*Acta Cryst.* (1977). A33, 512–518

# Minimum-Variance Fourier Coefficients from the Isomorphous Replacement Method by Least-Squares Analysis

By J. Sygusch

*Department of Biochemistry and the Medical Research Council Group on Protein Structure and Function, University of Alberta, Edmonton, Alberta, Canada T6G 2H7*

(*Received* 30 *June* 1976; *accepted* 27 *December* 1976)

A theory for the simultaneous least-squares refinement of protein phases and heavy-atom parameters is presented. Weights are utilized which include experimental error in the heavy-atom structure amplitude as well as in the protein structure amplitude. Direct refinement of the cosine and sine functions of the phase, $\alpha$, automatically furnishes the best phase thereby avoiding the calculation of the best phase *via* the Blow–Crick probability frequency function postulated for phase error. The refinement of protein phases is constrained such that $\cos^2\alpha+\sin^2\alpha=1$ for all protein phases. A minimum-variance Fourier synthesis analogous to the best Fourier synthesis is formulated which has as weights a figure of merit not only based upon phase error as calculated by the new method but also including experimental error in the protein structure amplitude. Approximations show that correlations between protein phases and heavy-atom parameters can be neglected provided the ratio of heavy-atom parameters to protein phases is sufficiently small. Comparison of the new refinement with a conventional refinement utilizing two heavy-atom derivatives measured to 3 Å resolution shows that the new refinement gives better closure and a stronger heavy-atom signal. Comparison of electron density maps shows that the minimum-variance Fourier synthesis yields an electron density map of improved resolution with respect to the conventional best Fourier synthesis.

## Introduction

The number of proteins whose structures have been determined by X-ray methods is rapidly increasing every year. In almost every case the structure determination has been achieved by the isomorphous replacement method. The objective of this method is to obtain a set of protein phases which together with the protein structure amplitudes enables calculation of an electron density map. The protein phases are derived from a mathematical analysis of differences between the native-protein structure amplitudes and the heavy-atom derivative structure amplitudes (Harker, 1956). The heavy-atom derivative is prepared in such a way as to be isomorphous to the native-protein structure (Green, Ingram & Perutz, 1954). For a variety of chemical and physical reasons the intensity differences between the heavy-atom derivative and native-protein structure are generally small. In large protein structures and especially at higher resolution, the magnitude of this difference is generally no greater than several standard deviations. Clearly for an accurately determined structure the protein must be properly refined.

The basis for protein phase refinement originates from the method of treatment of errors first introduced by Blow & Crick (1959) for the isomorphous replacement technique and applied to refinement of protein phases and electron density calculation by Dickerson, Kendrew & Strandberg (1961). Several years later North (1965) and Matthews (1966*a*) included anomalous scattering measurements in the refinement of protein phases and electron density calculation.

The usual method of refinement is iterative and consists of alternate cycles of protein phasing and least-squares refinement of heavy-atom parameters. For the heavy-atom parameters, the heavy-atom structure-amplitude differences between observed and calculated structure amplitudes called the lack of closure are minimized directly whereas for protein phases a distinction is made, the lack of closure is assumed to follow a Gaussian function whose distribution determines the protein phase (Blow & Crick, 1959). The root-mean-square (r.m.s.) lack of closure not only serves to define the standard deviation of the Gaussian distribution (Blow & Crick, 1959) but also can be employed for the calculation of weights in the heavy-atom refinement (Lipscomb *et al.*, 1966; Adams *et al.*, 1969). The r.m.s. lack of closure is recalculated after each cycle of refinement. Although this procedure of weighting in heavy-atom refinement has been successful, it has been shown in $\alpha$-chymotrypsin to lead to slow convergence and to bias towards the starting heavy-atom parameters (Blow & Matthews, 1973). This bias and slow convergence could be reduced upon introduction of a weighting scheme wherein weights were assigned to each structure amplitude according to the corresponding figure of merit. Individual weights in heavy-atom refinement which are based upon counting statistics have been used also by Reeke, Becker & Quiocho (1971) and Ten Eyck & Arnone (1976).

The interpretability of the electron density map depends upon the quality of the protein phases from which the map has been computed. The quality of the protein phase is assessed by the figure of merit (Dickerson, Kendrew & Strandberg, 1961) computed for each protein phase which is then used as a weighting factor in the Fourier synthesis. Dickerson, Weinzierl & Palmer (1968) have pointed out that the figure of merit of a particular phase is strongly dependent upon the dispersion of the probability distribution rather than upon its mean value. The figure of merit in effect is a measure of the precision of a phase, not of its accuracy. The dispersion which is recalculated from the residual errors of the individual reflexions after each cycle of refinement can also potentially suffer from the same sort of bias as do the heavy-atom parameters since the protein phase is conditional upon the heavy-atom parameters. More recently, Ten Eyck & Arnone (1976) have devised a procedure which would improve the correlation between the figure of merit and the accuracy of the data used to phase a particular reflexion. The procedure would remove the experimental errors from the heavy-atom errors in the dispersion term of the probability distribution function. An undesirable consequence of this procedure in some instances could be to make the phasing procedure even more sensitive to bias in the heavy-atom parameters. However, they showed that this procedure gave somewhat improved results with weak data (including anomalous dispersion data) but made little difference to the treatment of isomorphous differences.

