Protein Science (1995), *4*:561–602. Cambridge University Press. Printed in the USA.

REVIEW

Principles of protein folding – A perspective from simple exact models

KEN A. DILL,' SARINA BROMBERG,[1] KAIZHI YUE,' KLAUS M. FIEBIG,[1,3]
DAVID P. YEE,[1,4] PAUL D. THOMAS? AND HUE SUN CHAN'
[1] Department of Pharmaceutical Chemistry, Box 1204, University of California, San Francisco, California 94143-1204
'Graduate Group of Biophysics, Box 0448, University of California, San Francisco, California 94143-0448

(Received September 23, 1994; Accepted January 9, 1995)

Abstract

General principles of protein structure, stability, and folding kinetics have recently been explored in computer simulations of simple exact lattice models. These models represent protein chains at a rudimentary level, but they involve few parameters, approximations, or implicit biases, and they allow complete explorations of conformational and sequence spaces. Such simulations have resulted in testable predictions that are sometimes unanticipated: The folding code is mainly binary and delocalized throughout the amino acid sequence. The secondary and tertiary structures of a protein are specified mainly by the sequence of polar and nonpolar monomers. More specific interactions may refine the structure, rather than dominate the folding code. Simple exact models can account for the properties that characterize protein folding: two-state cooperativity, secondary and tertiary structures, and multistage folding kinetics—fast hydrophobic collapse followed by slower annealing. These studies suggest the possibility of creating "foldable" chain molecules other than proteins. The encoding of a unique compact chain conformation may not require amino acids; it may require only the ability to synthesize specific monomer sequences in which at least one monomer type is solvent-averse.

Keywords: chain collapse; hydrophobic interactions; lattice models; protein conformations; protein folding; protein stability

We review the principles of protein structure, stability, and folding kinetics from the perspective of simple exact models. We focus on the "folding code"—how the tertiary structure and folding pathway of a protein are encoded in its amino acid sequence. Although native proteins are specific, compact, and often remarkably symmetrical structures, ordinary synthetic polymers in solution, glasses, or melts adopt large ensembles of more expanded conformations, with little intrachain organization. With simple exact models, we ask what are the fundamental causes of the differences between proteins and other polymers—What makes proteins special?

One view of protein folding assumes that the "local" interactions among the near neighbors in the amino acid sequence, the interactions that form helices and turns, are the main determinants of protein structure. This assumption implies that isolated helices form early in the protein folding pathway and then assemble into the native tertiary structure (see Fig. 1). It is the premise behind the paradigm, primary → secondary → tertiary structure, that seeks computer algorithms to predict secondary structures from the sequence, and then to assemble them into the tertiary native structure.

Here we review a simple model of an alternative view, its basis in experimental results, and its implications. We show how the nonlocal interactions that drive collapse processes in heteropolymers can give rise to protein structure, stability, and folding kinetics. This perspective is based on evidence that the folding code is not predominantly localized in short windows of the amino acid sequence. It implies that collapse drives secondary structure formation, rather than the reverse. It implies that proteins are special among polymers not primarily because of the 20 types of their monomers, the amino acids, but because the amino acids in proteins are linked in ***specific sequences.*** It implies that the folding code resides mainly in global patterns of contact interactions, which are nonlocal, and arise from the arrangements of polar and nonpolar monomers in the sequence.

We review here the simple exact models that can address these questions of general principle. Such questions are often difficult to address by other means, through experiments, atomic simulation, Monte Carlo partial sampling, or approximate theoretical models. "Simple" models have few arbitrary parameters.

Reprint requests to: Ken A. Dill, Department of Pharmaceutical Chemistry, Box 1204, University of California, San Francisco, California 94143-1204; e-mail: dill@maxwell.ucsf.edu.

[3] Present address: New Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QY, UK.

'Present address: Department of Molecular Biotechnology, University of Washington, GJ-10, Seattle, Washington 98195.

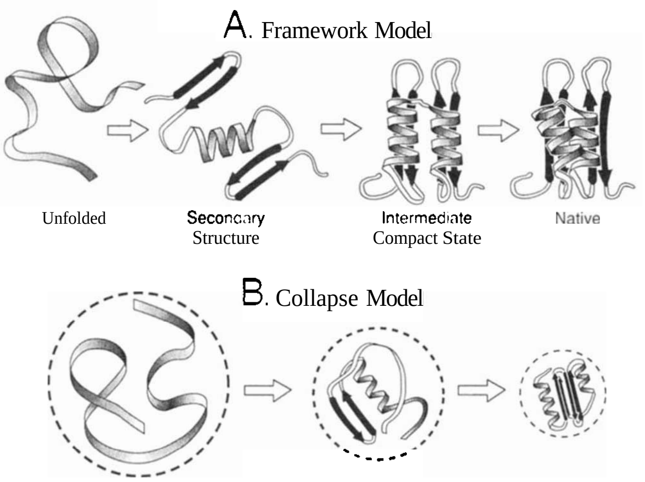

Fig. 1. **A:** If folding is driven by local interactions, secondary structure formation precedes collapse (adapted from Ptitsyn, 1987). R: If folding is driven by nonlocal interactions, collapse drives concurrent secondary structure formation.

"Exact" models have partition functions from which physical properties can be computed without further assumptions or approximations. Simple exact models are crude low-resolution representations of proteins. But while they sacrifice geometric accuracy, simple exact models often adequately characterize the collection of all possible sequences of amino acids (sequence space) and the collection of all possible chain conformations (conformational space) of a given sequence. For many questions of folding, we believe that complete and unbiased characterizations of conformational and sequence spaces are more important than atomic detail and geometric accuracy. We do not review the broader field of protein structure prediction, simulations with sparse sampling of conformational space, or simulations with multiple parameters. Excellent reviews of those areas are available elsewhere (Skolnick & Kolinski, 1989; Covell & Jernigan, 1990; Finkelstein & Reva, 1991; Hinds & Levitt, 1992; Kolinski & Skolnick, 1994; Merz & LeGrand, 1994).

Assumptions

The fundamental constraints on proteins

Interactions among spatially neighboring amino acids can be divided into ***local*** interactions among monomers that are close together in the chain sequence and ***nonlocal*** interactions among monomers that are widely separated in the sequence (Chan & Dill, 1991a). Both local and nonlocal interactions play a role in protein folding. The energies of the ϕ–ψ angles of the peptide bond (local interactions) dictate what backbone conformations are possible. But chains have a great deal of freedom to configure within the accessible bond angles and to satisfy steric constraints. The entropy of all this remaining freedom must be overcome by the forces of folding.

Local interactions and hydrogen bonding

The first view held that protein folding is dominated by hydrogen bonding (Mirsky & Pauling, 1936; Eyring & Stearn, 1939). In crystallographic and model studies of amino acids, Pauling and his colleagues postulated the existence of hydrogen bonded a-helices and β-sheets (Pauling et al., 1951; Pauling & Corey, 1951a, 1951b, 1951c, 1951d). The first crystal structures of globular proteins confirmed the presence and the importance of α-helices (Kendrew et al., 1960). Experimental measurements of the helix–coil transitions of synthetic polypeptides in solution were successfully modeled to account for the cooperativity of the helix–coil transition (Schellman, 1958; Zimm & Bragg, 1959; Poland & Scheraga, 1970; Scholtz & Baldwin, 1992). Studies of model peptides have established a quantitative understanding of the energetic contributions of the different amino acids to helix–coil transitions in water (Sueki et al., 1984; Lyu et al., 1990; O'Neil & DeGrado, 1990; Chakrabartty et al., 1991; Scholtz et al., 1991; Dyson et al., 1992a, 1992b; Scholtz & Baldwin, 1992; Vila et al., 1992; Padmanabhan et al., 1994), their capping interactions (Harper & Rose, 1993; Lyu et al., 1993), and the stabilities of turns in model peptides in solution (Wright et al., 1988). Interestingly, there is some evidence that the helical propensities in water are somewhat different than helical propensities in nonpolar environments, which may be better models of a protein interior (Waterhous& Johnson, 1994; Shiraki et al., 1995). Although hydrogen bonding and helical propensities have a strong historical link, they are not identical. Hydrogen bonding occurs in both local and nonlocal interactions, whereas helical and turn propensities describe only local interactions. Local interactions are strong determinants of the conformations of short peptides and fibrous proteins, but the following evidence suggests that they are weaker determinants of the conformations of globular proteins.

Nonlocal forces: hydrophobic interactions are important

In the 1950s, Walter Kauzmann argued that hydrogen bonds may not be the principal determinant of the structures of globular proteins, reasoning that the strength of the hydrogen bonds between the denatured protein chain and surrounding water molecules would be approximately the same as the intrachain hydrogen bonds in the native protein (Kauzmann, 1954, 1959). He argued that a strong force for folding proteins was the tendency of nonpolar amino acids to associate in water.

Although it seemed clear that hydrophobicity could drive proteins to become compact and acquire nonpolar cores, hydrophobicity seemed to be too nonspecific to drive the formation of specific native protein folds. By 1975, the synthesis of the two perspectives, one based on helical propensities and the other on nonpolar interactions, led to the view that hydrophobicity was mainly a globularization force that stabilizes compact conformations but does little to craft the specific and sequence-dependent secondary and tertiary architectures of proteins. For instance, Anfinsen and Scheraga (1975) stated that:

> "Evidence is now accumulating to suggest that nearest-neighbor, short-range interactions play the dominant role in determining conformational preferences of the backbones of the various amino acids, but that next-nearest neighbor (medium-range) interactions and, to a lesser extent, long-range interactions involving the rest of the protein chain are required to provide the incremental free energy to stabilize the backbone of the native structure."

This view gained credence from the partial success of (1) models, both computational and experimental, of peptides and protein pieces, and (2) database methods that rationalize helical, sheet, and turn propensities in proteins (Chou et al., 1972; Anfinsen & Scheraga, 1975; Montelione & Scheraga, 1989). Hydrophobicity was seen as "nonspecific," and hydrogen bonding and helical propensities were seen as the "specific" components of the folding code that directs a protein to fold to its unique native structure. A common view until recently (Dill, 1985, 1990) appears to have been that there was no single dominant force in folding.

Here we describe an alternative view, namely that both compactness ***and*** the specific architectures of globular proteins are encoded mainly in nonlocal interactions, as is the folding pathway. We first review experimental evidence for the importance of nonlocal interactions. Then we review predictions from simple exact models based on that premise and corresponding experiments.

Experimental evidence that nonlocal interactions are dominant

1. The water-to-oil transfer free energy, a measure of the interactions among monomer contacts, is large and negative for nonpolar amino acids, consistent with their burial in the protein core to avoid water. The average transfer free energy of a nonpolar amino acid is about −2 kcal/mol (Nozaki & Tanford, 1971).

2. Large positive changes in heat capacities result from unfolding most proteins (Privalov, 1979; Privalov & Gill, 1988), consistent with the solvation of nonpolar molecules in water. Transfers of some polar amino acids to water also have large heat capacity changes, but of opposite sign. Transferring the backbone groups into the folded protein may also involve heat capacity changes and contribute significantly to stability (Makhatadze & Privalov, 1993; Privalov & Makhatadze, 1993; and references therein). But backbone interactions, even if they are strong, cannot be the basis for the folding code, because they are not sequence dependent.

3. The free energies for helix formation are small (Sueki et al., 1984; Chakrabartty et al., 1991, 1994; Scholtz et al., 1991; Scholtz & Baldwin, 1992). For example, a recent free energy scale shows that only alanine is a helix-former (favorable free energy), leucine and arginine are helix-indifferent, and all other amino acids are helix-breakers (Chakrabartty et al., 1994). Site-directed mutagenesis studies show that helical propensities contribute less to the variance of changes in stability than hydrophobic interactions (Alber et al., 1988; Zhang et al., 1991; Pinker et al., 1993; Blaber et al., 1994). Helix stability increases with chain length and with reduced temperature (to near 0 °C) (Poland & Scheraga, 1970). But most helices in proteins are too short (Kabsch & Sander, 1983), and room temperature is too high, for protein helices to be stable by themselves. Predictions of helices in proteins are only around 60–70% correct, where 33% correct is expected from random choice in predicting three categories: helix, sheet, and other (Rooman & Wodak, 1988). Moreover, whereas most studies have been performed on water-soluble helices, most helices in globular proteins are arnphipathic (J. Thornton, pers. comm.), indicating that hydrophobic interactions are important factors stabilizing helices in globular proteins.

4. β-Sheet proteins have few local interactions, and those few are only at the turns, so helical propensities cannot explain the folding of sheet proteins. Nonlocal interactions must dominate the folding of sheet proteins.

5. Electrostatic interactions in proteins generally contribute little to structure and stability, as determined by the general insensitivity of the native structure to pH and salt, except in highly acidic or basic solutions (Dill, 1990). Mutational studies show that varying the charge on T4 lysozyme from +9 to +I leads to no change in structure (Sun et al., 1991). Goto and coworkers (Goto & Nishikiori, 1991; Hagihara et al., 1994) see no change in the native structure or in the native features of the cytochrome ***c*** molten globule at pH 7 with replacement from 0 to 19 positive charges by random acetylation of lysines.

6. Polypeptides can be designed to fold to apparently helical bundles by designing only the sequence of hydrophobic and polar residues, averaging over a variety of helical propensities, side-chain packing, and charge placements (Kamtekar et al., 1993; Munson et al., 1994). Amino acids in native state turn positions can be chosen largely randomly in some cases (Brunet et al., 1993). The tendencies to form helices or strands are more dependent on the solvent than on the amino acid sequence (Zhong & Johnson, 1992; Reed & Kinzel, 1993; Waterhous & Johnson, 1994).

Models

Based on the results above, we take as our premise that proteins are chain molecules that have specific monomer sequences and are driven to fold mainly by nonlocal interactions subject to steric constraints. There is currently no accurate analytical theory that can account for chain connectivity, excluded volume in the compact states, and specific sequences of monomer units. Simple exact models have been developed to explore such properties.

What are simple exact models?

There is more than one simple exact model of proteins. Figure 2 shows examples of model protein conformations in the two- and three-dimensional HP (H: hydrophobic, P: polar) lattice models (Lau & Dill, 1989), as well as conformations of a 27-mer cube "perturbed homopolymer" model (Shakhnovich & Gutin, 1993a; Socci & Onuchic, 1994). In simple exact lattice models, each amino acid is represented as a bead. Connecting bonds are represented by lines. The background lattice simply serves to divide space into monomer-sized units. A lattice site may be either empty or filled by one bead. Bond angles have only a few discrete values, dictated by the structure of the lattice. Many different types of lattices are possible, in both two and three dimensions. In some cases, models in two dimensions (2D) offer physical and computational advantages over models in three dimensions (3D) (see below). For most properties tested so far, 2D and 3D models give similar qualitative results. In the HP model, HH contacts are favorable. In the perturbed homopolymer model, all monomers are strongly attracted to each other, and effects of monomer sequence are treated as relatively small perturbations to this large net attraction. More detailed descriptions of the individual models are given at the end of this review.

The disadvantages of lattice models are clear. Resolution is lost. The details of protein structures and energetics are not accurately represented. Model chain lengths have often been un-

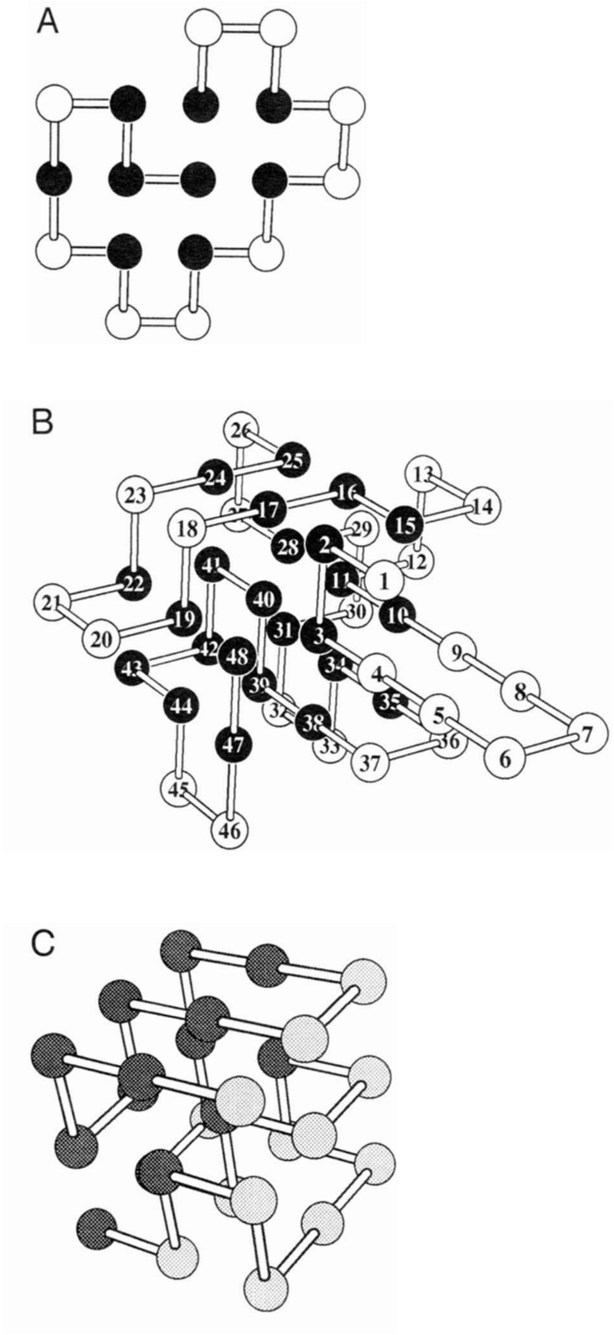

Fig. 2. Examples of native conformations in (A) the 2D HP model (Chan & Dill, **1994),** (B) the 3D HP model (Yue et al., **1995),** and (C) a perturbed homopolymer model (Shakhnovich & **Gutin, 1993a).** In A and B, black and white beads represent H (hydrophobic)and P (polar) monomers, respectively. In the HP model, relative energies of (HH, PP, HP) contacts are (−1, 0, **0).** The chain in C has two types of monomers, "A" and **"B."** Contacts (AA, BB, AB) have energies (−3, −3, −1). Native structures of HP models have hydrophobic cores, whereas those of the perturbed homopolymer model tend to separate into two sides with different monomer types.

realistically short, although this limitation is rapidly being overcome. On the other hand, lattice models have certain virtues. First, atomic-level simulations can currently explore only the small conformational changes that occur in very short times (typically picoseconds to nanoseconds). Lattice models can explore the larger conformational changes and the longer times involved in protein folding. Second, atomic force-field energies include covalent terms, so small conformational changes require computation of very small differences (a few kilocalories) between large energy terms (megacalories). Lattice models avoid this problem by omitting covalent energies. Third, atomic resolution models require many parameters, approximations, and involve incomplete conformational sampling. Simple exact models do not. Fourth, simple exact models can test the assumptions and approximations in analytical models. To our knowledge, all existing analytical theories of proteins make approximations such as those based on mean-field treatments (Dill, 1985; Chan & Dill, **1991a;** Dill & Stigter, **1995),** or approximations from the theory of spin glasses (Edwards & Anderson, 1975; Derrida, 1981; Binder & Young, 1986; Mezard et al., 1986; Fischer & Hertz, 1991) such as the random-energy assumption **(Bryngelson** & Wolynes, 1987; **Garel** & **Orland,** 1988; Shakhnovich & **Gutin, 1989a, 1989b),** and cannot treat specific monomer sequences. The predictions of these theories can be tested by exact models, which correctly account for these factors.

To study molecular properties of models requires computing entropies, energies, and free energies, which are derived from statistical mechanical partition functions. The basic process in computing partition functions is the counting of conformations. Lattice models allow direct enumeration of the conformations, for sufficiently short chains. Counting can be done by computer, taking "excluded volume" fully into account by forbidding any conformation in which two beads occupy the same lattice site. Exact models provide complete or near-complete knowledge of all the relevant conformations, without any approximations beyond those intrinsic to the model itself. In contrast, in molecular dynamics and Monte **Carlo** methods, the relevant conformations are sampled very locally or very sparsely.

Simple exact models have played an important role in polymer science. The first exact enumeration studies of short **homopolymer** chains on the square lattice in 2D and the cubic lattice in 3D were carried out by W.J.C. Orr (1947). Subsequent exact lattice model studies of homopolymers have provided the basis of major modern developments in polymer theory, particularly scaling laws and renormalization group methods (Barber & **Ninham,** 1970; de Gennes, 1979; Freed, 1987; des **Cloizeaux** & **Jannink,** 1990).

Lattice methods were first applied to protein stability and kinetics in the pioneering **"Gō** models" (Taketomi et al., 1975; **Gō** & Taketomi, 1978). **Gō** et al. studied folding kinetics using hypothetical potential functions with Metropolis Monte **Carlo** sampling in 2D and 3D lattice models. In their "strong specificity limit," the native structure is guaranteed to be the lowest-energy state by an ad **hoc** potential function. This potential function counts intrachain attractions only when a pair of monomers is arranged as in the native conformation. Such native forcing potentials are not intended to represent physical interactions because pairs of amino acid residues cannot switch on their attractions only when they are in their native arrangement. **Gō** et al. also studied an "intermediate specificity" case in which some nonnative contacts were permitted to be favorable. Gō models are not simple exact models because the potentials are not physical, and sampling is sparse.

Simple exact models for proteins were initiated in 1989 (Chan & Dill, **1989a, 1989b;** Lau & Dill, 1989) in 2D and for the **max-**

imally compact 27-mer cube and other chains in 3D (Chan & Dill, 1990a) to explore the consequences of physical potentials without approximation, and to study the properties of the full conformational and sequence spaces. Through those and subsequent studies (Chan & Dill, 1990b, 1991b, 1993b, 1994; Lau & Dill, 1990; Shakhnovich & Gutin, 1990a, 1990b, 1993a; Lipman & Wilbur, 1991; Shakhnovich et al., 1991; Leopold et al., 1992; Miller et al., 1992; O'Toole & Panagiotopoulos, 1992; Shortle et al., 1992; Yue & Dill, 1992, 1993, 1995; Camacho & Thirumalai, 1993a, 1993b, 1995; Fiebig & Dill, 1993; Gutin & Shakhnovich, 1993; Stillinger et al., 1993; Thomas & Dill, 1993; Unger & Moult, 1993; Bromberg & Dill, 1994; Gupta & Hall, 1994; Šali et al., 1994a, 1994b; Socci & Onuchic, 1994; Stolorz, 1994; Chan et al., 1995), many properties of simple exact models of proteins are now well understood.

The view that emerges from these studies is that polymers with specific sequences of at least two monomer types can collapse to stable compact states that resemble proteins in several respects. For some sequences the stable states under "native" conditions are compact and unique, with secondary and tertiary structures and nonpolar cores, even in the absence of local interaction biases. The stable structures are often neutral to mutations, more so at the surface than in the core. For many sequences, collapse involves sharp sigrnoidal transitions with corresponding peaks in heat absorption. Some sequences show two-state cooperativity. The denatured states can be compact and complex, depending on external conditions and monomer sequence. Folding kinetics can be multistaged, with concurrent development of compactness and secondary structure followed by slow "annealing" to native states. Sometimes the kinetics manifests itself as many paths and sometimes as particular sequences of events, depending on the property observed. Here we divide our account of these protein properties into three main parts: structure, thermodynamics, and folding kinetics.

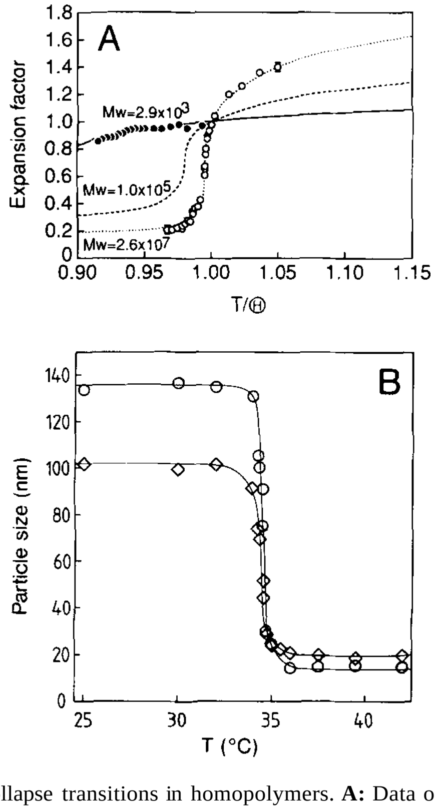

Fig. 3. Collapse transitions in homopolymers. **A:** Data of Sun et al. (1980) for polystyrene in cyclohexane are shown for three chain lengths, indicated by the molecular weight (Mw). Horizontal scale indicates temperature. Numbers of monomers in the chains are approximately **30,** 1,000, and 250,000. Only very long chains show sigmoidal transitions. B: Hydrodynamic radius (O)and radius of gyration (O) of PNIPAM as a function of temperature, in a dilute aqueous solution containing a small amount of surfactant to suppress aggregation. Numbers of monomers in the chains are approximately 62,000. Data from Meewes et al. (1991).

Protein structures

Nonlocal interactions drive collapse transitions, whereas local interactions drive helix transitions

Hydrophobic homopolymers collapse to compact states in water

Homopolymer behavior is the simplest model of chain collapse. Homopolymers are predicted to collapse when they are put into "poor" solvents (i.e., solvents that prefer phase separation to mixing with monomers of the type that comprise the homopolymer) (Anufrieva et al., 1968; Ptitsyn et al., 1968; de Gennes, 1975; Post & Zimm, 1979; Sanchez, 1979; Williams et al., 1981). It is observed experimentally that polystyrene, a chain of nonpolar monomers, collapses to a compact globule in a poor organic solvent (Sun et al., 1980) and poly-(N-isopropylacrylamide) (PNIPAM) collapses very sharply (with *increasing* temperature) in water (Fujishige et al., 1989; Rička et al., 1990; Meewes et al., 1991; Tiktopulo et al., 1994), resembling the renaturation of cold-denatured proteins (Privalov & Gill, 1988; see Fig. 3). As with proteins, PNIPAM collapse is accompanied by a peak in heat absorption (Tiktopulo et al., 1994).

Compactness in chain molecules stabilizes secondary structures

Exact lattice simulations predict that the collapse of polymer chains helps drive the formation of secondary structure, both helices and sheets (Chan & Dill, 1989b, 1990b; see Figs. 4, 5). This conclusion is confirmed in more realistic off-lattice models that show, however, that compactness-induced stabilization is not very structurally specific. For example, Gregoret and Cohen (1991), using a rotational isomeric model of protein chains constrained within ellipsoids of different volumes, show that compactness induces some, but not much, secondary structure if helices and sheets are defined by strict criteria. The results of Hao et al. (1992) and Socci et al. (1994) show that both compactness and intraresidue interactions are needed to approach the bond vector correlations of real proteins. Yee et al. (1994) confine random self-avoiding polyalanine chains to spheres of various diameters using a distance geometry procedure (Havel, 1990) and find that conclusions about compactness-induced secondary structure are strongly dependent on the criteria used to define helices and sheets. This study and one by Hunt et al. (1994) confirm that compactness stabilizes secondary structures (see Fig. 6), but in the absence of hydrogen bonding, helices and sheets only weakly resemble those in proteins. These "vague" helices (i.e., involving broader regions of ϕ–ψ angles than those

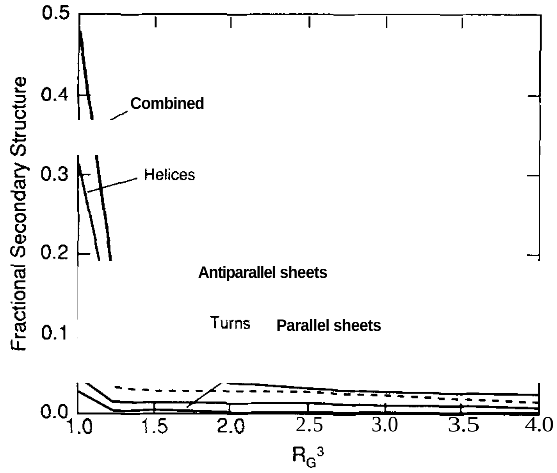

Fig. 4. Model prediction that secondary structure increases with chain compactness. Full ensemble of conformations of 12-segment chains configured on 3D simple cubic lattices, as a function of R_G^3, where R_G is the radius of gyration of the chain (in units of the minimum radius possible for the chains). Data from Chan and Dill (1990b).

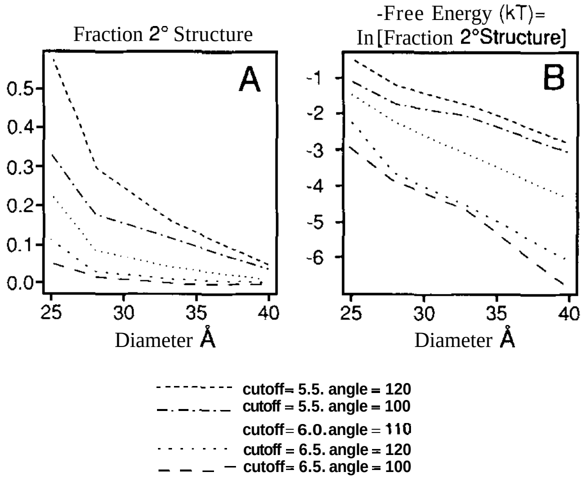

Fig. 6. Off-lattice studies confirm that compactness stabilizes secondary structure. **A:** Fraction of residues in secondary structures, determined by a topological contact method, as a function of the diameter of the sphere confining polyalanine 100-mers. But amounts of secondary structure depend strongly on the criteria used to define them. Cutoffs (maximum separation) defining a contact are varied from 5.5 to 6.5 A, the minimum bond angle required for residues to be assigned to a sheet is varied from 100° to 120°. B: Entropic stabilization due to chain compactness (natural logarithm of fraction of residues in secondary structure) is relatively independent of the criterion used to identify secondary structures (Yee et al., 1994).

of well-defined helices in globular proteins) can be pushed into "good" α-helices by the introduction of small hydrogen bonding forces, but "good" sheets require larger perturbations. Thus, compactness stabilizes ensembles of conformations that are roughly helix-like and sheet-like, but hydrogen bonding or other interactions are needed to "lock in" specific a-helices and β-sheets. Consistent with this picture, recent results from molecular dynamics simulations of chymotrypsin inhibitor 2 by **A.** Li and V. Daggett (submitted) show a correlation of compactness with increasing amounts of secondary structure (Fig. 7).

Consistent with the model predictions, experiments show that secondary structure is correlated with protein compactness. (1)

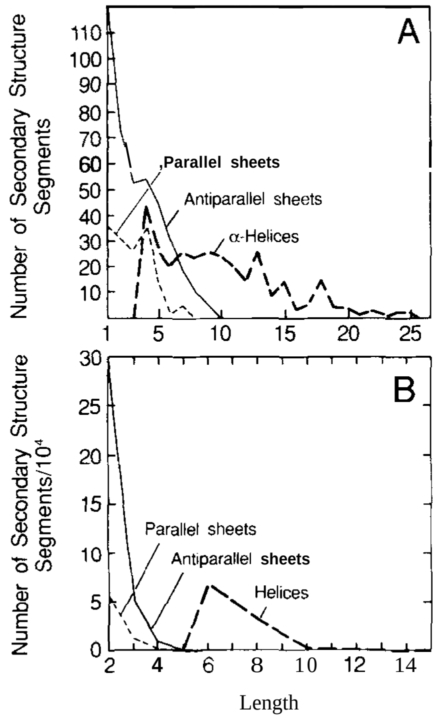

Fig. 5. Length distribution of secondary structures. **A:** Database observations of Kabsch and Sander (1983). B: Exhaustive simulations of maximally compact chains of 26 residues on 2D square lattices (Chan & Dill, 1990b).

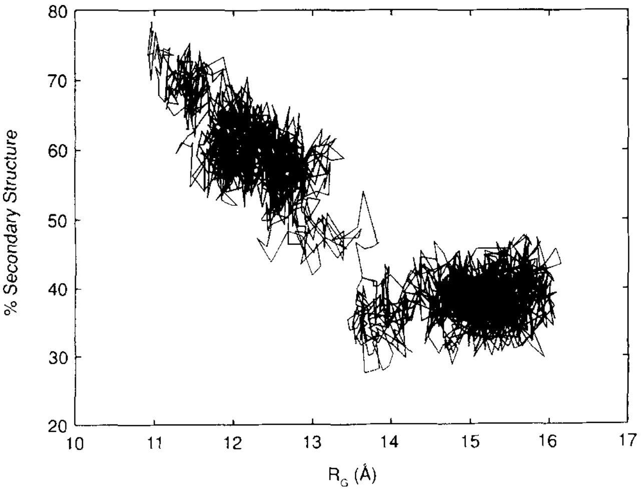

Fig. 7. Molecular dynamics trajectory of chymotrypsin inhibitor 2, showing increasing amount of secondary structure with decreasing radius of gyration R_G (**A.** Li & V. Daggett, unpubl. results, reproduced with permission; see also Li & Daggett [1994]).

