J. Mol. Biol. (1997) **268**, 209–225

JMB

Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions

Kim T. Simons[1], Charles Kooperberg[2], Enoch Huang[3] and David Baker[1]*

[1]*Department of Biochemistry*
Box 357350 University of
Washington, Seattle, WA
98195, USA

[2]*Department of Statistics*
Box 354322, University of
Washington, Seattle
WA 98195, USA

[3]*Beckman Laboratories for*
Structural Biology, Department
of Structural Biology, Stanford
University School of Medicine
Stanford CA 94305-5400, USA

**Corresponding author*

We explore the ability of a simple simulated annealing procedure to assemble native-like structures from fragments of unrelated protein structures with similar local sequences using Bayesian scoring functions. Environment and residue pair specific contributions to the scoring functions appear as the first two terms in a series expansion for the residue probability distributions in the protein database; the decoupling of the distance and environment dependencies of the distributions resolves the major problems with current database-derived scoring functions noted by Thomas and Dill. The simulated annealing procedure rapidly and frequently generates native-like structures for small helical proteins and better than random structures for small β sheet containing proteins. Most of the simulated structures have native-like solvent accessibility and secondary structure patterns, and thus ensembles of these structures provide a particularly challenging set of decoys for evaluating scoring functions. We investigate the effects of multiple sequence information and different types of conformational constraints on the overall performance of the method, and the ability of a variety of recently developed scoring functions to recognize the native-like conformations in the ensembles of simulated structures.

Keywords: protein folding; computer simulation; multiple sequence alignment; structure prediction; knowledge-based scoring functions

Introduction

In the last several years there has been exciting progress in the development of algorithms for *ab initio* protein folding: the generation of protein structures starting from amino acid sequence information alone (Kolinski & Skolnick, 1994; Bowie & Eisenberg, 1994; Yue & Dill, 1996, Srinivasan & Rose, 1995). Because of the many formidable problems facing *ab initio* folding simulations, such algorithms are not likely to become useful methods of structure prediction for any but the smallest proteins for quite some time. However, such efforts are of crucial importance because they highlight, as few other experiments can, the challenges facing current studies of protein folding.

Our primary interests in this area are twofold: first, to develop a computational model to complement biophysical and molecular biological studies of the folding of very small protein domains currently underway in our laboratory, and second, to build upon our studies of local sequence-structure relationships (Han & Baker, 1996), which are currently limited by a lack of treatment of non-local interactions. A working model for protein folding is that local amino acid sequence propensities bias each local segment of a folding polypeptide chain towards a small number of alternative local structures and that non-local interactions preferentially stabilize native-like arrangements of these otherwise transient local structures. The weak nature of the local propensities has complicated both the prediction of secondary structure from sequence and the search for structure in protein fragments (Bystroff *et al.*, 1996). Here, we use a knowledge-based treatment of local interactions related to that of our previous studies: short sequence segments are restricted to the local structures adopted by the most closely related sequences in the protein structure database.

Once the local structural preferences of portions of a sequence have been characterized, a method is required for generating structures consistent with these local preferences and for subsequently evaluating the quality of the structures using a measure of non-local interactions. Two quite different approaches to treating non-local interactions have been used in recent work: knowledge-based potential functions derived from the protein database (Miyazawa & Jernigan, 1996; Sippl, 1990; Wilson & Doniach, 1989; Jernigan & Bahar, 1996), which typically contain large numbers of parameters, and much simpler potentials based on chemical intuition (Srinivasan & Rose, 1995; Yue & Dill, 1996; Huang *et al.*, 1995), which can potentially provide clearer insights into sequence-structure relationships. We chose the first approach for the experiments described here because although there are many more parameters, all are completely determined by the structures in the protein database (excluding the proteins being folded); thus the danger of crafting a scoring function specific for a particular class of proteins can be avoided.

The shortcomings of current approaches to extracting scoring functions from the protein database have been highlighted by recent work (Thomas & Dill, 1996). Because of the critical importance of scoring functions to the success of any structure prediction effort, we begin by presenting a detailed derivation of scoring functions from a purely statistical point of view with particular attention to the interplay between solvation and residue pair specific effects.

Results

Scoring functions

In this section, we present a derivation of knowledge-based scoring methods which is considerably simpler than standard derivations and leads to a systematic series expansion of the residue distributions in the protein database. The derivation does not require the assumption that the protein database (the ground states of a set of molecules of different sizes and chemical compositions) constitutes some sort of Boltzmann distribution and avoids the ambiguities associated with the choice of a reference state.

We seek the most probable structure for a protein given the amino acid sequence and the large number of examples of sequences with known structures in the protein database. Using Bayes theorem, the probability of a structure given the amino acid sequence (and the information in the protein database) is

$$P(structure \mid sequence) = P(structure) \times \frac{P(sequence \mid structure)}{P(sequence)} \quad (1)$$

In comparisons of different structures for the same sequence, $P(sequence)$ is constant, and will be neglected in the following analysis. In the threading problem, it is simplest to assume that every structure in a representative protein set is equally probable: $P(structure) = 1/(\text{number of structures})$. As shown below, this together with the assumption of independence of residue pairs leads directly to standard expressions for distance-dependent scoring functions.

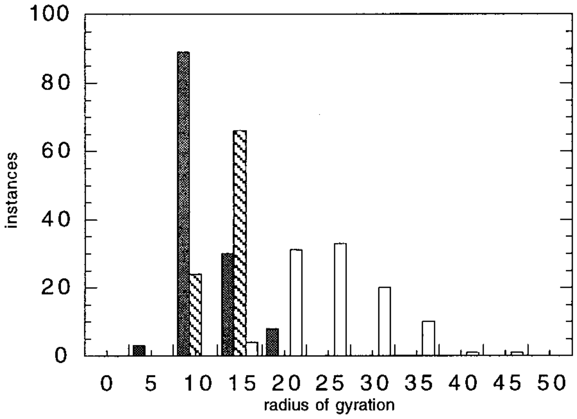

Figure 1. Comparison of the radii of gyrations of simulated and native structures. 100 structures were generated for chains of 100 residues by splicing together protein fragments as described in Methods using either no scoring function (open bars), or the square of the radius of the gyration as the scoring function (hatched bars). Histograms were computed using 5 Å bins. The distribution of radii of gyrations for the small (50 to 150 residue) proteins in the pdbselect 25 set is shown for comparison (filled bars).

In contrast, for the *ab initio* folding problem, not all generated structures are equally likely to be proteins; for example, highly expanded conformations and conformations with unpaired β strands occur frequently in randomly generated ensembles but not in real proteins. $P(structure)$ in this case captures all the features that distinguish folded protein structures from random chain configurations. In this study, a very simple form for $P(structure)$ is used: $P(structure)$ is zero for configurations with overlaps between atoms, and is proportional to $exp(-radius\ of\ gyration^2)$ for all other configurations. An empirical justification for this expression is given in Figure 1. Configurations generated by randomly splicing together protein fragments are considerably more expanded than native proteins (Figure 1, open bars), while conformations generated using $exp(-radius\ of\ gyration^2)$ for $P(structure)$ (hatched bars) in conjunction with simulated annealing have radii of gyrations comparable to those of native proteins of the same length (filled bars). Since the configurations are generated from protein fragments, their helix and strand content is similar to that of proteins, and thus $P(structure)$ in our case is independent of helix/strand content. A notable shortcoming of this expression for $P(structure)$ is that structures with paired β strands are no more probable than structures with unpaired β strands, and thus nothing in the simulation favors the formation of β sheets from β strands. We are currently developing an improved expression for $P(structure)$ which explicitly depends on the relative orientations of local structural elements:

$$P(structure) \cong \prod_{i<j} P(r_{ij}, \theta_{ij}, \varphi_{ij}, \omega_{ij} \mid ss_i, ss_j) \quad (2)$$

The r_{ij}, θ_{ij}, ϕ_{ij}, and ω_{ij} describe the separation and relative orientation of local structural elements ss_i and ss_j. Preliminary tests with fixed secondary structure simulations show that such an expression is sufficient to generate β sheet structures for short β strand containing chains.

Evaluation of the second term in equation (1), $P(sequence \mid structure)$, usually involves the assumption of independence of individual positions or pairs of positions. In the profile method (Bowie *et al.*, 1991),

$$P(sequence \mid structure) \cong \prod_{i} P(aa_i \mid E_i) \quad (3)$$

where E_i is the structural environment at position i (usually defined in terms of solvent accessibility and/or secondary structure) and the score of the structure is the product over all residues (aa_i). Other approaches involve the assumption of independence of pairs of positions rather than individual positions:

$$P(sequence \mid structure) \cong \prod_{i<j} P(aa_i, aa_j \mid r_{ij}) \quad (4)$$

where r_{ij} is the distance between residues i and j and the score is the product over all pairs of residues. While the assumption of independence of residue identities is certainly wrong in detail, it is considerably less drastic than the assumption that residue separations are independent of one another usually made in the derivation of potentials from the structure database (this amounts to a complete neglect of chain connectivity).

Equation (4) is very similar to expressions for potentials of mean force (Sippl, 1990; Kocher *et al.*, 1994) derived using the assumption that the protein database is some sort of Boltzmann distribution. Using Bayes theorem again for a particular pair of residues i and j,

$$P(aa_i, aa_j \mid r_{ij}) = P(aa_i, aa_j) \times \frac{P(r_{ij} \mid aa_i, aa_j)}{P(r_{ij})} \quad (5)$$

The first factor on the right is independent of structure, and the ratio is the expression derived by Sippl and others.

