Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Methodology and applications of automatic electron-density map interpretation by six-dimensional rotational and translational search for molecular fragments

Frantisek Pavelcik,[a,b]* Jan Zelinka[a,c] and Zbyszek Otwinowski[a]

[a]The University of Texas, Southwestern Medical Center at Dallas, Department of Biochemistry, 5323 Harry Hines Boulevard, Dallas, Texas 75235, USA, [b]Department of Inorganic Chemistry, Faculty of Natural Sciences, 84215 Bratislava, Slovak Republic, and [c]Institute of Neuroimmunology, Slovak Academy of Sciences, Dubravska cesta, 84000 Bratislava, Slovak Republic

Correspondence e-mail: pavelcik@fns.uniba.sk

Received 10 September 2001
Accepted 15 November 2001

The electron density for a search fragment and for the crystal is expanded in the space of spherical harmonics Bessel functions. The fast rotation function is evaluated for each grid point to test if the fragment can be orientated there. For the best scoring points, the six-dimensional coordinates of the fragment are refined by the second-derivative block-diagonal procedure. The method is able to locate fragments precisely over a wide range of resolutions for structure types from small organic molecules to proteins.

1. Introduction

Electron density is easily interpreted if well phased structure-factor data are available to atomic resolution. A Gaussian or parabolic curve is used to fit peak shape (Rollett, 1965; Pavelcik, 1994). If the data are below this resolution or the phasing is poor, then 'bones' skeletonization and chicken-wire representation of the electron density are calculated. It is necessary to use computer graphics to interpret the electron density [*e.g. FRODO* (Jones, 1978), *O* (Jones & Kjeldgaard, 1997), *QUANTA* (Oldfield, 1996) and *XtalView* (McRee, 1999)]. Scene analysis is another method for map interpretation (Fortier & Glasgow, 1998). Map interpretation is a time-limiting step in the protein structure determination and is quite subjective for low-resolution maps. Its automation is an important step in the overall automation of the protein structure determination and is essential for the success of structural genomic projects.

Interpreting electron density by directly fitting molecular fragments by six-dimensional rotation and translation searches has been attempted many times, but so far no fully automatic method or computer program for this purpose has been developed. The template method is used in the program *O* to position secondary-structure fragments of 5–7 residues (Kleywegt & Jones, 1997). A phased translation function reviewed by Bentley (1997) was designed to position a correctly oriented model in the unit cell. Cowtan (1998) has developed a fast Fourier based method to locate molecular fragments; this method is implemented in the program *FFFEAR*. Recently, Friedman (1999) described the interconversion between the spherical harmonics Bessel representation and the Fourier representation of objects in three-dimensional space. This is used to perform an exhaustive grid search in six-dimensional relative rotational and translational space of empirical energy functions and to estimate the contents of the asymmetric unit in sparsely packed non-centrosymmetric crystalline arrays.

We have used the spherical harmonics Bessel expansion of the fragment density and that of the electron density about

© 2002 International Union of Crystallography
Printed in Denmark – all rights reserved

each grid point of the asymmetric unit as input to the fast rotation function (Crowther, 1972) for positioning molecular fragments. Subsequent refinement using second derivatives gives precise solutions. Satisfactory test results for small fragments with rigid conformations of 5–20 atoms have been obtained using data with a resolution as low as 3–4 Å and with a phase error of up to 90°. This method should be easy to apply to the automatic interpretation of electron-density maps and further results will be presented in a subsequent paper. A database of molecular fragments in PDB-like format is being built for this purpose.

2. Theory

2.1. Spherical harmonics Bessel expansion of electron density

A rotation function is used to correlate a spherical volume of a density function with another similar function. Since we are dealing with the rotation of approximated spherical volumes, it is natural to expand any ordinary function $\rho(r)$ using spherical harmonics around the origin of the coordinate system,

$$\rho(x, y, z) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} c_{lm}(r) Y_l^m(\theta, \varphi), \tag{1}$$

where (r, θ, φ) are polar coordinates of the vector $\mathbf{r} = (x, y, z)$ and $c_{lm}(r)$ is the radial function (not dependent on orientation). $Y_l^m(\theta, \varphi)$ are spherical harmonics defined according to Wigner (1959). For $m > 0$,

$$Y_l^m(\theta, \varphi) = s\left[\frac{(2l+1)(l-m)!}{4\pi(l+m)!}\right]^{1/2} P_l^m(\cos\theta)\exp(im\varphi), \tag{2}$$

where $P_l^m(\cos\theta)$ is the associated Legendre polynomial function and s is the phase factor, which is either 1 or $(-1)^m$ depending on convention used.

Spherical harmonics constitute an orthogonal complete system defined on a spherical surface with unit radius. On such a sphere whose surface element is $\mathrm{d}\sigma = \sin(\theta)\,\mathrm{d}\theta\mathrm{d}\varphi$, the orthogonal normalization relation is

$$\int_{4\pi} Y_l^{m*}(\theta, \varphi) Y_l^{m'}(\theta, \varphi)\,\mathrm{d}\sigma = \int_{\varphi=0}^{2\pi}\int_{\theta=0}^{\pi} Y_l^{m*}(\theta, \varphi) Y_{l'}^{m'}(\theta, \varphi)\sin\theta\,\mathrm{d}\theta\mathrm{d}\varphi = \delta_{ll'}\delta_{mm'}. \tag{3}$$

