Journal of
**Applied
Crystallography**

ISSN 0021-8898

Refinement of proteins at subatomic resolution with *MOPRO*

Benoit Guillot, Laurence Viry, Regis Guillot, Claude Lecomte and Christian Jelsch

Journal of
Applied
Crystallography

ISSN 0021-8898

Received 6 November 2000
Accepted 24 January 2001

Refinement of proteins at subatomic resolution with *MOPRO*

Benoit Guillot,[a] Laurence Viry,[b] Regis Guillot,[a] Claude Lecomte[a] and Christian Jelsch[a]*

[a]Laboratoire de Cristallographie et Modélisation des Matériaux Minéraux et Biologiques, UMR CNRS 7036, Faculté des Sciences BP 239, 54506 Vandoeuvre-lès-Nancy, France, and [b]Centre Charles Hermite, Bâtiment LORIA, Faculté des Sciences, 54506 Vandoeuvre-lès-Nancy, France. Correspondence e-mail: jelsch@lcm3b.uhp-nancy.fr

Crystallography at subatomic resolution permits the observation and measurement of the non-spherical character of the atomic electron density. Charge density studies are being performed on molecules of increasing size. The *MOPRO* least-squares refinement software has thus been developed, by extensive modifications of the program *MOLLY*, for protein and supramolecular chemistry applications. The computation times are long because of the large number of reflections and the complexity of the multipolar model of the atomic electron density; the structure factor and derivative calculations have thus been parallelized. Stereochemical and dynamical restraints as well as the conjugate gradient algorithm have been implemented. A large number of the normal matrix off-diagonal terms turn out to be very small and the block diagonal approximation is thus particularly efficient in the case of large structures at very high resolution.

1. Introduction

The electron cloud around the atoms of a molecule is deformed as a result of chemical bonding and non-bonding interactions (notably hydrogen bonds) between the atoms. Accurate electron density distribution in the crystalline state can be derived from an ultra-high resolution X-ray diffraction experiment (Coppens, 1967). Our laboratory has been involved in charge density studies on molecules of increasing size: the octapeptide LBZ of helical structure (Jelsch *et al.*, 1998) and a scorpion toxin (Housset *et al.*, 2000). With the combined use of synchrotron radiation sources and crystal cryocooling, the number of protein structures refined at a resolution higher than 1.0 Å is increasing continuously.

Recently, the charge density of the protein crambin (46 amino acids) was analysed (Jelsch *et al.*, 2000) at ultra-high resolution (0.54 Å). The crystal structure was refined with a model for charged non-spherical multipolar atoms in order to describe the molecular electron density distribution accurately. The initial multipoles and charges were transferred from our database of average parameters (Pichon-Pesme *et al.*, 1995) derived from the analysis of several crystals of amino acids and small peptides. The average electron density parameters of the protein main chain were then refined against the crambin diffraction data. The crambin electron density deformation is illustrated in Fig. 1 (Jelsch *et al.*, 2000). The analysis of the electron density distribution of human aldose reductase is also underway. Crystals diffracting to very high resolution (Lamour *et al.*, 1999) have been grown for this enzyme of 315 amino acids and diffraction data have been collected to 0.65 Å resolution at the APS synchrotron (Mitschler *et al.*, 2000).

The least-squares computer program *MOPRO* was developed for the charge density analysis of proteins. The software

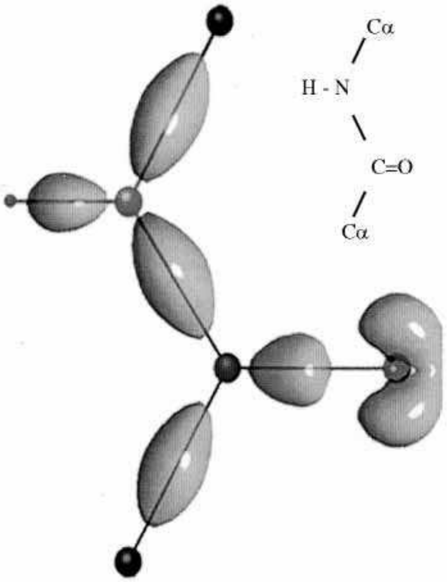

Figure 1
Iso-contour surface of the experimental deformation electron density (level 0.4 e Å^{-3}) along the polypeptide chain of crambin. The deformation density represents the difference between the actual electron density of the molecule and the density calculated for the promolecule, made up of isolated spherical neutral atoms. The density is 'static' in that it is computed for atoms at rest.

is derived from the charge density analysis program *MOLLY* (Hansen & Coppens, 1978). The atomic electron density is described in terms of core, valence and multipolar electron density:

$$\rho_{\rm atom}(\mathbf{r}) = \rho_{\rm core}(r) + P_{\rm val}\kappa^3\rho_{\rm val}(\kappa r) + \sum_l \kappa'^3 R_l(\kappa' r)\sum_m P_{lm} y_{lm\pm}. \qquad (1)$$

The first two terms on the right describe spherically symmetric core-plus-valence density and the third term describes the non-spherical multipolar distribution of the valence electron density of the atoms. The valence-shell electron populations $P_{\rm val}$ account for interatomic charge transfer and the multipole populations P_{lm} account for non-spherical intra-atomic redistribution of valence electron density. The R_l are Slater-type radial functions and the $y_{lm\pm}$ are real spherical harmonic angular functions. The parameters κ and κ' model the expansion and contraction, respectively, of the spherical and multipolar parts of the valence electron density.

2. Restraints

In structure refinement by least squares (LS), one usually seeks to minimize a residual function E, as a weighted sum over the reflections,

$$E = \sum_{\mathbf{H}} W_{\mathbf{H}}(|F_{\mathbf{H}}^{\rm obs}| - |F_{\mathbf{H}}^{\rm calc}|)^2, \qquad (2)$$

and assumes that a structure factor F is a linear function of each parameter for small changes.

When refining the structure of biological macromolecules, because of the limited resolution and the low variables/observations ratio, it is necessary to incorporate stereochemical knowledge into the refinement. Such restrained reciprocal-space least-squares refinement (Konnert, 1976; Konnert & Hendrickson, 1980) is very effective and increases the robustness of the convergence. The minimized function is composite, consisting of the crystallographic error term and other residual components which reflect target geometry as anticipated from small-molecule structures:

$$E = E_{\rm X\text{-}ray} + E_{\rm restrain}. \qquad (3)$$

2.1. Stereochemistry

The geometric restraints implemented in *MOPRO* are the distance, angle and planarity (Table 1). Planarity restraints were programmed by considering the eigenvalues λ_i of the 3 × 3 matrix $\mathbf{V}$ (Urzhumtsev, 1991):

$$V_{ij} = \sum_{\rm atoms}(X_i - \langle X_i\rangle)(X_j - \langle X_j\rangle), \qquad (4)$$

where X_1, X_2 and X_3 represent respectively the x, y and z coordinates of the atoms belonging to the plane. This makes it possible to refine the optimum plane orientation at the same time as the atomic coordinates. The function minimized in *MOPRO* is the dimensionless quantity $\lambda_1\lambda_2\lambda_3/(\lambda_1 + \lambda_2 + \lambda_3)^3$ (Table 1) which models planarity more accurately than the function $\lambda_1\lambda_2\lambda_3$ proposed by Haneef *et al.* (1985) and is simpler to implement than the minimization of the smallest eigenvalue $\lambda_{\rm min}$, proposed by Urzhumtsev (1991).

