PROTEINS: Structure, Function, and Genetics 37:592–610 (1999)

A Method for the Improvement of Threading-Based Protein Models

Andrzej Kolinski,[1,2]* Piotr Rotkiewicz,[1,2] Bartosz Ilkowski,[1,2] and Jeffrey Skolnick[1]*
[1]*Laboratory of Computational Genomics and Bioinformatics, Danforth Plant Science Center, CET, St. Louis, Missouri*
[2]*Department of Chemistry, University of Warsaw, Warsaw, Poland*

ABSTRACT **A new method for the homology-based modeling of protein three-dimensional structures is proposed and evaluated. The alignment of a query sequence to a structural template produced by threading algorithms usually produces low-resolution molecular models. The proposed method attempts to improve these models. In the first stage, a high-coordination lattice approximation of the query protein fold is built by suitable tracking of the incomplete alignment of the structural template and connection of the alignment gaps. These initial lattice folds are very similar to the structures resulting from standard molecular modeling protocols. Then, a Monte Carlo simulated annealing procedure is used to refine the initial structure. The process is controlled by the model's internal force field and a set of loosely defined restraints that keep the lattice chain in the vicinity of the template conformation. The internal force field consists of several knowledge-based statistical potentials that are enhanced by a proper analysis of multiple sequence alignments. The template restraints are implemented such that the model chain can slide along the template structure or even ignore a substantial fraction of the initial alignment. The resulting lattice models are, in most cases, closer (sometimes much closer) to the target structure than the initial threading-based models. All atom models could easily be built from the lattice chains. The method is illustrated on 12 examples of target/template pairs whose initial threading alignments are of varying quality. Possible applications of the proposed method for use in protein function annotation are briefly discussed. Proteins 1999;37:592–610.** © 1999 Wiley-Liss, Inc.

Key words: protein modeling; homology modeling; Monte Carlo simulations; threading; lattice protein models; protein structure prediction

INTRODUCTION

In the limit of no apparent sequence similarity to any solved protein structure, threading methods are perhaps the most powerful tool for the approximate fold determination of new protein sequences.[1–6] Given a library of known protein structures, threading methods attempt to find the fold that is the most compatible with a probe sequence. Various sequence-to-structure alignment methods and a variety of scoring functions (simplified potentials) have been developed to achieve this goal.[1–4,7,8] When the probe sequence is aligned onto the structure of another protein, an approximate three-dimensional model results. Unfortunately, the resulting models obtained from threading approaches are usually of very low quality, with gaps and insertions in threading alignments that somehow have to be connected or closed. Some fragments are significantly misaligned with respect to the true structure of the test protein. This may happen for several reasons. First, despite some general similarity in their folds, the structure of the query sequence and the protein detected by a threading method may be quite different. Thus, even after the best possible superposition, the two structures may differ significantly. Second, various threading methods and their associated scoring functions only focus on aspects of protein structure and a subset of their possible interactions. Consequently, the resulting sequence-to-structure alignments are, in general, quite different from those which result from the optimal structural superposition.[6] Since standard threading procedures employ a rigid "template" protein scaffold, the design of a scoring function that would give rise to the best possible structural superposition is, in practice, extremely difficult.

Nevertheless, even crude and incomplete threading-based three-dimensional models are sometimes sufficient for the functional annotation of newly sequenced proteins.[9,10] However, as the quality of the model decreases, active site identification becomes more and more problematic. With regard to the immense number of new protein sequences of unknown functions that come from sequencing various genomes, any method that improves protein structure predictions would be of great scientific as well as practical importance.

When the threading alignments are of good quality, then standard homology modeling tools can be used to build useful molecular models.[11,12] In contrast, when the alignments are poor, it is rather unlikely that classical homology modeling can significantly correct alignment errors.

Grant sponsor: National Institutes of Health; Grant Number: GM-48835; Grant sponsor: KBN (Poland); Grant number: grant GR-919; Grant sponsor: Howard Hughes Medical Institute; Grant number: 75195-543402.
*Correspondence to: Jeffrey Skolnick, Laboratory of Computational Genomics and Bioinformatics, Danforth Plant Science Center, CET, 4041 Forest Park Avenue, St. Louis, MO 63108 or Andrzej Kolinski, Department of Chemistry, University of Warsaw, Pasteura 1, 02–093, Warsaw, Poland.
Received 8 February 1999; Accepted 20 August 1999

One possible way to deal with this problem was recently proposed by Jaroszewski et al.[13] They have shown that detailed molecular models built from various threading alignments (obtained by changing the parameters of the threading algorithm) can be evaluated using knowledge-based potentials and the best alignments could be detected in the majority of cases. This method detects and rejects these cases where the target and template structures differ significantly. However, it does not deal with the problem of improving the alignment subsequent to the identification of the best probe-template pair.

In this work, we attempt to build and refine protein molecular models. Threading-based target-template alignments have been obtained from one standard threading method;[14] but in principle any could be used.[15] The modeling technique employs a lattice model recently developed by us and tested in a different context. This SICHO (SIde Chain Only) model employs a very simple, computationally very efficient, yet quite accurate, representation of protein structure and dynamics.[16,17] For the purpose of the present application, the model has been refined by incorporating evolutionary information into the interaction scheme. Starting from an initial conformation of the model lattice chain that approximately follows the threading template, a Monte Carlo annealing procedure attempts to find a conformation that maintains some (but not all) features of the original template and at the same time, optimizes packing and intra-protein interactions, as defined by the reduced model of the probe protein. This could also be visualized as a folding simulation in a soft tube built around the threading template.

The new method has been applied to 12 target/template protein pairs that produce various quality models. The parameters of the lattice model force field (more precisely, the balance between the intrinsic force field and the template-related biases) have been adjusted by a trial and error method for three of the 12 target/template protein pairs. The obtained parameters were subsequently used in the other nine simulations. As will become apparent after analysis of the simulation results, the obtained models for the three proteins used for tuning the potential are among the best. This may suggest that the method was strongly tuned to these three examples. This is not the case. First, these three proteins belong to completely different structural classes, so the tuning should be rather general, i.e., applicable to the majority of single domain proteins. Second, when the tuning procedure is performed on just a single case (the plastocyanin/azurin pair) almost the same results are obtained; this suggests that the optimal balance between the template-related soft restraints and the intrinsic force field of the model is similar for various proteins. Finally, the poorer results obtained for most of the remaining nine test proteins are simply due to the very poor quality of the initial threading models.

The remainder of this paper is outlined as follows. In Methods, we present the reduced lattice protein model used in the Monte Carlo sampling procedure. We describe the protein representation, the model of stochastic dynamics, the interaction scheme and the template-related biases and restraints. Then, in the Results section, the molecular models obtained from Monte Carlo simulated annealing and subsequent refinement procedures are compared with the initial crude, threading-based models. In the Discussion, we analyze the improved models and attempt to identify typical underlying structural rearrangements. Possible developments and applications to large-scale sequence-to-structure, and structure-to-function computational projects are also briefly discussed. Finally, the Conclusions section summarizes the main findings of this work. Some details that are important on a technical level, but not necessary for following the main features of this work, are found in the Appendix.

METHODS

Lattice Model

The reduced modeling of protein structure and dynamics usually employs an alpha carbon main chain representation.[18,19] Side chains are either completely neglected or treated at various levels of simplification. The choice of the alpha carbon representation is mostly motivated by the high level of geometric regularity of the main chains in folded proteins.[19] On the other hand, the packing and interactions between the side chains are perhaps much more sequence specific than are those of the main chain.

Motivated by the above reasons, we recently proposed a very simple lattice model in which we only explicitly treat the protein side chains. Elsewhere, it has been shown that it is possible to incorporate many protein-like features into such a representation.[16,17] These include local conformational propensities and the characteristic packing regularities of protein side chains. The advantage of this model is that the entire conformational space of quite large proteins can be efficiently sampled. For example, assuming loose knowledge of the secondary structure and a few long-range side chain contacts (about N/7, where N is the number of residues), which may come from sparse NMR data or other experimental techniques, low-resolution protein structures could be reproducibly and rapidly assembled for proteins containing up to 250 amino acids.[17] This compares rather favorably with other attempts to build low-resolution models from a small number of long-distance restraints.[19]

The model employed here is very similar to that previously described. There are small updates to the protein representation that slightly increase the geometric fidelity of the model. For the reader's convenience, the design of the model is outlined below.

Reduced representation of polypeptide chains

The model chain consists of a string of virtual bonds connecting the interaction centers that correspond to the center of mass of the side chains and the backbone alpha carbons. All heavy atoms have the same weight in this averaging. Thus, the center of glycine coincides with its C_α, the center of alanine is located in the middle of the $C_\alpha - C_\beta$ bond, the center of valine roughly coincides with the C_β atom, etc. These interaction centers (beads) are projected onto an underlying cubic lattice with a lattice spacing of 1.45 Å. Obviously, the virtual bonds resulting from such a

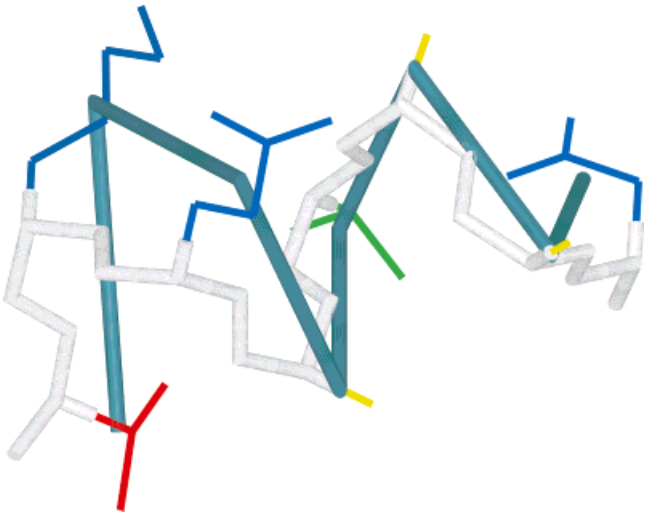

Fig. 1. Schematic illustration of the protein representation employed in this work. The fragment of a detailed protein structure (main chain backbone is shown in gray and the side chains in thinner sticks). The blue sticks correspond to the virtual bonds of the model chains, connecting the centers of mass of groups of atoms consisting of side chains and alpha carbons.

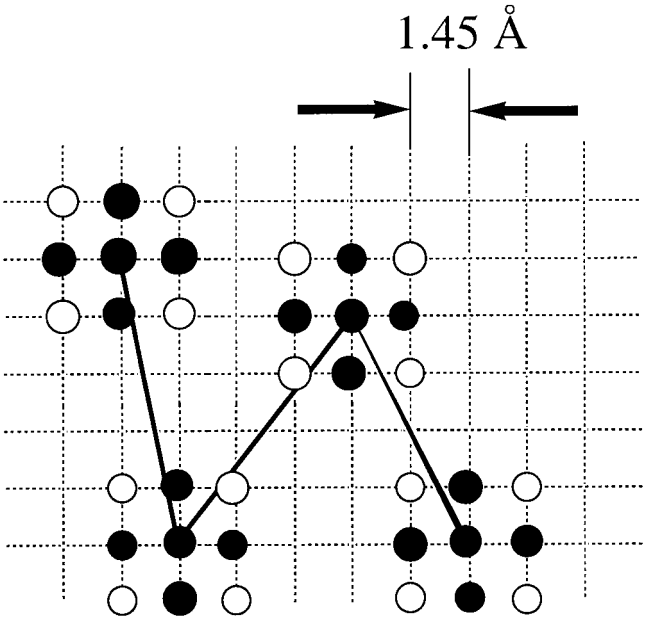

Fig. 2. Lattice representation of the model chain and its excluded volume. The sticks correspond to the model chain virtual bonds. Excluded volume of each model amino acid is represented by 19 points on the underlying cubic lattice with the mesh size equal to 1.45 Å. The black dots correspond to three lattice points along the axis orthogonal to the picture plane (one in the plane, one below, and one above the plane). The open circles correspond to single lattice points in the picture plane.

projection are of various lengths that depend on the identity of the two corresponding residues, the main chain conformation and the rotameric state of the side chain (see Fig. 1). A change in any of these variables may change the corresponding virtual bonds (the chain vectors **v**). In proteins, these distances have a quite broad distribution, ranging from 3.8 Å for a pair of glycines to about 10 Å for some pairs of large side chains in their anti-parallel orientation and expanded conformations. The corresponding set of lattice vectors covers this distribution with good fidelity. The shortest vectors are of the form of (±2, ±2, ±1) or (±3, 0, 0) vectors, including all possible permutations. The length of these vectors corresponds to a distance of 4.35 Å. The longest lattice vectors are of the (±5, ±2, ±1) type and their length corresponds to 7.94 Å. Thus, the wings of the distribution are cut off. This should not have any noticeable effect on the model's fidelity because the small-distance cut-off error is well below the resolution of the model, and the long-distance cut-off error is not important due to very rare occurrences of distances above 8 Å. As a result, the set of allowed lattice bonds consists of 646 vectors. For technical reasons, sequentially adjacent vectors must not be identical.

