Conference Proceedings
DL-CONF-97-001

Recent Advances in Phasing

Proceedings of the CCP4 Study Weekend
January 1997

K S Wilson G Davies A W Ashton and S Bailey

August 1997

COUNCIL FOR THE CENTRAL LABORATORY OF THE RESEARCH COUNCILS

The Central Laboratory of the Research Councils
Chadwick Library
Daresbury Laboratory
Daresbury
Warrington
Cheshire
WA4 4AD
Tel: 01925 603397 Fax: 01925 603195
E-mail library@dl.ac.uk

ISSN 1362-0193

RECENT ADVANCES IN PHASING

Proceedings of the CCP4 Study Weekend
January 1997

Compiled by:

K. S. Wilson University of York
G. Davies University of York
A. W. Ashton Daresbury Laboratory
S. Bailey Daresbury Laboratory

Also available at:
http://www.dl.ac.uk/CCP/CCP4/main.html

CCLRC
Daresbury Laboratory
1997

CONTENTS

INTRODUCTION

Solving the phase problem is the crucial step in obtaining a crystal structure. For macromolecular crystallographers it may also be the most difficult and time-consuming step. The traditional methods used are Multiple Isomorphous Replacement and Molecular Replacement. This meeting covered aspects of some of the newer methods that are in use and some potential methods for the future.

Multiwavelength Anomalous Dispersion (MAD) was covered in some detail. In the first session, John Helliwell described the basic principles of anomalous scattering, Andy Thompson explained the beamline requirements of MAD and Janet Smith introduced us to MAD on proteins. This introduction to the method was followed by several case studies (Valerie Biou, Mark Peterson, Ian Glover and Harry Powell). Eric de la Fortelle described the use of the program SHARP for MAD phasing and Phil Evans talked about multiple wavelength simultaneous scaling. Rather more unusual uses of anomalous diffraction were described by Bill Shepard and Edgar Weckert.

The second day of the meeting was more theoretical. An introduction to direct methods was given by Zbyszek Dauter followed by two talks on *ab-initio* phasing of proteins (and the limitations of the method) by the Shake-and Bake method (Charles Weeks) and using ShelX (George Sheldrick). Gerard Bricogne covered maximum entropy techniques and Abraham Szoke introduced to us the application of holographic methods to crystallography. Case studies of using wARP to improve your phasing were presented by Anastassis Perrakis. The final presentations were on low resolution phasing strategies, based on solution scattering (Dimitri Svergun) or crystallographic images (Alexandre Urzhumtsev). Kevin Cowtan summarised what we had learnt in the meeting and led the discussion.

The meeting, this year, was held at York University for the first time. There were 418 participants in total, including 111 participants from Europe, 16 from the Americas and 2 from Japan. Bursaries, covering the cost of registration and accomodation were given to 232 young scientists, and an additional contribution was made to the travel costs of 23 young scientists from outside the UK. The speakers comprised 8 from the UK, 8 from elsewhere in Europe and 3 from the USA.

The meeting was organised and supported by the BBSRC Collaborative Computational Project in Protein Crystallography (CCP4). We thank the invited speakers for sharing their expertise with us and for the contributions to this booklet. We are greatful to Daresbury Laboratory for providing organisational support, with particular thanks to Diane Travers, Val Matthews and the rest of the SAS team who ensured that the meeting ran to plan.

Keith Wilson
Gideon Davies
Sue Bailey
Alun Ashton

ACKNOWLEDGMENTS

CCP4 would like to thank the following companies for their financial contributions to the CCP4 project in the year 1996. This support was an essential contribution to the costs of the meeting.

Abbott Laboratories
Amgen Incorporated
ARIAD
Banyu Pharmaceutical Company Ltd.
Bristol Myers Squibb
Chugai
Du Pont
Eli Lilly
Genetics Institute
Gentech, Incorporated
Glaxo UK Ltd
Glaxo USA
Green Cross Corporations
Hoffman la Roche and Co
Hoescht
Japan Tobacco Incorporated
Kirin Brewery
Kissei
KT GmbH
Kyowa Hakko
Mitsubushi
Monsanto
Norvatis
Pfizer Ltd
Pharmacia
Pharmacia SpA
Procter and Gamble
Rhone-Poulenc
Schering
Shionogi
SmithKline
Sumitomo
Vertex
Wyeth-Ayerst Laboratories
Yamanouchi
Zeneca U.S.A.
Zeneca Pharmaceuticals

Basic Principles of Anomalous Scattering and MAD

Professor John R Helliwell
Department of Chemistry
University of Manchester
M13 9PL, U.K.

1. Introduction

The phase problem in macromolecular crystallography is amenable to the multiwavelength anomalous dispersion method, which is now known as the MAD method. The theory of various approaches has been developed over a long period (see e.g. Okaya and Pepinsky (1956), Mitchell (1957), Herzenberg and Lau (1967), Karle (1967, 1980), Hoppe and Jakubowski (1975)). The harnessing of synchrotron radiation and the chance thereby to finely tune the X–ray wavelength around the X–ray absorption edge of a target (heavy) atom, so as to vary the scattering in amplitude and phase of that atom, has made the method practical. If an anomalous scatterer can certainly be introduced, as with the seleno–methionine labelling of proteins (Hendrickson (1991)), then relatively speedy structure determination of a new protein structure is practicable. The protein size that is tractable is advancing as synchrotron machine, and beamline, as well as detector capability, improves. Experimental knowledge of anomalous dispersion effects has also improved. This involves fundamental studies whereby very fine profiles (Arndt et al. (1982)) or anisotropies (Templeton and Templeton (1985)) have been investigated for bound atoms. Moreover different strategies in macromolecular crystal phase determination (Peterson et al. (1996)) as well as phase improvement and extension methods (Chayen et al. (1996)) have given practical ideas of these capabilities. For other examples see the companion papers in this Proceedings. The growth of the number of SR instruments in the last 15 years is testimony to a confidence in these approaches. For an overview see Helliwell (1992) and for an update see Chayen et al. (1996), as well as ideas and plans at the outset see Helliwell (1979). The purpose of this paper is to set down the principles and basis of anomalous scattering for MAD.

2. X–ray absorption and scattering

An electron of an atom can be ejected when a photon has sufficient energy. A heavy atom has K and L, or even M, edges in the wavelength range which is useful for X–ray crystallography. The atomic scattering factor for X–rays of that atom in the resonant condition becomes complex i.e. alters the normal scattering factor in amplitude and phase. The anomalous dispersion coefficients f' and f" are used to describe this effect and which are wavelength dependent. Hence for the heavy atom we have

$$f = f_o + f'(\lambda) + if''(\lambda) \qquad (1)$$

This equation thereby serves to correct for the standard, more simple, model of X–ray scattering. 'Normal scattering' is basically determined by the total number of electrons in the atom and which takes no account of absorption edge resonance effects. For a heavy atom this is not the situation for the wavelengths we use! For the light atoms (C, N, O and H) their corrections to the normal scattering are negligible for our purposes. A free atom (i.e. without neighbours) has a relatively simple form for the variation with wavelength of f' and f" (figure 1a). The edge wavelength is then where the scattering factor becomes complex. A bound atom has neighbours which can scatter back the ejected photoelectron and thereby seriously modulate the absorption effect and also alter therefore the X–ray scattering anomalous dispersion coefficients (figs. 1b and c). Furthermore the values can become dependent on direction (so called dichroism effects) as there can be for example a high density of neighbours in one direction or plane over another (for example see Templeton and Templeton (1985) for $K_2\,PtCl_4$). Moreover the exact edge position depends on the oxidation state of the atom as the inner shell electron can be more tightly bound when valence outer shell electrons are removed. Finally the X–ray wavelength bandpass needs to be considered because inherent effects can be masked if this is broader than the linewidths of the features naturally present. To guide our understanding of these phenomena the classical treatment of the 'forced harmonic oscillator with damping' will be reviewed in the next section as it forms the basis for the free atom case.

3. A mathematical model: Forced harmonic oscillator with damping

An inner shell electron can be modelled as having vibrational motion in an alternating electric field subject to a restoring force with associated spring constant k and a damping force proportional to the velocity (see Woolfson (1970, 1997)). The equation of motion is then described by the equation

$$m\ddot{x} + g\dot{x} + kx = E_o e\, e^{i\omega t} \qquad (2)$$

Since we can expect that the oscillation of the electron will not in general be in phase with the driving electric field $E_o\, e^{i\omega t}$ then the amplitude will prove to be complex. Hence the amplitude and phase angle between the driving force and the resultant oscillation must be considered. The undamped resonant frequency is given by

$$\omega_0^2 = k / m \qquad (3)$$

The amplitude solution to the forced damped case is then

$$x_o = \frac{E_o e / m}{\left(\omega_o^2 - \omega^2 + i\frac{g\omega}{m}\right)} \tag{4}$$

Hence the modulus of the amplitude is obtained by multiplying x_o by its complex conjugate x_o^* and taking the square root to give

$$|x_o| = \frac{E_o e / m}{\left(\left(\omega_o^2 - \omega^2\right)^2 + \frac{g^2\omega^2}{m^2}\right)^{1/2}} \tag{5}$$

and the phase angle of the oscillation relative to the driving force is then

$$\tan\varphi = -\frac{g}{m} \quad \frac{\omega}{\left(\omega_o^2 - \omega^2\right)} \tag{6}$$

Equations 5 and 6 tell us much about the basic resonance condition. This is obviously according to the mechanical oscillator model rather than where, at resonance, the electron is ejected as a photoelectron! So this is basically referred to as the classical treatment rather than the full quantum treatment. One aspect of the approximation then which is incorrect is that the imaginary part of the complex amplitude varies symmetrically about the resonance frequency but of course f", which is the analogous quantity, is only finite on one side of the frequency range (namely the high frequency, short wavelength range in the atomic case). Nevertheless a plot of the amplitude and phase versus ω, for various degrees of damping (expressed numerically as the Q factor of the resonance where $Q = \omega_o / \gamma$ and $\gamma = g / m$) shows that the resonance when free of damping (infinite Q) is instantaneous. At finite Q the resonance is not sharp and can be characterised by the frequency band ($\delta\omega$) over which the amplitude has a value $1/\sqrt{2}$ of the resonant amplitude. In fact

$$\omega_o / \delta\omega \approx Q \tag{7}$$

In the case of radiation, rather than a mechanical oscillator, the damping effect is usually attributed to radiation damping, which is weak and thereby Q is very large and the resonance is very sharp (see figure 1a) and whereby the core hole lifetime defines the line width. Note also that if the application of the electric field involves a frequency bandpass then this will degrade the resonance sharpness measured experimentally.

At low Q there is a marked difference between the frequency for maximum amplitude when undamped than when it is damped i.e.

$$\omega(\text{max. amplitude}) = \omega_o\left(1-\frac{1}{2Q^2}\right)^{1/2} \qquad (8)$$

In this treatment, as well as the limitation of ignoring the ejection of the photo–electron, we are assuming that the atom is a free atom and that of course is not true in the bound atom case.

4. X–ray anomalous dispersion coefficients and the KKT

The quantities f' and f" are related by the Kramers–Kronig transform (KKT for short). These are

$$f'(\omega) = \sum_i \int_o^\infty \frac{\omega^2 (dg/d\omega)_i \, d\omega'}{\left(\omega^2 - \omega'^2\right)} \qquad (9a)$$

$$f''(\omega) = \frac{1}{2}\pi\omega \sum (dg/d\omega)_i \qquad (9b)$$

Since the mass absorption coefficient is proportional to $(dg/d\omega)$ (i.e. f") the measured absorption or fluorescence spectrum from the sample can be used to derive f' also.

These equations give the minimum of f' at the half way point up the edge i.e. f" for the simple step function absorption edge case (note that this is not as shown in the paper Hoppe and Jakubowski (1975) page 445 figure 4 which places f' min at f" max and which is correct for a resonance of infinite Q; see also the tabulated theory cases i.e. as per Sasaki (1989) which is similar for the interval used of $10^{-4}\ \delta\lambda/\lambda$.

5. Maximal dispersive and anomalous signals in the 'white line' case

The presence of 'white line' features alters the simple edge shape (fig. 1b and c); in such a situation 'half way up' the edge (λ_A) is between the floor of the pre–edge region and the top of the spike and is the f' minimum; the top of the absorption spike is a value of f' mid way between f' min and f' max (λ_B); the point half way down the absorption spike to the post edge plateau is an f'

maximum (λ_C). Clearly the maximal $\Delta f'$ and a maximal f'' to be stimulated can be probed by use of the wavelengths λ_A and λ_C as well as λ_B respectively. Use of λ_A and λ_C alone, whilst yielding maximal $\Delta f'$ would (via λ_C) involve a slight reduction in size of f'' harnessed.

6. Concluding remarks and the scope of the method

Anomalous scatterers can now be purposely inserted into a protein such as seleno–methionine labelling (Hendrickson (1991)) or the wide range of accessible absorption edges for different types of metal atom derivative utilised. Anomalous dispersion effects are optimised when the X–ray wavelength is tuned right to an absorption edge. Further enhancements are possible experimentally (for a recent review see Chayen et al. (1996)) by use of high brilliance insertion devices such as tunable undulators, in conjunction with high heat load optics, whereby a high flux onto the sample is preserved even with a tiny bandpass (i.e. 10^{-5} to 10^{-4}) so as to explore sharper linewidth features and possibly dichroism as well. Moreover hybrid experimental strategies involving a 'quick pass' data collection at a minimum of two wavelengths followed by a 'slow pass' higher resolution data collection at one wavelength could allow novel theoretical atomic resolution direct methods to be combined with the MAD methods. These then are the sort of structure determination tools that can form the basis to tackle the tens of thousands of protein structures of interest on genome scales.

References

Arndt, U.W., Greenhough, T.J., Helliwell, J.R., Howard, J.A.K., Rule, S.A. and Thompson, A.W. Nature, *298* (1982) 835–838.

Chayen, N.E., Boggon, T.J., Cassetta, A., Deacon, A., Gleichmann, T., Habash, J., Harrop, S.J., Helliwell, J.R., Nieh, Y.P., Peterson, M.R., Raftery, J., Snell, E.H., Hadener, A., Niemann, A.C., Siddons, D.P., Stojanoff, V., Thompson, A.W., Ursby, T. and Wulff, M. Quarterly Reviews in Biophysics, *28* (1996) 227–278.

Helliwell, J.R. Daresbury Study Weekend Proceedings DL/SCI, *R13* (1979) p1–6.

Helliwell, J.R. Reports on Progress in Physics, *47* (1984) 1403–1497.

Helliwell, J.R. "Macromolecular Crystallography with Synchrotron Radiation" Cambridge University Press (1992).

Hendrickson, W.A. Science, *254* (1991) 51–58.

Herzenberg, A. and Lau, H.S.M. Acta Cryst, *22* (1967) 24–28.

Hoppe, W. and Jakubowski, U. in Anomalous Scattering. Edited by Abrahams, S.C. and Ramaseshan, S. (1975) pp 437–461.

Karle, J. Applied Optics, *6* (1967) 2132–2135.

Karle, J. Int. J. of Quantum Chemistry, *7* (1980) 356–367.

Mitchell, C.M. Acta Cryst, *10* (1957) 475–476.

Okaya, Y. and Pepinsky, R. Phys. Rev., *103* (1956) 1645.

Peterson, M., Harrop, S.J., McSweeney, S.M., Leonard, G.A., Thompson, A.W., Hunter, W.N. and Helliwell, J.R. J. Synchrotron Rad., *3* (1996) 24–34.

Sasaki, S. KEK Report (1989) 88–14. National Laboratory for High Energy Physics, Tsukuba, Japan.

Templeton, D.H. and Templeton, L.K. Acta Cryst, *A41* (1985) 365–371.

Woolfson, M.M. X–ray crystallography (1970) [2nd Edition 1997] Cambridge University Press.

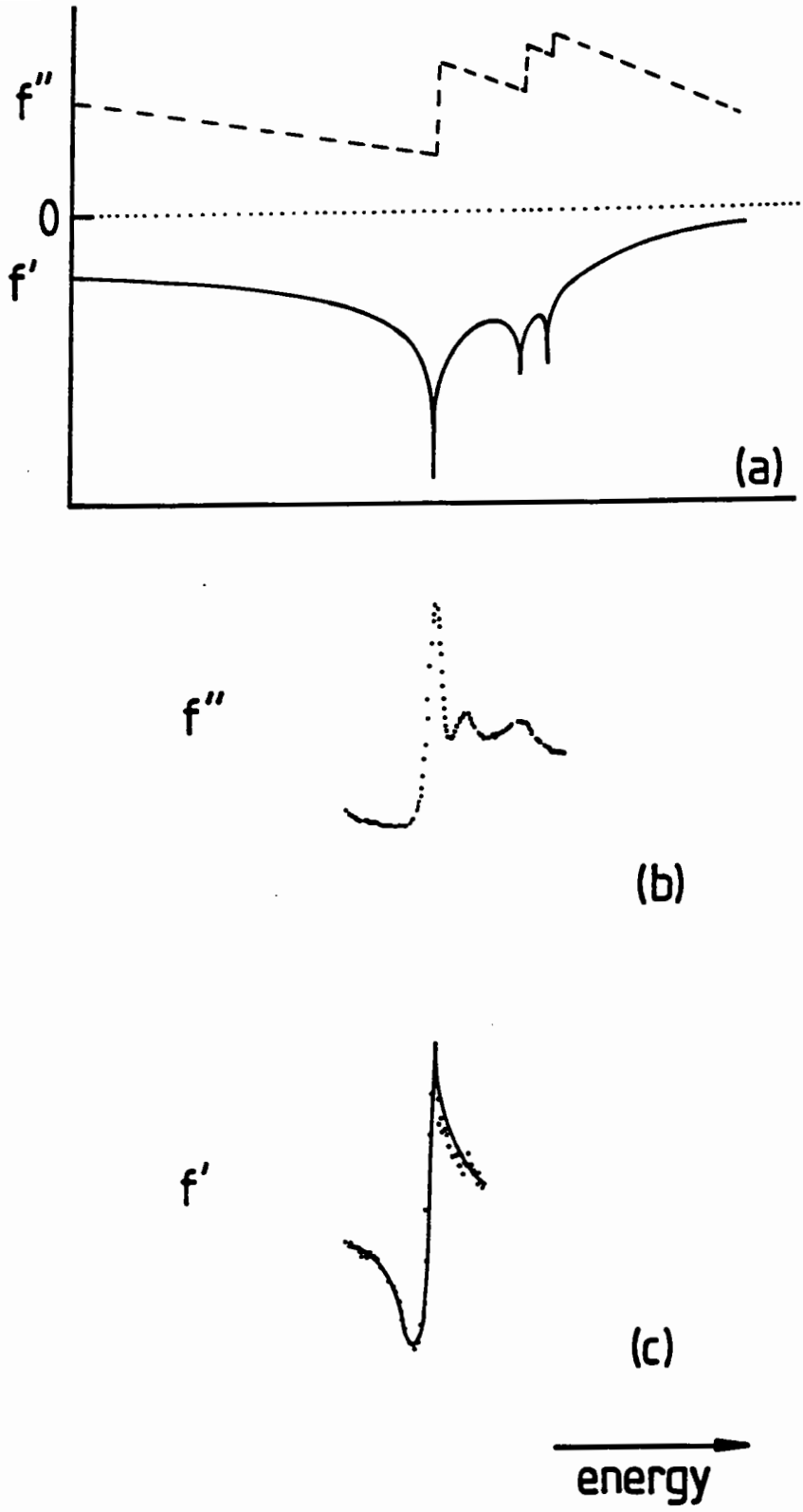


Figure 1

(a) Theoretical values of f' and f'' with wavelength for platinum around its L_I, L_{II} and L_{III} edges illustrates the typical situation for a free atom. From Sasaki (1989) tabulated in intervals of 0.0001Å very near to the edges and in intervals of 0.01Å remote from the edges.

(b) Measured 'monochromatic step scan' absorption curve for the L_{III} edge of $K_2\,Pt(CN)_4$ (Helliwell (1984)).

(c) Measured f' (by a polychromatic profile approach) for rhenium L_{III} edge (from Arndt et al. (1982)).

Design of Synchrotron Beamlines for MAD Protein Crystallography - ESRF BM14.

A. Thompson *

24th April 1997

*EMBL Grenoble Outstation

Introduction

The method of Multiple Wavelength Anomalous Diffraction [MAD] has been developed extensively in recent years by Hendrickson [1], Fourme [2] and Smith [3]. Small variations in intensity of diffraction spots due to resonant scattering of a heavy atom excited by the frequency of the incoming X-rays can be used in order to solve the crystallographic phase problem. The techniques and physics involved are described in the above reviews. The method is best pursued using the variable wavelength of radiation available at synchrotron sources, although in cases where the anomalous scattering is strong with respect to the size of the molecule allied techniques may be used on rotating anode sources ([4],[5] and [6]).

The "anomalous" intensity changes produced in the diffraction patterns are typically of the order of a few percent, and are (unfortunately) of a similar size to the errors (random and systematic) in *good* data! The methods of designing a beamline to maximise the chance of being able to collect MAD data by being aware of where error may creep into measurements are discussed in conjunction with the design of the MAD beamline (BM14) at the ESRF, its advantages and drawbacks.

Generalities

It is very important to appreciate that a synchrotron beamline (whether for protein crystallography or anything else) is a single instrument designed for a specific function or type of measurement. The more flexible the beamline, the more compromises are likely to have been made in its design! Moreover almost any decision about a beamline parameter affects all the rest. For example if you require to study extremely small crystals and the X-ray source size is large compared to the crystal, substantial source demagnification (and hence increased divergence through the sample) may be required. This will have a knock on effect on the type of 2-D detector

required - a large detector set far away from the sample would not necessarily be a good choice because of the increased divergence of the beam after the sample.

Beamline optimisation should then be discussed in terms of the source, optics, sample environment and detector, although I will discuss only the first two. The requirements of MAD will be discussed in the light of each of these. For a rigorous and well thought out discussion of beamline design, see Nave [7].

The Properties of the Source - how they help or hinder!

Some General Principles

Basically 3 types of source are available at modern synchrotrons - bending magnets (dipoles) or one of two insertion devices (wigglers or undulators). Their properties have been extensively discussed elsewhere. Table 1 gives several typical properties of these devices based on the ESRF source. The vertical and horizontal source size and divergences are matched to the sample size and mosaicity via conditioning optics. To achieve a narrow monochromatic bandpass with maximal flux, the divergence of the synchrotron beam in the dispersive direction must be matched to the natural monochromator bandpass.

Wavelength Range

MAD experiments have been reported at various absorption edges between the U M_V edge, 3.482 Å [8] and the Xe K edge, 0.358 Å [9]. Table 2 gives a copy of the periodic table showing absorption edges which may in principle at least be used - almost all of them! The source should be tunable between these limits, therefore. Some devices on 4th generation synchrotrons (bending magnets, wigglers) fulfil this requirement easily. Undulators, however, give a high peak brilliance in a narrow energy range (figure 1) with several harmonics of this energy, and therefore have to be tuned to give optimal intensity for a 3 wavelength MAD experiment. In this case, "tuning" refers to changing the gap between the permanent magnets in order to change the field and shift the energy spectrum. Wiggler and bending magnet sources are, by their nature, smoother, and hence require no "tuning" over the above wavelength range.

The above tuning of the undulator gap can perturb the electron beam position in the storage ring if the undulator field is not sufficiently symmetric (due to errors in construction or the increasing distortion of its frame due to the increasing magnetic field). Undulator "shimming" techniques developed at the ESRF have now improved to such an extent that this field inquality is much reduced and "tuning" is routinely possible L10]. In addition, shorter wavelength experiments which will need much tighter gap setting (for example for a MAD experiment at the Xe

edge using the 5th harmonic of a "typical" undulator, a gap of 17 mm would be required) will soon be possible at the ESRF [11] and [12]). It is worth noting, however that a beamline *optic* that is optimised for 0.35 Å will not necessarily be optimised for 3.5 Å! In other words multiple lines may be needed to cover the whole wavelength range.

Source size and divergence

The source characteristics should be matched to the sample. Nowadays a "typical" crystal seems to be more like 150 μm than 300 μm on edge. Source sizes and the corresponding convergence angles for a bending magnet, wiggler and high and low β undulator on the ESRF are given in table 3, assuming the source is demagnified to a 150 μm image. A typical sample mosaicity (after freezing) may be 0.3° (5 mrad). The divergence of the beam at focal spot should ideally be less than this.

Stability and Reproducibility

A drifting wavelength causes different values of f' and f'' to be "mixed" together, whereas a drifting intensity causes errors in inter filmpack scaling. The stability required depends on the actual bandpass chosen and the type of experiment performed. On BM14, the bandpass has been measured to be 2.4×10^{-4} from an Si (111) monochromator. For the anomalous signal to vary only 5% during the experiment with this bandpass and a wavelength of 1 A, the energy must be stable to 0.2 eV over the period of the experiment. An energy stability of 0.5 eV rms can thus be regarded as good enough for all practical purposes, and this is easily within the measured source stability of the ESRF.

The beamline optics can also contribute to beam instability by heating, vibration or other obscure effects! The source power should thus not exceed practical limits of what modern (cooled) optics can tolerate. Typical power densities at 30m from the ESRF source with a single section insertion device (up to 3 possible) and 100 mA stored beam followed by the power ACTUALLY USED in recording a diffraction pattern (assuming 2.5×10^{-4} bandpass) are given in Table 4 [13] Many papers have been written on high power beam optics, and several successful approaches are available for both undulator and wiggler beams. Even so, when the requirement is for stability, reproducibility and reliability, it seems logical to use the most efficient source available, ie the undulators.

Source Intensity

This has to be adequate to perform the experiment in a "reasonable" period of time, both from the point of view of getting good data ("systematic errors" can be time dependent, and sufficient couting time per image has to be spent to have adequate statistics) and getting your experiment scheduled when there is so much pressure on synchrotron beamlines! At the ESRF (even on the bending magnet beamline) the detector readout time is almost always longer than the required exposure time (for image plate detectors). The overall source intensity chosen therefore depends only on the thermal properties of the beam (see above) and the capacity of the sample to withstand the beam.

Optics.

Required Wavelength Resolution

What is the appropriate energy bandpass for MAD experiments? These limits are defined by the fineness of the sharpest features ("white lines") in your absorption spectrum. Krause and Oliver [14] tabulate the I; and L level X-ray line widths for Z from 1 to 110 . The widths result from the lifetime of excited states, limited by the uncertainty principle $\Delta E \Delta t = \frac{h}{2\pi}$ "White line" effects (Lye et al [15], Brown et al [16] and Lytle et al [17]) are present in L "edges" (where a transition from a core level to one of a high possible density of final states gives the "amplified" nature of the feature) or K edges (which may be due to different transitions to exciton levels). The line widths of some L edge white lines have been reported by Arp et al [18] and are in close agreement with the above tabulated values for $Z \geq 50$. Hence the required energy bandpass for a MAD beamline should be such that the instrument broadening due to the monochromator bandpass, vibration etc do not obscure the natural widths of features on the absorption edges. Table 5 gives (in percent) the ratio between the "core-hole" lifetime as given in [14] and energy of the X-ray absorption edge. Features are broadened by the convolution of the instrument resolution and the width of the feature. In the case of the fine Se absorption edge, a bandpass of 1×10^{-4} would give effectively no broadening. It can be inferred that a bandpass lower than around 1.0×10^{-4} will limit intensity without making features significantly sharper. However it is also clear that in some cases (for example Hg) such a narrow bandpass will give no useful increase in anomalous signal. It may be then useful to have a facility with a changeable bandpass between 1.0×10^{-4} and 3.0×10^{-4} such as could be achieved, for example, by an interchangeable pair of Si (111) and Si(311) monochromators such as are available on BM14 (allowing for increased bandpass due to the intrinsic Si quality, surface finish etc).

Possible Optical Configurations for MAD - their advantages and drawbacks.

The combination of a vertical focussing mirror and a horizontal focussing triangular monochromator as described in [19] is one of the most commonly used optics for synchrotron radiation. This geometry can be used for MAD measurements where the beamline bandpass (Emitted by the monochromator bandpass and beam horizontal divergence) is small enough, for example in [10]. It offers several important advantages:

- Can focus a large horizontal acceptance, depending on the exact geometry.
- Angle of monochromator, sample and collimation defines wavelength therefore changes in machine vertical and horizontal orbit or mirror vibration cause only intensity (not wavelength) changes.
- Easy to get fine focal spot - Si or Be give adequate bandpass.
- Permits "Simultaneous" MAD t20] if used with a large crossfire angle and hence bandpass, but can be used for ordinary MAD [10] with a narrow bandpass monochromator and pencil beam.

On the other hand the system has the following disadvantages:

- A nuisance to change wavelength, and hard to make wavelength change reproducible (need to move monochromator and sample and detector).
- Crossfire from the focussing - either the beamline is very long, or has a large defocussing ratio.

A more recent approach to a MAD beamline is evidenced by SRS 9.5 [21] which uses a toroidal mirror and channel cut monochromator to provide a beam whose position changes very little with wavelength, and a tight focus. This very simple optic has two major disadvantages :

- Need to limit vertical aperture (and hence intensity) to get a narrow energy bandpass.
- Small horizontal acceptance (less than 2 mrad).

It is interesting to note that both these disadvantages are almost eliminated with the narrow vertical and horizontal divergences of an undulator source.

Of course there are many variations on the theme of focussing beam with narrow bandpass and constant focal spot position with wavelength, for example the extremely successful Howard Hughes Medical Institute line at Brookhaven [22] which uses a horizontally focussing constant height monochromator and a vertically focussing mirror.

The ESRF beamline D2AM [23] uses a similar horizontal focussing monochromator, sandwiched between two X-ray mirrors. The first of these collimates the incoming radiation in the vertical ie there is (almost) no beam divergence component to the energy bandpass of the monochromator. The monochromatic beam is subsequently focussed by the second mirror.

The horizontally focussing monochromator arrangement has two main disadvantages in producing a fine focal spot -

- Its main advantage lies in focussing large horizontal divergences. However unless the surface shape of the monochromator can be made conical rather than cylindrical, it has to be used close to a 3:1 demagnifying configuration to give maximum throughput into the narrow bandpass [24]. This demagnifying ratio causes significant beam crossfire and hence reduces the unit cell size resolvable for a given detector.

- The bending radius has to be continually changed with energy in order to keep a tight focal spot (and hence stable intensity with wavelength variation on the sample). In fact the change in intensity over a typical experimental energy range can be large [25,26].

The approach adopted on ESRF beamline BM14 is a combination of the above approaches, with a collimating mirror, channel cut monochromator and a vertical and horizontal focussing mirror. The great disadvantage of this system are the optical aberrations due to the double mirror arrangement (the so called "slope error" contributes twice to the vertical broadening of the focal spot) and the differing object and image distances for the vertical (source at "infinity" as the beam is collimated) and horizontal (source at "source" position) foci. This latter effect is in practise very small for a narrow horizontal beam acceptance and small mirror grazing angle. A recent scheme for the proposed British Beamline at the ESRF [25] suggested using parabaloidal mirrors to limit this aberration.

Other geometries exist, for example the novel trichromator [27] principle proposed for the RIKEN beamline on SPRING-8. The success of this approach is for the future to reveal!

BM14

The beamline BM14 was designed to give a high stability beam with maximum intensity into a narrow wavelength bandpass.

- Wavelength Range 0.6 - 1.8 Å
- Intensity approximately 2.5 x 10^{12} photons per second at 1 Å wavelength into a focal spot of 0.8 mm x 0.3 mm.
- Rapid tunability - a few minutes to scan the whole wavelength range.
- Wavelength stability 0.1 eV rms.
- Intensity stability 0.8 % rms.
- Fast readout detectors (Offline FUJI plate, CCD and MAR 345)

To date a fifteen new structures have been solved on BM14 by MAD phasing, with several others in various stages of data analysis. Amongst these the weakest anomalous signal was 2.5%, and a typical signal is more like 3.5%. (Data quality (and hence crystal quality) has to be extremely good to use signals below 3% for MAD).

Beam intensity and wavelength stability can be good enough to collect data from stable (cryo-cooled) crystals in a random orientation, and a substantial fraction of the successful MAD phasing experiments on BM14 have been performed in this way. Care must of course be taken in order to record complete data including all measurements of Bijvoets, and to have a large redundancy of measurement. Inverse beam data collection or setting of a crystal around a mirror plane are preferable when searching for an extremely small signal, with an unstable beam or with uncooled crystals. A typical experiment on BM14 takes between 24 and 48 hours, depending on whether image plate or CCD detector is used.

The largest number of anomalous scatterers (that I am aware of) used in ab initio MAD phasing was 12 (a bacterial synthetase on the French ESRF Beamline D2AM by Fanchon and Bertrand). Examples of 6-10 anomalous scatterers are relatively common (for example Matias [28], Ceska [29]). Larger structures (J. A. Smith) have been shown to give good phases based on already refined Se signals, and direct methods has been used to identify 18 out of 22 Se sites in the asymmetric unit from a MAD data set collected on BM14 [30]

Recently some attention has been paid to the advantages of measuring phases to very high resolution [31], the ability to do this without worries of lack of isomorphism being one of the major advantages of MAD.

What the Future Holds for MAD

Interest is now being focussed on undulator beamlines for MAD protein crystallography (Westbrook(APS), Thompson et al [32], Shapiro et al [10]), where smaller crystals can be studied, or higher resolution data should be improved due to the improved beam collimation. Data collection times will become much smaller, and with rapid readout CCD detectors, fine phi oscillations will improve data quality due to the improved background per reflection. Finally, with extensive work by Schiltz et al [33] and increasing interest in soft "M" absorption edges, MAD will soon be extended to new wavelength ranges.

Conclusions

- MAD is a routine technique on modern beamlines, and many technical problems have solved (only to reveal subtle new ones!)
- An undulator is the best currently available source for a MAD beamline.
- Several optics solutions exist which should be well adapted to an undulator source.

Acknowledgements

Although the comments in the above paper are by me, the success of BM14 is due to the work of many people, namely :
V. Biou (IBS Grenoble), G. Blattman (ESR.F), L. Claustre (EMBL Grenoble), A. Gonzalez (EMBL Hamburg), S. Laboure, G. Leonard, G. Marot and M. Mattenet (ESRF), V. Stojanoff and O Svensson (ESRF) and P. Thorander (Siemens).

References

1. Wayne A Hendrickson. Science **254** (1991) pp 51- 58.

2. Roger Fourme, William Shepard and R.ichard Kahn. Progress in Biophysics and Molecular Biology **64** (1996) pp 167 - 199.
3. Janet L. Smith. Current Opinion in Structural Biology **1** (1991) pp 1002 - 1011.
4. Wayne A Hendrickson and Martha M Teeter. Nature **290** (1981) pp 107 - 113.
5. S-E Ryu, P D Kwong, A Truneh, T G Porter, J Arthos, M Rosenberg, X Dai, N Xuong, R Axel, R W Sweet and W A Hendrickson. Nature **348** (1990) pp 419 - 426.
6. Mariusz Jaskolski and Alexander Wlodawer. Acta Crystallographica **D52** (1996) pp 1075 - 1081.
7. Colin Nave. In preparation.
8. Hendrickson, personal communication.
9. Roger Fourme, personal communication.
10. L. Shapiro, A. Fannon, P. Kwong, A. Thompson, M. Lehmann, G. Grubel, J-F. Legrand, J Als-Nielsen, D. Colman and W. Hendrickson. Nature **374** (1995) pp 327 - 337.
11. ESRF Internal Memo, Ropert (1996).
12. ESRF Annual Report (1995) p 103.
13. ESRF "Red Book" (1991).
14. M. O. Krause and J. H. Oliver. Journal of Physical and Chemical Reference Data **8** no. 2 (1979) pp 329 - 338.
15. Richard C. Lye, James C. Phillips D. Kaplan, S. Donaich and Keith O. Hodgson. Proceeding of the National Academy of Science (USA) **77** (1980) pp 5883 - 5888.
16. M. Brown, R. E. Peierls and E. A. Stern. Physical Review B **15** (1977) pp 738 - 744.
17. F. W. Lytle, P.S.P. Wei, R. B. Greegor, G. H. Via and J. H. Sinfelt. Journal of Chemical Physics **70** (1979) pp 4849 - 4855.
18. U. Arp, G. Materlik, M. Meyer, M. Richter and B. Sonntage. in X-Ray Absorption Fine Structure. Edited by S S Hasnain (1991) pp 44 - 47.
19. M. Lemmonier, R. Fourme, F. Rousseaux and R. Kahn. Nuclear Instruments and Methods **152** (1978) pp 173 - 177.
20. Lee and C. Ogata. Journal of Applied Crystallography **28** (1995) pp 661 - 665.
21. A. Thompson, J. Habash, S Harrop, .J. R. Helliwell, C. Nave, P. Atkinson, S. S. Hasnain, I. D. Glover, P. R. Moore, N. Harris, S. Kinder and S. Buffey. Review of Scientific Instruments **63** (1992) pp 1062 - 1064.
22. J-L. Staudenmann, W. Hendrickson and Abramowitz. Review of Scientific Instruments **60** (1989) pp 1939 - 1942.
23. J-P. Simon, E. Geissler. A-M. Hecht, F. Bley, F. Livet, M. Roth, J-L. Ferrer, E. Fanchon, C. Cohen-Addad, J-C Thierry. Review of Scientific Instruments **63** (1992) pp 1051 - 1054.
24. C. Sparks, Borie and J. Hastings. Nuclear Instruments and Methods **172** (1980) pp 237- 242.
25. D. Paul, M. Cooper and W. Stirling. Review of Scientific Instruments **66** no. 2 (1995) pp 1741 - 1744.
26. A. Thompson and Å. Kvick. Report to the ESRF Scientific Advisory Committee (1992).

27. M. Yamamoto, T. Fujisawa, M. Nagasako, T. Tanaka, T. Uruga, H. Kimura, H. Yamaoka, Y. Inoue, H. Iwasaki, T. Ishikawa, H. Kitamura and T. Ueki. Review of Scientific Instruments **66** no. 2 (1995) pp 1833 - 1835.

28. P. Matias, J. Morais, A. Coelho, R. Meyjers, A. Gonzalez, A. Thompson, L. Sieker, J. LeGall and M-A. Carrondo in preparation.

29. T. A. Ceska, J. R. Sayers, G. Stier and D. Suck. Nature **382** (1996) pp 90 - 93.

30. J. A. Smith (personal communication) using software developed by Woolfson et al (RANTAN).

31. F.T. Burling, W.I. Weis, K.M. Flaherty and A.I. Brunger. Science **271** (1996) pp 72 - 77.

32. V. Biou, A. Gonzalez, J. Helliwell, S. McSweeney, J.A. Smith and A. Thompson (1995). Report to the ESRF Science Advisory Committee.

33. M. Schiltz. W. Shepard, R. Fourme, T. Prange, E. De la Fortelle and G. Bricogne. Acta Crystallographica D**52** (1997) pp 78 - 92.

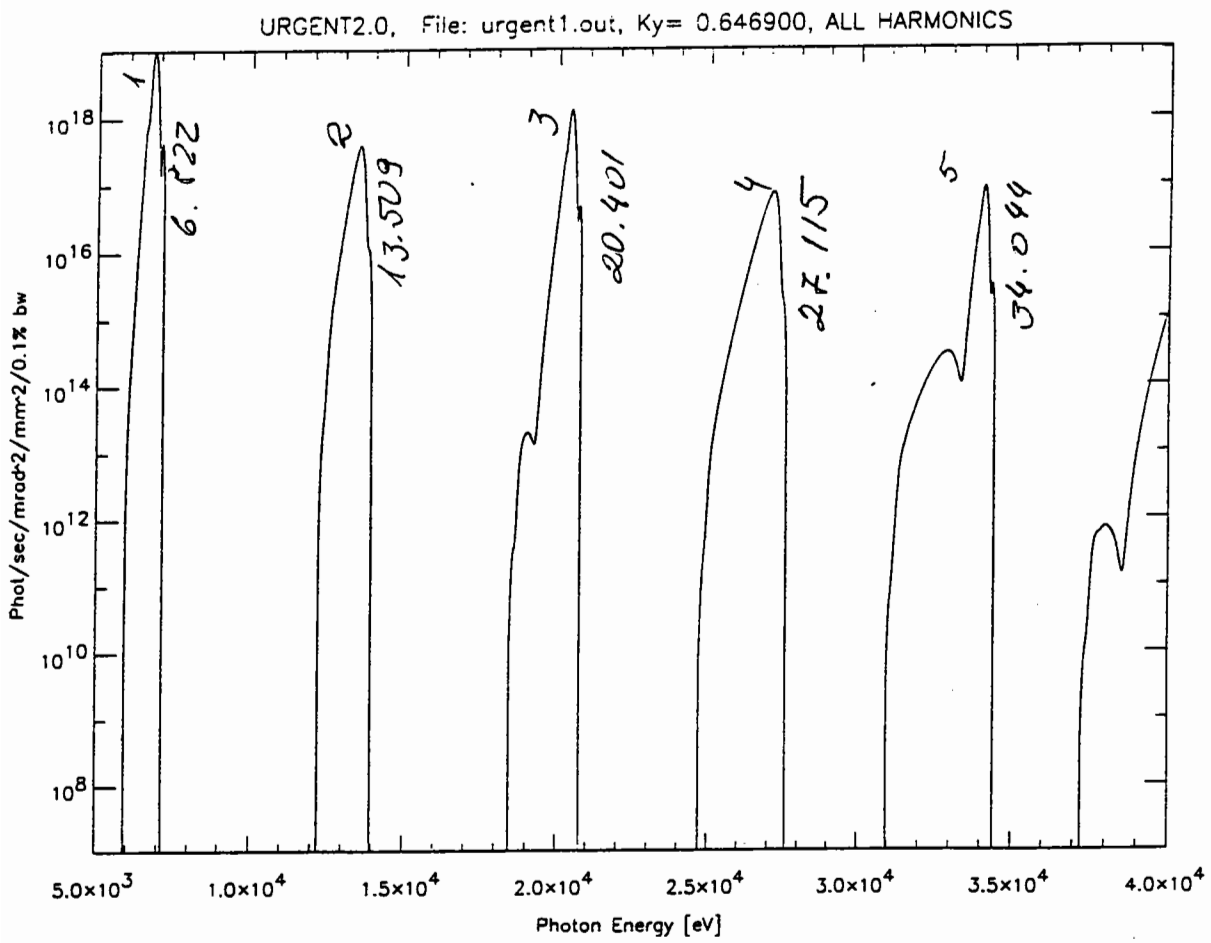


Figure 1

A calculated ESRF Undulator spectrum showing the discrete and narrow peaks of X-ray intensity. Brilliance is plotted against energy.

Table 1

ESRF Source Properties At beam exit from the shield wall Taken from ESRF "Red" Book Storage Ring at 200 mA					
	Power KW	Horiz Size mm	Vert Size mm	Power / length W/mm	Power / area W/mm**2
B.M.	0.84	126	2.3	6.7	2.9
Und.	2.4	5.5	3.3	495	149
Wigg.	14	52	3.3	252	76

Table 2

Note on table 2:-

Normally the larger edge transitions are better for MAD ie an L edge gives a bigger signal than a K edge etc. For this reason only the L_3 and M_5 edges are quoted. The usefulness of the absorption edges also (on average) increaases with increasing Z. However the difficulty in collecting accurate data at long wavelengths (low energies) should be borne in mind when preferring, for example the Xe L_3 edge to the K edge. Certain elements give "white lines" (for example Se) which "amplifies" the signal available. Where "white lines" occur for a K edge, they should also be present for L_1. When they occur for L_3 they should also be present for L_2.

Key to table 2:-

11.111 – Accessible for MAD on most beamlines

11.111 – Accessible for MAD on some beamlines

11.111 – Not accessible for MAD

*** 11.111 *** – Already successfully used for MAD

** 11.111 ** – Of potential future interest for MAD

Energies are given in KeV

Periodic Table Showing Absorption Edges			
Element	K	L_3	M_5
H	0.016		
He	0.025		
Li	0.055		
Be	0.112		
B	0.188		
C	0.284		
N	0.410		
O	0.543		
F	0.697		
Ne	0.870		
Na	1.071		
Mg	1.303		
Al	1.559		
Si	1.839		
P	*2.149*		
S	*2.472*		
Cl	*2.833*		
Ar	*3.206*		
K	*3.608*		
Ca	*4.039*		
Sc	*4.492*		
Ti	*4.966*		
V	*5.465*		
Cr	*5.989*		
Mn	**6.539**		
Fe	***7.112***		
Co	**7.709**		
Ni	**8.333**		
Cu	***8.979***		
Zn	***9.659***		
Ga	**10.367**		
Ge	**11.103**		
As	**11.867**		
Se	***12.658***		
Br	***13.474***		
Kr	**14.326**		
Rb	**15.200**		
Sr	**16.105**		
Y	**17.038**		
Zr	**17.998**	*2.233*	
Nb	**18.986**	*2.371*	
Mo	*20.000*	*2.520*	
Tc	*21.044*	*2.677*	
Ru	*22.117*	*2.838*	

Rh	*23.220*	*3.004*	
Pd	*24.350*	*3/.173*	

Periodic Table Showing Absorption Edges			
Element	K	L_3	M_5
Ag	*25.514*	*3.351*	
Cd	*26.711*	*3.538*	
In	*27.940*	*3.730*	
Sn	*29.200*	*3.929*	
Sb	*30.491*	*4.132*	
Te	**31.814**	**4.341**	
I	**33.169**	**4.557**	
Xe	**34.561**	**4.782**	
Cs	*35.985*	*5.012*	
Ba		*5.247*	
La		*5.483*	
Ce		*5.723*	
Pr		*5.964*	
Nd		*6.208*	
Pm		*6.459*	
Sm		**6.716**	
Eu		**6.977**	
Gd		**7.243**	
Tb		**7.514**	
Dy		**7.790**	
Ho		**8.071**	
Er		**8.358**	
Tm		**8.648**	
Yb		**8.944**	
Lu		**9.244**	
Hf		**9.561**	
Ta		**9.881**	
W		**10.207**	
Re		**10.535**	
Os		**10.871**	
Ir		**11.215**	
Pt		**11.564**	
Au		**11.919**	
Hg		**12.284**	*2.295*
Tl		**12.658**	*2.485*
Pb		**13.055**	*2.586*
Bi		**13.419**	*2.580*
Po		**13.814**	*2.683*
At		**14.214**	*2.787*
Rn		**14.619**	*2.892*
Fr		**15.031**	*3.000*
Ra		**15.444**	*3.105*

Ac		**15.871**	*3.219*
Th		**16.300**	*3.332*
Pa		**16.733**	*3.442*
U		***17.166***	*3.552*

Table 3

Beam Size and Convergence of ESRF Straight Sections				
Horiz size	Horiz Div	Horiz Conv	Vert Size	Vert Div
Low β Position				
0.97	37.6	243	0.23	13.9
High β Position				
0.11	197	145	0.47	15.5
All sizes in mm, divergences in μrad, FWHM.				

Table 4

Power emitted by "typical" insertion devices at the ESRF, 100mA stored current		
Source	Power Density 30m	Percentage Power Used
Bending Magnet	1.1	0.0013
Undulator (high β)	20	0.01
Wiggler	20	2×10^{-6}

Table 5

Key to table 5:-

Energy of edge in KeV, width of absorption edge in eV, "WL" width is the"white line width" in eV.

Theoretical Natural widths and measured values of several absorption edges				
Element	Edge	Width	Ratio	"WL" width
Fe	7.112	1.25	1.76×10^{-4}	
Se	12.658	2.33	1.84×10^{-4}	
Hg	14.209	5.5	3.87×10^{-4}	
Pt	13.273	5.86	4.41×10^{-4}	5.2×10^{-4}
Yb	8.944	5.14	4.6×10^{-4}	4.5×10^{-4}

Multiwavelength Anomalous Diffraction in Macromolecular Crystallography

Janet L. Smith
Department of Biological Sciences
Purdue University
West Lafayette, Indiana 47907
USA

Introduction

Multiwavelength anomalous diffraction (MAD) is the fastest growing method of structure determination in macromolecular crystallography. At least twenty-five new structures solved with MAD were published in the past year. Many factors contribute to the growth of MAD, and its future is extremely bright. The experience gained over the past several years is now being generalized to make MAD more accessible. This paper aims to present a practical overview of MAD. I first review the observational equation for MAD and describe the basis of the phasing signal and how it is estimated for specific problems. This is followed by a discussion of the design of a MAD experiment, schemes for data analysis and phasing, and considerations in solving the anomalous-scatterer partial structure. Finally, there is a discussion of selenomethionine as a phasing vehicle. More comprehensive reviews of MAD have been published by W. A. Hendrickson, who pioneered its development and application in macromolecular crystallography (Hendrickson, 1991; Hendrickson & Ogata, 1997).

Theoretical Basis

Electrons bound in atomic orbitals have specific resonant frequencies corresponding to allowed transitions. Anomalous scattering is the manifestation in X-ray diffraction of these resonance effects. The resonant frequencies of most chemical elements in biological macromolecules are far below the energies used for diffraction experiments, and their anomalous scattering is thus negligible. However, elements of atomic number 24 through 92 have resonant frequencies between 6 keV ($\lambda = 2$Å) and 40 keV ($\lambda = 0.3$Å), which give rise to detectable effects in X-ray scattering from macromolecular specimens labeled with these elements. Information about the phase of the scattered X-rays can be derived from the resonance effects, or anomalous scattering. Anomalous scattering is an atomic property and thus enters the equations for X-ray diffraction in the expression for the atomic scattering factor (f), which is the sum of "normal" atomic scattering factor f^0 and a complex "anomalous" correction having real (f') and imaginary (f") components:

$$f = f^0 + f' + if".$$

The breakdown of Friedel's law caused by the imaginary component of anomalous scattering (f") has been used for many years as a source of phase information in macromolecular crystallography. Wavelength-tunable synchrotron radiation allows the real

component (f') to be used as well, providing the opportunity for direct phasing through combination of the orthogonal effects of f' and f". MAD exploits differences in the observed diffraction intensities caused by differential f' and f" values at different X-ray wavelengths to achieve such direct phasing.

