Acta Cryst. (1997). A53, 564–575

Improved Estimation of Structure-Factor Difference Amplitudes from Poorly Accurate Data

T. URSBY[a]*† AND D. BOURGEOIS[a,b]

[a]*ESRF, BP 220, 38043 Grenoble CEDEX, France, and* [b]*UPR 9015/IBS, 41 Avenue des Martyrs, 38027 Grenoble CEDEX 1, France. E-mail: thomas.ursby@mbfys.lu.se*

(*Received 23 September 1996; accepted 20 March 1997*)

Abstract

The influence of measurement errors on structure-factor difference amplitudes are discussed and a formula is derived using Bayesian statistics that gives better estimates of the difference amplitudes and that reduces the noise in difference maps. The formula is of importance for reflections with poor signal-to-noise ratio. Significant improvement is obtained for poor data sets, such as those recorded in fast time-resolved experiments, and also for subsets in any data set that normally contains poorly accurate data, *e.g.* close to the high-resolution limit.

1. Introduction

The difference Fourier technique, owing to its simplicity and its capability to detect small modifications in electron density, is extensively used in the field of crystallography. Errors in difference-electron-density maps are appreciably smaller than in corresponding Fourier maps, such as $2F_{obs} - F_{calc}$ syntheses (see Table 1 for the notation used) and subtle features may be picked out more easily (Henderson & Moffat, 1971). Also, from one known parent crystal structure, it is possible to determine a number of closely related isomorphous structures by only measuring the amplitudes of their structure factors. The parent structure might be an atomic model at an intermediate stage of refinement, what we call a partial model with errors, or one of two related structures. In the latter case, one is for example seeking the position of heavy atoms or other reagents such as enzyme inhibitors or the determination of an intermediate structure in the field of time-resolved crystallography.

The technique was pioneered by Cochran (1951) and effects of errors were analysed in detail by Henderson & Moffat (1971). A general solution has been proposed to minimize the effect of errors in the phases that are experimentally determined and/or calculated from a model (Blow & Crick, 1959). Although the consequences of experimental errors in the measurement of structure-factor amplitudes have been pointed out (Henderson & Moffat, 1971), no remedy to this problem has been proposed. This may seem surprising given that the features in a difference Fourier synthesis arise exclusively from structure-factor-amplitude differences (the same phase is used for the two structure factors) but it can be explained by two facts: Firstly, the main source of noise in a difference map usually results from errors in the phases. In fact, when the difference is small ($f \ll F$), the amplitude difference is the projection of the difference vector along the two structure factors (which both have approximately the same phase in this case), see Fig. 1. The unmeasured orthogonal component, which on average is as large as the parallel component, is neglected. This results in peak heights in difference density maps of half their true values on average (Henderson & Moffat, 1971) and noise that is large relative to the amount of true density in the map (but on an absolute scale the error is much smaller than it would be in a $2F_{obs} - F_{calc}$ map). The noise is further enhanced owing to the errors of the parent structure phases. Secondly, in a majority of cases, static structures are studied and exposure times are chosen so that structure-factor amplitudes are measured with adequate accuracy.

However, cases occur where the statistical noise in the measurements appreciably degrade the quality of

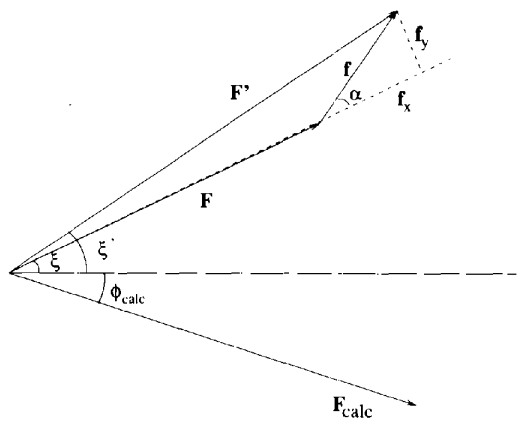


Fig. 1. The relation between the structure factors $\mathbf{F}_{calc}$ (of the parent structure model), $\mathbf{F}$ (of the parent structure) and $\mathbf{F}'$ (of the related structure). f_x is the projection along $\mathbf{F}$ of $\mathbf{f} = (\mathbf{F}' - \mathbf{F})$. When $f \ll F$ then $(F' - F) \simeq f_x$.

† Present address: Molecular Biophysics, Center for Chemistry and Chemical Engineering, Lund University, Box 124, S-22100 Lund, Sweden.

Table 1. *Notation*

The P of σ_P and N of σ_N refer to the P atoms in the model, which is assumed to lack Q atoms, and the total N atoms in the true structure ($N = P + Q$) and where the P atoms have errors in their coordinates. Here, F_N is thus the same as F [and F_N and F_P have different meanings for the primed parameters (see text)]. For the distributions, the subscripts C and N refer to centrosymmetric and non-centrosymmetric, respectively. For values of ε, see for example Stewart & Karle (1976).

$\mathbf{V}$, V	a vector and its amplitude
$\mathbf{F}$	structure factor of parent structure
$\mathbf{F}_{\rm calc}$	model structure factor of parent structure
$\mathbf{F}'$	structure factor of related structure
$\mathbf{f}$	$\mathbf{F}' - \mathbf{F}$
ΔF	$F' - F$
ΔI	$I' - I$
$I = F^2$	intensity of parent structure
$I_{\rm obs}$	observed value of I
$\sigma^2_{{\rm obs},I}$	variance of $I_{\rm obs}$
$F_{\rm obs}$	estimate of F from $I_{\rm obs}$ using *e.g.* the methods of French & Wilson (1978)
$\sigma^2_{{\rm obs},F}$	variance of $F_{\rm obs}$
$\Delta F_{\rm best}$	the best estimate of ΔF
$\mathbf{f}_{\rm best}$	the best estimate of $\mathbf{f}$ for a difference map
$\mathbf{f}_{\rm true}$	the true value of $\mathbf{f}$
σ^2_D	$\sigma^2_N(1 - \sigma^2_A)$
σ^2_A	$(D\sigma_P/\sigma_N)^2$
σ^2_P	$\langle F^2_P/\varepsilon\rangle$
σ^2_N	$\langle F^2_N/\varepsilon\rangle$
ε	correction factor for the expected intensity in a reciprocal-lattice zone
D	$\langle\cos(2\pi\mathbf{H}\cdot\Delta\mathbf{r}_j)\rangle$ (average over all atoms, $\mathbf{H}$ is the reciprocal-lattice vector and $\Delta\mathbf{r}_j$ is the coordinate error of atom j in the model) (Luzzati, 1952)
$\mathbf{m}$	$\int P(\xi)\exp(i\xi)\,d\xi$ = figure of merit
w	$(\varepsilon\sigma^2_D/2)/(\sigma^2_{{\rm obs},F} + \sigma'^2_{{\rm obs},F} + \varepsilon\sigma^2_D/2)$, the factor derived in (14)
$P(A/B)$	the conditional probability of A knowing B
$\bar{A}$	the average of A
$\dfrac{\sum(AB-\bar{A}\bar{B})}{\left[\sum(A^2-\bar{A}^2)\sum(B^2-\bar{B}^2)\right]^{1/2}}$	correlation coefficient between A and B
$R_{\rm sym}$	$\dfrac{\sum_{hkl}\sum_i\lvert I_{i,hkl}-\overline{I_{hkl}}\rvert}{\sum_{hkl}\sum_i I_{i,hkl}}$
$R_{\rm cryst}$	$\dfrac{\sum_{hkl}\lvert F_{\rm obs}-F_{\rm calc}\rvert}{\sum_{hkl}F_{\rm obs}}$

the difference maps. This will generally concern the contribution from reflections close to the high-resolution limit as well as data sets that for some reason have been collected with significant statistical error. One example is real-time-resolved experiments, where the acceptable exposure time often limits the statistical precision of the data. Further, the aim of following all the steps in a reaction implies that the differences between adjacent structures along the reaction coordinate will be subtle. As a consequence, not only is the noise high but also the signal is low.

A common practice is to throw away measured differences that are larger than a reasonable threshold. For true outliers, this is a correct procedure but it is also frequently applied to removal of unreasonably large differences in statistically inaccurate data. This might improve the data quality since these large differences have a significant influence on the map. However, a more adequate procedure should use the information given by the measurements, whatever it is, together with other information at hand and deduce estimates of the difference amplitudes.

In this paper, we propose a method that utilizes *a priori* knowledge to improve the estimates of structure-factor-amplitude differences. The estimates can be used for difference maps as well as for refinement. It is also in principle applicable to any other situation where difference amplitudes are used, *e.g.* for difference Patterson maps or with anomalous data. Theoretical considerations are developed in §2 and a simplified and efficient formula is suggested in §2.4. The technique has been tested on several fast time-resolved Laue data sets and results are presented in §3. It should be pointed out that the simplified formula follows directly from assuming Gaussian errors around the observed ΔF and a Gaussian distribution of ΔF. The somewhat complicated derivation serves the purpose of justifying these assumptions and to point out how a more exact expression can be obtained, if desired.

