

Statistical Modification of Anomalous-Scattering Differences

BY MITCHELL LEWIS

The Gibbs Chemical Laboratory, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA

AND D. C. REES

Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90024, USA

(Received 6 January 1982; accepted 7 January 1983)

Abstract

Statistical methodology originally introduced by French & Wilson [*Acta Cryst.* (1978). A34, 517–525] for the treatment of negative-intensity observations is shown to be useful for the estimation of anomalous-scattering differences. The general principles of this method are described, along with applications to data from a lead derivative of the enzyme carboxypeptidase A.

Introduction

The use of X-ray anomalous-scattering measurements in the determination of macromolecular structures has become a standard technique in protein crystallography (Blundell & Johnson, 1976). Although data-collection strategies have been developed which minimize the influence of systematic errors (Hendrickson & Teeter, 1981), a serious problem with the use of anomalous-scattering data in macromolecular crystallography is that Bijvoet differences are generally of the same magnitude as errors in the intensity measurements. As a result, the utility of anomalous-scattering measurements in phase determination is strongly influenced by the quality of the data. In this paper, we describe a method for processing anomalous-scattering data based upon statistical methodology originally introduced by French & Wilson (1978) for the treatment of negative intensity observations. This procedure serves to minimize the influence of measurement errors in the estimation of Bijvoet differences. Experimental results using data from a lead derivative of the protein carboxypeptidase A are presented, which demonstrate the improved quality of anomalous differences processed by this approach.

Theoretical background

Using the nomenclature of French & Wilson (1978), the conditional joint probability function $P_{\Delta}(\Delta|\delta)$,

which describes the probability that the true value for a Bijvoet difference, Δ , has an experimentally observed value, δ , may be expressed as the product of two probability distributions:

$$P_{\Delta}(\Delta|\delta) \propto P_{\delta}(\delta|\Delta) P_{\Delta}(\Delta). \quad (1)$$

$P_{\Delta}(\Delta)$ is the probability that the anomalous difference Δ occurs in the complete absence of experimental error. An experimental observation δ is related to the true difference, Δ , by a sampling distribution with a density function $P_{\delta}(\delta|\Delta)$. The proportionality in (1) expresses the requirement that $P_{\Delta}(\Delta|\delta)$ is normalized. The best value $E_{\Delta}(\Delta|\delta)$ for the Bijvoet difference may be calculated:

$$E_{\Delta}(\Delta|\delta) = \int_{-\infty}^{\infty} \Delta P_{\Delta}(\Delta|\delta) d\Delta. \quad (2)$$

The limits of integration extend from $-\infty$ to $+\infty$, since, unlike the intensity distributions described in French & Wilson (1978), positive and negative anomalous differences occur with equal probability.

Theoretical expressions for the probability distribution of anomalous differences, $P_{\Delta}(\Delta)$, have been reviewed by Srinivasan & Parthasarathy (1976). For the case of a non-centrosymmetric structure with two anomalous scatterers per unit cell, the appropriate function is

$$P_{\Delta}(\Delta) = \frac{\sqrt{2}}{8c\pi^{3/2}} K_0\left(\frac{\Delta^2}{64c^2}\right) \exp\left(\frac{-\Delta^2}{64c^2}\right). \quad (3)$$

K_0 is the modified Bessel function of the second kind of order zero, while c is given by

$$c = \frac{k'' \sigma_1 \sigma_2}{[1 + (k'' \sigma_1)^2]}, \quad (4)$$

where k'' is the ratio of the imaginary to the total real part of the atomic scattering factor of the anomalous scatterer, and σ_1^2 and σ_2^2 are the fractional contribution to the local mean intensity for the anomalous and non-anomalous scatterers, respectively. It should be emphasized that Δ in these expressions refers to the

differences between the normalized intensities of Bijvoet pairs.

