Determination of the Spatial Distribution of Protein Structure Using Solution Data

Russ B. Altman[1], Bruce S. Duncan[2], James F. Brinkley[1], Bruce G. Buchanan[1] & Oleg Jardetzky[2]

There has recently been an increased interest in non-crystallographic methods for determining protein structure. This interest arises in part from the difficulty of procuring high quality crystals of many proteins, and the expense of crystallographic studies. However, there are also factors extrinsic to crystallography which make alternative methods attractive. First, there is interest in studying the ways in which solution structures might differ from crystal structures. Second, the large number of crystallographically determined structures established to date has offered valuable lessons and intuition about the ways in which proteins fold – yet has allowed few successful applications of this knowledge to the determination of other structures, suggesting that the picture of protein structure and dynamics observed from crystallography is incomplete. Finally, the increased ability to gain high resolution data from protein Nuclear Magnetic Resonance (NMR) spectra offers a rich alternative source of information for protein structure studies (Jardetzky & Roberts 1981, Wüthrich 1986). The NMR data can normally be interpreted in terms of allowed distance ranges (on the order of 2 to 4 Angstroms) between hydrogen atoms within the protein, and are, for the present, the major non-crystallographic source of detailed structural information.

1. TWO PARADIGMS FOR INTERPRETATION OF SOLUTION DATA

There is a critical difference between the study of protein structure in solution and in crystalline form. For the most part, all the individual protein molecules in a crystal (especially those that have been studied at high resolution) assume

[1]Knowledge Systems Laboratory, [2]Stanford Magnetic Resonance Laboratory, Stanford University, Stanford, California 94305, U.S.A.

NMR SPECTROSCOPY IN DRUG RESEARCH, Alfred Benzon Symposium 26, page 209–232.
Editors: Jerzy W. Jaroszewski, Kjeld Schaumburg, Helmer Kofod, Munksgaard, Copenhagen 1988.

the same conformation. It is therefore appropriate to model the structure as a rigid conformer with some parameter, such as the temperature factor, to indicate areas of uncertainty.

It may not be reasonable to expect such rigid pictures from solution data. Since solution data are averaged over a population of mobile molecules and are integrated over time, they present a more difficult problem of interpretation. This is further complicated by the relative sparseness of high resolution data which makes the structure underconstrained from the start. A unique single set of atomic coordinates from solution data (especially for larger proteins) is therefore often impossible to obtain and may in fact be misleading.

The most general and accurate structural description is a spatial distribution function for each atom. A spatial distribution function describes the probability that a given atom is at a certain location in space. In the case of crystals, the distribution function for most atoms is a very sharp normal curve with variance implied by the temperature factor. Solution studies however, for the reasons outlined, may imply a non-normal distribution for many atoms. It is desirable to characterize these distributions in some detail in order to understand the ways in which concerted motion may occur in the solution structure.

The most common form for non-crystallographic data is distances (or ranges of distances) between atoms in a structure. Such data can come as the distance between protons measured by NMR (NOE measurements) or as the distance between chemical groups measured by fluorescence energy transfer and chemical cross-linking experiments (Altman & Jardetzky 1986). Thus, methods for deriving solution structures must have efficient ways to use distance data. However, the problem of determining protein structure in solution is still underconstrained even with the best experimentally obtainable data sets and requires one (or more) of the following in order to more completely define the structure:

1. addition of non-distance experimental constraints
2. addition of theoretical constraints *or*
3. a method for generating a complete set of all conformations that satisfy the given constraints, generating the spatial distribution of allowed atomic positions.

Two general paradigms for the determination of protein structure have emerged: the *adjustment* paradigm and the *exclusion* paradigm.

Methods within the *adjustment paradigm* use a refinement cycle of:

1. Hypothesize initial conformation
2. Repeat:

evaluate structure
adjust structure

3. End when adjustments are small.

Two principal examples of currently used methods within the adjustment paradigm are distance geometry (Havel & Wüthrich 1984, Wüthrich *et al.* 1984, Frayman 1984, Gariepy *et al.* 1986, Braun & Gō 1985) and molecular dynamics (Levitt 1983a, Clore *et al.* 1986). The distance geometry method is based on minimizing a global error function for the constraints on the distances between atoms derived from experimental data. The evaluation function can vary in detail, but generally measures the degree to which the structure satisfies these constraints – usually using a least squares criterion. Its gradient is used to make incremental adjustments in the structure. The adjustments can be small corrections in the coordinates of individual atoms (Havel & Wüthrich 1984), or small changes in the dihedral angles of the backbone peptide and amino acids (Braun & Gō 1985). When the adjustments become trivially small, the iterations are ended. Distance geometry, therefore, augments distance constraints with the assumptions that (1) the minimum of its error function corresponds to the correct structure, and that (2) the gradient of the error function is a proper indicator of the best adjustment.

Molecular dynamics methods use an evaluation function which measures the total free energy of the molecule (Levitt 1983a, Clore *et al.* 1986). These methods are based on the hypothesis that the actual protein conformation will correspond to the minimum of a free energy function. This hypothesis adds a strong theoretical constraint to the problem. The evaluation function usually has a number of distinct terms corresponding to each of the forces that determine atomic position (such as van der Waals, ionic, and hydrogen bonding). The functional form and parameters of these terms are in general not known, and so empirical functions and parameter values must be used (often taken from gas phase experiments). The gradient of the free energy function gives an indication of the path towards a more favorable energy state, and is used for adjustment. Since the final minimal value of the energy function is not known *a priori,* there is a variety of ending conditions for these methods. For example, the convergence of multiple starting structures to the same (or similar) "equilibrium" structures is taken as good evidence that the equilibrium structure is correct.

Methods within the adjustment paradigm have been successful in estimating the conformation of a number of small proteins and peptides. They can be hampered, however, by the existence of local minima in the evaluation functions which do not allow for structural adjustment towards the actual global minimum.

There is also no theoretical guarantee that the conformation corresponding to the minimum of the evaluation function is actually the correct structure. In fact, there is evidence that the adjustment methods may not always produce structures which consistently explain the data (Gariepy *et al.* 1986, Lefevre *et al.* 1987, Madrid *et al.* 1988). It seems that the evaluation criteria may add "hidden constraints" to the system that are dependent on the evaluation function and the method of search and select different final structures according to their specific bias. In addition, these methods only generate single structures – even when more than one conformation may be consistent with the data. The only way to keep track of multiple plausible structures is to run the programs repetitively with different initial starting structures. However, it is difficult to ensure that the resulting sample of conformations is fully representative of the range of conformations.

Methods within the *exclusion paradigm* use a refinement cycle of:

1. Hypothesize all conformations
2. Repeat (for each constraint or set of constraints):
   evaluate structures (individually or by classes)
   exclude unacceptable structures
3. End when all constraints have been used.

This class of methods does not adjust errant structures, but instead excludes them from future consideration. The accepted structures can be rated by the degree to which they satisfy the constraints. These methods rely on a large set of initial structures many of which are eventually excluded, but some of which satisfy the constraints.

Methods within the exclusion paradigm use evaluation functions based on some measure of the agreement with experimental or theoretical data. They differ from the functions used in adjustment methods in at least three ways:

1. The selection of structures is not guided by the results of previous evaluation. The functions themselves can range from simple "accept/exclude" predicates to more complicated functions which return a measure of the degree to which the constraints are satisfied.
2. Since the functions can be applied sequentially to evaluate structures, they need not simultaneously embody all possible constraints on the structure.
3. Since the functions are independent, they can be applied in parallel.

There is thus considerably more freedom in the form and use of these functions. Constraints which are naturally expressed as distances between two atoms (e.g., NOEs, fluorescence transfer, cross-linking data) can be tested as distances, but

those that are better expressed as global constraints on the size or shape of a molecule (e.g., from small angle x-ray scattering, low resolution crystallography) can be tested with non-distance calculations.

