Acta Cryst. (1974). A**30**, 180

Least-Squares Phase Refinement. II. High-Resolution Phasing of a Small Protein

BY D. SAYRE*

Mathematical Sciences Department, IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598, *U.S.A.*

(*Received* 2 *August* 1973; *accepted* 17 *September* 1973)

A recently described technique for direct least-squares refinement of phases has been successfully applied to the problem of phasing a 1·5 Å data set of observed structure-factor magnitudes for *C. pasteurianum* rubredoxin (M.W. $\simeq$6100), given as a starting point a set of phases to 2·5 Å resolution as determined from heavy-atom derivatives. The result is a 1·5 Å X-ray structure for rubredoxin which generally confirms that recently obtained by a method involving the production and refinement of an approximate atomic model of the protein. In the present work the structure was obtained without chemical assumptions and with a considerable saving of effort compared with the previous determination. The technique also avoided certain small errors in the assignment of side-group structures made in the previous determination. Final high-resolution phasing would result from conventional refinement of the direct-method structure. An additional point of interest is that the structure of a protein at the atomic level, elucidated on purely physical principles, has been found to agree with the deductions from chemistry.

A difficulty in the use of X-ray analysis in the study of protein structure at an atomic level of resolution has been the inability of the heavy-atom phasing methods to provide phases for the X-ray structure factors beyond approximately 2·0 or 1·9 Å resolution. Watson *et al.* (1963) appear to have had a limited success in the phasing of high-resolution myoglobin data by refinement of an atomic model based upon study of a 2 Å experimentally phased electron-density map. More recently Watenpaugh, Sieker, Herriott & Jensen (1973) have succeeded with this technique in the case of *C. pasteurianum* rubredoxin. Even here, however, subsequent chemical sequencing (McCarthy, 1972) indicates that the model refined was slightly in error in the identity of a few of the side groups. The difficulty is essentially one of producing an atomic model from experimentally phased maps of a resolution not really sufficient for the purpose.

Data for the present study consisted of the Watenpaugh *et al.* data set for rubredoxin, containing 5033 structure-factor amplitudes for rubredoxin to 1·54 Å resolution as measured by those workers at the University of Washington and kindly supplied to the author by them. Magnitudes of the F's appearing in the system of equations

$$a_{\mathbf{h}}F_{\mathbf{h}}=\sum F_{\mathbf{k}}F_{\mathbf{h}-\mathbf{k}} \tag{1}$$

were produced by placing the data set on an absolute scale (simultaneously adding F_{000}) and processing it to correspond to Gaussian atoms of shape $\exp(-4r^2)$. Magnitudes of the cross products of these terms appearing in equations (1) were computed and saved for later use.

Rubredoxin occurs in space group $R3$, with hexagonal cell parameters $a=b=64{\cdot}45$, $c=32{\cdot}68$ Å.

A set of 1608 experimentally derived phases to 2·5 Å resolution, based on heavy-atom derivatives of rubredoxin, was also kindly supplied by the investigators at the University of Washington. Insertion of these into the right-hand sides of equations (1) provided a rough initial extension of the phases from 2·5 to 1·5 Å.

The entire set of phases (with the exception of 000 and 003) was then subjected to 11 cycles of direct phase refinement, as recently described (Sayre, 1972). In this type of refinement the phases are adjusted to produce a minimum in the value of the expression

$$R=\sum|a_{\mathbf{h}}F_{\mathbf{h}}-\sum F_{\mathbf{k}}F_{\mathbf{h}-\mathbf{k}}|^2.$$

Since R is a measure of the degree to which equation-system (1) is not satisfied, the minimization process causes the phases to draw as nearly as possible to a solution of those equations. The time per cycle of refinement was approximately 50 min on a 360 Model 91 computer.

