Monte Carlo Simulations of Protein Folding. I. Lattice Model and Interaction Scheme

Andrzej Kolinski[1,2] and Jeffrey Skolnick[1]
[1]*Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037 and* [2]*Department of Chemistry, University of Warsaw, 02-093 Warsaw, Poland*

ABSTRACT **A new hierarchical method for the simulation of the protein folding process and the de novo prediction of protein three-dimensional structure is proposed. The reduced representation of the protein α-carbon backbone employs lattice discretizations of increasing geometrical resolution and a single ball representation of side chain rotamers. In particular, coarser and finer lattice backbone descriptions are used. The coarser (finer) lattice represents Cα traces of native proteins with an accuracy of 1.0 (0.7) Å rms. Folding is simulated by means of very fast Monte Carlo lattice dynamics. The potential of mean force, predominantly of statistical origin, contains several novel terms that facilitate the cooperative assembly of secondary structure elements and the cooperative packing of the side chains. Particular contributions to the interaction scheme are discussed in detail. In the accompanying paper (Kolinski, A., Skolnick, J. Monte Carlo simulation of protein folding. II. Application to protein A, ROP, and crambin. Proteins 18:353–366, 1994), the method is applied to three small globular proteins.** © 1994 Wiley-Liss, Inc.

Key words: **tertiary structure prediction, reduced protein model, lattice protein models, dynamic Monte Carlo simulations, potentials of mean force**

INTRODUCTION

The inability to predict the three-dimensional structure of globular proteins from protein sequence is one of the most important unsolved problems of contemporary theoretical molecular biology.[2–4] There have been various approaches to solve this problem. Probably the most successful to date are methods based on sequence and structure homology matching[5–14] to proteins for which the three-dimensional structures have been previously solved. These methods, when combined with molecular modeling, can predict the three-dimensional structures of some proteins or at least plausible structures of protein fragments. However, there is an obvious limitation of a homology-based approach to the protein folding problem. Namely, the number of known sequences is growing many times faster than the number of known three-dimensional structures. It is unclear what fraction of these sequences will have folds with representatives in the library of known three-dimensional structures. Some theoretical estimates suggest that the number of types of protein folds may be much larger than the number of distinct topologies seen in the present structural data base.[15] Thus, while very powerful in principle, homology modeling will always be restricted to comparing new sequences to extant topologies.

More straightforward approaches to the protein folding problem employ various computer simulation methods. Here, the ultimate aim is to build a computer model which, starting from sequence alone, generates the three dimensional structure of a given protein. These methods, which simulate the time evolution of the modeled system, could also provide information concerning the folding pathway(s). Consequently, the properties of the folding intermediates could be also studied. Some other methods, employing various types of energy minimization[16,17] and conformational search protocols[4] could also address the nature of folding intermediates. While the computer simulation methods are in principle more general than homology-based methods, the practical realization of a successful folding algorithm has proven to be extremely difficult.

Various computer models of protein folding employ very different levels of molecular detail.[3,4,18,19] Two extreme approaches may be considered. On one side, there are full atom models whose time evolution is simulated using molecular dynamics (MD) with a very detailed force field.[20,21] Rigorous implementations of MD studies of globular proteins usually require the explicit simulation of a large number of solvent molecules.[22,23] These detailed MD simulations are extremely expensive, and the longest simulations correspond to the real times in the range of nanoseconds. This is to be compared with

Received April 19, 1993; revision accepted December 20, 1993.

Address reprint requests to Dr. Jeffrey Skolnick, Department of Molecular Biology, The Scripps Research Institute, 10666 N. Torrey Pines Rd., La Jolla, CA 92037.

the time scale of real protein folding, which is of the order of milliseconds to seconds. Consequently, taking into consideration the present state of computing art, this approach can be used only for simulating rather fast local rearrangements. Simulations of the entire folding pathway would perhaps be even more expensive than the above comparison suggests due to the larger volume of the denatured protein in solution. There is one additional question sometimes addressed in the context of the more general validation of the MD methodology. It is not obvious that potentials currently used for MD simulations are capable of distinguishing (based on an energy comparison) between topologically different folds of the same sequence.[24] Nevertheless, MD simulations have proven extremely useful in studies of various local and/or small distance rearrangements of protein structures.[20,21,25] For example, the relaxation of the structures of site-specific mutants with respect to the wild type structures has been successfully investigated. Another type of MD study involves the small distance relaxation of entire globules,[20,22,26] as for example, the hinge bending motion of human lysozyme.[22] Recently, it has also become possible to predict the three-dimensional structure of apoproteins based on relaxation of the known structure containing the prosthetic group.[23] These are just some examples of applications based on MD, the most mature technique for the study of biomolecular systems. Detailed force field Monte Carlo simulations of the biomolecular systems have been much less popular.

Recognizing the time scale limitations inherent in detailed molecular dynamics simulations of protein dynamics, various reduced representations of the conformations of proteins were proposed.[3,18,19] Applications of reduced simplified models to the study of protein stability, dynamics, and folding rely on the assumption that the fundamental properties of proteins are rather robust, and independent of many atomic details. To what extent this assumption is valid should also be addressed in studies of simplified models. Typical reduced representation models ignore most of the atomic details. Usually a single amino acid fragment of the main chain backbone is represented as a single united atom. Similar simplifications are frequently invoked to account for the side groups. This class of models can explore time intervals much larger than the longest relaxation time of such model polypeptide chains. Various types of semiempirical potentials,[3,18,19,27,28] usually derived from the statistical properties of the known three-dimensional structures of the globular proteins,[29,30] have been used in the context of reduced models. Due to the very flat free energy surface of such models (which lack sufficient interaction specificity), the predicted three-dimensional structures of even very small proteins were of rather low accuracy.[3,18] In most applications, they are just on the border of being randomly packed, dense collapsed structures. In all cases, the pattern of side chain packing hardly resembled the specific arrangements seen in real native proteins. When a target potential was employed in these simplified models, the exploration of the effect of various forces on the stability of the globular state and cooperativity of the folding process became possible.[18] Related lattice models of the protein conformational space, which employed Monte Carlo dynamics (MCD) to simulate protein motion, generated the folding pathways of quite large globular proteins.[31-33] Using local conformational propensities consistent with the native structure, these model proteins very quickly folded to the proper unique "native state," thereby providing an example of how real proteins might "beat" the Levinthal paradox.[3,34] However, these simplified models, while quite helpful in understanding various aspects of protein folding dynamics and thermodynamics, also failed to predict three-dimensional structures of good accuracy when sequence information alone was used.

In this series of papers, we present a novel hierarchical approach to the protein folding problem. At least for some small proteins, it can predict their three-dimensional structure with an accuracy in the range of 2–4 Å rms (coordinate root mean square deviation after the best superposition) from the known (or expected) native state, using amino acid sequence as the only protein-specific input information. The hierarchy of simulation modules is as follows: First, a coarser, but rather flexible, lattice model is used to fold the protein of interest to a family of three-dimensional structures. The precision of the model is in the range of 3–4 Å rms for the $C\alpha$ trace, and is about 4–5 Å rms for all the heavy atoms. In these structures, the reproducibility of the packing pattern of the side chains is low, but it is still much higher than was observed in other simulations of simplified models. Then, the obtained folds are subject to refinement by a more precise, but still discrete, lattice model. This finer lattice representation produces a well-defined pattern of hydrogen bonds and protein-like side chain packing. The resulting finer lattice conformations provide a set of secondary structure and tertiary contact constraints which can be used in target MD folding, employing an all atom representation and a detailed force field. Using this hierarchical protocol, several simple proteins have been successfully folded with rather high reproducibility, and good accuracy of the obtained folds. In the accompanying paper,[1] we describe three examples: the B domain of staphylococcal protein A, a designed monomeric, 120 residue, version of *Escherichia coli* ROP dimer, and the 46 residue protein crambin (lcrn). The present paper provides a detailed description and discussion of the reduced model and simulation method. The force field incorporated into the model consists of several terms.

Some of these terms are similar to various knowledge-based potentials used in other studies of model polypeptides. However, since a large part of the interaction scheme is rather novel, it requires more detailed analysis; this is especially true since the various contributions to the secondary and tertiary interactions have to be appropriately balanced.

