Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

New ways of looking at experimental phasing

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 14 April 2003
Accepted 11 August 2003

In the original work by Blow and Crick, experimental phasing was formulated as a least-squares problem. For good data on good derivatives this approach works reasonably well, but we now attempt to extract more information from poorer data than in the past. As in many other crystallographic problems, the assumptions underlying the use of least squares for phasing are not satisfied, particularly for poor derivatives. The introduction of maximum likelihood (and more powerful computers) has led to substantial improvements. For computational convenience, these new methods still make many assumptions about the independence of different measurements and sources of error. A more general formulation for the probability distributions underlying likelihood-based methods for both experimental phasing and molecular-replacement phasing is reviewed. In the new formulation, all the structure factors associated with a particular *hkl* are considered to be related by a complex multivariate normal distribution. When it is assumed that certain errors are independent, the general formulation reduces to current likelihood targets. However, the new formulation makes the necessary assumptions more explicit and points the way to improving phasing using both isomorphous and anomalous differences.

1. Experimental phasing before likelihood

The history of experimental phasing in macromolecular crystallography goes back over 40 years. In the 1950s, Perutz and colleagues realised that the isomorphous replacement method that had been used for some time to determine phases in small-molecule crystallography could also be applied to proteins (Green *et al.*, 1954; Perutz, 1956). It was quickly apparent that the errors would be comparable to the signals and that a statistical treatment would be required. Blow & Crick (1959) proposed a least-squares treatment for the determination of phases and argued that the combined effect of heavy-atom model errors and measurement errors for both native (F_P) and the derivative (F_{PH}) structure-factor amplitudes could be approximated by attributing the combined error to the measurement of F_{PH}. In retrospect, this is a reasonable approximation for the case of single isomorphous replacement (SIR). For practical reasons, severe approximations were initially applied in practice: the combined error (called the lack-of-closure error) was estimated at only the single best phase angle, without taking phase ambiguity into account, and the SIR treatment was generalized by pairing each derivative in turn with the native measurement and then treating the resultant phase probabilities as independent.

As new statistical ideas have been introduced into crystallography and computers have become more powerful, approaches to experimental phasing have become more sophisticated. Raiz & Andreeva (1970) showed that phases could be estimated more accurately if the native amplitude was just treated as another observation and not paired with each derivative. Einstein (1977) pointed out that the practice of combining phase information from different derivatives by multiplying SIR phase-probability distributions gave too much weight to the measurement of the native amplitude. Otwinowski (1991) showed that it was important to average the lack-of-closure error over all possible phase angles during phasing and refinement of heavy-atom positions. As shown below, all of these considerations plus more are satisfied when experimental phasing is carried out by maximum-likelihood methods (Bricogne, 1991; Read, 1991, 1994; de La Fortelle & Bricogne, 1997).

2. The principle of maximum likelihood

Whenever we fit a model to data, we want the model to account as well as possible for the data. In maximum likelihood, the consistency of the model with the data is assessed by the likelihood function, defined as the probability of making the set of measurements given the model. The basic concept of likelihood is fairly intuitive, as McCoy (2002) illustrates with a simple thought experiment using dice.

Likelihood can be justified in terms of Bayes' theorem, which tells us that the probability of the model given the data (what we really want to measure) is proportional to the probability of the data given the model (*i.e.* the likelihood function) multiplied by the prior probability of the model. According to this interpretation, Bayesian reasoning justifies the use in structure refinement of geometric restraints, which reflect the plausibility (prior probability) of the atomic model.

Strictly, the likelihood function should be the joint probability distribution for the entire set of observations. It is often possible to assume that the observations are independent, so that the likelihood function becomes the product of all the probabilities of the individual observations. This is the case for likelihood-based structure refinement (Pannu & Read, 1996; Bricogne & Irwin, 1996; Murshudov *et al.*, 1997). For experimental phasing, the phase information comes from correlations among the structure factors with the same *hkl* from different crystals, at different wavelengths or with opposite hand, so it is necessary to consider the joint distributions of these structure factors. Correlations are much weaker among reflections with different *hkl* indices, so they are still assumed to be independent in experimental phasing. Nonetheless, these weak correlations are the basis of phase improvement by density-modification methods (reviewed by Kleywegt & Read, 1997) and could in principle be exploited directly in more sophisticated phasing strategies (Bricogne, 1993).

2.1. Probabilistic background: structure-factor probabilities

The likelihood function is the probability distribution for the observations, which are intensities or structure-factor amplitudes in crystallography. To define a likelihood function, it is necessary to understand the sources of error in predicting the observations from the model and how they propagate. To be of practical use, a likelihood function must also be defined in a form that can be computed effectively.

Crystallographic phasing methods all turn out to involve essential steps where related structure factors are compared: model with observed, native with derivative or derivative with calculated heavy-atom contribution. Information about one structure factor comes from its correlations with another structure factor. This information is statistical, so that it is necessary to consider the underlying probability distributions. As outlined in previous publications (reviewed in Read, 1997), the correlation between two structure factors will depend on the size of the common substructure (how many atoms the structures share) and the size of the coordinate differences between the substructures.

We are often interested in the conditional distribution of one structure factor (say $\mathbf{F}_1$) given another ($\mathbf{F}_2$). A proportion (D) of $\mathbf{F}_2$ will be correlated to $\mathbf{F}_1$, so that $D\mathbf{F}_2$ is the expected value of $\mathbf{F}_1$ given $\mathbf{F}_2$. D will tend to decrease with resolution, as coordinate differences become larger relative to the Bragg spacing. For differences between structure factors that are the sums of large numbers of atomic contributions, the central limit theorem will apply and we can assume that the distribution of $\mathbf{F}_1$ is a complex Gaussian or, equivalently, a symmetric two-dimensional Gaussian in the complex plane for an acentric structure factor or a one-dimensional Gaussian for a centric structure factor (Read, 1990). (For simplicity, the discussion below will concentrate on acentric structure factors. Similar reasoning applies to centric structure factors.) The conditional distribution for an acentric structure factor is given by

$$p(\mathbf{F}_1; \mathbf{F}_2) = \frac{1}{\pi\varepsilon\sigma_\Delta^2}\exp\left(-\frac{|\mathbf{F}_1 - D\mathbf{F}_2|^2}{\varepsilon\sigma_\Delta^2}\right). \qquad (1)$$

In this equation, σ_Δ^2 is the complex variance term that accounts for the combined effects of missing atoms and differences in the common substructure and ε accounts for the statistical effects of symmetry on expected intensities. (A complex variance is the expected value of the square of the magnitude of the deviation in the complex plane; it is equivalent to the sum of the variances for the real and imaginary components of a complex number, assuming that the real and imaginary components are independent and have equal variances.) Fig. 1(*a*) shows a schematic representation of the distribution in (1). This distribution applies, of course, to the phased structure factor, whereas the observations are unphased amplitudes.

To obtain the probability distribution of the amplitudes, it is necessary to integrate over the unknown phase angle to obtain (2), known as the Rice distribution in statistical literature, but also occurring frequently in crystallographic literature (*e.g.* Sim, 1959),

$$p(|\mathbf{F}_1|;\mathbf{F}_2)=\frac{2|\mathbf{F}_1|}{\varepsilon\sigma_\Delta^2}\exp\left(-\frac{|\mathbf{F}_1|^2+D^2|\mathbf{F}_2|^2}{\varepsilon\sigma_\Delta^2}\right)I_0\left(\frac{2|\mathbf{F}_1|D|\mathbf{F}_2|}{\varepsilon\sigma_\Delta^2}\right). \quad (2)$$

As discussed below, the considerations that apply to pairs of structure factors can be generalized to collections of structure factors. This will be required for the most general possible treatments of phasing by isomorphous replacement and anomalous dispersion.

