Acta Cryst. (1980). A**36**, 211–219

A Model-Building Procedure with Particular Application to Proteins

BY SUSAN FITZWATER AND HAROLD A. SCHERAGA

Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853, *USA*

(*Received* 29 *November* 1978; *accepted* 10 *September* 1979)

Abstract

A procedure for fitting a molecular model with fixed bond lengths and bond angles to a set of Cartesian coordinates has been developed. This procedure is designed to fit coordinates obtained from an X-ray electron-density map, although it can also fit coordinate sets from other sources. It differs from other model-building methods which use fixed bond lengths and bond angles in that it takes cognizance of certain information present in the electron-density map, which is ignored in other methods. The inherent uncertainty in the values of the coordinates determined from the map influences the fitting; *i.e.*, if the computed location of an atom of the model is within the uncertainty assigned to the corresponding coordinate derived from the map, the fit of the atom of the model is assigned as exact. Also, the fitting process may begin anywhere along the molecular chain, in contrast to other procedures which commence fitting at or near one end of the chain. Thus, the present procedure can avoid errors that might arise if the starting point were chosen in a region where the atoms are poorly defined, as they often are at the ends of the chain. The procedure is designed to be used with a large amount of operator intervention, making it fairly flexible. Complete mathematical details of the method are given. A Fortran IV computer program using the method to fit a polypeptide model to a set of Cartesian coordinates has been written. The program has been used to fit a model of bovine pancreatic trypsin inhibitor to Cartesian coordinates derived from the 2·5 Å resolution electron-density map. The r.m.s. deviation between the model and the weighted coordinates from the map was 0·49 Å. As a preliminary step in a refinement of the 2·5 Å structure by potential-energy-constrained model building, the model obtained here was subjected to energy minimization with the atomic coordinates constrained to remain 'close' to the original guide points.

Introduction

The utility of model-building procedures (methods for fitting a chain with fixed bond lengths and possibly fixed bond angles to arbitrary Cartesian coordinates) has been recognized for some years (Diamond, 1966). The most extensive use of these has been made in the refinement of X-ray structures of proteins, where the electron-density map is the source of the Cartesian coordinates to which the model is to be fitted (Diamond, 1971; Epp *et al.*, 1974; Huber *et al.*, 1974; Deisenhofer & Steigemann, 1975). The use of a model, constrained to be stereochemically reasonable to some degree, is now a standard feature in both real- and reciprocal-space refinements; a simple, powerful model-building procedure is a necessary tool for protein X-ray crystallography. Model-building procedures also have application outside the field of structural refinement; for example, calculational studies of enzyme–substrate interactions (Platzer, Momany & Scheraga, 1972; Pincus, Zimmerman & Scheraga, 1976, 1977; Pincus & Scheraga, 1979) may require a different standard geometry (set of fixed bond lengths and bond angles) for use in energy-computation algorithms than the one used to obtain the X-ray structure from the electron-density map. In this case, model building can be used to provide a structure with bond lengths and bond angles that are compatible with the computational methods to be used.

Several model-building procedures for proteins have been developed in recent years. The first and most widely used of these is the procedure of Diamond (1966). Another is that of Warme, Gō & Scheraga (1972) (hereafter referred to as WGS), which is not as versatile or as powerful as Diamond's (1966) method (since the latter uses a weighting scheme and can start at either end), but which is probably easier to use because it dispenses with techniques such as filtering (choosing the combination of parameters which 'best' improve the fit) and the use of probes (lengths of protein chain) of varying length which are designed to make the basic least-squares minimizer operate more efficiently. However, the WGS procedure appears to require more operator intervention than does the Diamond procedure. The Hermans–McQueen (1974) method of local change of atomic coordinates can be applied to protein model building, but use of this method in a model-building scheme which assumes fixed bond lengths and bond angles could require much

0567-7394/80/020211-09$01.00

adjustment of force constants to obtain fixed geometry. Several other methods which use *restraints* on bond lengths and bond angles have been developed in recent years (Dodson, Isaacs & Rollett, 1976; Ten Eyck, Weaver & Matthews, 1976). We believe, however, that there are advantages to using a model-building procedure which assumes fixed bond lengths and bond angles, particularly in structure refinement; these advantages will be discussed in a future paper (Fitzwater & Scheraga, 1980). At the present time, the Diamond or WGS procedures have been used most widely for applications of model building to proteins.

This paper will describe a new model-building procedure, which remedies certain problems common to both the Diamond (1966) and WGS methods, employs a powerful minimizer (Dennis & Mei, 1975), and retains most of the ease of use of the WGS procedure. In practice, the new method requires a certain amount of operator intervention in decision-making processes, but the intervention is largely routine and, in some cases, enables the improvement of the fitting to be greatly accelerated.

[We recognize that most of the problems with the Diamond (1966) model-building procedure discussed below are remedied by the Diamond (1971) real-space refinement procedure, in which a model with fixed bond lengths (but with certain bond angles variable) is fitted directly to the electron-density map. However, given the complexity and expense of real-space refinement, it is not anticipated that many investigators will find it preferable to simple model building in applications in which the latter will suffice. The procedure described here is designed for model building, *not refinement* (although certain elements of the method can be carried over into a refinement procedure, as will be discussed in future papers); therefore, it will be compared here *only* with other model-building methods.]

