Acta Cryst. (1990). A**46**, 537-540

A Probability Representation for Phase Information from Multiwavelength Anomalous Dispersion

BY ARNO PÄHLER,* JANET L. SMITH† AND WAYNE A. HENDRICKSON

Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, *USA*

(*Received* 15 *September* 1989; *accepted* 12 *February* 1990)

Abstract

A probability distribution function, cast in the representation of Hendrickson & Lattman [*Acta Cryst.* (1970), B**26**, 136-143], has been derived for the phase information from measurements of multiwavelength anomalous diffraction (MAD). This probability function readily permits one to determine figure-of-merit weights similar to those used in isomorphous replacement, and the coefficients that characterize this distribution function facilitate the combining of MAD phasing with results from other sources of phase information. This probability representation was derived in the course of a structural analysis of selenobiotinyl streptavidin from MAD data and applications have also been made in the structure determinations of interleukin-1α and a drug complex with brominated DNA.

Introduction

The likelihood with which the phase angle of a structure factor $F=|F|\exp(i\varphi)$ attains a value between 0 and 2π can be characterized by a phase probability distribution $P(\varphi)$. Blow & Crick (1959) introduced a description of such phase probability assuming a Gaussian distribution for the errors $\varepsilon(\varphi)$ from lack of closure in phase triangles in the method of isomorphous replacement. Thus

$$P(\varphi)=N\exp[-\varepsilon^2(\varphi)/2E^2] \qquad (1)$$

where E is the expectation value for errors at the true phase angles. Somewhat later, Hendrickson & Lattman (1970) gave an alternative description whereby a Gaussian distribution is ascribed to errors of closure in $|F|^2$ rather than in $|F|$ values. For the method of isomorphous replacement,

$$\varepsilon(\varphi)=|F_P(\varphi)+F_H|^2-|F_{PH}|^2 \qquad (2)$$

where F_P, F_{PH} and F_H stand for structure factors that correspond respectively to the native protein, a heavy-atom complex with the protein, and the heavy atom substructure. More generally, F_H can represent a known incremental source of phase information and F_{PH} can stand for a particular observational quantity. Distributions constituted in this way can then be represented in the functional form

$$P(\varphi)=N\exp(K+A\cos\varphi+B\sin\varphi+C\cos 2\varphi+D\sin 2\varphi). \qquad (3)$$

It is the phase probability coefficients A, B, C and D which uniquely characterize the distribution. The factor due to K is irrelevant to the form of the phase distribution and it can be combined into the normalization factor N to obtain unit integrated probability.

A conspicuous advantage of using the functional form described by (3), apart from having a convenient representation of $P(\varphi)$ by the four values A, B, C and D, is the possibility of combining phase information from various sources. If the distributions are mutually independent, by statistical arguments the distribution for combined information is just the product of the individual distributions. This combined distribution, by virtue of its exponential form, is then characterized by accumulated coefficients of A_{total} equal to the sum of the individual A values, and similarly for the other coefficients.

Multiwavelength anomalous dispersion

In recent years a new method for phase determination, based primarily on synchrotron radiation, has been developed for application in macromolecular crystallography. This technique, known as multiwavelength anomalous-dispersion (MAD) phasing, makes use of the wavelength dependence of the anomalous-dispersion part of the atomic form factor $f=f^0+f'(\lambda)+if''(\lambda)$ and has been cast into an algebraic form by Karle (1980). An algebraic formulation by Hendrickson (1985), which differs in the unknown quantities but is equivalent to Karle's approach (Karle, 1989), has been successfully applied in the structure determination of the protein streptavidin complexed with selenobiotin, an analog of its natural ligand biotin. The selenium atoms in this complex served as the source of anomalous scattering, and data from one crystal, collected at three different

* Present address: Protein Engineering Research Institute, 6-2-3, Furuedai, Suita, Osaka 565, Japan.

† Present address: Department of Biological Sciences, Purdue University, Lafayette, IN 47907, USA.