In this article we wish to present a formalism for the simultaneous least-squares refinement of both heavy-atom parameters and protein phases. This formalism takes into account the experimental error and can include the implicit correlation existing among protein phases and heavy-atom parameters. It is free of bias since all weights are based only upon experimental error and these weights are used for both the heavy-atom parameters and protein phase refinement. The weights allow for both error in the native structure amplitudes and error in the heavy-atom structure amplitudes. The formalism does not require a probability distribution to determine the protein phase since it directly yields the expected values of the cosine and sine functions of the phase which are required for electron density calculation. A figure of merit is derived for Fourier synthesis which takes into account not only the phase error but also the native-protein structure-amplitude error.

## Theory

### (*a*) *Refinement*

The least-squares sum that is minimized in order that protein phases and heavy-atom parameters be refined is shown in the following equation.

$$S=\sum_{H,i}\omega_{H_i}(|F_{PH}|_i^o-|F_{PH}|_i^c)^2+\omega'_{H_i}(\Delta^o_{PH_i}-\Delta^c_{PH_i})^2 \quad (1)$$

where subscripts $H$ and $i$ denote a particular reflexion and derivative respectively, with subscript $PH$ being used to distinguish between heavy-atom derivative structure amplitude $|F_{PH}|_i$ and native structure amplitude $|F_H|$, $\omega_{H_i}$ and $\omega'_{H_i}$ represent appropriate weighting factors and $|F_{PH}|_i^o$ and $\Delta^o_{PH_i}$ are the experimental observations representing isomorphous replacement and anomalous scattering measurements. Their observational equations are defined below.

$$|F_{PH}|_i^c=q_i[(|F_H|\cos\alpha_H+a_{Hi})^2+(|F_H|\sin\alpha_H+b_{H_i})^2]^{1/2}$$
(Blow & Crick, 1959) (2)

$$\Delta^c_{PH_i}=2q_i(|F_H|\delta_i/|F_{PH}|_i^c|f_H|_i)\,(b_{H_i}\cos\alpha_H-a_{H_i}\sin\alpha_H)$$
(Matthews, 1966*a*), (3)

where $q_i$ represents a scaling factor, $\alpha_H$ the protein phase, $f_{H_i}$ the heavy-atom structure factor having real and imaginary components $a_{H_i}$ and $b_{H_i}$ respectively and $\delta_i=ik|f_H|_i$, representing the ratio of the imaginary scattering component to the real scattering component of the heavy atom $i$. Equation (1) has been generalized to more than one type of anomalous scatterer by Matthews (1966*b*) but this will not affect the theory given below.

In conventional least-squares analysis the weighting factors, $\omega_{H_i}$ and $\omega'_{H_i}$ would simply be related to inverses of the variances of the observables, in this instance $|F_{PH}|_i^o$ and $\Delta^o_{PH_i}$. However, when the observational equations themselves are a function of an observable, in this case $|F_H|$, Deming (1938) has shown that the weights in the least-squares sum $S$ must be modified to take into account the random nature of the observable; then

$$\omega_{H_i}=1\Bigg/\left(\sigma^2_{PH_i}+\left|\frac{\partial|F_{PH}|_i^c}{\partial|F_H|}\right|^2\sigma^2_H\right) \quad (4)$$

and

$$\omega'_{H_i}=1\Bigg/\left(\sigma'^2_{PH_i}+\left|\frac{\partial\Delta^c_{PH_i}}{\partial|F_H|}\right|^2\sigma^2_H\right) \quad (5)$$

where $\sigma^2_{PH_i}$, $\sigma'^2_{PH_i}$ and $\sigma^2_H$ are the experimental variances associated with the observables $|F_{PH}|_i^o$, $\Delta^o_{PH_i}$ and $|F_H|$ respectively.

### (*b*) *Electron density*

A striking similarity exists between equations (2) and (3) and the calculated electron density shown below:

$$\varrho(\mathbf{r})=\frac{2}{V}\sum_H|F_H|\,(\cos\alpha_H\cos\mathbf{H}.\mathbf{r}-\sin\alpha_H\sin\mathbf{H}.\mathbf{r}), \quad (6)$$

where $\varrho(\mathbf{r})$ represents the electron and $\mathbf{r}$ is a position vector. In both sets of equations, the protein phase $\alpha_H$ is present only in terms of the trigonometric functions, $\cos\alpha_H$ and $\sin\alpha_H$. There is considerable advantage in the direct determination of the expected values of $\cos\alpha_H$ and $\sin\alpha_H$ by least squares over the more conventional approach where determination of the phase $\alpha_H$ is involved. By determining from least squares the expected values of $\cos\alpha_H$ and $\sin\alpha_H$, the best phase required for Fourier synthesis is automatically computed and no longer must be calculated *via* the Gaussian frequency function postulated by Blow &

Crick. Simultaneously the figure of merit, which previously had been found to be sensitive to the shape of this frequency function, as a result of the least-squares approach no longer depends upon knowledge of the phase error distribution. Furthermore, where before the figure of merit was calculated assuming a constant phase error for an entire class of reflexions, least-squares refinement does not require any such postulate.

