Distance geometry generates native-like folds for small helical proteins using the consensus distances of predicted protein structures

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, Missouri 63110
[2]Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305-5400

Abstract

For successful ab initio protein structure prediction, a method is needed to identify native-like structures from a set containing both native and non-native protein-like conformations. In this regard, the use of distance geometry has shown promise when accurate inter-residue distances are available. We describe a method by which distance geometry restraints are culled from sets of 500 protein-like conformations for four small helical proteins generated by the method of Simons et al. (1997). A consensus-based approach was applied in which every inter-$C\alpha$ distance was measured, and the most frequently occurring distances were used as input restraints for distance geometry. For each protein, a structure with lower coordinate root-mean-square (RMS) error than the mean of the original set was constructed; in three cases the topology of the fold resembled that of the native protein. When the fold sets were filtered for the best scoring conformations with respect to an all-atom knowledge-based scoring function, the remaining subset of 50 structures yielded restraints of higher accuracy. A second round of distance geometry using these restraints resulted in an average coordinate RMS error of 4.38 Å.

Keywords: ab initio folding; distance geometry; energy functions; protein structure prediction

How the sequence of a polypeptide determines its three-dimensional structure remains one of the most important unanswered questions in molecular biology. So-called "ab initio" computational approaches seek the overall fold of the polypeptide, often by starting with a random or extended chain and searching for the most energetically favorable, or statistically probable, conformation for the sequence. An alternative approach is first to sample conformational space as exhaustively as possible, given computational limits, then to apply a scoring function to assess the fitness of each candidate structure. In either case, reduction of the available conformational space is achieved by discretization on a lattice (Covell, 1992, 1994; Hinds & Levitt, 1992, 1994; Kolinski & Skolnick, 1994; Vieth et al., 1994) or sampling in torsion space (Wilson & Doniach, 1989; Bowie & Eisenberg, 1994; Dandekar & Argos, 1994, 1996; Monge et al., 1995; Mumenthaler & Braun, 1995; Srinivasan & Rose, 1995; Sun et al., 1995, Yue & Dill, 1996; Simons et al., 1997). However, reduction of the search space also decreases the fidelity with which the native fold can be represented. This is problematic, since knowledge-based scoring functions that recognize native folds have difficulty separating near-native folds from non-native folds (Park et al., 1997). Whereas ab initio

minimization methods can produce native-like tertiary structures, even without knowledge of the correct secondary structures, complete convergence to a native-like fold has not been achieved (Simons et al., 1997; Skolnick et al., 1997). Clearly, devising a procedure to select a single native-like fold from the pool of plausible, low-energy folds is an important hurdle remaining to be surmounted.

Distance-based techniques are a powerful way to process and optimally satisfy an overdetermined set of inter-residue and inter-atomic distances resulting in the generation of a single structure or a tightly clustered family of structures. Early work by Kuntz et al. (1979) used known disulfide bridges and simple methods to predict turns and hydrophobic contacts, coupled with minimization of distance matrix error, to generate structures as close as 5 Å root-mean-square (RMS) from the native BPTI fold. In the context of more recent ab initio folding studies, these methods have harnessed inter-atomic distance information using one of two approaches. The first is the incorporation of inter-residue distance information as terms in a variable target function to be used in a specialized buildup and minimization procedure (Hanggi & Braun, 1994; Mumenthaler & Braun, 1995) or as a term in a model force field with many components (Skolnick et al., 1997). The second approach, metric matrix distance geometry, utilizes a mathematical projection from distance space to three-dimensional space known as embedding (Havel et al., 1983) to yield a single structure con-

sistent with the distances, chiralities, and other geometric information taken as input. Using this type of algorithm, which will henceforth be referred to as "distance geometry," Taylor and coworkers determined that a subset of native inter-Cα distances, combined with the appropriate local geometry and a hydrophobicity term expressed as predicted inter-residue distances, can correctly generate native-like tertiary structures (Aszodi et al., 1995). Despite the significant advances represented in these recent studies, distance-based methods applied to ab initio structure prediction still require external information, such as the correct secondary structure or a subset of native inter-residue distances, or else fail to converge completely to a native-like fold (Aszodi et al., 1995; Mumenthaler & Braun, 1995; Skolnick et al., 1997).