0108-7673/90/070537-04$03.00 © 1990 International Union of Crystallography

wavelengths, sufficed to obtain an unambiguous chain tracing at 3·3 Å resolution for this protein for which no structural precedent was available (Hendrickson, Pähler, Smith, Satow, Merritt & Phizackerley, 1989). Other applications include cucumber basic blue protein which is structurally related to plastocyanin (Guss, Merritt, Phizackerley, Hedman, Murata, Hodgson & Freeman, 1988), lamprey hemoglobin (Hendrickson, Smith, Phizackerley & Merritt, 1988), and a bacterial ferredoxin (Murthy, Hendrickson, Orme-Johnson, Merritt & Phizackerley, 1988). The underlying idea is not restricted to proteins that naturally contain anomalous scatterers (*e.g.* metalloproteins) or ones into which intrinsic anomalously scattering ligands (*e.g.* selenobiotin or heavy metals) can be introduced. It has been generalized by constructing selenomethionyl proteins (Hendrickson, 1985; Hendrickson, Horton & LeMaster, 1990), in which methionine is replaced by its analog selenomethionine. Crystals of selenomethionyl variants of thioredoxin (Horton & Hendrickson, 1989), interleukin-1α (Hatada, Miller, Graves, Hendrickson & Satow, 1989) and ribonuclease H (Yang, Hendrickson & Crouch, 1989) could all be grown isomorphously with crystals obtained from the natural protein under similar conditions. The technique can also be used in the determination of nucleic acid structures as shown by Ogata, Hendrickson, Gao, Patel & Satow (1989) for a brominated DNA.

The phase equations for the algebraic MAD analysis in the case of a single kind of anomalous scatterer as formulated by Hendrickson (1985) are

$$|^{\lambda}F(\mathbf{h})|^2 = |^0F_T|^2 + a(\lambda)|^0F_A|^2 + b(\lambda)|^0F_T||^0F_A|\cos(^0\varphi_T - ^0\varphi_A) + s(\mathbf{h})c(\lambda)|^0F_T||^0F_A|\sin(^0\varphi_T - ^0\varphi_A) \quad (4)$$

with coefficients a, b and c defined by

$$a(\lambda) = [f'(\lambda)^2 + f''(\lambda)^2]/(f^0)^2 \quad (5a)$$

$$b(\lambda) = 2[f'(\lambda)/f^0] \quad (5b)$$

$$c(\lambda) = 2[f''(\lambda)/f^0]. \quad (5c)$$

Here $^0F_T = |^0F_T|\exp(i^0\varphi_T)$ and $^0F_A = |^0F_A|\exp(i^0\varphi_A)$ designate the normal scattering contributions to the structure factors from the total crystal structure and from the anomalous-scattering centers alone, respectively. The factor $s(\mathbf{h})$ stands for the sign of the Miller-index vector $\mathbf{h}$ and takes the value +1 or −1 depending on whether a reflection or its Friedel mate is being evaluated.

Equations (4) and (5a) to (5c) constitute the backbone of the method. Diffraction measurements made at different wavelengths and of Friedel mates can be used in a least-squares procedure based on (4) to obtain the wavelength-invariant quantities $|^0F_T|$, $|^0F_A|$ and $\Delta\varphi = {}^0\varphi_T - {}^0\varphi_A$. This is implemented in the program *MADLSQ*. The structure of anomalous-scattering centers must then be solved from the $|^0F_A|$ data, either from Patterson maps or by direct methods. This structure, once refined, can then be used to calculate $^0\varphi_A$ and hence $^0\varphi_T = \Delta\varphi + {}^0\varphi_A$. It has been observed that occasionally unrealistic values for ^0F_A are calculated in this process. A different approach, taken in the structure determination of selenobiotinyl streptavidin (Smith & Pähler, unpublished results) is to use the then known atomic positions of the anomalous scatterers for the calculation of $|^0F_A|$ as well as $^0\varphi_A$. However, this approach involves a scaling problem not inherent in the other approach (Hendrickson, 1985) and needs careful analysis of the data.