2.2. Other applications of likelihood in crystallography

One early use of likelihood in crystallography was to estimate the σ_A parameter of phase-probability distributions (Lunin & Urzhumtsev, 1984; Read, 1986), adjusting the value of σ_A for each resolution shell to maximize the likelihood of observing the structure-factor amplitudes in that shell given the structure factors calculated from a model. The probability distributions can be used to combine phase information from different sources or to compute maps that reduce model bias (Read, 1986, 1997). Essentially, the same likelihood function is used in the maximum-likelihood refinement of protein structures (Pannu & Read, 1996; Bricogne & Irwin, 1996; Murshudov *et al.*, 1997). Recently, I have followed up on a suggestion by Bricogne (1992) to use likelihood as a target for molecular replacement, implementing a new rotation likelihood function and using multivariate statistics (as discussed below) to allow for multiple models (Read, 2001). Experiences with the initial implementation in the program *Beast* confirm that likelihood is a more sensitive target than Patterson overlap or correlation scores for difficult molecular-replacement problems (Read, 2003).

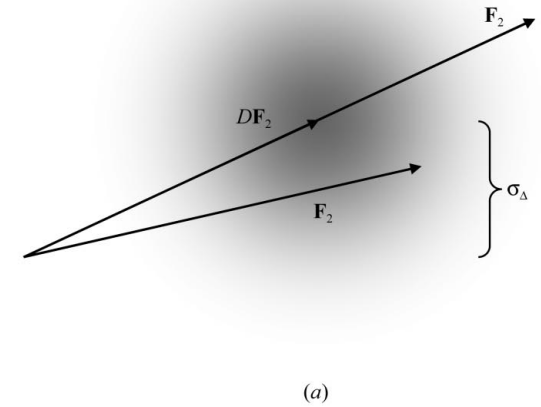
3. Current likelihood-based approaches to experimental phasing

The theoretical background for the application of likelihood to experimental phasing was outlined in the early 1990s (Bricogne, 1991; Read, 1991, 1994), leading to the development of the program *SHARP* (de La Fortelle & Bricogne, 1997). Although the likelihood functions used for phasing by isomorphous replacement initially appear daunting, we will see that they have an intuitive relationship to the familiar Harker (1956) construction.

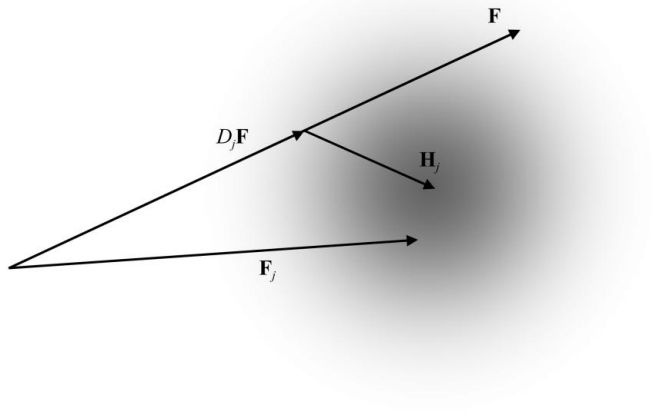
A likelihood function should be the joint probability distribution for the entire set of measurements as a function of the parameters being optimized. In the case of isomorphous replacement phasing, the observations are structure-factor amplitudes for all crystals and there are parameters describing the heavy-atom substructures (coordinates and B factors) as well as the lack-of-isomorphism errors. As discussed above, we can assume to a good approximation that structure factors for different hkls are independent, but the phase information comes from correlations among structure factors with the same hkl from native and derivative crystals. The required likelihood function is thus a product over all hkl indices of joint distributions for structure-factor amplitudes from all N crystals, which is obtained by integrating over all possible phases of phased structure factors,

$$\begin{aligned}L&=\prod_{hkl}p(|\mathbf{F}_0|,|\mathbf{F}_1|\ldots|\mathbf{F}_N|;\mathbf{H}_0,\mathbf{H}_1\ldots\mathbf{H}_N)\\&=\prod_{hkl}\int_0^{2\pi}\ldots\int_0^{2\pi}\int_0^{2\pi}p(|\mathbf{F}_0|,\alpha_0,|\mathbf{F}_1|,\alpha_1\ldots|\mathbf{F}_N|,\alpha_N;\\&\qquad\mathbf{H}_0,\mathbf{H}_1\ldots\mathbf{H}_N)\,\mathrm{d}\alpha_0\,\mathrm{d}\alpha_1\ldots\,\mathrm{d}\alpha_N. \quad (3)\end{aligned}$$

In (3), the index zero refers to the native crystal. In principle, all crystals may have associated heavy-atom models, from which structure factors $\mathbf{H}_j$ can be computed, but we will



(*a*)



(*b*)

Figure 1
Schematic representation of conditional probability distributions of structure factors. (*a*) Probability distribution (represented by grey shading) of structure factor $\mathbf{F}_1$ given a related structure factor $\mathbf{F}_2$. The distribution, given in (1), is centred on $D\mathbf{F}_2$, where D represents the fraction of $\mathbf{F}_2$ that is correlated with $\mathbf{F}_1$. The width of the distribution is described by the parameter σ_Δ. (*b*) Probability distribution (represented by grey shading) of $\mathbf{F}_j$ (structure factor for derivative j) given an assumed true native structure factor $\mathbf{F}$ and the structure-factor contribution $\mathbf{H}_j$ computed from a heavy-atom model. The distribution, given in (5), is centered on $D_j\mathbf{F}+\mathbf{H}_j$. The width of the distribution includes contributions from lack of isomorphism, errors in the heavy-atom model and measurement error.

assume for now that there is no heavy-atom model associated with the native crystal, so that $\mathbf{H}_0$ is zero.

Because the native and derivative crystals share their protein components, the native and derivative structure factors (and their phases) are highly correlated, which makes it extremely difficult to integrate over the multiple phases. Fortunately, it is often possible to remove the correlations, at the expense of introducing a dummy variable that must be eliminated by integration. It turns out that the *conditional* probabilities of the structure factors are independent if we make a number of reasonable assumptions. Firstly, we assume that we know the true value of the native structure factor. (In fact, its assumed value is the dummy variable that must later be eliminated.) Given the native and heavy-atom structure-factor contributions, the uncertainties in predicting the various structure factors come from errors in the heavy-atom models, lack of isomorphism and measurement error. If these sources of error are independent, the conditional probabilities of the individual structure factors are independent. Each phase can then be integrated out from the independent distribution, as performed for (2).

The dummy variable $\mathbf{F}$, which can be interpreted as the true value of the native structure factor, is introduced into an expanded joint probability distribution. (In fact, if $\mathbf{H}_0$ were not zero, $\mathbf{F}$ would be interpreted as the part of the native structure factor not accounted for by $\mathbf{H}_0$.) The original joint probability distribution of observed structure factors is then a marginal distribution that can be obtained by integrating over all possible values of $\mathbf{F}$ in the expanded distribution. The expanded distribution can in turn be expressed as the product of the prior distribution of $\mathbf{F}$ (a Wilson distribution; Wilson, 1949) and the distribution of the observed structure factors conditional on $\mathbf{F}$. Because the conditional distributions of the individual structure factors are independent, this becomes a product of conditional distributions,

$$\begin{aligned} p(&\mathbf{F}_0, \mathbf{F}_1 \ldots \mathbf{F}_N; \mathbf{H}_0, \mathbf{H}_1 \ldots \mathbf{H}_N) \\ &= \int p(\mathbf{F}, \mathbf{F}_0, \mathbf{F}_1 \ldots \mathbf{F}_N; \mathbf{H}_0, \mathbf{H}_1 \ldots \mathbf{H}_N)\ \mathrm{d}\mathbf{F} \\ &= \int p(\mathbf{F}) p(\mathbf{F}_0, \mathbf{F}_1 \ldots \mathbf{F}_N; \mathbf{F}, \mathbf{H}_0, \mathbf{H}_1 \ldots \mathbf{H}_N)\ \mathrm{d}\mathbf{F} \\ &= \int p(\mathbf{F}) \prod_{j=0}^{N} p(\mathbf{F}_j; \mathbf{F}, \mathbf{H}_j)\ \mathrm{d}\mathbf{F}. \end{aligned} \tag{4}$$