The formulation of the MAD observational equation used here is based on that of Karle (1980) as modified by Hendrickson *et al.* (1985).

$$|^{\lambda}\mathbf{F}_{obs}{}^{\pm}|^2 = |\mathbf{F}_T|^2 + a_\lambda|\mathbf{F}_A|^2 + b_\lambda|\mathbf{F}_T||\mathbf{F}_A|\cos(\phi_T-\phi_A) \pm c_\lambda|\mathbf{F}_T||\mathbf{F}_A|\sin(\phi_T-\phi_A), \quad [1]$$

where $a_\lambda = (f''_\lambda{}^2+f'_\lambda{}^2)/(f^0)^2$,

$b_\lambda = 2f'_\lambda/f^0$

and $c_\lambda = 2f''_\lambda/f^0$.

This formulation is distinguished from many others relating phases to anomalous scattering by Karle's insight that the real ($|^{\lambda}\mathbf{F}_A'|$) and imaginary ($|^{\lambda}\mathbf{F}_A''|$) structure amplitudes, due to f' and f", respectively, can be expressed as products of scattering factor ratios and normal structure amplitudes, due to f^0:

$$|^{\lambda}\mathbf{F}_A'| = (f'_\lambda/f^0)|\mathbf{F}_A|$$

and $$|^{\lambda}\mathbf{F}_A''| = (f''_\lambda/f^0)|\mathbf{F}_A|.$$

Wavelength-dependence and structure-dependence are thus separated into different quantities. All wavelength dependence is in the anomalous scattering factors, f'_λ and f''_λ, which do not depend on atomic positions, and all structure dependence is in the normal structure factors $\mathbf{F}_T$ and $\mathbf{F}_A$, which do not depend on wavelength. The structure factor $\mathbf{F}_T$ represents normal scattering from the total structure, and $\mathbf{F}_A$ represents normal scattering from the partial structure of anomalous scatterers. An Argand diagram showing the relationships of these structure factors has been published (Smith, 1991). Eq. 1 describes the case for one type of anomalous scatterer. In general, Eq. 1 will relate experimental observations to unknown quantities whose number equals twice the number of anomalous-scatterer types plus one, here $|\mathbf{F}_T|$, $|\mathbf{F}_A|$ and $(\phi_T-\phi_A)$ for one anomalous-scatterer type.

The MAD observational equation (Eq. 1) involves no approximations, and the accuracy of MAD phases is limited only by the precision of the diffraction data. This is in contrast to isomorphous replacement where phase accuracy is limited most severely by breakdown of the assumption of isomorphism of native and derivative crystals. The new prominence of MAD is due primarily to a significant improvement in the quality of diffraction data in general. This comes from the ability to measure better data faster thanks to widespread adoption of cryocooling techniques and to improvements in synchrotron sources and X-ray detectors.

Anomalous scattering factors

Anomalous scattering factors in the region of an absorption edge are sensitive to the chemical environment of the absorbing atom, and are significantly enhanced by sharp spectral features in many cases. Therefore, f" and f' for anomalous scatterers in macromolecules cannot be calculated as free-atom anomalous scattering factors (Cromer & Liberman, 1970a,1970b), which are accurate estimates for all chemistries at energies away from absorption edges. Several laboratories have schemes for extracting anomalous scattering factors f' and f" from X-ray spectra, none of which has been published in rigorous detail. However, all exploit the fact that the imaginary component of anomalous scattering f" is proportional to the atomic absorption coefficient μ_a, which can be obtained easily from raw X-ray fluorescence or transmission data. The scheme of Hendrickson *et al.* (1988) is described briefly here and illustrated in Fig. 1. The X-ray spectrum of the labeled macromolecule, typically a macromolecule single crystal, is measured as fluorescence through the edge of interest (Fig. 1a). Regions of the experimental spectrum slightly away from the edge are fit to theoretical values using the program XASFIT in order to place the experimental spectrum on an absolute scale (Fig. 1b). Theoretical values are obtained from a program by Don Cromer, modified by Wayne Hendrickson to produce spectra rather than f' and f" at single energies and variously called FPRIME, SPECTRUM or CROMER. Care must be taken to measure enough edge-remote points for reliable fit of the experimental spectrum, which may be quite noisy. A narrow region around the absorption edge is then cut from the scaled experimental spectrum and spliced into the theoretical spectrum. From the hybrid spectrum of f" values thus obtained, f' values are calculated by Kramers-Kronig transformation:

$$f'(E) = \frac{2\delta}{\pi} \sum_{i=0}^{\infty} \frac{E_i f_i''}{E^2 - E_i^2} \qquad [2]$$

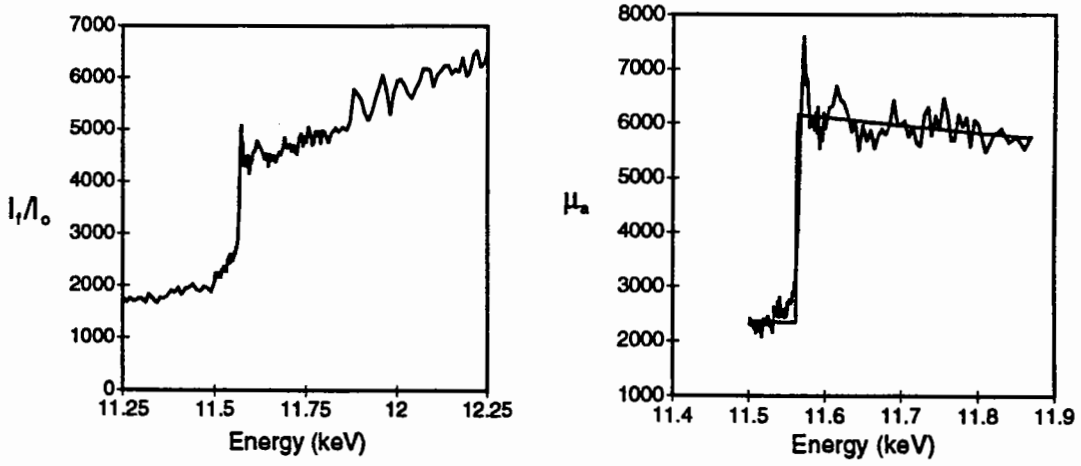
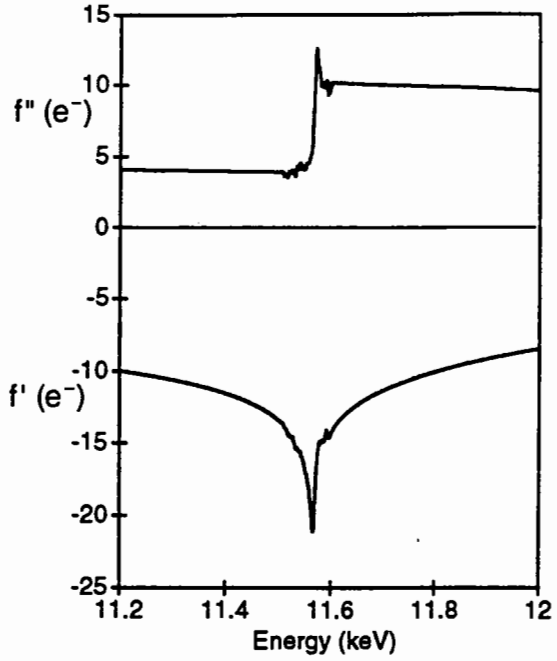


Figure 1

A. Fluorescence spectrum (I/I_0 on an arbitrary scale) through the Pt L_{III} absorption edge from a single crystal of β-hydroxydecanoyl thiolester dehydrase (Leesong *et al.*, 1996). A single methionine amino acid of the crystalline protein was labeled with Pt by soaking in a solution of K_2PtCl_4.

B. Scaling of fluorescence data to theoretical atomic absorption coefficients (μ_a). The raw fluorescence spectrum was fit to the theoretical spectrum for the Pt L_{III} edge using the program XASFIT. The scaled experimental spectrum is shown superimposed on the theoretical free-atom spectrum.



C. Hybrid f" and f' spectra for the Pt L_{III} edge. Using the program KRAMIG, the edge region has been cut from the experimental spectrum in B and spliced into the theoretical spectrum, μ_a converted to f", and f' calculated from f" by Kramers-Kronig transformation (Eq. 2).

where E is energy in eV and δ is the energy increment of the f" spectrum being transformed. In practice, the point of singularity for each f' ($E_i = E$) is not included in the summation, and a transformation range of ~500 eV beyond the f' being computed is sufficient to eliminate truncation effects. Splicing and f' calculation (Fig. 1c) are done with the program KRAMIG.

Typical anomalous scattering factors, f''_{max} and f'_{min}, estimated from X-ray spectra of protein crystals taken at MAD experimental stations, are given in Table 1 for several elements. In addition to the electronic environment of the anomalous scatterer, the energy dispersion of the incident X-ray beam also influences the values of anomalous scattering factors in the edge region.

Table 1. Typical anomalous scattering factors

Element	f^0 (e^-)	Edge	λ (Å)	f'_{min} (e^-)	λ (Å)	f''_{max} (e^-)	Reference
Fe	26	K	1.7402 1.7425	-9 -8	1.7380 1.7390	5 4	Hendrickson *et al.*, 1988 Smith *et al.*, 1994
Cu	29	K	1.3790	-8	1.3771	4	Guss *et al.*, 1988
Zn	30	K	1.2826	-9	1.2818	4	Zhang *et al.*, 1995
Se	34	K	0.9793	-11	0.9792	6	Wu *et al.*, 1994
Br	35	K	0.9207	-7	0.9196	4	Ogata *et al.*, 1989
Sm	62	L_{II}	1.6959	-16	1.6952	17	Tomchick *et al.*, 1996
Ho	67	L_{III}	1.5363	-28	1.5356	20	Weis *et al.*, 1991
Yb	70	L_{III}	1.3857	-33	1.3853	35	Shapiro *et al.*, 1995
W	74	L_{III}	1.2136	-24	1.2123	19	Egloff *et al.*, 1995
Os	76	L_{III}	1.1402	-23	1.1397	20	Cate *et al.*, 1996
Pt	78	L_{III}	1.0720	-21	1.0714	13	Fig. 1c
Hg	80	L_{III}	1.0094 1.0095	-18 -25	1.0057 1.0063	10 12	Tesmer *et al.*, 1994 Krishna *et al.*, 1994
U	92	L_{III}	0.7213	-21	0.7208	12	Glover *et al.*, 1995

Energy (keV) = 12.39854/λ (Å)

Estimation of the Magnitude of the MAD signal

Knowledge of anomalous scattering factors allows estimation of the MAD signal for a specific anomalous scatterer in a specific macromolecule. The orthogonal components of the phasing signal, due to the real and imaginary anomalous scattering factors f' and f", are estimated separately because both are required for phase determination. The maximum MAD Bijvoet signal is due to Bijvoet differences at the energy of peak absorption, or f''_{max}, and is proportional to $2f''_{max}$ of Table 1. The maximum MAD dispersive signal is due to wavelength differences between structure amplitudes at the energy of the inflection point of the edge (f'_{min}) and at a remote energy(f'_{remote}), and is proportional to $|f'_{min}-f'_{remote}|$.

The magnitude of the MAD phasing signal is estimated as the ratio of expected Bijvoet or dispersive difference to expected total scattering of the macromolecule. This is based on calculation of expected structure amplitudes $\langle|F|\rangle$, where $\langle|F|\rangle = \sqrt{\sum f_i^2}$ and $\langle|F|\rangle = \sqrt{N}f$ for N atoms of identical f (Crick & Magdoff, 1956). The diffraction ratios of interest to MAD (Hendrickson, 1985) are, for the dispersive signal,

$$\frac{\left\langle\left||F_{\lambda 1}|-|F_{\lambda 2}|\right|\right\rangle}{\langle|F_T|\rangle} \approx \sqrt{\frac{N}{2}}\frac{|f'_{\lambda 1}-f'_{\lambda 2}|}{\langle|F_T|\rangle} \qquad [3]$$

for N anomalous-scatterer sites with λ1 chosen at f'_{min} and λ2 chosen for $|f'_{\lambda 1}-f'_{\lambda 2}|_{max}$, and, for the Bijvoet signal,

$$\frac{\left\langle\left||F^+_\lambda|-|F^-_\lambda|\right|\right\rangle}{\langle|F_T|\rangle} \approx \sqrt{\frac{N}{2}}\frac{2f''_\lambda}{\langle|F_T|\rangle} \qquad [4]$$

with λ chosen at f''_{max}. These diffraction ratios are analogous to the usual calculation of isomorphous signal from experimental data in which

$$\frac{\left\langle\left||F_{PH}|-|F_P|\right|\right\rangle}{\langle|F_P|\rangle} \approx \sqrt{\frac{N}{2}}\frac{f^0}{\langle|F_P|\rangle} \qquad [5]$$

where f^0 is for the heavy atom. Values for f^0, f'_{min} and f''_{max} are those in Table 1. The denominator of all diffraction ratios is the expected total scattering of the macromolecule, which can be estimated for $2\theta = 0$ with the expressions in Table 2.

Table 2. Estimates of scattering strength for macromolecules, $\langle|F_T|\rangle$

Macromolecule	NA = # atoms (e^-)	NR = # residues (e^-)	MW = molecular weight (e^-)
Protein	$6.70\ (NA)^{1/2}$	$(346\ NR)^{1/2}$	$(3.14\ MW)^{1/2}$
DNA	$7.20\ (NA)^{1/2}$	$(1128\ NR)^{1/2}$	$(3.87\ MW)^{1/2}$
RNA	$7.26\ (NA)^{1/2}$	$(1183\ NR)^{1/2}$	$(3.89\ MW)^{1/2}$

A hypothetical example illustrates the issue of signal size in MAD *vs.* isomorphous replacement. Consider a 500-residue protein and the MAD signal generated by 10 Se anomalous scatterers. If $f''_{max} = 6\ e^-$, $f'_{min} = -11\ e^-$ and $f'_{remote} = -4\ e^-$, then by Eq. 4 the maximum Bijvoet signal will be ~6% of $|F_{obs}|$ and by Eq. 3 the maximum dispersive signal will be ~4% of $|F_{obs}|$. By comparison, the isomorphous replacement signal generated by one fully occupied Hg site ($f^0 = 80\ e^-$) in the same protein will be ~14% of $|F_{obs}|$ by Eq. 5. For many typical examples the MAD signal is near the noise level of moderate-quality diffraction data sets, whereas the isomorphous replacement signal is easily detectable in data of moderate quality. On the other hand, detection of the MAD signal is limited only by data quality whereas lack of isomorphism will pollute the isomorphous replacement signal with systematic error that cannot be removed. It is clear from the large number of successful MAD experiments that a relatively weak phasing signal is by no means an insurmountable problem.

MAD experimental design

Three important considerations distinguish the design and execution of a MAD experiment from more familiar monochromatic experiments in macromolecular crystallography. These are wavelength selection, data completeness and data quality. A discussion of the design of beamline components for MAD experiments is presented in another paper in this volume by A. W. Thompson.

The largest MAD phasing signal is obtained at energies with the most extreme values of f' and f", which correspond to the sharpest features of the absorption edge. Therefore, it is critical to determine the position of the absorption edge experimentally from the labeled macromolecule at the time of a MAD experiment. Even when the position of the edge is well known, small unanticipated chemical changes in the sample or energy changes in the X-ray beam can reduce the MAD signal very significantly if the sharp edge features are missed in selecting energies for data collection. Energies are selected at the peak of sample absorption just above the edge ("E_{peak}" for f''_{max}) to optimize the Bijvoet signal and at the inflection point of the edge ("E_{dip}" for f'_{min}) to optimize the orthogonal dispersive signal. The dispersive signal is further optimized if a third energy remote from the edge ("E_{remote}") is chosen. The choice of E_{remote} is experiment dependent, although it is typically above rather than below the edge due to the larger Bijvoet signal. E_{remote} may also be chosen to avoid complications from other edges or to obtain data at a wavelength optimal for model refinement.

There has been much debate about the optimal number of data-collection energies for successful phase determination by MAD. In the commonest MAD experiment $|F^+|$ and $|F^-|$ are measured at each of E_{dip}, E_{peak} and E_{remote}. If the difference in f' is large enough to produce a detectable signal, then one could in principle obtain phases from three measurements: $|F^+|$ and $|F^-|$ at E_{peak} and either $|F^+|$ or $|F^-|$ at E_{dip} (Peterson *et al.*, 1996). However, redundancy is one of the best ways to minimize the effects of measurement error in macromolecular crystallography. In the full three-energy experiment, the Bijvoet signal is redundant because the remote energy is above the edge. The orthogonal dispersive signal is redundant because two measurements are taken at each of E_{dip} and E_{remote}. There are several examples of even more redundant four- or five-wavelength MAD experiments. While greater redundancy is desirable, it should not be gained at the

cost of good counting statistics. Unfortunately, considerations of available beam time frequently preclude MAD experiments with more than three energies.

The MAD phasing signal is derived from intensity differences that may be similar in magnitude to measurement errors. Thus a general philosophy in the design of a MAD experiment is to equalize systematic errors among the measurements whose differences will contribute to each phase determination. This is achieved for each single reflection by recording Bijvoet measurements at all wavelengths from the same asymmetric unit of the same crystal at nearly the same time. Bijvoet mates can be recorded simultaneously by alignment of the crystal with a mirror plane perpendicular to the rotation axis, or Friedel images can be recorded in an "inverse beam" experiment. (Friedel images are related by 180° rotation of the crystal about any axis perpendicular to the incident beam, usually the data-collection axis). If crystal decay is a problem, small blocks of Bijvoet data can be recorded at each of the selected wavelengths before moving to another block of reciprocal space. When such a data collection strategy is followed, the resulting MAD data set will be complete with respect to recording all multiwavelength, Bijvoet measurements for all regions of the reciprocal lattice that are covered in the experiment. Coverage of reciprocal space can be monitored during the experiment by a strategy program, if available, or by reduction of diffraction images to integrated intensities for data from at least one wavelength. Completeness of the MAD data set is at least as important as for any diffraction experiment that will be used for phasing. If data, and hence phase information, are incomplete, it may be difficult to reproduce the same beam and sample conditions during a subsequent experiment, which is likely to occur only after some weeks or months.

Measurement errors are of major importance in all areas of macromolecular crystallography, but are the limiting factor in phase determination by MAD. MAD data should be of high quality by the usual measures (R_{sym}, redundancy, completeness), especially in experiments where the phasing signal is weak. In the hypothetical 500-residue protein with 10 Se anomalous scatterers, a 5% MAD signal will become undetectable as it is exceeded by R_{sym} "noise". Thus data with good counting statistics are of paramount importance. In a carefully designed experiment, the effect of increasing R_{sym} with increasing θ is mitigated somewhat by equalizing systematic errors. Nevertheless, if R_{sym} (I) is 30% for the outer shells of data, there will be virtually no detectable MAD phasing signal for these reflections in the hypothetical example. Disappearance of the phasing signal into R_{sym} noise is the major reason that useful MAD phases generally are not obtained to the diffraction limit of crystals even though anomalous scattering does not fall off with increasing θ.

Data processing and scaling

Concerns about signal size dominate special schemes for handling MAD data. Scaling strategies for MAD are discussed in detail elsewhere in this volume by P. R. Evans. Special computer programs for scaling MAD data have been developed (Hendrickson *et al.*, 1988; Friedman *et al.*, 1994). Two general approaches to data handling for MAD have been employed.

The approach originally proposed by Hendrickson, known as "phase first, merge later," represents the extreme interpretation of the scheme for equalizing systematic errors – the individual observations constituting a multiwavelength Bijvoet set, as determined by the data-collection strategy, are grouped together and scaled as usual, but are merged with redundant measurements only after phases are determined. Error estimates from the phasing or the agreement of redundant phase determinations can be incorporated into weights for averaging, or can be used to reject outliers. This approach involves complicated, experiment-dependent bookkeeping to assemble exactly the correct observations for each unmerged set.

A second approach, "merge first, phase later," is to scale and merge data at each wavelength, keeping Bijvoet pairs separate, and then to scale data at all wavelengths to one another. This is most easily and reliably done by scaling all data against a common standard data set, which can be the unique data from one wavelength with Bijvoet mates averaged. If the data collection followed one of the strategies outlined above, then measurements for each unique reflection are identically redundant, which itself minimizes systematic errors in the amplitude differences used for phasing. The second approach is computationally simpler than the first because it is experiment independent. However, unanticipated, minor experimental disasters may be more difficult to overcome in the "merge first, phase later" approach to data handling.

Approaches to MAD phasing

There are two general approaches to MAD phasing. One is to treat the problem explicitly and solve the MAD observational equation (Eq. 1). This explicit approach is embodied in the MADSYS package from the Hendrickson laboratory (Wu & Hendrickson, 1996), in particular in the phasing program MADLSQ. The other approach is to treat MAD phasing as a special case of multiple isomorphous replacement (MIR). The pseudo-MIR approach is discussed elsewhere in this volume by V. Biou and in two recent publications (Ramakrishnan & Biou, 1997; Terwilliger, 1997). Both approaches have been quite successful, and there are no hard-and-fast rules for which sorts of problems are more amenable to which approach, rumors in the community notwithstanding. There are advantages and disadvantages to both approaches.

The explicit approach provides the quantities $|F_T|$, $|F_A|$ and $(\phi_T-\phi_A)$. Estimates of the anomalous scattering factors at the wavelengths of data collection are required to fit the observations to the MAD phase equation. These estimates can be refined within MADLSQ, so they need not be highly accurate. A second calculation is required to obtain ϕ_T from the phase differences $(\phi_T-\phi_A)$. There are two advantages to the explicit approach. First, it is amenable to the "phase first, merge later" scheme of data handling because refinement of the anomalous-scatterer partial structure is entirely separate from phase calculation. In this case redundancies are merged to produce a unique data set at the level of the derived quantities $|F_T|$, $|F_A|$, $(\phi_T-\phi_A)$ and their error estimates. These error estimates or the agreement of redundant phase determinations can be used to weight terms in a Fourier synthesis from $|F_T|$ and ϕ_T. Phase probability coefficients (ABCDs) have been derived from the MAD phase equation (Pähler *et al.*, 1990). The second principle advantage of the explicit approach is calculation of an experimentally derived estimate of the normal structure amplitude $|F_A|$ for the anomalous scatterer. This is the quantity with

which the partial structure of anomalous scatterers is most directly solved and refined, and therefore should be highly sought. However, while MADLSQ is quite successful in the least-squares fit of the MAD phase equation to $|\mathbf{F}_{obs}|$ for high-quality data, it is poorly conditioned to extracting $|\mathbf{F}_A|$ from noisy data and requires careful pruning of outliers from the $|\mathbf{F}_A|$ values produced. A Bayesian method of $|\mathbf{F}_A|$ estimation (Terwilliger, 1994) should be more robust than the least-squares procedure.

In the pseudo-MIR approach data at one wavelength are designated as "native" data, which include anomalous scattering, and data at the other wavelengths as "derivative" data. This approach has the advantage that nothing need be known about the anomalous scattering factors prior to phasing. These quantities are incorporated into heavy-atom atomic "occupancies" and refined along with other parameters. Of course, the partial structure of anomalous scatterers must be known, and refinement of the partial structure is concurrent with phasing. In refinement of the "heavy atom" parameters, greater weight is given to the data set selected as "native." This bias should be removed by the new maximum-likelihood refinement of de La Fortelle and Bricogne (1997), which treats data at all wavelengths as statistically equivalent. The amplitudes $|\mathbf{F}_A|$ are not determined in the pseudo-MIR approach, and the partial structure is solved from Bijvoet differences between $|\mathbf{F}^+|$ and $|\mathbf{F}^-|$ or dispersive differences between $|\mathbf{F}_{\lambda 1}|$ and $|\mathbf{F}_{\lambda 2}|$, with wavelengths selected to optimize the signal. The pseudo-MIR approach is used more frequently than the explicit approach due to the greater familiarity of crystallographers with software for isomorphous replacement.

Determination of the anomalous-scatterer partial structure

A prerequisite for MAD-phased electron density, regardless of the phasing technique, is determination of the partial structure of anomalous scatterers. As described above, the optimal quantities for solving and refining the partial structure of anomalous scatterers are the normal scattering amplitudes $|\mathbf{F}_A|$. Frequently $|\mathbf{F}_A|$ values are not extracted from the MAD measurements, and the largest Bijvoet or dispersive differences are used instead. This involves the usual approximation of representing structure amplitudes ($|\mathbf{F}_A|$) as the subset of larger differences ($||\mathbf{F}^+|-|\mathbf{F}^-||$ or $||\mathbf{F}_{\lambda 1}|-|\mathbf{F}_{\lambda 2}||$). The approximation is accurate for only a small fraction of reflections because there is no correlation between the phase of $\mathbf{F}_P$ and the phase of $\mathbf{F}_A$. However, it suffices for a suitably strong signal and a suitably small number of sites. For virtually all structures determined by MAD, the anomalous-scatterer sites have been located by Patterson methods. However, the problem quickly becomes intractable by Patterson methods when there are more than a handful of sites. This is a current challenge for MAD, where the aim is to solve the macromolecule structure from one MAD data set using any number of anomalous scatterer sites. Statistical direct methods clearly hold the answer to this problem. Recent results are promising in this regard. Bertrand *et al.* (1997) have solved a 12-atom Se partial structure in a 437-residue protein by direct methods using $|\mathbf{F}_A|$s, and S. Doublié (personal communication) has solved a 15-atom Se partial structure in an asymmetric unit of 108 kDa using dispersive differences, also by direct methods. These results open the door for routine MAD determination of quite large structures with many anomalous scatterer sites. New direct methods techniques, such as described in this volume in papers by G. M. Sheldrick, by C. M. Weeks and by G. Bricogne, hold great promise for a major expansion in the complexity of anomalous-scatterer partial structure that can be solved.

The correct enantiomorph for the anomalous-scatterer partial structure must be determined (ϕ_A *vs.* $-\phi_A$) in order to obtain an electron-density image of the macromolecule. However, it cannot be determined directly from MAD data. The correct hand is chosen by comparison of electron density maps based on both enantiomorphs of the partial structure. Unlike the situation for MIR, the density based on the incorrect hand of the anomalous-scatterer partial structure is not the mirror image of that based on the correct hand and contains no image of the macromolecule. The correct map is distinguished by features such as a clear solvent boundary, positive correlation of redundant densities, and a macromolecule-like density histogram. If the anomalous scattering centers form a centric array, then the two enantiomorphs are identical and both maps are correct.

Selenomethionine

The most successful MAD phasing vehicle to date has been selenium in the form of selenomethionine (SeMet). This particularly clever experiment was devised by Wayne Hendrickson (1985), who also pioneered its use (Yang *et al.*, 1990; Hendrickson *et al.*, 1990). Briefly, proteins are labeled with Se by biological substitution of SeMet for methionine amino acids. This is achieved by blocking methionine biosynthesis in the cells in which the protein is produced and substitution of SeMet for Met in the growth medium. The generality of the labeling scheme for proteins is the root of its success. SeMet labeling technology is discussed in a recent review by Doublié (1997).

SeMet incorporation has been done most frequently for proteins expressed in *E. coli* strains that are auxotrophic for Met (strain DL41, Hendrickson *et al.*, 1990; strain B834, Leahy *et al.*, 1994 and Doherty *et al.*, 1995; strain LE392, Ceska *et al.*, 1996; strain MIC88, Shamoo *et al.*, 1995). Nearly complete incorporation has also been reported in nonauxotrophic bacterial strains (*E. coli* strain BL21, Harrison *et al.*, 1994; *E. coli* strain XA90, Labahn *et al.*, 1996), in a mammalian cell line (Lustbader *et al.*, 1995) and in baculovirus-infected insect cells (Chen & Bahl, 1991). Special precautions must be taken to prevent oxidation of SeMet proteins. In almost all cases, somewhat higher-than-normal concentrations of disulfide reducing agents, such as dithiothreitol or β-mercaptoethanol, are sufficient to protect SeMet from air oxidation to the selenoxide (Brot *et al.*, 1984). In a few cases, crystallization in an inert atmosphere has been necessary (Dyda *et al.*, 1994; Wu *et al.*, 1994). Because Se is a light element, the position of the K absorption edge moves to slightly higher energy upon oxidation, and a mixture of oxidation states in a sample crystal is predicted to obliterate the MAD signal.

Methionine is a particularly attractive target for anomalous scatterer labeling. The hydrophobic side chain of methionine, which carries the sulfur atom to be substituted by selenium, is usually buried in the hydrophobic core of globular proteins and is therefore relatively better ordered than are surface side chains. Evidence for isostructuralism of Met and SeMet proteins comes from the labeling experiment itself. All proteins in the biological expression system have SeMet substituted for Met at levels approaching 100%. The cells are viable, therefore the proteins are functional and isostructural with their unlabelled counterparts to the extent required by function.

The natural abundance of methionine in soluble proteins is approximately one in fifty amino acid residues. The N-terminal Met is not included in this estimate because, if present, it is usually disordered. Using Eqs. 3 and 4, this provides a maximal MAD phasing signal of 4-6% of |F|, easily detectable in strongly diffracting protein crystals and detectable with careful data collection from crystals of moderate quality. To improve the phasing signal, in a few cases Met has been substituted for other amino acids by site-directed mutagenesis (Leahy *et al.*, 1994, 1996; Skinner *et al.*, 1994; Tong *et al.*, 1996).

SeMet labeling is now part of the repertoire of protein crystallography, and has broader applicability than for MAD phasing alone. This comes from the relative ease of incorporation of the SeMet label, from the remarkable structural similarity of SeMet and wild type proteins, and from the uniformity and completeness of labeling. Crystals of SeMet proteins are usually isomorphous with those of the wild type, and consequently can be used as isomorphous derivatives. The isomorphous signal comes from the excess of 18 electrons in Se relative to S, making the SeMet isomorphous phasing signal (~10% of |F|, Eq. 5) about twice as strong as the SeMet MAD phasing signal (4-6% of |F|). In most cases SeMet derivatives are more isomorphous, and certainly more rationally produced, than are heavy-atom derivatives produced by the usual soaking procedures. Prior knowledge of exactly how Se labels the protein is itself a powerful tool. For example, the SeMet mutation is an extremely useful amino acid label for fitting a protein sequence to electron density. Also, noncrystallographic symmetry operators usually can be defined more reliably from Se positions in SeMet protein than by heavy-atom positions in conventional derivatives due to the uniformity and completeness of labeling (Tesmer *et al.*, 1996).

An analogous label is available for nucleic acids in the form of brominated bases, particularly 5-bromouridine, which is isostructural with thymidine. Iodinated bases are commonly used as isomorphous derivatives ($f^0 = 53\ e^-$) for nucleic acids, but the X-ray edges of I ($\lambda = 0.38$Å for K, $\lambda = 2.56$-2.72Å for L) occur at energies less favorable for accurate macromolecular data collection than does the K edge of Br ($\lambda = 0.92$Å).

Conclusion

Why is the enthusiasm for MAD so high today? There are three primary reasons. First, cryocrystallography has improved data quality to the point that the precision required for MAD is usual rather than exceptional. Second, new synchrotron sources and new beamlines provide intense, reliably tunable X-ray beams and the instruments to exploit them. Third, MAD works extremely well and very quickly. For many problems, the experimentally phased electron density is of stellar quality. Crystallographers are only beginning to appreciate the value of nearly error-free, model-independent phases (Burling *et al.*, 1996). The remaining challenges are in two areas. The greatest impediment to growth of MAD today is access to suitable experimental facilities. This non-technical problem may be solved only be a concerted effort of the community. The greatest technical challenge is to develop methods for solving large partial structures of anomalous scatterers. Here recent results with statistical direct methods are very promising, and MAD applied to large macromolecules no longer seems such a heroic undertaking. MAD has at last taken its place as a standard tool of macromolecular crystallography.

Acknowledgment

Work in the author's laboratory has been supported by grants from the U.S. National Institutes of Health (DK42303), and from the Lucille P. Markey Foundation to the Structural Studies Group at Purdue University. Collaboration with the scientific staffs at synchrotron facilities is gratefully acknowledged, especially A. W. Thompson of the European Synchrotron Radiation Facility, and S. E. Ealick of the Cornell High Energy Synchrotron Source.

References

Bertrand, J., Auger, G., Fanchon, E., Martin, L., Blanot, D., van Heijenoort, J. and Dideberg, O. (1997) Crystal structure of UDP-N-acetylmuramoyl-L-alanine:D-glutamate ligase from *Escherichia coli*. *EMBO J.* **16**: in press.

Brot, N., Fliss, H., Coleman, T. and Weissbach, H. (1984) Enzymatic reduction of methionine sulfoxide residues in proteins and peptides. *Meth. Enzymol.* **107**: 352-360.

Burling, F. T., Weis, W. I., Flaherty, K. M. and Brünger, A. T. (1996) Direct observation of protein solvation and discrete disorder with experimental crystallographic phases. *Science* **271:** 72-77.

Cate, J. H., Gooding, A. R., Podell, E., Zhou, K., Golden, B. L., Kundrot, C. E., Cech, T. R. and Doudna, J. A. (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* **273:** 1678-1685.

Ceska, T. A., Sayers, J. R., Stier, G. and Suck, D. (1996) A helical arch allowing single-stranded DNA to thread through T5 5'-exonuclease. *Nature* **382:** 90-93.

Chen, W. and Bahl, O. P. (1991). Recombinant carbohydrate and selenomethionyl variants of human choriogonadotropin. *J. Biol. Chem.* **266:** 8192-8197.

Crick, F. H. C. and Magdoff, B. S. (1956) The theory of the method of isomorphous replacement for protein crystals. I. *Acta Cryst.* **9:** 901-908.

Cromer, D. T. and Liberman, D. (1970a) Relativistic calculation of anomalous scattering factors for X-rays. *J. Chem. Phys.* **53**: 1891-1898.

Cromer, D. T. and Liberman, D. (1970b) Relativistic calculation of anomalous scattering factors for X-rays. *Los Alamos National Laboratory Report LA-4403.*

de La Fortelle, E. and Bricogne, G. (1997) Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Meth. Enzymol.* **276**: 472-494.

Doherty, A. J., Ashford, S. R., Brannigan, J. A. and Wigley, D. B. (1995). A superior host strain for the over-expression of cloned genes using the T7 promoter based vectors. *Nucleic Acids Res.* **23:** 2074-2075.

Doublié, S. (1997) Preparation of selenomethionyl proteins for phase determination. *Meth. Enzymol.* **276:** 523-530.

Dyda, F., Hickman, A. B., Jenkins, T. M., Engelman, A., Craigie, R. and Davies, D. R. (1994) Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science* **266:** 1981-1986.

Egloff, M.-P., Cohen, P. T. W., Reinemer, P. and Barford, D. (1995) Crystal structure of the catalytic subunit of human protein phosphatase 1 and its complex with tungstate. *J. Mol. Biol.* **254:** 942-959.

Friedman, A. M., Fischmann, T. O., Shamoo, Y. and Ealick, S. (1994) MADPRB: a new suite of programs for MAD data analysis incorporating robust estimation, maximum likelihood and Bayesian inference. *Abstracts of the Amer. Crystallogr. Assn. Series 2* **22**: 39.

Glover, I. D., Denny, R. C., Nguti, N. D., McSweeney, S. M., Kinder, S. H., Thompson, A. W., Dodson, E. J., Wilkinson, A. J. and Tame, J. R. H. (1995) Structure determination of OppA at 2.3 Å resolution using multiple-wavelength anomalous dispersion methods. *Acta Cryst.* **D51:** 39-47.

Guss, J. M., Merritt, E. A., Phizackerley, R. P., Hedman, B., Murata, M., Hodgson, K. O. and Freeman, H. C. (1988) Phase determination by multiple- wavelength X-ray diffraction: crystal structure of a basic "blue" copper protein from cucumbers. *Science* **241:** 806-811.

Harrison, C. J., Bohm, A. A. and Nelson, H. C. M. (1994). Crystal structure of the DNA binding domain of the heat shock transcription factor. *Science* **263:** 224-227.

Hendrickson, W. A. (1985) Analysis of protein structure from diffraction measurement at multiple wavelengths. *Trans. Amer. Crystallogr. Assn.* **21:** 11-21.

Hendrickson, W. A. (1991) Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254:** 51-58.

Hendrickson, W. A. and Ogata, C. M. (1997) Phase determination from multiwavelength anomalous diffraction measurements. *Meth. Enzymol.* **276**: 494-523.

Hendrickson, W. A., Smith, J. L. and Sheriff, S. (1985) Direct phase determination based on anomalous scattering. *Meth. Enzymol.* **115**: 41-55.

Hendrickson, W. A., Smith, J. L., Phizackerley, R. P. and Merritt, E. A. (1988) Crystallographic structure analysis of lamprey hemoglobin from anomalous dispersion of synchrotron radiation. *Proteins: Struct., Funct., Genet.* **4**: 77-88.

Hendrickson, W. A., Horton, J. R. and Lemaster, D. M. (1990). Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *EMBO J.* **9:** 1665-1672.

Karle, J. (1980) Some developments in anomalous dispersion for the structural investigation of macromolecular systems in biology. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **7:** 357-367.

Krishna, T. S. R., Kong, X.-P., Gary, S., Burgers, P. M. and Kuriyan, J. (1994) Crystal structure of the eukaryotic DNA polymerase processivity factor PCNA. *Cell* **79:** 1233-1243.

Labahn, J., Schärer, O. D., Long, A., Ezaz-Nikpay, K., Verdine, G. L. and Ellenberger, T. E. (1996) Structural basis for the excision repair of alkylation-damaged DNA. *Cell* **86:** 321-329.

Leahy, D. J., Erickson, H. P., Aukhil, I., Joshi, P. and Hendrickson, W. A. (1994). Crystallization of a fragment of human fibronectin: introduction of methionine by site-directed mutagenesis to allow phasing via selenomethionine. *Proteins: Structure, Function, and Genetics* **19:** 48-54.

Leahy, D. J., Aukhil, I. and Erickson, H. P. (1996) 2.0 Å crystal structure of a four-domain segment of human fibronectin encompassing the RGD loop and synergy region. *Cell* **84:** 155-164.

Leesong, M., Henderson, B. S., Gillig, J. R., Schwab, J. M. and Smith, J. L. (1996) Structure of a dehydratase-isomerase from the bacterial pathway for biosynthesis of unsaturated fatty acids: two catalytic activities in one active site. *Structure* **4:** 253-264.

Lustbader, J. W., Wu, H., Birken, S., Pollak, S., Kolks-Gawinowicz, M. A., Pound, A. M., Austen, D., Hendrickson, W. A. and Canfield, R. E. (1995). The expression, characterization and crystallization of wild-type and selenomethionyl human chorionic gonadotropin. *Endocrinology* **136:** 640-650.

Ogata, C. M., Hendrickson, W. A., Gao, X. and Patel, D. J. (1989) *Abstracts of the Amer. Cryst. Assoc., Ser. 2* **17**: 53.

Pähler, A., Smith, J. L. and Hendrickson, W. A. (1990) A probability representation for phase information from multiwavelength anomalous dispersion. *Acta Crystallogr. sect. A* **46:** 537-540.

Peterson, M. R., Harrop, S. J., McSweeney, S. M., Leonard, G. A., Thompson, A. W., Hunter, W. N. and Helliwell, J. R. (1996) MAD phasing strategies explored with a brominated oligonucleotide crystal at 1.65 Å resolution. *J. Synchrotron Rad.* **3:** 24-34.

Ramakrishnan, V. and Biou, V. (1997) Treatment of multiwavelength anomalous diffraction data as a special case of multiple isomorphous replacement. *Meth. Enzymol.* **276**: 538-557.

Shamoo, Y., Friedman, A. M., Parsons, M. R., Konigsberg, W. H. and Steitz, T. A. (1995) Crystal structure of a replication fork single-stranded DNA binding protein (T4 gp32) complexed to DNA. *Nature* **376:** 362-366.

Shapiro, L., Fannon, A. M., Kwong, P. D., Thompson, A., Lehmann, M. S., Grübel, G., Legrand, J.-F., Als-Nielsen, J., Colman, D. R. and Hendrickson, W. A. (1995) Structural basis of cell-cell adhesion by cadherins. *Nature* **374:** 327-337.

Skinner, M. M., Zhang, H., Leschnitzer, D. H., Guan, Y., Bellamy, H., Sweet, R. M., Gray, C. W., Konings, R. N., Wang, A. H. and Terwilliger, T. C. (1994) Structure of the gene V protein of bacteriophage f1 determined by multiwavelength x-ray diffraction on the selenomethionyl protein. *Proc. Natl. Acad. Sci. U.S.A.* **91:** 2071-2075.

Smith, J. L. (1991) Determination of three-dimensional structure by multi-wavelength anomalous diffraction. *Curr. Opinion Struc. Biol.* **1:** 1002-1011.

Smith, J. L., Zaluzec, E. J., Wery, J.-P., Niu, L., Switzer, R. L., Zalkin, H. and Satow, Y. (1994) Structure of the allosteric regulatory enzyme of purine biosynthesis. *Science* **264:** 1427-1433.

Terwilliger, T. C. (1994) MAD phasing: Bayesian estimates of F_A. *Acta Cryst.* **D50:** 11-16.

Terwilliger, T. C. (1997) Multiwavelength anomalous diffraction phasing of macromolecular structures: analysis of MAD data as single isomorphous replacement with anomalous scattering data using the MADMRG program. *Meth. Enzymol.* **276**: 530-537.

Tesmer, J. J. G., Stemmler, T. L., Penner-Hahn, J. E., Davisson, V. J. and Smith, J. L. (1994) Preliminary X-ray analysis of *Escherichia coli* GMP synthetase: Determination of anomalous scattering factors for a cysteinyl mercury derivative. *Proteins: Structure, Function and Genetics* **18:** 394-403.

Tesmer, J. J. G., Klem, T. J., Deras, M. L., Davisson, V. J. and Smith, J. L. (1996) The crystal structure of GMP synthetase reveals a novel catalytic triad and is a structural paradigm for two enzyme families. *Nature Structural Biology* **3**: 74-86.

Tomchick, D. R., Smith, J. L., Turner, R. J. and Switzer, R. L. (1996) PyrR, a bifunctional RNA-binding transcriptional attenuation protein and uracil phosphoribosyltransferase. *Acta Cryst.* **S52:** C-163.

Tong, L., Qian, C., Massariol, M.-J., Bonneau, P. R., Cordingley, M. G. and Lagacé, L. (1996) A new serine-protease fold revealed by the crystal structure of human cytomegalovirus protease. *Nature* **383:** 272-275.

Weis, W. I., Kahn, R., Fourme, R., Drickamer, K. and Hendrickson, W. A. (1991) Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing. *Science* **254:** 1608-1615.

Wu, H. and Hendrickson, W. A. (1996) The analytical approach of phasing by multiwavelength anomalous diffraction (MAD). *Acta Cryst.* **S52**: C-55.

Wu, H., Lustbader, J. W., Liu, Y., Canfield, R. E. and Hendrickson, W. A. (1994) Structure of human chorionic gonadotropin at 2.6 Å resolution from MAD analysis of the selenomethionyl protein. *Structure* **2:** 545-558.

Yang, W., Hendrickson, W. A., Crouch, R. J. and Satow, Y. (1990) Structure of ribonuclease H phased at 2 Å resolution by MAD analysis of the selenomethionyl protein. *Science* **249**: 1398-1405.

Zhang, G., Kazanietz, M. G., Blumberg, P. M. and Hurley, J. H. (1995) Crystal structure of the Cys2 activator-binding domain of protein kinase Cδ in complex with phorbol ester. *Cell* **81:** 917-924.

phasing MAD data using MIR programs

by
Valérie Biou
Laboratoire de Cristallographie Macromoléculaire, Institut de Biologie Structurale
41 avenue des Martyrs
F-38027 Grenoble cedex, France
and European Synchrotron Radiation Facility, BP 220, F-38043 Grenoble cedex France.
e-mail biou@ibs.fr

Introduction

Many structures have been solved using MAD data during the last few years, and their number is increasing exponentially. The aim of this paper is to give a practical approach to MAD, and in particular to the use of MIR programs to phase MAD data, and to discuss the limitations and advantages of the method.

In the presence of anomalously scattering atoms in the protein crystal, one can use two types of signal to calculate phases from a diffraction data set : (i) dispersive difference signal : due to the contribution of F'a to the structure factor, the intensity of a given reflection changes with the wavelength. (ii) anomalous signal : the intensity of symmetry related reflections is different due to the contribution of F"a (fig 1).

These signals can be used in a multiple wavelength dispersion (MAD) experiment with tuneable synchrotron radiation, so that both the dispersive and anomalous differences are maximised. This takes at least 3 wavelengths, which we shall define as follows : $\lambda 1$ is measured at the minimum of f', i.e., the inflection point of the fluorescence spectrum ; $\lambda 2$ is taken at the maximum of f" (and of the fluorescence spectrum) ; $\lambda 3$ is taken on the high energy side of the spectrum. Thus, that $\lambda 1$ and $\lambda 3$ maximise the dispersive difference signal, and $\lambda 2$ maximises the anomalous signal. A fourth wavelength, remote on the low energy side of the edge, can also be useful.

The advantages and disadvantages of MAD have been explained elsewhere (see for example Reid, 1996). Briefly, it is obvious that one overcomes anisomorphism problems between native and derivative by using MAD. One can collect three data sets on a single, flash frozen crystal containing an appropriate element. On the other hand, the anomalous signal is generally much less intense than the isomorphous signal for the same element. Just consider the example of the replacement of sulphur by selenium in selenomethionine. The K edge of selenium contributes 10 electrons at the maximum dispersive difference, whereas it gives 18 electrons isomorphous signal compared to sulphur. Even for such a light atom as selenium, the isomorphous difference will be roughly twice as large as the dispersive difference. In the case of mercury, the difference between the anomalous and the isomorphous contributions is even larger.

Therefore, the problem is to measure small differences between large figures. This has been said before, but it should be stressed : it is vital for a MAD experiment to get accurate measurements. Synchrotron beamlines have been developed that allow to do this in a shorter and

shorter time, and in the next few months there should be less shortage of beam time for MAD (see A.W. Thompson's paper in this issue).

Data collection and its preliminaries

feasibility assessment before going to the synchrotron

In order to properly plan an experiment, it is important to **evaluate the theoretical signal** one can expect to obtain from a given heavy atom derivative: these are the dispersive ratio, and the anomalous ratio, which give the proportion of the maximum anomalous or dispersive signals vs. the total scattering power of the macromolecule. Dispersive ratio = $q \times |f'_{\lambda 1} - f'_{\lambda 3}|$ Anomalous ratio = $q \times 2f''_{\lambda 2}$ where $q = \sqrt{\frac{Na}{2Np}} \times \frac{1}{Z_{eff}}$. Na = number of anomalously diffracting atoms in the unit cell, Np = number of protein atoms, and $Z_{eff} = 6.7$ electrons for a protein crystal (mean effective normal scattering on protein atoms), λ1, λ2 and λ3 are defined as in the introduction. In practice, a signal of 2.5% with very good data may be enough for phasing. 3.5 to 4% gives a good signal.

It is just as essential to have good knowledge of your crystals : mosaicity, resolution, diffracting power. Too high a mosaicity will make the data harder to integrate, and reduce the signal to noise ratio. MAD structures have been solved with mosaic crystals (up to 1° as defined in DENZO), but 0.4° or less gives better signal. If the crystal diffracts to high resolution, it is worth spending more time to collect high resolution at three wavelengths, to get accurate experimental phases at higher resolution. This can be achieved if the crystal diffracts strongly : the anomalous signal does not decay with resolution, but if the spot intensities become too low, the measurements will be more noisy, hindering the extraction of the anomalous signal.

data collection practice

It is essential to measure a **fluorescence spectrum** on your crystal (or 2 with 2 perpendicular crystal orientations). The absorption edge can shift due to anisotropy of the heavy atom chemical environment. This will determine the strength and position of the fluorescence spectrum and will allow you to decide at which wavelengths to collect. In case of a beam reinjection during the course of the measurements, it is wise to collect a fluorescence spectrum again.

The second step is to collect one image to determine the crystal orientation. From this, one can run a **data collection strategy** program in order to plan how much data needs to be collected. We routinely use Andrew Leslie's STRATEGY option in MOSFLM (Leslie, 1996). From a given crystal orientation, it gives the most convenient rotation range to run and predicts the expected completeness, both for individual reflections and for Bijvoet mates. If the crystal can be oriented so that it rotates around a mirror axis, it is better to do so, as it allows to collect Bijvoet mates in the same image. In the case where it is necessary to collect data from an additional crystal, the program gives the best rotation range to complete the datasets. Once you have set up the

strategy and the best exposure time, start the actual data collection, and measure 3 wavelengths, four if possible.

Finally, it is important to integrate and scale data carefully. A first run can be done on the first wavelength, while it is being collected. It will give information about the data quality and the anomalous signal to be expected from the whole data set. Several integration and scaling runs are usually necessary in order to get the best out of the data set (see P.R. Evans's contribution in this issue).

Phasing methods

Both phasing systems imply the **location of heavy atoms positions** in the unit cell. This can be done using Patterson maps or direct methods. Three types of Patterson maps can be used : dispersive difference Pattersons between two wavelengths, or anomalous difference Patterson for one wavelength, or a Patterson map calculated using the Fa's derived from the algebraic method (see below). This last method seems to be the one that gives the least noisy Pattersons, because systematic errors have been removed before. Similarly, the same types of differences can be used in direct methods to solve the heavy atom structure when the number of heavy atoms is too high (Bertrand *et al.*, 1997). This is probably going to be common practice in the near future, as it will allow to phase larger and larger structures with MAD. Starting from the location of heavy atoms, the next step is then to refine those and calculate phases. Two types of methods are available for phasing MAD data.

Algebraic method

The first method used to phase a novel structure using MAD data (Guss *et al.*, 1988), is based on the algebraic derivation of phases using a set of linear equations (Karle, 1980). This method allows to derive accurate values for the heavy atom structure factors (Fa), and gives an elegant solution of the phase problem. However, though it is being made more user friendly (Wu and Hendrickson, 1996), it has long been difficult to use, in particular because it required a careful bookkeeping for equivalent reflections. It works on unmerged, scaled individual reflections.