The key part of the proposed protocol for tertiary structure prediction is the discrete, reduced representation model of protein conformation and dynamics. What, then, are the major differences between the present approach and other applications of reduced models to the protein folding problem? First, our model of protein conformations uses high coordination lattice representations of the polypeptide chains.[27,28,35] The coarser (finer) lattice has an underlying grid spacing equal to 1.70 (1.22) Å. Due to the "protein-like" geometry of the lattice Cα backbone, the accuracy of the finer lattice side chain representation is about two times better than the coarser lattice description. These high coordination lattices reproduce, with a high degree of fidelity, various angular correlations seen in real proteins. For example, the virtual bond angle between two consecutive Cα–Cα vectors, and the angles between the main chain Cα backbone and the vectors defining the centers of mass of the attached side groups can be reproduced with an average error smaller than 10°. This level of accuracy is probably necessary to make any meaningful simulations of real proteins. At this level of resolution, various collective effects related to the fine details of protein packing start to become manifest. Furthermore, the discretized model is about two orders of magnitude faster in computer realization when compared to equivalent off-lattice models in the framework of a similar Metropolis scheme of the Monte Carlo dynamics. This is because the lattice model enables the "in front," prefabricated computation of various geometric transformations (elemental moves of the dynamic scheme) as well as numerous contributions to the energy.

A second distinguishing feature of the present approach is that the model force field contains several novel elements designed to mimic, as closely as possible, various interactions in globular proteins. Some terms of the potentials are sequence independent and are designed to keep the system in the portion of conformational space that is "protein-like." For example, the model system most frequently samples the valleys in an "averaged" Ramachandran map. This is achieved by the introduction of an energetic bias that regularizes the main chain backbone by enforcing the proper distribution of the short-range Cα–Cα distances and chirality. Then, a highly cooperative model of the hydrogen bond network drives the system to protein-like secondary structure. This could be either helical, β-sheet,[27,28] or a mixture of regular fragments with less regular ones. The important feature is that large, completely irregular globules are very unlikely, even in the absence of any amino acid specific interactions. The amino acid specific part of the conformational energy consists of several potentials of mean force describing the short- and long-range interactions. These potentials are derived from a statistical analysis of a database of high resolution, three-dimensional structures. They contain at least two terms that were absent in other studies. These terms are crucial for the predictive strength of the model. The first is an amino acid pair specific, mean force potential describing the angular correlation between side groups down the chain. This short-range potential triggers the formation of a particular type of secondary structure, when permitted by other interactions. The second is a multibody potential which reflects the specific, regular packing of the side groups. Thus, not only is secondary structure formation cooperative, but the transition from a globule with a loosely defined hydrophobic core to the globular native-like state possessing well organized packing pattern is also quite cooperative.

In what follows, we will try to provide a detailed physical justification for the various potentials used in this work, although the discretized model could be also viewed as a system whose physical meaning is justified by the ex post facto correct mapping of amino acid sequences to their respective three dimensional structures. In fact, recent simulations[28] of two proteins designed by DeGrado and co-workers[36,37] provide strong evidence for the validity of these potentials, at least for simple folds of globular proteins. The predicted folds, and the striking differences between the nature of the compact states (molten globule versus native-like) of these proteins, are in agreement with all known experimental facts.[36–38] Since these proteins have not been included in the database used for constructing the statistical potentials, the possibility of specific "target" biases in the folding simulations can be safely eliminated. Of course, there is always the possibility that the force field is biased toward a particular class of globular proteins, as characterized by size, content of regular secondary structure, etc. Due to the intrinsic character of reduced models and model potentials, the possibility of this kind of bias can only be examined by computer folding experiments on a variety of proteins.

The remainder of this paper is organized as follows: First, we describe the discretized representation of the protein main chain backbone on the coarser and finer lattices and the representation of the side groups in the framework of these models. Second, the Monte Carlo dynamics scheme is presented. The description of the geometric representation is followed by a detailed discussion of the interaction scheme and its implementation within the Metropolis sampling procedure. We conclude with a discussion of possible future refinements of the proposed method. The results of the folding of two helical proteins, and the folding of crambin, a small α/β protein, with a rather unique topology, are described in the accompanying paper.[1]

LATTICE MODELS AND MONTE CARLO DYNAMICS

The lattice model in the coarser representation is very similar to that used previously.[27,28,39] For both discretizations, the interaction scheme has been updated and refined.

Our reduced models use an α-carbon, lattice representation of the main chain backbone, i.e., every single amino acid segment of the main chain is treated as an united atom. The Cα trace serves as a reference frame for the definition of the side chain positions, and the orientation of the hydrogen bonds. The side groups are also treated as single united atoms. The location of the center of the side group depends on the amino acid identity, the local conformation of the Cα trace, and the actual rotamer of the side chain. As indicated previously, two lattice models are used in the folding algorithm. The first model has a coarser underlying grid. This model is employed when simulating the folding from an expanded, random coil state. The second, finer lattice model is used for the simulation of the later stages of folding. In principle, the finer lattice could be (and in a limited number of cases has been) used to simulate the entire folding process; the hierarchical approach was elected for the practical reason of making the entire simulation less CPU intensive. The descriptions of the geometric properties and Monte Carlo dynamics scheme are given separately for both models. Since the interaction scheme is essentially the same for both lattice representations, it is presented in a separate section.

Coarser Lattice Model

The set of basis vectors consists of all cyclic permutations of the x, y, and z (including sign permutations) coordinates of the following vectors: (2,1,1), (2,1,0), and (1,1,1). There are a total of 56 such vectors. Suppose that a lattice path is fit to the set of Cα Cartesian coordinates of a real protein. In order to obtain a good overall rms deviation from native as well as a good local angular correspondence to the real chain, some restrictions on the basis vectors are necessary. Namely, the valence angle for the model Cα trace is restricted to the range (78.5°, 143.1°). The boundaries were selected to cover the distribution seen in real proteins. The best fits to three-dimensional structures from the protein database are obtained when the spacing of the underlying cubic lattice (1,0,0) equals 1.70 Å.

This lattice is quite flexible. Large helical motifs can be represented with an accuracy of 0.7 Å rms, β-sheet motifs with an accuracy of 0.6 Å, and the average rms for the entire database is slightly below 1.0 Å rms for the Cαs. Moreover, the estimated angular error in the definition of the Cβ direction (see ref. 35 for more details) is in the range of 15–25°. These rms deviations and angular distortions are much smaller than in other simplified lattice models of proteins.[35]

The excluded volume of the model chain backbone is slightly exaggerated. The distance of the closest approach for a pair of nonbonded α-carbons is equal to 4.78 Å [the length of a lattice vector of the type (2,2,0)].

For each amino acid, a library of side chain rotamers was built within the framework of a single sphere representation. The number of model rotamers depends on the amino acid identity, the actual conformation of the main chain backbone, defined by two consecutive vectors of the Cα trace, and the assumed resolution for the side group representation. For the N-terminal (as well as for C-terminal) amino acid, the definition of the side group orientation is provided by a dummy backbone segment, which may also be treated as an N-terminal (C-terminal) cap of the polypeptide. The centers of interaction for the side groups have off-lattice coordinates, except for glycine whose center of interaction is located at the Cα position. The resolution of the model for the side groups equals 1.7 Å. Each side group has a strongly repulsive, square well core and a weaker square well interaction sphere. The cut-off distances for these envelopes are amino acid pair specific, reflecting the possibility of different packing of a given side group with various other side groups. Figure 1 schematically shows a short fragment of the model chain. The average side group diameters are drawn for the sake of clarity.

The dynamics of the model system is simulated by a stochastic process of small, random micro modifications of the chain conformation. The process is controlled by the asymmetric Metropolis scheme.[40] Monte Carlo dynamics (MCD) is the natural choice for discretized models and is to a large extent equivalent to an off lattice, Brownian dynamics simulation with a relatively long time step and a large random force. This means that the obtained trajectories are numerical solutions of a stochastic equation of motion, provided that the set of elemental moves spans the entire space of possible conformational transitions and that the probabilities for the elemental moves satisfy detailed balance. MCD has a physical meaning for those dynamic properties whose characteristic time scales are considerably larger than the time scale of the elemental micro modifications implemented in the algorithm. Lattice MCD has proven to be a very efficient method of studying long time dynamics of polymer systems.[41–44]

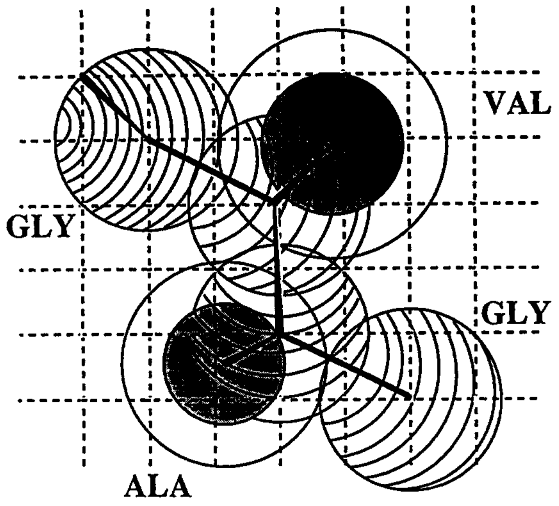

Fig. 1. Schematic drawing of a short fragment of the model polypeptide chain on the coarse hybrid lattice. The spacing of underlying cubic lattice grid is equal to 1.7 Å. The spheres centered on the vertices of the main chain correspond to the main chain portion of the excluded volume. The side chains have repulsive cores (shaded spheres) and square well attractive or weakly repulsive regions (open spheres). The radii shown in the figure are approximate, since the cut-off distances are amino acid pair specific.

