Theoretical studies of protein-folding thermodynamics and kinetics

Eugene I Shakhnovich

Recently, protein-folding models have advanced to the point where folding simulations of protein-like chains of reasonable length (up to 125 amino acids) are feasible, and the major physical features of folding proteins, such as cooperativity in thermodynamics and nucleation mechanisms in kinetics, can be reproduced. This has allowed deep insight into the physical mechanism of folding, including the solution of the so-called 'Levinthal paradox'.

Address
Department of Chemistry and Chemical Biology, Harvard University,
12 Oxford Street, Cambridge, MA 02138, USA

Current Opinion in Structural Biology 1997, 7:29–40

Electronic identifier: 0959-440X-007-00029

Abbreviations

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
CI2	chymotrypsin inhibitor 2
DP	designability principle
HP	hydrophobic-polar
MC	Monte Carlo
RC	reaction coordinate
RHP	random heteropolymers
T-jump	temperature jump
TS	transition state

Introduction

It is commonplace to begin a review or a paper about protein folding with the phrase 'despite all the efforts... understanding of protein folding mechanism remains elusive'. It is the purpose of the present review to claim the opposite: thanks to the efforts of many workers in the field, both experimentalists and theoreticians, we are reaching a better understanding of protein folding. The pieces of the folding puzzle are beginning to fit together into a meaningful picture of the physical mechanisms "that govern folding of polypeptide chains" [1].

Considerable progress has been made in the past year in both theoretical and experimental studies of protein folding. The signature of the present state of the field is a remarkable convergence (and interaction) between theory and experiment. In this review I will discuss the developments in the field from a theorist's perspective as the experimental work is reviewed by others in this issue, for example, see A Fersht (pp 3–9) and W Eaton (pp 10–14).

Protein-folding models

Theoretical studies of protein folding have focused on a number of issues: first, what are the sequence requirements for proteins to fold rapidly and be stable in their native conformations? Second, what are the thermodynamic mechanism(s) of protein stabilization and the kinetic mechanism(s) of folding? Third, are there special native structures (structural motifs) that are more likely to correspond to the native structures of foldable proteins? Fourth, what is the best approximation for protein-folding energetics (potentials)? These are interrelated topics, which makes the division somewhat arbitrary but it can serve as a useful framework for discussion. In this review I will focus mostly on the first and second points. While a number of important papers addressing the third [2••,3,4•,5•] and fourth points [6•,7,8•,9•,10,11] have appeared recently, space does not allow the provision of a consistent discussion of these important works.

From the beginning, the theoretical study of protein folding has relied heavily on computer simulations, although important analytical studies have been carried out as well. The early effort to model protein folding attempted numerically to represent real proteins, and the interactions between their components in the greatest possible detail. We can describe this as a 'top-down' approach. Through their realism, these models sought to reduce the likelihood of neglecting features crucial to the folding process and hence their great appeal. Protein folding occurs, however, on timescales that are computationally unreachable via top-down simulations; therefore, such detailed models cannot be used to study folding, either now or in the foreseeable future.

To circumvent the computational barrier, an alternative approach proceeds from the bottom up. It starts from the simplest model that still bears some resemblance to a protein, while being complex enough to pose nontrivial theoretical questions and having the potential to reproduce certain fundamental aspects of protein folding. Examples of this strategy are analytical studies using heteropolymer 'beads on a string' models [12–17], simulations using lattice [18–20,21•,22•], or off-lattice models [23,24,25•]. Moreover, in order to be useful, folding simulations must allow for a large number of runs, including many folding–unfolding events, to permit distinguishing between intrinsic features and statistical fluctuations. At present, such simplified models appear to be the only candidates for the computational study of protein folding.

However, one should also be cautious in using simplified protein models. The 'art' in using such models amounts to being able to distinguish which insights one should expect from them and which results may be a bit of a stretch or an artifact. It is important to appreciate that simplified models provide a coarse-grained description, and, as such, they may be adequate to describe effects taking place on longer than microscopic timescales and distances (1 ms and ≥10 Å, respectively). In other words, cubic or square lattice models are too crude to faithfully reproduce short-scale structural and chemical details of protein structure, such as the location and size distribution of secondary structure, similarities between lattice 'sequences' and protein sequences etc. At the most, such similarities can be superficial. An example of an overinterpretation of lattice model results is given in a recent paper [26], in which the authors discuss the so-called 'designability principle' (DP). The DP hypothesis states that the observed architectures of natural proteins have evolved because they can be encoded by large number of sequences. The DP was suggested in 1993 by Finkelstein *et al.* [27], on the basis of a simple analytical theory of heteropolymer thermodynamics. It was reinvented in [26] on the basis of the observation that some 27-mer structures can be encoded by a greater number of 'two-letter' sequences (that have a very special interaction potential) than other 27-mer structures. In an imaginative extrapolation the authors of [26] further speculated that the 27-mer structures which are more encodable have 'secondary structure' typical to that which one finds in natural proteins. As the physical reason for the observed behavior was not given, it is not at all clear whether this conclusion is an artifact of the special interaction scheme for the two-letter model and other particulars of their model (lattice type, coordination number etc).

While simple models are unlikely to depict all the details of protein structure, when properly formulated, they can reproduce most of the essential aspects of the protein-folding phenomenon: unique native structure (i.e. only one conformation as the global energy minimum); a large number of conformations (the 'Levinthal paradox'), and fast folding to the native state at conditions in which the native state is thermodynamically stable; and a cooperative-folding (first-order like) transition, occurring at the level of domains (independently folding units with 50–100 amino acids).

The requirement of cooperativity is necessary to reproduce the most universal feature of the thermodynamics of real proteins [28,29•]. The cooperative character of a folding transition has crucial implications for protein stability and for folding kinetics (see below). Therefore, any model of protein folding (simplified or 'realistic') which claims any relation to reality should reproduce this particular feature of folding thermodynamics.

Given that the interactions between model 'amino acids' are drastically schematized, it becomes crucial to rule out features of the observed behavior that are merely a consequence of the details of this schematization, in particular, the specific values used for pairwise interactions, the so-called 'parameter set'. This question was addressed in [30,31••,32]. It was shown that while the actual details of folding sequences depend on the potentials used, the generic features of a folding mechanism in the model structure, such as cooperativity, nucleation and even the location of the folding nucleus (see below), do not depend on the particular parameter set [31••]. Different models (and parameter sets), however, may lead to somewhat different energetic properties of model proteins that, in turn, can provide greater or lesser stability to native conformation. Moreover, in some models, the native state may not be stable at all. An example of this kind is the so-called 'HP' model in which amino acids can be of two types only: 'hydrophobic' and 'polar'. While such HP models have the clear advantage of their utmost simplicity (only one energetic parameter is involved — the energy of interaction between hydrophobic groups), they fail the very first 'feasibility' test: always more than one conformation (for 3D chains of reasonable length of 30–80 amino acids) correspond to the global energy minimum, that is, the native state is not unique [33]. It was argued in [34] that such a multiplicity of global energy minima in the HP model reproduces small deviations (e.g. small loop fluctuations) from the one unique native conformation observed in real proteins. This argument would have been convincing if the global minimum conformations in HP models were structurally similar; however, this is not the case: global minimum conformations in their model have entirely different structures [33].

It turns out that the degeneracy of the global energy minimum for random sequences in HP models makes such models 'undesignable', in other words, no sequence (whether randomly chosen or designed) can have a unique global energy minimum. The reason for this was explained in [33] in terms of energy-ladder diagrams. The factors that affect the degeneracy of the ground state in two-letter models were also addressed analytically in [35]. The degeneracy of the ground state conformations and the resulting 'undesignability' are specific to HP models; they are not present in other models which consider many types of monomers and/or different than HP interaction schemes [20,36•]. One other crucial shortcoming of HP models is that they fail to exhibit the cooperative-folding transition observed in both real proteins and many-letter models [20,28,37•].

