Article No. jmbi.1999.2861 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 290, 267–281

JMB

Ab Initio Fold Prediction of Small Helical Proteins using Distance Geometry and Knowledge-based Scoring Functions

Enoch S. Huang[1], Ram Samudrala[2] and Jay W. Ponder[1]*

[1]*Department of Biochemistry and Molecular Biophysics Washington University School of Medicine, Saint Louis MO 63110, USA*

[2]*Department of Structural Biology, Stanford University School of Medicine, Stanford CA 94305-5400, USA*

**Corresponding author*

The problem of protein tertiary structure prediction from primary sequence can be separated into two subproblems: generation of a library of possible folds and specification of a best fold given the library. A distance geometry procedure based on random pairwise metrization with good sampling properties was used to generate a library of 500 possible structures for each of 11 small helical proteins. The input to distance geometry consisted of sets of restraints to enforce predicted helical secondary structure and a generic range of 5 to 11 Å between predicted contact residues on all pairs of helices. For each of the 11 targets, the resulting library contained structures with low RMSD *versus* the native structure. Near-native sampling was enhanced by at least three orders of magnitude compared to a random sampling of compact folds. All library members were scored with a combination of an all-atom distance-dependent function, a residue pair-potential, and a hydrophobicity function. In six of the 11 cases, the best-ranking fold was considered to be near native. Each library was also reduced to a final *ab initio* prediction *via* consensus distance geometry performed over the 50 best-ranking structures from the full set of 500. The consensus results were of generally higher quality, yielding six predictions within 6.5 Å of the native fold. These favorable predictions corresponded to those for which the correlation between the RMSD and the scoring function were highest. The advantage of the reported methodology is its extreme simplicity and potential for including other types of structural restraints.

Keywords: protein structure prediction; distance geometry; helix packing

Introduction

Most applications of computational chemistry to biopolymers reduce to finding an appropriate compromise between the accuracy of the underlying potential energy model and the ability to sample important configurations adequately. In the context of *ab initio* prediction of protein structure, these coupled concerns can be largely separated into two distinct problems: searching conformational space (fold generation) and devising a scoring function (near-native fold selection). Minimization-based methods attempt to solve the search and selection problems concomitantly (Bowie & Eisenberg, 1994; Cui *et al*., 1998; Dandekar & Argos, 1996; Jones, 1997; Monge *et al*., 1995; Mumenthaler & Braun, 1995; Ortiz *et al*., 1998; Pedersen & Moult, 1997; Simons *et al*., 1997; Srinivasan & Rose, 1995; Sun *et al*., 1995; Wilson & Doniach, 1989). Alternatively, one can first sample conformational space using combinatorial methods to produce a library of protein-like folds (Chelvanayagam *et al*., 1998; Cohen *et al*., 1979; Covell, 1992; Hinds & Levitt, 1994; Park & Levitt, 1996). This library is then screened in order to select the most native-like of the candidate folds. The clean separation of the two problems has two advantages. First, the conformational search cannot be trapped in unproductive minima on a particular surface. Second, a fold library gives any scoring function a chance to select the best folds. Even if present methods do not suffice, the independent development of scoring functions remains a viable option.

Abbreviations used: DME, distance matrix error; RAPDF, residue-specific all-atom probability discriminatory function; HCF, hydrophobic compactness function.

E-mail address of the corresponding author: ponder@dasher.wustl.edu

0022-2836/99/260267–15 $30.00/0

In order for the library approach to be useful for *ab initio* fold prediction, certain criteria should be met. First, the library must contain at least one native-like fold. Successful representation of the native fold should not depend on knowing structural information about the target, such as native secondary structure, disulfide bonds, and radius of gyration. For small proteins, a coordinate root-mean-square deviation (RMSD) of about 6 Å relative to the native structure has been suggested as a target value (Reva *et al*., 1998). The correlation between score and RMSD tends to be weak (Park & Levitt, 1996; Park *et al*., 1997), but is somewhat stronger at lower RMSD (Samudrala & Moult, 1998). Hence, the odds that a near-native fold will be selected as the best-scoring fold varies directly with the concentration of near-native folds and inversely with the RMSD of the best folds in the library.

Because of the requirement that at least one near-native fold be present, combinatorial approaches such as enumeration on a lattice have served well in the past. These methods have the advantage that conformational space can be broadly and evenly sampled by a coarse-grained search. For many discrete-state models, the average RMSD of the best possible folds is related to n, the number of states per residue, by the following relationship: $\langle \mathrm{RMSD} \rangle \propto n^{0.5}$ (Park & Levitt, 1995). However, even though a 2 Å model is attainable by a four-state discrete model, this is very difficult in practice because correct secondary structures must be enforced and the freely rotatable residues carefully selected in order for the library size to remain tractably small. A tetrahedral lattice model that does not require *a priori* secondary structure or loop information can visit dozens of folds in the 6 Å range, but only after $\sim 10^7$ walks are completed (Hinds & Levitt, 1992). Thus, the concentration of near-native folds in the library is limited by the systematic nature of the search and the coarseness of the discretization, which together prevent the revisitation of promising folds.

Distance-based methods have been widely explored as a tool for protein structure prediction. These methods fall into two broad categories: metric matrix distance geometry, which uses a mathematical projection from distance space to three-dimensional space known as embedding (Aszodi *et al*., 1995; Havel, 1991), and minimization against distance constraints (Chelvanayagam *et al*., 1998; Lund *et al*., 1996; Mumenthaler & Braun, 1995; Ortiz *et al*., 1998; Smith-Brown *et al*., 1993). Similar to many energy minimization methods, distance-based approaches typically start with rigid or semi-rigid secondary structure elements; these elements are then assembled into compact structures using the distance information. Previous studies have estimated the minimum number of native or correct distances required to build native-like models. These native distances are usually applied in conjunction with force-fields generically used for proteins, predicted distance information derived from multiple sequence alignments, and hydrophobicity considerations. One study showed that only three native distances between each predicted secondary structure unit were sufficient to fold proteins to 3-5 Å RMSD (Smith-Brown *et al*., 1993), and two others suggested that as few as one correct inter-residue distance per residue in the protein may be sufficient to build near-native models (Aszodi *et al*., 1995; Lund *et al*., 1996). More recent methods report the successful folding of proteins using only predicted distances. The algorithm by Mumenthaler & Braun (1995) successfully folded six small helical proteins, starting with correct secondary structure assignments. Finally, Ortiz *et al*. (1998) extracted contact information from correlated mutation analysis enriched by a threading procedure. These predicted restraints are incorporated into an elaborate force field, which assembles native-like folds on a lattice with final RMSD between 3-7 Å.

Here we describe a novel library approach towards the *ab initio* fold prediction of helical proteins. In the absence of *a priori* distance information specific to the target, we specify a fixed distance range between each pair of predicted helices. These generic inter-helical distances are set from 5 to 11 Å and span those commonly observed in small helical proteins. Because these distance constraints are too few to determine a unique three-dimensional structure, we generate many different models consistent with the inter-helical distances. Unlike combinatorial approaches, distance geometry, in principle, need not suffer from the problems of coarse sampling and exhaustive enumeration; loop conformations and helix packing arrangements are sampled in continuous space and selected at random. We show that it is an efficient method, successfully generating many native-like topologies after only 500 trials. Finally, fold selection by previously published scoring functions is presented and the implications for *ab initio* fold prediction are discussed in detail.

Results

Prediction of secondary structure

The secondary structure and solvent accessibility predictions for each target was obtained by the PHD PredictProtein Server (http://www.embl-heidelberg.de) (Rost *et al*., 1994). Considering only protein sequences under 100 residues, we collected the first ten for which secondary structure prediction consisted entirely of alpha-helices and coil. An additional target, 1nkl, was added to test the effects of incorrect secondary structure prediction and misprediction of the number of helices (see below).

