Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 1 August 2003
Accepted 3 October 2003

On a fast calculation of structure factors at a subatomic resolution

P. V. Afonine[a,b] and A. Urzhumtsev[a]*

[a]LCM[3]B, UMR 7036 CNRS, Faculté des Sciences, BP 239, Université Henri Poincaré, Nancy 1, Vandoeuvre-lès-Nancy 54506, France, and [b]Centre Charles Hermite, LORIA, Villers-lès-Nancy 54602, France. Correspondence e-mail: alexander.ourjoumtsev@lcm3b.uhp-nancy.fr

In the last decade, the progress of protein crystallography allowed several protein structures to be solved at a resolution higher than 0.9 Å. Such studies provide researchers with important new information reflecting very fine structural details. The signal from these details is very weak with respect to that corresponding to the whole structure. Its analysis requires high-quality data, which previously were available only for crystals of small molecules, and a high accuracy of calculations. The calculation of structure factors using direct formulae, traditional for 'small-molecule' crystallography, allows a relatively simple accuracy control. For macromolecular crystals, diffraction data sets at a subatomic resolution contain hundreds of thousands of reflections, and the number of parameters used to describe the corresponding models may reach the same order. Therefore, the direct way of calculating structure factors becomes very time expensive when applied to large molecules. These problems of high accuracy and computational efficiency require a re-examination of computer tools and algorithms. The calculation of model structure factors through an intermediate generation of an electron density [Sayre (1951). *Acta Cryst.* **4**, 362–367; Ten Eyck (1977). *Acta Cryst.* A**33**, 486–492] may be much more computationally efficient, but contains some parameters (grid step, 'effective' atom radii *etc*.) whose influence on the accuracy of the calculation is not straightforward. At the same time, the choice of parameters within safety margins that largely ensure a sufficient accuracy may result in a significant loss of the CPU time, making it close to the time for the direct-formulae calculations. The impact of the different parameters on the computer efficiency of structure-factor calculation is studied. It is shown that an appropriate choice of these parameters allows the structure factors to be obtained with a high accuracy and in a significantly shorter time than that required when using the direct formulae. Practical algorithms for the optimal choice of the parameters are suggested.

1. Introduction

Nowadays, macromolecular crystals diffracting to the resolution of 1.1–0.9 Å are no longer an exceptional case (Fig. 1). At the same time, progress in crystallization techniques allows more and more macromolecular crystals to be obtained that diffract to a resolution of 0.9–0.8 Å or even higher. In what follows, we refer to a resolution higher than 0.9 Å as a subatomic resolution. This relatively fluid limit marks a new level in structural studies. For example, it becomes possible to analyse the density redistribution due to bond formation and to use more sophisticated structural models of macromolecules (Jelsch *et al.*, 2000; Afonine *et al.*, 2002) to describe this redistribution.

When studying crystals of small molecules at a subatomic resolution, traditional procedures and algorithms, for example those realized in program *SHELX* (Sheldrick & Schneider, 1997), use direct formulae to calculate accurate values of structure factors. These direct calculations are computationally possible because of a small number of parameters and structure factors. However, these numbers grow approximately as a cube with the unit-cell dimensions, and such calculations become an obstacle in a high-resolution refinement of macromolecular models.

Sayre (1951) suggested a fast algorithm that was realized later by Ten Eyck (1973, 1977) essentially using the speed of the fast Fourier transform algorithm (Cooley & Tukey, 1965). This algorithm calculates structure factors through an intermediate generation of the electron density. While an extra computational step of density generation is introduced, the total computational cost for macromolecular crystals is reduced drastically in comparison with the direct formulae to calculate structure factors from an atomic model. Agarwal (1978) has shown that a similar intermediate generation of a

density-like map may be used for fast calculation of the gradient of the traditional least-squares criterion, and Lunin & Urzhumtsev (1985) showed that this procedure could be generalized for an efficient calculation of the gradient of any crystallographic criterion. Currently, practically all programs developed specifically for refinement of macromolecular models use the intermediate density generation and fast schemes of gradient calculation.

The decomposition of all principal calculations in a suite of transitions from one level of model description to another (Lunin & Urzhumtsev, 1985; Urzhumtsev *et al.*, 1989; Urzhumtsev & Lunin, 2001) makes each transformation independent of the others. In particular, as was noted by Urzhumtsev *et al.* (1989), a modular organization of refinement programs does not make it difficult to replace isotropic atoms by anisotropic ones [this was practically realized in *REFMAC* (Murshudov *et al.*, 1997, 1999) and in *CNX* (2002)] or to substitute one reciprocal-space criterion for another. Similarly, the atomic model may be composed of other types of scatterers, for example of multipole atoms (Hansen & Coppens, 1978) and, again, this modification concerns only two program modules, namely the transition from the model parameters to the grid density values and the inverse transition from the derivatives of the criteria with respect to grid density values to the derivatives with respect to the model parameters.

On the other hand, such a fast algorithm introduces computational errors in the values of structure factors because the intermediate density is calculated at a finite grid and because limited atomic radii are used when generating the density. Theoretical studies and numerical tests (Ten Eyck, 1977; Agarwal, 1978; Lunin, 1982; Brünger, 1989) suggested a practical choice for the grid step and for the atomic radii to obtain structure factors with a reasonable accuracy at the resolution traditional for macromolecular crystallography, 1.5–4 Å. The introduction of an artificial additional displacement factor, the same for all atoms of the crystal (Ten Eyck, 1977), allowed more CPU time to be gained.

The structure-factor calculation at both limiting cases, at low and high resolutions, has some features that should be taken into account and we addressed the second of these two cases. When most of our tests had finished, Navaza (2002) published a theoretical analysis of the accuracy of structure factors calculated by the Sayre–Ten Eyck algorithm. While formally speaking his analysis is valid for all resolution ranges, it is implicitly aimed rather at molecular-replacement problems at middle and low resolutions. For example, the program *AMoRe* (Navaza, 1994) uses this method to calculate structure factors for some given search model. Traditionally, this is done at the resolution d of 3–4 Å and the default grid-step value of $h_{\rm gr} \sim d/4$ ensures precise enough values of calculated structure factors. However, when carrying out molecular replacement at a low resolution (for example, see Urzhumtsev & Podjarny, 1995; Ban *et al.*, 1998; Jamrog *et al.*, 2003; Liu *et al.*, 2003), the default value causes very significant errors that required the use, for example, of a 1 Å grid step throughout the whole series of tests in Urzhumtsev & Podjarny (1995). Navaza (2002) has shown that a proper choice of the additional displacement factor can reduce the grid step to the value $h_{\rm gr} \sim d/3$ even for these cases, 'up to any desired maximal resolution'. Navaza (2002) also indicated a way to estimate the 'radius' for Gaussian diffractors that ensures the required accuracy. Unfortunately, the suggested way is not convenient in a practical application. Most importantly, the discussion concerns the choice of sufficient, but not necessary the optimal, parameters. This is not a limiting criterion for molecular replacement when *AMoRe* (Navaza, 1994) executes it only once but becomes crucial for refinement at a subatomic resolution.

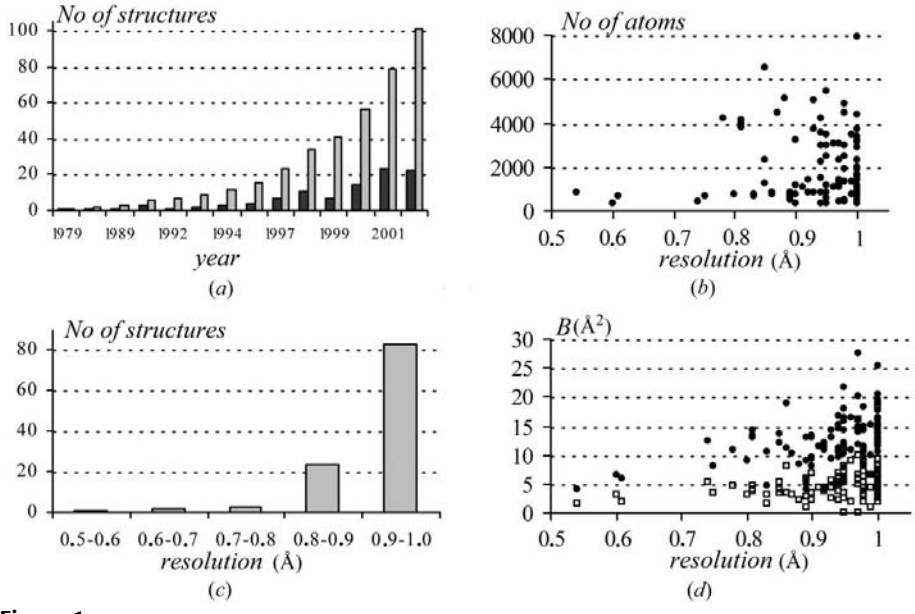

Figure 1
Statistical analysis of macromolecular structures determined at subatomic resolution (models deposited with the PDB to December 2002): (*a*) their number, (*b*) their size, in number of atoms, (*c*) their distribution with the resolution at which they were solved, (*d*) the distribution of minimal (square) and mean (diamond) B values in these models. Dark bars in (*a*) show the number of structures solved per year and grey bars show the cumulative number of structures.

At a subatomic resolution, the CPU time necessary to calculate structure factors and linked values (for example, corresponding gradient) can grow drastically because of the increased number of grid points, important both for density generation and for Fourier transformation. The attempts to refine relatively large protein structures [for example, the refinement of the structure of catalase at 0.89 Å as announced by Murshudov *et al.* (1999)] show that even one cycle of the refinement may require a significant time. The relative compu-

tational cost of algorithms depends on the size of objects and on the resolution at which they are studied.

The aim of this paper is to discuss several questions that, to our knowledge, are still open and not discussed in the literature:

(i) what is the optimal choice of the grid step, of the atomic radii and of the additional displacement parameter, sufficient to calculate structure factors at a subatomic resolution with a given high accuracy?