For known electron density, the radial functions $c_{lm}(r)$ can be determined utilizing the orthonormality condition

$$c_{lm}(r) = \int_{\sigma} \rho(r) Y_l^{m*}(\theta, \varphi)\,\mathrm{d}\sigma, \tag{4}$$

where σ is a sphere of radius r (Giacovazzo, 1998). There are many possible radial functions and it is not obvious which one is optimal. A typical choice is the spherical Bessel function (Crowther, 1972),

$$c_{lm}(r) = \sum_{n=1}^{\infty} a_{nlm} g_l(k_{nl} r), \tag{5}$$

where $g_l(k_{nl}r)$ is the normalized spherical Bessel function of order l,

$$g_l(k_{nl}r) = \frac{2^{1/2}}{a^{3/2} j_{l-1}(k_{nl}a)} j_l(k_{nl}r). \tag{6}$$

n represents various roots of the 'oscillatory' spherical Bessel function $j_l(x)$. k_{nl} is such as that $k_{nl}a = x_n$, where x_n are zero Bessel values $j_l(x_n) = 0$, $(n = 1, 2, \ldots, n_{\max})$ and a is the radius of the chosen sphere of expansion. Normalized spherical Bessel functions again form an orthonormal basis for expansion with respect to n,

$$\int_0^a g_l(k_{nl}r) g_l(k_{n'l}r) r^2\,\mathrm{d}r = \delta_{nn'}. \tag{7}$$

The complete orthonormal spherical harmonics Bessel functions used for the expansion of electron density are then

$$S_{nlm}(r, \theta, \varphi) = g_l(k_{nl}r) Y_l^m(\theta, \varphi). \tag{8}$$

The electron density can be expressed as

$$\rho(r, \theta, \varphi) = \sum_{n=1}^{\infty}\sum_{l=0}^{\infty}\sum_{m=-l}^{l} a_{nlm} S_{nlm}(r, \theta, \varphi). \tag{9}$$

The coefficients of expansion are

$$a_{nlm} = \int_{\varphi=0}^{2\pi}\int_{\theta=0}^{\pi}\int_{r=0}^{a} \rho(r, \theta, \varphi) S_{nlm}^* r^2 \sin\theta\,\mathrm{d}\theta\mathrm{d}\varphi\mathrm{d}r. \tag{10}$$

The coefficients have an important normalization property, which can be used to study the convergence of the expansion,

$$\sum_{n=1}^{\infty}\sum_{l=0}^{\infty}\sum_{m=-l}^{l} |a_{nlm}|^2 = \int_{r<a} \rho^2(r)\,\mathrm{d}V = \text{const}. \tag{11}$$

2.2. Calculation of coefficients in reciprocal space

Fourier representation of the electron density has to be included into these formulae in order to use the available phase information. Direct calculation from the electron density would require numerical integration. The basic formulae are given by Friedman (1999), but because we found it more convenient to use their modified form, the full derivation is presented here:

$$\rho(\mathbf{r}) = (1/V)\sum_h F_h \exp[-2\pi i(\mathbf{h}\cdot\mathbf{r})]. \tag{12}$$

F_h is the structure factor, $F_h = A_h + iB_h$. (10) becomes

$$a_{nlm} = \frac{1}{V}\sum_h F_h \int_{r<a} \exp(-2\pi i\mathbf{h}\cdot\mathbf{r}) g_l(k_{nl}r) \times Y_l^{m*}(\theta, \varphi) r^2 \sin\theta\,\mathrm{d}\theta\mathrm{d}r\mathrm{d}\varphi. \tag{13}$$

Any plane function may be expanded (see Landau & Lifschitz, 1972) according to

$$\exp(-2\pi i\mathbf{h}\cdot\mathbf{r}) = 4\pi\sum_{l=0}^{\infty}\sum_{m=-l}^{l} (-i)^l j_l(2\pi hr) Y_l^m(\theta, \varphi) Y_l^{m*}(\theta_h\varphi_h), \tag{14}$$

where (r, θ, φ) are polar coordinates of vector $\mathbf{r}$ in direct space and (h, θ_h, φ_h) are polar coordinates of vector $\mathbf{h}$ in reciprocal space. Combining (13) and (14) and utilizing orthonormalization conditions (for details, see Giacovazzo, 1998; Dodson,

1985) the integral (13) is factorized into radial and angular parts, with the radial part

$$R_{nl}(h, a) = \frac{4\pi}{V}\frac{(2a)^{1/2}k_{nl}j_l(2\pi ha)}{4\pi^2h^2 - k_{nl}^2}. \quad (15)$$

The final formula for electron-density expansion coefficients is

$$a_{nlm} = \sum_h F_h(-i)^l Y_l^{m*}(\theta_h, \varphi_h)R_{nl}(h, a). \quad (16)$$

2.3. Three-dimensional expansion in grid points of the unit cell

(16) represents the expansion of electron density at the origin within a sphere of radius a. This equation can be generalized to give the expansion at any point of the unit cell by applying an appropriate coordinate shift to make the point (x, y, z) of the electron-density map the origin for the spherical harmonics expansion. This shift can be viewed as a change of structure-factor phases,

$$a_{nlm}(x, y, z) = \textstyle\sum_h F_h(-i)^l Y_l^{m*}(\theta_h, \varphi_h)R_{nl}(h, a) \times \exp[-2\pi(hx + ky + lz)]. \quad (17)$$

Coefficients a_{nml} can be calculated conveniently by FFT on a selected grid using complex-to-complex Fourier transforms. However, because for given l there are $2l + 1$ complex angular functions and because these can be constructed from the same number of real components $(a^R, a^I)Y = \mathrm{Re}(Y) + i\mathrm{Im}(Y) = a^R + ia^I$, then the complex-to-complex FFT can be replaced by the same number of complex-to-real FFT, with a saving in the computer time. F_h need to be expanded only for one half of the reciprocal sphere in this case. Friedman (1999) formulae have been modified,

$$\begin{aligned} a_{nlm} &= a^R_{nlm} + ia^I_{nlm} \\ &= \sum_h F_h(-i)^l[\mathrm{Re}(Y_l^{m*}) + i\mathrm{Im}(Y_l^{m*})]R_{nl}(h, a)\exp(-2\pi i\mathbf{h}\cdot\mathbf{x}) \\ a^R_{nlm} &= \sum_h F_h(-i)^l\mathrm{Re}(Y_l^{m*})R_{nl}(h, a)\exp(-2\pi i\mathbf{h}\cdot\mathbf{x}) \\ a^I_{nlm} &= \sum_h F_h(-i)^l\mathrm{Im}(Y_l^{m*})R_{nl}(h, a)\exp(-2\pi i\mathbf{h}\cdot\mathbf{x}). \end{aligned} \quad (18)$$

Because spherical harmonics have centrosymmetric or antisymmetric properties with respect to reciprocal vector inversion, the Fourier coefficients are not in general Hermitian. The term $(-i)^l$ directly included into the Fourier coefficients removes the problem of Hermiticity of the Fourier coefficients.

