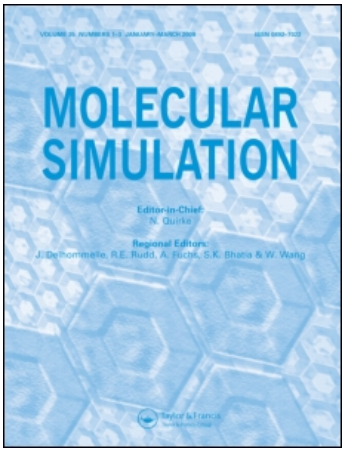


Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Predicting the Structure of the Flavodoxin from *Escherichia coli* by Homology Modeling, Distance Geometry and Molecular Dynamics

T. F. Havel[a]
[a] Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA

To cite this Article Havel, T. F.(1993) 'Predicting the Structure of the Flavodoxin from *Escherichia coli* by Homology Modeling, Distance Geometry and Molecular Dynamics', Molecular Simulation, 10: 2, 175 — 210
To link to this Article: DOI: 10.1080/08927029308022164
URL: http://dx.doi.org/10.1080/08927029308022164

Molecular Simulation, 1993, Vol. 10(2-6), pp. 175-210
Reprints available directly from the publisher
Photocopying permitted by license only

PREDICTING THE STRUCTURE OF THE FLAVODOXIN FROM *Eschericia coli* BY HOMOLOGY MODELING, DISTANCE GEOMETRY AND MOLECULAR DYNAMICS[1]

T.F. HAVEL

Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115

(Received August 1992, accepted October 1992)

As part of our on-going development of a method, based upon distance geometry calculations, for predicting the structures of proteins from the known structures of their homologues, we have predicted the structure of the 176 residue Flavodoxin from *Escherichia coli*. This prediction was based upon the crystal structures of the homologous Flavodoxins from *Anacystis nidulans*, *Chondrus crispus*, *Desulfovibrio vulgaris* and *Clostridium beijerinckii*, whose sequence identities with *Escherichia coli* were 44%, 33%, 23% and 16%, respectively. A total of 13,043 distance constraints among the alpha-carbons of the *Escherichia coli* structure were derived from the sequence alignments with the known structures, together with 8,893 distance constraints among backbone and sidechain atoms of adjacent residues, 978 between the alpha-carbons and selected atoms of the flavin mononucleotide cofactor, 116 constraints to enforce conserved hydrogen bonds, and 452 constraints on the torsion angles in conserved residues. An ensemble of ten random *Escherichia coli* structures was computed from these constraints, with an average root mean square coordinate deviation (RMSD) among the alpha carbons of 0.85 Ångstroms (excluding the first 1 and last 6 residues, which have no corresponding residues in any of the homologues and hence were unconstrained); the corresponding average heavy-atom RMSD was 1.60 Å.

Since the distance geometry calculations were performed without hydrogen atoms, protons were added to the resulting structures and these structures embedded in a $50 \times 50 \times 40$ Å solvent box with periodic boundary conditions. They were then subjected to a 20 picosecond dynamical simulated annealing procedure, starting at 300 K and gradually reduced to 10 K, in which all the distance and torsion angle constraints were maintained by means of harmonic restraint functions. This was followed up by 1000 iterations of unrestrained conjugate gradients minimization. The goal of this energy refinement procedure was not to drastically modify the structures in an attempt at *a priori* prediction, but merely to improve upon the predictions obtained from the geometric constraints, particularly with regard to their local backbone and sidechain conformations and their hydrogen bonds. The resulting structures differed from the respective starting structures by an average of 1.52 Å in their heavy atom RMSD's, while the average RMSD among the heavy atoms of residues 2–170 increased slightly to 1.66 Å. We hope these structures will be good enough to enable the phase problem to be solved for the crystallographic data that is now being collected on this protein.

KEY WORDS: *Escherichia Coli*, homology modeling, distance geometry, molecular dynamics.

INTRODUCTION

Because the rate at which DNA and proteins are being sequenced vastly exceeds the rate at which the corresponding protein structures can be determined, the problem of predicting the tertiary structure of proteins from their primary structure

[1]This work was supported by NIH grants GM-38221 and GM-47467.

continues to grow in importance [1]. Although considerable progress has been made at finding minimum energy conformations of small peptides, and even for larger proteins with sufficiently simple structural models and potential functions [2,3,4], it seems safe to say that it will be many years before these methods can achieve the reliability available from established experimental methods like nuclear magnetic resonance (NMR) and X-ray crystallography (cf. [5]). The steadily increasing gap between the supply of and the demand for the structures of new proteins, together with the observation that most protein domains fall into a relatively small number of "structural classes", has led many investigators to develop empirical methods of predicting the structural class from the sequence (see e.g [6, 7, 8, 9, 10]. Coordinates for the backbone atoms may be derived from such predictions by taking them from idealized models representing the corresponding class, after which the sidechain conformations may be added on by various energy minimization techniques [11, 12, 13, 14].

Another, related approach that has generally proven more effective is known as "homology modeling" [15, 16, 17, 18, 19]. This approach relies on having the structure of at least one, and preferably several homologous proteins available, from either previous NMR or crystallographic studies. Although the details of this procedure vary, its broad outline is now well-established and may be broken down into the following basic steps:

- First, the "structurally conserved regions" (SCR's) of the protein family in question are identified, generally by searching for segments of well-defined secondary structure that can be closely superimposed upon each other in all of the known homologous structures.
- Second, the sequences of the known structures are aligned so that the residues in each SCR match, and the sequence of the unknown protein is aligned with them so as to maximize the homology subject to the constraint that no insertions or deletions occur in any region that is aligned with an SCR.
- Third, the known structures are aligned in space by minimizing the RMSD between corresponding alpha-carbons in all of the SCR's together. The backbone coordinates in the SCR whose sequence is judged most similar to the unknown protein's are then copied to obtain backbone coordinates for the SCR's of the unknown protein. As a rule, the coordinates of any sidechain atoms that are the same as in the unknown protein are also copied to obtain coordinates for the corresponding atoms in the unknown.
- Fourth, coordinates are obtained for those regions, generally loops between regions of well-formed secondary structure, that do not lie in an SCR by means of either database searches or by various modeling techniques. Coordinates are also obtained for mutated sidechains and sidechains not in any SCR by similar methods.
- Finally, the resulting coordinates are refined by energy minimization and/or dynamics, with the goal of eliminating distortions in the covalent structure between SCR's and any steric clashes between the added sidechains that result from the foregoing "cut-and-paste" procedure.

Recently [20] it has been shown that distance geometry algorithms [21] can be used to construct not just one, but a large variety of initial model structures, and with substantially less effort than that required for a single structure by the cut-and-paste procedure. In addition, the method provides an estimate of the precision of

the structure prediction, and simultaneously searches for loop and sidechain conformations whose geometry is compatible with the relatively well-defined conformations of the SCR's. In order to evalulate the potential of this approach to homology modeling, structures were built for each of the four members of the Kazal family of trypsin inhibitors whose structures have been determined by crystallography and/or NMR, using structural information derived from one or more of the other experimentally determined structures. This structural information was expressed in the form of distance and chirality constraints among the alpha and gamma-carbons of the amino acids, which in turn were derived from alignments of the sequence of the computed structure with the sequences of the other structures. The conclusions of this study may be summarized as follows:

- When the sequence homology is sufficiently high, the procedure used in this study reliably delivered a precise and accurate backbone structure. The positions of the sidechain atoms were less precisely determined, although the gamma-carbons of conserved residues usually had the correct orientation.
- In cases of low homology (less than ca. 50%), the precision of the prediction declined substantially. In addition, the structures were biased away from the correct structure particularly when the constraints were derived from only one of the other structures, or when the other structures were substantially more similar to each other than to the computed structure.

Although this pilot study demonstrated the promise of the "distance geometry" approach to homology modeling, further work is needed to determine the precision to use for the constraints as a function of the degree of structural similarity of the known structures and their sequence similarity with the unknown. In addition, the sidechain constraints used were clearly inadequate, and the distance geometry program (DISGEO) that was used does not sample conformation space as well as more recent programs [21]. In order to gain further experience with such predictions and thereby refine the methodology, in the present paper we have used the distance geometry approach to predict the structure of the Flavodoxin from *Escherichia coli* from the crystal structures of four of its homologues, the Flavodoxins from *Anacystis nidulans, Clostridium beijerinckii, Chondrus crispus* and *Desulfovibrio vulgaris*. We believe that the resulting structures will be of sufficiently high quality to serve as valuable aids in solving the phase problem for the crystallographic data that should soon be available for the *E. coli* form, now that it has been cloned [22]. By this means we also hope to demonstrate that, although further improvements in the protocol are no doubt possible, the predictions obtained from the distance geometry approach are already good enough to be of substantial practical value.

METHODS AND PROCEDURES

The distance geometry approach to homology modeling consists of the following steps:

1. The SCRs are identified and the sequence of the unknown is aligned with the sequences of the known structures, much as in the usual procedure. Note, however, that in this paper the term "SCR" refers to any region of our unknown

protein that is believed to be similar to at least one of the known homologues, rather than to all of them as in its usual usage.

2. Distance (and possibly other) geometric constraints among the atoms of the unknown structure are derived by looking at the ranges in value of the corresponding distances in the known structures, and expanding these ranges to account for the fact that even correctly aligned residues may have somewhat different conformations.
3. An *ensemble* of conformations for the unknown protein is computed, where each member of the ensemble is compatible with the distance constraints but otherwise "random". This ensemble is then analyzed to determine which geometric features are uniquely determined by the constraints, and which remain underdetermined.
4. Each structure in the ensemble isenergy minimized in order to eliminate structural irregularities (such as eclipsed rotatable bonds) that are difficult to exclude by means of simple distance constraints.

Structure and Sequence Alignments

In the present case, the SCR's were identified by taking the published structural

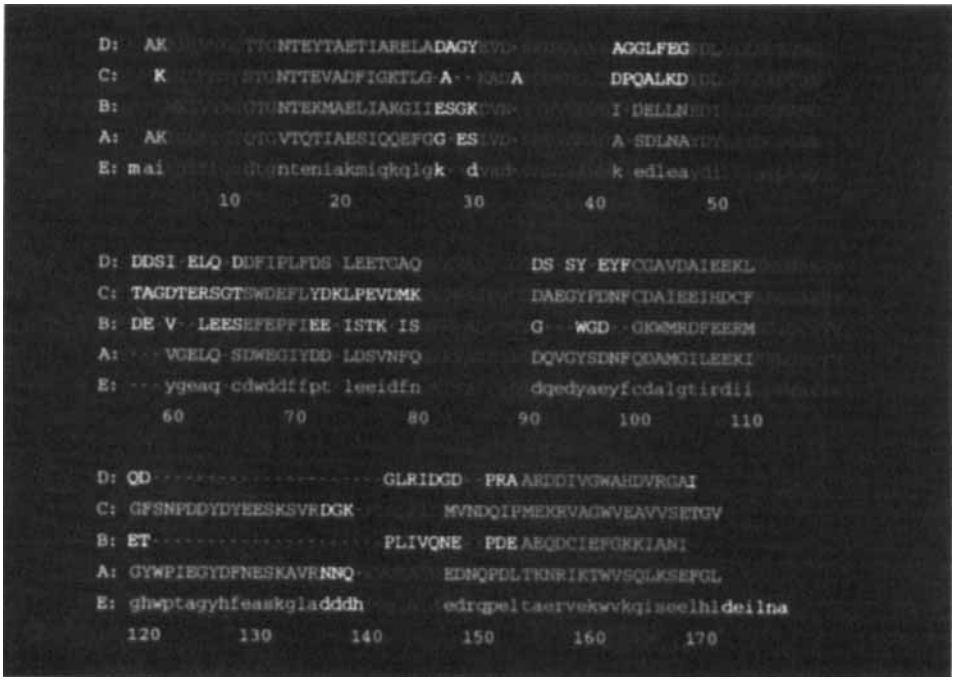


Figure 1 The Flavodoxin sequence alignment used throughout this study. The yellow letters at the beginning of each line indicate the organism for each sequence: **A** for *A. nidulans*, **B** for *C. beijerinckii*, **C** for *C. crispus*, **D** for *D. vulgaris*, and **E** for *E. coli*. The remaining letters on each line are the one-letter amino acid codes. The colors indicate the secondary structure element of each SCR identified: **red** for beta strands, **green** for alpha helices, **violet** for turns and **cyan** for random coil. **White** is used for those regions of the homologues that are not structurally homologous to the *E. coli* sequence, and for *E. coli* regions that are not homologous to any of the homologues. The sequence numbers shown refer to the *E. coli* sequence. (See Colour Plates)

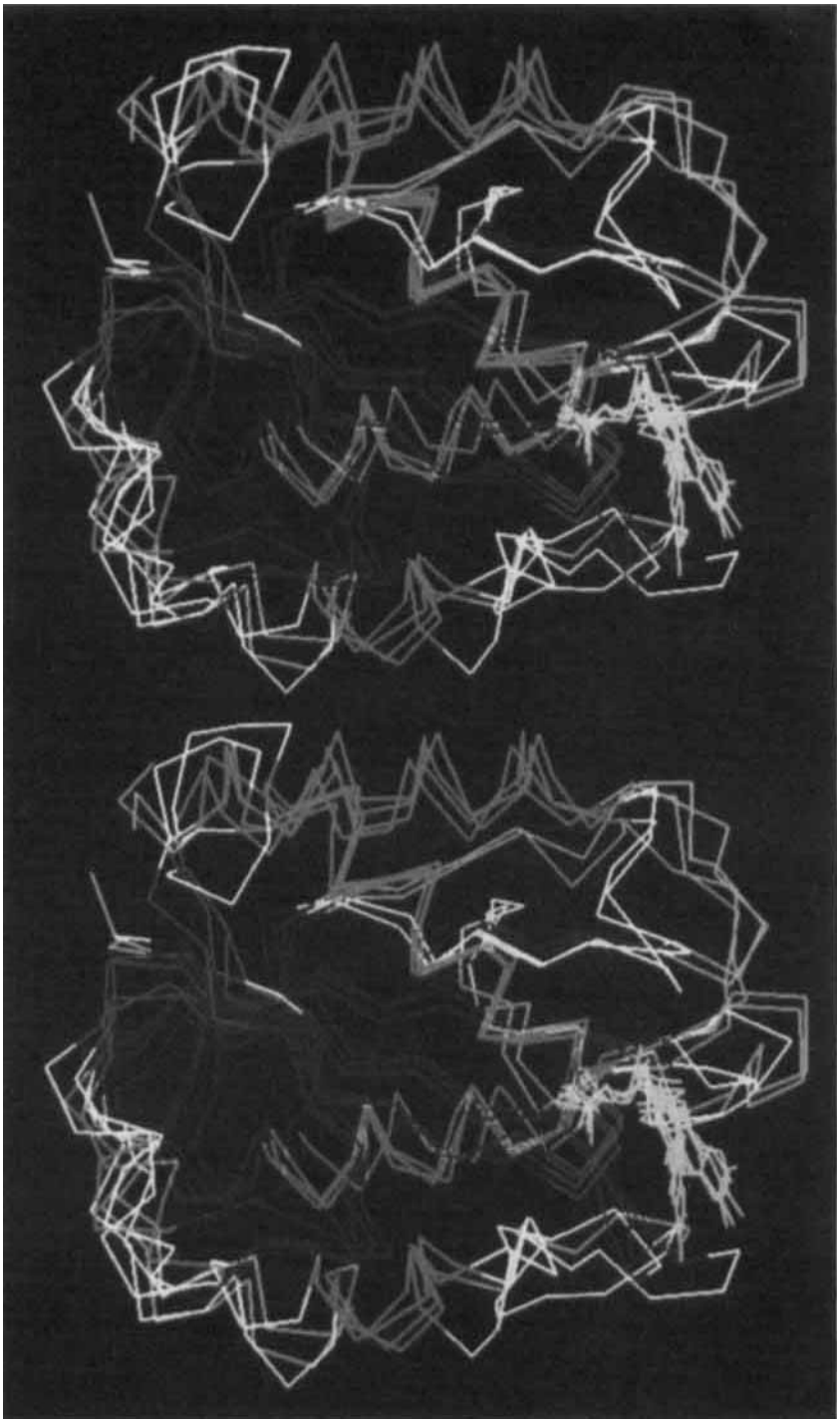


Figure 2 Stereoview of the alpha-carbon trace of the four homologues, *A. nidulans*, *C. beijerinckii*, *C. crispus* and *D. vulgaris*, where the alpha-carbons in the SCR's of the last three have been superimposed on the alpha-carbons of any corresponding SCR's in the *A. nidulans* structure so as to minimize the RMSD. The colors have the same meaning as in Figure 1, except that the flavin mononucleotide cofactors are all shown in **yellow**. This and all other computer graphics images in this paper were produced using the *Insight* program available from BIOSYM Technologies, Inc. (See Colour Plates)

alignments of the *A. nidulans*, *C. beijerinckii* and *D. vulgaris* sequences [23], and aligning the *C. crispus* structure to them. On the basis of the information present in the *C. crispus* structure, it was deemed advisable to adjust the alignment of the other three structures in a few places, so as to obtain overall consistency. These structural alignments were made using the *Homology* modeling program that is available from BIOSYM Technologies, Inc. [19]. The resulting four-way structural alignment is shown in Figures 1 and 2, with the SCR's indicated by colors; the crystallographic structures on which this alignment is based were first published in [24, 25, 26, 27]. The reader may also wish to refer to Figure 3, which shows a schematic drawing of the *A. nidulans* Flavodoxin with various relevant residue numbers added. Since the numbering of the *E. coli* Flavodoxin differs from the *A. nidulans* numbering by at most one at all positions, this same diagram can also be used as a guide to the figures showing the *E. coli* structures.

