A surprising simplicity to protein folding

David Baker

Department of Biochemistry, University of Washington, J567 Health Sciences Building, Box 357350, Seattle, Washington 98195, USA

The polypeptide chains that make up proteins have thousands of atoms and hence millions of possible inter-atomic interactions. It might be supposed that the resulting complexity would make prediction of protein structure and protein-folding mechanisms nearly impossible. But the fundamental physics underlying folding may be much simpler than this complexity would lead us to expect: folding rates and mechanisms appear to be largely determined by the topology of the native (folded) state, and new methods have shown great promise in predicting protein-folding mechanisms and the three-dimensional structures of proteins.

Proteins are linear chains of amino acids that adopt unique three-dimensional structures ('native states') which allow them to carry out intricate biological functions. All of the information needed to specify a protein's three-dimensional structure is contained within its amino-acid sequence. Given suitable conditions, most small proteins will spontaneously fold to their native states[1].

The protein-folding problem can be stated quite simply: how do amino-acid sequences specify proteins' three-dimensional structures? The problem has considerable intrinsic scientific interest: the spontaneous self-assembly of protein molecules with huge numbers of degrees of freedom into a unique three-dimensional structure that carries out a biological function is perhaps the simplest case of biological self-organization. The problem also has great practical importance in this era of genomic sequencing: interpretation of the vast amount of DNA sequence information generated by large-scale sequencing projects will require determination of the structures and functions of the encoded proteins, and an accurate method for protein structure prediction could clearly be vital in this process.

Since Anfinsen's original demonstration of spontaneous protein refolding, experimental studies have provided much information on the folding of natural proteins[2–4]. Complementary analytical and computational studies of simple models of folding have provided valuable and general insights into the folding of polymers and the properties of folding free-energy landscapes[5–7]. These studies of idealized representations of proteins have inspired new models, some described here, which attempt to predict the results of experimental measurements on real proteins.

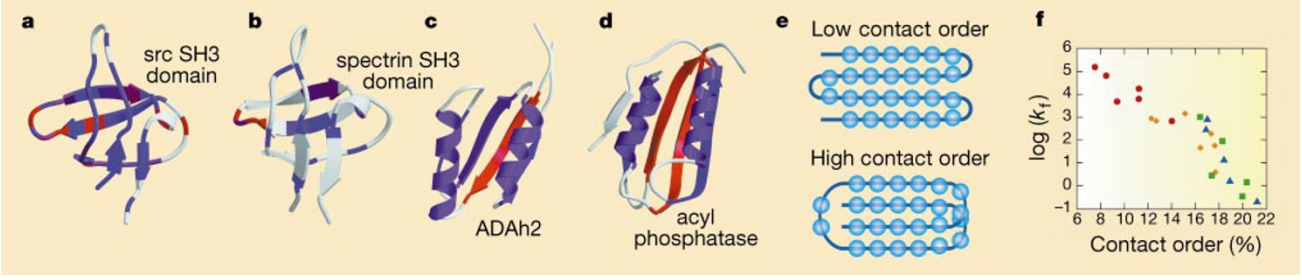
Because the number of conformations accessible to a polypeptide chain grows exponentially with chain length, the logical starting point for the development of models attempting to describe the folding of real protein is experimental data on very small proteins (fewer than 100 residues). Fortunately, there has been an explosion of information about the folding of such small proteins over the last ten years[3]. For most of these proteins, partially ordered non-native conformations are not typically observed in experiments, and the folding reactions can usually be well modelled as a two-state transition between a disordered denatured state and the ordered native state. In contrast, the folding kinetics of larger proteins may in some cases be dominated by escape from low-free-energy non-native conformations. The folding of larger proteins is also often facilitated by 'molecular chaperones'[8] which prevent improper protein aggregation.

To pass between the unfolded and native low-free-energy states, the protein must pass through a higher-free-energy transition state. In the unfolded state the protein can take up any one of many conformations, whereas in the native state it has only one or a few distinct conformations. The degree of heterogeneity of conformations in the transition state has thus been the subject of much discussion[9–11]. For example, one of the main differences between the

Box 1

Dependence of folding mechanisms on topology

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



'old' and 'new' views of protein folding is that the 'new' view allows for a much more heterogeneous transition state—really a transition state ensemble—than the 'old' view, which concentrated on a single, well defined folding 'pathway'.

