126. Shalloway, D., In Floudas, C.A. and Pardalos, P.M. (Eds.) Recent Advances in Global Optimization, Vol. 1, Princeton University Press, Princeton, NJ, 1991, pp. 433–648.
127. Clearwater, S.H., Huberman, B.A. and Hogg, T., Science, 254(1991)1181.
128. Kaesar, C. and Elber, R., J. Phys. Chem., 99(1995)11550.
129. Pedersen, J. and Moult, J., Curr. Opin. Struct. Biol., 6(1996)227.
130. McCammon, J.A., Gelin, B.R. and Karplus, M., Nature, 267(1977)585.
131. Powell, M.J.D., Math. Programming, 12(1977)241.
132. Noguti, T. and Go, N., Biopolymers, 24(1985)527.
133. Ripoll, D.R. and Scheraga, H.A., J. Protein Chem., 8(1989)263.
134. Skolnick, J. and Kolinski, A., Science, 250(1990)1121.
135. Go, N. and Scheraga, H.A., Macromolecules, 3(1970)178.
136. Ring, C.S. and Cohen, F.E., Isr. J. Chem., 34(1994)245.
137. Elofsson, A., Le Grand, S.M. and Eisenberg, D., Proteins, 23(1995)73.
138. Kang, H.S., Kurochkina, N.A. and Lee, B., J. Mol. Biol., 229(1993)448.
139. Dunbrack, R.L. and Karplus, M., J. Mol. Biol., 230(1993)543.
140. Borchert, T.V., Abagyan, R.A., Kishan, K.V.R., Zeelen, J.Ph. and Wierenga, R.K., Structure, 1(1993)205.
141. Metropolis, N.A., Rosenbluth, A.W., Rosenbluth, N.M., Teller, A.H. and Teller, E., J. Chem. Phys., 21(1953)1087.
142. Blanco, F.J., Jimenez, M.A., Herranz, J., Rico, M., Santoro, J. and Nieto, J., J. Am. Chem. Soc., 115(1993)5887.
143. Struthers, M.D., Cheng, R.P. and Imperiali, B., Science, 271(1996)342.
144. Abagyan, R.A. and Argos, P., J. Mol. Biol., 225(1992)519.
145. Wodak, S.J. and Janin, J., J. Mol. Biol., 124(1978)323.
146. Kuntz, I.D. and Blaney, J.M., Oatley, S.J., Langridge, R., and Ferrin, T.E., J. Mol. Biol., 161(1982)269.
147. Connolly, M.L., Biopolymers, 25(1986)1229.
148. Warwicker, J., J. Mol. Biol., 206(1989)381.
149. Goodsell, A.S., and Olson, A.J., Proteins Struct. Funct. Genet., 8(1990)195.
150. Cherfils, J., Duquerroy, S. and Janin, J., Proteins, 11(1991)271.
151. Jiang, F. and Kim, S.-H., J. Mol. Biol., 219(1991)79.
152. Chou, K.-C. and Carlacci, L., Protein Eng., 4(1991)661.
153. Bacon, D.J. and Moult, J., J. Mol. Biol., 225(1992)849.
154. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C. and Vakser, I.A., Proc. Natl. Acad. Sci. USA, 89(1992)2195.
155. Walls, P.H. and Sternberg, M.J.E., J. Mol. Biol., 228(1992)277.
156. Pellegrini, M. and Doniah, S., Proteins, 15(1993)436.
157. Vakser, I.A., Protein Eng., 8(1995)371.
158. Fischer, D., Lin, S.L., Wolfson, H.L. and Nussinov, R., J. Mol. Biol., 248(1995)459.
159. Goodsell, D.S., Morris, G.M. and Olson, A.J., J. Mol. Recog., 9(1996)1.
160. Janin, J., Prog. Biophys. Mol. Biol., 64(1995)145.
161. Diamond, R., J. Mol. Biol., 82(1974)371.
162. Sheriff, S., Silverton, E.W., Padlan, E.A., Cohen, G.H., Smith-Gill, S.J., Finzel, B.C. and Davies, D.R., Proc. Natl. Acad. Sci. USA, 84(1987)8075.
163. Borchert, T.V., Kishan, K.V.R., Zeelen, J.Ph., Schliebs, W., Thanki, N., Abagyan, R.A., Jaenicke, R. and Wierenga, R.K., Structure, 3(1995)669.
164. Thanki, N., Zeelen, J.Ph., Mathieu, M., Jaenicke, R., Abagyan, R.A., Wierenga, R.K. and Schliebs, W., Protein Eng., 10(1997)159.

Monte Carlo lattice dynamics and the prediction of protein folds

Jeffrey Skolnick[a] and Andrzej Kolinski[a,b]

[a]*Department of Molecular Biology, The Scripps Research Institute, 10666 North Torrey Pines Road, La Jolla, CA 92037, U.S.A.*
[b]*Department of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland*

Introduction

The general solution to the protein folding problem demands that two very difficult problems be concomitantly solved [1]. An energy function whose global minimum is in the native conformation of the protein must be developed [2]. Simultaneously, an efficient strategy to search through the myriad of local energy minima for the desired global minimum must be formulated [3]. One way to attack both problems is to reduce the complexity of the model being considered [4]. Rather than treat the model at the level of atomic detail, the representation of the protein can be simplified. Various extents of this simplification have been explored. They range from highly simplified models that treat the native conformation of proteins as points on a small cube to high coordination lattice models that describe the native conformation of proteins with high geometric fidelity [5–14]. While highly idealized models have been useful in providing a number of qualitative insights into some general features of protein folding [10,15,16], they cannot be used to fold a real protein. This chapter focuses on results from high coordination lattice models of proteins that have been developed over the past several years and which are complementary to the simplified model studies [17–31]. These high coordination models not only provide insights into the thermodynamics of the protein folding process [25], but in a number of cases can predict the native conformation of a number of proteins at the level of 2–4 Å root-mean-square deviation (rms) from native [20,21].

The outline of this chapter is as follows. We begin with a discussion of the geometric model of a protein and the interaction scheme. In particular, we focus on the reasons why the various contributions to the potential are included and describe what happens if the individual terms are considered in isolation. We then describe the two types of Monte Carlo sampling schemes that have been employed, namely classical Metropolis Monte Carlo (MMC) [32] and the novel entropy sampling Monte Carlo technique (ESMC) due to Hao and Scheraga [33–35]. Next, results from the folding of some idealized protein sequences are presented [18,23,25]. These studies enable an exploration of the possible origins of the cooperativity of the protein folding process. We then summarize results on the folding of a number of small globular proteins [20,36], as well as some predictions for protein redesign [29,30]. Subsequently, a novel algorithm for the prediction of the locations where the protein chain reverses

global direction, i.e. 'U'-turns and the dominant secondary structure found in the regions between the U-turns, is described [27]. This is followed by a review of folding results when a relatively small number of tertiary constraints (which may come from NMR experiments) are provided to the model [26]. Then, results from the *de novo* folding of the GCN4 leucine zipper, which adopts a dimeric coiled coil in solution, are summarized [21]. Next, an overview of the general formalism designed to predict the state of association of coiled coils [22,31], and comparison with experimental data on a variety of sequences, are presented [37,38]. We conclude with a discussion of the weaknesses of the present generation of lattice models and a perspective on the outlook for future progress.

Lattice models of proteins

As indicated in Fig. 1, the C^{α} coordinates of the protein backbone are confined to a set of lattice points which reside on an underlying cubic lattice, whose lattice spacing $a = 1.22$ Å [19]. Successive C^{α} atoms are connected by virtual bond vectors $a \cdot \mathbf{v}$, with $\{\mathbf{v}\} = \{(\pm 3, \pm 1, \pm 1), \ldots, (\pm 3, \pm 1, 0), \ldots, (\pm 3, 0, 0), \ldots, (\pm 2, \pm 2, \pm 1), \ldots, (\pm 2, \pm 2, 0), \ldots\}$. On considering all possible permutations of the coordinates, $\{\mathbf{v}\}$

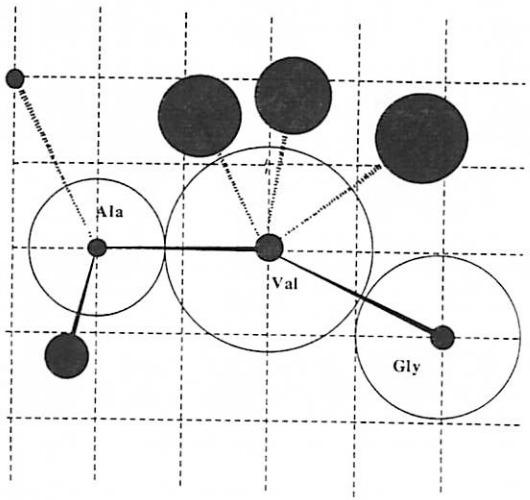


Fig. 1. Schematic representation of the geometry of the protein model. The C^{α} vertices are confined to high coordination lattice points. The side-chain centers of mass are located off-lattice. Ala, Pro, and Gly have a single rotamer for a given backbone virtual bond angle. All the other residues have multiple rotamers.

contains 90 basis vectors; thus, we refer to it as the 90-neighbor lattice. However, when the virtual bond angles are restricted to realistic values, the number of possible continuations of the C^{α} trace, given a pair of preceding C^{α}'s, is about 30. Thus, the intrinsic conformational entropy of the backbone is comparable to real proteins. The geometric accuracy of the C^{α} representation is in the range of 0.6–0.7 Å rms with respect to high-resolution PDB structures [39]. This is true regardless of protein size and orientation of the protein on the lattice. The fact that space is essentially isotropic and that all structures can be represented at comparable geometric resolution is the reason why this high coordination lattice is used.

Side chains are represented as a set of pseudoatoms located at the side-chain center of mass. For all amino acids except Gly, Pro and Ala, there are multiple rotamers. These rotamers are chosen so that the center of mass of a side chain in real proteins will be no farther than 1 Å from another member of the rotamer library. The side-chain rotamers are not confined to lattice points; however, the C^{α} backbone defines the reference frame for the rotamer coordinates. In a similar fashion, it is possible to rapidly and quite accurately reconstruct the peptide backbone and C^{β} atoms given a set of three virtual backbone bond vectors; the former can be used in an explicit atom hydrogen bond scheme [28,40]. More generally, many geometric and energetic quantities can be rapidly accessed from the set of virtual backbone vectors that define the instantaneous conformation of the chain. Thus, many quantities can be precalculated in advance. This allows for a two-order-of-magnitude speed-up over the corresponding model described in a continuous space representation [41]. The possibility of such a speed-up is absolutely essential to be able to adequately explore conformational space and is the principal reason why lattice models are used.

Interaction scheme

The key aspect of any successful model for protein folding is the nature of the terms that define the potential. Recently, it has once again become very popular to consider a very simple interaction set [42,43]; for example, all hydrophobic residues are treated as having the same interactions. While this simplicity is appealing, it belies the fact that such an approach generates many essentially isoenergetic chain conformations, many of which are geometrically unlike folded proteins. Typically, such an approach results in native-like states being in the best several hundred structures as ranked according to their energy [43]. While this selection may be somewhat better than random, in practice, there are far too many conformations to be of practical use. If, for example, one could predict a handful of different topologies, then such topologies might be differentiated experimentally. However, when there are several hundred possible answers, it is very unlikely that the correct fold can be fished out from the myriad of possibilities.

