Protein Folding
*In Vitro**

Biochemistry 412

February 20, 2007

[*Note: includes computational (*in silico*) studies]

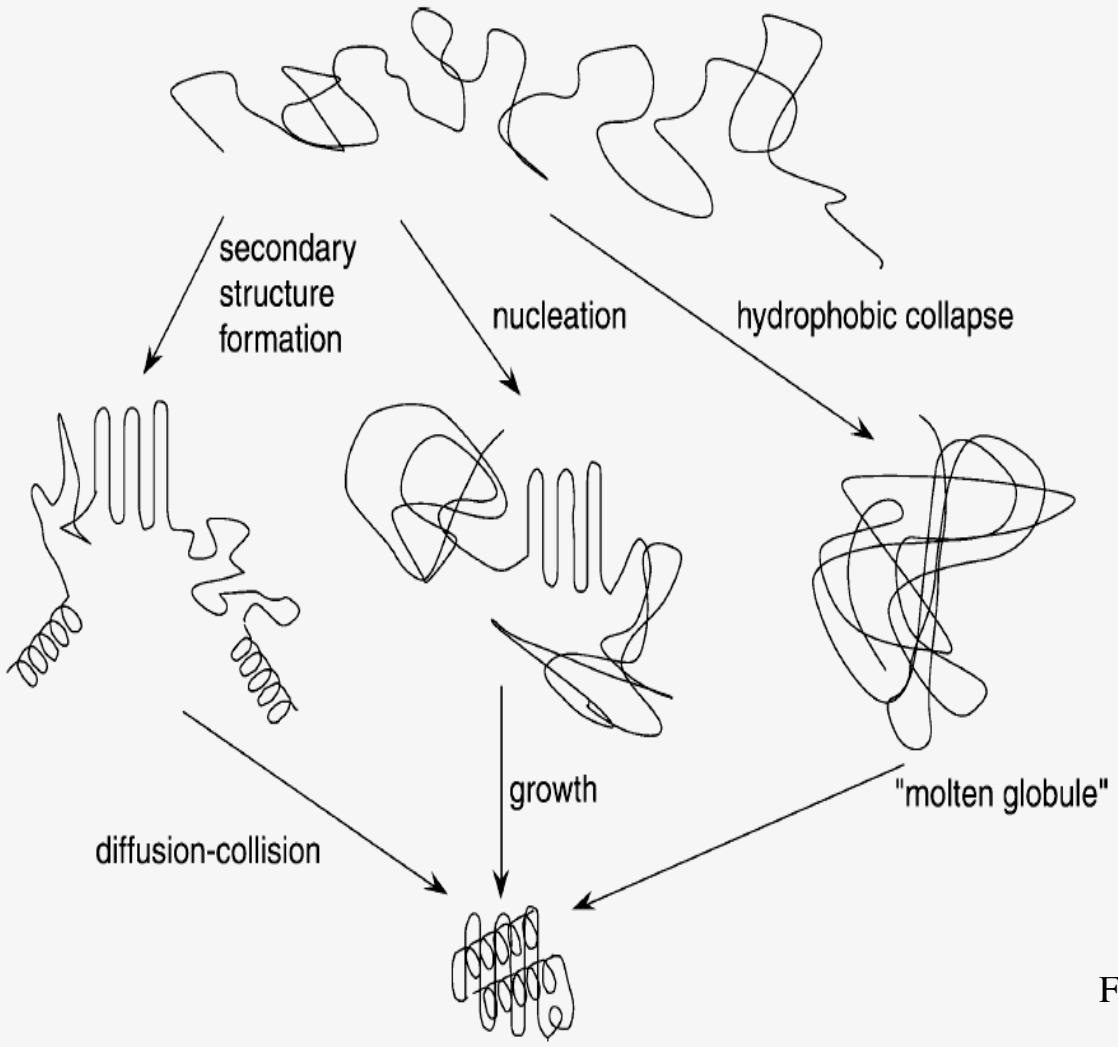


Figure 6. The Three Classical Mechanisms for Protein Folding

At left is pure framework or hierarchical, with secondary structure formed before tertiary. Center is nucleation. At right is tertiary interactions initiating secondary structure formation (modified from Fersht, 1999).

Fersht & Daggett (2002) *Cell* **108**, 573.

Some folding-related facts about proteins:

• Many small, single domain proteins exhibit simple two-state folding behavior

• Most proteins are only marginally stable (5 - 15 kcal/mol) under physiological conditions

• Small proteins generally fold very rapidly, often in less than a second

• During folding, proteins sample only very few of the total number of possible conformations (see Levinthal's Paradox, below)

And...

• It is assumed that a protein's amino acid sequence uniquely determines its native 3D structure

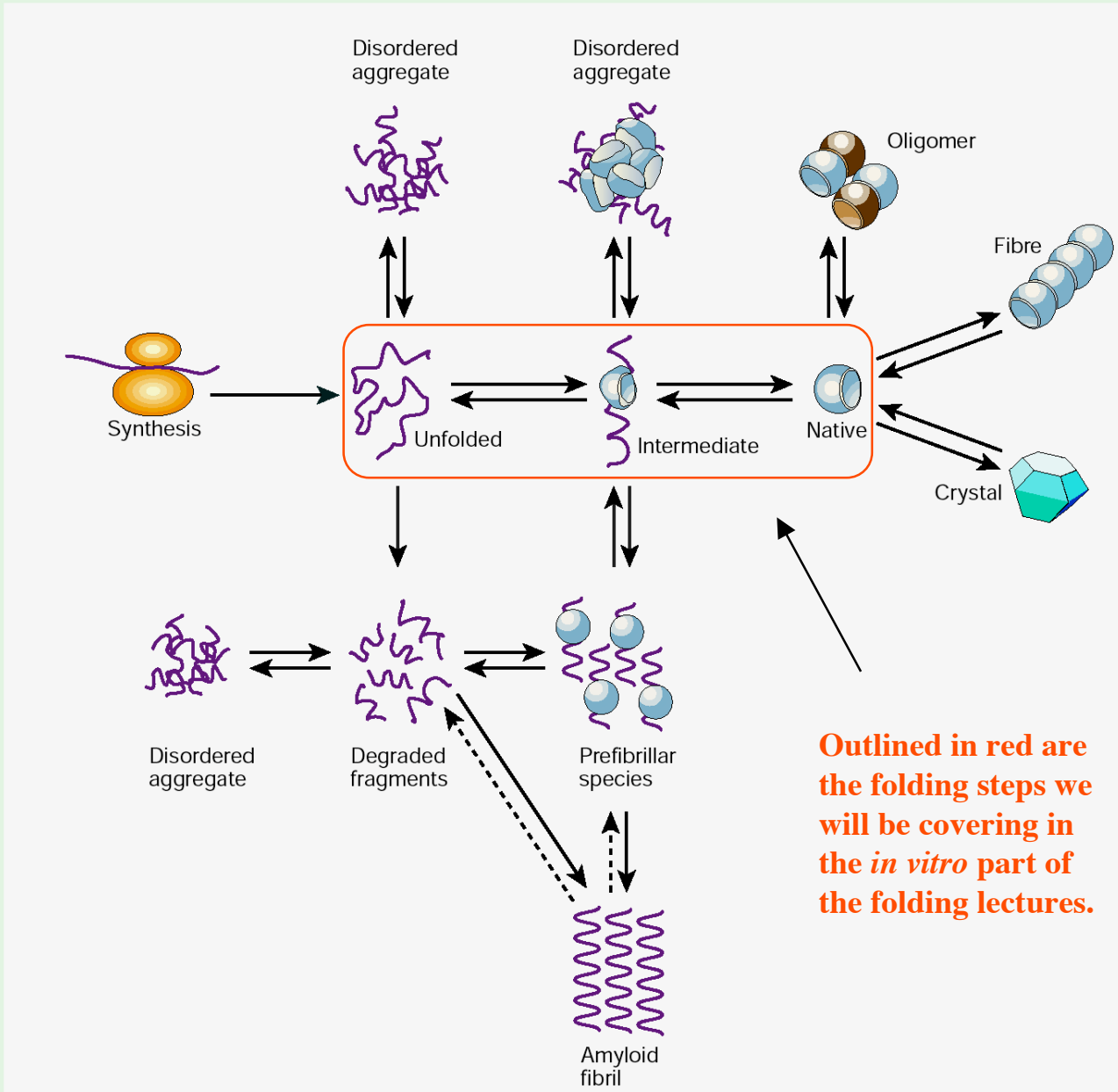


Figure 4 A unified view of some of the types of structure that can be formed by polypeptide chains. An unstructured chain, for example newly synthesized on a ribosome, can fold to a monomeric native structure, often through one or more partly folded intermediates. It can, however, experience other fates such as degradation or aggregation. An amyloid fibril is just one form of aggregate, but it is unique in having a highly organized 'misfolded' structure, as shown in Fig. 3. Other assemblies, including functional oligomers, macromolecular complexes and natural protein fibres, contain natively folded molecules, as do the protein crystals produced *in vitro* for X-ray diffraction studies of their structures. The populations and interconversions of the various states are determined by their relative thermodynamic and kinetic stabilities under any given conditions. In living systems, however, transitions between the different states are highly regulated by the environment and by the presence of molecular chaperones, proteolytic enzymes and other factors. Failure of such regulatory mechanisms is likely to be a major factor in the onset and development of misfolding diseases. Adapted from ref. 54.

Dobson (2003) *Nature* **426**, 884.

