International Tables for Crystallography (2006). Vol. F, Section 15.2.3, pp. 325–327.

15.2. Model phases: probabilities, bias and maps

BY R. J. READ

15.2.1. Introduction

The intensities of X-ray diffraction spots measured from a crystal give us only the amplitudes of the diffracted waves. To reconstruct a map of the electron density in the crystal, the unmeasured phase information is also required. In fact, the phases are much more important to the appearance of the map than the measured amplitudes. When phases are supplied by an atomic model, therefore, some degree of model bias is inevitable.

The optimal use of model phase information requires an estimate of its reliability, specifically the probability that various values of the phase angle are true. Such a probability distribution can be derived, starting first with the relationship between the structure factor (amplitude and phase) of the model and that of the true crystal structure. The phase probability distribution can then be obtained from this and used, for instance, to provide a figure-of-merit weighting that minimizes the r.m.s. error from the true electron density.

Even with figure-of-merit weighting, model-phased electron density is biased towards the model. The systematic bias component of model-phased map coefficients can be predicted, allowing the derivation of map coefficients that give electron-density maps with reduced model bias. With the help of a few simple assumptions, a correction for bias can also be made when different sources of phase information are combined.

Finally, the refinement of a model against the observed amplitudes allows a certain amount of overfitting of the data, which leads to an extra 'refinement bias'. Fortunately, the use of appropriate refinement strategies, including maximum-likelihood targets, can reduce the severity of this problem.

15.2.2. Model bias: importance of phase

Dramatic illustrations of the importance of the phase have been published. For instance, Ramachandran & Srinivasan (1961) calculated an electron-density map using phases from one structure and amplitudes from another. In this map there are peaks at the positions of the atoms in the structure that contributed the phase information, but not in the structure that contributed the amplitudes. Similar calculations with two-dimensional Fourier transforms of photographs (Oppenheim & Lim, 1981; Read, 1997) show that the phases of one completely overwhelm the amplitudes of the other.

These examples, though dramatic, are not completely representative of the normal situation, where the structure contributing the phases is partially or even nearly correct. Nonetheless, model phases always contribute bias, so that the resulting map tends to bear too close a resemblance to the model.

15.2.2.1. *Parseval's theorem*

The importance of the phase can be understood most easily in terms of Parseval's theorem, a result that is important to the understanding of many aspects of the Fourier transform and its use in crystallography. Parseval's theorem states that the mean-square value of the variable on one side of a Fourier transform is proportional to the mean-square value of the variable on the other side. Since the Fourier transform is additive, Parseval's theorem also applies to sums or differences.

If ρ_1 and ρ_2 are, for instance, the true electron density and the electron density of the model, respectively, Parseval's theorem tells us that the r.m.s. error in the electron density is proportional to the r.m.s. error in the structure factor. (The structure-factor error is a vector error in the complex plane.)

$$\langle \rho^2 \rangle = (1/V^2) \sum_{\text{all } \mathbf{h}} |\mathbf{F}(\mathbf{h})|^2,$$

$$\left\langle (\rho_1 - \rho_2)^2 \right\rangle = (1/V^2) \sum_{\text{all } \mathbf{h}} |\mathbf{F}_1(\mathbf{h}) - \mathbf{F}_2(\mathbf{h})|^2.$$

This understanding of error in electron-density maps explains why the phase is much more important than the amplitude in determining the appearance of an electron-density map. As illustrated in Fig. 15.2.2.1, a random choice of phase (from a uniform distribution of all possible phases) will generally give a larger error in the complex plane than a random choice of amplitude [from a Wilson (1949) distribution of amplitudes].

15.2.3. Structure-factor probability relationships

To use model phase information optimally, the probability distribution for the true phase (or, equivalently, the distribution of the error in the model phase) needs to be known. Such a distribution can be derived by first working out the probability distribution for the true structure factor (or the distribution of the vector difference between the model and true structure factors). Then the phase probability distribution is obtained by fixing the known value of the structure-factor amplitude and renormalizing.