It is quite conceivable, especially in the initial stages of protein-phase and heavy-atom refinement, that there is systematic error present in the observables which cannot be accounted for by the observational equations. An outcome of this problem can be that the sum of the squares of the expected values of the cosine and sine of the phase exceeds one. To avoid this problem, a constraint represented by the following equation

$$\cos^2 \alpha_H + \sin^2 \alpha_H = 1 \tag{7}$$

is imposed upon each protein phase during refinement. The consequence of this constraint however is to eliminate any information about the dispersion of the phase errors which otherwise would have been expressed by the figure of merit. Since the best electron density depends critically upon these weights, *i.e.* figures of merit assigned from a knowledge of the phase errors (Dickerson, Kendrew & Strandberg, 1961), this loss of information is crucial and must be compensated for. By assigning a weight $a_H$ to each term in the Fourier series such that the mean square error of the calculated electron density is a minimum, not only will the calculated Fourier synthesis have the proper functional form, but it also will be analogous to the best Fourier synthesis as originally derived by Blow & Crick. Their derivation for the best Fourier synthesis, as previously noted, requires that a probability distribution be postulated for the phase errors, whereas the analogous best or minimum-variance Fourier synthesis discussed above will be shown to be simply related to the errors of the least-squares refinement and of the experimental data.

Mathematically the minimization can be expressed as

$$U = \left\langle \int_v [\varrho_T(\mathbf{r}) - \varrho_c(\mathbf{r})]^2 \mathrm{d}v \right\rangle \text{ minimum,} \tag{8}$$

where $\varrho_T(\mathbf{r})$ is defined in (6) and

$$\varrho_c(\mathbf{r}) = \frac{2}{v} \sum_H a_H |F_H| (\cos\alpha_H \cos \mathbf{H}.\mathbf{r} - \sin \alpha_H \sin \mathbf{H}.\mathbf{r}). \tag{9}$$

Assuming that the phase $\alpha_H$ and structure amplitude $|F_H|$ are independent random variables, minimization of $U$ yields the following functional form for the weight, $a_H$, after some lengthy manipulations:

$$a_H = \frac{1}{[1 + (\sigma_H/|F_H|)^2](1 + \sigma_T^2)}, \tag{10}$$

where

$$\sigma_T^2 = \sigma^2(\cos \alpha_H) + \sigma^2(\sin \alpha_H) \tag{11}$$

$$\simeq \sigma^2(\cos \alpha_H)/\langle \sin \alpha_H \rangle^2 = \sigma^2(\sin \alpha_H)/\langle \cos \alpha_H \rangle^2 \tag{12}$$

with $\sigma^2(\cos \alpha_H)$ and $\sigma^2(\sin \alpha_H)$ being the variances of $\cos \alpha_H$ and $\sin \alpha_H$ respectively. It should be noted that this new figure of merit is more general than the original figure of merit since it also takes into account the experimental errors in protein structure amplitudes. If the majority of the protein structure amplitudes are measured with a precision considerably greater than their standard deviation, the new figure of merit will not be significantly influenced by the experimental measurements.

### (*c*) *Approximations*

The theory, as it has been presented to this point, cannot be readily implemented without some computational approximations. Full-matrix refinement which utilizes the entire normal-equation matrix becomes quickly prohibitive because of the large computer storage capacity necessitated by the normal-equation matrix. The increase in storage requirement is quadratic in the dimension of the normal-equation matrix where the dimension is equal to the sum of the number of refinable protein phases plus heavy-atom parameters. Detailed consideration of the normal-equation matrix however shows that off-diagonal elements which couple heavy-atom parameters and protein phases need not be stored and can be safely neglected provided the ratio of heavy-atom parameters to the number of protein phases is sufficiently small (see Appendix). The full matrix then breaks down into block-diagonal form in the heavy-atom parameters and diagonal form in the protein phases. This matrix is then readily handled on any reasonable-size computing machine. As discussed in the Appendix the diagonal elements of the inverted normal-equation matrix pertaining to the protein phases may be underestimated, especially if the ratio of number of the heavy-atom parameters to the protein phases is not sufficiently small. To compensate for the diminished diagonal terms and consequently smaller variance $\sigma_T^2$, a different refinement procedure can be employed. Only one of the trigonometric functions of the phase is refined and the other function is treated as a dependent variable. Then the first-order Taylor expansion of the functional relationship between the dependent and independent variables will tend to overestimate the variance of the dependent variable (Papoulis, 1965). Although this sort of compensation is not rigorous, it should be noted that where the phase has a high standard deviation, *i.e.* small diagonal element and thereby greatest sensitivity to the ratio of heavy-atom parameter to protein phases, the formula for the variance $\sigma_T^2$ (equation 12) will overestimate by the largest amount. Whereas for small values of $\sigma_T^2$, *i.e.* large diagonal elements and least sensitivity to this ratio, the formula becomes exact.

## Results

The new least-squares refinement was tested on data measured to 3 Å resolution for two heavy-atom derivatives of glycogen phosphorylase *a*. To assess the performance of the new refinement objectively, it was compared to a conventional refinement procedure set forth in a paper by Adams *et al.* (1969). A computer program initially written by M. Rossmann and others and based upon the above paper was rewritten to take into account the details of the new least-squares refinement. The unmodified program was employed as representative of the conventional refinement. This program has been used to solve numerous protein structures, lactate dehydrogenase (Adams *et al.*, 1969), yeast hexokinase (Fletterick, Bates, Steitz, 1975), serine protease SGPB (Delbaere, Hutcheon, James & Thiessen, 1975) and others. Since the new refinement requires weights based upon the errors in the measured data, every effort was made to allow for all possible sources of error to contribute to the weights. It should be noted that after data reduction, the weights were employed unaltered throughout subsequent calculations. Briefly, the sources of error which were considered and taken into account were counting statistics, short-term fluctuations (Sygusch, 1976), absorption errors, crystal decomposition errors, as well as intercrystal scaling errors.