The exclusion methods avoid the problems of local minima since there are no optimizations or adjustments. In addition, if the constraints do not determine a unique structure (but a family of structures) the exclusion methods will not exclude any members of the family. Finally, the relatively free functional form of the evaluation functions allows them to be developed, tested and validated independently. Indeed, the evaluation functions used by distance geometry and molecular dynamics can be incorporated into the exclusion paradigm by specifying the ranges of values which are considered acceptable and unacceptable.

The major problem for methods within the exclusion paradigm is to find tractable ways to keep track of all structures until the evaluation functions have significantly pruned down the search space. It is clearly impossible to enumerate all possible conformations exhaustively, and so techniques for compaction and expansion of structural representations must be developed.

For example, cytochrome-b562, discussed below, is a medium sized protein with approximately 1600 atoms. In the absence of *any* information about the connectivity or experimental constraints on these atoms, we can say that each atom can be anywhere, and there are an infinite number of conformations. The introduction of a volume constraint shows us that the atoms must be in a roughly elliptical volume that encompasses less than 20 cubic Angstroms. The number of positions per atom is now bounded by 20^3, and the number of conformations (sampled at 1 Angstrom) becomes $(20 \cdot 20 \cdot 20)^{1600}$. The introduction of covalent bonding information severely limits the number of positions for each atom, and the number of conformations also falls. In principle, the number of conformations that need to be considered at any point depends on the number of atoms in the structure and the number of feasible positions for each atom, as defined by known constraints. For proteins, the number remains very large.

2. THE PROTEAN SYSTEM

The PROTEAN system is a protein structure determination system which demonstrates the feasibility of the exclusion paradigm for obtaining detailed atomic structures. The input to the program is the primary structure of the protein, distance constraints based on NMR data, and volume and shape constraints from small angle x-ray scattering, light scattering or hydrodynamic measurements. The output of the program is a representative set of conformations which are compat-

ible with the data. The general strategies of the method, including hierarchical model building and the sequential application of constraints, have been described previously (Jardetzky 1984, Altman & Jardetzky 1986, Duncan *et al.* 1986, Jardetzky *et al.* 1986). In this paper, we summarize the essential features of PROTEAN and its performance on cytochrome-b562.

Cytochrome-b562 is a single chain protein composed of 103 amino acid residues. Its crystal structure is known (Lederer *et al.* 1981). It contains 4 alpha-helices forming a bundle and connecting coil elements. We generated artificial NOE data from the crystal structure of cytochrome-b562 and used PROTEAN to reconstruct the original crystal structure using these data. This procedure represents a further validation of the method and extends the validation previously reported by Lichtarge *et al.* (1987) on sperm whale myoglobin to the atomic level.

2.1. *Generate test data*

Hydrogens were added to the crystal structure determined by Lederer *et al.* (1981), using a modified version of the ADDHYDROGEN program in the MIDAS package (Jarvis *et al.* 1986). Distance constraints were generated by finding all pairs of hydrogen atoms in the protein that were less than 4 Angstroms apart. This would constitute the best data set that could be obtained by NMR. The heme group was not included in this calculation, although PROTEAN has successfully placed heme groups in test structures such as myoglobin (Lichtarge *et al.* 1987). There are 1600 atoms in cytochrome-b562, and $1600^2 - 1600/2$ possible inter-atomic distances. Using a 4 Angstrom cutoff, we generated only 729 NOES to be considered in addition to the covalent distances, and so the problem remains underspecified. The amino acid sequence and constraint network of b562 are given in Fig. 1.

2.2. *Define secondary structure within the backbone*

In an initial processing step, NMR data can be used to identify the approximate locations of helical and beta-strand elements. This greatly reduces the computational complexity of the problem and is justified because essentially non-overlapping subsets of NMR data are used for the determination of secondary and tertiary structures, respectively. Furthermore, the secondary structure can be defined on the basis of qualitative arguments and symbolic reasoning alone, whereas the definition of tertiary structure is based on numerical computation. Details of the analysis of NMR data to infer the position of helical and beta-

2	3	6	7	8	9	10	11	12	13	14	15	16	17	18
D	L	D	M	Q	T	L	N	D	N	L	K	V	I	E

19	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
K	K	A	N	D	A	A	L	V	K	M	R	A	A	A	L

39	40	41	42	56	57	58	59	60	61	62	63	64	65	66
N	A	Q	K	K	D	F	R	H	G	F	D	I	L	V

67	68	69	70	71	72	73	74	75	76	77	78	82	83	84	85
E	G	I	D	D	A	L	K	L	A	N	E	K	E	A	Q

86	87	88	89	90	91	92	93	94	95	96	97	98	99	101
A	A	A	E	Q	L	K	T	T	R	N	A	Y	H	K

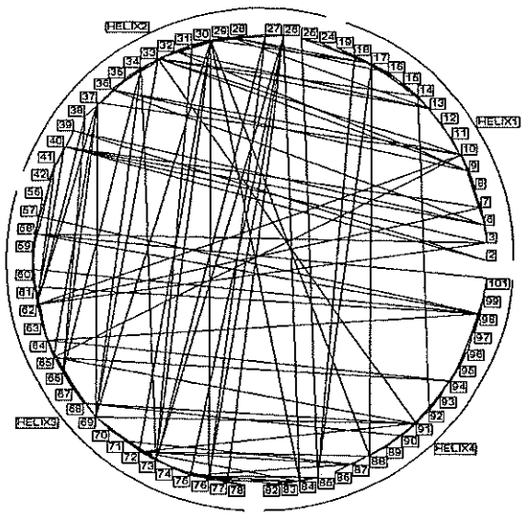


Fig. 1. Schematic diagram of the primary sequence of cytochrome-b562. The lines represent constraints between the amino acids.

strand elements have been reported elsewhere (Wüthrich 1986, Brugge *et al.* 1988). Briefly, secondary structures are recognized by the appearance of repeating patterns of distances between and among protons in the peptide backbone of the protein and proton exchange rates along the backbone. The identification of secondary structure has been done manually for a number of proteins, but has recently been automated in a system described elsewhere (Brugge 1987, Brugge *et al.* 1988).

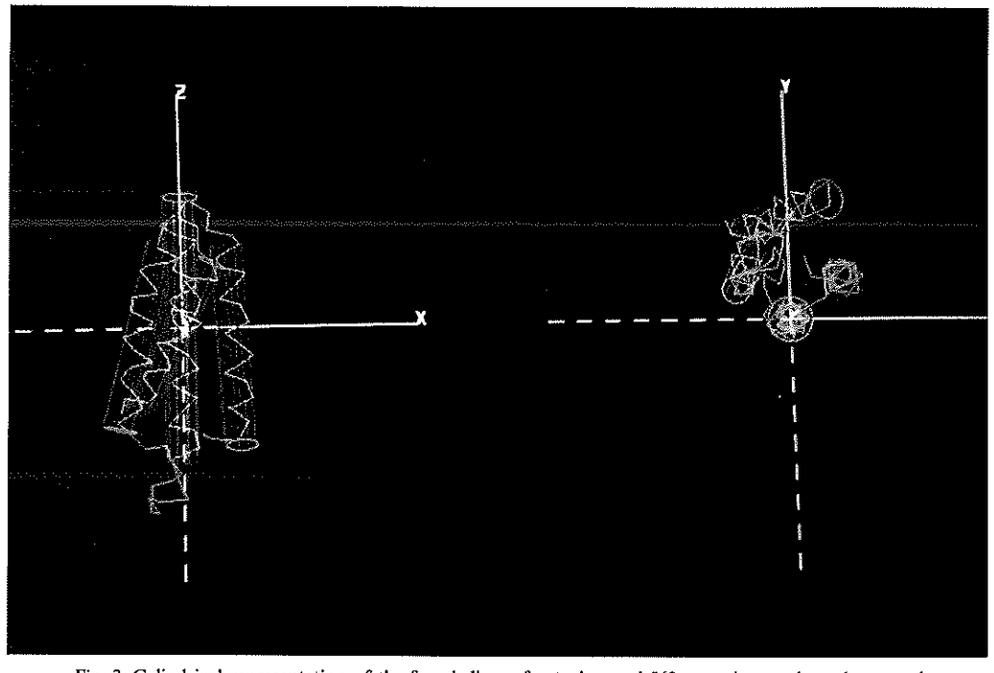


Fig. 2. Cylindrical representation of the four helices of cytochrome-b562, superimposed on the crystal structure. The two images are two views of the same helices rotated 90 degrees.