(ii) with this optimal choice of parameters, is this algorithm still more rapid than the traditional calculation through the direct formulae?

A special study is devoted to the case of models composed of anisotropic atoms.

2. Models and test conditions

2.1. Macromolecular structures resolved at subatomic resolution

A statistical analysis of macromolecular models obtained at a resolution of 1 Å or higher deposited in the PDB (Bernstein *et al.*, 1977; Berman *et al.*, 2000) shows that the number of such structures increases rapidly and that the size of these structures can be quite large (Fig. 1). While currently the best macromolecular crystals, such as crambin (Teeter *et al.*, 1993), antifreeze protein RD1 (Ko *et al.*, 2003) and aldose reductase (Lamour *et al.*, 1999), diffract to 0.54, 0.62 and 0.64 Å, respectively, most of the ultra-high-resolution models were obtained at a resolution somewhere between 0.7 and 0.9 Å (Fig. 1). This made our choice to study the structure-factor calculation at resolutions of 0.5, 0.7 and 0.9 Å.

At such resolutions, atomic displacement parameters are usually taken as anisotropic. An equivalent isotropic displacement parameter can be defined as

$$B_{\rm iso} = \kappa(U_{11} + U_{22} + U_{33})/3, \qquad (1)$$

where the scale coefficient κ depends on the definition of the anisotropic tensor $\mathbf{U}$ (Grosse-Kunstleve & Adams, 2002); when working in fractional coordinates, $\mathbf{U} = \mathbf{U}^*$, following the notation by Grosse-Kunstleve & Adams (2002), and $\kappa = 8\pi^2$. For most of the structures determined at a subatomic resolution, the minimal value of the parameter $B_{\rm iso}$ varies between 1 and 10 Å^2 and the mean value varies between 5 and 20 Å^2 (Fig. 1*d*). The maximal value in some models grows up to 100–500 Å^2 (not shown), indicating either a partial disorder of such crystals or possibly some problems in the model.

2.2. Test models

A multi-Gaussian approximation of atomic scattering factors is extremely convenient for crystallographic analysis because corresponding density distributions in direct space are also multi-Gaussian functions, thus allowing a simple density generation and an easy correction for harmonic atomic displacement. Depending on the problem, different numbers of Gaussians can be used. A two-Gaussian approximation is sufficient for crystallographic studies at 1.5–3 Å resolution (Agarwal, 1978). The available five-Gaussian approximation of scattering factors is valid up to a very high resolution, 0.25 Å (Maslen *et al.*, 1992), and this approximation was used in our tests. It can be noted that single-Gaussian scatterers can be used in some special cases like dummy bond electrons (Afonine *et al.*, 2002, 2004) at subatomic resolution or artificial atoms or blobs at middle and low resolution (Agarwal & Isaacs, 1977; Lunin & Urzhumtsev, 1984; Lunin *et al.*, 1995).

In our tests, a number of models were used. All these models were placed in a unit cell in space group $P1$ in order to exclude any correlation of results, especially the CPU time, with the number and type of crystal symmetries. For simplicity, the unit cell was taken orthogonal, with $\alpha = \beta = \gamma = 90^\circ$. The size of the cell varied with the size of the model.

Initially, several synthetic structures, composed of one or few atoms, were studied. This unit cell with $a = b = c = 10$ Å was large enough to consider such models as isolated.

First, a single Gaussian 'atom' with the scattering factor

$$f(\mathbf{s}) = f(s) = A\exp(-Bs^2/4) \qquad (2)$$

was analysed. Here and in what follows, s stands for the modulus of the vector $\mathbf{s} = (h,k,l)$. For similarity with well ordered C atoms, the parameter A was taken equal to 6, and B varied between 1 and 5 Å^2. Then a C atom was taken with the scattering factor in the form of the five-Gaussian approximation (Maslen *et al.*, 1992):

$$f(\mathbf{s}) = f(s) = \sum_{k=1}^{5} a_k \exp(-b_k s^2/4). \qquad (3)$$

In the following tests, two C atoms placed at a C—C bond distance were studied. Isotropic displacement parameters of these atoms varied also in the limits shown above. Then, one of the C atoms was replaced by O and shifted to the distance typical for the C—O bond. The next model was a peptide group (serine residue), placed in the unit cell defined above. Finally, an artificial structure composed from a tripeptide Val–Ser–Ser and one water molecule was built. The atomic composition of this model corresponds very closely to the mean composition of models of structures discussed in §2.1, *i.e.* 39% of C, 11% of N, 25.5% of O, 24.5% of H, approximately. The model was placed in the unit cell with $a = b = c = 16.5$ Å, space group $P1$ as previously.

The analysis of these simple models allows the determination of all parameters necessary to generate the density distribution.

The obtained results were verified with three real structural models, those for enkephalin (Aubry *et al.*, 1989; Wiest *et al.*, 1994), crambin (Teeter *et al.*, 1993; Jelsch *et al.*, 2000) and aldose reductase (Podjarny *et al.*, 2003), composed of 5, 46 and 320 amino acid residues, respectively (Table 1).

For the simplicity of the study and the presentation of results, most of the tests were done with an isotropic displacement factor. The results for all these models are quite similar to each other and only the most representative ones are shown below.

The final tests were done using models with the available anisotropic displacement parameters and confirmed the

Table 1
Parameters of the macromolecular models used for test calculations.

The lines with * show the parameters of the unit cell of the natural crystals, the line with # gives the parameters of simulated crystals in space group $P1$.

Molecule	Enkephalin	Crambin	Aldose reductase
No. of residues	5	46	320
No. of atoms, total	86	831	6631
% of H, C, N, O, S	50, 32.5, 11.5, 6, 0	50, 32, 10, 7.5, 0.5	40, 26, 27, 6.75, 0.25
Space group*	$P2_12_12_1$	$P2_1$	$P2_1$
Unit cell (a, b, c; β) (Å, °)*	10.9, 13.1, 21.2; 90.0	40.8, 18.5, 22.4; 90.5	49.4, 66.8, 47.4; 92.4
B_{ave}, B_{min}, B_{max} (Å^2)	1.2, 0.5, 3.0	3.9, 1.4, 16.7	10.9, 2.4, 81.5
Resolution of data (Å)	0.46	0.54	0.64
Unit cell $P1$ (test)#	15, 10, 15	30, 30, 30	55, 50, 55

suggested algorithms and the choice of parameters for practical situations.

The quality of the calculated structure factors was estimated using a conventional crystallographic R factor that is close to a relative error of the structure-factor magnitudes. The whole set of structure factors up to a given resolution was divided in shells of reciprocal space with an approximately equal number of reflections per shell; this number varied with tests but in any case it was larger than 50. The R factor was calculated both for the whole set of reflections and for each resolution shell. Usually, the R factor was maximal for the highest-resolution shell and this value and the total R factor were used as the criterion of the quality of the calculated structure factors.

3. Fast calculation of structure factors at subatomic resolution

3.1. Structure-factor calculation at subatomic resolution through direct summation

The formula

$$\mathbf{F}(\mathbf{s}) = \sum_{n=1}^{N_{at}} f_n(\mathbf{s}) \exp(-2\pi^2 \mathbf{s}^T \mathbf{U}_n^* \mathbf{s}) \exp(2i\pi \mathbf{r}_n \mathbf{s}), \qquad (4)$$

where each atom contributes to each structure factor, shows that the computational cost to obtain a set of structure factors by (4) is proportional to $M_{sf}N_{at}$. Here M_{sf} is the number of structure factors and N_{at} is the number of atoms in the model; it is supposed for simplicity that the crystal belongs to space group $P1$. The functions $f_n(\mathbf{s})$ in formula (4) are atomic scattering factors that can be represented by a sum of Gaussian functions (3) or by a combination of spherical harmonic functions (Stewart, 1969; Hansen & Coppens, 1978). For isotropic atoms, a symmetric matrix $\mathbf{U}_n^*$ is replaced by a scalar value B_n as in (1) giving

$$\mathbf{F}(\mathbf{s}) = \sum_{n=1}^{N_{at}} f_n(\mathbf{s}) \exp(-B_n s^2/4) \exp(2i\pi \mathbf{r}_n \mathbf{s}). \qquad (5)$$

For crystals of small molecules, both N_{at} and M_{sf} are of the order of a few hundreds. For macromolecular crystals taken at the same resolution, both N_{at} and M_{sf} are hundreds or thousands times larger, giving a total increase of millions in the number of operations in (4) or (5), which makes this method of computing structure factors very time consuming, especially for model refinement where it is repeated many times.

3.2. Sayre–Ten Eyck approach

An alternative way to calculate structure factors (Sayre, 1951) is based on the fundamental formula

$$\mathbf{F}(\mathbf{s}) = \int_{V_{cell}} \rho(\mathbf{r}) \exp[2i\pi(\mathbf{s}\mathbf{r})] \, dV_{\mathbf{r}}. \qquad (6)$$

and the simplest approximate formula to calculate this integral numerically is

$$\begin{aligned}\mathbf{F}(\mathbf{s}) &= \mathbf{F}(h,k,l) \\ &= \frac{V_{cell}}{N_X N_Y N_Z} \sum_{j_X=0}^{N_X-1} \sum_{j_Y=0}^{N_Y-1} \sum_{j_Z=0}^{N_Z-1} \rho(j_X, j_Y, j_Z) \\ &\quad \times \exp[2i\pi(hj_X + kj_Y + lj_Z)]. \end{aligned} \qquad (7)$$

Here V_{cell} is the volume of the unit cell and N_X, N_Y, N_Z are the numbers of grid points along each of the axes, $K_{grid} = N_X N_Y N_Z$ is the total number of grid points. If we suppose that our task is to calculate the full set of structure factors up to some resolution, then usually K_{grid} is of the same order as the number M_{sf} of these structure factors. The right-hand expression in (7) is the discrete Fourier transform of the function represented by the values $\rho(j_X, j_Y, j_Z)$ of the density distribution calculated at the grid points. Therefore, the whole procedure consists of two steps:

(*a*) starting from atomic parameters, calculation of electron-density values at the points of a regular grid of the unit cell;

(*b*) calculation of the discrete Fourier transform of this grid function.

