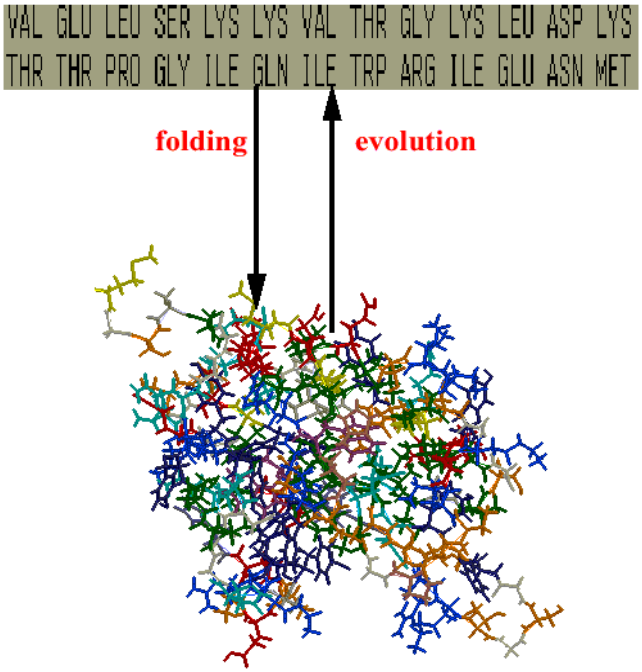
# Proteins can fold in vivo and in vitro (Anfinsen’59): protein folding problem

**PROTEIN FOLDING: THE INTELLECTUAL CHALLENGE.**

**HOW IS BIOLOGICAL 1d ORDER (INFORMATION) TRANSLATED INTO PHYSICAL 3d ORDER (STRUCTURE)?**



# THEORETICAL EFFORT IN PROTEIN FOLDING: STATISTICAL-MECHANICAL STUDIES.

## I Homopolymer microscopic models:('64-'82)

Ptistyn'64 (Flory-type), Lifshitz'68 (rigorous mean-field),DeGennes'75, Lifshitz, Grosberg, Khohlov'79, ES and A.Finkelstein'82 (theory of molten globules:homopolymers with side-chain freezing)

## II Phenomenological Model of N.Go ('69-)

Go'69'72- lattice models with strong bias ''by hands'' to the native state (''Principle of maximal consistency'') ,Jernigan'72, Dinner and Karplus'95, Pande nad Rokshar'97.

## III Exact theory (Zimm-Bragg) of local interactions:

Flory'58, Volkenstein'59, Scheraga'64-72, Finkelstein'75-82, Serrano'94....

## IV. Phenomenological ''landscape'' approach

(Bryngelson and Wolynes'87,89'90; Wolynes and coworkers '92-) An assumption about density of states is made at the beginning and its implications are explored.
''Principle of minimal frustartions'' suggested.

## V Phenomenological heteropolymer model: Dill'85

## VI.Microscopic Models of heteropolymers:analytical studies using replica approach.

The study starts from a microscopic hamiltonian. Density of states (''landscape'') emerges as a result, not as assumption.
Garel and Orland'88, ES and A.Gutin'89,
Sasai and Wolynes'91, ES and Ramanathan'94, Grosberg and Pande'95, ES and Archontes'94, ..... - still very active.

## VII Lattice model simulations: Ball and Fink,Broglia, Dill,Goldstein, Grosberg, Karplus, Onuchic, ES, Skolnick, Tang,Thirumalai, Unger,Wolynes.....

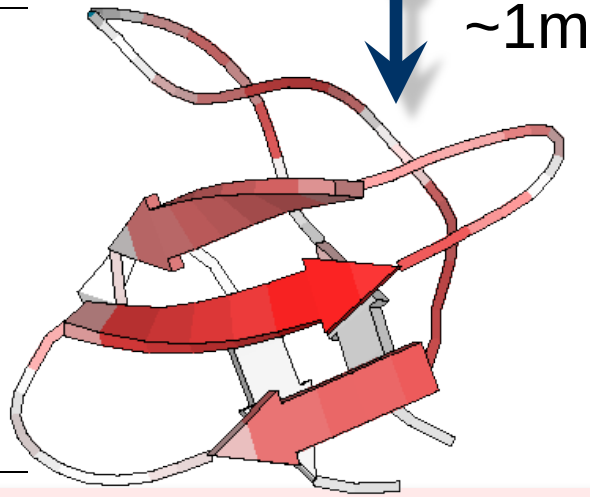
## IN 1994 THESE APPROACHES WERE NOTED BY BIOCHEMISTS AND CALLED ''A NEW VIEW''

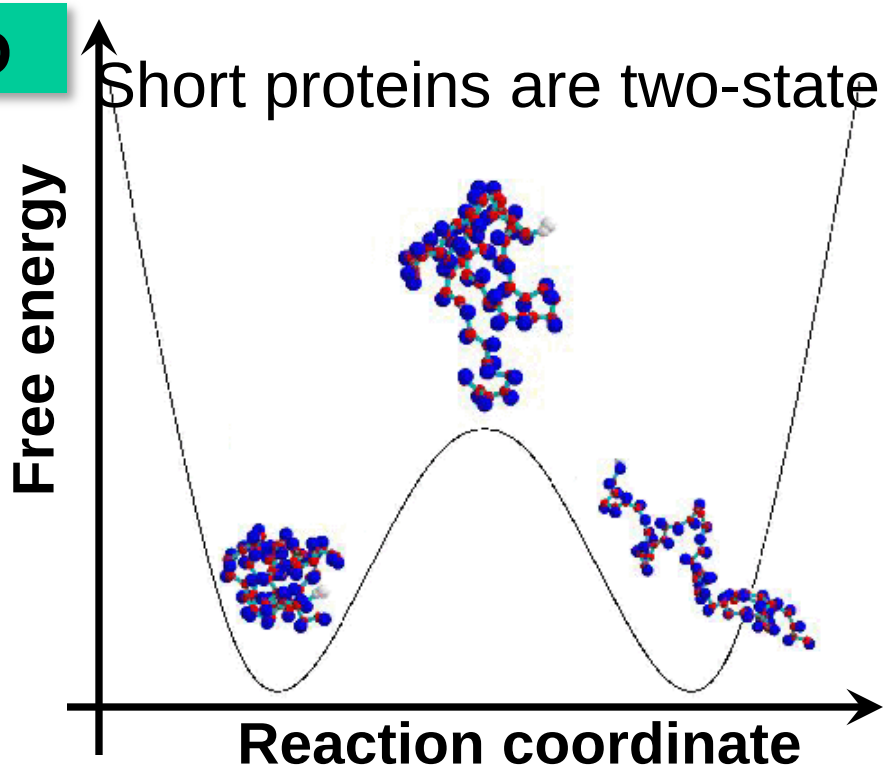
# PROTEINS: BASIC FACTS AND ENIGMAS

LDAPSQIEVKDVTDTTLAI...

~1ms - 1s

~10Å





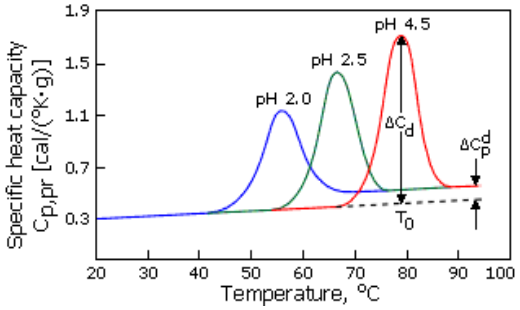

1. Protein sequence uniquely defines protein native (ground state) structure

2. Folding transition is cooperative (resembling first order transition)

3. Native state is thermodynamically stable

4. Native state is kinetically accessible --- reachable in a biologically reasonable time

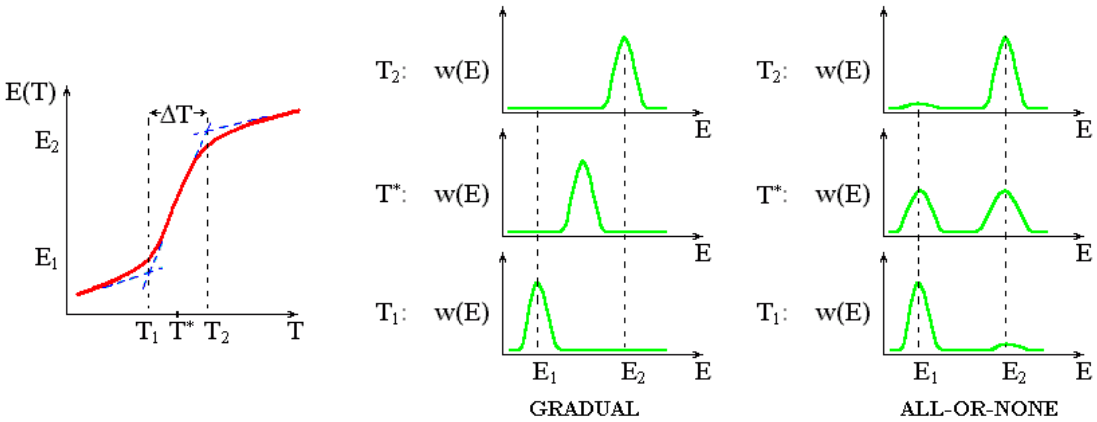
# Calorimetry: an important experimental test of cooperativity




**Calorimetric study of the lysozyme heat denaturation at various pH. The position of the heat capacity ($C_p$) peak determines transition temperature $T_0$, the peak width gives the transition width $\Delta T$, and the area under the peak determines heat $\Delta H$ absorbed by a gram of the protein. The values $\Delta T$, $\Delta H \times$protein's_M.W., and $T_0$ satisfy van't Hoff equations indicating hat the denaturation occurs as an *"all-or-none" (first order) transition*. The increased heat capacity of the denatured protein ($\Delta C_p$) originated from the enlarged interface between its hydrophobic groups and water after denaturation.**

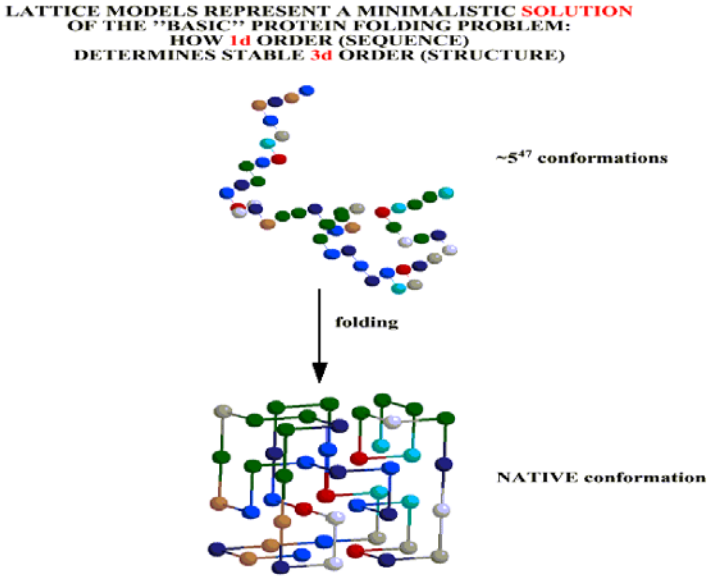
**Adapted from P.L.Privalov & N.N.Khechinashvili, J. Mol. Biol. (1974) 86:665-684.**

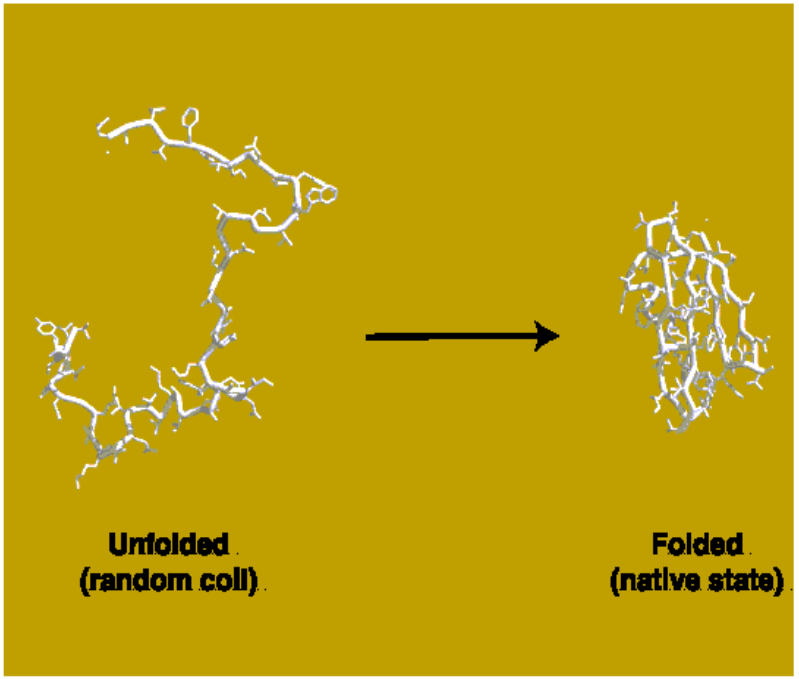
# ‘’All-or-none’’ first-order like transitions: is S-shape of a transition curve a good evidence?