Cooperativity of protein folding: a sequence specific feature

Earlier analytical studies of the thermodynamics of random heteropolymers (RHP) [12,13] showed that whereas they can undergo a folding transition into a unique conformation, this transition is not cooperative from a thermodynamics standpoint; in other words, it does not have a latent heat as it takes place over a wide temperature range. This theoretical prediction was confirmed by simulations [20,38] and experiments [39], and this feature of RHP renders it a poor model of protein-folding thermodynamics (and hence dynamics).

Protein-like models should account for the possibility of evolutionary sequence selection. It was pointed out in [40,41] that sequences that have 'unusually' low energies in their native conformation (i.e. much lower than the native energy of a typical random sequence) fold cooperatively. The simple qualitative explanation for this is given in Figure 1 of [42], and a detailed explanation is given in [43]. Briefly, the reason why low energy sequences fold cooperatively is because when the native state is well separated in energy from the bulk of misfolded conformations, the transition occurs between a free-energy minimum corresponding to the low energy native conformation and a free-energy minimum to which many higher energy denatured conformations belong (entropic advantage). At the same time, partly folded conformations are thermodynamically unfavorable: their energy is much higher than the energy of the native state but their number is still too small (for 3D models) to considerably contribute entropically. As a result, such intermediate conformations are higher in free energy than both the native and denatured states, that is a free-energy barrier between them exists.

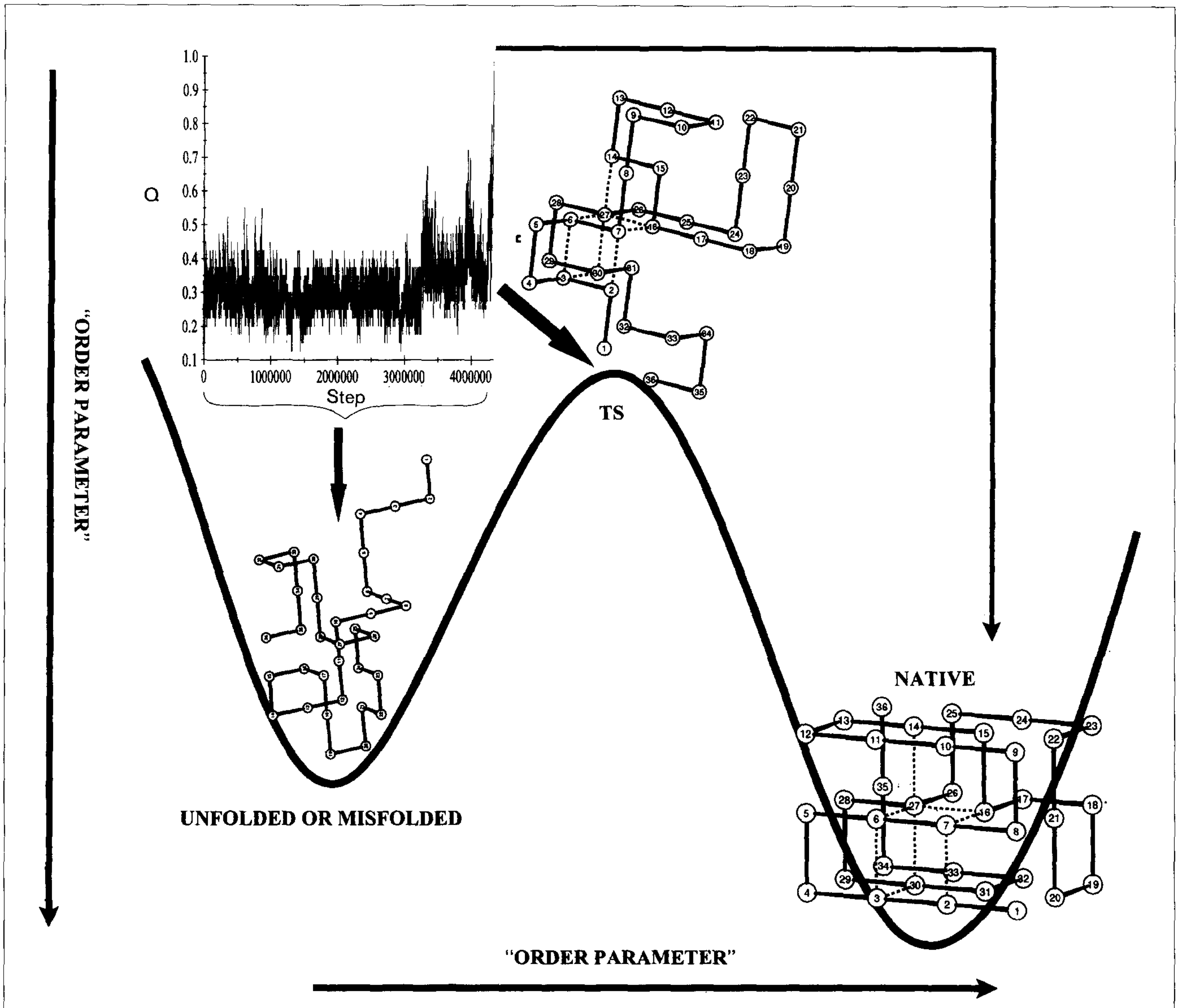
A more rigorous analytical study [16,17] verified the basic conclusions of the simple analysis presented in [42], with one important caveat (which was also mentioned in [42]): cooperative-folding transitions are only possible in three-dimensions. The physical explanation of this was given in [2••]: in two dimensions, polymeric bonds impose too strong restrictions on conformational freedom so that a significant fraction of all the contacts in compact 2D polymers are formed by monomers that are close in sequence (the so-called 'local' contacts). Structures that are formed predominantly by local contacts are called 'crumpled globules' [44]; most 2D compact polymers and a small fraction of 3D compact polymers are 'crumpled' globules (for an explanation see the Appendix in [2••]). Many partly folded conformations exist that share structural features with a particular (native) crumpled globule conformation: one can locally unfold, for example, half of a crumpled globule leaving the other half intact, because different fragments of the sequence are also spatially separated in crumpled globule substructures. This factor makes the thermodynamic properties of sequences that fold into crumpled globule conformation very different from those of normal 3D structures, and has profound implications for both their folding and their design [2••,3,45•]. Most importantly, there is no cooperative-folding transition for such structures [2••], as entropy (which is the logarithm of the number of conformations) is, in this case, a convex function of the energy. The latter is the signature of non-first-order transition [46,47]. This calls for caution in interpreting the results from 2D models [21•,48]. In my view, those working with 2D models must make it a high priority to give solid evidence for cooperative folding.

Such evidence may include bimodal histograms that show the distribution P(Q) of a folding-order parameter Q obtained from long-equilibrium simulations. The folding-order parameter Q may be defined, for example, as the fraction of contacts common to both the current conformation and the native state [19]. In such analyses, one should clearly distinguish between physical cooperativity and possible lattice artifacts. For example, lattice constraints completely rule out certain values of Q that are close to 1. Hence, for lattice models, a more reliable indicator of cooperativity would be a sharp minimum in the P(Q) histogram at Q values less than about 0.8, for which the lattice allows, in principle, a vast multitude of conformations. In this case, the existence of the minimum in the P(Q) distribution signals the existence of a free-energy barrier between native and unfolded states.