A few details may be added at this point. The 5033 structure factors in the Watenpaugh *et al.* data set are those, out of the 7345 which lie within the 1·54 Å sphere, which had peak counts greater than 2σ. The scale factor and average temperature factor for the data set were determined by requiring that the data set, after application of these factors, should give a Wilson plot resembling as nearly as possible the plots calculated for several artificial structures containing 436 point atoms of unit weight* placed in the asymmetric unit of the rubredoxin unit cell. The $a_{\mathbf{h}}$ in equations (1) are given by $V(p-qR-rR^2)\ (f^{sq}/f)$, where $V=$ volume of the unit cell, $R=|\mathbf{h}|$, p, q, r are parameters (Sayre, 1972) which depend upon the degree of in-

* Temporary address: Laboratory of Molecular Biophysics, Department of Zoology, South Parks Road. Oxford, England.

* The actual number of non-hydrogen atoms in rubredoxin is 424. The information available to Watenpaugh *et al.* at the outset of their high-resolution work, however, suggested that the number was 436, and it was thought to be fairest, for the purposes of the present work, to adopt the same number.

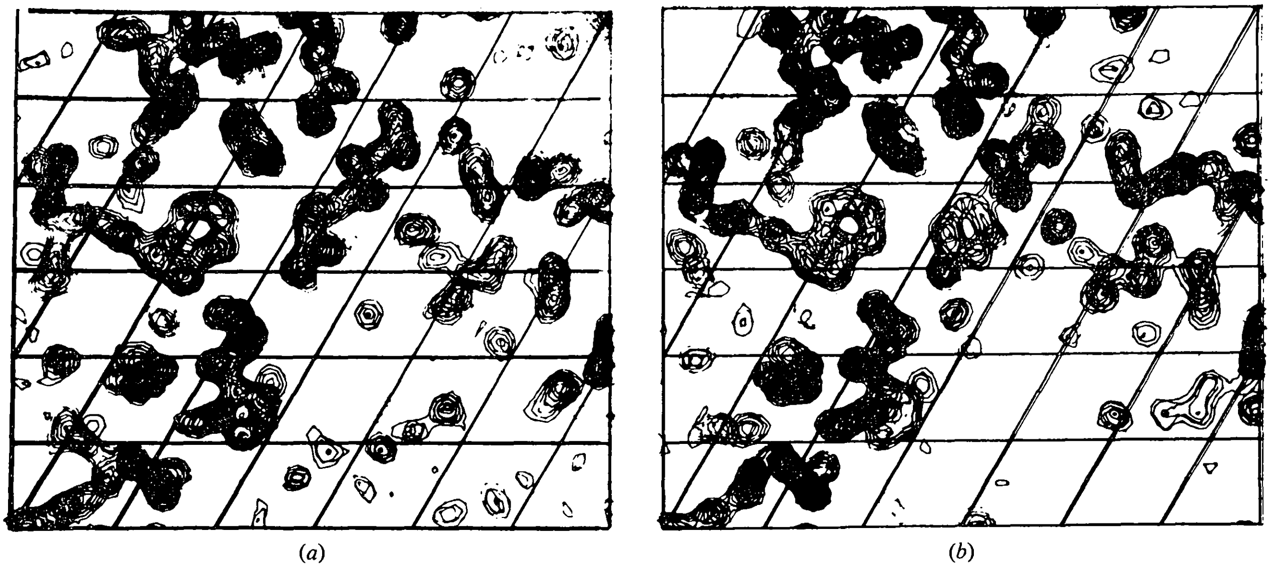

Fig. 1. Rubredoxin, sections $z=17/60$ through $23/60$. (*a*) 1·5 Å, direct-method phases. (*b*) 1·5 Å, phases obtained by Watenpaugh *et al.* (*c*) 2·5 Å, heavy-atom phases. The maps were calculated using the structure-factor amplitudes as prepared for use in the direct-method procedure (*i.e.* amplitudes are considerably sharpened).