In this case, we used the same helical elements as found in the crystal structure in order to facilitate comparison of our final structure with the crystal structure. These correspond to amino acids 2-19, 24-42, 56-80, and 99-102.

2.3. *Represent the secondary structures as rigid objects*

In searching for the allowed tertiary structures, PROTEAN positions objects roughly with coarse sampling to define an initial region of occupancy and later refines the sample grid for more precise location of structures. For the initial coarse sampling, we model the known secondary structures as having ideal backbone phi/psi ($-57/-47$ for helices, $113/-119$ for parallel beta-strands and $135/-139$ for anti-parallel beta-strands) angles as a default. If there are data indicating that there are bends in the ideal backbone, then we can change the default. Nevertheless, we are able to use *some* default to produce a local coordinate system for each secondary structure in which the approximate position of all atoms can be specified. For large, regular secondary structures the coordinate

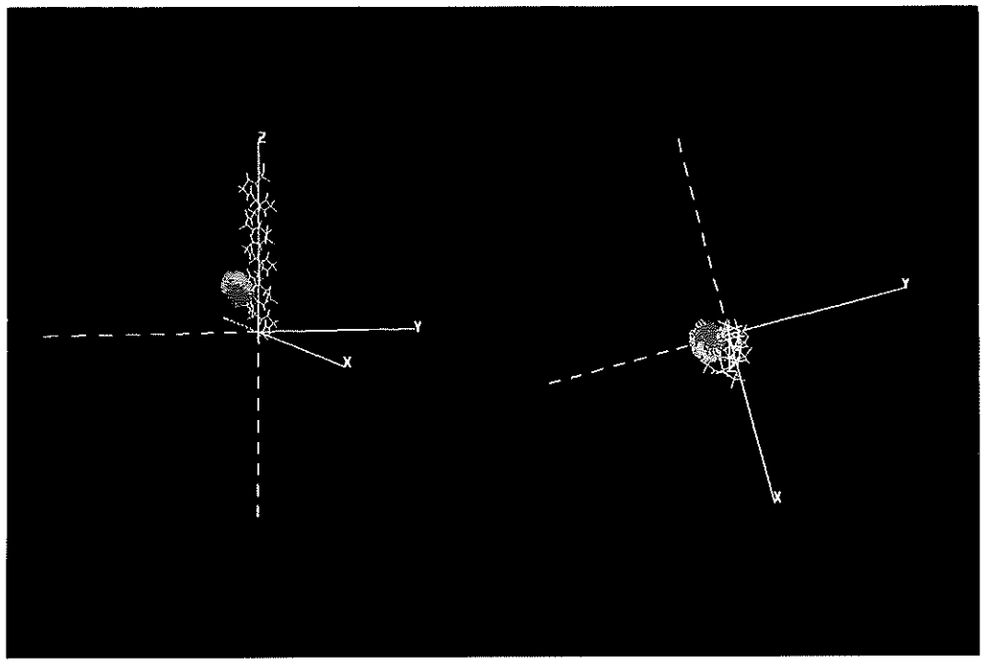


Fig. 3. Sidechain accessible volume of leucine 30 relative to helix 2. The orange indicates the accessible volume without taking into account constraints between the sidechains in the helix. The green indicates the accessible volume after taking these constraints into account.

system can contain many amino acids. For irregular or unpredictable secondary structures, a coordinate system must be defined for each amino acid separately.

For cytochrome-b562, each of the four helices was defined in a local coordinate system, and the approximate location of all backbone atoms was generated by assuming the standard phi/psi angles of $-57/-47$. For amino acids that were not part of regular secondary structure, a separate coordinate system was created centered on the fixed peptide backbone. The abstraction of the b562 structure to the rigid solid representation is shown in Fig. 2.

Having defined these coordinate systems, we are able to position each of the structures in space by simply specifying a transform (translation and rotation) relating the local coordinate system of the secondary structure to an arbitrary global coordinate system.

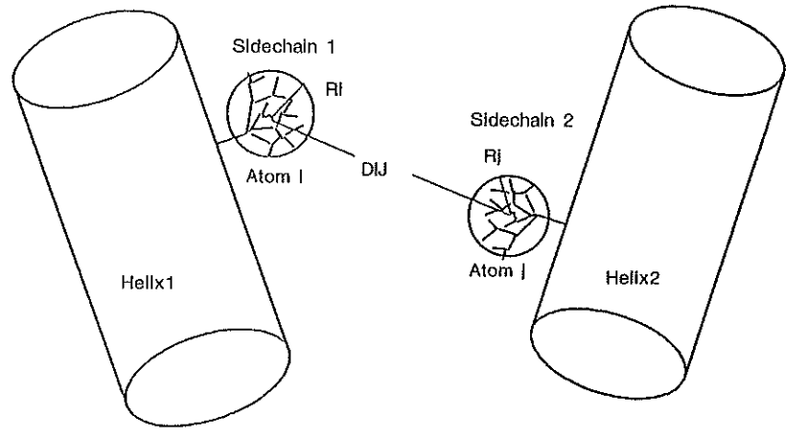


Fig. 4. Abstract representation of constraints between sidechains on different secondary structures. The positions of an atom in the accessible volume of a sidechain is represented as a sphere which encloses all the positions.

2.4 *Determine the initial sidechain accessible volumes for sidechains*

The *accessible volume* for a sidechain is first defined as the set of conformations that are compatible with the van der Waals forces between and among the atoms within the sidechain. The set of complete sidechain conformations has been precalculated and stored in a library. Each non-ordered, rotatable bond is sampled every 10 degrees to enumerate the possible side chain conformations. The number of legal conformations ranges from 1 (for Glycine) to 243 (for Arginine).

2.5. *Refine sidechains in context of specific amino acid sequence*

The sidechain conformations can be placed in the local coordinate systems defined around each secondary structure by transforming them so that the fixed alpha-carbon, beta-carbon and amide-nitrogen of the amino acid are coincident with their ideal backbone locations in the secondary structure. Certain sidechain conformations that are stored in the library will violate the van der Waals radii of fixed backbone atoms or other side chain atoms, and so must be excluded. In addition, there may be inter-sidechain NMR constraints which can be tested and used to exclude incompatible sidechain conformations. Thus, the refined accessible volume for a given side chain in a given secondary structure will be smaller than that stored in the library.

The result of the complete set of these calculations are sets of secondary

TABLE I

Cytochrome-b562: Numbers of helix locations at three stages of processing

Stage	Helix 1	Helix 2	Helix 4
Initial	3115	119	265
Intermediate	1874	117	153
Final	254	3	3

structural elements with sidechain conformations reduced to those that are compatible with internal NOE constraints and van der Waals restrictions, as shown in Fig. 3.

2.6. *Approximate the constraints between sidechains in different secondary structures*

Having sampled the accessible volumes for each sidechain, we have a large number of possible locations for each atom within a sidechain. It would be too expensive to consider *all* these possible locations during initial positioning. Therefore, the location of each atom relative to the local coordinate system of the secondary structure to which it belongs is summarized as a central point of the sampled locations and a radius describing the largest distance of any location to the central point. This radius is, therefore, a measure of uncertainty in the position of the atom. These summaries can be used for coarse definition of the locations of the secondary structures (see Fig. 4). For instance, if atom i and atom j are known to be at a distance of 7 Angstroms or less, then a conformation of the protein can be tested by calculating the distance between the two central locations, d_{ij}. If $d_{ij}+r_i+r_j$ is less than 7 (where r_i and r_j are the radii of the spheres enclosing the atoms), then the conformation satisfies the constraint, otherwise it is excluded. In this way, atomic constraints can be summarized and checked inexpensively in order to reduce the number of locations before making computationally expensive tests – which involve actually testing each pair of precise locations.