The published alignment of the *E. coli* sequence with the *A. nidulans* sequence (with which it is distinctly most similar) was then used to obtain an initial alignment with the other Flavodoxin sequences, and the alignment of the *E coli* sequence further adjusted to take account of the information present in the other sequences (see Figure 1). In particular, the site of the one-residue deletion in the *E. coli* sequence relative to *A. nidulans* was moved from between residues 31 and 32 to between residues 28 and 29, since it was clear from the structural alignments that this is where the indels (i.e. insertions/deletions) had been concentrated in the other structures. The following features of this alignment are worth noting:

1. The *N*-terminal methionine and *C*-terminal six residues of the *E. coli* sequence have no corresponding residues in any of the known homologues, so that nothing can be said about their conformations on the basis of the alignment.
2. The only other regions left completely unconstrained were a segment of two residues at 28–29 where the deletion occurred, and the residues 135–137 which flank the inserted histidine at position 138. Otherwise, there are no indels in the alignment involving the *E. coli* sequence.
3. In several regions, e.g. residues 59–64, indels occurred between the *E. coli* sequence and some but not all of the other sequences. In such cases the regions of those homologues with indels in them were not considered part of the SCR, so that those homologues had no corresponding residues in that region of the *E. coli* sequence.
4. In this alignment, the percent identities of the *E. coli* sequence with the *A. nidulans*, *C. beijerinckii*, *C. crispus* and *D. vulgaris* sequences are 44%, 16%, 33% and 23%, respectively. The sequence identities among the known structures, on the other hand, ranged from 17% to 33% (Table 1), which is comparable and hence implies that the variations among the known structures can be used as a reasonable estimate of magnitude of the deviation of the *E. coli* structure from them.

Geometric Constraints from the Alignment

The basic assumption on which the distance geometry approach to homology modeling is based is that the distances among pairs of atoms in the unknown structure should be similar to the distances between the corresponding pairs in the known structures. Moreover, it is reasonable to use the variations in the distances among the known structures to estimate the uncertainty of the corresponding

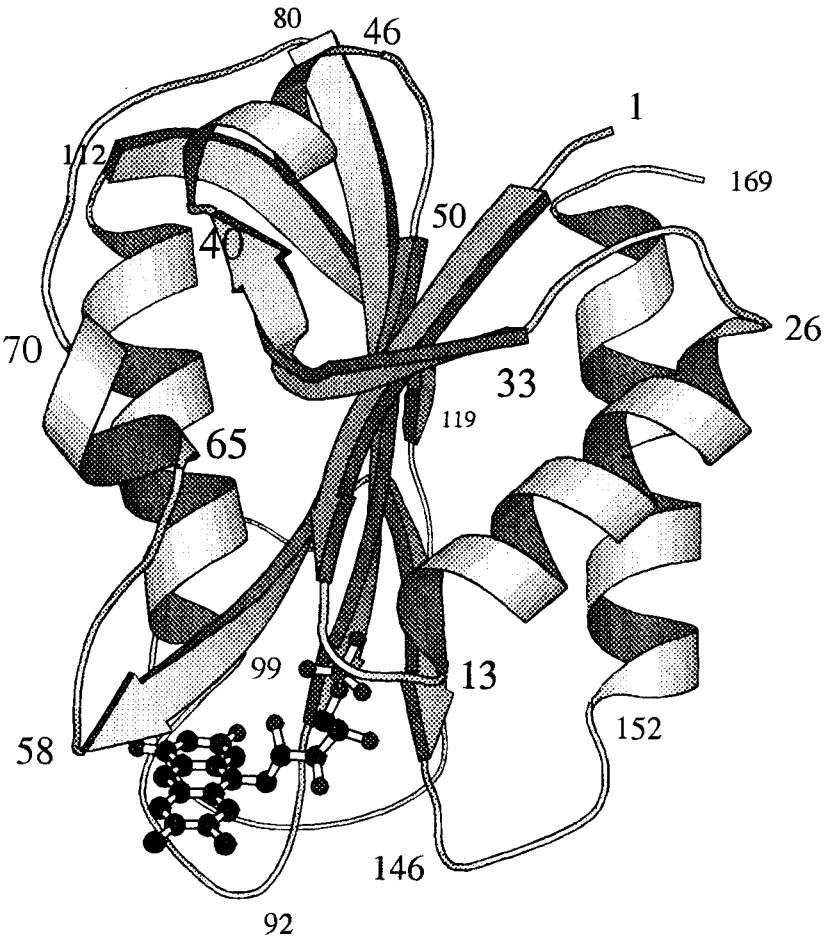


Figure 3 Schematic "ribbon" diagram of the Flavodoxin from *A. nidulans*, with the sequence numbers of the residues at the beginning and end of each element of secondary structure shown. This picture was produced using the *MolScript* program of P. Kraulis.

distances in the unknown. Here, the term "corresponding" means that the pair of residues in which the pair of atoms involved in the distance occur correspond in the alignment. For distances between atoms that are found in all amino acids, i.e. backbone atoms, the correspondence of the residues in the alignment is wholly sufficient to determine the correspondence of the atoms. Similarly, if both residues of the pair of atoms involved in the distance have the same type as in the homologues, then the correspondence between the atoms is obvious. If the distance is between a pair of sidechain atoms that are contained in residues that have mutated to residues of different types of some or all of the homologues, however, the corresponding distance in the unknown protein is not unambiguously defined.

It should be noted, however, that (unlike the usual cut-and-paste approach to model building), these ambiguities in our definition of the correspondence among the atoms will not necessarily lead to inaccuracies in the computed structures;

Table 1 Sequence identities (below diagonal), number of residues (diagonal) and alpha-carbon RMSD's between the common SCR's (above diagonal) of the four homologues.

	A. nidulans	*C. beijerinckii*	*C. crispus*	*D. vulgaris*
A. nidulans	169	1.90 Å	1.40 Å	1.29 Å
C. beijerinckii	20%	138	1.92 Å	1.67 Å
C. crispus	33%	17%	173	1.25 Å
D. vulgaris	27%	25%	21%	147

instead, what they will usually lead to is less precise constraints (because the distances between pairs of atoms that do not play identical structural roles will tend to vary more among the known structures). Similarly, errors in the alignment of the sequences of the known structures will tend to reduce the precision of the constraints rather than their accuracy. This in turn will reduce the precision of those features of the unknown structure that depend upon by these constraints, meaning that the members of the structural ensemble obtained from the distance geometry calculations will show larger variations in those features. Thus the statistical analysis of the ensemble will at least not be led to erroneous conclusions; rather it will conclude (correctly) that it is not possible to say anything about these features on the basis of the alignment. This automatic compensation for errors in the sequence alignment and ambiguities in the definition of the correspondence among the atoms is in fact one of the most significant strengths of the distance geometry approach to homology modeling.

For each pair of atoms i and j whose distance was constrained in the *E. coli* structure, the following procedure was used to determine the lower and upper bounds on that distance to use as input for the distance geometry calculations. First, if i_k and j_k denotes the corresponding pair of atoms in the k-th known structure (*A. nidulans*, *C. beijerinckii*, *C. crispus* or *D. vulgaris*) and $d(i_k, j_k)$ is the distance between them, we define the lower and upper limits on the corresponding distance in the *E. coli* structure as:

$$\bar{l}(i, j) := \min_k (d(i_k, j_k))$$

$$\bar{u}(i, j) := \max_k (d(i_k, j_k))$$

Then, the distance bounds used for the calculations were obtained from these limits by the following formulae:

$$l(i, j) := (\bar{u}(i, j) + \bar{l}(i, j))/2 - \rho(\bar{u}(i, j) - \bar{l}(i, j))/2 - \delta/n_{ij}$$

$$u(i, j) := (\bar{u}(i, j) + \bar{l}(i, j))/2 + \rho(\bar{u}(i, j) - \bar{l}(i, j))/2 + \delta/n_{ij}$$

where the parameter $\rho \geq 0$ is called the *precision* of the constraint and the parameter $\delta \geq 0$ is the *tolerance*. The variable n_{ij} is the number of homologues in which the *E. coli* atoms i and j have a corresponding atom, so that the corresponding distance exists in that homologue and hence was used in computing the limits $\bar{l}$, $\bar{u}$. Thus, when all four homologues were used in computing the limits, the range $(\bar{u} - \bar{l})/2$ times the precision was taken as a good estimate of the uncertainty and only one fourth the full tolerance was used, whereas if the corresponding distance existed in only one of the homologues, the uncertainty about that single distance $\bar{l} = \bar{u}$ was exactly $\pm\delta$.

A Brief Description of the Constraints

In order to determine the overall chain fold and secondary structure without having to deal with the ambiguities that arise with nonconserved residues, we extracted all the information we could concerning the distances among the alpha-carbons. In this case the precision ρ was set to 2.0, while the tolerance δ was set to 1 Å if both alpha-carbons were contained in the same SCR (short-range alpha-carbon constraints) and 2 Å if they were in different SCR's (long-range constraints). All told, this gave 575 short-range constraints and 12 468 long-range ones. In addition, 978 constraints were derived on the distances between the alpha-carbons and the *N1, C4, C6, C9, C2', C4'* and *P* atoms of the flavin mononucleotide cofactor (see Figure 4), using a precision $\rho = 2.0$ and a tolerance $\delta = 2.0$. Collectively, these will be referred to as *global* distance constraints; the root mean square gap between their lower and upper bounds was 4.20 Å.

In order to determine the finer details of the sidechain and backbone

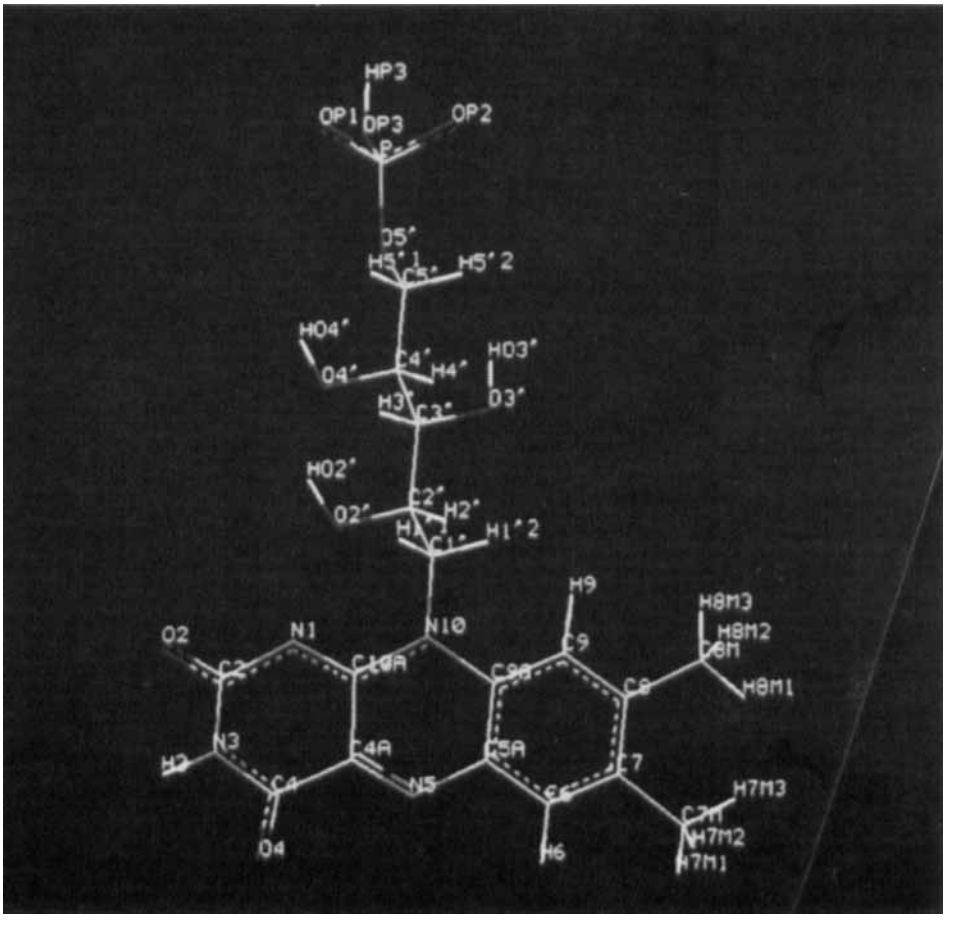


Figure 4 Drawing of the oxidized form of the flavin mononucleotide cofactor in the tautomer used for all the calculations in this paper, and with the atom labels used draw in in **yellow**. (See Colour Plates)

conformations, 8893 *local* constraints were also imposed on the distances between the nonhydrogen atoms within each residue and between the backbone and sidechain nonhydrogen atoms of adjacent residues, providing these were separated by more than three rotatable bonds. These were derived using a precision $\rho = 1.5$, a tolerance $\delta = 0.5$ Å, and the following table of equivalences between the side chain positions of nonidentical residues (Table 2). In this table, the rows and columns are indexed by the one-letter amino acid codes, and the entries in the upper half of the table contain the greek letter of the position in the sidechain out to which the atoms of each pair of residues were considered equivalent. A little more precisely, if at a given position in the *E. coli* sequence the corresponding residue in at least one of the homologues of known structure had the same type, then the bounds were derived from only those structures wherein the residue had the same type. Otherwise, the intersection over all sets of equivalent atoms in the *E. coli* sidechain and the corresponding sidechains in the homologues was determined, and bounds were generated only among the atoms in this intersection and the backbone atoms of adjacent residues. The root mean square gap between these bounds was 1.13 Å.

Additional constraints were imposed upon the torsion angles, using the same recipe as used for the distance constraints to derive lower and upper bounds on the angles with a precision $\rho = 1.5$ and a tolerance $\delta = 30°$. The angles constrained included all ϕ and ψ angles in residues that were aligned with any of the known homologues, together with all sidechain angles that were equivalent, with substitutions handled as shown in the lower half of Table 2. In all, this gave 452 torsion angle constraints with a root mean square gap of 77.5°. Finally, each of the homologues was energy minimized using the *Discover* molecular mechanics program (available from BIOSYM Technologies, Inc.), and a list of hydrogen bonds in the resulting structures printed. These lists were then scanned manually

Table 2 Sidechain atom (above and on diagonal) and angle (below diagonal) correspondences between residue types.

