Computational Methods for Protein Folding: Advances in Chemical Physics, Volume 120.
Edited by Richard A. Friesner. Series Editors: I. Prigogine and Stuart A. Rice.

ISBNs: 0-471-20955-4 (Hardback); 0-471-22442-1 (Electronic)

INSIGHTS INTO SPECIFIC PROBLEMS IN PROTEIN FOLDING USING SIMPLE CONCEPTS

D. THIRUMALAI, D. K. KLIMOV, AND R. I. DIMA

Institute for Physical Science and Technology and Department of Chemistry and Biochemistry, University of Maryland, College Park, MD, U.S.A.

CONTENTS

I. INTRODUCTION

Protein folding is a process by which a polypeptide chain made up of a linear sequence of amino acids adopts a well-defined three-dimensional native structure [1]. Single-domain proteins reach their biologically active native conformations on time scales that are typically on the order of 10–1000 milliseconds [2]. Since Anfinsen's pioneering experiments it has been known that protein folding is a self-assembly process in which the information needed to determine the three-dimensional native structure is contained in the primary sequence [3]. Given this, the next important question is how the native state is kinetically reached in such a short time scale [4]. This issue was first emphasized by Levinthal, who wondered how a protein of a reasonable length can navigate the astronomically large conformational space so efficiently [5]. Seeking to resolve the paradox, Levinthal suggested that certain preferred pathways must guide the chain to the native state. For years the Levinthal paradox has served as an intellectual impetus in our quest to understand the mechanisms by which a polypeptide chain reaches the native conformation.

The last decade has witnessed considerable advances in our understanding of how a polypeptide chain folds starting from an ensemble of denatured states [6–13]. In recent years, protein folding kinetics has become increasingly important, largely because misfolding (i.e., errors in refolding) has been implicated in a number of diseases [14]. As a result, several advances have been made to probe the factors that govern the normal folding of proteins. Fast-folding experiments [2,8,15–19] and single-molecule methods [20–24] are beginning to provide direct glimpse into the early events in the assembly of proteins. Protein engineering in conjunction with the Φ-value analysis has become the cornerstone technique in deciphering the structures of the elusive transition state ensemble of two-state folders [25–27]. Although these tools have helped us to understand folding of individual proteins, considerable progress still needs to be made before the complex processes in misfolding and assembly of proteins with increasing complexity are well understood. In particular, to translate the functional genomics efforts into practical applications, it is important to solve rapidly the proteomics problem, namely, the determination of protein structures.

These multifaceted activities have ushered all aspects of protein folding at the center stage of molecular biology.

The major focus has been in understanding the folding mechanisms of proteins that display two-state behavior [28]. A variety of factors that determine the plausible folding scenarios have been identified [6,9–12,29–36]. A number of distinct folding mechanisms emerge depending on the characteristics temperatures that determine the phases of the polypeptide chain [10,34]. These findings explicitly link the underlying thermodynamic properties of proteins and their folding mechanisms. Several studies have focused on the factors that determine the folding rates of two-state proteins. Plausible relationships between folding rates and the contact order [37] (which emphasizes the role of structures involving proximal residues), stability [34,38], and Z score [34] have been established. Because many of these conceptual ideas have been described in recent reviews [6,9–12,33,39–41], we will not discuss these here.

A variety of computational and phenomenological approaches have been employed to obtain the general principles that control the folding rates and mechanisms of single-domain globular proteins [6,10,33]. It may be naively thought that the computational protocol for describing protein folding is straightforward. Indeed, the folding dynamics is well-described by the classical Newton equations of motion, and folding may be directly monitored from an appropriately long trajectory. However, there are two drastic limitations that prevent this approach to study the folding of proteins. First, the force fields for such a complex system are not precisely known. As a result, one needs to rely on the transferability hypothesis that interactions derived for small molecules can be used in larger systems, such as proteins. The second problem is simple: the limitations of current CPU power. Repeated folding of even a single-domain protein requires generating of multiple trajectories in a millisecond time scale. Even creative use of massively parallel simulations does not entirely solve this severe numerical constraint [42].[1]

In light of these difficulties, various simplified models of proteins have been suggested [10,39,41]. Most of the insights from computations came from the systematic studies of folding using coarse-grained models. The main rationale for their use is that a detailed study of such models will reveal general principles, if any, that govern the folding of proteins [10,39,41,43,44]. Such an approach

[1]We have recently achieved extraordinary speed-up of folding simulations for several β-hairpin sequences using distributed computing. In collaboration with Parabon Computations Inc., we have shown that distinct folding scenarios emerge even in the formation of β-hairpins. For the hairpin taken from the C-terminal of the immunoglobulin binding protein (GB1), the folding mechanisms and the time scales depend on the location of the hydrophobic cluster (D. Klimov, D. Newfield, and D. Thirumalai, unpublised results).

has yielded considerable insights into the mechanisms, time scales, and pathways in the folding of polypeptide chains.

The purpose of this chapter is to describe applications of simple concepts and computations to three specific problems in protein folding: (i) Are the requirements that folded states of proteins be compact and have low energy sufficient to explain the emergence of the finite number of folds from a very dense sequence space? An affirmative answer to this question, at the conceptual level, can be given using lattice models of proteins [44]. (ii) Phenomenological theory and lattice model computations are used to clarify the role of disulfide bonds in protein folding. The theory based on the proximity rule [45] and the lattice models investigating disulfide bonds formation [46] provided clarifications of the expected pathways in the refolding of bovine pancreatic trypsin inhibitor (BPTI). Recent calculations have explained quantitatively the effect of intact S–S bonds on the folding and stability of barnase. (iii) We describe a simple model of chaperonin-assisted folding [47]. Specific predictions about the coupling between conformational change of the chaperone molecule and the folding of the substrate protein emerge from the calculations. These predictions were subsequently tested experimentally.

To make this chapter as self-contained as possible, we briefly describe lattice models and the commonly employed computational methods. This is followed by a brief description of how a monomeric protein folds. The contents of this section are important to better appreciate the role of chaperones in the rescue of proteins. The chapter is concluded with brief comments about the challenges we face in the straightforward all-atom simulations of protein folding.

II. LATTICE REPRESENTATIONS OF PROTEINS

A. Basic Assumptions

Lattice models (LM) of single chains have long been used in polymer physics to obtain a number of universal properties (scaling of the size of the polymer with N, distribution of end-to-end distances, etc.) of real homopolymer chains [48]. For these issues the universal properties are unaffected by the precise interactions between monomers as long as they are short-ranged. It is not clear *a priori* that lattice models can be used to investigate general features of folding (e.g., cooperativity of transition from unfolded **U** to native **N** states). Single-domain proteins are finite-sized with the number of amino acid residues, N, not typically exceeding much beyond 200. Specific interactions that leads to the unique architecture of the **N** state cannot be fully represented using LM. The dynamics of the folding process can clearly depend on the precise move sets, so that the correspondence between the Monte Carlo simulations and the kinetics in

aqueous solution is ambiguous at best. Nevertheless, a series of studies from several groups have yielded a number of predictions many of which have been affirmed experimentally [6,39,41,47].

In the context of protein folding, lattice models were first introduced by Gō and co-workers [49]. The insights brought by Dill and Chan in the late 1980s have had a great influence on the development of LM for understanding protein-folding kinetics [50]. Dill and co-workers argued that protein folding can be studied using short enough chains so that exact enumeration of all allowed conformations becomes possible. Exact enumeration enables precise computations of thermodynamic characteristics. Monte Carlo (MC) simulations, based on physically motivated move sets, can be used to monitor folding kinetics.

In the simplest LM, amino acids are represented by a single atom (treated as a backbone α-carbon) and the side chains are not explicitly considered. As a result, only a few basic interactions found in real proteins can be modeled. In the most popular version of LM the polypeptide chain adopts a self-avoiding walk on a cubic lattice [32,51,52]. The heterogeneity of interactions in amino acids is mimicked by having several interaction energy scales between the beads of the chain. In general, only short-range interactions between nonbonded residues that are nearest neighbors on the cubic lattice are taken into account. Thus, a generic energy function for such a model includes three components: (i) connectivity of the chain is preserved through rigid bonding of successive beads; (ii) a self-avoidance condition is imposed by the restriction that a given lattice site can be occupied only once; (iii) the contact interactions between the side chain beads i and j $B_{ij}(|i-j| \geq 1)$ are given by pairwise potentials. The energy of a conformation is

$$E = \sum_{i<j} B_{ij} \delta_{|\vec{r}_i - \vec{r}_j|,a} \tag{1}$$

where $\delta_{r,a}$ is the Kronecker delta function and a ($=3.8$ Å) is the lattice spacing.

1. *Contact Energies*

There are several models for the interaction matrix elements B_{ij} which take into account the diversity of interactions between amino acids. Because these models are at best a simple representation of the potentials in real proteins, it is not *a priori* clear that any particular model is better than the other. In the literature several different interaction schemes have been utilized [32,34,39,47,51,52]. These include HP model [39,51], random bond (RB) model [32], and the pairwise potentials derived from the statistical analysis of contacts between different amino acids in the protein structures [53–55]. In what follows we give a brief description of these models.

2. HP Model

This model reduces the set of 20 naturally occurring amino acids to two kinds, namely, hydrophobic (H) and polar (P) [39,51]. A sequence is given by the nature of the amino acid residue at a given position. For example, HPHPH is a sequence with $N = 5$. There are 2^N total sequences for a given N. In the HP model the interactions are given by a 2×2 matrix, whose elements are $B_{HH} = -\epsilon$ and zeros, otherwise. Despite the simplicity of the model, it is not exactly solvable due to the chain connectivity and excluded volume effects. Because the HP model can lead to microphase separation, variations in the interaction energies have been introduced. Various aspects of folding observed in the HP model (two-letter code) have been investigated by Dill et al. [39] and others [51].

3. Random Bond Model

In the RB model [32] the interaction elements are drawn from the Gaussian distribution

$$P(B_{ij}) = \frac{1}{\sqrt{2\pi}B} \exp\left(- \frac{(B_{ij} - B_0)^2}{2B^2} \right) \tag{2}$$

where B_0 is the average interaction that specifies the strength of the drive toward forming compact structures at low temperatures, and the dispersion B gives the extent of diversity of the interactions among beads. Energy is measured in terms of B which is set to unity. The choice of $B_0 = -0.1$ [32] ensures that the fraction of hydrophobic residues in a sequence (specified by the interaction matrix elements B_{ij}) is about 0.55, which roughly coincides with the fraction of hydrophobic residues in real proteins. A sequence is specified by the matrix of contact energies B_{ij}.

4. Statistically Derived Pairwise Potentials

In this case, the energies B_{ij} are given by pairwise statistical potentials computed by analyzing the frequency of amino acids interactions in the experimentally determined protein structures. Several sets of such potentials are currently available. These includes the potentials calculated by Miyazawa and Jernigan (MJ) [53], Kolinski, Godzik, and Skolnick (KGS) [54], Mirny and Shakhnovich [56], Tobi and Elber [57], and Betancourt and Thirumalai [55]. The major advantage of the such potential sets is that the model lattice sequence may now be described in terms of ''real'' amino acid composition, assuming that the contact energies reproduce the nature of interactions between amino acids.

5. Gō Model

The Gō model does not directly introduce a new force field, but modifies the existing energy function by tuning it to the known native structure [58]. Specifically, the Gō model considers only the interactions between residues

(beads on the lattice) that are present in the native (ground) state. In other words, only native contacts are taken into account. The major advantage of the Gō model is almost a complete elimination of frustration in a model protein and, as a result, a substantial increase in the folding rates [59]. The severe shortcoming is that the energy function and the native structure cannot be decoupled; consequently the Gō model, despite being topologically frustrated, is "foldable" by definition.

B. Lattice Models with Side Chains

The cubic lattice models described above is the simplest version of the coarse-grained model. One obvious way to make it more realistic is to incorporate the explicit representation of side chains [60]. In this case, a polypeptide chain is modeled by a sequence of N backbone beads, representing the C_α carbons of a protein backbone. Side-chain beads, which mimic amino acid residues, are attached to each backbone bead. In all, there are $2N$ beads in the model, all of which occupy the vertices of cubic lattice. The conformation of a protein is specified by $2N$ vectors $\vec{r}_{b,i}, \vec{r}_{s,i}, i = 1, 2, \ldots, N = 15$, where $\vec{r}_{b,i}$ and $\vec{r}_{s,i}$ are the positions of backbone and side-chain beads, respectively. The energy function used for the side-chain model is typically the same as employed in the model without side chains. These models provide a more realistic description of cooperativity of folding, because they include effects of side-chain packing [39].

C. Computational Methods

1. *Exhaustive Enumeration*

The conformational space of short lattice sequences can be exhaustively enumerated. All conformations for a polypeptide chains with $N \lesssim 20$ on a cubic lattice can be enumerated using the Martin algorithm [61]. This algorithm successively generates all self-avoiding conformations for a given N, which allows *exact* calculation of any thermodynamic quantity. In order to reduce the sixfold symmetry on the cubic lattice, the direction of the first monomeric bond may be fixed in all conformations. The remaining conformations are still related by the eightfold symmetry on the cubic lattice (excluding the cases when conformations are completely confined to a plane or straight line). To decrease further the number of conformations, the Martin algorithm may be modified to reject all conformations related by this symmetry [32]. For longer model sequences the CPU time required to enumerate all conformations becomes prohibitively long. With constant upgrade in computer power this limitation is being steadily overcome.

