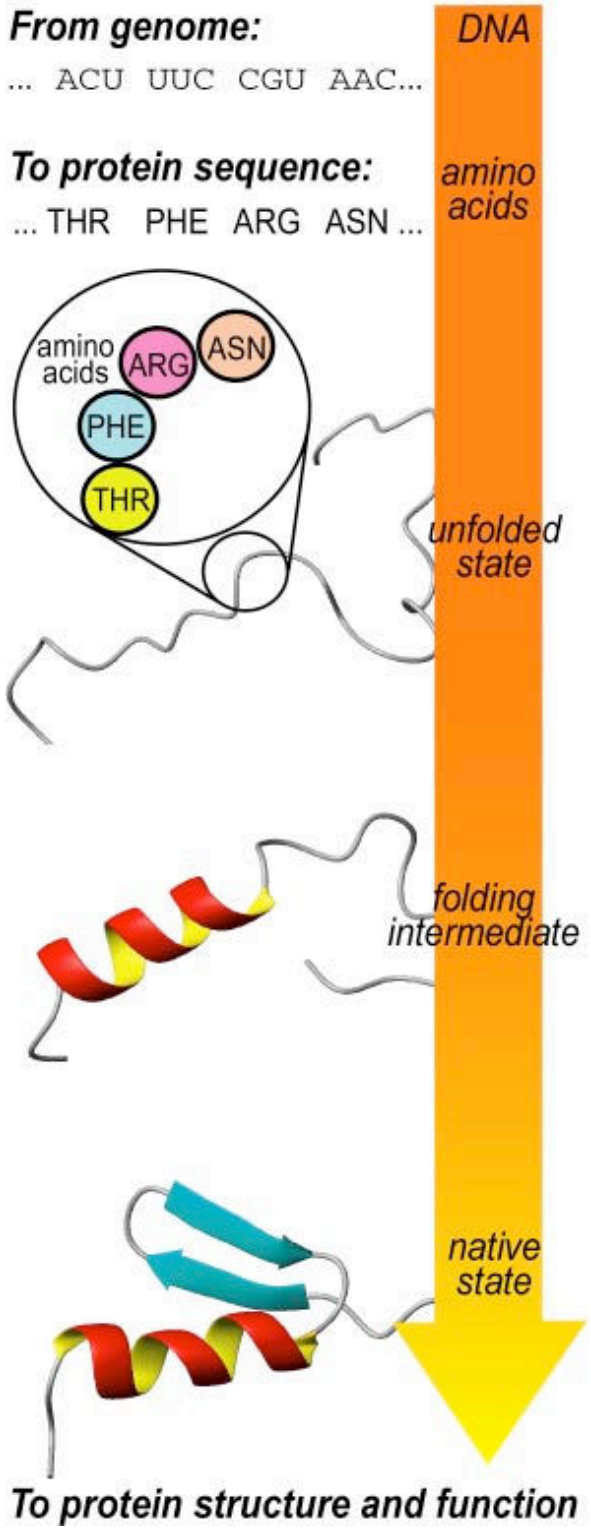
Tertiary Structure: from Random Coil to Stable Fold

Levinthal's Paradox and Anfinen's Theorem
Folding / Misfolding
Thermodynamics / Kinetics
Folding Simulations and Predictions


From genome:
... ACU UUC CGU AAC...
To protein sequence:
... THR PHE ARG ASN ...
DNA
amino acids
amino acids
ARG
ASN
PHE
THR
unfolded state
folding intermediate
native state
To protein structure and function

What Is Folding?

- Folding is the transition of a protein chain from a random conformation to a highly ordered, low-energy conformation.
- Folding is a dynamic self-recognition process.
- The folded protein conformation is (mostly) the native and functional state.
- Time range is around micro-second to second. The speed is depending on the ratio of local and distant side-chain contacts, the so-called contact order. Proteins with predominantly local contacts fold faster; the complexity of the transition state is important.
- The folded state is just slightly more stable than the unfolded state, although the contributing energy components are large. This makes all calculations and predictions very difficult.
- Protein sequences have been selected by evolution to fold into the correct conformation.

Levinthal's Paradox

... in a protein structure containing 150 amino acids ... there would be 10^{300} possible configurations in our theoretical protein. In nature, proteins apparently do not sample all of these possible configurations since they fold in a few seconds, and even postulating a minimum time for going from one conformation to another, the proteins would have time to try on the order of 10^{8} different conformations at most before reaching their final state.

Cyrus Levinthal, 1969

http://www-wales.ch.cam.ac.uk/~mark/levinthal/levinthal.html

Anfinsen's Theorem

The primary structure determines the tertiary structure.

In the mid 1950's Anfinsen began to concentrate on the problem of the relationship between structure and function in enzymes. On the basis of studies on ribonuclease with Sela and White, he proposed that the information determining the tertiary structure of a protein resides in the chemistry of its amino acid sequence. Investigations on reversible denaturation of several proteins served to verify this proposal experimentally. It was demonstrated that, after cleavage of disulfide bonds and disruption of tertiary structure, many proteins could spontaneously refold to their native forms. This work resulted in general acceptance of the 'thermodynamic hypothesis' (Nobel Prize 1972).

Anfinsen performed un-folding / re-folding experiments.

http://www.nobel.se/chemistry/laureates/1972/anfinsen-bio.html

How many possible sequences?

$$N^{seq} = 20 * 20 * \ldots = 20^{length}$$

The number of possible sequences is nearly infinite. Remember the models for evolution: divergent and convergent. This and the fact that many sequences can adopt the same structure is the reason why sequences are dissimilar in convergent evolution.

The probability of finding twice the same sequence is

$$p = 1/N * 1/N$$

That is an upper bound, because evolution selects fewer sequences.

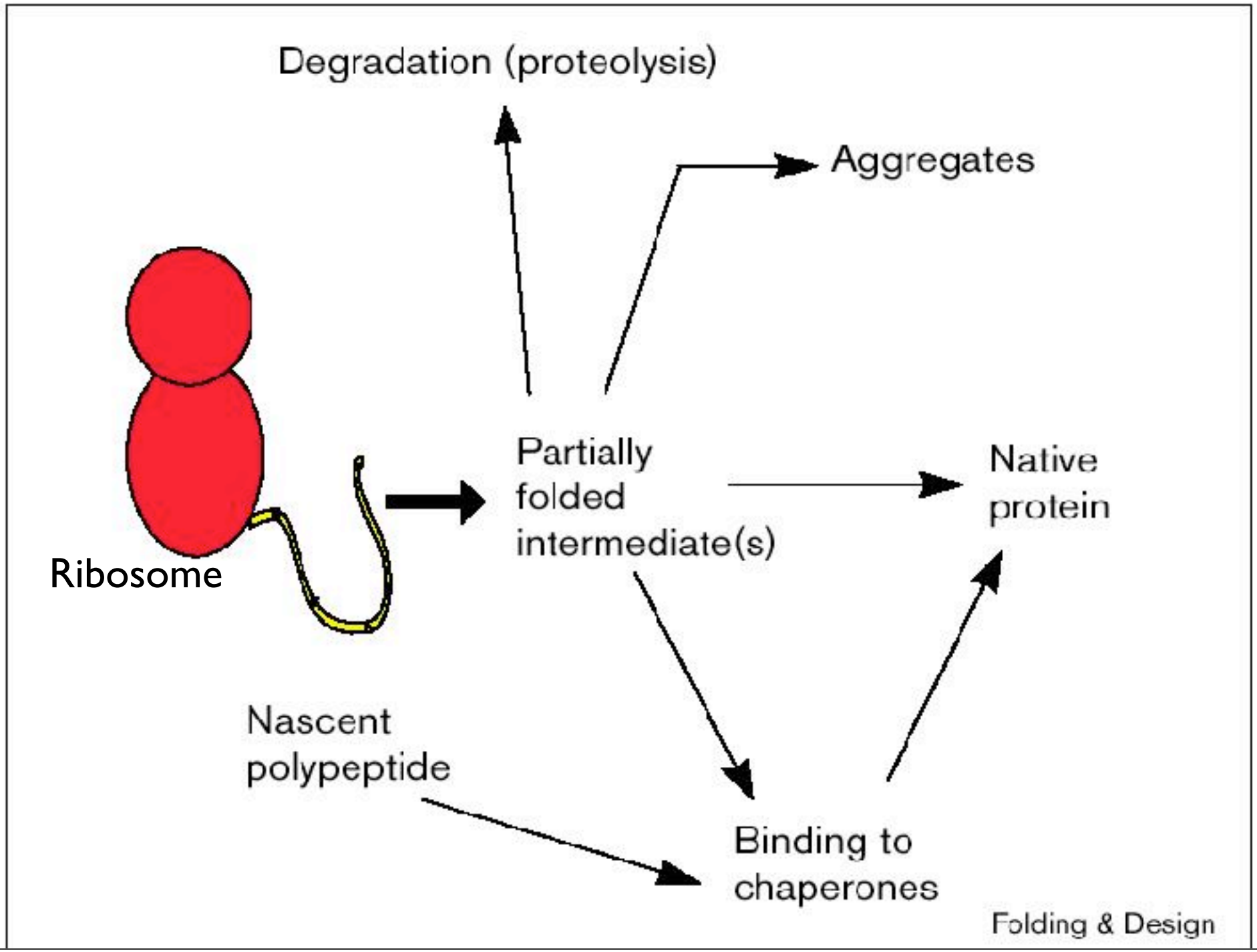
Folding / Unfolding

Most proteins are maximally stable at ≈30°C.

