PROTEINS: Structure, Function, and Genetics 44:133–149 (2001)

Generalized Comparative Modeling (GENECOMP): A Combination of Sequence Comparison, Threading, and Lattice Modeling for Protein Structure Prediction and Refinement

A. Kolinski,[1,2] M.R. Betancourt,[1] D. Kihara,[1] P. Rotkiewicz,[2] and J. Skolnick[1]*
[1]*Laboratory of Computational Genomics, Donald Danforth Plant Science Center, St. Louis, Missouri*
[2]*Department of Chemistry, University of Warsaw, Warsaw, Poland*

ABSTRACT **An improved generalized comparative modeling method, GENECOMP, for the refinement of threading models is developed and validated on the Fischer database of 68 probe–template pairs, a standard benchmark used to evaluate threading approaches. The basic idea is to perform ab initio folding using a lattice protein model, SICHO, near the template provided by the new threading algorithm PROSPECTOR. PROSPECTOR also provides predicted contacts and secondary structure for the template-aligned regions, and possibly for the unaligned regions by garnering additional information from other top-scoring threaded structures. Since the lowest-energy structure generated by the simulations is not necessarily the best structure, we employed two structure-selection protocols: distance geometry and clustering. In general, clustering is found to generate somewhat better quality structures in 38 of 68 cases. When applied to the Fischer database, the protocol does no harm and in a significant number of cases improves upon the initial threading model, sometimes dramatically. The procedure is readily automated and can be implemented on a genomic scale. Proteins 2001;44:133–149.**

Key words: protein structure prediction; threading; protein structure refinement; comparative modeling; knowledge-based potentials; lattice protein models; Replica Exchange Monte Carlo

INTRODUCTION

In this postgenomic era, when knowledge of the full proteome (the entire set of protein sequences) of an organism is becoming commonplace, the theoretical prediction of protein structure from the sequence of amino acids has also become a very urgent task for computational biology.[1] For a given protein sequence of unknown structure, there are three possible theoretical approaches to structure prediction. The most welcome situation occurs when, for a query protein, it is possible to find (using standard sequence comparison tools) another protein that is highly homologous (30% or larger sequence similarity) and for which the structure has been already solved experimentally.[2–5] In such cases, classical comparative modeling methods allow for the construction of molecular models, whose accuracy is sometimes (depending on the sequence similarity and "completeness" of the sequence alignment) close to that of experimental methods. Another quite typical situation is when sequence methods, or threading procedures, can detect only weakly similar sequences of protein(s) of known structure.[6–10] Consequently, the similarity of the (unknown) three-dimensional structure of the query sequence to the template structure cannot be quantified a priori. The two structures may have identical topology; however they may differ in the details of their loop conformations. Also, particular secondary structure elements may be of different size, and there may be different packing angles between secondary structural elements. Frequently, the actual structural similarity may be limited to only that part of the structure having a common structural motif (or motifs) while the remainder of the structure is completely different.[10,11] In the past, application of comparative modeling tools to such cases led to molecular models that differed substantially from the true structure of the query protein.[5]

Finally, the third possible situation occurs when the (unknown) fold of the query protein is significantly different from any of the known protein folds or when the existing sequence comparison and threading tools are unable to detect the proper structural template. In such a situation, one needs to apply *ab initio* type approaches of protein structure prediction.[12–15] Although progress in the field of *ab initio* structure prediction has recently become quite significant (as demonstrated by a number of groups during the CASP3 exercise), successful applications remain limited to relatively small proteins with rather simple fold topologies.[15] Moreover, even when successful, *ab initio* methods usually provide structures of low resolution. Quite often such predicted structures can be helpful

Grant sponsor: National Institutes of Health (Division of General Medical Sciences); Grant number: GM-48835.

*Correspondence to: Jeffrey Skolnick: Laboratory of Computational Genomics, Donald Danforth Plant Science Center, 893 North Warson Rd., St. Louis, MO 63141. E-mail: Skolnick@danforthcenter.org or Andrzej Kolinski, Kolinski@chem.vw.edu.pl

Received 9 January 2001; Accepted 15 March 2001

in predicting the biological functions, in particular the biochemical functions, of proteins.[1,16–19]

The focus of this article is on those cases that fall within the second protein structure prediction category outlined above. For ~30–50% of protein sequences in a newly sequenced genome, it is possible to detect a weakly homologous protein of known structure.[20–22] Our goal is to provide a method that enables construction of moderate-resolution molecular models for these sequences, in spite of the often significant structural differences between the detected template and the actual structure of the query protein whose structure is unknown.

To achieve this goal, computational tools capable of meeting the following requirements are necessary: First, one needs an efficient threading algorithm (or a combination of a sequence-based approach and threading) that detects plausible template proteins and provides an alignment of as good as possible quality.[9] The meaning of "good-quality alignment" in the context of this approach is discussed in some detail below. Let us point out that the structural template provided by the alignment of the query sequence to the structure of a known protein does not need to be complete (a substantial portion of the target proteins may remain undefined).

Moreover, the template can differ substantially from the equivalent portion of the (unknown) probe structure; in many cases, a template/probe root-mean-square deviation (RMSD) (coordinate from native after the best superposition) in the range of 10 Å remains an acceptable starting conformation. This leads to the second requirement. Namely, a molecular modeling tool must be developed that is capable of rearranging the initial threading-based model in such a way that the final structure is closer (sometimes much closer) to the query protein's "true" structure than it is to the template. This can be achieved only when the employed modeling tool permits very efficient sampling of protein conformational space (large-scale structural rearrangements) and when the force field of the employed molecular model is capable of selecting native-like structures of the query protein (at least when the search is limited to a portion of conformational space in a neighborhood of the template that is wide enough to comprise the query protein's structure). Finally, a means of a priori estimating the expected accuracy of the obtained molecular models needs to be provided.

Recently, we reported a method for the improvement of threading-based molecular models of proteins.[23] Using Monte Carlo simulations of a lattice model of polypeptide chains in which the conformations were restricted to a "tube" surrounding the template structure, it was possible to achieve substantial improvement in about 50% of the threading-based models in a small set of 12 test proteins. Although the goal of this work is similar, and we use essentially the same protein representation, there are also qualitative differences between that work and the present approach. First, we employ a different, recently developed threading algorithm called PROSPECTOR.[9] This new threading algorithm (described elsewhere) has three features that are very important for the present approach. It detects even remotely related pairs of proteins and produces very good alignments, sometimes close to the best structural alignments. Second, the force field of the lattice model has been refined and a more efficient Monte Carlo sampling scheme has been adopted. Third, the threading algorithms are used to derive a large set of restraints employed in the modeling,[24] as described below.

The threading template coordinates comprise a subset of the restraints. Threading-based predictions of tertiary contacts and local chain geometry are incorporated into the refinement algorithm. In addition, pair potentials and short-range restraints (distance restraints reflecting secondary structure preferences) are derived from a statistical analysis of sequentially similar protein fragments. These subsets of potentials and restraints are complementary, and they encode structural biases of very different origins.

To address the issues of fold identification, probe–template alignment and subsequent refinement, the modeling procedure consists of a hierarchy of sequence- and structure-based threading algorithms, Monte Carlo simulations,[23] distance–geometry-based averaging, as well as clustering of the lattice models, with subsequent construction of a detailed atomic model. In spite of the variety of computational tools used, the entire procedure is quite straightforward, relatively fast, and easy to use in an automated fashion for large-scale protein structure prediction.

It should be noted that the approach can also be employed in those cases where threading, or sequence comparisons, fails to detect a related global fold (or folds) but indicates possible local structural similarity to protein(s) of known structure. In such cases, a small structural motif (a long helix, helical hairpin, fragment of a β-sheet) can be used as a modeling "template." Such a template provides a folding scaffold, thereby reducing the conformational space to be searched in order to assemble the remaining portions of the structure of the query protein. Thus, there is a continuous transition from a type of comparative modeling (when the highly homologous structural template can be detected), through folding in a restricted space around a fragmentary template provided by the threading algorithm to weakly restrained essentially ab initio folding. As would be expected, the success rate of correct fold assembly and the average accuracy of the obtained molecular models decrease with the decaying quality and length of the template protein alignment. Reasonably good templates enable large, even multidomain proteins to be modeled, while the ab initio dominant approach is limited to small (no larger than 100–140 residues), single-domain proteins.

In this work, we describe the proposed methodology and analyze its performance using the Fischer database[25] as a test set. This test set is commonly used to benchmark threading approaches and is probably quite representative of larger sets of proteins. The similarity level of the related pairs of proteins from this database ranges from closely to very remotely related. Of course, our purpose is not only to detect related pairs of proteins, but also to test the ability

to obtain good molecular models for a substantial fraction of the test proteins. In this context, we briefly discuss the perspectives for the large-scale structural (and functional) annotation of genomic data. In a forthcoming publication, the method will be applied to structural annotation of proteins from a small genome, *M. genitalium*.[26]

METHODS

The present approach is a combination of a number of recently developed methods for protein recognition,[9] *ab initio* folding, and comparative modeling.[24] An overview is presented in Figure 1. Detailed descriptions and analysis of some of the components have been recently published.[9,23,24] For these algorithms, a brief summary is presented for the reader's convenience. Newer and less documented methods and algorithms are described in more detail.

The entire procedure consists of the following subsequent steps:

1. Thread the query sequence through the structural database and derive the sequence-specific long-range and short-range potentials.[9]
2. Build the starting lattice model using the partial threading alignment as a structural scaffold. As previously, the side-chain-only (SICHO) lattice model is used.[27]
3. Optimize the lattice model by the replica exchange Monte Carlo (REMC) method, using the template as a source of weak spatial restraints.[28,29]
4. Calculate the average lattice model by means of a clustering[30] or distance geometry (DG) procedure.[31]
5. Rebuild the atomic details (optional).

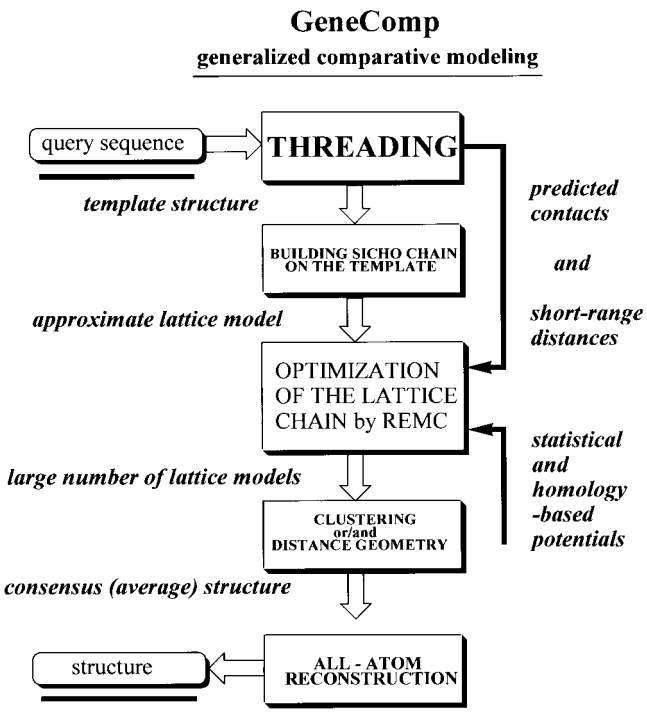


Fig. 1. Overview of the GENECOMP method. Starting from a sequence one first does multiple sequence alignments and then threads the sequence through a library of known protein structures. If a template is found, threading not only provides an alignment of the probe sequence to the template, but also of predicted contacts and secondary structure that may also include nonaligned regions. After an initial lattice model is built, *ab initio* folding in the vicinity of the template is done, and the structure is selected by either distance geometry or clustering, with the latter providing somewhat better models. Finally, atomic detail is added.