Although none of the model-building methods interact with the X-ray intensity data directly, all of them make indirect use of the data: they fit molecular models to atomic coordinates [Diamond's (1966) guide points] derived from the electron-density map which was computed from the intensity data. The major advantage of the new method is that it uses certain information present in the electron-density map which is ignored by the others. Even a low-resolution map offers some idea of the accuracy with which a given atom or group can be located. Therefore, as is true in certain structure-refinement/model-building procedures (Waser, 1963; Dodson, Isaacs & Rollet, 1976; Jack, 1977; Sussman, Holbrook, Church & Kim, 1977), the method provides for each guide point to be weighted. The user may insure that atomic positions which are well defined in the map will have a greater influence on the overall fit than those which are less well defined; poorly defined atoms may have no influence at all. In this respect, the new method is a definite improvement over the WGS procedure, in which all atomic positions are weighted equally, and provides more flexibility than the Diamond (1966) method, in which all positions have a weighting factor of zero or one. The choice of weighting scheme is left to the user, providing extensive control over the goodness of fit so that the user may produce a model which is best suited to his particular application. For example, the active site of an enzyme could be fitted very well at the expense of other regions of the molecule, which might be desirable for enzyme–substrate studies. Another feature of protein electron-density maps arises from limited resolution; the peaks, even those with a fairly high electron density, are diffuse enough so that one cannot locate an atom at a point but can locate it only within a circle whose center is in the area of the density peak. Therefore, our procedure allows for some play in the fit of the computed to the experimental positions, which, in this treatment, are necessarily points. If the computed position is within a pre-assigned distance P of the experimental one, the procedure treats the fit as exact. The degree of peak diffusivity seen in the map will govern the choice of P; the weights may be taken from the values of the electron density at the atomic locations. The use of the 'play parameter' P may be a cause for concern in some cases because it introduces a discontinuity which may create difficulty in a minimization procedure. In our experience, this difficulty was not encountered even though we did observe several 'crossovers' from within the circle of radius P to without. If the problem does arise, the discontinuous function can be replaced by a steep continuous one that can be differentiated in the minimization procedure.

Finally, most electron-density maps show that the protein is better defined in the middle of the chain than it is at either end. However, the WGS procedure dictates that the fitting must begin at or near the N terminus (see also Rasse, Warme & Scheraga, 1974 and Swenson, Burgess & Scheraga, 1978) while the Diamond (1966) procedure uses either the N or C terminus as the starting point (a non-terminal starting point may be used with the Diamond method, but this entails some extra work). Since the choice of the starting point exerts considerable influence over the whole fitting process, it seems ill-advised to choose as a starting point an atom whose experimental position is poorly defined. Because the ends of the chain are often diffusely defined in the electron-density map, the use of a terminus as a starting point should generally be avoided. Therefore, our method allows the user to choose a starting point for the fitting anywhere along the chain, and to pursue the fitting toward both the N and C termini simultaneously.

Our procedure is described in detail below. A computer program implementing the procedure has been written. The program can be used alone when the sole objective of the investigator is model building. How-

ever, it is also designed to interlock with a program for the refinement of protein structures which is being developed at present.

Method

Given a set of Cartesian coordinates, a molecular model is fitted to these coordinates by the following procedure:

(1) a user-specified conformation is generated;

(2) the molecule is oriented in the unit cell according to user-specified values for translational distances along the x, y and z axes and for rigid-body rotations in the directions described by the Euler angles, α, β, and γ;

(3) the differences between the computed and the X-ray coordinates are taken, squared, weighted and summed;

(4) the sum is minimized with respect to the user-specified parameters (dihedral angles, translational distances and Euler angles).

In outline, the new method does not vary from either the Diamond or the WGS methods. The major differences between this method and the others lie in the generation of the molecular conformation from an arbitrary position in the chain and in the inclusion of estimates of the quality of the electron-density map in the function to be minimized. In order to focus attention on these differences, each step of the method will be discussed below.

The first step is the generation of the molecular conformation. For this purpose, a standard geometry is adopted from *UNICEPP* (United Atom Conformational Energy Program for Peptides; Dunfield, Burgess & Scheraga, 1978) and a zero-point model is generated by setting all φ's, ψ's, ω's, and χ's equal to arbitrary values, except that ω preceding proline equals 180° (or 0°, if it is *cis*) and φ of proline equals −75°. The zero-point model is then altered to the desired conformation (the fitting-program model) by setting the φ's, ψ's, χ's and ω's to user-specified values, using standard rotation-matrix procedures as given in *UNICEPP*.

The δ's (defined in the legend to Fig. 1) are applied as follows to generate the molecule in the desired conformation. The origin and direction of the bond spindle vector for rotation about each bond are chosen according to whether the rotation is carried out in the direction from the N to the C terminus or *vice versa*, as will be discussed below. The positions of only those atoms that are attached directly or indirectly to the end of the spindle vector are affected by rotation around the bond (see Fig. 2). The positions of the atoms whose coordinates are affected by rotation by an angle δ are given by

$$\begin{bmatrix} x_f + x_s \\ y_f + y_s \\ z_f + z_s \end{bmatrix} = \begin{bmatrix} \cos\delta + l_1^2(1-\cos\delta) & l_1 l_2(1-\cos\delta) - l_3\sin\delta & l_1 l_3(1-\cos\delta) + l_2\sin\delta \\ l_1 l_2(1-\cos\delta) + l_3\sin\delta & \cos\delta + l_2^2(1-\cos\delta) & l_2 l_3(1-\cos\delta) - l_1\sin\delta \\ l_1 l_3(1-\cos\delta) - l_2\sin\delta & l_2 l_3(1-\cos\delta) + l_1\sin\delta & \cos\delta + l_3^2(1-\cos\delta) \end{bmatrix} \begin{bmatrix} x_i - x_s \\ y_i - y_s \\ z_i - z_s \end{bmatrix}, \quad (1)$$