By varying solvent conditions, DnaK (Palleros et al., 1993), apomyoglobin, and ferricytochrome *c* (Nishii et al., 1994; M. Kataoka, I. Nishii, T. Fujisawa, T. Ueki, F. Tokunaga, & Y. Goto, in prep.) can each be caused to have different radii. The amount of secondary structure increases with chain compactness (see Fig. 8). (2) Measuring the CD of random terpolymers of lysine, alanine, and glutamic acid, Rao et al. (1974) found that the highly compact conformations of random sequences are 46% helix. (3) In several proteins, equilibrium compact denatured states have much secondary structure (reviewed by Kuwajima, 1989; Ptitsyn & Semisotnov, 1991; Ptitsyn, 1992). In apocytochrome *c*, it appears that secondary structure is lost sharply when the radius of the molecule has expanded to somewhere between 18 and 22 A (Hamada et al., 1993). Interestingly, Jeng and Englander (1991) observed considerable helix in acid-denatured cytochrome *c*, even when it has a large radius at low salt in deuterated solutions. Because helices are not stable in isolation, Jeng and Englander attribute the helix formation to localized clustering. Although lattice model studies suggest that some proteins might assemble through isolated domains in this way (Lattman et al., 1994), experiments by Goto et al. (1993) were unable to confirm the observations of Jeng and Englander.

The perspective described above is that secondary structures develop as an indirect consequence of hydrophobic collapse, due to steric and compactness constraints. But another consequence of hydrophobic collapse is the decrease of the internal dielectric constant, which would strengthen the hydrogen bonding, helical dipoles, and other electrostatic interactions within the core. Thus, collapse might stabilize secondary structures through both specific and nonspecific mechanisms.

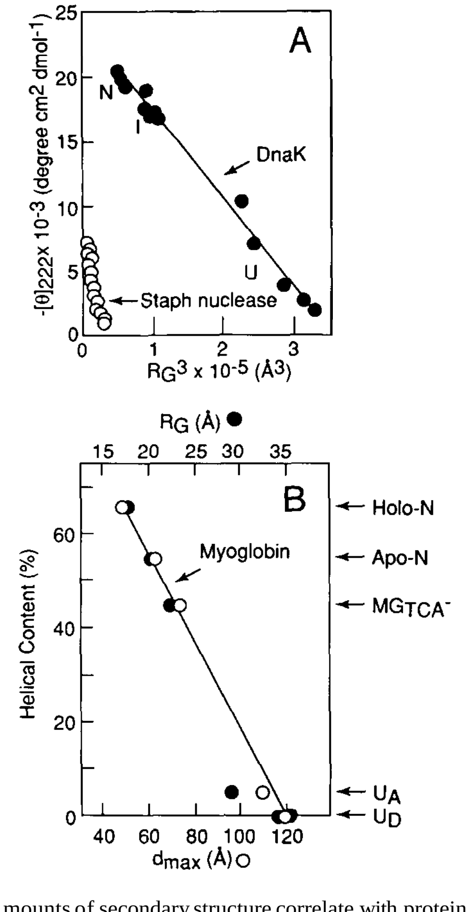

Fig. 8. Amounts of secondary structure correlate with protein compactness. A: Ellipticity at 222 nm versus R_G^3 for DnaK from data of Palleros et al. (1993), and staphyloccocal nuclease fragments and mutants, from data of Shortle and Meeker (1989). B: The data of M. Kataoka, I. Nishii, T. Fujisawa, T. Ueki, F. Tokunaga, and Y. Goto (in prep., reproduced with permission) show the content of α-helix estimated by CD increases with chain compactness (see also Nishii et al., 1994). Labels U_D, U_A, MG_{TCA}-, Apo-N, and Holo-N represent denaturant-induced unfolded state, acid-induced unfolded state, TCA$^-$-stabilized molten globule state, native state of apomyoglobin, and holomyoglobin native state, respectively. Radius of gyration R_G (●) and largest linear dimension d_{max} (O) of the proteins are determined by solution X-ray scattering.

Homopolymers do not collapse to unique states

How does an amino acid sequence encode only a ***single*** native conformation and exclude all others? We call this the encoding problem. Homopolymer models do not account for encoding because homopolymer collapse does not lead to a unique configuration. Although the maximally compact conformations of a polymer constitute only an infinitesimal fraction of all conformations, around 10^{-40} to 10^{-30} for 100-mers (from a mean-field theory: Dill, 1985; in exact 2D models: Chan &Dill, 1989b; Camacho & Thirumalai, 1993b; in exact 3D models: Chan & Dill, 1990a, 1991a), the absolute number of compact conformations is still quite large, and it grows exponentially with chain length (see Fig. 9). By itself, the steric exclusion in a compact chain cannot account for the encoding of a unique protein fold in an amino acid sequence.

Protein folding is better modeled as heteropolymer collapse

More accurate models recognize that proteins are not homopolymers, but heteropolymers, composed of different types of monomers. The simplest protein model divides the amino acids into two categories: hydrophobic (H) and ionic or polar (P)

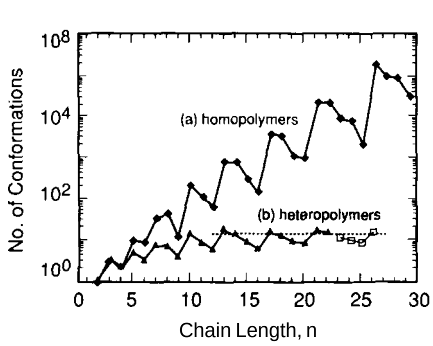

Fig. 9. Numbers of maximally compact conformations versus chain length. Upper curve (a): The number of maximally compact conformations on 2D square lattices grows exponentially as a function of chain length *n* (Chan & Dill, 1989b; Camacho & Thirumalai, 1993b). Lower curve (b): The number of maximally compact conformations that have the maximum number of HH contacts, averaged over HP sequences, is relatively small and becomes relatively independent of chain length. Data for chain length $n \le 22$ are obtained by exhaustive enumerations; data for n = 23–26 are estimated using randomly generated HP sequences. Long chain length limit is shown by the dotted line (Camacho & Thirumalai, 1993b).

(Dill, 1985; Lau & Dill, 1989, 1990; Chan & Dill, 1991b). Such a "two-letter alphabet" is the most elementary approximation to the true 20-letter amino acid alphabet. H P model chains have specific sequences of H-type and P-type monomers. The attraction between H monomers models the tendency of H P polymers in water to collapse to minimize the exposure of their H monomers to solvent and to P monomers (Dill, 1985; Lau & Dill, 1989; Chan & Dill, 1991b). Such collapse processes lead to compact states with nonpolar cores. The P monomers tend to the surface, driven by the HH attraction. Because compactness enhances ordered structure formation, as noted above, the collapse of heteropolymers induces secondary structure (Chan & Dill, 1991b). Another two-letter alphabet model is the "AB" model (Shakhnovich & Gutin, 1993a), in which there are strong AA and BB attractions and a weak AB attraction. This is not a physical model of hydrophobic and polar interactions: it leads to a "left–right" separation of monomers (see Fig. 2), rather than to an interior core and polar surface. Nevertheless, some folding predictions are similar in H P and AB models (see below).

Encoding unique native protein structures

Heteropolymers collapse to very few structures

Remarkably, whereas model homopolymers collapse to very many compact conformations, most model heteropolymer sequences collapse to very few lowest-energy conformations (Lau & Dill, 1989, 1990; Chan & Dill, 1991b, 1994; Camacho & Thirumalai, 1993b). What fraction of H P sequences have unique native structures? We use the term "degeneracy," g_N (≥ 1), to denote the number of lowest-energy (native) conformations of a sequence. When $g_N = 1$, a sequence has only a single conformation of lowest free energy, a unique "native state." Real proteins generally have small degeneracies. The fraction of H P sequences that fold to unique conformations ($g_N = 1$) in the 2D model is about 2.1–2.4%, depending slightly on chain length (Fig. 10). The 2D H P model predicts that most sequences have relatively small degeneracies (Chan & Dill, 1991b). Even though 5,808,335 conformations are accessible to each sequence with 18 monomers, more than half of the 18-mer H P sequences have g_N less than 50 (H.S. Chan & K.A. Dill, unpubl. results). Camacho and Thirumalai (1993b) have extended these conclusions to longer chain lengths by limiting their exhaustive enumerations to maximally compact conformations (see Fig. 9).

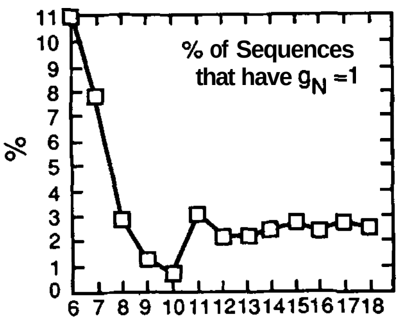

Fig. 10. Percentage of HP sequences that have unique native structures (g = 1) on 2D square lattices as a function of chain length *n*. (Data from Chan and **Dill** [1991b, 1994].)

Similar conclusions appear to hold for longer model chains (30-mers to 88-mers) in 3D, but 3D studies are less complete than 2D studies. In the longer chain 3D studies, some 88-mer H P sequences encode fewer than five native (lowest energy) conformations (Yue & Dill, 1995). This is a very small number compared to the maximum possible degeneracy available (10^{60}) to a sequence of that length. This enormous reduction of conformations indicates that the essentials of the folding code may be given by the sequence of hydrophobic and polar monomers.

Larger code alphabets promote uniqueness

In models with larger alphabets, or more types of interactions, more of the possible chain sequences have unique native conformations (O'Toole & Panagiotopoulos, 1992; Shakhnovich, 1994). Also, the alphabet size may determine the kinetic and thermodynamic difficulty of folding, with certain sets of larger alphabets favoring faster folding and greater stability. How protein-like are the alphabets and interaction energies used in model studies? The H P model represents minimal encoding, using only two "letters." Gō models (Taketomi et al., 1975; Go & Taketomi, 1978) and codes that allow independent variation of every contact energy (Shakhnovich et al., 1991; Šali et al., 1994a, 1994b) represent maximal encoding, where the number of different letters in the alphabet can be as large as the chain length and considerably greater than 20, the number of amino acid types. Real proteins undoubtedly fall somewhere between these extremes. It is not known what percentage of all possible amino acid sequences fold to unique native states, although experimental methods that extensively sample sequences are becoming feasible (Kaiser et al., 1987; Reidhaar-Olson et al., 1991; Kamtekar et al., 1993; Vuilleumier & Mutter, 1993; Davidson & Sauer, 1994). It is valuable to find the minimal alphabet size required for fast and stable folding in order to learn how simpler polymers might be designed to fold like proteins.

Native states are not spheres: Their deviations from maximal compactness are importanf

Protein native structures are not perfect spheres (Goodsell & Olson, 1993). They are highly, but not maximally compact (see Fig. 11). Deviations from maximal compactness in global shape, surface cavities, and active sites are intrinsic to protein structure and function. To assume that hydrophobicity is the dominant force in protein folding is not to imply that native structures are spherical, or that all hydrophobic residues are fully buried, because chain connectivity is a complex constraint. Native states of H P model proteins are often not maximally compact. In the HP model, the shapes of native proteins depend on their monomer sequences. Native H P model proteins often have H monomers at the surface and sometimes have P monomers inside, as real proteins do (Lee & Richards, 1971).

Is side-chain packing a major part of the folding code?

The Protein Data Bank (Bernstein et al., 1977; Abola et al., 1987) shows that protein interiors are tightly packed (Richards, 1974, 1977; Richards & Lim, 1993; Harpaz et al., 1994). Moreover, one of the few ways that "designed" proteins do not yet

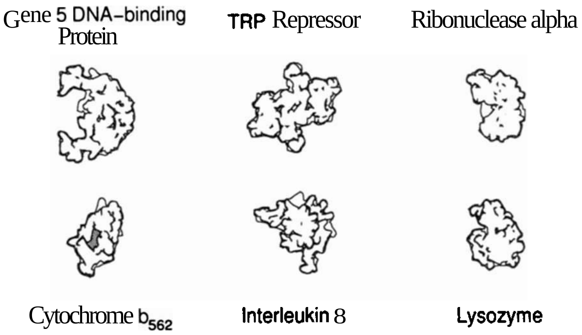

Fig. 11. Native protein structures are relatively but not maximally compact. Proteins are not spheres (Goodsell & Olson, 1993).

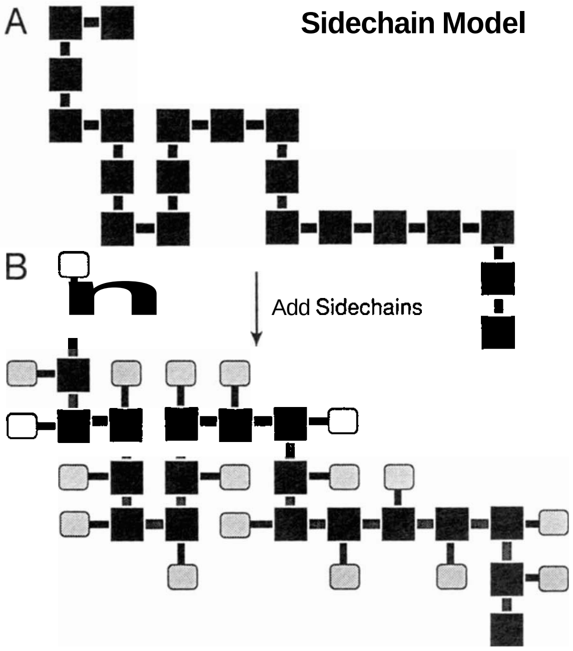

Fig. 12. Simple exact side-chain model. Taking the linear chain lattice model (LCM) to represent the main chain (A), a sidechain model (SCM) is created by attaching a single side-chain unit to each main-chain monomer. To represent side-chain rotameric degrees of freedom, each side-chain unit has the freedom to occupy any one empty lattice site adjacent to its corresponding main-chain monomer (B) (see Bromberg & Dill, 1994).

look like natural proteins is that they do not have the same tight side-chain packing (Betz et al., 1993), except when ligands like Zn^{2+} are added (Handel et al., 1993), suggesting a need to design precise fits into the packing of the side chains. Large cavity-creating perturbations are often destabilizing (Lim et al., 1992), and side-chain fits are important determinants of the structures of coiled coils (Harbury et al., 1993). Is the essence of protein structure and stability encoded in the microscopic details of jigsaw puzzle-like side-chain packing? If so, the single monomer representations of amino acids in simple exact models could not account for protein organization.

Other evidence suggests that side-chain packing is not the dominant component of the folding code: (1) Behe et al. (1991) found little preference of side chains to conjointly bury surface area. (2) Singh and Thornton (1990, 1992) found little preference of pairwise side-chain orientations among hydrophobic core residues. (3) Sosnick et al. (1994) have noted that the kinetic bottleneck to the folding of cytochrome c is probably not side-chain packing, because a single nonnative heme interaction slows folding by orders of magnitude. (4) Proteins show considerable structural tolerance for mutations that change side-chain size and shape (B.W. Matthews, 1987, 1993; Lim & Sauer, 1991). (5) Topologically similar proteins can have differently packed cores (Swindells & Thornton, 1993). (6) Proteins can maintain native topology in states that lack native-like packing (Hughson et al., 1990, 1991; Feng et al., 1994; Peng & Kim, 1994). (7) The fold of some proteins, such as globins, can be achieved by sequences that are less than 20% identical (Bashford et al., 1987).

What then, is the role of side chains in structure and stability?

We have extended the "string-of-beads" model to represent side chains simply as single pendant beads attached to each backbone bead (Bromberg & Dill, 1994; see Fig. 12). This is another simple exact model, intended to address questions of principle, not to accurately represent microscopic detail. This approach has two virtues. First, although higher resolution studies have computed rotational isomeric side-chain entropies neglecting excluded volume contributions (Pickett & Sternberg, 1993) and explored the effects of side-chain entropy on helix formation (Creamer & Rose, 1992), the simplified exact model has been the only way so far to study the significance of side-chain excluded volume. Second, the simplicity of this model allows us to explore the linkage between backbone and side-chain degrees of freedom (Baldwin et al., 1993; Richards & Lim, 1993), whereas earlier studies were caused by computational limits (Ponder & Richards, 1987; Lee & Subbiah, 1991) or theoretical premises (Shakhnovich & Finkelstein, 1989) to assume fixed backbone conformations.

The model studies show: (1) that side chains contribute a large excluded volume entropy that opposes folding, and (2) that side-chain and backbone degrees of freedom are strongly coupled. By exhaustive enumeration of short chains, and Monte Carlo sampling of chains up to 50 backbone monomers long in 2D and 3D, the excluded volume entropy contributed by the side chains has been determined as a function of backbone compactness (Bromberg & Dill, 1994; see Fig. 13). The results show that side chains "freeze," i.e., there is a steep loss of side-chain conformational entropy at the last stages of collapse to the native state. Coupling implies that if the chain is driven strongly enough to collapse, it will cause the side chains to freeze into place. These model results are consistent with the PNIPAM homopolymer collapse experiments of Binkert et al. (1991), showing that side-chain fluorescent labels have dramatically slowed motions at the collapse transition (see Fig. 14). The model also predicts that small expansions from the native state should lead to large increases in entropy (the opposite of rubber-like elasticity). This is consistent with experiments in which protein crystals that are mechanically stretched by 5% at T = 300 K are found to have TAS = +27 kcal mol^{-1} (Morozov & Morozova, 1993), although other entropy components might also be contributing.

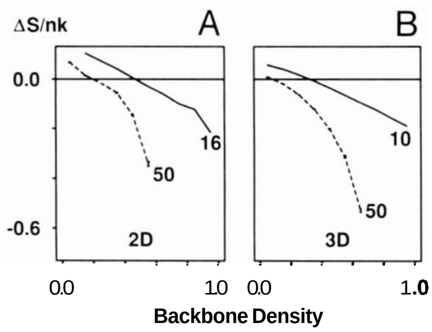

Fig. 13. Side-chain entropies depend on compactness. Excess entropies due to adding side chains, $\Delta S/(nk)$, where n is the chain length and k is Boltzmann's constant, versus backbone compactness ρ_l, in 2D (**A**) and 3D (R) by exhaustive enumeration for $n = 16$ in **2D**, and $n = 10$ in 3D, and by Monte Carlo (---) sampling for n = 50 in both (A) and **(B)** **(Bromberg** & Dill, **1994).** Increased slope at high densities is described as side-chain "freezing."

This model appears to be consistent with side-chain packing in proteins—cores can pack tightly, yet side chains have no packing preferences in their orientations or buried areas (see Fig. 15). But the protein-like aspects of side-chain packing that are achieved in this model do not arise from a complex jigsaw puzzle-like fit. The model side-chain packing is more random and nonspecific, more like nuts and bolts in a jar, a sort of ad hoc jumble, than like a jigsaw puzzle with precise pairwise shape complementarity between amino acids (see Fig. 15). A very lucid overview of types of packing and jigsaw-puzzle folding kinetics is given by Richards (1992). What are the implications of distinguishing a jigsaw-puzzle model of side chains from a nuts-and-bolts model? First, a jigsaw-puzzle model implies that if a native-like chain were systematically expanded, side chains would remain locked until a critical disjuncture point, estimated in one model (Shakhnovich & Finkelstein, 1989) to be around a 25% increase in volume. In contrast, the nuts-and-bolts model implies that different side chains will unfreeze at different expansions, but that, on average, most side-chain freedom will be gained upon expansion of only a few percent in volume. Second, because the nuts-and-bolts model predicts no step increase of side-chain entropy at relatively large chain expansion, it implies that side-chain packing is not the basis for the two-state cooperativity observed in proteins (see below).

To the extent that this uniform side-chain model is a reasonable first approximation to the variable side-chain sizes in proteins, it provides no basis for a folding code in which side-chain packing would somehow encode the difference between lysozyme and ribonuclease. Although native states may thus be destabilized by excluded volume side-chain entropies, they may be stabilized by energies of tight packing (Harpaz et al., 1994). But in order to contribute to the folding code, packing must differ strongly from one side chain to the next and be sequence dependent. Refined models are required to explore sequence-dependent packing differences, particularly for coiled coils, where steric details clearly play a role in structural differences (Harbury et al., 1993).

Tertiary structures can be encoded in minimally degenerate sequences

Protein tertiary structures can be remarkably symmetrical (Levitt & Chothia, 1976; Richardson, 1981; Branden & Tooze, **1991),** involving bundles of helices, stacks of β-**sheets,** or repeating α/β

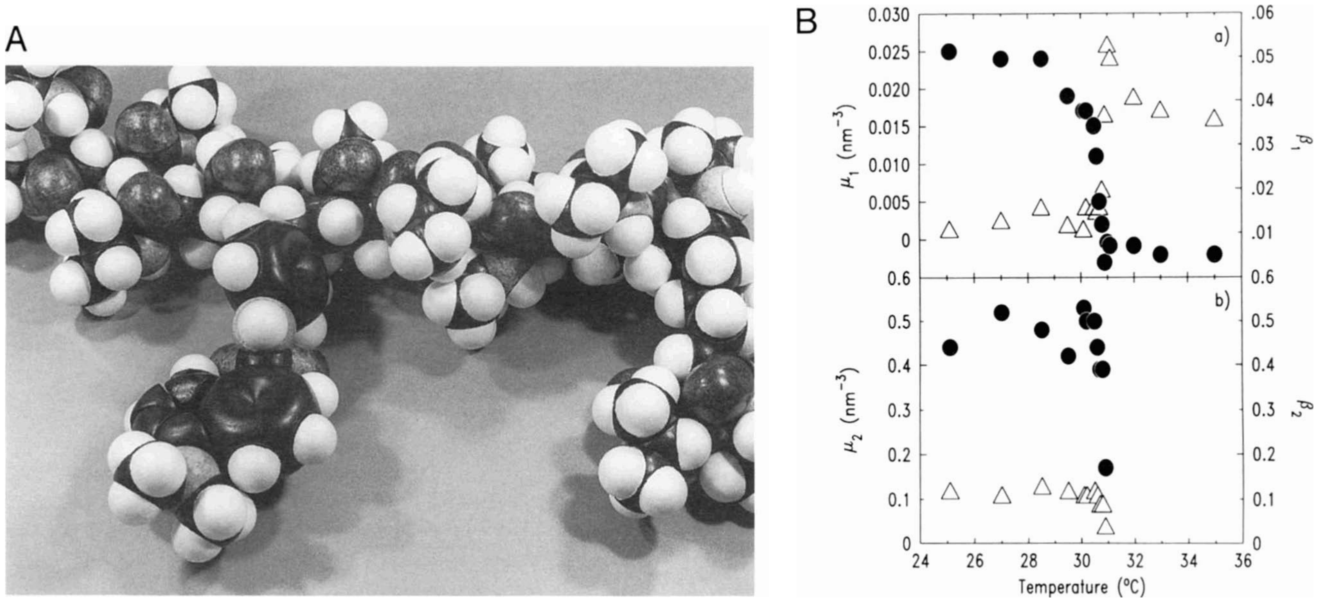

Fig. 14. Homopolymer collapse freezes out motions of side chains. Data of Binkert et al. (1991) using time-resolved measurements of fluorescence polarization anisotropy to monitor side-chain motions due to the temperature-induced conformational transition of PNIPAM in aqueous solutions. A: Space-filling model of a section of the polymer. B: Reduced reorientational relaxation rates μ_i and their amplitudes β_i versus temperature. a: μ_1 (left, ●) and β_1 (right, A), reflecting backbone motions. b: μ_2 (left, ●) and β_2 (right, A), reflecting mainly local side-chain motions. Both types of motions undergo a "freezing" transition at the same temperature around 31 °C. Chains have approximately 3,100 monomer units.

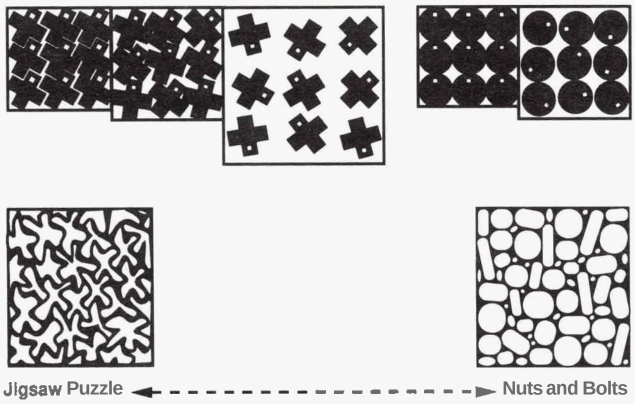

Fig. 15. Models for packing of side chains in proteins range from the jigsaw-puzzle type, requiring pairwise shape complementarity, to the nuts-and-bolts type, which is random like the jumble of nuts and bolts in a jar. Jigsaw-puzzle packing involves a point of critical disjuncture: as a protein expands in denaturation, the side chains gain no rotational freedom until they are separated by a critical distance. Nuts-and-bolts models gain rotational freedom as the protein first expands from its maximally compact state (see Bromberg & Dill, 1994).

units. What causes the symmetries? Unlike secondary structures, tertiary architectures are global properties. It is possible that tertiary structures arise as a sum of many small interactions. But then why is there so much global symmetry? Some pattern of intramolecular interactions must distributed broadly throughout the sequence. The HP model gives a simple explanation. Tertiary structure encoding has been explored using a new conformational search procedure, called Constrained Hydrophobic Core Construction (CHCC), applied to the HP model (Yue & Dill, 1993, 1995). This method finds globally optimal conformations with the maximum number of HH contacts. A certain small class of HP sequences has been found to produce native states with protein-like tertiary symmetries.

The HP sequences that produce protein-like tertiary structures are distinguished by their minimal degeneracy. HP sequences that have the fewest possible lowest energy states, the smallest values of g_N, have highly symmetrical protein-like native structures (Yue & Dill, 1995). We have found four-helix bundles, α/β-barrels, and parallel β-helices (Yoder et al., 1993) by seeking such minimally degenerate sequences with the CHCC procedure (see Fig. 16). In contrast to other models that produce protein-like

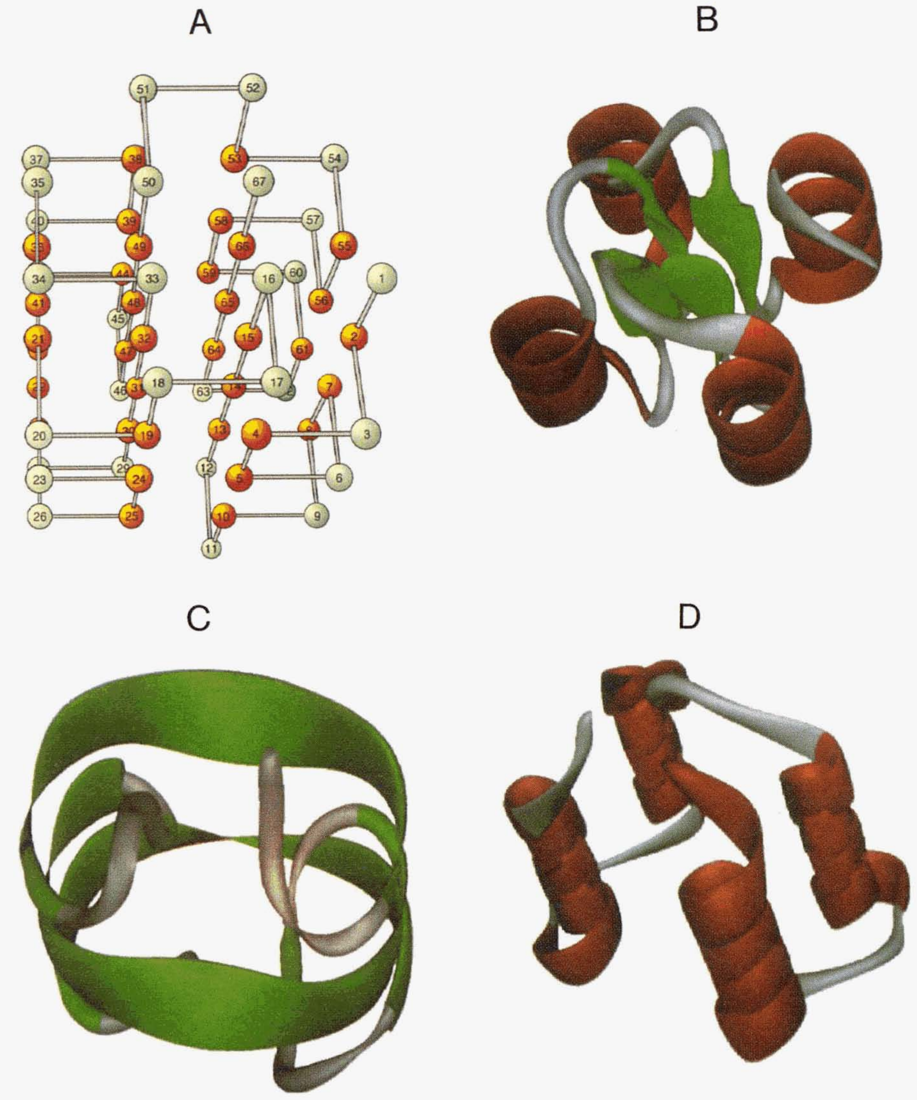

Fig. 16. Tertiary symmetries arise in the lattice model from finding the native states (maximum number of HH contacts) of HP sequences that have a minimal number of native states. A An HP lattice conformation resembling an α/β-barrel in real proteins. B: Same conformation as (A), but using ribbon diagrams. C, D: Ribbon diagrams for the 3D HP lattice model of a parallel P-helix and a 4-helix bundle, respectively, obtained in the same way (see Yue & Dill, 1995).

structures (Skolnick & Kolinski, 1991), this HP model study involves no parameters or energies. It just seeks conformations with a maximum number of HH contacts, from sequences that have a minimum number of native states. These model studies predict that the HP sequence is sufficient by itself to encode general tertiary architectures. In elegant experiments, Kamtekar et al. (1993) have engineered molecules that appear to fold to helix bundles using just an HP code. The degeneracies of their native states are not yet known.

Most surprisingly, the model predicts that the essence of the high symmetries in tertiary structures of native proteins goes beyond the relationship between a sequence and its native fold. High tertiary symmetries also depend on an implicit negative design: an encoding within the amino acid sequence of an ability not to fold to other conformations. In these few instances studied, the highest symmetries arise from sequences with the greatest degree of negative design. Such negative encoding has been studied so far only in the HP model, because it is the only model at present for which there is complete knowledge of the conformational and sequence spaces.

Sequence design: The hard part is uniqueness

Designing an amino acid sequence to fold to a desired ("target") conformation has two aspects: (1) positive design, ensuring that the sequence will fold to the target structure (reviewed in Richardson & Richardson, 1989), and (2) negative design (DeGrado et al., 1989; Hecht et al., 1990; Hill et al., 1990; Yue & Dill, 1992, 1995), ensuring that the sequence does not fold to stable alternative conformations. Designed proteins appear to have more conformational diversity than real native proteins (Betz et al., 1993; Handel et al., 1993; Sasaki & Lieberman, 1993; Tanaka et al., 1994), an indication that the negative design problem is not yet solved. It appears to be much easier to design into a sequence the ability to fold to a desired native structure (as one of several low-energy structures) than to design out an ability to fold to all of the other approximately 10^{68} incorrect structures (for a 100-mer).

We have used lattice models to study negative design (Yueet al., 1995). In a Harvard/UCSF collaboration, the Harvard group chose 10 different 3D 48-mer lattice target conformations. They designed HP sequences to fold to those structures by a Monte Carlo method without explicit negative design (Shakhnovich & Gutin, 1993a, 1993b; Shakhnovich, 1994). This method starts with random labels, H or P, painted onto each amino acid "bead." It then iteratively permutes the labels of the beads to reduce the energy. The negative design in this method was limited to maintaining a fixed monomer composition to avoid designing a homopolymer sequence, i.e., to avoid labeling all beads as H monomers. The UCSF group then used two different HP lattice conformational search strategies, the CHCC algorithm (Yue & Dill, 1993, 1995) and hydrophobic zippers (Dill et al., 1993; Fiebig & Dill, 1993) (see below), to seek the native conformation(~)of each sequence. The result was that the Monte Carlo procedure failed to adequately design HP sequences. For 9 of the 10 sequences designed by the Monte Carlo method, CHCC was able to find conformations of lower free energy than the target conformations to which they were designed to fold. Although the target structures were chosen to be maximally compact, the designed sequences invariably folded to more stable conformations that were not maximally compact (see Fig. 17). Thus, even when a sequence is designed to have an apparently good hydrophobic core, the molecule can usually fold to a structure with an even better hydrophobic core. This study indicates the importance of negative design for the HP model, and by inference, for real proteins. Design procedures without sufficient attention to negative design have also been found inadequate in another simple folding model based on contact and helical interactions (M. Ebeling & W. Nadler, submitted).

