J. Mol. Biol. (1997) 265, 217–241

JMB

MONSSTER: A Method for Folding Globular Proteins with a Small Number of Distance Restraints

Jeffrey Skolnick[1]*, Andrzej Kolinski[1,2] and Angel R. Ortiz[1]

[1]*The Scripps Research Institute, Department of Molecular Biology 10555 North Torrey Pines Road, La Jolla, California 92037, USA*

[2]*Department of Chemistry University of Warsaw Pasteura 1, 02-093 Warsaw Poland*

**Corresponding author*

The MONSSTER (**MO**deling of **N**ew **S**tructures from **S**econdary and **TE**rtiary **R**estraints) method for folding of proteins using a small number of long-distance restraints (which can be up to seven times less than the total number of residues) and some knowledge of the secondary structure of regular fragments is described. The method employs a high-coordination lattice representation of the protein chain that incorporates a variety of potentials designed to produce protein-like behaviour. These include statistical preferences for secondary structure, side-chain burial interactions, and a hydrogen-bond potential. Using this algorithm, several globular proteins (1ctf, 2gbl, 2trx, 3fxn, 1mba, 1pcy and 6pti) have been folded to moderate-resolution, native-like compact states. For example, the 68 residue 1ctf molecule having ten loosely defined, long-range restraints was reproducibly obtained with a C^{α}-backbone root-mean-square deviation (RMSD) from native of about 4. Å. Flavodoxin with 35 restraints has been folded to structures whose average RMSD is 4.28 Å. Furthermore, using just 20 restraints, myoglobin, which is a 146 residue helical protein, has been folded to structures whose average RMSD from native is 5.65 Å. Plastocyanin with 25 long-range restraints adopts conformations whose average RMSD is 5.44 Å. Possible applications of the proposed approach to the refinement of structures from NMR data, homology model-building and the determination of tertiary structure when the secondary structure and a small number of restraints are predicted are briefly discussed.

Keywords: NMR structure refinement; protein modeling; protein folding; lattice models; reduced protein models

Introduction

Recently, several investigators have addressed the problem of obtaining a low to moderate-resolution protein structure from a limited number of distance restraints and some knowledge of the secondary structure. By providing an initial guess of a structure for subsequent refinement, a successful method could be a very useful tool in NMR-based model building (Aszodi *et al*., 1995; Smith-Brown *et al*., 1993). It could be used in homology modeling procedures, as well as in the construction of models based on a set of experimentally provided or theoretically predicted distance restraints. This class of models is complementary to more standard NMR structure determination algorithms (Braun & Go, 1985; Clore *et al*., 1993; Gronenborn & Clore, 1994; Guentert *et al*., 1991; Havel & Wuthrich, 1985). There, one starts from a very large number of restraints and, while explicitly neglecting intraprotein interactions, the structure is derived from purely geometrical considerations. The most prevalent variants of this approach are distance geometry (Havel & Wuthrich, 1985) and distance geometry supplemented by molecular dynamics refinement (Guentert *et al*., 1991). When one has a large number of distance restraints, these are clearly the techniques of choice. On the other hand, especially in the early stages of NMR structure refinement when a relatively small number of restraints per residue are known, or when significant line broadening at the NMR spectrum is present, one requires a protein model that already has many built-in protein-like regularities and interactions (Kolinski & Skolnick, 1994b). This might prove useful as part of an automated NMR spectrum assignment procedure (Mumenthaler & Braun, 1995). In such situations, one might expect that a small number of restraints would facilitate the

0022–2836/97/020217–15 $25.00/0/mb960720

relatively fast convergence to compact, moderate-resolution protein-like structures. The conformation and quality of such structures would be dictated in part by the model force-field as well as by the particular sets of restraints. Of course, the set of restraints has to be in reasonable agreement with the target structure; however, it may be possible to override some errors in the restraints by various consistency requirements. Here, we describe the application of a high-coordination lattice model of protein structure and dynamics to the problem of structure determination from a small number of geometrical restraints (Kolinski & Skolnick, 1994a,b; Olszewski *et al*., 1996a,b; K. A. Olszewski, A. Kolinski & J. Skolnick, unpublished results). We term this the MONSSTER (MOdeling of New Structures from Secondary and TErtiary Restraints) method.

Similar reduced models were recently used in studies of unrestrained, i.e. *de novo*, folding of several small proteins from random conformations (Kolinski & Skolnick, 1994a,b; Olszewski *et al*., 1996a,b; K. A. Olszewski *et al*., unpublished results; Vieth *et al*., 1994). For several small proteins, using sequence information alone, the force-field resulted in native-like structures, whose root-mean-square deviation (RMSD) of the C^{α} atoms ranged from 2 to 5 Å from native. These proteins had a relatively simple native topology. For larger proteins having topologically more complex folds, the methodology failed. The origin of this failure has various causes: (1) the global fold was obtained with very low reproducibility; (2) only a portion of the structures was predicted correctly; or (3) the energy difference between native-like and misfolded structures was too small to confidently select the native structure. Nevertheless, in all cases, large fragments of these more complex proteins frequently assemble with reasonable accuracy, thereby demonstrating that the model has a strong tendency towards protein-like behavior. These observations provide the inspiration for the present series of studies. If the model is basically correct, but is unable to find the protein fold due to small (but nevertheless very important) deficiencies in the force-field, sampling realization, or from the simulation time being too short, then with some bias in the form of NMR-like restraints, it should be relatively easy to obtain moderate-resolution, native structures. In such a case, the model force-field would be expected to ''interpolate'' between sparsely distributed restraints.

There have been a number of studies with a similar objective; namely, given the correct secondary structure and a limited number of long-range restraints, predict the global fold of a globular protein. In particular, Smith-Brown *et al*. (1993) have modeled proteins as a chain of glycine residues. Local bond length, bond angle and Lennard-Jones interactions are implemented in the context of the program IMPACT. NMR restraints are encoded *via* a biharmonic potential, and the secondary structure is assigned to the ideal ϕ,ψ angles. Folding is forced to proceed sequentially *via* successive implementation of the restraints. A number of proteins was examined; by way of example, flavodoxin, a 138 residue α/β protein, was folded to a structure whose backbone RMSD from native was 3.18 Å for 147 restraints. This is to be contrasted to results reported here where structures at the level of 3.98 Å are obtained for a set of 35 randomly chosen restraints. Part of the difficulty Smith-Brown *et al*. (1993) experienced may be due to the lack of knowledge-based information in their force-field that can encode protein-like behavior. The incorporation of such knowledge-based potentials represents an important aspect of the current line of investigation.

Another effort to predict the global fold of a protein from a limited number of distance restraints is due to Aszodi *et al*. (1995). Their approach is based on distance geometry, where a set of experimental tertiary distance restraints are supplemented by a set of predicted interresidue distances. These distances are obtained from patterns of conserved hydrophobic amino acids that have been extracted on the basis of multiple sequence alignments. The protein is modeled in a C^{α}-C^{β} representation, where the distances between the C^{α}-C^{β} atoms and the van der Waals radii of the C^{β} atoms depend on the amino acid types. In general, they find that to assemble structures below 5 Å RMSD, on average, typically more than $N/4$ restraints are required, where N is the number of residues. Even then, the Aszodi *et al*. (1995) method has problems selecting out the correct fold from competing alternatives. For the case of thioredoxin, 2trx (Katti *et al*., 1990) considered in detail in both their work and here, when they use 30 restraints, half of their models exhibited an RMSD from native of 4.5 Å, while the other half had an RMSD up to 9 Å. They lack a method of selecting between the two structural families. The present method gives an RMSD from native of 3.4 Å when the lowest-energy structures are considered. Thus, a more sensitive energy function is provided by the present approach that allows one to choose the correct topology. However, the advantage of the Aszodi *et al*. (1995) approach is that it is very rapid, with a typical calculation taking on the order of minutes on a typical contemporary workstation. In contrast, the method developed here is relatively CPU-intensive, with each folding simulation typically requiring an overnight run on an HP-735 workstation, and for ambiguous cases, multiple runs have to be performed.

The method described in what follows differs from those of previous investigators in a number of important respects. Assembly occurs in the context of a reduced protein representation that includes an α-carbon backbone, as well as a single-ball, multiple rotamer representation of the side-chains. The underlying, knowledge-based force-field contribution dominates the assembly process. Thus, we need only fragmentary information about the secondary structure (only regular fragments of

β-sheets and α-helices experience short-range, soft restraints that are in fact quite broad) and a relatively small number of loosely defined, long-range restraints. We have chosen a number of small to moderately sized proteins to examine the sensitivity of the approach to protein size and secondary structural class.

Algorithm

Protein model and the force-field

The protein model is based on a lattice representation of the C^{α}-backbone trace. The details of the model have been described, and here, we give a brief summary for the reader's convenience (Kolinski & Skolnick, 1994a,b; Olszewski *et al*., 1996a,b; K. A. Olszewski *et al*., unpublished results). The reader who is familiar with our previous work or who is more interested in the specific implementation of the MONSSTER method may skip the rest of this section.

Lattice model of protein chain

The C^{α} backbone is a string of vectors of the type $a\cdot\mathbf{v}$ with $\{\mathbf{v}\} = \{(3, 1, 1), \ldots (3, 1, 0), \ldots (3, 0, 0), \ldots (2, 2, 1), \ldots (2, 2, 0), \ldots\}$. The value of the parameter $a = 1.22$ Å has been optimized to obtain the best fit of the lattice representation to high-resolution, sequentially non-homologous protein structures from the Brookhaven Protein Data Bank (PDB). All protein structures can be fit to the lattice with an accuracy between 0.6 and 0.7 Å RMSD (Godzik *et al*., 1993). The virtual bond angles between successive C^{α} atoms are restricted to reproduce a protein-like distribution. The definition of the end side-groups is facilitated by the addition of two dummy residues (or N and C caps). Consequently, a chain of N real amino acids is built from $N + 1$ vectors and has $N + 2$ total residues. Two successive backbone vectors provide the reference frame for the definition of a set of model (single interaction center for each side-group) rotamers that cover the conformational space of the side-chains with a 1 Å grid. A pseudo, single-rotamer side-chain located at a very small distance (0.3 Å) from the C^{α} atom is assumed for glycine.

The sampling of conformational space occurs *via* a standard asymmetric Monte Carlo Metropolis scheme (Metropolis *et al*., 1953). The conformational updates are composed of several types of local conformational micromodifications of the chain backbone and their associated side-groups, side-group equilibration cycles, and rare (small distance) motions of larger chain fragments.

Force-field of the protein model

The force-field contains potentials of mean force (predominantly of statistical origin) that account for the short-range interactions, long-range interactions and hydrogen-bond interactions (which could be short-range or long-range). All contributions to the potential are available *via* anonymous ftp (Skolnick, 1996).

The sequence-dependent, local conformational propensities were factorized into the following form:

$$E_{14} = \Sigma\epsilon_{14}(A_i, A_{i+1}, r^{2*}_{i-1,i+2}) \tag{1a}$$

with:

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

This component depends on the sequence through the pair of neighboring amino acid residues A_i and A_{i+1}, and controls the local chain geometry. In practice, $r^2_{i-1,i+2}$ is divided into six conformational states that roughly correspond to extended right-handed and extended left-handed states, wide right and wide left turns, and right-handed and left-handed helical states, respectively.

Long-range interactions are encoded *via* two potentials of mean force. The first term is a one-body, centrosymmetric potential that reflects the tendency of some amino acid residues to be buried and some to be exposed. It is important to note that this potential is applicable only to single-domain proteins. If the number of restraints is not too small, then this contribution to the potential could be neglected and, consequently, this methodology could then be applied to multidomain proteins:

$$E_1 = \Sigma\epsilon_1(r(A_i)/S_0) \tag{2}$$

with:

$$S_0 = 2.2N^{0.38}\ \text{Å}$$

where S_0 is the expected radius of gyration of a single-domain protein consisting of N amino acids in its native conformation. $r(A_i)$ is the distance of the center of mass of the ith side-group from the center of mass of the entire chain. Additionally, if the first and last residues of the protein are found in the interior shells, they are energetically penalized by $4k_BT$, with k_B Boltzmann's constant and T the absolute temperature.

Pairwise tertiary interactions are neglected between nearest neighbors down the chain, since these short-range interactions are already accounted for by the hydrogen-bond potential (see below) and secondary structure preferences. The pair interactions beyond the fourth neighbor are derived from the statistics of the database. For residues i and j:

$$E_{ij} = \begin{cases} E^{\mathrm{rep}}, & \text{for } r_{ij} < R^{\mathrm{rep}}_{ij} \\ \epsilon_{ij}, & \text{for } R^{\mathrm{rep}}_{ij} < r_{ij} < R^{\mathrm{con}}_{ij}, \text{ and } \epsilon_{ij} > 0 \\ f\epsilon_{ij}, & \text{for } R^{\mathrm{rep}}_{ij} < r_{ij} < R^{\mathrm{con}}_{ij}, \text{ and } \epsilon_{ij} < 0 \\ 0 & \text{otherwise} \end{cases} \tag{3a}$$

where: r_{ij} is the distance between the side-chain centers of mass, R^{rep}_{ij} and R^{con}_{ij} are the cut-off values

for hard core excluded volume interactions and for square-well, soft pairwise interactions, respectively. The amino acid pairwise specific parameters, ϵ_{ij}, are different from those used previously (Kolinski & Skolnick, 1994b). The present version of the potential has a more rigorous definition of the random reference state. Here, we note that the updated version of the potential performs better in inverse folding tests both with and without gaps. A detailed discussion of the validation of this new potential is described elsewhere (Skolnick *et al.*, 1996). Here, however, to facilitate assembly, the magnitude of this interaction for a repulsive pair of residues is decreased by a factor of 2.

E^{rep} is the repulsive energy associated with the overlap of α-carbon atoms and side-chains. It is of the order of $5k_BT$, which is relatively large in comparison to the magnitude of other interactions. The pairwise interactions for the nearest neighbors down the chain are neglected. f is an angular factor that weakly favors almost parallel or almost antiparallel orientations of the secondary structure elements such as occur in all-β and mixed α/β proteins.

If β or mixed α/β motif proteins are considered, then:

$$f = 1.0, \text{ for} |i-j| < 6 \qquad (3b)$$

$$2\cos^2(\mathbf{u}_i, \mathbf{u}_j)\cos^2(20°) - \cos^4(20°), \text{ otherwise}$$

The vector $\mathbf{u}_i = \mathbf{r}_{i+2} - \mathbf{r}_{i-2}$, with $\mathbf{r}_i$ the coordinate of the ith C^α, defines the backbone orientation. The maximum strength of the attractive side-chain interactions is for local conformations that are either parallel or antiparallel. If a helical protein is being considered, then $f = 1$.