Table 1 reports the three-state accuracy (Q3) of the prediction as compared to the secondary structure assignment of the native structure (Kabsch & Sander, 1983) for the 11 targets. The number of secondary structure units predicted by PHD (2°_{PHD})

Table 1. Secondary structures of the target proteins

PDB	len	Q3	PHD	PDB	nBur	nDist	nGood
1aca	86	90.7	4	4	3	6	4
1c5a	65	90.8	4	4	4	6	4
1fc2	43	76.7	3	3	0	3	3
1hdd	57	87.7	3	3	3	3	2
1nkl	78	83.3	4	5	4	6	4
1pou[a]	71	83.1	4	4	4	6	3
1r69	63	87.3	5	5	5	10	8
1trl	62	96.8	3	3	3	3	2
1xbl	75	86.7	4	4	3	6	3
2utg	70	92.9	4	4	3	6	3
4icb	76	85.5	4	4	4	6	4

[a] A predicted helix was broken into two helices at the position of low confidence in the PHD reliability index (Rel sec).

matched that of the native structure (2°_{PDB}) in every case, except for 1pou and 1nkl. Helices 3 and 4 of 1pou were merged into one long predicted helix. We severed this predicted third helix at the position of low confidence in the PHD reliability index (Rel sec), forming a total of four helices. For 1nkl, PHD incorrectly joined helices 3 and 4 and mispredicted a β-strand for helix 5. No attempt was made to correct these erroneous assignments.

Assignment of inter-helical distances

A distance range of 5 to 11 Å was specified between each pair of helices. These distances were measured between the C^{α} atoms of designated contact residues, one on each helix. These contact residues were chosen for their proximity to the center of each predicted helix and the likelihood of being buried away from solvent as predicted by the PHD server (see Methods). The accuracy of the solvent accessibility prediction is assessed in Table 1 as the number of predicted residues that were buried in the native structure (Hubbard & Thornton, 1993). Given the assigned contact residues on each helix, corresponding distances were then measured in the native structures. For each target, the number of native inter-helical distances that did not fall within the specified bounds is also listed in Table 1. Out of the total 61 distances, 40 were within 11 Å and 48 within 14 Å; all were greater than 5 Å.

The suitability of the generic distance bounds may also be assessed by counting pairs of helices for which any residue on one helix was within 5-11 Å of any residue on the other helix. For the native structures with three helices, all pairs of helices had some inter-helical distance between 5-11 Å. This was generally not the case for the four and five-helical native structures. Only 1pou had all its pairs of helices within these bounds. Next, for each native structure we queried if there existed any set of residues, one on each of the helices, such that every inter-helical distance specified by the set was within 5-11 Å. We found many such sets for all the three-helical targets, but only 1pou had any sets out of the four- and five-helical targets.

Generation of the fold libraries

For each target protein in the set, 500 structures were generated by the distance geometry protocol described in Methods. The accuracy of the method is assessed by the average RMSD, the minimum RMSD, and the number of folds within a series of RMSD cutoffs (Table 2). We generated structures between 3 and 4 Å of the target in four cases (1c5a, 1fc2, 1hdd-C, and 1trl-A) and structures within 5 Å for seven cases (Table 2). The average minimum RMSD over all 11 targets was 4.53 Å. Near-native representations for all 11 proteins were achieved after 500 trials, with 1aca being the most difficult case at 6.17 Å. The frequency of structures

Table 2. Fold libraries generated by distance geometry

PDB	⟨RG⟩	RGnat	RMSD range	Mean	<4	<5	<6	<7	<8	Log-odds
1aca	13.16	12.21	6.17-15.43	10.97	0	0	0	8	34	−5.01
1c5a	11.54	10.88	3.77-13.00	8.38	1	21	84	145	203	−6.14
1fc2	10.34	9.40	2.59-10.77	6.64	53	134	211	256	340	−5.56
1hdd-C	12.11	11.03	3.78-12.52	7.76	4	27	86	204	311	−5.52
1nkl	12.52	10.94	5.72-12.99	9.59	0	0	6	33	86	−4.99
1pou	12.05	10.91	4.67-13.11	9.35	0	2	12	39	108	−5.61
1r69	10.00	10.06	4.56-11.96	8.90	0	1	14	47	118	−5.17
1trl-A	12.33	10.89	3.88-13.78	7.66	2	29	104	213	325	−5.79
1xbl	12.27	13.43	5.57-13.48	10.05	0	0	2	10	55	−4.96
2utg-A	11.98	12.76	4.02-14.23	9.58	0	10	32	64	120	−6.22
4icb	11.88	11.33	5.09-14.11	9.60	0	0	8	37	96	−5.50

generated within 7 Å RMSD ranged from 2% in 1aca to over 50% in 1fc2, with an average of about 20%. Overall, the radii of gyration were slightly greater than the native on average. The folds with lowest RMSD for ten of the targets are shown in Figure 1(a)-(j).

Fold prediction by score selection

We used a hybrid energy function (see Methods) to select a near-native fold from each fold library. Prior to scoring, all-atom models were constructed using the C^{α} traces taken directly from the distance geometry procedure using the program SegMod (Levitt, 1992). Final RMSD ranges are shown in Table 3. The RMSD of each top-scoring structure is shown. Also listed is the log-odds of selecting a structure with lower RMSD than the best-scoring one. A 6 Å or better structure was selected in five out of ten cases (no such structure existed for 1aca); these were 1c5a, 1fc2, 1hdd-C, 1trl-A, and 4icb. A sixth target (1r69) was reasonably near-native with RMSD of 6.38 Å. In only one case (1xbl) did the function perform worse than chance expectation, or log-odds greater than −0.3. Also listed are the correlation coefficients between score and RMSD, which are usually weakly positive. A representative scatter plot (4icb) is shown in Figure 2.

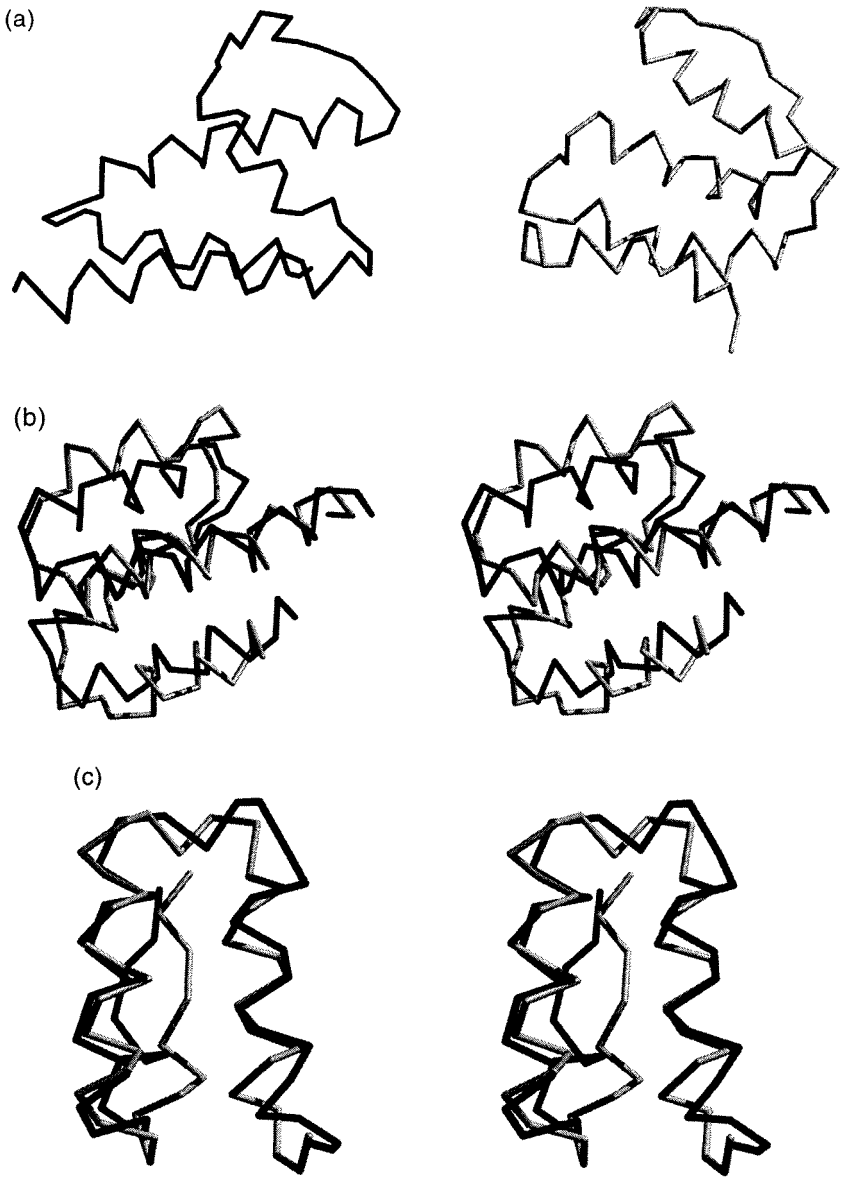

Figure 1 (*Legend shown on page 272*)