2.4. Fragment expansion

Electron density of a fragment is expanded in complex conjugated basis functions in order to utilize the orthogonality relation (3). Two methods for calculating coefficients of fragment expansion were developed. In both methods, the geometric center of the fragment is placed at the origin of the coordinate system and no phase shift is applied for expansion,

$$\rho_f(r, \theta, \varphi) = \sum_n\sum_l\sum_m b^*_{nlm}g_l(k_{nl}r)Y_l^{m*}(\theta, \varphi), \quad (19)$$

$$b_{nlm} = \int_{\varphi=0}^{2\pi}\int_{\theta=0}^{\pi}\int_{r=0}^{a}\rho_f(r, \theta, \varphi)g_l(k_{nl}r)Y_l^m(\theta, \varphi)r^2\sin\theta\,\mathrm{d}\theta\mathrm{d}\varphi\mathrm{d}r. \quad (20)$$

2.4.1. Ec fragments. The method is based on normalized calculated structure factors and (16) is used for calculation. (i) This representation has been developed for fragments which are not well represented by a small set of uniquely defined atom coordinates. This may involve fragments with disordered conformations, various averaged fragments and fragments involving an averaged environment with the potential for a large number of atoms. These fragment data are stored as precalculated structure factors in an artificial $P1$ cell. (ii) Alternatively, the fragment represented by atomic coordinates is positioned in a unit cell of the same size as the analyzed electron density and structure factors are calculated for the same set of reflection data as those used for electron-density expansion. These fragment data are stored in a coordinate file.

2.4.2. Dirac fragments. The method is based on atomic scatters. ρ_j is the shape function,

$$\rho(\mathbf{r}) = \sum_{j=1}^{N} Z_j\rho_j(\mathbf{r} - \mathbf{r}_j). \quad (21)$$

Numerical integration may be used to calculate the expansion coefficients. The shape of the electron density of the fragment can be represented by a Dirac δ function. A simple formula was derived,

$$b_{nlm} = \sum_{j=1}^{N} Z_j g_l(k_{nl}r)Y_l^m(\theta_j, \varphi_j). \quad (22)$$

2.5. Normalization and sorting of coefficients

The coefficients for electron-density expansion are calculated for all grid points and specified n, l, m in one FFT step. For the purpose of rotation and translation searches, they have to be rearranged for each particular grid point and all n, l, m. Furthermore, using normalized coefficients allows us to use as a target function the correlation coefficient between expansions instead of a simple overlap function. Rearrangement and normalization are in fact performed within one computing step,

$$a^N_{nlm} = a_{nlm}\Big/\left(\sum_{nlm}|a_{nlm}|^2\right)^{1/2}. \quad (23)$$

The normalization factor (denominator) represents the mean electron density in the volume of the radius of expansion. As it depends only on n, l, m it is calculated for all fragment/electron-density expansions only once. The normalization coefficient can be used also in envelope modelling or skeletonization.

2.6. Translation search

The expansion of electron density at each grid point of the asymmetric unit is equivalent to the translation search.

2.7. Rotation search

In the rotation search the particular grid point is formally placed at the origin of the unit cell and fragment rotation is performed by Crowther's fast rotation function. For a rotation Ω defined as Eulerian angles (α, β, γ), rotated spherical harmonics of degree l may be expressed as a weighted sum of the $(2l + 1)$ unrotated harmonics of degree l (Wigner, 1959; Altmann & Bradley, 1963),

$$\hat{\Omega}(\alpha, \beta, \gamma) Y_l^m(\theta, \varphi) = \sum_{m'=-l}^{l} Y_l^{m'}(\theta, \varphi) D^l_{m'm}(\alpha, \beta, \gamma), \qquad (24)$$

where the $D^l_{m'm}(\alpha, \beta, \gamma)$ matrix has the form

$$D^l_{m'm}(\alpha, \beta, \gamma) = \exp(m'\gamma) d^l_{m'm}(\beta) \exp(m\alpha). \qquad (25)$$

We prefer the original definition of D matrix given by Wigner (1959). Different definitions can be found in the literature. The Hermitian conjugate matrix is used with molecular replacement (Dodson, 1985; Navaza, 1990) or the complex conjugate of Giacovazzo (1998). The actual values of the d matrix (and its derivatives, see below) depend on definition of the spherical harmonics functions, the Eulerian angle convention and on the order of matrix multiplication. The elements of the d matrix are conveniently calculated by the recurrence formula of Navaza (1990) or by the original formulae of Wigner (1959).

Once the electron densities of the crystal $\rho(r, \psi, \varphi)$ and the electron density of the fragment $\rho_f(r, \theta, \varphi)$ are expanded within the spherical volume $r < a$ using (9) and (19), than the relative rotation overlap as the fragment is rotated can be defined as

$$R(\Omega) = \int\int\int \rho(r, \theta, \varphi) \hat{\Omega}(\alpha, \beta, \gamma) \rho_f(r, \theta, \varphi) r^2 \sin\theta \mathrm{d}r\mathrm{d}\theta\mathrm{d}\varphi. \qquad (26)$$

$\hat{\Omega}(\alpha, \beta, \gamma)\rho$ is the rotated version of ρ resulting from the rotation by (α, β, γ). Combining (24)–(26), (9) and (19) and using the orthogonality relation (see Dodson, 1985) for the expansion function (3) reduces this to

$$R(\Omega) = \sum_n \sum_l \sum_m \sum_{m'} a_{nlm'} b^*_{nlm} D^{l*}_{m'm}(\Omega). \qquad (27)$$

The radial summation n may be performed independently of the rotation Ω, so finally the rotations can be expressed as

$$R(\alpha, \beta, \gamma) = \sum_m \sum_{m'} \left[\sum_l c_{lm'm} d^l_{m'm}(\beta) \right] \exp[-i(m\alpha + m'\gamma)]. \qquad (28)$$

Owing to the form of the rotation coefficient, two of the three summations in (28) occur as Fourier series. For the purpose of FFT it is better to define rotation in cycles rather than in radians

$$R^\beta(u, v) = \sum_m \sum_{m'} f_{mm'} \exp[-i2\pi(mu + m'v)], \qquad (29)$$

where $u = \alpha/2\pi$ and $v = \gamma/2\pi$. Defining rotation on scale 0–1 leads to a formula very similar to that used in X-ray crystallography.

