Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Pushing the boundaries of molecular replacement with maximum likelihood

Randy J. Read

Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Hills Road, Cambridge CB2 2XY, England

Correspondence e-mail: rjr27@cam.ac.uk

Received 27 April 2001
Accepted 18 July 2001

The molecular-replacement method works well with good models and simple unit cells, but often fails with more difficult problems. Experience with likelihood in other areas of crystallography suggests that it would improve performance significantly. For molecular replacement, the form of the required likelihood function depends on whether there is ambiguity in the relative phases of the contributions from symmetry-related molecules (*e.g.* rotation *versus* translation searches). Likelihood functions used in structure refinement are appropriate only for translation (or six-dimensional) searches, where the correct translation will place all of the atoms in the model approximately correctly. A new likelihood function that allows for unknown relative phases is suitable for rotation searches. It is shown that correlations between sequence identity and coordinate error can be used to calibrate parameters for model quality in the likelihood functions. Multiple models of a molecule can be combined in a statistically valid way by setting up the joint probability distribution of the true and model structure factors as a multivariate complex normal distribution, from which the conditional distribution of the true structure factor given the models can be derived. Tests in a new molecular-replacement program, *Beast*, show that the likelihood-based targets are more sensitive and more accurate than previous targets. The new multiple-model likelihood function has a dramatic impact on success.

1. Introduction

Since the pioneering work by Rossmann & Blow (1962), molecular replacement has grown to be one of the most powerful tools of the macromolecular crystallographer. It will become even more important as the emerging structural genomics efforts generate structural models for an increasing fraction of possible folds. However, there is a need for methods to improve. Coverage of fold space would increase substantially if lower homology models could be tolerated. Even with good models, molecular replacement can be difficult if there are many copies in the unit cell. More sensitive scores for judging molecular-replacement solutions would help and likelihood is an excellent candidate.

The principle of maximum likelihood is quite simple: the best model is most consistent with the observations. Consistency is measured statistically by the probability that the observations should have been made. If the model is changed to make the observations more probable, the likelihood goes up, indicating that the model is better. When the probability distributions for the observations are Gaussian, maximum likelihood is equivalent to least squares. Maximum likelihood

© 2001 International Union of Crystallography
Printed in Denmark – all rights reserved

has become prominent in protein crystallography because the probability distributions of the observations are rarely Gaussian so that least-squares methods are rarely justified. Indirectly, the phase problem underlies the importance of likelihood. Many important probability distributions for phased structure factors (complex numbers for acentric structure factors, real numbers for centric) are indeed Gaussian, but we measure only amplitudes or intensities. The change of variables and integration to eliminate the unknown phase changes the form of the distributions.

Likelihood has been used for some time in macromolecular crystallography. The program *SIGMAA* (Read, 1986) computes model phase probabilities using σ_A parameters optimized by maximizing a likelihood function; Lunin & Urzhumtsev (1984) first suggested estimating phase probabilities by maximizing a similar likelihood function. In structure refinement, likelihood has been demonstrated to be much better than the traditional least-squares target (Pannu & Read, 1996; Murshudov *et al.*, 1997; Bricogne & Irwin, 1996). The improvement is even more striking if experimental phase information is exploited (Pannu *et al.*, 1998). (The unphased refinement likelihood target is essentially identical to the *SIGMAA* likelihood target, if one ignores the small effect of observation errors.) The introduction of likelihood into experimental phasing by isomorphous replacement or anomalous dispersion, implemented in the program *SHARP* (de La Fortelle & Bricogne, 1997), has improved both the quality of phases and the estimates of their accuracy.

Molecular replacement can be considered as a hypothesis-testing problem, in which different hypotheses about the orientation, position and (possibly) quality of the search model are tested against the data. As Bricogne (1997) has pointed out in this and other crystallographic contexts, likelihood is an ideal criterion for hypothesis testing. Bricogne (1992, 1997) first suggested applying likelihood to molecular replacement, but did not deal with the specific problems of a rotation likelihood function or of multiple models discussed below and had not reported any details of implementation at the time this work was carried out. Some of the ideas described here have been tested through a preliminary implementation (Read, 1999) in a modified version of *BRUTE* (Fujinaga & Read, 1987). To test new ideas, such as a multiple-model likelihood function, and to improve performance and ease of use, a new program, *Beast*, has now been written and is described here.

2. Likelihood functions for molecular replacement

Although the principle of maximum likelihood is simple, it can be difficult to derive appropriate probability distributions on which to base the likelihood targets. Complications often arise because of ambiguities: unknown phase angles or (as discussed below) unknown *relative* phase angles between contributions from symmetry-related molecules. What is needed is the probability distribution of the measurements, given as a function of model parameters and sources of error. The sources of error include errors in measuring the diffraction data, but for crystallographic applications the effects of errors in the atomic model are usually much larger. For this reason, measurement errors have been neglected in this work. A variety of types of error in the model can be shown to give rise to a Gaussian probability distribution for the true structure factor (Read, 1990, 1997), but it is important to note that these Gaussian distributions apply to the *phased* structure factor, not to its amplitude.

2.1. Likelihood function for translation or six-dimensional search

Traditionally, molecular replacement has been carried out with a divide-and-conquer approach, in which the dimensionality of the problem is reduced by separating the search for one molecule into two separate three-dimensional searches: a rotation search for orientation and a translation search for position (Rossmann, 1972). With modern computers, a six-dimensional search can now be applied if necessary, either as a grid search (Sheriff *et al.*, 1999) or using stochastic methods (Chang & Lewis, 1997; Kissinger *et al.*, 1999; Glykos & Kokkinidis, 2000). A full six-dimensional search can be thought of as testing a series of hypotheses about the orientation and position of the model. Similarly, for a translation search one is testing a series of hypotheses about the position of the model for a given orientation. The same likelihood function is appropriate for both searches, where the best solution will place all the atoms of the model in approximately the correct position and the calculated structure factor will be a reasonable approximation of the true structure factor.

