International Tables for Crystallography (2006). Vol. F, Chapter 15.2, pp. 325–331.

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

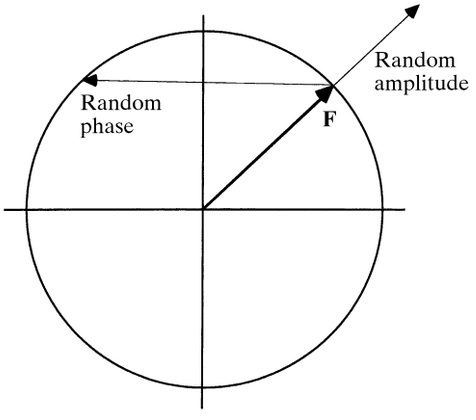

Fig. 15.2.2.1. Schematic illustration of the relative errors introduced by a random choice of phase or a random choice of amplitude. The example has been constructed to represent the r.m.s. errors introduced by randomization (computed by averages over the Wilson distribution). Phase randomization will 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.

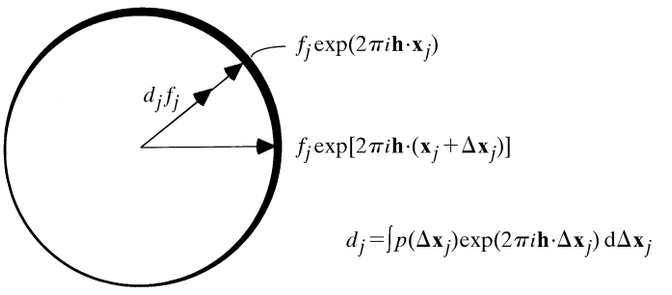

Fig. 15.2.3.1. Centroid of the structure-factor contribution from a single atom. The probability of a phase for the contribution is indicated by the thickness of the line.

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).

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

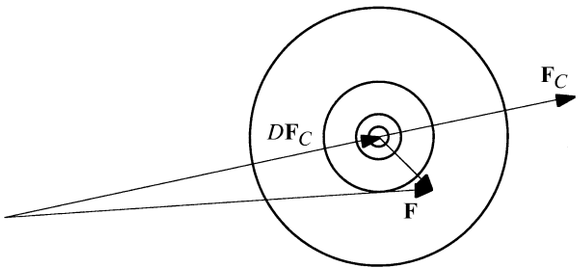

Fig. 15.2.3.2. Schematic illustration of the general structure-factor distribution, relevant in the case of any set of independent random errors in the atomic model.

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).

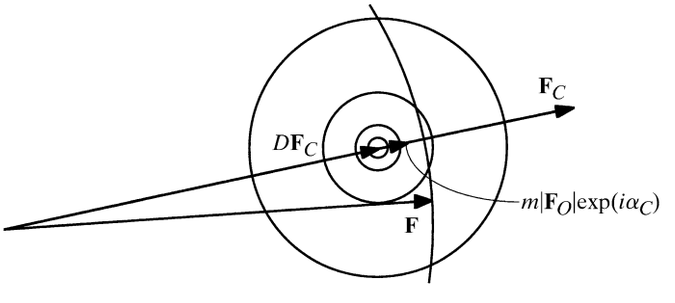

Fig. 15.2.4.1. Figure-of-merit weighted model-phased structure factor, obtained as the probability-weighted average over all possible phases.

In Main's approach, the cosine law is used to introduce the cosine of the phase error, which is converted into a figure of merit by taking expected values. Some manipulations allow us to solve for the figure-of-merit weighted map coefficient, which is approximated as a linear combination of the true structure factor and the model structure factor (Main, 1979; Read, 1986). Finally, we can solve for an approximation to the true structure factor, giving map coefficients from which the systematic model bias component has been removed.

$$m|\mathbf{F}_O|\exp(i\alpha_C) = F/2 + D\mathbf{F}_C/2 + \text{ noise terms},$$
$$F \simeq (2m|\mathbf{F}_O| - D|\mathbf{F}_C|)\exp(i\alpha_C).$$

A similar analysis for centric structure factors shows that there is no systematic model bias in figure-of-merit weighted map coefficients, so no bias correction is needed in the centric case.