Combining equations (1), (4), and (5), and using $P(structure) \sim \exp(-radius\ of\ gyration^2)$ leads to

$$P(structure \mid sequence) \cong e^{-radius\ of\ gyration^2} \times \prod_{i<j} \frac{P(r_{ij} \mid aa_i, aa_j)}{P(r_{ij})} \quad (6)$$

The negative logarithm of this expression was used as the scoring function in the initial generation of structures. The correction for small sample size suggested by Sippl (1990) was used in the second term and, to increase the number of counts in each bin, the order of the two interacting residues in the sequence was ignored. The scoring functions were calculated separately for the sequence separation bins described by Bauer & Beyer (1994).

Scoring functions based on residue pair distributions are often supplemented with a neighbor density or solvent accessible surface area term which is supposed to represent solvation effects (Sippl, 1993; Jones *et al.*, 1992). However, solvation forces make no less of a contribution to the observed pair distributions than do any other physical interactions; dominant features of the pair distributions such as the relatively high frequency of contacts between hydrophobic residues reflect the properties of the solvent. Because solvation is included implicitly in the pair distributions, such additional terms are better viewed as corrections for the lack of independence of the pair distributions (in physical terms, a lack of pair additivity of the interaction energies).

To determine whether such a correction is warranted, we calculated the expected neighbor density around each residue (the number of C^{β} atoms of other residues within 10 Å of the C^{β} atom of the residue) in the protein set based on the residue pair distributions. The calculated neighbor density distributions for aspartate and isoleucine (Figure 2, open symbols) resemble those actually observed (Figure 2, filled symbols), except that the calculated distribution for isoleucine extends to larger numbers of neighbors probably because excluded volume is neglected. The similarity between the calculated and observed distributions indicates that the density term used in previous treatments involves considerable overcounting; a proper correction for many body effects would primarily compensate for the lack of explicit treatment of excluded volume (Miyazawa & Jernigan, 1996).

An obvious difficulty with database derived scoring functions is that residue distributions are sensitive to protein size. For example, C^{β} densities (Figure 3A) and histograms of distances between hydrophobic residues (Figure 3B) are significantly different in small proteins (circles) than in the database overall (squares). However, the constraints of excluded volume and connectivity inherent in the method of generating structures (see Methods) together with the $P(structure)$ term appear to correct partially for such undesirable biases: distributions calculated from a large set of simulated structures (Figure 3, triangles) resemble the small protein distributions much more closely than the overall database distributions from which the scoring functions were derived. In particular, an additional excluded volume correction does not seem to be necessary.

The problems with the assumption of independence of the pair distributions have been elegantly demonstrated by Thomas & Dill (1996). The partitioning of hydrophobic residues to the interior and hydrophilic residues to the exterior of proteins

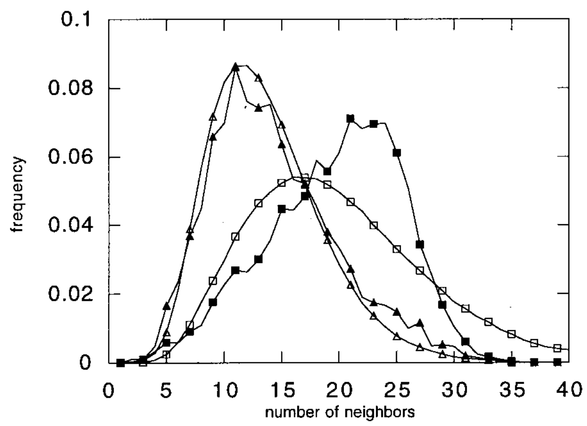

Figure 2. Comparison of observed C^β densities with C^β densities calculated from the pair distributions assuming complete independence. C^β densities in protein structures were tabulated for isoleucine (filled squares) and aspartate (filled triangles). The C^β distributions around each residue expected from the pair distributions alone were simulated assuming complete independence of the different pairs: each other residue in the sequence was considered to be in contact with the residue in question if a randomly chosen number between 0 and 1 was less than the contact frequency for the pair of residues calculated from the protein database (a function of the identity of the residues and their separation along the chain). Estimates of the distributions converged after 1000 trials for each residue. To compare with the observed C^β distributions, the simulated distributions for all isoleucine (open squares) or all aspartate (open triangles) residues were combined. The mean number of neighbors for each residue type is the same in the observed and simulated distributions.

dominate the scoring functions: $P(r_{ij}\,|\,\mathit{inside, inside})$, $P(r_{ij}\,|\,\mathit{inside, outside})$ and $P(r_{ij}\,|\,\mathit{outside, outside})$ are very similar to the corresponding distributions for hydrophobic pairs, hydrophobic and hydrophilic residues, and hydrophilic pairs, respectively. The dependence of the scoring functions on protein size and the number of hydrophobic residues in the sequence also stem primarily from the partitioning of hydrophobic residues into the interior.

We now consider a series expansion of $P(\mathit{sequence}\,|\,\mathit{structure})$ which incorporates solvation and residue pair interactions in a non-redundant manner. Any joint probability function can always be expanded in terms of marginal probability functions that depend on fewer variables:

$$P(x_i, x_2, x_3, \ldots, x_n) \cong \prod_i P(x_i) \prod_{i<j} \frac{P(x_i, x_j)}{P(x_i)P(x_j)} \cdots \quad (7)$$

The second order term introduces corrections due to pair interactions; higher order terms (not shown) incorporate corrections for third and higher order interactions. Expanding $P(\mathit{sequence}\,|\,\mathit{structure})$ in such a fashion, and replacing the entire three-dimensional structure by only those features of the structure related to the amino acids in question yields

$$P(aa_1, aa_2, \ldots, aa_n \mid \mathit{structure}) \cong \prod_i P(aa_i \mid E_i)$$
$$\times \prod_{i<j} \frac{P(aa_i, aa_j \mid r_{ij}, E_i, E_j)}{P(aa_i \mid r_{ij}, E_i, E_j)P(aa_j \mid r_{ij}, E_i, E_j)} \quad (8)$$

As in equation (3), E_i can represent a variety of features of the local structural environment around residue *i*. Estimation of the third order term, which would incorporate corrections due to residue triplet interactions, is very difficult due to database size limitations. The important feature of equation (8) is that residue-environment and residue pair interactions are both treated without redundancy and without blurring the specific residue pair interactions with the overall partitioning of residues into the protein core.

The second order term in equation (8) is compared to equation (5) for pairs separated by more than ten residues along the chain in Figure 4. Two environment classes are used: a residue is considered buried if there are more than 16 other C^β atoms within 10 Å of the C^β atom of the residue in question; otherwise, it is exposed. For all residue pairs and all environment classes, the second order term in equation (8) decays to near zero by 12 Å as expected for specific interactions. The leucine-isoleucine interaction (Figure 4A) is characteristic of the interactions between pairs of hydrophobic residues: the environment independent function (equation (5), continuous line) is attractive at short distances and repulsive at long distances, whereas the environment dependent functions in equation (8) are weakly attractive at ~8 Å and decay rapidly to zero at longer separations. The interactions between pairs of surface hydrophobic residues (Figure 4A, broken line) are considerably stronger than those between pairs of buried residues (Figure 4A, dotted line); the overall partitioning of hydrophobic residues to the core captured by the first order environment term more completely accounts for the proximity of hydrophobic residues in the core than the clustering of hydrophobic residues in surface patches. The interactions between surface and buried residues are intermediate between the two extremes (data not shown). The environment specific scoring function for glutamate-lysine pairs is attractive at short range for all environment classes (Figure 4B; again, only buried-buried and exposed-exposed pairs are shown), and is considerably closer to physical intuition than the environment independent function, which becomes attractive at large separations because of the partitioning of polar residues to protein surfaces. The aspartate-aspartate pair interaction is repulsive at short distances as expected for surface pairs (Figure 4C, broken line), but weakly attractive for buried pairs separated by ~9 Å (Figure 4C, dotted line), perhaps reflecting the presence of multiple

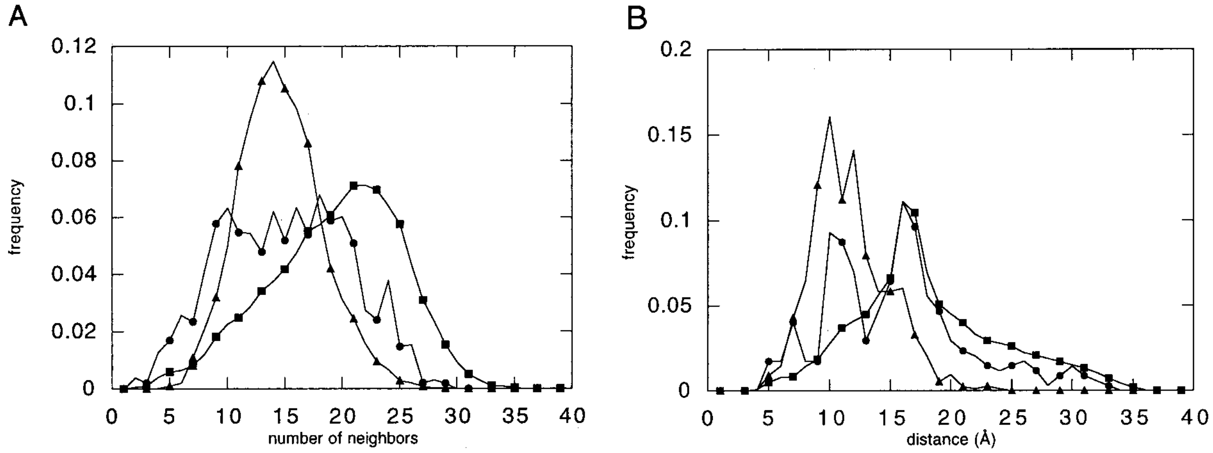

Figure 3. The protein size dependence of database derived scoring functions is partially corrected by the constraints in the simulation. A, The C^{β} density distributions of hydrophobic residues (ILV) of simulated structures (triangles) resemble those of small proteins (circles) more so than those of the entire database (squares). An extra solvation term is not required to reproduce the observed distributions. The C^{β} density around a given residue is defined as the number of C^{β} atoms within 10 Å of the C^{β} atom of that residue. B, The distribution of frequencies of different distance separations for pairs of hydrophobic residues separated by more than 20 residues in the sequence. The distribution for the simulated structures (triangles) resembles that of small proteins (circles) more closely than that of the entire protein database (squares), despite the fact that the latter distributions were used to score structures during the simulations.

buried aspartate residues in metal binding sites and enzyme active sites.

