A method for predicting protein structure from sequence

Jeffrey Skolnick*, Andrzej Kolinski*†, Charles L. Brooks III*‡, Adam Godzik* and Antonio Rey§

*Department of Molecular Biology, The Scripps Research Institute, 10666 N. Torrey Pines Road, La Jolla, CA 92037, USA, †Department of Chemistry, University of Warsaw, 02-093 Warsaw, Poland, ‡Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA 15213, USA and §Departmento de Quimica Fisica, Facultad de Ciencias Quimicas, Universidad Complutense de Madrid, Madrid, Spain.

Background: The ability to predict the native conformation of a globular protein from its amino-acid sequence is an important unsolved problem of molecular biology. We have previously reported a method in which reduced representations of proteins are folded on a lattice by Monte Carlo simulation, using statistically-derived potentials. When applied to sequences designed to fold into four-helix bundles, this method generated predicted conformations closely resembling the real ones.
Results: We now report a hierarchical approach to protein-structure prediction, in which two cycles of the above-mentioned lattice method (the second on a finer lattice) are followed by a full-atom molecular dynamics simulation. The end product of the simulations is thus a full-atom representation of the predicted structure. The application of this procedure to the 60 residue, B domain of staphylococcal protein A predicts a three-helix bundle with a backbone root mean square (rms) deviation of 2.25–3 Å from the experimentally determined structure. Further application to a designed, 120 residue monomeric protein, mROP, based on the dimeric ROP protein of *Escherichia coli*, predicts a left turning, four-helix bundle native state. Although the ultimate assessment of the quality of this prediction awaits the experimental determination of the mROP structure, a comparison of this structure with the set of equivalent residues in the ROP dimer crystal structure indicates that they have a rms deviation of approximately 3.6–4.2 Å.
Conclusion: Thus, for a set of helical proteins that have simple native topologies, the native folds of the proteins can be predicted with reasonable accuracy from their sequences alone. Our approach suggest a direction for future work addressing the protein-folding problem.

Current Biology 1993, 3:414–423

Background

One of the most important unsolved problems in contemporary molecular biology is deceptively easy to state — given the linear sequence of amino acids that comprise a given globular protein, predict the three dimensional structure of the biologically active conformation. The solution to this problem has proved elusive for a number of reasons. It requires a free energy function that can differentiate the native state from the misfolded conformations. This is a non-trivial problem, but substantial progress has been made recently by using empirically determined free energy functions [1–5]. One then has to solve the multiple free energy minima problem [6]. Recognizing that the free energy functions are likely to be approximate, criteria for assessing the success of the procedure are required. This method should only require the amino-acid sequence as input, with no information whatsoever about the folded conformation being provided; in other words, the method should fold the protein *de novo*. It should be able to reproduce known structures at an acceptable level of spatial resolution (3–4 Å rms deviation, or better, from the actual structure). The compatibility of the resulting predicted structures with their sequences, as assessed by inverse protein folding algorithms that address which sequences are compatible with a given structure, should be comparable to that of the experimentally determined native structures. Ultimately, it should be able to predict hitherto unsolved protein structures. In this paper, we describe a hierarchical approach to the protein-folding problem, which satisfies the above criteria for a number of proteins having simple native state topologies.

The basic idea of our hierarchical approach is to employ a reduced model that will fold a given sequence to produce an approximately correct native structure, and then subsequently to build the full heavy atom description. The use of a reduced model is motivated by practical considerations. Obviously, a reduced description decreases the number of degrees of freedom. More importantly, it permits a discretized (lattice) description of the backbone, which uses two orders of magnitude less computer time than the corresponding non-lattice model. In addition, discretization tends to smooth the free energy surface, thereby making configurational sampling more efficient. The use of reduced models to describe the geometry of, and interactions in, globular proteins has a long history [7–12]. They generally yield a manifold of compact states that, depending on the local interactions that are introduced, may or may not have any secondary structure. The resulting manifold of collapsed states is very liquid-like, with very non-specific tertiary contact patterns [8–12]. This contrasts with real proteins, the native states of

Correspondence to: Jeffrey Skolnick

which have side-chain contact patterns that are specific and well defined.

To eliminate this fundamental flaw and to reproduce the cooperative transition from the molten globule intermediate state, where substantial regions of native-like secondary structure occur but where side-chain packing is non-specific [13–14], to the native state, in the model we introduce cooperative side-chain–side-chain packing interactions [15]. These cooperative terms are generic in the sense that the patterns of contacts that are favored can occur in α, β, or mixed α/β proteins, and allow native-like patterns to emerge when appropriate. Our model has the following additional novel features. First, there is a cooperative hydrogen bonding scheme between α-carbons (Cαs)[15–16] that reproduces 90% of the hydrogen bonds in real native proteins, as defined by Kabsch and Sander [17]. Second, there is an amino-acid pair-specific potential that describes the angular correlation between side-chain centers of mass. This term is responsible for the predicted local conformational preferences [18]. The aforementioned set of phenomenological potentials allows the *de novo* folding of the sequences described below.

In what follows, we use our hierarchical method to predict the three-dimensional structures of the B domain of staphylococcal protein A [19–20], which adopts a three-helix bundle in solution [21], and the structure of mROP [22], which is an engineered version of the ROP dimer. The native state of the ROP dimer is a four-helix bundle comprising two helical hairpins, each containing 63 residues [23]. The modified 120 residue-long sequence of mROP is designed to adopt the left turning four-helix bundle geometry. For protein A, the predicted structures are on the level of 2.25–3 Å rms deviation from the known NMR solution structure for the backbone atoms. This marks the first time that structures of this quality have been predicted simply from their sequence, without recourse to any sequence or structural homology modeling techniques. In the case of mROP, the crystal structure is not yet solved, so we have compared the predicted mROP structure with the set of equivalent residues in the ROP dimer, and we find that the two differ by 3.6–4.2 Å rms deviation for the Cαs. The two structures differ mainly in their degree of supertwist, with the simulations predicting that less supertwist will be observed in mROP. This prediction awaits experimental verification.

