PROTEINS: Structure, Function, and Genetics 32:475–494 (1998)

Assembly of Protein Structure From Sparse Experimental Data: An Efficient Monte Carlo Model

Andrzej Kolinski*[1,2] and Jeffrey Skolnick[1]
[1]*Department of Molecular Biology, The Scripps Research Institute, La Jolla, California*
[2]*Department of Chemistry, University of Warsaw, Warsaw, Poland*

ABSTRACT **A new, efficient method for the assembly of protein tertiary structure from known, loosely encoded secondary structure restraints and sparse information about exact side chain contacts is proposed and evaluated. The method is based on a new, very simple method for the reduced modeling of protein structure and dynamics, where the protein is described as a lattice chain connecting side chain centers of mass rather than Cαs. The model has implicit built-in multibody correlations that simulate short- and long-range packing preferences, hydrogen bonding cooperativity and a mean force potential describing hydrophobic interactions. Due to the simplicity of the protein representation and definition of the model force field, the Monte Carlo algorithm is at least an order of magnitude faster than previously published Monte Carlo algorithms for structure assembly. In contrast to existing algorithms, the new method requires a smaller number of tertiary restraints for successful fold assembly; on average, one for every seven residues as compared to one for every four residues. For example, for smaller proteins such as the B domain of protein G, the resulting structures have a coordinate root mean square deviation (cRMSD), which is about 3 Å from the experimental structure; for myoglobin, structures whose backbone cRMSD is 4.3 Å are produced, and for a 247-residue TIM barrel, the cRMSD of the resulting folds is about 6 Å. As would be expected, increasing the number of tertiary restraints improves the accuracy of the assembled structures. The reliability and robustness of the new method should enable its routine application in model building protocols based on various (very sparse) experimentally derived structural restraints. Proteins 32:475–494, 1998.**

Key words: protein assembly; protein structure; protein reduced models; lattice models; Monte Carlo simulations; fold prediction

INTRODUCTION

The ability to predict the three-dimensional structure of a protein from its amino acid sequence is of great theoretical[1,2] and practical importance.[3] In practice, structure prediction could be attempted on various levels, ranging from purely de novo approaches to those that incorporate restraints derived from experimental data. The latter aspect of protein structure modeling has recently attracted significant attention[4–6] due to its possible application to model building based on structural restraints provided by nuclear magnetic resonance (NMR)[7] or other experimental methods. This paper describes a new, relatively rapid and straightforward algorithm for structure assembly that employs partial knowledge of known protein secondary structure and a small number of tertiary restraints. "Partial knowledge" means that the method does not require a detailed description of the local secondary structure in terms of its ϕ, ψ angles or their lattice equivalent. Instead, a three-letter code for secondary structure (H-helix, E-extended, and (−) everything else) is used as an input. This is translated in the program into loosely defined preferred ranges of local intrachain distances. In many respects, the present approach is very similar to our recently published MONSSTER (**MO**deling of **N**ew **S**tructures from **S**econdary and **T**ertiary **R**estraints) algorithm.[6] MONSSTER provides a well-defined protocol for the identification of moderate-resolution nativelike structures from known secondary structure and a small number of tertiary restraints.

What is the range of application of such models? Certainly, when a large number of distance restraints obtained from NMR[8] and possibly from homology-based theoretical considerations is available, more standard algorithms[9–12] are the tools of choice. These algorithms are based on purely geometrical considerations, followed by restrained molecular dynamics refinement of the model structures.[13] However, in many real life situations, the

Grant sponsor: National Institutes of Health; Grant number: GM-37408.
*Correspondence to: A. Kolinski, Department of Molecular Biology, TPC5, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037.
Received 17 July 1997; Accepted 13 April 1998

number of available restraints is relatively small and limited in the early stages of NMR-based structure determination. When the available restraints are too sparse to define even a moderate resolution structure, it is necessary to have a model that uses a reasonable force field capable of providing an overall protein-like bias. In such a case, even a small number of distance restraints could be sufficient to guide folding to the correct structure. Due to the necessity of sampling a substantial part of the protein conformational space, such an algorithm has to be computationally effective. Moreover, the force field of the model should be able to correct for some errors in the provided set of restraints. The MONSSTER method satisfied all these requirements. However, the computational cost of this method limited its application to proteins containing no more than 150 residues. Thus, our goal here is to remove this limitation so that a larger class of proteins can be treated using a relatively smaller number of long-range restraints.

In the past several years, there also have been a number of other studies that incorporate known, correct secondary structure and a limited number of known, correct tertiary restraints to predict the global fold of a globular protein. In particular, Smith-Brown et al.[4] have modeled a protein as a chain of glycine residues. Tertiary restraints are encoded via a biharmonic potential, with folding forced to proceed sequentially via successive implementation of these restraints. In practice, they find that a substantial number of tertiary restraints are required to assemble a protein structure. Another effort to predict the global fold of a protein from a limited number of distance restraints is due to Aszodi et al.[5] Their approach is based on distance geometry, where a set of experimental tertiary distance restraints is supplemented by a set of predicted interresidue distances. These distances are obtained from patterns of conserved hydrophobic amino acids that have been extracted from multiple sequence alignments. In general, they find that to assemble structures below 5 Å cRMSD, on average, more than N/4 restraints are required, where N is the number of residues. Even then, the Aszodi et al. method has problems selecting out the correct fold from competing alternatives. However, an advantage of the Aszodi et al. approach is that it is very rapid, with a calculation taking on the order of minutes on a typical contemporary workstation.

In this paper, we describe a new lattice protein model, the **SI**de **CH**ain **O**nly (**SICHO**) model, that focuses explicitly on the side chain center of mass positions and implicitly treats the protein backbone. As in MONSSTER, the present force field consists of short-range interactions that reflect secondary propensities and some short-range packing biases, a geometrically implicit model of cooperative hydrogen bonds, and explicit burial, pair and multibody tertiary interactions. The increased robustness of the present method is due to a more efficient protein representation and a new definition of the model force field, which when combined with a small number of long-range harmonic restraints (known side chain contacts), result in rapid collapse and assembly. Due to the way the model and force field are formulated, in comparison to MONSSTER, the computational cost scales with a lower power of the chain length.

The outline of the remainder of this paper is as follows. First, we describe the new method. Included is a detailed discussion of the geometric properties of the model, the force field and the conformational sampling protocol. Next we describe results on the folding of 8 representative proteins having a number of common protein motifs. We then compare our results with that of previous work.[4–6] Finally, in the Conclusions section, we summarize our results and outline the direction of future research.

METHODS

Protein Representation

The protein is modeled as a lattice chain connecting points restricted to an underlying simple cubic lattice whose mesh size equals 1.45 Å. By way of illustration, Figure 1 depicts short fragments of a β-strand and an α-helix in this particular lattice representation. This figure also shows the corresponding Cα-traces, which are not explicitly modeled. The distance between two consecutive side chain units is variable and is assumed to be in the range of $11^{1/2}$–$30^{1/2}$ lattice units or equivalently 4.8–7.9 Å. The length distribution roughly covers typical distances between two consecutive side chain centers of mass seen in real proteins.[14] The resulting number of side chain side chain vectors, |v|, is equal to 592. Similar limitations are superimposed on the distances between the i-th and i + 2^{nd} α-carbons, i-th and i + 3th α-carbons, etc. As a result, some implicit limitations are superimposed onto the range of planar angles defined by the positions of three consecutive side chains. Some possible three-vector local conformations are shown in Figure 2A.

As shown in Figure 2B, the excluded volume cluster defined for each side chain consists of the central lattice point (coinciding with the hypothetical center of mass of the side chain) and the 16 surrounding points located at positions (±1,0,0) and (±1, ±1,0), including all permutations of these vectors. With such a hard-core definition, the distance of closest approach of two residues is equal to three lattice units (4.35 Å). This nicely corresponds to the equivalent hard core in real proteins. There are also 30 possible lattice positions at which the closest approach, side chain-side chain contact, can occur. These are defined by six vectors of the (3,0,0) type and 24 vectors of the (2,2,1) type emanating from the side chain of interest. For bigger residues, a wider, finite magnitude, repulsive core is also included, and the number of "contact positions" is even larger.

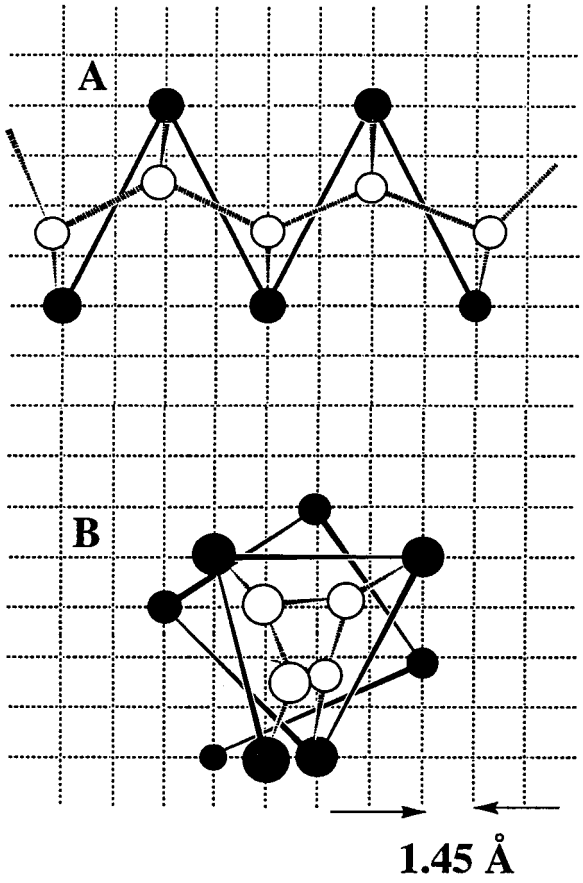

Fig.1. Illustration of the protein chain representation. (A) For a short expanded fragment and (B) for a helical fragment. The solid circles correspond to explicitly simulated side chain centers of mass. The open circles indicate the expected positions of the α-carbons.

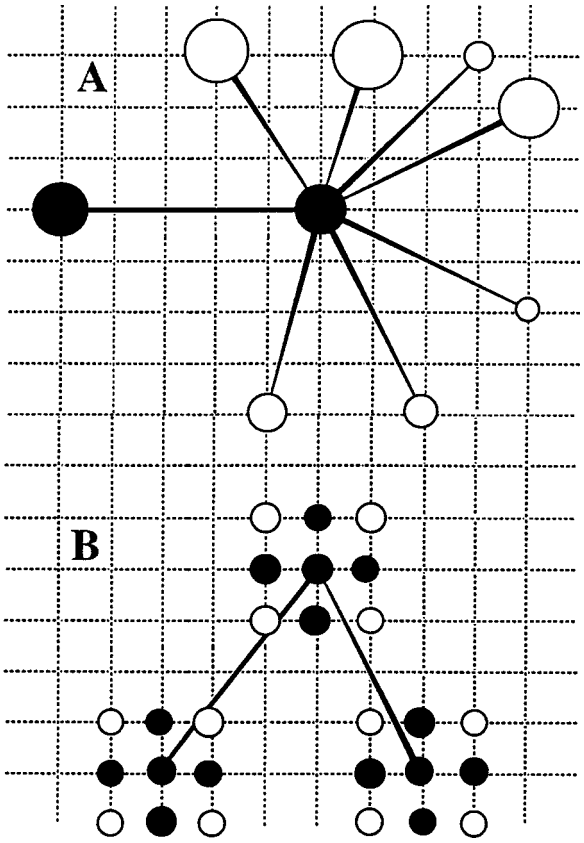

Fig. 2. Some examples of bonds connecting three successive side-chain united atoms. (A) The open circles in the upper panel correspond to a subset of possible positions of a third side chain given that the positions of the two preceding units (solid circles) are fixed and (B) illustration of excluded volume clusters. The solid dots correspond to the three lattice points along the axis orthogonal to the displayed slice. The open circles correspond to a single point in the plane.