There is a strong correlation between α and γ angles if the β angle is close to 0 or π. This correlation is also reflected in $f_{mm'}$ terms, which are not all independent. The two-dimensional R^β map effectively turns into a diagonal line for $\beta = 0$ or π. To achieve more uniform sampling of the angular space, the rotation search can alternatively be performed on transformed coordinates (Lattman, 1972). Transformed coordinates and indices for the Fourier coefficient f_{pq} are defined by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}$$

$$(p \quad q) = (m \quad m') \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \qquad (30)$$

The transformation leaves the scalar product $mu + m'v = px + qy$ unchanged,

$$R^\beta(x, y) = \sum_p \sum_q f_{pq} \exp[-i2\pi(px + qy)]. \qquad (31)$$

Whereas the periodical space of α, γ angles is represented by primitive square plane cell, the space of x, y angles is a centered rectangular cell (square for $\beta = \pi/2$), which for the case of $\beta = 0$ or $\beta = \pi$ reduces to a single line. The number of grid points in each direction can be calculated as a function of the β angle,

$$N_x = 2^{1/2} N \cos\beta/2,$$
$$N_y = 2^{1/2} N \sin\beta/2, \qquad (32)$$

where N would be the number of grids for the (α, γ) square cell.

In addition to the simple product function $R(\alpha, \beta, \gamma)$ we have chosen the correlation coefficient between the electron density and the electron density of the fragment as the target function. The correlation coefficient better reflects 'the shape of the electron density' than simple overlap function, but the information on absolute value of the overlap (product of two electron densities) is lost,

$$\mathrm{CC} = \frac{\langle \rho_1 \rho_2 \rangle}{\langle \rho_1^2 \rangle^{1/2} \langle \rho_2^2 \rangle^{1/2}}. \qquad (33)$$

In this case, in (24)–(29) the coefficients of expansion should be replaced with normalized coefficients.

The rotation function is calculated for each grid point. The result of the calculation is a six-dimensional map of overlap or correlation coefficients, which is analyzed by a six-dimensional peak-picking procedure.

3. Refinement of fragment position and orientation

The correlation coefficient and the overlap function are functions of six parameters $\mathrm{CC}(x, y, z, \alpha, \beta, \gamma)$. The translation search is given by expanding the electron density at fixed grid coordinates. The fragment is rotated about each grid point. For this purpose, there are many sets of electron-density

coefficients and only one set of coefficients for the fragment. The refinement of angular position is performed in more-or-less the opposite way. There is one set of electron-density expansion and several sets of fragment expansions. Interpolated (using peak picking) angles, positions and the nearest electron-density expansion grid point are used as the starting point. The advantage of using interpolated coordinates is that the starting positions are near the maximum of the overlap function (the second derivatives are negative). For this reason, the second-derivative method was selected for optimization. For simplicity, correlation among certain rotation and translation parameters were neglected.

3.1. Refinement of fragment orientation only

The electron-density expansion for the grid point which is the nearest to the peak is used. Derivatives of overlap function with respect to angles are calculated analytically. Analytical derivatives for Euler angles (in cycles) are given by

$$\frac{\partial}{\partial u}\mathrm{CC}(u,\beta,v) = \sum_m \sum_{m'} -2\pi i m \left[\sum_l c_{lm'm} d^l_{m'm}(\beta)\right] \times \exp[-i2\pi(m'v + mu)], \quad (34)$$

$$\frac{\partial^2}{\partial u^2}\mathrm{CC}(u,\beta,v) = \sum_m \sum_{m'} -4\pi^2 m^2 \left[\sum_l c_{lm'm} d^l_{m'm}(\beta)\right] \times \exp[-i2\pi(m'v + mu)], \quad (35)$$

$$\frac{\partial^2}{\partial u \partial v}\mathrm{CC}(u,\beta,v) = \sum_m \sum_{m'} -4\pi^2 m m' \left[\sum_l c_{lm'm} d^l_{m'm}(\beta)\right] \times \exp[-i2\pi(m'v + mu)]. \quad (36)$$

For simplicity, correlation between (u, v) and β (in radians) is ignored,

$$\frac{\partial}{\partial \beta}\mathrm{CC}(u,\beta,v) = \sum_m \sum_{m'} \left[\sum_l c_{lm'm} \frac{\partial}{\partial \beta} d^l_{m'm}(\beta)\right] \times \exp[-i2\pi(m'v + mu)], \quad (37)$$

$$\frac{\partial^2}{\partial \beta^2}\mathrm{CC}(u,\beta,v) = \sum_m \sum_{m'} \left[\sum_l c_{lm'm} \frac{\partial^2}{\partial \beta^2} d^l_{m'm}(\beta)\right] \times \exp[-i2\pi(m'v + mu)]. \quad (38)$$

Shifts are calculated from a Taylor expansion neglecting higher than quadratic terms. The 2×2 matrix-inversion method is used for α and γ and the diagonal approximation is used for β. Analytical derivatives of the $d(\beta)$ matrix are calculated using the same recurrence formulae as for calculation of the values of the d matrix (Navaza, 1990); for the special values $\beta = 0$ and $\beta = 180^\circ$ derivatives of the original Wigner formula (Wigner, 1959) are used.

3.2. Refinement of Dirac fragments

In this procedure the fragment is moved within the sphere of expansion. The coordinates of the nearest grid point (fractional coordinates) are determined and electron-density expansion coefficients are selected. The rotation matrix for the fragment is calculated from pre-refined Euler angles ($R_{\rm EU}$ matrix). The differences between interpolated coordinates and the nearest grid-point coordinates are transformed to Cartesian space (starting $\delta\mathbf{x}$ vector). Refinement of fragment position and orientation is carried out by a small shift in the origin of the fragment and by small rotations about each Cartesian coordinate axis. Because these are small, they can be considered to be independent. For each small movement of the fragment new expansion coefficients are calculated,

$$\mathbf{x}' = R \cdot R_{\rm EU}\mathbf{x} + \delta\mathbf{x}. \quad (39)$$

$R = R_x, R_y, R_z$ are small rotations about the Cartesian axes by angles φ_x, φ_y, φ_z. Derivatives of $\mathrm{CC}(x, y, z, \varphi_x, \varphi_y, \varphi_z)$ are calculated numerically. CC is estimated at three points $\zeta - \Delta$, ζ and $\zeta + \Delta$, where ζ is the parameter to be refined and Δ is a step. The step was fixed at one quarter of the grid step (for both positional and angular parameters). The first and second derivatives are calculated numerically by fitting a parabolic curve. Each iteration is started with $R = I$, where I is the unit matrix. The positional parameters are refined first. Each rotation angle is then optimized and the R matrix is upgraded. After each iteration, the $R_{\rm EU}$ matrix is multiplied by the R matrix. As the refinement converges the R matrix approaches the identity matrix.

