(A.3) reduces to

$$|F(\mathbf{H})|^2_{\rm exp} = \varepsilon(\mathbf{H})^2 \sum_{p=1}^{m/\varepsilon_{\mathbf{H}}} |S_p(\mathbf{H})|^2$$

$$= \varepsilon(\mathbf{H}) \sum_{p=1}^{m} |S_p(\mathbf{H})|^2. \qquad (A.5)$$

Now imagine that the complete molecular group of size N/m is replaced by a smaller fragment of size n, so that the unit cell also contains $m(N/m-n)$ randomly distributed atoms. If (A.5) is modified to include their contribution, it follows that

$$|F(\mathbf{H})|^2_{\rm exp} = \varepsilon(\mathbf{H}) \sum_{j=1}^{N-nm} Z_j^2 + \varepsilon(\mathbf{H}) \sum_{p=1}^{m} |S_p(\mathbf{H})|^2. \qquad (A.6)$$

Finally, expression of $|S_p(\mathbf{H})|^2$ in terms of the interatomic vectors $\mathbf{r}_j-\mathbf{r}_k$ results in the desired expression (4).

References

Amigó, J. M., Ochando, L., Abarca, B., Ballesteros, R. & Rius, J. (1992). *Mater. Sci. Forum.* In the press.

Declercq, J. P., Germain, G. & Van Meerssche, M. (1972). *Cryst. Struct. Commun.* **1**, 13-15.

Grigg, R., Kemp, J., Sheldrick, G. & Trotter, J. (1978). *J. Chem. Soc. Chem. Commun.* pp. 1109-1111.

Harada, Y., Lifchitz, A., Berthou, J. & Jolles, P. (1981). *Acta Cryst.* A**37**, 398-406.

Hovmöller, S. (1980). *Rotation Matrices and Translation Vectors.* IUCr/Univ. College Cardiff Press.

Jones, P. G., Sheldrick, G. M., Glüsenkamp, K. H. & Tietze, L. F. (1980). *Acta Cryst.* B**36**, 481-483.

Rius, J. & Miravitlles, C. (1986). *Acta Cryst.* A**42**, 402-404.

Rius, J. & Miravitlles, C. (1987). *J. Appl. Cryst.* **20**, 261-264.

Rius, J. & Miravitlles, C. (1988). *J. Appl. Cryst.* **21**, 224-227.

Rius, J., Miravitlles, C., Molins, E., Crespo, M. & Veciana, J. (1990). *Mol. Cryst. Liq. Cryst.* **187**, 155-161.

Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24-31.

Stewart, J. M., Karle, J., Iwasaki, H. & Ito, T. (1977). *Acta Cryst.* A**33**, 519.

Stubbs, M. & Huber, R. (1991). *Acta Cryst.* A**47**, 521-526.

Tollin, P. & Cochran, W. (1964). *Acta Cryst.* **17**, 1322-1324.

Wilson, C. C. & Wadsworth, J. W. (1990). *Acta Cryst.* A**46**, 258-262.

Acta Cryst. (1993). A**49**, 504-513

Molecular Dynamics in Refinement against Fiber Diffraction Data

By Hong Wang and Gerald Stubbs

Department of Molecular Biology, Vanderbilt University, Nashville, TN 37235, *USA*

(*Received* 5 *June* 1992; *accepted* 22 *October* 1992)

Abstract

The molecular dynamics (MD) method has been adapted for refinement of the structures of helical macromolecular aggregates against X-ray fiber diffraction data. To test the effectiveness of the method, refinements of the tobacco mosaic virus structure were carried out against a set of simulated fiber diffraction intensities using the MD method as well as the conventional restrained least-squares (RLS) method. The MD refinement converged to a very low R factor and produced a structure with generally satisfactory stereochemistry, while the RLS refinement was trapped at a local energy minimum with a larger R factor. Results suggest that the effective experimental radius of convergence of the MD method is significantly greater than that of the RLS method. Even when the initial structure is too far from the true structure to allow direct refinement, the MD method is able to find local minima that resemble the true structure sufficiently to allow improved phasing and thus lead to interpretable difference maps for model rebuilding.

Introduction

Fiber diffraction has been a very effective method for the determination of the molecular structures of filamentous macromolecular assemblies such as viruses, cytoskeletal elements, nucleic acids and polysaccharides. The component parts of these assemblies are often difficult or impossible to crystallize because of their natural tendency to form filaments and, even if they can be crystallized, the crystal structures do not usually reveal the important intermolecular interactions that are often the most biologically significant aspect of the molecular structure. Fiber diffraction is therefore the preferred method of analysis for these systems.

The defining property of a fiber diffraction specimen is that the diffracting units are randomly oriented about an axis, the fiber axis. Specimens may in fact be fibers, oriented gels or even stacks of sheet-like structures such as membranes. As a result of the random orientation about the axis, the fiber diffraction pattern is the cylindrical average of the diffraction pattern to be expected from one particle (in the

0108-7673/93/030504-10$06.00

absence of interference effects) or from a fully ordered array of particles (in the case of a crystalline fiber). The amount of information lost because of this averaging depends on the size and symmetry of the diffracting particles and on the resolution of the data. For tobacco mosaic virus (TMV), the effective number of observable diffraction data at 3 Å resolution is reduced by a factor of about 2.5; for the bacteriophage Pf1 at 3 Å resolution, the corresponding factor is 1.7 (Makowski, 1982). These factors can be much higher for lower-symmetry systems such as microtubules but for relatively symmetric systems such as helical viruses they are not so high as to preclude structural analysis and refinement.

Analysis of fiber diffraction patterns may conveniently be considered in two stages: the separation of the cylindrically averaged intensities and determination of the corresponding phases to obtain an initial model, and the refinement of that model to maximize agreement between the calculated and observed diffracted intensities. The first stage may be omitted in cases where the diffracting asymmetric unit is of relatively low molecular weight (for example, nucleic acids and polysaccharides); in these cases, an effective approach has been to postulate one or more models and to refine competing models separately. This approach is completely impracticable, however, for large structures such as viruses. Solutions better suited to large asymmetric units have included a multidimensional analog of protein crystallographic isomorphous replacement (Stubbs & Diamond, 1975; Namba & Stubbs, 1985), sometimes supplemented by data from the fine splitting of the layer lines in the diffraction patterns (Stubbs & Makowski, 1982), and the use of neutron scattering from proteins with specific amino acids deuterated (Nambudripad, Stark & Makowski, 1991).

