Protein Engineering vol.6 no.6 pp.605–614, 1993

Global folding of proteins using a limited number of distance constraints

Maureen J.Smith-Brown, Dorothea Kominos[1] and Ronald M.Levy

Department of Chemistry, Rutgers University, New Brunswick, NJ 08903, USA

[1]Present address: Department of Molecular Biophysics, The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA

A Monte Carlo method is presented which can obtain the correct tertiary fold of a protein given the secondary structure and as few as three interactions between each secondary structure unit. This method was used to fold hemerythrin, flavodoxin, bovine pancreatic trypsin inhibitor and a variable light domain from an immunoglobulin using the known secondary structures of these proteins. Each of the proteins was successfully folded to obtain a structure resembling the initial X-ray structure. Reasonable success was also achieved when using a secondary structure prediction algorithm to assign secondary structure. The r.m.s. deviations between the folded proteins and the crystal structures are in the order of 3–5 Å for the backbone coordinates. Evaluation of the r.m.s. deviations between members of the globin family indicates that two equivalent overall folds may have r.m.s. deviations of this or even larger magnitude. The limiting number of constraints necesssary to achieve the correct fold is discussed.

Key words: distance constraints/Monte Carlo/protein folding/tertiary structure

Introduction

X-ray crystallography and NMR spectroscopy have both been successful in determining the 3-D structures of proteins. Currently approximately 600 protein structures have been solved using X-ray crystallographic techniques and approximately 100 structures have been determined using NMR data. The number of protein sequences is growing at a much greater rate than the number of protein structures (Bowie *et al.*, 1991). Currently, the number of sequences outnumbers that of 3-D structures by some three orders of magnitude. Not all proteins are amenable to structural analysis by either single crystal X-ray diffraction techniques or NMR techniques. Molecular modeling techniques also fail when no homologs of known structure exist. The prediction of a protein structure from its primary sequence and the pathway by which the protein folds into its native structure, collectively known as the 'protein folding problem', have yet to be solved. The protein folding problem greatly limits the utility of sequence information in determining structural and functional aspects of proteins. The combined use of molecular modeling techniques with limited experimental information may help bridge this gap.

Current structure prediction techniques [see Schultz (1988) or Garnier (1990) for a review] cannot accurately predict protein structures or derive a 3-D structure given only the amino acid sequence. Statistical algorithms, such as the Chou and Fasman method (Chou and Fasman, 1974a,b, 1978; Prevelige and Fasman, 1989) and the Garnier–Osguthorpe–Robson (GOR) method (Garnier *et al.*, 1978; Gibrat *et al.*, 1987; Levin and Garnier, 1988; Garnier and Robson, 1989) are based on the statistical occurrence of a particular amino acid residue in a type of secondary structure. Statistical methods have only been successful to an accuracy of ~65%. Moreover, it has been shown that the same penta- and hexapeptide sequences may exist in many different structural forms in different proteins (Kabsch and Sander, 1984; Unger *et al.*, 1989), thus pointing to the importance of long-range interactions in determining structure and inherently limiting the accuracy of statistical methods. Similar in spirit to such methods are the neural network methods (Qian and Sejnowski, 1988; Holley and Karplus, 1989) which use neural networks to determine the likelihood of an amino acid being in a given conformation. The information currently available from these neural network algorithms is the relationship between primary sequence and secondary structure. In general these methods have slightly improved accuracy over that of the statistical algorithms. So-called knowledge-based methods depend on analysis of protein structure that goes beyond statistical occurrences of residues in secondary structure, i.e. lengths of typical α-helices, hydrophilicity of residues, etc. These methods have been used to determine accurately protein structure classes (Taylor and Thornton, 1983, 1984; Rooman *et al.*, 1990) and the placement of turns in many different sequences (Cohen *et al.*, 1986). It has been suggested (Rooman and Wodak, 1988) that although knowledge-based methods are intuitively more accurate than statistical methods, the knowledge database, i.e. the number of protein structures solved to atomic resolution, is still not large enough to determine accurately new structures. However, the concept of proteins consisting of a few well defined modular units (Baron *et al.*, 1991) suggests that perhaps there is enough structural information but that the correct approach to protein structure prediction has yet to be found. It is evident that currently available prediction schemes alone are not capable of accurately determining the 3-D structure of a protein from its amino acid sequence.

Various attempts to predict the 3-D fold of a protein of known secondary structure have been made. Ptitsyn and Rashin (1975) represented the α-helices of myoglobin as cylinders and examined the degree of exposure of the hydrophilic and hydrophobic groups in various conformations. Following the work of Richmond and Richards (1978) on the packing of the α-helices in myoglobin, Cohen *et al.* (1979, 1980, 1981, 1982) used a combinatorial approach to study the possible folds of proteins in a variety of protein classes. Recently, Taylor (1991) has attempted to predict the tertiary fold of all-α proteins by first predicting the secondary structure from multiply aligned sequences using pattern matching methods. Taylor then placed the most strongly predicted secondary structures on an idealized framework and generated chains tracing over this framework through a combinatorial process. Those folds that ranked well in terms of hydrophobic interactions and other constraints were evaluated in greater detail.

Yet another approach of tertiary structure prediction is that taken by Bowie *et al.* (1991) of studying the inverse protein

folding problem, i.e. which amino acid sequences can fold into a known 3-D structure. This method was successful in a number of cases, but unfortunately cannot be used to predict a structure for which no previous example is available. Goldstein *et al.* (1992) have used associative-memory Hamiltonians in combination with sequence alignment and simulated annealing techniques to derive the tertiary structures of a number of different proteins.

It is often possible to obtain some limited information from experimental techniques, such as fluorescence energy transfer or NMR. This information may include distances between groups in a protein, although complete sequence-specific resonance assignments may not have been obtained. One example is the protein α-lactalbumin, for which certain resonances were attributable to aromatic side chains, but for which sequence-specific resonance assignments proved difficult to obtain (Baum *et al.*, 1989). Very strong structural constraints are also available in the form of disulfide bridge information, which can be obtained experimentally. Our study was undertaken to investigate whether knowledge of the secondary structure, coupled with limited distance information is adequate to predict the tertiary fold of a protein. This type of study lies in contrast to high resolution structure generation and refinement based on a large number of experimental distance constraints. In order to obtain the global fold of a protein, we are using the whole protein backbone explicitly, coupled with an internal coordinate Monte Carlo procedure. Along with demonstrating that this methodology can be used to obtain a global fold, we also address several other points. Can a structure be folded if an incorrect secondary structure is assumed? What are the implications for determining the tertiary fold if a prediction scheme is used to assign the secondary structure? What is the smallest number of interactions necessary to define uniquely the global fold via this procedure?

Materials and methods

Protein structures as building blocks

One way of conceptualizing protein structures is to use a building block analogy. Proteins can be thought of as building blocks (i.e. secondary structure elements) of differing lengths, consisting of stiff cylindrical rods (α-helices), stiff extended chains (β-strands) and floppy regions that join the building blocks into the correct orientation. These conceptual building blocks can either be determined through knowledge of the structure (Kabsch and Sander, 1983) or through a secondary structure prediction scheme (Qian and Sejnowski, 1988). Once the approximate length and position of the secondary structure elements have been determined, the correct orientation between the secondary structure elements may be obtained by the use of two or more distances between each element. Such distance information may possibly be obtained from, for example, NMR or fluorescence energy transfer experiments.

