Prediction of the structure of proteins using related structures, energy minimization and computer graphics

David E. Stewart,* Paul K. Weiner† and John E. Wampler*

*Department of Biochemistry, University of Georgia, Athens, GA 30602, USA
†Department of Chemistry, Rutgers University, New Brunswick, NJ 08903, USA

Insight into the functions and interactions of proteins may be gained by correlating a variety of types of experimental data (including kinetics, spectroscopy, biophysical measurements, among others) with three-dimensional structural models displayed and manipulated using interactive computer graphics. Although tertiary structures have been determined for a large number of proteins, one limiting factor in structure-function studies is the lack of availability of the structural coordinates of specific proteins for which other types of detailed experimental data are known. However, as the data base of known structures grows, it becomes more and more likely that the structure of a closely related protein will be available. Here we present a method for predicting structures by (1) careful alteration of a known structure of a homologous, functionally analogous protein followed by (2) energy minimization to optimize the predicted structure. This method provides a rapid and effective solution to the initial problem of obtaining a working structure for modeling studies.

Received 22 April 1987
Accepted 25 June 1987

INTRODUCTION

Molecular modeling with computer graphics is becoming an increasingly used tool in the study of biological macromolecules and their interactions.[1-6] As powerful as this tool may prove to be, its use has been limited primarily to molecules with known three-dimensional structures. A major depository of such data, the Brookhaven Protein Data Bank,[7] contains over 300 sets of coordinate data for a wide variety of macromolecules, the vast majority being proteins. However, if none of these entries is the protein of interest, what role can this tool then play? One obvious possibility is to use the structure of a functionally related protein in the data base as a model for its functional analog. However, before this can be a widely used approach, it is necessary to be able to determine if the related protein will make a suitable model, and, if so, how to make it as close a structural analog of the protein of interest as possible.

Model building has been used previously for predicting the structures of unknown proteins. The approaches used can be classified into three categories:

(1) Predictions of the secondary and tertiary structure directly from amino acid sequence data[8-15]
(2) Models constructed by combining structural elements extracted from proteins in the data base where the sequence of amino acids in each segment is homologous to a section from the unknown protein[16-19]
(3) Models built by modifying a known structure of a homologous protein[20-27]

There are drawbacks to each of these methods, which should be recognized. Methods that attempt to predict the three-dimensional structure of proteins from their amino acid sequences have not yet been successful in obtaining the level of detail and certainty needed for studying molecular interactions and structure-function relationships.[8-15] While they can often demonstrate the ability to predict secondary structure within a class of proteins or within a test data base, their predictive ability for tertiary structure and for molecules outside of the developmental data bases falls considerably short.[14]

There are unavoidable hazards encountered in using an approach that involves linking together peptides extracted from several different proteins. The peptide sequences needed may not all be present in the data base, and those that are found will most likely not be located in exactly the same type of environment as that of the protein being modeled; neighboring group interactions, surface exposure and other variables will be different. Probably the most significant problem concerns the lack of functional homology between the proteins being used to supply the structural elements and the protein being modeled. Binding sites and catalytic sites — the sites for functionally specific interactions of a protein — are generally composed of residues from many different strands of the peptide chain. It is unlikely

0263-7855/87/030133-08 $03.00

Table 1. Proteins used in this study

Protein resolution (Å)	Organism	No. residues	Reference	Abbreviation
Azurin 2.0	*A. denitrificans*, strain NCTC 8582	129	33	AZAD
Azurin 2.7	*Pseudomonas aeruginosa*	128	34	AZPA
Phospholipase A2 2.1	Bovine pancreas	122	35	PA2B
Phospholipase A2 2.6	Porcine pancreas	124	36	PA2P
Rubredoxin 1.5	*D. vulgaris*	52	37	RBDV
Rubredoxin 1.2	*C. pasteurianium*	54	38	RBCP
Rubredoxin 1.5	*D. desulfuricans*, strain 27774	47	39	RBDD

that a model pieced together from fragments of many disparate proteins will correctly predict such important structural organization.

The third and most promising approach takes advantage of the similar tertiary structures of closely related proteins.[16,17,21,28,29] It involves superimposing the amino acid sequence of the protein being modeled onto the backbone of a functionally homologous protein. The first use of homology modeling was the prediction of the structure of α-lactalbumin based on the structure of hen egg white lysozyme by Browne *et al.*[21] Another use of this approach was the prediction of the structures of the calcium–binding proteins troponin-C[22,23] and calmodulin[26] based on the known structures of other calcium–binding proteins, such as parvalbumin. Recently, the three-dimensional structures of these proteins were solved from X-ray crystal studies.[30,31] It was found that the calcium–binding regions had been predicted accurately as was the overall helix-loop-helix supersecondary structure.[30,31] The regions away from the calcium–binding sites were predicted less well, but, even so, the models served as good tools for interpreting experimental results prior to solution of the X-ray structures.[30,31]