3.3. Analytical derivative method

A fast sophisticated procedure based on analytical derivatives is necessary for refinement of more general fragments (*e.g.* those whose expansion is based on calculated structure factors). Fragment positions can be refined by small shifts of the center of the electron-density expansion. These shifts can be obtained by calculating derivatives of the target function (CC) with respect to the coordinates,

$$\frac{\partial \mathrm{CC}}{\partial x} = \sum_n \sum_l \sum_m \sum_{m'} \frac{\partial a_{nlm}}{\partial x} b^*_{nlm} D_{mm'}(u,\beta,v), \quad (40)$$

$$\frac{\partial^2 \mathrm{CC}}{\partial x^2} = \sum_n \sum_l \sum_m \sum_{m'} \frac{\partial^2 a_{nlm}}{\partial x^2} b^*_{nlm} D_{mm'}(u,\beta,v), \quad (41)$$

$$\frac{\partial a_{nlm}}{\partial x} = -2\pi i \sum_h h F_h (-i)^l Y^{m*}_l(\theta_h,\varphi_h) R_{nl}(h) \times \exp[-2\pi(hx + ky + lz)], \quad (42)$$

$$\frac{\partial^2 a_{blm}}{\partial x^2} = -4\pi^2 \sum_h h^2 F_h (-i)^l Y^{m*}_l(\theta_h,\varphi_h) R_{nl}(h) \times \exp[-2\pi(hx + ky + lz)], \quad (43)$$

$$\frac{\partial^2 a_{nlm}}{\partial xy} = -4\pi^2 \sum_h hk F_h (-i)^l Y^{m*}_l(\theta_h,\varphi_h) R_{nl}(h) \times \exp[-2\pi(hx + ky + lz)] \quad (44)$$

and similarly for other coordinates. Calculation of analytical derivatives requires summation over all reflections. Euler angles are refined as previously described. Shifts are calculated from a Taylor expansion, neglecting higher than quadratic terms. The 3×3 matrix-inversion method is used for the

coordinates, the 2×2 matrix-inversion method for α and γ and the diagonal approximation for β. The electron-density expansion is recalculated at new (x, y, z) points after each iteration.

4. Test calculations and discussion

Test calculations were carried out on structures of various symmetry and complexity, ranging from small organic structures to proteins. The main objectives of these tests were to find the limitations of the method, the accuracy with which fragments could be fitted and to determine the best default parameters. We present here only selected results that highlight some of the method's properties. Calculated phases from refined structures and observed structure-factor amplitudes (F_{obs}, φ_{calc}) were used if not specified explicitly. The grid spacing for FFT was 0.4 Å.

Each fragment was taken from the refined structure conserving its original conformation. This means that one of solutions is expected to be unique. The fit was characterized by

$$\mathrm{ERR} = \left(\sum d^2/n\right)^{1/2}. \quad (45)$$

d is the difference between the atom position in the crystal structure and the refined position from the six-dimensional search. Another criterion was the correlation coefficient between calculated electron density at atomic positions and the atomic numbers (Z) of fragment atoms.

$$\mathrm{CCZ} = \frac{\langle Z\rho \rangle}{\langle Z^2 \rangle^{1/2} \langle \rho^2 \rangle^{1/2}}. \quad (46)$$

This criterion was sensitive to missing atoms. The exact fit means that all positioned fragment atoms are at reasonable distances (less than 0.2 Å in general) from crystal structure atoms, irrespective of atomic type. Dirac fragments were used in calculations if not explicitly specified otherwise.

Calculations were carried out using a 1 GHz Pentium III with 512 Mb of RAM. The CPU times ranges from 2 min for a small structure to 8 h for the largest protein. The disk space necessary to store the coefficients of the electron-density expansion for the protein structure is of the order of several gigabytes.

4.1. SIGI

$P\bar{1}$, $C_7H_{11}NO_3$, $Z = 2$, $a = 6.652$, $b = 7.758$, $c = 8.147$ Å, $\alpha = 73.09$, $\beta = 75.99$, $\gamma = 68.40^\circ$, resolution 0.84–4.70 Å (test distributed with *SHELXL*93). This structure was used to test trigonal and centrosymmetric space group, despite the fact that the expected applications are in non-centrosymmetric protein structures. The structure consists of a substituted [═O, $-CH_2-OH$, $-C(=NH)-CH_3$] tetrahydrofuran ring. A five-membered ring with radius 1.30 Å was selected as the search fragment. The parameters for expansion were $n(\mathrm{max}) = 7$, $l(\mathrm{max}) = 7$, $a = 2.4$ Å. After the refinement, the five top peaks gave CC values of 0.623, 0.682, 0.634, 0.654 and 0.629, with related ERRs of 0.120, 0.048, 0.086, 0.113 and 0.111, respectively. CCZs were in the range 0.99–0.95. All positions found were ring atoms with different numbering. The next peak gave the considerably smaller CC of 0.48 and a CCZ of 0.875. The correct and false solutions were well separated.

Table 1
Influence of phase quality on ability of the method to find the fragment position for RKSA1.

P is the randomization parameter ($P = 0$ for original phases, $P = 1$ for random phases). RR is the R factor resulting from randomization; DPHI is the phase error in ° (for definition, see *Appendix A*). CC and ERR are defined in the main text.

