PROTEINS: Structure, Function, and Genetics Suppl 5:149–156 (2001)
DOI 10.1002/prot.1172

Ab Initio Protein Structure Prediction via a Combination of Threading, Lattice Folding, Clustering, and Structure Refinement

Jeffrey Skolnick,[1]* Andrzej Kolinski,[1,2]* Daisuke Kihara,[1] Marcos Betancourt,[1] Piotr Rotkiewicz,[2] and Michal Boniecki[2]
[1]*Donald Danforth Plant Science Center, Saint Louis, Missouri*
[2]*Faculty of Chemistry, Warsaw University, Warsaw, Poland*

ABSTRACT **A combination of sequence comparison, threading, lattice, and off-lattice Monte Carlo (MC) simulations and clustering of MC trajectories was used to predict the structure of all (but one) targets of the CASP4 experiment on protein structure prediction. Although this method is automated and is operationally the same regardless of the level of uniqueness of the query proteins, here we focus on the more difficult targets at the border of the fold recognition and new fold categories. For a few targets (T0110 is probably the best example), the ab initio method produced more accurate models than models obtained by the fold recognition techniques. For the most difficult targets from the new fold categories, substantial fragments of structures have been correctly predicted. Possible improvements of the method are briefly discussed. Proteins 2001;Suppl 5:149–156.** © **2002 Wiley-Liss, Inc.**

Key words: CASP4; protein folding; protein structure prediction; lattice models; Monte Carlo methods; threading; structure clustering

INTRODUCTION

Prediction of three-dimensional protein structure from the amino acid sequence is one of the most challenging problems in theoretical structural biology.[1,2] Although the task is relatively easy when a new protein exhibits significant sequence similarity to one (or more) proteins for which the structure is already known,[3–5] with decreasing sequence similarity, the problem becomes more difficult and the predictions less certain and less accurate.[6–10] However, to be relevant in the postgenomic era, any contemporary structure prediction method, in addition to its prediction fidelity, must be capable of automated, large-scale predictions, associated with the structural annotation of entire genomes.[11]

To address these issues, over the last 2 years, we have developed a new methodology for protein structure prediction that consists of a hierarchy of sequence comparison, threading,[12] lattice folding,[13] and fold selection procedures. What seems to be important is that the proposed methodology remains qualitatively the same regardless of the sequence similarity or sequence-structure compatibility to proteins of known structure.[14] There are, however, some intrinsic limitations to the new method. Because structure assembly is performed by using a reduced, limited-resolution model, the methodology is not competitive with standard comparative modeling in the limit of a very high sequence similarity. In such cases where more traditional comparative modeling methods produce models with a coordinate root-mean-square deviation (cRMSD) from experimental structures in the range of 1.5–2.5 Å, our method will usually give somewhat less accurate models. This is because, depending on protein size, the inherent resolution of the model is in the range of 1.5–3 Å. For moderate sequence similarity, where fold recognition is relatively easy, but where the structural similarity between the target and template is limited, the method frequently produces molecular models that are closer to the true structure of the query protein than to the structure of the template protein.[14,15] Finally, in the range of nondetectable sequence similarity, the success ratio for the method strongly depends on protein size. For large proteins, ab initio protein structure assembly becomes significantly more computationally expensive, with fold selection being more uncertain. It is of interest that ab initio sometimes produces better models than those obtained via fold recognition methods.

MATERIALS AND METHODS

An outline of the methodology applied to all (except one) of the targets in the CASP4 experiment is given in the flow chart (Fig. 1).

Initial Parameter Derivation

The process starts from a standard multiple sequence alignment. The alignment facilitates the derivation of protein-dependent statistical short-range potentials, pairwise potentials, and secondary structure predictions. The short-range potentials define the distribution of distances between the i-th and i+k (with k = 1, 2, 3, and 4) centers of the side groups. The pairwise potentials for side groups

Grant sponsor: Division of General Medical Sciences of the National Institutes of Health; Grant numbers: GM-37408 and GM-48835.

*Correspondence to: Jeffrey Skolnick or Andrzej Kolinski, Donald Danforth Plant Science Center, 893 North Warson Rd., St. Louis, MO 63141. E-mail: skolnick@danforthcenter.org