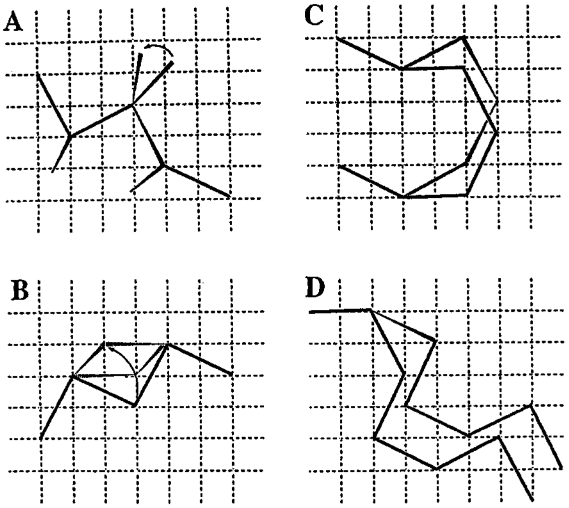

Fig. 2. Schematic representation of elemental moves employed in the MCD scheme on the coarser lattice. (**A**) An example of motion of the center of the model side group, simulating an internal isomeric transition for a flexible side chain. (**B**) Two bond, spike moves; for clarity, the rotamer displacements are not shown. (**C**) Four bond move. (**D**) Small distance, rigid body displacement of a large fragment of the model chain, starting from the rotation of a randomly selected single bond up to the chain terminus.

MCD on the coarser lattice employs the following set of elemental moves:

A. A random modification of the rotamer representation of the randomly selected amino acid. The Cα trace remains unaffected (Fig. 2A).

B. A virtual two bond spike move (Fig. 2B), subject to the bond angle restrictions mentioned above. The end segments are treated separately, and a new orientation of two affected bonds is selected randomly (not displayed).

C. A four bond kink move (Fig. 2C). Similar to the two bond moves, a prefabricated library of all possible four bond moves is employed.

D. Eight and 10 bond moves, where a fragment of the model chain moves a distance that is small enough to prevent crossing another portion of the model chain (not displayed).

E. A small random displacement of a large part of the chain, starting from the randomly selected segment up to the chain end (Fig. 2D). In contrast to the elemental moves (A–D), most of the side chain rotamers remain the same, due to the rigid body-like translation of the subchain.

The unit of the model time of a chain of length, n, is the time required for on average n attempts at moves A and B, $n-4$ attempts at moves C, $n-8$ plus $n-10$ attempts at the two kinds of moves of type D, and one attempt at move E. The moves are randomly mixed. The scaling of the model time to real time based on the frequency of local conformational transitions in real proteins would be rather ambiguous. Rather, attempts to relate the model time to real time should be based on longer relaxation phenomena. In this way, one may obtain a qualitative estimate of the time scales of various stages of the modeled protein dynamics and folding. The acceptance ratio of a particular move depends on the stage of the folding process and the system's temperature. With decreasing temperature, there is a slow down in the frequency of various processes. The model of dynamics allows for the slow diffusion of assembled fragments of secondary and supersecondary structure. Of course, these assembled fragments can also dissolve and reassemble in a different place. Therefore, various possible mechanisms of protein assembly are not a priori excluded.

Finer Lattice Model

For this lattice, the set of basis vectors is built from all the permutations of vectors of the type (3,1,1), (3,1,0), (3,0,0), (2,2,1), and (2,2,0). There are 90 vectors in the set. The mesh size of the underlying simple cubic lattice (1,0,0) is equal to 1.22 Å. The backbone valence bond angle lies in the range (72.5°, 154°), and the distance of closest approach for two Cαs is equal to 3.45 Å. In contrast to the somewhat exaggerated excluded volume of the coarser lattice backbone, the backbone of the finer lattice slightly underestimates the excluded volume of the main chain. These differences are in the range of the resolution of the finer lattice. Moreover, a small fraction of the proper volume of a given residue could be associated either with the main chain, united atom or the side group, united atom. In general, the new

TABLE I. Maximum Number of Side Group Rotamers for Lattice Models

Amino acid		Lattice	
		Coarser	Finer
Alanine	(A)	1	1
Serine	(S)	1	1
Cysteine	(C)	2	3
Valine	(V)	1	2
Threonine	(T)	1	3
Isoleucine	(I)	2	6
Proline	(P)	1	1
Methionine	(M)	5	13
Aspartic acid	(D)	3	8
Asparagine	(N)	3	10
Leucine	(L)	4	7
Lysine	(K)	9	25
Glutamic acid	(E)	8	21
Glutamine	(Q)	7	18
Arginine	(R)	13	58
Histidine	(H)	4	6
Phenylalanine	(F)	5	5
Tyrosine	(Y)	4	6
Tryptophan	(W)	6	6

discretization is somewhat more permissive; however, it is closer to the geometry of real proteins. The rms of fitted lattice backbones from Cα traces of PDB structures[30] is about 0.8 Å. The quality of fit is somewhat better than for the coarser lattice. The accuracy of the side chain representations improves significantly and is in the range of 1.0 Å for the centers of mass of particular rotamers. The number of side group rotamers for a given backbone conformation is on average two to three times larger than the corresponding number for the coarser lattice model (see Table I).

The model of the Monte Carlo dynamics is simplified. This seems to be acceptable due to the larger inherent flexibility of the finer lattice. On the other hand, one has to take into account the larger number of basis vectors. Consequently, tabularization of the four- (and more) bond moves is rather impractical on most computers. Taking these facts into consideration, the following set of elemental moves is used:

A. Random change of the rotamer of the randomly selected residue.

B. Three bond moves (Fig. 3A). The precalculated set of moves contains only those micro modifications that preserve the geometrical restrictions mentioned above. For a sequence of three backbone vectors, there are up to 168 acceptable new three bond sequences. Rotamers of the affected residues are randomized.

C. The small distance motion of a large, randomly selected, part of the chain (Fig. 3B). It is reconstructed by successive application of the three bond moves to the adjacent parts of the chain.

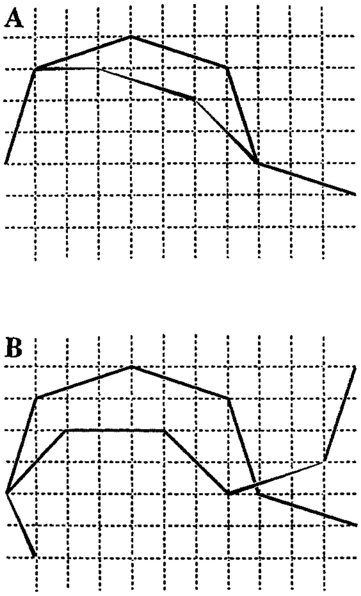

Fig. 3. Examples of finer lattice elemental moves. (**A**) Three bond kink move—there are up to 168 such modifications, given a particular starting geometry. (**B**) Larger distance moves, generated as a sequence of three bond moves applied to the randomly selected subchain. A conformational bias is applied in order to increase the acceptance ratio. The full conformational energy change is computed after the entire trial rearrangement is completed.

The definition of the model time unit and organization of the Monte Carlo algorithm are similar to that for the coarser lattice model. Both lattice models in the high temperature limit exhibit Rouse-type dynamics and the proper scaling of random coil dimensions with chain length.[27,45] This provides some additional evidence that MCD mimics to a large extent the qualitative features of the long time dynamics of real polypeptides.