These building blocks can fold together in different ways to produce distinct protein structure classes as described by Levitt and Chothia (1976). The structural classes include proteins which contain only α-structure (all-α proteins), proteins which contain only β-structure (all-β proteins), proteins with alternating regions of α- and β-structures (α/β proteins) and proteins which contain discrete regions of α- and β-structures ($\alpha + \beta$ proteins). We have attempted to fold one protein in each of these structural classes including hemerythrin (an all-α protein), the variable light domain of an immunoglobulin (an all-β protein), flavodoxin (an α/β protein) and bovine pancreatic trypsin inhibitor (PTI, an $\alpha + \beta$ protein).

Secondary structure assignments

In order to demonstrate whether we could determine the tertiary fold of a protein using known secondary structure and a few distances, we used the secondary structure assignments as defined by Kabsch and Sander (1983) for all the proteins listed above with the exception of hemerythrin. In the case of hemerythrin, the secondary structure assignments as noted in the Brookhaven Databank file were incorporated, as the secondary structure was not reported in the Kabsch and Sander paper.

The secondary structure prediction technique employed was that of Qian and Sejnowski (1988). They used neural networks to determine a set of weights which will give the most likely conformation of an amino acid based on a window of the six amino acids on either side. As with most prediction schemes, their method is generally better at determining the middle of a region of secondary structure than at pinpointing the end-points. In choosing a prediction scheme which is not based on homology to previously solved structures, we are not limiting ourselves to only studying proteins for which homologs have been found and studied.

Computational procedure

The calculations were executed via the program IMPACT (Kitchen *et al.*, 1990). IMPACT can be used to perform Monte Carlo, molecular dynamics or molecular mechanics calculations. For the purposes of our current Monte Carlo calculations, we used a modified version of the all-atom force field of Weiner *et al.* (1986). The bonds and angles were kept fixed and the energy function was as follows:

$$E(\mathbf{r}) = \sum_{\text{all dihedrals}} k_\phi[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} \right] + \sum_{\substack{\text{distance} \\ \text{constraints}}} w_d(d - d_0)^2 + \sum_{\substack{\text{torsion} \\ \text{constraints}}} w_\tau(\tau \times \tau_0)^2$$

where E is the total energy and k_ϕ is the dihedral angle constant. The second sum represents the Lennard–Jones contribution to the potential. The last two terms represent the distance and torsional constraints which are placed on the system, with w_d and w_τ being the weights on the constraint terms. The target distance and torsion values, d_0 and τ_0, represent either the upper or lower bounds of the range, depending on whether the value of d or τ lies above or below the range. The distance constraints are incorporated via a biharmonic potential. Therefore, the energy contribution is zero if a calculated distance is within the prescribed range. Torsional constraints are implemented in an analogous fashion. For the simulations in this paper, no electrostatic or hydrogen bond terms were included in the potential energy function.

The proteins were modeled using chains of glycine residues of the appropriate length. This greatly decreased computation time and allowed us to model the protein backbones without the complication of non-bonded repulsions caused by incorrect side chain conformations. One potential disadvantage to modeling just the polypeptide backbone, is that any distance information derived from NOE type data would have to be converted to that

appropriate for the atoms in the model (in this case, Cα atoms). This means that, for α-helices, where the strongest interactions are typically between side chains, large ranges on distance constraints would be needed. As a result, incorrect conformations might appear to satisfy a particular subset of the constraints but would interfere with the ability of the molecule to satisfy the entire set.

The secondary structure is first modeled into an extended protein chain. Two adjacent regions of secondary structure were given their ideal ϕ, ψ and ω angles. In some cases, the loop region in between two units of secondary structure was placed into a pseudohelical starting conformation, i.e. a helix-like structure with non-ideal torsion angles. This helps to keep the loop regions more compact and prevents some unfavourable non-bonded interactions, which can prevent the correct folding of these structures. In the early stages of the computation, only ϕ and ψ dihedral angles in the loop region and at the ends of the secondary regions were searched using a Metropolis Monte Carlo algorithm which has been previously described (Levy *et al.*, 1989). Generally, the initial variation in the dihedral angles was allowed to be in the order of 2°. This maximum variation automatically decreased throughout the course of the stimulation as the constraints became closer to being satisfied. A default temperature of 300 K was used throughout the simulations. The dihedral angles at the end-points were constrained to remain within 60° of their ideal values. The peptide ω angles were kept fixed at 180°. After the constraints were nearly satisfied or after the search produced few structural changes, the ϕ and ψ angles within the secondary structural units were varied, but were constrained to within 30° of ideality. It was found that preventing this variation could prohibit the proper folding of the protein. It must be kept in mind that the torsional angles of secondary structural units in real proteins show significant deviations from ideality (Barlow and Thornton, 1988). Typically, a minimum of 2000 Monte Carlo steps were used for initial placement of two adjacent secondary structure units and another 2000 steps of Monte Carlo was used while allowing the secondary structure to relax. The search was terminated when further steps of Monte Carlo made negligible improvement on the structure. Upon completion of this stage, another adjacent portion of secondary structure was placed into its ideal state. The constraints between that piece and the nearest section were added first. After those constraints were reasonably satisfied, constraints to other previously modeled sections were added in. This process was continued until either the entire protein was modeled or until it became impossible to satisfy the more recently added constraints. The modeling process is represented in Figure 1. The primary determination of the degree of success of a particular fold was through a visual comparison of the superimposed backbones of the model and crystal structures. The r.m.s. deviations between these structures were also calculated. Cohen and Sternberg (1980) studied the r.m.s. deviations between random protein segments of varying lengths. They determined that the relationship is approximately linear with the r.m.s. deviation being 0.0468 × (number of residues) + 9.25. We have used this equation to estimate the values of the r.m.s. deviations between random structures and the crystal structures and used these results as a basis for comparison.