Some general features have emerged from a comparison of the behavior of the two types of refinements. Convergence by the new least-squares approach was found to be more rapid; the new refinement converged after 5 cycles of refinement whereas the conventional refinement was stopped after 8 cycles when various statistical indicators discussed below exhibited stability.

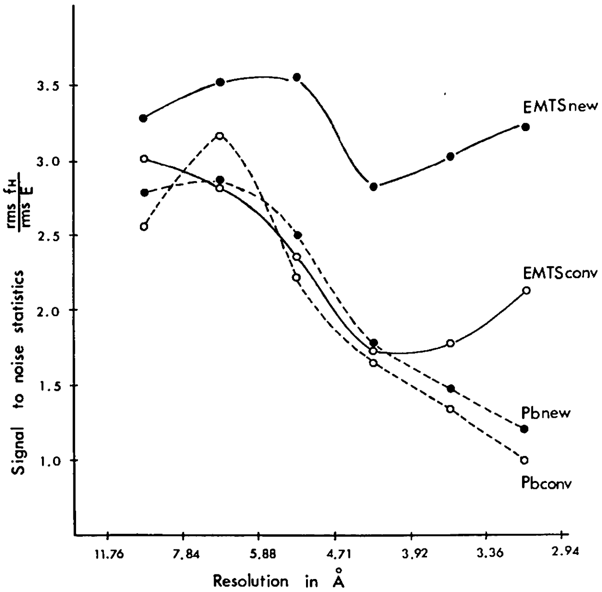


Fig. 1. Signal-to-noise ratios of two heavy-atom derivatives of glycogen phosphorylase *a* plotted as a function of resolution. NEW denotes results based upon the new refinement and CONV refers to results obtained from a conventional refinement. See text for heavy-atom derivative abbreviations.

Since the cosine and/or sine of the phase now is a refinable parameter, the evaluation of the phase at regular intervals* around the phasing circle can be eliminated, thereby resulting in a potentially substantial saving of computational time. Reflexions whose minima were judged to be too shallow and thus not suitable for refinement (*i.e.* low figure of merit) were however examined at regular intervals about the phasing circle at each cycle of refinement to determine whether the minimum had shifted. And lastly, the new refinement does not require any special precautions to be taken for variables which are highly correlated among themselves; all variables can be refined simultaneously, including occupancy and thermal motion.

The depth of the least-squares minima obtained for the two refinements as expressed by the statistical quantity, the goodness of fit, are very different, 9·92 for the new refinement and 0·41 for the conventional refinement. The goodness of fit is defined by the quantity $S/n-p$ where $S$ is the value of the least-squares sum $S$ of equation (1), $n$ represents the total number of unique heavy-atom measurements and $p$ the dimension of the normal-equation matrix. This quantity can be shown from statistical considerations for the proper set of weights to be equal to one (Hamilton, 1964). Whenever this quantity exceeds one, systematic error may be present in the data. In the case of conventional refinement where the value for the goodness of fit is substantially less than one, 0·41, either there exists a scaling factor error in the weights employed (Hamilton, 1964) or there is significant systematic error in the weights themselves. Since it is not possible to distinguish between either of these two possibilities, it is believed that a more meaningful comparison can be obtained between the two refinements if only unweighted statistical quantities are considered.

In Fig. 1 the signal-to-noise ratios are presented for the last cycle of each refinement for the two derivatives, lead nitrate (PB) and ethyl mercury thiosalycilate (EMTS). It is evident that in both cases the new least-squares refinement gives a much better signal-to-noise ratio than does the conventional refinement, especially for EMTS in the high-angle region. A breakdown of the signal and noise into various ratios comparing the new refinement with respect to the conventional refinement is presented in Fig. 2. In each case the new refinement results are superior to the conventional refinement. The r.m.s. $|f_H|$ ratio is always greater than one and the r.m.s. closure ($E$) ratio is virtually always less than one for both derivatives. The signal-to-noise ratio has improved in the new refinement with respect to the conventional refinement but not at the expense of

* The interval chosen for the evaluation of the phase about the phasing circle was 18°. This number was chosen as a compromise between reasonable computing time and cost. There are some 20 000 unique reflexions to 3 Å resolution in glycogen phosphorylase *a*.

simultaneously increasing or decreasing both r.m.s. $|f_H|$ and r.m.s. closure. In Table 1 are summarized some additional statistical quantities and final parameters for the final cycle of each refinement. $R_{\text{MOD}}$, the overall signal-to-noise ratio summarizes the results of Fig. 1; the new refinement is substantially better. This is further supported by the fact that $R_{\text{MOD, CEN}}$ is indeed lower for the new refinement than for the conventional refinement. The anomalous-dispersion parameters, $\delta_{\text{Pb}}$ and $\delta_{\text{EMTS}}$, which are both refinable, are both larger in the new refinement than in the conventional refinement, the ideal value being for both 0·12 approximately. This again points out that the heavy-atom signal is much better resolved in the new refinement.