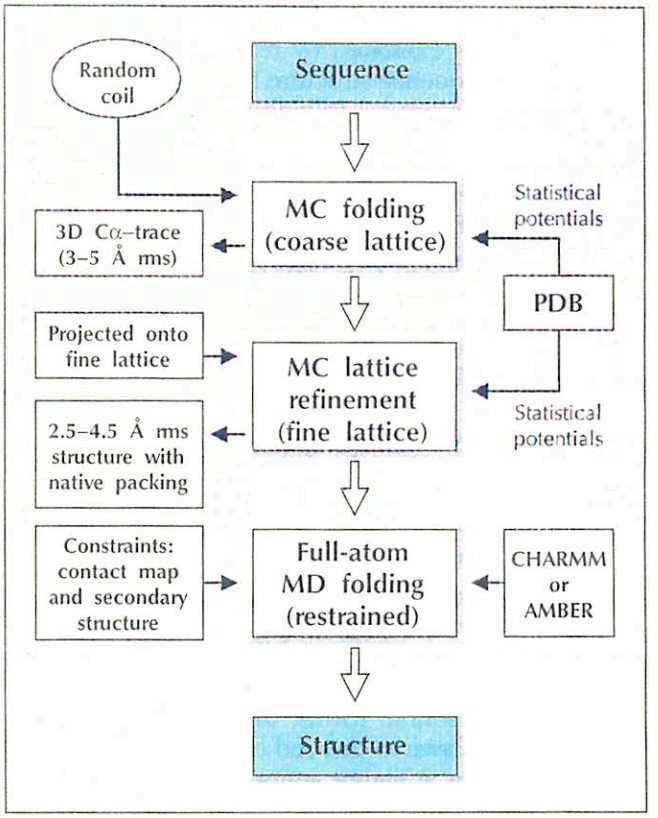

Fig. 1. Flow chart depicting the hierarchical approach. Folding commences on the coarse lattice using the protein sequence and a random chain conformation as input. The set of folded structures is then refined using the fine lattice to produce better packed, native structures. The system is then pulled off lattice, the virtual Cα bonds are regularized, an analytic procedure rebuilds the backbone heavy atoms, and tentative side-chain conformations are assigned based on the position of the side chains in the parent lattice structure. The lattice structure also provides the set of predicted side-chain contact pairs and secondary structure assignments (helix, turn, and so on) that serve as target functions for the full atom rebuilding procedure. The initial compact conformations are subjected to molecular dynamics (MD) simulations that produce a family of full-atom models. The average conformation after a series of ten such refinements is the final output of the *de novo* folding protocol.

Results and discussion

Description of the method

Our hierarchical approach starts with a reduced description of the unfolded protein and takes advantage of the greatly improved conformational sampling efficiency that is afforded by using reduced models to produce folded structures with Cα rms deviations in the range 3–5 Å from the native structure. In practice, the protein consists of a set of Cα backbone atoms that are restricted to a set of lattice points and a set of side chains that are not confined to the lattice. Each side chain is represented by a single ball located at the side-chain center of mass. As discussed in greater detail in Materials and methods, and as outlined in the flow diagram of Figure 1, Monte Carlo dynamics on a coarse lattice are used to generate the approximate fold [15–16,24]. Then, before building a complete atomic model, the structure is refined using a lattice with finer spacing between the lattice points. The resulting structures exhibit better side-chain packing and backbone geometry and provide a set of predicted secondary structure and side-chain contact constraints that serve as target functions in molecular dynamics simulations [25–26], which are designed to produce whole-atom models. Finally, to ascertain whether

the predicted structures exhibit the sequence-structure specificity that is exhibited by real protein structures, the quality of sequence-structure fitness is evaluated in the context of an inverse folding algorithm [4].

Folding simulations of Protein A

Recently, the solution structure of the 60-residue fragment comprising the B domain of protein A has been determined [21]. It adopts a three-helix bundle in solution, with the first helix (I) running from residues 10–19, the second helix (II) running from 25–37, and the third helix (III) running from 42–55, with the remainder of the sequence structurally undefined. Interestingly, helices II and III adopt an antiparallel hairpin conformation, whereas helix I crosses the carboxy-terminal hairpin at an angle of about 30°. This sequence should provide a good test of the folding algorithm. The protein is relatively small, the topology is simple and the structure is known.

Successful *de novo* folding to a long lived three-helix bundle is observed in about 2/3 (30) of all trajectories. Sometimes metastable, partially collapsed states about 50–60 k_BT higher in energy are obtained. Other times, the three-helix bundle forms, dissolves and reforms but it doesn't survive to the end of the simulation run. Much longer runs at higher temperatures might relieve this problem and increase the folding efficiency; this objective, however, has not yet been achieved.

Of the total of 30 successful, independent folding simulations, 19 folds have the correct topology and 11 the incorrect topology, where the amino-terminal helix is on the other side of the carboxy-terminal hairpin. The alternative topology may reflect any inadequacies in the model, or may be caused by the fact that a relatively minor reorientation of the carboxy-terminal hairpin can accommodate the amino-terminal helix on the other side of the hairpin. In addition, the protein A fragment is part of a series of four units that interact with each other [19]; thus, the ambiguity in interaction with the amino-terminal helix may be partly physical. The average energy of the correctly folded conformation is about -181 k_BT, and the minimum energy conformation found is -225 k_BT. In contrast, the average energy of the incorrectly folded conformation is -153 k_BT, with a minimum observed value of -198 k_BT. Thus, on the basis of energetic considerations, we conclude that the native three-helix bundle topology is correctly chosen. The rms deviation between conformations of the incorrect fold is 4.3 Å, with a contact overlap of roughly 33 %. This contrasts the results obtained for the native fold, which are described below.

Each simulation begins from a randomly generated initial state. Successful folding typically occurs by the on-site assembly of two helices (in many but not all cases, it is the amino-terminal hairpin), followed by formation of the final helix. The resulting state has many of the characteristics of a molten globule [14]. There is much, if not all, of the secondary structure, but the tertiary contacts are poorly-defined. Structural fixation, accompanied by the formation of a molecule-spanning pattern of tertiary contacts, is the rate-determining step in the folding procedure. In Figure 2, the upper triangle shows a representative contact map, in which contacts that are present for more than 75 % of a simulation time of 40 000 time steps are indicated by black dots. In the lower triangle, we present the first dissolution time of a contact. Black indicates that the contact lived the entire simulation time, and the various colors shown in the color bar indicate shorter contact lifetimes. A molecule-spanning collection of long lived side-chain contacts is clearly apparent. In agreement with experiment, this clearly demonstrates that the folded state of protein A is predicted to be native-like, with fixed side-chain contacts. It should be pointed out that side-chain fixation is not an obligatory feature of this model. In analogous studies of helical bundles, one of the sequences designed by DeGrado and Raleigh [27] is predicted to be a four-helix bundle topology but without side-chain fixation [15] (it is thus molten globule-like [14]). This agrees with experimental results on this system.