The combined expression (8), which will be referred to as the dist_env function for the remainder of the paper, is related to scoring functions described previously in the literature. Kocher *et al.* (1994) developed environment dependent versions of the expression in equation (5), and Miyazawa & Jernigan (1996) described an environment independent version of the second term in equation (8). As noted previously, the first term in equation (8) has been used extensively by Eisenberg's group (Bowie *et al.*, 1991) and others, and the description of the environment in terms of the number of nearby C^{β} atoms has been utilized by several groups (Huang *et al.*, 1995; Flockner *et al.*, 1995). Furthermore, combining different types of scoring functions has been explored in several recent studies (Kocher *et al.*, 1994; Park & Levitt, 1996). A major contribution of the analysis leading to equation (8) is the recognition that the two terms in (8) form the first part of a systematic expansion of $P(sequence\,|\,structure)$; this provides a rigorous justification for combination of the two terms. A virtue of equation (8) from a practical standpoint is that both the hydrophobic effect driven sequestration of non-polar residues in the core and more specific residue pair interactions are captured simultaneously (in the first and second terms, respectively). From a physical point of view, neither a residue-environment nor a residue-residue based description alone would be expected to adequately describe the dominant interactions in folding since the former clearly neglects specific pair interactions, and the latter requires an assumption of pair additivity likely to break down for the hydrophobic interaction (Rank & Baker, 1997).

Generation of structures

The method of generating structures is described in detail in Methods. Briefly, three-dimensional structures are generated by splicing together fragments of proteins of known structure with similar local sequences and evaluated initially using equation (6) above. Low scoring conformations with distributions of residues similar to those of known proteins are identified by simulated annealing in conjunction with a simple move set that involves replacing the torsion angles of a segment of the chain with the torsion angles of a different protein fragment with a related amino acid sequence. These conformations are then evaluated using the dist_env scoring function (equation (8)). Equation (6) is used in the initial generation of structures rather than the more data intensive equation (8) because of noise due to the limited size of the database; noise in scoring functions is much more problematic in folding simulations than in evaluating a relatively small number of fixed structures because there is much more freedom to "fit" the noise.

To evaluate the performance of folding simulation methods, it is essential to describe not only the best structures produced, but also the frequency with which such structures are generated. To facilitate comparisons with other work in this area, we describe the results with different simulation conditions in some detail: for each experiment, 100 simulated annealing runs were carried out, and the number of structures with less than

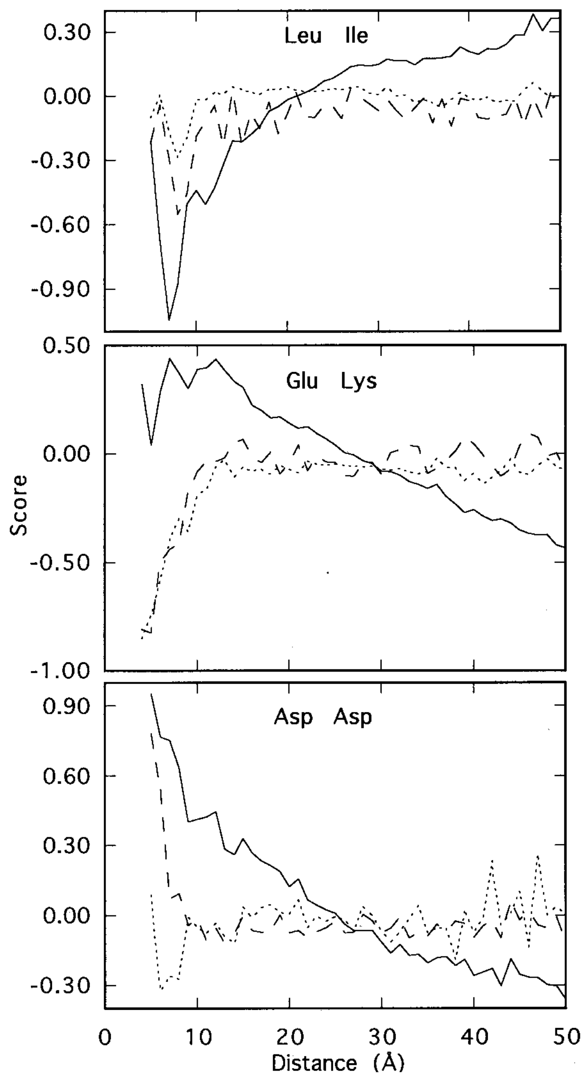

Figure 4. Comparison of the negative logarithms of equation (5) and the residue pair specific second term in equation (8) for sequence separations greater than ten. Residues with greater than 16 neighbors were considered buried. Continuous lines, equation (5); dotted lines, equation (8) both residues buried; broken line, equation (8) both residues exposed.

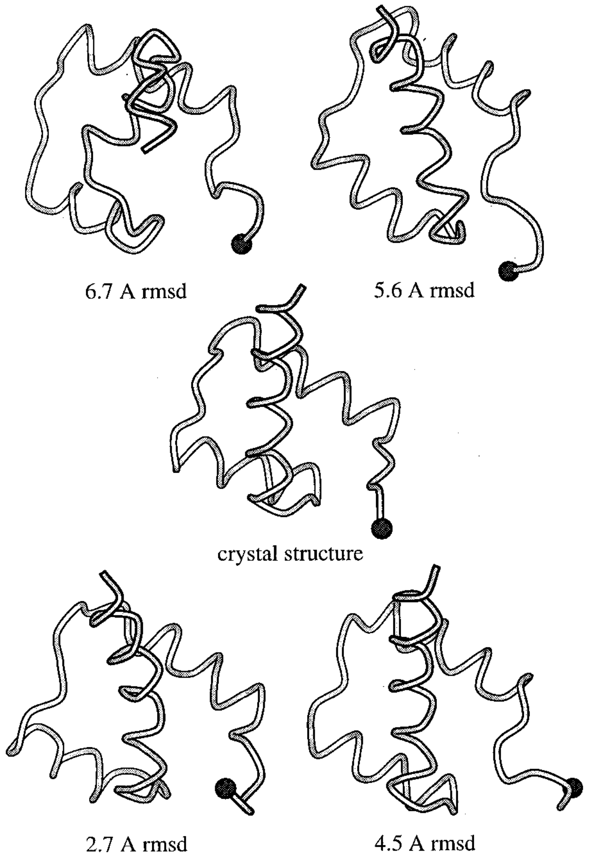

Figure 5. Simulated homeodomain structures with different rms deviations from the native structure. The N termini are displayed as black spheres.

7 Å, 6 Å, 5 Å and 4 Å C^{α} rmsd (root mean square deviation) from the native structure was recorded. The degree of topological similarity for the different rmsd ranges is illustrated in Figure 5. Because the average rmsd from the native structure increases with chain length, we also computed a length independent average quality factor Q for the 100 structures generated for each run condition using a variant of the procedure of Cohen & Sternberg (1980; see Methods). The more negative Q, the more native-like the simulated structures relative to random compact structures of the same length.

Structures within 5 Å rmsd from the native structure were readily generated for several small helical proteins (Table 1). The use of equation (8) to evaluate structures generated using equation (5) significantly increased the number of native like structures (dist_env filter in Table 1). The success of the folding simulations correlated with the length of the sequence being folded and the number of turns in the native structure. A large number of reasonable structures were obtained for protein A (43 residues), and relatively few for calbindin (76 residues). The cro repressor has two more turns than the homeodomain and folded considerably less well (Table 1). Most of the conformations with large rmsd from the native conformation had secondary structure and solvent accessibility patterns similar to those of the native structure (Figure 6). While no structures with less than 5 Å rmsd from the native state were obtained for either of the two β sheet containing proteins that were studied, the simulated structures have quality factors significantly less than zero and thus are more native-like than randomly generated structures (Table 1).