2.7. *Determine accessible volumes for each secondary structure using the constraints defined in 2.5.*

Once the spatial distribution of each atom relative to its secondary structure is known, we can characterize the locations of the secondary structures themselves relative to one another. In the cytochrome-b562 example, Helix-3 residues 53-80

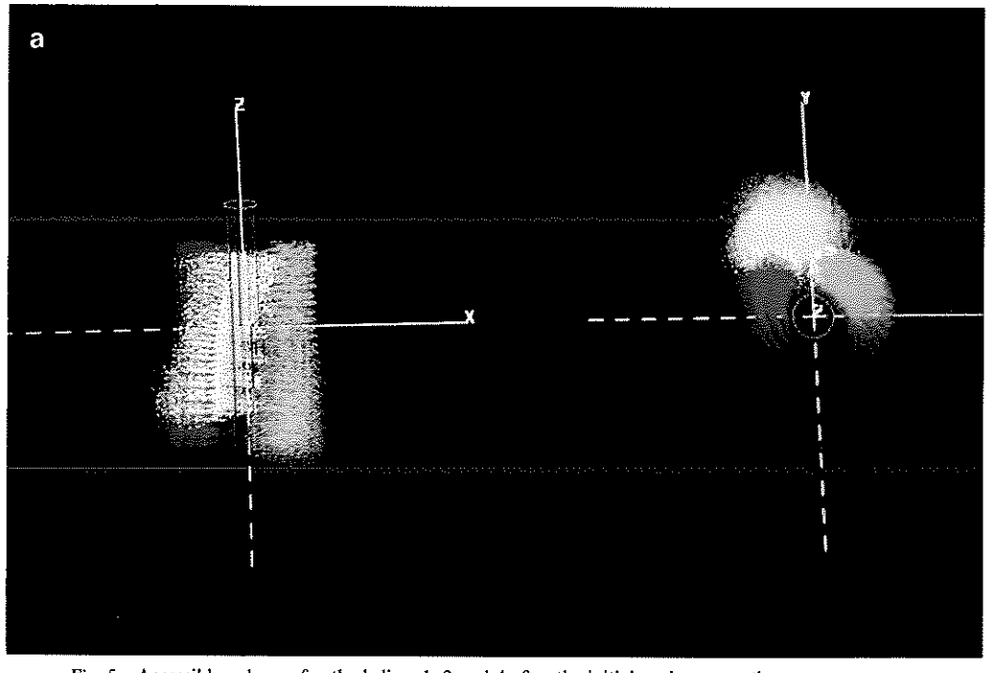


Fig. 5a. Accessible volumes for the helices 1, 2 and 4 after the initial anchor operation.

were arbitrarily chosen as the center of a fixed cartesian coordinate system which we call the anchor. The position of other helices relative to Helix-3 can then be described as a position (x y z) and 3 Euler angles denoting orientation (φ ψ ω). These 6 parameters specify a transformation in the space of the fixed structure (the anchor) which positions the object centered coordinate of the movable structure (the anchoree) – and therefore also positions the constituent atoms of the anchoree. These positions can be systematically sampled by varying the 6 parameters, and checked for compatibility with the constraints. This procedure is referred to as *anchoring*.

Helices 1, 2 and 4 were sampled at a positional resolution of 1 Angstrom and an orientational resolution of 10 degrees. Each sampled point was checked with respect to its ability to satisfy all the NOE constraints between the helices (as described in section 2.6) as well as a rough van der Waals check that prevent cylinders enclosing the peptide backbones from overlapping (more strict van der Waals exclusion criteria are used in the next section). Helix-1 has the least

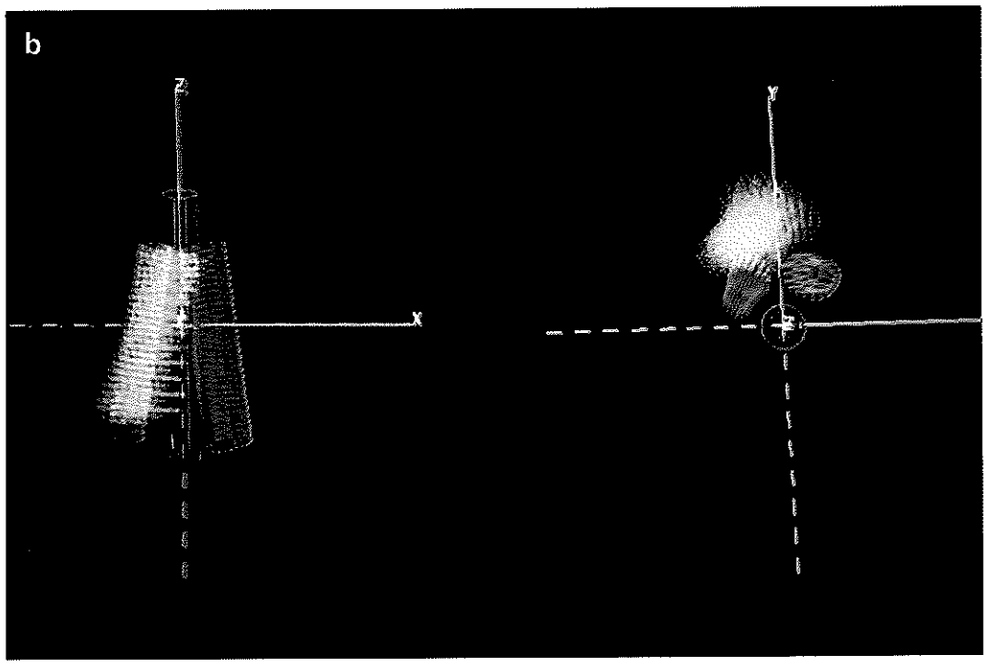


Fig. 5b. Final accessible volumes for helices 1, 2 and 4.

constraints with Helix-3 and so had the largest "accessible volume". The size of the initial location lists is given in the first row of Table I and is graphically represented in Fig. 5a.

At this point the constraints between the mobile helices could be used to further exclude some locations, a procedure called *yoking*. If a single location in the location list for Helix-1 was incompatible with all locations for Helix-2, then that location was excluded. Thus, we applied all the constraints to pairs of helices to obtain the smallest possible lists of locations for each object, as shown in the second row of Table I.

2.8. *Refine the accessible volumes for each secondary structure using the detailed sidechain conformations defined in 2.5.*

The representation of atomic positions as mean and radius, as described in 2.6, is useful for initial location of secondary structures, but can be made more precise. For each helix, we can enumerate all the possible locations for the atoms

(as sampled in section 2.5), and test these detailed locations for compatibility with the distance constraints to atoms on other helices. In addition, we made a more refined van der Waals interaction check which tests the positions of all backbone atoms, and not simply an enclosing volume. For cytochrome-b562, the size of these pruned location lists for the helices is shown in the third row of Table I and graphically in Fig. 5b.

2.9. *Select coherent sets of secondary structures to define coherent conformations*
Having defined the set of accessible volumes for each helix, we want the set of helix topologies that simultaneously satisfy all the constraints. These can be generated by checking every combination of the four helices (sampled from their respective accessible volumes), and checking for compatibility with the constraints. For cytochrome-b562, we had a total of $3 \times 3 \times 1 \times 254 = 2286$ possible conformations to consider (the cross product of all locations in the helix accessible volumes as shown in Table I). Of these, only 964 were found to be internally consistent. We refer to these as coherent instances, and they define the range of helix packing topologies consistent with the data. A representative set of 10 of these coherent instances was then selected for further processing.