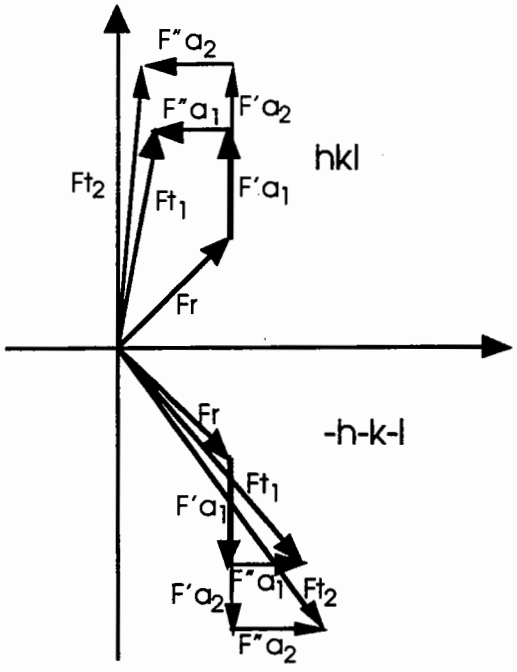


Figure 1 Vectorial representation of structure factors in the presence of anomalous scatterers. Subscripts 1 and 2 refer to two different wavelengths. Ft = total structure factor for reflection hkl.; Fr = contribution from the non anomalously scattering atoms;
F'a = contribution from the real part of anomalously scattering atoms;
F"a = contribution from the imaginary part of anomalously scattering atoms; Ft=Fr+F'a+iF"a.

MIR-like

From fig 1, it is visible that the different wavelengths can be considered as different heavy atom derivatives, and that multiple isomorphous replacement phasing methods should be usable in this context. Ramakrishnan *et al.* (1993) were the first to use an MIR program to solve a new structure using MAD data. Last year, about half the structures solved using MAD data were phased using an MIR program. It is more familiar to most protein crystallographers, and it allows to easily bring together all sorts of phasing information. A number of different programs can be used to do this, the most popular being probably MLPHARE (Otwinowski, 1991).

All of those programs refine the heavy atom positions and temperature factors, and refine phases against the lack of closure error. Most of the programs available (see Table I and Ramakrishnan and Biou (1997)) rely on a reference wavelength data set as the "native", and use the dispersive differences between this reference wavelength and the others, as well as the anomalous differences for all data. The differences lie in the statistical description of the phase and amplitude spaces. MLPHARE and the maximum likelihood option of PHASES use a maximum likelihood description of the phase space, thereby implying that most of the error comes from the phases and not from the amplitudes. On the other hand, SHARP uses a maximum likelihood description of the whole complex space, both amplitudes and phases. For a better description, see Eric de la Fortelle's paper in this issue. X-PLOR also offers a MAD phasing option (Burling *et al.*, 1996).

program	author	distribution	usage	principle
MLPHARE (Otwinowski, 1991)	Z.Ottwinovski	ccp4 suite , Daresbury	1 reflection file, 1 list of atomic scattering factors	choose one wavelength as "native" ; refines heavy atom parameters (different occupancy for real and anomalous parts), based on **maximum likelihood on the phase circle.**
PHASIT (Furey and Swaminathan, 1997)	W. Furey	phases suite, author	several reflection files ; atomic scattering factors are entered as parameters	choose one wavelength as "native" ; refines heavy atom parameters against origin-removed patterson, or using **maximum likelihood,** similarly to mlphare.

MADMRG + HEAVY (Terwillinger, 1994b; Terwillinger, 1994a)	T. Terwillinger	author	*madmrg* merges all MAD reflections into a "SIRAS"-like data set. *heavy* refines heavy atom parameters and calculates phases.	choose one wavelength as "native" ; refines heavy atom parameters against origin-removed patterson; one single occupancy.
SHARP (de la Fortelle and Bricogne, 1997)	E. de la Fortelle, G. Bricogne	author	http interface with user friendly data input. One reflection file.	no reference wavelength ; refines heavy atom parameters using anisotropic B factors and **maximum likelihood in the whole complex space.**
X-PLOR V 3.8.5 (Burling *et al.*, 1996)	A. Brunger	x-plor package, Yale university	distributed template macros, merged reflection file	still under development. choose one wavelength as "native"

Table I Some of the programs which can be used for both MIR and MAD phasing.

Table II gives a list of some structures solved using MAD data. This represents about a half of all structures solved this way. Besides the exponential increase with time, several striking points can be derived from this table. The molecular weights are increasing with time. Selenium from selenomethionine is by far the most used anomalous scatterer. Iron and mercury are next. This reflects the ease of introduction or the natural occurrence of those three elements in protein crystals. There is also a tendency towards measuring MAD data to higher resolution, rather than getting medium resolution phases and extending them with a native data set. The last column shows it is common use to mix MAD and MIR, and that about half of the recent year structures have been phased using an MIR program.

A number of practical points have been addressed in Ramakrishnan and Biou (1997). I would like to go back to one point which seems to be difficult to grasp in the beginning, namely the parallel use of f' and f" values and heavy atom occupancies. The structure factor for reflection h in the presence of anomalously scattering atoms of the same sort, can be written as the sum of a normal, F_h^n and a wavelength-dependent anomalous, ${}^{\lambda}F_h^a$ structure factors : $F_h = F_h^n + {}^{\lambda}F_h^a$ with ${}^{\lambda}F_h^a = \sum_j o_j \left({}^{\lambda}f'_j + i {}^{\lambda}f''_j \right) \exp(2\pi i h.r_j) = \sum_j \left({}^{\lambda r}o_j {}^{\lambda}f'_j + {}^{\lambda a}o_j i {}^{\lambda}f''_j \right) \exp(2\pi i h.r_j)$, where o_j is the occupancy of atom j, and ${}^{\lambda r}o_j$ and ${}^{\lambda a}o_j$ are the real and anomalous occupancies, respectively. If one sets both f' and f" to an arbitrary value, the refinement of anomalous and dispersive occupancy factors will adjust the relative values of ${}^{\lambda}F_h^a$. Thus, it does not make a difference whether one inputs reasonable values for f' and f", or if one inputs fake ones and lets the program refine occupancies. However, I feel more comfortable with inputting reasonable values of the anomalous scattering factors, because one gets occupancy values which "make sense" : in this case, they should be the same for a given heavy atom position throughout the data sets and then reflect the physical occupancy of the site. In the other case, the occupancy will vary according to the values of Δf' or f", and it should do so in a similar way for all sites at a given wavelength. Therefore, the anomalous occupancy should be highest at the maximum f" value, and

the dispersive occupancy should be highest for the difference between the minimum f' and the remote wavelength.

Quality criteria for phasing evaluation

It is important to include as much data as possible in the phasing process. The following criteria can be used to keep or select data : $R_{Cullis} = \frac{\sum_{centrics} \left\| F_{PH}(obs) \right| \pm \left| F_P(obs) \right\| - \left| F_H(calc) \right|}{\sum_{centrics} \left\| F_{PH}(obs) \right| \pm \left| F_P(obs) \right\|}$;

$R_{Kraut} = \frac{\sum_{acentrics} \left\| F_{PH}(obs) \right| - \left| F_{PH}(calc) \right\|}{\sum_{acentrics} \left\| F_{PH}(obs) \right\|}$ (isomorphous case)

$R_{Kraut} = \frac{\sum_{acentrics} \left\| F_{PH+}(obs) \right| - \left| F_{PH+}(calc) \right\| + \left\| F_{PH-}(obs) \right| - \left| F_{PH-}(calc) \right\|}{\sum_{acentrics} \left\| F_{PH+}(obs) \right| + \left| F_{PH-}(obs) \right\|}$ (anomalous case). R-Kraut should be as low as possible, and R-Cullis should ideally be close to 0.5, and "typical" values are between 0.8 and 0.6.

The figure of merit is the weighted mean of the cosine of the phase angle deviation from Φ_{best}. It is calculated as $m = \frac{\left| F_{hkl_{best}} \right|}{\left| F_{hkl} \right|}$ with $F_{hkl_{best}} = \frac{\sum_{\phi} P(\phi) F_{hkl}(\phi)}{\sum_{\phi} P(\phi)}$. The phasing power is defined as $\left[\frac{\sum_n \left| F_H \right|^2}{\sum_n \left| E \right|^2} \right]$ with $\sum_n \left| E \right|^2 = \sum_n \left(\left| F_{PH}(obs) \right| - \left| F_{PH}(calc) \right| \right)^2$ = rms lack of closure error. Both figure of merit and phasing power are plotted as a function of resolution, and a given data set should ideally be cut-off at a resolution where its phasing power drops below 1.

When MAD is not enough : how to incorporate everything you can in order to calculate phases

When the MAD phases are not sufficient to give an interpretable map, it is straightforward to introduce other phasing information. A "native" data set must be defined for all programs, except SHARP. All other data sets should be scaled with respect to this native. Derivatives should be screened for phasing power in order to keep only the useful data. Annex I shows an input and excerpts of an output file from PHASES, illustrating the introduction of a native, a mercury MAD data set and a single wavelength selenomethionine.

when is SAD enough

Several structures have been solved using a single heavy atom derivative anomalous signal (e.g., Biou *et al.*, 1995), and Eric de la Fortelle showed that SHARP was quite able to solve structures this way. It takes cases where a single heavy atom derivative (Pb in the mentioned case) gave a strong anomalous signal. The phase ambiguity can then be resolved using solvent flattening alone or solvent flattening and non crystallographic symmetry when applicable. It is of course more difficult and more risky, but it may work when one has no other choice.

Summary: MAD works if ...

You measure, process and scale data carefully on as good crystals as you can.

You try and minimize mosaic spread (work hard on cryoprotectants, use smaller crystals).

All modern phasing methods work, it is more important to use one you're familiar with, or you can get help with.

Then you can have an excellent experimental map to trace your chain automatically, and excellent phases to refine your model against.

I apologise to all of the authors whose structures were omitted from the list in Table 1. For lack of space I could not possibly include all of the relevant references.

pdb entry - protein (a)	reference	asymm. unit content	heavy atom (b)	res. (c)	data used - phasing method (d)
1CBP - blue copper protein	(Guss *et al.*, 1988)	10 kDa	Cu 1	2.5Å	MAD 4l- madsys
? - streptavidin	(Hendrickson *et al.*, 1989)	126 aa	Se 2	3.1Å	MAD 3l- madsys
1RNH - RNase H	(Yang *et al.*, 1990)	156 aa	Se 4 (6, 13, 37, 36 / 16)	2.2Å (2.0)	MAD 3l- madsys
1MSB - lectin domain from rat mannose-binding protein	(Weis *et al.*, 1991)	110 aa	Ho 4	2.5Å	MAD 3l- madsys
1TEN - fibronectin type III domain	(Leahy *et al.*, 1992)	91 aa	Se 1 (53, 39 / 21)	3Å (1.8)	MAD 4l- madsys
1ITH - homotetrameric hemoglobin	(Kolatkar *et al.*, 1992)	2x141 aa	Fe 1	5Å (2.5)	MAD 4l + MIR - madsys
1HST - histone H5 globular domain	(Ramakrishnan *et al.*, 1993)	2x90 aa	Se 2 (14, 15 / 21)	2.6Å	MAD 3l - mlphare
1HCN - HCG	(Wu *et al.*, 1994)	200 aa	Se 4 (61, 55, 56, 80 / 42)	2.6Å	MAD 4l- madsys
1BGH - gene V protein	(Skinner *et al.*, 1994)	87 aa	Se 1 (37/ 21) & 2	2.5Å	MAD 3l - heavy
1IRK - insulin receptor tyr kinase domain	(Hubbard *et al.*, 1994)	306 aa	Hg 2	2.5Å (2.1)	MAD 3l - madsys
1GPH - PRPP purine synthase	(Smith *et al.*, 1994)	4x350 aa	Fe 4	5 then 3Å	MAD 3l - madsys
1OLA - OppA	(Glover *et al.*, 1995)	58.8 kDa	U 8	2.3Å	MAD 4l- mlphare
1CNT - ciliary neutrophic factor	(McDonald *et al.*, 1995)	185 aa	Yb 1	2.4Å	MAD 4l- madsys
? - protein phosphatase 1	(Egloff *et al.*, 1995)		W + Hg	2.5Å	MAD 3l + MIR + 2-fold NCS- phases
1ASU - avian sarcoma virus integrase	(Bujacz *et al.*, 1995)	155 aa	Se 4 (23, 46, 41, 16 / 33)	2.2Å (1.7)	MAD 3l- phases
1TIG - IF3 C-terminal domain	(Biou *et al.*, 1995)	94 aa	Se 2 (40, 22 /20)	2 Å	MAD 3l - phases
1GEO* - sulfite reductase	(Crane *et al.*, 1995)	456 aa	Fe 5	2.5Å (1.6)	MAD 3l + MIRAS - madsys
1VHH - sonic hedgehog N-terminal domain	(Tanaka Hall *et al.*, 1995)	200 aa	Se 3 (19, 43, 47/11)	1.7Å	MAD 4l - madlsq
1IDO - integrin CR3 A domain	(Lee *et al.*, 1995)	192 aa	Se 3 (17, 17, 8/15)	2Å (1.7)	MAD 3l - mlphare
1SVC - NFkB p50 homodimer with DNA	(Müller *et al.*, 1995)	364 aa + 19 bp	Se 5 (98, 58, 49, 59, 66/ 70)+ I	3.4Å (2.6)	MAD 3l + MIR + crystal averaging - mlphare + madlsq
1NCG - cadherin	(Shapiro *et al.*, 1995)	110 aa	Yb 1	2.1Å	MAD 4l - madlsq
? - mannose-binding protein	(Burling *et al.*, 1996)	230 aa	Yb 1	1.8Å	MAD 4l - xplor
1RIE - rieske Fe-S protein fragment	(Iwata *et al.*, 1996)	120 aa	Fe 2	2.8Å (1.5)	MAD 3l - mlphare
1TBG* - G protein ßγ dimer	(Sondek *et al.*, 1996)	4x139	Gd 6	2.8Å (2.1)	MAD 3l - mlphare
1FBT* - fructose-2,6-biphosphatase	(Lee *et al.*, 1996)	220 aa	Se 4	2.8Å (2.5)	MAD 4l - mlphare
1GSS - glutathione S-transferase	(Reinemer *et al.*, 1996)	2x211 aa	Se 4 (16, 22, 28, 22 / 26) + I	3Å (2.2)	MAD 2l + MIR + 2-fold NCS- mlphare

? - TFIIA/ TBP/ DNA complex	(Geiger *et al.*, 1996)	300 aa + 18 bpDNA	Se / Br 5	3Å	MAD 5l + MR - mlphare
1WHI - ribosomal protein L14	(Davies *et al.*, 1996)	124 aa	Se 2 (32, 21 / 14)	2 Å (1.5)	MAD 3l + MIR - phases
1DKX - DnaK chaperone + peptide	(Zhu *et al.*, 1996)	218 + 7aa	Se 6	2.3Å	MAD 4l - madsys
1UMU - UmuD' protein	(Peat *et al.*, 1996)	2x116 aa	Se 4 (26, 48, 25, 31 / 24)	2.5Å	MAD 4l - madsys + multan
1TEN - fibronectin type III repeat	(Leahy *et al.*, 1996)	90 aa	Se 1 (53 /)	1.8Å	MAD 4l - madsys
1ZEN - class II aldolase	(Cooper *et al.*, 1996)	39 kDa	Se 6 (15, 33, 26, 31, 44, 23/ 36)	2.5Å	MAD 3l + MIR - mlphare

Table II Non exhaustive list of MAD structures to date.
(a) Pdb entry code followed by *: coordinates release still pending at time of writing. When replaced with ? : entry not found in pdb; (b) heavy atom : type, number and temperature factors ($Å^2$) of the corresponding SD or SE atoms in the released pdb entry for selenomethionine protein, followed with the mean overall temperature factor. (c) second figure between parentheses gives resolution used for refinement when different from the MAD experiment resolution.
(d) References for phasing programs : Heavy (Terwillinger, 1994a &b), Mlphare (Otwinowski, 1991), Madsys (Hendrickson *et al.*, 1988; Hendrickson, 1991), Phases (Furey and Swaminathan, 1997), Xplor version 3.8x (Burling *et al.*, 1996).

References

Bertrand, J.A., Auger, G., Fanchon, E., Martin, L., Blanot, D., van Heijenoort, J. and Dideberg, O. (1997) crystal structure of UDP-*N*-acetylmuramoyl-L-alanine:D-glutamata ligase from *Escherichia coli*. *EMBO J.*, (In Press)

Biou, V., Shu, F. and Ramakrishnan, V. (1995) X-ray crystallography shows that translational initiation factor IF3 consists of two compact alpha/beta domains linked by an alpha-helix. *EMBO J.*, **14**, 4056-4064.

Bujacz, G., Jaskolski, M., Alexandratos, J., Wlodawer, A., Merkel, G., Katz, R.A. and Skalka, A.M. (1995) High-resolution structure of the catalytic domain of avian sarcoma virus integrase. *J.Mol.Biol.*, **253**, 333-346.

Burling, F.T., Weis, W.I., Flaherty, K.M. and Brunger, A.T. (1996) Direct observation of protein solvation and discrete disorder with experimental crystallographic phases. *Science*, **271**, 72-77.

Cooper, S.J., Leonard, G.A., McSweeny, S.M., Thompson, A.M., Naismith, J.H., Qamar, S., Plater, A., Berry, A. and Hunter, W.N. (1996) The crystal structure of a class II fructose-1,6-biphosphate aldolase shows a novel binuclear metal-binding active site embedded in a familiar fold. *Structure*, **4**, 1303-1315.

Crane, B.R., Siegel, L.M. and Getzoff, E.D. (1995) Sulfite reductase structure at 1.6 A: evolution and catalysis for reduction of inorganic anions. *Science*, **270**, 59-67.

Cusack, S. , 1996. (UnPub)

Davies, C., White, S.W. and Ramakrishnan, V. (1996) The crystal structure of ribosomal protein L14 revels an important organisational component of the translational apparatus. *Structure*, **4**, 55-65.

de la Fortelle, E. and Bricogne, G. (1997) Maximum likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multivavelength anomalous diffraction methods. In Carter, C.W. and Sweet, R.M. (ed.)*Methods in Enzymology vol 276*, Academic Press, Orlando, Fl: pp. 472-494.

Egloff, M.P., Cohen, P.T., Reinemer, P. and Barford, D. (1995) Crystal structure of the catalytic subunit of human protein phosphatase 1 and its complex with tungstate. *J.Mol.Biol.*, **254**, 942-959.

Furey, W. and Swaminathan, S. (1997) Phases-95 : a program package for the processing and analysis of diffraction data from macromolecules. In Carter, C. and Sweet, R.M. (ed.)*Methods in Enzymology*, Academic Press, Orlando, Fl:

Geiger, J.H., Hahn, S., Lee, S. and Sigler, P.B. (1996) Crystal structure of the yeast TFIIA/TBP/DNA complex . *Science*, **272**, 830-836.

Glover, I.D., Denny, R.C., Nguti, N.D., McSweeny, S.M., Kinder, S.H., Thompson, A.M., Dodson, E.J., Wilkinson, A.J. and Tame, J.R. (1995) Structure determination of OppA at 2.3Å resolution using multiple-wavelength anomalous dispersion methods. *Acta Cryst.*, **D51**, 39-47.

Guss, J.M., Merritt, E.A., Phizackerley, R.P., Hedman, B., Murata, M., Hodgson, K.O. and Freeman, H.C. (1988) Phase determination by Multiple wavelength X-ray diffraction : crystal structure of a basic "blue" copper protein from cucumbers. *Science*, **241**, 806-811.

Hendrickson, W.A., Pähler, A., Smith, J.L., Satow, Y., Merritt, E.A. and Phizackerley, R.P. (1989) Crystal structure of core streptavidin determined from multiwavelength anomalous diffraction of synchrotron radiation. *Proc.Natl.Acad.Sci.U.S.A.*, **86**, 2190-2194.

Hendrickson, W.A. (1991) Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science*, **254**, 51-58.

Hendrickson, W.A.H., Smith, J.L., Phizackerley, R.P. and Merritt, E.A. (1988) Crystallographic structure analysis of lamprey hemoglobin from anomalous dispersion of synchrotron radiation. *Proteins*, **4**, 77.

Hubbard, S.R., Wei, L., Ellis, L. and Hendrickson, W.A. (1994) Crystal structure of the tyrosine kinase domain of the human insulin receptor . *Nature*, **372**, 746-754.

Iwata, S., Saynovits, M., Link, T.A. and Michel, H. (1996) Structure of a water soluble fragment of the 'Rieske' iron-sulfur protein of the bovine heart mitochondrial cytochrome bc1 complex determined by MAD phasing at 1.5Å resolution. *Structure*, **4**, 5678-579.

Karle, J. (1980) Some developments in anomalous dispersion for the structural investigation of macromolecular systems in biology. *Int.J.Quant.Chem.*, **7**, 357-367.

Kolatkar, P.R., Ernst, S.R., Hackert, M.L., Ogata, C.M., Hendrickson, W.A., Merritt, E.A. and Phizackerley, R.P. (1992) Structure determination and refinement of homotetrameric hemoglobin from Urechis caupo at 2.5 A resolution. *Acta Crystallogr.B*, **48**, 191-199.

Leahy, D.J., Hendrickson, W.A., Aukhil, I. and Erickson, H.P. (1992) Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science*, **258**, 987-991.

Leahy, D.J., Hendrickson, W.A., Aukhil, I. and Erickson, H.P. (1996) Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science*, **258**, 987-991.

Lee, J.O., Rieu, P., Arnaout, M.A. and Liddington, R. (1995) Crystal structure of the A domain from the alpha subunit of integrin CR3 (CD11b/CD18). *Cell*, **80**, 631-638.

Lee, Y.H., Ogata, C., Pflugrath, J.W., Levitt, D.G., Sarma, R., Banaszak, L.J. and Pilkis, S.J. (1996) Crystal structure of the rat liver fructose-2,6-bisphosphatase based on selenomethionine multiwavelength anomalous dispersion phases. *Biochemistry*, **35**, 6010-6019.

Leslie, A.G.W. *Program ipmosflm version 5.4*, 1996. (UnPub)

McDonald, N.Q., Panayotatos, N. and Hendrickson, W.A. (1995) Crystal structure of dimeric human ciliary neurotrophic factor determined by MAD phasing. *EMBO J.*, **14**, 2689-2699.

Müller, C.W., Rey, F.A., Sodeka, M., Verdine, G.L. and Harrison, S.C. (1995) Structure of the NF-Kappa B P50 homodimer bound to DNA. *Nature*, **373**, 311-317.

Otwinowski, Z. (1991) . In Wolf, W., Evans, P.R. and Leslie, A.G.W. (ed.)*Isomorphous replacement and anomalous scattering*, Daresbury Laboratory, Warrington: pp. 80.

Peat, T.S., Frank, E.G., McDonald, J.P., Levine, A.S., Woodgate, R. and Hendrickson, W.A. (1996) structure of the UMUD' protein and its regulation in response to DNA damage . *Nature*, **380**, 727.

Ramakrishnan, V., Finch, J.T., Graziano, V., Lee, P.L. and Sweet, R.M. (1993) Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature*, **362**, 219-223.

Ramakrishnan, V. and Biou, V. (1997) Treatment of MAD as a special case of MIR. In Carter, C. and Sweet, R.M. (ed.)*Methods in Enzymology vol 276*, Academic Press, Orlando, Fl: pp. 538-557.

Reid, R.J. (1996) As MAD as can be. *Structure*, **4**, 11-14.

Reinemer, P., Prade, L., Hof, P., Neuefeind, T., Huber, R., Zettl, R., Palme, K., Schell, J., Koelln, I., Bartunik, H.D. and Bieseler, B. (1996) Three-dimensional structure of glutathione S-transferase from Arabidopsis thaliana at 2.2 A resolution: structural characterization of herbicide-conjugating plant glutathione S-transferases and a novel active site architecture. *J.Mol.Biol.*, **255**, 289-309.

Shapiro, L., Fannon, A.M., Kwong, P.D., Thompson, A.M., Lehman, M.S., Grubel, G., Legrand, J.-F., Als-Nielsen, J., Colman, D.R. and Hendrickson, W.A. (1995) structural basis of cell-cell adhesion by cadherins. *Nature*, **374**, 327.

Skinner, M.M., Zhang, H., Leschnitzer, D.H., Guan, Y., Bellamy, H., Sweet, R.M., Gray, C.W., Konings, R.N., Wang, A.H. and Terwilliger, T.C. (1994) Structure of the gene V protein of bacteriophage f1 determined by multiwavelength x-ray diffraction on the selenomethionyl protein. *Proc.Natl.Acad.Sci.U.S.A.*, **91**, 2071-2075.

Smith, J.L., Zaluzec, E.J., Wery, J.P., Niu, L., Switzer, R.L., Zalkin, H. and Satow, Y. (1994) Structure of the allosteric regulatory enzyme of purine biosynthesis. *Science*, **264**, 1427-1433.

Sondek, J., Bohm, A., Lambright, D.G., Hamm, H.E. and Sigler, P.B. (1996) Crystal structure of a GA protein beta gamma dimer at 2.1A resolution. *Nature*, **379**, 369-374.

Tanaka Hall, T.M., Porter, J.A., Beachy, P.A. and Leahy, D.J (1995) A potential catalytic site revealed by the 1.7Å crystal structure of the amino-terminal signalling domain of Sonic hedgehog. *Nature*, **378**, 212-216.

Terwillinger, T.C. (1994a) MAD phasing : treatment of dispersive differences as isomorphous replacement information. *Acta Crystallogr.D*, **D50**, 17-23.

Terwillinger, T.C. (1994b) MAD phasing : Bayesian estimates of FA. *Acta Crystallogr.D*, **D50**, 11-16.

Weis, W.I., Kahn, R., Fourme, R., Drickamer, K. and Hendrickson, W.A. (1991) Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing. *Science*, **254**, 1608-1615.

Wu, H., Lustbader, J.W., Liu, Y., Canfield, R.E. and Hendrickson, W.A. (1994) Structure of human chorionic gonadotropin at 2.6Å resolution from MAD analysis of the selenomethionyl protein. *Structure*, **2**, 545-558.

Wu, H. and Hendrickson, W.A. (1996) The analytical approach of phasing by multiwavelength anomalous dispersion. *IUCR abstracts*, C55.(Abstract)

Yang, W., Hendrickson, W.A., Crouch, R.J. and Satow, Y. (1990) Structure of ribonuclease H phased at 2 A resolution by MAD analysis of the selenomethionyl protein. *Science*, **249**, 1398-1405.

Zhu, X., Zhao, X., Burkholder, W.F., Gragerov, A., Ogata, C.M., Gottesman, M.E. and Hendrickson, W.A. (1996) Structural analysis of substrate binding by the molecular chaperone DnaK. *Science*, **272**, 1606-1614.

Annex I

Example for an input file to PHASES where the MAD data has been scaled to a native data set, and an additional mercury derivative collected elsewhere with a higher occupancy was also used.

```
hgmad.pam
 0
   29.564100   18.059999   12.837400
    1.211520    7.056390     .284738   20.748199
-14.40000
   29.564100   18.059999   12.837400
    1.211520    7.056390     .284738   20.748199
-23.00000
   29.564100   18.059999   12.837400
    1.211520    7.056390     .284738   20.748199
-10.87000
   29.564100   18.059999   12.837400
    1.211520    7.056390     .284738   20.748199
 -4.99000
  6  1
hgmad5.hkl
hgmad l1 anomalous
natl1_ano.hkl
 4.00 5.00 2  .9957    .0 2.9489    .5462E-01  -.1775E+03   .1063E+07   .5844E+06
  2
Hg   -.11140   -.18788   -.08980  20.00000   1.51119   21
Hg   -.36478   -.16421   -.52978  20.00000 1.20566   21
hgmad l2 anomalous
natl2_ano.hkl
 4.00 5.00 2 1.0027    .0 2.2748    .6032E-01  -.2386E+03   .1642E+07   .1047E+07
  2
Hg   -.11616   -.18952   -.09076  20.00000   1.43557   22
Hg   -.36362   -.16569   -.52917  20.00000   1.12660   22
madc l1 isomorphous
natl1_iso.hkl
 4.00 5.00 0 1.0046    .0 4.3930    .7547E-01  -.1094E+03   .1181E+07   .9751E+06
  2
Hg   -.11017   -.18803   -.09007  20.00000   1.23379   21
Hg   -.36553   -.16387   -.53047  20.00000   1.00647   21
madc l2 isomorphous
natl2_iso.hkl
 4.00 5.00 0 1.0000    .0 4.6351    .5751E-01   .1684E+03   .7555E+06   .1108E+07
  2
Hg   -.10891   -.18749   -.08981  20.00000   1.20896   22
Hg   -.36537   -.16366   -.53040  20.00000   1.00433   22
madc l3 isomorphous
natl3_iso.hkl
 4.00 5.00 0 1.0000    .0 4.0734    .5735E-01   .2727E+03   .5969E+06   .1344E+07
  2
Hg   -.10885   -.18771   -.08972  20.00000   1.22450   23
Hg   -.36536   -.16375   -.53036  20.00000   1.00473   23
madc hg hamburg isomorphous
nathgderiv_iso.hkl
 4.00 5.00 0 1.0000    .0 6.9056    .7391E-01  -.1897E+03   .7521E+07   .6809E+07
  2
Hg   -.11096   -.18758   -.08918  20.00000   1.18599   24
Hg   -.36646   -.16443   -.52977  20.00000    .96625
  2   .20 18  0  1  0
  1     SET
 0 0 0 1
 0 0 0 0
 0 0
  2     SET
 0 0 0 1
 0 0 0 0
 0 0
......etc.
```

Excerpts from the PHASIT log file from the above input file.
The breakdown of phasing power vs resolution is given only for one dataset.

```
            STATISTICS FOR SET   1 AFTER REFINEMENT
       R KRAUT  =    .045 FOR  12662 ACENTRIC REFLECTIONS
            STATISTICS FOR SET   2 AFTER REFINEMENT
       R KRAUT  =    .056 FOR  10920 ACENTRIC REFLECTIONS
            STATISTICS FOR SET   3 AFTER REFINEMENT
       R CULLIS =    .558 FOR    319  CENTRIC REFLECTIONS
       R KRAUT  =    .038 FOR   3764 ACENTRIC REFLECTIONS
            STATISTICS FOR SET   4 AFTER REFINEMENT
       R CULLIS =    .620 FOR    834  CENTRIC REFLECTIONS
       R KRAUT  =    .045 FOR   5958 ACENTRIC REFLECTIONS
            STATISTICS FOR SET   5 AFTER REFINEMENT
       R CULLIS =    .623 FOR    770  CENTRIC REFLECTIONS
       R KRAUT  =    .049 FOR   5793 ACENTRIC REFLECTIONS
            STATISTICS FOR SET   6 AFTER REFINEMENT
       R CULLIS =    .513 FOR    648  CENTRIC REFLECTIONS
       R KRAUT  =    .110 FOR   5315 ACENTRIC REFLECTIONS

--------------- START OF NEXT PHASING CYCLE ---------------

INDIVIDUAL DATA SET RESULTS BASED ON UPDATED HEAVY ATOM AND E VALUES

 SET   1     madhg l1 anomalous

        MEAN FIGURE OF MERIT =   .389 FOR   6331 REFLECTIONS

 SET   2     madhg l2 anomalous

        MEAN FIGURE OF MERIT =   .148 FOR   5460 REFLECTIONS

 SET   3     madc l1 isomorphous
MEAN FIGURE OF MERIT =   .508 FOR   4083 REFLECTIONS
MEAN FIGURE OF MERIT =   .733 FOR    319 CENTRIC REFLECTIONS
MEAN FIGURE OF MERIT =   .488 FOR   3764 ACENTRIC REFLECTIONS
   SET   4     madc l2 isomorphous
MEAN FIGURE OF MERIT =   .474 FOR   6792 REFLECTIONS
MEAN FIGURE OF MERIT =   .677 FOR    834 CENTRIC REFLECTIONS
MEAN FIGURE OF MERIT =   .446 FOR   5958 ACENTRIC REFLECTIONS
   SET   5     madc l3 isomorphous
MEAN FIGURE OF MERIT =   .468 FOR   6563 REFLECTIONS
MEAN FIGURE OF MERIT =   .654 FOR    770 CENTRIC REFLECTIONS
MEAN FIGURE OF MERIT =   .443 FOR   5793 ACENTRIC REFLECTIONS
SET   6     madc hg hamburg isomorphous
MEAN FIGURE OF MERIT =   .396 FOR   5963 REFLECTIONS
MEAN FIGURE OF MERIT =   .569 FOR    648 CENTRIC REFLECTIONS
MEAN FIGURE OF MERIT =   .375 FOR   5315 ACENTRIC REFLECTIONS

********** RESULTS FROM COMBINED PROBABILITY DISTRIBUTIONS **********
ACENTRIC REFLECTIONS INCLUDED IF   1 OR MORE DATA SETS CONTRIBUTED IN PHASE CALCULATION
MEAN FIGURE OF MERIT =   .716 FOR    7538 PHASED REFLECTIONS
MEAN PHASE SHIFT FROM PREVIOUS CYCLE =   1.22 DEGREES
MEAN FIGURES OF MERIT AS FUNCTION OF FP MAGNITUDE
MEAN FOM =    .585 MEAN FP =   1558.76 NUM REFL =   753
MEAN FOM =    .702 MEAN FP =   2395.57 NUM REFL =   753
MEAN FOM =    .747 MEAN FP =   3105.68 NUM REFL =   753
MEAN FOM =    .731 MEAN FP =   3784.38 NUM REFL =   753
MEAN FOM =    .752 MEAN FP =   4456.45 NUM REFL =   753
MEAN FOM =    .744 MEAN FP =   5162.72 NUM REFL =   753
MEAN FOM =    .735 MEAN FP =   6018.64 NUM REFL =   753
MEAN FOM =    .723 MEAN FP =   7012.22 NUM REFL =   753
MEAN FOM =    .732 MEAN FP =   8504.27 NUM REFL =   753
MEAN FOM =    .708 MEAN FP =  11671.32 NUM REFL =   753
MEAN FIGURES OF MERIT AS FUNCTION OF RESOLUTION

     MEAN FOM =    .723  MEAN D =      4.07 NUM REFL =   753
     MEAN FOM =    .694  MEAN D =      4.24 NUM REFL =   753
```

```
     MEAN FOM =     .699  MEAN D =      4.42 NUM REFL =   753
     MEAN FOM =     .705  MEAN D =      4.64 NUM REFL =   753
     MEAN FOM =     .705  MEAN D =      4.90 NUM REFL =   753
     MEAN FOM =     .718  MEAN D =      5.24 NUM REFL =   753
     MEAN FOM =     .716  MEAN D =      5.71 NUM REFL =   753
     MEAN FOM =     .714  MEAN D =      6.40 NUM REFL =   753
     MEAN FOM =     .750  MEAN D =      7.60 NUM REFL =   753
     MEAN FOM =     .734  MEAN D =     11.87 NUM REFL =   753


   PHASING POWER BREAKDOWN BASED ON CURRENT PROTEIN PHASES

SET   1     madhg l1 anomalous

  MEAN D =    8.63 PHASING POWER =     2.00 MEAN BIAS =  91.4 REFL=    633
  MEAN D =    6.20 PHASING POWER =     2.93 MEAN BIAS =  91.9 REFL=    633
  MEAN D =    5.52 PHASING POWER =     3.06 MEAN BIAS =  86.9 REFL=    633
  MEAN D =    5.13 PHASING POWER =     2.64 MEAN BIAS =  88.6 REFL=    633
  MEAN D =    4.85 PHASING POWER =     2.38 MEAN BIAS =  85.7 REFL=    633
  MEAN D =    4.63 PHASING POWER =     2.06 MEAN BIAS =  93.8 REFL=    633
  MEAN D =    4.45 PHASING POWER =     2.24 MEAN BIAS =  89.6 REFL=    633
  MEAN D =    4.30 PHASING POWER =     2.06 MEAN BIAS =  91.4 REFL=    633
  MEAN D =    4.17 PHASING POWER =     2.15 MEAN BIAS =  93.6 REFL=    633
  MEAN D =    4.05 PHASING POWER =     1.86 MEAN BIAS =  93.3 REFL=    633
  MEAN D =    4.00 PHASING POWER =      .98 MEAN BIAS =  62.0 REFL=      1


OVERALL  MEAN D=  5.19 PHASING POWER =   2.29  M.R.E. =  .73 MEAN BIAS =  90.6 REFL=  6331

   UPDATED E VALUES BASED ON NEW PROTEIN PHASES


 NRFL       <F>       RMS E      E FIT      DEL E

 316      1433.3    586459.1   953553.1  -367094.1
 316      1927.6   1238142.4   915128.3   323014.1
 316      2296.8    700552.8   905066.5  -204513.8
 316      2650.5    744915.5   910378.9  -165463.4
 316      2940.9    920513.4   925675.9    -5162.6
 316      3221.9   1935495.5   949859.1   985636.4
 316      3504.3    912075.1   983473.8   -71398.6
 316      3803.2    904681.1  1029200.4  -124519.3
 316      4102.3   1017892.9  1085436.8   -67543.9
 316      4436.8   1125683.5  1160689.8   -35006.3
 316      4739.0   1246982.5  1239951.1     7031.4
 316      5026.7   1377596.4  1325317.4    52279.0
 316      5372.2   1303630.9  1440618.3  -136987.4
 316      5731.4   1461523.4  1575311.5  -113788.1
 316      6210.0   1776409.8  1778172.0    -1762.3
 316      6661.4   1947772.3  1994115.0   -46342.8
 316      7304.2   2016219.8  2342693.0  -326473.3
 316      8054.2   2656859.8  2810452.8  -153593.0
 316      9120.4   4160179.3  3588630.8   571548.5
 316     11456.4   5638618.0  5758275.0  -119657.0
(...)

  SET   3     madc l1 isomorphous
OVERALL  MEAN D=  5.59 PHASING POWER =  3.20  M.R.E. =  .52 MEAN BIAS =  87.7 REFL=  4083
 UPDATED E VALUES BASED ON NEW PROTEIN PHASES

 SET   4     madc l2 isomorphous
OVERALL  MEAN D= 5.68 PHASING POWER =  2.36  M.R.E. = .53 MEAN BIAS =  88.3 REFL=  6792

  SET   5     madc l3 isomorphous
OVERALL  MEAN D=  5.69 PHASING POWER =   2.35  M.R.E. =  .51 MEAN BIAS =  88.0 REFL= 6563

  SET   6     madc hg hamburg isomorphous
OVERALL  MEAN D=    6.12 PHASING POWER = 1.63  M.R.E. = .64 MEAN BIAS =  84.6 REFL= 5963
```

Advances in MIR and MAD phasing : Maximum-Likelihood Refinement in a Graphical Environment, with SHARP

E. de La Fortelle, J. Irwin
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH England
eric@mrc-lmb.cam.ac.uk, ji10@mrc-lmb.cam.ac.uk
http://Lagrange.mrc-lmb.cam.ac.uk

and G. Bricogne
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH England and LURE, bât 209d, F-91405 Orsay Cedex, France
gb10@mrc-lmb.cam.ac.uk
http://gerard2.mrc-lmb.cam.ac.uk

Abstract

The problem of estimating heavy-atom parameters (esp. occupancies) from acentric reflexions in the MIR method has a long history of difficulties, and a conceptually satisfactory solution allowing bias-free refinement of all parameters (including the lack of isomorphism) has only recently been obtained by a recourse to the method of maximum-likelihood estimation. The situation is essentially identical in the case of MAD phasing. The maximum-likelihood method needs to be invoked to exploit incomplete phase information in a heavy-atom parameter refinement while preventing that information from biasing the results.

We have designed and written from scratch a computer program - SHARP (Statistical Heavy-Atom Refinement and Phasing) - that fully implements the maximum-likelihood approach. It can refine simultaneously scale, a model for the lack of isomorphism and all heavy-atom parameters from MIR and MAD data, or any mixture of them. The program has been systematically tested, both on synthetic and on measured data, and compared to MLPHARE. The results show the superiority of our approach, especially in cases of low signal-to-noise ratio. The likelihood function has also been used as a detection tool to plot residual Fourier maps and probe for minor sites, and for the calculation of phase probability distributions encoded in Hendrickson-Lattman coefficients.

1. Introduction

Bias-free refinement of heavy-atom parameters in the MIR and MAD methods, which is an essential step towards obtaining the best possible electron-density maps given the available data, has remained for a long time a troublesome issue in macromolecular crystallography. The conventional approach to this problem was originally conceived [1],[2] as a straightforward adaptation of the least-squares method previously used on centric data by Hart [3] : the "most probable" or the "best" estimates of the phases, as defined by Blow & Crick [4], were simply made to play a rôle analogous to that of the signs of centric reflexions. Dickerson, Weinzierl & Palmer [5] pointed out that more than two derivatives were needed for this type of refinement, and Blow & Matthews [6] found this method to have poor convergence properties unless steps were taken to ensure that the acentric phase estimates used in the refinement were independent of the parameters that were being refined. With hindsight, these difficulties are easily

rationalised : this 'phased' least-squares refinement was, in effect, violating the first cardinal rule of the least-squares method, namely that any quantity involved in the observational equations should be either a *model parameter* or an *observation*. Treating the native phase as a known constant within each cycle, but recalculating it after each refinement step, introduces bias on the parameters, especially in the case of mostly bimodal phase distributions.

At the same time as the first attempts were being made to use phase estimates, an alternative refinement scheme was devised by Rossmann [7], based on a difference-Patterson correlation criterion, and evolved towards the "FHLE method" [8],[9], and finally the "origin-removed Patterson-correlation function" [10]. Here the use of acentric phase estimates is avoided altogether, but at the price of impoverishing the available information in the sense that multiple derivatives are not allowed to assist each other's refinement through the generation of phase information.

Sygusch [11] recognized that a middle-ground could perhaps be found if the acentric phases were no longer deemed to be "estimates", but were instead treated as extra parameters and refined along with the others. Unfortunately, the enormous increase in the number of variables dictated the use of a diagonal approximation, which rather defeated the original purpose of accommodating the correlations between phases and parameters. Bricogne [12], [13] proposed a solution that partially overcame these difficulties. The main idea was that structure-factor estimates for acentric reflexions are *implicit functions* of the parameters that are being refined. This dependence was shown to result (*via* the chain rule) in a correction to the partial derivatives from which the normal equations of the least-squares method are to be formed. Many previously observed pathologies, such as the rapid divergence of the site occupancies of good derivatives, were cured by this analysis, but slower-moving instabilities were observed that resulted in divergent behaviour of the estimates for the *lack of isomorphism* of the various derivatives. Moreover, the problem of bimodality remained.

At this point, compliance with the first cardinal rule of the least-squares method had been essentially restored, but attention was drawn to the violation of a second cardinal rule : the inverse-variance 'weights' in the expression for the least-squares residual should be kept *fixed* as if they were part of the observed data. Since the method of least-squares is a special case of the maximum-likelihood method when errors are normally distributed with fixed (co)variances, it is clear that the problem of properly estimating the lack-of-isomorphism parameters demanded a fully-fledged maximum-likelihood treatment rather than least-squares.

Perusal of the literature shows that two-dimensional statistical 'phasing' (probability distribution on the phase and on the modulus of the native structure factor) had been considered as early as 1970 [14], leading to the first mention of likelihood in this context by Einsein [15]. The first mention of parameter estimation by maximum-likelihood, in a very limited context, is found in Green [16]. Maximum-likelihood (ML) refinement for heavy-atom parameters was then advocated by Bricogne [17],[18],[19], Read [20], and an approximation to it was implemented by Otwinowski [21] in the program MLPHARE. This program is only a partial implementation of ML refinement - best described as 'phase-integrated least-squares' - in the sense that (i) it integrates the exponential of the least-squares residual and its partial derivatives only over the phase of the native structure factor (not over its modulus) ; and that (ii) the lack of isomorphism is still *re-estimated* at the end of each refinement cycle rather than being *refined*, and may often converge to non-optimal values. Nevertheless, this approach has been shown in numerous cases to yield better results than earlier refinements, drawing attention to the potential of maximum-likelihood methods.

The maximum-likelihood formalism outlined in Bricogne [22] for the MIR and SIR cases forms the basis of the present work. We will describe here its extension to probability distributions incorporating anomalous diffraction effects as well as measurement error and non-isomorphism. Integrating these distributions in the whole complex plane leads to likelihood functions that can be used for heavy-atom detection and refinement, and for producing phase probability distributions. We will also describe the current implementation of this formalism in a computer program, named SHARP (for Statistical Heavy-Atom Refinement and Phasing) [23].

2 Likelihood functions for parameter refinement

2.1. Outline

Generally speaking, bias is introduced in a model incorporating some degree of randomness whenever a *distribution* for a random quantity is replaced by a *value* for that quantity. The likelihood formalism avoids this pitfall by consistently emphasizing that *distributions* are involved.

More specifically, a least-squares (LS) model is always formulated as a prescription for turning given values of model parameters into 'calculated' (error-free) values to be compared with the observables. Error estimates are obtained *a posteriori*, by examining the residual discrepancy between the 'calculated' and the 'observed' quantities. By contrast, a likelihood-based model casts its predictions directly in the form of probability distributions for the observables, the quantities called 'calculated' in the LS formalism usually appearing as parameters in these distributions.

2.2. The native structure factor

The most important thing to bear in mind when building up the likelihood function for heavy-atom refinement is that the complex value of the native structure factor $\mathbf{F}^P(\mathbf{h})$ is not known. The measurement of a native amplitude for an acentric reflexion **h**, if present, gives rise to a two-dimensional probability distribution $p(\mathbf{F}^P(\mathbf{h}))$. A measurement for the structure factor of a derivative crystal will also give rise to a two-dimensional probability distribution $p(\mathbf{F}^P(\mathbf{h}) \mid \{g\})$ for the *native* structure factor, conditional to the values $\{g\}$ of the set of global parameters for the heavy-atom model, for the scaling model and for the lack-of-isomorphism model.

For a centric reflexion, the probablity distribution becomes one-dimensional, but the theory is essentially similar.

2.3. The likelihood function

For a given reflexion **h**, the probability distribution of the native complex-valued structure factor, conditional to all the information available, is obtained by multiplying the probability distributions of $\mathbf{F}^P(\mathbf{h})$ for independent measurements.

This probability distribution is then transformed into a likelihood distribution for that reflexion, via the simple rule (in the absence of prior phase information) :

$$\Lambda\ (\{g\}, \mathbf{F}^{P}_{*}(\mathbf{h})) = p(\mathbf{F}^{P}_{*}(\mathbf{h}) \mid \{g\})$$

Note that this equation is valid at each *trial point* $\mathbf{F}^{P}_{*}(\mathbf{h})$ in the Harker plane. In order to have a likelihood function that is independent of assumptions on the native complex structure factor, we must now integrate the likelihood function over all possible values of $\mathbf{F}^{P}_{*}(\mathbf{h})$:

$$\Lambda\ (\{g\}) = \iint \Lambda\ (\{g\}, \mathbf{F}^{P}_{*}(\mathbf{h}))\, d^{2}\mathbf{F}^{P}_{*})$$

In the case of a centric reflexion, the integration is one-dimensional only, along the axis defined by the centric phase.

3. Parametrisation

3.1. Heavy-atom structure factors

This parametrisation amounts to a physical description of diffraction properties , involving heavy-atom coordinates, occupancies, isotropic and (if need be) anisotropic temperature factors, as well as normal and anomalous scattering factors. This was prefered to 'isomorphous' and 'anomalous' occupancies, because the physical parameters f' and f'' are either known precisely from physical tables (MIR experiment off an absorption adge) or can be measured from fluorescence scans (MAD experiment). Our implementation uses a hierarchical organisation for these parameters, that enables common attributes to be shared appropriately (Fig. 2). A list of site coordinates is determined that contains all known sites in all derivatives, and for each level of the hierarchy, these sites are 'qualified' (by a chemical identity, by an occupancy etc.). In this way, the long-standing problem of the same site being refined independently at each wavelength of a MAD experiment *cannot* occur, and common sites in a MIR experiment are parametrised correctly.

Future developments will incorporate a parametrisation of the anisotropy of anomalous scattering [24],[25] and will allow a refinement of the corresponding parameters from unmerged data carrying suitable goniometric information for each measurement.

3.2. Scale factors

Currently, scale factors are parametrised by a constant scale K^{sc}_{j}, an isotropic relative temperature factor B^{sc}_{j}, and six anisotropic increments $b^{p,q}_{j}$ to B^{sc}_{j} :

$$k_{j}(\mathbf{h}) = K^{sc}_{j} \exp[-1/4\ B(^{sc}_{j})(d^{*})^{2}) \exp[-(\sum\ b^{p,q}_{j} h^{p} h^{q})]$$

3.3. Lack-of-isomorphism variance

Differences between native and derivative structure factors are explained by a heavy-atom model, and by an error model. In the 'null hypothesis' where we know nothing about the heavy-atom structure, all

the differences are on average attributed to the error, and this error will be refined to smaller values as the heavy-atom model becomes more accurate.

This error can be broken down in three main components :

* The measurement error, that is part of the crystallographic data and not refined.

* The physical lack-of-isomorphism error.

In the absence of structural evidence for 'localised' lack-of-isomorphism, our assumption will be that of Luzzati [26] that there is a random isotropic positional perturbation, with spatially uniform mean amplitude and normal distribution, over the whole asymmetric unit. Based on this hypothesis, following the work of Read [27] and Dumas [28], we used a one-parameter model for the physical lack-of-isomorphism variance, increasing with resolution.

* The model error.

This error has the same effect on the statistical distribution of the native structure factor as the previous one, but its variance is approximately decreasing with resolution as the mean intensity of remaining heavy atoms. We used a two-parameter model (a constant and a temperature factor) for this error.

A similar parametrisation is used for the error on the anomalous differences. Although there is no physical basis for adopting the same model, it was thought flexible enough as a function of resolution to fit to more diverse functions of resolution.

4. Other uses for the likelihood function

4.1. Residual maps for model updates

The likelihood formalism also provides the opportunity of checking for significant systematic disagreement betweeen the data and the substitution model. For each reflexion **h**, we calculate the gradients of the log-likelihood function with respect to the real and imaginary parts of the various heavy-atom structure factors $\mathbf{F}^{H}_{j}(\mathbf{h})$. These numbers are then used in Fourier syntheses to produce residual maps, that have the symmetry of the crystal. Similarly, in the case where there is significant anomalous diffraction, the gradients with respect to $(\mathbf{F}^{H}_{j+} + \mathbf{F}^{H}_{j-})$ become coefficients for isomorphous residual maps, and those with respect to $(\mathbf{F}^{H}_{j+} - \mathbf{F}^{H}_{j-})$ for anomalous residual maps.

These maps enable the detection of minor sites, and perform this task in an optimal fashion because they take into account the full unbiased phase information available from the data at the current stage of refinement. They are essentially Fourier syntheses calculated from inverse-variance weighted difference coefficients between the derivative and native data. Their enhanced sensitivity to any departure from the current heavy-atom model (when the data are accurate enough, and to high enough resolution) makes them the instrument of choice to detect more subtle features, such as anisotropy in the heavy-atom temperature factors or structural disorder at certain sites.