How can we eliminate conformational diversity when designing proteins? Handel et al. (1993) suggest that the lack of a unique structure arises from poor side-chain packing, and that more attention must be paid to designing cores that lock side chains in better steric fits. But conformational diversity can also arise from poor hydrophobic/polar sequence design. As noted above, a simple "hydrophobic inside, polar outside" rule is not an adequate design strategy (Yue & Dill, 1992). The difficulty encountered in hydrophobic/polar design (Shakhnovich, 1994) can be more reasonably ascribed to flaws in design strategy (Yue et al., 1995), rather than to the simplicity of the model. The fact that some HP sequences fold to multiple native states implies only that those sequences are not good folders. It does not imply that hydrophobicity is too nonspecific as a driving force to produce native structures. For example, any maximally compact conformation can be encoded by the sequence HHHH ... H. This sequence encodes all maximally compact structures and thus folds with great conformational diversity, so it would be a very poor design. Other sequences do fold to unique native

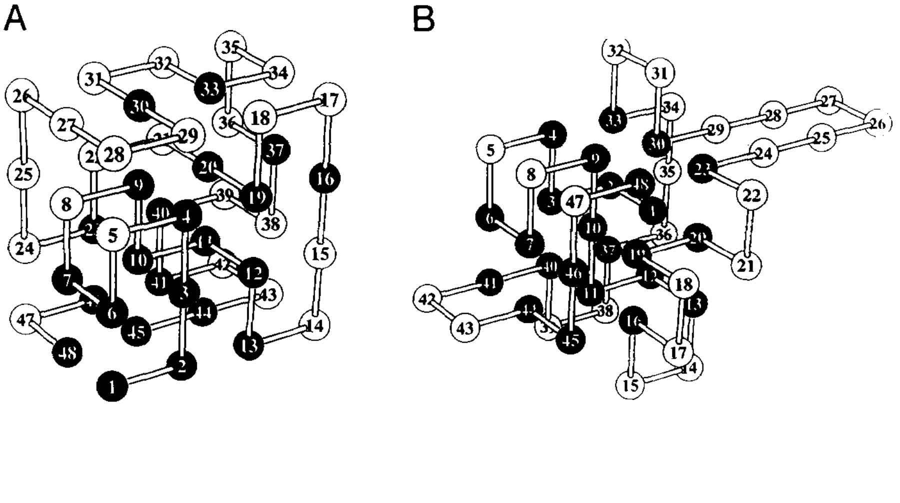

Fig. 17. A: Lattice model-designed protein and its HP sequence, designed by the Monte Carlo method of Shakhnovich and Gutin (1993a) (H: black bead; P: white bead), with limited "negative design." B: One of the many lower energy (globally optimal) structures of the same sequence was found by the CHCC conformational search method (Yue et al., 1995). True ground-state conformation (right) is not maximally compact. This indicates the importance of negative design in the HP model.

states, so problems with this one sequence, or any other particular sequence, do not imply that hydrophobicity is too nonspecific. Within the H P model there are a few good sequences and many bad sequences (Yue & Dill, 1995). Finding sequences with low degeneracy in the H P model requires a more sophisticated design strategy than just finding sequences that encode good, low-energy, hydrophobic cores (Yue & Dill, 1992, 1993, 1995; Yue et al., 1995), although perhaps such simple design strategies may be more successful with certain sets of larger monomer alphabets (Shakhnovich, 1994). Regarding the design of side-chain packing, it is interesting that the collapse of the homopolymer PNIPAM, which has side chains not too different from amino acids in size and conformational freedom, leads to much slower motion, if not the freezing, of side chains (Binkert et al., 1991), without the need to design "steric fit."

Adding hydrophobic monomers to stabilize native states can be a poor design strategy

H P model studies show, contrary to naive expectations, that attempting to stabilize a protein by adding extra hydrophobic contacts in the target structure can increase conformational diversity (Chan & Dill, 1991b; Yue & Dill, 1992). This can lead to thermodynamic instability of the single target structure, due to more stable denatured conformations. The best protein designs do not seek to maximize favorable native interactions but to minimize excess stabilizing interactions—as long as there is enough stability to hold the protein together. Any excess H monomers beyond those required to stabilize the desired native state as the lowest-energy structure increase the possibility that the chain will fold to alternative low-energy conformations (Chan & Dill, 1991b; Yue & Dill, 1992). Model studies noted above show that molecules designed to have apparently good hydrophobic cores can generally also have many equally good alternative conformations (Yue et al., 1995). Owing to constraints imposed by chain connectivity and intrachain interactions, well-designed proteins might at best only have marginal stability. Protein stability and genetic engineering experiments are consistent with this view: (1) real proteins are marginally stable (around 5–10 kcal/mol-protein, or about 100 cal/mol-amino acid), and (2) proteins designed to have a large number of favorable interactions have conformational diversity (Regan & DeGrado, 1988; Handel et al., 1993).

Not all proteins fold to unique native structures

In simple models, most of the possible sequences do not fold to unique states (Lau & Dill, 1989; Honeycutt & Thirumalai, 1990, 1992; Chan & Dill, 1991b). To achieve a unique fold requires some sequence selection. Are the native states of natural proteins unique? We regard any native structure as "unique" if it can be fully resolved in X-ray crystallography or NMR experiments, and we neglect small fluctuations and dynamic motions. (For discussion of these smaller motions, see Frauenfelder et al. [1991] and Straub and Thirumalai [1993].) Native proteins are often not unique, even by this low-resolution definition: where loops are unresolved in crystal or NMR structures, this indicates conformational diversity (Faber & Matthews, 1990; Frauenfelder et al., 1990; Chacko & Phillips, 1992; Engh et al., 1993; Shirakawa et al., 1993). In proteins such as insulin (Hua et al., 1992, 1993), a state with conformational diversity may be the functional form, rather than a unique native state. Thus, not all natural proteins are necessarily strongly selected to fold to unique conformations.

Mutational and evolutionary change

Like real proteins, the H P model responds to mutational and evolutionary change (Lau & Dill, 1990; Chan & Dill, 1991b; Lipman & Wilbur, 1991; Shortle et al., 1992; Chan & Dill, 1994). (1) For a considerable fraction of amino acid sequences, the native structures of H P proteins are tolerant to mutation, like real proteins (Bowie et al., 1990; Lim & Sauer, 1991; Heinz et al., 1992; B.W. Matthews, 1993), in that the mutant chain folds to the same native fold as the wild type. (2) The core is more highly conserved than the surface, i.e., mutations are more readily tolerated at the surface than in the nonpolar core, consistent with experiments of Reidhaar-Olson and Sauer (1988), Lim and Sauer (1991), and B.W. Matthews (1993). This implies a greater role for nonpolar interactions in driving folding because the hydrophobicity of amino acids measured from transfer experiments correlates with degree of burial in protein structures (Rose et al., 1985; Lawrence et al., 1987; Miller et al., 1987). (3) Convergence, the encoding of a given native structure by different sequences, is observed in the H P model (Lau & Dill, 1990; Chan & Dill, 1991b).

(4) Lipman and Wilbur (1991) have shown that the evolutionary fitness landscape, modeled with the 2D H P model, has a "connectedness" property. A sequence is considered to be functional if its native state has, as a "phenotype," a single contact map. A mutation is "nonlethal" when the mutant is functional, and "lethal" otherwise. Lipman and Wilbur found that there are large evolutionary networks linked by nonlethal mutational steps (H → P or P → H), satisfying a critical requirement of evolutionary space proposed by Maynard Smith (1970), viz., "if evolution by natural selection is to occur, functional proteins must form a continuous network which can be traversed by unit mutational steps without passing through non-functional intermediates." They also found that neutral mutations that do not change the phenotype are necessary for traversing the evolutionary networks, implying that "neutral mutations can act as a significant constraint on positive selection" (Lipman & Wilbur, 1991).

Protein folding thermodynamics

Folding is cooperative

What do simple exact models tell us about the thermodynamics of protein folding? Here we explore (1) the basis for folding cooperativity, (2) the absorption of heat in protein transformations, and (3) the nature of one-state and two-state transitions. We start by introducing a Tetramer Toy Model (TTM) of folding' to give a simple picture of the physical basis for the cooperativity and heat absorption of folding. We use the TTM to illustrate the meaning of an energy ladder or spectrum, the density of states, and a stability or energy gap.

The TTM is a 2D square lattice model of a four-monomer chain that has two H monomers at the ends and two P monomers in the middle (see Fig. 18). This short chain has only five

[5] With thanks to Walter Englander for motivating it.

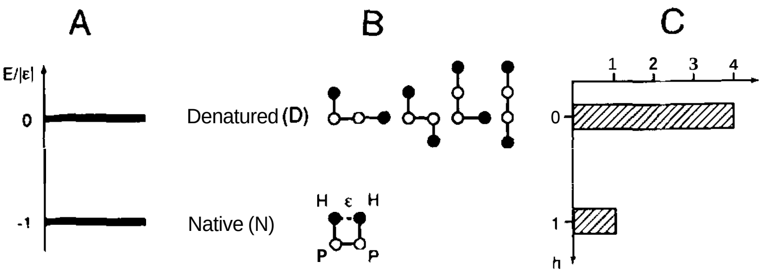

Fig. 18. Tetramer toy model (TTM). **A:** Two energy levels. B: One native and four denatured conformations. *C:* Density of states $g(h)$

possible conformations: one conformation, the "native" state, has $h = 1$ HH contact. The other four conformations have $h = 0$ HH contacts; collectively they are the "denatured" state (Fig. 18). The native state has lower energy by virtue of its one HH "bond" or contact. The most important quantity is the density of states, $g(h)$, the number of conformations as a function of the number of of HH contacts, $h = 0, 1, \ldots, h_N$, where h_N is the maximum number of HH contacts (Chan et al., 1992; Shortle et al., 1992; Stolorz, 1994). For this toy model, the number of native conformations is $g_N = g(h_N) = g(1) = 1$, and the number of denatured conformations is $g(0) = 4$.

The two states, native and denatured, can be represented by an energy level diagram (Fig. 18). The energy of each denatured conformation is higher than the native state by an amount $-\epsilon$, where $\epsilon < 0$, which represents the breaking of the HH contact "bond." We are not concerned for now with the subtle aspects of hydrophobic interactions: ϵ simply represents a favorable contact free energy. Rather, we simply take as an experimental fact that nonpolar association in water is favorable and its free energy is nearly independent of temperature over the wide range 0–100 °C (Privalov & Gill, 1988). That is, a first treatment recognizes that oil and water do not mix. A second treatment would go beyond this to recognize that the basis for the positive free energy of oil/water association is a large heat capacity, a negative entropy near room temperature, and a positive enthalpy at higher temperatures. It is common practice in these types of models to work at this first level of treatment and to simply regard ϵ as an "energy," and neglect the fact that it is more correctly a free energy. We follow that spirit here. What this treatment will miss is cold denaturation. An example of the second approach, treating the temperature dependence, is given by Dill et al. (1989).

Now we use statistical mechanics to compute the properties of this simple exact model. We require the partition function, Q, which is the sum of Boltzmann factors over all the conformational states:

$$Q = \sum_{h=0}^{h_N} g(h) e^{-h\epsilon/(kT)} = 4 + e^{-\epsilon/(kT)}, \qquad (1)$$

where k is the Boltzmann constant and T is absolute temperature. The probability, $P_N(T)$,that the chain is in its native state is defined by:

$$P_N(T) = \frac{e^{-h_N\epsilon/(kT)}}{Q} = \frac{e^{-\epsilon/(kT)}}{4 + e^{-\epsilon/(kT)}}, \qquad (2)$$

and the probability that the protein is in the denatured state $P_D(T)$ is given by:

$$P_D(T) = 1 - P_N(T) = \frac{4}{Q}. \qquad (3)$$

Figure 19 shows the sigmoidal thermal denaturation profile predicted by Equation 2. If we define cooperativity as a sigmoidal transition, then this model has cooperativity. (A more subtle distinction is whether cooperativity is one state or two state; see below.) At low temperatures, the native state (N) is stable but at high temperatures the four denatured conformations are more populated. We can express the denaturation in terms of the free energy of folding,

$$\Delta G_{fold} = -kT \ln\left(\frac{P_N}{P_D}\right) = \epsilon + kT \ln 4. \qquad (4)$$

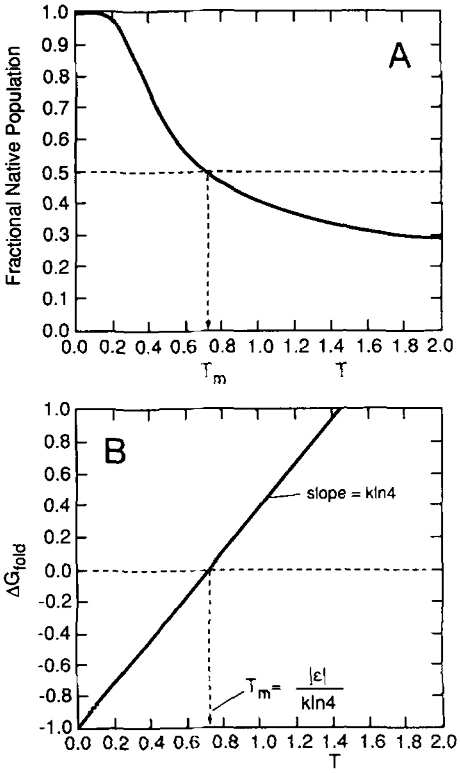

Fig. 19. Denaturation in TTM. Fractional native population **(A)** and free energy of folding ΔG_{fold} (B). Native state is stable at low temperatures. Protein denatures with increasing temperature. Absolute temperature T is in units of $|\epsilon|/k$, where k is Boltzmann's constant. ΔG_{fold} is in units of $|\epsilon|$. T_m is the mid-point temperature at which half of the chain population is native.

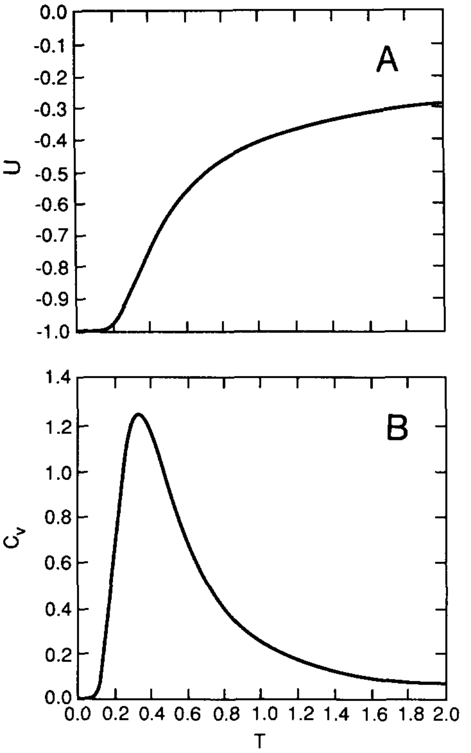

Fig. 20. Thermodynamics of TTM. Contact (free) energy E (A) and heat capacity C_V (B) as functions of temperature T, for a temperature-independent ϵ. Units for T, U, and C_V are, respectively, $|\epsilon|/k$, $|\epsilon|$, and k.

Hence, assuming ϵ is independent of temperature, and neglecting pressure–volume effects, which are normally small, the folding enthalpy and entropy are $\Delta H_{fold} = \epsilon$, and $\Delta S_{fold} = -k \ln 4$, respectively. Figure 19 shows ΔG_{fold} versus T. The midpoint temperature for denaturation is identified by $\Delta G_{fold} = 0$, so that $T_m = -\epsilon/(k \ln 4)$ in this model. (The native population never reaches zero in this model because the chain is so short that the native state competes with only four denatured configurations; the model becomes more realistic for longer chains.)

The definition of the total energy per molecule is

$$U \equiv \langle \epsilon h \rangle = \frac{1}{Q} \sum_{h=0}^{h_N} g(h) \epsilon h e^{-h\epsilon/(kT)} = \frac{\epsilon e^{-\epsilon/(kT)}}{4 + e^{-\epsilon/(kT)}}, \quad (5)$$

where $\langle \ldots \rangle$ denotes the average over all states. The specific heat is the derivative

$$C_V \equiv \left(\frac{\partial U}{\partial T}\right)_V = \frac{\epsilon^2}{kT^2} (\langle h^2 \rangle - \langle h \rangle^2) = \frac{\epsilon^2}{kT^2} \frac{4e^{-\epsilon/(kT)}}{[4 + e^{-\epsilon/(kT)}]^2}. \quad (6)$$

Figure 20 shows that this model predicts a peak of heat absorption upon denaturation. The heat absorption peak reflects the increased energy upon breaking the native noncovalent HH "bond." At low temperatures, a small amount of heat will not be absorbed because it is not sufficient to break the HH contact. At intermediate temperatures, heat is absorbed to break the HH contact and denature the protein. At high temperatures, the molecule is already fully denatured so no further heat can be absorbed to break additional HH contacts.

This is a toy model. But it shows that the cooperativity of protein folding can be captured simply and need not arise from coupled interactions. A sigmoidal transition can be as simple as the breaking of noncovalent contacts. Protein folding cooperativity could have many origins—in hydrogen bonding, hydrophobic interactions, in electrostatic interactions, in side-chain packing, or in combinations of these. To be more protein-like, models should treat longer chains, sharpening the cooperativity due to a better hydrophobic core (see below), and include the temperature dependence of the hydrophobic interaction to represent enthalpic and entropic components more accurately. Heat capacities of unfolding proteins are large, indicating that a single hydrophobic HH bond in the model may arise from a change in multiple hydrogen bonds in the solvent.

To get more insight into the complexity of the denatured state and the denaturation transition, we now consider a slightly better model, the Hexamer Toy Model (HTM), with three energy levels. Figure 21 shows the conformations, energy diagram, and density of states function $g(h)$ for the 6-mer HTM sequence. As in both the TTM and HTM, an important property of real proteins is that g generally increases as h decreases from h_N:

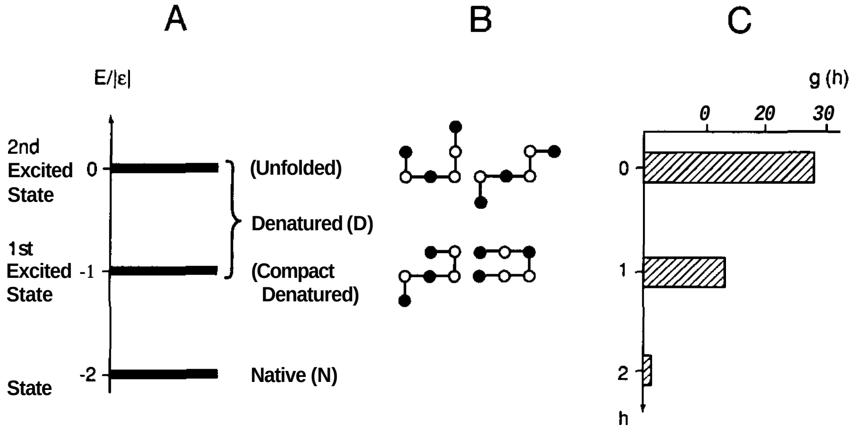

Fig. 21. Hexamer toy model (HTM). A: Three energy levels. B: Native and two representatives each of first excited (compact denatured) state and second excited (unfolded) state. C: Density of states $g(h)$.

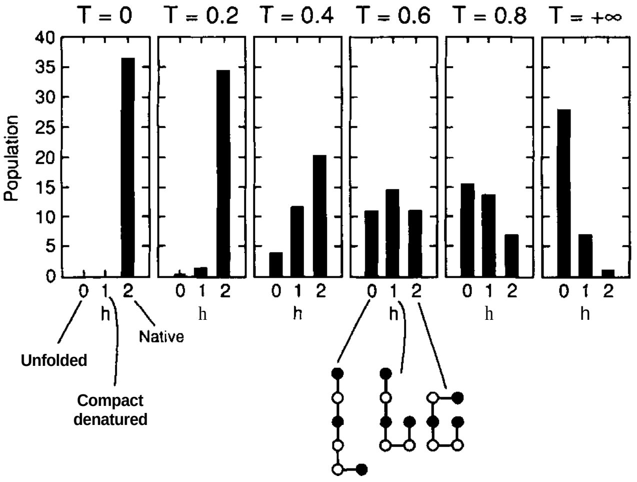

Fig. 22. Populations of states versus temperature (Tin units of $|\epsilon|/k$) in HTM. $T = +\infty$ distribution corresponds to $g(h)$. This model has "intermediate" states (see T = 0.6); hence, it is a one-state transition.

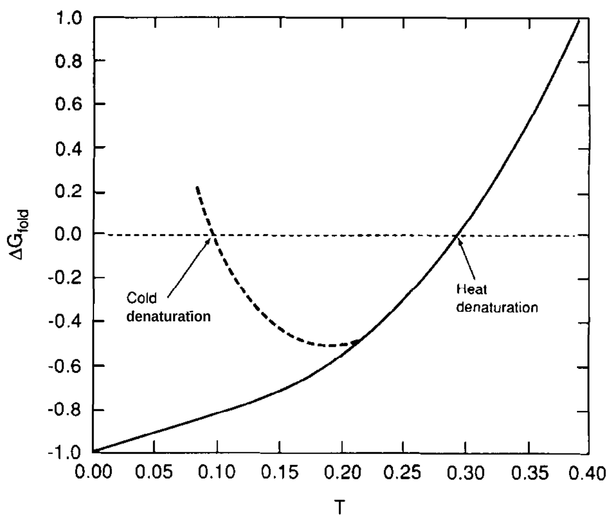

Fig. 23. Free energy of folding ΔG_{fold} (in units of $|\epsilon|$) versus temperature T (in units of $|\epsilon|/k$) for HP sequence studied by Gupta and Hall (1995), where ϵ is a constant (continuous curve). Sequence is shown in its unique native structure in Figure **24C** (ii). Curvature in ΔG_{fold} is caused by shifting subpopulations in the denatured ensemble, as in HTM. Dashed curve indicates that cold denaturation would occur if $|\epsilon|$ decreases with decreasing temperature.

there are more denatured (high-energy) conformations than native (low-energy) conformations. In analogy with quantum mechanical energy level diagrams, the "ground state" is the native conformation (lowest energy level), the "first excited state" (see Fig. 21) (next energy level) is a "compact" denatured state in the HTM, and the "second excited state" is an expanded or unfolded state (highest energy level in this model). Now we can explore the balance of native, compact denatured, and unfolded states.

When the temperature changes, it shifts not only the balance between native and denatured states, but also the distribution of subpopulations of the denatured state (see Fig. 22). At low temperatures (native conditions), the main denatured species is compact (because those conformations have more HH bonds), and at high temperatures, the main denatured species is expanded (because the larger number of expanded conformations leads to greater entropy). This causes curvature of the folding free energy versus temperature. At low temperature, stability is mainly the difference between the native state (2 HH contacts) and the compact denatured state (1 HH contact), a difference of 1 HH contact. At high temperature, the difference is 2 HH contacts. Figure 23 shows this curvature for a 2D HP 20-mer. Thus, protein stability under native conditions would be considerably overestimated by assuming the native structure is in equilibrium with a fully exposed denatured state (Shortle et al., 1992). This shifting compactness, entropy, and free energy of the denatured state with temperature is not an artifact of the simplicity of this toy model; it should arise in any statistical mechanical theory that takes chain connectivity into account.

Like the TTM and HTM, longer chain models using Monte Carlo sampling (O'Toole & Panagiotopoulos, 1992; Socci & Onuchic, 1994) and exact enumerations (Chan et al., 1992; Shortle et al., 1992; Chan & Dill, 1994; Gupta & Hall, 1995; Chan et al., 1995) show sigmoidal thermal transitions, specific heat absorption (see Fig. **24)**, and a shift in denatured state populations. The slope of the sigmoidal transition is steeper for some sequences than for others, corresponding to sharper peaks in specific heat, depending on their densities of states $g(h)$ (see Table 1). Note that the sequence with a broad transition in Figure 24A shows two peaks in its specific heat (see Fig. **24B)**, similar to the transitions of proteins with more than one melting domain (Privalov, 1982; Brandts et al., 1989). Because $\epsilon/(kT)$ is a measure of the strength of the HH interaction, which can be varied for real proteins by denaturants such as guanidinium hydrochloride or urea, these simple models can also be readily adapted to explore denaturant effects in the same way they explore thermal effects (Thomas &Dill, 1993; see also Alonso & Dill, 1991).

Cooperativity can be divided into one-state or two-state transitions (Dill & Shortle, 1991). Single domain proteins are often understood to fold with two-state cooperativity (Privalov, 1979). What type of cooperativity arises from these models? We return to the HTM model. Figure 22 shows the populations of the native, first excited, and second excited states versus temperature. One-state behavior means that the distribution over all states has only a single peak, as in Figure 22. Two-state behavior means that near the denaturation midpoint (of temperature in this example) the distribution of states will have two peaks, indicating two predominant "states," to use thermodynamic terminology (see Fig. 25). Each thermodynamic state corresponds to an ensemble of different microscopic chain configurations. "Two-state" behavior implies a free energy barrier between the two states. It is not the sharpness of a cooperative transition that distinguishes one-state from two-state behavior, but the number of identifiable populations. Very sharp one-state homopolymer collapse transitions (1–2 °C widths) are observed in PNIPAM (Tiktopulo et al., 1994; see Fig. 3). It is also important to note that two-state behavior does not imply that there is a single imperturbable denatured state unaffected by temperature or solvent conditions. Population shifts in the denatured state are predicted to occur in proteins with two-state behavior (Alonso et al., 1991; Dill & Stigter, 1995). For many HP sequences of chain lengths 18 or less in the 2D model, folding cooperativity

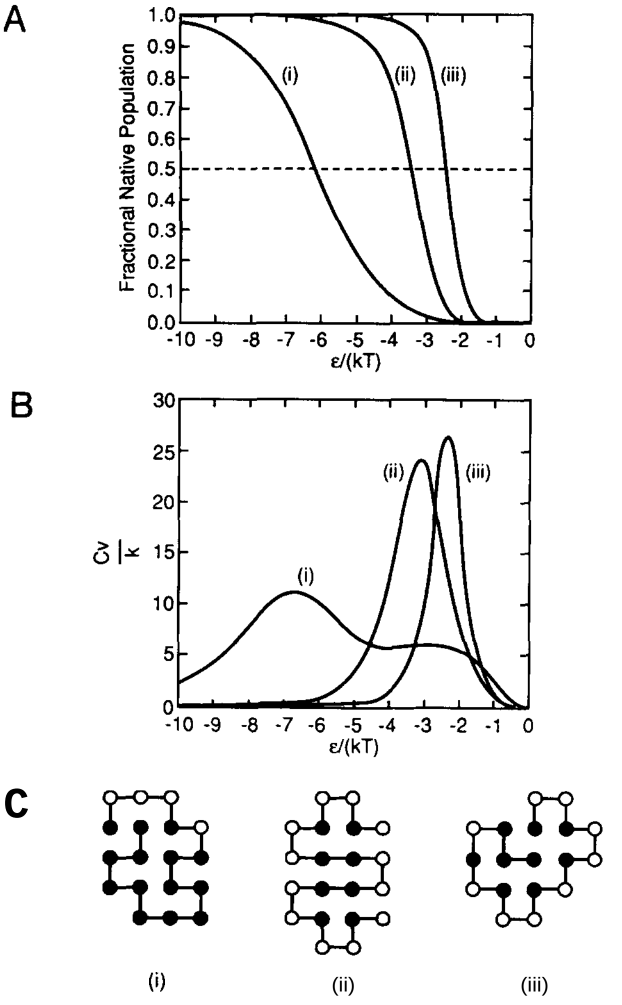

Fig. 24. **A:** Denaturation curves for three HP sequences on 2D square lattices. Their $g(h)$s are given in Table 1. B: Corresponding heat capacity C_V in units of k. C_V curves have sharper and higher peaks for sequences with sharper transitions (ii) and (iii). Sequence (i), with a broad transition, shows a double hump in its C_V plot (see text). C: Three sequences are shown in their unique native structures.

is one state, but Figure 25 shows an example of a sequence with two-state cooperativity. It is not clear what fraction of real amino acid sequences have two-state behavior, nor is it clear what fraction of long-chain HP model sequences have two-state behavior.

Within HP models, one-state behavior can be distinguished from two-state behavior by the shape of the $g(h)$ function. Two-state behavior requires that $-\ln g(h)$ versus h is concave upward near native energies h $= h_N$. When $g(h) = 0$ $(-\ln g(h) = \infty)$ for h $= h_N - 1, h_N - 2, \ldots, h_N - j$ $(j \geq 1)$, it has been called an "energy gap" (Chan & Dill, 1994; Šali et **al., 1994a, 1994b**; Shakhnovich, 1994). An example HP sequence with an energy gap is sequence (iii) in Figure 24A and B. The $g(h)$ and sole native structure of this sequence are given in Table 1 and Figure **24C,** respectively. Systems with energy gaps have been described by Guo et al. (1992) for an off-lattice model (Honeycutt & Thirumalai, 1990, 1992) and by Shakhnovich and **Gutin (1990b)** and **Šali** et al. **(1994a, 1994b)** for lattice models restricted to maximally compact conformations. Experimentally, one-state behavior can be distinguished from two-state behavior by determining the distribution of chain conformations and determining whether the distribution has two identifiable populations or only one. Transport methods, such as size exclusion chromatography (Uversky, **1993),** can be particularly useful for resolving slowly exchanging populations.

Protein folding cooperativity: A simplest hypothesis

The basis for protein folding cooperativity is not yet known. Many different models and types of interactions could lead to cooperativity. What is the simplest model for the two-state nature of protein folding? Helix–coil processes are "less cooperative" one-state transitions (see the discussion of one-dimensional Ising models in Stanley, 1987). Do homopolymers collapse with two-state transitions? This has been a matter of contention **(Ptitsyn** et **al.,** 1968; de Gennes, 1975; Post & Zimm, 1979; Sanchez, 1979; Grosberg & Khokhlov, 1987). Although Ptitsyn et **al.** (1968) argued that homopolymer collapse should be two-state in the limit of infinite chain length, it now appears that the collapse of a flexible homopolymer chain of finite length is a **one**-state transition (Sun et **al.,** 1980; Tiktopulo **et al., 1994),** unless chain stiffness is high, as in DNA (de Gennes, 1975; Post & Zimm, 1979; reviewed by Chan & Dill, **1991a, 1993a).** In this regard the collapse of flexible homopolymers is less cooperative than the two-state folding attributed to small globular proteins. Interestingly, a de novo design of an α/β protein shows that even when an amino acid sequence folds with two-state thermodynamics, as indicated by the equality of van't Hoff and calorimetric enthalpies, it does not imply that the sequence folds to a unique native state **(Tanaka** et al., 1994).

It has been proposed (Dill, 1985) that two-state protein folding cooperativity could arise simply because certain HP copolymer sequences can collapse to states that are not only compact, but also have good hydrophobic cores (reviewed in Chan & Dill, **1991a;** Dill & Stigter, 1995; Chan et al., 1995). The two-state nature of folding cooperativity was attributed to the ability of a sequence to partition its monomers into a folded structure consisting of a mostly hydrophobic core and a mostly polar surface. Homopolymers do not have this freedom. The hypothesis that two-state cooperativity can arise in such a simple model now has rigorous confirmation in the exact model results shown in Figure 25. Exact models show, however, that two-state behavior is a property of only selected HP sequences and would not be observed in random heteropolymers.

But based on a different assumption, namely that proteins resemble random heteropolymers, for which collapse is not two-state (Grosberg & Shakhnovich, **1986),** Shakhnovich and Finkelstein (1989) instead sought the basis for the two-state cooperativity of protein folding in side-chain packing (reviewed by Karplus & Shakhnovich, 1992). Their model led to the idea that compact denatured states are separated from native states by two-state transitions in which the side chains unfreeze, whereas the backbone remains native-like. We refer to this as the "side-chain molten globule" model (Ptitsyn, 1987; **Shakhnovich** & Finkelstein, 1989) to distinguish it from the term "molten globule," which is now commonly taken as an operational definition of a broad class of experimentally observed compact denatured states.