Received 5 April 2001; Accepted 13 July 2001

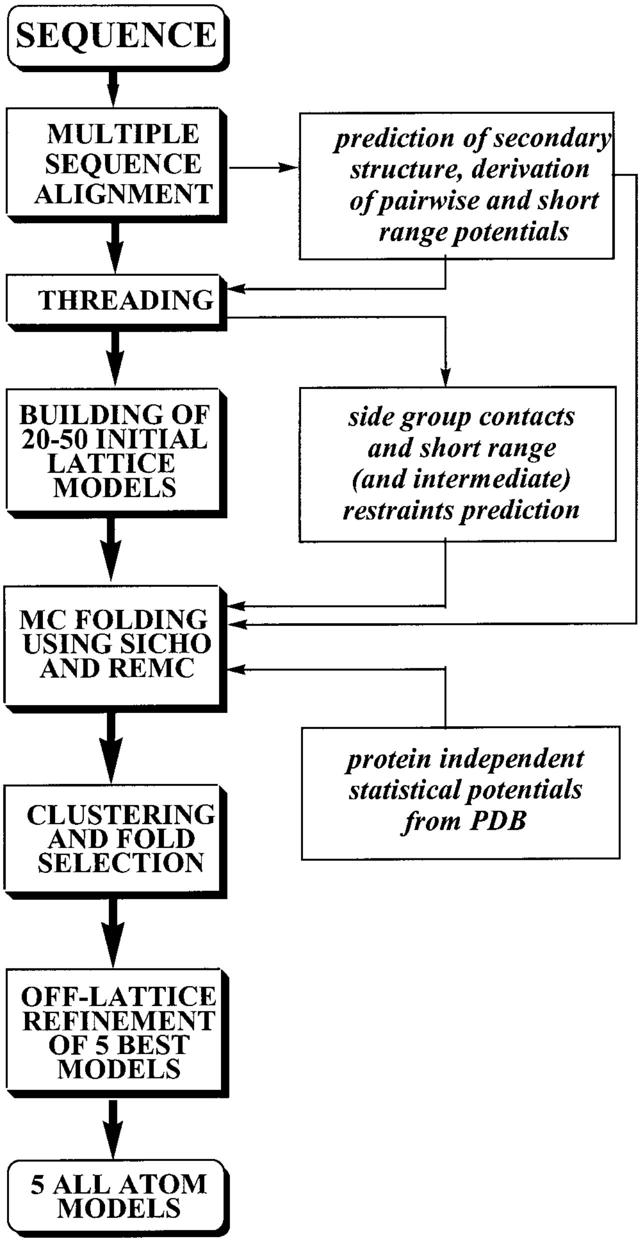
SEQUENCE
MULTIPLE SEQUENCE ALIGNMENT
prediction of secondary structure, derivation of pairwise and short range potentials
THREADING
BUILDING OF 20-50 INITIAL LATTICE MODELS
side group contacts and short range (and intermediate) restraints prediction
MC FOLDING USING SICHO AND REMC
protein independent statistical potentials from PDB
CLUSTERING AND FOLD SELECTION
OFF-LATTICE REFINEMENT OF 5 BEST MODELS
5 ALL ATOM MODELS

Fig. 1. Flowchart of the fold prediction methodology. The procedure is qualitatively the same for easy and difficult targets. The only difference is that for easy targets (comparative modeling and easier fold recognition), the set of restraints also contains an explicit template treated as an additional subset of soft restraints for the Monte Carlo folding program.

depend on the mutual orientation of the interacting chain units.[13–16] This potential is also used in our threading algorithm, PROSPECTOR.[12] Predicted secondary structure provides a bias for short-range conformational propensities and some weak restrictions on the hydrogen bond network in the lattice model's force field.

Threading-Based Prediction of Side-Group Contacts and Short-Range Restraints

A very important aspect of our approach is the threading stage in which we used our recently developed program PROSPECTOR.[12] The threading procedure is used as a tool to identify templates, when appropriate, but also for the side-group contact prediction and for the prediction of short-range and intermediate-range distances between the side groups along the polypeptide chain. The accuracy of the contact prediction depends on the presence of existing homologous proteins in the structural database. When the score for the top scoring sequences is high, the contact prediction and the distance prediction are very accurate, because the contribution from closely related folds is large. In such cases, the remaining protocol resembles comparative modeling with multiple templates (of various "weights") and "fuzzy" spatial restraints. Moreover, in such a situation the starting structures are already quite close to the target structure. When the top scoring structure is additionally used as a spatial template (Monte Carlo folding occurs in a portion of conformations strongly restricted to a neighborhood of the template) the procedure switches to the Generalized Comparative Modeling (GENECOMP) mode.[14] Of course, this variant of the methodology is not used for the "novel folds" or for the difficult cases of "fold recognition." In these situations, the predicted contacts are culled from templates that can have a different global fold, with corresponding lower accuracy. Similarly, the short-range distance restraints originate from numerous weakly similar protein fragments and are of a lower accuracy. Moreover, the starting lattice models are built from very fragmentary templates provided by PROSPECTOR and are usually very far from the template structure. Sometimes, the starting structures may contain small elements of a supersecondary structure resembling structural motifs present in the target structure. Others may provide a folding nucleus.

