Acta Cryst. (1976). A32, 311

A Method for Fitting Satisfactory Models to Sets of Atomic Positions in Protein Structure Refinements

BY E. J. DODSON AND N. W. ISAACS*
Laboratory of Molecular Biophysics

AND J. S. ROLLETT
Computing Laboratory, University of Oxford, England

(*Received* 10 *July* 1975; *accepted* 22 *September* 1975)

The formal refinement methods of least-squares adjustment or difference-map analysis give atomic positions in protein structures with standard deviations which are large compared with the standard deviations of accepted molecular dimensions. This paper describes a method of adjusting the Cartesian coordinates to obtain a properly weighted fit to both the positions from the refinement and the molecular parameters. The equations which have to be solved have many unknowns but few coefficients, and an effective iterative method can be used. The results of applications of the method to insulin are summarized.

Introduction

At various stages in the determination of the crystal structure of a protein, an electron density map is interpreted to produce a set of atomic positions. These are usually chosen to give a good fit to the electron density, but it is not usually easy to ensure that all the bond distances, angles and other molecular parameters corresponding to the chosen positions agree with those found in smaller structures.

A similar situation arises when position shifts are deduced either by analysis of a difference map, or by the method of least-squares.

The need arises for a method by which the best interpretation can be made, either for the calculation of structure factors and further refinement or for publication. It has been common practice to build a model in which certain types of bond distances and angles agree exactly with pre-set values. Usually a technique such as that of Diamond (1966) is used, in which model-building starts at one end of a polypeptide chain and successive dihedral angles are adjusted in turn to give a good fit to positions obtained by interpretation of the density. The difficulty can arise that no set of positions obeying the distance and angle conditions fits the density well enough. If attempts are made to put matters right by allowing certain angles to vary, then these angles are liable to take on improbable values.

These difficulties are not surprising. Molecular dimensions are the results of interactions between interatomic forces for atoms bonded together and those for atoms further away in the structure. We can reasonably expect that the forces due to secondary and tertiary structural features will cause small variations in primary structural dimensions.

These variations are hard to predict and we cannot expect to do more than to take them into account by assigning standard deviations to the primary dimensions. It is not feasible to carry out a detailed spectroscopic analysis for force constants in a molecular system of the size of a protein. The standard deviations for molecular parameters which we use are, therefore, approximate estimates. Fortunately this appears to be adequate for our purposes.

The primary dimensions become inexact observations with the same status as the observations on the atomic positions. All the observables are functions of the Cartesian coordinates of the atoms, and an optimal fit can be obtained by a least-squares adjustment of these.

The structure of insulin contains about 800 unique atomic positions, so that there are about 2400 Cartesian coordinates to be adjusted. The situation can be improved by adjusting separate polypeptide chains independently, but the systems of equations are still very large. We have solved them successfully by use of the conjugate gradient algorithm of Hestenes & Stiefel (1952). For this, the normal equations need not be computed explicitly, and full advantage is gained from the small number of non-zero coefficients in the observational equations.

The method which we describe here is very similar to those used by Levitt & Lifson (1969), Levitt (1974) and Hermans & McQueen (1974). They have obtained their minimization functions by discussing the energies required for molecular deformations, while we have derived a minimization function by considering force constants, but these two approaches are essentially equivalent. Our work differs from that of these other authors in that we do not attempt to include the effects of van der Waals forces except by assigning standard deviations to constrained interatomic distances. Accordingly the function that we minimize is quadratic in

* Present address: IBM Thomas J. Watson Research Laboratories, Yorktown Heights, N. Y., U.S.A.

the atomic coordinates to a good approximation and we obtain rather rapid convergence with the conjugate gradient algorithm. Because of this our method can conveniently be applied to adjust the coordinates given by each stage of difference-map refinement, and adds little to the total computing time.

Another difference between our work and that of previous authors lies in the technique we have used to maintain approximate planarity for aromatic groups and peptide linkages.