In these cases, the likelihood function used in *SIGMAA* or in maximum-likelihood structure refinement is the appropriate choice. This likelihood function is based on the structure-factor probability distributions given in (1), where p_a in (1*a*) describes the two-dimensional Gaussian distribution for acentric structure factors and p_c in (1*b*) describes the one-dimensional Gaussian distribution for centric structure factors,

$$p_a(\mathbf{F}_O;\mathbf{F}_C) = \frac{1}{\pi\varepsilon\sigma_\Delta^2}\exp\left[-\frac{|\mathbf{F}_O - D\mathbf{F}_C|^2}{\varepsilon\sigma_\Delta^2}\right] \qquad (1a)$$

$$p_c(\mathbf{F}_O;\mathbf{F}_C) = \frac{1}{(2\pi\varepsilon\sigma_\Delta^2)^{1/2}}\exp\left[-\frac{|\mathbf{F}_O - D\mathbf{F}_C|^2}{2\varepsilon\sigma_\Delta^2}\right], \qquad (1b)$$

where $\sigma_\Delta^2 = \Sigma_N - D^2\Sigma_P$, $\Sigma_N = \langle F_O^2/\varepsilon\rangle$, $\Sigma_P = \langle F_C^2/\varepsilon\rangle$, ε is the expected intensity factor and D is the Luzzati (1952) weighting factor.

Fig. 1 presents a schematic illustration of (1*a*) as applied to a translation search. In (1), the effect of measurement error is neglected and the measured amplitude, F_O, is assumed to be equal to the true amplitude. Measurement error generally has much less impact than the effect of model errors, particularly for difficult molecular-replacement problems, and it will be ignored in what follows. Nonetheless, the effect of measurement error could be included by using likelihood targets such as MLF1 and MLF2 (Pannu & Read, 1996) or by incrementing the variances (Murshudov *et al.*, 1997; Bricogne & Irwin, 1996). Note that uncertainty is increased by either incompleteness of the model (difference between Σ_N and Σ_P) or errors in the model (leading to lower values of D).

It is often convenient to work with normalized structure factors or E values because the probability distributions can then be expressed in terms of a single parameter σ_A instead of the two parameters σ_Δ and D,

$$p_a(\mathbf{E}_O;\mathbf{E}_C) = \frac{1}{\pi(1-\sigma_A^2)}\exp\left[-\frac{|\mathbf{E}_O-\sigma_A\mathbf{E}_C|^2}{1-\sigma_A^2}\right] \qquad (2a)$$

$$p_c(\mathbf{E}_O;\mathbf{E}_C) = \frac{1}{[2\pi(1-\sigma_A^2)]^{1/2}}\exp\left[-\frac{|\mathbf{E}_O-\sigma_A\mathbf{E}_C|^2}{2(1-\sigma_A^2)}\right], \qquad (2b)$$

where $\mathbf{E}_O = \mathbf{F}_O/(\varepsilon\Sigma_N)^{1/2}$, $\mathbf{E}_C = \mathbf{F}_C/(\varepsilon\Sigma_P)^{1/2}$ and $\sigma_A = D(\Sigma_P/\Sigma_N)^{1/2}$.

The likelihood functions require probabilities of amplitudes or intensities, so the unknown phase angle must be eliminated by integrating it out (acentric case) or summing over the two possible phase choices (centric case), giving

$$p_a(E_O;E_C) = \frac{2E_O}{1-\sigma_A^2}\exp\left(-\frac{E_O^2+\sigma_A^2E_C^2}{1-\sigma_A^2}\right)I_0\left(\frac{2E_O\sigma_AE_C}{1-\sigma_A^2}\right) \qquad (3a)$$

$$p_c(E_O;E_C) = \left[\frac{2}{\pi(1-\sigma_A^2)}\right]^{1/2}\exp\left[-\frac{E_O^2+\sigma_A^2E_C^2}{2(1-\sigma_A^2)}\right] \times \cosh\left(\frac{E_O\sigma_AE_C}{1-\sigma_A^2}\right). \qquad (3b)$$

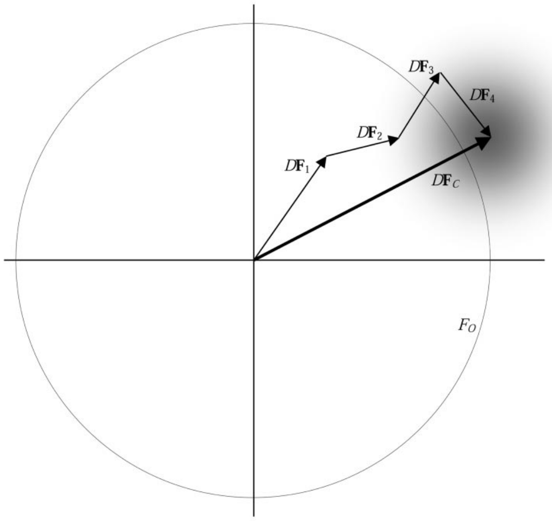

Figure 1
Schematic illustration of translation likelihood function for acentric structure factors. As a molecule is translated, the molecular-transform contributions from the symmetry-related copies (four in this example) will change in phase but not in amplitude. For the correct translation, the true structure factor will be found within a two-dimensional Gaussian distribution (shown as grey shading) centered on the total calculated structure factor, scaled by the factor D to obtain the centroid of the distribution (Read, 1990). The contribution of a single structure factor to the likelihood function is obtained by integrating around a circle with a radius given by the observed amplitude, F_O, so the likelihood will be high when this circle intersects regions of high probability in the two-dimensional Gaussian. For a combined rotation/translation search, both the amplitudes and phases of the molecular-transform contributions will vary.

2.2. Rotation likelihood function

Compared with a translation search, a rotation search differs in that the position of the molecule is considered to be unknown, so that the relative phases of the symmetry-related contributions of each molecule to the total structure factor are unknown. Given a trial orientation, we only have an estimate of the *amplitudes* of the molecular-transform contributions. The hypothesis we are testing, for each orientation, is that the set of observed structure factors could be obtained by adding up the molecular-transform contributions with some set of unknown relative phases, possibly with an additional contribution from unmodeled structure.

This is a random-walk problem like that of the Wilson (1949) distribution. In the rotation likelihood function, the symmetry-related molecular transforms (which vary in magnitude with orientation) play the role of the atomic scattering factors in the Wilson distribution. One significant difference is that each molecular-transform contribution has an associated uncertainty arising from model errors. The molecular-transform contribution of a single copy of a single molecule can be considered as a structure factor in $P1$, for which the distribution in (1a) applies. The effect of model errors is to downweight the contribution by the factor D for that molecule and to increase the variances by a factor of $(1 - D^2)$ times the total scattering power of the molecule (Read, 1990). Note that because the molecular transform has $P1$ symmetry, symmetry-related contributions to the structure factor lack the crystal symmetry and are in general independent. (Corrections using the expected intensity factor ε must be made in zones of the reciprocal lattice where contributions of symmetry-related molecules are constrained to be equal.)