Folding With Reduced Model (SICHO) and the Replica Exchange Monte Carlo Technique

Initial models from threading provide the seed structures for the Monte Carlo folding stage of the method. The folding algorithm uses a reduced representation of protein conformational space where the center of mass of the α-carbon and side chain heavy atoms define the interaction centers; we term this the SIde CHain Only or SICHO model.[16] Such sites are restricted to a high-coordination lattice. Besides the various restraints discussed above, the force field consists of several potentials derived from a statistical analysis of structural regularities of known protein structures. These potentials mimic short-range conformational propensities, pairwise side-chain interactions, three-body interactions of the side chains, and multibody interactions, which account for the averaged effect of the solvent and a cooperative network of main-chain hydrogen bonds. Details of the protein representation, conformational updating, and the force field of the model can be found elsewhere.[17]

The Replica Exchange Monte Carlo (REMC) technique was used to sample conformational space.[18,19] In REMC, a number of copies (replicas) of the model system are simulated simultaneously at various temperatures by using a traditional (in this case the asymmetric Metropolis scheme) sampling scheme. A relatively wide range of temperatures is covered, with the estimated folding temperature somewhere in the middle of this range. Occasionally, the replicas are swapped with a probability dependent on their relative conformational energy and temperature. This way the copies of the system sample not only conformational space but also various temperatures. Exchanges of the replicas between various temperatures facilitate fast transitions between local minima and significantly accelerate the convergence to the global minimum. This way, in REMC, a number of distinct conformations are in direct competition with each other. Several recent works on the application of Monte Carlo methods to protein-like molecular models indicated that the REMC technique is a much more efficient conformational search tool than, for example, simulated annealing or generalized ensemble versions of MC.[18,19] For a few large proteins, we applied a "cascade" approach. First, a number of REMC simulations were performed by using a different set of starting structures. Then, the best structures from each of the REMC runs were selected from the combined set of replicas for the subsequent simulation. Such a procedure might partially address the problem of the slow relaxation of large systems.

Fold Selection and Refinement

A large number of structures generated from REMC simulations were subsequently subjected to a clustering procedure.[20] The five best structures from the clustering algorithm were then subjected to a refinement process. An intermediate resolution off-lattice model and a Monte Carlo simulated annealing technique were used in this stage. Modeled structures were kept near the structures obtained from clustering. Although off-lattice refinement does not improve the overall accuracy of the prediction, it does regularize the local geometry.

Finally, all atom structures were rebuilt by using MaxSprout,[21] and the coordinates of the main chain were submitted to the CASP4 Prediction Center (http://predictioncenter.llnl.gov/casp4/Casp4.html).

RESULTS AND DISCUSSION

Predictions were made for all except for one of the targets of CASP4 (the 800-residue T0116). Here, we discuss the results of the more difficult targets, classified as new folds (NF) or new folds/fold recognition (FR) by the assessors. In addition, we present our predictions for a couple of targets (T0102 and T0110) that could also be predicted by threading approaches (and were properly matched to known folds by others). This is justified because our ab initio methodology produced lower RMSD models than were obtained by others via model building on threading-based templates.

Summary of the Prediction Results

In Table I, we compile some of our more significant predictions. Relatively large fragments of structures have been properly predicted for a number of difficult targets. However, for very difficult targets, classified by the CASP4 assessors as "new folds," our method failed to correctly predict the entire structure. In most cases, secondary structure elements of the fold and elements of supersecondary structure were predicted correctly; however, they had topological errors that led to a large overall cRMSD from the experimental structures. In other cases, as in the C-terminal domain of T0096, a quite accurate fold corresponding to the mirror image topology of the native structure was obtained. Figure 2 illustrates some of our predictions compared with the experimental structures.

TABLE I. Summary of Structure Predictions for Selected Targets

Target	Model	Position[a]	Length[b]	RMSD
T0087	1	12	42	4.84
T0087	1	1	30	3.80
T0087	4	218	44	4.73
T0087	4	219	37	3.55
T0089	1	161	96	5.52
T0089	2	280	64	3.98
T0094	1	45	28	4.80
T0094	1	13	24	3.78
T0094	3	48	39	4.88
T0094	3	40	31	3.75
T0096	3	131	91	5.30
T0096	3	167	55	3.15
T0097	1	37	55	4.78
T0097	1	37	42	3.83
T0098	1	1	42	4.71
T0098	1	1	38	3.33
T0102	1	1	70	3.61
T0106	4	73	38	4.94
T0106	4	72	25	3.34
T0110	1	1	95	5.09
T0110	1	50	43	3.40
T0110	3	1	95	4.21
T0110	3	44	47	3.55
T0114	4	1	87	8.56
T0114	4	44	43	4.62
T0115	2	171	46	4.71
T0115	5	171	42	3.88
T0120	1	115	28	3.86
T0124	1	71	77	3.95

[a] The residue number from which the fragment of a given length starts.

