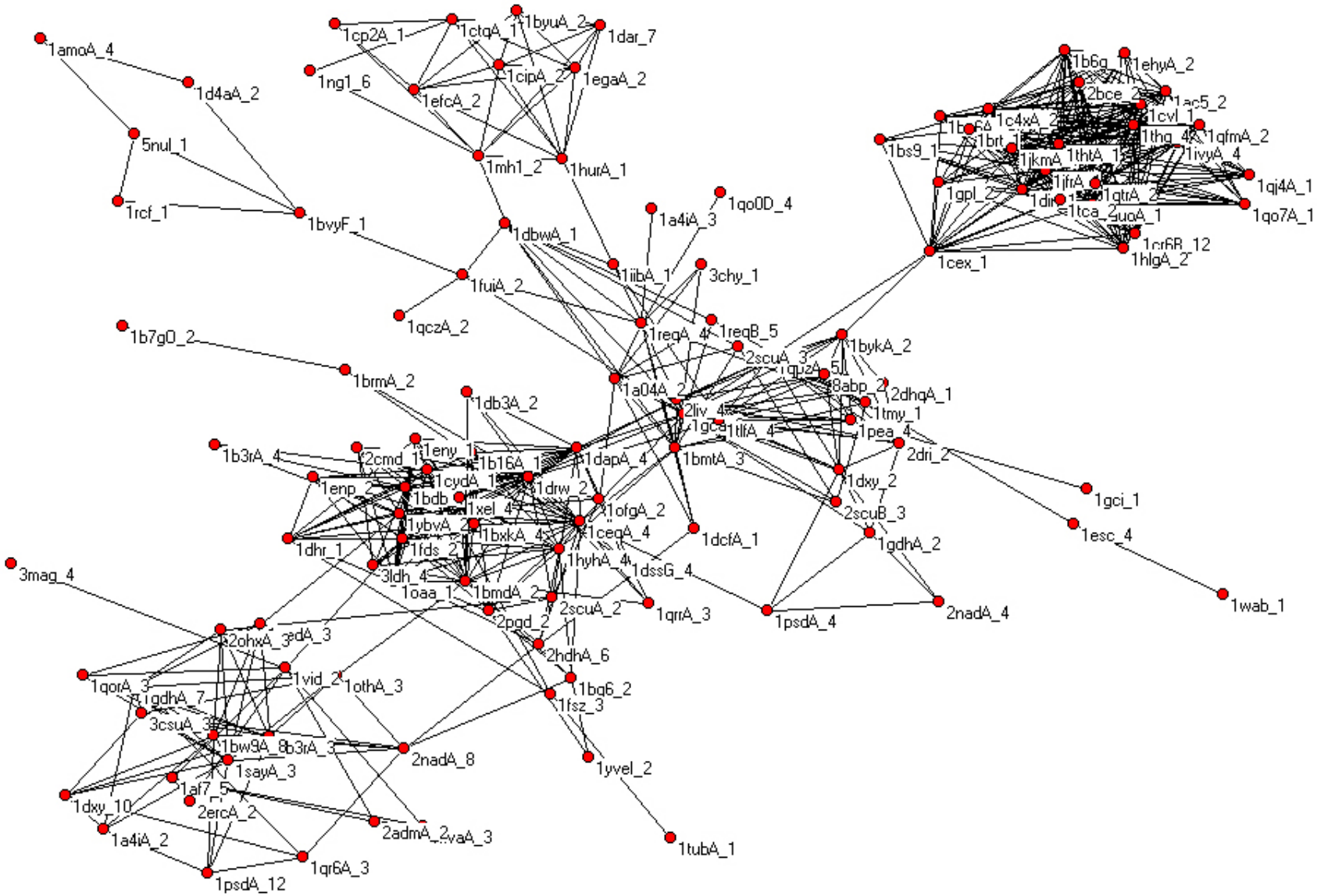
From DALI database we construct a graph of relations between non-homologous proteins:

1. **each node represents a domain**
2. **we draw an edge between any two proteins that are structurally similar.**
3. **to assign structural similarity to any two proteins, we use the *Z-score* --- the significance level of structural similarity.**
4. **introduce a cut-off Z_{min} for the significance level of the structural similarity.**
5. **identify connected components (clusters) in the protein conformation space – these are families of structurally related proteins (folds)**



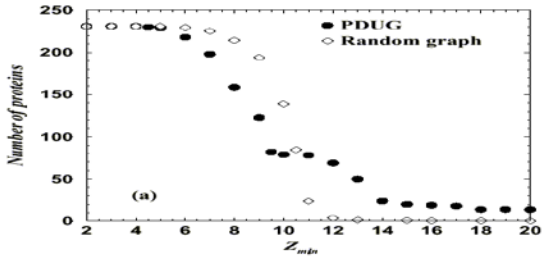
Proc. Natl. Acad. Sci. USA (2002)