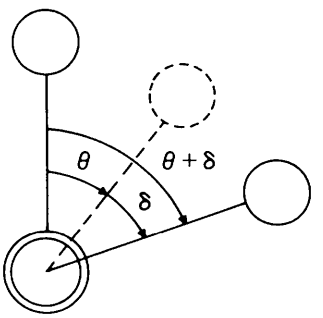

Fig. 1. Dihedral angle in the model. If θ is the value of the rotation applied about the bond in the zero-point model, and δ is the value of the rotation applied about the bond when the model is put into the desired conformation, the value of the model dihedral angle is $\theta + \delta$. By setting $\theta = 0$, the model dihedral angle becomes equal to δ; if the θ's (except as specified in the text) are set equal to zero, the model dihedral angles are equal to the δ's that are output by the fitting program.

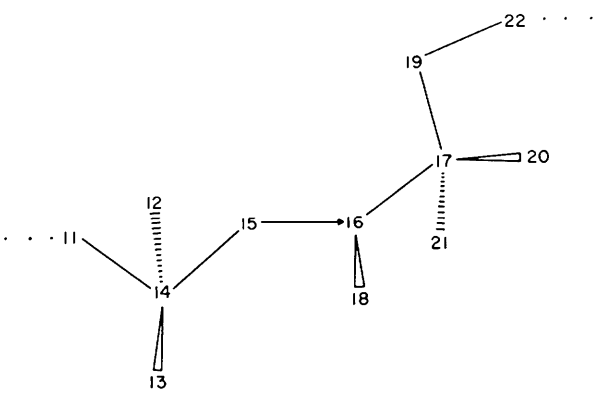

Fig. 2. Segment of an arbitrary chain molecule. Rotation around the spindle vector pointing from atom 15 to atom 16 (origin at atom 15) will alter the positions of atoms 17–22. The positions of atoms 11–16 are unaffected by this rotation.

where (x_i,y_i,z_i) are the coordinates before rotation, (x_f,y_f,z_f) are the coordinates after rotation, (x_s,y_s,z_s) are the coordinates of the end of the spindle vector which can be any bond in the molecule, and the l's are the direction cosines of the spindle vector. The dihedral rotation matrix in (1) is the transpose of the matrix given by Patterson (1959). The method distinguishes rotations around three types of spindle vectors: forward-backbone (*i.e.* in the direction from the N to the C terminus); side-chain (*i.e.* in the direction from the C^α to the last atom of the side chain); and backtwist-backbone (*i.e.* in the direction from the C to the N terminus). [As is the case with other model-building procedures (Diamond, 1966; Warme, Gō & Scheraga, 1972), the disulfide links are not formally treated as bonds. If they were considered as bonds, it would be necessary to treat most proteins as closed loops having fixed bond lengths and bond angles, and a rigorous treatment of this situation is a formidable problem (Gō & Scheraga, 1970). It is assumed that, when the model is fitted to the experimental coordinates, the distances between linked sulfur atoms will be suitably close to the experimentally observed disulfide-bond length. This is probably a fairly good assumption, since one would expect most disulfide bonds to be well defined in the electron-density maps of proteins; hence, the sulfur coordinates derived from the maps should be fairly accurate. In the event that the sulfurs are poorly defined, one might find it advantageous to include *restraints* on the disulfide-bond lengths and perhaps the C–S–S bond angles. A method for including such restraints has been used in several model-building or structure-refinement procedures (Hermans & McQueen, 1974; Dodson, Isaacs & Rollett, 1976; Sussman, Holbrook, Church & Kim, 1977) to force the model structure to retain approximately idealized short-range geometry.] For both the forward- and backtwist-backbone vectors, the spindle is chosen to coincide with a backbone bond. The definitions of the forward and the backtwist spindle vectors differ in their choice of vector origin: that of the forward vector is at the N-terminal end of the bond (the atom closer to the N terminus), while that of the backtwist vector is at the C-terminal end of the bond. The side-chain spindle vectors coincide with the side-chain bonds; the origin of such a spindle is the bond end closer to the C^α atom to which the side-chain is attached. For rotation, the origin of the cone of rotation is moved to the end of the spindle vector (x_s,y_s,z_s) as each rotation is performed.

The reason for defining both forward- and backtwist-backbone dihedral angles is to enable the procedure to start at a given residue (where the electron density is well defined) and to fit in both directions. The following description of the model-building process makes the advantages of this capacity to move in two directions clear. Typically, one starts by positioning a small portion of the model; other parts of the model are then moved into place with the original small portion (the starting segment) held close to its original position. The fit of the whole model depends strongly on the proper positioning of the starting segment; thus, it is important that the values of the parameters (dihedral angles, translational distances and Euler angles) which describe the position of the starting segment be close to the true values. One strategy that helps to obtain good values is to choose a *small* starting segment, whose position can be described by relatively few parameters: compensations for large errors in the values of the parameters cannot be made within a small set. The choice of the particular starting segment is of great importance: the guide points to which the model is being fitted should be clearly visible in the electron-density map, so that the starting segment of the model will match the corresponding portion of the real molecule. It is in choosing the starting segment that the new method offers a great advantage: the starting segment may lie anywhere in the molecule. The model may then be built from the starting segment, using forward rotations to position atoms between the C termini of the segment and of the molecule, and backtwist rotations to position atoms between the N termini of the segment and of the molecule. The user can choose a starting segment in a region where the electron density is well defined, rather than being forced to start the fitting at or near one end of the molecule, where the electron density is apt to be diffuse.