A number of related structure-factor distributions have been derived, differing in the amount of information available about the structure and in the assumed form of errors in the model. These range from the Wilson distribution, which applies when none of the atomic positions is known, to a distribution that applies when there are a variety of sources of error in an atomic model.

15.2.3.1. *Wilson and Sim structure-factor distributions in P1*

For the Wilson distribution (Wilson, 1949), it is assumed that the atoms in a crystal structure in space group $P1$ are scattered randomly and independently through the unit cell. In fact, it is sufficient to make the much less restrictive assumption that the

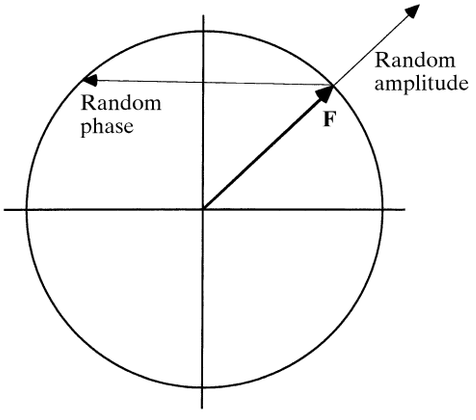

Fig. 15.2.2.1. Schematic illustration of the relative errors introduced by a random choice of phase or a random choice of amplitude. The example has been constructed to represent the r.m.s. errors introduced by randomization (computed by averages over the Wilson distribution). Phase randomization will introduce r.m.s. errors of $(2)^{1/2}$ ($\simeq 1.41$) times the r.m.s. structure-factor amplitude $|\mathbf{F}|$. By comparison, map coefficients weighted by figures of merit of zero would have r.m.s. errors equal to the r.m.s. $|\mathbf{F}|$, so a featureless map would be more accurate than a random-phase map. Amplitude randomization will introduce r.m.s. errors of $[(4-\pi)/2]^{1/2}$ ($\simeq 0.66$) times the r.m.s. $|\mathbf{F}|$, so a map computed with random amplitudes will be closer to the true map than a featureless map.

atoms are placed randomly with respect to the Bragg planes defined by the Miller indices. The assumption of independence is somewhat more problematic, since there are restrictions on the distances between atoms, large volumes of protein crystals are occupied by disordered solvent and many protein crystals display noncrystallographic symmetry; as discussed elsewhere (Vellieux & Read, 1997), the resulting relationships among structure factors are exploited implicitly in averaging and solvent-flattening procedures. The higher-order relationships among structure factors are used explicitly in direct methods for solving small-molecule structures and are being developed for use in protein structures (Bricogne, 1993). For the purposes of simpler relationships between the calculated and true structure factors for a single *hkl*, however, the lack of complete independence does not seem to create serious problems.

When atoms are placed randomly relative to the Bragg planes, the contribution of each atom to the structure factor will have a phase varying randomly from 0 to 2π. The overall structure factor can then be considered to be the result of a random walk in the complex plane, which can be treated as an application of the central limit theorem. The structure factor is the sum of the independent atomic scattering contributions, each of which has a probability distribution defined as a circle in the complex plane centred on the origin, with a radius of f_j. The centroid of this atomic distribution is at the origin, and the variance for each of the real and imaginary parts is $\frac{1}{2}f_j^2$. The probability distribution of the structure factor that is the sum of these contributions is a two-dimensional Gaussian, the product of the one-dimensional Gaussians for the real and imaginary parts. Because the variances are equal in the real and imaginary directions, it can be simplified, as shown below, and expressed in terms of a single distribution parameter, Σ_N.