Consequently, effects of lattice anisotropy are essentially nonexistent.

Side chain overlaps and interactions are readily detected by inspection of the occupancy status of the appropriate collection of lattice points in the Monte Carlo working box. As a result, for a given residue, the computational cost for calculating the short- and long-range interactions does not depend on chain length.

Monte Carlo Model and Conformational Updating

The Monte Carlo move set consists of single residue "kink" moves, chain-end moves, two-residue moves and small "rigid-body" displacements of a larger portion of the model chain. Examples of these moves are schematically illustrated in Figure 3A–D. A single "time step" consists of N attempts at kink moves, 2 attempts at chain-end moves, N-1 attempts at two-bond moves and one attempt at a randomly selected, large fragment displacement. Before any energy computation, the test for excluded volume violation is always performed, and trial conformations that would lead to steric collisions of chain units are rejected. Also, conformations that would result in nonphysical distances between two consecutive side chain units are a priori rejected.

Interaction Scheme

The interaction scheme consists of short-range interactions, hydrogen bond interactions, and long-range interactions. All types of interactions have generic (i.e., sequence-independent), sequence-dependent, and target (i.e., resulting from superimposed short- and long-range restraints) components. Below, we first discuss the generic and sequence dependent terms, and then describe those terms arising from the restraint contributions.

Sequence dependent short-range interactions

The potentials were derived from the geometric statistics of known protein structures. We considered pairwise-specific distances between nearest neighbors, up to the fourth neighbor, along the polypeptide chain. These distances depend on amino

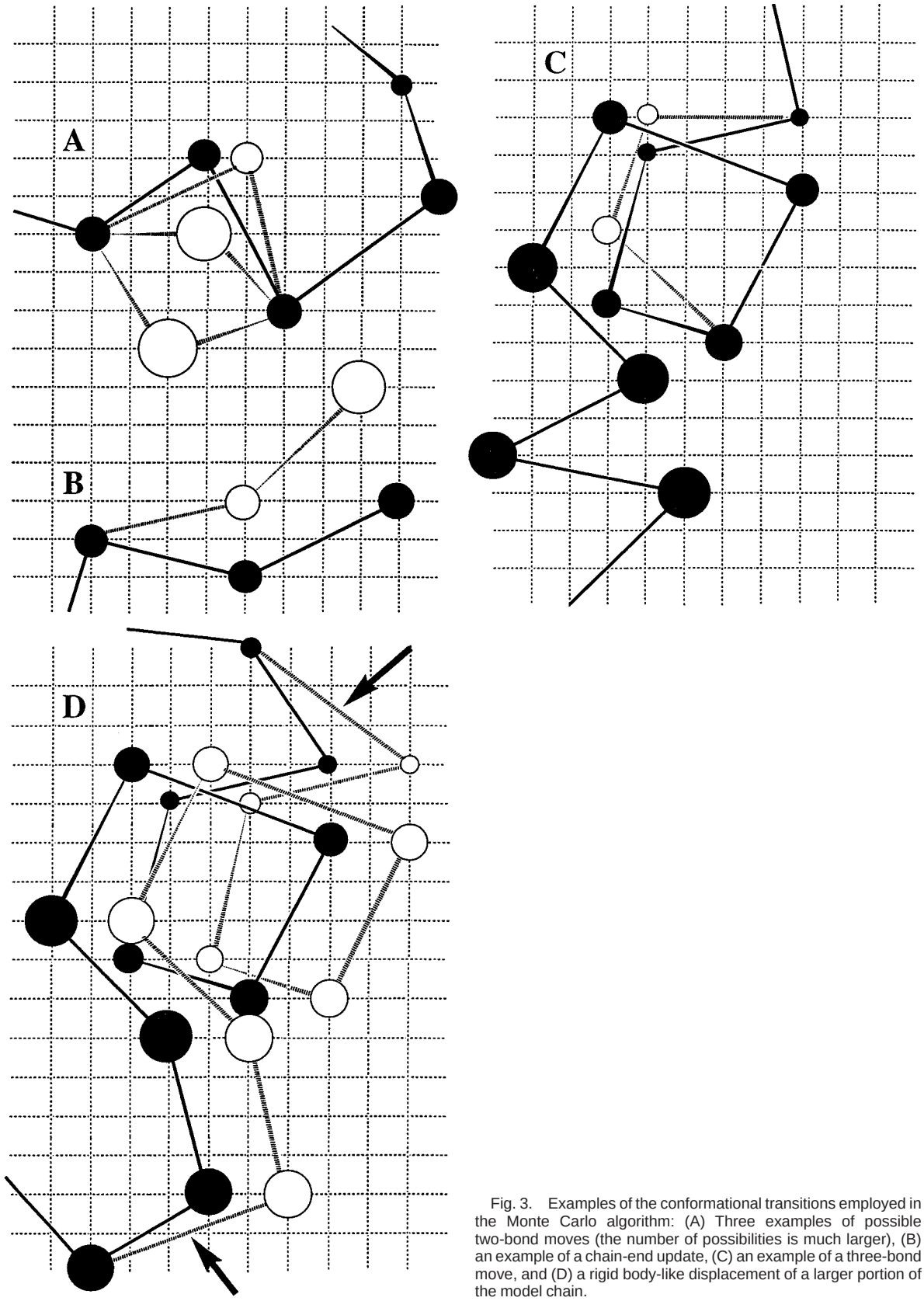

Fig. 3. Examples of the conformational transitions employed in the Monte Carlo algorithm: (A) Three examples of possible two-bond moves (the number of possibilities is much larger), (B) an example of a chain-end update, (C) an example of a three-bond move, and (D) a rigid body-like displacement of a larger portion of the model chain.

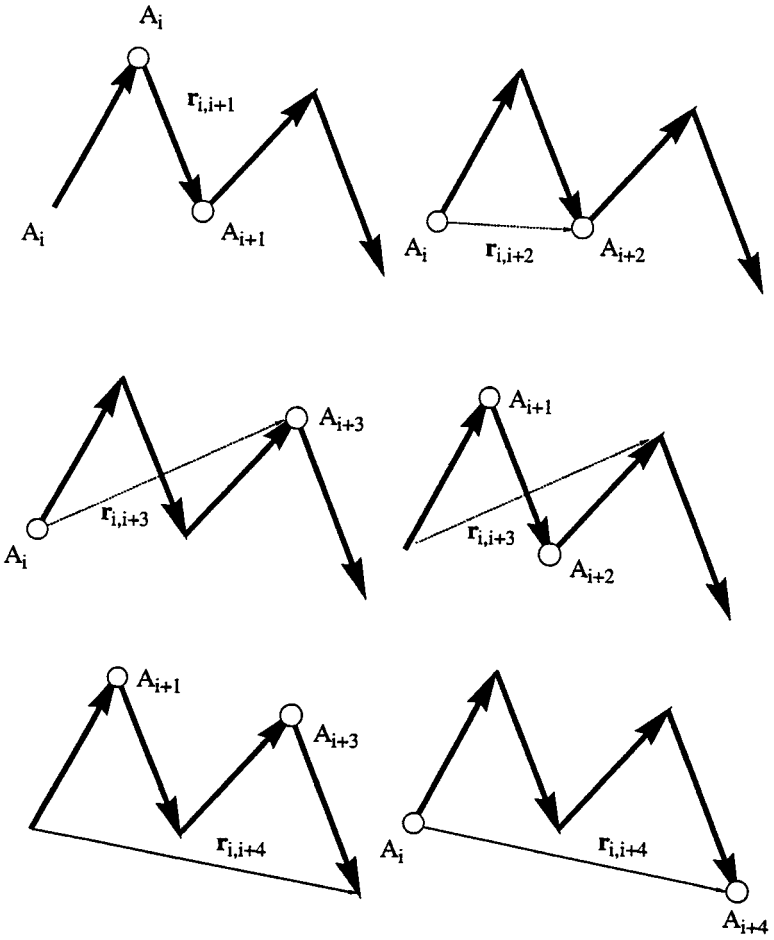

Fig. 4. Explanation of geometry used for the definition of the six terms describing the sequence-specific short-range interactions.

acid composition and the local chain geometry. Six bins, covering the majority of distances observed in proteins, have been used for all components of the short-range interactions. For a given pair of amino acids in a relevant position, the distribution of associated distances between side chain centers of mass is extracted from the statistical analysis of the structural database of nonhomologous proteins. When compared to an average distribution (ignoring the sequence information), this leads to a statistical potential. The technique is similar to that employed in other studies.[15] A detailed description of the derivation of strictly related potentials can be found elsewhere.[16] As schematically illustrated in Figure 4, the resulting potential could be expressed as follows:

$$\begin{aligned} E_{short} = &\sum E_{12}(r^2_{i,i+1}, A_i, A_{i+1}) \\ &+ \sum E_{13}(r^2_{i,i+2}, A_i, A_{i+2}) \\ &+ \sum E_{14}(r^{2*}_{i,i+3}, A_{i+1}, A_{i+2}) \\ &+ \sum E'_{14}(r^{2*}_{i,i+3}, A_i, A_{i+3}) \\ &+ \sum E_{15}(r^2_{i,i+4}, A_{i+2}, A_{i+3}) \\ &+ \sum E'_{15}(r^2_{i,i+4}, A_i, A_{i+4}). \end{aligned} \quad (1)$$

The summation is performed along the chain; E_{1d} refers to energy associated with interactions between the residue of interest and its d-1st neighbor down the chain. A_i denotes the amino acid identity at position i, and $r_{i,i+k}$ is the distance between residues i and i + k. The terms for the three-bond fragments include the effects of local chain chirality via a "chiral"-distance-squared term

$$r^{2*}_{i-1,i+2} = r^2_{i-1,i+2}\,\mathrm{sign}((\mathbf{v}_{i-1} \otimes \mathbf{v}_i) \cdot \mathbf{v}_{i+1}). \quad (2)$$

All terms are amino acid pair specific because the presently available structural database does not support meaningful statistics for higher order terms. Thus, there is a single energy term for one-bond and two-bond fragments, and two types of binary potentials for three-bond and four-bond fragments. These sequence dependent short-range interactions also provide some information about short-range packing regularities, e.g., the propensities for a particular side chain arrangement on a helical surface. It is not obvious what the optimal relative scaling of these interactions should be. For simplicity, we take the relative scaling of all terms equal to one. This scaling generates a reasonable level and identity of second-

ary structure. Since there are a large number of numerical values for these short-range potentials (six components, each having $20 \times 20 \times 6$ pairwise values for 6-bin histograms), the data are available via anonymous ftp[17], or upon request from the authors.