The primary measurements that can be made experimentally of the highly cooperative folding reactions of small proteins are: the folding rate; the distribution of structures in the transition state ensemble, inferred from the effects of mutations on the folding rate (Box 1); and the structure of the native state. Here I focus on recent progress in predicting these three features.

Topology determines folding mechanisms

Are simple models likely to be able to account for the overall features of the folding process, given the many possible inter-atomic interactions in even a small protein? Recent data indicate that the fundamental physics underlying the folding process may be simpler than was previously thought.

The complexity of protein structure emerges from the details of how individual atoms in both a protein's peptide backbone and its amino-acid residues interact. However, the general path that the polymer chain takes through space—its topology—can be very similar between proteins. Three independent lines of investigation indicate that protein-folding rates and mechanisms are largely determined by a protein's topology rather than its inter-atomic interactions[12].

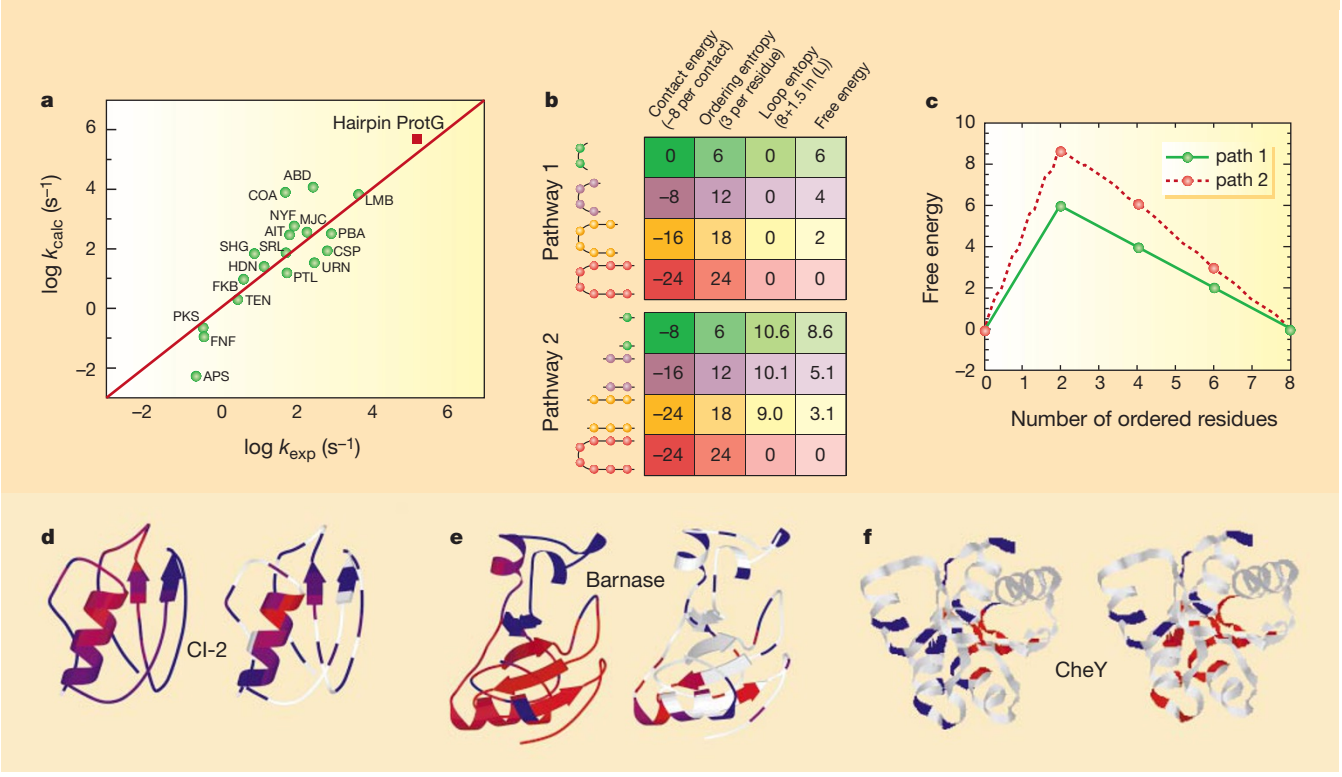
First, large changes in amino-acid sequence, either experimental[13,14] or evolutionary[15], that do not alter the overall topology of a protein usually have less than tenfold effect on the rate of protein folding[15]. This suggests evolution has not optimized protein sequences for rapid folding, an encouraging result for simple model development.