The random-walk problem of the rotation likelihood function can be treated at various levels of approximation. At the crudest level, we could assume that the central limit theorem applies to obtain a Wilson-like approximation to the rotation likelihood function, illustrated schematically in Fig. 2(a) and defined by

$$p_a(\mathbf{F}_O;\{\mathbf{F}_{jk}\}) = \frac{1}{\pi\varepsilon\Sigma_W}\exp\left(-\frac{F_O^2}{\varepsilon\Sigma_W}\right), \qquad (4)$$

where $\{\mathbf{F}_{jk}\}$ is the set of contributions of symmetry copies k of molecules j,

$$\Sigma_W = \left[\Sigma_N - \sum_j\sum_k D_j^2\Sigma_j\right] + \sum_j\sum_k D_j^2|\mathbf{F}_{jk}|^2$$

and $\Sigma_j = \langle F_{jk}^2\rangle$ for each of the symmetry copies k.

The component of Σ_W in square braces is the random error arising from model incompleteness and model errors. (4) allows for the possibility of more than one molecule in the

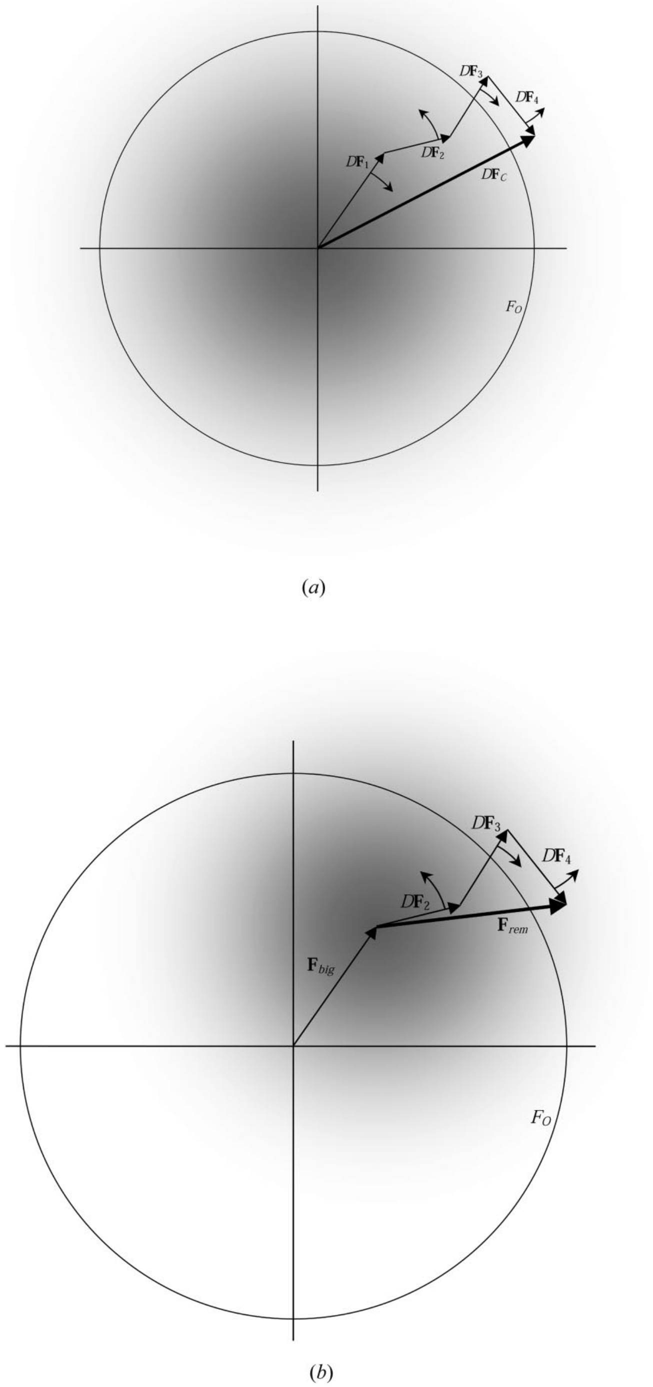

Figure 2
Schematic illustration of rotation likelihood functions for acentric structure factors. (*a*) In the Wilson-like approximation, the distribution is assumed to be a two-dimensional Gaussian arising from the sum of molecular-transform contributions with unknown phase angles, together with random errors resulting from model incompleteness and model error. (*b*) In the Sim-like approximation, the contribution from the single largest molecular transform ($\mathbf{F}_{\rm big}$) has an arbitrary phase and the distribution is assumed to be a two-dimensional Gaussian arising from the sum of the remaining molecular-transform contributions ($\mathbf{F}_{\rm rem}$) with unknown phase angles relative to the phase of $\mathbf{F}_{\rm big}$, together with random errors resulting from model incompleteness and model error.

asymmetric unit of the crystal. Only the acentric unnormalized case is given, but the centric case follows easily by analogy and normalization requires only a simple change of variables, as above. The likelihood function requires the probability of the amplitude (or intensity), obtained by integrating out the unknown phase,

$$p_a(F_O;\{\mathbf{F}_{jk}\}) = \frac{2F_O}{\varepsilon\Sigma_W}\exp\left(-\frac{F_O^2}{\varepsilon\Sigma_W}\right). \qquad (5)$$

For this to be a good approximation, the assumptions of the central limit theorem must apply, *i.e.* there must be a sufficient number of contributions to the sum and none may dominate. However, the number of molecular-transform contributions is often small. Interestingly, the Wilson-like approximation tends to become more valid as molecular-replacement problems become more difficult, either because there is a larger number of molecules in the unit cell (combination of non-crystallographic and crystallographic symmetry) or because the model is poorer or less complete (the Gaussian noise contribution becomes proportionately greater, so that the overall distribution is better modeled as Gaussian). For easier molecular-replacement problems, it may not matter that the approximation is poorer. An advantage of the Wilson-like approximation (compared with the Sim-like approximation discussed below) is that it is continuously differentiable and may lend itself to rapid approximations that can be computed by FFT methods.

Nonetheless, it is possible to derive better approximations. Shmueli and coworkers have addressed the question of structure-factor probability distributions in situations where the central limit theorem approximation is poorly justified, *i.e.* for small numbers of atoms or heterogeneous compositions (Shmueli & Weiss, 1995). They have derived probability distributions as Fourier–Bessel series, effectively by performing the convolution of the probability distributions of individual atomic contributions. The atomic distributions, for acentric structure factors, are non-zero on circles in the complex plane. The distribution for sums of molecular transforms can be derived by analogy, with the additional consideration that the Gaussian noise contribution from model error adds an additional convolution step, which introduces an exponential falloff term. Carrying this factor through, the probability distribution for acentric structure factors is

$$p_a(F_O) = \frac{2F_O}{F_{\max}^2}\sum_{m=1}^{\infty} D_m J_0\left(\frac{\gamma_m F_O}{F_{\max}}\right) \text{ for } 0 < F_O < F_{\max}, \qquad (6)$$

where $F_{\max}$ is the maximum possible F_O, γ_m is the mth zero of the J_0 Bessel function,

$$D_m = \frac{1}{J_1^2(\gamma_m)}\exp\left(-\frac{\gamma_m^2\sigma_\Delta^2}{4F_{\max}^2}\right)\prod_j J_0\left(\frac{\gamma_m DF_j}{F_{\max}}\right)$$

and F_j is the contribution from symmetry copy *j*.