P	RR	DPHI	CC	ERR
0.0	0.000	0.0	0.818	0.053
0.1	0.104	4.9	0.817	0.054
0.2	0.204	11.2	0.814	0.056
0.3	0.295	21.5	0.806	0.059
0.4	0.372	37.1	0.789	0.062
0.5	0.427	53.6	0.754	0.070
0.6	0.458	68.2	0.670	0.093
0.7	0.468	80.8	0.505	0.118
0.75	0.466	86.5	0.387	0.125
0.8	0.462	91.7	—	—

4.2. RKSA1

$P2_1$, $C_{13}H_{20}N_2O_5$, $Z = 2$, $a = 8.73$, $b = 8.59$, $c = 9.65$ Å, $\beta = 104.6^\circ$, resolution 0.84–9.66 Å (Steiner *et al.*, 1998). This sugar structure consists of 20 atoms and contains substituted (3Me-, CH_3O-, NC-, MeCONH-) fused six-membered and five-membered rings. A large fragment of 13 atoms with radius 3.1 Å was used for the search. At first, good phases were used for electron-density expansion. The radius for electron-density expansion was chosen to be 3.2 Å. The refinement for the top peaks gave a CC of 0.82, with a corresponding ERR of 0.053 Å, meaning an exact fit. The main conclusion from this test was that for a large fragment with a sphere of expansion similar to the radius of fragment and an electron-density expansion calculated with good phases, an exact fit could be achieved. The mathematical model seems to be good and the use of Bessel functions for radial expansion is appropriate. Structure RKSA1 was then used for a preliminary test of the relationship between the phase quality and the ability of the procedure to move the fragment into the correct position. Data sets with different phase qualities were obtained by randomization using the formula given in *Appendix A*. The results are given in Table 1. The procedure failed to position the fragment if the electron density was calculated with a mean phase error of 90°.

4.3. RKSA5

$P6_2$, $C_{13}H_{20}N_2O_5$, $Z = 6$, $a = 15.61$, $b = 15.61$, $c = 10.63$ Å, resolution 0.90–13.52 Å (Koos *et al.*, 2000). This sugar structure is an isomer of RKSA1 with a different conformation and with a different crystal symmetry. The expansion parameters were the same as in the previous case. The test focused on the influence of the data resolution on small-molecule positioning. Results are given in Table 2. It can be seen that 50 reflections were enough to place the fragment in the correct position.

Table 2
Influence of resolution on fragment positioning for RKSA5.

Res is resolution, N_{ref} is related number of reflections used. The other symbols are defined in the text.

Res	N_{ref}	CC	CCZ	ERR
0.90	5111	0.885	0.9605	0.020
1.00	3887	0.844	0.9701	0.028
1.20	2219	0.707	0.9918	0.043
1.40	1385	0.591	0.9876	0.054
1.60	932	0.528	0.9866	0.103
1.80	646	0.472	0.9913	0.143
2.00	478	0.438	0.9899	0.148
2.20	349	0.377	0.9918	0.159
2.40	277	0.350	0.9910	0.166
2.60	208	0.346	0.9613	0.293
2.80	178	0.327	0.9614	0.264
3.00	142	0.329	0.9173	0.485
3.20	121	0.327	0.9204	0.579
3.40	106	0.323	0.9329	0.552
3.60	87	0.293	0.8915	0.535
3.80	66	0.277	0.9064	0.777
4.00	63	0.227	0.9049	0.587
4.20	51	0.227	0.8425	0.617
4.40	45	0.260	0.9255	0.646
4.60	42	0.235 (no solution)	—	—

There were fluctuations in the calculated parameters when the number of reflections was less than 100.

4.4. MGHEX

$P3_1$, $Z = 3$, $C_{56}H_{80}Cl_2MgN_{16}O_2$, Mg(*cyclo*-Gly-Pro-Pro-Gly-Pro-Pro)$_2$(ClO_4)$_2$.4CH_3CN, a = 15.74, c = 24.00 Å, resolution 0.93–13.63 Å (Karle & Karle, 1981). Only one radius, a = 3.5 Å, with parameters n(max) = 7, l(max) = 7 was used for expansion. The aim of the test was to investigate the sensitivity of the method to the size of the fitted fragment *versus* the radius of expansion. The following fragments were selected from the crystal structure: Gly1, Gly2, Pro1, Pro2, Pro3, Pro4, GlyPro1, GlyPro2 and ProPro1. The search was carried out for all symmetry-equivalent positions.

The property of the trigonal system where the grid space is deformed and the symmetry-equivalent positions are numerically interpolated to close but different values with corresponding different CCs was utilized in tests. Fragments at different equivalent positions were refined to the practically the same value of CC and to the same differences in crystal atom positions (MGHEX and also RKSA5). The refinement can compensate for some errors introduced by grid sampling and interpolation. Four-atom Gly1 and Gly2 fragments were fitted to various groups in the molecule. There were 33 exact fits with ERR of 0.03–0.21 Å with CC in the range 0.60–0.50 and CCZ in the range 0.996–0.948. There was no clear distinction between correct and false peaks in the sorted peak list. CCZ was a better criterion for selection of correct peaks than CC. A simple overlap function produced many false peaks involving Mg atoms. The seven-atom fragments Pro1, Pro2, Pro3 and Pro4 were fitted into 3–5 different proline sites in the molecule with a CC in the range 0.77–0.64, distinguishable from false fittings. CCZ was in the range 0.99–0.91 and ERRs were in the range 0.02–0.15 Å. 11-atom GlyPro1 and GlyPro2 fragments were large enough to dominate on the peak list, with CC in the range 0.84–0.71. GlyPro1 was fitted in four different positions and GlyPro2 in only one position. The 14-atom ProPro1 fragment was fitted to only one position with ERR = 0.05. The conclusion can be drawn that smaller fragments are not easily distinguishable from false solutions if the radius of expansion is too large in comparison with the fragment radius.

4.5. 1pen (conotoxin)

$P2_1$, $Z = 2$, a = 15.0, b = 19.8, c = 16.5 Å, β = 113.4° (H-Gly Cys-Cys-Ser-Leu-Pro-Pro-Cys-Ala-Ala-Asn-Asn-Pro-Asp-Tyr-Cys-OH; Hu *et al.*, 1996), resolution 1.10–15.14 Å. As a test, single amino-acid residue fragments were used. The name of the search fragment is composed of the amino-acid code and the residue sequence number. The structure factors were calculated by automatic structure solution and refinement (Pavelcik, 1998), rather than from the PDB coordinates, in order to further simulate experimental conditions. A few atoms were missing and some atomic types were wrong. The 1pen fragments were divided into three groups with respect to their radii and expanded with parameters a = 2, 3 and 4 Å, l(max) = 7, n(max) = 7. 100 top peak positions from the six-dimensional search were refined.