For constant σ_1 and σ_2 , the dependence of $P_\Delta(\Delta)$ on the number of anomalous scatterers is not very significant. With many anomalous scatterers in a non-centrosymmetric arrangement, $P_\Delta(\Delta)$ takes the form:

$$P_\Delta(\Delta) = \frac{1}{4c} \exp\left(-\frac{|\Delta|}{2c}\right). \quad (5)$$

Test calculations described below gave essentially identical results using either (3) or (5) for $P_\Delta(\Delta)$.

With these probability distribution functions, an upper limit, Δ_{\max} , for the acceptable magnitude of a Bijvoet difference may be derived. This limit may be used to reject excessively large differences, which are most likely due to erroneous measurements. The fraction Q of Bijvoet differences whose magnitude is less than Δ_{\max} may be evaluated from the relationship

$$Q = \int_{-\Delta_{\max}}^{+\Delta_{\max}} P_\Delta(\Delta) d\Delta. \quad (6)$$

Using the expression for $P_\Delta(\Delta)$ in (5) to evaluate analytically this integral and solving for Δ_{\max} gives

$$\Delta_{\max} = -2c \ln(1 - Q). \quad (7)$$

In the experimental work to be described, Q was given the value 0.993, which corresponds to $\Delta_{\max} = 10c$.

The distribution of observed intensities about their true value is generally assumed to follow a normal distribution. If δ is the Bijvoet difference between intensities I_+ and I_- for two Friedel mates, each with an associated standard deviation σ_+ and σ_- , then the variance of δ is given by

$$\sigma_\delta^2 = \sigma_+^2 + \sigma_-^2. \quad (8)$$

Consequently, the distribution of δ about the true value Δ may be written

$$P_\delta(\delta|\Delta) = \frac{1}{\sqrt{2\pi}\sigma_\delta} \exp\left[-\frac{1}{2}\left(\frac{\delta - \Delta}{\sigma_\delta}\right)^2\right]. \quad (9)$$

In order to evaluate $P_\delta(\delta|\Delta)$, it is necessary to have a meaningful estimate of the standard deviation of each intensity observation. Standard deviations determined from counting statistics are often observed to be systematically underestimated. An accurate estimate of the variance may be obtained from a normal probability-plot analysis (Abrahams & Keve, 1971), if duplicate observations of the intensity measurements are available. These plots may be used to provide modified estimates, σ_{mod} , for standard deviations based on the experimental values, σ_c , derived from counting statistics. In the present work, the standard deviations were modified according to the expression

$$\sigma_{\text{mod}}^2 = \alpha\sigma_c^2 + \beta I^2. \quad (10)$$

The term in I^2 allows for errors proportional to the net count, such as X-ray source instability (Busing & Levy, 1957). The coefficients α and β are determined by a least-squares procedure, so that the normal probability plot exhibits the expected behavior.

Expressions for both $P_\Delta(\Delta)$ and $P_\delta(\delta|\Delta)$ having been obtained, $E_\Delta(\Delta|\delta)$ may be calculated by numerical integration of (2), using Gauss-Legendre quadratures (Stroud & Secrest, 1966). From $E_\Delta(\Delta|\delta)$, the normalized Bijvoet difference based on intensities, the corresponding difference between the structure-factor amplitudes, $F_\Delta(\Delta|\delta)$, may be calculated:

$$F_+ - F_- = \frac{F_+^2 - F_-^2}{F_+ + F_-} \equiv F_\Delta(\Delta|\delta) = \frac{E_\Delta(\Delta|\delta)\langle I \rangle}{F_+ + F_-}, \quad (11)$$

where $\langle I \rangle$ is the local average intensity (calculated in resolution shells) and F_+ and F_- are the structure-factor amplitudes of the hkl and $\bar{h}\bar{k}\bar{l}$ reflections, respectively.