**With the same temperature dependence of energy E (or any other observable parameter), the cooperative (“S-shaped”) transition can be either of the “all-or-none” type (example: protein denaturation), or gradual (example: helix-coil transition in polypeptides). The difference is displayed in the shape of the function W(E) showing distribution of molecules over energy (or over any other observable parameter) rather than in the shape of the curve E(T). Dashed lines in the left drawing explain a graphical determination f ΔT, the width of temperature transition.**

# Important remark – on protein models




**Proper selection of a model is a key to success! It depends on the Question you want to ask!**

# PHYSICAL QUESTIONS:

**A protein stays in its ground (micro) state at room temperature at which characteristic energy scale of all interactions is ~kT. (Thermodynamic uniqueness of protein structure)**

## Q1:

***a)*** **Is it a generic feature of any (i.e. random) sequence or some specially selected (designed) sequences only can fold?**

***b)*** **What features should be optimized to achieve thermodynamic uniqueness of protein structure?**

***c)*** **How many sequences can fold thermodynamically?**

## Q2:

**Obviously, proteins can find their ground state much faster than by exhaustive search (Kinetic accessibility of the native state).**

***a)*** **Which sequences can fold fast?**

***b)*** **What is the kinetic mechanism(s) of folding? – e.g. hierarchical or nucleation? (i.e. Second-order like vs First-order like)**

# Heteropolymer Hamiltonian: a basic model for study protein folding thermodynamics and kinetics

**The partition function:**

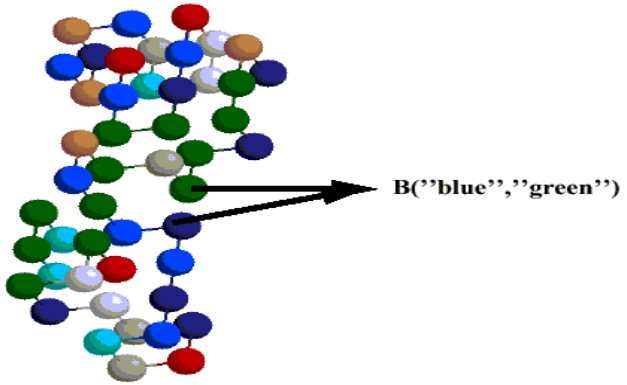
$$Z=\Sigma_{conf}\exp(-H_v(seq,conf))\prod_i g(r_{i+1}-r_i)$$

$g(r_{i+1}-r_i)$ describe local interactions along the sequence.

**The nonlocal Hamiltonian:**

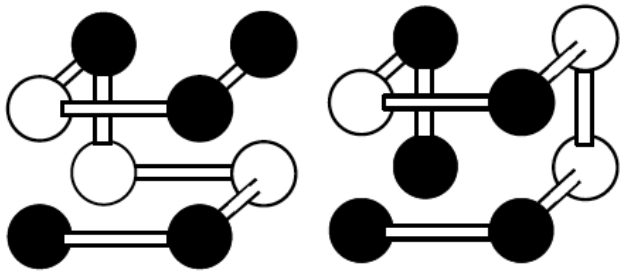
$$H_v(seq,conf)=\Sigma_{ij} B(\sigma_i,\sigma_j)U(r_i-r_j)$$

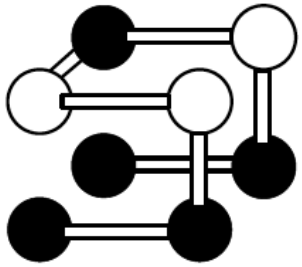
$\{\sigma_i\}$ describes polymer sequence



## Frustration in Polymers

- In a $2 \times 2$ cube there are only three different conformations

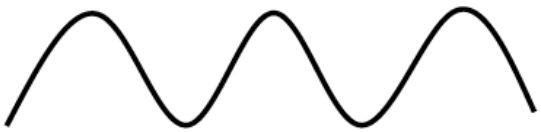



- A quenched sequence of monomers leads to frustration even for the two-letter case.

## Effect of Frustration

- In a homopolymer all different conformations with the same total number of contacts have exactly the same energy.



- On a lattice with coordination number $z$ the number of conformations is $(z/e)^N$
- In a frustrated system all conformations have unsatisfied contacts.
- The various conformations have different energies
- Freezing. Thermodynamic preference.

## The Polymer Hamiltonian

- Interactions between monomers in space

$$U_{ij} = B_{ij}\delta(\mathrm{r}_i - \mathrm{r}_j)$$

- In the two-letter problem (an A, B random sequence) $B_{ij}$ will be given by

$$B_{ij} = \chi\sigma_i\sigma_j + A\left(\frac{\sigma_i + \sigma_j}{2}\right) + B_0$$

$$B_{\mathrm{AA}} = B_0 + A + \chi$$
$$B_{\mathrm{BB}} = B_0 - A + \chi \qquad\qquad \sigma_i = \pm 1$$
$$B_{\mathrm{AB}} = B_0 - \chi$$

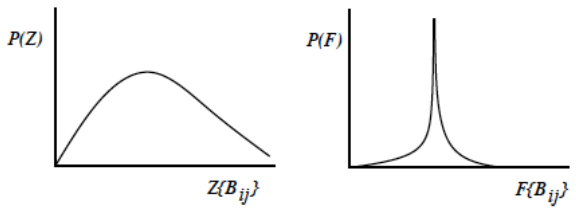
$$\chi = \frac{B_{\mathrm{AA}} + B_{\mathrm{BB}}}{2} - B_{\mathrm{AB}}$$

- Include polymeric bonds via

$$g(\mathrm{r}_{j+1} - \mathrm{r}_j) = \frac{1}{(2\pi a^2)^{3/2}} \exp\left[-\frac{(\mathrm{r}_{j+1} - \mathrm{r}_j)^2}{2a^2}\right]$$

## Averaging over Randomness

- Need to average over disorder.
- Self-averaging.
- Rare sequences with low energies (e.g. homopolymer) contribute significantly to the partition function.
- The contribution of these sequences to free energy is minimal.




- Replica Trick

$$\langle \ln Z \rangle_{av} = \lim_{n \to 0} \frac{1}{n} \left( \left[ Z^n \right]_{av} - 1 \right)$$

# Main steps in analytical theory
# (just an outline: for more detail see 2006 ChemRev)

**The key issue**: **How to average over sequences?**

**Prescription: Consider only self-averaging (SA) quantities – whose average is representative of majority of realizations.**

**Bad news: partition function is not SA (why???)**

**(Sort of) Good news: Free energy (i.e. log of Z) is SA!!!!**

$$< F(T) >= -kT \sum_{\{\sigma\}} P(\{\sigma\}) \log(Z(\{\sigma\}, T)$$

Where summation is taken over all $20^N$ sequences

and P ({s}) is the probability to find a sequence {s} in the ensemble

$P_r(\{\sigma\}) = P(\{\sigma\}) = \frac{1}{20^N}$ **For Random Sequences!**

However, **for evolutionary selected** sequences $P(\{\sigma\}) \neq P_r(\{\sigma\})$

**Therefore: Evolution enters theory via P({σ})!**

**ES and Gutin, Europhys Lett' 89, ibid Biophys Chem' 89, ibid JPhysA' 89 (d<=2 case)**

## Order Parameters

- Density of monomers

$$\rho(\mathrm{R}) = \sum_i \delta(\mathrm{r}_i - \mathrm{R})$$

- Phase separation of A from B monomers.

$$m(\mathrm{R}) = \rho_{\mathrm{A}} - \rho_{\mathrm{B}} = \sum_i \sigma_i \delta(\mathrm{r}_i - \mathrm{R})$$

- Overlap between conformations $\alpha$ and $\beta$

$$Q_{\alpha\beta}(\mathrm{R}_1, \mathrm{R}_2) = \sum_i \delta(\mathrm{r}_i^{\alpha} - \mathrm{R}_1)\delta(\mathrm{r}_i^{\beta} - \mathrm{R}_2)$$

  with

$$\int d\mathrm{R}_1 Q_{\alpha\beta}(\mathrm{R}_1, \mathrm{R}_2) = \rho_{\beta}(\mathrm{R}_2)$$

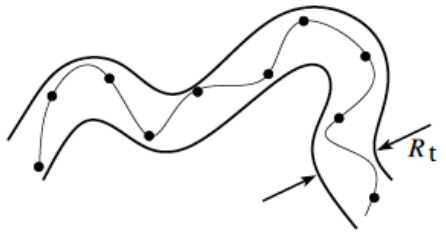
- For maximally compact state

$$Q_{\alpha\beta}(\mathrm{R}_1, \mathrm{R}_2) = Q_{\alpha\beta}(\mathrm{R}_1 - \mathrm{R}_2)$$

- $Q_{\alpha\beta}$ can be expressed as

$$Q_{\alpha\beta}(\mathrm{R}_1 - \mathrm{R}_2) = \frac{\rho}{R_t^3}\varphi_{\alpha\beta}\left(\frac{\mathrm{R}_1 - \mathrm{R}_2}{R_t}\right)$$

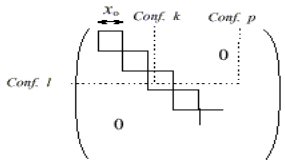
- Polymer in a tube



# Solution for the Q-order parameter

**Results for $Q_{\alpha\beta}$**

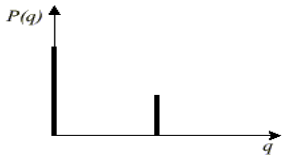
- The form of the solution for $Q_{\alpha\beta}$ gives a map of common contacts between different conformations.




- The matrix elements are

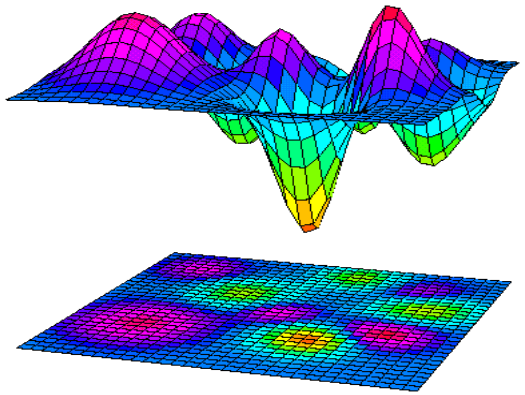
$$Q_{\alpha\beta} = \begin{cases} \rho\delta(\mathrm{R}_1 - \mathrm{R}_2) & \text{for } \alpha,\beta \text{ in the same group} \\ 0 & \text{for } \alpha,\beta \text{ in different groups} \end{cases}$$

- Conformations of different groups have no overlap
- In the frozen phase $P(q) = x_0\delta(q) + (1 - x_0)\delta(q - 1)$



# Potential energy surface for Random Heteropolymers

**Potential Energy Surface**



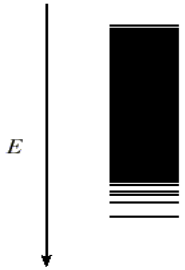
- Define $P(q) = \sum_{\alpha,\beta} P_\alpha P_\beta \delta(Q_{\alpha\beta} - q)$
- At high temperature $P(q) = \delta(q)$

# Density of states (spectrum) of the RHP

**Energy spectrum**

- Resulting energy spectrum and the R.E.M.