Phase distribution coefficients

Our approach in determining phase distribution coefficients for multiwavelength anomalous dispersion assumes that measurements for Friedel mates and measurements made at different wavelengths represent independent pieces of information. Given values for $|^0F_T|$, $|^0F_A|$ and $\Delta\varphi$ from the least-squares MAD evaluation and $^0\varphi_A$ values calculated from the anomalous-scatterer model, an error function in $^0\varphi_T$, analogous to (2), can be constructed from (4). Alternatively, as discussed above, the system of MAD equations can be solved for $|^0F_T|$ and $\Delta\varphi$ with $|^0F_A|$ values fixed at the calculated value, and these parameters can be used in the error function. In either case, an expansion of this $\varepsilon(\varphi)$ function in (1), using trigonometric identities as for the isomorphous replacement problem (Hendrickson & Lattman, 1970), then generates the desired *ABCD* coefficients for (3) along with the less-relevant scale parameter K. If the following abbreviations are introduced:

$$Q = |^{\lambda}F(\mathbf{h})|^2 - |^0F_T|^2 - a(\lambda)|^0F_A|^2 \quad (6a)$$

$$R = b(\lambda)|^0F_T||^0F_A| \quad (6b)$$

$$S = s(\mathbf{h})c(\lambda)|^0F_T||^0F_A|, \quad (6c)$$

then the coefficients for a specific observation j, made at a particular wavelength λ for the Bijvoet mate of reflection $\mathbf{h}$ having the sign $s(\mathbf{h})$, are

$$A_j = Q(R\cos{}^0\varphi_A - S\sin{}^0\varphi_A)/E^2 \quad (7a)$$

$$B_j = Q(R\sin{}^0\varphi_A + S\cos{}^0\varphi_A)/E^2 \quad (7b)$$

$$C_j = [(S^2 - R^2)\cos 2{}^0\varphi_A + 2RS\sin 2{}^0\varphi_A]/4E^2 \quad (7c)$$

$$D_j = [(S^2 - R^2)\sin 2{}^0\varphi_A - 2RS\cos 2{}^0\varphi_A]/4E^2. \quad (7d)$$

Coefficients for the complete MAD phase probability distribution associated with the structure factor ^0F_T for reflection $\mathbf{h}$ [formally $\mathbf{h}' = s(\mathbf{h})\mathbf{h}$ to account for the negative Bijvoet mate] are simply obtained by summation over all observations j. Thus, $A_{\mathrm{MAD}} = \sum A_j$ *etc.* Centroid phases and figure-of-merit weights can then be obtained from these distributions in complete analogy with procedures used in the isomorphous replacement method. This procedure has been

implemented in a new program *MADABCD* based on the *MADLSQ* program.

The additivity of coefficients used in this formulation is based on the assumption of independence of information associated with the individual observations. While the $|^{\lambda}F(\mathbf{h})|^2$ observations themselves are independent, the $|^{0}F_T|$, $\Delta\varphi$ and $^{0}F_A$ parameters derive collectively from all of the observations and this implies an interdependence of information. A similar lack of strict independence also afflicts the combination of phase probabilities from multiple isomorphous derivatives where, although F_H values are specific to the derivative, each piece of information does share a common $|F_P|$ observational component and heavy-atom sites are often in common among derivatives. A possible consequence of such interdependence would be falsely inflated figures of merit.

A problem that remains is how best to evaluate E, the Gaussian expectation value for errors. In the *MADABCD* program for producing phase distribution coefficients we make two passes through the whole process: a first pass with arbitrary initial estimates for E during which statistics are accumulated to determine $E = \langle \varepsilon^2 (\Delta\varphi_{MADLSQ} + {}^{0}\varphi_A) \rangle^{1/2}$, and a second pass in which these values are used. As expected from theoretical considerations (Hendrickson & Lattman, 1970), E estimates from this procedure proved to be a strongly varying function of structure-factor amplitude as well as scattering angle. Accordingly, as for related isomorphous replacement calculations (Hendrickson, Love & Karle, 1973), we overcame this problem by compiling E estimates in categories as a function of both $(\sin\theta)/\lambda$ and $|^{0}F_T|$. This approach models the functional behavior of E as a step function. An alternative procedure, not pursued by us at the time when we developed the program, would be to adopt a formulation similar to one proposed by Blundell & Johnson (1976). Here one would evaluate the Blow & Crick analog, E', as the r.m.s. discrepancy between observed and calculated $|^{\lambda}F(\mathbf{h})|$ amplitudes (rather than $|F|^2$ values) simply as a conventional function of scattering angle and then obtain the required values for (7) from $E^2 = 3E'^4 + 4|^{\lambda}F(\mathbf{h})|^2 E'^2$. Yet another alternative, described previously in connection with the treatment of anomalous scattering in isomorphous replacement experiments (Hendrickson, 1979), would be to evaluate the phasing coefficients for a normalized error model having $\varepsilon'(\varphi) = \varepsilon(\varphi)/|^{0}F_T|$. Both of these alternative procedures will produce E' expectation values that are essentially independent of structure-factor amplitude.

