Acta Cryst. (1996). D**52**, 693-711

A Pseudo-Cell Based Approach to Efficient Crystallographic Refinement of Viruses

DAVID H. JACOBSON,[a,b] JAMES M. HOGLE [b,c] AND DAVID J. FILMAN[b]*

[a]*Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093, USA,* [b]*Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, MA 02115, USA, and* [c]*Committee for Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02138, USA. E-mail: djf@vp2.med.harvard.edu*

(*Received 20 October 1995; accepted 22 January 1996*)

Abstract

Strategies have been developed for the inexpensive refinement of atomic models of viruses and of other highly symmetric structures. These methods, which have been used in the refinement of several strains of poliovirus, focus on an arbitrary-sized parallelepiped (termed the 'protomer' box) containing a single complete averaged copy of the structural motif which forms the protein capsid, together with the fragments of other symmetry-related copies of the motif which are located in its immediate neighborhood. The Fourier transform of the protomer box provides reference structure factors for stereochemically restrained crystallographic refinement of the atomic model parameters. The phases of the reference structure factors are based on the averaged map, and are not permitted to change during the refinement. It is demonstrated that models refined using the protomer box methods do not differ significantly from models refined by more expensive full-cell calculations.

1. Introduction

The crystallographic refinement of large highly symmetric protein structures such as viruses presents several challenges. The magnitude of the computation, because of the size of the unit cell and its large complement of atoms, often necessitates compromises in the refinement method. In most currently used refinement methods, structures are refined to optimize the agreement between observed and calculated structure factors, while maintaining the consistency of the atomic models with sequence information, stereochemical standards, and non-crystallographic symmetry (NCS) constraints. These physical constraints or restraints typically are included in refinements of atomic protein model to reduce the effective number of independent atomic parameters, thus increasing the number of observations per parameter, and thereby improving the convergence properties of the refinement. A restrained refinement typically minimizes a weighted combination of experimental and empirical terms,

$$Q_{\text{total}} = Q_{\text{crystallographic}} + Q_{\text{stereochemical}} + Q_{\text{other}}, \quad (1)$$

where $Q_{\text{stereochemical}}$ may include empirical restraints on bond lengths and angles, planarity, chiral volume, van der Waals forces, electrostatic forces and hydrogen bonding; and Q_{other} could include NCS restraints (Konnert & Hendrickson, 1980; Hendrickson, 1985; Brünger, 1992*a*). $Q_{\text{crystallographic}}$ has the general form,

$$Q = \sum_{hkl} w_{hkl} |\tilde{F}_{\text{obsd}}(hkl) - k_{\text{resol}} \tilde{F}_{\text{calc}}(hkl)|^2, \quad (2)$$

where w_{hkl} are reflection-specific weights, the k_{resol} are resolution-dependent scale factors, and $\tilde{F}_{\text{calc}}$ and $\tilde{F}_{\text{obsd}}$ are the model-based and reference complex-valued structure factors, respectively. The phases assigned to the reference structure factors of (2) are commonly obtained from the most current set of model-based structure factors, though experimentally determined phases are sometimes used instead.

The decision to include observed phases in the residual depends on one's estimate of whether experimental or model-dependent phases better approximate the true phases, which is sometimes indicated by a figure of merit (Rees & Lewis, 1983). In addition, schemes for using weighted combinations of both types of residuals have been suggested (Hendrickson & Lattman, 1970). In structures with extensive NCS, such as icosahedral viruses, the refinement of phases by the iterative application of NCS constraints to the observed structure-factor magnitudes can provide a phase set of exceptionally good quality (Rossmann & Blow, 1963; Main, 1967; Crowther, 1969; Bricogne, 1974, 1976). Indeed, the reliability of these phases suggest that they be given full weight in constructing the reference set $\{\tilde{F}_{\text{obsd}}\}$. (2) may then be expanded as,

$$Q_{\text{vec}} = \sum_{hkl} w_{hkl} \{ [A_{\text{obsd}}(hkl) - k_{\text{resol}} A_{\text{calc}}(hkl)]^2 + [B_{\text{obsd}}(hkl) - k_{\text{resol}} B_{\text{calc}}(hkl)]^2 \}, \quad (3)$$

where A and B are the real and imaginary components of the observed and calculated structure factors, respectively. When all of the terms are present, minimization of this residual is equivalent to a real-

space refinement which seeks to maximize a density overlap integral.

Refinements of several high-resolution atomic models of virus crystal structures have been reported (Jones & Liljas, 1984; Arnold & Rossmann, 1988; Filman *et al.*, 1989; Fry, Acharya & Stuart, 1993; Wu, Keller & Rossmann, 1993). In some cases, only the observed structure-factor magnitudes were regarded as standards, whereas in other cases both the experimental amplitudes and NCS-constrained phases were targets for the refinement.

The first serious effort to optimize the parameters of the atomic model of a virus was reported by Jones & Liljas (1984) in the structure determination of satellite tobacco necrosis virus (STNV). Given the limited computational resources at the time, it was necessary to separate the overall structure refinement into two distinct phases. Determination of the reference phases by the method of Bricogne (1974, 1976) was alternated with refinement of the parameters of the atomic model using the graphics program *FRODO* (Jones, 1985). Model refinement was approached as a real-space model-building exercise, finding locally optimal ways to fit pieces of a single capsid protein subunit to a portion of the averaged electron-density map. Rigid-body refinement and stereochemical optimization were repeatedly applied to small fragments of the models, both interactively and as batch processes under the control of command scripts.

Focusing on a single copy of the capsid protein represented an important short-cut, both in terms of the number of calculations and the amount of computer memory (expensive at the time) required to carry out the refinement. Since then, advances in computer technology have made it feasible to consider all of the atomic parameters at once, with the simultaneous application of crystallographic and stereochemical restraints.

Silva & Rossmann (1985) reported the first global refinement of a complete virus capsid (southern bean mosaic virus) using reciprocal-space methods. This refinement used a version of the stereochemically restrained least-squares refinement program *PROLSQ* (Konnert & Hendrickson, 1980) which had been modified to incorporate strict 20-fold NCS. In principle, this approach represented an improvement on piecemeal local optimization, both because the program explicitly considered non-bonded interactions, and because the global nature of the refinement made convergence to a single well defined minimum a possibility. This method followed the common practice of using observed structure-factor magnitudes (rather than the complex-valued transform of the icosahedrally constrained structure) as a reference standard. However, subsequent refinements from that group (Arnold & Rossmann, 1988; Wu, Keller & Rossmann, 1993) have included the option to minimize the residual in (3), which sums vector-valued discrepancies, or to minimize a figure-of-merit-weighted combination of scalar and vector terms.

In principle, such a global refinement should make it possible to spend computing resources to reduce the effort of the interactive model builder. Unfortunately, however, use of the program proved to be so expensive that only a randomly selected subset of the observed reflections could be used at one time to form the reference set for refinement. The practice of discarding most of the observations creates a dangerous reduction in the number of observations per parameter, which is especially problematic because of the inherently noisy quality of data from large unit cells.

Years ago, it was recognized that the model-based structure factors and estimates of partial derivatives for use in least-squares refinement procedures could both be obtained most efficiently for large structures by using fast Fourier transforms (Ten Eyck, 1973, 1977; Agarwal, 1978). However, when the maps to be transformed are larger than the available physical computer memory, disk access is required, either using explicit scratch files or so-called 'virtual memory'. Repeated large transforms during the course of a refinement are consequently very expensive, even when steps are taken to minimize disk access.

Recent versions of the *X-PLOR* package capable of handling the stereochemically restrained full-cell refinement of very large structures (Brünger, 1989, 1992*a*) have implemented such steps, though the program is still extremely slow whenever repeated virtual-memory access is required. Typically, strict NCS constraints are enforced to reduce the number of independent parameters in the refinement of icosahedral virus structures. Among the virus structures refined in this way are a number of foot-and-mouth disease virus (FMDV) strains (Acharya *et al.*, 1989), P3/Sabin poliovirus (Syed, Filman & Hogle, 1995), three variants of type 1 poliovirus (see below), and satellite tobacco mosaic virus (Larson *et al.*, 1993). In general, the FMDV refinements used observed structure-factor magnitudes as reference standards, while most of the poliovirus refinements used the transforms of the phase-constrained map as a standard.

The crystals of FMDV, P3/Sabin poliovirus, and P1/Mahoney poliovirus express fivefold, 15-fold, and 30-fold NCS, respectively. Accordingly, the size and expense of the calculation of the structure factors and their derivatives have risen with the number of atoms per asymmetric unit. Unlike the FMDV refinement, which used a full-cell *X-PLOR* calculation in all refinement cycles, the poliovirus structure determinations in this laboratory have needed to rely on much less expansive 'preliminary refinement' methods (which are described in the present manuscript) to produce essentially correct models. Sometimes, full-cell *X-PLOR* calculations were used in the very last stages of the refinements to further optimize these models.

Data presented below demonstrate that even though the atomic models were already close to convergence, straightforward minimizations using full-cell *X-PLOR* still required weeks on modern multi-processor workstations.

Conventional crystallographic model refinement algorithms [such as those in *X-PLOR*, *PROLSQ* and *TNT* (Tronrud, Ten Eyck & Matthews, 1987)] are very inefficient when applied to large highly symmetric virus structures. The most expensive aspects of refinement are the costs of calculating structure factors and of estimating their derivatives. These costs are proportional, both in CPU time and machine-memory requirements, to the number of atoms in the asymmetric unit (or viewed in a sightly different way, its volume). By contrast, the calculation of the stereochemical potential and its derivatives is remarkably inexpensive, being merely proportional to the size of a single copy of the unique structural motif (as shown below). In poliovirus, each of the 60 icosahedrally unique building blocks of the capsid is referred to as a protomer, and contains about 850 crystallographically ordered amino-acid residues.

The key to an efficient crystallographic refinement, then, is to make the structure-factor calculation similarly proportional to the volume of a single chemically continuous protomer, plus whichever fragments of symmetry-related protomers happen to be located in its immediate neighborhood. Arbitrarily defining this neighborhood to be parallelepiped, termed the 'protomer box', makes it straightforward to calculate its Fourier transform, implicitly treating it as if it were a periodic object. A pseudo real-space refinement procedure which minimizes the 'vector' residual in (3) then can be used to modify the atomic parameters of the single protomer so as to improve the agreement of model-based electron density (after appropriate filtering and scaling) with the corresponding portion of the authentic symmetry-constrained electron-density map. (Note that the phases of the reference structure factors do not change during the refinement.) Carrying out a stereochemically restrained refinement that converges to a well defined minimum simply requires combining the gradient of the vector residual (which is evaluated inexpensively in the protomer box) linearly with the gradient of the stereochemical potential.

The refinement methods described in this manuscript have been tested and evaluated on three variants of type 1 poliovirus. These variants differ in the amino-acid sequence of the *BC* loop of capsid protein VP1 and include P1/Mahoney, a well characterized neurovirulent laboratory strain of type 1 (Hogle, Chow & Filman, 1985), V510, a mouse-virulent chimera of serotypes 1 and 2 (Yeates *et al.*, 1991), and VD9, a recently solved type 1 mutant with a loop deletion (Jacobson, Filman, Martin, Girard & Hogle, 1996). The first P1/Mahoney atomic model was refined using protomer-box-based methods in the mid-1980's (Filman *et al.*, 1989), and reported as the coordinate set '2PLV' in the Protein Data Bank (PDB). At that time, non-bonded contacts and hydrogen bonds were omitted from the set of stereochemical constraints, in the interest of making the refinement affordable, and furthermore, the stereochemical standards differed from those commonly in use today. Similar refinements were carried out later in the determination of several other picornavirus structures, including V510 and Theiler's virus (Grant, Filman, Fujinami, Icenogle & Hogle, 1992). More recently, the intention to run molecular-dynamics calculations on poliovirus models has made it necessary to refine the models using the same set of tight stereochemical standards that will be used in dynamics.