The computational cost of the direct summation in (7) is proportional to $M_{sf}K_{grid}$ or, which is the same, to K_{grid}^2. This high cost is the main reason why this approach came to crystallographic practice (Ten Eyck, 1973, 1977) only after the fast Fourier transform algorithm had been developed (Cooley & Tukey, 1965). When using the FFT algorithm, the corresponding computational time T_{FFT} is proportional to $K_{grid} \ln K_{grid}$, *i.e.* almost linear with respect to the number of grid points.

The number of computer operations necessary to generate the electron-density values at the grid points (step *a*) is proportional to $K_{grid}N_{at}$ if a straightforward procedure is used, *i.e.* if for each grid point the contribution of each atom is added. However, the computer time may be reduced drastically if it is supposed that each atom has its electron density equal to zero everywhere outside a sphere of radius R with the

centre at the atomic position. In what follows, we call this value R the effective radius of the atom and denote the volume of the corresponding sphere by $V_{\rm atom}$.

Now the calculation of the density may be organized as a cycle through the list of all $N_{\rm at}$ atoms. For each of them, its contribution to the electron density is taken into account only for the grid points that are closer than R to the atomic centre. The number of such points per atom can be estimated through the volume of the atom and that of a grid pixel as $V_{\rm atom}/V_{\rm pixel}$, where $V_{\rm pixel} = V_{\rm cell}/K_{\rm grid}$, giving the total number of operations (and the time necessary to generate grid density values) as

$$T_{\rm dens} \propto (V_{\rm atom}/V_{\rm cell})K_{\rm grid}N_{\rm at} \propto (V_{\rm atom}/V_{\rm cell})M_{\rm sf}N_{\rm at} \qquad (8)$$

with the same factor $M_{\rm sf}N_{\rm at}$ as for the calculation by the direct formula (4) or (5). However, since $V_{\rm atom}$ is a fixed value and the ratio $V_{\rm atom}/V_{\rm cell}$ decreases with the crystal size, the coefficient before $M_{\rm sf}N_{\rm at}$ becomes smaller for large molecules showing for them an advantage of this method of calculation.

As a result, the total number of operations $T_{\rm total}$ required to calculate a set of structure factors from an atomic model through an intermediate generation of the electron density may be estimated as

$$T_{\rm total} = C_1(R^3/V_{\rm cell})K_{\rm grid}N_{\rm at} + C_2K_{\rm grid}\ln K_{\rm grid}, \qquad (9)$$

with the constants C_1 and C_2 independent of other parameters of the algorithm. The value $\ln K_{\rm grid}$ is small in comparison with $K_{\rm grid}$ so that the second term in (9) is practically proportional to the number of structure factors calculated. The first term in (9), while being formally proportional to $M_{\rm sf}N_{\rm at}$, contains now the factor $R^3/V_{\rm cell}$, which is small for macromolecular structures. This factor plays the most important role in the reduction of the computational time when using the Sayre–Ten Eyck approach to calculate structure factors (for example, see the row $T_{\rm form}/T_{\rm total}$ in Table 4, §3.10).

It is worth noting that formulae (8) and (9) may also be presented in a different way (V. Lunin, personal communication) if we introduce a mean crystal volume per atom as

$$V_{\rm crys} = V_{\rm cell}/N_{\rm at}. \qquad (10)$$

Then (9) becomes

$$T_{\rm total} = C_1K_{\rm grid}(V_{\rm atom}/V_{\rm crys}) + C_2K_{\rm grid}\ln K_{\rm grid}, \qquad (11)$$

The values $V_{\rm atom}$ and $V_{\rm crys}$ do not increase with the size of the studied structure, therefore formula (11) indicates that for the given model the computational cost of structure-factor calculation is practically proportional to the number of grid points or, as remarked above, to the number of structure factors.

While the Sayre–Ten Eyck algorithm for structure-factor calculation is currently implemented in most macromolecular refinement programs, it is not regularly used for calculations at a very high resolution where the problem of CPU is even more acute. The main reason for this is the much higher accuracy of structure factors required for this high-resolution analysis. Generally speaking, structure factors calculated through an intermediate generation of electron density are expected to be quite accurate even at a very high resolution when a very fine grid and large atomic radii are used. However, such a choice of parameters may significantly raise the number of operations to calculate the Fourier transform and to calculate the density increasing the number of grid points where atomic contribution should be taken into account. As a consequence, a gain in CPU time is not evident *a priori* when using this algorithm under such conditions.

3.3. Atomic radius for a continuous density distribution

As discussed above, when generating grid density values it is supposed that the electron density corresponding to an individual atom is equal to zero outside the sphere of some radius, called below an 'effective' atomic radius or simply atomic radius (it should not be confused with the atomic radius used for various physical and chemical considerations; the effective radius is a purely computational parameter). For the given grid step, the number of operations to generate electron density grows as the cube of atomic radius [see (9)], making it one of the principal parameters that define the speed of the algorithms. At the same time, an insufficient radius may cause high errors in the structure factors calculated from this density [for a recent and detailed result, see Navaza (2002)]. A variety of tests to study the role of this parameter was done, and the first example is a density analysis for the synthetic model composed from the C and O atoms at the distance of a double bond. The C atom was placed at the origin of the unit cell and the bond was aligned with the Ox axis of the unit cell described above. The density variation along this axis represents the behaviour of the density both at the bond (between atoms) and near the terminal atoms (outside the bond). The displacement parameter was taken isotropic and equal for both atoms.

The electron density was calculated as a sum of contributions from these two atoms, each contribution being in the form

$$\rho(\mathbf{r}, \mathbf{r}_0, B) = \sum_{k=1}^{5} a_k \left(\frac{4\pi}{b_k + B}\right)^{3/2} \exp\left(-\frac{4\pi^2|\mathbf{r} - \mathbf{r}_0|^2}{b_k + B}\right) \qquad (12)$$

with the spherically symmetric scattering factor $f(s)$ in the five-Gaussian form (3). Here r_0 is the position of the corresponding atom and B is its isotropic displacement parameter. The constants a_k and b_k, $k = 1, \ldots, 5$, are the same as in (3) and are different for O and C atoms. Fig. 2 shows the decrease of the atomic density in the centre of an atom when B varies from 1 to 5 Å^2; the most significant change happens when B increases from 1 to 2 Å^2. Since the total amount of the atomic electron density is conserved, this automatically means the growth of the density beyond some distance from the atomic centre. (It can be noted that, while the value of the density on the atomic centre decreases drastically, this happens inside a sphere of very small radius and the corresponding recompensating correction at large distances is relatively small.)

The amount of the electron density outside a sphere of a given radius is one of the key values defining the accuracy of the calculations when the electron-density generation is used

as an intermediate tool to obtain structure factors. This quantity and its influence on the accuracy of structure factors were studied previously by Agarwal (1978) and Lunin (1982) for the case of the two-Gaussian approximation to the atomic scattering factor; see also Bricogne (1993).

For an isotropic atom with a displacement parameter B and with a five-Gaussian approximation [formulae (3) and (12)] for its scattering factor, if the density is cut at a distance R, the relative accuracy $\varepsilon(\mathbf{s},B,R)$ of the structure factor $\mathbf{F}_R(\mathbf{s},B)$ can be estimated, similarly to Lunin (1982), as

$$\varepsilon(\mathbf{s},B,R)=\frac{|\mathbf{F}(\mathbf{s})-\mathbf{F}_R(\mathbf{s})|}{|\mathbf{F}(\mathbf{s})|}=\frac{\left(\frac{2}{s}\right)\left|\int_R^\infty r\sum_{k=1}^5 a_k\left(\frac{4\pi}{b_k+B}\right)^{3/2}\exp\left(-\frac{4\pi^2\mathbf{r}^2}{b_k+B}\right)\sin(2\pi\mathbf{sr})\,\mathrm{d}r\right|}{\sum_{k=1}^5 a_k\exp[-(b_k+B)s^2/4]}. \quad (13)$$

The integral in (13) was calculated numerically for various types of atoms and for different isotropic displacement factors. Fig. 3 shows $\varepsilon(\mathbf{s},B,R)$ in the highest-resolution zone (for the resolutions 0.5 and 0.7 Å, respectively) for the atoms C, O, N or S as a function of an atomic radius R. Indeed, for the same radius the relative accuracy differs by an order of magnitude (Figs. 3*a*, *b*, *c*) if B changes from 1 to 5 Å^2. In other words, these figures show by how much an increase of the displacement parameters increases the cut-off radius necessary to calculate structure factors with the same accuracy. One more observation is that for the given radius the accuracy improves tenfold when the resolution is decreased from 0.5 to 0.7 Å (Figs. 3*c* and *d*).

Fig. 3 shows that, after some limit value, increasing the radius does not improve significantly the accuracy of structure factors. For the following tests, the relative accuracy of 0.5% of structure factors in all resolution zones, including the zone of the highest resolution, was chosen as a target value. This value is lower than the usual R factor, which for studies of small molecules at subatomic resolution in best cases reaches 1–3% but for macromolecules is higher.

For the radii shown, the relative error for a C atom is notably higher than the error for O or N atoms. This is partially explained by the fact that the magnitude of the structure factors, the denominator in (13), decreases faster for C than for O or N atoms. Since a real structure contains a mixture of atoms of different types, the use of (13) for practical calculations should be revised. In more accurate estimations of the atomic radius, the denominator in the formula for the relative error should be taken equal to the magnitude of the structure factor calculated from the whole model rather than that of the structure factor for a chosen type of atom.

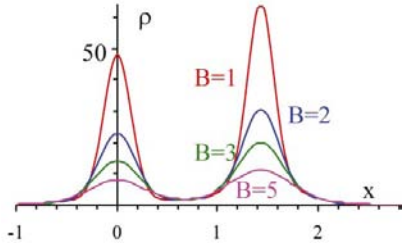

Figure 2
Electron-density distribution for an artificial diatomic model; the atomic displacement parameter varies from 1 to 5 Å^2.