Stereochemical information is generally no longer necessary for the non-hydrogen atoms in small-molecule crystallography. However, for medium-size molecules and macromolecules at atomic and subatomic resolution, stereochemical information may still be necessary, as some regions can be disordered. For example, in crambin, as much as 30% of the protein atoms have multiple conformations (Yamano *et al.*, 1997; Jelsch *et al.*, 2000). The programming of distance, angle and planarity restraints was therefore deemed necessary for the structure and charge density refinement of crambin.

On the other hand, less electron density is associated with the hydrogen atoms compared with the heavier atoms as the hydrogen atoms lack the core electrons. Thus, the hydrogen atoms may still need to have their positions defined by stereochemical information in small-molecule crystallography. In crystallographic refinement software like *SHELXL97* (Sheldrick & Schneider, 1997) or *CNS* (Brünger *et al.*, 1998), the positions of the hydrogen atoms can be constrained according to standard geometries. Alternatively, this can be achieved using restraints as in *MOPRO*. Restraints are a convenient and smoother way to refine properly the positions of hydrogen atoms in charge density studies of small molecules. Notably, the H—X distance to the neighbouring atom can be restrained to standard values derived from neutron diffraction studies (Allen, 1986).

Table 1
List of restraints implemented in *MOPRO*.

Keyword	Description	Function minimized
XYZRES	Coordinates X, Y, Z	$(X - X_r)^2$
DISTAN	Distance d between two atoms	$(d - d_r)^2$
ANGLER	Angle θ between three atoms	$(\theta - \theta_r)^2$
PLANAR	Planarity	$\lambda_1\lambda_2\lambda_3/(\lambda_1 + \lambda_2 + \lambda_3)^3$
UIJRES	Thermal displacement parameters	$(U^{ij} - U_r^{ij})^2$
RIGIDB	Rigid bond	$(Z_U - Z_{U'})^2$
URATIO	Isotropic thermal displacement parameter riding on bonding atom	$(U - U_r)^2$
ISOTRO	Thermal displacement ellipsoid limited anisotropy	$\sum(U^i - U^j)^2/\langle U^{ii}\rangle^2 = 3[\sum_i(U^{ii} - \langle U^{ii}\rangle)^2 + \sum_{i\neq j}U^{ij2}]/\langle U^{ii}\rangle^2$
KAPPAR	Expansion coefficient κ and κ'	$(\kappa - \kappa_r)^2$

2.2. Thermal motion

The values of the coordinates and thermal displacement parameters U^{ij} can also be restrained in *MOPRO*. Such restraints can be applied if neutron diffraction data are available for the crystal studied. In the refinement *versus* the X-ray diffraction data, the target values for the coordinates and thermal displacement parameters can be set to their values in the neutron structure, the allowed deviation being set to the neutron experimental error. However, as combined X-ray and neutron diffraction studies often display a thermal difference between the two data sets, an additional 'thermal scale factor' may be introduced. Alternatively, in the case of small compounds, thermal displacement parameters can be derived from an *ab initio* calculation of the harmonic force field (Flaig *et al.*, 1998) and could be used as targets in *MOPRO* for the U^{ij} restraints. These restraints, which are more flexible than constraints, permit the combination of information from different sources and are helpful in the deconvolution of the deformation density and the thermal motion.

Several restraints on the thermal displacement parameters have been implemented. Hydrogen atoms are often assigned an isotropic displacement parameter proportional to the equivalent B factor of the connecting atom. The multiplier coefficient generally used is respectively 1.5 or 1.2 for hydrogen atoms with or devoid of a rotation degree of freedom. The isotropic B factor of the hydrogen atoms can be restrained in *MOPRO* with a chosen proportionality coefficient. This restraint was applied to the main-chain hydrogen atoms in the charge density refinement of the main-chain polypeptide in crambin (Jelsch *et al.*, 2000). Such information can also be useful in small-molecule X-ray electron density studies, as the thermal motion of hydrogen atoms is not well defined.

A restraint limiting the anisotropy of the atomic thermal motion is also implemented. When the thermal tensor $\mathbf{U}^{ij}$ of an atom has three identical eigenvalues, the thermal motion is isotropic. The anisotropy of the $\mathbf{U}^{ij}$ tensor can be conveniently limited with the use of the dimensionless quadratic function $\sum(U^i - U^j)^2/\langle U^i\rangle^2$ describing the global discrepancy between the three eigenvalues U^i (Table 1). This represents a model between isotropy and anisotropy and avoids non-positive defined $\mathbf{U}^{ij}$ thermal motion tensors. Such isotropy restraints were applied in the case of crambin (Jelsch *et al.*, 2000) to the partially occupied water molecules and disordered protein atoms, for which the electron density map ($2F_o - F_c$ at the 5σ level) showed a well defined thermal displacement ellipsoid. Other partially occupied atoms were set to be isotropic.

A rigid-bond restraint, which renders the mean-square displacement for two covalently bonded atoms similar along the bond direction, is also programmed. Hirshfeld (1976) considered that a rigid-bond discrepancy of $\Delta_{ZU} < 10^{-3}$ Å^2 was a good criterion to assess the reliability of the thermal displacement parameters in a crystal structure. The deviation from the rigid bond in actual crystal structures devoid of crystallographic errors and uncertainties can presumably be expected to be one order of magnitude smaller than this value. This additional information on the thermal motion can facilitate its deconvolution from the deformation density. As shown by Rosenfield *et al.* (1978), the rigid-bond criterion may be extended to any pair of atoms belonging to a rigid group, such as aromatic cycles.

Table 2
Target values for the application of restraints on the expansion/contraction parameters κ and κ' of protein main-chain atoms and the phosphate group.

The values are from the updated experimental database of transferable charge density parameters (Pichon-Pesme *et al.*, 1995). The root-mean-square deviations are shown in parentheses. The values of the ξ and n_l coefficients up to the maximal multipolar expansion (dipoles, quadrupoles, ...) applied in the Slater-type radial function are also given. The radial function is of the form $R_{n_l}(r) = r^{n_l}\exp(-\kappa'\xi r)$.