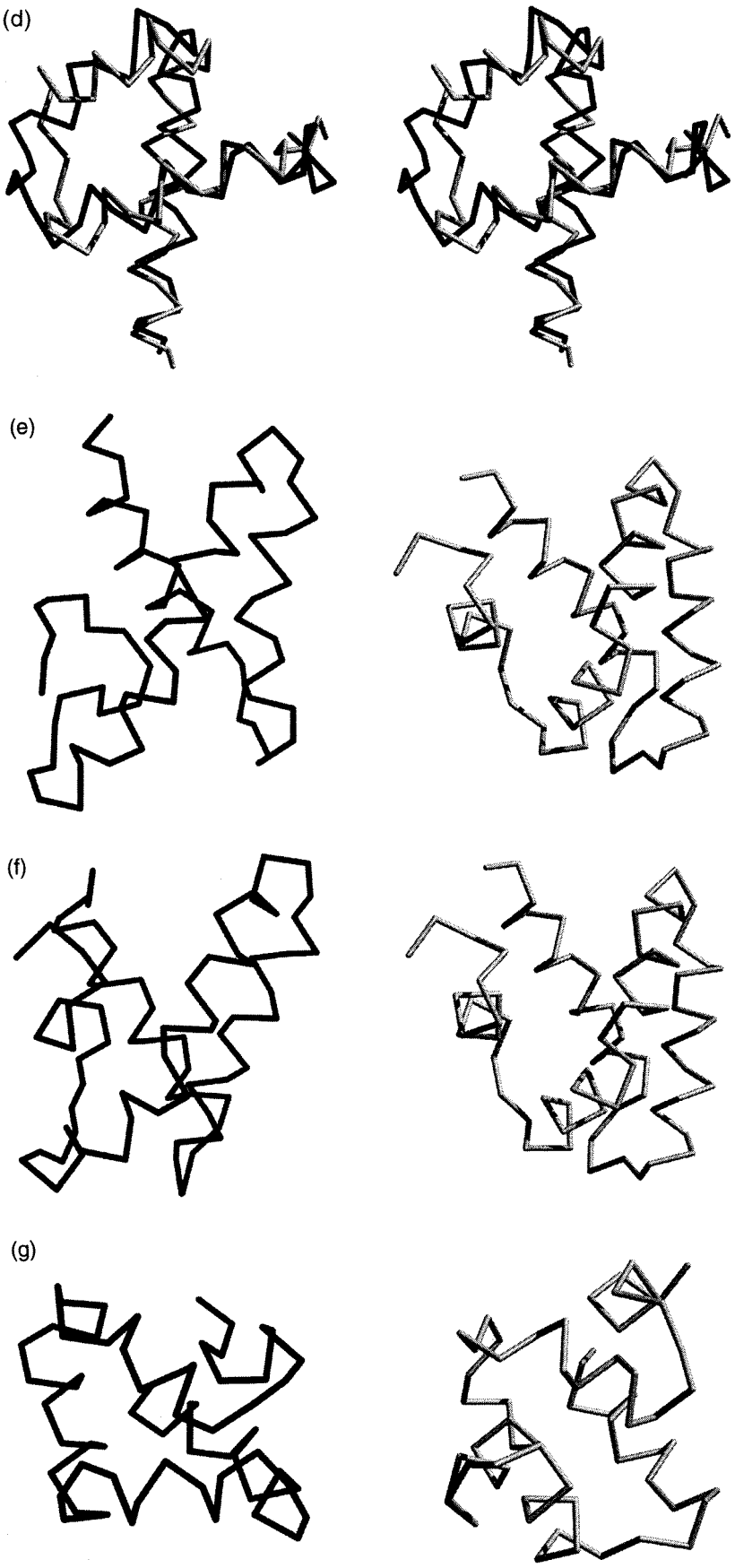

Figure 1 (*Legend shown on page 272*)

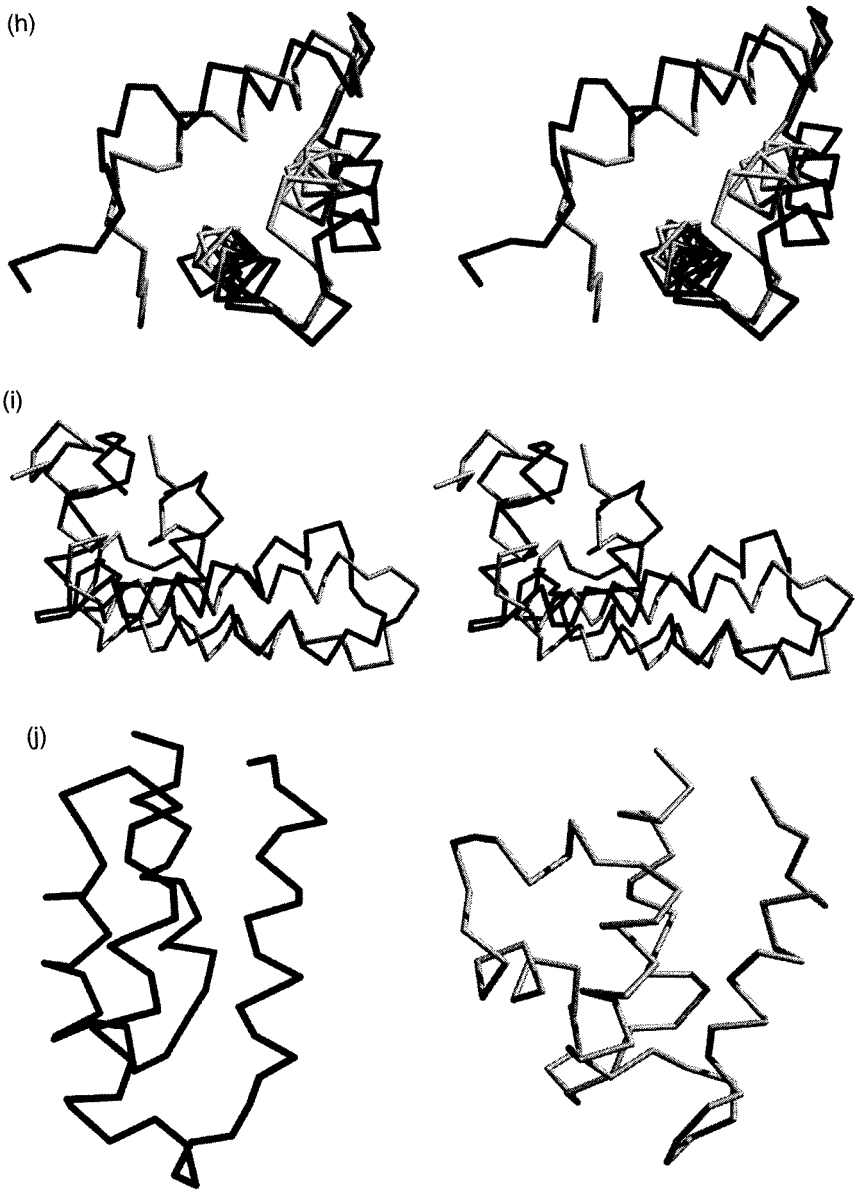

Figure 1. Best distance geometry models. For each Figure, the distance geometry model is drawn in dark grey and the experimentally determined structure in white. (b)-(d) and (h)-(i) are stereo pairs, while the remaining panels are shown side by side. (a) Bovine acyl-coenzyme A binding protein (1aca). (b) Porcine Des-Arg74-Complement C5a (1c5a). (c) Protein A (1fc2-C). (d) *Drosophila melanogaster* homeodomain (1hdd-C). (e) Porcine NK-lysin (1nkl). (f) Porcine NK-lysin using disulfide bonds. (g) Phage 434 repressor (1r69). (h) *Bacillus thermoproteolyticus* thermolysin fragment (1trl-A). (i) *Escherichia coli* DnaJ chaperone (1xbl). (j) Bovine calbindin D9 K (4icb). The MidasPlus software system from the Computer Graphics Laboratory, University of California, San Francisco (Ferrin *et al.*, 1988) was used for all images.