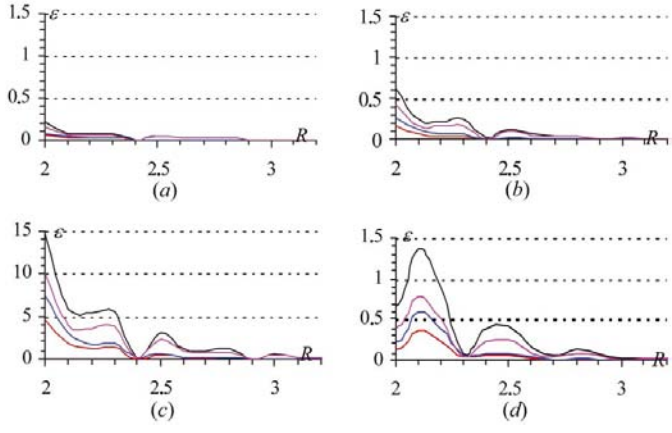

Figure 3
Relative error, in %, in the magnitude of an atomic structure factor as a function of atomic radius (see text for details). The curves are obtained for the resolution d and atomic displacement parameter B taken as: (*a*) d = 0.5 Å, B = 1 Å^2; (*b*) d = 0.5 Å, B = 2 Å^2; (*c*) d = 0.5 Å, B = 5 Å^2; (*d*) d = 0.7 Å, B = 5 Å^2. The curves are given for C (black), N (blue), O (red) and S (magenta).

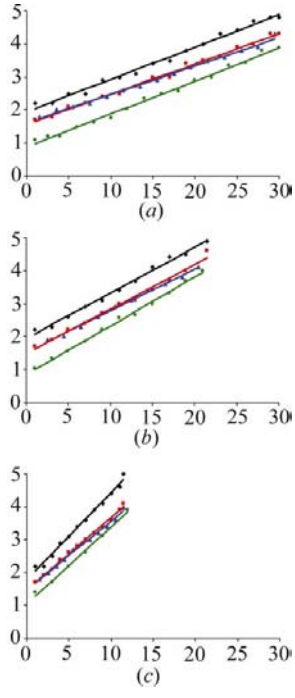

Figure 4
Minimal atomic radius R as a function of displacement parameter B (Å^2), necessary for the calculation of structure factors with a relative accuracy of 0.5% at a resolution of (*a*) 0.9, (*b*) 0.7 and (*c*) 0.5 Å. The radius is shown for several principal types of macromolecular atoms and is indicated as: H (green), C (black), N (blue), O (red).

Table 2
Coefficients α and β of the linear regression $R = \alpha B + \beta$ for atomic radii R as a function of displacement parameter B.

The radii are obtained in order to ensure a relative accuracy of 0.5% of calculated structure factors.

	H		C		N		O	
Resolution	α	β	α	β	α	β	α	β
0.9 Å	0.105	0.792	0.097	1.948	0.088	1.574	0.092	1.553
0.7 Å	0.149	0.807	0.140	1.926	0.122	1.563	0.129	1.537
0.5 Å	0.231	1.072	0.259	1.789	0.204	1.506	0.219	1.466

3.4. Atomic radius for different types of atoms

In order to get an estimation for the minimal atomic radius as a function of the resolution and of the value of the displacement parameter, the test with the tripeptide model described above has been done. The same value of displacement parameter B, varying from 1 to 30 Å^2, was assigned to all the atoms of this model. For each such model, the structure factors $\mathbf{F}(\mathbf{s},B) = F(\mathbf{s},B)\exp[i\varphi(\mathbf{s},B)]$ were calculated by formula (5).

The results of the previous section show that the atomic radius R of the order of 3.0–3.5 Å gives a very small computational error in structure factors. In the current test, first of all, the radii were kept at the value of 5 Å for all the atoms except C atoms for which it was varied, this being the only source of error in the calculated structure factors. With these parameters, the electron density was generated with a very fine grid, structure factors were calculated as Fourier coefficients of this function and their magnitudes $F_{\rm mod}(\mathbf{s},B;R)$ were compared with $F(\mathbf{s},B)$. The atomic radius $R_{\rm C}(B)$ for carbon was defined as the minimal value for which the R factor between $F_{\rm mod}(\mathbf{s},B;R)$ and $F(\mathbf{s},B)$ was below 0.25%.

Similar calculations were done for other types of atoms, N, O and H. After the limit radius was determined for each of these types, the values obtained were used together to check the total accuracy of structure factors when all atomic radii are limited. Considering structure factors from each type of atom to be independent, we expected to get an R factor of order $0.5\% \approx 0.25\% \times 4^{1/2}$, where 4 is the number of atomic types. In fact, the obtained R factor was significantly lower than 0.5% in all resolution zones. This margin allowed a reconsideration of the radius for hydrogen (its radius initially was smaller, but not significantly, than the radii for N and O atoms; since the H atoms may represent up to half of the model, see Table 1, a smaller radius for them can reduce the CPU time). For each B value, the radius for C, N and O atoms was taken as found in the previous tests and the atomic radius for hydrogen decreased up to the value when the relative accuracy of structure factors in the highest-resolution zone reached the 0.5% limit. The results of the search are summarized in Fig. 4.

As can be expected, the minimal radius grows with the B value. Under the same conditions, the atomic radius for the C atom is still larger than the radii for N and O atoms while, after the errors in the structure factor were normalized by the total structure factor, it became relatively close to them.

For a given type of atom, the radius R is practically the same at different resolutions when B is small and increases practically linearly with B (Fig. 4). The growth is higher for higher resolutions. Corresponding coefficients of the linear regression are shown in Table 2. A minor discrepancy between experimental data and corresponding linear functions is observed only for the smallest B values, those of 1–3 Å^2, where the radii obtained from the linear equations should be increased by 0.05–0.10 Å.

As has been mentioned, both a limited atomic radius and a finite step of the grid are responsible for the errors in the structure factors calculated through the density generation. We wanted to check first whether these parameters could be studied separately in some interval of their values. If this is the case, the same values of the effective radius obtained above can be used for any small enough grids, simplifying the analysis of the computational efficiency.

On the one hand, a density generation at a very fine grid allows the estimation of a minimal atomic radius (for a given accuracy of structure factors) for a 'practically continuous function'. On the other hand, it is clear that, when the grid step becomes larger than some limit, the computational error in structure factors will be too large for any radius. We hoped that for the intermediate values of the grid step a variation (owing to possible 'border effects') of the effective atomic radius, sufficient to reproduce structure factors with the chosen accuracy, should not be significant. In order to check this, the tests described above on a very fine grid were repeated at different grids and only minor fluctuations were observed in the limit radius values for all studied types of atoms (several examples are illustrated in Fig. 5).

Therefore, the tests described above allowed the determination of the minimal atomic radius (for the relative accuracy of structure factors of 0.5%) for each type of atom as a function of the resolution d and the B value.

3.5. Atomic displacement parameter and maximal grid step

When calculating structure factors through an intermediate generation of density, a grid step $h_{\rm gr}$ is traditionally taken as

$$h_{\rm gr} \sim d/3 - d/4, \qquad (14)$$

where d is the resolution of the data set (see Ten Eyck, 1977; Agarwal, 1978; Brünger, 1989; Navaza, 2002, and references therein). Use of a finer grid increases the CPU time, both for density generation and for the Fourier transform. A coarser grid cannot be used because of the computational errors introduced.

The Fourier coefficients $\mathbf{F}(h,k,l)$ of a continuous function differ from the Fourier coefficients $\mathbf{F}_g(h,k,l)$ for the corresponding function calculated at a grid. After a proper scaling,

$$\mathbf{F}_g(h,k,l) = \mathbf{F}(h,k,l) + \sum_{j_x,j_y,j_z=-\infty}^{\infty} \mathbf{F}(h+j_xN_x, k+j_yN_y, l+j_zN_z), \quad (15)$$

where N_x, N_y, N_z are grid numbers and the sum is taken over all possible integers j_x, j_y and j_z, different from 0 simultaneously. Since structure-factor magnitudes decrease in line with h,k,l, such a correction can be neglected for a fine enough grid but is significant for small grid numbers (as an example of research when such a small grid is justified computationally, see Lunin *et al.*, 2002). The mean value of the Fourier coefficients decreases with increasing indices, and this drop is faster for smoother functions. This means that the sum in (15), in other words the difference between $\mathbf{F}_g(h,k,l)$ and $\mathbf{F}(h,k,l)$, is large for sharp functions. (In crystallographic terms, this can be presented as the fact that the atoms with a small B may be 'too narrow' for a given grid, 'fall down between grid nodes' and the corresponding grid function does not reproduce correctly enough the density distribution and, as a consequence, its structure factors. This also gives an idea of how to smooth this function, as is discussed below in §3.6.) The search for the maximal grid step h_{gr}, which provides one with the desired structure-factor accuracy, as a function of B was done for a series of tripeptide models with the same B parameter assigned to all atoms as described previously. The results for several resolutions are shown in Fig. 6. The atomic radii were chosen from the previous analysis and the grid step was expressed through the resolution d as

$$h_{gr} = d/n, \quad (16)$$

where n is a real number. For medium B values, those between about 4 and 15 Å^2, the traditional estimation of $n = 3$ is quite satisfactory. The value of n can be decreased for atoms with larger B ($n = 2.5$) but should be larger than 3 for atoms with B smaller than 4 Å^2. Interestingly, this number decreases with the resolution.

The results of this test allowed an estimation $h_{gr} = h_{gr}(d)$ in the form of d/n for the maximal grid step which was used in the tests below. It is worth noting that the value of n decreases with the resolution.

