

***CHOOCH*: a program for deriving anomalous-scattering factors from X-ray fluorescence spectra**

Gwyndaf Evans and Robert F. Pettifer

Copyright © International Union of Crystallography

Author(s) of this paper may load this reprint on their own web site provided that this cover page is retained. Republication of this article or its storage in electronic databases or the like is not permitted without prior permission in writing from the IUCr.

CHOOCH: a program for deriving anomalous-scattering factors from X-ray fluorescence spectra

Gwyndaf Evans^{a*} and Robert F. Pettifer^b^aMRC Laboratory of Molecular Biology, Hills Road, Cambridge, England, and ^bDepartment of Physics, University of Warwick, Coventry, England. Correspondence e-mail: gwyndaf@mrc-lmb.cam.ac.uk

A Fortran program *CHOOCH*, which derives experimental values of the anomalous-scattering factors f'' and f' from X-ray fluorescence spectra, is described. The program assumes knowledge of theoretical values for the imaginary term, f'' , of the anomalous-scattering factor away from the absorption edge to scale the experimental fluorescence spectrum and thus derive values of f'' near the absorption edge, where tabular data are inappropriate. The Kramers–Kronig relation is used to calculate the real part, f' , of the anomalous-scattering factor. The program aids the decision-making process in macromolecular crystallographic experiments where optimal wavelength selection is required. Magnitudes of f' and f'' at selected wavelengths can later be used as starting values for heavy-atom refinement with crystallographic phasing programs.

1. Introduction

Anomalous scattering (AS) provides an invaluable source of phase information in macromolecular crystallographic structure determination. AS is usually measured from heavy atoms which are incorporated into the crystal lattice to form derivative crystals. This is performed either by soaking in a suitable heavy-atom salt solution (Blundell & Johnson, 1976) or by direct modification of amino acid residues (Doublé, 1997; Hendrickson *et al.*, 1990) or nucleic acid bases (Todd *et al.*, 1999; Golden *et al.*, 1996). Metallo-enzymes possess a natural source of significant AS.

AS is used in the single isomorphous replacement and anomalous scattering (SIRAS) (Baker *et al.*, 1990; North, 1965) method as a means of breaking the phase ambiguity in cases in which only one heavy-atom derivative is available. A special case of SIRAS, where the wavelength is optimized for derivative data acquisition, is known as SIROAS (Helliwell, 1992; Baker *et al.*, 1990) and is a means of maximizing the anomalous signal to noise ratio.

In the multiple-wavelength anomalous diffraction (MAD) method (Hendrickson *et al.*, 1985; Phillips & Hodgson, 1980; Okaya & Pepinsky, 1956), AS is used as the sole source of phase information in the form of multiple measurements at a number of different wavelengths about the absorption edge of a heavy atom incorporated into a macromolecular crystal lattice. Examples of phasing using only derivative data measured at a single wavelength (single-wavelength anomalous diffraction, SAD) also exist (Wallace *et al.*, 1990).

In each of these methods, the observed anomalous signal is typically very small (<5%). Maximizing the anomalous signal is therefore crucial to the experiment. Both f'' and f' have their respective local maxima and minima in the vicinity of characteristic absorption edges. Therefore, data are commonly measured at X-ray wavelengths near these edges. It follows that knowledge of the AS factors about the absorption edge of the heavy atom is required if full advantage is to be taken of the AS signal. This information may be derived from measurements of X-ray fluorescence generated when incident X-ray photons are absorbed by heavy atoms.

Theoretical calculations by Cromer (1983) provided tables of AS factors for all the elements, but it has long been understood that near characteristic absorption edges, theoretical values may differ significantly from those determined experimentally (Hartree *et al.*, 1934). Cromer theory assumes that the atoms are isolated and in vacuum, such that the excited photoelectrons above an absorption edge are ejected into a structureless continuum. In reality, strong deviations from isolated-atom calculations are observed as X-ray absorption near-edge structure (XANES) and extended X-ray absorption fine structure (EXAFS).

At the onset of absorption (the XANES region of the spectrum), where the AS is strongest, the transitions involve valence states. Thus charge, anisotropy, spin and orbital occupancy can all affect AS. EXAFS features are observed above the absorption edge and arise from interference with neighbouring atoms of a photo-electron ejected from the atom.