Although this procedure may appear somewhat complex, it is relatively straightforward and is easy to automate.

Threading

Our new threading approach, PROSPECTOR, uses a set of close and distant sequence profiles to generate first pass alignments.[9] The second pass for each also uses multiple-sequence-averaged pair potentials and secondary structure propensities, where the partners for the evaluation of the pair interactions are extracted from the alignment generated by the respective first pass sequence profiles. For the top 20 scoring structures (four scoring functions times the best five scoring functions for each structure) if a contact is present in 25% of the structures, it constitutes a predicted contact. Then, using a previously derived formalism,[32] these predicted contacts are also converted to a threading-based, protein-specific pair potential that is used in a subsequent iteration of threading, PROSPECTOR2. Additional predicted contacts are then collected; the threading-based, protein-specific pair potential, PAIR2, is recalculated using these new contacts, and the process is iterated for a third time in PROSPECTOR3. We term the resulting pair potential PAIR3. The resulting set of contacts from PROSPECTOR1–3 is pooled to form the predicted contacts used in subsequent simulations. We have found that, on average, PROSPECTOR3 gives the best alignments, as well as the best set of predicted local distances. The final template has the best Z-score between the distant sequence profile-based threading alignment and the alignment generated from the combined sequence profile, PAIR3, and the secondary structure profile. The entire set of predicted contacts are used as tertiary restraints. PROSPECTOR3 also provides a set of local chain geometry predictions that are extracted from the average geometry from the top scoring structures and that are subsequently incorporated into the lattice-based folding algorithm.

Multiple sequence alignments and the threading of short test sequence fragments are also used to derive protein-specific, additional short-range distance restraints (up to several residues along the chain), as well as orientation-dependent, protein-specific pair potentials.

Building the Starting Lattice Model

Proteins are modeled as lattice chains connecting vertices corresponding to the centers of mass of the "side"-chains (in reality, the center of mass of the side-chain heavy atoms plus the Cα). This model is named the side-chain-only, or SICHO, model. The grid of the underlying simple cubic lattice is equal to 1.45 Å. The distribution of distances between two subsequent chain units mimics

the distribution seen in the structural databases of proteins. The various distances reflect the different amino acid sizes, different conformations of the main chain and different side-chain rotamers (when applicable). Any given protein structure can be fitted to the corresponding lattice model with an average accuracy of ~0.8 Å RMSD. The model force field contains generic protein-like terms that convert the random coil into an average protein. Included in these generic terms are local stiffness, hydrogen bonding, and side-chain packing terms that generate protein-like side-chain packing patterns. There are also sequence-specific terms that reflect local conformational preferences, local side-chain burial, and orientation dependent pair interactions. Additional details about this lattice representation of proteins, a description of the modeling of protein dynamics and the basic force field of the model can be found in several recent publications.[23,24,27]

The alignment of the query sequence to the template structure is usually incomplete and contains insertions and gaps. The building procedure for the starting model takes this into consideration. First, the aligned residues of the template are projected onto the lattice, with the proper (lattice model-based) restrictions for the distances between the two subsequent units and planar angles for three subsequent units. An excluded-volume envelope (the distance of closest approach for two model residues is 3 lattice units, i.e., 4.35 Å) is then built around the alignment. In the next step, the nonaligned parts of the chains are successively added, taking into account the excluded volume of the already existing chain. In those cases when the number of residues of the query sequence is too small to connect a gap in the template, the closest fragments of the template are relaxed to accommodate such an alignment artifact. Nonaligned ends are attached in a semirandom fashion. For "very good" alignments, this procedure builds a quite accurate lattice model of the query protein. When the alignment becomes worse and covers a small fraction of the test sequence, the starting model could be very far from the target structure.

Restrained Lattice Folding—Optimization of the Initial Model

Lattice folding occurs in a restrained space using the Replica Exchange Monte Carlo method as the conformational search tool.[28] For this purpose, a number of copies of the initial model are created and placed at various temperatures, according to the REMC scheme. This Monte Carlo simulation consists of two parts: In the first stage, a short annealing run is performed using rather high temperatures in dimensionless units (T = 2.5–1.5). Then, in the second stage, the temperature range is set to T = 2.0–1.0 and a 5–10-times longer run compared to the first stage is performed. One simulation trajectory takes 1–2 days on a Pentium III 733 MHz PC. Twenty copies guarantee a very fast and efficient swapping of conformations among the various temperature levels (the temperature increment between replicas has been assumed temperature independent—a linear temperature set). Those conformations seen at the lowest temperature of the REMC scheme rapidly find energy minima. The minimum in many cases (as shown later) corresponds to near-native conformations. REMC proves to be much more efficient and faster than the conventional simulated annealing procedure that we recently used in a very similar context of lattice-based comparative modeling. Another difference between the previous and the present attempts to improve threading-based protein models is in the way the template and other restraints are implemented.

Previously, sampling was restricted to a "tube" surrounding the template structure.[23] The lattice model was free to move inside this tube and only occasionally to leave the tube, as a result of a substantial energy penalty. Consequently, the degree of improvement of the initial model was limited. Here, we apply a much softer and more diverse set of restraints for the lattice folding. Three sets of restraints are applied during the folding/optimization procedure.

The first set is associated with the template. For an aligned residue where the equivalence between the target and template residues was established by threading, the following potential is used:

$$V_{\text{templ}} = V_5 + V_1 \tag{1a}$$

with

$$V_5 = 0 \quad \text{for}$$

$$r_{\min} < 2 \text{ (in lattice units or 2.9 in Ångstroms)} \tag{1b}$$

$$V_5 = 0.5 \cdot \varepsilon_{\text{rest}} \cdot r_{\min} \quad \text{for} \quad r_{\min} \geq 2 \tag{1c}$$

and

$$V_1 = -\varepsilon_{\text{rest}} \quad \text{for} \quad r_{i,j} < 2 \tag{1d}$$

where $r_{i,j}$ is the distance between the ith Cα of the template structure and the jth Cα of the modeled target structure. The first index corresponds to the template residue and the second to the probe, and the aligned pairs have, by definition, the same indices, and $r_{\min} = \min\{r_{i-2,j}\ r_{i-1,j}, r_{i,j}, r_{i+1,j}, r_{i+2,j}\}$, the smallest distance between the target Cα and the five-residue Cα fragments of the template. The last condition allows for two-residue shifts of the target chains along the template structure, enabling "corrections" of the initial alignment. ϵ_{rest} is a constant scaling factor that sets the strength of the restraints (see below).

The second set of the restraints originates from the contact prediction procedure. Only a fraction of predicted contacts are exact, i.e., they are native for the target. A much larger fraction of the predicted contacts are "almost" correct, i.e., they are shifted by ±1 or ±2 residues with respect to native. This was taken into consideration in the design of the restraint potential, i.e.,

$$V_{\text{cont}} = -\varepsilon_{\text{rest}c} \quad \text{for} \quad d_{\min} < 5 \text{ (in lattice units)}$$

$$+ \varepsilon_{\text{rest}c} \cdot (d_{\min} - 6) \quad \text{for} \quad d_{\min} > 6 \text{ (in lattice units)} \tag{2}$$

where $d_{\min} = \min\{r_{i\pm k, j\pm k}, k = 0,1,2\}$, and the index (i,j) is the predicted contact in the template structure. The value

TABLE I. Comparison of the Accuracy of the Models for a "Tuning" Set of 12 Small Proteins Produced by GENECOMP With the Previous Version of Generalized Comparative Modeling*

Probe/ template proteins	GENECOMP (PROSPECTOR + lattice modeling + DG averaging)		Previous comparative approach	Generalized modeling
1aba_/1ego_	4.75	(90.8)	4.86	(79.3)
1bbhA/2ccy_	3.07	(93.9)	6.82	(88.5)
1cewI/1molA	7.79	(70.4)	14.38	(63.9)
1hom_/1lfb_	1.57	(97.7)	3.70	(58.8)
1stfI/1molA	7.07	(69.5)	5.95	(84.7)
1tlk_/2rhe_	3.42	(95.8)	4.17	(83.5)
256bA/1bbh_	2.44	(84.9)	4.36	(98.0)
2azaA/1paz_	7.87	(62.8)	10.77	(62.0)
2pcy_/2azaA	4.03	(88.9)	4.41	(94.5)
2sarA/9rnt_	5.76	(91.7)	7.83	(76.0)
3cd4_/2rhe_	7.15	(92.8)	6.39	(85.4)
5fd1_/2fxd_	11.99	(55.7)	12.40	(65.1)

*Root-mean-square deviations (RMSD) from native in Ångstroms (% of the length of the alignment to the template).

of the ε_{restc} is scaled relative to the number N_c of predicted contacts as follows:

$$\varepsilon_{restc} = \varepsilon_{rest} \qquad \text{for} \qquad N_c < N \tag{3a}$$

where N is the number of residues in the target protein and:

$$\varepsilon_{restc} = \varepsilon_{rest} \cdot N/N_c \qquad \text{for} \qquad N_c \geq N \tag{3b}$$

The positive part of the above potential enters into the total energy when it exceeds a threshold value $N_c/2$, allowing for the significant violation of a small fraction of the restraints.

The third set of restraints contains the target distances predicted from the fragment threading procedure. The corresponding potential could be expressed as follows:

$$\begin{aligned} V_{\text{dist}} = -\varepsilon_{restd} \qquad & \text{for} \qquad |r_{i,j} - R_{i,j}| < 1 \quad \text{(in lattice units)} \\ & + \varepsilon_{restd}\, 4(|r_{i,j} - R_{i,j}| - 1)/(4 + |i-j|) \\ & \text{for} \qquad |r_{i,j} - R_{i,j}| \geq 1 \quad \text{(in lattice units)} \end{aligned} \tag{4}$$

where r denotes the actual distance and R the predicted one. Those terms corresponding to distances that could be larger than the diameter of the target protein, as estimated from the distance $|i-j|$ along the chain, are ignored. Similar to the contact restraints, the strength of the distance restraints is scaled to account for the various numbers, N_d, of predicted restraints

$$\varepsilon_{restd} = \varepsilon_{rest} \qquad \text{for} \qquad N_d < N \tag{5a}$$

where N is the number of residues in the target protein and:

$$\varepsilon_{restd} = \varepsilon_{rest} \cdot N/N_d \text{ (for } N_d \geq N) \tag{5b}$$

The positive part of the above potential enters into the total energy when it exceeds a threshold value of $N_d/5$, allowing for the significant violation of a small fraction of the restraints. This reflects the structure of the data for predicted distances, which are similar to the contact-based restraints. Most of them are almost exact, while a fraction of the predicted set could be qualitatively wrong.

The total energy of the restraints is the sum of the above three components for all relevant residues (aligned) or pairs of residues (predicted contacts or distances). There is one adjustable parameter ε_{rest} in this scheme. The results are not very sensitive to the specific value; however, in the context intrinsic to the model force field, a value of ~0.5 appears to be close to optimal.

