Quarterly Reviews of Biophysics **35**, 2 (2002), pp. 111–167. © 2002 Cambridge University Press
DOI: 10.1017/S0033583502003761 Printed in the United Kingdom

Understanding protein folding with energy landscape theory
Part I: Basic concepts

Steven S. Plotkin* and José N. Onuchic†

Department of Physics, University of California at San Diego, La Jolla 92093, USA

1. Introduction

The current explosion of research in molecular biology was made possible by the profound discovery that hereditary information is stored and passed on in the simple, one-dimensional

* Address for correspondence: Steven Plotkin, Department of Physics and Astronomy, University of British Columbia at Vancouver, 6224 Agricultural Road, Vancouver, B.C. V6T 1Z1.
Tel.: 604-822-8813; Fax: 604-822-5324; E-mail: steve@physics.ubc.ca

† Address for correspondence: José N. Onuchic, University of California at San Diego, La Jolla, CA 92093, USA.
Tel.: 858-534-7067; Fax: 858-822-4560; E-mail: jonuchic@ucsd.edu

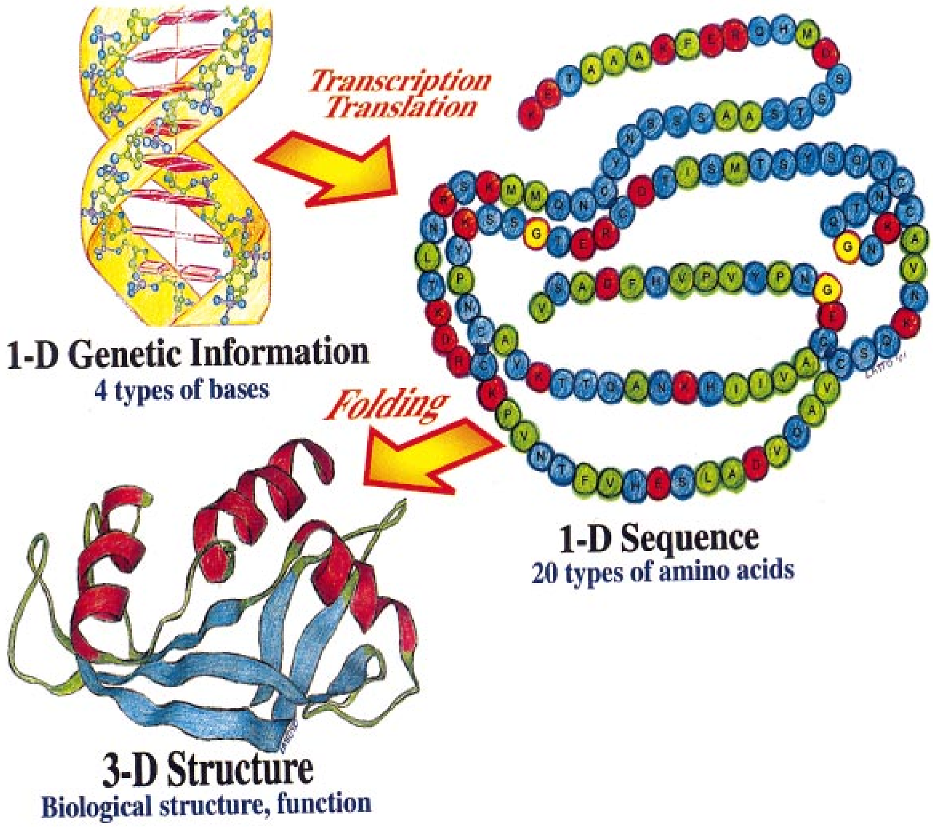

Fig. 1. All information necessary for the structure and function of a living organism is contained in the 1D sequence of base pairs in the DNA molecule. This information is eventually translated to the specific sequences of amino acids in protein chains. This 1D information encodes, through the complicated process of folding, for the 3D information contained in the native structure in which the protein is functional. Folding is non-local in that it involves bringing parts of the chain remote in sequence close together in space.

(1D) sequence of DNA base pairs (Watson & Crick, 1953). The connection between heredity and biological function is made through the transmission of this 1D information, through RNA, to the protein sequence of amino acids. The information contained in this sequence is now known to be sufficient to completely determine a protein's geometrical 3D structure, at least for simpler proteins which are observed to reliably refold when denatured *in vitro*, i.e. without the aid of any cellular machinery such as chaperones or steric (geometrical) constraints due to the presence of a ribosomal surface (for example Anfinsen, 1973) (see Fig. 1). Folding to a specific structure is typically a prerequisite for a protein to function, and structural and functional probes are both often used in the laboratory to test for the *in vitro* yield of folded proteins in an experiment.

Further understanding of the molecular description of life thus requires answering the deceptively simple question of how the 1D sequence of amino acids in a protein chain determines its 3D folded conformation in space.[1] Although this problem is now considered central to molecular biology, only after about 40 years of study is it beginning to yield to the combined efforts of molecular biologists, chemists, physicists, and mathematicians. The difficulty of the problem lies in the unfamiliar nature of 1D to 3D information transcription in going from protein sequence to protein structure: the information processing cannot go in a sequential symbol by symbol fashion, but must operate simultaneously using remote parts of the sequence, and hence is essentially a non-local, collective process rather than the trivial translation of a message.

Initial progress was primarily descriptive. Proteins were observed to be covalently bonded, linear polymers with a specific primary sequence of side-chains or amino acids attached at regular intervals, constituting the 1D information pattern. While a given sequence uniquely determines a folded structure, a given structure may be highly degenerate in the sequences that fold to it, or can be 'designed' for it. Finding these sequences is known as the inverse protein folding or design problem. It is not uncommon for sequences with less than 50% identity of amino acids to fold to the same or similar structure, with nearly the same folding rate at a given stability.

The first X-ray crystallographic structure observed, that of myoglobin, revealed a surprising amount of irregularity and apparent lack of symmetry as compared to the simple double-stranded DNA structure observed 5 years earlier (Kendrew *et al.* 1958) (see PDB entry 1mbn). The structures appeared as densely packed random walks of the backbone chain, soon understood to be driven at least in part by the hydrophobic interaction of some of the buried side chains with the surrounding water (Kauzmann, 1959). However at smaller length scales (several angstroms) structural regularities are seen in the form of helical segments of polymer interspersed throughout the globule (α-helices) (Perutz, 1951). This 'secondary' structure, predicted earlier to be stabilized by hydrogen bonds (Pauling *et al.* 1951), allows hydrophobic parts of the protein to be buried while hydrophilic parts, including the backbone chain itself, are exposed to solvent, crudely analogous to micellar formation.

A later analysis showed that the space of sterically allowed, local rotational angles of the backbone chain was quite restricted (Ramachandran & Sassiekharan, 1968). One of the allowed regions corresponded to the α-helix, another to a pleated 2D structure of parallel or anti-parallel strands (β-sheets). These structures can persist indefinitely in effectively infinite protein structures such as wool (the α-helix) or silk (the β-sheet), but in globular proteins ($\sim$ 10–50 Å in linear dimension) they are broken up by turns of dense, semi-rigid random-coil. The secondary structural elements of a protein are determined through the collective interactions of the elements with the rest of the molecule: the identity of an amino acid does not by itself exclusively determine what secondary structure it will be found in. For example, negatively charged Glutamic acid (Glu) occurs in 5 places in the chain A subunit of bovine pancreatic ribonuclease (PDB code 1C0B) depicted in Fig. 1: in residue position 2 (near one end) it is in the coil configuration, in position 9 it is in an α-helix, in position 49, coil, in

[1] More precisely the set of near native conformations, for example under biological conditions, conformational fluctuations around the folded structure in myoglobin render the protein transparent to O_2, even though there is no access channel to the heme evident in the crystal structure (Kendrew *et al.* 1958). In depth studies of the structure and dynamics within the native ensemble of conformational states have been made (Austin *et al.* 1975) which we will not consider in detail here, since we concentrate on the folding transition to this ensemble.

position 111, β-sheet. The corresponding residues in human ribonuclease, determined by structural alignment (Holm & Sander, 1996), are: Glu2, a polar residue which is uncharged but has a similar amino-acid structure; however it maintains an α-helical conformation; Leu10, a hydrophobic residue, Glu49, Asp84, another negatively charged amino acid; and Glu109. The corresponding amino acids are at different residue indices because the functioning protein is amenable to omissions or insertions in sequence. The two proteins are overall highly similar in structure as determined by their C_α–C_α distances (Holm & Sander, 1996). Thus the corresponding residues in the functioning protein from another species, or even of the same species at different stages of its evolution, may or may not be the same, or even the same type. Moreover they tend to, but need not have the same secondary structure since a protein can still function for a class of native structures with some non-zero degree of structural variance. The upshot from this example is that the formation of the folded structure is governed by the collective effects of non-covalent interactions in essentially the whole molecule: A theory that is local cannot solve the protein folding problem (Wolynes, 1992; Ngo *et al.* 1994). This is fundamentally different from the self-organization occurring in biological systems without fixed disorder, for example the hierarchical helix-formation involved in DNA folding, where the disorder in the sequence is suppressed by complementary base pairs, and the folding (supercoiling) mechanism is local. Collective interactions slow folding while enhancing stability by involving non-local parts of the chain in the folding nucleus.

The interactions stabilizing the native structure tend to also be cooperative: the energetic gains in forming native structure are achieved only when several parts of the protein are in spatial proximity and interacting (Perutz, 1970; Finkelstein & Shakhnovich, 1989; Murphy & Freire, 1993; Plotkin *et al.* 1997; Sorenson & Head-Gordon, 1998; Lum *et al.* 1999). There is a larger entropic barrier for this kind of process than for one involving simple pair interactions. We investigate cooperative effects later in Part II of this review series.

In forming a protein which is compact overall, the elements of secondary structure must themselves be packed together into what is called tertiary structure. β-Sheets can only be oriented in certain ways when stacked on top of each other, as in the packing of 2D layers into crystals (Chothia *et al.* 1977). Similarly at lower temperatures it becomes favorable for α-helices to align nematically as in liquid crystal ordering of rigid rods (Onsager, 1949; Flory, 1956; de Gennes, 1975), an effect also studied in nematic polymers and known as induced rigidity (de Gennes & Pincus, 1977).

High-resolution determinants such as NMR and crystallization have now revealed a prevalence of spatial group symmetries in native protein structures (Richardson, 1981; Chirgadze, 1987; Murzin & Finkelstein, 1988). For example cytochrome B_{662} has a 4-helix bundle motif with the approximate symmetry group of a square with undirected sides (the dihedral group D_4), plus a reflection through the plane orthogonal to the long axis. The location of the active site is unrelated to the structural symmetry of the protein, implying that the symmetry exists to assist folding rather than function. The current understanding of this phenomenon is that a symmetric structure has lower overall energy, just as the close packing of simpler molecules into a crystal having lower group symmetry than the liquid maximizes the number of interactions (Wolynes, 1996). The heteropolymeric nature of proteins introduces effective defects into the structure, making the symmetries approximate. Evolutionary pressure for folding to a stable structure may impart *en-passant* crystal-like group symmetry properties to the ground state of the system.

Larger proteins undergo still higher levels of organization in the form of domains, which tend to be contiguous in primary sequence and compact three-dimensionally (Wetlaufer, 1973). Finally, functioning biomolecules are often multi-protein complexes held together by van der Waals interactions or salt bridges, as in the case of the four myoglobin units comprising hemoglobin. This level of organization is usually referred to as quaternary structure, and allows for even larger scale collective motion. For example the allosteric function of hemoglobin results from cooperative effects on oxygen-binding affinity when the subunits are arranged in their quaternary structure (Monod *et al.* 1965; Perutz, 1970). This effect allows oxygen to be released from the molecule at venous oxygen pressure.

Some good reviews of protein folding, structure, function, and related issues are given in the References (see Gō, 1983; Dill, 1990; Frauenfelder *et al.* 1991; Wolynes, 1992; Karplus & Shakhnovich, 1992; Creighton, 1992; Bryngelson *et al.* 1995; Dill *et al.* 1995; Baldwin, 1995, 1999; Ball *et al.* 1996; Onuchic *et al.* 1997; Veitshans *et al.* 1997; Dill & Chan, 1997; Pande *et al.* 1997, 2000; Brooks *et al.* 1998; Dobson *et al.* 1998; Garel *et al.* 1998; Gruebele, 1999; Fersht, 1999; Wales & Scheraga, 1999; Honig, 1999; Onuchic *et al.* 2000). In the next two sections, we give an overview of the deeper conceptual issues involved in folding. We will often bring in equations derived later in the text to elucidate a point, always citing the future place they appear within the article. Terms are typically defined when introduced, however the reader may find the glossary in the Appendix helpful. We will also reference a review to appear in a subsequent issue of this journal, where many of the subjects mentioned here are treated more fully, as Part II.

2. Levinthal's paradox and energy landscapes

The observation mentioned earlier that proteins can fold reversibly *in vitro* without any external cellular machinery means that the folding mechanism can be theoretically and experimentally studied for a single isolated protein molecule interacting with solvent. In this review we study the theoretical aspects of the folding process based on this assumption. We will not treat in any detail those aspects of the folding problem related to predicting the actual folded structure from a given specific sequence – instead we will concentrate on elucidating general features present in the folding mechanism which are common to all proteins, as well as general principles which allow one to predict specific properties common to a set of proteins having a given structural or energetic feature.

The reversible *in vitro* folding of a single protein means that the protein in the native state is thermodynamically stable, and therefore that the native state has the global minimum free energy of all kinetically accessible structures (Epstein *et al.* 1963; Levinthal, 1969). Furthermore since the folded state is a small ensemble of conformational structures compared to the conformational entropy present in the unfolded ensemble, at a coarse-grained level of description the folded structure must then have the lowest internal energy of all kinetically accessible conformational structures.[2] We can define the energy landscape for this system as a mapping of the chain conformation to its internal energy, along with rules defining what

[2] Internal (free) energy is defined here as the free energy of a single backbone conformation, i.e. backbone conformational entropy has been subtracted out of the system.

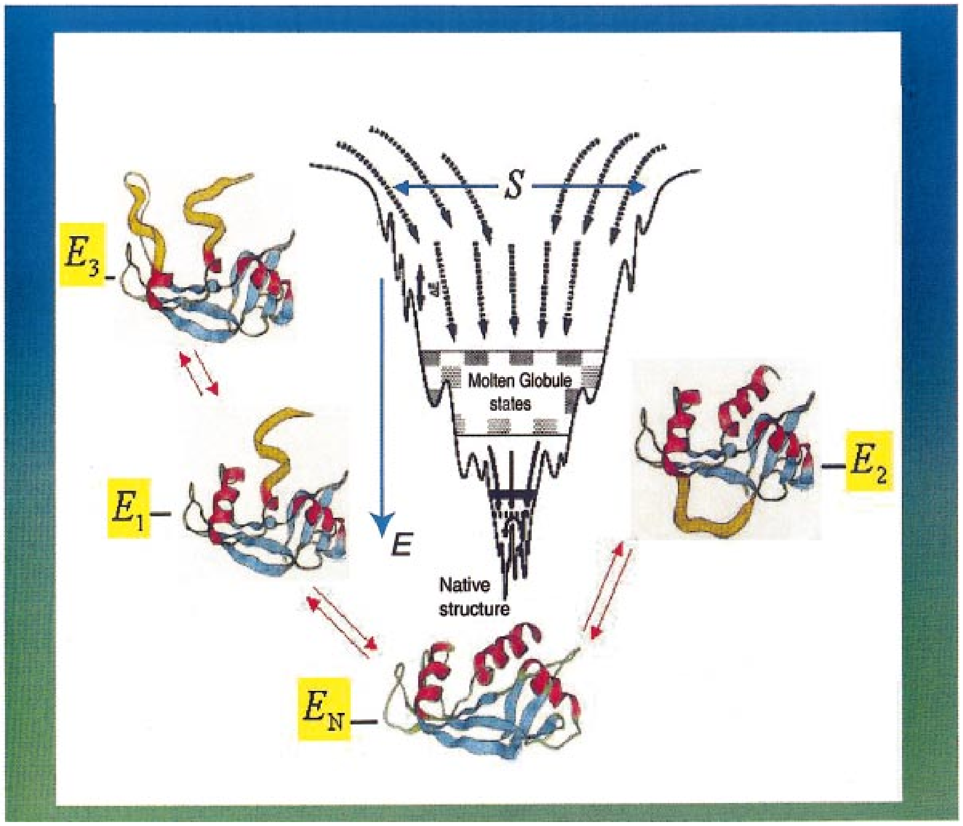

Fig. 2. Folding amounts to configurational diffusion on an energy landscape which has an overall funnel topography. The landscape is inherently many-dimensional corresponding to the intricate connectivities between configurational states, so the funnel schematic is clearly a projection. The depth represents the free energy of a conformational state, the width is typically taken to be a measure of the configurational entropy. Different states have different energies. Each state has a transition matrix describing transition rates to configurations locally connected to it. The set of states along with their energies and transition matrices fully determine the folding dynamics for the system.

configurations are accessible from a given configuration and how the protein can move between them (see Fig. 2).

Proteins have been evolutionarily designed to perform a specific biological function [or functions, for example cylin-dependent kinases have at least four functional sites (Dreyer *et al*. 2001)]. Essential to this function for many proteins appears to be the existence of a well-defined conformational structure under biological conditions. Thus part of the evolution process must involve design to fold to a well-defined structure. The co-evolution of function and native stability is non-trivial, since, e.g. preserving functions involving large-scale conformational changes may frustrate stability of the folded structure (see, for example Brown & Sauer, 1999; Garcia *et al*. 2000). The primary force opposing the transition to a well-defined structure is the necessary loss of conformational entropy of the polypeptide chain, since solvent entropy increases upon burial of non-polar side-chains, as well as for the organization of the polar groups (see Fig. 3, inset).

As can be seen from the experimental data in the inset of Fig. 3 (Makhatadze & Privalov, 1996), the conformational entropy gives the largest contribution to the entropy change of the

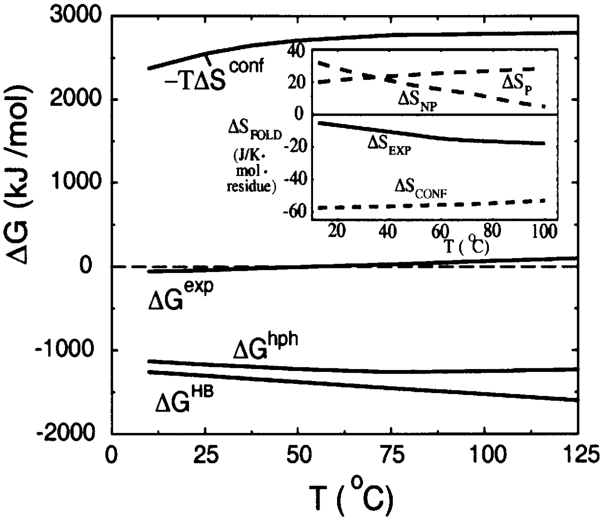

Fig. 3. (Inset) The entropy change in folding, shown here for Barnase (Makhatadze & Privalov, 1996), has four main contributions, organization of non-polar and polar residues which are both positive on folding, and conformational entropy of the backbone and side-chains, which are negative on folding. Much of the side-chain entropy is believed to be retained upon folding, while much of the backbone entropy is believed to be lost. The total entropy change on folding may be positive or negative; the total free energy change on folding is negative if the folded state is the equilibrium state, as seen below about 55 °C. The main figure shows the contributions to the Gibbs free-energy. Hydrogen bonding and hydrophobic forces drive folding, while conformational entropy loss opposes it. All contributions are large in magnitude compared to the net free-energy change. (Adapted from Makhatadze & Privalov, 1996.)

protein upon folding. The entropy change of hydration of polar and non-polar groups are both positive on folding and so drive the folding process. Consideration of both of these entropic factors significantly increases the value for the conformational entropy loss from previously believed values (Makhatadze & Privalov, 1996), although the numbers used here may be an over-estimate. This value includes the backbone conformational entropy loss, as well as the entropy loss to pack side-chains. The latter number is estimated to be about 14 J/(K mol residue) (McCammon *et al.* 1977; Karplus & McCammon, 1983; Doig & Sternberg, 1995). Much residual side-chain entropy is believed to be present in the folded state (McCammon *et al.* 1977; Karplus & McCammon, 1983). According to the experimental data in Fig. 3, the backbone entropy loss is about a factor of 3 greater than that of side-chain packing, so that protein folding is largely a backbone conformational transition. This is the assumption prevalent in many theoretical models of folding, and most computational models of folding. The favorable entropy to order hydrophobic and polar residues is typically incorporated into the effective free energies of interactions between residues in the native conformation.

The extent of the entropy lost in folding is enough to eliminate the possibility of an exclusively random search for the native structure: From Fig. 3 (inset) the conformational entropy loss per residue $\Delta s^{(c)}$ for Barnase folding at 25 °C is approximately

$$\Delta s^{(c)} = s_{\mathrm{u}}^{(c)} - s_{\mathrm{f}}^{(c)} \approx k_{\mathrm{B}} \ln\left(\frac{\Omega_{\mathrm{u}}}{\Omega_{\mathrm{f}}}\right) \approx 57\frac{\mathrm{J}}{\mathrm{K\ mol\ residue}}. \tag{2.1}$$

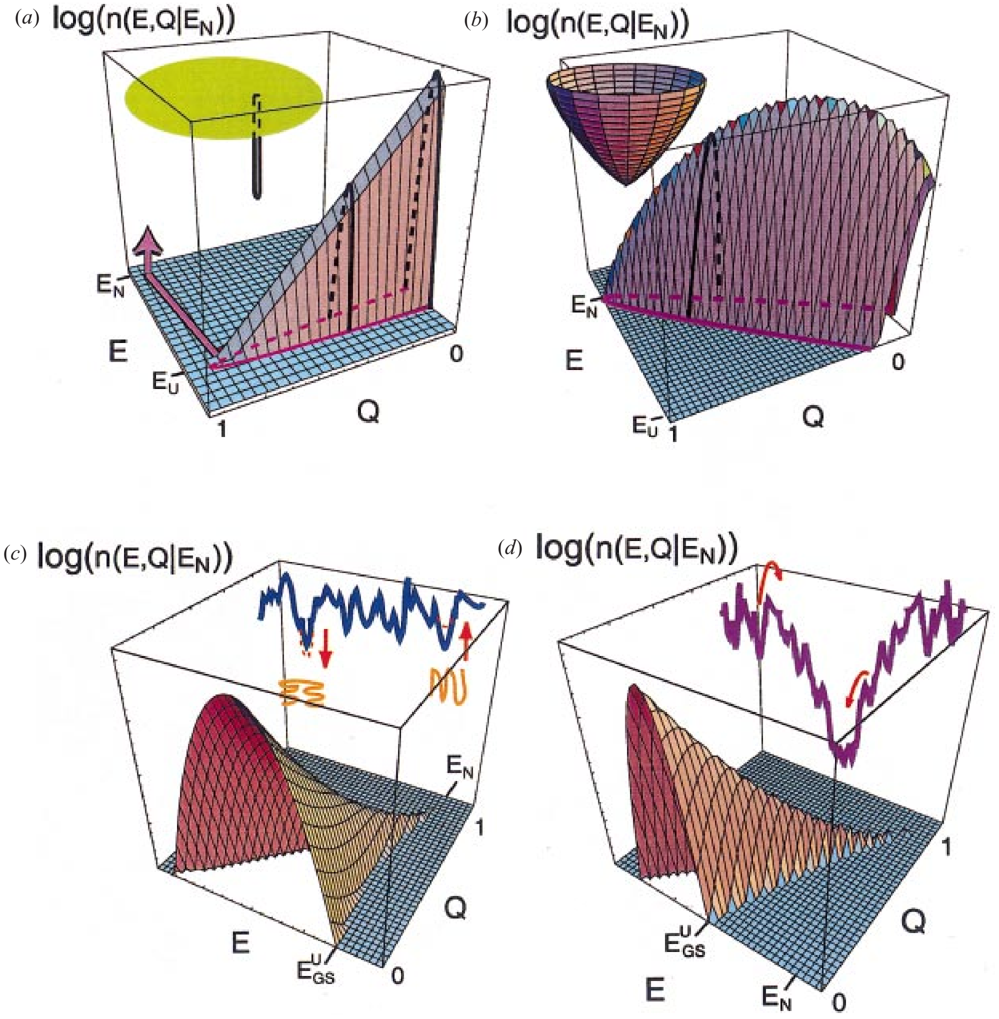

Fig. 4. Schematic density of states $n(E, Q|E_N)$ *versus* the energy E and order parameter Q measuring native similarity, given a state at $Q = 1$ has native energy, E_N. Surfaces are shown for several different model landscapes, to illustrate the importance of both energy and entropy function. (*a*) Golf-course or Levinthal landscape. The energy is well defined at Q, but does not drop down until $Q \approx 1$. The entropy decreases as Q increases, which results in a pathologically large entropy barrier to fold. The corresponding golf-course or Levinthal landscape is shown in the inset. (*b*) Independent residue model. The energy correlates well with Q and has a narrow width due to the absence of non-native heteropolymer interactions. The landscape is energetically funneled, but the independent residue entropy function leads to pathologies in stability and folding barrier (see Section 2). The inset shows the corresponding 'wine-glass' funnel. (*c*) Random heteropolymer landscape. Random interactions lead to a broad distribution of state energies. The ground state energy E_{GS}^{U} at $Q \approx 0$ is almost as low as the true ground state E_N at $Q = 1$, and small perturbations in protein sequence or environment may cause the ground state conformation at $Q \approx 0$ to become the new true ground state, as illustrated schematically with the corresponding energy landscape. The configurational entropy is shown here to decrease sublinearly due to polymer topological constraints (see Part II). (*d*) Funneled landscape of a minimally frustrated heteropolymer or protein. Random interactions are present which broaden the density of states, however minimal frustration leads to a residual correlation of mean energy with Q (see e.g. Gutin *et al*. 1995). The ground-state energy (E_{GS}^{U}) of the states at $Q \approx 0$ is significantly higher than the true ground state energy E_N at $Q = 1$. The incomplete cancellation of entropic losses and (negative) energetic gains as Q increases leads to a residual free-energy barrier to fold. The thermodynamic properties of each of these landscapes are described in Fig. 5.