The principal concern in the refinement of structures against fiber diffraction data is, as one might expect, the relatively small number of independent data available. Unrestrained refinement does not lead to stable stereochemically reasonable solutions, so it is necessary to incorporate stereochemical information such as bond lengths and angles into the refinement procedure. Two algorithms in particular have been effective: the linked-atom least-squares (LALS) method of Arnott and his collaborators (Arnott & Wonacott, 1966; Smith & Arnott, 1978) has been extensively used in the refinement of nucleic acids and polysaccharides, while the restrained least-squares (RLS) method of Hendrickson and Konnert (Hendrickson, 1985), which has been widely used in protein crystallography, has been adapted for use with fiber diffraction data (Stubbs, Namba & Makowski, 1986) and used in the determination of the structures of TMV (Namba, Pattanayek & Stubbs, 1989) and Pf1 (Nambudripad & Makowski, 1989). These refinement methods have met with considerable success but they suffer from the disadvantage that successful refinement depends heavily on the accuracy of the starting model. In practice, we have found that the radius of convergence of RLS refinement is significantly less than it is in crystallography. Even in crystallography, the limited radius of convergence of RLS refinement is a problem (Brünger, Kuriyan & Karplus, 1987) and for fiber diffraction, in which the data sets are significantly smaller than crystallographic data sets from comparably sized structures, the problem appears to be much more serious.

In recent years, the use of molecular dynamics in conjunction with refinement against diffraction data has been very effective in crystallography (Brünger, Kuriyan & Karplus, 1987; Brünger, Karplus & Petsko, 1989). In particular, molecular dynamics refinement has been reported to increase significantly the radius of convergence for refinement of crystallographically determined protein structures (Brünger, Kuriyan & Karplus, 1987). In molecular dynamics simulations, Newton's equations of motion are solved for the atoms in a molecule, using forces derived from potential functions that describe the bonding and nonbonding interactions between the atoms. X-ray diffraction data can be used as effective additional potential terms, thus restraining the structure to agree with the observed data. In the now widely used procedure of simulated-annealing refinement (Brünger, Kuriyan & Karplus, 1987), the energy of the protein structure is minimized then the process of heating the protein is simulated. At high temperatures, energy barriers between the starting model and structures of lower potential can be overcome; in this way, the radius of convergence of the refinement is increased. Finally, the structure is cooled ('annealed'). In this paper, we describe an adaptation of the simulated-annealing application of molecular dynamics, using the program *X-PLOR* (Brünger, Kuriyan & Karplus, 1987), for use with fiber diffraction data.

Theory

Molecular dynamics refinement

The potential energy of a molecule may be described by an empirical energy function, for example

$$E = E_b + E_a + E_t + E_{\text{nb}}, \tag{1}$$

where E_b is the potential energy due to deviations from ideal covalent bond lengths, E_a is due to deviations from ideal covalent bond angles, E_t (torsion) is due to rotations about bonds and E_{nb} is due to interactions between non-bonded atoms (Brooks *et al.*, 1983; Karplus & Petsko, 1990). E_b and E_a are usually modeled by simple harmonic potentials,

$$E_b = \sum k_b(r-r_0)^2,$$

$$E_a = \sum k_a(\theta-\theta_0)^2,$$

where the summations are over all bonds and bond angles, r and θ are the bond lengths and angles, and r_0 and θ_0 are ideal bond lengths and angles. E_t for rotations (φ) about single bonds is modeled by a cosine function such as

$$E_t = \sum k_t[1+\cos(n\varphi)],$$

where n is a small integer, depending on the nature of the atoms forming the bond. For restricted rotations, contributions to E_t may be modeled as harmonic terms; these terms ensure planarity of rings, peptide bonds and other fixed geometric structures. The constants k_b, k_a and k_t (force constants) determine the flexibility of the corresponding molecular features. $E_{\rm nb}$, the non-bonded interaction term, includes contributions from electrostatic interactions and from van der Waals interactions. One of the simplest representations of $E_{\rm nb}$ is

$$E_{\rm nb} = \sum (A/r^{12} - B/r^6 + q_1q_2/Dr),$$

in which r is the distance between atoms, q_1 and q_2 are the electrostatic charges of the atoms, D is an effective dielectric constant and A and B are constants depending on the interacting atoms. This term combines the Lennard-Jones 6–12 potential (having a minimum at the sum of the van der Waals radii of the two atoms) with the electrostatic attraction or repulsion between the atoms.

E is a function of the atomic coordinates of the molecule under consideration. A molecular dynamics simulation of the atomic motions may be carried out by solution of Newton's equations of motion for the atoms using forces derived from E. The energy E may be minimized as a function of the atomic coordinates to find the most stable molecular conformation. E is, of course, a very complicated function and without more information a molecule as complex as a protein would generally find its way into a local minimum in the potential-energy function, rather than the global minimum. This problem can often be overcome by adding 'effective potential energy' terms to E, in which the difference between experimental observations and their values calculated from the atomic coordinates is considered to contribute to the potential energy. The use of X-ray crystallographic diffraction data in this way has been particularly effective for the refinement of the structures of crystalline proteins. In the program *X-PLOR* (Brünger, Kuriyan & Karplus, 1987), the effective potential-energy term

$$E_{\rm crystal} = S\sum[F_{\rm obs}(hkl) - F_{\rm calc}(hkl)]^2 \qquad (2)$$

is added to E in the program *CHARMM* (Brooks *et al.*, 1983). The summation is over observed and calculated crystallographic structure factors; the scale factor S is chosen so that the gradient of $E_{\rm crystal}$ is comparable to the gradient of E in (1).

Simulated annealing, which involves minimization of potential-energy terms such as those in (1) and (2), has been used with great effect in the refinement of the structures of numerous crystalline proteins in recent years (Karplus & Petsko, 1990).

Fiber diffraction theory

Fiber diffraction specimens give rise to X-ray diffraction patterns characterized by layer lines, indexed by l, whose separation depends inversely on the length c of the repeating structural unit along the fiber axis. A particularly good example is shown in Fig. 1 of Namba & Stubbs (1985). Because of the cylindrical averaging, a fiber diffraction pattern is two-dimensional: all of the information is contained in a plane of reciprocal space. In the absence of interparticle interference effects, the diffracted intensities are distributed continuously along the layer lines.