The following experiments demonstrate that the protomer-box refinement (called 'XX12' in it current version) reduces the expense of the calculation by orders of magnitude relative to full-cell *X-PLOR*, both by decreasing the number of computations required, and by eliminating disk access related to the memory requirements of the full-cell calculation. Although the process is not yet fast enough to use during interactive model-building sessions, it requires only a few hours, rather than weeks, on common SGI R3000 workstations to globally optimize the results of a session of model building, thereby providing an improved starting point for the next interactive session. Further increases in efficiency are expected once the method (currently implemented as a UNIX C-shell script) is rewritten as an integrated program. In the present manuscript, direct comparisons of the models obtained from the full-cell and pseudo-cell calculations show that the two refinements yield very similar results, and indicate that using the less expensive XX12 procedure alone is probably sufficient.

It is worth emphasizing that much of the complexity of the procedure outlined below has been introduced deliberately to avoid artifacts in the refinement. These precautions ensure that the refined atomic parameters are not affected by the presence of enveloping artifacts near the model, and that all neighboring atoms are taken into account when the chemical potentials are evaluated. The strategies implemented here require different symmetries to be assumed by the crystallographic and stereochemistry routines, a contingency that common refinement programs are not designed to address. At the option of the user, a simpler more approximate version of positional refinement can be run using commonly available refinement packages, once a reference set of pseudo structure factors has been prepared as explained below. Although the simplified procedures produce results less closely similar to those of full-cell refinement, some users may find them satisfactory.

2. Methods

2.1. *Overview of the XX12 protocol*

The XX12 procedure is initiated by creating a suitable electron-density map to serve as the standard for the refinement For poliovirus, this map is an arbitrarily defined orthogonal pseudo-cell with $P1$ symmetry, the protomer box, which is of sufficient size ($105 \times 120 \times 90$ Å) to contain one complete chemically continuous protomer in the center of the box, together with any fragments of its eight neighboring protomers that happen to lie within its limits. Reference electron-density values for the protomer box are obtained by interpolation from an icosahedrally averaged electron-density map previously produced during NCS-based phase refinement. Then, in each cycle, the atomic model of the central protomer is expanded using icosahedral operators to fill the protomer box, and model-based electron density is generated from this expanded list. Roughly 3.5 protomers worth of model-based density is contained within the box limits, which is far less than the 30 protomers which occupy the $P2_12_12$ asymmetric unit of type 1 poliovirus. The refinement, then, is based on minimizing the following (primed) approximation to (1),

$$Q'_{\text{total}} = Q_{\text{pseudo-crystallographic}} + Q_{\text{stereochemical}}. \quad (4)$$

In each step of the refinement, the derivatives of Q'_{total} with respect to the parameters of the atomic model are simply a linear combination of the pseudo-crystallographic and stereochemical gradients, each evaluated at present by a separate program. The crystallographic component is estimated inexpensively by a variant of the method of Freer, Alden, Carter & Kraut (1975) from the gradient at the atom positions in 'phased' Fourier difference maps. These maps represent the resolution-bin-scaled difference between the Fourier transforms of symmetry-constrained protomer box maps and model-based electron density. The stereochemical component is determined using the unmodified version 3.1 of *X-PLOR* (Brünger, 1992*a*) by running eight cycles of Powell minimization of commonly used empirical stereochemical energy functions over the authentic crystallographic asymmetric unit, in the absence of crystallographic terms, and with strict NCS invoked. As the crystallographic and stereochemical components of Q'_{total} are linearly independent of one another, it is permissible to evaluate their gradients separately, and then combine them using any chosen weighting scheme.

The XX12 command script includes several additional procedures that also use the Fourier transform of the protomer box volume as a reference standard. These routines permit the inexpensive optimization of solvent occupancies and temperature factors, as detailed below, and are also capable of supporting crystallographically restrained simulated annealing.

2.2. *Details of the method*

The following sections detail critical steps and key files in the XX12 refinement procedure. A flow chart of positional refinement in XX12 is provided in Fig. 1.

2.2.1. *The wedge map.* An averaged map of the icosahedrally unique volume is produced in each cycle of the double-interpolation phase-constraint procedure. [The phase-constraint procedure is a locally developed implementation of the method of Bricogne (1974) designed to run efficiently on computers with moderately large physical memories.] The particular choice of icosahedral unique volume used here, termed the 'wedge', is bounded by three planes which connect two adjacent threefold axes, and a neighboring fivefold axis. In early stages of refinement, the maps are calculated with a simple spherical envelope, typically using generous inner and outer radial cutoffs of 80 and 180 Å, respectively. The advantage of generous spherical envelopes is that they are unlikely to truncate density that may deserve interpretation. Their disadvantage is that density artifacts corresponding to neighbouring particles may appear in the wedge map. Consequently, in the later stages of the structure solution, the simple envelope usually is replaced by a tighter one which excludes points in the map located further than a specified distance, typically 3 Å, from the nearest atom of the atomic model.

When data with a maximum resolution in the 2.6–3.1 Å range are included, the wedge map is calculated on a uniform 0.75 Å Cartesian grid. Because it is intended to be the source map for a $4 \times 4 \times 4$ (64-point) non-linear interpolation (described below), all points inside the geometric wedge are calculated, together with those additional points just outside of the wedge which might be required for interpolation. In space group $P2_12_12$, each value in the wedge map lying inside the envelope represents the average of 30 interpolated values from the 'unaveraged' crystallographic asymmetric unit.

2.2.2. *Non-linear interpolation.* Following a suggestion of Bricogne (1976), it was recognized that errors as a result of interpolation can be limited to acceptable levels in either of two ways. When using linear interpolation schemes (which derive each interpolated value from a $2 \times 2 \times 2$ array of input points), errors can be limited by making the grid of the source maps finer, and commonly, grid spacing of $d/5$ and $d/6$ are used (where d represents the nominal resolution limit of the data used to produce the source map). Alternatively, the use of a more expensive non-linear interpolation scheme (here deriving each interpolated value from a $4 \times 4 \times 4$ array of input points) permits

satisfactory results to be obtained from more coarsely sampled input grids (typically $d/3.5$ to $d/4$). This trade-off between map size and CPU cycles is particularly beneficial when the largest of the source maps becomes small enough to hold in physical memory, and algorithms which require random access to the map become feasible.

The specific data-smoothing non-linear interpolation used here is equivalent to a weighted least-squares function fit which minimizes the residual,

$$\sum_{xyz} [\zeta_{\text{obsd}}(x, y, z) - \zeta_{\text{calc}}(x, y, z,)]^2 W(x, y, z), \quad (5)$$

where x, y, and z represent the integer coordinates of the surrounding grid points, defined to range from 0 to 3, $\zeta_{\text{obsd}}(x, y, z)$ represent values from the source map, $\zeta_{\text{calc}}(x, y, z)$ is a 27-term polynomial (including products of all powers of x, y, and z between 0 and 2), and $W(x, y, z)$ is a normalized Gaussian weight for each grid point based on its distance from $(1 + u, 1 + v, 1 + w)$, the point whose value is to be interpolated. Here, u, v, and w each represent some fraction of a grid unit. For any particular advance choice of u, v, and w, an origin shift permits the variables of the polynomial to be taken to be $(x - 1 - u)$, $(y - 1 - v)$, and $(z - 1 - w)$, so that the solution to the interpolation problem becomes the zero-order coefficient of the origin-shifted polynomial, ζ'_{calc}. This value is calculable simply as a 64-term inner product between the list of ζ_{obsd} and a vector of 64 'interpolation coefficients' whose values depend only on u, v, w, and on the choice of grid. (Linear interpolation can be viewed analogously as an eight-term inner product, though with the grid points unweighted, the calculation of the 'interpolation coefficients' is considerably simpler.)

To make interpolation affordable, it was assumed that positional errors of 1.5% of a grid unit were tolerable. Hence, a $32 \times 32 \times 32$ look-up table was constructed, representing values of u, v, and w rounded to the nearest 1/32 of a grid unit. Each table entry is loaded at the

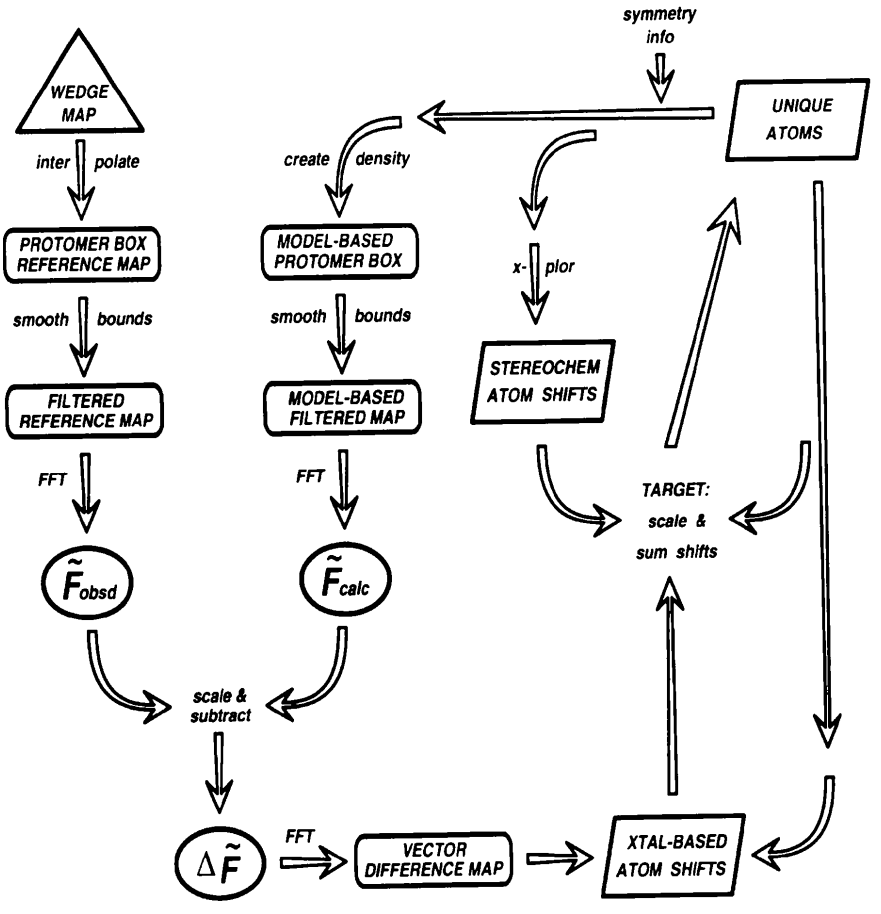


Fig. 1. A flow diagram summarizing the refinement of atomic positional parameters in the XX12 procedure. Detailed descriptions of most of the steps are given in the text. At the outset, electron-density values from a current version of the *WEDGE MAP* (upper left corner), obtained from the phase-constraint procedure, are expanded using icosahedral operators to fill the protomer-box volume, filtered, and Fourier transformed to create a set of reference structure factors (F_{obsd}) to be used as complex-valued standards for the refinement. Each refinement cycle begins with the most current copy of the atomic model (upper right corner). Expansion of these atom coordinates, using icosahedral operators, to fill the protomer-box volume, leads to a set of model-based structure factors (F_{calc}), which are scaled to the reference standards using resolution-dependent bin scales. The *VECTOR DIFFERENCE MAP* (bottom center) is obtained by Fourier transformation of the scaled differences between the two sets of structure factors. One set of *ATOMIC SHIFTS* (lower right corner) is obtained from the difference map by interpolation at the atom positions. Simultaneously, a second set of *ATOMIC SHIFTS* based solely on *STEREOCHEMISTRY* (center) is obtained by running the *X-PLOR* package, given knowledge of the unit cell and of icosahedral and crystallographic symmetry operators. Each cycle concludes with the *TARGET* routine (right center), which scales the two sets of shifts appropriately and updates the *ATOMIC MODEL* (upper right) with an improved set of coordinates.

beginning of the interpolation program with a vector of the 64 interpolation coefficients appropriate for the position (u, v, w). Carrying out this non-linear interpolation is then very similar to linear interpolation, as it requires converting the requested coordinate into map grid units, using the integer portion of the result to identify which portion of the input map the source values come from, and using the fractional portion of the result to determine the interpolation coefficients. The procedural difference, in this case, is that the interpolation vector is looked up, instead of calculated at run time.