The resulting total pair energy is given by:

$$E_{\text{pair}} = \Sigma\Sigma E_{ij} \qquad (3c)$$

Finally, the model has a cooperative hydrogen-bond scheme that is very much in the spirit of Levitt & Greer (1977). The model hydrogen-bond potential operates only between α-carbon atoms. Each α-carbon atom can participate in at most two hydrogen bonds (the α-carbon atom of proline is an exception and can participate in only one hydrogen bond), and there is no directionality (donor-acceptor) in the scheme. Two residues, i and j, form a hydrogen bond when they satisfy the following distance and orientational criteria:

$$|i-j| \geqslant 3$$
$$4.6 \leqslant r_{ij} \leqslant 7.3 \text{ Å}$$
$$|(\mathbf{v}_{i-1} - \mathbf{v}_i)\cdot\mathbf{r}_{ij}| \leqslant 13.4^2 \text{ Å}^2$$
$$|(\mathbf{v}_{j-1} - \mathbf{v}_j)\cdot\mathbf{r}_{ij}| \leqslant 13.4^2 \text{ Å}^2 \qquad (4a)$$

where $\mathbf{r}_{ij}$ is the vector connecting the α-carbon atoms of residues i and j, and the v_k ($k = i-1, i, j-1, j$) are the corresponding bond vectors, see above. This scheme reproduces most (about 90%) of the main-chain hydrogen bonds assigned to the structure by DSSP Kabsch-Sander algorithm (Kabsch & Sander, 1983). An explicit cooperativity is introduced into the hydrogen-bond scheme. When two neighboring pairs of residues are hydrogen bonded, then the system gets an additional favorable energy (cooperativity). Note that we ignore all side-chain–side-chain and side-chain–backbone hydrogen bonding.

The total hydrogen bond contribution to the potential is:

$$E_{\text{H-bond}} = \Sigma\Sigma\{E^{\text{H}}\delta(i,j) + E^{\text{HH}}\delta(i,j)\delta(i\pm1, j\pm1)\}\omega(\text{A}_i, nc(i)) \qquad (4b)$$

where $\delta(i,j)$ is equal to one (zero) when residues i and j are (not) hydrogen bonded. The values of energetical parameters $E^{\text{H}} = -0.75$ and $E^{\text{HH}} = -1.0$ were adjusted so that the secondary structure in the denatured states was marginal and the folded states had a well-defined network of hydrogen bonds. $\omega(\text{A}_i, nc(i))$ adjusts the strength of a hydrogen bond depending on whether the residue is hydrophobic or hydrophilic. For hydrophobic residues, it depends also on whether the residue is buried or exposed. $nc(i)$ is the number of side-chain contacts made by residue i. In particular, if residue i is hydrophilic, then:

$$\omega(\text{A}_i, nc(i)) = 0.5 \qquad (4c)$$

For hydrophobic residues:

$$\omega(\text{A}_i, nc(i)) = \min(1, q(nc(i)) \qquad (4d)$$

where:

$$q(nc) = 2^*nc^2/(nc + m(\text{A}_i))^2 \qquad (4e)$$

$m(\text{A}_i)$ is the side-chain contact threshold that determines whether a residue is buried or exposed. The following are considered to be hydrophobic residues: G, A, C, V, I, M, L, F, Y and W, and their associated contact thresholds are 1, 2, 3, 3, 4, 5, 4, 4, 5 and 5, respectively.

Restraint contributions

The present work is designed to test the basic methodology. Certainly, experimentally obtained NMR restraints could be used and translated into more rigorous geometrical criteria than those used here. Here, we opt for tests that incorporate as little information as possible, assuming that the addition of more specific restraints would only improve the model's predictive power. Results from real NMR (but limited in number) restraints will be addressed in future work.

In what follows, we have implemented two kinds of geometrical restraints that simulate those obtained experimentally. We assume that there is some knowledge of the regular elements of secondary structure and a limited number of tertiary restraints. The implementation of each kind of restraint is discussed in turn.

Short-range restraints. Since the secondary structure of the protein of interest is usually identified

in the early stages of NMR structural determination, we assume that the regular fragments of secondary structure are known, in the sense that a region of the structure can be described as helical or β-strand or turn. We further assume that the pattern of hydrogen bonds within the β-sheets is unknown; but if such assignments have been made, they could readily be incorporated as restraints. Such local structural information could either be provided by experiment or could be assigned from theoretical considerations (Kolinski *et al*., 1996; Rost & Sander, 1993)

In particular, the short-range restraints are encoded as follows: (1) For those residues assigned to be helical, the algorithm does not allow hydrogen bonds beyond the fifth residue along the chain. Similarly, a β-assigned residue cannot have a helix hydrogen-bond pattern nor can it hydrogen bond to a residue that has been assigned to be in a helical region of the molecule. (2) A given residue can be in one of five conformational states labeled 0 to 5. These conformational states are assigned on the basis of the value $r^2_{i-1,i+2}$, i.e. on the basis of local chain geometry. Coil is assigned to state 1. A right-handed helix is assigned to state 2. A turn is assigned to state 3. A β/extended conformation is assigned to state 4. An unknown conformational state is defined as 5. For residues that have an assigned secondary structural type, energetic biases for the various allowed conformational states are assigned. For left-handed helical states, i.e. states of type 0, all residues experience a repulsion of 1, in k_BT units.

For the conformational state of the kth residue, all $\mathrm{sec}(k, j) = 0$, unless:

(a) For consecutive residues that are β-coil or coil-β:

$$\mathrm{sec}(k, 2) = 1 \tag{5a}$$

$$\mathrm{sec}(k, 0) = 1 \tag{5b}$$

(b) For consecutive pairs of residues that are extended:

$$\mathrm{sec}(k, 4) = -1 \tag{5c}$$

$$\mathrm{sec}(k, 0) = 1 \tag{5d}$$

(c) For helix-coil or coil-helix states:

$$\mathrm{sec}(k, 0) = 1 \tag{5e}$$

$$\mathrm{sec}(k, 4) = 0 \tag{5f}$$

(d) For consecutive pairs of helical states:

$$\mathrm{sec}(k, 2) = -1 \tag{5g}$$

$$\mathrm{sec}(k, 0) = 1 \tag{5h}$$

(e) For coil-coil states:

$$\mathrm{sec}(k, 4) = -1 \tag{5i}$$

$$\mathrm{sec}(k, 3) = -0.5 \tag{5j}$$

$$\mathrm{sec}(k, 0) = 1 \tag{5k}$$

(f) For consecutive pairs of turn states:

$$\mathrm{sec}(k, 3) = -0.5 \tag{5l}$$

$$\mathrm{sec}(k, 2) = -1 \tag{5m}$$

Thus, turns are encoded on a generic basis, i.e. their chirality is not specified. Rather they behave as flexible joints between regular secondary structural elements. In the above, all energies are in k_BT units.

The resulting background target local conformational energy is:

$$E_{\mathrm{target},14} = \sum_{k=2}^{N-2} \left\{ 2.5\,\mu(k) \sum_{j=1}^{5} \mathrm{sec}(k, j) + (1 - \mu(k)) \sum_{j=1}^{6} ag(k, j) \right\} \tag{6}$$

where $\mu(k) = 1$ if the secondary structure is assigned *a priori*, and it is equal to zero otherwise. $ag(k, j)$ is an amino acid pair specific analog of equation (1), that depends on the identity of A_k and A_{k+1}. In contrast to $\mathrm{sec}(k,j)$, it is defined with respect to the six conformational bins, left and right-handed extend/β states, left and right-handed wide turns, and left and right-handed helices/tight turns. It depends on both $r^{2*}_{i-2,i+1}$ and $r^{2*}_{i-1,i+2}$ (see equation (1)) and acts to propagate secondary structural elements (Olszewski *et al*., 1996a).

Long-range restraints. In addition to the aforementioned restrictions superimposed on the hydrogen-bond network, long-range restraints operate on the level of distances between the centers of mass of the side-groups. Due to the fuzzy representation of the side-chains in our model and the ambiguity in NMR restraints as applied to our side-chains, long-range restraints are implemented as follows.

If residues i and j are predicted to be in contact, then the residue-based pair potential of equation (3) is modified so that $\epsilon_{ij} = -1.25$. The long-range tertiary restraints are as follows. Let $d_{ij} = r_{ij} - R^{\mathrm{con}}_{ij}$:

$$\begin{aligned} \Psi_{\mathrm{long}}(r_{ij}) &= 0 && \text{if } r_{ij} < R^{\mathrm{con}}_{ij} \\ &= \gamma d^2_{ij} && \text{If } R^{\mathrm{con}}_{ij} < d_{ij} < 34.5 \text{ Å} \\ &= \gamma(34.5)^2 && \text{Otherwise} \end{aligned} \tag{7a}$$

Typically, the value of γ ranges from 0.5 to 2.

In folding from random compact states, the restraints are not implemented simultaneously. Rather, they start from the central pair of secondary elements, starting at residue n_{low} and ending at residue n_{upp}. For each target region where the restraints are operative, the prefactor of the $\mathrm{sec}(k, j)$ term in equation (6) is augmented from 2.5 to 3.25. Once 80% of the target contacts are satisfied between n_{low} and n_{upp}, then restraints involving the next set of secondary elements are implemented. This appears to increase the folding efficiency by decreasing the extent of kinetic trapping. Thus:

$$V_{\mathrm{long}} = \Sigma\Sigma\psi(r_{ij}) \tag{7b}$$

where the sum is over all target tertiary restraints located between n_{low} and n_{upp}.

The total energy is given by:

$$E = 0.5E_{14} + 1.5E_1 + 2.75E_{pair} + E_{H\text{-}bond} + E_{target,14} + V_{long} \quad (8)$$

Finally, for a limited set of proteins, the effect of incorporating knowledge-based rules, which are further detailed in the Appendix, were explored. For example, consecutive βαβ elements have a strong tendency to be right-handed (Chothia & Finkelstein, 1990), while consecutive ββα elements in small proteins have a very strong tendency to be left-handed (Kajava, 1992). This knowledge-based information, while not necessary to fold from a limited number of restraints, acts to reduce the number of misfolded structures. In particular, it can help to eliminate the problem of topological mirror-image states, and thereby enhances the folding efficiency. By a topological mirror image, we refer to structures where the chirality of the secondary structure connections is reversed. (Pastore *et al*., 1991) Helices, when present, are right-handed in both the correct and mirror-image topology.

It should be pointed out that results of comparable resolution have been obtained using an alternative version of the model that incorporates explicit amide hydrogen atoms and backbone hydrogen atoms, and a different realization of the long-range restraints (Kolinski & Skolnick, 1996). This suggests that the results reported here are robust and quite insensitive to the details of implementation.

Folding protocol

A good simulation strategy is to perform the folding simulation in two stages. In the first stage, a simulated thermal annealing protocol is followed. Swollen, random coil fully extended chains, uncorrelated with the native conformation, are selected as initial structures. Each simulation starts at a reduced temperature in the range of 2 to 3.5, then the temperature is slowly lowered to between 1.9 and 1.8. Each of the final structures then serves as the input for an isothermal stability run at a lower reduced temperature, typically set to 1.4. Typically, three kinds of structures result. There are misfolded states of substantially higher energy that can be trivially dismissed. The remaining structures fall into two categories, the correct, native-like folds with various root RMSD of the α-carbon atoms from native and the topological, mirror-image folds. These pseudo mirror-image folds could be identified in two ways. First, they may exhibit an *a priori* violation of the known connectivity handedness rules for some supersecondary connections. In a number of cases, simulations are run with the knowledge-based rules included in the energy term, which aids in the rejection of such structures. More generally, such pseudo mirror-image folds could simply be rejected *a priori*. While this is not done here, this possibility should be kept in mind. Second, the average energies over all simulations for the mirror-image structures are higher than that for the correct folds. In addition, for a given simulation, the lowest minimum and average energies are observed for the native as opposed to the mirror-image topology. However, especially when knowledge-based rules are not included or none exist (as in helical proteins), in some cases, such pseudo mirror-image structures have an energy spectrum that partially overlaps with that of the native conformation. Since the number of independent simulations that are computationally practical ranges from 10 to about 50, conformational sampling has not converged. As such, we cannot unambiguously preclude the possibility that if additional simulations were performed, the relative ordering of the structures could change. However, as a practical matter, when multiple topologies are found, the predicted structure is that with the lowest average and minimum energies.

Results

Folding experiments for seven proteins of known tertiary structure were performed with a set of short-range restraints that are consistent with the DSSP assignment of their secondary structure. The long-range restraints (between residues that are separated by at least four other residues along the chain) were selected randomly from the set of binary side-chain contacts observed in the native state.

Folding of 1ctf

Table 1A illustrates the distribution of a sample set of long-range restraints along the chain used in this work for the case of 1ctf (Leijonmarck & Liljas, 1987). The secondary structure assignments of the restrained residues are given for easy reference. The restraints operate between the centers of mass of the side-chains according to the definition given in the previous section.

Preliminary folding simulations were performed without the use of knowledge-based rules. Roughly 10% of these simulations produced the native fold; all these structures had the lowest average energy. To further enhance conformational sampling, a series of 46 folding and refinement runs were undertaken where knowledge-based rules for the handedness of βαβ fragments were encoded into the folding algorithm. Again, such rules are not required, but do enhance the yield of native-like structures. As shown in Table 1B, in 25 cases the native topology is recovered. In five cases, the β strands were arranged sequentially. That is, the C-terminal and N-terminal strands were not adjacent; we refer to this fold as the 123 topology. This topology was observed in runs 4, 6, 13, 18 and 41. Based on their lower average energy, runs 4 and 41 were subjected to additional refinement tests. To

Table 1. 1ctf data

A. *Representative set of restraints used in the folding of 1ctf*

Restraint	Identity of first partner	Identity of second partner
1	2 Phe (β 1)	48 Lys (helix 3)
2	4 Val (β 1)	51 Ala (helix 3)
3	6 Leu (β 1)	34 Leu (helix 2)
4	9 Ala (turn 1)	16 Val (helix 1)
5	13 Lys (helix 1)	35 Val (helix 2)
6	20 Val (helix 1)	31 Ala (helix 2)
7	20 Val (helix 1)	42 Leu (β 2)
8	31 Ala (helix 2)	40 Ala (β 2)
9	42 Leu (helix 2)	58 Leu (helix 3)
10	58 Leu (helix 3)	65 Val (β 3)