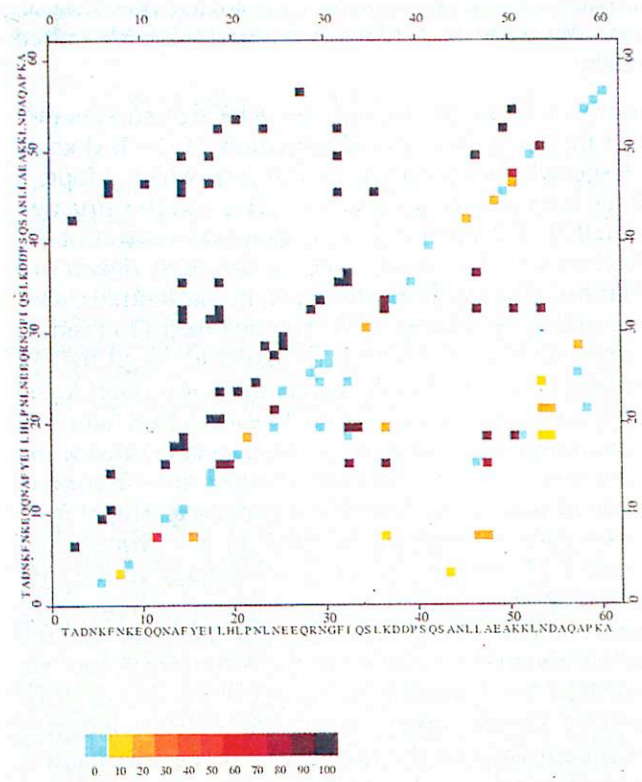

Fig. 2. For the B domain of protein A, the upper triangle presents a representative contact map showing those contacts present for at least 75 % of the simulation time of 40 000 MC time steps. The lower triangle depicts the first dissolution time of the contacts. A black square indicates that the contact survived over the entire simulation, and the various colors, shown in the color key, indicate shorter lifetimes. Note that for a given backbone conformation, the side chains relax with respect to the backbone in less than five time steps. Thus, side-chain fixation is demonstrated.

To characterize the folded state further, we performed a cluster analysis of 400 snapshots that were extracted from the above trajectory and grouped using a metric

that required 60% side-chain contact map overlap. Consistent with the idea that the predicted native conformation is well defined, a single family of structures is observed.

Of the 19 independent folds obtained on the coarse lattice, the five lowest energy parent structures were subjected to a series of at least three independent refinement simulations on the finer lattice. The lowest energy conformation seen in each simulation was extracted; their average , $<E_{min}>$, was -209 k_BT, with a standard deviation of 8 k_BT. (Note that there is a slightly different energy scale related to the different lattice representation.) Starting from a given parent structure, the average Cα rms deviation 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. Focusing on the refined structures with minimum energies at or below $<E_{min}>$, of which there are 19, the average rms deviations between pairs of structures is 2.83 Å. This level of agreement between independently refined structures is indicative of the model's precision.

Next, we turn to the accuracy of prediction. In all cases, residues 13–19, 25–37 and 42–55 are predicted to be helical. In many cases, residues 10–12 are helical as well. Residues 56–60 were unstructured. Depending on the particular run, the amino-terminal and carboxy-terminal helices may propagate slightly towards the ends; in other cases the ends may adopt an extended conformation. Thus, the level of agreement of the secondary structure with experiment is excellent. The 19 fine-lattice refinement runs described above have an average rms deviation of 3.3 Å from residues 13–55 in the native structure. In a given run, the rms deviation of the predicted structures relative to native ranges from 2.55 to 3.42 Å, with a standard deviation ranging from 0.2 –0.3 Å. Thus, the reproducibility of the model is slightly better than its accuracy. Subsequent refinements on the finer lattice at low temperature produced structures with an average 2.25 Å rms deviation from the native structure.

We then constructed an all-atom model from one of the lattice structures with Cα rms deviations of 3 Å from the NMR solution structure. To obtain an all-atom model, we subjected it to the rebuilding protocol described in the Materials and methods section. Using the set of predicted secondary structure and side-chain contact constraints provided by the lattice structure, the resulting all-atom models exhibited a backbone heavy atom rms deviation of 2.97 Å from the experimental structure. For the Cαs, this is the same rms deviation as achieved for the starting lattice structure and demonstrates that the rebuilt full-atom structures are at the same level of resolution as the lattice structures from which they are derived. In other words, the predicted lattice structures do not contradict full-atom models. This is a very important result, which is demonstrated here for the first time.

In Figure 3, we depict the various stages in the all-atom rebuilding procedure. Figure 3a shows residues 10–55 of the lattice structure, from which the constraints were derived, the remainder of the sequence being unstructured both in simulation and in solution. Figure 3b shows the conformation after the structure has been removed from the lattice and the backbone heavy atoms inserted. Figure 3c shows the superimposed

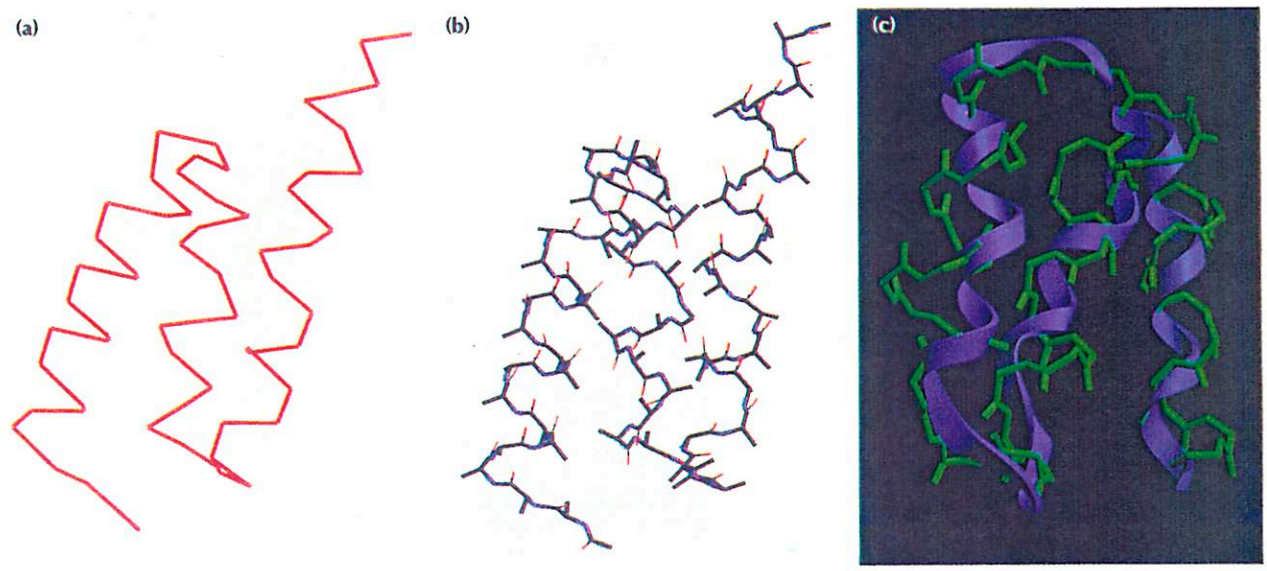

Fig. 3. The various stages of the all-atom rebuilding procedure for residues 10–55 of protein A: **(a)** lattice backbone Cα conformation obtained at the end of a refinement run; **(b)** the resulting conformation after it has been removed from the lattice, the virtual Cα bonds regularized, and all the other backbone heavy atoms inserted; **(c)** the backbone heavy atom trace of the predicted structure obtained after MD refinement, shown in green, superimposed onto a ribbon model of the NMR solution structure, shown in purple.

backbone atom trace of residues 10–55 in the NMR structure and the predicted structure. The amino-terminal helix crosses the carboxy-terminal helical hairpin at an angle of roughly 30°. The 3 Å rms deviation difference between structures arises from the helices being offset by a third to a half of a helical turn. This undoubtedly reflects the limitations of the lattice model, mostly from treating side chains in the single ball representation, and the inadequacies in the potential. To assess the quality of the predicted structures, the lattice structures were subjected to a recently developed inverse folding algorithm, which contains a related, but not identical, subset of the interactions used in the lattice models [28]. The lattice structure has an energy of $-14.9\ k_BT$ compared to the native value of $-19.6\ k_BT$. Thus, the model is of good quality.