Generic short-range conformational biases

Next, terms that do not depend on amino acid sequence have been introduced into the model force field. Thus, the energy contribution from these terms depends only on specific chain geometry (regardless of protein sequence) and its magnitude is controlled by a single adjustable energetical parameter ϵ_{gen}. Their purpose is to enforce a proteinlike distribution of short-range conformations.

The first of these components accounts for the characteristic stiffness of polypeptide chains, which builds on the observation that there is a characteristic orientation pattern in folded proteins. The local orientation of protein chain could be conveniently defined by a vector orthogonal to a triangle formed by three consecutive centers of the side chains. The corresponding conformational bias could be defined as follows:

$$E_{stiff} = -0.25\, \epsilon_{gen} \Sigma\, (\mathbf{w}_i \cdot \mathbf{w}_{i+4}) \qquad (3)$$

where $\mathbf{w}_i$ is a vector orthogonal to the plane formed by the two consecutive virtual chain bonds $\mathbf{v}_{i-1}$ and $\mathbf{v}_i$, ϵ_{gen} is an arbitrarily chosen energetical parameter, equal to 1 k_BT in all potentials described in this section, here scaled by a factor equal to -0.25. The length of the orthogonal vectors $\mathbf{w}_i$ is about 4 lattice units, and they are also used for detection of "hydrogen bonds." The dot product in the above equation is near its maximum value for expanded, β-like states and for helices. The high value of this product is significant in a majority of typical turns and loop-type local conformations. Thus, the potential provides a bias towards these relatively rigid elements of protein secondary structure.

The second generic term provides a bias towards regular arrangements of secondary structure. In a random lattice chain, the distribution of distances between the i-th and $i + 4^{th}$ bead would be unimodal and close to a Gaussian distribution. On the other hand, the corresponding distance distribution between residues in native proteins is bimodal. The shorter distance peak corresponds to helical and turn conformations, while the more diffuse, longer distance peak corresponds to expanded conformations. A term that adjusts the model to this bimodal distribution could be expressed as follows, with all distances in lattice units:

$$E_{struct} = \Sigma\, Es(i) \qquad (4a)$$

with:

$$Es(i) = -2\epsilon_{gen}, \quad \text{for } r^2_{i,i+4} < 33 \text{ and } (\mathbf{v}_i \cdot \mathbf{v}_{i+3}) > 0 \qquad (4b)$$

or

$$Es(i) = -2\epsilon_{gen}, \quad \text{for } 48 < r^2_{i,i+4} < 145 \text{ and } (\mathbf{v}_{i+1} \cdot \mathbf{v}_{i+2}) < 0. \qquad (4c)$$

The first set of conditions describes a loosely defined, helical conformation, while the second describes an expanded, β-type fragment. Thus Equation 4b states that the distance between the i-th and $i + 4^{th}$ side chain in a helix has to be small (here below ca. 8 Å). The second condition states that the chain has to make a tight turn. A corresponding set of conditions is defined for β-type expanded states. In both cases, the cut-off distances and the angular restrictions are selected in a very permissive way based on the observed distributions for native proteins. The permissive definition of local conformational biases drives the model system towards a loosely defined proteinlike chain geometry, yet it still allows substantial local mobility. As mentioned before, in all simulations, the value of ϵ_{gen} has been assumed to be equal to 1 k_BT.

"Hydrogen bonds" and generic packing biases

The model hydrogen bonds provide similar structure-regularizing biases with respect to tertiary interactions as the generic short-range interactions do for secondary structural regularities. Residue i is considered to be hydrogen bonded to residue j when the orthogonal vector $\mathbf{w}_i$ (originating from the bead i) touches any of the 17 points of the excluded volume cluster of residue j. Two (and only two, due to the above definition) hydrogen bonds can originate from a given residue. The geometry of hydrogen bonds is depicted in Figure 5. Only residues that are "in contact" could be hydrogen bonded. That is, there is the same long-range cut-off for side group pair interactions and for hydrogen bonding. The energy of the hydrogen bond network is defined as follows:

$$E_{H\text{-}bond} = -\epsilon_{H\text{-}bond} \Sigma\, (\delta^+ + \delta^- + \delta^{+,-}) \qquad (5)$$

where δ^+ , δ^- , $\delta^{+,-}$ are equal to 1 when the "right handed," the "left handed," and both hydrogen bonds originating from residue i are satisfied, respectively. Otherwise, the corresponding terms are equal to zero. The last term, $\delta^{+,-}$, is a cooperative hydrogen bond energy gained only upon local saturation. The numerical value of this parameter was assumed to be equal to 1.0–1.25 k_BT. The lower value of this parameter tends to accelerate folding, while the higher value tends to build structures of slightly

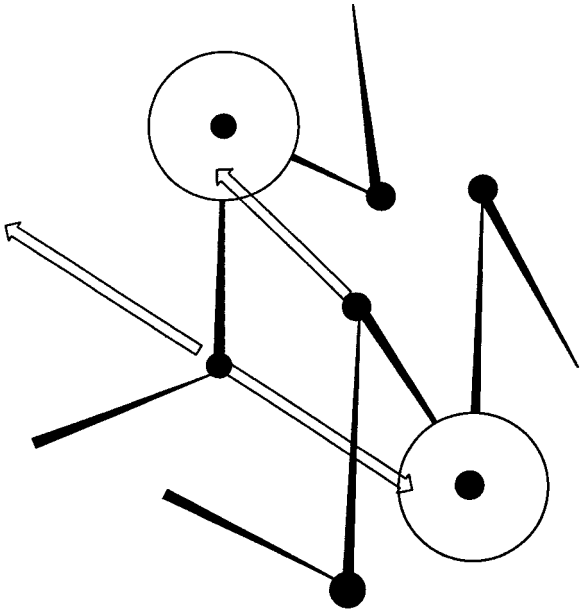

Fig. 5. Illustration of model hydrogen bond geometry. The hydrogen bonds are shown by open arrows.

better quality. These effects are small. In all isothermal Monte Carlo runs used for energy comparisons, the same value (1.0) has been employed.

Two other generic terms that enforce proteinlike packing regularities also have been introduced. The first one is a "contact map propagator" that reflects the most common patterns seen in all side chain contact maps of globular proteins.[18] It is defined in the following way:

$$E_{map} = -\epsilon_{gen}(\Sigma\Sigma(\delta_{i,j} \cdot \delta_{i+1,j+1} \cdot \delta_{i-1,j-1})\delta_{par} + \Sigma\Sigma(\delta_{i,j} \cdot \delta_{i-1,j+1} \cdot \delta_{i+1,j-1})\delta_{apar}) \quad (6)$$

where $\delta_{i,j}$ is equal to 1 (0) when residues i and j are (not) in contact. δ_{par} is equal to 1 only when the corresponding chain fragments are oriented in a parallel fashion, i.e., $(\mathbf{v}_{i-1} + \mathbf{v}_i) \cdot (\mathbf{v}_{j-1} + \mathbf{v}_j)$ 0. Similarly, δ_{apar} is equal to 1 when the chain fragments are anti-parallel. In the above equation (and in Eq. 7), $\epsilon_{gen} = 1$ is exactly the same parameter as the one used in the short-range generic terms.

A second packing regularizing term provides an additional cohesive energy between secondary structure elements by favoring the parallel packing of pairs of hydrophilic residues and the anti-parallel packing of pairs of hydrophobic residues. Consequently, since it exploits sequence information, this term is not purely generic; however, it is reduced to a two-letter (HP) code.

$$E_{packing} = -\epsilon_{gen}\,\Sigma\Sigma\,(\delta_{PP} \cdot \delta_{pp} + \delta_{HH} \cdot \delta_{app}) \quad (7)$$

where δ_{PP} (δ_{HH}) is equal to 1 when both residues in contact are hydrophilic, P, (hydrophobic, H), according to the Kyte-Doolittle hydrophobicity scale.[19] The value of δ_{pp} is equal to 1 only when the packing of the side chain pair is parallel; i.e., $(\mathbf{v}_{i-1} - \mathbf{v}_i) \cdot (\mathbf{v}_{j-1} - \mathbf{v}_j)$ 0. Similarly, δ_{app} is equal to 1 only when the packing of the side chain pair is antiparallel; i.e., $(\mathbf{v}_{i-1} - \mathbf{v}_i) \cdot (\mathbf{v}_{j-1} - \mathbf{v}_j)$ 0.

Various structure regularizing terms described in this and the previous section reflect the various structural regularities seen in globular proteins. Each one accounts for a different correlation that could be easily detected by statistical analysis of the geometry of the side-chain-only representation of protein structures. Except for the last one (which depends on some sequence features), they are sequence independent: the underlying regularities are true for all types of structural motifs of globular proteins. During the Monte Carlo simulations, these generic potentials provide a very strong bias against nonsensical, non proteinlike conformations. Such conformations would otherwise be quite frequent due to the reduced character of the protein representation. In the presence of these generic contributions to the model force field, the requirements for sequence-specific potentials are lower; they have to select between various proteinlike conformations, which makes the selection easier (and computationally less expensive) than in the much broader conformational space of an unrestricted model chain.

Sequence-specific long-range interactions

These interactions are defined as follows:

$$E_{pair} = \Sigma\Sigma E_{ij} \quad (8a)$$

where:

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

where ϵ_{ij} are the pairwise interaction parameters described in our previous work,[6,20] and the interactions are counted for all pairs, except the first nearest neighbors along the chain. There is a strong, soft-core repulsive energy, $E^{rep} = 4.0\ k_BT$, in all simulations. It provides a slightly larger excluded volume for larger amino acids than that defined by the hard core. The values of the cut-off distances $R_{i,j}^{rep}$ and $R_{i,j}$ are given in Table I. The values of $R_{i,j}$ were adjusted to approximately mimic the contact distances employed in the derivation of binary interactions parameters.[20] (Here, we employ the "native" interaction scale from Skolnick et al.[20].)

TABLE I. Compilation of Pairwise Cut-off Distances in Angstroms

A_i	A_j	$R_{i,j}^{rep}$	$R_{i,j}$ (attractive)[a]	$R_{i,j}$ (repulsive)
Small[b]	Small	4.35	7.03	6.32
Small	Large	4.57	7.03	6.32
Large	Large	4.83	7.50	7.03

[a]Attractive pair of amino acids.
[b]Small amino acids are: Gly, Ala, Ser, Cys, Val, Thr, Pro.