The following evidence argues against the side-chain molten globule model of compact denatured states. First, side chains

Table 1. ***Density of states g(h) of the sequences (i), (ii), and (iii) shown in Figure 24***[a]

	$g(h)$									
Sequence	h = O	h = 1	h = 2	h = 3	$h = 4$	h = 5	h = 6	h = 7	h = 8	h = 9
(i)	1,332,266	1,740,324	1,359,214	789,070	380,601	152,773	46,905	6,714	467	1
(ii)	21,146,335	15,348,238	4,526,737	779,973	82,065	5,766	457	6	1	–
(iii)	2,815,469	2,100,897	706,075	156,218	25,761	3,530	344	40	0	1

[a] Sequence (i) is a one-state sequence. Two-state sequences satisfy the condition that $-\ln g(h)$ concaves upward near the native h = h_N, i.e., $-\mathrm{d}^Z \ln g(h)/(dh^2)|_{h=h_N} < 0$ if g is a continuous function of h, and $-\ln g(h_N) + \ln g(h_N - 1) < -\ln g(h_N - 1) + \ln g(h_N - 2)$ if g is defined only for discrete h's, as in the lattice HP model. Because $\ln g(h_N) = 0$ in these examples, this condition is equivalent to requiring $g(h_N - 2) > g(h_N - 1)^2$. Sequences (ii) and (iii) are two-state sequences.

also "freeze" upon collapse of PNIPAM homopolymers (Binkert et al., 1991), but this does not result in two-state behavior (Tiktopulo et al., 1994), indicating that side-chain freedom is not the origin of two-state behavior, at least in PNIPAM. Second, small-angle X-ray scattering experiments described below indicate much broader conformational diversity of backbones in compact denatured states than is expected from the native-like backbones of the side-chain molten globule model (summarized in Lattman et al., 1994). Third, the side chain of cysteine 166 in the compact denatured state of β-lactamase is nearly as restricted as in the native state (Calciano et al., 1993). Fourth, there is evidence that compact denatured states are not a single backbone conformation, with fixed secondary structures, but are ensembles that vary with external conditions (for a comprehensive review, see Fink, 1995). For example, Figure 26 shows that varying [KCI] at pH 2 in the compact denatured state of β-lactamase can change the helix content over a wide range. Other examples are shown in Figure 8; see also Seshadri et al. (1994). Fifth, for at least some compact denatured states, electrostatics plays an essential role, because those compact states are observed at low pH as a function of salt concentration (Goto & Fink, 1990). We believe this results from a combined balance of hydrophobic and electrostatic interactions (Stigter et al., 1991). Hence, we believe that a simplest model for the two-state

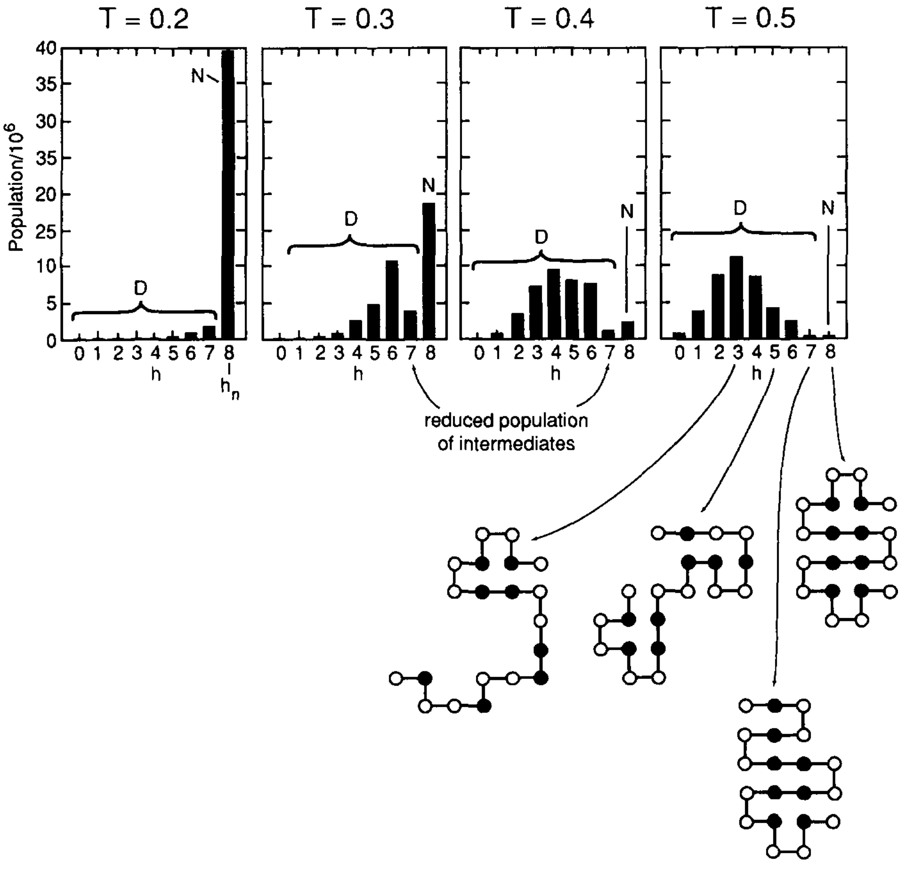

Fig. 25. Two-state thermal behavior of HP sequence in Figure 24C (ii), by exact enumeration. Populations of states versus temperature (T in units of $|\epsilon|/k$) show two population peaks corresponding to the native (N) and the denatured (D) states around the transition temperature (T ≈ 0.3–0.4); "intermediates" are less populated.

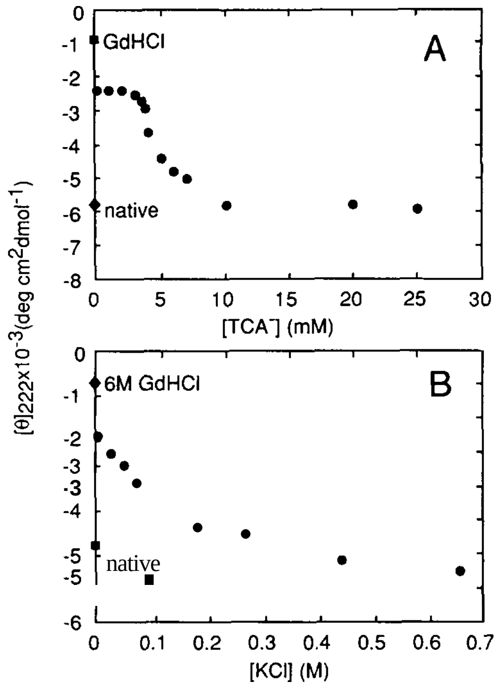

Fig. 26. Molten globule is not a single state; it can change with conditions. Anion-induced transitions of acid-unfolded β-lactamase to the compact denatured ("A") state at pH = 2.0. Anions were trichloroacetate (TCA⁻, A) and chloride (B). Transition was followed by ellipticity at 222 nm; data from Calciano et al. (1993). At low KCl, secondary structure content can be varied continuously in this compact denatured state.

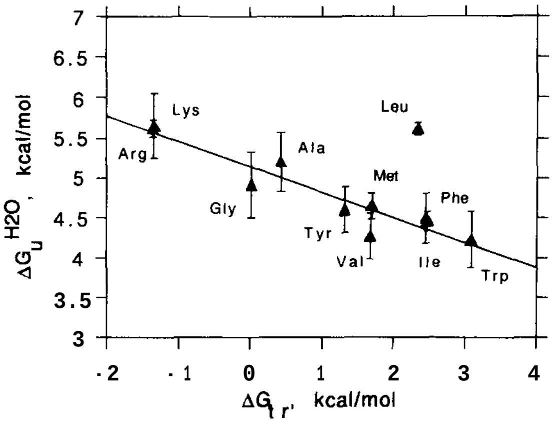

Fig. 27. "Reverse hydrophobic effect," interpreted as mutations that affect the denatured state more strongly than the native state. Plot shows the extrapolated free energy of guanidine-HCl unfolding $\Delta G_u^{H_2O}$, for 10 mutants of iso-1-cytochrome c at position 73 versus n-octanol to water transfer free energies, AG,, (Fauchère & Pliska, 1983). Position 73 is lysine in the wild-type protein (Bowler et al., 1993; L. Herrmann, B.E. Bowler, A. Dong, & W.S. Caughey, in prep., reproduced with permission).

cooperativity of protein folding is the encoding of good hydrophobic cores in HP sequences, rather than specific side-chain packing. We believe compact denatured states are different than predicted by the side-chain molten globule model, as we describe below.

The "structures" of denatured states

Denatured states of proteins are often compact and complex

What are the denatured conformations of proteins? HP lattice models predict that denatured states are broad ensembles of conformations that respond to changes in external conditions (Dill & Shortle, 1991; Shortle et al., 1992). There is no single denatured state. In strongly denaturing conditions, the most populated denatured species are highly unfolded. In native conditions, the most populated denatured species are compact (Ptitsyn, 1987, 1992; Dill & Shortle, 1991). The compact denatured states have some structure that is sequence dependent and native-like.

One indication of the complexity of the denatured state is the "reverse hydrophobic effect" (Pakula & Sauer, 1990; Bowler et al., 1993; L. Herrmann, B.E. Bowler, A. Dong, & W.S. Caughey, in prep.), whereby some replacements of polar by hydrophobic residues at the protein surface destabilize the folded state. This would seem to be the reverse of what is expected if hydrophobic forces fold proteins. Figure 27 shows the experimental evidence of Bowler et al. According to the HP model, those surface residues play a "reverse" role: they do not form hydrophobic contacts in the native state, but they do form hydrophobic contacts in the significant (i.e., low-energy) compact denatured conformations, hence, they destabilize the native protein (see Fig. 28). Their replacement by P monomers leads to stabilization of the native state. The HP model predicts that, at high H compositions, adding more H monomers often destabilizes, or only minimally stabilizes native proteins (Shortle et al., 1992).

A remarkable observation is that denatured states can be altered by single site mutations (Shortle & Meeker, 1986; Shortle et al., 1990; Flanagan et al., 1993). The HP model predicts that these mutations are at crucial sites in the small ensemble of the most important compact denatured conformations; a mutation at those positions changes the conformations of the relatively small number of dominant compact nonnative states. These mutations can also change the numbers of dominant denatured conformations and thus affect the conformational entropies of the denatured states. An experimental test (Fig. 29) shows how the denaturation slope, m, the change in stability with change in denaturant concentration, can be altered by mutation. This distribution is much wider than would be expected if denatured states were insensitive to mutations. The figure shows that the 2D HP model predicts a distribution similar in shape and width to the experimental observation on staphylococcal nuclease (Shortle et al., 1992).

Compact denatured states are broad ensembles of backbone conformations

Where is the disorder in compact denatured states? The side-chain molten globule model (Ptitsyn, 1987; Shakhnovich & Finkelstein, 1989) holds that the disorder in compact denatured states is in the side chains, whereas the backbone has a native-like structure. But simple exact lattice model studies predict that side-chain degrees of freedom are coupled to those of the backbone, and that compact denatured states have disorder in both

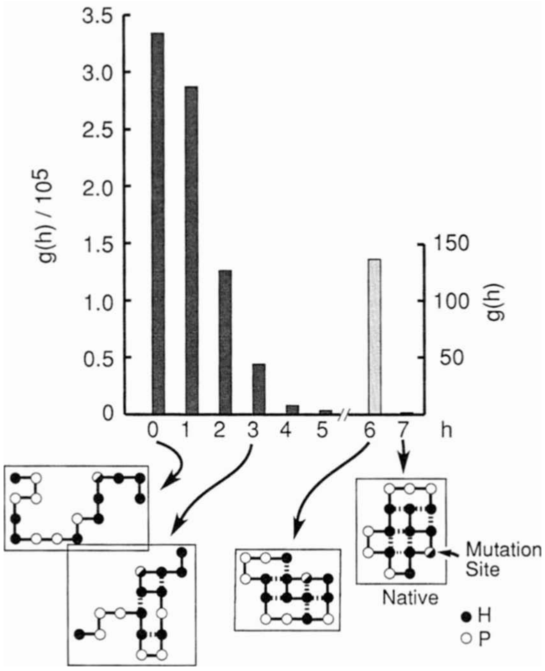

Fig. 28. How do mutations affect the denatured state? The $g(h)$ is shown for a particular 2D HP sequence. Open, filled, and half-filled circles represent P and H monomers and the mutation site, respectively. Example conformations shown for different values of h, the number of HH contacts. Mutation site is a position that has no H interaction in the N state (a "corner" site). The same monomer does have one H interaction (an "edge" site) in some of the dominant D conformations (the $g(6) = 141$ conformations are dominant under strong folding conditions, the $g(h)$ given is for the HP sequence with an H at the mutation site). Thus, H at that position destabilizes the N state. Note that the H/P sequence determines the number and the structure of D conformations. (Modified from Shortle et al. [1992].)

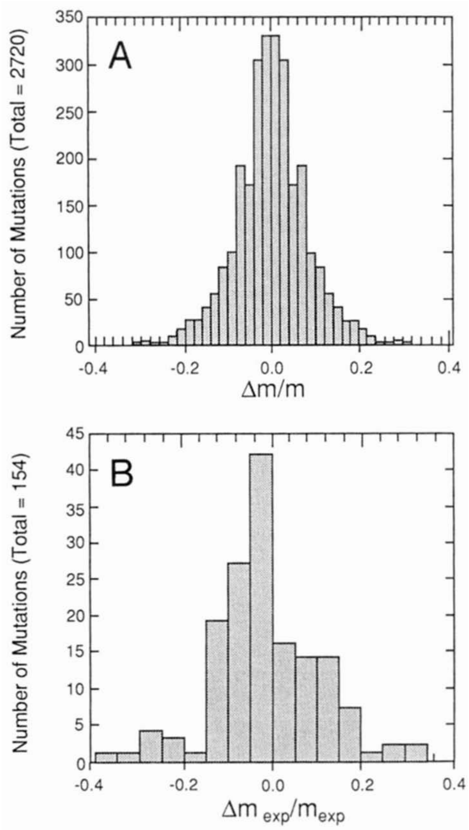

Fig. 29. Mutations affect denatured states. Distribution of denaturant slope "m" values for single mutations is much broader than expected if mutations affect only the native state: A: 2D HP lattice model, over all possible mutations (e = −4kT). B: Experiments on 154 single mutations on staphylococcal nuclease. This includes substitutions of phenylalanine, isoleucine, leucine, methionine, asparagine, proline, glutamine, serine, threonine, valine, and tyrosine residues to both alanine and glycine, as well as substitutions of alanine to glycine and glycine to alanine (modified from Shortle et al., 1992).

the side chains and the backbone (Bromberg & Dill, 1994; Lattman et al., 1994; see Fig. 30). In this view, compact denatured states are broad but limited ensembles of backbone conformations. Experimentally, the compact denatured state of guinea-pig α-lactalbumin has been characterized by hydrogen exchange as highly heterogeneous, in terms of the stability and specificity of both backbone and side-chain interactions (Chyan et al., 1993). NMR line broadening, indicating backbone conformational mobility (attributed to motions on the millisecond time scale), has been observed in the compact denatured states of bovine lactalbumin (Alexandrescu et al., 1993). guinea pig lactalbumin (Baum et al., 1989) and cytochrome *c* (Jeng & Englander, 1991). Thermally denatured ribonuclease A is compact, with about half the secondary structure of the native state (Seshadri et al., 1994), but with considerable solvent penetration as indicated by amide exchange (Robertson & Baldwin, 1991) and calorimetry (Privalov et al., 1989).

A basis for predicting backbone conformations in HP model compact denatured states is the assumption that collapse occurs by a process of "hydrophobic zipping^{n} (see below) until the chain reaches a state of "entropy catastrophe." At the point of the entropy catastrophe, a chain cannot gain HH contacts without large losses in conformational entropy, so it becomes trapped. The trapped states have the characteristics of compact denatured states (Kuwajima, 1989): radii slightly greater than the native protein, common local (helical and turn) contacts, and much hydrophobic clustering, but few nonlocal contacts in common. Small angle X-ray scattering (SAXS) experiments on denatured states of ribonuclease A (Sosnick & Trewhella, 1992) and staphylococcal nuclease (Flanagan et al., 1992) have suggested that bimodal distributions of pairwise interatomic distances, $P(r)$ (Fig. 31), may be a fingerprint of at least some compact denatured states, distinguishing them from native or highly unfolded states. Some hydrophobic zipper endstates have similar bimodal SAXS patterns (Lattman et al., 1994). Figure 32 shows examples of such conformations and their $P(r)$ curves. Earlier spin-glass models of Bryngelson and Wolynes (1987, 1989, 1990) have also predicted a similar "entropy crisis" leading to "misfolded frozen" compact denatured states.

Compact denatured states have common local interactions and hydrophobic clustering

Despite the considerable diversity predicted for the backbone conformations, the hydrophobically zipped compact denatured

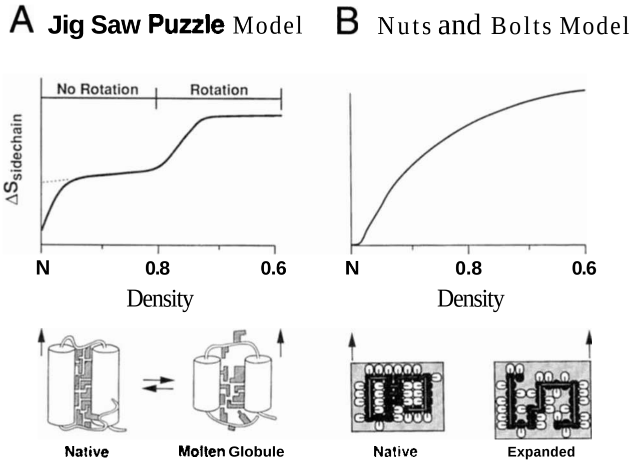

Fig. 30. A: Side-chain molten globule model: a first-order transition is proposed (Shakhnovich & Finkelstein, 1989; Ptitsyn, 1992) to arise from the native to molten globule state because of a sharp increase in side-chain rotational entropy at a critical disjuncture point. Backbone and secondary structures are assumed fixed in native-like conformations and thus assumed to be independent of side-chain freedom. B: By exact enumeration in a simplified model of side chains, the side-chain rotational entropy is shown to increase most sharply even at the earliest expansions from the native state, implying no critical disjuncture point (Bromberg & Dill, 1994). Side chains and backbone are found to be strongly coupled. It is proposed that a critical disjuncture point is not the defining characteristic of the compact denatured states.

conformations share common characteristics (Lattman et al., 1994). They have multiple or diffuse hydrophobic clusters, but no well-defined hydrophobic core. Hydrophobic ***clustering*** involves a much larger solvent-exposed hydrophobic surface area than the hydrophobic ***core*** of a native structure. Hydrophobic clustering in compact denatured states predicts high heat capacities resembling those of unfolded molecules (Ptitsyn, 1987). and low hydrogen exchange protection factors, consistent with observed protection factors of 10's–100's (Hughson et al., 1990; Jeng et al., 1990), where denatured states have protection factors around 1 (Buck et al., 1994) and native states can have protection factors as high as 10^8 (Jeng et al., 1990). Although hydrophobic zipping involves many random and opportunistic steps, nevertheless the many different chain conformations that result often have common locations of hydrophobic clusters, as well as helical and turn contacts, depending on the monomer sequence. These common characteristics have also been detected experimentally in denatured states. Hydrophobic clusters have been observed in equilibrium expanded denatured states of lysozyme (Evans et al., 1991), tryptophan synthase (Saab-Rincon et al., 1993), α-lactalbumin (Alexandrescu et al., 1993; Chyan et al., 1993), and pancreatic trypsin inhibitor (Lumb & Kim, 1994). In the urea unfolded state of 434-Repressor (Neri et al., 1992). the clustered residues are nearly contiguous in the sequence, consistent with local zipping.

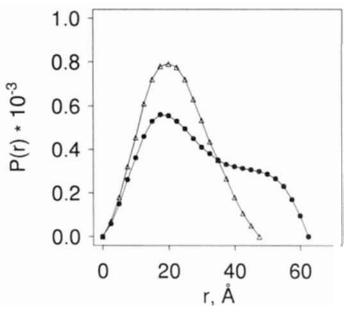

Fig. 31. Bimodal $P(r)$ curves observed for the 1–136 fragment of staphylococcal nuclease (Flanagan et al., 1992), possibly reflecting a general property of compact denatured states. Data were measured in the presence (A) and absence (●) of Ca^{2+} and inhibitor pdTp (3′,5′-biphospho-2'-deoxythymidine).

The gemisch state of proteins is not the molten globule

Incorrectly designed amino acid sequences fold to ensembles of compact conformations, sometimes resembling the desired target structure, but conformationally more diverse. Are these folded states of designed sequences the same as molten globules or compact denatured states? Not necessarily. To distinguish them, we define the gemisch state (which means "mixture" in German), to refer to a model of the native states of incorrectly designed sequences. The distinction between gemisch states and compact denatured states is shown in Figure 33. Gemisch states are native, not denatured, states. Compact denatured states are conformations of sequences that can reach a less diverse distribution of conformations, namely the native structure, under native conditions. Gemisch states are the multiple lowest energy conformations of sequences that can never achieve less diversity, under any conditions; hence, they are multiple native states. That is, gemisch molecules are bad folders, whereas molten globules are denatured states of good folders.

The experimental distinction between gemisch and molten globule states is that a gemisch molecule undergoes no transition to a more ordered state by varying experimental conditions, whereas a molten globule can be folded to a native state by changing conditions. If a molecule that folds uniquely, say at a temperature of 298 K in water at its isoelectric pH, can be caused to expand and increase its conformational diversity by a change in conditions, we would call this a compact denatured state. But if a molecule does not fold uniquely under conditions such as 298 K in water at its isoelectric pH, this would be a gemisch molecule. Homopolymers of H monomers fold to gemisch states: a polyethylene molecule will collapse to a large ensemble of compact conformations and can never achieve a unique fold even in a very poor solvent like water. Sequences with too much hydrophobicity and too many favorable potential contacts are likely to fold to gemisch states; an example may be the four-helix bundle of Handel et al. (1993). The structures of gemisch molecules may not differ from the structures of molten globules: the difference is in the capacity of a sequence to fold uniquely under appropriate conditions.

Some pieces of natural proteins may also be in gemisch states. For example, Peng and Kim (1994) have dissected α-lactalbumin to produce a molecule that consists only of the a-helical domain, which they call a-Domain^{ox}. a-Domain^{ox} does not fold to a native state, but resembles the A-state of α-lactalbumin. Peng and Kim suggest that a-Domain^{ox} has a native-like fold without extensive side-chain packing. Because a-Domain^{o}" does not fold uniquely, it may be an example of the gemisch state. Of course, reattaching the rest of the protein would give a completely different energy landscape. Gemisch state energy landscapes may

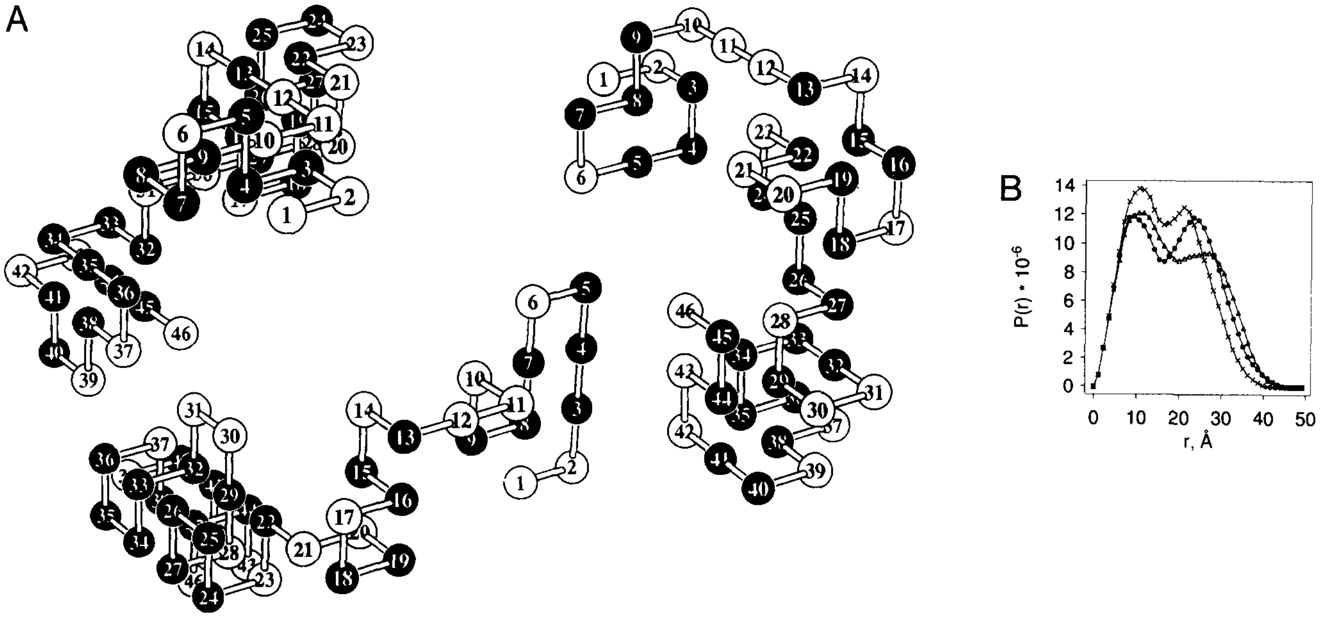

Fig. 32. **A:** Three different **3D** HP lattice model conformations from a single ensemble representing a compact denatured state. B: Corresponding bimodal $P(r)$ curves (from **Lattman et al.**, 1994).

be difficult to distinguish from those with deep kinetic traps (see below). Some proteins may divide into separate domains, some parts being native-like, and some parts having conformational diversity.

Some denatured states are taken as models for denatured states of other proteins, or for other conditions, or for kinetic intermediates. Is the equilibrium acid-denatured state the same as a kinetic intermediate for folding at neutral pH, for example? Model studies suggest caution in equating one nonnative state with another, unless the states are of the same protein and characterized by multiple methods (Ptitsyn et al., 1990). Model studies show that nonnative states are ensembles that shift with conditions (Dill & Shortle, 1991; Shortle et al., 1992). They can be as variable as the conditions that cause them (Calciano et **al.**, 1993; Palleros et al., 1993; Dobson, 1994; Nishii et **al.**, 1994; **Redfield** et **al.**, 1994). Chemical denaturants, temperature, pH, ionic strength, ligands, mutations, and truncations of sequence can change the balance of forces in different ways, as shown in recent thermodynamic and structural studies of denatured states (Tamura et al., **1991a, 1991b**; Damaschun et **al.**, 1993; Carra et **al.**, **1994a, 1994b)** and others reviewed by Shortle (1993). We see no reason to expect the cold-denatured ensemble of structures to be subject to the same balance of forces as the **acid**-denatured ensemble, for example.

Conformational switching: The actions of denaturants and alcohols

A simple exact model has been used to explore the relative importance of local and nonlocal interactions and the effects of solvents and denaturing agents on proteins (Thomas & Dill, 1993). The helical-HP model includes two types of **interaction** – a (nonlocal) HH contact interaction, as in the HP model described above, and a (local) helical propensity. Figure 34 shows an example conformation and energetic interactions in the helical-HP model. When helical propensities are dominant in the helical-HP model, chains undergo helix–coil transitions, and when HH interactions are dominant, chains collapse to compact native states. With the 2D helical-HP model we addressed two questions. (1) Does adding helical propensities cause the HP model to more closely mimic real proteins? (2) What are the mechanisms of denaturing agents such as urea, guanidinium hydrochloride, and trifluoroethanol (TFE) and other alcohols, that might act on both helical and hydrophobic interactions? For example, alcohols denature proteins and induce helical structure **(Tanford** et **al.**, 1960; **Tamburro** et al., 1968). Do they act primarily by strengthening helical propensities (Nelson & **Kallenbach,** 1986) or by weakening hydrophobic interactions (von **Hippel** & Wong, 1965; Brandts & Hunt, **1967)?**

The model makes several predictions. First, if solvents affect both helical and HH interactions, then chains can undergo **"con**formational switching" transitions. For example, a native conformation may switch to a state with more helix and fewer HH contacts (see Fig. 35). This may model the denaturation of globular proteins, including sheet proteins, to helical states in alcohols. The transition from the aqueous native state to the **"TFE** state" of hen egg-white lysozyme has been shown by NMR to be a conformational switch (Buck et al., 1993). At least partially stable alcohol-induced states have also been observed for 0-lactoglobulin (Dufour & Haertle, **1990)**, ubiquitin (Wilkinson & Mayer, 1986; Harding et al., **1991), monellin** (Fan et **al., 1993)**, and the low-pH form of α-**lactalbumin** (Alexandrescu et al., 1994). A 0-sheet to a-helix transition of β-**lactoglobulin** has been observed by Shiraki et al. (1995) in 20% TFE (Fig. **36B)**.

Second, comparison of the helical HP model (Fig. 35) with experimental alcohol titrations of protein solutions (Fig. 36) suggests that TFE acts primarily by weakening hydrophobic interactions in proteins, and that the strengthening of helical propensities

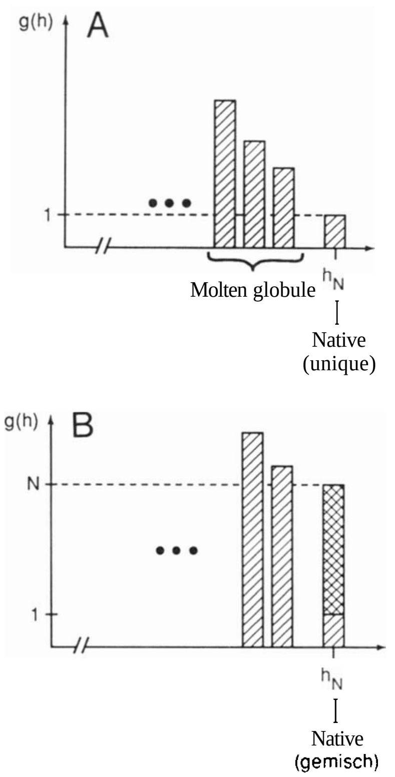

Fig. 33. Example densities of states $g(h)$, indicating how "gemisch" sequences differ from uniquely folding sequences. **A:** Sequence with a unique native structure. **B:** Gemisch-state sequence with multiple ($N > 1$) ground-state conformations.

happens only to a much smaller degree. In the same way, urea denaturation is best modeled as mainly weakening hydrophobic interactions, and to a much smaller degree, weakening helical propensities.

Third, the helical HP model predicts that the internal length distributions of helices and sheets in globular proteins (Kabsch & Sander, 1983) are best reproduced by the model native states only if the model helical propensities are negligible compared to the HH contact interaction (Thomas & Dill, 1993). These comparisons suggest that helical propensities are only weak determinants, relative to nonlocal interactions, of the structures of globular proteins in water. This is consistent with observations of Waterhous and Johnson (1994), shown in Table 2, and others (Rosenblatt et al., 1980; Zhong & Johnson, 1992; M.H. Hecht, pers. comm.), indicating that the conformations of certain 15–18-residue peptides are more strongly determined by the solvent than by their intrinsic helical propensities.

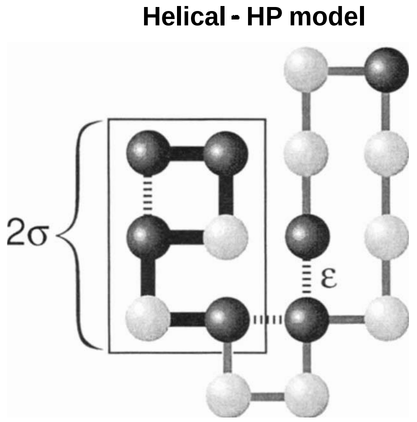

Fig. 34. 2D helical-HP model conformation. Energy per HH contact is ϵ, and energy per helical bond is a (from Thomas & Dill, 1993).

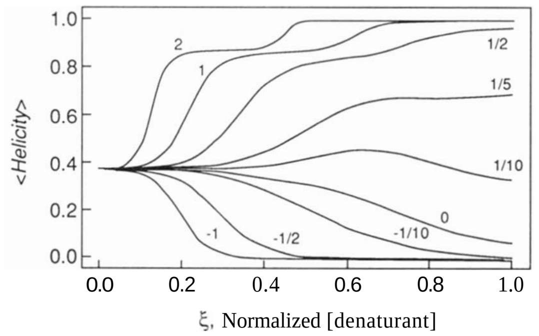

Fig. 35. Denaturation curves for different solvents, from 2D helical-HP model. The coordinate (corresponds to increasing amounts of some denaturant or chemical agent added to native protein; $\xi = 0$ for water. Curves are labeled by numbers indicating the relative importance of hydrophobicity to helical propensities for different solvents: 2, each unit of added solvent increases the helical propensity of the chain twice as much as it weakens hydrophobicity; −1/2, each unit of added solvent decreases the helical propensity half as much as it weakens hydrophobicity. A protein that begins with about 40% helicity in water can undergo transitions increasing its helicity in solvents that favor helical propensities or that weaken hydrophobic interactions. Curves labeled 1/2 to 1/5 are most representative of effects of alcohols and TFE; those labeled <0 are most representative of urea and guanidinium hydrochloride (see Fig. 36).

Hydrogen bonding may play a prominent role in structures in fibrous proteins and in membrane-spanning regions of proteins. Under conditions favoring both helical propensities and contact interactions, the helical-HP model predicts that a large fraction of all monomer sequences (but not all) will fold into helical bundles. It has been shown that membrane-spanning regions of several integral membrane proteins are helical (Deisenhofer et al., 1985; Yeates et al., 1987). The peptidesgramicidin A (Killian, 1992) and Lam B (Wang et al., 1993) undergo a conformational change to an a-helix upon insertion into membranes.