15.2.5.2. *Model bias in combined phase maps*

When model phase information is combined with, for instance, multiple isomorphous replacement (MIR) phase information, there will still be model bias in the acentric map coefficients, to the extent that the model influences the final phases. However, it is inappropriate to continue using the same map coefficients to reduce model bias, because some phases could be determined almost completely by the MIR phase information. It makes much more sense to have map coefficients that reduce to the coefficients appropriate for either model or MIR phases, in extreme cases where there is only one source of phase information, and that vary smoothly between those extremes.

Map coefficients that satisfy these criteria (even if they are not rigorously derived) are implemented in the program *SIGMAA*. The resulting maps are reasonably successful in reducing model bias. Two assumptions are made: (1) the model bias component in the figure-of-merit weighted map coefficient, $m_{\rm com}|\mathbf{F}_O|\exp(i\alpha_{\rm com})$, is proportional to the influence that the model phase has had on the combined phase; and (2) the relative influence of a source of phase information can be measured by the information content, H (Guiasu, 1977), of the phase probability distribution. The first assumption corresponds to the idea that the figure-of-merit weighted map coefficient is a linear combination of the MIR and model phase cases.

MIR:	$m_{\rm MIR}\lvert\mathbf{F}_O\rvert\exp(i\alpha_{\rm MIR})$	$\simeq \mathbf{F}$
Model:	$m_C\lvert\mathbf{F}_O\rvert\exp(i\alpha_C)$	$\simeq \mathbf{F}/2 + D\mathbf{F}_C/2$
Combined:	$m_{\rm com}\lvert\mathbf{F}_O\rvert\exp(i\alpha_{\rm com})$	$\simeq [1-(w/2)]\mathbf{F} + (w/2)D\mathbf{F}_C,$

where

$$w = H_C/(H_C + H_{\rm MIR})$$

and

$$H = \int_0^{2\pi} p(\alpha)\ln\frac{p(\alpha)}{p_0(\alpha)}\,{\rm d}\alpha; \quad p_0(\alpha) = \frac{1}{2\pi}.$$

Solving for an approximation to the true $\mathbf{F}$ gives the following expression, which can be seen to reduce appropriately when w is 0 (no model influence) or 1 (no MIR influence):

$$\mathbf{F} \simeq \frac{2m|\mathbf{F}_O|\exp(i\alpha_{\rm com}) - wD\mathbf{F}_C}{2-w}.$$

15.2.6. Estimation of overall coordinate error

In principle, since the distribution of observed and calculated amplitudes is determined largely by the coordinate errors of the model, one can determine whether a particular coordinate-error distribution is consistent with the amplitudes. Unfortunately, it turns out that the coordinate errors cannot be deduced unambiguously, because many distributions of coordinate errors are consistent with a particular distribution of amplitudes (Read, 1990).

If the simplifying assumption is made that all the atoms are subject to a single error distribution, then the parameter D (and thus the related parameter σ_A) varies with resolution as the Fourier transform of the error distribution, as discussed above. Two related methods to estimate overall coordinate error are based on the even more specific assumption that the coordinate-error distribution is Gaussian: the Luzzati plot (Luzzati, 1952) and the σ_A plot (Read, 1986). Unfortunately, the central assumption is not justified; atoms that scatter more strongly (heavier atoms or atoms with lower B factors) tend to have smaller coordinate errors than weakly scattering atoms. The proportion of the structure factor contributed by well ordered atoms increases at high resolution, so that the structure factors agree better at high resolution than if there were a single error distribution.

It is often stated, optimistically, that the Luzzati plot provides an upper bound to the coordinate error, because the observation errors in $|\mathbf{F}_O|$ have been ignored. This is misleading, because there are other effects that cause the Luzzati and σ_A plots to give underestimates (Read, 1990). Chief among these are the correlation of errors and scattering power and the overfitting of the amplitudes in structure refinement (discussed below). These estimates of overall coordinate error should not be interpreted too literally; at best, they provide a comparative measure.