Computing an Average Structure by Means of Distance Geometry

As a result of deficiencies in the force field, the optimal structure prediction in a folding simulation is not always the lowest-energy structure. To this end, it is preferable to obtain an average or consensus structure among the low-energy folds. Thus, from each of 20 simulations in the second pass described above, 200 conformations were stored in a constant interval of simulation time. The collected structures were averaged using a two-step distance geometry, DG, procedure.[31] After the first pass, those structures far away from the average were rejected, and the final DG conformation was constructed from the remaining set of structures. Interestingly, DG averaging always led to a lower RMSD from the native than the average RMSD for the original set of conformations from the lattice simulations. Sometimes the structures from DG were close to the best structures seen in the folding simulations.

Computing an Average Structure by Clustering

Folding simulations near a template can generate clusters of structures with significant differences between them as compared to the differences between the structures within each cluster.[30] When present, the different clusters arise mainly from the nonaligned portions in the low energy folds. Therefore, it is sometimes useful to cluster the structures into groups of different folds before obtaining the average structures.

The clustering of structures is carried out through a partitioning clustering technique.[30] This method arranges a collection of folds in a multidimensional space defined by a metric given by the relative root-mean-square deviation (RRMSD). The RRMSD is defined as the RMSD divided by a quantity that depends on the radius of gyration of the two structures involved, which when applied to random structures has a mean value approaching one as the chain length increases. The clusters are initially selected by determining the structures with a high probability of being at the center of the cluster, and assigning to them the structures that are significantly close to the center. The centroids for each cluster are determined by optimally aligning the structures in each cluster and then computing their average. The clusters are then refined by an iterative

process that consists of centroid calculations followed by the recalculation of the cluster members until a measure of cluster quality is optimized. The resulting clusters are compared to eliminate redundant ones. Finally, the centroid structures are refined by minimizing a harmonic potential constructed from the average distances and standard deviations between every pair of residues for each cluster. This final step is similar to the distance geometry method.

For the problem at hand, the structures in each trajectory are clustered to eliminate or reduce unwanted correlations. The resulting centroids from all trajectories are clustered once again to determine the significant folds. Because the structures are significantly similar due to the fact that we are doing folding in the vicinity of a template, the criterion that determines the size (in terms of RRMSD) of each cluster is determined based on the distribution of the RRMSD between each pair of structures. In particular, two structures with an RRMSD above the average plus two standard deviations of the RRMSD distribution are not allowed in the same cluster. The average energy of the structures of the cluster is assigned to the centroids and used to rank-order them.

Rebuilding the Atomic Details

A fast procedure was designed for reconstruction of the atomic details from the known positions of the Cα and the side-chains. (Given the side-chain center of mass position, it is quite easy to obtain a quick, approximate location of the Cα) The only constraints are the positions of the side-chain centers of mass. The initial local Cα trace geometry built as a geometric function of the side-chain centers of mass is not perfect. Therefore, the positions of Cα are optimized in the first step. This is done by a gradient optimization procedure using a very simple force field. There are several harmonic terms in the force field, including the distance between consecutive Cα atoms, the distance between the Cα atom and the side-chain center of mass, and a term that regularizes the angular correlation of the Cα. Thus, an improvement in local geometry occurs. In the next stage, the positions of the backbone atoms are reconstructed according to the local Cα trace conformation. In this step, the vector normal to the plane defined by three consecutive Cα is calculated. This vector is almost parallel to the peptide bond plane. Thus, the remaining atoms of the peptide bond can be positioned quite accurately. Next, the positions of the side-chain atoms are rebuilt. The conformations of the side-chains are chosen from a representative rotamer database. For rigid amino acids (e.g., phenylalanine), there is a single conformation in the database. There are up to 20 conformations for large, flexible side-chains (e.g., lysine). The conformation of the rotamer depends on the distance between the Cα atom and the center of mass of the side-chain, and the local chain conformation (i.e., the Cα–Cα–Cα angle). Next, as a final stage of the reconstruction procedure, the side-chains are rotated around a virtual Cα center of mass bond to avoid excluded volume conflicts. We also note that an alternative, but slower method of comparable accuracy was previously developed.[33]

This procedure yields reasonable structures; however, the packing of side-chains after the all-atom reconstruction is not optimized. This could be accomplished using one of the standard molecular mechanics procedures. Here, since we are focusing on the quality of the models generated by the GENECOMP procedure, this step has been omitted.

Test Proteins

Two sets of proteins were selected for the purpose of tuning and benchmarking the modeling method described in the previous section. The first set, containing 12 pairs of target and template proteins, is identical to the set analyzed in our previous work on the application of lattice models for the refinement of threading models.[23] For some of these test pairs, the new threading method, PROSPECTOR,[9] detects different template structures. However, for the purpose of comparison, the same templates as before were used, regardless of whether their threading scores are lower. The set of 12 proteins was also used to tune the scheme proposed here for the implementation of restraints in the lattice model. The second set of proteins was generated by the threading procedure applied to the Fischer database of sequences and structures.[25]

Application to the set of 12 proteins and comparison with previous work

The test set of 12 proteins was used to "tune" the strength (and their functional form) of various restraints employed in the lattice model. This was done by comparing the average quality of the models resulting from a series of simulations with various scaling factors of particular restraints generated by PROSPECTOR (see the previous sections). The results of the modeling for the present version of GENECOMP are compared with the pervious version of the generalized comparative modeling in Table I. The present GENECOMP models are better in 10 of 12 test cases. In five cases, the improvement of the models is of a qualitative nature. Improvement of the models is not only attributable to refined lattice modeling, but to the DG or cluster averaging of the final models and to the (on average) better starting models (alignments to templates) provided by PROSPECTOR as well.

Application to the Fischer database

The list of 68 target–template protein pairs in the Fischer database is shown in Table II together with the nomenclature of structure type assigned by SCOP.[25] This standard database contains a wide variety of structural types: 13 α proteins, 27 β, 18 α/β, 8 α+β, and 2 small proteins (which have small secondary structure content). The lengths of the proteins vary from 62 to 581 amino acid residues. Note that the correct template protein that has the best structural superposition in the Fischer database to the probe structure is always used for all the probe proteins, even when PROSPECTOR fails to assign the correct template in the first position (PROSPECTOR cor-

TABLE II. Target/Template Protein Sets in the Fischer Database

Target protein			Template protein		
PDB code	Name	Length	Structure type by SCOP	PDB code	Length
1aep_	Apolipopholin III	153	α; apolipopholin III	256bA	106
1bbhA	Cytochrome C	131	α; 4 helical up and down bundle	2ccyA	127
1bgeB	Granulocyte colony-stimulating factor	159	α; 4 helical cytokines	1gmfA	119
1c2rA	Cytochrome C2	116	α; cytochrome	1ycc	108
1cpcL	C-phycocyanin	172	α; globin-like	1colA	197
1dsbA	Disulfide bond formation protein	188	α; disulfide-bond formation facilitator (DSBA), insertion domain	2trxA	108
1dxtB	Hemoglobin	147	α; globin-like	1hbg_	158
1hom_	Antennapedia protein	68	α; DNA/RNA binding 3 helical bundle	1lfb_	77
1lgaA	Lignin peroxidase	343	α; heme-dependent peroxidase	2cyp_	293
1osa_	Calmodulin	148	α; EF hand-like	4cpv	108
1rcb_	Interleukin-4	129	α; 4 helical cytokines	1gmfA	119
2hpdA	Cytochrome P-450	457	α; cytochrome P-450	2cpp_	405
2sas_	Sarcoplasmic calcium-binding protein	185	α; EF hand-like	2scpA	174
1aaj_	Amicyanin	105	β; cupredoxins	1paz_	120
1arb_	Achromobacter protease I	263	β; trypsin-like serine protease	4ptp_	223
1bbt1	Foot-and-mouth disease virus	186	β; viral coat and capsid	2plv1	288
1cauB	Canavalin	184	β; double-stranded β-helix	1cauA	181
1cid_	CD4	177	β; immunoglobulin-like β sandwich	2rhe_	114
1fc1A	Immunoglobulin FC fragment	207	β; immunoglobulin-like β sandwich	2fb4H	229
1ltsD	Heat-labile enterotoxin	103	β; OB-fold	1bovA	69
1mdc_	Fatty acid binding protein	132	β; lipocalin	1ifc_	131
1mup_	Major urinary protein	157	β; lipocalin	1rbp_	174
1pfc_	Immunoglobulin p/Fc fragment	111	β; immunoglobulin-like β sandwich	3hlaB	99
1sacA	Serum amyloid component	204	β; Con A-like lectin/glucanases	2ayh_	214
1ten_	Tenascin	90	β; immunoglobulin-like β sandwich	3hhrB	195
1tie_	Erythrina trypsin inhibitor	166	β; β trefoil	4fgf_	124
1tlk_	Telokin	103	β; immunoglobulin-like β sandwich	2rhe_	114
2azaA	Azurin	129	β; cupredoxins	1paz_	120
2afnA	Nitrite reductase	331	β; cupredoxins	1aozA	552
2fbjL	Immunoglobulin FAB fragment	213	β; immunoglobulin-like β sandwich	8fabB	214
2mtaC	Methylamine dehydrogenase	147	β; cupredoxins	1ycc_	108
2omf_	OMPF porin	340	β; transmembrane β barrels	2por_	301
2pia_	Phosphatidylinositol 3-kinase	321	β; reductase/isomerase/elongation factor common domain	1fnr_	296
2sga_	Proteinase A	169	β; trypsin-like serin proteases	4ptp_	223
2sim_	Sialidase	381	β; 6 bladed β propeller	1nsbA	390
2snv_	Sindbis virus capsid protein	151	β; trypsin-like serine proteases	4ptp_	223
3cd4_	CD4	97	β; immunoglobulin-like β sandwich	2rhe_	114
3hlaB	Class I histocompatibility antigen A2.1	99	β; immunoglobulin-like β sandwich	2rhe_	113
4sbvA	Southern bean mosaic virus coat protein	199	β; viral coat and capsid	2tbvA	286
8i1b_	Interleukin-1 β	146	β; β trefoil	4fgf_	124
1aba_	Glutaredoxin	87	α/β; thioredoxin fold	1ego_	85
1atnA	Doxyribonuclease I	372	α/β; ribonuclease H-like motif	1atr_	383
1chrA	Chloromuconate cycloisomerase	370	α/β; mandelate racemase	2mnr_	357
1crl_	Lipase	534	α/β; α/β hydrolase	1ede_	310
1eaf_	Dihydrolipoyl transacetylase	243	α/β; CoA-dependent acyltransferases	4cla_	213
1gal_	Glucose oxidase	581	α/β; FAD/NAD(P) binding domain	3cox_	500
1gky_	guanylate kinase	186	α/β; p-loop containing nucleotide tryphosphate hydrolase	3adk_	194
1gp1A	Glutathione peroxidase	184	α/β; thioredoxin fold	2trxA	108
1hrhA	Ribonuclease H domain of HIV-1 reverse transcriptase	125	α/β; ribonuclease H-like motif	1rnh_	148
1mioC	Nitrogenase molybdenium–iron protein	525	α/β; nitrogenase iron molybdenium protein α- and β-chains	3minB	522
1npx_	NADH peroxidase	447	α/β; FAD/NAD(P) binding domain	3grs_	461
1tahA	Lipase	318	α/β; α/β-hydrolase	1tca_	317
2ak3A	Adenylate inase isoenzyme-3	226	α/β; p-loop containing nucleotide tryphosphate hydrolases	1gky_	186
2cmd_	Malate dehydrogenase	312	α/β; NAD(P) binding Rossmann fold	6ldh_	329
2gbp_	D-Galactose/D-glucose binding protein	309	α/β; periplasmic binding protein-like I	2liv_	344
2mnr_	mandelate racemase	357	α/β; TIM α/β barrel	4enl_	436
3chy_	cheY	128	α/β; flavodoxin-like	4fxn_	138
3rubL	ribulose 1,5-bisphosphate carboxilase/ oxygenase	442	α/β; TIM α/β barrel	6xia_	387
1cewI	Cystatin	108	α + β; cystatin-like	1molA	94
1fxiA	Ferredoxin I	96	α + β; β-grasp(ubiquitin-like)	1ubq_	76
1onc_	P-30 protein	104	α + β; RNase-like	7rsa	124
1stfI	Inhibitor stefin B	95	α + β; cystatin-like	1molA	94
2hhmA	Inositol monophosphatase	272	α + β; sugar phosphatases	1fbpA	316
2pna_	Phosphatidylinositol 3-kinase	104	α + β; SH2-like	1shaA	103
2sarA	Ribonuclease SA	96	α + β; microbial ribonucleases	9rnt_	104
5fd1_	Ferredoxin	106	α + β; ferredoxin-like small proteins; high-potential iron protein (HIPIP)	2fxb_	81
1hip_	Oxidized high-potential iron protein	85		2hipA	71
1isuA	High-potential iron–sulfur protein	62	Small proteins; HIPIP	2hipA	71