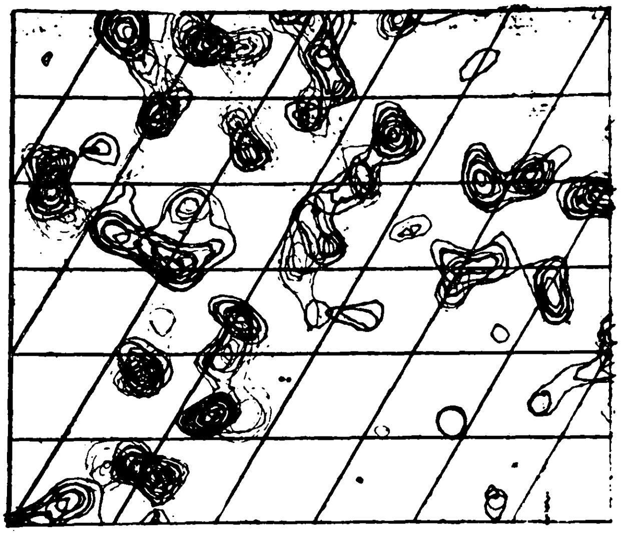

Fig. 1 (*cont.*)

completeness of the data set and which in the present case were found, by calculations on artificial structures with similar data incompleteness, to be 0·737, 0·085, and 0·716 respectively, and $f^{sq}/f=8^{-1/2}$ exp $(\pi^2R^2/8)$ for the Gaussian atoms employed. The magnitude and other information precomputed for each F_kF_{h-k} cross product were stored in a data set of 14 magnetic tapes (with data packing this could have been reduced to seven). In the refinement the initially given 1608 heavy-atom phases were permitted to change only slowly in the first few cycles. In addition, the equations arising from the 29 structure factors out to 10 Å resolution were not used in the refinement, as being probably strongly influenced by the presence of solvent. The value of $\sum|a_hF_h-\sum F_kF_{h-k}|^2$ after 11 cycles was $1{\cdot}09\times10^{15}$. The resulting phase set, despite a mean difference of 46·6° compared with the 1·5 Å phase set produced by Watenpaugh *et al.* after four cycles of ΔF syntheses and four of least-squares refinement (see Table 2 of their paper), provides a readily interpretable electron-density map. Several factors may enter into the rather large phase difference, including the tendency of the direct method to produce an equi-atom structure, whereas the Watenpaugh *et al.* phases, following the real structure, reflect the presence of an Fe and several S atoms, plus C, N, and O atoms with a large range of thermal motions. When a set of phases was produced by minimization starting from the Watenpaugh *et al.* phase set itself, the mean phase difference was 31·7°, although the electron-density map showed little obvious change except in peak heights.

Fig. 1(*a*) shows a portion of the 1·5 Å electron-density map resulting from the extension and refinement. For comparison, Fig. 1(*b*) shows the same portion of the 1·5 Å map with phases as obtained by Watenpaugh *et al.* Fig. 1(*c*) shows the starting point for the process of extension and refinement, *i.e.* the 2·5 Å electron-density map containing the 1608 terms with heavy-atom phases. All three maps were computed identically and contoured at the same absolute levels. Details of the maps are compared in Figs. 2 and 3.

In general quality the direct-method map is not quite as clean as the Watenpaugh *et al.* map, but it is nevertheless sufficient to allow approximately 400 of the 424 atoms in the molecule to be correctly located. This figure includes 356 atoms (205 main-chain and 151 side-group) which can be seen quite clearly in the map and 44 atoms (11 main-chain and 33 side-group) which are not seen so clearly but which can be correctly filled in from the indications of the map and the positions of other atoms. There are 24 atoms (three main-chain atoms at the ends of the chain, and 21 side-group atoms belonging to 10 residues) which are unassignable or doubtfully assignable on the basis of the map.