Numerical simulations support this form of the probability distribution, but it can take a large number of terms (up to $m = 100$) to converge and is relatively expensive to compute. However, there is an intermediate level of approximation, analogous to a suggestion of Shmueli *et al.* (1984). They found that for heterogeneous compositions with a single heavy atom, the Sim (1959) distribution is a good approximation, with the heaviest atom forming the partial structure and the remaining atoms comprising the missing structure. The Sim distribution has the same functional form as (1*a*), with the centric case (1*b*) corresponding to the Woolfson (1956) distribution. A Sim-like approximation to the rotation likelihood function is defined in (7), in which the single largest molecular-transform contribution plays the role of $\mathbf{F}_C$ in (1*a*) and the variance term is incremented by the sum of the squares of the remaining molecular-transform contributions,

$$p_a(\mathbf{F}_O; \{\mathbf{F}_{jk}\}) = \frac{1}{\pi\varepsilon\Sigma_S}\exp\left(-\frac{\left|\mathbf{F}_O - \mathbf{F}_{\rm big}\right|^2}{\varepsilon\Sigma_S}\right), \qquad (7)$$

where $F_{\rm big} = \max\{D_j\mathbf{F}_{jk}\}$ is the biggest molecular transform contribution,

$$\begin{aligned}\Sigma_S &= \left[\Sigma_N - \sum_j\sum_k D_j^2\Sigma_j\right] + \sum_j\sum_k D_j^2|\mathbf{F}_{jk}|^2 - F_{\rm big}^2\\ &= \Sigma_W - F_{\rm big}^2\end{aligned}$$

and $F_{\rm big} = |\mathbf{F}_{\rm big}|$.

This distribution is illustrated schematically in Fig. 2(*b*). Integration over the unknown phase angle gives

$$p_a(F_O; \{\mathbf{F}_{jk}\}) = \frac{2F_O}{\varepsilon\Sigma_S}\exp\left(-\frac{F_O^2 + F_{\rm big}^2}{\varepsilon\Sigma_S}\right)I_0\left(\frac{2F_OF_{\rm big}}{\varepsilon\Sigma_S}\right). \qquad (8)$$

Numerical simulations comparing the two approximations to the more exact form in (6) verify that the Sim-like approximation defined by (8) is better than the Wilson-like approximation defined by (5). However, as the parameters are adjusted to reflect difficult molecular-replacement problems (poor models and/or many molecules in the unit cell) the two approximations converge more closely to the exact form of the distribution. In the program and tests described below, the Sim-like approximation (and its centric analogue) are used for the rotation likelihood function.

2.3. Likelihood functions with partial ambiguity

Apart from the rotation problem, there are other cases in which there will be at least partial ambiguity of the relative phases of the contributions of different molecules. For complexes or crystals with non-crystallographic symmetry, the orientation and/or position of a subset of the molecules may be known and it would be helpful to use this information in computing rotation or translation functions for the remaining molecules. If only the orientation of a fixed molecule is known, then the individual symmetry-related molecular transforms all have unknown relative phases.

It may also be useful to define only part of the position vector, leaving the rest undetermined and thereby reducing the dimensionality of the translation search. For example, in the space group *P*622 each molecule takes 12 symmetry-related orientations and positions. If one searches in the *xy* plane, the relative positions of each set of six molecules related by the sixfold axis (and thus their relative phases) are defined. The *z* component of the translation only changes the relative position (and phase) of the two sets of six molecules. This is illustrated schematically in Fig. 3.

Finally, there will be ambiguities arising from coarseness of the search grids, which can be accounted for by using expected values and incrementing the variances (Bricogne, 1997). If the translation search is carried out on a coarse grid, there will be partial ambiguity of the relative phases of the contributions of symmetry-related molecules. This can be dealt with in the same way as positional uncertainty of individual atoms (Read, 1990) by reducing the expected value of the molecular-transform contribution and incrementing the variance correspondingly. A coarser rotation grid could also be used, accounting for the increased uncertainty in the orientation by averaging the molecular transform over the rotational uncertainty and incrementing the variances.

Searches with intermediate dimensionality (*e.g.* five-dimensional search of orientation and position in a plane for *P*622) may be important for improving signal-to-noise in difficult cases. This will be particularly true when the molecules in the crystal take on many orientations, through the combination of crystallographic and non-crystallographic symmetry. In such a case, the rotation likelihood function will have many molecular-transform terms of comparable weight. Each molecular-transform term is itself drawn from a Wilson distribution: the more terms there are, the more the overall

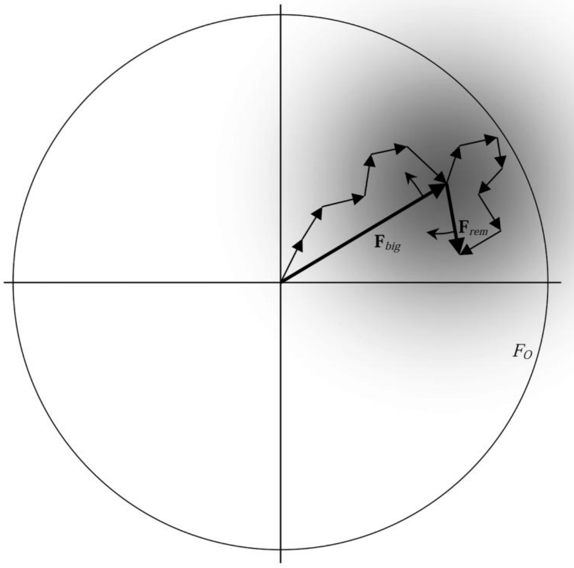

Figure 3
Schematic illustration of likelihood function for partial translational ambiguity. This example illustrates the uncertainty in an acentric structure factor in space group *P*622 when a translation search is conducted over the *xy* plane, leaving the *z* coordinate undefined. For any particular *xy* combination, varying *z* will change the phases of two groups of six molecular transforms in concert. At the correct *xy* translation, the uncertainty in *z* corresponds to uncertainty in the relative phase angle between the two sums of six molecular transforms, shown as heavy arrows. This uncertainty is modeled as a Sim-like probability distribution, similar to that shown in Fig. 2(*b*).

likelihood distribution will tend towards the same mean for all reflections, thus losing sensitivity. Increasing the dimensionality to (for instance) five in $P622$ reduces the number of separately phased contributions by a factor of six, greatly reducing the averaging effect that dilutes out the signal in the likelihood function. More generally, when the hypothesis is made more specific by reducing ambiguity, the probability distributions become sharper and the likelihood functions become more informative. This can be understood by comparing the schematic illustrations presented in Figs. 1 and 2.