Second, using the consequences of mutations on folding kinetics to probe the transition states of proteins with similar structures but very different sequences has shown that the structures of these transition states are relatively insensitive to large-scale changes in sequence[16–18]. For example, in Box 1 there are two examples of pairs of structurally related proteins with little or no sequence similarity that have very similar folding transition-state ensembles.

Box 2

Prediction of protein-folding mechanisms

Munoz and Eaton[24] computed folding rates by solving the diffusion equation of motion on the one-dimensional free-energy profiles that result from projection of the full free-energy landscape onto a reaction coordinate corresponding to the number of ordered residues. **a** shows the accuracy of their prediction by plotting computed folding rates (k_{calc}) against experimentally measured rates (k_{exp}). To predict folding transition state structure, the lowest free energy paths to the native state can be identified. For example, a β-hairpin (**b**) has two possible paths to the native state, beginning at the hairpin (pathway 1) or at the free ends (pathway 2; ordered residues only are indicated; L is loop length). The Table gives the contributions to the free energy of each configuration (total free energy is the sum of the first three columns). Plotting the free energy as a function of the number of ordered residues (**c**) shows that the transition state for both pathways consists of configurations with two of the residues ordered. Calculations on real proteins (**d**–**f**) have considered all possible paths: the folding rate and transition state structure are determined from the lowest free-energy paths. Galzitskaya and Finkelstein[25] and Alm and Baker[26] predicted the folding transition state structure of CheY (**f**), and CI-2 (**d**) and barnase (**e**), respectively. They identified the transition-state ensemble by searching for the highest free-energy configurations on the lowest free-energy paths between unfolded and folded states. The effects of mutations on the folding rate were predicted on the basis of the contribution of the interactions removed by the mutations to the free energy of the transition state ensemble, or by directly determining the change in folding rate. The predicted effects of mutations on the folding rates are shown on the native structure (left); the measured effects, on the right (the colour scheme is as in Box 1; grey, regions not probed by mutations; experimental results for CI-2 and barnase, ref. 4; CheY, ref. 38).



Third, the folding rates of small proteins correlate with a property of the native state topology: the average sequence separation between residues that make contacts in the three-dimensional structure (the 'contact order'; Box 1). Proteins with a large fraction of their contacts between residues close in sequence ('low' contact order) tend to fold faster than proteins with more non-local contacts ('high' contact order)[12,19]. This correlation holds over a million-fold range of folding rates, and is remarkable given the large differences in the sequences and structures of the proteins compared. Simple geometrical considerations appear to explain much of the difference in the folding rates of different proteins.

The important role of native-state topology can be understood by considering the relatively large entropic cost of forming non-local interactions early in folding. The formation of contacts between residues that are distant along the sequence is entropically costly, because it greatly restricts the number of conformations available to the intervening segment. Thus, interactions between residues close together in sequence are less disfavoured early in folding than interactions between widely separated residues. So, for a given topology, local interactions are more likely to form early in folding than non-local interactions. Likewise, simple topologies with mostly local interactions are more rapidly formed than those with many non-local interactions. More generally, the amount of configurational entropy lost before substantial numbers of favourable native interactions can be made depends on the topology of the native state. The importance of topology has also been noted in studies of computational models of folding[20–23].

As proteins' sequences determine their three-dimensional structures, both protein stability and protein-folding mechanisms are ultimately determined by the amino-acid sequence. But whereas stability is sensitive to the details of the inter-atomic interactions (removal of several buried carbon atoms can completely destabilize a protein), folding mechanisms appear to depend more on the low-resolution geometrical properties of the native state.