Atom type	$\langle\kappa\rangle$	$\langle\kappa'\rangle$	n_l	ξ (bohr^{-1})
C	0.994 (4)	0.93 (2)	2, 2, 3	3.0
O	0.975 (6)	0.95 (3)	2, 2, 3	4.5
N	0.985 (3)	0.86 (2)	2, 2, 3	3.8
Cα	0.991 (5)	0.89 (2)	2, 2, 3	3.0
Cα (Gly)	0.992 (3)	0.94(1)	2, 2, 3	3.0
Hα	1.16 (2)	1.07 (3)	1	2.26
H (N)	1.20 (4)	0.98 (3)	1	2.26
P	1.05 (2)	1.04 (3)	6, 6, 6, 6	3.6
O (O–P)	0.969 (1)	0.94 (3)	1, 2, 4	4.5

2.3. Expansion/contraction of the valence density

The use of external information obtained from accurate small-molecule crystallographic analyses as restraints in a macromolecular refinement can be extended to charge density parameters. The case of the parameters κ and κ' describing the radial expansion/contraction of the valence electron density is noteworthy, these features often being the most difficult to refine in electron density studies, even for small molecules. Peres *et al.* (1999) and Volkov *et al.* (1999, 2000) focus particularly on the radial expansion κ' of the multipolar deformation density and suggest that this parameter be constrained. The expansion/contraction parameters can be restrained in *MOPRO* during the multipolar refinement. For instance, average values from the database (Pichon-Pesme *et al.*, 1995) of experimental multipole parameters can be the selected targets (Table 2). Targets for the κ and κ' variables can also be deduced from theoretical calculations (Volkov *et al.*, 1999). Unlike the constrained refinement where the value is fixed, the refinement with restraints allows adjustment of the parameters with respect to the specific chemical environment of the considered atom. For instance, restraining the κ coefficient of the pyrophosphate atoms was deemed necessary in the multipolar refinement of NAD$^+$, the oxidized form of the nicotinamide adenine dinucleotide molecule (Guillot *et al.*, 2000), for which charge density analysis is underway. In the unrestrained refinement, the electron density maps (Fig. 2*a*) show a strong density on the phosphorus nucleus; this is caused by the κ parameter, which refined to an unusual value of 0.917 (3). The charges are unrealistic on the oxygen atoms (slightly positive) and the phosphorus atom (−1.7 e). The

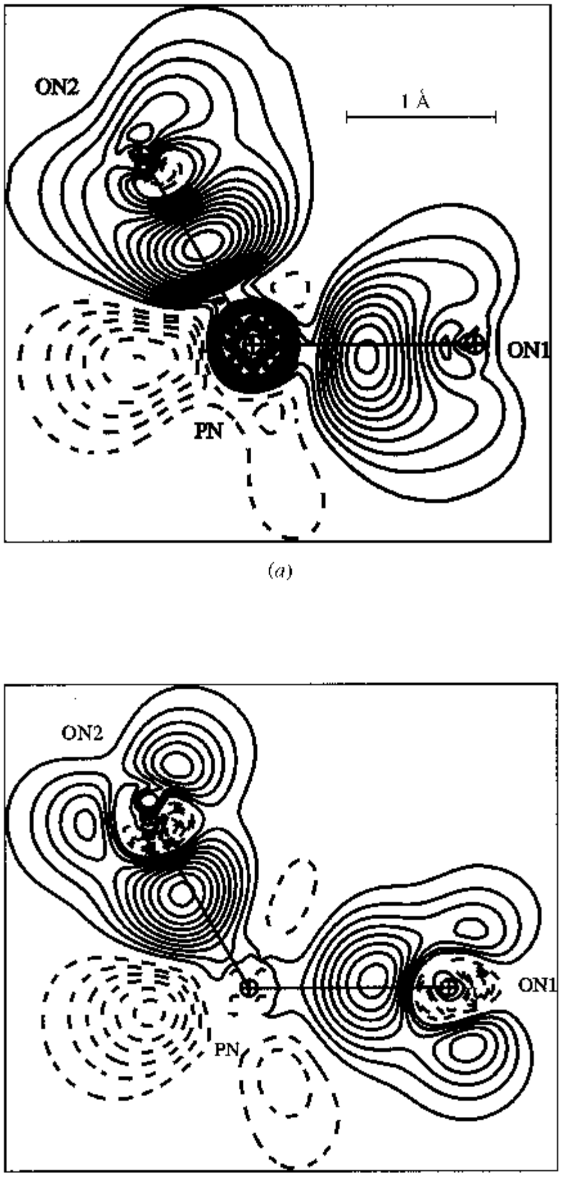

Figure 2
Static deformation electron density map in the ON1–PN–ON2 plane of the NAD$^+$ pyrophosphate group: (*a*) without restraints; (*b*) with the expansion/contraction coefficients κ and κ' of the phosphorus and pyrophosphate oxygen atoms restrained to the values extracted from the multipolar-parameters database (Table 2).

application of restraints on the radial expansion parameter κ leads to a physically meaningful charge density on the NAD$^+$ pyrophosphate (Fig. 2*b*).

3. Algorithmic optimizations

3.1. Conjugate gradients

In a cycle of least-squares minimization of the function E [equation (2)], the n parameters vector shift $\delta\mathbf{x}$ to be applied to the refined parameters is obtained by solving a system of n linear equations of the form

$$\mathbf{A}\,\delta\mathbf{x} = \mathbf{b}, \qquad (5)$$

where $\mathbf{A}$ is the n^2 symmetric positive definite matrix of normal equations and $\mathbf{b}$ is a vector of dimension n.

The equation system is solved in *MOLLY* (Hansen & Coppens, 1978) by inverting the normal matrix $\mathbf{A}$; this is computationally prohibitive for large systems with many variables. This procedure can be replaced by the conjugate gradient algorithm, which is an iterative procedure, first described by Hestenes & Stiefel (1952) and Fletcher & Reeves (1964), that avoids the inversion of the normal matrix and is less sensitive to matrix singularities.

The solution of the equation system (5) is approximated by successive displacements along $\mathbf{A}$ conjugate directions in the parameter space until convergence. The convergence is usually reached in a number of iterations that is much smaller than the number of variables n. The procedure, as implemented in *MOPRO*, is described in Appendix *A*.

3.2. Matrix preconditioning

The convergence of the conjugate gradient algorithm can be very slow. For example, in the refinement of aldose reductase and crambin, the number of necessary iterations can reach several hundreds (Figs. 3*a* and 3*b*). The rate of convergence of the conjugate gradient algorithm is related to the ratio between the largest and the smallest eigenvalue, also called the condition number of the normal matrix $\mathbf{A}$ (Tronrud, 1992).

To render the convergence faster and more robust, the normal matrix $\mathbf{A}$ can be preconditioned (Tronrud, 1992) in order to obtain a condition number closer to unity. Using this property, it is legitimate to transform the normal matrix $\mathbf{A}$ into $\mathbf{P}^{-1}\mathbf{A}$, where the matrix $\mathbf{P}^{-1}$ is a symmetric positive definite matrix called the preconditioner. In the preconditioned case, equation (5) becomes

$$\mathbf{P}^{-1}\,\mathbf{A}\,\delta\mathbf{x} = \mathbf{P}^{-1}\,\mathbf{b}. \qquad (6)$$

The general preconditioning procedures are detailed in Appendix *B*. The preconditioner $\mathbf{P}$ used in *MOPRO* is simply the diagonal part of the matrix $\mathbf{A}$. The matrix $\mathbf{P}^{-1}$ is then a rough approximation of the inverse matrix $\mathbf{A}^{-1}$ and the product $\mathbf{P}^{-1}\mathbf{A}$ is close to the identity matrix. This is especially true with high-resolution diffraction data (see the next paragraph), where the magnitudes of the non-diagonal elements of the matrix $\mathbf{A}$ are often small compared with the diagonal elements and, *a fortiori*, when the block-diagonal approximation of the matrix $\mathbf{A}$ is used.

