TANNER, L. (1968). Tech. Note 33, Ledgemont Laboratory, Kennecott Copper Corporation, Lexington, Massachusetts.
TANNER, L., CLAPP, P. C. & TOTH, R. S. (1968). *Mater. Res. Bull.* **3**, 855–861.
TIBBALLS, J. E. (1974). Ph. D. Thesis, Univ. of Melbourne, Melbourne.
TOWERS, G. R. (1972). Ph.D. Thesis, Univ. of Melbourne, Melbourne.
WALKER, C. B. & CHIPMAN, D. R. (1970). *Acta Cryst.* A**26**, 447–455.
WALKER, C. B. & KEATING, D. T. (1963). *J. Appl. Phys.* **34**, 2309–2312.
WARREN, B. E., AVERBACH, B. L. & ROBERTS, B. W. (1951). *J. Appl. Phys.* **22**, 1493–1496.
WARREN, B. E. & MOZZI, R. L. (1966). *Acta Cryst.* **21**, 459–461.
WILCHINSKY, Z. W. (1944). *J. Appl. Phys.* **15**, 806–812.
WILKINS, S. (1970). *Phys. Rev.* **2**, B3935–B3942.
WILLIAMS, R. O. (1970). Oak Ridge National Laboratory Report ORNL-TM-2866.
WOLFF, P. M. DE (1956). *Acta Cryst.* **2**, 682–683.

Acta Cryst. (1976). A**32**, 614

A Restrained-Parameter Structure-Factor Least-Squares Refinement Procedure for Large Asymmetric Units

BY JOHN H. KONNERT

Laboratory for the Structure of Matter, Naval Research Laboratory, Washington, D.C., 20375, U.S.A.

(*Received* 2 *September* 1975; *accepted* 27 *October* 1975)

A rapidly converging method for refining approximate atomic models is presented. It combines the conditional structure-factor least-squares procedure described by Waser [*Acta Cryst.* (1963). **16**, 1091–1094] with the conjugate gradient method for solving linear systems [Hestenes & Stiefel, *J. Res. Natl. Bur. Stand.* (1952). **49**, 409–436]. The method allows simultaneous variation of all of the structural parameters, although less than 1% of the derivative matrix need be calculated and stored for large systems and less than $\frac{1}{64}$th of the diffraction data accessible with Cu radiation need be used. Applications involving a 240 atom mineral and an 812 atom protein are mentioned.

Introduction

Approximate atomic models for large structures such as complex minerals and biological macromolecules may be obtained by various means. These methods include consideration of sub-cell symmetry and/or direct methods in the case of complex minerals and isomorphous replacement and anomalous dispersion techniques for macromolecules. The cost of conventional least-squares refinement of these trial structures, which may contain from several hundred to thousands of atoms, is in many cases prohibitive. A notable exception is the refinement of rubredoxin (Watenpaugh, Sieker, Herriott & Jensen, 1973). In some instances, particularly with macromolecules, the intensity data may be too limited. Several alternative approaches have been reported. In the field of protein crystallography, a constrained-model refinement is often employed to give a best fit to an electron-density Fourier map that has been calculated with approximate phases (Diamond, 1971). This model may then be used to compute improved phases from which a new Fourier map may be calculated and the procedure is cycled (Deisenhofer & Steigemann, 1975). Difference Fourier map refinements have also been used (Watenpaugh *et al.*, 1973), and combined with cycles of model idealization (Moews & Kretsinger, 1973; Freer, Alden, Carter & Kraut, 1975). For some minerals, refinement of trial structures employing only distance restraints has proved valuable (Barrer & Villiger, 1969; Meier & Villiger, 1969). These minerals have been of such a size as to permit the models obtained from a least-squares refinement of the distances to be further refined by conventional least squares.

An alternative to these techniques that simultaneously employs both intensity data and distance restraints is presented in this paper. The method is an extension of the conditional structure-factor least-squares technique described by Waser (1963). Subsidiary conditions in this technique are treated as observational equations; *i.e.*, the sum of squared residuals to be minimized is a function of not only observed and calculated intensities, but also ideal and calculated distances. Other subsidiary conditions, such as those involving thermal parameters, may also be included. The extension of the technique utilizes specific properties of the conjugate gradient method for solving linear systems (Hestenes & Stiefel, 1952). Two important features affecting the efficiency of this method are the choice of the elements of the derivative matrix to be retained and the selection of a subset of intensity data. Although less than 1% of the derivative matrix needs to be calculated and stored for large systems, and as little as 1/64th of the intensity data accessible with Cu

radiation used, it is possible to vary simultaneously all of the structural parameters, retain known geometry, and obtain rapid, meaningful convergence.