INTERACTION SCHEME FOR LATTICE MODELS

The interaction scheme is divided into three parts. A part of the interaction scheme is sequence independent and is designed to keep the model system in that portion of conformational space which resembles proteins. Then, there are short- and long-range interactions that are amino acid specific, pairwise amino acid specific, and finally, there are multibody interactions. Schematically, the energy of the model polypeptide can be written as follows:

$$E = E_{\text{C}\alpha\text{-trace}} + E_{\text{H-bond}} + E_{\text{rot}} + E_{\text{sg-local}} + E_{\text{one}} + E_{\text{pair}} + E_{\text{tem}} \quad (1)$$

where $E_{\text{C}\alpha\text{-trace}}$ is the sequence independent statistical potential for the main chain Cα-trace conformation, $E_{\text{H-bond}}$ is the cooperative potential simulat-

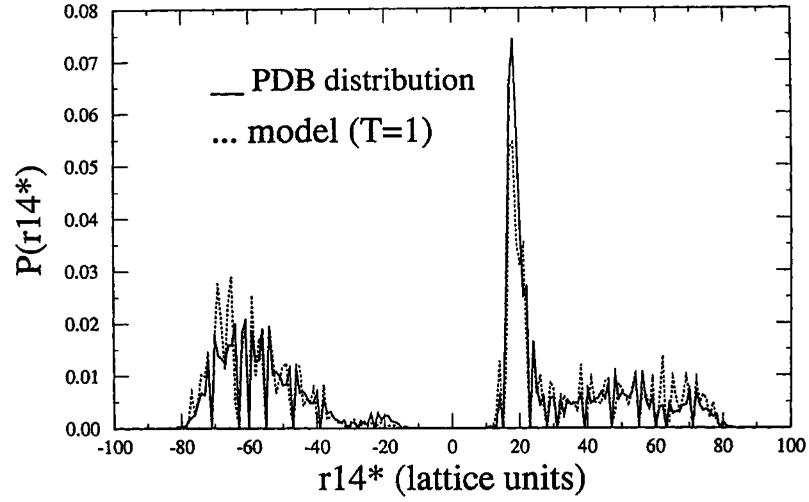

Fig. 4. Comparison of the distribution of chiral distances between the *i*th and *i*+3th α-carbons in the database (solid line) and in the finer lattice which has been regularized by an effective Ramachandran background potential, but without tertiary interactions (dashed line). Thus, the lattice discretization is consistent with the PDB distribution.

ing the hydrogen bond network in real proteins, E_{rot} is the rotamer energy for the single ball representation of the side chains, $E_{\text{sg-local}}$ is the energy dependent on the local angular correlations of the side group orientations, E_{one} is the amino acid specific centrosymmetric force, E_{pair} is the pairwise interaction of the side groups, and E_{tem} is the four body, side group contact map template interaction. The particular terms are described below. The method of derivation of the potentials, their statistical mechanical justification, their implementation in the Monte Carlo algorithms, and their different effects on the behavior of the model proteins are also discussed. The numerical values of various parameters of the proposed force field are either published[28] or in the case of large files, they are available by e-mail from the authors.[46]

Sequence Independent Potentials

Effective Ramachandran potential

Because various atomic details are neglected, lattice models have their own distribution of intrachain distances that are typical of random coil polymers. Therefore, protein like chain geometry has to be introduced into the model. To achieve this, the distribution of the distances between the ith and $i+3$th α-carbon vertices in the model chain, $r_{i,i+3}$, and the chirality of these three bond fragments is compared to the corresponding distribution for real proteins. Then, the bins in the discrete lattice distribution of these states are weighted by the appropriate Boltzmann factors in order to mimic the average distribution in proteins. The resulting set of energy parameters, $E_{C\alpha\text{-trace}}(\mathbf{b}_{i-1},\mathbf{b}_i,\mathbf{b}_{i+1})$, where $\mathbf{b}_i$ denotes the ith Cα backbone vector (from Cα i to $i+1$), was subsequently used in all folding simulations. This term plays the role of an effective average Ramachandran potential whose contribution to the total conformational energy for entire polypeptide is calculated by summation along the chain.

In Figure 4, we compare the finer lattice distribution of the intrachain distances, at a temperature $T=1.0$ (the temperature is dimensionless, since energy is always expressed in k_BT units), with the corresponding distribution for a set of 56 high-resolution globular proteins found in the Brookhaven Protein Data Bank[29,30] (PDB). The values of $r^2_{i,i+3}$ for the right-handed fragments are plotted along the positive x-axis, while the values for left-handed fragments are plotted along the negative x-axis. It may be noticed that the database distribution and the time averaged distribution from the long lattice simulation practically overlap. Therefore, the plot shows clearly that there is no bias toward any particular conformation (helices, extended, etc.) in the model lattice chain, when all other interactions are turned off. In this kind of "generic protein," the helix content is equal to the average helix content seen in the entire structural database. The highest narrow peak in the plot corresponds to helical conformations (along with a contribution of tight turns), while the two broader peaks correspond to less unique (with respect to backbone conformation measured by $r^2_{i,i+3}$) extended β-strand and coil conformations. In the presence of other interactions, this potential has to be suppressed by a factor in the range of 0.5. This is because other forces also tend to favor the proper local geometry of the model protein.

Hydrogen bonds

The second, sequence independent contribution to the potential implemented in the reduced model is the hydrogen bond potential with explicit cooperativity. The model hydrogen bond is designed to simulate some aspects of the hydrogen bond network of real proteins. The positions of backbone atoms participating in hydrogen bonds are not explicitly defined for lattice models. In principle, a reasonable approximation for all the heavy atom backbone coordinates can be generated based on the value of the bond angle between two consecutive virtual Cα–Cα bonds.[47] However, taking into consideration the limited accuracy of such a procedure as applied to lattice chains as well as the computational cost, a simplified scheme is proposed. The model H-bonds are generated explicitly from the Cα vertices, with account for different lengths and directionality. The model H-bond can be formed between two main chain beads i and j, provided that $|i-j| \geq 3$, and that the following geometrical restrictions are fulfilled:

$$R_{\min} \leq r_{i,j} \leq R_{\max} \tag{2a}$$

$$|(\mathbf{b}_{i-1}-\mathbf{b}_i) \cdot \mathbf{r}_{i,j}| \leq a_{\max} \tag{2b}$$

$$|(\mathbf{b}_{j-1}-\mathbf{b}_j) \cdot \mathbf{r}_{i,j}| \leq a_{\max} \tag{2c}$$

where $\mathbf{b}_i$ is the backbone vector, and $\mathbf{r}_{i,j}$ is the vector between "H-bonded" Cα vertices. $R_{\min}=4.8$ Å (4.6 Å) and $R_{\max}=7.0$ Å (7.3 Å) for the coarser (finer) lattice, respectively, and $a_{\max}$ is equal to 17.3 Å^2 (13.4 Å^2) for the coarser (finer) lattice. The different cut-off values reflect the different resolution of the two lattices. There is no asymmetry (donor versus acceptor) of the model H-bonds. Every model amino acid, except proline, can participate in at most two H-bonds, and proline can participate in one H-bond. These limitations suppress the number of possible realizations of the model H-bonds given only by Cα coordinates. The degeneracy of the model H-bond network is further suppressed by its implicit cooperativity. Allowing for lattice fluctuations, the above definition nicely corresponds to the geometry of the hydrogen bonded network of real proteins. In fact, the model definition recovers about 90% of the main chain hydrogen bonds as assigned by the Kabsch and Sander[48] method when applied to real proteins. Since the elements of secondary structure (helices, β-hairpins, and larger fragments) are reproduced by the lattice Cα traces with an accuracy in the range of 0.3–0.9 Å rms from native,[35] the method of Levitt and Greer could be also used for secondary structure assignment of the local and global level.[49] Indeed, allowing for some fluctuations of the lattice models, the geometric constraints given in Eq. (2) are very close to the Cα based distance definition proposed by Levitt and Greer.[49]

The model hydrogen bonds are cooperative. The system is additionally stabilized when each pair of adjacent H-bonds forms a helical or β-sheet type of hydrogen bond pattern. It has been noted that due to the Cα based definition of the H-bond pattern, there is no qualitative difference between the pattern seen in model parallel versus antiparallel β-sheets. The model definition neglects peptide bond orientations. The energy of the hydrogen bond network can be expressed as follows:

$$E_{\text{H-bond}} = \Sigma\Sigma\, E^{\mathrm{H}}\delta(i,j) + \Sigma\Sigma\, E^{\mathrm{HH}}\delta(i,j)\delta(i \pm 1, j \pm 1) \tag{3}$$

where i and j indicate the two residues of interest, E^{H} and E^{HH} are the energy of a single hydrogen bond and the cooperative contribution when a consecutive set of residues i,j and $i\pm1,j\pm1$ are hydrogen bonded, and $\delta(i,j)=1$ when the geometric criterion for H-bonds between Cαs i and j are satisfied. The geometric definition of hydrogen bond and its cooperativity are illustrated in Figure 5. In the absence of fine atomic details and the lack of explicit account of electrostatic interactions, the model cooperative network of hydrogen bonds plays the important role of a structure regularizing factor. The numerical values of E^{H} and E^{HH} were selected based on two criteria. First, the secondary structure assignment in the folded state (lattice realizations of the native state of plastocyanin and flavodoxin have been used) has to be as accurate as possible, and, on the other hand, the amount of secondary structure in the unfolded state (near the folding temperature) has to be marginal. $E^{\mathrm{H}}\sim0.5$, and $E^{\mathrm{HH}}\sim0.75$ have been used in the folding simulations (all the numerical values correspond to $T=1$). This is of the same range as the values used previously in MCD simulations of cooperative coil–helix and coil–β-globule transitions in much simpler systems.[27]