B. *Simulation results for 1ctf*

Run no.[a]	E_{last}	E_{min}[c]	$\langle E \rangle$[d]	r_{sat}/r_{tot}[e]	⟨cRMSD⟩[f]	⟨dRMSD⟩[g]	Topology
1a	–257.7	–302.4	–275.3	7/10	6.87	4.24	Misfold
2a	–252.4	–291.9	–263.7	7/10	10.42	5.39	Misfold
3a	–298.7	–328.9	–295.5	10/10	4.34	3.26	Native
3c	–297.5	–331.4	–297.3	7/10	4.41	3.51	Native
4a	–294.0	–325.9	–292.5	7/10	7.63	4.07	123
4a2	–328.5	–341.0	–313.4	8/10	7.51	3.90	123
4c2	–314.0	–333.1	–301.7	8/10	7.58	4.10	123
4c3	–281.7	–328.4	–299.7	8/10	7.83	4.22	123
5a	–314.6	–329.0	–288.9	8/10	3.79	3.08	Native
5a2	–307.6	–335.2	–307.4	8/10	3.99	3.13	Native
5c2	–312.9	–337.4	–308.3	10/10	3.96	3.04	Native
5d	–325.6	–343.0	–314.4	10/10	3.95	3.02	Native
5e	**–313.9**	**–343.8**	**–319.4**	**10/10**	**3.90**	**2.96**	**Native**
6a	–286.4	–312.4	–281.6	9/10	7.84	4.46	123
7a	–308.2	–329.0	–300.0	9/10	3.91	3.21	Native
7c	–311.1	–334.6	–307.3	9/10	3.91	3.25	Native
8a	–232.1	–273.1	–229.7	7/10	7.62	4.49	Misfold
9a	–258.7	–295.7	–266.7	8/10	5.12	3.72	Native
10a	–277.6	–303.1	–268.7	10/10	4.81	3.62	Native
11a	–298.3	–318.0	–289.5	8/10	4.81	3.52	Native
12a	–301.4	–316.9	–290.1	7/10	5.06	3.77	Native
13a	–291.3	–310.0	–281.4	10/10	8.89	4.58	Native
14c2	–317.2	–320.9	–302.9	7/10	4.55	3.75	Native
14c3	**–323.7**	**–346.2**	**–317.7**	**8/10**	**4.23**	**3.44**	**Native**
14c4	–281.2	–337.7	–307.0	9/10	4.38	3.46	Native
14c5	–306.9	–338.3	–303.51	10/10	4.23	3.23	Native
10c6	–314.4	–330.1	–300.2	10/10	4.60	3.42	Native
15a	–286.0	–311.7	–282.0	7/10	6.79	3.89	Misfold
16a	–298.2	–317.2	–285.0	10/10	4.62	3.53	Native
17a	–265.3	–302.8	–269.9	9/10	8.34	3.93	Misfold
18a	–288.8	–311.8	–282.2	8/10	6.47	3.83	123
19a	–308.9	–324.4	–287.9	9/10	4.26	3.39	Native
20a	–298.0	–312.7	–284.0	7/10	6.78	3.94	Misfold
21a	–305.1	–328.0	–292.7	10/10	4.31	3.28	Native
22a	–290.5	–318.5	–288.6	8/10	4.26	3.30	Native
23a	–305.1	–320.9	–301.4	10/10	10.11	3.72	Mirror
23c	**–291.5**	**–328.3**	**–294.2**	**10/10**	**10.13**	**3.78**	**Mirror**
23d	–291.1	–325.0	–294.7	9/10	10.21	3.80	Mirror
24a	–310.3	–327.8	–296.8	8/10	4.25	3.31	Native
25a	–249.8	–277.4	–243.6	6/10	8.62	4.96	Misfold
26a	–264.6	–288.0	–253.7	7/10	7.96	4.11	Misfold
27a	–225.6	–274.4	–218.2	8/10	7.12	4.77	Misfold
28a	–324.4	–344.6	–318.5	10/10	4.12	3.14	Native
28c	–314.6	–347.8	–314.3	10/10	4.02	3.02	Native
28c2	**–311.3**	**–353.2**	**–322.4**	**9/10**	**3.81**	**2.90**	**Native**
29a	–298.9	–335.5	–299.4	10/10	3.81	2.92	Native
30a	–286.7	–311.9	–280.5	8/10	9.89	4.15	Misfold
31a	–310.2	–328.3	–293.1	10/10	4.76	3.64	Native
32a	–317.0	–332.1	–307.3	9/10	4.22	3.13	Native
32c	–309.9	–338.3	–303.6	10/10	4.42	3.22	Native
32c2	–298.8	–326.3	–300.8	10/10	4.99	3.62	Native
33a	–286.8	–319.1	–289.2	8/10	8.00	4.23	Misfold
34a	–260.3	–283.5	–250.9	7/10	9.25	4.23	Misfold
35a	–245.5	–281.8	–248.8	9/10	11.28	5.20	Misfold
36a	–292.1	–320.4	–292.0	8/10	4.57	3.65	Native

continued overleaf

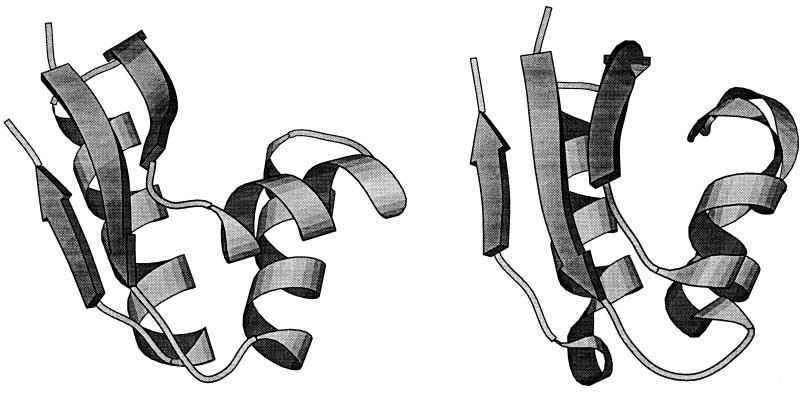
Table 1 *(continued)*

Run no.[a]	E_{last}	E_{min}[c]	$\langle E \rangle$[d]	r_{sat}/r_{tot}[e]	⟨cRMSD⟩[f]	⟨dRMSD⟩[g]	Topology
37a	–299.9	–330.0	–295.7	9/10	9.99	3.88	Mirror
38a	–300.3	–327.3	–303.1	8/10	4.52	3.25	Native
38c	–294.8	–343.7	–299.1	7/8	3.68	2.89	Native
39a	–301.1	–327.5	–303.0	8/10	3.79	3.07	Native
39c	–307.4	–331.0	–303.3	10/10	4.47	3.47	Native
40a	–322.3	–336.9	–306.5	7/10	3.90	3.11	Native
40c	–289.2	–305.7	–280.8	10/10	6.33	4.14	Native
41a	–308.0	–324.8	–301.5	8/10	7.36	4.17	Native
41c	–303.6	–332.7	–308.0	10/10	7.17	4.20	Native
42a	–301.2	–331.4	–292.6	7/10	4.94	3.58	Native
43a	–317.1	–337.9	–310.8	9/10	4.34	3.45	Native
44a	–287.7	–310.8	–281.8	9/10	7.54	4.01	Native
46a	–269.3	–295.9	–263.0	7/10	9.00	5.22	Misfold
46a	–289.0	–315.4	–283.6	7/10	10.07	3.82	Mirror

[a] A series of simulated annealing runs from a reduced temperature of 2.4 to 1.8 was done. Runs designated by an a, e.g. 1a, were refined at a reduced, temperature of 1.4 with $\gamma = 2.0$. Runs indicated by an a2 were preceded by a simulated annealing cycle over a temperature range of 1.9 to 1.4 with $\gamma = 2$. Runs of type c and higher followed the same protocol but with $\gamma = 10$. Runs indicated by a d or an e omit the simulated annealing step and are twice as long.
[b] E_{last} is the energy of the final conformation.
[c] E_{min} is the minimum energy observed during the course of the simulation.
[d] $\langle E \rangle$ is the average energy of the given simulation.
[e] Ratio of the restraints satisfied, r_{sat}, in the final structure obtained at the end of the simulation to the total number of restraints, r_{tot}.
[f] The coordinate RMSD is calculated with respect to the α-carbon atoms.
[g] The distance RMSD is calculated with respect to the α-carbon atoms.

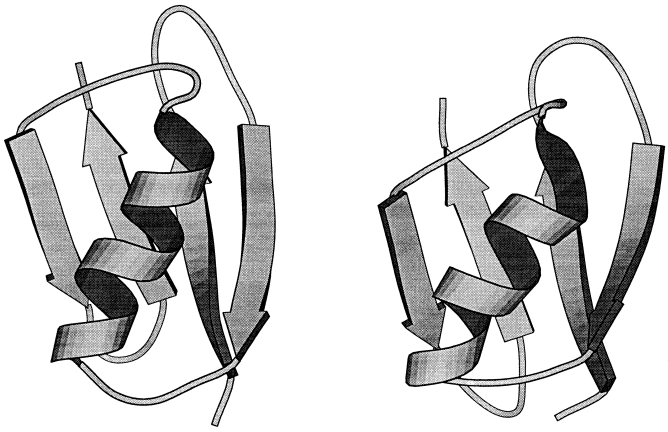
ensure the stability of the results, a series of simulations with moderate ($\gamma = 2$) and strong ($\gamma = 10$) restraints were used. For correct folds, the energy should be rather insensitive to the value of γ. However, for those cases where the fold is not particularly compatible with the restraint set, the energy should be quite sensitive to the value of γ. The resulting lowest average energy and minimum energies for the 123 topology are $-308k_BT$ and $-333.1k_BT$, respectively. In three cases, runs 23, 37 and 46, the topological mirror image is adopted. For the mirror-image topology, the corresponding average and minimum energy values, –295.7 and –330.0, respectively, are higher.

In contrast, for the native topology, the lowest average is observed for run 28c2, where the average and minimum energy values are –322.4 and $-353.2k_BT$, respectively. This run exhibits an average coordinate and distance RMSD of 3.81 Å and 2.90 Å, respectively, from the native structure. This resulting final structure is shown Figure 1(b). The next best simulation is run 5e, whose average and minimum energy values are –319.4 and $-343.8k_BT$, respectively. For this simulation, the average coordinate RMSD is 3.90 Å. This is followed by run 14c3 with an average and minimum energy of –317.7 and $-346.2k_BT$, respectively. Here, the average coordinate RMSD is 4.23 Å. Thus, on the basis of both the average and minimum energy, we conclude that the native topology is selected. The average coordinate RMSD for the three lowest-energy, independent simulations is 3.98 (±0.22) Å. Similarly, the average distance RMSD is 3.1 (±0.30) Å.



(a) **(b)**

Figure 1. (a) Fold of L7/L12 ribosomal protein, 1ctf, obtained from the X-ray structure. (b) Representative predicted conformation of 1ctf obtained using ten tertiary restraints. Figures are drawn using MOLSCRIPT (Kraulis, 1991).



(a) (b)

Figure 2. (a) Fold of protein G, 2gb1, obtained from the NMR solution structure. (b) Representative predicted conformation of 2gb1 obtained using eight tertiary restraints. Figures are drawn using MOLSCRIPT (Kraulis, 1991).

Protein G

Domain B1 of *Streptococcus* protein G, 2gb1, is a 56 residue protein that has a native conformation consisting of two β hairpins that form a single β sheet and a single helix (Gronenborn *et al*., 1991). The secondary structure is of the type ββαββ, and the native structure shown in Figure 2(a). 2gb1 has been subject to ten independent folding simulations from arbitrary random coil states in which eight long-range restraints are assumed. These restraints are summarized in Table 2A. In all cases, the simulated annealing runs were followed by lower temperature refinement with full implementation of the tertiary restraints. The results of these simulations are summarized in Table 2B. To explore its range of applicability, the knowledge-based rules described in the Appendix were used in both folding and refinement studies. The family of structures, run 10a, exhibiting lowest average energy had a coordinate RMSD of 3.81 Å and an average distance RMSD of 2.72 Å from native. The next higher energy structures had an average coordinate (distance) RMSD of 3.30 (2.37) and 5.03 (3.25) Å from runs 8a and 9a, respectively. The final structure from run 10a has an RMSD of 3.75 Å. In contrast, the final structure from run 9a has a C^{α} RMSD of 4.94 Å. This structure has a broken helix and might be tentatively dismissed on the basis that the secondary structure assigned to this region is not fully satisfied. Both structures were subjected to a second round of isothermal refinement at a reduced temperature of 1.2. In both cases, the correct, full helix is adopted. The average RMSD for runs 9b and 10b are 4.36 and 3.29 Å, respectively. On the basis of average energy, these two structures are identical, but if one uses the minimum energy (-261.3 *versus* -266.0, respectively) and number of satisfied restraints (7/8 *versus* 8/8) as the criterion to differentiate the two sets, then the lower RMSD family is chosen. The final structure from run 10b is shown in Figure 2(b).

If the lowest average energy over a simulation run is used to assign the native fold, then reasonably good structures whose RMSD from native is about 3.8 Å are predicted. A more conservative estimate would assign the native state to one of two closely related, structural families. Indeed, taking the set of five lowest energy structural families, the average coordinate RMSD is 4.17 Å, and the average distance RMSD is 2.77 Å from native. Thus, structures whose RMSD is better than about 5 Å are consistently predicted for 2gb1 using knowledge of the secondary structure and eight long-range restraints.

Thioredoxin

Thioredoxin (2trx) is a 108 residue mixed motif protein (Katti *et al*., 1990) whose folding has been examined also by Aszodi *et al*. (1995). The native conformation is shown in Figure 3(a). The present simulations for thioredoxin employed a set of 30 restraints; these are listed in Table 3A. A series of seven independent topology assembly simulations were performed without the use of knowledge-based rules. As summarized in Table 3B, three simulations yielded the native topology. Two simulations produced what is essentially the mirror-image topology, and two were completely misfolded structures. Once topology assembly had occurred, explicit knowledge-based rules were implemented. It is very easy to differentiate the native structure from its topological mirror image. Next, a series of runs designated by a in Table 4 using the knowledge-based rules detailed in the Appendix were undertaken. The native fold (from run 2a) and the topological mirror-image structure (from run 6a), whose distance RMSD was closest to native, were then subjected to a simulated annealing run over a reduced temperature range of 1.6 to 1.3. This was followed by an isothermal run at a reduced temperature of 1.3. In the latter two series of runs, no supersecondary structure knowledge-based rule was employed. Again, it is evident that the native fold is strongly preferred on an energetic basis, even when the knowledge-based rules are absent. The resulting structures had an

average coordinate RMSD from native of 3.41 Å and an average distance RMSD from native of 2.61 Å. The structure at the end of the simulation satisfied 26 of the 30 long-range contact restraints. This conformation is shown in Figure 3(b). In contrast, in the Aszodi *et al*. (1995) approach, half of the models with 30 restraints had an RMSD around 4.5 Å, while the other set of models had an RMSD up to about 9 Å. An advantage of the current approach is that energetic criteria can be used to identify acceptable models, and this is profitably exploited in the thioredoxin case.

Flavodoxin

Flavodoxin (3fxn) is a 138 residue α/β protein (Smith *et al*., 1977), whose native conformation is shown in Figure 4(a). In all cases, we undertook folding and refinement without using any knowledge-based rules, but we did assume general knowledge of the regular fragments of secondary structure and 35 long-range restraints; the latter are listed in Table 4A. To investigate the effect of the relative strength of the long-range restraints, we undertook a series of simulations with $\gamma = 2.0$, 1.0 and 0.5 (see equation (7a)). In all cases, we did not employ any knowledge-based rule in the calculation of the energy. Table 4B to D summarize the results. In the case of Table 4B and C, based on energy alone, the native topology is overwhelmingly favored over the mirror-image fold and these incorrect structures can be eliminated on the basis of energy.