The procedure has been applied to two structures. Highly twinned terrestrial low tridymite with 240 atoms in space group $P1$ has been refined to a conventional R of 6·4%. The details of the structure will be reported elsewhere (Konnert & Appleman, in preparation). A carp calcium-binding protein that has been previously refined by other techniques (Moews & Kretsinger, 1973) has been treated utilizing the 1370 data with a maximum d spacing of 5 Å and a minimum d spacing of 3 Å. The atomic model used was the same as the initial model for the previous refinement. Two cycles of refinement lowered R for this data from 42 to 20%. Each cycle required 65 min on a CDC 3800 which has a 100 K memory and a 0·9 μs cycle time. Further refinement is in progress and will be reported elsewhere (Konnert, Hendrickson & Karle, in progress).

The object of this paper is to describe the basic principles involved in the technique. Least-squares refinement with subsidiary conditions and the conjugate gradient method for solving linear systems will be reviewed first. Discussion of the derivative matrix elements to be stored, the selection and quantity of intensity data, and relative weighting will follow. Certain details of the refinements of tridymite and the protein will be mentioned.

Least-squares refinement with subsidiary conditions

The reader is referred to the paper by Waser (1963) on this subject. The topic will be discussed here briefly. The function minimized is of the form

$$\theta=\sum_i w_i(|F_o|_i-|F_c|)^2+\sum_l w_l(d_l'^2-d_{c,l}^2)^2 \qquad (1)$$

where i may range over all or just a portion of the intensity data, and l ranges over the distances to be restrained. An ideal distance is designated as d_l' and a calculated one as d_c. The weight assigned to an observation is w. If desired, additional sums derived from observational equations of different types may be included.

The normal equations are given in matrix notation.

$$\mathbf{Ah}=\mathbf{k} \qquad (2)$$

where **A** is the derivative matrix

$$A_{n,m}=\sum_i w_i\frac{\partial F_{c,i}}{\partial x_n}\cdot\frac{\partial F_{c,i}}{\partial x_m}+\sum_l w_l'\frac{\partial d_c^2}{\partial x_u}\cdot\frac{\partial d_c^2}{\partial x_m}+\text{terms from other observational equations.} \qquad (3)$$

$$h_n=\text{desired shift in the } n\text{th parameter} \qquad (4)$$

$$k_n=\sum_i w_i(|F_o|_i-|F_c|_i)\frac{\partial F_{c,i}}{\partial x_n}+\sum_l w_l(d_l'^2-d_{c,l}^2)^2\frac{\partial d_{c,l}^2}{\partial x_n}+\text{terms from other observational equations.} \qquad (5)$$

The index l in equation (3) runs only over those restraining equations that are functions of parameters n and m.

Method of conjugate gradients (c-g method) for solving linear systems

Only those portions of the paper by Hestenes & Stiefel (1952) which are presently used in the refinement technique will be discussed. The c-g method is an algorithm for solving a system $\mathbf{Ah}=\mathbf{k}$ of n linear equations in n unknowns. The solution is given in n iterative steps. However, since each iteration yields a better approximation to the solution, acceptable values for the shifts are obtained in many fewer than n iterations. One starts with an initial estimate of the parameter shifts, x_0 (all usually taken to be zero). Successive iterations determine new estimates $x_{0_n}, x_{1_n}, x_{2_n}\ldots$ of h_n. For the cases of interest here the matrix **A** is symmetric and positive definite. Therefore, the following equations may be used to obtain the solution to **h**. They are equations 3:1a to 3:1f in the paper by Hestenes & Stiefel.

$$p_0=r_0=\mathbf{k}-\mathbf{A}x_0 \quad (x_0 \text{ arbitrary, usually}=0) \qquad (6)$$

$$a_i=\frac{|r_i|^2}{(p_i,\mathbf{A}p_i)} \qquad (7)$$

$$x_{i+1}=x_i+a_ip_i \qquad (8)$$

$$r_{i+1}=r_i-a_i\mathbf{A}p_i \qquad (9)$$

$$b_i=\frac{|r_{i+1}|^2}{|r_i|^2} \qquad (10)$$

$$p_{i+1}=r_{i+1}+b_ip_i \qquad (11)$$

where (x,y) is the scalar product of x and y and $|x|+(x,x)^{1/2}$. An important feature of the c-g method is that the matrix **A** is retained unchanged during the procedure. Thus, only the non-zero elements need be stored and retrieved for the matrix multiplications, $\mathbf{A}p_i$. The elements, $A_{n,m}$, may be stored in any order as long as a scheme is devised for cataloging the indices. It should be emphasized that it is this property of the c-g method that makes possible an efficient conditional structure-factor least-squares refinement technique without employing a full-matrix approach.