2. Theory

2.1. *Background*

French & Wilson (1978) treated the problem of calculating structure-factor amplitudes from intensity measurements using Bayesian statistics. They were addressing the problem of what to do with weak measurements. The fact that the observed intensity can even be negative owing to statistical fluctuations points out that going from the intensity to the structure-factor amplitude is not simply a question of taking the square root. In the same manner, by using the prior knowledge we have about the distribution of, in our case, the differences between two sets of related structure-factor amplitudes and of the magnitude of the errors, we can make a better estimate of ΔF than simply $(F'_{\rm obs} - F_{\rm obs})$. The basic idea is the following: The measurement error is larger, on an absolute scale, for strong reflections than for weak reflections. Therefore, the distribution of the measured differences for the strong observations will be broader than for the weak ones even if the distributions of the real differences are the same (imagine the case where there are no differences, only measurement errors), leading to overestimation of the differences for the strong observations compared with the weaker ones. Since the errors are large for the strong reflections, the resulting noise can easily dominate the map. More accurately expressed, the larger the estimated uncertainty for an observed difference amplitude compared with the expected width of the amplitude difference distribution, the less dependent on the measurement and the more dependent on the prior expected value the estimate

should be. For example, since the expected amplitude-difference distribution is narrower at higher resolution, the same measured difference amplitude, with the same estimated uncertainty, should be reduced for a higher-resolution reflection compared with a lower-resolution one.

We will first derive the expressions for the best possible difference map and for an optimum estimate of the difference amplitude using the information of the expected distributions of $\mathbf{F}$ and $\mathbf{F}'$. The expression for the best difference map includes the structure-factor phase probability, which has been treated previously assuming no measurement errors (*e.g.* Read, 1986). When computing electron-density maps, common practice is to first calculate Fs from measured intensities, preferably by using methods such as described by French & Wilson (1978), and then to calculate a phase weight to reduce the noise in the map. By doing this, one is separating the probabilities of the phase and the amplitude, which is not strictly correct since the probability distribution of the phase depends on the amplitude. The separation simplifies the calculations and is in most cases justified (see §2.4). Similarly, we will show that the expressions obtained here can be simplified, under some reasonable assumptions, to a formula that is more practical than the rigorous expression and that can be used both for difference maps and for estimates of difference amplitudes.

There is at present great interest in using Bayesian approaches for the estimation of measurands. Examples of related work are, apart from French & Wilson (1978), Terwilliger (1994), who estimates a difference vector in the case of multiwavelength anomalous-diffraction (MAD) data, and Terwilliger & Berendzen (1996), who apply a similar method for weighting in structure refinement.

2.2. *Best difference map*

Blow & Crick (1959) showed that, in order to minimize the r.m.s. (root mean square) error in an electron-density map, the centre of gravity of the probability distribution of the vector in question should be used, thus defining the best map:

$$\mathbf{F}_{\rm best} = \int \mathbf{F} P(\mathbf{F})\, \mathrm{d}\mathbf{F}, \qquad (1)$$

where $P(\mathbf{F})$ is the probability distribution of $\mathbf{F}$. As pointed out by Srinivasan (1968), $P(\mathbf{F})$ is the conditional expectation value of $\mathbf{F}$, *i.e.* the expected value of $\mathbf{F}$ given the information we have. Similarly, in the case of a difference map, one should use as coefficients

$$\mathbf{f}_{\rm best} = \int \mathbf{f} P(\mathbf{f})\, \mathrm{d}\mathbf{f}, \qquad (2)$$

where $\mathbf{f} = (\mathbf{F}' - \mathbf{F})$. We assume that we have a model of the parent structure and observations of both the parent and the related structure. The probability distribution of $\mathbf{f}$ can be expressed as the joint conditional probability of $\mathbf{F}$ and $\mathbf{F}'$ given the observed intensities and E, which symbolizes our prior assumed knowledge, *i.e.* the assumed distribution of the measurement errors (including the estimations of the variances of the observations), the assumed probability distributions of $\mathbf{F}$ and $\mathbf{F}'$, and possibly the model, *i.e.* $\mathbf{F}_{\rm calc}$. The choice of $\mathbf{F}$ and $\mathbf{F}'$, instead of *e.g.* $(\mathbf{F} + \mathbf{F}')/2$ and $(\mathbf{F}' - \mathbf{F})$ (*cf.* Terwilliger, 1994), is motivated by the more direct relation to the measured quantities. The expressions become asymmetric in $\mathbf{F}$ and $\mathbf{F}'$ if we assume $\mathbf{F}'$ to depend on $\mathbf{F}$ and $\mathbf{F}$ to depend on some external prior knowledge, such as a model, but, in the absence of any external prior knowledge, $\mathbf{F}$ and $\mathbf{F}'$ are of course interchangeable. Expression (2) will thus be

$$\mathbf{f}_{\rm best} = \iint (\mathbf{F}' - \mathbf{F}) P(\mathbf{F}, \mathbf{F}'/I_{\rm obs}, I'_{\rm obs}, E)\, \mathrm{d}\mathbf{F}\, \mathrm{d}\mathbf{F}'. \qquad (3)$$

If only observations of the related structure are available, which is equivalent to examining the difference between a model and the corresponding set of observations, the same expressions can be used, replacing $(\mathbf{F}' - \mathbf{F})$ by $(\mathbf{F} - \mathbf{F}_{\rm calc})$ and by leaving out the probability distribution of $\mathbf{F}'$ (and changing the normalization constants appropriately). The difference between $\mathbf{F}$ and $\mathbf{F}_{\rm calc}$ could be due to a model that might not be complete and contains errors in the atomic positions. The difference between $\mathbf{F}'$ and $\mathbf{F}$ is formally equivalent to the difference between $\mathbf{F}$ and $\mathbf{F}_{\rm calc}$, though in this case it corresponds to changes in the structure and not to errors in the model.

By using Bayes's theorem (see *e.g.* French & Wilson, 1978), which states that

$$P(A/B) \propto P(B/A) P(A), \qquad (4)$$

we can express the probability distribution in (3) as

$$P(\mathbf{F}, \mathbf{F}'/I_{\rm obs}, I'_{\rm obs}, E) \propto P_1(I_{\rm obs}, I'_{\rm obs}/\mathbf{F}, \mathbf{F}', E) P_2(\mathbf{F}, \mathbf{F}'/E). \qquad (5)$$

It is reasonable to assume that the observed intensities are normally distributed around their true value and that their variances are known (French & Wilson, 1978). As pointed out by Hendrickson & Lattman (1970), the distribution around F can also be well approximated by a Gaussian even though the errors around I and F cannot both be strictly Gaussian. Assuming a Poisson distribution of the measured intensity and using the central limit theorem, it is theoretically more satisfying to use a Gaussian distribution of the intensity and we will do so in the following. If we assume that the measurement errors of $I_{\rm obs}$ and $I'_{\rm obs}$ are independent, we have

$$P_1(I_{\rm obs}, I'_{\rm obs}/\mathbf{F}, \mathbf{F}', E)\, \mathrm{d}I\, \mathrm{d}I' \propto \exp\left[-\frac{(I_{\rm obs} - I)^2}{2\sigma^2_{{\rm obs},I}}\right] \exp\left[-\frac{(I'_{\rm obs} - I')^2}{2\sigma'^2_{{\rm obs},I}}\right] \mathrm{d}I\, \mathrm{d}I'. \qquad (6)$$

Basic probability theory states that

$$P_2(\mathbf{F}, \mathbf{F}'/E) = P_3(\mathbf{F}/E)P_4(\mathbf{F}'/\mathbf{F}, E). \quad (7)$$

The probability distribution of $\mathbf{F}$ can be obtained from Luzzati (1952) and from the Wilson distribution (Wilson, 1949) as derived by Srinivasan & Ramachandran (1965). In the non-centrosymmetric case, the distribution is a circular symmetric Gaussian around $D\mathbf{F}_{\text{calc}}$:

$$P_{3N}(\mathbf{F}/E)\,\mathrm{d}\mathbf{F} = \frac{1}{\pi\varepsilon\sigma_D^2}\exp\left[\frac{-|\mathbf{F} - D\mathbf{F}_{\text{calc}}|^2}{\varepsilon\sigma_D^2}\right]\mathrm{d}\mathbf{F}, \quad (8)$$

with σ_D defined by

$$\sigma_D^2 = \sigma_N^2(1 - \sigma_A^2), \quad (9)$$

where σ_A is defined by Srinivasan & Ramachandran (1965). D (Luzzati, 1952), σ_A, σ_N and ε are defined in Table 1. The subscript N for the probability distribution stands for non-centrosymmetric.

The probability distribution of $\mathbf{F}'$ given $\mathbf{F}$ is, as already discussed, formally equivalent to the probability of $\mathbf{F}$ given $\mathbf{F}_{\text{calc}}$. The same expression can be used but the parameters should be changed appropriately and we represent this by adding a prime to the symbols (D' corresponds to the coordinate differences between the structures).*

In the probability distribution of $\mathbf{F}$, other sources of phase information can be included. Further, to minimize the bias towards the model, the amplitude distribution can be chosen as the Wilson distribution, thus not using any information about the amplitude of $\mathbf{F}_{\text{calc}}$ or even only using the information that F is positive. For example, using the Wilson distribution and denoting the phase probability by $P_\xi(\xi/F, E)$, P_3 can be written

$$P_{3N}(\mathbf{F}/E)\,\mathrm{d}\mathbf{F} = P_\xi(\xi/F, E)(1/\pi\sigma_N^2)\exp(-F^2/\sigma_N^2)\,\mathrm{d}\mathbf{F}. \quad (10)$$

The integral (3) can now be evaluated and $\mathbf{f}_{\text{best}}$ obtained: First rewrite (8) for P_4, choose (10) for P_3 and insert these in (7); then insert (6) and (7) in (5). We obtain, expressing $\mathbf{F}$ and $\mathbf{F}'$ in polar coordinates [(F, ξ) and (F', ξ'), respectively]:

$$\begin{aligned}\mathbf{f}_{\text{best},N} = K \int_0^\infty F\,\mathrm{d}F \int_0^{2\pi} \mathrm{d}\xi \int_0^\infty F'\,\mathrm{d}F' \int_0^{2\pi} \mathrm{d}\xi'(\mathbf{F}' - \mathbf{F}) \\ \times \exp[-(I_{\text{obs}} - F^2)^2/2\sigma_{\text{obs},I}^2] \\ \times \exp[-(I'_{\text{obs}} - F'^2)^2/2\sigma_{\text{obs},I}'^2]\, P_\xi(\xi/F, E) \\ \times \exp(-F^2/\sigma_N^2)\exp(-|\mathbf{F}' - D'\mathbf{F}|^2/\varepsilon\sigma_D'^2),\end{aligned} \quad (11)$$

where K is a normalization constant.