The intensity at reciprocal-space radius R on layer line l is

$$I(R,l) = \mathscr{G}(R,l)^2 = \sum_n G_{n,l}(R)G^*_{n,l}(R) \qquad (3)$$

(Waser, 1955; Franklin & Klug, 1955), where n is the order of the Bessel functions J_n that contribute to the complex Fourier–Bessel structure factor G (Klug, Crick & Wyckoff, 1958). The number of significant terms contributing to the intensity in (3) is limited by the selection rule (see below) and depends on the symmetry and dimensions of the diffracting particle and on the value of (R, l). For example, for TMV at 2.9 Å resolution, there can be as many as eight terms. As a convenient notation, the vector $\mathscr{G}$ is defined as the $2M$-dimensional vector whose components are the real and imaginary components of the M significant G terms contributing to a particular intensity $I(R, l)$ in (3).

Equation (3) can be compared with the crystallographic equation

$$I(h,k,l) = F^2_{hkl} = F_{hkl}F^*_{hkl}. \qquad (4)$$

For a helical structure, the integer n is restricted by the selection rule $l = tn + um$, where m is an integer and there are u subunits in t turns of the helix (Cochran, Crick & Vand, 1952). If the symmetry of the helical structure includes an N-fold rotation about the helix axis, n is further restricted to be a multiple of N (Klug, Crick & Wyckoff, 1958). Helical symmetry has been discussed by Klug *et al.* (1958).

The Fourier–Bessel structure factor can be expressed in terms of the atomic coordinates (r, φ, z),

$$G_{n,l}(R) = \sum_j f_j J_n(2\pi r_j R)\exp(-\varphi_j n + 2\pi l z_j/c), \qquad (5)$$

where f_j is the scattering factor for atom j and the summation is over all atoms in the helical asymmetric unit (Klug *et al.*, 1958). If each G is known, an electron density map may be obtained from the relationships

$$\rho(r, \varphi, z) = (1/c) \sum_{l=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} g_{n,l}(r) \times \exp[i(n\varphi - 2\pi l z_j / c)] \qquad (6)$$

and

$$g_{n,l}(r) = \int_0^{\infty} G_{n,l}(R) J_n(2\pi R r) 2\pi R \, \mathrm{d}R, \qquad (7)$$

where $\rho(r, \varphi, z)$ is electron density, r, φ and z are cylindrical coordinates in real space, and c is the repeat distance in the diffracting structure. The Fourier–Bessel transform represented by (6) and (7) is analogous to the Fourier transform of the F terms in crystallography.

Implementation

Adaptation of molecular dynamics refinement to use fiber diffraction data requires two major changes from the crystallographic application. The more obvious is that an effective potential-energy term, functionally equivalent to E_{crystal}, must take account of the different forms of crystalline and fiber diffraction data, as expressed by (3) and (4). The other change stems from the fact that, in the filamentous structures that usually give rise to fiber diffraction patterns, the helically repeating asymmetric units are often covalently connected. These covalent connections must be taken into account in calculating E_b, E_a and E_t.

We have developed a package of Fortran subroutines for the refinement of the structures of helical macromolecular aggregates against X-ray fiber diffraction data. The package has been incorporated into *X-PLOR*, a macromolecular-refinement program that uses X-ray crystallographic diffraction data or nuclear magnetic resonance data (Brünger, Karplus & Petsko, 1989; Brünger, 1990). The fiber diffraction package was desgined to be a fully compatible module of *X-PLOR*, so that most of the features in *X-PLOR* suitable for crystal-structure refinement and analysis can be utilized for fiber-structure refinement and analysis. These features include topology (connectivity) and parameter definitions, empirical potential-energy calculations, structural restraints and constraints, energy minimization, molecular dynamics refinement, structural and energetic analyses and data manipulation. The fiber diffraction package also includes a subroutine to deal with covalent bonds between symmetrically equivalent subunits in the empirical potential-energy calculations.

To include fiber diffraction information in a molecular dynamics refinement or energy minimization, the discrepancy between observed and calculated intensities in fiber diffraction is treated as an effective energy,

$$E_{\text{fiber}} = S_f \left[\sum_l \int_R w(\mathcal{G}_{\text{obs}} - k\mathcal{G}_{\text{calc}})^2 \, \mathrm{d}R \right] \times \left(\sum_l \int_R w\mathcal{G}^2_{\text{obs}} \, \mathrm{d}R \right)^{-1}, \qquad (8)$$

where k is a scale factor, S_f is a weight that makes the gradient of the effective energy comparable to the gradient of the empirical potential energies, w is the individual weight for each observed intensity I_{obs} and I_{calc} is a calculated intensity [equations (3) and (5)]. The scale factor k can be calculated by least-squares techniques,

$$k = \int_R w\mathcal{G}_{\text{obs}}\mathcal{G}_{\text{calc}} \Big/ \int_R w\mathcal{G}^2_{\text{obs}}.$$

In fiber diffraction, the deconvoluted intensities are usually measured at small sampling intervals along the layer lines, which are constant in reciprocal space. The integrals in (8) then become summations.

Calculations of Fourier–Bessel structure factors and their derivatives are most easily carried out in cylindrical coordinates. The orthogonal atomic coordinates are therefore transformed into cylindrical coordinates before the structure-factor calculations. After these calculations, the derivatives of a structure factor ($\partial G/\partial r_j$, $\partial G/\partial \varphi_j$, $\partial G/\partial z_j$) must be transformed back to orthogonal coordinates for molecular dynamics calculations or energy minimizations using the chain rule,

$$\frac{\partial G}{\partial x_j} = \left(\frac{\partial G}{\partial r_j}\right)\left(\frac{\partial r_j}{\partial x_j}\right) + \left(\frac{\partial G}{\partial \varphi_j}\right)\left(\frac{\partial \varphi_j}{\partial x_j}\right)$$

and

$$\frac{\partial G}{\partial y_j} = \left(\frac{\partial G}{\partial r_j}\right)\left(\frac{\partial r_j}{\partial y_j}\right) + \left(\frac{\partial G}{\partial \varphi_j}\right)\left(\frac{\partial \varphi_j}{\partial y_j}\right).$$

The most time-consuming part of the refinement algorithm is the calculation of Fourier–Bessel structure factors and their derivatives, since there is no algorithm for Fourier–Bessel structure factors of comparable efficiency to the fast Fourier transform for Fourier structure factors in crystallography. The major problem is the evaluation of Bessel functions. To reduce computational time, a Bessel function look-up table is used and the function is evaluated by linear interpolation from the table entries. A complete Bessel look-up table would be impracticably large, so the table is generated for one layer line at a time. The intensities on any one layer line are derived from a relatively small number of Bessel functions, depending on the resolution limit of the diffraction data and the symmetry and radius of the diffracting particle. Use of look-up tables can be suppressed by not setting the look-up flag, as in the crystallographic application of *X-PLOR*.