There is another reason why such a simplistic approach will not work. Consider the recent studies of Harbury et al. [37] on GCN4 leucine zippers and a number of mutants. They mutated the residues in the core to various combinations of Val, Leu,

and Ile. Depending on the identity and location in the sequence of the hydrophobic residues, the equilibrium shifted from dimers to trimers and then to tetramers. Since for these sequences the residues in the core are always hydrophobic, an interaction scheme based on just two types of residues, hydrophobic and hydrophilic, could not possibly predict the state of association. More generally, it is possible to build structures of different topologies that have the same pair interaction as assessed by the number of interacting hydrophils and hydrophobes. While complexity for complexity's sake is to be avoided, it is precisely to reduce the number of possible low-energy topologies that more complicated interaction schemes have been developed.

Contributions to the potential

In what follows, we describe the qualitative features of the interaction scheme which we have developed. The origin of these models goes back to very simple HP-type models where there are only two kinds of residues, polar and nonpolar [4,6–8,14,44–50], but additional complexity has been added to reproduce essential features of the physics which would be absent if such terms were excluded [18,19,21,23,36]. We wish however to emphasize that the force field is constantly being improved and modified in order to enable us to fold a broader class of proteins; it reflects the ongoing process of our increased understanding of interactions in proteins. Each time the potential changes, we go back and repeat the folding simulations on those proteins already folded so as to ensure that the 'improvements' permit an ever-increasing set of proteins to be folded.

The potential must be designed so as to capture both generic and sequence-specific features of proteins. The nature of the individual contributions is listed below.

Hydrogen bonds

The most important generic term involves hydrogen bonds. Whether the relative intraprotein hydrogen bond free energy is less favorable or more favorable than that of hydrogen bonds to water is not the crux of the effect. What is most salient is that the presence of unsatisfied hydrogen bonds within a protein is energetically very unfavorable. Since hydrogen bonds are both distance- and orientation-dependent, they are an extremely important structural regularizing term. They serve to greatly restrict the manifold of accessible compact conformations.

Two versions of hydrogen bonds have been implemented. One is C^{α}-based [19] and the other uses an explicit backbone amide hydrogen and carbonyl oxygen representation [28]. Both reproduce about 90% of the hydrogen bonds as assigned by Kabsch and Sander [51]. The former is very much in the spirit of Levitt and Greer [52], whereas the latter was introduced to improve the hydrogen bonding in β-structures. In the absence of other contributions to the potential, at low temperature, they tend to generate helices punctuated by breaks where the prolines are located [53]. The choice of helices over β-states is due to entropic reasons.

Intrinsic secondary preferences

Next, there are amino acid pair-specific contributions that reflect the statistical preference of individual amino acids to adopt a given type of secondary structure. Both cooperative and noncooperative versions of this potential have been used. Basically, similar behavior is observed, but the former version yields a better defined interface between secondary structural elements [19,21,23,24,29,30,36]. This contribution to the energy of the folded state is typically about 20–25% of the total. When used alone, this term produces fragments of secondary structure, with a very diffuse and continuous conformational transition. When combined with terms that account for the generic stiffness of polypeptide backbones, the accuracy of secondary structure prediction is comparable to more standard methods, i.e., depending on the sequence, between 50 and 70% of the residues are correctly assigned [24]. Finally, by providing for a small, but nonnegligible, amount of secondary structure in the denatured state, these terms assist in the early states of folding and also act to reduce the configurational entropy of compact states. On average, they determine which type of secondary structure a protein adopts, but they can be overridden by tertiary interactions [20].

Burial as a one-body term

The next class of terms reflects the individual preference of a given residue to be buried or exposed. One-body burial terms serve to generate compact structures where on average the hydrophobic residues are in the interior and the hydrophilic residues are exposed. But, they generate nonspecific side-chain packing arrangements, and multiple topologies can have an essentially identical burial energy. In the absence of other terms in the interaction scheme, when reasonable coordination number lattices are used and compact structures at protein-like densities are generated, contrary to the hypothesis of Dill and co-workers [20], secondary structure is not enhanced on compaction [17,53–56]. Many of these random conformations would in reality have very high energies because these conformations would not be hydrogen bonded.

For single-domain globular proteins, a centrosymmetric potential that describes the tendency for an amino acid side-chain side to be located at a given relative distance from the center of mass of the protein has been used [19,57]. This formulation offers the advantage that it can accommodate the fact that residues such as tyrosine prefer to be located near the protein surface. However, it suffers from the disadvantages that there may be problems if the protein is very asymmetric, it has trouble differentiating edge from interior strands in β-structures, and it cannot be applied to multimeric or multidomain proteins. To address these concerns, a potential based on the number of side-chain contacts has also been introduced [21]. Whenever the environment of a residue exceeds the contact threshold, it is counted as buried. A possible problem with this approach is that the situation can arise where a substantial amount of the surface is actually exposed, but where the contacts are clustered

over a relatively small portion of the surface. Improvement in the formulation of the burial potential is clearly necessary.

Pair potentials

These potentials of mean force help to select out the preferred topology and are operative in the molten globule and the native state. However, they do not provide a sufficient energetic separation between the native conformation and alternative higher energy structures, many of which have the native fold but different side-chain packing arrangements. Thus, they do not yield a unique native state with long-lived side-chain contacts. The best pair potentials of mean force have attractive or neutral interactions between hydrophils and attractive interactions between hydrophobes [23]; obviously, it is essential that hydrophilic and hydrophobic residues experience net repulsive interactions. When a pair interaction scale in which hydrophilic residues are repulsive is applied to a β-protein, then highly curved β-sheets are generated so as to minimize the number of hydrophilic–hydrophilic contacts. At a bare minimum, in order for twisted, quasiplanar β-sheets to be stable, the hydrophilic pair interaction should be no worse than neutral. However, in those scales where pairs of hydrophilic residues are attractive as are pairs of hydrophobic residues, then this contribution by itself cannot create the phase segregation where hydrophobic residues are on average found in the protein interior.

Many investigators, ourselves included, have developed statistical potentials of mean force between pairs of residues i and j obtained from expressions of the type [19,23,58,59]:

$$\varepsilon_{i,j} = -kT \ln(n_{obs}(i,j)/n_{exp}(i,j)) \quad (1)$$

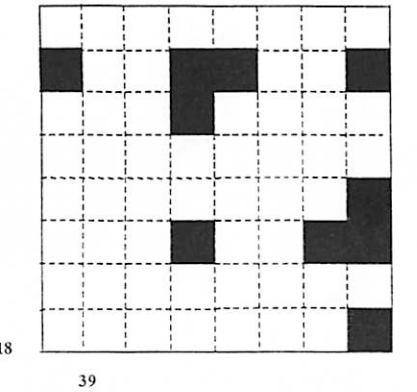
with $n_{obs}(i,j)$ the observed number of contacts between pairs of amino acids i and j. k is Boltzmann's constant and T is the absolute temperature. This quantity is directly obtained from a set of Protein Data Bank protein structures [60]. Here, $n_{exp}(i,j)$ is the expected number of contacts if interactions between i and j are random. It is in this term that the difference between all statistical contact potentials resides [59].

To date, the most sensitive residue-based pair potentials have been derived assuming that the quasichemical approximation holds for groups of heavy atoms; then, the average residue interaction based on the interaction between such groups is calculated [23]. The problem with the quasichemical approximation is that it ignores chain connectivity and the presence of regular secondary structure. Recently, a more general approach which includes these effects has been developed. Even at the level of interacting residues, it is the best inverse folding pair potential derived to date by our group [61].

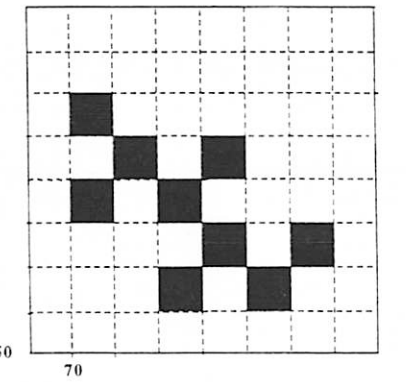
Effective multibody interactions

If one defines a contact as occurring when any pair of heavy atoms is less than 4.2 Å apart, then interacting supersecondary structural elements in globular proteins exhibit

well-defined side-chain contact patterns [18,19,62]. Typical helix-to-helix and beta-to-beta packing patterns are shown in Figs. 2A and B, respectively. Furthermore, it is possible to define a set of side-chain center of mass contact distance thresholds so that 82% of the heavy-atom contacts are recovered with a Matthews coefficient of



A



B

Fig. 2. Representative side-chain packing contact maps for an interacting pair of (A) antiparallel helices and (B) β-strands.

0.85 [61]. Thus, contact maps in the single ball side-chain description can essentially recover the heavy-atom contact description. In the absence of higher order multibody terms beyond pair contributions, we find that the predicted packing patterns of the resulting ensemble of structures exhibit essentially random overlap with the native state. However, the overlap with native contacts of the lowest energy structures is substantial [25]. Unfortunately, these very low energy states are rarely populated, and the folding transition is only very weakly cooperative. Furthermore, the models have much in common with the molten globule state of proteins [63–66]. They have substantial native-state secondary structure, but there is no fixation of tertiary contacts, and the manifold of structures tend to be swollen relative to the native state [18,25]. This can be rationalized as follows. Both one-body and pair interaction terms lack sufficient specificity to produce a native conformation that has a substantial energy gap with respect to other relatively nearby conformations. This results in an almost continuous transition from the unfolded state.

The higher order multibody component of the potential is only important when one has dense compact states; it permits, but does not require, side-chain fixation. Furthermore, without such terms, microphase separation of the side chains results, with an unphysical number and pattern of side-chain contacts. However, a key question is whether such potentials are physical or arise simply because reduced models are considered. The presence of reduced models certainly suggests that to some extent one must modify the interactions to reproduce the finer details of side-chain packing. However, even in molecular dynamics simulations of full-atom protein models, on starting from the crystal structure, the simulations tend to diffuse native side-chain packing towards a more liquid state [67]. Thus, the problem with extant potentials may be much deeper. Finally, we note that the potentials we are using are potentials of mean force. It is a well-known result from the statistical mechanics of small-molecule liquids that higher order correlation functions (for example, the three-body radial distribution function) are not simply factorizable into lower order distribution functions (this approximation is the Kirkwood superposition approximation), even if the naked potential is pairwise additive [68].

In order to introduce the possibility of side-chain fixation, Kolinski et al. [18,25,29,30,53,62] examined two classes of multibody terms. The first is of the form

$$E_4 = \sum(\varepsilon_{ij} + \varepsilon_{i+k,j+n})C_{ij}C_{i+k,j+n} \qquad (2)$$

with $|k| = |n|$. $C_{ij} = 1$ when side groups i and j are in contact; otherwise $C_{ij} = 0$. $\varepsilon_{i,j}$ is the pair potential between amino acids i and j. We have considered models with $n = 3$ and 4. As indicated in Figs. 2A and B, such patterns are typical of both helix-to-helix and beta-to-beta side-chain packing patterns. We have also included $n = 1$ terms which are typical of beta-to-beta contacts. A second implementation of the multibody potential involves the use of a neural network to recognize whether or not 7-residue by 7-residue subfragments of dense regions of contact maps are native-like [62]. Such a formulation can in many cases recognize misfolded proteins based on contact maps alone. It offers the advantage that many more kinds of contact patterns are considered

than are possible based on Eq. 2. On the other hand, the neural network does not consider the identity of participating amino acids. Although it was later modified to include the average pair potential of such interacting subfragments [29], it ignores the effect of side-chain size and may result in nonphysical packing arrangements.