The consequences are that the magnitude and X-ray energy of important features in an AS spectrum can be strongly dependent on the environment of a heavy atom when it is bound to a protein molecule or nucleic acid. An example of this effect is the strong anisotropy observed at the *K* absorption edge of selenium in selenomethionine (Hendrickson *et al.*, 1989). Selenium in this form also happens to be the most extensively used source of AS for MAD experiments in macromolecular crystallography (Hendrickson, 1999).

In some cases, heavy atoms bind reproducibly to similar environments. A typical example is the binding of mercury to the sulfur atom of the amino acid cysteine. It has been suggested that, in this particular case, values of the L_{III} absorption edge X-ray energy and magnitudes of the AS factors could be transferable to other cysteinyl mercury proteins (Tesmer *et al.*, 1994). However, the transferability of such data relies completely on the energy calibration of the beamline where the original fluorescence data were measured and on the energy calibration of the beamline at which the data are to be used. Beamlines would need to be equipped with absolute-calibration apparatus (Evans & Pettifer, 1996; Pettifer & Hermes, 1985) or some standard calibration techniques using metal foils, for example, would

have to be established for all protein crystallography beamlines capable of multiple-wavelength data acquisition. Such standards are not presently in place.

For these reasons, it is important that fluorescence measurements be taken from the same or a similar heavy-atom-labelled macromolecular crystal sample and on the same beamline that is to be used for diffraction measurements. Only then can deductions about correct wavelengths for the measurement of diffraction data be made.

In macromolecular crystallographic phase determination by MAD or SIROAS, the values of the AS factors f'' and f' near an absorption edge are typically included as refined parameters (de La Fortelle & Bricogne, 1997). Sensible starting values for these parameters can often result in quicker convergence of the refinement procedure. In cases where the observed anomalous signal is very small, reasonably accurate starting values are essential for allowing the refinement to proceed normally.

CHOOCH was written in order to allow fast and easy transformation of raw fluorescence data into AS factors while at the beamline, prior to performing a SIROAS, MAD or SAD experiment.

2. Theoretical background

The effects of AS are described mathematically by two correction terms which are applied to the normal atomic form factor or Thompson scattering factor f_0 . The modified scattering factor is described by $f = f_0 + f' + if''$, where f' is the real part and f'' the imaginary part of the AS correction term.

These AS factors vary most rapidly near characteristic absorption edges of atoms when the energy of the incident X-rays is similar to the binding energy of the absorbing electrons.

The optical theorem (James, 1969) relates the imaginary term f'' directly to the atomic absorption coefficient for an atom by

$$f'' = m_e c \epsilon_0 E \mu_a / e \hbar, \quad (1)$$

where μ_a is the atomic absorption coefficient, E the X-ray energy, and the other symbols have their usual meanings. As in other resonance phenomena, the real part of the dispersive term is related to the imaginary part by a Kramers–Kronig (K–K) transformation. In the case of X-ray scattering, the K–K transform takes the following form:

$$f'(E_0) = (2/\pi) \int_0^\infty [E f''(E)/(E_0^2 - E^2)] dE. \quad (2)$$

Hoyt *et al.* (1984) showed that f' in equation (2) could be obtained numerically by use of a Taylor expansion, replacing the singularity as follows:

$$\begin{aligned} f'(E_0) = & (2/\pi) \int_0^a [E f''(E)/(E_0^2 - E^2)] dE \\ & + (2/\pi) \int_b^\infty [E f''(E)/(E_0^2 - E^2)] dE \\ & + (1/\pi) \left\{ \int_a^b [f''(E)/(E_0 - E)] dE \right. \\ & - (\ln |b - E_0| - \ln |a - E_0|) - \left. \frac{df''}{dE} \right|_{E_0} (b - a) \\ & - \left. \sum_{n=2}^\infty \frac{1}{(n)!} \frac{d^n f''}{dE^n} (b - E_0)^n \right|_{E_0} - (a - E_0)^n \left. \right\}. \end{aligned} \quad (3)$$

Cromer & Liberman (1970) observed that an additional correction term equal to $5E_{\text{TOT}}/3mc^2$, where E_{TOT} is the total energy of the atom, needs to be added to the result of equation (3).