2.2.3. *The protomer-box reference map.* A convenient map volume for interactive viewing and model building is a rectangular box measuring 150 × 120 × 90 Å (with 0.75 Å per grid unit) centered around a single chemically continuous copy of the poliovirus protomer. The box is about 15–20 Å larger than the protomer in each dimension so that when a smoothing filter (that is described below) is applied to the edges of the box, only the neighboring protomers will be affected, and not the central one. Furthermore, the number of grid units on each edge is chosen to have small prime factors to facilitate fast Fourier transformation of the contents of the box. After convergence of the phase-refinement procedure (that is, the application of NCS constraints to the observed amplitudes and starting phases), the protomer-box map is created by 64-point non-linear interpolation from the final wedge map. As detailed below, the protomer box map is also intended to be used as the reference standard for a global real-space refinement of the atomic model.

2.2.4. *The model-based protomer-box map.* The corresponding model-based electron-density map is produced by first applying 60 icosahedral coordinate transformations to the list of atoms which define the current atomic model of the protomer, and then discarding symmetry-generated atoms which are too far outside the limits of the protomer box to contribute to it. Electron density is placed on the 0.75 Å grid by a locally developed implementation of an algorithm devised by Ten Eyck (1977). Each atom contributing to the map is represented by the convolution of its individual isotropic temperature factor with a five-term Gaussian approximation to the atomic form factors (Ibers, Templeton, Vainshtein, Bacon & Lonsdale, 1985). The maximum extent of any such Gaussian permitted by the program is a 16-grid unit cube; thus symmetry-related atoms further than eight grid units from the edges of the protomer box can safely be ignored.

2.2.5. *Filtering the maps.* Prior to the use of the reference and model-based protomer-box maps in the refinement of the atomic model, both maps are multiplied by identical half-cosine-shaped weighting functions (Fig. 2). The filter leaves the central parallelepiped containing the protomer of interest entirely unaffected, but it gradually reduces the density values on the periphery of the map to zero. The filter is intended to remove some of the artifacts from the transform of the box map which result from modeling a non-periodic object as a periodic one. In addition, this procedure reduces the influence of those symmetry-related points which are present in the protomer box in high copy number. Although this filter has a profound effect on the low-resolution pseudo-structure factors, it is not expected to affect the relative scaling of the transforms, insofar as it is applied identically to both the observed and calculated maps.

2.2.6. *Scaling transforms and reporting statistics.* Prior to the calculation of the difference map, the Fourier transform of the model-based map must be scaled to the transform of the observed electron density. For this purpose, the data are divided into several (typically 16) shells at exactly equal intervals of d^{*2}, and thus nearly equal reciprocal volume. In each shell, the model-dependent data are assigned a single linear scale factor, using either the expression,

$$k_{\text{resol}} = \left\{ \text{Re}\left[\sum_{hkl} \tilde{F}_{\text{obsd}}(hkl) F^*_{\text{calc}}(hkl) \right] \right\} \div \left[\sum_{hkl} |F_{\text{calc}}(hkl)|^2 \right], \quad (6)$$

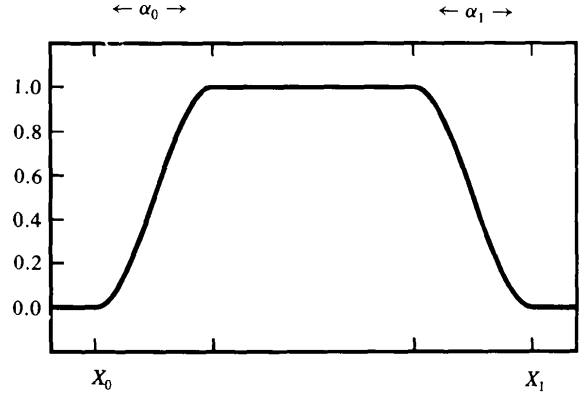


Fig. 2. The border-smoothing function, filter (x, y, z), which is applied to the reference and model-based protomer-box maps is the product of three one-dimensional functions: filter(x), filter(y), and filter(z). As an example, filter(x) has been plotted, assuming that x_0 and x_1 represent the limits of the protomer-box volume. The border widths, α_0 and α_1, which are not necessarily identical, have been chosen to be as large as possible, subject to the requirement that the filter is not permited to modify density values within a box tightly enclosing the central protomer [here represented in one dimension by the range from $(x_0 + \alpha_0)$ to $(x_1 - \alpha_1)$]. Density values located between x_0 and $(x_0 + \alpha_0)$ and those between x_1 and $(x_1 - \alpha_1)$ are damped when multiplied by the filter.

$$\text{filter}(x) = \begin{cases} 0 & \text{for } x \le x_0 \\ \frac{1}{2} - \frac{1}{2}\cos\{\pi(x - x_0)/\alpha_0\} & \text{for } x_0 \le x \le x_0 + \alpha_0 \\ 1 & \text{for } x_0 + \alpha_0 \le x \le x_1 - \alpha_1 \\ \frac{1}{2} - \frac{1}{2}\cos\{\pi(x_1 - x)/\alpha_1\} & \text{for } x_1 - \alpha_1 \le x \le x_1 \\ 0 & \text{for } x \ge x_1. \end{cases}$$

which minimizes in a least-squares sense the vector residual in (3), or the expression,

$$k_{\text{resol}} = \left[\sum_{hkl} |F_{\text{obsd}}(hkl)|\,|F_{\text{calc}}(hkl)|\right] \Big/ \left[\sum_{hkl} |F_{\text{calc}}(hkl)|^2\right], \tag{7}$$

which minimizes (2) when the phase of $\tilde{F}_{\text{obsd}}$ is assumed to be the same as the phase of $\tilde{F}_{\text{calc}}$. The resulting resolution dependence of the scaling function is satisfactory, provided that the scale factors vary sufficiently slowly from bin to bin, as is generally the case. In principle, rapid variation among neighboring scale factors would signal a need to change the number of bins (though in practice, this has never been necessary for virus structures at high resolution). A more common problem has occurred when attempting to use X-ray data near the diffraction limit of a particular crystal form. In that circumstance, the lack of correlation between the observed and calculated transforms causes the vector-based scale factor [(6)] in the highest resolution bins to tend to vanish. This is diagnostic, and provides a clear indication for decreasing the resolution limit of the calculation.

The use of resolution-dependent bin scaling in the refinement of atomic models represents a departure from the more common approach (*e.g.* in *PROLSQ* and *X-PLOR*) of using overall temperature factors to model the fall-off of intensity with resolution. Because a slightly larger number of parameters is used, it may also compensate for the non-Gaussian fall-off in intensity that is characteristic of macromolecules. In the refinement, the purpose of resolution-dependent bin scaling is to eliminate as much as possible of the discrepancy between the pseudo-observed and model-dependent transforms. One would like to avoid having any of the individual atomic parameters shift to compensate for what is essentially a scaling error.

In the refinement of virus models, the resolution-dependent bin scales compensate automatically for several factors besides the isotropic thermal motion of the molecule, including the distribution and relative contrast of bulk solvent and the presence of icosahedrally disordered nucleic acids. The use of 'averaged' electron-density maps as refinement standards also has an effect on scaling because the averaging process de-sharpens the reference image, because of both interpolation effects and imperfect knowledge of the position and orientation of the virus. In addition, relative scales in every resolution range are influenced by choosing how much partially ordered structure, both protein and solvent, to include explicitly in the atomic model. A simple two-parameter Gaussian scaling algorithm would necessarily be less effective in compensating for these influences.

2.2.7. *Vector-valued difference coefficients.* Once the appropriate resolution-dependent scale factors have been determined, it is simple to construct Fourier coefficients which express the discrepancy between the observed and model-based electron densities. Thus, assume that $\tilde{F}_{\text{pseudo-obsd}}$ and $\tilde{F}_{\text{pseudo-calc}}$ are defined as the Fourier transforms of, respectively, the filtered observed and filtered model-based protomer-box maps, where the filter is half-cosine smoothing function described above. Then the scaling program can be instructed to output scale-weighted vector difference coefficients of the form,

$$\Delta\tilde{F} = k_{\text{resol}}(\tilde{F}_{\text{pseudo-obsd}} - k_{\text{resol}}\tilde{F}_{\text{pseudo-calc}}). \tag{8}$$

Observe in (8) that the scale factor k_{resol} is included twice in these difference coefficients only when the vector difference map obtained from them is intended to be used in estimating the derivatives of $Q_{\text{pseudo-crystallographic}}$ with respect to the atomic parameters.

2.2.8. *Calculation of atomic positional shifts.* Atomic positional shifts for each cycle of XX12 are formed by a linear combination of the scaled pseudo-crystallographic and stereochemistry-based shifts. The derivatives of the phased quadratic residual [(3)] with respect to the parameters of the atomic model are approximated cheaply using a variant of the method of Freer *et al.*, (1975) by linear interpolation from the protomer-box difference map in the vicinity of each atom of the central protomer. The stereochemistry-based shifts are estimated using *X-PLOR* without crystallographic terms in the energy function by running eight cycles of Powell minimization. Except for the omission of crystallographic terms, the empirical energy-based minimization is set up exactly as it would be running *X-PLOR* refinement on the full virus structure, using the true crystallographic asymmetric unit and strict NCS constraints.

Once the energy-based and crystallographically based atomic shifts have been estimated, as above, a linear combination of the two sets of shifts serves as an adequate approximation to the gradient of Q_{total}, as defined in (4). The specific linear combination of the two shift vectors must be chosen to ensure, firstly, that the input model is not made worse by excessively large shifts in any cycle, secondly, that the average shift per cycle decreases nearly to zero as the process converges, and thirdly, that, at convergence, the model exhibits a root-mean-square (r.m.s.) bond-length error in the 0.012–0.014 Å range, with the other stereochemical restraints proportionately restrictive. To achieve this behavior in the refinement of poliovirus, an empirical schedule was developed for weighting the respective stereochemical and crystallographic contributions.

The desired schedule was developed by recognizing, firstly, that the *X-PLOR* stereochemical shift is the result of downhill steps in the stereochemical potential. Therefore, the use of its actual value is a safe way to

prevent over-shifting. Secondly, the pseudo-crystallographic shift is expected to be approximately proportional to the gradient of the pseudo-crystallographic potential, though the correct proportionality constant is unknown *a priori*. The proportionality constant is chosen such that the r.m.s. crystallographic shift remains smaller than the r.m.s. stereochemical shift for as long as the r.m.s. magnitude of the stereochemical shift exceeds the desired r.m.s. bond-length error. This should gradually improve the stereochemistry of the model. Once the stereochemical shift becomes smaller than the targeted bond length error, the r.m.s. magnitude of the pseudo-crystallographic shift is then scaled to match the r.m.s. stereochemical shift exactly. This anticipates that at convergence the two components of the gradient exactly cancel one another.*

When the two sets of shifts have been scaled as described, summed together, and applied to the input model, the resulting output model has necessarily been improved, and can serve as a suitable input model in the next refinement cycle. Typically, between six and 20 cycles are required to achieve convergence, depending on the quality of the starting atomic model.