3. Calibrating the likelihood functions

The likelihood functions depend on the values assumed for σ_A as a function of resolution. In principle, for each trial rotation and translation the σ_A curve could be adjusted to maximize the likelihood function, but this would be computationally very demanding. Nonetheless, σ_A values should be refined with the *SIGMAA* (Read, 1986) algorithm as part of the final scoring of potential solutions. During the search, a good *a priori* estimate of σ_A values can be made, with this forming part of the hypothesis to be tested.

The *a priori* estimates of σ_A are based on strong correlations between sequence identity and r.m.s. coordinate error (Chothia & Lesk, 1986). With a number of simplifying assumptions, the variation of σ_A as a function of resolution can be expressed as a function of the Fourier transform of the coordinate-error probability distribution. This behaviour is complicated by the effect of unmodelled or poorly modelled bulk solvent, which causes σ_A to fall off at low resolution. The behaviour of σ_A as a function of resolution can be modeled by the four-parameter functional form used in *REFMAC* (Murshudov *et al.*, 1997),

$$\sigma_A = \{f_p[1 - f_{\rm sol}\exp(-B_{\rm sol}\sin^2\theta/\lambda^2)]\}^{1/2} \times \exp\left(-\frac{8\pi^2}{3}\sigma_r^2\sin^2\theta/\lambda^2\right), \qquad (9)$$

where $f_{\rm sol}$ and $B_{\rm sol}$ describe the low-resolution solvent-related falloff, f_p is the fraction of ordered structure comprised by the model and σ_r is the r.m.s. coordinate error of the model. The two solvent-related terms affect a minority of data and standard values can be chosen. Inspection of σ_A curves suggests that suitable values for $f_{\rm sol}$ range from 0.8 to 0.95 and for $B_{\rm sol}$ from 100 to 250 Å^2. The current program defaults are 0.95 and 150, whereas the tests described below used values of 0.8 and 100. As expected, the choice of these parameters has only a small impact on the quality of results. The completeness of the model is generally known before molecular replacement is carried out and the r.m.s. coordinate error can be estimated using an equation derived by Chothia & Lesk (1986),

$$\sigma_r = 0.40\exp[1.87(1 - s)], \qquad (10)$$

where s is the fractional sequence identity.

Although (10) was derived by fitting data for r.m.s. deviation of main-chain atoms only, it works well in tests such as those described below. It would be preferable to choose the parameters in such an equation by optimizing likelihood functions; work is in progress to do this by comparing structure factors from related structures (R. B. Dodd & R. J. Read, unpublished work). Still better would be to estimate coordinate errors varying over the molecule as a function of local sequence identity and (perhaps) surface exposure. Such estimates could be used to weight the relative contributions of different atoms by adjusting their B factors (Read, 1990) and to compute better σ_A estimates.

4. Multivariate distributions for multiple models

As the database of known protein structures expands, one often has several choices of molecular-replacement model and the number of choices increases as the threshold for acceptable sequence identity levels is relaxed. In a number of cases, difficult molecular-replacement structures have been solved by using averaged electron density computed from several models that individually were not good enough (*e.g.* the test case discussed below of Pieper *et al.*, 1998). However, using multiple models in a likelihood function requires deriving the probability of the true structure factor given a collection of calculated structure factors. This must account for correlations between pairs of models. Two highly correlated models will provide less independent information than two uncorrelated models. The statistical framework that considers factors such as this is based on the complex multivariate normal distribution.

It is only necessary to consider the acentric case because the molecular transforms are computed in space group $P1$. The acentric structure-factor distribution in (1a) can be considered either as a bivariate normal distribution of the real and imaginary parts of the structure factor, with equal variances and zero covariances, or as a complex normal distribution. Such a complex normal distribution can be generalized to the multivariate case (Wooding, 1956), with properties similar to those of the real multivariate normal distribution. Since acentric structure factors for proteins are sums of large numbers of complex atomic contributions, it is reasonable to assume that the central limit theorem applies. As Tsoucaris (1970) points out, such an assumption is supported by general results by Klug (1958) on multivariate structure-factor distributions.

In a multivariate normal distribution applied to real numbers (such as centric structure factors), the variance term found in the univariate normal distribution is replaced by a covariance matrix which is symmetric. The diagonal terms are variances and the off-diagonal terms are covariances defined for variables x_i and x_j with means μ_i and μ_j as

$$\sigma_{ij} = \langle(x_i - \mu_i)(x_j - \mu_j)\rangle. \qquad (11)$$

In the complex multivariate normal distribution, the covariance matrix is in general Hermitian (meaning that $\boldsymbol{\sigma}_{ji}$ is the complex conjugate of $\boldsymbol{\sigma}_{ij}$ or that the matrix is equal to its Hermitian transpose). The covariance terms for complex variables $\mathbf{z}_i$ and $\mathbf{z}_j$ with means $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ are defined as

$$\boldsymbol{\sigma}_{ij} = \langle (\mathbf{z}_i - \boldsymbol{\mu}_i)(\mathbf{z}_j - \boldsymbol{\mu}_j)^* \rangle. \qquad (12)$$

The joint probability distribution is defined in terms of the covariance matrix $\boldsymbol{\Sigma}$ as

$$p(\mathbf{z}) = \frac{1}{|\pi\boldsymbol{\Sigma}|} \exp[-(\mathbf{z} - \boldsymbol{\mu})^H \Sigma^{-1} (\mathbf{z} - \boldsymbol{\mu})], \qquad (13)$$

where $(\mathbf{z} - \boldsymbol{\mu})$ is a column vector and superscript H indicates its Hermitian transpose (a row vector of complex conjugates) and vertical bars indicate the determinant of the matrix.