Thus, once a spectrum of f'' has been obtained, it is straightforward to calculate f' .

2.1. Normalization of fluorescence data

The raw fluorescence signal $\mathcal{R}(E)$ measured from a protein crystal sample is on an arbitrary scale. It is dependent on the incident X-ray energy, the absorption properties of the heavy atom being probed, the X-ray flux incident on the sample, the size and shape of the sample, the concentration of the heavy atom in the sample and, of course, the geometry and type of detection system being used. In addition there may be significant contamination of the signal from elastic scatter of the incident X-ray beam, as well as from other inelastic effects. The measured spectra must therefore be carefully analysed in order to remove unwanted background and then scaled to produce an 'experimental' spectrum of absorption and, thence, f'' .

The first stage involves a background subtraction and a subsequent renormalization step, resulting in a spectrum which is on average zero below the edge and one above the edge. The aim of the normalization step is to remove slowly varying (low-order) signal from the data, such that what remains is purely a result of the XANES and EXAFS character of the absorption edge. In this way, contamination by unwanted elastic and inelastic scatter can be removed. During this step, one necessarily removes the slowly varying component of the fluorescence created by the slow variation of the atomic cross section of the particular atom under investigation. Later this atomic component can be reincorporated using theoretical values. The normalization proceeds as follows.

The form of the observed fluorescence below the edge is fitted using a low-order polynomial $\mathcal{P}_{\text{below}}$ of a degree dependent on the experimental conditions. This fit is extrapolated over the whole spectrum and subtracted from the observed data. Similarly, above the edge, a low-order polynomial $\mathcal{P}_{\text{above}}$ is fitted to the above-edge spectrum and then extrapolated over the whole spectrum.

A normalized fluorescence spectrum $\mathcal{N}(E)$, which is equal to zero well below the edge and unity well above the edge, is thus obtained by

$$\mathcal{N}(E) = [\mathcal{R}(E) - \mathcal{P}_{\text{below}}(E)] / [\mathcal{P}_{\text{above}}(E) - \mathcal{P}_{\text{below}}(E)]. \quad (4)$$

The spectrum at this stage should contain only the XANES and EXAFS character of the absorption edge being probed, normalized such that the edge jump is unity.

2.2. Determination of $f''(E)$

The spectrum calculated using equation (4) is then used to create an experimental spectrum of f'' assuming that theoretical values of f'' (Cromer, 1983) are valid far away from the absorption edge where the effects of XANES and EXAFS are negligible. Within a few hundred electron-volts of an absorption edge, theoretical $f''(E)$ values can be well described by two linear functions $f''_{\text{below}}(E)$ and $f''_{\text{above}}(E)$, describing the variation of $f''(E)$ below and above the absorption edge. These functions can be used to obtain an experimental spectrum f''_{exp} using

$$f''_{\text{exp}} = \mathcal{N}(E)[f''_{\text{above}}(E) - f''_{\text{below}}(E)] + f''_{\text{below}}(E). \quad (5)$$

2.3. Determination of $f'(E)$

Calculation of $f'(E)$ using equation (3) requires integration with respect to E to be carried out between 0 and ∞ . Hoyt *et al.* (1984) determined practical integration limits for use in calculating K -edge $f'(E)$ values. They suggest an upper limit of $50 \times E_K$ and a lower limit of E_{L_1} . Measured fluorescence data typically extend to only a few

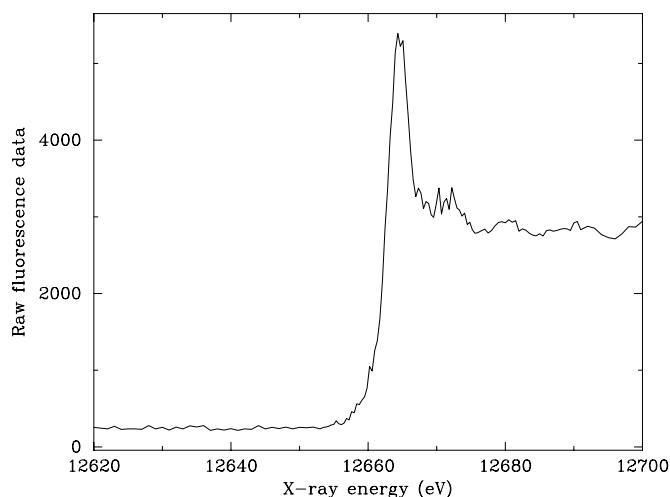


Figure 1
Raw fluorescence spectrum from a seleno-methionine substituted protein measured at the Se *K* edge and displayed using *CHOOCH*.