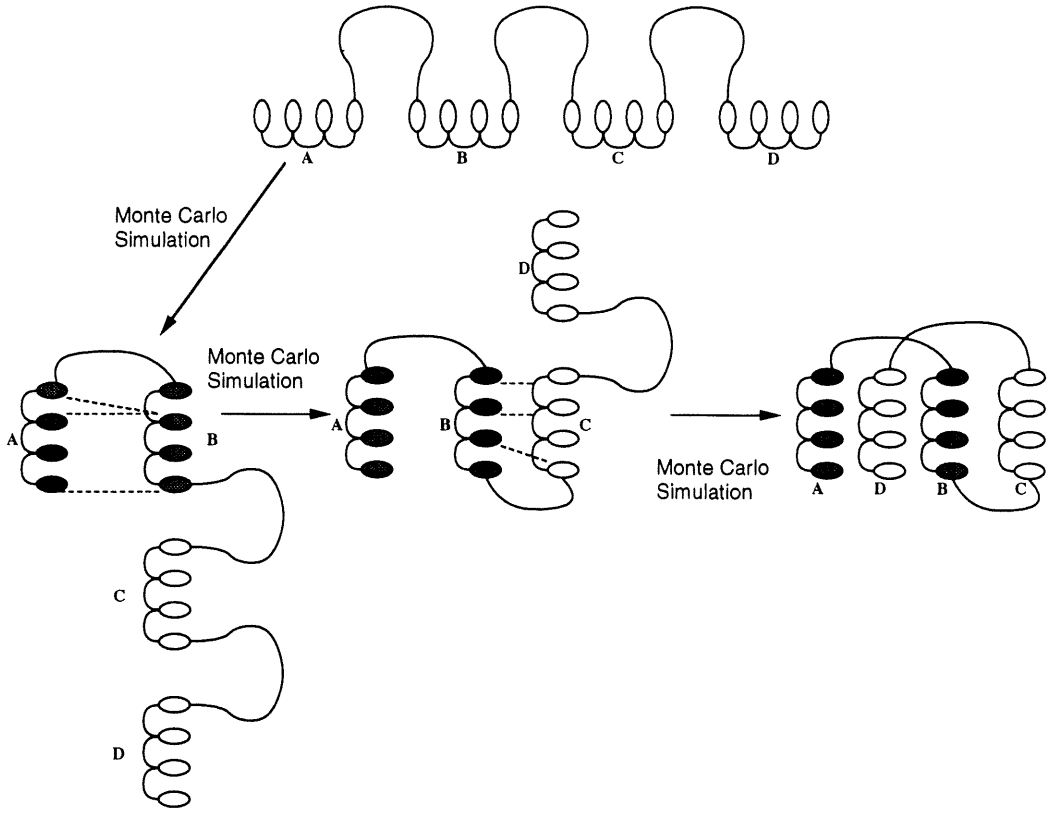

Fig. 1. Schematic diagram describing the folding of hemerythrin. Initially, the secondary structure is built into a polyglycine chain. Using three constraints between helices A and B, the relative orientation between the helices is defined. Then, the C- and D-helices are placed in analogous manner. Both redundant and non-redundant constraints are illustrated: when placing helices A and B, two constraints are shown to the same residue in helix B, while the constraints between helices B and C are unique. The interactions among non-adjacent helices, while present in the simulation, are not shown in the diagram for the sake of clarity.

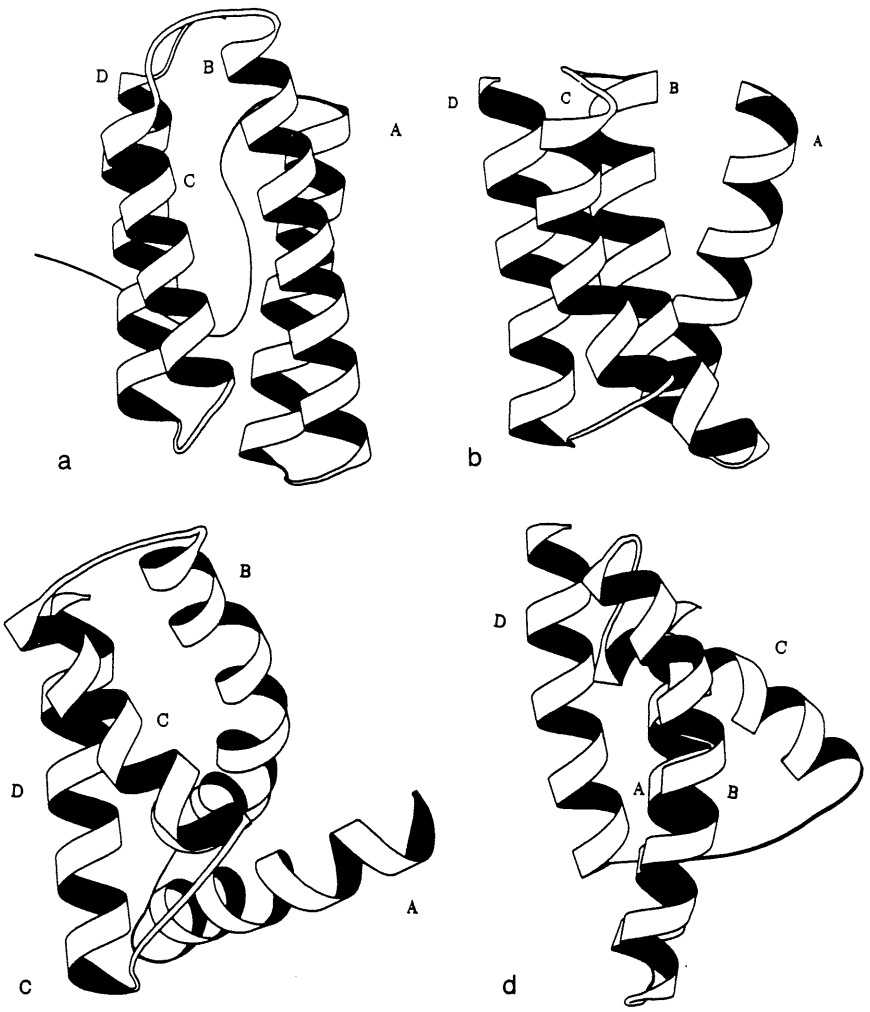

Fig. 2. Diagram of the protein hemerythrin. (**a**) Minimized crystal structure, (**b**) folded using constraint set CS1, (**c**) folded using constraint set CS2, (**d**) folded using constraint set CS3.

Results and discussion

Known secondary structure

Hemerythrin. Hemerythrin is a four-helix bundle protein whose coordinates are available through the Brookhaven Protein Databank (1HMQ; Bernstein *et al.*, 1977; Stenkamp *et al.*, 1983). A schematic diagram of hemerythrin is shown in Figure 2a. This crystal structure and those of the other proteins in this study were energy minimized after hydrogen atoms were added.

Three constraint lists were calculated for hemerythrin. The constraints were calculated based on the shortest Cα − Cα distances between secondary structure elements. These distances were calculated using the minimized X-ray structure. A range of ± 1 Å was placed on the constraints throughout the stimulations. Three different sets of constraints which each contained three constraints between each piece of secondary structure were used in order to study whether the choice of constraints used to generate the structures had a large effect on the resultant structure (see Table I). First, a list of 30 possible constraints was generated, containing the five shortest Cα − Cα distances between each element of secondary structure. From this list, three combinations were chosen of three interhelical constraints to yield three sets, each with 18 total constraints, sets CS1, CS2 and CS3. Using less than three constraints between pieces of secondary structure in hemerythrin caused incorrect helix orientations and therefore the structure could not be correctly folded. For set CS1, non-redundant constraints, i.e. ones not involving the same atoms within each interhelical constraint triad, were generated. Sets CS2 and CS3 were random samples of constraints.

The protein was folded in sections using the protocol described earlier. Initially the A and B helices were folded together, residues 19−37 and 41−64 respectively. This entailed 2000 steps of Monte Carlo using a scaled down Lennard−Jones 6−12 potential to represent the non-bonded repulsions. During the initial folding, only the loop and the end residues were sampled. A second set of 2000 steps was performed allowing the internal ϕ and ψ dihedral angles to vary within 30° of their ideal values. The C-helix was added in next, by first adding in the B−C constraints and then adding in the C−A constraints. Finally, the D-helix was added in an analogous fashion.