Here $\mathbf{\Omega}_{\mathrm{u,f}}$ is the number of states per residue in the unfolded and folded states respectively, and k_{B} is Boltzmann's constant. While the number of Eq. (2.1) may be an over-estimate, since it is comparable to the conformational entropies of many liquids, it is worthwhile to use the number in the following argument. Regardless of the number of states $\mathbf{\Omega}_{\mathrm{f}}$ in the folded ensemble, the number of states per residue to be searched through before finding a conformational state consistent with native structure is

$$\exp \Delta s^{(\mathrm{c})}/k_{\mathrm{B}} \approx 900 \, \frac{\text{states}}{\text{residue}} \tag{2.2}$$

or an entropy of about 6·8 k_{B} per residue.[3] Thus there is a huge conformational state space available to the system. Since Barnase has ~ 110 residues, $\mathbf{\Omega}_{\mathrm{tot}} \sim 900^{110} \sim 10^{325}$ states, of which a particular one or a very small subset must be found in folding. The estimate in Eq. (2.2) assumes the same Boltzmann weight for the probability of each state in the unfolded ensemble, however removing this assumption only makes the configurational entropy larger as determined from the measurement. Moreover, estimates of the water-accessible surface area of unfolded Barnase indicate a state similar to a random-coil, so enthalpic differences between conformations should not be too large,[4] and the approximation of equal occupation probability for many of the states is not obviously poor.

Letting $\nu = \mathbf{\Omega}_{\mathrm{u}} + 1$ be the total number of states of each residue and assuming one state per residue in the folded structure, the probability in a trial from a random search that any residue finds its native state is $1/\nu$. Then the probability $P_{\mathrm{F}}^{\mathrm{rand}}$ that all N residues independently find their native states is ν^{-N}. The mean number of trials $\bar{N}$ needed to find the full native conformation is then $1/P_{\mathrm{F}}^{\mathrm{rand}} = \nu^{N}$. The mean first passage time to find this state is the sampling time τ_0 times the mean number of required trials $\bar{N}$. The faster isomerization rates k_0 of a protein chain are of order $(1 \text{ ps})^{-1}$, so within a picosecond all residues made an attempt to find their native state.[5] Then the mean first passage time is

$$\tau_{\mathrm{F}}^{\mathrm{random}} \approx \tau_0 \nu^{N} = (10^{-12} \text{ s}) \, 900^{110} \sim 10^{305} \text{ years} \tag{2.3}$$

or about 10^{295} times the age of the universe! The argument showing the impossibility of protein folding by random search was first made in passing by Levinthal (1969) in an article questioning the necessity of Anfinsen's thermodynamic hypothesis mentioned above (Anfinsen, 1973).

In this formulation of the problem, finding the native structure is analogous to a drunken golfer finding the hole on a vast green (see inset of Fig. 4*a*), and it is useful to think of folding

[3] These entropies are somewhat higher than previous estimates for conformational entropy (of $\sim 2{\cdot}3$ k_{B}/residue or ~ 10 states/residue) basically because previous estimates did not take into account the loss in entropy on unfolding due to the hydration of polar groups (Makhatadze & Privalov, 1996), (see Fig. 3, inset). These entropy losses, while possibly an over-estimate, are still significantly less than say an ideal gas, which has $\sim 10^{7}$ states/particle at STP.

[4] This is not necessarily true in the collapsed molten globule state, where strongly repulsive residues may be in contact in some configurations and strongly attractive residues may contact in others.

[5] In Monte-Carlo simulations of folding where local moves in random positions are attempted, the 'time' associated with a Monte Carlo step scale like $\sim 1/N$, where N is the size of the system, since these moves are occurring concurrently throughout the real system. Molecular dynamics simulations, which integrate Newton's equations with noise, do not have this problem.

in this scenario as occurring on an energy landscape with a golf-green topography: the system obtains no energetic gain from ordering any residues until it stumbles upon the complete native structure,[6] so sampling is unbiased. The log density of states looks schematically as in Fig. 4*a*, and the energy, entropy, and free energy as a function of the amount of folded structure are given in Fig. 4.

Real proteins have *in vitro* folding times of ~ μs – s, so only a tiny fraction of the available conformational space is actually sampled in a given folding event. Since proteins are thermodynamically stable in their native structure, it is instructive to redo the Levinthal argument above assuming each residue has an average stability δg in its native state (Zwanzig *et al.* 1992). Figure 3 shows the various free-energy contributions upon folding of Barnase (Griko *et al.* 1994). The transition is opposed by the loss in conformational entropy, and driven by gains in free energy due to hydrophobic burial and hydrogen bond formation. These two potentials are comparable in magnitude. The modest total free-energy change upon folding arises from the cancellation of these large terms. At a temperature of 300 K, the overall stability of one molecule of Barnase is only 50 kJ mol^{-1} or 20 $k_B T$: each amino acid is stable by only $\delta g \approx 0{\cdot}18\, k_B T$ on average ($N = 110$ for Barnase). If each residue is treated as independent, and imparted the average stability, the probability p_N each residue is native at say 300 K is only

$$p_N = \frac{1}{1+e^{-\delta g/T}} \approx 0{\cdot}55, \tag{2.4}$$

again under the critical assumption that residue fluctuations are independent. Following the argument that led to Eq. (2.3), the mean first passage time is[7]

$$\tau_F^{\text{indep}} \approx \tau_0\,(1+e^{-\delta g/T})^N \approx (10^{-12}\text{ s})\,(1{\cdot}8)^{110} \sim 10^9 \text{ years.} \tag{2.5}$$

To obtain a folding time of 1 ms, a stability of 430 kJ mol^{-1} is needed; about 8 times the experimentally determined value. Explicitly solving for the mean first passage time from the master equation gives essentially the same result (Zwanzig *et al.* 1992). At the temperature where each residue is equally likely to be native or unfolded, i.e. where $\delta g = 0$,

$$\tau_F^{\text{indep}} \approx \tau_0 2^N = \tau_0 e^{N \ln 2} \sim 10^{13} \text{ years.} \tag{2.6}$$

The marginal stability of the native structure also poses a structural problem for independently fluctuating residues. Protein function often depends on a well-defined native structure, however for independent residues, the probability n independent residues are native is given by the binomial distribution, and the average number $\bar{n}$ of native residues at 300 K is $\bar{n} = Np_N \approx 60$. It is dubious there is enough native structure here to preserve function, and such a low value is not consistent with many experimental structural probes.

Even though the topography of the energy landscape here is that of a funnel with the correct overall slope (energetic bias), correlations in the nativeness of residues, and the resulting collective conformational motions, must be accounted for to give reasonable kinetic predictions, as well as the required stability necessary for function. See Fig. 4*b* for illustrations of the independent residue landscape. Accounting for the coupling of nativeness between

[6] Problems with energy landscape of this nature (the golf course), when viewed as optimization problems to find the ground state, have been shown to be NP complete (Baum, 1986); it takes a computer an exponentially long time ($\tau \sim \exp N$) to find the ground state as well.

[7] Note that, when there is no energetic bias, $\delta g = -T \ln \Omega_u$ and the Levinthal result in Eq. (2.3) is recovered.

residues when say n residues overall are native amounts to correctly enumerating the number of states at n. Then one can calculate the transition kinetics for the accessible states. In the following sections we pursue this approach to calculate the correct reaction surface, then the kinetics on that reaction surface.

To illustrate what is happening physically in the independent fluctuation model above, each residue can be considered to have 1 native state with energy ϵ ($\epsilon < 0$) and Ω_{u} non-native states with zero energy. Then in the model the energy gain in folding is proportional to the number of native residues, so the energy is a linear function of n:

$$E(n) = n\epsilon. \tag{2.7}$$

The entropy is composed of two terms by construction of the model. There are $\binom{N}{n}$ ways to choose n native residues, and each of these configurations have $\Omega_{\mathrm{u}}^{N-n}$ states. So the entropy as a function of n is

$$S(q) \approx N\ [(1-q)\ \ln\ \Omega_{\mathrm{u}} - q\ \ln\ q - (1-q)\ \ln\ (1-q)], \tag{2.8}$$

where $q \equiv n/N$ is a number between 0 and 1 (the entropy is extensive). At the temperature T_{F}, where each residue is 50% native, $\epsilon = -T_{\mathrm{F}}\ \ln\ \Omega_{\mathrm{u}}$, and the free energy $F(q) = E(q) - T_{\mathrm{F}} S(q)$ becomes

$$\frac{F(q)}{T_{\mathrm{F}}} = N\ [q\ \ln\ q + (1-q)\ \ln\ (1-q)] \tag{2.9}$$

up to a constant ($-N\ \ln\ \Omega_{\mathrm{u}}$). The free energy $F(q)$ has a single basin with a minimum at $q = \frac{1}{2}$ and maxima at $q = 0$ and $q = 1$ (see Fig. 5*b*). The depth ΔF of the minimum is N ln 2. This explains the long folding time at T_{F}, as well as the weak stability of the native structure. The native structure is stable only when the overall free-energy profile is downhill, which does not occur until very low temperature:

$$T_{\mathrm{stable}} \approx \frac{|\epsilon|}{\ln\ N}. \tag{2.10}$$

In this model, solving the kinetics amounts to reproducing an Arrhenius equation for the folding time:

$$\tau_{\mathrm{F}} = \tau_0 \mathrm{e}^{\Delta F/T}, \tag{2.11}$$

which at T_{F} above is τ_0 exp (N ln 2), as in Eq. (2.6).

This example is instructive in that it shows the importance of accurately calculating the entropy of the landscape as well as the energy, and moreover that there is an intimate connection between folding kinetics and the thermodynamic free-energy landscape. Several of the following sections will be devoted to treating these aspects in detail. The above model also neglects the potentially important effects of a temperature-dependent prefactor $\tau_0(T)$, which may arise due to non-native trapping. Both barrier heights and prefactors are important in determining the rate.

2.1 Including randomness in the energy function

In the previous example, the low energy native state was baked into the model *a priori*, and the other states were assumed to all have the same energy. It is unlikely any real sequence of amino acids has this landscape, since real protein sequences have likely evolved either from a nearly random sequence or a nearly uniform sequence in selecting for stability in a

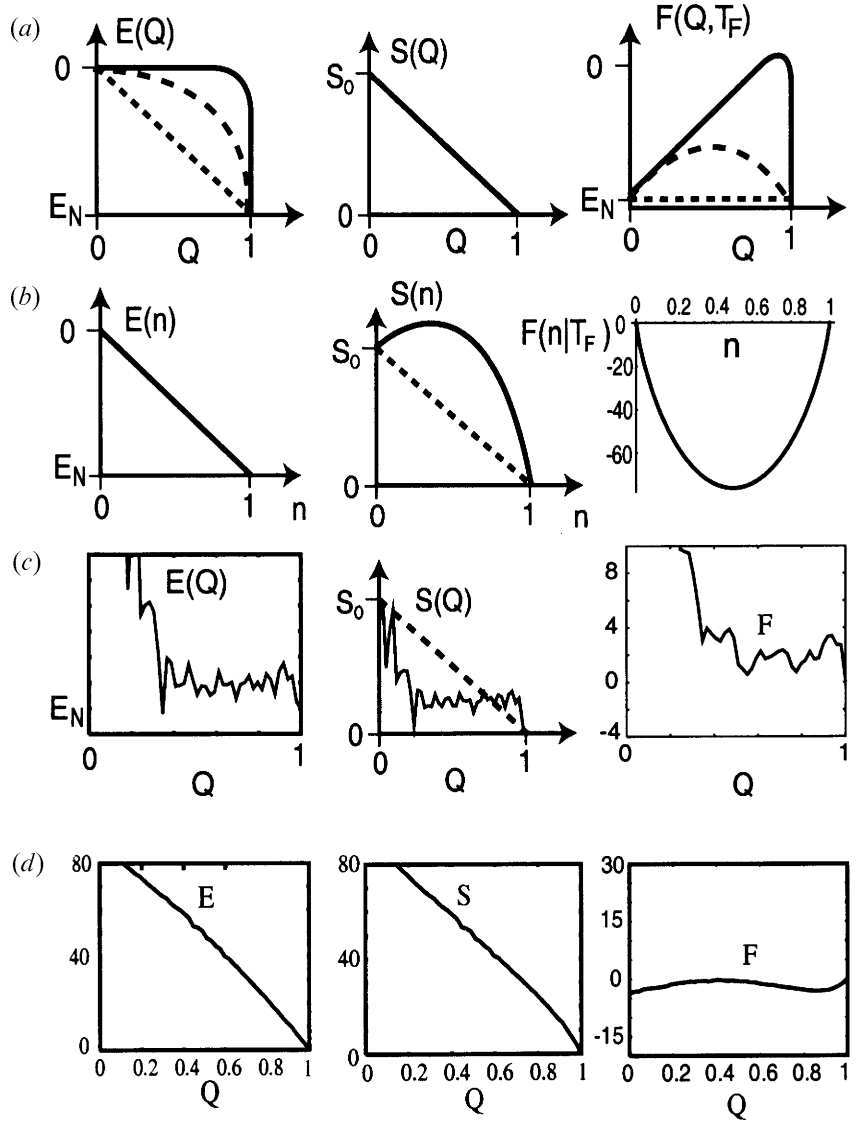

Fig. 5. Energy, entropy, and free-energy surfaces as a function of native similarity and at the folding temperature (T_F), for various models considered in the text. (*a*) For various degrees of cooperativity in the energy function. In the high-cooperativity limit where all the residues must be in spatial proximity to obtain energetic gain (solid lines), the Levinthal golf-course landscape is recovered. (*b*) For the independent-residue model, where amino acids are allowed to make independent fluctuations between native and non-native conformations. The U-shaped free-energy profile gives an unstable native conformation except at very low temperatures. (*c*) For a system modeling a random heteropolymer. Curves are obtained from an off-lattice HP model (Nymeyer *et al.* 1998). The thermal entropy is seen to be rugged and is numerically much less than the configuration entropy, indicating non-self-averaging effects are present. (*d*) For a system modeling a well-designed protein. Curves are obtained from an off-lattice Gō model (Nymeyer *et al.* 1998).

functional conformation and rapid and reliable folding to that particular conformation. For a given sequence of amino acids, the energy of a configuration is given by the sum of the energies of interactions between residues. Because of the numerous species of different amino acids (20) and their varied steric factors affecting hydrogen bonding and varied hydophobicities, proteins are characterized by a diversity of interaction energies. The landscape cannot be flat as assumed in the Levinthal argument, but must have some ruggedness to reflect this diversity of interaction energies. By the central limit theorem the total energies of configurations for a random sequence are Gaussianly distributed, and the energy landscape and density of states looks like those in Fig. 4*c*.

A result from applying spin-glass theory to a random heteropolymer (Bryngelson & Wolynes, 1987; Shakhnovich & Gutin, 1989) is that even the lowest energy states have sets of interactions that are frustrated: there are both attractive and repulsive interactions between parts of the chain. A random sequence typically cannot satisfy all of its structural constraints without bringing some relatively repulsive parts of the chain in proximity. Short chains with only a few allowable interaction energies (*n*-letter codes) may have several unfrustrated ground states. The problem in folding then becomes one of energetic discrimination between the native and competing states.

Another feature arising from the application of spin-glass theory to the heteropolymer is that the nearly degenerate ground states of the random heteropolymer are structurally dissimilar: they are disparate enough so that they share very few common interactions (Bryngelson & Wolynes, 1987; Shakhnovich & Gutin, 1989). This is a general feature of landscapes with frustrated interactions (Derrida, 1981; Mézard et al. 1984, 1986). To go from one nearly lowest energy state to another demands that we choose an almost completely different set of contacts, corresponding to a reshuffling of nearly all the interacting residue pairs and thus a globally different conformational state. For example the ground-state energy of the unfolded ensemble at $Q \approx 0$ in Fig. 4*c* is approximately equal to the ground-state energy of the whole system at $Q = 1$.

An obvious problem with sequences having the landscape topography of Fig. 4*c* is the stability of the native structure: small perturbations in the environment or sequence result in a new ground state with globally different topology. Large regions of the phase space would have to be restricted kinetically for the native state to maintain stability. If the protein randomly picked a basin to fall into, as in an un-chaperoned folding mechanism on a rugged unbiased landscape, a large fraction of proteins would have to be eliminated by degradation. In Part II.3, we calculate the number of basins on the landscape at the bulk glass transition temperature to be about the square root of the total number of collapsed states $\Omega_{\text{basin}} \approx \Omega^{0.5}$ (Plotkin *et al*. 1996). Since the helix-coil transition is well above this temperature, the protein will generically have transient helical structures which reduce the conformational entropy and renormalize the interaction energies. Thus the folding transition in protein sequences of length 60–80 residues may be roughly analogous to the folding transition in smaller minimalist models such as the 27-mer or 36-mer lattice model (Onuchic *et al*. 1995). Taking the respective number of collapsed conformational states for each, $\sim 10^5$ or $\sim 10^8$ (Flory, 1949, 1953; Sanchez, 1979; Bryngelson & Wolynes, 1990), the approximate number of basins is between 300 and 10000. So only about 1 protein in a thousand would be functional under this mechanism, meaning it is unlikely that this scenario would persist if evolution is concurrently selecting for reliable folding.

Another problem with typical random sequences is that the energetic variance of interactions is apparently not large enough: at biological temperatures, no conformational state is low enough in energy to be stable, and the system simply wanders configurationally about the landscape. The glass temperature (T_G), where a random sequence would be stable in one of the ground states, depends on the energetic variance and entropy in the system. One can estimate the energetic variance by comparing the timescale of reconfigurational dynamics in the collapsed molten globule (τ_{MG}) to the timescale of reconfigurations in the random coil, τ_{RC}, (Onuchic *et al.* 1995). At T_{F}, the mean time for reconfigurations on a rugged landscape is approximately

$$\tau_{\mathrm{MG}} \approx \tau_{\mathrm{RC}} \exp \left(S_{\mathrm{MG}} \frac{T_{\mathrm{G}}^2}{T_{\mathrm{F}}^2} \right) \tag{2.12}$$

[see Eq. (2.28)]. Reconfigurational timescales τ_{MG} in molten globule α-lactalbumin are about 10^{-3} s-10^{1} s (Kim *et al.* 1999); similar timescales are observed in apocytochrome b_{562} (Feng *et al.* 1994). On the other hand, reconfigurations of secondary structure in a free chain occur on timescales of about $\tau_{\mathrm{RC}} \approx 10^{-9}$ s. α-Lactalbumin has a chain length of 123, so the molten globule backbone entropy may be better estimated by a 64-mer minimalist chain, assuming secondary structure reduces the degrees of freedom by approximately 50 % as above. The chain entropy of a collapsed walk for the 64-mer is about 40–50 k_{B}, assuming a number of states per residue slightly less than 10. Here we assume a smaller, more commonly accepted value for the entropy of the collapsed chain than for the case of Barnase where the chain is in the coil. From Eq. (2.12) $T_{\mathrm{F}}/F_{\mathrm{G}}$ can be estimated:

$$\frac{T_{\mathrm{F}}}{T_{\mathrm{G}}} = \sqrt{\frac{S_{\mathrm{MG}}}{\ln(\tau_{\mathrm{MG}}/\tau_{\mathrm{RC}})}} \approx 1{\cdot}4 - 1{\cdot}9. \tag{2.13}$$

The T_{F} for α-lactalbumin is approximately 340 K ($T_{\mathrm{F}} \approx 330$ K for apocytochrome b_{562} is comparable). This gives a T_{G} of approximately

$$T_{\mathrm{G}} \approx 180\ \mathrm{K} - 240\ \mathrm{K}. \tag{2.14}$$

This temperature is well below biological temperatures, meaning that a typical random sequence is not particularly stable in any conformation at temperatures where it must function. Assuming about 100 k_{B} of entropy in the molten globule gives $T_{\mathrm{F}}/T_{\mathrm{G}} \approx 2{\cdot}3$ or $T_{\mathrm{G}} \approx 150$ K. Larger entropy in the molten globule only further reduces the temperature where the ground state is stable.

Put another way, a protein sequence must have enough extra energetic stability (E_{F}) in its native configuration to balance the free energy of the molten globule at the T_{F}. From Eq. (5.15) for the free energy derived later, the free energy of the unfolded or molten globule phase is given by

$$F_{\mathrm{MG}} = \bar{E} - TS_{\mathrm{MG}} - \frac{\Delta^2}{2T} \tag{2.15}$$

where $\bar{E}$ is the mean energy of the globule, S_{MG} is the configurational entropy of an unfolded but collapsed globule, and Δ^2 is the total variance of non-native interactions in the globule. We have neglected the presence of any native structure in the unfolded phase for now, and we are also ignoring any effects of concurrent collapse and folding, a subject treated later in Part II. The free energy in the folded phase is $\bar{E} + E_{\mathrm{F}}$, where E_{F} is the extra energetic stability in the folded structure. From Eq. (2.15) and $S = -\partial F/\partial T$, the thermal entropy in the molten globule can be obtained as

$$S(T) = S_{\mathrm{MG}} - \frac{\Delta^2}{2T^2} \tag{2.16}$$

and thus the thermal energy as

$$E(T) = \bar{E} - \frac{\Delta^2}{T}. \tag{2.17}$$

If there were no folded state of low energy, as the temperature is lowered, the thermal entropy would vanish at a T_G, given by

$$T_G = \frac{\Delta}{\sqrt{2\,S_{MG}}}, \tag{2.18}$$

where the system would be frozen into a low-lying state of energy

$$E_{GS} = E(T_G) = -\Delta\sqrt{2\,S_{MG}} \tag{2.19}$$

below the mean energy $\bar{E}$. Equations (2.18) and (2.19) are the relevant transition temperature and ground-state energy (E_{GS}) if the sequence were scrambled, i.e. for a random heteropolymer (RHP) of the same composition. Equating the free energies of the folded and unfolded states at T_F, and using Eq. (2.18) for the T_G in the unfolded ($Q \approx 0$) phase, gives finally:

$$\frac{E_F}{T_F} = -\left[1 + \left(\frac{T_G}{T_F}\right)^2\right] S_{MG} \approx -1{\cdot}4\,S_{MG}, \tag{2.20}$$

using an estimate for T_F/T_G around the average in Eq. (2.13).

However if the protein sequence were scrambled and reassembled along a chain, the energy of the ground state (in units of the original T_F) would be [see Eq. (2.19)]

$$\frac{E_{GS}}{T_F} = -2\,S_{MG}\frac{T_G}{T_F} \approx -1{\cdot}2\,S_{MG}, \tag{2.21}$$

so the ratio of a typical protein E_{GS} to the RHP E_{GS} is approximately

$$\frac{E_F}{E_{GS}} = \frac{1 + (T_F/T_G)^2}{2(T_F/T_G)} \approx 1{\cdot}1 \tag{2.22}$$

for $T_F/T_G \approx 1{\cdot}65$. Modest minimal frustration in the ground state, in this case resulting in only about 10% extra energy, can dramatically change the phase diagram and corresponding phase transition temperatures. Note that when $T_F = T_G$, $E_F = E_{GS}$ by Eq. (2.22).

Since T_F is well above T_G, the disorder should be self-averaging in the sense that the ground states of a random sequence should be about the same energy as the lowest energy states of the protein sequence when the protein is thread through the ensemble of collapsed structures unrelated to the native structure (the ensemble having essentially no native contacts). Equations (2.20) and (2.21) give

$$E_F - E_{GS} = -\left(1 - \frac{T_G}{T_F}\right)^2 T_F S_{MG} \tag{2.23}$$

or in other words the low-energy competing structures are about 10 $k_B\,T_{300\,K}$ or 25 KJ mol^{-1} above the energy of the native structure, using the numbers in the above example. This number may seem small; a factor of 2 in the entropy doubles the energy gap. On the other hand, the calorimetric measurements for the entropy loss on folding of Barnase result in an energy gap that is quite large (Fig. 3 inset). From Eq. (2.1), the extra stability is of the order 100 $k_B\ T_{300\ K}$. However the entropy inferred from the calorimetric data in Barnase may be closer to the coil entropy than the collapsed molten globule entropy relevant when considering competing states with the ground state.

2.2 Some effects of energetic correlations between structurally similar states

The above calculation overestimates T_F/T_G because it assumes the states on the landscape are uncorrelated in energy regardless of their structural similarity.[8] Inferred from the measured diffusion timescales, the energetic variance between states and thus T_G are much higher for a correlated landscape. Accounting for correlations in the landscape gives a significantly lower value of T_F/T_G as follows. First, a new feature when correlations are accounted for is the existence of a dynamical glass temperature (T_A) where reconfigurational barriers vanish. The fact that the diffusion timescales in many proteins are orders of magnitude different in the molten globule and random coil indicates that for these proteins the folding transition temperature (T_F) (roughly where the measurements are made) is below T_A. Below T_A, the mean reconfigurational diffusion time $\langle\tau\rangle$ is calculated as a barrier-hopping process, and can be expressed in terms of the fraction of entropy remaining at the barrier peak $\tilde{s}^{\ddagger}$, the position of the barrier peak $q^{\ddagger}$, measured as the fraction of residue–residue contacts shared with the local minimum configuration, and the reduced temperature $\tilde{T}$ in units of T_G. The result for the mean reconfiguration time is (see Part II)

$$\langle\tau(T)\rangle \approx \tau_0 \exp\left[S_{MG}\left((1-q^{\ddagger}-\tilde{s}^{\ddagger})-(2-q^{\ddagger})\,(1-\tilde{T}^{-1})^2\right)\right]. \tag{2.24}$$

Using $\langle\tau(T_F)\rangle \approx 10^{-3}$ s and $S_{MG} \approx 50\ k_B$ in Eq. (2.24), and the estimates within the correlated landscape theory of $\tilde{s}^{\ddagger} \cong 0{\cdot}4$ and $q^{\ddagger} \cong 0{\cdot}3$ (Wang *et al.* 1997), gives a value for T_F/T_G of only about 1·1, and thus $T_G^{corr} \approx 300$ K. This estimate (which strongly smoothes the landscape) and the REM estimate (which over-represents the ruggedness) probably bracket the true value for T_G. Nevertheless, the GREM estimate suggests that some sequence compositions may have T_G significantly higher than previous estimates. Assuming entropies of about 100 k_B in Eq. (2.24) gives $T_F/T_G \approx 1{\cdot}4$ or $T_G \approx 240$ K.