At an early stage of the refinement the positions obtained from a difference map are of low accuracy. They are therefore given low weight, and the model produced by our method obeys the heavily weighted constraints on bond lengths almost exactly. As the accuracy of the map coordinates improves, this situation changes, and any departures from constrained molecular dimensions which then develop can reveal discrepancies between the assumptions which have been made and the actual behaviour of the molecule. We show later how this was used to improve our model for the planarity of the peptide groups of insulin.

Minimization function and observational equations

We minimize a sum of squared terms of two kinds.

(*a*) The weighted squares of the distances between the indicated positions (*e.g.* positions of peaks on a map) and those for the adjusted model.

(*b*) The weighted squares of the differences between target interatomic distances and those for the adjusted model.

The interpretation of (*a*) needs no comment, nor that of (*b*) for constraints on bond lengths, but we also constrain bond angles, and this is done by defining target interatomic distances for pairs of atoms which are next-nearest neighbours in the structure. Constraints on bond distances and angles are not adequate to ensure reasonable planarity for groups such as peptide links and aromatic side chains. For each such group we place an artificial null-atom (since it makes no contribution to any structure factor) 20 Å from the plane of the group along some convenient line normal to this plane. Constraints on the distances between the null-atom and those in the group then maintain planarity. It has been found to be satisfactory to evaluate target distances for these constraints by calculation for an idealized planar group, so that only one set of target distances needs to be found for each type of group. The positions of the null-atoms are not constrained in any way except by their distances from the atoms of the groups.

The weight attached to each map coordinate is the inverse of the square of its estimated standard deviation. Similarly the weight for a distance is the inverse square of its e.s.d. Because the distances are better defined than the positions of the atoms, the equations are somewhat ill-conditioned in such a way that the fitted model can move as a rigid body without making any great difference to the minimization function. This leads to some slowness in the convergence of the conjugate gradient algorithm.

It can be shown that the various terms in the minimization function correspond to observational equations of the following kinds. For an atomic positional coordinate, *e.g.* x_r with an indicated value ξ_r (ξ_r may be given by a peak position on a Fourier map) we get

$$\sqrt{W_r}\,\delta x_r = \sqrt{W_r}(\xi_r - x_r)\,,$$

where δx_r is the correction to be applied to the present value of x_r, and W_r is the reciprocal of the square of the estimated standard deviation of ξ_r.

For an interatomic distance calculated as l_{rs} from the model, with a target value d_{rs}, we get

$$\sqrt{W_{rs}}\frac{\partial l_{rs}}{\partial x_r}\delta x_r + \sqrt{W_{rs}}\frac{\partial l_{rs}}{\partial y_r}\delta y_r + \ldots + \sqrt{W_{rs}}\frac{\partial l_{rs}}{\partial z_s}\delta z_s = \sqrt{W_{rs}}\,(d_{rs} - l_{rs})\,,$$

where, again, $\delta x_r \ldots \delta z_s$ are the corrections to be applied to the six positional coordinates for atoms r and s, and W_{rs} is the reciprocal of the square of the estimated standard deviation of d_{rs}. The derivative $\partial l_{rs}/\partial x_r$ is given by

$$\partial l_{rs}/\partial x_r = (x_r - x_s)/l_{rs}\,,$$

with similar equations for the other derivatives. These equations are written as if (x_r, y_r, z_r) are referred to orthogonal axes and measured in Å, but the modifications for other situations are not hard to devise.

The whole collection of observational equations can be written in matrix notation as

$$B\delta\mathbf{x} = \mathbf{d}\,,$$

where B is a large matrix having very few non-zero coefficients (no more than six in any one row). The corresponding normal equations are

$$B^T B\delta\mathbf{x} = B^T\mathbf{d}\,,$$

and we could, in principle, form and solve these. In practice this is inconvenient because B^TB has a large number of non-zero coefficients and it is therefore troublesome to store and manipulate. We avoid the difficulty by use of the conjugate gradient algorithm, which can be written