The kinetics and pathways of folding

How do proteins find their native states? Are there few or many pathways? What are the folding transition states? How do the amino acid sequences specify the folding pathways? How do mutations affect folding kinetics?

The way we understand folding kinetics depends in part on whether we believe folding is dominated by local or nonlocal interactions. Assuming local interactions are important factors in reducing conformational searching, as in diffusion/collision (Karplus & Weaver, 1976, 1994) or framework models (Ptitsyn et al., 1972; Kim & Baldwin, 1982; Baldwin, 1989; Ptitsyn, 1991; Karplus & Weaver, 1994), has led to the view that partially stable helices form early through fluctuations, reducing the conformational search, so they can then assemble into tertiary

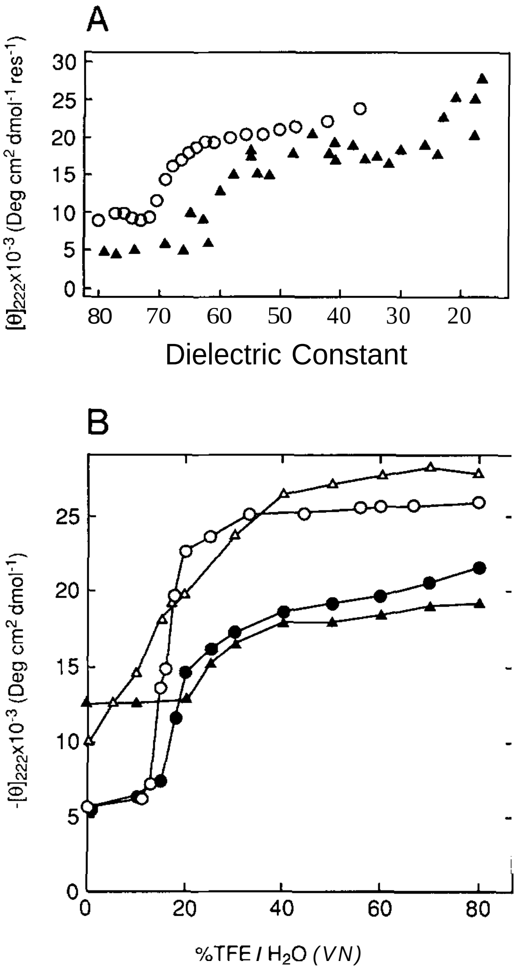

Fig. 36. Experimental alcohol denaturations of proteins. **A:** Helicity as measured by molar ellipticity at *222* nm, as a function of the dielectric constant of the solvent, for different alcohols. TFE denaturation of hen egg-white lysozyme is shown with circles and denaturation of ubiquitin using methanol, ethanol, isopropanol, and butanol is shown with triangles (Wilkinson & Mayer, 1986). B: TFE denaturation of intact (3-lactoglobulinat pH *2* (O)and pH 6 (O)and of RCM-β-lactoglobulin at pH 2 (A) and pH 6 (A) (data of Shiraki et al., 1995, reproduced with permission). These data may be compared to the HP model results in Figure 35.

structures. In this view, secondary structure fluctuations precede collapse and assembly (see Fig. 1).

On the other hand, the kinetics will be different if folding is dominated by nonlocal interactions. With collapse as the driving force, models indicate one or more stages involving: (1) a fast collapse in which hydrophobic clusters, helices, and sheets are driven to form through a zipping process, which can result in a broad ensemble of compact conformations, and (2) a slow process of breaking incorrect (nonnative) HH contacts to proceed to the native structure.

The slow process that overcomes the transition state energy barriers requires an opening of the chain to break incorrect HH contacts. There are multiple paths and transition states, but the ensemble of folding trajectories may have common features for a given protein, providing support for the apparently paradoxical view that proteins fold both by multiple paths and by specific sequences of events. In fact, these views are not mutually exclusive (see below). A principal conclusion from these studies is that protein folding has no simple reaction coordinates of the type used to describe small molecule reactions.

Energy landscapes

Folding kinetics can be described in terms of "energy landscapes." Figure 37 shows a few possible candidate landscapes for protein folding. The landscape of a sequence with "gemisch" ground states (see above) is shown in Figure 38. Folding would be slower if proteins had "golf-course" landscapes (Fig. 37A) than if they had "smooth funnel" (Dill, 1987, 1993; Leopold et al., 1992; Zwanzig et al., 1992; Bryngelson et al., 1995; Chan & Dill, 1994) landscapes (Fig. 37B). In smooth funnels, any conformation can proceed through a series of downhill energy steps to the native state, with no energy barriers. In Figure 37A, the landscape is flat, so all nonnative conformations have the same free energy, and the native state can only be found by random search. The "Levinthal paradox" (Levinthal, 1968) stems from estimating the difficulty of folding proteins by random search, by assuming a golf-course-like landscape. On the golf-course landscape, the search problem depends only on the size of the conformational space. But for proteins under native conditions, different conformations have different energies, implying that the flat golf-course landscape is not a good folding model. In our

Table 2. ***Data reproduced from Waterhous and Johnson (1994)***

Peptide sequence	Chou–Fasman[a]	Crystal[b]	%α in TFE[c]	%β in SDS[d]
IIPTAQETWLGVLTIMEHTV	β	a	*72*	65
LSGGIDVVAHELTHAVIDY	β	a	*72*	98
PAVHASLDKFLSSVSTVL	β	a	65	95
GYQCGTITAKNVTAN	β	β	64	94
$(VAEAK)_3$	*a*	–	79	80
$Y(EAAAK)_3A$	α	–	69	–

[a] Structure predicted by the algorithm of Chou and Fasman (1974a, 1974b).
[b] Structure observed in native protein.
'Amount of helical structure in isolated peptide in 90% TFE, determined by CD.
[d] Amount of sheet structure in isolated peptide in *2-6* mM SDS, determined by CD.

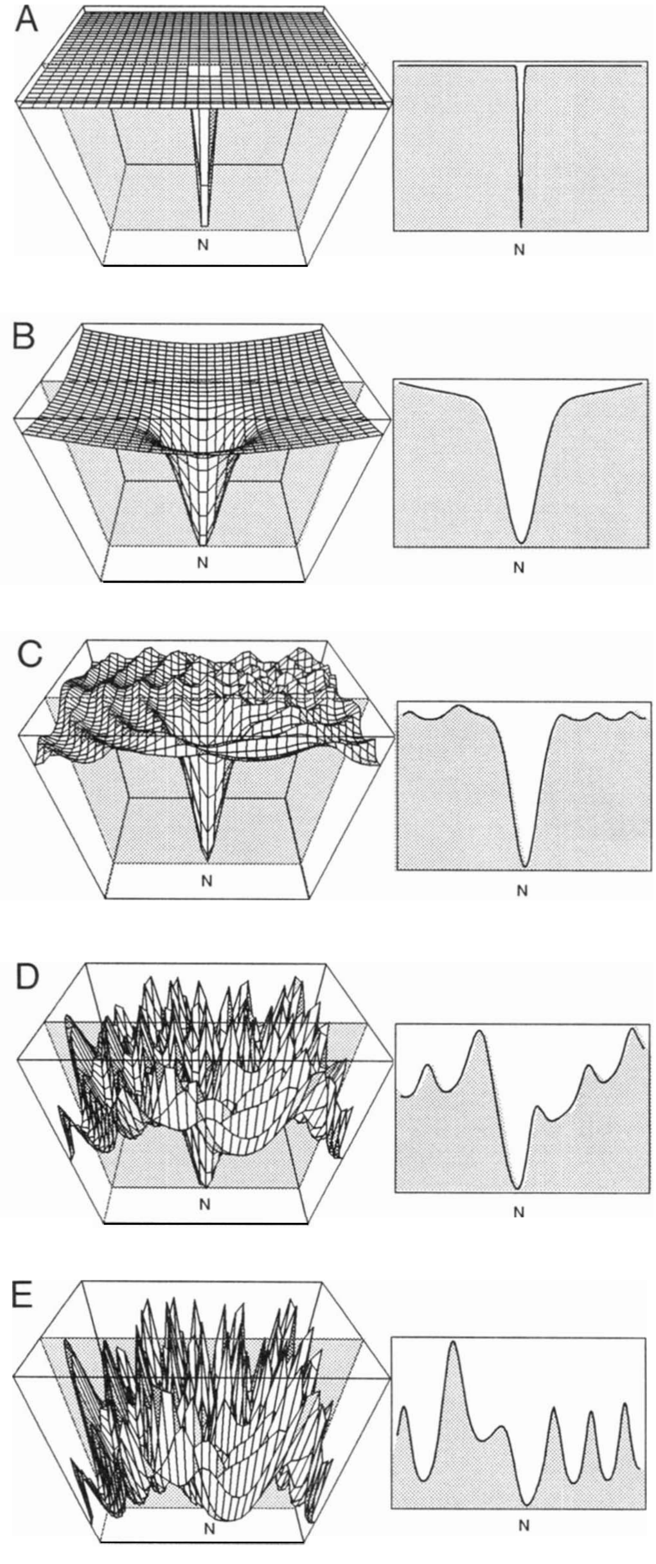

Fig. 37. *Schematic* drawing of multidimensional conformational energy landscapes. Energy is on the vertical axis and the other axes represent conformational degrees of freedom. N is the native structure. **A:** "Golf-course" landscape. **B:** Smooth funnel landscape in which every conformation can reach N without encountering energy barriers. C: Both smooth and rough landscape aspects. Overall, there is a broad, smooth funnel leading to the native state, but there is also some roughness superimposed on this funnel. D, E: "Rugged" landscapes. Local minima and barriers are higher in E.

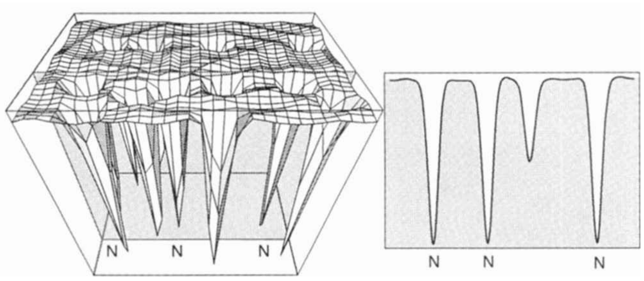

Fig. 38. Schematic drawing of the energy landscape of a sequence with gemisch ground states (see Fig. 33).

view, the Levinthal paradox is not a satisfactory description of the protein folding problem. Proteins with funnel-like folding landscapes are sometimes said to be under "thermodynamic" control, and those with rugged folding landscapes are said to be under "kinetic" control (Baker & Agard, 1994). Under folding conditions, comparison of Figure 37A and B (the landscapes of which are of the same size) suggests it is not the size but the shape of the landscape that matters (Chan & Dill, 1993b; Dill, 1993). Using random-energy models, Bryngelson and Wolynes (1987, 1989) first suggested that the landscape for protein folding must have some "ruggedness" (Fig. 37C,D,E). What shape is it?

An energy landscape is a multidimensional surface of the (free) energy versus the degrees of freedom. Two factors characterize the shape of an energy landscape: (1) the density of states $g(h)$, and (2) a measure of structural similarity or kinetic "nearness" of one conformation to another. Figure 39 shows how $g(h)$ is related to the ruggedness of a landscape. If there are many low-energy conformations it means that g is large when h, the number of HH contacts, is large (h near h_N), and the

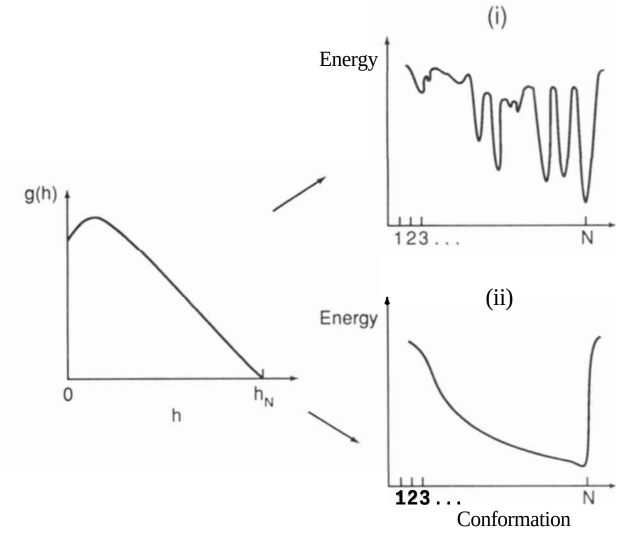

Fig. 39. Relation between density of states $g(h)$ and energy landscape. The energy landscape always defines $g(h)$, but $g(h)$ alone does not fully specify the energy landscape (see text for details). If g is large when h is large, the energy landscape may be rugged (i), or it may resemble a wide-bottom smooth funnel (ii).

landscape could be rugged, or it could be shaped like a wide-bottom smooth funnel, for example. On the other hand if g is small when h is large, the landscape cannot be very rugged. It would be more like a golf course. But the shape of a landscape is determined not only by $g(h)$; it also depends on some measure of conformational "distance" along a kinetic reaction coordinate (Chan & Dill, 1993b, 1994). Consider two hypothetical landscapes with identical $g(h)$ (Fig. 40). Suppose the first has reaction coordinate h, the number of HH contacts. Then because lower energies correspond to larger h, this landscape will be shaped like a funnel, and folding would be fast (Fig. 40A). Now instead suppose we define a different reaction coordinate by rearranging the conformations along the horizontal axis, as in Figure 40B. In this case, we would have a "reverse funnel," implying slow folding, because the native state can only be reached by uphill energy steps from most of the denatured conformations. The $g(h)$ is identical in both cases.

This comparison raises two points. First, kinetics goes beyond thermodynamics: Figure 40A and B represents exactly the same thermodynamic model (i.e., the $g(h)$'s are identical so the partition functions are identical), but they represent completely different kinetics. Second, the comparison indicates how precariously dependent kinetic modeling is upon the seemingly arbitrary choice of conformational adjacency and distance (Chan & Dill, 1993b, 1994). What is the appropriate model for conformational distance? The landscape in Figure 40B would seem to have an unphysical definition of reaction coordinate. However, the main point here is that neither Figure 40A nor Figure 40B show suitable reaction coordinates. We distinguish between an order parameter, a thermodynamic measure of progress from one state to another, and a reaction coordinate, a kinetic measure of progress. Figure 40A defines a legitimate order parameter, because the horizontal-axis quantity defines a relevant measure of progress from denatured to native states for computing the free energy. (The horizontal axis in Figure 40B is not a good order parameter, because it cannot be construed as a measure of progress from one state to another.) But neither quantity is a good reaction coordinate. Why not?

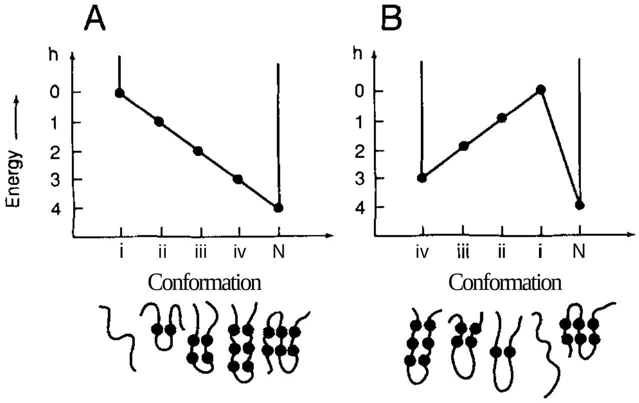

Fig. 40. Two hypothetical energy landscapes with identical conformations, and hence, identical densities of states $g(h)$. Different landscapes result from different definitions of conformational adjacencies or nearness (reaction coordinates). **A:** Conformations are in order of increasing number of HH contacts. B: Conformations are in the reverse order, except for the native state.

Although any progress variable is suitable as an order parameter for thermodynamic purposes, a kinetic reaction coordinate requires more. A reaction coordinate must be not only a measure of progress, but of kinetically achievable progress. That is, a suitable reaction coordinate must define a series of small-conformational-change steps, between kinetically adjacent states, that can lead from one conformation to another (Chan & Dill, 1993b, 1994). The essential difference is that, for a reaction coordinate, nearby regions on the horizontal axis must represent conformations that are structurally similar enough to interconvert rapidly. An order parameter does not require this. Such small steps are defined by "move sets" in dynamic Monte Carlo simulations (Fig. 41). Some quantities, such as counts of native-like contacts, have been used as reaction coordinates (Shakhnovich & Gutin, 1990a; Šali et al., 1994b; Shakhnovich, 1994). Although these are valid order parameters, simulations (Miller et al., 1992) and exact studies (Chan & Dill, 1994) show that such quantities do not satisfy the requirements of a legitimate reaction coordinate. For a given h, some conformations can get to the native state through downhill moves, but others will be in energy traps. Figure 42 shows how different conformations of the same h have different kinetic access to the native state. Exact studies also show that energy landscapes inferred from models are strongly dependent on the choice of move sets, which are arbitrary constructs, and thus Monte Carlo dynamics must be interpreted with caution (Chan & Dill, 1993b, 1994).

With those caveats, Figure 43 shows a protein folding energy landscape from exhaustive enumeration using a simple exact model, the 2D HP lattice model. It illustrates many of the features of model protein folding landscapes. The main results discussed below are not limited to this model; they are common to a wide range of protein models. Because HH contacts have a favorable free energy under folding conditions, lower free energies correspond to more HH contacts. The native state, the lowest point on the landscape, has 6 HH contacts, indicated by h = 6 on the vertical axis. The horizontal axis indicates conformations differing by a single Monte Carlo move.

A kinetic pathway of folding (path I) is indicated by the sequence of conformations: a, b, c, d, e', f, g', h', N in Figure 43. This is a "throughway" path, a funnel-like part of the landscape, in which the chain never encounters an energy barrier in this model. For throughway paths and funnel-like landscapes, the folding process might involve many small barriers that are below the level of resolution of simple exact lattice models. Most of the recent statistical mechanical models indicate multiple folding paths (Miller et al., 1992; Camacho & Thirumalai, 1993a; Bryngelson et al., 1995; Chan & Dill, 1994; Thirumalai, 1994), as was suggested by Harrison and Durbin (1985) (see Fig. 44). Path II involves a kinetically trapped "local minimum" conformation, B. Kinetic traps are low-energy nonnative states. Because they are low energy, they have many HH contacts, hence, they are usually compact, as B is. Thus, the main kinetic traps to folding are generally the most compact denatured states.

What are the transition states? The slow bottleneck step along path II is from the trapped conformation B to the transition state h. Conformation h is one of many on a "transition state plateau." The step from B to h involves a breaking of incorrect (nonnative) HH contacts and a corresponding opening up and expanding of the chain, at least locally. The trapped states are compact; the transition states are more open. Transition states represent increased conformational entropy and contact free en-

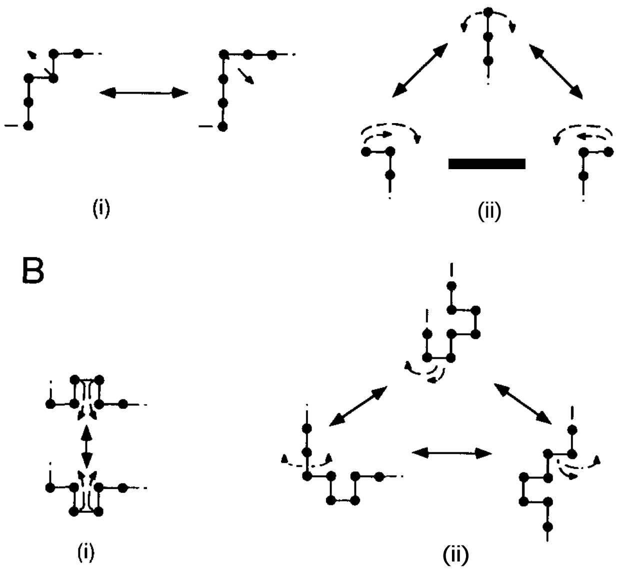

Fig. 41. Move sets used in 2D exact lattice enumeration studies of chain dynamics (Chan & Dill, 1993b, 1994). Double-headed arrows show which conformations are adjacent. Dashed arrows show monomer moves. A: Move set 1 (MS1); (i) a three-bead flip; (ii) end flips. B: Move set 2 (MS2) includes those in MS1 and also (i) crankshaft moves; (ii) rigid rotations.

ergy relative to the traps. In general, there are large ensembles of both traps and transition states. This example pathway further illustrates that the number of native contacts is not a viable reaction coordinate; both conformation B and n along path II have three native contacts. However, B is a deep local minimum, whereas from n the chain need not surmount any energy barrier to reach the native structure N. The following sections describe model folding kinetics in more detail.

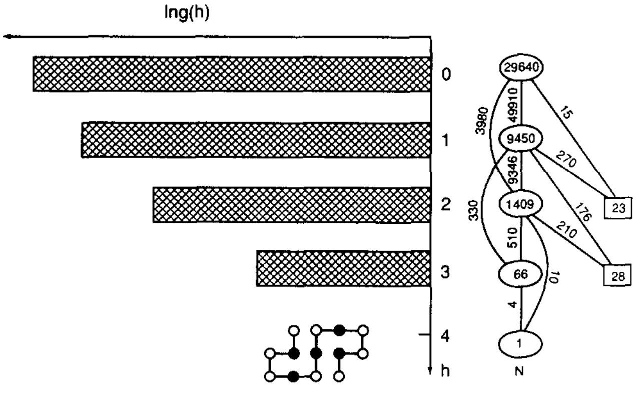

Fig. 42. Conformational "flow" diagram (right) and corresponding $g(h)$ showing the kinetic accessibilities of states (H.S. Chan & K.A. Dill, unpubl. results). This sequence has one native state, indicated at the bottom. Numbers in ovals represent conformations along throughway paths, whereas numbers in squares represent trapped conformations that must first go uphill before going downhill. Numbers along lines represent adjacencies, i.e., number of conformational pairs separated by a single move. Comparing square 28 with oval 66, it is clear that, even though both "thermodynamically" resemble the native state to the same degree (same h = 3), 66 conformations have direct downhill access to native, whereas 28 conformations are in a trap and must first climb an energy barrier; hence, they have very different kinetic relationships to the native state (see Chan & Dill, 1994).

Proteins collapse rapidly to compact states, then rearrange slowly to the native state by crossing energy barriers

Many simulations predict that polymer and protein collapse can occur in multiple stages (Abe & Gō, 1981; Gō & Abe, 1981; Shakhnovich et al., 1991; Honeycutt & Thirumalai, 1992; Camacho & Thirumalai, 1993a; Chan & Dill, 1994; Šali et al., 1994a, 1994b; Socci & Onuchic, 1994). In models of HP chains (Chan & Dill, 1994), or of homopolymers in poor solvents such as a chain of hydrophobic monomers in water (Chan & Dill, 1993b), for most sequences there is a general fast collapse to a broad distribution of compact denatured states, with much hydrophobic clustering and many incorrect (i.e., nonnative) contacts, followed by slow rearrangements and barrier-crossing processes to reach the lowest energy states. A few sequences fold in a single fast process. There are many paths the chains follow (Miller et al., 1992; Camacho & Thirumalai 1993a; Chan & Dill, 1993b, 1994). Folding kinetics is strongly sequence dependent. This applies to HP sequences (Chan & Dill, 1994), as well as to sequences with more interaction types (Shakhnovich et al., 1991; Leopold et al., 1992; Šali et al., 1994a, 1994b; Socci & Onuchic, 1994). Recent simulations using a perturbed homopolymer model indicate that the slow stage is more sequence dependent than the fast stage (Socci & Onuchic, 1994). Folding times of different unique sequences can differ by many orders of magnitude. Depending on the sequence, the relative time scale between the fast and slow stages may vary over a large range (Chan & Dill, 1994). A simple classification scheme of time scales is provided by Bryngelson et al. (1995).

Consistent with theory, experiments show that real proteins often fold with at least two distinct time scales, often with a transient population of nonnative compact conformations with significant hydrophobic clustering (Kuwajima, 1989, 1992; Chaffotte et al., 1992a, 1992b; Baldwin, 1993), but significant solvent exposure (Lu & Dahlquist, 1992; C.R. Matthews, 1993). The collapse process is rapid (Garvey et al., 1989; Radford et al., 1992; Barrick & Baldwin, 1993; Jennings &Wright, 1993; Briggs & Roder, 1994; Feng & Widom, 1994; Itzhaki et al., 1994; Uversky & Ptitsyn, 1994). Different populations of protein molecules fold by different pathways (Englander & Mayne, 1992; Radford et al., 1992; Englander, 1993; Fersht, 1993; Jennings et al., 1993; Miranker et al., 1993; Elöve et al., 1994).

The fast process may occur by hydrophobic zipping, with concurrent formation of secondary structure

Chain collapse can proceed by "zipping" together hydrophobic contacts (Fig. 45). Suppose a chain is highly unfolded when native conditions are "turned on," as when denaturant is jumped to zero concentration. Such native conditions cause hydrophobic residues to become "sticky." Two H monomers that are near neighbors in the sequence will contact because the free energy decrease for forming the contact outweighs the loss of chain conformational entropy of that particular HH link. If this HH contact then brings other H monomers into spatial proximity, then they too can contact without much further loss of conforma-

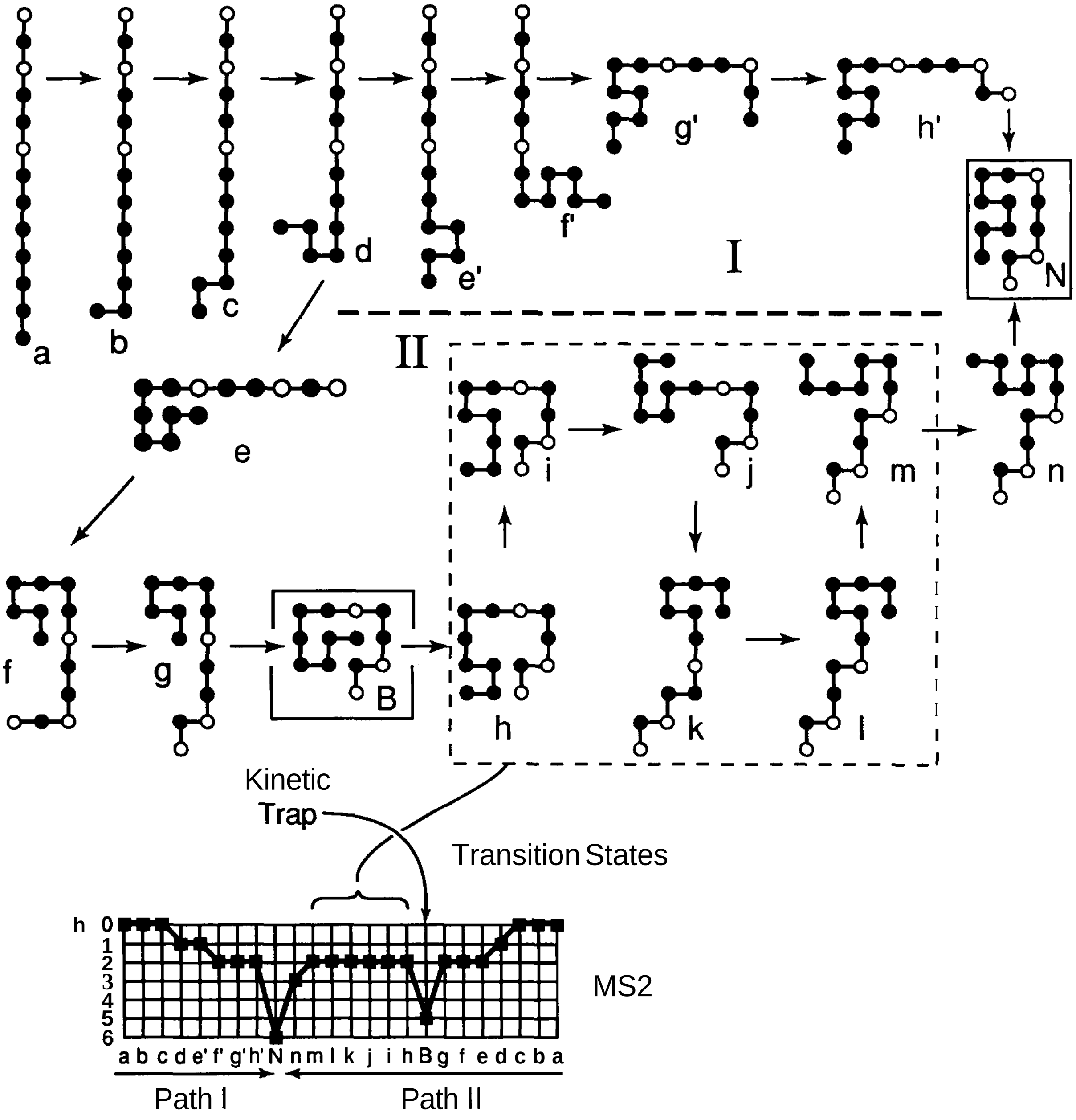

Fig. 43. Typical folding paths and their energy landscapes. Chains begin at conformation "a" and proceed to the native structure N. Path I has no barriers to N (a "throughway" path), but path I1 passes through local-minimum conformation B, then uphill across transition-state conformations to N. Bottom plot gives the history of the number of HH contacts *h*, and is the energy landscape along the two paths. (From Chan and Dill [1994].) Both paths begin with a hydrophobic zipper collapse.

tional entropy (Dill et al., **1993; Fiebig &** Dill, **1993)** and gain a net free energy advantage. This opportunistic process can continue as a zipping together of HH contacts, with only minimal loss of conformational entropy at each step. Hydrophobic zippers do not explore much of the total conformational space. Nevertheless, model studies show zippers are capable of finding globally optimal conformations **(Fiebig &** Dill, 1993), although most zipper "endstates" terminate in nonnative conformations (see Fig. 46). In this regard, we believe hydrophobic zipping models how proteins collapse rapidly to nonnative states. The slow annealing to the native structure then requires unzipping incorrect contacts.

Zipping implies that hydrophobic collapse will be concurrent with the development of helices and sheets **(Fiebig &** Dill, **1993;** Lattman et al., **1994).** Although the forces causing collapse, the hydrophobic interactions, are postulated to be stronger than the helical propensities, it does not follow that collapse precedes secondary structure formation in time. As hydrophobic zipping assembles nonpolar monomers into a core, it progressively stabilizes ensembles of helices and sheets. In this regard, collapse

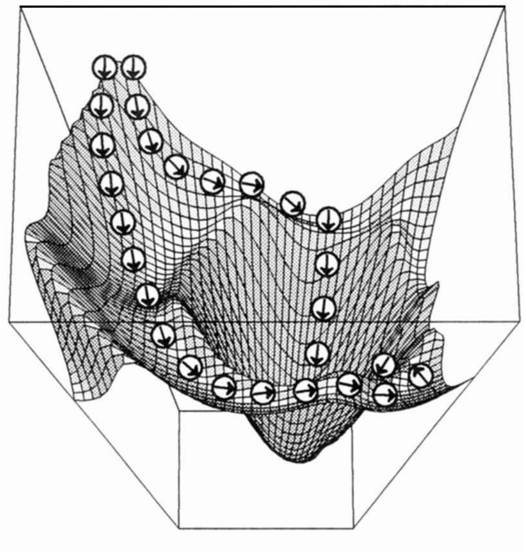

Fig. 44. A hypothetical conformational energy landscape illustrating the difficulty in defining a reaction coordinate. Even two nearly identical conformations going to the same final state can take very different paths.

is not entirely nonspecific: although there may be much disorder in the collapsed states, there is also much sequence-dependent order (Lattman et al., 1994). Experiments confirm that considerable collapse and secondary structure happen quickly in folding (Gilmanshin & Ptitsyn, 1987; Semisotnov et al., 1987; Briggs & Roder, 1992; Chaffotte et al., 1992a, 1992b; Elove et al., 1992; Serrano et al., 1992; Baldwin, 1993; Barrick & Baldwin, 1993; Jennings & Wright, 1993; Itzhaki et al., 1994; Nishii et al., 1994).

Evidence from Gast et al. (1993) appears to conflict with the view that collapse drives secondary structure formation. They have shown that the refolding of yeast phosphoglycerate kinase upon jumping the temperature from 0 to 30 °C is accompanied by a fast formation of about 40% of the secondary structure by CD measurements and a slower process of collapse plus the remaining secondary structure formation. The interpretation of these experiments is complicated because they involve multidomain proteins, so the radii measured (by light scattering) may be those of the largest components in solution. If there were a fast collapse of a small domain of the chain, the CD might see it, whereas the light-scattering would not.

Zipping does not imply that nonlocal processes are slow. On the contary, zipping is an explanation for how nonlocal contacts can be made so rapidly. Zipper simulations show that chain ends can come together quickly for some monomer sequences (Lattman et al., 1994). The N- and C-terminal helices of cytochrome c are observed to assemble on the fast collapse time scale (Roder et al., 1988). Because zipping is an hypothesis about kinetics, it implies that if proteins fold this way, then some proteins may reach only metastable states and not achieve their global minima in free energies. Some proteins appear to be in metastable states (Baker & Agard, 1994).