Another aspect of cooperativity which is not readily obvious from simplified thermodynamic analyses [41,42] concerns the behavior of longer chains. It is well known from experiments [29•] that longer proteins may have more complex folding–unfolding transitions because they may have several thermodynamic domains that unfold autonomously. Such behavior has been observed in simulations of longer chains [49]. It was shown that some sequences of 48 monomers fold cooperatively (shown by the fact that only the unfolded and the native states are significantly populated at temperatures near transition) whereas other sequences, designed to fold to the same native conformation, have an equilibrium intermediate, that is, a populated partly folded state. Similar results were obtained in the study of longer chains on the lattice in a different model, in which native contacts are assumed *a priori* to be more attractive [5•]. Theoretical analysis [37•] showed that the degree of cooperativity of the folding transition in longer chains may be modulated via an additional energetic parameter, δ, that presents heterogeneity of interaction energies among contacts formed in the native structure. Using this observation, an advanced sequence design procedure was proposed in [37•] that consistently allowed the design of longer cooperatively folding sequences.

The reason why cooperativity is a crucial feature of folding proteins was pointed out in [50,51] and has recently been analyzed in more detail [2••]. The folding properties of two sequences were compared in [2••]: one sequence folding to the 'crumpled' globule native state without cooperative transition; and another folding cooperatively to the 'normal' 3D native conformation, having a large number of nonlocal contacts. It was found that both

Figure 1



Schematic, highly oversimplified representation of the free-energy profile of cooperative protein folding. The insert shows a typical folding simulation MC trajectory for 36-mer lattice model [30]. The horizontal axis is the MC step and the vertical axis is the degree of folding measured as parameter Q. The parts of the folding trajectory (simulation) corresponding to fluctuations in the unfolded state, TS and descent to the native state are shown by arrows. Native conformation as well as examples of structures, typical to the unfolded and the TS are shown schematically.

sequences were able to find their native conformations. Real proteins, however, must not only find their native conformations but remain in them for a long time, in other words, they must fold at the conditions in which their native states are thermodynamically stable. To this end, it was found that more cooperative folding (as determined by the narrowness of the transition region in the temperature scale) corresponded to faster folding at temperatures in which the native states is thermodynamically stable. In contrast, sequences having a noncooperative-folding transition fold very slowly at the condition in which their native states is stable (a very low temperature is required to stabilize native conformation in this case) [2••,52•]. This factor is most crucial for HP models, which do not have a cooperative-folding transition and 'stability gaps'. Under the conditions at which the native state is stable, the folding of short 2D HP chains taken much longer than 'Levinthal time' [53,54]. In other words, for such models, an exhaustive conformational search is faster than simulating folding at conditions in which their native conformation is stable.

Implications of folding-cooperativity for kinetics: nucleation, intermediates and all that

The thermodynamics and kinetics of folding of small disulfide-free proteins follow a two-state mechanism

[55–61] which makes such proteins very attractive as experimental models from a theoretical standpoint. While they probably do not tell the whole story, their study is important because they yield a 'minimalistic' answer to how the folding problem can be solved by nature, and are thus likely to point out to necessary key elements of the folding mechanism. The folding of longer proteins may incorporate these key elements (as well as adding important new features, such as multidomain behavior [see above], intermediates and/or kinetic traps [62••,63–65]). It is reasonable to expect that an understanding of the folding of large complicated proteins would be impossible before attaining a solid understanding of the folding of the simplest ones.

The essential features of the 'free-energy landscape' for small proteins can be presented in a schematic 1D diagram (Fig. 1), which has a number of immediate implications.

What happens during the ultrafast stages?

The double-well free-energy profile suggests that the folding dynamics features two characteristic times: the relaxation time for the motion of a polypeptide chain in the free-energy minimum corresponding to the unfolded state; and the characteristic time to overcome the free-energy barrier between unfolded and folded states. The latter is identified with the experimentally observed folding time. An interesting implication of this feature is the prediction of the character of relaxation after an abrupt change of conditions from those favoring folding to those favoring unfolding (e.g. a T-jump for cold-denatured proteins [66•], or ultrafast mixing [67•]). Indeed, an ultrafast jump leads to 'instant' deformation of the free-energy profile making the 'unfolded' well higher than the folded well. Instantly, relaxation takes place adjusting to the new conditions in the 'unfolded' well [68•,69•]. This process is barrier free, and its analysis belongs to the realm of polymer dynamics [67•,69•,70,71•]. Much slower relaxation then takes place after a protracted lag, due to crossing of the barrier to the native state. This was indeed observed in recent experiments [66•,72•], in which the dynamics of folding was monitored after ultrafast laser triggering using a T-jump of cold-denatured species.

Intermediates: do proteins need them?

The fact that denatured and unfolded states are separated by the barrier of a first-order-like transition implies that there may be no (and there need not be) structural similarity between the unfolded state and the native state. Fersht [62••] analyzed general features of folding thermodynamics and kinetics on the basis of diagrams analogous to the one shown in Figure 1. He concluded that the formation of strong contacts in the unfolded state decreased its free energy, which may, under certain circumstances, increase the kinetic-folding barrier. The implication is that it may be beneficial to the stability and folding of a protein to retain as little as possible similarity to the native conformation in the unfolded state.

Similar points were made in theoretical papers [68•,73••], in which different simulation conditions (the absence or presence of average attraction between monomers [68•]) or sequences designed to fold into the same structure (but using different design techniques which yielded a different stability of misfolded conformations [73••]) yielded different folding scenarios excluding and including intermediates. It was found that folding is generally faster, and the native state is more stable for sequences that fold without detectable intermediates, in other words, via a simple two-state mechanism. It was pointed out in [68•] that, on the one hand, the formation of a compact intermediate decreases the entropic cost of the subsequent stages, and this is indeed advantageous. On the other hand, compactness also induces numerous interactions in the methastable intermediate, some of them inevitably non-native ones, which are not present in the transition state (TS). On balance, it may be favorable for rate optimization to eliminate intermediates.

Different authors emphasized different aspects of the intermediates: entropic advantages [74,75]; and enthalpic disadvantages [29•,62••,68•,73••]. It is possible that the relative roles of these factors may vary from protein to protein, and may depend on the chain length and folding temperature. The latter point is clear from the analysis of the entropic and energetic contributions to the free-energy barrier of folding [32,36•,43]. At high temperature, the barrier is mostly entropic, as enthalpically the TS is more favorable than the unfolded state. At higher temperatures, in which the entropic contribution to the free energy is relatively more pronounced, a partly structured (i.e. low entropy) intermediate may indeed result in faster folding. But at lower temperatures, in which the enthalpic contribution to the free-energy barrier becomes dominant [36•,43], a low energy intermediate would retard the folding, and hence would be evolutionarily disfavored.

What makes protein sequences fold rapidly?

The role of intermediates is the one aspect of the general problem of which factors determine the fast folding of model and of real protein sequences. It was suggested in [41], and shown for a simple model in [76] that a large energy gap between the native conformation and the closest to it in energy but structurally unrelated conformation may be one of the important signatures of fast-folding sequences. Subsequent studies fully confirmed this result. The *experimentum crucis* (which at present can only be carried out computationally) for this 'gap hypothesis' is to design sequences that have such large and small energy gaps, and to show that those having the larger gaps do fold more rapidly. This was done in [20], using lattice-model 80-mers, and in [23], using an off-lattice model, and the results fully agreed with the gap hypothesis.

An even more striking proof of the relevance of the 'energy gap' for folding kinetics came from a recent study [77•]

in which an evolution-like algorithm was developed to select fast-folding sequences for a lattice model. The only selection criterion used by this algorithm was that of rapid folding; no energetic criteria were used at all. It was found that, in accord with theoretical analyses [41,76], sequences that were selected as fast folders do indeed exhibit a pronounced energy gap (see Fig. 6 and the Discussion in [77•] for a detailed explanation and the correct meaning of 'energy gap').