Selection of elements of A to be retained

The c-g method is ideally suited for efficiently storing all of the elements related to the restraining equations. For the refinement of a structure of N atoms and M distance restraints, the number of related elements in one half of the symmetric matrix that need to be stored

is $6N+9M$. For the protein refinement involving 812 atoms and 2030 distance restraints, the number of locations required is 23142. Since M is roughly linear with N (typically $M \simeq 3N$), storage space required varies linearly with N. The question arises as to how many additional matrix elements are required in order to insure rapid convergence. Experience has indicated that only the elements involved in the restraining equations need be retained. All other matrix elements are relatively small and may be set equal to zero without hindering convergence. When all of the restraint-related terms are retained in the matrix, the resulting shifts are correlated so as to maintain approximately the known geometry. To ensure rapid convergence, it is necessary to determine the optimum value by which to scale the computed shifts. A convenient method for doing this is to determine the value that minimizes the R value for a small sample selected randomly from the data set. In order to retain the desired geometry, all shifts should be modified by the same factor.

Selection of data subset and weighting

Restraining bond distances and angles greatly reduces the number of structural parameters to be determined by the intensity data. Correlation of the shifts, due to the make-up of the matrix, reduces the problem to one largely of torsions. For this reason, a much smaller intensity subset is required than would be the case without the restraints.

In the initial stages of refinement, when positions are only approximately known, it is advantageous to use a subset of intensity data consisting of a low-angle shell of data. The lower-angle data correspond to larger interplanar spacings, and therefore, the derivatives are valid over a greater range of atomic coordinates. As the refinement proceeds, it is desirable to incorporate higher-angle data both for its greater resolution concerning coordinates and also for its power in determining thermal parameters. A particularly useful way for choosing such a subset appears to be to include the data with both the largest and the smallest $|E|$ values.

The relative weighting among the distance restraints is determined by the degree to which the calculated values are to be forced to the ideal values. Whereas a distance discrepancy of only several hundredths of an Å may be tolerable for covalently bonded distances, as much as several tenths of an Å may be acceptable for the larger distances related to bond angles. Relative weighting of the intensity data and the distance restraints must be monitored at each stage.

One possible method for setting the relative weights is to equate a representative term from the first sum of equation (1) with a term in the second sum. Then, since $\Delta d^2 \simeq 2d\Delta d$,

$$w_i(\Delta F_i)^2 = w_i(\Delta d_i^2)^2 \simeq w_i(2d_i'\Delta d_i)^2. \qquad (12)$$

If in the initial stages of refinement the average $|\Delta F|$ is 50 e, it might be reasonable to equate this $|\Delta F|$ with a $|\Delta d|$ of 0·1 Å, $w_i(50)^2 \simeq w_i(2d_i' \times 0{\cdot}1)^2$. In this way equation (12) affords a basis for setting the relative weighting of intensities and distance restraints.

Refinement of terrestrial low tridymite

The crystals which were investigated appeared to be in the orthorhombic system (Gardner & Appleman, 1974). During the course of the investigation, it became apparent that twinning is responsible for the orthorhombic diffraction pattern and that the true symmetry is lower. Because the refinement technique reported here requires only limited intensity data, it was possible to introduce a generalized twin capability into the refinement without prohibitively increasing the cost. Each intensity was expressed as the sum of four calculated intensities.

$$I_c(hkl) = k_1 I_c'(hkl) + k_2 I_c'(\bar{h}kl) + k_3 I_c'(h\bar{k}l) + k_4 I_c'(hk\bar{l}) .$$

Intensities of the ith twin component are scaled with k_i.