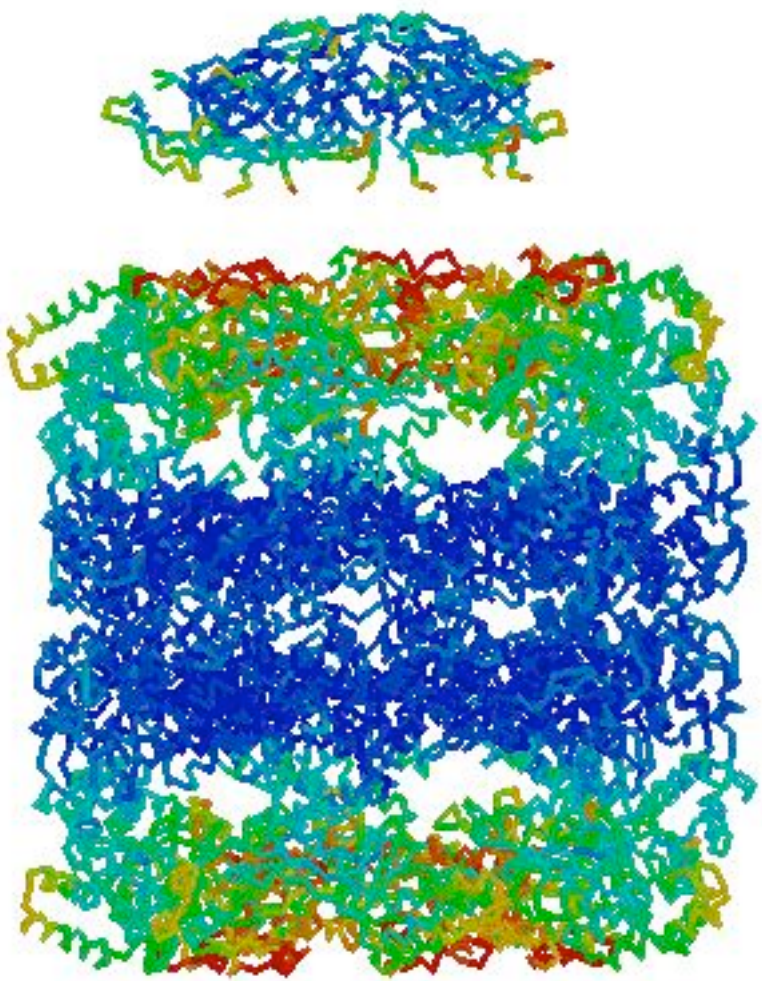
Unfolding can be induced by

- heating or cooling (thermal unfolding)
- chemical denaturants as urea, guanidinium hydrochloride, ions or organic solvents
- high pressure (high pressure unfolding)
- mutation

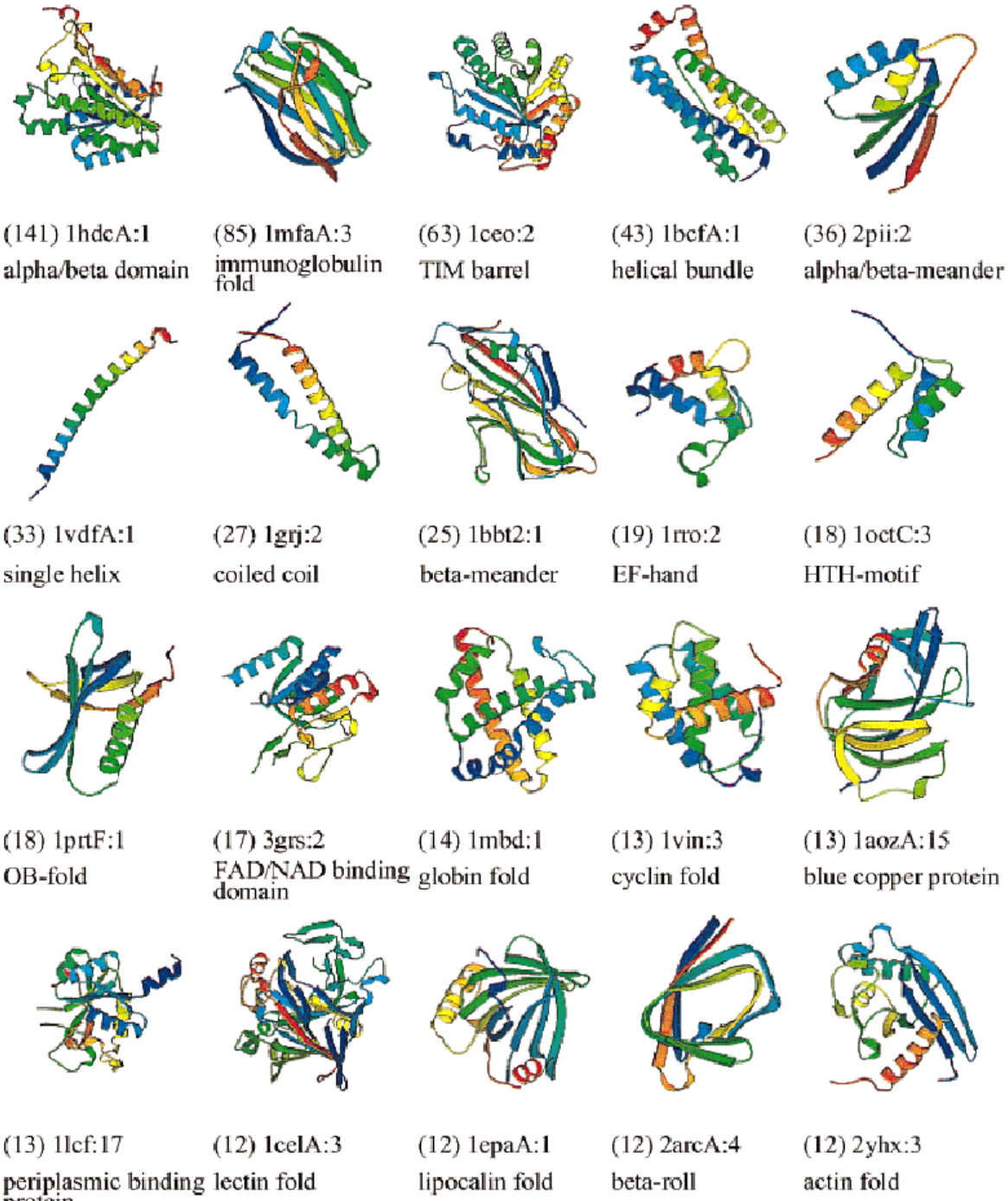
Folding versus Aggregation



GroEL / GroES Chaperone

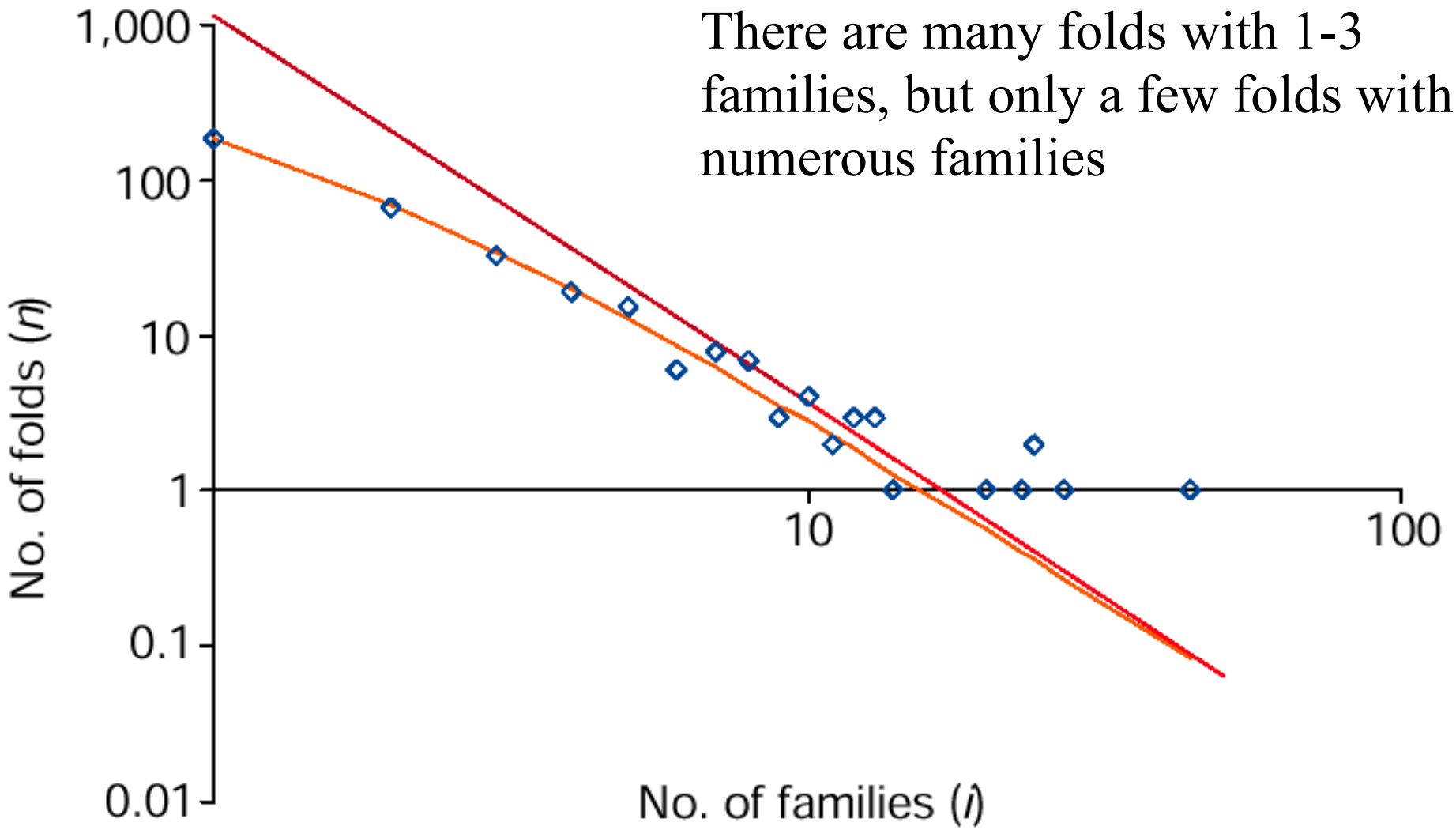


Dominant domain fold types.



Holm and Sander. PROTEINS: Structure, Function, and Genetics 33:88–96 (1998)