$$\begin{aligned}
&\text{Get}\quad \delta\mathbf{x}_0 = \mathbf{0}\\
&\text{Get}\quad \mathbf{p}_0 = \mathbf{r}_0 = B^T\mathbf{d}\\
&\text{For}\quad k = 0, 1, 2, \ldots\\
&\text{Get}\quad \mathbf{q}_k = B^TB\mathbf{p}_k\\
&\qquad \alpha_k = \mathbf{r}_k^T\mathbf{p}_k/\mathbf{p}_k^T\mathbf{q}_k\\
&\qquad \delta\mathbf{x}_{k+1} = \delta\mathbf{x}_k + \alpha_k\mathbf{p}_k\\
&\qquad \mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\mathbf{q}_k\\
&\qquad \beta_k = -\mathbf{r}_{k+1}^T\mathbf{q}_k/\mathbf{p}_k^T\mathbf{q}_k\\
&\qquad \mathbf{p}_{k+1} = \mathbf{r}_{k+1} + \beta_k\mathbf{p}_k\,.
\end{aligned}$$

It will be noticed that each step requires the multiplication of a vector by the sparse matrix B followed by multiplication of the resulting vector by B^T. The elements of B can either be stored in a compact way, or they can, quite economically, be computed each time they are required. Apart from this, the process requires storage for the vectors $\mathbf{p}_k$, $\mathbf{q}_k$, $\mathbf{r}_k$ and $\delta\mathbf{x}_k$, each of which can be overwritten by its successor.

The iteration can be continued for as many values of k as are necessary to provide convergence. Theoretically, perfect convergence is achieved when k is equal to the number of parameters in $\delta\mathbf{x}$, but this of little practical value if that is about 2400. In fact the process converges quite fast and we have chosen to terminate when $k=10$ if the largest element of $\alpha_k\mathbf{p}_k$ has not become smaller than a pre-set limit. The elements of B are then recalculated to take account of any alterations in the direction cosines of interatomic distances and the process is re-started. Four or five such outer iterations have normally provided an adequate fit, and the speed of the process has not been sensitive to the number of inner iterations in an outer iteration.

Most computing installations are likely to be able to provide a library routine to carry out the conjugate gradient process. It may prove difficult to use such a routine, however, because of the need to use the factorized form B^TB for the matrix of the equations, or because of the storage requirements for this problem.

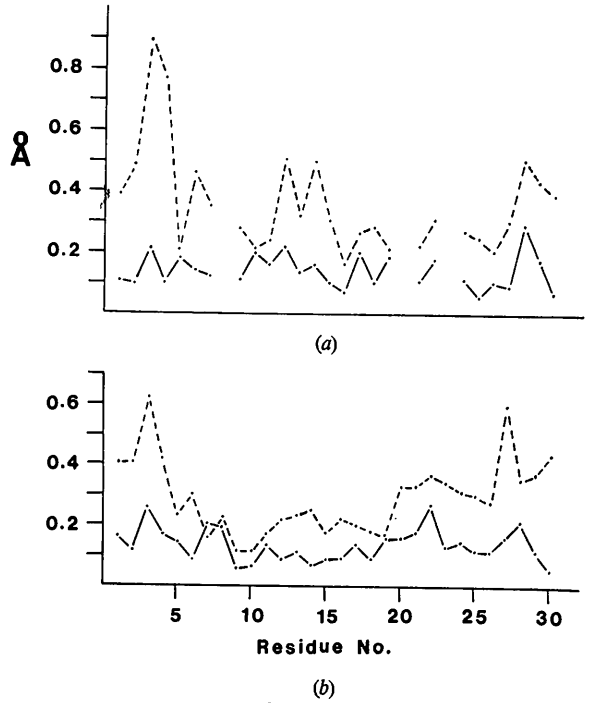

Fig. 1. R.m.s. deviations in Å from initial coordinates derived from a 2·8 Å resolution map for a B chain of 4-Zn insulin. Each point represents the r.m.s. deviation for the atoms of a single residue. Graph (*a*) is for the atoms of side chains only. Graph (*b*) is for the main-chain atoms only. The broken line is for the fitting procedure of Diamond (1966) and the chain dotted line is for the procedure of this paper.