Following refinement of the atomic positions, the user has the option to refine individual atomic temperature factors for the protein atoms, to refine individual occupancy parameters for ordered solvent, to return to interactive model building, or to re-initiate NCS averaging with phases based on the refined model to produce a new wedge map standard. It is worth emphasizing that the phases of the reference standard do not change while real-space refinement is going on. However, whenever model-based phases are used to re-initiate averaging, the model does have an impact on the reference phase set.

2.2.9. *Refinement of solvent occupancy factors.* Occupancies of bound solvent typically are refined after convergence of the positional parameters. (Fixed solvent molecules are assigned individual occupancies but not individual temperature factors.) The XX12 procedure normally refines solvent occupancies differently from the way that *X-PLOR* does. In each cycle of occupancy-factor refinement, a simple routine called *QREFIN* was used to change occupancy factors by a small pre-determined increment, with the sign of the change dependent upon the sign of the vector difference map when sampled at the atom position. The increment decreases as convergence is approached.

2.2.10. *Refinement of individual temperature factors.* The refinement of individual temperature factors can be carried out within the *X-PLOR* program, using a protomer-box-based approach, provided that an appropriate set of pseudo-structure factors has been constructed to use as standards for the refinement.* Unfortunately, it is not possible to use all of the electron density within the volume of the protomer box because that would require taking into account all of the atoms contributing to the box. *X-PLOR* would not deal gracefully with the haphazard collection of symmetry-related polypeptide fragments that would result. Instead, a workable compromise was developed which involved creating a tight model-based envelope around the central protomer to discard all of the electron-density values beyond a fixed distance from the central protomer, and any values located closer to another protomer than to the central one. It then is reasonable to pretend that the atoms of the central protomer, with no symmetry operators applied, entirely account for the electron density lying within the envelope. (The success of the method, detailed below, demonstrates that this approximation is adequate: temperature-factor refinement tolerates envelope-based artifacts and the lack of symmetry-generated neighbors much better than positional refinement does, partly owing to the irrelevance of non-bonded contact energies.)

Once the electron density of the protomer box had been masked in this way, it was simple to Fourier transform the contents of the box to produce a set of reference structure factors. The *X-PLOR* program was supplied with this reference set as a standard, together with the coordinates of the central protomer; and then was asked to determine what choices of restrained individual atomic temperature factors would minimize the 'phased' residual in (3).

The restraints, which limit the discrepancy in temperature factor (B) between covalently associated atoms, are controlled by assignable weights described in the *X-PLOR* manual (Brünger, 1992*a*). However, a problem was found with the default values for these weights determined automatically by *X-PLOR*, when using either the authentic or protomer-box-based reflection standards. As shown below, many individual B values dropped to the artificial lower limit of 2.00 Å^2. This made it impossible to assess how closely the results of full-cell and protomer-box-based refinements were correlated. To compensate, either of two approaches permitted temperature factors which were ordinarily

* As currently implemented, the gradient-scaling routine, referred to as '*TARGET*' in Fig. 1, accepts two control parameters as inputs to the program. The values used here, 0.02 Å for parameter 1 and 0.01 Å for parameter 2, were determined empirically to produce the desired behavior in poliovirus. First, if the r.m.s. stereochemical shift would exceed parameter 1, then all stereochemical shifts would be scaled down proportionately until the r.m.s. level matched parameter 1. This prevents large stereochemical errors from dominating the refinement in its earliest cycles. Second, the population of pseudo-crystallographic shifts is scaled linearly to an r.m.s. distance equal to the lesser of parameter 2 or the r.m.s. shift of the stereochemical population. To prevent refinement from being influenced excessively by outliers in the population, the '*TARGET*' routine also imposes a ceiling on individual stereochemical and crystallographic atomic shifts of three times the r.m.s. distance of their respective populations.

* Solvent occupancy factors can be refined by *X-PLOR* in a similar way. However, the *QREFIN* procedure yields better results.

Table 1. *Space requirements and computational expense of refining poliovirus atomic models by different methods*

Method	*X-PLOR*	XX12 positional	XX12 thermal
Cell (Å)	323 × 358 × 380	105 × 120 × 90	105 × 120 × 90
Grid size (Å)	0.72	0.75	0.75
Grid points in map	1.2×10^8	2.7×10^6	2.7×10^6
Storage required per FFT (Mbytes)	~900	~10	~10
Memory allocated (Mbytes)*	48	10	48
Crystallographic symmetry operators (No.)	4	1	1
NCS operators tested per atom	30	60	1
NCS operators applied per atom	30	3.85	1
Atoms needed for SF calculation	862560	27671	7188
Unique structure factors calculated	962646	97635	97635
CPU minutes for initial set-up	125 to generate H atoms §	22 for XX12,† 125 for *X-PLOR*	45 for XX12†
CPU minutes per refinement cycle	600–780 (positional) or 1100–1200 (B_{indiv})	45–60‡	7–9
Cycles to convergence for 2PLV	Positional, 50–60 Thermal, 20	50–60	20

* Tests of *X-PLOR* v3.1 often used three or four processors in parallel, so that actual elapsed times were smaller than those listed. Memory allocations, in a multi-user environment, were constrained to be smaller than those suggested by *X-PLOR* for optimal performance. † Set-up times for XX12 mainly are used for interpolation from the wedge map to create the protomer-box map, and for Fourier transformation to generate reference structure factors. Extra time for the set-up of temperature-factor refinement was used for calculating and applying a model-based envelope. ‡ Each cycle of XX12 included 28 cycles of *X-PLOR* positional refinement with no crystallographic terms: 20 cycles in which only H atoms were permitted to move, followed by eight additional cycles in which all atoms were refined. § A 10° φ-step was used in hbuild. Use of a 45° step reduces the time by a factor of 4.

below the artificial limit to become distinguishable. First, the structure-factor standards (based on either the protomer box or the authentic cell) could be de-sharpened by applying an artificial temperature factor to them prior to their use (see below). Alternatively, the individual atomic temperature factors could be prevented from dropping to physically unreasonable values by significantly increasing the restraining weights (data not shown).

2.3. *Crystallographic system under study*

The refinements in this paper were carried out on three strains of type 1 poliovirus: P1/Mahoney, V510 and VD9. All three strains crystallize in space group $P2_12_12$ with nearly identical cell dimensions ($a = 323, b = 358$ and $c = 380$ Å). Each unit cell contains two identical virus particles (with the second particle generated from the first by a 2_1 screw). Each particle has an icosahedrally symmetric protein shell consisting of 60 identical protomers, with each protomer including about 7200 ordered non-H atoms (including bound solvent). The particle is positioned in the unit cell such that one icosahedral twofold is coincident with a crystallographic twofold, which implies 30-fold NCS per crystallographic asymmetric unit.

2.4. *Full-cell refinements with X-PLOR*

The full-cell *X-PLOR* tests were run on a multi-processor SGI RISC3000 computer with 256 Mbytes of shared physical memory, operated as a time-sharing system with other users present. In these cases, *X-PLOR* was given atomic parameters for the atoms of a single protomer, the four $P2_12_12$ crystallographic symmetry operators, the true unit-cell parameters and the set of 30 NCS operators. Structure-factor calculations were carried out by *FFT*, with a grid 1/4 the resolution of the data actually observed. For 2PLV, V510 and VD9, the set of reflection standards was complete from 11 to 2.88, to 2.60, and to 2.80 Å resolution, respectively, requiring grid spacings of 0.72, 0.65 and 0.70 Å. The *FFT* routines were allowed to use 48 Mbytes of memory (with the remainder of the required space assigned as virtual memory) and three processors in parallel. Model-based structure factors and their derivatives with respect to the atomic parameters were re-computed for each minimization step, strict 30-fold NCS was enforced, and a 'phased' quadratic residual [(3)] was minimized, using the Fourier transform of the authentic icosahedrally constrained electron density as a standard.

Because this procedure was intended to improve the geometry of a model that already was essentially correct, the relative weight of the crystallographic term was taken as 1/4 of the value suggested by *X-PLOR*. This ensured that refinement improved the geometry of the model, possibly at the expense of the crystallographic residual.

2.5. *Experiments*

Two sets of experiments tested whether the XX12 procedure operating over the protomer box produced results consistent with those of *X-PLOR* operating over the full crystallographic asymmetric unit. In the first set of tests, the 'parallel' experiments, *X-PLOR* and XX12 were used independently to refine either atomic positional parameters or individual atomic temperature factors and solvent occupancies for a P1/Mahoney model whose atomic coordinates were taken from the PDB entry 2PLV. In the second set of tests, the 'series' experiments, XX12 positional refinement was followed by full-cell *X-PLOR* refinement, and then by XX12 again, using atomic models for the three strains of poliovirus discussed above. In every experiment, minimizations were carried out using the same set of stereochemical potentials, but with different crystallographic residuals. In principle, crystallographic residuals for the full-cell and small-box versions of the refinement might differ from one another slightly

because electron-density values in the crystallographic and pseudo-crystallographic asymmetric units are sampled with different multiplicities in the two refinements, because different methods are used for scaling the model-based structure factors to the observations, or because of differences in the particulars of sampling and interpolation.

Both the parallel and the series experiments were necessary to assess the usefulness of the XX12 protocol The series experiments asked whether the 'converged' parameters of refined models remained stable when subjected to the alternative procedure. That stability would be valuable if XX12 were used as an aid to interactive model-building, with the less expensive calculation run after each model-building session, and the more expensive procedure run only intermittently. In contrast, the parallel experiment asked whether both procedures would arrive at similar final answers when the starting model was somewhat further from the convergence point. That was intended to assess whether the less expensive procedure, by itself, would produce results sufficiently close to those of the full-cell refinement as to make the latter unnecessary. Collectively, these tests investigate how much variability is seen in the atomic models, determine which portions of the models exhibit the largest differences, and provide relative estimates of the computational costs.

3. Results

3.1. *Comparison of computation time*

The primary benefit of running XX12 rather than *X-PLOR* on the full crystallographic asymmetric unit is the significant reduction in the computation time, physical memory, and disk storage required (Table 1). In practice, a refinement that runs easily overnight using XX12 may take several days (or weeks) using a multiprocessor time-shared RISC3000 computer system. XX12 is therefore useful not only for crystallographic refinement *per se*, but also as an aid to model building, when run between model-building sessions. In principle, if the small-box algorithm were integrated into a model-building package, and a small enough box was chosen at each stage, the procedure also could be useful as an interactive model building tool.

3.2. *The refinement of positional parameters in parallel*

In this test, XX12 and full-cell *X-PLOR* were each used to refine the atomic positions of the 2PLV model. Previously, the 2PLV structure had been refined using a more primitive version of the protomer-box refinement which ignored van der Waals contacts, hydrogen bonds and variable torsion energies, and used a substantially different dictionary of stereochemical standards. Individual solvent occupancies in this starting model were determined by *QREFIN*. In the XX12 refinement, overall isotropic thermal motion was compensated for by applying optimized resolution-dependent bin scales to the model-based pseudo-crystallographic structure factors. For full-cell *X-PLOR*, an overall temperature factor ($B = 7.5$ Å^2) was calculated. The coordinates of this starting model were refined to convergence using both XX12 (over the protomer box) and *X-PLOR* (Over the $P2_12_12$ asymmetric unit).