2. *Monte Carlo Method*

The standard method for studying thermodynamics and kinetics of folding in the context of lattice models is the Monte Carlo (MC) algorithm [62]. Several types

of moves are commonly used [32]. These are (i) corner moves (a flip of the residue across the diagonal of the square formed by the neighboring bonds), (ii) crankshaft rotations (rotation of the beads $i+1$ and $i+2$, while keeping the adjacent beads i and $i+3$ fixed), and (iii) rotation of the end beads. Although the precise choice of moves or their probabilities affects the local structural dynamics, it is commonly believed that the general thermodynamic properties and even kinetic characteristics remain unchanged as long as the moves are ergodic. Even with the choice of physically motivated move sets their influence on the results must be tested.

3. *Multiple Histogram Technique*

The thermodynamic quantities for longer chains may be effectively computed using the multiple histogram method [51,60,63]. The method is based on the collection of a set of histograms at different values of the external parameter and combining them by reweightening the contribution from individual histograms. The thermal average of any quantity may then be calculated. Technically, multiple slow-cooling MC trajectories, each starting from different conditions, are needed to obtained the histograms. Each trajectory starts at a high temperature ($T_h > T_\theta$) and ends at the temperature $T_l < T_F$, where T_θ and T_F are the collapse and folding temperatures, respectively. In the course of a trajectory the temperature is changed periodically by small decrements, and the portions of simulations at a given fixed temperature (after quick equilibration intervals) are used for histogram collection. Usually, histograms for the values of energy, number of native contacts, radius of gyration, and so on, are obtained. There is no general prescription for choosing the lengths of the trajectory and of the equilibration interval because they depend strongly on the sequence and on the temperature. The number of trajectories is determined by the condition that the thermodynamics of the system should not change significantly with subsequent increase in sampling. Thus, by using multiple histogram technique, one can completely characterize the thermodynamics of the system by calculating the average of any quantity as a function of external parameter as well as the free energy profiles. Using the histograms, we can generate free energy profiles, provided that a useful reaction coordinate is chosen.

4. *Folding Kinetics*

The kinetics of folding of a lattice sequence is obtained using multiple folding trajectories at a fixed temperature. Each trajectory starts from a different high-temperature conformation. After a sudden quench of the temperature to T_s, the chain kinetics is monitored. Typically, the folding kinetics is characterized by time dependence of folding probes averaged over the total number of trajectories considered. The first passage to the native structure τ_{1i} is also recorded. From the distribution of τ_{1i} P_{fp}, the fraction of trajectories that have not reached the native

conformation at a time t is calculated using

$$P_u(t) = 1 - \int_0^t P_{fp}(s)\, ds \tag{3}$$

The integral of $P_u(t)$ determines the average passage time τ_F as

$$\tau_F = \int_0^\infty P_u(t)\, dt \tag{4}$$

Accurate results require generation of hundreds of folding events.

III. REDUCTION IN CONFORMATIONAL SPACE

A. Importance of Excluded Volume Interactions

The impetus to examine the size of the conformational space of proteins comes from Levinthal [5], who wondered how can a polypeptide chain, even though it is relatively small, navigate the vast number of allowed conformations in search of the unique native state? A popular resolution of this argument suggests that fundamental constraints, notably the excluded volume (EV) interactions between atoms, so vastly reduces the conformations that only a very limited number is ever sampled. This idea can be precisely tested using appropriate models.

The number of independent conformations for a chain with N beads on a cubic lattice is $C_{IND} = Z^N$, where $Z\,(=6)$ is the lattice coordination number. If excluded volume interactions (also referred to as steric clashes [64]) are taken into account, then the number of allowed conformations is

$$C_{EV} \simeq Z_{eff}^N N^{\gamma-1} \tag{5}$$

where the universal exponent $\gamma \approx 1.16$, and $Z_{eff} = 4.684$ in a cubic lattice. Both C_{IND} and C_{EV} scale exponentially with N. However, it might be argued that the finite size of the proteins might make the reduction, due to EV interactions, so significant that the ''entropy price'' to adopt native-like conformations is not very large. In a cubic lattice the entropy change, ΔS, upon going from S_{IND} to S_{EV} is $\Delta S/k_B \approx N \ln(Z/Z_{eff})$. For $N = 10, \Delta S \approx 12.8$ eu, which is substantial. However, the absolute entropy associated with S_{EV} is $N \ln Z_{eff}$. Neglecting logarithmic corrections we get $S_{EV} = k_B \ln C_{EV} \approx 15.4$ eu. Thus, considering steric clashes alone *does substantially reduce the size of the conformational space*. However, this reduction is not sufficiently large to solve the ''search problem'' envisioned by Levinthal.

In a recent interesting article, Pappu et al. [64] have reemphasized the importance of excluded volume interactions by enumerating the allowed conformations for blocked all-atom polyalanine chains, Al–$(Ala)_n$–N′-methylamide

for $n \leq 7$. By coarse graining the (ϕ, ψ) angles, they showed that the conformational space due to EV interactions is less than it would be if the (ϕ, ψ) angles are considered independent as suggested by Flory. This result is in qualitative accord with the estimates for lattice models given above. As pointed out above, this reduction is not sufficient to provide a qualitative explanation of the central kinetic issue raised by Levinthal.

Pappu et al. [64] suggest that EV interactions or steric clashes "bias" the conformations so that even in the unfolded state there is a significant tendency to form local structures. This is certainly the case in off-lattice models of proteins [10]. Typically, these fluctuating structures are stabilized by additional interactions (say, hydrogen bonding). If the favorable biasing interactions are too strong (greater than 2–3 $k_B T$), then the local interactions would become incompatible with the tertiary interactions. This has been shown to increase the topological frustration [65] see below, which in turn can lead to the dominance of kinetic traps. Thus, arguments that are based solely on the reduction of conformational space of proteins cannot account for the global folding mechanisms. Harmony (or consistency) between local and nonlocal interactions is necessary for efficient folding of proteins.

If only EV interactions are included in polypeptide chains, the chain cannot undergo a "phase transition" to any specific conformation. The effective mean-field one-body potential describing EV interactions is known to be long-ranged (scaling as $r^{-4/3}$). Consequently the polypeptide chain would adopt a random coil state at all temperatures, if only EV interactions are included. However, the chain can be induced to adopt a preferred structure (native conformation), if an additional attractive energy $-\epsilon$ between residues (hydrogen bond interactions, for example) is introduced. This is the basis of the popular HP model for proteins [39]. In a model, which takes into account the EV and attractive interactions, a phase transition into a native-like structure can occur at T such that $T \approx C_N \epsilon / S_U$, where C_N is the number of favorable native interactions and S_U is the entropy of the unfolded state. Pappu et al. [64] showed that by including an attractive energy term to mimic backbone hydrogen bonding, an apparent two-state transition from a stretched state to a contracted state takes place (Fig. 1). This kind of apparent two-state transitions, similar to those found in proteins, has been observed in simple lattice models as well [6]. The interesting feature of the calculations by Pappu et al. [64] is that a realistic model of even a short polypeptide chain with only one attractive energy scale can exhibit protein-like behavior.

IV. EMERGENCE OF STRUCTURES FROM THE DENSE SEQUENCE SPACE

The sequence space of proteins is extremely dense as the number of possible sequences for proteins of length N scales as 20^N. However, not all these

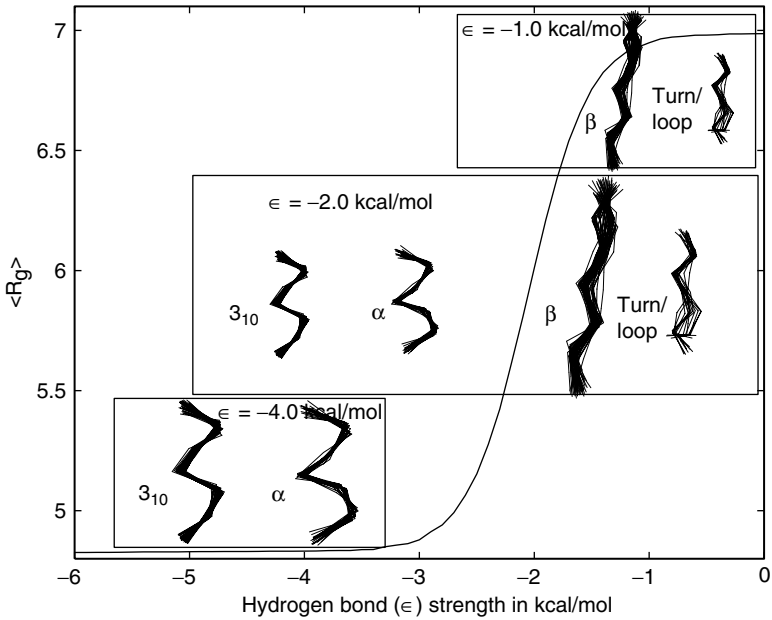


Figure 1. Dependence of the radius of gyration $\langle R_g \rangle$ for polyalanyl chain of length $n = 7$ [64] on the hydrogen bond length ϵ. As ϵ increases, compact conformations are populated preferentially. The transition from the extended conformations at higher values of ϵ to the contracted conformations occurs rather cooperatively in an apparent "two-state" manner. The radius of gyration is computed using a coarse-grained thermally weighted density of states (see Ref. 64 for details). Conformations that make the most significant contributions at different values of ϵ are also shown.

sequences encode for foldable protein structures, which for functional purposes are constrained to have specific physical characteristics. How do viable protein structures emerge from the dense sea of sequence space [66]? The extraordinary thinning of the sequence space as one gets to the structure space may be understood purely on the basis of accepted physical properties of proteins. To this end, two interrelated physical features of folded proteins must be taken into account. (i) Native proteins are compact. (ii) The interior of proteins consists mainly of hydrophobic residues, while the hydrophilic residues are typically found on the surface. This gives rise to a maximum number of favorable interactions making the native state very low in energy.

Lattice models are remarkably useful in answering the conceptual question posed above. To infer the sequence to structure mapping, we performed an exhaustive enumeration of all self-avoiding conformations for the sequences confined to cubic lattice with $N = 15$ [44]. The RB model has been used in the energy function with the parameters $B_0 = -0.1$ and $B = 1$. Protein-like structures are not only compact but also have low energy. We first computed the

number of compact structures (CSs) for a given N, C_N, (CS). The number of CSs, in its most general form, is expected to scale as

$$C_N(\mathrm{CS}) \simeq \bar{Z}^N Z_1^{N\frac{d-1}{d}} N^{\gamma_c - 1} \tag{6}$$

where $\ln \bar{Z}$ is the conformational free energy (in units of $k_B T$), Z_1 is the surface fugacity, d is the spatial dimension, and γ_c measures possible logarithmic corrections to the free energy. The number of natural protein folds is limited (perhaps a few thousands), and their number is expected to grow at rates much smaller than those predicted in Eq. (2). To explore this we calculated by exact enumeration the number of minimum energy structures (MES), C_N(MES), as a function of N.

We define MES as those conformations whose energies lie within the energy interval Δ above the lowest energy E_0, corresponding to the native state. Several values for Δ (1.2 or 0.6) were used to ensure that no qualitative changes in the results are observed. We also tested another definition for $\Delta = 1.3|E_0 - tB_0|/N$, where t is the number of nearest-neighbor contacts in the ground state. It is worth noting that in the latter case Δ increases with N. Nevertheless, both definitions yield equivalent results. The computational technique involves exhaustive enumeration of all self-avoiding conformations for $N \leq 15$ on a cubic lattice. We calculated the energies of all conformations according to Eq. (1) and then determined the number of MES and CS. Each quantity, such as C_N(MES), C_N(CS), the lowest energy E_0, or the number of nearest-neighbor contacts t in the lowest energy structures, is averaged over 30 sequences. To test the reliability of the computational results, an additional sample of 30 RB sequences was generated. Note that in the case of C(MES) we computed the quenched average as $C_N(\mathrm{MES}) = \exp\left[\overline{\ln [c(\mathrm{MES})]}\right]$, where c is the number of MES for one sequence.

The number of MES C(MES) is plotted as a function of the number of residues N in Fig. 2 for $\Delta = 0.6$. A pair of squares for a given N represents C(MES) computed for two independent runs of 30 sequences each. For comparison, the number of self-avoiding walks C(SAW) and the number of CS C(CS) are also plotted in this figure (diamonds and triangles, respectively). As expected on general theoretical grounds, C(SAW) and C(CS) grow exponentially with N, whereas the number of MES C(MES) exhibits drastically different scaling behavior. There is no variation in C_N(MES) (normally, associated with the variation of shapes of compact structures) and its value remains steady within the entire interval of N starting with $N = 7$. We find (see Fig. 2) that $C_N(\mathrm{MES}) \approx 10^1$. This result further validates our earlier finding for the two-dimensional model [46]. The results strongly suggest that C_N(MES) scales only as $\ln N$. Thus, the dual restriction of compactness and low energy of the native states may impose an upper bound on the number of distinct protein folds.