Figure 2b shows the result of the hemerythrin molecule folded using the constraint set CS1. This conformation contains the four α-helices in approximately the correct orientation. The ends, the N-terminus of the A-helix and the C-terminus of the D-helix, are further apart in the folded structure than in the crystal structure. The backbone r.m.s. deviation between residues 19−109 in the crystal and the folded structure is 5.26 Å. A large part of this deviation is caused by the separation of the two ends in the CS1 structure. A random structure of this length would be expected to have a 13.5 Å deviation from the crystal structure.

Table I Hemerythrin constraint sets

Set CS1 Cα−Cα	d_{lower}	d_{upper}	Set CS2 Cα−Cα	d_{lower}	d_{upper}	Set CS3 Cα−Cα	d_{lower}	d_{upper}
A−B constraints			A−B constraints			A−B constraints		
31−47	3.3	5.3	34−43	4.8	6.8	31−47	3.3	5.3
35−43	4.5	6.5	24−54	5.0	7.0	35−43	4.5	6.5
24−54	5.0	7.0	31−46	5.1	7.1	34−43	4.8	6.8
B−C constraints			B−C constraints			B−C constraints		
62−70	5.2	7.2	62−70	5.2	7.2	55−77	5.3	7.3
55−77	5.3	7.3	59−70	5.8	7.8	48−84	5.6	7.6
48−84	5.6	7.6	63−70	6.0	8.0	63−70	6.0	8.0
C−D constraints			C−D constraints			C−D constraints		
73−105	6.4	8.4	73−105	6.4	8.4	73−106	7.3	9.3
80−97	7.3	9.3	73−106	7.3	9.3	80−97	7.3	9.3
76−101	7.3	9.3	80−94	7.6	9.6	80−94	7.6	9.6
A−C constraints			A−C constraints			A−C constraints		
32−84	9.7	11.7	35−84	10.4	12.4	32−84	9.7	11.7
35−84	10.4	12.4	31−84	10.5	12.5	31−84	10.5	12.5
28−84	11.1	13.1	28−80	11.2	13.2	28−80	11.2	13.2
A−D constraints			A−D constraints			A−D constraints		
36−91	5.1	7.1	36−91	5.1	7.1	36−91	5.1	7.1
32−95	5.9	7.9	32−91	6.0	8.0	32−95	5.9	7.9
35−91	6.0	8.0	32−94	6.2	8.2	32−91	6.0	8.0
B−D constraints			B−D constraints			B−D constraints		
62−109	6.9	8.9	62−109	6.9	8.9	58−106	8.4	10.4
58−106	8.4	10.4	51−98	9.2	11.2	61−109	9.0	11.0
61−109	9.0	11.0	62−106	9.3	11.3	51−98	9.2	11.2

Figure 2c shows the hemerythrin molecule folded using the CS2 constraints. This structure shows less of a resemblance to the crystal structure. In the crystal, the A- and D-helices are roughly parallel to each other while in the CS2 structure, the angle between the two helices is closer to 90°. This may be due, in part, to the redundancy of the constraints between the C- and D-helices (73Cα − 105Cα and 73Cα − 106Cα) and the A- and D-helices (36Cα − 91Cα, 32Cα − 91Cα and 32Cα − 94Cα). The other interactions with the D-helix, i.e. the B−D interactions, are not sufficient to correctly define the orientation of the D-helix. The backbone r.m.s. deviation between this structure and residues 19−109 of the crystal structure is 8.41 Å, reflecting the incorrect tertiary fold of the CS2 structure.

The structure generated with constraint set CS3 is clearly incorrect (see Figure 2d). The orientation of the A-helix with respect to the B-helix is incorrect and is aggravated by a kink in the B-helix at residue 52. Surprisingly, residue 52 is not involved in any of the constraint interactions. Furthermore, the B- and C-helices are oriented approximately 90° with respect to one another and this in turn throws off the orientation of the D-helix with respect to all of the other helices. The backbone r.m.s. deviation of this structure to residues 19−109 of the minimized crystal structure of hemerythrin is 8.21 Å. Structure CS3 demonstrates the inability of the distance constraints alone to distinguish correct folding from incorrect folding; CS3 has a lower distance constraint violation than CS1 and CS2, but structure CS1 is most like that of the crystal structure.

Flavodoxin. Flavodoxin is an α/β protein of 138 residues. Its X-ray coordinates are available through the Brookhaven Databank (4FXN; Smith *et al.*, 1977). A diagram of flavodoxin is shown in Figure 3. Again, the constraints were based on Cα − Cα distances between secondary structure elements. For flavodoxin, more emphasis was placed on choosing three non-redundant constraints between secondary structures and the distances chosen were not necessarily among the shortest five, as a result of our experience with hemerythrin.

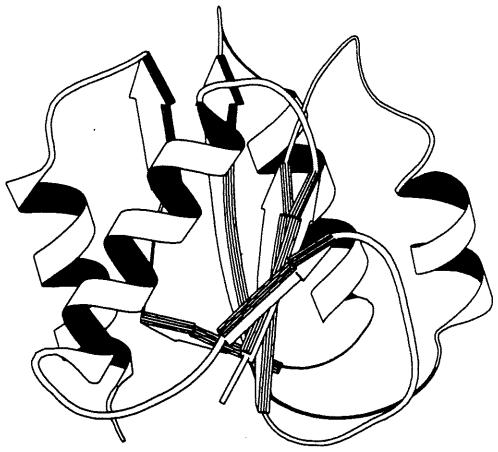

Fig. 3. Schematic diagram of flavodoxin.

During the course of folding flavodoxin it became apparent that the presence of a large number of residues between pieces of secondary structure greatly impedes the folding procedure. The greater number of degrees of freedom available to the loop region results in two difficulties. First, the greater flexibility inherent in long loop regions means that incorrect folds can appear to satisfy the constraints, but can interfere with the addition of subsequent sections of the protein. Second, if the loop region is too extended, it may buckle out in order to correctly position the secondary structure elements and may then have unfavorable non-bonded repulsions with other regions of the protein. The first problem can be overcome by increasing the number of constraints in troublesome regions. In the region of flavodoxin containing sheet 2 (residues 31−34), sheet 3 (residues 48−53) and helix

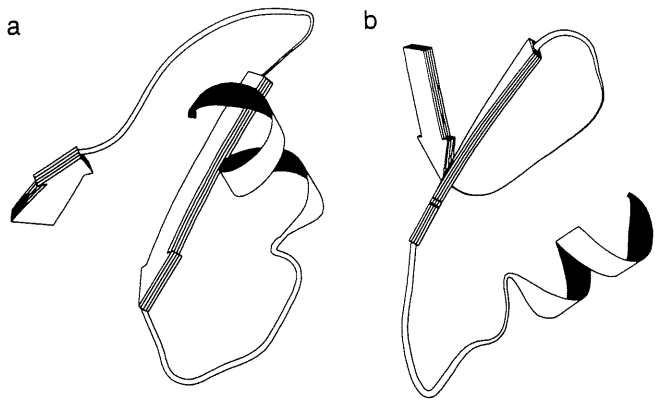

Fig. 4. The region of flavodoxin between residues 31 and 73. (**a**) Incorrect fold with 11 constraints, (**b**) the correct fold from the X-ray structure.