TABLE III. Summary of Contact Prediction Results for the Fischer Database

Name	N_c[a]	δ = 0[b]	δ = 1[b]	δ = 2[b]	δ = 3[b]	δ = 4[b]
1aaj_	84	0.64	0.8	0.92	0.94	0.99
1aba_	59	0.53	0.68	0.76	0.88	0.9
1aep_	18	0	0.06	0.22	0.5	0.67
1arb_	11	0.36	0.82	0.82	0.91	1
1atnA	134	0.28	0.54	0.75	0.82	0.86
1bbhA	58	0.5	0.59	0.64	0.72	0.83
1bbt1	15	0.2	0.27	0.73	0.87	0.93
1bgeB	9	0.67	0.78	1	1	1
1c2rA	88	0.65	0.83	0.91	0.95	0.95
1cauB	94	0.56	0.74	0.83	0.9	0.93
1cewI	34	0.44	0.71	0.76	0.85	0.88
1chrA	279	0.52	0.77	0.89	0.95	0.98
1cid_	36	0.42	0.56	0.72	0.83	0.89
1cpcL	54	0.07	0.37	0.48	0.57	0.69
1crl_	67	0.28	0.42	0.57	0.72	0.81
1dsbA	21	0.29	0.43	0.57	0.62	0.71
1dxtB	132	0.67	0.81	0.83	0.95	0.98
1eaf_	145	0.45	0.66	0.75	0.86	0.93
1fc1A	50	0.66	0.82	0.88	0.92	0.96
1fxiA	43	0.16	0.47	0.65	0.79	0.86
1gal_	217	0.6	0.77	0.86	0.9	0.93
1gky_	75	0.35	0.6	0.72	0.81	0.83
1gp1A	9	0	0.11	0.33	0.56	0.78
1hip_	52	0.62	0.75	0.87	0.9	0.94
1hom_	47	0.49	0.66	0.85	0.91	0.96
1hrhA	54	0.28	0.65	0.8	0.91	0.94
1isuA	31	0.26	0.61	0.9	0.94	0.94
1lgaA	145	0.62	0.82	0.86	0.95	0.97
1ltsD	31	0.16	0.45	0.77	0.87	0.9
1mdc_	14	0.36	0.5	0.64	0.64	0.64
1mioC	178	0.54	0.78	0.86	0.9	0.96
1mup_	98	0.57	0.8	0.91	0.97	0.98
1npx_	237	0.62	0.74	0.86	0.93	0.95
1onc_	99	0.68	0.85	0.92	0.95	0.96
1osa_	116	0.43	0.53	0.62	0.67	0.68
1pfc_	102	0.31	0.55	0.76	0.82	0.87
1rcb_	41	0.46	0.66	0.78	0.85	0.98
1sacA	67	0.16	0.31	0.45	0.54	0.66
1stfI	22	0.45	0.45	0.64	0.77	0.82
1tahA	12	0.42	0.92	0.92	1	1
1ten_	42	0.12	0.6	0.71	0.86	0.9
1tie_	28	0.43	0.64	0.75	0.79	0.82
1tlk_	65	0.74	0.83	0.94	0.97	0.98
2afnA	31	0.06	0.35	0.45	0.68	0.71
2ak3A	92	0.23	0.45	0.62	0.74	0.79
2azaA	105	0.15	0.28	0.53	0.6	0.7
2cmd_	230	0.58	0.75	0.83	0.94	0.96
2fbjL	47	0.55	0.7	0.81	0.89	1
2gbp_	72	0.58	0.79	0.89	0.96	0.97
2hhmA	47	0.4	0.66	0.77	0.83	0.96
2hpdA	87	0.48	0.74	0.84	0.91	0.92
2mnr_	45	0.42	0.51	0.71	0.76	0.84
2mtaC	1	1	1	1	1	1
2omf_	17	0.06	0.29	0.47	0.76	0.82
2pia_	14	0.14	0.36	0.64	0.64	0.71
2pna_	26	0.81	0.85	0.88	0.96	0.96
2sarA	39	0.31	0.72	0.9	0.97	0.97
2sas_	155	0.43	0.62	0.7	0.77	0.84
2sga_	0	0.43	0.62	0.7	0.77	0.84
2sim_	10	0.4	0.6	0.6	0.7	0.9
2snv_	30	0.33	0.6	0.7	0.83	0.93
3cd4_	99	0.42	0.58	0.74	0.78	0.83
3chy_	33	0.36	0.73	0.82	0.88	0.97
3hlaB	36	0.53	0.69	0.89	0.89	0.94
3rubL	53	0.06	0.28	0.58	0.77	0.85
4sbvA	58	0.36	0.64	0.84	0.88	0.9
5fd1_	32	0.09	0.34	0.53	0.75	0.75
8i1b_	59	0.46	0.73	0.85	0.92	0.95
average		0.41	0.61	0.75	0.83	0.88

[a] N_c is the total number of predicted contacts.
[b] Fraction of contacts correctly predicted within δ = ± m residues.

rectly places 61 of 68 pairs in the top position). We focus on the correct probe–template pairs because the aim of this article is to demonstrate the ability of the GENECOMP algorithm to refine the initial alignments generated by PROSPECTOR.

Accuracy of Threading-Based Contact Prediction

Table III presents the results of our threading-based approach to contact prediction. On average, 41% of the contacts are correctly predicted and 75% are correctly predicted within ±2 residues of a correct contact. Since contacts are predicted on an iterative basis, when the fraction of contacts is an appreciable fraction of the entire structure, it is quite likely that the probe–template alignment is significant. This is confirmed in Figure 2, where the ratio R_c of the number of contacts to the number of residues exceeds 50%. In 16 of 19 cases, the RMSD of the aligned regions is <8 Å and for 9 cases it is <5 Å. Thus, one can be reasonably confident of a good starting model when R_c is >0.5.

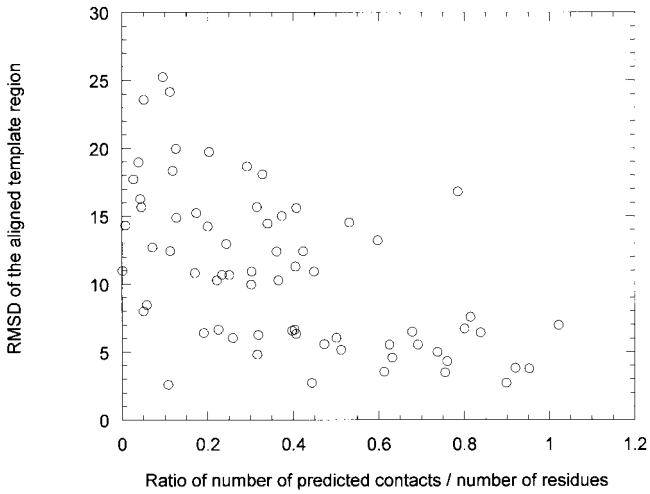


Fig. 2. Root-mean-square deviation (RMSD) of the aligned template region with respect to the probe native structure is plotted as a function of the ratio of the number of predicted contacts to the number of residues in the protein.

Optimization of the SICHO Lattice Model by the Replica Exchange Monte Carlo Method

Table IV presents the properties of the starting structures. As shown in column three, although correct templates are used in the initial threading, the threading alignments do not always cover a large portion of the probe proteins. Indeed, for some cases, this coverage drops around 50%. This is why the RMSD of the aligned region from the native structure and that of the whole protein sometimes differ significantly.

The simulation results are summarized in Table V. We have shown both the lowest-energy structure and the smallest RMSD structure (i.e. the best possible prediction) in the table but, in what follows, the results of the lowest-energy structures are reported unless otherwise indicated, because in general there is no way to guess what the lowest RMSD structure is in a blind prediction. A total of 31 out of 68 targets resulted in structures whose RMSD (whole protein) is <10 Å from the native structures. If only the targets with good threading results (those with <10 Å RMSD from the native are shown with asterisks in Table V) are counted, 29 of 31 remain <10 Å RMSD (exceptions are 1ltsD, which is 10.25 Å and 2sgaA, which is 12.00 Å from native, respectively) and 21 resulted in structures whose RMSD from native is <6 Å.

The dependence of the results on the quality of the initial threading alignment is shown in Figure 3A,B. Figure 3A compares the RMSD of the aligned region obtained from threading with the RMSD of the same region in the lowest-energy structure extracted from the simulations. Here the slope of the best fitting line is 0.96; and some of the lower initial RMSD structures exhibit the most dramatic improvement (lower left-hand corner). Figure 3B plots the RMSD of the entire initial structure versus the RMSD of the entire final structure that is extracted from the clustering algorithm. This is well described by a line with a slope of 1.03 and an intercept of −1.56. This curve provides a crude estimate of the likely global RMSD after refinement. On average, the structures improve, as evidenced by the average of the ratio of the final RMSD to the initial RMSD that is 0.86. For those targets with >10 Å RMSD starting structures, 26 out of 37 (70.3%) showed an improvement in terms of the RMSD of the whole protein. By contrast, for targets with <10 Å RMSD starting structures, 27 out of 31 (87.1%) showed some improvement. Especially, 15 of the 16 targets whose RMSD of the starting structure lies between 6 Å and 10 Å showed a positive improvement, and 6 out of 16 improved by >2 Å. Generally, it may be stated that a significant improvement occurs when the threading result is rather poor but not so bad as to be nonsensical. If the initial threading result is very accurate, there is no significant room for improvement. Another important thing that should be noted here is that in almost all cases, the resulting structure has either improved, or when this is not the case, then the level of deterioration is very small. This observation also holds for the aligned region of the proteins, which are never >1 Å worse than their threading results (Fig. 3A). This degradation of structure quality is insignificant and recoverable by the clustering or DG procedure. In this sense, these simulations do no harm to the threading-based structure and can be applied to all situations with impunity.