One-body burial interactions

To facilitate a rapid collapse of the model chain, we employ a centrosymmetric, density regularizing term based on a statistical analysis of single domain proteins. This is the only term that uses the assumption that the protein has a single domain. For some increase in computational cost, these terms could be omitted. The radius of gyration of the protein is given by:

$$S = (N^{-1}\Sigma(r_{CM} - r_i)^2)^{1/2} \quad (9)$$

where r_{CM} is the position of the center of mass of the globule, and r_i is the position of the center of mass of the i-th side chain. The size of a single domain protein is strongly correlated with the number of residues N according to:

$$S = 1.52\ N^{0.38} \quad \text{in lattice units.} \quad (10)$$

The exponent 0.38, obtained from the statistical analysis of single domain globular proteins,[21] is very close to the value of 1/3 expected for a long, collapsed polymer chain.[22] The corresponding potential has the following form[23]:

$$E_b = \epsilon_b \Sigma |m_{o,i} - m_i| \quad (11)$$

where $m_{o,i}$ is the target number of amino acids in a given spherical shell centered at the protein's center of mass. There are three equal thickness shells within a distance S, and they contain somewhat more than half of the protein residues. The entire protein is essentially contained in a sphere of radius equal to 5/3 S. The value of the parameter ϵ_b was equal to 0.25–1.0 k_BT, depending on protein size. Larger proteins tend to exhibit a larger absolute deviation from the above target distribution of mass, and consequently, a lower penalty for such deviations should be employed.

To further enhance rapid collapse, those residues that are within a radius of 2/3 S (a very conservative estimate of the hydrophobic core of a single domain globular protein) contribute $\epsilon_{KD}(i)/16$ to the total energy, where $\epsilon_{KD}(i)$ is the Kyte-Doolittle hydrophobicity parameter of the i-th residue.[19,24] The scaling factor 1/16 is arbitrary chosen. This potential (and its scaling with respect to other interactions) has very little effect on the folded structure, but it improves folding kinetics.

Multibody surface exposure term

Amino acid side groups have a different size and shape. Thus, when a given side chain is in contact with another amino acid, the fraction of its surface that is covered depends on the identity of the contacting partner. Appropriate parameters reflecting this observation (i.e., the surface coverage of particular types of side chains and associated statistical-type potential) could be derived from the statistics of known protein structures. In the present algorithm, each residue has 30 surface contact points. A subset of these contact points becomes occupied upon contact with other side chains or main chain Cα atoms. The Cα atom positions are approximated from the positions of three consecutive side chain beads and have their own excluded volume and contribution to surface coverage. Due to shadowing, some contact points could be multiply occupied by different residues (usually 1 or 2, or sometimes 3, but very rarely 4). The fraction of occupied surface points defines the fraction of buried area of a given side chain. The total energy of a model protein is computed as:

$$E_{surface} = \epsilon_s \Sigma\ E_b(A_i, a_i) \quad (12)$$

where a_i is the covered fraction of sites of amino acid side chain A_i and $E_b(A_i, a_i)$ is the statistical potential for amino acids A_i that are covered by a_i contact points, i.e., its coverage fraction is $a_i/30$. The reference state for this statistical potential is "an average" amino acid with average (over structural database) coverage. The scaling factor ϵ_s for this term has been assumed to be 0.25.

The above approach to the hydrophobic interactions allows suppression of our previously employed centrosymmetric one-body potentials[6] and thereby opens up the possibility of extending the present approach to multidomain and multimeric proteins. Here, we used both models of mean field hydrophobic interactions in parallel.

The force field designed for this model is entirely of a "knowledge-based" origin. Some terms, such as the generic short- and long-range potentials, provide a bias toward proteinlike short- and long-range correlations in the model chain. These potentials generalize regularities seen in native structures of all globular proteins. The sequence-specific terms were derived as statistical potentials with a rather careful selection of the reference state.[20,25,26] When several statistical potentials are combined in a relatively complex reduced model, an a priori derivation of the relative scaling factors is virtually impossible. Some double counting of particular physical interactions may occur. Thus, these scaling factors have to be adjusted to reproduce a reasonable balance between

the short- and long-range interactions. A proper balance should lead to a low secondary structure content in the denatured state and a well-packed and ordered collapsed state. The collapse transition should be as abrupt as possible, mimicking an all-or-none folding transition. This has been achieved in the present model with the given scaling of particular interactions. Folding experiments for several proteins of various structural classes were performed with no short- or long-range restraints. The force field described above fails to produce a unique folded state, except for very simple folding motifs. For more complex motifs, the folded states always had a secondary structure very close to the native, with good packing of the hydrophobic core; however, the arrangement of the secondary structure elements (connection of helices, order of β-strands in sheets, etc.) almost always had topological errors. As designed, the model with its force field is very efficient at generating proteinlike compact conformations. The model is not sensitive to the particular scaling of the various interactions within a broad range around the set used in this work. For example, removal of all generic terms also led to collapsed structures (although at lower temperatures) with good overall fidelity of the secondary structure, but the geometrical accuracy of the secondary structure and packing pattern was more irregular. A detailed discussion of the interplay between the generic and sequence-specific short-range potentials will be published elsewhere.[27] It could be expected that when the proposed force field is supplemented by some structural restraints, a proper fold should be easily selected.

Since a similar Cα-based model was successful in reproducing quite complex aspects of protein dynamics and thermodynamics,[6,15,16,23,28–36] we believe that the present realization of the force field approximately reproduces the main features of globular proteins. However, here, due to a different geometrical context, most of the short-range interactions had to be rederived. The present model is somewhat simpler (simpler representation and simpler definition of the force field) and computationally more efficient than our Cα-based model. Thus, larger proteins could be simulated.

Physical Basis of the Model Interaction Scheme

The proposed interaction scheme may appear rather complex and requires some explanation. The main difficulty is related to the fact that we are attempting to model quite realistic protein structures (as seen on the level of an entire fold) with an extremely simple representation of the protein conformational space. Only the side chains are explicitly modeled. The assumption of a single interaction unit per residue is computationally very efficient. Why were side chains used rather than (for example) alpha carbons? The choice is dictated by the fact that specific interactions in proteins involve side chains, while the main chain interactions are much less dependent on amino acid sequence. Due to this very simple representation and requested specificity of the model, several features have to be built into the model force field. First, the assumed protein representation, with a single center of interaction per protein side chain, allows too much conformational freedom. This is because there is no explicit backbone connectivity in the model chains. However, in real proteins, the backbone connectivity and conformational stiffness control, to some extent, the distances between the centers of mass of the side groups near each other along the polypeptide chain. The backbone effect is moderated by the side chains' internal degrees of freedom. It is reasonable to assume that for a short polypeptide fragment, the local geometry of the side chain centers of mass is mostly dictated by short-range interactions with a somewhat lesser effect from long-range (tertiary) interactions. The correct, proteinlike distance geometry of the side chain centers of mass would imply a correct, proteinlike geometry of the main chain. This provides a conceptual background for the sequence-specific short-range potential of mean force (discussed previously and defined in Eq. 1). This potential drives the system towards a local geometry (characterized by distances between side chains) that is characteristic of locally similar sequences.

At first glance, it may appear that such defined sequence-specific secondary propensities are sufficient for modeling proteinlike local geometry. This is not the case for a couple of reasons. First, the discussed statistical potentials are not very accurate due to the limited size of the database of known protein structures. However, more importantly, the assumed simplified representation of the polypeptide chains exhibits excessive flexibility. With respect to the assumed model of excluded volume, a substantial fraction of the model chain conformations that are otherwise allowed are conformations that cannot possibly occur in any protein or even in other polymers. It is not a good strategy to make the sequence-specific interaction so strong that such nonphysical geometries would be practically prohibited. This would lead to dynamic frustration of the model system due to very frequent trapping in the local conformational energy minima; thus, providing a generic bias towards proteinlike geometry is computationally more efficient. Then, much less is required from the sequence-specific part of the potential (selection within the proteinlike part of conformational space instead of selection within the much larger conformational space of a freely joined polymer chain). Moreover, a properly defined generic potential can "interpolate" proteinlike conformations for those fragments of a given polypeptide chain where the information content of the sequence-

specific potential is low (due to lack of examples in the database or balanced contradictory examples). As discussed above (see Eqs. 3 and 4), sequence-independent potentials exactly play such a role. The first such term provides a bias towards the protein-like stiffness of the model chain by an energetic preference for either expanded zigzag or helical conformations. The second term provides a bias towards a bimodal distribution of the distances between the i-th and $i + 4^{th}$ side chain units. Our definition of these potentials mimics some of the most general structural regularities seen in all folded proteins. They also provide a bias against nonphysical local conformations in the unfolded state.

Similar to the short-range interactions, there are sequence-specific and generic terms of the model tertiary interaction scheme. The pairwise contact potentials and the model of hydrophobic burial potentials of mean force derived from the statistics of the structural database do not require additional discussion. The procedures of derivation and implementation of such potentials are rather standard and commonly used in all reduced models of proteins.[15–17,21,25,26] Such statistical potentials encode some interaction preferences in real proteins. In the majority of cases, they are accurate enough to select a proper fold for a given sequence from a collection of other folds of natural proteins. However, in our case, the requirements are more stringent. The proper fold must be selected from a much larger number of conformations, most of them never observed in real proteins (but possible in the model due to the reduced representation). Thus, it is important to construct a generic potential that provides a bias toward proteinlike tertiary interaction patterns. Such patterns could be postulated as a generalization of structural regularities seen in known protein structures. An important feature of all protein structures is their very regular network of main chain hydrogen bonds. Our model lacks an explicit protein backbone. Nevertheless, an analysis of protein structures shows that the presence of hydrogen bonds between residues translates with high reproducibility into a pattern of contacting side chains. Indeed, the string of hydrogen bonds along a helix implies the existence of continuous (or almost continuous) strings of side group contacts along the helix surface. Similarly, a string of hydrogen bonds in a β-hairpin implies two strings of side group contacts, one on each side of the hairpin. Thus, a bias towards such a string (see Eq. 5 and discussion of the potential) could be used as an ersatz copy of the hydrogen bond interactions. Furthermore, such strings of contacts lead to a characteristic pattern of side chain contacts. The generic potential given in Eq. 6 provides a bias towards the most general feature of such patterns.[16,18]

Angular packing preferences for various types (hydrophobic or hydrophilic) of residues also could be used as a bias toward proteinlike side chain packing patterns (see Eq. 7 and the description of this term).

Such a defined model of the force field could be tested and the relative weights of the sequence-specific versus generic terms could be adjusted by a trial and error method. In our studies, we performed a long series of isothermal simulations of various proteins. While the nativelike structures were sometimes obtained only for very simple, small proteins, the accuracy (cRMSD from native) of the emerging elements of secondary and supersecondary structure elements (helices, helical hairpins, β-hairpins or α–β–α motifs) could be used as a convenient criterion.

The presented force field is not accurate enough for reproducible folding simulations of the majority of (even small) proteins. At the same time, it discriminates against a vast majority of nonsensical conformations and the nativelike structures always belong to a relatively small number of low energy conformations. Thus, when some long-range restraints of experimental origin are superimposed on top of the discussed force field, the nativelike conformation could be easily obtained in Monte Carlo simulations, as shown in the remainder of this paper.