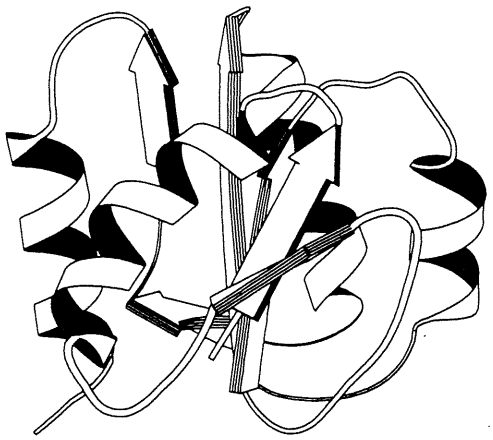

Fig. 5. Diagram of flavodoxin folded with 147 constraints.

2 (residues 66–73) there are two long loops, one 13 residues long and the other 14 residues long. Tests using three to four exact distances between each piece of secondary structure for the constraints in this region failed to yield the proper fold even though there was essentially no violation in the constraints (a total distance violation of 0.4 Å) (see Figure 4). In particular, note that the two β-strands are roughly parallel in the X-ray structure, but are antiparallel in the incorrectly folded structure. The use of six additional constraints allowed the proper fold to be achieved. The second problem, of extended loop structures causing unfavourable non-bonded interactions, can partially be alleviated by coiling up the loop region into a pseudohelical conformation at the beginning of the Monte Carlo simulation. During the course of the simulation, the loop will become more extended as torsion space is sampled, but is less likely to buckle out than if the initial starting structure is extended.

Flavodoxin has 10 different segments of secondary structure compared to hemerythrin, which has four. This larger number of structural units has both advantages and disadvantages. The most obvious disadvantage is the larger amount of computing time that is necessary to fold the protein. Another aspect is more subtle. When a segment is correctly placed, the addition of subsequent sections can improve the relative position of the segments. If a segment is incorrectly placed, however, it may make it impossible to satisfy future constraints and to position the ensuing segments. The folding of flavodoxin had to be restarted with slightly different conditions at varying points along the folding process after getting stuck in incorrect conformations,

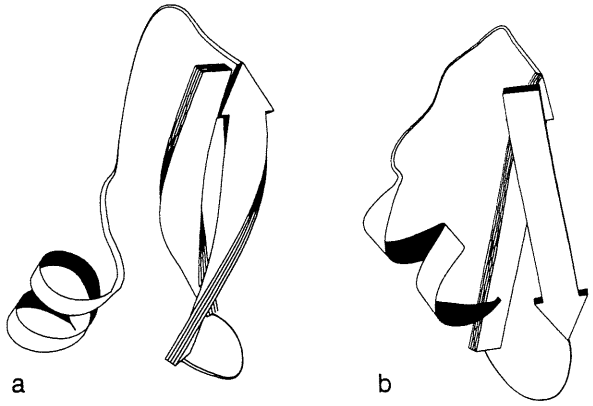

Fig. 6. Schematic diagram of PTI between residues 18 and 55. (**a**) Minimized X-ray structure, (**b**) folded using nine constraints.

but the end result, seen in Figure 5, has a backbone r.m.s. deviation of 3.18 Å, which is significantly better than the best result attained for hemerythrin. The expected random value between two protein segments 138 residues in length would be 15.7 Å. It should be noted that when using three constraints between each pair of units of secondary structure, there will be $3n(n-1)/2$ constraints, where n is the number of units, so that a much larger number of constraints was used for flavodoxin. There were a total of 147 constraints, which included extra constraints in the regions which were separated by large loops.

PTI. Bovine pancreatic trypsin inhibitor (PTI) is an $\alpha + \beta$ protein containing 58 amino acid residues. Its X-ray coordinates are available through the Brookhaven Protein Databank, (4PTI; Marquart *et al.*, 1983). A diagram of PTI is shown in Figure 6a. PTI has three distinct units of secondary structure; β-strands at residues 18–24 and 29–35 and a short α-helix at residues 48–55. A set of nine non-redundant Cα–Cα constraints were obtained from the minimized X-ray crystal structure. Once again, distance ranges of ±1 Å were placed on the constraints.

It was possible to obtain a reasonable fold of PTI using these nine constraints (see Figure 6b). The twist in the β-strands of PTI is not adequately represented in the model structure. The resultant structure has a backbone r.m.s. deviation of 5.1 Å to the crystal structure for the modeled region, between residues 18 and 55. A random structure of this length would be expected to have an 11.0 Å deviation. By allowing the secondary structure to be more flexible in the Monte Carlo simulation, it was expected that the twist would evolve during the simulation. However, the nine constraints were satisfied and any changes in the structure by Monte Carlo simulation were minimal. Furthermore, the twist in the β-structures arises naturally as a result of the presence of side chains (Salemme, 1983) which are not represented in this simulation. The addition of five new constraints still did not adequately represent the twist and, in fact, the r.m.s. deviation to the minimized crystal structure was slightly worse, 5.7 Å, even though the constraints were met without violation.

Immunoglobulin. The variable light domain of a human immunoglobulin (3FAB; Saul *et al.*, 1978) was modeled using our methodology. This domain is approximately a β-barrel structure of 101 residues (see Figure 7a).

As defined by Kabsch and Sander (1984), there are eight β-strands, at residues 9–12 (strand 1), 19–23 (strand 2), 35–40 (strand 3), 47–48 (strand 4), 57–62 (strand 5), 65–70 (strand 6), 79–87 (strand 7) and 90–101 (strand 8). The two β-sheets are made up of strands 1, 2, 5 and 6 and strands 3,

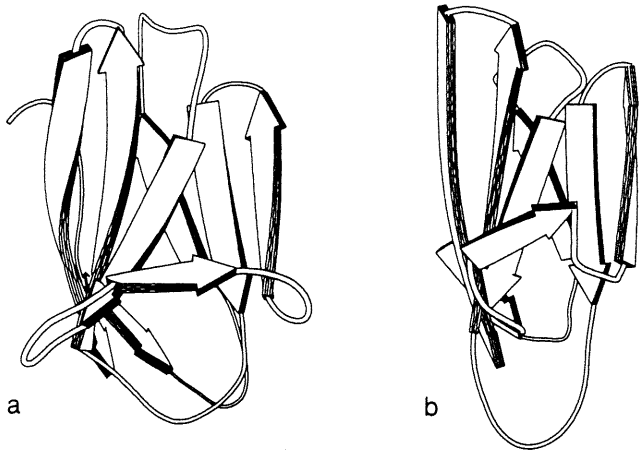

Fig. 7. Schematic diagram of the variable light domain of a human immunoglobulin. (**a**) Minimized crystal structure, (**b**) folded using 90 distance constraints.

Table II. Secondary structure of hemerythrin

X-ray structure[a]	Neural net prediction
–	17–20β
19–37α	28–35α
41–64α	42–50α and 58–64α
70–85α	69–72α and 80–82α
91–109α	92–100α

[a]Secondary structure as determined crystallographically (Stenkamp *et al.*, 1983).
[b]Prediction determined using the weights of Qian and Sejnowski (1988).