Folding simulations of mROP

mROP, a 120 residue polypeptide, is a redesigned version of the dimeric *E. coli* protein 1ROP [22]. The ROP dimer adopts a four-helix bundle native state [23]. Exceptionally long helices and the packing pattern of the hydrophobic core make this fold somewhat similar to a bundle formed by two supercoiled, coiled coils. The sequence of mROP, which apparently also folds into a similar, left turning, four-helix bundle is the following:

MTKQEKTALNMARFIRSQTLTLLEKLNELDADEQADIC-ESLHDHADELYRCSLASFKKPGQIDEQADICESLHDHA-DELYRSCLARFGGSKQEKTALNMARFIRSQTLTLLEKLN-ELAKG.

The folding simulations described below represent a *de novo* structural prediction for this engineered protein. Representative snapshots of the folding trajectory on the coarse lattice are given in Figure 4a–d. In all runs, the system rapidly developed an early intermediate that usually consisted of a helix hairpin, in most cases comprising the central helices. These elements of supersecondary structure dissolved and reformed several times during a run. The better defined, and much longer lasting, intermediate consisted of three helices, with the fourth helix many times partially folding, mostly by an on-site mechanism [7]. Eventually, the fourth helix assembled and a four-helix bundle finally developed. In a total of 12 independent folding simulations, 11 of these resulted in left turning bundles, and once a right turning four-helix bundle formed. In contrast to protein A, the model system never, except in the single case of the right turning bundle, adopted long-lived misfolded topologies. At temperatures about 10 % below the transition temperature, these bundles remain stable. The average conformational energy of the right turning bundle, $-429\ k_BT$, was higher than the energy of the left turning bundles, $-463\ k_BT$.

The resulting folded structures have been projected onto the finer lattice and subjected to long isothermal runs. In all cases, the system developed a rather well defined side-chain packing pattern, with subsets of contacts that never dissolved during very long runs. The left and right turning bundles have an average energy of $-415\ k_BT$ and $-370\ k_BT$, respectively. A typical situation is shown in Figure 5, where the final contact map and the first dissolution time are given for comparison. The overlap of the contact maps from independent runs is in the range of 45–55 %, reflecting reasonably consistent side-chain packing. Although the rms deviation from the dimer-based native structure is in the range 3.6–4.2 Å, the rms deviation distance for Cα traces between refined folds varies from 2.7–4.3 Å. How close the native monomer structure is to the dimer structure remains to be established, but it may turn out that the structure predicted here is closer to the monomer than to the dimer. Finally, in Figure 4e, we show the predicted full-atom model of mROP; the rms deviation of the Cαs between four independently generated full-atom models is about 3.0 Å from the lattice structure, and they have, on average, a 1.7 Å rms deviation from each other. Thus, for mROP the reconstruction protocol yields a rather well defined final structure. Finally, the Cα rms deviation in the full-atom structure is from 4.4–4.8 Å from the ROP dimer; this

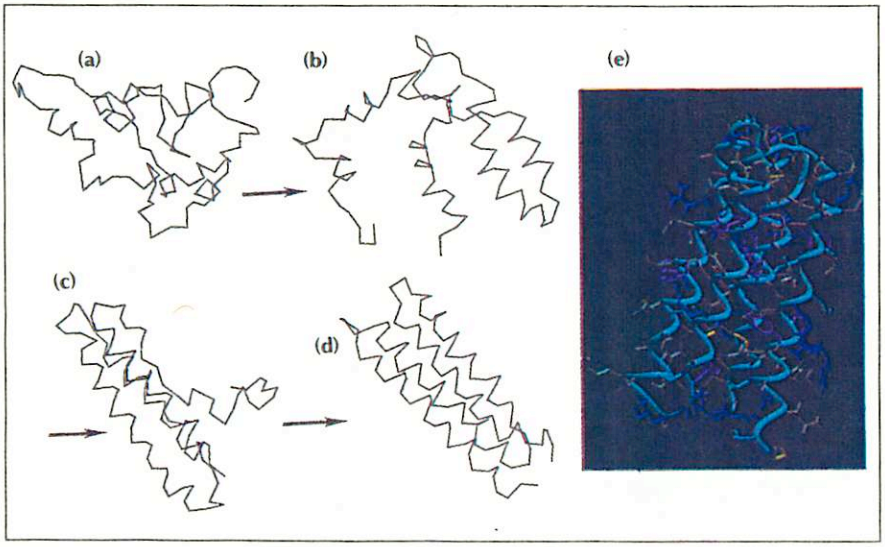

Fig. 4. Representative snapshots along the predicted folding trajectory of the mROP: **(a)** denatured state, initial conformation; **(b)** formation of the central helical hairpin; **(c)** formation of the three-helix bundle intermediate; **(d)** formation of the four-helix bundle topology; **(e)** full-atom structure obtained from the MD refinement procedure.

again indicates that the lattice structures are compatible with full-atom models.

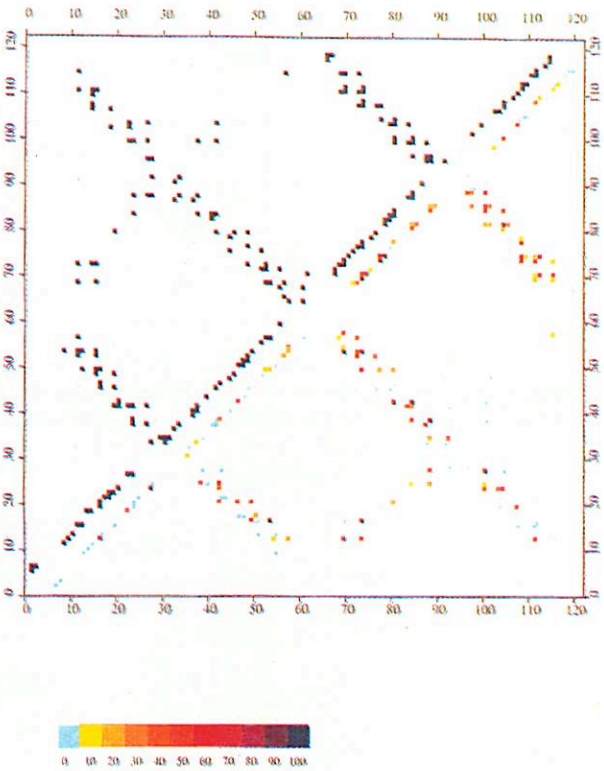

Fig. 5. For mROP, the upper triangle presents a representative contact map obtained at the end of a refinement trajectory of 100 000 Monte Carlo steps on the finer lattice. The lower triangle depicts the first dissolution time of the contacts. A black square indicates that the contact survived over the entire simulation, and the various colors shown in the color key indicate shorter lifetimes. Thus, side-chain fixation is demonstrated.