There should be little difficulty in producing a program which will serve, but it may be worth while to remark here that the value of α_k is such that $\mathbf{r}_{k+1}^T\mathbf{p}_k=0$, and that β_k is such that $\mathbf{p}_{k+1}^T\mathbf{q}_k=0$, for all k. These relations serve as a useful check on the correctness of the algorithm.

Applications of the method to insulin structures

We have used this method to fit acceptable models to sets of atomic positions obtained from:

(i) a 2·8 Å resolution map of 4-zinc insulin

(ii) a 1·5 Å resolution map of 2-zinc insulin, where the phases used were obtained by the Sayre least-squares phase refinement process (Sayre, 1972) and

(iii) a difference synthesis of 2-zinc insulin at 1·5 Å resolution.

(i) The initial coordinates for the 4-Zn molecule (100 residues, 812 atoms) were obtained in the usual way by fitting a model to the electron density map in a Richards comparator.

These coordinates gave a model which deviated from ideal geometry by up to 1·4 Å in a bond distance 45° in a valence angle and 17° in the ω torsion angle. We assigned standard deviations of 0·2 Å to the X and Y and 0·3 Å to the Z coordinates (parallel to the section axis of the electron density map) and refined the model for four cycles, treating each of the four chains as a separate entity except for the disulphide bridges. The overall geometry of the model based on these refined coordinates gave bond distances usually within 0·01 Å of the ideal values, (maximum difference 0·02 Å), valence angles usually within 5° of ideal (maximum difference 24°) and ω torsion angles greater than 176°. The largest deviations from ideal values are commonly found to be at C_α atoms.

The efficiency of this method of refinement may be evaluated by a comparison with the results of a model fitted to the original coordinates by the Diamond (1966) procedure. Both operations were carried out on an ICL 1906A computer with CPU times estimated at 4 min for this method and 25 min for the Diamond method. Fig. 1 shows the RMS deviations of coordinates derived by each of the two methods, from the initial coordinates, for one of the chains of the insulin molecule.

It can be seen that the use of the method described here constrains the model to lie more closely to the initial set of coordinates than the model given by the rigid-body method and this is perhaps reflected in the calculated structure factors which gave R values of 0·409 for the original coordinates, 0·407 for coordinates derived by the method described here and 0·413 for coordinates derived from the Diamond method.

(ii) Coordinates were obtained from the 1·5 Å Sayre map of 2-Zn insulin by choosing positions as close as possible to the maxima on the electron density map. This crude approach led to a model with severe discrepancies in the molecular geometry, *e.g.* by as much as 1·1 Å in bond lengths, 70° in valence angles

and 54° in ω torsion angles. These coordinates were assigned standard deviations reflecting our confidence in their positions. For example carbonyl O atoms appeared very strongly in the map and were assigned relatively low standard deviations, typically of 0·08 Å in the X and Y coordinates and 0·12 Å in Z. After five cycles of refinement the model coordinates showed a geometry where the largest deviations from ideal values were 0·03 Å in bond lengths, 23° in bond angles and 5° in the ω torsion angle. The overall computation time was of the order of 1·4 s per residue.

In Fig. 2 a typical section of the electron density map is shown with the initial and refined coordinates superimposed.

(iii) We have also used the method to correct the geometry of the coordinates obtained for 2-Zn insulin at 1·5 Å resolution after two rounds of difference Fourier refinement. The geometry of this model deviated from the ideal by up to 0·6 Å in bond lengths, 50° in valence angles and 40° in the ω torsion angle. After three cycles of refinement the deviations from the ideal geometry were of the same order as for the previous examples. A possible danger in using a model-fitting method to correct the geometry of coordinates derived from a difference map is that such a correction may produce shifts in the coordinates in directions which negate the improvements brought about by the difference-map refinement.