Table 2. *R.m.s. deviation of the atomic coordinates from idealized stereochemical standards before and after refinement 'in parallel'*

Atomic model*	Before refinement	Refined using XX12	Refined using full-cell *X-PLOR*
Bond lengths (Å)	0.043‡	0.012	0.014
Angles (°)	4.4	2.5	2.7
Variable torsions (°)	22.1	19.1	21.0
Fixed dihedrals (°)	2.5	1.3	1.4
R_{cryst} (to 2.88 Å)†	0.244	0.224	0.227

* Models used in this comparison included overall temperature factors, not individual ones, though the model-based amplitudes were scaled to the observations using resolution-specific bin scales calculated as shown in (7). † The value of R_{cryst} reported here compares the Fourier transform of the atomic model with actual observed structure-factor magnitudes, rather than with the transform of the averaged map used as a refinement standard. $R_{cryst} = (\sum_{hkl} ||F_{obsd}| - |F_{calc}||)/(\sum_{hkl} |F_{obsd}|)$. ‡ Stereochemical discrepancies reported for the starting model, 2PLV, are unusually large because the stereochemical dictionary used to refine that model was substantially different from a more current dictionary used for this Table and for the XX12 and *X-PLOR* refinements (namely the PARAM19X parameter set). If instead, bond lengths were compared with the mean length of chemically similar bonds from the same model, using the program *GEOM* (Cohen, 1986), the r.m.s. deviations from the mean were similar in all three structures. This implies that all three structures were restrained to a similar extent.

Complex-valued structure-factor standards for both refinements were derived from the same 30-fold NCS-constrained electron-density map of the icosahedrally unique volume. This 'wedge' map was obtained after convergence of the phase-constraint procedure. As described in *Methods*, a 64-point non-linear interpolation from the wedge map was used to reconstruct the electron density of both the authentic asymmetric unit and the protomer-box pseudo-cell. For XX12, the $P1$ pseudo-cell was constructed using only NCS operators, with the application of a border-damping filter to the electron density prior to its Fourier transformation. Reference structure factors for *X-PLOR* were calculated by a space-group specific FFT from an authentic $P2_12_12$ asymmetric unit that was constructed using both crystallographic and NCS operators. Crystallographic R factors and r.m.s. deviations of the starting and final atomic parameters from idealized stereochemical standards are listed in Table 2.

The resulting models were compared pair-wise by calculating the distances between corresponding atoms from the input 2PLV model, and the XX12- and *X-PLOR*-refined models (Fig. 3). Clearly, the overwhelming majority (97.8%) of the 854 residues are extremely similar in the two refined models, having average differences of 0.20 Å or less in the main chain (Fig. 3*c*). These differences are so small that the

corresponding atoms cannot readily be distinguished from one another by visual inspection of computer graphics. In contrast, a very small fraction of the residues, corresponding to the labeled peaks in Fig. 3, can be distinguished visually, though the overall conformations of these residue are still qualitatively quite similar. Not coincidentally, all of these residues have been built into relatively disordered regions of the electron-density map. In particular, Thr1010, Glu2005 and Ser4016 are each the last residue of a polypeptide chain to be considered sufficiently reliable to be included in the model. [In fact, the correct sequence numbering for the polypeptide segment 1006–1010, which includes Thr1010, is not known because the density is of insufficient quality to warrant sequence assignment (Filman *et al.*, 1989).] These movable

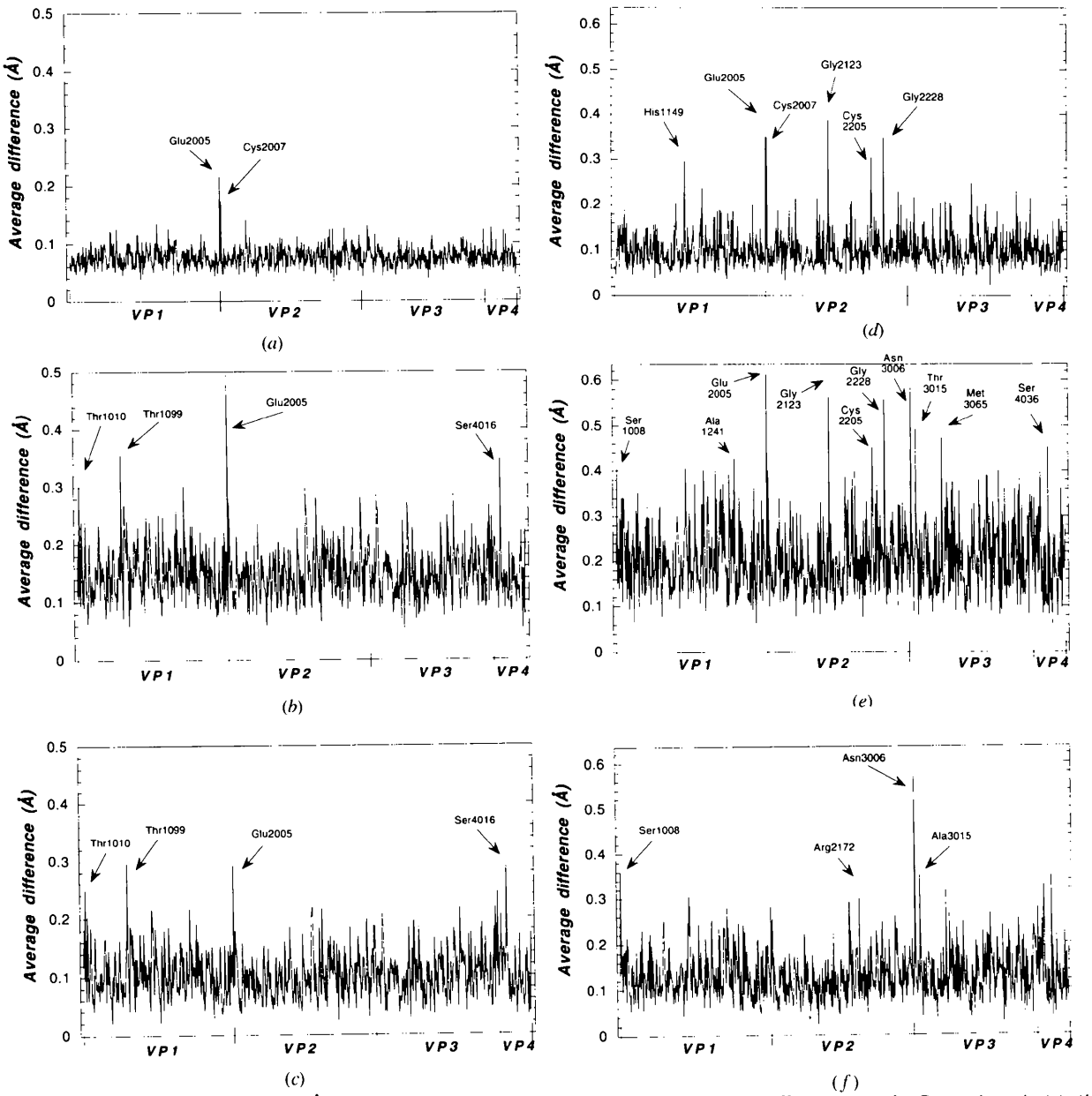


Fig. 3. Average positional differences (Å) per residue between P1/Mahoney models refined using different protocols. Comparisons in (*a*), (*b*) and (*c*) involve main-chain atoms, and those in (*d*), (*e*) and (*f*) involve side-chain atoms. (*a*) and (*d*) show differences caused by refining the starting model, 2PLV, using full-cell *X-PLOR*. (*b*) and (*e*) show differences as a result of refining 2PLV with XX12 instead. In (*c*) and (*e*) these two refined models have been compared with one another. Residue labels are arranged along the abscissa in numerical order. Labels beginning with a 1, 2, 3 or 4 belong to capsid proteins VP1, VP2, VP3 and VP4, respectively. Specific residues with atypically large differences have been identified, and are discussed in the text. Most of these large differences occur in partially ordered portions of the crystal structure.

residues are all unusual in that their positions are not strongly tethered either by well defined electron-density features or by the presence of modeled residues on both sides. Significant main-chain differences from the middle of a polypeptide chain include Thr1099, which is located in a portion of the solvent-exposed *BC* loop with unusually poorly defined electron density.

To confirm that only very weakly ordered portions of the structure are involved in these differences between the refined models, individual isotropic temperature factors for the affected atoms were determined in a separate experiment. Thr1099 and Ser4016 had refined *B* values of 51 and 60 Å^2, respectively. Thr1010 and Glu2005 had somewhat lower refined *B* values of 29 and 35 Å^2, primarily because their sites are thought to be only partially occupied, and therefore arbitrary occupancies of 0.5 had been assigned to them prior to temperature-factor refinement.

The side chains that *X-PLOR* changed the most (His1149, Glu2005, Cys2007, Gly2123, Cys2205 and Gly2228) were also changed significantly by XX12 (Figs. 3*d*–3*f*). In visual comparisons of the models with electron density, the XX12 model provided a noticeably better fit than either 2PLV or the *X-PLOR* model for residues Gly2123, Cys2205 and Gly2228. A number of additional side chains were shifted significantly by XX12, but not by full-cell *X-PLOR*. These included Ala1241, Asn3006, Thr3015 and Met3065 (Fig. 3*e*). The XX12 refinement produced a slightly better fit to the electron density than the 2PLV and *X-PLOR* models did in the vicinity of Ala1241, Thr3015 and Met3065. However, the shift of the Asn3006 side chain by XX12 clearly makes the model worse, and represents the single largest discrepancy (a 0.58 Å average difference) between the *X-PLOR* and XX12-refined models (Fig. 3*f*).

Unlike other main-chain differences and several of the side-chain differences, which can be explained easily by the presence of weak electron density and/or lack of tethering at chain termini, the difference for the side chain of Asn3006 represents the sole example of an undesirable behavior by an XX12 refinement. Errors in the refinement of Asn3006 were avoidable (in retrospect) and were caused by unusual circumstances. Briefly, five symmetry-related copies of the Asn3006 side chain form a ring, surrounding a density peak on