The program accommodates simple helical symmetry and cyclic symmetry about the fiber axis. The only other possible symmetry element in a helical structure is a twofold rotation about an axis perpendicular to the helix axis; this element has not yet been included in the symmetry elements considered by the program. Interparticle interactions and interactions with bulk solvent have not been included. The symmetry of the helical structure is utilized to reduce the computational time required for calculations of non-bonded interactions between atoms in different subunits. In crystallography, the search for non-bonded atom pairs is carried out throughout the entire unit cell. However, the number of symmetry-related subunits in a helix repeating unit can be very large. If the shape of the individual subunits is reasonably simple, it is possible first to identify the subunits that are within range for non-bonded interactions and consider only those subunits in the calculations. In TMV, for example, there are 49 subunits in the unit cell but a given subunit makes contacts with only 6 neighboring subunits.

Interactions between subunits in some helical systems involve not only non-bonded interactions but also covalent bonds. For example, in TMV, a single-stranded RNA molecule of approximately 6400 nucleotides follows the right-handed helix of coat protein subunits, with three nucleotides bound to each protein subunit. We refer to this type of linkage as a symmetric linkage. During the structure refinement, the continuity of the RNA molecule in the structure can be preserved by the introduction of restraints on the bond lengths and angles involving atoms in symmetry-related subunits. This is done by modifying the topology table and the table of non-bonded intersubunit interactions so that the energies associated with symmetric linkages are included in E_b, E_a and E_t.

A modified parameter file was used in place of the parameter file PARAM16.DNA in *X-PLOR*. In applications of the program to the refinement of virus structures (see below), the original parameters were found to be insufficient to maintain the planarity of nucleotide bases. In the replacement file, force constants for torsion angles and improper rotations similar to those of tryptophan and tyrosine were used; these force constants are much larger than those of the nucleotides in the original PARAM16.DNA parameter file. The modified parameter file produced stereochemically satisfactory structures.

The fiber diffraction package has been installed and tested on VAX/VMS and SGI IRIS/340 computers. Several rod-shaped helical plant viruses (Wang, Pattanayek & Stubbs, 1992) and a filamentous bacteriophage (Nambudripad & Makowski, 1992) have been refined using this package.

Refinement of tobacco mosaic virus

The structure of TMV was used to test the refinement procedure, since this virus structure has been determined by fiber diffraction methods at relatively high resolution (Namba, Pattanayek & Stubbs, 1989). TMV is a rod-shaped RNA virus 3000 Å in length and 180 Å in diameter. Approximately 2100 identical coat protein subunits form a right-handed helix with 49 subunits in three turns. The axial repeat distance is 69.0 Å. A single-stranded RNA molecule follows the basic helix between the coat protein subunits at a radius of about 40 Å. The TMV structure was refined at 2.9 Å resolution to an R factor of 0.097 using RLS by Namba, Pattanayek & Stubbs (1989). Fiber diffraction R factors are inherently lower than crystallographic R factors, because of the cylindrical averaging of the data but, for a structure having the symmetry and dimensions of TMV at 2.9 Å resolution, the R factor to be expected from a set of atoms randomly distributed within the radial limits of the virus would be about 0.32 (Stubbs, 1989; Millane, 1989). The final TMV structure included 158 amino acid residues, three nucleotides, 71 water molecules and two calcium ions.

For testing purposes, the model of Namba, Pattanayek & Stubbs (1989) with solvent molecules and calcium ions excluded for simplicity was used as a target structure. It is referred to here as NPS. A simulated intensity data set was generated between 10.0 and 2.9 Å resolution from the NPS model. To simulate errors such as might be contained in an unrefined virus structure, NPS was perturbed in a 0.4 ps molecular dynamics simulation at 500 K, without the restraint imposed by the fiber diffraction effective energy. In addition, the conformation of the loop between residues 103 and 106 was altered (Fig. 1*a*), and the side chain of Arg 122 was substantially perturbed (Fig. 1*b*). The R factor for this model, called the initial model, was 0.292. The r.m.s. differences between the initial model and the NPS atomic coordinates are shown for each residue in Fig. 2(*a*); the average r.m.s. difference was 1.79 Å (see Table 1).

Before molecular dynamics (MD) refinement was carried out, the initial model was idealized by energy minimization without the fiber diffraction effective energy. This released the tension caused by intra- and intersubunit non-bonded close contacts. A weight of 2.2×10^6 for the fiber diffraction effective energy [S_f in (8)] was estimated using Brünger's method (Brünger, 1990). Unit weight w [(8)] was assigned to all the diffraction data. Molecular dynamics refinement was carried out using the slow-cooling protocol of Brünger, Krukowski & Erickson (1990); the temperature was slowly decreased from 4000 to

200 K over a period of 0.76 ps. The temperature was controlled using the T coupling method (Berendsen, Postma, van Gunsteren, DiNola & Haak, 1984; Brünger, Krukowski & Erickson, 1990). After the MD refinement, the structure was further refined by 200 steps of energy minimization. Simulated data between 10.0 and 2.9 Å resolution were used in the refinement. Intersubunit interactions required consideration of six additional subunits, located −17, −16, −1, +1, +16 and +17 subunits from the originating subunit in the viral helix. In the TMV structure, these subunits are sufficient to cover all possible interactions. The continuity of the RNA molecule was maintained during refinement by restraining the one bond length (P–O3′) and four bond angles (C3′–O3′–P, O3′–P–O1P, O3′–P–O2P and O3′–P–O5′) affected by the symmetric linkage. During the refinement, the R factor was reduced from 0.292 to 0.070 (Fig. 3).

The initial model was also refined against the simulated intensity data by the restrained least-squares (RLS) method (Hendrickson, 1985; Stubbs, Namba & Makowski, 1986). Again, six additional subunits were considered in order to restrain close contacts and the bond distances around the symmetric linkage were restrained to preserve the continuity of the RNA molecule. This refinement (without manual rebuilding) converged to an R factor of 0.143, with comparable stereochemistry to the MD-refined model.

The R-factor distributions for the initial, RLS-refined and MD-refined models are shown in Fig. 4.