3.6. Additional displacement factor

Owing to the observation that atoms with a larger B do not need so fine a grid as atoms with a small B, Ten Eyck (1977) introduced an additional displacement parameter B_{add} to gain more CPU time for the structure-factor calculation. At the step of density generation, the atomic displacement parameter B is increased by this artificial factor B_{add}, the same for all atoms. As a consequence, a larger density grid can be used. After the corresponding Fourier coefficients are calculated, this modification of B can be taken into account multiplying each coefficient by $\exp(B_{add}s^2/4)$, s being the modulus of the vector $\mathbf{s} = (h,k,l)$. Recently, Navaza (2002) showed that such an approach is efficient not only at the traditional resolution between 1 and 4 Å but allows the use of the grid step $d/3$ even at much lower resolutions (see for example his test case at 15 Å resolution).

On the other hand, increasing B makes the atom 'larger' (see §§3.3, 3.4) and increases the number of grid points to which its contribution should be calculated and the CPU time requested. Therefore, for the given resolution and atomic displacement parameters available, some optimal combination of the grid step and B_{add} should be found.

It can be noted that, if all atomic displacement parameters are large, a negative value for the parameter B_{add} may be used,

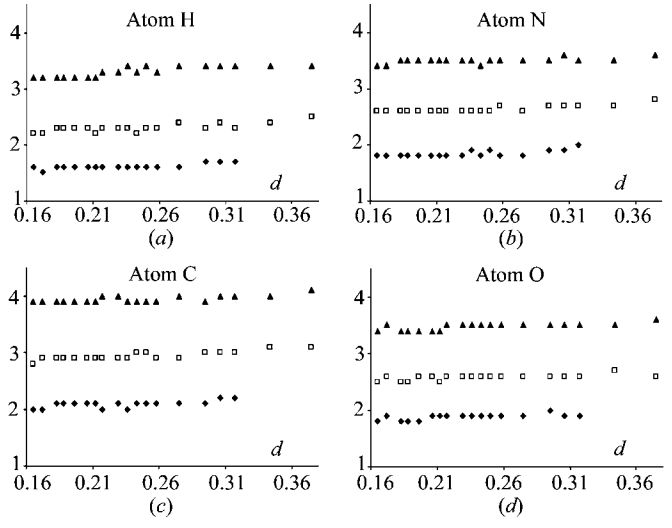

Figure 5
Minimal atomic radius R for H, C, N, O, necessary to calculate structure factors with an accuracy of 0.25% at a resolution of 0.9 Å as a function of the grid step d (in Å). The results are shown for the model with the displacement parameter B equal to 7 Å^2 (diamond), 15 Å^2 (square) or 25 Å^2 (triangle). When B = 7 Å^2, a grid step larger than 0.32 Å does not allow the accurate calculation of structure factors for any atomic radii. See text for details of the choice of the atomic radii.

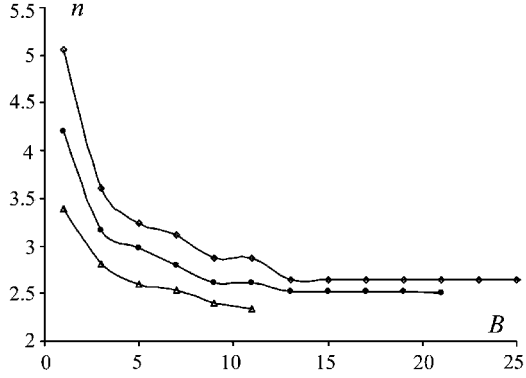

Figure 6
Maximal grid step to generate density sufficient to reproduce structure factors with R factor lower than 0.5% as a function of atomic displacement parameter (Å^2). The highest resolution d of structure factors is 0.9 Å (diamond), 0.7 Å (circle) or 0.5 Å (triangle). The grid step is expressed through the parameter n: $h = d/n$.

decreasing the atomic radii required for the accurate calculation of structure factors.

3.7. Choice of optimal parameters

The principal parameters that define the accuracy of structure factors calculated through a density generation are the grid step h_{gr}, atomic radii $R_{at,n}$ and atomic displacement parameters B_n, which can be corrected using the additional displacement factor B_{add}. These parameters define the main computational characteristic of the algorithm, the number N_{points} of grid points to which each atom contributes when generating the electron density. The CPU time of the first step of the Sayre–Ten Eyck algorithm is practically proportional to this number.

Let us suppose for the analysis below that all atoms of the model have the same displacement parameter value B and that the model composition corresponds to the mean composition discussed above in §2.2. For each value of the additional displacement parameter B_{add}, the effective displacement parameter can be calculated as $B_{eff} = B_{add} + B$. For the obtained B_{eff}, the optimal value of the grid step $h_{gr}(B_{eff}; d, \varepsilon)$ can be determined as well as the atomic radius $R_T(B_{eff}; d, \varepsilon)$ for all principal types of atoms, T = H, C, N, O. The number of operations to generate an electron density at a chosen grid is estimated as

$$N_{points}(B_{eff}; d, \varepsilon) \propto N_{mod} \sum_T \mu_T \left[\frac{R_T(B_{eff}; d, \varepsilon)}{h_{gr}(B_{eff}; d, \varepsilon)}\right]^3, \qquad (17)$$

where N_{mod} is the number of atoms in the model and μ_T is the relative share of atoms of the type T in proteins (see the discussion in §2.2). The total time to calculate structure factors is the sum of time T_D to generate the electron-density distribution, which is proportional to $N_{points}(B_{eff}; d, \varepsilon)$, and of time T_{FFT} to calculate the FFT, decreasing with $h_{gr}(B_{eff}; d, \varepsilon)$. Our previous experience (Urzhumtsev *et al.*, 1989), confirmed by the current calculations (Table 4, §3.10), shows that these two values have roughly the same magnitude but the latter is usually a few times smaller. Therefore, the reduction of the total computational time requires mainly the minimization of the number $N_{points}(B_{eff}; d, \varepsilon)$.

Fig. 7 shows $N_{points}(B_{eff}; d, \varepsilon)$ as a function of B_{eff} for several different resolutions d and $\varepsilon = 0.5\%$ chosen above. The minimum of this function indicates the optimal values of B_{eff} which consecutively defines the additional displacement parameter B_{add}, the optimal grid step $h_{gr}(B_{eff}; d, \varepsilon)$ and atomic radii for different types of atoms.

It is important to note that in this test the optimal value of B_{eff} does not depend on the B value for individual atoms. In particular, this means that, if $B > B_{eff}$, a negative additional displacement factor B_{add} should be introduced. In all cases, the grid step and atomic radii defined as discussed above are adequate to obtain structure factors with the required accuracy.

Fig. 7, taken together with Fig. 6, confirms that for such a subatomic resolution the optimal step can be chosen as $h_{gr} \simeq d/3$ similar to the case of lower resolutions (Navaza, 2002). A further increase of the grid step could accelerate the second step of the Sayre–Ten Eyck procedure, the Fourier transform. However, this would increase B_{eff} and, as Fig. 7 shows, the number of points used for density generation would slow down the first step. Decreasing the grid step is not necessary if the required accuracy of structure factors is already attained because that would only increase the number of computer operations at both steps.

The choice of step $h_{gr} \simeq d/3$ gives an estimate for the size of the array of the electron density. For crambin, the density array for the whole unit cell at a resolution of 0.6 Å will occupy about 14 Mbyte, for aldose reductase at 0.9 Å it requires 22 Mbyte, and for the same protein at 0.7 Å it requires 48 Mbyte, currently available with modern computers, and thus allows the calculation of the Fourier transformation directly in memory for most cases.

3.8. Limit radius value

The previous sections suggest a way to choose the atomic radii when all atoms of the model have the same displacement parameter, which is not true in a practical situation (see §2.1; see also Parthasarathy & Murthy, 1997, for the analysis of the distribution of atoms with B value).

When atoms of the model have different B values, those with a larger B value give a weaker contribution to higher-resolution structure factors and therefore require less absolute accuracy. As a consequence, for such atoms there is no need to increase their radii $R(B)$ linearly with B (Fig. 4) but some limit value R_{lim} can be used instead. Probably, such a value can be estimated theoretically knowing the distribution of atoms with the B values (Parthasarathy & Murthy, 1997), however we restricted ourselves by the following numerical analysis.

A model of aldose reductase was taken as a representative case with a large variation of B. In order to ensure the necessary accuracy of 0.5% when generating the density at the grid with the step $h = d/3$, an additional displacement parameter B_{add} was used as discussed in the previous section. Each atomic radius was taken following the linear regression

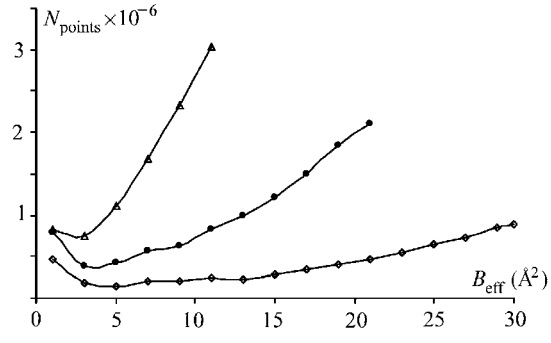

Figure 7
The number of grid points N_{points} to which the atoms of the model contribute when the electron density is generated at a grid and with atomic radii sufficient to calculate structure factors with a relative accuracy below 0.5%. Curves for resolutions of 0.9 Å (diamond), 0.7 Å (circle) and 0.5 Å (triangle) are shown for the tripeptide model with the standard share of the basic types of atoms (N_{mod} = 28; see §2.2).

Table 3
The limiting atomic radius allowing the accurate calculation of structure factors at resolutions of 0.5, 0.7, 0.9 Å.

The radius is shown for tests with aldose reductase, crambin and the crambin model with the distribution of the B parameter similar to that in the aldose reductase model. See text for more details.