The score and the rmsd from the native state as a function of cycle number are shown in Figure 7 for a successful run with the homeodomain. The score and rmsd both decrease rapidly and then undergo considerable uncorrelated fluctuations. As demonstrated in Table 2, the program is not simply

Table 1. Folding simulation results

	<7 Å rmsd	<6 Å rmsd	<5 Å rmsd	<4 Å rmsd	Lowest rmsd	*Q*
A. *Unconstrained simulations*						
Homeodomain						
dist_env filter + msa (100)	65	47	31	17	2.75	−1.7
dist_env filter − msa	63	45	31	16	2.75	−1.8
No filter	63	48	38	8	2.75	−1.5
Random sequence	31	11	1	0	4.89	−0.2
Random fragments	16	4	1	0	4.73	−0.6
Random all	6	2	0	0	5.82	0
Calbindin						
dist_env filter + msa (64)	31	17	2	0	4.70	−1.7
dist_env filter − msa	24	14	1	0	4.70	−1.9
No filter	17	3	2	0	4.86	−1.4
Random sequence	3	0	0	0	6.18	−0.2
Random fragments	6	1	0	0	5.71	−0.4
Random all	0	0	0	0	7.63	0
Protein A						
dist_env filter	96	95	93	41	3.29	−2.3
No filter	86	85	77	41	3.16	−2.0
Random sequence	33	25	8	1	3.52	−0.2
Random fragments	48	32	9	1	3.97	−0.6
Random all	32	14	1	0	4.58	0
Cro repressor						
dist_env filter + msa (4)	39	18	8	0	4.20	−1.7
dist_env filter − msa	35	20	10	0	4.20	−1.9
No filter	24	11	4	0	4.26	−1.5
Random sequence	7	1	0	0	5.95	−0.3
Random fragments	5	0	0	0	6.14	−0.7
Random all	0	0	0	0	7.26	0
Protein G						
dist_env filter + msa (5)	3	0	0	0	6.33	−1.5
dist_env filter − msa	2	0	0	0	6.33	−1.5
No filter	1	0	0	0	6.89	−1.2
Random sequence	0	0	0	0	8.43	−0.4
Random fragments	0	0	0	0	7.80	−0.6
Random all	0	0	0	0	8.35	0
Ribosomal fragment						
dist_env filter + msa (59)	16	6	0	0	5.26	−1.6
dist_env filter − msa	18	6	0	0	5.26	−1.5
No filter	8	3	0	0	5.85	−1.2
Random sequence	0	0	0	0	8.78	−0.2
Random fragments	0	0	0	0	7.46	−0.4
Random all	0	0	0	0	7.41	0
B. *Effect of conformational constraints*						
No constraint	24	11	4	0	4.26	−1.5
2° struct constant	23	13	3	0	4.44	−1.1
3rd turn constant	54	41	22	0	3.03	−2.4
C. *Folding of homologues*						
434 repressor	27	21	8	1	3.8	−1.4
Cro repressor	21	3	2	0	4.4	−0.9

For each of the proteins and run conditions indicated in the left column, 100 independent simulations were performed. The number of structures within the indicated rmsd of the corresponding native structure is given for each set of 100 runs in columns 2 to 5. The rmsd of the most native-like structure is given in column 6. The average value of the structure independent quality factor *Q* (see Methods) over the 100 runs is given in column 7. A, Random sequence simulations utilized scrambled versions of the native sequence; random fragment simulations, randomly selected fragments; and random all simulations, scrambled sequences and randomly selected fragments. For the + msa simulations scores were averaged over all sequences in multiple sequence alignments for the proteins being folded (the number of sequences are shown in parentheses). For the filter entries, the 100 conformations with the best scores according to the dist_env function (equation (8)) were chosen from sets of 500 conformations generated for each of the proteins using the no filter conditions. Multiple sequence information was used to identify the starting fragments in each set of simulations. B, Effect of conformational constraints on the folding of Cro repressor. In the simulation runs with fixed secondary structure, residues 1 to 14, 18 to 24, 30 to 38, 47 to 53, and 58 to 65 were constrained to be helical; in the second set of simulations, the torsion angles of residues in the third turn (residues 39 to 46) were fixed at their values in the native structure. C, Results for 434 and Cro repressors using single sequence information in fragment identification and scoring.

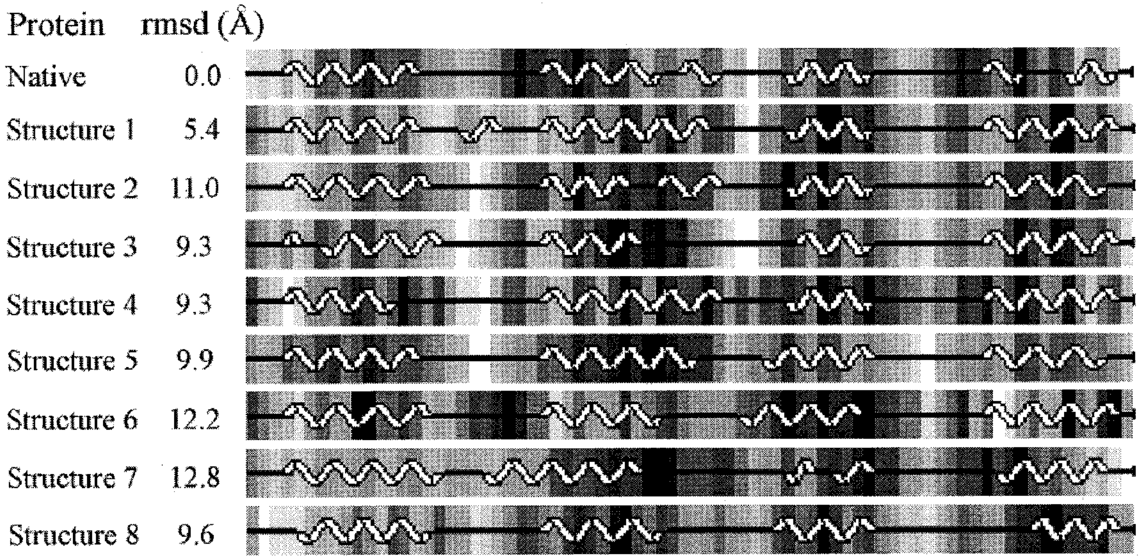

Figure 6. Solvent accessibility and secondary structure of a number of simulated non-native calbindin structures as depicted by PROCHECK (Laskowski *et al.*, 1993). The structures were randomly drawn from the simulated structure set prior to filtering. The rmsd to the native structure is shown in the second column; the rmsd between all pairs of structures is greater than 5 Å. White, solvent accessible; black, buried.

doing homology modeling; the fragments are derived from a wide variety of different, non-homologous protein structures. Proteins homologous to the protein being folded were removed from the fragment library and excluded from the construction of the scoring functions in all simulations.

In principle, both the method of selecting fragments and the method of scoring configurations should favor the native configuration relative to the majority of possible configurations. To assess the relative importance of the constraints on the search space and the scoring function in arriving at reasonable structures, two additional sets of 100 simulations were carried out in which either the starting sets of fragments were chosen randomly or the scoring function used a scrambled version of the native amino acid sequence. As indicated in Table 1, the number of good structures declined sharply in both cases. As a further control, structures were generated using random starting fragments and scored using a scrambled amino acid sequence. Many fewer structures of less than 6 Å rmsd were found for all of the proteins (Table 1).

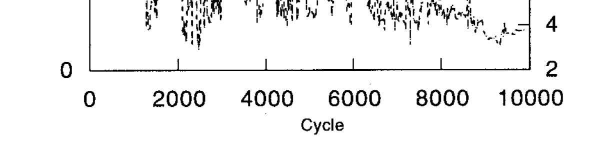

Figure 7. Progression of a homeodomain folding simulation. Continuous line, score; broken line, rmsd from the native structure. A cycle is an attempted replacement of the current torsion angles of a segment of the structure with the torsion angles of a fragment from the protein database with similar local sequence.

Although both the method of choosing fragments and the scoring function contribute to the generation of good structures, the simulations were relatively insensitive to moderate changes in either the method of choosing fragments or the scoring function. Multiple sequence information improved the resemblance of the starting fragments to the true structure and reduced the amount of noise in the scoring functions (Table 4 and Figure 11, see Methods), but neither significantly improved the yield of good structures. For example, the number of good structures obtained using single sequence information alone for the 434 repressor (Table 1C) was comparable to that obtained using multiple sequence information for the whole repressor family (Table 1A). A simple but perhaps not optimal method of using multiple sequence information for scoring, averaging together the scores for all sequences in a family being folded, slightly improved the recognition of native-like structures relative to that with single sequences alone (Tables 1 and 3, msa). Other methods of utilizing multiple sequence information may be more effective (Bowie *et al.*, 1990 and Defay & Cohen, 1996). An exception to the overall robustness of the structure generation procedure was the requirement for fragments of greater than five residues early in the simulations: considerably fewer good structures were generated when only

Table 2. Origins of fragments contributing to final simulated structures

Residue	Structure I (2.7 Å rmsd, 2.1 Å dme)	Structure II (3.0 Å rmsd 2.1 Å dme)
1	Methyltransferase (1hmy)	Endonuclease III (1abk)
2	Creatinase (1chm)	Endonuclease III (1abk)
3	Cytochrome *c* (1ccr)	Endonuclease III (1abk)
4	Cytochrome *c* (1ccr)	Recoverin (1rec)
5	Cytochrome *c* (1ccr)	Recoverin (1rec)
6	Barley seed protein (1bw4)	Recoverin (1rec)
7	Hydrolase inhibitor (1hle)	3-isopropyl malate DH (1hex)
8	Ribose binding protein (2dri)	3-isopropyl malate DH (1hex)
9	HIN recombinase (1hcr)	Proteinase inhibitor (1cew)
10	HIN recombinase (1hcr)	Proteinase inhibitor (1cew)
11	HIN recombinase (1hcr)	Proteinase inhibitor (1cew)
12	Aspartate aminotransferase (1ars)	Histidine binding protein (1hsl)
13	Apolipoprotein-E3 (1lpe)	Cutinase (1cus)
14	Apolipoprotein-E3 (1lpe)	Leghemoglobin (1gdm)
15	Apolipoprotein-E3 (1lpe)	Leghemoglobin (1gdm)
16	Glutathione transferase (1gst)	Leghemoglobin (1gdm)
17	Glutathione transferase (1gst)	Uteroglobin (1utg)
18	Acyl transferase (3cla)	Uteroglobin (1utg)
19	Interleukin-10 (1ilk)	Uteroglobin (1utg)
20	Thermolysin (8tln)	Alpha-parvalbumin (1rtp)
21	Immunoglobin FC (1fc2)	Adenovirus fiber protein (1knb)
22	Immunoglobin FC (1fc2)	Adenovirus fiber protein (1knb)
23	Immunoglobin FC (1fc2)	Adenovirus fiber protein (1knb)
24	Dihydrofolate reductase (3dfr)	Alpha-parvalbumin (1rtp)
25	Dihydrofolate reductase (3dfr)	Phosphotransferase (1npk)

The proteins from which the final torsion angles of two simulated homeodomain structures originate are indicated for residues 1 to 25 of both structures.

three residue fragments were used (data not shown). Longer fragments may prevent overly frequent changes in the direction of the collapsed chain.