Structure models built by these methods may be improved by using energy minimization to optimize bond lengths and angles and to remove unfavorable interactions. Many of these approaches have compared the energetics of the predicted structure with those of the original structure. If the two were comparable, then the prediction was considered valid. Novotny *et al.*[32] showed that this criteria is unacceptable by using it with two proteins of entirely different structures, *Themiste dyscritum* hemerythrin (mainly an α-helical protein) and mouse myeloma immunoglobin (mainly β-sheets). The two proteins have five residues conserved out of a total of 113. Imposing the sequence of one on the backbone of the other followed by minimization resulted in a completely different folding. However, using the criteria of comparable energetics, these predictions were acceptable.

The most successful approach for modeling using homologous structures is the method of Bruccoleri and Karplus,[20] which samples the conformational space of short polypeptide segments in proteins using an empirical energy function. However, this method is computationally intensive and therefore limited to fairly short peptide chains.

Table 2. Comparison of crystal structures

Structures compared	Homology (%)	RMS Deviation—main chain atoms
Rb – Cp vs. Dd	60*	2.6758
Rb – Dv vs. Dd	69*	2.6651
Rb – Cp vs. Dv	67	0.5181
Az – Pa vs. Ad	63	0.8702
P2 – P. vs. B.	83	1.4818

*Does not include seven-residue deletion

In this study, we investigate a method to develop quickly a structural model using the homology modeling approach starting with a functionally related protein. The approach changes a known structure into the unknown by first making the necessary sequence changes with minimal disturbance in three-dimensional structure and then applying energy minimization to optimize the mutated structure. The accuracy of this approach is tested using known structures and a simple criteria for success based on the root-mean-square (RMS) difference between predicted and actual structures. It is our hope that this approach will increase the usefulness of molecular modeling to the study of proteins by enlarging the data base of structures that can be used to investigate structure-function relationships.

METHODS

The coordinate data for the proteins used, shown in Table 1, were obtained from the Brookhaven Protein Data Bank with the exception of the *Desulfovibrio desulfuricans* rubredoxin (RBDD).[39,40] The pairs used were chosen on the basis of a good degree of sequence homology and not on structural similarities (Table 2). Sequence alignment was based on the obvious criteria of the greatest degree of homology for the azurins and phospholipase A2s. The alignment of *D. desulfuricans* rubredoxin with the other rubredoxins was made as discussed in reference 39. The sequence alignments are shown in Figures 1–3.

```
                                               10
PA2B  *** LEU TRP GLN PHE ASN GLY MET ILE LYS CYS LYS ILE
PA2P  ALA --- --- --- --- ARG SER --- --- --- --- ALA ---

                                   20
PA2B  PRO SER SER GLU PRO LEU LEU ASP PHE ASN ASN TYR GLY
PA2P  --- GLY --- HIS --- --- MET --- --- --- --- --- ---

                       30
PA2B  CYS TYR CYS GLY LEU GLY GLY SER GLY THR PRO VAL ASP
PA2P  --- --- --- --- --- --- --- --- --- --- --- --- ---

           40                                      50
PA2B  ASP LEU ASP ARG CYS CYS GLN THR HIS ASP ASN CYS TYR
PA2P  GLU --- --- --- --- --- GLU --- --- --- --- --- ---

                                       60
PA2B  LYS GLN ALA LYS LYS LEU ASP SER CYS LYS VAL LEU VAL
PA2P  ARG ASP --- --- ASN --- --- --- --- --- PHE --- ---

                           70
PA2B  ASP ASN PRO TYR THR ASN ASN TYR SER TYR SER CYS SER
PA2P  --- --- --- --- --- GLU SER --- --- --- --- --- ---

               80                                      90
PA2B  ASN ASN GLU ILE THR CYS SER SER GLU ASN ASN ALA CYS
PA2P  --- THR --- --- --- --- ASN --- LYS --- --- --- ---

                                          100
PA2B  GLU ALA PHE ILE CYS ASN CYS ASP ARG ASN ALA ALA ILE
PA2P  --- --- --- --- --- --- --- --- --- --- --- --- ---

                              110
PA2B  CYS PHE SER LYS VAL PRO TYR ASN LYS GLU HIS LYS ASN
PA2P  --- --- --- --- ALA --- --- --- --- --- --- --- ---

                  120
PA2B  LEU ASP LYS LYS ASN CYS ***
PA2P  --- --- THR --- LYS TYR CYS

--- = IDENTICAL
*** = NOT PRESENT IN CRYSTAL STRUCTURE
```