Table 1. *Refinement summary*

| | Refinement | |
|---|---|---|
| Statistics | New | Conventional |
| $R_{\text{MOD}}$* | 0·304 | 0·406 |
| $R_{\text{MOD, CEN}}$† | 0·659 | 0·719 |
| $\delta_{\text{Pb}}$‡ | 0·064 | 0·051 |
| $\delta_{\text{EMTS}}$‡ | 0·063 | 0·040 |
| $\overline{m}$ | 0·58 | 0·67 |

$$* \quad R_{\text{MOD}} = \frac{\sum_{H,i} \left|(|F_{PH}|_i^o - |F_{PH}|_i^c)\right|}{\sum_{H,i} |f_H|_i}.$$

† Summation $H$ is over centric zones only.

‡ Defined *via* equation (3).

A notable exception in Table 1 is the average figure of merit $\overline{m}$ which is substantially smaller for the new refinement ($\overline{m}_n = 0·58$) than it is for the conventional refinement ($\overline{m}_{c,f} = 0·67$). The figures of merit as a function of resolution for the two refinements are shown in Fig. 3. It is evident from Fig. 3 that the figures of merit from the two respective refinements cannot be solely related through some suitable scaling factor owing to the more rapid decrease with increasing resolution of the new figure of merit when compared to the conventional figure of merit. It is noteworthy that if the figure of merit ($\overline{m}_{c,n}$) is calculated *via* the conventional refinement from parameters and phases obtained from the new refinement, this figure of merit is substantially higher on the initial cycle but after cycles of conventional refinement tends to the figure of merit obtained from the final cycle of the true conventional refinement.

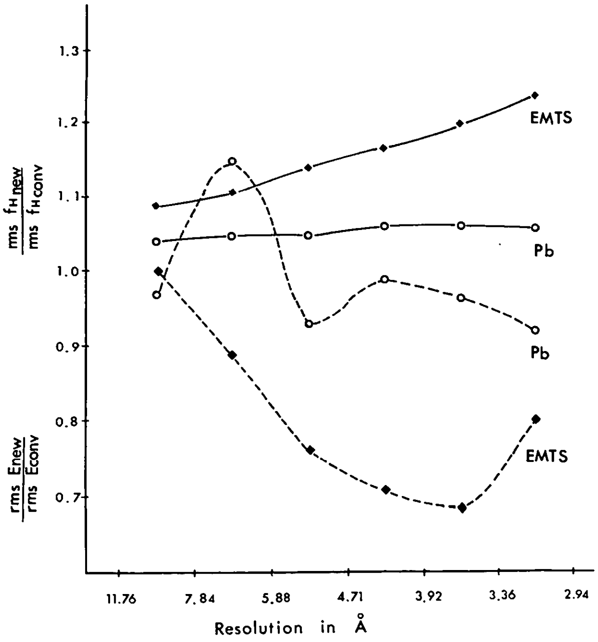


Fig. 2. Comparison as a function of resolution of identical quantities obtained from the new refinement and a conventional refinement. The continuous line represents the ratio of heavy-atom signals obtained from the new refinement with respect to the conventional refinement for two heavy-atom derivatives of glycogen phosphorylase *a*. Similarly, the broken line represents ratios of lack of closure for the two refinements and heavy-atom derivatives. See text for heavy-atom derivative abbreviations.