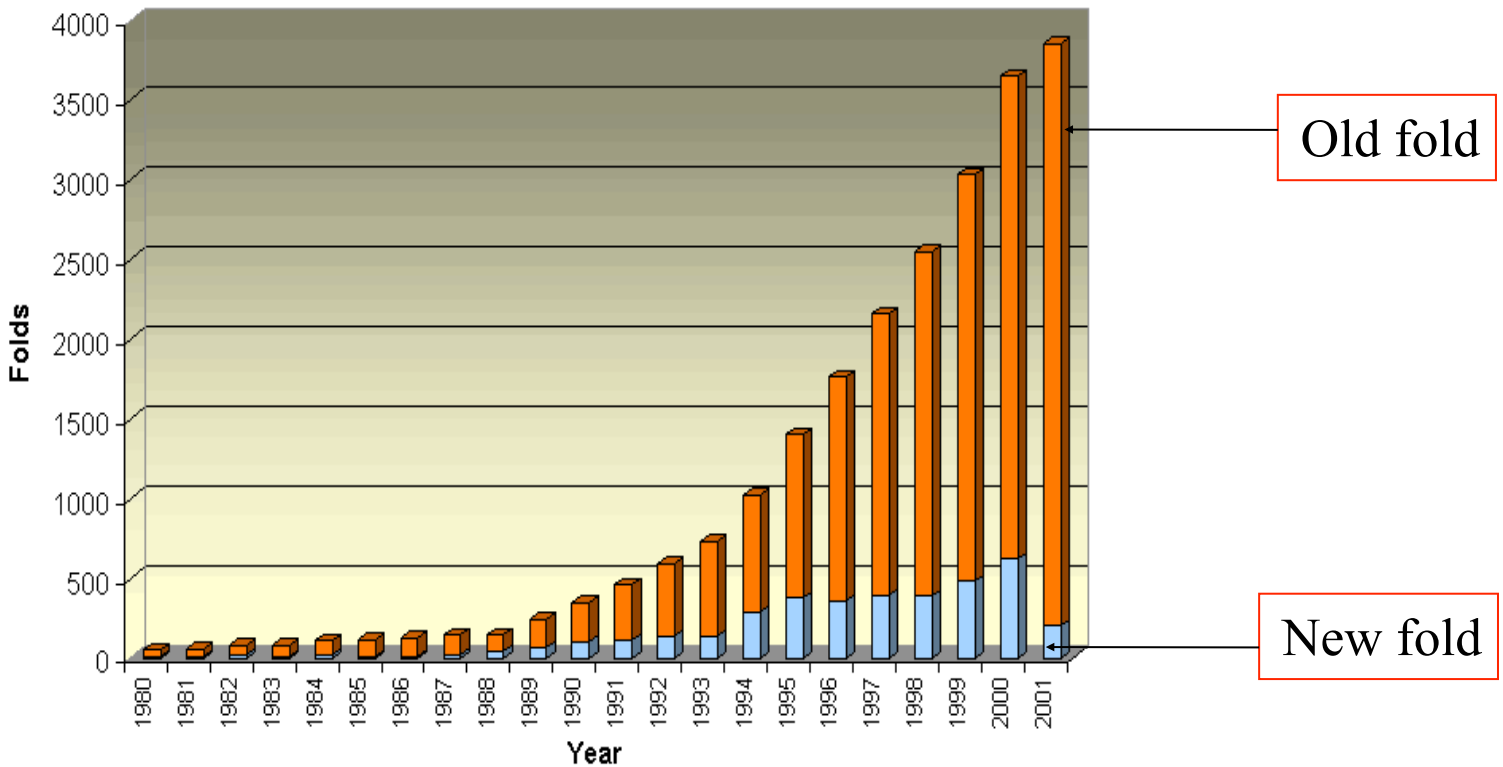
Distribution of protein folds by the number of families



There are many folds with 1-3 families, but only a few folds with numerous families

From Koonin et al. *Nature*. 420, 218-223 (2002).
See also: Zhang, C. & DeLisi, C. *J. Mol. Biol.* 284, 1301–1305 (1998).

PDB New Fold Growth



- Only a few thousand unique folds in nature
- 90% of new structures deposited to PDB in the past three years have similar structural folds

Fold Number Estimates

The total number of folds in globular, water- soluble proteins is estimated at about 1000.

The sequenced genomes of unicellular organisms encode from approximately 25%, for the minimal genomes of the Mycoplasmas, to 70-80% for larger genomes, such as *Escherichia coli* and yeast, of the total number of folds.

The number of protein families with significant sequence conservation was estimated to be between 4000 and 7000, with structures available for about 20% of these.

Wolf et al. JMB (2000) 299, 897-905

Hierarchy and Relationship

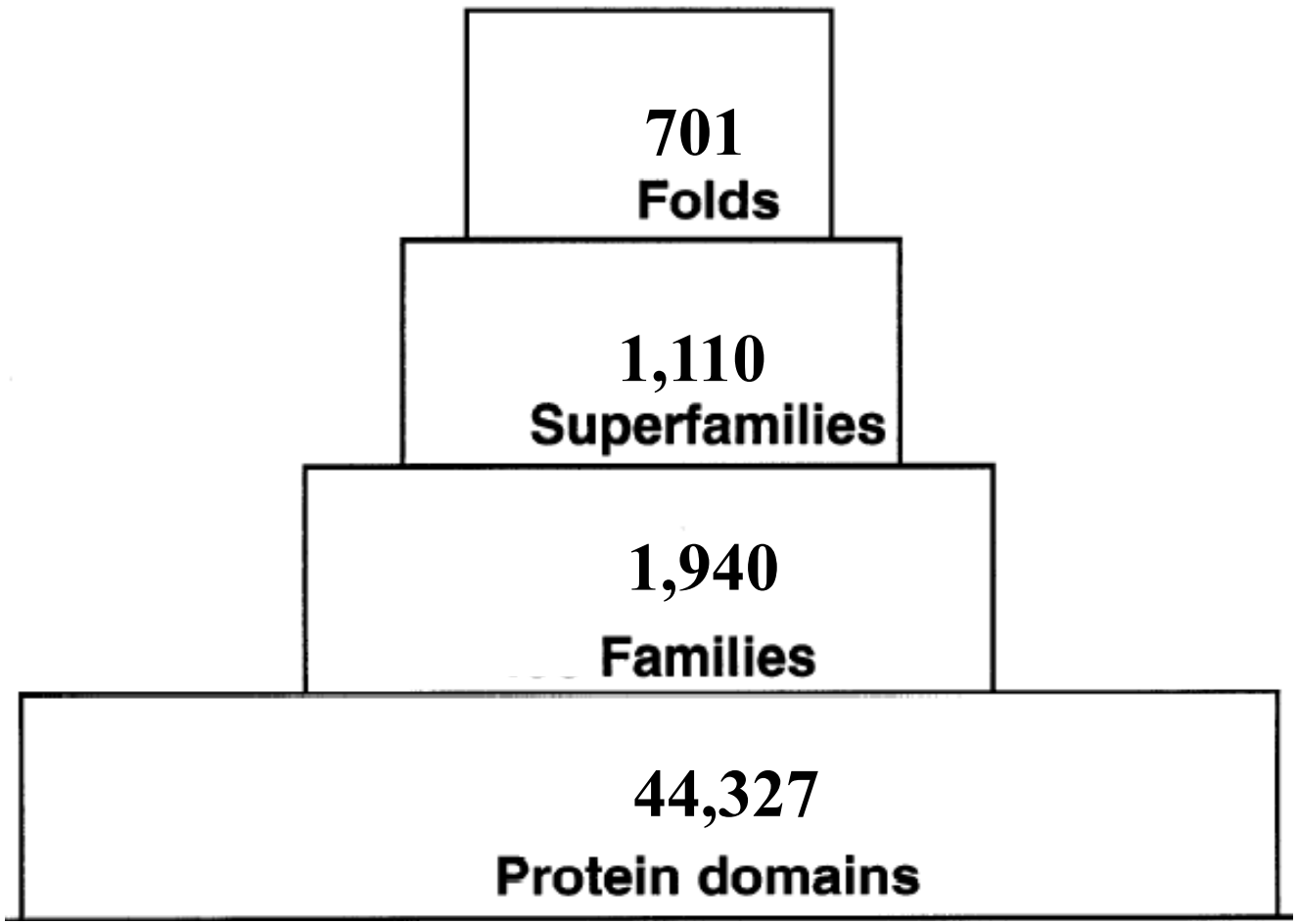
Family: Clear evolutionarily relationship; pairwise residue identities between the proteins are 30% and greater.

Superfamily: Probable common evolutionary origin; structural and f unctional features suggest that a common evolutionary origin is probable.

Fold: Major structural similarity; the same major secondary structures in t he same arrangement and with the same topological connections.

Proteins placed together in t he same fold category may not have a common evolutionary origin.

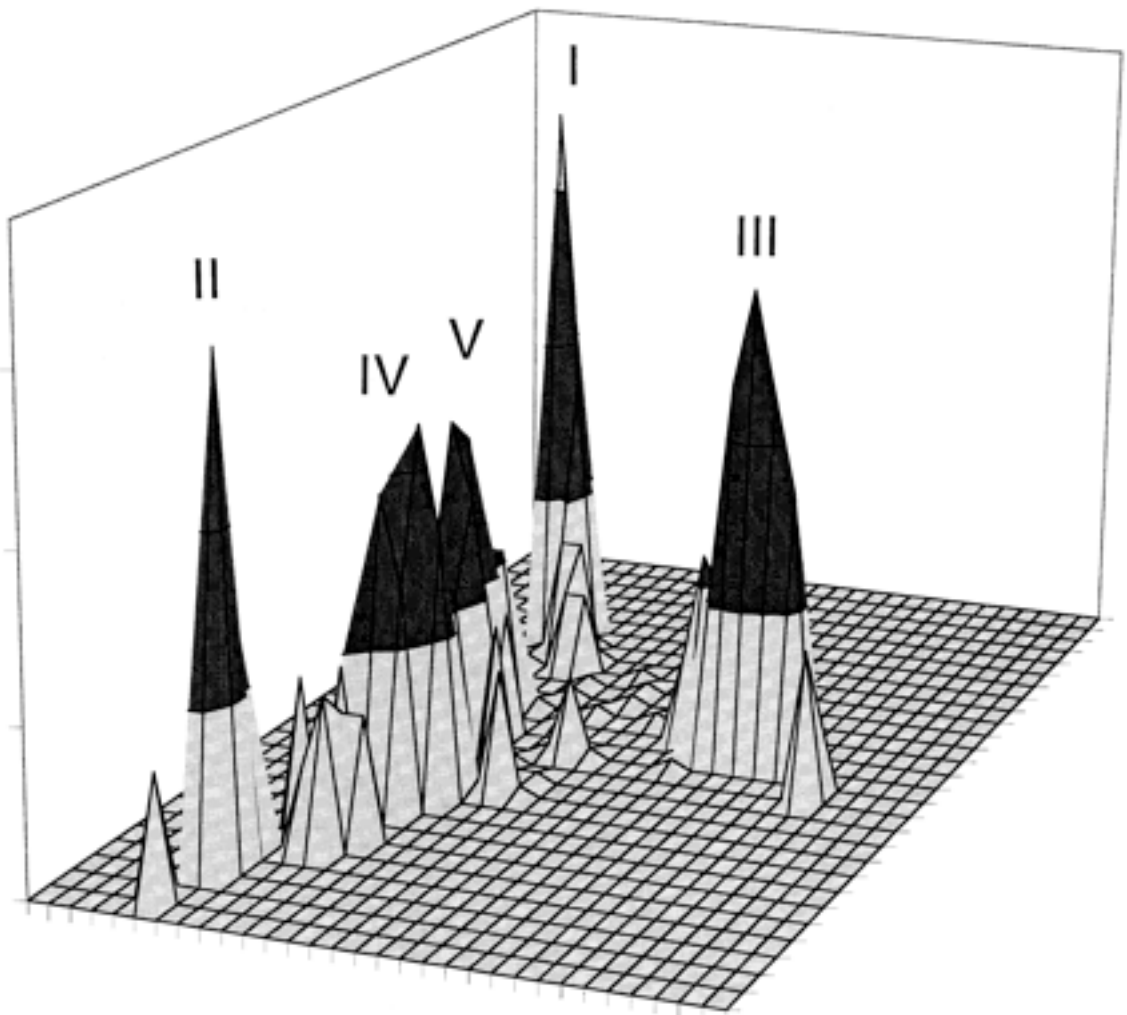
SCOP – a structural classification of proteins