15.2.7. Difference-map coefficients

The computer program *SIGMAA* (Read, 1986) has been developed to implement the results described here. Apart from the two types of map coefficient discussed above, two types of difference-map coefficient can also be produced:

(1) Model-phased difference map: $(m|\mathbf{F}_O| - D|\mathbf{F}_C|)\exp(i\alpha_C)$;

(2) General difference map: $m_{\rm com}|\mathbf{F}_O|\exp(i\alpha_{\rm com}) - D\mathbf{F}_C$.

The general difference map, it should be noted, uses a vector difference between the figure-of-merit weighted combined phase coefficient (the 'best' estimate of the true structure factor) and the calculated structure factor. When additional phase information is available, it should provide a clearer picture of the errors in the model.

15.2.8. Refinement bias

The structure-factor probabilities discussed above depend on the atoms having independent errors (or at least a sufficient number of groups of atoms having independent errors). Unfortunately, this assumption breaks down when a structure is refined against the observed diffraction data. Few protein crystals diffract to sufficiently high resolution to provide a large number of observations for every refinable parameter. The refinement problem is, therefore, not sufficiently overdetermined, so it is possible to overfit the data. If there is an error in the model that is outside the range of convergence of the refinement method, it is possible to introduce compensating errors in the rest of the structure to give a better, and misleading, agreement in the amplitudes. As a result, the phase accuracy (hence the weighting factors m and D) is overestimated, and model bias is poorly removed. Because simulated annealing is a more effective minimizer than gradient methods (Brünger *et al.*, 1987), it is also more effective at locating local minima, so structures refined by simulated annealing probably tend to suffer more severely from refinement bias.

There is another interpretation to the problem of refinement bias. As Silva & Rossmann (1985) point out, minimizing the r.m.s. difference between the amplitudes $|\mathbf{F}_O|$ and $|\mathbf{F}_C|$ is equivalent (by

Parseval's theorem) to minimizing the difference between the model electron density and the density corresponding to the map coefficients $|\mathbf{F}_O| \exp(i\alpha_C)$; a lower residual is obtained either by making the model look more like the true structure, or by making the model-phased map look more like the model through the introduction of systematic phase errors.

A number of strategies are available to reduce the degree or impact of refinement bias. The overestimation of phase accuracy has been overcome in a new version of *SIGMAA* that is under development (Read, unpublished). Cross-validation data, which are normally used to compute $R_{\rm free}$ as an unbiased indicator of refinement progress (Brünger, 1992), are used to obtain unbiased σ_A estimates. Because of the high statistical error of σ_A estimates computed from small numbers of reflections, reliable values can only be obtained by exploiting the smoothness of the σ_A curve as a function of resolution. This can be achieved either by fitting a functional form or by adding a penalty to points that deviate from the line connecting their neighbours. Lunin & Skovoroda (1995) have independently proposed the use of cross-validation data for this purpose, but as their algorithm is equivalent to the conventional *SIGMAA* algorithm, it will suffer severely from statistical error.

The degree of refinement bias can be reduced by placing less weight on the agreement of structure-factor amplitudes. Anecdotal evidence suggests that the problem is less serious, in structures refined using *X-PLOR* (Brünger *et al.*, 1987), when the Engh & Huber (1991) parameter set is used for the energy terms. In this new parameter set, the deviations from standard geometry are much more strictly restrained, so in effect the pressure on the agreement of structure-factor amplitudes is reduced. The use of maximum-likelihood targets for refinement (discussed below) also helps to reduce overfitting.

If errors are suspected in certain parts of the structure, 'omit refinement' (in which the questionable parts are omitted from the model) can be a very effective way to eliminate refinement bias in those regions (James *et al.*, 1980; Hodel *et al.*, 1992).

If MIR or MAD (multiwavelength anomalous dispersion) phases are available, combined phase maps tend to suffer less from refinement bias, depending on the extent to which the experimental phases influence the combined phases. Finally, it is always a good idea to refer occasionally to the original MIR or MAD map, which cannot suffer at all from model bias or refinement bias.