[b] The fragment of this length is compared with the experimental structure (coordinate RMSD after the best superposition).

An Ab Initio Approach Can Produce More Accurate Models Than Fold Recognition

In some cases (T0102, T0110, and some others, as well as T0114, to a lesser extent), quite accurate models were obtained despite the fact that our threading procedure did not recognize remotely similar folds present in the PDB.[22] For target T0102, a cyclic polypeptide consisting of 70 amino acids, our procedure produced a very good model with a cRMSD equal to 3.6 Å from the native structure for the first model submitted. There was only one more prediction of the same accuracy (model 4 from the Baker group); however, a few other groups produced models of comparable quality in the range of 4.0–4.3 Å from native

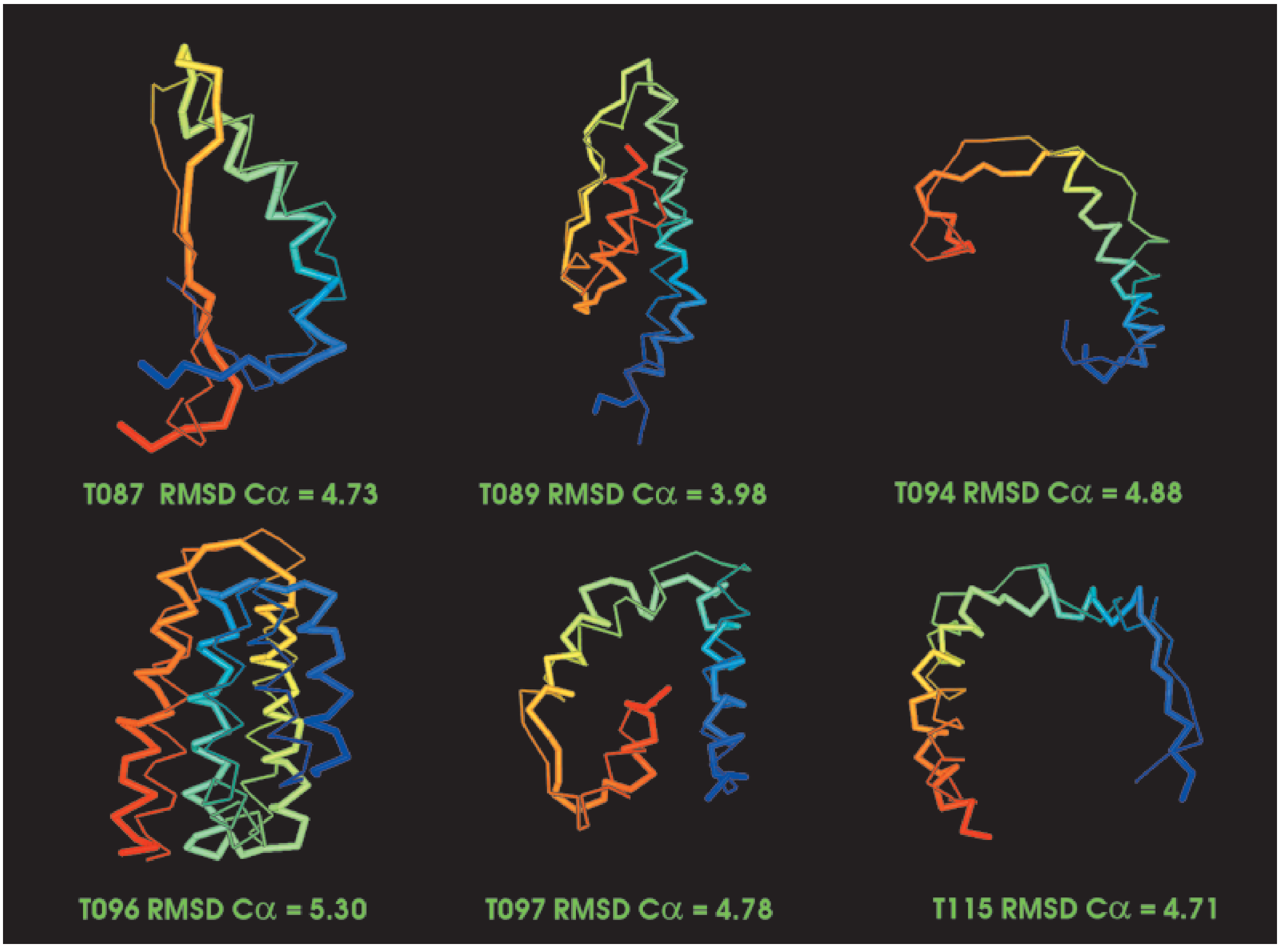

Fig. 2. Examples of well-predicted protein fragments (thick lines) superimposed onto equivalent fragments of experimental structures (thin lines).