4.2. Final phasing and calculation of Hendrickson-Lattman coefficients

Once the global parameters have been refined to convergence, the likelihood function ^{A} $(\mathbf{F}^{P}_{*},\{g\})$ considered as a function of the trial native structure factor $\mathbf{F}^{P}_{*}$ only, becomes (after suitable normalisation) the probability distribution function of the modulus *and* phase of the native structure factor (this is a simple application of Bayes's theorem). The two-dimensional centroids $\mathbf{F}^{P}_{best}(\mathbf{h})$, used as Fourier coefficients of the electron-density map, and the Hendrickson-Lattman 'ABCD' coefficients [29] of the marginal phase distribution can be easily derived from this likelihood function.

4.3. Future developments and perspectives

A natural extension of the quantitative use of residual maps based on log-likelihood gradients is the refinement of heavy-atom clusters of known geometry by real-space techniques of the Agarwal-Lifchitz type (e.g. as implemented in the TNT package). This is currently underway.

In order to offer *ab initio* detection capability, another type of map will be added to the existing program. Its coefficients will initially involve second-order derivatives of the log-likelihood function associated to the null hypothesis defined by "all intensity differences between data sets are caused by lack of isomorphism". This map will have the character of a Buerger sum function over a weighted difference-Patterson function [30]. As major sites are detected and included in the substitution model, the log-likelihood function will develop first-order derivatives giving rise to a difference-Fourier component in the residual map, while the revised second-order derivatives will keep contributing a component with the character of a sum function over a residual difference-Patterson.

The whole process of detecting sites and of assessing their significance quantitatively can thus be automated, using the log-likelihood gain referred to the null hypothesis as a scoring criterion for the peak-search. The procedure will stop when the highest remaining peak in the residual maps is essentially at the level of the noise.

Once all heavy atoms have been detected and refined, the remaining features in the 'isomorphous' residual maps, if they are significant, can provide the basis for a systematic study of lack of isomorphism. This could improve the rather crude way in which 'global' and 'local' lack of isomorphism have hitherto been described.

5. The Graphical User Interface

Because the program can accomodate data from many different experimental procedures (MIR, with or without anomalous scattering, MAD, or a blend of the two), it was necessary to guide the user during the buildup of a hierarchical parameters file describing this experiment. This was achieved by means of an HTML browser-based Graphical User Interface. The same system was used to facilitate inspection of the output of the program.

5.1. Choice of tools

Our approach was based on a client-server philosophy, in order to make best use of the World Wide Web as a communication tool. As a result, once SHARP is installed on a server (a powerful computer, workstation or other, that will actually do the calculations), any authorised user can run the program from any terminal connected to the Internet. This has proved invaluable during the beta-testing stage, and provides high flexibility for all users. On the other hand, if this 'universal access' becomes a security issue, it can be reduced to an Internet subdomain, or to a single machine.

The result is a forms-based interface, written in HTML language and processed by Perl scripts, that exactly mirrors the hierarchy of parameters during the buildup of the parameters file, and that connects automatically to Graphical Helper Applications to facilitate inspection of the output.

5.2. Input

The input pages consist in a series of embedded forms, that guide the user through our parametrisation of the experiment (list of sites, compounds, crystals, wavelengths, batches). Because the options taken in the higher levels condition the structure of the lower levels, the setting of the parameter tree is unidirectional (i.e. coming back up the tree erases what has been set further down).

5.3. Output

Maximum advantage is derived, in the output, from the hyper-link facility of the HTML language. A mouse-click on a hyper-link opens another file, accessible from the Internet. In practice, the information created by the program, instead of being stored in a single massive log-file, consists in a large number of small files stored in an 'output directory'. All these small 'explanatory' files can be accessed from a master file, called 'SHARP output', by means of specific hyperlinks. The master file contains the minimal information needed to follow the progress of the refinement, and all details are accessed through hyperlinks.

In the same way, the documentation can be accessed in a context-sensitive manner by clicking on hyperlinks called 'explanation', scattered at all points of interest in the master file and in secondary details files.

Graphical applications are triggered through a Unix "mailcap" mechanism, that relies on the extension of a file name to determine what program to use for visualising the contents. All statistics relative to the data (histograms of intensity, of isomorphous and anomalous differences) and to the phasing (lack of isomorphism statistics, phasing power and Rcullis) can be visualised that way, and maps can be plotted in programs npo [31] or O [32] by pressing a button in the interface, without having to program specific instructions for these graphical tools.

6. Applications

6.1. MAD dataset : IF3-C

One of the first experimental (as opposed to synthetic) datasets that we processed using SHARP was the IF3-C [33],[34] (C-terminal part of translational initiation factor 3). The two methionine residues of this 94-residue protein were substituted for selenomethionines and a three-wavelength anomalous diffraction experiment was performed at the Selenium K edge.

The starting heavy-atom model consisted in two selenium atoms with isotropic thermal motion. Refinement of this model showed that, consistently with the results of other refinement procedures, the second selenium atom had a high temperature factor (around 60). Once the refinement was completed, the residual maps showed strong anisotropic features for the first selenium site and weaker anisotropy for the second. We consequently updated the heavy-atom model by allowing an anisotropic temperature factor for both seleniums atoms. The resulting residual map showed much less features above the noise level, except for a 10σ peak at 1.8 Å distance from the first selenium site. The second update of the heavy-atom model allowed for a third selenium atom with an isotropic temperature factor, that refined to a low occupancy (0.2). The remarkable result was that the added occupancies of site 1 and site 3 were equal to the the occupancy of site 2 within the standard deviation of this parameter. This observation, added to the small distance between site 1 and site 3, shows that this methionine residue has a double conformation.

We then used the density modification program SOLOMON [35] to improve the phases, assuming that it would yield better results when the input phase probability distributions (encoded as Hendrickson-Lattman coefficients) are statistically more accurate. The density modification prodedure for both SHARP and MLPHARE was exactly similar. The results are summarized in Table 1.

6.3. SIRAS dataset : U2

This dataset had just been collected at the Trieste synchrotron source, at a wavelength that optimised the anomalous signal of the mercury atoms. The protein is a ternary complex of two proteins (U2A'/U2B'') and an RNA hairpin (U2snRNA hairpin IV) involved in the spliceosome [36]. The total molecular weight is 50kD. There are two molecules in the asymmetric unit, but the non-crystallographic symmetry was not used in the model-building stage, due to the very high quality of the maps.

The starting heavy-atom model consisted of two mercury sites, for which coordinates, occupancy and temperature factor were determined in a first round of refinement. The residual map plotted at the end of this refinement showed strong anisotropy features for both sites, and a suspicion of a double position for site 1. The anisotropy was refined first, and the subsequent residual map clearly showed that the cysteine residue to which the first mercury was bound had a double conformation. Once this was taken into account in the heavy-atom model in a third round of refinement, the residual map showed no more significant features, thus proving that the refinement was complete. The resulting map, after density modification in SOLOMON, was of high quality (see Table 2).

Interestingly, in this case the anomalous residual map yielded a much clearer information that the isomorphous redsidual map. This was confirmed by the phasing power statistics, which showed that, due to significant lack of isomorphism, the anomalous signal contributed far more to the phasing than the isomorphous signal. The whole procedure of refinement and phasing was then started again from the same initial assumptions, but *without using the native data.* Heavy-atom refinement yielded the same results, and the residual maps allowed unambiguous detection of both the anisotropic thermal motion and the double conformation. Phasing of this "Single-Wavelength Anomalous" dataset, followed by the same solvent-flattening procedure, yielded an interpretable electron-density map, although of a lesser

quality than the SIRAS map (see Table 2).

7. Conclusion

The maximum-likelihood refinement in SHARP, coupled with the very sensitive log-likelihood gradient maps used to detect residual features of the heavy-atom model, produces phase probability distributions for all measured reflexions that are an optimal starting point for density-modification procedures.

The test of using the anomalous scattering of a derivative by itself, in the second example, is of special interest. It was not useful for the determination of the structure in that particular case because the isomorphism was relatively good between the native and mercury derivative crystals. It shows nonetheless that, in cases of very strong non-isomorphism, a well-substituted derivative can be used *by itself* to provide phase information, if the anomalous signal is strong. In such a case of complete bimodality in the phase distribution of acentric reflexions, the main purpose of the density modification procedure is to select the right mode. SOLOMON seems to perform this task for most reflexions thanks to the envelope constraints.

References

[1] R. E. Dickerson, J. C. Kendrew & B. E. Strandberg "The Phase Problem and Isomorphous Replacement Methods in Protein Structures" In *Symposium on computer methods and the phase problem*, p. 84. Glasgow : Pergamon Press, 1960.

[2] R.E. Dickerson, J.C. Kendrew & B.E. Strandberg . In *Computing Methods and the Phase Problem in X-ray Crystal Analysis*, edited by R. Pepinsky, J.M. Robertson & J.C. Speakman, pp.236-251. Oxford : Pergamon Press, 1961.

[3] R.G. Hart *Acta Cryst.* **14**, pp.1194-1195, 1961.

[4] D.M. Blow & F.H.C. Crick "The Treatement of Errors in the Isomorphous Replacement Method" *Acta Cryst.* **12**, 794-802, 1959.

[5] R. E. Dickerson, J. E. Weinzierl & R. A. Palmer "A Least-Squares Refinement Method for Isomorphous Replacement" *Acta Cryst.* B**24**, 997-1003, 1968.

[6] D.M. Blow & B.W. Matthews "Parameter Refinement in the Multiple Isomorphous-Replacement Method" *Acta Cryst.* A**29**, 56-62, 1973.

[7] M. G. Rossmann "The Accurate Determination of the Position and Shape of Heavy-Atom Replacement Groups in Proteins" *Acta Cryst.* **13**, 221, 1960.

[8] G. Kartha "Comparison of Multiple Isomorphous Replacament ans Anomalous Dispresion Data for Protein Structure Determination.III. Refinement of Heavy Atom Positions by the Least-Squares Method"*Acta Cryst.* **19**, 883-885, 1965.

[9] E. J. Dodson, P. R. Evans & S. French "The Use of Anomalous Scattering in Refining Heavy Atom

Parameters in Proteins" *Anomalous Scattering*, pp. 423-436. Edited by S. Ramaseshan and S. C. Abrahams. Copenhagen : Munskgaard, 1975.

[10] T. C. Terwilliger & D. Eisenberg "Unbiased Three-Dimensional Refinement of Heavy-Atom Parameters by Correlation of Origin-Removed Patterson Functions" *Acta Cryst.* A**39**, 813-817, 1983.

[11] J. Sygusch "Minimum-Variance Fourier Coefficients from the Isomorphous Replacement Method by Least-Squares Analysis" *Acta Cryst.* A**33**, 512-518, 1977.

[12] G. Bricogne "Multiple Isomorphous Replacement : The Problem of Parameter Refinement from Acentric Reflexions" In *Computational Crystallography*, edited by D. SAYRE, pp. 223-230. New York: Oxford University Press, 1982.

[13] G. Bricogne "Application of Isomorphous Replacement and Anomalous Dispersion Techiques to Proteins" In *Methods and Applications in Crystallographic Computing*, edited by S.R. HALL & T. ASHIDA, pp. 141-151. Oxford : Clarendon Press, 1984.

[14] V. SH. Raiz & N. S. Andreeva "Determining the Coefficients of the Fourier Series of the Electron-Density Function of Protein Crystals"*Sov. Phys. Crystallogr.* **15**, 206-210. Translated from *Kristallografiya* **15**, 246-251, 1970.

[15] R. J. Einstein "An Improved Method for Combining Isomorphous Replacement and Anomalous Scattering Diffraction Data for Macromolecular Crystals" *Acta Cryst.* **A33**, 75-85, 1977.

[16] E. A. Green "A New Statistical Model for Describing Errors in Isomorphous Replacement Data : The Case of One Derivative"*Acta Cryst.* **A35**, 351-359, 1979.

[17] G. Bricogne Unpublished lecture given at the Bischenberg conference on the Crystallography of Molecular Biology, 1985.

[18] G. Bricogne "A Bayesian Theory of the Phase Problem. I. A Multichannel Maximum-Entropy Formalism for Constructing Generalized Joint Probability Distributions of Structure Factors" *Acta Cryst.* A**44**, 517-545, 1988.

[19] G. Bricogne "A Maximum-Likelihood Theory of Heavy-atom Parameter Refinement in the Isomorphous Replacement Method" In *Isomorphous Replacement and Anomalous Scattering* Proc. Daresbury Study Weekend, pp. 60-68. SERC Daresbury Laboratory, Warrington, England, 1991.

[20] R. J. Read "Dealing with imperfect isomorphism in multiple isomorphous replacement" In *Isomorphous Replacement and Anomalous Scattering* Proc. Daresbury Study Weekend, pp. 69-79. SERC Daresbury Laboratory, Warrington, England, 1991.

[21] Z. Otwinowski "Maximum Likelihood Refinement of Heavy Atom Parameters" In *Isomorphous Replacement and Anomalous Scattering* Proc. Daresbury Study Weekend, pp. 80-85. SERC Daresbury Laboratory, Warrington, England, 1991.

[22] G. Bricogne *op. cit,* 1991.

[23] E. de La Fortelle & G. Bricogne. "Maximum-Likelihood Heavy-Atom Parameter Refinement in the MIR and MAD Methods" In *Methods in Enzymology,* (C.W Carter & R.M. Sweet, eds), **276**, Chapter 27, pp472-494, Academic Press, 1997.

[24] D.H. Templeton & L.K. Templeton *Acta Cryst.* A**38**, 62-67, 1982.

[25] L.K. Templeton & D.H. Templeton *Acta Cryst.* A**44**, 1045-1051, 1988.

[26] V. Luzzati *Acta Cryst.* **5**, 802-810, 1952.

[27] R. J. Read "Improved Coefficients for Maps Using Phases from Partial Structures With Errors" *Acta Cryst.* A**42**, 140-149, 1986.

[28] P. Dumas "The Heavy-Atom Problem : a Statistical Analysis. I. *A Priori* Determination of Best Scaling, Level of Substitution, Lack of Isomorphism and Scaling Power" *Acta Cryst.* A**50**, 526-537, 1994.

[29] W. A. Hendrickson & E. E. Lattman "Representation of Phase Probability Distributions for Simplified Combination of Independant Phase Information" *Acta Cryst.* B**26**, 136-143, 1970.

[30] G. Bricogne. "Bayesian Statistical Viewpoint on Structure Determination : Basic Concepts and Examples" In *Methods in Enzymology,* (C.W Carter & R.M. Sweet, eds), **276**, Chapter 23, pp. 361-423, Academic Press, 1997.

[31] Collaborative Computational Project, Number 4 "The CCP4 suite : Programs for Protein Crystallography" *Acta Cryst.* D**50**, 760-763, 1994.

[32] T.A. Jones, J.Y. Zou, S.W. Cowan & M. Kjeldgaard, "Improved Methods for Building Protein Models in Electron Density Maps and the Location of Errors in these Models" *Acta Cryst.* A**47**, 110-119, 1991.

[33] V. Biou, F. Shu & V. Ramakrishnan "X-Ray Crystallography Shows that Translational Initiation Factor IF3 Consists of 2 Compact alpha/beta Domains Linked by an alpha-helix" *EMBO. J.* **14**, 4056-4064, 1995.

[34] V. Ramakrishnan &V. Biou. "Treatment of Multiwavelength Anomalous Diffraction Data as a Special Case of Multiple Isomorphous Replacement" In *Methods in Enzymology,* (C.W Carter & R.M. Sweet, eds), **276**, Chapter 31, pp. 538-557, Academic Press, 1997.

[35] J.P. Abrahams & A. G. W. Leslie "Methods used in the structural determination of bovine mitochondrial F1 ATPase" *Acta Cryst.* D**52**, 30-42, 1996.

[36] S. Price & K. Nagai, unpublished results.

TABLES

Glossary :

FOM is the mean figure of merit in that resolution bin.
DELTAPHI is the mean phase difference, *weighted by amplitude and FOM*, in that resolution bin.
CORREL is a reciprocal-space correlation coefficient between complex structure factors. By Parseval's theorem it is equivalent to a real-space correlation coefficient in that resolution bin.

Resolution (Å)	ALL	50.0	5.25	3.73	3.05	2.64	2.36	2.16	2.00	1.87
SHARP refinement and phasing, density modification with SOLOMON										
FOM	0.90	0.84	0.91	0.91	0.90	0.90	0.90	0.90	0.89	
<DELTAPHI>	30.5	39.1	25.3	29.5	32.0	30.6	29.3	30.9	32.2	
CORREL	0.80	0.70	0.86	0.81	0.77	0.80	0.82	0.80	0.78	
MLPHARE refinement and phasing, density modification with SOLOMON										
FOM	0.90	0.84	0.91	0.91	0.90	0.90	0.90	0.90	0.89	
<DELTAPHI>	30.5	39.1	25.3	29.5	32.0	30.6	29.3	30.9	32.2	
CORREL	0.80	0.70	0.86	0.81	0.77	0.80	0.82	0.80	0.78	

Table 1 : Quality of IF3-C MAD phasing, in comparison with the refined model

---

Resolution (Å)	ALL	50.0	7.83	5.57	4.56	3.95	3.54	3.23	2.99	2.80
SHARP refinement and phasing, density modification with SOLOMON - SIRAS data										
FOM	0.90	0.89	0.95	0.96	0.96	0.95	0.92	0.89	0.78	
<DELTAPHI>	43.3	38.9	35.9	32.6	36.8	42.6	50.8	57.8	64.6	
CORREL	0.66	0.64	0.74	0.78	0.73	0.66	0.55	0.46	0.37	

SHARP refinement and phasing, density modification with SOLOMON - SAD data

FOM	0.90	0.86	0.93	0.95	0.94	0.95	0.92	0.88	0.82
<DELTAPHI>	57.0	58.2	50.0	48.2	50.7	55.8	62.9	68.0	72.2
CORREL	0.49	0.45	0.57	0.60	0.56	0.49	0.39	0.32	0.26

Table 2 : Quality of U2 SIRAS phasing and SAD (Single-Wavelength Anomalous Diffraction) phasing, with SHARP

Multiwavelength anomalous dispersion phasing strategies investigated with a brominated oligonucleotide.

* *Mark R. Peterson*

Structural Chemistry Section, Department of Chemistry, University of Manchester, Oxford Road, Manchester, M13 9PL, England, U.K.

Abstract

Multiwavelength anomalous dispersion methods were used to analyse the crystal structure of d(CGCG^{Br}CG) in extension of the work presented in Peterson, Harrop, McSweeney, Leonard, Thompson, Hunter and Helliwell (1996) *J. Synch. Rad.* **3**, 24-34. The brominated oligonucleotide d(CGCG^{Br}CG) of chemical formula $C_{114}N_{48}O_{68}P_{10}Br_2$ crystallises in space group $P2_12_12_1$ with unit cell dimensions a=17.97, b=30.98, c=44.85Å, $\alpha=\beta=\gamma=90^o$. It was chosen as a test crystal to evaluate the MAD method itself and to commission station PX9.5 for several reasons; it was radiation insensitive; it had a very good concentration of anomalous scatterers, i.e. two bromines in two hundred and forty light atoms; and the bromine K edge was very near to the critical wavelength flux output of the SRS wiggler. It also diffracted strongly, due to the relatively small unit cell, in spite of the rather small crystal volume. Data to a resolution of 1.65Å were collected at four wavelengths about the bromine atom K absorption edge using synchrotron radiation at Station PX9.5, SRS, Daresbury.Traditionally, the maximum of f '' is not coincident with the minimum in f ', however, in this case both are observed on the same data set, λ_2. Hence Δ_{anom} and $\Delta f'$ could be maximised using only two wavelengths. Various wavelength combinations phasing strategies were then studied, ranging from 4 to 2 wavelengths. DM phase improvement procedures were also employed on these combinations giving highly interpretable maps even for unoptimised 2 wavelength cases.

Data Collection

Data collection was conducted at Station PX9.5 at the Synchrotron Radiation Source (SRS) in Daresbury, England. The single crystal selected for the data collection had a pseudo-hexagonal plate morphology of dimensions 0.2 X 0.1 X 0.01 mm. As the anomalous scattering factors are derived from the atomic absorption coefficient, a XANES (X-ray absorption near edge structure) experiment was also carried out on station PX9.5 to decide upon the precise wavelengths to be used in the data collection. The $\delta\lambda/\lambda$ for the beam was set at 4.4×10^{-4}, by restricting the vertical divergence of the beam by a factor of two with the use of slits upstream of the focussing mirror.

Upon inspection of a test diffraction image, it could be seen that the crystal was relatively well aligned, i.e. the Bijvoet mates could be measured on the same or adjacent images. No further crystal alignment was undertaken. The wavelengths for the diffraction measurements were chosen to optimise the phasing power by (a) maximising the f '' effect and (b) $\Delta f'$ for different wavelengths for each hkl. Hence, four wavelengths were chosen: (1) a reference on the long wavelength side of the edge (λ_1=0.9323 Å); (2) at the absorption edge inflection point (λ_2=0.9192 Å); (3) at the "white line" absorption maximum (λ_3=0.9185 Å); (4) a reference on

* Current Address: Wellcome Sciences Institute, Department of Biochemistry, University of Dundee, Dundee, DD1 4HN, Scotland, U.K.

the short wavelength side of the edge (λ_4=0.8983 Å). The choices of λ_2 and λ_3 follow what are known as f ' dip and f " max respectively.

For each wavelength the crystallographic data were collected, each involving a 4° rotation of the crystal. For each 4° sweep the total exposure time was 60 seconds. In total 120° of data were collected for each of the four wavelengths. Another 120°, a fifth data set was then collected on the same crystal, immediately after the MAD data, at the "white line" (i.e. λ_3) but with the crystal misaligned by offsetting one of the goniometer head arcs by approximately 30°. This allowed reflections previously in the blind region to be measured and combined with the λ_3 data set.

Merging statistics from the five data sets are displayed in Table 1.

Analysis Vs. Intensity	λ_1	λ_2	λ_3	λ_4	$^5\lambda_3$
R_{merge}	2.3%	2.7%	2.9%	2.5%	2.9%
R_{anom}	2.0%	8.6%	7.2%	5.8%	6.3%
Total No. of Reflections	11372	11598	11573	11655	11506
No. of Ind. Reflections	3028	3052	3054	3065	2988
Completeness	93.1%	93.7%	93.8%	93.6%	94.5%
Multiplicity	3.8	3.8	3.8	3.8	3.9

Table 1. Merging statistics for the five data sets from AGROVATA (CCP4 (1994)).

The weak and negative intensities were made consistent with a Wilson distribution of structure factor amplitudes using TRUNCATE (CCP4). The computer programs CAD and SCALEIT (CCP4) were employed to combine the five data sets into one file and to put them on an overall common scale. This was done with respect to λ_2, it was treated as the 'native'. It was indeed found that λ_2 had the largest MFID between all other data sets.

	MFID	MFAD.	kemp	ktheor
λ_1	5.7 (3.3)%	2.7%	3.67	11.71
λ_2	0 (3.5)%	11.5%	0	0
λ_3	3.5 (0)%	9.8%	0.64	0.84
λ_4	6.9 (5.0)%	7.8%	1.48	1.89
$^5\lambda_3$	6.1 (5.3)%	9.3%	1.04	0.84

Table 2. SCALEIT (CCP4) statistics between wavelengths treating λ_2 as the 'native', bracketed values are statistics treating λ_3 as the 'native'.

SCALEIT provides useful estimates of the largest acceptable dispersive and absorptive differences between and within the different λ data sets. Due to the sensitivity of Patterson methods to spurious, large, differences it was important to reject any unacceptably large differences as outliers. The final SCALEIT statistics are shown in Table 2.

Dispersive and absorptive Patterson maps were then generated with FFT. Identification of the bromine sites could be readily found using both the anomalous (i.e. $\lambda 2F^+ - \lambda 2F^-$) and dispersive (i.e. $F\lambda_4 - F\lambda_2$) Patterson maps. From the three Harker sections in both maps, two consistent bromine sites could be easily found. The quality of these Patterson maps can be seen in Figures 1. and 2. The positions of the two bromine sites were 0.3241, 0.2009, 0.0100 (Site A) and 0.5010, 0.1807, 0.2310 (Site B) respectively.

Figure 1. Harker sections (U=1/2, V=1/2, W=1/2) of the Patterson map (shown from 0 to 1/2 in each section for U, V and W as appropriate) calculated with Patterson coefficients based upon anomalous differences recorded at λ_2 (0.9192Å).

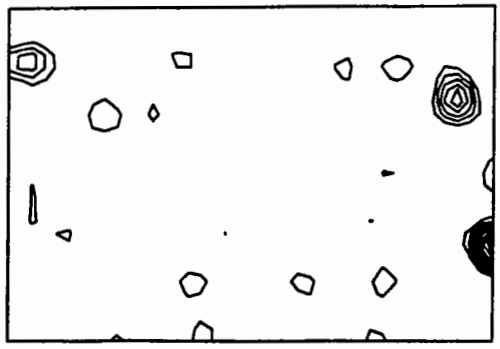


Figure 2. Harker sections (U=1/2, V=1/2, W=1/2) of the Patterson map (shown from 0 to 1/2 in each section for U, V and W as appropriate) calculated with Patterson coefficients based upon dispersive differences between data sets λ_2 and λ_4, i.e. the f ' dip and the short wavelength reference data set.

Phase calculations

Each bromine atom had its co-ordinates, temperature factors and occupancies (both real and anomalous) refined in MLPHARE (CCP4)for ten cycles. The refined positions of the two bromine atoms were used in MLPHARE on both hands (i.e. x y z and $\bar{x}$ $\bar{y}$ $\bar{z}$). MLPHARE also treats the data sets collected at different wavelengths as isomorphous derivatives with one data set being chosen as the 'native'. To maintain a consistent positive dispersive difference between the other data sets, the f ' dip data set (λ_2) was chosen as the native. Dispersive differences between λ_2 and the other data sets give rise to isomorphous differences, especially λ_1 and λ_4, with respect to λ_2 which were treated as apparent real occupancies of the anomalous scatterers. For the 'native' data set (λ_2), the real occupancies of the anomalous scatterers were fixed to zero initially. The figures of merit of the MAD phases, using all four wavelengths (excluding the $^5\lambda_3$ data set), were 0.86/0.82 to 1.65Å resolution for the acentric/centric data respectively for both hands. The f ' and f " anomalous scattering factors were added to the form factor list, both being arbitrarily set equal to one electron so that the real and anomalous occupancies corresponded to the number of electrons involved in the dispersive and absorptive differences respectively, as the data sets were on a common absolute scale previously via SCALEIT and TRUNCATE. Table 3 gives the relevant phasing statistics for each derivative against the native (λ_2), and also compares the theoretical values of the anomalous coefficients f ' and f " (Sasaki, 1989) at each wavelength with the coefficients extracted at each wavelength via the occupancies in MLPHARE.

The phases from MLPHARE were then combined with the structure factor amplitudes from the λ_2 native data set, enabling a MAD electron density map was calculated via FFT (CCP4). The MAD maps were calculated on both hands (Figs. 3 (a) and (b)) at 1.65 Å resolution. The figures of merit for both sets of phases do not distinguish between correct and incorrect enantiomers. The problem is only resolved upon inspection of the MAD electron density maps for "chemical sense". That is the map calculated on the correct hand (Fig. 3(a)) showed the bases clearly and building of the model with O (Jones et al (1989)) could be easily started from the known heavy atom positions. The map calculated on the wrong hand was totally uninterpretable (Fig. 3 (b)).

(a) (b)

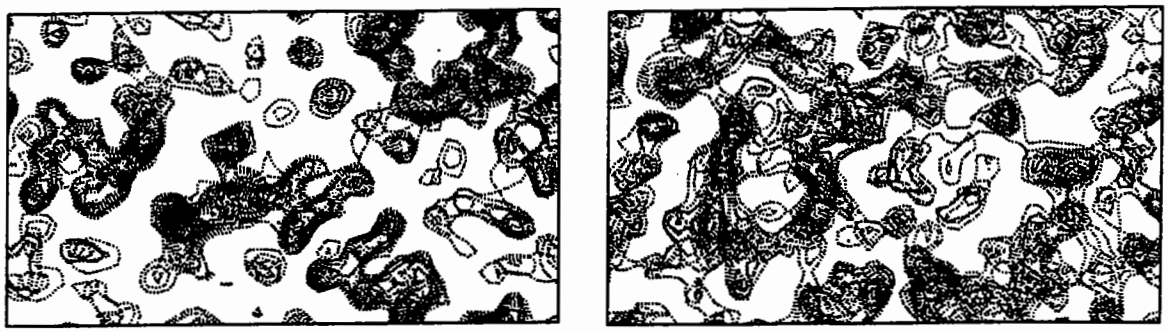


Figure 3. A portion of the MAD electron density map (shown from 0 to 1 in x and y for a 2.8Å thick slab in z) calculated at 1.65Å resolution contoured at 0.5 rms intervals and commencing at 0.5 rms (a) calculated on the correct hand (b) calculated on the wrong hand.

		Theory (SASAKI (1989))				Experimental (MLPHARE)			
	λ	λ'	f '	f "	Δf '	$\Delta f'_A$	$\Delta f'_B$	f''_A	f''_B
λ_1	0.9323	0.9332	-3.83	0.515	6.12	5.64	5.5	0.45	0.45
λ_2	0.9192	0.9201	-9.95	3.823	0	0	0	5.68	3.55
λ_3	0.9185	0.9194	-6.76	3.817	3.19	3.05	2.31	3.83	3.17
λ_4	0.8983	0.8992	-3.08	3.644	6.87	6.58	6.37	2.99	2.72
$^5\lambda_3$	0.9185	0.9194	-6.06	3.817	3.19	1.96	2.70	2.55	2.40

Table 3. The theoretical values of the anomalous scattering factors at the wavelengths (λ) used (from Sasaki (1989) corrected for a shift of 11 eV (λ') due to the absolute setting of monochromator being incorrect) and also those experimentally derived from MLPHARE. MLPHARE has anomalous scattering factors for both bromine sites A and B.

Key observations on the MAD Work.

In the variation of f ' and f " with wavelength, only two wavelengths need to be measured to yield a Δ_{anom} at one wavelength and a change via Δf ' of F_{hkl} between the two wavelengths (Okaya and Pepinsky (1956); Hoppe and Jakubowski (1975); and Helliwell (1979)). The choice of wavelengths to maximise Δ_{anom} and Δf ' was made with reference to the fluorescence spectrum. A key objective is to make the centres of the phasing circles in the Harker phasing diagram well separated and non-collinear; which is a necessary and sufficient condition for phasing (Helliwell (1984)). Traditionally, the maximum of f " is not coincident with the minimum in f '. Hence, three wavelengths would be needed in such a situation for fully moving the centres of the phasing circles apart. In this study however, although λ_3 was expected to have the largest Friedel anomalous difference, in fact that was the case for the λ_2 (f ' dip) data set (e.g. see R_{anom} values in Table 1). In light of λ_2 being the f " maximum, λ_3 was taken as 'native' to comfirm if λ_2 was indeed at the f ' dip. This was done by comparing MFID's between data sets where λ_2 then λ_3 are taken as the 'native' data sets. It was indeed found that λ_2 had the largest MFID between all other data sets (see Table 2). In such a case then, where both the f " maximum and the f ' minimum case are both observed on the same data set, i.e. λ_2, one data set becomes essentially redundant i.e. λ_3 in making the biggest anomalous differences. Hence, various alternative strategies of λ combinations were investigated.

Phase Information and Electron Density Map Quality from Various Wavelength Combinations.

The following analysis can essentially be split up into three categories involving data sets recorded at: respectively four, three, and two wavelengths in a variety of combinations to explore both experimental strategies for phasing and theoretical/computational strategies of phase improvement (See Figure 4 and Table 4 for respective map quality and FOM's). The experimental straregies were published in Peterson *et al.* (1996).

Case 1: λ_1, λ_2, λ_3, λ_4
This combination of wavelengths is the case described previously where the f " anomalous effects of each wavelength are all utilised along with the isomorphous effects between λ_2 and each of the other three wavelengths. The map was of excellent quality and structural moieties could be easily characterised.

Case 2: λ_1, λ_2, λ_3
This three wavelength case, and the next, is to compare the two possible choices of reference wavelength. Sometimes, due to lack of SR beam time and/or prolonged exposure times, it may be only feasible to collect data at three wavelengths. The reference wavelength, λ_1, has no anomalous signal as it is situated on the long wavelength side of the Br K edge. The map, however, was of excellent quality and could be easily characterised.

Case 3: λ_2, λ_3, λ_4
The reference wavelength, λ_4, has a good anomalous signal as it is situated on the short wavelength side of the absorption edge, unlike λ_1. The overall figure of merit was certainly improved compared with case 2. The map was again of excellent quality and could be easily characterised.

Case 4: λ_2, λ_4
The theoretical minimum case for unique phase determination involves two wavelengths. This is akin to the 'two-short-wavelength-method' of Hoppe and Jakubowski (1975). It is required that the centres of the phasing circles be well separated and non-collinear and this is achieved well here (Helliwell (1984)). The λ_2, λ_4 pairing has the largest dispersive difference, whilst, λ_2 also has the maximum Friedel difference. The electron density map was of high quality and totally interpretable.

Case 5: λ_2 , λ_3
This combination of wavelengths stimulated by the correspondence from D. H. Templeton, was used to see if the map could be phased with two extremely close wavelengths (i.e. only 0.0007Å apart!) that might be adversely affected by dichroism effects. Also the λ_2, λ_3 pairing has half the dispersive signal compared to the theoretical minimum, case 4, λ_2, λ_4. However, λ_2 has the largest anomalous difference whereas λ_3 has the next largest anomalous difference..

Density Modification Procedures for Improvement of Phase Quality.

The principle of density modification (DM) is to improve the experimental phases by imposing restrictions on the density in real space and then using the phases of the modified map to alter or replace the experimental phases. In protein crystallography these are important methods for phase improvement. Moreover they may be applied so as to reduce the number of wavelengths needed in a MAD phase determination experiment and/or use wavelengths very close in value, but with reduced (less optimal) values of f " or Δf '. The map modification process embroided

in the program DM (Cowtan (1994)) was used on the various wavelength phasing combinations.

Case 1: Density Modified λ_1, λ_2, λ_3, λ_4
The quality of the original map was very good, however, DM improved the map quality around all the bases. All bases now had well defined, complete electron density apart from base 7 which still had a lack of connectivity at one bond.

Case 2: Density Modified λ_1, λ_2, λ_3
Seven bases (1, 3, 8, 9, 10, 11 and 12) that had incomplete density (side chains missing or lack of connectivity) originally, sufficiently improved to now show well resolved connected density. The remainder of the bases, which had previously suffered from a lack of connectivity, were still not significantly altered.

Case 3: Density Modified λ_2, λ_3, λ_4
Eight bases (1, 3, 4, 8, 9, 10, 11, and 12) that had incomplete density (side chains missing or lack of connectivity) originally, sufficiently improved via to now show well resolved connected density. The remainder of the bases which suffered from a lack of connectivity were not significantly altered.

Case 4: Density Modified λ_2, λ_4
Eight bases (3, 4, 6, 8, 9, 10, 11 and 12) which were defined by density with a lack of connectivity at a least one bond now showed well defined connected density after DM. The remaining four bases showed a clear improvement in density quality, e.g. base 1 now has the nitrogenous side chain defined.

Case 5: Density Modified λ_2 , λ_3
The original map had most structural moieties in the correct position. DM further increased the map quality considerably, so much so that all the bases are easily characterised. Bases 3, 4, 5, 10 and 11 now had well defined connected density compared to the lack of connectivity experienced in the original map at these positions. Bases 1, 6, 9 and 12 showed improved density, whereas bases 7, 8 were still interpretable, but were slightly better defined in the original map. Base 2 showed no significant change in density. As might be expected this modified map was not of a high quality as compared to modified case 4.

(a) (b)

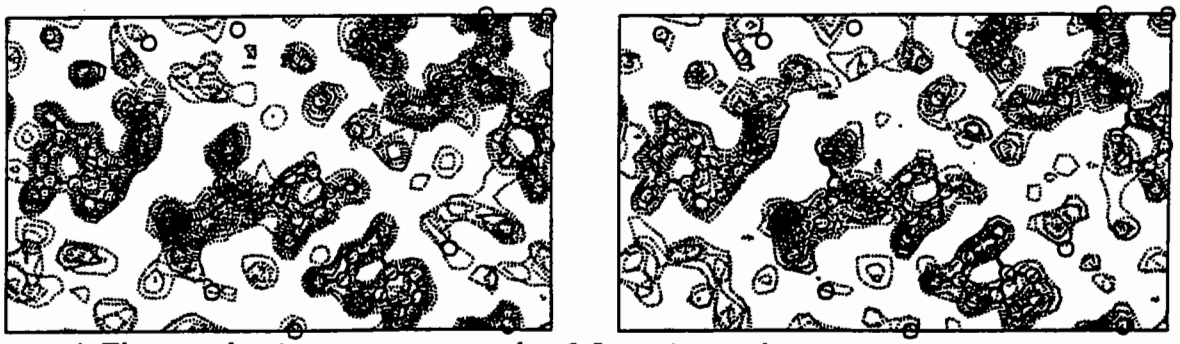


Figure 4. Electron density maps contoured at 0.5 rms intervals commencing at 0.5 rms shown from 0 to 1 in x and y for a 2.8 Å thick slab in z. The final refined molecular model is superimposed so as to compare the maps for.a) Case 4: λ_2, λ_4.b) Case 5: λ_2, λ_3

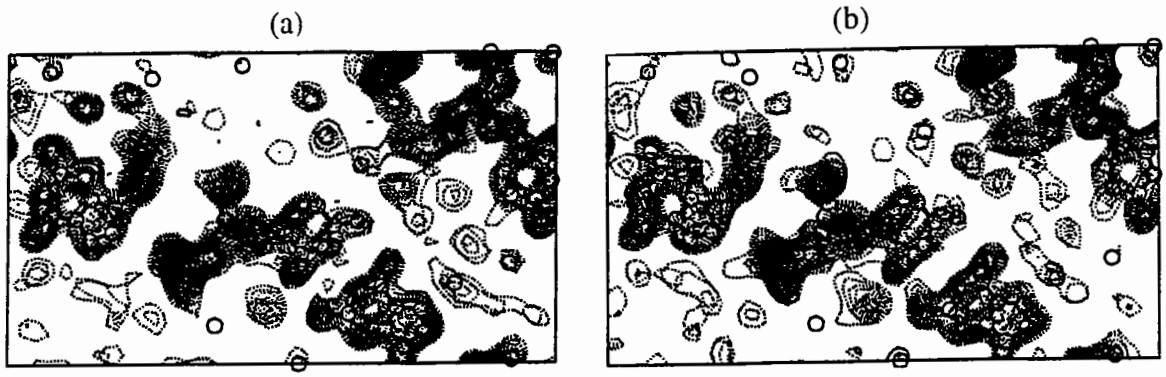


Figure 5. Density modified electron density maps contoured at 0.5 rms intervals commencing at 0.5 rms shown from 0 to 1 in x and y for a 2.8 Å thick slab in z. The final refined molecular model is superimposed so as to compare maps corresponding to a) Case 4; Density Modified λ_2, λ_4 b) Case 5; Density Modified λ_2, λ_3

Case	Description	Mean FoM_a	Mean FoM_c	Overall FoM	FOM (DM)	$<\Delta\phi>$
1	$\lambda_1, \lambda_2, \lambda_3, \lambda_4$	0.857 (2399)	0.825 (636)	0.850 (3035)	0.905 (3034)	9.36
2	$\lambda_1, \lambda_2, \lambda_3$	0.778 (2399)	0.722 (636)	0.766 (3035)	0.863 (3034)	15.66
3	$\lambda_2, \lambda_3, \lambda_4$	0.826 (2399)	0.770 (636)	0.814 (3035)	0.887 (3034)	11.23
4	λ_2, λ_4	0.771 (2399)	0.700 (636)	0.756 (3035)	0.862 (3034)	14.60
6	λ_2, λ_3	0.629 (2399)	0.460 (636)	0.593 (3035)	0.787 (3034)	31.07

Table 4. Figures of merit for the various wavelength combinations described in this section for acentric (a), centric (c) and overall cases (number of reflections in brackets) where the figure of merit (m) = cos (mean phase error).[1]

Discussion and Concluding Remarks.

λ_2 alone yields the largest f " value, as expected from theory, if not the Kronig-Kramers transform curve. Hence, the choice of two wavelengths, a reference wavelength, λ_1 or λ_4 with λ_2, whilst being the theoretical minimum number of wavelengths, also yielded the biggest Δf ' and f " differences in the diffraction data. The use of 2-λ's may be of interest when the concentration of anomalous scatterers is high in the system, and when a three or four wavelength data set collection strategy is not favourable (e.g. due to restricted beam time, and long exposure times per diffraction image are needed).

Density modification was then considered for the various wavelength scenarios. There is a special interest in the two wavelength cases which simplify the experimental and beamline needs. Key points are further discussed now. The already good map quality in the λ_2, λ_4 phasing combination was reinforced further after the DM procedure and structure solution became even easier. The isomorphous difference between data sets λ_2 and λ_3 is half that of the previous two cases mentioned above, 3.68 electrons, but this is generated by a change in wavelength of only 0.0007Å! The advantage of this is that beam position incident onto the sample would be essentially identical for the two wavelengths. The original λ_2, λ_3 phases and

[1] Compared with Peterson et al. 1996 the total number of reflections are now 2399, 636, and 3035 throughout. In the previous publication a coding error in CCP4 MLPHARE had lead to the rejection of some especially large -ve Δ_{anoms}. This coding error has been rectified in a new release of the program. There was no visible impact of this error on the map quality and comparisons no impact on the figures of merit values of the reflections that were phased, and which also constituted a large fraction of the total available in any case.

map were of only reasonable quality before DM procedures. The DM phases produced a highly interpretable map in which the structure could be easily solved. Structure solution can then even be obtained when the isomorphous signal was not optimised, due to these modification procedures. Overall, DM could perhaps be further enhanced if the electron density 'data bank' used for histogram matching actually consisted of nucleic acid density instead of protein density (which had to be used here). In essence, a key result, cases 1 to 4 become equally comparable in terms of FOM's of the phases after DM.

In Peterson *et al.* (1996), it was reasoned that dichroism effects were not evident in the f ' and f " values, in essence because the maximum induced f " and Δf ' differences were induced with respect to λ_2 in agreement with theory but somewhat unexpected. However, it was pointed out by David Templeton (pers comm), that for the two independent Br sites (A and B) in the crystallographic asymmetric unit, there did appear to be a variation between the two sites f ', and f " values which had a maximum at λ_2. Hence, at λ_2 the effect of different atomic environments of the A and B sites might explain this, in a similar way to the previously reported bromide example of Templeton and Templeton (1995), in which there was a very marked edge shift, on edge, for the parallel and perpendicular polarisation components of 0.00031Å (estimated from figure 3 of that paper). Therefore, the λ_2, λ_3, pair in the analysis would be the most to suffer if dichroism were present to a large degree. Since Figure 5 (b) shows good quality phasing and electron density map quality, it can be concluded that dichroism was not a major factor in the f ', f " values that we have encountered. Nevertheless further experiments are planned to explore the values of f ', and f " at finer $\delta\lambda$ sampling and for dichroism which must be present to some degree.

In summary, this work successfully evaluated and compared a variety of MAD experimental and computational procedures for phase improvement. It provides guidance in planning future experiments and/or new instruments, and is therefore a significant contribution to the methods of protein crystal structure determination. Aspects of the work are published in Peterson *et al.* (1996).

Acknowledgements

Thanks for discussions with J. R. Helliwell, W. N. Hunter and G. A. Leonard. Thanks also to S. J. Harrop and S. M. McSweeney for data collection assistance at Station 9.5 SRS, Daresbury. Correspondence on possible dichroism in the f' and f " values at λ_2 and λ_3 was between D. H. Templeton and J. R. Helliwell.

References

CCP4 (1994) *Acta Cryst.* **D50**, 760-763.
Sasaki, S. (1989) KEK Report 88-14, Tsukuba 305, Japan.
Okaya, Y. and Pepinsky, R. (1955) *Phys. Rev.* **98**, 1857-58.
Hoppe, W. and Jakubowski, U. (1975) In Anomalous Scattering, 437-61.
Helliwell, J. R. (1979) Daresbury study weekend, DL/SCI/R13,1-6
Helliwell, J. R. (1984) Reports on Progress in Physics **47**, 1403-1409.
Peterson, M. R. *et al.* (1996) *J. Synch. Rad.* **3**, 24-34.
Cowtan, K.(1994) Newsletter on protein crystallography, **31**, 34-38.
Templeton, D. and Templeton, L. (1995) *J. Synch. Rad.* **2**, 31-35.

Case Study: MAD phasing of desulphoredoxin, an Fe metalloprotein.

Ian D. Glover and Don Nguti.
Physics Department, Keele University, Keele, Staffs. ST5 5BG.

Introduction.

Desulphoredoxin is a small iron containing metalloprotein, consisting of a dimer of 36 residue chains each coordinating an iron atom. Data collected about the Fe absorbtion edge of 1.74Å, wavelengths being set with reference to XANES spectra recorded from a single crystal, were used to determine the positions of the anomalously scattering Fe atom and hence, using MLPHARE calculate an electron density map.

Desulphoredoxin is an Fe-S protein isolated from *Desulphovibrio gigas*, (Mouri *et al.*, 1977, Bruschi *et al.*, 1979) comprised of two 36 residue monomers, each coordinating an iron atom, which form a dimers with an M_r of 7740. Each of the monomers has four Cys residues expected to coordinate the iron atom. Most biochemical and spectroscopic evidence points to a similar coordination of iron but in relation to rubredoxin, higher symmetry in Fe binding is anticipated.

Good quality crystals of desulphoredoxin were first reported in 1980 (Seiker et al., 1980), but no suitable derivatives have been prepared. As a small metalloprotein it presented a good case for structure determination using MAD methods. With two Fe atoms in a small protein significant anomalous scattering contributions are expected, the maximal anomalous diffraction ratios (Hendrickson, 1991) of 5% and 4.8% for the absorptive and dispersive contributions respectively.

Data Collection.

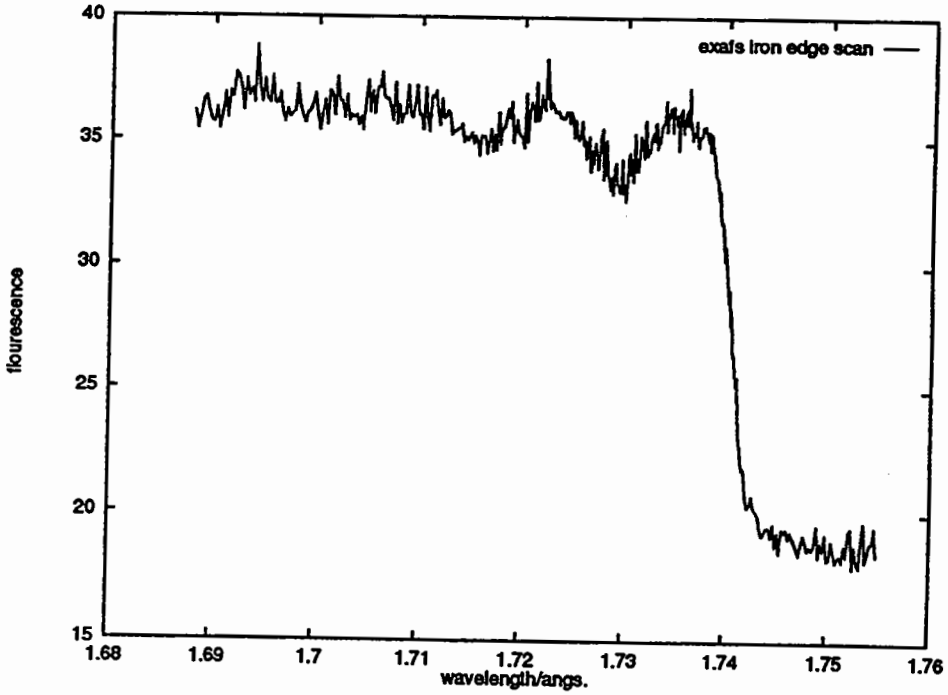
Desulphoredoxin crystallises in space group $P3_121$ (or its enantiomer) with cell dimensions a = b = 42.28Å, c = 72.46Å, and $\gamma = 120°$. The crystals grow to approximately 0.3mm in the largest dimensions and are relatively radiation stable. All data were collected on station 9.5(Thompson *et al.*, 1992) at the Daresbury SRS using an 18cm diameter MAR image plate detector and a channel cut Si(111) double crystal monochromator. MAD data were collected at four wavelengths, three close to the Fe-K edge, determined from XANES scans from a crystal and a fourth, higher resolution, dataset recorded at a remote wavelength. As the data were collected at room temperature all measurements contributing to a particular phase determination were collected as close together in time as possible. Initial calibration of the incident X-ray wavelengths was performed using the iron edge in a piece of magnetic tape, and thereafter x-ray wavelengths calculated using the monochromator angle. Due to the goniometer geometry the closest possible approach of the detector limited data collected at the longer wavelengths to approximately 3Å resolution.