15.2.9. Maximum-likelihood structure refinement

In the past, conventional structure refinement was based on a least-squares target, which would be justified if the observed and calculated structure-factor amplitudes were related by a Gaussian probability distribution. Unfortunately, the relationship between $|\mathbf{F}_O|$ and $|\mathbf{F}_C|$ is not Gaussian, and the distribution for $|\mathbf{F}_O|$ is not even centred on $|\mathbf{F}_C|$. Because of this, it was suggested (Read, 1990; Bricogne, 1991) that a maximum-likelihood target should be used instead, and that it should be based on probability distributions such as those described above.

Three implementations of maximum-likelihood structure refinement have now been reported (Pannu & Read, 1996; Murshudov *et al.*, 1997; Bricogne & Irwin, 1996). As expected, there is a decrease in refinement bias, as the calculated structure-factor amplitudes will not be forced to be equal to the observed amplitudes. Maximum-likelihood targets have been shown to work much better than least-squares targets, particularly when the starting models are poor.

Prior phase information can also be incorporated into a maximum-likelihood target (Pannu *et al.*, 1998). Tests show that even weak phase information can have a dramatic effect on the success of refinement, and that the amount of overfitting is even further reduced (Pannu *et al.*, 1998).

Acknowledgements

This chapter is a revised version of a contribution to *Methods in Enzymology* (Read, 1997).

References

15.1

Abrahams, J. P. (1997). *Bias reduction in phase refinement by modified interference functions: introducing the γ correction*. *Acta Cryst.* D**53**, 371–376.

Abrahams, J. P. & Leslie, A. G. W. (1996). *Methods used in the structure determination of bovine mitochondrial F_1 ATPase*. *Acta Cryst.* D**52**, 30–42.

Agarwal, R. C. & Isaacs, N. W. (1977). *Method for obtaining a high resolution protein map starting from a low resolution map*. *Proc. Natl Acad. Sci. USA*, **74**(7), 2835–2839.

Baker, D., Bystroff, C., Fletterick, R. J. & Agard, D. A. (1993). *PRISM: topologically constrained phase refinement for macromolecular crystallography*. *Acta Cryst.* D**49**, 429–439.

Baker, D., Krukowski, A. E. & Agard, D. A. (1993). *Uniqueness and the ab initio phase problem in macromolecular crystallography*. *Acta Cryst.* D**49**, 186–192.

Bhat, T. N. & Blow, D. M. (1982). *A density-modification method for the improvement of poorly resolved protein electron-density maps*. *Acta Cryst.* A**38**, 21–29.

Blow, D. M. & Rossmann, M. G. (1961). *The single isomorphous replacement method*. *Acta Cryst.* **14**, 1195–1202.

Bricogne, G. (1974). *Geometric sources of redundancy in intensity data and their use for phase determination*. *Acta Cryst.* A**30**, 395–405.

Bricogne, G. (1976). *Methods and programs for direct-space exploitation of geometric redundancies*. *Acta Cryst.* A**32**, 832–847.

Brünger, A. T. (1992). *Free R value: a novel statistical quantity for assessing the accuracy of crystal structures*. *Nature (London)*, **355**, 472–475.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Crystallographic R factor refinement by molecular dynamics*. *Science*, **235**, 458–460.

Bystroff, C., Baker, D., Fletterick, R. J. & Agard, D. A. (1993). *PRISM: application to the solution of two protein structures*. *Acta Cryst.* D**49**, 440–448.

Chapman, M. S., Tsao, J. & Rossmann, M. G. (1992). *Ab initio phase determination for spherical viruses: parameter determination for spherical-shell models*. *Acta Cryst.* A**48**, 301–312.

Cowtan, K. D. (1999). *Error estimation and bias correction in phase-improvement calculations*. *Acta Cryst.* D**55**, 1555–1567.

Cowtan, K. D. & Main, P. (1993). *Improvement of macromolecular electron-density maps by the simultaneous application of real and reciprocal space constraints*. *Acta Cryst.* D**49**, 148–157.

Cowtan, K. D. & Main, P. (1996). *Phase combination and cross validation in iterated density-modification calculations*. *Acta Cryst.* D**52**, 43–48.

Cowtan, K. D. & Main, P. (1998). *Miscellaneous algorithms for density modification*. *Acta Cryst.* D**54**, 487–493.

Cowtan, K. D. & Zhang, K. Y. J. (1999). *Density modification for macromolecular phase improvement*. *Prog. Biophys. Mol. Biol.* **72**, 245–270.