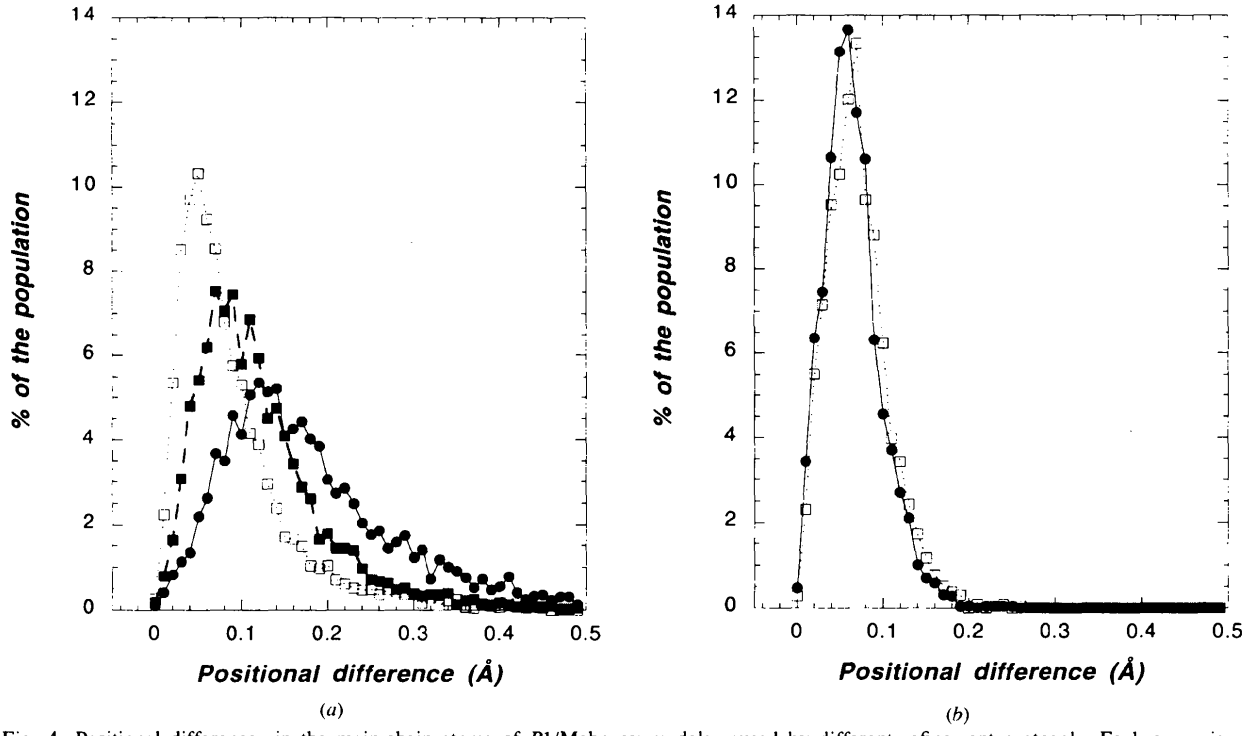


Fig. 4. Positional differences in the main-chain atoms of P1/Mahoney models caused by different refinement protocols. Each curve is a histogram, representing a pair-wise comparison between two atomic models, and showing the frequencies of occurrence of various positional differences among the population of atoms. In (*a*), which illustrates a 'parallel' experiment, the open squares represent changes in the 2PLV model caused by 60 cycles of refinement with full-cell *X-PLOR*. Filled circles show changes as a result of the refinement of 2PLV using XX12, and filled squares indicate differences between the two final models. Though all of the changes were small, on average, XX12 caused the most extensive changes, partly due to more stringent stereochemical restraints, but also conceivably due to the failure of full-cell *X-PLOR* to converge. (*b*) To assess whether the full-cell *X-PLOR* refinement had been terminated prematurely, atomic positions from the final model (open squares) were compared with positions ten cycles earlier (filled circles). These distributions were nearly identical, suggesting that additional refinement was not required.

the fivefold axis. The peak is quite large, taking up a volume more typical of a solvated ion than a water molecule. Though suspected to be a rotationally disordered ion, its true chemical identity is unknown, and it has been modeled as a water whose adjustable occupancy parameter refines to a value significantly greater than 1.0 (when five superimposed symmetry-generated copies are taken into account). Conformational adjustments in Asn3006 making the ring smaller were coupled with an obviously incorrect translation of the ion outward along the fivefold axis, away from its central position in the density peak.

Tests conducted in retrospect have shown that if an appropriate individual temperature factor had been assigned to this unusual peak to account for its breadth, then the XX12 results would have closely approximated the 2PLV and *X-PLOR* models in this area. Assigning a single overall temperature factor to the entire solvent population only caused a problem here because the occupancy of the site and the discrepancy in *B* both were very large at the same time. The resulting ripple in the difference map led to spurious shifts in the side chains, followed by incorrect, but compensating changes in the coordinates of the ion. Here, the use of difference maps to estimate partial derivatives for the refinement (a computational short-cut) caused an anomaly that an XX12 user could easily identify and correct.

One surprising result of the parallel experiment was that XX12 appears to have moved the atoms of *P*1/Mahoney further from their starting positions, on average, than full-cell *X-PLOR* did (Figs. 3 and 4*a*). This result is somewhat counter-intuitive, since one might expect that 2PLV, a model which had been refined using the original protomer-box protocol, using different stereochemical standards, would be most dissimilar to the model resulting from a full-cell *X-PLOR* refinement. It should be noted, however, that in all cases, the overall r.m.s. differences were small with respect to the 0.2–0.3 Å expected error in a 2.9 Å structure.

One explanation for this behaviour is that both refinement methods shifted many atoms in the same direction, but that full-cell *X-PLOR* shifted them to a less extent. This is consistent with the tendency of the XX12-refined model to resemble the *X-PLOR* model more closely than it does 2PLV (Fig. 4*a*), and with the slightly tighter restraints applied to the XX12 model (Table 2). Alternatively, it is conceivable that 60 cycles of full-cell refinement were insufficient to reach convergence. This is less likely, however, given the marked similarity between the *X-PLOR*-refined models after 50 and 60 cycles of refinement (Fig. 4*b*).

3.3. *The refinement of positional parameters in series*

The second set of experiments was designed to evaluate the extent to which atomic models refined to convergence by the XX12 procedure would remain stable upon further refinement with *X-PLOR*. Atomic models of *P*1/Mahoney, V510, and VD9 which had undergone several rounds of NCS-based phase constraints, atomic parameter refinement with XX12, interactive model building, and a concluding cycle of XX12 positional refinement were used as starting points for the test. Unlike the parallel refinements reported above, these refined models included individual isotropic temperature factors for the atoms of the protein and lipid ligands, in addition to the usual assignment of individual occupancy factors to the fixed solvent molecules. To assess the stability of the model parameters refined by the XX12 procedure, each of these three starting models (labeled 'stage 1') was subjected to 30–40 additional cycles of positional refinement with *X-PLOR* operating over the crystallographic asymmetric unit (yielding the models designated 'stage 2'). A final refinement of the 'stage 2' models using XX12 then yielded the models designated as 'stage 3'. Table 3 presents crystallographic and stereochemical agreement statistics for each refinement stage of the three structures.

Table 3. *Crystallographic and stereochemical statistics for the atomic positional refinements of PLV, V510 and VD9 by the sequential use of XX12 (stage 1), X-PLOR (stage 2) and again XX12 (stage 3)*

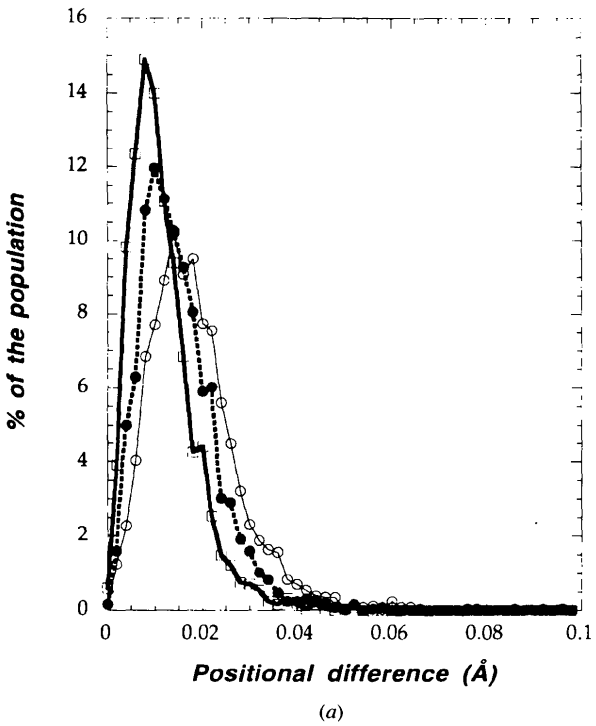
Virus	Stage	Bond lengths	Bond angles	Variable torsions	Fixed torsions	R_{cryst}*
PLV	1	0.012	2.472	18.335	1.242	.1885
	2	0.011	2.429	18.276	1.252	.1867
	3	0.012	2.453	18.272	1.253	.1882
V510						
	1	0.013	2.513	18.348	1.281	.2532
	2	0.012	2.525	18.375	1.229	.2519
	3	0.013	2.524	18.315	1.268	.2515
VD9						
	1	0.012	2.515	18.341	1.251	.2585
	2	0.012	2.507	18.375	1.236	.2565
	3	0.012	2.512	18.311	1.305	.2577

* R_{cryst} compares the Fourier transform of the atomic model with observed structure-factor magnitudes, and is defined in the legend of Table 2. For PLV, V510, and VD9, respectively, the R_{cryst} calculation includes data from 30 Å to 2.88, 2.60 and 2.80 Å.

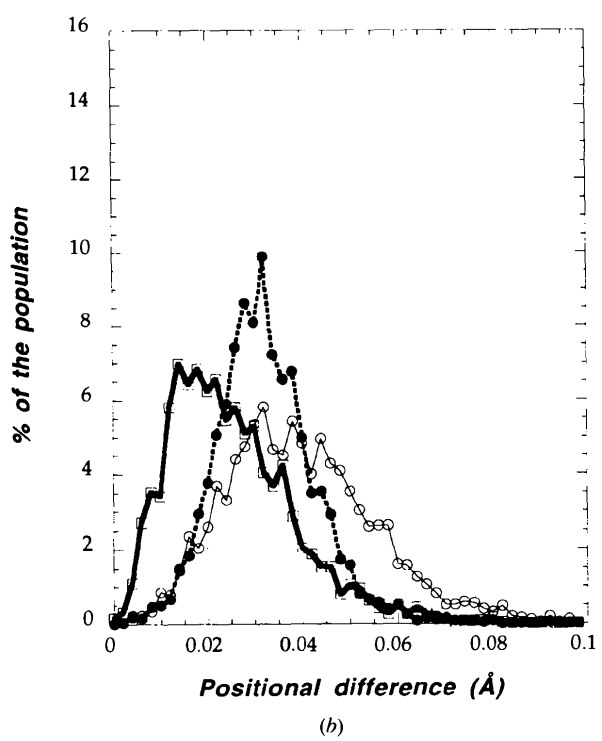
The models produced at each stage of the refinement test were quite similar to one another. Each part of Fig. 5 is a histogram that describes the refinement of one of the three virus structures by showing pairwise comparisons between models at successive stages of refinement. For 2PLV and VD9, shifts in excess of 0.04 Å were very rare; and most of the shifts occurred in the 0.01–0.02 Å range. For V510 (Fig. 5*b*), the shifts were consistently larger, though still quite small compared to the expected error of the experiment. The vast majority of atoms were shifted much less than 0.1 Å, an insignificant difference in the atomic positions given the resolution of the data. The small sizes of these positional differences is consistent with

the fact that XX12 invokes *X-PLOR* to provide stereochemical restraints, and it confirms that the crystallographic potentials are very similar, even though XX12 and *X-PLOR* implement their gradients differently.

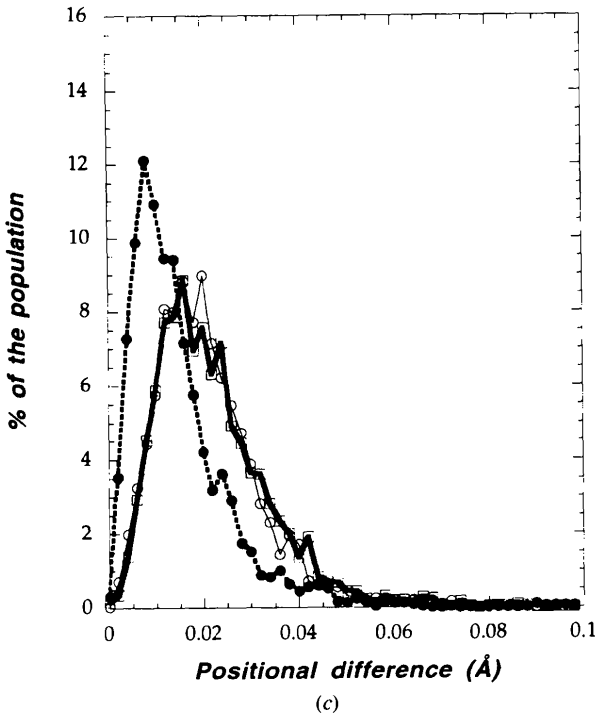
For all three viruses, the plots comparing stages 1 and 2 were skewed further to the right than the plots comparing stages 1 and 3. Thus, the full-cell *X-PLOR* refinement produced atomic shifts in the XX12 model which were partially reversed by re-refining with



(*a*)



(*b*)



(*c*)

Fig. 5. Changes in the positions of main-chain atoms of various poliovirus models caused by the application of different refinement protocols in series. Each plot is a histogram, indicating the fraction of the atoms whose positional shifts fall within a given range of distances. (*a*), (*b*) and (*c*) represent the *P*1/Mahoney, V510 and VD9 structures, respectively. Each of the 'series' experiments began by refining some initial model to convergence with the XX12 procedure, continued with a further refinement using full-cell *X-PLOR*, and concluded with a final XX12 stage. In each plot, open circles represent changes in the initial XX12-refined models caused by full-cell *X-PLOR*. Open squares represent further changes in *X-PLOR*-refined models caused by the final refinements using XX12. Closed circles show the cumulative differences between the initial and final XX12-refined models, and thus indicate whether the effects of full-cell refinement were reversible. Although the magnitudes of the atomic shifts varied among these three examples, observe that changes caused by *X-PLOR* (open circles) were consistently the largest. Understandably, the smaller changes seen in examples (*a*) and (*c*) proved to be more readily reversible than the larger changes seen in (*b*).