	A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	β	β	β	β	β	β	β	α	β	β	β	β	β	β	β	β	β	β	β	β
R	—	ζ	β	γ	β	β	δ	α	γ	β	β	δ	β	β	β	β	β	β	β	β
D	—	—	δ	γ	β	β	β	α	β	β	β	β	β	β	β	β	β	β	β	β
N	—	χ^1	χ^1	δ	β	β	β	α	γ	β	β	β	β	β	β	β	β	β	β	β
C	—	—	—	—	β	β	β	α	β	β	β	β	β	β	β	β	β	β	β	β
E	—	—	—	—	—	ϵ	δ	α	β	β	β	β	β	β	β	β	β	β	β	β
Q	—	χ^2	—	—	—	χ^2	β	α	γ	β	β	β	β	β	β	β	β	β	β	β
G	—	—	—	—	—	—	—	α	α	α	α	α	α	α	α	α	α	α	α	α
H	—	χ^1	—	χ^1	—	—	χ^1	—	ϵ	β	β	β	β	β	β	β	β	γ	β	β
I	—	—	—	—	—	—	—	—	—	δ	β	β	β	β	β	β	β	β	β	γ
L	—	—	—	—	—	—	—	—	—	—	δ	β	γ	β	β	β	β	β	β	β
K	—	χ^2	—	—	—	—	—	—	—	—	—	ζ	β	β	β	β	β	β	β	β
M	—	—	—	—	—	—	—	—	—	—	χ^1	—	ϵ	β	β	β	β	β	β	β
F	—	—	—	—	—	—	—	—	—	—	—	—	—	ζ	β	β	β	γ	δ	β
P	—	—	—	—	—	—	—	—	—	—	—	—	—	—	δ	β	β	β	β	β
S	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	γ	β	β	β	β
T	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	γ	β	β	β
W	—	—	—	—	—	—	—	—	χ^1	—	—	—	—	χ^1	—	—	—	η	γ	β
Y	—	—	—	—	—	—	—	—	—	—	—	—	—	χ^2	—	—	—	χ^1	ζ	β
V	—	—	—		—	—	—	—	—	χ^1	—	—	—	—	—	—	—	—	—	γ

Table 3 Hydrogen bond constraints (donor acceptor).

5–G:N	49–I:O	6–I:N	32–D:O	7–F:N	51–L:O
8–F:N	34–H:O	9–G:N	53–G:O	11–D:N	FMN:OP2
12–T:N	FMN:OP2	12–T:OG1	FMN:OP1	13–G:N	FMN:OP1
14–N:N	FMN:OP1	14–N:ND2	FMN:OP1	15–T:N	FMN:OP3
16–E:N	10–S:OG	18–I:N	14–N:O	19–A:N	15–T:O
20–K:N	16–E:O	21–M:N	17–N:O	11–I:N	18–I:O
23–Q:N	19–A:O	23–Q:NE2	31–A:O	24–K:N	20–K:O
25–Q:N	21–M:O	26–L:N	22–I:O	27–G:N	23–Q:O
32–D:N	4–T:O	34–H:N	6–I:O	34–H:ND1	32–D:OD1
37–A:N	35–D:OD1	38–K:N	35–D:O	40–S:N	43–D:OD1
47–Y:OH	32–D:OD2	50–L:N	82–L:O	51–L:N	5–G:O
52–L:N	84–A:O	53–G:N	7–F:O	54–I:N	86–F:O
56–T:N	88–C:O	57–W:NE1	FMN:OP2	59–Y:N	FMN:N5
60–G:N	FMN:O4	63–Q:NE2	65–D:OD1	65–D:N	63–Q:OE1
66–W:NE1	53–G:O	66–W:N	63–Q:OE1	70–F:N	66–W:O
70–F:N	67–D:O	72–T:OG1	68–D:O	77–D:N	45–E:OE1
80–G:N	113–G:O	81–K:NZ	45–E:O	81–K:NZ	47–Y:O
83–V:N	115–T:O	84–A:N	50–L:O	86–F:N	52–L:O
87–G:N	142–L:O	89–G:N	144–I:O	90–D:N	FMN:O2
91–Q:N	145–D:OD1	91–Q:NE2	145–D:OD1	91–Q:NE2	149–Q:OE1
92–E:N	90–D:OD1	93–D:N	90–D:OD1	94–Y:N	90–D:O
98–F:N	129–A:O	99–C:N	FMN:O2	101–A:N	56–T:OG1
102–L:N	99–C:O	104–T:N	100–D:O	105–I:N	101–A:O
106–R:N	102–L:O	107–D:N	103–G:O	108–I:N	104–T:O
109–I:N	105–I:O	110–E:N	106–R:O	112–R:N	109–I:O
113–G:N	110–E:O	115–T:N	81–K:O	117–V:N	83–V:O
120–W:NE1	125–Y:OH	125–Y:N	122–T:O	125–Y:OH	143–A:O
126–H:N	91–Q:OE1	128–E:N	95–A:O	129–A:N	95–A:O
130–S:OG	98–F:O	132–G:N	130–S:OG	140–V:N	132–G:O
142–L:N	85–L:O	144–I:N	87–G:O	147–D:N	145–D:OD1
148–R:NE	90–D:OD1	148–R:N	145–D:OD1	153–T:OG1	146–E:OE1
156–R:NH1	149–Q:OE1	156–R:NH2	149–Q:OE1	156–R:N	153–T:O
157–V:N	153–T:O	158–E:N	154–A:O	159–K:N	155–E:O
160–W:NE1	118–G:O	160–W:N	156–R:O	161–V:N	157–V:O
162–K:N	158–E:O	163–Q:N	159–K:O	164–I:N	160–W:O
164–I:N	161–V:O	165–S:N	161–V:O	166–E:N	162–K:O
166–E:N	166–E:OE1	FMN:OP3	10–S:OG	FMN:OP3	15–T:OG1
FMN:O2′	56–T:O	FMN:N3	97–Y:O	FMN:O3′	147–D:OD1
FMN:O3′	FMN:O4′	FMN:O4′	FMN:O5′		

and the hydrogen bonds that occurred in all of the homologues (except those in which a substitution had occurred that eliminated the donor or acceptor atoms) were identified. Providing that analogous donor and acceptor atoms were also present at the aligned positions in the *E. coli* sequence, a distance range of 2.5 Å to 3.3 Å was imposed on the donor/acceptor distance. (The proton/acceptor distance could not be constrained, since no protons were included in the distance geometry calculations.) This gave the 116 hydrogen bond constraints shown in Table 3.

Distance Geometry Calculations

Using the *NMRchitect* package available from BIOSYM Technologies, Inc., the distance and torsion angle constraints derived from the sequence alignment with the homologous structures were combined with the bond lengths, bond angles, hard sphere radii etc. that follow from the chemical identities of the atoms and the arrangement of covalent bonds in *E. coli* Flavodoxin. The entire set of constraints was then used as input to the *DG-II* distance geometry program [21],

and an ensemble of ten structures that satisfied the constraints but were otherwise "random" computed. In order to reduce the computer time required, no hydrogen atoms were included in these calculations, leaving a total of 1423 atoms.

The overall computational protocol followed was similar to that used in [21]:

1. First, the initial bounds matrix was subjected to triangle inequality bound smoothing followed by sequential tetrangle bound smoothing, to obtain a set of distance bounds with a root mean square gap between the bounds of 15.13 Å. This is only about half the root mean square gap that is typically obtained with calculations using NMR data, which in turn will tend to make the present calculations more reliable and efficient. The bound smoothing calculations themselves required approximately 13 hours of CPU time on a Sun SparcStation II desktop workstation.
2. Next, ten distance matrices were chosen using a process known as *metrization*, which yields highly random distance matrices whose elements obey both the distance bounds as well as the triangle inequality [28, 29]. Each full metrization required approximately one and one half hours on the SparcStation II computer.
3. These distance matrices were then converted into metric matrices, i.e. matrices of dot products among the vectors from the center of mass to each of the atoms, and these metric matrices converted into coordinates for the atoms as described in [30, 31]. Each such "embedding" of the distance matrices required less than five minutes of CPU time on the SparcStation II.
4. The resulting coordinates were refined so as to give a weighted least-squares fit to the original distance matrix using a procedure known as *majorization* [32], where the weights were chosen according to the range-plus-average scheme described in [21]. In contrast to the procedure followed in this latter reference, however, the least-squares equations were solved by a linear conjugate gradients algorithm rather than by computing a generalized inverse, and a larger step size was used at each iteration. This new procedure is substantially more efficient, and required only about 20 minutes for the twenty-five iterations used for each structure on the SparcStation II.
5. Finally, the torsion angle constraints and chirality constraints implicit in the stereochemistry of the amino acids were converted into oriented volume constraints [21], and the coordinates refined via the usual simulated annealing procedure used by the *DG-II* package to full consistency with the distance and volume constraints. Each of these calculations required between two and a half and five hours on the SparcStation II, depending on if the correct mirror image was chosen at the onset or not.

Energy Minimization

Although the structures obtained from distance geometry calculations inevitably have the correct stereochemistry and are free of any serious steric problems, they often exhibit eclipsed comformations at rotatable bonds and may have unfavorable electrostatic interactions, e.g. hydrogen bond geometries, that could easily be eliminated by small changes in their conformations. This latter problem is particularly noticeable in the calculations reported here because the distance geometry calculations were done without any protons present for efficiency's sake.

Thus, although the hydrogen bond distance constraints enforced the spatial proximity of the donor and acceptor atoms, they could not guarantee an approximately optimum angle of 180° about the intervening hydrogen atom. For these reasons the ten structures obtained from the distance geometry calculations were energy minimized.

In our experience, however, ordinary energy minimization by standard descent techniques such as conjugate gradients usually fails to change the conformations significantly. Thus we have used a simulated annealing procedure in which the distance geometry structures are subjected to molecular dynamics simulation at gradually decreasing temperatures starting from 300 *K*. Because molecular dynamics tends to unfold proteins unless done with great care, the full set of constraints used in the distance geometry calculations were enforced using a harmonic restraint function ("pseudo-potential"). Finally, the structures were polished off with an unrestrained conjugate gradients energy minimization. Since the distance geometry structures already constituted a diverse sample of the "conformation space" compatible with the constraints, the goal of this mild energy refinement procedure was not to drastically alter their conformations in a search for new ones, but rather to attain a good energy minimum that was largely compatible with the geometric constraints derived by homology.

These calculations were done using BIOSYM's *Discover* molecular mechanics program, together with the associated *CVFF* energy function. Parameters were obtained for the flavin mononucleotide cofactor using the automatic parametrization procedure in BIOSYM's *Insight* program, assuming the tautomer for the oxidized form shown in Figure 4. To make these calculations as realistic as computationally possible, water was accounted for explicitly using a solvent box along with periodic boundary conditions. The dielectric constant was set to 1.0 and a sigmoidal scaling of the nonbonded interactions was used starting at 6 Å and going to zero at the cutoff of 8 Å. The refinement procedure itself consisted of the following steps:

1. Protons were added to each structure so that all its nonhydrogen atoms had their correct valences and charges for a pH of 7.0.
2. The resulting molecule was embedded in a $50 \times 50 \times 40$ Å solvent box containing ca. 2500 water molecules.
3. The homology restraints were added to the potential function using a harmonic restraint function with a force constant of 10.0 kcal/Å^2 for the hydrogen bond restraints and 1.0 for all other distance restraints. The torsion constraints were enforced using a harmonic potential with a force constant of 30 kcal/rad^2.
4. The system was energy minimized briefly to eliminate atomic overlaps, first using 100 steps of steepest descents with all nonhydrogen, nonsolvent atoms fixed, and then another 100 steps with no fixed atoms.
5. The system was heated rapidly to 300 K, and cooled in gradual steps to 10 K over 20 psec. of molecular dynamics simulation using a leap-frog algorithm with a stepsize of 1.0 fsec.
6. The restraints were removed and 1000 iterations of conjugate gradients energy minimization applied.

Each of these refinements consumed approximately three and one half days of CPU time on a single processor of an Silicon Graphics 480 computer, or roughly a factor of ten more than the distance geometry calculations.

Structural Analysis

The statistical analyses used to study the structural ensembles computed from the homology constraints are similar to those commonly used for NMR constraints [5]. The two most important geometric properties analyzed are:

1. The *root mean square deviation* (RMSD) in the coordinates [33]. This difference measure may be computed for segments of the polypeptide chain or other subsets of the atoms (e.g. alpha-carbons), in order to measure the difference in their relative positions [5]. The mean value of the RMSD over all pairs of structures in the ensemble then provides a convenient measure of the variability of those coordinates over the ensemble.
2. The torsion angles about single bonds. The variations in the ϕ and ψ angles of proteins may be visualized via a Ramachandran map [34]. More recently, a torsion angle "order parameter" has been introduced that measures how well an angle is defined and at the same time defines an average value for the angle [35] (see below).

The use of the RMSD for comparing structures and parts of structures is well-established, and does not need to be explained here (see e.g. [21]). It should nonetheless be pointed out that the optimal superposition of the coordinates obtained by computing the RMSD between them allows one to easily visualize the detailed differences via molecular computer graphics, which is perhaps more important than the value of the RMSD itself. Many hours were spent studying the ensembles computed in this paper by these techniques, in order to locate differences between them and their homologues that could then be quantitated by other means.

The order parameter of a torsion angle in an ensemble of conformations is obtained by placing a unit vector in the plane $\mathbf{R}^2$ for each conformation of the ensemble, where the angle of the unit vector with the x-axis is equal to the value of the torsion angle in that conformation. The unit vectors are then summed and divided by their number, to obtain a vector whose length is the order parameter and whose angle with the x-axis is the average angle by definition. In equations, if θ_{i_k} is the value of the i-th angle in the k-th member of an ensemble of N conformations, then the order parameter of θ_i is

$$\xi_i := \frac{\sqrt{(\Sigma_k \cos(\theta_{i_k}))^2 + (\Sigma_k \sin(\theta_{i_k}))^2}}{N}$$

and its average is

$$\bar{\theta}_i := \arctan\left(\frac{\Sigma_k \sin(\theta_{i_k})}{\Sigma_k \cos(\theta_{i_k})}\right),$$

where the sign of $\bar{\theta}_i$ is the same as that of $\Sigma_k \sin(\theta_{i_k})$. Note that ξ_i is always between zero and one, and is one only if the angle has precisely the same value in all members of the ensemble. In practice, we have found that the order parameter is usually close to unity, i.e. it is not very sensitive to differences in the precision of moderately well-defined angles. Thus we have used instead the $\log_{10}$ of the

transformed order parameter $\hat{\xi}_i$:

$$\log_{10}(\hat{\xi}_i) := -\log_{10}(1-\xi_i)$$

as a measure of the precision of the angle.