(Osguthorpe, Friesner, Scheraga). Figure 3 shows the predicted α-carbon trace superimposed on the experimental structure of T0102.

For T0110, a 95-residue α/β protein of a quite complicated fold, our ab initio prediction produced the most accurate model, quantitatively better than the models resulting from predictions based on fold recognition approaches or on various ab initio techniques (again, several groups predicted qualitatively correct structures with a somewhat lower overall accuracy). The third model had a cRMSD of 4.2 Å from native. The remaining models (from different clusters of the MC simulations) were also quite good, with cRMSDs of 5.1 Å, 5.0 Å, and 7.2 Å for models 1, 2, and 5, respectively. Model 4 with a cRMSD equal to 10.0 Å had the mirror image topology, with an otherwise qualitatively correct contact map for the side groups and correct secondary structure.

Figure 4 illustrates various stages of the model building of T0110. The first panel shows the virtual chain connecting the centers of mass of the side groups, as provided by the SICHO based Monte Carlo simulations. The second panel shows the full atom model after rebuilding the alpha carbon trace, the off-lattice refinement, and after reconstruction of the side-chain atoms. In the third panel, the final predicted alpha carbon trace is superimposed onto the experimental structure.

What Drives Fold Assembly?

The model interaction scheme consists of three distinct components. First, it has a force field composed of various potentials describing short-range conformational propensities, pairwise interactions of the side group, a model of hydrogen bonds, and multibody terms mimicking an averaged hydrophobic effect. This force field enables the ab initio folding of only very small proteins of a simple topology. For example, target T0102 has been folded by using this minimal set of interactions. Implementation of the remaining components of the force field to this special case of a cyclic polypeptide was considered unnecessary, but it works as well when used. For larger proteins with more complex topology, such a straightforward approach would be insufficient. The two remaining components of the model force field are based on the local (or global) sequence similarity and are protein specific. By using a criterion of local sequence similarity, a protein-specific short-range potential could be derived and combined with the generic short-range potentials. Similarly, by taking into account the local sequence similarity of two peptide

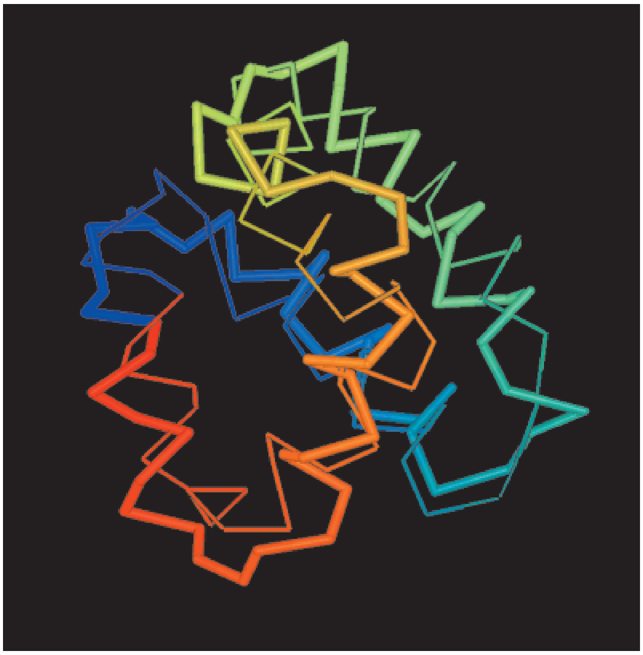

Fig. 3. Predicted structure of T0102 (thick line) superimposed on the alpha carbon trace of the experimental structure (thin line).

fragments associated with long-range side group contacts, one may build a protein-specific pairwise potential. These potentials, additionally combined with a conservative prediction of protein secondary structure, noticeably improve the model's specificity. However, the most important influence on the likelihood of good structure prediction is the quality of the contact prediction obtained from threading.

Table II summarizes the results of the contact prediction procedure for selected targets. There seems to be a clear correlation between the accuracy of contact prediction (and the number of predicted contacts) and the quality of our prediction. For example, for T0110, the contact prediction is very good; this resulted in a very good prediction of the entire structure of this protein. This contact prediction was based on interacting regions in proteins that did not have the global topology of T0110. The early version of our threading algorithm used during the CASP4 experiment did not recognize any globally similar fold to T0110; thus, the contact prediction emerged from many proteins with structurally similar fragments. In other words, the algorithm did not use any global template, but a number of partial templates contributed to contact and distance predictions. The quality of contact prediction for this target was good enough to reproduce some subtle structural details, including the bending of the long helix. It is of interest that the accuracy of the secondary structure prediction appears to be of lesser importance. Overpredictions or underpredictions of helices or β-strands are not very harmful; however, grossly wrong predictions (helix instead of beta, or vice versa) are dangerous. For this reason, we used very conservative predictions that consisted of a consensus of high fidelity predictions from various methods.[24] It should also be noted that contact predictions and local distance predictions are to some extent redundant, with the protein-specific potentials derived from analysis of local sequence similarity to proteins of known structure. The latter are less specific for the sequence of interest.