Figure 1. Comparison of amino acid sequences of bovine and porcine pancreatic phospholipase A2

```
                                           10
AzPA  ALA GLU CYS SER VAL ASP ILE GLN GLY ASN ASP GLN MET
AzAD  --- GLN --- GLU ALA THR --- GLU SER --- --- ALA ---

                               20
AzPA  GLN PHE ASN THR ASN ALA ILE THR VAL ASP LYS SER CYS
AzAD  --- TYR ASP LEU LYS GLU MET VAL --- --- --- --- ---

                   30
AzPA  LYS GLN PHE THR VAL ASN LEU SER HIS PRO GLY ASN LEU
AzAD  --- --- --- --- --- HIS --- LYS --- VAL --- LYS MET

       40                                      50
AzPA  PRO LYS ASN VAL MET GLY HIS ASN TRP VAL LEU SER THR
AzAD  ALA --- SER --- --- --- --- --- --- --- --- THR LYS

                                   60
AzPA  ALA ALA ASP MET GLN GLY VAL VAL THR ASP GLY MET ALA
AzAD  GLU --- --- LYS GLU --- --- ALA --- --- --- --- ASN

                       70
AzPA  SER GLY LEU ASP LYS ASP TYR LEU LYS PRO ASP ASP SER
AzAD  ALA --- --- ALA GLN --- --- VAL --- ALA GLY --- THR

           80                                      90
AzPA  ARG VAL ILE ALA HIS THR LYS LEU ILE GLY SER GLY GLU
AzAD  --- --- --- --- --- --- --- VAL --- --- GLY --- ---

                                      100
AzPA  LYS ASP SER VAL THR PHE ASP VAL SER LYS LEU LYS GLU
AzAD  SER --- --- --- --- --- --- --- --- --- --- THR PRO

                          110
AzPA  GLY GLU GLN TYR MET PHE PHE CYS THR PHE PRO GLY HIS
AzAD  --- --- ALA --- ALA TYR --- --- SER --- --- --- ---

              120
AzPA  SER ALA LEU MET LYS GLY THR LEU THR LEU LYS ***
AzAD  TRP --- MET --- --- --- --- --- LYS --- SER ASN

--- = IDENTICAL
*** = DELETION
```

Figure 2. Comparison of amino acid sequences of Pseudomonas aeruginosa *and* A. denitrificans *azurin*

In a protein coordinate file, the entry for each coordinate point, or atom, includes the residue name, number and atom type. The proteins were mutated by selecting the nonconserved residues and changing the residue name and then deleting any nonappropriate atoms. Revised structures were completed by allowing a subroutine within the energy minimization suite AMBER[41] to replace missing atoms of altered residues and then minimizing the structure using the appropriate parameters.

```
                                           10
RBDV  MET LYS LYS TYR VAL CYS THR VAL CYS GLY TYR GLU TYR
RBCP  --- LYS --- --- THR --- THR --- --- --- --- ILE ---
RBDD  --- GLN --- --- VAL --- ASN --- --- --- --- GLU ---

                               20
RBDV  ASP PRO ALA GLU GLY ASP PRO THR ASN GLY VAL LYS PRO
RBCP  --- --- GLU ASP GLY --- --- ASP ASP --- --- ASN ---
RBDD  --- --- ALA GLU HIS --- ASN VAL *** *** *** *** ***

                   30
RBDV  GLY THR SER PHE ASP ASP LEU PRO ALA ASP TRP VAL CYS
RBCP  --- --- ASP --- LYS ASP ILE --- ASP --- --- --- ---
RBDD  *** *** PRO --- ASP GLN LEU --- ASP --- --- --- ---

       40                                      50
RBDV  PRO VAL CYS GLY ALA PRO LYS SER GLU PHE GLU ALA ALA
RBCP  --- LEU --- --- VAL GLY --- ASP GLU --- GLU GLU VAL
RBDD  --- VAL --- --- VAL SER --- ASP GLN --- SER PRO ALA

RBDV  *** ***
RBCP  GLU GLU
RBDD  *** ***

--- = IDENTICAL
*** = DELETION
```

Figure 3. Comparison of amino acid sequences of rubredoxins from D. vulgaris, C. pasteurianium, *and* D. desulfuricans

After the molecules were modified and minimized, RMS deviations were calculated by a program included in AMBER based on the method of Ferro and Hermans.[42] This program also outputs a coordinate file for the best fit superimposition.

Calculations were performed on a Digital Equipment Corporation VAX 11/780 running the VMS v4.4 operating system. Modeling was performed on an Evans & Sutherland PS340 utilizing the Evans & Sutherland MOGLI molecular modeling software.

RESULTS

The method of mutating a residue and revising the residue structure was varied to deduce the best overall method of prediction. Since a subroutine in the energy minimization package used, AMBER, adds missing atoms to a residue, there is substantial freedom in the choice of atoms deleted prior to completion and minimization of the revised structure.