The efficiency of the preconditioning is illustrated in Figs. 3(*a*) and 3(*b*) for the two protein systems at subatomic resolution. In the refinement cycle described for aldose reductase, the convergence requires 370 iterations, while with the preconditioned matrix only 19 iterations are necessary (Fig. 3*a*). In both the preconditioned and the non-preconditioned case, the parameter shifts $\delta\mathbf{x}_i$ converge to the value obtained by matrix inversion.

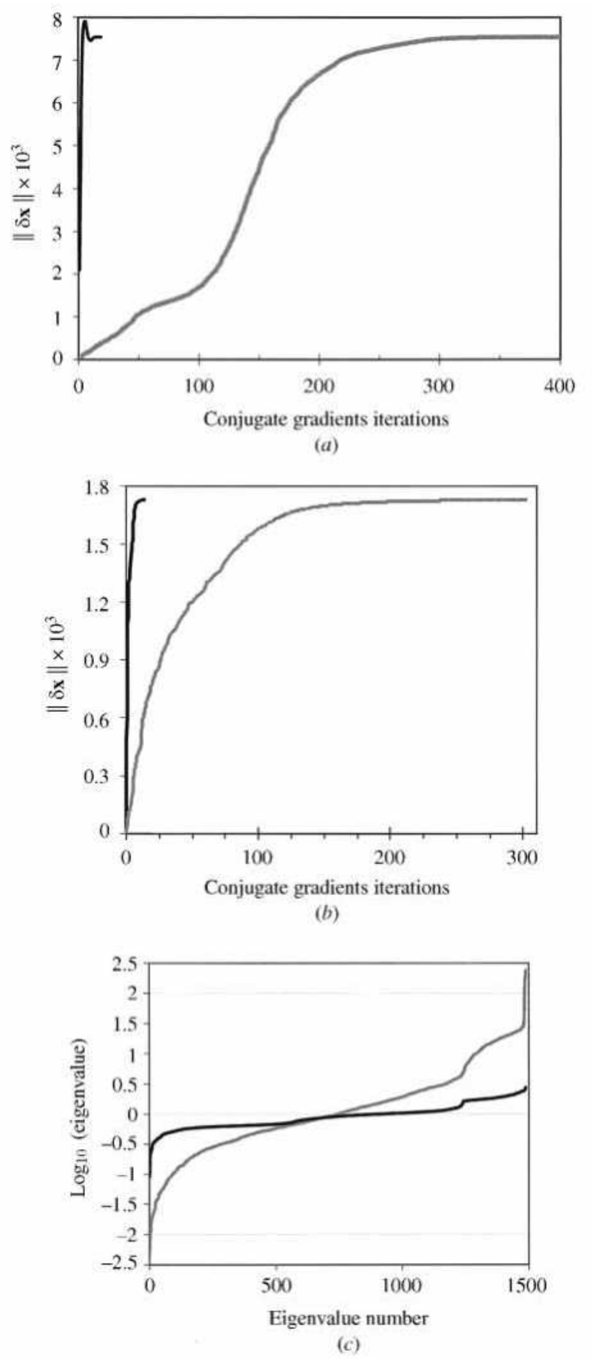

Figure 3
Effect of the normal matrix preconditioning on the rate of convergence of the conjugate gradient procedure. The norm of the shift vector $\delta\mathbf{x}_i$ at each conjugate gradient cycle is represented until convergence is reached ($\varepsilon < 10^{-7}$). (*a*) In this minimization cycle, 7500 thermal displacement parameters U^{ij} of aldose reductase atoms were refined against 311 000 high-resolution reflections ($d < 1$ Å). Black curve: preconditioned normal matrix. Grey curve: non-preconditioned matrix. (*b*) Refinement of 1488 U^{ij} thermal displacement parameters for the non-disordered atoms of the protein crambin at 0.54 Å resolution. Black curve: preconditioned normal matrix. Grey curve: non-preconditioned matrix. (*c*) Corresponding eigenvalues spectrum of the preconditioned matrix $\mathbf{P}^{-1}\mathbf{A}$ (black curve) and the non-preconditioned normal matrix $\mathbf{A}$ (grey curve) in the crambin refinement. The condition numbers (ratio of the largest and the smallest eigenvalues) are 31 and 52031, respectively. The eigenvalues of $\mathbf{A}$ represented on the diagram have been divided by the median eigenvalue of $\mathbf{A}$.

The eigenvalue spectra of matrices $\mathbf{A}$ and $\mathbf{P}^{-1}\mathbf{A}$ have been compared for the refinement cycle of crambin described in Fig. 3(*b*). The *SSPEV* routine from the *LAPACK* library (http://www.netlib.org/lapack/) is able to compute the eigenvalues of symmetric definite positive matrices. The matrix $\mathbf{P}^{-1}\mathbf{A}$ is not symmetric but has the same eigenvalues as $\mathbf{P}^{-1/2}\mathbf{A}\mathbf{P}^{-1/2}$, which is symmetric. To determine the eigenvalue spectrum of $\mathbf{P}^{-1}\mathbf{A}$, the *SSPEV* routine was thus applied to the $\mathbf{P}^{-1/2}\mathbf{A}\mathbf{P}^{-1/2}$ matrix.

As can be seen in Fig. 3(*c*), the eigenvalue spectrum of matrix $\mathbf{P}^{-1}\mathbf{A}$ is considerably narrower (on a logarithmic scale) than the spectrum of matrix $\mathbf{A}$. The condition number is reduced by more than three orders of magnitude from 52031 to 31. Correlatively, 303 iterations are necessary to reach convergence criterion with the non-preconditioned matrix, and only 15 for the preconditioned one (Fig. 3*b*). Moreover, the major variations of $\delta\mathbf{x}_i$ occur in the first ten steps.

3.3. Normal matrix sparsity

The normal matrix element concerning the refined parameters p_i and p_j is obtained from the weighted summation of the structure-factor derivative products over the reflections $\mathbf{H}$:

$$A_{ij} = \sum_{\mathbf{H}} W_{\mathbf{H}}[\partial F^{\rm calc}(\mathbf{H})/\partial p_i][\partial F^{\rm calc}(\mathbf{H})/\partial p_j]. \qquad (7)$$

The derivative products concerning the restraints are similarly added to the matrix elements.

If n is the number of refined parameters, the normal matrix contains $n(n + 1)/2$ independent elements. In the case of macromolecules containing several thousands of atoms, it is recommended to omit the off-diagonal elements with small values, as storing the full normal matrix can become computationally prohibitive. In all the crystallographic refinements of macromolecules at atomic and subatomic resolution performed in the laboratory, the normal matrix turns out to be very sparse, as nearly all of the off-diagonal elements have very small values (Jelsch, in preparation). A normal matrix element A_{ij} is associated with a pair of parameters p_i and p_j, and consequently with a pair of atoms a_i and a_j.

The elements of the normal matrix have a global tendency to decrease rapidly with the Patterson vector length between the concerned atoms, as already observed by Templeton (1999). To evaluate the relative magnitude of the elements, the normalized normal matrix $\mathbf{A}'$ shall be considered:

$$A'_{ij} = A_{ij}/[(A_{ii}\,A_{jj})/2]. \qquad (8)$$

The diagonal elements of $\mathbf{A}'$ are all equal to unity and the magnitude of an off-diagonal element A'_{ij} can then be assessed by comparison with unity. For example, in the case of aldose reductase at 0.65 Å resolution, the positional and displacement parameters of the non-hydrogen protein atoms were refined. The distances below 1 Å involve disordered atoms. The matrix elements are on average very small ($|A'_{ij}| < 0.05$) for atomic distances longer than 1 Å (Fig. 4). Thus, matrix elements with a significant magnitude correspond generally to pairs of parameters of the same atom.