Implementation of the Restraints

Encoding short-range conformational propensities

In the test of structure assembly described in this work, we assume knowledge of secondary structure[37] in the form of a three-letter code: E-extended, H-helix and (-) everything else. Such a three-letter code is then translated onto a set of biases towards a corresponding range of local intrachain distances and angular correlations. Only E and H states have some conformational biases, and their definitions are geometrically very permissive. The set of secondary structural restraints are as follows:

1. An H-state cannot be hydrogen bonded to an E-state. (When detected, such bonds are ignored and do not contribute to the conformational energy.)

2. A residue in a continuous stretch of H-states can hydrogen bond only to residues $i - 3$ and $i + 3$. (When hydrogen bonds are associated with Cα's or side chains, this is the canonical helix pattern.)

3. The system gains an additional energy equal to $-\epsilon_{gen}$ (over the previously defined generic contributions, and ϵ_{gen} is of the same exact value as that used in the definition of various generic terms of the model force field) in all the following cases (a-c): As shown in Figure 6, for helical type states when:

$$\text{for} \qquad r^2_{i,i+4} < 33 \tag{13a}$$

a) residues $i + 1$ and $i + 2$ are assigned as helical if $(\mathbf{v}_i \cdot \mathbf{v}_{i+2}) < 0$

b) residues $i + 2$ and $i + 3$ are assigned as helical if $(\mathbf{v}_{i+1} \cdot \mathbf{v}_{i+3}) < 0$

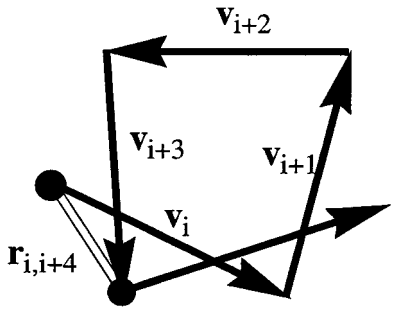

Fig. 6. Geometry employed in the definition of the helical bias (see text).

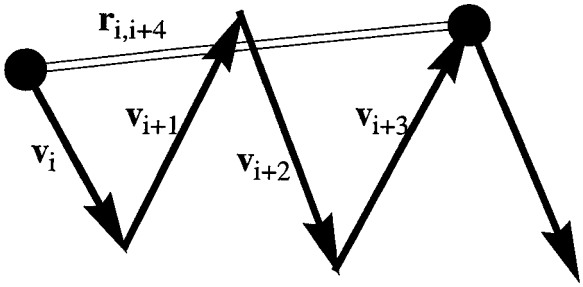

Fig. 7. Geometry employed in the definition of the extended, β-state bias (see text).

c) residues i + 1, i + 2 and i + 3 are assigned as helical if $(\mathbf{v}_i \cdot \mathbf{v}_{i+3}) > 0$.

As shown in Figure 7, for expanded states when

$$\text{for} \qquad 48 < r^2_{i,i+4} < 145 \tag{13b}$$

a) residues i + 1 and i + 2 assigned as extended if $(\mathbf{v}_i \cdot \mathbf{v}_{i+2}) > 8$

b) residues i + 2 and i + 3 assigned as extended if $(\mathbf{v}_{i+1} \cdot \mathbf{v}_{i+3}) > 8$

c) residues i + 1, i + 2 and i + 3 assigned as extended if $(\mathbf{v}_i \cdot \mathbf{v}_{i+3}) < 0$.

The set of conditions given in Equations 13a and 13b describe various geometrical boundaries for the local conformation of the model chain that are characteristic for helical and expanded states, respectively. In each case, they were split into three sets of conditions to make the energy landscape as smooth as possible (otherwise, a single condition could be applied). In the present realization, the model system gains some energetical stabilization when even a nucleus of a helix or expanded state forms. On the other hand, the conditions are rather permissive, allowing substantial fluctuations of the secondary structure without an energetical penalty. This is the reason for certain cut-offs for intrachain distances and dot products of the relevant side-chain vectors. Of course, these cut-offs are consistent with the vast majority of helical or β-type geometries seen in globular proteins.

Long-range restraints

Long-range restraints are implemented in the form of a distorted harmonic potential. Additionally, the contact energy for such side chain pairs is modified as well below:

$$\begin{aligned} &E_{ij,\text{restrained}} \\ &= \infty, && \text{for } r_{ij} < 3 \\ &\quad E^{rep}, && \text{for } 3 \le r_{ij} < R^{rep}_{i,j} \\ &\quad \epsilon_{ij} - 0.5 && \text{for } R^{rep}_{i,j} \le r_{ij} < R_{i,j} \\ &\quad \epsilon_{res}(R^2_{i,j} - R^{2\,rep}_{i,j}) && \text{for } R_{i,j} < r_{ij} < 10 \\ &\quad \epsilon_{res}(100 - (r^2_{ij} - 100)/3 && \text{for } 10 < r_{ij}. \end{aligned} \tag{14}$$

The value of parameter ϵ_{res} in structure assembly runs was set equal to 1/8; while during the low temperature refinement run, it was set equal to 1/4. The meaning of other parameters is the same as in Equation 8. In the first three ranges, the above function is consistent with the definition of pairwise interactions defined in the previous section. For restrained residues, the pairwise potential has been enhanced (line 3 of Eq. 14). The two remaining lines define a pseudoharmonic long-distance potential. For longer distances (line 5), it is slightly suppressed. This is done for purely technical reasons because the weaker function facilitates the somewhat faster assembly of model protein chains.

Folding Procedure

The sampling procedure employed for protein assembly is based on Monte Carlo simulated thermal annealing. The stages are outlined below:

1. In the first step, a random expanded chain conformation is subjected to Monte Carlo simulated thermal annealing[38] over a broad range of temperature from T = 6 (T = 4 for smaller proteins) to T = 1. After annealing, the number of satisfied long-range restraints in each folded protein is inspected. Those folds with more than 1/7 of their restraints significantly violated are rejected without further inspection (i.e., when the corresponding side chain-side chain distance is larger than 7 lattice units for proteins smaller than 100 residues and 8 lattice units for proteins larger than 100 residues) This particular choice has been motivated by our previous studies of a similar problem[6] and by preliminary tests of the present model. Allowing a larger number of violated restraints may lead to topologically wrong folds, while requesting all restraints to be satisfied would decrease the efficiency of the method (some good folds with small local distortions would be rejected). The success ratio at this stage depends on the protein and the number of long-range restraints. For example, in 1gb1, protein G, with eight restraints, more than 75% of short assembly runs

(5–15 minutes of CPU time on a HP C-110 workstation) are successful. In the case of 1pcy, plastocyanin, with 15 restraints, the corresponding success rate is about 30% for 4-hour-long simulations on an HP C-110 workstation. Of course, a slower annealing protocol increases the fraction of assembled structures that satisfy the restraints. However, it appears that use of a larger number of shorter simulations is a more effective sampling protocol because a greater number of structures are collected for each protein.

2. All structures obtained via the rapid annealing procedure are subject to a refinement process. Each structure is duplicated and subjected to two independent Monte Carlo annealing runs over the temperature range $T = 2-1$. The lowest conformational energy structure (from the last snapshot of the corresponding trajectories) is accepted for further analysis.

3. For each protein, both the lowest energy conformation and the lowest energy alternative conformation are subject to isothermal runs. The purpose is to establish whether the proper fold could be automatically selected based on the choice of the lowest average energy structure.

4. The Cα coordinates of the final, lowest energy structures are rebuilt. The reconstruction procedure is based on Monte Carlo annealing of a phantom lattice model chain that has two united atoms per residue: one centered on the Cα and the other at the side chain center of mass. This Cα plus side chain center of mass, CAPLUS, model (described elsewhere[6,16,29–31,39]) only employs statistical potentials describing short-range interactions and side chain rotamer preferences. The positions of the side chains in the CAPLUS model are driven by a harmonic potential to the predicted side chain positions from the side chain only model.

RESULTS AND DISCUSSION

The test set employed in this work is representative of single domain water-soluble proteins[40] and consists of the following proteins that were previously studied[6] in the CAPLUS model: the small (structured) protein fragment of 6pti, chosen for comparison with the work of Smith-Brown et al.,[4] the all-α protein myoglobin (1mba), the α/β motifs of protein G, thioredoxin, flavodoxin, and an all-β protein, 1 pcy. In addition, we also examined the folding of a 247-residue TIM barrel, Atim, and the β-protein 4fab. The set of restraints used in this work is exactly the same as in our previous studies[6] in those cases where the studied system is the same. When a smaller number of restraints are used, these are randomly chosen from the larger restraint set. For the two proteins not studied previously, we report the set of employed long-range restraints in Tables II and III. The short-range restraints, as before, come from the three-letter code of the DSSP assignment[37] of the native secondary structure, and are as described in the Methods section.

TABLE II. Tertiary Restraint Lists for 4fab

27 restraints		16 restraints
1 PRO	ASN-100	xx
4 GLN	MET-95	
8 THR	PRO-107	xx
8 ILE	PRO-21	xx
11 ILE	LEU-21	
11 THR	LEU-107	
15 ILE	LEU-111	xx
19 LEU	ALA-109	xx
20 LYS	SER-79	xx
23 TRP	SER-40	
23 PHE	SER-76	
29 HIS	LEU-98	xx
34 LYS	GLY-55	
37 LYS	TYR-55	
39 TYR	ARG-54	xx
39 SER	ARG-94	xx
40 LEU	TRP-52	xx
44 GLY	LYS-89	xx
40 LEU	TRP-78	xx
42 ASP	LEU-87	
43 VAL	GLN-90	
53 LEU	ILE-78	xx
56 PHE	VAL-67	xx
67 ASP	PHE-87	
80 LEU	ILE-109	
92 GLY	PHE-106	xx
95 TRP	GLN-101	xx

Results of Monte Carlo Simulated Annealing

The results of stage 2 are compiled in Table IV. The numbers of restraints are given next to protein PDB codes.[14] An estimate of the cRMSD from the PDB structure and conformational energy (in dimensionless k_BT units) is given for the last snapshot of each trajectory. The cRMSD is measured between the Cα's of the real structure and the roughly estimated position of the Cα's of the model chain. The latter are obtained according to the following definition: $\mathbf{r}^C_{\alpha i} = (4\mathbf{r}_i + \mathbf{r}_{i-1} + \mathbf{r}_{i+1})/6$, where the sum in the brackets is over the corresponding side chain coordinates of the model chain. The exact agreement of the secondary structure of the predicted fold and the experimental structure was not examined in detail; however, in all runs, it was very close to the target with a small tendency for extension (by one or two residues) of helical fragments in some cases (e.g., the short helix of plastocyanin). The cRMSD and the energy (in dimensionless k_BT units) correspond to the last snapshots of the second simulated thermal annealing runs.