In the current study we describe a general application of metric matrix distance geometry in ab initio structure prediction. Our method considers a set of structures collected from an ab initio folding protocol and derives a consensus set of inter-Cα distances from the set. Here we show that distance geometry is able to take the consensus distance information and produce a fold with accuracy better than the mean average in the set. Moreover, if there are near-native folds in sufficient concentration present in the collection, the set of consensus distances will result in a near-native fold. Consensus methods have also been used to enhance the predictions of side-chain conformation when many near-native templates are available (Tuffery et al., 1997; Huang et al., 1998).

Results and discussion

We chose as our test case the structures of Baker and coworkers (Simons et al., 1997), who employed a Bayesian scoring function to anneal the locally favored conformations of a protein chain into protein-like structures. We found this test set particularly interesting as the structures were built without knowledge of the secondary structure, yet many near-native folds were generated. Four small helical proteins were chosen as targets. Their identities, Protein Data Bank (PDB) (Bernstein et al., 1977) identifiers, size, and secondary structures follow: Protein A (1fc2:C; 43 residues, 3 helices), the homeodomain protein (1hdd:C; 57 residues, 3 helices), 434 cro repressor (2cro; 65 residues, 5 helices), and calbindin (4icb; 76 residues, 5 helices). A set of 500 independently generated structures was available for each protein. There is a clear inverse relationship between protein length and the overall quality of the predictions in each set (Table 1). The structures overall show the packing density, secondary structure elements, and solvent acces-

sibility patterns of native structure, even if the tertiary arrangements are often incorrect (Simons et al., 1997).

For the 500 proteins in each set, all inter-Cα distances were measured except for those between consecutive residues. The inter-Cα distances were stored in 1 Å bins. The most frequently occurring distance bin for each inter-residue pair was taken as the consensus value for that inter-residue distance. These constraints, along with the local bonded geometry, chiralities, and standard van der Waals atomic radii, constituted the only input for the distance geometry module in the TINKER software package (Ponder, 1998; http://dasher.wustl.edu/tinker/). Distance restraints were specified as a 1 Å range between lower and upper bounds and corresponded exactly with the consensus bins, also 1 Å in width. Details regarding the embedding algorithm, the metrization procedure, annealing protocol, and performance are presented in Materials and methods. We saved the first structure that presented the correct global chirality (e.g., the presence of right-handed helices) and compared it with the experimentally determined structure of the target protein. In all cases, further distance geometry structures produced from the same set of input restraints were essentially identical to the first structure chosen.

