PROTEINS: Structure, Function, and Genetics 18:353–366 (1994)

Monte Carlo Simulations of Protein Folding. II. Application to Protein A, ROP, and Crambin

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 **The hierarchy of lattice Monte Carlo models described in the accompanying paper (Kolinski, A., Skolnick, J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins 18:338–352, 1994) is applied to the simulation of protein folding and the prediction of 3-dimensional structure. Using sequence information alone, three proteins have been successfully folded: the B domain of staphylococcal protein A, a 120 residue, monomeric version of ROP dimer, and crambin. Starting from a random expanded conformation, the model proteins fold along relatively well-defined folding pathways. These involve a collection of early intermediates, which are followed by the final (and rate-determining) transition from compact intermediates closely resembling the molten globule state to the native-like state. The predicted structures are rather unique, with native-like packing of the side chains. The accuracy of the predicted native conformations is better than those obtained in previous folding simulations. The best (but by no means atypical) folds of protein A have a coordinate rms of 2.25 Å from the native Cα trace, and the best coordinate rms from crambin is 3.18 Å. For ROP monomer, the lowest coordinate rms from equivalent Cαs of ROP dimer is 3.65 Å. Thus, for two simple helical proteins and a small α/β protein, the ability to predict protein structure from sequence has been demonstrated.**

Key words: tertiary structure prediction, protein folding pathways, molten globule state, protein A, crambin

INTRODUCTION

In the previous paper of this series,[1] we described a hierarchical method for the simulation of protein folding. The method employs high coordination lattices of increasing resolution and fast Monte Carlo dynamics (MCD) algorithms. A coarser lattice model is used for fast assembly of protein topology, and the resulting folding simulations exhibit many features of real proteins. For some designed sequences,[2,3] even the most characteristic features of the transition state have been reproduced.[4] However, for naturally occurring proteins, it appears that the resolution of the coarser lattice is somewhat too low. Therefore, after assembly of the global topology, the simulation is continued on a finer lattice[1] with considerably better geometric accuracy.[5] With the appropriate rescaling of geometric parameters, the force field is essentially the same for both lattices.[1] Because the finer lattice models have a different chain entropy and better side chain packing, the magnitudes of the contributions from the various potentials to the conformational energy are slightly different. In this paper, 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 crambin (1 crn). The first two are helical, while crambin is a small α/β protein.

The remainder of this paper is organized as follows: In the next section, the folding protocol is briefly summarized. Then, the results of simulations on three proteins are presented in detail. In the case of ROP monomer (mROP), we analyze the properties of the folding intermediate closely resembling the molten globule state. The long time dynamics of the mROP bundle is also examined. In the case of crambin, we demonstrate how the efficiency of the folding algorithm could be enhanced for proteins having crosslinks. The paper concludes with a discussion of these results and their implications for the solution of the protein folding problem.

METHOD

Since the details of the model and force field are described in the previous paper,[1] here only a short summary is given. In order to eliminate the possibility of target bias of the MCD simulations neither protein A, ROP dimer, nor crambin was included in the database used in the derivation of the statistical potentials. The folding experiments were performed according to the following protocol:

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.

1. For each folding experiment, an initial conformation (an expanded random coil on the coarser lattice[1,4]) is generated using either a separate algorithm or the folding MCD algorithm run at very high temperature.

2. Preliminary simulations are done to obtain a crude estimate of the folding temperature.

3. Employing temperature annealing, simulations of folding on the coarser lattice are performed for various starting conformations. In those cases where the system tends to be trapped in various (nonunique and therefore presumably misfolded) long-living intermediates, an alternative strategy may be used. First, one runs the system at a constant temperature, somewhat above the renaturation temperature. This way the model system samples a broad range of conformational space, frequently visiting various regions corresponding to partly assembled folds. During this run, the statistics on local conformational preferences are collected. This *predicted secondary structure information* can be then used as a supplementary contribution to the short range interactions, thereby shifting the folding mechanism toward a diffusion-collision mode of assembly.

4. The results of a set of independent, identical length runs are analyzed. All structures having the same topology and secondary structure are grouped together to form a family. The minimum and average energies of each group are compared, along with the rms deviation of all members of the family. The nonreproducible folds of higher energy are dismissed as misfolds. In the case of multiple families having different topologies, the lowest energy members of all topologies are subject to further refinement.

5. For a given family, the refinement procedure involves the projection of a set of the lowest energy coarser lattice folds onto the finer lattice. After a short equilibration period that removes some incompatibilities between the two lattices, long simulations are performed. For the situation when multiple families are found, the family with the lowest energy and smallest mean rms between members is identified as the putative native fold.

RESULTS

Folding of Protein A

Recently, a very accurate three-dimensional structure of the 60 residue fragment comprising the B domain of protein A in solution has been determined[6] from NMR. The fold has a three helix bundle topology. Helices II (residues 25 to 37) and III (residues 42 to 55) adopt an antiparallel hairpin conformation, with helix I (residues 10 to 19) crossing the C-terminal hairpin at an angle of about 30°. The remainder of the sequence has a poorly defined conformation. This simple structure should provide a good test of the folding algorithm.

For this protein, 45 folding simulations on the coarser lattice have been performed. The simulated thermal annealing procedure has been used, scanning a rather broad range of temperatures. Successful folding to a long lived three helix bundle is observed in 2/3 (30) of the trajectories. The remaining, unsuccessful runs can be divided into two categories. The first consists of those runs in which irregular collapsed structures are obtained. Their energy is about 50 to 60 k_BT larger than in correctly folded structures. These misfolded states can also be disregarded due to lack of reproducibility; each is different. Other times, the three helix bundle forms, dissolves, and reforms but does not survive to the end of the simulation run. Protein A is a small protein with only 29 native side group contacts. Consequently, folding may occur over quite a broad range of temperature. In order to have a high folding efficiency, one should simulate the model system dynamics over long times and at temperatures close to the lower limit of the transition range. This, however, requires much longer runs, due to the large number of local free energy minima which are present. The model system can easily spend a large fraction of time in metastable states. We opted here for fast folding, even at the expense of a substantial fraction (1/3) of unsuccessful experiments.

In spite of the above flaw, the annealing experiment rather clearly shows that the model protein folds to a three helix bundle, with very well-defined secondary structure. However, there remains the problem of topology. In 19 independent folding simulations, three helix bundles with the correct topology were obtained. In the remaining 11 simulations, the incorrect topology assembled, with the N-terminal helix on the other side of the C-terminal hairpin. The alternative topology may reflect inadequacies in the model, or may be due to the fact that a relatively minor reorientation of the C-terminal hairpin can accommodate the N-terminal helix. In addition, the protein A fragment is part of a series of four such units that interact with each other[6]; thus, the ambiguity in interaction with the N-terminal helix may be to some extent physical. The average conformational energy of the correct folds is about -181 k_BT, and the minimum energy is -225 k_BT. In contrast, the average energy of the incorrect folds is -153 k_BT, with a minimum observed value of -198 k_BT. The numbers refer to the same temperature and the same scaling factors for the various terms of the potential.[1] Thus, on the basis of energetic considerations, we conclude that the native three helix bundle topology is correctly chosen. The reproducibility of the nonnative three helix bundle topology is rather low. The rms between Cα traces of the incorrectly folded bundles is 4.3 Å, with contact map overlap in the range of 33%. This has to be compared with much better defined native folds of the model protein described below.