It is only in the case of $\gamma = 2$ that some ambiguity results, with the one mirror-image topology having an energy comparable to that of some of the higher-energy native folds. If we assign the native fold on the basis of the set of structures having the

Table 2. Protein G data

A. *Representative set of restraints used in folding of protein G*

Restraint	Identity of first partner	Identity of second partner
1	3 Tyr (β 1)	50 Lys (turn 3)
2	5 Leu (β 1)	16 Thr (β 2)
3	8 Asn (β 1)	55 Thr (β 4)
4	18 Thr (β 2)	30 Phe (helix 1)
5	23 Ala (helix 1)	45 Tyr (β 3)
6	26 Ala (helix 1)	52 Val (β 4)
7	45 Tyr (β 3)	31 Ala (helix 2)
8	39 Val (turn 3)	54 Val (β 4)

B. *Simulation results for the B1 domain of protein G*

Run no.[a]	E_{last}[b]	E_{min}[c]	$\langle E \rangle$[d]	r_{sat}/r_{tot}[e]	⟨cRMSD⟩[f]	⟨dRMSD⟩[g]	Topology
1	–162.2	–205.8	–106.0	7/8	5.2	3.44	Native
1a	–287.8	–305.0	–245.6	8/8	3.67	2.49	Native
2	–165.5	–184.5	–98.0	6/8	6.81	3.92	Native
2a	–264.4	–283.3	–231.0	7/8	5.17	3.27	Native
3	–178.1	–200.1	–103.8	8/8	5.45	3.6	Native
3a	–287.8	–297.7	–249.6	7/8	5.06	3.01	Native
4	–127.6	–187.9	–100.0	4/8	5.22	3.57	Native
4a	–260.8	–272.2	–222.8	7/8	6.31	4.46	Native
5	–156.9	–184.9	–93.6	6/8	5.90	3.85	Native
5a	–253.7	–269.3	–223.8	6/8	5.92	3.75	Native
6	–160.9	–196.2	–100.1	6/8	5.81	3.72	Native
6a	–264.8	–291.5	–238.8	7/8	6.93	3.89	Native
7	–174.9	–202.9	–99.77	8/8	6.0	3.00	Native
7a	–263.2	–281.3	–240.8	8/8	4.06	2.88	Native
8	–201.6	–203.4	–98.0	7/8	5.70	3.77	Native
8a	–283.9	–310.1	–259.8	8/8	3.30	2.37	Native
9	–188.1	–199.0	–102.9	8/8	5.43	3.62	Native
9a	–303.6	–315.1	–259.9	8/8	5.03	3.25	Native
9b	**–247.6**	**–261.3**	**–226.6**	**7/8**	4.36	3.22	Native
10	–199.8	–228.4	–171.7	8/8	3.93	2.82	Native
10a	–296.6	–311.4	–262.9	8/8	3.81	2.72	Native
10b	**–247.2**	**–266.0**	**–227.1**	**8/8**	**3.29**	**2.62**	Native

[a] A series of simulated annealing runs from a reduced temperature of 2.5 to 1.8 was done, with the exception of run 10, which covered a temperature range of 1.8 to 1.5. Runs designated by an a, e.g. 1a started at a reduced temperature of 1.2 and ended at a reduced temperature of 1.0. Run 10a was 2.5 times longer than the previous refinement runs. All runs included the knowledge-based rules described in the Appendix. Runs designated by a b were isothermal at $T = 1.2$ and used $\gamma = 0.5$.

[b] E_{last} is the energy of the final conformation.

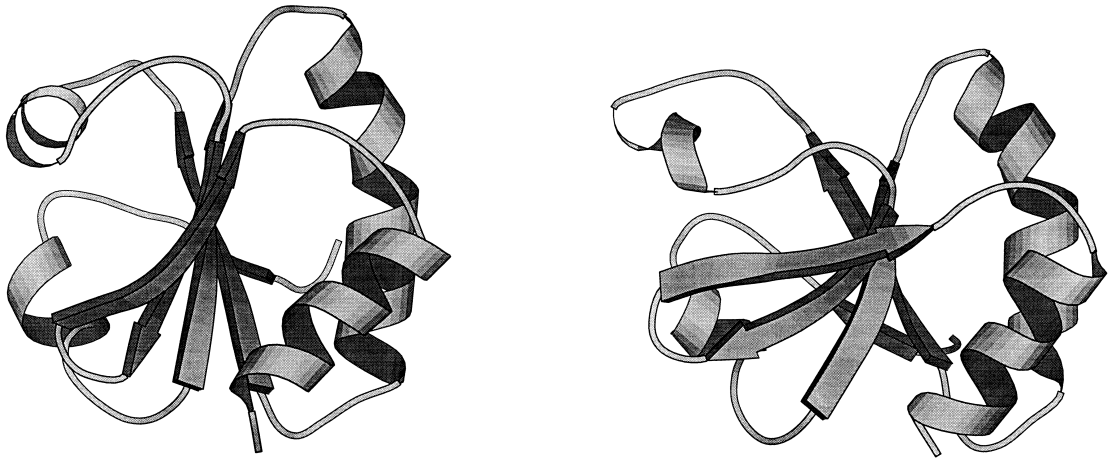
[c] E_{min} is the minimum energy observed during the course of the simulation.

[d] $\langle E \rangle$ is the average energy of the given simulation.

[e] Ratio of the restraints satisfied, r_{sat}, in the final structure obtained at the end of the simulation to the total number of restraints, r_{tot}.

[f] The coordinate RMSD is calculated with respect to the α-carbon atoms.

[g] The distance RMSD is calculated with respect to the α-carbon atoms.



(a) **(b)**

Figure 3. (a) Fold of thioredoxin, 2trx, obtained from the X-ray structure. (b) Representative predicted conformation of 2trx obtained using 25 tertiary restraints. Figures are drawn using MOLSCRIPT (Kraulis, 1991).

lowest average and minimum energy, the native topology again is selected with values of -659.2 and $-712.6k_BT$, respectively. In contrast, the mirror-image topology has average and minimum energies of -577.5 and $-643.0k_BT$. In Figure 4(a), we plot the energy *versus* RMSD for the set of pooled conformations listed in Table 4C and D. Not surprisingly, a portion of the energy spectra of the

Table 3. Thioredoxin data

A. *Restraint list for thioredoxin*

4-46	23-81	30-36	53-107
4-56	24-99	38-76	60-75
5-15	26-42	38-94	70-81
8-63	27-79	41-99	76-93
12-66	28-35	45-103	80-107
13-18	28-75	46-55	81-86
16-81	29-59	46-104	90-102
		49-103	94-99

B. *Isothermal simulation results for thioredoxin*

Run no.[a]	E_{last}[b]	E_{min}[c]	$\langle E \rangle$[d]	r_{sat}/r_{tot}[e]	⟨cRMSD⟩[f]	⟨dRMSD⟩[g]	Topology
1a	–514.7	–557.6	–416.8	19/30	7.64	4.87	Native
2a	–577.8	–620.8	–492.0	24/30	4.69	3.34	Native
2b	–511.0	–571.5	–464.7	27/30	3.44	2.61	Native
2c	**–518.1**	**–588.1**	**–552.2**	**26/30**	**3.41**	**2.61**	**Native**
3a	–490.6	–528.4	–411.7	10/30	11.94	5.77	Misfold
4a	–370.2	–388.1	–297.6	13/30	12.36	6.34	Misfold
5a	–512.6	–558.0	–457.5	26/30	4.16	3.09	Native
6a	–476.7	–597.2	–375.2	22/30	11.80	3.75	Mirror
6b	–261.7	–327.6	–287.0	18/28	11.16	3.78	Mirror
6c	**–270.1**	**–329.8**	**–290.6**	**19/30**	**12.24**	**6.09**	**Mirror**
7a	–454.6	–500.7	–385.2	19/30	9.51	5.31	Mirror
7a	–297.9	–330.8	–243.9	21/30	12.34	6.11	Mirror

[a] A series of simulated annealing runs from a reduced temperature of 2.2 to 1.9 was undertaken. Knowledge-based energetic biases were not used in the assembly portion of these simulations. All runs designated by an a, e.g. 1a started at a reduced temperature of 1.4 and finished at a reduced temperature of 1.0. These runs had knowledge-based rules for βαβ and ββα-type states. Runs designated by a b started from the final configuration of an a type run and covered a temperature range of 1.6 to 1.3. Isothermal runs are designated by a c, and started from the minimum energy configuration of a b type run and were done at a reduced temperature of 1.3. No knowledge-based rule is implemented in b or c-type runs.

[b] E_{last} is the energy of the final conformation.

[c] E_{min} is the minimum energy observed during the course of the simulation.

[d] $\langle E \rangle$ is the average energy of the given simulation.

[e] Ratio of the restraints satisfied to the total number of restraints.

[f] The coordinate RMSD is calculated with respect to the α-carbon atoms.

[g] The distance RMSD is calculated with respect to the α-carbon atoms.

Table 4. Flavodoxin data

A. *Restraint list for flavodoxin*

3-16	11-87	19-50	48-83	81-111
3-32	11-120	19-85	50-135	81-138
3-50	12-54	19-132	51-99	84-99
3-85	15-85	23-135	53-61	87-117
4-69	15-115	28-136	53-95	100-114
5-13	15-128	33-38	61-70	111-138
6-56	17-32	38-69	61-99	117-124

B. *Isothermal refinement simulation results for flavodoxin: $\gamma = 0.5$ for both folding and refinement*

Run no.[a]	E_{last}[b]	E_{min}[c]	$\langle E \rangle$[d]	r_{sat}/r_{tot}[e]	$\langle$cRMSD$\rangle$[f]	$\langle$dRMSD$\rangle$[g]	$\langle$dRMSD$\rangle$[g]
1a	–616.4	–654.4	–590.7	29/35	4.38	3.18	Native
2a	–651.4	–676.2	–630.5	28/35	3.81	2.73	Native
3a	–650.5	–690.5	–633.0	30/35	3.87	2.90	Native
4a	–375.6	–435.1	–340.6	16/35	11.99	6.71	Misfold
5a	–556.7	–605.1	–548.1	26/35	4.62	3.42	Native
6a	–517.6	–554.9	–497.0	20/35	8.67	5.13	Misfold
7a	–650.7	–699.3	–643.7	29/35	4.40	3.10	Native
8a	–622.5	–663.3	–606.6	29/35	4.18	3.06	Native
9a	–578.2	–643.9	–585.5	26/35	4.30	3.33	Native
10a	–615.3	–654.1	–594.6	27/35	4.75	3.50	Native
11a	–581.1	–645.9	–572.3	29/35	4.26	3.08	Native
12a	–284.8	–324.0	–255.4	13/35	12.05	6.27	Misfold
13a	–352.0	–410.1	–335.2	10/35	11.17	5.30	Misfold
14a	–592.7	–642.5	–587.5	23/35	4.75	3.46	Native

[a] Each of a series of simulated annealing runs was performed over a reduced temperature range of 3.0 to 1.9. During these runs, $\gamma = 0.5$. Runs a were done at a reduced temperature of 1.4, with $\gamma = 0.5$. No knowledge-based rule is implemented.
[b] E_{last} is the energy of the final conformation.
[c] E_{min} is the minimum energy observed during the course of the simulation.
[d] $\langle E \rangle$ is the average energy of the given simulation.
[e] Ratio of the restraints satisfied, r_{sat}, to the total number of restraints, r_{tot}.
[f] The coordinate RMSD is calculated with respect to the α-carbon atoms.
[g] The distance RMSD is calculated with respect to the α-carbon atoms.

C. *Isothermal refinement simulation results for flavodoxin: $\gamma = 1.0$ for folding and $\gamma = 0.5$ for refinement*

Run no.[a]	E_{last}[b]	E_{min}[c]	$\langle E \rangle$[d]	r_{sat}/r_{tot}[e]	$\langle$cRMSD$\rangle$[f]	$\langle$dRMSD$\rangle$[g]	Topology
1a	–630.0	–662.9	–602.1	28/35	4.92	3.40	Native
2a	–318.2	–392.2	–291.8	16/35	11.75	6.42	Misfold
3a	–547.3	–566.6	–506.9	18/35	13.83	4.48	Mirror
4a	–502.5	–566.5	–508.2	24/35	13.34	3.67	Mirror
5a	–673.7	–703.9	–648.8	31/35	3.93	2.83	Native
6a	–643.2	–696.6	–627.8	27/35	4.10	3.05	Native
7a	–523.1	–563.9	–481.3	23/35	13.93	4.45	Mirror
8a	–642.9	–667.9	–621.9	28/35	3.62	2.72	Native

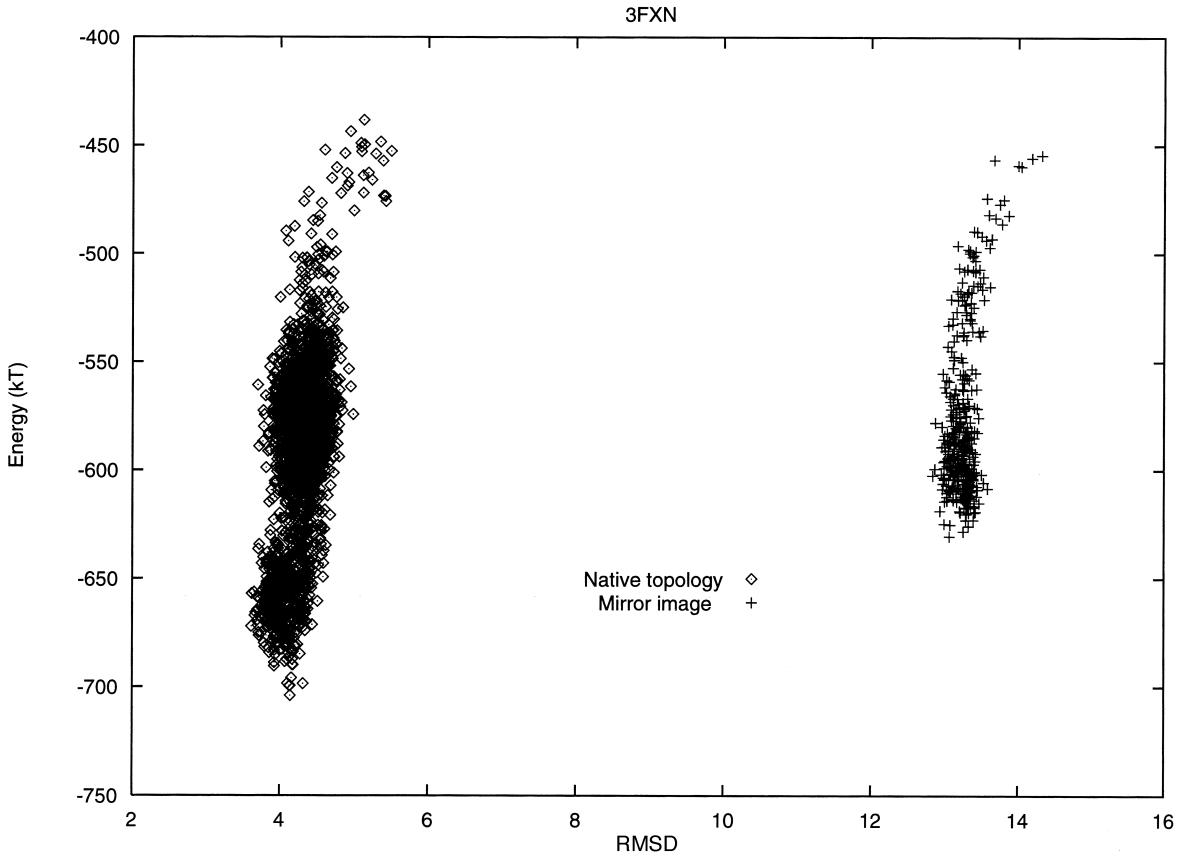
[a] Each of a series of simulated annealing runs was performed over a reduced temperature range of 3.0 to 1.9. During these runs, $\gamma = 1.0$. Runs a were done at a reduced temperature of 1.4, with $\gamma = 0.5$. No knowledge-based rule is implemented.
[b] E_{last} is the energy of the final conformation.
[c] E_{min} is the minimum energy observed during the course of the simulation.
[d] $\langle E \rangle$ is the average energy of the given simulation.
[e] Ratio of the restraints satisfied, r_{sat}, to the total number of restraints, r_{tot}.
[f] The coordinate RMSD is calculated with respect to the α-carbon atoms.
[g] The distance RMSD is calculated with respect to the α-carbon atoms.