Two-State Behavior
Energetic and Kinetic Formalisms

Let U signify the unfolded state and N signify the native state:

$$U \underset{}{\overset{K_{eq}}{\rightleftharpoons}} N$$

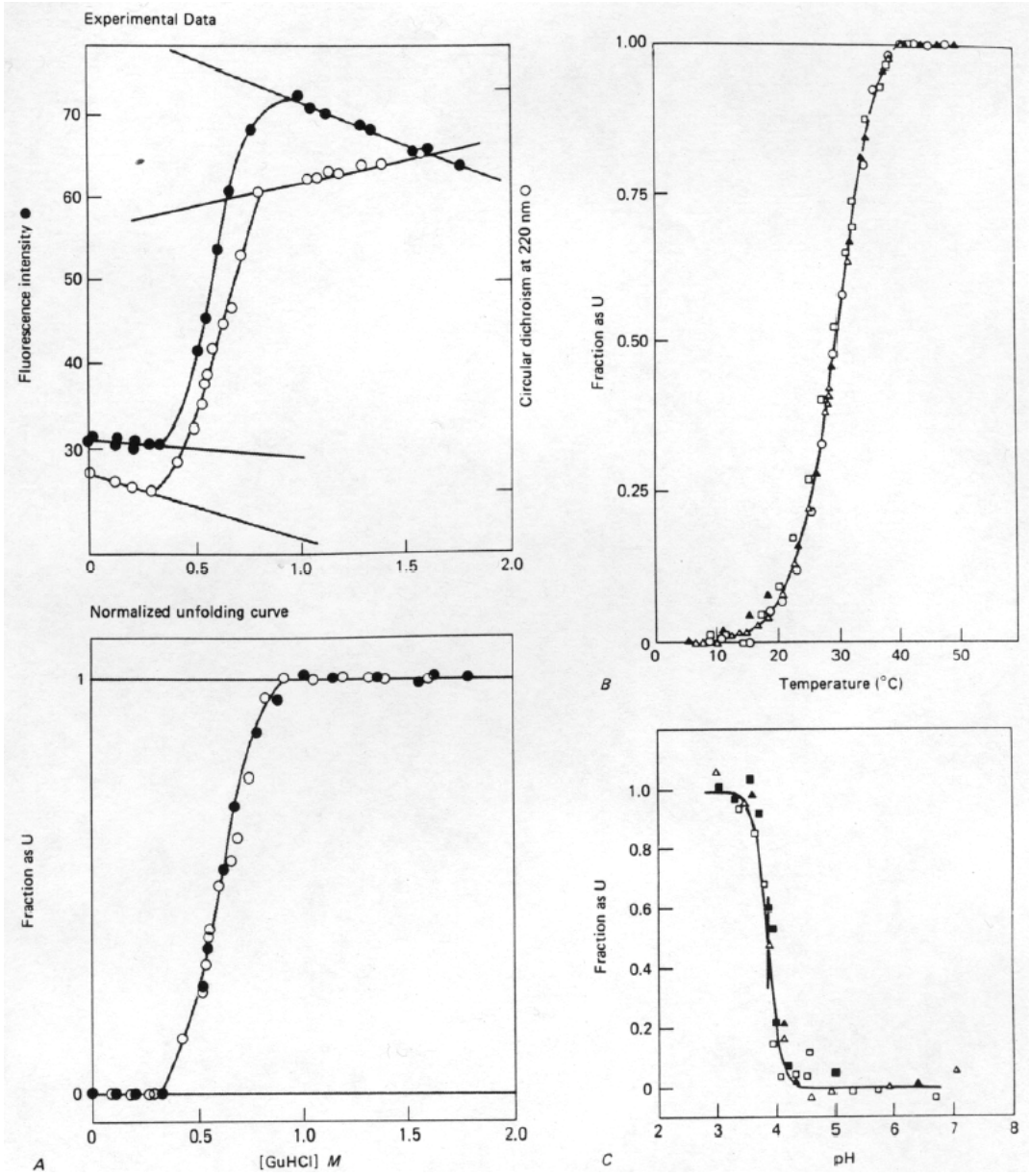
At equilibrium $k_{fold}[U] = k_{unfold}[N]$

So $K_{eq} = [N]/[U] = k_{fold}/k_{unfold}$

And likewise, the stabilization free energy can be expressed as

$$\Delta G^\circ = G_N^\circ - G_U^\circ = -RT\ln K_{eq}$$

Experimental (equilibrium) unfolding of proteins



Creighton
Proteins
W. H. Freeman
1984, p. 288

Note that K_{eq} is a function of the denaturant concentration, since denaturants *by definition* shift the equilibrium toward the unfolded state.

In fact, $\ln K_{eq}$ can be approximated as a linear function of the denaturant concentration, i. e.,

$$\ln K_{eq} = \ln K_{eq}^{H2O} - c[\text{denaturant}]$$

Where *c* is a constant for a given protein and set of conditions.

The reason this is important is that linear plots enable you to accurately measure stabilization free energy differences between two different proteins (e. g., between a wild type protein and its mutant).

The Protein Folding Problem: Levinthal's Paradox

- Ribonuclease (124 residues) can potentially form about 10^{50} conformations. If it tries a different conformation every 10^{-13} seconds, it would take $10^{50}/10^{13} = 10^{37}$ seconds or ~10^{30} years to try all conformations, yet ribonuclease folds in ~1 minute.

- There must be pathways of folding with sequential, dependent steps (intermediates), instead of a random "sampling" of all possible conformations.

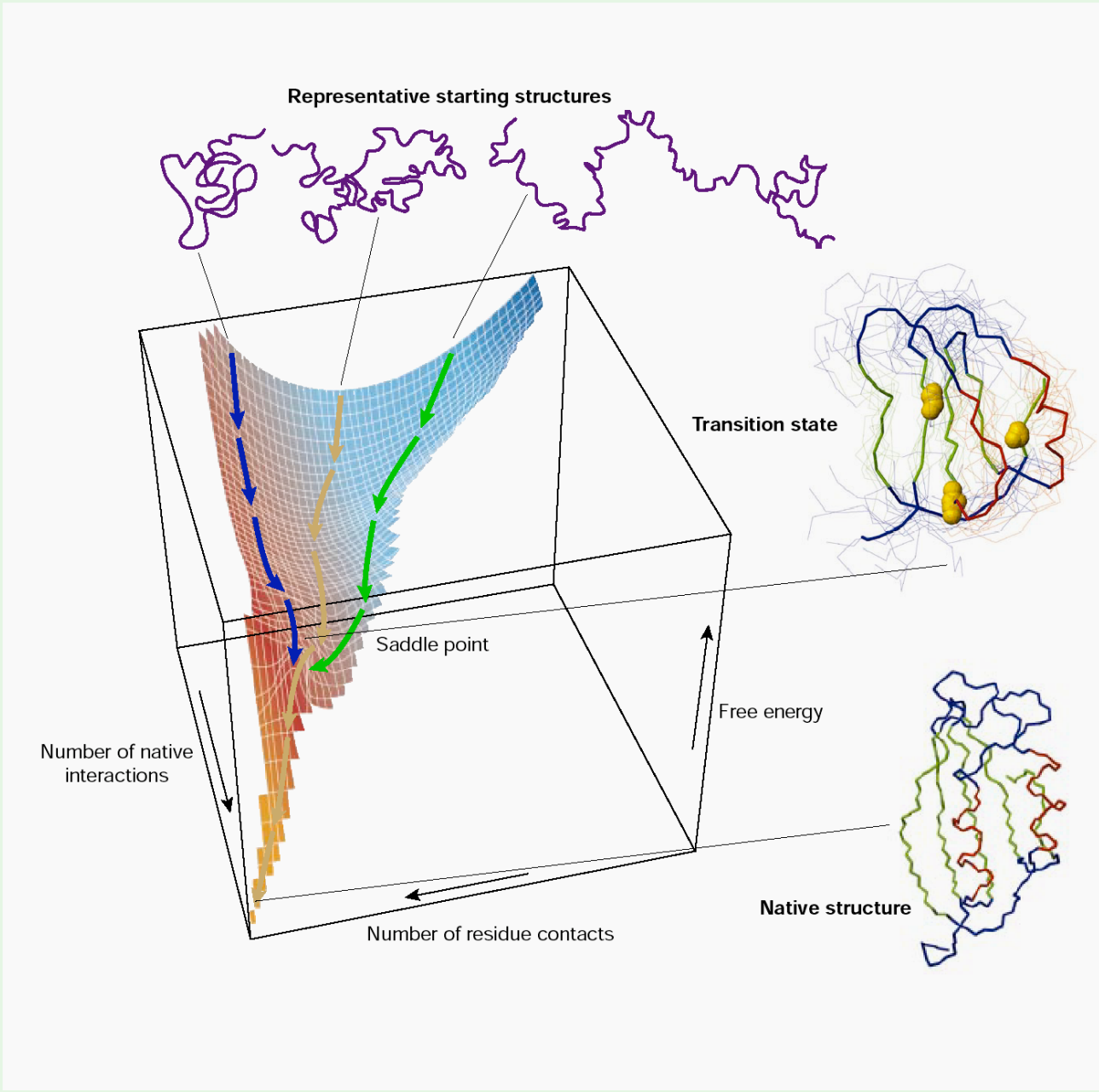
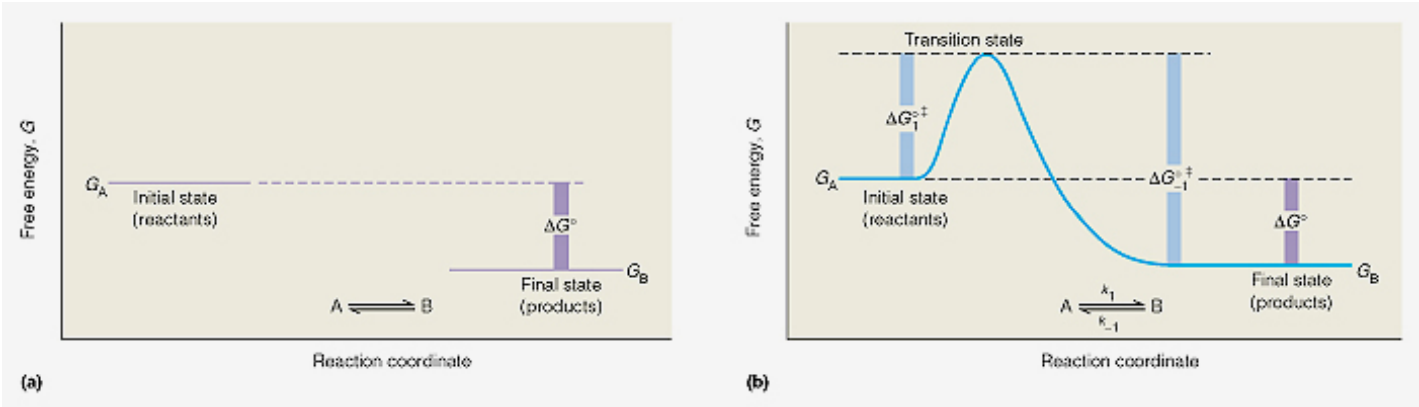


Figure 1 A schematic energy landscape for protein folding. The surface is derived from a computer simulation of the folding of a highly simplified model of a small protein. The surface 'funnels' the multitude of denatured conformations to the unique native structure. The critical region on a simple surface such as this one is the saddle point corresponding to the transition state, the barrier that all molecules must cross if they are to fold to the native state. Superimposed on this schematic surface are ensembles of structures corresponding to different stages of the folding process. The transition state ensemble was calculated by using computer simulations constrained by experimental data from mutational studies of acylphosphatase[18]. The yellow spheres in this ensemble represent the three 'key residues' in the structure; when these residues have formed their native-like contacts the overall topology of the native fold is established. The structure of the native state is shown at the bottom of the surface; at the top are indicated schematically some contributors to the distribution of unfolded species that represent the starting point for folding. Also indicated on the surface are highly simplified trajectories for the folding of individual molecules. Adapted from ref. 6.