2.10. *Iterate steps 6 through 8 for coils*
Given the 10 coherent instances for the helices, the next step is to determine the set of positions for the coil amino acids for each of them. In a procedure quite analogous to that outlined for placement of the helices in section 2.6 through 2.9, we define the initial coarse accessible volume for each coil coordinate system, refine the accessible volume with detailed tests of sidechain conformations, and then generate coherent sets of coil locations for each set of helices.

In the current case, the result of the previous operations was a complete set of 10 coherent instances with a single location for each of the helix coordinate systems. These represent a systematic sample of the conformations of the protein allowed by the initial data.

2.11. *Generate consistent backbone traces for each coherent conformation*
As outlined in section 2.3, we divided the molecule into independent coordinate systems for purposes of efficiency in determining their location. Since we discretely sample space at an interval of 1 Angstrom (for translations) and 10 degrees (for orientation), the final coherent instances may not accurately match the covalently connected polypeptide backbone due to the imprecision of sampling.

TABLE II

Cytochrome-b562: Best thread RMS deviations of C-alphas from crystal C-alphas (angstroms)

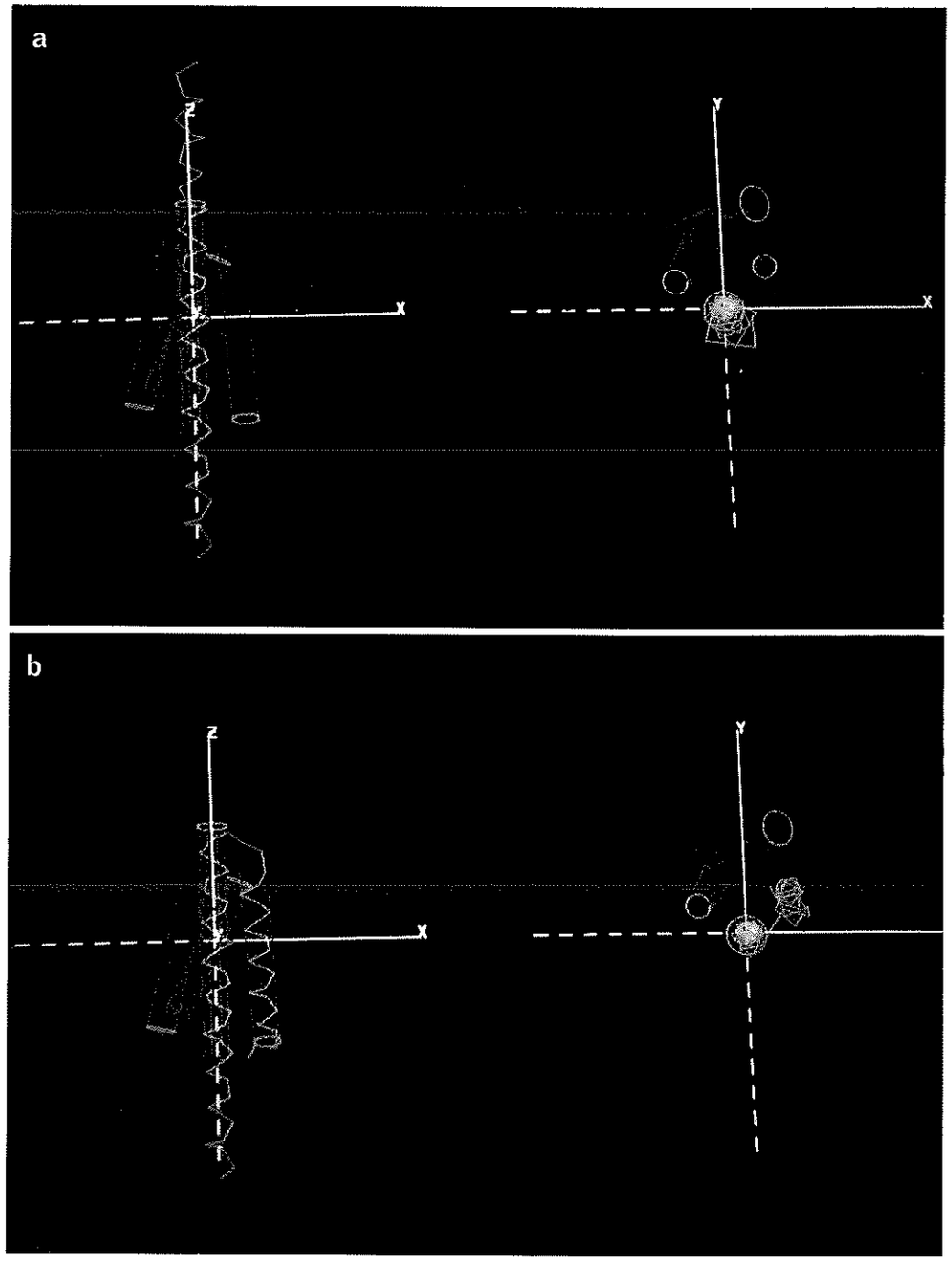
Instance	With Coils	Without Coils
1	3.2	2.7
2	4.1	2.4
3	5.4	3.7
4	4.2	3.0
5	3.7	2.9
6	3.5	2.5
7	3.3	2.5
8	4.1	2.7
9	4.4	2.8
10	5.1	2.6

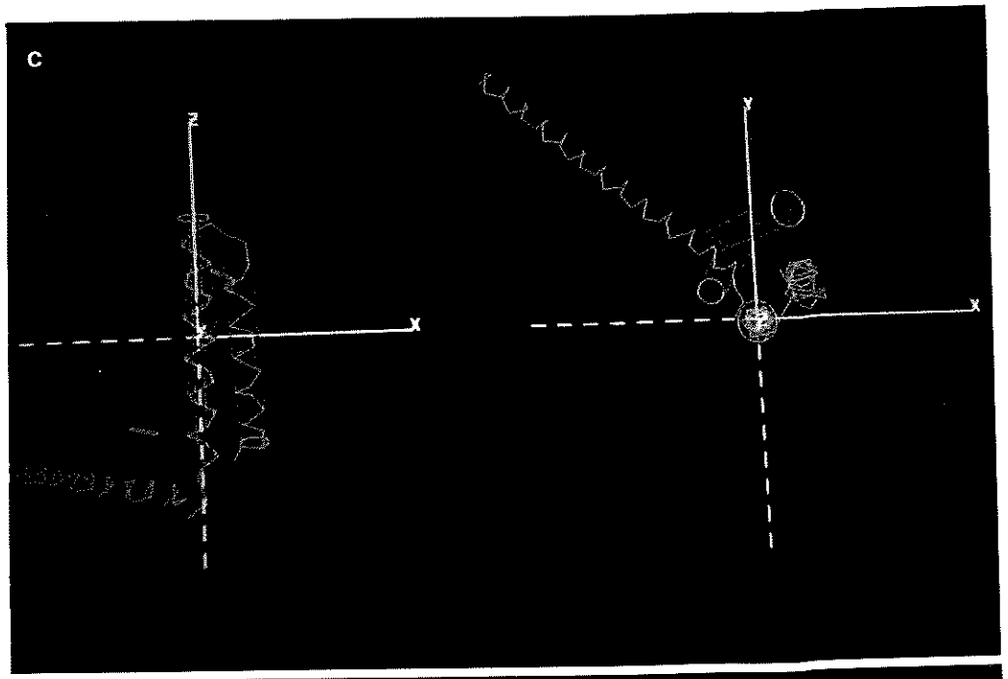
In order to make small adjustments to the backbone, and remove these inaccuracies, we employ a procedure called *atomic threading*. We first generate a continuous polypeptide of the same length as the coherent instance to be threaded, with the phi/psi angles of the secondary structure regions set to proper default values, and the phi/psi angles of the intervening coils set to random values within the legal Ramachandran range of an extended beta-strand. We then superimpose the polypeptide backbone over the fixed anchor, Helix-3, using a least squares distance between the atoms on the backbone and the positions in the ideal Helix-3. Next we vary the phi/psi angles along the polypeptide towards the C and N terminus by performing a linear best fit optimization of the backbone peptide atoms to the positions of atoms within the coherent instance. It is important to underscore that we only adjust local phi/psi angles to reduce local errors, so that there is *no* global optimization of the entire structure at any stage of the structure determination. Since the coherent instance has been defined using the full set of constraints, it is important that our optimization bring the polypeptide backbone thread as close to the target coherent instance as possible.