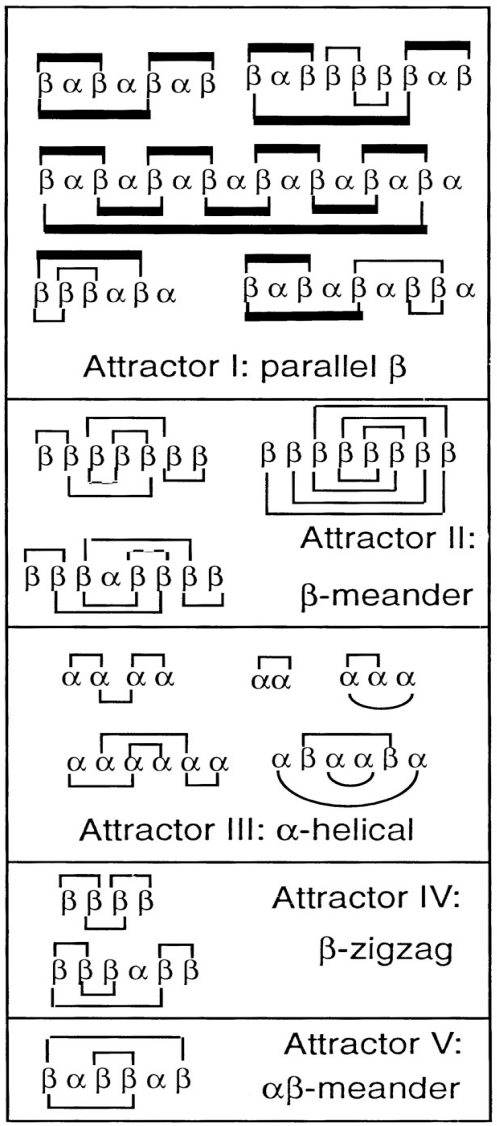


A fold is a topology of the folded protein backbone.
Unknown whether superfamilies of the same fold are monophyletic

Updated from Murzin et al. *J. Mol. Biol.* 247, 536-540.

What does fold space look like?

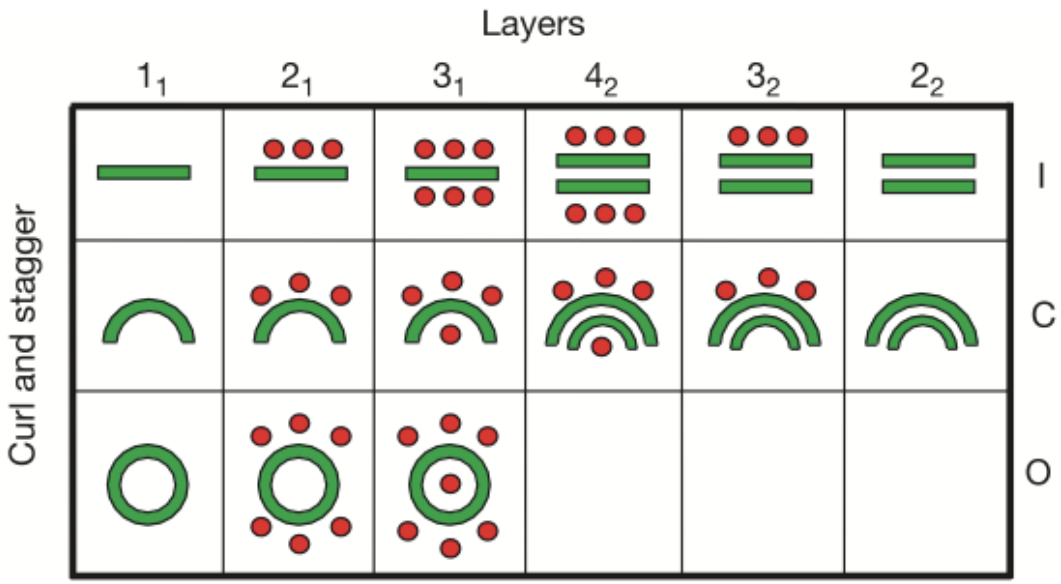


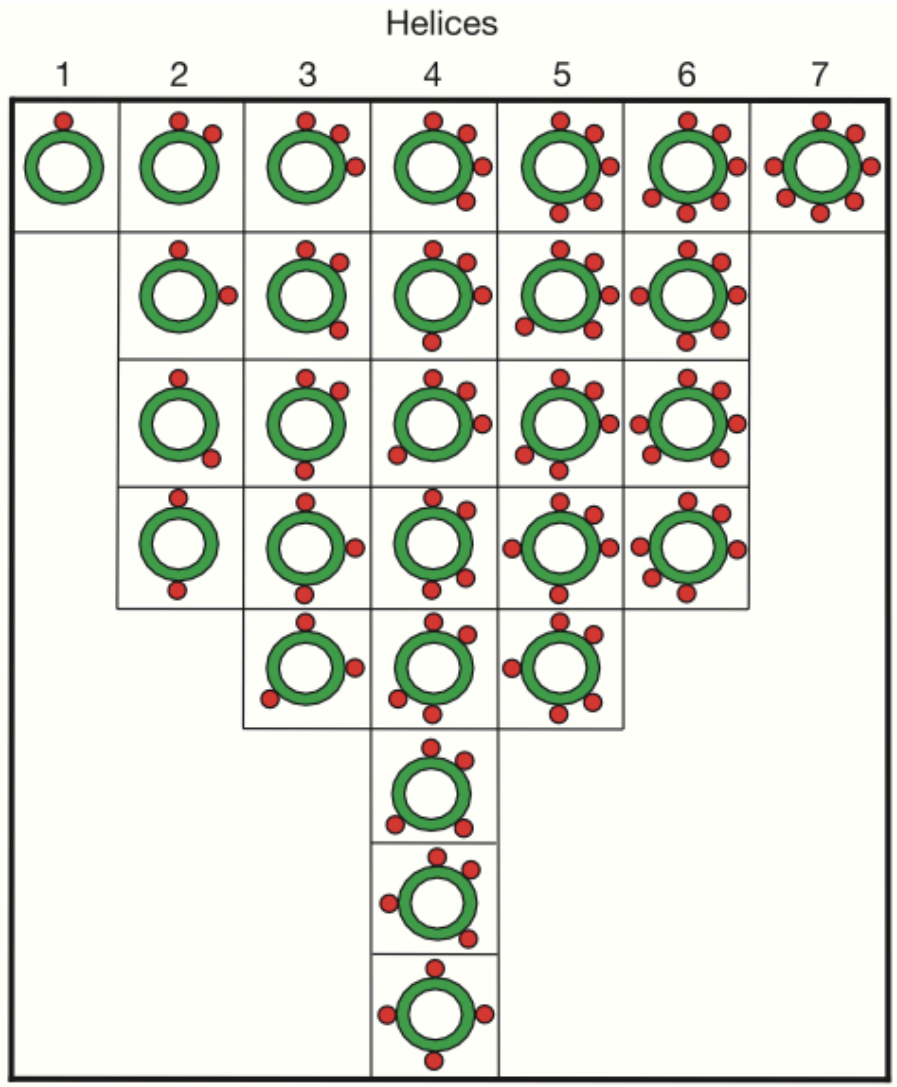


- An All vs. All comparison was done
- A multivariate scaling method to summarize relationships
- 5 attractors are detected corresponding to 40 of fold space

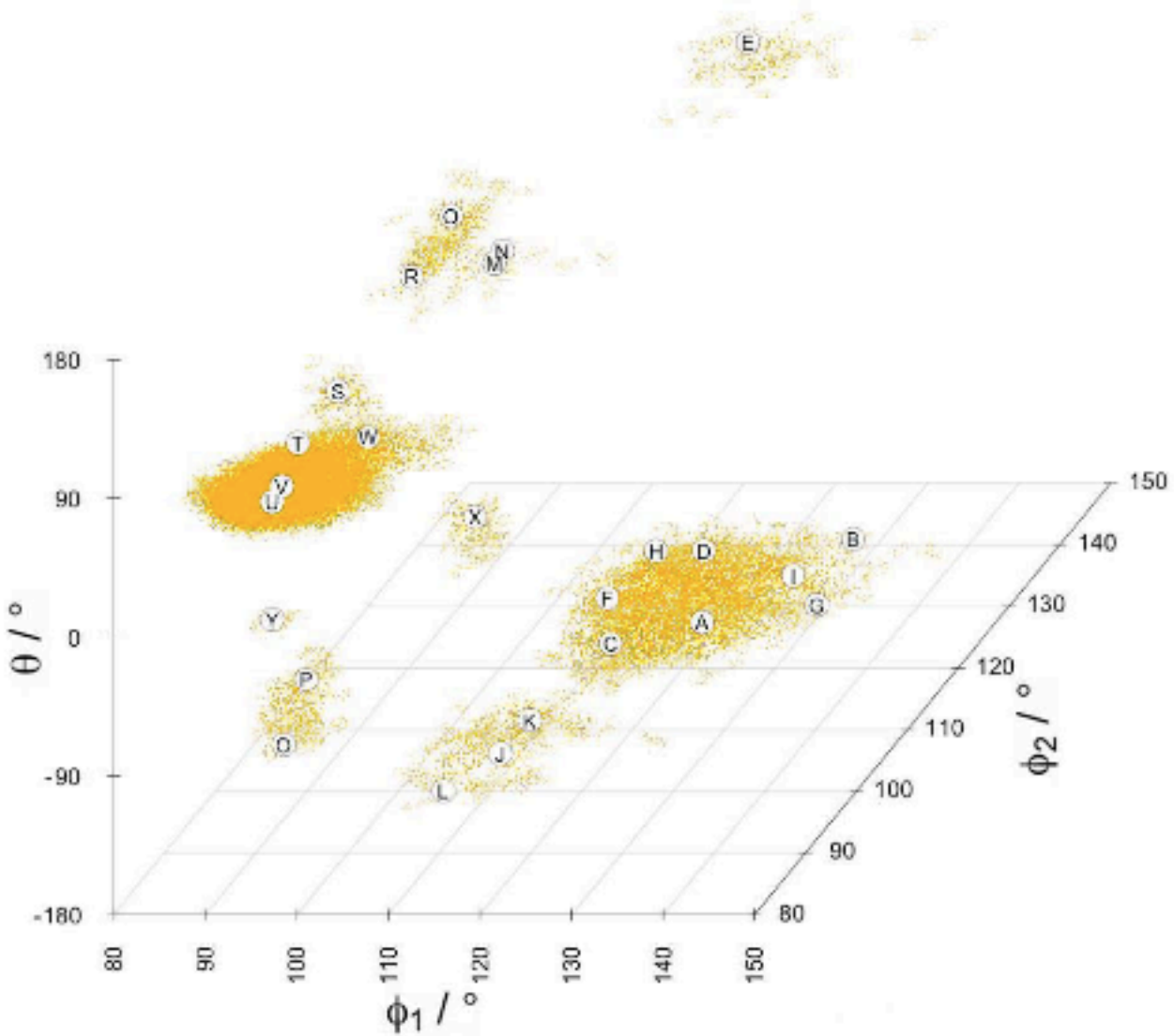
Holm and Sander. Science (1996) 273 595-