Dobson (2003) *Nature* **426**, 884.

Generalized Free Energy Diagrams

[for folding, let N (native state) = "B" = "P" and U (unfolded state) = "A" = "S"]

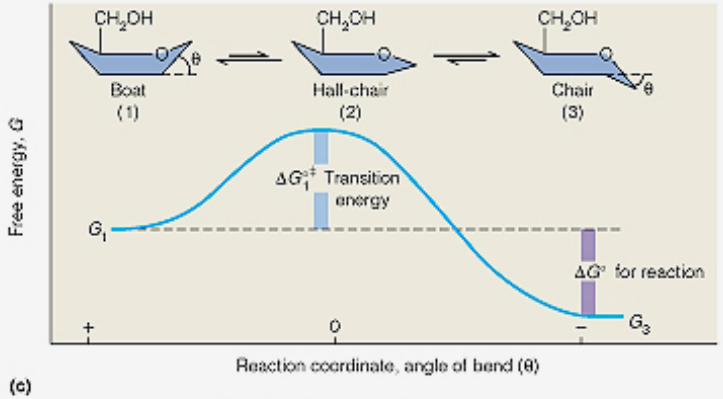


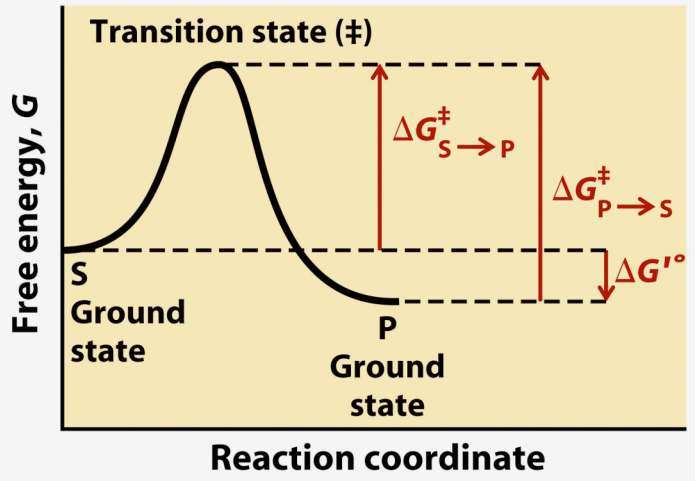
$K = e^{-\Delta G°/RT}$

For A ⇔ A‡

$[A]^{\ddagger}/[A]_o = e^{-\Delta G°\ddagger/RT}$

$[A]^{\ddagger} = [A]_o\, e^{-\Delta G°\ddagger/RT}$





K = equilibrium constant

‡ = transition state

$[A]^{\ddagger}$ = concentration of molecules having the activation energy

$[A]_o$ = total concentration

$-\Delta G°\ddagger$ = standard free energy change of activation (activation energy)

Note that the transition state (TS) energy, $G^{\ddagger}$, can be indirectly measured based on its difference with the unfolded and native state free energies.

Thus,

$$\Delta G_{TS-U} = G^{\ddagger} - G_U^{\circ} \text{ and } \Delta G_{N-TS} = G_N^{\circ} - G^{\ddagger}$$

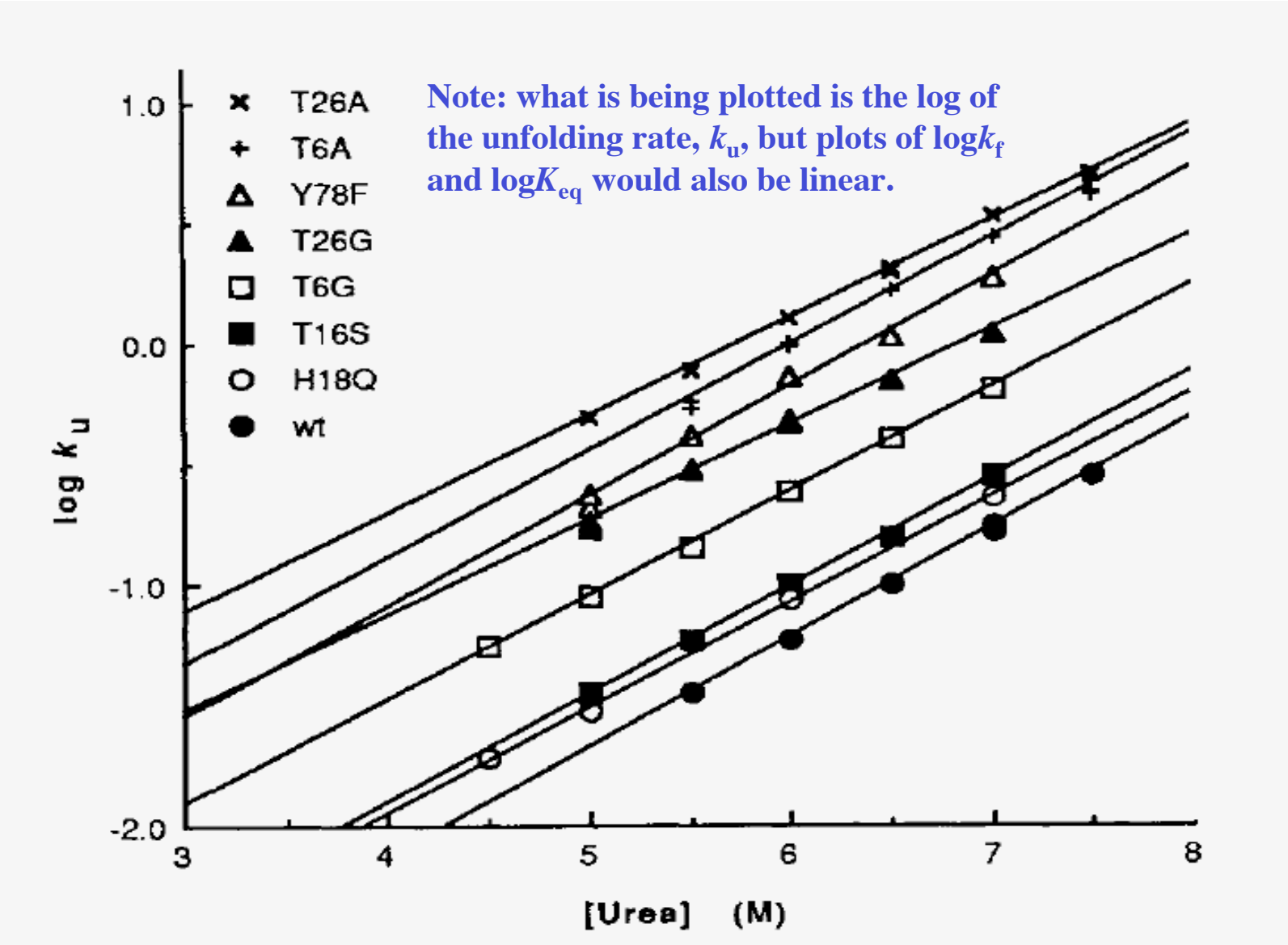
And

$$\Delta G_{TS-U} = -RT\ln k_{fold} \text{ and } \Delta G_{N-TS} = RT\ln k_{unfold}$$

Now….

Protein engineering rears its head!

Denaturation Data for Barnase Mutants



Matouschek et al (1989) *Nature* **340**, 122.

TABLE 1 Changes in activation ($\Delta\Delta G_U^{\ddagger}$) and free energies of unfolding ($\Delta\Delta G_U$) on mutation of barnase

Mutation	Function of position	$\Delta\Delta G_U$ kcal mol^{-1}	$\Delta\Delta G_U^{\ddagger}$ (at 4 M urea) kcal mol^{-1}	$\Delta\Delta G_U^{\ddagger}$ (in H_2O) kcal mol^{-1}	$\Delta\Delta G_U^{\ddagger}/\Delta\Delta G_U$ (at 4 M urea)	$\Delta\Delta G_U^{\ddagger}/\Delta\Delta G_U$ (in H_2O)
Thr → Gly 6	N cap*	1.34	0.86	1.01	0.64	0.76
Thr → Ala 6	N cap*	2.23	1.66	1.76	0.74	0.79
Thr → Gly 26	N cap*	1.58	1.33	1.67	0.84	1.06
Thr → Ala 26	N cap*	2.14	1.91	2.19	0.89	1.02
His → Gln 18	C cap—charge/helix dipole	1.60	0.23	0.36	0.14	0.22
Thr → Ser 16	Hydrophobic on helix surface	1.87	0.28	0.37	0.15	0.20
Tyr → Phe 78	Bridges loop	1.50	1.39	1.41	0.92	0.94
Leu → Ala 14	Hydrophobic core	4.80	1.91	1.86	0.40	0.39
Ile → Val 88	Hydrophobic core	1.49	0.30	0.28	0.20	0.19
Ile → Ala 88	Hydrophobic core	4.46	0.68	0.33	0.15	0.07
Ile → Val 96	Hydrophobic core	0.98	0.48	0.55	0.49	0.56
Ile → Ala 96	Hydrophobic core	3.52	0.74	0.60	0.21	0.17

All data at 25 °C and pH 6.3. Values of $\Delta\Delta G_U$ are taken from refs 11, 12 and unpublished data (L.S. and A.R.F; manuscript in preparation), apart from those for Thr → Ser 16 and Tyr → Phe 78 which were measured as described in ref. 12.* The nomenclature of ref. 21 for residue that occurs at the N terminus of α-helix and can make hydrogen bonds with the backbone NH groups not involved in intrahelical backbone hydrogen bonding[20].

Matouschek et al (1989) *Nature* **340**, 122.