For all four proteins, the consensus distances yield a structure that has a coordinate RMS error less than the mean structure in the respective set of 500 structures (Table 1). In the case of 1fc2, a significant fraction (35%) of the structures is near-native, i.e., within 4 Å coordinate RMS error of the correct structure. Hence, it is perhaps not surprising that a near-native structure results from distance geometry after the consensus distances are considered. However, the example of 1hdd demonstrates that even if only 8% of the original structures contain near-native folds, these conditions are sufficient for our method to recover a near-native structure (3.78 Å coordinate RMS error). Moreover, the model constructed by distance geometry is more accurate than 95% of the structures in the set. On the other hand, in the cases where the parent structures are less accurate, such as for 2cro and 4icb, the distance geometry models also reflect the errors in the inter-residue distances. The model of 4icb resembles many of the very best structures present in the set of 500, and the correct topology of the structure is recognizable, even if it is more expanded than the native structure (Fig. 1D).
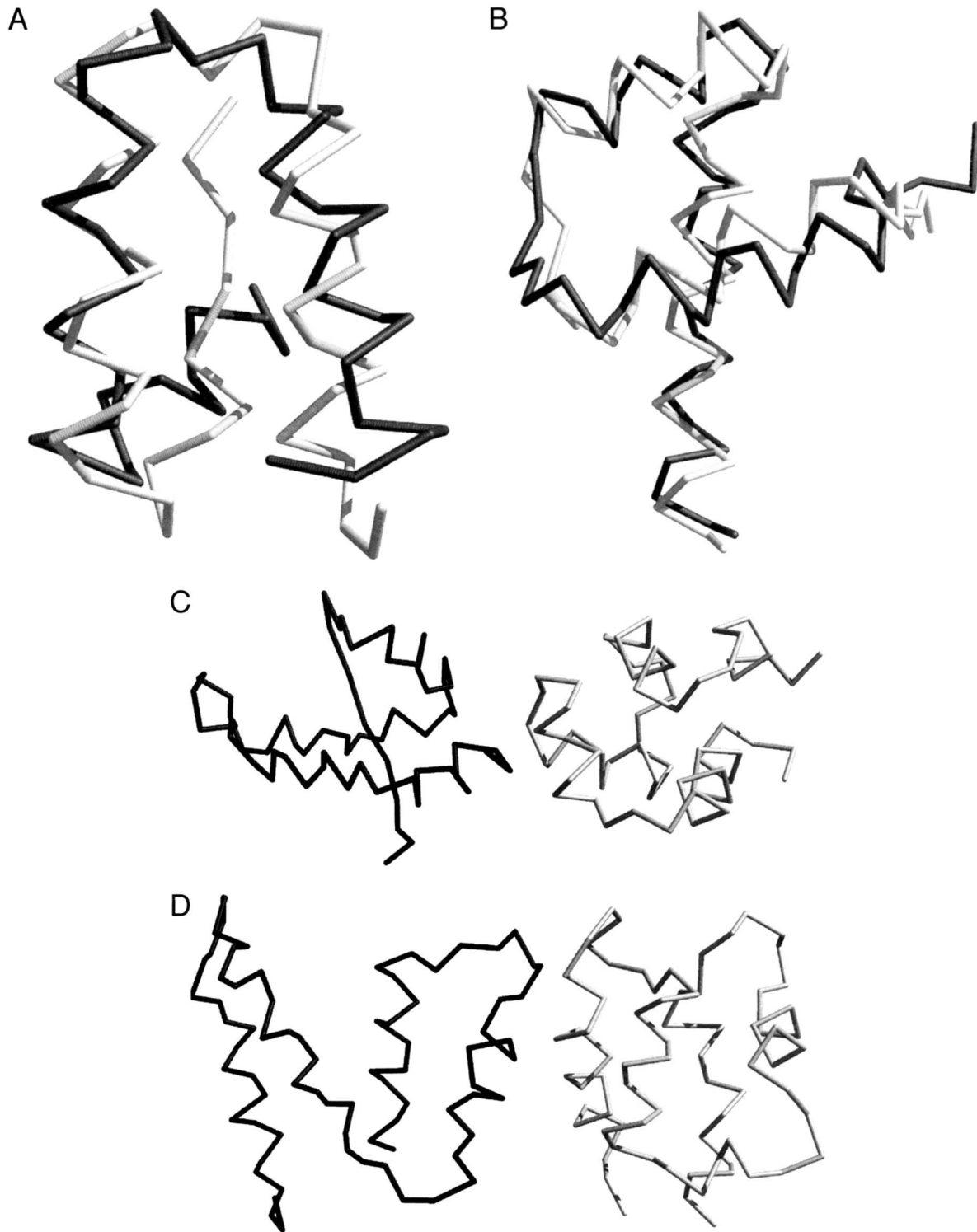
The distance geometry models of 1fc2 and 1hdd present well-formed helical structure (Fig. 1A,B). In contrast, the model of 2cro exhibits deformed secondary structure (Fig. 1C). This distortion is caused by a set of consensus constraints that are incompatible with each other, the possibility of which is an inherent danger associated with a consensus-based method. Table 1 lists the RMS error between the inter-Cα distance constraints (taken as the center of the consensus bin) and the actual inter-Cα distance found in the respective distance geometry structure. The integrity of the secondary structure elements in 4icb is qualitatively intermediate, consistent with the trend seen in the RMS restraint error.