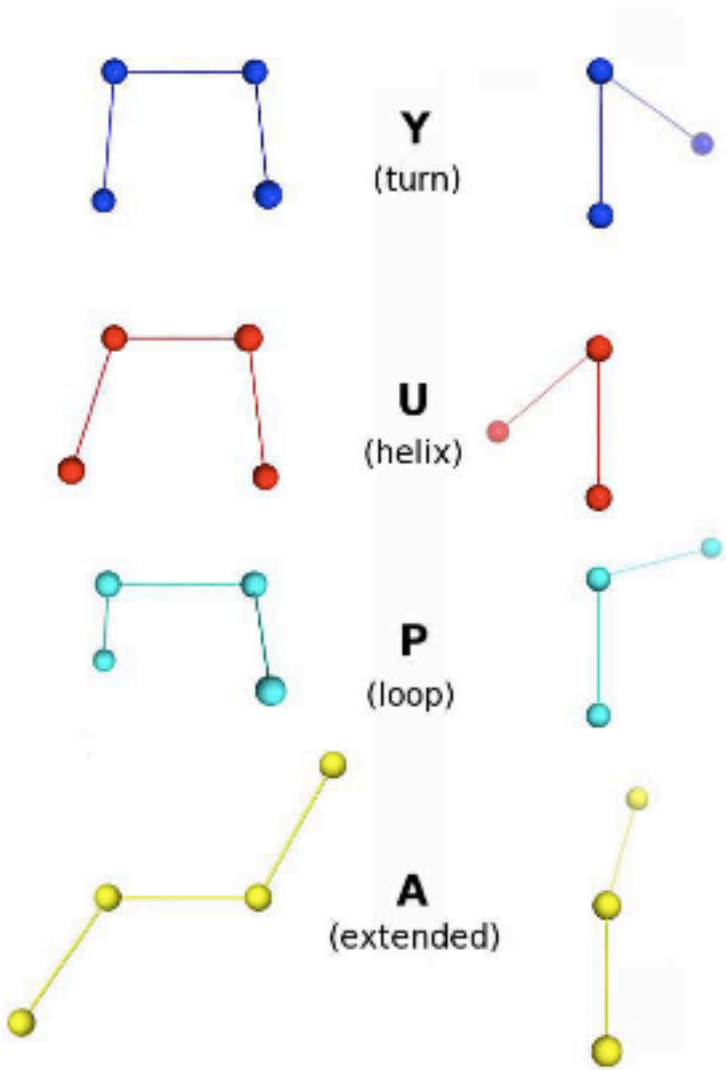
Fold Alphabet



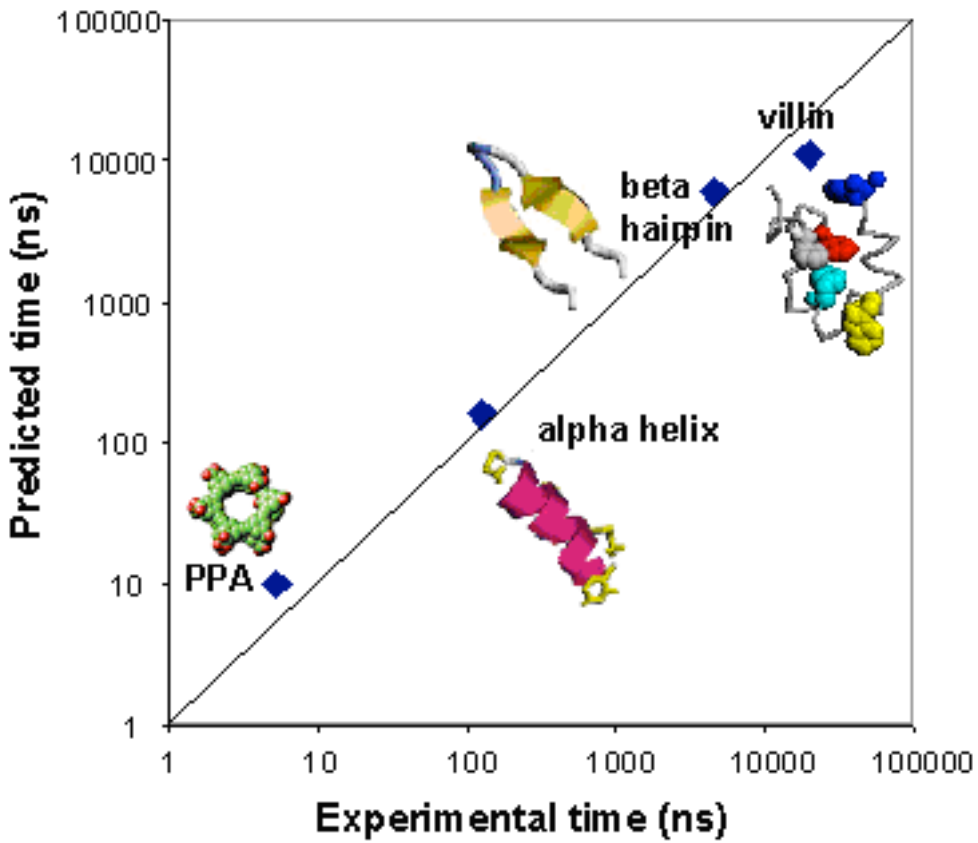


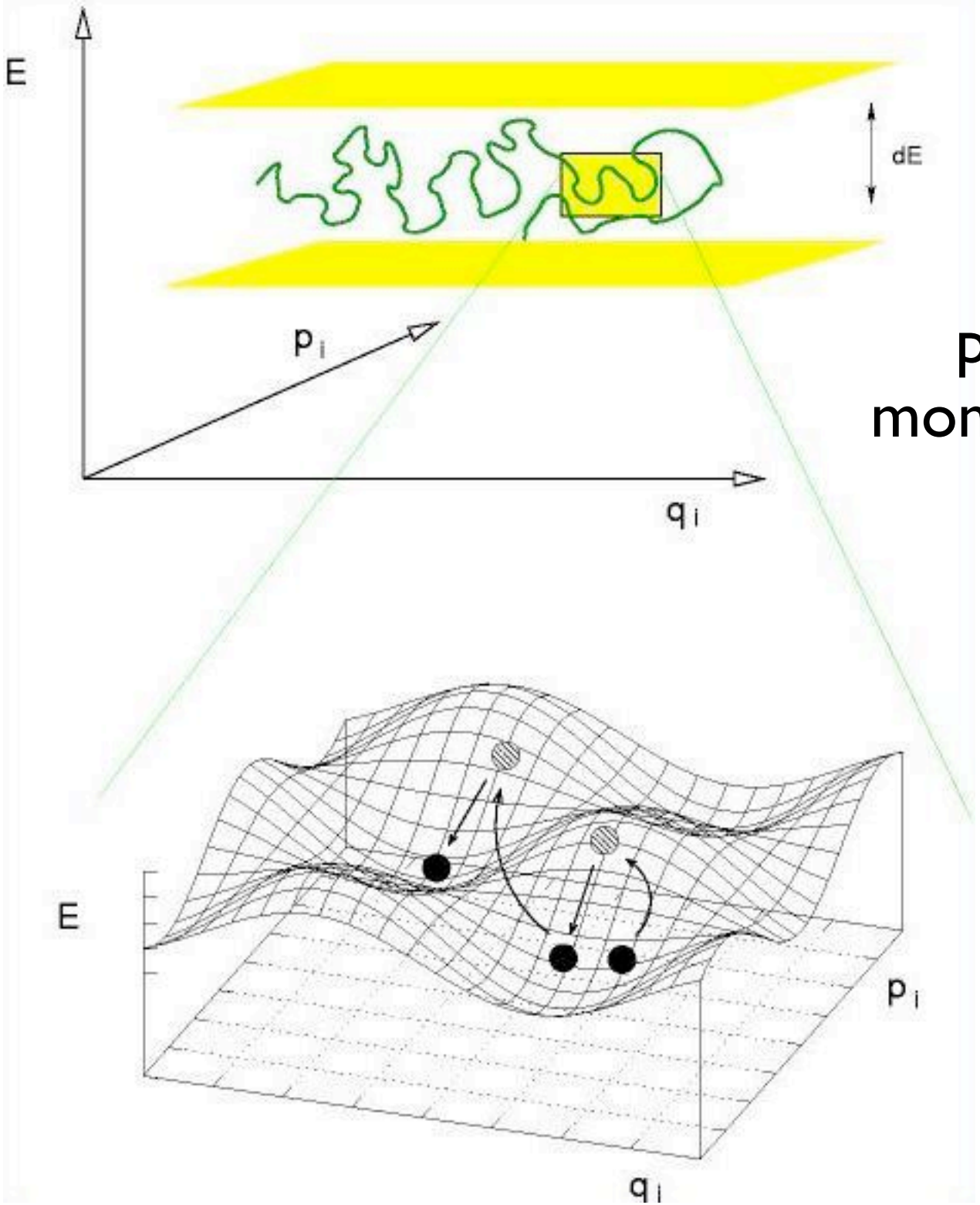
Fragment Alphabet





Folding Time Scales



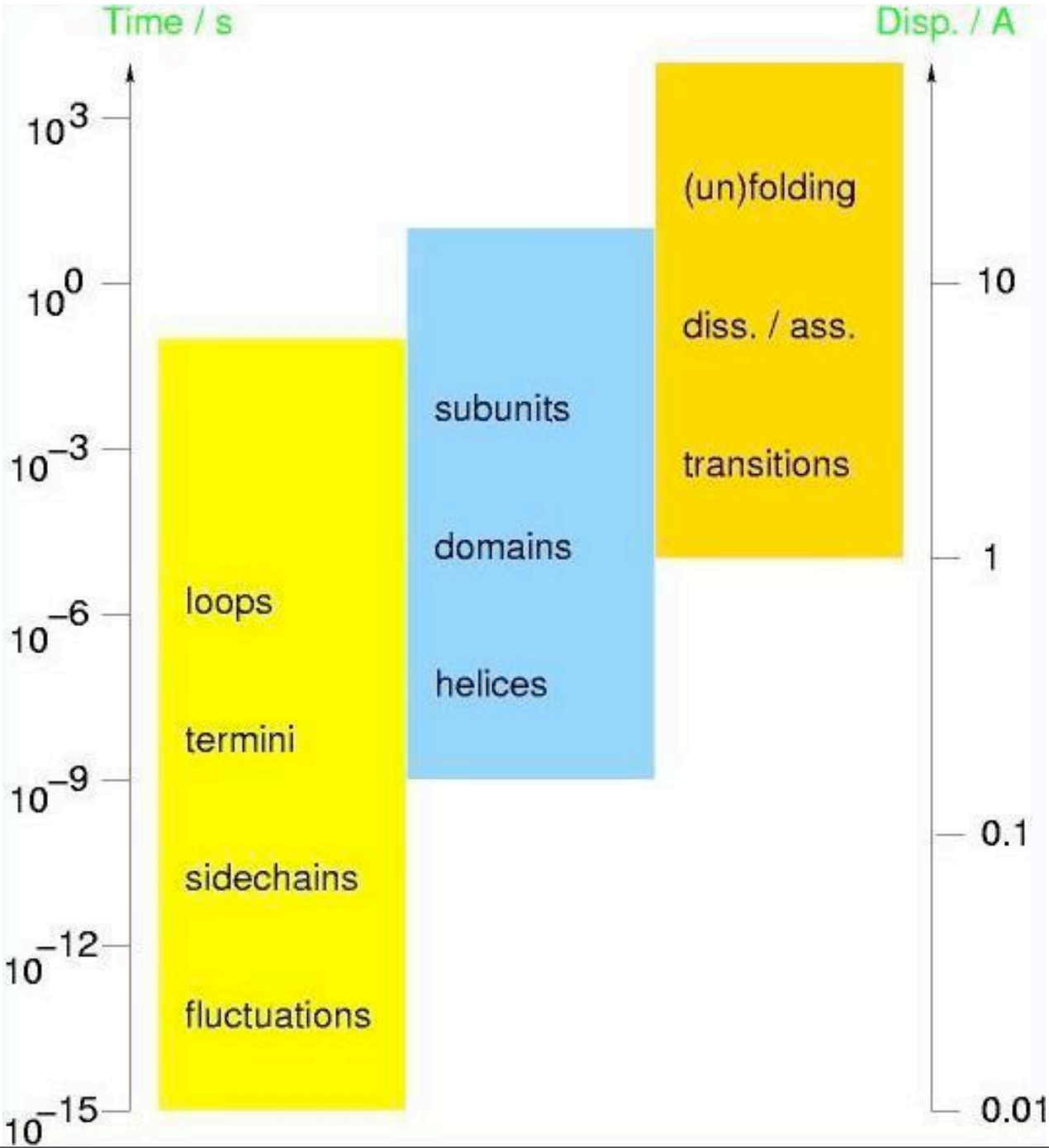


Energy hyper-surface

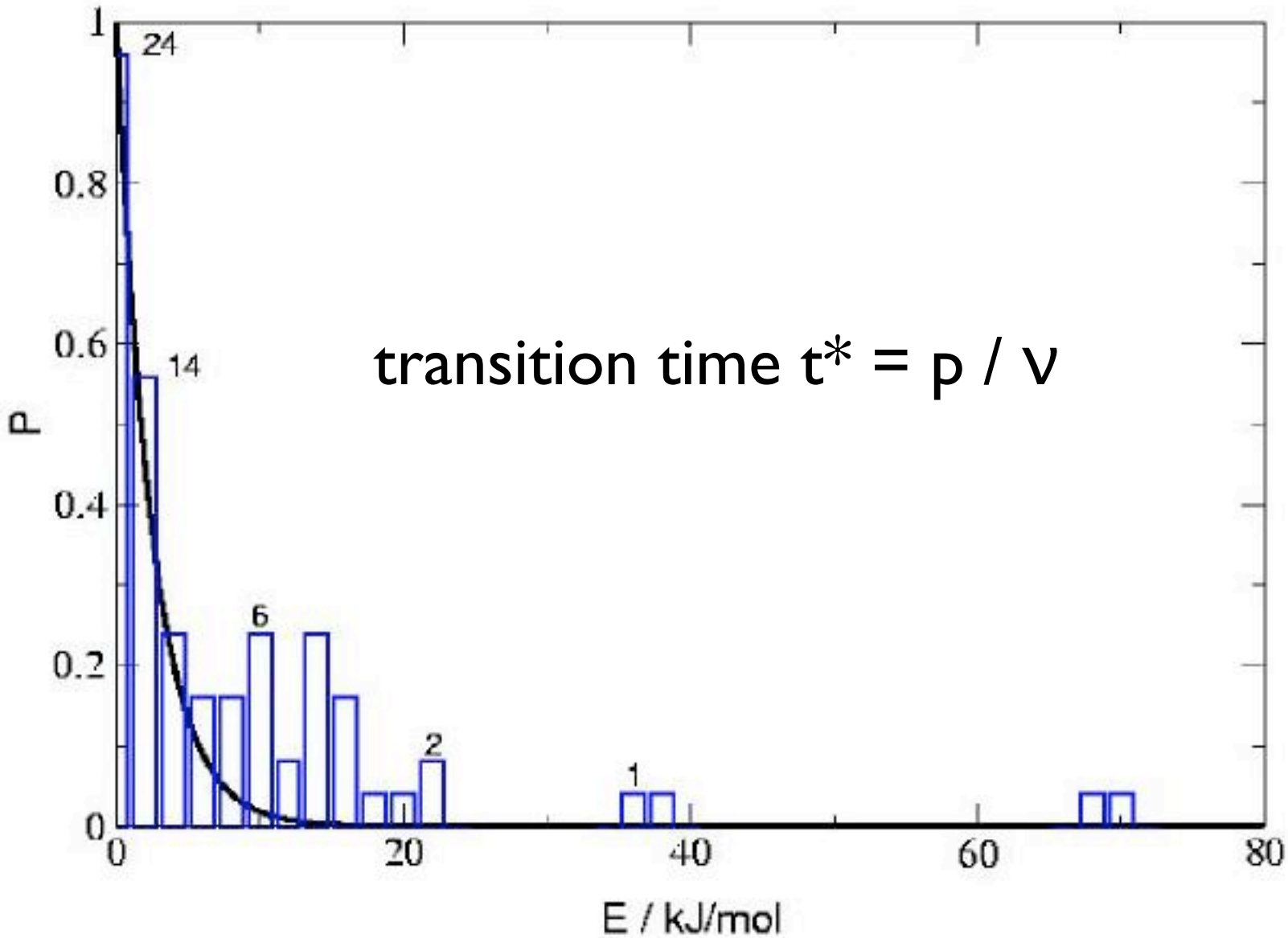
p_i, q_i (phase space): momenta and coordinate