A cluster of excluded volume points is associated with each bead of the model chain. Each cluster consists of 19 lattice points: the central one, six points at positions (±1, 0, 0), (0, ±1, 0) and (0, 0, ±1) with respect to the central one, and 12 points at positions (±1, ±1, 0), including all permutations. Thus, the closest approach positions of another cluster with respect to a given cluster are of the form (±2, ±2, ±1) and (±3, 0, 0), as measured between the cluster centers. Consequently, there are 30 closest approach positions. The distance of the closest approach nicely corresponds to the smallest values of the inter-residue distances in real proteins. Since the average "contact distances" (see the following sections) of the model residues are somewhat larger than the distance of the closest approach, there are much more than 30 spatial orientations of two residues being in contact. Consequently, such a representation of protein structure entirely avoids various anisotropy effects typically seen in the lower resolution lattice protein models. Figure 2 shows a small fragment of the model chain confined to the underlying cubic lattice with a lattice spacing equal to 1.45 Å. The excluded volume points are denoted by the solid and open circles. The solid circles indicate the three lattice points along the direction orthogonal to the plane of the figure: one in the plane below and one in front of the plane. The open circles denote points in the plane. With the above geometric restrictions, all PDB structures[20] could be represented with an average root mean square deviation (rmsd) of about 0.8 Å. Again, the accuracy of the fit does not show any systematic dependence on protein length or on the orientation of the crystallographic structure with respect to the lattice coordinate system. Some features of the model chain are illustrated in Figure 1.

Conformational updating

The simplicity of the model protein representation facilitates the very rapid sampling of conformational space. The Monte Carlo algorithm employs three types of conformational transitions. The first type is a single bead, two-chain vector move. A random displacement of a randomly selected bead is generated and approved provided that the vector lengths and the excluded volume are not violated. The range of a random displacement is from 1 to $5^{1/2}$ lattice units. When accepted by the Metropolis criterion[21] (see the next section), such a move is equivalent to a collective

rearrangement of the main chain and/or the side chain internal coordinates in a real polypeptide chain. The force field of the model, especially its generic components, prevents the acceptance of nonsensical, non protein-like conformations.[16] The second type of motion involves the permutation of three chain vectors. This is a larger scale move that is relatively rarely accepted due to possible steric interactions. The last type of move involves a randomly selected fragment consisting of several chain units. This fragment moves as a rigid body due to appropriate small changes in the two flanking chain vectors. For instance, such a move may translate a helical segment by a small distance, thereby slightly changing the conformation of the corresponding turn or loop regions.

Interaction scheme

The model force field consists of several types of potentials. The first are generic, sequence-independent, biases that penalize against non protein-like conformations. Sequence specific contributions to the force field consist of knowledge-based, two-body and multibody potentials extracted from a statistical analysis of known protein structures. Finally, there are two kinds of potentials that contain evolutionary information extracted from multiple sequence alignments. In all cases, all PDB structures whose sequences are similar to the query sequence have been removed from the structural database used in the derivation of the potential (greater than 25% sequence identity). All potentials were derived from PDB structures and then translated into proper lattice discretized form.

The generic protein stiffness potential and secondary structure bias. As defined above, the model chain is intrinsically very flexible. A substantial fraction of its conformations that are allowed due to the assumed simplified hard core interactions do not correspond to any real polypeptide chain conformation. In reality, proteins are relatively stiff polymers exhibiting very characteristic distributions of certain short-range distances. For example, the bimodal distribution of the distances between the i-th and i+4th residues reflects the tendency to adopt either of two types of conformations. These correspond to expanded (β-type or expanded coil) or very compact conformations (as within helices or turns). Such generic features need to be included in the model. We proceed in a similar fashion, as described elsewhere.[16] The details are different due to the particular protein representation that we employ.

First, for all possible two-vector sequences of the model chain, let us define a direction **w** that is almost perpendicular to the plane formed by the fragment. A small systematic deviation from the exactly orthogonal direction is introduced into **w** to obtain vectors that are on average parallel to the helix axis and that also account for the average supertwist of β-strands.

$$\mathbf{u}_i = (\mathbf{v}_{i-1} \otimes \mathbf{v}_i - \mathbf{v}_{i-1} - \mathbf{v}_i) \quad (1)$$

$$\mathbf{w}_i = \mathbf{u}_i / |\mathbf{u}_i| \quad (2)$$

where $\mathbf{v}_i$ is the i-th vector (or virtual bond) of the model chain, the symbol "⊗" denotes the vector cross product and

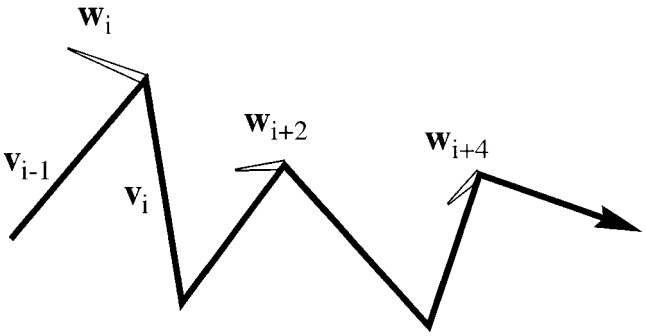

Fig. 3. A fragment of the model chain and a set of vectors **w** employed in the definition of the short-range polypeptide chain stiffness (see the text for details).

$|\mathbf{u}_i|$ is the length of vector $\mathbf{u}_i$. These "directions of secondary structure" (the vectors **w** point along a helix or across a β-sheet) are normalized so that their length equals unity. The idea is explained in Figure 3, where the model chain virtual bonds are shown in solid lines and the vectors **w**i are shown in open arrows.

The stiffness/secondary structure bias term has the following form:

$$E_{stiff} = -\epsilon_{gen} [\Sigma \min \{0.5, \max (0, \mathbf{w}_i \bullet \mathbf{w}_{i+2})\}] - \epsilon_{gen} [\Sigma \min \{0.5, \max (0, \mathbf{w}_i \bullet \mathbf{w}_{i+4})\}]. \quad (3)$$

Where ϵ_{gen} is a constant energy parameter, common for all generic potentials, and Σ means the summation along the chain. The above formulation means that the system is energetically stabilized when pairs of the "direction of secondary structure" vectors are parallel (positive dot product). As can be read from the above equation, the stabilization energy increases in the range between 90° and 30° (the angle between appropriate vectors **w**) and then maintains its extreme value. Thus, small fluctuations of secondary structure have no influence on the value of this potential.

Additionally, a weak bias has been introduced towards helix-like and β-type expanded states. All conformations are, of course, allowed; the purpose of this bias is to mimic a protein-like (average) distribution of local conformations. Symbolically, this could be written as follows:

$$E_{struct} = \Sigma \{\delta H1(i) + \delta H2(i) + \delta E1(i) + \delta E2(i)\} \quad (4)$$

with:

$$\delta H1(i) = -\epsilon_{gen}, \quad \text{for } r^2_{i,i+4} < 36 \text{ and } (\mathbf{v}_i \bullet \mathbf{v}_{i+3}) > 0 \text{ and } (\mathbf{v}_i \bullet \mathbf{v}_{i+2}) < -5$$
$$0, \quad \text{otherwise} \quad (4a)$$

$$\delta H2(i) = -\epsilon_{gen}, \quad \text{for } r^2_{i,i+4} < 36 \text{ and } (\mathbf{v}_i \bullet \mathbf{v}_{i+3}) > 0 \text{ and } (\mathbf{v}_{i+1} \bullet \mathbf{v}_{i+3}) < -5$$
$$0, \quad \text{otherwise} \quad (4b)$$

$$\delta E1(i) = -\epsilon_{gen}, \quad \text{for } 56 < r^2_{i,i+4} < 135 \text{ and } (\mathbf{v}_i \bullet \mathbf{v}_{i+2}) > 5$$
$$0, \quad \text{otherwise} \tag{4c}$$

$$\delta E2(i) = -\epsilon_{gen}, \quad \text{for } 56 < r^2_{i,i+4} < 135 \text{ and } (\mathbf{v}_{i+1} \bullet \mathbf{v}_{i+3}) > 5$$
$$0, \quad \text{otherwise} \tag{4d}$$

The numerical values are in lattice units and are selected to define a broad range of helical/turn conformations (for the δH1 and δH2 contributions) or expanded conformations (for the δE1 and δE2 contributions). Due to the exclusive character of the two subsets of geometrical conditions for specific chain conformations, the minimum contribution from a residue is equal to $-2\epsilon_{gen}$ (either the first two conditions or the two last conditions can be simultaneously satisfied). Let us express the last condition a bit differently. Equation (4d) says that the system gains an energy equal to $-\epsilon_{gen}$ for being in an expanded β-type conformation. For a four-vector fragment of the chain, this requires that the distance between the i-th and $i+4^{th}$ beads (the centers of mass of the side chain plus Cα units) has to lie between 10.7 and 16.8 Å, and the chain vectors $\mathbf{v}_{i+1}$ and $\mathbf{v}_{i+3}$ have to be oriented in a parallel-like fashion (the dot product > 5). Additional stabilization is gained when, for the same fragment, another pair of vectors is parallel (see Eq. (4c)). The broad ranges allow for substantial fluctuations around an ideal expanded state and accommodate the variations of the model chain geometry caused by differences in side chain size.

We have performed computational experiments where all interactions except the ones defined above, were turned off. At low temperature, the model chain forms rapidly fluctuating local clusters of expanded and helix-like states. The persistence length and the distributions of the short-range distances along the chains mimic protein-like geometry.

Generic packing cooperativity. We introduce two terms that enforce some of the most general regularities of the dense packing of protein structures.[22] In all the more regular elements of secondary structure (within helices and β-sheets, but not between helices) and, to a lesser extent, in some coil-type fragments and turns, given a contact between a pair of reference residues, there is a very strong preference to have contacts between the preceding and the following residues. Indeed, the contact maps of globular proteins contain very characteristic strips.[23] Those near the diagonal correspond to the intra-helical contacts, those farther from the diagonal (parallel or antiparallel to the diagonal) correspond to contacts between β-strands within β-sheets. Thus, we introduce the following energetic bias towards such a mode of packing:

$$E_{map} = -\epsilon_{gen} [\Sigma\Sigma (\delta_{i,j} \bullet \delta_{i+1,j+1} \bullet \delta_{i-1,j-1}) \delta_{par} + \Sigma\Sigma (\delta_{i,j} \bullet \delta_{i-1,j+1} \bullet \delta_{i+1,j-1}) \delta_{apar}] \tag{5}$$

where the summations are over all pairs of residues i, j, and $\delta_{i,j}$ is equal to 1 (0) when residues i and j are (are not) in contact. δ_{par} is equal to 1 only when the corresponding chain fragments are oriented in a parallel manner, i.e., when the chain vectors satisfy the following condition $(\mathbf{v}_{i-1} + \mathbf{v}_i)\bullet(\mathbf{v}_{j-1} + \mathbf{v}_j) > 0$, otherwise $\delta_{par} = 0$. Similarly, δ_{apar} is equal to 1 when the chain fragments are antiparallel, and it is equal to zero otherwise. For a given contact of a pair of residues, the maximal energetic stabilization due to regular side chain packing is therefore equal to $-\epsilon_{gen}$, which has the same value as in the previously defined potentials.