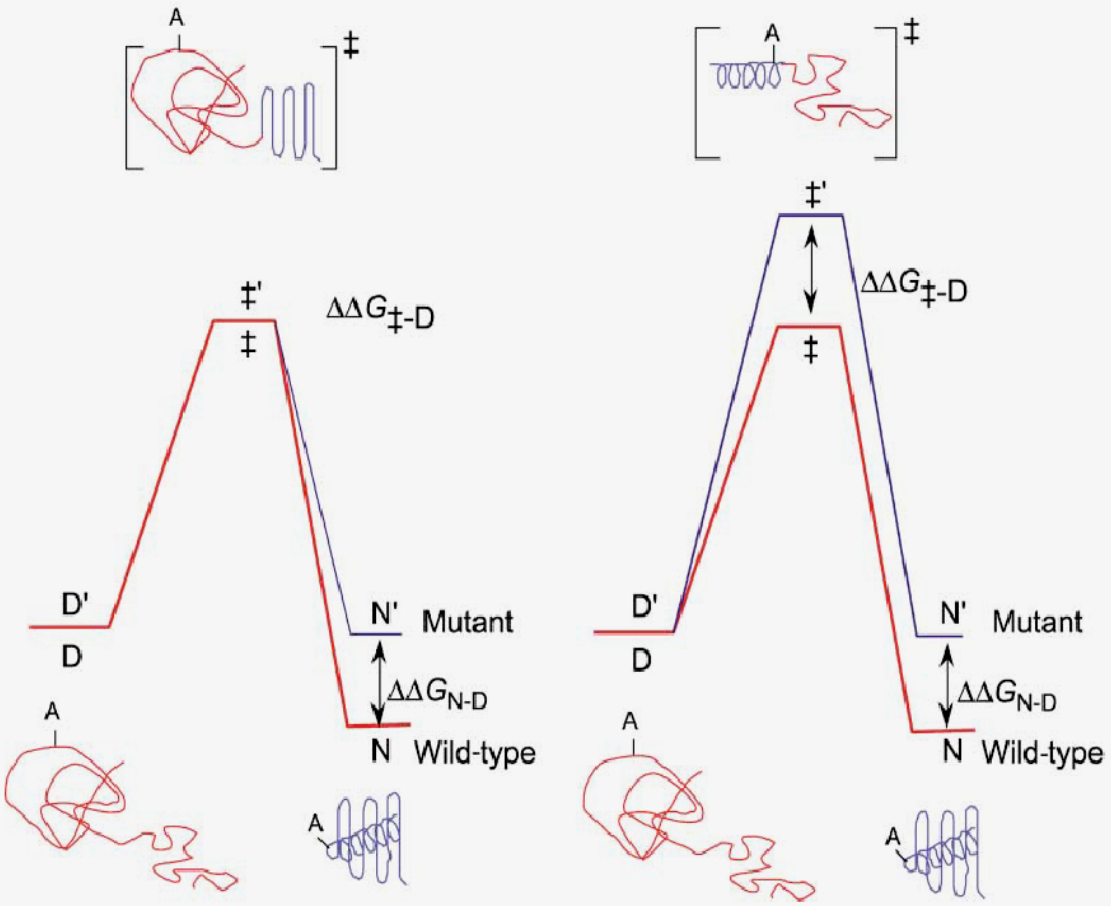


Figure 1. Reaction Profile and ϕ Value Analysis

Schematic profiles are sketched in red for a protein that has an alanine residue (A) in a helix, and in blue for a mutant in which the alanine is mutated to a glycine.

(Left) The transition state (‡), at the top of the energy profile, has the helical region as denatured as in the denatured state D. The energy of the transition state is affected by A to G by the same energy as in D, and so the change in energy of ‡ relative to that of D, $\Delta\Delta G_{\ddagger-D}$, is 0. Thus, $\phi = \Delta\Delta G_{\ddagger-D}/\Delta\Delta G_{N-D} = 0$.

(Right) The opposite case when the helix is fully structured in the transition state has $\Delta\Delta G_{\ddagger-D} = \Delta\Delta G_{N-D}$, and so $\phi = 1$ (modified from Fersht, 1999). The value of $\Delta\Delta G_{\ddagger-D}$ is calculated from the ratio of rate constants for folding of wild-type ($k_{f(wt)}$) and mutant ($k_{f(mut)}$) proteins [$\Delta\Delta G_{\ddagger-D} = RTln(k_{f(wt)}/k_{f(mut)})$]. The value of $\Delta\Delta G_{N-D}$ is calculated by subtracting the free energy of folding of wild-type protein ($\Delta G_{N-D(wt)}$) from that of mutant ($\Delta G_{N-D(mut)}$). The free energies of folding are usually measured from urea-, guanidinium chloride-, or thermal-denaturation curves. ϕ value analysis requires measuring rate and equilibrium constants.

Fersht & Daggett (2002) *Cell* **108**, 573.

Φ-Value Analysis

$$\Phi = (\Delta G_{TS-D} - \Delta G'_{TS-D})/(\Delta G_{N-D} - \Delta G'_{N-D}) = \Delta\Delta G_{TS-D}/\Delta\Delta G_{N-D} \quad \text{(Eqn1)}$$

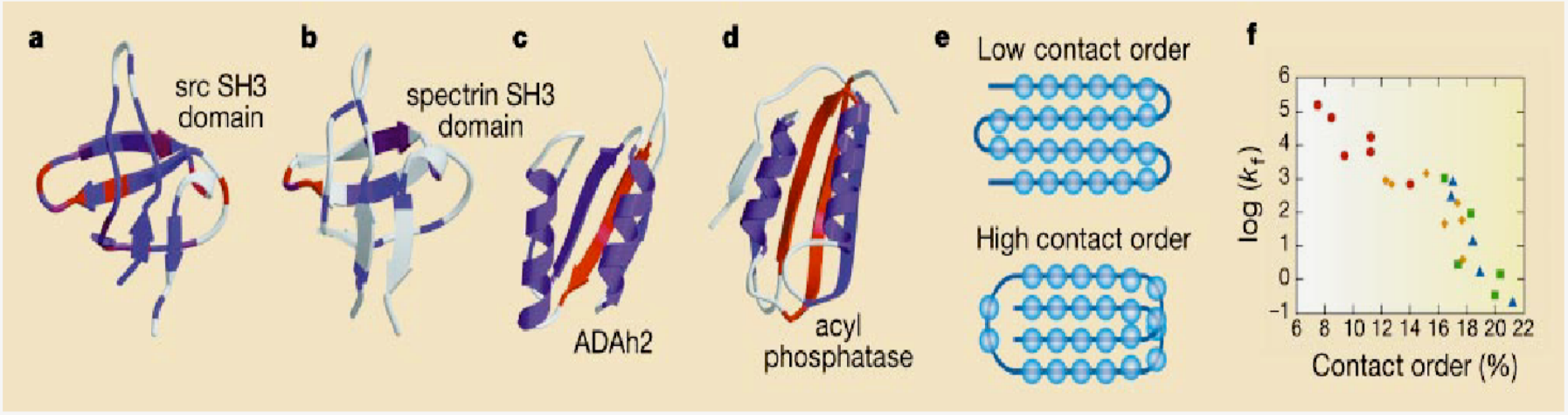
where ΔG_{TS-D} and ΔG_{N-D} are the free energies of the transition state (this could also be an intermediate state) and the native state, respectively, relative to the denatured state for the wild-type protein (Fig. l).

Corresponding terms for the mutant are indicated by a prime. $\Delta\Delta G_{N-D}$ and $\Delta\Delta G_{TS-D}$ are the destabilization energies of the native state and transition states of interest, respectively, caused by the mutation.