In terms of side groups, 36 can be seen in their entirety and eight more (11, 17, 32, 40, 46, 48, 50, 51) become assignable on the basis of filling in. As stated above there are 10 residues (1, 2, 3, 14, 16, 21, 31, 47, 53, 54) in which there is at least one atom which is doubtful or unassignable and whose side-group structure cannot therefore be entirely decided from an examination of the map. There are few if any atoms in false sites, the phase refinement appearing to have the property that while it may sometimes cause an atom to be omitted it seldom creates a false one.

Of the 54 residues of rubredoxin, six have a particular interest here in that their side-group structures have been shown, by the subsequent completion of the chemical sequencing of rubredoxin, to have been slightly incorrectly assigned* in the atomic model formulated for refinement by Watenpaugh *et al.* The details are summarized in Table 1. At five of these six residues the side-group structure indicated by the direct-method map is in agreement with the chemical evidence [see for example Figs. 2(*h*) and 2(*i*)]. The remaining side group (21) is one of those which is uncertain in the direct-method map.

Table 1. *Six side-group structures*

	Seattle model (*a*)	Chemical sequencing (*b*)	Direct method
8	Ile	Val	*b*
12	Val	Ile	*b*
21	Glu	Asp	
24	Ile	Val	*b*
41	Ile	Leu	*b*
44	Ile	Val	*b*

The effect of the direct method in driving the structure toward equality of atoms can be noticed, for example, in Fig. 2(*f*), where the Fe atom and one of the 4 cysteine S atoms to which it is bonded are shown; it is evident, by comparison with other portions of the structure shown in that Figure, that the Fe and S peak heights are not much greater than the average peak height.

In addition to the features already discussed, the direct-method map shows approximately a dozen fairly strong peaks lying in the solvent part of the structure. These all occur close to the surface of the molecule at sites where a significant degree of attachment of solvent might be expected, and with one exception they coincide with strong peaks in the Watenpaugh *et al.* map. It appears likely, therefore, that certain of the more pronounced features of the solvent structure are reflected in the direct-method phases.

Attempts to carry out extensions from 3 Å (967 heavy-atom phases) produced maps which although nicely atomic do not appear to be interpretable in terms of protein structure, indicating that the initial

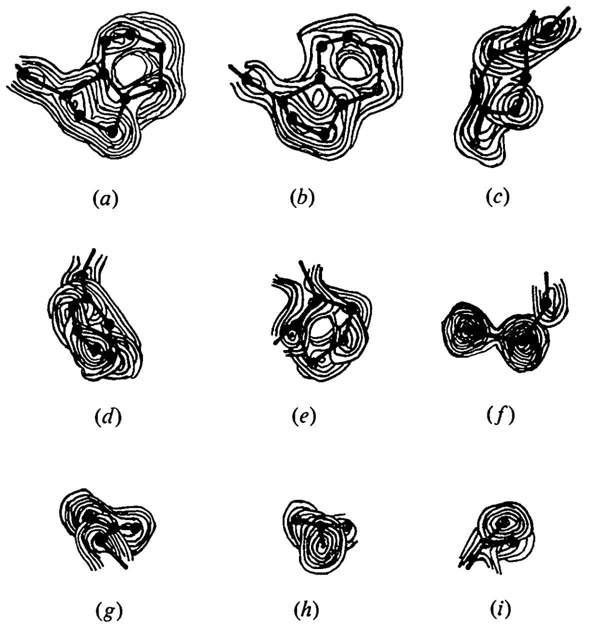

Fig. 2. Various side groups. All are taken from the direct-method map except (*b*), which is from the map using the phases of Watenpaugh *et al.* (*a*) Trp 37. (*b*) Trp (37). (*c*) Tyr 11. With an atom missing, this is one of the more poorly rendered groups. The Watenpaugh *et al.* map gives very good definition for this group; see their Fig. 3. (*d*) Phe 49. (*e*) Pro 20. (*f*) Cys 9, showing also one of the Fe–S bonds. The peak heights of the Fe and S atoms are lower than they would be in nature, reflecting the attempt on the part of the phase refinement to produce an equi-atom structure. (*g*) Asp 35. (*h*) Val 8. (*i*) Val 44. In (*a*)–(*g*) the atomic coordinates shown are those of Watenpaugh *et al.* (their Table 3). In (*h*) and (*i*) their coordinates, which are for isoleucine, are not used.