Generally, the predicted structures cluster into two well defined groups, one of which dominates on the basis of energy, and which is taken to approximate native structure. The remaining, misfolded

TABLE III. Tertiary Restraint Lists for Atim

Set of 62 restraints	Set of 50 restraints	Set of 37 restraints	Set of 62 restraints	Set of 50 restraints	Set of 37 restraints
2–228	xx	xx	91–125	xx	87–120
4–37	xx		91–231	xx	90–122
4–206	xx	xx	93–125	xx	
6–123	xx	xx	94–166	xx	
6–89			95–168	xx	
6–162			98–126	xx	xx
7–248	xx	xx	98–145	xx	xx
10–94	xx		105–145	xx	
11–64	xx	xx	105–148	xx	
11–237	xx	xx	109–152	xx	
15–46	xx	xx	112–149	xx	xx
20–49			112–161	xx	xx
23–237		24–54	116–153	xx	121–160
27–59			127–145	xx	
27–241		32–59	127–165	xx	
30–245	xx	xx	128–142		
36–58	xx	xx	128–165	xx	xx
26–248	xx	41–91	130–175	xx	xx
37–89	xx		133–181	xx	
39–123	xx	44–82	142–165	xx	
47–63			142–189	xx	xx
47–87			143–192	xx	
51–86	xx	xx	150–197	xx	
59–245	xx	xx	155–200	xx	xx
60–89			162–208	xx	xx
63–90	xx		165–189	xx	xx
66–79		67–111	165–209	xx	xx
68–114	xx		183–225	xx	xx
79–114	xx		193–205	xx	xx
89–162		82–120	215–244	xx	xx
90–122	xx	xx	230–248	xx	

structures (when observed more than once) were also similar to each other. They represent the topological mirror structure where the chirality of the connections between secondary structural elements (helices and β-strands) is reversed, but the chirality of the secondary structure elements is the same as in the native state, e.g., helices remain right handed. Several interesting observations emerge from the results presented in Table IV. First, in the majority of the runs, the native fold is recovered. The accuracy depends on protein size and number of restraints, but only slightly on protein type. Generally, the accuracy increases with decreasing protein size. The best accuracy is observed for the 56-residue, B1 domain of protein G,[41] where in most simulations the obtained structures had cRMSD from native below 3 Å. Interestingly, for the smaller 6pti fragment with a larger number of restraints, the accuracy was systematically somewhat worse. This reflects the effect of protein "regularity". The fold of protein G has a high content of regular secondary structure, while in the 6pti fragment, a substantial fraction of the chain is classified as a loop or coil. The analysis of other cases shows a tendency towards higher accuracy for more regular folds. The accuracy of helical and α/β proteins is greater than for all β-proteins. This is clearly demonstrated on comparison of 1pcy with 2trx. While both proteins are of comparable size, for 2trx with 16 restraints, structures with a cRMSD below 3.5 Å are produced, but for 1pcy with 15 restraints, structures above 5.2 Å result.

In the above cases, based on the conformational energy of just one (the last in a trajectory) snapshot, it was possible in all cases to identify the proper fold. However, it should be also noted that this very simple criterion may not always work. Indeed, in the case of the fourth set (S4) of long-range restraints for 6pti, the difference between the energy of a misfolded state and the lowest energy of properly folded states (simulation #3) is marginal. Moreover, the three remaining properly folded conformations have a higher energy than the misfolded one does. Fortunately, for bigger proteins, the situation is much better. The energy gap between the proper fold and misfolded states is usually quite large, except for the cases of all β-proteins with the smallest number of long-range restraints.

The reasons for the apparently lower reliability of the β-protein prediction are complex. At the present stage of development, our model and its force field

TABLE IV. Coordinate cRMSD and Conformational Energy of the Final Structure at the End of the Simulated Thermal Annealing Procedure

Name	Run no.	cRMSD (Å)	Energy
6pti (18)[a]	1	3.3[b]	−321.9
41 res	2	3.8	−313.2
(18–56 fragment)	3	4.1	−302.8
6pti (9/S1)	1	4.1	−336.4
	2	4.2	−345.4
	3	3.6	−318.9
	4	3.8	−385.9
6pti(9/S2)	1	3.8	−331.8
	2	4.3	−320.2
	3	4.0	−341.6
	4	4.4	−353.6
6pti(9/S3)	1	3.4	−303.1
	2	4.0	−318.7
	3	4.8	−324.5
	4	MI[c]	−323.2
	5	MI	−322.5
6pti(9/S4)	1	MI	−319.1
	2	3.8	−312.8
	3	4.0	−320.8
	4	4.2	−280.4
	5	4.1	−302.0
6pti(9/S5)	1	3.9	−370.0
	2	MI	−324.7
	3	MI	−283.3
	4	4.4	−355.2
	5	4.2	−338.2
1gb1(8)	1	2.4	−539.6
56 res	2	2.6	−527.7
	3	2.6	−530.2
	4	2.7	−562.3
	5	2.7	−548.0
	6	2.7	−542.0
	7	3.0	−550.5
	8	3.2	−586.7
	9	3.5	−563.7
	10	4.0	−551.0
	11	MI	−535.3
1ctf(10)	1	3.4	−710.5
68 res	2	3.6	−758.3
	3	3.7	−720.9
	4	3.7	−746.6
	5	4.1	−622.5
	6	MI	−655.1
	7	4.6	−700.0
	8	3.8	−692.0
	9	3.2	−727.2
	10	3.3	−749.4
1pcy(46)	1	3.1	−841.8
99 res	2	3.6	−824.5
	3	3.5	−787.2
	4	3.8	−783.3
	5	3.9	−834.7
	6	3.5	−848.0
	7	MI	−744.2
1pcy(25)	1	4.7	−944.3
	2	4.8	−786.7
	3	4.5	−898.0
	4	5.2	−928.4
1pcy(15)	1	MI	−870.7
	2	5.6	−860.8
	3	5.2	−874.7
	4	5.3	−925.1
2trx(16)	1	3.8	−1098
108 res	2	3.5	−1089
	3	4.5	−1022
2trx(30)	1	2.8	−1036
	2	3.2	−1037
	3	3.7	−1041
	4	MI	−844
4fab(27)	1	4.5	−959
111 res	2	4.4	−1006
	3	4.9	−1037
	4	4.1	−1031
	5	MI	−984
4fab(16)	1	5.0	−1042
	2	5.1	−1062
	3	4.8	−1041
	4	5.5	−953
	5	4.9	−1035
	6	5.8	−1090
	7	MI	−1005
	8	MI	−1033
	9	MI	−1062
3fxn(35)	1	3.6	−1441
138 res	2	3.8	−1447
	3	3.6	−1432
	4	3.5	−1485
	5	4.5	−1409
	6	3.5	−1493
	7	4.4	−1464
	8	4.1	−1533
	9	MI	−1289
3fxn(20)	1	3.7	−1447
	2	3.9	−1464
	3	3.3	−1511
	4	4.2	−1515
	5	4.2	−1503
	6	4.5	−1499
1mba(20)	1	3.5	−1705
146 res	2	3.7	−1733
	3	4.1	−1705
	4	5.6	−1605
	5	4.2	−1849
	6	5.0	−1570
	7	4.1	−1741
Atim(62)	1	5.0	−2412
247 res	2	5.6	−2357
	3	5.7	−2417
	4	5.8	−2491
	5	5.4	−2499
Atim(50)	1	5.9	−2428
	2	6.5	−2507
	3	5.9	−2540
	4	6.2	−2509
Atim(36)	1	6.6	−2469
	2	6.4	−2599
	3	6.3	−2558
	4	MI	−2593
	5	6.5	−2526
	6	6.5	−2643

[a] In parentheses, the number of long-range restraints, S1, S2, . . . S5, various sets of restraints for 18–56 residue 6pti fragment.

[b] Coordinate root mean square deviation between crystallographic coordinates of Cα's and the approximate positions of the model Cα's calculated as: $\mathbf{r}_i^{C\alpha} = (4\mathbf{r}_i + \mathbf{r}_{i-1} + \mathbf{r}_{i+1})/6$.

[c] MI, misfolded structure that has satisfied the long-range restraints, generally the topological mirror image fold.

are not capable of folding complex β-type natural proteins without the assistance of some long-range restraints. These proteins have a larger number of building blocks (compare the number of β-strands in an all-β protein with the number of helices in a helical structure of similar size) and, consequently, more complex folds. Thus, the same number of long-range restraints provides relatively less structural information for β-proteins. As a result, the demands on the force field with a given (small) number of restraints are greater. While in all cases examined here, the proper fold could be identified by choosing the lowest energy final conformation, for a number of cases, this was just by good luck. In reality, in these doubtful cases, the magnitude of the energy fluctuations was larger than the observed energy difference for the final states. Certainly, a more dependable protocol for the selection of the proper fold is necessary, and one such protocol is described in the next section.

Structure Refinement and Selection of Native Folds

As described in the Methods section, the lowest-energy final structures from simulated annealing, representing the putative proper fold and the closest competing alternative topology, were subjected to isothermal Monte Carlo runs using the same force field and sets of restraints. The results of these stage 3 runs are summarized in Table V. All simulations were done at $T = 1$. The average cRMSD from native and the average energy are computed from 200 snapshots of the Monte Carlo trajectory. In all cases, the proper fold can be identified based on the average conformational energy. Thus, a combination of fast-simulated annealing and long isothermal runs allows the dependable selection of the proper fold. Indeed, during rapid assembly via Monte Carlo-simulated annealing, a fine-tuning of structural details is not always achieved. In long isothermal runs, the misfolded (topological mirror image conformations) states could always be detected as those of higher average conformational energy. For the case of 6pti, where five different sets of restraints were examined, the lowest-energy misfolded structure has a higher conformational energy than the highest-energy proper fold, regardless of the set of restraints. On average, the accuracy of the predicted native fold improved slightly during the isothermal runs and ranges between 3 and 5 Å cRMSD (for the estimated positions of Cα's), except for the Atim barrel where it is about 6 Å. By way of illustration, in Figures 8 and 9 we present a representative conformation (generated using the MOLMOL[42] procedure) of 3fxn and 4fab obtained from the isothermal refinement runs (employs 20 and 16 restraints, respectively) with a cRMSD of 4.4 Å and 5.5 Å, respectively.

Increasing the number of long-range restraints, on average, leads to some increase in the compactness of the obtained structures, as assessed by their average root-mean-square radius of gyration. There is no obvious systematic difference between the dimensions of the native and misfolded states. Since in both cases the majority of restraints are always satisfied, the difference in conformational energy arises from the underlying force field that has a reasonable level of specificity for nativelike structures. Unfortunately, the non-restraint contributions to the potential are not sufficiently specific to fold the protein (except for a few small proteins) without the assistance of the restraints. On the other hand, within the limit of N/7 restraints, if the restraints are used alone without the remainder of the potential, the resulting structures are essentially random. Thus, it is the synergism of the restraints with the underlying contributions to the potential that permits the folding of these proteins. For some of the test proteins, good folds could be obtained with a smaller than N/7 restraints (e.g., 4 restraints for protein G). The value of N/7 is a conservative estimate of a safe lower bond for all proteins. This number is smaller than required by related methods.[4–6]

Next, the side-chain-based lattice models serve as a target for building a model with two united atoms per residue, i.e., in the CAPLUS model. Table VI displays the cRMSD data for such reconstructed main chains. Somewhat surprisingly, there is no significant difference between the average quality of the rebuilt Cα chains and that roughly estimated from a simple linear combination of three successive side chain centers of mass. This shows that the side chain model is consistent with the CAPLUS model used previously. The Cα reconstruction process employed here neglects all the long-range interactions (except of course the target harmonic restraints). This is done for the sake of computational efficiency. Since this is a rather marginal point for the present studies, where a rapid algorithm for topology assembly is discussed, we refrain from further discussion of various possible methods of refinement of the structures produced by this protocol.