In addition to distorting the secondary structures, mutually inconsistent restraints can cause errors in the global fold description. This phenomenon is also seen in the modeling of 2cro, in which the structure generation is dominated by a few distances defined by residues distant in the sequence. For example, consider the distribution of inter-Cα distances between residues 35 and 65 of 2cro. It is rather broad, with an average value of 15.4 Å and a standard deviation of 4.8 Å (Fig. 2A). The corresponding distance in the native structure is 18.4 Å, while the consensus distance bin is 13–14 Å. Distance geometry satisfies this incorrect constraint, and

Table 1. *Results for distance geometry modeling* [a]

| Protein	Min	Max	Mean	SD	DG	<DG	ConErr
1fc2:C	3.11	10.48	4.91	1.63	3.95	0.30	0.38
1hdd:C	2.75	12.81	6.82	2.69	3.78	0.05	0.99
2cro	4.20	12.49	8.72	1.99	7.17	0.22	2.00
4icb	4.70	14.00	9.40	2.40	6.50	0.16	1.73

[a]The columns Min, Max, Mean, and SD refer to the coordinate RMS error of the folds provided by Simons et al. (1997). The RMS error of the structure generated by distance geometry is listed in the DG column. The fraction of structures with RMS error less than the distance geometry model is shown under <DG. The column ConErr refers to the RMS error of the input constraints relative to the distance geometry model.

Fig. 1. Distance geometry models of four small helical proteins. In every figure, the distance geometry model is depicted in dark grey and the crystal structure in white. **A:** Protein A (1fc2:C). **B:** The homeodomain protein (1hdd:C). **C:** Cro repressor (2cro). **D:** Calbindin (4icb). For clarity, the models shown side by side for 2cro and 4icb. Molecular graphics images were produced using the MidasPlus software system from the Computer Graphics Laboratory, University of California, San Francisco (Ferrin et al., 1988).

others akin to it, by projecting a Cartesian structure with a inter-Cα distance of 12.3 Å. Inspection of the structure (Fig. 1C) shows that the C-terminal tail protrudes through a four-helical ring, rather than forming a helix and packing properly against the other helices. Therefore, proper placement of the C-terminal segment would result in a structure grossly inconsistent with the input constraint.
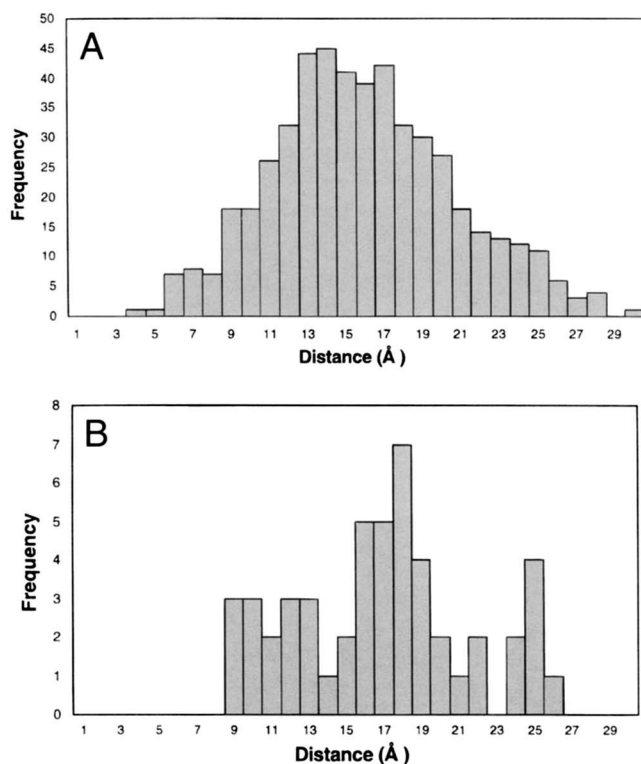
Fig. 2. A: Histogram of inter-Cα distances for residues 36 and 65. This histogram represents all the inter-Cα distances for residues 36 and 65 in the set of 500 parent structures of 2cro generated by Simons et al. (1997). The consensus distance is the bin ranging from 13 to 14 Å. **B:** Histogram of a subset of inter-Cα distances for residues 36 and 65. This histogram plots the inter-Cα distances for residues 36 and 65 in the subset of 50 best-scoring structures of 2cro. The consensus distance, the bin ranging from 18 to 19 Å, closely matches that of the native structure.