In (4), each derivative structure factor is the sum of a protein component and a heavy-atom component. Because of lack of isomorphism, only a proportion (D_j) of the native structure factor will be correlated with the protein component of the derivative structure factor, as discussed for equation (1). A similar factor is required in principle to account for the proportion of the calculated heavy-atom structure factor that is correlated with the true heavy-atom contribution, but we can assume that the variation of this factor with resolution will be modelled by the maximum-likelihood refinement of heavy-atom occupancies and B factors (Read, 1991). The distribution of $\mathbf{F}_j$ will thus be centred on $(D_j\mathbf{F} + \mathbf{H}_j)$ and the variance of this probability distribution will include contributions from lack-of-isomorphism errors and heavy-atom model errors. To a good approximation, error in the measurement of the amplitude $|\mathbf{F}_j|$ can be accounted for by increasing the complex variance (Green, 1979; Murshudov *et al.*, 1997; de La Fortelle & Bricogne, 1997), giving

$$p(\mathbf{F}_j; \mathbf{F}, \mathbf{H}_j) = \frac{1}{\pi(\varepsilon\sigma_\Delta^2 + \sigma_j^2)} \exp\left[-\frac{|\mathbf{F}_j - (D_j\mathbf{F} + \mathbf{H}_j)|^2}{\varepsilon\sigma_\Delta^2 + \sigma_j^2} \right]. \tag{5}$$

Fig. 1(*b*) shows a schematic representation of the distribution in (5). To obtain the probability distribution of the measured amplitude, required for (3), the unknown phase is integrated out, giving

$$\begin{aligned} p(|\mathbf{F}_j|; \mathbf{F}, \mathbf{H}_j) = &\frac{2|\mathbf{F}_j|}{\varepsilon\sigma_\Delta^2 + \sigma_j^2} \exp\left(-\frac{|\mathbf{F}_j|^2 + |D_j\mathbf{F} + \mathbf{H}_j|^2}{\varepsilon\sigma_\Delta^2 + \sigma_j^2} \right) \\ &\times I_0\left(\frac{2|\mathbf{F}_j||D_j\mathbf{F} + \mathbf{H}_j|}{\varepsilon\sigma_\Delta^2 + \sigma_j^2} \right). \end{aligned} \tag{6}$$

(3), (4) and (6) are combined to obtain

$$L = \prod_{hkl} \int p(\mathbf{F}) \prod_{j=0}^{N} p(|\mathbf{F}_j|; \mathbf{F}, \mathbf{H}_j)\ \mathrm{d}\mathbf{F}. \tag{7}$$

Finally, we note that the log of the likelihood has its maximum at the same point as the likelihood itself. Since it is much more convenient to deal with a sum than a product of small numbers, we maximize the log of the likelihood,

$$\mathrm{LL} = \ln L = \sum_{hkl} \ln\left[\int p(\mathbf{F}) \prod_{j=0}^{N} p(|\mathbf{F}_j|; \mathbf{F}, \mathbf{H}_j)\ \mathrm{d}\mathbf{F} \right]. \tag{8}$$

(7) and (8) can be understood intuitively in terms of the Harker (1956) construction. In the Harker construction, each circle represents complex values of the native structure factor that are consistent with one of the observed amplitudes. Derivative circles are offset from the origin by the negative of the heavy-atom structure-factor contribution, because the derivative structure factor should be the sum of the heavy-atom contribution plus the native protein contribution. If there are no errors, all circles intersect uniquely at the native structure factor. Similarly, each of the conditional probability distributions in the product in (8) is a function of the assumed native structure factor. Plotted on an Argand diagram, they are circularly symmetric distributions offset from the origin by a vector related to the negative of the heavy-atom contribution (Fig. 2). The width of each distribution reflects the combined effect of measurement error, lack of isomorphism and errors in the heavy-atom model. When the distributions are multiplied together, peaks in the resulting function represent plausible values for the native structure factor.

The likelihood function is the integral over $\mathbf{F}$, which is the volume under the surface defined by the product of the circular distributions and the prior distribution for $\mathbf{F}$. The effect of the prior distribution is neglected by de La Fortelle & Bricogne (1997). Although this is not strictly correct, $p(\mathbf{F})$ varies slowly over regions of interest so it will have relatively little impact (Read, 1991). However, if the native crystal were known to contain atoms with significant scattering at the positions of heavy atoms in the derivative crystals, the infor-

mation from this should be reflected in the prior distribution of **F**, which should be offset from the origin. Depending on the relative scattering of these atoms, the effect could be significant.

It is interesting to consider what happens as the likelihood function is maximized as a function of the parameters. The heavy-atom parameters will change the values of $\mathbf{H}_j$, which will move the centres of the circular distributions. The likelihood will be maximized when these circular distributions overlap optimally, as shown in Fig. 3. In the presence of errors, however, not all of the circles for all reflections can be made to overlap. The lack-of-isomorphism variance will broaden the

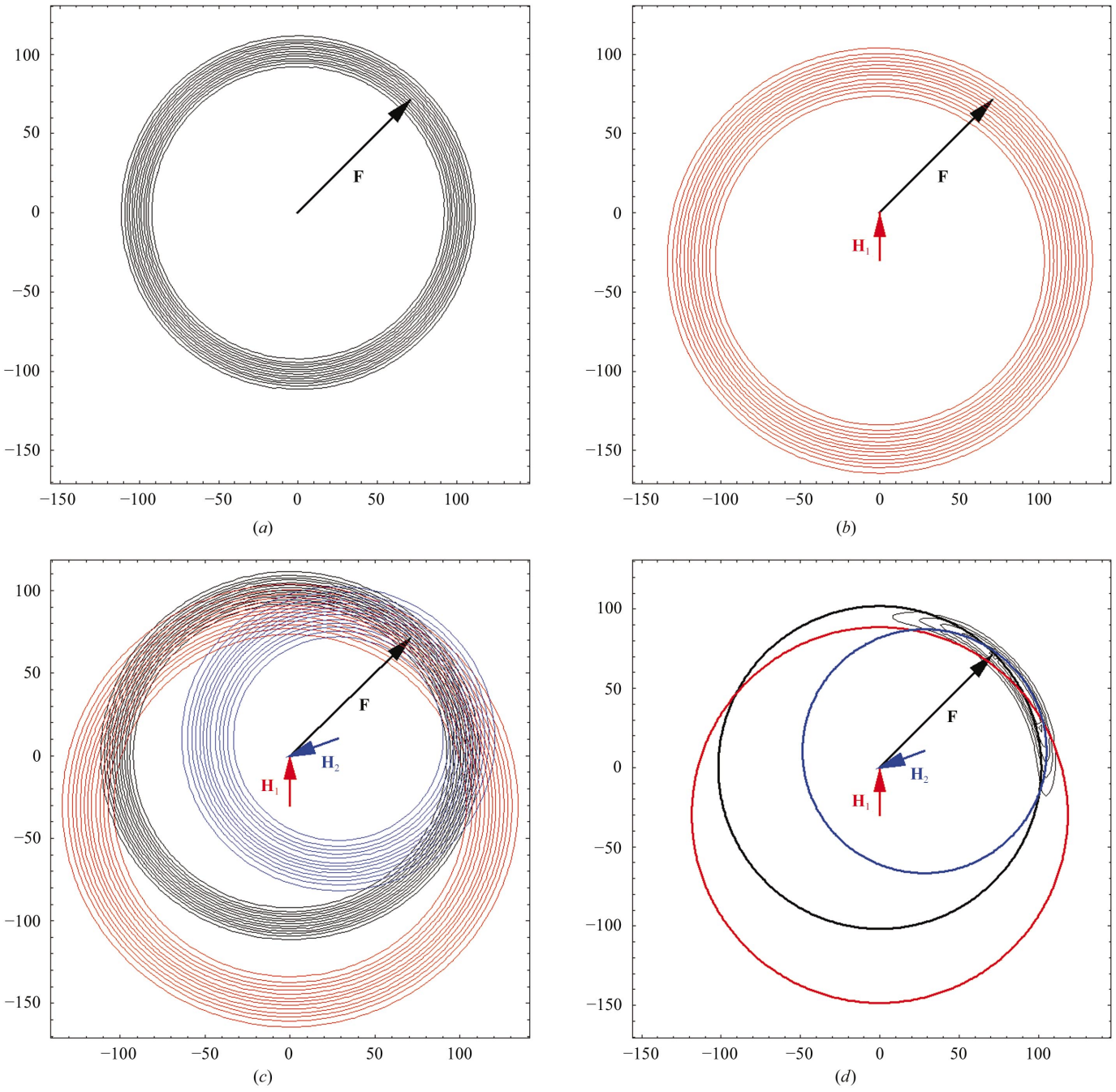


Figure 2
Schematic representation of the interpretation of (7) as a probabilistic version of the Harker (1956) construction. Each panel shows a contour plot of a probability distribution. (*a*) shows the probability distribution for the observed native amplitude as a function of possible values for the true native structure factor. The width of this distribution represents the estimated error in measurement of the native amplitude. (*b*) shows the probability distribution for the observed amplitude for derivative 1, again as a function of possible values for the true native structure factor. Because the derivative structure factor must be the sum of the heavy-atom contribution and the protein contribution, the centre of the circularly symmetric distribution is offset by minus the heavy-atom contribution. (*c*) overlays the first two distributions, as well as the distribution for a second derivative, and (*d*) shows the product of all three distributions as a function of the assumed true native structure factor. The combined distribution has significant values where the circles from the Harker construction come close to intersecting.