Synergism of the contributions to the potential

Based on a large variety of simulations, we conclude that there is no single dominant interaction responsible for protein folding. In agreement with Hao and Scheraga [33–35], we conclude that the contribution to the stability of a protein due to interactions reflecting intrinsic secondary structure propensities (local hydrogen bonding plus local conformational preferences) is roughly equal to that of tertiary interactions [18,23,53]. Hydrogen bonding acts to restrict the manifold of compact states to those which are almost maximally hydrogen bonded, thereby reducing the conformational entropy of compact states. Similarly, intrinsic secondary structural preferences, although inherently weak, bias the system towards the secondary structure found in the native state. Of course, these can be overridden by tertiary interactions. Hydrophobic interactions create the average phase segregation of the amino acids. That is, in a typical protein roughly 75% of all hydrophobic residues are buried. However, since all hydrophobic residues are not buried, this argues that there are interactions (e.g. the resulting compact structure might not be hydrogen bonded) that oppose the burial of all hydrophobic residues. Pair interactions help reduce the configurational entropy of compact states by acting to break the degeneracy of compact structures and may serve to destabilize alternative conformations as well as to stabilize the native fold. Finally, higher order packing interactions might be responsible for the fixation of structure on passage from the molten globule to the native state [25].

Our simulations argue that a protein is a system under tension in the sense that while the system is in a global free-energy minimum, this minimum arises as a compromise between all the above terms [53]. The native conformation lies in that portion of conformational space consistent with the interplay of the interactions that comprise a globular protein. By eliminating any given class of terms, an important physical feature of a protein is removed. Thus, in these models, there is no single dominant term driving protein folding; rather the stability of the native state arises from the consensus and interplay of a number of terms representing different physical effects.

Monte Carlo sampling schemes

The well-known Metropolis Monte Carlo (MMC) procedure randomly samples conformational space according to the Boltzmann distribution of (distinguishable) conformations [32]:

$$P_i = \exp(-E_i/kT) \qquad (3)$$

In order to generate this distribution, the transition probability $p_{i,j}$ from an 'old' conformation i to a 'new' conformation j (for the asymmetric scheme) is controlled by the energy difference $\Delta E_{ij} = E_j - E_i$ via

$$p_{i,j} = \min\{1, \exp(-\Delta E_{ij}/kT)\} \quad (4)$$

Obviously, this technique is very sensitive to the presence of energy barriers. To ensure adequate sampling, typically a collection of elemental backbone moves involving end moves, and collective motions of two to four bonds are randomly performed. In addition, small-distance motions of a large, randomly selected part of the chain are employed. Side chains can also independently move. The key to a successful dynamic Monte Carlo protocol is to include a sufficiently large move set so that no element of structure is artificially frozen in space.

To enhance the sampling efficiency, Hao and Scheraga [33–35] have employed the entropy sampling Monte Carlo method (ESMC) in their study of simplified protein models. ESMC was originally proposed by Lee [69] in the context of a simple Ising model and is closely related to the multicanonical MC technique of Berg and Neuhaus [70]. Since the formulation of Hao and Scheraga is the most straightforward and has been applied to both simplified and higher resolution models, we briefly review their approach.

Unlike MMC, ESMC generates an artificial distribution of states that is controlled by the conformational entropy as a function of the energy of a particular conformation E_i:

$$P_i^{ESMC} = \exp(-S(E_i)/k) \quad (5)$$

The transition probability can be formally written as

$$p_{i,j}^{ESMC} = \min\{1, \exp(-\Delta S_{i,j}/k) \quad (6)$$

with $\Delta S_{i,j}$ being the entropy difference between energy levels i and j, respectively.

At the beginning of the simulation, the entropy is not known. However, from a density-of-states energy histogram, H(E), an estimate, J(E), for the entropy S(E) can be iteratively generated. The kth iteration consists of an ESMC simulation run with S(E) approximated by $J_{k-1}(E)$. Here,

$$J_k(E) = J_{k-1}(E) + \ln(\max\{1, H_k(E)\}) \quad (7)$$

After a sufficient number of runs, all the states are sampled with the same frequency. Then, the histogram of H(E) becomes flat, and the curve of J(E) + constant approaches the true S(E) curve.

Folding protocol

For each sequence considered, starting from arbitrary random conformations, a series of independent simulated annealing experiments are performed. In many cases, at least 10, and more recently at least 20, independent simulations are

performed, and the resulting minimum-energy structures are clustered according to global topology. If the dispersion in topologies is large, then the sequence is viewed as being nonfoldable using that generation of the model and its associated potentials. For those sequences that produce a handful (less than four topologies), then each of the topologies is subjected to an isothermal stability run. In a number of cases, the topologies which result are the native fold and its topological mirror image. For example, as shown in Fig. 3, there are left- and right-turning four-helix bundles. In both cases, the helices are right-handed, but the chirality of the topology is reversed. The structure with the lowest average and minimum energy is assigned to be the predicted native state. The resulting lattice model with side chains is then pulled off-lattice, and the backbone and side chains are reconstructed using the procedure described in Ref. 21. To date, the reduced and full atom models are completely compatible [21,29,30,36].

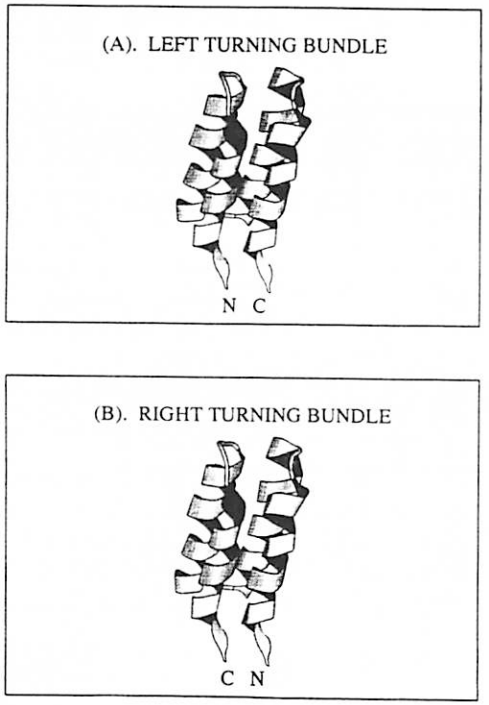


Fig. 3. Schematic illustration of (A) left- and (B) right-turning four-helix bundles.

Folding of exaggerated helical protein sequences

Using a lower coordination lattice, an early set of *de novo* simulations [18] (i.e. folding without any encoded knowledge of the native conformation) was performed on two 73-residue sequences designed by DeGrado and co-workers [71,72]. The first sequence contained an all-leucine core. In excellent agreement with experiment, this sequence is predicted to form a thermodynamically very stable four-helix bundle, but one with nonunique side-chain packing. The simulated sequence had many of the properties of a molten globule state. It had substantial secondary structure, and its average mean-square radius of gyration was about 15% larger than that found in the native state of a redesigned sequence (see below). Moreover, the simulations predicted that the right- and left-handed four-helix bundles should be isoenergetic. This prediction was subsequently confirmed by experiment [73]. Finally, within each topology, the molecule migrates among a few distinct families of structures which share the same global topology, but which differ in the identity of the residues which stabilize them.

A second sequence designed by DeGrado had 14 amino acid substitutions in the hydrophobic core [72]. In contrast to the first sequence, due to sequence heterogeneity in the hydrophobic core, differential pair interactions break the degeneracy of the various structures, and this molecule is predicted to prefer the right-turning, four-helix bundle topology. Following rapid assembly using standard MMC to a four-helix bundle topology, the molecule slowly relaxed to a more compact structure which is unique at the level of resolution of this class of models. Moreover, since the energy monotonically decayed as a function of time during the relaxation process, this implied the existence of entropic barriers between the compact molten globule-like state and the predicted native conformation. Fixation of the side chains was observed to occur when higher order multibody terms of the type given by Eq. 2 are included in the model, but it does not happen if such terms are deleted. These simulations pointed out the importance of including a cooperative protein-like interaction scheme into the potential used in folding.

Factors responsible for the uniqueness of the native structure

The full lattice model described above was used to explore the requirements for the *de novo* folding from an arbitrary random conformation of idealized sequences of four- and six-stranded β-barrels [23,25]. Of particular interest is the design of a putative 45-residue, six-stranded β-barrel which adopts the schematic topology shown in Fig. 4A. Simulations using MMC were used to test various possible conjectures about the factors responsible for the structural uniqueness of the native state [23]. Among these were the relative importance of generic hydrophilic/hydrophobic amino acid patterns, and the possible role of polar amino acids in destabilizing misfolded conformations [37].

A simple alternating pattern of valines and serines in the putative β-strand regions, when punctuated by appropriate turn-forming residues, is found to produce

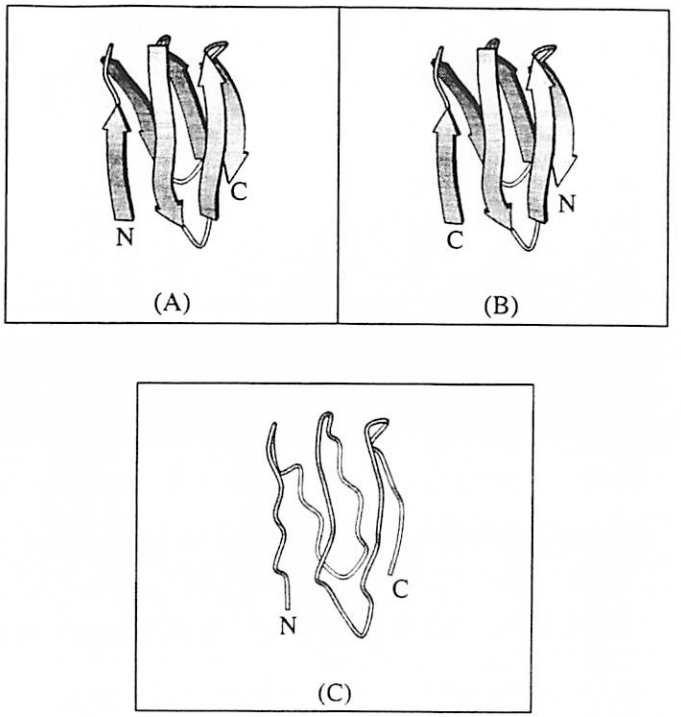


Fig. 4. Schematic illustration of (A) the desired six-stranded β-barrel and (B) the mirror image barrel. The predicted C^{α} trace of the native structure of the designed sequence betamod is shown in (C).

a manifold of six-stranded β-barrels having different topologies. This implies that a simple HP (nonpolar polar) model is not sufficient to yield a structurally unique native state when systems having conformational entropy on the order of that of real proteins are considered. Furthermore, the packing of the resulting hydrophobic core is very diffuse. Thus, to enhance the stability of the hydrophobic core and to partially break the degeneracy of the various topologies, four Phe residues were introduced into the sequence. This reduced the number of observed topologies; however, the topology was not uniquely defined. Substitution of Asp for Ser residues was done at positions designed to destabilize incorrect topologies. The resulting sequence adopted the desired as well as the mirror image topology shown in Fig. 4B. Analysis of the energetic contributions indicated that the packing interactions favored the desired fold, but that the residues introduced in the turns favored the mirror image topology.