Predicting folding mechanism from topology

The results described above indicate that simple models based on the structure of the native state should be able to predict the coarse-grained features of protein-folding reactions. Several such models have recently been developed, and show considerable promise for predicting folding rates and folding transition-state structures. Three approaches[24–26] have attempted to model the trade-off between the formation of attractive native interactions and the loss of configurational entropy during folding. Each assumes that the only favourable interactions possible are those formed in the native state. This neglect of non-native interactions is consistent with the observed importance of native-state topology in folding, and dates back to the work of Go on simple lattice models[27].

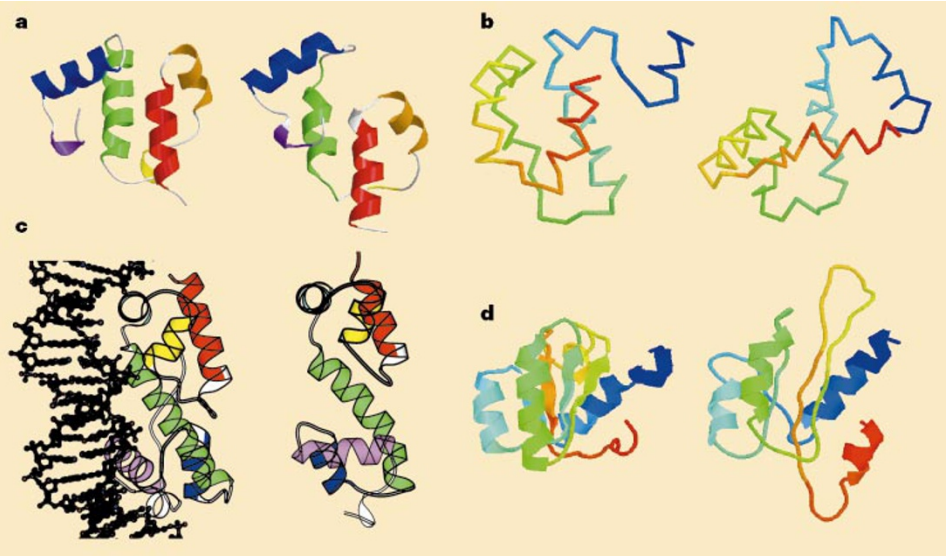
Although the approaches differ in detail, the fundamental ideas are similar. All use a binary representation of the polypeptide chain in which each residue is either fully ordered, as in the native state, or completely disordered. To limit the number of possible configurations, all ordered residues are required to form a small number of segments, continuous in sequence. Attractive interactions are taken to be proportional to the number of contacts, or the amount of buried surface area, between the ordered residues in the native structure, and non-native interactions are completely ignored. The entropic cost of ordering is a function of the number of residues ordered and the length of the loops between the ordered segments. Folding kinetics are modelled by allowing only one residue to become ordered (or disordered) at a time. As the number of ordered residues increases, the free energy first increases, owing to the entropic cost of chain ordering, and then decreases, as large numbers of attractive native interactions are formed.

Such simple models can potentially be used to predict experimentally measurable quantities such as the folding rate, which depends on the height of the free-energy barrier, and the effects of mutations on the folding rate, which depend on the region(s) of the protein ordered near the top of the barrier. Predictions of both

Box 3

Ab initio structure predictions

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



folding rates and folding transition-state structures using these simple models are quite encouraging (Box 2; other recent models have also yielded good results[28–33]).

The success of these models in reproducing features of real folding reactions again supports the idea that the topology of the native state largely determines the overall features of protein-folding reactions and that non-native interactions have a relatively minor role. Incorporation of sequence-specific information into these models, either in the inter-residue interactions or in the free-energy costs of ordering different segments of the chain, should improve their accuracy to the point where they may be able to account for much of the experimental data on the folding of small proteins.

Ab initio structure prediction

Predicting three-dimensional protein structures from amino-acid sequences alone is a long-standing challenge in computational molecular biology. Although the preceding sections suggest that the only significant basin of attraction on the folding landscapes of small proteins is the native state, the potentials used in *ab initio* structure-prediction efforts have not had this property, and until recently such efforts met with little success. The results of an international blind test of structure prediction methods (CASP3; ref. 34) indicate, however, that significant progress has been made[35,36].