distributions to the degree required to make them overlap significantly. Fig. 4 shows how the likelihood varies with the lack-of-isomorphism variance for a derivative. Animated versions of Figs. 3 and 4 have been deposited as supplementary material.[1]

As noted above, the maximum-likelihood treatment of MIR refinement and phasing circumvents the problems of earlier treatments. The measurement of the native amplitude is just treated as another measurement and no longer plays a privileged role and the estimation of the lack-of-isomorphism error automatically considers all possible choices for the true phase.

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: ba5044). Services for accessing this material are described at the back of the journal.

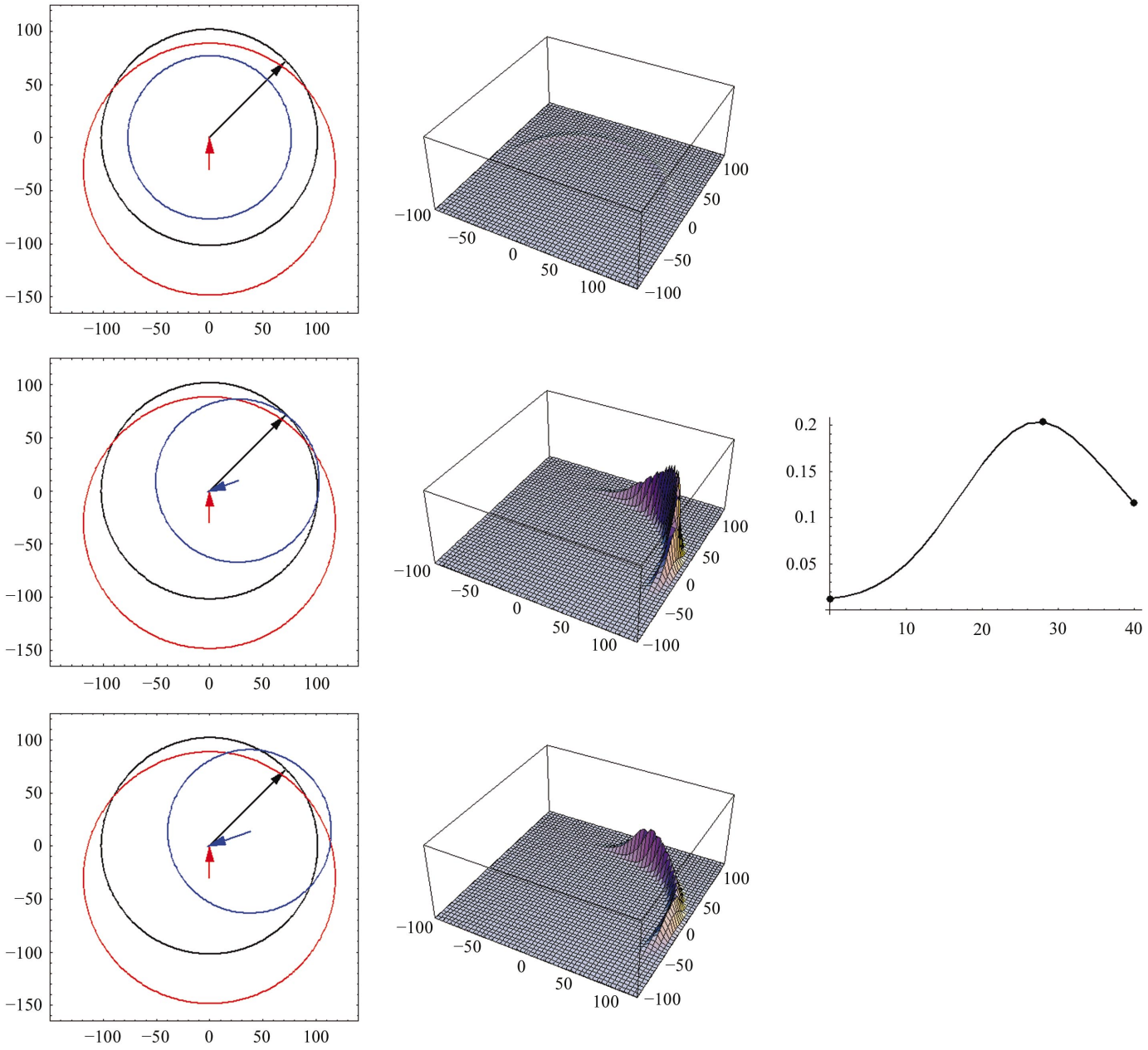


Figure 3
Representation of likelihood as function of heavy-atom contribution to the structure factor for one derivative. The panels on the left show three possible choices of the heavy-atom contribution. Since a change in the heavy-atom contribution shifts the centre of the probability distribution for the derivative amplitude (see Fig. 2), it changes the amount of overlap of the distributions from different crystals. The combined distribution from the native and two derivative amplitudes (multiplied by the prior probability of the true structure factor, as in equation 7) is shown as a surface plot in the three middle panels. The likelihood is the volume under this surface and the panel on the right shows the likelihood as a function of the size of the heavy-atom contribution, with dots highlighting the values corresponding to the three pairs of plots to the left. An animated version of this figure is available as supplementary material[1].

4. A new view: multivariate structure-factor distributions

In the derivation of a likelihood function for isomorphous replacement above, multivariate distributions are avoided by casting the expression in terms of a product of independent conditional univariate distributions. However, it is possible to obtain the same likelihood function starting from a multivariate distribution (Pannu *et al.*, 2003). This approach has the advantage of making the necessary assumptions clearer. It also points the way towards new likelihood targets that require fewer assumptions.

Complex normal distributions of single variables, such as the distribution of a single structure factor in (1), can be generalized to the multivariate case (Wooding, 1956). The central limit theorem justifies applying such distributions to collections of structure factors that are sums of many small atomic contributions (Tsoucaris, 1970). The parameters of a multivariate complex normal distribution of structure factors