The probability distribution functions $P_\Delta(\Delta)$ [(5)], $P_\delta(\delta|\Delta)$ [(9)] and $P_\Delta(\Delta|\delta)$ [(1)] are illustrated in Fig. 1 for the case $\delta = 0.10$, $\sigma_\delta = 0.03$ and $c = 0.033$. The best estimate for the anomalous difference, $E_\Delta(\Delta|\delta)$, is calculated from (2) to be 0.086. An important aspect of the modification procedure is the generality of the relationship $|E_\Delta(\Delta|\delta)| \leq |\delta|$. The value of a Bijvoet difference may either be reduced or remain essentially unchanged, but can never increase or have a sign change as a result of this method. As the ratio δ/σ_δ decreases, the ratio $E_\Delta(\Delta|\delta)/\delta$ will also decrease. The effect of this procedure is similar to the omission of small Bijvoet differences from the anomalous Patterson maps calculated by Hendrickson & Teeter (1981). The present method provides a smoother and more objective method, however, for dampening the contribution of these terms.

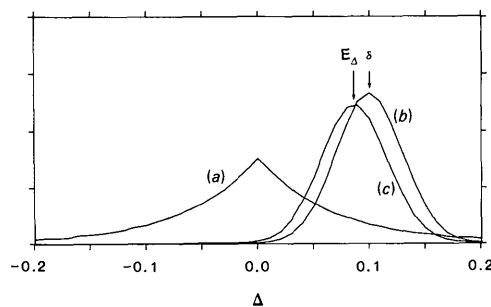


Fig. 1. Variation of the probability distribution functions $P_\Delta(\Delta)$ [curve (a), equation (5)], $P_\delta(\delta|\Delta)$ [curve (b), equation (9)] and $P_\Delta(\Delta|\delta)$ [curve (c), equation (1)] with the true anomalous difference Δ , for the case $\delta = 0.10$, $\sigma_\delta = 0.03$ and $c = 0.033$. Positions for the experimental difference, δ , and the best estimate for the difference, $E_\Delta(\Delta|\delta)$, are indicated. The relative ordinates of the probability functions are not drawn to scale.

Results and discussion

The data-processing procedure was tested using anomalous-scattering data collected from the lead derivative of carboxypeptidase A. Preparation of this derivative has been described (Lipscomb *et al.*, 1966). Crystals of carboxypeptidase A grow in space group $P2_1$ ($a = 51.60$, $b = 60.27$, $c = 47.25$ Å, $\beta = 97.27^\circ$), with one protein molecule (molecular weight of 34 000) in the asymmetric unit. Data were collected from a single crystal to 3.5 Å resolution on a Nicolet $P2_1$ diffractometer equipped with a Cu X-ray source and a graphite monochromator. Intensities were estimated using the Wyckoff step scan (Wyckoff *et al.*, 1967) to sum individual steps located between the half-heights of a peak. The data were corrected for absorption (North, Phillips & Mathews, 1968) and Lorentz-polarization effects. The overall merging $R = \sum |I_i - I_j| / \sum (I_i + I_j)$, where I_i and I_j are symmetry-related reflections (including Friedel mates), was 0.022 for centric reflections and 0.045 for acentric reflections. Standard deviations based on counting statistics were corrected using the normal probability-plot analysis described in the preceding section.

There are two lead sites per asymmetric unit in the lead derivative of carboxypeptidase A, with coordinates (0.094, 0.0, 0.089) and (0.089, 0.040, 0.147), and occupancies of approximately 58 and 53 electrons, respectively (Quiocho & Lipscomb, 1971). Consequently, $\sigma_1^2 = 0.07$ and $\sigma_2^2 = 0.93$. With $k'' = 0.13$ (lead atom, Cu radiation), (4) yields $c = 0.033$. Modified Bijvoet differences were calculated for 3354 Friedel mates from the experimental measurements using (2). The average values for the experimental and modified differences (based on normalized intensities) were 0.082 and 0.053, respectively, with a standard deviation of 0.065. An additional 66 reflections (2% of the acentric data), for which the experimental Bijvoet differences exceeded a Δ_{\max} of 0.33, were excluded from all calculations. This rejection criterion is approximately equivalent to a maximum allowable Bijvoet difference of $5.5\sigma_8$, similar to the $5\sigma_8$ cutoff used in the crambin structural studies (Hendrickson & Teeter, 1981).