Furthermore, careful analysis of numerous 'evolved' fast-folding sequences (V Abkevich, L Mirny and E Shakhnovich, unpublished data) showed that whereas random (i.e. nonselected) sequences folded via a collapsed 'burst-phase' intermediate, sequences that evolved to be fast-folding clearly showed a sharp two-state transition that has no burst-phase intermediates. Interestingly, the selection process eliminated the burst-phase intermediate by making local (in sequence) contacts less favorable: the algorithm tends to select sequences in which strongly attracting residues are far from each other along the sequences. Furthermore, correlations, such as the ones implied by these results, were found recently in real protein sequences [78•]: hydrophobic residues were found to be anticorrelated along sequences; in other words, hydrophobic residues tend to be further apart from each other in real protein sequences than in random ones. Making local contacts less favorable primarily destabilizes the unfolded conformation and makes the polymer chain stiffer. Stiffer chains are known to have a more pronounced cooperative-folding transition [79,80].

The results of the lattice-model analysis and the statistical analysis of real protein sequences are consistent with a general understanding of the statistical mechanics of folding, but they are in direct disagreement with recent claims that local contacts play a major role in folding kinetics [81]. The major flaw of the analysis made in [81] is that they did not consider fast folding to the stable native conformation, but were more concerned by the rate of 'hitting' the lowest energy state regardless of its stability. In contrast, an earlier paper [76], as well as a more recent paper [2••], was concerned with rapid folding under the conditions in which the native state is thermodynamically stable. It was shown that the abundance and strength of local contacts becomes irrelevant (or even counter-productive) when folding is studied under conditions in which the native state is thermodynamically stable [2••]. This is also consistent with what is found in the statistics of protein sequences, exhibiting anticorrelation of hydrophobic residues [78•], and structures [2••], exhibiting dominance of nonlocal contacts, almost always including contacts between N-terminal and C-terminal regions.

The gap criterion was also criticized by Klimov and Thirumalai [82], who defined the gap as the energy difference between the native state and the second lowest energy conformation, which typically turns out to be different from the native state by only one monomer flip. Their definition is in contrast with the more physically meaningful definition of the energy gap (or 'stability gap' [41]) as the energy difference between the native state and the lowest energy misfolded (i.e. structurally distinct from the native state) conformation [41,76,77•]. The two definitions had been compared already (see Fig. 17 in [76], which was essentially reproduced as Fig. 1 in [82]) whereby it was made clear that no correlation between the gap, as defined by Klimov and Thirumalai [82], and the folding rate can be expected. More recently, the full density of states of slow-folding and fast-folding sequences were given (see Fig. 6 in [77•]) showing a clear correlation between the folding rate and the properly defined gap. Figure 6 in [77•] explains why such a correlation should exist and why the 'gap' defined by Klimov and Thirumalai [82] is irrelevant for folding.

Klimov and Thirumalai [82] suggested a parameter, σ, as a criterion of fast folding where $\sigma = (T_\Theta - T_f)/T_\Theta$, and where sequences with a small σ fold rapidly. In their words, T_f is the temperature of the folding transition and T_Θ is the temperature of the 'collapse transition'. Klimov and Thirumalai [82] obtain T_Θ as the maximum of the curve of dependence of heat capacity versus temperature. The σ criterion, as it stands, is somewhat confusing. In all previous simulations (e.g. see [20,48]), only one peak in the plot of temperature dependence versus heat capacity was observed, which occurred at the folding-transition temperature T_f. This is fully consistent with the fact that the folding transition is cooperative. The definition of T_Θ from the heat capacity peak [82] implies either that the authors see two peaks of heat capacity (one peak at T_f, as in previous simulations, and another at $T_\Theta \neq T_f$), or that they confuse T_Θ with T_f. In the former case, an explanation is needed as to why Klimov and Thirumalai's curves of heat capacity differ from those obtained by others. In the latter case, T_Θ [82] is actually T_f, the temperature of the midfolding transition, whereas T_f is the temperature at which the transition is complete and the chain is in its native conformation, in other words, the σ criterion relates the folding rate to the width (in terms of temperature) of the folding transition. If this second interpretation of the σ criterion is correct, then it is fully equivalent to the 'gap' criterion because the gap (correctly defined) is related to the specific heat of the first-order folding transition [41,42], which in turn is related to its width via the van't-Hoff relation [28,79]. As mentioned above, the similar relation between the folding rate and cooperativity of the folding transition was previously observed and discussed in earlier publications [20,76].

Transition-state ensemble: the nucleation mechanism

As with any other system, the order of the folding transition determines its kinetic features [83]. The cooperativity of the protein-folding transition suggests that kinetically it must follow a nucleation mechanism as seen in first-order phase transitions. According to the nucleation mechanism in physical kinetics, a system fluctuates in the 'old' phase (unfolded state) until a critical fluctuation creates an island of a 'new' phase (e.g. the folded state) that is sufficiently large (or specific) to grow, proceeding downhill in free energy. The simplest manifestation of such a mechanism is vapor condensation, whereby liquid droplets spontaneously form and the ones whose sizes exceed a critical threshold, determined by the interplay of bulk and surface energies, grow further into the liquid phase. Such critical nuclei correspond to the TS for the vapor condensation reaction.

Returning to the realm of proteins, we should expect that a folding TS contains a (partly) assembled fragment of the native structure, which can be identified as a critical nucleus for folding. Such kinetic behavior was indeed found in folding simulations, and has been described in detail [30]. Furthermore, a nucleation mechanism was experimentally discovered independently, using a thorough protein-engineering analysis of the folding of a small protein, CI2 [84]. This nucleation mechanism was called 'nucleation-condensation' in [62••,84••] to distinguish it from earlier hypotheses [85,86] in which authors had speculated about the nucleation of folding via local (in sequence) contacts. The latter hypothesis was based on the assumption that such contacts would probably form early in the process of folding. The qualitative description in Figure 1 suggests that this view is dramatically different from the nucleation mechanism of a first-order folding transition. Indeed, TS conformations appear late in time, just prior to the fast descent to the native state. The analysis based on this premise allowed the identification of particular structural elements corresponding to the TS(s) in folding simulations [30]. It was pointed out in [30], and in subsequent discussions of nucleation-condensation mechanism [62••,84••], that the folding nucleus should contain at least a few nonlocal contacts corresponding to formation of long loops. The reason for this is that the nucleus is the first set of conformations, after which the chain 'descends' downhill to the native state. The nucleus is the lowest of all the free-energy barriers, that is, it is a saddle point in the free-energy landscape. This implies that the TS ensemble should not only satisfy the requirement of being relatively low in free energy, but also, more importantly, it should not with overwhelming probability descend back to the unfolded conformation or to a misfolded trap. Therefore, TS conformations must be sufficiently dissimilar to the unfolded state or traps, in other words, they must carry distinctive features of the native state that are not shared with the unfolded state. Local contacts are dominant in the unfolded state [2••,87], and some contacts are present both in the TS-ensemble conformations and in the native state. Certain nonlocal native contacts that are induced by long loops, however, play a crucial role in the TS, because after they are formed, a critical fragment of structure unique to the native state appears, and subsequent dynamics lead unidirectionally to the native state. In other words, the process of pre-TS fluctuations can be viewed as the fast formation and dissolution of numerous local contacts until certain nonlocal contacts are formed which stay intact until the native state is reached. It follows from this analysis that efforts to identify nucleation sites as being low energy local native-like elements in the denatured state [88] may be futile as these belong to the small perturbation of the unfolded state.