RESULTS AND DISCUSSION

The Distance Geometry Ensemble

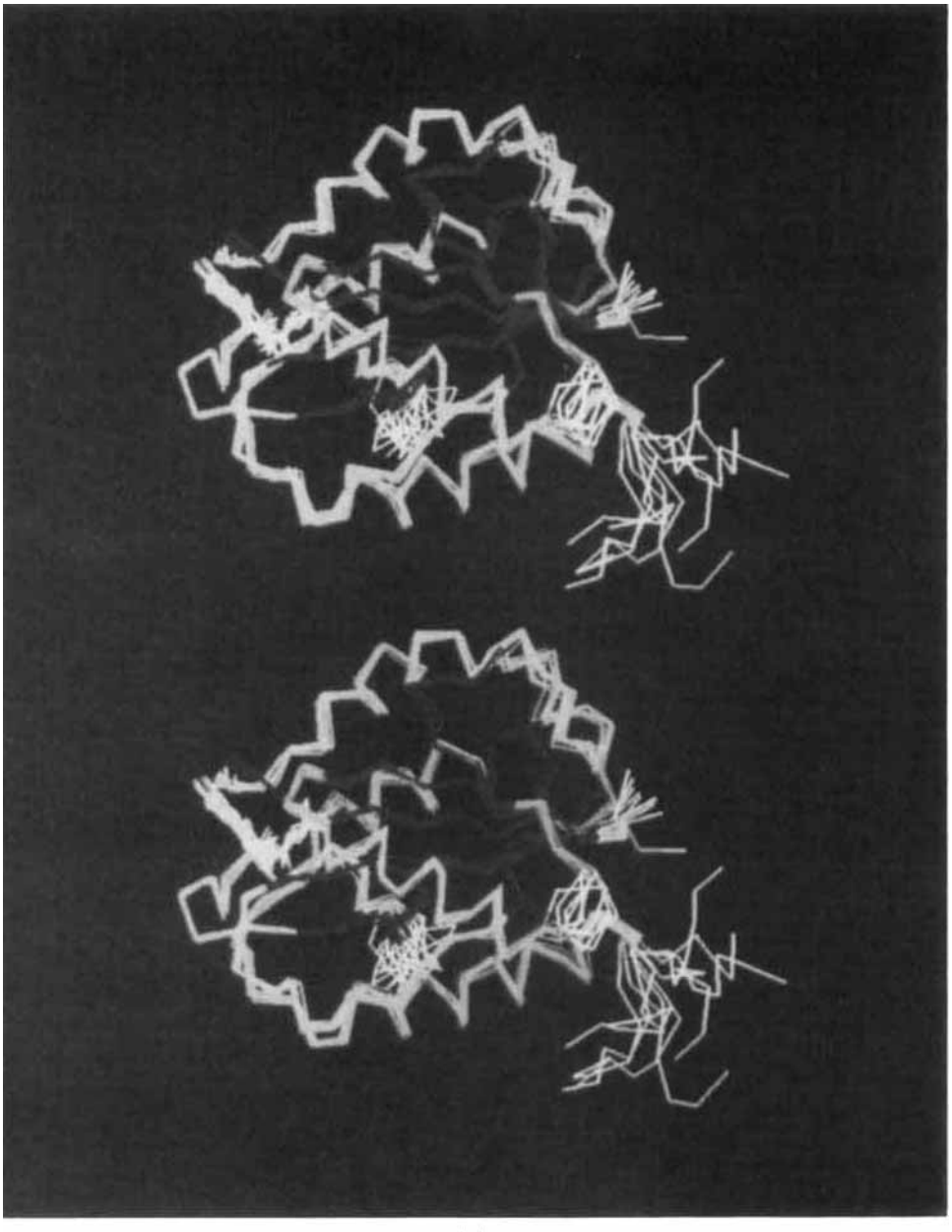
Each of the ten distance geometry calculations converged smoothly to structures with maximum distance constraint violations ranging from 0.27 to 0.59 Å. The number of distance constraint violations exceeding 0.1 Å ranged from 8 to 38 with an average of 14.7; the number of distance constraint violations exceeding 1% of the violated bound ranged from 19 to 73 with an average of 28.0. These figures indicate that the constraints as a whole were largely consistent even though they were pieced together from four different structures, which in turn means our choice of precisions and tolerances was not drastically too "tight".

One structure in particular exhibited numerous small violations involving the sidechain of the phenylalanine at position 86, even though its maximum distance constraint violation was only 0.27 Å. This sidechain was found to be separated from its position in the remaining structures by glycine-53 in the center strand of the beta-sheet, which it evidently failed to cross during the optimizations. This new position for the sidechain was accommodated by twisting the stand in which it itself occurs rather than by a change in its local conformation. Otherwise, no local minima were encountered that were obviously different from the usual conformations.

The number of torsion angle constraint violations exceeding 5.0° ranged from 8 to 11. A few of these violations were quite large, however, exceeding 120° in some cases. These turned out to be an artifact of the method by which torsion angle constraints are usually enforced in the *DG-II* package, coupled with the reduced structure representation that was used. The torsion angle constraints are usually translated into oriented volume constraints, which individually permit ambiguities in the angle range [36]. As a rule, these ambiguities can be eliminated by defining the range in all possible ways, i.e. with respect to all pairs of 1,4 atoms, but without protons the number of such oriented volume constraints per angle was too small to guarantee a unique solution in all cases.

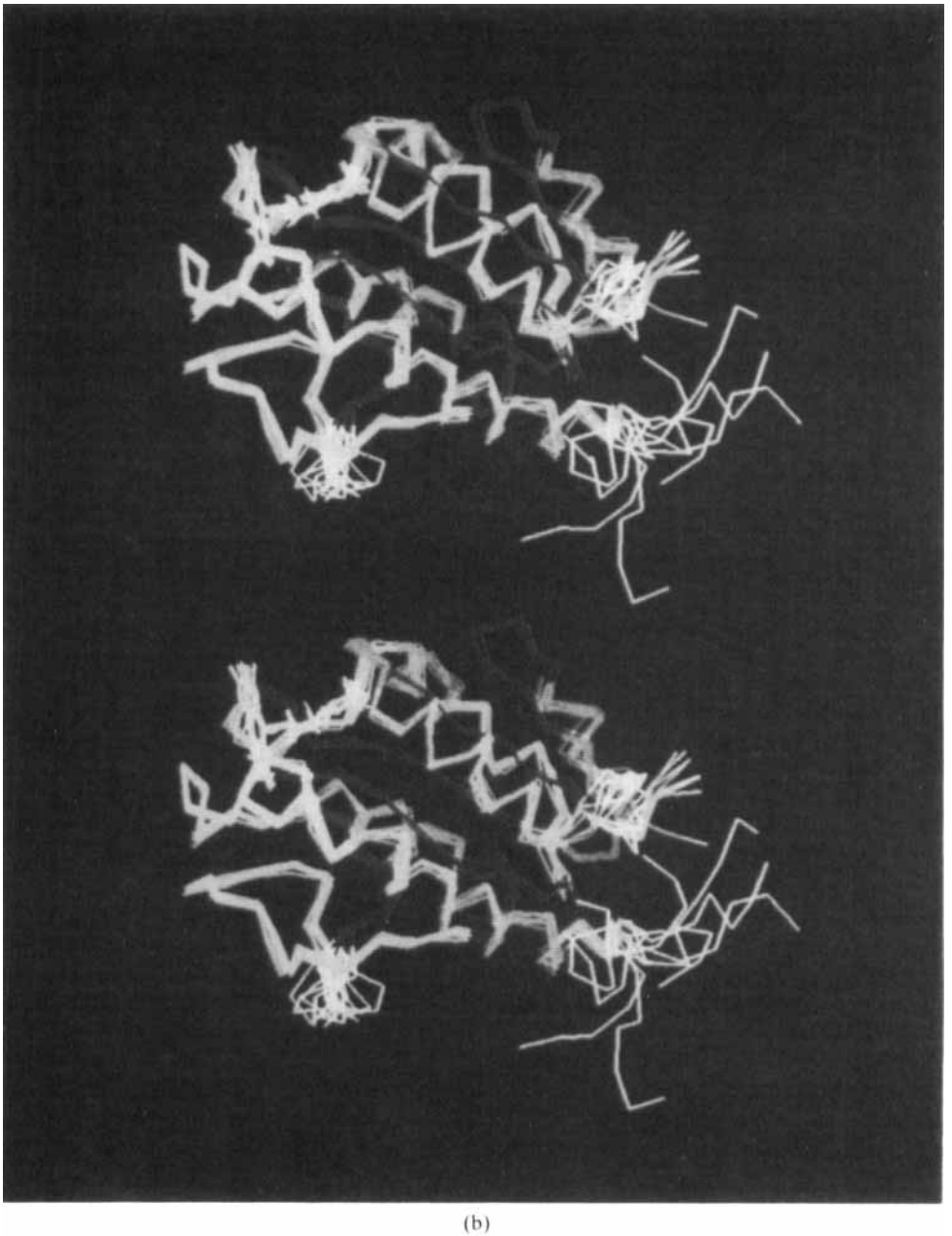
The problem could probably be eliminated by also using 1,4 distance constraints to restrict the range of torsion angle. Fortunately, the number of such torsion angle violations was small, ranging from 2 to 5 over the ten structures and involving only 10 torsion angles in all. Of these, one was the ϕ angle of glycine-27, four were χ^1 angles and the rest were higher χ angles; none were violated in more than six of the ten structures, and three were violated in only one of the structures. Thus, the precision of the ensemble would probably not have been improved significantly had these angles been properly constrained, and in any case these ranges were properly enforced during the dynamical energy minimization performed subsequently.

The RMSD among the alpha-carbons, excluding the unconstrained residues 1 and 171–176, averaged only 0.85 Å over all pairs of structures in the ensemble (maximum 1.07 Å). The average heavy atom RMSD among these same residues averaged 1.60 Å (maximum 1.88 Å). These numbers are comparable to those



(a)

Figure 5 The alpha-carbon trace plus flavin mononucleotide cofactor of the ten structures obtained by distance geometry calculations, all aligned so as to minimize the RMSD with the alpha-carbons of residues 2–170 of the first structures in the ensemble. The colors have the same meanings as in Figures 1 and 2. Figure (a) shows an edge-on view of the central beta sheet; (b) shows a front view of the sheet. (See Colour Plates)



(b)

obtained in a high-quality NMR structure determination, and indicate that the variations in the relative positions of the atoms tends to be substantially less than the range allowed by the individual constraints. This phenomenon is a consequence of the way in which the constraints were pieced together from diverse structures and applied to yet another structure, so that the range of solutions compatible with

the combined homology, stereochemical and steric restrictions is considerably smaller than the precision of the individual homology constraints might suggest.

Two different stereoviews, turned by 90°, of the alpha-carbons of the distance geometry ensemble are shown in Figure 5 (a) and (b), where all members of the ensemble have been colored by SCR. This not only confirms the overall similarity of the backbone conformations, but also illustrates how the distance geometry algorithm automatically "searches" for a diverse set of conformations in those regions of the chain that could not be aligned with any of the homologues. These regions include the C-terminal six residues as well as the loops at residues 28–29 and 136–138, where the deletion and insertion of the *E. coli* sequence with respect to all four homologues are found, respectively. Diverse sets of possible conformations for those sidechains that had been substituted in *E. coli* with respect to all four homologues were also obtained, as will presently be discussed in detail.

A Ramachandran plot in the ϕ, ψ plane of the residues in all ten structures exhibited the usual clustering of the residues in the energetically favorable alpha and beta regions (see Figure 6(a)). There was also a noticeable tendency of the residues to clump together at certain points. These points were found to be occupied by the same residues in all ten of the structures, showing that the ϕ and ψ angles of these residues were within a few degrees of each other. Careful examination of a few such clumps revealed that the corresponding angles in the homologues also tended to be within about ten degrees of each other in all four homologues and usually differed by less than 0.1 Å in their backbone RMSD. Thus, the local distance and torsion constraints in these regions tended to be relatively tight, and this combined with the exact distance constraints that define an ideal covalent geometry (as opposed to the range that occurs in the homologues) was evidently sufficient to give the very high degree of similarity obtained.

This similarity was also visible in the plots of the order parameters versus sequence (see Figure 7). Although some of the really extreme spikes may be due in part to roundoff error, the local structure as a whole shows excellent definition almost everywhere except at the ends and in the regions where the insertion and deletion occurred. One should be a little careful not to read too much into these plots: the abrupt dip in the ψ plot at residue 117 and the ϕ plot at residue 118, for example, is due primarily to a single structure wherein these angles are turned nearly 180° from the rest. The χ^1 angles with transformed order parameters greater than 10 ($\log_{10} = 1$ in the plots) correlate well with the sidechains that are identical between *E. coli* and at least one of its homologues, and therefore had constraints on all of their atoms. The exceptions we have looked at are easy to understand. For example, the sidechain of histidine-34 is involved in a hydrogen bond with the aspartate at 32 and hence has a relatively well-defined χ^1 angle. The leucine at position 50, on the other hand, is rather poorly defined because the corresponding leucines in the homologues fall into two structural classes: in *A. nidulans* and *C. beijerinckii* χ^1 is about $-60°$, while in *C. crispus* it is $-150°$. Thus the distance and torsion constraints generated by our algorithm included this entire range, and in four of the computed *E. coli* structures the χ^1 angle of leucine-50 is near $-150°$ while it is in the neighborhood of $-60°$ in the remaining six structures.

The Energy Minimized Ensemble

Since the dynamical energy minimization used a relatively "soft" harmonic

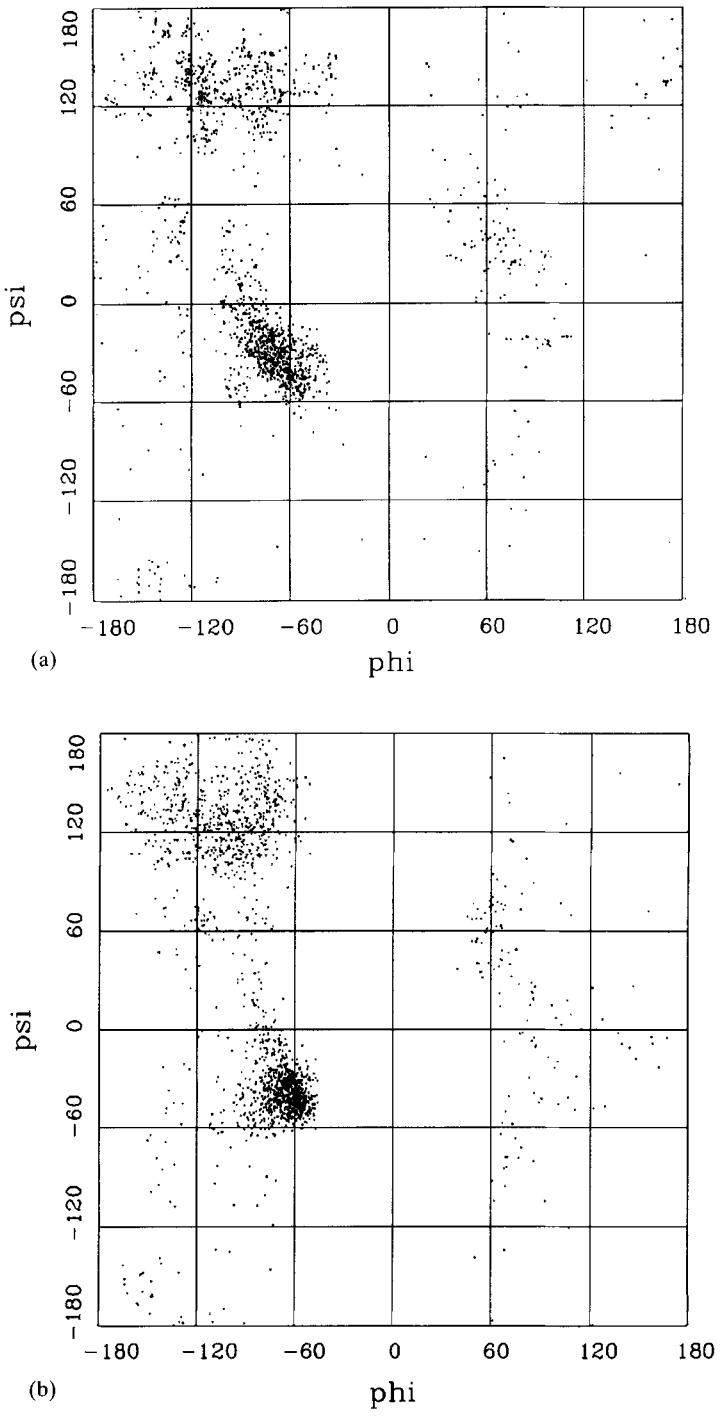
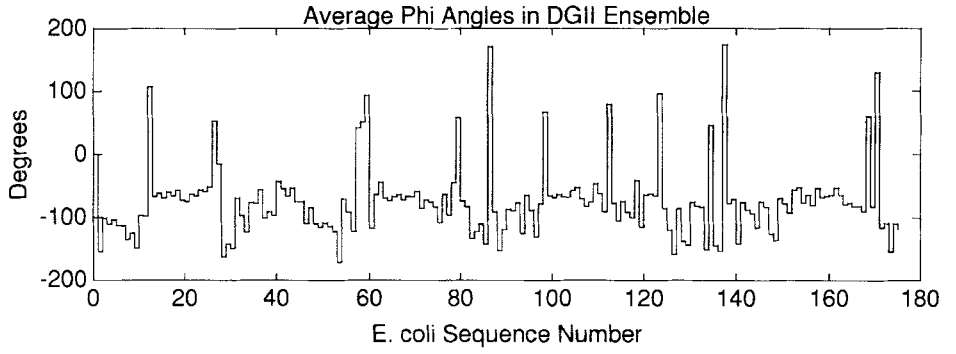
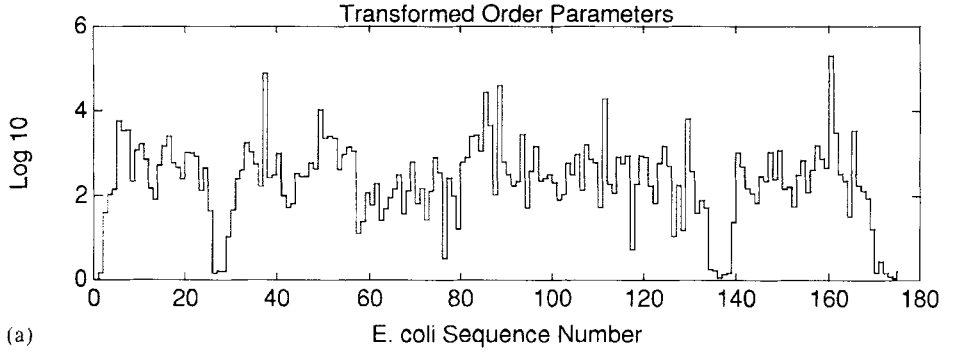
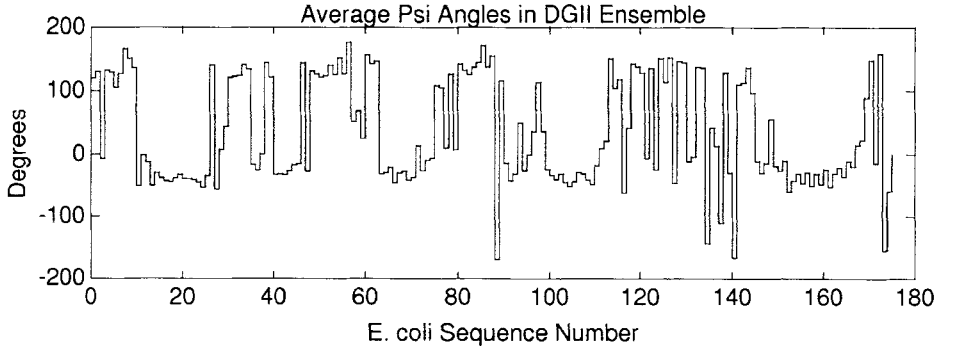