Table 2. *Distance geometry models using constraints from a score-filtered subset*[a]

| Protein	Min	Max	Mean	SD	DG	<DG
1fc2:C-50	3.30	5.98	4.17	0.42	4.08	0.38
1hdd:C-50	2.75	7.98	4.74	1.23	3.38	0.12
2cro-50	4.20	12.49	7.61	2.29	4.27	0.04
4icb-50	4.85	13.13	8.51	2.40	5.78	0.10

[a]The columns Min, Max, Mean, and SD refer to the coordinate RMS error of the top-scoring 50 folds. The RMS error of the structure generated by distance geometry is listed in the DG column. The fraction of structures with RMS error less than the distance geometry model is shown under <DG.

The 500 structures of 2cro do comprise several folds of moderately good quality, e.g., the 44 structures between 4 and 6 Å RMS error. In the hopes of amplifying the signal carried by these near-native structures, we filtered the 500 structures via the all-atom scoring function RAPDF (see Materials and methods) and saved the 50 best-scoring structures, i.e., the top 10%. Prior to scoring the folds, it was first necessary to construct all-atom models of the proteins. Side chains were added by the software of Bower et al. (1997). Steric clashes were removed by steepest-descent minimization under the CHARMM22b force field parameters (Brooks et al., 1983; A. McKerell, Jr., pers. comm.) and using the TINKER software package.

The top-scoring 50 structures with respect to RAPDF were generally of better quality than the parent set of 500 (Table 2). The mean RMS error decreased more than 1 Å for 2cro. More importantly, the consensus distance between residues 36 and 65 increased to 17–18 Å, very close to the actual distance of 18.4 Å. Another round of distance geometry yielded a structure with 4.26 Å coordinate RMS error, essentially the limit presented by the parent set of 500 (Table 1). As shown in Figure 3, the resulting structure appears less deformed in the regions of secondary structure. Repeating the filter for the other three proteins also improves the distance geometry models overall (Table 2), indicating that scoring with RAPDF has the desired effect of improving the quality of the input restraints. Further improvement in the quality of the

secondary structures, but not in the overall coordinate RMS error, can be achieved by widening the distance between the upper and lower bounds for all nonlocal interactions while keeping the window centered at the same consensus distance (data not shown). The softening of conflicting long-distance restraints effectively allows the helices to regularize.

Encouraged by the results of the filtering procedure, we assessed the accuracy of the top-scoring structure with respect to RAPDF. Table 3 shows that the top-scoring structures are amongst the best structures available in their respective sets. In terms of overall RMS accuracy, the top-scoring structures and the distance geometry structures are comparable. However, the distance geometry structure of 4icb reflects the use of distances taken from structures better than the one proposed by the scoring criterion alone: the coordinate RMS fell from 6.82 to 5.78 Å. Evaluation of the consistency of this function with regards to other ab initio test sets is currently underway (Samudrala & Moult, 1998).