Applications

The probability formalism described here was developed in the context of our structure determination of selenobiotinyl streptavidin using MAD data. One motivation that led us to this development was the observation that quantitative measures for quality of phasing in early versions of *MADLSQ* did not correlate well with phase discrepancies in the case of lamprey hemoglobin (J. L. Smith, 1986, unpublished results; Hendrickson *et al.*, 1988). The current implementation of the least-squares approach produces better measures. Figure-of-merit weights from the *ABCD* formulation do however provide a convenient and conventional measure of phase reliability. More importantly, the phase distribution coefficients that we have derived here proved invaluable for the purpose of combining MAD results with phase information from other sources.

To have a stringent test of the MAD phasing method, we determined the chain fold for streptavidin without resorting to other sources of phase information. The phases for this map at 3·3 Å resolution were produced from the *ABCD* formulation and had a mean figure of merit of 0·78. This phase set was essentially the same as that directly obtained from the *MADLSQ* algebraic fitting procedure (Hendrickson *et al.*, 1989). For the building of a complete model including side chains we made use of the fact that the asymmetric unit of these crystals contains two basically identical subunits. Here the MAD phase probability formulation was used in phase combination with the information from molecular averaging and solvent flattening. This gave a significantly enhanced overall appearance to the map at 3·1 Å resolution and it facilitated model building. It was accompanied by an average shift of 23° in phase angle but only a slight improvement from 57 to 54° in average phase discrepancy compared with a model refined to $R = 0{\cdot}177$ at 2 Å resolution (Hendrickson *et al.*, 1989).

This probability formulation for MAD phasing has also been used recently in other applications. In the case of interleukin-1α (Graves, Hatada, Hendrickson, Miller, Madison & Satow, 1990), MAD phase information based on rather inaccurate data had been obtained from a selenomethionyl variant of the protein and isomorphous replacement information from a mercury derivative was also available. Neither one alone sufficed to resolve the phase ambiguities, but phase combination resulted in an interpretable map. The third application involves a drug complex with a brominated DNA oligomer (Ogata *et al.*, 1989). Here the MAD phasing directly from *MADLSQ* did lead to an interpretation for the nucleic-acid structure. However, model rebuilding during refinement and the interpretation of drug components of the structure were facilitated by the probabilistic combination of phase information from the partially refined model with that from the MAD experiment.

We thank Marcos Hatada and Craig Ogata for sharing the results of their applications with us and

Eric Fanchon for helpful discussions. This work was supported in part by a grant (GM-34102) from the US National Institute of Health.