LARGEST CLUSTER AT Z_c



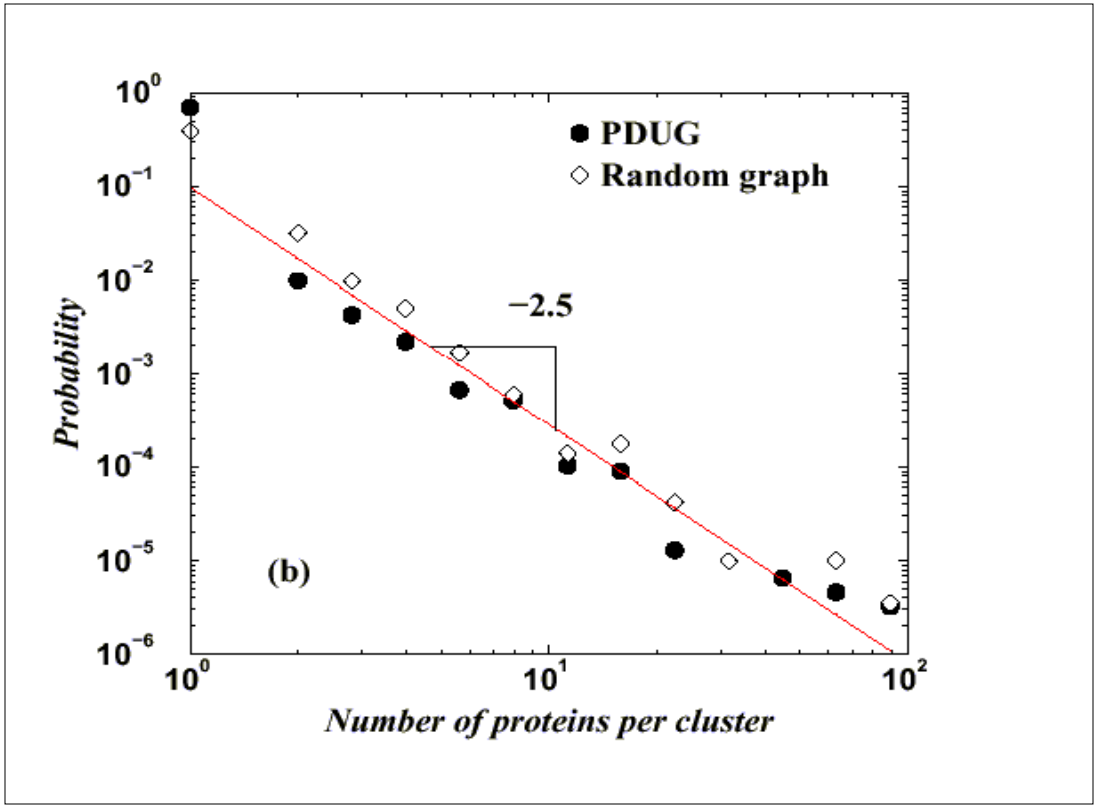
Proc. Natl. Acad. Sci. USA (2002)

How do properties of PDUG depend on Zmin?