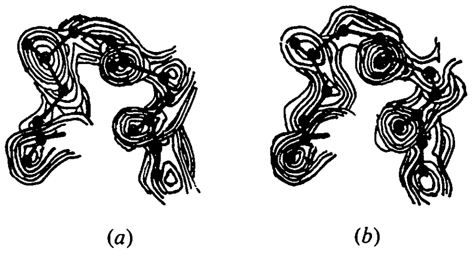

Fig. 3. A section of main chain, running from residue 29 (lower right) to residue 32 (lower left). (*a*) Direct-method map. (*b*) Map using the phases of Watenpaugh *et al.* Atomic positions taken from their Table 3.

* A private communication from L. H. Jensen indicates that the correct identities of residues 8, 12, 24, 41, 44 had in fact become known to the University of Washington investigators by the conclusion of their work.

extension process had in these cases started the minimization on the walls of non-physical, or more accurately non-chemical, minima. Fig. 4 shows the result of one such attempt, to be compared with Fig. 1(*a*) and (*b*). Starting the process from 2 Å (2813 heavy-atom phases), on the other hand, produced a result similar to that obtained by starting from 2·5 Å.

In a minimization problem with approximately 5000 parameters it is not at present feasible to form and solve the system of normal equations $A^T A\mathbf{d} = -A^T\mathbf{r}$ from which the increment **d** to be applied to the parameters is commonly derived.[*] One approach to reducing the cost of the minimization is to ignore everything in the normal equations except the terms on the principal diagonal. This technique greatly reduces the size of the computation but also lowers the rate of convergence considerably; it nevertheless accomplished the minimization successfully in several refinements on the rubredoxin data. A better approach,[†] however, is based upon the observation that **d** is also the vector which minimizes $\|A\mathbf{d}+r\|$, and that even a few steps of a conjugate-gradient minimization process (Hestenes & Stiefel, 1952) will develop a fairly good approximation to **d**. This technique, with five steps of conjugate-gradient minimization, was used in the refinements from 2·5 Å and 2 Å; it costs more per cycle, but shows rapid and steady convergence, as may be seen in Fig. 5. Using the latter technique, the complete extension and refinement process for rubredoxin at a typical large computing installation would cost today approximately $7500.

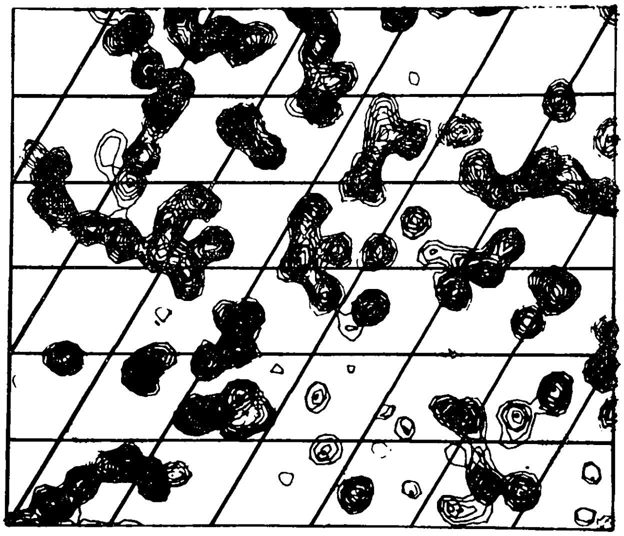

Fig. 4. Rubredoxin, sections $z = 17/60$ through 23/60. Non-chemical structure resulting from attempt to extend and refine phases from 3 Å heavy-atom phases.