Fold prediction by consensus distance geometry

Consensus distance geometry was used to generate a structure for each target (Table 4). Details for the procedure are found in Methods and other published work (Huang *et al.*, 1998). Structures within 6 Å were generated for five targets: 1fc2, 1hdd-C, 1r69, 1trl-A, and 4icb (Figure 3(a), (b), (d)-(f)). A sixth model, 1nkl, with RMSD of 6.53 Å, is also reasonably native-like. It is noteworthy that these six models corresponded to those cases for which the correlation coefficients were the six highest. The lowest correlation coefficient for a successful target was 0.234 for 4icb. Consensus distance geometry computed structures near-native structures despite the relatively sparse sampling of conformations near the native fold (Tables 2 and 3). For instance, the fraction of the libraries within 6 Å RMS of the respective native fold was approximately 3 % for 1r69, 2 % for 4icb, and 1 % for 1nkl.

Table 3. Folds selected by the hybrid energy function

PDB	RMSD range	RMSD	Select	Log-Odds	CC
1aca	6.07-15.45	10.89	7.61	−1.469	0.109
1c5a	3.67-12.88	8.26	4.63	−1.585	0.172
1fc2	2.60-10.81	6.58	5.33	−0.471	0.321
1hdd-C	3.64-12.34	7.63	4.27	−1.745	0.415
1nkl	5.60-13.27	9.71	9.43	−0.389	0.252
1pou	4.58-13.01	9.26	7.71	−0.721	0.207
1r69	4.58-12.24	8.91	6.38	−1.252	0.277
1trl-A	3.73-13.52	7.53	3.84	−2.222	0.414
1xbl	5.49-13.41	9.99	10.37	−0.281	0.029
2utg-A	3.88-14.22	9.48	9.26	−0.418	0.152
4icb	4.93-14.03	9.55	5.07	−2.398	0.234
Avg	4.43-13.20	8.89	6.71	−1.177	0.235

Inclusion of disulfide-bonding information

To test the effect of correct disulfide-bonding information on the structure generation and selection on our methods, we submitted the appropriate information as covalently bonded atoms in 1nkl and 1c5a. It is noteworthy that protein 1nkl was a target for the second meeting on the Critical Assessment of protein Structure Prediction methods (CASP2), and that the disulfide bond pairs were made available to the predictors. The results for library construction and fold prediction are shown in Table 5.

For 1c5a, the combination of accurate secondary structure prediction and three disulfide bonds yielded structures of 4 Å RMS or better in nearly a third of the library. The highest ranking structure with respect to the combination function exhibited a 5.44 Å RMS error, and the consensus distance geometry structure improved to 4.79 Å.

The prediction of 1nkl was once again hampered by poor secondary structure assignment, but did improve markedly with the use of disulfide bonds. The best-scoring structure and consensus distance geometry structure had RMS errors of 5.91 Å and 4.61 Å, respectively. Figure 3(c) shows that correctly positioned disulfide bonds can partially compensate for errors in secondary structure prediction. For example, the long third helix is bent by the disulfide bonds in the best predicted structure, causing it to more closely resemble the native fold which has two helices over the same residue range.

Discussion

Quality of fold libraries

We have shown that distance geometry is an effective way to generate accurate representations of small helical proteins. By deliberately specifying an under-determined distance matrix using a few reasonable restraints, the procedure is able to suggest an arbitrary number of plausible alternatives, of which some are native-like. The method is very efficient: after 500 trials, we find on the order of a 100 near-native folds for our three-helical targets. For our four and five-helical targets, the concentration is somewhat less. This is partly due to the inverse relationship between protein size

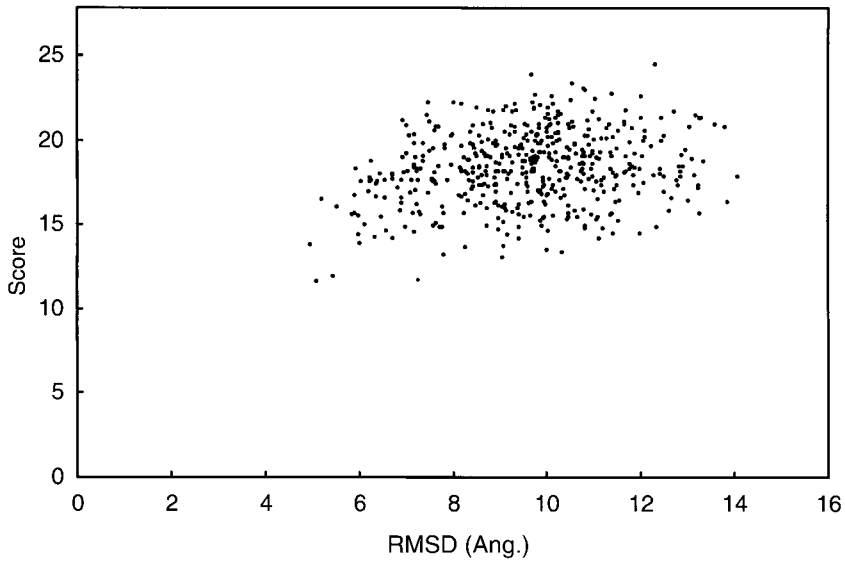

Figure 2. Plot of score *versus* coordinate RMS error. The score of a given conformation is weakly correlated with its C^{α} RMS error relative to the native fold. Here the plot for 4icb is shown (correlation coefficent = 0.23, 500 folds). The relationship is more pronounced for conformations within ~8 Å from the target.

Table 4. Consensus distance geometry results

PDB	RMSD
1aca	8.55
1c5a	7.92
1fc2	3.93
1hdd-C	3.84
1nkl	6.53
1pou	9.13
1r69	5.22
1trl-A	4.04
1xbl	9.99
2utg-A	7.11
4icb	5.37
Avg	6.51

Consensus distances for use in distance geometry were generated by Boltzmann weighting ($kT = 10$) the distances from the top-50 scoring folds in each library.

and the likelihood of generating structures within a given RMSD cutoff. Table 2 shows the log-odds of a chance prediction of the structure with lowest RMSD for each target, computed as log $(1/N_R)$, where N_R is the estimated number of protein-like structures of a particular size required to find one structure at RMSD cutoff R. N_R is computed as equation (4) by Reva *et al*. (1998):

$$N_R = \frac{(\sigma\sqrt{2\pi})}{\int_{-\infty}^{R} \exp(-(x - \langle R\rangle)^2/2\sigma^2)dx}$$

where $\langle R\rangle$ is the mean of the RMSD distribution and σ is its standard deviation. Following Reva *et al*. (1998), we set $\sigma = 2.0$ and $\langle R\rangle = 3.333N^{1/3}$, where N is the number of residues in the protein. The log-odds that are observed (-5 to -6) correspond to values of N_R between 1×10^5 and 1×10^6 random structures.

The quality of the folded structures is perhaps even more striking when one considers that a significant fraction of the native inter-helical distances were not strictly obeyed by our models. Many of the native folds had pairs of helices for which no inter-C^α atom was in the generic range of 5-11 Å. When such distances did exist, errors can still arise from the choice of contact residues, caused perhaps by inaccurate solvent accessibility predictions or the assumption that the contact residue should be centered in each predicted helix. For example, the inter-C^α atom distance between residues 10 and 50 should be 18.2 Å for the protein 4icb. An examination of the 500 models of 4icb revealed that the maximum distance was 12.4 Å, and the distance in the best structure of the library (5.09 Å RMSD) was 11.3 Å. Clearly the method is sufficiently robust so as to overcome errors in the assigned distance ranges.