Four approaches were tested. Structures were modified by (a) deleting the entire side chain of differing residues and constraining all common atoms, (b) deleting only the atoms that differ between the two residues with the correction for branched chain residues described below and constraining all remaining atoms, (c) deleting the side chain to the first common main side chain atom and constraining the remaining atoms, or (d) deleting atoms as in c, but applying constraints only to atoms of unaffected residues. An example of this can be seen in a change from valine to threonine. In the first approach the entire side chain is deleted, with AMBER adding all of the side chain atoms. With the second method, only atoms that are different are deleted, in this case one of the γ-carbons of valine. The last two methods involve deleting both of the γ-carbons, which leaves the first main side chain atom, the β-carbon.

As indicated, there are two variables in this method. The second is the degree to constrain the model to the original coordinates. Since this predictive method is based on the premise that proteins with similar sequences will have similar structures, it is desirable to use the original coordinates as constraints. The question arises as to how much weight the starting coordinates should have. AMBER includes an option that allows one to vary this weight, expressed in kcal mole^{-1}.

With AMBER, the constraints may be applied in several ways. The two most pertinent to this study were (1) constraining all the coordinates present after the differing atoms were deleted, or (2) constraining the coordinates for all but the affected residues.

After minimization, the RMS deviations were calculated for four different combinations of structural components: Type I, where all atoms present in both the crystal structure and the predicted structures were compared; Type II, where all residues except the ones that were changed were compared; Type III, where atoms present in both the crystal structure and the predicted structure were compared, with the exception of the atoms with RMS deviations much larger than the others; and Type IV, where only the main chain nitrogen, α-carbon and carboxy terminal atoms were compared. The justification for the Type III calculation is that manual manipulation of exterior side chains, where these deviations almost always occur, generally eliminates this high deviation. Color Plate 1 illustrates how the variable residues in rubredoxin are primarily found on the exterior and how large RMS position differences tend to be found in residues that are relatively free to move about. Since this freedom of movement exists and proteins are, after all, dynamic entities, some error of this type might be acceptable.

One point noted in examining the RMS deviations is that for some residues, namely, aspartic acid, glutamic acid, arginine and valine, the branched chain atoms in the "V," when they are replaced by AMBER, are named oppositely as compared to the crystal structure. This gives a falsely high RMS deviation, which is corrected by simply switching the position of these atoms in the data file. This also occurred with asparagine, glutamine and threonine. However, in this case, it may not be acceptable to switch the atoms, since an energetically unfavorable change might be made. The solution is to keep this source of error in mind when modeling and to correct empirically the configuration of these side chains, since they are flexible and can usually be manipulated.

A more extensive study of the variables of this method was carried out, using a combination of different constraining weights, constrained atoms and deleted atoms to determine the best approach. For this study a single pair of proteins were used with the structure of the rubredoxin from *Clostridium pasteurianum* (RBCP) being changed into that of *D. vulgaris* (RBDV). This choice was made because rubredoxins are fairly small proteins, thus allowing many calculations without using an inordinate amount of computer time.

After the best parameters were determined for the conversion from RBCP to RBDV, a control was run using RBDV as a starting point instead of RBCP. Atoms were deleted from the same residues so that the starting point for AMBER involved replacement of the same numbers of atoms as with RBCP. In this case, of course, the residue names were not changed. This control was done to determine the amount of deviation inherent in allowing AMBER to add atoms and optimize.

The modified structures were computed by AMBER with the results shown in Table 3. As can be seen from the data, the lowest RMS deviation appears to result from deleting the atoms to the first common main side chain atom and constraining all coordinates except for the affected residues with a constraning weight of 100 kcal mol^{-1}. The other molecules were changed using this scheme along with some random spot checking of other methods to ensure that this was indeed the most efficient method. The results are given in Table 4.

DISCUSSION

To assess the accuracy of this method, the RMS deviation between a crystal structure and its predicted structures was calculated. In addition, a criteria was needed to judge the accuracy of a prediction based on these RMS values. It was expected, and also observed, that as the resolution of the crystal structure decreased, the RMS deviation between the predicted and crystal structures increased. This is expected, since there is less certainty in the position of atoms in a lower resolution structure and therefore energy refinement would have a larger effect, even between the pair of crystal structures. Therefore, only a relative limit can be set, which depends upon the resolution of the structure.