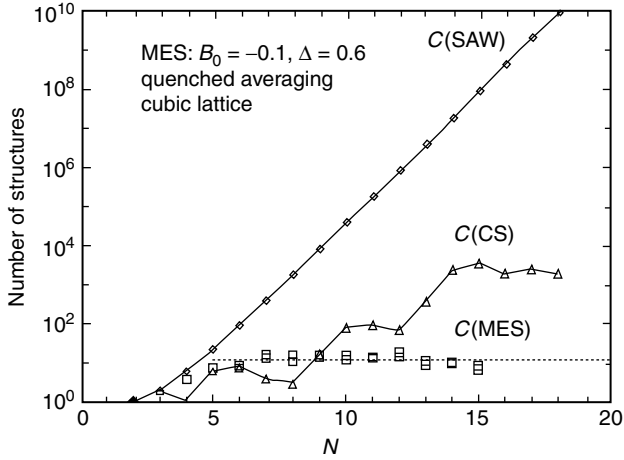
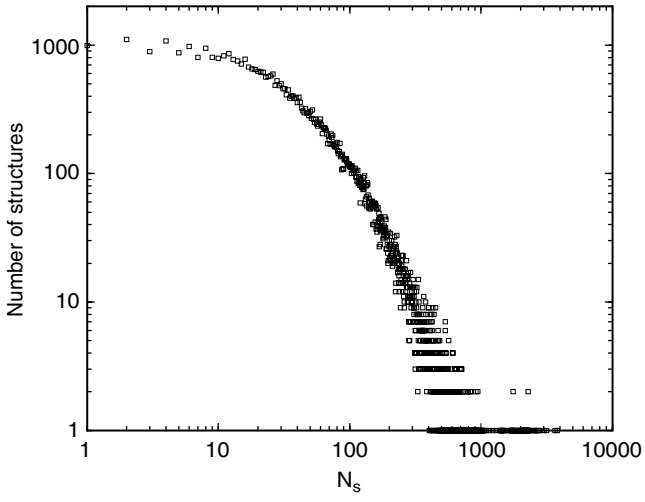


Figure 2. Scaling of the number of MES C_N(MES) (squares) on a cubic lattice. The data are obtained for $B_0 = 0.1$ and $\Delta = 0.6$. The pairs of squares for each N represent the quenched averages for different samples of 30 RB sequences. The number of compact structures C_N(CS) and self-avoiding conformations C_N(SAW) are plotted to highlight the dramatic difference in scaling behavior. It is clear that C(MES) remains practically flat; that is, it grows no faster than ln N.

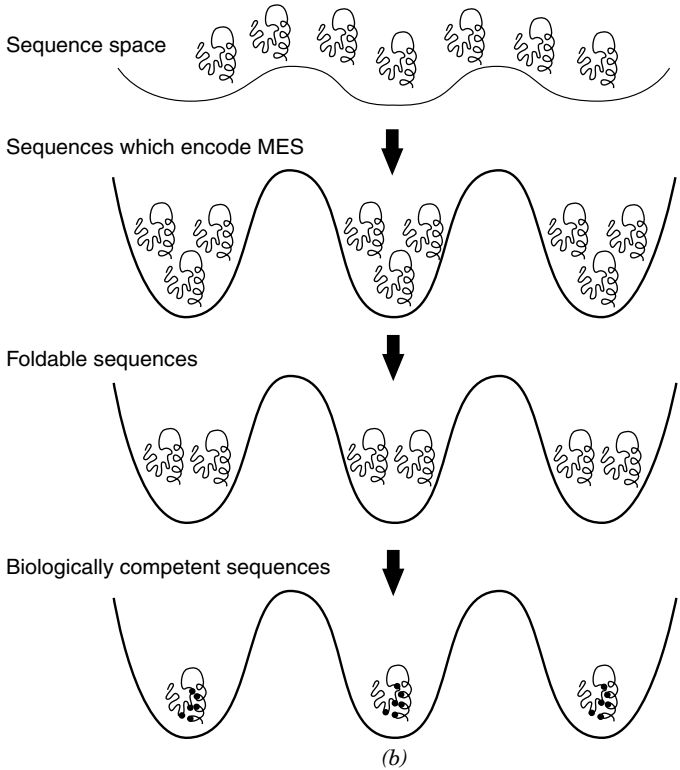
A. Designability of Protein Folds

The computations described above indicate that minimal restrictions on the structures (compactness and low energies) make the structure space sparse. Consequently, each basin of attraction in the structure space must contain numerous sequences [66]. The way these sequences are distributed among the very slowly growing number (with respect to N) of conformations—that is, the density of sequences in structure space—is another important question. Li et al. [67] considered a three dimensional cubic lattice proteins with $N = 27$. By using the HP model and restricting themselves to only maximally compact structures as tentative candidates for protein native states, they showed that certain folds (i.e., structures) accommodate much larger number of sequences (see Fig. 3a) than the others. In one example, they found the NBA (the structure) that serves as a ground state for 3794(!) (when the total number is 2^{27}) sequences and, hence, was considered most designable. The precise distribution of sequences among NBAs is a function of the particular energy function.

An important conclusion of Li et al. [67] is that one can define, at least operationally, a designability index for every fold found in PDB. The structural characteristics of a given fold determine its designability. Several authors have suggested that if the fold has even an approximate symmetry, then it would be more designable [67,68]. This might explain the preponderance of TIM barrel


Number of structures
1000
100
10
1
1
10
100
1000
10000
N_s

(a)


Sequence space
Sequences which encode MES
Foldable sequences
Biologically competent sequences

(b)

structures. If the symmetry argument is extended to RNAs, then we would conclude that certain symmetries should be hidden at the sequence level of mRNAs and ultimately the genes themselves encoding a given protein [44].

Because the number of NBA for the entire sequence space is very small, it is likely that proteins could have evolved randomly. Natural folds must correspond to one of the native basins of attraction in the structure space so that many sequences have these folds as the native conformations. In other words, natural protein folds, especially those with approximate symmetries, represent highly designable structures [67]. Further support for these ideas comes from the study of Lindgard and Bohr [69]. These authors showed that among maximally compact structures there are only very few folds that have protein-like characteristics. It was also estimated that the number of distinct protein folds is on the order of 10^3. Thus, each fold can be designed by many candidate sequences. However, there is also evolutionary pressure for sufficiently rapid folding to avoid aggregation. This kinetic requirement further restricts the possible sequences that can serve as biologically viable proteins (Fig. 3b).

V. PROTEIN FOLDING MECHANISM

Using lattice models with side chains we describe the most commonly found scenarios observed in protein folding. Because this topic has been subject of numerous reviews [6,9–12,41], we will stress a few points that are relevant in considering chaperonin-mediated protein folding that is discussed in Section VII.

A. Two-State Folders

Thermodynamics for the sequence with the native state shown in Fig. 4 with the contact interaction potentials B_{ij} taken from Table III of Ref. 54 reveals that it folds cooperatively in an apparent two-state manner. This is also reflected in the thermal distribution of the overlap function values $h(\chi)$ at the folding transition temperature T_F (Fig. 4). A nearly bimodal distribution of $h(\chi)$ with the peaks at $\chi \lesssim 0.2$ (NBA) and $\chi \sim 0.6$ (unfolded state) is observed. There is also nonnegligible contribution from the intermediate values of χ representing partially folded structures. Experiments that probe in more detail the thermal unfolding of proteins are beginning to reveal the possible importance of these

Figure 3. (a) A log–log plot of the histogram for number of structures with respect to the number of associated sequences N_s for 27-mer maximally compact cubic lattice conformations [67]. The plot illustrates a dramatic heterogeneity among structures in terms of their ability to encode protein sequences. (b) Schematic illustration of the mapping of vast sequence space onto the limited number of protein folds. This mapping involves drastic reduction in sequence space as polypeptide sequences evolve into functionally competent proteins.

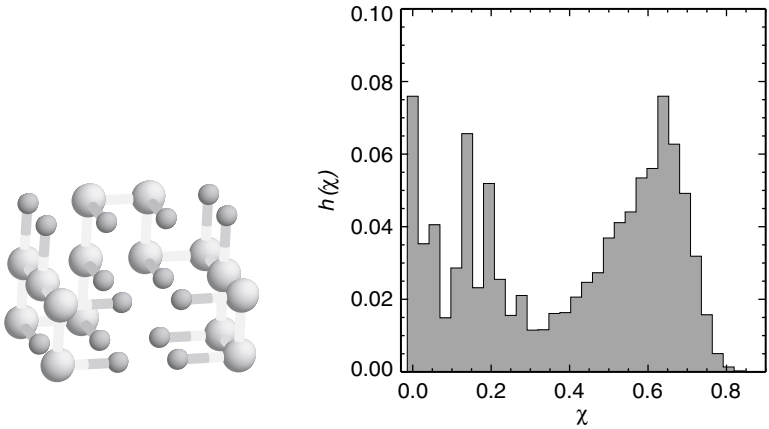


Figure 4. Native structures of sequences A generated using [126] is shown in the left panel. Backbone and side-chain beads are shown in light and dark gray, respectively. Native conformation is compact and has a well-defined hydrophobic core. The figure is generated using program RasMol [126]. The right panel displays the thermal distribution of states $h(\chi)$ calculated at $T \approx T_F$ for sequence A. $h(\chi)$ is approximately bimodal so that only NBA ($\chi \lesssim 0.2$) and unfolded state $\mathbf{U}(\chi \sim 0.6)$ are significantly populated. Although small, the population of intermediate states nevertheless makes a sizable contribution to thermodynamics (affecting mainly cooperativity of folding).

conformations [70]. Due to substantial contribution from the partially folded structures, thermal unfolding cannot be quantitatively described as two-state.

The folding kinetics can be probed using the distribution of the first passage times, τ_{1i}. Several hundred ($\sim$600) folding events are used to obtain the distribution of τ_{1i}, from which the fraction of unfolded molecules $P_u(t)$ may be readily obtained. In addition to $P_u(t)$, we have computed the time dependence of the radius of gyration $\langle R_g(t)\rangle$, where the average is taken over 100 folding trajectories.

The sequence, whose native state is shown in Fig. 4, displays two-state kinetics for the temperatures $T \geq 0.8T_F$; that is, $P_u(t) \sim \exp(-\frac{t}{\tau_F})$, where τ_F is the folding time. To probe the sequence of events en route to the native conformation, we computed $\langle R_g(t)\rangle$, which reveals two stages in collapse. Initial rapid burst phase is followed by a gradual chain compaction (Fig. 5). The overall collapse time τ_c is associated with the second characteristic time. From the approach to the native conformation we draw the following general conclusions regarding two-state folders:

(a) The ratio τ_F/τ_c for two-state folders is typically less than 10. This is consistent with the fast-folding experiments on several two-state folders, which

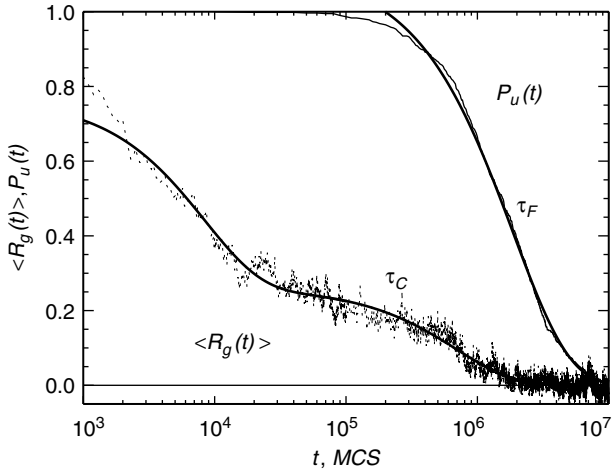


Figure 5. The time dependence of the normalized radius of gyration $\langle R_g(t)\rangle$ and the fraction of unfolded molecules $P_u(t)$ for sequence A at $T = 0.94T_F$. Data are averaged over 100 [for $\langle R_g(t)\rangle$] and 600 [for $P_u(t)$] trajectories. $P_u(t)$ decays exponentially with the time scale $\tau_F = 2.07\times 10^6$ MCS. The approach of $\langle R_g(t)\rangle$ to equilibrium is biexponential with the times scales 0.083×10^6 MCS and 0.698×10^6 MCS. The first time scale is due to extremely rapid burst-phase partial collapse. The second time scale, which is associated with the collapse time τ_c, corresponds to the final compaction. The ratio τ_F/τ_c is approximately 3.0.

show that proteins rapidly collapse and reconfigure themselves to reach the native state. For the sequence in Fig. 4, $\tau_F/\tau_c \approx 3$. This ratio is in the range 5–10 for proteins.

(b) Analysis of the collapsed conformations shows that they are native-like; that is, the initial collapse in two-state folders is ''specific'' with very few nonnative interactions present. The overall scheme for reaching the NBA for two-state folders, which was predicted using theoretical arguments, is

$$\mathbf{U} \rightarrow \{\mathbf{I_N}\} \rightarrow \mathbf{N} \tag{7}$$

where $\{\mathbf{I_N}\}$ is a collection of native-like structures. Fast-folding experiments on cyt-c and tendamistat [71] have been interpreted using this picture. Because the initial collapse is specific, the ensemble of native-like intermediates can be likened to an ''on-pathway'' intermediate. Lattice simulations (without side chains) using Gō model have come to a similar conclusion [72]. In the Gō model the only possible nonnative ''interaction'' comes from the topological entanglements, which are highly unlikely given the relatively small (48-mer) well-designed sequence.