The refinement was then carried out in the space group $F1$ with 240 atoms in the asymmetric unit. A point to note is that all four I_c' components of each observation may refine to the same value if the true symmetry is orthorhombic. They may refine to pairs of equal values if the true symmetry is monoclinic, or they may, as turned out to be the case, all be different if the symmetry is triclinic. The centroid was constrained with three Lagrangian multipliers, *i.e.*, the sum of the shifts of the x coordinates was constrained to zero as were the sums of the y and z shifts. Four thermal factor parameters were used in describing the system. All Si atoms were given the same isotropic B which refined to 0·61 Å^2. A single anisotropic ellipsoid was used to describe the oxygen atoms. The orientation of this ellipsoid for each oxygen atom was fixed by the atomic environment of that atom. The refined values for the axes describing that ellipsoid are 0·60 Å^2 for the axis fixed to be parallel to the line connecting the bridged Si atoms, 2·20 Å^2 for the axis perpendicular to the Si–O–Si plane, and 1·65 Å^2 for the third axis. Although four scale factors were included, all refined to the same value in agreement with the observed orthorhombic diffraction pattern. With the inclusion of one isotropic extinction parameter, a total of 732 parameters were used to describe the model. The 320 Si–O distances were restrained to 1·61 Å, the 480 O–O distances of tetrahedra to 2·63 Å and the 160 shortest Si–Si distances to 3·08 Å. Of the 3280 diffraction data collected, a useful sub-set of data for initial refinement cycles were the $\sim$600 data with $\sin\theta/\lambda < 0{\cdot}67$. After several cycles with $w_i = 1/75^2$ and $w_i = (6/d')^2$, the conventional R was 11% for the 600 data that were refined and 13% for all of the data. The full shifts obtained from the least-squares procedure were utilized although modification by an overall factor has been found at times to improve the convergence. Subsequent refinement employing higher-angle data and larger sub-sets reduced R to 6·4%. The average deviation

from ideal values was ~0·01 Å for the Si–O distances and ~0·02 Å for the others. The details of this structure will be reported elsewhere (Konnert & Appleman, in preparation).

Refinement of calcium-binding protein

As previously stated the 1370 data with d spacing between 5 and 3 Å were used. Since the 812 atom model did not include solvent, the data with d spacing greater than 5 Å were excluded. 2030 observational equations were employed to restrain distances related to bonded distances, bond angles, and planar groups. The model included 2436 positional parameters that were simultaneously varied. The weights for equation (1) were $w_l = 1/75^2$ for all reflections, $w_l = (6/d'_l)^2$ for bonded distances and $w_l = (4/d'_l)^2$ for the others. The first cycle reduced the conventional R from 42 to 30%. A scaling of 0·4 times the calculated shifts was found to be optimal. The second cycle reduced R to 20% with 0·7 scaling of the shifts. The average deviation from 'ideal' values was 0·04 Å for the bonded distances and 0·06 Å for the others. A single overall thermal factor was used. Further refinement is in progress involving restrained thermal parameters and higher-angle data. Full details will be reported elsewhere (Konnert, Hendrickson & Karle, in progress).

The author wishes to thank Dr Jerome Karle for discussions and encouragement, Dr Wayne A. Hendrickson for general discussions and suggestions concerning the c-g procedure, Dr Daniel E. Appleman for discussions, and Dr Robert Kretsinger for kindly affording the use of the protein data and for providing the initial coordinates.

References

BARRER, R. M. & VILLIGER, H. (1969). *Z. Kristallogr.* **128**, 352–370.

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

DIAMOND, R. (1971). *Acta Cryst.* A**27**, 436–452.

FREER, S. T., ALDEN, R. A., CARTER, W. C. JR & KRAUT, J. (1975). *J. Biol. Chem.* **250**, 46–54.

GARDNER, S. P. & APPLEMAN, D. E. (1974). ACA Abstracts, Summer.

HESTENES, M. R. & STIEFEL, E. (1952). (1952). *J. Natl. Bur. Stand.* **49**, 409–436.

MEIER, W. M. & VILLIGER, H. (1969). *Z. Kristallogr.* **129**, 411–423.

MOEWS, P. C. & KRETSINGER, R. H. (1975). *J. Mol. Biol.* **91**, 201–228.

WASER, J. (1963). *Acta Cryst.* **16**, 1091–1094.

WATENPAUGH, K. D., SIEKER, L. C., HERRIOTT, J. R. & JENSEN, L. H. (1973). *Acta Cryst.* B**29**, 943–956.