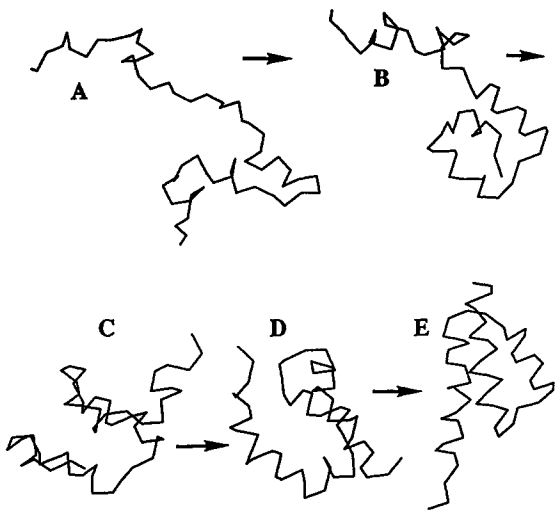

Fig. 1. Representative folding pathway of protein A. The entire folding process takes place on the finer lattice.

Folding of the native state typically proceeds by the on-site assembly of helices (in many, but not all cases, it is the N-terminal hairpin), followed by formation of the final helix. A folding pathway is schematically illustrated in Figure 1, where representative snapshots of the Cα trace from a single MCD trajectory are shown. The structures obtained at the end of successful folding simulations have many of the characteristics of a molten globule. There is much, if not all, of the secondary structure but poorly defined tertiary contacts. In other words, successful assembly of the native fold topology does not conclude the folding process. Structural fixation accompanied by the formation of a cross molecular pattern of tertiary contacts is the rate-determining step in folding.

Long isothermal runs at a temperature just below the transition temperature facilitate the process of collective adjustment of the side group packing in the protein core. In Figure 2, the upper triangle shows a representative contact map in a simulation of 3,000,000 time steps. Black indicates that the contact was present during the entire simulation and decreasing shades of gray indicate that the contact was present for a smaller fraction of the total simulation time. In the lower triangle, we present the first dissolution times of side group–side group contacts. A black square indicates that the contact lived the entire simulation time, and decreasing shades of gray indicate shorter lifetimes. In agreement with experiment, this clearly demonstrates that the folded state of protein A is predicted to be native like, with essentially fixed side chain contact.

Next, we selected for further refinement the five, lowest energy, native-like folds obtained from independent runs on the coarser lattice. These structures were projected onto the finer lattice, and for each of them, at least 3 independent refinement simulations were performed. After a very short relaxation of minor artifacts of the lattice to lattice projection, the finer lattice systems adopted well-defined folds. The lowest energy conformation seen in each simulation was extracted; their average, $\langle E_{min} \rangle$, was -209 k_BT, with a standard deviation of 8 k_BT. Starting from a given parent structure, the average Cα rms of the lowest energy states within a given family is 2.4 Å. A total of 27 refined structures were generated with an average rms for all 351 unique pairs of 3.1 Å. There are 19 refined folds for which the minimum conformational energy is below $\langle E_{min} \rangle$. The average rms between pairs of these structures is 2.83 Å. The average fraction of the same pairwise contacts seen in independent refinement runs ranges from 40 to 75%. This level of agreement between independently folded and then refined structures, indicates the precision (or reproducibility) of the model simulations.

The accuracy of the prediction is on the same level. In all cases, residues 13–19, residues 25–37, and residues 42–55 are predicted to be helical. Residues 1–9 and 56–60 lack any specific structure. In many cases, residues 10–12 were helical as well. Depending on the particular run, the helices may extend slightly beyond these helical regions; but in other cases, these regions assume extended conformations. Thus, the level of agreement of the secondary structure with experiment is excellent.

The 19 finer lattice refinement runs described above have an average Cα trace rms of 3.3 Å from residues 13–55 in the native structure. In a given run, the rms relative to native ranges from 2.55 to 3.42 Å, with a standard deviation ranging from 0.2 to 0.3 Å. Thus, the precision of the model is slightly better than its accuracy. Subsquent refinements on the finer lattice at low temperature produced structures averaging 2.25 Å rms from native. Figure 3A shows superimposed residues 10–55 of four lattice structures, obtained in independent simulations. Figure 3B shows the predicted Cα trace in green superimposed on the experimentally determined native backbone conformation shown as a purple ribbon. The rms between these structures is 2.25 Å.

Folding of ROP Monomer

The native state of ROP dimer consists of two helical hairpins, each 59 residues long, arranged in the form of a four helix bundle.[7] The helix-to-helix pattern of side group packing within this exceptionally long helix bundle resembles that seen in known examples of two chain, coiled coils. Therefore, the ROP dimer fold may be considered as an example of a supercoiled, coiled coil. This interesting protein has been redesigned into monomeric form by reengineering the loop connections between helices II and III, and between helices III and IV. The sequence of

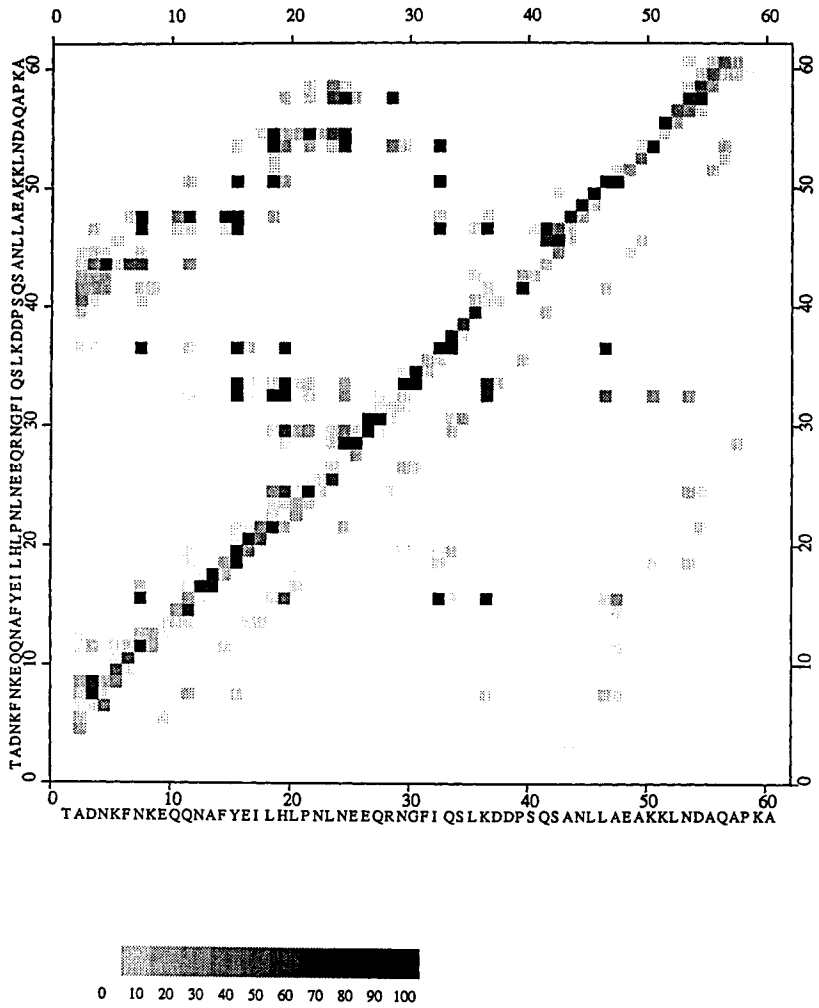

Fig. 2. Typical contact map for the protein A fold (above the diagonal) and the first contact dissolution time (below the diagonal). In the upper part of the diagram, the black dots correspond to the contacts seen in more than 90% of the trajectory snapshots, and the various shades of gray represent less frequently seen contacts. In the lower part of the diagram, the black dots correspond to those contacts whose lifetime is longer than the simulation time. The various shades of gray reflect average, first dissolution lifetimes of less stable pairwise contacts.

mROP composed of 120 amino acids is given below.[8] The underlined symbols represent substituted, or inserted residues.

```
123456789 123456789 123456789
MTKQEKTALNMARFIRSQTLTLLEKLNELD   helix I
ADEQADICESLHDHADELYRCSLASFKKPG   helix II
DIDEQADICESLHDHADELYRSCLARFGGS   helix III
KQEKTALNMARFIRSQTLTLLEKLNELAKG   helix IV
```

The solution to the crystal structure of ROP monomer (mROP) has not yet been published; however, it was designed to adopt the topology of a left turning, four helix bundle similar to the fold of ROP dimer. Therefore, the simulations presented here have the character of a tertiary structure prediction.