B. Moderate Folders, Topological Frustration, and Kinetic Partitioning Mechanism

Many qualitative aspects of the folding kinetics of moderate folders can be understood in terms of the concept of topological frustration [10]. On average, about 55% of residues in proteins are hydrophobic, and their density along the sequence is roughly constant. As a result, on any local length scale there is a propensity for the hydrophobic residues to form tertiary contacts (structures) under folding conditions; that is, proximal residues adopt preferred structures. The assembly of the resulting structures would most likely be in conflict with the global native fold. The incompatibility of the low free-energy structures on local and global scales leads to a phenomenon called topological frustration. Topological frustration is an *intrinsic property of all foldable sequences* and arises due to the polymeric nature of proteins and the heterogeneity of amino acids. It follows that even the Gō model is topologically frustrated because residue connectivity can render certain favorable local structure incompatible with the global fold. An important physical outcome of topological frustration is that the free-energy folding landscape is rough, consisting of many minima that are separated by barriers of varying heights.

One of the principal consequences of topological frustration is that the folding kinetics follows the kinetic partitioning mechanism (KPM) [10]. Imagine an ensemble of unfolded molecules in search of the native conformation (Fig. 6). Due to the heterogeneity of folding pathways, a fraction of molecules, Φ, would reach the NBA (or **N**) rapidly without being kinetically trapped in the low-lying free-energy competing basins of attraction (CBA). The remaining fraction,

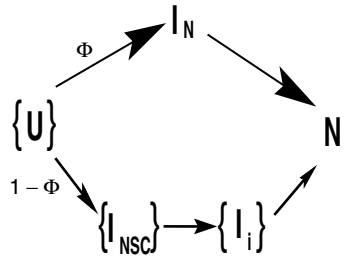


Figure 6. The sketch of the protein folding pathways. The fast (upper) folding pathway includes the formation of native-like collapsed states $\{\mathbf{I_N}\}$, which rapidly convert into the native state **N**. The fraction of protein molecules, folding along this pathway, is Φ. For two-state folders, $\Phi \approx 1$. The lower track (followed by $1 - \Phi$ molecules) represents slow pathway(s), which fold by a three-stage kinetic mechanism. At the first stage, nonspecific collapse species $\mathbf{I_{NSC}}$ form, which later convert into a collection of discrete native-like intermediates $\{\mathbf{I_i}\}$. The transition from $\{\mathbf{I_i}\}$ to the native state is slow and represents the rate-limiting step in the slow pathway. The degree of heterogeneity in the folding pathways depends on the sequence and external conditions.

$(1-\Phi)$, would be trapped, and only on longer time scales would thermal fluctuations enable the chain to reach the NBA through an activated process. The value of the partition factor Φ depends on the sequence and external conditions. Thus, topological frustration leads to a separation of the initial ensemble of denatured molecules into fast- and slow-folding phases (Fig. 6). For two-state folders, which have a funnel-like free energy landscape, $\Phi = 1$.

According to the KPM $P_u(t)$ [see Eq. (3)] is given by

$$P_u(t) = \Phi \exp\left(-\frac{t}{\tau_N}\right) + \sum_k a_k \exp\left(-\frac{t}{\tau_k}\right) \tag{8}$$

where τ_N is the time scale for reaching the native state by the fast (direct) process (presumably by the nucleation-collapse), and τ_k is the time scale for indirect folding pathways, in which the native state is reached after escaping a local free-energy minimum (trap) k. Prefactors a_k are related to the ''volumes'' associated with the k^{th} CBA. Thus, folding trajectories can be divided into those that reach the native conformation rapidly (their fraction or partition factor is Φ) and those that follow indirect off-pathway routes (Fig. 6).

The validity of the KPM has been demonstrated in several protein-like models beginning with the studies of Guo and Thirumalai [73]. More importantly, refolding experiments, on lysozyme [74] and large ribozymes [75] have confirmed the KPM. Using interrupted folding experiments, Kiefhaber [74] was the first to show that $\Phi \approx 0.15$ in lysozyme. Subsequent studies of lysozyme by Dobson and coworkers [76] show that $\Phi \approx 0.25$ in lysozyme. The difference is presumably due to changes in folding conditions. Perhaps the most direct demonstration of the validity of the KPM comes from the single-molecule FRET measurements on the L-21 Sca I ribozyme [77]. The results of these experiments analyzed by us showed that $\Phi \approx 0.06$, which is consistent with the estimates from ensemble measurements. These experiments show that KPM offers an unified picture of folding for a class of proteins and RNA [78].

VI. DISULFIDE BONDS IN FOLDING

A. Refolding of BPTI

Bovine pancreatic trypsin inhibitor (BPTI), a small protein with 56 amino acid residues (Fig. 7), is the first one for which a detailed map of the refolding pathways was deciphered. The native state of BPTI contains three disulfide (S–S) bonds formed between six Cys residues. Native state is specified by [30–51; 5–55;14–38] bonds. This notation indicates that Cys^{30} forms an S–S bond with Cys^{51}, and so on. Reduction of the S–S bonds unfolds BPTI. By using S–S bond formation as a ''progress variable,'' Creighton [79–83] devised ingenious methods to trap the disulfide-bonded intermediates along the folding pathway.

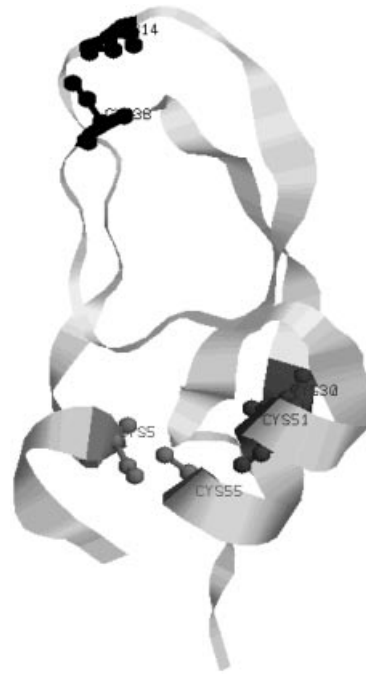


Figure 7. See also color insert. The native-state conformation of the bovine pancreatic trypsin inhibitor (BPTI). The figure was produced with the program RasMol 2.7.1 [126] from the PDB entry 1bpi. There are three disulfide bonds in this protein: Cys5–Cys55 shown in red, Cys14–Cys38 shown in black, and Cys30–Cys51 shown in blue. The corresponding Cys residues are in the ball-and-stick representation and are labeled. The two helices (residues 2–7 and 47–56) are shown in green.

The refolding pathways were described in terms of the nature of the intermediates that accumulate during folding. There are 75 distinct intermediates containing one or more disulfide bonds that can be formed from six Cys residues. On the time scale of the experiments, Creighton discovered that only eight intermediates could be detected. These experiments were among the earliest to show that in the folding reaction only a small number of partially folded intermediates accumulates.

The most surprising discovery made by Creighton [79–83] was that in the refolding of BPTI, three non-native states—namely, the intermediates with disulfide bonds not present in the native state—are well-populated. More importantly, two of the non-native species, [30–51;5–14] and [30–51;5–38], are involved in the productive pathway; that is, folding proceeds through either of these two kinetically equivalent intermediates. The detection method employed by Creighton involves quenching the folding reaction using chemistry to stop the reaction. To isolate only the intermediate that would naturally occur in the refolding process, the quench rate must exceed rates of formation of other products. The chemistry of the quench method determines the time required to stop the reaction from progressing. Creighton's findings were challenged by

Weissman and Kim (WK) [84–87], who used pH changes (acid quenching) to disrupt the folding reaction. The most glaring difference between the two series of studies is that WK showed that, in the productive pathway, *only native intermediates* play a significant role. Non-native intermediates may only be involved as required by disulfide chemistry in the last stages of folding of BPTI; that is, they play a role in the formation of the precursor [30–51;5–55] from [30–51;14–38] (denoted by N_{sh}^{sh} and N', respectively).

In an attempt to resolve the apparent controversy between the findings of Creighton and WK, we introduced a phenomenological theory, referred to as the *proximity rule* [45], to predict the folding pathways in globular proteins. Our theory accounts for entropic effects analytically and energetic effects only approximately. The premise of the proximity rule is that local events, governed mainly by entropic considerations, dictate the initial events in protein folding. The importance of local events is the basis of the hierarchic mechanism of folding [11,12] and is also emphasized in the notion that contact order [37] is the primary determinant of folding rates of proteins. Just as in the applications of proximity rule, we expect that theories that rely largely on local events *can only* account for the early processes in folding. However, such theories often ''work'' in regimes for which they are not, in principle, applicable.

B. Proximity Rule

The major conformational changes in disulfide bonded proteins, such as BPTI and ribonuclease A [88], can be understood in terms of disulfide bond rearrangement. Thus, the conformations of the intermediates that determine the folding pathways are specified in terms of the S–S bonds. In such proteins the S–S bonds serve as a surrogate ''reaction coordinate.'' These observations enable us to develop the proximity rule based on the following general principles.

1. *Loop Formation Probability*

We assume that the initial intramolecular disulfide bond rearrangement is a random process governed largely by entropic considerations. The probability of forming a disulfide bond under oxidizing conditions depends only on the loop length $l = |i - j|$, where i and j are the positions of the Cys residues along the polypeptide chain. The probability of simultaneous loop formation of lengths l_1 and l_2, $P(l_1, l_2)$, is assumed to be proportional to $P(l_1)P(l_2)$. The absence of correlation limits the theory to the prediction of only the earliest events in BPTI refolding. Similarly, theories that are based on local propensities alone can only describe the formation of secondary structures and initial tertiary structures in the folding of globular proteins. Despite this limitation, the utility of the proximity rule to predict the refolding pathways of BPTI was extended using parameters determined from experiments [45]. The loop formation probability P(l) may be computed by modeling the polypeptide chain as a semiflexible chain.

2. *Folding Kinetics*

For slow-folding proteins, which require reconfiguration of partially folded structures, folding follows a three-stage kinetics [45]. These stages are as follows: (i) There is a rapid collapse of the chain to a set of compact conformations. At this stage, most of the free energy arises from a competition between hydrophobic forces and loop entropy. In BPTI this stage is characterized by the need to have proper loop contacts between Cys residues, so that a single S–S bond can form. At the end of this stage the most stable single disulfide species accumulate. (ii) The rearrangement of the single disulfide bonds leads to the formation of the native two-disulfide species. (iii) The rate-determining step involves the transition from the stable two-disulfide species to the native conformation. In this sequential progression bifurcations in the folding pathways are possible resulting in the parallel pathways to the native state.

Loop formation probability $P(l)$ may be obtained approximately using statistical mechanics of stiff chains [89]. Here, we provide the physical requirements. For chains with an effective persistence length l_p, we expect $P(l)$ to be negligible for $l < l_p$.[2] This is because the requisite self-avoidance criterion, bond angle, and dihedral angle constraints are violated for the loop lengths less than l_p. In the denatured conformations, excluded volume interactions are predominant; therefore for large enough l we expect $P(l)$ to decay as $\approx l^{-\theta_3}$ with $\theta_3 \approx 2.2$. Combining these requirements, we write $P(l)$ as

$$P(l) \approx \frac{1 - \exp(-l/l_p)}{l^{\theta_3}} \tag{9}$$

Experiments by Darby and Creighton [91], who measured the rates of formation of single disulfide intermediates in BPTI, can be understood using Eq. (9) for $P(l)$. The higher probability of forming loops between the ends of the chain is neglected in obtaining $P(l)$. This approximation should not have an effect in predicting the rates of single S–S bond formation in BPTI, but will be relevant in getting estimates of time scales for forming loops in polypeptide chains.

The general scheme described above has been applied to obtain approximately the refolding pathways in BPTI using experimentally determined rearrangement time τ_i for the transition from the single S–S intermediates to the double S–S species. Our results showed [45] that on a relatively long time scale, comparable to that used in the experiments by Creighton or WK, only native-like species should be populated. It may be that in the process of forming these native-like intermediates, certain non-native species identified by Creighton are transiently involved. Based on our estimate of τ_i, the transient

[2]In certain protein structures, loops with $l < l_p$ can form. However, such loops are stiff and often have very high strain energy [90].

population of non-native intermediates occurs on the time scales less than 30 seconds.

Because our theory is most accurate for predicting the ordering of single disulfide species, we focus on their rates and extents of accumulation. Considerations based on $P(l)$ suggest that only a small subset of the single disulfide intermediates can form. From $P(l)$ it follows that the probability of forming [14–38] is considerably greater than that of [5–55]. However using the kinetic constraints we have shown that although [14–38] forms rapidly and early in the folding process, its concentration decreases rapidly at subsequent times, whereas those of [30–51] and [5–55] increase. This specific prediction is one of the *striking outcomes* of the proximity rule [45]. The distinct kinetic behavior of the native [14–38] compared to the other two native single S–S intermediates is related to stability reasons [45,92]. The partially folded solvent-exposed state [14–38], which perhaps is the molten globule form of BPTI, can form without burying the hydrophobic core of the protein. On the other hand, the intermediates [5–55] and [30–51], in which the four Cys residues are in the interior, require the formation of the hydrophobic core of the protein (Fig. 7). The burial of hydrophobic residues that brings the Cys residues in proximity so those S–S bonds can form requires overcoming free energy barriers. This delays their formation compared to that of [14–38].