hundred electron-volts below and above the absorption edge. Theoretical values of $f''(E)$ must therefore be employed to extrapolate to the required limits. Extrapolation is carried out using absorption data published by McMasters *et al.* (1969).

3. Program description

CHOOCH has been written in Fortran 77 and uses calls to the *PGPLOT* library of graphical subroutines to provide a visual interface for manipulation of the experimental fluorescence data (*PGPLOT* is freely available for non-commercial use from <http://astro.caltech.edu/%7Eetjp/pgplot/>). The program also makes use of a number of freely available subroutines for spline (Woltring, 1986) and polynomial (Shampine *et al.*, 1974) fitting, plus the *mucal* subroutine (written by Pathikrit Bandyopadhyay and available from <http://ixs.csrii.iit.edu/database/programs/mcmaster.html>) which calculates the total absorption cross section according to McMasters *et al.* (1969).

Input to *CHOOCH* is in the form of an ASCII data file containing a title, the number of data points from the fluorescence measurement and a list of data points, consisting of the X-ray energy in electron-volts and the recorded fluorescence signal on an arbitrary scale.

On executing the program, the user is provided with a graphical view of the observed raw fluorescence spectrum and is prompted to select energy ranges between which low-order polynomial fits will be performed in order to perform background fitting below and above the absorption edge, as described in §2.1. The user has the option to fit a horizontal line manually to the spectrum or have the computer automatically fit a first-, second- or third-order polynomial to the spectrum. The program displays the normalized curve, which the user may accept in order to proceed, or reject and try again. On proceeding, the program fits a smooth spline to the raw normalized spectrum to remove noise. This also allows the spectrum to be interpolated, producing a uniformly increasing energy scale, which is required in the next stage of the program.

CHOOCH then calculates the spectrum of f'' and f' AS factors as described in §§2.2 and 2.3. The integration limits used by *CHOOCH* in the determination of f' are read from an ASCII file (*atom.11b*), which is distributed with the program. The choice of values is based on the competing characteristics of speed and accuracy. The values

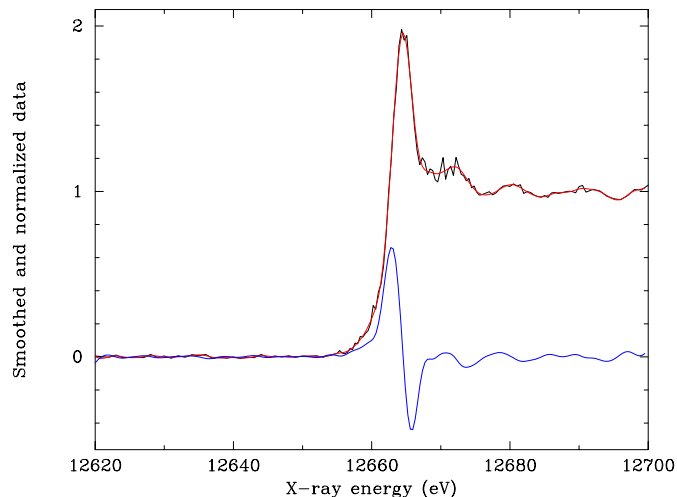


Figure 2
Raw spectrum from Fig. 1 after normalization (black) and after being smoothed using a spline fit (red). The blue line represents the first derivative of the splined data.

selected typically give accuracy to within a few tenths of an electron, which is comparable to or less than the expected systematic errors that may be introduced during the normalization stage. If higher accuracy is required, the associated ASCII data file may be modified manually.