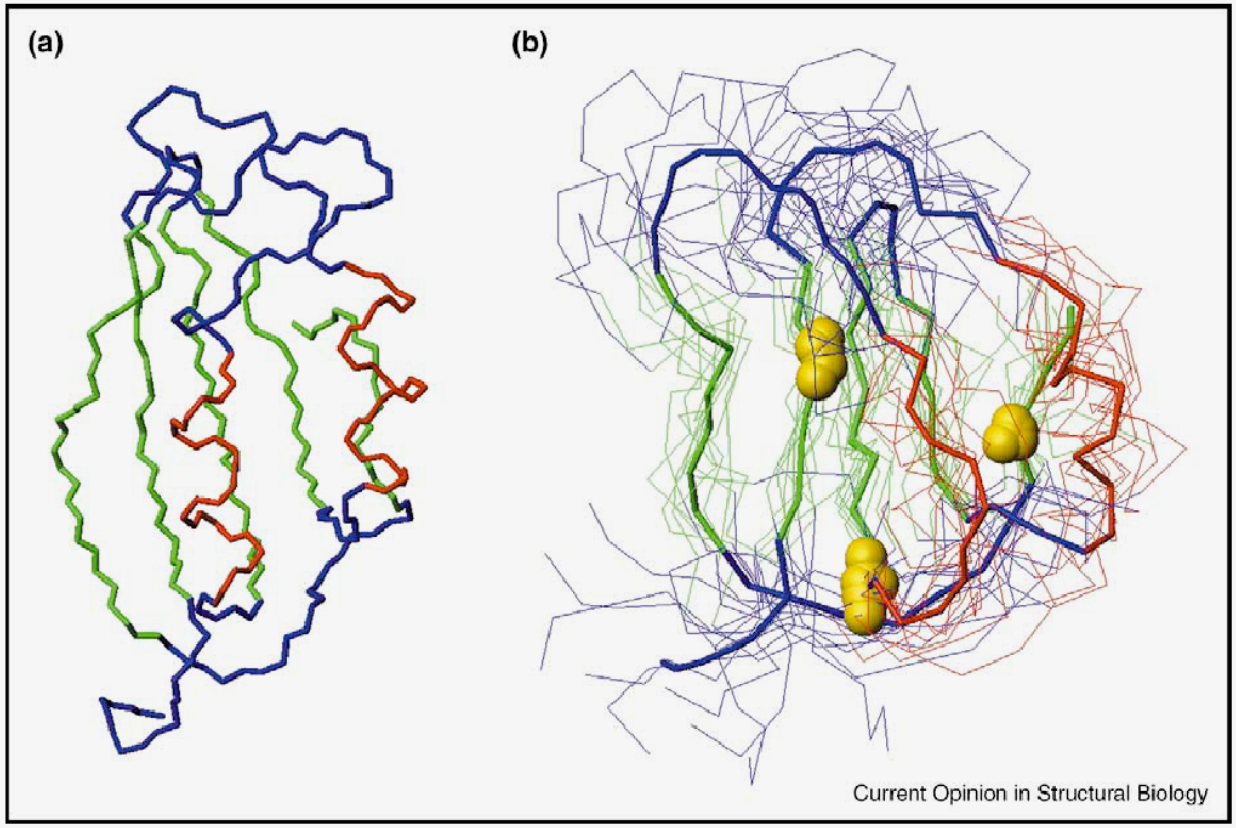
Box 1

Dependence of folding mechanisms on topology

The structures of folding transition states are similar in proteins with similar native structures. The distribution of structure in the transition state ensemble can be probed by mutations at different sites in the chain; mutations in regions that make stabilizing interactions in the transition state ensemble slow the folding rate, whereas mutations in regions that are disordered in the transition state ensemble have little effect[4]. For example, in the structures of the SH3 domains of src[18] (**a**) and spectrin[17] (**b**), and the structurally related proteins Adah2 (ref. 37; **c**) and acyl phosphatase[16] (**d**), the colours code for the effects of mutations on the folding rate. Red, large effect ; magenta, moderate effect; and blue, little effect. In the two SH3 domains, the turn coloured in red at the left of the structures appears to be largely formed, and the beginning and end of the protein largely disrupted, in the transition state ensemble. (To facilitate the comparison in **c** and **d**, the average effect of the mutations in each secondary structure element is shown.) This dependence of folding rate on topology has been quantified by comparing folding rates and the relative contact order of the native structures. The relative contact order is the average separation along the sequence of residues in physical contact in a folded protein, divided by the length of the protein. **e**, A low- and high-contact-order structure for a four-strand sheet. In **f**, black circles represent all-helical proteins, green squares sheet proteins and red diamonds proteins comprising both helix and sheet structures. The correlation between contact order and folding rate (k_f) is striking, occurring both within each structural subclass and within sets of proteins with similar overall folds (proteins structurally similar to the α/β protein acyl phosphatase[16] are indicated by blue triangles).



Baker (2000) *Nature* **405**, 39.



Comparison of **(a)** the native state structure and **(b)** the most representative structures of the transition state ensemble of acylphosphatase, determined by all-atom molecular dynamics simulations [26]. Native secondary structure elements are shown in colour (the two α helices are plotted in red and the β sheet is in green). The three key residues for folding are shown as gold spheres [5•,26].

Vendruscolo & Paci (2003) *Curr. Opin. Struct. Biol.* **13**, 82.

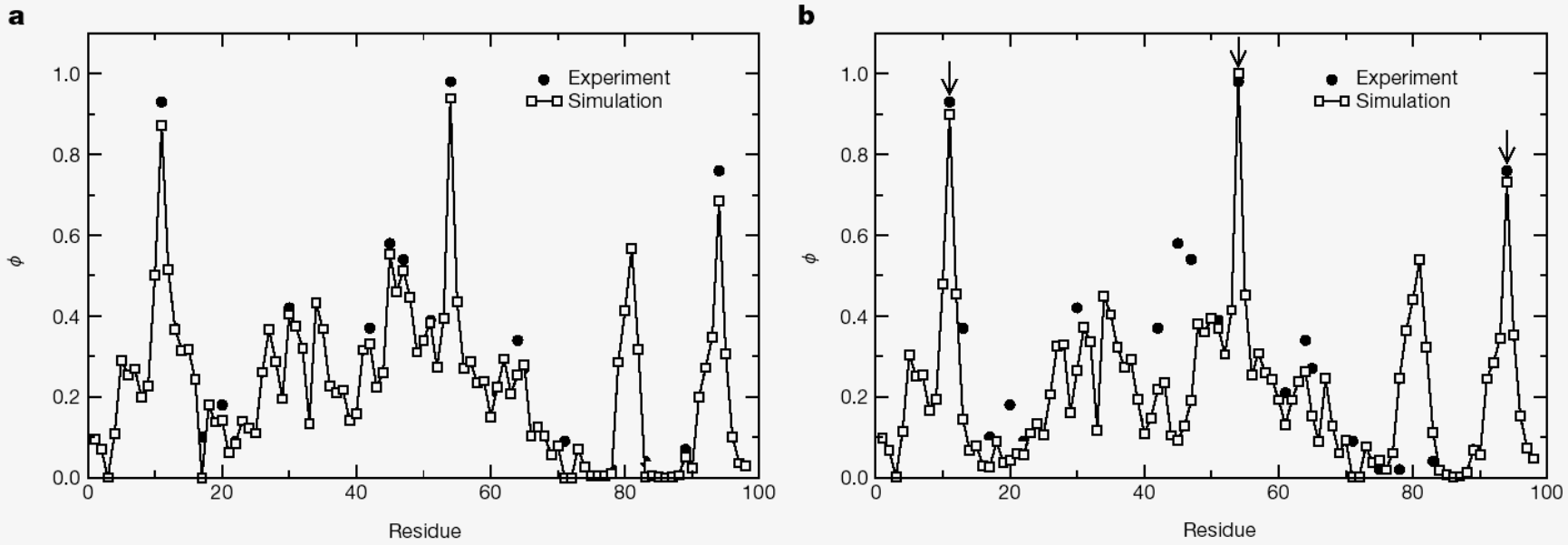


Figure 2 Comparison of ϕ_i^{calc} with ϕ_i^{exp}. **a**, When all 24 of the latter are used as restraints (filled circles); **b**, when only Y11, P54 and F94, are used as restraints (open squares, see text). Only four ϕ_i^{exp} are significantly underestimated; they are V13, T42, G45 and V47. The reason for the underestimation is that this subset of residues forms long-range contacts mainly to each other (for example, V13 with G45) and are, therefore, almost totally unrestrained in the primitive model when Y11, P54 and F94 are used as restraints; the only common contact is between residues Y11 and V47. Except for V13, which is close to Y11, they are not part of the folding core.

Vendruscolo et al (2001) *Nature* **409**, 641.

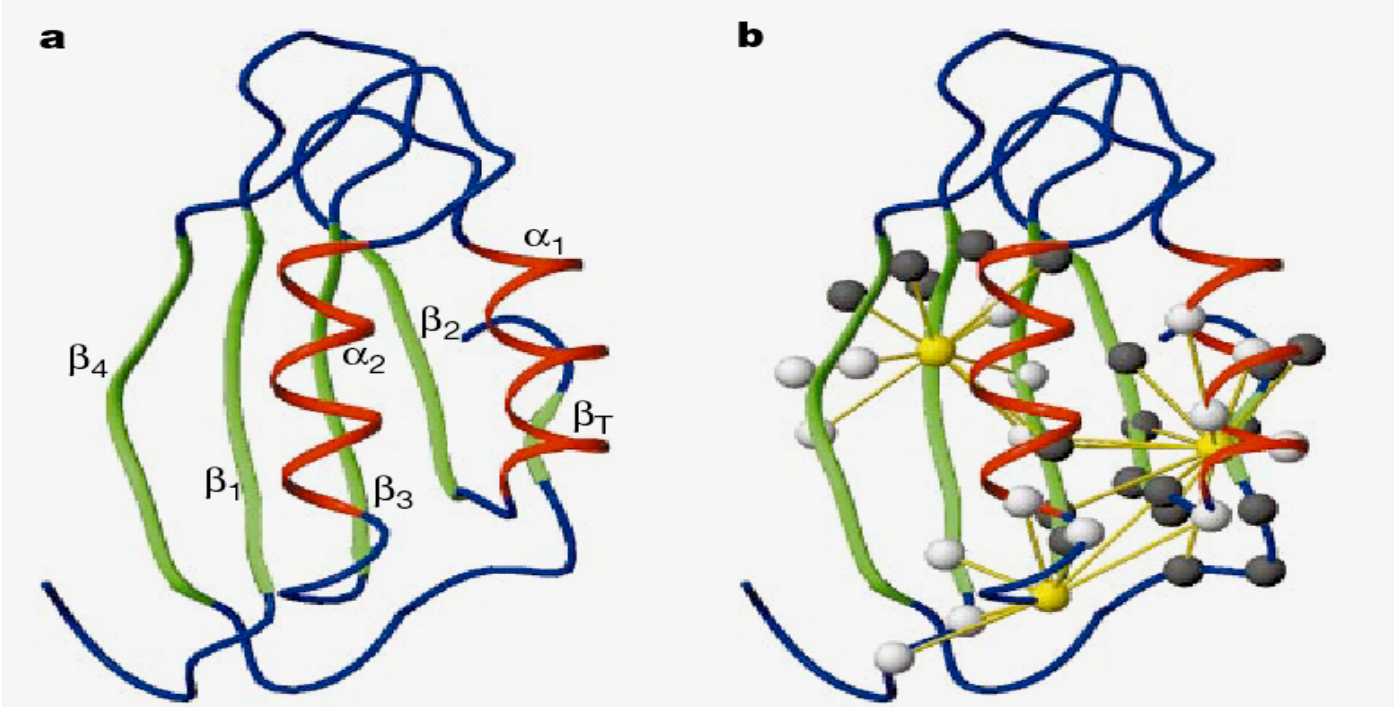


Figure 1 Native structure of acylphosphatase, AcP. **a**, The secondary structure elements are: β_1 (residues 7–13), α_1 (residues 22–33), β_2 (residues 36–42), β_3 (residues 46–53), α_2 (residues 55–56), β_4 (residues 77–85), β_T (residues 93–97); residues between these regions are parts of loops. **b**, the key residues 11, 54 and 94 found from the transition-state analysis are shown as gold spheres on the native structure (see text). They are connected by gold-coloured bonds to the residues (white and black spheres) forming a native contact with them (Y11 has long-range contacts with 47–52 and 78–81, P54 with 5–7 and 34–35, and F94 with 26–31, 36–39 and 50–52). Residues forming the transition-state core identified in the present work are shown as black spheres. They fall into two groups in spatial proximity; there is a larger group (residues 28, 35–39, 51–54 and 90–95) comprising parts of α_1, β_2, α_2 and β_T, and a smaller group (residues 11–13, 47 and 78, 79) comprising parts of the β_1, β_3 and β_4 strands in the native structure.

Vendruscolo et al (2001) *Nature* **409**, 641.