random walk on hyper-surface

Motion/Displacement Time Scales

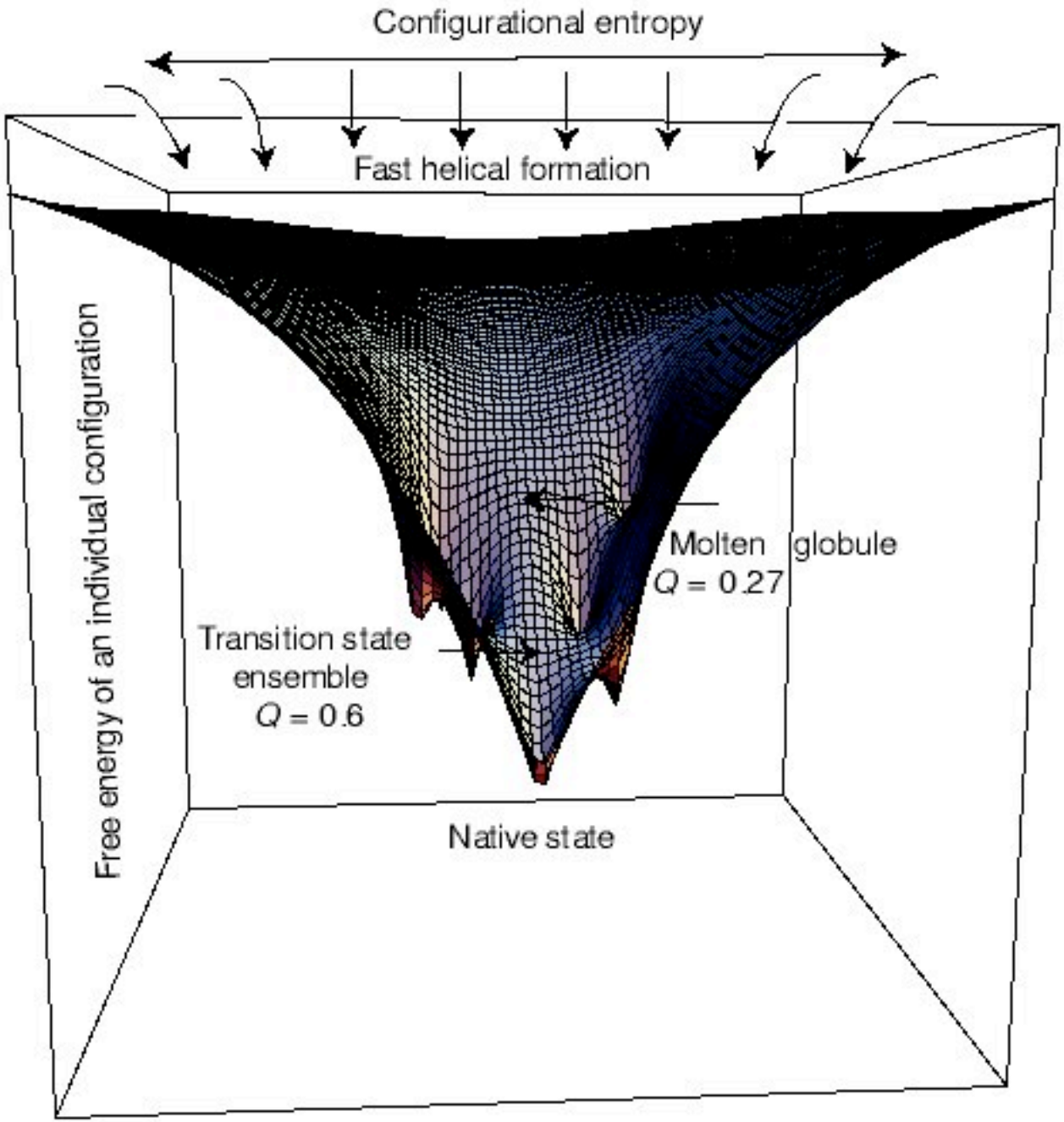


Energy Barriers



transition time $t^* = p / \nu$

Folding Funnel



Folding / Unfolding Reaction

Reaction

$$F(olded) \underset{k^{\leftarrow}}{\overset{k^{\rightarrow}}{\rightleftharpoons}} U(nfolded)$$