Crowther, R. A. & Blow, D. M. (1967). *A method of positioning a known molecule in an unknown crystal structure*. *Acta Cryst.* **23**, 544–548.

15.1 (*cont.*)

Greer, J. (1985). *Computer skeletonization and automatic electron-density map analysis*. In *Diffraction methods for biological macromolecules*, edited by H. W. Wyckoff, C. H. W. Hirs & S. N. Timasheff, Vol. 115, pp. 206–224. Orlando: Academic Press.

Harrison, R. W. (1988). *Histogram specification as a method of density modification*. *J. Appl. Cryst.* **21**, 949–952.

Hauptman, H. (1986). *The direct methods of X-ray crystallography*. *Science*, **233**, 178–183.

Hendrickson, W. A., Klippenstein, G. L. & Ward, K. B. (1975). *Tertiary structure of myohemerythrin at low resolution*. *Proc. Natl Acad. Sci. USA*, **72**(6), 2160–2164.

Hendrickson, W. A. & Lattman, E. E. (1970). *Representation of phase probability distributions for simplified combination of independent phase information*. *Acta Cryst.* B**26**, 136–143.

Hoppe, W. & Gassmann, J. (1968). *Phase correction, a new method to solve partially known structures*. *Acta Cryst.* B**24**, 97–107.

Karle, J. (1986). *Recovering phase information from intensity data*. *Science*, **232**, 837–843.

Lamzin, V. S. & Wilson, K. S. (1997). *Automated refinement for protein crystallography*. *Methods Enzymol.* **277**, 269–305.

Leslie, A. G. W. (1987). *A reciprocal-space method for calculating a molecular envelope using the algorithm of B. C. Wang*. *Acta Cryst.* A**43**, 134–136.

Lunin, V. Yu. (1988). *Use of the information on electron density distribution in macromolecules*. *Acta Cryst.* A**44**, 144–150.

Lunin, V. Yu. & Skovoroda, T. P. (1991). *Frequency-restrained structure-factor refinement. I. Histogram simulation*. *Acta Cryst.* A**47**, 45–52.

Main, P. (1979). *A theoretical comparison of the* β, γ' *and* $2F_o - F_c$ *syntheses*. *Acta Cryst.* A**35**, 779–785.

Main, P. (1990*a*). *A formula for electron density histograms for equal-atom structures*. *Acta Cryst.* A**46**, 507–509.

Main, P. (1990*b*). *The use of Sayre's equation with constraints for the direct determination of phases*. *Acta Cryst.* A**46**, 372–377.

Main, P. & Rossmann, M. G. (1966). *Relationships among structure factors due to identical molecules in different crystallographic environments*. *Acta Cryst.* **21**, 67–72.

Matthews, B. W. (1968). *Solvent content of protein crystals*. *J. Mol. Biol.* **33**, 491–497.

Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *On the application of the minimal principle to solve unknown structures*. *Science*, **259**, 1430–1433.

Navaza, J. (1994). *AMoRe: an automated package for molecular replacement*. *Acta Cryst.* A**50**, 157–163.

Perrakis, A., Sixma, T. K., Wilson, K. S. & Lamzin, V. S. (1997). *wARP: improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models*. *Acta Cryst.* D**53**, 448–455.

Podjarny, A. D. & Yonath, A. (1977). *Use of matrix direct methods for low-resolution phase extension for tRNA*. *Acta Cryst.* A**33**, 655–661.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors*. *Acta Cryst.* A**42**, 140–149.

Read, R. J. & Schierbeek, A. J. (1988). *A phased translation function*. *J. Appl. Cryst.* **21**, 490–495.

Reynolds, R. A., Remington, S. J., Weaver, L. H., Fisher, R. G., Anderson, W. F., Ammon, H. L. & Matthews, B. W. (1985). *Structure of a serine protease from rat mast cells determined from twinned crystals by isomorphous and molecular replacement*. *Acta Cryst.* B**41**, 139–147.

Rossmann, M. G. & Blow, D. M. (1962). *The detection of sub-units within the crystallographic asymmetric unit*. *Acta Cryst.* **15**, 24–31.