Conclusions

This paper has described an application of a general methodology for folding proteins using only the amino-acid sequence. For a small, three-helix bundle protein of 60 residues, structures are predicted that have a 2.2–3.0 Å rms deviation from the NMR solution structure. For a redesigned, monomeric version of the ROP dimer, the present simulations predict a structure that is about 3.6–4.2 Å from the set of equivalent residues in ROP. The two structures differ mainly in the degree of supertwist that they exhibit, with the monomer predicted to be less supertwisted than the dimer. This series of simulations marks the first time that protein structures of this quality have been predicted by an algorithm that starts from a random chain conformation, with no extrinsic information about the native conformation, and finishes with a full heavy atom model. Moreover, the refinement protocol indicates that lattice structures are compatible with full atom, continuous space conformations. That is, there are no major contradictions between discretized and continuous space representations of proteins. For sequences designed by DeGrado [27] and others [28,29], this methodology has also been proved to be capable of independently predicting a number of experimental observations [15]. The most salient feature for the DeGrado sequences is the way in which a molten globule equilibrium intermediate can develop the more rigid structure characteristic of native proteins by sequence mutation.

Our current approach extends previous attempts to predict protein structure in the following important respects. First, this is a hierarchical approach that combines both coarse and fine resolution lattices to generate the folded conformation from the denatured state. Second, the methodology predicts local conformational preferences based on the angular correlation of side-chain centers of mass. These sequence-specific terms act as the trigger for the formation of secondary structure in the contex of sequence-independent cooperative hydrogen bonds and a generalized Ramachandran potential that describes the average local distribution of protein backbone distances and chain chirality. Finally, by introducing cooperative side-chain packing templates, we incorporate the fact that side chains in globular proteins exhibit specific packing patterns.

Although this methodology works for a number of proteins having simple topologies with short turns between secondary structure elements, it is uncertain how broad the class of proteins is that can be treated. It is clear that the protein folding problem is not yet solved. At this point, improvements in the potential function to better account for the correlated side-chain packing preferences are required. It is our belief that the set of side-chain contact templates giving rise to the possibility of side-chain fixation that we now use are incomplete and insufficiently cooperative, and that the current hydrogen-bond scheme is too permissive. Work to eliminate these deficiencies in the model are in progress. Nevertheless, even as it stands, our approach has considerable promise for the *de novo* design of proteins as well as for the prediction of tertiary structure in small proteins. For a number of proteins, these simulations demonstrate that starting from sequence information it is possible to predict structures that bear a close resemblance to observed native folds. Although much work is required to solve the full protein folding problem, the progress made to date provides encouraging results that imply that such a solution is possible.

Materials and methods

Overview of the protein folding protocol

A flow chart depicting the entire procedure is presented in Figure 1. We begin with randomly generated, unfolded conformations of the protein restricted to a coarse lattice that is described in further detail below [15–16,24]. The molecule consists of the on-lattice set of backbone α-carbons plus an off lattice set of side-chain rotamers; each side chain is represented by a single ball located at the side-chain's center of mass. The time course of the folding process is simulated using Monte Carlo

dynamics (MCD); the algorithm permits the relative motion of all secondary and supersecondary elements, thereby avoiding artificial kinetic trapping of partly folded structures. A detailed discussion of the move set may be found elsewhere [15]. After a set of preliminary runs designed to identify the transition temperature, a series of folding simulations are run for fixed time intervals. These typically consist of 50 000–100 000 Monte Carlo time steps, and the final conformation at the end of the simulation is saved. The simulations are run typically over a narrow temperature range (or they may be isothermal). A set of the lowest energy conformations is refined on a finer lattice, which has a more accurate geometric description of the α-carbon backbone and better side-chain packing. Then, all the heavy atoms of the backbone and side chains are rebuilt [30]. This comprises the initial conformation for the full atom relaxation protocol. Here, the predicted lattice structures provide secondary structure and side-chain contact constraints that act as target functions in molecular dynamics simulations using either the CHARMM [25] or AMBER [26] force fields. To test the quality of the predicted native conformation, sequence structure fitness is evaluated using our inverse folding algorithm [4].

Lattice model

Geometrical considerations

The underlying spatial grid is taken to be a cubic lattice. For the coarser lattice used for folding from the denatured state, consecutive α-carbons are joined by vectors chosen from cyclic permutations of vectors of the type (± 1, ± 1, ± 1), (±2, ± 1, 0) and (±2, ± 1, ± 1), where the length in lattice units between adjacent cubic lattice sites is unity (based on the average Cα virtual bond lengths, the distance between cubic lattice points corresponds to 1.7 Å). On the finer lattice used for refinement, consecutive α-carbons are joined by vectors chosen from the set (± 2, ±2, 0), (± 2, ± 2, ± 1), (±3, 0, 0), (± 3, ± 1, 0) and (± 3, ± 1, ± 1); here, the distance between neighboring sites on the underlying cubic lattice corresponds to 1.22 Å. The distance of closest approach of all non-bonded α-carbons on the coarse and fine lattices are 4.78 and 3.45 Å, respectively. Unlike many other lattices used to describe proteins, these do not suffer from any significant orientational biases that could arise from underlying lattice symmetries [24].

Interactions

Both the coarse and fine hybrid lattices have an equivalent set of interaction terms that are adjusted for the metric of the underlying cubic lattice grid. For all proteins discussed in this paper, the identical parameter set has been employed and is available from the authors by E-mail [18]. The various contributions to the energy are: (1) a local Ramachandran-like backbone potential; (2) a hydrogen-bond potential; (3) a local amino-acid pair specific potential reflecting the orientational correlations of the center of mass vectors of the first to the fourth neighbors down the chain; (4) a side-chain rotamer energy; (5) a one body amino-acid specific potential that reflects the preference for that side-chain center of mass to be located at a given distance from the center of mass of the protein; (6) amino-acid pair specific tertiary interaction potentials; and (7) cooperative side-chain packing templates. Each of these terms is discussed in greater detail below.