Assignment of this limit is arbitrary and subject to individual viewpoint. When X-ray crystal structures are refined, significant movements of individual residues of several angstroms are common. Bruccoleri and Karplus[20] estimate that the experimental accuracy of X-ray crystallography and energy refinement is on the order of 1 Å. In addition, some of the differences noted between a modeled structure and an X-ray structure are due to differences in environment, such as crystal packing or solvent conditions. Lesk and Chothia[43] examined this in detail and found that these differences cause RMS deviations in the range of 0.25 to 0.40 Å for the main chain atoms. Temperature studies of sperm whale myoglobin[44] indicate that vibrational motion of side chain residues accounts for an average uncertainty in position of about 0.2 Å, while the main chain atoms show much less displacement (< 0.1 Å). Recent molecular dynamics simulations for myoglobin by Elber and Karplus,[45] on the other hand, suggest more flexibility with RMS differences between minimized structures of about 2 Å. The control experiment reported in Table 3 gave a RMS difference of 0.5 Å when a single protein was manipulated by the procedures outlined above and minimized. Thus we should not expect the method to give results closer than about 0.5 Å to the "true" structure, and, from a practical point of view considering the uncertainties of the actual X-ray structures, environmental effects and the contributions of dynamics to structural uncertainty, a reasonable criterion for testing the accuracy of this approach should be an RMS difference of around 1Å to 1.5 Å.

The best way to understand the relationship of the RMS deviations reported to differences in structure is to observe the superimposed structures with the dynamic graphics display. The program module that calculates the RMS deviations also outputs a coordinate file, allowing direct display of the two images at their minimum

Table 3. Prediction of the structure of rubredoxin from *D. vulgaris* using the crystal structure of rubredoxin from *C. pasteurianium*

Method	Constraint, Kcal mole^{-1}	RMS Deviation			
		I	II	III	IV
A	100	0.9345	0.6069	0.8644	0.5139
B	100	0.8571	0.6060	0.7554	0.5114
B	1000	0.8401	0.6044	0.7550	0.5163
C	50	0.8348	0.6093	0.7298	
C	100	0.8349	0.6072	0.7300	0.5163
C	1000	0.8375	0.6049	0.7331	0.5159
C	10000	0.8393	0.6045	0.7351	
D	50	0.8263	0.6307	0.7202	0.5162
D	100	0.8230	0.6073	0.7175	0.5140
D	1000	0.8263	0.6307	0.7202	0.5162
D	10000	0.8263	0.6037	0.7202	0.5162
D	100000	0.8263	0.6037	0.7202	0.5162
		CONTROL: *D. vulgaris to D. vulgaris*			
D	100	0.5171	0.0000	0.4609	0.2015

Columns

I = Comparison of all atoms after switching atoms named backward (see text)
II = Comparison using only atoms of common residues
III = Comparison of all atoms except for those with larger than average rms deviations
IV = Main chain atoms only

Methods

A: Delete all residues except N, CA, C of differing residues and allow AMBER to add the appropriate atoms which are missing. Constrain all atoms
B: Delete as few atoms as possible—rename if possible. For Asn-Asp, Glu-Gln, and Thr-Val delete just one of "V."
C: Delete as few as possible—but do not rename and delete both wings of "V." Constrain all atoms
D: As with C, but constrain only nonaffected residues

RMS orientation. Here it is clear that while the most successful modeling still gives side chains with slightly different orientations from that in the crystal structure, the main chain atoms have a lower deviation. This is expected and consistent with both experimental and theoretical studies.[20,43,44,45] The main chain atoms are the base, or anchor, for the more flexible side chains.

The structures of bovine and porcine phospholipase A2 (Color Plate 2A) clearly illustrate the overall approach and its results. Panel A shows main chain atoms of the superimposed crystal structures of these two proteins aligned for minimum main chain RMS deviation. Clearly, while the two structures are analogous with similar arrangements and regions of secondary structure, a variety of minor positioning differences are found, along with one major region of distinctly different folding. In the bovine structure, residues 59 through 67 form a helix of approximately two turns, whereas the porcine structure folds as a pleated sheet in this same region.

Upon comparing the superposition of the X-ray structure of bovine phospholipase A2 (PA2B) with its predicted structure from the porcine protein (Color Plate 2B) and the superposition of the X-ray structure of porcine phospholipase A2 (PA2P) with the structure predicted for it (not shown), the superimposed main chains are positioned nearly identically. Most of the small differences have been removed. However, the major difference in the region of residues 59 to 67 is still seen. The PA2B–based structures tend to retain the helix, while the PA2P–based structures keep the pleated sheet conformation. This highlights a weakness of this method in predicting major changes in the structure. The forces that change this region are probably not due to local influences, since most of the residues here are conserved except for the residues at either end—valine (PA2B) to phenylalanine (PA2P) at position 59 and asparagine to glutamate at position 67. Closeup examination of this region with surfaces displayed shows that the entire section is highly exposed on the exterior of the protein, but under the influence of the charges of residues of the adjacent helix. It may be that the difference in configuration in this region is influenced by the large number of altered residues in the adjacent helix (residues 50 to 58). Aside from this one area, the X-ray and predicted structures superimpose well.