Wavelength selection
The XANES spectrum were recorded from a single crystal of desulphoredoxin is shown in figure 1. The spectrum was

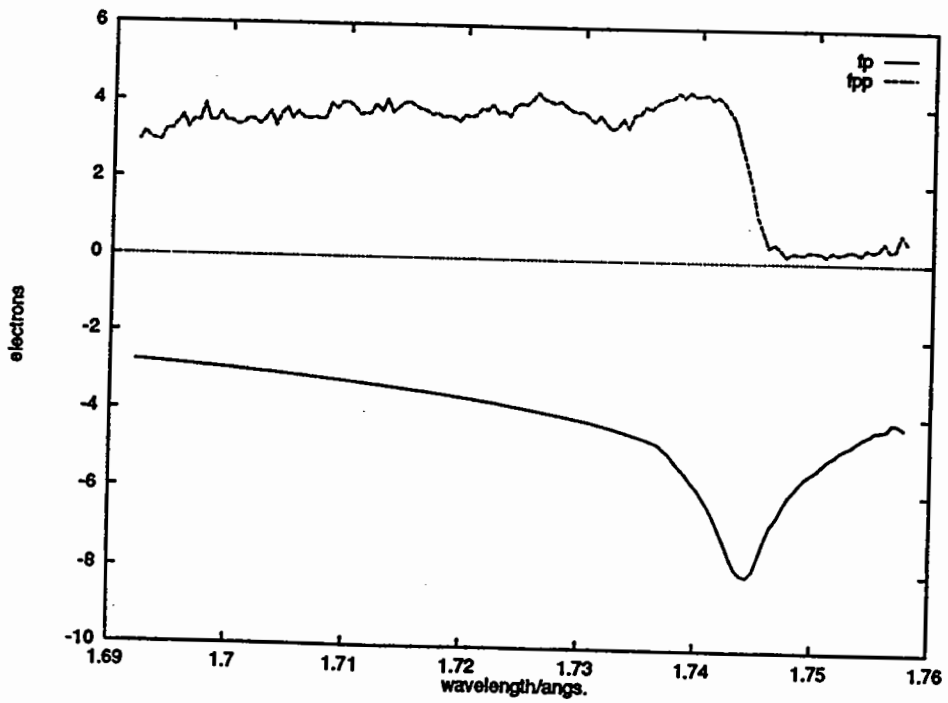
Figure 1.a) The fluoresence XANES spectrum recorded from a single crystal of desulphoredoxin using a single wire proportional counter on station 9.5 at Datesbury.

b) The transformed spectrum showing the values of f′ and f′′ in electrons as a function of incident x-ray wavelength

a)



b)



transformed using the Kramers-Kronig (Kronig & Kramers, 1928) transform, to obtain experimental values of f′ and f" (table 1). The values of the anomalous scattering coefficients were used to select the nominal wavelengths, λ_1 at 1.744Å, the first point of inflection on the f" curve, and therefore the minimum or most negative value on the f′ curve. The second wavelength, λ_2 was selected at 1.740Å, the maximum on the f" curve, this data set will yield the greatest Bijvoet or Friedel differences. The third wavelength, λ_3, was selected at 1.7285Å, remote from the edge. The fourth wavelength was collected at 0.9Å, where the incident flux on station 9.5 is significantly higher and with the same data collection geometry allowed much higher resolution data (1.8Å) to be collected. During data collection at the longer wavelengths the monochromator second crystal was detuned to avoid harmonic contamination of the incident beam.

Dataset	Wavelength (Å)	f′	f′′
λ_1	1.7444	-8.091	1.993
λ_2	1.7405	-6.096	4.337
λ_3	1.7284	-4.054	3.975
λ_4	0.9000	-1.100	2.900

Table 1. *The anomalous scattering factors for iron in desulphoredoxin at the wavelengths selected for data collection, the first three are derived from the Kramers-Kronig transform of the recorded XANES spectrum shown in figure 1.*

One crystal was used in the collection of the three near edge data sets, λ_1, λ_2 and λ_3, and a second crystal used to record the fourth, 0.9Å wavelength, λ_4 data set. The crystals were accurately aligned with the c* axis parallel to the spindle axis. In this orientation there were no mirror related reflection recorded on the same image, all mirror related reflections were recorded by inverting the crystal, i.e. recording data at ϕ and $\phi + 180^\circ$. A total of 94° ($\Delta\phi$ = 3 or 4°)of data were collected at wavelengths 1,2, and 3 and 70° ($\Delta\phi$ = 2°) at the fourth wavelength

Scaling and merging of the data.

Initial data reduction was carried out using the MOSFLM (Leslie, 1992) suite of programs after the determination of the initial orientation matrix using REFIX. Regardless of the phasing approach to be used, MADSYS or MLPHARE, once collected the data must be scaled, both within datasets and for the MAD analysis, between datasets to reduce differences due to crystal decay, absorbtion and any variation in detector response. Scale factors were calculated initially using ROTAVATA (CCP4, 1994) which calculated a single scale factor (Fox & Holmes, 1966) that is applied to all reflections in a particular batch, usually a single image. This means that symmetry related reflection falling

on consecutive batches can have very different scale factors. Since the scaling is based on all symmetry related reflections within a dataset whose intensities are expected to be equal a continuously varying scale factor may be more appropriate, such as the approach used in SCALA (P.R. Evans, this volume) where the scale factor is a continuous function of rotation angle and detector position.

1) ROTAVATA
Scale and temperature factors between batches within each dataset were initially calculated using ROTAVATA and applied using AGROVATA. The results are set out in detail in tables 2 and 3. Taking the three datasets collected at wavelengths close to the iron edge, the overall R_{SYMM} values are 14.6%, 15.2% and 13.8% respectively for the λ_1, λ_2 and λ_3 datasests which compare very unfavourably with the dataset recorded at 0.9Å wavelength. This poor scaling is clearly seen in the tables of batch scale and temperature factors calculated by ROTAVATA which show a large variation in scale factors and very significant variation in temperature factors. This poor scaling contributes to the mediocre quality of the merged data. Few batches had low R_{SYMM} values and the signal to noise, as judged by the value of $I/\sigma(I)$ was poor, averaging 3.5. Contrasting with this is the λ_4 dataset where the scale factors follow a regular progression, the biggest variations occurring either side of a beam refill and the temperature factors vary only slightly. The R_{SYMM} values are significantly lower and the signal to noise better with an average $I/\sigma(I)$ of 17.5.

This variation is seen despite the fact that the data were collected from similar sized crystals of the same shape. Furthermore it should be noted that the R_{SYMM} value of the λ_4 data at 3.05Å resolution is only 1.8%. The only difference between the data is that the λ_1, λ_2 and λ_3 data were collected at longer wavelengths and that higher absorbtion at these wavelengths is having a significant effect on the internal consistency. Data of this quality is clearly going to present problems for the subsequent MAD analysis when the expected values for the largest anomalous and dispersive diffraction ratios are 5% and 4.8% respectively.

Dataset	Wavelength (Å)	I_{MEAN}/σ	R_{SYMM}	N_{obs}
λ_1	1.7444	3.54	0.146	6857
λ_2	1.7405	3.29	0.152	6889
λ_3	1.7284	3.95	0.138	6766
λ_4	0.9000	17.46	0.034	20308

Table 2. *Summary of the overall batch symmetry R-factors for the four MAD datasets. Note that the fourth wavelength extends to 1.8Å resolution.*

	DATASET λ_1		DATASET λ_4	
BATCH	SCALE	B	SCALE	B
1	1.000	0.0	1.000	0.0
2	1.445	5.0	0.965	-0.2
3	1.970	-0.7	1.007	-0.1
4	1.499	6.0	0.937	-0.6
5	1.957	7.9	0.951	-0.6
6	2.345	-2.9	0.988	-0.7
7	2.520	3.6	0.992	-0.8
8	2.201	1.4	1.032	-0.8
9	2.505	0.3	1.060	-0.8
10	2.113	4.6	1.114	-0.9
11	2.827	5.0	1.077	-0.9
12	3.504	4.0	1.099	-1.0
13	2.134	4.4	1.092	-1.1
14	2.574	5.7	1.156	-1.4
15	2.833	6.5	1.185	-0.7
16	2.883	4.3	1.235	-1.0
17	2.925	4.1	1.222	-1.3
18	3.081	3.3	1.263	-1.1
19	3.193	1.1	1.248	-1.3
20	3.385	-0.6	1.268	-1.2
21	3.758	0.3	1.298	-1.4
22	4.021	-1.2	1.305	-1.4
23	4.434	-2.6	1.412	-1.5
24	4.811	-4.1	1.380	-1.5
25	2.453	-2.8	1.404	-1.6
26	2.918	-2.6	1.423	-1.4
27	3.153	3.6	1.445	-1.9
28			1.053	-1.6
29			1.061	-1.6
30			1.069	-2.2
31			1.066	-2.0
32			1.082	-2.3
33			1.076	-2.2
34			1.073	-2.3
35			1.089	-2.2

Table 3. **a**)*The scale and temperature factor (B) for the datasets λ_1 and λ_4, recorded at 1.7444Å and 0.900Å wavelengths calculated using the program ROTAVATA. The abrupt change in scale factors in the short wavelength data at batch 28 is due to a beam refill.*

b) Values for the λ_1 dataset after scaling using SCALA.

BATCH	SCALE	B
1	0.279	0.0
2	0.334	-0.47
3	0.416	-1.16
4	0.534	-3.72
5	0.333	-1.95
6	0.414	-1.45
7	0.594	-3.298
8	0.5624	-4.62
9	0.652	-5.37
10	0.504	-2.65
11	0.657	-2.87
12	0.878	-2.61
13	0.535	-1.61
14	0.583	-2.30
15	0.644	-2.87
16	0.705	-2.32
17	0.711	-3.09
18	0.771	-4.31
19	0.872	-4.69
20	0.970	-5.93
21	1.014	-7.12
22	1.149	-7.91
23	1.334	-9.61
24	1.506	-11.00
25	0.7614	-8.83
26	0.873	-10.21
27	0.940	-10.68

2) SCALA.
The program SCALA was used to calculate scale and temperature factors for each dataset prior to merging in AGROVATA. SCALA differs in methodology in that it calculates a three dimensional scale factor for each reflection taking into account rotation angle and its position on the detector. This methodology has significant benefits when applied to this case where sample absorbtion is anticipated to have a large effect on the internal consistency of the data. The results from scaling and merging

with SCALA/AGROVATA (tables 3 and 4) show a very significant improvement for the data collected at long wavelengths. The signal to noise ratios have increased considerably and the consistency, typically from approx 12% to 3%. The short wavelength data, however, shows very little improvement.

Table 4. a) *Summary of the overall batch symmetry R-factors for the four MAD datasets scaled using SCALA (data compared to 3.05Å resolution) and b) the merging statistics and multiplicity (Mult.) from AGROVATA.*

a)

Dataset	Wavelength (Å)	I_{MEAN}/σ	R	N_{obs}
λ_1	1.7444	9.44	0.033	5820
λ_2	1.7405	10.51	0.034	5723
λ_3	1.7284	11.14	0.034	5446
λ_4	0.9000	21.20	0.029	4447

b)

Dataset	R_{MERGE}	D_{MIN}	N_{UNIQUE}	$\%_{COMPLETE}$	Mult.
λ_1	0.045	3.05	1510	96.3	4.4
λ_2	0.048	3.04	1519	96.1	4.4
λ_3	0.036	3.03	1524	95.6	4.1
λ_4	0.029	1.78	7155	95.5	3.1

Phasing using MLPHARE.

The program MLPHARE (CCP4, 1994) is now a widely used option in the approach to the phase determination in MAD methods. Although designed for MIR phasing it can be viewed intuitively as taking one dataset as a native (with anomalous scattering) and the other datasest as derivatives, all conveniently isomorphous. In the process the real and anomalous occupancies may be refined either as relative values or as scattering factors by supplying unitary scattering factors to the lookup table, for data on an approximately absolute scale. One dataset, λ_4, was chosen as the native, it has the least significant anomalous scattering contributions, and the other three datasets scaled to this native using SCALEIT. Date were previously put on an approximately absolute scale using Wilson statistics as implemented in TRUNCATE. In common with MIR phasing the heavy atom, or in this case anomalous scattering, partial structure must first be located using Patterson maps or direct methods. In the MAD case Patterson maps may be calculated with a wide variety

of coefficients, the most important being the anolaous difference Pattersons, usually calculated exploiting the dataset with the maximum expected f" signal and the dispersive difference Patterson calculated using the differences between datasests with the largest and least f' contribution.

Patterson maps calculated using anomalous differences and dispersive differences are shown in fig. 2. The anomalous scattering partial structure was interpreted in terms of two independent Fe sites. A calculated Patterson is also shown, confirming the interpretation of the anomalous scattering partial structure.

Phasing.
MLPHARE was used to refine each of the two Fe sites independently and then used together in phasing and site refinement. Initial real occupancies were estimated in the ratios of the real, f' components of the anomalous scattering and refined against centric data before anomalous occupancies were estimated and refined. The two sites were then refined using real and anomalous occupancies simultaneously against all data to 3.05Å resolution. The overall figures of merit were 0.82 and 0.74 for centric and acentric reflections respectively.

a)

Parameter	λ_1	λ_2	λ_3
Phasing power (acentric)	2.6	2.2	2.2
(centric)	1.6	1.3	1.3
R_{CULLIS} (acentric)	0.53	0.59	0.59
(centric)	0.53	0.63	0.63
(anomalous)	0.70	0.70	0.80

b)

	λ_1	λ_2	λ_3	λ_4
SITE 1				
Real Occupancy	0.404	0.313	0.301	0.0
Anom. Occupancy	0.909	1.197	1.051	0.339
SITE 2				
Real Occupancy	0.441	0.340	0.330	0.0
Anom. Occupancy	0.862	1.086	0.962	0.327

Table 5. **a)** Summary of the statistics for the refinement of the two Fe sites in MLPHARE and **b)** real and anomalous occupancies for the two sites after refinement.

The anomalous scattering partial structure had been solved using Patterson methods and the ambiguity in the hand of the partial

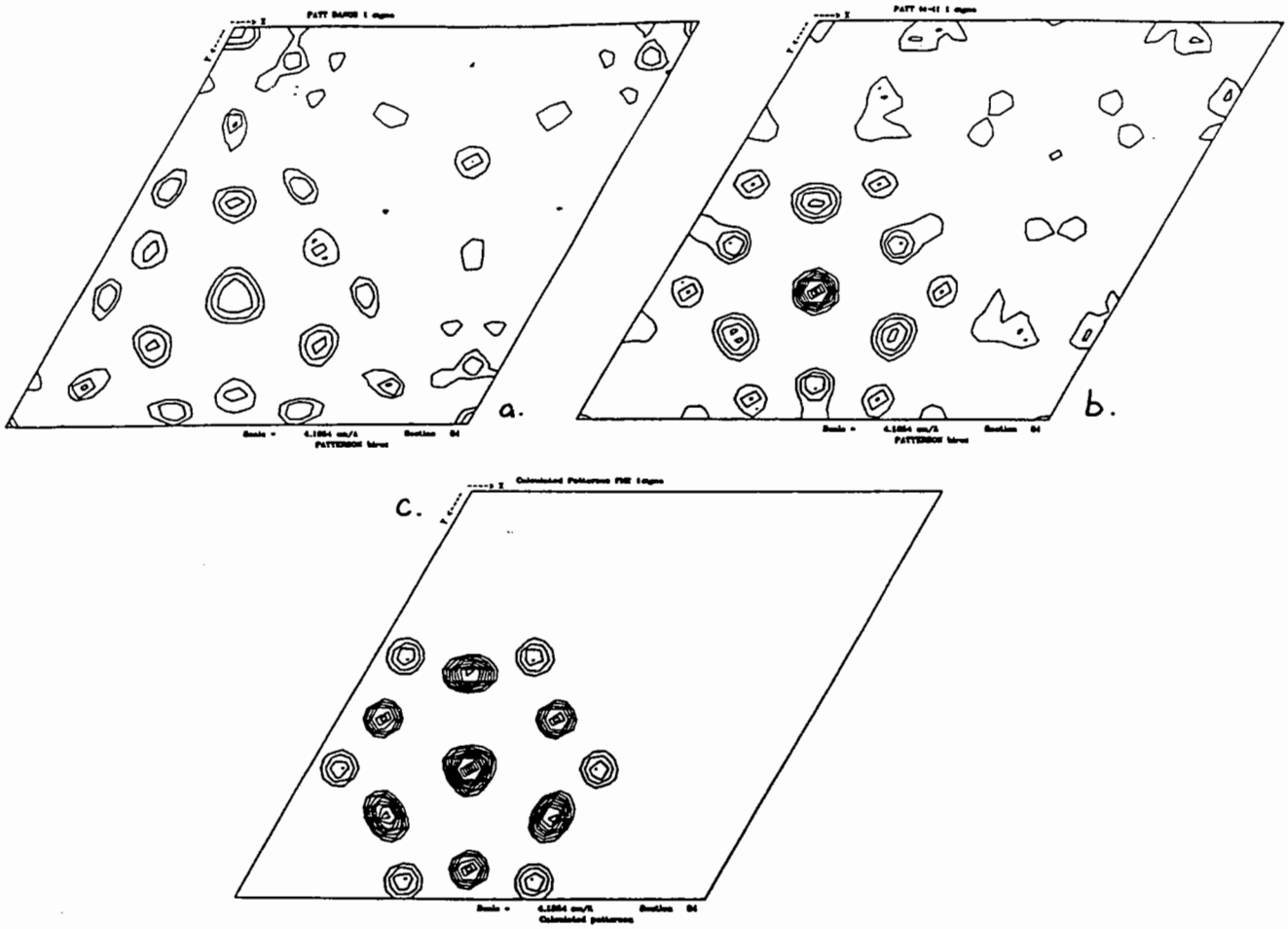
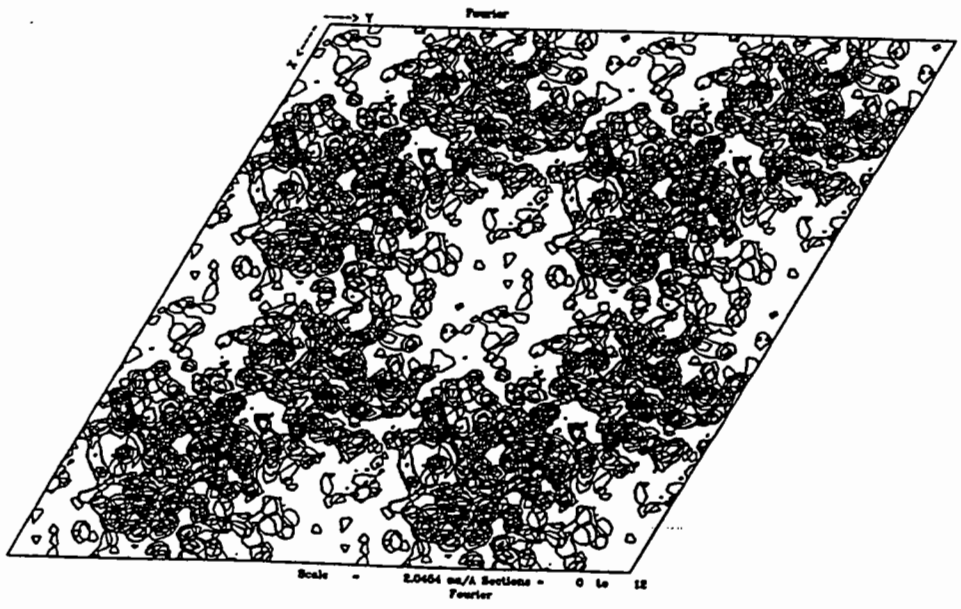


Figure 2. a) Anomalous difference Patterson map calculated using the λ_2 (maximised f'') dataset, **b)** Dispersive difference Patterson calculated using the difference in tructure factors between λ_1 and λ_4 **c)** Calculated Patterson map using refined Fe site positions.

Figure 3. The calculetd electron density map, showing 1/6th of the unit cell in c the section direction, two unit cells in each other direction.



structure was resolved by calculating the two alternate maps, in this case by calculating the maps in the alternate space groups $P3_221$ and $P3_121$. The former showed clear molecular boundaries and the iron sites could be readily located along with clear density for the iron ligands. Away from the iron sites however no clear contiguous density was observed so the map was subjected to iterative cycles of density modification, solvent flattening and histogram matching using the program DM. Map improvement was monitored using the free R flag as shown in table n, and the increase in the overall figure of merit from 0.69 to 0.81 for all data accomplished with a mean change in phase angle of 15.5^0. The calculated electron density map had improved significantly with evidence of contiguous density, showing the iron site to be in a distorted tetrahedral geometry coordinated through four cysteinyl sulphurs and clear strands of density including the short loop between Cys 9 and 12, figure 3.

Interwavelength scaling and scattering factors.

Although the MLPHARE approach to phasing has been used in this case the MADSYS suite of programs may alternatively be used. In this case the datasets , scaled using SCALA as before, were merged to give one '+' and one '-' reflection for each hkl. After local scaling (ANOSCL) the datasets recorded at each wavelength were put on the same relative, quasi-absolute scale using WVLSCL. In the course of the program the anomalous scattering factors f' and f" are refined from the crystallographic data, giving what should be analogous results to the refinement of occupancies (both real and imaginary) from MLPHARE. The results are shown in table 6, and it is clear that the refinement of the scattering factors from WVLSCL is more satisfactory than that from MLPHARE, apparently preserving the variation in the anomalous scattering contributions at values closer to those obtained from the Kramers-Kronig transform of the observed XANES spectrum from the crystal, suggesting that the inter-wavelength scaling using in this program maintains a more consistent representation of the anomalous scattering contributions in the scaled data.

Dataset & wavelength	f'	f''
λ_4	-0.31	1.11
λ_1	-8.03	2.92
λ_2	-5.47	4.03
λ_3	-5.40	3.34

Table 6. The values of the refined f' and f" contributions at the four wavelengths from WVLSCL.

Acknowledgments.
We are grateul to M. Carrondo, M. Archer and P. Matias at CTQB in Portugal for their collaboration and contribution in the work described in this report, CCLRC Daresbury for the provision of synchrotron radiation and Keele University for suport.

References.

Bruschi, M., Moura, I., LeGall, J., Xavier, A.V. & Seiker, L.C. (1979) Biochem. Biophys. Res. Comm. **90** 596-600

CCP4 (1994) Acta Cryst **D50** 760-763

Fox G.C. & Holmes, K.C. (1966) Acta Cryst. **A34** 886-889

Hendrickson, W.A. (1991) Science **254** 51-58

Kronig, R.de L. & Kramers, H.A. (1928) Z. fur Physik **28** 174

Leslie, A.G.W. (1992) In CCP4-ESF-EACMB Newsletter for Protein Crystallography. Vol 26.

Thompson, A.W., Habash, J., Harrop, S., Helliwell, J.R., Nave, C., Atkinson, P., Hasnain, S.S., Glover, I.D., Moore, P.R., Harris, N., Kinder, S. & Buffey, S. (1992) Rev. Sci. Instrum. **63** 1062-1064

MAD-DM At Elettra; A Case Study

Harold R. Powell, University of Cambridge

Introduction

MAD is an extremely demanding technique which can yield good phases from high quality crystals and data. However, in combination with DM, usable maps can be obtained from datasets which are little better than average. The present work is intended to show that provided some care is taken in the early stages of the process, it is a straightforward technique which is of particular applicability to oligonucleotide crystallography.

Here I concentrate on the aspects of the technique as I have applied it, treating the problem as a variation on MIR using MLPHARE for heavy atom refinement.

The data for the three structures discussed here were all collected at the new synchrotron in Trieste, Italy, on the protein crystallography beamline 5.2R on visits in February and May 1996; they were the first three MAD datasets that I collected, and among the first to be collected at Elettra.

The beamline at Trieste is well suited to MAD because of the easily tunable X-ray source from ~0.62Å to ~3.1Å [1]. It supplies 10^{12} to 10^{13} monochromatic photons per second; although the X-rays are not quite as well focussed as at the ESRF, it is still an extremely bright source, and the reliability and stability are very high.

It is necessary to process the diffraction data as well as possible; small errors can lead to failure of MAD-DM as it uses extremely small differences between Bijvoet pairs, which are expected to be only slightly larger than the errors in the data themselves. Without concentrating on the data processing here, it should nevertheless be remembered that any outliers flagged in the output from scaling should be noted and if the deviations are particularly large, these reflections should be omitted manually from further processing, at least until the heavy atoms have been located; Patterson maps in particular are very sensitive to the presence of rogue reflections. The SCALEIT statistics for the merging R factors of and between datasets should also be examined; if the differences between the datasets are all about the same, then location of heavy atoms is unlikely to be successful by any means.

The majority of the calculations performed in these analyses were carried out with standard CCP4 [2] programs; the data for the first example have been made available as part of a worked example on the CCP4 server. Data reduction from raw images was carried out with Denzo and Scalepack [3]; processing with other programs (e.g.MOSFLM and SCALA) will yield data of similar quality. The general scheme followed is outlined in Table 1.

Determination of the X-ray Absorption Edge

Oligonucleotides are often available in much lower quantities than proteins, and this is eqspecially true of those species containing anomalous scatterers; also, crystallization is often difficult and thus few crystals are available. However, the monomer nucleotides or even nucleo*sides* are available pure in large quantities, so in these experiments the XRF spectra were obtained for 5-bromo-2'-deoxyuridine and used to determine the appropriate wavelengths for data collection. The chemical environment of the bromine in 5-bromo-2'-deoxyuridine (the nucleoside) is very similar to that in 5-bromo-uracil (the free base) or even in an oligonucleotide containing 5-bromo-2'-deoxyuridine-5'-phosphate, hence XRF spectra obtained from these species are all extremely similar, and in general similar to that in Mark Peterson's in this Report.

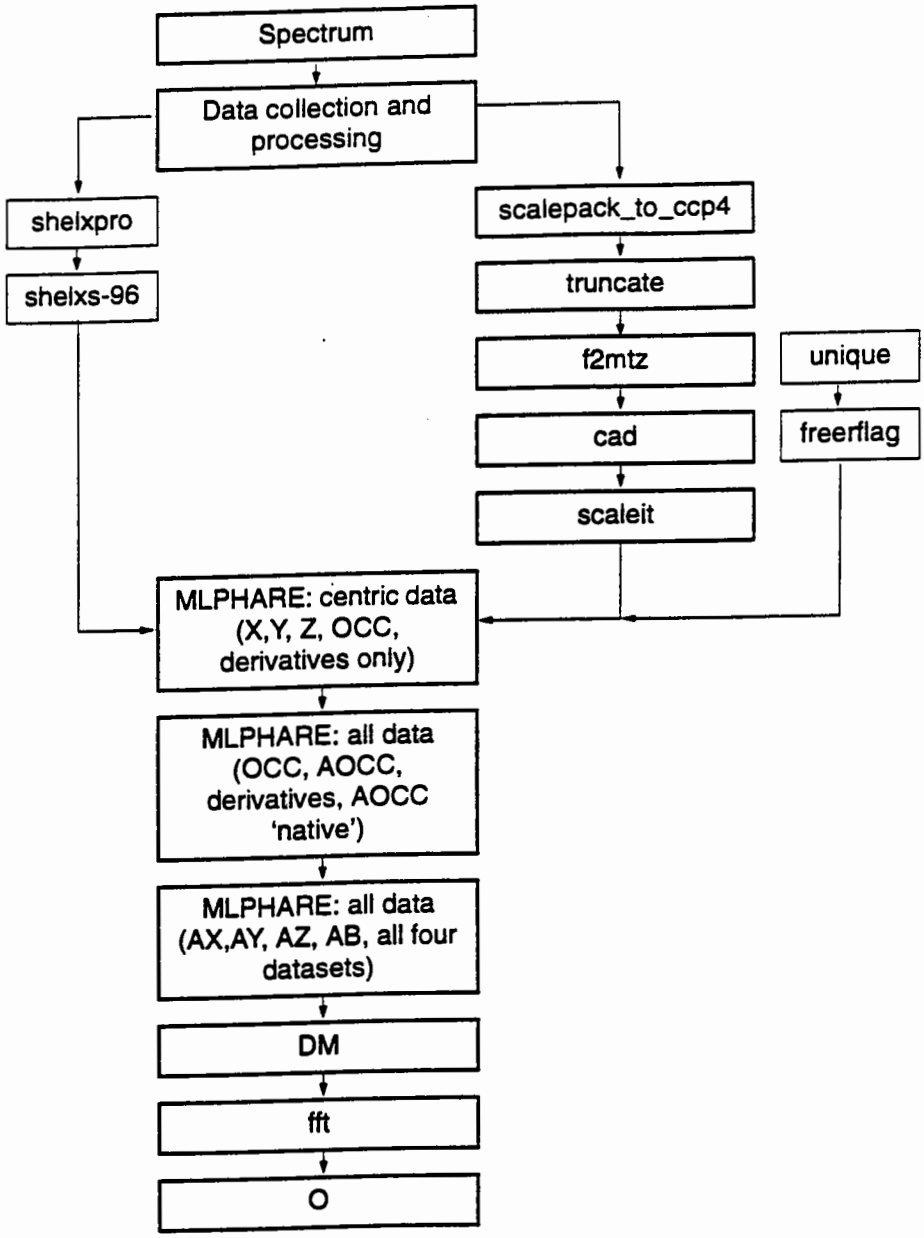


Table 1: Flowchart of general procedure

Data Collection

The most important point is that the crystals containing the anomalous scatterer must be of high quality. Small crystals help avoid problems due to absorption; as the ΔFs are very small, a poor absorption correction could mask completely any effect being exploited.

At the synchrotron, the quality of the optics is paramount; it is essential that not only is the wavelength what you think it is, but also that it can be reliably and repeatedly reselected. The X-rays must be stable for extended periods, both in terms of intensity and wavelength. Small variations can easily accumulate into significant errors.

The advent of cryo-cooling of macromolecular crystals is one of the features that has made MAD-DM data collection reasonably straightforward recently. The ability to collect several complete datasets on a single crystal has increased the chance of success of this method considerably.

Many crystallographers make life more difficult for themselves by not trying the 'oil drop' technique, but instead search for cryoprotectants that may well contribute to increased

mosaicity and reduction in data quality. Much of the degradation in crystal quality on freezing is due to surface moisture freezing rather than ice formation in the solvent channels inside the crystal [4]. The oil drop method, because it removes this surface moisture, will in many cases prevent crystal damage; it has *never* failed for me on either DNA or protein crystals. It has the added advantage that the crystal is coated in a hydrophobic layer, so it does not dry out and can be handled for some minutes outside its sitting or hanging drop.

I prefer to mount the crystal in a random orientation; this is advantageous in that the completeness of the datasets is increased over that obtainable from an aligned crystal. With a stable crystal and stable X-rays, there is little to be gained from the careful alignment of the crystal on an axis. The advantage of measuring Bijvoet pairs close together in time seems to be relatively unimportant, in DNA crystallography at least.

Location of Heavy Atoms

Atomic coordinates for the anomalous scatterers in each example were determined using the direct methods option in SHELXS-96 [5] (F^2 data from Scalepack were processed with SHELX-PRO [6] to yield anomalous ΔF values). An example of the results of this strategy for the first sample is in Table 2; it can be seen that this route should be considered as the first choice for heavy atom determination. Direct Methods seem to be more 'robust', and resistant to the presence of outliers in the data than the Patterson method, and give answers in negligible time.

The reliability of direct methods can be judged from several criteria; chief amongst these in

	dataset	x	y	z	CPU (s)
Direct methods	inflexion	0.1908	0.0150	0.1676	15.3
		0.1877	0.1562	0.1896	
	white-line	0.8088	0.4842	0.1673	17.0
		0.8105	0.3441	0.1892	
	high E offset	0.6912	0.0155	0.1678	13.9
		0.6887	0.1572	0.1903	
Patterson	inflexion	0.6895	0.5170	0.6679	347.0
		0.6859	0.6551	0.6897	
	white-line	0.8094	0.9809	0.8324	224.7
		0.8160	0.8456	0.8101:	
	high E offset	-	-	-	-

Table 2: SHELXS-96 anomalous ΔF results. Crystal 1; Space group I222, so the positions found in each solution are equivalent by space group symmetry.Times are for an SG Indigo2, R4K, 150MHz.

my view is that if the same results are obtained from each of the datasets with an anomalous contribution but not from the long wavelength offset, the answer is probably correct. Once (if!) they have failed it may be necessary to calculate Patterson maps, plot Harker sections and interpret these, but in the general case this will not be necessary. In my eagerness to look at electron density, I tend to glance over the SCALEIT statistics while the program output is scrolling past on screen, and only return to it later if difficulties have arisen.

MAD by itself will rarely provide enough phase information to be able to produce

interpretable electron density maps; some kind of additional phase extension is usually required in addition. We have used the CCP4 program DM, which applies solvent flattening and histogram matching to the data, and this leads to maps which can be of very high quality.

Structure Solutions

Sample 1

A crystal of the cyclic DNA octamer CAT-BrU-CAT-BrU, which has the 5' and 3' ends joined, was used in this study.

Four datasets were collected , one each at a long wavelength offset, at the inflexion point, the white-line maximum and a short wavelength offset (Table 3). Processing of these data showed that they were reasonably complete, and using Scalepack's 'linear R-factor' and 'square R-factor' as guides, they were of reasonable but not exceptional quality.

Sequence	CATBrUCATBrU			
Crystal System	orthorhombic		Space Group	I222
Cell dimensions	a = 22.627	b = 26.002	c = 70.045	
Crystal to detector	120mm		Frames	60 x 3.0°
Max. resolution	~ 1.5Å			
Dataset (Å)	0.8993	0.9198	0.92054	0.9334
Total data	36470	36152	36277	36223
Unique data	3552	3530	3529	3527
R_{merg} (1)	0.066	0.071	0.060	0.040
R_{merg} (2)	0.058	0.074	0.067	0.048
Completeness (%):	99.2	98.6	98.5	98.6

Table 3: Data collection statistics for Sample 1.

Direct methods gave two possible bromine positions (see Table2), which was expected from the unit cell dimensions and space group.

Heavy atom refinement according to the scheme in Table 1 gave the results in Table 4. It is worth spending a little time looking at the various figure of quality produced. For the *Figures of Merit*, values greater than 0.6 can be considered encouraging, and if > 0.8, the problem can be considered well on the way to being solved. The *Cullis R-factors*, which are calculated for each derivative should become smaller for a correct answer; final values of $R_{Cull}(cen) < 0.9$ and $R_{Cull}(acen) < 0.6$ for the white-line maximum and short wavelength offset datasets should be seen as encouraging, and an $R_{cull}(ano) < 0.5$ for the datasets with an anomalous contribution seems a good indicator that the correct answer is being approached

Another measure of the correctness of the refinement process can be found by inspection of the refined values of Occ and AOcc (the real and anomalous occupancies), as they should be proportional to $\Delta f''$ and f'' respectively; even in the best collected datasets, there will be deviations from these relationships which reflect the fact that datasets have not been collected exactly at the inflexion point and whiteline maximum (Table 5). However, as long as the proportions are roughly correct, it is important not to worry too much

The main thing to be remembered about the various measures of quality associated with heavy atom refinement is that they are *only guides*; the best, and only *sure* way of knowing that the MAD-DM process has been successful is when calculated electron density is studied and model fitting can begin.

			(+x, +y, +z)	(-x, -y, -z)
ML-PHARE	Totals	FoM (ace)	0.8521	0.8524
		FoM (cen)	0.6646	0.6717
		FoM (all)	0.8164	0.8179
	Deriv #1	Cullis R (ace)	0.58	0.57
		Cullis R (cen)	0.61	0.60
		Cullis R (ano)	0.89	0.89
	Deriv #2	Cullis R (ace)	0.85	0.85
		Cullis R (cen)	0.87	0.86
		Cullis R (ano)	0.30	0.30
	Deriv #3	Cullis R (ace)	0.51	0.50
		Cullis R (cen)	0.54	0.54
		Cullis R (ano)	0.36	0.36
	"Native"	Cullis R (ace)	1.46	1.46
		Cullis R (cen)	1.00	1.00
		Cullis R (ano)	0.35	0.34
DM	FoM-DM		0.881	0.886
	R_{free}		0.557	0.502
	Real Space R_{free}		0.349	0.202

Table 4: Selected MLPHARE and DM statistics for Sample 1.

DM was run in a more-or-less default mode of solvent flattening with histogram matching; the only required information from the crystallographer is a reasonable estimate of the solvent fraction of the unit cell. The figure that seems most informative from DM is the *Real Space* Free R; this can give good information on the correct hand of the structure (which cannot be obtained from MLPHARE), and is also a further indication that the whole process has worked. Note that it is only after processing with DM that there is a significant difference between the two hands, and it is apparent in this case that originally the wrong hand was chosen. Phases are also calculated for many reflections unphased in previous steps, and this phase extension is important in being able to calculate electron density. .

	λ (Å)	0.8993	0.91980	0.92054*	0.9331
Occ	Δf' (exp)	5.372	1.279	0	4.489
	Br(1)	0.154	0.043	0	0.129
	Br(2)	0.186	0.058	0	0.158
AOcc	f" (exp)	3.641	3.826	2.167	~0.5
	Br(1)	3.028	3.098	2.998	0.461
	Br(2)	3.306	3.670	3.240	0.491

Table 5: Refined occupancies for bromine atoms: Crystal 1:(Occ ∝Δf', AOcc ∝ f")

* inflexion point dataset (reference)

The phases calculated by DM can be used directly by FFT to produce an F(obs) map which can be viewed on a graphics workstation after suitable translation.Electron density showing obvious base stacking and in the region of an A-T base pair can be seen in Figure 1

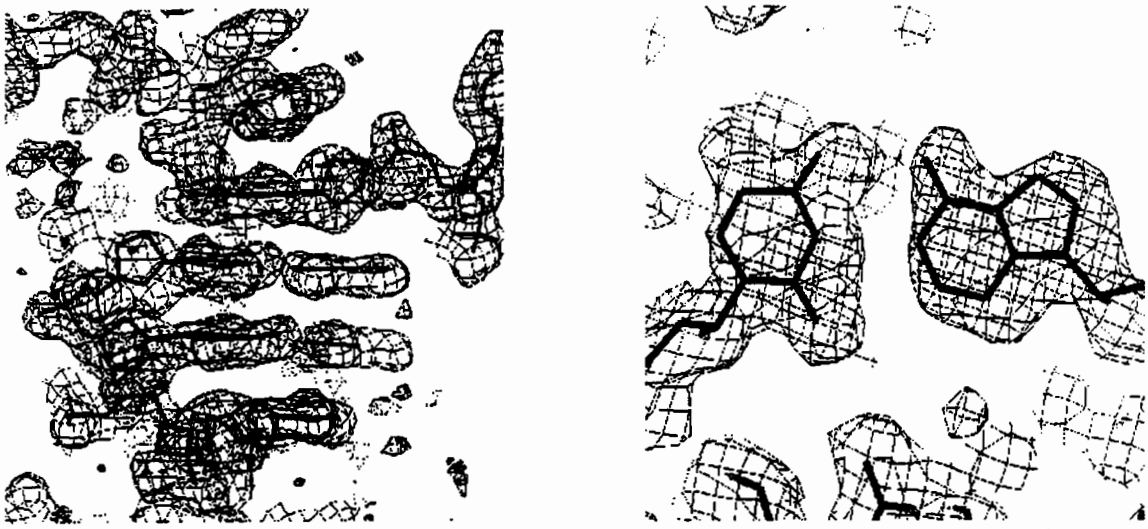


Figure 1: electron density plots showing (i) base stacking and (ii) a T-A base pair.

Sample 2

The second sample was isomorphous with the native structure solved elsewhere. In this case, instead of four datasets, seven were collected; the extra three were collected with wavelengths at -1eV (#5), +1eV (#6) and +2eV (#7) from the measured inflexion point of the nucleotide. This experiment was intended to ensure that we had a dataset as close as possible to the true inflexion point of the oligonucleotide. As it turned out, the real value was between the measured IP and #5.

Sequence	cyclic CATBrUCATBrU						
Crystal System	orthorhombic		Space Group		$P2_12_12_1$		
Cell dimensions (Å)	a = 22.80		b = 27.86		c = 55.06		
Crystal to detector	120mm		Frames	35 @ 3° (long λ offset 27 @ 3°)			
Max. resolution	~1.5 Å						
Dataset (Å)	0.92155	0.92079	0.9003	0.9334	0.92162	0.92148	0.92141
Total data	23555	23707	23409	18392	23170	23809	23846
Unique data	9753	9789	9597	8834	9610	9837	9865
R_{merg} (1)	0.070	0.069	0.072	0.059	0.064	0.085	0.091
R_{merg} (2)	0.083	0.082	0.098	0.073	0.079	0.096	0.108
Completeness (%)	85.0	85.1	83.6	76.8	83.7	85.7	85.7

Table 6:Data collection statistics for Sample 2.

The data collected were not of the same quality as for Crystal 1 (Table 2), but Direct Methods revealed the presence of four heavy atoms in the asymmetric unit, with roughly the same coordinates as those for the four non-base-paired thymine methyl groups in the native.

DM	FoM-DM	0.924	0.922
	R_{free}	0.434	0.440
	Real Space R_{free}	0.269	0.255

Table 7: Selected DM statistics for Sample 2.

Examination of an F(obs) map in the region of an A-T base pair (Figure 2) reveals that the electron density is interpretable, but less easily than for sample 1. However, with some work the molecule could be successfully fitted even without prior knowledge of the correct

structure.

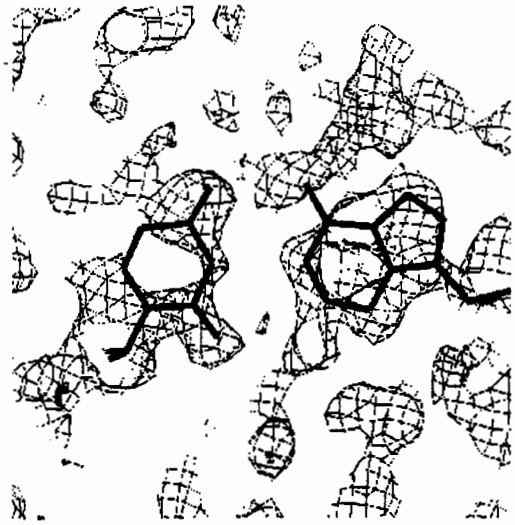


Figure 2: Electron Density in the region of an A-T base pair showing the lower quality of this solution compared to Sample 1.

Sample 3

This work is part of an ongoing project led by Dr Christine Cardin of Reading University, and I was in the fortunate position of helping her in this study. The crystal used was grown by Dr Adrienne Adams of Trinity College, Dublin.

The whole analysis from raw images to first electron density map took about one and a half working days, and only took that long because we took our time over it!.

The data collected appeared comparable at all stages of the processing to those from sample 1.

Direct methods found one heavy atom in the asymmetric unit. Heavy atom refinement proceeded smoothly, and examination of the measures of quality from DM show that there is little to choose between the correct and incorrect hand for this structure. However, note that the Real Space R_{free} values for both hands are far worse than for the previous two samples; this should emphasize the point that all the numbers output by the programs should only be taken as guides!. .

Sequence	ACGTACG-BrU			
Crystal System	tetragonal	Space Group		$P4_32_12$
Cell dimensions	a = 41.991	c = 25.301		
Crystal to detector	120mm	Frames	30 @ 3° (*15 @ 3°)	
Max. resolution	~ 1.6Å			
Dataset	0.9344*	0.9216	0.9208	0.9003
Total data	12171	25725	26727	28390
Unique data	4276	5672	5867	6301
Rmerg (1)	0.051	0.071	0.058	0.061
Rmerg (2)	0.074	0.105	0.100	0.096
Completeness (%):	74.3	98.7	98.6	98.2

Table 8: Data Collection Statistics for Sample 3.

Electron density in an F(obs) map revealed that the solution from DM with the worse statistics was actually correct. Figure 4(i) shows the spectactularly good density for the oligonucleotide revealed in the first map calculated; it is not necessary to include a model of the structure to see in Figure 4(ii) the positions of the Br in a BrU-A base pair and most of the base atoms as well

Conclusions

	(+x, +y, +z)	(-x, -y, -z)
FoM-DM	0.871	0.868
R_{free}	0.526	0.541
Real Space R_{free}	0.354	0.399

Table 9: Selected DM statistics for Sample3.

The take-home message from this work is that the facilities to collect data for a MAD-DM experiment and the programs to process these data are available now. MAD-DM is straightforward provided that data are collected carefully from the best available crystals; it is capable of giving excellent electron density which allows rapid and relatively easy structure building. The comparison of F(obs) maps for crystals 1 and 3 shows that it is necessary to examine the electron density rather than rely on the statistics; there can be a marked difference even between apparently similar data.

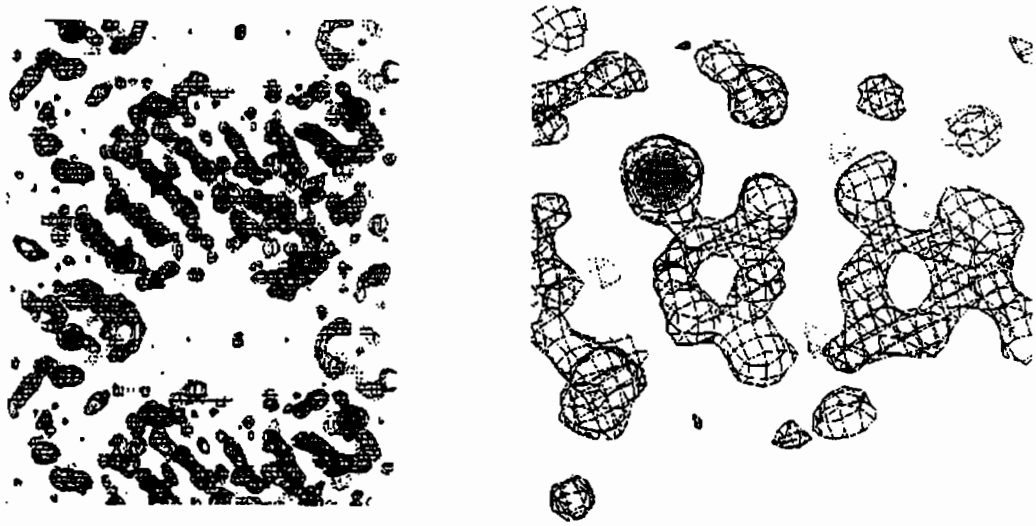


Figure 3: Electron density for Sample 3 (i) showing obvious base stacking and (ii) in the region of a BrU-A base pair.

Acknowledgements

Eleanor Dodson (York), Christine Cardin, Alan Todd (Reading), Adrienne Adams (Dublin), Stephen Salisbury, Sarah Wilson (CCDC) and the CCDC and University Library, Cambridge.

References

[1] see the WWW page http://elettra.trieste.it

[2] CCP4 (1996) Colaborative Computational Project Number 4. The CCP4 Suite: programs for protein crystallography. Acta Crystallogr. D50, 760 - 763.

[3] Z. Otwinowski, Denzo and Scalepack, film processing programs for macromolecular crystallography. Yale University, New Haven, 1995.

[4] see, for example, http://www-structure.llnl.gov/Xray/cryo-notes/Cryonotes.html

[5] SHELXS-96, G.M.Sheldrick, Universität Göttingen, 1996

[6] SHELXPRO, G.M.Sheldrick, Universität Göttingen, 1996

SCALING OF MAD DATA

Philip R. Evans, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH

The integrated intensities from any data collection experiment are not all on the same scale, because of various systematic differences in the collection procedure. It is the task of the "data reduction" protocol to place all observations on a common scale, to detect and reject outliers (reflections for which the data collection has gone badly wrong), and to produce a list of |F| and σ(|F|) for the structure determination. There are some special considerations in the optimum treatment of data intended for MAD phasing, in that we want very accurate *differences* between amplitudes, for the anomalous differences $\Delta F_{\pm}$ and the dispersive differences ΔF_{λ}, rather than the most accurate absolute values. This means a difference both in data collection strategy, designing the experiment to minimize the systematic errors in the differences, and in the scaling strategy, in which *relative* scaling can reduce, though probably not eliminate, the systematic errors. In the MAD phasing method, we need accurate differences because the small signal is easily swamped by systematic errors, and we also need to be careful about eliminating outliers, since a small number of spurious large differences can confuse both Patterson and direct methods of locating the anomalous diffracting centres.

To aid designing data collection and scaling strategies, it is helpful to enumerate the reasons for the observed intensities not being on the same scale. These factors can be roughly divided into those that can be in principle calculated, and those that must be determined empirically from the data.

(1) Calculable scale factors

• Lorentz factor – this is uncertain close to the rotation axis, but is not normally a problem

• Polarization – this may be uncertain for synchrotron radiation, but the error is small

• Corrections arising from deficiencies in the integration program – if the geometrical parameters used by the integration program are inaccurate, the prediction of which spots are partially recorded will also be inaccurate. The estimated partiality may be improved by post-refinement (eg in Scalepack or Mosflm)

• Different truncation of the tails of reflections caused by diffuse scattering – partially recorded reflections are measured over at least twice the rotation range of fully recorded reflections, so if the spots have long tails in the rotation direction, more of the tails will be included in partials than in fulls (the TAILS correction in Scala is an attempt to correct for this, see appendix below).

(2) Empirical scale factors

These are usually subsumed into general scaling.

• Change of incident beam intensity – mainly on synchrotrons

• Change of detector sensitivity – the variation of sensitivity across the detector is best determined in a separate calibration (flood-field correction), but the overall "sensitivity" may be taken up in the scaling, particularly for film or off-line image plates

• Different crystals

• Illuminated volume – if the crystal is larger than the beam. This is indistinguishable from absorption in the incident beam

• Absorption – less of a problem at short wavelengths, but hard to correct for satisfactorily

• Radiation damage – serious on unfrozen crystals

• Wavelength-dependent factors – mainly for the Laue method

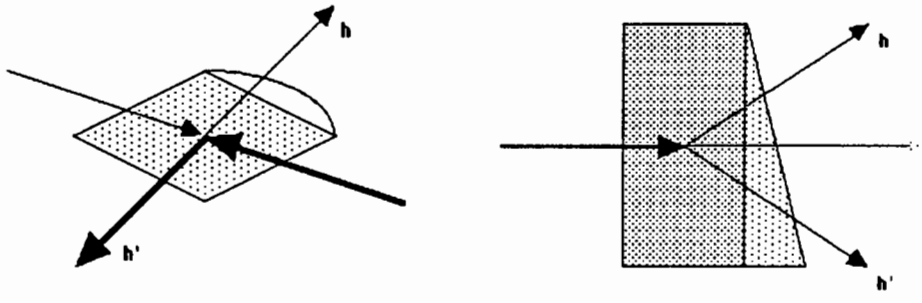
It is possible to design the data collection strategy for MAD data collection such that many of these systematic errors can be made equal, so that they cancel out in the dispersive and anomalous differences. Note that this is the opposite of the optimum collection strategy for data intended for structure refinement, when ideally we should try to maximize the systematic differences between observations, so that the scaling procedure can determine the different corrections for different parts of the data, or least can average out the systematic errors.

(a) Dispersive differences (ie between different wavelengths) – measurements of the same reflection at different wavelengths will normally be made in the same way, so that the systematic arrors should be the same. The main difference is that they are necessarily measured at different times: radiation damage is the only difficult time-dependent scale, hence the great advantage of using frozen crystals. On unfrozen crystals, the strategy must be to collect different wavelengths close together in time (eg as images interleaved at each wavelength).