$$\mathbf{F} = \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{h}\cdot\mathbf{x}_j) = A + iB; \quad \langle A\rangle = \langle B\rangle = 0;$$

$$\sigma^2(A) = \sigma^2(B) = \tfrac{1}{2}\sum_{j=1}^{N} f_j^2 = \tfrac{1}{2}\Sigma_N, \text{so}$$

$$p(A) = [1/(\pi\Sigma_N)^{1/2}]\exp\left(-A^2/\Sigma_N\right),$$

$$p(B) = [1/(\pi\Sigma_N)^{1/2}]\exp\left(-B^2/\Sigma_N\right),$$

$$p(\mathbf{F}) = p(A,B) = (1/\pi\Sigma_N)\exp\left(-|\mathbf{F}|^2/\Sigma_N\right).$$

The Sim distribution (Sim, 1959), which is relevant when the positions of some of the atoms are known, has a very similar basis, except that the structure factor is now considered to arise from a random walk starting from the position of the structure factor corresponding to the known part, $\mathbf{F}_P$. Atoms with known positions do not contribute to the variance, while each of the atoms with unknown positions (the '*Q*' atoms) contributes $\frac{1}{2}f_j^2$ to each of the real and imaginary parts, as in the Wilson distribution. The distribution parameter in this case is referred to as Σ_Q. The Sim distribution is a conditional probability distribution, depending on the value of $\mathbf{F}_P$,

$$p(\mathbf{F};\,\mathbf{F}_P) = (1/\pi\Sigma_Q)\exp\left(-|\mathbf{F}-\mathbf{F}_P|^2/\Sigma_Q\right).$$

The Wilson (1949) and Woolfson (1956) distributions for space group $P\bar{1}$ are obtained similarly, except that the random walks are along a line and the resulting Gaussian distributions are one-dimensional. (The Woolfson distribution is the centric equivalent of the Sim distribution.) For more complicated space groups, it is reasonable to assume that acentric reflections follow the $P1$ distribution and that centric reflections follow the $P\bar{1}$ distribution. However, for any zone of the reciprocal lattice in which symmetry-related atoms are constrained to scatter in phase, the variances must be multiplied by the expected intensity factor, ε, for the zone, because the symmetry-related contributions are no longer independent.

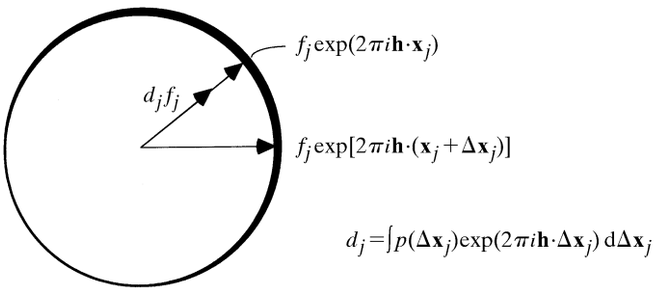

Fig. 15.2.3.1. Centroid of the structure-factor contribution from a single atom. The probability of a phase for the contribution is indicated by the thickness of the line.

15.2.3.2. *Probability distributions for variable coordinate errors*

In the Sim distribution, an atom is considered to be either exactly known or completely unknown in its position. These are extreme cases, since there will normally be varying degrees of uncertainty in the positions of various atoms in a model. The treatment can be generalized by allowing a probability distribution of coordinate errors for each atom. In this case, the centroid for the individual atomic contribution to the structure factor will no longer be obtained by multiplying by either zero or one. Averaged over the circle corresponding to possible phase errors, the centroid will generally be reduced in magnitude, as illustrated in Fig. 15.2.3.1. In fact, averaging to obtain the centroid is equivalent to weighting the atomic scattering contribution by the Fourier transform of the coordinate-error probability distribution, d_j. By the convolution theorem, this in turn is equivalent to convoluting the atomic density with the coordinate-error distribution. Intuitively, the atom is smeared over all of its possible positions. The weighting factor, d_j, is thus analogous to the thermal-motion term in the structure-factor expression.