There have been a number of reports in the last several years of folding simulations in which the secondary structure was kept fixed at that of the native structure (Monge *et al.*, 1994; Sun *et al.*, 1995). To investigate the effects of constraining different parts of the protein, we carried out further sets of simulations with different parts of the cro repressor fixed. Fixing the secondary structure (71% of the structure) had relatively little effect (Table 1B). This suggested that the simulation procedure might yield fairly accurate secondary structure predictions, but although the procedure did lead to a considerable increase in the native secondary structure content relative to the starting fragments (83% *versus* 61% correct in a three state secondary structure prediction), the well established PhD secondary structure method (Rost *et al.*, 1994) had an accuracy of 85% on the same set of proteins. In contrast to the result with constraining secondary structure, the fixing of even one turn (14% of the structure) substantially increased the yield of good structures (Table 1). Thus, a relatively small conformational constraint can compensate for deficiencies in scoring functions.

Table 3. *Z*-scores for native-like conformations with different scoring functions

	1FC2A	1HDD	2CRO	4ICB	Average
Surface	−0.52	−0.23	−0.38	−0.48	−0.40
HF	−0.46	−0.68	−0.04	−0.69	−0.47
Contact(HL)	−0.41	−0.19	0.08	−0.38	−0.23
Contact(MJ)	−0.30	−0.13	0.08	−0.59	−0.24
Shell	−0.41	−0.48	−0.55	−1.05	−0.63
Shelltop	−0.39	−0.37	−0.42	−1.02	−0.55
Histogram	0.00	−0.04	−0.70	−0.48	−0.31
VdW(HL4)	−0.36	−0.69	−0.39	−1.31	−0.69
Shellm	−0.43	−0.54	−0.66	−0.59	−0.56
Shelltopm	−0.38	−0.56	−0.64	−0.89	−0.62
Eq(8)	−0.32	−0.69	−1.12	−0.87	−0.75
Eq(8) + msa	−0.32	−0.79	−1.08	−1.29	−0.87

The cutoff below which conformations were taken to be native-like was 4 Å rmsd for protein A and the homeodomain, and 5 Å rmsd for calbindin and cro repressor. The *Z*-scores (the number of standard deviations separating the scores of the native-like conformations from the ensemble average) were calculated over ensembles of 500 conformations for each protein generated using the "no filter" condition of Table 1.

Although drawing comparisons between simulations and experiments in this area is undoubtedly premature, there are some interesting parallels that are worth noting. First, a number of mutations in small proteins which increase helical propensities have been found not to increase the rate of folding (López-Hernández *et al.*, 1997; Munoz & Serrano, 1996), consistent with the failure to significantly improve the yield of good structures in the simulations by fixing the secondary structure. Second, the observation that fixing a single turn dramatically increases the yield of good structures is interesting in the light of the finding that mutations which disrupt particular turns can greatly slow the rate of folding (H. Gu, D. Kim & D.B., unpublished results).

Over the next several years it may be possible to gain considerable insight by bringing together results from experiments and folding simulations.

Evaluation of scoring functions using the ensembles of simulated structures

A variety of different knowledge-based scoring methods/potential functions have been developed in the last several years (Park & Levitt, 1996; Jernigan & Bahar, 1996). One of the best methods for evaluating scoring functions is to test their ability to recognize near-native conformations in large sets of "decoy" structures (Park & Levitt, 1996). The ensembles of conformations generated in the course of this work constitute a particularly challenging set of decoys because of their largely native-like solvent accessibility and secondary structure patterns. To determine whether the simulated structures had obvious non-protein-like features as well as test different scoring functions, we evaluated structures using scoring functions developed over the last several years primarily in Levitt's group.

The scoring functions show a modest degree of discrimination between native-like and non-native configurations (Table 3 and Figure 8). The functions are described in detail by Park & Levitt (1996); Huang *et al.* (1995); and Park *et al.* (1997). Very briefly, the HF (hydrophobic fitness) function assesses the extent to which the hydrophobic residues are sequestered into the protein core, the Contact(HL), Contact(MJ), Shell, and Shelltop functions are contact based scoring functions of the form *log* [(*number of contacts observed between residues i and j*)/(*number of contacts expected*)] which differ in the way contacts are defined and the expected number of contacts are estimated, and the histogram, VdW(HL4), Shellm and Shelltopm are different formulations of distance dependent scoring functions. The average Z-scores of native-like structures range from −0.23 to −0.62 for the different functions. The discriminatory power of the dist_env scoring function (equation (8)) was slightly better than that of the other functions (Z score of −0.75). Computing scores using multiple sequence information by averaging together the scores for each sequence in a family gave a further modest improvement in performance (Z score of −0.87). The significance of Z-scores in this range is por-

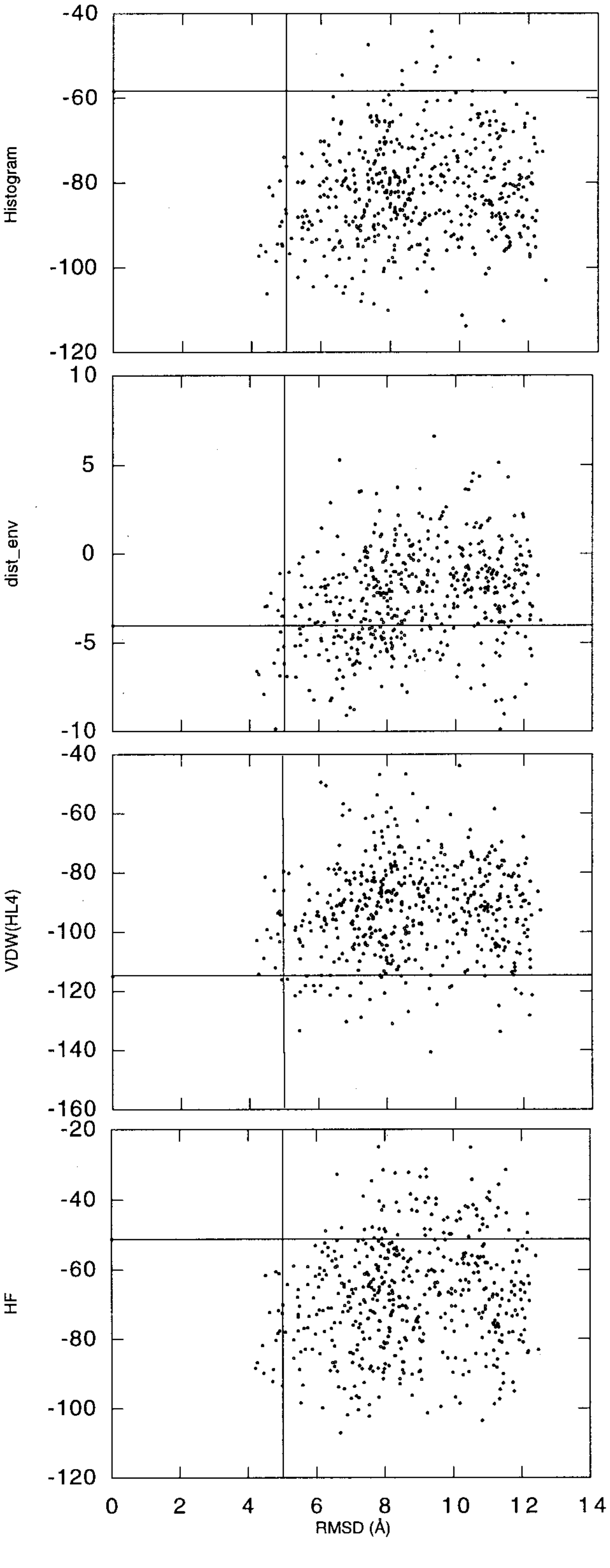

Figure 8. rmsd *versus* score for simulated cro structures. The vertical line indicates the rmsd cutoff for native-like structures used in Table 3; the horizontal line, the score of the native structure.

trayed by the examples of individual score *versus* rmsd plots shown in Figure 8.

Discussion

Derivation of scoring functions

The systematic derivation of scoring functions presented here has a number of useful features. First, the separation of sequence dependent and sequence independent contributions in equation (1) divides the problem into two more manageable subproblems that can be approached separately. As made clear in the derivation leading to equation (5), many current scoring functions consider only the sequence dependent term and thus should not be expected to be sufficient for the *ab initio* folding problem. Second, the series expansion of $P(sequence\,|\,structure)$ given in equation (8) provides a recipe for combining environment and residue pair specific effects in a systematic and non-redundant manner.

Comparison to other folding simulation studies

Previous *ab initio* folding studies have focused on either small proteins (Bowie & Eisenberg, 1994; Wilson & Doniach, 1989; Yue & Dill, 1996; Kolinski & Skolnick, 1994; Monge *et al.*, 1994) or protein fragments (Srinivasan & Rose, 1995; Avbelj & Moult, 1995). The work in this paper was largely inspired by the pioneering work of Bowie and Eisenberg, which showed that reasonable protein structures can be generated from sequence information alone without disulfide constraints or large numbers of free parameters. In the following section our results are compared to those obtained in earlier studies. In general, the results are comparable to or better than those of previously described methods.