$$H = \sum_{ij} B_{ij}\delta(\mathbf{r}_i - \mathbf{r}_j)$$

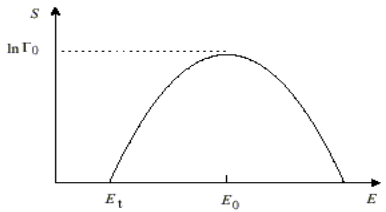



$$P(q) = x_0\delta(q) + (1 - x_0)\delta(q - 1)$$

- Conformations in the bottom of the spectrum are entirely different

# Random Energy Model

**The Random Energy Model**




- The energy levels follow a random distribution

$$P(E) = \Gamma_0 \exp[-(E - E_0)^2/\sigma^2]$$

- The entropy for each level is given by

$$S(E) = \ln \Gamma_0 - \frac{(E - E_0)^2}{\sigma^2}$$

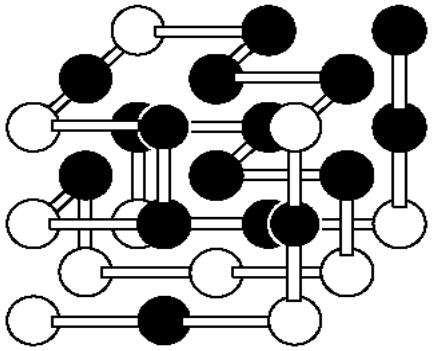
- Freezing occurs when $S \sim 0$
- The transition temperature is given by the slope

$$\frac{1}{T} = \frac{dS}{dE}$$

at $E_t$

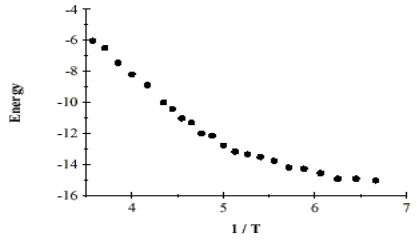
# Lattice Model to test of the results of Random Heteropolymer Theory

**Lattice Model**

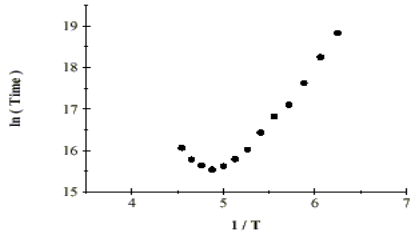


- $\Gamma_0 = 103346$, $K = 28$, $B = -1$, $B_0 = 1$
- $E_{\text{conf}} = -(K_{bb} + K_{ww})$

# Lattice Model analysis of RHP: the results



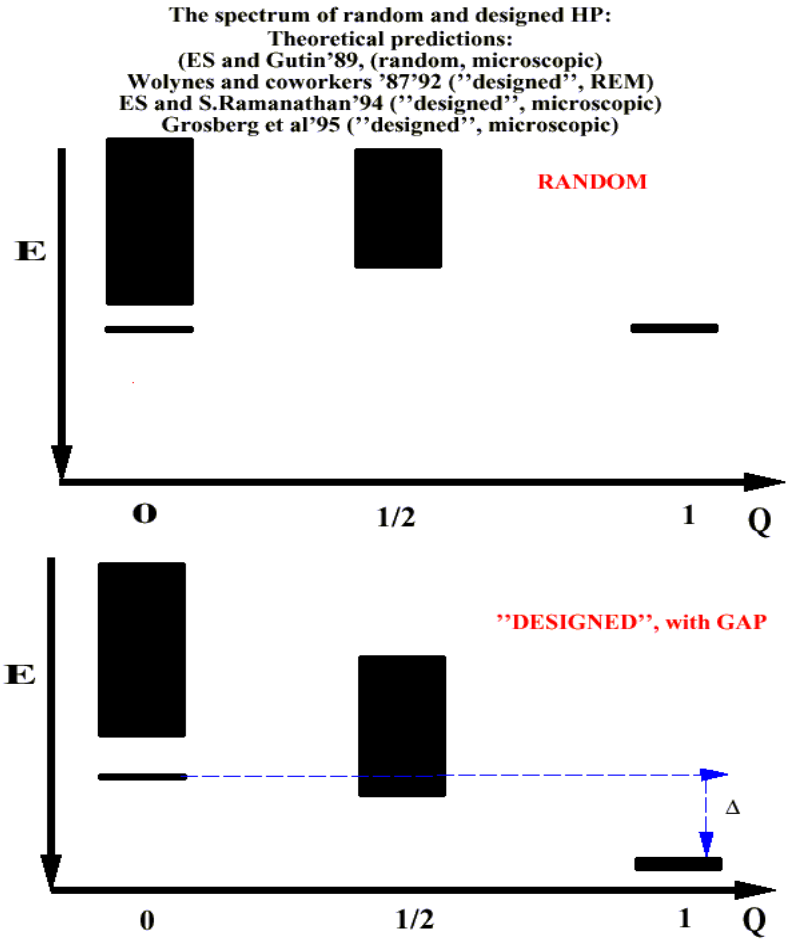

a




b

- Experiment: Quasirandom sequence of F2 fragment of tryptophane synthase (Chafotte *et al.*, 1991) the transition is noncooperative.
- Experiment: "Binary design" (M.Hecht *et al*, 1995)

# Why is Random HP a bad model of protein: The case for Evolution




**This picture is valid only for d>**

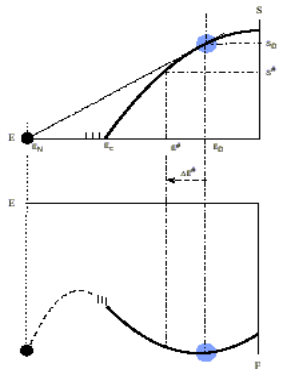
**The ''freezing'' transition in RHP is *non-cooperative***

# Solution: Designed Sequences!

IS RANDOM HP A GOOD MODEL FOR PROTEIN:

- RHP Exhibit thermodynamically noncooperative "freezing" transition.
- Folding transition in proteins is cooperative (Privalov '73,'79..).

A POSSIBLE SOLUTION: DESIGNED SEQUENCES



WARNING: THIS IS A CARTOON. IT MAY BE VALID FOR A "TYPICAL" 3D CONFORMATION, BUT NOT FOR ALL CONFORMATIONS AND NOT FOR $D \leq 2$

# Statistical Mechanics of Protein-Like Sequences with Gaps

- **Now we understand (or anticipate) that protein-like sequences should be selected to have large energy gap to the native state**
- **Thus the sequence ensemble (i.e. P({σ}) should be shifted accordingly:**

$$P_{T_{sel}}(\{\sigma_i\}) = \exp\left(-\frac{H\left(\{\sigma\},\{r^0\}\right)}{T_{sel}}\right) = \exp\left(-\frac{\sum_{i<j} B(\sigma_i,\sigma_j)U(r_i^0 - r_j^0)}{T_{sel}}\right)$$

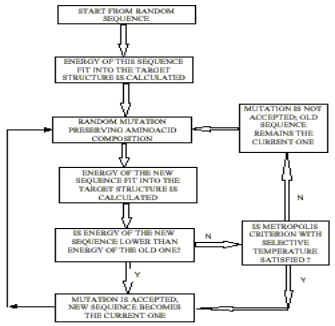
**Where $\{r^0\}$ determines native conformation, $T_{se}$ is a measure of evolutionary selection**

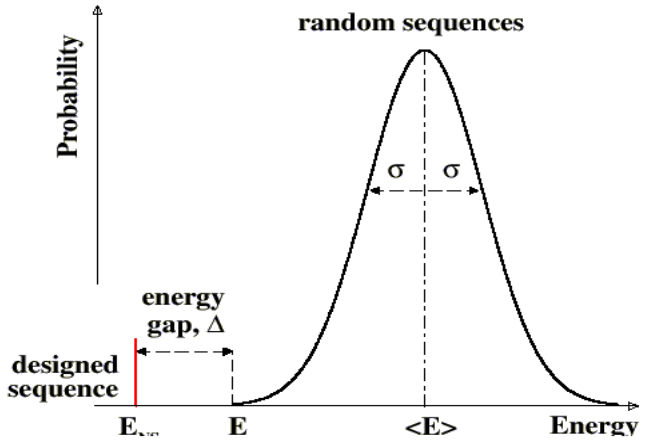
Statistical mechanics *and dynamics* in Sequence Space: **Evolution**
(See lecture 1)

# Algorithm of design: Monte-Carlo in Sequence Space

### Algorithm of Selection & Design

- Microcanonical ensemble vs
- Canonical ensemble with evolutionary temperature: MC procedure in sequence space:
- $E = E_0(\{\sigma_i\}, \{r_i^0\})$
- $\{\sigma_i\}$ is a sequence ($\sigma_i = 1...20$ is a type of monomer i)

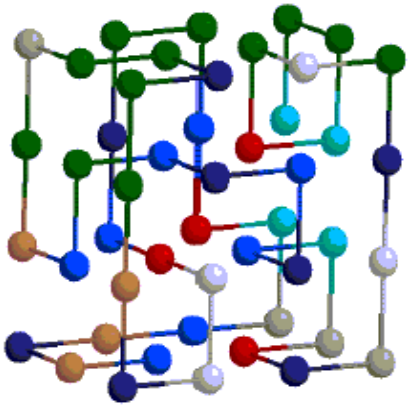


# LATTICE MODELS REPRESENT A MINIMALISTIC SOLUTION OF THE ''BASIC'' PROTEIN FOLDING PROBLEM: HOW 1d ORDER (SEQUENCE) DETERMINES STABLE 3d ORDER (STRUCTURE)

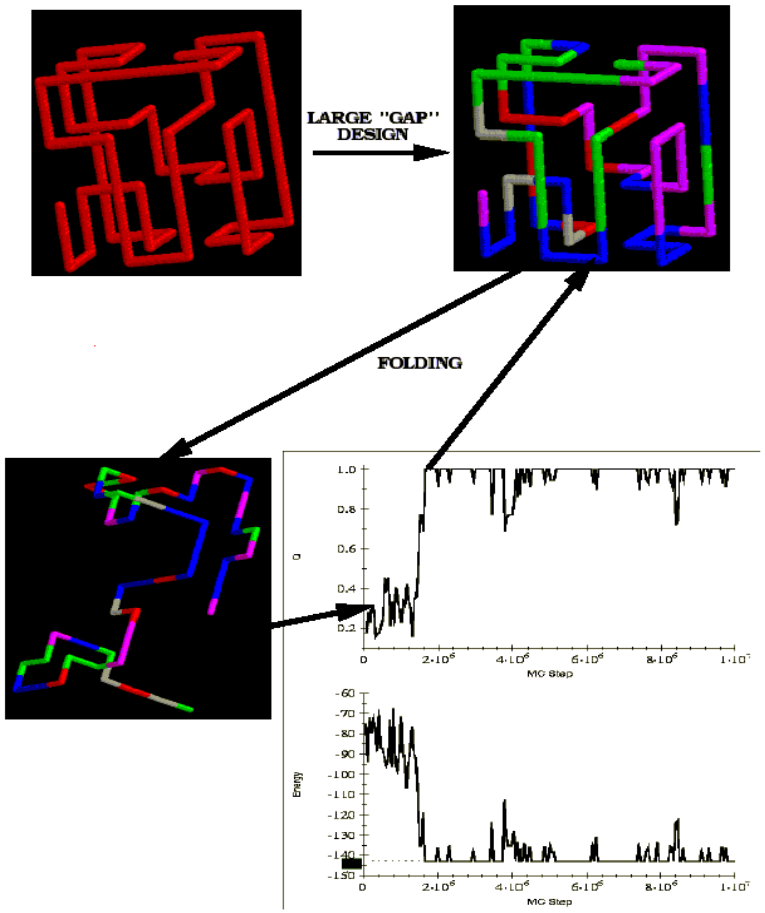
~$5^{47}$ conformations

folding



NATIVE conformation

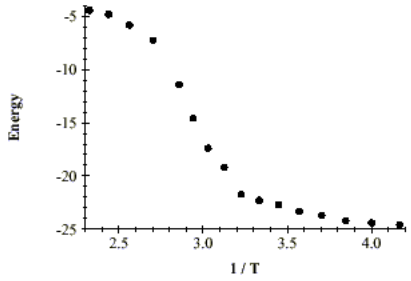
# Testing the design idea in simulations: the design-folding paradigm