Short-Range Interactions

Rotamer energy

For a given local backbone conformation defined by two consecutive Cα–Cα vectors, there is a set of side group rotamers. Each rotamer is represented by a single ball. The number of rotamers in the set depends on amino acid identity. For alanine, there is always only one rotamer; for the bigger amino acids, the number of rotamers is larger. The maximum number of rotamers for various amino acids are presented in Table I. The rotamer library was constructed as follows: First, for each residue in all the proteins from the structural database, the best fit of two lattice vectors was calculated. The resulting projections were grouped according to backbone conformation and the amino acid identity. Then, the average center of mass of all heavy atoms (equal mass assumed) of the most populated side chain rotamer was calculated. If the next most populated rotamer's average center of mass was within a specified distance threshold, then that rotamer was not

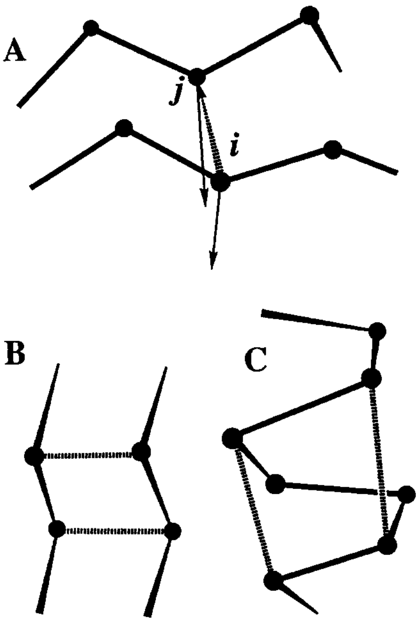

Fig. 5. Illustration of the geometry of the model H-bonds (**A**). $\mathbf{r}_{ij}$ from Eq. (2), the vector between two Cαs, is shown as dashed arrow, the vectors $\mathbf{b}_{k-1}-\mathbf{b}_k$, with $k=i$ or j are shown as thin solid arrows. (**B**) β-Sheet fragment, with model H-bonds shown in dashed lines. (**C**) Helical fragment. In cases A and B, there is one cooperative contribution due to the regular ordering of the two H-bonds.

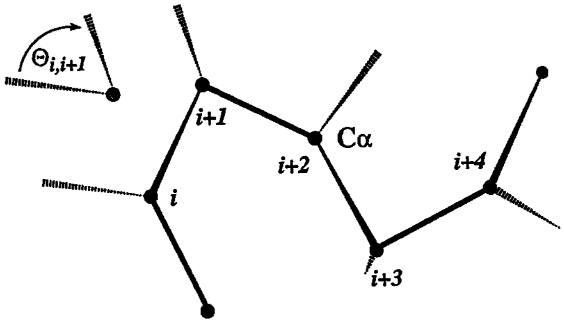

Fig. 6. Illustration of the geometry associated with the short-range angular correlation of the side group vectors. See the text for details.

added to the rotamer library, but was counted as belonging to the first rotamer. If the distance between the two centers of mass was beyond the threshold, then the second rotamer was added to the library. The process is repeated until all side chain rotamers have been compared to the existing library, subsumed as belonging to a previous rotamer or added as a new member. The distance threshold was set equal to 1.7 Å (1. Å) for the coarser (finer) lattice. The energy of a given rotamer is defined as $-\ln(f_{\mathrm{rot}} \times N_{\mathrm{rot}})$, where for a given backbone conformation, f_{rot} is the frequency of occurrence of the rotamer in the library ($\Sigma f_{\mathrm{rot}} = 1$), and N_{rot} is the number of different rotamers of the residue under consideration.

Local side chain orientational coupling

At least for the reduced representation employed here, a statistical analysis of the structural database seems to indicate that the most specific information about sequence-dependent local conformational propensities is encoded in the angular correlations between the orientations of the side groups. Figure 6 schematically shows a small fragment of the protein Cα backbone, with arrows pointing toward the average center of mass of the side group. Statistics have been the collected for pairs of amino acids at position i and $i+1$, i and $i+2$, i and $i+3$, and i and $i+4$, using 10 bins for the $\cos(\Theta_{i,i+k})$. The energy (dimensionless) associated with a particular bin has been determined by comparing the observed population with respect to a "random" population, i.e., one with a uniform distribution in all the bins, in the usual way by:

$$\epsilon[\cos(\Theta)] = -k_BT \times ln(observed/random), \text{ with } k_BT = 1 \quad (4)$$

The short-range interactions for the entire chain read as follows:

$$\begin{aligned} E_{\mathrm{short}} &= E_{\mathrm{rot}} + E_{\mathrm{sg-local}} \\ &= \sum_i \left\{ E_{\mathrm{rot}}(a_i) + \sum_{k=1}^{4} E_k(\Theta_{i,i+k}, a_i, a_{i+k}) \right\} \quad (5) \end{aligned}$$

where a_i is the amino acid at position i down the chain, $\Theta_{i,i+k}$ is the angle between the side group vectors for the actual rotamers. $E_{\mathrm{rot}}(a_i)$ is the energy of a particular rotamer. Again, this contribution to the potential of mean force comes from the statistics of three dimensional structures in the PDB database.[29,30]

General observations

The short-range angular interaction potential triggers the formation of secondary structure elements. In the absence of any sequence specific, long-range interactions, although in presence of the generic interactions discussed above, the resulting secondary structure is on average about 70% correct, when backbone distance criteria are applied. This means that on average helical (or turn) conformations, as measured by the Cα–Cα distances down the chain (and backbone handedness), are correctly recovered. A similar level of accuracy is observed for expanded states. The implicit cooperativity of this potential and its explicit coupling with the side group degrees of freedom facilitate a much better accuracy of secondary (and supersecondary) structure prediction when moderated by long-range interactions. This interesting application of the proposed protein model will be exploited elsewhere.[50]

The best results are obtained when the strength of the short-range interactions is scaled by a factor of 0.75. Otherwise, the chain mobility is strongly suppressed, and the system tends to be locked in local minima on the conformational energy landscape. The necessity of scaling down the various potentials emerges from the incomplete separation of the contributions of the various potentials and the intrinsic cooperativity of intraprotein interactions.

Long-Range Interactions

One body, centrosymmetric burial potential

There are three contributions to the long-range interactions. The first one is a central, one body, amino acid specific potential. It is based on the observation that some amino acids tend to be buried in the interior of the globular protein, others tend to be just below the surface, while still others like to be exposed to the surrounding solvent. The potential for each amino acid has been derived from the statistics of single domain globular proteins. It assumes spherical symmetry of the compact globule, and the energy of each amino acid is a function of its identity and the distance of the center of the side group (Cα in the case of glycine) from the center of gravity of the protein chain.[51] Application of this potential requires an estimate of the radius of gyration (S) of the modeled (n residue) protein in its folded state. Since all globular proteins are more or less closely packed and have on average the same density, the requisite estimate can be done with sufficient accuracy.

$$S = 2.2\, n^{0.38} \qquad (\text{in Å}). \tag{6}$$

The above equation is derived from the fit to a set of single domain structures from PDB. Note that the exponent is somewhat larger than 0.33 expected for a closely packed long polymer,[52] reflecting a finite size effect. The one body potential makes a marginal contribution to the energy of the native states; however, for denatured, expanded states, it can have quite a large positive contribution. Of course, this simplified potential cannot be applied to larger proteins having well-separated domains. In such cases, a different approach is required, the detailed discussion of which is beyond the scope of the present paper. However, we note that an alternative approach may invoke an energy penalty for strongly hydrophobic side groups being exposed (whose degree of exposure is measured by comparison of the actual number of contacts for a given side group with the expected average number of binary contacts for this amino acid type). In some refinement runs employing the finer lattice discretization, this kind of one body, amino acid specific burial term has been introduced in addition to the central, one body force. This part of the potential has been also derived from the statistics of the database. For a given side group, and a given actual number of contacts with other side groups, the energy is assumed to be proportional to $-\ln$(number of contacts/average number of contacts for this amino acid). Local contacts (up to the fourth nearest neighbors along the chain) and nonlocal contacts have been treated separately. The effect of this update, however small, tends to generate better defined conformations of loops and chain ends.