In some cases the sampling was clearly inadequate, for example in the case of 1xbl. The native fold of 1xbl, unlike most of the other folds tested in this work, is rather elongated, built from a pair of two short helices and a pair of long helices (Figure 1(i)). The radius of gyration of this protein is 13.43 Å, the largest in the set of 11 proteins, even though it is exceeded in length by 1aca, 4icb, and 1nkl. Visual inspection of the C^α trace shows that the helices that are not as closely packed in 1xbl as in the other targets, and three of the six predicted inter-residue distances exceed 11 Å in the native structure (Table 1).

Limitations of the method

Even though the method presented here is very promising as a library generation method, its scope is limited to small helical proteins. Regarding the size limitation, one certainly cannot assign a fixed generic distance between every helical pair for proteins much larger than 100 residues. A medium-sized protein such as myoglobin, for example, has inter-helical distances of 20 to 30 Å. Although in principle distance geometry can handle any upper bound, it still remains to be seen whether a near-native structure would result after a reasonable number of trials. We are now searching for patterns in the inter-helical distances of larger native proteins in order to bias the distance selection towards productive combinations. The second limitation to our method is its dependence on the accurate prediction of helices. Indeed the Q3 accuracy for ten targets (all but 1nkl) was excellent overall. The nature of the errors is illustrative: rather than misassign helical residues to a β-strand, nearly all of the mispredictions were shifts in the helical boundaries relative to their correct positions. For instance, the N terminus of the third helix of 1aca was shifted by five residues, and shifts of three residues or more were tolerated in four other cases. We are further encouraged by our experience with 1nkl, which suggests that the method can even approximate a five-helical protein with only four secondary structure elements, one of which is a strand.

The corresponding problem for β-rich proteins is more difficult, not only because secondary structure prediction is less reliable, but also because of the complex geometries arising from various sheet topologies. Towards this end, Chelvanayagam *et al*. (1998) describe a combinatorial distance-based approach to set the inter-residue distances found in

Table 5. Fold libraries with disulfide bond information

PDB	RMSD range	Mean	<4	<5	<6	<7	<8	Select	cDG
1c5a	2.62-11.37	5.87	152	210	262	317	408	5.44	4.79
1nkl	4.21-12.98	7.58	0	27	124	223	319	5.91	4.61

Data reflect the folds in the library after side-chain construction, energy minimization, and scoring.

possible sheet topologies. Moreover, the recent report by Zhu & Braun (1999) suggests that specialized residue pair potentials for contacts across β-strands might allow improved discrimination of correct strand pairings. It would be interesting to see if such approaches would allow the construction of a greater variety of folds. An alternative to our fixed 5-11 Å distance range between helices would be to allow use of a general probability distribution over all possible distances for certain key distances. A fold library would then be able to explore the combinatorial sets of mutually exclusive distances needed to handle β-sheet and larger helical structures within the present protocol. A formalism for including such ''disjunctive'' constraints within a probabilistic least-squares algorithm is outlined in a recent report by Chen *et al*. (1999). Their ideas could be incorporated into distance geometry *via* a preprocessing step that chooses key distances from distance probability distributions before proceeding with selection of the remaining trial distances.

Comparison to other fold libraries

The quality of the library is greatly enhanced compared to earlier work, both in terms of the low-end RMSD and the concentration of near-native folds. We first compare our results to those of two different library methods prior to any score-based filtering (Hinds & Levitt, 1994; Park & Levitt, 1996). In the study by Hinds & Levitt (1994), models were exhaustively enumerated on tetrahedral lattice to an average minimum RMSD of 5.4 Å (Park & Levitt, 1995). Two targets, 4icb and 1r69, overlapped with our test set. For 1r69, the number of compact, self-avoiding structures was 8.0×10^5. In this library, there were 20 structures with distance matrix error (DME) of 3.46 Å or better. Our 1r69 library had 1 structure out of 500 within this cutoff, corresponding to a sampling efficiency that is roughly 80 times better. In the lattice library of 4icb, there were 20 structures out of 1.5×10^6 within 3.92 Å DME. After measuring the DME values in our library, we found three folds out of 500, a sampling efficiency increase of 450-fold. However, the tetrahedral lattice model has the distinct advantage that it may be applied to targets irrespective of secondary structure class and size (though exhaustive enumeration has only been tested for proteins of ~100 residues). In exchange for general applicability, our method returns accuracy and sampling efficiency. The comparison of our library to that of Park & Levitt (1996) is less straightforward. Both methods depend on the enforcement of rigid secondary structures, but the Park & Levitt results reflect correct secondary structure assignments and optimized ϕ/ψ states for selected loop residues. Nevertheless, we note that in the Park & Levitt (1996) 1r69 library, there were 301 structures out of 199,943 within 4.5 Å RMSD. Our library contained one structure at 4.56 Å RMSD, suggesting that the sampling efficiencies are roughly equal. If one takes into consideration that we pre-filter our structures for mirror images, the effective sampling efficiency of the Park & Levitt method is nominally about twofold higher. However, the extent to which the inherent sampling efficiency is enhanced by correct secondary structures and torsion optimization is unclear. Finally, the computational complexity of the Park & Levitt model scales exponentially with the number of freely rotatable residues. As discussed in the Introduction, this is an inherent limitation on combinatorial approaches to library construction of targets of 100 residues or less.

Next, we compare our results to those by Simons *et al*. (1997). This method is a minimization-based approach to protein structure prediction, but rather than reporting a single predicted structure for each target, the authors elected to describe the accuracy of the method in terms of sets of 500 folds. These sets effectively served as libraries from which different scoring methods can select final predicted structures (Simons *et al*., 1997; Huang *et al*., 1998). Three targets from the Simons *et al*. library overlapped with ours: 1fc2, 1hdd-C, and 4icb. In the Simons *et al*. libraries for 1fc2 and 1hdd-C, there were 41 and eight folds, respectively, within 4 Å RMSD. For 4icb, there were two folds within 5 Å and three within 6 Å RMSD. The minimum RMSD values for the three libraries were 3.16 Å, 2.75 Å, and 4.86 Å, for 1fc2, 1hdd-C, and 4icb, respectively. At face value, these figures of merit closely match those reported here. Again, the pre-selection of the correct mirror image effectively makes our method nominally less efficient by a factor of two, though the score-biased nature of the Simons *et al*. search strategy undoubtedly enriches the sets for near-native folds. In contrast, our library is essentially a collection of compact, self-avoiding arrangements of helices, assembled without regard to statistically probable contacts or packing angles.

Performance of the scoring function

The hybrid scoring function used to select the best-ranking fold as the predicted structure was moderately successful: six of the 11 best-scoring folds might be considered to be near-native. The same scoring function played an integral role in the consensus distance geometry approach, which also yielded six near-native structures of somewhat higher quality. Over the five near-native structures in common between the best-scoring set and the consensus distance geometry set (1fc2, 1hdd-C, 1r69, 1trl-A, 4icb) the mean RMSD fell from 4.98 to 4.48 Å). In all five cases in which the distance geometry structure was unable to project a near-native structure, the correlation between score and RMSD was weak, leading to the incorporation of incorrect distances. There are a few reasons for the variability in correlation coefficient. First, errors in the structure generation protocol result in a lower