4, 7 and 8 respectively. This domain is quite challenging because of the presence of three loops which cross between the sheets. Similar to our experience of modeling flavodoxin, additional constraints were needed in order to model correctly the fold in the region of the long loop from residues 23 to 35 (between strands 2 and 3, which cross between the sheets). Without these additional constraints, folds were achieved which satisfied the constraints between strand 3 and strands 1 and 2, but which were incorrect folds and therefore prevented the successive folding of the molecule.

The variable light domain was successfully folded with a total of 90 constraints (see Figure 7b). This structure has a backbone r.m.s. deviation of 4.56 Å to the minimized crystal structure of this domain. A random structure would be expected to have a 13.1 Å deviation. No structural information about the CDR (complementarity determining regions) should be inferred from this modeled structure because the goal of this study is not to model the loops, but rather to obtain the correct orientation of the β-strands with respect to one another.

Using secondary structure from a prediction scheme

The neural network prediction scheme of Qian and Sejnowski (1988) was used to derive the secondary structure for the proteins we were modeling. Only two of the three regions were predicted for PTI, so not much would be accomplished by attempting to fold it using this secondary structure. For hemerythrin, flavodoxin and the immunoglobulin domain, the predicted secondary structure was treated as if it was believed to be the correct structure, so that the ϕ and ψ angles appropriate to the prediction

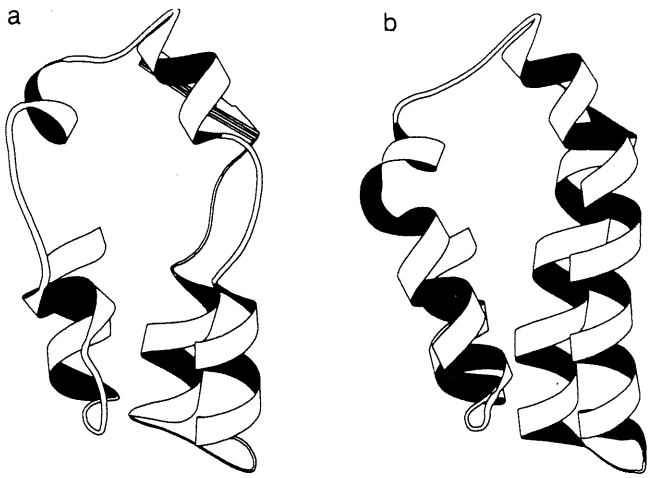

Fig. 8. Diagram of hemerythrin between residues 17 and 100 with secondary structure from a prediction scheme and 63 constraints. (**a**) Plotted using the predicted secondary structure, (**b**) plotted with placement of the helices along the chain as if the correct secondary structure were present in the model.

Table III. Secondary structure of flavodoxin

X-ray structure[a]	Neural net prediction[b]
2–6β	3–7β
11–25α	13–24α
31–34β	31–33β
–	40–44α
48–53β	49–51β
66–73α	70–75α
81–88β	81–84β
94–105α	97–102α
109–110β	108–110β
115–118β	–
125–136α	128–131α

[a]From Kabsch and Sander (1983) Table AIII, entry 12.
[b]Prediction determined using the weights of Qian and Sejnowski (1988).

were built into the individual regions. The constraints were once again based on the Cα distances of the crystal structure.

Hemerythrin. Table II compares the true secondary structure of hemerythrin to that determined by the prediction scheme. It is important to note that seven different regions are predicted, while only four exist in the actual structure. In using the same folding algorithm that was used earlier, there are now 63 constraints rather than 18. The seven predicted regions include a small β-region (residues 17–20) which does not exist in the X-ray structure, as well as two pairs of regions which when joined approximately comprise helix B and helix C respectively. The distance constraints are again derived from the actual distances in the X-ray structure. The structure obtained using these constraints is illustrated in Figure 8; it has approximately the correct fold and a backbone r.m.s. deviation from the crystal structure of 4.22 Å. This r.m.s. deviation is smaller than that of the corresponding structure folded with known secondary structure due to the greater number of constraints. It is interesting to note that the gaps in the B- and C-helices and the addition of a small sheet region did not prevent the structure from achieving a reasonable fold. This is not to say, however, that much larger errors in predicted secondary structure would be overcome.

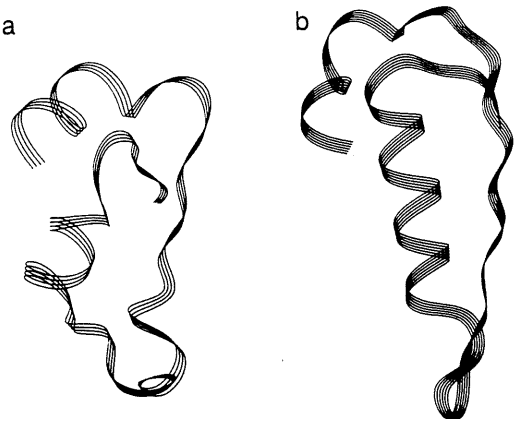

Fig. 9. Ribbon diagram of flavodoxin between residues 70 and 110. (**a**) Secondary prediction plus 18 distance constraints, (**b**) X-ray structure.

Table IV. Secondary structure of the variable light domain of a human immunoglobulin

X-ray structure[a]	Neural net prediction[b]
9–12β	–
19–23β	17–22β
35–40β	–
47–48β	49–51β
57–62β	57–59β
65–70β	66–70β
79–87β	80–81α
90–101β	90–92β / 100–101β

[a]From Kabsch and Sander (1983) Table AIII, entry 36.
[b]Prediction determined using the weights of Qian and Sejnowski (1988).

Flavodoxin. Table III compares the actual secondary structure of flavodoxin to that predicted by the neural network scheme. The prediction appears to be somewhat better than that for hemerythrin. The region of flavodoxin between residues 70 and 110 was folded using this secondary structure and 18 constraints. Figure 9 compares the fold achieved by this method with that of the same region of the crystal structure. The fold is quite good and has a backbone r.m.s. deviation from the minimized crystal structure of 4.96 Å.

Immunoglobulin. Table IV compares the known secondary structure of the variable light domain with that predicted using the neural network methodology. The prediction is good in some regions and poor in others. A two residue α-helix (residues 80–81) is predicted within an area where a sheet is present in the crystal structure (residues 79–87). The sheet between residues 90 and 101 is predicted at both end-points but not in the interior. Two of the earlier sheets are not predicted at all. Because of the sparsity of the prediction below residue 49, only the region between residues 49 and 101 was modeled. The resulting fold is reasonably good and has a backbone r.m.s. deviation of 4.99 Å (see Figure 10). It might be expected that the predicted helical region would interfere with the ability to fold up the all-β protein. The fact that it did not inhibit the folding is probably due to the fact that the region is so short. In fact, it is too short to define even one turn in a true helix.