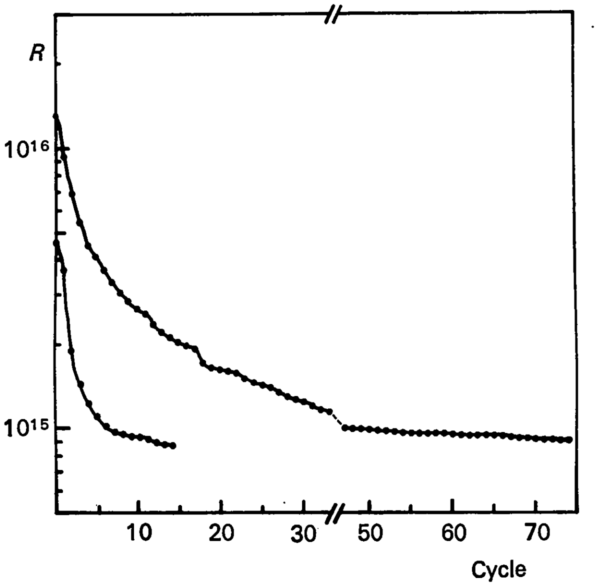

Fig. 5. R as a function of the number of cycles of refinement. Upper curve: principal diagonal approximation, extension from 3 Å, time per cycle 25 min (360 Model 91). Lower curve: conjugate-gradient technique, extension from 2 Å, time per cycle 50 min.

Summary

High-resolution phasing of protein data by the direct method studied yields phases which appear to have slightly more random noise superposed than phases derived from an atomic model, but it is likely also to produce fewer systematic errors arising from mistaken positions. This suggests that a two-stage process may provide a path for the most accurate high-resolution phasing for proteins: use of the direct method for producing the initial high-resolution maps, followed by conventional refinement of the resulting structural model. This process has also the advantage of rapidity, objectivity, and convenience, and would reduce considerably the number of intensities required to be collected from the heavy-atom derivatives of the protein.

Because of the size of the minimization problem involved, the practical limit for this type of phase refinement probably lies at present in the 15–30000 molecular weight range.

The author wishes to express his thanks to Professor Dorothy Hodgkin for her kind invitation to spend a year with her group at the Laboratory of Molecular Biophysics, Oxford, where the analysis of the direct-method maps was carried out, and to Professor L. H.

[*] Let $w_\mathbf{h} = u_\mathbf{h} + iv_\mathbf{h} = a_\mathbf{h}F_\mathbf{h} - \sum F_\mathbf{k}F_{\mathbf{h}-\mathbf{k}}$, $\mathbf{h} = \mathbf{h}_1, \ldots, \mathbf{h}_R$. Let $r_{2j-1} = u_{\mathbf{h}j}$ and $r_{2j} = v_{\mathbf{h}j}$. Then $\mathbf{r}$ is the $1 \times 2R$ vector of residuals (r_i). Let the parameters (phases) be φ_i, $i = 1, \ldots, P$. Then A is the $P \times 2R$ Jacobian matrix $(\partial r_i/\partial\varphi_j)$. In the work on rubredoxin, R and P were approximately equal ($R = 5005$, $P = 5032$). A (probably) preferable procedure would have employed the reflections observed to be weaker than 2σ to increase R to 7317, thereby increasing the number of observational equations. Improvement could also have been secured by making use of a non-uniform weighting scheme ($R = \sum w_\mathbf{h}|a_\mathbf{h}F_\mathbf{h} - \sum F_\mathbf{k}F_{\mathbf{h}-\mathbf{k}}|^2$).

[†] Suggested to the author by Dr Philip Wolfe of the IBM Research Center, Yorktown Heights, New York.

Jensen and his colleagues at the University of Washington for their help and interest in this study and for their generosity in allowing the use of their data on rubredoxin.

References

Hestenes, M. R. & Stiefel, E. (1952). *J. Res. Natl. Bur. Std.* **49**, 409–436.