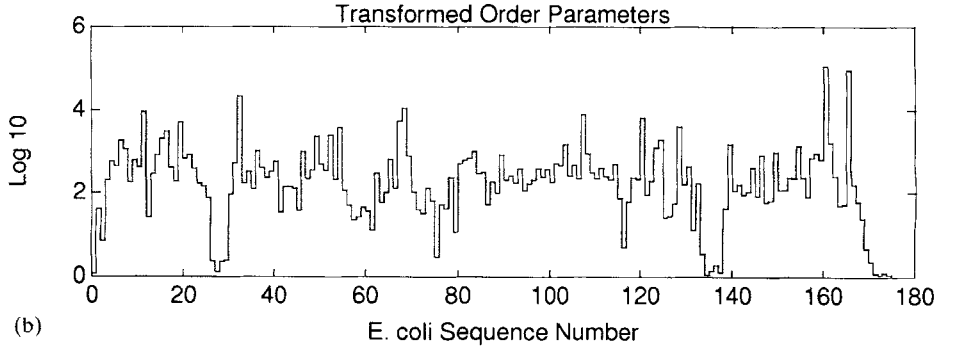
Figure 6 Ramachandran plot of $10 \times (176-2) = 1740$ residues in the ensemble of 10 distance goemetry structures (a) and the 10 structures after refinement versus the *CVFF* potential (b).





(a)





(b)

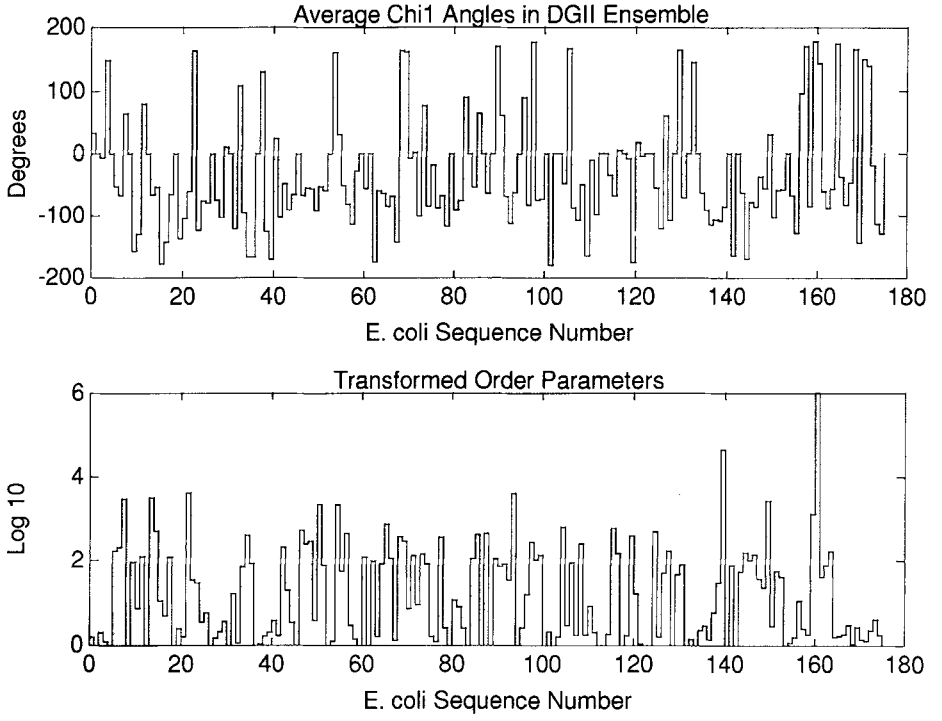
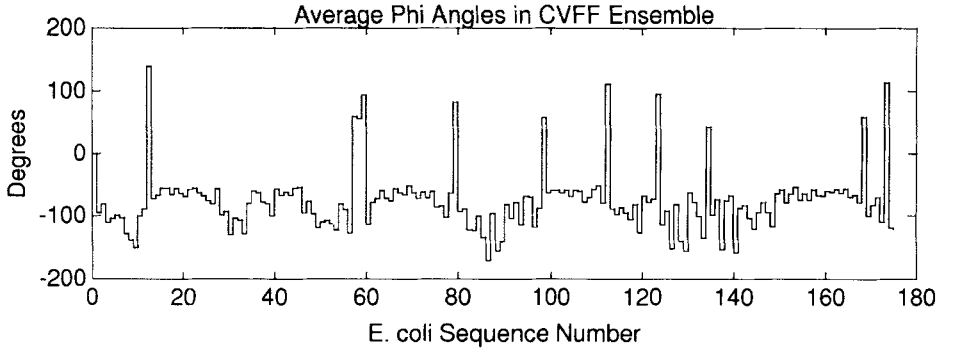
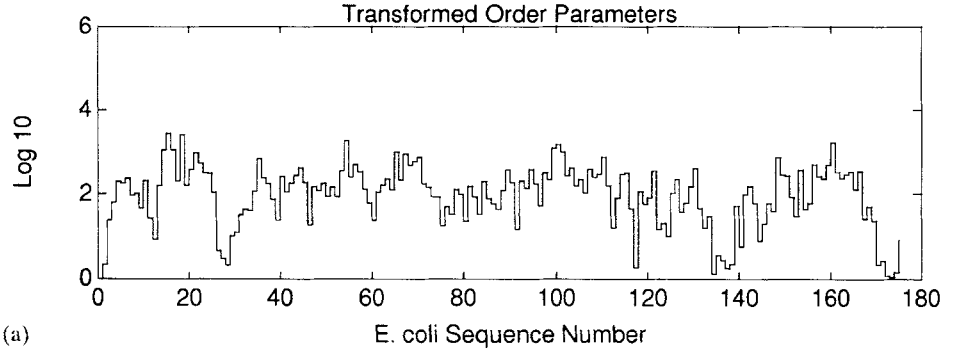


Figure 7 The average angles and corresponding transformed order parameters plotted versus the sequence number for the ϕ angles (a), ψ angles (b) and χ^1 angles (c) of the distance geometry ensemble. For those residues that have no χ^1 angle, the average angle was arbitrarily set to zero and the transformed order parameter to one.

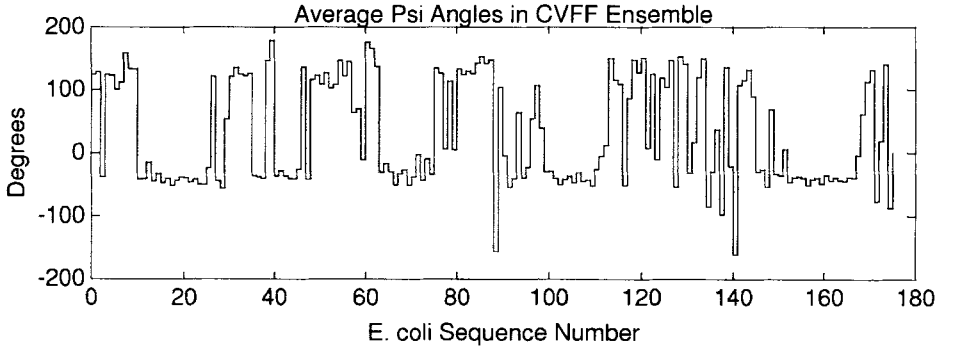
potential to enforce the constraints, the number and the magnitude of the violations increased substantially even before the final unconstrained energy minimization. Even so, large violations were not numerous, with only 1 to 10 distance violations over 1.0 Å and 0 to 3 torsion violations exceeding 10° in the ten structures. The constraint violations increased further in size and number during the final unconstrained minimization, after which the number of distance violations exceeding 1.0 Å ranged from 43 to 135 with an average of 84.1, while the number of torsion violations exceeding 10° ranged from 26 to 46 with an average of 32.7. The maximum distance violation ranged from 1.79 to 4.38 Å with an average of 2.56 Å, and the maximum torsion violations ranged from 35° to 104° with an average of 54°. Although these violations are much larger than those in the distance geometry structures, given the heuristic procedure by which the constraints were obtained it is not unreasonable to allow such violations if necessary to reduce the energy.

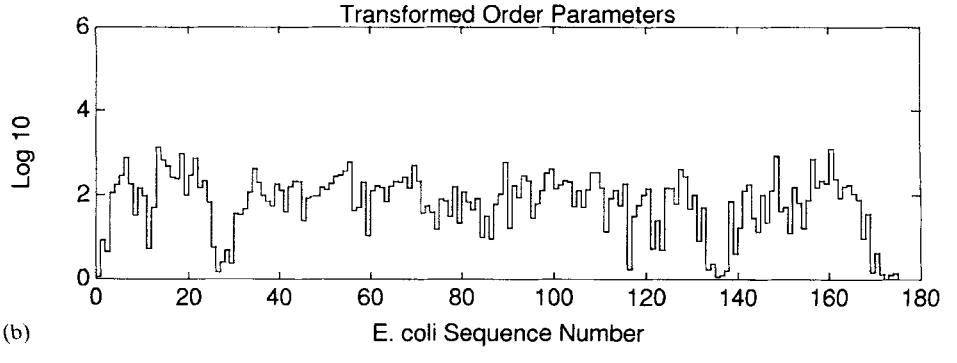
The energies attained after the final minimization ranged from −29259 to −29648 kcal, which ran about 400 kcal less than the energy after the dynamical annealing. The change in the structures during the final minimization ranged from 0.34 Å to 0.50 Å in the heavy atom RMSD's of residues 1 through 176. The overall change in the structures during the entire refinement (annealing plus minimization) ranged from 1.45 Å to 1.60 Å in heavy atom RMSD, or 0.95 Å to 1.17 Å for the alpha-carbons. These numbers are comparable to the RMSD's among the members of the distance geometry ensemble, which together with their similar energies





(a)





(b)

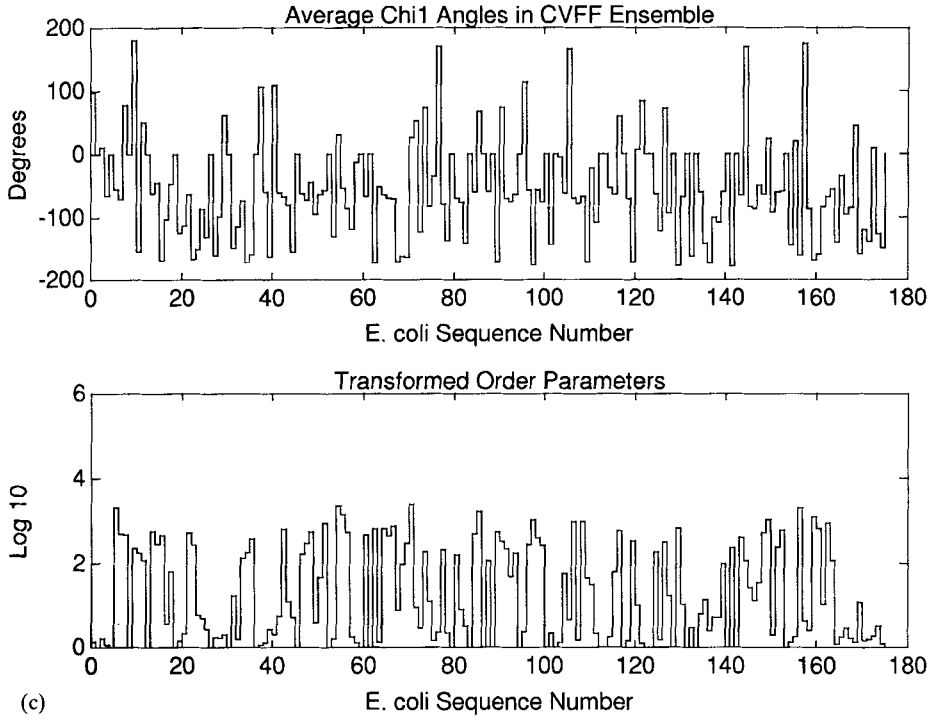


Figure 8 The same plots as in Figure 7 for the ensemble after energy minimization versus the *CVFF* potential.

implies that the simulated annealing procedure was powerful enough to search a substantial region of the conformation space around each starting structure for an energy minimum that was (largely) consistent with the constraints.

The alpha-carbon RMSD between residues 2 through 170, averaged over all pairs of the ten refined structures, was 1.00 Å (maximum 1.28); the corresponding heavy atom RMSD averaged 1.66 Å (maximum 1.94 Å). These numbers are slightly higher than the corresponding numbers for the unrefined *DG-II* ensemble, in part because the structures expanded slightly on refinement in order to accommodate

Table 4 Results of dynamical energy refinement.