Comparison With Other Work

As mentioned in the Introduction, there have been several other attempts to use known secondary structure and some tertiary restraints in the prediction of protein three-dimensional structures. However, the closest studies of other workers who used both known secondary structure and exact tertiary restraints are those of Smith-Brown and coworkers[4] and Aszodi and Taylor[5]. Smith-Brown et al. have examined a number of proteins. By way of example, flavodoxin, a 138-residue α/β protein, was folded to a structure whose backbone cRMSD from native was 3.18 Å for 147 restraints. In contrast, with just 20

TABLE V. Compilation of the Results of the Isothermal Simulations

Name	Fold	Average cRMSD	cRMSD (A) Cα-fit	Average energy	$\langle S \rangle^{1/2}$ (Å)
6pti(9/S1)	NAT[a]	3.69 (0.21)	4.28 (0.29)[c]	−323.4	10.6
41 res	MI[b]	Not observed			
18–56					
6pti(9/S2)	NAT	3.67 (0.22)	3.60 (0.11)	−326.1	10.6
	MI	Not observed			
6pti(9/S3)	NAT	4.41 (0.12)	4.23 (0.07)	−325.0	9.9
	MI	8.01 (0.21)		−301.0	10.6
6pti(9/S4)	NAT	4.01 (0.29)	4.05 (0.22)	−327.6	10.6
	MI	8.11 (0.34)		−309.8	9.6
6pti(9/S5)	NAT	4.06 (0.24)	4.27 (0.14)	−349.7	10.8
	MI	8.34 (0.23)		−318.4	10.2
1gb1(8)	NAT	3.11 (0.13)	3.39 (0.14)	−582.9	10.9
56 res	MI	8.61 (0.13)		−567.2	11.5
1ctf(10)	NAT	3.48 (0.25)	3.21 (0.08)	−699.9	11.2
68 res	MI	8.68 (0.16)		−656.9	11.7
1pcy(46)	NAT	3.44 (0.11)	3.80 (0.08)	−856.5	12.7
99 res	MI	11.34 (0.08)		−796.9	12.7
1pcy(25)	NAT	4.87 (0.12)	4.88 (0.04)	−952.5	13.1
	MI	Not observed			
1pcy(15)	NAT	5.27 (0.06)	5.70 (0.16)	−891.7	13.0
	MI	7.70 (0.08)		−841.8	12.9
2trx(30)	NAT	3.63 (0.15)	3.11 (0.14)	−1013	13.0
108 res	MI	Not observed			
2trx(16)	NAT	3.43 (0.12)	3.52 (0.06)	−1082	13.3
	MI	11.88 (0.11)		−888	13.2
4fab(27)	NAT	4.77 (0.06)	4.42 (0.07)	−1040	13.8
111 res	MI	11.49 (0.13)		−1011	13.8
4fab(16)	NAT	5.53 (0.08)	5.92 (0.10)	−1137	14.1
	MI	12.76 (0.09)		−1033	13.8
3fxn(35)	NAT	3.91 (0.12)	4.06 (0.08)	−1514	14.3
138 res	MI	12.94 (2.33)		−1311	14.3
3fxn(20)	NAT	4.44 (0.22)	4.12 (0.14)	−1401	14.3
	MI	Not observed			
1mba(20)	NAT	4.44 (0.23)	4.34 (0.05)	−1698	15.0
146 res	MI	Not observed			
Atim(62)	NAT	5.19 (0.10)	5.08 (0.11)	−2423	17.4
247 res	MI	Not observed			
Atim(50)	NAT	5.77 (0.06)	5.96 (0.04)	−2483	17.7
	MI	Not observed			
Atim(36)	NAT	6.66 (0.09)	6.74 (0.15)	−2622	17.9
	MI	9.97 (0.05)		−2549	17.9

[a]Native structure.
[b]Misfolded (generally the topological mirror image fold) structure.
[c]The number in parentheses is the standard deviation of the coordinate root-mean-square distance (in Angstroms) between the crystallographic and predicted α-carbon traces (see also the legend for Table IV).

restraints, here structures whose cRMSD from native is 4.2 Å are generated. Similarly for 3fab, they required 90 restraints to produce a model whose cRMSD is 4.6 Å. For 4fab in the present approach, use of just 27 restraints yields a model whose cRMSD is 4.4 Å. Their requirement for a large number of restraints is perhaps due to the lack of knowledge-based, proteinlike background potential.

Another effort to predict the global fold of a protein from a limited number of distance restraints is due to Aszodi et al.[5] In general, they find that to assemble structures below 5 Å cRMSD, on average, typically more than N/4 restraints are required, where N is the number of residues. Even then, the Aszodi et al. method has problems selecting out the correct fold from competing alternatives. While their best folds are of acceptable accuracy, the competing misfolded structures could be disregarded based on energetic considerations. In contrast, in the simulations presented here, the nativelike fold could be easily detected as the lowest energy structure, and just N/7 restraints are required to produce structures of comparable accuracy. However, the Aszodi et al. approach is very rapid, with a typical calculation taking on the order of minutes on a typical contemporary workstation.

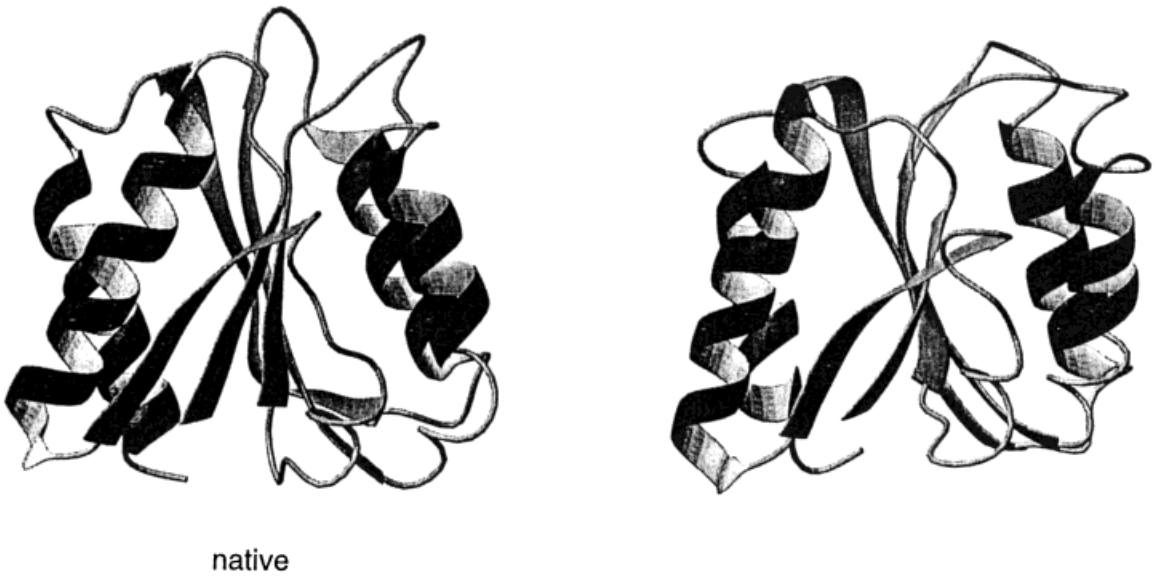

Fig. 8. Fold of 3fxn obtained using 20 tertiary restraints compared with the native structure. The picture has been prepared using MOLMOL.[42] The native secondary structure boundaries (helices and β-strands) have been superimposed on the predicted structure. A slight distortion of one helix (bottom right of the figure) and some distortions of the central β-sheet are noticeable.

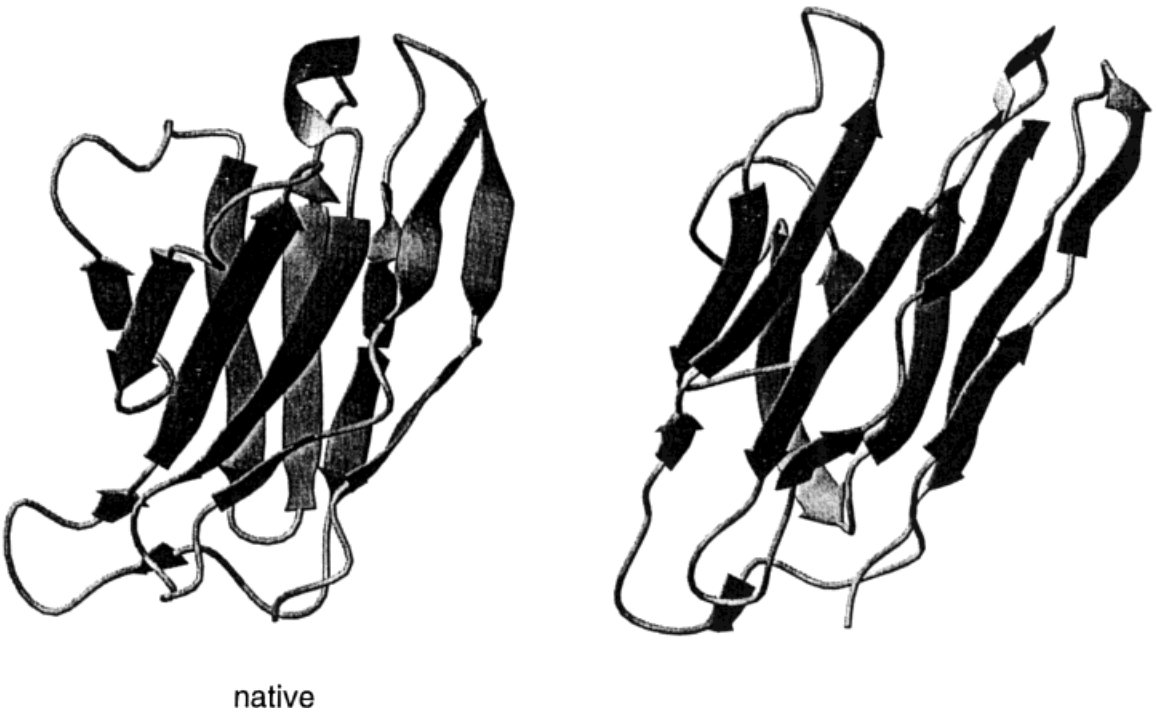

Fig. 9. Representative structure of 4fab obtained using 16 tertiary restraints compared with the native structure.