	Aldose					Crambin					Crambin → aldose				
d (Å)	R_{ass} (%) [B_{ass} (Å^2)]	H	C	N	O	R_{ass} (%) [B_{ass} (Å^2)]	H	C	N	O	R_{ass} (%) [B_{ass} (Å^2)]	H	C	N	O
0.9	0.63 [17]	2.6	3.6	3.1	3.1	0.65 [8]	1.6	2.7	2.3	2.3	0.65 [25]	3.4	4.4	3.8	3.9
0.7	0.42 [14]	2.9	3.9	3.3	3.3	0.40 [6]	1.7	2.8	2.3	2.3	0.44 [18]	3.5	4.4	3.8	3.9
0.5	0.22 [09]	3.2	4.1	3.3	3.4	0.25 [5]	2.2	3.1	2.5	2.6	0.25 [13]	4.1	5.1	4.2	4.3

$R_T(B)$ (Fig. 4) for the corresponding atomic type T and was equal to $R_{T,\mathrm{lim}} = R_T(B_{\mathrm{lim}})$ for all atoms with the value $B > B_{\mathrm{lim}}$. The parameter B_{lim} was varied and the R factor with the exact structure-factor magnitudes was calculated in resolution zones. The maximal value (corresponding as a rule to the highest-resolution zone) was defined. These calculations were repeated at the resolutions of 0.5, 0.7 and 0.9 Å, and several representative examples are shown in Fig. 8(*a*).

At a resolution of 0.5 Å, the R factor in the highest-resolution zone is quite sensitive to the variation of B_{lim} and the corresponding radius near the chosen limit of 0.5%. On the other hand, it can be noted that, confirming the analysis in §3.3, the variation of the radius above some value does not increase further the accuracy of structure factors. This asymptotic residual error, owing to the chosen grid step, is below 0.5% (Fig. 8*a*), and this gives an idea for another estimation for the limiting radius of the given type of atom (Table 3). The value of B_{ass} when the residual error becomes constant can define such a radius as $R_{\mathrm{ass}} = R_T(B_{\mathrm{ass}})$

A similar test was done with the model of crambin where the atomic displacement parameters are much smaller. In the next test, they were generated randomly in the same range as for the aldose reductase model. The study of these models and several with an intermediate situation shows that the asymptotic R factor is practically independent of the protein and of the distribution of B values (Fig. 8*a*), but the value of B_{ass} when the asymptotic accuracy was achieved clearly correlates with the mean displacement parameter B_{mean} (Fig. 8*b*). The interpolation curves (Fig. 8*b*) can serve to estimate the B_{ass} value and corresponding radii R_{ass} for other models.

When the resolution decreases, the behaviour of the R factor does not change. The major difference is that the asymptotic R factor approaches the chosen limit of 0.5% (at 0.7 Å, not shown) or even slightly overcomes it (Fig. 8*a*). Nevertheless, its value, maximal for all resolution zones, is still reasonable, and the mean R-factor value is still significantly lower than 0.5%. At these resolutions, the parameters B_{ass} and R_{ass} behave also in the same way as they do at 0.5 Å and for a given atomic model their values can be estimated from Fig. 8 as shown in Table 3.

3.9. Choice of parameters in a practical situation

To summarize the previous analysis: the following algorithm to choose the parameters for density generation can be suggested [as expected, the first steps are similar to those discussed by Navaza (2002)].

The minimal value of the atomic displacement parameter $B_{\mathrm{min}} = \min\{B_n\}$, $n = 1, N_{\mathrm{at}}$, is calculated.

For the chosen resolution d, the optimal value of $B_{\mathrm{eff}}(d)$ and the grid step $h_{\mathrm{gr}}(B_{\mathrm{eff}}; d)$ are defined; in general, an estimation $d/3$ for the grid step is appropriate.

B_{add}, positive or negative, is taken such that it shifts the minimal atomic displacement parameter to $B_{\mathrm{eff}}(d)$:

$$B_{\mathrm{min}} + B_{\mathrm{add}} = B_{\mathrm{eff}}(d). \qquad (18)$$

The value of B_{add} is added to each atomic displacement parameter B_n, $n = 1, N_{\mathrm{at}}$, and for each of them the corresponding minimal value of the radius $R_T(B_n + B_{\mathrm{add}}; d)$ is defined depending on the type of atom as a linear function of B corrected for small values (1–3 Å^2) and limited by the corresponding R_{ass} for large B values.

The electron density is generated with the chosen parameters; Fourier coefficients are calculated; structure factors are recovered from these coefficients, multiplying them by $\exp(B_{\mathrm{add}}s^2/4)$.

Generally speaking, the optimization of the CPU time requires that not the atom with minimal B but 'most of the

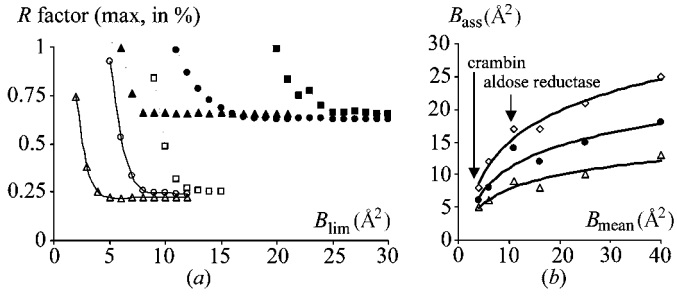

Figure 8
Study of the limiting atomic radius. (*a*) The relative accuracy, in %, of structure factors in the highest-resolution zone as a function of the limiting atomic radius calculated as $R_{\mathrm{lim}} = R(B_{\mathrm{lim}})$. The curves are shown for aldose reductase (circles), crambin (triangles) and the crambin model with the distribution of the B parameter similar to that in the aldose reductase model (squares). Open symbols correspond to a resolution of 0.5 Å, and the black symbols correspond to that of 0.9 Å. The grid step used is $d/3$. See text for more details. (*b*) The value of B_{ass} when the R factor reaches its asymptotic value as a function of the mean B value of the displacement atomic parameters. The calculations are done for the crambin model (B_{mean} = 4 Å^2; B_{max} = 17 Å^2), for aldose reductase (B_{mean} = 11 Å^2; B_{max} = 81 Å^2), and for several crambin models with modified B values so that B_{mean} varied from 6 to 41 Å^2. Triangles, circles and rhombi show the results at 0.5, 0.7 and 0.9 Å resolution.

atoms' of the model have a B value equal to $B_{\rm eff}$ thus prescribing a value for $B_{\rm add}$ different from that obtained by (18). However, since $B_{\rm eff}$ is quite small, this could make displacement factors for some atoms equal to zero or to some negative value that may cause large computational errors.

3.10. Verification of the procedure with isotropic structural models

The suggested scheme was applied to three crystals, the atomic structure of which has been previously solved at subatomic resolution (see §2.2). For all these models, at this first test the anisotropic displacement parameter was replaced by the equivalent value of the isotropic parameter (1), and a set of structure factors $F_{\rm form}(\mathbf{s})$ at a resolution of 0.5 Å was calculated by the direct formula (5).

For each resolution d, equal to 0.5, 0.7 or 0.9 Å, respectively, the parameter $B_{\rm add}$ and atomic radii were chosen for each of the three models as discussed above (§3.9). The electron density was generated at a corresponding grid with the step $h(B_{\rm eff}; d) = d/3$ and structure factors $F_{\rm dens}(\mathbf{s})$ were obtained as its Fourier coefficients multiplied by $\exp\{B_{\rm add}s^2/4\}$.

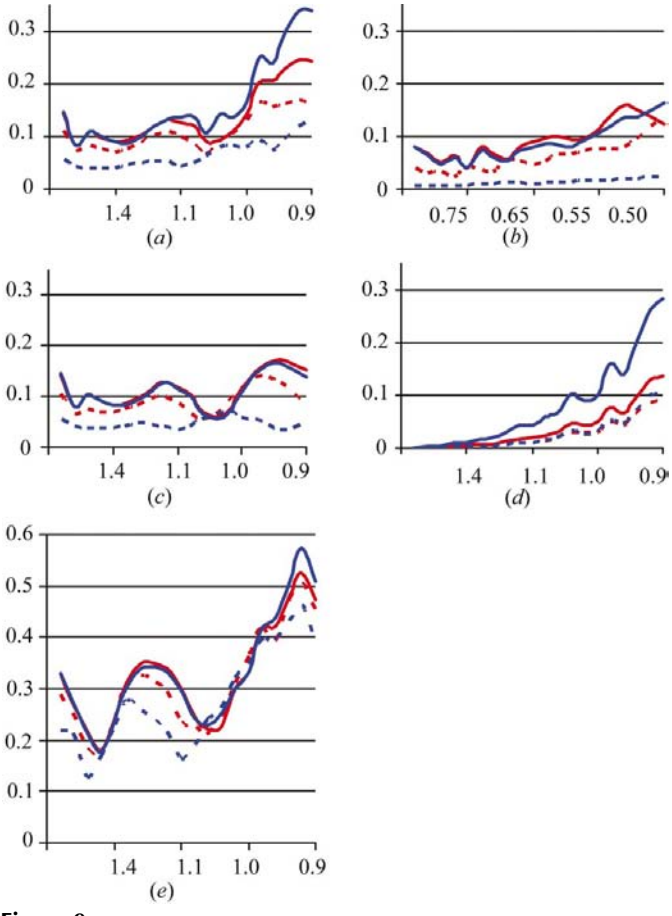

Figure 9
R factor, in %, calculated as a function of resolution d (Å), for structure factors calculated through density generation for the enkephalin model with anisotropic displacement parameter. The parameters for density generation were estimated through an equivalent isotropic value (full curves) or through minimal/maximal eigenvalues (broken curves). The results are shown for the model with its natural displacement parameters (red curves) and for the model with an artificially increased anisotropy (blue curves). (*a*) 0.9 Å, standard parameters; (*b*) 0.5 Å, standard parameters; (*c*) 0.9 Å, small grid step of $d/12$; (*d*) 0.9 Å, large radii; (*e*) 0.9 Å, isotropic atoms included.