D. *Isothermal refinement simulation results for flavodoxin: $\gamma = 2.0$ for folding and $\gamma = 0.5$ for refinement*

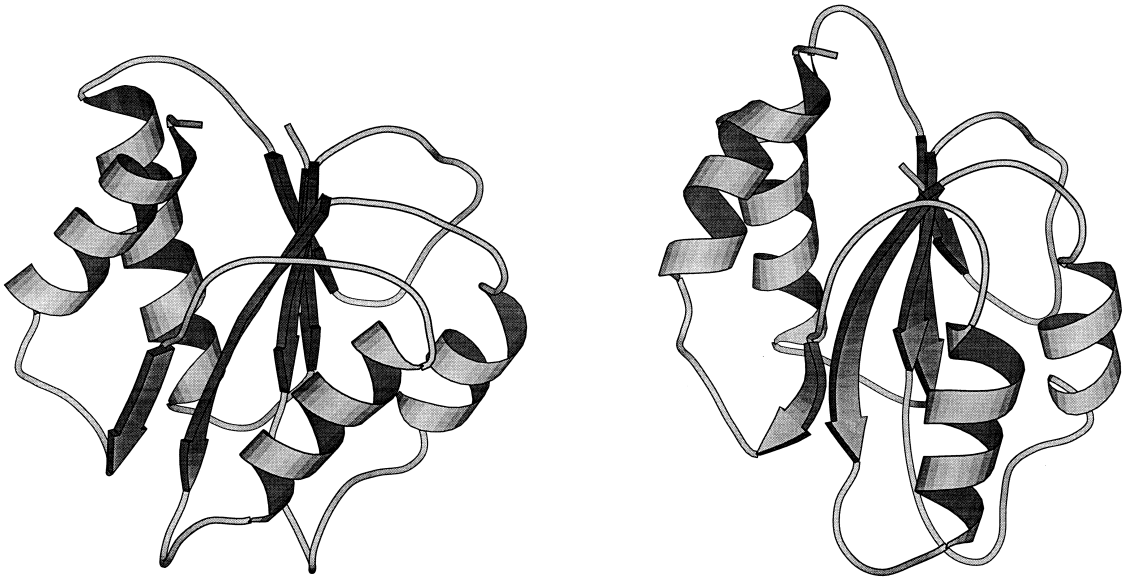
Run no.[a]	E_{last}[b]	E_{min}[c]	$\langle E \rangle$[d]	r_{sat}/r_{tot}[e]	$\langle$cRMSD$\rangle$[f]	$\langle$dRMSD$\rangle$[g]	Topology
1a	–599.4	–642.4	–586.1	26/35	4.27	3.13	Native
2a	–651.3	–682.0	–624.5	26/35	4.29	2.98	Native
3a	–599.7	–626.0	–563.8	28/35	4.46	3.35	Native
4a	–591.5	–641.9	–580.9	25/35	4.45	3.19	Native
5a	–608.4	–643.0	–577.5	26/35	13.26	3.57	Mirror
6a	–599.0	–631.4	–582.3	25/35	4.27	3.16	Native
7a	–664.7	–712.6	–659.2	29/35	3.98	2.90	Native
8a	–537.1	–604.3	–552.7	26/35	4.29	3.08	Native

[a] Each of a series of simulated annealing runs was performed over a reduced temperature range of 3.0 to 1.9. During these runs, $\gamma = 2.0$. Runs a were done at a reduced temperature of 1.4, with $\gamma = 0.5$. No knowledge-based rule is implemented during the course of these simulations.
[b] E_{last} is the energy of the final conformation.
[c] E_{min} is the minimum energy observed during the course of the simulation.
[d] $\langle E \rangle$ is the average energy of the given simulation.
[e] Ratio of the restraints satisfied, r_{sat}, to the total number of restraints, r_{tot}.
[f] The coordinate RMSD is calculated with respect to the α-carbon atoms.
[g] The distance RMSD is calculated with respect to the α-carbon atoms.

two topologies overlap. The native and mirror-image topologies share many but not all structural features in common. Many of the side-chain contacts are the same, as are the secondary structures in the regions between turns. Where they differ is in the chirality of the turns and in a subset



(a)



(b) **(c)**

Figure 4. (a) Energy *versus* RMSD for the set of structures whose average properties are summarized in Table 4C and D. (b) Fold of flavodoxin, 3fxn, obtained from the X-ray structure. (c) Representative predicted conformation of 3fxn obtained using 35 tertiary restraints. Figures are drawn using MOLSCRIPT (Kraulis, 1991).

Table 5. Myoglobin data

A. *Restraint list for myoglobin*

2-78	70-91
6-129	71-112
13-116	74-137
24-68	88-140
28-108	91-105
29-46	92-141
36-107	93-144
40-54	102-142
42-104	105-138
46-63	113-130

B. *Isothermal refinement simulation results for myoglobin*

Run no.[a]	E_{last}[b]	E_{min}[c]	$\langle E \rangle$[d]	r_{sat}/r_{tot}[e]	⟨cRMSD⟩[f]	⟨dRMSD⟩[g]
02191a	–494.0	–438.7	–357.12	13/20	7.81	4.57
02191b	**–676.4**	**–723.7**	**–661.3**	**18/20**	**6.60**	**4.11**
02192a	–552.9	–462.0	–383.8	16/20	5.27	3.71
02192b	**–686.0**	**–705.0**	**–662.6**	**17/20**	**4.89**	**3.46**
02193a	–562.9	–515.2	–435.6	15/20	5.45	4.02
02193b	**–614.3**	**–716.6**	**–665.3**	**17/20**	**5.52**	**3.85**
02194a	–570.0	–542.8	–435.4	16/20	5.77	4.09
02194b	**–681.2**	**–716.4**	**–663.7**	**19/20**	**5.59**	**3.68**

[a] Each of a series of simulated annealing runs was performed over a reduced temperature of 2.5 to 2.0. Then, all runs designated by an a, e.g. 1a, started at a reduced temperature of 1.4 and finished at a reduced temperature of 1.0. No knowledge-based rule was used. Runs designated by a b started from the final configuration of an a-type run and had a reduced temperature of 1.4. The prefactor γ in equation (7a) was 0.5.
[b] E_{last} is the energy of the final conformation.
[c] E_{min} is the minimum energy observed during the course of the simulation.
[d] $\langle E \rangle$ is the average energy of the given simulation.
[e] Ratio of the restraints satisfied, r_{sat}, to the total number of restraints, r_{tot}.
[f] The coordinate RMSD is calculated with respect to the α-carbon atoms for residues 5 to 145.
[g] The distance RMSD is calculated with respect to the α-carbon atoms for residues 5 to 145.

of the side-chain contacts. If all the structures from Table 4C and D are pooled, then the native conformation has an average energy of $-592.8k_BT$ and the mirror-image topology has an average energy of $-522.6k_BT$. Thus, on this basis as well, the native fold is selected. Of course, we cannot preclude the possibility that with additional simulations, the mirror-image fold would be selected, but employing energy-based criteria to choose the topology, the native fold is preferred. Ultimately, we could choose the native over the mirror-image topology on the basis that a very strongly conserved knowledge-based rule is violated, but here we use energetic criteria that do not include such information.

Of a total of 30 simulations, nine produced either the mirror-image fold or an incorrect topology, and the remaining 21 adopted the correct fold. The conformations of lowest average and minimum energy have a coordinate and distance RMSD from native for the α-carbon atoms of 3.98 and 2.90 Å, respectively. The average distance and coordinate RMSD from native for all 21 sets of independently folding simulations of native-like 3fxn are 4.28 and 3.12 Å, respectively. These results should be compared with those of Smith-Brown *et al*. (1993). Using a total of 148 restraints, they produced structures whose backbone RMSD from native is 3.18 Å. In contrast, using just 35 restraints that have been arbitrarily chosen from the native contact map, structures whose average RMSD from native is 3.98 Å are generated. We show the predicted fold obtained at the end of run 7c of Table 4D in Figure 4(c). Rather good agreement with the native conformation is evident (Figure 4(b)). Thus, in spite of its quite substantial size, the ability to reproducibly fold an α/β protein from a fairly small number of long-range restraints has been demonstrated.

Myoglobin

Next, we considered the folding of the 146 residue, helical protein myoglobin, 1mba, using just 20 long-range restraints (Bolognesi *et al*., 1989), which are summarized in Table 5A. The native conformation of 1mba is shown in Figure 5(a). In a series of 27 independent simulations, a variety of conditions were explored. In 25 of 27 cases, the native topology was adopted. This set of structures had an average coordinate RMSD of 5.89 Å and an average distance RMSD of 3.92 Å from native. The two sets of misfolded structures could be dismissed on the basis of their average energies and the fact that they satisfied less than half of the total restraints. In Table 5B, we present representative results from a set of four independent simulations. In each case, the starting conformation was an arbitrary random-coil state. The average coordinate and distance RMSD for this set of four independent

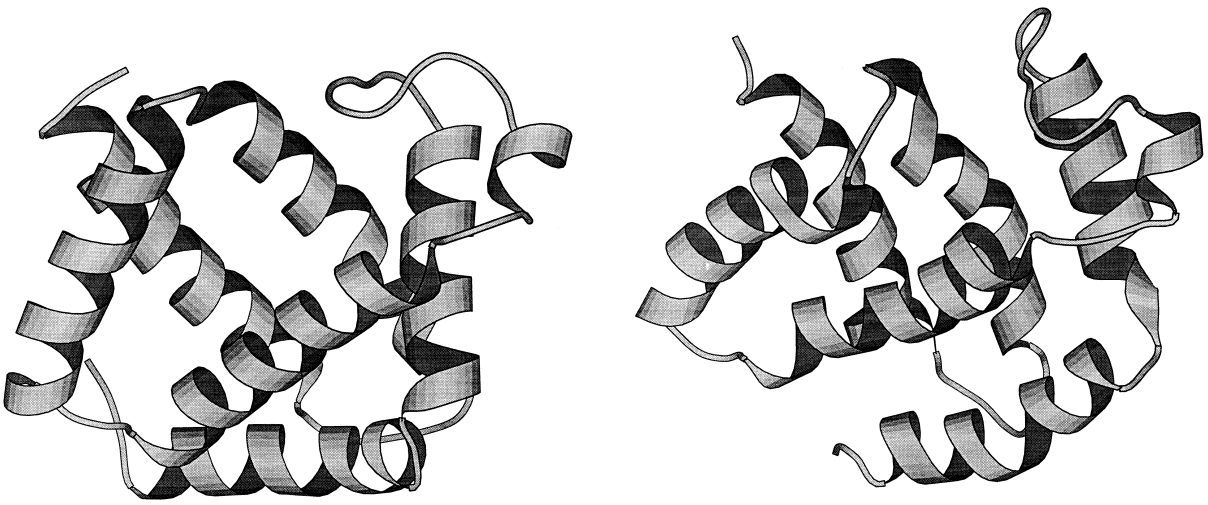
runs is 5.65 and 3.79 Å, respectively. Figure 5(b) shows a typical predicted structure. Reasonable agreement with experiment is provided. Typical of this class of models, the folding of helices gives the best results, in that one restraint every 7.3 residues, on average, gives reasonably good structures. In contrast, for the larger α/β proteins thioredoxin and flavodoxin, we found that one restraint every four residues produces models on the level of 3.4 to 4 Å RMSD.

Plastocyanin

Plastocyanin (1pcy) is a 99 residue, Greek key, β-barrel protein (Guss & Freeman, 1983), whose native structure is shown in Figure 6(a). Simulations of plastocyanin with 46 (25) restraints listed in Table 6A (C), were undertaken, and the results are summarized in Table 6B (D), respectively. For the case of 46 restraints, a series of five independent simulations were performed. Three resulted in the native fold, and two produced the topological mirror-image. In this case, on the basis of the energy, it is straightforward to select the native topology. A typical resulting structure is shown in Figure 6(b). The run with lowest average energy, run 1a, produces structures whose coordinate and distance RMSD from native are 3.51 and 2.50 Å, respectively. These structures are actually quite close to the geometric resolution of the lattice model when side-chains are included.

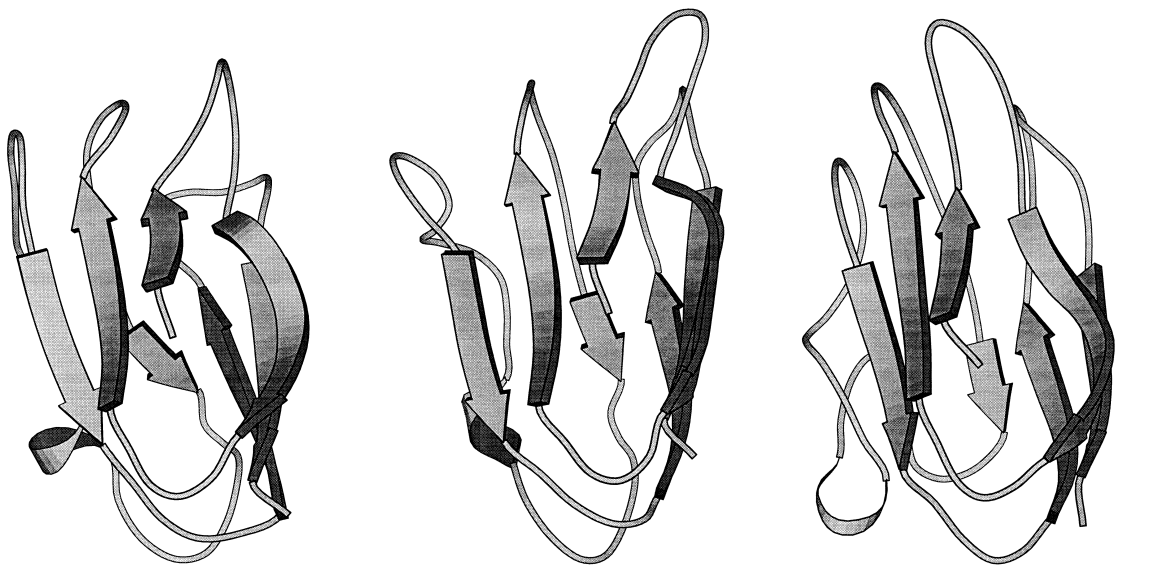
Next, we considered the situation when there are 25 long-range restraints. A series of ten independent assembly runs were undertaken, and in seven of ten cases, the native fold is recovered. In two cases, the mirror-image topology is produced, and in one case, a misfolded structure whose topology is related to the native fold is predicted. Table 6D summarizes the results based on the preliminary set of refinement runs of each of these ten structures. The native-like topology (run 3a) and mirror image topologies (run 9a) of lowest average energy were subject to a pair of isothermal stability runs. Based on the average minimum energy for the set of two runs and the lowest minimum energy observed, it is apparent that the native fold is favored. In Table 6E, we describe the results of additional simulations on the native topology (in runs 3b and 3c) and the mirror-image topology (in runs 9b and 9c). The energy *versus* RMSD for these runs is presented in Figure 6(c). The average energy on pooling the results of the pair of runs for the native and mirror-image topology are -365.3 and $-350k_BT$, respectively.