In general, the geometry of the MD-refined model is excellent. The average r.m.s. deviations of covalent bond lengths (0.015 Å) and angles (2.0°) from ideal values are comparable with those found in well refined structures using either crystal or fiber diffraction data. The deviations are plotted as a function of residue number in Fig. 5. Some of the C_α–C_β bonds in the refined model are systematically longer than the ideal value of a carbon-carbon single bond

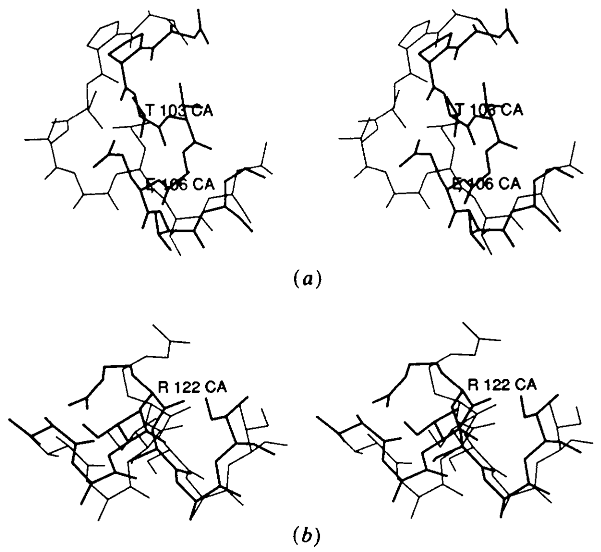


Fig. 1. Stereoviews of parts of the reference structure (thick lines), called NPS in the text, and the initial model for the refinement tests (thin lines). (*a*) The 103–106 loop and (*b*) residue Arg 122.

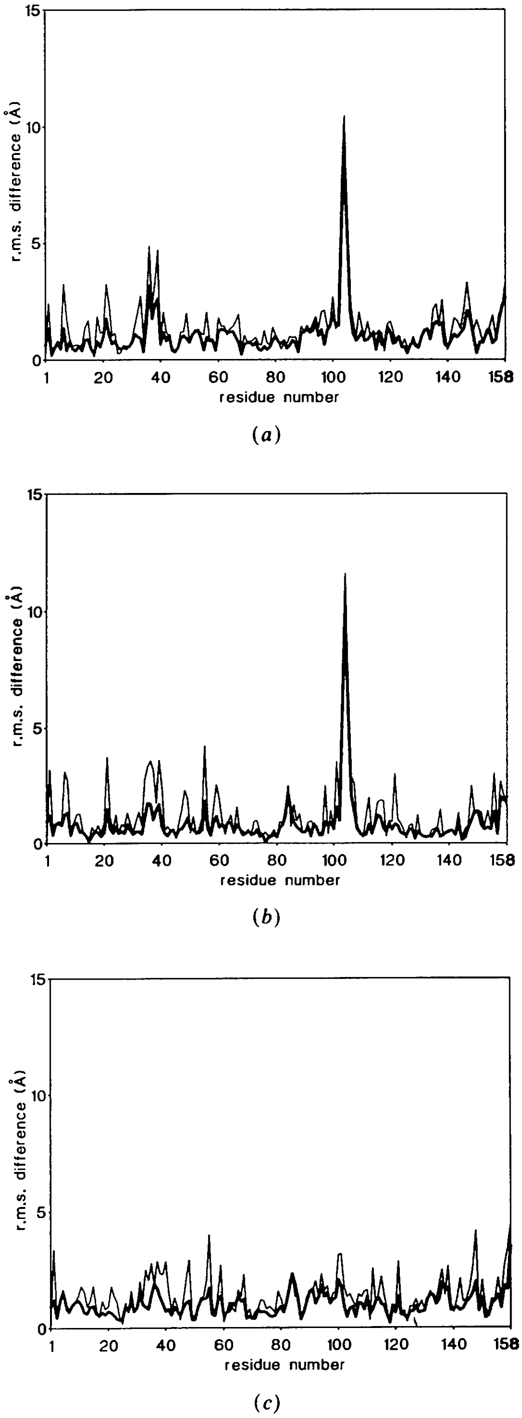


Fig. 2. R.m.s. differences between the atomic coordinates of different models as a function of residue number in the protein. Thick lines: main-chain differences. Thin lines: side-chain differences. (*a*) Initial model and NPS. (*b*) MD-refined model and NPS. (*c*) Initial model and MD-refined model.

Table 1. *R.m.s. differences between atomic coordinates in different TMV models*

NPS is the 'true' structure, the target of the refinements. The initial model was obtained by perturbing NPS and then refined by both the MD and RLS methods. MD is the MD-refined model; RLS is the RLS-refined model. R.m.s. distances between corresponding atoms in the two models listed on each line are in Å. H atoms are not included.

	Models		Protein main chain	Protein side chain	RNA	All
Distance from target before refinement	Initial	NPS	1.39	1.91	2.79	1.79
Distance from target after refinement	MD	NPS	1.28	1.85	1.65	1.64
	RLS	NPS	1.30	1.83	2.34	1.69
Distance moved during refinement	MD	Initial	1.06	1.64	2.80	1.53
	RLS	Initial	0.36	0.52	1.27	0.53

(1.540 Å). The bond angles around these elongated C_α–C_β bonds also deviate from ideal values. Most of the long bond distances are found in residues for which the main-chain dihedral angles are in forbidden regions of the Ramachandran plot.

The r.m.s. difference between the atomic coordinates in the MD-refined model and NPS was 1.64 Å, reduced slightly from 1.79 Å. The difference between the RLS-refined model and NPS was 1.69 Å. Differences for main-chain protein, side-chain protein and RNA atoms, as well as r.m.s. shifts during refinement, are given in Table 1. The r.m.s. differences between the MD-refined model and NPS are plotted as a function of residue number in Fig. 2(*b*). The differences between the initial model and the MD-refined model are shown in Fig. 2(*c*). As may be seen in Fig. 2, the r.m.s. differences for the altered loop between residues 103 and 106 were smaller than those in the initial model, but the overall conformation was not corrected [compare Figs. 1(*a*) and 7(*a*)]. The r.m.s. deviations of residues in the α-helical regions of the

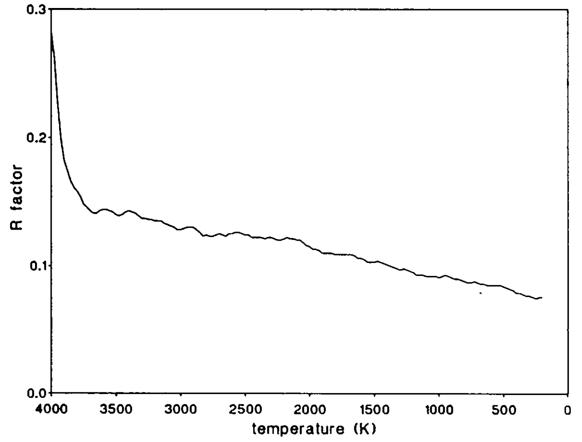


Fig. 3. *R*-factor changes during the molecular dynamics refinement using the slow-cooling protocol.