Figure 4A–C presents typical examples of trajectories of the three classes of targets. When the threading result is good as in Figure 4A (1tlk_), then both the RMSD and energy do not change much during the simulation. Basically the structure fluctuates near the initial structure. In the case of 2azaA, (Fig. 4B), whose threading result lies in the intermediate range of RMSD, a reduction of both RMSD and energy are observed in the early time steps and the RMSD gradually continues to reduce. Generally, this drop of RMSD in the early stage of the simulation corresponds to the relaxation of the initial structure including that of the unaligned region of the threading template. For 1cid_ (Fig. 4C), which has poor initial threading results,

TABLE IV. Initial Properties*

Target protein	No. of aligned residues	Alignment coverage (%)	RMSD from threading result (aligned region)	RMSD of initial structure (aligned region)	RMSD of initial structure (whole protein)
1aaj_	87	82.86	6.74	8.57	13.37
1aba_	79	90.81	6.52	6.89	7.09
1aep_	98	64.05	18.36	18.73	18.13
1arb_	213	80.99	16.32	16.55	17.93
1atnA	280	75.27	12.42	12.40	14.13
1bbhA	123	93.89	2.74	2.99	3.34
1bbt1	207	97.18	12.73	10.92	10.82
1bgeB	110	62.86	8.50	6.02	10.29
1c2rA	99	85.35	4.35	4.77	5.91
1cauB	147	89.63	5.18	4.33	4.92
1cewI	76	70.37	4.85	5.77	9.59
1chrA	344	92.97	3.50	4.91	5.76
1cid_	99	55.93	19.76	19.99	20.55
1cpcL	140	81.40	15.71	15.27	15.20
1crl_	255	47.75	20.01	20.43	23.04
1dsbA	94	51.65	12.46	11.77	15.39
1dxtB	136	92.52	2.74	2.97	3.62
1eaf_	175	78.13	13.25	11.29	11.37
1fc1A	200	96.62	12.99	13.07	13.08
1fxiA	59	61.46	10.94	11.34	11.84
1gal_	430	74.01	15.03	12.45	14.59
1gky_	159	85.48	6.68	6.70	7.86
1gp1A	104	52.79	8.03	8.31	13.63
1hip_	68	80.00	3.55	3.71	4.92
1hom_	43	97.73	5.56	2.92	2.94
1hrhA	117	86.03	6.59	5.26	8.12
1isuA	59	95.16	6.06	6.32	6.29
1lgaA	246	77.60	12.45	10.49	14.87
1ltsD	59	59.00	9.99	10.29	12.65
1mdc_	128	96.97	2.62	3.17	3.27
1mioC	464	88.38	14.48	14.57	16.07
1mup_	147	93.63	5.56	5.53	5.84
1npx_	412	92.17	14.56	14.59	14.42
1onc_	102	98.08	3.81	3.85	4.00
1osa_	104	70.27	16.84	17.86	20.81
1pfc_	91	89.22	3.84	5.51	6.51
1rcb_	92	71.32	6.28	6.13	7.17
1sacA	156	76.47	18.13	17.89	18.93
1stfI	75	76.53	6.66	5.55	8.52
1tahA	181	56.92	19.00	19.36	19.93
1ten_	84	93.33	5.60	4.70	4.79
1tie_	103	59.88	10.85	10.97	12.49
1tlk_	92	95.83	4.61	4.36	4.37
2afnA	299	95.83	25.27	23.56	23.60
2ak3A	162	78.26	15.63	19.83	19.49
2azaA	81	62.79	7.605	8.78	11.00
2cmd_	299	95.83	5.016	5.24	5.43
2fbjL	201	94.37	10.30	10.59	10.68
2gbp_	242	80.94	10.72	10.57	12.23
2hhmA	195	71.69	15.26	15.81	20.24
2hpdA	378	85.33	6.44	6.11	7.44
2mnr_	341	95.52	14.92	15.06	15.27
2mtaC	96	65.31	14.35	14.53	15.68
2omf_	279	82.06	23.61	23.69	23.47
2pia_	255	79.44	15.72	15.88	18.03
2pna_	27	46.55	10.69	9.19	10.50
2sarA	88	91.67	6.36	6.51	6.94
2sas_	160	86.49	6.45	6.88	7.56
2sga_	179	98.90	11.02	10.92	11.78
2sim_	252	66.14	17.74	22.18	22.29
2snv_	127	84.11	14.28	13.64	14.70
3cd4_	90	92.78	7.02	6.97	7.49
3chy_	111	86.72	6.07	6.20	6.70
3hlaB	74	83.15	10.30	10.02	9.88
3rubL	315	66.04	24.19	23.93	24.45
4sbvA	194	97.49	18.68	18.82	18.75
5fd1_	59	55.66	10.95	11.03	13.80
8i1b_	108	73.97	11.31	12.44	13.44

*All root-mean-square deviation (RMSD) values are in Ångstroms.

TABLE V. Results of Lattice Simulations

Target protein[a]	Lowest-energy structure RMSD (whole protein)	Lowest-energy structure RMSD (aligned region)	Smallest RMSD structure RMSD (whole protein)	Smallest RMSD structure RMSD (aligned region)
*1aaj_	8.42	6.17	6.15	4.92
*1aba_	5.58	5.62	3.55	3.31
1aep_	18.34	19.23	18.32	17.98
1arb_	17.30	16.62	15.78	15.78
1atnA	13.33	12.06	12.00	11.22
*1bbhA	3.65	3.38	2.71	2.53
1bbt1	10.81	10.92	9.57	9.33
*1bgeB	6.27	6.02	5.04	4.93
*1c2rA	5.37	4.77	4.31	3.85
*1cauB	5.69	4.47	4.04	3.63
*1cewI	7.35	4.35	4.10	3.54
*1chrA	5.11	3.79	3.77	3.36
1cid_	18.64	18.57	14.05	13.55
1cpcL	13.15	13.02	12.30	12.25
1crl_	24.21	20.13	21.35	19.67
1dsbA	15.94	13.01	11.58	8.13
*1dxtB	3.53	3.15	2.91	2.60
1eaf_	10.09	12.65	9.27	12.06
1fc1A	12.89	12.97	12.43	12.50
1fxiA	10.28	10.80	8.53	9.20
1gal_	17.00	14.44	14.05	11.58
*1gky_	7.76	6.27	6.13	5.75
1gp1A	14.75	8.98	9.08	9.63
*1hip_	4.86	4.40	3.92	3.33
*1hom_	5.00	4.48	1.50	2.82
*1hrhA	5.50	5.11	4.90	4.25
*1isuA	4.23	4.31	3.20	3.19
1lgaA	17.13	12.51	13.10	11.81
*1ltsD	10.25	9.95	8.11	7.79
*1mdc_	3.12	2.95	2.55	2.51
1mioC	15.19	14.35	14.05	13.57
*1mup_	4.46	4.55	4.14	4.20
1npx_	13.75	13.97	13.61	13.76
*1onc_	3.53	3.29	3.08	3.01
1osa_	17.57	17.34	16.56	16.18
*1pfc_	4.48	4.25	3.81	3.63
*1rcb_	5.52	5.06	3.91	3.50
1sacA	18.21	17.43	16.89	15.51
*1stfI	7.38	4.84	4.97	4.40
1tahA	21.60	18.57	18.90	18.28
*1ten_	3.96	4.01	3.16	3.17
1tie_	12.88	12.82	10.74	10.73
*1tlk_	3.19	3.33	2.35	2.35
2afnA	23.23	24.83	22.60	24.20
2ak3A	15.51	15.44	14.65	14.57
*2azaA	8.40	6.85	6.33	5.70
*2cmd_	4.74	4.72	4.22	4.19
2fbjL	8.67	8.67	7.77	7.52
2gbp_	10.46	9.64	9.50	8.96
2hhmA	17.09	14.98	15.22	13.30
*2hpdA	6.07	5.48	5.41	5.14
2mnr_	14.07	13.98	13.55	13.55
2mtaC	15.84	13.98	14.04	12.84
2omf_	23.51	23.06	21.82	21.99
2pia_	17.29	15.25	15.64	14.47
2pna_	11.31	9.52	7.27	7.80
*2sarA	5.93	5.70	4.88	4.77
*2sas_	5.78	5.75	5.51	5.31
*2sga_	11.87	12.00	9.78	10.68
2sim_	19.79	17.06	16.52	15.50
2snv_	13.72	13.23	12.78	12.09
*3cd4_	7.26	6.93	5.98	5.63
*3chy_	4.53	4.53	3.58	3.30
3hlaB	8.96	9.18	4.72	4.59
3rubL	24.19	23.72	22.26	21.73
4sbvA	18.12	18.19	17.73	17.75
5fd1_	11.64	10.52	10.70	9.34
8i1b_	12.58	12.01	10.77	10.59

*Asterisk (*) indicates those proteins for which the lowest-energy structure has an RMSD of <10 Å from native. All root-mean-square deviation (RMSD) values are in Ångstroms.

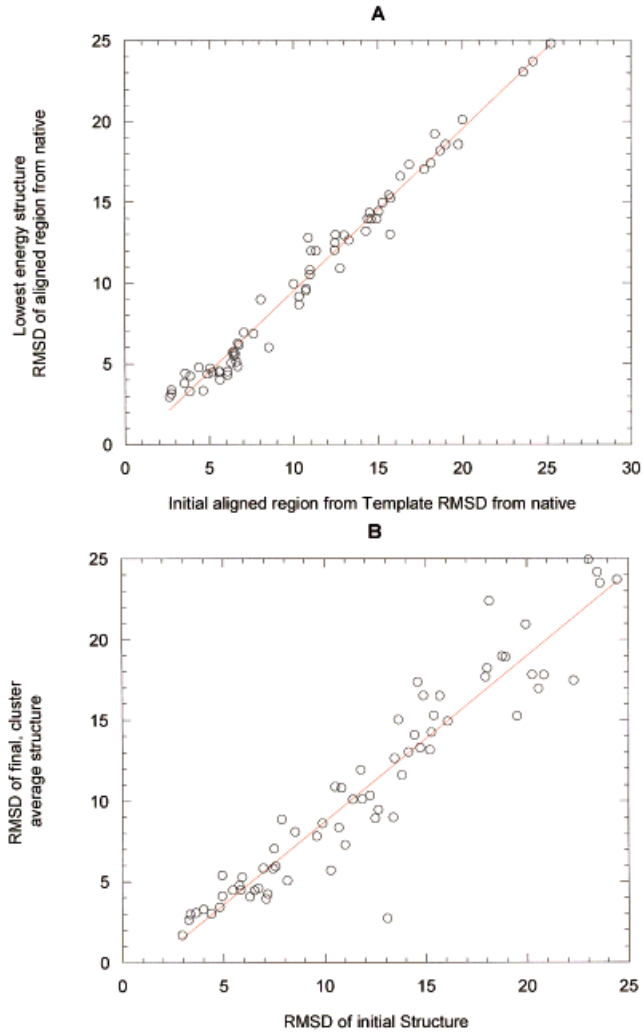


Fig. 3. **A:** Comparison of the root-mean-square deviation (RMSD) of the initial aligned region with the template with respect to the native structure versus the RMSD with respect to native of this same region in the lowest-energy structure. **B:** Comparison of the RMSD with respect to native of the entire initial structure with the RMSD with respect to native of the average structure extracted from clustering.

the structure did not improve much during the simulation, although the drop of energy in the early stage is still observed. In this particular case, the energy terms describing the generic stiffness, amino acid pair potential, hydrogen bond and amino acid burial (for a detailed explanation of these energy terms, see, for example, the previous article[23]) are reduced, but this was not enough to drive the structure in the correct direction.