The packing cooperativity of the model protein is further enhanced by a term that mimics main-chain hydrogen bonds. The geometry of protein hydrogen bonds is translated into a specific range of the model chain geometry. First, let us define a vector that is likely to connect the model beads that are within motifs that represent regular secondary structure elements. Such a vector should connect beads i and i+3 in a helix and the appropriate beads in a β-sheet. An optimization procedure leads to the following definition of this vector:

$$\mathbf{h}_i = 3.3(\mathbf{v}_{i-1} \otimes \mathbf{v}_i)/|(\mathbf{v}_{i-1} \otimes \mathbf{v}_i)| - \mathbf{v}_{i-1}/|\mathbf{v}_{i-1}|. \tag{6}$$

The value of the 3.3 pre-factor has been found to be optimal for reproducing the internal main chain hydrogen bonding in the lattice projected PDB structures. However, due to the wide distribution of the model chain bond lengths, there are always some hydrogen bonds that are missed in the model. The coordinates of the vectors $\mathbf{h}_i$ are rounded-off to the nearest integer value. Thus, in a helix the $\mathbf{h}_i$ vectors have a component whose length is about 3 lattice units in the direction perpendicular to the three-residue plane (the first term in the above sum). They are also tilted back by a lattice unit (the last term of Eq. (6)). The projection along the helix axis is also about 3 lattice units; this nicely coincides with the 1.5 Å longitudinal increment per residue in a real helix. Residue i is considered to be hydrogen bonded with residue j when the vector $\mathbf{h}_i$ points to any of the 19 points of the excluded volume cluster of residue j. Correspondingly, the vector $-\mathbf{h}_i$ may point to another cluster. Such a situation is illustrated in Figure 4, where residue i is hydrogen bonded with residues j and k because the hydrogen bond vectors coincide with the excluded volume of both residues. The excluded volume clusters are symbolically represented by open spheres. Since the excluded volume clusters never overlap, the maximum number of these "hydrogen bonds" originating from residue i is equal to 2. The total energy of the "hydrogen bond network" can be written as:

$$E_{H\text{-}bond} = -\epsilon_{H\text{-}bond} \Sigma (\delta^+ + \delta^- + \delta^{+,-}) \tag{7}$$

where δ^+ (δ^-) equals 1 when the vector $\mathbf{h}_i$ ($-\mathbf{h}_i$) connects with an excluded volume cluster, and $\delta^{+,-} = 1$ when both vectors connect to some clusters, respectively. Otherwise, the corresponding terms are equal to zero. The cooperative contribution, $\delta^{+,-}$, corresponds to the local saturation of the hydrogen bond network.

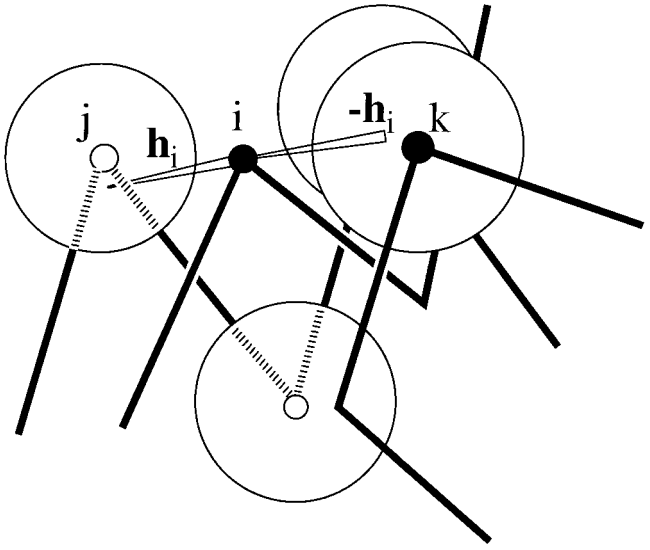

Fig. 4. Schematic illustration of the main chain's "hydrogen bonds." Residue i is hydrogen bonded to residue j and k because the vectors $\mathbf{h}_i$ and $-\mathbf{h}_i$ (see the definition in the text) connect with any of the points forming of the excluded volume clusters (the clusters are symbolically shown as large spheres) of these residues.

Again, a computational experiment has been done to check the effect of these generic potentials on the behavior of the model system. When only the interactions outlined up to this point are included (all the above short- and long-range generic potentials), the model lacks sequence specific information. At sufficiently low temperatures, the chain adopts either of the following two types of structures: a long (sometimes broken) helical structure, or a β-sheet with a right-handed supertwist. These motifs fluctuate and are not structurally unique. In a long chain, these two classes of secondary structure elements sometimes form separate domains.

Sequence specific short-range interactions. For the sequence of interest from the structural database, one may extract the statistics of distances between a pair of amino acids (with their interaction centers as defined in the model) A_i and B_{i+k}, where A and B denote the identities of the amino acids and i is the position in the chain. Here, we consider k = 1, 2, 3, 4, 6 and 8. The terms for k = 3 and k = 6 are treated as chiral variables. This means that the distance between A_i and B_{i+3} is stored as a positive or negative number, depending on the handedness of the corresponding three-bond segment. For the k = 6 case, the chirality is defined for three subsequent supervectors (the doublets of vectors between beads i and i+2, i+2 and i+4, and from i+4 to i+6). As was done here, the sequence of interest can be divided into overlapping short fragments. These could be aligned to the sequences of known structures. The highest scoring fragments provide a set of structural templates. The obtained statistics could be related to a random distribution and the statistical potential of mean force could be appropriately derived. The k = 1, 2, 3, and 4 terms were weighted equally, while the terms for k = 6 and k = 8 had weights reduced by a factor of two, with respect to the lower order terms. Homologous proteins were always excised from the structural database for the purpose of these test calculations. As previously shown, this type of potential very nicely reproduces the local conformational propensities of globular proteins.[16]

TABLE I. Compilation of Pairwise Cut-Off Distances for Pairwise Interactions

A_i	A_j	$R_{i,j}^{rep}$ (Å)	$R_{i,j}$ (Å)
Small[a]	Small	4.35[b]	5.97
Large[c]	Large	4.83	6.80
Other	Combinations[d]	4.57	6.32

[a]Small amino acids are: Gly, Ala, Ser, Cys.
[b]This value corresponds to the excluded volume radius of three lattice units; therefore, for pairs of small amino acids, the soft-core envelope does not exist.
[c]Large amino acids are Phe, Tyr, Trp.
[d]Small-large, other (than small or large)-large, other-small.

The short-range potentials could be made even more sequence specific when one employs evolutionary information encoded in homologous sequences. In such a case, the aligned fragments of highly homologous sequences (from the sequence database) are treated as the original test sequence, thereby increasing the strength of the statistics. The details of the derivation procedure are given in Appendix 1. Encoding such evolutionary information improves performance of the proposed method; however, it is not crucial. Simulations without homology-enhanced potentials lead to slightly worse results. Most of the test sequences employed in this work belong to relatively large families of proteins; however, the criterion of the number of similar sequences was not taken into consideration in the selection process. Also, in this respect, the selected set is rather representative of all small single domain proteins.

Sequence specific pairwise interactions. The pairwise interactions between model residues are defined by contact potentials in the form of a square well function.

$$E_{ij} = \begin{cases} \infty, & \text{for} \quad r_{ij} < 3 \\ E^{rep}, & \text{for} \quad 3 \le r_{ij} < R_{i,j}^{rep} \\ \epsilon_{ij}, & \text{for} \quad R_{i,j}^{rep} \le r_{ij} < R_{i,j} \\ 0, & \text{for} \quad R_{i,j} < r_{i,j} \end{cases} \tag{8}$$

where ϵ_{ij} are the pairwise interaction parameters, r_{ij} is the distance between chain beads i and j, $E^{rep} = 3kT$ is a constant repulsive term operating at very short distances, and $R_{i,j}^{rep}$ and $R_{i,j}$ are the cut-off values that depend on amino acid type. The values of these cut-off parameters are provided in Table I.

Because the derivation of the potentials uses evolutionary information, the interaction parameters depend not only on amino acid identity, but also on their positions in the polypeptide chain. A more detailed description of the

derivation of these potentials may be found elsewhere.[17] The total energy contribution from the pairwise interactions is therefore calculated as follows:

$$E_{pair} = \Sigma \Sigma E_{ij} \quad (9)$$

where the summations are over all $j > i$ pairs of residues.

Multibody potentials. The hydrophobic interactions in our model are partially accounted for by pairwise interactions between residues; however, this is not sufficient to generate well-packed proteins. Thus, a surface exposure based statistical potential has been developed. The scheme is as follows: Each model residue has been assigned 24 surface contact points. A specific subset of these contact points becomes occupied upon contact with other residues. The main-chain Cα atoms contribute separately to the coverage of a given residue. The positions of the Cα atoms could be quite well approximated given the positions of three consecutive side chain beads.[16] Some contact points could be multiply occupied. The fraction of non-occupied surface points defines the exposed fraction of a given side chain. Potentials could be derived from a statistical analysis of the protein structures for which the solvent exposure has been determined on the atomic level. The total surface energy is computed as follows:

$$E_{surface} = \Sigma E_b(A_i, a_i) \quad (10)$$

where a_i is the covered fraction of the residue A_i and $E_b(A_i, a_i)$ is the statistical potential when amino acid type A has a_i of its surface points occupied, i.e., the covered fraction of its surface is equal to $a_i/24$.

Studying the distribution of inter-residue contacts in globular proteins, we have found that various amino acids have different tendencies to pack in a parallel or antiparallel fashion. A contact between residues i and j is considered to be "parallel" when $(\mathbf{v}_{i-1} - \mathbf{v}_i)\bullet(\mathbf{v}_{j-1} - \mathbf{v}_j) > 0$, and "antiparallel" otherwise. Moreover, for a given residue there are strong correlations between the number of parallel and antiparallel contacts given the total number of contacts. Due to the reduced character of our model, the other contributions to the force field do not properly account for such effects. Therefore, the model force field has been supplemented by the following multibody potential:

$$E_{multi} = \Sigma E_m(A,n_p,n_a) \quad (11)$$

where $E_m(A,n_p,n_a)$ is the value of the statistical potential for residue type A having n_p parallel and n_a antiparallel contacts. The reference state is a random distribution of contacts. The values along particular diagonals ($n_p + n_a = n_c$) have been re-normalized such that the lowest energy for a diagonal was exactly equal to the value of statistical potentials derived from the distribution of the total number of contacts n_c for a given type of residue.

Total intrinsic conformational energy. The total internal conformational energy of the model chain was equal to:

$$E_{total} = E_{stiff} + E_{map} + 0.875E_{H\text{-}bond} + 0.75E_{short} + 1.25E_{pair} + 0.5E_{surface} + 0.5E_{multi} \quad (12)$$

with the value of generic parameter $\epsilon_{gen} = 1$ kT.

The relative scaling of various potentials has been adjusted by a trial and error method in ab initio folding experiments performed for the following selected small proteins: 1fna, the B domain of protein A and the B1 domain of protein G. The objective was to maintain low secondary structure content in the random coiled state and dense packing with a proper level of secondary structure in the collapsed globular state. For instance, the small 56-residue α/β protein G domain folded ab initio in about 30% of simulated annealing Monte Carlo simulations to a native-like structure with an rmsd from native in the range of 4 Å. The majority of the remaining misfolded conformations had native-like secondary structures, but they had topological errors, usually involving the wrong order of β-strands in the four-member β-sheet. The model is not sensitive to small variations in these scaling parameters.

Building the Starting Lattice Model

A separate algorithm was used to build an initial lattice model from a given target sequence alignment to a template structure. Such alignments contain gaps and insertions. First, interaction centers are computed from the template. Then, starting from the first aligned position, the lattice chain is sequentially built. At each step in the aligned region, the new vectors are selected so as to minimize the distance of the lattice chain from the equivalent template points. In the gap regions, the distance from the last residue of the preceding aligned fragment to the first residue of the next is divided to generate a set of checkpoints. The number of these checkpoints is equal to the number of target sequence residues that have to be mounted to span the gap. The checkpoints outside the entire alignment are randomly generated. The set of all checkpoints provides the target for the starting lattice model. The model chain maintains the excluded volume and satisfies the other geometric restrictions discussed before.