Substitution with Gly linkers resulted in the desired native fold becoming the most stable topology. The resulting designed sequence, called betamod, is given by

GVDVDV-GGG-**VDVDV**-GGG-**FRFRV**-GGG-**VRFRF**-GG-**VDVDV**-GGG-**VDVDV**

The residues in bold indicate the location of the putative β-strands. Strands 1, 4, and 5 form the first β-sheet, while strands 2, 3, and 6 form the second sheet. The loop/turn regions are composed of flexible Gly connectors. A representative conformation obtained from the simulations is shown in Fig. 4C.

A question remains as to whether this sequence would in reality adopt a unique native conformation, a molten globule state or would not fold at all. Thus, experimental examination of this sequence is currently underway in the laboratory of Dr. Derek Woolfson [74]. In the interim, the results of these simulations suggest that these models might prove to be useful tools in protein design.

Origin of the cooperativity of protein folding

A very important question is whether simplified models can reproduce the thermodynamic behavior of proteins. Experimentally, in a number of proteins, the cooperativity in protein folding arises on passage from the molten globule state to the native conformation [63,64,75]. Such molten globules or compact intermediates have a volume which is about 50% larger than native, a substantial amount of native secondary structure, but diffuse tertiary contacts. These observations suggest that the fixation of side chains accompanying the transition to the native conformation is involved in the cooperativity of protein folding.

To investigate the possibility of a first-order transition in protein folding, Hao and Scheraga [33] employed the ESMC method to examine the folding thermodynamics of a 38-residue protein confined to the 210-lattice introduced by Kolinski et al. [14] and Skolnick and Kolinski [76]. Subsequently, they examined the sequence requirements for an all or none transition [34,35]. They conclude that designed or optimized sequences exhibit a cooperative folding transition which is long-range (i.e. involves tertiary interactions), whereas random sequences fold to compact states by what is an essentially continuous transition. These are very important studies, because they show for the first time in a nontrivial model that adoption of a unique low-energy state depends on the interplay of long- and short-range interactions. These model proteins included a local conformational bias for native-like secondary structure and a single side-chain rotamer for each residue.

Subsequently, Kolinski et al. [25] employed the ESMC method to investigate the folding thermodynamics of betamod. These studies build on the Hao–Scheraga work in the following ways. Now, a much higher coordination lattice is used, there is no target bias for the native state's secondary structure, and multiple side-chain rotamers are present so that the possibility of side-chain fixation exists. These three differences result in a model which has a considerably higher entropy in the compact state.

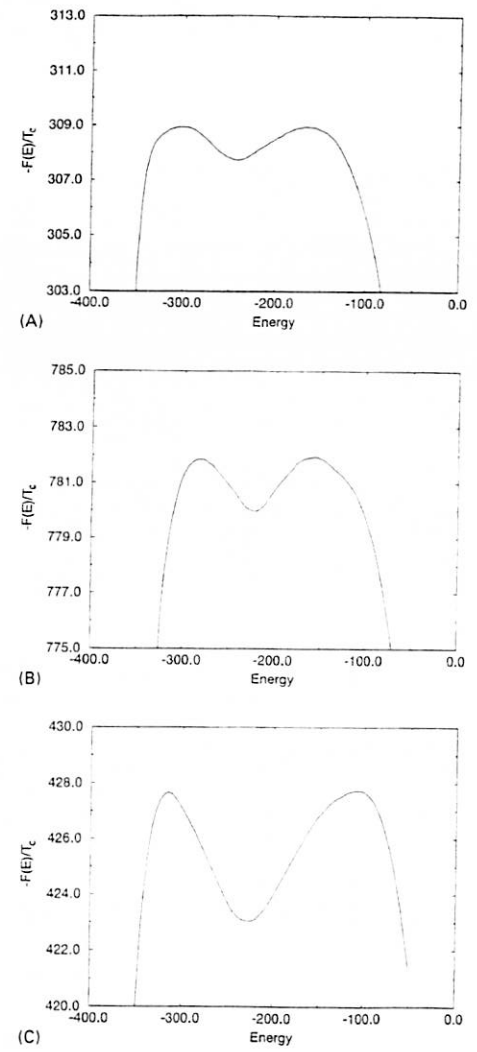


Fig. 5. Plots of the free energy, F/T_c, versus energy, E, for models I–III at the transition midpoint temperature, T_c.

Three distinct versions of tertiary interactions were considered. In the first, model I, only pair potentials are used. Model II also includes n = 3 and n = 4 type terms given by Eq. 2. Model III extends model II to include beta-type n = 1 terms. The scale factors for the pair and multibody interactions have been adjusted so that the tertiary interaction energy in the putative native fold of all three models is essentially the same.

As shown in Fig. 5, where the reduced free energy, F/T_c, versus energy, E, is plotted at the folding transition temperature, T_c, qualitatively different behavior is seen on passage from model I to model III. Model I, lacking high-order multibody interactions, essentially has a continuous thermodynamic transition. With the inclusion of higher order multibody packing interactions, the conformational transition becomes all or none. Interestingly, the lowest energy states in all three models correspond to the same manifold of structures (i.e. structures which are unique at the level of resolution of the lattice models) and correspond to the native fold shown in Fig. 4C. What differs in the three models is the separation of the low-energy native-like state from the manifold of other conformations that contribute to the partition function.

Nature of the transition state

The nature of the conformations located at the free energy versus energy maximum, viz. the transition state, was examined. In models II and III, which exhibit two-state

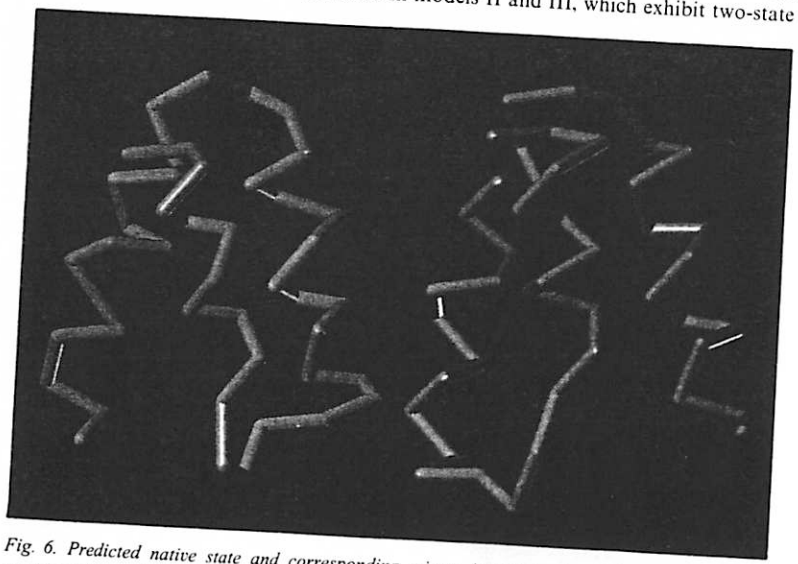


Fig. 6. Predicted native state and corresponding mirror image topology of the B domain of protein A shown in green and magenta, respectively.

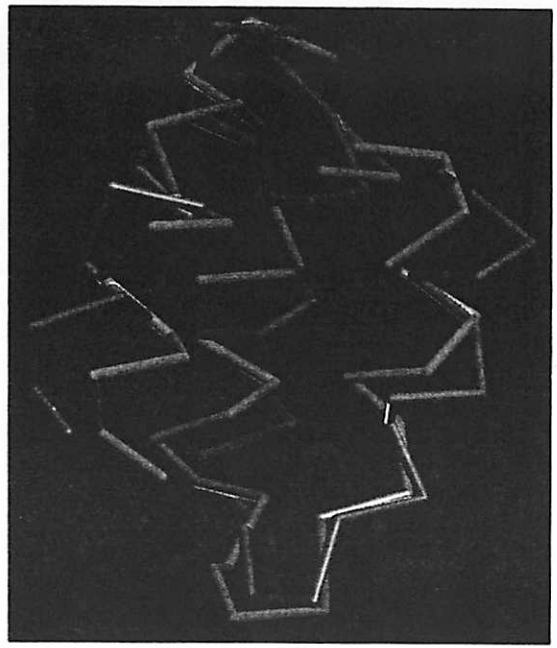


Fig. 7. Predicted backbone trace of protein A in magenta, superimposed on the experimental NMR solution structure in green.

thermodynamics, the transition state is comprised of structures having about 60% of the native state's secondary structure, about 50% of the side-chain contacts which are native, and a volume which is about 50% larger than native. This description of the transition state supports Kuwajima's [63] critical substructure model, where the activated state has a partial amount of native secondary structure, there are a subset of native contacts and the molecule is swollen relative to the native conformation. Such a range of physical properties has also been experimentally observed in a number of systems including α-lactalbumin and calcium-binding parvalbumin [77]. These simulations suggest that cooperative many-body interactions involving protein side-chains are the dominant factor responsible for the cooperativity of protein folding in models where the side-chains have internal degrees of freedom and perhaps in real proteins as well.

Folding of domains of protein A

In solution, the B domain of protein A is a 55-residue protein in which residues 10–55 adopt a three-helix bundle geometry [78]. Because of its structural simplicity

and small size, this protein is a natural testing ground for *de novo* prediction methods, and a variety of generations of the model have been applied to this molecule [20,29,36]. Initially, folding of this molecule was attempted on a coarser lattice, followed by refinement on the 90-neighbor (finer) lattice described above [20,36]. Based on both the average and minimum energy, the correct topology is chosen over the mirror image, both of which are depicted in Fig. 6 in green and magenta, respectively. For this sequence, all contributions to the energy favor the native state. The resulting structures have an rms from native for residues 13–55 of 3.3 Å. A typical predicted conformation superimposed on the solution NMR structure of Gouda et al. [78] is shown in Fig. 7. Subsequent refinement of the model showed that the folding on a coarser lattice followed by refinement on the finer lattice (which permits better helix-to-helix packing) is unnecessary; rather, direct folding on the finer lattice is a more straightforward and simpler procedure [20,29,36].

In the original simulations, folding tended to occur by the preferential formation of the N-terminal hairpin, followed by assembly of the final helix [20]. Subsequent simulations on a more refined model suggest that the C-terminal hairpin assembly is more likely. There is also an indication that the C-terminal helix of the B domain of protein A may be stable in solution. There is some experimental indication that this might be the case [79]. In agreement with experiment, the simulations predict that the folded state is native-like with very long-lived side-chain contacts.

The B domain is but one of five highly homologous, extracellular domains of protein A designated as E, D, A, B, and C, respectively. These five domains all bind to immunoglobulin. Thus, it is a reasonable conjecture that all have the same solution structure. In addition, Montelione and co-workers have determined the NMR solution structure of the Z domain of protein A, which differs from the B domain by the single-point mutation G30A [80,81]. They find that it has a very similar fold to that of the B domain. In order to investigate the ability of the folding algorithm to fold homologous sequences, the folding of all five wild-type domains and the Z domain was successfully undertaken. In all cases, the native topology is energetically favored over the mirror image topology, with the C domain exhibiting the smallest energetic preference for the native over the mirror image fold. For all six sequences, the final structures are within 3.5 Å rms of the predicted B domain conformation.