2.3. *Best difference amplitude*

The expression derived above gives the Fourier coefficients to use to obtain the best difference map according to the definition of Blow & Crick (1959). For reciprocal-space refinement, an estimate of $\Delta F = (F' - F)$, not influenced by phase probabilities, is needed. The refinement can then be carried out by calculating F' as $F + \Delta F_{\text{best}}$ [or F as $F_{\text{calc}} + (F - F_{\text{calc}})_{\text{best}}$ in the case where there is no related structure, where $(F - F_{\text{calc}})_{\text{best}}$ is the best estimate of $F - F_{\text{calc}}$]. Another possibility is to use difference refinement (Fermi, Perutz, Dickinson & Chien, 1982). F' would then be calculated as $F_{\text{calc}} + \Delta F_{\text{best}}$ and F' used in the refinement program. This will minimize $[\Delta F_{\text{best}} - (F'_{\text{calc}} - F_{\text{calc}})]$.

To estimate ΔF, we use (3) again but we now calculate the expectation value of ΔF instead of $\mathbf{f}$. This will minimize the probability-weighted r.m.s. error between the true ΔF and the estimate in the same way as (3) minimizes the probability-weighted r.m.s. error between $\mathbf{f}_{\text{true}}$ and the estimate (in the complex plane). The integral is

$$\Delta F_{\text{best}} = \iint \Delta F P(\mathbf{F}, \mathbf{F}'/I_{\text{obs}}, I'_{\text{obs}}, E)\,\mathrm{d}\mathbf{F}\,\mathrm{d}\mathbf{F}'. \quad (12)$$

By expressing the integral of ξ' with the zero-order modified Bessel function, one can rewrite this as (Appendix A)

$$\begin{aligned}\Delta F_{\text{best},N} = K \int_0^\infty F\,\mathrm{d}F \int_0^{2\pi} \mathrm{d}\xi \int_0^\infty F'\,\mathrm{d}F' \times \Delta F \\ \times \exp[-(I_{\text{obs}} - F^2)^2/2\sigma_{\text{obs},I}^2] \\ \times \exp[-(I'_{\text{obs}} - F'^2)^2/2\sigma_{\text{obs},I}'^2] \\ \times P_\xi(\xi/F, E)\exp(-F^2/\sigma_N^2) \\ \times \exp[-(F'^2 + D'^2F^2)/\varepsilon\sigma_D'^2] \\ \times I_0(2D'F'F/\varepsilon\sigma_D'^2).\end{aligned} \quad (13)$$

The expressions corresponding to (8), (10), (11) and (13) for the centrosymmetric case are given in Appendix B.

* There is however one difference: Between the two structures, atoms can be added or removed whereas atoms are always added to a partial model compared with the true structure. The probability distribution for the new structure factor is not the same when atoms are removed as when they are added so the derivation of (8) is not valid when atoms are removed. This can be formally solved by assigning the structure with less scattering matter as the parent structure and regarding the remaining exchanged scattering matter as 'coordinate errors', which holds as long as the added atoms are of the same type as the removed atoms. The often assumed Gaussian distribution of coordinate errors would of course not hold in this case and methods for estimating parameters depending on this assumption will in principle be invalid. However, numerical tests show that the distribution of $\mathbf{f}$, when atoms are removed from a model, is well described by (8) with D and σ_D estimated from the simulated data by the program *SIGMAA* (Read, 1986). (Results not shown.)

2.4. *Simplified formula*

Expression (11) can be simplified if a few assumptions are made, namely: (i) the measurement errors have a Gaussian distribution around F (instead of around $I = F^2$ as assumed so far); (ii) $\bar{f} \ll \bar{F}$; (iii) $f \ll F$; and (iv) $\sigma_{\text{obs},F} < F$ (not required for difference amplitudes). These four assumptions lead to (Appendix *C*)

$$\begin{aligned}\mathbf{f}_{\text{best},N} &= \mathbf{m}[(\varepsilon\sigma_D^2/2)/(\sigma_{\text{obs},F}^2 + \sigma_{\text{obs},F}'^2 + \varepsilon\sigma_D^2/2)] \\ &\quad \times (F'_{\text{obs}} - F_{\text{obs}}) \\ &= \mathbf{m}w(F'_{\text{obs}} - F_{\text{obs}}), \end{aligned} \tag{14}$$

thus defining w (we drop the prime on σ_D since there is only one σ_D in the formula). If instead an expression for the amplitude difference is sought, then **m** should be omitted. The corresponding expression for the centrosymmetric case of equation (14) is obtained by changing σ_D^2 to $2\sigma_D^2$.

In the following, we discuss assumptions (i) to (iv).

The first assumption is in general acceptable, as already mentioned in §2.2.

When the change between the two structures is small, the second assumption, $\bar{f} \ll \bar{F}$, will be true.

The third assumption, $f \ll F$, will be true for most reflections if assumption (ii) is true. For those reflections not fulfilling assumption (iii), f is in general still small on an absolute scale (F and F' both small) and these reflections will give a small contribution to the total differences. In the case of (difference) maps, these reflections will also have an uncertain phase, *i.e.* a small figure of merit, and their influence will thus be reduced further.

The fourth assumption is commonly made when first an estimate of F is calculated and *then* a phase probability weight. This separation between amplitude and phase weighting in (11) is the basis for the simplification leading to (14). In general, this separation is not strictly correct, since the phase probability depends on the value of F. However, if assumption (iv) is fulfilled, this dependence can be neglected and the full probability distribution of F can be replaced by a point estimate from it. The variation of the phase probability with F can be illustrated by the variation of m with F as shown in Fig. 2. If the maximum acceptable relative change of m is p then it can easily be shown that the variation of F should be less than pF (in the limiting case of a small F; for larger Fs, the condition is weaker). As long as $\sigma_{F,\text{obs}}$ is not large compared with F, the relative variation of m is not too large, which implies that the variation of P_ξ over the range of F where the probability distribution of F is non-negligible is small. In general, F is determined with a reasonable accuracy (*e.g.* in a static experiment recorded prior to a less-accurate time-resolved measurement of F'), and assumption (iv) is fulfilled, allowing F to be set to a constant value in (11). In practice, there are always reflections for which $\sigma_{\text{obs},F}$ is not much smaller than F. However, this would mostly occur for the weakest reflections, for which the wide distribution of ξ will result in a small $\mathbf{f}_{\text{best}}$ (or small **m** if the ξ integral is separated) and a small contribution to the difference map.

To reduce the influence of systematic errors, data sets of both F and F' from the same crystal, under the same

Table 2. *Comparison for MBS (see text) between non-optimized data ($F'_{obs} - F_{obs}$), optimized data from (14) using σ_D estimated from the simulated data and (†) from the model structure factors*

The first line ('No f.o.m.') shows the correlation for the map without figure-of-merit weighting (Read, 1986) which is used for all the other maps. The use of (14) is thus more important in this case than figure-of-merit weighting. The last line shows the effect of setting the high-resolution limit at 2.7 Å instead of at 2.3 Å. It is seen that the improvement cannot be obtained by a simple resolution cut-off. Note that the correlation coefficient for amplitudes in this case is calculated only with data above 2.7 Å. The correlation coefficient is defined in Table 1. It is calculated between the true difference data (see text) and the data derived from the simulation. The average w in (14) is given in the column $\bar{w}$.

Map	Correlation maps	Coefficient amplitudes	$\bar{w}$
No f.o.m.	0.513		
Non-optimized	0.530	0.540	1
Optimized	0.626	0.666	0.529
Optimized†	0.633	0.685	0.367
Resolution cut-off	0.557	(0.671)	1

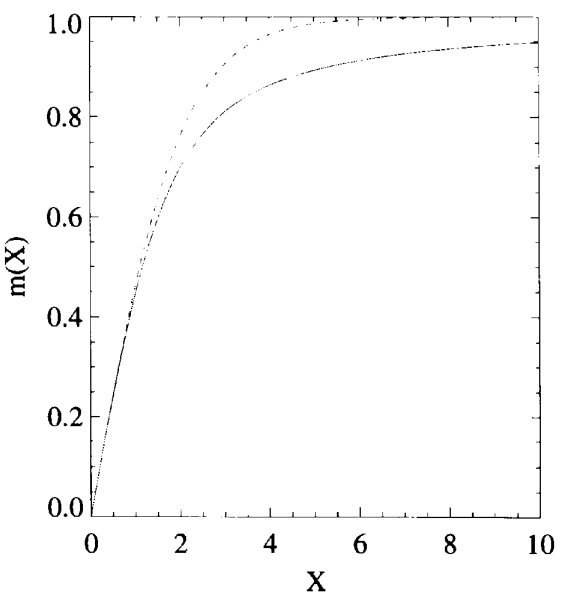


Fig. 2. The figure of merit, m, as a function of $X = 2F'F/\varepsilon\sigma_D^2$. In the non-centrosymmetric case (solid line), $m = I_1(X)/I_0(X)$ and, in the centrosymmetric case (dashed line), $m = \tanh(X)$ (Srinivasan, 1966). $I_0(X)$ and $I_1(X)$ are the zero- and first-order modified Bessel functions.

conditions, might be used even if there is a separate more accurate data set of F. In this case, assumption (iv) concerns the more accurate data set. Even though condition (iv) is then not fulfilled for the F used, one can justify use of the simplified formula (using the more accurate F for calculating m).