Implementation of the Template Restraints

The template (more precisely the structural fragments of the template protein that correspond to the aligned residues of the probe sequence) is projected onto the underlying cubic lattice. The corresponding three-dimensional array, initially filled with zeros, is then updated to store a loose trace of the template. All elements of the array that are closer than $6^{1/2}$ lattice units from template residues are assigned the corresponding residue numbers. When a lattice point is within a distance of $6^{1/2}$ from two or more residues, the number of the closest residue is as-

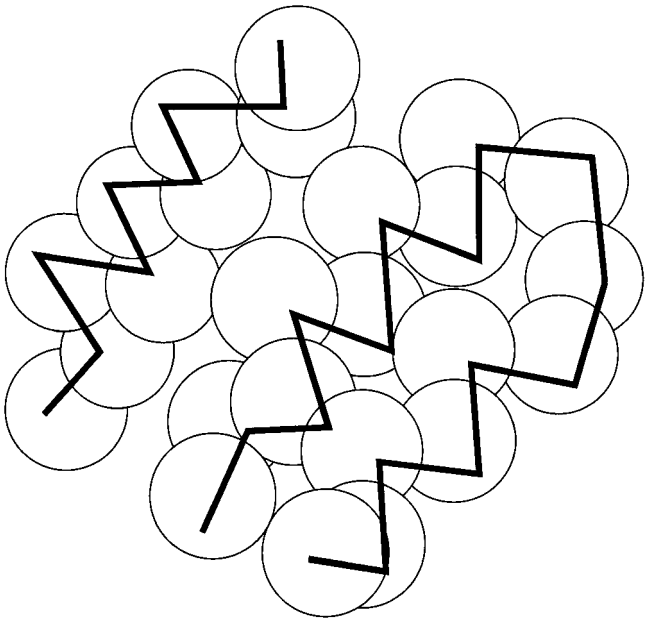

Fig. 5. Fragment of the model template chain (shown in the black sticks) and the template tube formed by the chain of spheres. The target chain (not shown in the drawing) is allowed to move in the tube with a penalty associated with all excursion from the tube.

signed to the corresponding element of the occupancy array. In the direction towards the center of mass of the template, the cut-off distance for creating the template "tube" is equal to $14^{1/2}$ instead of the $6^{1/2}$ value in the other direction. This fills in most of the volume occupied by the template structure. Figure 5 schematically shows such tubes surrounding the aligned fragments of the template chain (in solid lines). To illustrate the above-mentioned different width of the tube in the directions towards the center (versus the outside) of the template structure, the blobs forming the tube are shifted towards the center of mass of the template. This facilitates the close packing of the query (target) chain that wanders within the tube.

As described in the previous section, the starting model is placed into the template tube. The initial alignment provides an equivalence list between the template and target residue indices. This is called "the old assignment" in contrast to the "new assignment" which will be generated by the program. Both the old and the new assignments are then evaluated and updated in the following way:

a) At the very beginning of the simulation process, the old assignment (the original alignment) is copied into the new assignment list. The entries of these lists identify the tube compartments and the equivalent residues of the template chain. Then, all residues for which the total number of long distance ($i - j > 4$) contacts for a three-residue fragment (with the residue of interest as a central one) is smaller than two become non-assigned both in the old and new assignment lists. This erases those template fragments that do not interact with the rest of the model protein. Thus, "non compact" fragments of the template are ignored.
b) The new assignment is then updated when, for a steric reason (or due to local stiffness), the initial query chain residue simultaneously satisfies the following two criteria: (i) the bead of the query chain is farther away than five lattice units from the corresponding template residue of the original equivalence assignment ("old assignment"), and, (ii) the position of the query chain residue (the central point of the excluded volume cluster) coincides with a lattice point that is assigned to any other template residue. The number read from the appropriate element (occupied by the lattice chain) of the occupancy array that corresponds to the bead coordinates becomes the updated entry of the new equivalence list.
c) For all residues of the starting query chain that are farther away than nine lattice units from the equivalent (according to the old assignment) template residues, both old and new assignments are erased. These residues also become non-assigned. All allowed updates of the old assignments can only remove some entries from the equivalence list, which means that some part of the threading alignment is erased. The new assignments are dynamic (due to the updates described in b), and they have the character of a structural superposition, which is not sequential in many places.

This updated pair of assignments of the query chain residues to the template defines a flexible tube around the template chain. To keep the moving query chain in the neighborhood of the template, a set of biases is introduced. First, the model chain is kept in the broad vicinity of the original template (according to the updated old assignment list) by

$$E_{temp,o} = \Sigma\, \delta_o(i) f_r \max\{0, (|\mathbf{r}_i - \mathbf{r}_{oi}| - 9)\} \qquad (13)$$

where f_r is a constant (equal to 1kT in all simulations), $\mathbf{r}_i$ is the position of the query chain, $\mathbf{r}_{oi}$ is the position of the template and $\delta_o(i)$ is equal to 1 (0) when residue i is assigned (non assigned) according the old alignment.

Then, the residues of the query chain are similarly bonded to the template residues in the new assignment by

$$E_{temp,n} = \Sigma\, \delta_n(i) f_r \max\{0, (|\mathbf{r}_i - \mathbf{r}_{ni}| - R_t)\} \qquad (14)$$

where $\mathbf{r}_{ni}$ is the position of the initial template according to the new assignment and $\delta_n(i)$ is equal to 1 (0) when residue i is assigned (non assigned) according the new assignment. The constant R_t is equal to 7 (4) when residue i occupies any point of the template tube (the residue is outside the tube, i.e., the occupancy array at position $\mathbf{r}_i$ has value 0).

Additional restraints are the following:

$$E_{tube} = -E^{rep}\, \Sigma\, \{\delta_o(i)\delta_3(i) + \delta_n(i)\delta_t(i) + \delta_n(i)\delta_c(i)\} \qquad (15)$$

where $\delta_3(i)$ is equal to 1 when residue i of the query chain is at a distance smaller than 3 lattice units from the template

according to the old assignment, otherwise $\delta_3(i) = 0$. The second component, $\delta_t(i)$, is equal to 1 (0) when the residue is anywhere in the template tube (is outside). $\delta_c(i)$ is equal to 1 for a "quasi-continuous" alignment on the tube, i.e., when $|al(i-1) + al(i+1)|/2 - al(i) < 2$, where al(i) is the value of occupancy array in the tube for residue i of the query chain, otherwise $\delta_c(i) = 0$.

A small energy reward is also provided when the secondary structure of the query chain is consistent with the template structure. For all residues that are in extended or helical states (as defined in the loose conformational definition used for the generic short-range potentials) and that are in agreement with the secondary structure read from the corresponding fragments of the template protein, the system is stabilized by an energy equal to $-\epsilon_{gen}$.

With the above restraints, the system only pays a small energetic penalty for moving along the template tube (shifts in the alignment with possible lateral adjustment); however, the penalty is large for escaping from the loosely defined volume occupied by the template. For instance, it is possible (and this happened in a couple of cases studied here) that continuous fragments of the original alignments permute (this cannot be called an alignment in the conventional sense) by swapping their original tube compartments. This only occurs when the potential strongly favors such a rearrangement of the topology. The two assignments, carried out by the algorithm, play a different role. The "old" one bonds the model chain to the broad vicinity of the threading-based template. The "new" dynamic assignment is a compromise between the template restraints and packing requirements of the model chain.

Summary of the Threading Model Refinement Protocol

The entire model building procedure is illustrated in a flow-chart (see Figure 6) and can be outlined as follows:

a) Generate the threading alignment between the query sequence and the template structure.
b) Derive the sequence similarity-based short and long-range pairwise potentials. (Structures of proteins homologous to the query sequence are excised from the structural database; however, multiple alignments with homologous sequences of unknown structures were used in the potential derivation procedures.)
c) Build the starting continuous model chain onto the lattice-projected template structure.
d) Build the tube around the aligned fragments of the template structure. Then, perform the first stage of Monte Carlo refinement, where simulated annealing is done over a temperature range of 2–1. Since the Monte Carlo algorithm corrects unlike fragments of the alignment, the simulated annealing run is repeated two times. Subsequent runs have no systematic effect on the obtained models.
e) Refinement of the structure. The model obtained from the above simulations is assumed to be the new template, with a full length, complete self-alignment. The distance restraints from the new template are narrowed to 4 lattice units, and simulated annealing is performed over a narrower temperature range (1.5 to 1.0).
f) Selection of the lowest energy structures, by short isothermal simulations at T = 1, followed by building all-atom models using MODELLER.[24]

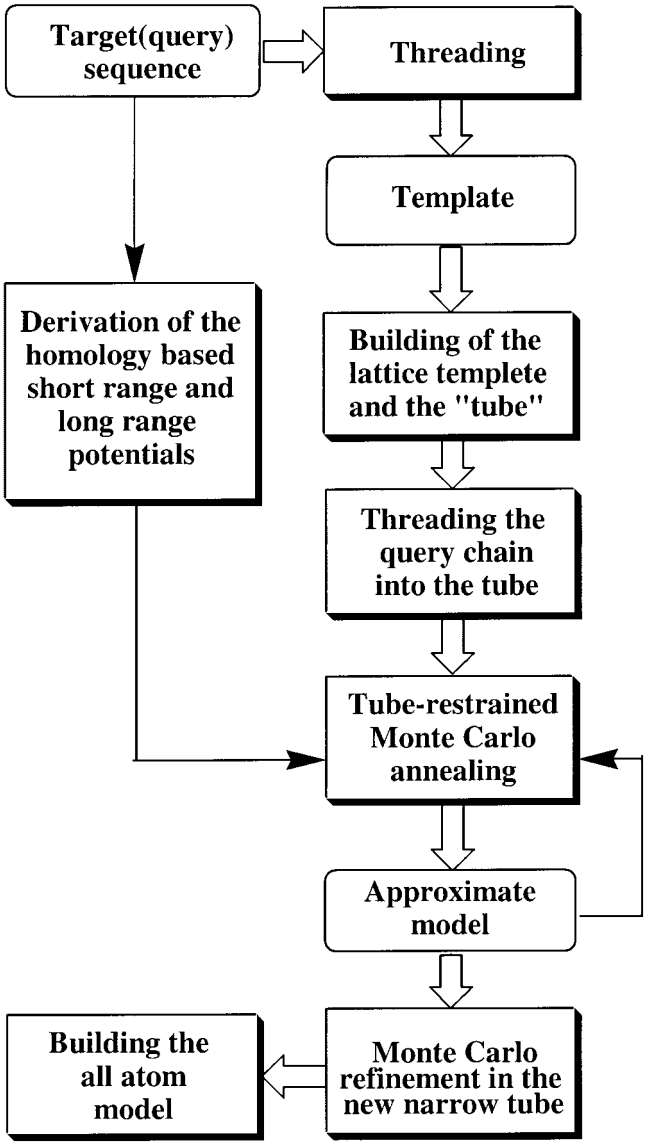

Fig. 6. Flow chart illustrating the molecular modeling procedure described in the text.

RESULTS

Test Proteins, Templates and Starting Alignments

Twelve pairs of target/template proteins of very low sequence similarity were selected for the present study. These proteins belong to various classes of small globular proteins, with the selected set being rather representative. As described in the Methods section, the relative scaling of the various potentials of the model force field has been adjusted in a series of ab initio folding simulations on

TABLE II. List of Target/Template Pairs Studied in This Work

Target protein			Template protein		
PDB code	Name	Length	PDB code	Name	Length
1aba_	Glutaredoxin	87	1ego_	Glutaredoxin	85
1bbhA	Cytochrome C	131	2ccy_	Cytochrome C	127
1cewI	Cystatin	108	1molA	Monellin	94
1hom_	Antennapedia protein	68	1lfb_	Transcription factor	77
1stfI	Papain	98	1molA	Monellin	94
1tlk_	Telokin	103	2rhe_	Immunoglobulin	114
256bA	Cytochrome c	106	1bbh_	Cytochrome C	131
2azaA	Azurin	129	1paz_	Pseudoazurin	120
2pcy_	Plastocyanin	99	2azaA	Azurin	129
2sarA	Ribonuclease	96	9rnt_	Ribonuclease	104
3cd4_	T-cell surface glycoprotein	178	2rhe_	Immunoglobulin	114
5fd1_	Ferrodoxin	106	2fxd_	Ferrodoxin	81

several (different from described here) small proteins. For the tuning of the template restraint contribution, we selected three proteins: 2pcy, 256b, and 1hom. These proteins belong to rather different structural classes: 2pcy is a quite irregular β-type protein with a very poor initial threading-based model, when the 2azaA template is used. 256b is a compact, four-helix bundle, where the original alignment appears to be quite good; however, the template and target structures have a different packing of helices that needs to be significantly readjusted to obtain a reasonable model. A very different example is 1hom. Here, the target fold is not very compact, and it is important to see if the proposed procedure can handle such small open structures. All proteins were subject to the previously described model building/refinement procedure. The list of these proteins is given in Table II. The threading alignments have been generated by a standard threading algorithm.[14] These alignments are compiled in Table III.