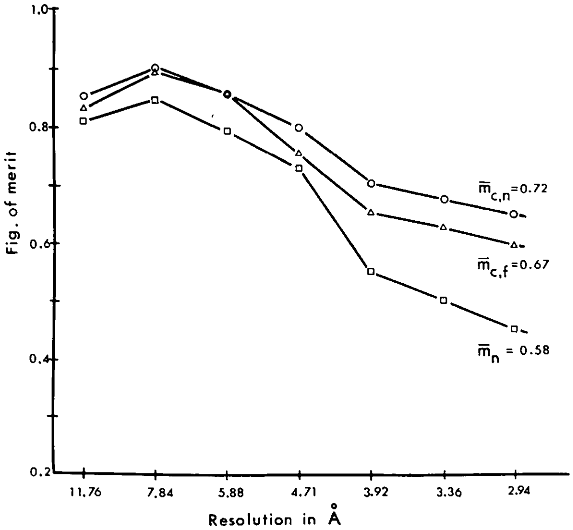
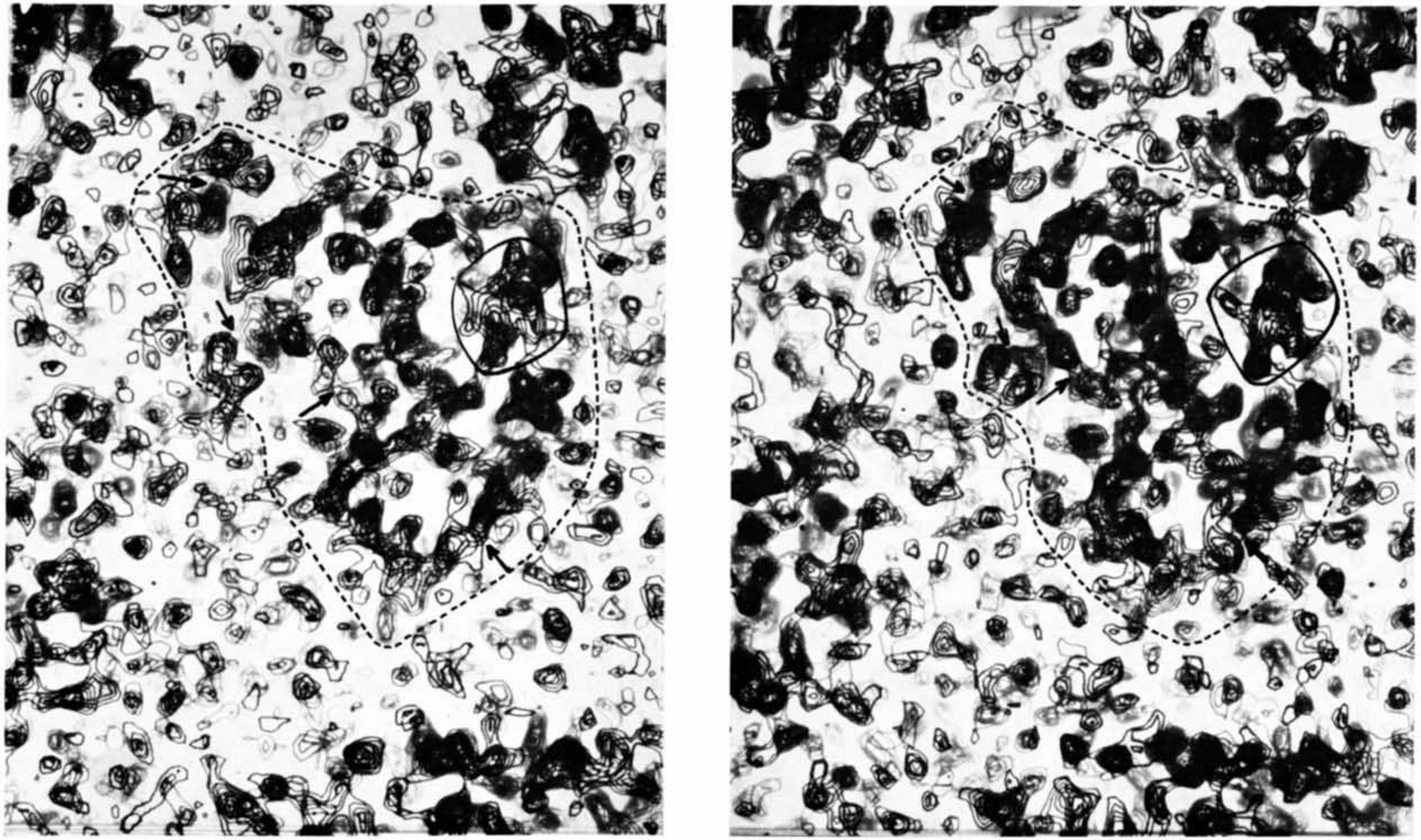


Fig. 3. Comparison as a function of resolution of various figures of merit obtained from the new refinement and a conventional refinement. The curves designated by average figures of merit $\overline{m}_n$, $\overline{m}_{c,f}$ and $\overline{m}_{c,n}$ represent calculations based upon the new refinement, the conventional refinement and the initial cycle of a conventional refinement using parameters and phases from the final cycle of the new refinement.

In Fig. 4 there is shown a 10 Å thick electron density section of glycogen phosphorylase *a* computed with phases and figure of merits obtained respectively from the two refinements. The contour levels have been chosen so that average value of the product of the figure of merit and the protein structure amplitude is identical for both maps. The new map (*b*) appears to be significantly sharper in resolution of the protein backbone than is the conventional map (*a*) even when the average figure of merit has been computed to be smaller in the new map (*e.g.* compare circled cross section of an α-helix on both maps). With the exception of a few cases (approximately 5%), whenever the protein backbone chain was traced through low density on the new electron density map the protein backbone chain went through even weaker or even non-existent electron density on the conventional map. It should be noted that if the lowest contour level were eliminated from both maps in Fig. 4 the new map would be considerably



(*a*) (*b*)

Fig. 4. Electron density maps of a 10 Å thick section through glycogen phosphorylase *a* from two sets of phases obtained from two different heavy-atom derivatives measured to 3 Å resolution. The map (*a*) on the left represents a best Fourier synthesis that has been computed from phases obtained from a conventional refinement. The map (*b*) on the right corresponds to a minimum-variance Fourier synthesis calculated from phases obtained from the new refinement. The broken line represents an envelope enclosing the main protein body. The arrows indicate electron density features which are weak or broken in one map when compared to the equivalent feature in the other map. The circled density represents a cross section through an α-helix.

improved in presentation and still be completely interpretable whereas the conventional map would definitely deteriorate in interpretability.

## Summary

A comprehensive theory of least-squares refinement of heavy-atom parameters and protein phases has been presented. The theory does not make any distinction between the manner of refining heavy-atom parameters and protein phases. The least-squares refinement allows input into the weighting factors for the observational equations representing isomorphous replacement and anomalous scattering, experimental error in the heavy-atom derivative structure amplitudes as well as in the protein structure amplitudes. By refining the cosine and sine values of the phase, no probability functions are required for the evaluation of the best phase. Although least-squares refinement allows independent refinement of the cosine and sine functions of the phase, experience has shown that refinement should be conducted such that the constraint of $\cos^2 \alpha_H + \sin^2 \alpha_H = 1$, where $\alpha_H$ is the protein phase, is satisfied in all reflexions.