In the case of *Pseudomonas aeruginosa* and *Azobacter denitrificans* azurins, while the sequence homology is lower than in the case of phospholipase A2 (63% versus 83%), the main chain RMS deviation is much better (0.87 versus 1.48; see Table 2). The large regions of substituted amino acids seem to have little effect on the position of the backbone. Thus, the main chain atom of the crystal and predicted structures in both cases are equally good fits (Table 4). When the main chain atoms of predicted structures are superimposed on the corresponding crystal structures in this case, little difference from the superposition of the crystal structures can be detected. With this pair of proteins the important differences lie in positioning of the side chain atoms. Comparing all common atoms, the RMS deviation is 2.59 Å, whereas after mutation and minimization this value falls to 1.6 (Table 4).

As can be seen from the sequence alignment of the rubredoxins in Figure 3, the proteins from all three sources have quite homologous sequences except for the

Table 4. Results of predicting structures

Method	Constraint Kcal mole^{-1}	RMS Deviation			
		I	II	III	IV
			Phospholipase A2, Bovine from Porcine		
D	100	2.1543	1.9948	1.0376	0.6706
			Phospholipase A2, Porcine from Bovine		
D	100	2.5367	1.9994	1.0295	0.6639
	Azurin (Ad) from Azurin (Pa)				
C	100	1.6338	1.5801	1.1539	0.8600
D	100	1.6207	1.5732	1.1489	0.8563
			Azurin (Pa) from Azurin (Ad)		
D	100	1.3572	1.1637	1.3001	0.8609
			Rubredoxin (Cp) from Rubredoxin (Dv)		
A	100	1.1570	0.5915	1.0984	0.5067
C	50	1.0792	0.5851	1.0048	0.5135
C	100	1.0769	0.5880	1.0031	0.5074
D	50	1.0668	0.6037	0.9975	0.5338
D	100	1.0668	0.6037	0.9975	0.5338
D	1000	1.0668	0.6037	0.9975	0.5338
			Rubredoxin, *D. desulfuricans* from *D. vulgaris*		
A	100	1.8399	0.6786	0.9788	0.5942
C	100	3.0610	0.6974	1.0905	0.6902
D	100	2.6123	0.6977	1.0749	0.6061
			Rubredoxin, *D. desulfuricans* from *C. pasteurianium*		
A	100	3.0188	0.8111	1.1775	0.7412
C	100	3.0327	0.8123	1.1409	0.7449
D	100	2.5916	0.8086	1.1300	0.6515

Columns

I: Comparison of all atoms after switching atoms named backward (see text)

II: Comparison using only atoms of common residues

III: Comparison of all atoms except for those with larger than average rms deviations

IV: Main chain atoms only

Methods

A: Delete all residues except N, CA, C of differing residues and allow AMBER to add the appropriate atoms which are missing. Constrain all atoms.

B: Delete as few atoms as possible—rename if possible. For Asn-Asp, Glu-Gln, and Thr-Val delete just one of "V." Constrain all atoms

C: Delete as few as possible—but do not rename and delete both wings of "V." Constrain all atoms

D: As with C, but constrain only nonaffected residues

seven-residue deletion in RBDD. This is also the only rubredoxin containing a histidine residue (residue 18). However, even in this case, the prediction of the structure of *D. desulfuricans* rubredoxin from the other two is surprisingly accurate (Color Plate 3). Panel A shows the common portion of the two crystal structures. Panel B shows the main chain atoms of the predicted structure of *D. desulfuricans* rubredoxin based on that from *D. vulgaris*. The two loose ends from the seven-residue deletion constitute a large portion of the RMS value. However, if the two ends are joined (Panel C), this difference is corrected. After being corrected for atoms that have RMS deviations much greater than the average, the value for the RMS deviation is below the limit of 1 Å. As can be seen (Table 4), the method that gave the best results in this case was not the same as the method that proved to be the most accurate in changes that did not involve gaps. The best results were achieved by allowing AMBER to add the entire side chain, not just a few atoms. This is not surprising, since this allows AMBER to place the side chains in the open space created by the missing residues and not constrained to the space occupied by the replaced residues.

The superposition of the X-ray structure of *D. vulgaris* rubredoxin (RBDV) and its predicted structure (not shown) shows essentially the same deviations as the superposition of the X-ray structure of RBCP and its predicted structure. Both pairs have similar RMS values to that of the main chains of the parent crystal structures. The X-ray and predicted structures have close alignment of the main chain atoms, and on close observation of the side chain atoms it is seen that with some manipulation they can also be superimposed very closely. The major difference between the X-ray and predicted structures for both pairs is at positions 34 and 35. In both, the residue at position 34 is proline and the residue in position 36 is aspartate. However, in RBDV the residue at position 35 is alanine and in RBCP it is aspartate. The difference could be due to the change from an uncharged, small side chain to a bulky, charged side chain.