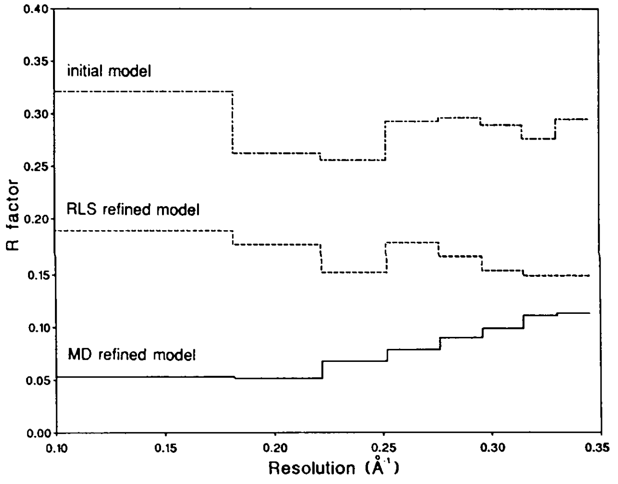
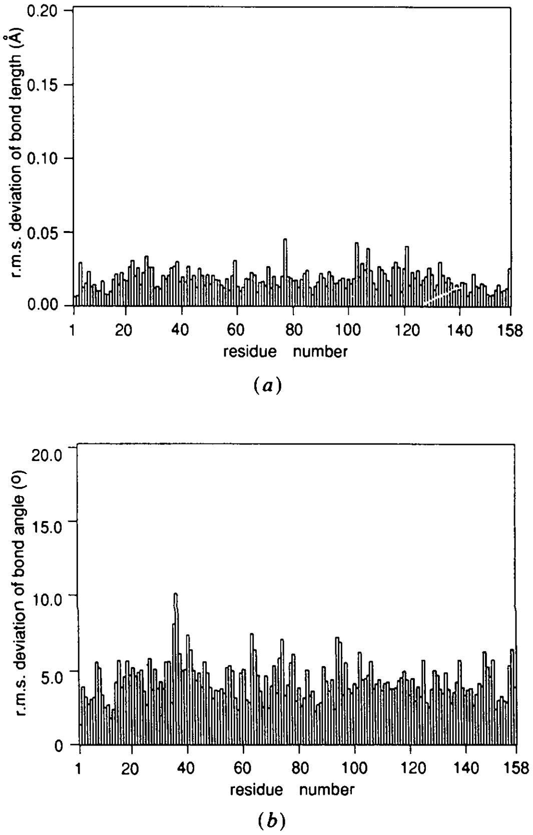


Fig. 4. *R*-factor distributions within refinement ranges between 10 and 2.9 Å for the initial model, the RLS-refined model and the MD-refined model.



(*a*)

(*b*)

Fig. 5. R.m.s. deviations from ideal values of bond lengths and angles in the MD-refined model as a function of residue number. (*a*) Bond lengths. (*b*) Bond angles.

MD-refined model were smaller than those in the initial model.

Electron density maps were calculated using coefficients $6\mathcal{G}_{\text{obs}}-5\mathcal{G}_{\text{calc}}$ and $\mathcal{G}_{\text{obs}}-\mathcal{G}_{\text{calc}}$, where $\mathcal{G}_{\text{obs}}$ was taken from the simulated intensity data set. $6\mathcal{G}_{\text{obs}}-5\mathcal{G}_{\text{calc}}$ maps are similar to crystallographic $2F_{\text{obs}}-F_{\text{calc}}$ electron density maps (Namba & Stubbs, 1987). $6\mathcal{G}_{\text{obs}}-5\mathcal{G}_{\text{calc}}$ maps of the region in the vicinity of Arg 90 (part of the RNA binding site), calculated at 2.9 Å resolution, are shown in Fig. 6 and $\mathcal{G}_{\text{obs}}-\mathcal{G}_{\text{calc}}$ maps of the 103–106 loop region are shown in Fig. 7. $\mathcal{G}_{\text{calc}}$ was taken from the MD-refined model, the RLS-refined model and the initial model.

Although the MD refinement did not correct the conformation of the 103–106 loop, the difference maps (Fig. 7*a*) clearly showed the correct position of the chain in this region. In contrast, difference maps based on the initial model (Fig. 7*c*) were completely uninterpretable and, while difference maps based on the RLS model (Fig. 7*b*) did show some indication of the error, it seems most unlikely that those maps could have been interpreted without knowledge of the true structure. Neither the refined structure nor the difference maps indicated the correct conformation of the Arg 122 side chain (not shown).