Proximity rule also predicts that the ratio of the maximum concentration of [30–51] to that of [5–55] is about 7 : 1, whereas the concentration of [14–38] is negligible on the same time scale. This ratio is in excellent agreement with the experiments of WK, who found a ratio of 6 : 1, and is in disagreement with Creighton's estimate of 20 : 1.

The theoretical prediction that [14–38] should be the first intermediate to accumulate was *subsequently confirmed* by Dadlez and Kim [92]. Using oxidized glutathione (GSSH) to initiate disulfide bond formation and acid quenches to trap intermediates, they noted that the earliest intermediate that accumulates is [14–38]. The tenfold rearrangement of [14–38] compared to [30–51] or [5–55] was rationalized in terms of stability (see arguments given above). These findings are also consistent with the results for synthetic models, in which the Cys except at the positions 14 and 38 were replaced by α-amino-*n*-butryic acid (*Abu*) [93]. The folding of $[14–38]_{Abu}$ is similar to the formation of [14–38] in the wild type. This reinforces the notion that entropic considerations and overall hydrophobicity of BPTI rather than specific native interactions between the remaining cysteines, perhaps on the collapse time scale, determine the early formation of [14–38].

Despite being intensively studied, there are several major questions in the refolding of BPTI that are not understood. We mention two of them: (a) The *in vitro* folding pathways show that there are dead-end kinetic traps [84], which completely block the folding reaction. Weissman and Kim [87] showed that

such kinetic traps are completely eliminated when the disulfide bonds rearrangements are catalyzed by protein disulfide isomerase (PDI). The presence of PDI, which may be viewed as an intramolecular chaperone, enhances the folding rate by several thousands. The mechanism of action of PDI has not been elucidated. (b) In a beautiful experiment, Zhang and Goldenberg [94] showed that the dead-end kinetic traps in the wild-type BPTI are entirely eliminated by a single amino acid substitution. The mutant Y35L (tyrosine at position 35 is replaced by leucine) results in a rapid sequential pathway in which only native intermediates are populated. The simplistic explanation of this spectacular experiment is that the nonproductive intermediates in this mutant are destabilized. A fuller molecular explanation is required.

C. Modeling the Role of S–S Bonds

A key disagreement between the early works [79–83] and the more recent studies WK on the refolding of BPTI is the role of non-native intermediates in directing the folding of BPTI. Creighton argued that not only were two non-native intermediates ([30–51;5–14] and [30–51;5–38]) accumulated substantially, but also they were equally involved in the productive folding pathways. WK showed that non-native states were not obligatory intermediates, and the only intermediates in the folding were native. Non-native intermediates may be involved in the transition state in the late stages of folding.

To clarify the relevance of non-native intermediates in the folding of proteins dictated by the formation of disulfide bonds Camacho and Thirumalai [45] used lattice models. While these models are merely caricatures of proteins, they contain the specific effects that can be studied in microscopic detail. We used a two-dimensional lattice sequence consisting of hydrophobic (H), polar (P), and Cys (C) residues. If two C beads are near neighbors on the lattice, they can form a S–S bond with an associated energy gain of $-\epsilon_s$ with $\epsilon_s > 0$. Thus, topological specificity is required for native S–S bond formation in this model. We have studied the folding kinetics of this model, which is perhaps the simplest model that can probe the characteristics of native and non-native disulfide bonded intermediates.

The sequence studied consists of $M = 23$ monomers, of which four represent C sites. The native conformation corresponds to [2–15;9–22] (Fig. 8a). The model sequence has six possible single and two disulfide intermediates including the native state. There are three *native* intermediates and two *non-native* intermediates. Even though the number of such intermediates are far less than the corresponding number in BPTI, it is sufficient to examine the crucial distinction between the roles played by native and non-native intermediates in the folding kinetics. Some of the questions that arise in the experimental studies of refolding of BPTI can be precisely answered using these simple models.

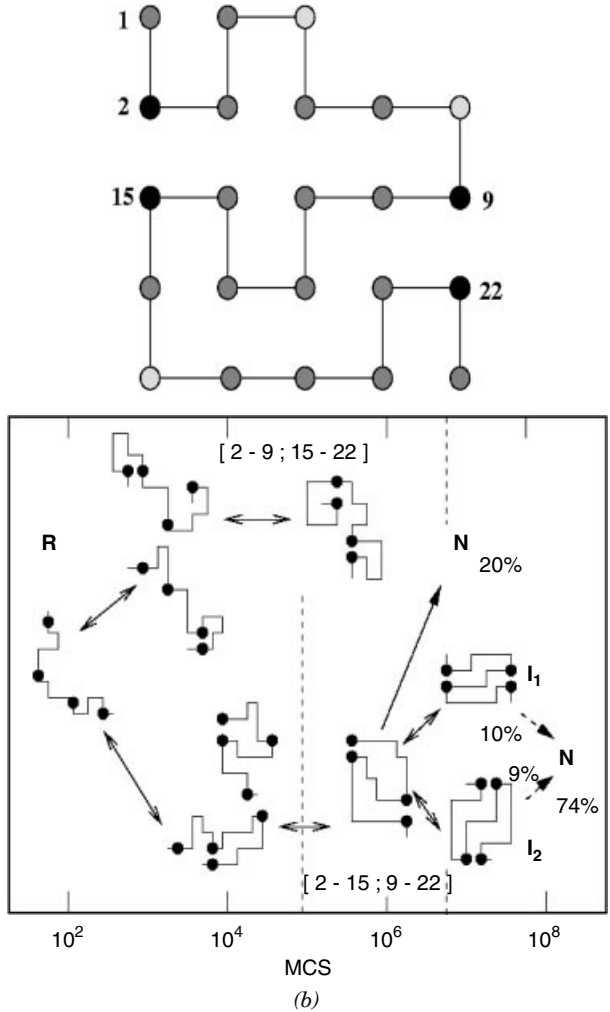


Figure 8. (See also color insert.) (a) The ground-state conformation of the two-dimensional model sequence with $M = 23$ beads and four covalent (S) sites. The red, green, and black circles represent, respectively, the hydrophobic (H), polar (P), and S sites. (b) Diagram of representative time snapshots along the main pathways of folding of the sequence in panel (a). The S sites are shown as black circles. Dotted lines delineate the three main folding regimes (random collapse, kinetics ordering and all-or-none). The arrows indicate the various transitions occurring in the system: the double-headed continuous arrows indicate backward and forward reactions where there is no substantial re-arrangement of the chain; the single-headed arrows indicate that the native-state is stable on the time scale of the simulations ($\sim 10^9$ MCS); the dashed arrows are for indirect transitions which occur by breaking the disulfide bonds and partial unfolding of the structure. The percentages indicate the concentration of the native and two native-like intermediates at the end of the second regime of kinetic ordering.

To probe the dynamical role played by the intermediates, we computed the time dependence of the concentration of the six species. The folding pathways are characterized in terms of the appearance of these intermediates (Fig. 8b). There are pathways that lead directly to **N** exclusively via native-like intermediates. In others, non-native intermediates are involved early in the folding process. For purposes of ascertaining the importance of the intermediates, all times are measured in terms of τ_F, the folding time. At the earliest time, $t < 10^{-5}\tau_F$ single disulfide species accumulate, whose probabilities of formation are determined by $P(l)$. At times that are roughly tenfold longer, the rearrangement of the nonnative single disulfide intermediates leads to the formation of two stable native single disulfide ([9–22] and [2–15]) species. These early intermediates act as seeds (nucleating sites) for subsequent formation of the native state [45]. At times on the order of about $10^{-4}\tau_F$, which coincide with the time at which native single disulfide species form, the concentration of these intermediates *cannot* be determined based on entropic considerations alone. Energetic considerations, such as favorable hydrophobic interactions, affect the formation of single disulfide intermediates.

In the second stage of the assembly we find that non-native two disulfide intermediate [2–9;15–22] can form transiently (Fig. 9b). Because this intermediate is unstable, it quickly rearranges to the more stable native **N** state. On relatively long time scales ($t \approx 0.01\tau_F$) we find that there are two native-like intermediates, in which the disulfides are in place but some other parts of the structure are not yet fully formed. This may be the analogue of the N_{sh}^{sh} state in BPTI which only needs the nearly solvent-exposed [14–38] bond to form. In the final stage of folding, structural fluctuations that transiently break the native S–S bonds enable the transition to **N**. This transition involves the transient formation of the non-native intermediate [2–9;15–22]. The two native-like intermediates I_1 and I_2 (Fig. 8b) rearrange almost exclusively via [2–9;15–22].

Even with an extremely simple model, several conclusions have been reached, which help clarify some of the issues in the refolding of BPTI. (1) Non-native species can form early in the folding process when bulk of the ordering is determined by entropic considerations. The current experiments on BPTI are far too slow to detect these early intermediates. On the time scale of collapse the more stable single disulfide species, which are native-like, form. (2) As the folding reaction progresses, native-like intermediates tend to form so that the productive pathways largely contain native-like intermediates. (3) The rate-determining step involves an activated transition from native-like species, via a high free-energy non-native transition state, to **N**. The transitions appear to involve rearrangement of the structure that does not involve the S–S bonds. These calculations suggest that although the folding pathways of BPTI can be described in terms of the disulfide intermediates, a complete description requires accounting for hydrophobic and charge effects as well. At present,

these effects have not been completely examined experimentally or theoretically. The profound effect of point mutations [94] in altering the folding rates and the pathways of BPTI folding suggests that there are strong couplings between S–S bond formation and other forces that drive the native structure formation.

D. Engineering Disulfide Bonds in Barnase

To probe the folding pathways in BPTI the S–S bonds were initially reduced that results in unfolding. Refolding is initiated under oxidizing conditions that enable S–S bond formation and restoration of the native state. Alternatively, the impact of S–S bonds can be studied by engineering them at specific locations. With the S–S bonds intact, protein can be unfolded using denaturants such as urea. The folding kinetics can be initiated by diluting the denaturant. The latter procedure, which was first used by Clarke and Fersht [95], enables the study of the effect of intact S–S bonds on the stability and kinetics of folding. Clarke and Fersht used this procedure to engineer S–S bonds at two specific locations in barnase, whose folding without disulfide bonds has been well-characterized. This allows for a comparison of folding characteristics of proteins with and without disulfide bonds.

Two positions in barnase were constrained by S–S bonds that were left intact [95a]. One of them, between residues 85 and 102, connects two loops that apparently form early in the folding pathway of the wild type protein. A second disulfide between residues 43 and 80 connects two secondary structural elements. Barnase containing disulfide bonds is more stable than the wild type because the introduction of the S–S bond increases the free energy of the unfolded states. From the native state of barnase it is clear that the enhanced stability upon introduction of the disulfide bond between 43 and 80 cannot be accounted for solely by lowered entropy of the unfolded state compared to the WT. Using the Flory estimate we expect that stability of $[43–80]_{Bar}$ should be $1.5\,RT\ln 38 \approx 3.2$ kcal/mol, whereas that of $[85–102]_{Bar}$ is ≈ 2.6 kcal/mol. These estimates do not compare favorably with the experimental values, which are 2.1 kcal/mol and 4.3 kcal/mol for $[43–80]_{Bar}$ and $[85–102]_{Bar}$, respectively. This suggests that the introduction of S–S bonds could also stabilize the native state to some extent.

Refolding kinetics of the mutated barnase depends strongly on the location of the S–S bond. Assuming that reduction in the conformation space leads to rate enhancement, we would predict that $[43–80]_{Bar}$ should fold faster than $[85–102]_{Bar}$. However, the opposite trend is found experimentally. The mutant with the shorter loop folds about five times more rapidly, whereas barnase with the disulfide between 43 and 80 folds two times slower than the wild type. Using lattice simulations, Abkevich and Shakhnovich [95b] argued that if S–S bonds are engineered into the regions highly structured in the transition state, refolding rates can be increased compared to the WT. The presence of S–S bonds

elsewhere in the protein can either increase or decrease folding rates depending on the external conditions. Because the region containing residues 83 and 102 forms early in the folding process, it may be part of the folding nuclei. This explains the enhanced rate of folding of $[85–102]_{Bar}$ compared to WT. Because residues 43 and 80 are not part of the folding nuclei, the folding of $[43–80]_{Bar}$ is about 1.7 times slower than the WT. Thus, the simulations of Abkevich and Shakhnovich using simple lattice models are consistent with experiments.

VII. CHAPERONIN-FACILITATED PROTEIN FOLDING

According to the Anfinsen's hypothesis [88] natural proteins fold spontaneously to their lowest free energy states. By analyzing the weights of proteins and protein synthesis rates under glucose feeding conditions, Lorimer [96] estimated that in *Escherichia coli* more than 90% of proteins fold to their native states as envisaged by Anfinsen. This is remarkable because one might imagine that trafic (due to other macromolecules) in the crowded cellular environment might lead to strong intermolecular interactions which could potentially interfere with monomeric folding. Nevertheless, it appears that many proteins assemble spontaneously to their functionally competent states *in vivo* as envisioned by Anfinsen. However, there are some proteins that require the assistance of molecular chaperones to fold to the native conformation. The functions of the class I chaperonins belonging to heat shock protein family are the most extensively studied [97–100]. In this chapter we focus on insights into their function using simple lattice models [47,101,102].