The use of (1) to determine the shifts in coordinates produced by variations of the dihedral angles dictates that the rotations must be applied in the same order every time the molecule is generated: most atoms are positioned by rotations around many bonds, which would lead to a matrix product in (1), and matrix multiplication is not commutative (Williams, 1972). Therefore, a fixed order of application of rotations must be defined; the choice of order is entirely arbitrary. All forward rotations are applied first, starting at the N-terminal point of the starting segment and running directly to the C-terminal point of the segment whose coordinates are to be fitted. The side-chain rotations are applied next. Finally, all backtwist rotations are applied, starting at the backtwist C-terminal point and running directly to the backtwist N-terminal point. Formally, of course, the dihedral angle around a given backbone bond may be described by *either* the value of a forward rotational angle *or* that of a backtwist rotational angle; there is no need to specify both values for one bond. However, in practice, it has been found that, by applying both forward and backtwist rotations to a *few* backbone bonds, the efficiency of the fitting process can be improved; hence, provision for the application of both types of rotation about a single bond has been made in a simple manner, *viz* by formally applying both types of rotation about every backbone bond. Of

course, specification of a zero value for a given backtwist rotation means that the dihedral angle around the bond is completely described by the value of the forward rotational angle, and *vice versa*.

Once the molecule has been generated in the desired conformation, it must be properly oriented in the unit cell before its coordinates can be compared with the experimental coordinates. Therefore, the molecule is translated as a rigid body along the x, y, and z axes according to

$$\begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix} = \begin{bmatrix} x_f + X \\ y_f + Y \\ z_f + Z \end{bmatrix}, \tag{2}$$

where (x_f,y_f,z_f) are the coordinates of every atom (from 1) before translation, (x_t,y_t,z_t) are the corresponding coordinates after translation and X, Y, Z are the translational distances. Finally, the molecule is rotated as a rigid body according to

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \begin{bmatrix} \cos\gamma\cos\beta\cos\alpha - \sin\gamma\sin\alpha & \cos\gamma\cos\beta\sin\alpha + \sin\gamma\cos\alpha & -\cos\gamma\sin\beta \\ -\sin\gamma\cos\beta\cos\alpha - \cos\gamma\sin\alpha & -\sin\gamma\cos\beta\sin\alpha + \cos\gamma\cos\alpha & \sin\gamma\sin\beta \\ \sin\beta\cos\alpha & \sin\beta\sin\alpha & \cos\beta \end{bmatrix} \begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix}, \tag{3}$$

where (x_t,y_t,z_t) are the coordinates of every atom before rotation, (x_c,y_c,z_c) are the corresponding coordinates after rotation (the final coordinates) and α, β, γ are the Euler rotational angles, chosen according to the convention generally used in the quantum theory of angular momentum (Arfken, 1970). The origin for rotation is chosen to be the origin of the unit cell. This choice differs from that used in other procedures for positioning the molecule in the unit cell (Nyburg, 1974; Ferro & Hermans, 1977), and means that the orientation process is not even approximately separable into translational and rotational components. It does have the merit of being computationally simple.

After these model coordinates (x_c,y_c,z_c) have been obtained, the function to be minimized (the sum of the weighted squares) is computed according to

$$F = \sum_{i=1}^{N} W^i[(x_c^i - x_e^i)^2 + (y_c^i - y_e^i)^2 + (z_c^i - z_e^i)^2], \tag{4}$$

where N is the number of atoms in the molecule, the W^i's are the weights assigned to the coordinates, and (x_e,y_e,z_e) are the experimental coordinates to which the model is being fitted. As was explained in the *Introduction*, the term in F corresponding to the ith atom is set equal to zero (*i.e.* the atom is considered to have been located *exactly*) if the position of this atom in the model is within a radius P of its experimental position. P is the uncertainty in the location of the guide points; a reasonable estimate of its value may be made after examination of the electron-density map. While the method of choice of the W^i's is entirely up to the user, a convenient one, used in the model building done here, assigns weights to the coordinates of each atom in proportion to the electron density at the atomic position.

The minimizer used by the procedure is *MINOP* (Dennis & Mei, 1975), a modification of Powell's (1970*a*,*b*) dog-leg strategy which appears to be powerful and quite efficient. Tests on simple functions have shown that, typically, *MINOP* minimizes to the same point as Powell's *MINFA* (1970*c*,*d*) in roughly two-thirds of the time (Dennis & Mei, 1975). The use of *MINOP* in several different applications, some of which involved the simultaneous minimization of over a hundred variables (unpublished results obtained in this investigation), indicates that it is more powerful than either the Davidon (1959) or classical least-squares minimizers, which are used in the earlier model-building procedures (Warme, Gō & Scheraga, 1972). *MINOP* does require values of gradients, as do the aforementioned minimizers. Given (3) and (4), the determination of the derivatives with respect to the Euler angles is trivial. The derivatives with respect to the dihedral angles are given by

$$\partial F/\partial\delta_j = \sum_{i=k}^{l} 2W^i[(x_c^i - x_e^i)\,\partial x_c^i/\partial\delta_j + (y_c^i - y_e^i)\,\partial y_c^i/\partial\delta_j + (z_c^i - z_e^i)\,\partial z_c^i/\partial\delta_j], \tag{5}$$

where

$$\partial x_c^i/\partial\delta_j = [\mathbf{n}_x \cdot (\mathbf{n}_j \times \mathbf{r}_{ij})][x_c^i/(x_c^{i2} + y_c^{i2} + z_c^{i2})^{1/2}] \tag{6}$$