We have also investigated the influence of other possible factors on the simulation results. The first question is whether some structural types of proteins are easier to refine (Table II). It may be worth mentioning that 22 out of 27 β proteins achieved some improvement regardless of the quality of the threading results. Looking at the trajectories of those proteins that exhibited some improvement in spite of a bad initial threading alignment, the RMSD of the final structures were better because of the regulariza-

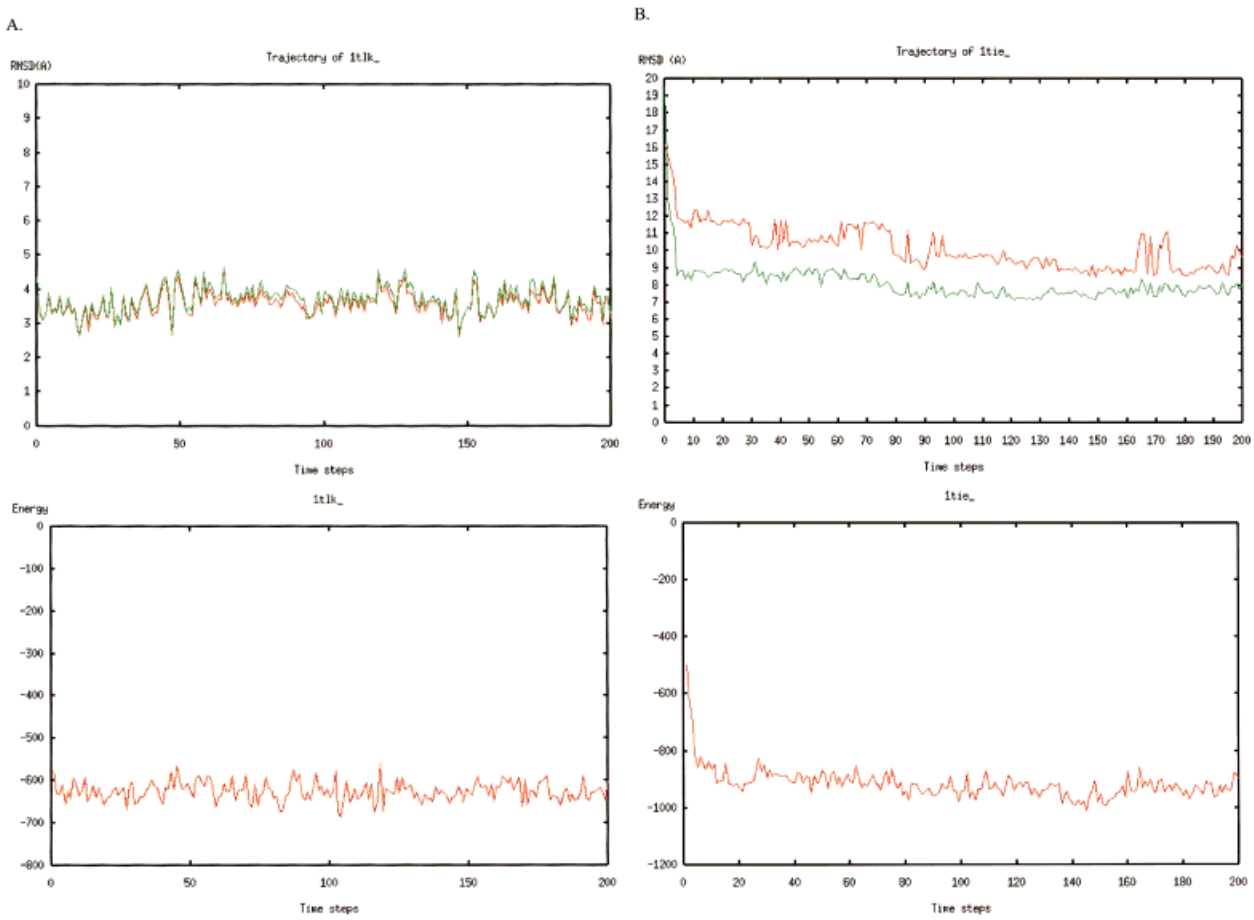


Fig. 4. Representative lattice simulation trajectories. **A:** The case of 1tlk_ whose initial threading result is 4.61 Å root-mean-square deviation (RMSD) from the native structure. **A-1:** The RMSD from native structure. The red line is whole protein RMSD and the green line is template region RMSD. **A-2:** The energy is shown for the same simulation shown in A-1. **B:** A case of 1tie_. RMSD of the threading result is 7.88 Å. **C:** A case of 1cid_. RMSD of the threading result is 19.8 Å. The trajectory that contains the lowest-energy structure is shown for these three targets.

tion of β-strands and sheets and because they are compact. The number of targets with good results (<10 Å RMSD) for each structural type can be readily explained by the fact that most have good initial threading results. There is a weak correlation between the length of the target chain and the final RMSD (the correlation coefficient between protein size and the whole protein RMSD is 0.53). But this is the same range as the correlation between the chain length and RMSD of the initial threading results (0.52). Finally, we observed that there was no correlation between the number of secondary or tertiary restraints used and the improvement in the RMSD during the simulation.

Figure 5A–E compares the initial, final (DG based structures), as well as the Modeller structures (see below) for 1aba_, 1rcb_, 1ten_, and 3chy_. In all five cases, the RMSD of the final structure is lower than that of the initial model, and Modeller is found to perform significantly poorer in this limited set; we return to this point below. Interestingly, sometimes even for rather good initial structures (e.g., 1onc_) the structures can improve somewhat. Other times, such as in 1rcb_, the improvement is minor, but there are cases, such as 1aba_, 1ten_, and 3chy_, where the improvement in the backbone RMSD is on the order of 2 Å.

Turning once again to Table V, none of the lowest-energy structures corresponds to the smallest RMSD structure. Thus, the potential energy needs to be improved. But in the meanwhile, a practical way to get the best possible structure out of the pool of structures generated by the series of simulations is needed. The following structure refinement procedures were undertaken to achieve this aim.

Selection of Structures by Distance Geometry

As shown in Table VI, application of distance geometry (DG) with the protocol outlined in the Methods section usually leads to a better structure selection than the structure of lowest conformational energy. In 53 of 68 cases, the structures have a lower RMSD from the native structure after the application of DG. Usually the improvement is small, in the range of 0.3 Å. However, in a number of cases, it is quite significant; in 10 cases the improvement is >1.0 Å and in 4 cases, it is >2.0 Å. Only in two cases were the structures after DG significantly

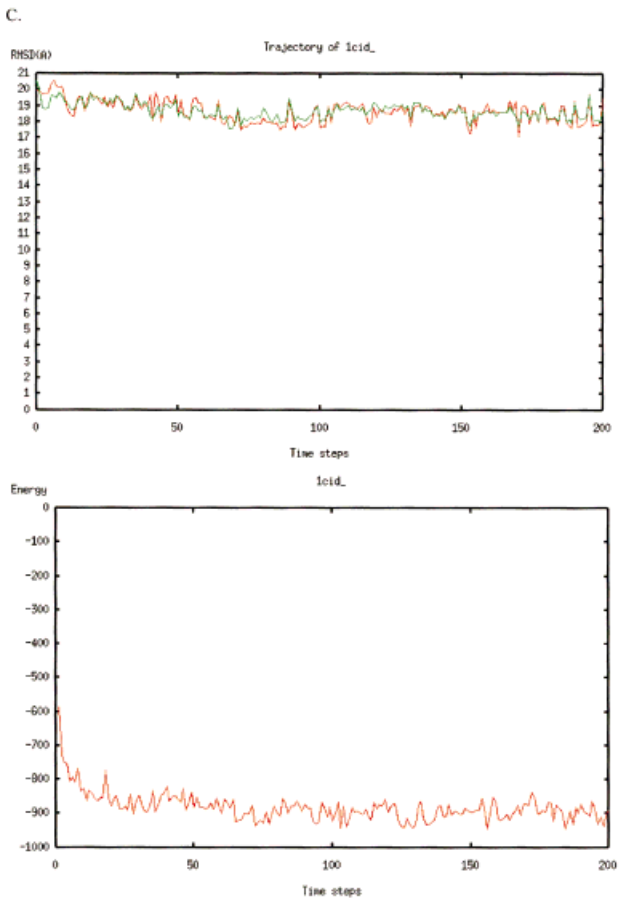


Figure 4. (Continued.)

worse than the lowest-energy structures (difference of >1.0 Å).

Results of Clustering

In general, comparison of the best centroid with the lowest-energy structure showed only marginal improvement. As shown in Table VI, for the 68 structures, the centroid RMSD improved on average by ~0.3 Å over the lowest-energy structures. Even though the centroids in most cases were similar in quality to the lowest-energy structures, only in a few cases were the centroids worse than the lowest-energy structures; however, in many cases the centroids were significantly better. In 52 of 68 cases, the structure generated by clustering has a lower RMSD from native than the lowest-energy structure. Two centroids (1stfI and 1aep) were worse by >1 Å as compared with the lowest-energy structures, while eight centroids (2fbjL, 1ltsD, 1fc1A, 1aba, 1cid, 1rcb, 3hlaB, 2azaA) were >1 Å better than the lowest-energy structures. Clustering is clearly the better procedure than distance geometry, as it generates (sometimes only slightly) better structures in 38 cases than distance geometry does, while distance geometry is better in 30 cases. Considering that clustering performs significantly better in *ab initio* folding, we will employ clustering in future work.

Comparison With Modeller

Several comparative modeling tools were recently developed. One of the most widely used is Modeller, developed by Sali and colleagues[2,3,34,35] Modeller allows for the high-throughput modeling of protein structures on a genomic scale.[3] Since the method presented here is more complex and more CPU intensive, (but high-throughput simulations are certainly possible), the key question is whether GENECOMP performs sufficiently better to justify the increased computational cost. To answer this question, we compared the structures generated by GENECOMP with Modeller in Table VII. Both procedures started from exactly the same templates and the same alignments generated by PROSPECTOR. If all models are considered, then GENECOMP performs better than Modeller in 53 cases, worse in 13 cases, and the same in two cases. Considering only templates whose RMSD is <10 Å, then GENECOMP performs better in 29 cases, Modeller performs better in 5 cases, and they perform the same in one case. However, when Modeller does perform better, the two structures differ by a small amount. In many cases of very good (or good) templates, the two methods generate models of similar quality. The situation changes when the level of homology becomes weaker and when, consequently, the threading models are more distant from the probe structure. Here, the models generated by GENECOMP are almost always of noticeably better accuracy. We can most likely ignore those cases when both methods lead to very bad models. As can easily be seen from the data compiled in Table VII within the range of 4–8 Å RMSD, GENECOMP almost always generated better models than Modeller. The typical difference is 1–2 Å; however, in a few cases it is as much as 4–5 Å. Thus, on average the proposed method leads to qualitatively better molecular models that will have significant consequences for structure-based protein function prediction and other aspects of proteomics.