Anomalous-difference Patterson maps were calculated with coefficients $(F_+ - F_-)^2$ (Rossmann, 1961), using both experimental and modified anomalous differences. These maps will be subsequently referred to as the 'unmodified' and 'modified' maps, respectively. The $v = \frac{1}{2}$ Harker section of these maps are illustrated in Fig. 2. While the self-vector peaks corresponding to the lead vector sites are obvious in both maps, the peaks in the modified map are approximately 22% higher relative to the origin peak than in the unmodified map. The height of the origin peak is proportional to the sum of the anomalous differences squared. An increase in peak height/origin ratio

consequently indicates a reduction in random errors in the data, since a proportionally larger fraction of the total anomalous scattering signal is present in the self-vector peaks. Consideration of the peak/background ratios for each map is also consistent with this conclusion, although the net improvement is less pronounced. While the modification procedure increases the root-mean-square value of the background in the modified map by 17% relative to the unmodified map, this change is still less than the accompanying increase in peak heights. As a result, there is an overall 4% improvement in the peak/background ratio with the modification procedure.

One possible explanation for the peak-height enhancement observed in the modified Patterson map concerns a possible decrease in effective resolution of this map. Since errors in the intensity measurements generally increase with resolution, Bijvoet differences calculated from (2) will become smaller at higher resolution. The net result is that the higher-resolution terms will contribute less to the Patterson map, thus decreasing the apparent resolution of the map and possibly enhancing the peak-to-noise ratio.

The resolution dependence of the anomalous differences was examined by the variation with $\sin^2 \theta / \lambda^2$ of the ratio k (Matthews, 1966):

$$k = \frac{2 \sum |F_p - F_{ph}|}{\sum |F_+ - F_-|}, \quad (12)$$

where F_p and F_{ph} are the native and derivative structure-factor amplitudes, respectively. k is twice the average isomorphous difference divided by the average anomalous difference, and ideally is equivalent to the

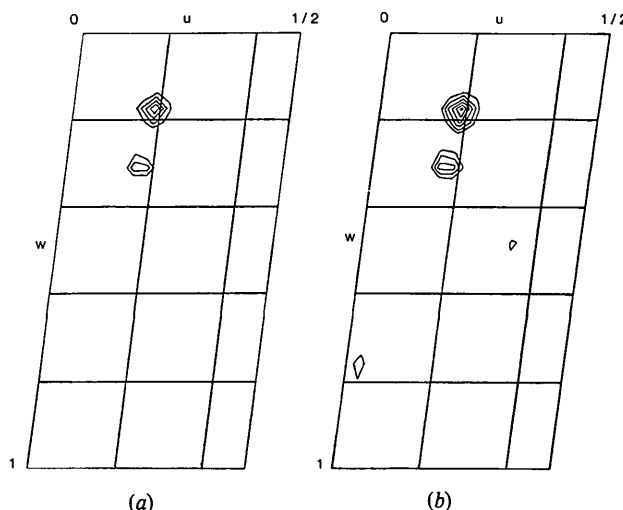


Fig. 2. $v = \frac{1}{2}$ Harker sections of anomalous-difference Patterson maps calculated from (a) experimental and (b) modified anomalous differences. Contour levels are drawn at intervals of 2.5% of the origin peak, omitting the first level above zero.

ratio of the real to imaginary parts of the heavy-atom scattering factor ($k = 1/k''$). In the resolution range used in this work, $k = 7.6$ for lead. The resolution dependence of k ratios calculated from experimental and modified differences are illustrated in Fig. 3. k ratios calculated with the modified anomalous differences are seen to agree much more closely with the theoretical value than k ratios calculated with the experimental differences. Since the same isomorphous differences were used for each calculation, the experimental Bijvoet differences are apparently too large, and the error in these measurements increases with resolution. Such behavior is not uncommonly observed with anomalous-scattering measurements (Blundell & Johnson, 1976). By contrast, the modified differences more closely exhibit the expected resolution dependence. The apparent decrease in the magnitude of the modified anomalous differences with resolution is due not to the modified differences being too small, but to the experimental differences being too large. Consequently, Patterson maps calculated from modified

differences should more accurately reflect the true resolution of the data.