and similarly for y_c^i and z_c^i. The (x_c^i,y_c^i,z_c^i) are the computed coordinates of the ith atom, (x_e^i,y_e^i,z_e^i) are the experimental coordinates of the ith atom, $\mathbf{n}_j$ is the unit spindle vector of the bond around which rotation δ_j is applied, $\mathbf{n}_x$ is the unit vector in the x direction, $\mathbf{r}_{ij}$ is the position vector of atom i with respect to the spindle origin, k is the first atom whose position is affected by a change in δ_j, and l is the last atom whose position is affected by such a change. The derivatives with respect to the translational distances are

$$\begin{bmatrix} \partial F/\partial X \\ \partial F/\partial Y \\ \partial F/\partial Z \end{bmatrix} = \sum_{i=1}^{N} 2W^i \begin{bmatrix} \cos\gamma\cos\beta\cos\alpha - \sin\gamma\sin\alpha & -\sin\gamma\cos\beta\cos\alpha - \cos\gamma\sin\alpha & \sin\beta\cos\alpha \\ \cos\gamma\cos\beta\sin\alpha + \sin\gamma\cos\alpha & -\sin\gamma\cos\beta\sin\alpha + \cos\gamma\cos\alpha & \sin\beta\sin\alpha \\ -\cos\gamma\sin\beta & \sin\gamma\sin\beta & \cos\beta \end{bmatrix} \begin{bmatrix} x_c^i - x_e^i \\ y_c^i - y_e^i \\ z_c^i - z_e^i \end{bmatrix}. \tag{7}$$

Computer program

A Fortran IV computer program implementing the method described above has been written. The program occupies 124 kbytes of storage on an IBM 370/168 machine (*H*-extended compilation). It has been tested by fitting a model of bovine pancreatic trypsin inhibitor (BPTI) to a set of experimental coordinates. A program package containing the fitting program, a program to generate the zero-point model, and complete user descriptions for both is available.* The fixed values of the bond lengths and bond angles supplied with the package are those of *UNICEPP* (Dunfield, Burgess & Scheraga, 1978), but the user can easily substitute other values if this is desired.

The program is designed to be used with a large amount of operator intervention. The program simply optimizes the fit between the experimental and model coordinates or a subset of coordinates by altering the variables (dihedral angles, translational distances, and Euler angles) specified by the user. The user must plan the fitting process (*i.e.* make such decisions as when to add a residue to be fitted, when to fit a certain portion of the chain, whether to refit a part of the molecule, and when to do so) himself. No decision-making procedures are automated; by leaving these decisions to the user, the procedure is made more flexible than it would be otherwise.

Because the fitting process is directed by the operator, it is very difficult to make any estimate of the time required to produce a satisfactory model for a protein of, say, 100 residues. During the fitting of a model of bovine pancreatic trypsin inhibitor (discussed below), it was observed that one cycle of minimization with 180 backbone dihedral angles as variables took 6 s on an IBM 370/168 computer (FORTG compilation). This figure is probably an upper limit on the time/cycle; in practice, the use of such a large set of 180 variables is less advantageous than serial minimizations using smaller subsets of variables. The program has several time-saving features which can reduce running time drastically if used in conjunction with a carefully chosen set of variables.

Efficacy of the method

The fit of the model to the experimental BPTI coordinates provides a good assessment of the power of the method. A summary of the results of the fitting process will be given here. All dihedral angles (including ω's) were allowed to vary. The coordinates to which the model was fitted were derived from an electron-density map of BPTI calculated from data with 2·5 Å resolution (Huber *et al.*, 1970; Steigemann, 1976); the coordinate weights were selected to be proportional to the electron density at the atomic position. These 'rough-map' coordinates are referred to as RM1. A Kendrew wire model which had been fitted to the 2·5 Å map earlier was used as a guide for the derivation of these coordinates (Steigemann, 1976). The only stereochemical constraint that we placed on the coordinates of RM1 was that each atom be between 1 and 2 Å from any atom to which it was bonded. The use of the coordinates of RM1 as guide points provides a fairly stringent test of the new method: the model-building process must impose many short-range stereochemical constraints which were not imposed during the derivation of the guide points, and still fit the guide points fairly closely.

The r.m.s. deviation of the coordinates for the weighted atoms (about 70% of the total number of non-hydrogen atoms) of the model from the RM1 guide points was 0·49 Å; if the seven atoms (roughly 2·5% of the weighted coordinates) with the worst individual r.m.s. deviations were ignored, the total r.m.s. deviation dropped to 0·41 Å. Figs. 3 and 4 provide a more detailed assessment of the fit. Fig. 3 is a histogram of

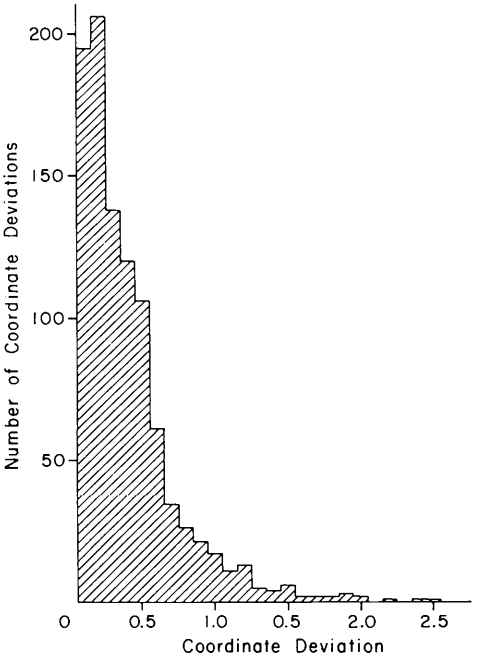

Fig. 3. Histogram of deviations (Å) between model coordinate components (x, y, and z coordinates) and RM1 coordinate components for BPTI.