Upon comparing the superposition of the X-ray structure of RBDD and the structure predicted from RBCP with the superposition of RBDD and the structure predicted from RBDV, few differences are observed. The structure predicted from RBDV is much closer to the crystal structure of RBDD than is the structure predicted from RBCP, as could be expected from the higher sequence homology. The major structural divergence noted in both predicted structures occurs in the gap caused by the seven-residue deletion. The main chain is interrupted here, a problem that is solved by using interactive computer graphics to rotate the peptide bonds to join the gap. The resulting structure is very close to the X-ray structure (Color Plate 1).

Within the limits of the sampling of structures and sequences we have examined, this method appears to work well. In all cases the proteins were homologous by several criteria: amino acid sequence, overall structural properties such as molecular weight, biochemical function and detailed structure. Perhaps the most important single aspect of homology that controls the probability for success with this approach is homology in biochemical function, not amino acid sequence. This fact is illustrated by comparisons of overall structureal homology versus amino acid sequence homology. Clearly, as shown in both Color Plates 2 and 3, functionally homologous proteins often have very homologous main chain structures. Even when amino acid sequence homology is low, structural homology can be strikingly consistent. For example, the known structures of cytochromes c_3 found in the Brookhaven files have low sequence homology, but very high structural homology. The *D. vulgaris* c_3 and the *D. desulfuricans* protein show only 24% sequence homology,[46] yet the four hemes of these two molecules can be almost exactly superimposed on each other. Thus, it is difficult to state a criterion by which one could judge the suitability of a particular protein for modeling by this approach. Certainly, if the structure of a functionally homologous protein is known and it exhibited similar physical and biochemical characteristics along with greater than 60% sequence homology, we can expect successful modeling. However, the approach might also be used in cases where there is less sequence homology, but where the other measures show considerable similarity. In particular, as seen in the previous studies with calcium–binding proteins, and in examining the structure of proteins containing large prosthetic groups, much less sequence homology may be allowable in some cases.

CONCLUSIONS

The following conclusions may be drawn from these observations.

1. Major changes in the conformation of the main chain on the exterior of the protein will not be predicted by this method. This was illustrated by the failure of the method to predict the change from pleated sheet conformation to α-helix, which was observed for the short exterior segment in phospholipase A2. In this method heavy constraints are placed on the coordinates of residues, which are not different between the two sequences; only very localized effects are modeled. Thus, changes induced by longer range interactions or crystal packing are not modeled. However, we might expect, as with phospholipase, that such errors will be a small fraction of the overall structural approximation.
2. The primary premise used here is that functionally homologous proteins with homologous amino acid sequences will have similar structures. In our tests all of the sequences used had homologies of 60% or higher.
3. The adjustments of structure by this method may not precisely compensate for large changes in charge and/or bulk. For example, an unresolved problem was observed when a small, uncharged side chain was replaced by a bulky, charged side chain, or vice versa.
4. Replaced side chains are generally folded into the same relative positions in the predicted and X-ray structures. The errors seen are primarily due to free rotation. When the crystal structure is known, these are easily corrected empirically. In the absence of such data, these alternate conformations are likely to be valid representations of low energy conformations that are part of the dynamics of the structure *in vivo*.
5. The method described can be used with proteins where the sequences alignment is 1:1, or where small deletions are present. It does not take into account insertions.

This method is based on the premise that functionally proteins with homologous sequences will have closely related structures. Subject to the limitations and guidelines enumerated, it does appear to allow the structure of a homologous protein to be used to develop an initial model. This model can then be used for selecting potential sites for site-directed mutagenesis and for interpreting biophysical and spectral results. It can thus serve the role that models always should serve of stimulating experimentation to validate, refute or refine the model. As more and more models of individual structures are developed and the X-ray data base expands, this approach should add to the knowledge of structure-function relationships and the degree to which evolutionary development conserves three-dimensional structure.

Further work is now in progress to expand these methods to account for insertion of residues. This technique is being used to investigate the structural implications of spectroscopic data obtained with homologous proteins where only some of the X-ray structures are available. We are also investigating how homologous sequences must be in order to use this method, in particular, whether other structural components, such as cofactors, act as structural constraints in compensation for lacks of homology.

ACKNOWLEDGEMENTS

The authors thank Steve Follin of ETA Systems, Inc., for his help with use of the Evans & Sutherland PS340 Molecular Graphics Workstation, and ETA Systems, Inc., for time on this instrument. We are also indebted to Dr. Charles Bender of the UGA Advanced Computational Methods Center for use of the VAX 11/780 and financial support. The coordinates for rubredoxin from *D. desulfuricans*, strain 27774, were supplied by Dr. Ronald Stenkamp and Dr. Larry Sieker of the Department of Biological Structure of the University of Washington, Seattle, WA, and Dr. Jean LeGall of the Department of Biochemistry, University of Georgia, Athens, GA.