Rossmann, M. G., McKenna, R., Tong, L., Xia, D., Dai, J.-B., Wu, H., Choi, H.-K. & Lynch, R. E. (1992). *Molecular replacement real-space averaging*. *J. Appl. Cryst.* **25**, 166–180.

Sayre, D. (1952). *The squaring method: a new method for phase determination*. *Acta Cryst.* **5**, 60–65.

Sayre, D. (1972). *On least-squares refinement of the phases of crystallographic structure factors*. *Acta Cryst.* A**28**, 210–212.

Sayre, D. (1974). *Least-squares phase refinement. II. High-resolution phasing of a small protein*. *Acta Cryst.* A**30**, 180–184.

Schevitz, R. W., Podjarny, A. D., Zwick, M., Hughes, J. J. & Sigler, P. B. (1981). *Improving and extending the phases of medium- and low-resolution macromolecular structure factors by density modification*. *Acta Cryst.* A**37**, 669–677.

Schuller, D. J. (1996). *MAGICSQUASH: more versatile non-crystallographic averaging with multiple constraints*. *Acta Cryst.* D**52**, 425–434.

Sim, G. A. (1959). *The distribution of phase angles for structures containing heavy atoms. II. A modification of the normal heavy-atom method for non-centrosymmetrical structures*. *Acta Cryst.* **12**, 813–815.

Swanson, S. M. (1994). *Core tracing: depicting connections between features in electron density*. *Acta Cryst.* D**50**, 695–708.

Tsao, J., Chapman, M. S. & Rossmann, M. G. (1992). *Ab initio phase determination for viruses with high symmetry: a feasibility study*. *Acta Cryst.* A**48**, 293–301.

Vellieux, F. M. D. (1998). *A comparison of two algorithms for electron-density map improvement by introduction of atomicity: skeletonization, and map sorting followed by refinement*. *Acta Cryst.* D**54**, 81–85.

Vellieux, F. M. D. A. P., Hunt, J. F., Roy, S. & Read, R. J. (1995). *DEMON/ANGEL: a suite of programs to carry out density modification*. *J. Appl. Cryst.* **28**, 347–351.

Wang, B. C. (1985). *Resolution of phase ambiguity in macromolecular crystallography*. In *Diffraction methods for biological macromolecules*, edited by H. W. Wyckoff, C. H. W. Hirs & S. N. Timasheff, Vol. 115, pp. 90–113. Orlando: Academic Press.

Weeks, C. M., DeTitta, G. T., Miller, R. & Hauptman, H. A. (1993). *Applications of the minimal principle to peptide structures*. *Acta Cryst.* D**49**, 179–181.

Wigley, D. B., Roper, D. I. & Cooper, R. A. (1989). *Preliminary crystallographic analysis of 5-carboxymethyl-2-hydroxymuconate isomerase from Escherichia coli*. *J. Mol. Biol.* **210**, 881–882.

Wilson, A. J. C. (1949). *The probability distribution of X-ray intensities*. *Acta Cryst.* **2**, 318–321.

Wilson, C. & Agard, D. A. (1993). *PRISM: automated crystallographic phase refinement by iterative skeletonization*. *Acta Cryst.* A**49**, 97–104.

Woolfson, M. M. (1987). *Direct methods – from birth to maturity*. *Acta Cryst.* A**43**, 593–612.

Zhang, K. Y. J. (1993). *SQUASH – combining constraints for macromolecular phase refinement and extension*. *Acta Cryst.* D**49**, 213–222.

Zhang, K. Y. J., Cowtan, K. D. & Main, P. (1997). *Combining constraints for electron density modification*. In *Macromolecular crystallography*, edited by C. W. Carter & R. M. Sweet, Vol. 277, pp. 53–64. New York: Academic Press.

Zhang, K. Y. J. & Main, P. (1988). *Histogram matching as a density modification technique for phase refinement and extension of protein molecules*. In *Improving protein phases*, edited by S. Bailey, E. Dodson & S. Phillips. Report DL/SCI/R26, pp. 57–64. Warrington: Daresbury Laboratory.