The chaperonin family of proteins, namely GroEL and GroES, that function as a nanomachine by utilizing ATP, assist misfolded substrate proteins to reach their native states [100,103,104]. The crystal structures of GroEL [105], GroES [106], and the complex GroEL/GroES/ADP [107] have provided considerable insights into the chaperonin action. The chaperonin GroEL is a double-ringed oligomer consisting of two back-to-back stacked heptameric rings. It has an overall cylindrical structure divided into two nonconnected cavities, in which the substrate protein (SP) can be sequestered. Each subunit of the GroEL particle consists of three domains, namely, the equatorial domain, the intermediate domain, and the apical domain [100]. The heaviest of these is the equatorial domain, which contains more than half of the molecular weight of GroEL. We have argued that the concentration of dense inertial mass in the equatorial domain is necessary to generate the requisite force to peel the initially captured substrate protein (SP) from the apical domain. The concentric assembly of the subunits produces a ring structure having an architecture with an unusual sevenfold symmetry (Fig. 9a).

The co-chaperonin GroES, containing seven subunits [106], caps the GroEL particle as a dome. A remarkable feature, which has mechanistic implications, is

that upon binding of GroES and ATP the volume of the cavity doubles [100]. This enhanced volume is accompanied by a series of concerted allosteric transition that the GroEL particle undergoes [108–110]. Because of the non-specificity of GroEL-SP interactions [111–113] and the plasticity of the architecture of the GroEL particle, this system acts as a ''one size fits all'' nanomachine.

Considerable progress in understanding the mechanism of this nanomachine has become possible due to a combination of an extraordinary body of experimental work [98,100] and some contributions from theoretical studies [114,115]. The hemicycle, which constitutes the fundamental functioning cycle of the GroEL machine [110], is schematically sketched in Fig. 9b. The process is initiated by the capture of the SP by the apical domain of the GroEL particle. To a first approximation, the mouth of the cavity can be thought of as a continuous hydrophobic surface formed by the helices in the apical domain. The nonspecific, but favorable, interaction between the SP and GroEL is due to the attraction between the exposed hydrophobic residues of the SP and the hydrophobic surface of the apical domain. Upon binding of ATP and GroES (in this specific order), significant concerted transitions occur in the GroEL particle. The series of transitions alters, in a fundamental way, the nature of interaction between GroEL and the SP [100]. Whereas in the process of capture the SP-GroEL interaction is attractive, the interaction is either neutral or even repulsive after encapsulation (step 2 in Fig. 9b). The surface remains hydrophilic until the restoration of GroEL to the initial state. This alteration between hydrophobic (H) and hydrophilic (P) surface enables this system to function as an annealing machine. The release of GroES and the encapsulated SP occurs when ATP and/or another SP molecule binds to the *trans*-ring [107].

Although the underpinnings of the cycle (Fig. 9b) are based on a number of experiments and theoretical arguments, several outstanding questions remain. A key issue is related to the coupling between the concerted allosteric transitions that the GroEL particle undergoes and the SP folding rate [47,116]. Consider the cycle displayed in Fig. 9b. Upon binding ATP to the upper ring, a cooperative transition $T \leftrightarrow R$ takes place. The terminology T and R are borrowed from the Monod–Wyman–Changeaux model [117] describing the binding of oxygen to hemoglobin. The tense state T has a higher afinity for ATP than the relaxed R state [109]. Upon binding ATP, the intermediate domain moves 25° toward the equatorial domain, which closes the ATP binding sites. Even with this relatively minor rigid body movement of the intermediate domain, the interaction between the SP and the walls (the apical domain) are weakened [108,109]. The weakened interaction is sufficient to enable the SP protein to unfold at least partially [118]. Subsequent binding of ATP and GroES leads to much larger domain movements in the GroEL particle. In particular, the apical domain moves upward by 60° and twists, with respect to the equatorial

domain, by 90° [100]. This large segmental motion, which results in the encapsulation of the SP, doubles the volume of the cavity. Upon encapsulation, the interaction between the SP and the walls is either neural or repulsive depending on the size of SP. At least five independent rate constants are required to describe these large-scale concerted allosteric transitions in the GroEL particle [110]. This makes the description of the coupling between allostery of GroEL and the SP folding rate very difficult.

To examine the coupling between the allosteric transitions and SP folding rates, a model system may be considered in which the action of GroEL and ATP

(a)

Figure 9. (See also color insert.) (a) Rasmol [126] view of one of the two rings of GroEL from the PDB file 1oel. The seven chains are indicated by different colors. The amino acid residues forming the binding site of the apical domain of each chain (199–204, helix H: 229–244 and helix I: 256–268) are shown in red. The most exposed hydrophobic amino acids that are facing the cavity and are implicated in the binding of the substrate as indicated by mutagenesis experiments [112, 127] are : Tyr199, Tyr203, Phe204, Leu234, Leu237, Leu259, Val263, and Val264. (b) A schematic sketch of the hemicycle in the GroEL–GroES-mediated folding of proteins. In step 1 the substrate protein is captured into the GroEL cavity. The ATPs and GroES are added in step 2, which results in doubling the volume, in which the substrate protein is confined. The hydrolysis of ATP in the *cis*-ring occurs in a quantified fashion (step 3). After binding ATP to the *trans*-ring, GroES and the substrate protein are released that completes the cycle (step 4).

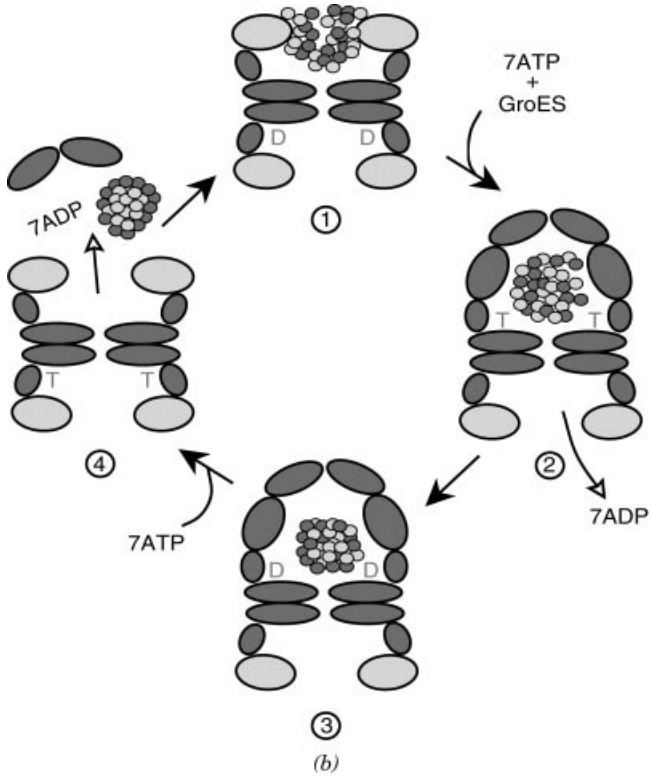


Figure 9 (*Continued*)

on SP can be investigated without the complication of the GroES interaction [116]. Many *in vitro* studies on the interaction between the GroEL and the SP have used only this subset [98]. In this reduced model system the equilibrium constant between T and R states and the time constants characterizing the SP folding are the only relevant parameters [110,116]. To examine the coupling in the reduced nanomachine, we modeled the central cavity as a cubic box, and a lattice model representation of the polypeptide chain was employed [47]. This, of course, is a highly simplified representation the GroEL–SP system. However, qualitative testable predictions of the coupling between allostery and the SP can be made using this caricature. Initially the interior walls of the GroEL particle (in the T state) are assumed to be hydrophobic. This description is reasonable, because in the T state the arrangement of the apical domain offers the SP a continuous lining of hydrophobic residues (Fig. 9a). We vary the wall hydrophobicity

of GroEL by letting one particular residue (leucine) describe the wall character. Thus, the interactions between the wall and the SP protein is

$$E_c = \sum_i hE_{wi} \tag{10}$$

where h $(0 \leq h \leq 1)$ gives the strength of the interaction, and E_{wi} is the contact interaction between the ith residue of the SP and the wall. The total energy of the encapsulated SP is given by the sum of Eq. (10) and the ''internal'' energy of the SP [Eq. (1)].

The key annealing action of the GroEL particle arises due to the changes in the hydrophobicity of its inner walls during the hemicycle [47]. In other words, during a single turnover the cavity changes from being able to capture the SP to that is which binding is not favored. This change in the wall character is accompanied by the allosteric transition of GroEL, resulting in the encapsulation of the SP. The effect of changing hydrophobicity is mimicked in our simplified model by letting the hydrophobicity of the confining cavity vary during the turnover time, τ_i. We divide τ_i into two subintervals. During a period τ_P the wall remains hydrophilic (P), and for the remainder $\tau_i - \tau_P$ the cavity is hydrophobic. Because the model does not include GroES, the situation we consider may serve as a model for the coupling between $T \leftrightarrow R$ transition and the SP folding. Here we have $\tau_P/(\tau_i - \tau_P) \sim L$, where L is the equilibrium constant between T and R [116]. By examining the effect of changing values of L on the rates of the SP folding the dependence of the SP folding rate on the allosteric equilibrium transitions can be examined. Simulations of the simplified lattice representation of the GroEL–SP system shows that there is an inverse correlation between the extent of the $T \leftrightarrow R$ transition and the folding rate of the SP. In other words, as the cooperativity of the $T \leftrightarrow R$ transition increases (higher values of L), the slower is the SP folding rate.

A. Unfolding Activity of GroEL

Although it is accepted that the GroEL nanomachine rescues the SP by stochastically enabling it to sample the rough free-energy landscape, the microscopic action on the SP has only recently become clear. A few experiments have shown that upon change in the wall characteristics of GroEL the SP unfolds partially, if not globally. Using hydrogen exchange labeling, Zahn et al. [118] showed that GroEL accelerates the exchange of highly protected amide protons. Because highly protected protons (high protection factor) are typically buried in the core of the SP, it follows that the SP unfolds in the presence of GroEL.

Nieba-Axmann et al. [119] also examined the plausible structural fluctuations in GroEL-bound cyclophilin A (CypA) using amide-proton exchange

measurements. In the absence of nucleotides and GroES, folding of CypA is extremely sluggish. Upon addition of ADP, the rate increases by a factor of about 2.5, whereas the addition of ATP leads to a threefold enhancement in the folding rate. However, when GroES is added, the rate increases by a factor of 14 at 6°C and by nearly 30-fold at $T = 30$°C. The near independence of the refolding on nucleotides suggests that the full recovery of CypA occurs within a single turnover.

Upon binding of ATP and GroES, the domain moves upward and twists by 90° about the equatorial domain [100]. This results in the weakening of the interaction between the SP and the walls of the cavity. The simple lattice model described above, in which the character of the wall changes from H to P, can model the chaperonin-assisted folding provided the folding reaction is complete in a single turnover. To examine the structure of the polypeptide chain due to alterations in the wall character, we computed the inherent structures of the chain in the on state (hydrophobic wall) and in the off state (hydrophilic wall). According to the iterative annealing mechanism [103,104], upon going from the on state to the off state the polypeptide chain should undergo kinetic partitioning [Eq. (8)]. The inherent structures prior to and immediately following the change in the cavity characteristics allows us to compute the degree of commitment of the SP to folding. Because the GroEL machine operates stochastically, there ought to be a distribution of states of the SP that are populated as the on–off transition takes place. The simulations show (see Fig. 9 of Ref. 47) that before the transition in the cavity, a fraction of molecules is committed to folding, while most of the conformations fall into basins of attraction corresponding to misfolded or unfolded states. After the transition to the off state the chain is largely unfolded. This shows that upon weakening of the SP–GroEL interaction, which occurs as the GroEL particle undergoes the allosteric transitions, the polypeptide chain globally unfolds. In other words, the chief mechanism operative in the GroEL-mediated folding is that chaperonins *help fold proteins by globally unfolding them!* This is consistent with the predictions of IAM and is also affirmed by several experiments [118–120].

The simulations using simplified models are entirely consistent with several experiments including the one reported by Nieba-Axmann et al. [119], who noted that amide protons that are highly protected from hydrogen exchange in the native state of CypA in the absence of GroEL become much less protected when bound to the chaperonin. The protection factor decreases by nearly two orders of magnitude upon binding to GroEL. Thus binding to GroEL shifts the equilibrium from compact native-like states to globally unfolded conformations. In this dynamical picture of GroEL action, as opposed to the static Anfinsen cage model, chaperonins unfold the SP. It also follows that efficient folding can be induced by repeated unfolding of the chain.

B. Unfolding by Stretching

If, in the course of the allosteric transitions of the GroEL particle, the SP is unfolded, a natural question is, what is the mechanism of the GroEL-SP interaction that globally unfolds the protein? Lorimer and co-workers [120] have explored this issue using hydrogen–tritium exchange experiments in chaperonin-assisted folding of RUBISCO. They observed that within the time scale of a single turnover (approximately 13 seconds), complete exchange of tritiums takes place. This shows that RUBISCO unfolds at least partially, if not globally. From the crystal structures of GroEL and the GroEL–ADP–GroES complex, it is known that upon undergoing a series of concerted allosteric transitions, two adjacent subunits that are about 25 Å apart in the T state are about 33 Å apart in the R'' state [107,120]. This large-scale movement is presumed to generate force on the SP [99]. Recent pulling experiments on several proteins [121] show that the native state can be fully unfolded if a force exceeding a threshold value is applied. The magnitude of the threshold force depends on the SP [122].