REFERENCES

1 Poulos, T. L. and Kraut, J. *J. Biol. Chem.* 1980, **255,** 10322
2 Getzoff, E. D., *et al. Nature* 1983, **306,** 287
3 Tainer, J. A., *et al. Nature* 1983, **306,** 284
4 Mauk, M. R., *et al. Biochemistry* 1986, **25,** 7085

5 Feldmann, R. J., *et al. Proc. Natl. Acad. Sci. USA* 1978 **75,** 5409

6 Leszczynski, J. F. and Rose, G. D. *Science* 1986, **234,** 849

7 Bernstein, F. C., *et al. J. Mol. Biol.* 1977, **112,** 535

8 Garnier, J., Osguthorpe, D. J. and Robson, B. *J. Mol. Biol.* 1978, **120,** 97

9 Burgess, A. W., and Scheraga, H. A. *Proc. Natl. Acad. Sci. USA* 1975, **72,** 1221

10 Chou, P. Y. and Fasman, G. D. *Biochemistry* 1974 **13,** 211

11 Wu,. T. T. and Kabat, E. A. *J. Mol. Biol.* 1973, **75,** 13

12 Hopp, T. P. and Woods, K. R. *Proc. Natl. Acad. Sci. USA* 1981, **78,** 3824

13 Sweet, R. M. and Eisenberg, D. *J. Mol. Biol.* 1983, **171,** 479

14 Lim, V. I. *J. Mol. Biol.* 1974, **88,** 873

15 Taylor, W. R. and Thornton, J. M. *Nature* 1983, **301,** 540

16 de la Paz, P., *et al. EMBO J.* 1986, **5,** 415

17 Sweet, R. M. *Biopolymers* 1986, **25,** 1565

18 Greer, J. *J. Mol. Biol.* 1981, **153,** 1027

19 Greer, J. *J. Mol. Biol.* 1981, **153,** 1043

20 Bruccoleri, R. E. and Karplus, M. *Biopolymers* 1987, **26,** 137

21 Browne, W. J., *et al. J. Mol. Biol.* 1969, **42,** 65

22 Kretsinger, R. H. and Barry C. D. *Biochim. Biophys. Acta* 1975, **405,** 40

23 Tufty, R. M. and Kretsinger, R. H. *Science* 1975, **187,** 167

24 Warme, P. K., *et al. Biochemistry* 1974, **13,** 768

25 Endres, G. F., Swenson, M. K. and Scheraga, H.A. *Arch. Biochem. Biophys.* 1975, **168,** 180

26 Aulabaugh, A., *et al. Eur. J. Biochem.* 1984, **143,** 409

27 Chothia, C., *et al. Science* 1986, **233,** 755

28 Chothia, C. and Lesk, A. *EMBO J.* 1986, **5,** 823

29 Lesk, A. M., Levitt, M. and Chothia, C. *Protein Engineering* 1986, **1,** 77

30 Babu, Y. S., *et al. Nature* 1985, **315,** 37

31 Herzberg, O. and James, M. N. G. *Nature* 1985, **313,** 653

32 Novotny, J. Bruccoleri, R. and Karplus, M. *J. Mol. Biol.* 1984, **177,** 787

33 Norris, G. E., Anderson B. F. and Baker, E. N. *J. Mol. Biol.* 1983, **165,** 501

34 Adman, E. T., Sieker, L. C. and Jensen, L. H. *Isr. J. Chem.* 1981, **21,** 8

35 Dijkstra, B. W., *et al. Biochemistry* 1984, **23,** 2759

36 Dijkstra, B. W., *et al. J. Mol. Biol.* 1983, **168,** 163

37 Adman, E. T. and Jensen, L. H., *Am. Cryst. Assoc., Abstr. Papers* 1979, **6,** 65

38 Watenpaugh, K. D., Sieker, L. C. and Jensen, L. H. *J. Mol. Biol.* 1980, **138,** 615

39 Sieker, L. C., *et al. Febs* 1986, **208,** 73

40 These coordinates were kindly provided to us by Dr. Ronald Stenkamp at the University of Washington (ref. 39) and Dr. Jean LeGall at the University of Georgia.

41 Weiner, P. K. and Kollman, P. A. *J. Comp. Chem.* 1981, **2,** 287

42 Ferro, D. R., and Hermans, J. *Acta Crystallog. Sect A* 1977, **33,** 345

43 Lesk, A. M. and Chothia, C. H. *Phil. Trans. R. Soc. London A* 1986, **317,** 345

44 Frauenfelder, H., Petsko, G. A. and Tsernoglou, D. *Nature* 1979, **280,** 558

45 Elber, R. and Karplus, M. *Science* 1987, **235,** 318

46 Haser, R., *et al. Nature* 1980, **282,** 806