The issue of folding-nucleus specificity is currently of considerable interest. Two extreme possibilities were outlined in [31••]: a nonspecific nucleus model in which any (even noncontiguous) native fragments of sufficient size serve to nucleate folding; and a specific-nucleus model in which a vast majority of TS conformations share a set of 'specific contacts'. The latter possibility suggests that TS conformation(s) for each protein molecule with a given sequence (or for each 'run' of simulations) share a specific set of contacts, although each of them may also feature a number of other native contacts not found in other TS conformations. (There has been a misleading opinion that the specific-nucleus model assumes a unique TS conformation. This is not correct: TS in the specific-nucleus model represents an ensemble of conformations, which share a certain set of dominant contacts.)

The specific-nucleus model predicts the existence of 'kinetically most important' residues, the mutation of which would have the most impact on the folding rate. Simulations [31••] and experimental protein-engineering analyses [84,89•,90] show that such kinetically critical positions exist in some proteins, supporting the specific-nucleus model. A simple method was suggested in [31••] that predicts the location of kinetically important amino acids [31••]. The method provided successful predictions for CI2 [31••,84], Che Y [89•] (E Shaknovich, unpublished data) and acyl-coenzyme binding protein (F Poulsen, personal communication). However, we should note that the 'specific'-nucleus and 'nonspecific'-nucleus models are extreme cases, and it is quite probable that the real situation lies in between these and may vary from protein to protein.

Currently, protein-engineering analysis is the only experimental tool for characterizing TSs. The method evaluates how a mutation changes the free energy of thermal stabilization of a protein ($\Delta\Delta G_{eq}$) and its folding rate. The impact of mutation on the folding rate can be interpreted using the TS theory in terms of the change in the free-energy barrier ($\Delta\Delta G^{\dagger}$). The ratio of the two quantities ($\phi = \Delta\Delta G^{\dagger}/\Delta\Delta G_{eq}$) characterizes the degree to which a mutated amino acid forms its contacts in the TS ensemble. The protein-engineering analysis allows the attainment of a number of illuminating results concerning the character of the TS-ensemble in small proteins. However, this approach has its limitations, most notable of which is the assumption that mutations do not affect the structure and energetics of the unfolded state. Although this probably is the case generally, local (native and non-native) contacts are present in the unfolded state. Rigorous computational analysis (V Abkevich, A Gutin, E Shakhnovich, unpublished data) suggests that the existence of native and non-native contacts in the unfolded state can considerably affect the precision of the protein-engineering method. For example, if an amino acid from the nucleus also participates in non-native interactions in the unfolded state, its apparent ϕ-value [91] can be small. The signature of this situation is that amino acids neighboring in sequence to the one being studied have negative ϕ-values. This was the case for Ile157 from CI2, which has a low ϕ-value but participates in the nucleus, as was established by thorough analysis using double mutants [84].

A similar argument may be used to explain the results of [90], in which it was concluded that permutation of the amino acids changes the structure of the folding TS to the same native conformation. Indeed, this conclusion is based on the observation of a pronounced change in ϕ-value upon the permutation of only one mutation (Val44→Ala) in the α-spectrin domain of SH3. The permutation, however, changes the local interactions in a significant part of the sequence so that Val44 can participate in non-native local interaction in the permutants. The fact that the neighboring Lys43 acquires a negative ϕ-value in the permutant suggests that this is probably the case. A more detailed analysis, such as the one made in the study of CI2 folding [84,91], is needed to see which of the effects of permutation on the TS of SH3 are real, and indeed how the TSs of the permutants are different.

Nucleus TSs were determined from the kinetic simulations of lattice models of protein folding [30,31••] using a minimal set of contacts whose formation causes unidirectional descent to the native state. In an earlier publication [74], the TS(s) were determined from equilibrium simulations. The same approach was used in subsequent publications [92•,93•] with similar conclusions (e.g. compare Fig. 4d in [74] with Fig. 1 in [92•]). The free-energy, energy, and entropy as a function of the order parameter Q (the fraction of native contacts) were obtained from simulations in [74] using the histogram method [94]. Similar functions were also presented in a later publication [93•]. In a recent study of a more detailed folding model [95••], the free-energy profile as a function of another order parameter, the total volume V, of the globule, was obtained from equilibrium molecular-dynamics simulations.

The maximum of free energy as a function of the order parameter (Q in [74,93•]; V in [95••]) was associated with the transition region. This procedure is equivalent to projecting a multidimensional free-energy landscape onto certain 'axes' corresponding to Q or to V. Clearly, in such a procedure, the true TS region (the saddle point in the multidimensional configurational space) does not necessarily project to the maximum of the F(Q) or F(V) curve. In other words, 'the transition region' $Q \approx Q^{\dagger}$, as evaluated from the maximum of the F(Q) curve, may not contain all the conformations that have a coefficient of dynamic transmission to the native state close to 0.5. The converse is true as well, not all conformations with $Q \approx Q^{\dagger}$ have a high transmission coefficient to the native state. In this sense, the reader must be warned that the plot in Figure 1 is given for illustrative purposes only (I cannot draw multidimensional figures): the F(Q) and F(V) plots may be very valuable in evaluating the thermodynamic properties of folding transition but they may be not sufficient to identify real TS(s). The only reliable way to identify the folding TS in simulations is to derive it from kinetics analysis, that is, to find a set of conformations such that simulations starting from these conformations have, for example, 50% probability of rapidly finishing up in the native state before unfolding (V Pande, A Grosberg and E Shakhnovich, unpublished data). For this reason, a recent study on the folding of lattice 125-mers, in which a non-nucleation mechanism of folding is proposed [5], should be supplemented by a kinetic analysis similar to the one described above and carried out in [30]. The analysis of equilibrium properties is not sufficient to support or rule out nucleation or other kinetic mechanisms for the model.

I now come to the broader question of what is a good 'reaction' coordinate (RC; or order parameter [for statistical mechanicians]) to describe folding kinetics. Using Q as the RC assumes that native contacts are spread randomly in TS conformations. This is the simplest assumption, and has been was made in a number of papers (e.g. see [74,93•,96]), and though it may be acceptable as a crude initial qualitative approximation for very short chains, it may not be correct for more realistic chain lengths and for a quantitative analysis. The main difficulty is that kinetic theory based on the assumption of Q as the RC, as well on as kinetic arguments presented in [96], predicts that the folding time grows exponentially with chain length, which is in dramatic disagreement with both simulations [97•] and experiment. A recent study [97•] showed that the dependence of folding time length, τ, is a power law, $\tau(N) \sim N^{\lambda}$ both for random sequences and designed ones with the exponent $\lambda \approx 6$ for random sequences and $\lambda \approx 3.5$ for highly designed sequences, which also suggests that the difference of folding rates between random and designed sequences, at the conditions of fastest folding, becomes more pronounced as the chains get longer. This conclusion contradicts that made in a recent paper [52•], in which it was found that random and designed sequences fold equally fast at their respective folding-rate optima. Only very short sequences were studied [52•], however, which may obscure the difference between random and designed sequences.

The nucleation mechanism explains the relatively weak power-law dependence of the folding rate on length for designed sequences [97•,98]. The entropic cost of loop formation around the assembled native fragment in a TS conformation plays the role of a 'surface energy' in the nucleation mechanism of folding [97•,98]. Interestingly, loop entropy may explain the important finding that a particular permutation of the SH3 sequence leads to a much faster folding protein than the wild-type [99]. Indeed, in the fast-folding permutant (Ser19–Pro20) the N and C termini are located at a position which in the wild-type belongs to a long loop.