Redesign of protein A to adopt the mirror image topology

A key question in understanding the principles of protein folding is the origin of the preference for a given topology as opposed to the topological mirror image. Two viewpoints have emerged: in one, the topology is dictated by the packing interactions in the hydrophobic core [82,83], and in the other the turns play a role in dictating the preferred topology [84,85]. To examine these questions, we attempted to redesign the sequences of the B and A domains of protein A so that they adopt the mirror topology shown in Fig. 6 in magenta. Both multiple mutations in the hydrophobic core and in the turn regions between helices I and II were made. To scan a large number of

mutations to search for sequences that favor the mirror image over the native fold, a sieve method was developed. Modifications in the hydrophobic core were made in three groups, each involving point mutations at six sites. At each site, the native residue was replaced by Ala, Val, Ile, Leu and Phe. Thus, 15 625 mutations were examined for each group of mutation sites. For those sequences which survived the sieve procedure, none was found to prefer the mirror image over the native fold. Therefore, these results are consistent with the idea that the fine details of hydrophobic packing do not constitute the sole driving force for the folding process, but may stabilize an already acquired motif [86].

Next, two-point mutations in the turn connecting helices I and II were examined. With the exception of glycine, proline and cysteine, all possible mutations of Asn^{22} and Asn^{24} were allowed. Again applying the sieve procedure, most mutations in the turn regions do not disrupt the preference for the native fold. This is qualitatively consistent with experiments which indicate that, in general, turns can be modified without qualitatively changing the global fold [87,88]. However, about 11 of the 289 mutants, i.e. about 4%, resulted in sequences having varying degrees of preference for the mirror image topology.

The most promising N22R and N24M double mutant was subject to further evaluation. While the probability of finding an Arg in the i+2 position of the turn is relatively high, Met is rarely seen in the N-caps of helices [89]. To confirm that the RM mutant is foldable (at least in computro), a series of 10 independent MMC folding simulations were undertaken. In 6 of 10 simulations, the mirror image topology is obtained, with the remainder adopting the native fold. The RM mutation modifies the intrinsic secondary structure preferences so that in contrast to the wild-type they now favor the mirror image turn. This tendency is further augmented by the burial of M in the mirror image, but not in the native fold. This produces a net favorable pair interaction for the mirror image topology. In other words, the predicted preference for the image topology arises from the favorable juxtaposition of intrinsic secondary preferences and tertiary interactions [85]. These models indicate that such a juxtaposition is what is responsible for the adoption of a particular fold. To ensure that the predicted lattice models are consistent with atomic resolution models, all-atom models were constructed and were found to be in complete qualitative agreement. The experimental test of this prediction is now underway in Dr. Peter Wright's [79] laboratory here at Scripps. Finally, to examine the robustness of the RM mutation, it was applied to the A domain of protein A. The simulations predict that this sequence should also be a likely candidate for adopting the mirror image topology.

Effect of amino acid order on folding

By reading the sequence of a naturally occurring protein backwards, i.e. generating a retroprotein, a sequence of the same composition and hydrophobicity as the wild-type protein results [90]. However, proteins are chiral systems, and there have been a number of conjectures as to whether retroproteins will fold, and, if so, what

topology will they adopt. Some authors have gone so far as to suggest that a retroprotein might adopt the mirror image structure, including left-handed helices [91]. While this conjecture is unlikely, there are a number of consequences accompanying the retroinversion of a sequence. All prediction methods based on composition will predict the same structural class for the wild type and retroprotein [92]. However, if the positions of the helices and turns remain the same as in the wild type, then in general the locations of the capping [93,94] and turn residues will not be optimal [89]. At a minimum, one might expect some rearrangement of the secondary structural elements. If side-chain packing and/or the distribution of nonpolar side chains is a dominant factor in determining the global fold, then one might expect the retroprotein to adopt the same topology as that of the native protein. There is always the possibility that an entirely different fold might be adopted or the retroprotein might not fold at all.

Because of the robustness of the lattice folding algorithm as applied to protein A, the retrosequence of the B domain was generated and subjected to a series of 15 folding experiments [30]. The results strongly suggest that the predicted native state of retroprotein A, shown in Fig. 8, is a three-helix bundle of the same topology as the wild-type sequence. It is important to emphasize here that the prediction that the retrosequence adopts the same topology as the wild-type sequence may be due to

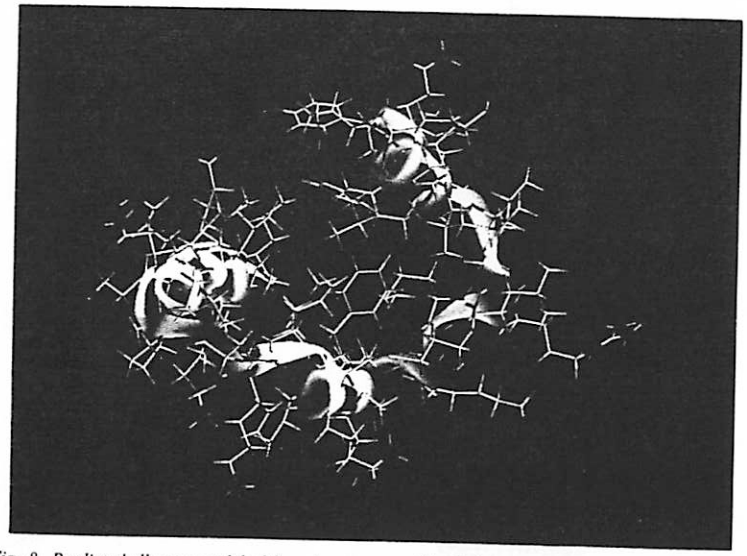


Fig. 8. Predicted all-atom model of the retrosequence of the B domain of protein A. The ribbon tube depicts the position of the backbone atoms and clearly indicates a three-helix bundle topology.

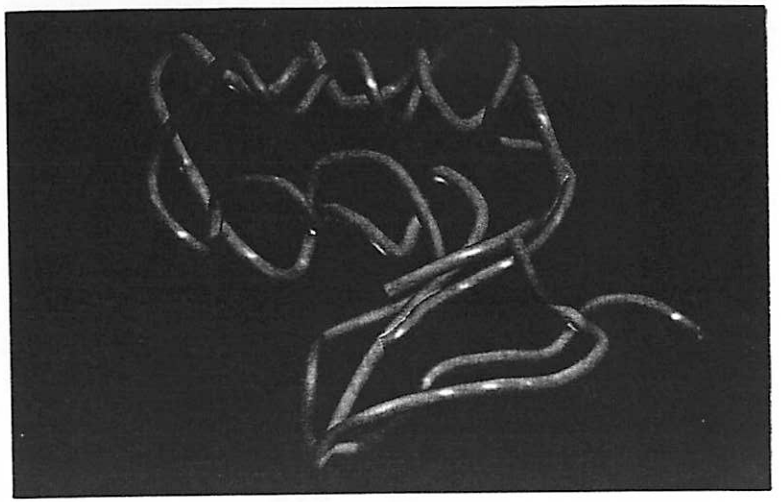


Fig. 9. Predicted C^{α} trace of the native conformation of crambin, in magenta, superimposed on the crystal structure, in green.

the high symmetry of the three-helix bundle fold, and it is very likely that this result is not true in general.

To accommodate the local secondary structural propensities, the secondary structural elements shift their positions with respect to the B domain. Among the most salient changes is the shift in the location of the C-terminal turn. This adjustment in position allows for the third helix to have N-cap residues that are favored. Furthermore, the predicted structure retains many of the hydrophobic core contacts as in the B domain. This suggests that hydrophobic interactions exert an important influence on driving the system to adopt a three-helix bundle topology. However, pair interactions alone are isoenergetic in the native and mirror image fold. What drives the system to favor the native fold is the difference in burial energies of the two topologies. This implies that in this case burial preferences select out the native over the mirror image topology.

Subsequently, atomic models were built from the lattice structures. In all cases, the hydrophobic core is well packed. Depending on the starting structure, the N- and C-terminal ends of the helices vary by about one residue. Overall, the lattice and all-atom models are consistent. Encouraged by these results, the structure of this sequence is now being determined by Dr. Chi-Huey Wong's group [95] at the Scripps Research Institute.

Folding of ROP monomer

The native structure of wild-type ROP is a dimer consisting of two antiparallel helical hairpins arranged in a coiled-coil geometry [96]. Sander and co-workers have redesigned this molecule to form a 120-residue, monomeric, left-turning, four-helix bundle [97]. Subsequently, Regan et al. have also redesigned the dimer to form a monomer using glycine linkers of various lengths [88]. In simulations done to date, the original Sander sequence has been used; work is in progress to fold those sequences designed by Regan et al. In the earlier simulations, folding commenced on a coarser lattice from random geometries [19,20]. A very strong preference for the designed, left-turning bundle was indicated. As in the case of protein A, the resulting low-energy structures were then projected onto the 90-neighbor lattice and refined. The predicted structures have a C^{α} rms ranging from 2.6 to 4.2 Å with respect to the set of equivalent residues in the ROP dimer crystal structure. What is striking is that the simulations predict that the molecule has less supertwist than is found in the ROP dimer structure. Whether these predictions are true or not awaits the experimental determination of the ROP crystal structure.

The folding simulations predict the existence of late, presumably molten globule, folding intermediates that are present prior to the formation of the native state. These metastable intermediates have the same global fold as native, but their radius of gyration is about 5% larger. Similar chain expansions have been observed in an apomyoglobin folding intermediate [98]. The secondary structure is essentially identical to native but there are much larger fluctuations in the turn regions and at the chain ends. In the molten globule, none of the side-chain contacts survives for the entire simulation run, while in the native state there are many such long-lived contacts. Furthermore, in the molten globule, the side-chain contact patterns are more diffuse, which is consistent with the observation that the helices are sloshing back and forth against each other. Detailed analysis of the dynamics of the molten globule state indicated that it is very liquid-like, and has much in common with the dynamics of a gel. In contrast, the native conformation is much less mobile, with the displacements (apart from global diffusion) limited to relatively small-scale motions.

Folding of crambin and the use of predicted secondary bias to enhance folding efficiency

To address the concern of whether the model can predict the tertiary structure of α/β proteins, the folding of crambin was undertaken [19,20]. This 46-residue protein has a native state comprising a helical hairpin and a three-stranded antiparallel β-sheet. It also contains three disulfide cross-links [99]. The simulations do not assume anything about a specific cross-link pattern, but rather that cystines of some sort are present. When straightforward folding from the random state was undertaken, the correct topology and disulfide pattern is always recovered, but in most cases the secondary structure, especially in the putative helical regions, is highly distorted.