As with the models for protein-folding mechanisms, most of the successful methods attempt to ignore the complex details of the inter-atomic interactions—the amino-acid side chains are usually not explicitly represented—and instead focus on the coarse-grained features of sequence–structure relationships. Problems in which the full atomic detail of interactions in the native state is important—such as the design of novel stable proteins, and the prediction of stability and high resolution structure—will almost certainly require considerably more detailed models.

Some of the most successful blind *ab initio* structure predictions made in CASP3 are shown in Box 3. In several of these predictions the root-mean-square deviation between backbone carbon atoms in the predicted and experimental structures is below 4.0 Å over segments of up to 70 residues. Several of these models can compete with more traditional fold-recognition methods. At least one case (Mar A) gave a model capable of providing clues about protein function[39].

The predictions are an encouraging improvement over those achieved in the previous structure-prediction experiment (CASP2), but improvements are still needed to the accuracy and reliability of the models. Improvements in *ab initio* structure prediction may allow these methods to generate reliable low-resolution models of all the small globular proteins in an organism's genome.

Emerging simplicity

The experimental results and predictions discussed here indicate that the fundamental physics underlying folding may be simpler than previously thought and that the folding process is surprisingly robust. The topology of a protein's native state appears to determine the major features of its folding free-energy landscape. Both protein structures and protein-folding mechanisms can be predicted, to some extent, using models based on simplified representations of the polypeptide chain. The challenge ahead is to improve these models to the point where they can contribute to the interpretation of genome sequence information. □