There is evidence for hydrophobic zipping in proteins. For proteins with considerable helix, it is difficult to distinguish whether helices are driven by local or nonlocal interactions. But sheet proteins have predominantly nonlocal interactions. In interleukin-1β, the first sheet protein for which detailed kinetic data are available, Gronenborn and Clore (1994) observe folding kinetics consistent with hydrophobic zipping (see also Varley et al., 1993). The fast process, which appears zipper-like, leads to an ensemble of different sheets without native-like hydrogen bonding patterns. Recently zipper-like ensembles have also been observed in sheet peptides taken from platelet factor-4 (Ilyina & Mayo, 1995; Ilyina et al., 1994).

What is the reaction coordinate for folding?

Figures 37 and 44 show that, for rugged multidimensional energy landscapes, there is no simple way to define a single reaction coordinate, i.e., a single lowest-energy sequence of events for the entire ensemble of folding molecules. Polymer collapse involves an ensemble of lowest energy trajectories through an energy

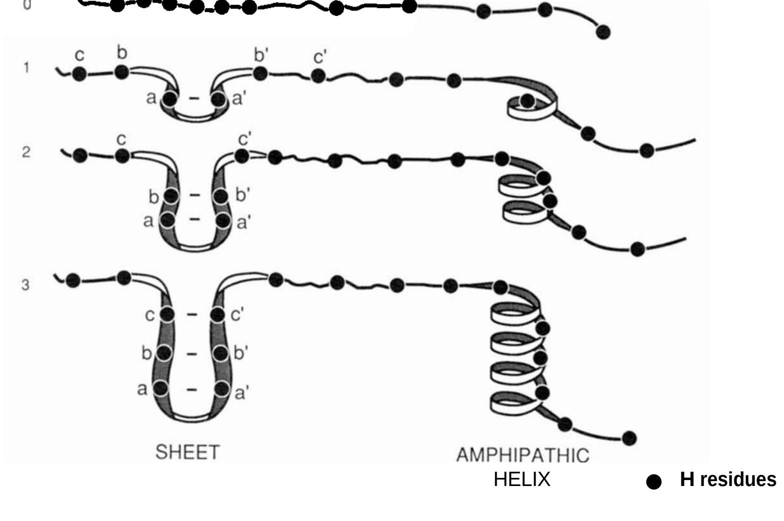

Fig. 45. Hydrophobic zipper model of protein-folding pathways. The closest hydrophobic (H) residues (●) in a sequence pair together first, e.g., a and a' in step 0. They constrain the chain and bring other H residues, such as the (b, b') pair, into spatial proximity. Now (b, b') further constrains the chain and brings the (c, c') pair into spatial proximity, etc. As H contacts form and develop a core, helices and sheets zip up if they have appropriate HP sequences. (From Dill et al. [1993].)

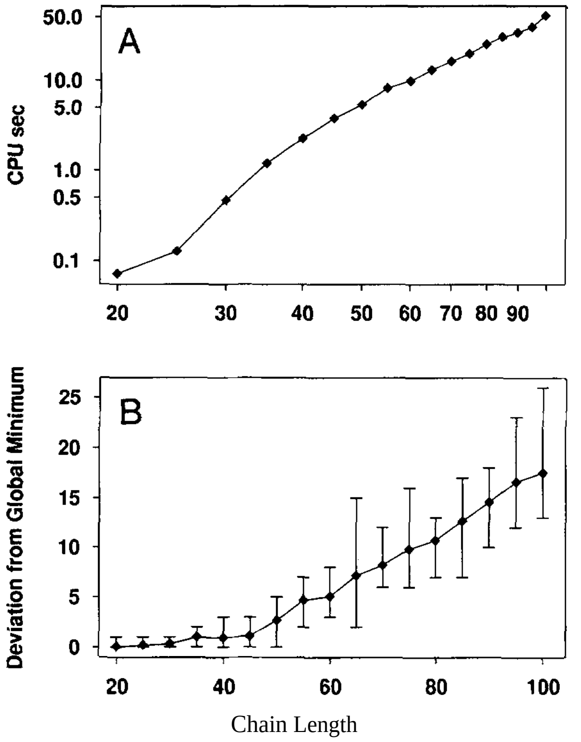

Fig. 46. Test of hydrophobic zippers as a conformational search strategy, for HP chains on 3D simple cubic lattices. Chain lengths are shown on the horizontal axes. About 20 different sequences were tested for each chain length. **A:** Computer time scaling with chain length. B: How far the hydrophobic zipper endstates are from globally optimal conformations (by the CHCC method), in units of HH contacts. Up to 50-mers, hydrophobic zippers can find native states in reasonable computer time for some sequences.

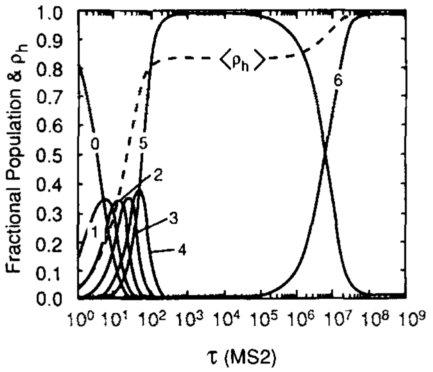

Fig. 47. Time evolution of folding in 2D HP Monte Carlo kinetics (Chan & Dill, 1994). Folding conditions are turned on at time $\tau = 0$; chains are started open, i.e., distributed uniformly among the $h = 0$ open conformations. Each curve is the fractional population of conformations versus time that have h HH contacts (native population corresponds to $h = h_N = 6$). Note that for this sequence the average hydrophobic burial $\rho_h \equiv h/h_N$ shows a fast collapse, then an annealing to the native state that is five orders of magnitude slower for this particular ϵ.

landscape. But multiple paths do not imply that folding properties are random functions of time. Ensemble-averages of time-dependent properties can readily be computed. For example, Figure 47 shows the time-dependent hydrophobic burial in one sequence in the 2D HP model: in a small number of time steps, chains reach a metastable (nonoptimal) hydrophobic burial, but only over a much larger number of time steps do they anneal to the native state.

That individual chains fold by multiple paths is not necessarily inconsistent with experiments showing specific pathways. Figure 48 illustrates how a "pathway" can be observed even when individual chains follow diverse routes. The distinction we draw is between: (1) the many different ways each individual chain gets to the native state, versus **(2)** the ensemble average of some experimentally observed quantity, taken over all the chains. Consider the relevant and irrelevant degrees of freedom. An experiment observes certain contacts or specific bond conformations in a part of the chain; these are the relevant degrees of freedom. The irrelevant degrees may be those for other parts of the chain, perhaps distant from the assembly of interest, or where there is too much conformational diversity to specify a given structure. As the chain folds, the experiment may show that the relevant degrees follow some particular sequence of events, on average. But because a microscopic pathway of an individual chain is defined in terms of *all* its degrees of freedom, the fact that other parts of the chain may have different conformations during that process implies that the individual molecules are traversing different microscopic pathways. Hence, whether we believe chains follow few or many paths depends in part on whether we define "paths" to mean: **(1)** what each molecule is doing, or (2) what experiments are observing. Thus, even when there are many diverse configurations that are traversed in statistical mechanical and computational models, they can readily lead to ensemble-averaged properties showing different macroscopic properties at different times in the folding process.

How should we define a reaction coordinate for folding? In Monte Carlo dynamics, "move sets" define allowable "steps" (see Fig. 41) along a process of conformational change. We have defined kinetic "distance" as the minimum number of moves required to get from one conformation to another along lowest energy paths (Chan & Dill, **1993b,** 1994). The use of lowest energy paths is a standard requirement for defining reaction coordinates, and the "minimum number of moves" is needed to satisfy the triangle inequality to give a proper measure of distance.

The exercise of defining a proper reaction coordinate along a lowest-energy or "minimum climb" pathway (Chan & Dill, 1994) to the native state leads to a most interesting and counter-intuitive conclusion: open chain conformations are kinetically "closer" to the native state than are many compact conformations (Chan & Dill, 1994). That is, there are usually fewer and lower barriers to reaching the native state from a more open state (if a minimal-climb path is taken) than from a more compact state. The process of folding is usually a process of first moving away from the native state in the fast collapse stage (in this kinetic sense), then toward the native state in the slow barrier-climbing steps (Chan & Dill, 1994).

Consistent with this view, Ikai and Tanford (1971) express their kinetic results on cytochrome *c* in terms of N ⇌ U ⇌ X, where N is the native state, U is the fully unfolded state, and X is incorrectly folded. X is sometimes called an "off-pathway" state: if it were made more stable, folding would be slower. Experiments confirm that initial "burst-phase" condensation of-

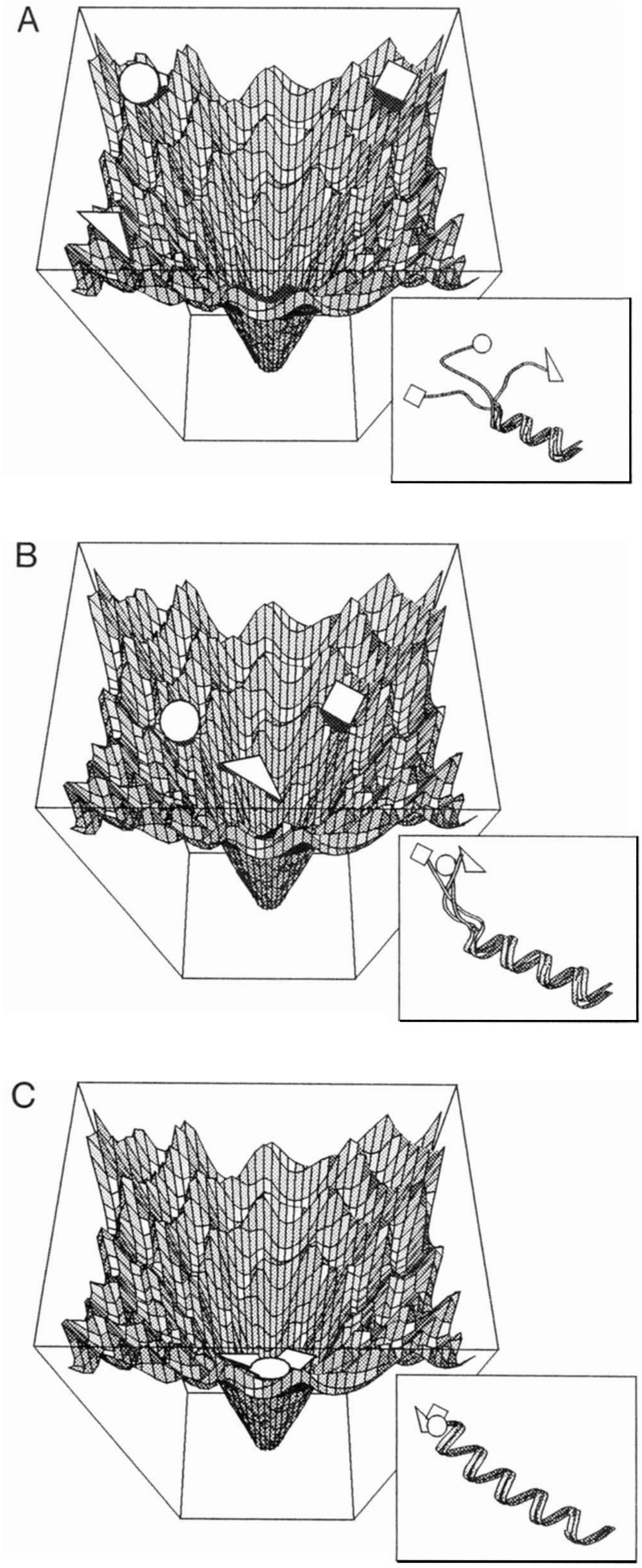

Fig. 48. Multiple pathways can be consistent with specific sequences of observable events in protein folding kinetics. Three different starting conformations are shown. Suppose only the helical parts are observable in the experiment. Conformations of other parts of the chains are "irrelevant" in that they are not resolved by the experiment at that stage. Each molecule traverses a different path downhill to the native helix, while the experiment "sees" a single "path," i.e., formation of a helix.

ten leads to some misorganization of the chain, resulting in major barriers to folding (Radford et al., 1992; Sosnick et al., 1994; reviewed by Creighton, 1994 and Dobson et al., 1994). Consistent with the theoretical prediction that there are many barriers with different heights (Camacho & Thirumalai, 1993a; Chan & Dill, 1994), the kinetics from "burst-phase" intermediates to the native state are multiphasic for some proteins (Matthews & Hurle, 1987; Kuwajima et al., 1991; Jennings et al., 1993).

Folding transition states involve an opening of the chain

The prediction that chains must open up at a late stage of folding before reaching the native structure is consistent with the "cardboard box" model of Goldenberg and Creighton (1985); with experiments on bovine pancreatic trypsin inhibitor (BPTI), that nonnative species accumulate transiently to a certain degree and some unfolding of a kinetic intermediate precedes formation of the native structure (Weissman & Kim, 1991, 1992a, 1992b; Creighton, 1992; Kosen et al., 1992); and with CD experiments on hen lysozyme indicating nonnative disulfide bonds or aromatic interactions in folding intermediates (Chaffotte et al., 1992a, 1992b; Hooke et al., 1994).

Transition states may involve very small local expansions from the compact trapped states (Chan & Dill, 1994); hence, this predicted opening of the chain need not conflict with experimental observations that the energetic properties of transition states are often close to those of their native states (Segawa & Sugihara, 1984; Chen et al., 1989; Serrano et al., 1992). This view is further supported by kinetics experiments on mutants of chymotrypsin inhibitor 2 (Jackson et al., 1993), which show that interactions at the edges of the hydrophobic core are significantly weakened or lost in the unfolding transition state.

Mutational effects on folding kinetics are subtle

Mutations alter folding kinetics (Matthews & Hurle, 1987; Fersht, 1993; C.R. Matthews, 1993), sometimes radically (Iwakura et al., 1993; Sosnick et al., 1994) and sometimes to a lesser extent (Hooke et al., 1994). Simple exact models also show that mutational effects can be very subtle and not predictable from knowledge of the native structure alone. Figure 49 shows two HP sequences that have identical native structures. They differ by only a single monomer. An H → P mutation distant from this position in the sequence speeds the folding for the first sequence and slows it for the second. In the first instance, an HH contact on the folding pathway must be broken (which is an uphill step in energy). Replacing the H by P removes this barrier and speeds folding (see Fig. 49A). In the second instance, an H serves to "fish" another H away from a nonnative HH contact along the pathway. Replacing that H with P now eliminates a way to break an HH contact and slows folding (Fig. 49B) (Chan & Dill, 1994).

Relationship between the thermodynamics and dynamics of protein folding

The most direct evidence that proteins do not fold along golf-course landscapes, and do not follow Levinthal-like random searches, is that folding rates depend on external conditions such

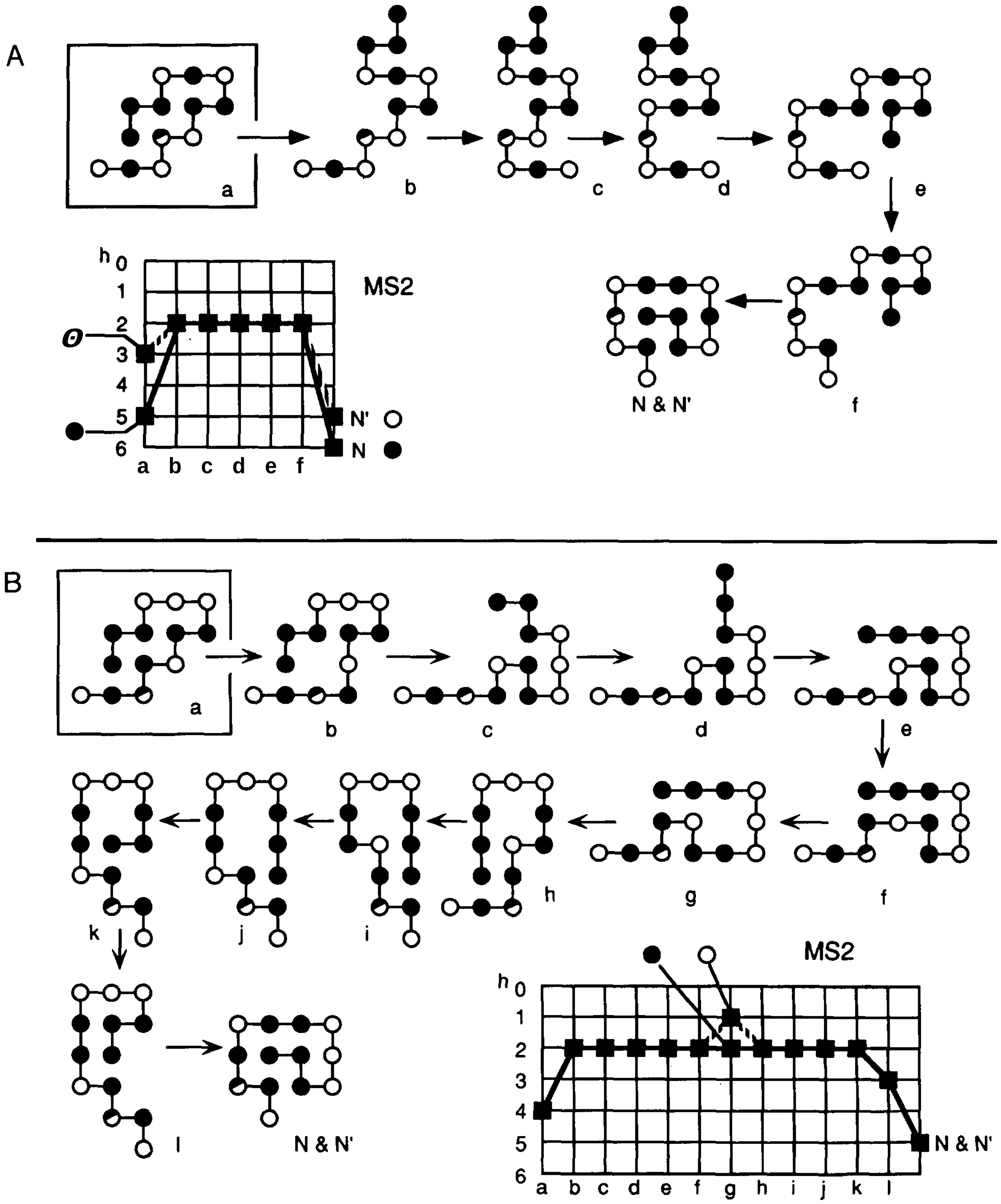

Fig. 49. Kinetic effects of mutations. Mutation sites are represented as half-filled circles. The H → P mutation in (A) decreases the barrier height, whereas the H → P mutation in (B) increases the barrier height. (From Chan and Dill [1994].)

as temperature and solvent (C.R. Matthews, 1993). External factors change the shapes, but not the sizes (i.e., numbers of degrees of freedom), of landscapes. Energy landscapes are flatter for denaturing conditions than for folding conditions. In the HP model, strongly denaturing conditions corresponds to an HH sticking energy, $\epsilon = 0$, for which all conformations are isoenergetic and the landscape is perfectly flat. To represent increasingly native conditions, ϵ is set increasingly negative (indicating

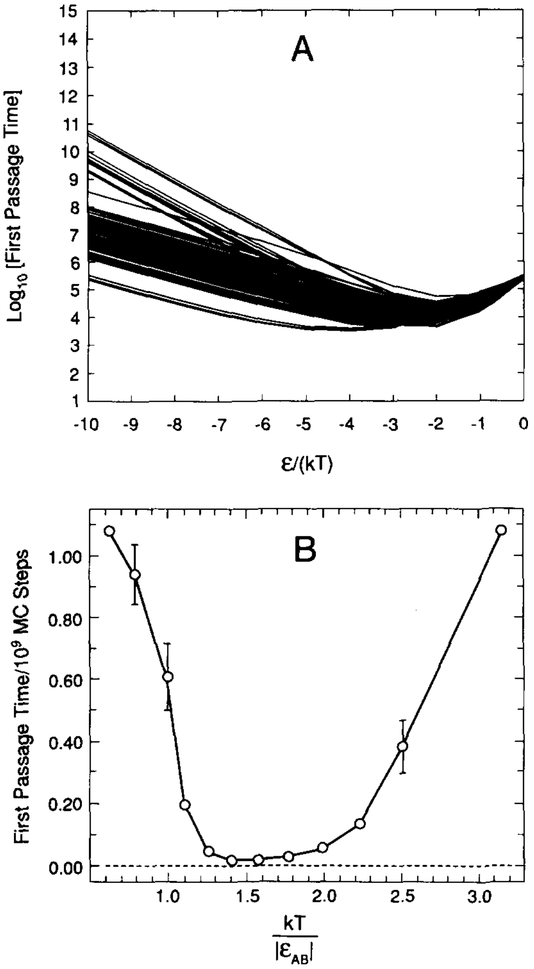

Fig. 50. Folding times. **A:** First passage time to reach native versus contact energy ϵ for all 13-monomer unique HP 2D sequences (Chan & Dill, 1994). Similar results were obtained in a more limited Monte Carlo study by Miller et al. (1992). **B:** Mean folding time in number of Monte Carlo (MC) steps versus temperature for one two-letter perturbed homopolymer sequence, from the 3D Monte Carlo simulation of Socci and Onuchic (1994). Error bars indicate standard deviation of the mean. Both plots show that there is an optimal range of intermediate contact energy or temperature at which first passage is fastest.

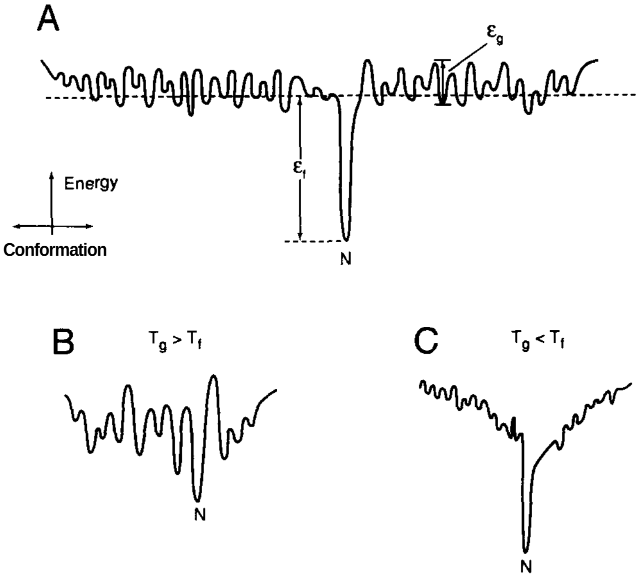

Fig. 51. Definitions of glass and folding temperatures in proteins. **A:** Folding temperature is defined by a folding energy, ϵ_f, representing the energy difference from the native state to some typical value of denatured state energy, taking into consideration also the entropy of the denatured state. Glass temperature is defined by an energy, ϵ_g, which is an average barrier height for kinetics. B: "Poor" folder: barriers are high relative to stability, $T_g > T_f$. Schematic drawings of energy landscapes of poor folders are given in Figure 37D and E. C: "Good" folder: barriers are small relative to stability, $T_f >$ T,. A schematic drawing of the energy landscape of a good folder is given in Figure 37C.

stronger HH attraction), and the landscape becomes increasingly rugged.

Studies of the HP model, using both simulation (Miller et al., 1992) and exact methods (Chan & Dill, 1994), and a study of a two-letter perturbed homopolymer model by Monte Carlo simulation (Socci & Onuchic, 1994) show that there is an optimal value of sticking energy (ϵ for the HH energy in the HP model) that maximizes the folding speed. Under strongly denaturing conditions, folding is slow for thermodynamic reasons (i.e., the native state is unstable), but under strong folding conditions, folding is slow for kinetic reasons. When $\epsilon = 0$ (golf-course), the search is essentially random, and the native state has the same very low probability of being populated as any other single conformation, of which there are a very large number, so the time required to access the native structure is very long. On the other hand, under strong HH sticking conditions, the search is highly directed toward the native state, but the kinetic traps are also very deep, so folding is slow because the time required to escape traps is prohibitive. Under intermediate sticking conditions, there is some direction toward the native state, but the barriers are surmountable, so the time needed to first arrive at the native structure (first passage time) is faster than in either of the extreme cases (see Fig. 50). Note that the first passage time, which is commonly used because it is easy to compute, is not equivalent to the folding time, more relevant to experiments. Experimental folding times also depend on native state stability, because they reflect the time required for a stable population of chains to reach the native state (Chan & Dill, 1994; Socci & Onuchic, 1994), as discussed below.

Folding speeds and barrier heights can be described in terms of a folding temperature, T_f, and a glass temperature, T,. The folding temperature, which can be defined for example as the midpoint of the equilibrium denaturation transition, is a simple measure of the folding free energy ϵ_f, $T_f - |\epsilon_f|/k$. When the temperature T > T_f, most molecules are denatured; when T < T_f, most molecules are native; and T = T_f is the temperature of equal native and denatured populations.

A glass is a system trapped in low-energy metastable states. The glass temperature for folding $T_g - |\epsilon_g|/k$ is defined by some average height $|\epsilon_g|$ of typical energy barriers (see Fig. 51). At high temperatures, T > T_g, the system can readily surmount barriers and traverse conformational space freely. At low temperatures, T < T_g, thermal energy is insufficient to cause the system to escape kinetic traps, and it behaves as a glass on the relevant time scale. These ideas originated in the random-energy and spin-glass models of Bryngelson and Wolynes (1987, 1989) and Goldstein et al. (1992a, 1992b) and have recently been developed more explicitly in lattice models by Socci and Onuchic (1994).

The value in defining folding and glass temperatures is that they provide a simple way to distinguish good (stable and fast folding) sequences from bad ones. Different amino acid sequences have different folding landscapes. T_f and T_g are properties of a landscape of a sequence. Good sequences have $T_f > T_g$: the barriers are small relative to the overall stabilization energy. Poor sequences have $T_f < T_g$. Consider what happens upon cooling, for both types of sequence. When the temperature T of a protein solution is higher than the intrinsic T_f and T_g, the protein is denatured. If a good sequence in solution is cooled to the point that $T_f > T > T_g$, then the native state is stable and the kinetic barriers are small enough that the system can find the native state. On the other hand, for a poor sequence, an intermediate temperature $T_f < T < T_g$ implies that even when the temperature is high enough to denature the molecules, it is still not high enough to surmount the kinetic barriers, so these molecules cannot achieve stable native populations in reasonable times (Goldstein et al., 1992a, 1992b; Socci & Onuchic, 1994). Examples of foldable and unfoldable perturbed homopolymer sequences are given by Socci and Onuchic (1994). Typical experimental T_f values are around 50–100 °C, whereas typical glass transition temperatures for good sequences of real proteins are not known but are likely to be below 0 °C.

In general, stability need not be related to folding speed, but for models studied so far, sequences that encode large energy gaps are fast folders, because they often have fewer deep kinetic traps (Shakhnovich & Gutin, 1993a; Bryngelson et al., 1995; Chan & Dill, 1994; Sali et al., 1994a, 1994b; Shakhnovich, 1994). Bryngelson et al. (1995) give an excellent review of this issue, and distinguish the "energy gap" from the related "stability gap." Contrary to the recent suggestion that a large energy gap is a "necessary and sufficient" condition to predict fast folders for one particular model (Shakhnovich & Gutin, 1993a; Sali et al., 1994a, 1994b; Shakhnovich, 1994), Figure 7 of Sali et al. (1994a) shows that the correlation is only weak – some "strongly folding sequences" have energy gaps as narrow as some "nonfolding sequences," indicating that designing folding speed into a sequence does not necessarily follow from designing stability (Camacho & Thirumalai, 1995; Chan, 1995). The studies of Šali et al. also raise other questions because: (1) deductions about stability were performed for a range of temperatures above those where the model proteins were actually stable (Chan, 1995), and (2) the "stability of the ground state" calculated by Shakhnovich and Gutin (1993a) and the "native concentration" calculated by Sali et al. (1994a) do not correspond to the true thermodynamic stability. They are based on a reference state that consists of only the maximally compact ensemble rather than the much larger true denatured ensemble (Chan & Dill, 1994; Socci & Onuchic, 1994). This leads to a considerable overestimation of the true folding temperature T_f of the model.

Some predictions of analytical spin-glass/random-energy models of proteins are in general agreement with results from simple exact models. They predict a rugged energy landscape, the possibility of a long-lived metastable glassy state (Bryngelson & Wolynes, 1987, 1989, 1990), and low average native degeneracies for certain random heteropolymers (Shakhnovich & Gutin, 1989a, 1990a, 1990b; Gutin & Shakhnovich, 1993; Wilbur & Liu, 1994). Spin-glass and related models of protein folding are reviewed in Karplus and Shakhnovich (1992), Bryngelson et al. (1995), Frauenfelder and Wolynes (1994), and Garel et al. (1995).

Critique of the models

We have reviewed the results of several simple exact models of proteins. At the level of generality treated here, the predictions of these models are largely in agreement, indicating that such predictions may be general and robust. But there are also important differences among the simple exact models we have reviewed. Here we compare them.

1. The HP model (Lau & Dill, 1989; Chan & Dill, 1991b; Shortle et al., 1992; Lattman et al., 1994) has the fewest parameters; it depends only on one quantity ϵ, the HH sticking energy. All other interactions (HP, PP) are zero relative to the solvated states of the monomers. For studies of native structures, ϵ is set to infinity; then the model has no parameters. This model has its physical basis in the dominance of the hydrophobic driving force (Dill, 1990). Having a single parameter has the advantage of simplicity. The obvious drawbacks are: chains are short, some studies are done in 2D, conformations are restricted to square or cubic lattices, other lattice geometries have not yet been explored, atomic detail is not included, and only a minimal set of interaction energies is considered.

But we regard the short-chain 2D model as also having one significant ***physical*** advantage over 3D models, a view that clearly requires some justification. Exact models, which are based on full conformational enumeration, have been restricted to short chains in either 2D or 3D. A principal factor in the physics of folding is the surface/volume ratio. To correctly model the exterior/interior ratio of myoglobin in 3D requires simulations of 150-mer chains, but in 2D requires simulations of only 16-mer chains (Chan & Dill, 1991b). Thus, we regard 2D studies of short chains as models of longer chains, whereas we regard 3D studies of short chains as models of short chains in 3D. For most properties we have tested, 2D and 3D models generally behave similarly. The use of hydrophobic zippers (Dill et al., 1993; Fiebig & Dill, 1993) and the CHCC search strategy (Yue & Dill, 1993, 1995) now allow studies of longer chains in 3D in the HP model. The HP model probably represents an extreme in the ruggedness and high glass transition temperature of an energy landscape and high degeneracy of native states because of its restriction to a two-letter alphabet.

2. The perturbed homopolymer model with independent intrachain contact interactions (Shakhnovich et al., 1991; Sali et al., 1994a, 1994b [the interactions among monomers i and j are defined by the quantities $B\{ij\}$]) is a model of 27-mers on cubic lattices, whose native structures are confined to a 3 × 3 x 3 cube. This model has the advantage of being 3D, and it is computationally tractable to find native states for certain forms of potential functions. All monomers are assumed to be strongly attracted to all others, so the physical basis for this model is different than hydrophobic and polar interactions. This model has two energy parameters: a mean attraction and a variance. A central feature of this model is that if the mean attraction is strong enough relative to the variance, then the native state of essentially any sequence is guaranteed to be maximally compact. In this way, the native state can be found by a search of only the 103,346 maximally compact conformations (Chan

and Dill, 1990a; Shakhnovich & Gutin, 1990a).[6] This is a "perturbed homopolymer" because the variation among monomers is small relative to the mean attraction. The drawbacks are: (1) The potential is not a good physical description of amino acids. Based on oil/water partitioning experiments, amino acids are not all strongly attracted to each other, nor are the variations small. (2) The denatured states are too many to be enumerated and have sometimes been erroneously estimated from only the maximally compact ensemble. Hence, stabilities are often not correctly estimated. (3) These models assume all native states are maximally compact, not accounting for variations in the overall shape of the native structure due to variations in sequence. (4) The contact interactions are assumed to be independent, unlike in real proteins (Chan, 1995; H.S. Chan & K.A. Dill, in prep.). Some of these models are also limited by their assumption that native states are maximally compact (Shakhnovich et al., 1991; Šali et al., 1994a, 1994b).

3. Perturbed homopolymers with two-letter codes have also been explored (Gutin & Shakhnovich, 1993; Shakhnovich & Gutin, 1993a; Socci & Onuchic, 1994). These authors use two-letter (A and B) sequences with relative energies of (AA, BB, AB) contacts set equal to (−3, −3, −1). All contacts are favorable, but contacts between the same types of monomers are more favorable. This is not a model for solvent-driven interactions: the monomers tend to phase separate into left and right domains (Fig. 2), rather than into an interior hydrophobic core and polar exterior, as proteins do. Two-letter sequences with energies of (AA, BB, AB) contacts equal to (−1, −1, 0) have been studied by O'Toole and Panagiotopoulos (1992).