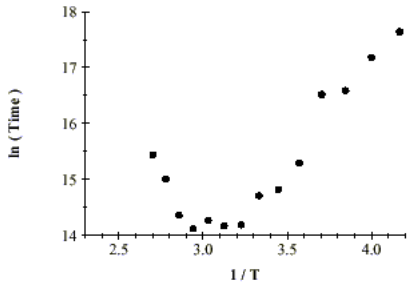
# Design works

THE DESIGN WORKS:

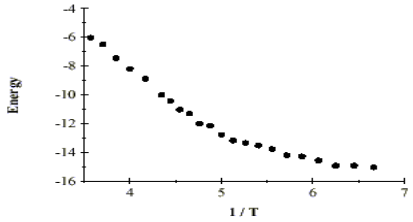
- Designed peptides fold cooperatively



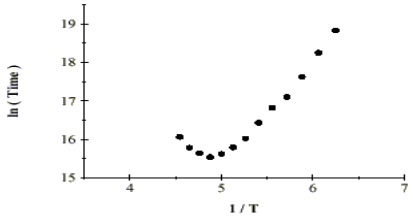

a




b



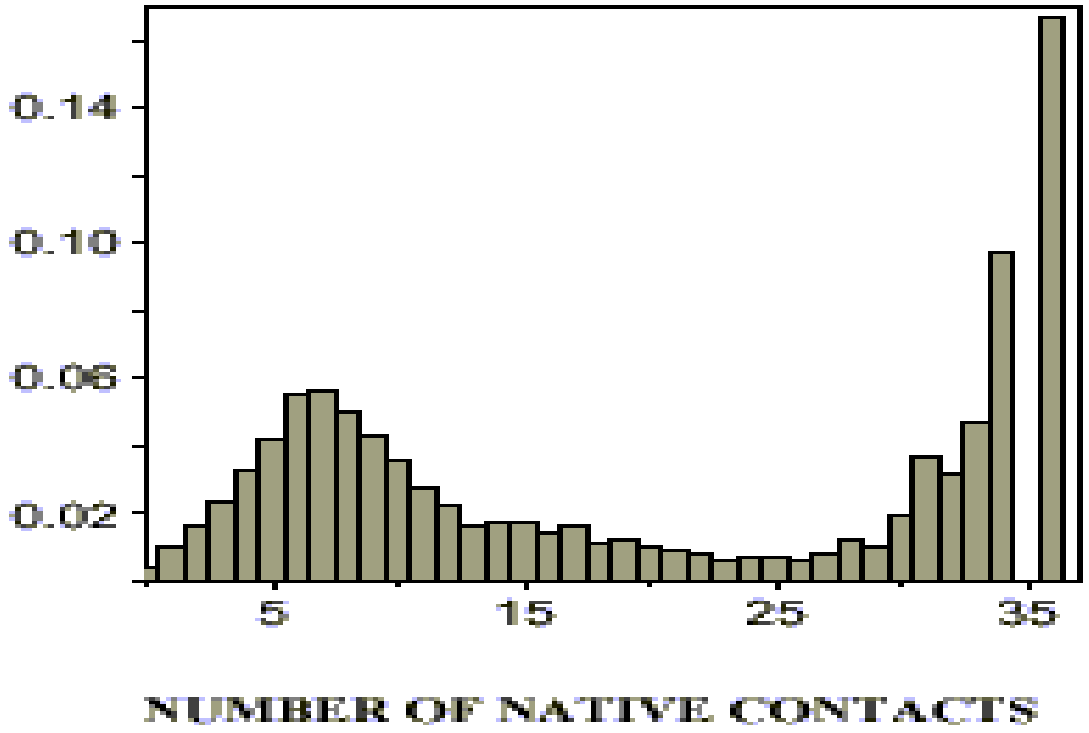

a




b

- Experiment: Quasirandom sequence of F2 fragment of tryptophane synthase (Chafotte *et al.*, 1991) the transition is noncooperative.
- Experiment: "Binary design" (M.Hecht *et al*, 1995)

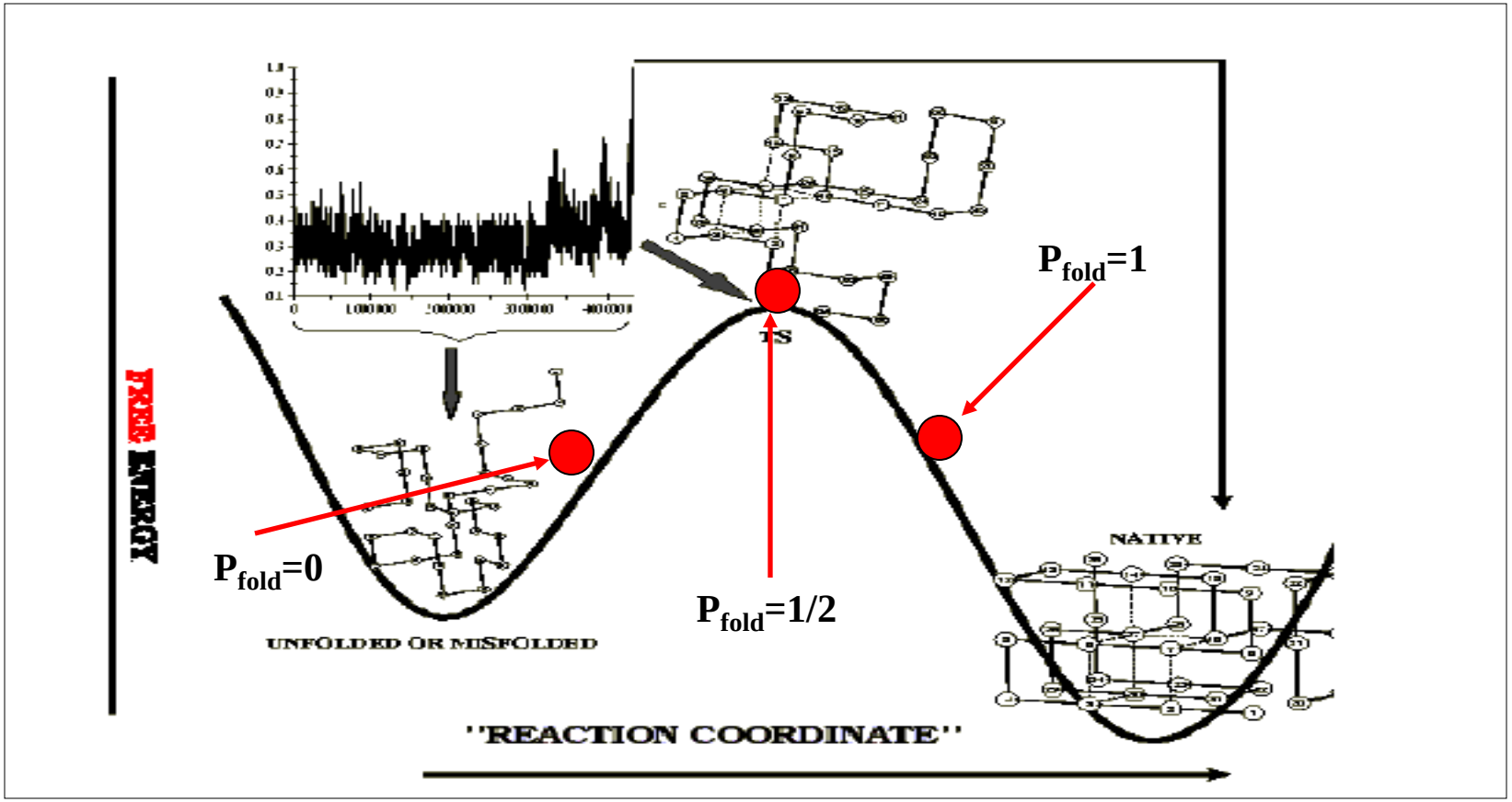
# Cooperativity of ‘’designed’’ (evolutionary selected) sequences



# Q: Why Does Large Gap Solve Kinetic Protein Folding?

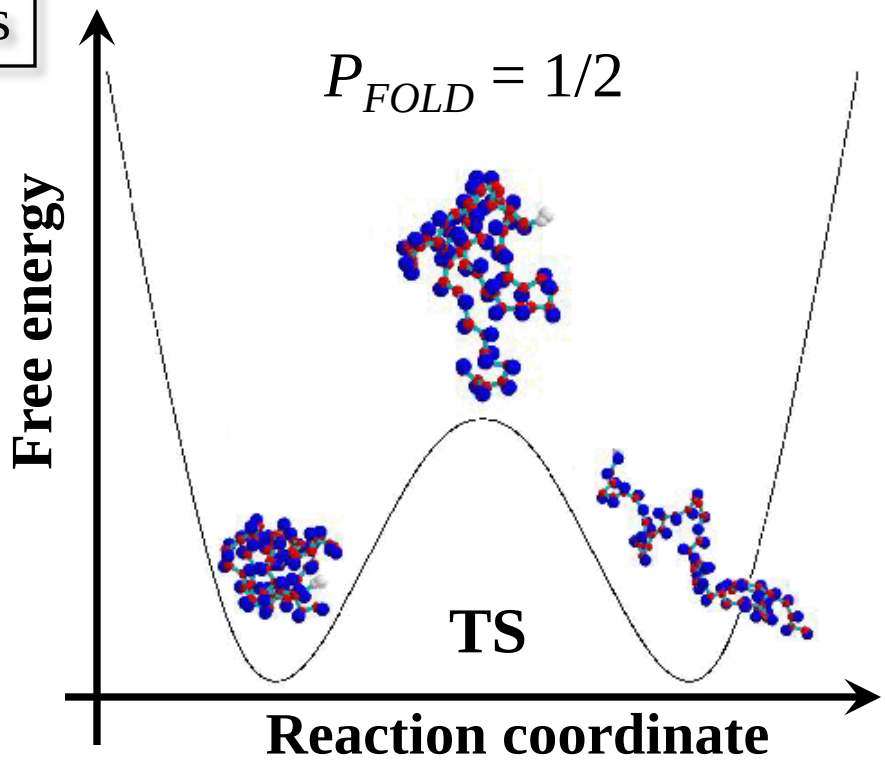
A: It Provides **Nucleation** (akin to first order phase transition kinetics)

# Where is folding Transition State on a trajectory and in microscopic picture?



# HOW DO PROTEINS FOLD?-kinetics

**Nucleation scenario – formation of obligatory contacts between amino acids in the *transition state*.**



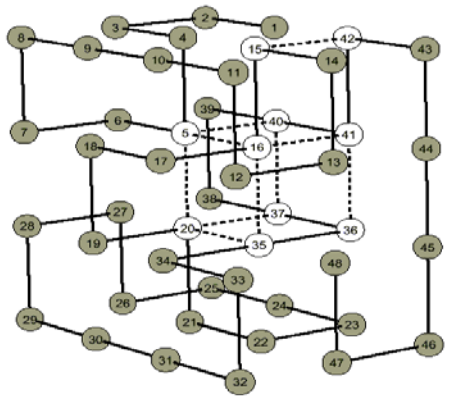

GOAL:

- Test nucleation scenario
- Identify transition state ensemble


Shakhnovich et al., *Nature* **379**, 96-98 (1996)
Abkevich et al., *Biochemistry* **33**, 10026-10036 (1994)
Dokholyan et al., *J. Mol. Biol.* **296**, 1183--1188 (2000)