Our consensus-based distance geometry approach shows promise in the context of native-like fold generation for two reasons. One, the sets of structures contain folds that are related to each other to varying degrees. In a case like 1fc2, where nearly a third are near-native, the consensus method mostly chooses the inter-residue distances reflecting the dominant subpopulation. Even for the 1hdd set, in which near-native folds are not present in high concentration, a consensus method still works because the distance errors in members of the fold set are distributed differently, such that the consensus method is able to largely preserve the correct distances and remove the incorrect distances as noise. The second reason why the distance geometry approach succeeds is that the problem of generating a single Cartesian structure from all the inter-Cα distances is greatly overdetermined. Even with as few as one or two correct inter-residue distances per residue, a native-like fold can be built by metric matrix distance geometry (Aszodi et al., 1995). The method described here provides $(n - 1)(n - 2)/2n$ constraints per residue, or approximately 30 constraints per residue for a small protein of 60 amino acids. Thus, any incorrect inter-residue distances that may result from the consensus procedure can be compensated by the other distances. Distance geometry is very powerful for generating a "best" structure from such an overdetermined set of conflicting constraints.

Our results suggest that when handled properly, distance geometry can circumvent one of the major problems that continues to plague ab initio folding, the selection of a single structure from the many promising candidates. Certainly the final accuracy of the
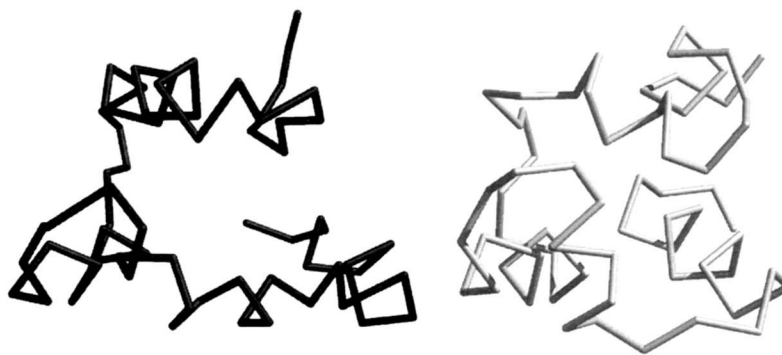
Fig. 3. Distance geometry model of 2cro using filtered restraints. The consensus restraints from the top 50 scoring structures with respect to RAPDF (Samudrala & Moult, 1998) were used to build a model of cro repressor, shown in dark grey. The model closely resembles that of the crystal structure, shown in white.

model is dependent on the quality of the parent structures used to obtain the distance restraints. However, this approach is sufficiently general such that irrespective of the accuracy of the ab initio method being used, a final structure that is better than average is generated. We have also shown that applying a more discriminating scoring function to the folds sets can enhance and improve the distance geometry models. In other words, a new method of fold generation to obtain the fold sets may not be necessary if one can simply filter the structures already obtained by the method of choice. Finally, this method is insensitive to the choice of algorithm for generating candidate structures, whether by minimization or enumeration, on a lattice or in torsion space. We expect that this consensus-based distance geometry method will prove useful as a tool for the ab initio prediction of protein structure.

Materials and methods

Generation of ab initio structures

Backbone-only structures for the four small helical proteins were generated by the method of Simons et al. (1997) and provided by the laboratory of David Baker (University of Washington, Seattle, Washington). Briefly, the procedure begins with an extended polypeptide chain. Guided by a Bayesian scoring function, it chooses random moves in torsion space under a simulated annealing schedule. The moves at each residue position are limited to those in a library of fragments of unrelated protein structures with similar local sequences (Bystroff et al., 1996; Simons et al., 1997). For

Table 3. *Highest ranking structures by an all-atom scoring function*

Structure	RMSD
1fc2:C	3.68
1hdd:C	2.85
2cro	4.26
4icb	6.82

each protein, 500 low-energy, compact folds were independently generated.