This phase constraint effectively sets the figure of merit equal to one for all reflexions, thereby necessitating the derivation of a new minimum-variance Fourier synthesis analogous to the best Fourier synthesis of Blow & Crick. The new refinement obtained is not only a function of the phase errors determined by the least-squares refinement but also a function of the measurement errors in the protein structure amplitudes.

The refinement of the protein phases and heavy-atom parameters can be simplified computationally provided the ratio of the total number of refinable heavy-atom parameters to protein phases is sufficiently small. Difficulties in refinement can occur if this ratio is not small enough.

Comparison of the new refinement and of a conventional refinement for a two-derivative case shows that the new refinement gives better closure and improved heavy-atom signals. The minimum-variance Fourier synthesis computed from phases obtained from the new refinement was of higher resolution than the best one computed from the conventional refinement.

The author wishes to thank Dr R. J. Fletterick for his support and many fruitful discussions, Dr N. B. Madsen for his enthusiastic interest and Dr M. N. G. James for initially pointing out this problem area. The author also wishes to thank the Medical Research Council of Canada for a Post-Doctoral Fellowship and the Protein Structure and Function Group at the University of Alberta for generous computing funds.

## APPENDIX

In least-squares refinement the parameter shifts and errors are dependent upon the inverted normal-equation matrix. Inversion of the normal-equation matrix in unreduced form is not possible for computational reasons. In the following a reduction of the normal-equation matrix will be presented which is adequate for most protein refinements.

The unreduced normal-equation matrix is of the following form:

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}. \qquad (1)$$

The matrix $A$ is a square-matrix block-diagonal form constructed from the heavy-atom parameter derivatives. A typical element $a_{ij}$ of $A$* is

$$a_{ij} = \sum_{H,H'} \frac{\partial(|F_{PH}|}{\partial P_i} \frac{\partial|F_{PH'}|}{\partial P_j} \delta_{PH,PH'}, \qquad (2)$$

where $P_i$ represents a heavy-atom parameter and $\delta_{PH,PH'}$ is one if the product pertains to the same derivative and reflexion. It is zero otherwise. The matrix $B$ is a rectangular matrix representing the interaction of the heavy-atom parameters and the protein phases. A typical element $b_{i,H}$ is

$$b_{i,H} = \frac{\partial|F_{PH}|}{\partial P_i} \frac{\partial|F_{PH}|}{\partial \cos \alpha_H}. \qquad (3)$$

The matrix $C$ is diagonal with elements,

$$c_{H,H'} = \sum_{H,H'} \frac{\partial|F_{PH}|}{\partial \cos \alpha_H} \frac{\partial|F_{PH'}|}{\partial \cos \alpha_{H'}} \delta_{PH,PH'}. \qquad (4)$$

The inverted normal equation matrix can be written in the same form as (1).†

$$\begin{pmatrix} D & E \\ E^T & F \end{pmatrix}. \qquad (5)$$

Of primary interest are the weights $a_H$ required for electron-density synthesis which are dependent upon $\sigma_T^2$. The variance $\sigma_T^2$, where $\sigma_T^2 = \sigma^2(\cos \alpha_H) + \sigma^2(\sin \alpha_H)$, is obtained from the diagonal elements of the inverted normal-equation matrix. Consequently the matrix $F$ is of most immediate interest.

Then

$$F = (I - C^{-1}B^TA^{-1}B)^{-1}C^{-1} \qquad (6)$$

where $I$ is the identity matrix. The dimension of the square phase-error matrix $F$ is equal to the number of refinable phases. Thus the inversion of the coefficient matrix $(I - C^{-1}B^TA^{-1}B)^{-1}$ is the limiting step. If however it were possible to show that the term $C^{-1}B^TA^{-1}B$ were small or negligible compared to $I$, the inversion of the coefficient matrix would be considerably simplified. The arguments to be presented below are by no means rigorous but are merely intended to convey order-of-magnitude considerations.

* Anomalous scattering will not be considered since for most instances the anomalous signal is one order of magnitude less than the isomorphous signal and thus will not affect the preceding analysis, as shown below.

† All equations referred to in this Appendix are equations in the Appendix and not in the main paper.

Consider first a term of $B^TA^{-1}B$:

$$(B^TA^{-1}B)_{H,H'}=\sum_{j,k}\frac{\partial|F_{PH}|}{\partial\cos\alpha_H}\,\frac{\partial|F_{PH}|}{\partial P_i}\times a_{jk}^{-1}\,\frac{\partial|F_{PH}|}{\partial P_k}\,\frac{\partial|F_{PH'}|}{\partial\cos\alpha_{H'}}\qquad(7)$$

$$=\frac{\partial|F_{PH}|}{\partial\cos\alpha_H}\left(\sum_{j,k}\frac{\partial|F_{PH}|}{\partial P_j}\times\sigma_jr_{jk}\sigma_k\frac{\partial|F_{PH'}|}{\partial P_k}\right)\frac{\partial|F_{PH'}|}{\partial\cos\alpha_{H'}},\qquad(8)$$

where $r_{jk}$ represents an element of the heavy-atom correlation matrix and $\sigma_j$ is defined by

$$a_{jk}^{-1}=\sigma_jr_{jk}\sigma_k.\qquad(9)$$

Provided the correlations among heavy-atom parameters are not unduly severe

$$\sigma_j\simeq\frac{1}{[\sum_H(\partial|F_{PH}|/\partial P_j)^2]^{1/2}}.\qquad(10)$$