Sequence independent contributions

On both the fine and coarse lattices, the local distribution of Cα–Cα distances and chain chirality is rather close to the mean distribution in structures from the Brookhaven Protein Data Bank (PDB) [31]. A small energetic correction allows this distribution to be recovered and helps to insure that generic protein-like states are sampled. This contribution plays the role of an average Ramachandran plot. In the native states of protein A and mROP, this term typically contributes about 16% of the total energy. This term and all other potentials of mean force, with the exception of the hydrogen-bond energy, are constructed by examining the statistics of occurrence of the relevant parameter in a library of high resolution PDB structures [18]. This structural library does not contain protein A, ROP nor any other protein subject to *de novo* structure prediction.

An empirical hydrogen bond potential was developed for all pairs of residues $|i - j| > 3$. Two residues are considered to be hydrogen bonded if the distance between α-carbons, r_{ij}, is smaller than 7.2 Å and the following geometrical requirements are satisfied:

$$|(b_i - b_{i+1}) \cdot r_{ij}|^{1/2} \leq 4.16\,\text{Å}$$
$$|(b_j - b_{j+1}) \cdot r_{ij}|^{1/2} \leq 4.16\,\text{Å}$$

where the b_k are the backbone vectors. Only main chain–main chain hydrogen bonds are permitted. At most, every residue, with the exception of proline, can participate in two such interactions. Proline is restricted to having only one hydrogen bond. This definition of hydrogen bonds allows for the proper pattern of hydrogen bonds to form within helices or across β sheets. Furthermore, the interaction is cooperative; consecutive pairs of hydrogen bonds down the chain experience additional stabilization. That is, the total hydrogen-bond energy is of the form

$$E_{H-bond} = \sum\sum E^{H}\delta(i,j) + \sum\sum E^{HH}\delta(i,j)\delta(i \pm 1, j \pm 1)$$

where E^{H} and E^{HH} are the energy of a single hydrogen bond and the excess energy when two consecutive hydrogen bonds are formed, when i and j are the residues of interest, and where $\delta(i,j)=1$ when the geometric criteria (see Equation 1) for the formation of a hydrogen bond between the ith and jth α-carbons are satisfied. The approximate magnitudes of E^{H} and E^{HH} were determined by performing a series of studies on polyalanine and polyvaline model systems [16]. We then fine-tuned these parameters by requiring that in the denatured state transitions between helix and non-helical states can readily occur, and that the unfolded state helix content is low (<20% for helical sequences such as protein A and mROP). In reality, a range of values can be chosen, but typical values of E^{H} and E^{HH} are $-1.25\ k_BT$ and $-1.5\ k_BT$ respectively. k_B is Boltzmann's constant. For the predicted native states of protein A and ROP, $E_{H\text{-}bond}$ contributes to roughly 25% and 36% of the total energy respectively.

Amino-acid pair specific short range interactions

To account for local, amino-acid pair specific conformational preferences, a free energy term of strength E_{hm}, which is based on the relative angular orientation of the side-chain centers of mass, was derived for the first to the fourth neighbors down the chain. In these models, we do not specify the secondary structure that the sequence will adopt. Rather, it arises from the interplay of the amino-acid pair specific preferences embodied in E_{hm} and the tertiary interactions. In particular,

$$E_{hm} = \sum_{\text{residues, i}} \sum_{k=1}^{4} \epsilon_{hm}\left[\cos\left(\theta_{i,i+k}\right)\right]$$

where $\Theta_{i,i+k}$ denotes the angle between the vectors pointing from the ith and i+kth α-carbons to the centers of mass of their corresponding side groups. ϵ_{hm} is a statistical potential derived from the structural database that is amino-acid pair specific and

is a function of k. A histogram has been constructed by dividing $\cos(\Theta_{i,i+k})$ into ten bins and comparing the observed number of cases in the 56 member structural database to the expected value assuming a uniform distribution in $\Theta_{i,i+k}$. The negative logarithm of this ratio is the energy ε_{hm} for that bin. In the predicted native conformations of protein A and mROP, this term represents 19% and 14% of the total energy respectively.

For each side-chain rotamer, there is an energy reflecting the frequency of occurrence in our structural database. For both protein A and mROP, this term contributes about 7% of the total energy in the predicted native state.

Long range interactions

There is a one-body, amino-acid specific potential of mean force that depends on the distance of the side groups center of mass from the center of mass of the protein chain [32]. This contribution reflects the tendency of some amino acids to be buried in the hydrophobic core, whereas others prefer to be exposed to solvent or to lie at the solvent–protein interface. It requires an *a priori* estimate of the radius of gyration, S. For single domain proteins, S is a well-defined function of the number of amino acids; thus, we only use the length of the sequence to determine the target value for the radius of gyration. This term typically makes a small contribution to the energy in the folded states, but can in many cases be used to match sequences to their corresponding structures.

Then, there are soft core repulsive regions, with widths that are amino-acid pair specific. The repulsion is on the order of $4k_BT$. Beyond the soft core repulsive region, there are pair potentials, ε_{ij}, between residues i and j constructed using the Bragg-Williams approximation [33] for the expected number of contacts, assuming random packing. These terms are used in the calculation of the non local interaction energy. A number of variants of the pair interaction energy have been employed, but the most successful form currently used for attractive pairs is

$$E_{long} = \epsilon_{ij} g_{ij} \text{ if } r_{ij} < d_{ij} \text{ and } \epsilon_{ij} < 0$$

with $g_{ij}=1$ if $|i-j| < 6$, and $g_{ij}=1-(f-0.88)^2$ otherwise. f is the cosine of the angle between vectors connecting α-carbons i-2 and i+2 and j-2 and j+2. r_{ij} and d_{ij} are the actual and the contact cutoff distances between the centers of mass of side chains i and j respectively. Interactions between all residues $|i-j| > 1$ are allowed.

For repulsive pairs, such that $r_{ij} < d_{ij}$ and $\varepsilon_{ij} > 0$,

$$E_{long} = \epsilon_{ij};\ |i-j| > 5.$$

The different treatment of attractive and repulsive pairs seems to accelerate folding. Moreover, the orientational term for attractive pairs reflects the fact that helices and β sheets tend to pack with a small crossing angle [34], but a non zero interaction term is permitted even when the two elements of the chain are orthogonal. This orientation term also partly accounts for the anisotropy in side-chain packing arising from the presence of the protein backbone. These pair potentials provide about 23% and 17% of the total energy in the predicted native conformation of protein A and mROP, respectively.