Hydrophobic residues in the src SH3 domain core that network with high Φ-value residues, the interactions of which putatively determine the topology of the transition state ensemble (TSE)

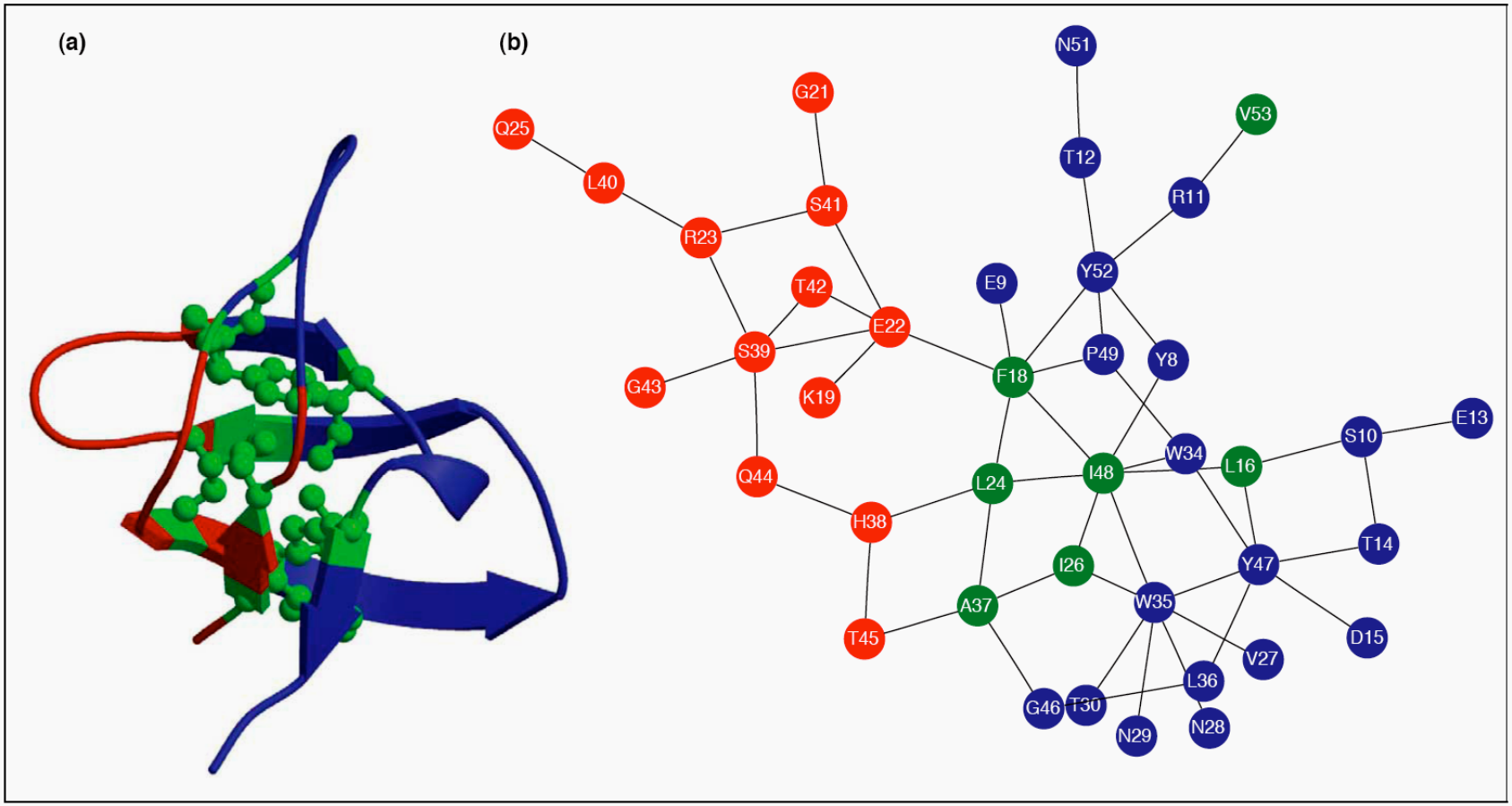


Figure 4. Definition of the topology in the transition state ensemble (TSE) of the Src tyrosine kinase Src homology 3 (SH3) domain by a network of interactions. **(a)** Native-state structure of the SH3 domain in which the residues in the hydrophobic core are shown in a ball-and-stick representation and are coloured green, and the left- and right-hand region of the protein structure in this view is coloured red and blue, respectively. **(b)** Network of interactions in the TSE of SH3. Each node corresponds to a single amino acid residue, each of which is colour-coded using the scheme in (a). The connecting lines indicate strong non-covalent interactions in the TSE between amino acids more than two residues apart [15]. A subset of residues in the hydrophobic core is required to define the overall network of interactions in the TSE. Furthermore, it has been shown that the formation of a network between the same six residues in the core is essential for defining the topology in the TSEs of three different SH3 domains [15].

Lindorff-Larsen et al (2005) *Trends Biochem. Sci.* **30**, 13.

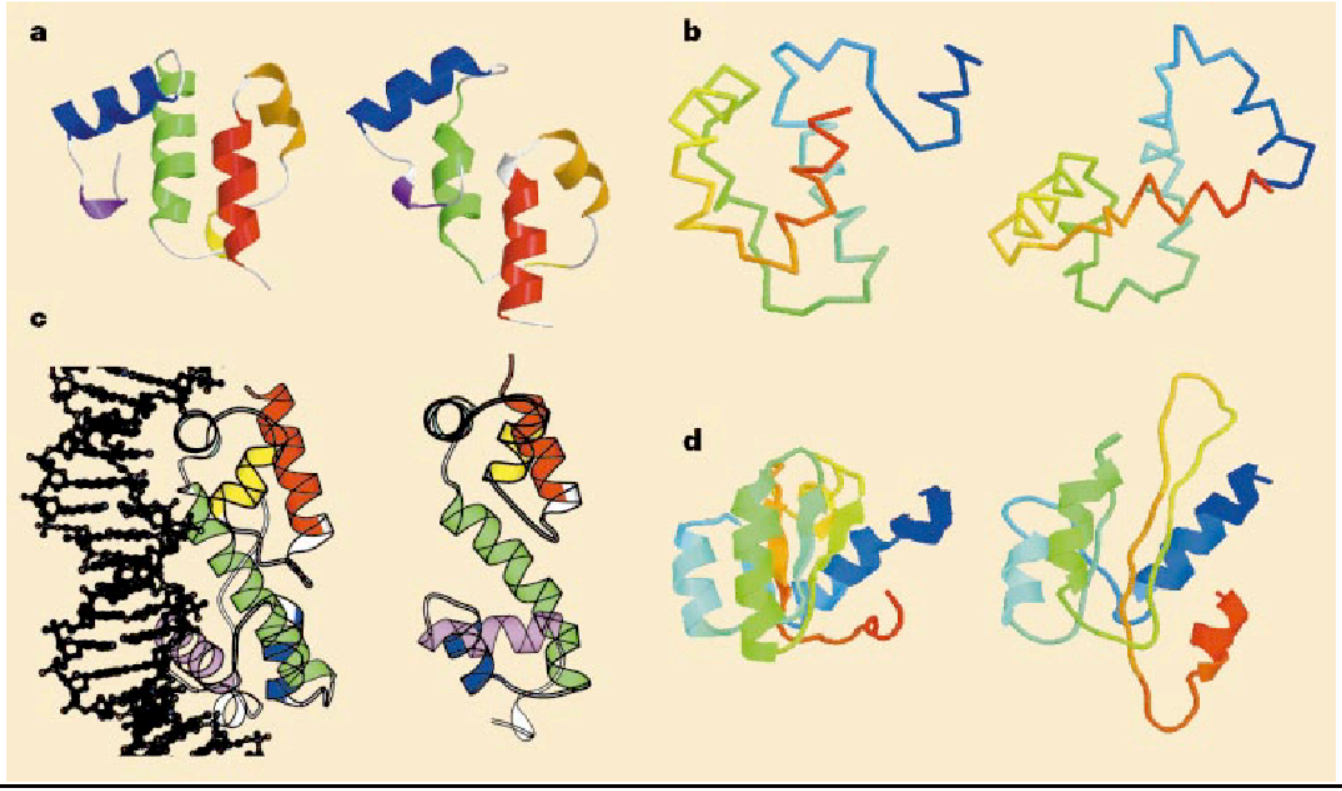
More Computational Protein Folding

What else have the theorists been doing lately?

Box 3

Ab initio structure predictions

Blind *ab initio* structure predictions for the CASP3 protein structure prediction experiment. For each target, the native structure is shown on the left with a good prediction on the right (predictions by Baker[39](**a**, **c**), Levitt[40](**b**) and Skolnick[41](**d**) and colleagues; for more information see http://predictioncentre.llnl.gov/ and *Proteins* Suppl. **3,** 1999). Segments are colour coded according to their position in the sequence (from blue (amino terminus) to red (carboxy terminus)). **a**, DNA B helicase[41]. This protein had a novel fold and thus could not be predicted using standard fold-recognition methods. Not shown are N- and C-terminal helices which were positioned incorrectly in the predicted structure. **b**, Ets-1 (ref. 43). **c**, MarA[44]. This prediction had potential for functional insights; the predicted two-lobed structure suggests the mechanism of DNA binding (left, X-ray structure of the protein–DNA complex). **d**, L30. A large portion of this structure was similar to a protein in the protein databank but the best *ab initio* predictions were competitive with those using fold-recognition methods. The three approaches that produced these predictions used reduced-complexity models for all or almost all of the conformational search process.



Baker (2000) *Nature* **405**, 39.