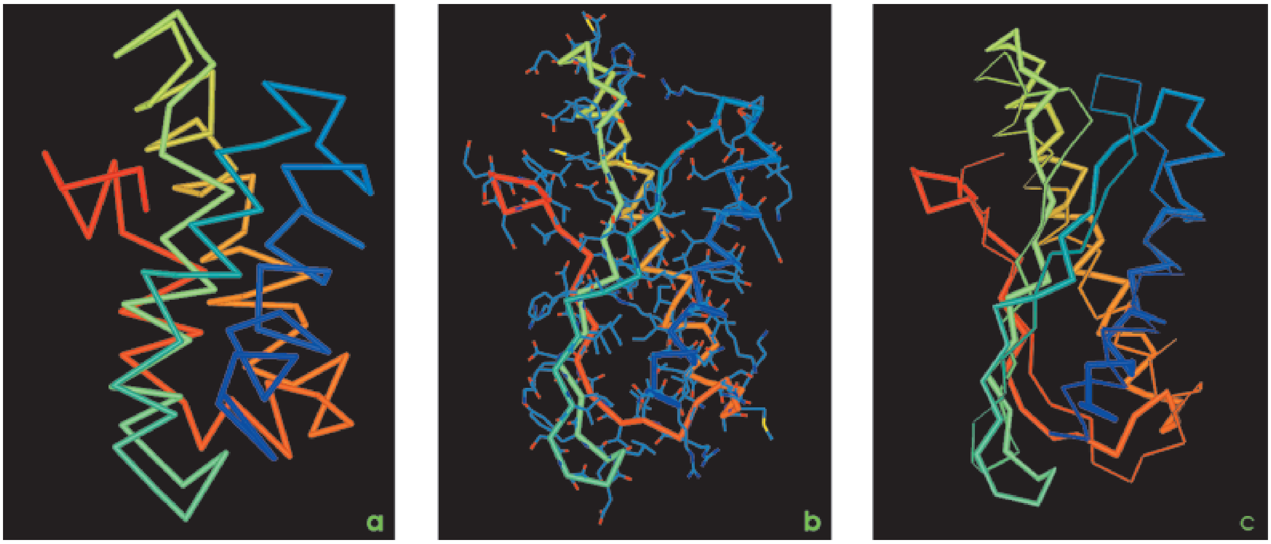

Fig. 4. Prediction of the T0110 structure during various stages of the method outlined in Figure 1. **A:** The best SICHO structure after clustering—the structure closest to the centroid of the best cluster. The thick line connects the side chain centers of mass. **B:** The structure after the rebuilding of the alpha carbon trace; refinement was done with an intermediate resolution off-lattice protein model and with reconstructed side chains. The centers of mass of the side chains from (A) are used as a guiding restraint for the side-group rotamer selections. Only a very brief optimization is needed after this stage to obtain good side-chain packing. **C:** The α-carbon trace (thick line) of the final structure is superimposed onto the experimental structure (thin line).

TABLE II. Summary of the Accuracy of the Contact Predictions for Selected Targets

Name of protein	No. of residues	No. of contacts predicted	Fraction of contacts correct within δ = 0 residues	Fraction of contacts correct within δ = ±1 residues	Fraction of contacts correct within δ = ±2 residues	Fraction of contacts correct within δ = ±3 residues	Fraction of contacts correct within δ = ±4 residues
t0087	309	44	0.16	0.45	0.57	0.73	0.84
t0096	155	22	0.18	0.5	0.64	0.82	0.82
t0097	105	38	0.11	0.42	0.55	0.71	0.79
t0098	119	52	0.15	0.33	0.56	0.71	0.79
t0102	70	34	0.12	0.5	0.68	0.79	0.88
t0105	94	25	0.16	0.6	0.84	0.92	0.92
t0106	125	44	0.07	0.18	0.32	0.48	0.64
t0110	95	54	0.33	0.5	0.65	0.7	0.83
t0114	87	29	0.21	0.48	0.76	0.9	0.97
t0115	296	20	0.15	0.4	0.65	0.7	0.9
t0120	203	8	0.25	0.5	0.63	0.75	0.75