The variances for the individual atomic contributions will differ in magnitude, but if there are a sufficient number of independent sources of error, we can invoke the central limit theorem again and assume that the probability distribution for the structure factor will be a Gaussian centred on $\sum d_j f_j \exp\left(2\pi i\mathbf{h}\cdot\mathbf{x}_j\right)$. If the coordinate-error distribution is Gaussian, and if each atom in the model is subject to the same errors, the resulting structure-factor probability distribution is the Luzzati (1952) distribution. In this special case, $d_j = D$ for all atoms, where D is the Fourier transform of a Gaussian and behaves like the application of an overall B factor.

15.2.3.3. *General treatment of the structure-factor distribution*

The Wilson, Sim, Luzzati and variable-error distributions have very similar forms, because they are all Gaussians arising from the application of the central limit theorem. The central limit theorem is valid under many circumstances; even when there are errors in position, scattering factor and B factor, as well as missing atoms, a similar distribution still applies. As long as these sources of error are independent, the true structure factor will have a Gaussian distribution centred on $D\mathbf{F}_C$ (Fig. 15.2.3.2), where D now includes effects of all sources of error, as well as compensating for errors in the overall scale and B factor (Read, 1990).

$$p(\mathbf{F};\,\mathbf{F}_C) = (1/\pi\varepsilon\sigma_\Delta^2)\exp\left(-|\mathbf{F}-D\mathbf{F}_C|^2/\varepsilon\sigma_\Delta^2\right)$$

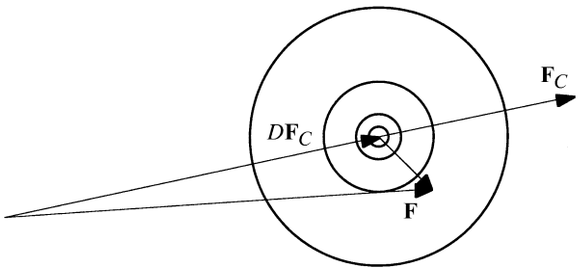

Fig. 15.2.3.2. Schematic illustration of the general structure-factor distribution, relevant in the case of any set of independent random errors in the atomic model.

in the acentric case, where $\sigma_\Delta^2 = \Sigma_N - D^2\Sigma_P$, ε is the expected intensity factor and Σ_P is the Wilson distribution parameter for the model.

For centric reflections, the scattering differences are distributed along a line, so the probability distribution is a one-dimensional Gaussian.

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

15.2.3.4. *Estimating σ_A*

Srinivasan (1966) showed that the Sim and Luzzati distributions could be combined into a single distribution that had a particularly elegant form when expressed in terms of normalized structure factors, or E values. This functional form still applies to the general distribution that reflects a variety of sources of error; the only difference is the interpretation placed on the parameters (Read, 1990). If $\mathbf{F}$ and $\mathbf{F}_C$ are replaced by the corresponding E values, a parameter σ_A plays the role of D, and σ_Δ^2 reduces to $(1 - \sigma_A^2)$. [The parameter σ_A is equivalent to D after correction for model completeness; $\sigma_A = D(\Sigma_P/\Sigma_N)^{1/2}$.] When the structure factors are normalized, overall scale and B-factor effects are also eliminated. The parameter σ_A that characterizes this probability distribution varies as a function of resolution. It must be deduced from the amplitudes $|\mathbf{F}_O|$ and $|\mathbf{F}_C|$, since the phase (thus the phase difference) is unknown.

A general approach to estimating parameters for probability distributions is to maximize a likelihood function. The likelihood function is the overall joint probability of making the entire set of observations, which is a function of the desired parameters. The parameters that maximize the probability of making the set of observations are the most consistent with the data. The idea of using maximum likelihood to estimate model phase errors was introduced by Lunin & Urzhumtsev (1984), who gave a treatment that was valid for space group $P1$. In a more general treatment that applies to higher-symmetry space groups, allowance is made for the statistical effects of crystal symmetry (centric zones and differing expected intensity factors) (Read, 1986).