Cooperative side-chain interactions

Many attempts to predict the structure of a protein from its sequence have failed because the models could not generate the well-defined patterns of side-chain packing that are exhibited by globular proteins. Moreover, detailed analysis of these packing patterns indicates that amino-acid packing preferences are context sensitive. They not only depend on the particular partners, but also on how far the interacting partners are separated down the chain. A final motivation for introducing cooperative side-chain packing terms comes from the experimental observation that the molten globule to native state transition is very cooperative [14]. Because in many molten globule intermediates there is a substantial amount of secondary structure [14], presumably a portion of the cooperativity results from side-chain fixation into the well-defined patterns found in the native state. Figure 6 shows examples of characteristic side-chain–side-chain contact patterns.

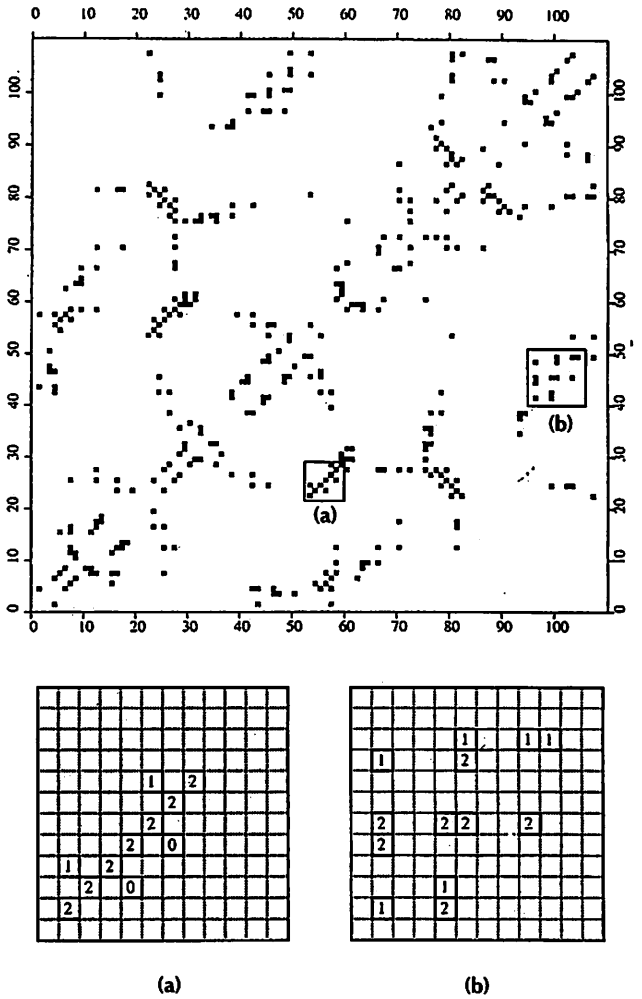

Fig. 6. An example of a side-chain contact map for the protein thioredoxin from which representative examples of contact maps for (a) a parallel β-hairpin motif and (b) a parallel helix–helix motif have been excised. The numbers in the boxes indicate the number of contact templates in which the particular interacting pair participates.

The novel and important component of our tertiary interaction scheme involves the inclusion of four body template interactions that bias for a subset of contact patterns seen in the native states of all kinds of globular proteins. This term is important only in the late stages of folding associated with side-chain fixation.

Suppose that a pair of side chains i and j are in contact. If another contact, say between residues j+4 and i-4 occurs, then there is an additional energetic contribution equal to the sum of the two

pairwise interactions. In fact, the complete set of cooperative side-chain interaction terms associated with an ij contact, the template energy, is taken to be

$$E_{tem} = \left(\epsilon_{ij} + \epsilon_{i+,kj+n}\right) C_{ij} C_{i+k,j+n};\ |k| = |n|, n = \pm 3, \text{ and } \pm 4.$$

$C_{ij} = 1(0)$ if side chains i and j are (not) in contact. The terms with n and k equal to 3 and 4 have been selected because they are consistent with subsets of all types of packing patterns seen in globular proteins. Clearly, this is a very naive implementation of such cooperative terms; in future work, these terms will be generalized. In the case of protein A and mROP, the contact template terms contribute on the order of 7% and 10% of the total energy of the native state respectively.

Whether these phenomenological terms would be really necessary if a whole-atom model were employed is unclear. One might imagine that they mimic the cooperative nestling of different rotamers that cannot be accounted for in a reduced side-chain representation. On the other hand, it is extremely difficult even in small molecule simulations to simulate crystallization from a liquid (this process may be analogous to the fixation of side chains on going from the molten globule to the native state); thus, these knowledge-based phenomenological potentials can help facilitate the process of side-chain fixation. However, as shown elsewhere [15], for one of the sequences designed by DeGrado and Raleigh [27], the presence of cooperative templates does not demand side-chain fixation, but on suitable modification of the sequence, side-chain fixation is predicted in apparent agreement with experiment [27].

Because of the choice of the diagonal ± 3 and ± 4 template type terms, there is the potential concern that these templates bias for helical states and thus might predict a helical topology independent of the sequence. This in fact is not the case. First of all, the template contribution to the energy of the folded state is small, being typically less than 10% of the native folded energy. It only contributes in the late stages of folding after topology assembly and secondary structure formation has occurred and acts to reduce the entropy of the side chains to produce a subset of protein-like patterns. Second, the fraction of residues in β proteins and α/β proteins that are involved in templates described by equation 5 is 27%, whereas 34% of residues in helical proteins participate. This might indicate a slight bias towards helical proteins, but the average number of template interactions per residue in β proteins is 2.3 versus 2.2 for helical proteins. Furthermore, the same interaction set which folded protein A and mROP can bring structures of the β protein plastocyanin, having a very loosely defined native topology and a Cα rms deviation from native of 11.0 Å, to structures whose rms deviations from native are below 4.0 Å. A similar experiment when applied to the α/β protein flavodoxin yields comparable results. When flavodoxin is folded from the denatured state, a mixed motif structure is predicted; however, there are errors at the level of the arrangement of the supersecondary elements. This indicates that either the simulations are not long enough, the interaction scheme requires improvement or that both conditions hold.

Perhaps, the most compelling evidence against the argument that there is a strong bias for the folding of helical proteins is provided by a series of *de novo* simulations on the folding of the minimal α/β protein crambin. This is a 46 residue protein whose native fold is composed of a small helical hairpin and three extended chains arranged into an antiparallel β sheet. It also contains three pairs of disulfide crosslinks which make the native conformation extremely stable. Starting from the unfolded state with no information whatsoever provided to the algorithm about the identity or number of crosslinks, and using the same parameter set as applied to protein A and mROP, the average coordinate rms deviation of the backbone Cαs is 3.6 Å from the set of corresponding atoms in the crystal structure. Furthermore, all disulfide bonds are correctly predicted. Thus, this strongly argues against the conjecture that our results for protein A and mROP arise from the use of a biased potential that can only generate helical bundles.