Compilation of the Modeling Results

Due to its stochastic character, the entire simulation procedure has been repeated several times for each case of the target template chains. The resulting structures were then subject to a refinement run. Namely, the algorithm employed in the first stage of the Monte Carlo modeling (starting from the initial, "old" threading-based alignment and performing all the updates of the alignment described in the "implementation of the template restraints" section) has been used in short isothermal runs at low ($T = 1$) temperature, with the final structure obtained at the end of the first stage of Monte Carlo used as input. At this temperature, the model does not change any of its global features, rather only local fluctuations are seen. The average conformational energy, which includes the intrinsic force field of the model and the effect of template restraints, was then used to select the "best" structure. The model has quite a strong, root-mean-square deviation, rmsd, versus energy correlation far from the native state. Closer to native state, the two quantities become uncorrected or the correlation is weak, depending on the particular case. It should be pointed that out that this refers to the entire force field (intrinsic terms and the template biases). A quite different situation is observed for the intrinsic force field alone; this has the strongest correlation of rmsd versus energy near the native structure (unpublished results). Since all our models are, at best, of moderate resolution, this criterion is no better than that based on the total energy. The lowest average (total) energy conformation from these short isothermal runs was selected for further consideration. For example, in the case of 1tlk, a structure that has a rmsd of 4.4 Å from native was selected, while several simulations resulted in structures about 3 Å from native.

Tables IV and V contain a compilation of the simulation results. In Table IV, the Cα rmsd from the native are compared for two kinds of molecular models. The first was generated using the initial threading template followed by automated modeling using MODELLER. We realize that the use of this homology-modeling tool in such a naive way is not the best practice; however, we wanted some means of comparison for two automated methods of model building from poor initial data. The second set of rmsd values is for the present lattice models, which for convenient comparison were converted into the full-atom models via an automatic application of MODELLER (with lattice models of the Cα backbones used as templates). As one may see, the most significant improvement of the model quality occurs when the threading alignment produces a rather poor but not nonsensical initial model (compare Table IV and Table V). As shown in Table V, for small globular proteins, such initial threading-based models have a rmsd in the range of 6–8 Å from native (over the aligned fragments). When the threading models are really bad, e.g., for 1cewI or 2azaA, the improvement is rather small. At the other extreme are those cases when the alignment is good and the resulting rmsd relatively small. Here also, the changes are small because the models are already good. Importantly, the procedure essentially does no harm to these models; thus, it can be applied to all situations with impunity. Moreover, as it is illustrated in Figure 7, the proposed modeling method systematically improves the entire threading-based model. The number of residues in structurally very accurate fragments (say less than 2 Å from native) is the same as for threading based models.

TABLE III. Starting Alignments Employed in Model Building

```
        12345678901234567890123456789012345678901234567890123456789012345678901234567890
1aba_: ....MFKVYGYDSNIHKCGPCDNAKRLL....TVKKQPFEFINI.MPEKGVFDDEKIAE..LLTKLGRDT
1ego_: MQTV...IFGRS....GCPYCVRAKDLAEKLSN.ERDDFQYQYVDIRAEGI.TKEDLQQKA.....GKPV
1aba_: QIGLTMPQVFAPDGSHIGGFDQLREYFK.....
1ego_: E...TVPQIFV.DQQHIGGYTDFAAWVKENLDA
        12345678901234567890123456789012345678901234567890123456789012345678901234567890
1bbhA: ..AGLSPEEQIETR....QAGYEFMG...WNMGKIKANLEGEYNAAQVEAAANVIAAIANSGMGALYGPG
2ccyA: QS...KPEDLLKLRQGLMQTLKSQWVPIAGFAAGKADLPADAAQRAENMAMVAKLAPIGWAKGTEAL.PN
1bbhA: .TDKNVGDVKTRVKPEFFQN..MEDVGKIAREFVGAANTLAEVAATGEAEAVKTAFGDVGAACKSCHEKY
2ccyA: G..........ETKPEAFGSKS.AEFLEGWKALATESTKLAAAAKAGP.DALKAQAAATGKVCKACHEEF
1bbhA: RAK
2ccyA: KQD
        12345678901234567890123456789012345678901234567890123456789012345678901234567890
1cewI: .....GAPVPVDE.NDEGLQRALQFAM.AEYNRASNDKYS.SRVVRVISA......KRQLVSGIK.YILQ
1molA: GEWEI...IDIGPF....TQNLGKFAVDEENKIGQYGRLTFNKVIRPCMKKTIYENERE....IKGYEYQ
1cewI: V...EIGRTTCPKSSGDLQSCEF....HDEPEMAKYTTCTFVVYSIP.WLNQIKLLESKCQ..
1molA: LYVY.........ASDKLFRADISEDY.......KTRGRKLLRFNGPV............PPP
        12345678901234567890123456789012345678901234567890123456789012345678901234567890
1hom_: MRKRGRQTYTRYQTLEL....EKEFHFNRYLTRRRR...........IEIAHALC..............
1lfb_: .............RFKWGPAS.QQI.LFQAYERQKNPSKEERETLVE.ECNRAECIQRGVSPSQAQGLGS
1hom_: ..LTERQIKIWFQNRRMKWKKENKTKGEPG
1lfb_: NLVTEVRVYNWFANR...RKEEAFRH....
        12345678901234567890123456789012345678901234567890123456789012345678901234567890
2pcy_: ...IDVLLGADDGSLAFVPSEFSISPGEKIVFK...........NNAGFPHNIVFDEDSIPSGVDASKIS
2azaA: AQC.EATIESND.AMQYDLKEMVVDKSCK.QFTVHLKHVGKMAKSAMG..HNWVLTKEADKEGVATDGMN
2pcy_: MSEEDLLNA..............KGETFEVAL.....SNKGEYSFY.CSPHQGAGMVGKVTVN..
2azaA: AGLAQDYVKAGDTRVIAHTKVIGGGESDSVTFDVSKLTPGEAYAYFCSFPGHWAMMKGTLKL.SN
        12345678901234567890123456789012345678901234567890123456789012345678901234567890
1stfI: .MMSGAPSATQPATAETQ.HIADQV.RSQLEE.KYNKK.FPV.FKAVSFK.....SQVVAGTNYFIKVHV
1molA: G.......EWEIIDIGPFTQNLGKFAVDEENKIGQYGRLTFNKVIRPCMKKTIYENEREIKG.YEYQLYV
1stfI: GDEDFVHLRVFQSLPHENKPLTLSNYQTNKAKHDELTYF
1molA: YASDKLFRADI.SEDYKTRGRKLLRF...NGPVPPP...
        12345678901234567890123456789012345678901234567890123456789012345678901234567890
1tlk_: VAEEKPHVKPYFTKTILDMD......VVEGSAARFDCKVEGY.........P.....DPEVMWFKDDNPV
2rhe_: ............ESVLTQPPSASGT..PGQRVTISCTGSATDIGSNSVIWYQQVPGKAPKLLIYYNDLL
1tlk_: KES.RHFQIDYDEEGNCSLTISEVCGDDDAKYTCKAVNSLGEAT.....CTAELLVETM..
2rhe_: PSGVSDRFSASKSGTSASLAISGLESEDEADYYCAAWNDSLDEPGFGGG..TKLTVLGQPK
        12345678901234567890123456789012345678901234567890123456789012345678901234567890
256b_: ..ADLEDNMETLNDNLKV..........IEKADNAAQVKDALTKMRAAALDAQKAT.PPKLEDKSPD.S
1bbhA: AGLSPEEQIETRQAGYEFMGWNMGKIKANLEGEYNAAQVEAAANVIAAIANSGMGALYGPGTDKNVGDVK
256b_: ....PEMKDFRHGFDIL......VGQIDDALKLANEGKVKEAQAAAEQLKTTRNAYHQKYR..
1bbhA: TRVKPEF..FQNMEDVGKIAREFVGAANTLAEVAATGEAEAVKTAFGDVGAACKSCHEKYRAK
        12345678901234567890123456789012345678901234567890123456789012345678901234567890
2azaA: AQCEATIESNDAMQYDLKEMVVDKSCKQFTVHLKHVGKMAKSAM.....GHNWVLTKEADKEG.....VA
1paz_: .........................ENIEVHM..LNKGAEGAMVFEPA...YI...KANPGDTVTFIPV
2azaA: TDGMNAGLAQDYVKAGDTRV.IAHTKVIGGGESDSVTFDVSKLTPGEAYAYFCS.FPGHWA..MMKGTLK
1paz_: DKGHNVESIKDMIPEGAEKFK........SKINENYVLTVTQ..PG.AYLVKCTP...HYAMGMI.ALIA
2azaA: LSN.........................
1paz_: VGDSPANLDQIVSAKKPKIVQERLEKVIA
        12345678901234567890123456789012345678901234567890123456789012345678901234567890
2sarA: ............DVSGTVCLSALPPEATDTLNLIAS.DGPFPYSQDGV....VFQNRESVLPTQSYGYY
9rnt_: ACDYTCGSNCYSS............SDVSTAQAAGYKL...HEDGETVGSNSY.PHKYNNYEGFDFSVSS
2sarA: HEYTV.........ITPGARTRGTRRIICGEATQEDYYTGDHYATFS...LIDQTC.
9rnt_: PYYEWPILSSGDVY..SGGSPGADRVVFN...ENNQLAGVITHTGASGNN..FVECT
        12345678901234567890123456789012345678901234567890123456789012345678901234567890
3cd4_: K.....KVVLGKKGDTVELTCTASQKKS....IQFHWK..NSNQIKILGNQGSFLTKGPSKLNDRAD..S
2rhe_: ESVLTQPPSASGTPGQRVTISCTGSATDIGSNSVIWYQQVPGKAPKLLIYY...NDLLPSGVSDRFSAS.
3cd4_: RRSLWDQGNFPLIIKNLK.....IEDSDTYICEVEDQKEEVQLLVFGLTANSDTHLLQGQSLTLTLESPP
2rhe_: ....KSGTSASLAISGLESEDEADYY...CAAWNDSLDEPG.............FGGGTKLTVLGQP
3cd4_: GSSPSVQCRSPRGKNIQGGKTLSVSQLELQDSGTWTCTVLQNQKKVEFKIDIVVLA
2rhe_: K.....................................................
        12345678901234567890123456789012345678901234567890123456789012345678901234567890
5fd1_: AFVVTDNCIKCKYTDCVEVCPVDCFYEGPNFLVIHPDEC.IDCALCEPECP.AQAIFSEDEVPEDM..QE
2fxd_: ........................PKYTIVDKETCI.....ACGACGAAAPDIYDYDEDGIAYV.T
5fd1_: FIQLNAE....LAEVWPNITE.KKDPLPDAEDWDGVKGKLQHLER...
2fxd_: LDDNQGIVEVP.DILIDDMMDA..FEGCPTDSIKVADEPFDGDPNKFE
```

TABLE IV. Alpha Carbon Rmsd From Native for Models Built From the Initial Threading Alignments and Refined by Lattice Simulations†

Target protein	Threading +MODELLER	SICHO +MODELLER
1aba_	4.43	4.86
1bbhA	6.77	6.82
1cewI	14.96	14.38
1hom_	7.82	3.70
1stfI	6.40	5.95
1tlk_	7.23	4.17
256bA	6.09	4.36
2azaA	21.95	10.77
2pcy_	6.56	4.41
2sarA	10.28	7.83
3cd4_	6.74	6.39
5fd1_	25.67	12.40

†Note: The threading +MODELLER models use the threading alignments (for the aligned residues) as the target for all-atom reconstruction. SICHO models are the reduced lattice models obtained by the method described in this work. The final all-atom model is also built by MODELLER using as a target the lattice model alpha carbon positions estimated from the SICHO lattice model. The values of the rmsd for alpha-carbon traces (in Å) are given for the structured parts of the target molecules (1hom_: residues 7–59, 1tlk_: residues 9–103, 3cd4_: residues 1–97 i.e., the first domain).