The σ_A values are estimated by maximizing the joint probability of making the set of observations of $|\mathbf{F}_O|$. If the structure factors are all assumed to be independent, the joint probability distribution is the product of all the individual distributions. The assumption of independence is not completely justified in theory, but the results are fairly accurate in practice.

$$L = \prod_{\mathbf{h}} p(|\mathbf{F}_O|;\, |\mathbf{F}_C|).$$

The required probability distribution, $p(|\mathbf{F}_O|;\, |\mathbf{F}_C|)$, is derived from $p(\mathbf{F};\, \mathbf{F}_C)$ by integrating over all possible phase differences and neglecting the errors in $|\mathbf{F}_O|$ as a measure of $|\mathbf{F}|$. The form of this distribution, which is given in other publications (Read, 1986, 1990), differs for centric and acentric reflections. (It is important to note that although the distributions for structure factors are Gaussian, the distributions for amplitudes obtained by integrating out the phase are not.) It is more convenient to deal with a sum than a product, so the log likelihood function is maximized instead. In the program *SIGMAA*, reciprocal space is divided into spherical shells, and a value of the parameter σ_A is refined for each resolution shell. Details of the algorithm are given elsewhere (Read, 1986).

The resolution shells must be thick enough to contain several hundred to a thousand reflections each, in order to provide σ_A estimates with a sufficiently small statistical error. A larger number of shells (fewer reflections per shell) can be used for refined structures, since estimates of σ_A become more precise as the true value approaches 1. If there are sufficient reflections per shell, the estimates will vary smoothly with resolution. As discussed below, the smooth variation with resolution can also be exploited through a restraint that allows σ_A values to be estimated from fewer reflections.

15.2.4. Figure-of-merit weighting for model phases

Blow & Crick (1959) and Sim (1959) showed that the electron-density map with the least r.m.s. error is calculated from centroid structure factors. This conclusion follows from Parseval's theorem, because the centroid structure factor (its probability-weighted average value or expected value) minimizes the r.m.s. error of the structure factor. Since the structure-factor distribution $p(\mathbf{F};\, \mathbf{F}_C)$ is symmetrical about $\mathbf{F}_C$, the expected value of $\mathbf{F}$ will have the same phase as $\mathbf{F}_C$, but the averaging around the phase circle will reduce its magnitude if there is any uncertainty in the phase value (Fig. 15.2.4.1). We treat the reduction in magnitude by applying a weighting factor called the figure of merit, m, which is equivalent to the expected value of the cosine of the phase error.

15.2.5. Map coefficients to reduce model bias

15.2.5.1. *Model bias in figure-of-merit weighted maps*

A figure-of-merit weighted map, calculated with coefficients $m|\mathbf{F}_O| \exp(i\alpha_C)$, has the least r.m.s. error from the true map. According to the normal statistical (minimum variance) criteria, then, it is the best map. However, such a map will suffer from model bias; if its purpose is to allow the detection and repair of errors in the model, this is a serious qualitative defect. Fortunately, it is possible to predict the systematic errors leading to model bias and to make some correction for them.

Main (1979) dealt with this problem in the case of a perfect partial structure. Since the relationships among structure factors are the same in the general case of a partial structure with various errors, once $D\mathbf{F}_C$ is substituted for $\mathbf{F}_C$, all that is required to apply Main's results more generally is a change of variables (Read, 1986, 1990).

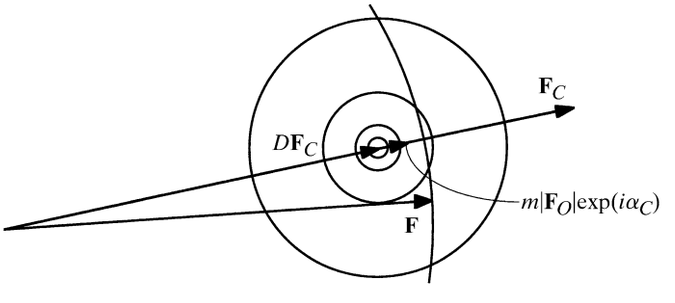

Fig. 15.2.4.1. Figure-of-merit weighted model-phased structure factor, obtained as the probability-weighted average over all possible phases.