Pair potential

Then, there are pairwise interactions. These can be written as follows:

$$E_{\text{pair}} = \begin{cases} \mathrm{E}_{\text{rep}}, & \text{for } r_{ij} < R^{\text{rep}}{}_{ij} \\ \epsilon_{ij}, & \text{for } R^{\text{rep}}{}_{ij} < r_{ij} < R_{ij}, \quad \text{and } \epsilon_{ij} \geq 0 \\ f\epsilon_{ij}, & \text{for } R^{\text{rep}}{}_{ij} < r_{ij} < R_{ij}, \quad \text{and } \epsilon_{ij} < 0 \end{cases} \tag{7a}$$

where i and j are the interacting amino acids separated down the chain at least by one residue (the nearest neighbors down the chain do not interact). The cut-off distances R^{rep} and R_{ij} are amino acid pair specific. E_{rep}, which is in the range of 4 k_BT, is a penalty for overlapping the repulsive cores of the side groups. The same repulsive force is applied to the side group-main chain overlaps. The ϵ_{ij} are pairwise, amino acid-specific interaction parameters and are derived from the statistics of a database of high resolution three dimensional structures (see Appendix for the details of the derivation of these parameters). The factor f depends on the angle between average orientations of the backbone fragment, defined by the vectors $\mathbf{u}_i = \mathbf{r}_{i+2} - \mathbf{r}_{i-2}$, with $\mathbf{r}_i$ being the Cartesian coordinate of the ith α-carbon. In particular:

$$f = 1.0 - \{\cos^2(\mathbf{u}_i,\mathbf{u}_j) - \cos^2(20°)\}^2. \tag{7b}$$

The above factor reflects the average angle between elements of secondary structure seen in globular proteins. The maximum occurs at 20°, and the minimum, which occurs when the chain elements are perpendicular, is about 0.22 of the maximum strength. Note that small deviations from perpendicular orientations make pairwise interactions much more favorable (e.g., for 70°, $f = 0.56$). Some interactions, like those between distant β-strands in TIM barrel motifs (and in some other folds), will be artificially suppressed. However, most binary interactions occur between adjacent strands or helices. Consequently, this bias is rather small. Identical results are obtained for the folding of protein A when Eq. (7b) is supplemented by the term, 1-$\cos^2(\mathbf{u}_i,\mathbf{u}_j)$, which has a maximum at 90°.

Some comment is required about the definition of the pairwise interaction contact cut-off. Two side groups in a real protein are considered to be in contact when any pair of their heavy atoms is "in contact," i.e., the distance between these atoms is smaller than 4.2 Å. A compilation of the database statistics on the pairwise contact distances reveals that they exhibit a rather sharp distribution (see

Table I of ref. 28). A strongly repulsive core is assumed up to a distance equal to the average contact distance minus two standard deviations. The soft, square well envelopes extend one standard deviation beyond the mean values. The numerical values of the one body potential, pairwise potentials, and the sizes of the spherical side groups can be found elsewhere.[28]

Multibody side chain packing interactions

The set of interactions discussed above can fold a limited set of small globular proteins. Moreover, it can distinguish between correct and incorrect folds of larger globular proteins. However, the obtained folds are of low resolution, and their side chain packing is rather nonspecific. These folds usually have the character of molten globules,[53,54] with well-defined secondary structure, a somewhat larger volume than close packed structures, and a liquid like hydrophobic core. In contrast, it is known that the pattern of side chain packing in native proteins is highly specific and is more solid than liquid-like. Moreover, experimental studies show that the transition from the molten globule state to the native state is very cooperative.[53,54] Therefore, somewhat in analogy to the cooperativity of the H-bond network, a cooperativity of the side chain packing is proposed. Since our cooperative model of H-bonds reproduced quite well the cooperative helix–coil transition, it is expected that a similar parameterization, when applied to side chain packing, could perhaps facilitate cooperative fixation of side chains in the native state. In the present and our previous simulations,[28] the cooperativity of the side chain packing is accounted for by generic multibody interactions of the following form:

$$E_{\text{tem}} = (\epsilon_{i,j} + \epsilon_{i+k,j+n})\, C_{i,j} \times C_{i+k,j+n}; \text{ with } |k| = |n|,\ n = \pm 3, \text{ and } \pm 4 \qquad (8)$$

where $C_{i,j} = 1$ (0) if side chains i and j are (not) in contact, i.e., $r_{i,j} < R_{i,j}$. This "template" contribution makes some patterns of the side chain packing explicitly more favorable. The templates used here are applicable to helical, as well as to β-sheet type patterns of the side chain contacts. We also note that the cooperative templates only make a substantial contribution to the total energy subsequent to formation of the topology found in the native state. Since these molten globule intermediates already have a substantial amount of the native state's secondary structure, it is evident that the inclusion of the cooperative packing templates is not responsible for the structural class (helix, beta, or mixed motif) that the sequence chooses to adopt.

Figure 7 shows, as typical examples, several pairs of contacts coupled according to the above cooperative term. The patterns of helix–helix and β-sheet contacts were literally taken from the X-ray contact map of the real protein, thioredoxin, (2trx), a small α/β globule consisting of 108 amino acids. In most proteins, these patterns are not so clean; there are usually some additions or deletions from such "ideal" templates. Nevertheless, very similar patterns can be seen many times in practically all globular proteins.[55] The templates embodied in Eq. (8) are generic in that they do not bias toward any specific secondary structure; however, they facilitate a specific side chain packing pattern when the secondary structure develops. Folding simulations[28,39] of two proteins designed by DeGrado and co-workers[36,37] showed that the proposed multibody cooperative interactions do not enforce side chain fixation when the real protein[36] does not undergo a transition from the molten globule to the native state. However, for a reengineered sequence,[37] in agreement with experimental data, the simulations[28] show native like side chain fixation. Parenthetically, we note that instead of a cooperative term that favors proper, protein-like, contact–contact correlations, it is possible to use an apparently equivalent approach where nonphysical clusters of the side groups are penalized. In the last case, the pairwise interactions have to be somewhat stronger.

Why are these kinds of cooperative terms necessary? First, the reduced representation model, due to the "fuzzy" description of the side groups, cannot reproduce the fine effects of close atomic packing, where perhaps the cooperative thermodynamics of the side group nestling occurs. In this respect, the contact templates simulate the observed fine packing. However, even in detailed MD simulations of protein structures, the specific patterns of the side chain packing seen in the native state seem to degenerate.[56] Therefore, it is possible that multibody interactions have a more fundamental physical justification than the practical one invoked in these reduced models. This question will be further addressed in the near future in the context of MD simulations.

The scaling of various contributions to the force field of the present model has been done in preliminary runs, by requiring a marginal level of secondary structure in the unfolded state and a high level of secondary structure in the collapsed not necessarily native states. Since the long- and short-range interactions are not strictly separated, they have to be properly balanced. While this procedure seems to be somewhat arbitrary, it at least allows us to fold several proteins using the same set of interaction parameters. Due to insights gained from the previous work,[27,28] the search was not completely blind.

OVERVIEW OF THE FOLDING PROCEDURE

The folding simulations start from randomly generated unfolded states of the model protein restricted to the coarser lattice. Folding proceeds by

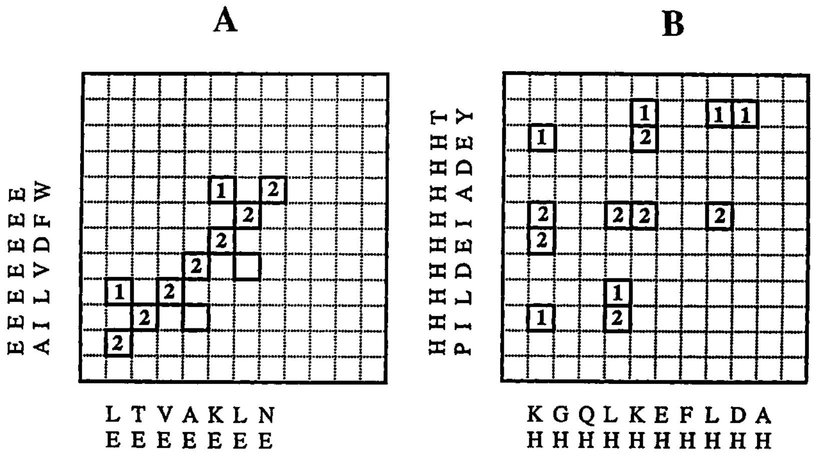

Fig. 7. Examples of contact maps for parallel β-strand–β-strand and parallel helix–helix motifs. The numbers indicate the number of cooperative "template" terms which stabilize particular contacts.

simulated temperature annealing, or under isothermal conditions, depending on whether the transition temperature has been already estimated. The coarser lattice model, with a mesh size equal to 1.7 Å, tends to adopt loosely defined folded conformations much faster than the finer lattice model does. It is unclear if use of the finer lattice model over the entire folding pathway would decrease the fraction of misfolded, kinetically trapped, compact states observed in some folding experiments. The answer to this question will require numerous lengthy simulations.

In most cases, the folds obtained from the coarser lattice simulations have the correct secondary structure and an average contact map describing the side chain packing which could be considered native-like. However, the instantaneous contact maps from various simulations differ substantially; the overlap between them is in the range of 20–35%. These values appear to be too small for a plausible representation of the unique native state, even taking into consideration the limited resolution of the model. Moreover, when the lifetimes of these binary contacts are measured, it becomes clear that the native-like side chain fixation is not always possible to achieve in the coarser lattice representation. Instead, the packing of the model protein interior is to some extent liquid-like, exhibiting high side chain mobility. On the other hand, for very simple designed helical proteins, the difference between molten globule and native-like states has been qualitatively reproduced.[28] This provides additional evidence that the generic side chain packing templates do not guarantee that a native-like state, with long lived side chain contacts is achieved.