1. Anfinson, C. Principles that govern the folding of protein chains. *Science* **181,** 223–227 (1973).
2. Baldwin, R. L. & Rose, G. D. Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem. Sci.* **24,** 26–33 (1999).
3. Jackson, S. E. How do small single-domain proteins fold? *Fold. Des.* **3,** R81–91 (1998).
4. Fersht, A. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (Freeman, New York, 1999).
5. Chan, H. S. & Dill, K. A. Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics. *Proteins* **30,** 2–33 (1998).
6. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21,** 167–195 (1995).
7. Dobson, C. M. & Karplus, M. The fundamentals of protein folding: bringing together theory and experiment. *Curr. Opin. Struct. Biol.* **9,** 92–101 (1999).
8. Horwich, A. L. Chaperone rings in protein folding and degradation. *Proc. Natl Acad. Sci. USA* **96,** 11033–11040 (1999).
9. Shakhnovich, E. I. Folding nucleus: specific or multiple? Insights from lattice models and experiments. *Fold. Des.* **3,** R108–111 (1998).
10. Pande, V. S., Grosberg, A. Y., Tanaka, T. & Rokhsar, D. Pathways for protein folding: is a new view needed? *Curr. Opin. Struct. Biol.* **8,** 68–79 (1998).
11. Thirumalai, D. & Klimov, D. K. Fishing for folding nuclei in lattice models and proteins. *Fold. Des.* **3,** R112–118 (1998).
12. Alm, E. & Baker, D. Matching theory and experiment in protein folding. *Curr. Opin. Struct. Biol.* **9,** 189–196 (1999).
13. Riddle, D. S. *et al.* Functional rapidly folding proteins from simplified amino acid sequences. *Nature Struct. Biol.* **4,** 805–809 (1997).
14. Kim, D. E., Gu, H. & Baker, D. The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl Acad. Sci. USA* **95,** 4982–4986 (1998).
15. Perl, D. *et al.* Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nature Struct. Biol.* **5,** 229–235 (1998).
16. Chiti, F. *et al.* Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.* **6,** 1005–1009 (1999).
17. Martinez, J. C. & Serrano, L. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Struct. Biol.* **6,** 1010–1016 (1999).
18. Riddle, D. S. *et al.* Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.* **6,** 1016–1024 (1999).
19. Plaxco, K. W., Simons, K. T. & Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277,** 985–994 (1998).
20. Shea J. E., Onuchic J. N. & Brooks, C. L. III. Exploring the origins of topological frustration: design of a minimally frustrated model of fragment B of protein A. *Proc. Natl Acad. Sci. USA* **96,** 12512–12517 (1999).
21. Onuch, J. N., Nymeyer, H., Garcia, A. E., Chaine, J. & Socci, N. D. The energy landscape theory of protein folding: insights into folding mechanism and scenarios. *Adv. Protein Chem.* **53,** 87–152 (2000).
22. Micheletti, C., Banavar, J. R., Maritan, A. & Seno, F. Protein structures and optimal folding from a geometrical variational principle. *Phys. Rev. Lett.* **82,** 3372–3375 (1999).
23. Abkevich, V., Gutin, A. & Shakhnovich, E. Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33,** 10026–10036 (1994).
24. Munoz, V. & Eaton, W. A. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA* **96,** 11311–11316 (1999).
25. Galzitskaya, O. V. & Finkelstein, A. V. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl Acad. Sci. USA* **96,** 11299–11304 (1999).
26. Alm, E. & Baker, D. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl Acad. Sci. USA* **96,** 11305–11310 (1999).
27. Go, N. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12,** 183–210 (1983).
28. Portman J. J., Takada, S. & Wolynes, P. G. Variational theory for site resolved protein folding free energy surfaces. *Phys. Rev. Lett.* **81,** 5237–5240 (1998).
29. Debye, D & Goddard, W. A. First principles prediction of protein-folding rates. *J. Mol. Biol.* **294,** 619–625 (1999).
30. Li, A. J. & Daggett, V. Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 by molecular dynamics simulations. *J. Mol. Biol.* **257,** 412–429 (1996).
31. Lazaridis, T. & Karplus, M. 'New view' of protein folding reconciled with the old through multiple unfolding simulations. *Science* **278,** 1928–1931 (1997).
32. Sheinerman, F. B. & Brooks, C. L. III A molecular dynamics simulation study of segment B1 of protein G. *Proteins* **29,** 193–202 (1997).
33. Burton, R. E., Myers, J. K. & Oas, T. G. Protein folding dynamics: quantitative comparison between theory and experiment. *Biochemistry* **37,** 5337–5343 (1998).
34. Moult, J., Hubbard, T., Fidelis, K. & Pederson, J. T. Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins* (Suppl.) **3,** 2–6 (1999).
35. Orengo, C. A., Bray, J. E., Hubbard, T. LoConte, L. & Sillitoe, I. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins* (Suppl.) **3,** 149–170 (1999).
36. Murzin, A. G. Structure classification-based assessment of CASP3 predictions for the fold recognition targets. *Proteins* (Suppl.) **3,** 88–103 (1999).
37. Villegas, V., Martinez, J. C., Aviles, F. X. & Serrano, L. Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J. Mol. Biol.* **283,** 1027–1036 (1998).
38. Lopez-Hernandez, E. & Serrano, L. Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI-2. *Fold. Des.* **1,** 43–55 (1996).
39. Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* (Suppl.) **3,** 171–176 (1999).
40. Samudrala, R. Xia, Y. Huang, E. & Levitt, M. Ab initio protein structure prediction using a combined hierarchical approach. *Proteins* (Suppl.) **3,** 194–198 (1999).
41. Ortiz, A. R., Kolinski, A., Rotkiewicz, P., Ilkowski, B. & Skolnick, J. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins* (Suppl.) **3,** 177–185 (1999).
42. Weigelt, J., Brown, S. E., Miles, C. S. & Dixon, N. E. NMR structure of the N-terminal domain of *E. coli* DnaB helicase: implications for structure rearrangements in the helicase hexamer. *Structure* **7,** 681–690 (1999).
43. Slupsky, C. M. *et al.* Structure of the Ets-1 pointed domain and mitogen-activated protein kinase phosphorylation site. *Proc. Natl Acad. Sci. USA* **95,** 12129–12134 (1998).
44. Rhee, S., Martin, R. G., Rosner, J. L. & Davies, D. R. A novel DNA-binding motif in MarA: the first structure for an AraC family transcriptional activator. *Proc. Natl Acad. Sci. USA* **95,** 10413–10418 (1998).