Conclusion

A recent review [100•] compared the 'classical view' that folding is a set of mechanistically defined steps proceeding via well defined intermediates with the 'new view' that posits that a protein is a system with many degrees of freedom, for which entropy is an essential factor in the free-energy balance and the kinetics. The classical view reflects a 'chemical' understanding of protein folding as being a complex reaction that proceeds via a mechanistically defined pathway. In contrast, the new view envisages folding as proceeding via a 'statistical pathway' that features a sequence of multiply populated, kinetically distinguishable macrostates: the unfolded state; intermediate states (if any); the TS; and finally, a more unique (lower entropy) native state. The formation of a TS conformation (containing the nucleus) is thus a statistical fluctuation that can occur in an innumerable number of ways. (Of course, the TS does not require a random multiparticle collision to form a nucleus, because amino acids that attract each other are more likely to interact in any state.) According to the new view, on the one hand, it is meaningless to ask what precise sequence of microscopic events leads to the formation of the nucleus. On the other hand, the descent from the nucleus TS to the native conformation can be a deterministic process with a markedly favored 'pathway'. However, this process, while definitely interesting, accounts for a negligable fraction of the total folding time and does not determine the folding rate.

The role of intermediates in protein folding continues to be of great interest and importance. It is interesting, however, that the discussion of this topic in the literature has taken a turn in a new direction. In contrast to the classically pervasive assumption of the necessity of intermediates for the solution of 'Levinthal's paradox', the recent experimental evidence [58,101] and theoretical analyses [62••,68•,73••] suggest that, while there may be cases in which intermediates facilitate faster folding, they do not represent a necessary or even vital feature of protein-folding dynamics, at least for small proteins. In the realm of the new view, what is required to understand basic folding kinetics is the experimental and theoretical characterization of the TS ensemble and the development of a statistical theory of how such TS(s) are reached via the thermal fluctuations of the polypeptide chain.

In conclusion, I would like to point out that recent experimental and theoretical (computational) developments have shown that if the energetics are correct (i.e. a cooperative-folding transition to a stable native state occurs, so that kinetics follow a nucleation mechanism), there is no more 'Levinthal paradox' in protein folding than there is in vapor condensation or any other first-order transition accompanied by a massive loss of entropy. Taking this into account, the nucleation mechanism of folding, when studied in more detail (including microscopic analytical theory), is likely to represent the solution of the general problem of protein-folding kinetics, at least for small proteins.

The legitimate question remains of how this understanding of folding kinetics helps to solve the most renowned aspect of the protein-folding problem — tertiary structure prediction. The answer to this is that our understanding focuses our thinking in the direction of what would be the best model that combines the right energetics with computational tractability, the one that will enable us to find native structure of a protein. At one end are simplified lattice models, which are computationally tractable, but in which the 'right' energetics are achieved by sequence design; at the other end are all-atom models (with solvent included), in which the energetics may be right (at least we know that they are right in natural proteins), but which are totally prohibitive for folding simulations. While the question of which models are best for describing protein energetics is the crucial one for structure prediction (e.g. see [9•,10,11•,102•]), its discussion may be the subject of another review.

Acknowledgements

I am deeply grateful to my colleagues and collaborators Martin Karplus, Alexander Grosberg, Oleg Ptitsyn, Alexander Gutin, Victor Abkevich, Leonid Mirny, Gabriel Berriz, Vijay Pande for their numerous insights and contributions. Oleg Ptitsyn, Victor Abkevich and Gabriel Berriz have critically read the manuscript and made many helpful illuminating comments. The financial support is provided by National Institutes of Health grant GM 52126

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Anfinsen C: **Principles that govern the folding of protein chains.** *Science* 1973, **181**:223–230.

2. •• Abkevich AI, Gutin AM, Shakhnovich EI: **Impact of local and non-local interactions on thermodynamics and kinetics of protein folding.** *J Mol Biol* 1995, **252**:460–471.
The structural basis for determining the order of folding transition is studied. For proteins with a large number of nonlocal contacts, the folding transition is first-order, whereas for ones with local contacts it is not. This difference in cooperativity is shown to have dramatic implications for folding kinetics and stability.

3. Govindarajan S, Goldstein R: **Searching for foldable protein structures using optimized energy functions.** *Biopolymers* 1995, **36**:43–51.

4. • Govindarajan S, Goldstein R: **Why are some protein structures so common?** *Proc Natl Acad Sci USA* 1995, **93**:3341–3345.
Different structures are analyzed in two papers [4•,6•] from the standpoint of their 'designability', and the ones that are more designable are attributed to commonly observed structures.

5. • Dinner A, Sali A, Karplus M: **The folding mechanism of larger model proteins: role of native structure.** *Proc Natl Acad Sci USA* 1996, **93**:8356–8361.
The folding of cubic lattice 125-mers is studied using interaction potentials that assign stronger attraction to native contacts than to non-native ones (the 'Go'-model). It is shown that the internal stable core forms first, and that the remaining part of the structure fluctuates around this, sometimes approaching but not completely reaching, the native structure.

6. • Finkelstein AV, Gutin AM, Badretdinov A: **Why are the same protein folds used to perform different functions?** *Proteins* 1995, **23**:142–149.
See annotation [4•].

7. Godzik A, Kolinski A, Skolnick J: **Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets.** *Protein Sci* 1995, **4**:2101–2117.

8. • Hao M-H, Scheraga H: **How optimization of potential function affects protein folding.** *Proc Natl Acad Sci USA* 1996, **93**:4984–4989.
A thorough study of the lattice model in which potentials are optimized to make a 'native' conformation for a given sequence a pronounced energy minimum.

9. • Thomas D, Dill KA: **Statistical potentials extracted from protein structures: how accurate are they?** *J Mol Biol* 1996, **257**:457–469.
A knowledge-based approach to deriving potentials is tested on a 2D HP model. The parameters are not completely recovered, and on the basis of this, the knowledge-based methods are criticized.

10. Myazawa S, Jernigan R: **Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading.** *J Mol Biol* 1996, **256**:623–644.

11. • Mirny L, Shakhnovich EI: **How to determine protein folding potential? A new approach to the old problem.** *J Mol Biol* 1997, in press.
A method to derive the protein-folding potential from a PDB file is suggested to be the optimization of the space parameters to make all proteins stable. The method is tested on lattice models with 20 types of amino acids in which it is shown to recover the 'true' potential with 91% accuracy. Other popular methods are also tested and evaluated.

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

13. Garel T, Orland H: **Mean-field model for protein folding.** *Europhys Lett* 1988, **6**:307–309.

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

15. Sasai M, Wolynes PG: **Unified theory of collapse, folding, and glass transitions in associative-memory hamiltonian models of proteins.** *Phys Rev A* 1992, **46**:7979–7997.

16. Ramanathan S, Shakhnovich E: **Statistical mechanics of proteins with 'evolutionary selected' sequences.** *Phys Rev E* 1994, **50**:1303–1312.

17. Pande V, Grosberg AYu, Tanaka T: **Freezing transition of random heteropolymers consisting of arbitrary sets of monomers.** *Phys Rev E* 1995, **51**:3381–3393.

18. Kolinski A, Skolnick J: **Monte-carlo simulation of protein folding. Lattice model and interaction scheme.** *Proteins* 1994, **18**:338–352.

19. Shakhnovich EI, Farztdinov GM, Gutin AM, Karplus M: **Protein folding bottlenecks: a lattice Monte-Carlo simulation.** *Phys Rev Lett* 1991, **67**:1665–1667.

20. Shakhnovich EI: **Proteins with selected sequences fold to their unique native conformation.** *Phys Rev Lett* 1994, **72**:3907–3910.

21. • Dill KA, Bromberg S, Yue K, Feibig KM, Yee DP, Thomas PD, Chan HS: **Principles of protein folding – a perspective from simple exact models.** *Protein Sci* 1995, **4**:561–602.
This paper presents a review of the studies of 2D HP models and a discussion of their relationship to protein folding.

22. • Hao M-H, Scheraga H: **Statistical thermodynamics of protein folding: comparison of a mean-field theory with Monte-Carlo simulations.** *J Chem Phys* 1995, **102**:1334–1339.
A comparison of the heteropolymer theory based on the Random Energy model with lattice simulation results is presented.