Lattice folding and refinement

Initial simulations of the Monte Carlo dynamics of mROP indicated a very cooperative transition to the globular state, with a strong propensity for helical conformations. When the model system is simulated at relatively high temperatures, safely above the collapse temperature, one may extract the helix probability profile along the sequence. The results are shown as the solid line in Figure 4. Clearly, the sequence has a strong preference to adopt a helical conformation. The central engineered loop breaks the helical pattern. The two remaining putative turns are less visible; their location has to be induced by tertiary interactions that are too weak at this temperature. This particular experiment indicates the possibility of applying the proposed reduced models to secondary structure prediction. This extension of the model will be explored elsewhere.[9]

Due to the highly cooperative helix–coil transition of mROP, the folding simulation has to be performed over a relatively narrow temperature range. This

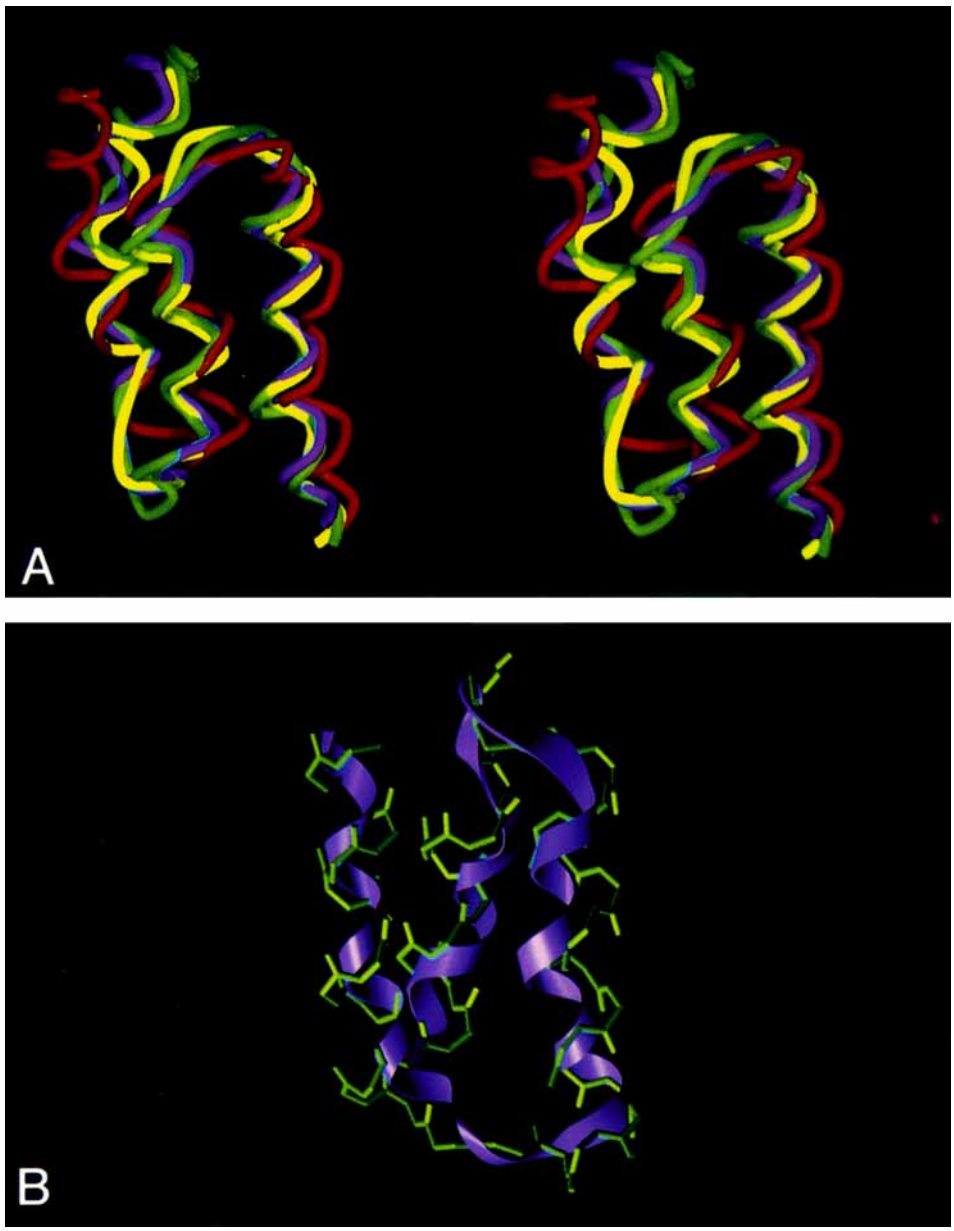

Fig. 3. (**A**) Superimposed α-carbon traces, shown as tubes of various colors, for high resolution, finer lattice folds of protein A. (**B**) The predicted Cα trace in green superimposed on the experimentally determined native backbone conformation, shown as a purple ribbon.

avoids the trapping of the model system in quenched, misfolded states. The transition temperature was estimated by measurement of the helix content in thermal annealing, MCD simulations over a relatively broad range of temperature. The proper folding simulations are performed using much slower "cooling." A typical change in the temperature during the folding process was in the range of 5%.

A representative folding trajectory on the coarser lattice is shown in Figure 5, where particular snapshots present Cα traces of various early intermediates seen during the folding process. A portion of the central helical hairpin usually serves as an initiation site for mROP folding. This is the first, very early, intermediate seen in most simulations. A partially or completely folded hairpin, most frequently consisting of helices II and III, dissolves and again forms several times during a typical run. Well-defined initiation of folding in either of the two other hairpin turns is very rare. The next early intermediate consists of a three helix bundle with the fourth

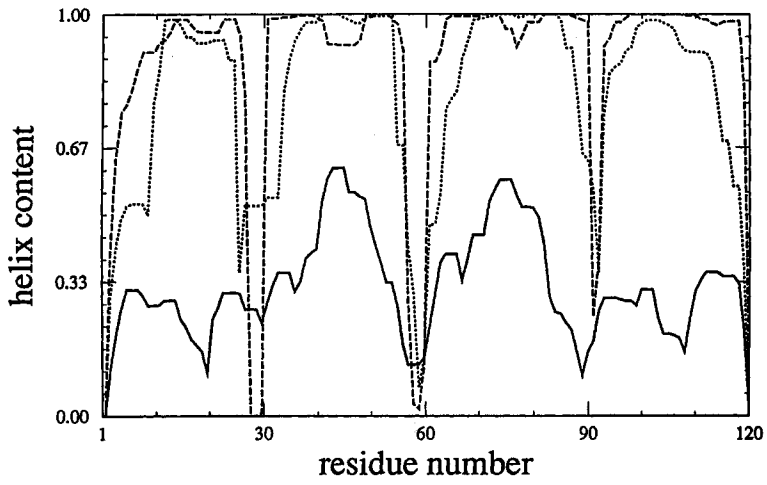

Fig. 4. Average helicity (averaged over a long run) as a function of position in the ROP monomer sequence in the denatured state as $T=2.5$ (solid line), the molten globule state at $T=2.25$ (dotted line), and the native state at $T=1.5$ (dashed line). See the text for a more detailed description of particular simulations.

helix rather poorly defined. The last helix assembles predominantly by an on-site mechanism. In most cases, the N-terminal helix assembles last; however, the statistics are too poor to draw any stronger conclusions. Several times, the four helix bundle dissolved (and assembled again). When the lowest energy states were subject to deeper quenching (with about a 10% change in temperature), the resulting folds remained stable.