To obtain the probability distribution of the true molecular-transform contribution for a particular molecule, we start with the joint distribution of the molecular transforms for the true structure and all the models. The structures (and hence the structure factors) are related, but before the models are fixed the positions of the atoms are considered unknown, so that the structure factors all have expected values of zero. The terms in the covariance matrix are then given by

$$\boldsymbol{\sigma}_{ij} = \langle \mathbf{F}_i \mathbf{F}_j^* \rangle. \qquad (14)$$

If we normalize the structure factors so that their mean-square values (complex variances) are one, the covariance matrix becomes a correlation matrix, with diagonal elements equal to one and off-diagonal elements given by

$$\rho_{ij} = \langle \mathbf{E}_i \mathbf{E}_j^* \rangle. \qquad (15)$$

In other applications of multivariate complex normal distributions to crystallography, the off-diagonal elements will have an imaginary component. However, in the case of multiple models there is no reason to expect a significant imaginary component unless the models are translationally misaligned, leading to a systematic phase shift. The off-diagonal terms of the correlation matrix are therefore real and are equivalent to σ_A values between pairs of models. Such an interpretation of σ_A in terms of a real correlation of structure factors has been proposed by Srinivasan & Chandrasekaran (1966). In practice, (15) is used to compute elements of the correlation matrix between structure factors from models for which both phases are known. Because the correlations will vary with resolution, separate correlation matrices are computed for resolution shells. Values of σ_A computed from the functional form given in (9) are used for the correlation terms between the true (numbered 0 in the following) and model (numbered 1 to n) molecular transforms.

Standard manipulations allow one to derive a conditional probability distribution from a multivariate normal distribution when some of the variables are known (Johnson & Wichern, 1998). The new distribution is also normal and has a new mean and covariance/correlation matrix derived from a partitioning of the original matrix. For the case of multiple models, where all but one of the variables is fixed, the correlation matrix is partitioned as follows

$$P = \begin{bmatrix} 1 & P_{01} \\ P_{10} & P_{11} \end{bmatrix}, \qquad (16)$$

where P_{01} is a row vector of σ_A values between the true and model molecular transforms, P_{10} is its transpose and P_{11} is the correlation matrix involving only models. The conditional probability distribution is obtained as

$$p(\mathbf{E}_0; \{\mathbf{E}_i\}) = \frac{1}{\pi\sigma^2(\mathbf{E}_0)} \exp\left(-\frac{|\mathbf{E}_0 - \langle \mathbf{E}_0 \rangle|^2}{\sigma^2(\mathbf{E}_0)}\right) \qquad (17)$$

where $\sigma^2(\mathbf{E}_0) = 1 - P_{01}P_{11}^{-1}P_{10}$, $\langle \mathbf{E}_0 \rangle = P_{01}P_{11}^{-1}\mathbf{E}$ and $\mathbf{E}$ is the vector of model $\mathbf{E}_i$ values. It is easy to verify that for the case of one model this equation reduces to (2*a*).

5. Implementation of likelihood-based molecular replacement

A preliminary implementation (Read, 1999) of the rotation and translation likelihood functions (lacking the treatment of multiple models) was carried out in a modified version of *BRUTE* (Fujinaga & Read, 1987). These likelihood functions and the multiple-model likelihood function have now been reimplemented in a new program, *Beast*, which is faster, easier to use and designed to form part of the *CCP*4 (Collaborative Computational Project, 1994) program suite. The name '*Beast*' is an acronym for '*b*rute-force molecular replacement with *e*nsemble-*a*verage *st*atistics'. For convenience, *Beast* computes the log of the likelihood. This is placed on an absolute scale by subtracting the log-likelihood for the uninformative Wilson (1949) distribution, giving the log-likelihood gain (LLG). Like *BRUTE*, *Beast* uses a brute-force search of possible molecular-replacement solutions, which are scored individually. In principle, approximations could be devised to allow rapid calculations with FFTs (Bricogne, 1992), but it seemed more important at this point to develop a 'gold standard' against which such approximations could be judged.

Structure factors are interpolated in *Beast* from finely sampled molecular transforms, as performed for instance in *AMoRe* (Navaza, 1994). If multiple models are available, a statistically weighted ensemble average molecular transform is computed as described above and then used in further calculations. For efficiency, searches are carried out on a hexagonal close-packed grid as performed in *FFFEAR* (Kevin Cowtan, personal communication) using the locally orthogonal Lattman angles (Lattman, 1972) for orientation searches and an orthogonal search space for translation searches. For multiple-molecule searches, known molecules can be fixed in orientation and, optionally, in position.

6. Test cases

6.1. *Streptomyces griseus* trypsin

The structure of *S. griseus* trypsin (SGT) was solved, with some difficulty, using bovine trypsin (Chambers & Stroud, 1979) as a search model. It was difficult in part because of inaccuracy of rotation parameters determined with the fast rotation function (Crowther, 1972). Most attempts to solve the translation problem used an orientation obtained from a rotation function computed with data to 2.8 Å resolution that turned out to be 6.9° in error compared with the final molecular-replacement solution (Read & James, 1988). In

Table 1
Rotation-function results for *S. griseus* trypsin.

Algorithm	Resolution range (Å)	Correct peak†	Orientation error‡ (°)
Crowther	10.0–2.8	5.32	6.9
Crowther	10.0–3.5	5.62	3.4
Likelihood	25.0–2.8	7.80	0.8

† Peak height expressed in terms of r.m.s. deviations from the mean. ‡ Compared with final orientation from molecular replacement after rigid-body refinement.

Table 2
Translation-function results for *S. griseus* trypsin.

Orientation error (°)	Correct peak†	Highest noise peak†	Mean of search†	R.m.s. from mean†
6.9	2.3	−0.7	−43.9	9.5
3.4	84.6	32.0	−27.0	10.4
0.8	128.1	49.2	−16.7	10.5

† Scores are expressed in terms of log-likelihood gain.

Table 3
Rotation-function results with *H. volcanii* dihydrofolate reductase.

	AMoRe peak number†		Likelihood peak number‡	
No. of models§	Molecule 1	Molecule 2	Molecule 1	Molecule 2
1¶	7	16	2	5
3††	–	–	1	2
5‡‡	6	9	1	2
7§§	3	13	1	2

§ Models were chosen from a set of five *E. coli* DHFR structures (PDB codes 4dfr_A, 4dfr_B, 5dfr, 6dfr and 7dfr) with 32% sequence identity, one *Lactobacillus casei* DHFR structure (3dfr) with 23% sequence identity and one chicken liver DHFR (8dfr) with 25% sequence identity. † Results of Pieper *et al.* (1998), computed in *AMoRe* (Navaza, 1994). Multiple models were superimposed into a common orientation and their density averaged for the molecular-replacement calculation. No result was given for the three models. ‡ Computed in *Beast* using data from 3–25 Å resolution. ¶ Single *E. coli* DHFR model: 4dfr_B. †† One representative of each of three species: *E. coli* 4dfr_B,

Table 4
Translation-function results for *H. volcanii* dihydrofolate reductase.