23. Irback A, Schwarze H: **Sequence dependence of self-interacting random chains.** *J Phys A* 1995, **28**:2121–2132.

24. Guo Z, Thirumalai D: **Nucleation mechanism for protein folding and theoretical predictions for hydrogen-exchange labeling experiments.** *Biopolymers* 1995, **35**:137–139.

25. • Berriz G, Gutin A, Shakhnovich E: **Langevin model for protein folding: cooperativity and stability.** *J Chem Phys* 1997, in press.
This paper presents simulations of an off-lattice folding model and an evaluation of cooperativity of the folding transition of different models.

26. Li H, Winfreen N, Tang C: **Emergency of preferred structures in a simple model of protein folding.** *Science* 1996, **273**:666–669.

27. Finkelstein AV, Gutin A, Badretdinov A: **Why are some protein structures so common?** *FEBS Lett* 1993, **325**:23–28.

28. Privalov PL: **Stability of proteins. Small single-domain proteins.** *Adv Prot Chem* 1979, **33**:167–170.

29. • Privalov PL: **Intermediate states in protein folding.** *J Mol Biol* 1996, **258**:707–725.
A review of experimental results is presented that has the conclusion that intermediates may not be essential for fast folding and stability.

30. Abkevich VI, Gutin AM, Shakhnovich EI: **Specific nucleus as the transition state for protein folding: evidence from the lattice model.** *Biochemistry*, **33**:10026–10036.

31. •• Shakhnovich EI, Abkevich VI, Ptitsyn OB: **Conserved residues and the mechanism of protein folding.** *Nature* 1996, **379**:96–98.
A method is presented to predict the most kinetically important residues in real proteins, which yields a successful blind prediction of the key nucleation residue in CI2.

32. Socci N, Onuchic JN: **Kinetics and thermodynamic analysis of protein-like heteropolymer: Monte Carlo histogram technique.** *J Chem Phys* 1995, **103**:4732–4744.

33. Yue K, Fiebig K, Thomas P, Chan H-S, Shakhnovich EI, Dill KA: **A test of lattice protein folding algorithms.** *Proc Natl Acad Sci USA* 1995, **92**:325–329.

34. Chan H-S, Dill KA: **Comparing folding codes of proteins and copolymers.** *Proteins* 1996, **24**:335–344.

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

36. • Betancourt M, Onuchic JN: **Kinetics of protein-like models: the energy landscape factors that determine folding.** *J Chem Phys* 1995, **103**:773–787.
A 2D-folding model is studied in detail, including the features of sequences and potentials that affect folding and stability.

37. • Abkevich VI, Gutin AM, Shakhnovich EI: **Improved design of stable and fast-folding proteins.** *Fold Des* 1996, **1**:221–232.
This paper presents an improved version of the MC design in sequence space that allows the design of long sequences having cooperative (single-domain) behavior.

38. Shakhnovich EI, Gutin AM: **Enumeration of all compact conformations of copolymers with quenched disordered sequence of links.** *J Chem Phys* 1990, **93**:5967–5971.

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

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

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

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

43. Abkevich VI, Gutin AM, Shakhnovich EI: **Free energy landscape for protein folding kinetics. Intermediates, traps and multiple pathways in theory and lattice model simulations.** *J Chem Phys* 1994, **101**:6052–6062.

44. Grosberg AYu, Nechaev SK, Shakhnovich EI: **The role of topological constraints in the kinetics of collapse of macromolecules.** *J Phys (France)* 1996, **49**:2095–2100.

45. • Pande V, Grosberg AYu, Joerg C, Tanaka T: **Is heteropolymer freezing well described by the random energy model?** *Phys Rev Lett* 1996, **76**:3987–3990.

A computational analysis of several popular heteropolymer models is carried out using exhaustive enumeration of conformations for cubic-lattice models.

46. Go N: **Theoretical studies of protein folding.** *Annu Rev Biophys Bioeng* 1983, **12**:183–210.

47. Hao M-H, Scheraga H: **Monte-Carlo simulation of a first order transition for protein folding.** *J Phys Chem* 1994, **98**:4940–4945.

48. Fukugita M, Lancaster D, Mitchard M: **A heteropolymer model study for the mechanism of protein folding.** *Biopolymers* 1997, in press.

49. Abkevich VI, Gutin A, Shakhnovich EI: **Domains in folding of model proteins.** *Protein Sci* 1995, 4:1167–1177.

50. Karplus M, Shakhnovich E: **Protein folding: theoretical studies of thermodynamics and dynamics.** In *Protein Folding.* Edited by Creighton T. New York: WH Freeman and Company; 1992:127–196.

51. Ptitsyn OB: **The molten globule state.** In *Protein Folding.* Edited by Creighton T. New York: WH Freeman and Company; 1992:243–300.

52. • Galzitskaya O, Finkelstein AV: **Folding of chains with random and edited sequences: similarities and differences.** *Protein Eng* 1995, **8**:883–892.

Folding of short chains with exhaustively enumerated conformations is studied using MC simulations. The authors assert that random and 'edited' sequences have the same optimal rates of folding.

53. Chan H-S, Dill KA: **Transition states and folding dynamics of proteins and heteropolymers.** *J Chem Phys* 1994, **100**:9238–9257.

54. Camacho C, Thirumalai D: **Modeling the role of disulfide bonds in protein folding: entropic barriers and pathways.** *Proteins* 1995, **22**:27–40.

55. Jackson SE, Fersht A: **Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition.** *Biochemistry* 1991, **30**:10428–10435.

56. Alexander A, Orban J, Bryan P: **Kinetic analysis of of folding and unfolding of 56 amino acid IGg-binding domain of *streptococcal* protein g.** *Biochemistry*, **31**:7243–7248.

57. Viguera A-R, Martinez JC, Filimonov VV, Mateo PL, Serrano L: **Thermodynamic and kinetic analysis of the SH3 domain of spectrin shows a two-state folding transition.** *Biochemistry* 1994, **33**:2142–2150.

58. Schindler T, Herrler M, Marahiel M, Schmidt F-X: **Extremely rapid protein folding in the absence of intermediates.** *Nat Struct Biol* 1995, **2**:663–673.

59. Kragelund BB, Robinson CV, Knudsen J, Dobson CM: **Folding of a four-helix bundle: studies of acetyl-coenzyme a binding protein.** *Biochemistry* 1995, **34**:7117–7124.

60. Englander W, Sosnick TR, Mayne L, Hiller R, Englander S: **The barriers in protein folding.** *Nat Struct Biol* 1994, 1:149–156.

61. Huang G, Oas T: **Submillisecond folding of monomeric λ-repressor.** *Proc Natl Acad Sci USA* 1995, **92**:6878–6882.

62. •• Fersht AR: **Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications.** *Proc Natl Acad Sci USA* 1995, **92**:10869–10873.

An insightful discussion of the mechanism(s) of cooperative-folding transitions and the factors affecting rate and stability of folding is presented. One of the main conclusions is that intermediates may slow down folding and decrease stability. Arguments are presented suggesting how a nucleation mechanism may 'solve' the dynamic-folding problem.

63. Radford S, Dobson C, Evans P: **The folding of hen lysozyme involves partially structured intermediates and multiple pathways.** *Nature* 1992, **358**:302–307.

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

65. Radford S, Dobson CM: **Insight into protein folding using physical techniques: studies of lysozyme and α-lactalbumin.** *Phil Trans R Soc London B* 1995, **348**:17–25.

66. • Ballew RM, Sabelko A, Gruebele M: **Direct observation of fast protein folding: initial collapse of myoglobin.** *Proc Natl Acad Sci USA* 1996, **93**:5759–5764.