TABLE V. Alpha Carbon Rmsd (in Å) From Native for Threading-Based Models and Lattice SICHO Models Built by MODELLER. Comparison for Threading-Aligned Fragments Only†

Target protein	Starting RMSD	Threading RMSD	SICHO RMSD	LENGTH
1aba_	4.37	3.89	4.40	69
1bbhA	7.03	6.35	6.69	116
1cewI	12.88	12.37	10.74	69
1hom_	5.59	5.34	3.45	40
1stfI	7.05	6.04	4.73	83
1tlk_	7.88	7.15	3.94	86
256bA	6.92	6.06	4.37	104
2azaA	11.04	13.53	9.94	80
2pcy_	7.64	6.65	4.36	94
2sarA	8.28	8.07	7.60	73
3cd4_	5.72	5.56	5.22	82
5fd1_	12.38	12.18	11.94	69

†Note: The starting rmsd is for the set of threading-aligned residues of the template from the equivalent native target coordinates. The MODELLER models use the threading alignments and an all-atom target. SICHO models are the all-atom models built by MODELLER using the lattice models (only Cα) as a target. The length of the alignments is given in the last column, and the same sets of residues are compared for both methods.

For the remaining residues, a large average decrease of the distance from target coordinates could be observed in all cases (see Fig. 7). We have also run a number of test simulations employing the structure of the query protein as a template (with full and randomly selected "align-

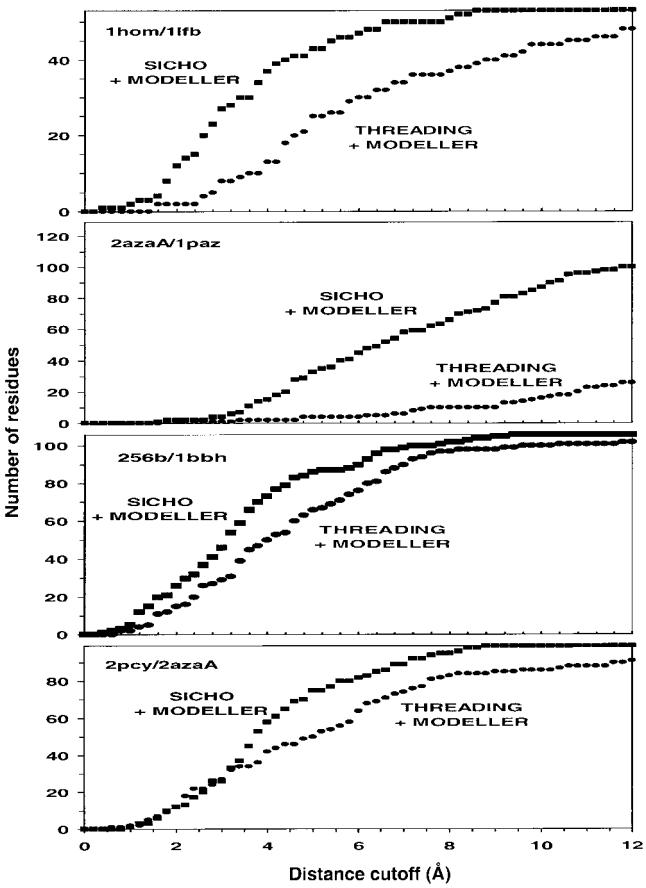

Fig. 7. Comparison of the accuracy of the threading/MODELLER structure with SICHO/MODELLER structures. The number of alpha carbon atoms whose distance from the native structure is less than a given cut-off is plotted as a function of the cut-off value for four example proteins.

ments" fragments). In all cases, the resulting structures were 2.5–4.0 Å rmsd (depending on protein size) from the native structures. Thus, given a good initial alignment, the model stays in the vicinity of that structure, thereby demonstrating that this approach "does no harm".

In summary, in 6 of 9 test cases (in 9 of 12, including the three proteins employed in the model tuning procedure), the models generated by the method proposed here give lower values of rmsd over the set of aligned residues than that found in the initial structure. In the three remaining cases, the changes in rmsd are insignificant (essentially in the range of the statistical fluctuations). In five cases, qualitative improvements were observed (for the aligned residues as well as for entire models; compare data given in Table IV): from 5.6 Å to 3.5 Å for 1hom, from 7.1 Å to 4.7 Å for 1stfI, from 7.9 Å to 3.9 Å for 1tlk, from 6.9 Å to 4.4 Å for 256b or from 6.6 Å to 4.4 Å for 2pcy. These numbers are for the initial threading and final lattice (refined with MODELLER) models, respectively. It should be noted that the MODELLER refinement of the final lattice models changes their rmsd very little (in the range of 0.2 Å), while the improvement of the initial threading models by the application of MODELLER is more noticeable.

It is very interesting to see how the proposed procedure deals with the non-aligned part of the model. Comparison of the rmsd values for the aligned parts (Table V) and for the entire structured parts (Table IV) of the model shows that the algorithm builds rather reasonable models of the entire structure, provided there is a well-defined fragment of good geometrical fidelity in the original alignment. Again, in all but two cases, the present method leads to more accurate models. For both the aligned part of the molecules and for entire chains (Table IV), good models are generated in about half of the studied cases (including all three proteins used in the model-tuning procedure). In the remaining cases, one may see models that are marginally improved as for 3cd4 or that remain rather poor final models as for 2azaA or 5fd1; this is true despite an rmsd decrease of more than 10 Å, as compared to models generated automatically by MODELLER from the initial threading results. However, as demonstrated in the second panel of Figure 7, it should be pointed out that in these cases, large fragments of the structure qualitatively improve. Indeed, about 40 residues of 2azaA in our model are closer than 6 Å from the native structure, while the threading/MODELLER model structure has only five residues in this range. A possible way to improve the performance of the method in the case of very bad templates is to loosen the template restraints and perform the annealing from higher temperature. This, however is closer to ab initio folding with weak homology based restraints and will be addressed elsewhere.

DISCUSSION

Means of the Model Improvement

There are several ways in which the described algorithm changes the protein model from the original fragmentary threading model. The first is rather trivial in that the non-aligned parts (mostly loops) are added and readjusted according to packing requirements and the preferences encoded in the force field. Then, the entire chain has some freedom of movement within the template tube without any changes in its template-target sequence assignment. Furthermore, parts of the chain can slide along the tube, thereby allowing for a quite substantial modification of the initial alignment and, consequently, the resulting structure. Finally, the aligned fragments can leave the tube in a lateral direction. These segments can enter a different part of the template tube or remain outside of it. Such motions of the model chain could result in a large change of the structure, or even change the fold topology. The last, rather radical mode of the model rearrangements, happened in several cases. In other words, the most effective way of model improvement was by neglecting a part of the threading alignment, even at the expense of various template-related energetical penalties. Interestingly, those sections of the threading-based model consistent with the target structure undergo only very minor changes in all cases, and the alignment remains unchanged. As discussed below, this observation may help identify those models that should be of good quality from those for which improvement of the starting threading model is not satisfactory.

Below, for three selected cases, we analyze in more detail specific rearrangements of the initial threading models that take place during the Monte Carlo simulations.

2pcy case

The threading alignment of the 2pcy sequence on 2azaA covers a substantial part of the sequence. There are gaps of substantial length. As a result, the threading model has the wrong topology, and two-edge strands of the eight-member β-barrel (one in each of the two β-sheets) are located in the wrong sheets. This is the reason for the resulting 7.6 Å rmsd from native for the models built solely from the threading alignment. During the simulations, the three C-terminal strands remain almost unchanged. Similarly, the three N-terminal strands undergo only small adjustments; however, in several models, one or two strands slide along the tube by a distance that sometimes changes the original alignment by one or two positions. The central fragment of the model chain (two putative irregular strands, with a couple of short helices breaking these strands) is responsible for the large rmsd in the initial model. The algorithm erases most of the template-target assignments in this part of the molecule. Partly this occurs because of the compactness criterion; several residues do not have any long-range contacts in the threading model. During the simulated annealing process, residues 30–37 (small differences in the extension of this fragment can be seen between the particular runs) switch their sheet assignment, and join the tube fragment associated with one of the C-terminal β-strands, the third one from the C-terminus. This is seen in the final "new assignment," or pseudo alignment. At the same time, the second strand (completely helical in the threading model) moves to the second sheet, and the long helix breaks and becomes distorted, as actually occurs in 2pcy's native structure. Most of the displaced residues join the tube fragments generated by various secondary structural elements of the template, but only a few maintain their original assignments to the template tube. This way the internal force field of the lattice model overrides the target interactions, significantly correcting the threading model. The initial model and the final model are compared with the native structure in Figure 8, where stereo alpha-carbon traces are displayed in their best mutual superposition, using the MOLMOL[25] drawing program.

256bA case

This molecule is a four-helix bundle and the threading alignment has a few gaps. The template structure is very similar to the target, but the threading model is not very good. During the simulations, most of the C-terminal helical hairpin remains almost unchanged, except for the loop region that is very mobile. The third (first helix of the C-terminal hairpin) helix of the model is the most stable. The N-terminal hairpin undergoes a large-scale rearrangement. The second helix undergoes a rotation that changes its packing angle with respect to the remainder of the

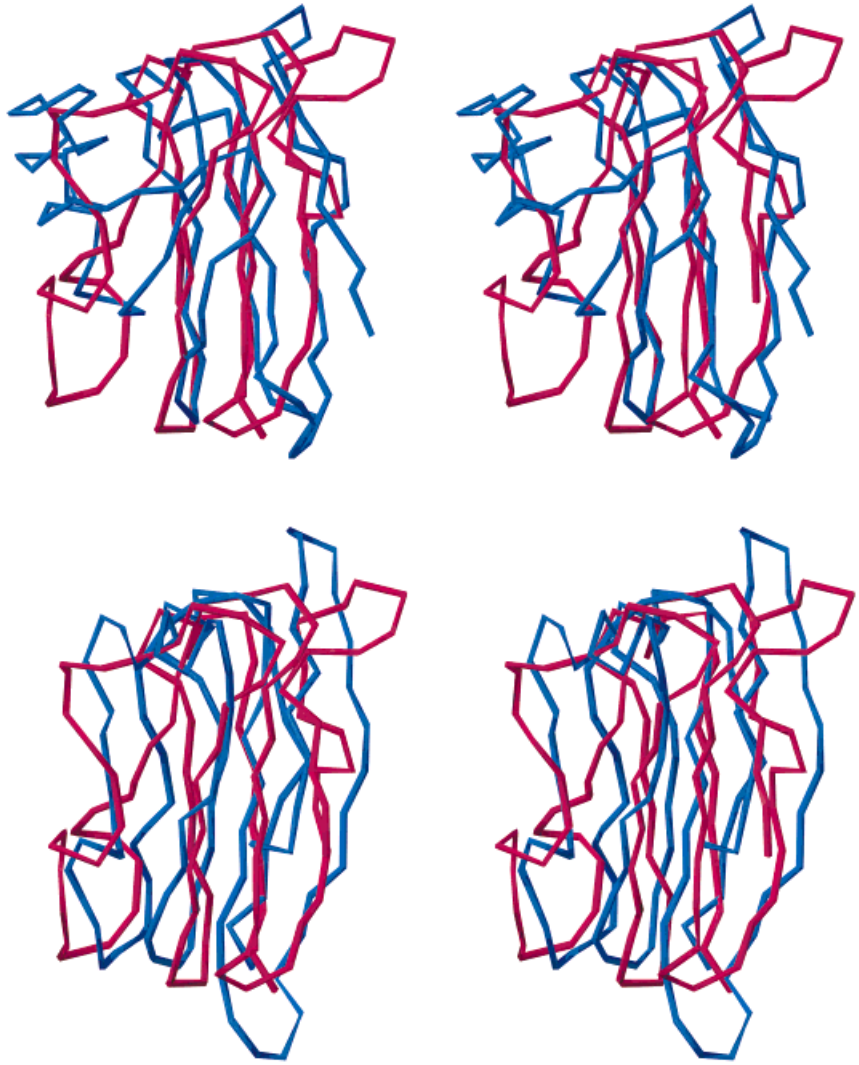

Fig. 8. Stereo drawings of the two models of plastocyanin (in red) superimposed onto crystallographic structure 2pcy (in green). The upper panel shows the model obtained by MODELLER from the threading alignment, the lower panel shows the model obtained by the procedure described in this work. For the sake of readability, only the alpha carbon traces are displayed.

molecule. As a result, the end of this helix moves by about 7 Å in a lateral direction, while the beginning of this helix stays close to its original position. The largest changes are observed for the first N-terminal helix. It moves along the tube, changing assignment indices by several (up to 8) residues; a lateral adjustment takes place as well. The initial model and the final model (superimposed onto the native structure) are compared in Figure 9. The helical regions of the final model are very close to the native structure; the largest errors that account for most of the structure errors are in the central turn/loop region.