The results of the calculation are written to an ASCII file and displayed as a graphical plot. The program determines the X-ray energies and AS factor magnitudes at the points corresponding to the f'' maximum and the f' minimum values, these being two energies which are typically measured in a MAD experiment. The program produces a brief summary file containing this information and can also on request produce a PostScript file of the AS curves.

4. Application

4.1. Program tests

The program has been in circulation for a number of years and is in use on MAD crystallography beamlines at the ESRF, HASYLAB, SRS, ELETTRA, SPring-8, APS, ALS, SSRL and NSLS synchrotrons,¹ as well as at numerous other crystallographic institutions. It has been used to calculate AS factors for most heavy-atom types used in protein crystallography and has provided the correct guidance as to the choice of wavelength for subsequently successful experiments, and also correct guidance for starting values for refinement. Indeed, comparisons between values of AS factors derived with *CHOOCH* and those refined at the stage of crystallographic phase determination by MAD have been made (Evans & Pettifer, 1996) and good agreement found.

The program has a number of limitations on accuracy, but the main contributing factor is the quality of the fluorescence data; the choice of tabular data is made in this light. Another limitation is that the AS may be anisotropic (Templeton & Templeton, 1988). If time and

¹ ESRF – European Synchrotron Radiation Facility, Grenoble, France; HASYLAB – Hamburger Synchrotronstrahlungslabor, Hamburg, Germany; SRS – Synchrotron Radiation Source, Daresbury, UK; ELETTRA – Trieste, Italy; APS – Advanced Photon Source, Chicago, USA; SPring-8 – Super Photon Ring - 8 GeV, West Harima, Japan; ALS – Advanced Light Source, Berkeley, USA; SSRL – Stanford Synchrotron Radiation Laboratory, Stanford, USA; NSLS – National Synchrotron Light Source, Brookhaven, USA.

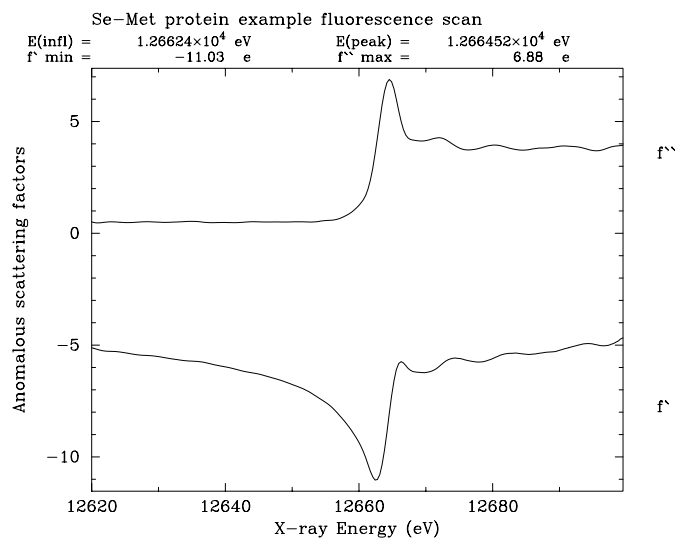


Figure 3
Final result as output by *CHOOCH*: experimental values of the anomalous-scattering factors about the *K*-absorption edge of Se in a seleno-methionine labelled protein.

radiation damage permit, angular-dependent fluorescence data could be obtained and the tensor components of AS derived using this program.

The value of a program of this type is that optimum conditions can be obtained for the actual crystal under study, at the beamline. This means that the resolution of the X-ray monochromator is automatically taken into account.

4.2. Example

A standard application of the program is in structure determination using seleno-methionine substituted proteins and the MAD method. An example of the use of *CHOOCH* is given here for the determination of the X-ray energies of the f'' maximum and the f' minimum at the selenium *K*-absorption edge in a small 16 kDa protein (Walsh *et al.*, 1999) solved by MAD at beamline 19ID of the Structural Biology Centre at the APS. Fig. 1 shows the raw fluorescence curve displayed when *CHOOCH* is run. The program prompts the user for parameters determining the type of fits to be used in normalization of the raw spectrum. Fig. 2 shows the normalized spectrum (black) and the smoothed spectrum (red), along with its first derivative (blue). Using the smoothed spectrum and its derivatives, *CHOOCH* calculates the f'' and f' spectrum and displays it (Fig. 3).