DISCUSSION

In this article, we present a refined generalized comparative modeling method, GENECOMP, designed to improve the quality of moderate-resolution threading models. The basic idea of the approach is to perform *ab initio* folding using a lattice protein model, SICHO,[23] in the vicinity of an alignment to a template provided by the threading algorithm PROSPECTOR.[9] PROSPECTOR also provides predicted contacts and secondary structure not only for the template-aligned regions but also possibly for the unaligned regions by garnering additional information from other structures. This information is incorporated into the refinement algorithm and can therefore improve the unaligned regions as well. Since the lowest-energy structure generated by the simulations does not necessarily have the lowest RMSD from the native structure, we employed two structure selection protocols: Distance Geometry[31] and clustering.[30] Clustering is found to generate somewhat better quality structures in 38 of 68 cases. The resulting structures can also be converted to atomic detail models. In general, when applied to the Fischer database[25] we

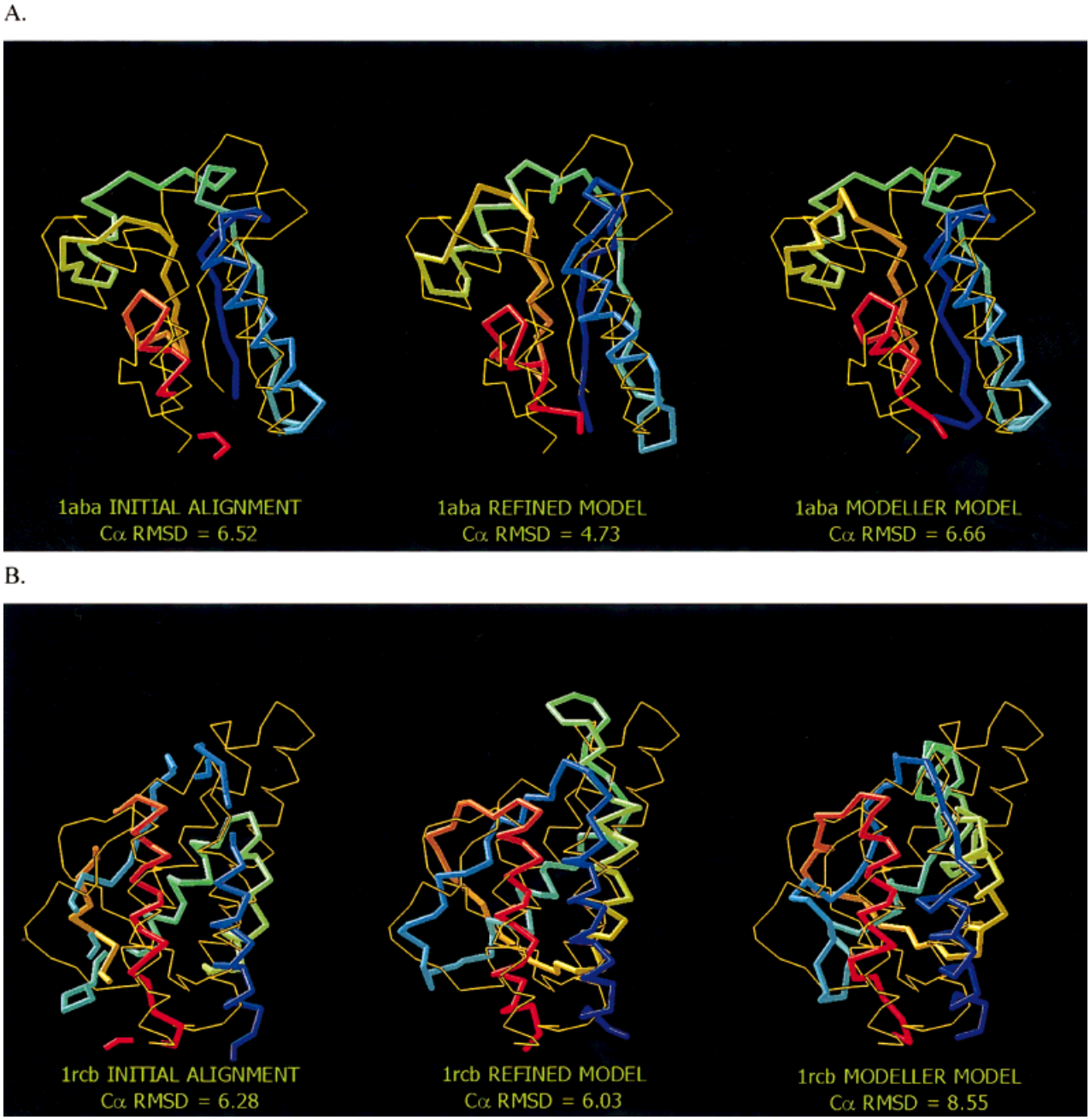
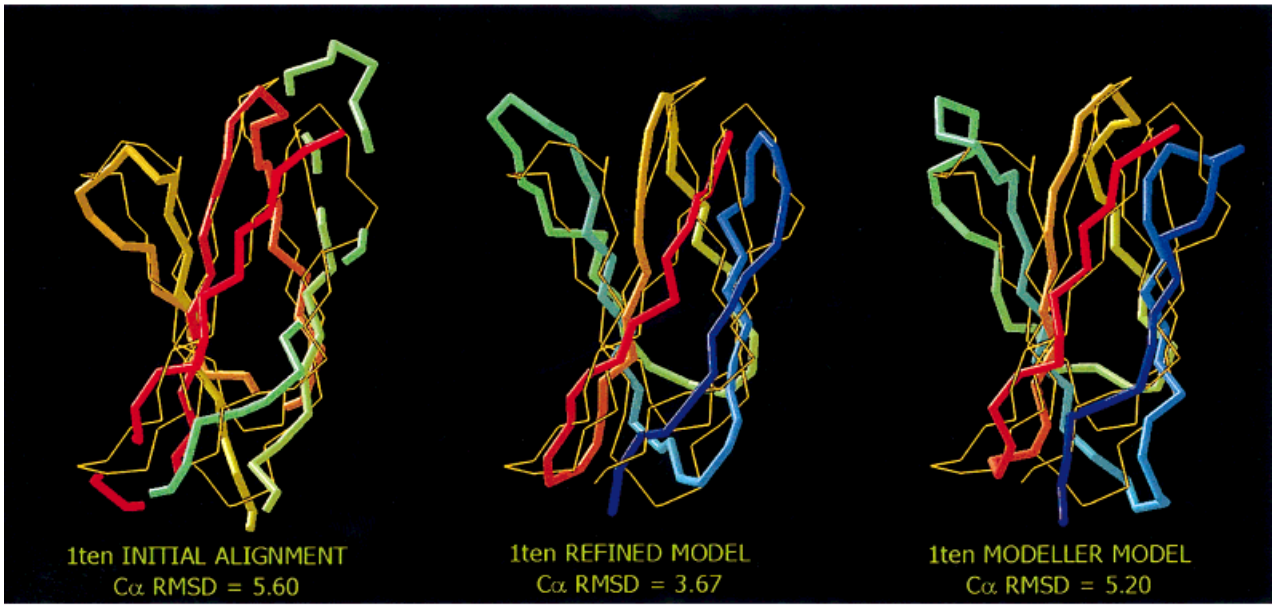


Fig. 5. Representative initial, final and Modeller structures for 1aba_(**A**), 1rcb_(**B**), 1ten_(**C**), and 3chy_(**D**). The native structure is shown in the thin tube, and the predicted structure is in the thick tube.

have found that the protocol does no harm and in a significant number of cases improves upon the initial threading model, sometimes dramatically. The procedure is readily automated and can be implemented on a genomic scale.

The question is why and when this method would work. Clearly the quality of the results depends of the initial alignment quality, with initial template alignments within the range of <10 Å, likely to see some improvement in the template alignment. Since the method also has the capacity of improving the quality of unaligned regions if these are not too long or if there are predicted contacts that are reasonably accurate then improvement in these regions can be expected. As in *ab initio* folding, for sufficiently large inserted regions in general the quality of the model is intimately tied to the quality of the predicted contacts. However, since the lattice model has a resolution of $\sim$3 Å RMSD from native (due to deficiencies in the force field; note that the geometric resolution of the SICHO model is $\sim$0.8 Å), templates below this RMSD cannot be effectively treated. Also we do not differ between good and poor templates; clearly this needs to be done. That is if one has a very good template, one would want to tightly bind the

C.



D.

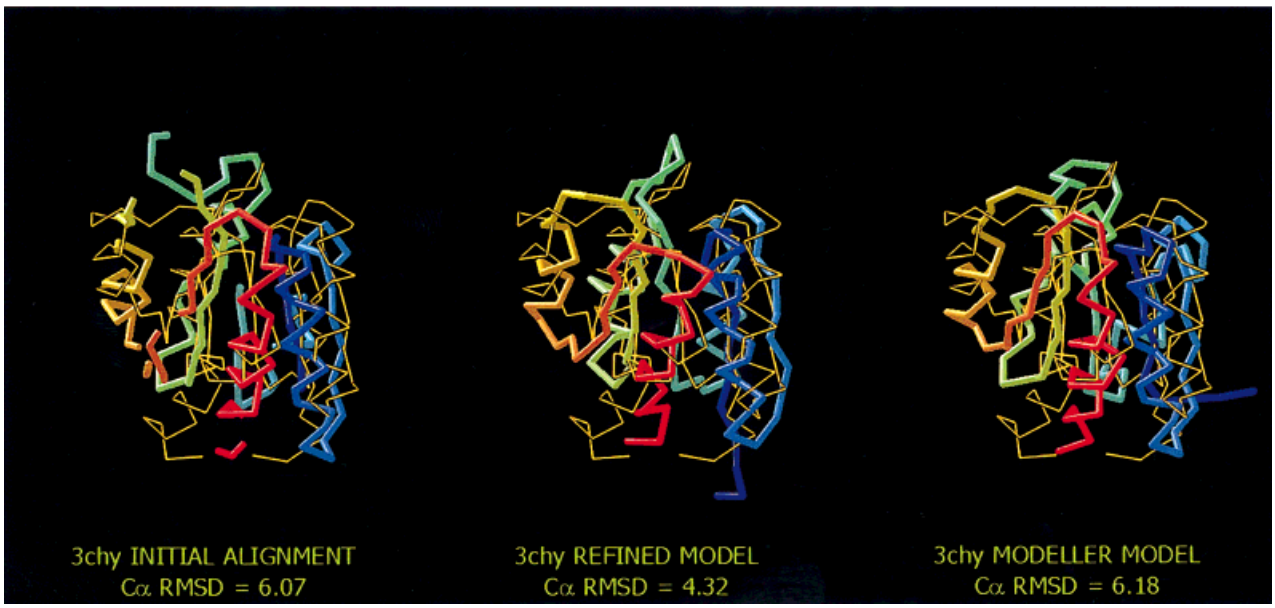


Figure 5. (Continued.)

probe sequence to it, whereas if the template is poor, then it should be loosely bound. Efforts are now underway to incorporate this feature into GENECOMP. Thus, there are a number of features of this protocol that demand improvement. Because the quality of the initial models depends on the initial alignment, threading algorithms that generate better alignments are required. Furthermore, a better procedure for the selection of lower RMSD structures needs to be developed. One promising way is to rebuild atomic models and then select the structure using heavy atom knowledge-based potentials.