If we confine ourselves to the run of lowest average energy, then structures whose coordinate and distance RMSD from native are 5.44 and 3.54 Å are predicted. The average coordinate and distance RMSD based on runs 3b and 3c are 5.62 and 3.64 Å, respectively. The final structure from run 3c is presented in Figure 6(d). The overall features of the plastocyanin fold are quite well recovered. The results here can be profitably compared to that of Levy and co-workers on a protein having a very similar topology, the variable light domain of a human immunoglobulin, 3fab (Saul *et al.*, 1978). They find that 90 restraints are required to produce a structure of backbone RMSD of 4.56 Å. In contrast, in this work, for a similar fold with 25 restraints, structures on the level of 5.44 Å RMSD are produced. If 46 restraints are used, then structures whose backbone RMSD is 3.51 Å can result. This is suggestive that the potential terms in the model can be used to supplement knowledge-based restraint information.

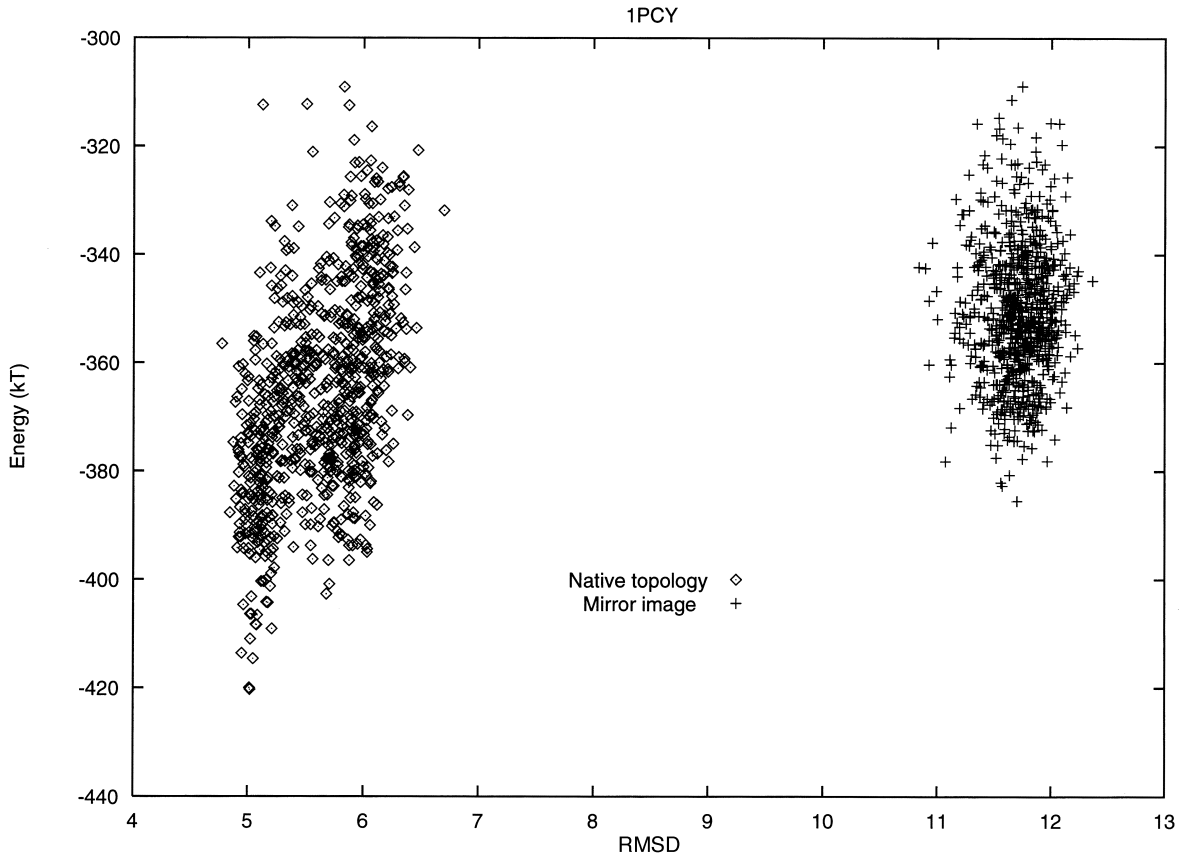


(a) **(b)**

Figure 5. (a) Correct fold of myoglobin, 1mba, obtained from the X-ray structure. (b) Representative predicted conformation of lmba obtained using 20 tertiary restraints. Figures are drawn using MOLSCRIPT (Kraulis, 1991).



(a) (b) (d)



(c)

Figure 6. (a) Fold of plastocyanin, 1pcy, obtained from the X-ray structure. (b) Representative predicted conformation of 1pcy obtained using 46 tertiary restraints. (c). Energy *versus* rms for the set of structures whose average properties are summarized in Table 6E. (d) Representative predicted conformation of 1pcy obtained using 25 tertiary restraints. Figures are drawn using MOLSCRIPT (Kraulis, 1991).

Bovine pancreatic trypsin inhibitor (BPTI)

To further compare the results of the present approach with those of Smith-Brown *et al*. (1993), the folding of the 18-55 fragment of bovine pancreatic trypsin inhibitor was undertaken for the five different sets of restraints listed in Table 7A. This fragment contains two β strands, associated with residues 18 to 24 and 29 to 35, and an α helix involving residues 48 to 55. In all simulations, nine restraints were used. The results are shown in Table 8B. A key observation is that the accuracy of the structure is different for different restraint combinations. We note that the total energy cannot be compared for different restraint sets, since different pair terms are redefined to be attractive, see above equation (7a). The average coordinate RMSD over the five restraint sets is 4.84 Å with a standard deviation of 0.8 Å. If the lowest average simulation is taken, then the average RMSD is 4.68 Å with a standard deviation of 0.9 Å. This is to be compared to the Smith-Brown *et al.* (1993) results of 5.1 Å based on a single structure.

Figure 7(a) shows the native structure of the 28-55 fragment (Wlodawer *et al*., 1987). Figure 7(b) and (c) show the lowest-energy predicted structure from run 3b using restraint set S2, with an RMSD of 4.98 Å and from run 4b using restraint set S5, with an RMSD of 5.40 Å. The characteristic twist of BPTI in this region is evident. However, in both cases the strands are too close when they cross. This is to be contrasted with the single conformation reported by Smith-Brown *et al*. (1993), where the twist is absent. They ascribe the lack of twist to the absence of side-chains. This conjecture is not inconsistent with the present results, but the origin of the twist needs additional investigation.

Another question of interest is just how much information is provided by the secondary and tertiary restraints and excluded volume alone. This

Table 6. Plastocyanin data

A. *Restraint list for plastocyanin folding with 46 restraints*

2-28	18-95	35-66	45-81
3-19	19-27	36-64	46-52
4-15	19-82	37-87	46-80
5-37	19-96	37-92	50-76
5-63	21-74	38-57	57-63
7-15	26-73	39-63	57-70
8-32	27-96	39-70	59-85
12-90	28-69	39-82	63-68
14-29	29-39	41-46	74-80
14-39	29-82	41-55	74-98
14-82	31-63	41-96	79-97
			83-88
			88-93

B. *Isothermal results of plastocyanin folding with 46 restraints*

Run no.[a]	E_{last}[b]	E_{min}[c]	$\langle E \rangle$[d]	r_{sat}/r_{tot}[e]	⟨cRMSD⟩[f]	⟨dRMSD⟩[g]	Topology
1a	–462.0	–499.3	–364.5	34/46	3.51	2.50	Native
2a	–107.6	–119.2	–50.36	19/46	9.24	5.30	Mirror
3a	–305.7	–329.7	–228.3	32/46	10.97	3.38	Mirror
4a	–431.9	–448.7	–342.4	34/46	3.33	2.28	Native
5a	–431.9	–453.9	–339.0	41/46	3.07	2.42	Native

[a] Each of a series of simulated annealing runs was performed over a reduced temperature of 2.1 to 1.8. Then, all runs designated by an a, e.g. 1a, started at a reduced temperature of 1.4 and finished at a reduced temperature of 1.0. No knowledge-based rule was used. Results are reported for the second set of refinement over this temperature range The prefactor γ in equation (7a), was 1.5.
[b] E_{last} is the energy of the final conformation.
[c] E_{min} is the minimum energy observed during the course of the simulation.
[d] $\langle E \rangle$ is the average energy of the given simulation.
[e] Ratio of the restraints satisfied, r_{sat}, to the total number of restraints, r_{tot}.
[f] The average coordinate RMSD is calculated with respect to the α-carbon atoms for residues 1 to 99.
[g] The distance RMSD is calculated with respect to the α-carbon atoms for residues 1 to 99.

C. *Restraint list for plastocyanin folding with 25 restraints*

2-28	19-27	46-55
3-16	19-82	46-80
5-37	23-98	55-72
5-63	26-71	74-98
5-93	29-39	80-98
10-33	29-82	
12-92	31-63	
14-29	37-84	
14-84	39-57	
14-94	41-82	

continued overleaf

Table 6 *(continued)*

D. *Isothermal results of plastocyanin folding with 25 restraints*

Run no.[a]	E_{last}[b]	E_{min}[c]	$\langle E \rangle$[d]	r_{sat}/r_{tot}[e]	⟨cRMSD⟩[f]	⟨dRMSD⟩[g]	Topology
1a	–423.5	–447.0	–341.3	18/25	6.40	4.23	Native
2a	–465.3	–486.8	–360.0	20/25	6.88	4.66	Native
3a	–472.8	–489.3	–360.3	22/25	6.22	4.22	Native
4a	–418.6	–437.2	–346.8	17/25	7.098	4.26	Misfold
5a	–414.5	–444.5	–341.4	18/25	7.26	5.11	Native
6a	–469.2	–479.7	–357.3	21/25	6.68	4.63	Native
7a	–423.4	–449.9	–334.9	19/25	7.22	5.03	Native
8a	–443.2	–464.5	–330.2	21/25	12.08	4.29	Mirror
9a	–418.1	–445.0	–344.4	20/25	12.18	4.64	Mirror
10a	–457.1	–474.0	–348.3	23/25	5.73	4.10	Native

[a] Each of a series of simulated annealing runs was performed over a reduced temperature of 2.1 to 1.8. Then, all runs designated by an a, e.g. 1a, started at a reduced temperature of 1.4 and finished at a reduced temperature of 1.0. No knowledge-based rules were used. Results are reported for the second set of refinements over this temperature range The prefactor γ in equation (7a) was 0.75 for folding and 1.5 for refinement.
[b] E_{last} is the energy of the final conformation.
[c] E_{min} is the minimum energy observed during the course of the simulation.
[d] $\langle E \rangle$ is the average energy of the given simulation.
[e] Ratio of the restraints satisfied, r_{sat}, to the total number of restraints, r_{tot}.
[f] The average coordinate RMSD is calculated with respect to the α-carbon atoms for residues 1 to 99.
[g] The distance RMSD is calculated with respect to the α-carbon atoms for residues 1 to 99.

E. *Further refinement of selected conformation from D*

Run no.[a]	E_{last}[b]	E_{min}[c]	$\langle E \rangle$[d]	r_{sat}/r_{tot}[e]	⟨cRMSD⟩[f]	⟨dRMSD⟩[g]	Topology
3b	–380.2	–414.0	–375.2	22/25	5.44	3.54	Native
3c	–351.0	–431.0	–356.5	22/25	5.79	3.74	Native
9b	–333.8	–390.3	–350.1	17/25	11.74	4.00	Mirror
9c	–318.5	–393.9	–351.6	18/25	11.69	4.02	Mirror

[a] Runs 3a and 9a were continued. Both b and c runs were done at a reduced temperature of 1.3. γ of equation (7a) equaled 1.5. Runs of type c are 3 1/3 times longer than runs of type b.
[b] E_{last} is the energy of the final conformation.
[c] E_{min} is the minimum energy observed during the course of the simulation.
[d] $\langle E \rangle$ is the average energy of the given simulation.
[e] Ratio of the restraints satisfied, r_{sat}, to the total number of restraints, r_{tot}.
[f] The average coordinate RMSD is calculated with respect to the α-carbon atoms for residues 1 to 99.
[g] The distance RMSD is calculated with respect to the α-carbon atoms for residues 1 to 99.

point is addressed in Table 7C, where for the five restraint sets, the simulations are performed without using any of the terms presented in equations (1) to (4). These simulations lack the sequence-specific terms reflecting intrinsic secondary structural propensities, burial terms, pair interactions and hydrogen bonding. Comparing the results of Table 7B with those of Table 8C, it is apparent that the RMSD from native is about 50% worse than when the remainder of the terms in the force-field are used. Thus, the inclusion of other terms in the potential that reflect the sequence specificity and protein-like features substantially enhances the accuracy of the predicted structures.

An analogous set of studies for the folding of the entire BPTI molecule was undertaken where the full potential was used. Simulations containing 12 restraints were employed, and the various restraint sets are listed in Table 7D. The resulting simulations are shown in Table 7E. Depending on the identity of the restraints structures whose RMSD from native ranging from about 6 to 9 Å are observed. The performance on BPTI of this algorithm is clearly not as good as in the other cases that have been studied. This may reflect the complexity of the BPTI topology as well as inadequacies in the present model.