Therefore, it is expected that using the simplified formula should always be at least as good as using only figure-of-merit weighting. This was confirmed by the experimental results presented in §3 (MBS), which show that the improvement when using (14) was significantly larger than the one resulting from figure-of-merit weighting only (see Table 2). Further improvement by use of the impractical rigorous expression is expected to be minor, although numerical tests using (11) have not been carried out for a complete data set.

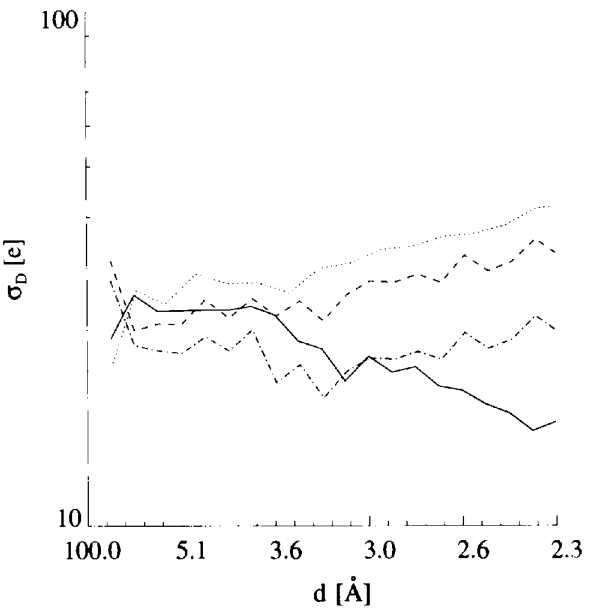


Fig. 3. Estimation of σ_D for simulated data (MBS, see text). The solid line represents the true σ_D (estimated from model data), the dashed line σ_D estimated from noisy data and the dashed-dotted line σ_D estimated from noisy data but with deconvolution with the observed variance. (The σ_D estimated from noisy data results from a convolution of the true differences and measurement errors.) The influence of the measurement errors is larger at higher resolution since the higher-resolution data are weaker and it can also be seen in the figure that the true and estimated σ_D diverge as the resolution increases. Using the estimated variances, the influence of the measurement errors can be removed by a deconvolution of the observed distribution with the measurement error distribution (see §2.5). In each bin, σ_D^2 was estimated as $n(F' - F)^2/\varepsilon$ (see §2.5). The deconvolution consisted of subtracting the average $(\sigma^2_{\text{obs},F} + \sigma'^2_{\text{obs},F})$ from $(F' - F)^2$ but not letting σ_D^2 be less than half of $n(F' - F)^2/\varepsilon$ [since $(\sigma^2_{\text{obs},F} + \sigma'^2_{\text{obs},F})$ is about as large as $(F' - F)^2$ and both are estimated only approximately]. The dotted line corresponds to the dashed line but with σ_D^2 estimated as $\sigma_N^2(1 - \sigma_A^2)$ using the program *SIGMAA* (Read, 1986) to estimate σ_A. The two ways of estimating σ_D thus give similar results.

2.5. *Estimation of parameters*

For expression (14), an estimate of σ_D (which is a function of resolution) is needed. To obtain σ_D, estimates of σ_N and σ_A are required, see (9). σ_N^2 is easily estimated from the data as F^2/ε and σ_A can be estimated as described by Read (1986). If $f \ll F$ then σ_D^2 can also be obtained as $n(F' - F)^2/\varepsilon$, where n is one for centric reflections and two for acentric reflections (Read, 1986). However, both methods will overestimate σ_D owing to the measurement errors and this has to be taken into account, see Fig. 3. If the variances of the observations are well estimated, the true σ_D can be obtained by deconvolution of the observed distribution with the measurement error distribution since the observed distribution results from a convolution of the true differences and the measurement errors. Since these are Gaussian distributions, this translates into

$$\sigma_D^2 = \sigma^2_{D,\text{obs}} - \overline{n(\sigma^2_{\text{obs},F} + \sigma'^2_{\text{obs},F})/\varepsilon}, \qquad (15)$$

where $\sigma^2_{D,\text{obs}}$ is the estimation obtained neglecting the presence of measurement errors.

Estimates of the structure-factor amplitudes and their variances are obtained by the usual methods. Correctly estimated variances are of course of major importance.

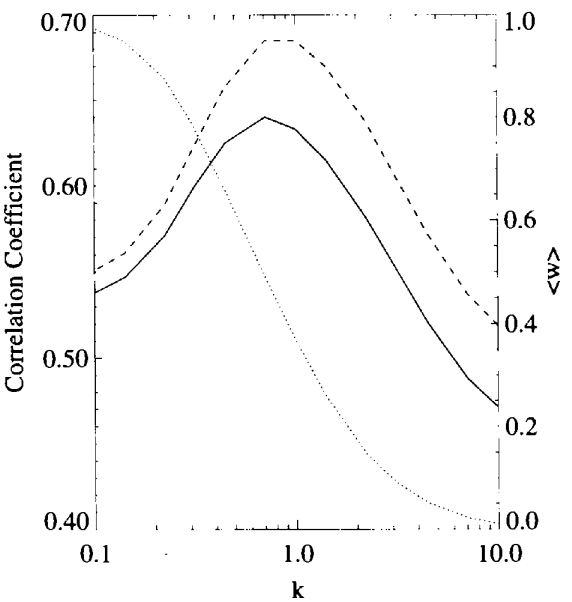


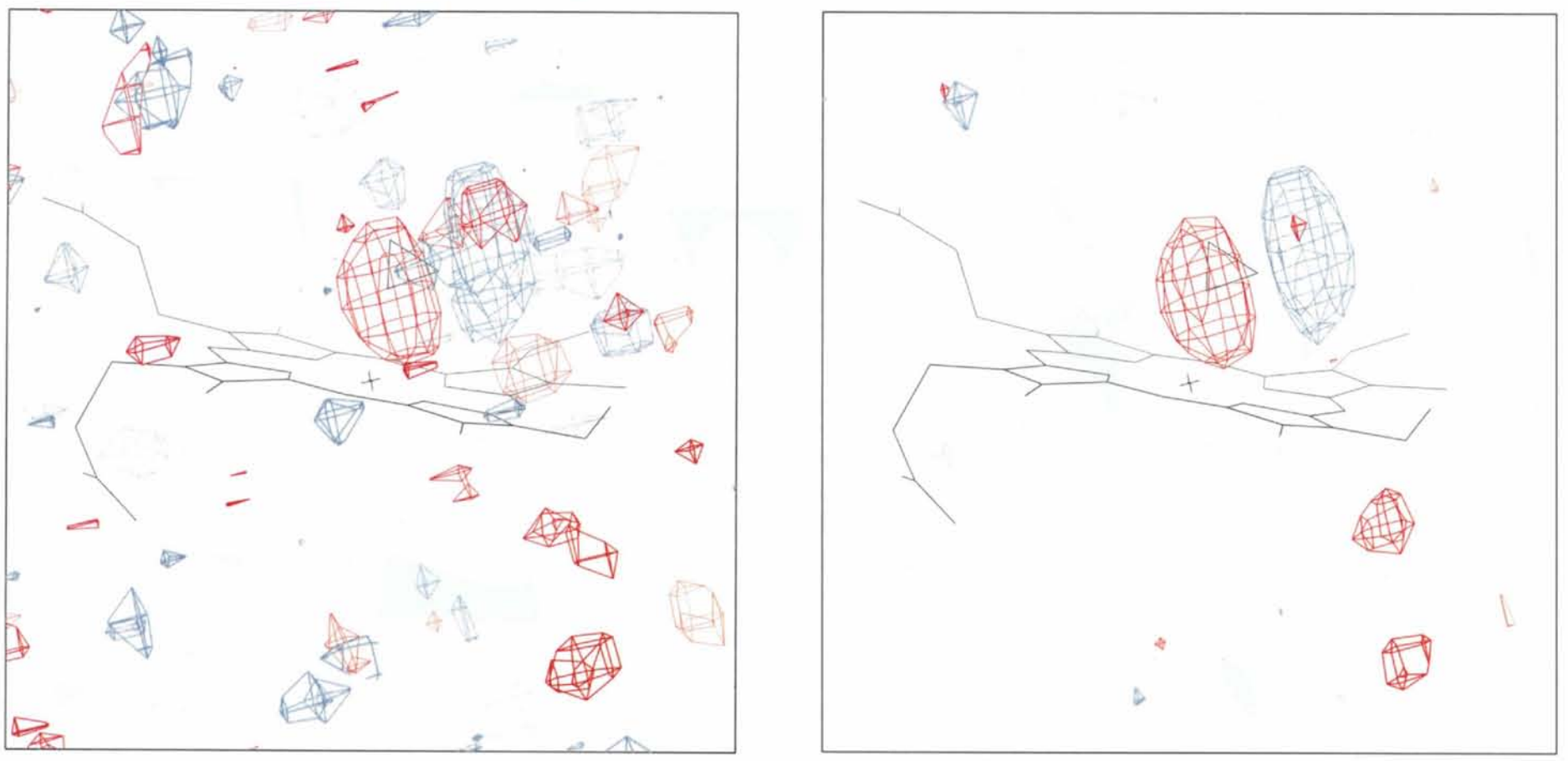
Fig. 4. Effect of badly estimated variances of the difference amplitudes (or badly estimated σ_D) for the case MBS (see text). The $\sigma_{\text{obs},F}$s were multiplied by k before (14) was applied. Thus, $k = 0$ corresponds to using $F'_{\text{obs}} - F_{\text{obs}}$ without modification. $k = 1$ corresponds to σ_D estimated from the models and $\sigma_{\text{obs},F}$ as output from the data-reduction program. Correlation coefficients were calculated between model data and simulated optimized data. The solid curve is the correlation coefficient between the difference maps and the dashed curve between the difference amplitudes. The dotted curve (scale on the right) shows the average of w in (14).