The refinement procedure

Once the sequence is folded on lattice, the resulting lattice conformation is relaxed using an off lattice Monte Carlo procedure that regularizes the Cα virtual bond lengths by permitting small local displacements of the α-carbons. Then, an analytic procedure reconstructs all the backbone heavy atoms as well as the β-carbons [30]. The side chains beyond Cβ are built using a heavy atom rotamer library. A genetic algorithm is used that penalizes heavy atom overlaps and favors side chains with centers of mass close to those obtained from the model on the fine lattice. Alternatively, an expanded chain conformation can be built. In both cases, the refined, fine lattice structure provides a set of side-chain–side-chain contacts (weighted by their importance and the mean distance between centers of mass) and secondary structure predictions that serve as input target functions for molecular dynamics simulations using the CHARMM [25] or AMBER [26] force fields. Several cycles of molecular dynamics refinement are carried out using protocols similar to those employed in the refinement of NMR structures [35]. The refined all atom structures are constructed from averages over these trials.

To test the validity of this approach, a set of analogous constraints was extracted from the experimentally determined structure of the B domain of staphylococcal protein A [21]. This 60 residue protein contains well-defined secondary structure between residues 10–55. Use of the MD refinement procedure produced a structure that has a 1.38 Å backbone heavy atom rms deviation from the solution structure. Furthermore, the family of structures generated from the refinement are reasonably unique with an average rms deviation of 0.95 Å from one another.

Acknowledgment: Stimulating discussions with Drs. W. Beers, J. Dyson, R.A. Lerner, M. Milik and P.E. Wright are gratefully acknowledged. We also thank Dr. C. Sander for providing us with the mROP sequence. This research was supported in part by NIH grants GM-37408 and GM-37554.

References

1. Bowie JU, Luethy R, Eisenberg D: A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 1991, 253:164–170.
2. Sippl MJ, Weitckus S: Detection of native-like models for amino-acid sequences of unknown three-dimensional structure in a database of known protein conformations. *Proteins* 1992, 13:258–271.
3. Maiorov VN, Crippen GM: Contact potential that recognizes the correct folding of globular proteins. *J Mol Biol* 1992, 277:876–888.
4. Godzik A, Skolnick J, Kolinski A: A topology fingerprint approach to the inverse folding problem. *J Mol Biol* 1992, 227:227–238.
5. Jones DT, Taylor WR, Thornton JM: A new approach to protein fold recognition. *Nature* 1992, 358:86–89.
6. Ripoll DR, Piela L, Velasquez M, Scheraga HA: On the multiple minima problem in the conformational analysis of polypeptides. *Proteins* 1991, 10:188–198.
7. Skolnick J, Kolinski A: Computer simulations of globular protein folding and tertiary structure. *Annu Rev Phys Chem* 1989, 40:207–235.
8. Levitt M, Warshel A: Computer simulation of protein folding. *Nature* 1975, 253:694–698.

9. KUNTZ ID, CRIPPEN GM, KOLLMAN PA, KIMELMAN D: Calculation of protein tertiary structure. *J Mol Biol* 1976, 106:983–994.
10. HAGLER AT, HONIG B: On the formation of protein tertiary structure on a computer. *Proc Natl Acad Sci USA* 1978, 75:554–558.
11. WILSON C, DONIACH S: A computer model to dynamically simulate protein folding: studies with crambin. *Proteins* 1989, 6:193–209.
12. YCAS M: Computing tertiary structures of proteins. *J Prot Chem* 1990, 9:177–200.
13. DOLGIKH DA, *ET AL.*: α-lactalbumin: compact state with fluctuating tertiary structure? *FEBS Lett* 1981, 136:311.
14. KUWAJIMA K: The molten globule. *Proteins* 1989, 6:87–103.
15. 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 helical proteins. *J Chem Phys* 1993, 98:7420–7433.
16. KOLINSKI A, SKOLNICK J: Discretized model of proteins. I: Monte Carlo study of cooperativity in homopolypeptides. *J Phys Chem* 1992, 97:9412–9426.
17. KABSCH W, SANDER C: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983, 22:2577–2637.
18. SKOLNICK J: Lattice parameter set. Available via anomymous ftp from scripps.ftp.
19. UHLEN M, *ET AL.*: Complete sequence of the Staphylococcal gene encoding protein. *J Biol Chem* 1984, 259:1695–1702.
20. DEISENHOFER J: Crystallographic refinement and atomic models of a human FC fragment and its complex with fragment B of protein A from *Staphylococcus areus* at 2.9 Å resolution. *Biochemistry* 1981, 20:2361.
21. GOUDA H, *ET AL.*: Three-dimensional solution structure of the B-domain of Staphylococcal Protein A: comparisons of the solution and crystal structures. *Biochemistry* 1992, 40:9665–9672.
22. SANDER C (EDS): *Protein Design Exercises 86.* Heidelberg: European Molecular Biological Laboratory; 1986, 1.
23. BANNER DW, KOKKINIDIS M, TSERNOGLOU D: Structure of the ColE1 rop protein at 1.7 Å resolution. *J Mol Biol* 1987, 196:657–675.
24. GODZIK A, KOLINSKI A, SKOLNICK J: Lattice representation of globular proteins: how good they are and how good are they? *J Comp Chem*, in press.
25. BROOKS BR, *ET AL.*: CHARMM: a program for macromolecular energy, minimization and dynamics calculations. *J Comp Chem* 1983, 4:187–217.
26. PEARLMAN DA, *ET AL.*: AMBER User Manual. San Francisco: University of California; 1991.
27. RALEIGH DP, DEGRADO WF: A *de novo* designed protein shows a thermally induced transition from a native to a molten globule-like state. *J Am Chem Soc* 1992, 114:10079–10081.
28. GODZIK A, KOLINSKI A, SKOLNICK J: *De novo* and inverse folding predictions of protein structure and dynamics. *J Comp Aided Mol Des* 1993, in press.
29. HANDEL T, DEGRADO WF: A designed 4-helical bundle shows characteristics of both molten globule and native state. *Biophysical J* 1992, 61:A265.
30. REY A, SKOLNICK J: Analytical reconstruction of a protein backbone from the α-carbon coordinates. *J Comp Chem* 1992, 13:443–456.
31. BERNSTEIN FC, *ET AL.*: The protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 1977, 112:535–542.
32. NIKISHAWA K, OOI T: Radial locations of amino-acid residues in a globular protein: correlation with the sequence. *J Biochemistry* 1986, 100:1043–1047.
33. HILL TL: *Statistical Mechanics.* New York: McGraw-Hill; 1956.
34. MURZIN AG, FINKELSTEIN AV: General architecture of the α-helical globule. *J Mol Biol* 1988, 204:749–769.
35. BRUNGER AT, CLORE GM, GRONENBORN AM, KARPLUS M: Three dimensional structure of proteins determined by molecular dynamics with interproton distance restraints. *Proc Natl Acad Sci USA* 1986, 83:3801–3805.

Received: **19 April 1993**; revised: **8 June 1993.**
Accepted: **8 June 1993.**