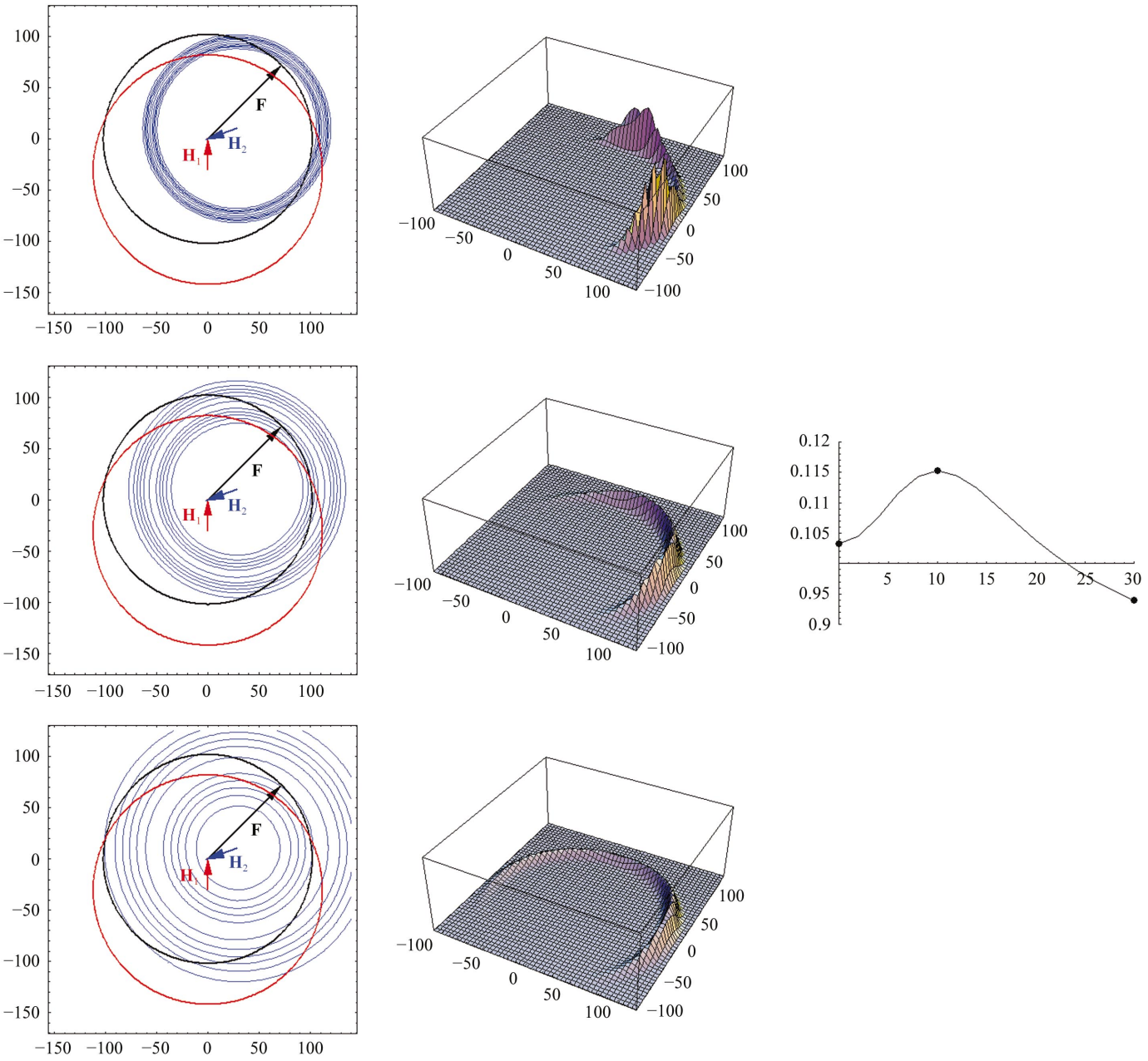


Figure 4
Representation of likelihood as a function of assumed lack-of-isomorphism error for one derivative. The panels on the left show probability distributions for three possible choices of lack-of-isomorphism error for the derivative, which changes the width of the distributions. The combined distribution from the native and two derivative amplitudes (multiplied by the prior probability of the true structure factor, as in equation 7) is shown as a surface plot in the three middle panels. The likelihood is the volume under this surface and the panel on the right shows the likelihood as a function of the lack-of-isomorphism error, with dots highlighting the values corresponding to the three pairs of plots to the left. Likelihood is maximized when the distribution is just broad enough to overlap with the other distributions. An animated version of this figure is available as supplementary material.

$\mathbf{F}$ are their expected values, $\langle\mathbf{F}\rangle$, and the elements σ_{ij} of the covariance matrix $\boldsymbol{\Sigma}$,

$$p(\mathbf{F}) = \frac{1}{|\pi\boldsymbol{\Sigma}|}\exp[-(\mathbf{F}-\langle\mathbf{F}\rangle)^H\boldsymbol{\Sigma}^{-1}(\mathbf{F}-\langle\mathbf{F}\rangle)]$$
$$\sigma_{ij} = \left\langle(\mathbf{F}_i-\langle\mathbf{F}_i\rangle)(\mathbf{F}_j-\langle\mathbf{F}_j\rangle)^*\right\rangle. \quad (9)$$

In this equation, $(\mathbf{F}-\langle\mathbf{F}\rangle)$ is a column vector and the superscript H indicates its Hermitian transpose (a row vector of complex conjugates). Note that σ_{ij} in this notation refers to a covariance element and not to a standard deviation.

As shown in the context of molecular replacement (Read, 2001), it can be useful to start from an extended joint distribution of observed and calculated structure factors. In this case, nothing is yet known about the structure factors, so that their expected values are zero and the covariance elements are given by $\sigma_{ij} = \langle\mathbf{F}_i\mathbf{F}_j^*\rangle$. When the calculated structure factors are fixed, the information they provide is reflected in non-zero expected values and reduced covariance elements.

The covariance elements will be dominated by contributions from atoms shared between the crystals, because terms relating unmatched atoms will tend to cancel (Read, 1990),

$$\mathbf{F}_i = \sum_{k=1}^{N_A} f_{ik}\exp(2\pi i\mathbf{h}\cdot\mathbf{x}_{ik})$$
$$\langle\mathbf{F}_i\mathbf{F}_j^*\rangle \simeq \sum_{\text{common atoms}} f_{ik}f_{jk}\langle\exp[2\pi i\mathbf{h}\cdot(\mathbf{x}_{ik}-\mathbf{x}_{jk})]\rangle. \quad (10)$$

Note that the covariances can be complex in general, but will be assumed to be real in the case of isomorphous replacement, where the scattering factors are real and there is no reason to expect a systematic phase shift. In the case of anomalous dispersion, the scattering factors are complex and the presence of an imaginary component will lead to imaginary terms in the off-diagonal covariance elements (Pannu *et al.*, 2003). The expected value of the phase-shift term (the exponential in equation 10) is like a figure of merit, so it will typically be less than one. For convenience, we will define D values below to describe the overall effective value of this phase-shift term and express the covariances in terms of the Wilson distribution parameter for the shared atoms. When the atomic scattering factors from corresponding atoms in the two crystals are identical, these D values are related to the Fourier transform of coordinate differences (Read, 1990).

If we assume that the crystals have independent lack-of-isomorphism errors, that there are no sites in common among derivatives and that there are no significant scatterers occupying the same position in the native crystal as in the heavy-atom sites (*i.e.* isomorphous addition rather than isomorphous replacement), then we find that when the heavy-atom contributions are fixed, the new conditional covariance matrix has no off-diagonal elements (Pannu *et al.*, 2003). This means that the joint distribution of observed structure factors (conditional on the heavy-atom model contributions) can be factored into a product of independent distributions of single structure factors, giving the same result as obtained above.

In fact, it is not necessary to make such stringent assumptions in order to obtain such a simple result. We still obtain a product of independent conditional distributions (or, equivalently, a diagonal covariance matrix) as long as the errors in the models are independent, regardless of whether there are common sites or heavy atoms replacing significant scatterers in the native crystal. In the following, we will see that the assumption of independent errors implies relationships among the elements of the covariance matrix for the observed and calculated structure factors. Defining these relationships will force us to be explicit about which errors are independent, but will then allow us to factor out common submatrices and drastically simplify the covariance matrix for the conditional distribution of observed structure factors.

We start from a joint distribution of the observed structure factors, the dummy structure factor and the calculated heavy-atom structure factors, $p(\mathbf{F}_0, \mathbf{F}_1 \ldots \mathbf{F}_N, \mathbf{F}, \mathbf{H}_0, \mathbf{H}_1 \ldots \mathbf{H}_N)$. In this joint distribution, there is no prior information before fixing the heavy-atom models, so all expected values are zero. The covariance matrix is given by (11), where it is partitioned into submatrices that will be manipulated when the conditional variables are fixed (Johnson & Wichern, 1998),

$$\boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}, \quad (11a)$$

$$\boldsymbol{\Sigma}_{11} = \begin{pmatrix}\langle\mathbf{F}_0\mathbf{F}_0^*\rangle & \langle\mathbf{F}_0\mathbf{F}_1^*\rangle & \cdots & \langle\mathbf{F}_0\mathbf{F}_N^*\rangle\\ \langle\mathbf{F}_1\mathbf{F}_0^*\rangle & \langle\mathbf{F}_1\mathbf{F}_1^*\rangle & \cdots & \langle\mathbf{F}_1\mathbf{F}_N^*\rangle\\ \vdots & \vdots & \ddots & \vdots\\ \langle\mathbf{F}_N\mathbf{F}_0^*\rangle & \langle\mathbf{F}_N\mathbf{F}_1^*\rangle & \cdots & \langle\mathbf{F}_N\mathbf{F}_N^*\rangle\end{pmatrix}, \quad (11b)$$