Reflections for which the variances do not correctly reflect the error will not be correctly treated and should be rejected. This can be achieved by evaluating the probability of the observed value and rejecting the reflection if this probability is too small. This is what is done when an outlier of equivalent intensity measurements is rejected and can here also be applied to the probability of an observed ΔF given the estimated variance and the *a priori* distribution of ΔF.

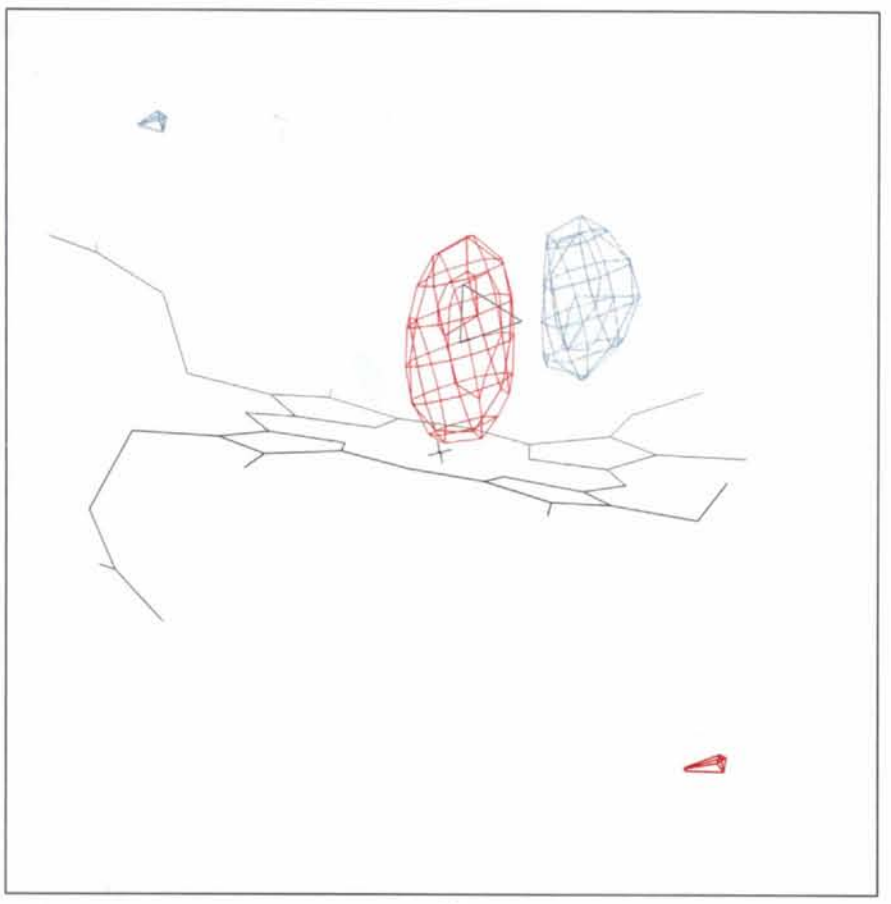
For difference maps, **m** should also be estimated and this can be done with the usual methods, *e.g.* with the program *SIGMAA* (Read, 1986). If phase information from a model is used then either F_{obs} or F'_{obs} [or $(F_{\text{obs}} + F'_{\text{obs}})/2$] can be used in the phase probability expressions since the difference is assumed to be small. Note that if *SIGMAA* is used to estimate σ_D then F_{obs} and F'_{obs} should be input to the program and not F_{obs} and F_{calc}.

The parameter D needed in (11) and (13) can be obtained as discussed by Read (1986).

It is important not to underestimate σ_D excessively, which can be understood in the following way: The



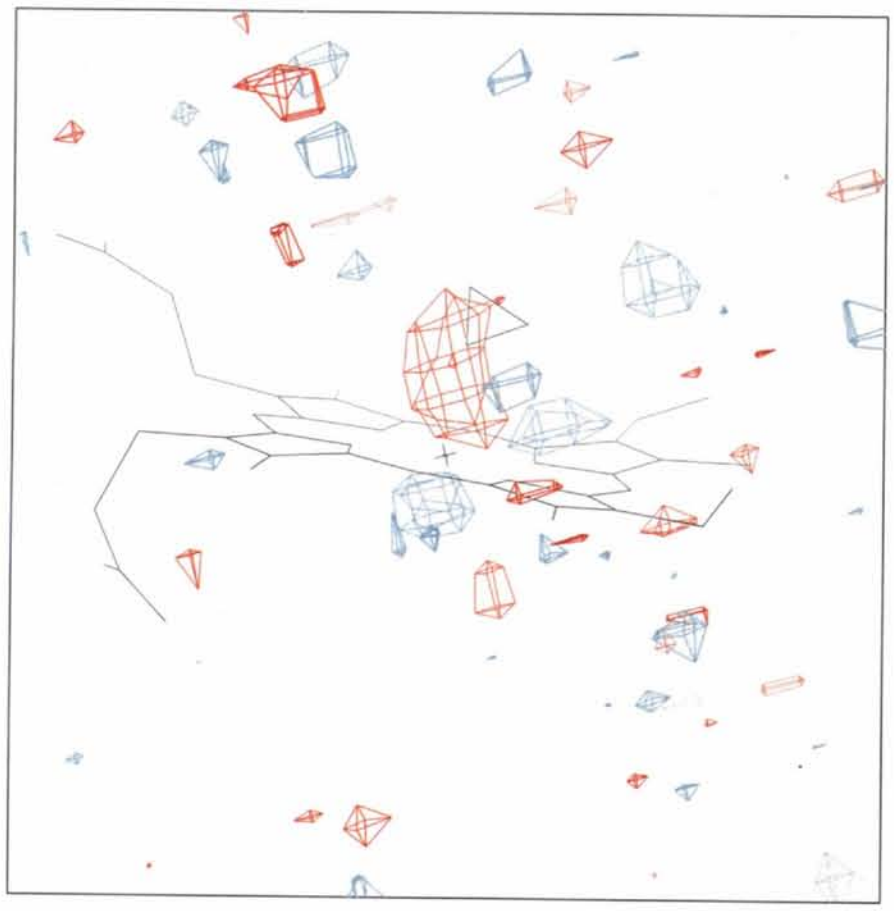
(*a*) (*b*)



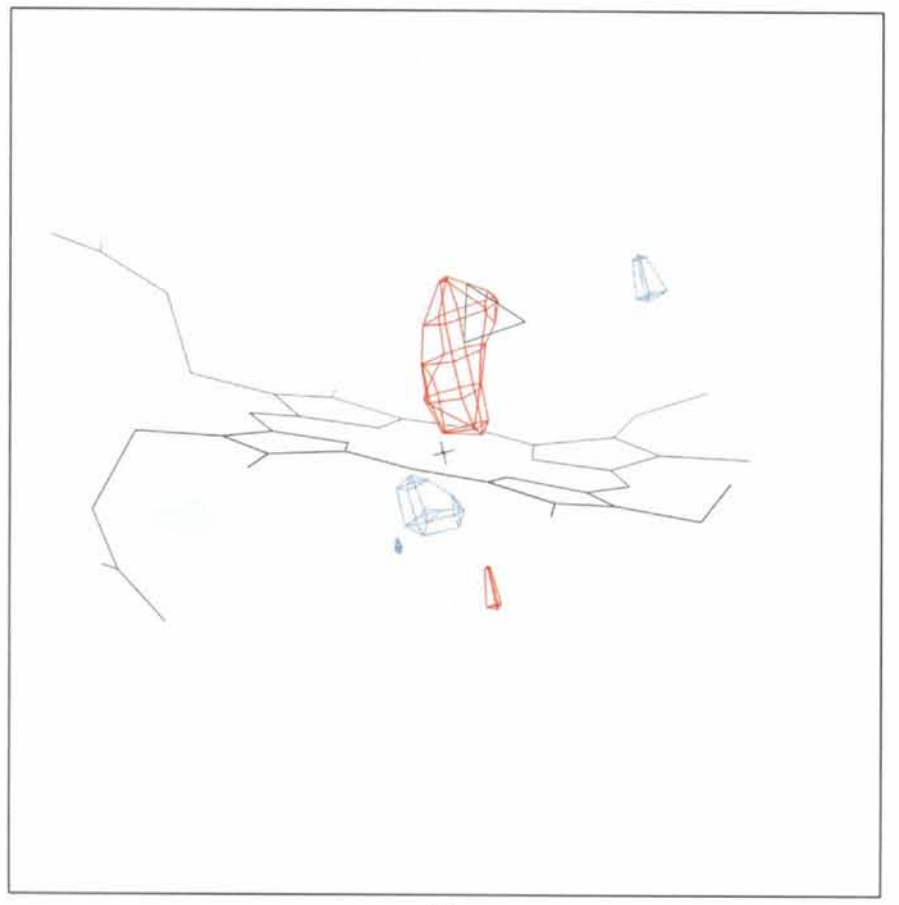
(*c*)

