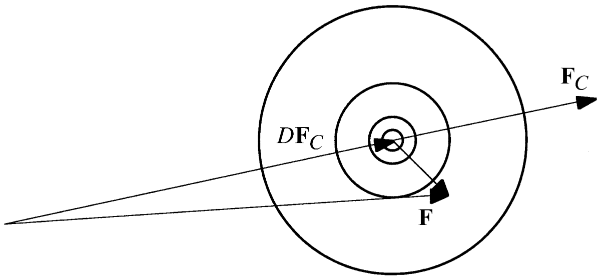

Fig. 15.2.3.2. Schematic illustration of the general structure-factor distribution, relevant in the case of any set of independent random errors in the atomic model.

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).

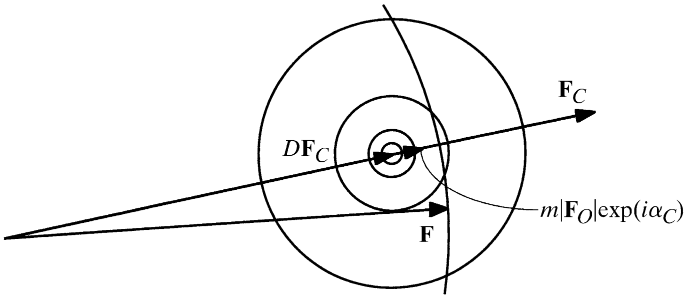

Fig. 15.2.4.1. Figure-of-merit weighted model-phased structure factor, obtained as the probability-weighted average over all possible phases.

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.

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

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, $m_{\text{com}}|\mathbf{F}_O|\exp(i\alpha_{\text{com}})$, 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.

$$\begin{array}{lll} \text{MIR:} & m_{\text{MIR}}|\mathbf{F}_O|\exp(i\alpha_{\text{MIR}}) & \simeq \mathbf{F}\\ \text{Model:} & m_C|\mathbf{F}_O|\exp(i\alpha_C) & \simeq \mathbf{F}/2 + D\mathbf{F}_C/2\\ \text{Combined:} & m_{\text{com}}|\mathbf{F}_O|\exp(i\alpha_{\text{com}}) & \simeq [1-(w/2)]\mathbf{F} + (w/2)D\mathbf{F}_C, \end{array}$$

where

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

and

$$H = \int\limits_0^{2\pi} p(\alpha)\ln\frac{p(\alpha)}{p_0(\alpha)}\,\text{d}\alpha; \quad p_0(\alpha) = \frac{1}{2\pi}.$$

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

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

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: $m_{\text{com}}|\mathbf{F}_O|\exp(i\alpha_{\text{com}}) - D\mathbf{F}_C$.

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

references