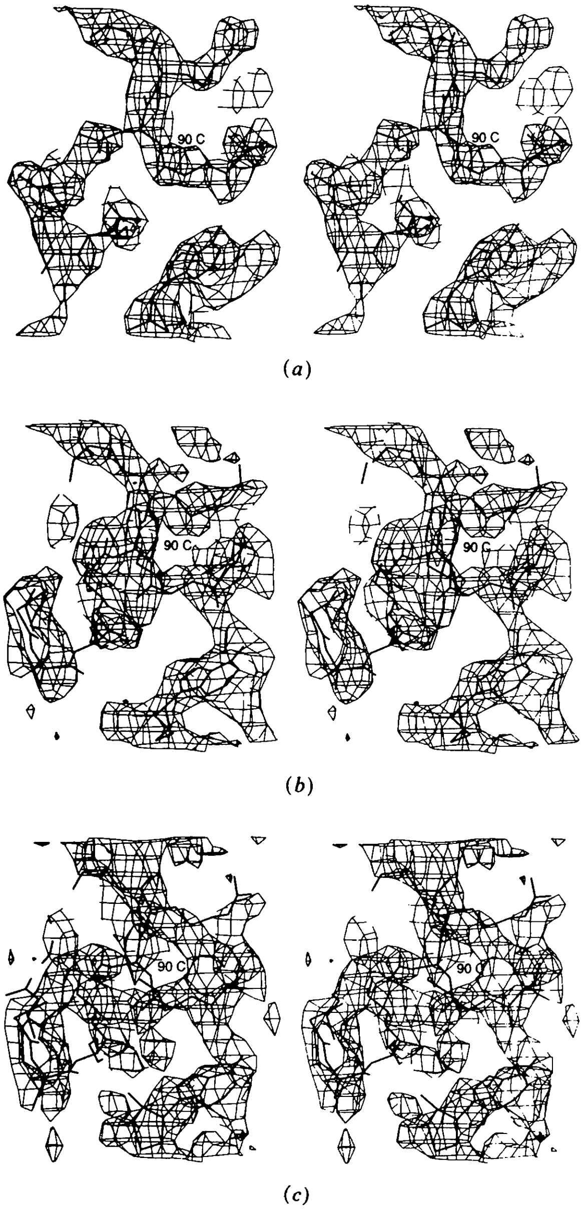


Fig. 6. Stereoviews of $6\mathcal{G}_{\text{obs}}-5\mathcal{G}_{\text{calc}}$ electron density maps at 2.9 Å resolution of the TMV coat protein in the vicinity of Arg 90. Thin lines represent the contoured electron density. Thick lines represent the corresponding models. (*a*) $\mathcal{G}_{\text{calc}}$ calculated from the MD-refined model. (*b*) $\mathcal{G}_{\text{calc}}$ calculated from the RLS-refined model. (*c*) $\mathcal{G}_{\text{calc}}$ calculated from the initial model.

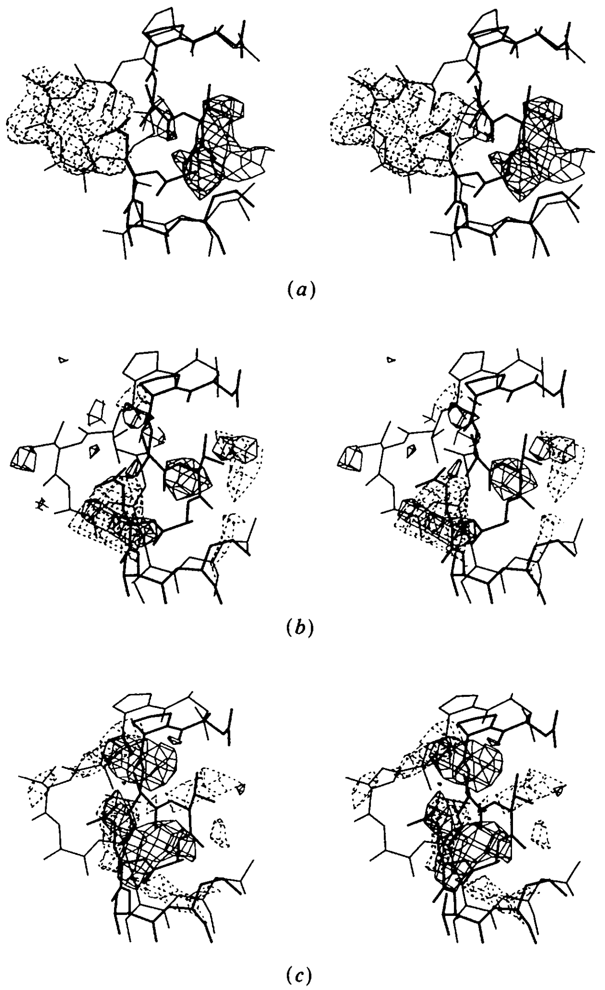


Fig. 7. Stereoviews of difference electron density maps $\mathcal{G}_{\text{obs}}-\mathcal{G}_{\text{calc}}$ of the TMV coat protein in the vicinity of the 103–106 loop. In the density maps, the solid lines represent the contoured electron density at the 5σ level; the dashed lines are at the -5σ level. Thick lines represent the NPS model. (*a*) $\mathcal{G}_{\text{calc}}$ calculated from the MD-refined model. Thin lines represent the MD-refined model; (*b*) $\mathcal{G}_{\text{calc}}$ calculated from the RLS-refined model. Thin lines represent the RLS-refined model. (*c*) $\mathcal{G}_{\text{calc}}$ calculated from the initial model. Thin lines represent the initial model.

Discussion

Molecular dynamics refinement of the TMV structure converged to a very low R factor and produced a structure with generally satisfactory stereochemistry. Gross structural errors in the model were not corrected automatically but the model was improved sufficiently to allow these errors to be corrected by reference to difference electron density maps.

The most significant effect of MD refinement on the structure determination was the improvement in the quality of the electron density maps (Figs. 6 and 7). In the $6\mathcal{G}_{\rm obs}-5\mathcal{G}_{\rm calc}$ map of the initial model (Fig. 6*c*), interpretable electron density generally appeared only in the main-chain regions of regular secondary structures. Side-chain structures could not be determined from these maps. In contrast, in the $6\mathcal{G}_{\rm obs}-5\mathcal{G}_{\rm calc}$ map of the MD-refined model (Fig. 6*a*), both the main chain and the side chains closely fit the electron density. The large errors in the conformation of the 103–106 loop could easily be recognized in the $\mathcal{G}_{\rm obs}-\mathcal{G}_{\rm calc}$ map (Fig. 7*a*). More subtle errors, such as the altered conformation of the side chain of Arg 122, were not detected at this stage of refinement but our experience with other virus structures (Pattanayek & Stubbs, 1992; Wang, Pattanayek & Stubbs, 1992) has been that errors such as these can be found and corrected after one or more cycles of rebuilding and further refinement.

Comparison of the R factors of the MD-refined model (0.070) and the RLS-refined model (0.143) as well as the difference maps obtained from the two refined models (Figs. 6 and 7) suggests that the effective experimental radius of convergence of the MD method is significantly greater than that of the RLS method. The r.m.s. difference between the atomic coordinates of the model and those of the true structure was reduced only slightly during MD refinement, from 1.79 to 1.64 Å, while RLS refinement reduced this difference to 1.69 Å. The RNA atomic coordinates were greatly improved during MD refinement (Table 1). This limited test did not directly address the question of radius of convergence, but in view of the greater ability of MD refinement to move atoms to positions corresponding to lower R factors, it appears to be very likely that the MD-refinement radius of convergence is greater than that of RLS refinement, as has been observed with crystallographic data (Brünger, Kuriyan & Karplus, 1987). The RLS method is not usually successful when an inaccurate starting model is refined against all available diffraction data; a common practice is to start the refinement using only low-resolution data and gradually to add the higher-resolution data as the refinement progresses. This procedure increases the initial radius of convergence but requires much more time and frequent model rebuilding. In contrast to RLS, MD refinement can be started with all available diffraction data without a significant reduction of the radius of convergence. Similar results have been obtained for MD refinement using crystallographic data (Gros, Fujinaga, Dijkstra, Kalk & Hol, 1989; Brünger, Krukowski & Erickson, 1990).