(b) Anomalous differences – it is not possible to collect I_+ and I_- in exactly the same way, on the same area of the detector. The most difficult correction is absorption, other corrections are likely to be the same. Absorption is a serious problem at the longer-wavelength edges (eg Fe), less of a problem for Se or Br edges.

There are two ways of minimizing the absorption differences, though neither will eliminate the problem:-

(i) inverse beam method – measure reflections at ϕ and ϕ+180°. This inverts the direction of the incident and diffracted beams. The absorption will only be the same if the crystal and its mount have a centre of symmetry



(i) (ii)

(ii) rotate the crystal about a two-fold axis, and collect Bijvoet pairs (eg hkl, hk-l for a crystal rotating around the c axis) on the same image. This requires the crystal to be aligned about an axis, at least approximately. The absorption is only the same if the crystal and its mount have a plane of symmetry perpendicular to the rotation axis.

To correct for absorption differences between Bijvoet pairs, the scaling model must be able to apply a different scale to I_+ and to I_-, so the scaling model must be anisotropic

and non-centrosymmetric. Suitable functions are 3-dimensional smoothed scales (local scales) and 3D functions such as spherical harmonics. The functions must not vary too much locally, otherwise the real differences will be scaled out.

How well are scale factors determined?

The problem with 3-dimensional scale functions is that they are typically ill-determined by the observed data. The empirical correction factors listed above may be divided into two categories:-

1) functions of the incident beam direction (illuminated volume, absorption in the incident beam) or of time,which is equivalent (beam intensity, radiation damage). With any area detector, these functions are well-determined, since many reflections are measured at the same time for each direction.

2) functions of the diffracted beam direction (absorption, radial dependence of radiation damage). These functions are poorly determined, since there are relatively few observations in each direction. The corrections are well-determined only:-

(a) with high symmetry (thus high redundancy of measurements made under different conditions)

(b) collection by rotation about more than one axis (to measure equivalent reflections with different beam paths in the crystals)

(c) scaling relative to a reference set – this gives relative rather than absolute scales, but is useful to reduce systematic errors in differences, as is required for MAD data.

A relative-scaling protocol

The following suggested protocol for scaling MAD data uses a reference data set, which provides a an anchor for the scaling parameters. Note that in the reference, I_+ and I_- are averaged, so that in the real datasets, systematic bias in the anomalous difference will be reduced (the mean $\Delta I_\pm$ should be zero). A similar protocol is also useful for scaling heavy-atom derivatives using the native dataset as reference, in the MIR method.

1. Choose reference set: this should be (in order of importance)
   (a) the most complete
   (b) the most accurate
   (c) remote from the anomalous edge

2. Scale and merge the reference set, merging I_+ and I_-, to get a unique set of merged intensities I_{ref}

3. Sort the reference set together with all unmerged data, for all wavelengths (including the set used as reference, if this is to be used in phasing).

4. scale all data together, perhaps in two passes

   (a) batch scaling (scale k & B-factor for each image ("batch")) to remove discontinuities between images. If all images are reliably on a similar scale with no discontinuities between images (stable source, collected by dose etc), this step may be omitted.

   (b) smooth scaling using a 3-dimensional anisotropic or local scaling model. This may be parameterized in camera space (x,y,ϕ or beam directions) or in crystal space (h,k,l). An example in Scala would be SCALES ROTATION SPACING 10 DETECTOR 3 3.

5. split out each wavelength, either averaging repeated and symmetry-related observations, or keeping them separate (depending on whether the phasing strategy uses merged or unmerged data)

Various programs allow scaling of this type, eg the XDS package (Kabsch 1988)), the CCP4 program SCALA (which took some inspiration from the Kabsch method), and X-GEN (Howard)

Results

Trials with a Se-methionine data set (thanks to Richard Pauptit) and a Br-uridine DNA set (thanks to Harry Powell and Christine Cardin) showed a small but significant improvement using this protocol, compared to scaling each dataset separately. The improvement is presumably only small because absorption, which causes the most serious systematic errors is small at the Se and Br edges. Absorption is much more serious at longer wavelength, so for MAD measurements on for example the Fe edge this scaling method would produce a much larger gain. However, since the MAD signal is so small, even a small improvement can make the difference between success and failure, and a small reduction in the difference between observations (as measured by reduced dispesive and anomalous differences) may make a substantial difference in phasing.

Appendix

A simple correction for the bias between fully-recorded and partially-recorded reflections caused by diffuse scattering

Many protein crystals show marked diffuse scattering, which is seen as long tails on spots in the "phi" direction, so that reflections often appear on the image before they are predicted. If the mosaicity is increased to include these tails, too many reflections may be rejected as overlaps. Fully-recorded reflections are integrated over a smaller phi width than partials, so more of the tails are chopped off for fulls than for partials. This leads to the typical negative partial bias, with partials systematically larger than equivalent fulls.

A correction has been introduced into SCALA which attempts to correct for the different truncation of diffuse scattering tails, using a simple model of thermal diffuse scattering, expressed as 2 or 3 parameters over the whole data set. This implementation does not attempt to correct for diffuse scattering itself, only for the different effect on fulls and partials. This correction reduces the partial bias substantially, and seems to improve the data generally, though sometimes the parameter refinement can be a little unstable.

<u>The method</u>

This algorithm was inspired by the correction described by Blessing (1987), but in his case full profiles of the diffraction spots were analysed to determine the diffuse scattering contribution. Data collected with relatively coarse rotation slices do not provide enough information to do this, and the typical crowded diffraction patterns of macromolecule crystals make it harder to extract full profiles, since the spots may overlap.

1. The thermal diffuse scattering contribution to the integrated intensity is proportional to the Bragg intensity J. If the complete profile is measured, the measured intensity I, including diffuse scatter is given by

$$I = J\,(1+\alpha)$$

where α is a proportionality constant

2. The proportionality constant α varies with resolution, and may be anisotropic. At present an isotropic model is implemented in Scala

$$\alpha = \alpha_0 + s^2 \alpha_1$$

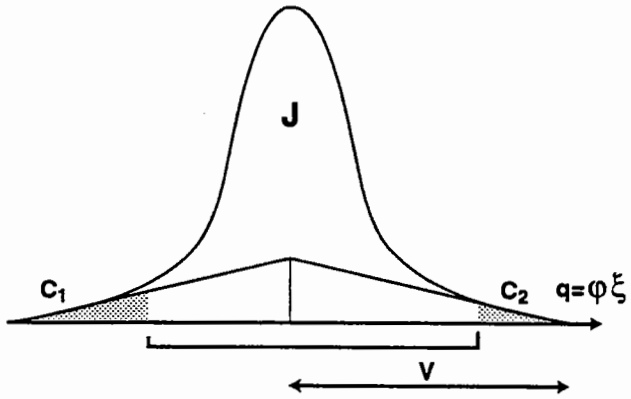
where $s^2 = (\sin \theta / \lambda)^2$, and α_0 is normally = 0. α_0 and α_1 are refinable parameters.

3. The width of the thermal diffuse scattering peak is assumed to be constant in reciprocal space, = v, a refinable parameter. The distance in reciprocal space travelled by a reflection rotated by an angle $\Delta\phi$ at a radius ξ from the rotation axis is given by

$$q = \xi \Delta\phi$$

4. The profile of the diffuse scattering peak is modelled as a triangle, with width v (in the reciprocal space coordinate q), and height h, where h is a function of α, since the area of the triangle is I-J = h v = J α, hence h = J α / v

5. If the scan width of an observation, including all parts of a partially recorded reflection, is less than 2v, the tails of the diffuse scattering peak may be truncated, clipping off areas C_1 and C_2 (≥ 0) (see figure). These areas may be calculated from the rotation angles at the start of the scan (the beginning of the first image contributing to the observation), the centre of the reflection (the predicted angle), and the end of the scan (the end of the last image contributing to the observation).



6. The correction factor for diffuse scattering if the full profile were measured would be given by

$$J = I / (1 + \alpha)$$

For the truncated profile

$$J = I / (1 + \alpha (1 - C_1 - C_2))$$

where C_1 and C_2 are expressed as fractions of the complete area of the triangle (h v)

Since I do not trust this simple formulation to correct properly for diffuse scattering, the correction used is

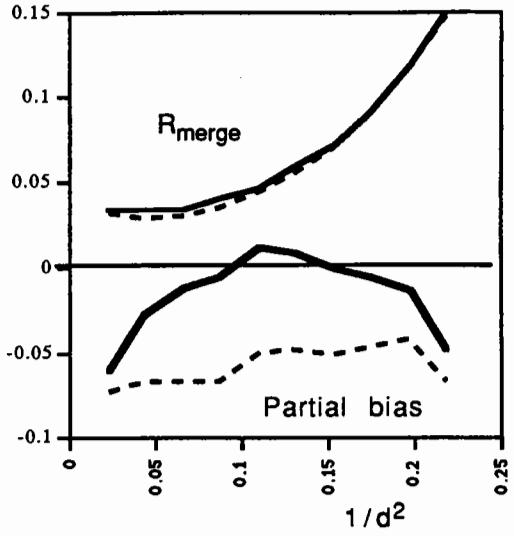
$$J = I (1 + \alpha) / (1 + \alpha (1 - C_1 - C_2))$$

This corrects for the different truncation of the peak for different spots, particularly the difference between observations made over 1, 2 ,3 etc images, but not for the diffuse scattering itself.

The parameters refined are α_0, α_1 and v (note that C_1 and C_2 are functions of v), though normally α_0 is fixed at 0.0.

Results

Application of the Tails correction to to datasets with visible diffuse scattering typically has a dramatic improvement on the partial bias, ie the systematic difference between fully recorded and partially recorded reflections (see figure), and often a significant improvement in R_{merge}. The correction is not well-determined if the diffuse scattering is small, nor if the mosaicity is badly underestimated in the integration process: in these cases, the parameters can take on unrealistic (eg negative) values.



Example of the improvement in partial bias (lower curves) and R_{merge} (upper curves), plotted against resolution. Solid lines: with Tails correction, dashed lines: without correction. The partial bias is $\Sigma_h(\langle I_{full}\rangle - I_{partial}) / \Sigma_h \langle I_{full}\rangle$, where the summations are over all reflections for which there are both fulls and partials, $\langle I_{full}\rangle$ is the mean of all the fully recorded observations of the reflection, and $I_{partial}$ is a summed partial observation

Acknowledgements: I thank Richard Pauptit, Harry Powell and Christine Cardin for the loan of datasets, Gérard Bricogne for helpful discussions on statistics, and Eric de la Fortelle for help in running SHARP.

References

Blessing, R.H. (1987), Data Reduction and Error analysis for Accurate Single Crystal Diffraction Intensities, Cryst. Rev. **1**, 3-58

Howard, A.J. X-GEN documentation

Kabsch, W. (1988) Evaluation of single-crystal X-ray diffraction data from a position sensitive detector, J. Appl. Cryst. **21**, 916-924

MASC:
A Combination of Multiple-Wavelength Anomalous Diffraction & Contrast Variation

W. Shepard[†], M. Ramin[†], R. Kahn[§], & R. Fourme[†],

[†] *LURE, B.209D, Université Paris Sud, 91405 Orsay, France.*
[§] *Institut de Biologie Structurale, 41 Avenue des Martyrs, 38027 Grenoble, France.*

1. Introduction

Contrast variation methods have primarily been applied and developed in low angle scattering studies as a means of extracting information on the shape of a particle dispersed in a solvent medium (for a review see Williams et al., 1994). This method deals with the changes invoked in the scattered intensities of a small angle scattering experiment when the density of the particle is varied relative to its solvent medium. The difference between the particle and solvent densities is defined as the "contrast" (Stuhrmann & Kirste, 1965; Ibel & Stuhrmann, 1975). The term "density" in this context refers to the electronic density in an X-ray scattering experiment, the isotopic substitution ratio (H/D) in a neutron scattering experiment, or any other physical density which scatters the incident beam.

Contrast variation techniques can be extended to macromolecular crystal systems since such crystals typically consist of 30-70% solvent, which is a phase of rapidly interchanging molecules. Bragg & Perutz (1952) applied such methods to a haemoglobin crystal and observed changes in the intensites of low resolution X-ray reflections after altering the electronic density of the mother liquor. In particular, they related these changes to the Fourier transform of the solvent accessible regions of the crystal. In other words, the data from a contrast variation series provides information on the macromolecular envelope.

Others have since applied contrast variation techniques in either X-ray or neutron diffraction experiments to glean low resolution structures from macromolecular crystals (Harrison, 1969; Jack, Harrison & Crowther, 1975; Moras et al., 1983; Roth et al., 1984; Bentley et al., 1984; Podjarny et al., 1987). In particular, Carter et al. (1990) used a formalism which separated the diffraction effects of the molecular envelope and the internal fluctuations (Bricogne, unpublished) in the direct phase determination of the molecular envelope of tryptophanyl-tRNA synthetase.

Anomalous dispersion has also been employed in small angle scattering experiments to produce contrast variation. Examples on biological systems are the harnessing of the iron K-edge in ferritin (Stuhrman, 1980) and the phosphorous K-edge in ribosomes (Hütsch, 1993). In crystallography, the use of anomalous scattering effects from the solvent has been suggested by Wyckoff and others where it could be used as a supplement to a standard contrast variation series (Dumas, 1988; Crumley, 1989; and Carter et al., 1990). However, in these cases, the anomalous scattering was still restricted at a single wavelength. The possibility of exploiting the

full potential of anomalous scattering at several wavelengths was orginally put forward by Bricogne (1993).

2. Theoretical principles of contrast variation

Here only an outline of the theoretical principles will be given. Readers wishing for a fuller account are referred to Fourme et al. (1995). The starting point of what we call MASC (Multiple-wavelength Anomalous Solvent Contrast) is the basic principles of contrast variation, where the macromolecular crystal lattice is assumed to be biphasic: one region of the unit cell is occupied by the macromolecule, domain U (Figure 1a), and the other, domain V-U (Figure 1b), is occupied by the solvent which is in a liquid-like state of rapid exchange. The domain containing the macromolecule is presumed to be ordered, whereas the solvent regions are presumed to be completely disordered.

We define ρ_s as the electronic density of the solvent volume, which is constant since this region is flat and featureless. $G_U(\mathbf{h})$ is the Fourier transform of the indicator function $\chi_U(\mathbf{r})$, defined as equal to 1 inside the volume U and 0 elsewhere (Bricogne, 1974). It should be noted that $-G_U(\mathbf{h}) = G_{V-U}(\mathbf{h})$ when $\mathbf{h}\neq\mathbf{0}$, such that $G_{V-U}(\mathbf{h})$ is the Fourier transform of the complementary indicator function $\chi_{V-U}(\mathbf{r})$ which corresponds to the region occupied by the solvent. The total structure factor, $F(\mathbf{h})$, can be written as the sum of two components: one from the ordered regions of the crystal, $F_p(\mathbf{h})$, and the other from the solvent, $\rho_s G_{V-U}(\mathbf{h})$. These two components are related since the volume occupied by either the marcomolecule or the solvent are by definition mutually exclusive.

$$F(\mathbf{h}) = F_p(\mathbf{h}) - \rho_s G_U(\mathbf{h})$$

$F_p(\mathbf{h})$ is also the Fourier transform of the macromolecule in a vacuum (Figure 1c), and it can be expressed as the sum of the $\langle\rho_p\rangle G_U(\mathbf{h})$ and $\Delta(\mathbf{h})$, the latter which is the Fourier transform of the internal density fluctuations from the mean density inside the domain U (i.e. $\langle\rho_p\rangle - \rho_p(\mathbf{r})$, see figure 1e).

$$F_p(\mathbf{h}) = \langle\rho_p\rangle G_U(\mathbf{h}) + \Delta(\mathbf{h}).$$

Substituting in this expression for $F_p(\mathbf{h})$ gives

$$F(\mathbf{h}) = (\langle\rho_p\rangle - \rho_s) G_U(\mathbf{h}) + \Delta(\mathbf{h}).$$

The term $(\langle\rho_p\rangle - \rho_s)$ is defined as the contrast (Stuhrmann & Kirste, 1965), and when it is equal to zero the system is said to be at the constrast matching point (see figure 1d), whereby only the internal electronic density fluctuations contribute to the overall structure factor. A demonstration of this expression can be found in Carter et al. (1990).

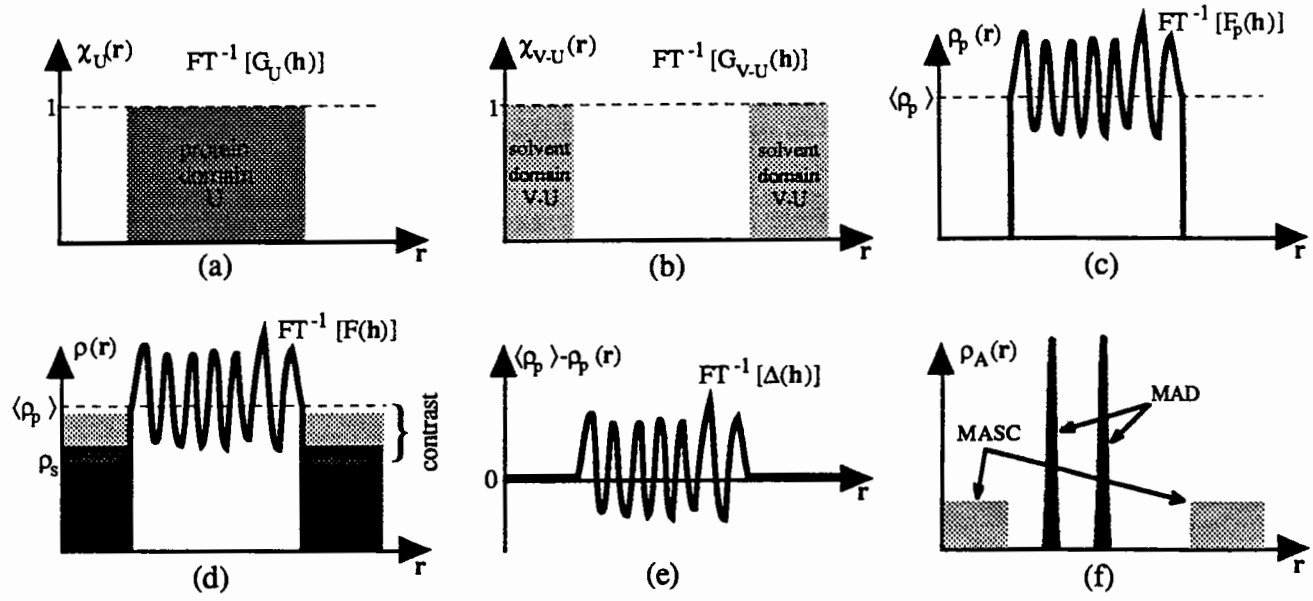


Figure 1. The 1-dimensional slices of different components in contrast variation theory: a) Indicator function of the ordered domain, U, containing the protein. b) Indicator function of the disordered domain, V-U, containing the solvent. c) The electronic density of only the ordered domain, U. This corresponds to the macromolecule in a vacuum. d) The electronic density for both the macromolecule and solvent regions. Three different electronic densities of the solvent are represented by the three shades of grey. The contrast is shown for one of these. e) The internal electronic density fluctuations inside the macromolecule. f) The anomalous electronic density for both the MAD and MASC cases.

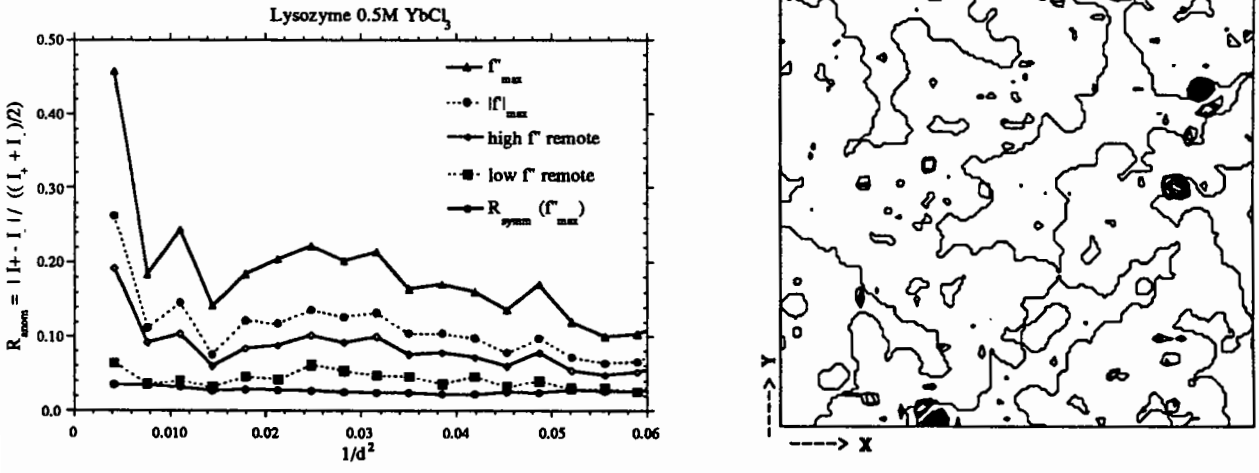


Figure 2. On the left, anomalous R-factors for MASC data of HEW lysozyme in 0.5M $YbCl_3$. Four wavelengths at the Yb L_{III}-edge plus the R-factor for true symmetry related reflections (i.e. respecting the differences between I+ and I-). On the right, ordered sites of Yb^{3+} ions in HEW lysozyme crystals. A phased anomalous Fourier map is superimposed on to a map of the protein envelope. Dark spots show Yb^{3+} positions in crevices and near the surface of the protein.

Table 1.

Protein	HEW Lysozyme			P64k	Xylose isomerase	
MW	14.3kDa			64kDa	173.2kDa	
unit cell & space group	a=b=78.48Å c=37.65Å $P4_32_12$	a=b=78.68Å c=37.05Å $P4_32_12$	a=b=78.18Å c=37.60Å $P4_32_12$	a=b=140.62Å c=77.02Å $P4_32_12$	a=b=141.91Å c=227.48Å $P3_221$	a=b=142A c=227Å $P3_221$
Anomalous Scatterer [A]	0.8M $YbCl_3$	0.5M $YbCl_3$	1.5M NaBr	3.5M $(NH_4)_2SeO_4$	2.0M $(NH_4)_2SeO_4$	1.35M Rb_2SO_4
Absorption edge	Yb L_{III}-edge	Yb L_{III}-edge	Br K-edge	Se K-edge	Se K-edge	Rb K-edge
wavelengths λ (Å)						
long-λ remote	-	1,39294	0,9222	0,99188	0.99188	0.8222
edge	1.38809Å	1,38593	0.9202	0.97954	0.97954	0.8178
peak	1.38751Å	1,38531	0,9195	0,97935	0.97935	0.8172
short-λ remote	1.38084Å	1,37762	0,9155	-	0.97912	0.8139
beamline station	D23	DW21	DW21	TROIKA	TROIKA	DW21
detector	MWPC	18cm IPS	18cm IPS	30cm IPS	30cm IPS	18cm IPS
Resolution limits	18.3-3.34Å	34.0-3.90Å	56.0-3.97Å	100.0-4.18Å	106.0-4.11Å	120.0-4.61Å

Table 2. Results obtained from MADLSQ and GFROMF. R-Factors quoted are based upon |Fobs(**h**)| & |Fcalc(**h**)|, where |Fobs(**h**)| is from MADLSQ or GFROMF and |Fcalc(**h**)| is calculated from the known model.

Protein	P64k	Xylose isomerase	Lysozyme
d_{min}	20Å	20Å	20Å
N_{meas}/N_{poss}	73 / 73	215 / 233	7 / 12
R-MADLSQ	32.5%	26.3%	32.1%
R-GFROMF	33.6%	29.7%	37.2%

Extension to the case where anomalous scatterers are present in the solvent can be done using the seminal idea of Karle (1980) where the structure factors are separated into wavelength independent (f°) and dependent parts ($^\lambda f' + i^\lambda f''$),

$$^\lambda f = f^\circ + {}^\lambda f' + i^\lambda f''.$$

For our discussion here, we assume that there is only one anomalous scatterer which is randomly dispersed in the solvent domain V-U such that it has a uniform distribution and no ordered sites bound to the surface of the macromolecule. For simplicity, we also assume that any scattering factor at low resolution is constant with respect to scattering angle to within a first order approximation. The density of the anomalous scatterers in the solvent can be treated as a complex quantity, $^\lambda\rho_{sA}$, which is dependent upon wavelength,

$$^\lambda\rho_{sA} = {}^\circ\rho_{sA}\,(1 + {}^\lambda f'/f^\circ + i^\lambda f''/f^\circ),$$

and where $^\circ\rho_{sA}$ is the normal electronic density of the anomalous scatterer. The total electonic density of the solvent, $^\lambda\rho_s$, becomes a function of the wavelength, and can be separated into wavelength independent and dependent parts,

$$^\lambda\rho_s = {}^\circ\rho_s + {}^\circ\rho_{sA}\,(^\lambda f'/f^\circ + i^\lambda f''/f^\circ).$$

Note that the term $^0\rho_s$ includes the normal scattering part of the anomalous scatterer. Thus one obtains,

$$^\lambda F(\mathbf{h}) = (\langle\rho_p\rangle - {}^\lambda\rho_s)\, G_U(\mathbf{h}) + \Delta(\mathbf{h})$$

$$^\lambda F(\pm\mathbf{h}) = \{\ (\langle\rho_p\rangle - {}^\circ\rho_s)\, G_U(\mathbf{h}) + \Delta(\mathbf{h})\ \} - \{\ {}^\circ\rho_{sA}\,(^\lambda f'/f^\circ \pm i^\lambda f''/f^\circ)\, G_U(\mathbf{h})\ \}$$

The terms in between the first set of brackets represent the wavelength independent part of the overall structure factor, denoted $^\circ F(\mathbf{h})$. It includes the envelope, constrast and fluctuation terms. The second set of brackets is wavelength dependent, and incorporates the envelope and the anomalous structure factors of A, $^\lambda f'$ and $^\lambda f''$. Note that the wavelength dependent contribution is substracted from the normal scattering part indicating that the anomalous and dispersive structure factors of A are applied to the Fourier transform of the indicator function of the solvent accessible domain, $-G_U(\mathbf{h})$.

By defining $\Gamma(\mathbf{h}) = -{}^\circ\rho_{sA}G_U(\mathbf{h})$ one generates a expression of the overall structure factor similar to the starting point used for the algebraic MAD method (Hendrickson, 1985), where $\Gamma(\mathbf{h})$ replaces the normal scattering component of the of the partial structure A, $^\circ F_A(\mathbf{h})$.

$$^\lambda F(\pm\mathbf{h}) = {}^\circ F(\mathbf{h}) + (^\lambda f'/f^\circ \pm i^\lambda f''/f^\circ)\, {}^\circ F_A(\mathbf{h}) \qquad \text{"MAD"}$$
$$^\lambda F(\pm\mathbf{h}) = {}^\circ F(\mathbf{h}) + (^\lambda f'/f^\circ \pm i^\lambda f''/f^\circ)\, \Gamma(\mathbf{h}) \qquad \text{"MASC"}$$

The substitution of $\Gamma(\mathbf{h})$ for ${}^{\circ}F_A(\mathbf{h})$ has an obvious physical meaning. The anomalous partial structure, A, which is a set of a few punctual and ordered scatterers in a MAD experiment is exchanged for an extended uniform electron density in a MASC experiment (see figure 1f). The separation of the effects of the anomalous partial structure A (and hence the Fourier transform of the solvent accessible volume) from the overall diffraction effects can be applied using a set of equations analogous to the MADLSQ equations, where they are solved for $|{}^{\circ}F_T(\mathbf{h})|$, $|\Gamma(\mathbf{h})|$ and phase difference between ${}^{\circ}F_T(\mathbf{h})$ and $\Gamma(h)$, $\Delta\phi = (\phi_T - \phi_\Gamma)$, i.e.

$$|{}^{\lambda}F(\pm\mathbf{h})|^2 = |{}^{\circ}F_T(\mathbf{h})|^2 + a(\lambda)\,|\Gamma(\mathbf{h})|^2 + b(\lambda)\,|{}^{\circ}F_T(\mathbf{h})|\,|\Gamma(\mathbf{h})|\cos(\Delta\phi) \pm c(\lambda)|{}^{\circ}F_T(\mathbf{h})|\,|\Gamma(\mathbf{h})|\sin(\Delta\phi)$$

where, $a(\lambda) = ({}^{\lambda}f^2 + {}^{\lambda}f''^2)/({}^{\circ}f^2)$, $b(\lambda) = 2\,{}^{\lambda}f'/{}^{\circ}f$, $c(\lambda) = 2\,{}^{\lambda}f''/{}^{\circ}f$.

A MASC experiment has an advantage over other contrast variation methods, since the contrast variation is generated by inducing a physical change. This eliminates the possibility of changes in the crystal lattice due to varying ionic strength, pH, precipitant concentration, etc... which can arise in a chemical contrast series, and thus enforces strict isomorphism.

3. Strength of the anomalous signal in MASC

The strength of the signal in an anomalous contrast variation series can be quantified in a similar way to those in the MAD method, i.e. by measuring differences between Bijvoet pairs (anomalous or ${}^{\lambda}f''$ contribution) and wavelengths (dispersive or ${}^{\lambda}f'$ contribution). Intuitively, the magnitude of the anomalous signal in a MASC experiment is expected to vary considerably with resolution, being very large in the lowest resolution shells and then diminishing rapidly with increasing resolution. One also expects the anomalous signal to be directly proportional to the concentration of the anomalous scatterer in the solvent accessible volume. Furthermore, the signal will be maximised at the point of contrast matching. By making a certain number of approximations, it is possible to derive expressions for and calculate the expected anomalous and dispersive ratios (Fourme et al., 1995), but for the purpose of succintness only the final expressions will be given here. Thus for anomalous and dispersive differences one gets[†] ,

$$\langle|{}^{\lambda}\Delta F(\pm\mathbf{h})|\rangle / \langle|{}^{\lambda}F(\mathbf{h})|\rangle = 3.44 \times 10^{-4}\,[A]\,(2\,{}^{\lambda}f''/f_{eff})\,(M_w{}^{1/12}\,s)^{-2}\exp(-Bs^2/4)$$

and

$$\langle|{}^{\Delta\lambda}\Delta F(\mathbf{h})|\rangle / \langle|{}^{\lambda}F(\mathbf{h})|\rangle = 3.44 \times 10^{-4}\,[A]\,(\Delta f'/f_{eff})\,(M_w{}^{1/12}\,s)^{-2}\exp(-Bs^2/4).$$

Clearly, the anomalous signal is dependent on a number of factors, such as the molar concentration of the anomalous scatterers, [A] and the magnitudes of f'' and Δf'. However, the resolution, s, has the strongest effect on the anomalous signal which drops away as a function of $1/s^2$ and $\exp(-Bs^2/4)$. The term $\exp(-Bs^2/4)$ represents a Gaussian smoothing of the

[†] Where $s = 2\sin\theta/\lambda$, $|{}^{\lambda}\Delta F(\pm\mathbf{h})| = |\,|{}^{\lambda}F(+\mathbf{h})| - |{}^{\lambda}F(-\mathbf{h})|\,| / \langle|{}^{\lambda}F(\mathbf{h})|\rangle$, $\langle|{}^{\lambda}F(\mathbf{h})|\rangle = |\,|{}^{\lambda}F(+\mathbf{h})| + |{}^{\lambda}F(-\mathbf{h})|\,|/2$, and $|{}^{\Delta\lambda}\Delta F(\mathbf{h})| = |\,|{}^{\lambda i}F(\mathbf{h})| - |{}^{\lambda j}F(\mathbf{h})|\,| / \{\,(\langle|{}^{\lambda i}F(\mathbf{h})|\rangle + \langle|{}^{\lambda j}F(\mathbf{h})|\rangle\,) / 2\,\}$.

envelope boundary, where B is pseudo-temperature factor which defines the thickness of the interface rather than the temperature factor of the macromolecule or solvent. The signals are also somewhat dependent upon the molecular weight, but a 100kDa protein will still produce 68% of the signal of a 10kDa protein. For a hypothetical case of a 50kDa protein in 3.5M $(NH_4)_2SeO_4$, where f"= $7.0e^-$, $\Delta f'$=$8.6e^-$ and B=100Å^2, expected anomalous and dispersive ratios (respectively) are 0.441 and 0.274 at 33Å resolution, 0.156 and 0.097 at 20Å resolution, and 0.032 and 0.020 at 10Å resolution. Hence one expects in the lowest resolution shells very large signals, which will decrease sharply with increasing resolution. If one wishes to obtain a measurable anomalous signal out to 10Å resolution, then one requires either multimolar quantities of a K-edge scatterer or molar quantities of a L-edge scatterer.

4. Experimental

As a MASC experiment utilises the variation of f' and f", data collection should ideally be carried out at X-ray wavelengths near absorption edges of the anomalous scatterer. Thus the requirements are very similar to a MAD experiment - i.e. tuneable X-rays with a narrow band pass ($\Delta\lambda/\lambda \approx 10^{-4}$), a X-ray fluorescence detector to determine precisely the wavelengths of f''_{max} and $|f'|_{max}$, the recording of Bijvoet mates or Friedel pairs close in time, etc... - but with the additional requirement that the experimental setup is designed to collect reflections at the lowest possible resolution. This often requires the mounting of a small beamstop just in front of the detector entrance window. Other practical considerations are to use an area detector with a large dynamic range to accomodate the accurate measurement of the most intense low resolution reflections with those weaker reflections at more moderate resolution.

A variety of anomalous scatterers may be used in a MASC experiment, and the most suitable ones will depend on the crystallisation conditions of the macromolecule. Analogues of the precipitating agent are good choices since such compounds are less likely to perturb the crystalline lattice (e.g. selenate for sulphate, bromide for chloride, tribromoacetate for acetate, etc...). To date, MASC data have been collected on crystals of three proteins of differing molecular weights and with a variety of different anomalous scatterers (see Table 1). In order to develop the MASC method, all of the cases are known crystal structures, which allows the experimental results to be compared with the correct envelope transform moduli and phases. In each of the experiments described below, the X-ray diffraction data were recorded at the wavelengths corresponding to the $|f'|_{max}$ and the f''_{max} which were determined from the X-ray fluorescence spectra from a solution of the anomalous scatterer, as well as for at least one wavelength remote of the absorption edge. A small beamstop (≈2-3mm) was mounted and aligned just in front of the entrance window of the detector. Where possible, the crystallographic axes were aligned so that Bijvoet pairs could be measured on the same image. Below, we describe in detail the experiments and the results for only two anomalous scatterers.

4.1 Hen egg white lysozyme co-crystallised in $YbCl_3$

The very first MASC experiment was conducted on single crystals of lysozyme directly crystallised from solutions of 0.3-0.5M $YbCl_3$. This combination was chosen because of the

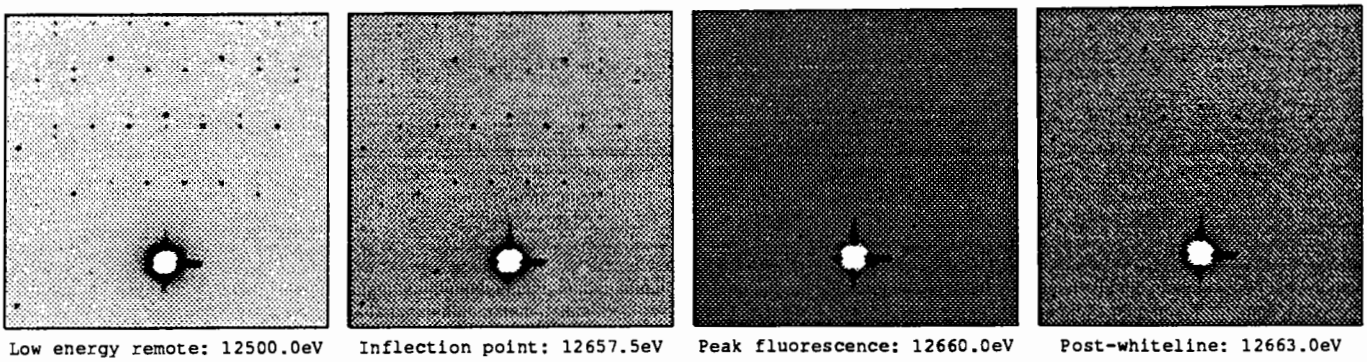


Figure 3. Fluorescence and absorption effects of the diffraction pattern from a xylose isomerase crystal in 2M $(NH_4)_2SeO_4$ recorded at four different wavelengths about the Se K-edge and of the same region of reciprocal space.

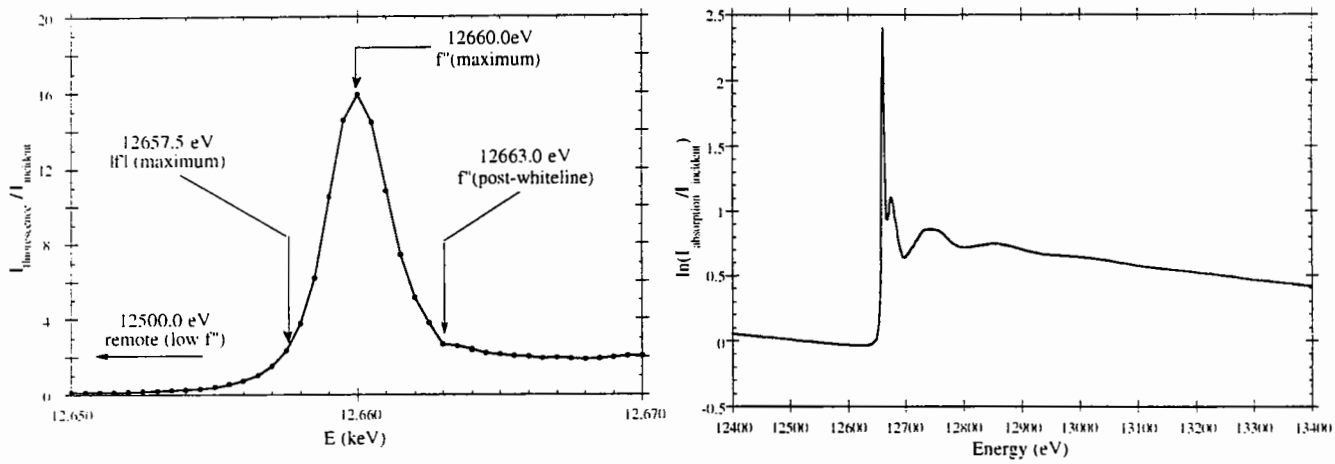


Figure 4. Whiteline structure of the Se K-edge of 0.1M $(NH_4)_2SeO_4$ recorded on the TROÏKA beamline station, ESRF, France (left), and the extended absorption spectra recorded at LURE, Orsay, France (right).

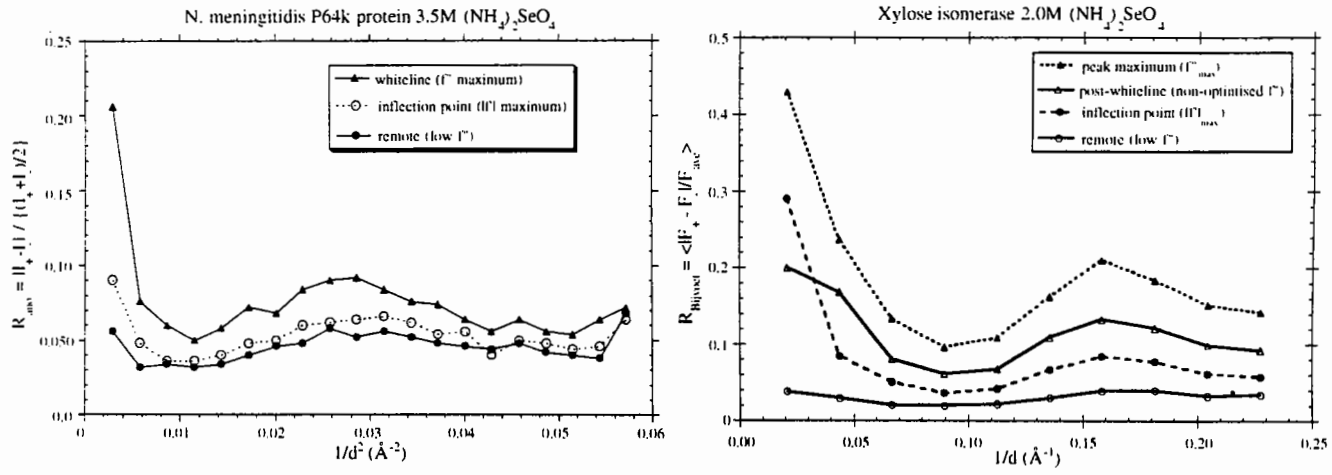


Figure 5. Anomalous R-factors as a function of resolution for P64k and xylose isomerase in $(NH_4)_2SeO_4$. Please note that the resolution is broken down as a function of $1/d^2$ for P64k and $1/d$ for xylose isomerase.

ease of obtaining crystals and their robustness, as well as for the white line structure of the Yb L_{III}-edge. The Yb^{3+} ion concentration could be increased to 0.8M using vapour diffusion techniques before the crystal quality would deteriorate. X-ray diffraction data were collected at three wavelengths including one remote on the high energy side of the Yb L_{III}-edge on the D23 station (Kahn et al., 1986) at LURE-DCI (Orsay, France). Bijvoet ratios are shown in Figure 2. The results confirmed the large anomalous signal at low resolution as expected by theory. At the wavelength corresponding to the maximum of f", the Bijvoet ratio reaches ≈50% for the lowest resolution shell and then diminishes sharply with increasing resolution. The internal agreement between true equivalent reflections is within ≈1-3%, implying that the anomalous signal is real, reproduceable and not artifact of either the data processing or the beamstop shadow. The anomalous signal however extends well beyond 10Å resolution indicating that some Yb^{3+} ions have bound to the protein. Anomalous difference Patterson maps did not reveal the positions of three bound Yb^{3+} ions, which were eventually found in a phased anomalous difference Fourier map (Fig. 2). The reason for this might be because the diffraction data is only ≈60% complete. This experiment has recently been repeated with 0.5M $YbCl_3$ on the DW21b station at LURE-DCI to obtain a complete MASC data set and also to investigate the possibility of using the ordered sites of the anomalous scatterer in an overall phasing and phase extention strategy.

4.2 P64k and xylose isomerase in $(NH_4)_2SeO_4$

P64k is a 64kDa outer membrane protein from *Neisseria meningitidis* currently under study in our lab (Li de la Sierra et al., 1994; Li de la Sierra et al., 1997), and it crystallises from ammonium sulphate solutions. Xylose isomerase also crystallises from ammonium sulphate solutions but as a tetramer (173.2kDa) in the asymmetric unit (Rey et al., 1988). Both of these proteins represent large macromolecular structures on the scale of those typically solved by the MAD method. Ammonium sulphate in the mother liquor of the crystals could be substituted with multimolar concentrations of ammonium selenate via simple soaking techniques. Crystals of both proteins could withstand 3.5M $(NH_4)_2SeO_4$, which brings the solvent electronic density equal to the average protein electronic density, i.e. the contrast matching point. This allowed us to collect MASC data at the Se K-edge of selenate which features a white-line structure at a wavelength near ≈1Å. In the first series of these experiments done on the TROIKA station at the ESRF (Grenoble, France), the diffracting power of the P64k crystals deteriorated rapidly under the intense radiation of the undulator beam despite cooling the sample at 4°C. In order to collect a complete MASC data set off of one crystal, the experiments were later repeated using flash cooling and cryogenic techniques. Images recorded at the Se K-edge showed a marked decrease in the diffraction intensities as well as a substantial increase in the overall background. Although such absorption and fluorescence effects have been previously noted, they were never so severe (see Figure 3). This could be understood once it was realised that selenate has an exceptionally large whiteline resonance (see Figure 4) which could only be revealed when using the finer energy resolution of the monochromated X-rays from the Si(333) crystal instead of the diamond C(111) crystal used in the previous run. To circumvent and

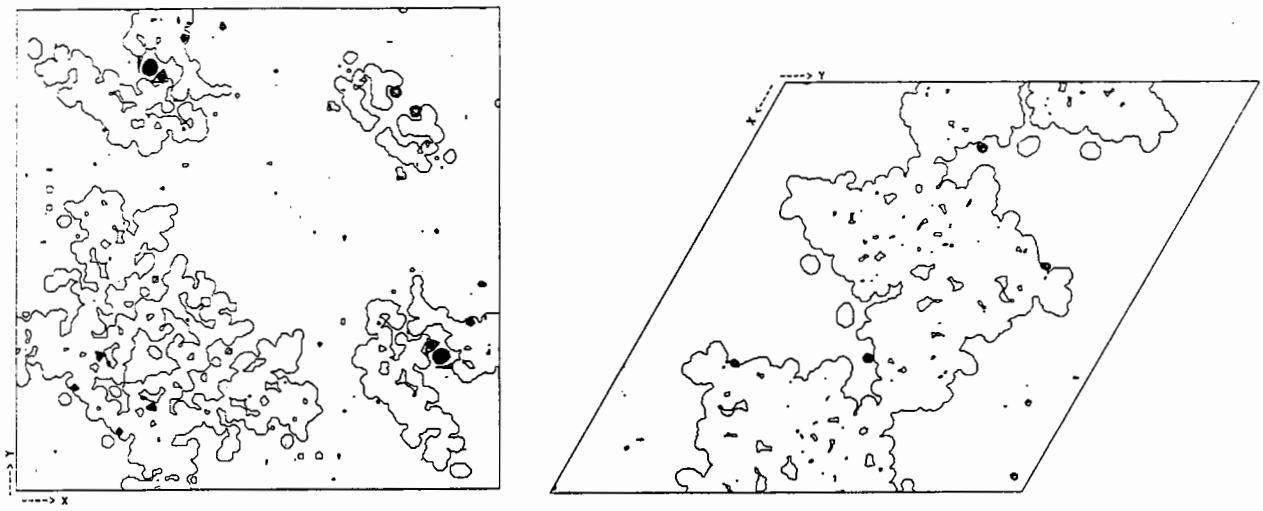


Figure 6. Ordered sites of selenate ions in P64k (left) and XI (right). In each case, a phased anomalous Fourier map is superimposed on to a map of the protein envelope. Dark spots on each map show selenate positions in crevices and near the surface of the protein.

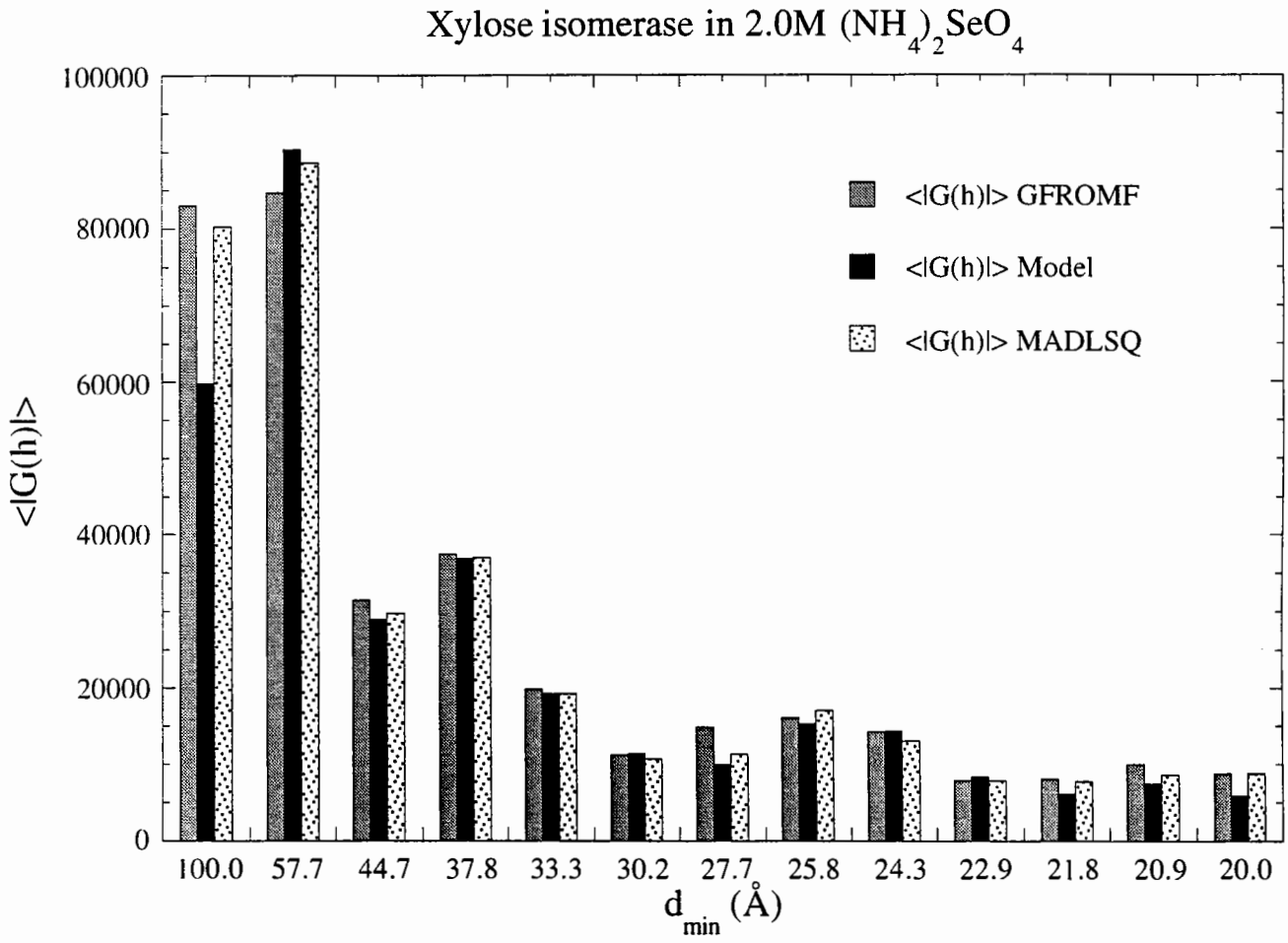


Figure 7. Agreement as a function of resolution of $|G_U(\mathbf{h})|$ values of xylose isomerase in 2.0M $(NH_4)_2SeO_4$ calculated from MADLSQ, GFROMF and its model.

minimise the absorption and fluorescence effects, a number of precautions were taken: i) Crystals were rapidly rinsed or washed in an analogous sulphate solution just prior to freezing, thus removing mother liquor containing any excess selenate surrounding the crystal, ii) smaller crystals were used to reduce the amount of absorption relative to the crystal volume, and iii) finer oscillation angles were recorded for the wavelength corresponding to the maximum of f" to improve the signal to background ratio. There are other tactics which could be employed to minimise fluorescence effects. For X-ray energies used in typical crystallography experiments (0.5Å - 2.0Å), the fluorescence yield after absorbing an incident photon is 2-3 times higher for the K-edges than for the L-edges (Kortright, 1986). For example, the fluorescence yield is ≈60% for Se at its K-edge, whereas the fluorescence yield is only ≈20% for Yb at one of its L-edges. Another method to reduce fluorescence effects is simply to increase the sample-to-detector (D) distance since fluorescence which is radiated isotropically will fall off as $1/D^2$, while the diffracted beams being quasi-parallel will remain essentially constant with D.