The derivatives of a given heavy-atom parameter generally do not vary among themselves by more than one order of magnitude. Then

$$\frac{\partial|F_{PH}|}{\partial P_j}\sigma_j\simeq n_H^{-1/2},\qquad(11)$$

where $n_H$ represents the number of observations per heavy-atom compound. Equation (7) then reduces to

$$(B^TA^{-1}B)_{H,H'}\simeq\frac{1}{n_H}\,\frac{\partial|F_{PH}|}{\partial\cos\alpha_H}\Big(\sum_{j,k}r_{jk}\Big)\frac{\partial|F_{PH'}|}{\partial\cos\alpha_{H'}}.\qquad(12)$$

In general,

$$\sum r_{jk}\simeq\sum_{PH}kn_{PH},\qquad(13)$$

where $n_{PH}$ is the number of refinable parameters per heavy-atom compound and $k$ is a constant known from experience to vary between two and four. If the entire second portion of the coefficient matrix is considered:

$$(C^{-1}B^TA^{-1}B)_{H,H'}\simeq\frac{1}{\sum_H(\partial|F_{PH}|/\partial\cos\alpha_H)^2}\times\frac{\partial|F_{PH}|}{\partial\cos\alpha_H}\,\frac{\partial|F_{PH'}|}{\partial\cos\alpha_{H'}}\,\frac{\sum_{PH}kn_{PH}}{n_H}\qquad(14)$$

and if the same sort of considerations are applied as for equations (10) and (11)

$$(C^{-1}B^TA^{-1}B)_{H,H'}\simeq k\bar{n}_{PH}/n_H\qquad(15)$$

where $\bar{n}_{PH}$ is the average number of heavy-atom parameters for the heavy-atom compounds.

Thus as the number of observations increases the term represented by equation (15) steadily becomes less important. For most proteins this term differs by at least two orders of magnitude from unity. For large proteins and few heavy-atom sites this term may differ by at least three orders of magnitude from unity. However, the converse of this result is that for small proteins and many sites, this term may not differ by more than one order of magnitude from unity and thus need not be negligible. The binomial expansion of $F$ to first order is

$$F=(I+C^{-1}B^TA^{-1}B)C^{-1}.\qquad(16)$$

The general effect of taking into account heavy-atom and protein-phase coupling is to increase the diagonal elements of the phase-error matrix $F$ with respect to the inverted diagonal matrix $C^{-1}$.

It should be noted that convergence problems can arise especially in cases where refinement based upon low-resolution data is concerned. In these cases the $k\bar{n}_{PH}/n_{PH}$ ratio may not be small enough and if the off-diagonal elements are neglected, the magnitudes of the calculated parameter shifts for both heavy-atom parameters and protein phases may be inaccurate and slow down convergence. In severe cases the signs of the parameter shifts may be wrong and the refinement may even diverge.

## References

Adams, M. J., Haas, D. J., Jeffrey, B. A., McPherson, A. Jr, Mermall, H. L., Rossmann, M. G., Schevitz, R. W. & Wonacott, A. J. (1969). *J. Mol. Biol.* **41**, 159–188.

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

Blow, D. M. & Matthews, B. W. (1973). *Acta Cryst.* A**29**, 56–62.

Delbaere, L. T. J., Hutcheon, W. L. B., James, M. N. G. & Thiessen, W. E. (1975). *Nature, Lond.* **257**, 758–763.

Deming, W. E. (1938). *Statistical Adjustment of Data.* New York: Dover.

Dickerson, R. E., Kendrew, J. C. & Strandberg, B. E. (1961). *Acta Cryst.* **14**, 1188–1195.

Dickerson, R. E., Weinzierl, J. E. & Palmer, R. A. (1968). *Acta Cryst.* B**24**, 997–1003.

Fletterick, R. J., Bates, D. J. & Steitz, T. A. (1975). *Proc. Natl. Acad. Sci. US*, **72**, 38–42.

Fletterick, R. J., Sygusch, J., Semple, H. & Madsen, N. B. (1976). *J. Biol. Chem.* **251**, 6142–6146.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *Proc. Roy. Soc.* A**255**, 287.

Hamilton, W. C. (1964). *Statistics in Physical Sciences.* New York: Ronald Press.

Harker, D. (1956). *Acta Cryst.* **9**, 1–9.

Lipscomb, W. N., Coppola, J. C., Hartsuck, J. A., Ludwig, M. L., Muirhead, H., Searl, J. & Steitz, T. A. (1966). *J. Mol. Biol.* **19**, 423–441.

Matthews, B. W. (1966*a*). *Acta Cryst.* **20**, 82–86.

Matthews, B. W. (1966*b*). *Acta Cryst.* **20**, 230–239.

North, A. C. T. (1965). *Acta Cryst.* **18**, 212–216.

Papoulis, A. (1965). *Probability, Random Variables and Stochastic Processes.* New York: McGraw-Hill.

Reeke, G. N., Becker, J. W. & Quiocho, F. A. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 277–284.

Sygusch, J. (1976). *Acta Cryst.* A**32**, 859–862.

Ten Eyck, L. F. & Arnone, A. (1976). *J. Mol. Biol.* **100**, 3–11.