Twelve long folding runs on the coarser lattice were undertaken and produced compact globular states. In 11 cases, the four helix bundle adopted the left turning topology, as expected for this engineered sequence. The only stable misfolded state had the right turning bundle topology. No other long lived, collapsed states were observed. This higher reproducibility of the overall topology of the mROP fold in comparison with protein A folds is probably related to the much stronger propensities for very regular secondary structure. The existence of a more cooperative transition within a narrow temperature range avoids grossly misfolded compact conformations. The right turning fold has a higher average conformational energy, (equal to $-429\ k_BT$) when compared with the average energy of left turning bundles (equal to $-463\ k_BT$) at the same temperature. Consequently, the simulations predict that the left turning topology is the correct fold.

The precision of the coarser lattice folds is less than in the case of protein A. First, the rms deviation from the equivalent positions of α-carbons in the known dimeric structure (the monomer is expected to adopt a rather similar structure) varies between 4 and 5.5 Å. This is mostly related to fluctuating errors in the helix-to-helix registration for one or two pairs of helices. The rms deviation between lattice structures obtained in different runs is

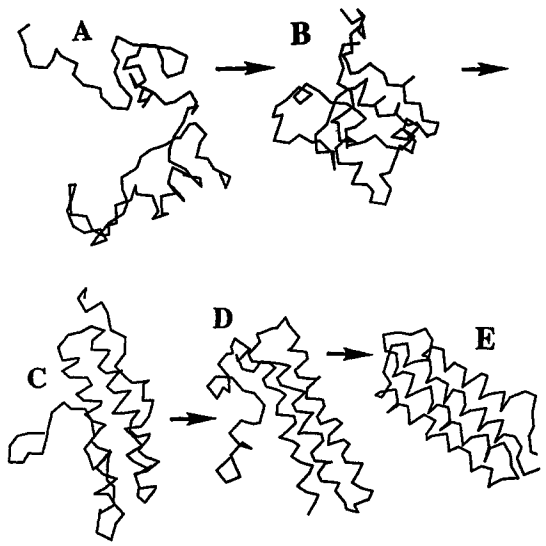

Fig. 5. Representative folding pathway of topology assembly for ROP monomer on the coarser lattice. (**A**) The initial expanded random coil state. (**B**) At a somewhat lower temperature, a frequently seen, more compact, but still expanded state, with a small amount of secondary structure. (**C**) A typical, very early folding intermediate consisting of the central helical hairpin. (**D**) Three helix intermediate with the N-terminal portion of the chain attempting "on site" assembly of the fourth helix. (**E**) Final, "native" Cα trace of the left turning, four helix bundle.

marginally smaller and varies between 4 and 5 Å. The packing of the side chains is not unique; the fraction of overlapping side chain contacts for independent folds is in the range of 25%. Moreover, only a small fraction of side group contacts is long lived. Consequently, the core of the coarser lattice model protein is to a large extent liquid like; there is no well-defined transition to the state having a fixed

pattern of side chain packing. In other words, the transition from the molten globule-like to native-like state of mROP is difficult to achieve on the coarser lattice. It is also possible that the simulation time is too short to achieve side chain fixation in a molecule of this size.

The five lowest energy folds were then projected onto the finer lattice, and the minor inconsistencies resulting from the projection were relaxed. The refinement runs were performed at constant temperature, slightly below the transition temperature, using a minimizing procedure which works as follows. At the end of every MCD cycle (the time unit of the model dynamics), the total energy is computed in order to store the Cα trace of the (approximately) lowest energy state from the beginning of the MCD run. Each simulation consisted of 20 subruns (an arbitrary number). At the beginning of each subrun, the system was restarted from the Cα trace corresponding to the lowest energy state; however, the rotamers were randomized. This procedure seems to be very effective in searching for the global minimum of the conformational free energy, provided that the system is already close to this minimum. For the present model, this free energy minimum corresponds to the well-defined Cα trace of the fold. Of course, this procedure could be used for folding from an expanded denatured state; however, here, the length of the single subrun should be about an order of magnitude longer. Otherwise, the system could oscillate around a deep local minimum of the free energy surface. The refinement runs lead to very well-defined folds with a fixed pattern of side chain packing. This can be illustrated by comparison of the instantaneous side chain contact map with the plot of the first dissolution time for these contacts, taken from subsequent long MCD runs. A typical example is given in Figure 6, where the same convention as in Figure 2 is used. There is clearly a well-defined subset of contacts which forms a native like packing pattern. Unlike the case of protein A, here the very short "absence" of a contact may remain undetected. In this context, it is noteworthy that the relaxation of thermalized rotamers must be very fast, at least for the essential contacts which form the network across the bundle. Otherwise, these contacts would not survive an appreciable fraction of, much less the entire, simulation time. These simulations once again indicate that side chain relaxation with respect to a given backbone is fast,[4,10,11] and that the very slow chain fixation from molten globule like states is a very long collective process entailing small adjustments of the backbone conformation and side chains. Certainly, to a large extent, the almost instantaneous relaxation of the side chains for native like conformations is an artifact of our reduced model (however, MD simulations show that side chain relaxations are rather fast,[11] with rearrangements occurring on the nanosecond time scale). On the other hand, this is perhaps the feature that allows the folding process to occur in our model in a reasonable amount of computer time.

On the finer lattice, the precision of the refined structures improves considerably. First, the fraction of binary contacts recovered in independent runs is in the range of 45 to 55%. The average rms between Cα traces of independent, finer lattice folds drops to 3.20 Å, with the smallest value equal to 1.63 Å, and the largest equal to 4.20 Å for the most distant pair. The refined folds are on average closer to the ROP dimer structure, with the Cα rms ranging from 4.06 to 4.80 Å. Thus, the "average" model structure of mROP does not coincide very well with the equivalent portions of the ROP dimer structure. The lowest energy folds were then subjected to very low temperature quenching. The quenched structures assumed rms values in the range of 3.6–4.2 Å from equivalent α-carbons of ROP dimer. The rms deviation of individual helices varied between 1.0 and 2.4 Å, depending on the run and their position in the model bundle. What is interesting is that the central helical hairpin usually has a much lower rms with respect to the two corresponding helices of the ROP dimer; in some simulations, it is below 2 Å. Also, the model side chain packing for these two helices is better. This may reflect physical reality, which could be confirmed with the forthcoming solution of the mROP crystal structure. The substantially lower rms of particular helices when compared to the higher rms for the entire bundle seems to indicate that the short-range interactions and the geometric representation of the model protein backbone (some helices have an rms close to the limit of the finer lattice resolution) are better modeled than the long-range interactions and the side group packing. Consequently, the errors in helix-to-helix packing could be considered as the major reason for the limited precision (1.6–4.2 Å) of mROP folding. These errors lead to small shifts of registration of the side chain packing, as well as to small, but nevertheless noticeable, differences in the twist of the entire bundle seen in different simulations. To rectify this, an improvement of the long-range interactions will be attempted in the near future. However, taken at face value, the simulations predict that the helices in the mROP structure should be less supertwisted than in the ROP dimer structure.