Zhang, K. Y. J. & Main, P. (1990*a*). *Histogram matching as a new density modification technique for phase refinement and extension of protein molecules*. *Acta Cryst.* A**46**, 41–46.

Zhang, K. Y. J. & Main, P. (1990*b*). *The use of Sayre's equation with solvent flattening and histogram matching for phase extension and refinement of protein structures*. *Acta Cryst.* A**46**, 377–381.

15.2

Blow, D. M. & Crick, F. H. C. (1959). *The treatment of errors in the isomorphous replacement method*. *Acta Cryst.* **12**, 794–802.

Bricogne, G. (1991). *A multisolution method of phase determination by combined maximization of entropy and likelihood. III. Extension to powder diffraction data*. *Acta Cryst.* A**47**, 803–829.

15.2 (*cont.*)

Bricogne, G. (1993). *Direct phase determination by entropy maximization and likelihood ranking: status report and perspectives. Acta Cryst.* D**49**, 37–60.

Bricogne, G. & Irwin, J. (1996). In *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). *Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature (London)*, **355**, 472–474.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Crystallographic R factor refinement by molecular dynamics. Science*, **235**, 458–460.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst.* A**47**, 392–400.

Guiasu, S. (1977). *Information theory with applications.* London: McGraw-Hill.

Hodel, A., Kim, S.-H. & Brünger, A. T. (1992). *Model bias in macromolecular crystal structures. Acta Cryst.* A**48**, 851–858.

James, M. N. G., Sielecki, A. R., Brayer, G. D., Delbaere, L. T. J. & Bauer, C.-A. (1980). *Structures of product and inhibitor complexes of Streptomyces griseus protease A at 1.8 Å resolution – a model for serine protease catalysis. J. Mol. Biol.* **144**, 43–88.

Lunin, V. Yu. & Skovoroda, T. P. (1995). *R-free likelihood-based estimates of errors for phases calculated from atomic models. Acta Cryst.* A**51**, 880–887.

Lunin, V. Yu. & Urzhumtsev, A. G. (1984). *Improvement of protein phases by coarse model modification. Acta Cryst.* A**40**, 269–277.

Luzzati, V. (1952). *Traitement statistique des erreurs dans la determination des structures cristallines. Acta Cryst.* **5**, 802–810.

Main, P. (1979). *A theoretical comparison of the β, γ' and $2F_o - F_c$ syntheses. Acta Cryst.* A**35**, 779–785.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method. Acta Cryst.* D**53**, 240–255.

Oppenheim, A. V. & Lim, J. S. (1981). *The importance of phase in signals. Proc. IEEE*, **69**, 529–541.

Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Incorporation of prior phase information strengthens maximum-likelihood structure refinement. Acta Cryst.* D**54**, 1285–1294.

Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood. Acta Cryst.* A**52**, 659–668.

Ramachandran, G. N. & Srinivasan, R. (1961). *An apparent paradox in crystal structure analysis. Nature (London)*, **190**, 159–161.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst.* A**42**, 140–149.

Read, R. J. (1990). *Structure-factor probabilities for related structures. Acta Cryst.* A**46**, 900–912.

Read, R. J. (1997). *Model phases: probabilities and bias. Methods Enzymol.* **277**, 110–128.

Silva, A. M. & Rossmann, M. G. (1985). *The refinement of southern bean mosaic virus in reciprocal space. Acta Cryst.* B**41**, 147–157.

Sim, G. A. (1959). *The distribution of phase angles for structures containing heavy atoms. II. A modification of the normal heavy-atom method for non-centrosymmetrical structures. Acta Cryst.* **12**, 813–815.

Srinivasan, R. (1966). *Weighting functions for use in the early stages of structure analysis when a part of the structure is known. Acta Cryst.* **20**, 143–144.

Vellieux, F. M. D. & Read, R. J. (1997). *Non-crystallographic symmetry averaging in phase refinement and extension. Methods Enzymol.* **277**, 18–53.

Wilson, A. J. C. (1949). *The probability distribution of X-ray intensities. Acta Cryst.* **2**, 318–321.

Woolfson, M. M. (1956). *An improvement of the 'heavy-atom' method of solving crystal structures. Acta Cryst.* **9**, 804–810.