Structure number	*Number and size of violations*				*Change (RMSD)*		*Energy (kcal)*
	>1 Å	*Max.*	*>10°*	*Max.*	*Alpha*	*Heavy*	
I	74	2.40	27	34.7	1.16	1.55	−29630
II	65	2.83	34	45.6	1.02	1.45	−29587
III	130	4.38	32	48.6	1.00	1.61	−29580
IV	135	2.58	34	82.3	0.95	1.53	−29604
V	100	1.79	35	103.7	0.96	1.56	−29414
VI	74	2.14	46	43.2	1.06	1.59	−29546
VII	59	3.05	26	42.5	1.17	1.55	−29523
VIII	96	2.45	27	45.7	1.08	1.56	−29456
IX	65	1.99	37	44.2	1.08	1.60	−29648
X	43	1.95	29	46.6	1.06	1.52	−29259

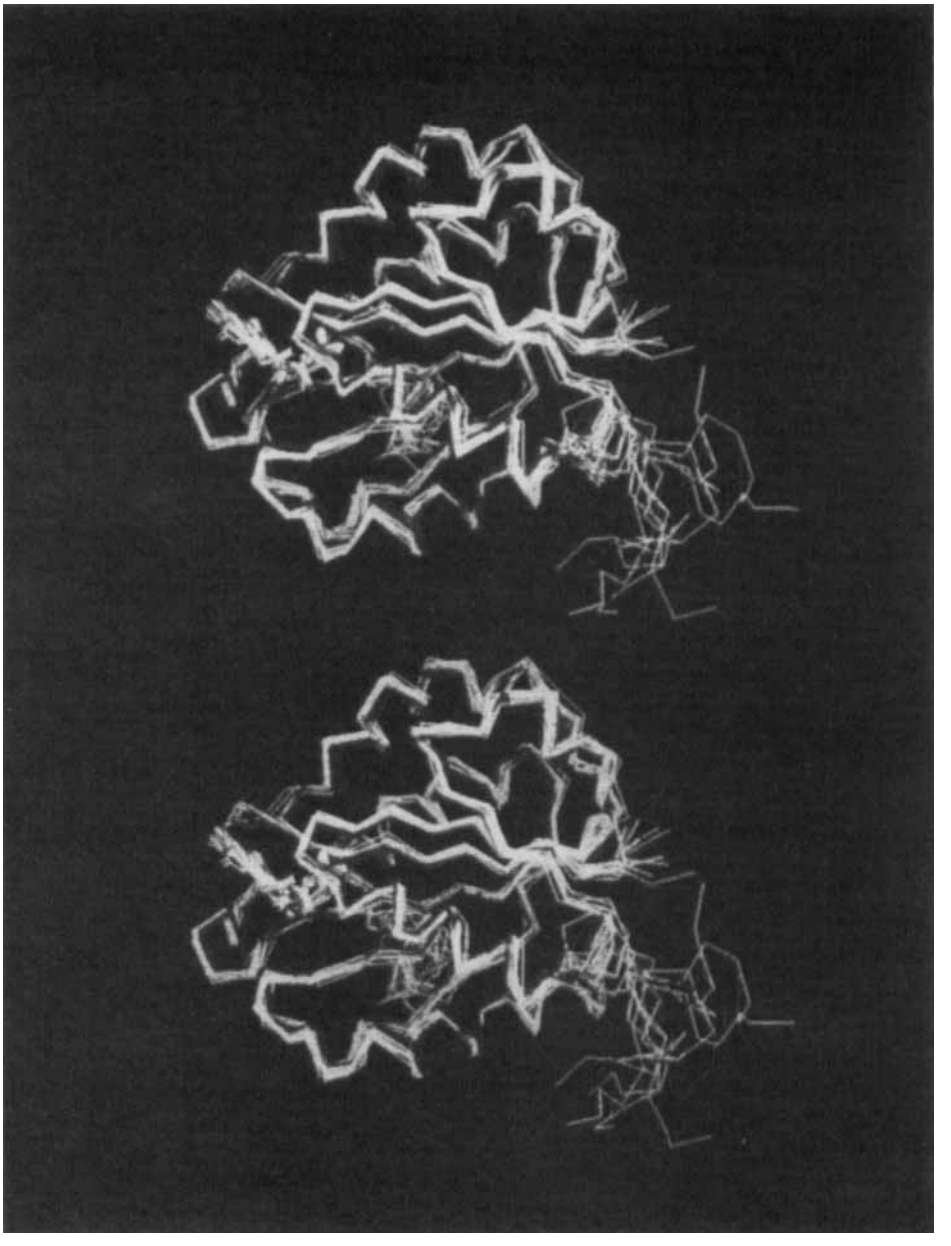


Figure 9 Stereoview comparing the alpha-carbon traces and flavin mononucleotide cofactors of the distance geometry and energy minimized ensembles. All structures have been superimposed so as to minimize the alpha-carbon RMSD to the corresponding SCR's of the *A. nidulans* Flavodoxin (not shown). The distance geometry ensemble is drawn in red, the energy minimized ensemble in green. (See Colour Plates)

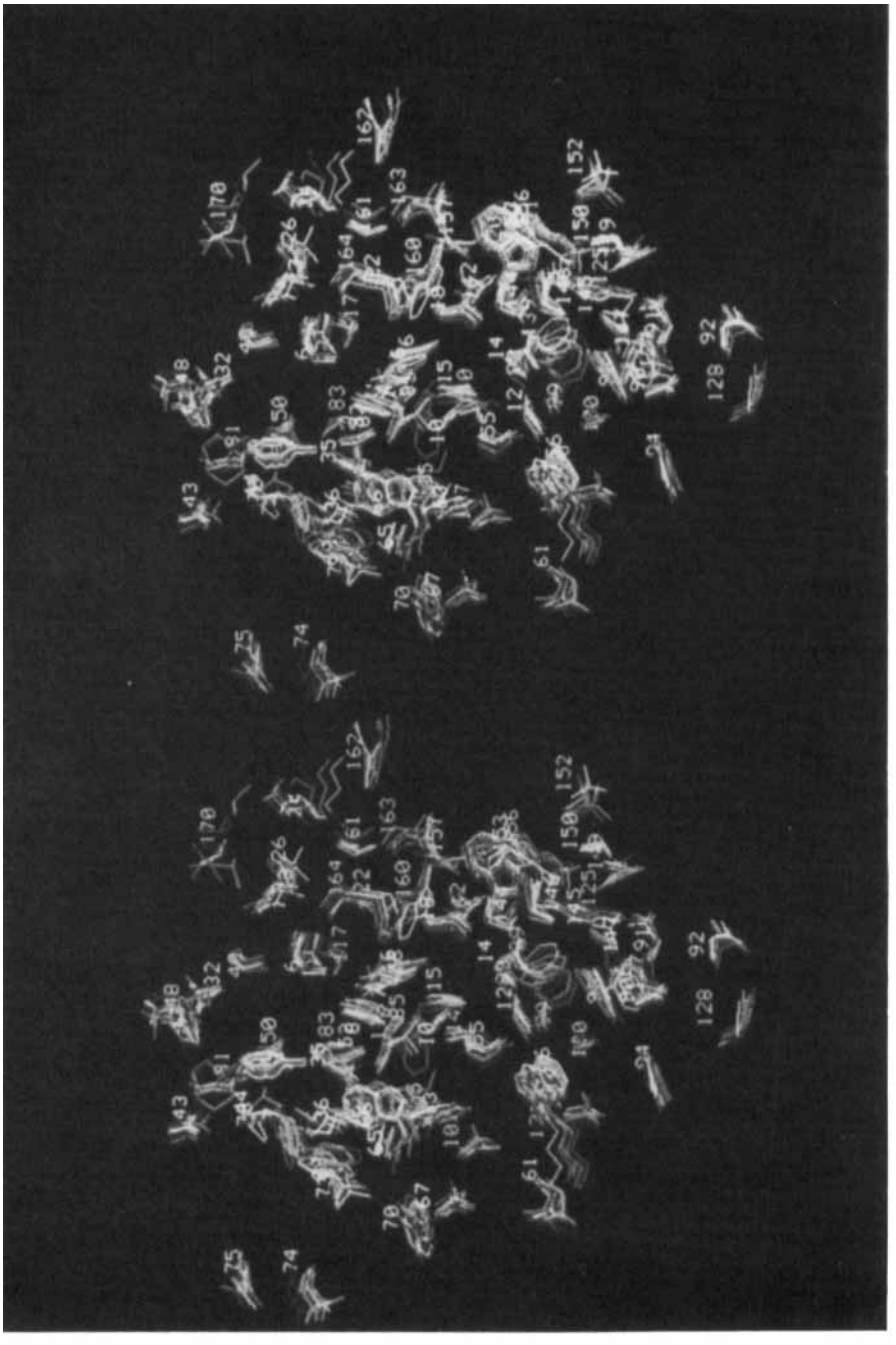
the added protons (the radius gyration of the alpha-carbons increased by 2.6%), but mostly because the constraints were not as strongly enforced. Figure 9 shows the alpha-carbons and flavin mononucleotide cofactor before and after energy minimization, from which it may be seen that there were systematic changes in the ensemble upon minimization. Despite the small increase in the heavy atom RMSD among the structures, the stereoview of the sidechains that were conserved between *E. coli* and at least one of its homologues (Figure 10(a) and (b)) shows that most of these sidechains were fairly well-defined, and that their spatial dispersion was decreased by the refinement.

As expected, the distribution in the ϕ/ψ angles improved on dynamical refinement versus the relaxed restraints (Figure 6(b)), and showed a distribution very similar to that found in crystal structures. The sequence dependence of any systematic changes in the distribution of the torsion angles on refinement, on the other hand, is revealed by a comparison of the plots of the average angles and transformed order parameters in the *DG-II* and *CVFF* ensembles (Figure 7 and 8). From these plots it may be seen that although the average ϕ and ψ angles did not change appreciably, the precision with which they were defined *decreased* slightly as a rule on refinement. This, once again, is probably due to the lower weight given to the restraints (including the covalent geometry), and to the fact that the restraints were often incompatible with the energy. The average χ^1 angles, on the other hand, did change appreciably in many cases, and they became better defined as a rule: 56% of the residues with χ^1 angles now have $\log_{10}(\hat{\xi}) > 1$. This is probably due to the elimination of eclipsed conformations by the refinement.

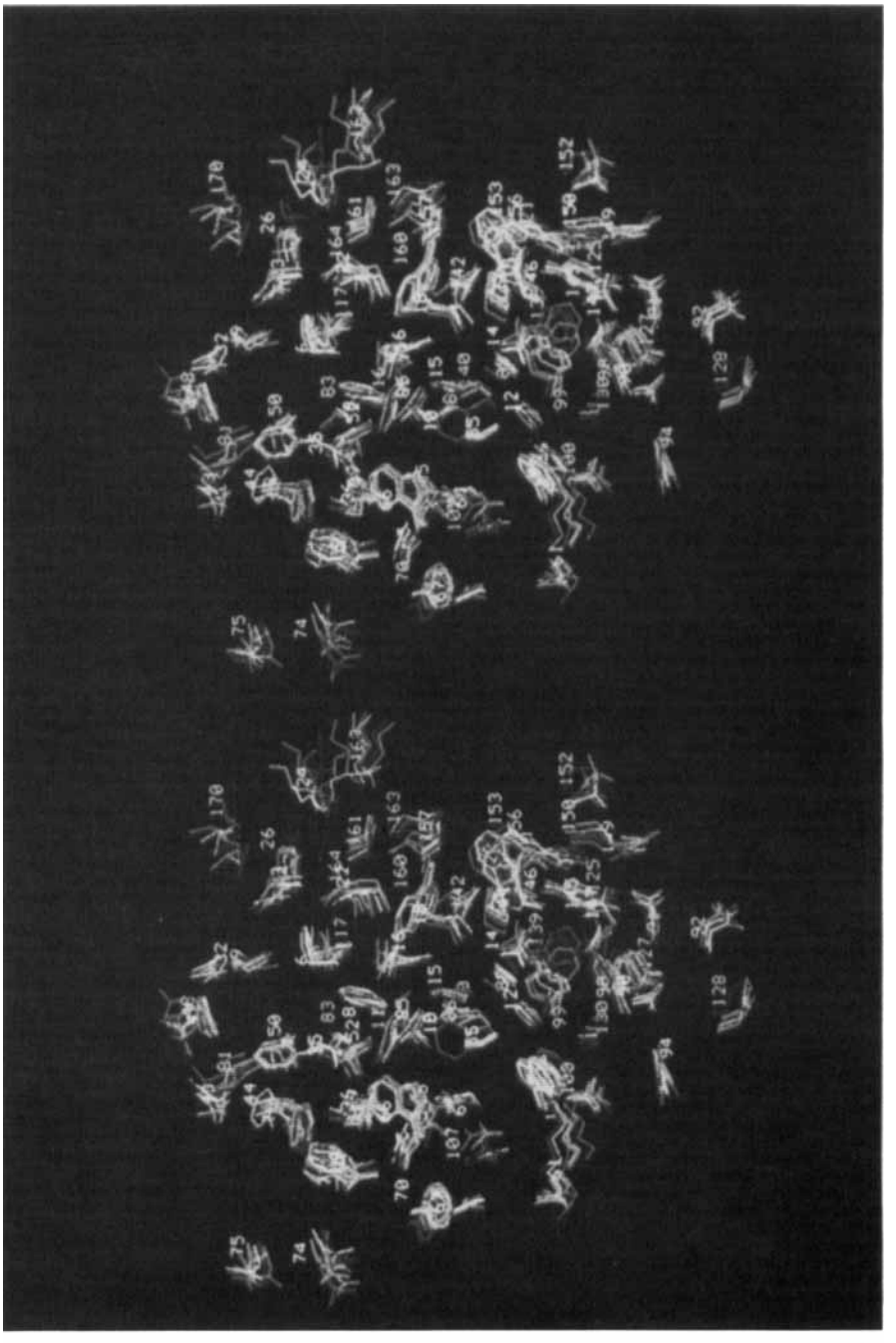
The listing of the hydrogen bonds in the structures produced by *Discover* implied that between 55 and 60% of those that were constrained in the distance geometry calculations were actually formed in the refined structures, and that 31 of 116 pairs of atoms involved in hydrogen bond constraints did not form hydrogen bonds in any of the refined structures. The criteria used by *Discover* for hydrogen bonds are a little on the stringent side, however. When all acceptor – proton pairs separated by less that 2.5 Å were considered to be hydrogen bonds, they were formed an average of 65% of the time, and only 15 did not occur in any of the refined structures; when all acceptor – proton pairs separated by less than 3.0 Å were considered to be hydrogen bonds they were formed an average of 82% of the time, and only six failed to be formed in any of the ten structures. One of these six was due to *Insight* using a tautomer of neutral histidine in the refinement, and another two were due to ambiguities in the labeling of the carboxyl oxygens of aspartate and glutamate. Thus, our strategy for obtaining a homologous hydrogen bonding network in the computed structures usually succeeded, even though no protons were used in the distance geometry calculations.

The Plausiblity of the Structures

A comparison of the alpha-carbons in the SCR's of the computed *E. coli* structures with the corresponding alpha-carbons of the homologues yielded the RMSD's shown in Table 5. Figure 11 shows a stereoview of the alpha-carbon traces and flavin mononucleotide cofactors of *DG-II* ensemble and the *C. beijerinckii, C. crispus* and *D. vulgaris* structures all superimposed on the *A. nidulans* structure so as to minimize the alpha-carbon RMSD among their common SCR's. The structures are most similar to the *A. nidulans* structure, and they diverged from it and the other homologues on refinement. Whether this trend was correct or not



(a)



(b)

Figure 10 Stereoview of the heavy atoms in the conserved sidechains of the distance geometry (a) and energy minimized (b) ensembles. The structures have been aligned so as to minimize the alpha-carbon RMSD between residues 2–170 of the first structure in each ensemble, and the structures have each been given a different color. (See Colour Plates)