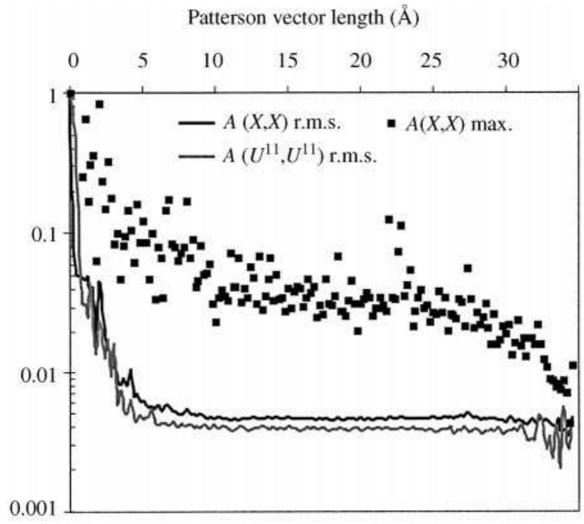

Figure 4
Evolution of the normalized matrix elements as a function of Patterson vector length in the crystallographic refinement of aldose reductase. The matrix elements have been ordered with increasing Patterson vector lengths and grouped in shells of 0.2 Å for the calculation of the root-mean-square (r.m.s.) and maximum values. Black curve: r.m.s. value of the $A(X, X)$ elements. Small black squares: maximum value of $|A(X, X)|$. Grey curve: r.m.s. value of $A(U^{11}, U^{11})$ elements.

The *a priori* knowledge that most of the matrix elements are negligible and the application of the conjugate gradient method in the software *MOPRO* avoids the computation and the inversion of the full normal matrix. The block-diagonal approximation is thus particularly efficient in the least-squares refinement of macromolecules at very high resolution.

3.4. Parallelization

Despite the considerable speed improvements of scalar computers during recent years, the use of a parallel architecture should be valuable for some heavy applications in crystallography. The charge density analysis of protein structures demands much computer time and memory, as the number of both observed reflections and refined parameters is large. For example, in the case of human aldose reductase, the number of unique reflections measured at 0.65 Å resolution reaches 511 000. Furthermore, the computation of structure factors takes longer when using a multipolar atom model rather than a spherical model. In the multipolar model case, the atomic density is decomposed into the core and valence electron density [equation (1)], which requires several products of spherical harmonic and Bessel functions. The number of parameters describing an atom is consequently increased (three coordinates, six anisotropic displacement parameters, one occupancy factor, three dipoles, five quadrupoles, seven octupoles, ...).

Fast Fourier transform (FFT) algorithms permit a more rapid computation of the structure factors and their derivatives. FFT algorithms are efficiently applied in the case of spherical atoms with anisotropic thermal motion, notably by modelling the atomic electron density as a sum of Gaussian functions. In the case of multipolar atoms, FTT procedures are less convenient to apply; they have not been implemented in *MOPRO*.

The software *MOLLY*, on which *MOPRO* is based, was designed for small-molecule refinement and was written in Fortran 77. Some serial optimizations were necessary to adapt the source code to a more easily parallelizable form and some Fortran 90 features, such as dynamical array handling, were implemented. These optimizations resulted in a speed-up factor of about three, prior to any parallelization. The principal steps in the program *MOPRO* are shown in Fig. 5.

The program *MOPRO* has been parallelized using the OpenMP Fortran Application Programming Interface (API) on an SGI Origin 2000 equipped with 250 MHz R10000 processors. OpenMP API allows calculations in a run to be shared among several processes (threads), by inserting directives in the original code. These directives are recognized by the compiler if the need for a parallel run is specified (`-mp` option); otherwise they are considered as Fortran comments. OpenMP API is independent of the programming language used and can be ported on any shared-memory multiprocessor computer, as long as a compatible compiler is available. There are two main ways to parallelize a code (Fig. 6), corresponding either to iterative or to non-iterative work sharing. The iterative work sharing applies when loop iterations (`DO` or `WHILE` statements) are distributed among several processes. In this case, each iteration or 'skunk' of iterations are executed simultaneously by different threads, instead of being executed consecutively as in standard serial loops (Fig. 6*a*).

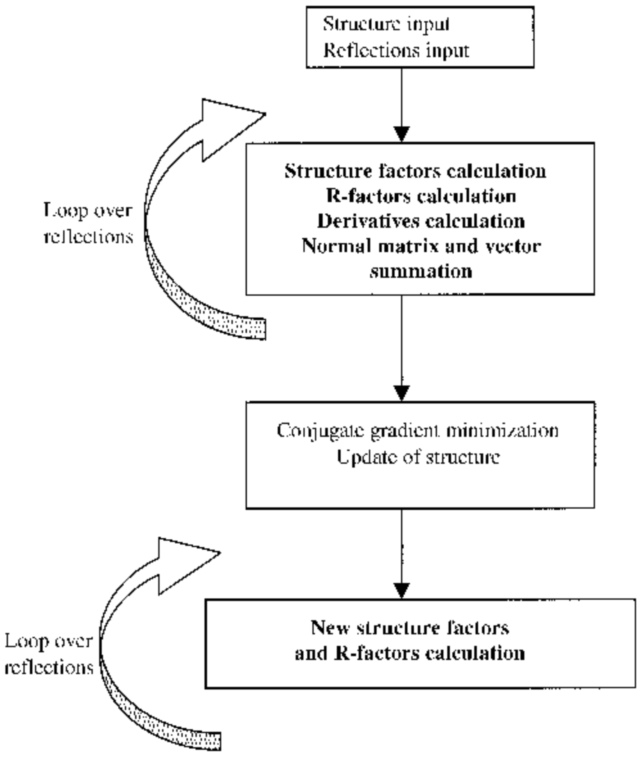
Structure input
Reflections input
Structure factors calculation
R-factors calculation
Derivatives calculation
Normal matrix and vector summation
Loop over reflections
Conjugate gradient minimization
Update of structure
New structure factors and R-factors calculation
Loop over reflections

Figure 5
Diagram describing the general architecture of the program *MOPRO*. The procedures that have been parallelized are in bold.