* The program package is available as document No. NAPS-03402 of the ASIA National Auxiliary Publication Service, c/o Microfiche Publications, PO Box 3513, Grand Central Station, New York, NY 10017, USA. A copy may be secured by citing the document number and by remitting $3.00 for microfiche or $35.00 for photocopies. Outside the United States and Canada, postage is $1.00 for a microfiche or $3.00 for a photocopy. Advance payment is required. Make check or money order payable to Microfiche Publications.

the differences between the coordinates of the model atoms and those of the guide points. 55% of the coordinate differences are less than 0·3 Å, the value chosen for the uncertainty in the locations of the coordinates on the basis of the map resolution and possible errors in reading the x,y,z coordinates from the wire model (for BPTI, $P = 0{\cdot}3\sqrt{3}$). The largest coordinate deviation is 2·5 Å. Fig. 4 shows the r.m.s. deviations for the individual residues. The fit of a few residues in the neighborhood of residue 20 is noticeably poorer than that of the rest of the chain; several of these, however, have bulky aromatic side chains, and it can be difficult to obtain a good fit for these (see Fig. 5).

The fitting process typically did not pull the locations of model atoms out of the regions of high electron density seen in the experimental map. Fig. 6 shows a portion of the electron-density map near residues 7, 8 and 9. The guide points and model atoms are also shown. Both are well embedded in the high-density region.

The fitting method developed here does not fix the disulfide-bond lengths or the C–S–S bond angles, nor does it restrain these quantities (although such restraints could easily be added). It is of interest, therefore, to compare the values of the disulfide-bond lengths and C–S–S bond angles given by the guide points with those obtained in the model. These are compared in Table 1. The bond lengths are, for the most part, far from experimental values, which would argue for the use of bond-length (and perhaps bond-angle) restraints at some time in the model-building pro-

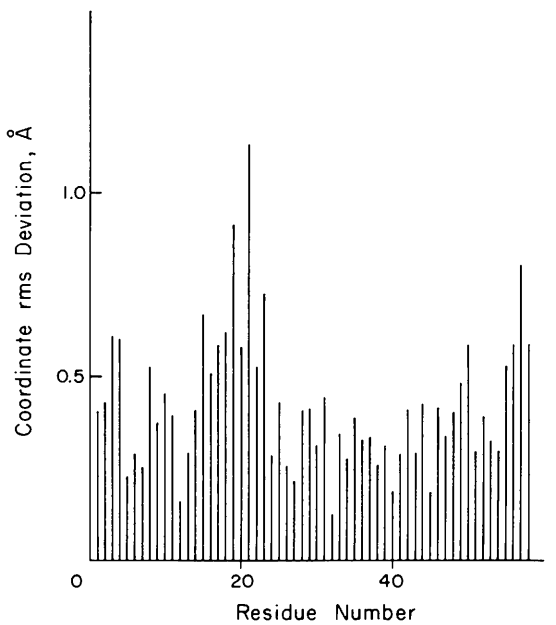

Fig. 4. The r.m.s. deviations between weighted model and RM1 coordinate components for the individual residues of BPTI. The r.m.s. deviation is

$$\left\{\sum_{i=1}^{n}[(x_c^i - x_e^i)^2 + (y_c^i - y_e^i)^2 + (z_c^i - z_e^i)^2]/3n\right\}^{1/2},$$

where n = number of non-hydrogen atoms in residue.

Table 1. *Disulfide-bond lengths* (Å) *and* C–S–S *bond angles* (°) *from guide points and after model building*

The average experimental values for the disulfide-bond length and C–S–S bond angle are 2·04 Å and 104°, respectively (in the documentation provided with the paper of Dunfield, Burgess & Scheraga, 1978).

	From guide points*	After model building
S(5)–S(55)	1·456	2·043
S(14)–S(38)	2·020	2·575
S(30)–S(51)	1·616	1·844
C(55)–S(55)–S(5)	101·9	112·5
C(51)–S(51)–S(30)	107·6	104·4

* C(5), C(14), C(30) and C(38) were not visible in the electron-density map and hence were not used as guide points.

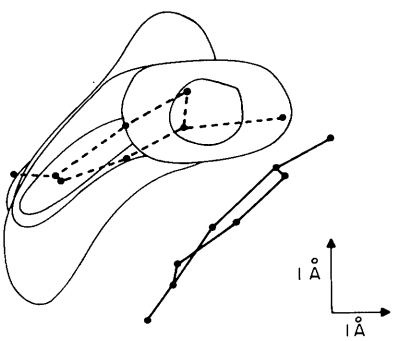

Fig. 5. Section of the electron-density map showing the side chain of Tyr 21. Dotted lines connect the RM1 coordinates, while solid lines connect the model coordinates. The model side chain has moved out of the high electron-density region entirely, due primarily to the poor fit of C^{β}. The side chain lacks the flexibility to move itself back into the proper region once the C^{β} atom has moved out of the proper position.