$$\boldsymbol{\Sigma}_{12} = \begin{pmatrix}\langle\mathbf{F}_0\mathbf{F}^*\rangle & \langle\mathbf{F}_0\mathbf{H}_0^*\rangle & \langle\mathbf{F}_0\mathbf{H}_1^*\rangle & \cdots & \langle\mathbf{F}_0\mathbf{H}_N^*\rangle\\ \langle\mathbf{F}_1\mathbf{F}^*\rangle & \langle\mathbf{F}_1\mathbf{H}_0^*\rangle & \langle\mathbf{F}_1\mathbf{H}_1^*\rangle & \cdots & \langle\mathbf{F}_1\mathbf{H}_N^*\rangle\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \langle\mathbf{F}_N\mathbf{F}^*\rangle & \langle\mathbf{F}_N\mathbf{H}_0^*\rangle & \langle\mathbf{F}_N\mathbf{H}_1^*\rangle & \cdots & \langle\mathbf{F}_N\mathbf{H}_N^*\rangle\end{pmatrix},$$

$$\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^H = \boldsymbol{\Sigma}_{12}^T, \quad (11c)$$

$$\boldsymbol{\Sigma}_{22} = \begin{pmatrix}\langle\mathbf{F}\mathbf{F}^*\rangle & \langle\mathbf{F}\mathbf{H}_0^*\rangle & \langle\mathbf{F}\mathbf{H}_1^*\rangle & \cdots & \langle\mathbf{F}\mathbf{H}_N^*\rangle\\ \langle\mathbf{H}_0\mathbf{F}^*\rangle & \langle\mathbf{H}_0\mathbf{H}_0^*\rangle & \langle\mathbf{H}_0\mathbf{H}_1^*\rangle & \cdots & \langle\mathbf{H}_0\mathbf{H}_N^*\rangle\\ \langle\mathbf{H}_1\mathbf{F}^*\rangle & \langle\mathbf{H}_1\mathbf{H}_0^*\rangle & \langle\mathbf{H}_1\mathbf{H}_1^*\rangle & \cdots & \langle\mathbf{H}_1\mathbf{H}_N^*\rangle\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \langle\mathbf{H}_N\mathbf{F}^*\rangle & \langle\mathbf{H}_N\mathbf{H}_0^*\rangle & \langle\mathbf{H}_N\mathbf{H}_1^*\rangle & \cdots & \langle\mathbf{H}_N\mathbf{H}_N^*\rangle\end{pmatrix}. \quad (11d)$$

To clarify the relationships involving the crystals and their corresponding heavy-atom models, we use f and g to represent atomic scattering factors and $\mathbf{x}$ and $\mathbf{y}$ to represent coordinates for the corresponding crystals and models, so that the observed and calculated structure factors are defined as

$$\mathbf{F}_i = \sum_{k=1}^{N_A} f_{ik}\exp(2\pi i\mathbf{h}\cdot\mathbf{x}_{ik}),$$
$$\mathbf{H}_i = \sum_{k=1}^{N_A} g_{ik}\exp(2\pi i\mathbf{h}\cdot\mathbf{y}_{ik}). \quad (12)$$

For simplicity, each crystal or model is considered to contain all the atoms that are present in any crystal or model, but assigning zero scattering factors to atoms that do not exist in that crystal or model. The dummy structure factor $\mathbf{F}$ is inter-

preted as the true native structure factor, minus the contribution of any atoms represented by $\mathbf{H}_0$. If we number the atoms from 1 to N_L for the 'light' atoms that are not present in any of the heavy-atom models, then from $N_L + 1$ to N_A for atoms in the heavy-atom models, we obtain

$$\mathbf{F} = \sum_{k=1}^{N_L} f_{0k} \exp(2\pi i \mathbf{h} \cdot \mathbf{x}_{0k}),$$
$$\langle \mathbf{F}\mathbf{F}^* \rangle = \sum_{k=1}^{N_L} f_{0k}^2 = \Sigma_{N_L}. \quad (13)$$

(For simplicity, the expected intensity factor ε that should appear in all covariance elements is omitted in equation 13 and subsequent equations.)

We will assume that heavy-atom model 0 contains all atoms in any of the heavy-atom models that replace atoms with significant scattering in the native crystal. **F** will then be uncorrelated to any of the heavy-atom models, so that the off-diagonal elements in the first column and row of $\mathbf{\Sigma}_{22}$ will all be zero. Off-diagonal elements of this submatrix relating heavy-atom models will be given by (14). In this and subsequent equations it is assumed that differences in heavy-atom positions are uncorrelated with differences in their scattering factors.

$$\begin{aligned}\langle \mathbf{H}_i \mathbf{H}_j^* \rangle &= \sum_{k=N_L+1}^{N_A} g_{ik} g_{jk} \langle \exp[2\pi i \mathbf{h} \cdot (\mathbf{y}_{ik} - \mathbf{y}_{jk})] \rangle \\ &= D_{H_{ij}} \sum_{k=N_L+1}^{N_A} g_{ik} g_{jk} \\ &= D_{H_{ij}} \Sigma_{H_{ij}}. \end{aligned} \quad (14)$$

In (14), the sum runs over all atoms present in any heavy-atom model, but only common sites contribute because atoms that are not present in a particular model are assigned an atomic scattering factor of zero. The parameter $D_{H_{ij}}$ is the effective overall value of the phase shift term arising from coordinate differences between common sites in the heavy-atom models and $\Sigma_{H_{ij}}$ accounts for the extent of overlap between heavy-atom models. If there are no common sites, then $\Sigma_{H_{ij}}$ will be zero. For the diagonal elements, $D_{H_{ii}} = 1$ and we use Σ_{H_i} to represent $\Sigma_{H_{ii}}$. A symbolic expression for $\mathbf{\Sigma}_{22}$ is given in (15), summarizing the results above,

$$\mathbf{\Sigma}_{22} = \begin{pmatrix} \Sigma_{N_L} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_{22} \end{pmatrix},$$
$$\mathbf{T}_{22} = \begin{pmatrix} \Sigma_{H_0} & D_{H_{01}}\Sigma_{H_{01}} & \cdots & D_{H_{0N}}\Sigma_{H_{0N}} \\ D_{H_{01}}\Sigma_{H_{01}} & \Sigma_{H_1} & \cdots & D_{H_{1N}}\Sigma_{H_{1N}} \\ \vdots & \vdots & \ddots & \vdots \\ D_{H_{0N}}\Sigma_{H_{0N}} & D_{H_{1N}}\Sigma_{H_{1N}} & \cdots & \Sigma_{H_N} \end{pmatrix}. \quad (15)$$

The submatrix $\mathbf{\Sigma}_{12}$ contains cross-terms relating the observed structure factors to the structure factors that will be fixed in the conditional distribution. Cross-terms involving the dummy structure factor **F** are simple,

$$\begin{aligned}\langle \mathbf{F}_i \mathbf{F}^* \rangle &= \sum_{k=1}^{N_L} f_{ik} f_{0k} \langle \exp[2\pi i \mathbf{h} \cdot (\mathbf{x}_{ik} - \mathbf{x}_{0k})] \rangle \\ &= D_i \Sigma_{N_L}. \end{aligned} \quad (16)$$

Since **F** is considered to be the unmodelled part of the true native structure factor, D_0 is equal to one. D_i values for other crystals provide an overall effective value for the phase-shift term, absorbing any systematic differences in scale factor or overall B factor.