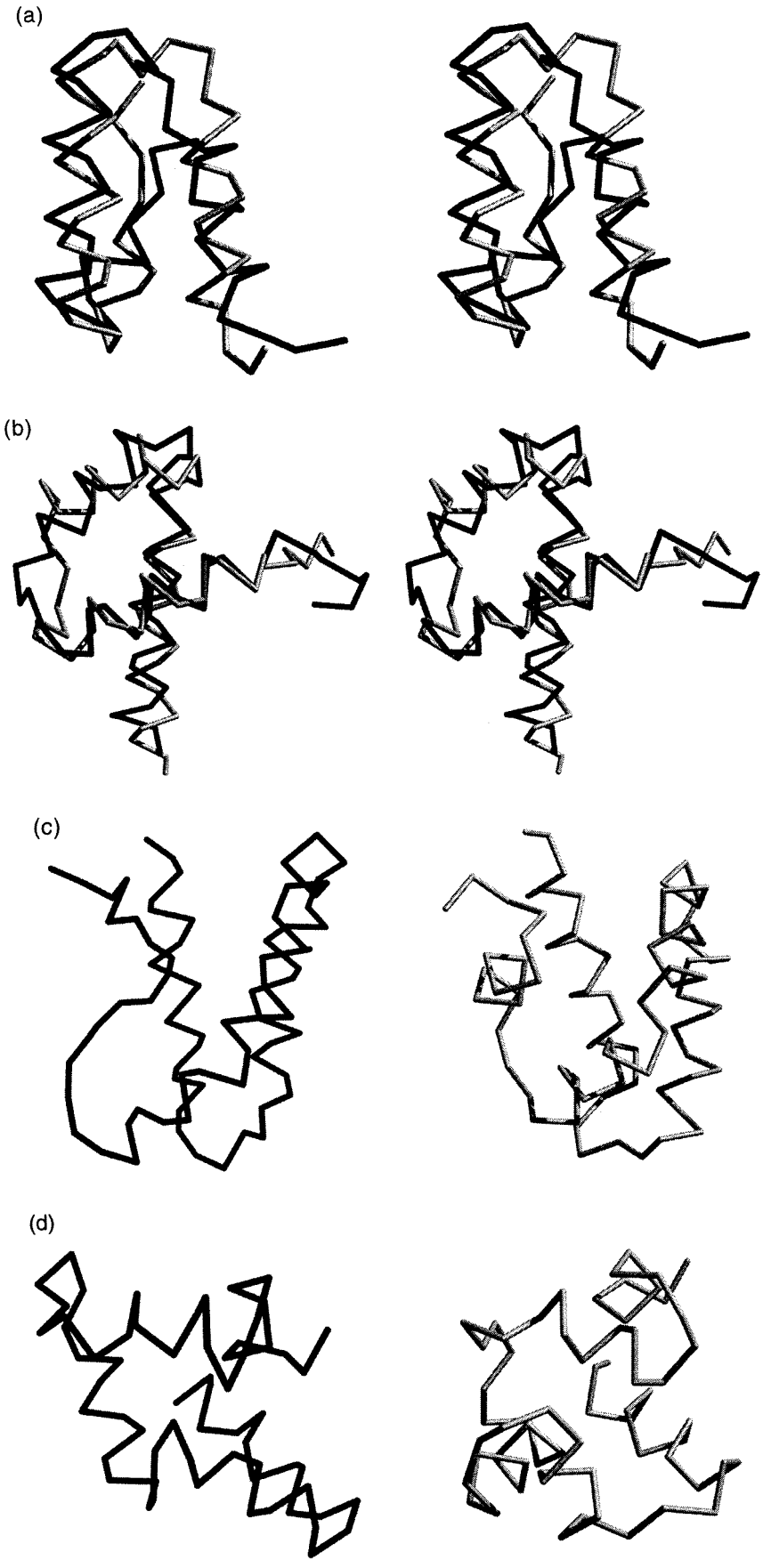

Figure 3 (*Legend opposite.*)

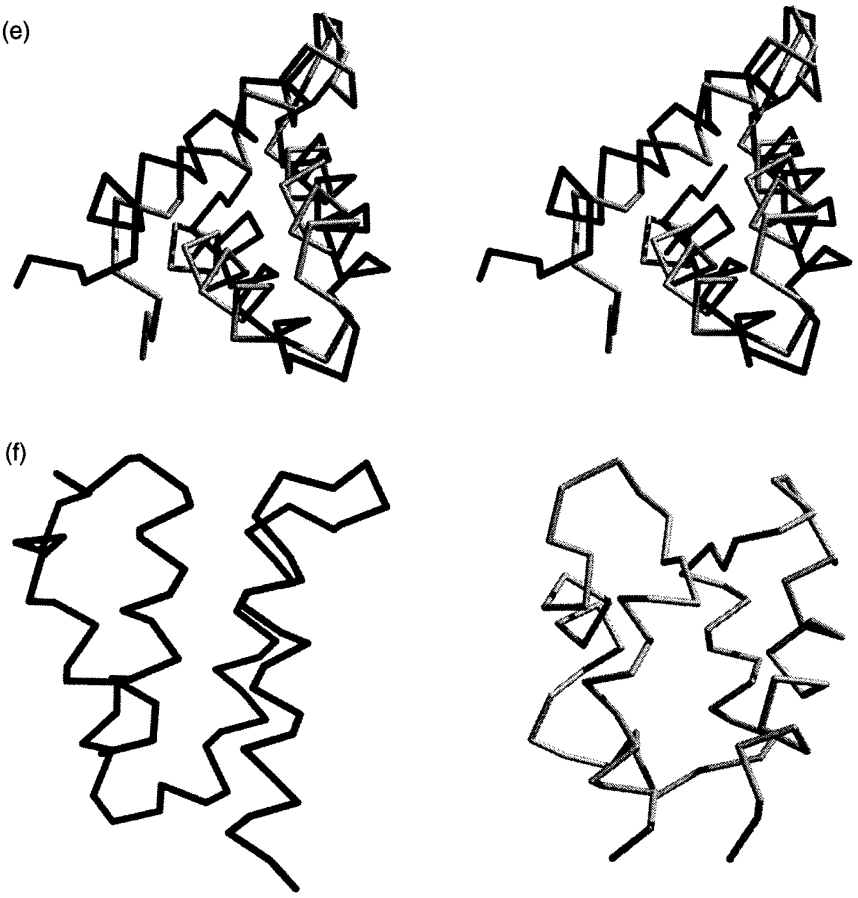

Figure 3. Consensus distance geometry models. The structures were modeled from the consensus inter-C^{α} distances from the top-scoring 10 % subset in each fold set. Each fold was assigned its Boltzmann weight ($kT = 10$) before the consensus was taken. Model structures are depicted in dark grey and the respective native structures in white. (a), (b) and (e) are stereo pairs, while the remaining panels are shown side by side. The PDB identifier for each Figure: (a) 1fc2-C; (b) 1hdd-C; (c) 1nkl, using disulfide bonds; (d) 1r69; (e) 1trl-A; and (f) 4icb.

density of near-native states. Since the correlation between score and RMSD tends to improve at lower RMSD ranges, the overall correlation coefficient suffers when sampling is poor. For instance, the failures of 1aca and 1xbl were related in part to the relative paucity of low RMSD structures. Second, the target in question may not be well-suited for the scoring function. Exposed hydrophobic interfaces in a native structure generally result in less favorable scores, since knowledge-based functions typically derive much of their discriminatory power from hydrophobic contacts (Huang *et al*., 1995; Thomas & Dill, 1996). In our set of targets, 2utg is a dimer in solution and presents a hydrophobic interface.

Comparison with other *ab initio* results

Here, we compare our final predicted structures with those reported in the literature. We restrict our discussion to those methods which were able to report a near-native fold as the single best-scoring structure.

The 1nkl model by Jones (1996) represents the ''blind'' prediction at CASP2 that had the lowest C^{α} RMSD error (6.2 Å). The location of the three disulfide bridges was used for this *bona fide* prediction. While we generated a 4.6 Å model with the aid of disulfide-bonding patterns (6.5 Å without disulfide bonds), it must be stressed that we did not participate in the blind prediction experiment. However, we have tested this method at the third meeting on the Critical Assessment of protein Structure Prediction (CASP3: http://predictioncenter.llnl.gov/casp3/Casp3.html; see below).

A recent report by Ortiz *et al*. (1998) reports the successful folding of several targets presented here. After multiple simulations, the fold with the lowest energy was chosen as the final model. The RMSD values were 3.1 Å for 1fc2, 3.5 Å for 1pou, 4.2 Å for 1c5a, 4.5 Å for 3icb, and 5.6 Å for 1nkl, the last of which included disulfide information. These results are somewhat better than the final structures generated by our method, though the best structures in the respective libraries were very close to the final structures reported by Ortiz *et al*. (Table 3). In other words, an ideal scoring function would have selected structures with RMSD 2.60 Å for 1fc2, 4.58 Å for 1pou, 3.67 Å for 1c5a, 4.93 Å

for 4icb, and 5.60 Å for 1nkl. We also note that the method by Ortiz *et al*. (1998) can only be applied to targets for which at least ten homologous sequences are available in the HSSP database (Sander & Schneider, 1991). Additional predicted distance constraints used by Ortiz *et al*. (1998) were obtained after the application of a threading protocol. Although native-like folds for our small helical proteins were constructed without the assistance of a structural database, it would nevertheless be interesting if one can enrich the near-native fraction in our libraries by a similar knowledge-based technique. Ortiz *et al*. also outline a method for predicting disulfide bonds when the correct assignments are not available. Our approach can easily express predicted disulfide bridges as additional distance constraints, and when this information is correct, the near-native fold fraction increases accordingly, leading directly to better predicted folds (Table 5).