Effect of using completely incorrect secondary structure

As one test of the necessity of having approximately the correct secondary structure, the four helical regions from the crystal

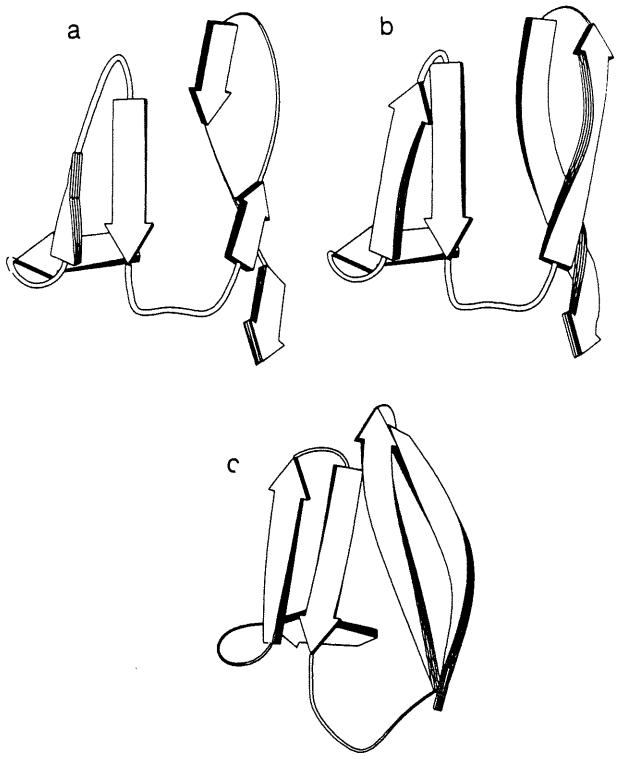

Fig. 10. Diagram of the immunoglobulin domain between residues 49 and 101 modeled with secondary structure from the prediction scheme. (**a**) Plotted using the predicted secondary structure, (**b**) plotted with the sheets located as if the correct secondary structure had been used, (**c**) region from the X-ray structure.

structure of hemerythrin were placed into the torsion angles appropriate for antiparallel β-sheets. An attempt was made to fold hemerythrin using the 18 constraints which were used earlier. The CS1 constraint set was chosen since this gave the most successful hemerythrin fold using the known secondary structure. For this set, it was impossible to satisfy the constraints using the Monte Carlo methodology presented. This is due to the fact that α-helices are much more compact than β-strands. Furthermore, the periodicity of an α-helix and a β-strand is different. The same face of an α-helix occurs approximately every four residues while for β-strands it occurs every two residues. This example clearly shows that incorrect secondary structure assignments will cause this algorithm to fail and thus indicate a problem with those assignments.

Limiting number of constraints

Our simulations thus far have used constraint sets which included interactions between each pair of secondary structure units. One would expect that some subset of these interactions would also be able to define uniquely the overall fold. A very simplistic model can be used to determine the minimum number of interactions that would be sufficient. We assume that each structural unit is merely a point in a plane and examine the number of distances among units necessary to pinpoint the location of a new point. Assume that the locations of points A, B and C are known. A new point D is to be placed in the plane. If only the C–D distance is known, D can be located anywhere on the circle surrounding C (see Figure 11a). If both the C–D and B–D distances are known, then the position of D is narrowed down to the two points where the circles surrounding B and C intersect (Figure 11b). Finally, if the A–D distance is also known

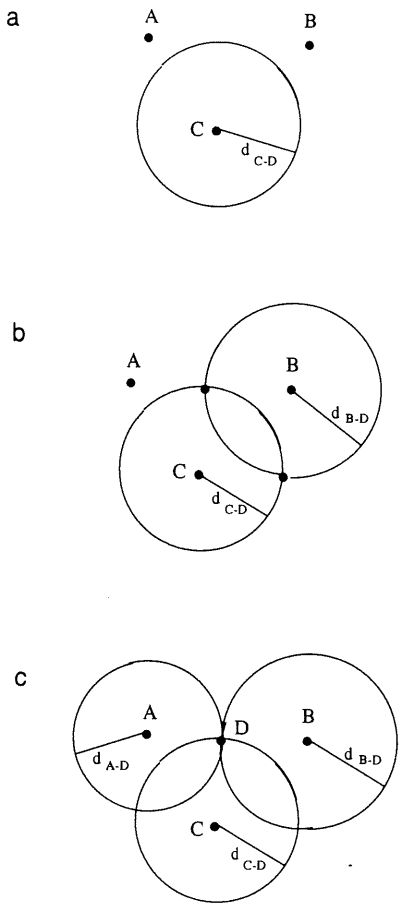

Fig. 11. Schematic diagram illustrating the minimum number of distances necessary to define the position of a fourth object when the locations of the previous three objects are known. (**a**) Only the C−D distance is known, so the location of D is somewhere on the circle of radius d_{C-D} surrounding C. (**b**) The C−D and B−D distances are known. This limits the location of D to the two points of intersection of the surrounding circles. (**c**) The A−D, B−D and C−D distances are all known. The position of D is now uniquely defined.

there is a unique possibility for the location of D, at the point where all three circles intersect (see Figure 11c).

We can expect from this illustration that we would need interactions to three previous secondary structure units to position uniquely a new unit. In practice, the presence of distance ranges might require additional interactions. A set of calculations was performed on flavodoxin, using the known secondary structure and a subset of the original constraint set. As each unit was added in, after the first three, constraints to only the previous three units were used rather than to all previous units. This cut the number of constraints used from 147 down to 84. The overall fold was still quite good, although the r.m.s. deviation to the crystal structure increased to 5.63 Å.

Our simple model indicated that using constraints to only the previous two units should limit the possible number of orientations at each step, but not uniquely define the structure, so that it would be possible for an incorrect fold to satisfy the constraints. A new set of calculations was performed on flavodoxin with this more limited constraint set, which now contains only 61 constraints. At the early stages in the simulation, the structure closely resembles that of the minimized X-ray structure. Eventually, however, the structure starts to deviate while still satisfying the distance constraints. The final structure has regions which match the minimized crystal structure but which are not correctly placed relative to each other. The total distance violation is only 0.136 Å, but the r.m.s. deviation is 11.86 Å.

Conclusions

It has been demonstrated that for representative proteins in each of the four different structural classes, it has been possible to achieve the correct tertiary fold using only secondary structure and a limited number of distance constraints. In the case of hemerythrin, three constraint sets were used to fold a model structure. The two constraint sets which were highly redundant, i.e. incorporated the same residue in more than one constraint between a pair of secondary structure units, led to structures which were incorrectly folded. It was also found that at least three constraints between adjacent secondary structure units were needed to define the relative position between two units. For a test case using two constraints between each pair of secondary structure units in hemerythrin, a flat four-helix layer was obtained. Two constraints were not enough to define the correct orientation of the helices with respect to each other and the final sets of constraints could not be satisfied due to incorrect placement of the second and third helices. In cases of long loop regions or very short secondary structural units, more than three constraints between units were sometimes needed to obtain the correct placement of units with respect to each other.

Our success using secondary structure from a neural network prediction scheme is particularly encouraging. The accuracy of the prediction has an effect on the final outcome, but as illustrated by our ability to fold hemerythrin with the somewhat faulty predicted secondary structure, the use of additional constraints can, in some cases, overcome inaccuracies in the prediction.