On the other hand, the value of T_G^{corr} is lowered when the energetic cooperativity of interactions is taken into account. The interaction energies arise from the free energy of burying hydrophobic surface area, and hydrogen bonding between side-chains (and backbone in the case of helices). The buried surface area for three hydrophobic residues is more than the sum of the surface area buried pair by pair, so there is an explicitly cooperative interaction in folding which reduces energetic gain due to native structural similarity until a larger fraction of native structure is present, more so than if pair interactions were accounted for alone. Moreover, models with Hamiltonians containing the sum of residue–residue interactions may have cooperativity present because a pair-wise potential overestimates the side-chain entropy lost in packing amino acids together: the entropy loss of packing a third

[8] Such an approximation is also referred to as the random energy model approximation or REM approximation. Accounting for correlations in the energetics of similar states on the landscape is sometimes referred to as the generalized random energy model approximation or GREM.

side-chain has already been mostly accounted for in the pairwise terms of three amino acids, so the free energy gain is larger as more residues are involved in the interaction. Renormalizing the degrees of freedom in the system to coarse-grained residue–residue interactions can introduce cooperativity in the effective interactions in the Hamiltonian. While these cooperative effects may depend on the reference state (conformation corresponding to the trap to be escaped from), making it difficult to apply to configurational diffusion, it is clear that such effects make the landscape more rugged, and estimates of T_{G} and search times are shifted back towards the REM values. In fact for a spin-glass model with p-spin interactions, the REM distribution of states was obtained in the $p \to \infty$ limit (Derrida, 1981), and for the p-contact Hamiltonian we introduce in Part II and applied to folding, the REM is obtained in the same limit.

Even if a sequence had sufficient energetic variance to have $T_{\mathrm{G}} \approx T_{\mathrm{bio}}$ folding would tend to be slow on an unbiased landscape. The escape rate from a trap with energy E_i is the sum of the rates of escape across saddle points on the landscape:

$$k_i = k_0 \sum_s \mathrm{e}^{-(E_{\mathrm{s}} - E_i)/T}. \tag{2.25}$$

The number of saddles should scale like the degree of connectivity on the landscape, which we postulate to scale linearly with N (as the number of dihedral angles in the protein for example). Then the sum in Eq. (2.25) is approximately the number of saddles times an average over the saddle-point energies, which we heuristically take to follow the same Gaussian distribution as the density of states. This amounts to a REM approximation for the saddles. Then the averaged rate is

$$\bar{k}_i = k_0 c N \, \mathrm{e}^{E_i/T} \int \mathrm{d}\, E_{\mathrm{s}} P(E_{\mathrm{s}}) \, \mathrm{e}^{-E_{\mathrm{s}}/T} \approx k_0 c N \, \mathrm{e}^{E_i/T} \, \mathrm{e}^{\Delta^2/2T^2}, \tag{2.26}$$

assuming a Gaussian distribution for $P(E_S)$ with variance Δ^2.

The barrier height for each of the saddles depends on the energy of the original state: at lower temperatures, lower energies tend to be occupied as $E(T) = -\Delta^2/T$. Taking the state i to have the most probable energy at temperature T,

$$\bar{k}(T) \approx k_0 c N \, \mathrm{e}^{-\Delta^2/2T^2} \tag{2.27}$$

$$\bar{\tau}(T) \equiv \frac{1}{\bar{k}} = \frac{\tau_0}{cN} \mathrm{e}^{S_{\mathrm{MG}} T_{\mathrm{G}}^2/T^2} \tag{2.28}$$

which at T_{G} essentially reproduces the Levinthal estimate in Eq. (2.3), now because the reconfigurational barriers to travel between globally different states scale extensively, as opposed to the golf-course scenario where an exponential number of states must be sampled to find a specific one.[9] For disordered systems such as the RHP, it is believed that there can exist polynomially many nearly degenerate ground states (Marinari *et al.* 1996, 2000) ($\Omega_{\mathrm{GS}} \sim N^{\alpha}$). Then the characteristic time at T_{G} to find the true ground state or native state for a RHP is

$$\bar{\tau}_{\mathrm{F}} \sim N^{\alpha-1} \, \mathrm{e}^{N s_{\mathrm{MG}}}, \tag{2.29}$$

where s_{MG} is the entropy per residue on the chain. The ground state of a typical sequence of amino acids strung together is not kinetically accessible on biological timescales at the

[9] This search time for a random sequence with rugged landscape should perhaps be called the Bryngelson–Wolynes search time (Bryngelson & Wolynes, 1987, 1989), since these authors first recognized the search problem is real for typical (non-protein-like) sequences.

temperatures where it would be thermodynamically stable; when the system is cooled down enough so that it stays in the ground-state conformation, it takes too long to get there. Note that the relaxation time in Eq. (2.28) has a stronger temperature dependence than an Arrhenius law. An equation of this form was used by Ferry over 50 yr ago to describe viscosity in liquids and polymers (Ferry, 1950), and many of the same themes as those in the above calculation arise in theories of reconfiguration in supercooled liquids (Kirkpatrick *et al*. 1989; Xia & Wolynes, 2000).

Reconfiguration barriers on a correlated landscape are significantly reduced from the estimate in Eq. (2.28) (Wang *et al*. 1997). Evaluating Eq. (2.24) at $T = T_{\mathrm{G}}$ gives

$$\langle\tau\rangle^{\mathrm{corr}}\,(T_{\mathrm{G}}) = \tau_0 \mathrm{e}^{S_{\mathrm{MG}}(1-q^{\ddagger}-\tilde{s}^{\ddagger})} \approx \tau_0\, \mathrm{e}^{0\cdot 3 S_{\mathrm{MG}}}. \tag{2.30}$$

The barriers are about a third of the REM barriers, or the effective chain length is only a third as long. Accounting for correlations on the landscape greatly reduces the search time, and it is worthwhile here to recalculate it. The search time on the uncorrelated rugged landscape is already reduced because collapse and partial helical content reduce the entropy to say $\sim 50\ k_{\mathrm{B}} - 100\ k_{\mathrm{B}}$ as above. Using Eq. (2.28) at $T = T_{\mathrm{G}}$ gives

$$\tau_{\mathrm{BW}} \approx (10^{-9}\ \mathrm{s})(\mathrm{e}^{50} - \mathrm{e}^{100}) \sim 10^{5}\ \mathrm{years} - 10^{27}\ \mathrm{years}. \tag{2.31}$$

The search time on a correlated landscape is

$$\tau_{\mathrm{corr}} \approx (10^{-9}\ \mathrm{s})(\mathrm{e}^{(0\cdot 3)(50)} - \mathrm{e}^{(0\cdot 3)(100)}) \sim 1\ \mathrm{ms} - 3\ \mathrm{h}, \tag{2.32}$$

which is within the folding times of real proteins. While the entropy here may be underestimated, this illustrates the dramatic effect that energetic correlations have on the dynamics. It also serves as an example of how reductions in the degrees of freedom due to local mechanisms such as transient helix formation, and generic effects such as sequence-independent collapse, can reduce the search space to a volume small enough that the native state may be found on biological timescales. This result as well as the GREM estimate for T_{G} suggest other possible mechanisms of folding such as diffusion on a nearly unbiased landscape with smaller than average barriers (Plotkin & Wolynes, 2002). However the experimental evidence of robustness to environmental or mutational perturbations still favors a funneled landscape description. Additionally the comments above on cooperative effects apply here to the estimates of diffusion times, thus increasing the value of τ_{corr}.

While pathologies must be present in smooth landscapes for NP completeness (e.g. the golf-course topography), NP completeness tends to be a generic property of systems with rugged landscapes (Kirkpatrick *et al*. 1983) such as the Ising spin glass, Potts spin glass, real (structural) glass, as well as RHPs for the reasons discussed above. Variations on the motif of long-range interactions inducing NP completeness in polymers are present here as well. For example the ferromagnetic random-field Ising model has polynomial complexity. (d'Auriac *et al*. 1985), while the ground state search for a spin glass is an NP complete problem in dimensions larger than 2 (Barahona, 1982; Bachas, 1984).

3. Resolution of problems by funnel theory

To have a new phase (the folded state) that is stable at biological temperatures, there must exist structural conformations of significantly lower energy than the ground states of a RHP. This is analogous to the low-energy, low-entropy crystal state which nucleates out of a super-cooled liquid at low enough temperature. However here such a low-energy state is only allowed by making mutations in the amino-acid sequence, through the process of evolution, until ground states are found that have most interactions satisfied, and their energetic frustration thus minimized (Bryngelson & Wolynes, 1987; Goldstein *et al.* 1992; Leopold *et al.* 1992; Shakhnovich & Gutin, 1993; Onuchic *et al.* 1995; Bornberg-Bauer & Chan, 1999; Buchler & Goldstein, 1999).

As far as protein function is a prerequisite for replication, and function is contingent on a fairly well-defined native structure, evolution then selects for native stability, which is tantamount to selecting for reliable folding.[10] Insofar as rapidly reproducing life may more readily adapt to its environment, more rapid dynamics, and consequently more rapid folding, may be selected for; slower folding simply sets a different timescale for cell dynamics and an organism's lifetime. However, given a set of timescales for an organism, slower folding proteins may be selected against because of rate-limiting effects on cellular dynamics. This is one factor among many which has an effect (in this case shortening) on the lifespan of a particular species over generations.

Since the native structure is the lowest in energy, and structurally similar configurations must be correlated energetically to it, we should expect the shape of the landscape for a protein to have a funnel topography, at least in the vicinity of the native structure, and correspondingly a funneling of the density of states, as in Fig. 4*d*. In the following sections and in Part II we will focus on how the energetic correlations due to structural similarity determine the energy landscape, and how the underlying energy landscape then governs the free-energy landscape, dynamics, and folding mechanism.

A funneled landscape is robust to environmental perturbations as well as sequence mutations, because potentially competing low-energy states are still similar in structure (see Fig. 4*d*). This may be quantified by noting that the density of states with energy E and fraction of shared native contacts Q with given state N having energy E_{N} is

$$n_Q(E\,|\,E_{\mathrm{N}}) \approx \exp\left(S(Q) - \frac{(E-QE_{\mathrm{N}})^2}{2\Delta^2(1-Q)}\right). \tag{3.1}$$

This equation, derived in Part II.4, is equivalent to the total number of states, exp $S(Q)$, times a gaussian with mean linearly correlated to E_{N}, and a variance due to non-native interactions which linearly decreases with similarity to the native state N. Equation (3.1) applies for an arbitrary landscape with pair-wise interaction energies, whether funneled or not, in that E_{N} need not be low. Here we consider state N to be the ground state or native state of the system. One approximation made here is that the total number of contacts remains roughly constant independent of Q. This approximation should be good below the collapse temperature, however special care should be taken in general to account for cases where collapse and folding are concurrent, as in well-designed proteins (see Part II.5), or for cases where new ground states emerge which are not entirely collapsed (Yue *et al.* 1995).

Let the energies now be perturbed, by the environment or sequence mutations, so that the new energy E_{i}' of state i is the old energy E_i plus the sum of perturbations on the contact energies:

$$E_i' = E_i + \sum_{\alpha<\beta} \delta\epsilon_{\alpha\beta}\delta_{\alpha\beta}(i). \tag{3.2}$$

[10] Many proteins are not fully structured *in vivo*; for these proteins the selection for function may involve the co-evolution of stability of the protein–substrate complex (Wright & Dyson, 1999; Shoemaker *et al.* 2000); folding and binding then occur concurrently.

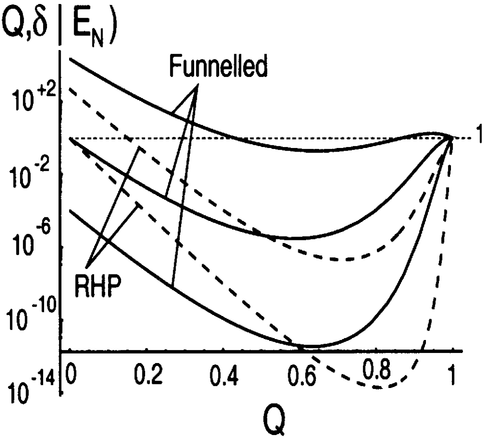

Fig. 6. Average number of states $n_<$ which dip below the putative ground state after a perturbation, as a function of Q, for originally funneled (solid curves) and random heteropolymer (dashed curves) landscapes. The plot is on a log scale; numbers below the dashed horizontal line are less than 1 and insignificant. From Eqs. (2.20) and (2.21), and using $S_{\mathrm{MG}} \approx 50\ k_{\mathrm{B}}$, the relative stability $\alpha \cong 0{\cdot}17$ for the funneled landscape ($\alpha = 0$ for the RHP landscape). The lower dashed curve is the RHP landscape before perturbation. After a perturbation of 10% of the original variance ($c = 0{\cdot}1$), the funneled ground state remains robust (lower solid curve); any new ground states which emerge are close to $Q = 1$ and thus similar in structure to the original native state. For perturbations $c = 0{\cdot}1$ new ground states emerge at $Q = 0$ for the RHP (upper dashed curve): the RHP landscape is not robust. For a funneled landscape at the critical value of perturbation, $c = \alpha$ from Eq. (3.7), and $\sim \mathcal{O}(1)$ new ground states emerge at $Q = 0$ (middle solid curve). Applying a sufficiently large perturbation (e.g. $c = 0{\cdot}35$ as in the top solid curve) eventually destroys the overall funnel topography for an originally funneled landscape.

The sum is over residue indices and $\delta_{\alpha\beta}(i) = 1$ if residues α and β are in contact in state i, zero otherwise. On average the perturbation in contact energy is taken to be zero, $\overline{\delta\epsilon_{\alpha\beta}} = 0$ so $\bar{E}' = \bar{E}$, but has a non-zero variance so that the new energetic variance is $\overline{\delta E'^2} = \overline{\delta E^2} + c\Delta^2 = (1+c)\Delta^2$ with the coefficient $c > 0$. Then the density of states after the perturbation is

$$n_Q(E'|E_{\mathrm{N}}) = \mathrm{e}^{S(Q)} P_Q(E'|E_{\mathrm{N}}) \approx \exp\left(S(Q) - \frac{(E' - QE_{\mathrm{N}})^2}{2\Delta^2(1-Q+c)} \right). \tag{3.3}$$

The average number of states at Q that dip below the original ground state (E_{N}) after the perturbation is

$$n_<(Q, c|E_{\mathrm{N}}) = \mathrm{e}^{S(Q)} \int_{-\infty}^{E_{\mathrm{N}}} \mathrm{d}E' P_Q(E'|E_{\mathrm{N}}). \tag{3.4}$$

We take the entropy to be a linear function of Q as a first approximation: $S(Q) = S_0(1-Q)$, and we let the original ground-state energy $E_{\mathrm{N}} = (1+\alpha)\ E_{\mathrm{GS}}$, where $E_{\mathrm{GS}} = -\Delta\sqrt{2S_0}$ is the ground-state energy of a RHP, and $\alpha \geqslant 0$. Then taking the integral in Eq. (3.4), the number of states below the original ground state after the perturbation is

$$n_<(Q, c|\alpha) = \mathrm{e}^{S_0(1-Q)} \mathrm{erfc}\left(\frac{(1+\alpha)(1-Q)}{1-Q+c} \sqrt{S_0} \right) \tag{3.5}$$

for a given stratum of states at Q, relative strength of perturbation c, and degree of minimal frustration α. Here erfc(x) is the complementary error function (Abramowitz & Stegun, 1972).

The function in Eq. (3.5) is plotted in Fig. 6 for funneled as well as RHP landscapes.

A funneled landscape is robust to an extensive amount of perturbation, while on the RHP landscape even small perturbations cause a new ground state to emerge which is structurally dissimilar ($Q = 0$) from the original ground state. For a funneled landscape, a large enough perturbation can introduce structurally dissimilar ground states. Evaluating Eq. (3.5) at $Q = 0$ for moderate to large system sizes gives

$$n_{<}(0, c|\alpha) \approx \exp\left(S_0\left[1-\left(\frac{1+\alpha}{1+c}\right)^2\right]\right), \tag{3.6}$$

which is less than 1 (i.e. the native state remains the ground state) so long as $c < \alpha$, or (put in terms of the original dimension-full quantities) so long as the relative variance in the perturbation is less than the relative extra stability of state N:

$$\frac{\Delta^2_{\mathrm{pert}}}{\Delta^2} < \frac{E_{\mathrm{N}} - E_{\mathrm{GS}}}{E_{\mathrm{GS}}}. \tag{3.7}$$

Similar conclusions of native stability to perturbation have been drawn in references (see Shakhnovich & Gutin, 1991; Bryngelson, 1994; Pande *et al.* 1995, 1997).

In addition to tolerating moderate perturbations in pressure, temperature, acidity, or denaturant, proteins are robust to a large variety of sequence mutations, pointing to a new problem of determining the ensemble of sequences which fold to (or are 'designable' to) a given structure. This problem, mentioned earlier in the introduction, is known as the inverse folding problem. Again one of the general principles which has emerged is that symmetric structures maximize the number of interactions and so may have deeper funnels, allowing for larger perturbations in sequence before new dissimilar ground-state structures emerge by Eq. (3.7). Such a result may be incorporated into the correlated landscape framework by allowing a different conformational entropy $S(Q)$ for different structures in Eq. (3.4), as well as incorporating the structural dependence of local contact densities into the energetic variance.

A consequence of a funneled landscape topography is that the native structure is kinetically accessible at the temperatures where it is thermodynamically stable. α-Lactalbumin is functional at about 310 K, which is below its folding temperature of 340 K, and well above the estimated T_{G} of ≈ 240 K for a random sequence of the same amino-acid composition.

Thus for any sequence to have a window in temperature where it is foldable, the sequence must satisfy the criterion

$$\frac{T_{\mathrm{F}}}{T_{\mathrm{G}}} > 1, \tag{3.8}$$

which again is another way of saying that the energy landscape must be funneled. To see this note that the T_{F} is where the free energies of the unfolded and folded states are equal, and equate Eq. (2.15) with the native energy $\bar{E} + E_{\mathrm{N}}$ to give:

$$T_{\mathrm{F}} = \frac{|E_{\mathrm{N}}|}{2S_0}\left(1+\sqrt{1-\frac{2\Delta^2 S_0}{E_{\mathrm{N}}^2}}\right). \tag{3.9}$$

Or by rearranging Eq. (2.22), the criterion (3.8) becomes

$$\frac{T_{\mathrm{F}}}{T_{\mathrm{G}}} = \frac{E_{\mathrm{N}}}{E_{\mathrm{GS}}} + \sqrt{\left(\frac{E_{\mathrm{N}}}{E_{\mathrm{GS}}}\right)^2 - 1} > 1, \tag{3.10}$$

where E_{GS} is the ground-state energy of a RHP, as in Eq. (2.19). By inspection Eq. (3.10) is equivalent to

$$|E_{\mathrm{N}}| > |E_{\mathrm{GS}}| \tag{3.11}$$

or the landscape is funneled. For a sequence to be thermodynamically stable at biological temperatures, $T_{\mathrm{bio}} < T_{\mathrm{F}}$, so biological temperatures must fall within the window between T_{G} and T_{F}, i.e. $T_{\mathrm{G}} < T_{\mathrm{bio}} < T_{\mathrm{F}}$.

The funneled shape of the landscape biases the average sampling of states so that transitions to native isomerization angles have higher rates k_{N} than those rates k_0 from correct to incorrect isomerizations (Bryngelson & Wolynes, 1989; Zwanzig *et al*. 1992). Thus not all sequences of reconfigurational events are equally probable: there is a drift component in the conformational diffusion towards the native state.

As in the case of crystallization, the minimally frustrated character of the ground state leads to a simple first-order folding transition for the system, as opposed to the multi-exponential relaxation present in RHPs, where many barriers contribute to the relaxation rate. Because $T_{\mathrm{F}} > T_{\mathrm{G}}$, the thermal entropy in the unfolded and transition state ensembles is large under folding conditions (T_{G} is where the thermal entropy vanishes); many states may be potentially occupied during folding at functional temperatures.

Under these conditions, the entropic force driving escape from low-energy traps is large, and relaxations are relatively fast compared to the folding time. Then the free energy $F(Q)$, at similarity Q to any given structure functions as an effective solvent-averaged potential of mean force, as in transition-state theory in condensed media (Miller, 1974; Pechukas, 1976; Caldeira & Leggett, 1983; Pollak, 1992). A given protein configurationally diffuses on this surface, and the probability distribution $P(Q, t|Q_0, t_0)$ obeys a Fokker–Planck equation, in principle modified by a Q-dependent, frequency-dependent and temperature-dependent diffusion coefficient, $D(Q, \omega, T)$.[11]

The Laplace-transformed Fokker–Planck equation for $P(Q, \omega) = \mathscr{L}P(Q, t)$ on the free-energy potential $F(Q, \omega, T)$ is described further in Part II (see also Bryngelson & Wolynes, 1989). However, some general simplifying principles emerge when overdamped diffusion on a potential of mean force is applicable: the mean first-passage time depends exponentially on the barrier height in units of the temperature, and the diffusion coefficient enters only into the prefactor. The rate for barrier crossing is then (see Kramers, 1940; Miller, 1974; Grote & Hynes, 1980; Pollak *et al*. 1990)

$$k(\Delta F^{\ddagger}, D) = k_0(D)\mathrm{e}^{-\Delta F^{\ddagger}/T}, \tag{3.12}$$

[11] The diffusion is overdamped in that the Reynolds number for residue motions is exceedingly small. Taking the velocity scale to be about 100 Å/1 ms from the measurements noted above, the length scale of the moving object to be about the size scale of a helical turn ~ 5 Å, and the kinematic viscosity ν of water 10^{-6} m^2 s^{-1}, the Reynolds number $R = VL/\nu$ is $\approx 10^{-8}$. Put another way, if a residue were set in motion at the velocity V above, it would come to rest by Stokes' frictional drag in a distance of about $\mathrm{mV}/\sigma\pi\eta L$. Using the mass of an average residue ($m = 110$ Da), and the dynamic viscosity of water $\eta = 10^{-3}$ N/s m^{-2}, this distance is only about a factor of 10^{-10} of the residue's own size – a number which is not particularly meaningful given the finite times between molecular collisions, except to indicate the reaction is well within the spatial diffusion limit. For fuller accounts of the dynamics of proteins in the folded state, see Karplus & McCammon (1983), McCammon & Harvey (1987), Smith (1991), Kitao *et al*. (1991).

which under strongly native conditions where the back-reaction may be neglected, controls the decrease in unfolded protein population. $N_{\mathrm{u}}(t)$ with time by

$$N_{\mathrm{u}}(t) = N_{\mathrm{u}}(0)\,\mathrm{e}^{-k(\Delta F^{\ddagger}, D)t}. \tag{3.13}$$

So folding rates may be determined by analyzing the free-energy potential $F(Q)$, for the system, as well as the friction for diffusion on $F(Q)$. If there is more than one dominant barrier contributing to the folding rate, the decrease in unfolded population may have a multi-exponential time-dependence.

The free-energy potential may be decomposed into energetic and entropic terms:

$$F(Q, T) = E(Q, T) - TS(Q, T). \tag{3.14}$$

When Q is measured relative to the native structure, the energy $E(Q)$ tends to be a monotonically decreasing function of Q, embodying the funneling mechanism to the native state in the presence of energetic correlations to the $Q = 1$ structure. The entropy is also in general a monotonically decreasing function of Q, and measures the decrease in number of conformational states as more structural constraints are imposed. Any nonlinearity in $F(Q)$ arises from a mismatch between these two terms. If there were no correlations, or equivalently if the interactions were so cooperative that all the residues had to be in place in the native structure to yield the required stabilization energy, the energy profile of Fig. 5*a* (solid line) is recovered. Then the folding free-energy barrier is T times the total entropy, and the mean folding time

$$\tau = \tau_0 \mathrm{e}^{\Delta F^{\ddagger}/T} = \tau_0 \mathrm{e}^{S_0} \tag{3.15}$$

is the Levinthal time once again.

Schematics of the energy, entropy, and free energy as a function of native similarity are shown in Fig. 5 for the various scenarios discussed previously.

3.1 Physical origin of free-energy barriers

Numerous experimental probes observe a cooperative, two-state-like folding transition at the folding temperature for many small proteins (Pohl, 1969; Privalov, 1979; Jackson & Fersht, 1991; Horovitz & Fersht, 1992; Huang & Oas, 1995; López-Hernández & Serrano, 1996; Schindler & Schmid, 1996; Burton *et al.* 1997; Eaton *et al.* 1997, 1998; Martinez *et al.* 1998; Plaxco *et al.* 1998; Chiti *et al.* 1999; Fulton *et al.* 1999). In larger proteins, as well as some smaller ones, partially native intermediates may tend to accumulate for various reasons (Creighton, 1974; Baldwin, 1975; Kim & Baldwin, 1990; Matouschek *et al.* 1990; Radford *et al.* 1992; Bai *et al.* 1995; Deng & Smith, 1998). One way to see why the free-energy profile of the independent residue model does not produce two-state behavior is to appeal to a nucleation picture of folding in the capillarity limit. In this scenario, native structure is formed which is contiguous in space, and the interface thickness l_{IF} between the native and non-native region is much less than the size scale of the protein: $l_{\mathrm{IF}} \ll L \sim 20-100$ Å. One can readily see the capillarity picture is at best a good approximation for protein-sized systems, however it often provides an accurate description of folding barriers (Finkelstein & Badretdinov, 1997; Wolynes, 1997b; Portman *et al.* 1998) and also provides a good starting point for understanding the origin of barriers by treating folding as a nucleation process, analogous to conventional first-order phase transitions (Gunton *et al.* 1983).

In the phenomenological picture of nucleation, the stable bulk phase grows into the metastable phase in such a way as to minimize the free energy of the intermediate configurations. So for example in an Ising system where there is an energetic gain for neighboring spins to be aligned and an equal in magnitude cost for them to be counter-aligned, the stable phase grows in spherically to maximize the relative number of favorable to unfavorable interactions. At finite temperature the interface is roughened by entropic driving forces and has some width (Weeks *et al.* 1973), but there is still a characteristic critical nucleus size and barrier height beyond which the growth process is thermodynamically downhill.