At early stages in the 2-Zn insulin refinement this did not occur. For the second difference map R for the output coordinates was 30·7%. The coordinates were then adjusted by the method of this paper and R rose to 31·1%. A difference map computed from these coordinates gave output coordinates yielding an R of 26·8%. At later stages the increase in R as a result of fitting the coordinates to molecular dimensions became as much as 2%. It was believed that this was due to the requirement that the peptide groups should be planar to high accuracy, and the standard deviations for α C atoms in the peptide planar groups were increased from 0·01 to 0·04 Å. In the next cycle the increase in R produced by the fitting process was reduced to 0·6%.

It is clear from these results that any excessive rigidity in the set of constraints will produce a significant increase in disagreement with the X-ray data once the refinement progresses to the point at which the accuracy is sufficient to reveal this. Careful examination of the results in such a situation can lead to an understanding of the problem and the effects of an appropriate relaxation of the constraints can then be tested. We consider that it is an advantage of the flexible constraint process that the validity of the assumptions can be tested in this way.

Description of the programs

The program is written in standard Fortran and a version currently in use on the ICL 1906A computer will accept up to 65 residues (650 atoms) with a core requirement of 90 K words (24 bit word length). The target distances used in the progam have been taken from *Molecular Structures and Dimensions* Vols. 1 to 5 (Kennard, Watson & Town, 1970*a*, *b*, 1971, 1973, 1974). The e.s.d.'s have been set as 0·01 Å for bond distance and planarity constraints (distance between the atoms in the planes and the null-atoms), and 1·5° for the valence angles. This gives a root weight of 100 for the bond and planarity constraints and of the order of 60 for tetrahedral angles with decreasing values for smaller angles. The null-atoms are given e.s.d.'s of 1000 Å, which effectively allows them to move freely from their original calculated positions (20 Å on a line normal to the least-squares plane and passing through the centroid of the group of atoms). Our present experience suggests that this rather arbitrary weighting scheme is adequate.

For the first cycle of refinement the model coordinates are set to the input map coordinates and, in order to prevent too large a shift from these positions, all the map coordinates which have e.s.d.'s greater than 0·02 Å are given an e.s.d. of that value. In later cycles the true e.s.d. for each coordinate is used.

The input to the program consists of a file of orthogonal coordinates, a connectivity file and a file specifying planar groups of atoms. These latter two files are created by a separate program which has been freely adapted by E. J. D. from the standard group section of a program written by Diamond (1966). Because of the gross errors which may arise in reading

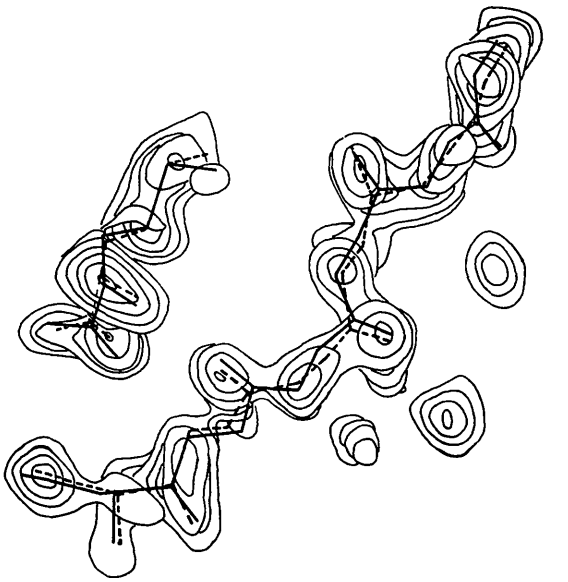

Fig. 2. Composite of sections of the map of 2-Zn insulin, phased by the method of Sayre (1972), showing electron density peaks corresponding to parts of one of the B chains. The broken lines connect positions estimated directly from the peaks. The full lines connect positions fitted by the method of this paper. It can be seen that the imposition of satisfactory molecular dimensions does not spoil the fit to the density.

protein coordinates from a map the connectivity has to be based on atom types rather than on interatomic distances. This program makes the connexions between atoms on the basis of the atom label and at the same time assigns for each connexion a code number which defines the target bond length. Where an atom has more than one connexion the combination of these codes defines the target angle. In a similar way atoms which should be planar are listed and coded. The program calculates bond lengths and angles and prints a list of these which should be checked before the fitting calculations are carried out.