4.5.1. G01 fragment. The top 62 peaks with CCZ > 0.90 were fitted to various atom groups with similar stereochemistry (not only to the N—C—C—O sequence). Only one of them was false (No. 56 based on CCZ). CCs were in the range 0.46–0.34. These fits were quite accurate, the maximum difference from refined atomic positions being in general about 0.15 Å (maximum ERR = 0.23 Å). The values of CC for fits with one atom missing were in the range 0.40–0.35. There was no correlation between the ERR and the CC for well fitted fragments. CCZ was again a better criterion than CC to distinguish between correct and false solutions. There was also no correlation with the difference in atomic types. The refined CC for position related to the crystal position of G01 (from which the fragment coordinates had been derived) was only 0.397. For this kind of fragment additional criteria for analysis should be developed.

4.5.2. A09, A10 fragments. CC = 0.42–0.30. The number of exact fits was 11 for A09 and 18 for A10. Again, the original fragment positions were not the highest on the list, but the presence of an extra asymmetric atom in comparison to glycine was the significant differentiation character. The fragment should be asymmetric for a unique search.

4.5.3. S04. The original fragment was at the top of the list with CC = 0.53. The next peaks were considerably worse. There was probably no group with similar conformation in this structure.

4.5.4. C02, C03, C08, C16 fragments. Cysteine groups were fitted with a highest of CC 0.67–0.53 in pairs (C02, C08), (C03, C16) of similar cysteine conformations. Some peaks with exact fit (but no S atoms) had a considerably lower CC. The heavy atom was the significant marker.

Table 3
Influence of data resolution for Y44 fitting in crambin.

Res	N_{ref}	CC	CCZ	ERR
0.83	37099	0.8497	0.9705	0.0728
1.00	28145	0.8255	0.9852	0.0709
1.40	10648	0.5116	0.9934	0.1018
1.80	5107	0.4072	0.9895	0.1302
2.20	2785	0.2949	0.9813	0.2175
2.60	1695	0.2655	0.9914	0.3255
3.00	1078	0.2394	0.9704	0.4404
3.40	723	0.2364	0.9490	No solution

4.5.5. L05, N12, D14 fragments. Only one exact fit was found for each fragment, corresponding to the peak with the highest CC. The highest/next CCs were 0.57/0.51, 0.53/0.43 and 0.60/0.49. Correct peaks were distinguishable from the next lowest peaks. For the N11 fragment the correct peak was No. 6, as two atoms of this fragment were not used to calculate structure factors.

4.5.6. P06, P07, P13 fragments. Only unique fits (0.53/0.41, 0.51/0.39 and 0.502/0.477) were found for proline residues.

4.5.7. Y15 fragment. A unique fit with CC = 0.680/0.563 was dominant. Only five peaks were found in the peak-picking procedure above the given threshold.

4.6. 1ab1 (crambin)

$P2_1$, $a = 40.763$, $b = 18.492$, $c = 22.333$ Å, $\beta = 90.61^\circ$, resolution 0.83–40.76 Å, Thr-Thr-Cys-Cys-Pro-Ser-Ile-Val-Ala-Arg-Ser-Asn-Phe-Asn-Val-Cys-Arg-Leu-Pro-Gly-Thr-Ser-Glu-Ala-Ile-Cys-Ala-Thr-Tyr-Thr-Gly-Cys-Ile-Ile-Ile-Pro-Gly-Ala-Thr-Cys-Pro-Gly-Asp-Tyr-Ala-Asn (Teeter *et al.*, 1993).

Crambin was used to test the influence of data resolution on the ability of the method to reveal atomic fragments for protein. Y44 was used as a test fragment, with expansion parameters a = 4.4 Å, l(max) = 7, n(max) = 7. Results are presented in Table 3. The program was not able to position the fragment of 12 atoms at a resolution lower than 3.0 Å.

4.7. ICA3

$P2_12_12_1$, a = 61.941, b = 70.425, c = 148.962 Å, resolution 2.3–60.0 Å. Protein dimer, 780 residues in total (Borek, 2001). n(max) = 5, l(max) = 7, grid 0.6 Å. One turn of poly-Ala α-helix (18 atoms) was positioned into many places in the electron density. The best peak (based on CC) had ERR = 0.178 Å in comparison with refined coordinates. There were clusters of peaks clearly indicating α-helix.

4.8. Fragment expansion

Coefficients of expansion have the important convergence property given by (11). For the RKSA1 fragment the table of the sum of expansion coefficients as function of n(max) and l(max) was calculated. The Ec fragment expansion was used. The important question was how many expansion coefficients were optimal for positioning the fragment. Results are shown in Fig. 1. The limiting extrapolated value of the sum was 695 000. The expansion practically converged for $n(\max) \simeq 6$ and $l \simeq 12$. However, even smaller numbers were sufficient to recover the fragment position from the electron density. Expansion coefficients of the Dirac fragment do not converge; E values instead of structure factors are more suitable for electron-density expansion if the fragment atoms are expressed as point scatterers.

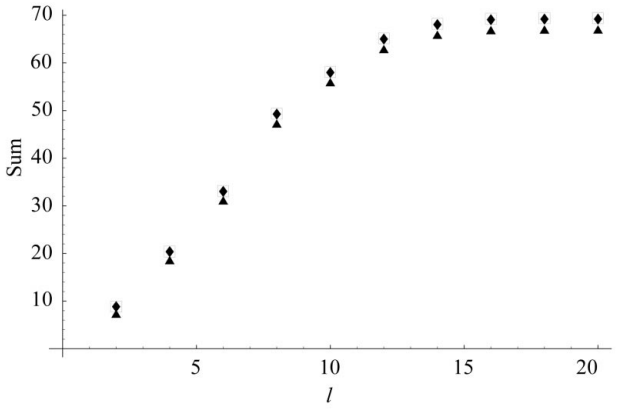

Figure 1
Convergence properties of expansion coefficients. Coordinates from the RKSA1 fragment. The sum is calculated according to (11). Filled triangles, n = 4; filled diamonds. n = 7; squares, n = 10.