The anomalous signal for both proteins follows the expected trend, being very large for the Bijvoet pairs at lowest resolution and decreasing rapidly with increasing resolution (see Figure 5). At higher resolution, the Bijvoet ratios for both proteins are of the order of the internal agreement, but despite this low anomalous signal, up to 12 possible selenate ion sites have been located from phased anomalous difference Fourier maps in the P64k crystals (see Figure 6). Similarily, several selenate ion sites have been found in the crystals of xylose isomerase. All sites are at or near the macromolecular boundary, often in crevices, and their relative occupancies vary considerably. The existence of ordered anomalous sites appears to be more general than expected, but it opens up a potential of phasing to higher resolution once a model for envelope is determined.

5. Extracting $|G_U(\mathbf{h})|$ from MASC data

Two methods have been utilised to extract the moduli of $G_U(\mathbf{h})$ from multiple-wavelength diffraction data. One uses the algebraic equations in the MAD method as implemented in the program MADLSQ (Hendrickson, 1985), and the other uses the program GFROMF (Carter & Bricogne, 1987), which is designed to extract the $|G_U(\mathbf{h})|$ from the $|^iF_{obs}(\mathbf{h})|$ of a chemical contrast variation series. Both methods give satisfactory results up to at least 20Å resolution.

Prior to using either method of extracting the $|G_U(\mathbf{h})|$, the X-ray data were set on a common scale using the program SCALA (Evans, 1993). The data were scaled in two steps: i) an internal scaling for each wavelength to correct for incident beam fluctuations and sample decay, and ii) a pseudo-local scaling between a reference wavelength (low f") and the other wavelengths to minimise absorption effects.

5.1 MADLSQ

As mentioned above, the program MADLSQ, which was originally designed for multiple-wavelength diffraction data, solves the set of equations by non-linear least-squares for $|^\circ F_T|$, $|^\circ F_A|$ and the phase difference $\Delta\phi_{T\text{-}A}$. For MASC data, $|\Gamma(\mathbf{h})|$ (or $\rho_{sA}|G_U(\mathbf{h})|$) replaces $|^\circ F_A(\mathbf{h})|$, and the phase difference becomes $\Delta\phi_{T\text{-}\Gamma}$. The program also has the ability to refine or fix the

values of f' and f'' of the different wavelengths. Results are shown in Figure 7 for the data collected on xylose isomerase crystals soaked in $(NH_4)_2SeO_4$ and as compared to the $|G_U(\mathbf{h})|$ calculated from the coordinates of the 3D structure deposited in the Protein Data Bank (Rey et al., 1988). Note the sharp asymptotic decrease in $|G_U(\mathbf{h})|$ with increasing resolution. The agreement between model and experiment deteriorates beyond 10-20 Å resolution for several reasons: i) the relative magnitudes of $|G_U(\mathbf{h})|$ are small, ii) the absorption effects are more pronounced at higher diffracting angles, and iii) the possibility of ordered anomalous scattering sites contributes to the partial structure extracted from the MADLSQ equations (i.e. $|\Gamma(\mathbf{h})|$ is more precisely defined as $|\Gamma(\mathbf{h})+^{\circ}F_A(\mathbf{h})|$).

5.2 GFROMF

In chemical contrast variation studies, the program GFROMF (Carter & Bricogne, 1987) extracts the $|G(\mathbf{h})|$ from the diffraction data $|^iF_{obs}(\mathbf{h})|$ for i=1,...,N where i corresponds to to a different solvent density, $^i\rho_s$. To extend this to multiple wavelength cases, we simply substitute in for the contrast series $|^{\lambda_i}F_{obs}(\mathbf{h})|$ where $\lambda_i = \lambda_1,..., \lambda_N$ and the solvent density becomes $^{\lambda_i}\rho_s$. The same formalism is used to describe the overall structure factor in terms of the Fourier transforms of the envelope ($G_U(\mathbf{h})$) and the internal denity fluctuations ($\Delta(\mathbf{h})$). If $X(\mathbf{h})$ and $Y(\mathbf{h})$ are the real and imaginary components of $\Delta(\mathbf{h})$ relative to $G_U(\mathbf{h})$, one has,

$$|^iF_{calc}(\mathbf{h})| = {}^iK \{ [X(\mathbf{h}) + (\langle\rho_p\rangle - {}^i\rho_s) |G_U(\mathbf{h})|]^2 + Y(\mathbf{h})^2 \}^{1/2}$$

The GFROMF scheme carries out the non-linear least-squares refinement of $|G_U(\mathbf{h})|$, $X(\mathbf{h})$ and $Y(\mathbf{h})$ from scaled data summed over all contrasts, and minimises the function,

$$\Sigma_i\Sigma_{hkl}\, \sigma_{obs}(\mathbf{h})^{-2} (|^iF_{obs}(\mathbf{h})| - |^iF_{calc}(\mathbf{h})|)^2,$$

where $\sigma_{obs}(\mathbf{h})$ is standard deviation of $|^iF_{obs}(\mathbf{h})|$. Note that $X(\mathbf{h})$ and $Y(\mathbf{h})$ represent both the magnitude and the phase difference between $G_U(\mathbf{h})$ and $\Delta(\mathbf{h})$. In practice, ^iK, a scale factor between the different data sets, should be refined for all but one contrast or wavelength.

The original program was modified to incorporate anomalous scattering contributions such that,

$$|^{\lambda_i}F_{calc}(\pm\mathbf{h})| = {}^{\lambda_i}K \{ [X(\mathbf{h}) + (\langle\rho_p\rangle - \rho_s - (^{\lambda_i}f'/f^{\circ})^{\circ}\rho_{sA}) |G_U(\mathbf{h})|]^2 + [Y(\mathbf{h}) \pm (-(^{\lambda_i}f''/f^{\circ})^{\circ}\rho_{sA}) |G_U(\mathbf{h})|]^2 \}^{1/2}$$

Tests executed on simulated MASC data of kallikreen (52kDa), at three different contrasts of selenate and three wavelengths per contrast, returned exact values of $|G_U(\mathbf{h})|$, $X(\mathbf{h})$ and $Y(\mathbf{h})$ of the simulated observed data. With experimental data, the results gave R-factors of ≈30-35% for P64k and xylose isomerase crystals (see Table 2). This level of agreement is satisfactory considering that many of the parameters are unrefined. In particular, the values of $^{\lambda_i}f'$ and $^{\lambda_i}f''$ employed were derived from previous runs of MADLSQ, and theoretical values of the contrast were used rather than allowing them to refine. The scale factors between different wavelengths

($^{\lambda i}$K) were set to unity since the data were already set on a common scale. In principle, all of these parameters should be refineable in the GFROMF scheme, even though the number of observations in the lower resolution shells is not overly large. What is certain is that prior precise knowledge of the values of the contrasts, $^{\lambda i}$f' and $^{\lambda i}$f'' is important to extract $|G_U(\mathbf{h})|$ values of satisfactory quality.

6. Phasing G-moduli

Previous methods of phasing $|G_U(\mathbf{h})|$ from either X-ray contrast variation series (e.g. Carter, et al) or H/D substitution contrast variation series (Moras et al., 1983; Roth et al., 1984; Bentley et al., 1984; Podjarny et al., 1987; Roth, 1991) employed the assumptions that the set of $|G_U(\mathbf{h})|$ behave much in the same way as the structure factors of small molecule crystal structures. Hence such attempts have used the programs of traditional direct methods of small molecule crystallography. As a starting point, we have also examined this strategy in preliminary trials for phasing a set of $|G_U(\mathbf{h})|$ from MASC data, but it is clear that the limited success with these methods necessitates a re-examination of the phasing methods used up to now.

Using 1664 calculated $|G_U(\mathbf{h})|$ up to 10Å resolution from a model of xylose isomerase, phase sets for the $|G_U(\mathbf{h})|$ were generated using the program MITHRIL (Gilmore, 1984). Normalisation was carried out empirically, by dividing the entire set of $|G_U(\mathbf{h})|$ by a constant which set the 397 largest $|G_U(\mathbf{h})|$ to greater than 1.3. Of the phase sets generated, using triplets, magic integers and statistically weighted tangent refinement, the best solutions gave correlation coeffients of ≈0.74 for 20Å resolution maps. However, none of the conventional figure-of-merits were capable of distinguishing a correct phase set.

The limited success obtained from the use of traditional direct methods is not surprising considering that such methods are based on a variety of assumptions which are not valid for a set of $|G(\mathbf{h})|$. For example, envelopes are not point scatterers, as can be assumed for atoms. Also an envelope also does not represent a random distribution of scatterers; quite the contrary, by definition of the biphasic model, the scatterers are confined inside the volume of the solvent. Consequently, a set of $|G(\mathbf{h})|$ does not follow Wilson statistics. In addition, normalisation of $|G(\mathbf{h})|$ can not be accomplished as in traditional methods because of the relatively few reflections at low resolution and their very large dynamic range. Despite these differences with small molecules, a set of $|G(\mathbf{h})|$ has the advantage in being complete with relatively few reflections (i.e. there are only a total of 73 unique reflections to 20Å resolution for P64k).

The problem of phasing a set of $|G_U(\mathbf{h})|$ clearly needs to be readdressed. We are currently considering other methods towards phasing $|G_U(\mathbf{h})|$, and the use of Maximum Entropy and Likelihood ranking to test envelope hypotheses. The literature shows an increasing interest in the field of low resolution phasing. Some of these methods approximate globular proteins as spheres or a few large Gaussian spheres (Andersson & Hovmöller, 1996; Harris, 1995; Lunin et al., 1995; Urzhumtsev et al., 1996), or as a gas of hard sphere point scatterers (Subbiah, 1991; Subbiah, 1993).

7. Conclusions & Perspectives

It has been demonstrated that contrast variation in macromolecular crystallography can be generated using anomalous dispersion techniques in a MAD-like experiment. The method benefits from the strict isomorphism imposed by the external physical change of the wavelength of the X-rays applied to a single sample. This is clearly advantageous over a chemical contrast series experiment which typically requires several samples soaked in different media, and which risks destroying any isomorphism.

From the studies presented here, large anomalous signals are observed in the lowest resolution shells. In all of the cases studied to date, the anomalous signal extends to higher resolution indicating the presence of ordered anomalous scattering sites. Such sites have little effect at low resolution, and they are a bonus in a MASC experiment because they may provide a path for phasing the 3D structure to higher resolution once the envelope is known. Extracting the set of $|G_U(\mathbf{h})|$ from MASC data can be accomplished using two different procedures; one based on the algebraic equations of multiple-wavelength diffraction data (MADLSQ) and the other based on the equations derived from a chemical contrast variation series (GFROMF).

The process of phasing a set of $|G_U(\mathbf{h})|$ needs further attention. Traditional direct methods, which are intended for small molecule structures, are not suitable for this type of phase problem. If the phasing step of a set of $|G_U(\mathbf{h})|$ can be dealt with, then the combination of anomalous dispersion and contrast variation techniques can lead to a general method for low resolution phasing of very large macromolecules including those beyond the scope of MIR and MAD methods. Finally, knowledge of the macromolecular envelope will help phase the structure to higher resolution.

Acknowledgements

We thank A.Thompson, A. Gonzalez, G. Grübel, D. Abernathy & M. Lehmann for support during the experiments on the TROÏKA beamline at the ESRF, Grenoble, France. We are also grateful to D. Ragonnet, D. Chandesris and the SEXAFS group at LURE for assistance with the DW21 beamline.

References

Andersson, K.M. & Hovmöller, S. (1996) Acta Cryst. **D52**, 1174-1180.

Bentley, G.A., Lewit-Bentley, A., Finch, J.T., Podjarny, A.D. & Roth, M. (1984). J. Mol. Biol. **176**, 55-75.

Bragg, W.L. & Perutz, M.F. (1952) Acta Cryst. **5**, 277-289.

Bricogne, G. (1974) Acta Cryst. **A30**, 395-405.

Bricogne, G. (1993) Acta Cryst. **D49**, 37-60.

Carter, C.W., Jr. & Bricogne, G. (1987) GFROMF: A computer program for scaling and estimating envelope structure factors from contrast variation data. Dept. of Biochemistry, CB#7260, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7260, USA

Carter, C.W., Crumley, K.V., Coleman, D.E., Hage, F. & Bricogne, G. (1990) Acta Cryst. **A46**, 57-68.

Crumley, K.V. (1989) M. Sc. Thesis, University of North Carolina, Chapel Hill, USA.

Dumas, C. (1988) Thèse de Doctorat des Sciences Naturelles, Université Paris-Sud, Centre Scientifique d'Orsay, France.
Evans, P. R. (1993) In *Proceedings of the CCP4 Study Weekend on Data Collection and Processing*. pp 114-122. Warrington: SERC Daresbury Laboratory.
Fourme, R., Shepard, W., Kahn, R., L'Hermite, G. & Li de La Sierra, I. (1995) J. Synchr. Rad. **2**, 36-48.
Gilmore, C.J. (1984) J. Appl. Cryst. **17**, 42-46.
Grübel, G (1994) ESRF Beamline Handbook, 55-59.
Harris, G.W. (1995) Acta Cryst. **D51**, 695-702.
Harrison, S.C. (1969) J. Mol. Biol. **42**, 457-483.
Hendrickson, W.A. (1985) Trans. Am. Crystallogr. Assoc. **21**, 11-21.
Hütsch, M. (1993) PhD Thesis, University of Hamburg, Germany.
Ibel, K. & Stuhrmann, H.B. (1975) J. Mol. Biol. **93**, 255-265.
Jack, A., Harrison, S.C. & Crowther, R.A. (1975) J. Mol. Biol. **97**, 163-172.
Kahn, R., Fourme, R., Bosshard, R. & Saintagne, V. (1986) Nucl. Instrum. Methods, **A246**, 596.
Karle, J. (1980) Int. J. Quantum Chem: Quantum Biol. Symp. **7**, 357-367
Kortright, J.B. (1986) in "Center for X-ray Optics: X-ray Data Booklet", pp. 2-19 - 2-20, ed.D. Vaughan
Kühnholz, O. (1991) J. Appl. Cryst. **24**, 811-814.
Li de la Sierra, I., Prangé, T., Fourme, R., Padron, G., Fuentes, P., Musacchio, A. and Madrazo, J. (1994) J. Mol. Biol. **235**, 1154-1155.
Li de la Sierra, I., Pernot, L., Prangé, T., Saludjian, P., Schiltz, M., Fourme, R. and Padron, G. (1997) J. Mol. Biol., in press.
Lunin, V.Y., Lunina, N.L., Petrova, T.E., Vernoslova, E.A., Urzhumtsev, A.G. & Podjarny, A.D. (1995) Acta Cryst. **D51**, 896-903.
Moras, D., Lorber, B., Romby, P., Ebel, J.-P., Giégé, R., Lewit-Bentley, A. & Roth, M. (1983) J. Biol. Struct. Dynam. **1**, 209-223.
Munk, B. (1988) PhD Thesis, University of Hamburg, Germany.
Podjarny, A., Bhat, T.N. & Zwick, M. (1987) Annu. Rev. Biophys. Biophys. Chem. **16**, 351-374, and references cited therein.
Rey, F., Jenkins, J., Janin, J., Lasters, I., Alard, P., Matthyssens, G. & Wodak, S. (1988) PROTEINS: Structure, Function and Genetics **4**, 165-172
Roth, M., Lewit-Bentley, A. & Bentley, G.A. (1984) J. Appl. Cryst. **17**, 77-84.
Roth, M. (1991) Crystallographic Computing 5: From Chemistry to Biology, pp 229-248. Oxford University Press. Eds. D. Moras, A.D. Podjarny & J.C. Thierry.
Stuhrmann, H.B. & Kirste, R. (1965) Z. für. Physik. Chem. **46**, 247-250.
Stuhrmann, H.B. (1980) J. Appl. Cryst. **A36**, 996-1001.
Subbiah, S. (1991) Science **252**, 128-133.
Subbiah, S. (1993) Acta Cryst. **D49**, 108-119
Urzhumtsev, A.G., Vernoslova, E.A & Podjarny, A.D. (1996) Acta Cryst. **D52**, 1092-1097.
Williams, C.E., May, R.P. & Guinier, A. (1994) in "Materials Science and Technology: A Comprehensive Treatment - Characterisation of Materials Part II" Volume Ed. E. Lifshin, Eds. R.W. Cahn & E.J. Kramer, Vol. 2B, 611-656.

Direct Phase Determination by Multi-Beam Diffraction

Edgar Weckert, Kerstin Hölzer, Klaus Schroer
Institut für Kristallographie, Universität Karlsruhe (TH)
Kaiserstr. 12, D-76128 Karlsruhe, Germany

1 Introduction

The knowledge of the three-dimensional structure of a molecule is for many questions in science extremely important since nearly all properties do not depend only on the chemical composition but also on the arrangement of the atoms. For the determination of the three-dimensional structure of such molecules X-ray crystallography plays an important role. However, other methods like NMR (Wüthrich, 1995) gained more importance recently.

The electron density $\rho(\mathbf{r})$ of a crystal is periodic in three dimensions. Therefore, the Fourier transform of $\rho(\mathbf{r})$ of the whole crystal is discrete if the crystal is assumed to be infinite. Owing to this fact it is sufficient to restrict all considerations to one unit cell. The coefficients of the Fourier transform of $\rho(\mathbf{r})$ of the unit cell are called structure factors and are given by:

$$F(\mathbf{h}) = \sum_{N_{atom}} \Big(\int_{V_{atom}} \rho(\mathbf{r}) e^{2\pi i \mathbf{hr}} dr^3 \Big) e^{2\pi i \mathbf{h} \mathbf{r}_{atom}} \tag{1}$$

Hereby, $\mathbf{h}$ denotes a reciprocal space vector and $\mathbf{r}$ are vectors in direct space. The integral in brackets represents the Fourier transform of the electron density of an atom which is commonly called atomic scattering factor $f_j(\mathbf{h})$. $F(\mathbf{h})$ is a complex number which can be separated in modulus and phase $F(\mathbf{h}) = |F(\mathbf{h})| \exp(i\phi(\mathbf{h}))$. The phase $\phi(\mathbf{h})$ of $F(\mathbf{h})$ depends on the origin of the unit cell that was chosen. For an origin shift of $\Delta\mathbf{s}$ the phase $\phi(\mathbf{h})$ will change by $-2\pi\mathbf{h}\Delta\mathbf{s}$. The intensity of the reflections measured during X-ray experiments is given by $I(\mathbf{h}) \propto F(\mathbf{h})F^*(\mathbf{h}) \approx F(\mathbf{h})F(\bar{\mathbf{h}})$. The latter term is only valid if absorption is small. This means, an X-ray experiment with a single reflection delivers only the modulus but not the phase of the structure factors. Therefore, no simple Fourier back transformation

$$\rho(\mathbf{r}) = \frac{1}{V_{uc}} \sum_{\mathbf{h}} F(\mathbf{h}) e^{-2\pi i \mathbf{hr}} \tag{2}$$

to reveal the electron density of the unit cell is possible. This is well-known as the phase problem of X-ray crystallography. Since the discovery of X-rays many methods have been worked out to surmount the phase problem for both small and macromolecular structures. For small molecules in general more structure amplitudes $|F(\mathbf{h})|$ are available than unknown parameters that are necessary to describe the basic structure, taking also into account that it consists of atoms with a positive electron density everywhere. This is true if reflections up to atomic resolution can be measured. For that case, powerful

computer programs are available to solve the structures directly from the measured structure amplitudes by statistical methods called 'Direct Methods' (Debaerdemaeker, Tate & Woolfson, 1988; Sheldrick, 1990; Altomare et al., 1994; Miller, Gallo, Khalak & Weeks, 1994). For structures where no information up to atomic resolution is available additional information has to be provided for a successful solution of the structure. This can be the position of one or more heavy atoms with or without anomalous dispersion contributions (SIR, MIR and MAD) or a significant part of the molecule (molecular replacement). All relevant methods have been reviewed in a recent text book (Woolfson and Fan, 1995) and are the object of continuous improvement.

It is the purpose of this contribution to show that under certain circumstances also phase information besides the amplitudes can be obtained directly in a X-ray diffraction experiment. As explained before the phase of a single structure factor has no physical meaning since it depends on the choice of the origin. However, the phase of the product of structure factors whose corresponding reciprocal lattice vectors join to a closed polygon is independent of the origin. Such a quantity is called invariant. The simplest invariant (besides $F(\mathbf{h})F(\overline{\mathbf{h}})$) is a three-structure factor invariant like

$$F(\overline{\mathbf{h}})F(\mathbf{g})F(\mathbf{h}-\mathbf{g}) = |F(\overline{\mathbf{h}})F(\mathbf{g})F(\mathbf{h}-\mathbf{g})|e^{i\Phi_T} \tag{3}$$

with the triplet phase

$$\Phi_T = \phi(\overline{\mathbf{h}}) + \phi(\mathbf{g}) + \phi(\mathbf{h}-\mathbf{g}). \tag{4}$$

Triplet phases play a key role in 'Direct Methods' and it will be shown that they are physically measurable quantities. They consist of a sum of three structure-factor phases. In some cases, however, it is even possible to measure the phase of a single reflection if one reflection is a seminvariant reflection and if the other two are correlated by symmetry. Seminvariant reflections do not change their phase if one of the symmetrically equivalent origins of the unit cell is chosen. Let (**R**,**t**) be the rotational and translational part of a space-group symmetry operation. If triplets of the kind

$$\Phi_T = \phi(\overline{\mathbf{h}}_S) + \phi(\mathbf{g}) + \phi(-\mathbf{g}\mathbf{R}) \tag{5}$$

can be found the phase of the **g** reflection cancels since $\phi(\mathbf{g}\mathbf{R}) = \phi(\mathbf{g}) - 2\pi\,\mathbf{t}\,\mathbf{h}$. If Φ_T is known, the calculation of $\phi(\mathbf{h}_S)$ is straightforward. In 'Direct Methods' triplets like (5) are called Σ_1 relationships.

The direct measurement of phase relationships between X-ray reflections is only possible by means of an interference experiment. Hereby, it is necessary to superpose two waves with exactly the same wave vector **K**. If two waves $A\,e^{i\alpha}$ and $B\,e^{i\beta}$ with amplitudes A, B and phases α, β interfere the resulting intensity is given by[1]

$$I = A^2 + B^2 + 2\,A\,B\,\cos(\pm(\alpha-\beta)). \tag{6}$$

Equation (6) shows that the intensity depends on the phase difference of the two waves. The idea how this can be achieved in a diffraction experiment by means of a three-beam case was born already in 1949 by Lipscomb.

[1]It has been assumed that they have the same **K** vectors, so the **K r** term in the complex exponential functions has already been omitted.

2 Three-beam interference

In a three-beam case there are besides the origin two other reciprocal lattice nodes on the Ewald sphere. This can for example be achieved by the so-called Ψ-scan technique. Hereby, one reciprocal lattice vector **h** is brought to its diffraction position. This reflection is considered to be the primary one. By means of a rotation around **h** a secondary reciprocal lattice node **G** is turned on to the Ewald sphere. This situation is depicted in Fig. 1. The secondary wavefield with **K(g)** can in part be diffracted by the reciprocal

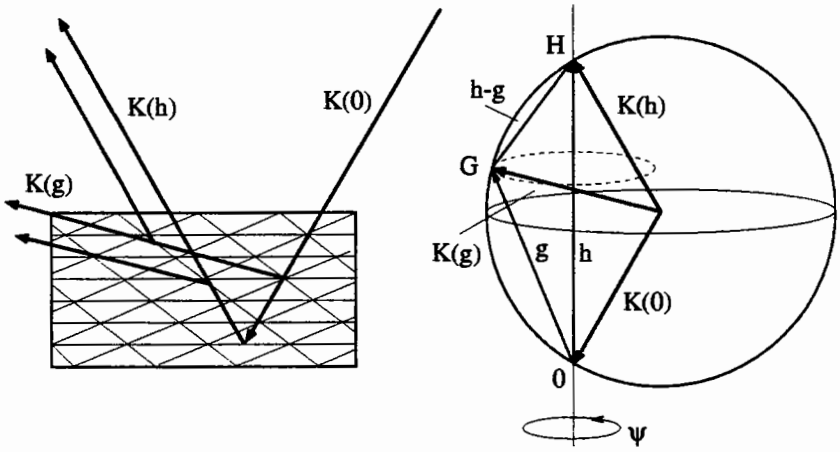


Figure 1: Three-beam case: schematical representation in crystal and reciprocal space with primary reflection **h** and secondary reflection **g**; for simplicity all three **K** vectors are drawn co-planar.

lattice vector **h-g** into **K(h)** direction[2]. Therefore, two wavefields are propagated in **K(h)** direction, the primary one scattered from the **h** net planes which has a phase shift $\phi(\mathbf{h})$ and the so-called Umweg wave (detour wave) scattered from **g** and **h-g** with the corresponding phase shift $\phi(\mathbf{g}) + \phi(\mathbf{h}-\mathbf{g})$. According to (6) the intensity in direction **K(h)** should depend on $\pm(\phi(\mathbf{g})+\phi(\mathbf{h}-\mathbf{g})-\phi(\mathbf{h}))$ and on the amplitudes of the primary and the Umweg wave. This qualitative interpretation already proposed by Lipscomb (1949) is however not sufficient to describe the intensities during a three-beam case. For an exact description the dynamical theory of diffraction has to be applied (Colella, 1974; Pinsker, 1978; Hümmer & Billy, 1982; Chang, 1984; Chang, 1987; Weckert & Hümmer, 1990; Weckert & Hümmer, 1997). It is beyond the scope of this article to give a full treatment of this theory for the three- or multi-beam case. For further reading please consult the cited literature. Here only the basic results for understanding the principles of three-beam diffraction will be summarized. In a perturbational approach (Bethe approximation (Bethe, 1928)) the amplitude in **K(h)** direction can be written as (Weckert & Hümmer, 1997)

$$\frac{\mathbf{D(h)}}{\mathbf{D(0)}} = N^{-1}R(\mathbf{h})\left(\alpha_{0h}\Gamma F(\mathbf{h}) + R(\mathbf{g})\alpha_{0g}\alpha_{hg}\Gamma^2 F(\mathbf{g})F(\mathbf{h}-\mathbf{g})\right) = N^{-1}R(\mathbf{h})F_{eff} \qquad (7)$$

hereby $N = 1 - \alpha_{hg}^2(\Gamma F(\mathbf{h}-\mathbf{g}))^2 R(\mathbf{g})R(\mathbf{h})$, **D** denote the amplitudes of the wave fields, α_{ij} are coupling scalar products, $\Gamma = r_e\lambda^2/\pi V_{uc}$ is a constant characterizing the coupling

[2]The same holds for the wavefield with **K(h)** via **g-h** into **K(g)** direction.

between the crystals electrons and X-rays, $r_e = 2.81 \cdot 10^{-15}\, m$ is the classical electron radius and the resonance terms $R(\mathbf{h}_m)$ are given by

$$R(\mathbf{h}_m) \approx \frac{\mathbf{K}(\mathbf{h}_m)^2}{\mathbf{K}_0^2 - \mathbf{K}(\mathbf{h}_m)^2} \approx |R(\delta)|\, e^{i\Delta\phi(\delta)}. \tag{8}$$

The angle δ represents either ω for $\mathbf{h}_m$=$\mathbf{h}$ or Ψ for $\mathbf{h}_m$ =$\mathbf{g}$, $\mathbf{K}_0$ is the wave vector of the incident radiation inside the crystal. Since $\mathbf{K}(\mathbf{h}_m) = \mathbf{K}_0 + \mathbf{h}_m$ holds from (8) it is obvious that if absorption is taken into account the resonance terms $R(\mathbf{h}_m)$ behave like Lorentzians with a phase shift from 0 to π if a reciprocal lattice vector moves from the inside ($|\mathbf{K}(\mathbf{h}_m)| < |\mathbf{K}_0|$) to the outside ($|\mathbf{K}(\mathbf{h}_m)| > |\mathbf{K}_0|$) of the Ewald sphere. For the further discussion we assume that $N \approx 1$ holds in (7). Comparing (7) with (6) the total phase difference for the interference of the two waves can be deduced:

$$\Phi_{tot}(\Psi) = \pm((\phi(\mathbf{g}) + \phi(\mathbf{h}-\mathbf{g}) + \Delta\phi(\Psi) - \phi(\mathbf{h})) \approx \Phi_T + \Delta\phi(\Psi). \tag{9}$$

In Fig. 2 a schematical drawing of amplitude and phase of the resonance term are shown. Suppose the triplet phase of a three-beam case $\mathbf{0/h/g}$ is zero: $\Phi_T = 0°$. Then, at the

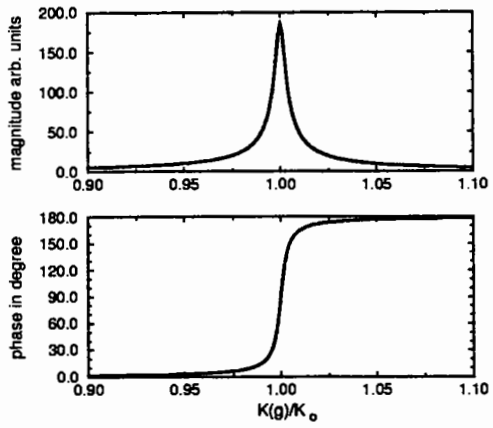


Figure 2: Schematical drawing of amplitude (magnitude) and phase of the resonance term R(g) close to the three-beam position.

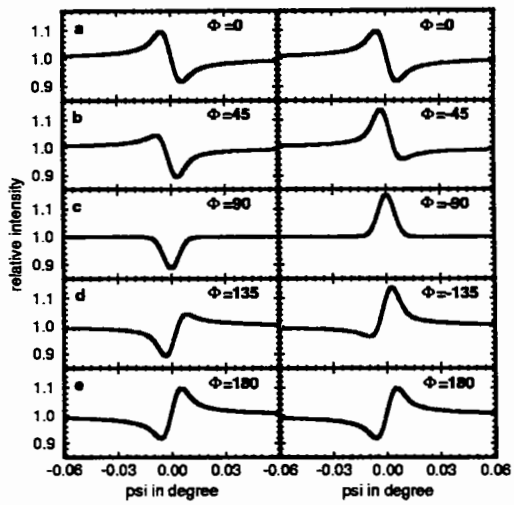


Figure 3: Interference-profiles for different triplet phases

beginning of the Ψ-scan $\Delta\phi(\Psi) = 0$ and $\Phi_{tot}(\Psi)$ is zero as well. The amplitude of the Umweg wave is small and the two-beam intensity for $\mathbf{h}$ is observed. Scanning towards the three-beam position the amplitude of the Umweg wave increases. The primary wave and the Umweg wave interfere in a constructive way which leads to an increase in the resultant amplitude of $\mathbf{D}(\mathbf{h})$. Near to the three-beam position $\Delta\phi(\Psi)$ shifts rapidly from 0 to 180°, then $\Phi_{tot}(\Psi) = 180°$. That means that the interference becomes destructive and the two-beam intensity is decreased. At the end of the Ψ-scan when the amplitude of the Umweg wave decreases, the two-beam intensity is observed again. A calculated profile of this type is shown in Fig. 3a. It reflects the fact that $\cos[\Phi_{tot}(\Psi)]$ changes its sign as $\Phi_{tot}(\Psi)$ varies from 0 to 180°. The profile forms for other triplet phases Φ_T also shown in Fig. 3 can be explained analogously. In Fig. 3 the ratios of the structure factor moduli have been chosen appropriately that for example the destructive interference for $\Phi_T = 90°$ is comparable to the constructive interference for $\Phi_T = -90°$. However, in general this is not

the case since additional symmetric effects which do not depend on the sign of the triplet phase occur which superpose the pure interference effects shown in Fig. 3 (Weckert & Hümmer, 1990; Weckert, Schwegle & Hümmer, 1993). By comparison of the profiles for two centrosymmetrically related three-beam cases **h, g, h-g** and $\overline{\mathbf{h}}, \overline{\mathbf{g}}, \mathbf{g}-\mathbf{h}$ shown in the left and right column of Fig. 3, respectively, these Umweganregung (increase) or Aufhellung (decrease) effects can be recognized and eliminated.

3 Experimental

The measured triplet-phase sensitive signal is the change of the intensity of a reflection due to the interference with a second additionally excited one. This means that the rotation around the primary reciprocal lattice vector **h** has to be very accurate as otherwise spurious intensity modulations will occur and spoil any interference pattern. For this purpose a special Ψ-circle diffractometer has been constructed which is able to perform a Ψ-scan by the rotation of a single axis only. The angular resolution of this diffractometer for those circles that move the crystal is $0.0002 - 0.00005°$. In Fig. 4 a schematical drawing of the diffractometer is shown. As the detector is mounted on two perpendicular

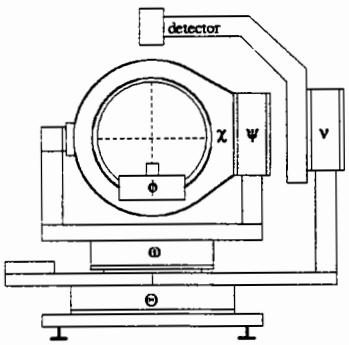


Figure 4: ψ-circle diffractometer

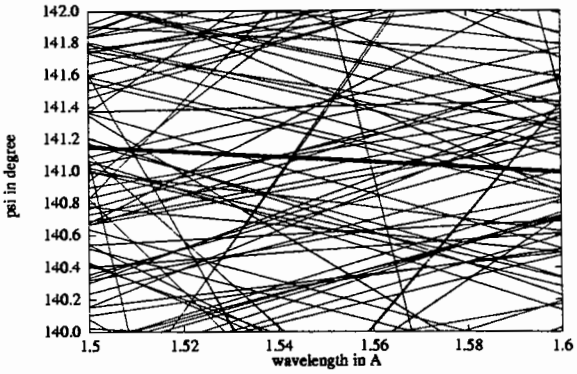


Figure 5: Three-beam positions in dependence on ψ and λ for tetragonal lysozyme with $V_c = 238000\text{Å}^3$. Only three-beam cases with $q > 0.25$ (see text) for the primary reflection 4 7 0 are shown. The thick line shows the position of the three-beam case $4\,7\,0/2\,5\,1/2\,\bar{2}\,\bar{1}$.

circles it can be moved to any direction in the upper half sphere. Thus, also the diffracted intensity of the secondary reflection **g** during the Ψ-scan can be measured which is very important for large structures to obtain the accurate three-beam position.

In Fig. 1 only one secondary reciprocal lattice vector is shown. In reality the number of secondary vectors can be very large. For the crystal structure of a small amino acid at $\lambda = 1.5405$ Å for a full turn in Ψ about 6000 three-beam cases occur. This means on average one three-beam case for a $\Delta\Psi$ of $0.05°$ which is too narrow so that the interference profiles of neighbouring three-beam cases would overlap. However, it is possible to find larger gaps for some three-beam cases since the Ψ positions are not equally spaced to measure an undisturbed interference profile. The Ψ positions of different three-beam cases to one particular primary reflection depend very sensitively on the wavelength. Hence, by searching for a suitable wavelength a three-beam case of interest can be separated from neighbours for small and medium size structures. In case of macromolecular

structures even by changing the wavelength overlap of different interference profiles can not be avoided. Owing to the fact that the number of weak reflections in macromolecular structures is large the wavelength for a given three-beam case with large structure factor moduli can be selected properly that all neighbouring three-beam cases have significant smaller structure factors. Assuming the interesting three-beam case is **h, g, h-g** with structure factors F(**h**), F(**g**) and F(**h-g**) then it has been shown experimentally as well as theoretically (Weckert, Schwegle & Hümmer, 1993; Weckert & Hümmer, 1997) that the interference effect of neighbouring case **h, g', h-g'** can be neglected if

$$q = \frac{F'(\mathbf{g}')F'(\mathbf{h}-\mathbf{g}')}{F'(\mathbf{g})F'(\mathbf{h}-\mathbf{g})} \lesssim 0.25. \tag{10}$$

The F' are structure factor moduli corrected for polarization. In these cases it is adequate to search for a suitable wavelength that all three-beam cases with $q > 0.25$ are sufficiently far away from the interesting one. In Fig. 5 an example for a particular triplet of tetragonal lysozyme is given. Due to the necessity to change the wavelength this experiments require synchrotron radiation which helps also due to its high brilliance to measure the comparable small interference effects.

The crystals used for experimental-phase determination by three-beam interferences are of normal size (0.05 - 1 mm). Protein crystals have been sealed in capillaries together with some mother liquid. The mosaic spread should be as small as possible. However, the crystals do not have to be perfect. As long as a crystal consists out of a few larger mosaic blocks whose reflection profile can be separated by the incident radiation[3] three-beam interference experiments with single mosaic blocks are possible. Difficulties arise if the mosaic distribution is smooth and wide.

In order to calculate the influence of possible neighbouring triplets and to search for suitable three-beam case an intensity data set as complete as possible is required. For protein crystals also all reflections at low resolution have to be measured.

4 Three-beam experiments with protein crystals

In the past years three-beam interference experiments with various proteins have been carried out (Hümmer, Schwegle & Weckert, 1991; Chang, King, Huang & Gao, 1991; Weckert, Schwegle & Hümmer, 1993; Weckert & Hümmer, 1997). The first interference experiments were observed with crystals from sperm whale myoglobine. Later other proteins were investigated like a Fab - fragment (space group: $P\,2_1\,2_1\,2_1, V \approx 280000\text{Å}^3$), triclinic and tetragonal hen-egg white lysozyme, proteinase K (space group: $P\,4_3\,2_1\,2, V \approx 500000\text{Å}^3$) and trypsin. All experiments were carried out with synchrotron radiation either at beam line C of HASYLAB in Hamburg or from an ESRF bending magnet (Swiss-Norwegian beamline, Grenoble).

In the very beginning wavelengths around 1.54Å were used. In this wavelength range radiation damage is severe. For this reason higher energies ($\approx$ 1-1.1Å) were selected for more recent experiments. Three-beam interference effects could be observed up to a unit cell size of $1.2 \cdot 10^6\text{Å}^3$ (catalase oxidoreductase). For triclinic lysozyme it was possible to measure triplet phases where reflections up to a resolution of 2Å were involved (2.5Å for tetragonal lysozyme). With the present set-up at an ESRF bending magnet about 6

[3]Using radiation from an ESRF bending magnet, mosaic blocks which are not more than $\approx 0.003°$ inclined towards each other can already be separated.

triplet phases per hour can be measured with a 600 μm crystal of tetragonal lysozyme in the resolution range of 3-6 Å. For a 150 μm crystal of proteinase K in the same resolution range about three triplet phases per hour are still possible. This number can be increased if a more brilliant source is available. The maximum number of triplet phases that could be measured from a single protein crystals of tetragonal lysozyme was about 150. before the radiation damage was too strong. If crystals of very small mosaicity are available the intensity changes owing to the interference effects are in the order of 5 to 15%. An example for three-beam interference profiles of tetragonal lysozyme is given in Fig. 6. In Fig. 7 the influence of the radiation damage on a three-beam interference profile is

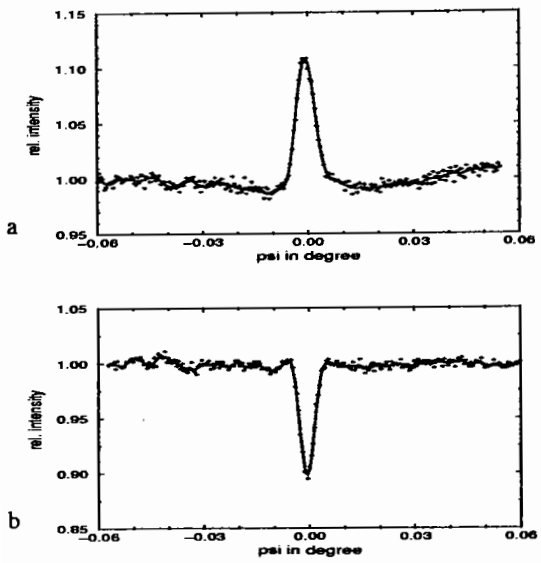


Figure 6: Measured three-beam Ψ-scan profiles with an estimated triplet phase of $\mp 90°$ from tetragonal lysozyme at λ = 1.3047Å; a: three-beam case: $8\,11\,\overline{8}/1\,4\,\overline{2}$, $\Phi_3^{calc} = -107°$ (entry 1lse of PDB), b: three-beam case: $\overline{8}\,\overline{11}\,8/\overline{1}\,\overline{4}\,2$; $|F(8\,11\,\overline{8})| = 727, |F(1\,4\,\overline{2})| = 1319, |F(7\,7\,\overline{6})| = 1157$, exp. conditions: ESRF, Si 1 1 1 monochromator, π-polarization.

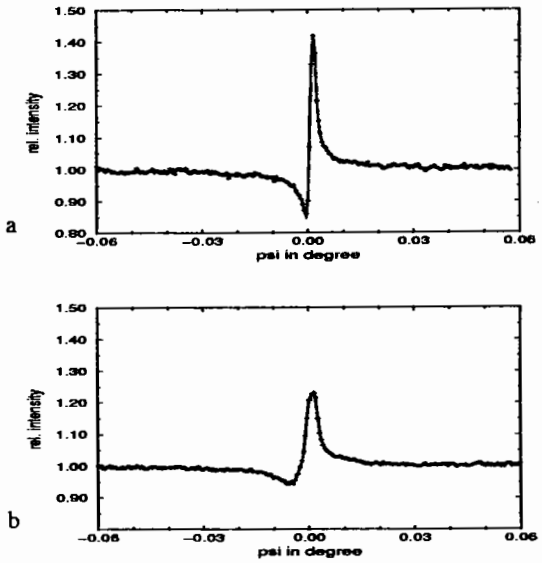


Figure 7: Measured three-beam Ψ-scan profiles with an estimated triplet phase of 180° and Umweganregung from tetragonal lysozyme at λ = 1.3302Å. a: three-beam case: $\overline{7}\,4\,0/\overline{5}\,2\,1/\overline{2}\,2\,\overline{1}$, $\Phi_3^{calc} = 173.3°$ (entry 1lse of PDB), b: same three-beam case as in (a) after 36h exposure to X-rays; $|F(\overline{7}\,4\,0)| = 1065, |F(\overline{5}\,2\,1)| = 2027, |F(\overline{2}\,2\,\overline{1})| = 2902$; exp. conditions: ESRF, Si 1 1 1 monochromator, π-polarization.

demonstrated. After 36 h of exposure the interference effect is only half as pronounced as for the undamaged crystal.

The mean error for the measured triplet phases of all investigated compounds compared to the known structure models was about 20°. In order to test the feasibility of triplet-phase data collection and also to develop a suitable strategy it was attempted to measure a larger number of triplet phases from tetragonal lysozyme. Meanwhile, more then 700 triplet-phases have been measured which contain about 630 different single phases. The distribution of the resolution of these phases is shown in Fig. 8. The maximum of this distribution is at about 4Å.

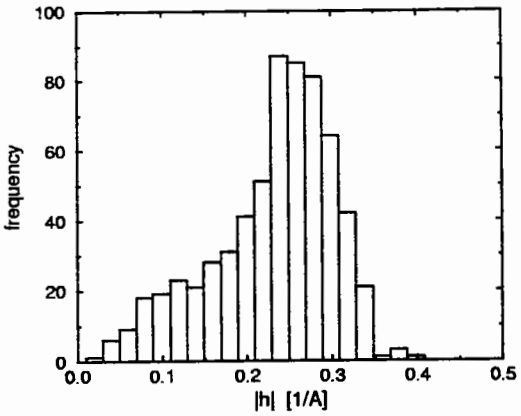


Figure 8: Distribution of the resolution of triplet reflection measured from tetragonal lysozyme.

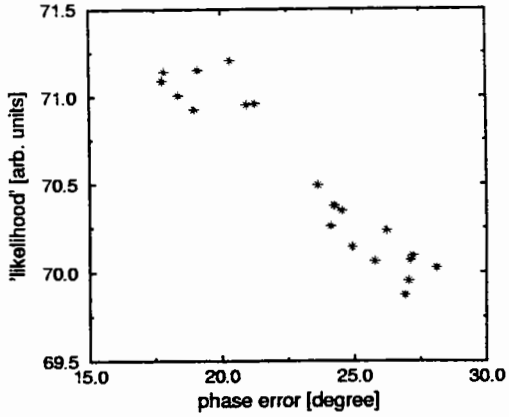


Figure 9: Correlation between mean weighted phase error and likelihood for different magic integer permutations.

5 Structure determination using experimental phases

In order to apply (2) to calculate an electron-density map single-structure factor phases are needed, which require the choice of an origin. From the 630 reflections two reflections which occur most frequently in different triplets were selected to fix the origin. The phases of this two reflections were taken from the known structure model for comparison. Among the 710 triplet phases were 24 Σ_1 relationships according to (5) which provide single phases. These 26 single phases can now be connected by other measured triplets to give further new single phases. To keep error propagation small no single phase should depend on a maximum number of measurements. For important reflections more than one phasing branch can be used to fix the phase. Since there are more than 700 triplet phases available for 630 reflections the dataset shows some redundancy. Nevertheless, not all reflections could be assigned with a phase. The phase of these reflections were treated as symbols which had to be permuted by a magic integer algorithm (Main, 1977). For each permutation a maximum entropy map and the corresponding likelihood was calculated (Bricogne & Gilmore, 1990). In Fig. 9 the likelihood as function of the mean-weighted phase error for each permutation is drawn. It is obvious that likelihood seems to be a suitable criteria to discriminate permutation with lower phase errors. An electron-density map calculated with the phases from one of the permutations with smaller phase error shows already large portions of the molecule despite the fact of some main chain breaks.

There should be a number of other possibilities to take advantage of experimental phase information. One of them is certainly in 'Direct Methods' where estimated triplet phases can be substituted by measured ones.

6 Conclusion

It has been shown that the direct determination of triplet phases even from crystals of small proteins is possible provided a stable tune able synchrotron-radiation source is available. The accuracy for the phases that can be achieved seems to be sufficient. Introducing the experimental measured triplets into a maximum entropy based approach is capable to provide single phase which can be used to calculate a map. Compared to

other experimental phasing methods like MAD the three-beam interference method is slower and crystals of very low mosaic spread have to be used. On the other hand the phase information provided by the three-beam interference method can be obtained from a native protein crystal without introducing heavy atoms. Protein crystals very often show a very small mosaic spread, however, the radiation damage can be quite severe. This seems to be the main problem since the shock freezing method which is successfully applied for intensity data collections produces in general a mosaic spread which is too wide for the application of the three-beam interference method.

References

Altomare, A., Cascarano, G., Giacovazzo, A., Burla, M., Polidori, G. & Camalli, M. . J. Appl. Cryst., *27* (1994) 435

Bethe, H. A. Ann. Phys.(Leipzig), *87* (1928) 55–129

Bricogne, G. & Gilmore, C. J. Acta Cryst., A*46* (1990) 284–297

Chang, S. L. *Multiple Diffraction of X-rays in Crystals*, (1984) Berlin, Heidelberg, New York: Springer Press.

Chang, S. L. Crystallogr. Rev., *1* (1987) 87–189

Chang, S. L., King, H. E., Huang, M. T. & Gao, Y. Phys. Rev. Lett., *67* (1991) 3113–3116

Colella, R. Acta Cryst., A*30* (1974) 413–423

Debaerdemaeker, T., Tate, C. & Woolfson, M. M. Acta Cryst., A*44* (1988) 353–357

Hümmer, K. & Billy, H. Acta Cryst., A*38* (1982) 841–848

Hümmer, K., Schwegle, W. & Weckert, E. Acta Cryst., A*47* (1991) 60–62

Lipscomb, W. N. Acta Cryst., *2* (1949) 193–194

Main, P. Acta Cryst., A*33* (1977) 750–757

Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. J. Appl. Cryst., *27* (1994) 613–621

Pinsker, Z. G. *Dynamical Scattering of X-rays in Crystals.* (1978) Berlin: Springer Press.

Sheldrick, G. M. Acta Cryst., A*46* (1990) 467–473

Weckert, E. & Hümmer, K. Acta Cryst., A*46* (1990) 387–393

Weckert, E. & Hümmer, K. (1997) Acta Cryst., A*53* (1997) in press

Weckert, E., Schwegle, W. & Hümmer, K. Proc. R. Soc. Lond. A, *442* (1993) 33–46

Woolfson, M. M. & Fan, H.-F. *Physical and Non-Physical Methods of Solving Crystal Structures.* (1995) Cambridge: Cambridge University Press

Wüthrich, K. Acta Cryst., D*51* (1995) 249–270

Direct methods - overview for macrocrystallographers

Zbigniew Dauter and Peter Main

Depts. of Chemistry and Physics, University of York, YO1 5DD

The term *´direct methods´* is used in small molecule crystallography to describe methods of structure solution, that is to say methods for phase derivation, by purely mathematical means *utilising the measured structure amplitudes only*. In a diffraction experiment it is only the structure factor amplitudes $|F_{hkl}|$ that are measured ($|F_{hkl}| = \sqrt{I_{hkl}}$). We can see that if we express the electron density as a Fourier transform of the structure factors:

$$\rho(xyz) = \Sigma_{hkl} |F_{hkl}| \exp(i\varphi_{hkl}) \exp 2\pi i(hx+ky+lz)$$

then the only unknowns are the phases of the structure factors, φ_{hkl}. The knowledge of the phases is much more important than that of the amplitudes, as can be seen from the following relationship, based on the principle, that the Fourier transform of the product of two functions is equal to the convolution of individual transforms:

$\|F_{hkl}\|$	x	$\exp(i\varphi_{hkl})$	=	F_{hkl}
FT		FT		FT
amplitude synthesis (≈Patterson)	*	phase synthesis	=	$\rho(xyz)$

The Fourier transform of the amplitudes gives a function very similar to the Patterson, which has a huge peak at the origin and does not correspond to the actual electron density $\rho(xyz)$. Most of the information about positions of the atoms in the crystal, or peaks in $\rho(xyz)$, must be contained in the phases of the structure factors. Therefore the fundamental problem in the crystallographic diffraction analysis is the *phase problem*.