Conclusions

Here, we have employed a lattice-based, reduced protein model for the simulation of native fold assembly using limited knowledge of the short-range and a small to moderate number of long-range restraints. In the MONSSTER algorithm, short-range restraints are encoded as loosely defined energetical biases that favor a geometry consistent with generic helical, β-type or turn backbone conformations. The chirality of turns is not specified; rather, they simply behave as flexible joints. When combined with tertiary interactions, the resulting potential is sufficient to produce folds of moderate resolution. These folds possess a protein-like pattern of hydrogen bonds, and a dense (but non-unique) packing of the side-chains. With $N/7$ (with N the number of residues) long-distance restraints operative, the present method permits the assembly of lattice protein models whose accuracy is in the range of 4 to 5.5 Å RMSD (from native) for C^{α}-traces of α and some

Table 7. 6pti data

A. *Summary of restraint sets for folding of fragment 18-55 of 6pti*

S1	S2	S3	S4	S5
13 34	13 34	13 34	13 34	13 34
1 18	2 17	1 18	2 18	1 18
3 27	4 28	3 27	3 29	3 27
5 14	6 26	5 31	5 16	5 16
6 38	12 35	6 13	6 38	7 14
12 35	16 27	7 14	7 14	12 35
18 27	26 38	16 27	12 35	16 27
28 34	28 34	26 38	18 27	26 38
32 36	28 38	28 38	28 38	28 34

B. *Simulation results on the 18-55 fragment of 6pti using the full potential*[a]

Run	E_{last}	E_{min}	$\langle E \rangle$	r_{sat}/r_{tot}	⟨cRMSD⟩	⟨dRMSD⟩
S1						
1b	–114.4	–128.5	–109.08	8/9	4/92	3.36
2b	–120.3	–129.9	–103.36	7/9	3.97	2.55
3b	–116.6	–127.9	–100.92	8/9	4.93	3.16
4b	–104.9	–135.5	–115.52	6/9	4.46	3.16
5b	–121.3	–133.3	–107.52	8/9	5.10	3.74
S2						
1b	–158.1	–184.8	–165.68	9/9	6.35	5.62
2b	–172.9	–182.5	–163.39	9/9	4.65	3.56
3b	–177.6	–189.0	–169.36	9/9	6.25	5.29
4b	–169.2	–180.3	–159.07	9/f9	6.38	5.97
5b	–155.6	–179.5	–160.31	9/9	5.92	5.02
S3						
1b	–72.0	–110.0	–89.84	8/9	4.17	2.06
2b	–84.2	–109.6	–80.07	7/9	6.30	4.40
3b	–83.3	–110.3	–85.59	6/9	5.11	3.40
4b	–92.1	–113.7	–90.47	8/9	4.00	2.14
5b	–127.4	–131.7	–114.75	7/9	4.16	2.61
S4						
1b	–183.4	–199.2	–171.15	9/9	4.48	3.00
2b	–175.3	–182.4	–157.79	7/9	4.57	3.01
3b	–147.5	–181.2	–156.46	7/9	4.40	2.64
4b	–171.4	–194.6	–168.67	8/9	5.16	3.12
5b	–153.6	–169.4	–154.47	8/9	4.37	3.19
S5						
1b	–104.2	–119.8	–101.40	8/9	4.06	3.20
2b	–82.0	–120.0	–93.69	8/9	4.99	3.88
3b	–93.8	–130.2	–107.42	8/9	4.38	3.11
4b	–103.0	–135.2	–117.97	8/9	4.03	2.95
5b	–112.1	–128.2	–106.57	7/9	4.13	2.78

[a] Simulated annealing was done over a temperature range of 1.95 to 1. This is followed by an isothermal refinement run at a reduced temperature of 1. For each set of restraints, five independent folding/refinement runs were performed.

C. *Simulation results on the 18-55 fragment of 6pti using only excluded volume interactions and secondary and tertiary restraint information*[a]

Run	E_{last}	E_{min}	$\langle E \rangle$	r_{sat}/r_{tot}	⟨cRMSD⟩	⟨dRMSD⟩
			S1			
1b	30.6	25.5	34.2	7/9	7.33	4.75
2b	31.5	–3.40	33.4	7/9	7.48	4.81
3b	33.6	–1.50	32.8	8/9	7.64	4.60
4b	32.7	–2.30	32.9	7/9	7.52	4.41
5b	33.3	25.0	32.8	7/9	7.43	4.78
			S2			
1b	–35.4	–39.7	–33.26	9/9	8.34	4.39
2b	–34.9	–112.5	–37.12	9/9	8.28	4.35
3b	–35.2	–82.1	–37.59	9/9	5.79	4.38
4b	–20.2	–72.4	–35.13	7/9	8.27	4.25
5b	–24.3	–159.9	–40.22	5/9	8.26	4.36
			S3			
1b	32.2	24.8	34.8	8/9	5.74	4.15
2b	32.2	25.7	34.4	7/9	6.03	4.27
3b	37.2	23.0	33.5	8/9	5.97	4.25

continued overleaf

Table 7 *(continued)*

Run	E_{last}	E_{min}	$\langle E \rangle$	r_{sat}/r_{tot}	⟨cRMSD⟩	⟨dRMSD⟩
4b	32.1	26.0	34.7	7/9	6.05	4.25
5b	35.4	24.7	34.2	8/9	6.01	4.24
			S4			
1b	–35.4	–37.8	–32.2	9/9	6.63	3.81
2b	–35.0	–39.8	–32.2	8/9	6.75	3.87
3b	–34.9	–39.8	–32.0	8/9	7.11	3.87
4b	–21.3	–39.9	–32.1	7/9	6.74	3.92
5b	–30.4	–68.3	–33.4	9/9	7.61	3.78
			S5			
1b	31.0	20.1	34.5	7/9	7.17	4.58
2b	38.7	24.4	34.3	6/9	6.75	4.53
3b	36.4	27.0	33.7	7/9	8.30	4.71
4b	32.4	26.6	33.6	7/9	7.34	4.62
5b	34.9	25.0	34.2	7/9	8.19	4.58

[a] Simulated annealing was done over a temperature range of 1.95 to 1. This is followed by an isothermal refinement run at a reduced temperature of 1. For each set of restraints, five independent folding/refinement runs were performed. All energetic contributions of equations (1) to (4) were deleted.

D. *Summary of restraint sets for folding of 6pti*

S1	S2	S3	S4
5-55	5-55	5-55	5-55
14-38	14-38	14-38	14-38
30-51	30-51	30-51	30-51
1-23	4-45	1-23	2-54
4-42	6-23	11-33	10-41
10-39	10-41	19-34	21-32
17-34	11-33	22-43	21-45
21-32	17-34	23-30	23-52
23-52	22-33	29-52	23-55
23-55	33-44	30-48	29-52
33-44	43-51	43-51	35-40
45-54	45-55	45-55	45-54

E. *Simulation results on 6pti using the full potential*[a]

Run	E_{last}	E_{min}	$\langle E \rangle$	r_{sat}/r_{tot}	⟨cRMSD⟩	⟨dRMSD⟩
			1			
1b	–164.7	–189.40	–159.96	10/12	10.21	6.13
2b	–139.0	–189.80	–165.41	10/12	10.00	5.14
3b	–225.7	–231.20	–208.06	11/12	9.75	4.59
4b	–188.3	–204.50	–176.91	10/12	9.58	4.57
5b	–213.0	–226.10	–203.03	10/12	10.42	5.15
			2			
1b	–182.4	–214.50	–190.91	11/12	8.74	4.56
2b	–227.0	–243.50	–216.59	11/12	9.85	4.76
3b	–195.0	–212.30	–187.59	9/12	9.51	4.57
4b	–217.2	–376.60	–191.14	10/12	9.49	4.38
5b	–215.1	–233.60	–217.16	11/12	9.92	4.91
			3			
1b	–149.5	–152.10	–130.06	10/12	6.50	4.19
2b	–152.1	–164.20	–143.11	11/12	6.00	4.19
3b	–138.1	–163.90	–134.09	10/12	6.15	4.24
4b	–151.6	–168.70	–140.66	10/12	5.89	3.94
5b	–165.4	–186.10	–160.93	11/12	6.26	4.75
			4			
1b	–238.1	–273.00	–246.39	11/12	9.10	3.85
2b	–243.4	–271.70	–244.19	11/12	9.36	4.22
3b	–221.6	–252.50	–225.14	11/12	9.18	4.96

[a] Simulated annealing was done over a temperature range of 1.95 to 1. This is followed by an isothermal refinement run at a reduced temperature of 1. For each set of restraints, five independent folding/refinement runs were performed.

α/β proteins. Reflecting their greater intrinsic flexibility, β-proteins require of the order of $N/4$ restraints to achieve comparable accuracy. When the number of tertiary restraints is increased to roughly $N/2$, then structures whose average coordinate RMSD from native is 3.5 Å result. For larger mixed motif proteins such as flavodoxin and thioredoxin, when roughly $N/4$ restraints are used, then structures whose coordinate RMSD is below 4.0 Å from native are predicted. This level of

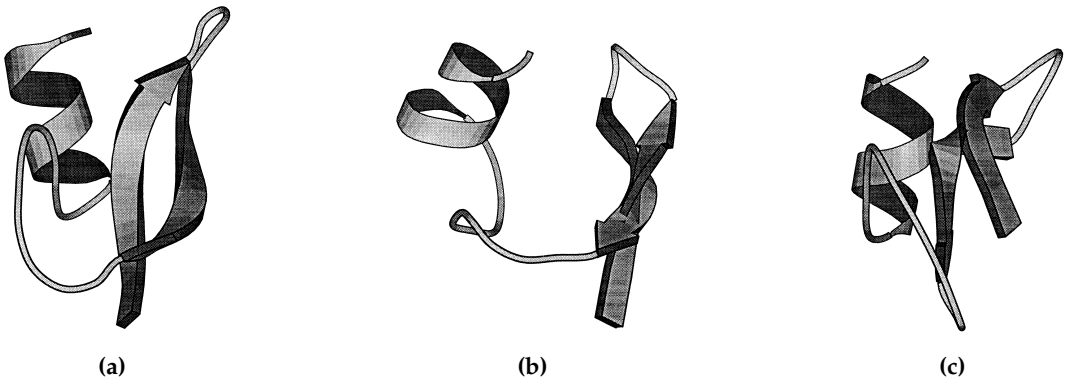


Figure 7. (a) Fold of the 18-55 fragment of bovine pancreatic trypsin inhibitor, 6pti, obtained from the X-ray structure. (b) Representative predicted conformation of 6pti obtained using nine tertiary restraints from restraint set S2. (c) Representative predicted conformation of 6pti obtained using nine tertiary restraints from restraint set S5. Figures are drawn using MOLSCRIPT (Kraulis, 1991).

accuracy is usually sufficient for full atom model-building. Of course, in practice, as was shown in the case of BPTI, the quality of the resulting structures depends significantly on the distribution of these restraints.

MONSSTER produces structures that are consistently of higher resolution than those obtained in previous investigations. This improvement comes from the fact that folding is driven by the distance restraints and by the underlying potential. As the number of restraints is reduced, the computations become more expensive due to the necessity of longer simulation runs and the greater fraction of unsuccessful folding events. Incorrect topologies can be dismissed on the basis of both energetic criteria and the fact that a larger number of restraints may be unsatisfied. However, in many cases when there are only a small number of restraints, then even if all or almost all such tertiary restraints are satisfied, the molecule may be misfolded. In such a situation, one is forced to rely solely on energetic criteria. Thus, at least ten independent assembly simulations should be performed. These simulated annealing, topology assembly runs need to be followed by at least one and preferably more, isothermal, dynamic Monte Carlo refinement runs. The latter define the average energy that should be used to select the fold. In all cases done to date, this protocol has been successful. However, in the limit of very few restraints, since the current version of the potential cannot fold all proteins, at some point, it must fail.

An important question that must be addressed is how well the model performs when a subset of the secondary and/or tertiary restraints provided to MONSSTER are incorrect. Preliminary results indicate that even when some topological elements that are not present in the native fold are assigned to be in contact, the method can predict the native fold. Thus far, while alternative topologies are generated, such incorrect structures can be eliminated on the basis of their minimum and average energies as compared to that of the native fold. Work is in progress to explore this issue further.

The proposed methodology may be useful in the early stages of model building that incorporate information based on a small number of restraints from NMR experiments or from other experiments (for example, knowledge of disulfide bonds could be used as additional, or the only, restraints). These restraints could be very loosely defined and incomplete. Only partial knowledge of secondary structure is assumed; both helical and β states are defined over a broad region of conformations, and turns are treated as flexible regions of unspecified chirality. It should further be noted that NMR docking simulations could be easily enhanced by introducing more specific restraint information. This possibility is now being explored. MONSSTER could be used to refine structures based on models built using classical homology modeling tools. How well this will work remains to be established. Finally, we are examining the use of this methodological protocol to fold small proteins *de novo*, where the secondary structure is predicted as are the tertiary restraints. Preliminary results are quite encouraging with all-helical, β, α/β and α + β motifs having been folded.

Acknowledgements

This work was supported by NIH grant no. GM-37408. Helpful discussions with Dr Jane Dyson, Dr Adam Godzik, and Dr Peter Wright are gratefully acknowledged, as is the very helpful guidance provided by the referees. A.K. acknowledges partial support from University of Warsaw grant BST-34/96, and is an International Research Scholar of the Howard Hughes Medical Institute (grant no. 75195-543402).

References

Aszodi, A., Gradwell, M. J. & Taylor, W. R. (1995). Global fold determination from a small number of distance restraints. *J. Mol. Biol.* **248**, 308–326.

Bolognesi, M., Onesti, S., Gatti, G., Coda, A., Ascenzi, P. & Brunori, M. (1989). *Aplysia limacina* myoglobin.

Crystallographic analysis at 1.6 Å resolution. *J. Mol. Biol.* **205**, 529–544.
Braun, W. & Go, N. (1985). Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. *J. Mol. Biol.* **186**, 611–626.
Chothia, C. & Finkelstein, A. (1990). The classification and origins of protein folding patterns. *Annu. Rev. Biochem.* **59**, 1007–1039.
Clore, G. M., Robien, M. A. & Gronenborn, A. M. (1993). Exploring the limits of precision and accuracy of protein structures determined by nuclear magnetic resonance spectroscopy. *J. Mol. Biol.* **231**, 82–102.
Godzik, A., Kolinski, A. & Skolnick, J. (1993). Lattice representation of globular proteins: how good are they? *J. Comp. Chem.* **14**, 1194–1202.
Gronenborn, A., Filpula, D. R., Essig, N. Z., Achari, A., Whitlow, M., Wingfield, P. T. & Clore, G. M. (1991). A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science,* **253**, 657–660.
Gronenborn, A. M. & Clore, G. M. (1994). Where is NMR taking us? *Proteins: Struct. Funct. Genet.* **19**, 273–276.
Guentert, P., Braun, W. & Wuthrich, K. (1991). Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the suporting programs CALIBA, HABAS and GLOMSA. *J. Mol. Biol.* **217**, 517–530.
Guss, J. M. & Freeman, H. C. (1983). Structure of oxidized poplar plastocyanin at 1.6 Å resolution. *J. Mol. Biol.* **169**, 521–563.
Havel, T. F. & Wuthrich, K. (1985). 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.
Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577–2637.
Kajava, A. V. (1992). Left handed topology of supersecondary structure formed by aligned α-helix and β-hairpin. *FEBS Letters,* **302**, 8–10.
Katti, S. K., LeMaster, D. M. & Eklund, H. (1990). Crystal structure of thioredoxin from *Escherichia coli* at 1.68 Å resolution. *J. Mol. Biol.* **212**, 167–184.
Kolinski, A. & Skolnick, J. (1994a). Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. *Proteins: Struct. Funct. Genet.* **18**, 353–366.
Kolinski, A. & Skolnick, J. (1994b). Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins: Struct. Funct. Genet.* **18**, 338–352.
Kolinski, A. & Skolnick, J. (1996). *Lattice Models of Protein Folding, Dynamics and Thermodynamics*, R. G. Landes Company, Austin, TX.
Kolinski, A., Skolnick, J. & Godzik, A. (1996). A method for the prediction of surface ''U''-turns and transglobular connections in small proteins. *Proteins: Struct. Funct. Genet*. In the press.
Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.
Leijonmarck, M. & Liljas, A. (1987). Structure of the C-terminal domain of the ribosomal protein 17/L12 from *Escherichia coli* at 1.7 Å resolution. *J. Mol. Biol.* **195**, 555–579.
Levitt, M. & Greer, J. (1977). Automatic identification of secondary structure in globular proteins. *J. Mol. Biol.* **114**, 181–293.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **51**, 1087–1092.
Mumenthaler, C. & Braun, W. (1995). Automated assignment of simulated and experimental NOESY spectra of proteins by feedback littering and self-correcting distance geometry. *J. Mol. Biol.* **254**, 465–480.
Olszewski, K. A., Kolinski, A. & Skolnick, J. (1996a). Does a backwardly read sequence have a unique native state? *Protein Eng.* **9**, 5–14.
Olszewski, K. A., Kolinski, A. & Skolnick, J. (1996b). Folding simulations and computer redesign of protein A three helix bundle motifs. *Proteins: Struct. Funct. Genet.* **25**, 286–299.
Pastore, A. Atkinson, R. A., Saudek, V. & Williams, R. J. P. (1991). Topological mirror images in protein structure computation: an underestimated problem. *Proteins: Struct. Funct. Genet.* **10**, 22–32.
Rost, B. & Sander, C. (1993). Prediction of secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599.
Saul, F. A., Amzel, L. M. & Poljak, R. J. (1978). Preliminary refinement and structural analysis of the Fab fragment from human immunoglobulin new at 2.0 Å resolution. *J. Biol. Chem.* **253**, 585–597.
Skolnick, J. (1996). ftp.scripps.edu in directory /pub/skolnick/nmr.
Skolnick, J., Jaroszewski, L., Kolinski, A. & Godzik, A. (1996). Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Sci*. In the press.
Smith, W. W., Burnett, R. M., Darling, G. D. & Ludwig, M. L. (1977). Structure of the semiquinone form of flavodoxin from *Clostridium hp*. Extension of 1.8 Å resolution and some comparisons of the oxidized state. *J. Mol. Biol.* **117**, 195–225.
Smith-Brown, M. J., Kominos, D. & Levy, R. M. (1993). Global folding of proteins using a limited number of distance restraints. *Protein Eng.* **6**, 605–614.
Vieth, M., Kolinski, A., Brooks, C. L., III & Skolnick, J. (1994). Prediction of the folding pathways and structure of the GCN4 leucine zipper. *J. Mol. Biol.* **237**, 361–367.
Wlodawer, A., Nachman, J., Gilliland, G. L., Gallagher, W. & Woodward, C. (1987). Structure of form III crystals of bovine pancreatic trypsin inhibitor. *J. Mol. Biol.* **198**, 469–480.