The R factor between structure-factor magnitudes for the two corresponding sets, $\{F_{\rm form}(\mathbf{s})\}$ and $\{F_{\rm dens}(\mathbf{s})\}$, is shown in Table 4. These calculations justify the choice of the parameters for the requested accuracy of 0.5%. The same table allows the comparison of the CPU time necessary for the direct computation of structure factors by formula (5) and the Sayre–Ten Eyck algorithms with the parameters defined as above (§3.9). It shows that the latter reduces the CPU time by 1–2 orders, depending on the size of the crystal. A CPU time is given for a Pentium III/733 MHz processor.

Table 4 gives one more observation worth noting. While the gain in computational efficiency is practically independent of the resolution, it increases with the size of the molecule being in agreement with theoretical considerations of §3.2.

3.11. Choice of optimal parameters for atoms with anisotropic scattering factors

The next test was similar to the previous one but atomic scattering factors were taken to be anisotropic as they are traditionally used at such a high resolution. Two different calculations were done. In order to see more clearly the effect of anisotropy, all isotropic atoms (H atoms) were removed from the model.

At the first test, the parameters for density generation were taken from the previous example being estimated from the equivalent isotropic displacement parameters. In the second test, two different B values were calculated for each atom, $B_n^{\rm min} = 8\pi^2 U_{\rm min}$ corresponding to the minimal and $B_n^{\rm max} = 8\pi^2 U_{\rm max}$ corresponding to the maximal eigenvalues of the matrix $\mathbf{U}_n^*$. These two values define the smallest and the largest size of the atom, respectively. The parameter $B_n^{\rm min}$ was used to define the value of $B_{\rm add}$ and $B_n^{\rm max}$ was used to define the corresponding atomic radius. Fig. 9 shows the results of these tests with the enkephalin data at 0.9 and 0.5 Å resolutions.

As can be expected, the second calculation, at a finer grid with larger radii, gives structure factors with a higher accuracy (Figs. 9*a* and *b*). However, both the gain in accuracy and the loss in CPU time, in comparison with the first calculation, are small. What is more important is that the accuracy obtained in the first calculation is sufficient and the usual anisotropy in displacement factors does not require a more complicated estimation of parameters (see, in particular, Navaza, 2002). In order to study more difficult cases, the anisotropy of all atoms was artificially increased as described in Appendix *A*. The tests repeated with such modified conditions give a higher difference in accuracy of structure factor obtained in these two ways. The second scheme gives slightly better results and the first scheme gives slightly worse results in comparison with the previous test. Nevertheless, the accuracy of structure factors obtained using parameters from equivalent isotropic B values is still sufficient for practical crystallographic studies. It is interesting to note that the calculations at 0.5 Å, where the grid step is small, give smaller errors and closer curves.

To check the role of the parameters, two additional test calculations were done at 0.9 Å. In the first one, the radius was left as it was in the initial test but a much smaller grid step $d/12$ was used to generate the density. The errors went down essentially for higher-resolution reflections and, in general, they behave similarly to the test at 0.5 Å (Fig. 9*c*). In the second test, the grid step was kept as $d/3$ but all atomic radii were increased to a very large value (5 Å). In comparison with the initial test, the R factor in the highest-resolution zone decreased insignificantly but decreased practically to 0 for the reflections of the lowest resolution (Fig. 9*d*) showing the role of the density cut-off at a large distance.

When the isotropic atoms (H atoms) were included back into the model, this increased systematically the error and the difference between the results obtained with the two schemes becomes even smaller (Fig. 9*e*). In summary, these tests essentially suggest that the estimation of the parameters for density generation from equivalent isotropic B values is sufficient for practical work.

4. Atomic model refinement at subatomic resolution

Calculation of structure factors from an atomic model is one of the most time-consuming steps of the refinement procedure. Another time-consuming step is the calculation of the gradient of the crystallographic criterion with respect to atomic parameters when gradient minimization methods are used.

As an application of the general theorem (Baur & Strassen, 1983; Kim *et al.*, 1984), it has been demonstrated (Lunin & Urzhumtsev, 1985) that for any crystallographic criterion the calculation of the gradient through intermediate steps of its calculation, first with respect to structure factors and then with respect to density values, needs the same number of operations as the calculation of a single set of structure factors, $T_{\rm grad} \simeq T_{\rm total}$, and not $T_{\rm grad} \simeq N_{\rm at} T_{\rm total}$ as could be expected. All parameters at these intermediate steps, in particular $h_{\rm gr}$, $R_{\rm at}$, $B_{\rm add}$, should be conserved as they are used for the calculation of structure factors (Lunin & Urzhumtsev, 1985).

It can be noted that, even when some variants of the conjugate-gradient method formally use the normal matrix (matrix of the second derivatives), in practice this use is reduced to the calculation of the product of the normal matrix by a given vector that can be calculated by the same time $T_{\rm prod} \simeq T_{\rm total}$ (Lunin & Urzhumtsev, 1985; Urzhumtsev *et al.*, 1989). Moreover, even if the second-order minimization methods are used for refinement and the exact normal is required, for most crystallographic criteria it can be calculated by the time proportional to $T_{\rm total}$ (Urzhumtsev & Lunin, 2001).

Therefore, for optimal refinement algorithms, as soon as a fast scheme of structure-factor calculation is suggested, the optimal schemes for calculation of all relevant quantities (gradient, derivative along a given direction, normal matrix, its product by a given vector) can be generated automatically. The CPU time to calculate these quantities should be close to the CPU time to calculate a set of structure factors.

The time $T_{\rm total}$ when the Sayre–Ten Eyck algorithm is used with the parameters defined above was estimated using non-optimized programs that generated the electron density in the tests described above. For comparison, the CPU time necessary to calculate structure factors by formula (4), also using a non-optimized program, is given. The CPU times (Table 4) are therefore overestimated but their ratio gives an idea of the gain in computation time for the macromolecular model refinement that can be obtained using the Sayre–Ten Eyck algorithm in comparison with the direct formulae.

Table 4
Comparison of structure factors, calculated by direct formula and through density generation, and the corresponding CPU time, in s.

In all cases, the density was generated at the grid with a step equal to 1/3 of the corresponding resolution. $B_{\rm add}$ was chosen following the algorithm described in §3.9. $R_{\rm total}$ and $R_{\rm highest}$ stand for the R factor for the whole data set and for the reflections in the highest-resolution zone. $T_{\rm total}$ stands for the CPU time used to calculate structure factors through the described algorithms where $T_{\rm dens}$ and $T_{\rm FFT}$ are the CPU time for density generation and for FFT; $T_{\rm form}$ stands for the CPU time to calculate corresponding structure factors through direct formulae. A Pentium III/733 MHz processor was used for the calculations.

	Enkephalin	Crambin	Aldose reductase
No. of atoms	86	831	3346
Resolution, d = 0.9 Å			
No. of reflections	6445	77665	434623
$B_{\rm add}$	4.0	3.1	2.1
$R_{\rm total}$ ($R_{\rm highest}$) (%)	0.43 (0.66)	0.3 (0.65)	0.24 (0.6)
$T_{\rm dens}$, $T_{\rm FFT}$	0.4, 0.1	4.6, 0.9	33.7, 9.4
$T_{\rm total}$	0.5	5.5	43.2
$T_{\rm form}$	1.5	176.5	3983
$T_{\rm form}/T_{\rm total}$	3	32	92
Resolution, d=0.7 Å			
No. of reflections	13754	164874	923661
$B_{\rm add}$	2.5	1.6	0.6
$R_{\rm total}$ ($R_{\rm highest}$) (%)	0.33 (0.54)	0.2 (0.41)	0.15 (0.43)
$T_{\rm dens}$, $T_{\rm FFT}$	0.8, 0.2	11.1, 1.8	84.5, 22.2
$T_{\rm total}$	1.0	12.9	106.7
$T_{\rm form}$	3.3	376	8486
$T_{\rm form}/T_{\rm total}$	3	29	80
Resolution, d = 0.5 Å			
No. of reflections	37554	452044	2533850
$B_{\rm add}$	1.5	0.6	−0.4
$R_{\rm total}$ ($R_{\rm highest}$) (%)	0.13 (0.33)	0.08 (0.22)	0.06 (0.26)
$T_{\rm dens}$, $T_{\rm FFT}$	2.7, 0.5	44.4, 6.9	278, 47
$T_{\rm total}$	3.1	51.3	325
$T_{\rm form}$	9	1034	23336
$T_{\rm form}/T_{\rm total}$	3	20	72

5. Conclusions

A recently started crystallographic study of macromolecules at a subatomic resolution poses a number of questions including the various computational aspects. One of the major computational problems is a fast and very precise calculation of structure factors from an atomic model. The current study demonstrates that an efficient calculation of structure factors at such a resolution through an intermediate generation of an electron-density distribution (Sayre–Ten Eyck algorithm) is possible, thus completing and extending the results by Ten Eyck (1977), Agarwal (1978), Lunin (1982) and Navaza (2002)

to this resolution range. The numerical tests show how much CPU time is gained with the growth in the molecular size.

The traditional models of isotropic or anisotropic atoms with a multi-Gaussian approximation of atomic scattering factors may be completed by Gaussian 'dummy-bond electrons' (Afonine *et al.*, 2002, 2004). This provides one with precise enough models approaching multipolar ones by the quality of details reproduced but presented by a much smaller number of parameters. With this Gaussian form of scattering factors of all model components, the displacement factors can be taken into account very simply and allow an easy application of such density-based calculation of structure factors and all corresponding derivatives. A substitution of this scheme for the scheme of structure-factor calculation by the direct formulae reduces the CPU time by 1–2 orders.

The accurate calculation of structure factors through a generation of an electron-density distribution requires an optimal choice of parameters of the method.

The choice of the correct additional displacement parameter, positive or negative, is a crucial point. In general, in order to determine these parameters for the given model and resolution, one needs to:

(i) find the minimal value of the atomic displacement parameter of the model;

(ii) estimate the optimal value of the effective displacement parameter and the grid step;

(iii) estimate the additional displacement parameter, positive or negative;

(iv) find the corresponding atomic radii.