To alleviate this problem, higher temperature simulations were undertaken where the S-S bond dissociation rate is sufficiently high. Then, statistics about secondary structure preferences (helix/turn or extended/loop) are collected. This prescreening predicts the location of the N-terminal helix with a shift of two to three residues towards the amino terminus, but the prediction for the second helix is more accurate. This stands in contrast to most standard secondary structure prediction methods which mostly predict β-strands in these regions [100,101]. With an approximate prediction of the helical regions in hand, a small energetic bias (proportional to the helicity of a given residue at the higher temperature) is added to the model. In about 50% of the folding simulations, low-energy conformations having a helical hairpin whose C^{α} rms from native is about 4 Å are predicted. The other conformations preserve the global topology of the native fold, but are 20% higher in energy. Subsequent refinement at low temperature produces structures whose average C^{α} rms is below 4 Å. For residues 3–42, the average coordinate rms is 3.6 Å, with a distance rms of 2.6 Å. As is evident from Fig. 9, which shows the predicted structure in magenta superimposed on the backbone of the crystal structure, in green, while the global fold is well reproduced, there are slight shifts in the position of one of the helices and the conformation of residues 43–46 is incorrect.

This protocol has also been applied to the folding of protein A, with comparable results obtained as when the predicted secondary structure bias is not incorporated into the folding algorithm. In addition, the protocol has been employed to predict the tertiary structure of the V-3 loop of gp-120 [102]. The resulting conformation is suggested to consist of three β-strands and a small C-terminal helix. The results of these simulations suggest that either experimental or predicted secondary structural constraints can be incorporated into the folding algorithm. Such predicted biases can greatly speed up the folding process. However, care has to be taken to ensure that if the prediction is uncertain, it can be overridden by other interactions.

Method for the prediction of surface U-turns and transglobular connections in small proteins

A knowledge of the locations where the chain changes its global direction, i.e. the U-turns, and of the dominant secondary structure of the intervening transglobular regions, i.e. the blocks, represents very useful information for a folding algorithm [27]. Thus, a simple method for predicting these building blocks in small single-domain proteins has been developed. Such an approach is complementary to more standard secondary structure prediction schemes [103]. Here, global rather than local information is desired; the structural assignments depend on the conformation of the entire chain. For example, if a given region favors helix, this tendency can be overridden because another part of the chain has a lower energy if it is helical and the first helical region is shifted to form a turn.

Table 1 *Summary of prediction statistics for the blocks and U-turns algorithm*

Protein name[a]	Surface U-turn prediction accuracy[b]	Errors of U-turn locations[c]	Secondary structure block prediction accuracy[d]	Comments on wrong assignment
1gb1	4/4	0-2/2*-3-0	5/5	–
proA	2/2	2-3	3/3	–
1fas	5/5	2-1-0-2-0	5/5	Terminal coiled assigned β, inserted β without a turn
1pou	3/3 1 over	4-3-2-0	4/4	Extended coil inserted
1tlk	7/7	5-3-2/1-1-2-4-7	7/8	Turn inserted into the C-terminal β-strand
1ris	5/5	3-5-6-3/4-4	5/6	Second β-strand predicted helical
1lpt	4/4	0-5-2-0	4/4	Shifted turns, hairpin-like C-terminus predicted as β
1ten	6/7	1/1-3-0-3-2-2/1	7/8	Shifted turns, one β-strand missed
1mjc	5/5	1-2-0-2-0	6/6	Long central coil added as β
–	41/42 = 98% 1 over	–	46/49 = 94%	Including overpredicted turn in 1tlk

[a] PDB descriptor.

[b] The ratio of the correctly predicted number of surface U-turns to the actual number in the protein. A turn is said to be correctly predicted if its boundaries at least partially overlap with the actual turn location. 'Over' indicates that an additional U-turn(s) is predicted which does not occur in the protein structure.

[c] i/j means that i residues of the preceding block and j residues of the following block have been incorrectly assigned as a part of the surface loop/turn. Otherwise, the number of overassigned residues of one of the transglobular blocks is given.

[d] The ratio of the correctly predicted number of secondary structure blocks to the actual number of surface turns in the protein. The secondary structure of a given block is said to be correctly predicted when the secondary structure of the three central residues agrees with the experimental structure.

The method consists of five basic steps:

(1) Estimate the radius of gyration of the protein. This imposes restrictions on the maximum and minimum length of the extended and helical fragments that can fit into the globule.
(2) The locations of the U-turns are randomly chosen and hairpins appropriate to the chain division are pulled from a database of protein structures. While a lattice realization of the structures is used for computational convenience, in principle, the approach is completely general.
(3) The energy, consisting of local secondary preferences, a centrosymmetric burial term, and a term which reflects the orientation of the hydrophobic face with respect to the core, is calculated.
(4) The division process is repeated many times.
(5) At the end of the selection process, the statistics on the set of lowest energy structures are performed. The location of the predicted U-turns is established, and the dominant secondary structure in the three central residues between U-turns is used to assign the secondary structure of the entire block.

Application has been made to a set of test proteins, and part of the results are summarized in Table 1. At least for the testing set, the method is quite accurate, with over 90% of the U-turns and blocks correctly predicted. In six of the nine test sequences, the number of U-turns and the secondary structure of the blocks are correctly predicted. These encouraging results suggest that the blocks and U-turns algorithm holds considerable promise in providing important information for three-dimensional modeling procedures. When successful, it provides sufficient information to propose a relatively small number of low-resolution alternative folds. Furthermore, it can be used as a filter or constraint in inverse folding algorithms either to predict the global topology or, in a potentially more powerful application, to predict the conformation of hairpin fragments. At present, when inverse folding algorithms are used to predict the structure of 15–20-residue pieces of a protein, mixed in with the correct, low-energy structures are a variety of comparable energy false positives. The blocks and U-turns algorithm can be used to filter out such false positives. Preliminary application of this combined approach has yielded promising results.

Folding with a small number of long-range restraints

A number of investigators have examined the problem of determining a low to moderate resolution protein structure given a relatively small number of distance restraints and some knowledge of the secondary structure [26]. The ability to predict such structures would aid in the early stages of NMR structural refinement when secondary structure information and a limited number of distance restraints are known. In contrast, when a large number of restraints are available, then the use of distance geometry or distance geometry supplemented by molecular dynamics are the methods of choice [104]. Here, we describe results when the lattice model of protein folding is supplemented by a rough knowledge of secondary structure and some

tertiary constraint information. Such an investigation can also clarify whether the present realization of the model is basically correct, but is simply in need of further refinement, or more substantial problems with the model exist which would require its fundamental reformulation.

From random extended states, the folding of L7/L12 ribosomal protein, 1ctf, protein G, 1gb1, and thioredoxin, 2trx, all of which are α/β proteins, plastocyanin, 1pcy, which is an all-beta protein, and the helical protein, sperm whale myoglobin, 1mba, were undertaken [60]. After a simulated annealing run, the resulting final conformation is subject to isothermal refinement. At least five, and in many cases 20, independent folding/refinement runs were performed. For the sake of brevity, the simulation results for the run having the lowest average energy are presented in Table 2.

For the three α/β proteins considered, it is apparent that reasonable structures are obtained when there is on the order of one long-range constraint every seven residues. Similar results are found for helical proteins. Reflecting inherent problems in the model, this class of models requires a greater number of restraints for β-proteins. For the β-protein plastocyanin, one tertiary constraint every four residues is required.

These results should be compared to those of Smith-Brown et al. [105]. To obtain results of comparable accuracy to plastocyanin folded with 46 restraints, they required 90 restraints to fold a variable light domain of human immunoglobulin, 3Fab [60]. Similarly, Smith-Brown et al. require 147 constraints to obtain a structure that is 3.18 Å rms from the native conformation of flavodoxin [60]. In contrast, preliminary results on the folding of flavodoxin with 35 tertiary restraints indicate that structures on the level of 4 Å rms are obtained.

Aszodi and co-workers [106] have applied a distance geometry protocol where the secondary structure is known and where correct constraints are supplemented by predicted interresidue distances based on multiple sequence alignments. They refold thioredoxin with about 48 restraints to structures whose rms is about 5.0 Å. However, for a smaller number of restraints, the structures are almost random, having an rms from native on the order of 10 Å. In contrast, with just 15 restraints, structures on the

Table 2 *Results from NMR docking simulations with a limited number of tertiary restraints*

Protein	Number of residues	Number of tertiary restraints	Average coordinate rms[a]	Average distance rms[a]
1gb1	56	8	3.81	2.72
1ctf	68	8	4.27	3.13
2trx	108	15	6.60	4.77
1pcy	99	46	3.32	2.46
1pcy	99	25	6.22	3.93
1mba	146	20	5.52	3.85

[a] Root-mean-square deviation of the C^{α} coordinates in Å.

level of 6.6 Å rms are obtained here. Moreover, in the case of helical proteins such as 1mba, the docking algorithms can assemble the approximate topology. This suggests that the present lattice-based approach could be used to generate low to moderate resolution structures from a rather small number of restraints. However, the present realization needs improvement. Short-range restraints are incorporated as a very soft energetic bias to helix, turn or extended conformations, as appropriate. Tertiary restraints are also very loosely defined as operating on the level of side-chain contacts. Better restraints that actually include the information contained in 2D and 3D NMR experiments should be implemented. Similarly, disulfide bond restraints, which have a very specific geometry, could also be included. Thus, the results from this very simple realization of tertiary restraints could possibly be improved by better use of experimental data.

Folding of the GCN4 leucine zipper

Because of their sequential and structural simplicity and biological importance, coiled coils are natural test systems for protein folding and multimer assembly algorithms. The simplest realization of the coiled-coil motif consists of two α-helices wrapped around each other with a left-handed supertwist. Furthermore, coiled-coil sequences are characterized by a quasirepeating heptad of residues designated by the letters a–g, where positions a and d occur in the coiled-coil interface [107]. A particularly well characterized coiled coil is the leucine zipper of the transcriptional activator, GCN4 [108,109]. Each protein chain contains 33 residues, and it has a high-resolution crystal structure. To predict the GCN4 quaternary structure, Nilges and Brunger [110,111] assumed an initial conformation consisting of an idealized, parallel coiled coil. They were able to refine the structure from an initial 3.1 Å rms on the backbone atoms to a level of 1.26 Å rms for the backbone atoms and 1.75 Å for all heavy atoms in the dimerization interface. To accomplish this refinement, they used molecular dynamics supplemented by imposed helical backbone hydrogen bond restraints and a number of distance restraints.