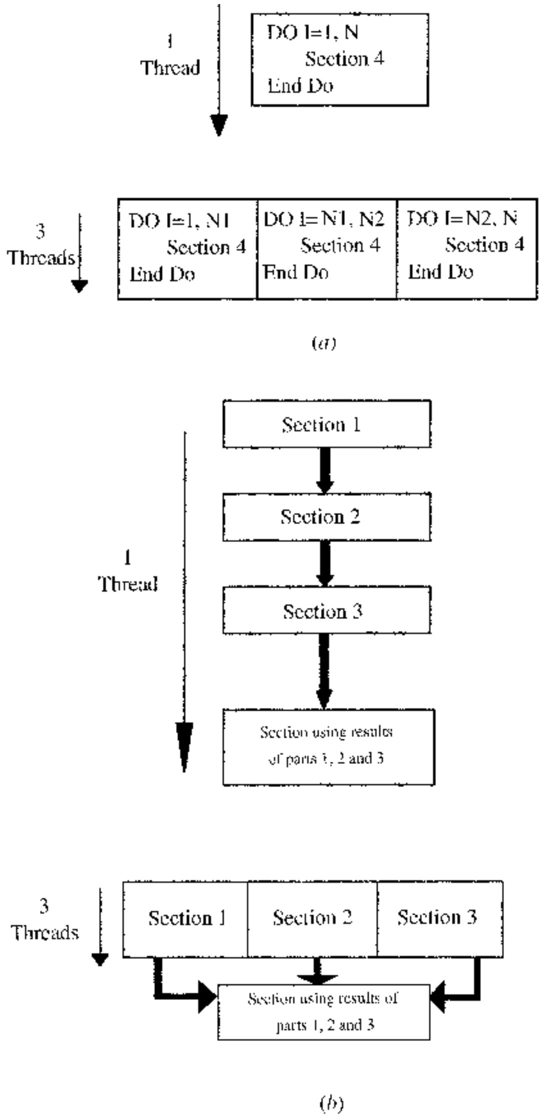

Figure 6
(*a*) Schematic representation of the iterative work-sharing method. A loop over *N* iterations is computed by one thread (top of the figure). After parallelization, it is chopped into three loops over *N*/3 iterations computed simultaneously by three threads (bottom). (*b*) Schematic representation of the non-iterative work-sharing construct. At the top of the figure, three independent sections of the code are executed consecutively by one thread. At the bottom, they are executed simultaneously by three threads in parallel mode.

In the non-iterative work sharing, two or more independent sections of the code are executed simultaneously by several threads in the parallel construct (Fig. 6*b*). In a sequential run, these sections would have been executed successively.

In both cases, the main process (master thread) of the sequential part of the program splits into several threads when a parallel directive is met. At the end of the parallel section or loop, the program returns to the serial mode. A program with frequent switches to parallel mode is denoted fine-grained, as opposed to coarse-grained programs. As the threads have to be synchronized in a parallel calculation and as the system needs some time to switch to parallel mode, the time saving in fine-grained programs may be limited. This is particularly true when many threads are used. More details about OpenMP are available at http://www.openmp.org. Optimal parallel calculations are performed with one thread per processor, as a number of threads larger than the number of available processors does not lead to additional gains in speed.

The efficiency of the parallelization can be estimated from the speed-up ratio $s(p) = t(p = 1)/t(p)$, where $t(p)$ is the execution time for p threads (Fig. 7*a*). Another criterion is the efficiency $e(p) = s(p)/p$ (Fig. 7*b*). In the current paper, the speed-up and efficiency are defined for the entire optimized program, *i.e.* the execution time including both the sequential and the parallel parts of the code.

In the serial execution of *MOPRO*, the structure factors, their derivatives and the normal matrix-element calculations, and secondly the conjugate gradients procedure, are the most computer-time-consuming parts of the program, especially for

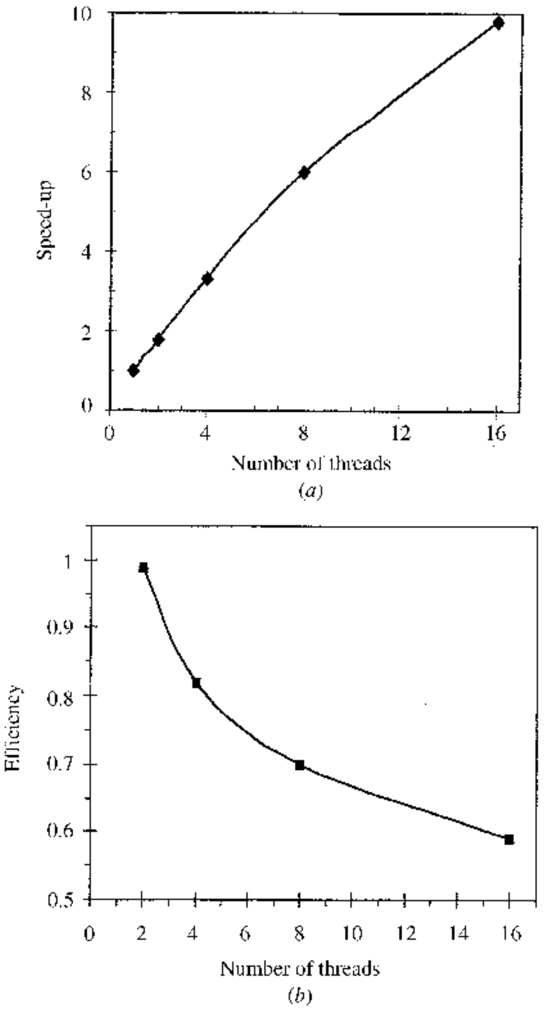

Figure 7
Speed-up (*a*) and efficiency (*b*) of the parallelization as a function of the number of threads during one cycle of coordinates refinement for aldose reductase at 0.65 Å resolution. The structure contains 5916 atoms; 10500 variables were refined against 470000 observed structure factors.

large systems with many refined parameters. The loop parallelization method (iterative work sharing) was applied to these parts of the software *MOPRO*. The computation of the structure factors, of their derivatives and of the normal matrix are built inside a loop over all the reflections. As each result is independent from the others, the reflections were distributed among several threads.

The speed-up and efficiency obtained by the parallelization in the case of aldose reductase are shown in Fig. 7. The refinement cycle includes the matrix construction from the structure factors and their derivatives (parallel loop), the conjugate gradient minimization (serial) and a final structure-factor calculation (parallel) (Fig. 5). The execution time when 16 processors are used is reduced by a factor of ten to about 2 h. The parallelization thus allows an appreciable time saving and renders the multipolar refinement with the software *MOPRO* applicable to macromolecules.

4. Electroneutrality constraint

The program *MOPRO* allows the refinement of the valence populations. An electroneutrality constraint on the atomic charges may be necessary to keep the total number of electrons constant. The electroneutrality constraint method initially implemented in *MOLLY* uses some properties of the variance–covariance matrix (Hamilton, 1964), which is the inverse of the normal matrix **A**. In *MOPRO*, with the use of the conjugate gradients algorithm, the variance–covariance matrix is not computed and this method is thus not applicable. The method described by Raymond (1972) has therefore been implemented in *MOPRO*. As the sum of the valence populations is set constant, the last parameter Pv_n can be set to a linear combination of the $n - 1$ precedent independent parameters and the number of variables is decreased by one. Then, if δPv_i is the shift on the ith refined valence population, the shift to apply to the last parameter is

$$\delta Pv_n = -\sum_{1,n-1} \delta Pv_i. \tag{9}$$

The derivatives of the structure factors F with respect to the $n - 1$ linearly independent Pv variables have then to be modified accordingly:

$$\partial F/\partial Pv_i \rightarrow \partial F/\partial Pv_i - \partial F/\partial Pv_n. \tag{10}$$

This method leads to the same valence population shifts as with the Hamilton (1964) method.