In a very general sense, our current method is most similar to the recently developed MONSSTER algorithm that uses the CAPLUS model.[6] It also employs a reduced lattice model of protein, a background knowledge-based force field and a simulated thermal annealing Monte Carlo procedure for fold assembly. Using MONSSTER, about N/4 restraints are required to assemble β-type and α/β-proteins, while helical proteins required N/7 restraints. Here, for a representative set of proteins, all types of folds

TABLE VI. Comparison of Results for the *CAPLUS* and *SICHO* Models With Exact Secondary and Tertiary Restraints

PDB name	Number of residues	Type	Number of restraints	cRMSD in Å from the *SICHO* model[a,b]	cRMSD in Å from the *CAPLUS* model[a]
1gb1	56	α/β	8	3.4	3.3
1ctf	68	α/β	10	3.2	4.2
1pcy	99	β	46	3.8	3.5
1pcy	99	β	25	4.9	5.4
1pcy	99	β	15	5.7	—
2trx	108	α/β	30	3.1	3.4
2trx	108	α/β	16	3.5	—
4fab	113	β	27	4.4	–
4fab	113	β	16	5.9	–
3fxn	138	α/β	35	4.1	3.9
3fxn	138	α/β	20	4.1	—
1mba	146	α	20	4.3	5.9
Atim	247	α/β	62	5.1	—
Atim	247	α/β	50	6.0	—
Atim	247	α/β	36	6.7	—

[a]Average cRMSD of the Cα over an isothermal stability run.
[b]The average cRMSD is reported from structures obtained after the *SICHO* model has been mapped into the *CAPLUS* model and relaxed.

could be assembled with knowledge of, on average, N/7 tertiary restraints. In addition, the results are less sensitive to the distribution of restraints. For example, in the case of the 18–55 fragments of 6pti in the CAPLUS model, the cRMSD was about 6–8 Å for different sets of restraints. In contrast, in the side-chain-based model, for all sets of examined restraints, it is about 4 Å. Furthermore, as evidenced by the ability to fold the 247-residue TIM from a fully expanded state, much larger systems can be treated. With an increasing number of long-range restraints, the accuracy of assembled structures increases and seems be consistently better than for the previously published methods. In addition, the resulting models are found to be less sensitive to the restraint distribution. The current model also offers the advantage of speed. For small proteins, the algorithm is now essentially interactive. It takes about 5–10 minutes of CPU time on a contemporary workstation to assemble the relatively complex motif of a 68-residue 1ctf fragment. Since the cost scales approximately as N^3, assembly of larger structures requires more time. Thus, a myoglobin folding simulation requires about 2 hours of CPU time.

CONCLUSIONS

In this work, we have described a new model for the assembly of protein structures from known secondary structure and a small number of exact tertiary restraints. While the model only explicitly considers side chain centers of mass, the effect of backbone atoms is implicitly built into the model force field, which also exploits the structural regularities seen in protein structures. Thus, the new model is fully compatible with more complex models that employ a larger number of united atoms per residue. In all respects, this new method compares favorably with previous approaches having a similar goal; the assembly of tertiary structure from loosely encoded secondary structural biases and a small number of tertiary restraints. The most important aspect of the new model is the very small number of tertiary restraints required to assemble moderate resolution folds. For a representative set of all types of single domain proteins (all-α, all-β, α/β motifs), the required number of restraints is about N/7, with N the number of residues in the protein. Furthermore, due to a new, rapid treatment of side chain burial, it should be possible to extend this method to multidomain proteins, which will be attempted in the near future.

Another advantage of the new method is associated with the relatively simple and reliable protocol of detecting a proper fold from less frequently generated misfolded structures. These misfolded structures are almost exclusively the topological mirror images of the proper fold. In all cases examined to date, the nativelike structure always has a lower conformational energy. This and the small number of required tertiary restraints suggest that the new model's underlying force field captures a number of the essential aspects of protein interactions. At the same time, the model is simpler and computationally more efficient than previously employed lattice models.[6,39] Due to a much lower computational cost (at least one order of magnitude), it is possible to assemble larger structures, including the 247-residue Atim domain.

Finally, we note that preliminary work indicates that it is possible to generate low resolution folds

using only a small set of probable side chain contacts[35] (predicted via correlated mutations analysis[43]) and somewhat more elaborate potentials describing short-range interactions (derived from geometrical analysis of sequentially similar protein fragments). Several example structures such as myohemerythrin (1hmd) and the complex β-type motif immunoglobulin fold (1fna), have been assembled, albeit with a lower fraction of successful experiments than reported in this work. While this application of the new model requires further refinement, it clearly provides a well-defined framework for addressing the more general aspects of protein structure prediction.

ACKNOWLEDGMENTS

Andrzej Kolinski acknowledges support from the University of Warsaw Grant BST-34/97 and is an International Scholar of the Howard Hughes Medical Institute.

REFERENCES

1. Friesner, R.A., Gunn, J.R. Computational studies of protein folding. Annu. Rev. Biophys. Biomol. Struct. 25:315–342, 1996.
2. Levitt, M. Protein folding. Curr. Opin. Struct. Biol. 1:224–229, 1991.
3. Anfinsen, C.B., Scheraga, H.A. Experimental and theoretical aspects of protein folding. Adv. Protein Chem. 29:205–300, 1975.
4. Smith-Brown, M.J., Kominos, D., Levy, R.M. Global folding of proteins using a limited number of distance restraints. Protein Eng. 6:605–614, 1993.
5. Aszodi, A., Gradwell, M.J., Taylor, W.R. Global fold determination from a small number of distance restraints. J. Mol. Biol. 251:308–326, 1995.
6. Skolnick, J., Kolinski, A., Ortiz, A.R. MONSSTER: A method for folding globular proteins with a small number of distance restraints. J. Mol. Biol. 265:217–241, 1997.
7. Kaptein, R., Boelens, R., Scheek, R.M., van Gunsteren, W.F. Protein structures from NMR. Biochemistry 27:5389–5395, 1988.
8. Gronenborn, A.M., Clore, G.M. Where is NMR taking us? Proteins 19:273–276, 1994.
9. Braun, W., Go, N. Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. J. Mol. Biol. 186:611–626, 1985.
10. Havel, T.F., Wuthrich, K. An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformation in solution. J. Mol. Biol. 182:281–294, 1985.
11. Havel, T.F. An evaluation of computational strategies for use in the determination of protein structures from distance constraints obtained by nuclear magnetic resonance. Prog. Biophys. Mol. Biol. 56:43–78, 1991.
12. Mumenthaler, C., Braun, W. Automated assignment of simulated experimental NOESY spectra of protein feedback littering and self-correcting distance geometry. J. Mol. Biol. 254:465–480, 1995.
13. Guentert, P., Braun, W., Wuthrich, K. Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. J. Mol. Biol. 217:517–530, 1991.
14. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Simanouchi, T., Tasumi, M. The protein data bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
15. Kolinski, A., Skolnick, J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins 18:338–352, 1994.
16. Kolinski, A., Skolnick, J. "Lattice Models of Protein Folding, Dynamics and Thermodynamics." Austin, TX: R.G. Landes Co., 1996.
17. Kolinski, A., Skolnick, J. Parameters of statistical potentials. Available by ftp from public directory scripps.edu(pub/andr/side_only/*). 1997.
18. Godzik, A., Skolnick, J., Kolinski, A. Regularities in interaction patterns of globular proteins. Protein Eng. 6:801–810, 1993.
19. Kyte, J., Doolittle, R.F. A simple method for displaying the hydrophatic character of protein. J. Mol. Biol. 157:105–132, 1982.
20. Skolnick, J., Jaroszewski, L., Kolinski, A., Godzik, A. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Sci. 6:676–688, 1997.
21. Kolinski, A., Godzik, A., Skolnick, J. A general method for the prediction of the three dimensional structure and folding pathway of globular proteins. Application to designed helical proteins. J. Chem. Phys. 98:7420–7433, 1993.
22. de Gennes, P.G. "Scaling Concepts in Polymer Physics." Ithaca, NY: Cornell University Press, 1979.
23. Kolinski, A., Skolnick, J. Determinants of secondary structure of polypeptide chains: Interplay between short range and burial interactions. J. Chem. Phys. 107:953–964, 1997.
24. Eisenberg, D., McLauchlan, A.D. Solvation energy in protein folding and binding. Nature 319:199–203, 1986.
25. Godzik, A., Kolinski, A., Skolnick, J. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. Protein Sci. 4:2107–2117, 1995.
26. Godzik, A. Knowledge-based potential for protein folding: What can we learn from known structures? Curr. Biol. 4:363–366, 1996.
27. Kolinski, A., Jaroszewski, L., Rotkiewicz, P., Skolnick, J. An efficient Monte Carlo model of protein chains. Modeling the short-range correlations between side group centers of mass. J. Phys. Chem. 102:4628–4637, 1998.
28. Kolinski, A., Skolnick, J. Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. Proteins 18:353–366, 1994.
29. Kolinski, A., Galazka, W., Skolnick, J. Computer design of idealized β-motifs. J. Chem. Phys. 103:10286–10297, 1995.
30. Kolinski, A., Milik, M., Rycombel, J., Skolnick, J. A reduced model of short range interactions in polypeptide chains. J. Chem. Phys. 103:4312–4323, 1995.
31. Kolinski, A., Galazka, W., Skolnick, J. On the origin of the cooperativity of protein folding. Implications from model simulations. Proteins 26:271–287, 1996.
32. Olszewski, K., Kolinski, A., Skolnick, J. Does a backwardly read protein sequence have a unique native state? Protein Eng. 9:5–14, 1996.
33. Olszewski, K., Kolinski, A., Skolnick, J. Folding simulations and computer redesign of protein A three-helix bundle motifs. Proteins 25:286–299, 1996.
34. Ortiz, A.R., Hu, W.-P., Kolinski, A., Skolnick, J. A method for prediction of the tertiary structure of small proteins. J. Mol. Graph, in press.
35. Ortiz, A.R., Hu, W-P., Kolinski, A., Skolnick, J. Method for low resolution prediction of small protein tertiary structure. In: "Proceedings of the Pacific Symposium on Biocomputing'97," Altman, R.B., Dunker, A.K., Hunter,L., Klein, T.E. (eds.). Singapore: World Scientific Pub., 1997: 316–327.
36. Skolnick, J., Kolinski, A. Monte Carlo lattice dynamics and the prediction of protein folds. In: "Computer Simulations

of Biomolecular Systems. Theoretical and Experimental Studies," van Gunsteren, W. F., Weiner, P.K., Wilkinson, A.J. (eds.). The Netherlands: ESCOM Science Pub. 395–429, 1997.
37. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
38. Binder, K. "Monte Carlo Methods in Statistical Physics." Berlin: Springer-Verlag, 1986.
39. Skolnick, J., Kolinski, A. Protein modelling. In: "Encyclopedia of Computational Chemistry," Schleyer, P., Kollman, P. (eds.). Sussex, England: John Wiley & Sons, in press.
40. Richardson, J. The anatomy and taxonomy of protein structure. Adv. Protein Chem. 34:167–339, 1981.
41. Gronenborn, A., Filpula, D.R., Essig, N.Z., Achari, A., Whitlow, M., Wingfield, P.T., Clore, G.M. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. Science 253:657–660, 1991.
42. Koradi, R. MOLMOL: a program for display and analysis of macromolecular structures. J. Mol. Graph. 14:51–55, 1996.
43. Goebel, U., Sander, C., Schneider, R., Valencia, A. Correlated mutations and residue contacts in proteins. Proteins 18:309–317, 1994.