For proteins there is a relative gain to have native parts clustered in space to maximize their interactions, as well as surface cost because residues have less contacts there. The contact density is somewhat reduced by the presence of the backbone chain. However there is also an entropic cost for the polymer halo dressing the native core: the entropy per residue of the unfolded state is lost for the residues in the native core, plus there is an additional entropy loss because the remaining polymer chain is not completely free, but rather emerges from and re-enters into the core in various places (see e.g. Fig. 10, inset). The reduction in entropy of a free chain by adding these topological constraints is the surface entropy cost of the nucleus. This results in a convex-down entropy function as a function of the order parameter, e.g. number of folded residues or number of native contacts, so there is an entropic contribution to the barrier, because again free-energy barriers arise here from incomplete cancellation of energy and entropy.

The large mixing entropy in the independent residue model is manifested in the capillarity approximation by the entropy of nucleus placement. This scales only as ln N and is comparatively negligible here. The true loss in polymer conformational entropy is not accounted for in the independent residue model: the non-independence of residues alters the form of $S(Q)$, and demanding a uniform field results in a higher free energy than the true free energy, which may be governed by configurations closer to capillarity. Accounting for polymer entropy losses in the mean-field limit, as well as for an inhomogeneous contact field, is treated in Part II and the result is also a convex-down entropy function. To illustrate heuristically how polymer entropy loss is a larger contribution than mixing entropy gain, consider the conformational loss to form a contact between two residues separated on the chain by sequence length l. To the first approximation this is the random flight chain conformational entropy loss: $\Delta S \approx \frac{3}{2} \ln (a/l)$, where a is a coefficient $\sim \mathcal{O}(1)$ (Shoemaker *et al.* 1999; Plotkin & Onuchic, 2000). At $Q = \frac{1}{2}$ the mixing entropy is the largest – there N residues have $N \ln 2$ mixing entropy. Comparing the entropy terms $|(N/2)\frac{3}{2} \ln l| > |N \ln 2|$ when $l \gtrsim 2{\cdot}5$. Since this is true for essentially all loops in a protein (even the shortest α-helices have $l \approx 3$–4) the polymer entropy loss is typically greater in magnitude than the combinatorial entropy gain, and the entropy will tend to decrease as Q increases. Moreover, the rate of entropy loss tends to be larger in an unconstrained protein, because polymer loops are allowed to be longer, whereas in a near-native protein, entropy losses are smaller because contacts are only zipping up remaining non-native polymer. Chain stiffness and excluded volume further enhance the dominance of polymer entropy loss over mixing entropy.

Another important point is that the binary fluid approximation for the mixing entropy used in Eq. (2.8) is in fact an over-estimate, since as a result of the connectivity of the backbone chain not all residues may be independently pinned down in the folded conformation without pinning other perhaps neighboring residues down as well. This effect is treated further in Part

II.2. Presently no definitive theory exists capable of distinguishing between the capillarity and mean-field limits in folding or heteropolymer reconfiguration, but theories accounting for heterogeneity in folding (Shoemaker *et al.* 1997, 1999; Portman *et al.* 1998; Plotkin & Onuchic, 2000) begin to address this issue.

Collecting the energetic and entropic terms into a bulk free-energy gain coefficient (per folded residue) $f(T)$ and a surface free-energy cost per folded residue $\sigma(T)$, the free energy as a function of the number of folded residues N_{F}, can be written as

$$F(N_{\mathrm{F}}, T) = F(0, T) - f(T)N_{\mathrm{F}} + \sigma(T)N_{\mathrm{F}}^{z}, \tag{3.16}$$

where $F(0, T)$ is the free energy of the unfolded state [$F(N) = E_{\mathrm{N}}$ is the free energy of the folded state]. The scaling exponent $z = \frac{2}{3}$ if there is no roughening of the interface, otherwise it may be smaller, producing a weaker surface cost.

As the system size $N \to \infty$, the bulk free-energy gain $f(T_{\mathrm{F}}) \to 0$ at the folding temperature, defined through $F(0, T_{\mathrm{F}}) = F(N, T_{\mathrm{F}})$ in Eq. (3.16). But for a finite-sized system such as a protein, there is still a non-negligible surface to volume ratio in the fully folded state at $N_{\mathrm{F}} = N$, so $-f(T_{\mathrm{F}})$ is slightly negative rather than zero, and the T_{F} is slightly depressed from the bulk value. Letting $f_{\infty}(T) \equiv (F(N) - F(0, T))/N$, the bulk coefficient is obtained from Eq. (3.16):

$$f(T) = -f_{\infty}(T) + \frac{\sigma(T)}{N^{1-z}}. \tag{3.17}$$

At T_{F}, $f_{\infty} = 0$ and $f(T_{\mathrm{F}})$ dies away as $\sigma(T_{\mathrm{F}})N^{-1+z}$ as explained above [the surface tension $\sigma(T)$ is intensive]. Roughening, when it occurs, introduces a scaling of the surface tension with N_{F}, given in three dimensions by Villain (1985), Kirkpatrick *et al.* (1989) and Wolynes (1997b)

$$\sigma_{\mathrm{tot}}(N_{\mathrm{F}}, T) = \frac{\sigma(T)}{N_{\mathrm{F}}^{\frac{1}{6}}}. \tag{3.18}$$

This reduces the exponent z from $\frac{2}{3}$ to $\frac{1}{2}$, and weakens the surface cost, decreasing the size of the critical nucleus. Since the surface of the folded protein is not particularly roughened, the exponent z may begin to approach $\frac{2}{3}$ again as $N_{\mathrm{F}} \to N$.

Setting $\partial F(N_{\mathrm{F}})/\partial N_{\mathrm{F}} = 0$ gives the critical nucleus size,

$$N_{\mathrm{F}}^{\ddagger} = N n_{\mathrm{F}}^{\ddagger} = \left(\frac{z\sigma}{f}\right)^{1/(1-z)}, \tag{3.19}$$

where $0 \leqslant n_{\mathrm{F}} \leqslant 1$, and free-energy barrier height:

$$\Delta F^{\ddagger} = \sigma(1-z)\left(\frac{z\sigma}{f}\right)^{z/(1-z)} = \sigma(1-z)n_{\mathrm{F}}^{\ddagger z}N^{z}. \tag{3.20}$$

Thus the barrier arises from surface cost, and scales like N^{z}. In some cases, the folding free-energy profile may be downhill and barrier-less, corresponding to a vanishing of the surface tension in the capillarity model.

The barrier is small, and leads to a relatively fast folding rate. For example, taking $N = 100$, $z = \frac{1}{2}$, and $\sigma = 1\ k_{\mathrm{B}}T_{\mathrm{F}}$, the barrier at T_{F} is located at $N_{F}^{\ddagger} = 25$ and has a height $\Delta F^{\ddagger}$ of $2\frac{1}{2}\ k_{\mathrm{B}}T_{\mathrm{F}}$. The folding rate at T_{F} is then given by Eq. (3.12):

$$k_{\mathrm{F}} \approx k_{0}\, \mathrm{e}^{-2{\cdot}5} \approx 0{\cdot}1\ k_{0}. \tag{3.21}$$

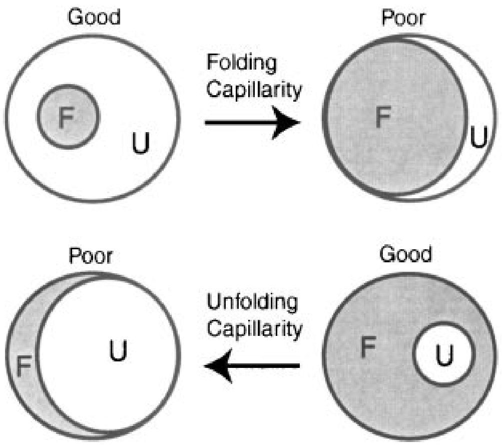

Fig. 7. Illustration of the nuclei configurations involved in folding and unfolding capillarity. The folding capillarity approximation is accurate at low degrees of nativeness, while the unfolding capillarity approximation is accurate at high nativeness. Insisting that folding capillarity holds at high nativeness results in configurations having anomalously high surface tension. Relaxing this condition at high nativeness results in configurations corresponding to unfolding capillarity which minimize the free energy.

A folding rate of 1 ms has a prefactor of (0·1 ms)$^{-1}$ in this model. More cooperative interactions correspond to a larger surface tension σ.

The stability problems of the independent residue model are not present here. Taking the values used earlier for Barnase, let the stability at 300 K (i.e. below T_{F}) be $\Delta F = 20\ k_{\mathrm{B}}T$ (i.e. only $\sim 0{\cdot}2\ k_{\mathrm{B}}T$ per residue on average). Then letting $\sigma \cong 1\ k_{\mathrm{B}}T$ and $N \cong 100$, the bulk coefficient $f = 0{\cdot}3\ k_{\mathrm{B}}T$ from Eq. (3.17). The average amount of native structure at this temperature is then given by

$$\bar{N}_{\mathrm{F}} = \frac{\int_0^N \mathrm{d}N_{\mathrm{F}} N_{\mathrm{F}} \mathrm{e}^{-\Delta F(N_{\mathrm{F}})/k_{\mathrm{B}}T}}{\int_0^N \mathrm{d}N_{\mathrm{F}} \mathrm{e}^{-\Delta F(N_{\mathrm{F}})/k_{\mathrm{B}}T}} = \frac{\int_0^N \mathrm{d}N_{\mathrm{F}} N_{\mathrm{F}} \mathrm{e}^{0\cdot 3N_{\mathrm{F}} - N_{\mathrm{F}}^{\frac{1}{2}}}}{\int_0^N \mathrm{d}N_{\mathrm{F}} \mathrm{e}^{0\cdot 3N_{F} - N_{\mathrm{F}}^{\frac{1}{2}}}} \approx 96, \tag{3.22}$$

which indicates a well-defined native structure.

Applying the phenomenological nucleation theory to folding is instructive, but may seem to sweep too many details of the folding mechanism under the rug. Such a description must also be incomplete: the resulting free-energy profile is asymmetric in shape, but must apply to both nucleation of the folded state out of the unfolded state and vice versa, however there are no coefficients that could make the profiles identical. To see this, the unfolding process may be written as

$$F(N_{\mathrm{F}},\ T) = F_{\mathrm{F}} - f'N(1-n_{\mathrm{F}}) + \sigma' N^z (1-n_{\mathrm{F}})^z, \tag{3.23}$$

with $f' = +f_{\infty}(T) + \sigma' N^{-1+z}$ by the same reasoning that led to Eq. (3.17). Equating Eqs. (3.16) and (3.23) for all n_{F} leads directly to the trivial solution $\sigma = \sigma' = 0$ as the only solution, with a resulting linear free-energy profile. The capillarity theory is good for describing the folding nucleus when it occurs early (at low nativeness), or the unfolding nucleus when it occurs late (at high nativeness). The folding capillarity model gives a poor approximation for the configurations involved in late stage folding events, and vice versa for the unfolding capillarity theory. This is illustrated in Fig. 7.

It is worthwhile then to investigate the nature of entropic losses and energetic gains that occur during folding in more detail. Such a microscopic theory of folding is developed further in Part II within the framework of the correlated energy landscape.

Numerous experimental studies focus on specific heterogeneities in the folding mechanism. For example, sequence mutation assays determine which residues are formed or unformed in the transition state (Fersht *et al*. 1992). Various experimental probes (Ptitsyn, 1992, 1994; Dael *et al*. 1993; Radford & Dobson, 1995; Roder & Colon *et al*. 1997; Schulman, 1997; Heidary *et al*. 2000) along with Gō model simulations (Clementi *et al*. 2000a, b) can determine which elements are structured or unstructured in intermediates, and whether these intermediates are misfolded energetic traps or en-route entropic traps. Folding times have been observed to correlate with the mean sequence length between residues in contact in a protein's native structure (Plaxco *et al*. 1998), motivating studies of structural and sequence modifications to examine their effect on folding rate (Ladurner & Fersht, 1997; Viguera & Serrano, 1997). This provides motivation to develop theories which can provide a general framework to understand and predict features characteristic of the structural and energetic heterogeneities which are present in the folding mechanism (see Part II.7).

Protein stability may be strongly affected by some selective mutations, in that some relatively small local changes in energy may destabilize the native state. A complete understanding of the relative importance of each residue for stability and folding for a given protein has not been achieved yet, however, if a residue destabilizes the native state we can say that the free energy of another ensemble of configurations at $Q < 1$ becomes the new minimum. As mentioned above, proteins are only marginally stable at biological temperatures. The entropic bias of the unfolded state competes against the stability of the native structure. Typical stabilities for single domain proteins of sizes approximately 50–100 residues are about 10 $k_B T$ at 300 K. Mutating some amino acids from their wild-type identities may frustrate the native state by disrupting packing, disallowing hydrogen bonds normally present, or perhaps exposing non-polar surface. One can see that if a mutation effects say the neighboring $\sim$ 5–10 residues by about (0·5–1) $k_B T$, while simultaneously stabilizing $\sim$ 0–5 non-native interactions by (0·5–1) $k_B T$, the resulting free energy shift of $\sim$ 10 $k_B T$ can make an ensemble with $Q < 1$ the new free-energy minimum at 300 K, destabilizing the native structure. It is important to note that the more cooperative the stabilizing interactions are, the more a single mutation affects stability. If the misfolded state is the new energetic minimum, it is nearly impossible to achieve any native yield. If the misfolded state is an entropic minimum, lowering the temperature to attempt to re-obtain the native structure may still not be possible since at the low temperatures required to reduce the unfolded entropic stability, the interactions driving folding will have weakened due to their temperature dependence and so no longer favor the native state as strongly.

Residues which contribute largely to stability are likely to be more strongly involved than average in the transition state on energetic grounds. A larger contribution to stability may also lead to an enhanced conservation of their sequence identity or residue type (polar, hydrophobic, charged, etc.) (Overington *et al*. 1992; Shakhnovich *et al*. 1996; Michnick & Shakhnovich, 1998; Mélin *et al*. 1999; Mirny & Shakhnovich, 1999; Zou & Saven, 2000), since mutations of these residues would more strongly destabilize the native structure. Of course there are functional reasons for residues to be conserved as well, which may provide even stronger sequence inflexibility among structural homologs (Garcia *et al*. 2000; Plaxco *et al*. 2000a).

While the minimal frustration principle solves Levinthal's paradox, it raises a new question: how could a search through the vast space of possible sequences for the subset of foldable sequences ever be accomplished on an evolutionary timescale (Lau & Dill, 1990; Wolynes, 1994)?[12] In contrast to the Levinthal search problem for the ground state in random sequences, the best folding sequence (fastest, most stable, or even most functional) need not be found. Although there is evidence that the sequence landscape is rugged (Govindarajan & Goldstein, 1997a, b; Bornberg-Bauer & Chan, 1999; Tianna *et al*. 2000), in that regions of random heteropolymer sequences may largely surround regions of minimally frustrated sequences and there is no extended neutral network as in the case of RNA sequence landscapes (Schuster *et al*. 1994; Schuster, 1997), a local minimum solution is sufficient to solve the sequence search problem: annealing a Hamiltonian to a local minimum on a rugged fitness landscape is an NP easy problem (Onuchic *et al*. 1997), and so can occur on geological timescales. Biological protein sequences have not come to equilibrium and found their global fitness minimum (and will probably never come to equilibrium due to the dynamic environment between species), in that many sequences may be found which fold faster, are more stable, or are more optimized functionally than wild-type sequences (Hecht *et al*. 1985; Kim *et al*. 1998). Functionally active proteins need not find their global fitness minimum to satisfy the requirements of stability, function, and sufficiently rapid folding. Issues of protein design have received much attention in the literature, and are treated in detail elsewhere (see e.g. Lau & Dill, 1990; Shakhnovich & Gutin, 1990; Goldstein *et al*. 1992; LaBean & Kauffman, 1993; Davidson & Sauer, 1994; Shakhnovich, 1994; Pande *et al*. 1994b; Desjarlais & Handel, 1995; Li *et al*. 1996; Wolynes, 1996; Deutsch & Kurosky, 1996; Cordes *et al*. 1996; Dahiyat & Mayo, 1996; Morrissey & Shakhnovich, 1996; Saven & Wolynes, 1997; Nelson *et al*. 1997; Nelson & Onuchic, 1998; Iba *et al*. 1998; Vendruscolo *et al*. 1999; Bornberg-Bauer & Chan, 1999; Mélin *et al*. 1999; Tianna *et al*. 2000; Zou & Saven, 2000; Buchler & Goldstein, 2000; Tokita *et al*. 2000).

4. Generic mechanisms in folding

A funneled folding mechanism is a well-defined physical solution to the Levinthal problem, however other mechanisms may operate in conjunction with funneling to affect and possibly accelerate the folding rate, as alluded to above. In condensation or crystal nucleation, the bias towards the low-energy, low-entropy state is local and non-specific. In protein folding, funneling induces a bias towards the low-energy, low-entropy native state which is non-local as well as local, and specific to the native structure. It is worthwhile giving some examples of generic ordering processes which may operate in parallel with funneling, and which can dramatically reduce the number of degrees of freedom in the system. These transitions may also add new timescales to the folding problem, which are separately observable when they are shorter than the folding timescale (Ballew *et al*. 1996; Kuwajima *et al*. 1996; Hamada *et al*. 1996; Eaton *et al*. 1998; Munoz *et al*. 1998; Houry *et al*. 1998; Metzler *et al*. 1998; Baldwin

[12] There are about $\sim 20^{100} \approx 10^{130}$ possible sequences for a protein of length 100. Thus only an extremely small fraction of sequence space has ever been sampled, e.g. the number of protons in the universe (upper bound to the number) times the age of the universe in seconds (approximately upper bound to the sampling time) is still only $10^{80} \times 10^{17} = 10^{97}$ or 10^{-33} of the full sequence space.

& Rose, 1999; Hummer *et al.* 2000, 2001; Sagnella *et al.* 2000; Fernández *et al.* 2000; Yao *et al.* 2001; Kuwata *et al.* 2001).

4.1 Collapse, generic and specific

Generic collapse due to net hydrophobicity in the protein increases the packing density, and results in an extensive amount of entropy reduction. Nucleation of this state from the coil is analogous to condensation from the gas phase, although the barriers for each process may be quite different (Lifshitz *et al.* 1978). A polymer chain having ν orientations per chain link and packing density η $(0 < \eta < 1)$ has conformational entropy (Flory, 1949, 1953; Sanchez, 1979; Bryngelson & Wolynes, 1990)

$$S(\eta) \approx N\left[\ln\frac{\nu}{\mathrm{e}} - \left(\frac{1-\eta}{\eta}\right)\ln(1-\eta)\right], \tag{4.1}$$

which is a monotonically decreasing function of the packing density η. When $\eta = 0$, $S(0) = N \ln \nu$, and when $\eta = 1$, $S(1) = N \ln (\nu/\mathrm{e})$. In going from $\eta \approx 0$ to $\eta \approx 1$, the polymer loses about Nk_{B} of entropy, and the Levinthal time is decreased by a factor of about e^{-N}. Collapse will occur at a temperature T_{c} of roughly $z\bar{\epsilon}/k_{\mathrm{B}}$, where $\bar{\epsilon}$ is the average free energetic gain per residue–residue (hydrophobic + hydrogen bond) interaction, and z is the average number of interactions per residue. This state, with a conformationally fluid backbone and reminiscent of a polymer below its θ-point, is thought to represent the disordered parts of the protein in the equilibrium molten globule phase (Griko *et al.* 1988; Ptitsyn, 1995; Onuchic *et al.* 1995; Duan & Kollman, 1998); many observations of molten globule intermediates may be simply generically collapsed states with minimal tertiary order (Onuchic, 1997).

On the other hand, algorithms which apply the constraints measured from the NMR spectra to every member of the ensemble of the intermediate infer in effect an averaged structure, which often reproduces the native structure. Any general bias towards a specific structure will give this result, even if the ensemble consists of numerous, partially structured configurations which are native-like in varied spatial regions (Onuchic, 1997). This indicates that funneling mechanisms may be in effect early in the folding process, in that signals biasing native structure are felt in the unfolded ensemble (Onuchic *et al.* 1995); if the folding mechanism were exclusively non-specific at this stage, no partial native structure would be observed.

Observations of non-native contacts in some of these intermediates support the notion that the tertiary structure of the intermediate is not well-defined (Hamada *et al.* 1996; Kuwajima *et al.* 1996; Balbach *et al.* 1997; Forge *et al.* 2000; Fernández *et al.* 2000). It will be interesting to determine whether such misfolded intermediates, as occur in β-lactalbumin, are high entropy traps with fluctuating non-native structure, or low entropy traps, with relatively fixed non-native structure. A funneled folding mechanism would tend to give rise to higher entropy traps, unless barriers were large.

4.2 Helix formation

Another mechanism of generic entropy reduction is the formation of transient helical structure in the unfolded state. Various studies have observed secondary structure in the unfolded state (Miranker *et al.* 1991; Radford *et al.* 1992; Flanagan *et al.* 1993; Lin *et al.* 1994;

Gladwin & Evans, 1996; Hamada *et al.* 1996; Kuwajima *et al.* 1996; Balbach *et al.* 1997; Schulman *et al.* 1997; Luisi *et al.* 1999; Fernández *et al.* 2000; Forge *et al.* 2000). Whether the helicity is native or non-native, it is likely to be flickering, transient structure in the high entropy unfolded state, rather than fixed and rigid. In any event partial secondary structure formation in a protein reduces the conformational entropy by renormalizing the statistical segments on the backbone chain undergoing configurational diffusion, effectively reducing their number (Luthey-Schulten *et al.* 1995; Onuchic *et al.* 1995; Saven & Wolynes, 1996). The chain entropy has a roughly linear dependence on the fraction of chain involved in helices f_{H} (Onuchic *et al.* 1995), so that one can consider a partially helical chain to have effective length of approximately $N(1-f_{\mathrm{H}})$:

$$S_{\mathrm{tot}} = N_{\mathrm{eff}}\ s_0 \approx N(1-f_{\mathrm{H}})s_0. \tag{4.2}$$

Transient helix formation will occur generically when the temperature for the helix-coil transition ($T_{\text{H-C}}$) is comparable or greater than the T_{F} for the system under study. It is in principle also possible to have transient initiation of β-sheet structure in the form of short, flickering zippers. Estimates for the amount of helical secondary structure in the unfolded state are about $f_{\mathrm{H}} \approx \frac{1}{2}$ (Onuchic *et al.* 1995), reducing the entropy by about $Ns_0/2 \sim Nk_{\mathrm{B}}/2$, and the Levinthal time by a factor of about $\mathrm{e}^{-N/2}$ or about e^{-50} for a 100-residue protein, perhaps more for flexible chains. Since the energy scales involved in the helix-coil transition and folding are not widely different $T_{\text{H-C}} \sim T_{\mathrm{F}}$, both mechanisms may occur concurrently. This is in contrast to the role of secondary *versus* tertiary formation in RNA folding, where the transition involving secondary structure is driven by hydrogen pair bonding interactions each of order 10 $k_{\mathrm{B}}T$ and occurs first, removing a large amount of conformational entropy (Sankoff *et al.* 1978; Gesteland & Atkins, 1993).

Removing the extra degrees of freedom of the full energy function renormalizes the energy and entropy scales on the energy landscape. Universal features surviving this rescaling of thermodynamic parameters are reminiscent of the well-known law of corresponding states in conventional phase transitions (Goldenfeld, 1992), e.g. although the energy and entropy scales may be very different for the boiling of say water and xenon, after rescaling the temperature and density by their critical values, the phase diagram relevant to boiling (the equilibrium regions and coexistence curve) can be superimposed. In the context of the helix-coil transition for a (finite-sized) protein, the corresponding renormalization is thought to make a system of length $N \approx$ 50–100 amino acids amenable to analysis by simulating coarse-grained models of about half the length, with $N_{\mathrm{eff}} \approx$ 25–50 (Onuchic *et al.* 1995).

Caution and some empirical knowledge must be used in generalizing the law of corresponding states to properties beyond the equilibrium phase diagram. For example it is well known that the heat capacity peaks of model proteins are all significantly broader than those of laboratory proteins, which could directly follow from an under-representation of the amount of entropy in the system.

The law of corresponding states applies to the structure of the phase diagram rather than the kinetics of the transition. Just as water and xenon have different nucleation barriers to boil or condense, so also do simulations and real proteins have different absolute folding barriers. However, folding simulations of minimalist models to a given native structure have been shown to capture the general trends in folding mechanism for a variety of proteins (Ueda *et al.* 1975; Levitt & Warshel, 1975; Leopold *et al.* 1992; Fiebig & Dill, 1993; Socci & Onuchic, 1994; Chan & Dill, 1994; Abkevich *et al.* 1994; S̃ali *et al.* 1994; Pande *et al.* 1994b; Onuchic *et al.* 1995; Shakhnovich *et al.* 1996; Socci *et al.* 1996; Shoemaker *et al.* 1997, 1999; Lazaridis & Karplus, 1997; Nymeyer *et al.* 1998; Shea *et al.* 1998, 1999; Klimov & Thirumalai, 1998a, b; Du *et al.* 1999; Galzitskaya & Finkelstein, 1999; Alm & Baker, 1999; Munoz & Eaton, 1999; Nymeyer *et al.* 2000; Clementi *et al.* 2000a, b), so, for example a full atomistic description may not be necessary to locate preferred nucleation sites, or determine whether folding is two-state or proceeds through intermediates. This is a kind of extension of corresponding states: the characterizable influence of native structure on folding mechanism for real-proteins, as well a wide class of coarse-grained to highly resolved simulation models. An extension of the law of corresponding states to folding thermodynamics may be applied by modeling proteins having a similar surface to volume ratio as real proteins, since a dominant force stabilizing the native structure is the burial of hydrophobic surface area. This approach has typically been carried out by simulating coarse-grained models containing hydrophobic and hydrophilic monomers in reduced dimension (Dill *et al.* 1995), where the enumeration of conformational states is less computationally demanding. Caution must be taken however in generalizing the analysis of the 2D lattice models to the kinetics of real 3D proteins.