Fig. 5. Difference maps MBS (see text). The map in (*a*) is calculated with the amplitudes $m(F'_{\text{obs}} - F_{\text{obs}})$, where m is the figure of merit as calculated with *SIGMAA* (Read, 1986) with F_{obs} and F_{calc} (from 1MBC) as input and with phases from 1MBC. (*b*) is obtained by using the same coefficients multiplied by w from (14). (*c*) shows what the difference map should look like (using the model difference amplitudes but the same completeness and phases as for the other maps). The heme group of the parent structure is shown with black lines (the Fe atom as a cross). Positive density is coloured blue and negative red. The large negative peak in (*c*) corresponds to the old position of the CO (and to some extent to the motion of the Fe atom) and the large positive peak to the new position. The maps are contoured at ±(the negative peak height divided by 3), where the negative peak height is the minimum density value in the map (which corresponds to the removed CO and which is the most prominent peak in the map). These levels show more clearly the effect of the method than contouring at a given σ level since in the latter case the amount of spurious noise peaks would be approximately the same in all maps (the map r.m.s. being essentially given by the noise). Expressed in map r.m.s., the peak is larger for the optimized map. The negative peak height expressed in σs is 7.1, 9.1 and 10.7 for the three maps, respectively.

observed differences consist of the true differences and noise, $\Delta F_{obs} = \Delta F_{true} + \Delta F_{noise}$. Since the Fourier transform is a linear operation, the density map will consist of the sum of the true density, $\Delta\rho_{true} = \mathrm{FT}[\Delta F_{true}]$, and the noise density, $\Delta\rho_{noise} = \mathrm{FT}[\Delta F_{noise}]$ (neglecting the phase error that will be the same for noise-free and noisy data), where FT symbolizes the Fourier transform. Modification of the observed amplitudes, as described, means that the density is modified in such a way that the reduction of the noise density is larger than the modification of the true density. If we underestimate σ_D, we will distort the true density excessively, whereas, if we overestimate σ_D, we will tend towards the non-optimized observed density, $\Delta\rho_{obs}$, which contains more noise than the optimized map but where the underlying $\Delta\rho_{true}$ is non-distorted. It is well known that random amplitudes and model phases will reconstruct the model (with added noise) so also $\Delta\rho_{obs}$ is biased. The conclusion is that it is important to correctly estimate the parameters.



(*a*)



(*b*)

Fig. 6. Difference maps MB (see text). As (*a*) and (*b*) in Fig. 5. Phases from 1MBC. The levels are ±(the negative peak height divided by 1.5) in these maps. The negative peak height expressed in σs is 4.3 and 6.1 for the two maps, respectively. It is known that the CO molecule is photolysed and that a downward motion of the heme Fe atom should follow. This agrees with the two largest peaks in (*b*). Other changes are expected to be smaller. In the improved map, (*b*), the major peaks are enhanced compared with the smaller peaks, many of which are chemically unreasonable, and it seems reasonable to interpret this as the noise being reduced and the true densities appearing more clearly.

The optimized difference vectors minimize the noise in the map but the true density is systematically reduced [since w is always between 0 and 1 when using (14)]. Therefore, the difference map has to be normalized with $(\sum w)^{-1}$ to better estimate peak heights in difference maps and to set the difference amplitudes on a correct scale.

2.6. *Other potential applications*

To apply (14) to other situations involving small differences, the validity of the assumptions on which (14) rely must be verified in each case. For example, anomalous differences might be the result of only one scatterer and in this case a Gaussian distribution of the differences is not a good approximation (Terwilliger, 1994), though this is not essential for the method. For other applications, one also needs to consider the validity of the suggested methods for estimating the parameters.

In the case of difference Patterson amplitudes, the expectation value of $(\Delta F)^2$ rather than ΔF should be used. This will minimize the probability-weighted r.m.s. error between the true $(\Delta F)^2$ and the estimate (and thus the r.m.s. error in the difference Patterson map) instead of the probability-weighted error of ΔF. Following the derivation in Appendix *C*, one finds that the expectation value of $(\Delta F)^2$ reduces to (29) with f_x replaced by f_x^2. This integral evaluates to $\mu^2 + \sigma^2$ with μ and σ given by (30):

$$
\begin{aligned}
&[(\Delta F)^2]_{\text{best, } N} \\
&\quad = [(\varepsilon\sigma_D^2/2)/(\sigma_{\text{obs, } F}^2 + \sigma_{\text{obs, } F}'^2 + \varepsilon\sigma_D^2/2)](F'_{obs} - F_{obs})^2 \\
&\qquad \times [(\varepsilon\sigma_D^2/2)/(\sigma_{\text{obs, } F}^2 + \sigma_{\text{obs, } F}'^2 + \varepsilon\sigma_D^2/2) \\
&\qquad + (\sigma_{\text{obs, } F}^2 + \sigma_{\text{obs, } F}'^2)/(F'_{obs} - F_{obs})^2] \\
&\quad = w^2(\Delta F_{obs})^2[1 + (\sigma_{\text{obs, } F}^2 + \sigma_{\text{obs, } F}'^2)/w(\Delta F_{obs})^2] \\
&\quad = (\Delta F_{\text{best, } N})^2[1 + w(\sigma_{\text{obs, } F}^2 + \sigma_{\text{obs, } F}'^2)/(\Delta F_{\text{best, } N})^2].
\end{aligned}
\tag{16}
$$

The best estimate of $(\Delta F)^2$ is thus larger than the square of the best estimate of ΔF.

3. Applications

The simplified formula (14) has been tested for three different Laue data sets, of which the first two are from fast time-resolved experiments. The Laue method is often employed in the field of fast time-resolved protein crystallography (*e.g.* Hajdu & Andersson, 1993), owing to its unique advantage of gathering a large amount of information in a minimal time. This is done at the expense of data quality since diffracted intensities lie on top of a polychromatic background, are often spatially overlapped and generally need to be normalized in a wavelength-dependent fashion before useful structure-factor amplitudes can be extracted.

The method has been tested using a home-made routine written with IDL (Research Systems Inc.) with input/output in MTZ format (Collaborative Computational Project, Number 4, 1994). A standard Collaborative Computational Project, Number 4 (1994) format version of the program will soon be written. Laue data processing was done using the *Daresbury Laue Software Suite* (Campbell, 1995), a home-made integration program (Bourgeois, Nurizzo, Kahn & Cambillau, 1997) and the Collaborative Computational Project, Number 4 (1994).

The first case, referred to as MBS, consists of simulated data from two models of myoglobin. The idea is to produce realistic data sets, including noise, and to examine the improvement by application of (14) as judged by comparison with the corresponding noise-free data sets. The parent structure is obtained from a refined model of carbonmonoxy myoglobin at 40 K and the related structure is the photolysed product (Teng, Šrajer & Moffat, 1994). In this structure, the bond between the CO molecule and the Fe atom of the heme group is broken and the CO is found in a docking site inside the heme pocket approximately 1.5 Å from the bound position. There are also further less-prominent rearrangements in the protein. Integrated intensities, based on structure factors from the two models, the X-ray spectrum and the sensitivity of the detector, were calculated for a data set typical for an experiment such as MB, described below. Based on experimentally observed variances (expressed as a function of reflection intensity), noise was then added to the intensities. The data were treated in the same way as experimental data. The data were 83% complete (∞–2.3 Å, using only singles, 60% in ∞–4.6 Å and 86% in 4.6–2.3 Å) and with an unweighted R_{sym} of 21%, see Bourgeois *et al.* (1996). The phases were calculated from the model 1MBC (Kuriyan, Wilz, Karplus & Petsko, 1986), where 1MBC is the entry label in the Brookhaven Protein Data Bank. It is a model of carbonmonoxy myoglobin determined under different conditions and refined independently. It is used here to mimic the differences between the model and the parent structure. σ_D^2 was estimated both from the noisy data and from the noise-free model data as $n(F' - F)^2/\varepsilon$, as described in §2.5. Fig. 3 confirms that σ_D obtained from the noisy data is overestimated and that the deconvolution improves the estimation. The variances of the difference amplitudes were as output from the data-reduction program. We can directly compare the obtained difference amplitudes and difference map with the true ones, *i.e.* difference amplitudes obtained directly from the models and the difference map calculated with these. We define the true difference map as the one calculated with the noise-free amplitude differences of the two models but with the same completeness and phases as for the simulated noisy data. Table 2 presents a comparison of the true data and the simulated data and shows the efficiency of (14). It is interesting to examine how sensitive the method is to correctly estimated parameters (*i.e.* σ_{obs} and σ_D). This is shown in Fig. 4. It can be seen that the correlation coefficients show large improvement for correctly estimated parameters but that 'overoptimization' ($k \gg 1$, *i.e.* σ_D underestimated) makes it worse than in the non-optimized case (*cf.* the discussion in §2.5). Note that parameters estimated directly from the data give nearly optimal improvement for this simulation (see Table 2). In Fig. 5 are shown the difference maps and it can be noted that the noise reduces considerably by applying (14).