The folding of the 434 repressor has been simulated in several previous studies. The best structure in the set of 100 configurations (Table 1C) generated without using multiple sequence information for the 434 repressor had an rmsd from native of 3.8 Å, and eight out of the 100 structures had an rmsd of less than 5 Å. For comparison, in simulations with fixed secondary structure and a genetic algorithm, ~7.5% of trials yielded structures within 7 Å rmsd of the native structure, but structures within 5 Å rmsd were not obtained (Monge *et al.*, 1995). In another approach using fixed secondary structure and a genetic algorithm, the final predicted structure of the 434 repressor was >10 Å rmsd from the native structure (Sun *et al.*, 1995). Better results were obtained by Bowie & Eisenberg (1994), 33% of 200 folding trials yielded structures with less than 4 Å dme (distance matrix error) from the native state, but several parameters in the scoring function were trained on this protein. For the homeodomain and protein A, the striking feature of our results is the large fraction of reasonably native-like structures; good structures were generated in previous studies, but the frequency of success was not always reported (Bowie & Eisenberg, 1994; Sun *et al.*, 1995).

To our knowledge, the only simulation attempt for calbindin utilized distance geometry with the secondary structure elements kept fixed (Mumanthaler & Braun, 1995); the average rmsd from the native state among the top ten structures was 7.2 Å. Our procedure yielded several reasonable structures for calbindin (Table 1). The best structure in the 100 runs described in Table 1 is shown in Figure 9; the similarity in topology is clear both in the ribbon drawing (Figure 9A) and in the contact map (Figure 9B).

The success in folding this set of helical proteins using a fairly broad range of methods is encouraging, but we, like others, have had much less success with β sheet containing proteins. Monge *et al.* (1994) and Dandekar & Argos (1996) did obtain good structures for a number of β strand containing proteins, but since the secondary structure was held fixed in most of these cases, the results are not directly comparable to ours. Three structures with less than 7 Å rmsd from the native state were obtained in 100 simulations with protein G, an α/β protein. Interestingly, the central helix was very often present in the simulated structures, and the two β hairpins were often partially formed, but the secondary structural elements were not properly packed. The ribosomal fragment, 1ctf, has three helices and a three-stranded β sheet in which the paired strands are not adjacent in the sequence. In simulations, the helices and strands frequently formed, but the strands did not come together to form the sheet; the best structure in 100 runs had an rmsd of 5.3 Å. The generation of β sheets from unpaired β strands will require an explicit β strand pairing term in $P(structure)$ (see equation (2)); there is nothing in our current expression for $P(structure)$ which favors strand pairing.

Evaluation of scoring functions using decoy sets

The evaluation of the simulated structures using alternative scoring functions provides insights into both the structures and the scoring functions. The scoring functions developed by Levitt's group capture a variety of different features of the residue distributions in protein structures. The average scores of close to native structures are better than those of the overall population for all of the functions, but the Z-scores are quite modest (−0.23 to −0.69). On the positive side, this indicates that the simulated structures have substantial protein-like properties. On the negative side, the scoring functions have relatively little discriminatory power: this is highlighted in the score *versus* rmsd plots shown in Figure 8. Among the contact based scoring functions, the Shell functions appear to be more discriminating than the Contact functions. The histogram function was less discriminating than the other distance dependent functions; this is

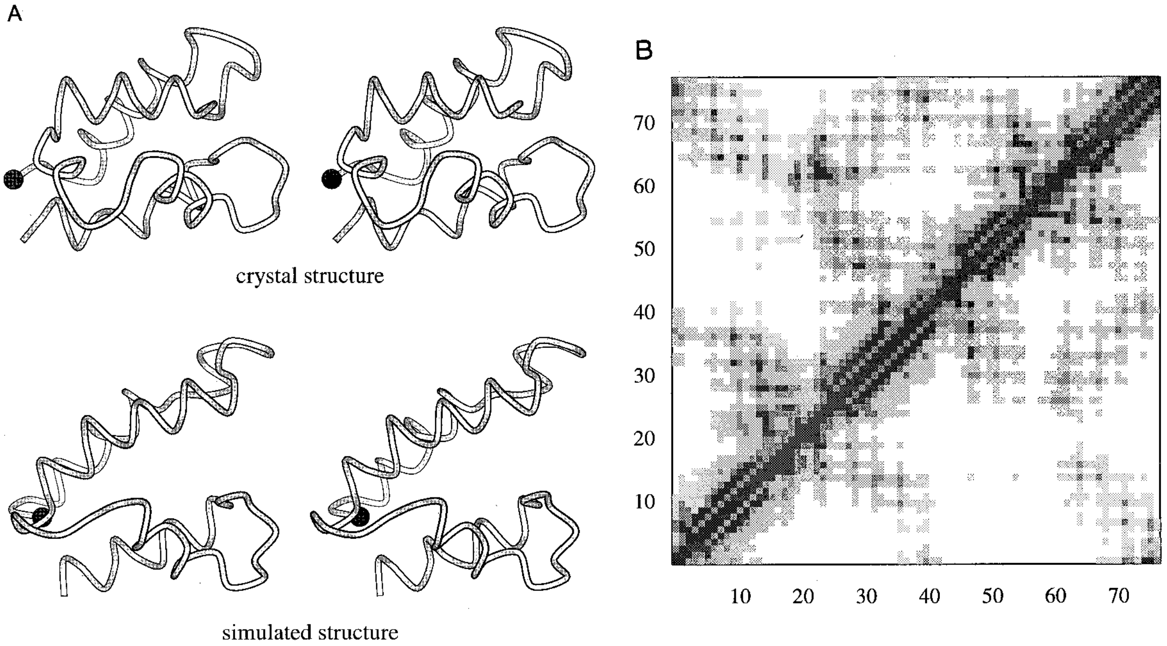

Figure 9. Comparison of the best calbindin structure generated in 100 simulations to the native structure. This structure has an rmsd from native of 4.9 Å, and a dme of 3.8 Å. A, Stereo view of ribbon diagrams as depicted by MOLSCRIPT (Kraulis, 1991). B, Contact maps. Native calbindin is in the upper right triangle, the simulated structure on the lower left. Residue pairs separated by less than 7, 10, 15 Å are indicated by black, dark grey and light grey squares, respectively.

probably related to the fact that the histogram function is very similar to that used in the generation of the simulated structures (equation (5)). Park & Levitt (1996) tested the ability of the functions to discriminate native-like from non-native structures in large sets of compact self-avoiding conformations with native-like secondary structure: the average Z-scores for native-like structures were considerably lower in their test case (~ – 1.5) than for our simulated structures. The simulated structures are a more stringent test of the functions since there are many fewer obviously incorrect configurations (with hydrophobic residues exposed, for example); because the functions are not orthogonal, conformations with good scores using one set of functions are more likely to have good scores using the other functions than are exhaustively generated structures.

Directions for future work

Current scoring functions evidently are not capable of consistently distinguishing native-like from non-native-like structures. What might be missing? The most obvious feature that might distinguish native structure from the incorrect structures is side-chain packing, but preliminary side-chain modeling studies have shown that many of the non-native structures accommodate the side-chains as well as the native-like structures (T. Lybrand, D. Alonso & V. Daggett, personal communication). An expert protein modeller concluded that low resolution crystal structures are more readily distinguished from true structures using standard molecular modeling tools than many of the non-native structures generated in this study (T. Lybrand, personal communication). Other possibilities include differences in configurational entropy (native structures may populate particularly broad energy minima) and kinetic accessibility.

There is considerable room for improvement of our current approach. A better expression for *P*(*structure*) (see equation (2)) should improve the results with β strand containing proteins. The expression for *P*(*sequence* | *structure*) should be improved by incorporating additional features of the residue environment such as secondary structure, avoiding the binary cutoff in the definition of residue burial (as indicated in Figure 3A, the cutoff at 16 neighbors misses most of the buried residues in small proteins; better results are obtained with lower cutoffs), and improving the treatment of the limited data problem. Finally, further improvements in the restriction of the conformational search space could partially compensate for imperfect scoring functions. Nearest neighbor methods are not optimal solutions to classification problems (Duda & Hart, 1973), and thus more sophisticated procedures may yield better sets of starting fragments; we are currently refining the sequence patterns which correlate with local structure toward this end (C. Bystroff & D.B., unpublished results). Improvements in the search procedure such as local moves (Elofsson *et al.*, 1995) and genetic algorithms are unfortunately not likely to help with the current scoring functions: even the crude simulated annealing method used here (10,000 cycles, ~three minutes of cpu time per structure) readily generated structures with better scores than the native structure.

As found in earlier studies (Mumenthaler & Braun, 1995; Park & Levitt, 1996), for any sequence there are many conformations with substantial amounts of native secondary structure in which almost all the hydrophobic residues are buried. Why the native state is so strongly preferred over plausible incorrect structures with buried hydrophobic residues and substantial secondary structure is an important challenge for current studies of protein folding.

Methods

Structure generation

Structures are represented using a simplified model consisting of the heavy atoms of the main-chain and the C^{β} atom of the side-chain. For glycine residues, a virtual C^{β} atom is used. All bond lengths and angles are held constant according to the ideal geometry of alanine (Engh & Huber, 1991); the only remaining variables are the backbone torsional angles.