4.3 Nematic ordering

Helices present in an unfolded protein tend to align generically, similar to nematic or cholesteric liquid crystal order (Onsager, 1949; Flory, 1956; de Gennes, 1975; de Gennes & Pincus, 1977). This assists folding since most helices in the folded state tend to be aligned. Aligned helices gain steric entropy relative to non-aligned helices, in that their excluded volume is reduced: the excluded volume of two randomly oriented helices of long axis L and diameter D ($D \gg L$) roughly has the shape of a parallelepiped of volume $\sim L^2D$, while aligned helices have the excluded volume of a cylinder of volume $\sim LD^2$. Since the ratio of excluded volumes $D/L \gg 1$, a reasonable first approximation is to modify Eq. (4.1) so that aligned helical residues suffer essentially no steric entropy loss upon collapse of the polymer to density η (Luthey-Schulten *et al.* 1995; Saven & Wolynes, 1996). Then the total number of residues N is reduced by the number of aligned helical residues $f_{\mathrm{LC}}N_{\mathrm{H}}$, where f_{LC} is an order parameter measuring liquid crystal ordering [$f_{\mathrm{LC}} = 0(1)$ in the isotropic (nematic) state]. The steric entropy loss $\Delta S_{\mathrm{steric}} = S(\eta) - S(\eta = 0)$ becomes

$$\Delta S_{\mathrm{steric}} = -[N - f_{\mathrm{LC}}N_{\mathrm{H}}]\left[\left(\frac{1-\eta}{\eta}\right)\ln(1-\eta)+1\right]. \tag{4.3}$$

Aligned helices also lose orientational entropy ΔS_{rot}, of an amount about the log of the rotational partition function[13]

$$\Delta S_{\mathrm{rot}} \approx -k_{\mathrm{B}} \ln\left(\frac{2k_{\mathrm{B}}T}{\hbar^2/I}\right) \approx -(9-11)k_{\mathrm{B}} \tag{4.4}$$

per helix aligned with a given one, so the total rotational entropy loss is

$$(N_{\mathrm{helices}} - 1)\Delta S_{\mathrm{rot}}. \tag{4.5}$$

[13] As in poly-atomic gases, the moments of inertia I of even the smallest helices are sufficiently large that the energy level spectrum is approximately continuous: $\hbar^2/I \sim (10^{-4} - 10^{-5})k_{\mathrm{B}}T_{\mathrm{room}}$ (helix lengths of $\sim$ 5–10 Å and masses $\sim$ 300–1000 Da were used for this estimate).

A somewhat smaller value of the rotational partition function was used in Luthey-Schulten *et al.* (1995) and Saven & Wolynes (1996), of approximately $\frac{1}{4}$ of the entropy loss in Eq. (4.4).[14] The total entropy $S_0 + \Delta S_{\text{steric}} + (N_{\text{helices}} - 1)\Delta S_{\text{rot}}$ results in the isotropic distribution having higher entropy than the aligned. However, by Eqs. (4.3) and (4.5) an aligned helix ($f_{\text{LC}} = 1$) may grow while reducing the steric entropy loss, and without suffering any more orientational entropy loss, since for aligned helices steric terms do not contribute and the rotational entropy loss has already been paid for. Moreover, as an aligned helix grows it gains an extra energy in hydrogen bonds proportional to its length. Thus there is a first-order transition where as the temperature is lowered, the order parameter f_{LC} jumps from 0 to 1, and concurrently there is a rapid growth of helices coupled to the nematic transition. There is a discontinuity in the entropy corresponding to the latent heat of the transition of about $(15\text{–}20)k_{\text{B}}$ for a system of size $N \approx 100$, so the Levinthal time is further reduced by a factor of about $e^{N/5}$ or about e^{-20} for a 100-residue helical protein. The analogous generic transition relevant to the stacking of β-sheets tends to be present only in aggregates, and so is not particularly as important (but may be important in studying misfolding), at least for smaller proteins.

4.4 Microphase separation

The protein sequence is composed of both hydrophilic and hydrophobic amino acids, which phase separate in water, analogous to micellar formation (Tanford, 1980). This process is not independent of folding – in smaller proteins it may remove enough entropy to leave only a small ensemble of nearly native states. In larger proteins, the inside and outside behave as polymer melts with constraints on the interface, and there is a significant amount of entropy left. Here we outline the basic physics behind the transition in the context of proteins. We take a heuristic approach, following an analysis similar to Dill's (Alonso *et al.* 1991; Dill & Stigter, 1995). A more thorough analyses of the transition can be found in these references, and in the literature (see e.g. Leibler, 1980; Leibler *et al.* 1983; Fredrickson *et al.* 1992; Sfatos *et al.* 1993, 1994, 1995; Garel *et al.* 1994). The simplest global order parameter to characterize micro-phase separation within a single protein is

$$\sigma = \frac{N_{\text{H}}^{\text{in}} - N_{\text{H}}^{\text{out}}}{N_{\text{H}}} = \frac{N_{\text{P}}^{\text{out}} - N_{\text{P}}^{\text{in}}}{N_{\text{P}}}, \tag{4.6}$$

where N_{H} and N_{P} are the numbers of hydrophobic (H) and hydrophilic (P) residues respectively ($N_{\text{H}} + N_{\text{P}} = N$), N_{H}^{in} and $N_{\text{H}}^{\text{out}}$ are the numbers of hydrophobic buried and exposed residues (likewise for N_{P}^{in} and $N_{\text{P}}^{\text{out}}$), and for simplicity we have let $N_{\text{H}} = N_{\text{P}}$, so $N_{\text{H}}^{\text{in}} = N_{\text{P}}^{\text{out}}$.[15] The order parameter can have values $-1 \leqslant \sigma \leqslant 1$; $\sigma = 0$ is the uniform state, around which we will obtain a free energy expansion. Microphase separation is driven by free energy gains to bury Hs and expose Ps, and is opposed by the loss in entropy to separate the

[14] The estimates in Luthey-Schulten *et al.* (1995) and Saven & Wolynes (1996) should be reasonably accurate quantitatively, because (1) real helices need not lose all their rotational phase space upon alignment, and (2) the entropy reduction actually applies to one less than the total number of helices, since there is still an arbitrary direction for alignment. Because there are only $\approx$ 3–4 helices in the models this factor also significantly reduces the entropy loss.

[15] This is a fairly good approximation in that the ratio of H to P residues is nearly independent of protein size. The fraction of non-polar, non-charged amino acids in typical proteins remains constant at about $N_{\text{H}}/N \approx 2/3$ for proteins with chain lengths between $N = 50$ and $N = 400$.

system into two un-mixed states. For a polymer system there is also an 'elastic entropy' loss coupled to the phase separation that opposes the transition. The mixing entropy reduction $\Delta S_{\text{mix}}(\sigma)$ for either the Hs or Ps is

$$\left.\begin{aligned}\Delta S_{\text{mix}}(\sigma) &= \ln\begin{pmatrix} N_{\text{H}} \\ N_{\text{H}}^{\text{in}} \end{pmatrix} - N \ln 2 \\ &= -N\left(\frac{1+\sigma}{2}\ln(1+\sigma)+\frac{1-\sigma}{2}\ln(1-\sigma)\right).\end{aligned}\right\} \tag{4.7}$$

Note the mixing entropy is an even function of the order parameter: $\Delta S_{\text{mix}}(-\sigma) = \Delta S_{\text{mix}}(\sigma)$. This must be true also for the elastic entropy term as well, and a Landau expansion must contain only even powers of σ. We approximate the elastic entropy loss to lowest order, as a harmonic spring:

$$\Delta S_{\text{P}}(\sigma) \approx Nb\sigma^2. \tag{4.8}$$

This approximation is very crude since in the phase-separated state $\sigma \approx 1$ and thus is not small. In our approximation then $\Delta S_{\text{P}}(1) = Nb$ is the polymer entropy loss upon full phase separation, which we estimate below.

The solvent-averaged potential $U(\sigma)$ driving the transition depends specifically on burying Hs and exposing Ps, and so is an odd function of σ. Using $U(\sigma) = -U(-\sigma)$ and expanding around $\sigma = 0$ gives, to third order:

$$U(\sigma) \approx -N(h\sigma + \epsilon\sigma^3) \tag{4.9}$$

with $h > 0$ and $\epsilon > 0$. Here h is a one-body energy term, and would play the role of an external field in the analogous mean field Ising spin model. A two-body potential could be accommodated by a term $\sim$ sign (σ) σ^2, but is not necessary for the derivation. The term $\propto \sigma^3$ is a three-body term which imparts cooperativity in the model. Higher order terms may be considered straightforwardly.

We estimate the parameter b in Eq. (4.8) from polymer physics as follows. As shown in the inset of Fig. 8, if there are N_i interfaces between runs of H and P residues ($N_i = 12$ in the figure), the polymer entropy loss (neglecting end effects) is the log of the probability for a chain segment of sequence length l_i to propagate from one place on the interface surface to any other place on the surface without crossing it first, summed over the number of interfaces:

$$\Delta S_{\text{P}}(1) = Nb \approx \sum_{i=2}^{N_i} \ln \oint d\mathrm{r}_i G_{l_i}(\mathrm{r}_i|\mathrm{r}_{i-1}). \tag{4.10}$$

A quick inspection of PDB protein sequences reveals that for $N \lesssim 200$ the number of interfaces $N_i \cong N/2$,[16] so the average sequence length $\bar{l}$ is only $\cong 2$. We approximate each l_i by its average $\bar{l} \cong 2$, thus runs of Hs and Ps are fairly short. Further approximating the integral by its largest value, i.e. where the chain returns to the origin, we obtain[17]

$$Nb \approx N_i \ln G_{\bar{l}}(0|0)$$

or

$$b \approx \tfrac{1}{2}\ln\left(\frac{3}{2\pi\bar{l}}\right)^{\frac{3}{2}} \approx \tfrac{3}{4}\ln\frac{3}{4\pi}. \tag{4.11}$$

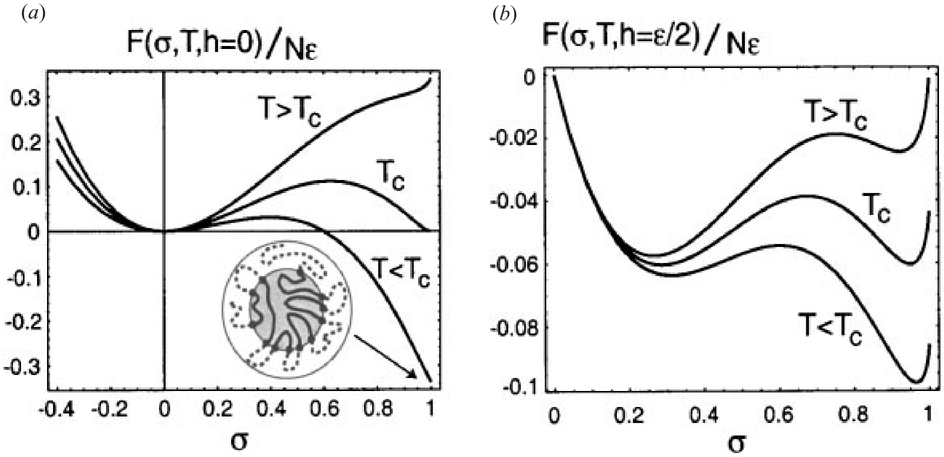

Fig. 8. Free-energy profiles *versus* microphase separation order parameter σ, from Eq. (4.12). (*a*) 'Zero field' case: the one body term in the potential is set to 0. The critical temperature $T_C \approx 0{\cdot}6\epsilon$. Then the free-energy minima are at $\sigma = 0$ above T_C and $\sigma = 1$ below T_C. The barrier is several $k_B T_C$ at the critical temperature. (Inset) Schematic of a microphase separated polymer. Hydrophobic residues (solid polymer) are buried, hydrophilic (dashed) exposed. The residues between hydrophobic and hydrophilic trains, denoted by black circles in the figure, are free to wander on the interface. (*b*) In the presence of a one-body term, the free-energy minima are shifted – particularly the mixed state is no longer at $\sigma = 0$. This is analogous to a ferromagnet in a weak field. The presence of a one-body term which favors partial separation raises the critical temperature (here for $h = \epsilon/2$, $T_C \approx 0{\cdot}86\epsilon$), and reduces the barrier height – the barrier for $h = \epsilon/2$ is only about 1/7 of the $h = 0$ case, i.e. just a couple $k_B T_C$.

The free-energy $F_{\mu\phi}(\sigma)$, from Eqs. (4.7), (4.8) and (4.9) is

$$\frac{F_{\mu\phi}(\sigma)}{N} = -h\sigma - \epsilon\sigma^3 + Tb\sigma^2 + \frac{T}{2}[(1+\sigma)\ \ln(1+\sigma) + (1-\sigma)\ \ln(1-\sigma)] \tag{4.12}$$

and is shown in Fig. 8. The equilibrium values of the order parameter are obtained by minimizing $F(\sigma)$ with respect to σ:

$$\sigma = \tanh\left(\frac{h + 3\epsilon\sigma^2}{T} + 2b\sigma\right). \tag{4.13}$$

The transition is first order, with a barrier height at the critical temperature of about a few $k_B T_C$.

From Eq. (4.12), the entropy loss upon microphase separation is

$$\Delta S_{\mu\phi} = S_{\mu\phi}(\sigma = 1) - S_{\mu\phi}(\sigma = 0) = -N(b + \ln 2)k_B, \tag{4.14}$$

or about 1·7 Nk_B.

As mentioned above, microphase separation is not decoupled from folding. In fact, folding of 2-letter code models with the appropriate energy function may be thought of as microphase separation, with the pattern of H and P residues along the sequence specifying the ground state structure (Fiebig & Dill, 1993; Sfatos *et al.* 1993; Camacho & Thirumalai, 1993; Pande *et al.* 1994a). Local concentrations of non-polar residues along the sequence have been

[16] This justifies *a posteriori* the assumption in Eq. (4.8) that the elastic entropy loss scales extensively. For larger ranges of protein sizes up to about $N = 500$, the number of interfaces begins to scale more strongly with the surface area of the protein $\sim N^{\frac{2}{3}}$.

[17] This entropy is calculated for the whole system $N_H + N_P$; technically we may wish to divide by two here since we are just dealing with the separation of a single phase, H or P.

suggested to be nucleation sites for folding (Matheson & Scheraga, 1978). On the other hand, it is also known that it is difficult to generate a funnel topography with only a 2-letter code energy function (Honeycutt & Thirumalai, 1992; Wolynes, 1997a; Nymeyer *et al.* 1998, 2000). This difficulty is partially reflected in real amino-acid sequences, even though a 2-letter amino-acid code may have more complexity than a 2-letter computational model because of solvent and many-body effects. The Sauer group has found sequences composed of one non-polar and two polar amino acids which cooperatively fold to helical structures (Davidson *et al.* 1995), however the Baker group did not find a sequence using a 3-letter amino-acid alphabet that was capable of folding to the src SH3 domain native structure, but needed at least a 5-letter code (Riddle *et al.* 1997). Because of the numerous low-energy states on a 2-letter code landscape, the temperatures required to give the actual ground state a thermodynamic weight or occupation probability comparable to 1 are sufficiently low, so that either (1) the solvent-averaged energy functions would change upon cooling and no longer stabilize the native state, or (2) even if the energy functions did not change, folding would be exceedingly slow. Moreover, the ground states for 2-letter codes are often degenerate – the probability a ground state in a random copolymer is k-fold degenerate decays slowly as $\sim k^{-1}$ (Gutin & Shakhnovich, 1993). Microphase separation, particularly with a term penalizing exposed hydrophobic residues as in Eq. (4.12), reduces this degeneracy. For short chains of a given stiffness it may remove the degeneracy completely as mentioned above. Hecht and colleagues have found cooperatively folding sequences to an α-helical structure based on a binary patterning of polar and non-polar residues, but of various residue identities (Hecht *et al.* 1985). However this does not preclude the possibility of folding true two-amino-acid letter sequences to an α-helical structure. More letters may be needed for β-structures to remove degeneracies corresponding to sliding motions between strands, and removing these degeneracies in an energy function is one of the main difficulties in β-protein structure prediction. A true two-amino-acid code sequence capable of folding and function has yet to be observed but may exist most likely for an α-helical structure. The landscape for this 2-letter-code sequence will likely be atypically rugged.

The upshot of the above discussions is that phase transition mechanisms which are generic, insofar as they are independent of energetic bias to a specific structure, can reduce the entropy sufficiently that the remaining phase space may be randomly searched on timescales much less than the Levinthal time [Eq. (2.3)]. The glassy search time among the low-energy states after all the above generic phase transitions have occurred in a hypothetical protein is

$$\left.\begin{aligned}\tau_{\text{gen}} &\sim \tau_0 \exp\,(Ns_0 - \Delta S_{\text{collapse}} - \Delta S_{\text{helix}} - \Delta S_{\text{liq-xtal}} - \Delta S_{\mu\phi})\\ &\sim (10^{-12}\ \text{s}) \exp N \left(\frac{s_0}{2} - 3\right) \sim 1\ \text{year}.\end{aligned}\right\} \tag{4.15}$$

While the numbers here are clearly approximate, the conclusion remains that generic mechanisms have clearly reduced the search problem. Specificity of course will act in conjunction with generic mechanisms in real proteins, and it may in fact be difficult to separate the two mechanisms in practice.

5. Signatures of a funneled energy landscape

On a funneled energy landscape, one structure has distinctly lower energy than all other dissimilar structures, as in Eq. (3.11). Several consequences of this have been discussed already, we will elaborate on a few more here.

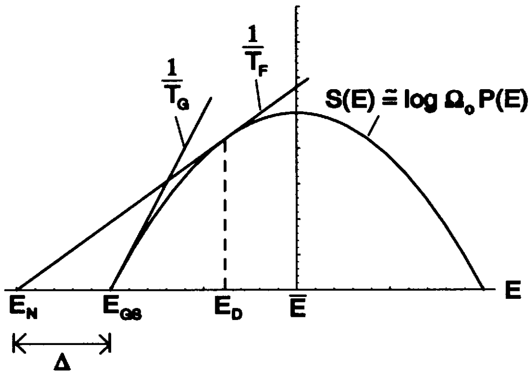

Fig. 9. Log density of states in the REM. The slope of the curve at the $Q = 0$ ground-state energy (E_{GS}) gives the reciprocal of the glass temperature (T_G) in the model. For designed sequences there is a minimally frustrated native state (E_N) with considerably lower energy Δ than the putative ground state of a RHP having the same composition. This state will be in thermodynamic equilibrium with the same Boltzmann weight as a manifold of unfolded states with higher average energy (E_D) at a temperature (T_F) above the REM glass temperature (T_G). (Figure adapted from Shakhnovich & Gutin, 1993.)

First, the low energy of the native structure means that it is occupied with equal weight as a collection consisting of an exponential number ($\sim e^{\alpha N}$) of higher energy dissimilar states at a temperature T_F above the temperature where the system loses its configurational entropy (see Fig. 9). The partition function of partially native structures contains the sum over states at a given partial degree of nativeness, which we can take to be the fraction of native interactions Q ($0 \leqslant Q \leqslant 1$). Each state has a native core with energy E_C and a non-native decoration dressing the core with energy E_{NN}. The partition function then becomes:

$$Z(Q, T) = \sum_{\text{states at } Q} e^{-(E_C + E_{NN})/T}. \tag{5.1}$$

The native core has essentially no configurational entropy, while the non-native dressing has substantial entropy (see e.g. inset of Fig. 10). Again since T_F is above T_G by the evolutionary shaping of the landscape, the entropy at T_F is large, and the entropy at Q is reduced mainly by the parts of the protein pinned down by native interactions, rather than by deep, non-native traps. As a consequence of the large entropy, the non-native part of the protein occupies energies in the continuous part of the spectrum of its density of states, $n(E)$. Then the free-energy of the non-native part of the protein is independent of the specific sequence, and only depends upon gross overall features of the interactions such as their mean and variance (Derrida, 1981; Gross & Mézard, 1984). Again this is because many states are sampled at the typical energies of the non-native dressing, so relative fluctuations in the number of states with energy E for different sequences from the average density of states (for the 'average' sequence) $\bar{n}(E)$, die away as the system size increases:

$$\frac{\sqrt{\langle (n(E) - \bar{n}(E))^2 \rangle}}{\bar{n}(E)} \approx \frac{1}{\sqrt{\bar{n}(E)}}. \tag{5.2}$$

Since $\bar{n}(E)$ scales exponentially in system size, fluctuations sequence to sequence may be neglected, and the density of states for the non-native dressing is to very good approximation the sequence-averaged density of states: $n(E) \approx \bar{n}(E)$. The free energy of the native core depends intrinsically on the specific sequence.

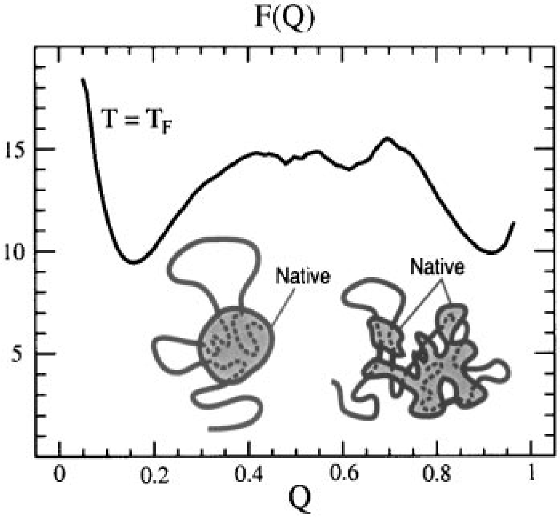

Fig. 10. Free energy *versus* fraction of native contacts Q, obtained from off-lattice simulations using a uniform Gō potential to the native structure of CheY (data from Clementi *et al*. 2000b). Fluctuations in the free-energy profile here arise from the effects of native topology on the entropy and energy at partial degrees of nativeness. (Inset) Schematic illustration of partially native configurations consisting of native cores and the surrounding polymer halo. The core may be globular or ramified.

For near native structures with $Q \approx 1$, the halo dressing the core is small, and so has relatively few states. In this regime it is possible that sequence-dependent (non-self-averaging) effects will be important. This may be particularly true for conformational motions related to the function of the protein.

Generally however the partition function for the non-native dressing, which is the Laplace-transformed density of states[18]

$$\bar{Z}(T) = \int_{-\infty}^{\infty} \mathrm{d}E \; \bar{n}(E)\,\mathrm{e}^{-E/T} \tag{5.3}$$

is independent of the specific sequence and only depends on the gross features of the sequence composition. It is then referred to as the sequence-annealed partition function.

To illustrate, let the non-native interactions be distributed with width b and mean c, and let there be M total native interactions in the native structure. Then, using the central limit theorem for the distribution of energies, the density of states for the polymer surrounding a given native core is a Gaussian:

$$\bar{n}(E) \approx \mathrm{e}^{S_\mathrm{C}} \frac{1}{\sqrt{2\pi\Delta^2}}\,\mathrm{e}^{-(E-\bar{E})^2/2\Delta^2}, \tag{5.4}$$

[18] Here the integration may be taken from $-\infty$ to ∞ even though the system is finite, because those energies in the tails of the bound spectrum do not contribute significantly to the partition function at temperatures where folding occurs.

where S_{C} is the configurational entropy of the non-native polymer dressing the native core, $\bar{E} = M(1-Q)c$, and $\Delta^2 = M(1-Q)b^2$. We are neglecting for now the coupling of folding, or Q, with density, which we reserve for Part II.5. Then from Eq. (5.1) the annealed partition function for the core,

$$\left.\begin{aligned}\bar{Z}_{\mathrm{C}}(Q,\, T) &= \mathrm{e}^{-E_{\mathrm{C}}/T}\int_{-\infty}^{\infty} \mathrm{d}E\,\bar{n}\,(E)\mathrm{e}^{-E/T} \\ &= \mathrm{e}^{-E_{\mathrm{C}}/T+S_{\mathrm{C}}}\exp\left(\frac{\Delta^2}{2T^2}-\frac{\bar{E}}{T}\right),\end{aligned}\right\} \tag{5.5}$$

only depends on the energy and entropy associated with the particular core, and the mean and variance of the sequence composition. The total partition function at Q is the sum of Eq. (5.5) over cores:

$$\bar{Z}(Q,\, T) = \sum_{\text{cores}} \bar{Z}_{\mathrm{C}}(Q,\, T) \cong \exp\left(\frac{\Delta^2}{2T^2}-\frac{\bar{E}}{T}\right)\sum_{\text{cores}} \mathrm{e}^{-F_{\mathrm{C}}/T}. \tag{5.6}$$

Now let us rewrite the core energy as $E_{\mathrm{C}} = QMc + \mathrm{E}'_{\mathrm{C}}$, where c is again the average energy of a contact (native or non-native) and E'_{C} is now the *extra* energetic stability gained by the native core; at Q all native cores are taken to have MQ contacts. The free-energy $F(Q,\, T)$ is then given by $-T \ln \bar{Z}(Q,\, T)$:

$$\frac{F(Q,\, T)}{M} = c - \frac{b^2}{2T}(1-Q) - \frac{T}{M}\ln\sum_{\text{cores}} \mathrm{e}^{-E'_{\mathrm{C}}/T+S_{\mathrm{C}}} \tag{5.7}$$

which is the potential of mean force that the system reconfigures and folds upon.

Equation (5.7) is the resulting free-energy of a funneled landscape. Note there is no assumption that the number of distinct folding routes or cores is large, only that the non-native part of the protein may be treated as annealed (over sequences). The specific energetics and entropics of partially native structures enters into the free-energy profile. For example, Fig. 10 shows the free-energy profile near the transition temperature for an off-lattice C_α model folding to the native structure of CheY (Clementi *et al.* 2000b), with parameters $c = b = 0$ and $E'_{\mathrm{C}} = MQ\bar{\epsilon}$ for all cores (a *uniform* Gō model to the native structure of CheY). Fluctuations in the free-energy profile then result from largely from entropic preferences for certain cores over others during folding, and the fluctuations in entropy as more native constraints are added.