1tlk case

Telokin is a quite regular β-protein. Again, due to gaps and insertions, the threading model has produced a model whose topology is wrong. During the simulations, one of the β-strands from the original model leaves the initial assignment and sticks to the tube of a strand from the opposite sheet. Two β-strands that are not in the threading model (lack of the alignment assignments) are built in the simulated annealing procedure, and they join tubes associated with existing strands. The entire structure, except for the last C-terminal β-strand that remains essentially unchanged, rearranges substantially. Mostly lateral (orthogonal to the local direction of the template tube) displacements occur in the range of 6 Å for about half of all the residues. As a result, the model improves its rmsd by almost 4 Å. The initial model and the final model (superimposed onto the native structure) are compared in Figure 10.

How to Identify Good Models

As mentioned before, the proposed method generates low to moderate resolution models of correct topology in those cases when the initial threading-based alignment leads to at least a partially correct structure, i.e., when a

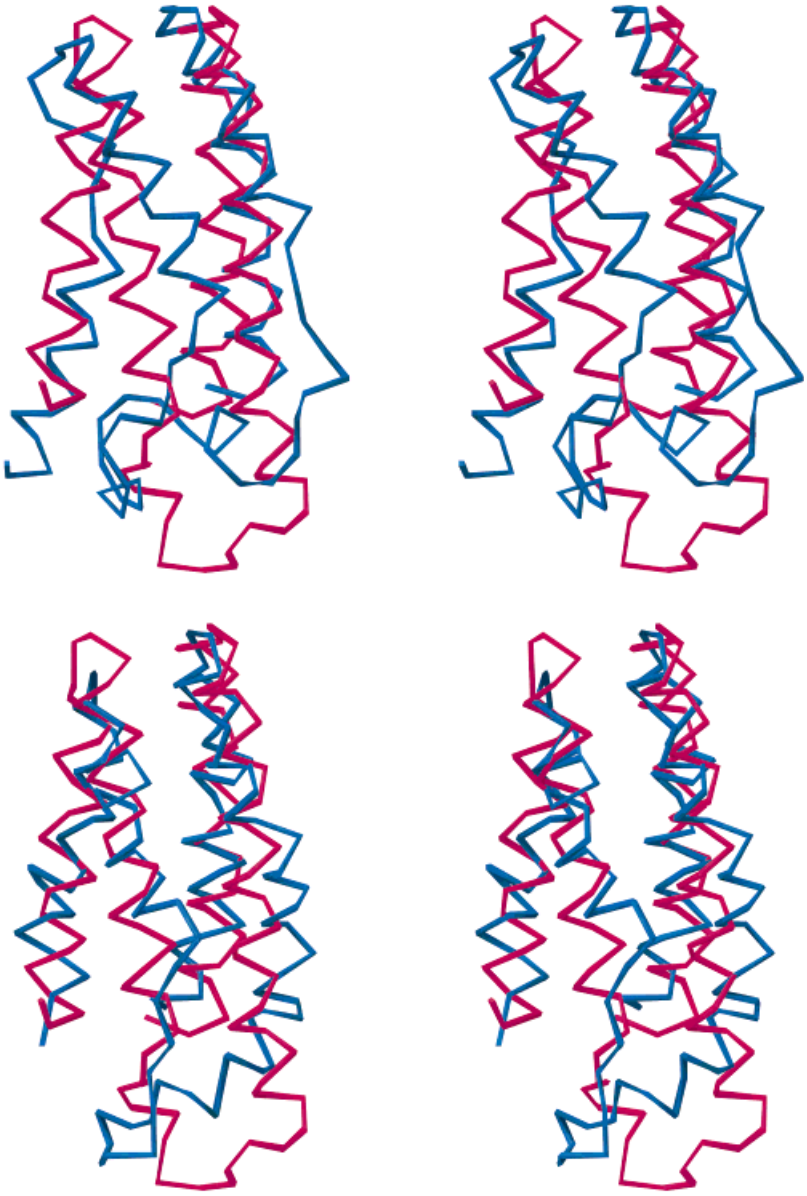

Fig. 9. Stereo drawings of the two models of the cytochrome 256b (in red) superimposed onto crystallographic structure (in green). The upper panel shows the model obtained by MODELLER from the threading alignment, the lower panel shows the model obtained by the procedure described in this work. For the sake of readability, only the alpha carbon traces are displayed.

part of the identified template is close to the target structure. How to (a priori) distinguish between a good (threading-based) alignment from a poor one is a non-trivial question. Unfortunately we do not have a general solution to this problem.

For example, it might appear that the cases where the initial alignment is more continuous should lead to better final model. In reality, no such correlation was observed. For instance similar qualitative (by more than 2 Å) improvements of the model were observed for two helical proteins, 1hom and 256bA. This is true in spite of the fact that in the first case the fraction of aligned residues is much less than in the second case where threading aligns 104 of the 106 residues in the target sequence.

The intrinsic force field of the reduced model correctly identifies the native structure (the lattice projection) as the lowest energy conformation when compared with the models generated by MODELLER from the initial threading alignments. The models obtained by lattice homology modeling are described in this work. In all cases, except one (1bbhA, where MODELLER gave a slightly better result than the present method) the energy of the models built by the present method is significantly lower than other worse models (including these built by automatic use

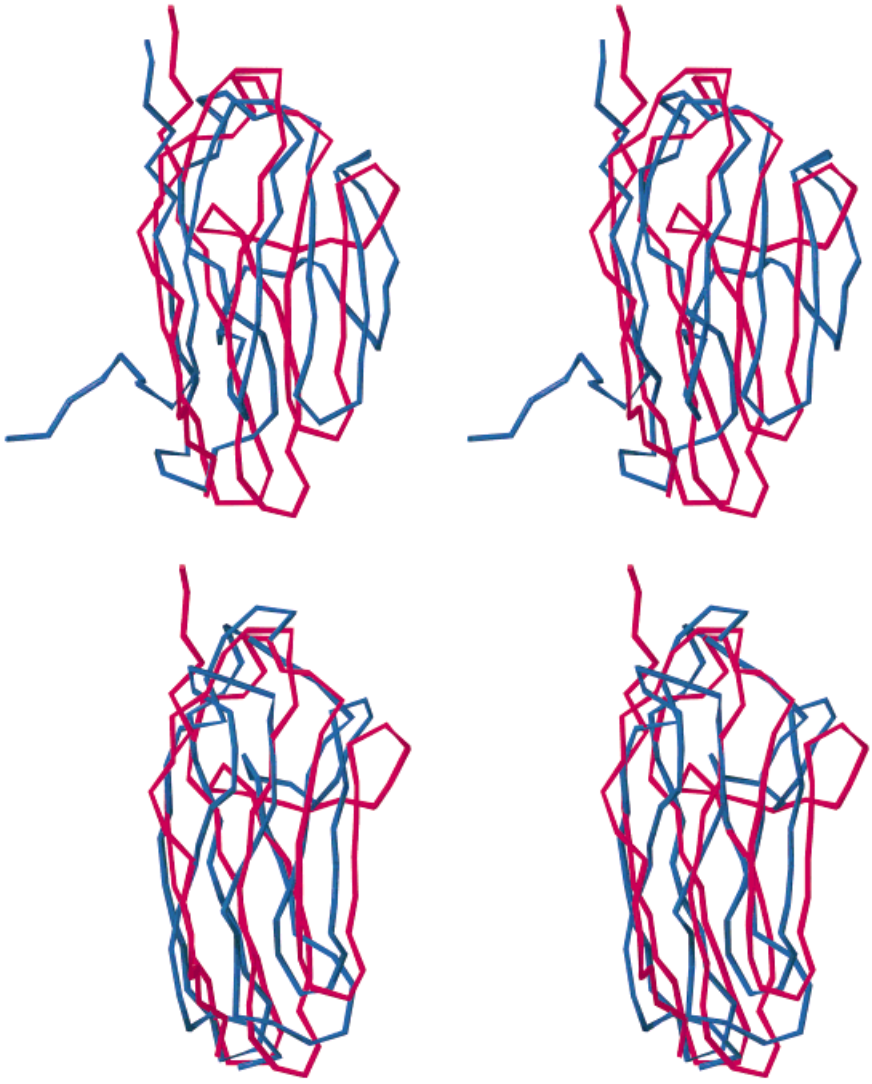

Fig. 10. Stereo drawings of the two models of telokin (in red) superimposed onto crystallographic structure 1tlk (in green). The upper panel shows the model obtained by MODELLER from the threading alignment, the lower panel shows the model obtained by the procedure described in this work. For the sake of readability, only the alpha carbon traces are displayed.

of MODELLER). These are interesting results; however, our goal is to identify those target/template pairs where the final model is of reasonable quality from those cases where, despite a sometimes large improvement of the initial models, the resulting structures are still far from the native target conformation. Unfortunately, simple energetic criteria (conformational energy per residue in the final model, decrease of energy from the starting model to the final model, etc.) do not enable identification of these poor quality structures.

In the previous section, we discussed how the proposed modeling procedure improves the initial, threading-based model. This could actually be used for a qualitative identification of better models. Consider the displacement of particular residues (as a function of their position along the chain) during the entire simulation procedure. In those cases where the final model is of good quality, the plots indicate relatively well separated regions where the chain modifications were small and also indicated regions of large modifications. This is consistent with the previously mentioned characteristic behavior of "good" models, for which some fragments of alignments are recognized by the procedure as being very good and behave as a scaffold for readjustment of the remainder of the protein. In contrast, poor models are characterized by random fluctuations of the spatial amino acid displacements along the sequence. In such cases there is no pattern. Perhaps there is a huge energy barrier between the starting model and the better, near native models that cannot be surmounted by partial readjustment of the initial alignment. Examples of both situations are given in Figures 11 and 12. The lowest (and locally similar) displacement (during the modeling proce-

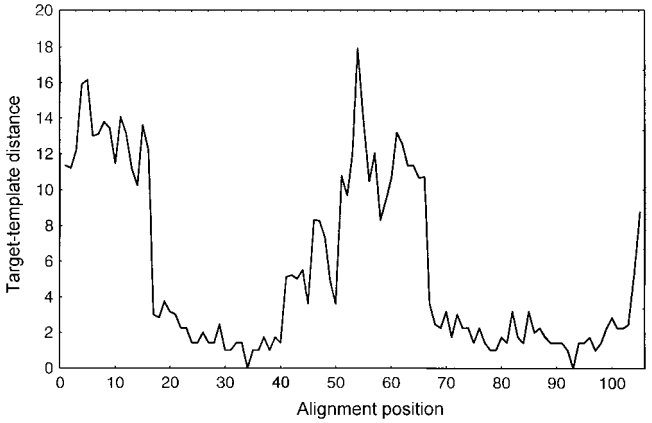

Fig. 11. Displacement of the model chain units during the Monte Carlo simulation as a function of the position along the chain for the aligned portion of the 256b molecule. The very stable (most of the second helix and C-terminal hairpin) regions and very mobile regions (the first helix and the central loop region) are clearly separated. This is the pattern typical for successful modeling (relatively low final rmsd from the native structure).