Progress in *de novo* protein structure prediction & design

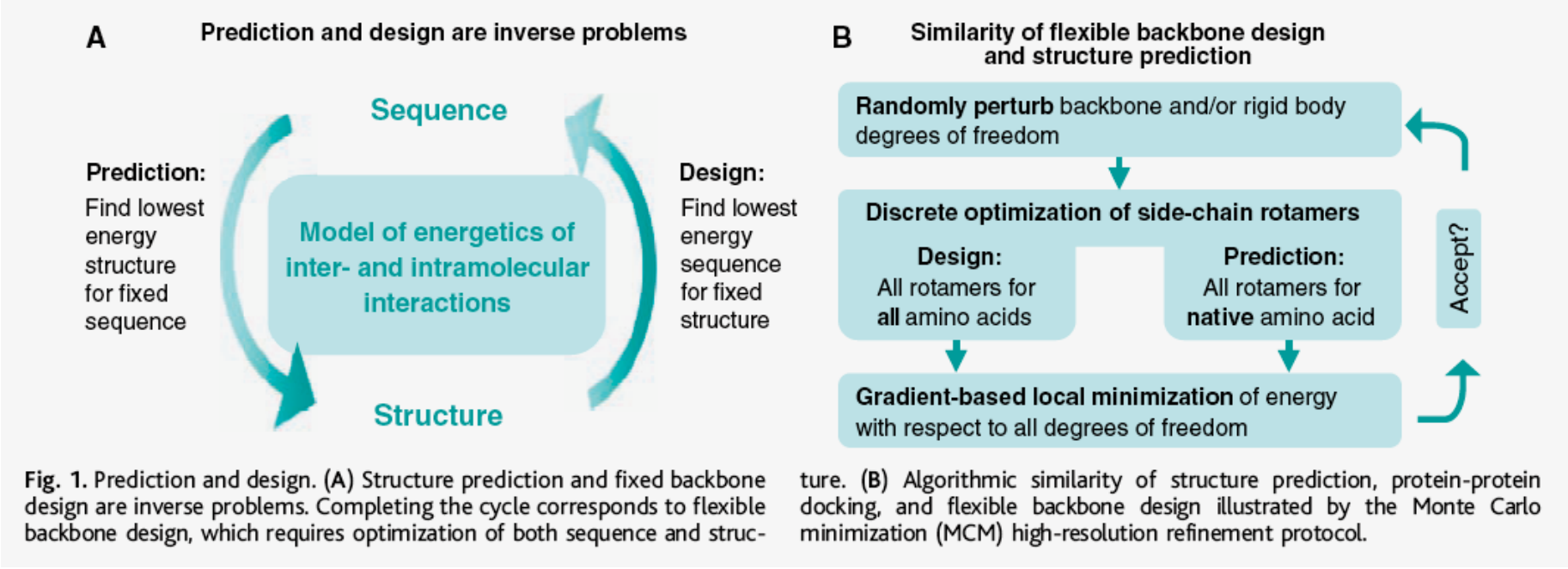


Fig. 1. Prediction and design. (A) Structure prediction and fixed backbone design are inverse problems. Completing the cycle corresponds to flexible backbone design, which requires optimization of both sequence and structure. (B) Algorithmic similarity of structure prediction, protein-protein docking, and flexible backbone design illustrated by the Monte Carlo minimization (MCM) high-resolution refinement protocol.

Schueler-Furman et al (2005) *Science* **310**, 638.

Present protein folding theory suggests that proteins must find the right polypeptide chain topology ("topomer") first, then they can form 2° structure and snap into the correct 3D conformation relatively quickly.

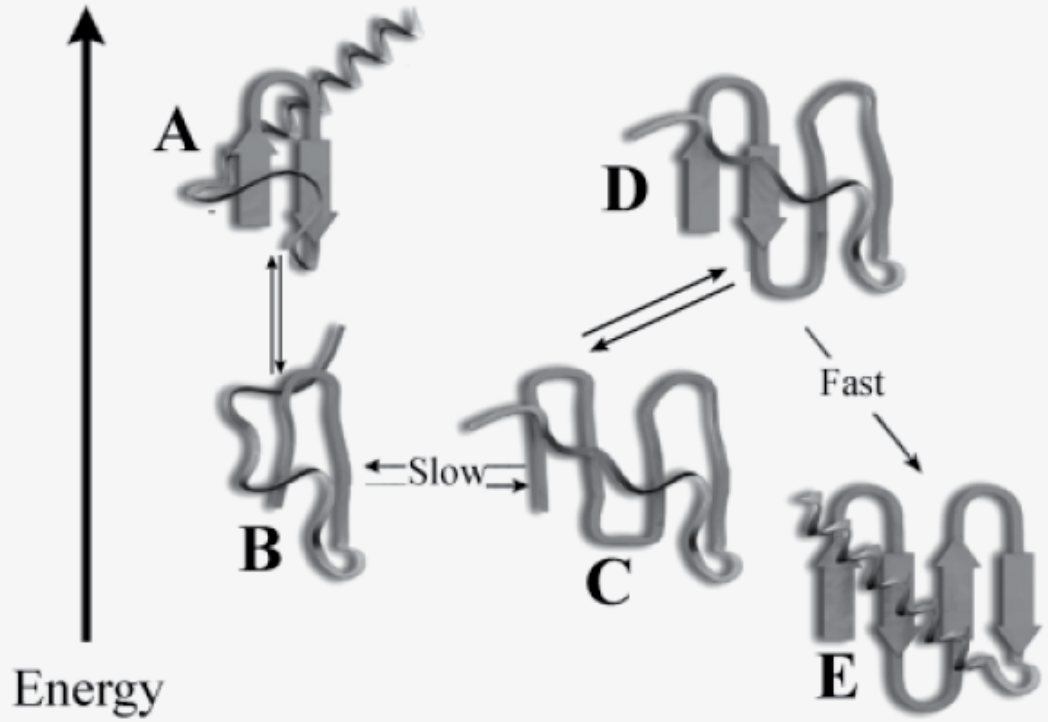


Figure 10 The essence of the topomer search model is that the rate with which an unfolded polymer diffuses between distinct topologies is much slower than the rate with which local structural elements zipper (and, critically, unzip) [reviewed in (59)].

Gillespie & Plaxco (2004) *Ann. Rev. Biochem.* **73**, 837.

Structures of hypothetical folding intermediates*, including a putative transition state, obtained by a molecular dynamics (MD) simulation of protein *un*folding

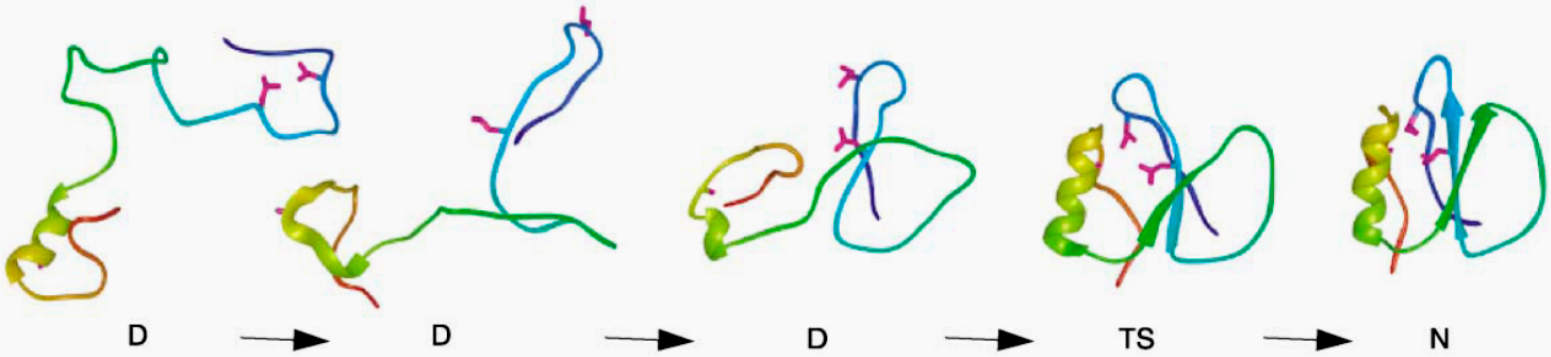


Figure 2. Folding Pathway of CI2

An MD unfolding simulations from the native state N to the denatured state(s) D at 225°C shown in reverse. The structures are colored from red at the N terminus to blue at the C terminus. The 3, 2, and 1 ns structural snapshots, shown from left to right, are members of the D ensemble. NMR, circular dichroism, and fluorescence spectroscopy on the denatured state and peptide fragments show that there is little residual structure apart from a very weak conformational bias to form the helical region. Fragments of increasing length from the N terminus do not form compact structures until residues 49–57 are added, which pack against and stabilize the native helix. The transition state (TS, 225 ps) is built around an extended nucleus, in which L49 and I57 pack against Ala16 (shown in magenta), toward the N terminus of the α helix.

***Note: structures shown are actually only representative structures of *ensembles* of intermediates.**

Fersht & Daggett (2002) *Cell* **108**, 573.