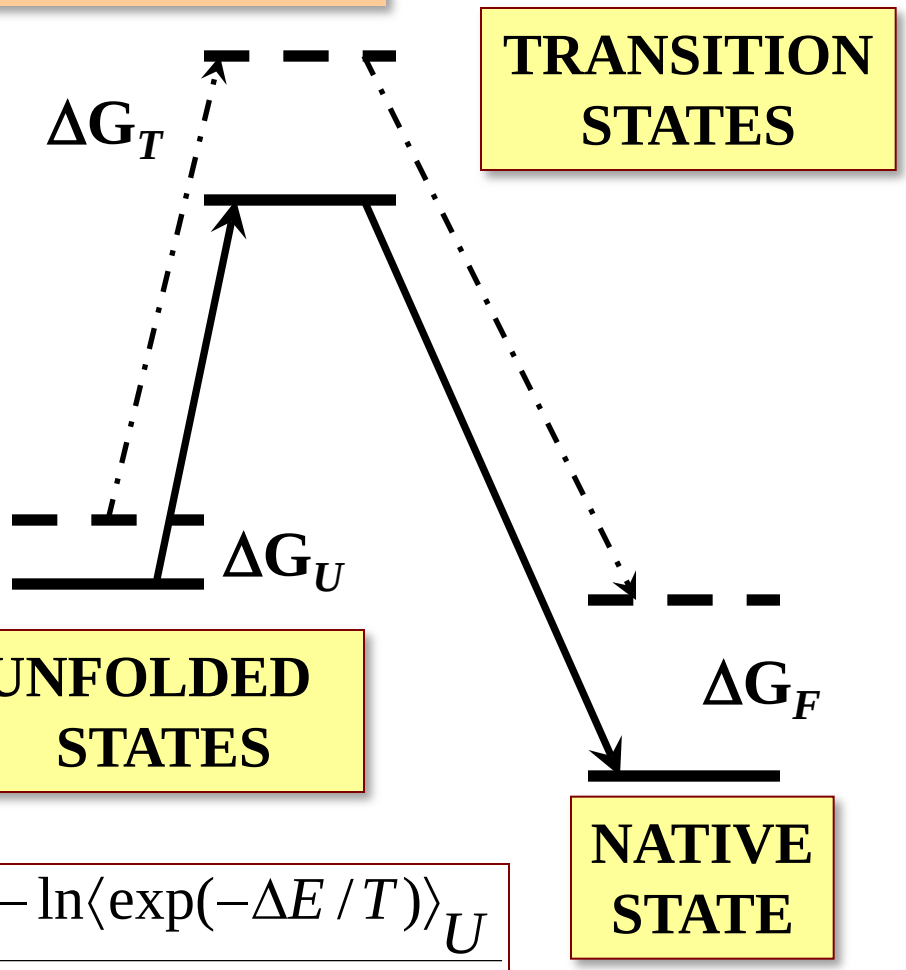
# Nucleation mechanism: Formation of set of Obligatory contacts (Folding Nucleus) that define TSE

(Abkevich et al’ Biochemistry’94)



# PROTEIN ENGINEERING: Φ-VALUE ANALYSIS

**Method:** **Engineer a protein with an altered amino acid at a target position and test to which extent the transition state is affected compare to the native state.**

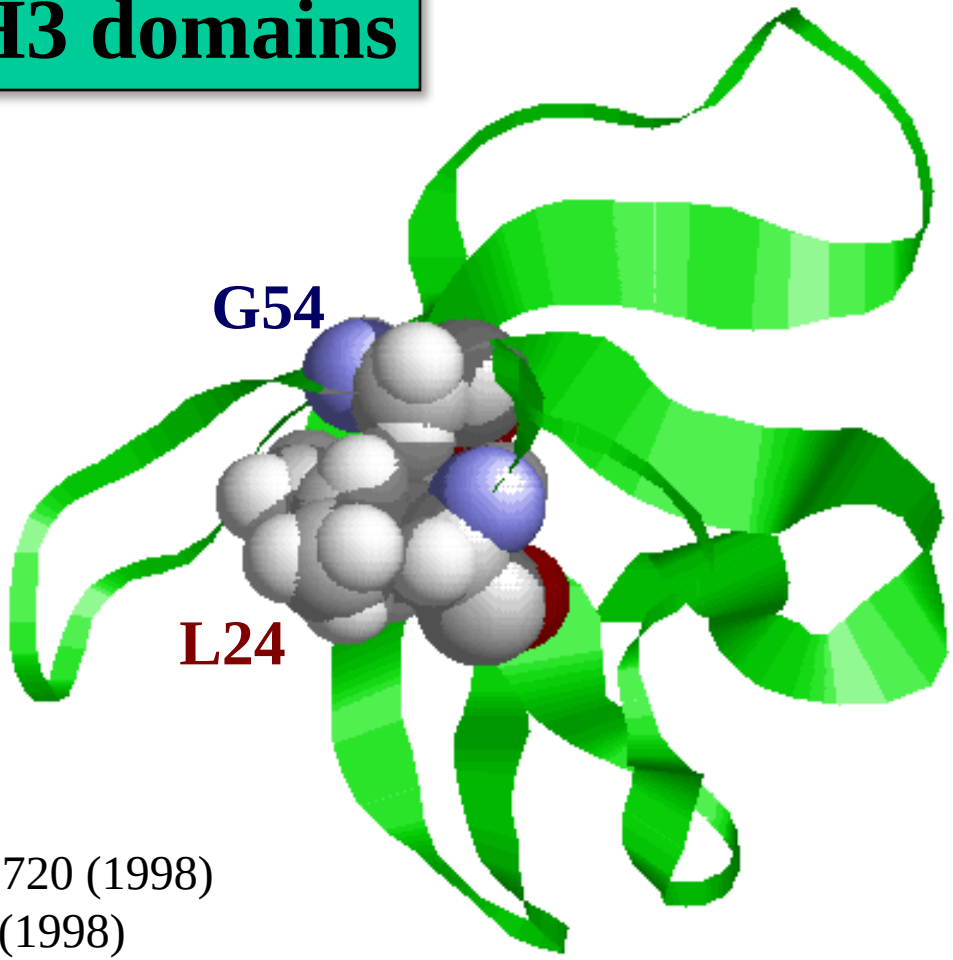



**A residue with**
**Φ=1 is kinetically important**
**Φ=0 is kinetically unimportant**

$$\Phi = \frac{\Delta G_T - \Delta G_U}{\Delta G_F - \Delta G_U} = \frac{\ln\langle \exp(-\Delta E / T)\rangle_T - \ln\langle \exp(-\Delta E / T)\rangle_U}{\ln\langle \exp(-\Delta E / T)\rangle_F - \ln\langle \exp(-\Delta E / T)\rangle_U}$$

Fersht, *Curr. Opin. Struct. Biol.* **7**, 3-9 (1997)

# Folding Nucleus in SH3 domains




Grantcharova et al., *Nat. Struct. Biol.* **5**, 714-720 (1998)
Martinez et al., *Nat. Struct. Biol.* **5**, 721-729 (1998)

```
                1       10        20        30        40        50
1ckaA  AEYVRALFDFNGNDEEDLPFKKGDILRIRDKPEEQWWNAEDS.EGKRGMIPVPYVEKY
1shg   KELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVN..DRQGFVPAAYVKKL
1nloC  ..TFVALYDYESRTETDLSFKKGERLQIVNNTEGDWWLAHSLTTGQTGYIPSNYVAPS
```

Ding et al., *Biophys. J. submitted* (2002)
Borreguero et al., *J. Mol. Biol. in press* (2002)

# Evolutionary control of folding rates and stability

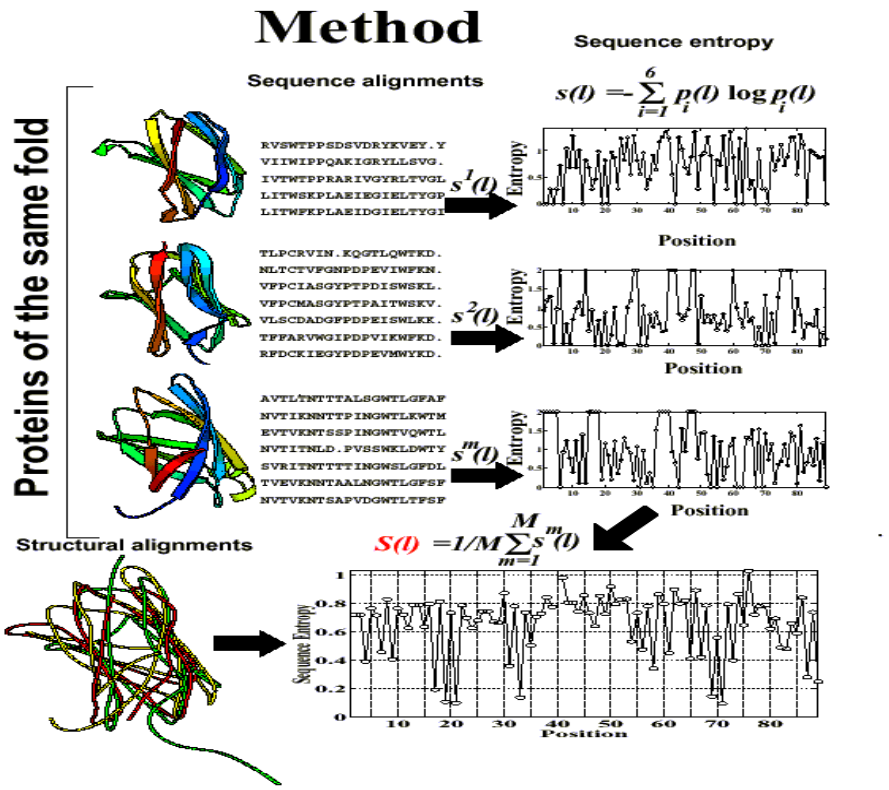
- **The idea: Nucleus residues may determine folding rate**

- **Therefore if evolution cared about folding it could have exerted extra pressure on nucleus residues.**

- **How to select ‘’ folding pressure’’ on the background of many other, strong pressures (function, stability…)?**

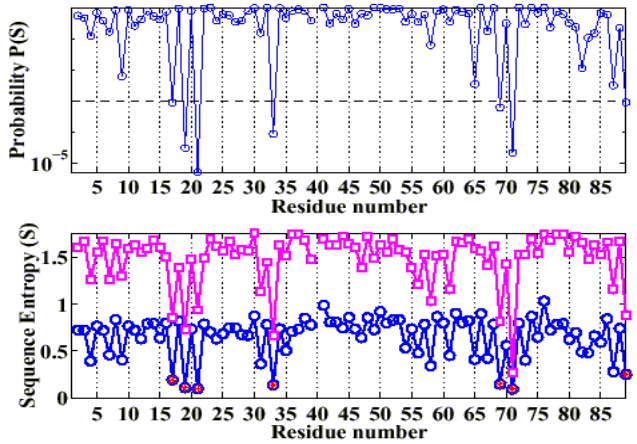
- **Look at proteins featuring similar structure but diverse functions!**

# Conservatism-of-Conservatism: The Idea



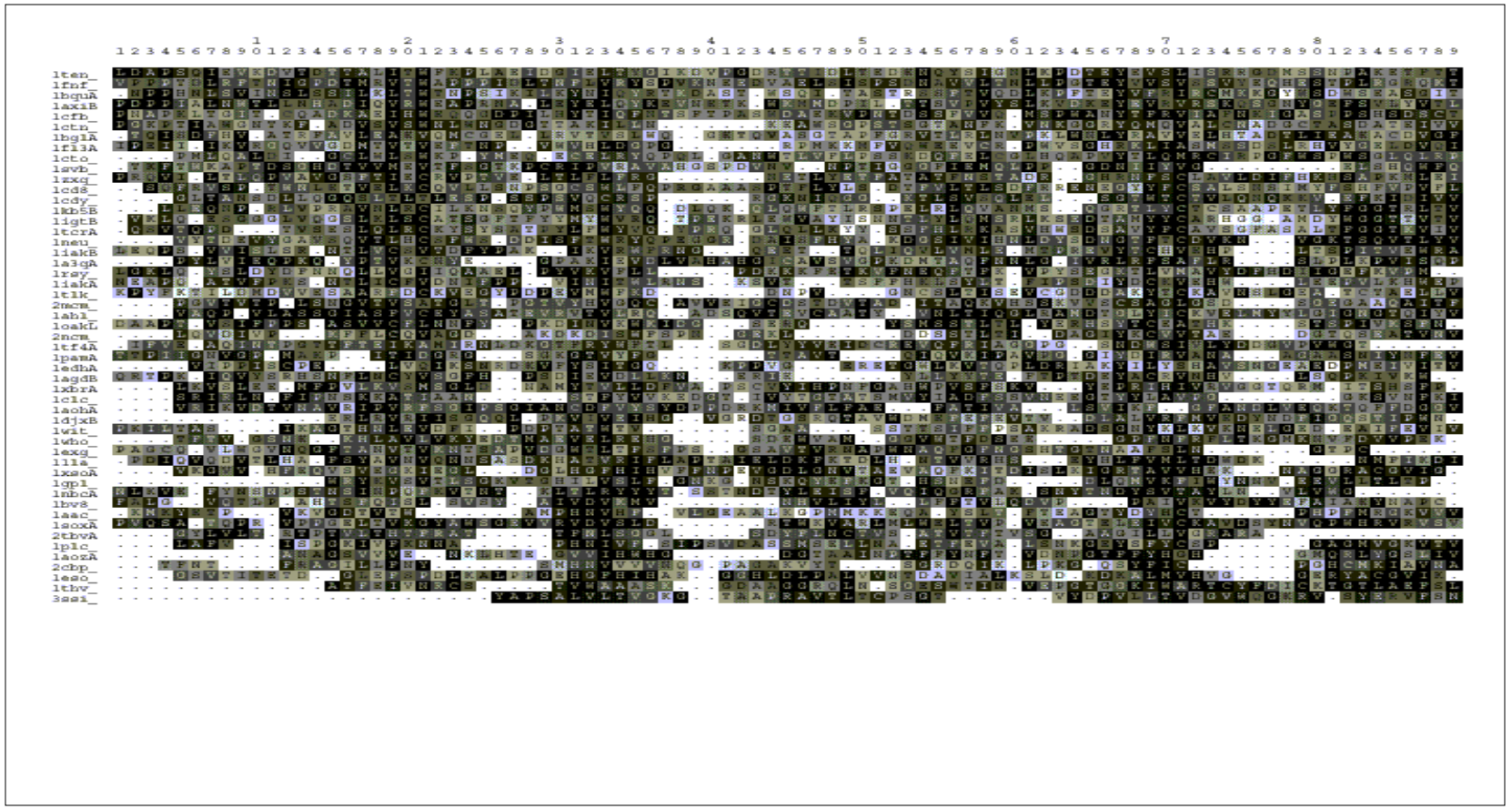
# CoC Analysis correctly predicts folding nucleus in Ig-fold proteins