Prior to scoring with the all-atom function (below), side chains were added by the software package SCWRL (Bower et al., 1997). The side-chain placement strategy utilizes a backbone-dependent rotamer library (Dunbrack & Karplus, 1993) to position side chains such that the total steric clash is minimized. Since there was usually residual steric clash, we subjected each structure to 500 steps of steepest descent minimization using the CHARMM22b forcefield (Brooks et al., 1983; A. McKerell, pers. comm.) Electrostatic terms were neglected and a cutoff of 12 Å was applied for nonbonded interactions. All minimization was performed in Cartesian space with the TINKER software package.

Metric matrix distance geometry with pairwise metrization

All metric matrix distance geometry calculations were performed with the program distgeom from the TINKER suite. The metrization technique, developed in its original form by Havel and Wuthrich (1984), greatly improves the sampling properties of the distance geometry algorithm (Havel, 1990). In this study, structures are generated using 10% random pairwise metrization, which is efficiently achieved via a fast shortest path update algorithm (Dionne, 1978) used to resmooth the lower and upper bounds matrices every time a trial interatomic distance is chosen. The trial distances are selected from approximately Gaussian distributions between the lower and upper bounds (Oshiro et al., 1991). The center of the distribution between the upper and lower bounds is automatically set by distgeom based on the number and type of input restraints. The initial distribution center is set to $D_{\text{CENTER}} = B_L + \alpha(B_U - B_L)$, where B_L and B_U are the lower and upper bounds for a particular atom pair. For all proteins considered here, the empirical factor $\alpha = 1.65/(B_{\text{max}})^{1/4}$ is used, where B_{max} is the maximal upper bound in the entire structure following the initial bounds smoothing. This mechanism consistently yields folds with an approximately correct overall radius of gyration. 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, and the input distance restraints. Further details of the algorithm and its application to NMR NOE structure determination can be found in Hodsdon et al. (1996).

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

The all-atom scoring function RAPDF (Samudrala & Moult, 1998) was used to calculate the probability of a native or "correct" structure given a set of interatomic distances. The conditional probabilities are compiled by counting frequencies of distances between pairs of atom types in a database of protein structures. All nonhydrogen atoms are considered, and the description of the atoms is residue specific, i.e., the $C\alpha$ of an alanine is different from the $C\alpha$ of a glycine. This results in 167 atom types. We divide the distances observed into 1 Å bins ranging from 3 to 20 Å. Contacts between atom types in the 0–3 Å range are placed in a separate bin, resulting in total of 18 distance bins. For observations of distances between pairs of atoms between the atoms of a side chain and the main-chain atoms of that residue, a separate table of frequencies is compiled using eighteen 1 Å bins ranging from 0–18 Å.

We compile tables of scores proportional to the negative log conditional probability that one is observing a native conformation given an interatomic distance for all possible pairs of the 167 atom types for the 18 distance ranges. Given a set of distances in a conformation, we can evaluate the probability that the conformation represents a "correct" fold by summing the scores for all distances and the corresponding atom pairs. A full description of this formalism is given in Samudrala and Moult (1998).

References

Aszodi A, Gradwell MJ, Taylor WR. 1995. Global fold determination from a small number of distance restraints. *J Mol Biol 251*:308–326.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD Jr, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Bower MJ, Cohen FE, Dunbrack RL Jr. 1997. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *J Mol Biol 267*:1268–1282.

Bowie JU, Eisenberg D. 1994. An evolutionary approach to folding small α-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc Natl Acad Sci USA 91*:4436–4440.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem 4*:187–217.

Bystroff C, Han KF, Simons KT, Baker D. 1996. Local sequence-structure correlations in proteins. *Curr Opin Struct Biol 7*:417–421.

Covell DG. 1992. Folding protein α-carbon chains into compact forms by Monte Carlo methods. *Proteins Struct Funct Genet 14*:409–420.

Covell DG. 1994. Lattice model simulations of polypeptide chain folding. *J Mol Biol 235*:1032–1043.