Equilibrium constant and kinetic constants

$$K = c(U) / c(F) = k^{\rightarrow} / k^{\leftarrow}$$

Thermodynamic equilibrium

$$\Delta G = \Delta H - T \Delta S$$

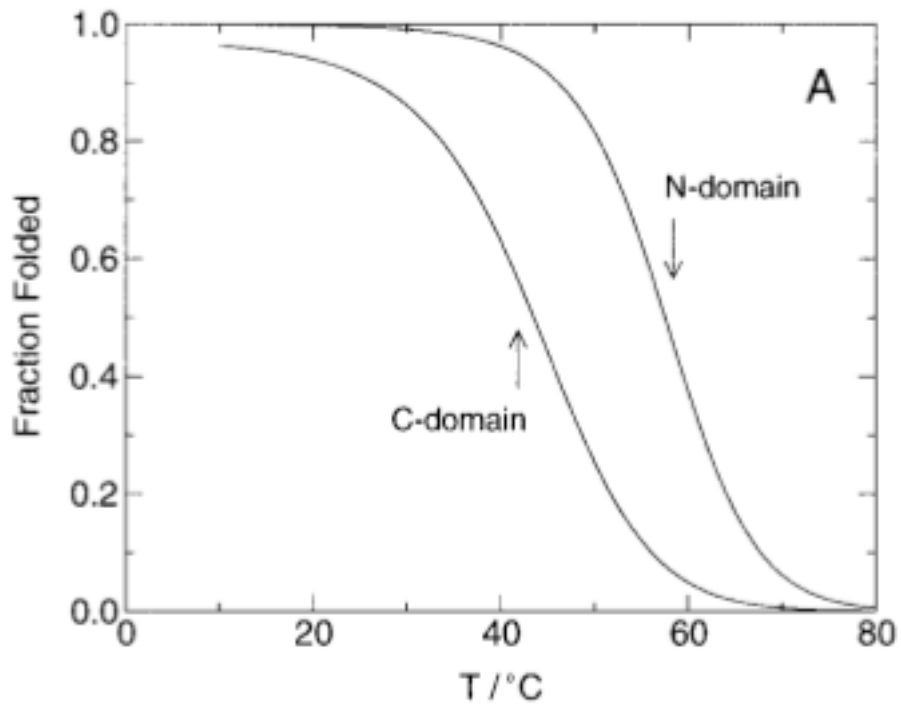
free enthalpy

enthalpy

entropy

Folding / Unfolding Equilibrium

Unfolding should be reversible, otherwise it is aggregation or some other process. That means in thermal unfolding, the heating and cooling curve should be identical.



The 0.5 transition point is called ‘unfolding temperature’.

Thermodynamic Stability of Proteins

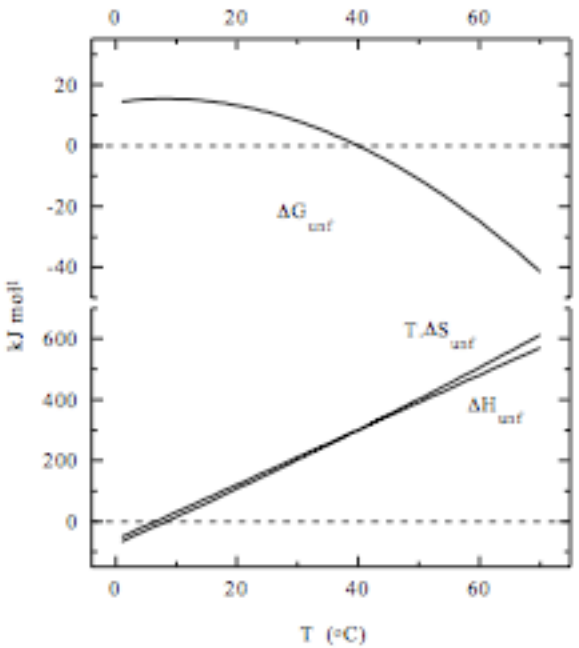


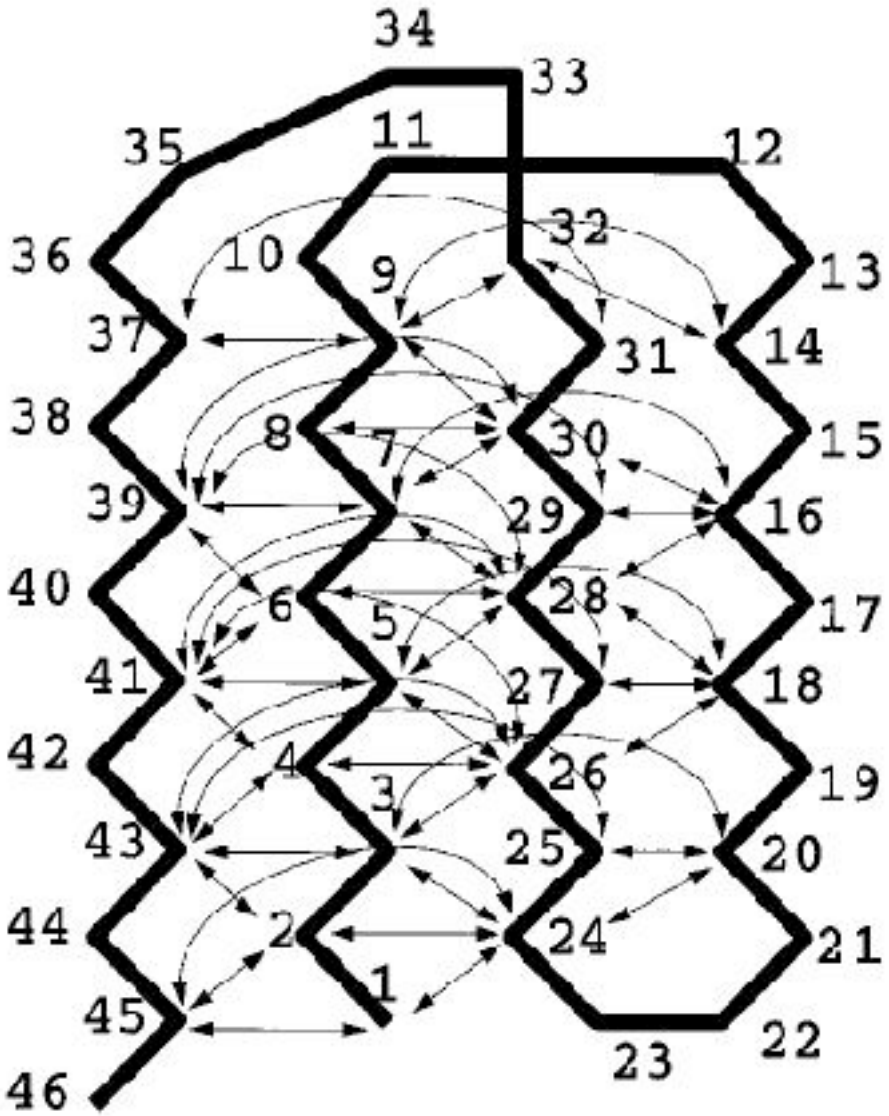
Fig.5: Characteristic temperature variation of thermodynamic parameters for unfolding of a small globular protein. Data are calculated for a typical protein unfolding at 40 ^{O}C (T_m) with $\Delta H_m = 300$ kJ mol^{-1} and assuming a constant $\Delta C_p = 9$ kJ K^{-1} mol^{-1} . Note how the relatively small unfolding free energy (ΔG_{unf}) is made up of the difference between relatively large enthalpic (ΔH_{unf}) and entropic (ΔS_{unf}) contributions. Temperature variation of ΔC_p would show as a curvature of the ΔH_{unf} and $T.\Delta S_{unf}$ lines.

A. Cooper, "Protein: A Comprehensive Treatise" , Volume 2, pp. 217-270 (1999)

(Folding) Simulations

Representation	Search (sampling)	Target Function
Cα	Grid (Go)	Free Energy
Cα, Cβ	off-Grid	Probability
Backbone	Monte-Carlo	% Native Contacts
All-atom	Dynamics	RMSD

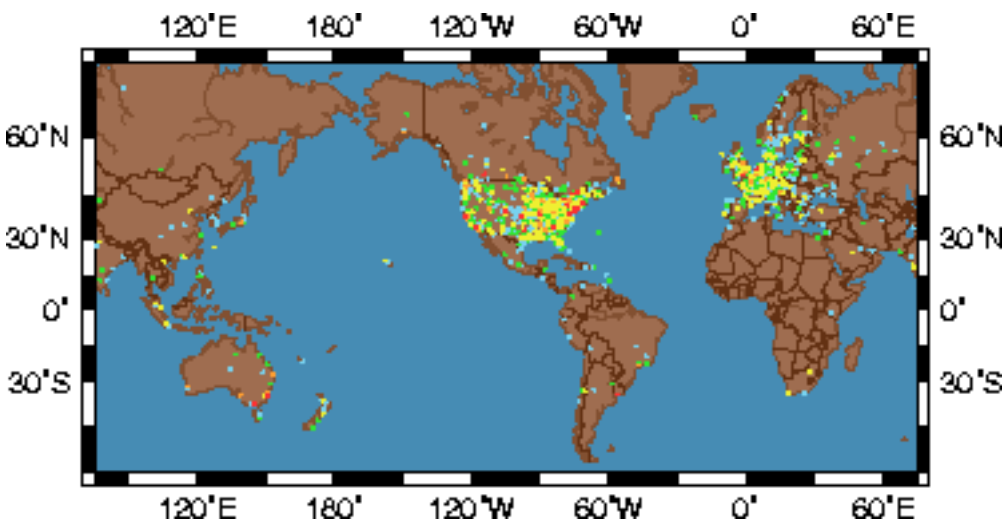
Go Model

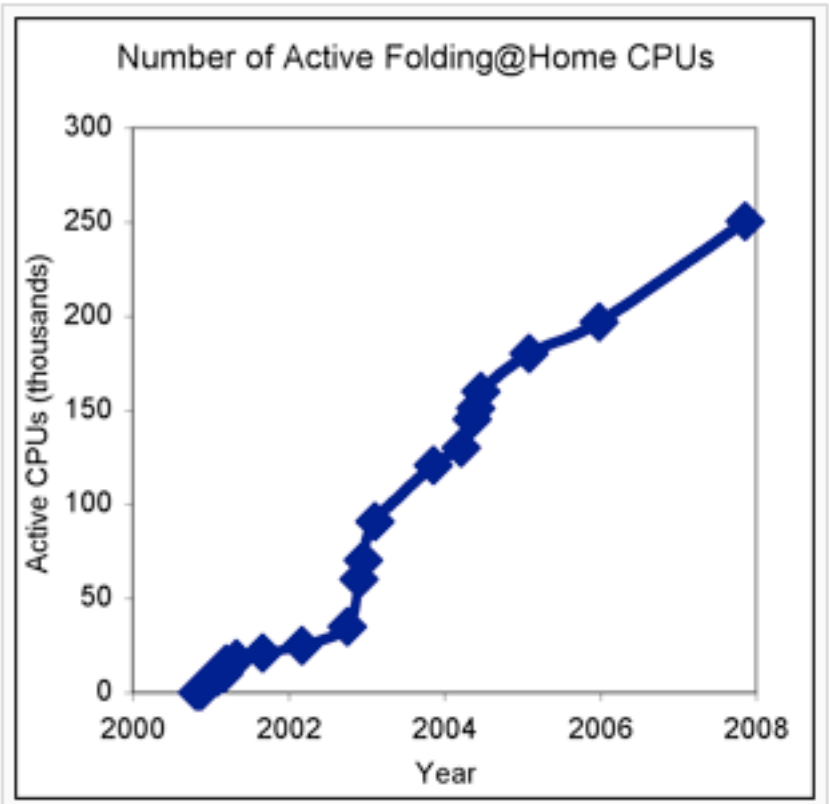


- highly reduced conformational space
- controllable sampling

Folding @ Home

Distributed Computing





ROSETTA - most successful *de novo* prediction method

- David Baker, U. Washington, Seattle
- Based on the idea that the possible conformations of any short peptide fragment (3-9 residues) are well-represented by the structures it is observed to adopt in the pdb
- Generate a library of different possible structures for each sequence segment
- Search the possible combinations of these for ones that are protein-like by various criteria