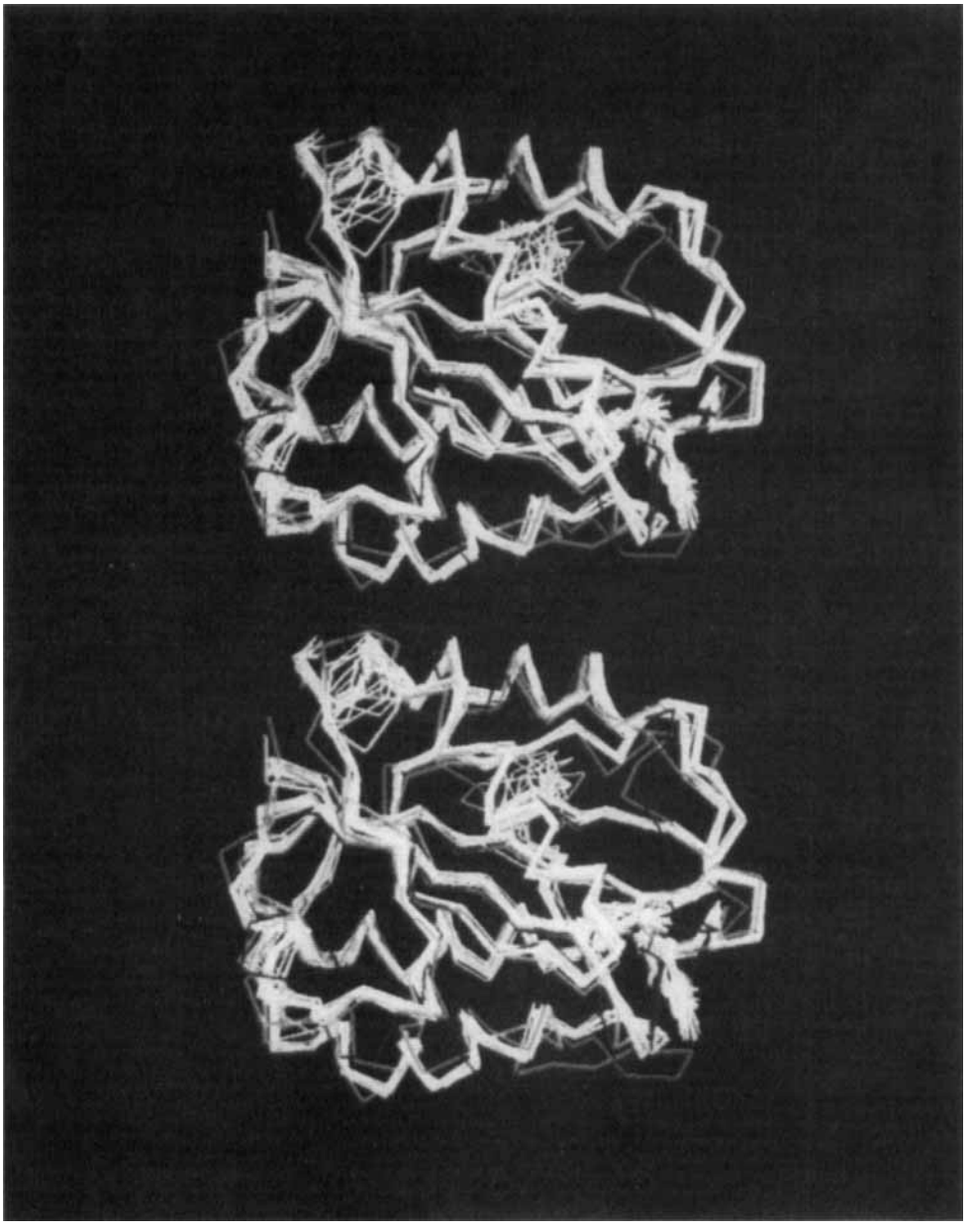


Figure 11 The alpha-carbons of residues 2–170 and flavin mononucleotide cofactor of the distance geometry ensemble along with the four homologous structures, all of which have been superimposed so as to minimize the RMSD between the alpha-carbons of the SCR's that they have in common with the *A. nidulans* structure. The structures in distance geometry ensemble are **white**, while the *A. nidulans* structure is **yellow**, the *C. beijerinckii* structure is **red**, the *C. crispus* structure is **green** and the *D. vulgaris* structure is **violet**. (See Colour Plates)

Table 5 Alpha-carbon RMSD's between SCR's of the *E. coli* structures and its homologues (Å).

DG-II	*I*	*II*	*III*	*IV*	*V*	*VI*	*VII*	*VIII*	*IX*	*X*
A. nidulans	0.66	0.75	0.72	0.63	0.67	0.67	0.66	0.76	0.68	0.66
C. beijerinckii	1.72	1.75	1.73	1.83	1.71	1.93	1.73	1.80	1.70	1.63
C. crispus	1.39	1.34	1.39	1.33	1.31	1.26	1.38	1.29	1.31	1.42
D. vulgaris	1.15	1.15	1.20	1.24	1.30	1.27	1.16	1.13	1.24	1.21
CVFF										
A. nidulans	0.97	1.09	1.14	0.97	1.05	1.04	1.01	1.05	0.97	0.97
C. beijerinckii	1.98	1.96	1.90	2.00	1.81	2.03	1.94	1.82	1.94	1.87
C. crispus	1.52	1.65	1.62	1.65	1.51	1.44	1.52	1.60	1.59	1.55
D. vulgaris	1.47	1.61	1.43	1.50	1.45	1.56	1.44	1.51	1.48	1.50

will have to wait for the structure to be determined experimentally. Nevertheless, the observed differences between the computed structures and their homologues, among the homologues themselves, and their sequence identities, make us fairly confident that the alpha-carbons of the SCR's are with 1.5 Å of the "correct" structure in all structures of both the distance geometry (*DG-II*) and refined (*CVFF*) ensembles.

One thing that is reasonably certain is that no single structure in our ensemble is correct in all its features. Instead, the structures are probably "chimeric", in that for example one structure might be correct in one region but wrong in another while the situation may be just the opposite in another of the structures. Since the C-terminal end of the molecule is extended and not folded back onto the rest of the protein in all of the computed structures, it is certainly incorrect – or at least different from what it will be in any crystal structure. Nonetheless, it has been found in protein structure determinations by nuclear magnetic resonance that nonconserved terminal regions of proteins tend to be unstructured in solution, even if they are well-ordered in a crystal structure [5]. Hence it is possible that there is no really correct structure for the C-terminus.

The turns at the site of the deletion and the insertion of the *E. coli* sequence with respect to its homologues were examined in some detail by the graphical method described in [35]. At the site of the deletion (residues 28–29 in *E. coli*), the homologues differ from each other quite substantially in their ϕ and ψ angles, particularly *C. crispus*. The *C. beijerinckii* and *D. vulgaris* structures have a type I turn [37] followed by a residue with positive ϕ angle, whereas in *A. nidulans* the positive ϕ angle precedes the type I turn. Six of the ten refined *E. coli* structures exhibit a type I turn at residues 28–29 (numbers III, IV, VI, VII, IX and X), which is preceded in all but one case by a positive ϕ angle (X). The remaining four structures do not exhibit any well-defined patterns, and it therefore seems probable that this turn will have the *A. nidulans* conformation in the crystal structure.

The site of the insertion at residues 135–138 does not exist in either *C. beijerinckii* or *D. vulgaris*, because both of these have a massive deletion extending from residues 120–140 of the *E. coli* sequence. In *A. nidulans*, this region contains a type II at the corresponding residues 134–135, which adjoins a type I′ at 135–136 and continues on in an extended conformation. In *C. crispus*, the type II turn at the corresponding residues 138–140 adjoins a distorted type II′ turn, and then goes into an extended conformation. Six of the ten *E. coli* structures (II through V, IX and X) exhibit a pattern like *C. crispus* in residues 134–136, but residues 137–138 show no trend whatsoever. In three of the structures (VIII, IX and X) these residues

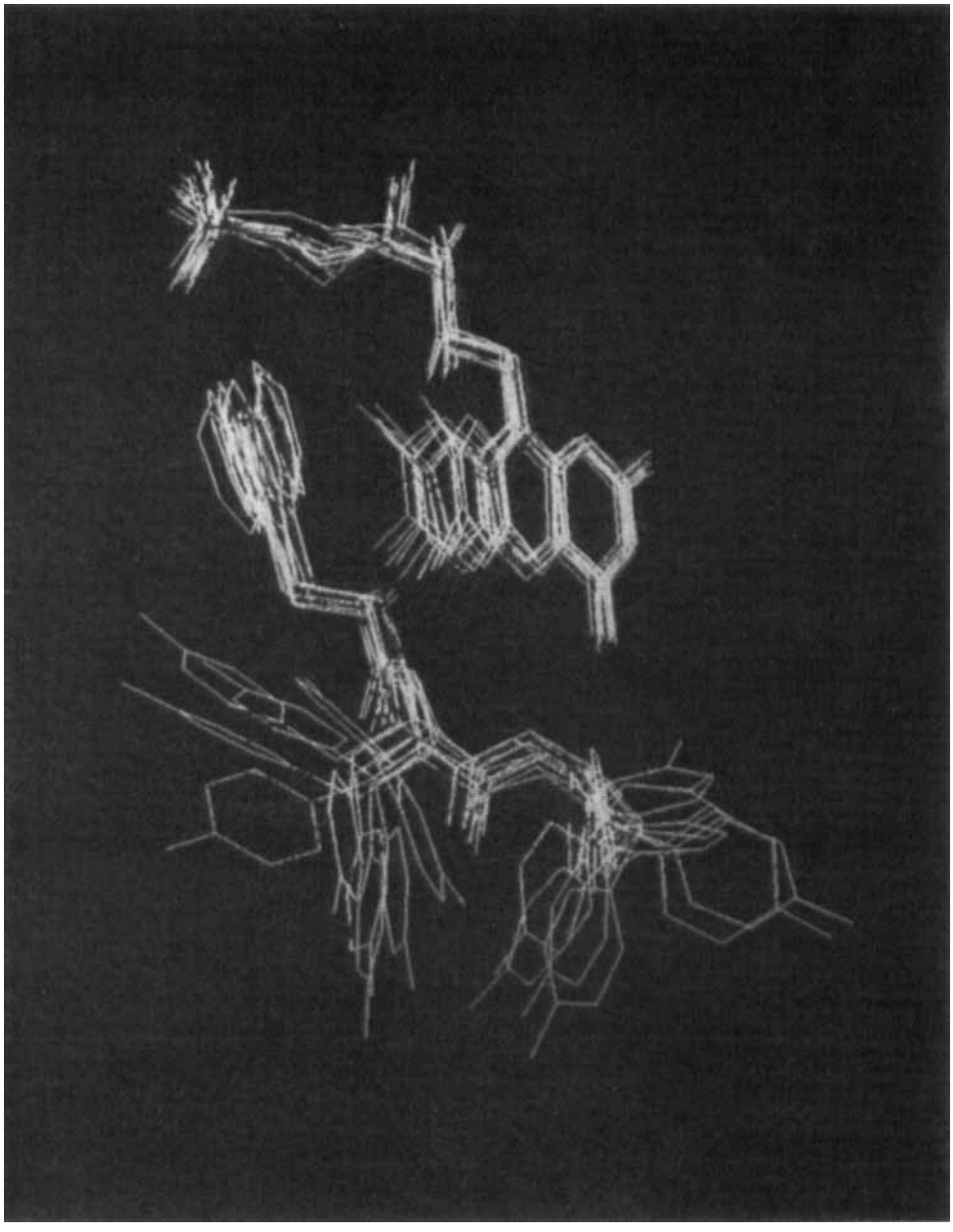


Figure 12 The residues 57 through 59 and the flavin mononucleotide cofactors of the *E. coli* structures before and after energy minimization. The structures have been aligned on the first structure in each ensemble so as to minimize the RMSD between the backbone atoms of residues 57–59; the colors are the same as in Figure 9. (See Colour Plates)

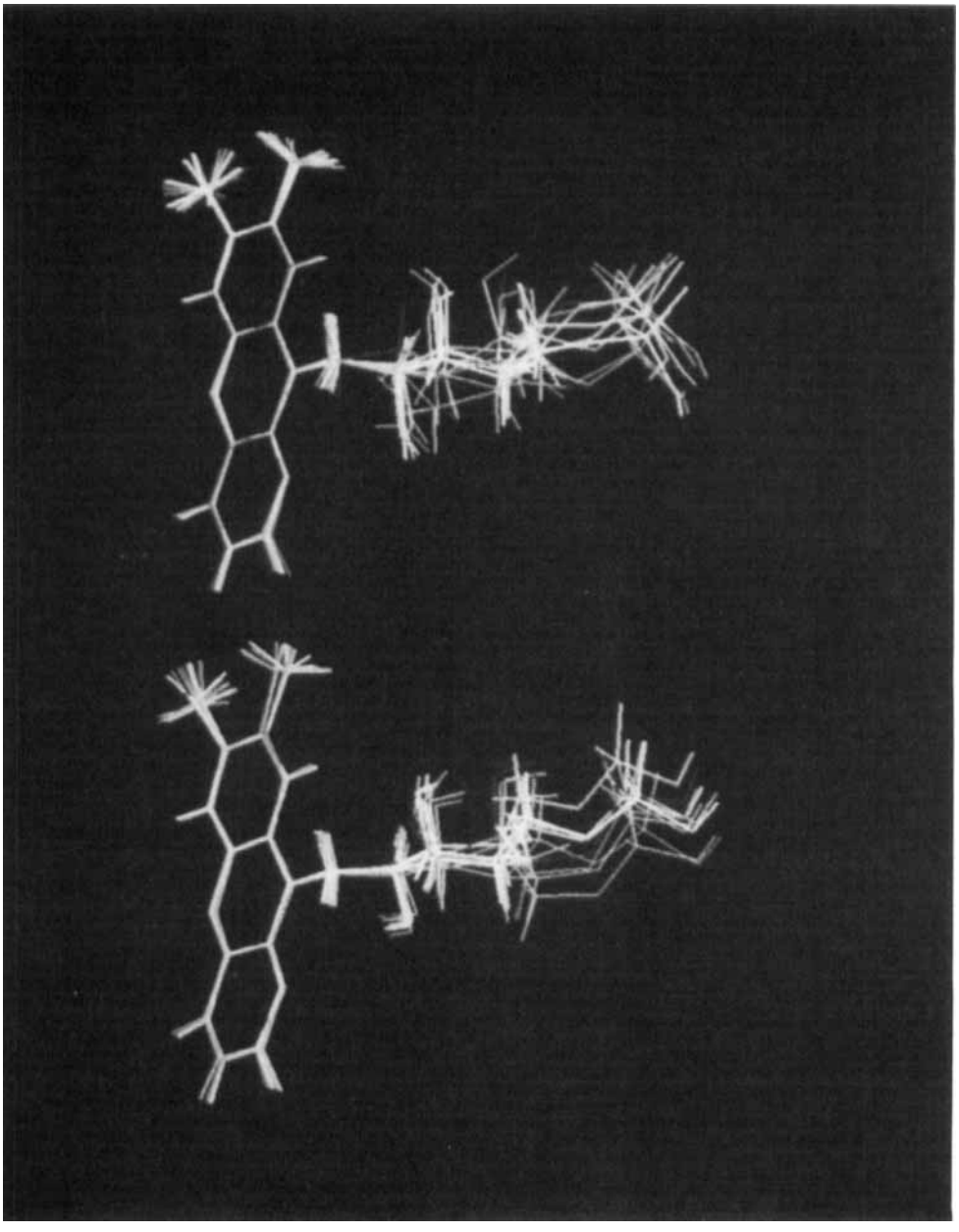


Figure 13 The flavin mononucleotide cofactor before (left) and after (right) energy minimization, where the structures in each ensemble have been aligned so as to minimize the RMSD between their flavin rings. The colors are the same as in Figure 10. (See Colour Plates)

exhibit positive ϕ angles, while residue 139 is the β region in all ten structures. It should also be mentioned that in the 2D nuclear magnetic resonance spectra of the *A. nidulans* Flavodoxin, the sequential nuclear Overhauser enhancements $d_{\alpha N}$ and d_{NN} were not observed in this region [38], which implies that it is disordered in solution. Given the residues 135–137 of the *E. coli* are all aspartate, it would be surprising if the same were not true for the *E. coli* form as well.

Another noteworthy feature of the Flavodoxin from *E. coli* is the presence of a pair of adjacent tyrosine residues at positions 58 and 59, which have no analogue in any of the homologues. In all ten of the refined *E. coli* structures this region has a type I′ turn followed by a positive ϕ angle at glycine-60, so we predict that this is the correct backbone conformation. As shown in Figure 12, however, the sidechain conformations are essentially random, although after refinement there is some preference for staggered rotomers and for the rings to lie flat against one of the many other aromatic sidechains in this region, or the methyl groups of the flavin ring. Thus, there is little we can say about their conformations at this time.

The final feature of the computed structures that we wish to consider is the conformation of the sugar chain of the flavin mononucleotide cofactor. In all of the homologous structures these angles are all approximately *trans* except for the angle about the *C3′–C4′* bond, which is *gauche*$^+$ in the six structures I, IV through VII and IX, but *gauche*$^-$ in the four structures II, III, VIII and X. Although the refinement reduced the spatial dispersion of the sugar chain substantially (see Figure 13), it did not succeed in changing the rotomeric state of the *C3′–C4′* angle. What it did change was the rotomeric state of the *C5′–O5′* angle from *trans* to *gauche*$^-$ in all of those structures in which the *C3′–C4′* angle was also *gauche*$^-$. Although this succeeded in making the chain as a whole follow a similar path through space as the *gauche*$^+$ conformations, because the *gauche*$^+$ conformation occurs in all four homologues we strongly suspect it is the correct conformation.