Copies of the Fortran text of this program, which includes the dimensions which we have used for the various standard groups, can be obtained from one of us (N.W.I.).

Conclusions

The method of this paper has invariably produced models for protein structures with acceptable bond lengths, bond angles, and planarities, and the positions obtained have fitted all the atomic positions derived from refinement processes, within the limits prescribed on the basis of the estimated accuracies of these positions. The cost of applying the method has been considerably less than that of computing a difference map or that for the corresponding set of structure factors, in the case of insulin. Since the starting model need not obey the constraints accurately, there is no difficulty in providing suitable starting coordinates, nor in obtaining sufficient convergence from them. The method therefore turns out to be simple to use and free from characteristics which tend to cause failures. This is in a great measure due to the possibility of generating the elements of the observational equations automatically.

The simplicity, power, economy and reliability of the method encourage us to recommend it for general use in protein structure refinement. In combination with methods such as that of Sayre (1972) for generating Fourier maps displaying near atomic resolution, it appears to offer a route to optimally refined models of proteins which requires less tedious precise human interpretation of electron density maps than has been needed up to now.

References

DIAMOND, R. (1966). *Acta Cryst.* **21**, 253–266.

HERMANS, J. & MCQUEEN, J. E. (1974). *Acta Cryst.* A**30**, 730–739.

HESTENES, M. R. & STIEFEL, E. (1952). *J. Res. N.B.S.* **49**, 409–436.

KENNARD, O., WATSON, D. G. & TOWN, W. G. (1970*a*, *b*, 1971, 1973, 1974). *Molecular Structures and Dimensions*, Vol. 1 to 5. Utrecht: Oosthoek, Scheltema & Holkema.

LEVITT, M. (1974). *J. Mol. Biol.* **82**, 393–420.

LEVITT, M. & LIFSON, S. (1969). *J. Mol. Biol.* **46**, 269–279.

SAYRE, D. (1972). *Acta Cryst.* A**28**, 210–212.

Acta Cryst. (1976). A**32**, 315

The Temperature Dependence of the Integrated X-ray Diffracted Intensities from Sodium Metal

BY D. W. FIELD AND B. BEDNARZ

Department of Physics, University of Adelaide, North Terrace, Adelaide, South Australia, 5001

(*Received* 11 *August* 1975; *accepted* 27 *September* 1975)

The temperature dependence of the integrated X-ray diffracted intensities in sodium metal has been determined for the 222, 400, 330, 411 and 332 reflexions in the temperature range 148 K to the melting point of 371 K. In the temperature range 148 K to about 300 K, all the data can be fitted using a quasi-harmonic approximation for the temperature factor. From room temperature to the melting point the intensities for all the reflexions were observed to decrease rapidly with temperature, and could not be fitted either with a quasi-harmonic or fourth-order anharmonic model for the temperature factor. There is no evidence for anisotropy in the intensities below room temperature, but from 293 K to the melting point, anisotropy increases rapidly. A qualitative explanation of the high-temperature phenomena in terms of a lattice relaxation around the vacancies has been advanced.

Introduction

Anisotropy in the room-temperature X-ray structure factors of body-centred cubic sodium has been reported recently (Field & Medlin, 1974, hereinafter referred to as I). In that paper, the integrated intensity data was analysed, following the formalism of Willis (1969), in terms of an anharmonic temperature factor derived from a fourth-order anisotropic expansion of the single-atom potential function appropriate to the body-centred cubic structure. No conclusion could be drawn from the single-temperature experiment about the contribution of possible isotropic anharmonic terms to the temperature factor because the isotropic parameters are strongly correlated with the harmonic parameters in the least-squares analysis. The experiment discussed in the present paper was carried out to investigate both the possible contribution of isotropic