Once a series of coarser lattice folds are obtained, each is subject to a refinement procedure. First, the Cα trace is projected onto the finer lattice, whose mesh size is 1.22 Å. After a short relaxation of the minor packing incompatibilities introduced by the projection procedure, the finer lattice systems gradually adopt well-defined packing, will all signatures of side chain fixation. These folds, when simulated well below the transition temperature, have a backbone rms from real native states in the range of 2 to 4 Å, depending on protein size and identity. In the final stage of the folding protocol, the entire full atom structures may be reconstructed.[57]

DISCUSSION AND CONCLUSION

The relatively high accuracy of our reduced models[28,57] has been achieved due to a sufficiently flexible lattice representation[35] of the main chain conformation and a careful implementation of the geometric structure of proteins. In addition, several terms, novel in the context of "simplified" model potentials, have been implemented, which apparently mimic in a reasonable way a protein-like force field. Of course, we realize that the potential which is presently used has not been self-consistently derived. That is, the various terms are introduced independently and are designed to represent particular aspects of the interactions controlling protein folding. Therefore, future work will be focus on the preparation of mean force potentials that are derived in a more consistent way. Using a genetic algorithm as an optimization procedure, it should be possible to obtain a more specific, and self-consistent potential consisting of one body, pairwise and multibody interactions of the united atoms employed in these reduced models. In addition, the model H-bond network scheme should also be reexamined. Obvious improvements may include hydrogen bonding with side groups, as well as accounting for the donor–acceptor asymmetry of the H-bond.

One may also wonder if the single ball representation of side group rotamers is not the weak link in the present model. Very likely, the resolution of the model could be improved by a finer side chain representation; however, it appears that the single ball rotamer representation is not yet fully exploited. The relatively good structures obtained for various helical bundles and the rather irregular fold of crambin suggest that the packing in the present model is surprisingly good. Moreover, application of a more elaborate set of side chain contact map templates may fix some ambiguities of the model protein packing and make the transition from the molten globule to the native state more cooperative. It also has to be kept in mind that a more accurate representation of the side chains could prohibitively increase the computational cost. This is another reason why improvements in the potential using the present level of discretization will be attempted first. Hopefully, this will allow the folding of the more complicated motifs of β-protein and larger α/β barrels. Preliminary attempts at folding these more complicated motifs indicate that while the number of secondary and supersecondary elements are in general correctly predicted, to date the native topology has not been recovered. Whether this merely reflects insufficient computer time, inadequate potentials or both remains to be established.

Another question that has to be addressed is the possibility of implicitly built-in biases in the proposed force field. This kind of bias could manifest itself as a hidden target potential. On the trivial level, the description of various potentials shows that it is not the case. On the other hand, up to now only a few simple and small proteins have been folded using the described method.[1,28,39,57] Therefore, one has always keep in mind the chance that the model and its potentials somehow favor these folds. Some evidence that there is no significant bias comes from stability tests and refolding experiments performed on more complicated β and α/β-proteins. These proteins (plastocyanin, flavodoxin) when started from conformation having a relatively large rms (in the range of 12 Å from native Cα trace), but with substantial memory of the native like overall fold, collapse to conformations having mostly correct secondary structure, an rms deviation from native in the range of 4–6 Å, and about 50% of the native side group contacts. Due to their simpler topology, it is very likely that small helical proteins are easier to fold on the computer than β-proteins (or α/β-proteins) of comparable size. Finally, it appears that the further justification of this reduced, but nontrivial, model of protein conformation and protein folding will have to be provided by expanding the set of tractable folds. Of course, each update of the force field, the Monte Carlo algorithm, or the folding protocol should not only allow us to fold new proteins, but should also improve the accuracy of the previously folded structures. This way one may learn about various factors controlling protein folding, therefore providing elements of a solution to the protein folding problem.

ACKNOWLEDGMENTS

Valuable discussions with Drs. William Beers, Charles L. Brooks, III, and Adam Godzik are gratefully acknowledged. This research was supported in part by grant GM-37408 of the Division of General Medical Sciences of the National Institutes of Health.

REFERENCES

1. Kolinski, A., Skolnick, J. Monte Carlo simulation of protein folding. II. Application to protein A, ROP, and crambin. Proteins 18:353–366, 1994.
2. Creighton, T.E. Protein folding. Biochem. J. 270:131–146, 1990.
3. Levitt, M. Protein folding. Curr. Opinion Struct. Biol. 1:224–229, 1991.
4. Dill, K.A. Folding proteins: Finding a needle in a haystack. Curr. Opinion Struct. Biol. 3:99–103, 1993.
5. Wodak, S.J., Rooman, M.J. Generating and testing protein folds. Curr. Opinion Struct. Biol. 3:247–259, 1993.
6. Bowie, J.U., Clarke, N.D., Pabo, C.O., Sauer, R.T. Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. Proteins 7:257–264, 1990.
7. Bowie, J.U., Luethy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three dimensional structure. Science 253:164–170, 1991.
8. Bryant, S.H., Lawrence, C.E. An empirical energy function for threading protein sequence through folding motif. Proteins 16:92–112, 1993.
9. Finkelstein, A.V., Reva, B.A. A search for the most stable folds of protein chains. Nature (London) 351:497–499, 1991.
10. Godzik, A., Skolnick, J., Kolinski, A. A topology fingerprint approach to the inverse folding problem. J. Mol. Biol. 227:227–238, 1992.
11. Godzik, A., Skolnick, J. Sequence-structure matching in globular proteins: Application to supersecondary and tertiary structure determination. Proc. Natl. Acad. Sci. U.S.A. 89:12098–12102, 1992.
12. Gribskov, M., McLachlan, M., Eisenberg, D.P. Profile analysis: Detection of distantly related proteins. Proc. Natl. Acad. Sci. U.S.A. 84:4355–4358, 1987.
13. Sippl, M.J., Weitckus, S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. Proteins 13:258–271, 1992.
14. Jones, D.T., Taylor, W.R., Thornton, J.M. A new approach to protein fold recognition. Nature (London) 358:86–89, 1992.
15. Finkelstein, A.V., Ptitsyn, O.B. Why do globular proteins fit the limited set of folding patterns? Prog. Biophys. Mol. Biol. 50:171–190, 1987.
16. Piela, L., Kostrowicki, J., Scheraga, H.A. The multiple-minima problem in the conformational analysis of molecules. Deformation of the potential energy hypersurface by the diffusion equation method. J. Phys. Chem. 93:3339–3346, 1989.
17. Kostrowicki, J., Scheraga, H.A. Application of the diffusion method for global optimization in oligopeptides. J. Phys. Chem. 96:7442–7449, 1992.
18. Skolnick, J., Kolinski, A. Computer simulations of globular protein folding and tertiary structure. Annu. Rev. Phys. Chem. 40:207–235, 1989.
19. Jernigan, R.L. Protein folds. Curr. Opinion Struct. Biol. 2:248–256, 1992.
20. Karplus, M., Petsko, G.A. Molecular dynamics simulations in biology. Nature (London) 347:631–639, 1990.
21. Brooks, C.L., III, Karplus, M., Pettit, B.M. Proteins: A theoretical perspective of dynamics, structure and thermodynamics. Adv. Chem. Phys. 71:259, 1988.
22. Brooks, C.L., III, Karplus, M. Solvent effects on protein motion and protein effects on solvent motion. Dynamics of

the active site region of lysozyme. J. Mol. Biol. 208:159–181, 1989.