**Immunoglobulin-like fold**
**Tenascin (fibronectin type III repeat)**

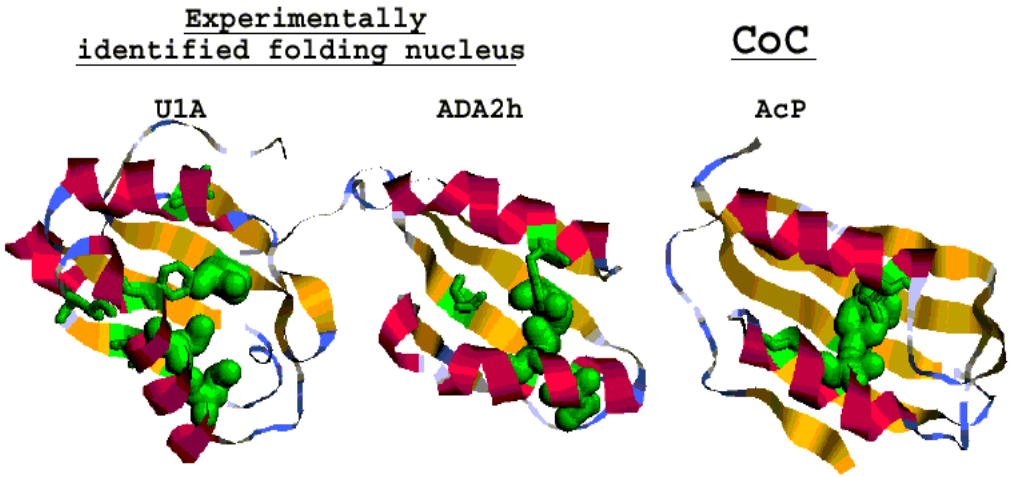




**Experiment:J.Clarke and coauthors (2001)**

# Example: Ig-Fold proteins



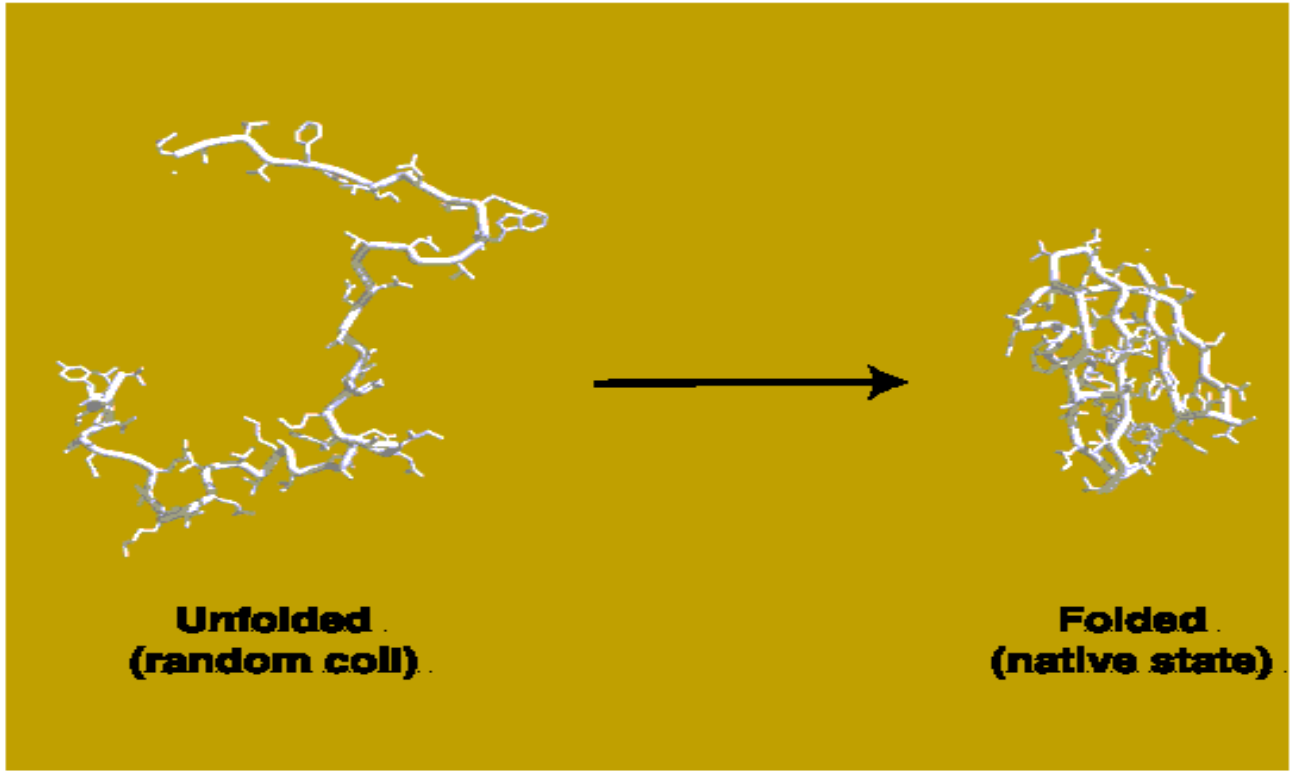
# Understanding diversity of protein folding rates….




Otzen DE, Kristensen O, Proctor M, Oliveberg M
*Biochemistry 1999 38(20):6499-511*

Villegas V, Martinez JC, Aviles FX, Serrano L
*J Mol Biol 1998 283(5):1027-36*

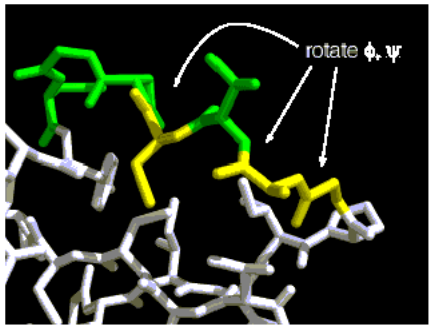
# An all-atom Monte-Carlo Folding Simulation



## Move set

1 MC step = 1 backbone move + 10 sidechain moves

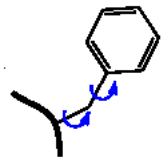
### Backbone




(1) Select a **window of six** consecutive residues.
(2) Randomly select 1, 2, or 3 residues in this window.
(3) Rotate the $\phi$ and $\psi$ angles of the **selected residues.**

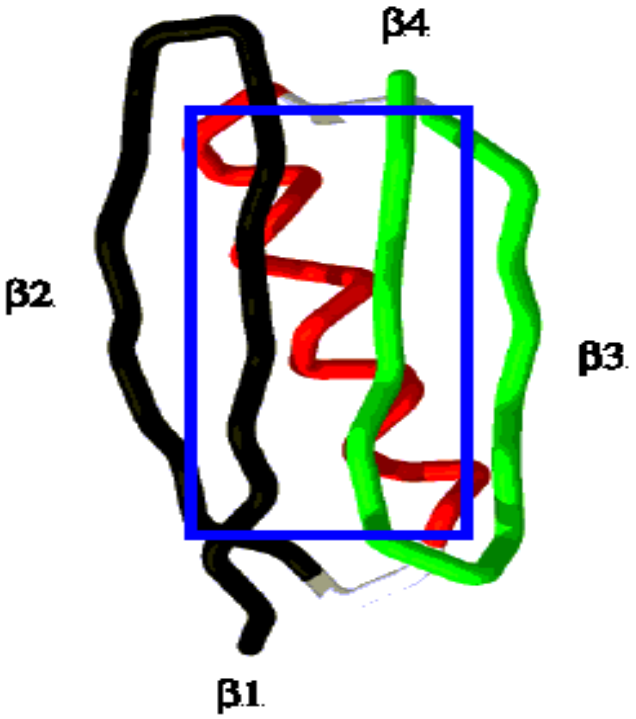
### Sidechain

(1) Randomly select a non-proline residue.
(2) Rotate each of the $\chi$ angles of the selected residue.



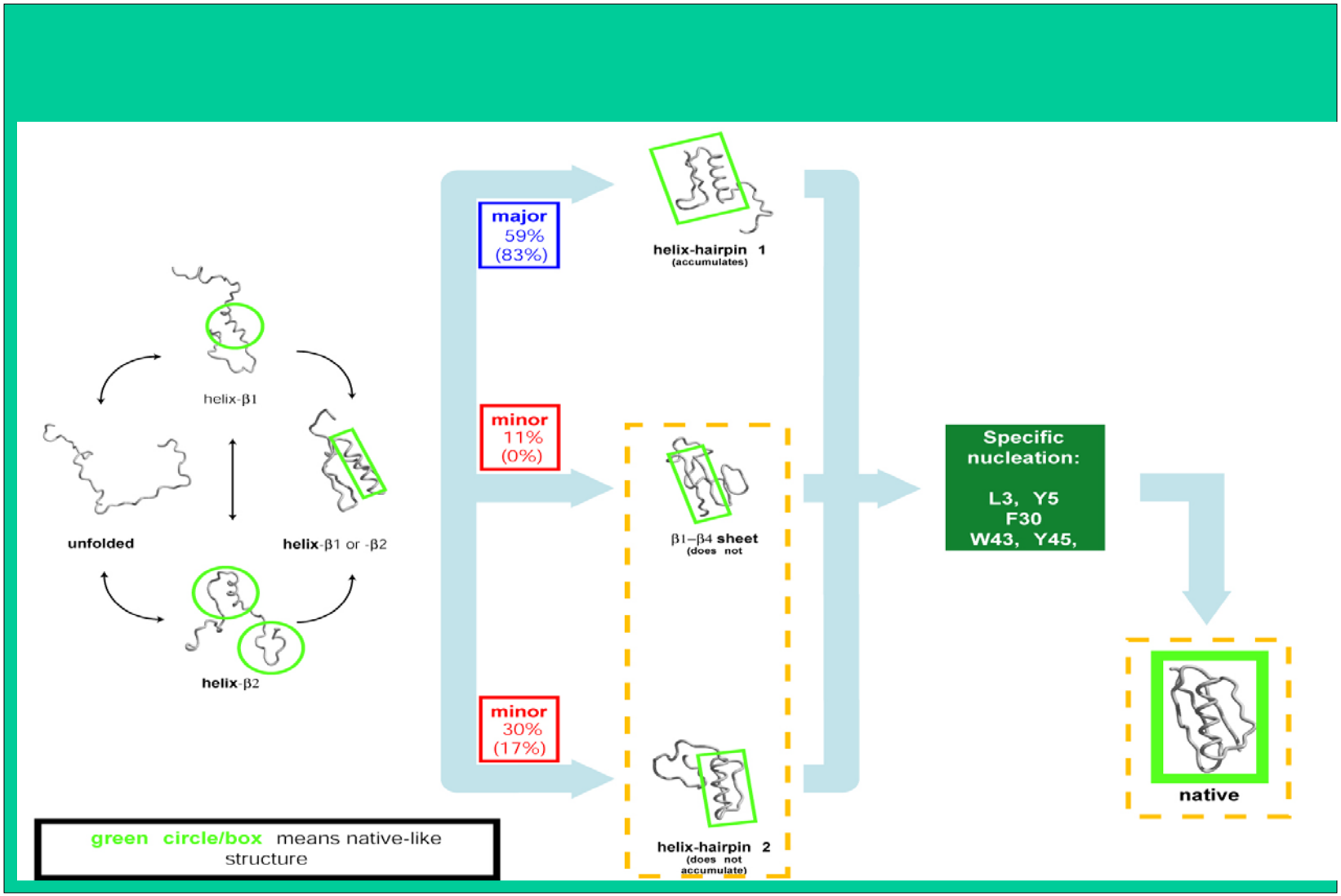
All angle sizes are chosen from a normal distribution with mean zero and standard deviation 2 for backbone and 10 for sidechains.