5. Estimated standard deviations

The properties of the estimated standard deviations (e.s.d.'s, or standard uncertainties) of the crambin atomic coordinates have been analysed at atomic (1 Å) and subatomic (0.54 Å) resolution. The e.s.d.'s discussed here were obtained by inversion of the (non-full) normal matrix during a coordinates refinement cycle. Fig. 8 shows the e.s.d.'s for the X coordinates as a function of the atomic thermal motion. No stereochemical restraints were applied in this analysis. When the data were further truncated at 1.5 Å resolution, the matrix turned out to be singular. This illustrates that geometrical restraints are necessary information to incorporate at that resolution.

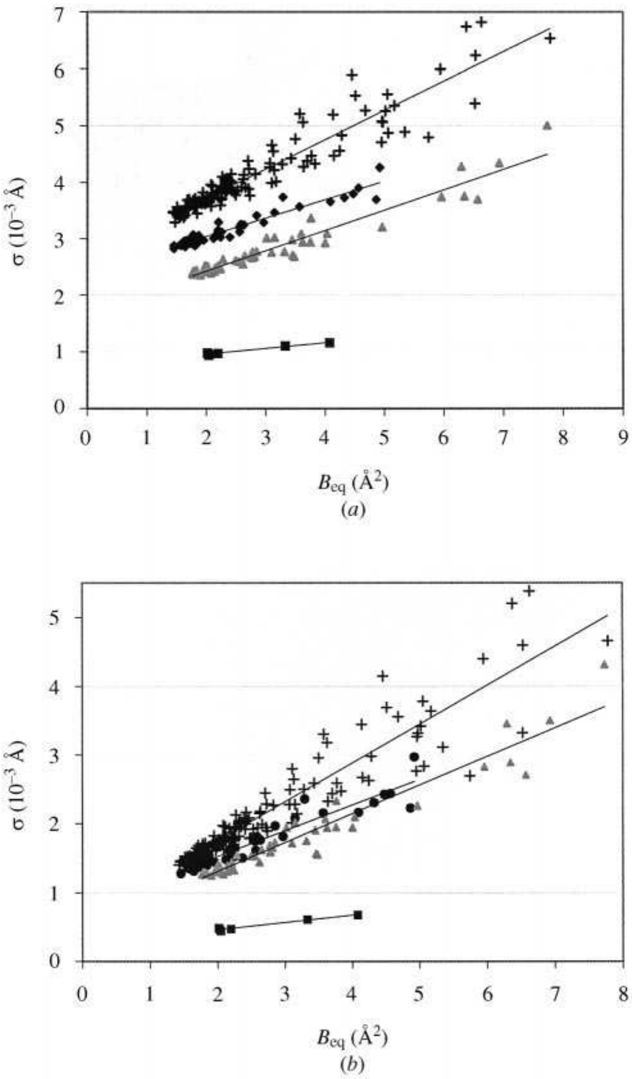

Figure 8
Estimated standard deviations of the X coordinates as a function of the equivalent isotropic thermal displacement parameter $B_{eq} = \langle B^{ii} \rangle$. The coordinates for the non-disordered parts of the protein crambin were refined. (*a*) All reflections to 0.54 Å resolution. (*b*) Truncation at 1 Å resolution. The different atom types can be distinguished (sulfur: squares; oxygen: triangles; nitrogen: circles; carbon: crosses).

The e.s.d.'s do increase, as expected, with the thermal motion of the atoms (Stanley, 1965) and there is a twofold increase of the precision of the coordinates from 1 to 0.54 Å resolution. The e.s.d.'s of the coordinates are also clearly dependent on the atom type, especially at 1 Å resolution (Fig. 8*a*). The positions of the six sulfur atoms of crambin are defined with the best precision. The higher the scattering factor of an atom in the considered resolution range, the more precise are the coordinates. The scattering factor of an atom at low resolution is proportional to total number of electrons (core plus valence) of the atom. Therefore, at 1 Å resolution, the carbon, nitrogen and oxygen atoms form three distinguished clusters in the B *versus* e.s.d. graph (Fig. 8*a*).

On the other hand, at very high resolution ($d < 0.9$ Å), the diffraction essentially arises from the core electrons (Jelsch *et al.*, 1998). Thus, in the upper resolution ranges, the scattering factors of the carbon, nitrogen and oxygen atoms are basically similar, as there are two core electrons in all three cases. Thus, when all reflections up to subatomic resolution are used, the e.s.d.'s are less dissimilar for these three atom types (Fig. 8*b*).

Moreover, the e.s.d.'s diminish more in relative value for the atoms with low thermal motion when the resolution is increased from 1 to 0.54 Å. This correlates well with the fact that the diffraction at subatomic resolution originates essentially from the atoms with low thermal displacement parameters ($B < 3$ Å^2).

6. Conclusions

In the past, precise analyses of electronic distribution have been restricted to small molecules, which generally satisfy the necessary conditions of subatomic resolution and low thermal motion. With the availability of third-generation synchrotron X-ray beamlines, some protein diffraction data can now be measured at subatomic resolution. The electron density analysis of such macromolecules has thus become a feasible challenge. Currently available multipolar refinement programs, dedicated to small molecules, would need a prohibitively excessive amount of CPU time if applied to large structures. Thus, the algorithmic optimizations and the parallelization implemented in *MOPRO* were necessary in order to refine the charge density of macromolecules, such as aldose reductase, in a reasonable amount of computer time.

Provided that the protein atoms have standard names, a transfer program is available to transform a *SHELXL* structure file into a *MOPRO* input file. The information contained in the database of multipolar parameters (Pichon-Pesme *et al.*, 1995), notably the definition of the local axis system, is directly transferred to the *MOPRO* input file containing the coordinates, the thermal displacement and charge density parameters. These improvements, in conjunction with the implementation of restraints necessary in macromolecular crystallography, make the program *MOPRO* a reliable tool for protein charge density analysis.

APPENDIX *A*
The conjugate gradient algorithm

The starting conditions of the iterative procedure applied in *MOPRO* to solve equation (5) are

$$\mathbf{p}_0 = \mathbf{r}_0 = \mathbf{b} - \mathbf{A}\,\delta\mathbf{x}_0, \tag{11}$$

where $\delta\mathbf{x}_0$ is the initial parameters vector shift, usually set to zero. The following iterative process is repeated to obtain successive estimations of the parameters shift $\delta\mathbf{x}_i$:

$$\alpha_i = (\mathbf{r}_i^T\,\mathbf{r}_i)/(\mathbf{p}_i^T\,\mathbf{A}\,\mathbf{p}_i), \tag{12}$$

$$\delta\mathbf{x}_{i+1} = \delta\mathbf{x}_i + \alpha_i\,\mathbf{p}_i, \tag{13}$$

$$\mathbf{r}_{i+1} = \mathbf{r}_i - \alpha_i\,\mathbf{A}\mathbf{p}_i, \tag{14}$$

$$\beta_{i+1} = (\mathbf{r}_{i+1}^T\,\mathbf{r}_{i+1})/(\mathbf{r}_i^T\,\mathbf{r}_i), \tag{15}$$

$$\mathbf{p}_{i+1} = \mathbf{r}_{i+1} + \beta_{i+1}\,\mathbf{p}_i, \tag{16}$$

where $\mathbf{p}_i$ is the direction of search in the parameter space at the ith iteration and $\mathbf{r}_i$ is equal to the residual $\mathbf{b} - \mathbf{A}\,\delta\mathbf{x}_i$ at the ith iteration.