Advantages of This Method

The method is fully automated, and the methodology is the same regardless of the existing homology between the query protein and the proteins in the structural database. Thus, it can be easily applied to the structural annotation on a genomic scale. A large success rate, which is competitive with other methods (a large fraction of correct and accurate predictions), could be expected for the following types of proteins. First is the broad class of proteins that have weakly homologous proteins in the structural database. However, they do not need to be significantly high scoring in the threading stage to contribute to the accuracy of the side-group contact prediction. Indeed, the ab initio approach with this method was more successful in a number of cases than fold recognition approaches were in the geometric accuracy of the predicted structures. The best examples are predictions for T0102 and T0110. In addition, for distant homology modeling cases, where the sequence similarity is easily detectable, but the structures of the template and the query protein differ significantly, the proposed method can modify the template structure to a larger extent than is possible in more traditional comparative modeling methods. Modeling of the loops allows for great flexibility, and the initial alignment is usually improved during the restrained Monte Carlo folding. (Relatively good models were thus generated for T0092, the N-terminal domain of T0096 and T0112.) The "New Fold" method performs well for small proteins of not too complex topology. Another potential advantage of the present approach is that, when needed, some information about folding intermediates, folding nuclei, and so forth can be extracted from the MC simulations.[23]

What Went Wrong? Problems With the Present Approach and Its Limitations

To increase prediction fidelity and the accuracy of obtained models for the most difficult cases of new folds, almost all components of our combined method need improvement or even major revision. The threading stage is used for the derivation of the tertiary restraints of the Monte Carlo folding algorithm. Chances for good predictions strongly depend on the number and the accuracy of predicted contacts. This situation needs to be improved and probably combined with other methods of contact prediction such as correlated mutation analysis[8] and related analyses of protein sequences.

The second problem is due to the Monte Carlo fold assembly procedure. Most CASP4 targets were relatively large proteins. The algorithm was not well tested in this range of protein sizes prior to CASP4. Clearly, simulations for numerous targets were simply too short. Because the adaptation of the REMC technique to parallel processing is relatively straightforward, future applications should include the exploration of this possibility. Sampling could be also improved in a different way. Frequently, especially for long-chain proteins, the imperfections of the reduced lattice models lead to chain "crumpling" or entanglements, which decrease the efficiency of the sampling. Of course, there is an apparent simple cure to the problem by increasing the relative weight of the force field components controlling the short-range interactions and the generic stiffness of the polypeptide chain. This unfortunately slows down the chain dynamics and subsequently increases the folding time. A number of methods shown by the CASP4 participants (especially Baker's very successful method[9]) were based on the concept of fold assembly from small fragments excised from known protein structures. Such methods, although fast and straightforward, have their disadvantages as well, due to a lack of local repacking and problems with handling long chains. On the other hand, a lot of local details, which need to be painstakingly built into our approach, are provided by fragment assembly techniques. Thus, it is very tempting to combine the basic attributes of both methods. Namely, during the Monte Carlo simulations, the higher temperature copies of the system could be occasionally subjected to a retracking procedure where the SICHO chain is replaced by a chain built from PDB-based fragments loosely fitting the original lattice chain. Subsequently, the "regularized" chain could be projected back onto the lattice. This should provide a very efficient means of purging non-protein-like conformations from the pool of replicas and a bias toward more regular and protein-like structures. At the same time, chain compactness and a reasonable force field should facilitate rapid fold assembly. Work along this line is already in progress.

The knowledge-based force field of our model (heuristic potentials derived from statistical analysis of the regularities seen in known protein structures) are far from perfect. Unfortunately, its incremental improvements, although systematically made, are time-consuming and costly (except the trivial one resulting from increasing the size of the PDB) due to the necessity of doing careful computational tests of all updates. On the other hand, in many cases, the Monte Carlo algorithm generates very good structures (low cRMSD from the experimental structure) that have conformational energies higher than those of the lowest energy conformations. Simply put, for many proteins the correlation between the energy and cRMSD is poor. Consequently, fold selection is not a trivial task and failed for several CASP4 targets. Parenthetically, let us note that in numerous recent approaches to protein structure prediction, the fold generation stage and the fold selection are separated in the sense that different force fields (or scoring functions) are used for each.[7,9] In our approach, the improvement of the fold selection should probably be based on the all-atom reconstruction of representative structures generated during MC sampling with the subsequent application of atom-based potentials of the mean force. The preliminary results in this direction are rather encouraging.