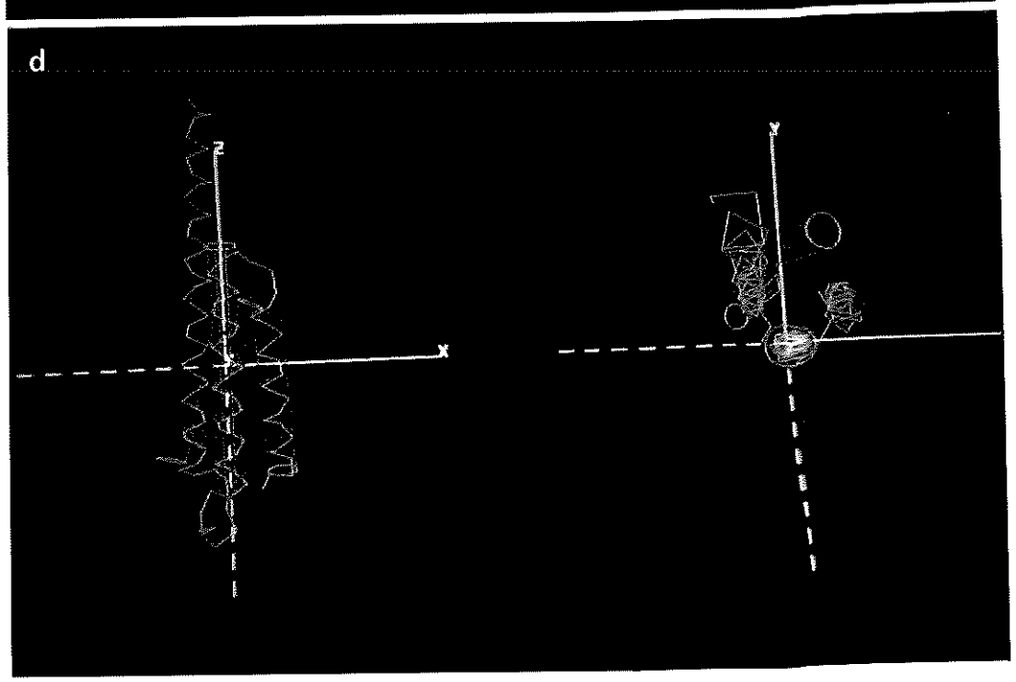
The threading procedure is illustrated in Fig. 6 which shows successive stages in threading to the best fit coherent instance of the crystal (Fig. 2).

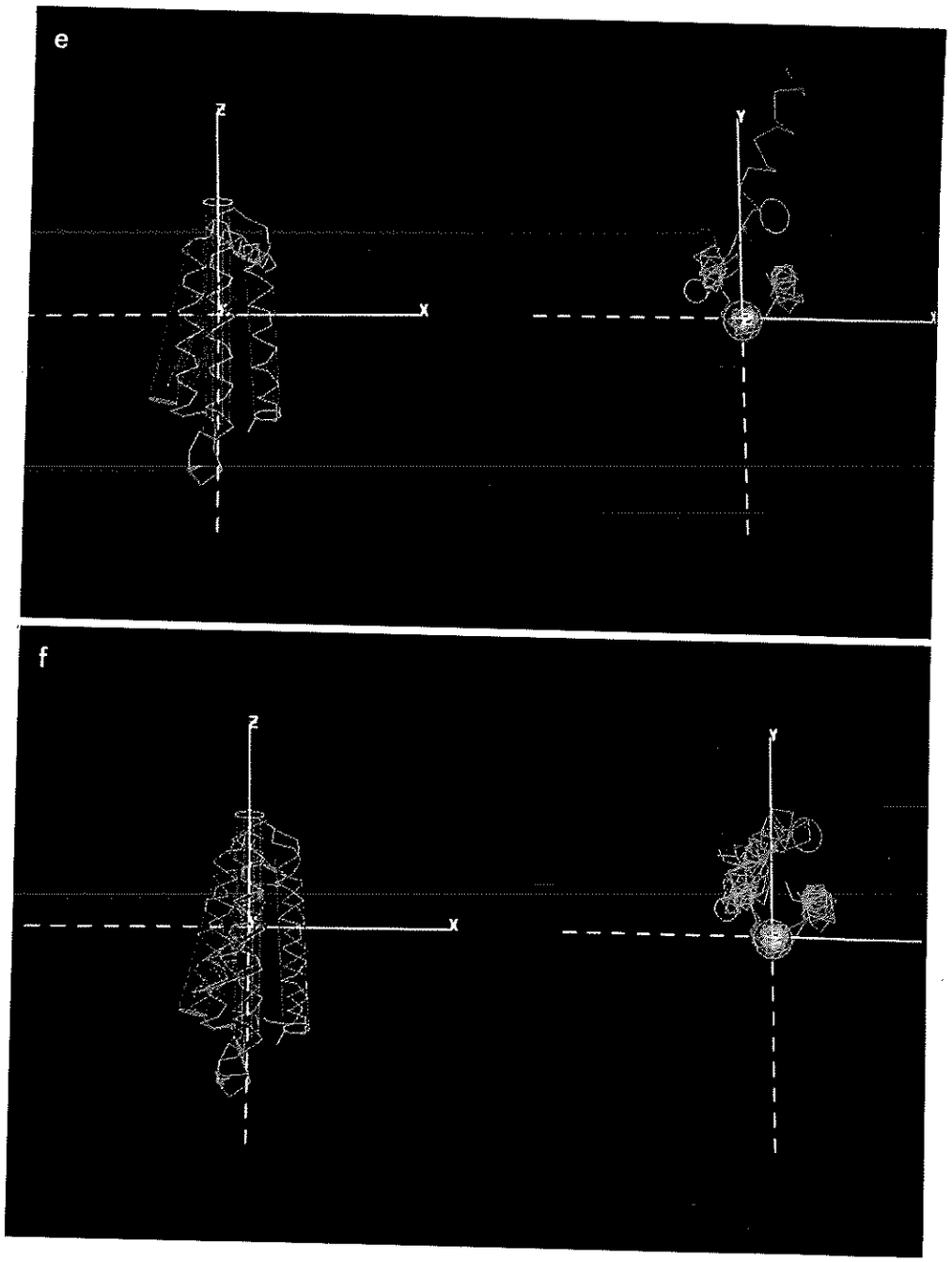
For each of the 10 coherent instances, a series of approximately 30 threads was obtained using different random starting values for the coil phi/psi angles. These threads were graphically displayed, and those that did not have significant

Fig. 6a-f. Successive stages in the threading of a covalently legal backbone to a coherent instance.


a
Z
X
Y
X
b
Z
X
Y


c
Z
X
Y
X


d
Z
X
Y


e
Z
X
Y
f
Z
X
Y

van der Waals violations were manually selected. Of these, the threaded backbone that had the best RMS fit to the coherent instance was chosen as the best thread.

The final result of threading is shown in Fig. 7. Fig. 7a shows a superposition of 38 threads which satisfy all constraints. Fig. 7b depicts the 10 best threaded backbones superimposed such that the RMS deviation between crystal c-alphas and threaded backbone c-alphas was minimized. The average RMS difference between the threaded structures and the crystal structure was 4.13 Angstroms when all c-alphas were considered (range 3.25 to 5.42) and 2.8 Angstroms when only helix c-alphas were considered (range 2.42 to 3.74) (see Table II).