Appendix

Detailed study of the native conformation of a large number of proteins indicates that β-strands and α-helices tend to pack at well-defined orientations (Chothia & Finkelstein, 1990). Furthermore, it is well established that in α/β and α + β proteins there is a very strong preference that βαβ fragments adopt a right-handed topology, as in Figure A1(a), as opposed to the left-handed topology shown in Figure A1(b) (Chothia & Finkelstein, 1990). Similarly, for small proteins, there is a very strong tendency for ββα fragments to adopt a left-handed topology as depicted in

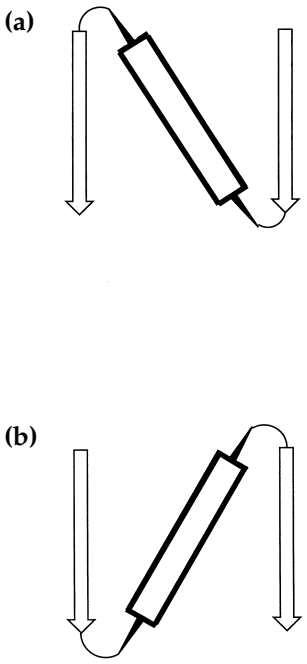


Figure A1. In globular proteins, βαβ fragments exhibit a strong preference to adopt a right-handed topology, as in (a), as opposed to the left-handed topology shown in (b) (Chothia & Finkelstein, 1990). In the Figure, the thin arrows represent strands in the paper plane, and the thick cylinder represents a helix above the paper plane.

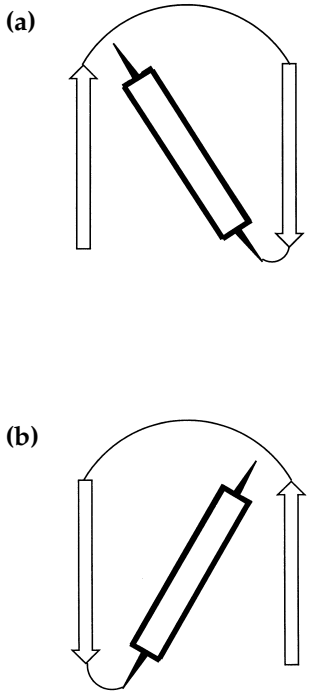


Figure A2. For small proteins, there is a very strong tendency for ββα fragments to adopt a left-handed topology as depicted in (a) as opposed to the right-handed topology of (b). In the Figure, the thin arrows represent strands in the paper plane, and the thick cylinder represents a helix above the paper plane.

Figure A2(a) as opposed to the right-handed topology of Figure A2(b) (Kajava, 1992).

In what follows, we describe a protocol for detecting the relative orientation of successive secondary elements as well the handedness of βαβ and ββα fragments, and then introduce a penalty term for supersecondary structural elements that violate such knowledge-based rules. In such an analysis, we shall implicitly assume that the secondary structure is known. The secondary structure could either be detected based on the instantaneous conformation of the chain, or could be assumed based on *a priori* knowledge of the secondary structure. We further assume here that secondary structural elements that have been identified lie between turn regions, which we call U turns (Kolinski *et al*., 1996), where the chain reverses global direction. Let there be N_{sec} such secondary elements in the chain. The secondary structure type of the N_{sec} elements is stored as a vector **S** of length N_{sec}. Let the vector that stores the residue at which the structural element begins and ends be **IB** and **IE**. Thus, the jth element begins and ends at $ib(j)$ and $ie(j)$, respectively.

We point out that folding without these knowledge-based terms has been successful for all proteins examined to date. However, there are practical advantages to incorporating these rules into the folding algorithm. Their inclusion eliminates the problem of mirror-image topologies, and thus they can substantially enrich conformational sampling. Such knowledge-based rules are especially useful when the number of restraints is very small. For the 1ctf and the B domain of protein G, folding from the denatured state was done both with and without such rules. For illustrative purposes, we report results based on a series of simulations that incorporated the knowledge-based rules (see Tables 1B and 2B of the main text, repectively). Topology assembly of thioredoxin occurred without such rules, but for a series of isothermal simulations (a-type runs) knowledge-based rules were employed. Note that even without such rules the average energy is lower for the native than the mirror-image topology (see Table 3B of the main text).

The key to the successful detection of the handedness of a supersecondary fragment is to first smooth the local chain backbone. This eliminates local irregularities that may cause problems with correct handedness identification.

Let $\mathbf{X} = \{\mathbf{x}_i\}$ denote the actual coordinates of the α-carbon atoms. To account for end effects, in a chain containing N residues labeled 1 to N, we set:

$$\mathbf{x}_k = \mathbf{x}_1 \text{ and } k = -4, -3, -2, -1, 0 \qquad \text{(A1a)}$$

and:

$$\mathbf{x}_{N+k} = \mathbf{x}_N \text{ and } k = 1, 2, 3, 4, 5 \qquad \text{(A1b)}$$

Then:

$$\mathbf{Y} = \mathbf{MX} \tag{A2a}$$

with $\mathbf{M}$ an $N + 10$ by $N + 10$ matrix:

$$\mathbf{M}(i, i) = 51/243$$
$$\mathbf{M}(i, i \pm 1) = 45/243$$
$$\mathbf{M}(i, i \pm 2) = 30/243$$
$$\mathbf{M}(i, i \pm 3) = 15/243$$
$$\mathbf{M}(i, i \pm 4) = 5/243$$
$$\mathbf{M}(i, i \pm 5) = 1/243$$
$$\mathbf{M}(k, l) = 0 \text{ otherwise} \tag{A2b}$$

and $\mathbf{Y} = \{\mathbf{y}_i\}$, with $\mathbf{y}_i$ coordinates of the ith α-carbon atom on the smoothed chain trace.

The first type of knowledge-based rule asks whether an adjacent pair of supersecondary elements is antiparallel. Helical hairpins need not be antiparallel (they can pack at all angles but in addition to the antiparallel orientation, perpendicular orientations can occur). Successive β strands are either antiparallel or can pack at roughly 90° if they correspond to a crossover between sheets. α/β or β/α elements almost always tend to be antiparallel. Thus, if at least one of two successive secondary elements is a β state, say involving secondary structural elements k and $k + 1$, then, we examine the following:

$$\text{Let } is1 = is(k-1),\ ie1 = ie(k-1),$$

and

$$is2 = is(k),\ ie2 = ie(k).$$

Then:

$$\mathbf{YB}(k) = \mathbf{y}(ie(k)) - \mathbf{y}(is(k)) \tag{A3a}$$

$$\cos(\theta(k)) = \mathbf{YB}(k)\cdot\mathbf{Y}(k-1)/|\mathbf{YB}(k)\|\mathbf{YB}(k-1)| \tag{A3b}$$

If $\cos(\theta(\mathrm{k}))\ > 0.77$, then:

$$v_{\text{anti}}(k) = 10\gamma(\cos(\theta(k)) + 0.77)^2 \tag{A4}$$

where γ is defined in equation (7a). $v_{\text{anti}}(k)$ penalizes successive vectors whose angle is less than 39.6°. Thus, states that are perpendicular are ignored, as are antiparallel vector orientations. Only orientations that are almost parallel are penalized.

The total contribution to the energy arising from orientational preferences of successive secondary elements is:

$$V_{\text{anti}} = \sum_{2}^{N_{\text{sec}}} v_{\text{anti}}(k)\lambda(k) \tag{A5a}$$

$$\lambda(k) = 0 \text{ if the } k-1 \text{ and } k \text{ secondary structural elements are helical} \tag{A5b}$$
$$\lambda(k) = 1 \text{ otherwise}$$

Next, we implement the rule that $\beta\alpha\beta$ elements tend to be right-handed. Suppose that we have established that the $k-2$, $k-1$ and kth regions of the chain are helical, β and helical, respectively. If elements $k-1$ and k, and k and $k+1$ are antiparallel, i.e. if the angle between successive elements is less than 45°, then we check for the handedness of the $\beta\alpha\beta$ fragment. Let:

$$\mathbf{a} = \mathbf{YB}(k) \tag{A6a}$$

Let $ie1 = ie(k-2)$, $ie3 = ie(k)$, $is3 = ib(k)$, then:

$$\mathbf{b} = \mathbf{y}(ie1) - \mathbf{y}(is3) \tag{A6b}$$

$$\mathbf{e} = \mathbf{a} \otimes \mathbf{b}/|\mathbf{a} \otimes \mathbf{b}| \tag{A6c}$$

Now:

$$m = (is3 + 1 + ie1)/2 \tag{A7a}$$

$$m3 = (ie3 + is3 + 1)/2 \tag{A7b}$$

$$\mathbf{d} = \mathbf{y}(m) - \mathbf{y}(m3) \tag{A8a}$$

$$dist_{\beta\alpha\beta} = \mathbf{d}\cdot\mathbf{e} \tag{A8b}$$

$|dist_{\beta\alpha\beta}|$ is the distance of the central helix from the plane formed by the two β strands. If $dist_{\beta\alpha\beta}$ is negative, then the chirality of the $\beta\alpha\beta$ connection is left-handed, and it is penalized by:

$$v_{\beta\alpha\beta}(k) = 20\cdot\gamma\cdot dist^2_{\beta\alpha\beta} \text{ if } 0 > dist_{\beta\alpha\beta} \tag{A9}$$

The total contribution to the energy from the incorrect handedness of $\beta\alpha\beta$ connections is:

$$V_{\beta\alpha\beta} = \sum_{k=3}^{N_{\text{sec}}} v_{\beta\alpha\beta}(k)\eta(k) \tag{A10a}$$

where:

$$\eta(k) = 1 \text{ if the secondary structure of elements } k-2, k-1, \text{ and } k \text{ are } \beta, \alpha, \beta \tag{A10b}$$
$$\eta(k) = 0 \text{ otherwise}$$

Next, we turn to the determination of the handedness of $\beta\beta\alpha$ connections. Suppose that we have established that the $k-2$nd, $k-1$st and kth secondary elements of the chain are β, β and α, respectively. If elements $k-1$ and k, and k and $k+1$ are antiparallel, i.e. if the angle between successive elements is less than 45°, then we check for the handedness of the $\beta\beta\alpha$ fragment.

Let $is1 = is(k-2)$ and $ie1 = ie(k-2)$; $is2 = is(k-1)$ and $ie2 = ie(k-1)$; $is3 = is(k)$ and $ie3 = ie(k)$, then:

$$m1 = (is1 + ie1 + 1)/2 \tag{A11a}$$

$$m2 = (ie2 + is2 + 1)/2 \tag{A11b}$$

$$\mathbf{a2} = \mathbf{YB}(k-1) \tag{A11c}$$

$$\mathbf{b2} = \mathbf{y}(m2) - \mathbf{y}(m1) \tag{A11d}$$

So that:

$$\mathbf{e2} = \mathbf{a2} \otimes \mathbf{b2}/|\mathbf{a2} \otimes \mathbf{b2}| \tag{A12}$$

Now:

$$m3 = (is3 + ie3 + 1)/2 \tag{A13a}$$

$$\mathbf{d2} = \mathbf{y}(m3) - \mathbf{y}(m2) \quad \text{(A13b)}$$

Then:

$$dist_{\beta\beta\alpha} = \mathbf{d2}\cdot\mathbf{e2} \quad \text{(A14)}$$

$|dist_{\beta\beta\alpha}|$ is the magnitude of the distance from the plane formed by the two β-sheets to the α-helix. If $dist_{\beta\beta\alpha} < 0$, then the chirality of the connection is left-handed and is preferred. Otherwise, it is disfavored by:

$$v_{\beta\beta\alpha}(k) = 20\cdot\gamma\cdot dist^2_{\beta\beta\alpha} \text{ if } dist_{\beta\beta\alpha} > 0 \quad \text{(A15)}$$

The total contribution to the energy due to the incorrect handedness of $\beta\beta\alpha$ connections is:

$$V_{\beta\beta\alpha} = \sum_{k=3}^{N_{sec}} v_{\beta\beta\alpha}(k)\mu(k) \quad \text{(A16a)}$$

where:

$\mu(k) = 1$ if the secondary structure of element $k-2$, $k-1$ and k are β, β,α (A16b)

$\mu(k) = 0$ otherwise

References

Chothia, C. & Finkelstein, A. (1990). The classification and origins of protein folding patterns. *Annu. Rev. Biochem.* **59**, 1007–1039.

Kajava, A. V. (1992). Left handed topology of supersecondary structure formed by aligned α-helix and β-hairpin. *FEBS Letters*, **302**, 8–10.

Kolinski, A., Skolnick, J., Godzik, A. & Hu, W. P. (1996). A method for the prediction of surface ''U'' turns and transglobular connections in small proteins. *Proteins: Struct. Funct. Genet.* In the press.

Edited by B. Honig

(Received 10 July 1996; received in revised form 3 October 1996; accepted 3 October 1996)