4.9. Refinement

Refinement of Euler angles by analytical derivatives is a fast procedure. We found that only a marginal improvement can be achieved by refinement of angles only, because the interpolated angles are close to the correct values. The value of the overlap function is more sensitive to the correct fragment position than to the orientation.

4.9.1. Dirac refinement. No more than four iterations were necessary to refine both fragment position and orientation. A considerable increase (up to 10%) in correlation coefficient could be achieved. The procedure is simple, convergent and fast, but it is not sufficiently general. The disadvantage of Dirac refinement is that results are dependent on selection of the grid for electron-density expansion (on the distance of the center of the fragment to the nearest grid-point position).

4.9.2. Analytical derivative refinement. The method is sufficiently general and can be applied to any type of fragments. In each iteration step it is necessary to make several summations of structure factors, which makes this procedure considerably slower in comparison with Dirac refinement. However, because the electron-density expansion is recalculated at each new fragment position, the refinement is not dependent on the sampling grid.

4.10. Perspectives

Preliminary results from protein and polypeptide structures suggest that fragments of secondary-structure elements (α-helix and β-strand) can be easily connected into poly-Ala chains. Side chains can be modeled with locally restricted multidimensional searches. Fitting of quite large fragments (domains) in the early steps of phase development is also

promising. Omitting solvent regions can increase the speed of calculations. The search algorithm described in this paper can be parallelized at the level of one process and also at the level of several coordinated processes. Taking into account the price and accessibility of current personal computers, this can lead to a substantial decrease in the necessary real computing time.

APPENDIX *A*
Randomization of phases.

Randomization is based on the idea of the partially known structure

$$F_c = \sum_P f_p \exp(2\pi i \mathbf{h}\cdot\mathbf{r}_p) + \sum_Q f_q \exp(2\pi i \mathbf{h}\cdot\mathbf{r}_q), \qquad (47)$$

where P are known and Q are unknown atoms. The unknown part has a limited maximal value. The structure factor $F_h = A_h + iB_h$ is modified,

$$\begin{aligned} A'_h &= (1-p)A_h + pKX \\ B'_h &= (1-p)B_h + pKY, \end{aligned} \qquad (48)$$

where p is the randomization parameter, X and Y are random numbers $(-1, 1)$ and K is a constant. We used $K = 1$ for randomization of E values.

$$\mathrm{RR} = \frac{\sum \big||F'| - |F|\big|}{\sum |F|}, \qquad (49)$$

where

$$F' = (A'^2_h + B'^2_h)^{1/2}, \quad tg\varphi' = B'/A'.$$

The related phase error is calculated by

$$\mathrm{DHPI} = \left[(1/N)\sum(\Delta\varphi)^2\right]^{1/2}. \qquad (50)$$

$\Delta\varphi$ is the phase difference (the smaller of two phase values on the circle).

FP and JZ are indebted to the University of Texas, Southwestern Medical Center at Dallas for financial support. The research was supported by grant NIH GM 53163 and by VEGA 1/7264/20 of the Slovak Republic. The authors are grateful to Professor E. J. Dodson for critical reading of the manuscript and valuable discussion.

References

Altmann, S. L. & Bradley, C. J. (1963). *Philos. Trans. R. Soc. London Ser. A*, **255**, 193–198.
Bentley, G. A. (1997). *Methods Enzymol.* **276** 611–619.
Borek, D. (2001). PhD thesis, A. Mickiewicz University.
Cowtan, K. D. (1998). *Acta Cryst.* D**54**, 750–756.
Crowther, R. A. (1972). *The Molecular Replacement Method*, edited by M. G. Rossmann, pp. 173–178. New York: Gordon & Breach.
Dodson, E. J. (1985). *Proceedings of the CCP4 Study Weekend. Molecular Replacement*, pp. 33–45. Warrington: Daresbury Laboratory.
Fortier, S. & Glasgow, J. (1998). *Direct Methods for Solving Macromolecular Structure*, edited by S. Fortier, pp. 339–359. Dordrecht: Kluwer Academic Publishers.
Friedman, J. (1999). *Comput. Chem.* **23**, 9–23.
Giacovazzo, C. (1998). *Direct Phasing in Crystallography*, pp. 645–651. Oxford University Press.
Hu, S. H., Gehrmann, J., Guddat, L. W., Alewood, P. F., Craik, D. J. & Martin, J. L. (1996). *Structure*, **4**, 417–423.
Jones, T. A. (1978). *J. Appl. Cryst.* **11**, 268–272.
Jones, T. A. & Kjeldgaard, M. (1997). *Methods Enzymol.* **227**, 173–230.
Karle, I. L. & Karle, J. (1981). *Proc. Natl Acad. Sci. USA*, **78**, 681–685.
Kleywegt, G. J. & Jones, T. A. (1997). *Acta Cryst.* D**53**, 179–185.
Koos, M., Steiner, B., Gajdos, J., Langer, V., Gyepesova, D., Smrcok, L. & Durik, M. (2000). *Molecules*, **5**, 219–226.
Landau, L. D. & Lifschitz, E. M. (1972). *Theorie Quantique Relativiste.* Moscow: MIR.
Lattman, E. E. (1972). *Acta Cryst.* B**28**, 1065–1068.
McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.
Navaza, J. (1990). *Acta Cryst.* A**46**, 619–620.
Oldfield, T. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 67–74. Warrington: Daresbury Laboratory.
Pavelcik, F. (1994). *Acta Cryst.* A**50**, 551.
Pavelcik, F. (1998). *J. Appl. Cryst.* **31**, 960–962.
Rollett, J. S. (1965). *Computing Methods in Crystallography*, edited by J. S. Rollett, pp. 35–37. Oxford: Pergamon Press.
Steiner, B., Koos, M., Langer, V., Gyepesova, D. & Smrcok, L. (1998). *Carbohydr. Res.* **311**, 1–9.
Teeter, M. M., Roe, S. M. & Heo, N. H. (1993). *J. Mol. Biol.* **230**, 292–311.
Wigner, E. (1959). *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra.* New York: Academic Press.