Molten globule versus native state of the ROP model

It has been previously[4] demonstrated that the coarser lattice model can reproduce to a large extent the basic properties of both the molten globule state[12] and the native state, including the latter's much better side chain packing. In fact, the simulation[4] of two helical proteins designed by DeGrado and co-workers is in accord with experiment.[2,3,13] For the sequence which was designed to have a

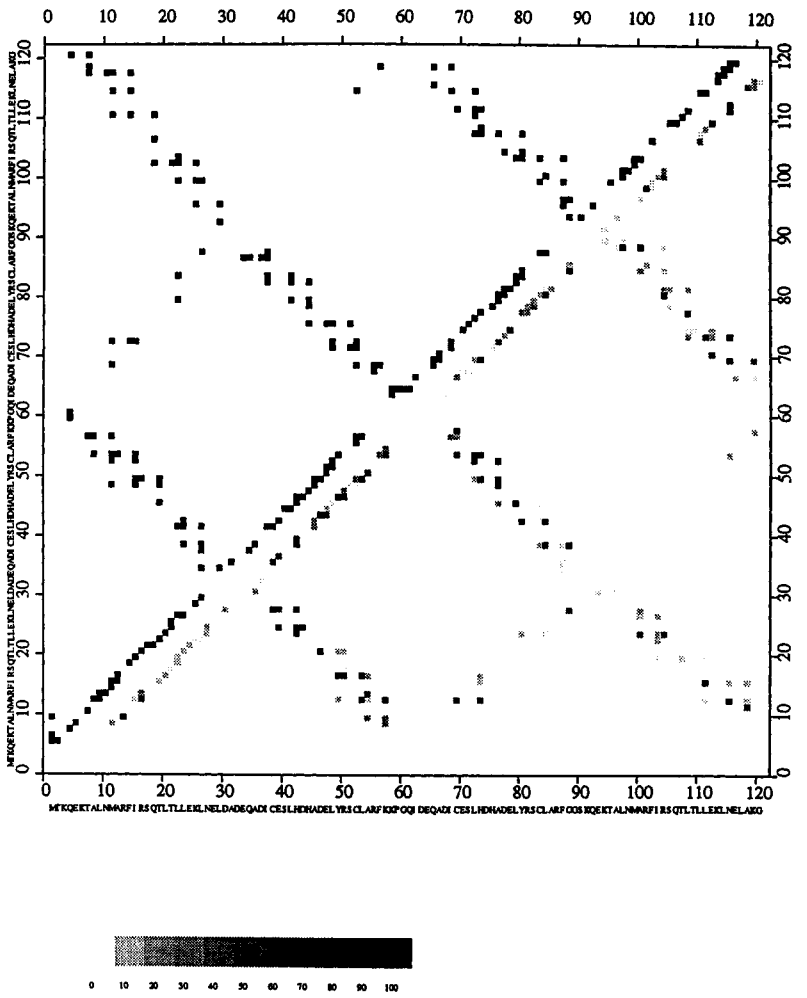

Fig. 6. Typical contact map obtained from the last snapshot of a trajectory for the ROP monomer fold (above the diagonal) and an illustration of the first dissolution time for side group contacts (below the diagonal). The black dots in the latter case correspond to those contacts whose lifetimes are longer than the simulation time, which is equal to 9×10^5 MCD time steps. The various shades of gray reflect lifetimes of less stable pairwise contacts.

strongly hydrophobic core, both simulation[4] and experiment[2] indicate a thermodynamically very stable fold; however, there is a liquid like, nonspecific packing of the hydrophobic core. The redesigned sequence[3] has been predicted[4] to adopt a very specific packing of the hydrophobic side chains.

In this context and in the case of an "almost natural" protein, mROP, it is interesting to examine how the model depicts the late, presumably molten globule,[14] intermediates, and the native state in the finer lattice model. For this purpose, a long simulation on the finer lattice is performed at a temperature where the folded topology is marginally stable. The previously described minimizing procedure has been applied in order to avoid complete unfolding in the case of a large random conformational fluctuation. At the same temperature, because of the all-or-none character of the folding transition, one may perform very long simulations of completely denatured states. The very long molten globule simulation run employing the "minimizer procedure" may be considered as a series of 20 shorter, but still long, runs starting from different conformations of the side chains, with Cα traces in the broad basin near the correct fold. The first important observations is that within a particular subrun there is no relaxation to a different state, which would be characterized by differing energies or by the dynamics and identity of contacts. The states at the beginning of particular subruns are on average of the same energy and the same rms as any other states seen in the simulation. Consequently, in spite of use of the "minimizer" protocol, a pseudo-equilibrium metastable state is simulated, with marginal if any bias.

This metastable state has properties that commonly are considered to be a signature of the molten

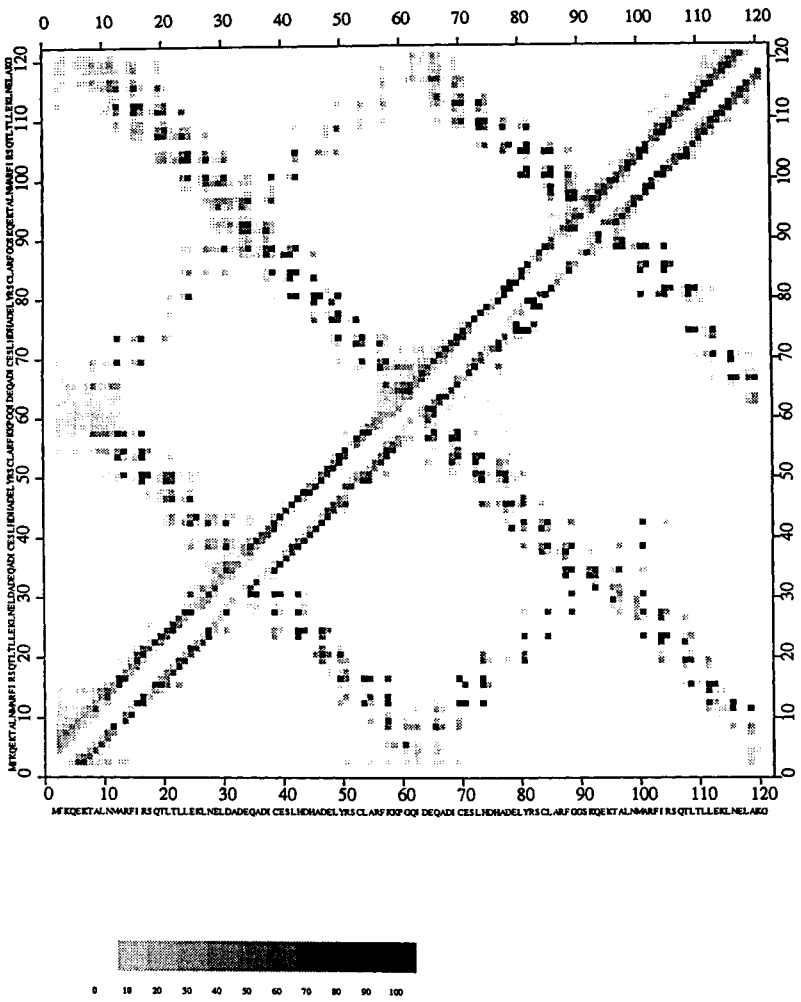

Fig. 7. Average frequency of occurrence (in various shades of gray) of pairwise contacts in the native state of ROP monomer (below the diagonal) and for the molten globule state (above the diagonal).

globule state.[12,14] The overall fold is the same as that observed at lower temperatures. The secondary structure is essentially the same as in the native fold, but there are larger fluctuations near the loops and the ends of the chain. This leads to a small decrease in the average helix content relative to the native state. This is illustrated in Figure 4 by the two upper curves for the helix probability profiles in the native and molten globule states respectfully. However, the molten globule molecule is swollen—the radius of gyration, as measured for the Cα trace, is noticeably larger, by about 5%, than for model native folds. This corresponds to a 15–20% increase of volume. Interesting, the average number of side chain contacts is close to that in the native state, although the entire structure is more mobile. In particular, the side chain contact pattern is very unstable; however, when it is time-averaged, the contact pattern is almost the same as that seen in the native state. This similarity is illustrated in Figure 7, which employs the same conventions as in Figure 2. In spite of this time average similarity, the collective dynamic properties are completely different. When the first dissolution time is analyzed for the molten globule-like state, no single contact survives the entire run. This stands in marked contrast to the substantial fraction of such contacts in the native-like state which survive (see Fig. 6). The model molten globule, in agreement with experimental work and theoretical interpretations,[12,14,15] has no fixed network of pairwise side chain contacts that are typical of the native state.