Summary of blind prediction results

We applied our method to five small targets for the CASP3 experiment: T0065 (31 residues, SinI protein), T0056 (114 residues, DnaB helicase), T0061 (76 residues, protein HDEA), T0079 (96 residues, MarA protein), and the first 75 residues of T0083 (cyanase). These sequences did not have any detectable sequence similarity with proteins of known structure, and were predicted or known to be all-helical. We submitted up to five models for each target using the methods described here. One target (T0065) has a simple alpha-hairpin structure; our model was correct to 3.8 Å. Two targets (T0079, T0083) had structural analogs in the database which were detectable by threading. For T0079, our best model comprised 70 contiguous residues with RMSD of 5.7 Å (11.4 Å overall), and the best model of T0083 had an RMSD of 5.4 Å over 60 contiguous residues (8.7 Å over all 75 submitted residues). The best model for T0061 was 6.7 RMSD from the native over 60 contiguous residues (9.8 Å overall). None of the submitted models of T0056 were native-like.

Conclusions

We have described a new method of generating fold libraries for a variety of helical proteins using distance geometry. Despite the simplicity of the method, in which only predicted secondary structures and generic inter-helical distances were specified, we were able to generate at least one native-like conformation for each target. In most cases, many near-native folds were generated. There appear to be two critical factors for successful fold prediction. The first is the adequate sampling of conformational space near the target fold. Although our minimalist approach is sufficient in many cases, it is also extensible, as it readily uses distance information gathered from disulfide bonds, multiple sequence alignment, and mutagenesis experiments. The second is the tendency for a scoring function to score the native-like folds as a group more favorably than the non-native folds. In most cases we observe that increased sampling near the native conformation enhances the performance of our knowledge-based function in this regard, but this was not the case for every protein. Overall, we were very encouraged by the results on small helical proteins, and are currently testing methods to extend the method to larger systems.

Methods

Prediction of secondary structure and contact residues

The sequence for each target was submitted to the PHD PredictProtein Server (http://www.embl-heidelberg.de) (Rost *et al*., 1993). Helix boundaries were those predicted by the PHDsec profile, and no manual adjustment was made other than to break a helix in 1pou. The third and fourth helices in 1pou were joined by the PHD prediction; a break in this helix was created where the helical prediction was less reliable at residue 51. For prediction of solvent accessibility, we considered the high-confidence SUBacc profile. If the predicted helix did not contain any residues that satisfied the burial criteria for SUBacc, then we considered the P_3acc profile for prediction of solvent accessibility. The contact residue was the residue predicted to be buried located nearest the center of each helical segment.

Distance geometry

All residues other than Gly and Pro were converted to Ala prior to distance geometry calculations. Distance geometry calculations were performed with the program distgeom from the TINKER suite using 10% random pairwise metrization (Hodsdon *et al*., 1996). Trial distances were selected from approximately Gaussian distributions between the lower and upper bounds. The center of the distribution between the upper and lower bounds is a function of the number and type of input restraints and is consistent with the expected radius of gyration of the structure. Following metrization, embedding and majorization, the generated structure is refined *via* 10,000 steps of simulated annealing against a set of penalty functions which enforce local geometry, chirality, excluded volume, input distance restraints, and torsion restraints. Prior to simulated annealing, the global enantiomer closest to the target structure was selected for refinement. This does not reduce the generality of our method, since both enantiomers could have been subjected to annealing at a cost of additional CPU time. Over the residues predicted to be helical, two types of restraints were specified: a virtual torsion angle defined by four consecutive C^{α} atoms, constrained to be between

40° and 60°, and intra-helical distances measured from a canonical alpha-helix with $\phi = -57°$ and $\psi = -47°$. This process was repeated until 500 structures with right-handed helices were generated.

All-atom model construction

The software SegMod (Levitt, 1992) was used to restore the full amino acid sequence to the C^α trace built from distance geometry. Each structure was subjected to 200 steps of energy minimization using ENCAD (Levitt *et al*., 1995) prior to scoring.

Scoring function

For evaluating sequence to structure compatibility, we used a hybrid scoring function that combined scores from three distinct functions: an all-atom distance-dependent conditional probability function, a hydrophobic compactness function, and an inter-residue contact function (shell). The final score was computed by summing the three components after dividing each by its respective standard deviation calculated over 500 conformations.

Residue-specific all-atom probability discriminatory function (RAPDF)

The all-atom scoring function, RAPDF, was used to calculate the probability of a conformation being native-like given a set of inter-atomic distances (Samudrala & Moult, 1998). The conditional probabilities were compiled by counting frequencies of distances between pairs of atom types in a database of protein structures. A set of 312 unique folds from the SCOP database (Hubbard *et al*., 1997) was used. All non-hydrogen atoms were considered, and a residue-specific description of the atoms was used, i.e. the C^α trace of an alanine residue is different from the C^α trace of a glycine residue. This resulted in a total of 167 atom types. The distances observed were divided into 1 Å bins ranging from 3 Å to 20 Å. Contacts between atom types in the 0-3 Å range were placed in a separate bin, resulting in a total of 18 distance bins. Distances within a single residue were not included in the counts.

We compile tables of scores s proportional to the negative log conditional probability that we are observing a native conformation given an inter-atomic distance d for all possible pairs of the 167 atom types, a and b, for the 18 distance ranges, $P(C|d_{ab})$:

$$s(d_{ab}|C) = -\ln\frac{P(d_{ab}|C)}{P(d_{ab})}\alpha - \ln P(C|\{d_{ab}^{ij}\})$$

where $P(d_{ab}|C)$ is the probability of observing a distance d between atom types a and b in a correct structure, and $P(d_{ab})$ is the probability of observing such a distance in any structure, correct or incorrect. The required ratios $P(d_{ab}|C)/P(d_{ab})$ are obtained as follows:

$$\frac{P(d_{ab}|C)}{P(d_{ab})} = \frac{N(d_{ab})/\sum_d N(d_{ab})}{\sum_{ab} N(d_{ab})/\sum_d \sum_{ab} N(d_{ab})}$$

where $N(d_{ab})$ is the number of observations of atom types a and b in a particular distance bin d, $\Sigma_d N(d_{ab})$ is the number of a-b contacts observed for all distance bins, $\Sigma_{ab} N(d_{ab})$ is the total number of contacts between all pairs of atoms types a and b in a particular distance bin d, and $\Sigma_d \Sigma_{ab} N(d_{ab})$ is the total number of contacts between all pairs of atom types a and b summed over all the distance bins d.

Given a set of distances in a conformation, the probability that the conformation represents a ''correct'' fold was evaluated by summing the scores for all distances and the corresponding atom pairs. A complete description of this formalism has been published elsewhere (Samudrala & Moult, 1998).

Hydrophobic compactness function (HCF)

The hydrophobic compactness function (HCF) score for a given conformation is calculated using the formula:

$$HCF = \frac{\sum_i^N (\bar{x} - x_i)^2 + (\bar{y} - y_i)^2 + (\bar{z} - z_i)^2}{N}$$

where N is the number of carbon atoms in the protein, and x, y, and z are the three-dimensional coordinates of those atoms. This measure is the square of the radius of gyration of the carbon atoms.