Anomalous-scattering methods have enjoyed increasing popularity for the phase determination of macromolecular structures. The recent determination of the crambin structure (Hendrickson & Teeter, 1981) and the availability of synchrotron-radiation sources for anomalous dispersion and multiple-wavelength measurements (Templeton, Templeton, Phillips & Hodgson, 1980) have contributed to this interest. The statistical procedure for processing anomalous difference measurements presented in this paper should facilitate application of these techniques to macromolecular structure determinations. As with all data-modification procedures, however, these methods can never substitute for careful data-collection techniques.

We would like to thank Professor W. N. Lipscomb for providing the data-collection facilities. ML was the recipient of an American Cancer Society Fellowship.

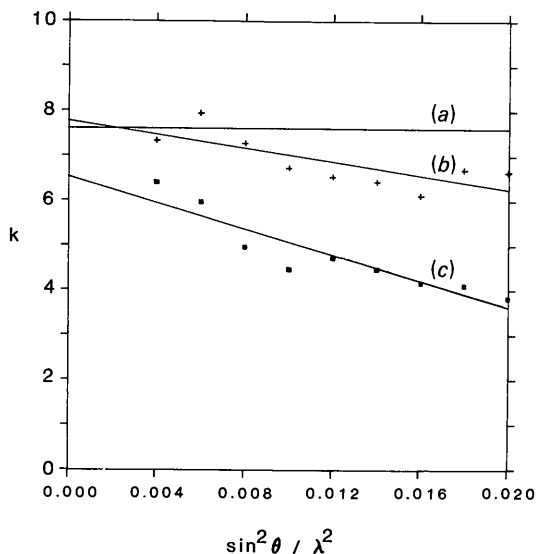


Fig. 3. Variation of the ratio k (equation 12) with $\sin^2 \theta / \lambda^2$. Curve (a) is the theoretical curve; curve (b) is the least-squares line through points calculated from modified anomalous differences (+); curve (c) is the least-squares line through points calculated from experimental anomalous differences (■).

References

- ABRAHAMS, S. C. & KEVE, E. T. (1971). *Acta Cryst.* **A27**, 157–165.
- BLUNDELL, T. L. & JOHNSON, L. N. (1976). *Protein Crystallography*. New York: Academic Press.
- BUSING, W. R. & LEVY, H. A. (1957). *J. Chem. Phys.* **26**, 563–568.
- FRENCH, S. & WILSON, K. (1978). *Acta Cryst.* **A34**, 517–525.
- HENDRICKSON, W. A. & TEETER, M. M. (1981). *Nature (London)*, **290**, 107–113.
- 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). *Acta Cryst.* **20**, 230–239.
- NORTH, A. C. T., PHILLIPS, D. C. & MATHEWS, F. S. (1968). *Acta Cryst.* **A24**, 351–359.
- QUIOCHO, F. A. & LIPSCOMB, W. N. (1971). *Adv. Protein Chem.* **25**, 1–78.
- ROSSMANN, M. G. (1961). *Acta Cryst.* **14**, 383–388.
- SRINIVASAN, R. & PARTHASARATHY, S. (1976). *Some Statistical Applications in X-ray Crystallography*. Oxford: Pergamon.
- STROUD, A. H. & SECREST, D. (1966). *Gaussian Quadrature Formulas*. Englewood Cliffs: Prentice-Hall.
- TEMPLETON, D. H., TEMPLETON, L. K., PHILLIPS, J. C. & HODGSON, K. O. (1980). *Acta Cryst.* **A36**, 436–442.
- WYCKOFF, H. W., DOSCHER, M., TSEBNOGLU, D., INAGAMI, T., JOHNSON, L. N., HARDMAN, K. D., ALLEWELL, N. M., KELLY, D. M. & RICHARDS, F. M. (1967). *J. Mol. Biol.* **27**, 563–578.