We use a simple nearest neighbor procedure to define the conformational search space. In previous studies, we found stronger correlation between local sequence and local structure for nine residue fragments than for other fragment lengths of less than 15 amino acids (Bystroff *et al.*, 1996); thus we chose to build structures from segments of nine residues. Frequency distributions for the 20 amino acids at each position in the protein being folded and in proteins of known structure (pdb select_25 Oct 95 list, Hobohm *et al.*, 1992) were generated using multiple sequence information when available from the HSSP database (Sander & Schneider, 1991) and substitution matrix based pseudo-counts (Henikoff & Henikoff, 1996). For each segment *S* of length 9 in the protein being folded, the 25 nearest sequence neighbors in the structure database were identified using a simple distance measure (Han & Baker, 1995) that compares the amino acid frequency distributions at each position in the two segments:

$$DISTANCE = \sum_{i}^{9} \sum_{aa}^{20} \mid S(aa, i) - X(aa, i) \mid \qquad (9)$$

where *S*(*aa*,*i*) and *X*(*aa*,*i*) are the frequencies of amino acid *aa* at position *i* in nine residue segments of either the sequence being folded (*S*) or of one of the proteins in the pdb_select_25 set (*X*). The test of the nearest neighbor strategy described in Table 4 shows that the percentage of neighbors structurally similar to the true structure is greater when multiple sequence information is available and is considerably greater when either single or multiple sequence information is used than expected by chance.

Table 4. Similarity in structure of sequence nearest neighbors to the experimentally observed structure

Sequence information used to find fragments	% Similar
Multiple sequence alignment	20.8
Single sequence	17.5
Random sequence	8.0

25 nearest neighbors of nine-residue fragments were chosen for every position in a set of ten proteins (1aaj, 1edt, 1ifc, 1rec, 2pcdA, 1cskA, 1htp, 1lpe, 2ayh, 5p21; 1519 positions in total) using the simple city block metric as described in Methods. The percentage of the 25 neighbors within 1.0 Å dme from the native structures are indicated. Pseudo-counts were added to sequence profiles using the BLOSUM62 substitution matrix (Henikoff & Henikoff, 1992).

The conformation of each segment is chosen from the ensemble of structures adopted by these sequence nearest neighbors. Because the torsion space representation assumes ideal bond lengths and angles, considerable inaccuracies result from the use of torsion angles of PDB structures directly. To minimize these problems, a random torsion space search around torsion angles calculated from the crystal structure was conducted for each PDB structure to identify a configuration with ideal bond lengths and angles with low rmsd of atomic coordinates from the experimental structure. Torsional angles for the nearest neighbors were taken from these idealized structures. All homologs of the proteins (greater than 25% sequence identity) being folded were removed from the data set to eliminate bias in both the fragments and the scoring function.

The consistency of structure among the sequence nearest neighbors of a segment reports on the consistency of the sequence to structure mapping around the segment in question (Han & Baker, 1995), and thus the reliability of a prediction based on a set of nearest neighbors can potentially be assessed even in the absence of knowledge of the true structure (Yi & Lander, 1993). The correlation between consistency of structure among the nearest neighbors and the average similarity of the neighbors to the true structure is quite striking (Figure 10). This correlation can be used to choose optimal fragment sets for building up a structure from the many local segments with different boundaries and lengths which cover each position in the sequence.

The starting configuration in all simulations was the fully extended chain. A move consists of substituting the torsional angles of a randomly chosen neighbor at a randomly chosen position for those of the current configuration. The junctions between fragments were not constrained and thus the local structure repertoire is somewhat larger than that defined by the nearest neighbor sets alone. Moves which bring two atoms within 2.5 Å are immediately rejected; other moves are evaluated according to the Metropolis criterion using equation (6). Simulated annealing was carried out by reducing the temperature from 2500 to 10 linearly over the course of 10,000 cycles (attempted moves). To verify that native-like structures exist in the conformational space defined by the neighbor sets, simulated annealing runs were carried out for each structure starting from an extended chain using the distance matrix error (dme) as the scoring function (Table 5; Bowie & Eisenberg, 1994). The quality of the match was considerably better for the shorter proteins, but still respectable for the longer proteins. This sets the "gold standard" for the structurally unbiased simulations described in Results. Local moves (Elofsson *et al.*, 1995) and genetic algorithms (Pederson & Moult, 1996) might substantially improve the generation of native-like structures.

Fidelis *et al.* (1994) suggested that building structures from fragments was likely to be unsuccessful because of the structural divergence of local fragments and the weakness of local sequence-structural correlation. The success of our procedure and others like it may be due to two reasons: one, reasonable tertiary structures can be assembled without near perfect local structural agree-

A) Variability

position		segment	length		
	3	5	7	9	11
30	31.81	32.91	23.54	23.76	19.54
31	39.82	24.62	18.42	20.83	23.17
32	38.07	25.30	19.64	21.53	27.17
33	37.47	28.97	22.95	21.92	29.59
34	56.93	35.24	40.00	31.42	30.97
35	38.11	45.75	45.61	38.67	36.66
36	59.49	55.11	43.81	39.81	38.38
37	63.96	54.75	43.08	45.21	43.75
38	57.03	62.82	45.29	40.16	49.58
39	46.27	33.27	52.45	36.20	49.16
40	50.67	35.05	45.45	35.16	40.57

B) rmsd vs. native

position		segment	length		
	3	5	7	9	11
30	26.97	26.15	18.66	18.32	16.54
31	32.16	19.59	15.01	16.35	19.71
32	30.67	20.33	15.92	16.72	21.67
33	28.98	22.52	17.96	17.41	22.58
34	46.93	28.46	31.95	23.11	26.42
35	47.30	45.22	44.80	42.19	43.05
36	56.05	55.42	48.51	47.40	45.33
37	66.81	53.57	47.99	49.51	47.85
38	50.44	49.17	34.00	29.63	43.42
39	37.70	22.86	37.09	29.75	40.97
40	42.45	29.73	32.13	29.51	33.87

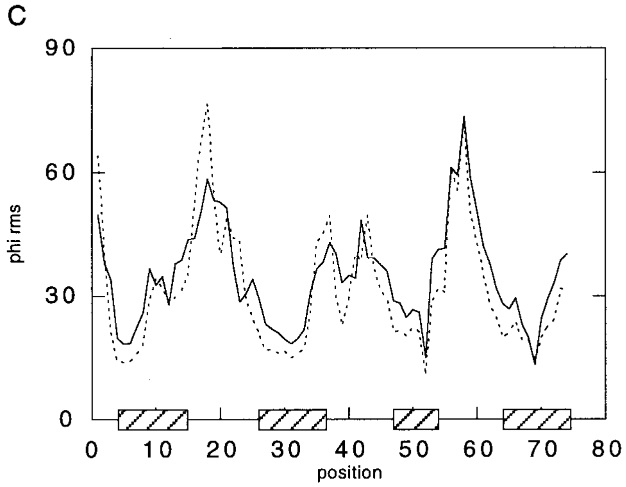

Figure 10. Correlation between structural variability and similarity to the true structure in nearest neighbor sets. For A and B, sequence nearest neighbors were identified for a series of different segment lengths for positions 30 to 40 of calbindin. A, Average rmsd in phi over residues i, $i+1$, and $i+2$ for the neighbor sets indicated at the top. B, Average rmsd from the true structure of calbindin for the same neighbor sets. The lowest values in each row are indicated in red; there is a strong correlation between low variability in the neighbor sets and low rmsd from the true structure. C, Variability in phi and rmsd from the native structure for entire calbindin sequence. Each position is represented by the segment with the lowest variability (red in A). The four helices in the native structure are indicated by hatched bars.

Table 5. Native-like structures can be generated from the nearest neighbor fragment sets in less than 10,000 cycles of simulated annealing

Protein	Fold	rmsd (Å)
Homeodomain	(α)	0.87
Calbindin	(α)	3.45
Protein A	(α)	0.42
Cro repressor	(α)	2.97
Protein G	(α/β)	4.61
Ribosomal fragment	(α/β)	3.78

The sum of squares of the differences in the distances between pairs of residues in the native and simulated structures was minimized by simulated annealing. Only distances between pairs of residues separated by less than 15 Å in the native structure were considered. The rmsd of the most native-like structure generated in 100 runs is indicated.

ment between simulated and native structures and two, even weak local sequence biases can significantly affect the likelihood of generating different tertiary structures.

Scoring functions

Implementation of equation (8) requires definition of the residue environments. There is an obvious tradeoff between the increase in quality of the representation with the number of different environments considered and the reduction in the amount of data available for the estimates in the second term in equation (8). For this paper we have adopted a compromise approach: the environment is defined in both terms of equation (8) by the number of C^{β} atoms of other residues within 10 Å of the residue in question, but in the first term where there is plenty of data, each number of neighbors is considered a separate environment, whereas in the second term, only two environments defined by a binary cutoff at 16 neighbors are used. Limited data problems were crudely treated by neglecting terms determined by fewer than ten observations in the database.