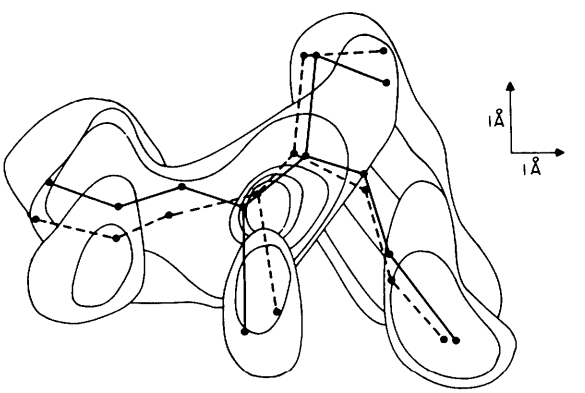

Fig. 6. Section of the electron-density map near residues 7, 8 and 9. Dotted lines connect the RM1 coordinates, while solid lines connect the model coordinates.

cedure. We prefer to add these restraints when we add potential-energy restraints to model building (Fitzwater & Scheraga, 1980).

The r.m.s. deviation obtained from the fitting of BPTI may be compared with that obtained in similar studies using different model-building methods. Some difficulty in making such a comparison arises: some model-building methods themselves are not directly comparable because different quantities are allowed to vary in different methods; also, most sets of guide coordinates had more reasonable stereochemical constraints than did our set. The most valid comparison that can be made is probably that with the fit of BPTI carried out by Swenson, Burgess & Scheraga (1978) using the method of Warme, Gō & Scheraga (1972). A standard-geometry model was fitted to the coordinates of the 1·5 Å refined structure (Deisenhofer & Steigemann, 1975) with the peptide-bond dihedral angles of the model fixed at 180°. The r.m.s. coordinate deviation was 0·3 Å (this figure, and the one cited next, are the r.m.s. deviation given in the relevant paper divided by $\sqrt{3}$ to put it on the same basis as the r.m.s. deviations obtained in this study). Deisenhofer & Steigemann (1975) fitted a model to the coordinates of the BPTI wire model obtained from the 2·5 Å electron-density model using Diamond's (1966) method, which allows the peptide-bond dihedral angles and the N–C^{α}–C′ bond angle to vary. They obtained a r.m.s. coordinate deviation of 0·21 Å. The r.m.s. coordinate deviations resulting from both of these studies are lower than the r.m.s. coordinate deviation obtained from the study reported here, but the guide points used here were not subject to the stringent stereochemical constraints imposed on the guide points used by Swenson, Burgess & Scheraga (1978), or even the milder ones embodied in the wire-model coordinates used by Deisenhofer & Steigemann (1975). The method developed here, then, appears to be capable of fitting a model with fixed bond lengths and bond angles to a set of guide points with reasonable accuracy. Given its power and utilization of certain previously ignored experimental information, the method should be a useful tool for various studies involving protein structure.

First step of structure refinement

As a preliminary step in a refinement of the 2·5 Å structure by potential-energy-constrained model building, the model obtained here was subjected to energy minimization with the atomic coordinates constrained to remain 'close' to the original guide points. Those atoms (13% of the total) which were not visible in the map and whose locations could not be inferred from the locations of atoms which were visible in the map were positioned by energy criteria alone; they were moved into low-energy positions, with the positions of all other atoms fixed by model building. Then, the following function was minimized:

$$G = E_{\text{pot}} + W_{\text{rel}} \sum_{i=1}^{N} W^i[(x_c^i - x_e^i)^2 + (y_c^i - y_e^i)^2 + (z_c^i - z_e^i)^2], \quad (8)$$

where E_{pot} is the *UNICEPP* potential energy (Dunfield, Burgess & Scheraga, 1978) and W_{rel} is a selected weighting factor. Equation (8) combines standard geometry and potential-energy information with structural information derived from the 2·5 Å electron-density map. The r.m.s. deviation between the model and 1·5 Å refined coordinates for the weighted atoms after the minimization of G rose to 0·64 Å; for all atoms (including those positioned by energy criteria alone), this figure is 0·93 Å. The R factor for all reflections with resolution between 7·5 and 2·5 Å is 0·56; this may be compared with the value of 0·52 given by the standard-geometry model obtained from the 2·5 Å electron-density map by Deisenhofer & Steigemann (1975). Examination of the details of the fit shows that the high values for the r.m.s. coordinate deviation are caused largely by poor fits of a few atoms; many of the atoms which were positioned entirely by energy criteria are far from the 1·5 Å refined locations. The best way to improve this situation is to improve the electron-density map, so that more atoms become visible. A description of the methods used to improve the map, and the refinement of the 2·5 Å BPTI structure, will be published elsewhere.

We are indebted to Dr W. Steigemann for providing us with the Kendrew wire-model coordinates and the 2·5 Å structure factors and phases, to the Brookhaven Data Bank for the 1·5 Å refined coordinates, and to Dr George Némethy for helpful comments on the manuscript.

This work was supported by research grants from the National Science Foundation (PCM75-08691), from the National Institute of General Medical Sciences of the National Institutes of Health, US Public Health Service (GM-14312), and from The Mobil Foundation.

SF was a Chaim Weizmann Postdoctoral Fellow, 1977–79.

References

ARFKEN, G. (1970). *Mathematical Methods for Physicists*, pp. 178–180. New York: Academic Press.

DAVIDON, W. C. (1959). *USAEC Res. Dev. Rep.* ANL-5990.

DEISENHOFER, J. & STEIGEMANN, W. (1975). *Acta Cryst.* B**31**, 238–250.

DENNIS, J. E. & MEI, H. H. W. (1975). *An Unconstrained Optimization Algorithm which Uses Function and Gradient Values*, Tech. Rep. No. TR 75-246, Dept of Computer Science, Cornell Univ., Ithaca, New York 14853.