4.3. Experimental data requirements

The reliability of f'' and f' values determined using this program is governed almost exclusively by the quality of the experimental data used for the calculations. Spectra measured from macromolecular samples are typically very noisy and may only extend to ± 50 eV about the absorption edge. Such data may be prone to systematic error introduced by inadequate normalization. The normalization procedure and subsequent f'' calculation relies completely on the assumption that the experimental values of f'' approach those of theory far away from the absorption edge. As such, experimental data far away from the absorption edge are required at the normalization stage in order to satisfy this assumption. So how far must the fluorescence data extend either side of the absorption edge? Inspection of EXAFS data from a number of different samples yielding *K*- and *L*-

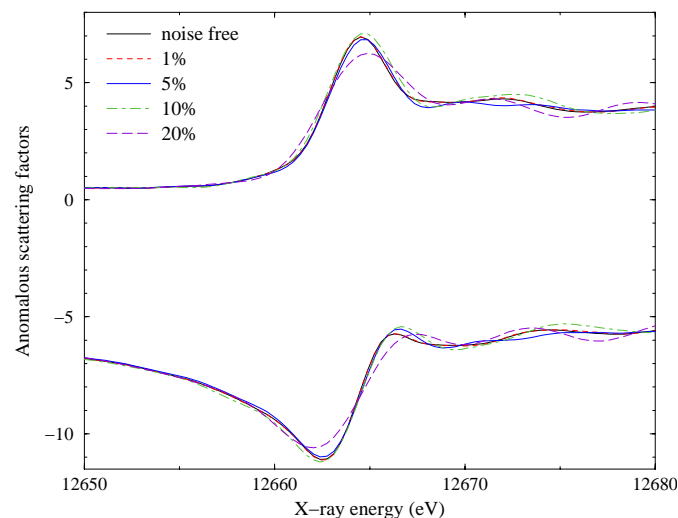


Figure 4
Effect of random noise in the experimental data on the results produced by *CHOOCH*. A smoothed fluorescence spectrum was used to simulate noise-free data; to this, random noise was introduced at the 1, 5, 10 and 20% levels. Normalization was carried out by fitting a first-order polynomial to the below- and above-edge regions over the same energy ranges in all cases. The resulting anomalous-scattering factor curves are displayed.

edge spectra provides an indication of the extent to which near-edge effects extend into the spectra below and above the edge. The conclusion is that the edge effects can sometimes be visible within ± 200 eV of an absorption edge. As such, fluorescence data points should ideally be recorded at energies >200 eV on either side of the edge and extending preferably out to ± 400 eV either side. These data need not be as finely sampled as the near-edge data and consequently should not significantly increase the time required to measure a spectrum or significantly increase the X-ray dose received by the sample. These are of course guidelines for an ideal case. Depending on the accuracy required for a particular experiment, some of the criteria may be relaxed.

The effect of random noise on the quality of results obtained with *CHOOCH* is depicted in Fig. 4, which shows results obtained from using a noise-free spectrum and spectra with noise introduced artificially at the 1, 5, 10 and 20% levels. The curves derived from spectra with 1 and 5% noise agree well with the noise-free result. At the level of 10% noise, we see a deviation in the predicted energy of the f'' maximum of ~ 0.2 eV and an increase in magnitude of ~ 0.2 e. The deviation is somewhat less at the f' minimum. Results derived with 20% noise show a more significant deviation from the noise-free result. The f'' maximum is found to be 0.5 eV higher in energy and ~ 0.7 e lower in magnitude. The f' minimum is 0.5 eV lower in energy and ~ 0.5 e higher in magnitude. The effects of noise therefore only appear to become significant at the 20% level. However, even at this level of error one can see clearly from Fig. 4 that the X-ray energies selected by *CHOOCH* would still correspond to points in the noise-free spectrum very close to the maxima and minima, and would therefore constitute successful use of the program. If the level of noise were higher than 20%, users of *CHOOCH* could run the risk of selecting X-ray energies which are incorrect by more than 0.5 eV.