References

BLOW, D. M. & CRICK, F. H. C. (1959). *Acta Cryst.* **12**, 794-802.
BLUNDELL, T. L. & JOHNSON, L. N. (1976). *Protein Crystallography.* New York: Academic Press.
GRAVES, B. J., HATADA, M. H., HENDRICKSON, W. A., MILLER, J. K., MADISON, V. S. & SATOW, Y. (1990). *Biochemistry*, **29**, 2679-2684.
GUSS, J. M., MERRITT, E. A., PHIZACKERLEY, R. P., HEDMAN, B., MURATA, M., HODGSON, K. O. & FREEMAN, H. C. (1988). *Science* **241**, 806-811.
HATADA, M. H., MILLER, J. K., GRAVES, B. J., HENDRICKSON, W. A. & SATOW, Y. (1989). Am. Crystallogr. Assoc. Abstr. Ser. 2, Vol. 17, 89.
HENDRICKSON, W. A. (1979). *Acta Cryst.* A**35**, 245-247.
HENDRICKSON, W. A. (1985). *Trans. Am. Crystallogr. Assoc.* **21**, 11-21.
HENDRICKSON, W. A., HORTON, J. R. & LEMASTER, D. M. (1990). *EMBO J.* In the press.
HENDRICKSON, W. A. & LATTMAN, E. E. (1970). *Acta Cryst.* B**26**, 136-143.
HENDRICKSON, W. A., LOVE, W. E. & KARLE, J. (1973). *J. Mol. Biol.* **74**, 331-361.
HENDRICKSON, W. A., PÄHLER, A., SMITH, J. L., SATOW, Y., MERRITT, E. A. & PHIZACKERLEY, R. P. (1989). *Proc. Natl Acad. Sci. USA*, **86**, 2190-2194.
HENDRICKSON, W. A., SMITH, J. L., PHIZACKERLEY, R. P. & MERRITT, E. A. (1988). *Proteins*, **4**, 77-88.
HORTON, J. R. & HENDRICKSON, W. A. (1989). Proc. Am. Crystallogr. Assoc. Meet. Abstr. Ser. 2, Vol. 17, p. 122.
KARLE, J. (1980). *Int. J. Quant. Chem. Symp.* **7**, 357-367.
KARLE, J. (1989). *Acta Cryst.* A**45**, 303-307.
MURTHY, H. M. K., HENDRICKSON, W. A., ORME-JOHNSON, W. H., MERRITT, E. A. & PHIZACKERLEY, R. P. (1988). *J. Biol. Chem.* **263**, 18430-18436.
OGATA, C. M., HENDRICKSON, W. A., GAO, X., PATEL, D. J. & SATOW, Y. (1989). Proc. Am. Crystallogr. Assoc. Meet. Abstr. Ser. 2, Vol. 17, p. 53.
YANG, W., HENDRICKSON, W. A. & CROUCH, R. J. (1989). Proc. Am. Crystallogr. Assoc. Meet. Abstr. Ser. 2, Vol. 17, p. 114.

Acta Cryst. (1990). A**46**, 540-544

Direct Low-Resolution Phasing from Electron-Density Histograms in Protein Crystallography

BY V. YU. LUNIN, A. G. URZHUMTSEV AND T. P. SKOVORODA

Research Computing Centre, USSR Academy of Sciences, Pushchino, Moscow Region 142292, *USSR*

(*Received* 12 *February* 1990; *accepted* 6 *March* 1990)

Abstract

An approach to direct phasing of low-resolution reflections is proposed. It is based on the generation of a large number of phase sets and selection of those variants whose electron-density-synthesis histograms are close to a prescribed standard. Classifying them into clusters and averaging them inside every cluster restricts their number to one to three usually, in which a phase set close to the standard is contained. The best variant can be recognized by the properties of its cluster. Test phasing of 29 low-resolution reflections has resulted in a correlation coefficient of 0·94 and a mean phase difference of 40° compared with the true phases.

1. Introduction

In previous years histograms corresponding to finite-resolution electron-density syntheses were shown to be a useful tool in macromolecular structure-factor determination (Lunin, 1986, 1988; Lunin & Skovoroda, 1990; Luzzati, Mariani & Delacroix, 1988; Mariani, Luzzati & Delacroix, 1988) and refinement (Harrison, 1988; Zhang & Main, 1990). Some methods of histogram prediction were suggested for proteins with unknown spatial structure. In this paper we discuss how histograms may be used to phase low-resolution reflections directly.

The idea of the approach is very simple on the face of it. One generates many (*e.g.* random) trial phase sets and separates those which lead to histograms close to the predicted one. It would be reasonable to expect that if the number of generated phase sets is large enough, one necessarily finds a variant close to the true one, which can be identified by a 'good' histogram of the corresponding synthesis. The actual situation is much more complicated, and there may exist several different phase sets leading to histograms close to the prescribed one. Since these histograms can always possess errors, we should consider all such variants as admissible.

Cluster-analysis methods permit a more thorough study of the set of admissible variants. These are classified into subsets grouped about different solutions of the phase problem, which then are averaged to 'extract' some (two or three) possible phase-problem solutions. In our tests one of these extracted variants was found to be sufficiently close to the true

0108-7673/90/070540-05$03.00 © 1990 International Union of Crystallography