XX12. Because the stereochemical potentials applied in both procedures are nearly identical (Table 3), this result suggests that the crystallographic residuals minimized by the two approaches are similar to one another, but not absolutely identical.

To illustrate how small the atomic shifts actually were, Fig. 6(*a*) shows the largest positional shift in V510, obtained from the comparison between stages 1 and 2. From this result it can be concluded that the models produced by the faster XX12 procedure were not altered in a very meaningful way by further refinement with *X-PLOR*. In addition, closer inspection revealed that some of the largest changes in V510 involved moving an incorrectly placed atom in the same direction in both the *X-PLOR* and XX12 refinements. Fig. 6(*b*) shows a picture of just such a progression of shifts for one atom in V510. Here, the 'correct' direction for the atomic shift can be assessed by visual inspection of the shape of the electron-density feature associated with the carbonyl O atom of Val1244. In this instance, *X-PLOR* had moved the atom in the correct direction, but not far enough, so the shift by the second run of XX12 proceeded in the same direction. Presumably, the slowness of the change in this particular case was because of some difficulty in propagating compensating structural adjustments to other portions of the structure.

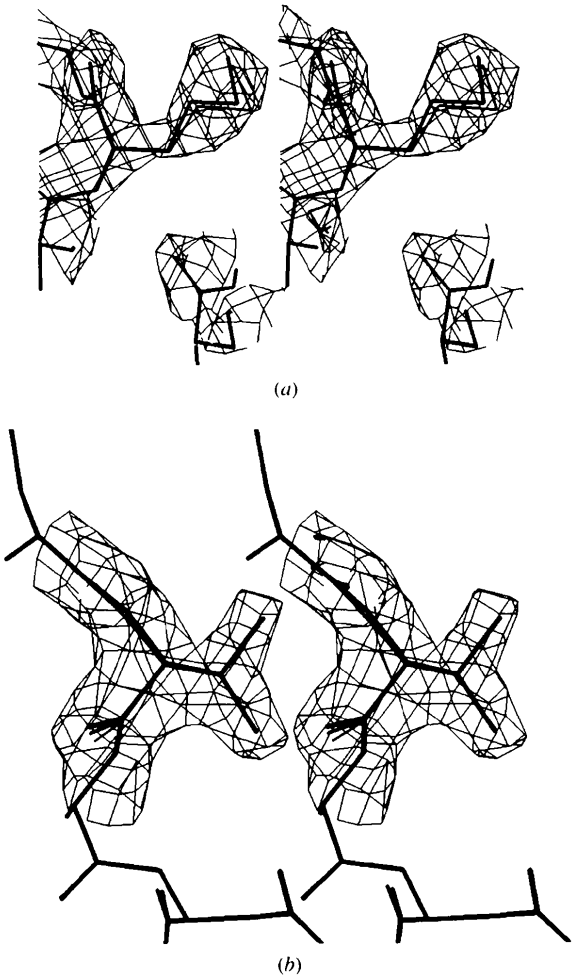


Fig. 6. Stereoviews of portions of the V510 atomic model, illustrating changes caused by refinement. (*a*) Met3149, which exhibits the single largest atomic difference (0.45 Å) caused by applying full-cell *X-PLOR* refinement (thick lines) to a model previously refined using XX12 (thin lines). Both conformations represent very similar interpretations of the electron density. (*b*) The conformation adopted by Val1244 after successive refinement steps. The thin, intermediate, and thick lines, respectively, show the model after an initial refinement with XX12, a subsequent refinement with full-cell *X-PLOR*, and a final refinement with XX12. Observe that all three refinement steps have tended to move the carbonyl O atoms in the same direction, towards the electron-density bulge at the left of the panel.

3.4. *Temperature-factor and solvent-occupancy tests*

To determine whether the protomer-box-based refinement of temperature factors was satisfactory, temperature factors for *P*1/Mahoney were refined in parallel experiments using either XX12 or full-cell *X-PLOR*, and then compared using scatter plots (Figs. 7*a* and 7*b*). Both unmodified and de-sharpened reference standards were tried. The values obtained from the two refinement protocols are correlated very strongly with one another, and give very similar indications of which portions of the structure are the most and the least mobile.

To verify that the occupancy-determining procedure was working, a more extensive refinement was carried out alternately improving individual atomic temperature factors for the protein atoms and individual occupancies for the solvent atoms, until the parameter values became stable. This test (Figs. 7*c* and 7*d*) had started with a *P*1/Mahoney model with atomic coordinates that had been refined to convergence with XX12. To minimize the effect of model bias, all of the solvent molecules were assigned initial occupancies of 0.20. From that starting point, individual temperature factors for the protein atoms and individual occupancies for the solvent molecules were refined in parallel experiments by *X-PLOR* over the full cell and by XX12 over the protomer box. The same icosahedrally constrained wedge map was expanded in different ways and Fourier transformed to provide complex-valued reference standards for both the *X-PLOR* and XX12 refinements. When *X-PLOR* was used to refine temperature factors, an artificial temperature-factor increment of 15 $Å^2$ was applied to the full-cell reference, or 25 $Å^2$ to the protomer-box reference. This allowed individual atomic temperature factors to be distributed over a wider range of values without being affected by the 2.00 $Å^2$ lower limit.

Once solvent occupancies were permitted to refine, individual temperature factors refined by the two methods showed a slightly improved correlation, partly

as a result of optimal de-sharpening of the standards. Figs. 7(*c*) and 7(*d*) each shows a scatter plot comparing individual atomic temperature factors (*B*) or refined solvent occupancies (*q*) refined by XX12 with those calculated by full-cell *X-PLOR*. A least-squares line is plotted in both cases, with a correlation coefficient of 0.94 for the occupancy test, and a correlation coefficient of 0.98 for the temperature-factor test. These high

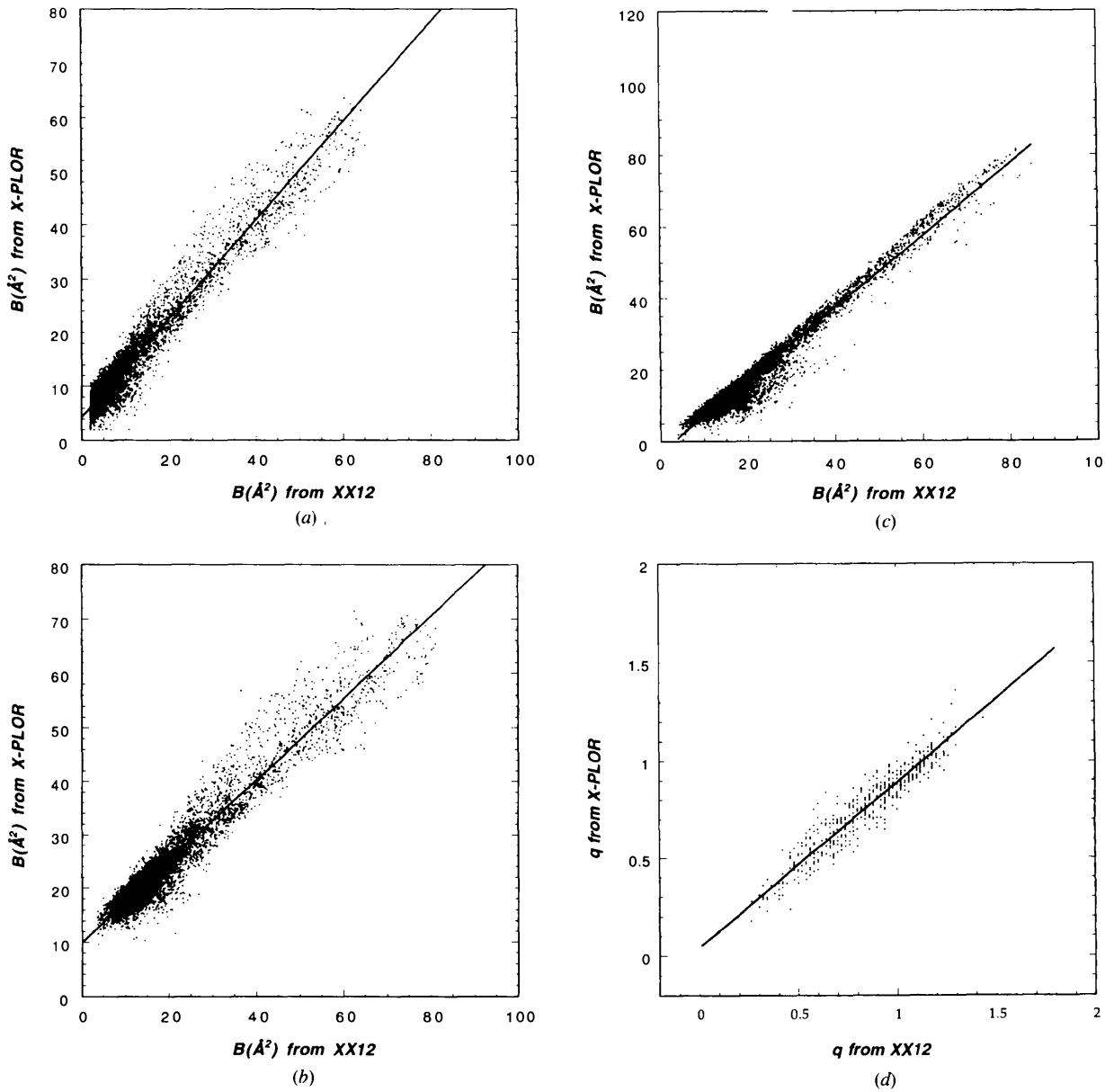


Fig. 7. Individual atomic temperature factors (*B*) and solvent occupancies (*q*) refined using XX12 are compared with those refined using full-cell *X-PLOR*. In each scatter plot, each point represents one non-H atom. In (*a*) and (*b*), solvent occupancies were fixed in advance, and then temperature factors for the same *P*1/Mahoney model were refined using either 20 cycles of XX12 (abscissa) or 12 cycles of full-cell *X-PLOR* (ordinate), with the restraining weight determined automatically by *X-PLOR* in both cases. A tight model-based envelope around the central protomer was always applied to the electron density before its use as a standard in XX12, as described in the text. Complex-valued structure-factor standards for both refinements were derived from the same NCS-based phase constraint procedure. In (*a*), no artificial overall temperature factor was applied to either set of structure-factor standards, while in (*b*) an overall temperature factor of 20.0 Å^2 was applied to both sets. (*c*) and (*d*) represent a separate experiment in which occupancies for the fixed solvent atoms were permitted to refine alternately with restrained individual isotropic temperature factors for the non-solvent atoms, until convergence was approached. The line shown in each plot was obtained by unweighted linear regression, and the goodness of fit was assessed using the standard linear correlation coefficient, r_{corr}. r_{corr} values were 0.97, 0.97, 0.98, and 0.94 in (*a*), (*b*), (*c*) and (*d*), respectively. $r_{corr} = (\sum_{xy} - \sum_x \sum_y)/\{[(\sum_{xx} - \sum_x \sum_x)(\sum_{yy} - \sum_y \sum_y)]^{1/2}\}$.

correlations indicate that the XX12 pseudo-cell procedure produces qualitatively similar results to those of the more expensive full-cell procedure, but not precisely identical numerical values. This should be sufficient for most purposes, as both methods give similar indications of the relative mobilities (or disorder) of different portions of the structure. Regardless, the individual temperature factors reported for a virus structure cannot be interpreted directly in terms of mean atomic displacements. Whenever the transform of the averaged map is used as a standard for refinement, as done here, the resolution-dependent fall-off in intensities is likely to be affected by the smoothing effects of interpolation during map averaging, and by subtle errors in particle position and orientation. Furthermore, the resolution-dependent bin scales used in XX12 can act as a temperature-factor increment or decrement, shifting the individual B values collectively.