A comparison of the numerical values of characteristic properties of the model system in the denatured, molten globule, and quenched native state is given in Table I for the finer lattice model. Therefore, the nature of the observed molten globule state is the same in both the coarser and finer lattice description. Rather, the differences between the molten globule-like state and native-like state are bet-

TABLE I. Properties of ROP Monomer in the Random Molten Globule and Native States*

	T	Energy (k_BT)	Radius of gyration (Å)	Helix content (%)	Contact lifetime (MC units)
Denatured state	2.5	−80.	33.2	33.	Below 10
Denatured state	2.25	−122.	25.6	45.	Below 100
Molten globule	2.25	−279.	15.5	81.	~100
Native state	1.75	−375.	14.9	88.	$\sim 10^6$
Native state	1.50	−502.	14.8	89.	Above 10^6

*The average (over long MCD runs) values of the reduced energy (Energy), root mean square radius of gyration of the Cα trace, helix content, and lifetime of pairwise contacts (contact lifetime), versus temperature (T). The underlying cubic lattice mesh size is 1.22 Å.

ter defined for the finer lattice model. On the other hand, it is unlikely that the molten globule-native state transition could be observed in a model employing a cruder discretization than our coarser lattice. The very low temperature, $T = 1.5$, (temperature is dimensionless since energy is expressed in k_BT units) was selected for refinement simulations in order to achieve a unique structure. However, side chain fixation, or the transition from the molten globule state to the native state, was observed at much higher temperatures, up to $T=2.0$, i.e., close to the temperature of the molten globule ($T=2.25$) or the denatured state isothermal simulations. Thus, the folding transition temperature for our model potential should be located between 2.0 and 2.25. At $T=2.0$, one can still observe a well-defined native like state, while at $T=2.25$ the model system has properties typical of a molten globule. The more exact estimation of the thermodynamic characteristics of the transition (the energy change, heat capacity curve, etc.) would require much longer and far more expensive simulations.

Dynamic properties of the mROP model

Due to the small distance elemental moves employed, the model of dynamics employed in the present MCD simulations appears to be quite realistic. In theoretical statistical physics studies of polymeric systems, much simpler models of MC lattice dynamics have been successfully applied to the examination of very complex relaxation processes in entangled multichain systems.[16–18] Moreover, even the folding pathways of simple model polypeptides are the same regardless of lattice MCD or off-lattice dynamics employed.[19] Thus, it seems worthwhile to examine the global dynamics of our protein model. The dynamics of the denatured state is the simplest. Not surprisingly, at high temperatures, the dynamics has all the general features of Rouse dynamics—the dynamics of flexible polymer chains with negligible hydrodynamic interactions. Identical behavior was previously observed in simpler polypeptide models.[20,21]

More interesting are the dynamics of the molten globule and the native states of the model mROP protein, examined here for the first time. Their dynamic properties are illustrated in Figure 8, where various autocorrelation functions are plotted on a log-log scale. The Cα average mean square displacements, $g(t)$, are plotted in the curves denoted by solid and open circles for the molten globule and native states, respectively. $g(t)$ is defined as follows:

$$g(t) = \Sigma<[\mathbf{r}_i(t + t_0) - \mathbf{r}_i(t_0)]^2>/ n \qquad (1)$$

with $\mathbf{r}_i(t)$ the Cartesian coordinate vector of the ith Cα vertex at time t, the brackets denote the average over the trajectory, and the summation is over the entire chain. Similarly, the motion of the center of mass of the model molecule can be measured. The mean square displacement of the center of mass, $g_{cm}(t)$ is plotted in the curves denoted by the boxes. The solid (open) symbols correspond to the molten globule (native state) dynamics. The solid straight line indicates the free diffusion limit when $g(t) \sim t$, the dashed lines have a slope of 1/2, and the dotted line has a slope of 1/4. Although the molten globule state has a marginally larger radius of gyration than the native state, the molten globule-like state (at $T=2.25$) has a dramatically larger mobility. As a matter of fact, the molecule undergoes slow translational diffusion, as indicated by the observed behavior of the center of gravity autocorrelation function in the long time limit (here, it is observed for $t>10^4$). Over the same time range, corresponding to center of gravity displacement on the order of a few Angstroms, the Cα beads move as a Rouse chain,[17,22] with $g(t)\sim t^{1/2}$. This is yet another demonstration of the liquid-like character of the model system at this temperature; the interactions between helices do not superimpose sufficiently strong restraints. The main restraints felt by the chain are related to covalent bonds.

A completely different situation is observed for the native state at $T=1.7$. The overall dynamics is limited to short distance relaxation, and since there is no rigid body translation elemental moves of the entire structure in the MCD model, both curves ap-

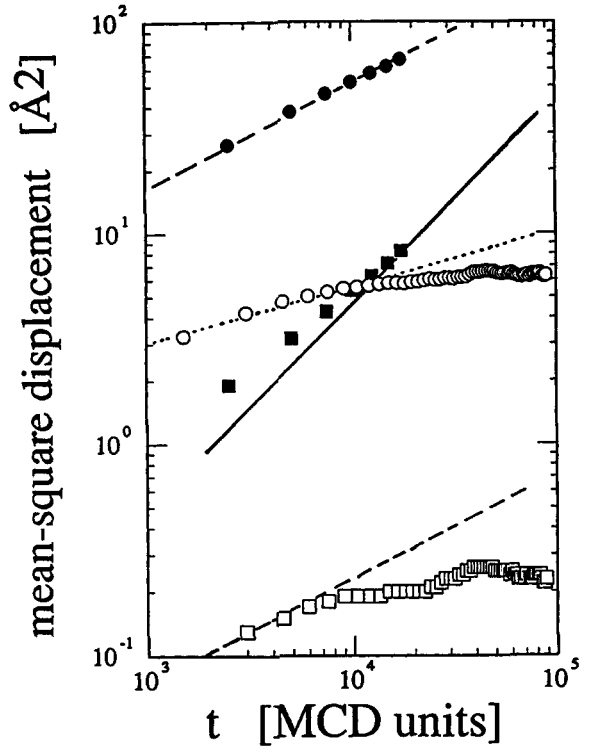

Fig. 8. Log–log plots of the average Cα mean square displacement versus time (circles), and the mean square displacement of the center of mass (boxes). The open symbols are for the "native" state of the model system, and the solid symbols correspond to motion seen for the molten globule state. The lines given for comparison correspond to free diffusion (solid line), the diffusion of a single segment of a Rouse chain at distances smaller than the radius of gyration of the model chain (dashed lines), and diffusion of a Rouse segment through random channels formed by strong constraints (dotted line).

proach a plateau. However, at short times, there seems to be a well-defined region where the single segment autocorrelation function is proportional to $t^{1/4}$ and the center of mass moves according to the 1/2 exponent. Analogous to the situation seen in highly entangled polymer systems,[17,19] this is indicative of a gel-like network; here, it results from the slowly relaxing tertiary interactions. In conclusion, the dynamic properties of the molten globular state differ dramatically from those of the native like state and are even more striking than the differences seen in the structural properties. Of course, due to the reduced character of our model and the use of Monte Carlo dynamics, the time scales of various relaxation processes may be considerably distorted. Nevertheless, the qualitative trends should reflect those seen in the long time dynamics of the real protein.