In spite of its inadequacies, the existing procedure represents a significant step toward the development of techniques that refine moderate resolution threading models. Such a protocol is required to enhance the yield of true positives for structure-based functional annotation. Active site descriptors such as the FFF developed by Fetrow and Skolnick[1,16–19,36] require that the active site residues be more or less correctly positioned (the backbone RMSD in the vicinity of the active site should be ~4–5 Å). If not, a false-negative will result. The current protocols offer the promise of improving the quality of the alignment and of

TABLE VI. Comparison of Lowest Energy, Distance Geometry-Generated, Cluster-Based and Best Possible Structures*

Target	Lowest energy	DG	Clustering	Best structure
1aaj_	8.42	9.37	9.04	6.15
1aba_	5.58	4.75	3.95	3.55
1aep_	18.34	21.45	22.38	18.32
1arb_	17.30	17.46	17.80	15.78
1atnA	13.33	13.16	13.26	12.00
1bbhA	3.65	3.07	2.99	2.71
1bbt1	10.81	10.70	10.80	9.57
1bgeB	6.27	5.45	5.71	5.04
1c2rA	5.37	5.34	5.30	4.31
1cauB	5.69	5.45	5.41	4.04
1cewI	7.35	7.79	7.85	4.10
1chrA	5.11	4.90	4.78	3.77
1cid_	18.64	18.44	17.36	14.05
1cpcL	13.15	13.58	13.19	12.30
1crl_	24.21	24.09	24.93	21.35
1dsbA	15.94	16.47	15.90	11.58
1dxtB	3.53	3.01	3.08	2.91
1eaf_	10.09	10.32	10.10	9.27
1fc1A	12.89	13.12	13.01	12.43
1fxiA	10.28	10.18	10.22	8.53
1gal_	17.00	17.80	17.38	14.05
1gky_	7.76	6.36	8.94	6.13
1gp1A	14.75	13.74	15.06	9.08
1hip_	4.86	4.26	4.13	3.92
1hom_	5.00	1.57	1.70	1.50
1hrhA	5.50	5.07	5.07	4.25
1isuA	4.23	5.07	4.09	3.20
1lgaA	17.13	15.59	16.58	13.10
1ltsD	10.25	10.21	9.52	8.11
1mdc_	3.12	2.66	2.65	2.55
1mioC	15.19	14.71	14.94	14.05
1mup_	4.46	4.38	4.51	4.14
1npx_	13.75	14.12	14.10	13.61
1onc_	3.53	3.51	3.29	3.08
1osa_	17.57	17.90	17.85	16.56
1pfc_	4.48	4.28	4.46	3.81
1rcb_	5.52	6.09	4.30	3.91
1sacA	18.21	18.81	19.16	16.89
1stfI	7.38	7.07	8.11	4.97
1tahA	21.60	21.51	21.47	18.90
1ten_	3.96	3.62	3.49	3.16
1tie_	12.88	12.98	12.55	10.74
1tlk_	3.19	3.42	3.32	2.35
2afnA	23.23	25.05	23.55	22.60
2ak3A	15.51	15.46	15.29	14.65
2azaA	8.40	7.87	7.27	6.33
2cmd_	4.74	4.44	4.49	4.22
2fbjL	8.67	8.78	8.71	7.77
2gbp_	10.46	10.07	10.37	9.50
2hhmA	17.09	17.57	17.31	15.22
2hpdA	6.07	5.83	5.81	5.41
2mnr_	14.07	14.28	14.27	13.55
2mtaC	15.84	16.49	16.64	14.04
2omf_	23.51	23.45	24.29	21.82
2pia_	17.29	16.77	18.41	15.64
2pna_	11.31	8.92	10.90	7.27
2sarA	5.93	5.76	5.85	4.88
2sas_	5.78	6.11	5.95	5.51
2sga_	11.87	10.49	11.94	9.78
2sim_	19.79	18.57	17.47	16.52
2snv_	13.72	13.84	13.38	12.78
3cd4_	7.26	7.15	7.05	5.98
3chy_	4.53	4.36	4.59	3.58
3hlaB	8.96	8.63	8.66	4.72
3rubL	24.19	24.15	23.71	22.26
4sbvA	18.12	18.53	18.99	17.73
5fd1_	11.64	11.99	11.75	10.70
8i1b_	12.58	12.88	12.82	10.77

*All root-mean-square deviation (RMSD) values are in Ångstroms.

TABLE VII. Comparison of Generalized Comparative Modeling with Modeller*

Probe	GENECOMP + DG	Modeller
1aaj_	9.37	10.13
1aba_	4.75	6.66
1aep_	21.45	21.56
1arb_	17.46	18.56
1atnA	13.16	15.61
1bbhA	3.07	3.02
1bbt1	10.7	10.21
1bgeB	5.45	10.34
1c2rA	5.34	5.84
1cauB	5.45	5.93
1cewI	7.79	8.47
1chrA	4.9	4.57
1cid_	18.44	20.19
1cpcL	13.58	15.62
1crl_	24.09	25.89
1dsbA	16.47	16.37
1dxtB	3.01	3.05
1eaf_	10.32	10.82
1fc1A	13.12	15.02
1fxiA	10.18	11.27
1gal_	17.8	18.86
1gky_	6.36	11.82
1gp1A	13.74	15.22
1hip_	4.26	4.06
1hom_	1.57	1.73
1hrhA	5.07	6.95
1isuA	5.07	5.84
1lgaA	15.59	14.72
1ltsD	10.21	10.88
1mdc_	2.66	2.66
1mioC	14.71	16.78
1mup_	4.38	4.93
1npx_	14.12	14.48
1onc_	3.51	5.14
1osa_	17.9	16.89
1pfc_	4.28	4.39
1rcb_	6.09	8.55
1sacA	18.81	18.78
1stfI	7.07	12.76
1tahA	21.51	23.47
1ten_	3.62	5.2
1tie_	8.6	9.3
1tlk_	3.42	4.31
2afnA	25.05	25.67
2ak3A	15.46	19.89
2azaA	7.87	9.48
2cmd_	4.44	5.13
2fbjL	8.78	11.47
2gbp_	10.07	10.5
2hhmA	17.57	21.08
2hpdA	5.83	6.96
2mnr_	14.28	14.5
2mtaC	16.49	17.9
2omf_	23.45	25.34
2pia_	16.77	16.56
2pna_	8.92	10.64
2sarA	5.76	6.59
2sas_	6.11	6.97
2sga_	10.49	13.45
2sim_	18.57	14.43
2snv_	13.84	12.95
3cd4_	7.15	7.25
3chy_	4.36	6.18
3hlaB	8.63	9.6
3rubL	24.15	25.6
4sbvA	18.53	18.93
5fd1_	11.99	15.03
8i1b_	12.88	13.76

*All root-mean-square deviation (RMSD) values are in Ångstroms. The same alignments (see Table IV) were used as starting templates for GENECOMP and for Modeller.

increasing the yield of true positives in biochemical function prediction.

Currently we are applying this methodology to refine the set of threading-based structures in the *M. genitalium* genome.[26] Here all structures with a threading *Z*-score of >1 will be subjected to GENECOMP, and all predicted models will be provided on the web. This is but a first effort at genomic-scale threading model refinement; an effort that will receive increased emphasis in the next few years.

REFERENCES

1. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. Nature Biotechnol 2000;18: 283–287.
2. Sanchez R, Sali A. Evaluation of comparative protein structure modeling by MODELLER-3. Proteins 1997;Suppl 1:50–58.
3. Sanchez R, Pieper U, Mirkovic N, de Bakker PI, Wittenstein E, Sali A. MODBASE, a database of annotated comparative protein structure models. Nucleic Acids Res 2000;28:250–253.
4. Sternberg MJ, Bates PA, Kelley LA, MacCallum RM. Progress in protein structure prediction: assessment of CASP3. Curr Opin Struct Biol 1999;9:368–373.
5. Alwyn Jones T, Kleywegt GJ. CASP3 comparative modeling evaluation. Proteins 1999;Suppl 3:30–46.
6. Bryant SH. Evaluation of threading specificity and accuracy. Proteins 1996;26:172–185.
7. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
8. Wilmanns M, Eisenberg D. Inverse protein folding by the residue pair preference profile method. Protein Eng 1995;8:626–639.
9. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR: a new approach to threading. Proteins 2001;42:319–331.
10. Panchenko AR, Marchler-Bauer A, Bryant SH. Combination of threading potentials and sequence profiles improves fold recognition. J Mol Biol 2000;296:1319–1331.
11. Panchenko A, Marchler-Bauer A, Bryant SH. Threading with explicit models for evolutionary conservation of structure and sequence. Proteins 1999;Suppl 3:133–140.
12. Lee J, Liwo A, Ripoll DR, Pillardy J, Scheraga HA. Calculation of protein conformation by global optimization of a potential energy function. Proteins 1999;Suppl 3:204–208.
13. Ortiz A, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. Ab initio folding of proteins using restraints derived from evolutionary information. Proteins 1999;(suppl 3):177–185.
14. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;37(suppl 3):171–176.
15. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. Proteins 1999;37:149–170.
16. Fetrow JS, Skolnick J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J Mol Biol 1998;281:949–968.
17. Fetrow JS, Godzik A, Skolnick J. Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. J Mol Biol 1998;282:703–711.
18. Skolnick J, Fetrow J. From genes to protein structure and function: novel applications of computational approaches in the genomic era. TIBTECH 2000;18:34–39.
19. Zhang L, Godzik A, Skolnick J, Fetrow JS. Functional analysis of the *Escherichia coli* genome for members of the alpha/beta hydrolase family. Fold Des 1998;3:535–548.
20. Gerstein M. Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. Proteins 1998;33: 518–534.
21. Hegyi H, Gerstein M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J Mol Biol 1999;288:147–164.
22. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D- PSSM. J Mol Biol 2000;299:499–520.
23. Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. A method for improvement of threading based models. Proteins 1999;37:592–610.
24. Skolnick J, Kolinski A. A unified approach to the prediction of protein structure and function. Adv Chem Phys 2001;in press.
25. Fischer D, Elofsson A, Rice D, Eisenberg D. Assessing the performance of fold recognition methods by means of a comprehensive benchmark. Pacific Symp Biocomput 1996;300–318.
26. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, et al. The minimal gene complement of *Mycoplasma genitalium*. Science 1995;270:397–403.
27. Kolinski A, Skolnick J. Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. Proteins 1998;32:475–494.
28. Swendsen RH, Wang JS. Replica Monte Carlo simulation of spin glasses. Phys Rev Lett 1986;57:2607–2609.
29. 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.
30. Betancourt M, Skolnick J. Finding the needle in a haystack: Educing protein native folds from ambiguous ab initio folding predictions. J Comp Chem 2001;22:339–353.
31. Huang ES, Samudrala R, Ponder JW. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. J Mol Biol 1999;290:267–281.
32. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins 2000;38:3–16.
33. Feig M, Rotkiewicz P, Kolinski A, Skolnick J, Brooks CL III. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins 2000;41:86–97.
34. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
35. Sali A, Potterton L, Yuan F, van Vlijmen H, Karplus M. Evaluation of comparative protein modeling by MODELLER. Proteins 1995;23:318–326.
36. Fetrow JS, Siew N, Skolnick J. Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. FASEB J 1999; 13:1866–1874.