One issue which has not been fully addressed to this point is the magnitude of the r.m.s. deviations between the folded proteins and the minimized X-ray structures. If a comparison were to be made between X-ray structures determined by two different laboratories or between an X-ray structure and an NMR structure, r.m.s. deviations in the order of 5−8 Å would certainly be considered unacceptably large. When the overall fold between two proteins of similar structure is being compared, these numbers are, however, quite reasonable. If one compares, for example, the globins, hemoglobin (1HCO, 287 residues; Baldwin, 1980), leghemoglobin (1LH1, 153 residues; Arutyunan *et al.*, 1980) and myoglobin (1MBS, 153 residues; Scouloudi, 1978), one finds very similar r.m.s. deviations upon comparing regions of similar structure. The backbone r.m.s. deviation of myoglobin from leghemoglobin is 5.2 Å. For hemoglobin residues 187−287, compared to myoglobin residues 53−153, a backbone r.m.s. deviation of 8.5 Å is observed. The backbone r.m.s. deviation of hemoglobin residues 187−287 from leghemoglobin residues 8−153 is 7.4 Å. Thus, the r.m.s. deviation obtained for our folded structures are acceptable for our goal of obtaining the correct global fold of a protein. In addition, the magnitudes of our r.m.s. deviations are consistent with those of other investigators in the field (Cohen *et al.*, 1982; Goldstein *et al.*, 1992). For a more complete discussion of the significance of the r.m.s. deviations, see Cohen and Sternberg's (1980) paper.

We have presented a computational methodology by which the correct tertiary fold of a protein may be obtained. One possible application of this method is the piecing together of proteins from individual subunits which have been solved through NMR or other techniques.

Acknowledgements

This project was supported by National Institutes of Health Grant GM-30580 and by the Center for Biomolecular Simulations at Columbia University (*NIH* 5 P41 RR06892-02).

References

Arutyunan,E.G., Kuranova,I.P., Vainshtein,B.K. and Steigemann,W. (1980) *Kristallografiya*, **25**, 80–103.
Baldwin,J.M. (1980) *J. Mol. Biol.*, **136**, 103–128.
Barlow,D.J. and Thornton,J.M. (1988) *J. Mol. Biol.*, **201**, 601–619.
Baron,M., Norman,D.G. and Cambell,I.D. (1991) *Trends Biochem. Sci.*, **16**, 13–17.
Baum,J., Dobson,C.M., Evans,P.A. and Hanley,C. (1989) *Biochemistry*, **28**, 7–13.
Bernstein,F.C., Koetzle,T.F., WIlliams,G.J.B., Meyer,E.F., Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Bowie,J.U., Lüthy,R. and Eisenberg,D. (1991) *Science*, **253**, 164–170.
Chou,P.Y. and Fasman,G.D. (1974a) *Biochemistry*, **13**, 211–222.
Chou,P.Y. and Fasman,D.G. (1974b) *Biochemistry*, **13**, 222–245.
Chou,P.Y. and Fasman,G.D. (1978) *Annu. Rev. Biochem.*, **47**, 251–276.
Cohen,F.E. and Sternberg,M.J.E. (1980) *J. Mol. Biol.*, **138**, 321–333.
Cohen,F.E., Richmond,T.J. and Richards,F.M. (1979) *J. Mol. Biol.*, **132**, 275–288.
Cohen,F.E., Sternberg,M.J.E. and Taylor,W.R. (1980) *Nature*, **285**, 378–382.
Cohen,F.E., Sternberg,M.J.E. and Taylor,W.R. (1981) *J. Mol. Biol.*, **148**, 253–272.
Cohen,F.E., Stenberg,M.J.E. and Taylor,W.R. (1982) *J. Mol. Biol.*, **156**, 821–862.
Cohen,F.E., Abarbanel,R.M., Kuntz,I.D. and Fletterick,R.J. (1986) *Biochemistry*, **25**, 266–275.
Garnier,J. (1990) *Biochimie*, **72**, 513–524.
Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) *J. Mol. Biol.*, **120**, 97–120.
Garnier,J. and Robson,B. (1989) In Fasman,G. (ed.), *Prediction of Protein Structure and the Principles of Protein Conformation*. Plenum Press, New York, pp. 417–465.
Gibrat,J.-F., Garnier,J. and Robson,B. (1987) *J. Mol. Biol.*, **198**, 425–443.
Goldstein,R.A., Luthey-Schulten,Z.A. and Wolynes,P.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 4918–4922.
Holley,L.H. and Karplus,M. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 152–156.
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Kabsch,W. and Sander,C. (1984) *Proc. Natl Acad. Sci. USA*, **81**, 1075–1078.
Kitchen,D.B., Hirata,F., Westbrook,J.D., Levy,R., Kofke,D. and Yarmush,M. (1990) *J. Comput. Chem.*, **11**, 1169–1180.
Levin,J.M. and Garnier,J. (1988) *Biochim. Biophys. Acta*, **955**, 283–295.
Levitt,M. and Chothia,C. (1976) *Nature*, **261**, 552–558.
Levy,R.M., Bassolino,D.A., Kitchen,D.B. and Pardi,A. (1989) *Biochemistry*, **28**, 9361–9372.
Marquart,M., Walter,J., Deisenhofer,J., Bode,W. and Huber,R. (1983) *Acta Crystallogr. B*, **39**, 480–490.
Prevelige,P.,Jr and Fasman,G.D. (1989) In Fasman,G. (ed.), *Prediction of Protein Structure and the Principles of Protein Conformation*. Plenum Press, New York, pp. 391–416.
Ptitsyn,O.B. and Rashin,A.A. (1975) *Biophys. Chem.*, **3**, 1–20.
Qian,N. and Sejnowski,T.J. (1988) *J. Mol. Biol.*, **202**, 865–884.
Richmond,T.J. and Richards,F.M. (1978) *J. Mol. Biol.*, **119**, 537–555.
Rooman,M.J. and Wodak,S.J. (1988) *Nature*, **335**, 45–49.
Rooman,M.J., Rodriguez,J. and Wodak,S.J. (1990) *J. Mol. Biol.*, **213**, 327–336.
Salemme,F.R. (1983) *Prog. Biophys. Mol. Biol.*, **42**, 95–133.
Saul,F.A., Amzel,L.M. and Poljak,R.J. (1978) *J. Biol. Chem.*, **253**, 585–597.
Schultz,G.E. (1988) *Ann. Rev. Biophys. Biophys. Chem.*, **17**, 1–21.
Scouloudi,H. and Baker,E.N. (1978) *J. Mol. Biol.*, **126**, 637–660.
Smith,W.W., Burnett,R.M., Darling,G.D. and Ludwig,M.L. (1977) *J. Mol. Biol.*, **117**, 195–225.
Stenkamp,R.E., Sieker,L.C, and Jensen,L.H (1983) *Acta Crystallogr. B*, **39**, 697–703.
Taylor,W.R. (1991) *Protein Engng*, **4**, 853–870.
Taylor,W.R. and Thornton,J.M. (1983) *Nature*, **301**, 540–542.
Taylor,W.R. and Thornton,J.M. (1984) *J. Mol. Biol.*, **173**, 487–514.
Unger,R., Harel,D., Wherland,S. and Sussman,J.L. (1989) *Proteins*, **5**, 355–373.
Weiner,S.J., Kollman,P.A., Nguyen,D.T. and Case,D.A. (1986) *J. Comput. Chem.*, **7**, 230–252.

Received on February 8 1993; accepted April 12 1993