Cross-terms involving the heavy-atom structure factors are given by (17), where it is assumed that errors in the heavy-atom coordinates for one derivative are uncorrelated with differences in the position of the corresponding atoms in other heavy-atom models.

$$\begin{aligned}\langle \mathbf{F}_i \mathbf{H}_j^* \rangle &= \sum_{k=N_L+1}^{N_A} f_{ik} g_{jk} \langle \exp[2\pi i \mathbf{h} \cdot (\mathbf{x}_{ik} - \mathbf{y}_{jk})] \rangle \\ &= \sum_{k=N_L+1}^{N_A} f_{ik} g_{jk} \langle \exp[2\pi i \mathbf{h} \cdot (\mathbf{x}_{ik} - \mathbf{y}_{ik} + \mathbf{y}_{ik} - \mathbf{y}_{jk})] \rangle \\ &= \sum_{k=N_L+1}^{N_A} f_{ik} g_{jk} \langle \exp[2\pi i \mathbf{h} \cdot (\mathbf{x}_{ik} - \mathbf{y}_{ik})] \rangle \\ &\quad \times \langle \exp[2\pi i \mathbf{h} \cdot (\mathbf{y}_{ik} - \mathbf{y}_{jk})] \rangle \\ &= D_{H_{ij}} \sum_{k=N_L+1}^{N_A} f_{ik} g_{jk} \langle \exp[2\pi i \mathbf{h} \cdot (\mathbf{x}_{ik} - \mathbf{y}_{ik})] \rangle \\ &= D_{H_{ij}} D_{H_i} \Sigma_{H_{ij}}. \end{aligned} \quad (17)$$

As in (14), $D_{H_{ij}}$ is defined as the effective overall value of the phase-shift term arising from differences in atomic coordinates in the models. D_{H_i} is defined similarly as the effective overall value of the phase-shift term arising from heavy-atom model errors, but absorbing any systematic difference in the scale of the f and g atomic scattering factors. In fact, when the heavy-atom occupancies and B factors are refined by maximum likelihood, the resulting g_{jk} values should be approximately equal to $f_{jk}\langle \exp[2\pi i \mathbf{h} \cdot (\mathbf{x}_{ik} - \mathbf{y}_{ik})] \rangle$ (Read, 1990, 1991), in which case the phase-shift component of D_{H_i} will cancel the scale component and D_{H_i} will be equal to one.

Comparing (14) and (17), we see that the terms relating observed and model structure factors differ from the model–model terms only by the factor D_{H_i}. This means that we can define a column vector **D** and a diagonal matrix $\mathbf{D}_H$, then specify $\mathbf{\Sigma}_{12}$ in terms of the earlier matrix $\mathbf{T}_{22}$,

$$\mathbf{D} = \begin{pmatrix} 1 \\ D_1 \\ \vdots \\ D_N \end{pmatrix},$$
$$\mathbf{D}_H = \begin{pmatrix} D_{H_0} & 0 & \cdots & 0 \\ 0 & D_{H_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_{H_N} \end{pmatrix}, \quad (18)$$

$$\boldsymbol{\Sigma}_{12} = (\mathbf{D}\Sigma_{N_L} \quad \mathbf{D}_H \mathrm{T}_{22}). \tag{19}$$

In the conditional distribution, the equations for the updated mean and covariance matrix both include the expression $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$. With the assumptions we have made about independence of errors, this expression assumes the simple form

$$\begin{aligned}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} &= (\mathbf{D}\Sigma_{N_L} \quad \mathbf{D}_H\mathbf{T}_{22})\begin{pmatrix} 1/\Sigma_{N_L} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_{22}^{-1}\end{pmatrix} \\ &= \begin{pmatrix}\mathbf{D} & \mathbf{D}_H\end{pmatrix}. \end{aligned} \tag{20}$$

The mean for the conditional distribution is given by (21), where we see that even though there may be correlations among the heavy-atom models, as long as the errors are independent each heavy-atom model only contributes to the expected value of the corresponding observed structure factor,

$$\begin{aligned}\mu &= \begin{pmatrix}\langle\mathbf{F}_0\rangle \\ \langle\mathbf{F}_1\rangle \\ \vdots \\ \langle\mathbf{F}_N\rangle\end{pmatrix} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\begin{pmatrix}\mathbf{F} \\ \mathbf{H}_0 \\ \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_N\end{pmatrix} \\ &= (\mathbf{D} \quad \mathbf{D}_H)\begin{pmatrix}\mathbf{F} \\ \mathbf{H}_0 \\ \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_N\end{pmatrix} = \begin{pmatrix}\mathbf{F} + D_{H_0}\mathbf{H}_0 \\ D_1\mathbf{F} + D_{H_1}\mathbf{H}_1 \\ \vdots \\ D_N\mathbf{F} + D_{H_N}\mathbf{H}_N\end{pmatrix}. \end{aligned} \tag{21}$$

As noted above, the D_{H_i} values will be equal to one if the heavy-atom occupancies and B factors are allowed to refine to maximize the likelihood target, so the expected values of the structure factors agree with the result in (5).

The matrix $\boldsymbol{\Sigma}_{11}$ contains terms relating the observed structure factors, which may be expressed as in (22). In addition to our earlier assumption that the errors in the heavy-atom model were not correlated with differences in common heavy-atom positions in pairs of models, we now assume that errors in pairs of models are uncorrelated. Again, the factors D_{H_i} and D_{H_j} incorporate the effect of overall differences in scale between the heavy-atom scattering factors in the crystals and the models,

$$\begin{aligned}\langle\mathbf{F}_i\mathbf{F}_j^*\rangle &= \sum_{k=1}^{N_A} f_{ik}f_{jk}\langle\exp[2\pi i\mathbf{h}\cdot(\mathbf{x}_{ik}-\mathbf{x}_{jk})]\rangle \\ &= \sum_{k=1}^{N_L} f_{ik}f_{jk}\langle\exp[2\pi i\mathbf{h}\cdot(\mathbf{x}_{ik}-\mathbf{x}_k+\mathbf{x}_k-\mathbf{x}_{jk}]\rangle \\ &\quad + \sum_{k=N_L+1}^{N_A} f_{ik}f_{jk}\langle\exp[2\pi i\mathbf{h}\cdot(\mathbf{x}_{ik}-\mathbf{y}_{ik}+\mathbf{y}_{ik} \\ &\quad - \mathbf{y}_{jk}+\mathbf{y}_{jk}-\mathbf{x}_{jk})]\rangle \\ &= \sum_{k=1}^{N_L} f_{ik}f_{jk}\langle\exp[2\pi i\mathbf{h}\cdot(\mathbf{x}_{ik}-\mathbf{x}_k)]\rangle\langle\exp[2\pi i\mathbf{h}\cdot(\mathbf{x}_k-\mathbf{x}_{jk})]\rangle \\ &\quad + \sum_{k=N_L+1}^{N_A} f_{ik}f_{jk}\langle\exp[2\pi i\mathbf{h}\cdot(\mathbf{x}_{ik}-\mathbf{y}_{ik})]\rangle \\ &\quad \times \langle\exp[2\pi i\mathbf{h}\cdot(\mathbf{y}_{ik}-\mathbf{y}_{jk})]\rangle\langle\exp[2\pi i\mathbf{h}\cdot(\mathbf{y}_{jk}-\mathbf{x}_{jk})]\rangle \\ &= D_iD_j\Sigma_{N_L} + D_{H_{ij}}\sum_{k=N_L+1}^{N_{\mathrm{atom}}} f_{ik}f_{jk}\langle\exp[2\pi i\mathbf{h}\cdot(\mathbf{x}_{ik}-\mathbf{y}_{ik})]\rangle \\ &\quad \times \langle\exp[2\pi i\mathbf{h}\cdot(\mathbf{y}_{jk}-\mathbf{x}_{jk})]\rangle \\ &= D_iD_j\Sigma_{N_L} + D_{H_{ij}}D_{H_i}D_{H_j}\Sigma_{H_{ij}}. \end{aligned} \tag{22}$$

The last thing we need to compute the conditional covariance matrix is the submatrix $\boldsymbol{\Sigma}_{21}$,

$$\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^T = \begin{pmatrix}\Sigma_{N_L}\mathbf{D}^T \\ \mathbf{T}_{22}\mathbf{D}_H\end{pmatrix}. \tag{23}$$

The expression for the conditional covariance matrix is given in (24), where results from (20), (22) and (23) are incorporated,

$$\begin{aligned}\boldsymbol{\Sigma} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ &= \boldsymbol{\Sigma}_{11} - (\mathbf{D} \quad \mathbf{D}_H)\begin{pmatrix}\Sigma_{N_L}\mathbf{D}^T \\ \mathbf{T}_{22}\mathbf{D}_H\end{pmatrix} \\ &= \boldsymbol{\Sigma}_{11} - \left[\Sigma_{N_L}\begin{pmatrix} D_0^2 & D_0D_1 & \cdots & D_0D_N \\ D_0D_1 & D_1^2 & \cdots & D_1D_N \\ \vdots & \vdots & \ddots & \vdots \\ D_0D_N & D_1D_N & \cdots & D_N^2\end{pmatrix} + \mathbf{D}_H\mathbf{T}_{22}\mathbf{D}_H\right] \\ &= \begin{pmatrix}\sigma_{\Delta 0}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{\Delta 1}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{\Delta N}^2\end{pmatrix}, \\ &\text{where } \sigma_{\Delta i}^2 = \Sigma_i - D_i^2\Sigma_{N_L} - D_{H_i}^2\Sigma_{H_i} + \sigma_i^2 \end{aligned} \tag{24}$$

Because the covariance matrix is diagonal, the joint conditional distribution can be factored into a product of univariate distributions of the form given in (5). Because the distributions are independent, the phases can each be integrated analytically, giving a product of distributions of the form given in (6). When the expected intensity factor ε omitted from (13) to (24) is reintroduced, the variance term is seen to be equivalent to that given in (6). As in (7), this product must be multiplied by the prior distribution of the dummy variable $\mathbf{F}$ before integrating over $\mathbf{F}$ to obtain the likelihood target, which is the conditional distribution of the observed amplitudes.