ROSETTA fragment libraries

- Remove all homologs of the protein to be modeled (>25% sequence identity)

- For each 9 residue segment in the target, use sequence similarity and secondary structure similarity (compare predicted secondary stucture for target to fragment secondary structure) to select ~25 fragments

- Because secondary structure is influenced by tertiary structure, ensure that the fragments span different secondary structures

- The extent to which the fragments cluster around a consensus structure is correlated with how good a model the fragment is likely to be for the target

ROSETTA search algorithm

Monte Carlo/Simulated Annealing

- Structures are assembled from fragments:
  - Begin with a fully extended chain
  - Randomly replace the conformation of one 9 residue segment with the conformation of one of its neighbours in the library
  - Evaluate the move:
    Accept or reject based on an energy function
  - Make another random move…
  - After a prescribed number of cycles, switch to 3-residue fragment moves

ROSETTA scoring function

$$P(\text{structure} \mid \text{sequence}) = \frac{P(\text{structure}) \times P(\text{sequence} \mid \text{structure})}{P(\text{sequence})}$$

sequence is constant
need to estimate for decoys built from fragments

Main contributions to P(structure):

- secondary structure packing
  (e.g. ensure β-strands form β-sheets)

-VdW packing

Simons et al. PROTEINS (1999) 34, 82-95

Success Stories

Design of the first artificial protein fold

Design of the first artificial functional enzyme

Design of multiple sequences for a given fold

Design of interfaces between proteins

CASP

9th Community Wide Experiment on t
Critical Assessment of Techniques for Protein Stru

Target List

Targets expire on specified date at noon (12:00) local time in California (GMT - 7 hours). If information leak occ
the models submitted within the initial 3 weeks only.
Yellow color - target expires for TS, AL, DR, RR, FN predictions within 48 hours
Orange color - target expires for TS, AL, DR, RR, FN predictions within 24 hours
Red color - target has expired for TS, AL, DR, RR, FN predictions, but is still open for QA predictions

View: All | Server only | Human and Server

#	Tar-id	Type	Residues	Method	Entry Date	Server Expiration	Human Expiration
1.	T0515	Human/Server	365	X-RAY	2010-05-03	2010-05-06	2010-05-24
2.	T0516	Server only	229	X-RAY	2010-05-03	2010-05-06	2010-05-24
3.	T0517	Human/Server	159	X-RAY	2010-05-04	2010-05-07	2010-05-25
4.	T0518	Server only	288	X-RAY	2010-05-04	2010-05-07	2010-05-25
5.	T0519	Server only	180	X-RAY	2010-05-04	2010-05-07	2010-05-25
6.	T0520	Human/Server	189	X-RAY	2010-05-05	2010-05-08	2010-05-26
7.	T0521	Server only	179	X-RAY	2010-05-05	2010-05-08	2010-05-26
8.	T0522	Server only	134	X-RAY	2010-05-05	2010-05-08	2010-05-26

Model Accuracy

HIGH ACCURACY

NM23
Seq id 77%

Cα equiv 147/148
RMSD 0.41Å

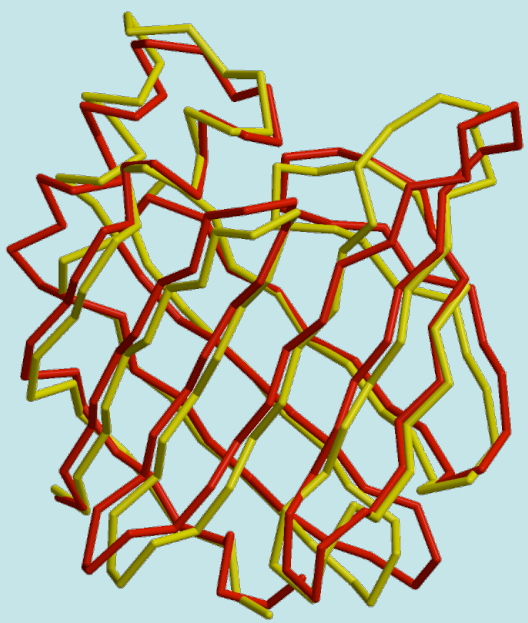
Side chains
Core backbone
Loops

X-RAY / MODEL

MEDIUM ACCURACY

CRABP
Seq id 41%

Cα equiv 122/137
RMSD 1.34Å

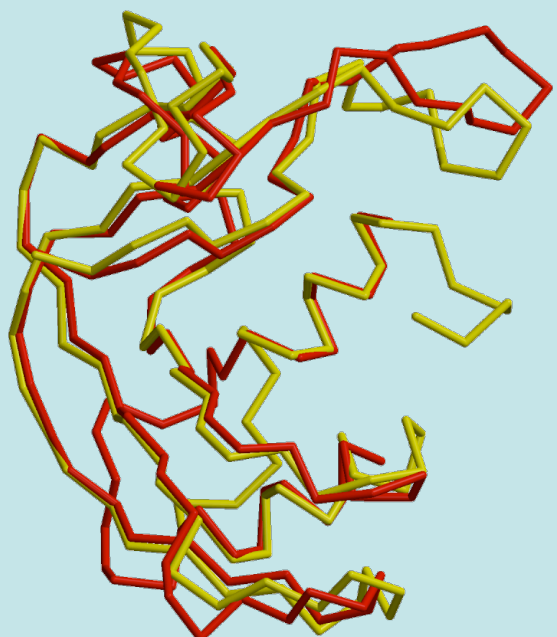


Side chains
Core backbone
Loops
Alignment

LOW ACCURACY

EDN
Seq id 33%

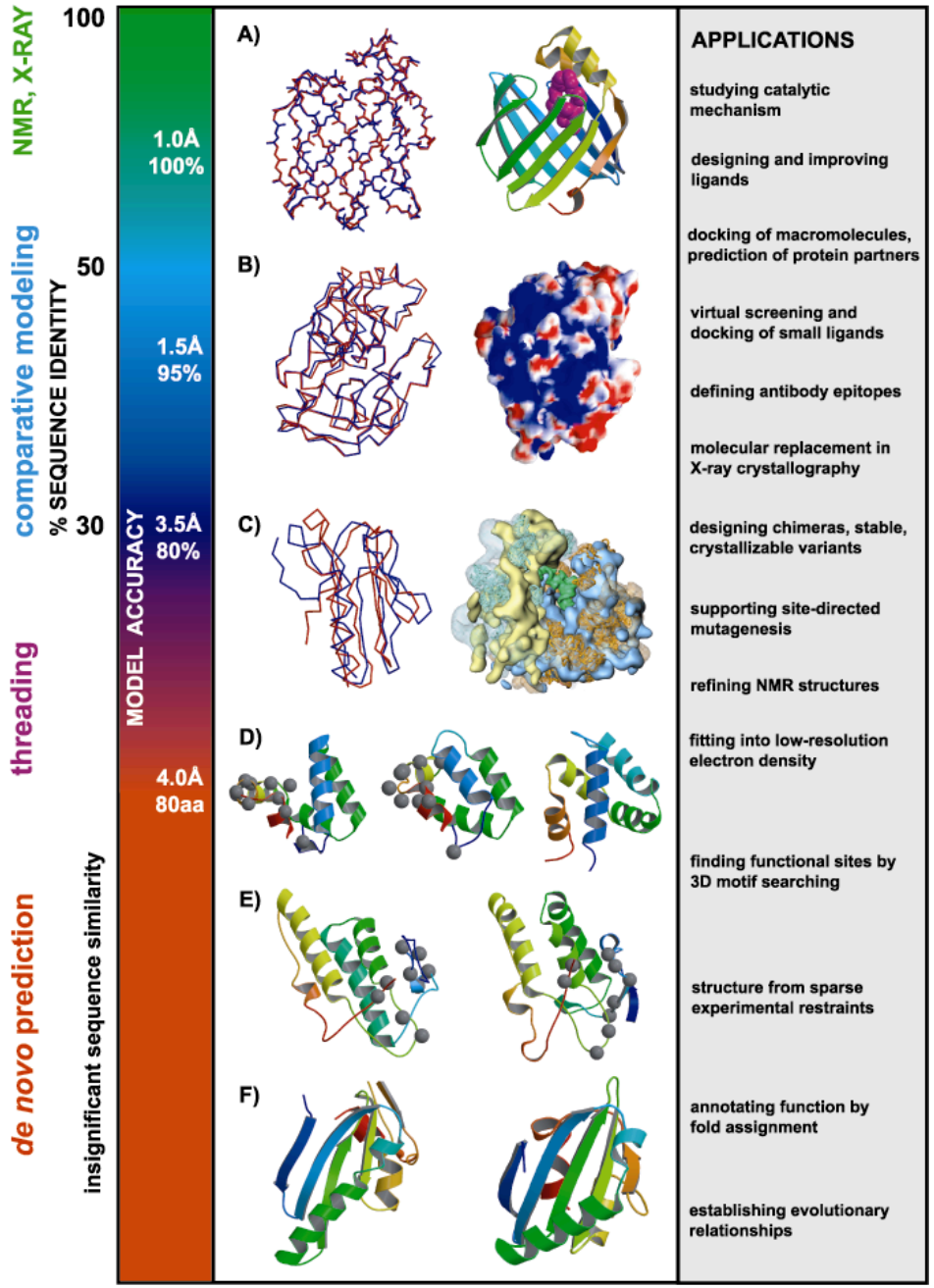
Cα equiv 90/134
RMSD 1.17Å



Sidechains
Core backbone
Loops
Alignment
Fold assignment

Marti-Renom et al. Annu.Rev.Biophys.Biomol.Struct. 29, 291-325, 2000.

Utility of protein structure models, despite errors



D. Baker & A. Sali. Science 294, 93, 2001.

Learning Outcomes

- Role of Chaperone
- Folding Funnel : Entropy versus Energy
- Kinetic Barriers and Fluctuations
- Folding as Reaction: Kinetics and Equilibrium
- Rosetta Method for Structure Prediction