§3.9 provides the reader with more detailed recommendations.

APPENDIX *A*
Amplification of atomic anisotropy

For an atom with the scattering factor represented as a sum of Gaussian functions (3), the introduction of an anisotropic displacement represented by an anisotropic tensor $\mathbf{U}_{\rm Cart}$ in the formula for electron density gives

$$\rho(\mathbf{r}-\mathbf{r}_0) = Q\sum_{j=1}^{5}\frac{a_j(4\pi)^{3/2}}{|8\pi^2\mathbf{U}_{\rm Cart}+b_j\mathbf{I}|^{1/2}}\exp[-4\pi^2(\mathbf{r}-\mathbf{r}_0)^{\rm T} \times (8\pi^2\mathbf{U}_{\rm Cart}+b_j\mathbf{I})^{-1}(\mathbf{r}-\mathbf{r}_0)], \quad (19)$$

easily comparable with (12). Here, $(\mathbf{r}-\mathbf{r}_0)^{\rm T}$ stands for the transposed vector $(\mathbf{r}-\mathbf{r}_0)$ expressed in Cartesian coordinates and Q is the occupancy of the atom. The symmetric positively defined matrix $\mathbf{U}_{\rm Cart}$ has three real positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 > 0$ that represent the atomic displacement in three mutually orthogonal directions, corresponding eigenvectors of $\mathbf{U}_{\rm Cart}$, which are principal axes of the displacement. The higher the ratio $\lambda_1/\lambda_3 > 1$, the more anisotropic is the atom. The matrix $\mathbf{U}_{\rm Cart}$ varies with the Cartesian basis in which atomic coordinates are given.

In particular, being represented in the basis of its eigenvectors, the matrix $\mathbf{U}_{\rm Cart}$ becomes a diagonal matrix Λ with the eigenvalues λ_1, λ_2, λ_3 on its diagonal. If the matrix $\mathbf{Q}$ corresponds to the transition from this basis in eigenvectors to the initial Cartesian basis, then the matrices are linked by the formula

$$\mathbf{U}_{\rm Cart} = \mathbf{Q}^{-1}\mathbf{\Lambda}\mathbf{Q}. \quad (20)$$

It can be noted that

$$(\mathbf{U}_{\rm Cart})^2 = \mathbf{U}_{\rm Cart}\mathbf{U}_{\rm Cart} = \mathbf{Q}^{-1}\mathbf{\Lambda}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{\Lambda}\mathbf{Q} = \mathbf{Q}^{-1}\mathbf{\Lambda}^2\mathbf{Q} \quad (21)$$

and, more generally,

$$(\mathbf{U}_{\rm Cart})^m = \mathbf{Q}^{-1}\mathbf{\Lambda}^m\mathbf{Q}. \quad (22)$$

A comparison of (20) and (22) shows that the matrix $(\mathbf{U}_{\rm Cart})^m$ corresponds to the anisotropic atom with the same eigenvectors (principal axes of displacement) and with the eigenvalues λ_1^m, λ_2^m, λ_3^m. The ratio

$$\lambda_1^m/\lambda_3^m = (\lambda_1/\lambda_3)^m > \lambda_1/\lambda_3 \quad (23)$$

shows that the anisotropy of this atom is more pronounced than for the initial situation represented by the matrix $\mathbf{U}_{\rm Cart}$. The larger m, the more the anisotropy is increased.

The substitution $(\mathbf{U}_{\rm Cart})^m$ for $\mathbf{U}_{\rm Cart}$ changes, however, the value of the equivalent isotropic parameter $B_{\rm iso}$ (1). Since the trace of the matrix is independent of the choice of the basis,

$$U_{11}+U_{22}+U_{33} = \lambda_1+\lambda_2+\lambda_3, \quad (24)$$

the matrix $(\mathbf{U}_{\rm Cart})^m$ should be multiplied by

$$(\lambda_1+\lambda_2+\lambda_3)/(\lambda_1^m+\lambda_2^m+\lambda_3^m), \quad (25)$$

previously used as an anisotropic tensor for an atom with the same equivalent $B_{\rm iso}$ and with an increased anisotropy.

This simple way to obtain a more anisotropic model was used in the tests described in §3.11. The curves in Fig. 9 are for $m = 6$.

The authors sincerely thank Dr V. Y. Lunin for numerous discussions and suggestions on both the scientific content of the manuscript and the presentation of the results and for his informal help in the preparation of the manuscript. PA and AU are members of GdR 2417 CNRS.

References

Afonine, P. V., Lunin, V. Y., Muzet, N. & Urzhumtsev, A. (2004). *Acta Cryst.*, D**60**. In the press.

Afonine, P. V., Pichon-Pesme, V., Muzet, N., Jelsch, C., Lecomte, C. & Urzhumtsev, A. (2002). CCP4 Newsletter on Protein Crystallography, 41. http://www.ccp4.ac.uk/newsletter41/00_contents.html.

Agarwal, R. C. (1978). *Acta Cryst.* A**34**, 791–809.

Agarwal, R. C. & Isaacs, N. W. (1977). *Proc. Natl Acad. Sci. USA*, **74**, 2835–2839.

Aubry, A., Bouhrakis, N., Sakarellos-Daitsiotis, C. & Marraud, M. (1989). *Biopolymers*, **28**, 27–40.

Ban, N., Freeborn, B., Nissen, P., Penczek, P., Grassucci, R. A., Sweet, R., Frank, J., Moore, P. B. & Steitz, T. A. (1998). *Cell*, **93**, 1105–1115.

Baur, W. & Strassen, V. (1983). *Theor. Comput. Sci.* **22**, 317–330.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucl. Acids Res.* **28**, 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bricogne, G. (1993). *International Tables for Crystallography*, Vol. B, edited by U. Shmueli, pp. 23–106. Dordrecht: Kluwer Academic Publishers.
Brünger, A. T. (1989). *Acta Cryst.* A**45**, 42–50.
CNX (2002). Accelrys, Inc., San Diego, USA. http://www.accelrys.com/Doc/life/CNX2002/CNX_2002_User_Guide/CNX2002UserGuide.pdf.
Cooley, J. W. & Tukey, J. W. (1965). *Math. Comput.* **19**, 297–301.
Grosse-Kunstleve, R. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 477–480.
Hansen, N. K. & Coppens, P. (1978). *Acta Cryst.* A**34**, 909–921.
Jamrog, D. C., Zhang, Y. & Phillips, G. N. Jr (2003). *Acta Cryst.* D**59**, 304–314.
Jelsch, C., Teeter, M. M., Lamzin, V., Pichon-Pesme, V., Blessing, R. H. & Lecomte, C. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 3171–3176.
Kim, K. M., Nesterov, Yu. E. & Cherkassky, B. V. (1984). *Dokl. Acad. Nauk SSSR*, **275**, 1306–1309.
Ko, T.-P., Robinson, H., Gao, Y.-G., Cheng, C.-H. C., DeVries, A. L. & Wang, A. H.-J. (2003). *Biophys. J.* **84**, 1228–1237.
Lamour, V., Barth, P., Rogniaux, H., Poterszman, A., Howard, E., Mitschler, A., Van Dorsselaer, A., Podjarny, A. & Moras, D. (1999). *Acta Cryst.* D**55**, 721–723.
Liu, Q., Weaver, A. J., Xiang, T., Thiel, D. J. & Hao, Q. (2003). *Acta Cryst.* D**59**, 1016–1019.
Lunin, V. Y. (1982). *Optimization of the Calculation of Structure Factors in Protein Crystallography*. NCBI, Pushchino, Preprint. (In Russian.)
Lunin, V. Y., Lunina, N. L., Petrova, T. E., Vernoslova, E. A., Urzhumtsev, A. & Podjarny, A. D. (1995). *Acta Cryst.* D**51**, 896–903.
Lunin, V. Y. & Urzhumtsev, A. (1984). *Acta Cryst.* A**40**, 269–277.
Lunin, V. Y. & Urzhumtsev, A. (1985). *Acta Cryst.* A**41**, 327–333.
Lunin, V. Y., Urzhumtsev, A. & Bockmayr, A. (2002). *Acta Cryst.* A**58**, 283–291.
Maslen, E. N., Fox, A. G. & O'Keefe, M. A. (1992). *International Tables for Crystallography*, Vol. C, edited by A. J. C. Wilson, pp. 476–516. Dordrecht: Kluwer Academic Publishers.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S. & Dodson, E. J. (1999). *Acta Cryst.* D**55**, 247–255.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Navaza, J. (2002). *Acta Cryst.* A**58**, 568–573.
Parthasarathy, S. & Murthy, M. R. N. (1997). *Prot. Sci.* **6**, 2561–2567.
Podjarny, A., Schneider, T. R., Cachau, R. E. & Joachimiak, A. (2003). *Methods Enzymol.* **374**, 324–344.
Sayre, D. (1951). *Acta Cryst.* **4**, 362–367.
Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.
Stewart, R. F. (1969). *J. Chem. Phys.* **51**, 4569–4577.
Teeter, M. M., Roe, S. M. & Heo, N. H. (1993). *J. Mol. Biol.* **230**, 292–311.
Ten Eyck, L. F. (1973) *Acta Cryst.* A**29**, 183–191.
Ten Eyck, L. F. (1977) *Acta Cryst.* A**33**, 486–492.
Urzhumtsev, A. & Lunin, V. Y. (2001). *Acta Cryst.* A**57**, 451–460.
Urzhumtsev, A., Lunin, V. Y. & Vernoslova, E. A. (1989). *J. Appl. Cryst.* **22**, 500–506.
Urzhumtsev, A. & Podjarny, A. D. (1995). *Acta Cryst.* D**51**, 888–895.
Wiest, R., Pichon-Pesme, V., Bénard, M. & Lecomte, C. (1994). *J. Phys. Chem.* **98**, 1351–1362.