Folding of Crambin

In recent work,[4,10] several designed helical proteins had been folded using the present (coarser lattice) reduced model. In this paper, two natural sequences that fold into helical motifs have also been examined. An obvious concern is whether the model is biased toward highly regular, helical structures. To address this point, we should also examine α/β proteins and β-proteins. Here, we present the result of crambin (1crn). This is a 46-residue protein with a rather unique fold composed of a small helical hairpin and three extended chains arranged into a minimal antiparallel β-sheet. Three pairs of cystine cross links make the fold very stable, despite the small size of the protein. It has been assumed in these simulations that in the native state of crambin all cysteines are in form of cystines. However, no specific pattern of crosslinks has been assumed. As in the case of mROP and protein A, crambin has been excluded from the database used for the derivation of the potentials.

The pairwise potential describing side group interactions distinguishes between Cys–Cys interactions, Cyx–Cyx bonds, and the Cys–Cyx interaction. From the statistics of the database, the Cys–Cyx interaction is strongly repulsive. Consequently, the folding of the very hydrophobic crambin sequence also simulates the equilibrium between the various oxidation states of pairs of cysteines and cystines. In order to achieve the correct native state, the six cysteines in the crambin sequence have to adopt the proper pattern of the covalent bonds. This pattern is not encoded into the input data; rather the MCD algorithm searches the various possibilities. The overall efficiency of the folding protocol (the one that has been used in the two cases described above) for crambin is rather low; only a fraction (1/10) of the folding simulation yields the proper native like state, having the generally correct secondary structure, the correct network of S–S bonds, and well-defined side chain packing. The others are misfolded conformations. Misfolded in the case of crambin means that the topology of the fold is correct; however, the secondary structure is highly distorted. These toplogical isomers are 4.3–5.5 Å from native and are not unique. In other words, the straightforward folding simulations of crambin in most cases led to various misfolded structures that correspond to deep local minima on the model system free energy landscape. Perhaps by generating a large number of folded and misfolded conformations, one can select the proper one based on an energy comparison. An additional tool which under certain circumstances might allow selection of the proper native-like state might be provided by an inverse folding algorithm.[10,23] Unfortunately, our inverse folding algorithm does not match the crambin sequence to its native structure. Thus, it is of marginal utility here.

Here, we present a somewhat different approach. First, let us note that in a total of about 20 long runs, we did not observe any grossly misfolded states with the wrong topology. The conclusion from

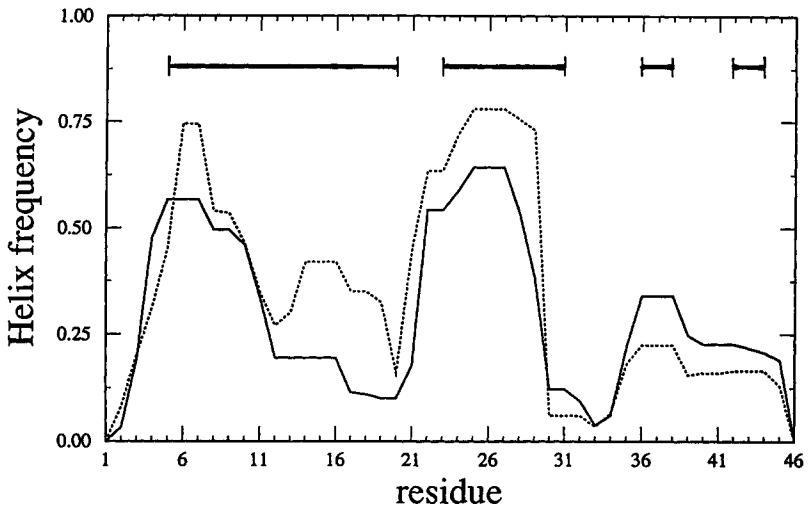

Fig. 9. The helicity (right-hand turning compact conformation of three consecutive backbone vectors) profiles of crambin extracted from two independent MCD simulations at two temperatures close to the renaturation temperature. The solid horizontal lines indicate the native helical and turn regions.

this observation is that the system is locked in long-lived metastable states, where the rather rigid S–S bonds slow down both the development of proper side chain packing and the formation of secondary structure. The model system has to wait a long time for dissociation of a disulfide bond in order to open a channel for local rearrangement. This suggests the following update to the folding protocol. First, the system is simulated at a constant temperature where the S–S bond dissociation rate is still sufficiently high. At this temperature, the model crambin samples a wide range of partly folded states. We then collect statistics about the secondary structure preferences (only helical propensities were extracted, or more precisely, two kinds of secondary structure propensities were extracted: helix/turn and extended/loop). The helicity profiles ("helical" residues occur when three consecutive backbone vectors form a well-defined right-handed turn corresponding to the highest peak in the distribution of $r^2_{i,i+3}$ shown in Figure 4 of ref. 1) from these isothermal, prescreening runs very clearly indicate two helical fragments in the crambin structure. The N-terminal helix (residues 7–17) is predicted with a shift by two to three residues toward the N terminus, and the second helix (residues 23–30) is predicted more accurately. The short helical fragment (residues 42–44) at the C-terminus is hardly visible. The results are shown in Figure 9, where helicity profiles extracted from two independent runs above the folding temperature are compared with the Kabsch–Sander[24] assignment of the secondary structure of crambin. It is worth noting that standard secondary structure prediction schemes (the Neural Network Model of Holley and Karplus,[25] the GOR method,[26] and the combined neural network and statistical methods expert systems[27]) fail to predict any helices in crambin. Instead, β-sheets are anticipated by most of these methods; however, somewhat more successful is the Chou and Fasman method[28] that predicts a portion of the middle helix. As mentioned in paper 1,[1,9] the crambin example illustrates how the proposed method can be used as a secondary structure predictor.

Having an approximate prediction of the helical regions, the folding algorithm is supplemented by an additional small energetic bias towards helical conformations, whose strength for a particular residue is proportional to its helicity at higher temperature. Because the elements of secondary structure (helices) form at higher temperature, there is a larger chance that the S–S bonds are formed after the proper supersecondary structure emerges. Consequently, the problem of long lived intermediates having out of register S–S bonds can to a large extent be eliminated. Extended fragments are always more mobile in our model and the proper minimal β-sheet fragment of crambin always assembles after the helical hairpin. The above strategy improves the folding efficiency. However, about 50% of the folds have a more or less distorted secondary structure, mostly in the putative helical regions. These poorly defined folds can be dismissed because their conformational energy is about 20% higher than the family of the best, native like folds. In Figure 10A, an example of the model crambin fold is shown superimposed on the native state. The reproducibility of the folding procedure is demonstrated in Figure 10B, where the best superpositions of the three Cα traces obtained from independent runs are presented.