Vieth and co-workers [21] have employed a hierarchical approach to fold the GCN4 leucine zipper from two chains that were initially in random conformations. No information about the global fold is assumed other than that there are two chains in a box. Thus, the possibility of higher order multimer formation was not considered there. First, a high-resolution lattice model is employed to assemble the topology. The lowest energy lattice structures have an rms from the C^{α} trace of the crystal structure ranging from 2.3 to 3.7 Å. Then, using these structures, detailed atomic models were built and relaxed using CHARMM all-atom building and MD-based simulated annealing in explicit water [112]. The average structure built from the entire family of five independently refined conformations has an rms deviation of 0.8 Å for the backbone atoms, 1.31 Å for the heavy atoms in the dimerization interface, and 2.29 Å for all heavy atoms. Figure 10 shows tube diagrams of the backbones of the five refined structures along with the crystal structure, which is shown in magenta. The

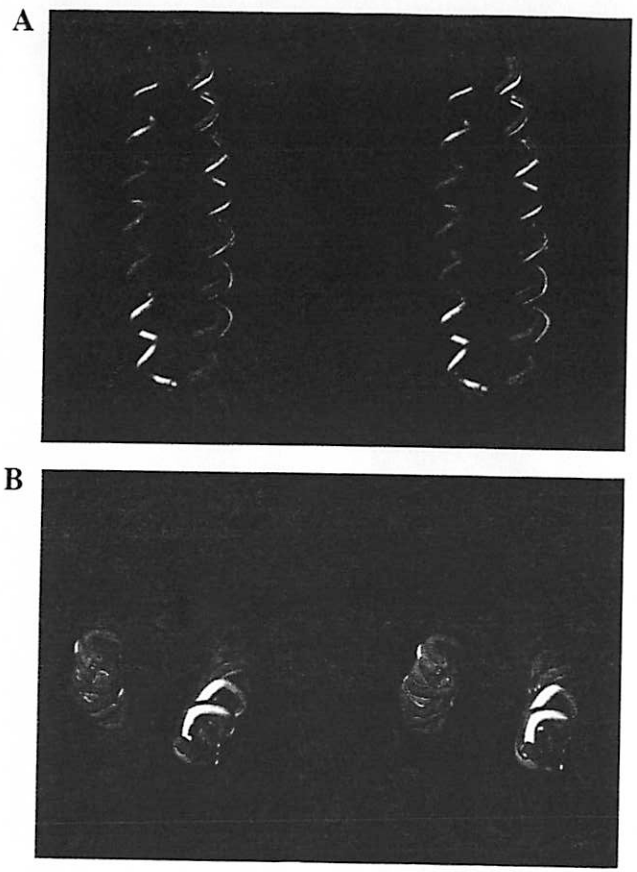


Fig. 10. Side and top views of the tube diagrams of five refined, predicted structures of the GCN4 leucine zipper, shown in red, yellow, green, cyan and white, along with the crystal structure, shown in magenta. Basically, there is one fused, six-color tube.

predicted positions of the side chains in the dimerization interface are essentially unique, but much greater variation is found in the positions of the surface residues.

These simulations also suggested a possible mechanism of the GCN4 leucine zipper coiled-coil assembly. Folding commences from the collision of two short helical stretches, generally located at the ends of the chain. These interacting helical stretches

then propagate along the molecule. After small adjustments in registration by an inch worm type mechanism, the final, parallel in register coiled-coil dimer forms. Among the last regions to lock into place are the Asn^{16} residues which are located in the dimerization interface and which are regions of predicted low intrinsic stability. Although a detailed study has not yet been performed, there is some indication that assembly from the N-termini is preferred.

Method for predicting the state of association of proteins

A limitation of the above calculation is the assumption that the oligomerization state has no higher order dimers. However, coiled coils can associate to tetramers or even higher order aggregates [37,113,114]. Due to computer time limitations, the straightforward simulation of multimer equilibria is far beyond contemporary computer resources. Furthermore, if MMC is used, then the folding process must be repeated tens, if not hundreds, of times to be statistically significant. Thus, to predict the state of association, we developed a methodology that estimates the equilibrium constants among a spectrum of assumed parallel and antiparallel oligomers [22,115]. Subsequently, a more refined, lattice-based method was developed that also permits the monomeric state to be included [31]. Since the second method gives essentially the same results as the original technique in the regime where the two approaches overlap, we summarize the results from this more general approach.

In order to calculate the equilibrium constant [116], the internal partition function, Z_{int}, is required. For the denatured state, to estimate Z_{int}, we developed a transfer matrix treatment that includes all interactions within five-residue fragments [31,76]. The disadvantage of this approach is that it ignores longer range interactions. Dimers and higher order multimers are treated somewhat differently. First, we note that, in general, P(E), the probability of being in an energy level E, is related to Z_{int} by

$$Z_{int} = N(E)\exp(-\beta E)/P(E) \tag{8}$$

Here, N(E) is the degeneracy of energy level E and $\beta = 1/kT$. Since E and P(E) are readily obtained from an MMC simulation, the remaining problem is to determine N(E). It may be estimated to within a constant by ESMC [33,34]; however, here we use a quasianalytic method. The basic idea is to focus on the most probable energy state. The Monte Carlo simulation provides the set of three consecutive C^{α} virtual bonds that are sampled by the ensemble of structures having the most probable energy, $\bar{E}$. $N(\bar{E})$ is obtained as the transfer matrix product of all such sets of three-bond vectors that are appropriately combined.

Initially, this approach was applied to calculate the monomer–dimer equilibrium in the GCN4 leucine zipper [108] and its fragments [117]. The results of the calculations and the comparison with experiment for the predicted dominant species are shown in Table 3. For the three cases for which experimental data are available, the prediction agrees with experiment. Examination of the stability of the GCN4 fragments indicates

Table 3 *Comparison with experiment of the predicted dominant species for the GCN4 leucine zipper*

Protein	Predicted dominant species	Experimental dominant species
GCN4 wild-type	2	2
GCN4 8–30	2	2
GCN4 11–33	1	1
GCN 4–26	2	Not yet measured

that the stability of a given fragment cannot be estimated from the stability of the parent molecule. A similar situation is obtained for other coiled coils such as tropomyosin. The origin of the lack of stability of the 11–33 fragment arises from the difficulty in burying Asn^{16} in a helical conformation in the core. In the 11–33 fragment, there would be a single helical turn at the N-terminus before the Asn. Since it is not stable enough to force the hydrophilic Asn to be helical, the entire fragment becomes disordered. This effect is due to loop entropy, which in coiled coils acts to prohibit random coiled conformations between interacting helical stretches [118]. In this and in all the systems studied, the simulations suggest that coiled coils are highly cooperative with many of the observed phenomena caused by nonadditive effects.

Next, we examined the stability of Fos and Jun coiled coils. In partial agreement with experiment, Fos without a GCG linker is predicted to be monomeric, whereas at high concentration both monomer and dimers are present [38]. In an equimolar mixture of cross-linked Fos and Jun homodimers, as in the experimental system, the simulation predicts that Fos heterodimers should preferentially form. The calculations suggest that the presence of Thr and Lys in the interfacial region of Fos homodimers gives rise to the relative instability of Fos homodimers. When an equimolar system of Fos and Jun are present, the system can lower its overall free energy by forming Fos–Jun heterodimers.

Coiled coils can also provide insights into the factors driving the formation of quaternary structure. In an elegant study, Harbury et al. [37] simultaneously replaced all four a and d residues of the GCN4 leucine zipper by Leu, Val, and Ile. In Table 4, the theory is compared with experiment. In five of eight cases, the simulations and experiment are in agreement over the entire concentration range, and, in another case, agreement is found over a portion of the experimental range. These calculations suggest that intrinsic secondary structural preferences and configurational entropy favor lower order species, while quaternary interactions favor higher order species. This conjectured origin of multimer stability is inconsistent with the suggestion of Harbury et al. [37] that the selection of a given species is due to the requirement that the lowest energy side-chain rotamer selects the particular interchain packing geometry. Such a level of detail is beyond the lattice models where the side chains are represented by soft core balls.

Table 4 *Comparison of the predicted and experimentally measured dominant species of GCN4 and seven mutants*

Residues at positions a	d	Dominant species from experiment	Dominant species from simulation
GCN4 wild-type		2	2
I	L	2	3
I	I	3	3
L	I	4	4
V	I	?	3
L	V	3	3
V	L	2,3	2
L	L	3	3

Weaknesses of the lattice models

While this chapter has presented a number of examples where folding of a protein from sequence alone has been achieved, the full solution of the protein folding problem is not in hand. There still remain problems with the potential. While the current generation can differentiate grossly incorrect folds from native, in many cases it is very difficult to differentiate topologies having substantial similarity to the native fold. These close topological cousins have essentially the same burial energy as native and differ by a relatively small number of side-chain contacts and differences in secondary structure. To some extent, this is a physical effect. Even if two topologies differ on average by 10kT in energy but have comparable configurational entropy, then the lower energy fold will be thermodynamically very favored [22,115]. However, the accurate portrayal of this difference is nontrivial. Furthermore, due to the representation of side chains as soft balls with a single interaction center and a relatively wide interaction basin, the high coordination lattice models overestimate the protein's entropy. Such an excess entropy also results in an increase in the backbone's flexibility. This is probably a major cause of the difficulty the models have with the folding of naturally occurring β-proteins. Part of this effect is inevitable in any reduced protein model. Clearly, improved side-chain representations are necessary. Another problem concerns adequate conformational sampling. Given that close topologies have overlapping energy spectra, a sufficient number of simulations must be done to ensure that the predicted low-energy structure is well characterized. At present, this is possible only for simple folds lacking reversals in chain direction. ESMC may be helpful in this regard, but such calculations can also be very expensive [25,33–35]. Thus, the model representation, potential and sampling protocols all require improvement.

Conclusions

In this chapter we have described the folding of protein A and seven homologous sequences, the putative retrosequence of protein A, two sequences designed by DeGrado, a putative ROP monomer, crambin, the V-3 loop of gp-120, and the GCN4 leucine zipper. For many of these sequences, structures which are in reasonable agreement with experiment have been predicted, and the remainder stand as predictions to be tested by experiment. Furthermore, assuming that the native state is located in a collection of parallel and antiparallel dimers, trimers and tetramers, the quaternary structure of GCN4, two of its fragments, five of eight wild-type mutants, Fos, Jun, and Fos–Jun heterodimers have been successfully predicted. In addition, the 4–26 fragment of GCN4 is predicted to be dimeric. The simulations argue that the native structure is a compromise among numerous contributions to the potential. The different terms such as hydrophobic interactions, hydrogen bonding and cooperative side-chain packing interactions give rise to different aspects of protein-like behavior. The results to date suggest that progress is being made in the *de novo* prediction of protein structure. Future advances are likely to result from better, more specific energy functions, better model realizations, and combined approaches such as the use of inverse folding to predict the structure of fragments followed by their assembly using reduced models such as have been discussed here. Overall, the prospect for future progress in the protein folding problem remains bright.

Acknowledgements

This work was supported by NIH Grant Nos. GM-37408, GM-38794, TW-00418, and by the University of Warsaw, BST-532-34/96. A.K. is an International Research Scholar of the Howard Hughes Medical Institute (Grant #75195-543402). The contributions of Drs. C.L. Brooks III, A. Godzik, M. Milik, M. Vieth, and K. Olszewski to the work described here are gratefully acknowledged.

References

1. Jernigan, R.L., Curr. Opin. Struct. Biol., 2(1992)248.
2. Anfinsen, C.B., Science, 181(1973)223.
3. Ripoll, D.R., Piela, L., Velasquez, M. and Scheraga, H.A., Proteins, 10(1991)188.
4. Skolnick, J. and Kolinski, A., Science, 250(1990)1121.
5. Go, N. and Taketomi, H., Proc. Natl. Acad. Sci. USA, 75(1978)559.
6. Go, N., Abe, H., Mizuno, H. and Taketomi, H., Protein Folding, Elsevier, Amsterdam, 1980, p. 167.
7. Kolinski, A., Skolnick, J. and Yaris, R., Biopolymers, 26(1987)937.
8. Skolnick, J. and Kolinski, A., Annu. Rev. Phys. Chem., 40(1989)207.
9. Godzik, A., Kolinski, A. and Skolnick, J., J. Comput.-Aided Mol. Design, 7(1993)397.
10. Bryngelson, J.D., Onuchic, J.N., Socci, N.D. and Wolynes, P.G., Proteins, 21(1995)167.

11. Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D. and Chan, H.S., Protein Sci., 4(1995)561.
12. Sali, A., Shakhnovich, E. and Karplus, M., J. Mol. Biol., 235(1994)1614.
13. Sali, A., Shakhnovich, E. and Karplus, M., Nature, 369(1994)248.
14. Kolinski, A., Milik, M. and Skolnick, J., J. Chem. Phys., 94(1991)3978.
15. Shakhnovich, E.I. and Gutin, A.M., Proc. Natl. Acad. Sci. USA, 90(1993)7195.
16. Socci, N.D. and Onuchic, J.N., J. Chem. Phys., 100(1994)1519.
17. Kolinski, A. and Skolnick, J., J. Phys. Chem., 97(1992)9412.
18. Kolinski, A., Godzik, A. and Skolnick, J., J. Chem. Phys., 98(1993)7420.
19. Kolinski, A. and Skolnick, J., Proteins, 18(1994)338.
20. Kolinski, A. and Skolnick, J., Proteins, 18(1994)353.
21. Vieth, M., Kolinski, A., Brooks III, C.L. and Skolnick, J., J. Mol. Biol., 237(1994)361.
22. Vieth, M., Kolinski, A., Brooks III, C.L. and Skolnick, J., J. Mol. Biol., 251(1995)448.
23. Kolinski, A., Galazka, W. and Skolnick, J., J. Chem. Phys., 103(1995)10286.
24. Kolinski, A., Milik, M., Rycombel, J. and Skolnick, J., J. Chem. Phys., 103(1995)4312.
25. Kolinski, A., Galazka, W. and Skolnick, J., Proteins (1997) in press.
26. Skolnick, J., Kolinski, A. and Ortiz, A.R., J. Mol. Biol., 265(1997)217.
27. Kolinski, A., Skolnick, J., Godzik, A. and Hu, N.P., Proteins, 27(1997)290.
28. Milik, M., Kolinski, A. and Skolnick, J., J. Comput. Chem., 18(1997)80.
29. Olszewski, K.A., Kolinski, A. and Skolnick, J., Proteins, 25(1996)286.
30. Olszewski, K.A., Kolinski, A. and Skolnick, J., Protein Eng., 9(1996)5.
31. Vieth, M., Kolinski, A. and Skolnick, J., Biochemistry, 35(1996)955.
32. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E., J. Chem. Phys., 51(1953)1087.
33. Hao, M.-H. and Scheraga, H.A., J. Phys. Chem., 98(1994)4940.
34. Hao, M.-H. and Scheraga, H.A., J. Phys. Chem., 98(1994)9882.
35. Hao, M.-H. and Scheraga, H.A., J. Chem. Phys., 102(1995)1334.
36. Skolnick, J., Kolinski, A., Brooks III, C.L., Godzik, A. and Rey, A., Curr. Biol., 3(1993)414.
37. Harbury, P.B., Zhang, T., Kim, P.S. and Alber, T., Science, 262(1993)1401.
38. O'Shea, E.K., Rutkowski, R., Stafford III, W.F. and Kim, P.S., Science, 245(1989)646.
39. Godzik, A., Kolinski, A. and Skolnick, J., J. Comput. Chem., 14(1993)1194.
40. Rey, A. and Skolnick, J., J. Comput. Chem., 13(1992)443.
41. Rey, A. and Skolnick, J., Proteins, 16(1993)8.
42. Srinivasan, R. and Rose, G.D., Proteins, 22(1995)81.
43. Dill, K.A. and Yue, K., Protein Sci., 5(1996)254.
44. Kolinski, A., Skolnick, J. and Yaris, R., J. Chem. Phys., 85(1986)3585.
45. Kolinski, A., Skolnick, J. and Yaris, R., Macromolecules, 20(1987)438.
46. Kolinski, A. and Skolnick, J., Proc. Natl. Acad. Sci. USA, 83(1986)7267.
47. Skolnick, J. and Kolinski, A., J. Mol. Biol., 212(1990)787.
48. Skolnick, J., Kolinski, A. and Yaris, R., Proc. Natl. Acad. Sci. USA, 86(1989)1229.
49. Skolnick, J., Kolinski, A. and Yaris, R., Biopolymers, 28(1989)1059.
50. Skolnick, J., Kolinski, A. and Yaris, R., Proc. Natl. Acad. Sci. USA, 85(1988)5057.
51. Kabsch, W. and Sander, C., Biopolymers, 22(1983)2577.
52. Levitt, M. and Greer, J., J. Mol. Biol., 114(1977)181.
53. Olszewski, K.A., Kolinski, A. and Skolnick, J., Protein Eng., 9(1996)5.
54. Gregoret, L.M. and Cohen, F.E., J. Mol. Biol., 219(1991)109.

55. Hunt, G.N., Gregoret, L.M. and Cohen, F.E., J. Mol. Biol., 241(1994)214.
56. Hao, M.-H., Rackovsky, S., Liwo, A., Pincus, M.R. and Scheraga, H.A., Proc. Natl. Acad. Sci. USA, 89(1992)6614.
57. Nikishawa, K. and Ooi, T., Biochemistry, 100(1986)1043.
58. Miyazawa, S. and Jernigan, R.L., Macromolecules, 18(1985)534.
59. Godzik, A., Kolinski, A. and Skolnick, J., Protein Sci., 4(1995)2107.
60. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112(1977)535.
61. Skolnick, J., Jaroszewski, L., Kolinski, A. and Godzik, A., Protein Sci., 6(1997)676.
62. Milik, M., Kolinski, A. and Skolnick, J., Protein Eng., 8(1995)225.
63. Kuwajima, K., Proteins, 6(1989)87.
64. Ptitsyn, O.B., Pain, R.H., Semisotnov, G.V., Zerovnik, E. and Razgulyaev, O.I., FEBS Lett., 262(1990)20.
65. Brooks, C.L., Curr. Opin. Struct. Biol., 3(1993)92.
66. Skolnick, J., Kolinski, A. and Godzik, A., Proc. Natl. Acad. Sci. USA, 90(1993)2099.
67. Eloffson, A. and Nilsson, L., J. Mol. Biol., 223(1993)766.
68. Barker, J.A. and Henderson, D., Rev. Mod. Phys., 48(1976)587.
69. Lee, J., Phys. Rev. Lett., 71(1993)211.
70. Berg, B.A. and Neuhaus, T., Phys. Rev. Lett., 68(1991)9.
71. Handel, T. and DeGrado, W.F., Biophys. J., 61(1992)A265.
72. Raleigh, D.P. and DeGrado, W.F., J. Am. Chem. Soc., 114(1992)10079.
73. Handel, T.M., Williams, S.A. and DeGrado, W.F., Science, 261(1993)879.
74. Woolfson, D.N. (1996) personal communication.
75. Ptitsyn, O.B., J. Protein Chem., 6(1987)273.
76. Skolnick, J. and Kolinski, A., J. Mol. Biol., 221(1991)499.
77. Kuwajima, K., Mitani, M. and Sugai, S., J. Mol. Biol., 206(1989)547.
78. Gouda, H., Torigoe, H., Saito, A., Sato, M., Arata, Y. and Schimada, I., Biochemistry, 40(1992)9665.
79. Wright, P.E. (1996) personal communication.
80. Nilsson, B., Moks, T., Jansson, B., Abrahamsen, L., Elmblad, A., Holmgren, E., Henrichson, C. and Jones, T.A., Protein Eng., 1(1987)107.
81. Lyons, B.A., Tashiro, M., Cedergren, L. and Montelione, G.T., Biochemistry, 32(1993)7839.
82. Bowie, J.U., Reidhaar, O.J.F., Lim, W.A. and Sauer, R.T., Science, 247(1990)1306.
83. Rose, G.D. and Wolfenden, R., Annu. Rev. Biophys. Biomol. Struct., 22(1993)381.
84. Dyson, J.H. and Wright, P.E., Curr. Biol., 3(1993)60.
85. Chou, K.C., Maggiora, G. and Scheraga, H.A., Proc. Natl. Acad. Sci. USA, 89(1992)7315.
86. Behe, M.J., Lattman, E.E. and Rose, G.D., Proc. Natl. Acad. Sci. USA, 88(1991)4195.
87. Brunet, A.P., Huang, E.S., Huffine, M.E., Loeb, J.E., Weltman, R.J. and Hecht, M.H., Nature, 364(1993)355.
88. Predki, P.F. and Regan, L.R., Biochemistry, 34(1995)9834.
89. Wilmot, C.M. and Thornton, J.M., J. Mol. Biol., 203(1988)221.
90. Goodman, M. and Chorev, M., Acc. Chem. Res., 12(1979)1.
91. Guptasarma, P., FEBS Lett., 310(1992)205.
92. Chou, K.C., Proteins, 21(1995)319.
93. Presta, L.G. and Rose, G.D., Science, 240(1988)1632.
94. Richardson, J.S. and Richardson, D.C., Science, 240(1988)1648.

95. Wong, C.H. (1996) personal communication.
96. Banner, D.W., Kokkinidis, M. and Tsernoglou, D., J. Mol. Biol., 196(1987)657.
97. Sander, C. (1993) personal communication.
98. Eliezer, D., Jennings, P.A., Wright, P.E., Doniach, S., Hodgson, K.O. and Tsuruta, H., Science, 270(1995)487.
99. Hendrickson, W.A. and Teeter, M.M., Nature, 290(1981)107.
100. Holley, L.H. and Karplus, M., Proc. Natl. Acad. Sci. USA, 86(1989)152.
101. Garnier, J., Osguthorpe, D.J. and Robson, B., J. Mol. Biol., 120(1978)97.
102. LaRosa, G.J., Davide, J.P., Weinhold, K., Waterbury, J.A., Profy, A.T., Lewis, J.A., Langlois, A.J., Dreesman, G.R., Boswell, R.N., Shadduck, P., Holley, L.M., Karplus, M., Bolognesi, D.P., Matthews, T.J., Emini, E.A. and Putney, S.D., Science, 249(1990)932.
103. Rost, B. and Sander, C., J. Mol. Biol., 232(1993)584.
104. Havel, T.F., Prog. Biophys. Mol. Biol., 56(1991)43.
105. Smith-Brown, M.J., Kominos, D. and Levy, R.M., Protein Eng., 6(1993)605.
106. Aszodi, A., Gradwell, M.J. and Taylor, W.R., J. Mol. Biol., 248(1995)308.
107. McLachlan, A.D. and Stewart, M., J. Mol. Biol., 98(1975)298.
108. O'Shea, E.K., Klemm, J.D., Kim, P.S. and Alber, T., Science, 254(1991)539.
109. Ellenberger, T.E., Brandl, C.J., Struhl, K. and Harrison, S.C., Cell, 71(1992)38.
110. Nilges, M. and Brunger, A.T., Proteins, 15(1993)133.
111. Nilges, M. and Brunger, A.T., Protein Eng., 4(1991)649.
112. Brooks, B.R., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4(1983)187.
113. Harbury, P.B., Kim, P.S. and Alber, T., Nature, 371(1994)80.
114. Chmielewski, J., J. Am. Chem. Soc., 116(1994)6451.
115. Vieth, M., Kolinski, A. and Skolnick, J., J. Chem. Phys., 102(1995)6189.
116. McQuarrie, A.D., Statistical Mechanics, Harper & Row, New York, NY, 1976.
117. Lumb, K.J., Carr, C.M. and Kim, P.S., Biochemistry, 33(1994)7361.
118. Skolnick, J., Macromolecules, 17(1984)645.