McCarthy, K. (1972). Ph.D. Thesis, George Washington Univ.

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

Watenpaugh, K. D., Sieker, L. C., Herriott, J. R. & Jensen, L. H. (1973). *Acta Cryst.* B**29**, 943–956.

Watson, H. C., Kendrew, J. C., Coulter, C. L., Brändén, C.-I., Phillips, D. C. & Blake, C. F. (1963). *Acta Cryst.* **16**, A81.

Acta Cryst. (1974). A**30**, 184

Thermal Diffuse Scattering Corrections for Single-Crystal Integrated Intensity Measurements

By E. D. Stevens*

Department of Chemistry, University of California, Davis, California 95616, *U.S.A.*

(*Received* 30 *July* 1973; *accepted* 15 *September* 1973)

A computer program has been developed for calculating one- and two-phonon thermal diffuse scattering corrections for integrated X-ray intensity measurements. The correction includes the anisotropy in the diffuse-scattering intensity distribution and the geometry of the scan for crystals of any symmetry type. The calculated two-phonon correction is not negligible and may be as large as the one-phonon correction for high-order reflections. The effects of slit size, scan range, crystal orientation, crystal misalignment, and neglect of phonon dispersion on the calculated corrections are investigated.

Introduction

The attainment of accurate structure factors from X-ray intensity measurements requires in many cases a correction for the thermal diffuse scattering (TDS) included in intensity scans. The difficulty of the calculations has prevented the routine application of TDS corrections in X-ray crystallography. The general approach to calculating TDS corrections and some of the approximate methods that have been used are described by Cochran (1969).

Until recently, most approximate methods have involved the assumption of a spherical TDS distribution about the Bragg reflection. Rouse & Cooper (1969) have developed a program to calculate the one-phonon correction which correctly includes the anisotropic TDS intensity distribution for crystals of any symmetry. Walker & Chipman (1970) have written two programs for calculating the one-phonon TDS correction which are restricted to cubic crystals. The first program (Walker & Chipman, 1970, 1971*a*) includes the primary-beam divergence, wavelength distribution, and anisotropy of the scattering. The second program (Walker & Chipman, 1970, 1971*b*) neglects the primary-beam divergence, but includes a simplification in the calculation which makes the program fast enough to be used routinely with intensity measurements.

In this paper a TDS correction program for θ: 2θ scans is described which is similar to the faster program of Walker & Chipman but is not restricted to cubic crystals. In addition, an approximate correction for two-phonon TDS intensity is included. The calculation of the two-phonon correction has also been simplified and requires little additional effort. The program has been used to investigate the effects of scan range, slit size, and crystal orientation and misalignment on the corrections.

Theory

For acoustic phonons, neglecting dispersion, primary-beam divergence, and mosaic spread, the ratio of one-phonon-included TDS intensity to the Bragg intensity for a scan is given by

$$\alpha_1 = \frac{I_1}{I_B} = \frac{k_B T}{v} \sum_i w_i \int J_1(\mathbf{q}) d^3q \tag{1}$$

where the integration is over the volume in reciprocal space swept out by the scan and

$$J_1(\mathbf{q}) = \frac{1}{q^2} \sum_{j=1} \frac{[\mathbf{H} \,.\, \mathbf{e}_j(\mathbf{q})]^2}{\varrho V_j^2(\mathbf{q})} \,. \tag{2}$$

Here $\mathbf{H}$ is the scattering vector, $q=|\mathbf{q}|$, k_B is Boltzman's constant, T is the temperature, ϱ is the density of the crystal, v is the unit-cell volume, and $V_j(\mathbf{q})$ is the velocity of the acoustic lattice wave $\mathbf{q}$. The $\mathbf{e}_j(\mathbf{q})$ ($j=1,2,3$) are unit vectors in the directions of polarization of the lattice wave.

* Present address: Department of Chemistry, State University of New York, Buffalo, N.Y. 14214, U.S.A.