DIAMOND, R. (1966). *Acta Cryst.* **21**, 253–266.
DIAMOND, R. (1971). *Acta Cryst.* A**27**, 436–452.
DODSON, E. J., ISAACS, N. W. & ROLLETT, J. S. (1976). *Acta Cryst.* A**32**, 311–315.
DUNFIELD, L. G., BURGESS, A. W. & SCHERAGA, H. A. (1978). *J. Phys. Chem.* **82**, 2609–2616.
EPP, O., COLMAN, P., FEHLHAMMER, H., BODE, W., SCHIFFER, M., HUBER, R. & PALM, W. (1974). *Eur. J. Biochem.* **45**, 513–524.
FERRO, D. R. & HERMANS, J. (1977). *Acta Cryst.* A**33**, 345–347.
FITZWATER, S. & SCHERAGA, H. A. (1980). In preparation.
GŌ, N. & SCHERAGA, H. A. (1970). *Macromolecules*, **3**, 178–187.
HERMANS, J. JR & MCQUEEN, J. E. JR (1974). *Acta Cryst.* A**30**, 730–739.
HUBER, R., KUKLA, D., BODE, W., SCHWAGER, P., BARTELS, K., DEISENHOFER, J. & STEIGEMANN, W. (1974). *J. Mol. Biol.* **89**, 73–101.
HUBER, R., KUKLA, D., RUHLMANN, A., EPP, O. & FORMANEK, H. (1970). *Naturwissenschaften*, **57**, 389–392.
JACK, A. (1977). *Acta Cryst.* A**33**, 497–499.
NYBURG, S. C. (1974). *Acta Cryst.* B**30**, 251–253.
PATTERSON, A. L. (1959). *International Tables for X-ray Crystallography*, Vol. II, p. 63. Birmingham: Kynoch Press.
PINCUS, M. R. & SCHERAGA, H. A. (1979). *Macromolecules.* **12**, 633–644.
PINCUS, M. R., ZIMMERMAN, S. S. & SCHERAGA, H. A. (1976). *Proc. Natl Acad. Sci. USA*, **73**, 4261–4265.
PINCUS, M. R., ZIMMERMAN, S. S. & SCHERAGA, H. A. (1977). *Proc. Natl Acad. Sci. USA*, **74**, 2629–2633.
PLATZER, K. E. B., MOMANY, F. A. & SCHERAGA, H. A. (1972). *Int. J. Pept. Protein Res.* **4**, 201–219.
POWELL, M. J. D. (1970*a*). *A Hybrid Method for Nonlinear Equations*, in *Numerical Methods for Nonlinear Algebraic Equations*, edited by P. RABINOWITZ, pp. 87–114. London: Gordon & Breach.
POWELL, M. J. D. (1970*b*). *A Fortran Subroutine for Solving Systems of Nonlinear Algebraic Equations*, in *Numerical Methods for Nonlinear Algebraic Equations*, edited by P. RABINOWITZ, pp. 115–161. London: Gordon & Breach.
POWELL, M. J. D. (1970*c*). *A Fortran Subroutine for Unconstrained Minimization, Requiring First Derivatives of the Objective Function*, UK At. Energy Auth., Res. Group, Culham Lab., Rep. AERE-R 6469, Harwell, England.
POWELL, M. J. D. (1970*d*). *A New Algorithm for Unconstrained Optimization*, in *Nonlinear Programming*, edited by J. B. ROSEN, O. L. MANGASARIAN & K. RITTER, pp. 31–65. New York: Academic Press.
RASSE, D., WARME, P. K. & SCHERAGA, H. A. (1974). *Proc. Natl Acad. Sci. USA*, **71**, 3736–3740.
STEIGEMANN, W. (1976). Personal communication.
SUSSMAN, J. L., HOLBROOK, S. R., CHURCH, G. M. & KIM, S. H. (1977). *Acta Cryst.* A**33**, 800–804.
SWENSON, M. K., BURGESS, A. W. & SCHERAGA, H. A. (1978). In *Frontiers in Physico-Chemical Biology*, edited by B. PULLMAN, pp. 115–142. New York: Academic Press.
TEN EYCK, L. F., WEAVER, L. H. & MATTHEWS, B. W. (1976). *Acta Cryst.* A**32**, 349–350.
WARME, P. K., GŌ, N. & SCHERAGA, H. A. (1972). *J. Comput. Phys.* **9**, 303–317.
WASER, J. (1963). *Acta Cryst.* **16**, 1091–1094.
WILLIAMS, D. E. (1972). *Acta Cryst.* A**28**, 629–635.

Acta Cryst. (1980). A**36**, 219–228

The Space-Group Determination of GaS and $Cu_3As_2S_3I$ by Convergent-Beam Electron Diffraction

BY P. GOODMAN AND H. J. WHITFIELD

CSIRO Division of Chemical Physics, PO Box 160, *Clayton, Victoria, Australia* 3168

(*Received* 1 *May* 1979; *accepted* 24 *September* 1979)

Abstract

Space-group determinations were carried out on GaS and $Cu_3As_2S_3I$ single crystals using convergent-beam electron diffraction data. The space groups found were $P6_3/mmc$ and *Cmcm*, respectively. The effectiveness of the specific test for centrosymmetry was examined. In both structures the test was made clearer by use of the inclined-axis technique, in which the incident beam is at an appreciable angle to the principal zone axis. It was concluded that the test was superior to optical methods in structures such as GaS which have a high dislocation density.

Introduction

Since the publication of a space-group identification procedure consisting of specific symmetry tests in convergent-beam electron diffraction (Goodman, 1975), a number of substances have been examined in

0567-7394/80/020219-10$01.00