The residual free-energy barrier arises from the incomplete cancellation of entropy by energy as Q increases from 0 to 1. Since in general the non-native contribution to the free-energy is a linear function of Q, the barrier governing the folding rate, which must arise from nonlinearity in Q, is determined by properties of the energies and entropies of partially native structures, which are in turn determined by the native topology and distribution of native stability throughout the protein. Thus minimal frustration and the corresponding funneled landscape has as a consequence the intrinsic connection between native structural and energetic properties and the folding barrier. Recently several observations (Plaxco *et al.* 1998) and proposed models (Munoz & Eaton, 1999; Alm & Baker, 1999; Riddle *et al.* 1999; Clementi *et al.* 2000a, b; Kim *et al.* 2000; Plotkin & Onuchic, 2000) indicate this important connection. The dramatic dependence of folding times on properties of the native structure such as contact order (Plaxco *et al.* 1998, 2000b) illustrates the importance of backbone topology in determining the folding free-energy profile. This justifies to some extent the modeling of folding through coarse-grained polymer chain models.

The probability a particular native core is sampled at $\mathcal{Q}$ is

$$p_{\mathrm{C}} = \frac{\mathrm{e}^{-F_{\mathrm{C}}/T}}{\mathrm{e}^{-F(\mathcal{Q},\, T)/T}}, \tag{5.8}$$

which depends on the core's energetic stability and the entropy dressing the core.

The probability a particular native contact is occupied is related to how much the free-energy changes when that contact's energy ϵ_i is changed. From Eq. (5.7),

$$\frac{\partial F}{\partial \epsilon_i} = \frac{\sum\limits_{\text{cores}} \left(\dfrac{\partial E_C}{\partial \epsilon_i} \right) \mathrm{e}^{-F_{\mathrm{C}}/T}}{\sum\limits_{\text{cores}} \mathrm{e}^{-F_{\mathrm{C}}/T}}. \tag{5.9}$$

Because the energy of a core is the sum of its native contact energies, $E_{\mathrm{C}} = \sum_{j \in \text{core}} \epsilon_j$, then it is true that

$$\begin{aligned} \frac{\partial E_{\mathrm{C}}}{\partial \epsilon_i} &= \begin{cases} 1 \text{ if contact } i \text{ is made in the core} \\ 0 \text{ otherwise} \end{cases} \\ &\equiv \delta(i, \text{ core}). \end{aligned} \tag{5.10}$$

Thus the change in free energy with respect to the energy of contact i is the thermally averaged probability that contact i is made,

$$\frac{\partial F}{\partial \epsilon_i} = \langle \delta(i, \text{ core}) \rangle \equiv \mathcal{Q}_i \tag{5.11}$$

or equivalently the fraction of proteins in a macroscopic sample at $\mathcal{Q}$ which contain contact i.

The set $\{\mathcal{Q}_i(\mathcal{Q})\}$ for all $\mathcal{Q}$ defines the thermodynamic folding mechanism, and is straightforwardly determined in principle from the free-energy function, Eq. (5.7). The folding mechanism is then a function of the energies and entropies of partially native structures, which are fully determined by the native topology and distribution of native stability throughout the folded protein. So minimal frustration and the corresponding funneled landscape predicts that the folding mechanism is most strongly determined by these native topological and energetic properties.

For landscapes not well-funneled, i.e. where $(E_{\mathrm{N}} - E_{\mathrm{GS}})/E_{\mathrm{GS}} \ll 1$, the annealed approximation for the non-native polymer dressing the cores is poor: the density of states is in the low-energy discrete part of the spectrum, and escape from individual traps determined by the specific sequence are important for determining diffusion and folding rates. In this case, the temperature is not large compared to the non-native variance, and there is little entropy left for the non-native polymer dressing the native cores. The total entropy summed over cores of the partially native protein is small, and consists mostly of the entropy of placement of native cores.

From Eq. (5.5), the free energy for a particular core on a funneled landscape is

$$\bar{F}_{\mathrm{C}}(\mathcal{Q}, T) = -T \ln \bar{Z}_{\mathrm{C}}(\mathcal{Q}, T) = \bar{E} + E_{\mathrm{C}} - TS_{\mathrm{C}} - \frac{\Delta^2}{2T}. \tag{5.12}$$

So the condition for the annealed approximation to be valid is then that the entropy of the non-native dressing be greater than zero:

$$\bar{S}_{\mathrm{C}}(Q, T) = -\frac{\partial \bar{F}_{\mathrm{C}}}{\partial T} = S_{\mathrm{C}} - \frac{\Delta^2}{2T^2} > 0 \tag{5.13}$$

or

$$T > T_{\mathrm{G}}(Q) = \frac{\Delta(Q)}{\sqrt{2S_{\mathrm{C}}(Q)}}, \tag{5.14}$$

for at least one of the possible folding cores. The Q dependence of the non-native variance and polymer dressing entropy have been noted in Eq. (5.14).

If a particular core has no entropy ($\bar{S}_{\mathrm{C}}(Q, T) \approx 0$) but others do, the other cores will dominate the folding mechanism since they are much more likely to be occupied by Eq. (5.8). However if the entropy for the typical folding core $\bar{S}_{\mathrm{core}}(Q, T)$ is not > 0, then many microscopic barriers control the rate. These are integrated over in obtaining $F(Q)$, so the order parameter Q, specifying structural similarity to the native state is then not a good reaction coordinate for folding, in that folding rates have little or nothing to do with the free-energy barriers that appear in $F(Q)$.

But generally for a minimally frustrated system the annealed approximation is valid, and the folding rate is governed by $F(Q)$. Moreover, if no particular cores are strongly preferred, the kinetic proximity to the native state will correlate well with Q, so long as the topology of the native structure does not too severely restrict the connectedness of the conformations. That is, the ensemble of transition states will be well-described by the ensemble of conformations having the overlaps $Q^{\ddagger}$ of the free-energy barrier peak.

The ensemble of kinetic transition states may be described as the ensemble of states that have a probability of $\frac{1}{2}$ of folding before unfolding (Du *et al.* 1998). For large barriers, the transition states are concentrated close to the barrier peak, but for wide, flat barriers as typically occur in simulational models of folding, the transition states tend to be spread out. The degree to which the kinetic transition states are localized to the barrier peak is related to how uniformly the different partially native structures are occupied at $Q^{\ddagger}$. This is closely related to the degree of uniformity in experimental ϕ-values measurements (Fersht *et al.* 1992). This uniformity may be adjusted in principle by varying the strength of different native interactions.

In any given folding event, a protein molecule samples numerous cores from the full ensemble of possible cores. Near biological conditions, each of these growth processes has a region which is thermodynamically uphill where energetic gains are not quite compensating for entropic losses, and a region thermodynamically downhill where energetic gains win over entropic losses. The transition between these two regions occurs in general at different amounts of native structure for different native core locations, depending on the different amounts of core energy and halo entropy dressing the core. Again, the transition regions for different growth processes are further localized together [and to the barrier peak of $F(Q)$] when the cores tend to be more uniformly occupied.

So for a funneled landscape, the transition states correlate well with the barrier peak so long as the fluctuations in the occupation probabilities of the various cores are not too large. By similar reasoning which led to Eq. (5.14) we can obtain a criterion for the barrier peak position $Q^{\ddagger}$ to be a good indicator for the position of the transition state ensemble for a funneled landscape. Let the log number of distinct cores at Q be $S_{\mathrm{rout}}(Q)$, a geometrical quantity.[19] Let the free energy of cores be Gaussianly distributed with mean $\bar{F}_{\mathrm{C}}(Q) = \bar{E}_{\mathrm{C}} - T\bar{S}_{\mathrm{C}}$ and variance $\delta F_{\mathrm{C}}^2(Q) = \delta E_{\mathrm{C}}^2(Q) + T^2 \delta S_{\mathrm{C}}^2(Q)$. Then from the partition function at Q [Eq. (5.6)], the free energy is obtained as

$$\left.\begin{aligned} \bar{F}(Q,\, T) &= -T \ln \bar{Z}(Q, T) = \bar{E} - \frac{\Delta^2}{2T} - T \ln \left(\mathrm{e}^{S_{\mathrm{rout}}(Q)} \int_{-\infty}^{\infty} P(F_{\mathrm{C}}) \mathrm{e}^{-F_C/T} \right) \\ &= \bar{E} + \bar{E}_{\mathrm{C}} - T\bar{S}_{\mathrm{C}} - \frac{\Delta^2}{2T} - TS_{\mathrm{rout}} - \frac{\delta F_{\mathrm{C}}^2}{2T}. \end{aligned}\right\} \tag{5.15}$$

The condition for the probability of folding to correlate well with Q is that the total route entropy be large. Then many cores are sampled in folding, and the Boltzmann weights of specific cores are all very small. Taking the derivative of Eq. (5.15) gives the total entropy as a sum of two expressions:

$$\left.\begin{aligned} \bar{S}(Q,\, T) &= -\frac{\partial \bar{F}}{\partial T} = \bar{S}_{\mathrm{C}}(Q, T) + S_{\mathrm{R}}(Q, T) \\ &= \left(\bar{S}_{\mathrm{C}}(Q) - \frac{\Delta^2(Q)}{2T^2} \right) + \left(S_{\mathrm{rout}}(Q) - \frac{\delta F_{\mathrm{C}}^2}{2T^2} \right). \end{aligned}\right\} \tag{5.16}$$

The first term in parentheses is the average amount of entropy for each core that we saw earlier in Eq. (5.13). As long as the system is minimally frustrated, this entropy is greater than zero. The second term is the thermodynamic route entropy of the system. This term is at most the log number of distinct geometrical cores $S_{\mathrm{rout}}(Q)$ as determined by the native structure the system folds to, and is reduced by fluctuations in the free energies of the various cores.

Given the landscape is funneled and the mean core entropy is large by Eq. (5.14), the criterion then for the route entropy to be large, and for the barrier peak to well represent the transition state ensemble, is then

$$S_{\mathrm{rout}}(Q) - \frac{\delta E_{\mathrm{C}}^2(Q)}{2T^2} - \frac{\delta S_{\mathrm{C}}^2(Q)}{2} > 0 \tag{5.17}$$

or

$$T > T_{Q\text{-rxn}} = \frac{\delta E_{\mathrm{C}}}{\sqrt{2S_{\mathrm{rout}} - \delta S_{\mathrm{C}}^2}}. \tag{5.18}$$

If fluctuations in the core weights are large enough the denominator vanishes. Then one core dominates the folding, and the degree of formation of its particular native contacts is the appropriate reaction coordinate for folding. From Eq. (5.18), for sufficient entropic variance in potential folding nuclei, the folding mechanism is through a specific nucleus. Since the entropic variance is a function only of the native topology, some native structures may inherently have specific folding mechanisms, while others may have folding mechanisms

[19] We use the notation S_{rout} because this is a thermodynamic way of quantifying the number of distinct 'routes' to the native structure.

involving a diffuse nucleus. For diffuse nucleus proteins, Q is a good progress coordinate. For specific nucleus proteins, those contacts involved in the specific native core, $\{Q_i\}_{\text{core}}$, are the best progress coordinate. For smaller entropic variance $\delta S_{\text{C}}^2 < 2S_{\text{rout}}$, variance in the energetic stability may still induce specificity. A similar temperature scale for route-like folding is obtained in Part II.5.

Various degrees of specificity or diffusivity have been observed in simulations (Daggett & Levitt, 1993; Abkevich *et al.* 1994; Chan & Dill, 1994; Guo & Thirumalai, 1995, 1997; Boczko & Brooks, 1995; Socci *et al.* 1996; Onuchic *et al.* 1996; Shakhnovich, 1997; Sheinerman & Brooks, 1998a, b; Du *et al.* 1998; Klimov & Thirumalai, 1998b; Nymeyer *et al.* 2000; Li *et al.* 2000) and experiments (Fersht *et al.* 1992; Radford *et al.* 1992; Itzhaki *et al.* 1995; López-Hernández & Serrano, 1996; Viguera *et al.* 1996; Daggett *et al.* 1996; Oliveberg *et al.* 1998; Martinez *et al.* 1998; Grantcharova *et al.* 1998; Otzen & Fersht, 1998; Chiti *et al.* 1999; Munoz & Eaton, 1999; Riddle *et al.* 1999; Martinez & Serrano, 1999; Plaxco *et al.* 2000a; Fersht, 2000). The question of whether a folding mechanism proceeds through a specific nucleus or a diffuse nucleus may be addressed by judicious application of criteria such as Eq. (5.18).

It is possible for specific non-native ('off-pathway') traps to exist during folding in some proteins as well. These may be able to be discerned by simulating Gō models to the same native structure, to see if the intermediate is still there (on-pathway) or absent (off-pathway). Explicitly accounting for chain connectivity properties may become important for folding to some native structures as well, particularly for larger proteins.

6. Statistical Hamiltonians and self-averaging

The complexity of a fully atomistic Hamiltonian for a protein molecule is prohibitive to analysis or simulation. Moreover, for many theoretical questions it is not important to know all the parameters of the Hamiltonian for a specific system under study: often there are many irrelevant degrees of freedom in the problem which are not worth keeping track of. For example it would be largely a waste of effort to fully describe the effects of non-native interactions between say helix 1 and helix 2 of wild-type monomeric λ-repressor, one of the fastest folding two-state folders. Often it is the overall gross parameters of a statistical ensemble of such Hamiltonians that are of theoretical interest, e.g. the universal properties of the set of Hamiltonians determined by the set of sequences that give rapid two-state folding to the native structure of λ-repressor. In these cases a complex potential function can be replaced by a coarse-grained stochastic one having the same statistical characteristics. Applications of such methods have been used in other fields of physics, notably in the treatment of highly excited states of nuclei with random Hamiltonians (Wigner, 1951), extensively in the theory of spin glasses (Mézard *et al.* 1986) and neural networks (Hopfield, 1982; Amit *et al.* 1985), and in the conformational sub-states of already folded proteins (Austin *et al.* 1975; Frauenfelder *et al.* 1991; Stein, 1985). The usage of statistical Hamiltonians was established in the theory of protein folding by Bryngelson & Wolynes (1987, 1989).

Quantities calculated from such statistical Hamiltonians are representative of the ensemble to the extent that they are self-averaging. In ordinary statistical mechanics the relative fluctuations of a thermodynamic quantity, such as the energy, die away as $\sim N^{\frac{1}{2}}$. Likewise we expect that sample-to-sample (or sequence-to-sequence) fluctuations of thermodynamic quantities go to zero in the limit of an infinitely large system. However, proteins are fundamentally different from systems such as spin-glasses because their sequences are *a priori* finite in length, and have special properties which are lost if scrambled. In protein folding, self-averaging translates to those properties which are independent of the specific characteristics of a given ensemble of sequences or Hamiltonians, but which depend only on some gross, overall feature of that ensemble. This depends strongly on the property under investigation. For example, log collapse times tend to be self-averaging, i.e. sequence-independent, but only over sequences of some fixed overall composition (fixed hydrophobicity). But for this sequence subset log folding times are non-self-averaging. If we take a further subset of these sequences having the same stability gap, folding times of this sub-sub-set are comparable, but can still be quite disparate. If we specify further the sub-ensemble of sequences having the same native structure, or same overall native topology or contact-order, then their corresponding folding barriers are comparable and may be said to be self-averaging. As mentioned above, the size of the sequence pool to select from is huge: a 100-mer has a sequence space $\sim 10^{50}$ times the number of protons in the universe, so there may still be large ensembles of sequences with self-averaging characteristics.

Underlying the theoretical approach is the conviction that many of the determining factors of folding rate and mechanism are physical parameters of an ensemble of protein sequences which are self-averaging in the above sense. Examples of these are contact-order, stability gap, overall ruggedness or energetic variance in the unfolded state, and chain stiffness or average entropy per length. Many quantitative features of protein folding may be described in terms of a few relevant thermodynamic parameters which most strongly govern the structure of the free-energy landscape.

To capture the physics of a reconfigurable polymer chain, with a quenched sequence of residues and a set of interactions between them, the following Hamiltonian is typically used:

$$\mathscr{H}_{\mathrm{RHP}}(\{s_i\},\{r_i\}) = \sum_{i<j} \epsilon(s_i, s_j)\Delta(r_i - r_j), \tag{6.1}$$

where i counts the monomers or residues along the chain, and the double sum is over residue indices, $s_i \in \{1, \ldots, p\}$ is the species of monomer i along the chain so that $\{s_i\}$ represents the monomer sequence, p is the number of species (1 for a homopolymer, 2 for a co-polymer, ∞ for a continuous distribution of interaction energies), and r_i is the position of monomer i so that $\{r_i\}$ represents the polymer conformation. $\Delta(r)$ is a function that includes an excluded volume repulsion at short distances and dies away at long distances; in lattice models, $\Delta(a) = 1$, $\Delta(0) = \infty$, and $\Delta(r > a) = 0$, where a is the lattice spacing. Thus the energy of a particular polymer conformation is determined by the matrix of species-species pair interactions $\epsilon(s_i, s_j)$. Extensions to many-body interactions are possible by introducing terms in the Hamiltonian $\propto \Sigma_{i<j<k} \epsilon(s_i, s_j, s_k, \ldots)\Delta(r_i - r_j)\Delta(r_j - r_k) \ldots$ (see Part II, sections 3 and 5).

It is implicitly assumed in Eq. (6.1) that the requirements of chain connectivity are met: r_i and r_{i+1} are always near each other in space for all i, so e.g. if the chain were on a lattice, the explicit constraint $\Pi_i \delta(r_{i+1} - r_i - a)$ must appear in the trace. Some theoretical models introduce harmonic spring constraints between consecutive residues along the chain, which is formally convenient. The presence of chain connectivity and excluded volume make a general description of reconfigurational kinetics difficult.[20] However the thermodynamics

[20] Models of folding usually assume a well-connected set of states, however for some native structures, for example those proteins with knotted native structures (Taylor, 2000), it may become important to explicitly account for chain topological constraints.

may be obtained if either the partition function or density of states is known. In Eq. (6.1) it is assumed that the solvent molecules equilibrate considerably faster than the polymer, i.e. the solvent degrees of freedom have been integrated over to yield the effective interaction matrix $\epsilon(s_i, s_j)$, just as integrating over the solvent coordinates leads to an effective attraction between two hydrophobic molecules. The averaging may lead to many-body effects mentioned above. The explicit temperature dependence of the hydrophobic effect may be inserted into the coupling energies in the model (see e.g. Dill *et al.* 1989) to investigate phenomena associated with temperature-dependent interactions such as cold-denaturation, but has traditionally been left out.

Due to the large number of possible values for pair interaction energies in a realistic protein Hamiltonian, it is common to let the number of effective monomer species $p \to \infty$ and to take the interactions $\epsilon(s_i, s_j)$ be independent, random variables, usually Gaussianly distributed, with mean $\bar{\epsilon}$ and variance b^2:

$$P(\epsilon) = (2\pi b^2)^{-\frac{1}{2}} \mathrm{e}^{-(\epsilon-\bar{\epsilon})^2/2b^2}. \tag{6.2}$$

The coarse-grained Hamiltonian [Eq. (6.1)] governs the folding of the backbone chain, the motivation being that folding is largely a transition of the backbone, with side-chain ordering slaved to backbone ordering. As mentioned above, the backbone carries most of the entropy. Experimental observations such as the correlation of rates with topological properties of the native backbone structure mentioned above in Section 5 further support this description. The Hamiltonian in Eq. (6.1) is coarse-grained in that it looks at effective residue–residue interactions: the residues are collective objects that may consist of many atoms, or even several amino acids, depending which degrees of freedom have been removed by phase transitions (Section 4), or are irrelevant. Removing extra degrees of freedom from the full energy function simply renormalizes the energy and entropy scales on the landscape (Onuchic *et al.* 1995): the model still exhibits folding, but with a rescaled transition temperature. Scaling coupling energies ϵ by the critical temperature gives a universal phase diagram, for systems falling into the universality class of backbone-entropy-driven unfolding as discussed previously.

The system governed by the Hamiltonian in Eqs. (6.1) and (6.2) with $\bar{\epsilon} = 0$ contains only a single energy scale b and a single entropy scale s_0, at fixed density. So there is a phase transition temperature, T_{G}, as described in Section 2.1. However even if $\bar{\epsilon} = 0$, coupling the dependence of entropy on polymer density can lead to a separate collapse transition from the glass transition, resulting in coil, molten globule, and frozen (glass) phases in the phase diagram. As elaborated earlier in Section 2, the RHP landscape described by Eqs. (6.1) and (6.2) (and Figs. 4*c* and 5*c*) does not describe the observed signatures of protein folding, whereas a minimally frustrated energy landscape does.

The physical manifestations of evolution to a funneled energy landscape may be embodied in a Hamiltonian which has *a priori* a low energy in a particular state, so that the coupling energies $\epsilon_{ij}^{\mathrm{N}}$ tend to be stronger in a given (native) conformation. The mean native interaction is taken to be stronger to embody minimal-frustration in the folded structure (Bryngelson & Wolynes, 1987). Then for the ensemble of sequences folding to that native conformation, the native contact energies should be chosen from a distribution with a lower mean $\bar{\epsilon}_{\mathrm{N}}$, and with a different variance b_{N}^2 that may be smaller:

$$\mathscr{H}_{\mathrm{P}}(\{s_i\}, \{r_i\}) = \sum_{i<j} [\epsilon_{ij}^{\mathrm{N}} \Delta_{ij} \Delta_{ij}^{\mathrm{N}} + \epsilon_{ij} \Delta_{ij}(1 - \Delta_{ij}^{\mathrm{N}})]. \tag{6.3}$$

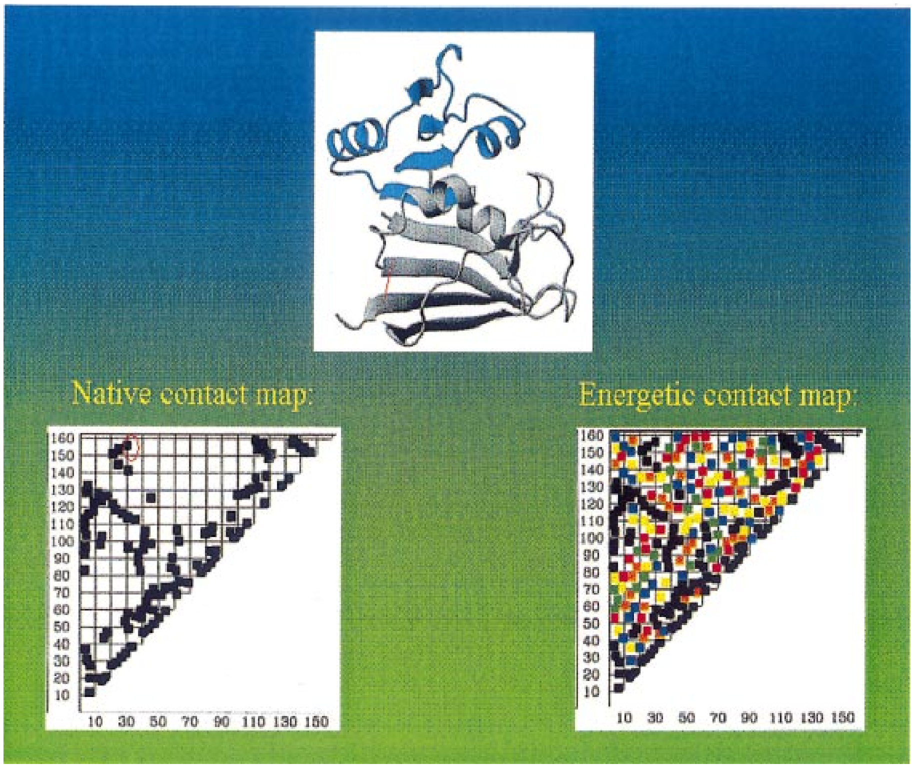

Fig. 11. The native contact map marks which residues are in contact (within a cut-off length) in the native structure, by denoting contacts as black squares on a plot with residue index on both axes (only half the plot is needed since the contact matrix is symmetric). For example the contact marked by the red bar in the native structure of DHFR, a two-domain α/β enzyme, is between residues 32 and 156, and is circled in red on the native contact map. The energetic contact map illustrated schematically here accounts for all contacts, native or non-native, and encodes (here by color) the contact energies by strength. (Adapted from Clementi *et al.* 2000a.)

Equation (6.3) gives the energy of a particular configuration defined by the set of contact interactions $\{\Delta_{ij}\} = \{\Delta\ (r_i - r_j)\}$. Energies $\epsilon_{ij}^{\mathrm{N}}$ given to native contacts are either chosen from a distribution, possibly Gaussian

$$P(\epsilon_{ij}^{\mathrm{N}}) = (2\pi b_{\mathrm{N}}^2)^{-\frac{1}{2}}\, \mathrm{e}^{-(\epsilon_{ij}^{\mathrm{N}} - \bar{\epsilon}_{\mathrm{N}})^2/2b_{\mathrm{N}}^2}. \tag{6.4}$$

or may be specified explicitly as the set $\{\epsilon_{ij}^{\mathrm{N}}\}$. Energies ϵ_{ij} of non-native contacts are chosen from the distribution in Eq. (6.2). The double sum in Eq. (6.3) is over residue indices, and $\Delta_{ij} = 1$ if residues i and j are in contact (within a cut-off length) in a configuration, $\Delta_{ij} = 0$ otherwise. $\Delta_{ij}^{\mathrm{N}} = 1$ if these residues are also in contact in the *native* configuration, and $\Delta_{ij}^{\mathrm{N}} = 0$ otherwise. In Part II we will calculate thermodynamics for the Hamiltonians [Eqs. (6.1) and (6.3)].

Here, instead of going into the formalism involved in computing a trace over the Hamiltonian in Eq. (6.3), we can get an intuitive feel for the what such a statistical Hamiltonian means by considering an energetic contact map as in Fig. 11. The native contact map counts which residues are interacting in the native structure. When all structures are considered, non-native as well as native contacts may be made. For the real system there is some particular realization of coupling energies for native and non-native contacts, which are shown in Fig. 11 by color-coding contacts so that for example darker squares indicate stronger contacts. The statistical Hamiltonian approach lets at least the non-native (and perhaps the native) interactions be chosen from the Gaussian distribution of Eq. (6.2) [and perhaps also Eq. (6.4)], to investigate universal features for the ensemble of sequences that fold to a given structure, say to the particular structure shown in Fig. 11.