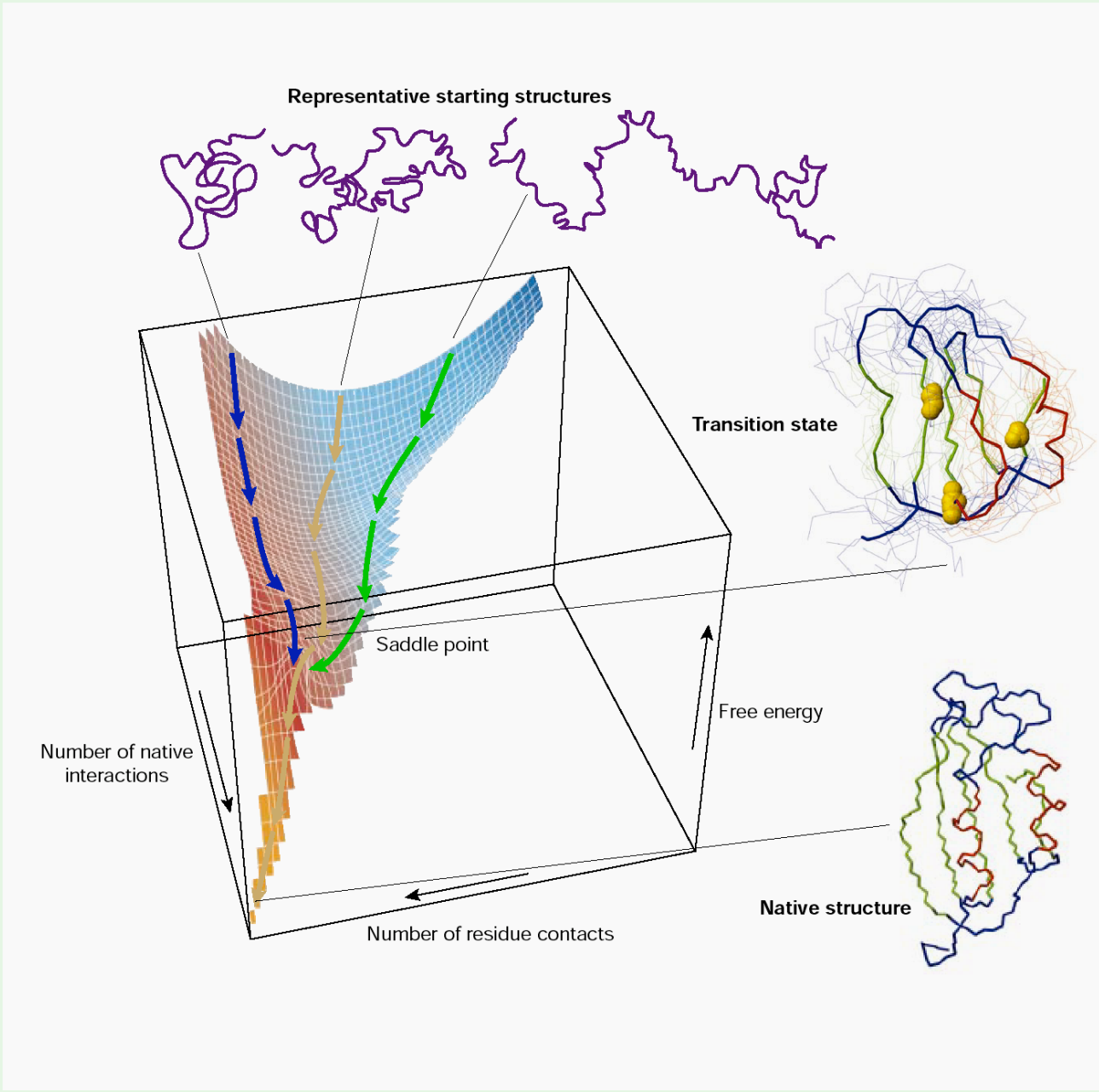


Figure 1 A schematic energy landscape for protein folding. The surface is derived from a computer simulation of the folding of a highly simplified model of a small protein. The surface 'funnels' the multitude of denatured conformations to the unique native structure. The critical region on a simple surface such as this one is the saddle point corresponding to the transition state, the barrier that all molecules must cross if they are to fold to the native state. Superimposed on this schematic surface are ensembles of structures corresponding to different stages of the folding process. The transition state ensemble was calculated by using computer simulations constrained by experimental data from mutational studies of acylphosphatase[18]. The yellow spheres in this ensemble represent the three 'key residues' in the structure; when these residues have formed their native-like contacts the overall topology of the native fold is established. The structure of the native state is shown at the bottom of the surface; at the top are indicated schematically some contributors to the distribution of unfolded species that represent the starting point for folding. Also indicated on the surface are highly simplified trajectories for the folding of individual molecules. Adapted from ref. 6.

Dobson (2003) *Nature* **426**, 884.

Calculating structures with their associated energies: Deep & narrow energy wells are a hallmark of near-correct structures

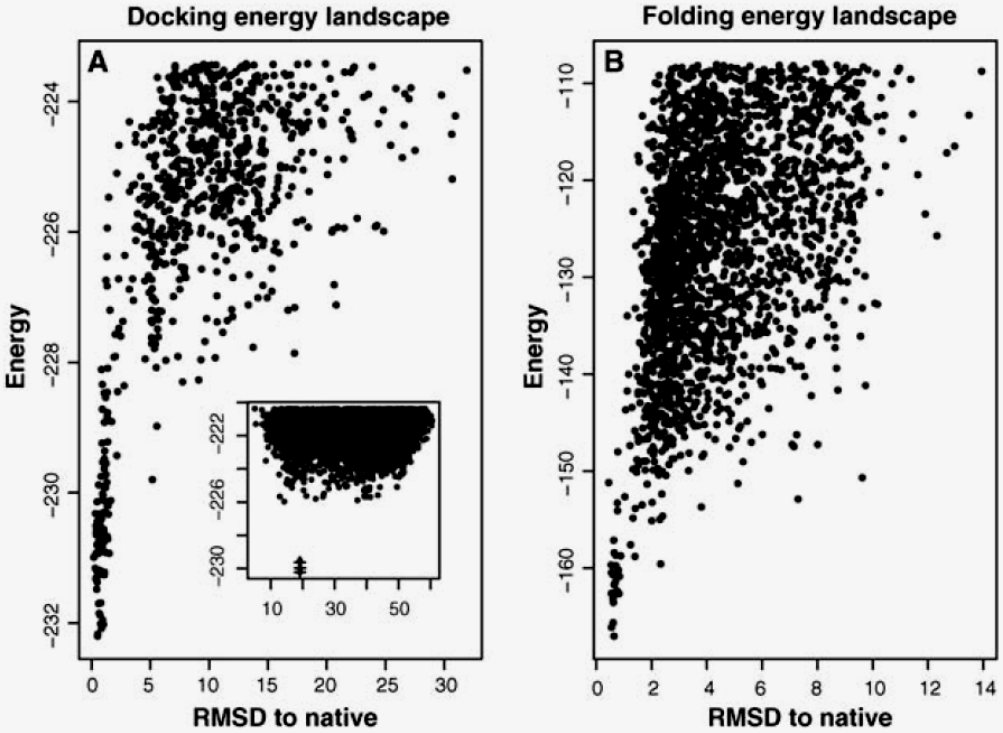


Fig. 2. Energy landscapes. Each point represents the lowest energy structure sampled in a single MCM trajectory. (**A**) Docking energy landscape for Capri Target 12 [cohesin-dockerin complex; PDB (Protein Data Bank) ID 1ohz (*66*)]. (**Inset**) In a large collection of trajectories starting from different random orientations that were carried out for the CAPRI experiment, a small number of structures (+) are distinguished from the background population by a significant energy gap. The *x* axis is the RMSD to an arbitrary reference orientation. (Main panel) Trajectories starting from these low-energy structures map out a narrow energy funnel. The *x* axis is the RMSD to the native structure. A deep energy funnel, as in this example, is a strong indicator that a prediction is correct. (**B**) Folding landscape for double-stranded RNA binding protein [PDB ID 1di2 (*67*)]. The backbone RMSD is to the native structure. The energy function (units are in kcal/mol) includes entropic contributions from solvation effects, but not the configurational entropy associated with protein vibrational and side-chain degrees of freedom, and hence is not the true free energy.

Schueler-Furman et al (2005) *Science* **310**, 638.

For small, single domain proteins, some of the predictions are getting very good!

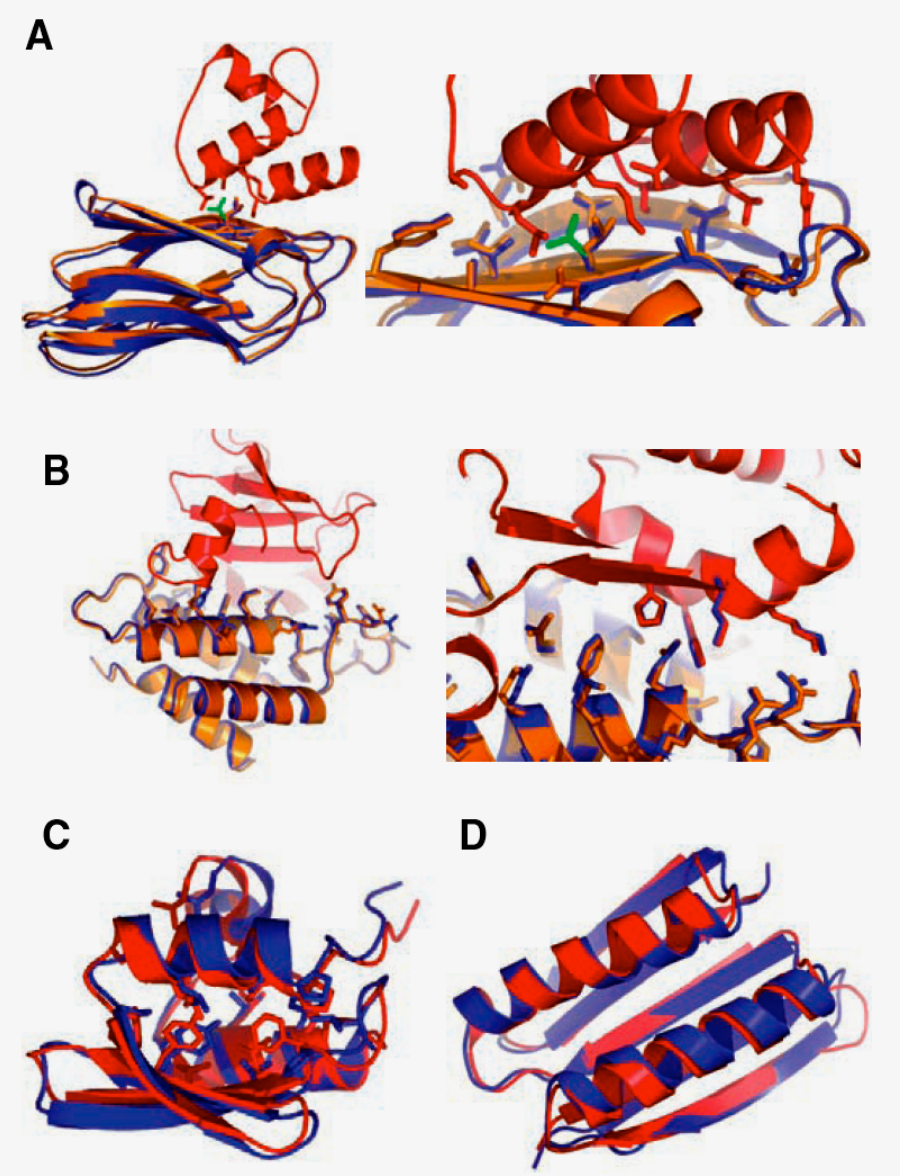


Fig. 3. Examples of high-resolution prediction and design. (**A**) CAPRI Target 12 [dockerin–cohesin (*66*); interface residue backbone RMSD = 0.27 Å]. The lowest energy structure in Fig. 2A, main panel, is shown here. The side chain of Leu-83 (green in the free monomer) changes conformation upon binding. Side-chain conformations in red were provided; those in blue were predicted. (**B**) CAPRI Target 15 [ColicinD–Immunity protein D (*68*); interface residue backbone RMSD = 0.23 Å]. No side-chain information was provided for either partner. (**C**) CASP6 de novo structure prediction Target 0281 [hypothetical protein from *Thermus thermophilus* Hb8, PDB ID 1whz (*69*); backbone RMSD = 1.59 Å]. (**D**) TOP7 (RMSD = 1.2 Å) (*42*). (A) and (B) are adapted from figure 1 of (*70*). Blue: models; red and orange: x-ray structures.

Schueler-Furman et al (2005) *Science* **310**, 638.

You, too, can do theoretical protein folding with David Baker!!

See http://boinc.bakerlab.org/rosetta

But make sure your computer doesn't overheat!

Folding@home

http://folding.stanford.edu/

(Vijay Pande lab)