International Tables for Crystallography (2006). Vol. F, Section 15.2.8, pp. 328–329.

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

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

where

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

and

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

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.