The essence of minimal frustration for the real system's energy function, formalized through such a statistical Hamiltonian as Eq. (6.3), is that the bias to the native structure can be seen as a signal through the non-native noise. In the language of neural networks, the 'network' here is trained to have one dominant memory. In the language of spin-glasses, there is a 'ferromagnetic' bias to the native structure super-imposed on a random spin-glass.

A tacit assumption in the illustration of Fig. 11 is that at some level, interaction energies may be decomposed into two-body terms. To account for many-body effects, a higher dimensional representation is needed.

7. Conclusions and future prospects

An understanding of protein folding provides a link between the genetic code in the DNA molecule, and the structure and function of a living organism. However a description of protein folding is impeded by the complexity of the process. Still much of this complexity can in fact be exploited by taking a statistical approach to the energetics of protein conformation, that is to the energy landscape. The energy landscape approach explains when and why self-averaging behavior will govern the folding process, and when sequence-specific behavior, such as specific folding pathways or intermediates, should be observed.

We considered various possible landscapes in this review and found that the most likely landscape topography for a simple protein is that of an overall funnel with some residual heteropolymer ruggedness present. Evolution of amino-acid sequences to those having such a landscape solves various problems seemingly present in folding. Specifically, the native structure can be found on biological timescales, and is stable at biological temperatures. The native structure is robust to most perturbations in environment or sequence, as well as to thermal fluctuations.

Many transitions can occur within proteins in addition to the folding transition. Examples include collapse, helix-coil, liquid-crystal, microphase separation, and glass transitions. Thus a rich phase diagram for the system may be constructed. The understanding of this diverse behavior is facilitated by applying the methods of statistical mechanics of disordered systems, polymer physics, and theories of phase transitions in finite-sized systems.

Because of minimal frustration and the resulting funneled energy landscape of a protein, many features of folding are fortunately self-averaging and thus universally applicable to a wide class of protein sequences. As a consequence of the minimally frustrated character of the landscape, folding can be projected onto one or few reaction coordinates without too much loss of kinetic information.

In Part II of this review, we continue by exploring some central topics to energy landscape theory, such as kinetics of disordered systems, reaction coordinates, many-body effects, generalizing scalar order parameters to a field theory, and the theory of polymers under topological constraints. These ideas have all come together to bring us to a fuller understanding of the physical processes involved in protein folding.

8. Acknowledgments

We are grateful to Peter Wolynes and Jin Wang for numerous insightful and enjoyable discussions, and we also thank Laurel Latto (laurel_latto@juno.com) for her help with the color figure sketches. This work was funded by NSF Bio-Informatics fellowship DBI9974199 and NSF Grant MCB0084797.

9. Appendix: Glossary of terms

Contact order – a measure of the mean sequence length separating interacting residues in the native structure.

Energy gap – the difference in energy between either the native and random heteropolymer ground states, or the native and the mean unfolded energy, depending on the context.

Energy landscape – the network of all conformational states of a protein, with an internal free-energy associated with each conformation, and with the connectivity of the network specified or assumed implicitly.

ϕ-values – a measure of the participation of residues in the folding transition state ensemble. A particular residue is mutated to one interacting similarly with its neighbors, and the effect on folding rates and stability treated as a small perturbation. The ratio of the change in log rate to the change in stability is the ϕ-value for that residue.

Folded state – thermodynamic state dominated by the correct native conformation (the native state).

Folding temperature (T_F) – the temperature at which the native state is in thermodynamic equilibrium with the unfolded state, i.e. at this temperature the free-energy typically has a double well structure of nearly equal depths for a first-order folding transition. For a sequence to be foldable to a stable structure in a reasonable time $T_F > T_G$.

Free-energy profile – the free-energy $E - TS$ as a function of an order parameter or parameters, which are usually taken to measure structural similarity to the native conformation. Serves as a reaction surface for folding.

Frustration – property of a heteropolymer configuration where some interactions are satisfied (low in energy) at the expense of necessarily making some unfavorable interactions.

Glass temperature (T_G) – the temperature at which a random heteropolymer is frozen into one of its nearly degenerate ground states (sometimes referred to as the freezing temperature).

Glass phase – phase of a heteropolymer below T_G dominated by one or very few conformations (sometimes referred to as the frozen phase).

Generalized Random Energy Model or GREM – exactly solvable model of a system accurate up to pair correlations between states, so that high-order probability distributions factor into products of joint-probability distributions. All states in the model are at the nodes of an ultrametric tree. The model exhibits a phase transition to a low-temperature, reduced entropy state consisting of a set of globally distinct basins. Lowering the temperature further results in a gradual freezing to a zero entropy state for the heteropolymer.

Minimal frustration – property of the ground state of protein-like sequences of amino acids where interaction energies are not as competing or frustrated as the majority of purely random sequences. This yields a native state which is a dominant, unique ground state, with lower energy than the putative ground state energy of a typical random sequence, and an energy landscape with a funnel topography.

Misfolded state – a protein configuration dominated by low-energy, non-native interactions.

Molten globule – collapsed, liquid drop-like state of a polymer which may have anywhere from negligible to a partial amount of native structure and typically a large conformational entropy. May be induced by adding acid or other denaturant, or raising the temperature.

Monomer – the statistical segment in a polymer which contributes to the backbone conformational entropy, and which may contain several amino acids depending on the degree of secondary structure present.

Native state – same as folded state.

Q – a measure of similarity to a given structure, usually the folded structure. Here *Q* is almost always taken to be the fraction of contacts or interactions in common with the native structure.

Random Energy Model or REM – exactly solvable model of a system taking energy levels to be random and uncorrelated, so that pair and higher-order joint probability distributions factorize into products of single-state distributions. The model exhibits a phase transition to a low-temperature phase of zero entropy.

Random heteropolymer or RHP – a polymer having a random sequence of amino acids.

Random coil – phase of a polymer behaving as a self-avoiding random coil. As a thermodynamic phase of a protein, it may contain some secondary helical structure.

Stability gap – see energy gap.

Structural dispersion – a measure of the variance in sequence length separating interacting residues in the native structure.

Tertiary structure – organization between units of secondary structure, the dominant ordering process of which is simulated lattice models (although some aspects of secondary structure formation may be captured in lattice models such as β-sheet formation).

Transition state – either the collection of configurations which constitute the maximum on the folding free-energy profile, or when considered kinetically, a member of the ensemble of configurations equally likely to fold or unfold.

Unfolded state – the high entropy state where protein function is lost, attributed to the loss of the 3D folded structure. The unfolded state may be globular or coil depending where parameters of the system lie on the phase diagram.

10. References

Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10026–10036.

Abramowitz, M. & Stegun, I. A. (1972). *Handbook of Mathematical Functions*, 9th edn. New York: Dover.

Alm, E. & Baker, D. (1999). Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. natn. Acad. Sci. USA* **96**, 11305–11310.

Alonso, D. O. V., Dill, K. A. & Stigter, D.. (1991). The three states of globular proteins: acid denaturation. *Biopolymers* **31**, 1631–1649.

Amit, D. J., Gutfreund, H. & Sompolinsky, H. (1985). Spin-glass models of neural networks. *Phys. Rev. (A)* **32**, 1007–1018.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science* **181**, 223.

Austin, R. H., Beeson, K. W., Eisenstein, L. & Frauenfelder, H. (1975). Dynamics of ligand binding to myoglobin. *Biochemistry* **14**, 5355–5373.

Bachas, C.. (1984). Computer-intractability of the frustration model of a spin-glass. *J. Phys.* (*A*) **17**, L709–L712.

Bai, Y., Sosnick, T. R., Mayne, L. & Englander, S. W. (1995). Protein folding intermediates: native-state hydrogen exchange. *Science* **269**, 192–197.

Balbach, J., Forge, V., Lau, W. S., Jones, J. A., van Nuland, N. A. J. & Dobson, C. M. (1997). Detection of residue contacts in a protein folding intermediate. *Proc. natn. Acad. Sci. USA* **94**, 7182–7185.

Baldwin, R. L. (1975). Intermediates in protein folding reactions and mechanism of protein folding. *Annu. Rev. Biochem.* **44**, 453–475.

Baldwin, R. L. (1995). The nature of protein folding pathways: the classical versus the new view. *J. biomol. NMR* **5**, 103–109.

Baldwin, R. L. (1999). Protein folding from 1961 to 1982. *Nature struct. Biol.* **6**, 814–817.

Baldwin, R. L. & Rose, G. D. (1999). Is protein folding hierarchic? II. Folding intermediates and transition states. *TIBS* **24**, 77–83.

Ball, K. D., Berry, R. S., Kunz, R. E., Li, F. Y., Proykova, A. A. & Wales, D. J. (1996). From topographies to dynamics on multidimensional potential energy surfaces of atomic clusters. *Science* **271**, 963–966.

Ballew, R. M., Sabelko, J. & Gruebele, M. (1996). Direct observation of fast protein folding: the initial collapse of apomyoglobin. *Proc. natn. Acad. Sci. USA* **93**, 5759–5764.

Barahona, F. (1982). On the computational-complexity of ising spin-glass models. *J. Phys.* (*A*) **15**, 3241–3253.

Baum, E. B. (1986). Intractable computations without local minima. *Phys. Rev. Lett.* **57**, 2764–2767.

Boczko, E. M. & Brooks, C. L. (1995). First principles calculation of the folding free energy for a three helix bundle protein. *Science* **269**, 393–396.

Bornberg-Bauer, E. & Chan, H. S. (1999). Modeling evolutionary landscapes: mutational stability, topology, and super-funnels in sequence space. *Proc. natn. Acad. Sci. USA* **96**, 10689–10694.

Brooks, C. L., Gruebele, M., Onuchic, J. N. & Wolynes, P. G. (1998). Chemical physics of protein folding. *Proc. natn. Acad. Sci. USA* **95**, 11037–11038.

Brown, B. M. & Sauer, R. T. (1999). Tolerance of arc repressor to multiple-alanine substitutions. *Proc. natn. Acad. Sci. USA* **96**, 1983–1988.

Bryngelson, J. D. (1994). When is a potential accurate enough for structure prediction? theory and application to a random heteropolymer model of protein folding. *J. chem. Phys.* **100**, 6038–6045.

Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995). Funnels, pathways and the energy landscape of protein folding. *Proteins* **21**, 167–195.

Bryngelson, J. D. & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. natn. Acad. Sci. USA* **84**, 7524–7528.

Bryngelson, J. D. & Wolynes, P. G. (1989). Intermediates and barrier crossing in a random energy model (with applications to protein folding). *J. phys. Chem.* **93**, 6902–6915.

Bryngelson, J. D. & Wolynes, P. G. (1990). A simple statistical field theory of heteropolymer collapse with applications to protein folding. *Biopolymers* **30**, 177–188.

Buchler, N. E. G. & Goldstein, R. A. (1999). Universal correlation between energy gap and foldability for the random energy model and lattice proteins. *J. chem. Phys.* **111**, 6599–6609.

Buchler, N. E. G. & Goldstein, R. A. (2000). Surveying determinants of protein structure designablity across different energy models and amino-acid alphabets: a consensus. *J. chem. Phys.* **112**, 2533–2547.

Burton, R. E., Huang, G. S., Daugherty, M. A., Calderone, T. & Oas, T. G. (1997). The energy landscape of a fast-folding protein mapped by Ala-Gly substitutions. *Nature struct. Biol.* **4**, 305–310.

Caldeira, A. O. & Leggett, A. J. (1983). Quantum tunneling in a dissipative system. *Ann. Phys.* **149**, 374–456. [Erratum **153**, 445 (1983).]

Camacho, C. J. & Thirumalai, D. (1993). Minimum energy compact structures of random sequences of heteropolymers. *Phys. Rev. Lett.* **71**, 2505–2508.

Chan, H. S. & Dill, K. A. (1994). Transition states and folding dynamics of proteins and heteropolymers. *J. chem. Phys.* **100**, 9238–9257.

Chirgadze, Y. N. (1987). Deduction and systematic classification of spatial motifs of the antiparallel-beta-structure in globular proteins. *Acta Crystallogr.* (*A*) **43**, 405–417.

Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature struct. Biol.* **6**, 1005–1009.

Chothia, F. C., Levitt, M. & Richardson, D. (1977). Structure of proteins – packing of alpha-helices and pleated sheets. *Proc. natn. Acad. Sci. USA* **74**, 4130–4134.

Clementi, C., Jennings, P. A. & Onuchic, J. N. (2000a). How native state topology affects the folding of dihydrofolate reductase and interleukin-1beta. *Proc. natn. Acad. Sci. USA* **97**, 5871–5876.

Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000b). Topological and energetic factors: what determines the structural details of the transition state ensemble and en-route intermediates for protein folding? An investigation for small globular proteins. *J. molec. Biol.* **298**, 937–953.

Cordes, M. H. J., Davidson, A. R. & Sauer, R. T. (1996). Sequence space, folding and protein design. *Curr. Opin. struct. Biol.* **6**, 3–10.

Creighton, T. E. (1974). Intermediates in the refolding of reduced pancreatic trypsin inhibitor. *J. molec. Biol.* **87**, 579–602.

Creighton, T. E. (Ed). (1992). *Protein Folding*. New York: W. H. Freeman.

Dael, H. V., Haezebrouck, P., Morozova, L., Arico-Muendel, C. & Dobson, C. M. (1993). Partially folded states of equine lysozyme, structural characterization and significance for protein folding. *Biochemistry* **32**, 11886–11894.

Daggett, V. & Levitt, M. (1993). Protein unfolding pathways explored through molecular dynamics simulations. *J. molec. Biol.* **232**, 600–619.

Daggett, V., Li, A., Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1996). Structure of the transition state for folding of a protein derived from experiment and simulation. *J. molec. Biol.* **257**, 430–440.

Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci.* **5**, 895–903.

d'Auriac, J. C. A., Preissmann, M. & Rammal, R. (1985). The random field ising model: algorithmic complexity and phase transition. *J. Physique Lett.* **46**, L173–L180.

Davidson, A. R., Lumb, K. J. & Sauer, R. T. (1995). Cooperatively folded proteins in random libraries. *Nature struct. Biol.* **2**, 856–863.

Davidson, A. R. & Sauer, R. T. (1994). Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. natn. Acad. Sci. USA* **91**, 2146–2150.

de Gennes, P.-G. (1975). *The Physics of Liquid Crystals*. Oxford: Clarendon Press.

de Gennes, P. G. & Pincus, P. (1977). Nematic polymers. *Polym. Prepr.* **18**, 161–172.

Deng, Y. & Smith, D. L. (1998). Identification of unfolding domains in large proteins by their unfolding rates. *Biochemistry* **37**, 6256–6262.

Derrida, B. (1981). Random-energy model: an exactly solvable model of disordered systems. *Phys. Rev.* (*B*) **24**, 2613–2626.

Desjarlais, J. R. & Handel, T. M. (1995). De novo design of the hydrophobic cores of proteins. *Protein Sci.* **4**, 2006–2018.

Deutsch, J. M. & Kurosky, T. (1996). New algorithm for protein design. *Phys. Rev. Lett.* **76**, 323–326.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry* **29**, 7133–7155.

Dill, K. A., Alonso, D. O. V. & Hutchinson, K. (1989). Thermal stabilities of globular proteins. *Biochemistry* **28**, 5439–5449.

Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D. & Chan, H. S. (1995). Principles of protein folding – a perspective from simple exact models. *Protein Sci.* **4**, 561–602.

Dill, K. A. & Chan, H. S. (1997). From levintal to pathways to funnels. *Nature struct. Biol.* **4**, 10–19.

Dill, K. A. & Stigter, D. (1995). Modeling protein stability as heteropolymer collapse. *Adv. Protein Chem.* **46**, 59–104.

Dobson, C. M., Sali, A. & Karplus, M. (1998). Protein folding: a perspective from theory and experiment. *Angew. Chem. Int. Ed. Engl.* **37**, 868–893.

Doig, A. J. & Sternberg, M. J. E. (1995). Side-chain conformational entropy in protein folding. *Protein Sci.* **4**, 2247–2251.

Dreyer, M. K., Borcherding, D. R., Dumont, J. A., Peet, N. P., Tsay, J. T., Wright, P. S., Bitoni, A. J., Shen, J. & Kim, S. H. (2001). Crystal structure of human cyclin-dependent kinase 2 in complex with the adenine-derived inhibitor h717. *J. med. Chem.* **44**, 524–530.

Du, R., Pande, V. S., Grosberg, A. Y., Tanaka, T. & Shakhnovich, E. S. (1998). On the transition coordinate for protein folding. *J. chem. Phys.* **108**, 334–350.

Du, R., Pande, V. S., Grosberg, A. Y., Tanaka, T. & Shakhnovich, E. S. (1999). On the role of conformational geometry in protein folding. *J. chem. Phys.* **111**, 10375–10380.

Duan, Y. & Kollman, P. A. (1998). Pathways to protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740–744.

Eaton, W. A., Munoz, V., Thompson, P. A., Chan, C. K. & Hofrichter, J. (1997). Submillisecond kinetics of protein folding. *Curr. Opin. struct. Biol.* **7**, 10–14.

Eaton, W. A., Munoz, V., Thompson, P. A., Henry, E. R. & Hofrichter, J. (1998). Kinetics and dynamics of loops, α-helices, β-hairpins, and fast-folding proteins. *Acc. Chem. Res.* **31**, 745–753.

Epstein, C. J., Goldberger, R. F. & Anfinsen, C. B. (1963). The genetic control of tertiary protein structure: studies with model systems. *Cold Spring Harbor Symp. Quant. Biol.* **28**, 439.

Feng, Y., Sligar, S. G. & Wand, A. J. (1994). The solution structure of apocytochrome b_{562}. *Nature struct. Biol.* **1**, 30–36.

Fernández, A., Colubri, A. & Berry, R. S. (2000). Topology to geometry in protein folding: β-lactoglobulin. *Proc. natn. Acad. Sci. USA* **97**, 14062–14066.

Ferry, J. D. (1950). Mechanical properties of substances of high molecular weight. VI. Dispersion in concentrated polymer solutions and its dependence on temperature and concentration. *J. Am. chem. Soc.* **72**, 3746–3752.

Fersht, A. R. (1999). *Structure and Mechanism in Protein Science*, 1st edn. New York: W. H. Freeman & Co.

Fersht, A. R. (2000). Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. natn. Acad. Sci. USA* **97**, 1525–1529.

Fersht, A. R., Matouschek, A. & Serrano, L. (1992). I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. molec. Biol.* **224**, 771–782.

Fiebig, K. M. & Dill, K. A. (1993). Protein core assembly processes. *J. chem. Phys.* **98**, 3475–3487.

Finkelstein, A. V. & Badretdinov, A. Y. (1997). Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Folding & Design* **2**, 115–121.

Finkelstein, A. V. & Shakhnovich, E. I. (1989). Theory of cooperative transitions in protein molecules. 2. Phase diagram for a protein molecule in solution. *Biopolymers* **28**, 1681–1694.

Flanagan, J. M., Kataoka, M., Fujisawa, T. & Engelman, D. M. (1993). Mutations can cause large changes in the conformation of a denatured protein. *Biochemistry* **32**, 10359–10370.

Flory, P. J. (1949). The configuration of real polymer chains. *J. chem. Phys.* **17**, 303–310.

Flory, P. J. (1953). *Principles of Polymer Chemistry*. Ithaca: Cornell University.

Flory, P. J. (1956). Phase equilibria in solutions of rod-like particles. *Proc. Roy. Soc. Lond.* (*A*) **234**, 73–89.

Forge, V., Hoshino, M., Kuwata, K., Arai, M., Kuwajima, K., Batt, C. A. & Goto, Y. (2000). Is folding of β-lactalbumin non-hierarchic? intermediate with native-like β-sheet and non-native α-helix. *J. molec. Biol.* **296**, 1039–1051.

Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. (1991). The energy landscapes and motions of proteins. *Science* **254**, 1598–1603.

Fredrickson, G. H., Milner, S. T. & Leibler, S. L. (1992). Multicritical phenomena and microphase ordering in random block copolymer melts. *Macromolecules* **25**, 6341–6354.

Fulton, K. F., Main, E. R. G., Daggett, V. & Jackson, S. E. (1999). Mapping the interactions present in the transition state for unfolding/folding of fkbp12. *J. molec. Biol.* **291**, 445–461.

Galzitskaya, O. V. & Finkelstein, A. V. (1999). A theoretical search for folding/unfolding nuclei in three dimensional protein structures. *Proc. natn. Acad. Sci. USA* **96**, 11299–11304.

Garcia, C., Nishimura, C., Cavagnero, S., Dyson, H. J. & Wright, P. E. (2000). Changes in the apomyoglobin folding pathway caused by mutation of the distal histidine residue. *Biochemistry* **39**, 11227–11237.

Garel, T., Leibler, L. & Orland, H. (1994). Random hydrophilic-hydrophobic copolymers. *J. Phys. II* (*France*) **4**, 2139–2148.

Garel, T., Orland, H. & Pitard, E. (1998). Protein folding and heteropolymers. In *Spin Glasses and Random Fields* (ed. A. P. Young). River Edge, NJ: World Scientific.

Gesteland, R. F. & Atkins, J. F. (1993). *The RNA World*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Gladwin, S. T. & Evans, P. A. (1996). Structure of very early protein folding intermediates: new insights through a variant of hydrogen exchange labelling. *Folding & Design* **1**, 407–417.

Gō, N. (1983). Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183–210.

Goldenfeld, N. (1992). *Lectures on Phase Transitions and the Renormalization Group*. Reading, MA: Addison-Wesley.

Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992). Optimal protein-folding codes from spin-glass theory. *Proc. natn. Acad. Sci. USA* **89**, 4918–4922.

Govindarajan, S. & Goldstein, R. A. (1997a). Evolution of model proteins on a foldability landscape. *Proteins Struct. Funct. Genet.* **29**, 461–466.

Govindarajan, S. & Goldstein, R. A. (1997b). The foldability landscape of model proteins. *Biopolymers* **42**, 427–438.

Grantcharova, V. P., Santiago, J. V., Baker, D. & Riddle, D. S. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src sh3 domain. *Nature struct. Biol.* **5**, 714–720.

Griko, Y. V., Makhatadze, G. I., Privalov, P. L. & Hartley, R. W. (1994). Thermodynamics of barnase unfolding. *Protein Sci.* **3**, 669–676.

Griko, Y. V., Privalov, P. L., Venyaminov, S. Y. & Kutyshenko, V. P. (1988). Thermodynamic study of the apomyoglobin structure. *J. molec. Biol.* **202**, 127–138.

Gross, D. J. & Mézard, M. (1984). The simplest spin glass. *Nuc. Phys.* (*B*) **240**, 431–452.

Grote, R. F. & Hynes, J. T. (1980). The stable states picture of chemical reactions. II. Rate constants for condensed and gas phase reaction models. *J. chem. Phys.* **73**, 2715–2732.

Gruebele, M. (1999). The fast protein folding problem. *Annu. Rev. Phys. Chem.* **50**, 485–516.

Gunton, J. D., Miguel, M. S. & Sahni, P. S. (1983). The dynamics of first order transitions. In *Phase Transitions and Critical Phenomena* (eds. C. Domb & J. L. Lebowitz), vol. 8, pp. 267–466. New York: Academic Press.

Guo, Z. & Thirumalai, D. (1995). Kinetics of protein folding: nucleation mechanism, time scales, and pathways. *Biopolymers* **36**, 83–102.

Guo, Z. & Thirumalai, D. (1997). The nucleation-collapse mechanism in protein folding: evidence for the non-uniqueness of the folding nucleus. *Folding & Design* **2**, 377–391.

Gutin, A. M., Abkevich, V. I. & Shakhnovich, E. I. (1995). Evolution-like selection of fast-folding model proteins. *Proc. natn. Acad. Sci. USA* **92**, 1282–1286.

Gutin, A. M. & Shakhnovich, E. I. (1993). Ground state of random copolymers and the discrete random energy model. *J. chem. Phys.* **98**, 8174–8177.

Hamada, D., Segawa, S. & Goto, Y. (1996). Non-native alpha-helical intermediate in the refolding of beta-lactalbumin, a predominantly beta-sheet protein. *Nature struct. Biol.* **3**, 868–874.

Hecht, M. H., Hehir, K. M., Nelson, H. C. M., Sturtevant, J. M. & Sauer, R. T. (1985). Increasing and decreasing protein stability – effects of refertant substitutions on the thermal-denaturation of phage lambda-repressor. *J. cell. Biochem.* **29**, 217–224.

Heidary, D. K., O'Neill, J. C., Roy, M. & Jennings, P. A. (2000). An essential intermediate in the folding of dihydrofolate reductase. *Proc. natn. Acad. Sci. USA* **97**, 5866–5870.

Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science* **273**, 595–602.

Honeycutt, J. & Thirumalai, D. (1992). The nature of the folded state of globular proteins. *Biopolymers* **32**, 695–709.

Honig, B. (1999). Protein folding: from the levinthal paradox to structure prediction. *J. molec. Biol.* **293**, 283–293.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. natn. Acad. Sci. USA* **79**, 2554–2558.

Horovitz, A. & Fersht, A. (1992). Co-operative interactions during protein folding. *J. molec. Biol.* **224**, 733–740.

Houry, W. A., Sauder, J. M., Roder, H. & Scheraga, H. A. (1998). Definition of amide protection factors for early kinetic intermediates in protein folding. *Proc. natn. Acad. Sci. USA* **95**, 4299–4302.

Huang, G. S. & Oas, T. G. (1995). Structure and stability of monomeric lambda-repressor – NMR evidence for 2-state folding. *Biochemistry* **34**, 3884–3892.

Hummer, G., García, A. E. & Garde, S. (2000). Conformational diffusion and helix formation kinetics. *Phys. Rev. Lett.* **85**, 2637–2640.

Hummer, G., García, A. E. & Garde, S. (2001). Helix nucleation kinetics from molecular simulations in explicit solvent. *Proteins* **42**, 77–84.

Iba, Y., Tokita, K. & Kikuchi, M. (1998). Design equation: a novel approach to heteropolymer design. *J. Phys. Soc. Jpn.* **67**, 3985–3990.

Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. molec. Biol.* **254**, 260–288.

Jackson, S. E. & Fersht, A. R. (1991). Folding of chymotrypsin inhibitor 2. 1. Evidence for a two state transition. *Biochemistry* **30**, 10428–10435.