Residue-residue contact function (shell)

The shell scoring function is described in detail elsewhere (Park *et al*., 1997). Briefly, it is a simple pairwise contact function with the form:

$$E = \sum_{i=1} \sum_{j>i+1} e_{ij}^{ab}$$

where e is the contact score for residues i and j of types a and b, respectively. $e_{ij}^{ab} = e^{ab}$ if $d_{ij} < 7.0$ Å and zero otherwise. All inter-residue distances d_{ij} were measured from an interaction center located 3 Å from the C^α atom along the C^α-C^β vector:

$$e^{ab} = -\ln n_{obs}^{ab}/n_{exp}^{ab}$$

where n_{obs}^{ab} is the number of residue types a and b within 7 Å in a database of proteins. n_{exp}^{ab} is the number of contacts expected in a random mixture of residue types in the database:

$$n_{exp}^{ab} = \sum_p C_p \frac{R_p^{ab}}{\frac{1}{2}(N_p - 2)(N_p - 1)}$$

For each protein p, C_p is the total number of contacts, R_p^{ab} is the number of residue pairs of type a and b separated by at least two residues in the sequence, and N_p is the number of residues.

Consensus distance geometry

A single Cartesian structure consistent with the most frequently observed inter-C^α atom distances in each low-energy subset was computed with the program distgeom of the TINKER suite (http://dasher.wustl.edu/tinker/). The inter-C^α distances for the 50 top-scoring folds are measured and saved in 1 Å bins. The upper and lower bounds for each distance was determined by jury process, in which each distance received a weight equal to the Boltzmann weight of the conformation from which it was measured, i.e.:

$$W_i = \frac{\exp(-E_i/kT)}{Q}$$

where E is the score of structure i, and Q is the partition function:

$$Q = \sum_i \exp(-E_i/kT)$$

Here, kT is set to 10. The distance bin that received the most weighted votes was used to set the constraints for distance geometry. Additional details are published elsewhere (Huang *et al*., 1998; Samudrala *et al*., 1999).

Calculation of accessible surface area

We used the software NACCESS (Hubbard & Thornton, 1993) to compute the relative solvent accessibility for the predicted contact residues in each native structure. We consider those side-chains with relative solvent accessibility of 25 % or less as buried (nBur; Table 1).

Acknowledgments

This work was supported by Department of Energy grant DE-FG07-96ER14693 to J.W.P. and a grant from The Jane Coffin Childs Memorial Fund for Medical Research. E.S.H. is a Fellow of The Jane Coffin Childs Memorial Fund for Medical Research. R.S. acknowledges support from the Burroughs Wellcome Fund/Program in Mathematics and Molecular Biology.

References

Aszodi, A., Gradwell, M. J. & Taylor, W. R. (1995). Global fold determination from a small number of distance restraints. *J. Mol. Biol.* **251**, 308-326.

Bowie, J. U. & Eisenberg, D. (1994). An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl Acad. Sci. USA,* **91**, 4436-4440.

Chelvanayagam, G., Knecht, L., Jenny, T., Benner, S. A. & Gonnet, G. H. (1998). A combinatorial distance-constraint approach to predicting protein tertiary models from known secondary structure. *Fold. Design,* **3**, 149-160.

Chen, C. C., Singh, J. P. & Altman, R. B. (1999). Using imperfect secondary structure predictions to improve molecular structure computations. *Bioinformatics,* **15**, 53-65.

Cohen, F. E., Richmond, T. J. & Richards, F. M. (1979). Protein folding: evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example. *J. Mol. Biol.* **132**, 275-288.

Covell, D. G. (1992). Folding protein alpha-carbon chains into compact forms by Monte Carlo methods. *Proteins: Struct. Funct. Genet.* **14**, 409-420.

Cui, Y., Chen, R. S. & Wong, W. H. (1998). Protein folding simulation with genetic algorithm and supersecondary structure constraints. *Proteins: Struct. Funct. Genet.* **31**, 247-257.

Dandekar, T. & Argos, P. (1996). Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions. *J. Mol. Biol.* **256**, 645-660.

Ferrin, T. E., Huang, C. C., Jarvis, L. E. & Langridge, R. (1988). The MIDAS display system. *J. Mol. Graph.* **6**, 13-27.

Havel, T. F. (1991). An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Prog. Biophys. Mol. Biol.* **56**, 43-78.

Hinds, D. A. & Levitt, M. (1992). A lattice model for protein structure prediction at low resolution. *Proc. Natl Acad. Sci. USA,* **89**, 2536-2540.

Hinds, D. A. & Levitt, M. (1994). Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.* **243**, 668-682.

Hodsdon, M. E., Ponder, J. W. & Cistola, D. P. (1996). The NMR solution structure of intestinal fatty acid-binding protein complexed with palmitate: application of a novel distance geometry algorithm. *J. Mol. Biol.* **264**, 585-602.

Huang, E. S., Subbiah, S. & Levitt, M. (1995). Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* **252**, 709-720.

Huang, E. S., Samudrala, R. & Ponder, J. W. (1998). Distance geometry generates native-like folds for small helical proteins using the consensus distances of predicted protein structures. *Protein Sci.* **7**, 1998-2003.

Hubbard, S. J. & Thornton, J. M. (1993). *'NACCESS' Computer Program*, Department of Biochemistry and Molecular Biology University College London.

Hubbard, T. J. P., Murzin, A. G., Brenner, S. E. & Chothia, C. (1997). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **25**, 236-239.

Jones, D. T. (1997). Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins: Struct. Funct. Genet.* **Suppl 1**, 185-191.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577-2637.

Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226**, 507-533.

Levitt, M., Hirshberg, M., Sharon, R. & Daggett, V. (1995). Potential-energy function and parameters for simulations of the molecular-dynamics of proteins and nucleic-acids in solution. *Comput. Phys. Commun.* **91**, 215-231.

Lund, O., Hansen, J., Brunak, S. & Bohr, J. (1996). Relationship between protein structure and geometrical constraints. *Protein Sci.* **5**, 2217-2225.

Monge, A., Lathrop, E. J., Gunn, J. R., Shenkin, P. S. & Friesner, R. A. (1995). Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models. *J. Mol. Biol.* **247**, 995-1012.

Mumenthaler, C. & Braun, W. (1995). Predicting the helix packing of globular proteins by self-correcting distance geometry. *Protein Sci.* **4**, 863-871.

Ortiz, A. R., Kolinski, A. & Skolnick, J. (1998). Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments. *J. Mol. Biol.* **277**, 419-448.

Park, B. H. & Levitt, M. (1995). The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* **249**, 493-507.

Park, B. H. & Levitt, M. (1996). Energy functions that discriminate X-ray and near native folds from well-constructed docoys. *J. Mol. Biol.* **258**, 367-392.

Park, B. H., Huang, E. S. & Levitt, M. (1995). Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* **266**, 831-846.

Pedersen, J. T. & Moult, J. (1997). Protein folding simulations with genetic algorithms and a detailed molecular description. *J. Mol. Biol.* **269**, 240-259.

Reva, B. A., Finkelstein, A. V. & Skolnick, J. (1998). What is the probability of a chance prediction of a protein structure with an rmsd of 6 A? *Fold. Design,* **3**, 141-147.

Rost, B., Sander, C. & Schneider, R. (1994). PHD - an automatic mail server for protein secondary structure prediction. *Comput. Appl. Biosci.* **10**, 53-60.

Samudrala, R. & Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **275**, 895-916.

Samudrala, R., Xia, Y., Levitt, M. & Huang, E. S. (1999). A combined approach for *ab initio* construction of low resolution protein tertiary structures from sequence. *Proc. Pac. Symp. Biocomput.*, 505-516.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56-68.

Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209-225.

Smith-Brown, M. J., Kominos, D. & Levy, R. M. (1993). Global folding of proteins using a limited number of distance constraints. *Protein Eng.* **6**, 605-614.

Srinivasan, R. & Rose, G. D. (1995). LINUS: a hierarchic procedure to predict the fold of a protein. *Proteins: Struct. Funct. Genet.* **22**, 81-99.

Sun, S., Thomas, P. D. & Dill, K. A. (1995). A simple protein folding algorithm using a binary code and secondary structure constraints. *Protein Eng.* **8**, 769-778.

Thomas, P. D. & Dill, K. A. (1996). Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* **257**, 457-469.

Wilson, C. & Doniach, S. (1989). A computer model to dynamically simulate protein folding: studies with crambin. *Proteins: Struct. Funct. Genet.* **6**, 193-209.

Zhu, H. Y. & Braun, W. (1999). Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of β-sheet formation in proteins. *Protein Sci.* **8**, 326-342.

Edited by F. Cohen

(Received 16 December 1998; received in revised form 27 April 1999; accepted 27 April 1999)