To estimate the value of the force imparted to the SP, it is necessary to obtain the interaction energy between the SP and the apical domain of the GroEL particle. The SP–GroEL interaction energy must exceed $\frac{3}{2}k_BTS_{mis}$, where S_{mis} is the translational entropy of the misfolded chain molecule for capture by GroEL to occur. Assuming that the subunits move apart by about 0.2 nm, we estimate that the minimum force required to peel off the SP from the apical domain is about 35 pN. A more precise estimate of the interaction energy between the SP and GroEL can be made by assuming that the inner lining of the GroEL cavity can be modeled as a hydrophobic wall onto which the SP is adsorbed [114]. By balancing the free energy gain due to favorable hydrophobic interaction between GroEL and SP and the entropy loss due to the pinning of the SP, we estimate that the interaction energy should not exceed about $10\,k_BT$ [114]. The force needed to overcome this interaction is about 200 pN. This value is large enough to unfold immunoglobin proteins with β-sandwich topology [121]. We suspect that at these values of the forces most substrate proteins can at least partially unfold. These estimates give credence to the notion that it is the generation of force in the power stroke of the chaperonin machinery that unfolds the SP [120].

The estimate of the force given above is only an average force. A given SP molecule can bind to a subset of the seven subunits [123]. Because of the heterogeneity of the conformations of the misfolded SP, we expect variations in the binding states from molecule to molecule. Thus, there should be a distribution of unfolding forces. Recent AFM experiments [124] show that this distribution is very broad, indicating that there is a large sample to sample variation in the unbinding forces. Such large variations for other SP can only

be measured using single-molecule measurements. Surely, these sample-to-sample variations in lifetimes of the complexes [125] and forces imparted to SP will require revisions of the iterative annealing mechanism [103].

VIII. CONCLUSIONS

Protein folding presents significant challenges because the parameter space is extremely large. From the myriad of experimental and theoretical studies it is not clear that there are many general principles that govern the kinetics of folding. Nevertheless, using simple models, several precise predictions have been made. In this chapter we have described the utility of simple lattice models and phenomenological theories in answering very specific questions in protein folding. It is remarkable that these simple ideas have been fruitful in enabling us to formulate conceptual questions such as the physical basis for the emergence of structures and their designability. Lattice models can also be used to understand qualitatively the importance of intermediates in the folding of proteins that are controlled by the stability of disulfide bonds. Experimentally testable predictions in the field of assisted folding also have been made using caricatures of the chaperonin systems. These practical applications attest to the utility of these models in providing a conceptual understanding of the basic principles in a variety of problems.

None of the applications described here can be tackled using a ''realistic'' all-atom representation of proteins. The precise predictions that we and others have made using coarse-grained models of proteins are currently beyond the reach of molecular dynamics simulations. In this sense, we hope that this chapter serves as a challenge to the practitioners of all-atom simulations. Even assuming that the interaction potentials are adequate, the severe restriction on the simulation time scales acts as a major constraint. The lack of reliable potentials and the accessible computer times has prevented straightforward use of molecular dynamics calculations from being a predictive tool. There is hope that, in the next few years, unlimited computer power may be unleashed to obtain a detailed picture of how proteins fold. This potential comes from developments in distributed computing that can, in principle, be used to generate several long trajectories. In the applications that we have carried out, a rather detailed picture of β-hairpin assembly for several sequences has been obtained (D. K. Klimov, D. Newfield, and D. Thirumalai, unpublished results). A different, but related, approach has also been undertaken by Pande and co-workers (V. Pande, private communication). The development of a high-performance computer by IBM also has raised the specter of hope that computational bottlenecks may be overcome in the next few years so that challenging problems such as biomolecular folding can be undertaken. Even if these tolls are routinely available, simple concepts will play a pivotal

role in formulating the issues in the study of biomolecules, because in the ultimate analysis protein folding (or any other problem in molecular biology) is not *merely* a computational problem.

Acknowledgments

We are grateful to Carlos Camacho, Zhuyan Guo, and George H. Lorimer for illuminating discussions and George Rose for providing us with the Fig. 1. We thank David Goldenberg for a critical reading of the article. This work was supported in part by the National Science Foundation grant CHE99-75150.

References

1. T. E. Creighton, *Proteins: Structures and Molecular Principles*, W. H. Freeman, New York, 1993.
2. T. Schindler, M. Herrier, M. A. Marahiel, and F. X. Schmid, Extremely rapid protein folding in the absence of intermediates. *Natur. Struct. Biol*. **2**, 663–673 (1995).
3. C. B. Anfinsen, Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).
4. R. L. Baldwin, Why is protein folding so fast. *Proc. Natl. Acad. Sci. USA* **93**, 2627–2628 (1996).
5. C. Levinthal, in *Mossbauer Spectroscopy in Biological Systems*, P. Debrunner, J. C. M. Tsibris, and E. Munck, eds., University of Illinois Press, Urbana, IL, 1968.
6. K. A. Dill and H. S. Chan, From Levinthal to pathways to funnels. *Natur. Struct. Biol.* **4**, 10–19 (1997).
7. A. R. Fersht, Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol*. **7**, 3–9 (1997).
8. W. A. Eaton, V. Munoz, P. A. Thompson, C. K. Chan, and J. Hofrichter, Submillisecond kinetics of protein folding. *Curr. Opin. Struct. Biol*. **7**, 10–14 (1997).
9. V. S. Pande, A. Y. Grosberg, T. Tanaka, and D. S. Rokhsar, Pathways for protein folding: Is a new view needed? *Curr. Opin. Struct. Biol.* **8**, 68–79 (1998).
10. D. Thirumalai and D. K. Klimov, Deciphering the timescales and mechanisms of protein folding using minimal off-lattice models. *Curr. Opin. Struct. Biol.* **9**, 197–207 (1999).
11. R. L. Baldwin and G. D. Rose, Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* **24**, 26–33 (1999).
12. R. L. Baldwin and G. D. Rose, Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem. Sci*. **24**, 77–83 (1999).
13. J. Rumbley, L. Hoang, L.Mayne, and S. W. Englander, An amino acid code for protein folding. *Proc. Natl. Acad. Sci. USA* **98**, 105–112 (2001).
14. P. T. Lansbury, Evolution of amyloids: What normal protein folding can tell us about fibrillogenesis and disease. *Proc. Natl. Acad. Sci. USA* **96**, 3342–3344 (1999).
15. S. Williams, T. P. Cosgrove, R. Gillmanshin, K. S. Fang, R. H. Callender, W. H. Woodruff, and R. B. Dyer, Fast events in protein folding: Helix melting and formation in a small peptide. *Biochemistry* **35**, 691–697 (1996).
16. G. S. Huang and T. G. Oas, Submillisecond folding of monomeric λ repressor. *Proc. Natl. Acad. Sci. USA* **92**, 6878–6882 (1995).
17. V. Munoz, P. A. Thompson, J. Hofrichter, and W. A. Eaton, Folding dynamics and mechanism β-hairpin formation. *Nature* **390**, 196–199 (1997).
18. A. G. Ladurner and A. R. Fersht, Upper limit of the time scale for diffusion and chain collapse in chymotrypsin inhibitor 2. *Nature Struct. Biol.* **6**, 28–31 (1999).

19. K. Kuwata, R. Shastry, H. Cheng, M. Hashino, C. A. Batt, Y. Goto, and H. Roder, Structural and kinetic characterization of early folding events in β-lactoglobulin. *Natur. Struct. Biol.* **8**, 151–155 (2001).
20. M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub, Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science* **276**, 1109–1112 (1997).
21. M. S. F. Kellermayer, S. B. Smith, H. L. Granzier, and C. Bustamante, Folding–unfolding transitions in single titin molecules characterized with laser tweezers. *Science* **276**, 1112–1116 (1997).
22. A. F. Oberhauser, P. E. Marszalek, M. Carrion-Vazquez, and J. M. Fernandez, Single protein misfolding events captured by atomic force microscopy. *Natur. Struct. Biol.* **6**, 1025–1028 (1999).
23. H. Li, A. F. Oberhouser, S. B. Fowler, J. Clarke, and J. M. Fernandez, Atomic force microscopy reveals the mechanical design of a modular protein. *Proc. Natl. Acad. Sci. USA* **97**, 6527–6531 (2000).
24. T. E. Fisher, A. F. Oberhauser, M. Carrion-Vazquez, P. E. Marszalek, and J. M. Fernandez, The study of protein mechanics with the atomic force microscope. *Trends Biochem. Sci.* **24**, 379–384 (1999).
25. A. R. Fersht, Characterizing transition states in protein folding. *Curr. Opin. Struct. Biol.* **5**, 79–84 (1995).
26. L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, The nature of the transition state of chymotrypsin inhibitor 2 analyzed by protein engineering methods: Evidence for a nucleation–condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260–288 (1995).
27. E. L. McCallister, E. Alm, and D. Baker, Critical role of β-hairpin in protein G folding. *Natur. Struct. Biol.* **7**, 669–673 (2000).
28. S. E. Jackson, How do small single domain proteins fold. *Fold. Des.* **3**, R81–R91 (1998).
29. R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, Optimal protein-folding codes from spin-glass theory. *Proc. Natl. Acad. Sci. USA* **89**, 4918–4922 (1992).
30. A. Sali, E. Shakhnovich, and M. Karplus, Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* **235**, 1614–1636 (1994).
31. J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Funnels, pathways and energy landscape of protein folding, *Proteins: Struct. Funct. Genet.* **21**, 167–195 (1995).
32. D. K. Klimov and D. Thirumalai, Factors governing the foldability of proteins. *Proteins: Struct. Funct. Genet.* **26**, 411–441 (1996).
33. J. N. Onuchic, Z. A. Luthey-Schulten, and P. G. Wolynes, Theory of protein folding: An energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600 (1997).
34. D. K. Klimov and D. Thirumalai, Linking rates of folding in lattice models of proteins with underlying thermodynamic characteristics. *J. Chem. Phys.* **109**, 4119–4125 (1998).
35. D. S. Riddle, V. P. Grantcharova, J. V. Santiago, E. Alm, I. Ruczinski, and D. Baker, Experiment and theory highlight role of native topology in SH3 folding. *Natur. Struct. Biol.* **6**, 1016–1024 (1999).
36. D. Baker, A surprising simplicity of protein folding. *Nature* **405**, 39–42 (2000).
37. K. Plaxco, K. T. Simons, and D. Baker, Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994 (1998).
38. J. Clarke, E. Cota, S. B. Fowler, and S. J. Hamill, Folding studies of immunoglobulin-like β-sandwich proteins suggest that they share a common folding pathway. *Struct. Fold. Des.* **7**, 1145–1153 (1999).

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

40. D. Thirumalai, From minimal models to real proteins: Time scales for protein folding. *J. Phys. I* **5**, 1457–1467, 1995.

41. A. R. Dinner, A. Sali, L. J. Smith, C. M. Dobson, and M. Karplus, Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci*. **25**, 331–339 (2000).

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

43. D. Thirumalai, D. K. Klimov, and M. R. Betancourt, Exploring the folding mechanisms using lattice models, in *Monte Carlo Approach to Biopolymers and Protein Folding*, P. Grassberger, G. T. Barkema, and W. Nadler, eds., Singapore, World Scientific, 1998, pp. 19–28.

44. D. Thirumalai and D. K. Klimov, Emergence of stable and fast folding protein structures, in S. Kim, K. J. Lee, and W. Sung, eds., *Stochastic Dynamics and Pattern Formation in Biological Systems*, American Institute of Physics, New York, 2000, pp. 95–111.

45. C. Camacho and D. Thirumalai, Theoretical predictions of folding pathways using the proximity rule with applications to BPTI. *Proc. Natl. Acad. Sci. USA* **92**, 1277–1281 (1995).

46. C. Camacho and D. Thirumalai, Modeling the role of disulfide bonds in protein folding: Entropic barriers and pathways. *Proteins Struct. Funct. Gen.* **22**, 27–40 (1995).

47. M. R. Betancourt and D. Thirumalai, Exploring the kinetic requirements for enhancement of protein folding rates in the GroEL cavity. *J. Mol. Biol*. **287**, 627–644 (1999).

48. W. J. C. Orr, Statistical treatment of polymer solutions at infinite dilution. *Trans. Faraday Soc.* **43**, 12–27 (1947).

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

50. H. S. Chan and K. A. Dill, Intrachain loops in polymers: Effects of excluded volume. *J. Chem. Phys*. **90**, 493–509 (1989).

51. N. D. Socci and J. N. Onuchic, Kinetic and thermodynamic analysis of protein-like heteropolymers: Monte carlo histogram technique. *J. Chem. Phys*. **103**, 4732–4744 (1995).

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

53. S. Miyazawa and R. L. Jernigan, Estimation of effective inter-residue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules* **18**, 534–552 (1985).

54. A. Kolinski, A. Godzik, and J. Skolnick, A general method for the prediction of the three dimensional structure and folding pathway of globular proteins: Application to designed helical proteins. *J. Chem. Phys*. **98**, 7420–7433 (1993).

55. M. R. Betancourt and D. Thirumalai, Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Prot. Sci*. **8**, 1–8 (1999).

56. L. A. Mirny and E. I. Shakhnovich, How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol*. **264**, 1164–1179 (1996).

57. D. Tobi and R. Elber, Distance-dependent, pair potential for protein folding: Results from linear optimization. *Proteins Struct. Funct. Gen*. **41**, 40–46 (2000).