The second case, referred to as MB, is similar to MBS but based on real data from an experiment performed at room temperature (Šrajer *et al.*, 1996). The parent structure is carbonmonoxy myoglobin and the related structure is the photolysed product 4 ns after the photolysing laser flash. The data consisted of 47 images, each with 450 ps accumulated exposure time. The data are highly redundant but with significant statistical errors. The data were 84% complete (∞–2.0 Å, using only singles, 64% in ∞–4.0 Å and 87% in 4.0–2.0 Å) and with an unweighted R_{sym} of 13% (14% for the related structure data set). In this case, the true difference amplitudes are not known but the non-optimized and optimized difference maps can be compared, Fig. 6, and again the (assumed) noise in the map reduces [the average w in (14), $\overline{w}$, is 0.340].

In the third test case, a modified model of native cutinase (Martinez, De Geus, Lauwereys, Matthyssens & Cambillau, 1992), a fungal lipolytic enzyme of 22 kDa,* was compared with monochromatic and Laue data of the native. The 1.0 Å resolution model (Longhi, Czjzek, Lamzin, Nicolas & Cambillau, 1997), with $R_{cryst} = 9.7\%$, was modified by removing one residue in the catalytic triad and by only keeping one of the conformations for the residues modelled in multiple conformations. In this case, there are only observed intensities of the related structure (the Laue data set of the native structure) and a model of the parent structure (the modified model) but no observations of the parent structure. The monochromatic data (of the native structure) only serves as a reference. Since the

* 1 Da = 1 Dalton = 1.66054×10^{-27} kg.

monochromatic data are of high quality ($R_{cryst} = 9.7\%$ to 1.0 Å but here used to 1.5 Å since the Laue data set extends only to this resolution), it can be taken as a good approximation of the true structure-factor amplitudes of the native [the refined Laue data set has $R_{cryst} = 19.3\%$ and $R_{free} = 24.2\%$ to 1.5 Å resolution (Bourgeois, Longhi, Wulff & Cambillau, 1997)]. If the technique works satisfactorily, applying (14) to the differences between the modified model and the Laue data should bring them closer to the differences between the modified model and the monochromatic data. The Laue data were in this case from 19 images, each with 450 ps accumulated exposure time but the data quality is better than MB because the crystals are better diffracting and the X-ray flux was higher ($R_{sym} = 10.4\%$). The completeness was 72% (18–1.5 Å, using only singles, 48% in 18–3.0 Å and 75% in 3.0–1.5 Å). Even though the modifications of the model are small, the true differences are expected to be larger since it is a comparison between a model and experimental data. Because of this fact and the higher data quality, the influence of statistical noise is smaller. We can thus expect the improvement to be less significant. The difference amplitudes and the corresponding difference maps were calculated for modified model/monochromatic data and for modified model/Laue data. With (14), the correlation coefficient between the maps increases from 0.478 to 0.491 and between the amplitude differences from 0.703 to 0.722 ($\overline{w} = 0.935$). Several different modifications of the model were tried to verify that the improvement of the data was not random. Residues were removed or shaken (randomly moved from their original position). The number of residues involved and the amplitudes of the shaking were also varied. The optimized differences were always better but the improvement was, as expected, smaller the larger the modification.

4. Conclusions

In the first two examples presented in §3, the simple expression (14) is shown to greatly improve the estimates of the structure-factor differences as judged by comparison with the true difference amplitudes and difference map in the case of the simulation, and by inspection of the difference maps in the case of real data. The method should always give improved estimates of difference structure factors, provided that the *a priori* information is reasonably correct, but the improvement is only substantial if the noise in the difference amplitudes is significant relative to the noise-free values. The method could be applied to any situation where difference structure factors are used, *e.g.* for difference Patterson maps or when anomalous data are used, but the simplified formula (14) might need some modification.

We recommend the expression derived here to be routinely applied for difference data if the subsequent analysis does not already take *a priori* information into account.

APPENDIX *A*

Following Sim (1958), we can express the phase integral in (12) using Bessel functions:

$$\begin{aligned}
&\int_0^{2\pi} \exp(-|\mathbf{F}' - D'\mathbf{F}|^2/\varepsilon\sigma_D'^2)\,\mathrm{d}\xi' \\
&= \int_0^{2\pi} \exp\{-[F'^2 + D'^2F^2 - 2F'D'F \\
&\quad \times \cos(\xi' - \xi)]/\varepsilon\sigma_D'^2\}\,\mathrm{d}\xi' \\
&= \int_0^{2\pi} \exp[-(F'^2 + D'^2F^2)/\varepsilon\sigma_D'^2] \\
&\quad \times \exp[2F'D'F\cos(\xi' - \xi)/\varepsilon\sigma_D'^2]\,\mathrm{d}\xi' \\
&= \exp[-(F'^2 + D'^2F^2)/\varepsilon\sigma_D'^2]2\pi I_0(2D'F'F/\varepsilon\sigma_D'^2),
\end{aligned} \tag{17}$$

where I_0 is the zero-order modified Bessel function.

APPENDIX *B*

B 1. Expression (8)

From the same references (Luzzati, 1952; Wilson, 1949; Srinivasan & Ramachandran, 1965), we obtain for the centrosymmetric case:

$$\begin{aligned}
P_{3C}(\mathbf{F}/E)\,\mathrm{d}\mathbf{F} &= [1/(2\pi\varepsilon\sigma_D^2)^{1/2}] \\
&\quad \times \exp(-|\mathbf{F} - D\mathbf{F}_{calc}|^2/2\varepsilon\sigma_D^2)\,\mathrm{d}\mathbf{F},
\end{aligned} \tag{18}$$

where vectors are used to denote the structure factors when the sign is included.

B 2. Expression (10)

For the centrosymmetric case, (10) becomes

$$\begin{aligned}
P_{3C}(\mathbf{F}/E)\,\mathrm{d}\mathbf{F} &= P_s(s/F, E)[1/(2\pi\sigma_N^2)^{1/2}] \\
&\quad \times \exp(-F^2/2\sigma_N^2)\,\mathrm{d}F,
\end{aligned} \tag{19}$$

where P_s is now the probability distribution of the sign of **F** extracted from a model and/or obtained by other means. The distribution of the amplitude is obtained by multiplying the distribution by two.

B 3. Expression (11)

Using (18) and (19) instead of (8) and (10), we obtain:

$$\begin{aligned}
\mathbf{f}_{best,\,C} = K \int_{-\infty}^{\infty} \mathrm{d}\mathbf{F} \int_{-\infty}^{\infty} \mathrm{d}\mathbf{F}' \times (\mathbf{F}' - \mathbf{F}) \\
\times \exp[-(I_{obs} - F^2)^2/2\sigma_{obs}^2] \\
\times \exp[-(I'_{obs} - F'^2)^2/2\sigma_{obs}'^2] \\
\times P_s(s/F, E)\exp(-F^2/2\sigma_N^2) \\
\times \exp(-|\mathbf{F}' - D'\mathbf{F}|^2/2\varepsilon\sigma_D'^2).
\end{aligned} \tag{20}$$

B 4. Expression (13)

For the weighted average of the difference amplitude, $(\mathbf{F}' - \mathbf{F})$ in (20) should be replaced by $|\mathbf{F}' - \mathbf{F}|$. However, the conditional probability of F given $F_{\rm calc}$ has been derived (Srinivasan & Ramachandran, 1965). Their expression (21) becomes in our notation:

$$P(F/E) = (2/\pi\varepsilon\sigma_D^2)^{1/2} \exp[-(F^2 + D^2F_{\rm calc}^2)/2\varepsilon\sigma_D^2] \times \cosh(-FDF_{\rm calc}/\varepsilon\sigma_D^2). \quad (21)$$

Use of this instead of (18) means that the limits of integration of (3) will be $0, \infty$ instead of $-\infty, \infty$ for both F' and F and the weighted average of $(F' - F)$ becomes

$$\begin{aligned}\Delta F_{{\rm best},\,C} = K \int_0^\infty {\rm d}F \int_0^\infty {\rm d}F' \times (F' - F) \\ \times \exp[-(I_{\rm obs} - F^2)^2/2\sigma_{{\rm obs},\,I}^2] \\ \times \exp[-(I'_{\rm obs} - F'^2)^2/2\sigma_{{\rm obs},\,I}'^2] \\ \times P_s(s/F, E) \exp(-F^2/\sigma_N^2) \\ \times \exp[-(F'^2 + D'^2F^2)/2\varepsilon\sigma_D'^2] \\ \times \cosh(D'F'F/\varepsilon\sigma_D'^2). \end{aligned} \quad (22)$$

APPENDIX *C*
Simplified formula

With the assumptions given in §2.4, (11) will be simplified.

We choose to use point estimates of F and F' and to replace the Gaussian distributions of F and F' by a Gaussian distribution of ΔF. That the measurement errors of ΔF have a Gaussian distribution (with the variance being the sum of the two separate variances) follows if the measurement errors of F and F' have Gaussian distributions and if these are independent. The observed structure-factor amplitudes are obtained *e.g.* with the method of French & Wilson (1978).