# Protein G – Folding of a small protein in all-atom detail – Go model



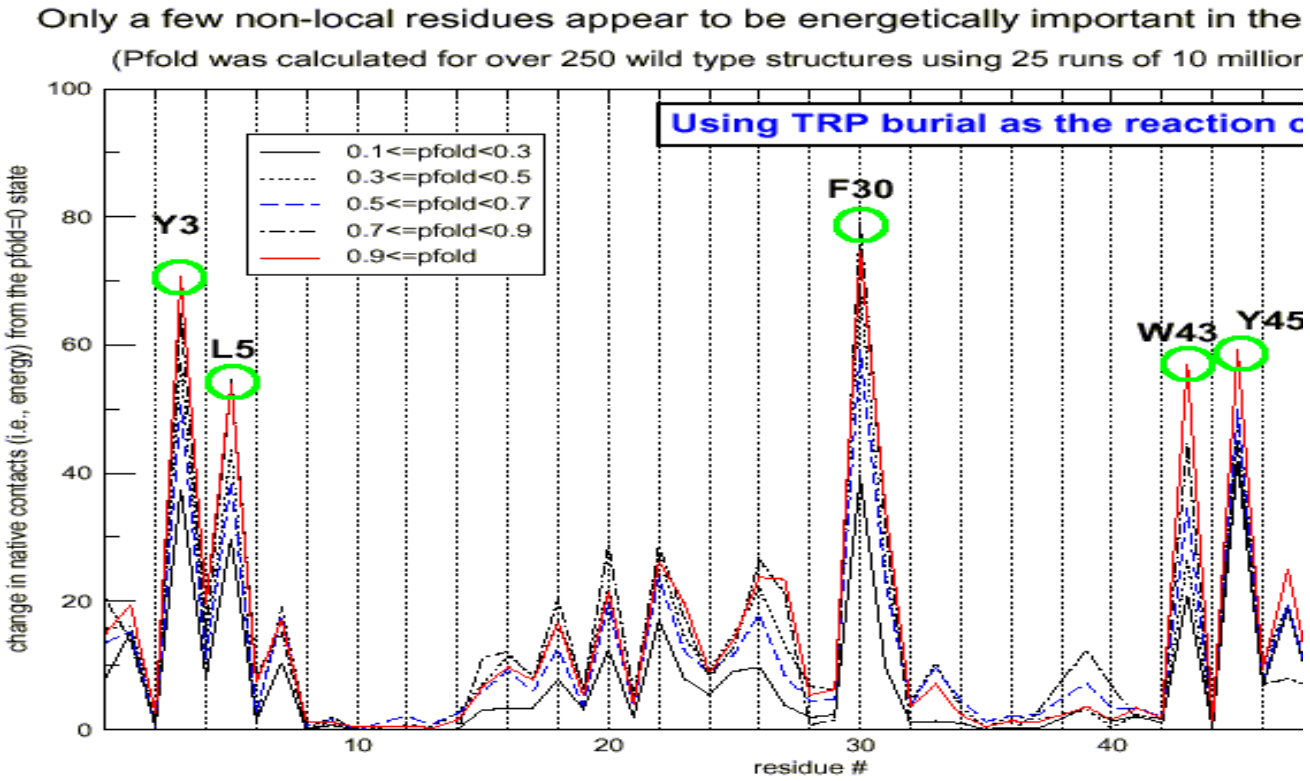

- Black – first beta-hairpin
- Red – alpha helix
- Green – second beta hairpin

# Protein G folding pathways: summary



# What about TSE in protein G?

## Same protocol with $P_{fold}$ identifies states that are committed to fold very fast, ‘’ downhill’ ’, in less than $10^7$ steps



$P_{fold} = 0.5$


green = important in WT
red = important also in mutant

# How to Fold A Protein?

- From Sequence to Structure (i.e. non-Go)
- All Atom
- Low RMSD….

(wishlist….)

**Approach:**

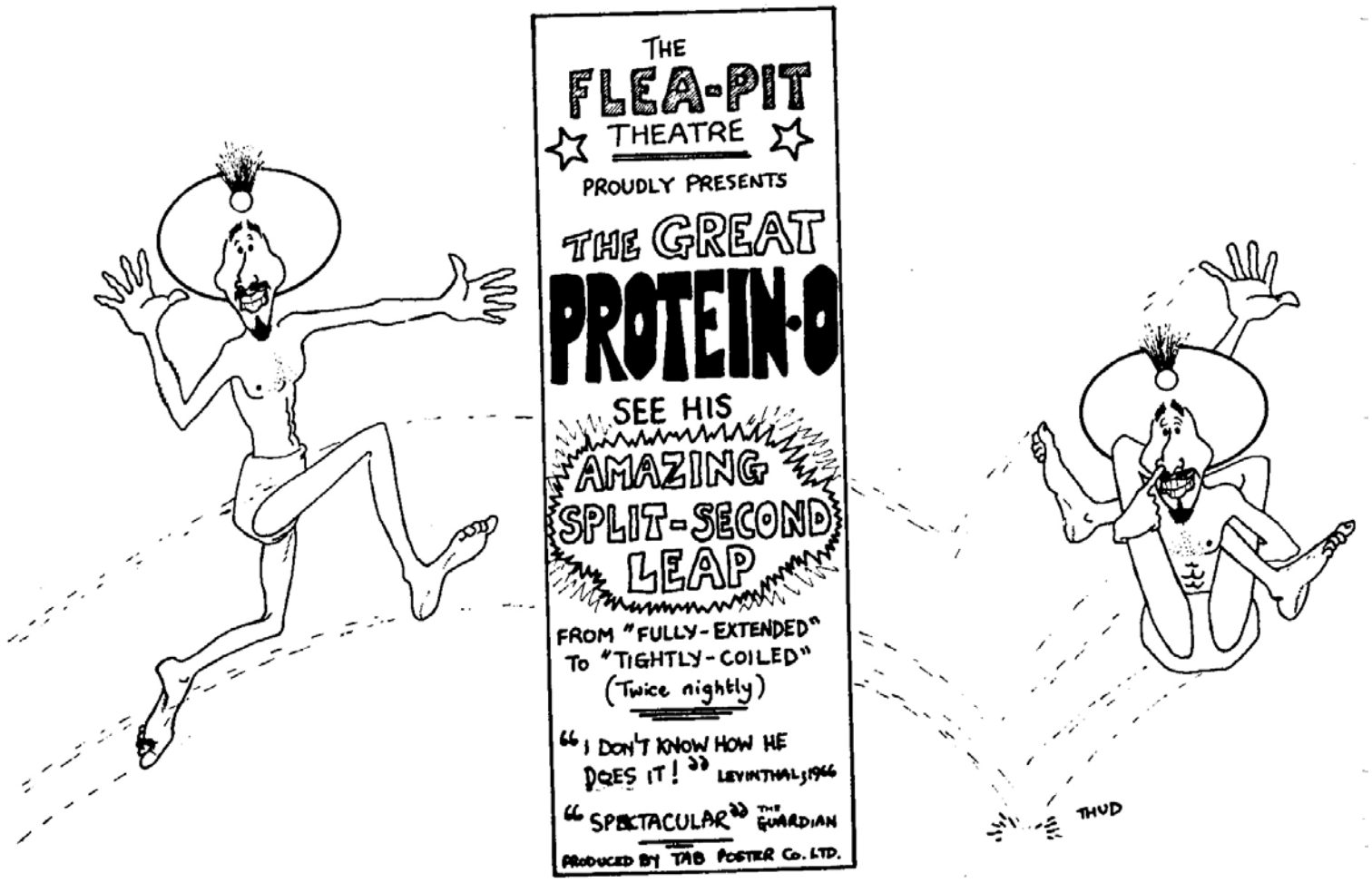
**All-Atom Statistical Potentials (2-body+HB+Local)**

**Kussell,ES PNAS’02, Hubner,Deeds,ES, PNAS’05**

**Dynamically balanced All-Atom MC:**
**global+local BB moves + sidechain moves**

**Shimada, Kussell,ES JMB’01**
**Shimada, ES PNAS’02**
**Hubner, Shimada, ES JMB’04**

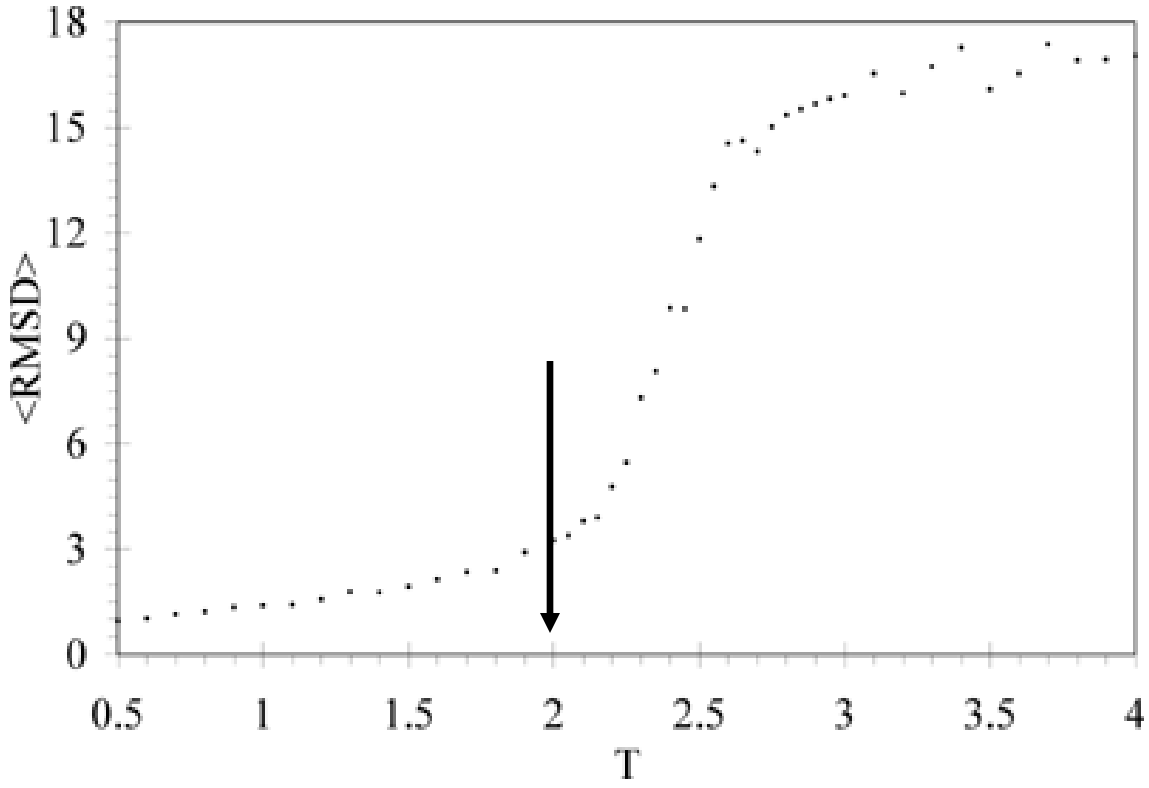
# HOW DO PROTEINS FOLD?
# (all-atom, full detail)