	AMoRe correlation†			*Beast* LLG		
No. of models‡	Molecule 1	Molecule 2	Noise	Molecule 1	Molecule 2	Noise§
1	0.158	0.169	0.156	23.9	26.8	24.4
3	—	—	—	38.2	31.4	19.7
5	0.181	0.179	0.150	32.2	36.6	20.4
7	0.189	0.187	0.154	42.5	36.9	15.6

‡ As for Table 3. † Results of Pieper *et al.* (1998) computed in *AMoRe* (Navaza, 1994). No result was given for the three models. § Highest translation peak for an incorrect orientation.

contrast, a rotation function computed using data to only 3.5 Å resolution gave a more accurate orientation, with an error of only 3.4°. As shown in Table 1, both signal-to-noise and accuracy improve dramatically in the likelihood-based rotation function. In the likelihood approach, it is not necessary to choose the correct resolution range because data at high resolution are automatically downweighted if necessary.

In the initial structure solution, translation searches were carried out in *BRUTE* (Fujinaga & Read, 1987) using as a score the correlation between E^2 values from 4 to 8 Å resolution. These searches failed with the orientation that erred by 6.9°. Eventually, the structure was solved using a limited six-dimensional search in which the orientation was varied for a series of translation searches (Read & James, 1988). As shown in Table 2, the likelihood-based translation function succeeds even with the worst orientation. [Note that the log-likelihood gain is barely positive, implying that the model is barely more informative than the Wilson (1949) distribution. This occurs because the presumed r.m.s. error of 1.4 Å, deduced using (10) from the sequence identity of 32%, is a severe underestimate when the orientation is so much in error.] As the orientation of the model improves, the likelihood score and the discrimination from incorrect translations both improve significantly.

6.2. *Haloferax volcanii* dihydrofolate reductase

The structure of *H. volcanii* dihydrofolate reductase (DHFR) was solved by Pieper *et al.* (1998) using *AMoRe* (Navaza, 1994), but only when they used a composite model comprised of seven different DHFR structures. One of the biggest difficulties they faced was determining the orientations of the two molecules in the asymmetric unit. With the single best model (molecule *B* from the *Escherichia coli* DHFR in PDB file 4dfr or model 4dfr_B), the correct orientations showed up as peaks 7 and 16 in the *AMoRe* rotation search. Even though a subsequent translation search with all the orientations brings these orientations to the top of the list, the discrimination from noise is very poor. As the results in Tables 3 and 4 show, *Beast* displays much better signal-to-noise in this problem, particularly for the rotation search (Table 3) where 4dfr_B comes up as peaks 2 and 5.

Adding information from more models improves the results for both programs, but has greater effect with *Beast*. With *AMoRe*, the correct orientations were never at the top of the list, even with up to seven models. However, they are at the top of the list with the likelihood-based rotation function, even with just three models (Table 3). The translation searches are successful with both programs (Table 4); as the number of models increases, the discrimination improves, particularly for *Beast*.

It is interesting that adding multiple models of the same *E. coli* protein (albeit in different ligation states) improves the signal-to-noise ratio. To the extent that these models resemble each other (as measured by high correlations in the correlation matrix), they will be downweighted in the statistical average, so adding multiple copies of similar models will not dilute the signal that comes from other less similar models.

6.3. Other results

Test versions of *Beast* and the earlier implementation in *BRUTE* have been distributed to a number of laboratories, some of which have reported success in solving structures that could not be solved otherwise. Two such structures have been published, both using the modified version of *BRUTE*: *Sulfolobus solfataricus* cytochrome P450 (Yano *et al.*, 2000) and a hexitol nucleic acid (Declercq, 2000).

7. Conclusions

The introduction of likelihood-based scores has increased the sensitivity of molecular-replacement searches compared with more traditional methods. The introduction of multivariate statistics allows the optimal use of multiple models. As the database of known structures expands, it will be more and more common to have several possible models to choose from.

Apart from the increase in sensitivity, a great advantage to the likelihood-based targets is the reduction of adjustable parameters. It is common in molecular-replacement trials to experiment with the integration radii for the rotation function, resolution limits, degree of sharpening of the data and choice of model. Often, several models are constructed by trimming off different amounts of the least-conserved portions. Like the Patterson correlation searches in *BRUTE* (Fujinaga & Read, 1987), *X-PLOR* (Brünger, 1992) and *CNS* (Brunger *et al.*, 1998), the likelihood-based approach avoids integration radii, as the structure factors are always referred to the crystal cell. If the model quality is estimated correctly, data to too high resolution will effectively be ignored, so resolution limits are not necessary. The way in which variances in the probability distributions vary with resolution is controlled as well by the model quality parameters; the resulting variation in the extent to which data at different resolutions are consulted is what the sharpening parameters attempt to mimick. In *Beast*, it is not necessary to choose among several possible models; in fact, they should all be used. Finally, instead of trimming the least-conserved portions of the model, it would be better to downweight their influence by increasing their *B* factors according to their expected r.m.s. error (Read, 1990).

7.1. Other applications of molecular-replacement likelihood functions

The rotation likelihood function could be used to refine incomplete molecular-replacement solutions before the translation vector had been completely defined. For instance, the relative orientations of domains or elements of secondary structure could be refined; in favourable cases, it may even be possible to refine finer details of the structure. This approach has been successful using Patterson correlation refinement in *X-PLOR* (Brünger, 1992) and *CNS* (Brunger *et al.*, 1998) and should be even more powerful using likelihood targets.

The multiple-model likelihood function could be applied in other circumstances where more than one atomic model is available. If a structure were solved with multiple molecular-replacement models, the combined probability distribution for the true structure factor could be used to define better phase-probability distributions and *SIGMAA* map coefficients (Read, 1986), replacing $D\mathbf{F}_C$ by the expected value of $\mathbf{F}_O$ given the multiple models and σ_Δ^2 by the conditional variance in a manner similar to that shown in (17). The multiple-model likelihood function could also be used for refinement. One intriguing possibility is to save the model before simulated-annealing refinement as a fixed model, the information from which would be used while refining the moving model. This might be useful because in the course of simulated-annealing refinement, the model temporarily becomes worse. It has been found that when combining simulated annealing with likelihood, it is necessary to freeze the σ_A values during the annealing run (Adams *et al.*, 1997); if they are updated to lower values, pressure to fit the diffraction data is reduced and the refinement diverges. Keeping the initial model information would allow the refinement to 'remember' what was known about the true phases initially, which would restrain such divergence.