The integrand in (7) is the joint distribution of the dummy structure factor **F** and the observed amplitudes. The posterior distribution of **F** can be obtained by fixing the observed amplitudes and renormalizing, so we see that the expected value of **F** is the centre of mass of the likelihood target before integration over **F**. Note that if there is a model associated with crystal 0 (the 'native' crystal), the dummy **F** represents what is unexplained by $\mathbf{H}_0$, so the expected value of the native structure factor itself will not be given by the expected value of **F**, but rather by $\langle\mathbf{F}\rangle + D_{H_0}\mathbf{H}_0$.

Although the multivariate derivation has given the same result as the earlier derivations based on independent conditional distributions of single structure factors (Bricogne, 1991; Read, 1991, 1994; de La Fortelle & Bricogne, 1997), this approach forces one to be more explicit about the assumptions that are made. In addition to assuming that the lack of isomorphism for different derivative crystals is independent, we have also assumed that all the errors in positions of atoms in the heavy-atom models are independent. This assumption may be violated by common practice in heavy-atom refinement to constrain the positions of common sites among derivatives (de La Fortelle & Bricogne, 1997). If, in fact, corresponding heavy atoms differ somewhat in position in two derivatives, constrained refinement will place them in an average position. We saw in (22) and (24) that it is necessary to assume that the positional errors $\mathbf{x}_{ik} - \mathbf{y}_{ik}$ and $\mathbf{y}_{jk} - \mathbf{x}_{jk}$ are independent in order for the off-diagonal elements to be eliminated from the conditional covariance matrix. However, these errors will be highly correlated when the positions $\mathbf{y}_{ik}$ and $\mathbf{y}_{jk}$ are constrained to be identical.

Another potential concern is whether refinement will proceed to eliminate correlated errors when it is assumed in the target that there are no correlations. For instance, if common heavy-atom sites are put into refinement with the same starting coordinates, will the assumption that their errors are independent hinder their refinement to slightly different positions?

5. New approaches to experimental phasing

The multivariate statistical analysis of experimental phasing gives us a deeper understanding of the assumptions that are made in current methods about the independence of sources of error. It also points the way to new approaches in which these assumptions can be relaxed.

Firstly, the analysis of multiple isomorphous replacement above shows that it is not necessary to interpret the dummy variable that is introduced in current approaches (Bricogne, 1991; Read, 1991, 1994; de La Fortelle & Bricogne, 1997) as the true value of one of the observed structure factors. It can be given any interpretation that eliminates (or at least minimizes) the off-diagonal covariance elements in the conditional covariance matrix, so that only two-dimensional numerical integration over the dummy variable is required. As shown above, one possible interpretation of the dummy variable is the 'light-atom' substructure of a native crystal that shares atomic sites with at least one derivative. We are exploring the implications of other interpretations of the dummy variable.

Secondly, when it is not possible to eliminate the covariance elements, the multivariate approach shows how to deal with remaining correlations. One source of correlation is in the lack-of-isomorphism errors, which have been assumed to be independent but may be similar among derivatives. This is particularly likely to arise when there are common sites among derivatives, introducing common perturbations of the protein structure. Another source of correlation is anomalous dispersion, in which the effects of errors in the model of anomalous scatterers are necessarily correlated among the Friedel mates for different wavelengths. In principle, the likelihood function could be computed by taking the distribution of all the observed structure factors conditional on models of heavy atoms and anomalous scatterers (but with no dummy variable) and integrating over all the observed phases. Only one such integral can easily be carried out analytically, although Bricogne (2000) has proposed a solution to the multiple integral in terms of 'generalized Bessel functions'.

In at least one special case, the approach of integrating over the observed phases is straightforward. If there are only two observations, one phase integral can be carried out analytically, with the second leaving only a one-dimensional numerical integration. There are a number of cases where we have two observations, such as the joint refinement of native and liganded crystals, single isomorphous replacement and single-wavelength anomalous dispersion (SAD). We have implemented a SAD likelihood target and applied it to a number of test cases (Pannu & Read, 2003). Initial results suggest that this approach will work very well and, given the renaissance of SAD phasing (Dauter *et al.*, 2002), may find significant application. In parallel work, Garib Murshudov (personal communication) has been exploring the use of a similar function to exploit the signal from intrinsic anomalous scatterers in the refinement of protein structures.

This review draws on the work of and discussions with past and present members of my group, particularly Airlie McCoy, Raj Pannu, Laurent Storoni and Hamsapriye. Comments by Gérard Bricogne and questions from Hamsapriye prompted a closer examination of the necessary assumptions of independence. Our research is supported by the Wellcome Trust (UK) and by NIH/NIGMS under grant No. 1P01GM063210.

References

Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
Bricogne, G. (1991). *Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 60–68. Warrington: Daresbury Laboratory.
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. (1993). *Acta Cryst.* D**49**, 37–60.
Bricogne, G. (2000). *Advanced Special Functions and Applications: Proceedings of the Melfi School on Advanced Topics in Mathe-*

matics and Physics, edited by D. Cocolicchio, G. Dattoli & H. M. Srivastava, pp. 315–232. Rome: Aracne Editrice.
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.
Dauter, Z., Dauter, M. & Dodson, E. J. (2002). *Acta Cryst.* D**58**, 494–506.
Einstein, J. R. (1977). *Acta Cryst.* A**33**, 75–85.
Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *Proc. R. Soc. London Ser. A*, **255**, 287–307.
Green, E. A. (1979). *Acta Cryst.* A**35**, 351–359.
Harker, D. (1956). *Acta Cryst.* **9**, 1–9.
Johnson, R. A. & Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis*, 4th ed. New Jersey: Prentice–Hall.
Kleywegt, G. J & Read, R. J. (1997). *Structure*, **5**, 1557–1569.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
Lunin, V. Y. & Urzhumtsev, A. G. (1984). *Acta Cryst.* A**40**, 269–277.
McCoy, A. J. (2002). *Curr. Opin. Struct. Biol.* **12**, 670–673.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Otwinowski, Z. (1991). *Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80–86. Warrington: Daresbury Laboratory.
Pannu, N. S., McCoy, A. J. & Read, R. J. (2003). *Acta Cryst.* D**59**, 1801–1808.
Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.
Pannu, N. S. & Read, R. J. (2003). *Acta Cryst.* D**59**. In the press.
Perutz, M. F. (1956). *Acta Cryst.* **9**, 867–873.
Raiz, V. Sh. & Andreeva, N. S. (1970). *Sov. Phys. Crystallogr.* **15**, 206–210.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.
Read, R. J. (1991). *Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 69–79. Warrington: Daresbury Laboratory.
Read, R. J. (1994). *Lecture Notes from the Workshop on Isomorphous Replacement Methods in Macromolecular Crystallography.* Am. Crystallogr. Assoc. Ann. Meet., Atlanta, GA, USA.
Read, R. J. (1997). *Methods Enzymol.* **277**, 110–128.
Read, R. J. (2001). *Acta Cryst.* D**57**, 1373–1382.
Read, R. J. (2003). *Crystallogr. Rev.* **9**, 33–41.
Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.
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.