23. Brooks, C.L., III, Characterization of "native" apomyoglobin by molecular dynamics simulation. J. Mol. Biol. 233: 521–527, 1993.
24. Novotny, J., Bruccoleri, R., Karplus, M. An analysis of incorrectly folded protein models. Implication for structure prediction. J. Mol. Biol. 177:787–818, 1984.
25. Brooks, C.L., III, Molecular simulation of peptide and protein unfolding: In quest of a molten globule. Curr. Opinion Struct. Biol. 3:92–98, 1993.
26. Elber, R., Karplus, M. Multiple conformational states of proteins: A molecular dynamics analysis of myoglobin. Science 235:318–321, 1987.
27. Kolinski, A., Skolnick, J. Discretized model of proteins. I. Monte Carlo study of cooperativity in homopolypeptides. J. Chem. Phys. 97:9412–9426, 1992.
28. Kolinski, A., Godzik, A., Skolnick, J. A general method for the prediction of the three dimensional structure and folding pathway of globular proteins: Application to designed proteins. J. Chem. Phys. 98:7420–7433, 1993.
29. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Simanouchi, T., Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
30. PDB Quarterly Newsletter, No. 63, January 1993.
31. Skolnick, J., Kolinski, A. Simulations of the folding of a globular protein. Science 250:1121–1125, 1990.
32. Godzik, A., Skolnick, J., Kolinski, A. Simulations of the folding pathway of triose phosphate isomerase-type α/β barrel proteins. Proc. Natl. Acad. Sci. U.S.A. 89:2629–2633, 1992.
33. Skolnick, J., Kolinski, A., Godzik, A. From independent modules to molten globules: Observations on the nature of protein folding intermediates. J. Mol. Biol. 90:2099–2100, 1993.
34. Levinthal, C. Are there pathways for protein folding? Chim. Phys 65:44–45, 1968.
35. Godzik, A., Kolinski, A., Skolnick, J. Lattice representations of globular proteins: How good are they? J. Comp. Chem. 14:1194–1202, 1993.
36. Handel, T., DeGrado, W.F. A designed 4-helical bundle shows characteristics of both molten globule and native states of proteins. Biophysical J. 61:A265, 1992.
37. Raleigh, D.P., DeGrado, W.F. A de novo designed protein shows a thermally induced transition from a native to a molten globule like state. J. Am. Chem. Soc. 114:10079–10081, 1992.
38. Handel, T.M., Williams, S.A., DeGrado, W.F. Metal ion-dependent modulation of the dynamics of a designed 4-helix bundle. Science 261:879–885, 1993.
39. Godzik, A., Kolinski, A., Skolnick, J. De novo and inverse folding predictions of protein structure and dynamics. J. Comp. Aided Mol. Design 7:397–438, 1993.
40. Binder, K., ed., "Monte Carlo Methods in Statistical Physics." Berlin: Springer-Verlag, 1986.
41. Baumgartner, A. Simulation of polymer motion. Annu. Rev. Phys. Chem. 35:419–435, 1984.
42. Kolinski, A., Skolnick, J., Yaris, R. Does reptation describe the dynamics of entangled, finite length polymer systems? A model simulation. J. Chem. Phys. 86:1567–1585, 1987.
43. Kolinski, A., Skolnick, J., Yaris, R. Monte Carlo studies on the long time dynamic properties of dense cubic lattice multichain systems. I. The homopolymeric melt. J. Chem. Phys. 86:7164–7173, 1987.
44. Kolinski, A., Skolnick, J., Yaris, R. Monte Carlo studies on the long time dynamic properties of dense cubic lattice multichain systems. II. Probe polymer in a matrix of different degrees of polymerization. J. Chem. Phys. 86:7174–7180, 1987.
45. Kolinski, A., Milik, M., Skolnick, J. Static and dynamic properties of a new lattice model of polypeptide chains. J. Chem. Phys. 94:3978–3985, 1991.
46. Kolinski, A., Skolnick, J. Parameters of statistical potentials. Available by ftp from the public directory: scripps.edu (pub/MCDP) 1993.
47. Rey, A., Skolnick, J. Efficient algorithm for the reconstruction of a protein backbone from the α-carbon coordinates. J. Compt. Chem. 13:443–456, 1992.
48. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
49. Levitt, M., Greer, J. Automatic identification of secondary structure in globular proteins. J. Mol. Biol. 114:181–293, 1977.
50. Rey, A., Kolinski, A., Skolnick, J. Application of a discretized protein model to secondary structure prediction, in preparation.
51. Nikishawa, K., Ooi, T. Radial locations of amino acid residues in a globular protein: Correlation with the sequence. J. Biochem. 100:1043–1047, 1986.
52. de Gennes, P.G. "Scaling Concepts in Polymer Physics." Ithaca, NY: Cornell University Press, 1979.
53. Kuwajima, K. The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. Proteins 6:87–103, 1989.
54. Ptitsyn, O.B., Pain, R.H., Semisotnov, G.V., Zerovnik, E., Razgulyaev, O.I. Evidence for a molten globule state as a general intermediate in protein folding. FEBS 262(1):20–24, 1990.
55. Godzik, A., Skolnick, J., Kolinski, A. Regularities in interaction patterns of globular proteins. Prot. Engineer. 6:801–810, 1993.
56. Elofsson, A., Nilsson, L. How consistent are molecular dynamics simulations? Comparing structure and dynamics in reduced and oxidized Escherichia coli thioredoxin. J. Mol. Biol. 233:766–780, 1993.
57. Skolnick, J., Kolinski, A., Brooks, C.L., III, Godzik, A. A method for prediction of protein structure from sequence. Current Biol. 3:414–423, 1993.
58. Hill, T.L. "An Introduction to Statistical Thermodynamics." New York: Dover, 1960.

APPENDIX

In the calculation of the pair potential, we divide the protein into classes consisting of the backbone and the side chains appropriate to the twenty amino acids; thus, there are 21 different types of interacting groups. A contact between group i and j occurs when any heavy atom in the side chain i (or backbone) is within 4.2 Å of a heavy atom of group j. We count the total number of observed individual contacts between group i and j, $N(i,j)_{\mathrm{obs}}$. Alternatively, one could simply count a contact as that when at least one side chain heavy atom of i lies within 4.2 Å of side chain j. We have opted for the former definition, because it may reflect the relative strength of contacts. That is, if a pair of side chains has on average many individual contacts, then the strength of the interaction should be stronger than if it makes just one contact on average. Of course, we have to correct the interaction scale for side group size, i.e., bigger side chains have more contacts simply because they are larger.

The pair potential between residues i and j is defined by

$$\epsilon_{ij} = -\ln[N(i,j)_{\mathrm{obs}}/\bar{N}(i,j)] \qquad \text{(A1)}$$

where $N(i,j)_{\mathrm{obs}}$ and $\bar{N}(i,j)$ are the observed and expected number of contacts if the distribution is random. The crux of the calculation is the estimation of $\bar{N}(i,j)$. In reality, the determination of $\bar{N}(i,j)$ is very complicated. In the environment of folded proteins, one desires the expected number of contacts between amino acid pairs having the shape and size of real amino acids, but where there are no interactions other than those which preserve the excluded volume. To estimate $\bar{N}(i,j)$, we adopt a Bragg Williams approximation and proceed[58] by analogy to the Flory Huggins theory for polymers, generalized here to a heterogeneous system including side chains. Each side chain and backbone heavy atom is assumed to have a total coordination number of z. (In what follows, we take the average coordination number of 5.) For each chemical bond formed, the remaining sites available for interaction is reduced by one. Thus, the backbone N has $z-2$ available sites, the carbonyl oxygen has $z-1$ sites, the alanine methyl group has $z-1$ sites, etc. Let Z_j be the total number of available sites of group j. (It is equal to the sum of the non bonded sites of all the heavy atoms comprising group type j.) Thus, if there are N_j molecules of type j, then the total number of possible interacting sites is

$$N = \sum_{j=1}^{21} N_j Z_j. \tag{A2}$$

Now, the total number of observed interactions is obtained by counting the total number of contacts in the system, $N_{\mathrm{T,obs}}$. If the calculation is done correctly $N \geq N_{\mathrm{T,obs}}$. The fraction of "holes" in the system is obtained by

$$\phi_{\mathrm{holes}} = 1 - N_{\mathrm{T,obs}}/N. \tag{A3}$$

For $z=5$, we find that $\phi_{\mathrm{holes}}=0.026$, a quite reasonable value for a densely packed system such as a protein.

The fraction of interaction sites (or the surface fraction) contributed by group type i is

$$\phi_i = \frac{N_i Z_i}{\sum_{j=1}^{21} N_j Z_j}. \tag{A4}$$

If all the heavy atoms are taken to be equal in size and having the same coordination number, then ϕ_i is the volume fraction of i. For molecule i, neglecting end effects, the total number of possible interacting sites is $N_i Z_i$. The probability that these sites can interact with group type j ($\neq i$) is ϕ_j. Thus, the expected number of ij contacts

$$\bar{N}(i,j) = N_i Z_i \phi_j = N \phi_i \phi_j. \tag{A5}$$

Similarly, the expected number of contacts between identical groups is

$$\bar{N}(i,i) = \frac{N\phi_i^2}{2}. \tag{A6}$$

The factor of two corrects for over counting.

At this juncture, a number of observations are appropriate. First, this treatment accounts for the fact that groups of different size will have a different number of interactions even if the ensemble is random. Because a site fraction, ϕ_i, is used, bigger groups have more interactions simply because they are larger. Note that ϕ_i is not equal to the mole fraction of residues; this would only hold if all groups contained an identical number of heavy atoms having an identical coordination number. Thus, the use of the mole fraction in the calculation of the expected number of contacts is incorrect; it makes larger groups more attractive simply on the basis of their size. Finally, this treatment could be generalized to include the actual surface fraction of different groups, thereby improving the accuracy of the approximation to the expected number of contacts.