Dandekar T, Argos P. 1994. Folding the main chain of small proteins with the genetic algorithm. *J Mol Biol 236*:844–861.

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

Dionne R. 1978. Étude et extension d'un algorithme de Murchland. *Infor 16*:132–146.

Dunbrack RL Jr, Karplus M. 1993. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J Mol Biol 230*:543–574.

Ferrin TE, Huang CC, Jarvis LE, Langridge R. 1988. The MIDAS display system. *J Mol Graphics 6*:13–27.

Hanggi G, Braun W. 1994. Pattern recognition and self-correcting distance geometry calculations applied to myohemerythrin. *FEBS Lett 344*:147–153.

Havel TF. 1990. The sampling properties of some distance geometry algorithms applied to unconstrained polypeptide chains: A study of 1830 independently computed conformations. *Biopolymers 29*:1565–1585.

Havel TF, Kuntz ID, Crippen GM. 1983. The theory and practice of distance geometry. *Bull Math Biol 45*:665–720.

Havel TF, Wuthrich K. 1984. A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular ¹H-¹H proximities in solution. *Bull Math Biol 46*:673–698.

Hinds DA, Levitt M. 1992. A lattice model for protein structure prediction at low resolution. *Proc Natl Acad Sci USA 89*:2536–2540.

Hinds DA, Levitt M. 1994. Exploring conformational space with a simple lattice model for protein structure. *J Mol Biol 243*:668–682.

Hodsdon ME, Ponder JW, Cistola DP. 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 ES, Koehl P, Levitt M, Pappu RV, Ponder JW. 1998. Accuracy of side-chain prediction upon near-native protein backbones generated by ab initio folding methods. *Proteins Struct Funct Genet.* Forthcoming.

Kolinski A, Skolnick J. 1994. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins Struct Funct Genet 18*:338–352.

Kuntz ID, Crippen GM, Kollman PA. 1979. Application of distance geometry to protein tertiary structure calculations. *Biopolymers 18*:939–957.

Monge A, Lathrop EJP, Gunn JR, Shenkin PS, Friesner RA. 1995. Computer modeling of protein folding: Conformational and energy 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.

Oshiro CM, Thomason J, Kuntz ID. 1991. Effects of limited input distance constraints upon the distance geometry algorithm. *Biopolymers 31*:1049–1064.

Park BH, Huang ES, Levitt M. 1997. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J Mol Biol 266*:831–846.

Ponder JW. 1998. *TINKER—Software tools for molecular design, User's Guide for Version 3.6.* Washington University School of Medicine.

Samudrala R, Moult J. 1998. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol 275*:893–914.

Simons KT, 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.

Skolnick J, Kolinski A, Ortiz AR. 1997. MONSSTER: A method for folding globular proteins with a small number of distance restraints. *J Mol Biol 265*:217–241.

Srinivasan R, Rose GD. 1995. LINUS: A hierarchic procedure to predict the fold of a protein. *Proteins Struct Funct Genet 22*:81–99.

Sun S, Thomas PD, Dill KA. 1995. A simple protein folding algorithm using a binary code and secondary structure constraints. *Protein Eng 8*:769–778.

Tuffery P, Etchebest C, Hazout S. 1997. Prediction of protein side chain conformations: A study on the influence of backbone accuracy on conformation stability in the rotamer space. *Protein Eng 10*:361–372.

Vieth M, Kolinski A, Brooks CL 3rd, Skolnick J. 1994. Prediction of the folding pathways and structure of the GCN4 leucine zipper. *J Mol Biol 237*:361–367.

Wilson C, Doniach S. 1989. A computer model to dynamically simulate protein folding: Studies with crambin. *Proteins Struct Funct Genet 6*:193–209.

Yue K, Dill KA. 1996. Folding proteins with a simple energy function and extensive conformational searching. *Protein Sci 5*:254–261.