Karplus, M. & McCammon, J. A. (1983). Dynamics of proteins: elements and function. *Annu. Rev. Biochem.* **53**, 263–300.

Karplus, M. & Shakhnovich, E. I. (1992). Protein folding: theoretical studies of thermodynamics and dynamics. In *Protein Folding.* (ed. T. E. Creighton) chap. 4, pp. 127–195. New York: W. H. Freeman.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14**, 1–64.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wycoff, H. W. & Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* **181**, 662–666.

Kim, D. E., Fisher, C. & Baker, D. (2000). A breakdown of symmetry in the folding transition state of protein 1. *J. molec. Biol.* **298**, 971–984.

Kim, D. E., Gu, H. & Baker, D. (1998). The sequences of small proteins are not extensively optimized for natural selection. *Proc. natn. Acad. Sci. USA* **95**, 4982–4986.

Kim, P. S. & Baldwin, R. L. (1990). Intermediates in the folding reactions of small proteins. *Annu. Rev. Biochem.* **59**, 631–660.

Kim, S., Bracken, C. & Baum, J. (1999). Characterization of millisecond time-scale dynamics in the molten globule state of alpha-lactalbumin in nmr. *J. molec. Biol.* **294**, 551–560.

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.

Kirkpatrick, T. R., Thirumalai, D. & Wolynes, P. G. (1989). Scaling concepts for the dynamics of viscous liquids near an ideal glassy state. *Phys. Rev.* (*A*) **40**, 1045–1054.

Kitao, A., Hirata, F. & Gō, N. (1991). The effects of solvent on the conformation and the collective motions of protein: normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chem. Phys.* **158**, 447–472.

Klimov, D. K. & Thirumalai, D. (1998a). Co-operativity in protein folding: from lattice models with sidechains to real proteins. *Folding & Design* **3**, 127–139.

Klimov, D. K. & Thirumalai, D. (1998b). Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism. *J. molec. Biol.* **282**, 471–492.

Kramers, H. A. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* (*Utrecht*) **7**, 284–304.

Kuwajima, K., Yamaya, H. & Sugai, S. (1996). The burst-phase intermediate in the refolding of β-lactalglobulin studied by stopped-flow circular dichroism and absorption spectroscopy. *J. molec. Biol.* **264**, 806–822.

Kuwata, K., Shastry, R., Cheng, H., Hoshino, M., Batt, C. A., Goto, Y. & Roder, H. (2001). Structural and kinetic characterization of early folding events in beta-lactoglobulin. *Nature struct. Biol.* **8**, 151–155.

LaBean, T. H. & Kauffman, S. A. (1993). Design of synthetic gene libraries encoding random sequence proteins with desired ensemble characteristics. *Protein Sci.* **2**, 1249–1254.

Ladurner, A. G. & Fersht, A. R. (1997). Glutamine, alanine or glycine repeats inserted into the loop of a protein have minimal effects on stability and folding rates. *J. molec. Biol.* **273**, 330–337.

Lau, K. F. & Dill, K. A. (1990). Theory for protein mutability and biogenesis. *Proc. natn. Acad. Sci. USA* **87**, 638–642.

Lazaridis, T. & Karplus, M. (1997). New view of protein folding reconciled with the old through multiple unfolding simulations. *Science* **278**, 1928–1931.

Leibler, L. (1980). *Macromolecules* **13**, 1602.

Leibler, L., Orland, H. & Wheeler, J. C. (1983). Theory of critical micelle concentration for solutions of block co-polymers. *J. chem. Phys.* **79**, 3550–3557.

Leopold, P. E., Montal, M. & Onuchic, J. N. (1992). Protein folding funnels – a kinetic approach to the sequence structure relationship. *Proc. natn. Acad. Sci. USA* **89**, 8721–8725.

Levinthal, C. (1969). How to fold graciously. In *Mossbauer Spectroscopy in Biological Systems* (eds. P. DeBrunner, J. Tsibris & E. Munck), pp. 22–24. Urbana, IL: University of Illinois Press.

Levitt, M. & Warshel, A. (1975). Computer simulation of protein folding. *Nature* **253**, 694–698.

Li, H., Helling, R., Tang, C. & Wingreen, N. (1996). Emergence of preferred structures in a simple model of protein folding. *Science* **273**, 666–669.

Li, L., Mirny, L. A. & Shakhnovich, E. I. (2000). Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus. *Nature struct. Biol.* **7**, 336–342.

Lifshitz, I. M., Grosberg, A. Y. & Khokhlov, A. R. (1978). Some problems of the statistical physics of polymer chains with volume interaction. *Rev. Mod. Phys.* **50**, 683–713.

Lin, L., Pinker, R. J., Forde, K., Rose, G. D. & Kallenbach, N. R. (1994). Molten globular characteristics of the native state of apomyoglobin. *Nature struct. Biol.* **447**, 452.

López-Hernández, E. & Serrano, L. (1996). Structure of the transition state for folding of hte 129 aa protein chey resembles that of a smaller protein ci-2. *Folding & Design* **1**, 43–55.

Luisi, D. L., Wu, W. J. & Raleigh, D. P. (1999). Conformational analysis of a set of peptides corespondin to the entire primary sequence of the N-terminal domain of the ribosomal protein 19: evidence for stable native-like secondary structure in the unfolded state. *J. molec. Biol.* **287**, 395–407.

Lum, K., Chandler, D. & Weeks, J. D. (1999). Hydrophobicity at small and large length scales. *J. phys. Chem.* **103**, 4570–4577.

Luthey-Schulten, Z., Ramirez, B. E. & Wolynes, P. G. (1995). Helix-coil, liquid crystal and spin glass transitions of a collapsed heteropolymer. *J. phys. Chem.* **99**, 2177–2185.

Makhatadze, G. I. & Privalov, P. L. (1996). On the entropy of protein folding. *Protein Sci.* **5**, 507–510.

Marinari, E., Parisi, G., Ricci-Tersenghi, F., Ruiz-Lorenzo, J. J. & Zuliani, F. (2000). Replica symmetry breaking in short-range spin glasses: theoretical foundations and numerical evidences. *J. Stat. Phys.* **98**, 973–1047.

Marinari, E., Parisi, G., Ruiz-Lorenzo, J. & Ritort, F. (1996). Numerical evidence for spontaneously broken replica symmetry in 3d spin glasses. *Phys. Rev. Lett.* **76**, 843–846.

Martinez, J. C., Pisabarro, M. T. & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nature struct. Biol.* **5**, 721–729.

Martinez, J. C. & Serrano, L. (1999). The folding transition state between sh3 domains is conformationally restricted and evolutionarily conserved. *Nature struct. Biol.* **6**, 1010–1016.

Matheson, R. R. & Scheraga, H. A. (1978). A method for predicting nucleation sites for protein folding based on hydrophobic contacts. *Macromolecules* **11**, 819–829.

Matouschek, A., Kellis, J. T., Serrano, L., Bycroft, M. & Fersht, A. R. (1990). Transient folding intermediates characterized by protein engineering. *Nature* **346**, 440–445.

McCammon, J. A., Gelin, B. R. & Karplus, M. (1977). Dynamics of folded proteins. *Nature* **267**, 585–590.

McCammon, J. A. & Harvey, S. C. (1987). *Dynamics of Proteins and Nucleic Acids.* Cambridge: Cambridge.

Mélin, R., Li, H., Wingreen, N. S. & Tang, C. (1999). Designability, thermodynamic stability, and dynamics in protein folding: a lattice model study. *J. chem. Phys.* **110**, 1252–1262.

Metzler, R., Klafter, J., Jortner, J. & Volk, M. (1998). Multiple time scales for dispersive kinetics in early events of peptide folding. *Chem. Phys. Lett.* **293**, 477–484.

Mézard, M., Parisi, E. & Virasaro, M. A. (1986). *Spin Glass Theory and Beyond*. Singapore: World Scientific Press.

Mézard, M., Parisi, G., Sourlas, N., Toulouse, G. & Virasoro, M. (1984). Replica symmetry breaking and the nature of the spin glass phase. *J. Physique* **45**, 843–854.

Michnick, S. W. & Shakhnovich, E. (1998). A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. *Folding & Design* **3**, 239–251.

Miller, W. H. (1974). Quantum mechanical transition state theory and a new semiclassical model for reaction rate constants. *J. chem. Phys.* **61**, 1823–1834.

Miranker, A., Radford, S. E., Karplus, M. & Dobson, C. M. (1991). Demonstration by nmr of folding domains in lysozyme. *Nature* **349**, 633–636.

Mirny, L. A. & Shakhnovich, E. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding, kinetics and function. *J. molec. Biol.* **291**, 177–196.

Monod, J., Wyman, J. & Changeux, J.-P. (1965). On the nature of allosteric transitions: a plausible model. *J. molec. Biol.* **12**, 88.

Morrissey, M. P. & Shakhnovich, E. I. (1996). Design of proteins with selected thermal properties. *Folding* & *Design* **1**, 391–405.

Munoz, V. & Eaton, W. A. (1999). A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. natn. Acad. Sci. USA* **96**, 11311–11316.

Munoz, V., Henry, E. R., Hofrichter, J. & Eaton, W. A. (1998). A statistical mechanical model for β-hairpin kinetics. *Proc. natn. Acad. Sci. USA*. **95**, 5872–5879.

Murphy, K. P. & Freire, E. (1993). Structural energetics of protein stability and folding cooperativity. *Pure Appl. Chem.* **65**, 1939.

Murzin, A. G. & Finkelstein, A. V. (1988). General architecture of the alpha-helical globule. *J. molec. Biol.* **204**, 749–769.

Nelson, E. D. & Onuchic, J. N. (1998). Proposed mechanism for stability of proteins to evolutionary mutations. *Proc. natn. Acad. Sci. USA*. **95**, 10682–10686.

Nelson, E. D., Teneyck, L. F. & Onuchic, J. N. (1997). Symmetry and kinetic optimization of protein-like heteropolymers. *Phys. Rev. Lett.* **79**, 3534–3537.

Ngo, J. T., Marks, J. & Karplus, M. (1994). Computational complexity, protein structure prediction, and the levinthal paradox. In *The Protein Folding Problem and Tertiary Structure Prediction* (eds. K. Merz & S. Le Grand), pp. 435–508. Boston: Birkhauser.

Nymeyer, H., Garcia, A. E. & Onuchic, J. N. (1998). Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. natn. Acad. Sci. USA*. **95**, 5921–5928.

Nymeyer, H., Socci, N. D. & Onuchic, J. N. (2000). Landscape approaches for determining the ensemble of folding transition states: success and failure hinge on the degree of minimal frustration. *Proc. natn. Acad. Sci. USA* **97**, 634–639.

Oliveberg, M., Tan, Y., Silow, M. & Fersht, A. (1998). The changing nature of the protein folding transition state: implications for the shape of the free energy profile for folding. *J. molec. Biol.* **277**, 933–943.

Onsager, L. (1949). The effects of shape on the interaction of colloidal particles. *Ann. N.Y. Acad. Sci.* **51**, 627–659.

Onuchic, J. N. (1997). Contacting the protein folding funnel with NMR. *Proc. natn. Acad. Sci. USA* **94**, 7129–7131.

Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. (1997). Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600.

Onuchic, J. N., Nymeyer, H., Garcia, A. E., Chahine, J. & Socci, N. D. (2000). The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. *Adv. Protein Chem.* **53**, 87–152.

Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z. & Wolynes, P. G. (1996). Protein folding funnels: the nature of the transition state ensemble. *Folding & Design* **1**, 441–450.

Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z. & Socci, N. D. (1995). Toward an outline of the topography of a realistic protein-folding funnel. *Proc. natn. Acad. Sci. USA* **92**, 3626–3630.

Otzen, D. E. & Fersht, A. R. (1998). Folding of circular and permuted chymotrypsin inhibitor 2. Retention of the folding nucleus. *Biochemistry* **37**, 8139–8146.

Overington, J., Donnelly, D., Johnson, M., Šali, A. & Blundell, T. L. (1992). Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.* **1**, 216–226.

Pande, V. S., Grosberg, A. & Tanaka, T. (1994a). Phase diagram of imprinted copolymers. *J. Phys. II* (*France*) **4**, 1771–1784.

Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1994b). Thermodynamic procedure to synthesize heteropolymers that can renature to recognize a given target molecule. *Proc. natn. Acad. Sci. USA* **91**, 12976–12979.

Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1995). How accurate must potentials be for successful modeling of protein folding? *J. chem. Phys.* **103**, 9482–9491.

Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1997). Statistical mechanics of simple models of protein folding and design. *Biophys. J.* **73**, 3192–3210.

Pande, V. S., Grosberg, A. Y. & Tanaka, T. (2000). Heteropolymer freezing and design: towards physical models of protein folding. *Rev. Mod. Phys.* **72**, 259–314.

Pauling, L., Corey, R. B. & Branson, H. R. (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. natn. Acad. Sci. USA* **37**, 205–211.

Pechukas, P. (1976). Statistical approximations in collision theory. In *Dynamics of Molecular Collision, Part B* (ed. W. H. Miller), pp. 269–322. New York: Plenum Press.

Perutz, M. F. (1951). New X-ray evidence on the configuration of polypeptide chains. Polypeptide chains in poly-g-benzyl-t-glutamate, keratin and haemoglobin. *Nature* **167**, 1053–1054.

Perutz, M. F. (1970). Stereochemistry of cooperative effects of hemoglobin. *Nature* **228**, 726.

Plaxco, K. W., Larson, S., Ruczinski, I., Riddle, D. S., Thayer, E. C., Buchwitz, B., Davidson, A. R. & Baker, D. (2000a). Evolutionary conservation in protein folding kinetics. *J. molec. Biol.* **298**, 303–312.

Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. molec. Biol.* **277**, 985–994.

Plaxco, K. W., Simons, K. T., Ruczinski, I. & Baker, D. (2000b). Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. *Biochemistry* **39**, 11177–11183.

Plotkin, S. S. & Onuchic, J. N. (2000). Investigation of routes and funnels in protein folding by free energy functional methods. *Proc. natn. Acad. Sci. USA* **97**, 6509–6514.

Plotkin, S. S., Wang, J. & Wolynes, P. G. (1996). Correlated energy landscape model for finite, random heteropolymers. *Phys. Rev.* (*E*) **53**, 6271–6296.

Plotkin, S. S., Wang, J. & Wolynes, P. G. (1997). Statistical mechanics of a correlated energy landscape model for protein folding funnels. *J. chem. Phys.* **106**, 2932–2948.

Plotkin, S. S. & Wolynes, P. G. (2002). On the evolution of protein folding landscapes. Preprint.

Pohl, F. M. (1969). On the kinetics of structural transition I of some pancreatic proteins. *FEBS Lett.* **3**, 60–64.

Pollak, E. (1992). Variational transition state theory for dissipative systems. In *Activated Barrier Crossing* (eds. P. Hanggi & G. Fleming), pp. 5–41. Singapore: World Scientific.

Pollak, E., Tucker, S. & Berne, B. J. (1990). Variational transition-state theory for reaction rates in dissipative systems. *Phys. Rev. Lett.* **65**, 1399–1402.

Portman, J. J., Takada, S. & Wolynes, P. G. (1998). Variational theory for site resolved protein folding free energy surfaces. *Phys. Rev. Lett.* **81**, 5237–5240.

Privalov, P. L. (1979). Stability of proteins; small globular proteins. *Adv. Protein Chem.* **33**, 167–241.

Ptitsyn, O. B. (1992). The molten globule state in protein folding. In *Protein Folding* (ed. T. E. Creighton), p. 243. New York: W. H. Freeman.

Ptitsyn, O. B. (1994). Kinetic and equilibrium intermediates in protein folding. *Protein Eng.* **7**, 593–596.

Ptitsyn, O. B. (1995). Structures of folding intermediates. *Curr. Opin. struct. Biol.* **5**, 74–78.

Radford, S. A. & Dobson, C. M. (1995). Insights into protein folding using physical techniques-studies of lysozyme and alpha-lactalbumin. *Phil. Trans. R. Soc. Lond.* (*B*) **348**, 17–25.

Radford, S. A., Dobson, C. M. & Evans, P. A. (1992). The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature* **358**, 302–307.

Ramachandran, G. N. & Sassiekharan, V. (1968). Conformation of polypeptides and proteins. *Adv. Protein Chem.* **23**, 283–437.

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34**, 167–339.

Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D. (1999). Experiment and theory highlight role of native state topology in sh3 folding. *Nature struct. Biol.* **11**, 1016–1024.

Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q. & Baker, D. (1997). Functional rapidly folding proteins from simplified amino acid sequences. *Nature struct. Biol.* **4**, 805–809.

Roder, H. & Colon, W. (1997). Kinetic role of early intermediates in protein folding. *Curr. Opin. struct. Biol.* **7**, 15–28.

Sagnella, D. E., Straub, J. E. & Thirumalai, D. (2000). Time scales and pathways for kinetic energy relaxation in solvated proteins: application to carbonmonoxy myoglobin. *J. chem. Phys.* **113**, 7702–7711.

Šali, A., Shakhnovich, E. I. & Karplus, M. (1994). How does a protein fold? *Nature* **369**, 248–251.

Sanchez, I. C. (1979). Phase transition behavior of the isolated polymer chain. *Macromolecules* **12**, 980–988.

Sankoff, D., Morin, A.-M. & Cedergren, R. J. (1978). Evolution of 5s-RNA secondary structures. *Can. J. Biochem. Cell.* (*B*) **56**, 440–443.

Saven, J. G. & Wolynes, P. G. (1996). Local conformation signals and the statistical thermodynamics of collapsed helical proteins. *J. molec. Biol.* **257**, 199–216.

Saven, J. G. & Wolynes, P. G. (1997). Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules. *J. phys. Chem.* **101**, 8375–8389.

Schindler, T. & Schmid, F. X. (1996). Thermodynamic properties of an extremely rapid protein folding reaction. *Biochemistry* **35**, 16833–16842.

Schulman, B. A., Kim, P. S., Dobson, C. M. & Redfield, C. (1997). A residue-specific NMR view of the non-cooperative unfolding of a molten globule. *Nature struct. Biol.* **4**, 630–634.

Schuster, P. (1997). Landscapes and molecular evolution. *Physica (D)* **107**, 351–365.

Schuster, P., Fontana, W., Stadler, P. F. & Hofecker, I. (1994). From sequences to shapes and back: a case study in rna secondary structures. *Proc. R. Soc. Lond. (B)* **255**, 279–284.

Sfatos, C. D., Gutin, A. M. & Shakhnovich, E. I. (1993). Phase diagram of random copolymers. *Phys. Rev (E)* **48**, 465–475.

Sfatos, C. D., Gutin, A. M. & Shakhnovich, E. I. (1994). Phase transitions in a 'many-letter' random heteropolymer. *Phys. Rev. (E)* **50**, 2898–2905.

Sfatos, C. D., Gutin, A. M. & Shakhnovich, E. I. (1995). Critical compositions in the microphase separation transition of random copolymers. *Phys. Rev. (E)* **51**, 4727–4734.

Shakhnovich, E. I., Abkevich, V. & Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. *Nature* **379**, 96–98.

Shakhnovich, E. I. (1994). Proteins with selected sequences fold to unique native conformation. *Phys. Rev. Lett.* **72**, 3907–3910.

Shakhnovich, E. I. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. struct. Biol.* **7**, 29–40.

Shakhnovich, E. I. & Gutin, A. M. (1989). Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. *Biophys. Chem.* **34**, 187–199.

Shakhnovich, E. I. & Gutin, A. M. (1990). Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature* **346**, 773–775.

Shakhnovich, E. I. & Gutin, A. M. (1991). Influence of point mutations on protein structure: probability of a neutral mutation. *J. theor. Biol.* **149**, 537.

Shakhnovich, E. I. & Gutin, A. M. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. natn. Acad. Sci. USA* **90**, 7195–7199.

Shea, J. E., Nochomivitz, Y. D., Guo, Z. & Brooks, C. L. (1998). Exploring the space of protein folding hamiltonians: the balance of forces in a minimalist β-barrel model. *J. chem. Phys.* **109**, 2895–2903.

Shea, J. E., Onuchic, J. N. & Brooks, C. L. (1999). Exploring the origins of topological frustration: design of a minimally frustrated model of fragment b of protein a. *Proc. natn. Acad. Sci. USA* **96**, 12512–12517.

Sheinerman, F. B. & Brooks, C. L. (1998a). Calculations on folding of segment b1 of streptococcal protein G. *J. molec. Biol.* **278**, 439–455.

Sheinerman, F. B. & Brooks, C. L. (1998b). Molecular picture of folding of a small α/β protein. *Proc. natn. Acad. Sci. USA* **95**, 1562–1567.

Shoemaker, B. A., Portman, J. J. & Wolynes, P. G. (2000). Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. natn. Acad. Sci. USA* **97**, 8869–8873.

Shoemaker, B. A., Wang, J. & Wolynes, P. G. (1997). Structural correlations in protein folding funnels. *Proc. natn. Acad. Sci. USA* **94**, 777–782.

Shoemaker, B. A., Wang, J. & Wolynes, P. G. (1999). Exploring structures in protein folding funnels with free energy functinals: the transition state ensemble. *J. molec. Biol.* **287**, 675–694.

Smith, J. C. (1991). Protein dynamics: comparison of simulations with inelastic neutron scattering experiments. *Quart. Rev. Biophys.* **24**, 227–292.

Socci, N. D. & Onuchic, J. N. (1994). Folding kinetics of protein-like heteropolymers. *J. chem. Phys.* **101**, 1519–1528.

Socci, N. D., Onuchic, J. N. & Wolynes, P. G. (1996). Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. chem. Phys.* **104**, 5860–5868.

Sorenson, J. M. & Head-Gordon, T. (1998). The importance of hydration for the kinetics and thermodynamics of protein folding: simplified lattice models. *Folding & Design* **3**, 523–534.

Stein, D. L. (1985). A model of protein conformational sub-states. *Proc. natn. Acad. Sci. USA* **82**, 3670–3672.

Tanford, C. (1980). *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*, 2nd edn. New York: John Wiley & Sons.

Taylor, W. R. (2000). A deeply knotted protein structure and how it might fold. *Nature* **406**, 916–919.

Tianna, G., Broglia, R. A. & Shakhnovich, E. I. (2000). Hiking in the energy landscape in sequence space: a bumpy road to good folders. *Proteins* **39**, 244–251.

Tokita, K., Kikuchi, M. & Iba, Y. (2000). Dynamical equation approach to protein design. *Progr. theor. Phys.* (Suppl.). **138**, 378–383.

Ueda, Y., Taketomi, H. & Gō, N. (1975). Studies on protein folding, unfolding, and fluctuations by computer simulation. *Int. J. Peptide Res.* **7**, 445–459.

Veitshans, T., Klimov, D. & Thirumalai, D. (1997). Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Folding & Design* **2**, 1–22.

Vendruscolo, M., Najmanovich, R. & Domany, E. (1999). Protein folding in contact map space. *Phys. Rev. Lett.* **82**, 656–659.

Viguera, A. R. & Serrano, L. (1997). Loop length, intramolecular diffusion and protein folding. *Nature struct. Biol.* **4**, 939–946.

Viguera, A. R., Serrano, L. & Wilmanns, M. (1996). Different folding transition states may result in the same native structure. *Nature struct. Biol.* **3**, 874–880.

Villain, J. (1985). Equilibrium critical properties of random field systems – new conjectures. *J. Physique* **46**, 1843–1852.

Wales, D. J. & Scheraga, H. A. (1999). Global optimization of clusters, crystals, and biomolecules. *Science* **285**, 1368–1372.

Wang, J., Plotkin, S. S. & Wolynes, P. G. (1997). Configurational diffusion on a locally connected correlated energy landscape; application to finite, random heteropolymers. *J. Phys. I* (*France*) **7**, 395–421.

Watson, J. D. & Crick, F. H. C. (1953). Genetical implications of the structure of deoxyribonucleic acid. *Nature* **177**, 964.

Weeks, J. D., Gilmer, G. H. & Leamy, H. J. (1973). Structural transition in the ising-model interface. *Phys. Rev. Lett.* **31**, 549–551.

Wetlaufer, D. B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. natn. Acad. Sci. USA* **70**, 697–701.

Wigner, E. P. (1951). On the statistical distribution of the widths and spacings of nuclear resonance levels. *Proc. Camb. Phil. Soc.* **47**, 790–798.

Wolynes, P. G. (1992). Spin glass ideas and the protein folding problems. In *Spin Glasses and Biology* (ed. D. L. Stein), pp. 225–259. Singapore: World Scientific.

Wolynes, P. G. (1994). Three paradoxes of protein folding. In *Proceedings of Symposium on Distance-based Approaches to Protein Structure Determination II* (eds. H. Bohr, S. Brunak & J. Keiding), Copenhagen, Denmark: CRC Press.

Wolynes, P. G. (1996). Symmetry and the energy landscapes of biomolecules. *Proc. natn. Acad. Sci. USA* **93**, 14249–14255.

Wolynes, P. G. (1997a). As simple as can be? *Nature struct. Biol.* **4**, 871–874.

Wolynes, P. G. (1997b). Folding funnels and energy landscapes of larger proteins in the capillarity approximation. *Proc. natn. Acad. Sci. USA* **94**, 6170–6175.

Wright, P. E. & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. *J. molec. Biol.* **293**, 321–331.

Xia, X. Y. & Wolynes, P. G. (2000). Fragilities of liquids predicted from the random first order transition theory of glasses. *Proc. natn. Acad. Sci. USA* **97**, 2990–2994.

Yao, J., Chung, J., Eliezer, D., Wright, P. E. & Dyson, H. J. (2001). NMR structural and dynamic characterization of the acid-unfolded state of apomyoglobin provides insights into the early events in protein folding. *Biochemistry* **40**, 3561–3571.

Yue, K., Fiebig, K., Thomas, P. D., Chan, H. S., Shakhnovich, E. I. & Dill, K. A. (1995). A test of lattice protein folding algorithms. *Proc. natn. Acad. Sci. USA* **92**, 325–329.

Zou, J. M. & Saven, J. G. (2000). Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. *J. molec. Biol.* **296**, 281–294.

Zwanzig, R., Szabo, A. & Bagchi, B. (1992). Levinthal's paradox. *Proc. natn. Acad. Sci. USA* **89**, 20–22.