3. DISCUSSION

The final set of structures must be interpreted with several considerations in mind. First, in demonstrating this method we have only used the covalent connectivity constraints within a polypeptide, van der Waals constraints, and a set of distance ranges representing NOEs generated from the crystal structure. We have shown elsewhere that other constraints can be used to exclude some of the remaining conformations: these include experimental constraints on shape, volume, surface and buried amino acids (Altman & Jardetzky 1986). Since we have not used these here, our reported spatial distributions for cytochrome-b562 represent, in one sense, intermediate results. We have also not used theoretical constraints on the preferred conformations of sidechains and backbone structures (except for Ramachandran compatibility). However, separate evaluation functions can be constructed to test these constraints and modify the best estimate of the spatial distribution.

Second, these results do not include NOE constraints to the coils, although the thread procedure (together with manual selection of van der Waals legal structures) does include covalent and van der Waals constraints. However, the superimposed structures in Fig. 7b show that tight placement of the helices provides a strong constraint on the coils, and that the random selection of coil phi/psi angles provides a clear picture of the remaining coil accessible volume. As additional coil constraints are added, the size of the coil accessible volume should decrease.

As more and more distance data are derived from the crystal structure, we can more severely limit the spatial distributions until we converge on a set of delta functions for most atoms. However, in the interpretation of actual solution data, this is an unlikely situation, and so we have focussed our research on useful

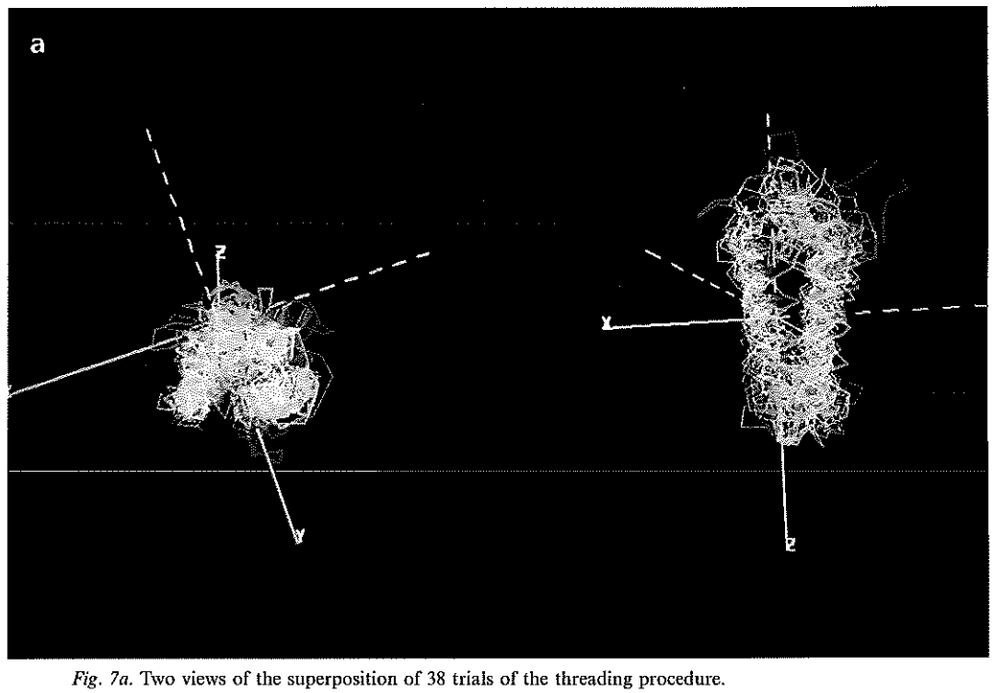


Fig. 7a. Two views of the superposition of 38 trials of the threading procedure.

representation of such apparently "intermediate" results. Recent results with repeated runs of other structure determination methods has shown that the *variation* from mean position is at least as important as the mean position itself. This is especially true with regard to solution data that are taken over a population of structures that may have important dynamic elements. The variation then defines the volume within which movement is likely to be occurring, although it does not further define the movement itself.

The exclusion method has been used in other applications of artificial intelligence methods to data interpretation. One of the oldest, and most similar, was the DENDRAL program (Lindsay *et al.* 1980) in which mass spectrometry (and other) data were used to determine structures of organic molecules. As in the present problem, the initial space of possible structures was too vast for exhaustive enumeration followed by testing single structures for exclusion. Instead, large classes of structures were excluded with a single test before enumeration of the individual instances, just as PROTEAN does, for instance, in excluding positions

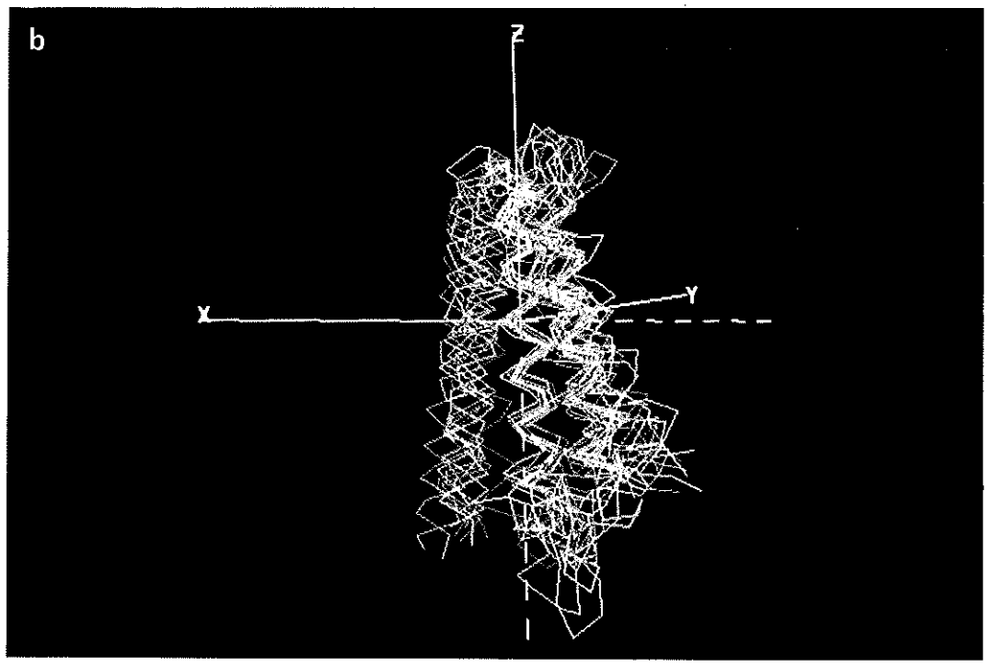


Fig. 7b. Superposition of the best 10 threaded backbones such that the RMS deviation between each backbone and the crystal structure is minimized.

of helices at a coarse level of description before describing the positions of their component atoms. Perhaps the single most important benefit of the exclusion method is the assurance that every possible structure has been considered – either explicitly or as an implicit member of a class considered explicitly – and that all structures are either included in the output or have been excluded for an explicit reason.

The results in this paper represent a successful use of the exclusion paradigm for the systematic definition of the entire family of protein structures that are compatible with a given data set. The exclusion paradigm relies on the use of compact data structures to represent initially large sets of conformations, but allows representations to become finer as the number of representations decreases. The chief technical obstacle in the development of these methods remains the large size of the search space – with its potential for combinatorial explosion – especially when a sample-based method is used. For this reason we are investigating the possibility of more parametric descriptions of spatial distribution.