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

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

60. D. K. Klimov and D. Thirumalai, Cooperativity in protein folding: From lattice models with side chains to real proteins. *Fold. Des*. **3**, 127–139 (1998).

61. J. L. Martin, Computer enumerations, in *Phase Transitions and Critical Phenomena*, C. Domb and M.S. Green, eds., Academic Press, New York, 1974, pp. 102–110.

62. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Equation of state calculations by fast computing machines. *J. Chem. Phys*. **21**, 1087–1092 (1953).

63. A. M. Ferrenberg and R. H. Swendsen, Optimized monte carlo data analysis. *Phys. Rev. Lett*. **63**, 1195–1198 (1989).

64. R. V. Pappu, R. Srinivasan, and G. D. Rose, The flory isolated-pair hypothesis is not valid for polypeptide chains: Implications for protein folding. *Proc. Natl. Acad. Sci. USA* **97**, 12565–12570 (2000).

65. Z. Guo and D. Thirumalai, Kinetics and thermodynamics of folding of a *de novo* designed four-helix bundle. *J. Mol. Biol*. **263**, 323–343 (1996).

66. S. Govindarajan, R. Recabarren, and R. Goldstein, Simulating the total number of protein folds. *Proteins Struct. Funct. Gen*. **35**, 408–414 (1999).

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

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

69. Per-Anker Lindgard and H. Bohr, Magic numbers in protein structures. *Phys. Rev. Lett.* **77**, 779–782 (1996).

70. M. E. Holtzer, E. G. Lovett, D. A. d'Avignon, and A. Holtzer, Thermal unfolding in a gcn4-like leucine zipper: $^{13}C^{\alpha}$ nmr chemical shifts and local unfolding curves. *Biophys. J.* **73**, 1031–1041 (1997).

71. A. Bachman and T. Kiefhaber, Apparent two-state tendamistat folding is a sequential process along a defined pathway. *J. Mol. Biol*. **306**, 375–386 (2001).

72. V. S. Pande and D. S. Rokhsar, Folding pathway of a lattice model for proteins. *Proc. Natl. Acad. Sci. USA* **96**, 1273–1278 (1999).

73. Z. Guo and D. Thirumalai, Kinetics of protein folding: Nucleation mechanism, time scales and pathways. *Biopolymers*, **36**, 83–103 (1995).

74. T. Kiefhaber, Kinetic traps in lysozyme folding. *Proc. Natl. Acad. Sci. USA* **92**, 9029–9033 (1995).

75. T. Pan and T. R. Sosnick, Intermediates and kinetic traps in the folding of large ribozyme revealed by UV and CD spectroscopies and catalytic activity. *Nature Struct. Biol*. **14**, 931–938 (1997).

76. A. Matagne, S. E. Radford, and C. M. Dobson, Fast and slow tracks in lysozyme folding: Insight into the role of domains in the folding process. *J. Mol. Biol.* **267**, 1068–1074 (1997).

77. X. Zhuang, L. E. Bartley, A. P. Babcock, R. Russell, T. Ha, D. Herschlag, and S. Chu, A single-molecule study of RNA catalysis and folding. *Science* **288**, 2048–2051 (2000).

78. D. Thirumalai, D. K. Klimov, and S. A. Woodson, Kinetic partitioning as a unifying theme in the folding of biomolecules. *Theor. Chem. Acc.* **1**, 149–156 (1997).

79. T. E. Creighton, Renaturation of the reduced bovine pancreatic trypsin inhibitor. *J. Mol. Biol*. **87**, 563–577 (1974).

80. T. E. Creighton, Conformational restrictions on the pathway of folding and unfolding of BPTI. *J. Mol. Biol.* **113**, 275–293 (1977).

81. T. E. Creighton, Energetics of folding and unfolding of pancreatic trypsin inhibitor. *J. Mol. Biol*. **113**, 295–312 (1977).

82. T. E. Creighton, Effects of urea and guanidine. Hcl on the folding and unfolding of pancreatic trypsin inhibitor. *J. Mol. Biol*. **113**, 313–328 (1977).

83. T. E. Creighton and D. P. Goldenberg, Kinetic role of meta-stable native like two disulphide species in the folding transition of bovine pancreatic trypsin inhibitor. *J. Mol. Biol*. **179**, 497–526 (1984).

84. J. S. Weissman and P. S. Kim, Reexamination of the folding of BPTI: Predominance of native intermediates. *Science* **253**, 1386–1393 (1991).

85. J. S. Weissman and P. S. Kim, Kinetic role of non-native species in the folding of bovine pancreatic trypsin inhibitor. *Proc. Nat. Acad. Sci. USA* **89**, 9900–9904 (1992).

86. J. S. Weissman and P. S. Kim, The pro region of BPTI facilitates folding. *Cell* **71**, 841–851 (1992).

87. J. S.Weissman and P. S. Kim, Efficient catalysis of disulphide bond rearrangements by protein disulphide isomerase. *Nature* **365**, 185–188 (1993).

88. C. B. Anfinsen, Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).

89. D. Thirumalai, Time scales for the formation of the most probable tertiary contacts in proteins with applications to cytochrome C. *J. Phys. Chem.* **103**, 608–610 (1999).

90. D. K. Klimov and D. Thirumalai, Mechanisms and kinetics of β-hairpin formation. *Proc. Natl. Acad. Sci. USA* **97**, 2544–2549 (2000).

91. N. J. Darby and T. E. Creighton, Dissecting the disulphide-coupled folding pathway of bovine pancreatic trypsin inhibitor—forming 1st disulphide bonds in analogues of the reduced protein. *J. Mol. Biol*. **232**, 873–886 (1993).

92. M. Dadlez and P. S. Kim, A third native one-disulphide intermediate in the folding of bovine pancreatic trypsin inhibitor. *Nature Struct. Biol.* **2**, 674–679 (1995).

93. M. Ferrer, G. Barany, and C. Woodward, Partially folded molten globule and molten coil states of bovine pancreatic trypsin inhibitor. *Nature Struct. Biol.* **2**, 211–217 (1995).

94. J. X. Zhang and D. P. Goldenberg, Amino acid replacement that eliminates kinetic traps in the folding pathway of pancreatic trypsin inhibitor. *Biochemistry* **32**, 14075–14081 (1993).

95a. J. Clarke and A. R. Fersht, Engineered disulfide bonds as probes of the folding pathway of barnase—increasing the stability of proteins against the rate of denaturation. *Biochemistry* **32**, 4322–4329 (1993).

95b. V. I. Abkevich and E. I. Shakhnovich. What can disulfide bonds tell us about protein energetics, function and folding: Simulations and bioinformatic analysis. *J. Mol. Biol.* **300**, 975–985 (2000).

96. G. H. Lorimer, A quantitative assessment of the role of the chaperonin proteins in protein folding *in vivo. FASEB J.* **10**, 5–9 (1996).

97. A. Richardson, S. J. Landry, and C. Georgopolulos, The ins and outs of a molecular chaperone machine. *Trends Biochem. Sci*. **23**, 138–143 (1998).

98. W. A. Fenton and A. L. Horwich, GroEL-mediated protein folding. *Prot. Sci*. **6**, 743–760 (1997).

99. G. H. Lorimer, Folding with a two-stroke motor. *Nature* **388**, 720–723 (1997).

100. Z. Xu and P. B. Sigler, GroEL/GroES: Structure and function of a two-stroke folding machine. *J. Struct. Biol*. **124**, 129–141 (1999).

101. C. D. Sfatos, A. M. Gutin, V. Abkevich, and E. I. Shakhnovich, Simulations of chaperone-assisted folding. *Biochemistry* **35**, 334–339 (1996).

102. H. S. Chan and K. A. Dill, A simple model of chaperonin-mediated protein folding. *Prot. Struct. Funct. Genet.* **24**, 345–351 (1996).

103. M. J. Todd, G. H. Lorimer, and D. Thirumalai, Chaperonin-facilitated protein folding: Optimization of rate and yield by an iterative annealing mechanism. *Proc. Natl. Acad. Sci. USA* **93**, 4030–4035 (1996).

104. F. J. Corrales and A. R. Fersht, Toward a mechanism of GroEL–GroES chaperone activity: An ATPase-gated and pulsed folding and annealing cage. *Proc. Natl. Acad. Sci. USA* **93**, 4509–4512 (1996).

105. K. Braig, Z. Otwinowski, R. Hegde, D. C. Boisvert, A. Joachimiak, A. L. Horwich, and P. B. Sigler, The crystal structure of the bacterial chaperonin at 2.8 Å. *Nature* **371**, 578–586 (1994).

106. J. F. Hunt, A. J. Weaver, S. J. Landry, L. Gierasch, and J. Deisenhofer, The crystal structure of the GroES co-chaperonin at 2.8 Å resolution. *Nature* **379**, 37–49 (1996).

107. Z. Xu, A. Horwich, and P. B. Sigler, The crystal structure of the asymmetric GroEL–GroES–$(ADP)_7$ chaperonin complex. *Nature* **388**, 741–750 (1997).

108. J. Ma, P. B. Sigler, Z. H. Xu, and M. Karplus, A dynamic model for allosteric mechanism for GroEL. *J. Mol. Biol*. **302**, 303–313 (2000).

109. O. Yifrach and A. Horovitz, Nested cooperativity in the ATPase activity of the oligomeric chaperonin GroEL. *Biochemistry* **34**, 5303–5308 (1995).

110. D. Thirumalai and G. H. Lorimer, Chaperonin-mediated protein folding. *Annu. Rev. Biophys. Biomol. Struct*. **30**, 245–269 (2001).

111. P. V. Viitanen, A. A. Gatenby, and G. H. Lorimer, Purified chaperonin 60 (GroEL) interacts with the non-native states of a multitude of *Escherichia coli* proteins. *Prot. Sci.* **1**, 363–369 (1992).

112. L. Chen and P. B. Sigler, The crystal structure of a GroEL/peptide complex: Plasticity as a basis for substrate diversity. *Cell* **99**, 757–768 (1999).

113. J. Chatellier, A. M. Buckle, and A. R. Fersht, GroEL recognizes sequential and nonsequential linear structural motifs compatible with extended β-strands and α-helices. *J. Mol. Biol.* **292**, 163–172 (1999).

114. D. Thirumalai, Theoretical perspectives on *in vitro* and *in vivo* folding, in S. Doniach, editor, *Statistical Mechanics, Protein Structure, and Protein–Substrate Interactions*, Plenum, New York, 1994, pp. 115–134.

115. K. Gulukota and P. G. Wolynes, Statistical mechanics of kinetic proof reading in protein folding *in vivo. Proc. Natl. Acad. Sci. USA* **91**, 9292–9296 (1994).

116. O. Yifrach and A. Horovitz, Coupling between protein folding and allostery in the GroE chaperonin system. *Proc. Natl. Acad. Sci.* **97**, 1521–1524 (2000).

117. J. Monod, J. Wyman, and J. P. Changeaux, On the nature of allosteric interactions: A plausible model. *J. Mol. Biol.* **12**, 88–118 (1965).

118. R. Zahn, S. Perrett, G. Stenberg, and A. R. Fersht, Catalysis of amide proton exchange by the molecular chaperones GroEL and SecB. *Science* **271**, 642–645 (1996).

119. S. E. Nieba-Axmann, M. Ottinger, K. Wuthrich, and A. Pluckthun, Multiple cycles of global unfolding of GroEL-bound cyclophilin A evidenced by NMR. *J. Mol. Biol.* **271**, 803–818 (1997).

120. M. Shtilerman, G. H. Lorimer, and S. W. Englander, Chaperonin function: Folding by forced unfolding. *Science* **284**, 822–825 (1999).

121. T. E. Fisher, P. E. Marszalek, and J. M. Fernandez, Stretching single molecules into novel conformations using the atomic force microscope. *Nat. Struct. Biol.* **9**, 719–724 (2000).

122. D. K. Klimov and D. Thirumalai, Native topology determines force-induced unfolding pathways in globular proteins. *Proc. Natl. Acad. Sci. USA* **97**, 7254–7259 (2000).

123. G. W. Farr, K. Furtak, M. B. Rowland, N. A. Ranson, H. R. Saibil, T. Kirchhausen, and A. L. Horwich, Multivalent binding of non-native substrate proteins by the chaperonin GroEL. *Cell* **100**, 561–573 (2000).

124. A. Vinckier, P. Gervasoni, F. Zaugg, U. Ziegler, P. Lidner, P. Groscurth, A. Pluckthun, and G. Semenza, Atomic force microscopy detects changes in the interaction forces between GroEL and substrate proteins. *Biophys. J.* **74**, 3256–3263 (1998).

125. M. B. Viani, L. I. Pietrasanta, J. B. Thompson, A. Chand, I. C. Gebeshuber, J. H. Kindt, M. Richter, H. G. Hansma, and P. K. Hansma, Probing protein–protein interactions in real time. *Nature Struct. Biol.* **7**, 644–647 (2000).

126. R. Sayle and E. J. Milner-White, Rasmol: Biomolecular graphics for all. *Trends Biochem. Sci.* **20**, 374–376 (1995).

127. W. A. Fenton, Y. Kashi, K. Furtak, and A. L. Horwich, Residues in chaperonin GroEL required for polypeptide binding and release. *Nature* **371**, 614–619 (1994).