The virtues of simplified exact models

Theoretical models need not mimic the atomic details of protein structures to be useful. The purposes of theoretical models are: (1) to extract essential principles, (2) to make testable and falsifiable predictions, and (3) to unify our understanding of the many different properties of a system. The protein models we describe are not microscopically accurate: proteins are treated as strings of beads, with discrete orientations determined by spatial lattices. Some of these lattice studies involve chains that are much shorter than real proteins (less than 20 monomers) and are sometimes configured only in two-dimensional space. They often use only two monomer types, rather than 20. Despite such shortcomings, these models offer some advantages:

1. Some properties cannot be predicted by other approaches. For example, we can study the folding code because we can study every possible sequence and the native conformation(s) of each one. It is not possible to explore sequence space broadly using models that have atomic resolution or 20 monomer types.

2. Exactness in a model is valuable. Models can be divided into two components: the physical model itself and the mathematical approximations required to study it. From the field of critical phenomena and phase transitions, it is known that physical principles can often be probed more deeply when the physics is appropriately simplified and the mathematics is accurate than when poor mathematical approximations are used to study an accurate model of the physics. For example, the Ising model is a simple lattice model widely used to study the physics of spins and magnetization, binary alloys, gases and liquids, and phase transitions and critical phenomena (Ising, 1925; Huang, 1987). An exact solution developed by Onsager produces very different behavior of the specific heat than was previously predicted by Bragg–Williams and Bethe mean-field approximations (Huang, 1987, chapters 16 and 17). The exact results are in good agreement with experiments despite the physical simplifications intrinsic to the model (Stanley, 1987). Keeping the number of parameters to a minimum allows us to understand the consequences of a model, rather than the consequences of the choices of parameters.

An exact model may predict genuine "surprises," but in ad hoc models, failures to agree with expectations can be dismissed as consequences of sparse sampling, inaccurate approximations, or adjustable parameters. In an exact model, predictions are direct consequences of the model. We can learn from their failures as well as from their successes. The idea that compactness in polymers stabilizes secondary structures was predicted from exact model studies (Chan & Dill, 1989a, 1989b, 1990a, 1990b). Because that result was not anticipated, it would have been difficult to recognize in a Monte Carlo simulation of a multiparameter model because any secondary structure observed in compact structures would have been attributed to hydrogen bonding or other terms. Once a new result is predicted, many other methods can confirm or reject it.

3. Models that involve the least microscopic detail and the greatest extraction of principle can teach us most broadly about how protein-like behavior is encodable in other types of chain molecules than proteins. If we study only models of 20 amino acids, we necessarily limit our understanding of foldable molecules to proteins.

4. Because simple exact models of proteins have the "folding problem"—very few native states in a conformational space that grows exponentially with the chain length—and because their global minima can be known exactly in some cases, they have been useful for testing conformational search algorithms (O'Toole & Panagiotopoulos, 1992; Fiebig & Dill, 1993; Unger & Moult, 1993; Stolorz, 1994).

5. Simple lattice models explicitly account for specific monomer sequences, chain connectivity, and excluded volume and are useful for testing analytical theories, such as mean-field treatments of heteropolymer collapse and spin-glass models, which often involve highly simplified approximations. For instance, simple exact models show how the "rugged landscape" envisioned in spin-glass treatments (Bryngelson & Wolynes, 1987, 1989) actually arises in a concrete model of chains (Camacho & Thirumalai, 1993a; Chan & Dill, 1993b, 1994).

Designing foldable polymers

What makes proteins special is less a matter of their monomer types and more a matter of their specific sequences

These model studies and related experiments (Blalock & Bost, 1986; Brunet et al., 1993; Kamtekar et al., 1993; Davidson & Sauer, 1994) imply that, at least at low resolution, protein structures and folding behavior may be encoded mainly in the order-

[6] **Because each maximally compact 27-mer conformation belongs to a set of 48 conformations related by rigid rotations and inversions, it is only necessary to enumerate 4,960,608/48 = 103,346 conformations.**

ing of hydrophobic and polar monomers along the chain, whereas helical and turn propensities and side-chain packing play a smaller role. Studies of structural databases indicate that other such factors, including helix end-capping, and propensities for glycines and prolines in turns, do contribute to the folding code of proteins. One example of the importance of other interactions is in leucine zippers. Although this involves an aggregation of multiple chains, rather than a folding of a single chain, studies with leucine zipper peptides show that: (1) they are held together principally by hydrophobic interactions, but (2) periodic repeat changes in steric packing can determine whether their 3D structure and hydrophobic "core" entails two, three, or four chains (Harbury et al., 1993). Hence, even if the sequence of hydrophobic and polar monomers is the major component of the folding code, it is surely not the only component.

Perhaps protein-like properties are designable into chain molecules that have monomers quite different from amino acids. This hypothesis has not yet been tested because polymer chemistry has lacked one of the most important capabilities available to biological syntheses: the ability to construct specific monomer sequences. Synthetic polymers have been either homopolymers or simple heteropolymers, with random sequences, alternating sequences (ABABABAB ...), or with "blocks" of monomer types (AAAABBBB ...). Until recently, the ability to construct specific monomer sequences has been possible only for biological molecules. But new methods for synthesizing specific monomer sequences (Simon et al., 1992; Cho et al., 1993) might now allow the design of other foldable polymers.

The folding code is not local

The protein folding problem has been referred to as the second half of the genetic code (King, 1989). The first half of the genetic code is like a dictionary: each amino acid in a protein is encoded as a specific triplet of nucleic acid bases in a DNA sequence. But to the extent that nonlocal interactions are dominant, simple exact model studies imply that the second half of the code, i.e., the encoding of the tertiary structure of a protein within its amino acid sequence, is more like a mystery novel: the full message is a global property. Change one or two amino acids, and the message does not change. Replace large sections of the sequence with others that retain the same general theme, and the message remains. In our view, the main information in the amino acid sequence is not primarily encoded in the relationships between each letter and its next neighbor in the sequence, but in the potential relationships for all the possible nonlocal pairings.

Summary

We have reviewed some principles deduced from simple exact protein models and related experiments. These models are based on two premises:

1. For some broadscale properties of proteins, it is more important to represent without bias the conformational and sequence spaces and less important to capture atomic details.

2. Proteins are polymers, free to distribute through large ensembles of possible conformations, but constrained by excluded volume, chain connectivity, and nearest-neighbor interactions, and specific monomer sequences. The dominant interactions are nonlocal and solvent mediated, and the folding code resides mainly in the arrangement of hydrophobic and polar monomers.

These model studies imply an alternative to the paradigm (primary → secondary → tertiary), which was based on an assumed primacy of local interactions, and from which it followed that: (1) folding kinetics was predicted to involve early and independent formation of helices and sheets followed by their assembly into tertiary structures, and (2) computer algorithms could be designed to predict native structures by first predicting secondary structures, which could then be assembled into tertiary structures. But the results reviewed here suggest instead how protein folding resembles a process of heteropolymer collapse, in which secondary structures are a consequence of folding, rather than its cause. In this view, secondary structures are not so much encoded within their ϕ–ψ propensities to have certain bond angles; they are more strongly encoded within the ability of a hydrophobic/polar sequence to form a good hydrophobic core, highly constrained by chain connectivity and steric exclusion. Only selected sequences will fold well. Proteins are neither homopolymers nor random heteropolymers; their specific sequences distinguish one chain fold, and one protein function, from another. It remains to be determined how the principles found in these simple models can be applied to the design and folding of proteins in more realistic models. Perhaps most interesting is the possible prospect of designing into completely different types of polymer molecules the ability to fold and function like proteins.

Acknowledgments

We thank Robert L. Baldwin, Th. Binkert, Bruce Bowler, Joe Bryngelson, Fred Cohen, Chris Dobson, Walter Englander, Tony Fink, Yuji Goto, Nick Pace, Oleg Ptitsyn, Fred Richards, J. Rička, Bob Sauer, Nick Socci, and Tobin Sosnick for helpful comments, and Danny Heap for technical and creative help with some of the figures. We thank the National Institutes of Health and the Office of Naval Research for financial support. P.D.T. is a Howard Hughes Medical Institute Predoctoral Fellow.

References

Abe H, Gō N. 1981. Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins. ***Biopolymers*** 20:1013–1031.

Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. 1987. Protein Data Bank. In: Allen FH, Bergerhoff G, Seivers R, eds. ***Crystallographic databases– Information content, software systems, scientific applications.*** Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography. pp 107–132

Alber T, Bell JA, Sun DP, Nicholson H, Wozniak JA, Cook S, Matthews BW. 1988. Replacements of Pro 86 in phage T4 lysozyme extend an α-helix but do not alter protein stability. ***Science 239***:631–635.

Alexandrescu AT, Evans PA, Pitkeathly M, Baum J, Dobson CM. 1993. Structure and dynamics of the acid-denatured molten globule state of a-lactalbumin: A two-dimensional NMR study. ***Biochemistry 32***:1707–1718.

Alexandrescu AT, Ng YL, Dobson CM. 1994. Characterization of a trifluoroethanol-induced partially folded state of a-lactalbumin. ***J Mol Biol 235***:587–599.

Alonso DOV, Dill KA. 1991. Solvent denaturation and stabilization of globular proteins. ***Biochemistry 30***:5974–5985.

Alonso DOV, Dill KA, Stigter D. 1991. The three states of globular proteins: Acid denaturation. ***Biopolymers*** 31:1631–1649.

Anfinsen CB, Scheraga HA. 1975. Experimental and theoretical aspects of protein folding. ***Adv Prot Chem 29***:205–300.

Anufrieva EV, Birshtein TM, Nekrasova TN, Ptitsyn OB, Sheveleva TV. 1968. The models of the denaturation of globular proteins. II. Hydrophobic interactions and conformational transition in polymethacrylic acid. ***J Polym Sci C 16***:3519–3531.

Baker D, Agard DA. 1994. Kinetics versus thermodynamics in protein folding. Biochemistry 33:7505–7509.

Baldwin EP, Hajiseyedjavadi O, Baase WA, Matthews BW. 1993. The role of backbone flexibility in the accommodations of variants that repack the core of T4 lysozyme. Science 262:1715–1718.

Baldwin RL. 1989. How does protein folding get started? Trends Biochem Sci 14:291–294.

Baldwin RL. 1993. Pulsed H/D-exchange studies of folding intermediates. Curr Opin Struct Biol 3:84–91.

Barber MN, Ninham BW. 1970. Random and restricted walks—Theory and applications. New York: Gordon and Breach.

Barrick D, Baldwin RL. 1993. Stein and Moore Award address. The molten globule intermediate of apomyoglobin and the process of protein folding. Protein Sci 2:869–876.

Bashford D, Chothia C, Lesk AM. 1987. Determinants of a protein fold. Unique features of the globin amino acid sequences. J Mol Biol 196:199–216.

Baum J, Dobson CM, Evans PA, Hanley C. 1989. Characterization of a partly folded protein by NMR methods: Studies on the molten globule state of guinea pig a-lactalbumin. Biochemistry 28:7–13.

Behe MJ, Lattman EE, Rose GD. 1991. The protein-folding problem: The native fold determines packing, but does packing determine the native fold? Proc Natl Acad Sci USA 88:4195–4199.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. J Mol Biol 112:535–542.

Betz SF, Raleigh DP, DeGrado WF. 1993. De-novo protein design—From molten globules to native-like states. Curr Opin Struct Biol 3:601–610.

Binder K, Young AP. 1986. Spin glasses: Experimental facts, theoretical concepts and open questions. Rev Mod Phys 58:801–976.

Binkert Th, Oberreich J, Meewes M, Nyffenegger R, Rifka J. 1991. Coil-globule transition of poly(N-isopropylacrylamide): A study of segment mobility by fluorescence depolarization. Macromolecules 24:5806–5810.

Blaber M, Zhang XJ, Lindstrom JD, Pepiot SD, Baase WA, Matthews BW. 1994. Determination of alpha-helix propensity within the context of a folded protein. Sites 44 and 131 in bacteriophage T4 lysozyme. J Mol Biol 235:600–624.

Blalock JE, Bost KL. 1986. Binding peptides that are specified by complementary RNAs. Biochem J 234:679–683.

Bowie JU, Reidhaar-Olson JF, Lim WA, Sauer RT. 1990. Deciphering the message in protein sequences: Tolerance to amino acid substitutions. Science 247:1306–1310.

Bowler BE, May K, Zaragoza T, York P, Dong A, Caughey WS. 1993. Destabilizing effects of replacing a surface lysine of cytochrome c with aromatic amino acids: Implications for the denatured state. Biochemistry 32:183–190.

Branden C, Tooze J. 1991. Introduction to protein structure. New York: Garland.

Brandts JF, Hunt L. 1967. The thermodynamics of protein denaturation. Ill. The denaturation of ribonuclease in water and in aqueous urea and aqueous ethanol mixtures. J Am Chem Soc 89:4826–4838.

Brandts JF, Hu CQ, Lin LN, Mos MT. 1989. A simple model for proteins with interacting domains. Applications to scanning calorimetry data. Biochemistry 28:8588–8596.

Briggs MS, Roder H. 1992. Early hydrogen-bonding events in the folding reaction of ubiquitin. Proc Natl Acad Sci USA 89:2017–2021.

Bromberg S, Dill KA. 1994. Side-chain entropy and packing in proteins. Protein Sci 3:997–1009.

Brunet AP, Huang ES, Huffine ME, Loeb JE, Weltman RJ, Hecht MH. 1993. The role of turns in the structure of an alpha-helical protein. Nature 364:355–358.

Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. 1995. Funnels, pathways and the energy landscape of protein folding: A synthesis. Proteins Struct Funct Genet 21:167–195.

Bryngelson JD, Wolynes PG. 1987. Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 84:7524–7528.

Bryngelson JD, Wolynes PG. 1989. Intermediates and barrier crossing in a random energy model with applications to protein folding. J Phys Chem 93:6902–6915.

Bryngelson JD, Wolynes PG. 1990. A simple statistical field theory of heteropolymer collapse with application to protein folding. Biopolymers 30:177–188.

Buck M, Radford SE, Dobson CM. 1993. A partially folded state of hen egg white lysozyme in trifluoroethanol: Structural characterization and implications for protein folding. Biochemistry 32:669–678.

Buck M, Radford SE, Dobson CM. 1994. Amide hydrogen exchange in a highly denatured state: Hen egg-white lysozyme in urea. J Mol Biol 237:247–254.

Calciano LJ, Escobar WA, Millhauser GL, Miick SM, Rubaloff J, Todd AP, Fink AL. 1993. Side-chain mobility of the β-lactamase A state probed by electron spin resonance spectroscopy. Biochemistry 32:5644–5649.

Camacho CJ, Thirumalai D. 1993a. Kinetics and thermodynamics of folding in model proteins. Proc Natl Acad Sci USA 90:6369–6372.

Camacho CJ, Thirumalai D. 1993b. Minimum energy compact structures of random sequences of heteropolymers. Phys Rev Lett 71:2505–2508.

Camacho CJ, Thirumalai D. 1995. Modeling the role of disulfide bonds in protein folding: Entropic barriers and pathways. Proteins Struct Funct Genet. Forthcoming.

Carra JH, Anderson EA, Privalov PL. 1994a. Thermodynamics of staphylococcal nuclease denaturation. 1. The acid-denatured state. Protein Sci 3:944–951.

Carra JH, Anderson EA, Privalov PL. 1994b. Thermodynamics of staphylococcal nuclease denaturation. 2. The A-state. Protein Sci 3:952–959.

Chacko S, Phillips GN Jr. 1992. Diffuse X-ray scattering from tropomyosin crystals. Biophys J 61:1256–1266.

Chaffotte AF, Cadieux C, Guillou Y, Goldberg ME. 1992a. A possible initial folding intermediate: The C-terminal proteolytic domain of tryptophan synthase beta chains folds in less than 4 milliseconds into a condensed state with non-native-like secondary structure. Biochemistry 31:4303–4308.

Chaffotte AF, Guillou Y, Goldberg ME. 1992b. Kinetic resolution of peptide bond and side chain far-UV circular dichroism during the folding of hen egg white lysozyme. Biochemistry 31:9694–9702.

Chakrabartty A, Kortemme T, Baldwin RL. 1994. Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. Protein Sci 3:843–852.

Chakrabartty A, Schellman JA, Baldwin RL. 1991. Large differences in the helix propensities of alanine and glycine. Nature 351:586–588.

Chan HS. 1995. Kinetics of protein folding. Nature 373:664–665.

Chan HS, Bromberg S, Dill KA. 1995. Models of cooperativity in protein folding. Philos Trans R Soc Lond. Forthcoming.

Chan HS, Dill KA. 1989a. Intrachain loops in polymers: Effects of excluded volume. J Chem Phys 90:492–509.

Chan HS, Dill KA. 1989b. Compact polymers. Macromolecules 22:4559–4573.

Chan HS, Dill KA. 1990a. The effects of internal constraints on the configurations of chain molecules. J Chem Phys 92:3118–3135.

Chan HS, Dill KA. 1990b. Origins of structure in globular proteins. Proc Natl Acad Sci USA 87:6388–6392.

Chan HS, Dill KA. 1991a. Polymer principles in protein structure and stability. Annu Rev Biophys Biophys Chem 20:447–490.

Chan HS, Dill KA. 1991b. "Sequence space soup" of proteins and copolymers. J Chem Phys 95:3775–3787.

Chan HS, Dill KA. 1993a. The protein folding problem. Phys Today 46(2): 24–32.

Chan HS, Dill KA. 1993b. Energy landscape and the collapse dynamics of homopolymers. J Chem Phys 99:2116–2127.

Chan HS, Dill KA. 1994. Transition states and folding dynamics of proteins and heteropolymers. J Chem Phys 100:9238–9257.

Chan HS, Dill KA, Shortle D. 1992. Statistical mechanics and protein folding. In: Bialek W, ed. Princeton lectures on biophysics. Singapore: World Scientific. pp 69–173.

Chen BL, Baase WA, Schellman JA. 1989. Low-temperature unfolding of a mutant of phage T4 lysozyme. 2. Kinetic investigations. Biochemistry 28:691–699.

Cho CY, Moran EJ, Cherry SR, Stephans JC, Fodor SP, Adams CL, Sundaram A, Jacobs JW, Schultz PG. 1993. An unnatural biopolymer. Science 261:1303–1305.

Chou PY, Fasman GD. 1974a. Conformational parameters for amino acids in helical, 0-sheet, and random coil regions calculated from proteins. Biochemistry 13:211–222.

Chou PY, Fasman GD. 1974b. Prediction of protein conformation. Biochemistry 13:222–245.

Chou PY, Wells M, Fasman GD. 1972. Conformational studies on copolymers of hydroxylpropyl-L-glutamine and L-leucine. Circular dichroism studies. Biochemistry 11:3028–3043.

Chyan CL, Wormald C, Dobson CM, Evans PA, Baum J. 1993. Structure and stability of the molten globule state of guinea-pig a-lactalbumin: A hydrogen exchange study. Biochemistry 32:5681–5691.

Covell DG, Jernigan RL. 1990. Conformations of folded proteins in restricted spaces. Biochemistry 29:3287–3294.

Creamer TP, Rose GD. 1992. Side-chain entropy opposes a-helix formation but rationalizes experimentally determined helix-forming propensities. Proc Natl Acad Sci USA 89:5937–5941.

Creighton TE. 1992. The disulfide folding pathway of BPTI. Science 256: 111–112.

Creighton TE. 1994. The energetic ups and downs of protein folding. Nature Struct *Biol 1*:135–138.

Damaschun G, Damaschun H, Gast K, Misselwitz R, Muller JJ, Pfeil W, Zirwer D. 1993. Cold denaturation-induced conformational changes in phosphoglycerate kinase from yeast. Biochemistry *32*:7739–7746.

Davidson AR, Sauer RT. 1994. Folded proteins occur frequently in libraries of random amino acid sequences. Proc *Natl* Acad *Sci* USA *91*:2146–2150.

de Gennes PG. 1975. Collapse of a polymer chain in poor solvents. *J Phys* Lett Paris *36*:L55–L57.

de Gennes PG. 1979. Scaling concepts in polymerphysics. Ithaca, New York: Cornell University Press.

DeGrado WF, Wasserman ZR, Lear JD. 1989. Protein design, a minimalist approach. Science *243*:622–628.

Deisenhofer J, Epp O, Miki K, Huber R, Michel H. 1985. Structure of the protein subunits in the photosynthetic reaction centre of *Rhodopseudomonas* viridis at 3 Å resolution. Nature *318*:618–624.

Derrida B. 1981. Random-energy model: An exact solvable model of disordered systems. Phys Rev B *24*:2613–2626.

des Cloizeaux J, Jannink G. 1990. Polymers in solution – Their *modelling* and structure. Oxford, UK: Clarendon Press.

Dill KA. 1985. Theory for the folding and stability of globular proteins. Biochemistry *24*:1501–1509.

Dill KA. 1987. The stabilities of globular proteins. In: Oxender DL, Fox CF, eds. Protein engineering. New York: Alan R. Liss, Inc. pp 187–192.

Dill KA. 1990. Dominant forces in protein folding. Biochemistry *29*:7133–7155.

Dill KA. 1993. Folding proteins: Finding a needle in a haystack. Curr Opin Struct *Biol 3*:99–103.

Dill KA, Alonso DOV, Hutchinson K. 1989. Thermal stabilities of globular proteins. Biochemistry *28*:5439–5449.

Dill KA, Fiebig KM, Chan HS. 1993. Cooperativity in protein folding kinetics. Proc *Natl* Acad Sci USA *90*:1942–1946.

Dill KA, Shortle D. 1991. Denatured states of proteins. *Annu* Rev Biochem *60*:795–825.

Dill KA, Stigter D. 1995. Modeling protein stability as heteropolymer collapse. Adv Protein Chem *46*:59–104.

Dobson CM. 1994. Protein folding: Solid evidence for molten globules. Curr *Biol 4*:636–640.

Dobson CM, Evans PA, Radford SE. 1994. Understanding how proteins fold: The lysozyme story so far. Trends Biochem Sci *19*:31–37.

Dufour E, Haertle T. 1990. Alcohol-induced changes of 0-lactoglobulin-retinol binding stoichiometry. Protein Eng *4*:185–190.

Dyson HJ, Merutka G, Waltho JP, Lerner RA, Wright PE. 1992a. Folding of peptide fragments comprising the complete sequence of proteins. Models for initiation of protein folding. I. Myohemerythrin. J Mol Biol *226*:795–817.

Dyson HJ, Sayre JR, Merutka G, Shin HC, Lerner RA, Wright PE. 1992b. Folding of peptide fragments comprising the complete sequence of proteins. Models for initiation of protein folding. II. Plastocyanin. *J Mol Biol 226*:819–835.

Edwards SF, Anderson PW. 1975. Theory of spin glasses. J Phys F 5: 965–974.

Elove GA, Bhuyan AK, Roder H. 1994. Kinetic mechanism of cytochrome c folding: lnvolvement of the heme and its ligands. Biochemistry 33: 6925–6935.

Elove GA, Chaffotte AF, Roder H, Goldberg ME. 1992. Early steps in cytochrome c folding probed by time-resolved circular dichroism and fluorescence spectroscopy. Biochemistry 31:6876–6883.

Engh RA, Dieckmann T, Bode W, Auerswald EA, Turk V, Huber R, Oschkinat H. 1993. Conformational variability of chicken cystatin. Comparison of structures determined by X-ray diffraction and NMR spectroscopy. J Mol *Biol 234*:1060–1069.

Englander SW. 1993. In pursuit of protein folding. Science *262*:848–849.

Englander SW, Mayne L. 1992. Protein folding studied using hydrogen-exchange labeling and two-dimensional NMR. *Annu* Rev Biophys *Biomol Structure*:243–265.

Evans PA, Topping KD, Woolfson DN, Dobson CM. 1991. Hydrophobic clustering in nonnative states of a protein: Interpretation of chemical shifts in NMR spectra of denatured states of lysozyme. Proteins Struct Funct Genet *9*:248–266.

Eyring H, Stearn AE. 1939. The application of the theory of absolute reaction rates to proteins. Chem Rev *24*:253–270.

Faber HR, Matthews BW. 1990. A mutant T4 lysozyme displays five different crystal conformations. Nature *348*:263–266.

Fan P, Bracken C, Baum J. 1993. Structural characterization of monellin in the alcohol-denatured state by NMR: Evidence for 0-sheet to a-helix conversion. Biochemistry *32*:1573–1582.

Fauchère JL, Pliska V. 1983. Hydrophobic parameters a of amino acid side chains from the partitioning of N-acetyl-amino acid amides. Eur J Med Chem Ther Chem *18*:369–375.

Feng HP, Widom J. 1994. Kinetics of compaction during lysozyme refolding studied by continuous-flow quasielastic light scattering. Biochemistry 33:13382–13390.

Feng YQ, Wand AJ, Sligar SG. 1994. Solution structure of apocytochrome *b*562. Nature Struct Biol 1:30–35.

Fersht AR. 1993. Protein folding and stability: The pathway of folding of barnase. FFBS Lett *325*:5–16.

Fiebig KM, Dill KA. 1993. Protein core assembly processes. J Chem Phys *98*:3475–3487.

Fink AL. 1995. Compact intermediate states in protein folding. *Annu* Rev Biophys *Biomol Struc*. Forthcoming.

Finkelstein AV, Reva BA. 1991. A search for the most stable folds of protein chains. Nature 351:497–499.

Fischer KH, Hertz JA. 1991. Spin glasses. Cambridge, UK: Cambridge University Press.

Flanagan JM, Kataoka M, Shortle D, Engelman DM. 1992. Truncated staphylococcal nuclease is compact but disordered. Proc Natl Acad *Sci* USA *89*:748–752.

Flanagan JM, Kataoka M, Fujisawa T, Engelman DM. 1993. Mutations can cause large changes in the conformation of a denatured protein. Biochemistry *32*:10359–10370.

Frauenfelder H, Alberding NA, Ansari A, Braunstein D, Cowen BR, Hong MK, lben IET, Johnson JB, Luck S, Marden MC, Mourant JR, Ormos P, Reinisch L, Scholl R, Schulte A, Shyamsunder E, Sorensen LB, Steinbach PJ, Xie A, Young RD, Yue KT. 1990. Proteins and pressure. *J Phys* Chem *94*:1024–1037.

Frauenfelder H, Sligar SG, Wolynes PG. 1991. The energy landscapes and motions of proteins. Science 254:1598–1603.

Frauenfelder H, Wolynes PG. 1994. Biomolecules: Where the physics of complexity and simplicity meet. Phys Today *47*(2):58–64.

Freed KF. 1987. Renormalization group theory of *macromolecules*. New York: Wiley.

Fujishige S, Kubota K, Anto I. 1989. Phase transition of aqueous solutions of poly(*N*-isopropylacrylamide) and poly(*N*-isopropylmethacrylamide). J Phys Chem *93*:3311–3313.

Garel T, Orland H. 1988. Mean-field model for protein folding. Europhys Lett *6*:307–310.

Garel T, Orland H, Thirumalai D. 1995. Analytical theories of protein folding. In: Elber R, ed. New developments in theoretical studies of proteins. Singapore: World Scientific Press.

Garvey EP, Swank J, Matthews CR. 1989. A hydrophobic cluster forms early in the folding of dihydrofolate reductase. Proteins Struct Funct Genet *6*:259–266.

Gast K, Damaschun G, Damaschun H, Misselwitz R, Zirwer D. 1993. Cold denaturation of yeast phosphoglycerate kinase: Kinetics of changes in secondary structure and compactness on unfolding and refolding. Biochemistry *32*:7747–7752.

Gilmanshin RI, Ptitsyn OB. 1987. An early intermediate of refolding α-lactalbumin forms within 20 ms. *FEBS* Lett *223*:327–329.

Gō N, Abe H. 1981. Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation. Biopolymers *20*:991–1011.

Go N, Taketomi H. 1978. Respective roles of short- and long-range interactions in protein folding. Proc *Natl* Acad Sci USA *75*:559–563.

Goldenberg DP, Creighton TE. 1985. Energetics of protein structure and folding. Biopolymers 24:167–182.

Goldstein RA, Luthey-Schulten ZA, Wolynes PG. 1992a. Optimal protein-folding codes from spin-glass theory. Proc *Natl* Acad Sci USA *89*:4918–4922.

Goldstein RA, Luthey-Schulten ZA, Wolynes PG. 1992b. Protein tertiary structure recognition using optimized hamiltonians with local interactions. Proc Natl Acad Sci USA *89*:9029–9033.

Goodsell DS, Olson AJ. 1993. Soluble proteins: Size, shape and function. Trends Biochem Sci *18*:65–68.

Goto Y, Fink AL. 1990. Phase diagram for acidic conformational states of apornyoglobin. J Mol *Biol 214*:803–805.

Goto Y, Hagihara Y, Hamada D, Hoshino M, Nishii I. 1993. Acid-induced unfolding and refolding transitions of cytochrome c: A three-state mechanism in H_2O and D_2O. Biochemistry *32*:11878–11885.

Goto Y, Nishikiori S. 1991. Role of electrostatic repulsion in the acidic molten globule of cytochrome c. J Mol *Biol 222*:679–686.

Gregoret LM, Cohen FE. 1991. Protein folding: Effect of packing density on chain conformation. J Mol *Biol 219*:109–122.

Gronenborn AM, Clore GM. 1994. Experimental support for the "hydrophobic zipper" hypothesis. Science *263*:536.

Grosberg AYu, Khokhlov AR. 1987. Physics of phase transitions in solutions of macromolecules. Sov Sci Rev A *8*:147–258.

Grosberg AYu, Shakhnovich EI. 1986. Theory of phase transitions of the coil-globule type in a heteropolymer chain with disordered sequence of links. *Zh Exp Teor Fiz (USSR) 91*:2159–2170.

Guo Z, Thirumalai D, Honeycutt JD. 1992. Folding kinetics of proteins: A model study. *J Chem Phys 97*:525–535.

Gupta P, Hall CK. 1995. Computer simulation studies of protein refolding pathways and intermediates. *Am Inst Chem Eng J*. Forthcoming.

Gutin AM, Shakhnovich EI. 1993. Ground state of random copolymers and the discrete random energy model. *J Chem Phys 98*:8174–8177.

Hagihara Y, Tan Y, Goto Y. 1994. Comparison of the conformational stability of the molten globule and native states of horse cytochrome *c*. Effects of acetylation, heat, urea and guanidine-hydrochloride. *J Mol Biol 237*:336–348.

Hamada D, Hoshino M, Kataoka M, Fink AL, Goto Y. 1993. Intermediate conformational states of apocytochrome *c*. *Biochemistry 32*:10351–10358.

Handel TM, Williams SA, DeGrado WF. 1993. Metal ion-dependent modulation of the dynamics of a designed protein. *Science 261*:879–885.

Hao MH, Rackovsky S, Liwo A, Pincus MR, Scheraga HA. 1992. Effects of compact volume and chain stiffness on the conformations of native proteins. *Proc Natl Acad Sci USA 89*:6614–6618.

Harbury PB, Zhang T, Kim PS, Alber T. 1993. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science 262*:1401–1407.

Harding MM, Williams DH, Woolfson DN. 1991. Characterization of a partially denatured state of a protein by two-dimensional NMR: Reduction of the hydrophobic interactions in ubiquitin. *Biochemistry 30*:3120–3128.

Harpaz Y, Gerstein M, Chothia C. 1994. Volume changes on protein folding. *Structure 2*:641–649.

Harper ET, Rose GD. 1993. Helix stop signals in proteins and peptides: The capping box. *Biochemistry 32*:7605–7609.

Harrison SC, Durbin R. 1985. Is there a single pathway for the folding of a polypeptide chain? *Proc Natl Acad Sci USA 82*:4028–4030.

Havel TF. 1990. The sampling properties of some distance geometry algorithms applied to unconstrained computed conformations. *Biopolymers 29*:1565–1585.

Hecht MH, Richardson JS, Richardson DC, Ogden RC. 1990. De novo design, expression, and characterization of felix: A four-helix bundle protein of native-like sequence. *Science 249*:884–891.

Heinz DW, Baase WA, Matthews BW. 1992. Folding and function of a T4 lysozyme containing 10 consecutive alanines illustrate the redundancy of information in an amino acid sequence. *Proc Natl Acad Sci USA 89*:3751–3755.

Hill CP, Anderson DH, Wesson L, DeGrado WF, Eisenberg D. 1990. Crystal structure of alpha 1: Implications for protein design. *Science 249*: 543–546.

Hinds DA, Levitt M. 1992. A lattice model for protein structure prediction at low resolution. *Proc Natl Acad Sci USA 89*:2536–2540.