These simulations of crambin prove that while the

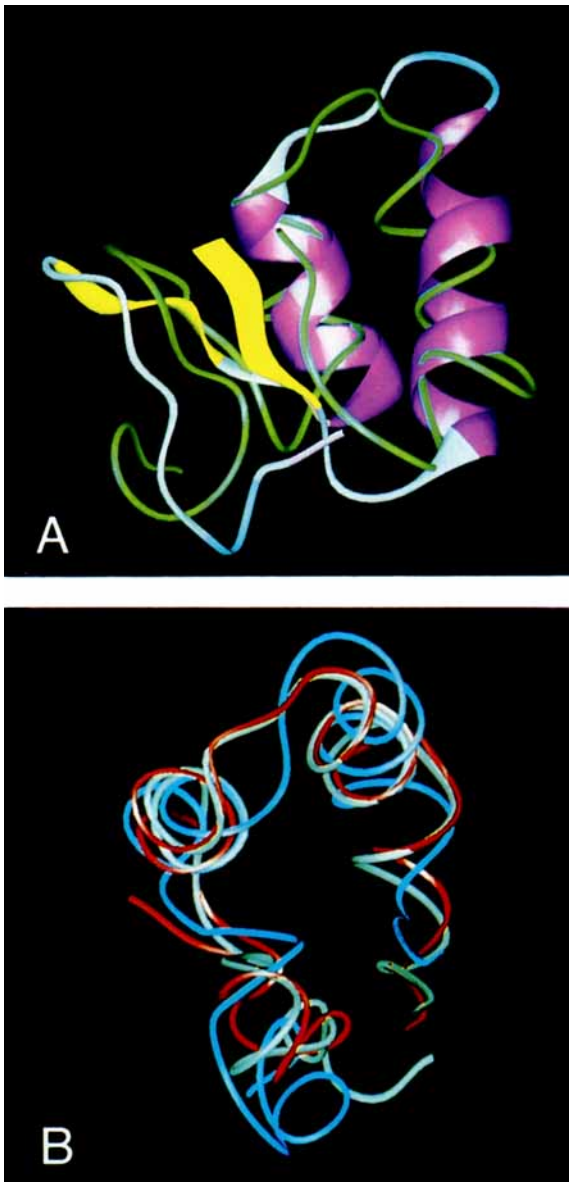

Fig. 10. The Cα trace of the model fold of crambin (green) superimposed on the native state in purple and white (**A**), and three independent folds at the best mutual superposition (**B**).

present methodology is still far from being an automatic three-dimensional structure predictor, in addition to some helical proteins, a very different class can also be successfully folded. Moderately accurate structures of crambin (~4.0 Å rms from native) can be obtained in about 50% of the simulations. A subset of three of these independently obtained structures was refined in very long runs and has an average rms below 4.0 Å. The average coordinate rms for residues 3–42 over long runs is 3.61 Å, with the smallest coordinate rms equal to 3.18 Å. (for residues 1–46, the smallest rms is 3.76 Å), and a distance rms of 2.6 Å from the native Cα-trace. This demonstrates the rather accurate (except for the minimal C terminal helix) assembly of the secondary structure elements. The pattern of S–S bond is also exactly predicted. Compared to the PDB structure, the distances between Cαs 3–40, 4–32, and 16–26 (S–S bonds) are reproduced with an accuracy of 1.1 ± 0.2 Å.

CONCLUSION

These simulations show that at least for three moderate size globular proteins having simple topological motifs, the protein folding problem can be solved in its most general sense. That is, one can simulate the entire folding process, starting from an expanded random coil state and ending up with three-dimensional structures having reasonable resolution. The predicted secondary structure, tertiary interactions, packing, and dynamic properties are much closer to the native state of real proteins than had been achieved in earlier reduced models. In fact, the structures obtained from the present lattice simulations are sufficiently accurate to permit a full atom reconstruction of the native structure.[11] The procedures starts from a very fast analytical reconstruction of the backbone and Cβ atoms.[29] Then, the remainder of the side groups can be inserted. The resulting full atom structures require relatively minor local readjustments[11] using restrained molecular dynamics simulations, similar to those used for NMR refinement.[30] Consequently, there are no major contradictions between the geometry of the reduced models and the full atom model obtained from a detailed force field.

The folding process seems to proceed along quite well-defined folding pathways. Some very early intermediates are seen more frequently than others. Later stage, partly folded intermediates are even better defined. Finally, in every simulation, the longest lived intermediates have the properties of a molten globule. The observed pathways are essentially independent of the small variation in the folding protocol associated with annealing versus isothermal folding simulations. While the general picture of the folding pathways seen in the model systems seems to be in agreement with experiment and other theoretical work,[12,14,15] it is too early for more detailed conclusions. Once more sequences have been folded, the correspondence of the model folding pathways to experiment (in vitro) will become better established. Additional experimental data for small proteins that undergo reversible thermal denaturation are also required.

As was previously discussed,[1] the force field used here, while of sufficient accuracy for use in folding simulations of small, simple proteins, is perhaps not capable of the de novo folding of larger and more complex proteins. The refinement of the various, mostly statistical, potentials should be possible due to the growing number of solved, high-resolution, 3D structures of globular proteins.[31] Other contri-

butions to the proposed force field will be reexamined as well.[1] In conclusion, while much work remains to be done, the possibility of at least a limited solution to the protein folding problem seems to have been demonstrated, and the directions of future developments seem to be well defined.

ACKNOWLEDGMENTS

Valuable discussions with Drs. William Beers, Charles L. Brooks, III, and Adam Godzik are gratefully acknowledged. We thank Michal Vieth for helpful assistance in preparation of some of the figures. 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. I. Lattice model and interaction scheme. Proteins 18:338–352, 1994.
2. 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.
3. 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.
4. 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.
5. Godzik, A., Kolinski, A., Skolnick, J. Lattice representations of globular proteins: How good are they? J. Comp. Chem. 14:1194–1202, 1993.
6. Gouda, H., Torigoe, H., Saito, A., Sato, M., Arata, Y., Schimada, I. Three-dimensional solution structure of the B-domain of staphylococcal protein A. Comparison of the solution and crystal structures. Biochemistry 40:9665–9672, 1992.
7. Banner, D.W., Kokkinidis, M., Tsernoglou, D. Structure of the Col*El Rop protein at 1.7 Angstroms resolution. J. Mol. Biol. 196:657–675, 1987.
8. Sander, C. private communication, 1993.
9. Rey, A., Kolinski, A., Skolnick, J. Application of a discretized protein model to secondary structure prediction, in preparation.
10. 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.
11. 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.
12. Kuwajima, K. The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. Proteins 6:87–103, 1989.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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.
19. Rey, A., Skolnick, J. Comparison of lattice Monte Carlo dynamics and Brownian dynamics folding pathways of α-helical hairpins. Chem. Phys. 158:199–220, 1991.
20. 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.
21. Kolinski, A., Skolnick, J. Discretized model of proteins. I. Monte Carlo study of cooperativity in homopolypeptides. J. Chem. Phys. 97:9412–9426, 1992.
22. de Gennes, P.G. "Scaling Concepts in Polymer Physics." Ithaca, NY: Cornell University Press, 1979.
23. Godzik, A., Skolnick, J., Kolinski, A. A topology fingerprint approach to the inverse folding problem. J. Mol. Biol. 227:227–238, 1992.
24. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
25. Holley, L.H., Karplus, M. Protein secondary structure prediction with a neural network. Proc. Natl. Acad. Sci. U.S.A. 86:152–156, 1989.
26. Garnier, J., Ousguthorpe, D.J., Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120:97–120, 1978.
27. Zhang, X., Mesirov, J.P., Waltz, D.L. Hybrid system for protein secondary structure prediction. J. Mol. Biol. 225: 1049–1063, 1992.
28. Chou, P.Y., Fasman, G.D. Prediction of protein secondary structure. Adv. Enzymol. 47:45–148, 1978.
29. Rey, A., Skolnick, J. Efficient algorithm for the reconstructions of a protein backbone from the α-carbon coordinates. J. Compt. Chem. 13:443–456, 1992.
30. Brunger, A.T., Clore, G.M., Gronenborn, A.M., Karplus, M. Three dimensional structure of proteins determined by molecular dynamics with interproton distance restraints. Proc. Natl. Acad. Sci. U.S.A. 83:3801–3805, 1986.
31. PDB Quarterly Newsletter, No. 63, January 1993.