The convergence is assumed to be reached when the residual norm $\mathbf{r}_i$ is a significantly small fraction ε of the initial residual norm $\mathbf{r}_0$. The tolerance term ε is defined by the user and was set to 10^{-6}.

APPENDIX *B*
Matrix preconditioning

To apply the matrix preconditioning, the matrix $\mathbf{P}$ can be decomposed according to the Cholesky factorization:

$$\mathbf{P} = \mathbf{L}\,\mathbf{L}^T. \tag{17}$$

The system of equations (5) can then be rewritten

$$\mathbf{L}^{-1}\,\mathbf{A}\,\mathbf{L}^{T-1}\,(\mathbf{L}^T\delta\mathbf{x}) = \mathbf{L}^{-1}\,\mathbf{b}. \tag{18}$$

The conjugate gradient procedure can be rewritten according to the previous equation, with $\mathbf{A}$, $\delta\mathbf{x}$ and $\mathbf{b}$ replaced by $\mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{T-1}$, $\mathbf{L}^T\delta\mathbf{x}$ and $\mathbf{L}^{-1}\mathbf{b}$, respectively, but these equations have the disadvantage that $\mathbf{L}$ and $\mathbf{L}^{-1}$ must be computed and that the solution $\delta\mathbf{x}'$ must be transformed into the original set of parameters by the transformation $\delta\mathbf{x} = (\mathbf{L}^T)^{-1}\,\delta\mathbf{x}'$. However, after application of the following variables substitution,

$$\delta\mathbf{x}' = \mathbf{L}^T\delta\mathbf{x}, \quad \mathbf{r}' = \mathbf{L}^{-1}\,\mathbf{r}, \quad \mathbf{p}' = \mathbf{L}^T\,\mathbf{p}, \tag{19}$$

and with $\mathbf{L}^{T-1}\mathbf{L}^{-1} = \mathbf{P}^{-1}$, the procedure can be rewritten in a form in which only $\mathbf{P}^{-1}$ is needed. The initial conditions are

$$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\,\mathbf{x}_0, \quad \mathbf{p}_0 = \mathbf{P}^{-1}\,\mathbf{r}_0. \tag{20}$$

The preconditioned conjugate gradient iterative procedure is then unchanged for equations (13) and (14), while the equations (12), (15) and (16) are modified into

$$\alpha_i = \mathbf{r}_i^T\,\mathbf{P}^{-1}\,\mathbf{r}_i/\mathbf{p}_i^T\,\mathbf{A}\,\mathbf{p}_i, \tag{21}$$

$$\beta_{i+1} = \mathbf{r}_{i+1}^T\,\mathbf{P}^{-1}\,\mathbf{r}_{i+1}/\mathbf{r}_i^T\,\mathbf{P}^{-1}\,\mathbf{r}_i, \tag{22}$$

$$\mathbf{p}_{i+1} = \mathbf{P}^{-1}\,\mathbf{r}_{i+1} + \beta_{i+1}\,\mathbf{p}_i. \tag{23}$$

Professor Niels Hansen is gratefully acknowledged for helpful discussions. We thank the French Ministry of Research and Technology for financial support to BG and RG.

References

Allen, F. H. (1986). *Acta Cryst.* B**42**, 515–522.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Coppens, P. (1967). *Science*, **158**, 1577–1579.

Flaig, R., Koritsanszky, T., Zobel, D. & Luger, P. (1998). *J. Am. Chem. Soc.* **120**, 2227–2238.

Fletcher, R. & Reeves, C. M. (1964). *Comput. J.* **7**, 149–154.

Guillot, B., Jelsch, C. & Lecomte, C. (2000). *Acta Cryst.* C**56**, 726–728.

Hamilton, W. C. (1964). *Statistics in Physical Science, Estimation Hypothesis Testing and Least Squares*, pp. 137–139. New York: Ronald Press.

Haneef, I., Moss, D. S., Stanford, M. J. & Borkakoti, N. (1985). *Acta Cryst.* A**41**, 426–433.

Hansen, N. K. & Coppens, P. (1978). *Acta Cryst.* A**34**, 909–921.

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

Hirshfeld, F. L. (1976). *Acta Cryst.* A**32**, 239–244.

Housset, D., Pichon-Pesme, V., Jelsch, C., Benabicha, F., Maierhofer, A., David, S., Fontecilla-Camps, J. C. & Lecomte, C. (2000). *Acta Cryst.* D**56**, 151–160.

Jelsch, C., Pichon-Pesme, V. & Lecomte, C. & Aubry, A. (1998). *Acta Cryst.* D**54**, 1306–1318.

Jelsch, C., Teeter, M. M., Lamzin, V., Pichon-Pesme, V., Blessing, R. H. & Lecomte, C. (2000). *Proc. Natl Acad. Sci. (USA)*, **97**, 3171–3176.

Konnert, J. H. (1976). *Acta Cryst.* A**32**, 614–617.

Konnert, J. H. & Hendrickson, W. A. (1980). *Acta Cryst.* A**36**, 344–350.

Lamour, V., Barth, P., Rogniaux, H., Poterszman, A., Howard, E., Mitschler, A., Vandorsselaer, A., Podjarny, A. & Moras, D. (1999). *Acta Cryst.* D**55**, 721–723.

Mitschler, A., Sanishvili, R., Joachimiak, A., Howard, E., Barth, P., Lamour, V., Guillot, B., Van Zandt, M., Sibley, E., Moras, D. & Podjarny, A. (2000). 19th European Crystallographic Meeting, 25–31 August 2000, Nancy. Poster s7.m0.p1.

Peres, N., Bouhkris, A., Souhassou, M., Gavoille, G. & Lecomte, C. (1999). *Acta Cryst.* A**55**, 1038–1048.

Pichon-Pesme, V., Lecomte, C. & Lachekar, H. (1995). *J. Phys. Chem.* **99**, 6242–6250.

Raymond, K. N. (1972). *Acta Cryst.* A**28**, 163–166.

Rosenfield, R. E. Jr, Trueblood, K. N. & Dunitz, J. D. (1978). *Acta Cryst.* A**34**, 828–829.

Sheldrick, G. M. & Schneider, T. (1997) *SHELXL*: *High-Resolution Refinement*, in *Methods in Enzymology*, Vol. 276, *Macromolecular Crystallography*, Part *B*, edited by C. W. Carter Jr & R. M. Sweet. New York: Academic Press.

Stanley, E. (1965). *Acta Cryst.* **19**, 1055–1056.

Templeton, D. H. (1999). *Acta Cryst.* A**55**, 695–699.

Tronrud, D. E. (1992). *Acta Cryst.* A**48**, 912–916.

Urzhumtsev, A. G. (1991). *Acta Cryst.* A**47**, 723–727.

Volkov, A., Abramov, Y., Coppens, P. & Gatti, C. (2000). *Acta Cryst.* A**56**, 332–339.

Volkov, A., Wu, G. & Coppens, P. (1999). *J. Synchrotron Rad.* **6**, 1007–1015.

Yamano, A., Heo, N. H. & Teeter, M. M. (1997). *J. Biol. Chem.* **272**, 9597–9600.