Honeycutt JD, Thirumalai D. 1990. Metastability of the folded states of globular proteins. *Proc Natl Acad Sci USA 87*:3526–3529.

Honeycutt JD, Thirumalai D. 1992. The nature of folded states of globular proteins. *Biopolymers 32*:695–709.

Hooke SD, Radford SE, Dobson CM. 1994. The refolding of human lysozyme: A comparison with the structurally homologous hen lysozyme. *Biochemistry 33*:5867–5876.

Hua QX, Kochoyan M, Weiss MA. 1992. Structure and dynamics of despentapeptide-insulin in solution: The molten-globule hypothesis. *Proc Natl Acad Sci USA 89*:2379–2383.

Hua QX, Ladbury JE, Weiss MA. 1993. Dynamics of a monomeric insulin analogue: Testing the molten-globule hypothesis. *Biochemistry 32*:1433–1442.

Huang K. 1987. *Statistical mechanics, 2nd ed*. New York: Wiley.

Hughson FM, Barrick D, Baldwin RL. 1991. Probing the stability of a partly folded apomyoglobin intermediate by site-directed mutagenesis. *Biochemistry 30*:4113–4118.

Hughson FM, Wright PE, Baldwin RL. 1990. Structural characterization of a partly folded apomyoglobin intermediate. *Science 249*:1544–1548.

Hunt NG, Gregoret LM, Cohen FE. 1994. The origins of protein secondary structure: Effects of packing density and hydrogen bonding studied by a fast conformational search. *J Mol Biol 241*:214–225.

Ikai A, Tanford C. 1971. Kinetic evidence for incorrectly folded intermediate states in the refolding of denatured proteins. *Nature 230*:100–102.

Ilyina E, Mayo KH. 1995. Multiple native-like conformations trapped via self-association-induced hydrophobic collapse of the 33 residue β-sheet domain from platelet factor-4. *Biochem J 306*:407–419.

Ilyina E, Milius R, Mayo KH. 1994. Synthetic peptides probe folding initiation sites in platelet factor-4: Stable chain reversal found within the hydrophobic sequence LIATLKNGRKISL. *Biochemistry 33*:13436–13444.

Ising E. 1925. The theory of ferromagnetism. *Z Phys 31*:253–258.

Itzhaki LS, Evans PA, Dobson CM, Radford SE. 1994. Tertiary interactions in the folding pathway of hen lysozyme – Kinetic studies using fluorescent probes. *Biochemistry 33*:5212–5220.

Iwakura M, Jones BE, Falzone CJ, Matthews CR. 1993. Collapse of parallel folding channels in dihydrofolate reductase from *Escherichia coli* by site-directed mutagenesis. *Biochemistry 32*:13566–13574.

Jackson SE, el Masry N, Fersht AR. 1993. Structure of the hydrophobic core in the transition state for folding of chymotrypsin inhibitor 2: A critical test of the protein engineering method of analysis. *Biochemistry 32*: 11270–11278.

Jeng MF, Englander SW. 1991. Stable submolecular folding units in a noncompact form of cytochrome *c*. *J Mol Biol 221*:1045–1061.

Jeng MF, Englander SW, Elöve GA, Wand AJ, Roder H. 1990. Structural description of acid-denatured cytochrome *c* by hydrogen exchange and 2D NMR. *Biochemistry 29*:10433–10437.

Jennings PA, Finn BE, Jones BE, Matthews CR. 1993. A reexamination of the folding mechanism of dihydrofolate reductase from *Escherichia coli*: Verification and refinement of a four-channel model. *Biochemistry 32*: 3783–3789.

Jennings PA, Wright PE. 1993. Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. *Science 262*:892–896.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Kaiser CA, Preuss D, Grisafi P, Botstein D. 1987. Many random sequences functionally replace the secretion signal sequence of yeast invertase. *Science 235*:312–317.

Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH. 1993. Protein design by binary patterning of polar and nonpolar amino acids. *Science 262*:1680–1685.

Karplus M, Shakhnovich E. 1992. Protein folding: Theoretical studies of thermodynamics and dynamics. In: Creighton T, ed. *Protein folding*. New York: Freeman. pp 127–195.

Karplus M, Weaver DL. 1976. Protein-folding dynamics. *Nature 260*:404–406.

Karplus M, Weaver DL. 1994. Protein folding dynamics: The diffusion-collision model and experimental data. *Protein Sci 3*:650–668.

Kauzmann W. 1954. Denaturation of proteins and enzymes. In: McElroy WD, Glass B, eds. *The mechanism of enzyme action*. Baltimore: Johns Hopkins Press. pp 70–120.

Kauzmann W. 1959. Some factors in the interpretation of protein denaturation. *Adv Protein Chem 14*:1–63.

Kendrew JC, Dickerson RE, Strandberg BE, Hart RG, Davies DR. 1960. Structure of myoglobin: A three-dimensional fourier synthesis at 2 Å resolution. *Nature 185*:422–427.

Killian JA. 1992. Gramicidin and gramicidin-lipid interactions. *Biochim Biophys Acta 1113*:391–425.

Kim PS, Baldwin RL. 1982. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu Rev Biochem 51*:459–489.

King J. 1989. Deciphering the rules of protein folding. *Chem Eng News 67*:32–54.

Kolinski A, Skolnick J. 1994. Monte-Carlo simulations of protein folding. 1. Lattice model and interaction scheme. *Proteins Struct Funct Genet 18*:338–352.

Kosen PA, Marks CB, Falick AM, Anderson S, Kuntz ID. 1992. Disulfide bond-coupled folding of bovine pancreatic trypsin inhibitor derivatives missing one or two disulfide bonds. *Biochemistry 31*:5705–5717.

Kuwajima K. 1989. The molten globule state as a clue for understanding the folding and cooperativity of globular protein structure. *Proteins Struct Funct Genet 6*:87–103.

Kuwajima K. 1992. Protein folding in vitro. *Curr Opin Biotechnol 3*:462–467.

Kuwajima K, Garvey EP, Finn BE, Matthews CR, Sugai S. 1991. Transient intermediates in the folding of dihydrofolate reductase as detected by far-ultraviolet circular dichroism spectroscopy. *Biochemistry 30*:7693–7703.

Lattman EE, Fiebig KM, Dill KA. 1994. Modeling compact denatured states of proteins. *Biochemistry 33*:6158–6166.

Lau KF, Dill KA. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules 22*:3986–3997.

Lau KF, Dill KA. 1990. Theory for protein mutability and biogenesis. *Proc Natl Acad Sci USA 87*:638–642.

Lawrence C, Auger I, Mannella C. 1987. Distribution of accessible surfaces of amino acids in globular proteins. *Proteins Struct Funct Genet 2*: 153–161.

Lee B, Richards FM. 1971. The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol 55*:379–400.

Lee C, Subbiah S. 1991. Prediction of protein side-chain conformation by packing optimization. *J Mol Biol 217*:373–388.

Leopold PE, Montal M, Onuchic JN. 1992. Protein folding funnels: A kinetic approach to the sequence-structure relationship. Proc Natl Acad Sci USA 89:8721-8725.

Levinthal C. 1968. Are there pathways for protein folding? J Chim Phys 65:44-45.

Levitt M, Chothia C. 1976. Structural patterns in globular proteins. Nature 261:552-558.

Li A, Daggett V. 1994. Characterization of the transition state of protein unfolding by use of molecular dynamics: Chymotrypsin inhibitor 2. Proc Natl Acad Sci USA 91:10430-10434.

Lim WA, Farruggio DC, Sauer RT. 1992. Structural and energetic consequences of disruptive mutations in a protein core. Biochemistry 31:4324-4333.

Lim WA, Sauer RT. 1991. The role of internal packing interactions in determining the structure and stability of a protein. J Mol Biol 219:359-376.

Lipman DJ, Wilbur WJ. 1991. Modelling neutral and selective evolution of protein folding. Proc R Soc Lond B 245:7-11.

Lu J, Dahlquist FW. 1992. Detection and characterization of an early folding intermediate of T4 lysozyme using pulsed hydrogen exchange and two-dimensional NMR. Biochemistry 31:4749-4756.

Lumb KJ, Kim PS. 1994. Formation of a hydrophobic cluster in denatured bovine pancreatic trypsin inhibitor. J Mol Biol 236:412-420.

Lyu PC, Liff MI, Marky LA, Kallenbach NR. 1990. Side chain contributions to the stability of or-helical structure in peptides. Science 250:669-673.

Lyu PC, Wemmer DE, Zhou HX, Pinker RJ, Kallenbach NR. 1993. Capping interactions in isolated alpha helices: Position-dependent substitution effects and structure of a serine-capped peptide helix. Biochemistry 32:421-425.

Makhatadze GI, Privalov PL. 1993. Contribution of hydration to protein folding thermodynamics. I. The enthalpy of hydration. J Mol Biol 232:660-679.

Matthews BW. 1987. Genetic and structural analysis of the protein stability problem. Biochemistry 26:6885-6888.

Matthews BW. 1993. Structural and genetic analysis of protein stability. Annu Rev Biochem 62:139-160.

Matthews CR. 1993. Pathways of protein folding. Annu Rev Biochem 62:653-683.

Matthews CR, Hurle MR. 1987. Mutant sequences as probes of protein folding mechanisms. BioEssays 6:254-257.

Maynard Smith J. 1970. Natural selection and the concept of a protein space. Nature 225:563-564.

Meewes M, Rička J, de Silva R, Nyffenegger R, Binkert Th. 1991. Coil-globule transition of poly(N-isopropylacrylamide). A study of surfactant effects by light scattering. Macromolecules 24:5811-5816.

Merz KM Jr, LeGrand SM, eds. 1994. The protein folding problem and tertiary structure prediction. Boston: Birkhauser.

Mezard M, Parisi G, Virasoro MA. 1986. Spin glass theory and beyond. Singapore: World Scientific.

Miller R, Danko CA, Fasolka J, Balazs AC, Chan HS, Dill KA. 1992. Folding kinetics of proteins and copolymers. J Chem Phys 96:768-780.

Miller S, Janin J, Lesk AM, Chothia C. 1987. Interior and surface of monomeric proteins. J Mol Biol 196:641-656.

Miranker A, Robinson CV, Radford SE, Aplin RT, Dobson CM. 1993. Detection of transient protein folding populations by mass spectrometry. Science 262:896-900.

Mirsky AE, Pauling L. 1936. On the structure of native, denatured, and coagulated proteins. Proc Natl Acad Sci USA 22:439-447.

Montelione GT, Scheraga HA. 1989. Formation of local structures in protein folding. Acc Chem Res 22:70-76.

Morozov VN, Morozova TYa. 1993. Elasticity of globular proteins. The relation between mechanics, thermodynamics and mobility. J Biomol Struct & Dyn 11:459-481.

Munson M, O'Brien R, Sturtevant JM, Regan L. 1994. Redesigning the hydrophobic core of a four-helix-bundle protein. Protein Sci 3:2015-2022.

Nelson JW, Kallenbach NR. 1986. Stabilization of the ribonuclease S-peptide a-helix by trifluoroethanol. Proteins Struct Funct Genet 1:211-217.

Neri D, Billeter M, Wider G, Wuthrich K. 1992. NMR determination of residual structure in a urea-denatured protein, the 434-repressor. Science 257:1559-1563.

Nishii I, Kataoka M, Tokunaga F, Goto Y. 1994. Cold denaturation of the molten globule states of apomyoglobin and a profile for protein folding. Biochemistry 33:4903-4909.

Nozaki Y, Tanford C. 1971. The solubility of amino acids and two glycine polypeptides in aqueous ethanol and dioxane solutions. J Biol Chem 246:2211-2217.

O'Neil KT, DeGrado WF. 1990. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. Science 250:646-651.

Orr WJC. 1947. Statistical treatment of polymer solutions at infinite dilution. Trans Faraday Soc 43:12-27.

O'Toole EM, Panagiotopoulos AZ. 1992. Monte Carlo simulation of folding transitions of simple model proteins using a chain growth algorithm. J Chem Phys 97:8644-8652.

Padmanabhan S, York EJ, Gera L, Stewart JM, Baldwin RL. 1994. Helix-forming tendencies of amino acids in short hydroxybutyl-L-glutamine peptides—An evaluation of the contradictory results from host-guest studies and short alanine-based peptides. Biochemistry 33:8604-8609.

Pakula AA, Sauer RT. 1990. Reverse hydrophobic effects relieved by amino-acid substitutions at a protein surface. Nature 344:363-364.

Palleros DR, Shi L, Reid KL, Fink AL. 1993. Three-state denaturation of DnaK induced by guanidine hydrochloride. Evidence for an expandable intermediate. Biochemistry 32:4314-4321.

Pauling L, Corey RB. 1951a. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. Proc Natl Acad Sci USA 37:235-240.

Pauling L, Corey RB. 1951b. The pleated sheet, a new layer configuration of polypeptide chains. Proc Natl Acad Sci USA 37:251-256.

Pauling L, Corey RB. 1951c. The structure of fibrous proteins of the collagen-gelatin group. Proc Natl Acad Sci USA 37:272-281.

Pauling L, Corey RB. 1951d. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. Proc Natl Acad Sci USA 37:729-740.

Pauling L, Corey RB, Branson HR. 1951. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci USA 37:205-211.

Peng Z, Kim PS. 1994. A protein dissection study of a molten globule. Biochemistry 33:2136-2141.

Pickett SD, Sternberg MJ. 1993. Empirical scale of side-chain conformational entropy in protein folding. J Mol Biol 231:825-839.

Pinker RJ, Lin L, Rose GD, Kallenbach NR. 1993. Effects of alanine substitutions in a-helices of sperm whale myoglobin on protein stability. Protein Sci 2:1099-1105.

Poland DC, Scheraga HA. 1970. Theory of the helix-coil transition. New York: Academic Press.

Ponder JW, Richards FM. 1987. Tertiary templates for proteins. J Mol Biol 193:775-791.

Post CB, Zimm BH. 1979. Internal condensation of a single DNA molecule. Biopolymers 18:1487-1501.

Privalov PL. 1979. Stability of proteins: Small globular proteins. Adv Protein Chem 33:167-241.

Privalov PL. 1982. Stability of proteins. Proteins which do not present a single cooperative system. Adv Protein Chem 35:1-104.

Privalov PL, Gill SJ. 1988. Stability of protein structure and hydrophobic interaction. Adv Protein Chem 39:191-234.

Privalov PL, Makhatadze GI. 1993. Contribution of hydration to protein folding thermodynamics. II. The entropy and Gibbs energy of hydration. J Mol Biol 232:660-679.

Privalov PL, Tiktopulo EI, Venyaminov SYu, Griko YuV, Makhatadze GI, Khechinashvili NN. 1989. Heat capacity and conformation of proteins in the denatured state. J Mol Biol 205:737-750.

Ptitsyn OB. 1987. Protein folding: Hypotheses and experiments. J Protein Chem 6:273-293.

Ptitsyn OB. 1991. How does protein synthesis give rise to the 3D-structure? FEBS Lett 285:176-181.

Ptitsyn OB. 1992. The molten globule state. In: Creighton T, ed. Protein folding. New York: Freeman. pp 243-300.

Ptitsyn OB, Kron AK, Eizner YuYe. 1968. The models of the denaturation of globular proteins. 1. Theory of globule-coil transitions in macromolecules. J Polym Sci C 16:3509-3517.

Ptitsyn OB, Lim VI, Finkelstein AV. 1972. Secondary structure of globular proteins and the principle of concordance of local and long-range interactions. In: Hemker HC, Hess B, eds. Analysis and simulation of biochemical systems. Proceedings of the Eighth FEBS Meeting. Amsterdam: North-Holland. pp 421-429.

Ptitsyn OB, Pain RH, Semisotnov GV, Zerovnik E, Razgulyaev 01. 1990. Evidence for a molten globule state as a general intermediate in protein folding. FEBS Lett 262:20-24.

Ptitsyn OB, Semisotnov GV. 1991. The mechanism of protein folding. In: Nall BT, Dill KA, eds. Conformations and forces in protein folding. Washington, D.C.: AAAS. pp 155-168.

Radford SE, Dobson CM, Evans PA. 1992. The folding of hen lysozyme involves partially structured intermediates and multiple pathways. Nature 358:302-307.

Rao SP, Carlstrom DE, Miller WG. 1974. Collapsed structure polymers. A scattergun approach to amino acid copolymers. Biochemistry 13:943-952.

Redfield C, Smith RAG, Dobson, CM. 1994. Structural characterization of a highly-ordered molten globule at low pH. Nature Struct Biol 1:23-29.

Reed J, Kinzel V. 1993. Primary structure elements responsible for the conformational switch in the envelope glycoprotein GP120 from human immunodeficiency virus type 1: LPCR is a motif governing folding. Proc *Natl* Acad Sci USA *90*:6761-6765.

Regan L, DeGrado WF. 1988. Characterization of a helical protein designed from first principles. Science *241*:976-978.

Reidhaar-Olson JF, Bowie JU, Breyer RM, Hu JC, Knight KL, Lim WA, Mossing MC, Parsell DA, Shoemaker KR, Sauer RT. 1991. Random mutagenesis of protein sequences using oligonucleotide cassettes. Methods *Enzymol 208*:564-586.

Reidhaar-Olson JF, Sauer RT. 1988. Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. Science 241:53-57.

Richards FM. 1974. The interpretation of protein structures: Total volume, group volume distributions and packing density. J Mol *Biol 82*:1-14.

Richards FM. 1977. Areas, volumes, packing and protein structure. *Annu* Rev Biophys Bioeng 6:151-176.

Richards FM. 1992. Folded and unfolded proteins—An introduction. In: Creighton T. ed. Protein folding. New York: Freeman. pp 1-58.

Richards FM, Lim W. 1993. An analysis of packiñg in the protein folding problem. Q Rev Biophys *26*:423-498.

Richardson JS. 1981. The anatomy and taxonomy of protein structure. Adv Protein Chem *34*:167-339.

Richardson JS, Richardson DC. 1989. The de novo design of protein structures. Trends *Biochem* Sci *14*:304-309.

Rifka J, Meewes M, Nyffenegger R, Binkert T. 1990. Intermolecular and intramolecular solubilization: Collapse and expansion of a polymer chain in surfactant solutions. Phys Rev Lett *65*:657-660.

Robertson AD, Baldwin RL. 1991. Hydrogen exchange in thermally denatured ribonuclease A. Biochemistry *30*:9907-9914.

Roder H, Elöve GA, Englander SW. 1988. Structural characterization of folding intermediates in cytochrome c by H-exchange labeling and proton NMR. Nature *335*:700-704.

Rooman MJ, Wodak SJ. 1988. Identification of predictive sequence motifs limited by protein structure data base size. Nature *335*:45-49.

Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science 229*:834-838.

Rosenblatt M, Beaudette NV, Fasman GD. 1980. Conformational studies of the synthetic precursor-specificregion of preproparathyroid hormone. Proc Natl Acad Sci USA *77*:3983-3987.

Saab-Rincon G, Froebe CL, Matthews CR. 1993. Urea-induced unfolding of the a subunit of tryptophan synthase: One-dimensional proton NMR evidence for residual structure near histidine-92 at high denaturant concentration. *Biochemistry* 32:13981-13990.

Šali A, Shakhnovich E, Karplus M. 1994a. Kinetics of protein folding—A lattice model study of the requirements for folding to the native state. J Mol *Biol 235*:1614-1636.

Šali A, Shakhnovich E, Karplus M. 1994b. How does a protein fold? Nature *369*:248-251.

Sanchez IC. 1979. Phase transition behavior of the isolated polymer chain. Macromolecules *12*:980-988.

Sasaki T, Lieberman M. 1993. Between secondary structure and tertiary structure falls the globule: A problem in de novo protein design. Tetrahedron *49*:3677-3689.

Schellman JA. 1958. The factors affecting the stability of hydrogen-bonded polypeptide structures in solution. J Phys Chem *62*:1485-1494.

Scholtz JM, Baldwin RL. 1992. The mechanism of a-helix formation by peptides. *Annu* Rev Biophys Biomol Struct *21*:95-118.

Scholtz JM, Qian H, York EJ, Stewart JM, Baldwin RL. 1991. Parameters of helix-coil transition theory for alanine-based peptides of varying chain lengths in water. Biopolymers 31:1463-1470.

Segawa SI, Sugihara M. 1984. Characterization of the transition state of lysozyme unfolding. I. Effect of protein-solvent interactions on the transition state. Biopolymers *23*:2473-2488.

Semisotnov GV, Rodionova NA. Kutyshenko VP, Ebert B, Blanck J, Ptitsyn OB. 1987. Sequential mechanism of refolding of carbonic anhydrase B. *FEBS* Lett *224*:9-13.

Serrano L, Matouschek A, Fersht AR. 1992. The folding of an enzyme. III. Structure of the transition state for unfolding of barnase analysed by a protein engineering procedure. J Mol *Biol 224*:805-818.

Seshadri S, Oberg KA, Fink AL. 1994. Thermally denatured ribonuclease A retains secondary structure as shown by FTIR. *Biochemistry 33*:1351-1355.

Shakhnovich E, Farztdinov G, Gutin AM, Karplus M. 1991. Protein folding bottlenecks: A lattice Monte Carlo simulation. Phys Rev Lett 67: 1665-1668.

Shakhnovich EI. 1994. Proteins with selected sequences fold into unique native conformation. Phys Rev Lett *72*:3907-3910.

Shakhnovich EI. Finkelstein AV. 1989. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. Biopolymers 28:1667-1680.

Shakhnovich EI, Gutin AM. 1989a. Formation of unique structure in polypeptide chains—Theoretical investigation with the aid of a replica approach. Biophys Chem 34:187-199.

Shakhnovich EI, Gutin AM. 1989b. The nonergodic "spin-glass-like"phase of heteropolymer with quenched disordered sequence of links. Europhys Lett *8*:327-332.

Shakhnovich EI, Gutin AM. 1990a. Enumeration of all compact conformations of copolymers with random sequence of links. J Chem *Phys 93*:5967-5971.

Shakhnovich EI, Gutin AM. 1990b. Implications of thermodynamics of protein folding for evolution of primary sequences. Nature *346*:773-775.

Shakhnovich EI, Gutin AM. 1993a. Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad *Sci* USA *90*:7195-7199.

Shakhnovich El, Gutin AM. 1993b. A new approach to the design of stable proteins. Protein *Eng 6*:793-800.

Shirakawa M, Fairbrother WJ, Serikawa Y, Ohkubo T, Kyogoku Y, Wright PE. 1993. Assignment of 1H, 15N, and 13C resonances, identification of elements of secondary structure and determination of the global fold of the DNA-binding domain of GAL4. Biochemistry *32*:2144-2153.

Shiraki K, Nishikawa K, Goto Y. 1995. Trifluoroethanol-induced stabilization of the a-helical structure of β-lactoglobulin: Implication for nonhierarchical protein folding. J Mol *Biol 245*:180-194.

Shortle D. 1993. Denatured states of proteins and their roles in folding and stability. Curr Opin Struct *Biol 3*:66-74.

Shortle D, Chan HS, Dill KA. 1992. Modeling the effects of mutations on the denatured states of proteins. Protein *Sci 1*:201-215.

Shortle D, Meeker AK. 1986. Mutant forms of staphylococcal nuclease with altered patterns of guanidine hydrochloride and urea denaturation. Proteins Struct Funct Genet *1*:81-89.

Shortle D, Meeker AK. 1989. Residual structure in large fragments of staphylococcal nuclease: Effects of amino acid substitutions. Biochemistry *28*:936-944.

Shortle D, Stites WE, Meeker AK. 1990. Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry 29*:8033-8041.

Simon RJ, Kania RS, Zuckermann RN, Huebner VD, Jewell DA, Banville S, Ng S, Wang L, Rosenberg S, Marlowe CK, Spellmeyer DC, Tan R, Frankel AD, Santi DV, Cohen FE, Bartlett PA. 1992. Peptoids: A modular approach to drug discovery. Proc Natl Acad Sci USA *89*:9367-9371.

Singh J, Thornton JM. 1990. T SIRIUS. An automated method for the analysis of the preferred packing arrangements between protein groups. J Mol *Biol 211*:595-615.

Singh J, Thornton JM. 1992. Atlas of protein side-chain interactions. New York: Oxford University Press.

Skolnick J, Kolinski A. 1989. Computer simulations of globular protein folding and tertiary structure. *Annu* Rev Phys Chem *40*:207-235.

Skolnick J, Kolinski A. 1991. Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J Mol Biol 221*:499-531.

Socci ND, Bialek W, Onuchic JN. 1994. Properties and origins of protein secondary structure. Phys Rev E *49*:3440-3443.

Socci ND, Onuchic JN. 1994. Folding kinetics of protein-like heteropolymers. J Chem Phys *100*:1519-1528.

Sosnick TR, Mayne L, Hiller R, Englander SW. 1994. The barriers in protein folding. Nature Struct Biol 1:149-156.

Sosnick TR, Trewhella J. 1992. Denatured states of ribonuclease A have compact dimensions and residual secondary structure. Biochemistry *31*:8329-8335.

Stanley HE. 1971, 1987. Introduction to phase transitions and critical phenomena. New York: Oxford University Press.

Stigter D, Alonso DOV, Dill KA. 1991. Protein stability: Electrostatics and compact denatured states. Proc Natl Acad *Sci* USA *88*:4176-4180.

Stillinger FH, Head-Gordon TH, Hirshfeld CL. 1993. Toy model for protein folding. Phys Rev E 48:1469-1477.

Stolorz P. 1994. Recursive approaches to the statistical physics of lattice proteins. In: Hunter L, ed. Proceedings *of* the 27th Hawaii international conference on system sciences, volume V. Los Alamitos, California: IEEE Computer Society Press. pp 316-325.

Straub JE, Thirumalai D. 1993. Exploring the energy landscape in proteins. Proc Natl Acad *Sci* USA *90*:809-813.

Sueki M, Lee S, Powers SP, Denton JB, Konishi Y, Scheraga HA. 1984. Helix-coil stability constants for the naturally occurring amino acids in water. 22. Histidine parameters from random poly [(hydroxybutyl) glutamine-co-L-histidine]. Macromolecules 17:148-155.

Sun DP, Soderlind E, Baase WA, Wozniak JA, Sauer U, Matthews BW. 1991. Cumulative site-directed charge-change replacements in bacteriophage

T4 lysozyme suggest that long-range electrostatic interactions contribute little to protein stability. J Mol *Biol 221*:873–887.

Sun ST, Nishio I, Swislow G, Tanaka T. 1980. The coil–globule transition: Radius of gyration of polystyrene in cyclohexane. J Chem Phys 73: 5971–5975.

Swindells MB, Thornton JM. 1993. A study of structural determinants in the interleukin-1 fold. Protein *Eng 6*:711–715.

Taketomi H, Ueda Y, Gō N. 1975. Studies on protein folding, unfolding and fluctuations by computer simulation. *Int* J Pept Protein Res *7*:445–459.

Tamburro AM, Scatturin A, Rocchi R, Marchiori F, Borin G, Scoffone E. 1968. Conformational transitions of bovine pancreatic ribonuclease S-peptide. *FEBS* Lett *1*:298–300.

Tamura A, Kimura K, Akasaka K. 1991a. Cold denaturation and heat denaturation of Streptomyces subtilisin inhibitor. 2. 1H NMR studies. Biochemistry *30*:11313–11320.

Tamura A, Kimura K, Takahara H, Akasaka K. 1991b. Cold denaturation and heat denaturation of Streptomyces subtilisin inhibitor. 1. CD and DSC studies. Biochemistry 30:11307–11313.

Tanaka T, Kuroda Y, Kimura H, Kidokoro SI, Nakamura H. 1994. Cooperative deformation of a de novo design protein. Protein Eng 7:969–976.

Tanford C, De PK, Taggart VG. 1960. The role of the a-helix in the structure of proteins: Optical rotatory dispersion of 0-lactoglobulin. J Am Chem *Soc 82*:6028–6034.

Thirumalai D. 1994. Theoretical perspectives on in vitro and in vivo protein folding. In: Doniach S, ed. Statistical mechanics, protein structure, and protein–substrate interactions, New York: Plenum. pp 115–134.

Thomas PD, Dill KA. 1993. Local and nonlocal interactions in globular proteins and mechanisms of alcohol denaturation. Protein *Sci 2*:2050–2065.

Tiktopulo EI, Bychkova VE, Rička J, Ptitsyn OB. 1994. Cooperativity of the coil–globule transition in a homopolymer—Microcalorimetric study of poly(*N*-isopropylacrylamide). Macromolecules *27*:2879–2882.

Unger R, Moult J. 1993. Genetic algorithms for protein folding simulations. J Mol *Biol 231*:75–81.

Uversky VN. 1993. Use of fast protein size-exclusion liquid chromatography to study the unfolding of proteins which denature through the molten globule. Biochemistry 32:13288-13298.

Uversky VN, Ptitsyn OB. 1994. "Partly folded" state, a new equilibrium state of protein molecules: Four-state guanidinium chloride-induced unfolding of β-lactamase at low temperature. Biochemistry *33*:2782–2791.

Varley P, Gronenborn AM, Christensen H, Wingfield PT, Pain RH, Clore GM. 1993. Kinetics of folding of the all-β sheet protein interleukin-1 β. Science *260*:1110–1113.

Vila J, Williams RL, Grant JA, Wojcik J, Scheraga HA. 1992. The intrinsic helix-forming tendency of L-alanine. Proc Natl Acad Sci USA *89*:7821–7825.

von Hippel PH, Wong KY. 1965. On the conformational stability of globular proteins: The effects of various electrolytes and non-electrolytes on the thermal ribonuclease transition. J Biol Chem *240*:3909–3923.

Vuilleumier S, Mutter M. 1993. Synthetic peptide and template-assembled synthetic protein models of the hen egg white lysozyme 87–97 helix: Importance of a protein-like framework for conformational stability in a short peptide sequence. *Biopolymers 33*:389–400.

Wang Z, Jones JD, Rizo J, Gierasch LM. 1993. Membrane-bound conformation of a signal peptide: A transferred nuclear Overhauser effect analysis. Biochemistry 32:13991–13999.

Waterhous DV, Johnson WC Jr. 1994. Importance of environment in determining secondary structure in proteins. Biochemistry *33*:2121–2128.

Weissman JS, Kim PS. 1991. Reexamination of the folding of BPTI: Predominance of native intermediates. Science 253:1386–1393.

Weissman JS, Kim PS. 1992a. Kinetic role of nonnative species in the folding of bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci USA *89*:9900–9904.

Weissman JS, Kim PS. 1992b. The disulfide folding pathway of BPTI. Science *256*:112–114.

Wilbur WJ, Liu JS. 1994. Energy distribution of the compact states of a peptide chain. Macromolecules *27*:2432–2438.

Wilkinson KD, Mayer AN. 1986. Alcohol-induced conformational changes of ubiquitin. Arch Biochem Biophys *250*:390–399.

Williams C, Brochard F, Frisch HL. 1981. Polymer collapse. *Annu* Rev Phys Chem *32*:433–451.

Wright PE, Dyson HJ, Lerner RA. 1988. Conformation of peptide fragments of proteins in aqueous solution: Implications for initiation of protein folding. Biochemistry *27*:7167–7175.

Yeates TO, Komiya H, Rees DC, Allen JP, Feher G. 1987. Structure of the reaction center from Rhodobacter sphaeroides R-26: Membrane–protein interactions. Proc *Natl* Acad Sci USA *84*:6438–6442.

Yee DP, Chan HS, Havel TF, Dill KA. 1994. Does compactness induce secondary structure in proteins? A study of poly-alanine chains computed by distance geometry. J Mol *Biol 241*:557–573.

Yoder MD, Keen NT, Jurnak F. 1993. New domain motif: The structure of pectate lyase C, a secreted plant virulence factor. *Science 260*:1503–1507.

Yue K, Dill KA. 1992. Inverse protein folding problem: Designing polymer sequences. Proc Natl Acad Sci USA *89*:4163–4167.

Yue K, Dill KA. 1993. Sequence structure relationship of proteins and copolymers. Phys Rev E *48*:2267–2278.

Yue K, Dill KA. 1995. Forces of tertiary structural organization of globular proteins. Proc Natl Acad *Sci* USA 92:146–150.

Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI, Dill KA. 1995. A test of lattice protein folding algorithms. Proc Natl Acad *Sci* USA *92*:325–329.

Zhang XJ, Baase WA, Matthews BW. 1991. Toward a simplification of the protein folding problem: A stabilizing polyalanine a-helix engineered in T4 lysozyme. Biochemistry *30*:2012–2017.

Zhong L, Johnson WC Jr. 1992. Environment affects amino acid preference for secondary structure. Proc Natl Acad *Sci* USA *89*:4462–4465.

Zimm BH, Bragg JK. 1959. Theory of the phase transition between helix and random coil in polypeptide chains. J Chem Phys *31*:526–535.

Zwanzig R, Szabo A, Bagchi B. 1992. Levinthal's paradox. Proc Natl Acad Sci USA *89*:20–22.