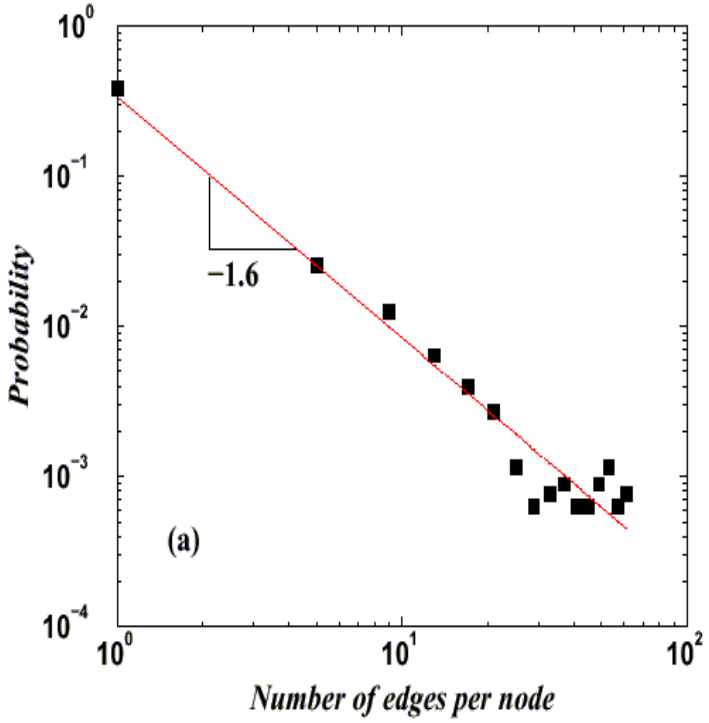


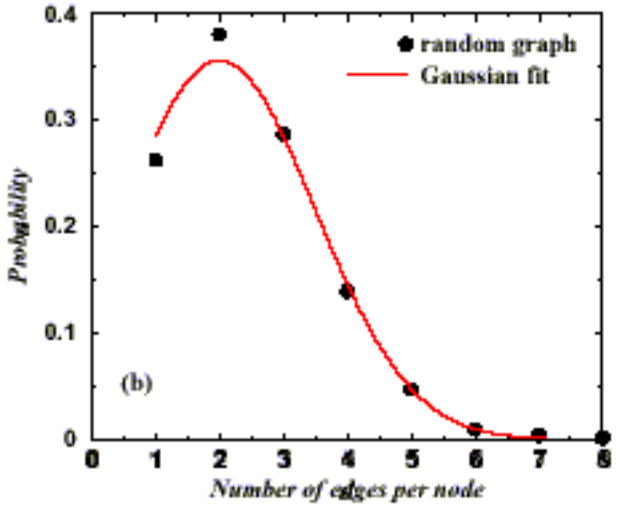
Note: control of random graph is crucial here

Critical behavior at Zmin: inhomogeneous fold distributions



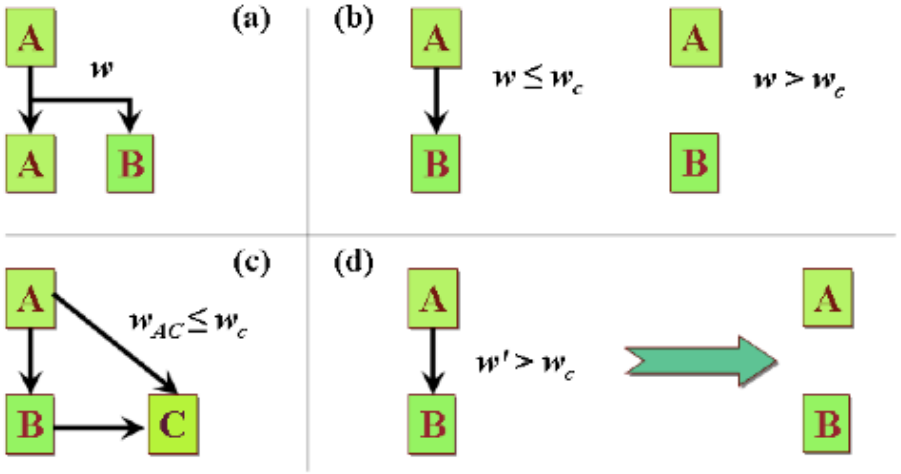
Is PDUG lake a random graph? NO! It is a scale-free network!



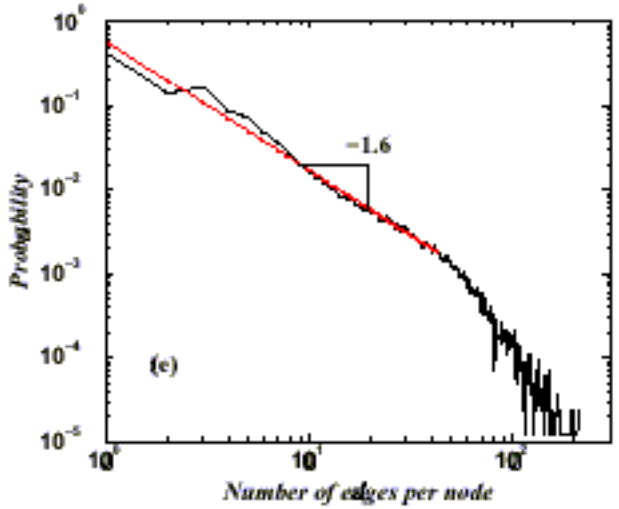


This observation shows that PDU is organized hierarchically…

Divergent evolution from the ‘’Big Bang’’: a dynamic model for protein creation via gene duplication and mutation



The divergent evolution model explains scale-free protein universe



Summary: Convergent or divergent structural evolution?

Designability principle is an elegant physics-based scenario of **convergent evolution**. However it has not been overly successful In explaining observed features of protein universe.

The dynamic **divergent evolution** scenarios are more successful in Explaining peculiar properties of protein universe. However their Burden at this point is to explain evolution of function.

Divergent evolution scenario is supported by data and analysis.

How many sequences fit a given structure?

Degeneracy of Protein Code: How Many Sequences Fit a Given Structure?

- The number of sequences that fit a given structure with a given energy E is simply an entropy in sequence space:

$$M(E) = \exp\left(S(E)\right)$$

- This number can be determined from $E(T_{eval})$ using standard statistical mechanical relationship:

$$S(T) - S(\infty) = \frac{E(T)}{T} - \int_T^\infty \frac{E(t)}{t^2} dt$$

with $S(\infty) = \ln 20$ and $S(\infty) = \ln 2$ for the two-letter code

- Having E(T) we obtain S(E) from S(T)

Protein structure and function are weakly connected- major challenge for structural genomics