One of the first observations of protein folding in an extended time range from microseconds to seconds.

67. • Hagen S, Hofrichter A, Szabo A, Eaton W: **Diffusion-limited contact formation in unfolded cytochrome c: estimating the maximum rate of protein folding.** *Proc Natl Acad Sci USA* 1996, **93**:11615–11617.

The relaxation processes in cytochrome *c* after ultrafast mixing are studied and the rate of diffusion-limited polymer collapse is estimated.

68. • Gutin AM, Abkevich VI, Shakhnovich EI: **Is burst hydrophobic collapse necessary for rapid folding?** *Biochemistry* 1995, **34**:3066–3076.

Two folding scenarios are simulated, including and excluding the burst intermediate. It is shown that the existence of burst intermediate does make folding faster. Different aspects of the burst intermediates are discussed in detail.

69. • Sosnick TR, Mayne L, Englander SW: **Molecular collapse: the rate-limiting step in two-state cytochrome c folding.** *Proteins* 1996, **24**:417–426.

Interesting experimental evidence and analysis in support of a nucleation mechanism are presented.

70. Doi M Edwards SF: *The Theory of Polymer Dynamics.* Clarendon Press: Oxford; 1986.

71. • Roan J-R, Shakhnovich EI: **Dynamics of heteropolymers in dilute solution: effective equation of motion and relaxation spectrum.** *Phys Rev E* 1996, **54**:5340–5357.

An analytical study of heteropolymer dynamics in the microscopic polymer model is presented. It is shown that there is no analogy between spin glasses and heteropolymers as far as kinetics is concerned.

72. • Nolting B, Golbik R, Fersht AR: **Submillisecond events in protein folding.** *Proc Natl Acad Sci USA* 1995, **92**:10668–10672.

73. •• Mirny LA, Abkevich VI, Shakhnovich EI: **Universality and diversity of the protein folding scenarios: a comprehensive analysis with the aid of lattice model.** *Fold Des* 1996, **1**:103–116.

This paper presents a systematic detailed comparison of folding mechanisms for two sequences designed to fold into the same lattice conformation. One sequence (Seq2) folds via burst intermediates and encounters traps, while the other (Seq1) folds in a two-state manner. It is found that the kinetic intermediate of the Seq2 is similar to the equilibrium intermediate that is present for this sequence at intermediate temperature.

74. Sali A, Shakhnovich EI, Karplus M: **How does a protein fold?** *Nature* 1994, **369**:248–251.

75. Karplus M, Sali A: **Theoretical study of protein folding and unfolding.** *Curr Opin Struct Biol* 1995, **5**:58–73.

76. Sali A, Shakhnovich EI, Karplus M: **Kinetics of protein folding. A lattice model study for the requirements for folding to the native state.** *J Mol Biol* 1994, **235**:1614–1636.

77. • Gutin AM, Abkevich VI, Shakhnovich EI: **Evolution-like selection of fast-folding model proteins.** *Proc Natl Acad Sci USA* 1995, **92**:1282–1286.

An algorithm which selects fast-folding sequences of model proteins is presented. It is shown that evolved fast-folding sequences have their native conformations as pronounced energy minima.

78. • Irback A, Peterson C, Potthast M: **Evidence for nonrandom hydrophobicity structures in protein chains.** *Proc Natl Acad Sci USA* 1996, **93**:9533–9538.

This paper presents a solid statistical evidence that hydrophobic amino acids are anticorrelated along protein sequences.

79. Lifshitz IM, Grosberg AYu, Khohlov AR: **Some problems of statistical physics of polymers with volume interactions.** *Rev Mod Phys* 1978, **50**:683–713.

80. Sfatos CM, Gutin AM, Shakhnovich EI: **Phase diagram of random copolymers.** *Phys Rev E* 1993, **48**:465–475.

81. Unger R, Moult J: **Local interactions dominate folding in a simple protein model.** *J Mol Biol* 1996, **259**:988–994.

82. Klimov D, Thirumalai D: **A criterion which determines foldability of proteins.** *Phys Rev Lett* 1996, **76**:4070–4073.

83. Lifshits EM, Pitaevskii LP: *Physical Kinetics.* Oxford: Pergamon; 1981.

84. Itzhaki L, Otzen D, Fersht AR: **The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding.** *J Mol Biol* 1995, **254**:260–288.
••
A first comprehensive protein engineering analysis of folding is presented, which provides strong evidence in favor of the nucleation-condensation mechanism.

85. Wetlaufer D: **Nucleation, rapid folding and globular intrachain regions in proteins.** *Proc Natl Acad Sci USA* 1973, **70**:697–701.

86. Moult J, Unger R: **An analysis of protein folding pathways.** *Biochemistry* 1991, **30**:3816–3824.

87. Grosberg AYu, Khohlov AR: *Statistical Mechanics of Macromolecules.* New York: AIP Press; 1994.

88. Avbelj F, Moult J: **Determination of the conformation of folding initiation sites in proteins by computer simulations.** *Proteins* 1995, **23**:129–141.

89. Lopez-Hernandez E, Serrano L: **Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI2.** *Fold Des* 1996, 1:43–55.
•
A protein engineering analysis of the folding of a longer protein that has a folding intermediate.

90. Viguera AR, Serrano L, Wilmanns M: **Different folding transition states may result in the same native structure.** *Nat Struct Biol* 1996, **3**:874–880.

91. Matouschek A, Kellis J Jr, Serrano L, Bycroft M, Fersht AR: **Transient folding intermediates characterized by protein engineering.** *Nature* 1990, **346**:440–445.

92. Wolynes PG, Onuchic JN, Thirumalai D: **Navigating the folding routes.** *Science* 1995, **267**:1619–1620.
•
A perspective article that summarizes earlier findings from simplified folding models.

93. Socci ND, Onuchic JN, Wolynes PG: **Diffusive dynamics of the reaction coordinate for protein folding funnels.** *J Chem Phys* 1996, **104**:58–60.
•
The kinetic analysis of folding for 27-mers on the basis of F(Q) plots, which assume that Q is a reaction coordinate and the maximum of F(Q) is a TS, is presented.

94. Ferrenberg AM, Swendsen RH: **Optimized Monte Carlo data analysis.** *Phys Rev Lett* 1989, **63**:1195–1197.

95. Boczko E, Brooks C: **First-principle calculation of the folding free energy of a three-helix bundle protein.** *Science* 1995, **269**:393–396.
••
An interesting analysis of folding in the model intermediate between lattice models and full-atom ones. The ingenious approach made it possible to evaluate free-energy profile as a function of the volume of the globule.

96. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG: **Funnels, pathways, and the energy landscape of protein folding: a synthesis.** *Proteins* 1995, **21**:167–195.

97. Gutin AM, Abkevich VI, Shakhnovich EI: **Chain length scaling of protein folding time.** *Phys Rev Lett* 1996, **77**:5433–5436.
•
It is shown that folding rates for model protein sequences follow power-law dependence on length at conditions of their respective fast folding.

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

99. Viguera AR, Blanco F, Serrano L: **The order of secondary structure elements does not determine the structure of a protein but does affect its folding kinetics.** *J Mol Biol* 1995, **247**:670–681.

100. Baldwin R: **The nature of protein folding pathways: the classical versus the new view.** *J Biomol NMR* 1995, **5**:103–109.
•
A thought-provoking review of experimental data comparing the 'classical' and the 'new' view of protein folding.

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

102. Pande V, Grosberg AYu, Tanaka T: **How accurate must potentials be for successful modeling of protein folding?** *J Chem Phys* 1995, **103**:1–10.
•
A degree of uncertainty in folding parameters that still allows the folding of well-designed sequences is estimated analytically.