Noise in the scoring functions is more of a problem in *ab initio* folding than in threading because configurations can relax to fit the noise. The HSSP database of multiple sequence alignments for proteins of known structure (Sander & Schneider, 1991) was used to increase the number of counts contributing to the scoring functions. Because the different sequences in a multiple sequence alignment are not independent, each pair of residues contributed a factor of *1/(number of sequences in the family)* to the relevant pairwise histograms; the total contributions of each family to the scoring functions are thus equal. To assess the quality of the additional information in multiple sequence alignments, we constructed scoring functions corresponding to equation (5) for a frequently occurring pair, isoleucine-leucine, using a subset of the protein families such that the total number of counts was roughly the same as that for very rare pairs of amino acids when all families were used. To clarify the differences between the functions, the small data correction was not made for this test. The scoring function constructed from the multiple sequence alignments for the small protein subset (Figure 11, open circles) is much closer to the scoring function derived from all of the families (Figure 11, filled circles) than is the scoring function constructed from single sequences (Figure 11, open triangles). Thus, multiple sequence alignments

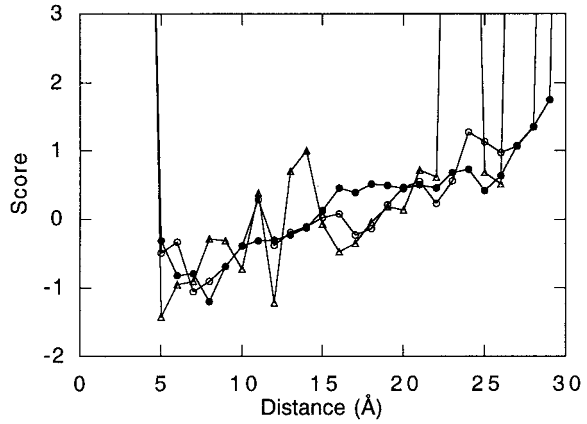

Figure 11. Multiple sequence information can reduce small sample size related noise in scoring functions. The negative logarithm of equation (5) for isoleucine-leucine pairs was calculated using either the entire pdbselect 25 set of proteins (filled circles), a randomly selected protein subset of the pdbselect set (triangles), or the same subset and the corresponding HSSP files to augment the numbers of residue pair counts (open circles). The multiple sequence information considerably reduces the noise in the small subset functions without biasing them substantially away from the best estimate obtained using all of the data.

should improve the quality of the scoring functions for rare amino acid pairs. Unless otherwise indicated, multiple sequence alignments were used for both fragment picking and in the generation of the scoring functions in all simulations.

Assessment of structure quality

For each protein under study, 100 simulations were performed using a variety of different conditions (summarized in Table 1). The average rmsd from the native state for random compact structures increases with chain length (Cohen & Sternberg, 1980) and thus the number of low rmsd structures is expected to be greater for shorter sequences. To derive a length independent measure of the success of simulations, we evaluated the mean and standard deviation of the rmsd from the native state for randomly generated compact structures with the same length as the native sequence (random all, Table 1). Following Cohen & Sternberg (1980) we define a length independent quality factor Q:

$$Q = \frac{\langle rmsd \rangle_{simulated} - \langle rmsd \rangle_{random\ compact}}{\sigma_{random\ compact}} \qquad (10)$$

where $\langle rmsd \rangle_{\text{simulated}}$ is the average rmsd from the native state among the 100 structures generated for each simulation condition, and $\langle rmsd \rangle_{\text{random}}$ and σ_{random} are the mean and the standard deviation in rmsd from the native state for the random compact structures.

Protein coordinates were taken from the Brookhaven National Archive (Bernstein *et al.*, 1977). Homeodomain, 1HDD chain C (Kissinger *et al.*, 1990); calbindin, 4ICB (Svensson *et al.*, 1992); protein A, 1FC2 chain C (Deisenhofer, 1981); cro repressor, 2CRO (Mondragon *et al.*, 1989b), 434 repressor, 1R69 (Mondragon *et al.*, 1989a); protein G, 2GB1 (Gronenborn & Clore, 1991); ribosomal fragment, 1CTF (Leijonmarck & Liljas, 1987).

Acknowledgments

We thank Christopher Bystroff, Karen Han, Tom Defay, Nathan Hunt, Paul Thomas, Kam Zhang, David Teller and Valerie Daggett for discussion and critical reading of the manuscript. K.T.S. was supported by PHS NRSA T32 GM07270 from NIGMS. C.K. was supported in part by ONR grant N-00014-96-0192 and by NSF grant DMS 9403371. This work was partially supported by NSF Cooperative agreement BIR-9214821 and young investigator awards to D.B. from the NSF and the Packard foundation.

References

Avbelj, F. & Moult, J. (1995). Determination of the conformation of folding initiation sites in proteins by computer simulation. *Proteins: Struct. Funct. Genet.* **23**, 129–141.

Bauer, A. & Beyer, A. (1994). An improved pair potential to recognize native protein folds. *Proteins: Struct. Funct. Genet.* **18**, 254–261.

Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein databank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

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

Bowie, J. U., Clarke, N. D., Pabo, C. O. & Sauer, R. T. (1990). Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins: Struct. Funct. Genet.* **7**, 257–64.

Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science,* **253**, 164–70.

Bystroff, C., Han, K. F., Simons, K. T. & Baker, D. (1996). Local sequence-structure correlations in proteins. *Curr. Opin. Struct. Biol.* **7**, 417–21.

Cohen, F. E. & Sternberg, M. J. (1980). On the prediction of protein structure: The significance of the root-mean-square deviation. *J. Mol. Biol.* **138**, 321–33.

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.

Defay, T. & Cohen, F. (1996). Multiple sequence information for threading algorithms. *J. Mol. Biol.* **262**, 314–323.

Deisenhofer, J. (1981). Crystallographic refinement and atomic models of a human FC fragment and its complex with fragment B of protein A from staphylococcus areus at 2.9 and 2.8 angstroms resolution. *Biochemistry,* **20**, 2361–2370.

Duda, R. O. & Hart, P. E. (1973). *Pattern Classification and Scene Analysis,* John Wiley & Sons, New York.

Elofsson, A., LeGrand, S. M. & Eisenberg, D. (1995). Local moves: an efficient algorithm for simulation of protein folding. *Proteins: Struct. Funct. Genet.* **23**, 73–82.

Engh, R. & Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta. Crystallog. sect. A*, **47**, 392–400.

Fidelis, K., Stern, P. S., Bacon, D. & Moult, J. (1994). Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng.* **7**, 953–960.

Flockner, H., Braxenthaler, M., Lackner, P., Jaritz, M., Ortner, M. & Sippl, M. J. (1995). Progress in fold recognition. *Proteins: Struct. Funct. Genet.* **23**, 376–386.

Gronenborn, A. M. & Clore, G. M. (1991). A novel, highly stable fold of the immunoglobin binding domain of streptococcal protein G. *Science,* **253**, 657–661.

Han, K. & Baker, D. (1995). Recurring local sequence motifs in proteins. *J. Mol. Biol.* **251**, 176–187.

Han, K. & Baker, D. (1996). Global properties of the mapping between local sequence and local structure in proteins. *Proc. Natl Acad. Sci. USA,* **93**, 5814–5818.

Henikoff, J. G. & Henikoff, S. (1996). Using substitution probabilities to improve position-specific scoring matrices. *CABIOS,* **12**, 135–143.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915–10919.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Sci.* **1**, 409–417.

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.

Jernigan, R. L. & Bahar, I. (1996). Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* **6**, 195–209.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature,* **358**, 86–89.

Kissinger, C. R., Liu, B., Martin-Blanco, E., Kornberg, T. B. & Pabo, C. O. (1990). Crystal structure of an engrailed homeodomain/DNA complex at 2.8 angstroms resolution: a framework for understanding homeodomain/DNA interactions. *Cell,* **63**, 579–590.

Kocher, J. P., Rooman, M. J. & Wodak, S. J. (1994). Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J. Mol. Biol.* **235**, 1598–613.

Kolinski, A. & Skolnick, J. (1994). Monte Carlo simulations of protein folding. II. Application to Protein A, ROP, and Crambin. *Proteins: Struct. Funct. Genet.* **18**, 353–366.

Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.

Laskowski, R. A., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291.

Leijonmarck, M. & Liljas, A. (1987). Structure of the C-terminal domain of the ribosomal protein L7/L12 from *Escherichia coli* at 1.7 A. *J. Mol. Biol.* **195**, 555–79.

López-Hernández, E., Cronet, P., Serrano, L. & Muñoz, V. (1997). Folding kinetics of CheY mutants with enhanced native [alpha]-helix propensities. *J. Mol. Biol.* **266**, 610–620.

Miyazawa, S. & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623–644.

Mondragon, A., Subbiah, S., Alamo, S. C., Drottar, M. & Harrison, S. C. (1989a). Structure of the amino-terminal domain of phage 434 repressor at 2.0 angstroms resolution. *J. Mol. Biol.* **205**, 189–200.

Mondragon, A., Wolberger, C. & Harrison, S. C. (1989b). Structure of phage 434 cro protein at 2.35 angstroms resolutions. *J. Mol. Biol.* **205**, 179–188.

Monge, A., Friesner, R. A. & Honig, B. (1994). An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. *Proc. Natl Acad. Sci. USA,* **91**, 5027–5029.

Monge, A., Lathrop, E. J. P., 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.

Munoz, V. & Serrano, L. (1996). Local versus nonlocal interactions in protein folding and stability – an experimentalist's point of view. *Folding Design,* **1**, R71–R77.

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

Park, B. H., Huang, E. S. & Levitt, M. (1997). Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* In the press.

Pederson, J. T. & Moult, J. (1996). Genetic algorithms for protein structure prediction. *Curr. Opin. Struct. Biol.* **6**, 227–231.

Rank, J. & Baker, D. (1997). A desolvation barrier to the hydrophobic cluster formation may contribution to the rate limiting step in protein folding. *Protein Sci.* **6**, 347–354.

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

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.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. *J. Mol. Biol.* **213**, 859–883.

Sippl, M. J. (1993). Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comp. Aided Mol. Design,* **7**, 473–501.

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. T. & Dill, K. A. (1995). A simple protein folding algorithm using a binary code and secondary structure constraints. *Protein Eng.* **8**, 769–778.

Svensson, L. A., Thulin, E. & Forsen, S. (1992). Proline cis-trans isomers in calbindin D9K observed by X-ray crystallography. *J. Mol. Biol.* **223**, 601–606.

Thomas, P. D. & Dill, K. A. (1996). Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* **257**, 457–69.

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

Yi, T. M. & Lander, E. S. (1993). Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol.* **232**, 1117–1129.

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

Edited by F. E. Cohen

(Received 14 August 1996; received in revised form 27 January 1997; accepted 27 January 1997)