5. System requirements, availability and documentation

CHOOCH has been successfully run on Unix platforms using DEC Alpha, SGI and Linux machines. The source code is available free of

charge and can be obtained by accessing <http://lagrange.mrc-lmb.cam.ac.uk/doc/gwyndaf/Chooch.html> or by contacting the authors directly. The program is distributed with full documentation describing its installation and use.

GE thanks the European Molecular Biology Laboratory for a pre- and post-doctoral fellowship.

References

- Baker, P. J., Farrants, G. W., Stillman, T. J., Britton, K. L., Helliwell, J. R. & Rice, D. W. (1990). *Acta Cryst.* **A46**, 721–725.
- Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*. London: Academic Press.
- Cromer, D. T. (1983). *J. Appl. Cryst.* **16**, 437–438.
- Cromer, D. T. & Liberman, D. (1970). *J. Chem. Phys.* **53**, 1891–1898.
- Doublé, S. (1997). *Methods Enzymol.* **276**, 523–530.
- Evans, G. & Pettifer, R. F. (1996). *Rev. Sci. Instrum.* **67**, 3428–3433.
- Golden, B. L., Gooding, A. R., Podell, E. R. & Cech, T. R. (1996). *RNA*, **2**, 1295–1305.
- Hartree, D. R., Kronig, R. D. L. & Petersen, H. (1934). *Physica*, **1**, 895–924.
- Helliwell, J. R. (1992). *Macromolecular Crystallography with Synchrotron Radiation*. Cambridge University Press.
- Hendrickson, W. A. (1999). *J. Synchrotron Rad.* **6**, 845–851.
- Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990). *EMBO J.* **9**, 1665–1672.
- Hendrickson, W. A., Pahler, 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. & Sheriff, S. (1985). *Methods Enzymol.* **115**, 41–55.
- Hoyt, J. J., de Fontaine, D. & Warburton, W. K. (1984). *J. Appl. Cryst.* **17**, 344–351.
- James, R. W. (1969). *The Optical Principles of the Diffraction of X-rays*. London: G. Bell.
- La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
- McMasters, W. H., Grande, D. N. K., Mallet, J. H. & Hubbell, J. H. (1969). *Compilation of X-ray Cross Sections*. Tech. Rep. UCRL-50174. Lawrence Radiation Laboratory, Livermore.
- North, A. C. T. (1965). *Acta Cryst.* **18**, 212–216.
- Okaya, Y. & Pepinsky, R. (1956). *Phys. Rev.* **103**, 1645–1647.
- Pettifer, R. F. & Hermes, C. (1985). *J. Appl. Cryst.* **18**, 404–412.
- Phillips, J. C. & Hodgson, K. O. (1980). *Acta Cryst.* **A36**, 856–864.
- Shampine, L. F., Davenport, S. M. & Huddleston, R. E. (1974). *Fit Discrete Data in a Least-Squares Sense by Polynomials in One Variable*. Fortran subroutine. Sandia National Laboratories, Albuquerque, USA.
- Templeton, L. K. & Templeton, D. H. (1988). *Acta Cryst.* **A44**, 1045–1051.
- Tesmer, J. J. G., Stemmler, T. L., Penner-Hahn, J. E., Davisson, V. J. & Smith, J. L. (1994). *Proteins Struct. Func. Gene.* **18**, 394–403.
- Todd, A. K., Adams, A., Powell, H. R., Wilcock, D. J., Thorpe, J. H., Lausi, A., Zanini, F., Wakelin, L. P. G. & Cardin, C. J. (1999). *Acta Cryst.* **D55**, 729–735.
- Wallace, B. A., Hendrickson, W. A. & Ravikumar, K. (1990). *Acta Cryst.* **B46**, 440–446.
- Walsh, M. A., Dementieva, I., Evans, G., Sanishvili, R. & Joachimiak, A. (1999). *Acta Cryst.* **D55**, 1168–1173.
- Woltring, H. J. (1986). *Test Programme for Generalized Cross-Validatory Spline Smoothing with Subroutine GCVSPL and Function SPLDER*. Fortran subroutine. University of Nijmegen and Philips Medical Systems, The Netherlands.