However, there could be other reasons for the failure of the proposed method. A good example is T0114, a relatively small protein of 87 residues. In this case, the contact prediction was quite successful. Of 29 contacts predicted, 76% were accurate within ±2 residues and 97% were accurate within ±4 residues. In addition, the Monte Carlo sampling should be very efficient for a protein of this size. Nevertheless, no good models were produced, and the results of simulations did not cluster well. Probably the number of predicted restraints and their accuracy was still too low to guide the fold assembly process. Indeed, despite its small size, T0114 has a quite complex β-barrel topology, with some rare structural "decorations," where with a large number of predicted contacts, the registration errors of the predictions cancel out to some extent. With a small number of predicted contacts, the registration errors might be very dangerous, especially for β-type proteins. All of the above-discussed deficiencies of various elements of our methodology led to failure in this case where one might expect better performance of the method; only small fragments (36 residues) were predicted correctly in model 4.

CONCLUSIONS

Our results show that it is possible for a fraction of globular proteins to assemble low- to moderate-resolution protein structures starting from a collection of essentially random conformations. Sampling occurs in a conformational space defined by a high-coordination lattice model of a polypeptide chain, and optimization is controlled by the Replica Exchange Monte Carlo scheme. The folding process is driven to a large extent by the accuracy of predicted tertiary contacts. The method is applicable to all variants of the protein structure prediction problem, from comparative modeling to the prediction of new folds. The procedure is automatic and can be applied on a genomic scale. The best results, compared with other approaches, are obtained in the case of very weak homology to known structures. This implies that the division of the protein structure prediction problem into various classes with respect to the level of similarity to already known structures becomes more and more diluted. The major weakness of the presented methodology, which needs to be addressed in the near future, lies in the poor sampling of larger proteins and problems with best fold selection.

REFERENCES

1. Montelione GT, Anderson S. Structural genomics: keystone for a Human Proteome Project. Nat Struct Biol 1999;6:11–12.
2. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. Nature Biotech 2000;18: 283–287.
3. Sali A, Overington JP, Johnson MS, Blundell TL. From comparison of protein sequences and structures to protein modeling and design. Trends Biochem Sci 1990;15:235–250.
4. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
5. Aszodi A, Tylor WR. Homology modeling by distance geometry. Folding Design 1996;1:325–334.
6. Lee J, Liwo A, Ripoll DR, Pillardy J, Scheraga HA. Calculation of protein conformation by global optimization of potential energy function. Proteins 1999;Suppl 3:204–208.
7. Samudrala R, Xia H, Huang E, Levitt M. Ab initio proteins structure prediction using a combined hierarchical approach. Proteins 1999;Suppl 3:194–198.
8. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. Ab initio folding of proteins using restraints derived from evolutionary information. Proteins 1999;Suppl 3:117–185.
9. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;Suppl 3:171–176.
10. Zhang B, Jaroszewski L, Rychlewski L, Godzik A. Similarities and differences between nonhomologous proteins with similar folds: evaluation of threading strategies. Folding Design 1997;2:307–317.
11. Clark MS. Comparative genomics: the key to understand the Human Genome Project. Bioessays 1999;21:121–130.
12. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. Proteins 2001;42:319–331.
13. Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. Protein folding: flexible lattice models. Prog Theor Physics (Kyoto) 2000;Suppl 138:292–300.
14. Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J. Generalized Comparative Modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. Proteins 2001;44:133–149.
15. Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. A method for the improvement of threading-based protein models. Proteins 1999;37: 592–610.
16. Kolinski A, Rotkiewicz P, Skolnick J. Application of high coordination lattice model in protein structure prediction. In: Grassberger P, Barkema GT, Nadler W, editors. Monte Carlo approach to biopolymers and protein folding. Singapore/London: World Scientific; 1998. p 100–130.
17. Skolnick J, Kolinski A, Ortiz AR. Derivation of protein specific pair potentials based on weak sequence fragment similarity. Proteins 2000;38:3–16.

18. Hansmann UHE. Parallel tempering algorithm for conformational studies of biological molecules. Chem Phys Lett 1997;281: 140–150.
19. Gront D, Kolinski A, Skolnick J. Comparison of three Monte Carlo search strategies for a proteinlike homopolymer model: folding thermodynamics and identification of low-energy structures. J Chem Phys 2000;113:5065–5071.
20. Betancourt MR, Skolnick J. Finding the needle in a haystack: educing protein native folds from ambiguous ab initio folding predictions. J Comput Chem 2001;22:339–353.
21. Holm L, Sander C. Database algorithm for generating protein backbone and the side chain coordinates from the Cα trace. J Mol Biol 1991;218:183–194.
22. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Simanouchi T, Tasumi M. The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
23. Kolinski A, Ilkowski B, Skolnick J. Dynamics and thermodynamics of beta-hairpin assembly: insight from various simulation techniques. Biophys J 1999;77:2942–2952.
24. Rost B, Sander C. Progress of 1D protein structure prediction at last. Proteins 1996;23:295–300.