# Methods

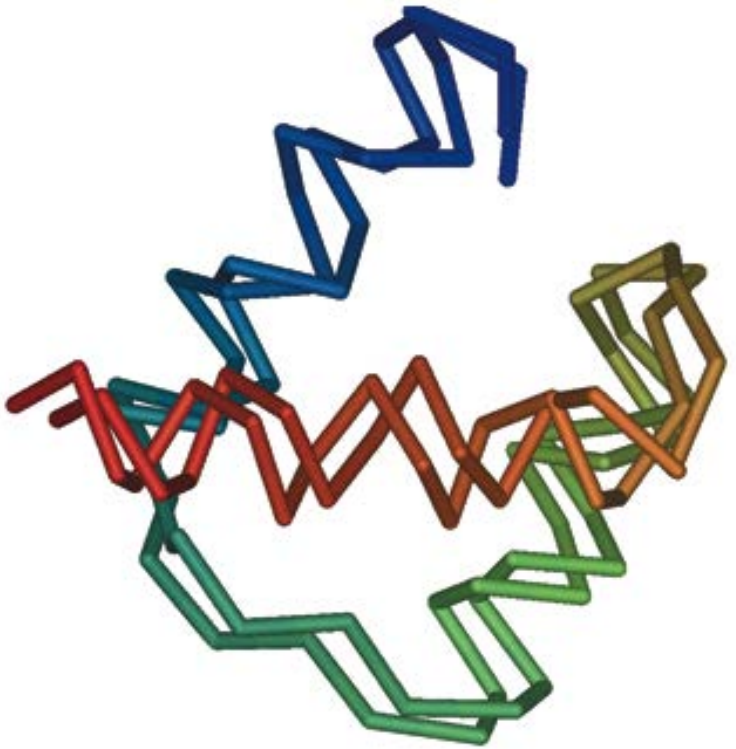
- 4000 folding runs from random coil chain
- At Constant T (NOT Replica Exchange!)
- TSE by $p_{fold}$
- Graph analysis of massive data (like PDUG etc.) instead of projecting on ‘’ reaction coordinates’’
  - $E_{min}$ and folding graphs
  - Clustering (PDUGs) in multiple Order Parameters: a Multidimensional view

# Folding at physiological T~25C



# Identifying the native state

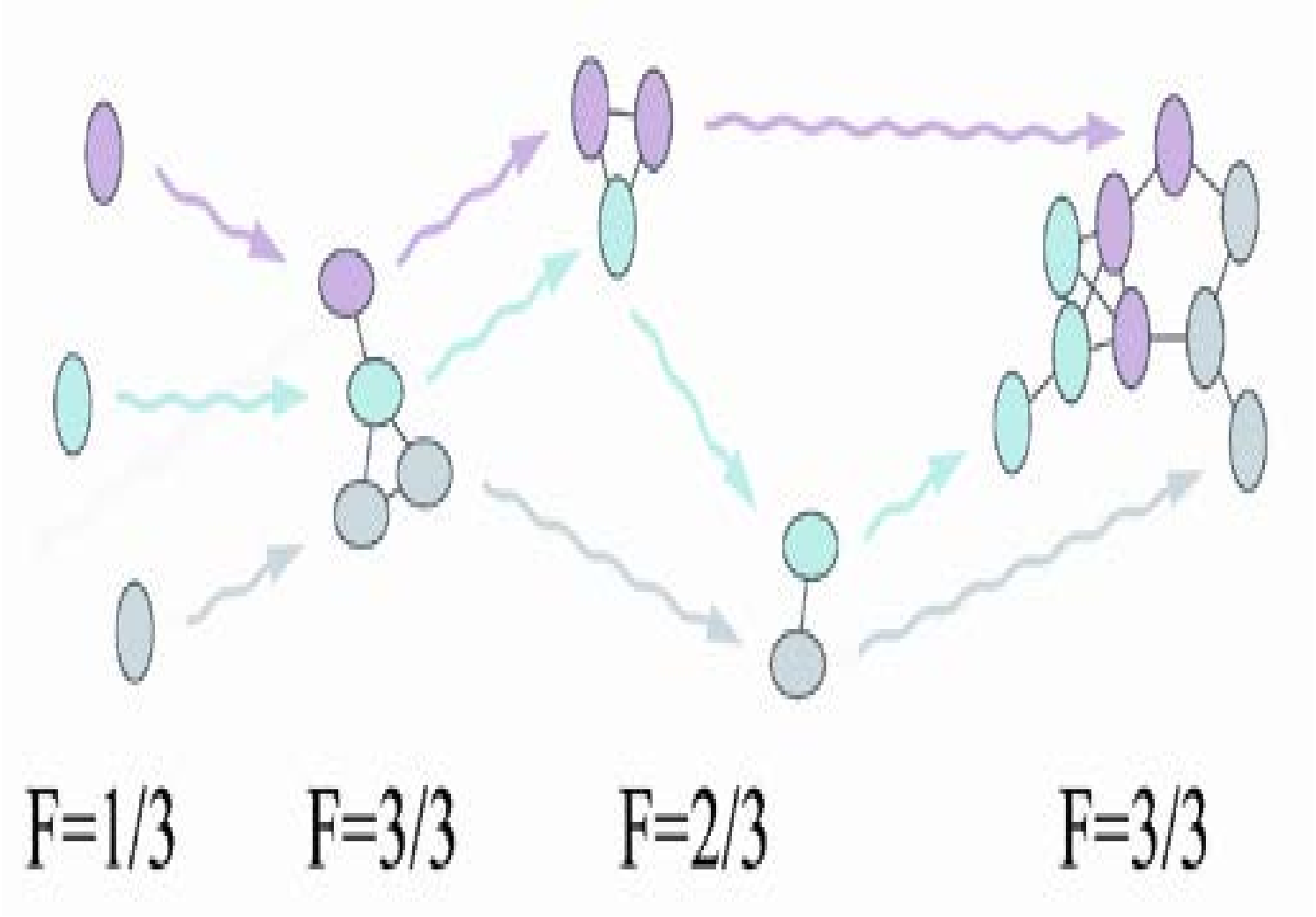
- **Lowest E predition is 2.44Å (best of 4000)**
- **a < 3Å prediction is statistically significant with p-value ~$10^{-7}$**
- **Of 4000, 44 trajectories sampled the ≤ 2Å range, 523 ≤ 3Å range, 1685 ≤ 4Å range, 2700 ≤ 5Å range, and 3331 ≤ 6Å range.**
- **This is consistent with usual exponential distributions of FPT**



# A network ensemble view folding

- Construct a *structural* (not kinetic, cf Caflisch) graph
- Allows combination of multiple trajectories
- Multidumensional view: Cluster Conformations based on various properties: RMSD, Rg, dRMS
- That will allow to fully characterize the folding mechanism, *while any single Order Parameter may be misleading*
- How to introduce Ensemble Kinetics into the graph description: ***idea of flux!***

# Flux:



# Ensemble kinetics

- MFPT = $63 \cdot 10^6$ MC steps, so ~15% of runs would not be expected to fold (~ number of 4000 runs that did not sample < 6Å)
- relaxation from the initial random, swollen-coil conformation to a semi-collapsed state with some hydrogen bonding and non-specific hydrophobic interactions in ~$8 \cdot 10^6$ MC steps
- laser T-jump experiments also reveal two-step folding kinetics at 25° C, with fast ($t_{1/2} \approx$ 1.5microseconds) and slow ($t_{1/2} \approx$ 15microseconds) phases. (Fersht et al)

# What is an intermediate?

- **Accumulative intermediate *vs* intermediary state**
- **A structural graph can identify ‘’parametrically’’ cohesive intermediates**
- **A non-GC cluster that is “visited” by multiple trajectories, i.e. having high flux=1**

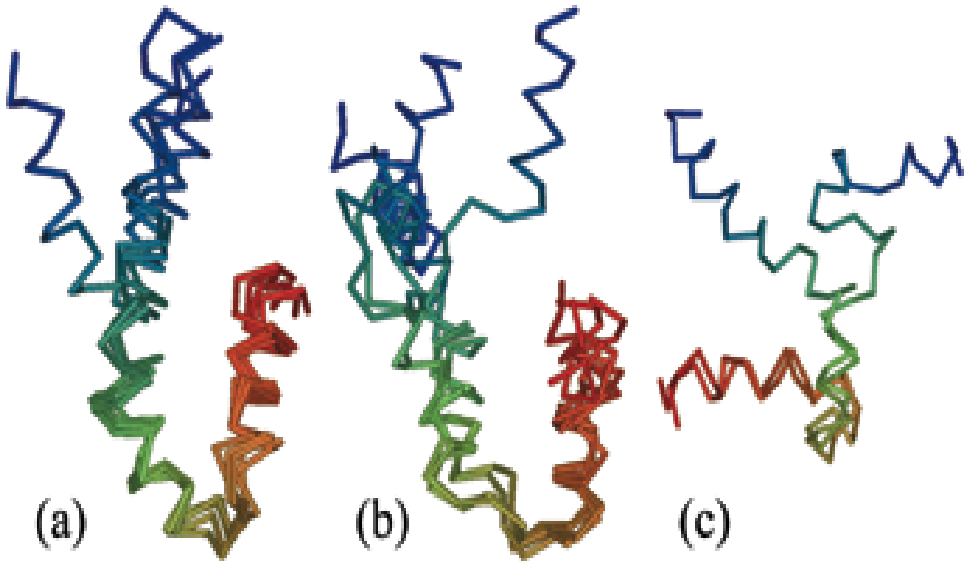
# Is there an intermediate?

- Measure cluster size and flux (fraction of trajectories that “visit” cluster)
- Measure at all cutoffs from fully connected to fully unconnected
- **Flux Analysis reveals one accumulating intermediate (Rg clustering) and one transient one (dRMS clustering)**

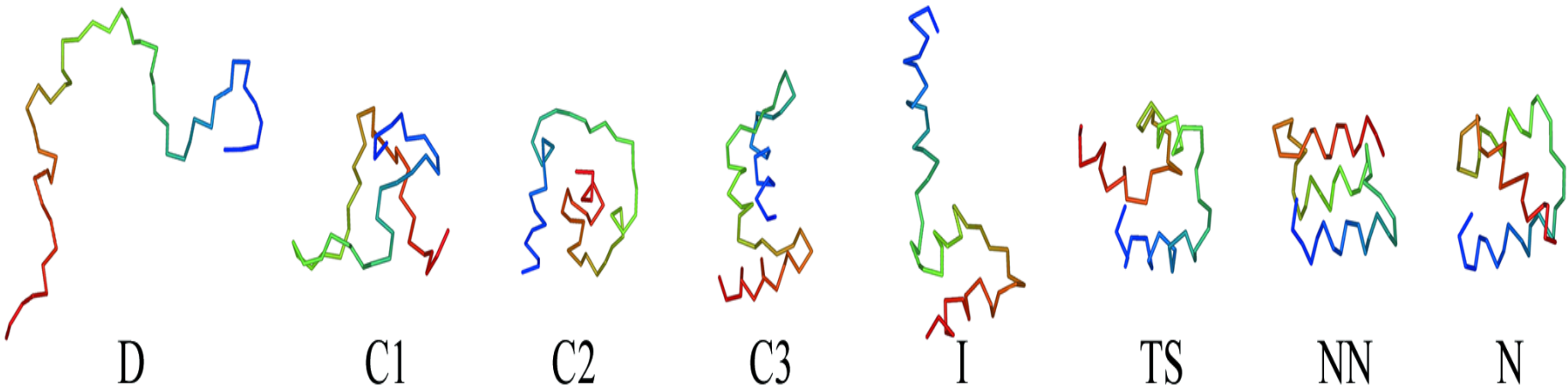
# Atomistically resolved structural intermediate

- **Conformations of the dRMS intermediary cluster perfectly with NMR-derived structures of L16A model of the intermediate**

  **(25 structures 1UZC)**
- **Average RMSD between dRMS cluster and 25 1UZC is 4.6A, some conformations as low as 1.5A**



# Folding Scenario: Summary



# Important new approaches to simulations of folding kinetics

**Markov Chain States:**

Distributed computing generates many relatively short trajectories, the data are then collected and analyzed using markov state models: analogous to Monte-Carlo (Pande et al)

**Simulations using specialized Anton computer (DE Shaw and colleagues)**

Very long trajectories for several proteins (hundred microseconds) were obtained and multiple folding-unfolding events were observed and analyzed (Amber FF04 force-field). The details of folding mechanisms for villin headpiece depend on the force-field. (Science 2009,2011, Biophys Journal 2011)