Starting from (11), replacing the measurement error distribution with a Gaussian distribution around $\Delta F_{\rm obs} = (F'_{\rm obs} - F_{\rm obs})$ and changing variables $F', \xi' \rightarrow f_x, f_y$ (see Fig. 1), we obtain:

$$\begin{aligned}\mathbf{f}_{{\rm best},\,N} = K \int_0^\infty \delta_F(F_{\rm obs})\,{\rm d}F \int_0^{2\pi} {\rm d}\xi \int_{-\infty}^\infty {\rm d}f_x \int_{-\infty}^\infty {\rm d}f_y \times \mathbf{f} \\ \times \exp[-(\Delta F_{\rm obs} - \Delta F)^2/2(\sigma_{{\rm obs},\,F}^2 + \sigma_{{\rm obs},\,F}'^2)] \\ \times P_\xi(\xi/F, E) \exp[(-|\mathbf{F}' - D'\mathbf{F}|^2)/\varepsilon\sigma_D'^2], \end{aligned} \quad (23)$$

where $\delta_F(F_{\rm obs})$ is a Dirac δ function indicating that F is equal to $F_{\rm obs}$. The Dirac δ function is introduced since a point estimate of F is used in P_ξ instead of the full probability distribution. (The probability distribution of F is already evaluated when the point estimate is determined). Note that

$$\Delta F = [(F + f_x)^2 + f_y^2]^{1/2} - F \quad (24)$$

and that $\mathbf{f}$ can be written as:

$$\mathbf{f} = (f_x \cos\xi - f_y \sin\xi) + i(f_x \sin\xi + f_y \cos\xi). \quad (25)$$

Since the probability distribution is even in f_y, the f_y terms of $\mathbf{f}$ will cancel and $\mathbf{f}$ can be replaced by $f_x \cos\xi + if_x \sin\xi = f_x \exp(i\xi)$.

The assumption $f \ll F$ implies that the differences between the two structures are small and in particular that the coordinate shifts are small. Thus, $D' \simeq 1$. Further, $f \ll F$ implies that $\Delta F \simeq f_x$. We evaluate the integral over F and rewrite (23):

$$\begin{aligned}\mathbf{f}_{{\rm best},\,N} \\ = K' \int_0^{2\pi} P_\xi(\xi/F_{\rm obs}, E) \exp(i\xi)\,{\rm d}\xi \\ \times \int_{-\infty}^\infty \exp(-f_y^2/\varepsilon\sigma_D'^2)\,{\rm d}f_y \\ \times \int_{-\infty}^\infty f_x \exp[-(\Delta F_{\rm obs} - f_x)^2/2(\sigma_{{\rm obs},\,F}^2 + \sigma_{{\rm obs},\,F}'^2)] \\ \times \exp(-f_x^2/\varepsilon\sigma_D'^2)\,{\rm d}f_x. \end{aligned} \quad (26)$$

The normalization constant can then be written

$$\begin{aligned}K'^{-1} = \int_0^{2\pi} P_\xi(\xi/F_{\rm obs}, E)\,{\rm d}\xi \\ \times \int_{-\infty}^\infty \exp(-f_y^2/\varepsilon\sigma_D'^2)\,{\rm d}f_y \\ \times \int_{-\infty}^\infty \exp[-(\Delta F_{\rm obs} - f_x)^2/2(\sigma_{{\rm obs},\,F}^2 + \sigma_{{\rm obs},\,F}'^2)] \\ \times \exp(-f_x^2/\varepsilon\sigma_D'^2)\,{\rm d}f_x \end{aligned} \quad (27)$$

and the integral over f_y in (26) will thus cancel with the normalization constant. The integral over ξ [with the integrand $P_\xi(\xi/F_{\rm obs}, E) \exp(i\xi)$] is the usual integral for the figure of merit, where F is taken as a constant, and we denote it $\mathbf{m}$. [The figure-of-merit integral arises from expression (14) in Blow & Crick (1959) when F is set to a constant.] Thus,

$$\begin{aligned}\mathbf{f}_{{\rm best},\,N} = K''\mathbf{m} \int_{-\infty}^\infty f_x \exp\left[-\frac{(\Delta F_{\rm obs} - f_x)^2}{2(\sigma_{{\rm obs},\,F}^2 + \sigma_{{\rm obs},\,F}'^2)}\right] \\ \times \exp(-f_x^2/\varepsilon\sigma_D'^2)\,{\rm d}f_x, \end{aligned} \quad (28)$$

which can be rearranged to

$$\begin{aligned}\mathbf{f}_{{\rm best},\,N} = K'''\mathbf{m} \int_{-\infty}^\infty f_x \exp\{-\tfrac{1}{2}f_x^2[1/(\sigma_{{\rm obs},\,F}^2 + \sigma_{{\rm obs},\,F}'^2) \\ + 2/\varepsilon\sigma_D'^2] + f_x[\Delta F_{\rm obs}/(\sigma_{{\rm obs},\,F}^2 + \sigma_{{\rm obs},\,F}'^2)]\}\,{\rm d}f_x. \end{aligned} \quad (29)$$

This distribution is a Gaussian with mean and variance

$$\mu = \frac{\Delta F_{\rm obs}/(\sigma^2_{{\rm obs},F} + \sigma'^2_{{\rm obs},F})}{1/(\sigma^2_{{\rm obs},F} + \sigma'^2_{{\rm obs},F}) + 2/(\varepsilon\sigma'^2_D)}$$
$$\sigma^2 = \frac{1}{1/(\sigma^2_{{\rm obs},F} + \sigma'^2_{{\rm obs},F}) + 2/(\varepsilon\sigma'^2_D)}. \quad (30)$$

Evaluating the integral gives

$$\begin{aligned}\mathbf{f}_{{\rm best},N} &= \mathbf{m}\mu \\ &= \mathbf{m}[(\varepsilon\sigma'^2_D/2)/(\sigma^2_{{\rm obs},F} + \sigma'^2_{{\rm obs},F} + \varepsilon\sigma'^2_D/2)] \\ &\quad \times (F'_{\rm obs} - F_{\rm obs}) \\ &= \mathbf{m}w(F'_{\rm obs} - F_{\rm obs}), \end{aligned} \quad (31)$$

which is equivalent to (14). The formula for the centrosymmetric case is obtained by changing σ'^2_D to $2\sigma'^2_D$.

In the case of difference amplitudes, **m** in (14) should be omitted. In this case, the assumption $\sigma_{{\rm obs},F} < F$ is not necessary since the integrals of F and ξ will cancel with the normalization constant even if P_ξ depends on F (*i.e.* the probability distribution of F is kept instead of setting $F = F_{\rm obs}$).

The authors would like to thank Keith Moffat for initiating the project from which this work resulted.

References

Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.

Bourgeois, D., Longhi, S., Wulff, M. & Cambillau, C. (1997). *J. Appl. Cryst.* **30**, 153–163.

Bourgeois, D., Nurizzo, D., Kahn, R. & Cambillau, C. (1997). *J. Appl. Cryst.* In the press.

Bourgeois, D., Ursby, T., Wulff, M., Pradervand, C., LeGrand, A., Schildkamp, W., Labouré, S., Šrajer, V., Teng, T.-Y., Roth, M. & Moffat K. (1996). *J. Synchrotron Rad.* **3**, 65–74.

Campbell, J. W. (1995). *J. Appl. Cryst.* **28**, 228–236.

Cochran, W. (1951). *Acta Cryst.* **4**, 408–411.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* **D50**, 760–763.

Fermi, G., Perutz, M. F., Dickinson, L. C. & Chien, J. C. W. (1982). *J. Mol. Biol.* **155**, 495–505.

French, S. & Wilson, K. (1978). *Acta Cryst.* **A34**, 517–525.

Hajdu, J. & Andersson, I. (1993). *Annu. Rev. Biophys. Biomol. Struct.* **22**, 467–498.

Henderson, R. & Moffat, J. K. (1971). *Acta Cryst.* **B27**, 1414–1420.

Hendrickson, W. A. & Lattman, E. E. (1970). *Acta Cryst.* **B26**, 136–143.

Kuriyan, J., Wilz, S., Karplus, M. & Petsko, G. A. (1986). *J. Mol. Biol.* **192**, 133–154.

Longhi, S., Czjzek, M., Lamzin, V., Nicolas, A. & Cambillau, C. (1997). In preparation.

Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.

Martinez, C., De Geus, P., Lauwereys, M., Matthyssens, G. & Cambillau, C. (1992). *Nature (London)*, **356**, 615–618.

Read, R. (1986). *Acta Cryst.* **A42**, 140–149.

Sim, G. A. (1958). *Acta Cryst.* **11**, 123–124.

Šrajer, V., Teng, T.-Y., Ursby, T., Pradervand, C., Ren, Z., Adachi, S., Schildkamp, W., Bourgeois, D., Wulff, M. & Moffat K. (1996). *Science*, **274**, 1726–1729.

Srinivasan, R. (1966). *Acta Cryst.* **20**, 143–144.

Srinivasan, R. (1968). *Z. Kristallogr.* **126**, 175–181.

Srinivasan, R. & Ramachandran, G. N. (1965). *Acta Cryst.* **19**, 1008–1014.

Stewart, J. M. & Karle, J. (1976). *Acta Cryst.* **A32**, 1005–1007.

Teng, T.-Y., Šrajer, V. & Moffat, K. (1994). *Nature (London) Struct. Biol.* **1**, 701–705.

Terwilliger, T. C. (1994). *Acta Cryst.* **D50**, 11–16.

Terwilliger, T. C. & Berendzen, J. (1996). *Acta Cryst.* **D52**, 743–747.

Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.