Although the geometry of the refined model was generally excellent, the systematic errors in the lengths of some of the C_α–C_β bonds associated with forbidden dihedral angles in the main chain were a matter of concern. Close contacts between side-chain atoms and main-chain atoms were apparently resolved by pushing the side chain away from the main chain, which elongated the C_α–C_β bond. In practice, it has been necessary to correct these dihedral angles and other minor geometric problems by manual intervention followed by further refinement (Wang, Pattanayek & Stubbs, 1992).

The refinement did not converge to an R factor of zero, even though no errors were introduced into the simulated intensity data set. In particular, the altered loop structure between residues 103 and 106 in the initial model was not corrected. This observation suggests that the energy barrier between the altered loop structure and the correct structure was too high to be overcome under the refinement conditions used. However, the MD refinement moved many atoms to positions close to the true positions of other atoms [compare Figs. 1(*a*) and 7(*a*)]. This phenomenon was seen in many parts of the structure. The MD refinement moved the structure to a local energy minimum rather than the global energy minimum but, in the local energy minimum, most of the atoms were located in the electron density of the true structure. The refined structure therefore allowed the calculation of greatly improved phases and thus improved electron density maps, as is evident in Figs. 6 and 7. Some of the errors of refined mislocated atoms could easily be corrected from the improved electron density maps and a model improved in this way could serve as a starting point for a second round of refinement.

The high energy barriers between the starting structure and the true structure may be attributed to the compactness of the TMV particle. Since this compactness is typical of many biological filamentous assemblies, these barriers should be expected in most fiber diffraction refinements. In crystals of macromolecules, a large portion of the space between protein subunits is occupied by solvent molecules. These solvent molecules are not usually considered in molecular dynamics refinements, so the empty spaces in the crystal lattice can allow thermal expansion of the protein molecules during refinement at high temperatures without explicitly considering thermal expansion of the crystal. In contrast, the coat protein molecules in helical viruses are usually folded into a relatively compact structure and the protein subunits in turn are packed into a very compact helical aggregate, with much less space for molecular dynamics refinement to move the atoms within the particle. The energy barriers caused by the close contacts are so high that increasing the starting temperature without allowing thermal expansion of the particle would not be expected to overcome them.

In summary, molecular dynamics refinement against fiber diffraction has been shown to be an effective means of structure determination. Even when the initial structure is too far from the true structure to allow direct refinement, the method is able to find local minima that resemble the true structure sufficiently to allow improved phasing and thus lead to interpretable difference maps.

We thank Axel Brünger, Lee Makowski and Rekha Pattanayek for valuable discussions. This work was supported by NSF grants DIR-8915800 and DIR-9011014.

References

Arnott, S. & Wonacott, A. J. (1966). *Polymer*, **7**, 157-166.

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. (1984). *J. Chem. Phys.* **81**, 3684-3690.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). *J. Comput. Chem.* **4**, 187-217.

Brünger, A. T. (1990). *X-PLOR Manual.* Version 2.1. Yale Univ., New Haven, USA.

Brünger, A. T., Karplus, M. & Petsko, G. A. (1989). *Acta Cryst.* **A45**, 50-61.

Brünger, A. T., Krukowski, A. & Erickson, J. (1990). *Acta Cryst.* **A46**, 585-593.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458-460.

Cochran, W., Crick, F. H. C. & Vand, V. (1952). *Acta Cryst.* **5**, 581-586.

Franklin, R. E. & Klug, A. (1955). *Acta Cryst.* **8**, 777-780.

Gros, P., Fujinaga, M., Dijkstra, B. W., Kalk, K. H. & Hol, W. G. J. (1989). *Acta Cryst.* **B45**, 488-499.

Hendrickson, W. A. (1985). *Methods in Enzymology*, Vol. 115, edited by H. W. Wyckoff, C. H. W. Hirs & S. N. Timasheff, pp. 252-270. Orlando: Academic Press.

Karplus, M. & Petsko, G. A. (1990). *Nature (London)*, **347**, 631-639.

Klug, A., Crick, F. H. C. & Wyckoff, H. W. (1958). *Acta Cryst.* **11**, 199-213.

Makowski, L. (1982). *J. Appl. Cryst.* **15**, 546-557.

Millane, R. P. (1989). *Acta Cryst.* **A45**, 573-576.

Namba, K., Pattanayek, R. & Stubbs, G. (1989). *J. Mol. Biol.* **208**, 307-325.

Namba, K. & Stubbs, G. (1985). *Acta Cryst.* **A41**, 252-262.

Namba, K. & Stubbs, G. (1987). *Acta Cryst.* **A43**, 533-539.

Nambudripad, R. & Makowski, L. (1989). *Biophys. J.* **55**, 417*a*.

Nampudripad, R. & Makowski, L. (1992). Unpublished results.

Nambudripad, R., Stark, W. & Makowski, L. (1991). *J. Mol. Biol.* **220**, 349-379.

Pattanayek, R. & Stubbs, G. (1992). *J. Mol. Biol.* **228**, 516-528.

Smith, P. J. C. & Arnott, S. (1978). *Acta Cryst.* **A34**, 3-11.

Stubbs, G. (1989). *Acta Cryst.* **A45**, 254-258.

Stubbs, G. & Diamond, R. (1975). *Acta Cryst.* **A31**, 709-718.

Stubbs, G. & Makowski, L. (1982). *Acta Cryst.* **A38**, 417-425.

Stubbs, G., Namba, K. & Makowski, L. (1986). *Biophys. J.* **49**, 58-60.

Wang, H., Pattanayek, R. & Stubbs, G. (1992). Unpublished results.

Waser, J. (1955). *Acta Cryst.* **8**, 142-150.

Acta Cryst. (1993). **A49**, 513-527

The Electron Distribution in Corundum. A Study of the Utility of Merging Single-Crystal and Powder Diffraction Data

By Anthony S. Brown and Mark A. Spackman

Department of Chemistry, University of New England, Armidale, NSW 2351, Australia

and Roderick J. Hill

CSIRO Division of Mineral Products, PO Box 124, Port Melbourne, Victoria 3207, Australia

(*Received 22 April 1992; accepted 22 October 1992*)

Abstract

Powder X-ray diffraction data for corundum were collected by a variety of methods and reduced to structure amplitudes by two profile-fitting techniques. The resulting averaged powder-data set was merged with three different single-crystal data sets to assess the improvements possible over least-squares modelling of extinction for accurate electron density analysis of minerals. With reference to the deformation electron density derived from multipole refinements, it is concluded that this strategy offers advantages over the *post facto* modelling of severe extinction effects commonly observed in such

0108-7673/93/030513-15$06.00