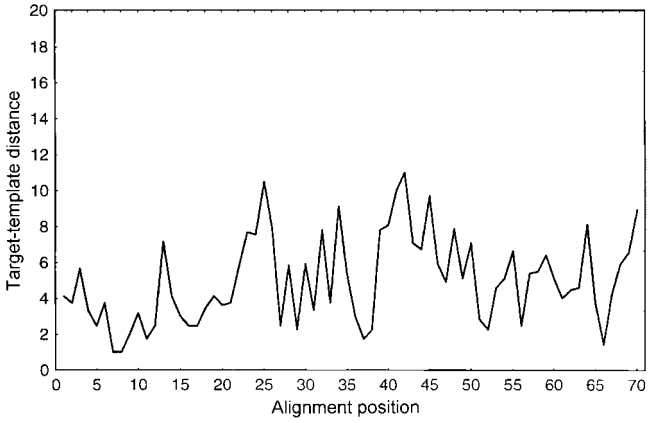

Fig. 12. Displacement of the model chain units during the Monte Carlo simulation as a function of the position along the chain for the aligned portion of the 5fd1 molecule. In contrast to the case of 256b (see Fig. 10) the displacements of the chain elements are essentially random. This kind of pattern suggests a rather poor quality final model.

dure) regions identify the regions of an optimal (or very close to optimal) alignment. While the above is not easy for a simple quantification, it still can be used as a heuristic criterion for the identification of cases where the method proposed in this work is likely to provide relatively good, low resolution models. Figure 13 shows the plot of model accuracy (measured as the alpha carbon rmsd from native) as a function of the variability in the model chain mobility during the simulations. Unfortunately, the correlation is not very strong. Consequently, the mobility criterion has to be used with caution. Rather, plots as given in Figures 11 and in 12 can be used to identify the best fragments of the threading models. Indeed, there are very strong correlations between the lowest mobility and the best structural fidelity (to the target structure) of the model chain fragments. This may have some other applications, where

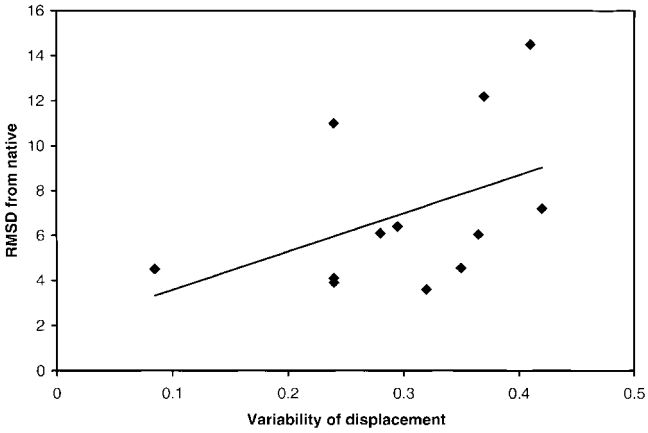

Fig. 13. Accuracy of the final models, measured as the C_α rmsd from the native structure, as a function of displacement variation. The variation is defined as a ratio of the number of passages of the residue displacement plot (as given in Figs. 10 and 11) through the line of average displacement to the total number of protein residues.

assessment of the reliability of various parts of a model structure is needed.

SUMMARY AND CONCLUSION

In this work, we proposed a novel approach to prediction of low-resolution protein structures that is based on homology or sequence-structure compatibility. The method employs templates obtained from threading procedures. It should be noted that the alignments used in this work belong to the best alignments available from threading procedures.[5] Of course, the alignments can also be obtained from sequence alignments. Such templates are used to guide Monte Carlo simulations that employ a reduced protein chain representation. In about a third of the studied cases, the procedure is capable of making large structural rearrangements that lead to qualitative improvements in the initial poor models. In some other cases, despite a huge decrease in the rmsd between the model and the target native structure, the final model was still unsatisfactory. Analysis of the simulation trajectories allows for a plausible identification of those cases where the final models qualitatively improved with respect to the initial, threading-based model.

While the described method needs further improvement (better resolution, better procedure for the model validation), even now it may be useful for large-scale protein structure and function prediction. A complete series of simulations for a single target/template pair could be now performed in 24 hours on a single state-of-the-art computer. The process itself could easily be automated. Thus, predictions on a genomic scale are quite feasible and will be attempted in the near future. In this regard, it is possible to identify the biochemical function of a protein function having a model with a 5–6 Å backbone rmsd.[9,10] Certainly, that would be much more difficult, if not impossible, for a model with an 8 Å Cα rmsd from native. For example, the model of plastocyanin (2pcy) generated by the proposed method has its four copper-binding residues

much closer to their native position than the threading-based model does. Thus, having a structural template of this active site, the model structure can be identified with high fidelity as a copper-binding protein. In a substantial fraction of cases, function annotation based on structures provided by the proposed method would certainly fail, due to the above-discussed problems in the identification of good quality models. Nevertheless, it appears from the present studies that for many new proteins that cannot be annotated by other simpler methods, their function could be identified. Thus, the proposed method is complementary to sequence-based and threading methods and provides a means for improvement of initially poor and incomplete models. On the other hand, it is also complementary to standard homology modeling tools, enabling homology modeling in those cases where the template is structurally very far from the target structure.

ACKNOWLEDGMENTS

This work was partially supported by NIH Grant No. GM-48835 and KBN (Poland) grant GR-919. AK is an International Scholar of the Howard Hughes Medical Institute (HHMI grant #75195-543402). We would like to express our thanks to an anonymous referee for his/her comments that helped us to greatly improve the presentation of this work.

REFERENCES

1. Bowie JU, Luethy R, Eisenberg D. A method to identify protein sequences that fold into a known three dimensional structure. Science 1991;253:164–170.
2. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
3. Godzik A, Skolnick J, Kolinski A. A topology fingerprint approach to the inverse folding problem. J Mol Biol 1992;227:227–238.
4. Miller RT, Jones DT, Thornton JM. Protein fold recognition by sequence threading tools and assessment techniques. FASEB 1996;10:171–178.
5. Zhang B, Jaroszewski L, Rychlewski L, Godzik A. Similarities and differences between nonhomologous proteins with similar folds: evaluation of threading strategies. Fold Des 1997;2:307–317.
6. Hu W-P, Godzik A, Skolnick J. On the origin of sequence-structure specificity. How does an inverse folding approach work? Protein Eng 1997;10:317–331.
7. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through folding motif. Proteins 1993; 16:92–112.
8. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. Proteins 1995;23:356–369.
9. 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 T_1ribonucleases. J Mol Biol 1998;281:949–968.
10. Fetrow J, Godzik A, Skolnick J. Functional analysis of the *Escherichia coli* genome using the sequence-structure-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxireductase activity. J Mol Biol 1998;282: 703–711.
11. Sali A, Overington JP, Johnson MS, Blundell TL. From comparison of protein sequences and structures to protein modeling and design. TIBS 1990;15:235–250.
12. Aszodi A, Tylor WR. Homology modeling by distance geometry. Fold Des 1996;1:325–334.
13. Jaroszewski L, Pawlowski K, Godzik A. Multiple model approach: exploring the limits of comparative modeling. J Mol Modelling 1998;4:294–309.
14. Jaroszewski L, Rychlewski L, Zhang B, Godzik A. Fold prediction by a hierarchy of sequence, threading, and modeling methods. Protein Sci 1998;7:1431–1440.
15. Wodak SJ, Rooman MJ. Generating and testing protein folds. Curr Opin Struct Biol 1993;3:247–259.
16. Kolinski A, Jaroszewski L, Rotkiewicz P, Skolnick J. An efficient Monte Carlo model of protein chains. Modeling the short-range correlations between side groups centers of mass. J Phys Chem 1998;102:4628–4637.
17. Kolinski A, Skolnick J. Assembly of protein structure from sparse experimental data. An efficient Monte Carlo model. Proteins 1998;32:475–494.
18. Kolinski A, Skolnick J. Lattice models of protein folding, dynamics and thermodynamics. Austin, TX: RG Landes; 1996. 200 p.
19. Skolnick J, Kolinski A, Ortiz AR. MONSSTER: A method for folding globular proteins with a small number of distance restraints. J Mol Biol 1997;265:217–241.
20. Bernstein FC, Koetzle TF, Williams GJB, et al. The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
21. Binder K. The Monte Carlo method in condensed matter physics: Institut Für Physik, Johannes Gutenberg-Universität, Mainz, 1991.
22. Godzik A, Skolnick J, Kolinski A. Regularities in interaction patterns of globular proteins. Protein Eng 1993;6:801–810.
23. Milik M, Kolinski A, Skolnick J. Neural network system for the evaluation of side chain packing in protein structures. Protein Eng 1995;8:225–236.
24. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
25. Koradi R. MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 1996;14:51–55.
26. Altschul SF, Madden TL, Schaefer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res 1997;25:3389–3402.
27. Hobohom U, Scharf M, Schneider R, Sander C. Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Sci 1992;1:409–417.
28. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.

APPENDIX A

Derivation of the Homology-Enhanced Short-Range Statistical Potentials

The input sequence is compared to the non-redundant sequence database (containing over 300,000 sequences) using the PSI-BLAST program.[26] This process is iterated until a stable multiple sequence alignment is obtained. Next, the sequence profile is built using this set of homologous sequences. The profile is a two-dimensional array of size 20×N where N is the input sequence length and each of 20 positions corresponds to the frequency of a given amino acid in the multiple sequence alignment. Let $o_a(i)$ will be a number of occurrences of amino acid type a at multiple sequence alignment position i. Then profile $P(a, i)$ can be defined in the following way:

$$P(a, i) = o_a(i)/\mathrm{M} \quad \mathrm{i} = 1,2, \ldots \mathrm{N} \tag{A1}$$

where M is a number of aligned sequences.

In the next step, a representative set of PDB files[27] of proteins of known structure (PDBSELECT'97) is scanned to find the fragments of sequences similar to the profile. The comparison is done using a 21-amino acid window of sequences and the BLOSUM80 mutation matrix[28] as the scoring function.

$$s_{i,j} = \Sigma_{\mathrm{k}=-10,10} P(D(j + k), i + k) \tag{A2}$$

where $s_{i,j}$ is the score of a given fragment (at position i of the profile and position j of database sequence D), D is a given database sequence, k denotes the scanning window position, $D(j + k)$ is the amino acid type at position $j + k$ of sequence D.

The 200 best scoring sequence fragments (and corresponding pieces of structures) are stored. These presumably exhibit significant structural similarity to the corresponding fragments of the query sequence. Then, the side group atom positions of these fragments are used in the calculation of the potentials. For each of these fragments, various distances (based on side chain + C_α centers of mass) are calculated. For computational convenience, these distances are discretized into a number of bins. Five bins have been assumed for the distances between the i-th and i+1st residue; six bins for the distances between i-th and i+2nd residue; eight bins for the distances between the i-th and i+3rd residues; seven bins have been assumed for the i, i+4 distances; nine bins for the i, i+6 distances and seven bins for the i, i+8 distances. The distances i, i+3 and i, i+6 have been assumed to be "chiral," i.e., negative values have been assigned to the left handed fragments and positive to right-handed fragments, respectively. For the two shortest distances, conservation of the identity of the flanking residues was enforced in the fragment alignment procedure. Then, weighted histograms are built by summation of the number of occurrences of a particular bin at a given sequence position (using the mutation matrix score as a weight).

$$H(r_{lx,b}) = 1/200\ \Sigma_{m=1,200} s_m(r_{lx,b}) \qquad \text{(A3)}$$

where $H(r_{1x,b})$ is the histogram value, $r_{1x,b}$ is the b-th bin of distance r_{1x}, ($1x$ is the short-hand notation for the distances between residues i and i+x), sm ($r_{1x,b}$) is the score of fragment that belongs to the bin b and m is the number of fragments.

Finally, the potentials for particular distances along the chain are calculated using the obtained histograms and a random statistical distribution as the reference state.

$$E = -ln\ (H(r_{1x,b})/H^o(r_{1x,b})) \qquad \text{(A4)}$$

where the denominator corresponds to the histograms averaged over the database (ignoring sequence similarity and amino acid identity).