CONCLUSIONS

In the preceding section the computed *E. coli* structures were subjected to a careful analysis with the goal of discovering which structural features were not well-determined by the geometric constraints derived from the alignment, along with anything else that might be wrong with the computed conformations. From this, it has become clear that there are some things we might have done better. For example, torsion angle constraints should have been applied to the sugar chain of the flavin mononucleotide cofactor, to ensure that its bonds came out in the same rotomeric states as in the homologues. In addition, the clustering observed in the Ramachandran map of the distance geometry structures implies that the tolerance used for deriving the local backbone and sidechain constraints was probably a little too low, and in future calculations we would recommend a value of $\delta = 1.0$ Å for these constraints.

On the positive side, we point out that we easily computed the ten complete *E. coli* Flavodoxin structures shown in Figure 14, all of which are certainly correct in their chain fold, the location of their secondary structure elements, the placement of the flavin mononucleotide cofactor, and the conformation of many of the essential sidechains – all without the benefit of any direct experimental data. The procedure required relatively little time and relatively few arbitrary decisions once the alignment had been decided upon, and the majority of it could readily be

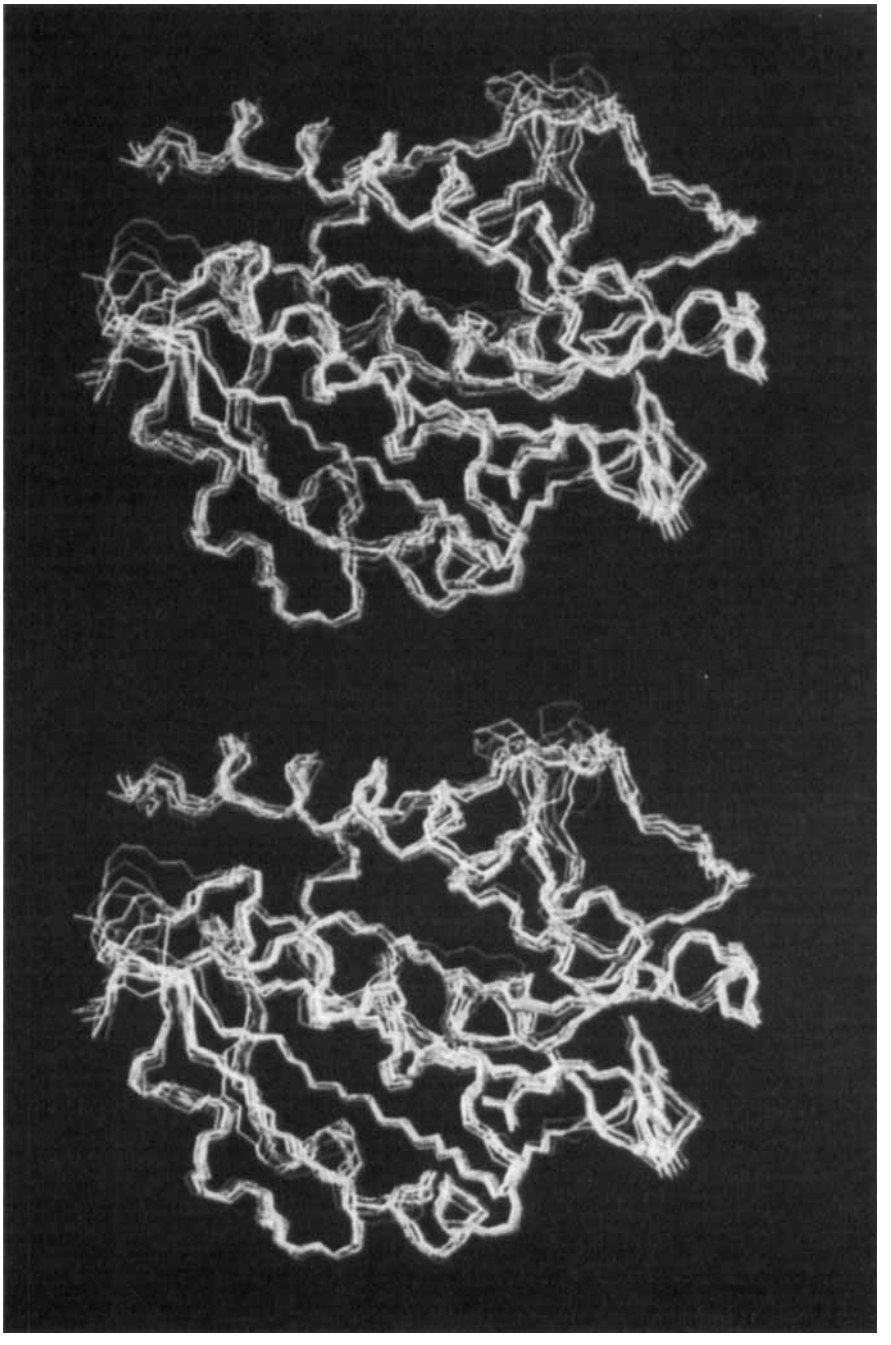


Figure 14 Stereoview of the final, energy minimized ensemble of ten *E. coli* Flavodoxin structures. Only the *N*, *CA* and *C* atoms in residues 2–170 are shown, together with the flavin mononucleotide cofactor's heavy atoms. The last nine structures have been superimposed on the first so as to minimize the RMSD between the alpha-carbons of residues 2–170, and each structure has been given a different color. (See Colour Plates)

automated. While the computations involved may seem undesirably long, most of the computer time was consumed by the refinement, because it was done rigorously with solvation accounted for by including all necessary water molecules explicitly. Given that the backbone conformation generally comes out so well, more limited systematic searches for the conformations of sidechains that cannot be determined from the available homologues could probably be used instead [11, 12, 13, 14]. This should both reduce the computer time required while at the same time possibly improving the predictions.

Although many detailed features of the conformation remain underdetermined, the distance geometry approach has the advantage that these features can be identified by comparing the members of the ensemble with one another to find out how much and in what ways they differ. This reduces the danger of making false predictions about the structure, and one can, if so inclined, proceed to resolve the underdetermined features by whatever means seems most suitable, experimental or theoretical. In the present case, a fluorescence transfer or NMR study might be used to obtain additional constraints on the sidechains of the double tyrosine at positions 58–59. The possibility of combining constraints derived from homology studies with constraints derived from experiments to determine a better structure than could be derived from either source alone is in fact another significant advantage of our approach. In this context, it is worth stressing that diverse geometric constraints often act together in surprising ways to determine conformational features about which one has no *a priori* information.

With the possible exception of the C-terminal end, all of the structures produced are in accord with the basic principles of protein conformation, and we doubt that an expert could easily distinguish the actual structure (once it has been determined) from any of the structures in the refined ensemble. It has recently been reported that various simple potential functions can identify incorrect protein folds [39, 40, 41, 42, 43], and we stand ready to provide any of these investigators with our structures to see if they can choose the "best". The structures computed in this paper have already been given to the laboratory of Prof. M. Ludwig at the University of Michigan, where the crystal structure of *E. coli* Flavodoxin is in progress. By computing the electron density from these structures, averaging it appropriately over the entire ensemble to eliminate contributions from underdetermined atoms, and Fourier transforming to obtain the complex structure factors, we hope that it will prove possible to obtain a reasonably good estimate of the phases and thereby eliminate the need for multiple isomorphous replacements.

Acknowledgements

This work was supported by NIH grants GM-38221 and GM-47467. Thanks are also due to Dr. K. Watenpaugh, Prof. K. Fukuyama and Prof. M. Ludwig for making the coordinates of their latest Flavodoxin crystal structures available to us.

References

[1] M.E. Goldberg, "The second translation of the genetic message", *Trend. Biol. Sci.*, Oct.: 388–391 (1985).

[2] G.M. Crippen and M.E. Snow, "A 1.8 Å resolution potential function for protein folding", *Biopolymers*, **29**, 1479–1489 (1990).

[3] G. Nemethy and H.A. Scheraga, "Theoretical studies of protein conformation", *FASEB J.*, **4,** 3189–3197 (1990).
[4] J. Skolnik and A. Kolinski, "Dynamic monte carlo simulations of a new lattice model of globular protein folding, structure and dynamics, *J. Mol. Biol.*, **221,** 499–531 (1991).
[5] G. Wagner, S. Hyberts, and T. Havel, "NMR structure determination in solution: A critique and comparison with X-ray crystallography", *Ann. Rev. Biophys. Biomol. Struct.*, **21,** 167–198 (1992).
[6] R.F. Smith and T.F. Smith, "Automatic generation of primary sequence patterns from sets of related protein sequences", *Proc. Natl. Acad. Sci.*, **87,** 118–122 (1990).
[7] B.L. Sibanda, T.L. Blundell, and J.M. Thornton, "Conformation of β-hairpins in protein structures, a systematic method of classification with applications to modeling by homology, electron density fitting and protein engineering", *J. Mol. Biol.*, **206,** 759–777 (1989).
[8] M.J. Rooman, J. Rodriguez, and S.J. Wodak, "Relations between protein sequence and structure and their significance", *J. Mol. Biol.*, **213,** 337–350 (1990).
[9] M. van Heel, "A new family of powerful multivariate statistical sequence analysis techniques", *J. Mol. Biol.*, **220,** 877–887 (1991).
[10] C.S. Ring, D.G. Kenller, R. Langridge, and F.E. Cohen, "Taxonomy and conformation analysis of loops in proteins", *J. Mol. Biol.*, **224,** 685–699 (1992).
[11] N.L. Summers and M. Karplus, "Construction of side-chains in homology modelling", *J. Mol. Biol.*, **209,** 785–811 (1989).
[12] L. Holm and C. Sander, "Fast and simple Monte Carlo algorithm for side chain optimization in proteins: Application to model building by homology", *Proteins: Struct. Funct. Genet.*, **14,** 213–233 (1992).
[13] C. Lee and S. Subbiah, "Prediction of protein side-chain conformation by packing optimization", *J. Mol. Biol.*, **217,** 373–388 (1991).
[14] J. Desmet, M. De Maeyer, B. Hazes, and I. Lasters, "The dead-end elimination theorem and its use in protein side-chain positions", *Nature*, **356,** 539–542 (1992).
[15] R.J. Feldmann, D.H. Bing, M. Potter, C. Mainhart, B. Furie, B.C. Furie, and L.H. Caporale, "On the construction of computer models of proteins by the extension of crystallographic structures," in *Macromolecular Structure and Specificity: Computer-Assisted Modeling and Applications*, B. Venkataraghavan and R.J. Feldmann, eds, Annals New York Academy of Sciences, **439,** 12–30 (1985).
[16] C.A. Schiffer, J.W. Caldwell, P.A. Kollman, and R.M. Stroud, "Prediction of homologous protein structures based on conformational searches and energetics", *Proteins*, **8,** 30–43 (1990).
[17] J. Greer, "Comparative modeling methods: Application to the family of the mammalian serine proteases", *Proteins: Struct. Funct. Genet.*, **7,** 317–334 (1990).
[18] J. Overington, M.S. Johnson, A. Sali, and T.L. Blundell, "Tertiary structural constraints on protein evolutionary diversity: Templates, key residues and structure prediction", *Proc. Royal Soc. London*, **241,** 132–145 (1990).
[19] R. Lee, "Protein model building using structural homology", *Nature*, **356,** 543–544 (1992).
[20] T.F. Havel and M. Snow, "A new method for building protein conformations from sequence alignments with homologues of known structure", *J. Mol. Biol.*, **217,** 1–7 (1991).
[21] T.F. Havel, "An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance," in *Progress in Biophysics and Molecular Biology*, D. Nobel and T.L. Blundell, eds, Pergamon Press, Oxford, England, **56,** 43–78 (1991).
[22] C. Osborne, L. Chen, and R.G. Matthews, "Isolation, cloning, mapping and nucleotide sequencing of the gene encoding flavodoxin in *Escherichia coli J. Bacteriol.*, **173,** 1729–1737 (1991).
[23] M.L. Ludwig and C.L. Luschinsky, *Chemistry and Biochemistry of Flavoenzymes, III*, chapter "Structure and Redox Properties of Clostridial Flavodoxin", CRC, 427–466, 1992. Press, Franz Müller, ed.
[24] K.D. Watenpaugh, L.C. Sieker, and L.H. Jensen, "The binding of riboflavin-5'-phosphate in a flavoprotein: Flavodoxin at 2.0 Å resolution", *Proc. Natl. Acad. Sci.*, **70,** 3857–3860 (1973).
[25] R.M. Burnett, G.D. Darling, D.S. Kendall, M.E. LeQuesne, S.G. Mayhew, W.W. Smith, and M.L. Ludwig, "The structure of the oxidized form of clostridial flavodoxin at 1.9 Å resolution: Description of the flavin mononucleotide binding site. *J. Biol. Chem.*, **249,** 4383–4392 (1974).
[26] W.W. Smith, K.A. Pattridge, M.L. Ludwig, G.A. Petsko, D. Tsernoglou, M. Tanaka, and K.T. Yasanobu, "Structure of oxidized flavodoxin from *Anacystis nidulans*", *J. Mol. Biol.*, **165,** 737–755 (1983).
[27] K. Fukuyama, S. Wakabayashi, H. Matsubara, and L.J. Rogers, "Tertiary structure of oxidized flavodoxin from an eukaryotic red algae *Chrondrus crispus* at 2.35 Å resolution", *J. Biol. Chem.*, **265,** 15804–15812 (1990).

[28] T.F. Havel and K. Wüthrich, "An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformation in solution", *J. Mol. Biol.*, **182,** 281–294 (1985).
[29] A.W.M. Dress and T.F. Havel, "Shortest-path problems and molecular conformation", *Discrete Applied Math.*, **19,** 129–144 (1988).
[30] G.M. Crippen and T.F. Havel, "Stable calculation of coordinates from distance information", *Acta Cryst.*, **A34,** 282–284 (1978).
[31] G.M. Crippen and T.F. Havel, *Distance Geometry and Molecular Conformation*, Research Studies Press, Letchworth, UK, ISBN 0-86380-073-4; J. Wiley and Sons, NY, ISBN 0-47192061-4 (1988).
[32] Jan de Leeuw, "Convergence of the majorization method for multidimensional scaling", *J. Classification*, **5,** 163–180 (1988).
[33] S.T. Rao and M.G. Rossmann, "Comparison of super-secondary structures of proteins", *J. Mol. Biol.*, **76,** 241–256 (1973).
[34] G.N. Ramachandran and V. Sasisekharan, "Conformation of polypeptides and proteins", *Adv. Protein Chem.*, **68,** 284–438 (1963).
[35] S.G. Hyberts, M.S. Goldberg, T.F. Havel, and G. Wagner, "The solution structure of eglin c based on measurements of many NOEs and coupling constants, and its comparison with X-ray structures", *Protein Sci.*, **1,** 736–751 (1992).
[36] T.F. Havel and K. Wüthrich, "A distance geometry program for determining the structures of small proteins and other macromolecules fromnuclear magnetic resonance measurements of ^1H–^1H proximities in solution", *Bull. Math. Biol.*, **46,** 673–698 (1984).
[37] J. Richardson, "The anatomy and taxonomy of protein structure", *Adv. Protein Chem.*, **34,** 167–339 (1981).
[38] R.T. Clubb, V. Thanabal, C. Osborne, and G. Wagner, ^1H and ^{15}N resonance assignments of oxidized flavodoxin from *Anacystis nidulans* with 3D NMR", *Biochemistry*, **30,** 7718–7730 (1991).
[39] J. Vila, R.L. Williams, M. Vasquez, and H.A. Scheraga, "Emprical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor", *Proteins*, **10,** 199–218 (1991).
[40] L. Holm and C. Sander, "Evaluation of protein models by atomic solvation preference", *J. Mol. Biol.*, **225,** 93–105 (1992).
[41] R. Luthy, J.U. Bowie, and D. Eisenberg, "Assessment of protein models with three-dimensional profiles", *Nature*, **356,** 83–85 (1992).
[42] V.N. Maiorov and G.M. Crippen, "A contact potential that recognizes the correct folding of globular proteins", *J. Mol. Biol.*, **227,** 876–888 (1992).
[43] C.E. Lawrence and S.H. Bryant, "Hydrophobic potentials from statistical analysis of protein structures", *Meth. Enzymol.*, **202,** 20–31 (1991).