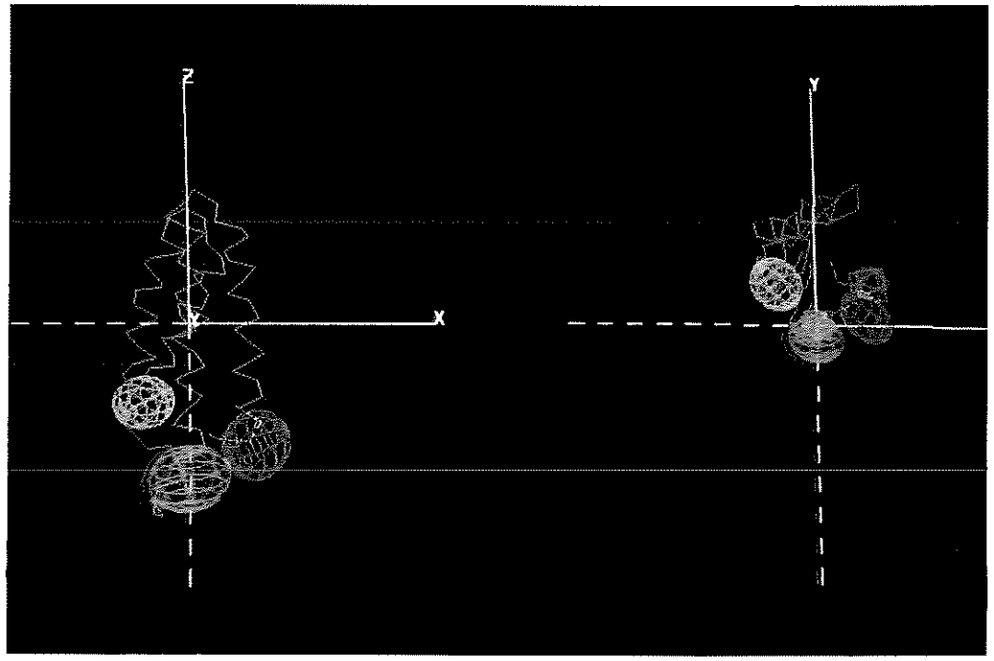


Fig. 8. Uncertainty in the positioning of the random coil residues Ala43, Leu48 and Met55 in cytochrome-b562, using the NOE constraints generated from the crystal structure and the probabilistic formulation described below.

Instead of sampling space discretely, it is possible to represent the accessible volume of an object as a parameterized distribution. For example, if we want to model all accessible volumes as three-dimensional Gaussian distributions, we can represent the position of each atom with a two-part *state representation*. The first part is a vector of an atom's mean position:

$$\bar{x} = (\hat{x}\ \hat{y}\ \hat{z})$$

The second part is a covariance matrix describing the uncertainty of the variables and the correlation between them in the position vector:

$$C(\bar{x}) = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix}$$

The accessible volume of an object can be drawn, using the mean position and

covariance, as an ellipsoid representing equiprobability contours. The contours drawn in this paper, with the sampled method, are essentially 99% confidence contours. The ellipsoid would be centered at mean position, x, and would have a shape and orientation determined by the covariance matrix as described in Smith & Cheeseman (1986). An illustration is given in Fig. 8 (see color plate).

There have been a number of reports of efficient methods for representing uncertain spatial locations for a large number of objects using this formalism. Methods based on the Kalman filter (Gelb 1984, Smith *et al.* 1986) can be used to update these state estimates by introducing new data. The key assumption in this particular approach is that the distributions that are being modelled are well-described by their first two moments (mean and variance). We are currently further investigating the suitability of these approaches in the context of PROTEAN.

ACKNOWLEDGMENTS

We would like to acknowledge support from NIH grants RR002300 and RR00785, NSF grant DMB-8402348, NASA-Ames Contract NCC2-274 and Boeing Computer Services Contract W271799.

REFERENCES

Altman, R. A. & Jardetzky, O. (1986) New strategies for the determination of macromolecular structure in solution. *J. Biochem. 100,* 1403–23.

Braun, W. & Gō, N. (1985) Calculation of protein conformations by proton-proton distance constraints: A new efficient algorithm. *J. Mol. Biol. 186,* 611–26.

Brugge, J. A. (1987) *ABC: A Knowledge-Based System for Determining Structural Components of Proteins.* M.S. Thesis, Computer Science Department, Stanford University, Stanford, CA.

Brugge, J. A., Buchanan, B. G. & Jardetzky, O. (1988) Toward automating the process of determining polypeptide secondary structure from ^{1}H NMR data. *J. Comput. Chem.* (in press).

Clore, G., Brünger, A., Karplus, M. & Gronenborn, A. (1986) Application of molecular dynamics with interproton distance restraints to three-dimensional protein structure determination: A model study of Crambin. *J. Mol. Biol. 191,* 523–51.

Duncan, B., Buchanan, B. G., Hayes-Roth, B., Lichtarge, O., Altman, R., Brinkley, J., Hewett, M., Cornelius, C. & Jardetzky, O. (1986) PROTEAN: A new method of deriving solution structures of proteins. *Bull. Mag. Res. 8,* 111–9.

Frayman, F. (1985) *PROTO: An Approach for Determining Protein Structures from Nuclear Magnetic Resonance data: An Excercise in Large Scale Interdependent Constraint-Satisfaction.* Ph.D. Thesis, Computer Science Department, Northwestern University, Evanston, Illinois.

Gariepy, J., Lane, A., Frayman, F., Wilbur, D., Robien, W., Schoolnik, G. K. & Jardetzky, O. (1986) Structure of the toxic domain of the Escherichia coli heat-stable enterotoxin ST 1. *Biochemistry 25,* 7854–66.

Gelb, A. (1984) *Applied Optimal Estimation.* M.I.T. Press.

Havel, T. & Wüthrich, K. (1984) A distance geometry program for determining the structures of small

proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular H-H proximities in solution. *Bull. Math. Biol. 46:4,* 673–698.

Jardetzky, O. (1984) A method for the definition of the solution structure of proteins from NMR and other physical measurements: The Lac-Repressor headpiece. In: *Progress in Bioorganic Chemistry and Molecular Biology,* ed. Yu. A. Ovchinnikov, pp. 55–63, Elsevier Science Publishers B.V., Amsterdam.

Jardetzky, O., Lane, A., Lefèvre, J-F., Lichtarge, O., Hayes-Roth, B. & Buchanan, B. (1986) Determination of macromolecular structure and dynamics by NMR. In: *NMR in the Life Sciences,* eds. Bradbury, E. M. & Nicolini, C., pp. 49–72, Plenum Press, New York.

Jardetzky, O. & Roberts, G. C. K. (1981) *NMR in Molecular Biology.* Academic Press, New York.

Jarvis, L., Huang, C., Ferrin, T. & Langridge, R. (1986) *UCSF MIDAS Molecular Interactive Display and Simulation: User's Manual.* University of California, San Francisco.

Lederer, F., Glatigny, A., Bethge, P., Bellamy, H. & Mathews, F. (1981) Improvement of the 2.5 Angstroms resolution model of cytochrome-b562 by redetermining the primary structure and using molecular graphics. *J. Mol. Biol. 148,* 427.

Lefèvre, J-F., Lane, A. N. & Jardetzky, O. (1987) Solution structure of the trp operator of Escherichia coli determined by NMR. *Biochemistry* (in press).

Levitt, M. (1983a) Molecular dynamics of native protein I. Computer simulation of trajectories. *J. Mol. Biol. 168,* 595–620.

Levitt, M. (1983b) Molecular dynamics of native protein II. Analysis and nature of motion. *J. Mol. Biol. 168,* 621–57.

Lichtarge, O., Cornelius, C. W., Buchanan, B. G. & Jardetzky, O. (1987) Validation of the first step of the heuristic refinement method for the derivation of solution structures of proteins from NMR data. *PROTEINS Structure, Function and Genetics 2,* 340–358.

Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A. & Lederberg, J. (1980) *Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project.* McGraw-Hill, New York.

Madrid, M. & Jardetzky, O. (1988) *Biochim. Biophys. Acta* (in press).

Smith, R. & Cheeseman, P. (1986) On the representation of spatial uncertainty. *Int. J. Robot 5,* 4.

Smith, R., Self, M. & Cheeseman, P. (1986) Estimating uncertain spatial relationships in robotics. *Proceedings of the Second Workshop on Uncertainty in AI.* Philadelphia, August, 1986.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acid.* John Wiley & Sons, New York.

Wüthrich, K., Billeter, M. & Braun, W. (1984) Polypeptide secondary structure determination by nuclear magnetic resonance observation of short proton-proton distances. *J. Mol. Biol. 180,* 715–40.