7.2. Future directions

In the most difficult molecular-replacement problems there are a large number of molecules in the unit cell, which reduces tremendously the signal in a rotation search. For such cases, the problem is not so much with the scoring function as the dimensionality of the search problem; once the answer has been found it is often clearly correct.

Stochastic search methods, such as Monte Carlo and genetic algorithms, are often very effective in such high-dimension problems. This can be seen, for instance, in the ligand-docking problem (Read *et al.*, 1995). Some success has already been achieved by such algorithms for molecular replacement (Chang & Lewis, 1997; Kissinger *et al.*, 1999; Glykos & Kokkinidis, 2000). The combination of these improved search methods with likelihood targets should make even more difficult problems tractable. This approach is presently being implemented (A. J. McCoy, N. S. Pannu & R. J. Read, unpublished work) within a new general phasing program under development in my laboratory.

An exciting possibility that will be explored is to gradually increase the dimensionality of the search space during optimization with a genetic algorithm. The initial search could define only the orientations of the molecules, scored by the rotation likelihood function. Two translation directions could be added, defining the positions of the molecules relative to the axis with highest rotational symmetry; finally, the last translation direction could be added. The effective size of the search space could be decreased and the convergence radius increased by allowing for uncertainty in the parameters. This would be performed by averaging the probability distributions over the uncertainty and incrementing the variances, as discussed in the context of coarse search grids. In the course of the search, the uncertainties would be gradually reduced to sharpen the score function.

Finally, in many molecular-replacement problems one has prior knowledge of the non-crystallographic symmetry operators, obtained from self-rotation and native Patterson functions (Navaza *et al.*, 1998). This information should also be exploited by coupling the parameters of copies of the search models.

Beast will be submitted for inclusion in the *CCP*4 (Collaborative Computational Project, Number 4, 1994) program suite after implementation and testing of the most important remaining options has been completed. In the meantime, it is available by request from the author.

It is a pleasure to acknowledge helpful discussions about the theory and algorithms with Navraj S. Pannu and Airlie J. McCoy. Garib Murshudov originally suggested that multivariate statistics should be relevant to the treatment of multiple models in *SIGMAA*, which inspired the use of such methods in this work. David Schuller provided useful feedback on features of the earlier implementation in *BRUTE*. Kay Diederichs provided the FFT routines that were used in *Beast* and contributed a code optimization that improved the speed of the translation search substantially. Osnat Herzberg kindly provided the dihydrofolate reductase test data. This research was supported by a Principal Research Fellowship from the Wellcome Trust, UK.

References

Adams, P. D., Pannu, N. S., Read, R. J. & Brünger, A. T. (1997). *Proc. Natl Acad. Sci. USA*, **94**, 5018–5023.

Bricogne, G. (1992). *Proceedings of the CCP4 Study Weekend. Molecular Replacement*, edited by W. Wolf, E. J. Dodson & S. Gover, pp. 62–75. Warrington: Daresbury Laboratory.

Bricogne, G. (1997). *Methods Enzymol.* **276**, 361–423.

Bricogne, G. & Irwin, J. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.

Brünger, A. T. (1992). *X-PLOR. Version 3.1. A System for X-ray Crystallography and NMR.* Yale University, Connecticut, USA.

Brunger, 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.

Chambers, J. L. & Stroud, R. M. (1979). *Acta Cryst.* B**35**, 1861–1874.

Chang, G. & Lewis, M. (1997). *Acta Cryst.* D**53**, 279–289.

Chothia, C. & Lesk, A. M. (1986). *EMBO J.* **5**, 823–826.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Crowther, R. A. (1972). *The Molecular Replacement Method*, edited by M. G. Rossmann, pp. 173–178. New York: Gordon & Breach.

Declercq, R. (2000). PhD thesis, Katholieke Universiteit Leuven, Leuven, Belgium.

Fujinaga, M. & Read, R. J. (1987). *J. Appl. Cryst.* **20**, 517–521.

Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst.* D**56**, 169–174.

Johnson, R. A. & Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis*, 4th ed. Upper Saddle River, NJ, USA: Prentice Hall.

Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.

Klug, A. (1958). *Acta Cryst.* **11**, 515–543.

La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.

Lattman, E. E. (1972). *The Molecular Replacement Method*, edited by M. G. Rossmann, pp. 179–185. New York: Gordon & Breach.

Lunin, V. Y. & Urzhumtsev, A. G. (1984). *Acta Cryst.* A**40**, 269–277.

Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.

Navaza, J., Panepucci, E. H. & Martin, C. (1998). *Acta Cryst.* D**54**, 817–821.

Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Acta Cryst.* D**54**, 1285–1294.

Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.

Pieper, U., Kapadia, G., Mevarech, M. & Herzberg, O. (1998). *Structure*, **6**, 75–88.

Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.

Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.

Read, R. J. (1997). *Methods Enzymol.* **277**, 110–128.

Read, R. J. (1999). *XVIIIth IUCr Congress and General Assembly.* Abstract No. M07.0A.002.

Read, R. J., Hart, T. N., Cummings, M. D. & Ness, S. R. (1995). *Supramol. Chem.* **6**, 135–140.

Read, R. J. & James, M. N. G. (1988). *J. Mol. Biol.* **200**, 523–551.

Rossmann, M. G. (1972). *The Molecular Replacement Method.* New York: Gordon & Breach.

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

Sheriff, S., Klei, H. E. & Davis, M. E. (1999). *J. Appl. Cryst.* **32**, 98–101.

Shmueli, U. & Weiss, G. H. (1995). *Introduction to Crystallographic Statistics.* Oxford University Press.

Shmueli, U., Weiss, G. H., Kiefer, J. E. & Wilson, A. J. C. (1984). *Acta Cryst.* A**40**, 651–660.

Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.

Srinivasan, R. & Chandrasekaran, R. (1966). *Indian J. Pure Appl. Phys.* **4**, 178–186.

Tsoucaris, G. (1970). *Acta Cryst.* A**26**, 492–499.

Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.

Wooding, R. A. (1956). *Biometrika*, **43**, 212–215.

Woolfson, M. M. (1956). *Acta Cryst.* **9**, 804–810.

Yano, J. K., Koo, L. S., Schuller, D. J., Li, H., Ortiz de Montellano, P. R. & Poulos, T. L. (2000). *J. Biol. Chem.* **275**, 31086–31092.