4. Discussion

Three distinct poliovirus strains were each real-space refined by two different methods, using structure-factor standards derived from the same wedge map in both procedures. In all three strains, the final refined atomic models were essentially identical throughout the well ordered portions of the map, and most of the less well ordered portions. Minor differences in atomic position may have arisen from several possible causes. For example, various points in the crystallographic asymmetric unit are sampled by the protomer box with differing multiplicities. In addition, XX12 differs from *X-PLOR* with regard to the algorithm it uses to approximate the derivatives of the stereochemical and crystallographic potentials. Perhaps the most important cause of differences in the models involves the unavoidable use of discrete sampling in the calculation by Fourier transformation of maps and derivatives. In particular, the grids used by XX12 and *X-PLOR* must remain relatively coarse for the calculations to be affordable, and the grids are not identical (Table 1). Moreover, *X-PLOR* interpolates values (*i.e.*, derivatives for each atom) using points from one particular portion of the asymmetric unit map. The exact result therefore depends on an arbitrary choice of which protomer is considered the canonical one. In contrast, XX12 uses maps that are calculated on regular Cartesian grids and created using an image-smoothing interpolation method. Nonetheless, the results of the 'parallel' experiments demonstrate that these differences in the mechanical aspects of the computation are too small to have a significant effect on well ordered portions of the structure, but may cause noticeable differences in the poorly ordered portions. Residues near chain termini are particularly affected because the stereochemical portion of the refinement residual expresses a much weaker preference than usual regarding their locations.

Even with recent advances in computing, the full-cell refinement of type 1 poliovirus still uses significant resources and requires weeks on modern work stations. In contrast, short-cuts used in the pseudo-cell approach enable the refinements of poliovirus models to run much faster while producing equally valid results. For viruses the size of poliovirus, this makes refinements inexpensive enough to run routinely between model-building sessions. Moreover, the affordability of the approach may be crucial for viruses having a larger number of subunits than poliovirus, such as SV40 (Liddington *et al.*, 1991) or polyomavirus (Stehle, Yan, Benjamin & Harrison, 1994), or for structures having a similar number of atoms, but more extensive NCS and correspondingly larger unit cells, such as coxsackievirus B3 (Muckelbauer *et al.*, 1995).

None of the refinements described in this manuscript resulted in an interpretation of electron density that was qualitatively different from that of the input models. In every instance, XX12 and *X-PLOR* improved the agreement with both crystallographic and stereochemical standards. The *X-PLOR* and XX12-refined models were substantially similar, though for any given atom, either model may have exhibited a better fit to the density. However, the actual differences were negligible, implying that the more expensive full-cell refinement usually can be dispensed with. Finally, the ability of the protomer box based approach to emulate the results of full-cell temperature-factor and solvent-occupancy refinement with *X-PLOR* has been demonstrated.

4.1. *A brief comparison with protein refinement*

The refinement of atomic models of virus structures differs from the refinement of most small proteins in several fundamental ways. In the conventional reciprocal-space refinement of a protein, only the structure-factor amplitudes are reliable, the number of observations per parameter is small, and the introduction of stereochemical restraints improves this ratio by decreasing the number of degrees of freedom of the model. Additionally, the phases of the reference structure factors are completely biased by the trial model. Omit calculations and cross validation [*e.g.*, the $R_{\rm free}$ statistic of Brünger (1992*b*)] represent two partially successful ways to detect and overcome the effects of model bias.

Refining a structure with a high degree of exact NCS is a wholly different matter. Because the number of unique parameters is quite small, the problem is vastly overdetermined. Experimental phase information is unusually reliable and can be incorporated into the structure-factor standards. This makes refinement more robust and makes model bias must less of a problem. In general, major errors in the model are

detected easily by comparisons with the averaged F_o map (or an averaged omit map if the density feature is weak). However, R_{free} is meaningless: the mere presence of NCS in the crystal induces a strong interdependence among the structure-factor amplitudes that makes the selection of a truly independent trial set impossible.

5. Conclusions

Crystallographic refinement of icosahedral viruses is a computationally intensive task when the contents of the entire crystal unit cell are considered. By restricting one's focus to a small portion of the averaged electron-density map, the protomer-box map, the positional, thermal and occupancy parameters of the atomic model can be refined to convergence in hours. This compares to several days or weeks required by *X-PLOR* operating on the actual unit cell of the crystal.

The treatment of the protomer box as a crystallographic pseudo-cell allows for the use of crystallographic refinement packages other than *X-PLOR* to provide empirical energy-based positional shifts. Furthermore, the use of a model-based enveloping routine has made it possible to take advantage of other refinement capabilities, such as temperature-factor refinement and simulated annealing, without requiring modification of the source code.

For ease in implementation, the XX12 procedure was assembled as a shell script* connecting several pre-existing programs that communicate *via* disk files. In spite of the significant inefficiencies which result from this method of communication, the XX12 approach is still an order of magnitude faster than optimized *X-PLOR* code. Further increases in the speed of the XX12 algorithm will certainly be realized.

This manuscript includes work carried out by DHJ in partial satisfaction of the requirements of the degree Doctor of Philosophy in Chemistry from the University of California, San Diego. Professor R. F. Doolittle is thanked for his patience and encouragement. The authors thank R. Grant, C. Hiremath, D. Joseph-McCarthy, R. Syed and R. Basavappa for testing the procedure and providing useful suggestions, and thank D. Joseph-McCarthy and M. Wien for critical reading of the manuscript. This work was supported by NIH grant AI20566 (to JMH) and by a grant from the Keck Foundation to Harvard Medical School to establish a structural biology facility in the Department of Biological Chemistry and Molecular Pharmacology.

* The shell script is available from the authors on request.

References

Acharya, R., Fry, E., Stuart, D., Fox, G., Rowlands, D. & Brown, F. (1989). *Nature (London)*, **337**, 709–716.
Agarwal, R. C. (1978). *Acta Cryst.* A**34**, 791–809.
Arnold, E. & Rossmann, M. G. (1988). *Acta Cryst.* A**44**, 270–282.
Bricogne, G. (1974). *Acta Cryst.* A**30**, 395–405.
Bricogne, G. (1976). *Acta Cryst.* A**32**, 832–847.
Brünger, A. T. (1989). *Acta Cryst.* A**45**, 42–50.
Brünger, A. T. (1992*a*). *X-PLOR. Version 3.1. A system for X-ray crystallography and NMR*. Yale University, Connecticut, USA.
Brünger, A. T. (1992*b*). *Nature (London)*, **355**, 472–475.
Cohen, G. H. (1993). *J. Appl. Cryst.* **26**, 495–496.
Crowther, R. A. (1969). *Acta Cryst.* B**25**, 2571–2580.
Filman, D. J., Syed, R., Chow, M., Macadam, A. J., Minor, P. D. & Hogle, J. M. (1989). *EMBO J.* **8**, 1567–1579.
Freer, S. T., Alden, R. A., Carter, C. W. & Kraut, J. (1975). *J. Biol. Chem.* **250**, 46–54.
Fry, E., Acharya, R. & Stuart, D. (1993). *Acta Cryst.* A**49**, 45–55.
Grant, R. A., Filman, D. J., Fujinami, R. S., Icenogle, J. P. & Hogle, J. M. (1992). *Proc. Natl. Acad. Sci. USA*, **89**, 2061–2065.
Hendrickson, W. A. (1985). *Methods Enzymol.* **115**, 252–270.
Hendrickson, W. A. & Lattman, E. E. (1970). *Acta Cryst.* B**26**, 136–143.
Hogle, J. M., Chow, M. & Filman, D. J. (1985). *Science*, **229**, 1358–1365.
Ibers, J. A., Templeton, D. H., Vainshein, B. K., Bacon, G. E. & Lonsdale, K. (1985). In *International Tables for X-ray Crystallography*, edited by C. H. MacGillavry, G. D. Rieck & K. Lonsdale, pp. 201–246. Dordrecht: D. Reidel.
Jacobson, D. H. (1995). PhD thesis, University of California, San Diego, USA.
Jacobson, D. H., Filman, D. J., Martin, A., Girard, M. & Hogle, J. M. (1996). In preparation.
Jones, T. A. (1978). *J. Appl. Cryst.* **11**, 268–272.
Jones, T. A. (1985). *Methods Enzymol.* **115**, 157–171.
Jones, T. A. & Liljas, L. (1984). *Acta Cryst.* A**40**, 50–57.
Konnert, J. H. & Hendrickson, W. A. (1980). *Acta Cryst.* A**36**, 344–350.
Larson, S. B., Koszelak, S, Day, J., Greenwood, A., Dodds, J. A. & McPherson, A. (1993). *Nature (London)*, **361**, 179–182.
Liddington, R. C, Yan, Y., Moulai, J., Sahli, R., Benjamin, T. L. & Harrison, S. C. (1991). *Nature (London)*, **354**, 278–284.
Main, P. (1967). *Acta Cryst.* **23**, 50–54.
Muckelbauer, J. K., Kremer, M., Minor, I., Diana, G., Dutko, F. J., Groake, J., Peaver, D. C. & Rossmann, M. G. (1995). *Structure*, **3**, 653–667.
Rees, D. C. & Lewis, M. (1983). *Acta Cryst.* A**39**, 94–97.
Rossmann, M. G. & Blow, D. M. (1963). *Acta Cryst.* **16**, 39–45.
Silva, A. M. & Rossmann, M. G. (1985). *Acta Cryst.* B**41**, 147–157.
Stehle, T., Yan, Y., Benjamin, T. L. & Harrison, S. C. (1994). *Nature (London)*, **369**, 160–163.

Syed, R., Filman, D. J. & Hogle, J. M. (1995). Coordinate set 1PVC in the Protein Data Bank, Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Ten Eyck, L. F. (1973). *Acta Cryst.* A**29**, 183–191.

Ten Eyck, L. F. (1977). *Acta Cryst.* A**33**, 486–492.

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489–501.

Wu, H., Keller, W. & Rossmann, M. G. (1993). *Acta Cryst.* D**49**, 572–579.

Yeates, T. O., Jacobson, D. H., Martin, A., Wychowski, C., Girard, M., Filman, D. J. & Hogle, J. M. (1991). *EMBO J.* **10**, 2331–2341.