Several methods of solving the phase problem exist, starting from trial and error modelling for the simplest of the structures, through interpretation of Patterson function with unknown or known (Molecular Replacement) structural models, to rather tedious methods utilising the signal from the heavy atoms either already present in the structure (e.g., MAD) or substituted into the structure (MIR). The term *direct methods* traditionally refer to methods of phase calculation which utilise analytical mathematical (probabilistic) *equations based only on the observed structure factor amplitudes*.

The structure factors, i.e. their amplitudes and phases, in general, depend on the distribution of atoms in the unit cell of the crystal:

$$F_{hkl} = |F_{hkl}| \exp(i\varphi_{hkl}) = \sum_j f_j \exp(^{-B}/_{4d^2}) \exp -2\pi i(hx_j+ky_j+lz_j)$$

The atomic coordinates, x_j,y_j and z_j are expressed as fractions of the cell edges and relate to a common reference point, the origin of the cell. It is convenient to fix the origin at symmetry positions such as a center of symmetry if it exists. In other space groups, such as $P2_1$, it may lie on the screw axis anywhere along b direction. Moreover in most space groups there are several special positions of the same symmetry and any of them can be selected as the origin of the cell. Change of the origin will not change the amplitude but in general may change the individual phases. The table shows how the phases (or in this case, signs) of reflections with different parity of their indices change when the origin is shifted between eight possible centres of symmetry in space group P-1.

origin shift parity	0,0,0	$^1/_2$,0,0	0,$^1/_2$,0	0,0,$^1/_2$	0,$^1/_2$,$^1/_2$	$^1/_2$,0,$^1/_2$	$^1/_2$,$^1/_2$,0	$^1/_2$,$^1/_2$,$^1/_2$
eee	+	+	+	+	+	+	+	+
eeo	+	+	+	-	-	-	+	-
eoe	+	+	-	+	-	+	-	-
oee	+	-	+	+	+	-	-	-
eoo	+	+	-	-	+	-	-	+
oeo	+	-	+	-	-	+	-	+
ooe	+	-	-	+	-	-	+	+
ooo	+	-	-	-	+	+	+	-

In this case only reflections with all three indices even do not change their phase when the origin is shifted. In the process of *ab initio* phase estimation it is necessary to ensure that all phases form a consistent set and refer to a common origin. Analysis of phase dependence on the selection of different origins, permissible in given space group, leads to the concept of *structure seminvariants*. They are the structure factors or their linear combinations, whose phase does not depend on the choice of the origin, under the condition that it is allowed for the particular space group. One of the simplest seminvariants is formed by a so-called Σ_2 triplet of three structure factors E_h, E_k and E_{-h-k}, for which the sum of indices is zero.

The selection of one of the several possible origins can be done by a free choice of phases for three (or less, for centered or higher symmetry cells) reflections, which do not form seminvariants.

The equation for the electron density does not provide a direct relationship between structure factor amplitudes and phases. If the electron density was completely unknown, the amplitudes and phases would need to be treated as completely independent. Fortunately, we have some expectations about the electron density which indirectly constrain the terms in the right hand side of the electron density equation. Since the amplitudes are known, those constraints can be utilised to formulate some phase restrictions. Many analytical or probabilistic relationships of different strength and usefulness have been proposed. For the pioneering work in this field, setting out the basis of the direct methods, Jerome Karle and Herbert Hauptman were awarded Nobel Prize in Chemistry in1985.

The features of the electron density which can be expressed mathematically and used in structure determination are set out here:

1.	atomicity of $\rho(x)$	normalised structure factors
2.	positivity of $\rho(x)$	inequalities and determinants
3.	equal atoms	Sayre´s equation
4.	$\int\rho^3(x)\,dV = max.$	tangent formula
5.	$-\int\rho(x)\ln\rho(x)\,dV = max.$	maximum entropy methods

6.	partial structure	modification of probability equations
7.	multiple motifs	molecular replacement
8.	$\rho(x) = \text{const.}$	solvent flattening and density modification

It is known from the principles of chemistry that the atoms cannot lie closer together than a certain distance. The electrons are concentrated to a certain volume around the atoms and the thermal vibration smears out the electron density to some extent around the average atomic positions, but in general the electron clouds of separate atoms do not overlap considerably. This can be utilised to remove the effect of the atomic or, rather, electron cloud shape (represented by the term $f_j \exp(^{-B}/_{4d^2})$) from the structure factor. Removal of the term $f_j \exp(^{-B}/_{4d^2})$ from the structure factors F_{hkl} , substituting them by the *normalised structure factors*, E_{hkl} leads to the deconvolution of the point atom structure, as can be seen from the following relationship:

E_{hkl}	x	$f_j \exp(^{-B}/_{4d^2})$	=	F_{hkl}
FT		FT		FT
point atom structure	*	real atom	=	$\rho(xyz)$

This can be done by dividing the structure amplitudes by their average value in the resolution ranges:

$$|E_{hkl}|^2 = I_{hkl} / \langle I \rangle$$

and can be represented by the Wilson plot (1942) which in average is horizontal. This procedure weights up the high resolution intensities, intrinsically small due to the atomic shape and its vibration and allows the selection of the relatively largest structure factors in all resolution ranges. Direct methods anyway usually utilise only a subset of the largest amplitudes in the process of phase estimation.

The electron density must not be negative, otherwise it has no physical meaning. This constraint leads to the formulation of inequality relations, which were the first of the

mathematical expressions connecting the phases and amplitudes of the structure factors, given by Harker and Kasper (1948). An example of such inequality in terms of unitary structure factors ($U_{hkl} = F_{hkl} / F_{000}$):

$$U^2_{hkl} \leq {}^1/_2 \, (1 + U_{2h2k2l})$$

is valid in P-1. If both U_{hkl} and U_{2h2k2l} are sufficiently large, the inequality relationship can prove that the sign of U_{2h2k2l} must be positive. Such relationships were generalised by Karle and Hauptman (1950) and also expressed in the form of determinants. However, inequalities are not powerful enough and are not used in practice any more.

If we neglect the hydrogens, the assumption that the crystals of organic compounds consists of equal atoms is a good approximation. The diffracting power of carbon, nitrogen and oxygen with 6, 7 and 8 electrons are similar. If we also take into account atomicity of the electron density, it leads to so-called Sayre´s equation. This was formulated in three papers published in 1952 by Sayre, Cochran and Zachariasen in the same volume of Acta Cryst.

If the electron density within the crystal consisting of equal atoms is squared, the resulting ´squared´ density is almost proportional to the original, except that the atomic peaks have a somewhat different shape. We can introduce the structure factors of the squared structure,

$$G_{hkl} = g \sum_j \exp 2\pi i \, (hx_j+ky_j+lz_j) = {}^g/_f \, f \sum_j \exp 2\pi i \, (hx_j+ky_j+lz_j) = {}^g/_f \, F_{hkl}$$

On the other hand, from the following convolution

$\rho(r)$	x	$\rho(r)$	=	$\rho^2(r)$
FT		FT		FT
F_h	*	F_h	=	G_h

it can be shown that $G_h = {}^1/_V \, F_h * F_h = {}^1/_V \sum_k F_k \, F_{h-k}$, and we obtain the Sayre´s equation:

$$F_h = {}^1/_V \, {}^f/_g \sum_k F_k \, F_{h-k}$$

which gives exact relationship among the structure factors. It is the most important equation in direct methods and forms the basis of phase propagation and refinement.

Closely related to the Sayre´s equation is the tangent formula, given by Cochran (1956), which can be expressed as follows:

$$\tan \varphi_h = \frac{\sum_k |E_k E_{h-k}| \sin (\varphi_k + \varphi_{h-k})}{\sum_k |E_k E_{h-k}| \cos (\varphi_k + \varphi_{h-k})}$$

The tangent formula is based on the probability considerations for the distribution for the unknown phase φ_h with the other phases known. The reliability of the formula depends on the value of:

$$\alpha_h = 2/\sqrt{N} \; |E_h| \; | \sum_k E_k E_{h-k} |$$

A simplified conclusion from the tangent formula and Sayre´s equation is that

for non-centrosymmetric crystals E_h has phase of $\{ \sum_k E_h E_k E_{h-k} \}$

and for centrosymmetric crystals E_h has sign of $\{ \sum_k E_h E_k E_{h-k} \}$

If the values of all three normalised structure factors of the Σ_2 triplet are large, there is a high probability that even for a single triplet their phases sum to zero, or (for centrosymmetric crystals) the product of their signs is positive. This is the basis of the *symbolic addition* procedure, introduced in the early 1960's by Isabella Karle. Symbolic addition was widely used in the era before automatic programs and fast computers became available.

In this method the phases of some reflections were represented by letter symbols, and together with the origin fixing reflections constitute the starting set. In non-centrosymmetric space groups the phase of one more reflections can be chosen to fix the enantiomorph. The symbols are then propagated through a number of Σ_2 relations, so that a number reflections have phases represented by symbols. Some reflections may have several different estimations

expressed by different symbols, which provides additional relations between symbols or allows to assign specific phase to a symbol. The example below illustrates the procedure for P-1 symmetry.

origin fixing reflections:	3	0	2	+
	2	-3	2	+
	1	2	-1	-
symbols	0	3	3	a
	3	2	4	b

phase propagation:

3	0	2	+		3	0	2	+	
2	-3	2	+		-1	-2	1	-	
5	-3	4	+		2	-2	3	-	
1	2	-1	-		-2	3	-2	+	
0	3	3	a		3	2	4	b	
1	5	2	-a		1	5	2	b	

From the last two relations it is evident that b = -a, and the number of symbols can be reduced. This procedure finally leads to a single combination of phases which hopefully provides an interpretable E-map.

In the *multisolution approach*, introduced in 1970's in York in the program MULTAN (Germain, Main & Woolfson, 1970), the phases of the reflections in the starting set are permuted and those combinations then propagated and refined, thus producing a number of potential solutions. The starting phases can be permuted in a simple way, with centrosymmetric reflections having 0 or 180° and non-centrosymmetric ones 45, 135, 225 or 335°, giving either 2^n or 4^n combinations. Another method of sampling the phase space more effectively with less permutations is based on the idea of magic integers (White & Woolfson, 1975). The phases of the starting set reflections are expressed as a function of a single variable:

$\varphi_i = m_i\ x \qquad \mathrm{mod}(2\pi)$

for a sequence of mutually prime integer numbers m_i.

A variation of the multisolution method, which gained more popularity with the increased availability of faster and larger computers, is the *random approach*. The random phases can be assigned to the limited phase set, as in the multisolution approach, and then propagated and refined as in the program RANTAN (Yao, 1981), or all phases given the random values and then refined to consistency as in SHELXS (Sheldrick, 1990). The last approach is gaining more popularity lately and is implemented in most of the contemporary direct methods programs.

The multisolution and random methods create a large number of phase sets, some of which are correct, leading to interpretable E-maps, and some incorrect. The identification of the correct set(s) is not easy and requires the use of reliable *figures of merit*, which test the quality of phases. Several different FOM's have been proposed and used in different programs.

ABSFOM checks the internal consistency of triplets, its value should be 1 for correct set and 0 for random phases. Rα checks the deviations from the expected values of α_h. Ψ_0 makes use of the triplets with E_h small. Obviously such reflections do not take part in the phase refinement and therefore provide an independent check of phase correctness. NQEST is based on negative quartets, for which E_h, E_k, E_k, E_{h+k+l} are large but E_{h+k} , E_{k+l} and E_{l+h} are small. The phase of such seminvariant $\varphi_4 = \varphi_h + \varphi_k + \varphi_l + \varphi_{h+k+l}$ is expected to be 180°, therefore NQEST should have a negative value for good phases. In practice often the direct methods programs use a combination of several figures of merit and select the best phase set according to combined figure of merit.

The structure determination by direct methods consists of the following steps:

1. Calculation of normalised structure factors from F_{obs} and selection of a set of large E's

2. Setting up Σ_2 phase relationships, $E_h\ E_k\ E_{h-k}$
3. Phase assignment to the starting set, including origin fixing
4. Phase propagation and refinement
5. Calculation of figures of merit
6. Computation and interpretation of the E-map.

In contemporary direct methods programs all these steps can be performed automatically, without manual intervention. Indeed, most of small structures with up to about 100 atoms can be solved using the default program options. This is not the case for larger molecules, including small proteins, for which the process of solving the structure by direct methods is far from routine and requires diffraction data to extend beyond 1.2 Å, involves generation of an enormous number of phase sets and the use of efficient figures of merit. Largest structures solved so far by direct methods have 400 - 500 atoms (Dauter, Lamzin & Wilson, 1995, Sheldrick et al., 1993). Hovever, recently Herbert Hauptman (1996) formulated an optimistic opinion that within 10 years we should be able to solve structures with up to 1000 atoms at somewhat lower resolution.

References

Cochran, W. (1952) Acta Cryst., 5, 65 - 67.

Cochran, W. (1955) Acta Cryst., 8, 473 - 478.

Dauter, Z., Lamzin, V.S. & Wilson, K.S. (1995) Curr. Op. Struct. Biol., 5, 784 - 790.

Harker, D. & Kasper, J.S. (1948) Acta Cryst., 1, 70 - 75.

Hauptman, H. (1996) IUCr Congress, Seattle, Abstract no. BL.01, p. C-7.

Hauptman, H. & Karle, J. (1953) ACA Monograph No. 3, Polycrystal Book Service.

Karle, J. & Hauptman, H. (1950) Acta Cryst., 3, 181 - 187.

Karle, J. & Karle, I.L. (1966) Acta Cryst., 21, 849 - 859.

Germain, G., Main, P. & Woolfson, M.M.(1970) Acta Cryst., B26, 274 - 285.

Sayre, D. (1952) Acta Cryst., 5, 60 - 65.

Sheldrick, G.M. (1990) Acta Cryst., A46, 467 - 473.

Sheldrick, G., Dauter, Z., Wilson, K., Hope, H. & Sieker, L. (1993) Acta Cryst., D49, 18 - 23.

White, P.S. & Woolfson, M.M. (1975) Acta Cryst., A31, 53 - 56.

Wilson, A.J.C. (1942) Nature(London), 150, 151.

Yao, J.-X. (1981) Acta Cryst., A37, 642 - 644.

Zachariasen, W.H. (1952) Acta Cryst., 5, 68 - 70.

Macromolecular Phasing by *Shake-and-Bake*

Charles M. Weeks[1] and Russ Miller[1,2]
[1]Hauptman-Woodward Med. Res. Inst., Inc., 73 High St., Buffalo, NY 14203 USA
[2]SUNY-Buffalo, Dept. of Comp. Sci., Buffalo, NY 14260 USA
weeks@hwi.buffalo.edu and miller@hwi.buffalo.edu

Abstract. *Shake-and-Bake* is a direct methods procedure which has provided *ab initio* solutions for protein structures containing as many as 600 independent non-H atoms, provided that good-quality diffraction data are available to 1.1Å resolution. Its ultimate potential is unknown. The *Shake-and-Bake* algorithm extends the range of conventional direct methods by repetitively, unconditionally, and automatically alternating reciprocal-space phase refinement with filtering in real space to impose constraints. An extensive web site for *SnB*, a computer program implementing *Shake-and-Bake*, can be found at URL http://www.hwi.buffalo.edu/SnB.

1 Introduction

The majority of small-molecule organic crystal structures having fewer than 100 independent non-H atoms are solved using reciprocal-space direct methods. Conventional direct methods begin to fail in the 100-200 atom range because the accuracy of the underlying probabilistic phase relationships decreases as the size of the structure increases. Fortunately, however, the size of molecular structures amenable to phasing by direct methods can be increased significantly if these methods are augmented by the imposition of physically-meaningful constraints in real space. Real-space phase-improvement methods, commonly known as density-modification methods, are widely used in macromolecular crystallography (see review by Podjarny *et al.*, 1987). The potential for real-space constraints to improve phases in the context of small-molecule direct methods was recognized by Jerome Karle (1968) who found that even a relatively small, chemically-sensible fragment extracted by manual interpretation of an electron density map could be parlayed into a complete solution by transformation back to reciprocal space and then performing additional iterations of phase refinement. The tremendous increases in computer speed in recent years have made it feasible to consider repeatedly cycling many trial structures back-and-forth between real and reciprocal space, while performing optimization alternately in each space. This compute-intensive process, which requires the use of two Fourier transforms during each cycle, forms the basis of the synergistic *Shake* (phase refinement) and *Bake* (density modification) procedure in which the power of reciprocal-space phase refinement is automatically augmented by filtering to impose the phase constraints implicit in real space (Weeks *et al.*, 1994).

Getting Started. Practitioners of conventional direct methods handle the problem of beginning a structure determination when no atomic positions are known by adopting a 'multisolution' approach in which multiple sets of trial phases are evaluated either symbolically (Karle & Karle, 1966) or numerically (Germain & Woolfson, 1968), with probable correct set(s) determined by ranking according to a suitable figure-of-merit. Since solving very large structures requires a large initial set of presumed-known phases, it has been found advantageous to generate trial phase sets by using a random-number generator to assign values to many or all of the required phases (Baggio *et al.*, 1978: Yao, 1981). In the *Shake-and-Bake* procedure, phases are assigned initial values by generating trial structures consisting of randomly positioned atoms (thereby imposing an atomicity constraint from the outset) and then computing structure factors.

Foundations of Phase Refinement. Direct methods are based on the fact that there exist linear combinations of phases, called structure invariants, whose values, in principle, depend only on the magnitudes of the normalized structure factors,

$$E_{\mathbf{H}} = |E_{\mathbf{H}}|\exp(i\varphi_{\mathbf{H}}) = (1/N^{1/2})\sum_{j=1}^{N}\exp\left(2\pi i\mathbf{H}\cdot\mathbf{r}_j\right) \qquad (1)$$

where $\mathbf{r}_j$ is the position vector of one of the N atoms, assumed identical, in the unit cell. The conditional probability distributions of the structure invariants permit individual invariant values to be estimated as first proposed by Hauptman and Karle (1953). The most useful phase relationships are the three-phase or triplet invariants,

$$T_{\mathbf{HK}} = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}}, \qquad (2)$$

which have the associated parameters (or weights),

$$A_{\mathbf{HK}} = \left(2/N^{1/2}\right)\left|\mathrm{E}_{\mathbf{H}}\mathrm{E}_{\mathbf{K}}\mathrm{E}_{\mathbf{H}+\mathbf{K}}\right|. \qquad (3)$$

Ab initio phase determination involves the derivation of individual phase values from a set of triplets having a sufficiently large triplet:phase ratio (e.g., 10:1). The tangent formula,

$$\tan(\phi_{\mathbf{H}}) = \frac{\sum_{\mathbf{K}}|E_{\mathbf{K}}E_{\mathbf{H}-\mathbf{K}}|\sin(\phi_{\mathbf{K}}+\phi_{\mathbf{H}-\mathbf{K}})}{\sum_{\mathbf{K}}|E_{\mathbf{K}}E_{\mathbf{H}-\mathbf{K}}|\cos(\phi_{\mathbf{K}}+\phi_{\mathbf{H}-\mathbf{K}})}, \qquad (4)$$

(Karle & Hauptman, 1956) provides a simple, but highly effective means for extracting phase values from the triplets. If several pairs of phases, $\phi_{\mathbf{K}}$ and $\phi_{\mathbf{H}-\mathbf{K}}$, and their associated $|E_{\mathbf{K}}|$, $|E_{\mathbf{H}-\mathbf{K}}|$ are known, equation (4) can be used to determine the most probable value for $\phi_{\mathbf{H}}$. Phase expansion and/or refinement in reciprocal space is accomplished through successive applications of this relationship. The tangent formula, in either its original or a weighted form, is at the heart of widely-used conventional multisolution phasing programs, including MULTAN (Main *et al.*, 1980) and SHELXS (Sheldrick, 1984), which refine multiple sets of trial phases by making many iterations or passes through the phase list.

Minimal Function. Minimization of an objective function like the minimal function,

$$R(\Phi) = \sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}}\left[\cos T_{\mathbf{HK}} - \frac{I_1(\mathrm{A}_{\mathbf{HK}})}{I_0(\mathrm{A}_{\mathbf{HK}})}\right]^2 \Bigg/ \sum_{\mathbf{H},\mathbf{K}}\mathrm{A}_{\mathbf{HK}} \qquad (5)$$

(Debaerdemaeker & Woolfson, 1983; Hauptman, 1991; DeTitta *et al.*, 1994) provides an alternative approach to phase refinement. $R(\phi)$ is a measure of the mean-square difference between the values of the triplets calculated using a particular set of phases and their expected values as given by the ratio of modified Bessel functions, and it is expected to have a constrained global minimum when the phases are equal to their correct values for some choice of origin and enantiomorph (the minimal principle). Equation (5) can also be written to include contributions from higher-order (quartet) invariants, but this option has not been shown, within the context of *SnB*, to be computationally efficient. Experimentation has thus far confirmed that, when the minimal function is used actively in the phasing process and solutions are produced, the final trial structure corresponding to the smallest value of $R(\phi)$ is a solution. Therefore, $R(\phi)$ is also an extremely useful figure-of-merit.

Parameter Shift. Parameter shift (Bhuiya & Stanley, 1963) is a seemingly simple search technique that has proven to be quite powerful as an optimization method when used to reduce the value of the minimal function, provided that appropriate choices of parameter values are made. The phases are considered in decreasing order with respect to the values of the associated |E|'s. When considering a given phase ϕ_i, the value of the minimal function (Eq. (5)) is initially evaluated three times. First, with the given set of phase assignments,

second with phase ϕ_i modified by the addition of the predetermined phase shift, and third with ϕ_i modified by the subtraction of the predetermined phase shift. If the first evaluation yields the minimum of these three values of the minimal function, then consideration of ϕ_i is complete, and parameter shift proceeds to ϕ_{i+1}. Otherwise, the direction of search is determined by the modification that yields the minimum value, and the phase is updated to reflect that modification. In this case, phase ϕ_i continues to be updated by the predetermined phase shift in the direction just determined so long as the value of the minimal function is reduced, though there is a user-defined predetermined maximum number of times that the shift is attempted. Based on extensive experimentation with these and related parameters, involving a variety of structures in several space groups, it has been determined that in terms of running time and percentage of trial structures that produce a solution, an excellent choice of parameters consists of the following: (i) perform a small number of passes through the phase set, (ii) evaluate the phases in order by decreasing |E|-values, and (iii) for each phase, perform a maximum of two 90° phase shifts. When the parameter-shift phase refinement is applied in centrosymmetric space groups, only a single shift of 180° is required for each phase. Surprisingly, higher success rates have been obtained if restricted phases in acentric space groups are treated as general phases (Weeks *et al.*, 1994).

Real-Space Constraints. Automatic real-space electron-density map interpretation in the *Shake-and-Bake* procedure consists of selecting an appropriate number of the largest peaks (typically equal to or less than the expected number of atoms) to be used as an updated trial structure without regard to chemical constraints other than a minimum allowed distance between atoms. If markedly unequal atoms are present, appropriate numbers of peaks (atoms) can be weighted by the proper atomic numbers during transformation back to reciprocal space. Thus, *a priori* knowledge concerning the chemical composition of the crystal is utilized, but no knowledge of constitution is required or used during peak selection. It is useful to think of peak picking in this context as simply an extreme form of density modification appropriate when atomic-resolution data are available. The entire dual-space refinement procedure is repeated for an appropriate number of cycles which have been determined empirically by experimentation with known datasets (Weeks *et al.*, 1994).

2 Methods

The *Shake-and-Bake* procedure has been implemented in an efficient and easy-to-use program, *SnB* (Miller *et al*, 1994). Pertinent information concerning *SnB* including the complete *User's Manual* may be accessed from the home page on the World Wide Web at URL:http://www.hwi.buffalo.edu/SnB. Stand-alone UNIX executables for SGI, SUN, IBM, and DEC alpha workstations as well as PC/Linux versions may be downloaded without cost to academic users. *SnB* has also been incorporated into Molecular Structure Corporation's teXsan package of crystallographic programs, and supercomputer versions have been installed on the Cray T3D and Cray C90 at the Pittsburgh Supercomputing Center, the CM-5 at NCSA, and the SP2 at the Cornell Theory Center.

Overview of the *SnB* Program. The main menu of *SnB* gives the user the options of (i) generating and processing trial structures to determine a structure, (ii) producing a histogram of $R(\phi)$ values for completed trial structures of a previously submitted structure-determination process, and (iii) displaying the best current trial structure (*i.e.*, lowest $R(\phi)$). A typical application of *SnB* consists of submitting a structure-determination process, monitoring the progress of the trial structures by occasionally viewing a histogram of final minimal-function values and, when a potential solution is identified, examining the geometry of this structure. The running time of the structure-determination procedure for large, difficult structures requiring many trials is substantial, and the ability to follow conveniently the course of such jobs is essential.

The flow chart presented in Figure 1 illustrates the basic operation of the *Shake-and-Bake* process as implemented in *SnB*. Triplet and (optionally) negative-quartet structure invariants,

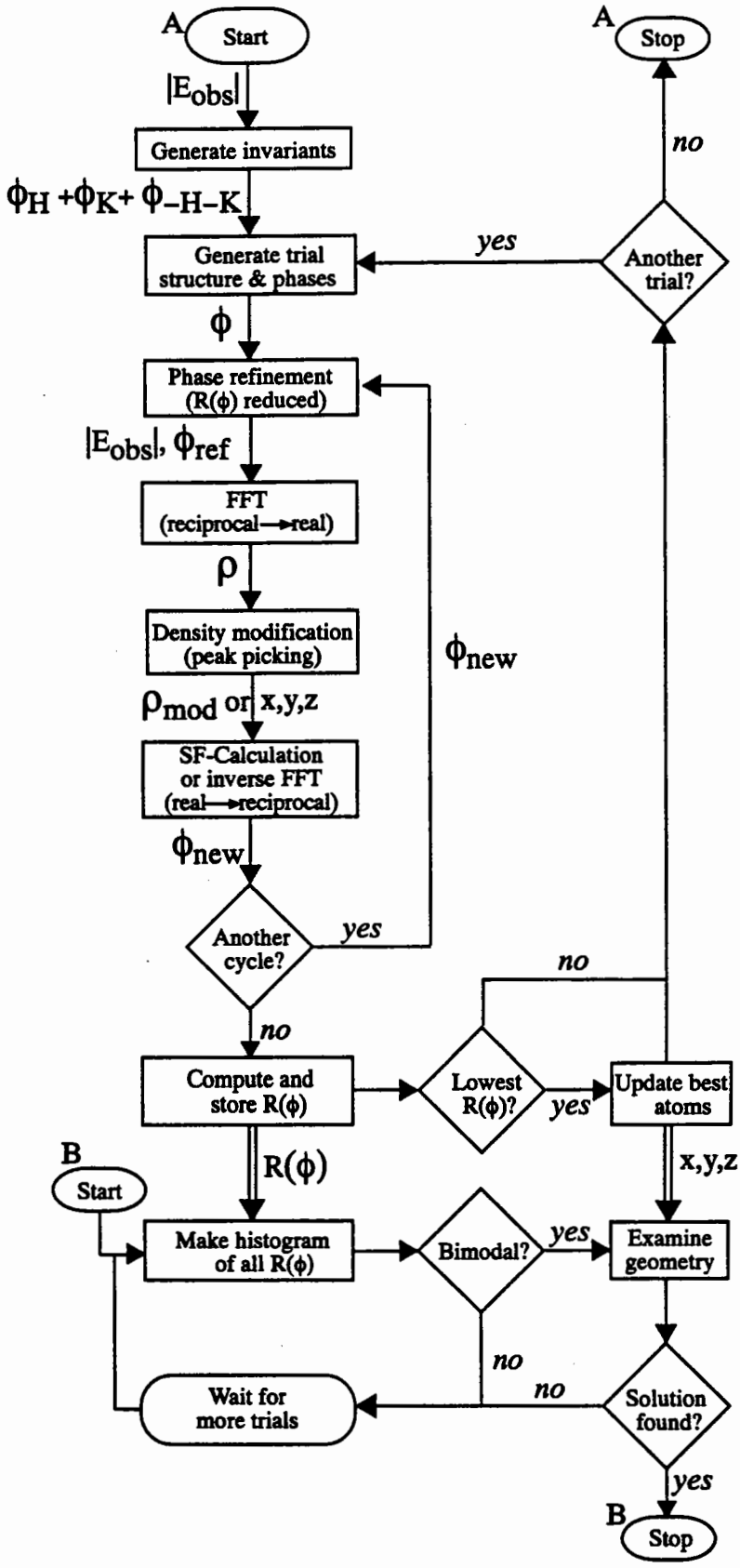


Figure 1. A flow chart for the *Shake-and-Bake* algorithm. Solid lines represent flow of control; double lines show movement of data. 'Start A' represents the beginning of a structure-determination process, and 'Start B' indicates the beginning of a session in which the R(ϕ) histogram and molecular geometry are checked.

as well as the initial coordinates for the trial structures, are generated. Once this information is available, every trial structure is subjected to the following *Shake-and-Bake* procedure. Initially, a structure-factor calculation is performed which yields phases corresponding to the trial structure. The associated value of the minimal function, R(ϕ), is then computed. At this point, the cyclical *Shake-and-Bake* phasing procedure is initiated, as follows. The phases are refined *via* the tangent formula or by parameter shift so as to reduce the value of R(ϕ). These phases are then passed to a Fourier routine which produces an electron-density map, but no graphical output is produced. Instead, the map is examined by a peak-picking routine which typically finds the *n* largest peaks (where *n* is the number of independent non-H atoms in the asymmetric unit) subject to the constraint that no two peaks are closer than a specified distance. These peaks are then considered to be atoms, and the process of structure-factor calculation, phase refinement, and density modification *via* peak selection is repeated for the predetermined number of *Shake-and-Bake* cycles.

For each completed trial structure, the final value of the minimal function is stored in a file, and the histogram routine can be run to determine whether or not a solution appears to be present in the set of completed trial structures. A bimodal distribution with significant separation is a typical indication that solutions are present (as shown in Figure 2), while a unimodal, bell-shaped distribution (*e.g.*, Figure 2 with the '0.467 to 0.470' row omitted) typically indicates a set of nonsolutions. Two options permit the user to view the current best structure. The first requires only a character-based terminal and produces a text plot suitable for printing on a line printer. The user can then manually 'connect the dots.' This routine also produces a list of the interpeak distances and angles. The second option makes use of GeomView, a graphical routine developed by the Geometry Center (Center for the Computation and Visualization of Geometric Structures at the University of Minnesota) and suitable for an X-Windows environment. These options are included to assist the user in deciding whether a solution has, in fact, been obtained. They are not intended to provide complete visualization, especially for larger structures. It is expected that the coordinates will be input into other graphical programs for more extensive display.

***SnB* Parameters.** The *SnB* user must supply (i) basic crystal data including space group, cell constants, and the contents of the asymmetric unit and (ii) an input reflection file consisting of *h, k, l* and the normalized structure-factor magnitudes, |E|. The program will automatically sort this data into descending order by |E|, eliminate systematic absences, and eliminate duplicate reflections. No selection based on σ(F) or F/σ(F) is performed. It is often critical that |E| values be calculated extremely carefully, and Blessing's programs (Blessing *et al.*, 1996) are recommended. Cost-effective default values for the control parameters (displayed following each query) are based on experience with several known test structures and are summarized in Table 1. Several parameters depend on structure size and can be expressed as a function of *n*. The user is free to override these recommendations, if desired.

A few comments concerning the parameters affecting the trial structures are in order. In practice, it is not necessary to use more than 100 randomly positioned atoms as an initial trial structure. During later cycles, choosing *n* peaks to recycle through the procedure gives optimum success rates for smaller structures. However, for large structures containing a significant number of atoms with low occupancy or high thermal motion, trial structures composed of less than *n* peaks (*e.g.*, 0.8**n*) give better performance. The geometry of trials that are solutions can be improved by E-Fourier recycling (Sheldrick, 1985), and the user can select the number of such Fourier refinement cycles (*i.e.*, *SnB* cycles with no phase refinement) and the number of peaks. Also, it is often useful to build, over the course of several cycles, from the number of peaks used during the *Shake-and-Bake* stage to the approximate total number of atoms expected in the structure. When atoms with atomic numbers greater than 10 are present, the user has the option of weighting the appropriate number of largest peaks in the structure-factor calculations. Unequal weighting has resulted in accelerated convergence to solution in cases where a small number of sulfur, iron, or chlorine atoms is present.

Table 1. Default parameter values for the *SnB* structure-determination procedure.

Parameter	Default
Non-H atoms in ASU	*n*
Invariant generation:	
Number of phases	10*n*
Number of triples	100*n*
Number of negative quartets	0
Starting atoms/random trial	min (*n*,100)
Number of *SnB* cycles:	
Parameter shift (PS)	*n*/2
refinement or Tangent formula refinement	*n*/4
PS phase refinement:	
Size of phase shift	90°
Max. number of shifts	2
Number of iterations	1
Exploit restricted phases?	No
Number of peaks to select	[0.8*n*,*n*]
Exploit heavy atoms?	Yes
Number E-Fourier steps	0

Figure 2. A 20-bucket histogram of the final minimal function values after 255 cycles for the 624-atom Tox II structure. 5000 phases and 50,000 triplet invariants were used. The separation between the single solution and the 1618 non-solutions is clearly shown.

R(Φ) Range	Trials in Range
0.467-0.470	1*
0.471-0.474	0
0.475-0.478	0
0.479-0.482	0
0.483-0.486	0
0.487-0.490	0
0.491-0.494	0
0.495-0.498	0
0.499-0.502	0
0.503-0.506	0
0.507-0.510	25*
0.511-0.514	135***
0.515-0.518	386**********
0.519-0.522	639**********
0.523-0.526	390**********
0.527-0.530	41*
0.531-0.534	2*
0.535-0.538	0
0.539-0.542	0

The relative efficiency of tangent-formula and parameter-shift phase refinement in *Shake-and-Bake* has been compared using known atomic-resolution datasets (Weeks *et al.*, 1997). In the case of tangent refinement the minimal function is also computed, but used only as a figure-of-merit. Regardless of which refinement method is used, optimization proceeds most rapidly when there is immediate feedback of each refined phase value. In general, the tangent formula solves small structures (<100 atoms) more cost-effectively, but the two phase-refinement methods are equally efficient for solving most of the tested structures with more than 100 independent atoms, including crambin. However, only parameter shift has produced recognizable solutions for gramicidin A although another figure-of-merit might be more reliable for tangent refinement. In addition, tangent-formula cost-effectiveness is highly dependent on the number of phase-refinement iterations (*i.e.*, the number of passes through the list of phases) per complete *Shake-and-Bake* cycle whereas parameter shift does not exhibit such strong dependency. The number of iterations per cycle must be chosen judiciously if high efficiency is, in fact, to be achieved. This is especially true for structures in space group P1 where it is never advisable to perform more than one iteration of tangent refinement per cycle. Given a fixed number of machine cycles, it is important to consider the trade-off between the number of trial structures processed and the number of cycles processed per trial structure. Experimentation has shown that, with a phase-refinement technique consisting of a single-iteration, two-step parameter shift of 90°, the point of diminishing returns is at approximately *n*/2 cycles. Therefore, the program defaults the number of cycles per trial to approximately this value. Overall recommendations for phase-refinement are given in Table 2.

After the dialogue is complete, the user is asked to review the information supplied and make any necessary changes. This information is then stored for use at a later time and for

Table 2. Phase-Refinement Recommendations.

Recommendation	Method	Cycles	*SnB* Iterations/Cycle)
$n < 100$ atoms	Tangent Formula	$n/4$	4 or 1 (P1)
$n > 100$ atoms	Parameter Shift	$n/2$	1
Always Safe	Parameter Shift	n	1

use by the histogram routine. Once a user decides that the parameters are satisfactory, the program automatically initiates the structure-determination procedure by spawning a batch job.

3 Applications

A list of successful *SnB* applications to protein structures is given in Table 3. Gramicidin A, crambin, and rubredoxin were previously known test structures re-solved at the Hauptman-Woodward Institute. The 64-residue scorpion toxin (Tox II) had been previously solved, but the number of residues and the amino acid sequence were deliberately withheld from the Buffalo group. The only information supplied was that the protein was composed of approximately 500 atoms and contained four disulfide bonds. The remaining structures (vancomycin, Er-1 pheromone, and alpha-1 peptide) were previously unknown, and these applications were made in other laboratories without direct involvement by the authors of *SnB*. All were solved routinely and automatically using essentially default parameters. Success rate (percentage of trial structures going to solution) depends on size and complexity of the structure, resolution and quality of data, the presence of heavier atoms (*e.g.*, S, Cl, Fe), and the space group as well as the number of refinement cycles. Success rate typically decreases as the size of the structure increases or the resolution or data quality decreases. Success rates for structures in P1 are significantly higher than for other space groups, a result which may be related to the fact that the origin position can be chosen arbitrarily in P1.

The application to Tox II was made on a network of SGI R4000 Indigo Workstations with *SnB* running as a background job for approximately six weeks. One morning, the histogram reproduced in Figure 2 was found during the daily progress check. After detecting that the histogram was now bimodal, the single trial in the 0.467 to 0.470 range was examined, and a conservative model consisting of five fragments and a total of 241 atoms was constructed. Following multiple cycles of Xplor refinement, the residual was 0.16 for 624 non-H atoms (Smith *et al.*, 1996).

It has been known for some time that conventional direct methods can be a valuable tool for locating the positions of heavy atoms using isomorphous ΔE's (Wilson, 1978) and anomalous scatterers using anomalous ΔE's (Mukherjee *et al.*, 1989). Thus, it is no surprise that the *Shake-and-Bake* algorithm can be fruitfully applied in this arena as well. The first application of this type was to native and Se-Met data for avian sarcoma virus integrase (Bujacz *et al.*, 1995). The four Se atoms were found using 189 ΔE values (>1.76) in the resolution range 20 to 3.7Å. The investigators report that the isomorphous difference Patterson map was impossible to deconvolute without the aid of direct methods.

4 Concluding Remarks

The ultimate potential of the *Shake-and-Bake* approach to the *ab initio* structure determination of macromolecules is unknown. The combination of this technique with increasingly more-powerful computers has recently permitted direct-method solutions in situations regarded as impossible only a few years ago. The combination of *Shake-and-Bake* methodology with alternative density-modification methods and supplemental phasing information from isomorphous replacement and single or multiple-wavelength anomalous dispersion may allow equally spectacular advances in the near future.

Table 3. Protein structures solved *ab initio* by *SnB*

Structure	Non-H Atoms/ASU	Space Group	Resolution	Success Rate	References
Vancomycin	255	$P4_32_12$	0.9Å	1/4200	P. Loll, pers. comm.
Gramicidin A	317	$P2_12_12_1$	0.86	0.25%	Hauptman, 1995
Er-1 Pheromone	328	C2	1.0	0.25%	Anderson *et al.*, 1996
Crambin	~400	$P2_1$	0.83	2-3%	Weeks *et al.*, 1995
Alpha-1 Peptide	471	P1	0.92	5%	Prive *et al.*, 1995
Rubredoxin	497	$P2_1$	1.0	2.7%	Hauptman, 1995
Tox II	624	$P2_12_12_1$	0.96	1/1619	Smith *et al.*, 1996

Acknowledgments. The *Shake-and-Bake* algorithm and the *SnB* program have been made possible by the financial support of grants GM-46733 from NIH and IRI-9412415 from NSF. The authors would like to acknowledge the guidance and inspiration provided by Prof. Herbert Hauptman throughout the development of *SnB*.

References

Anderson, D.H., Weiss, M.S., & Eisenberg, D. Acta Cryst, *D52*, (1996) 469.
Baggio, R., Woolfson, M.M., Declercq, J.-P., & Germain, G. Acta Cryst, *A34* (1978) 883.
Bhuiya, A.K. and Stanley, E. Acta Cryst, *16* (1963) 981.
Blessing, R.H., Guo, D.Y., & Langs, D.A. Acta Cryst, *D52* (1996) 257.
Bujacz, G., Jaskolski, M., Alexandratos, J., Wlodawer, A., Merkel, G., Katz, R.A., & Skalka, A.M. Jour. Mol. Biol, *253* (1995) 333.
Debaerdemaeker, T. and Woolfson, M.M. Acta Cryst, *A39* (1983) 193.
DeTitta, G.T., Weeks, C.M., Thuman, P., Miller, R., & Hauptman, H.A. Acta Cryst, *A50* (1994) 203.
Germain, G. and Woolfson, M.M. Acta Cryst, *B24* (1968) 91.
Hauptman, H.A. "A Minimal Principle in the Phase Problem", in *Crystallographic Computing 5: From Chemistry to Biology*, D. Moras, A.D. Podnarny & J.C. Thierry (Eds.), IUCr Oxford Univ. Press, 1991, pp. 324-332.
Hauptman, H.A., Acta Cryst, *B51* (1995) 416.
Hauptman, H.A. and Karle, J. *Solution of the Phase Problem. I. The Centrosymmetric Crystal*, ACA Monograph No. 3, Dayton, OH:Polycrystal Book Service (1953).
Karle, J. Acta Cryst, *B24* (1968) 182.
Karle, J. and Hauptman, H. Acta Cryst, *9* (1956) 635.
Karle, J. and Karle, I.L. Acta Cryst, *21* (1966) 849.
Miller, R., Gallo, S.M., Khalak, H.G., & Weeks, C.M. J. Appl. Cryst, *27*, (1994) 613.
Mukherjee, A. K., Helliwell, J. R., & Main, P. Acta. Cryst, *A45*, (1989) 715-718.
Podjarny, A.D., Bhat, T.N., & Zwick, M. Ann. Rev. Biophys. Biophys. Chem, *16* (1987) 351.
Prive, G., Ogihara, N., Wesson, L., Cascio, D., & Eisenberg, D. Abstr. W008, Proc. Am. Crystallogr. Assoc. Meeting, Montreal (1995).
Sheldrick, G.M. "SHELX-84", in *Crystallographic Computing 3: Data Collection, Structure Determination, Proteins, and Databases*, G M. Sheldrick, C. Kruger & R. Goddard (Eds.), Clarendon Press, Oxford, 1985, pp. 184-189.
Smith, G. D., Blessing, R. H., Ealick, S. E., Fontecilla-Camps, J. C., Hauptman, H. A., Housset, D., Langs, D. A., & Miller, R. Abstr. E1146, IUCr Meeting, Seattle (1996).
Weeks, C.M., DeTitta, G.T., Hauptman, H.A., Thuman, P., & Miller, R., Acta Cryst, *A50* (1994) 210.
Weeks, C.M., Hauptman, H.A., Chang, C.-S., & Miller, R. "Structure Determination by Shake-and-Bake with Tangent Refinement", *ACA Transactions Symposium 30* (1997).
Weeks, C.M., Hauptman, H.A., Smith, G.D., Blessing, R.H., Teeter, M.M., & Miller, R. Acta Cryst, *D51* (1995) 33.
Wilson, K. S. Acta Cryst, *B34*, (1978) 1599.

Direct Methods based on real / reciprocal space iteration

By George M. Sheldrick

Institut für Anorganische Chemie, D37077 Göttingen, Germany

Abstract

It appears that direct methods inspired by *Shake & Bake* involving iteration between real and reciprocal space are able to solve structures with several hundred independent atoms, but still require data to *atomic resolution* (say 1.2Å). Applications to the *ab initio* phasing of proteins (given very high resolution data) and to the location of anomalous scatterers from lower resolution ΔF or MAD F_A data are discussed.

1. Introduction

A feature that probably contributed significantly to the rapid acceptance of the conventional direct methods program SHELXS-86 was the E-Fourier recycling (Sheldrick, 1982, 1990) shown in Fig. 1 that was used to complete the structure obtained from direct methods.

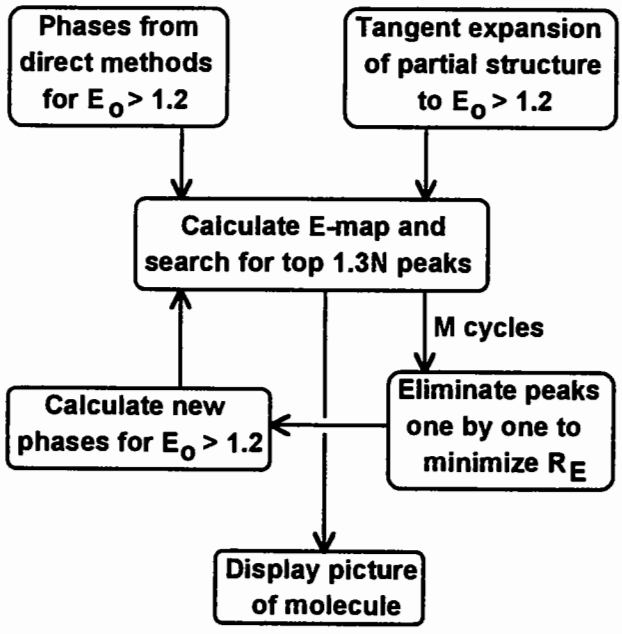


Fig. 1. E-Fourier recycling as used in SHELXS-86 to improve phases from direct methods.

Usually a couple of cycles were sufficient. Since the E-Fourier recycling was only applied to the 'best' solution, and only E-values greater than (say) 1.2 were employed, the computing requirements were modest. Very often this procedure was able to find every atom (except perhaps disordered solvent molecules), which users found very convenient. On a few occasions the E-Fourier recycling succeeded in extracting the solution from a rather dubious set of direct methods phases, but despite this strong hint, it did not occur to me that it could itself be effective as a 'direct method'. This required the development of the *Shake & Bake* philosophy by Weeks, Miller & Hauptman at Buffalo (Miller, DeTitta, Jones, Langs, Weeks & Hauptman, 1993; Miller, Gallo, Khalak & Weeks, 1994), which inspired much of the work reported here.

2. Peaklist optimisation

Fine tuning of the E-Fourier recycling method since SHELXS-86 was distributed included Sigma-A weighted difference Fourier maps (Read, 1986) and the use of the correlation coefficient (Fujinaga & Read, 1987) between E_c^2 and E_o^2 to decide which atoms to delete:

$$CC = [\, \Sigma w E_o^2 E_c^2 \cdot \Sigma w - \Sigma w E_o^2 \cdot \Sigma w E_c^2 \,] \; / \; \{\, [\, \Sigma w E_o^4 \cdot \Sigma w - (\Sigma w E_o^2)^2] \cdot [\Sigma w E_c^4 \cdot \Sigma w - (\Sigma w E_c^2)^2 \,] \,\}^{1/2}$$

The correlation coefficient is more sensitive in the important early stages, and appears to give a very good indication of the true phase error (e.g. Fig. 2).

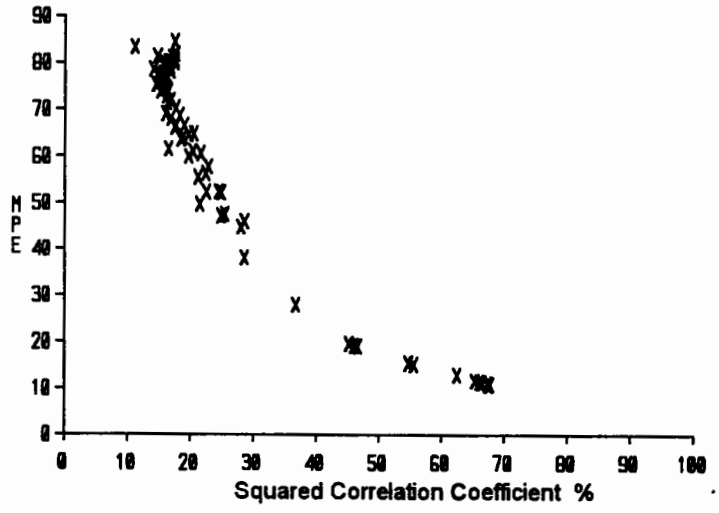


Fig. 2. E-weighted mean phase error (MPE) from direct methods attempts as a function of the square of the correlation coefficient between E_c^2 and E_o^2 for crambin (0.92Å data kindly provided by Håkon Hope).

Tests on rubredoxin by Sheldrick & Gould (1995) showed that the elimination of atoms to improve the correlation coefficient (*peaklist optimisation*) was very effective at expanding the structure from the iron and four sulfur atoms to all ca. 400 atoms, provided that the resolution was better than 1.3Å.

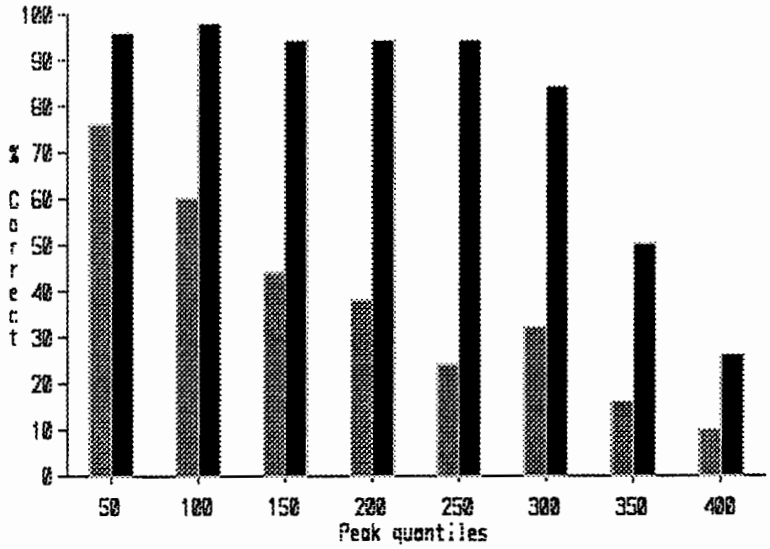


Fig. 3. Histogram showing the percentage of correct peaks (within 0.3Å of the true atomic positions) before (grey) and after (black) iterative *peaklist optimisation* for rubredoxin with the data truncated to 1.2Å. There are 50 peaks in each quantile, so the first pair of columns are for peaks 1-50 sorted on peak height, the second pair for peaks 51-100, etc.

3. Application to the solution of an unknown protein

Provided that a few heavier atoms can be located by for example Patterson interpretation, *peaklist optimisation* can be used to complete the structure and so - in the exceptionally favourable case of a small macromolecule that diffracts to atomic resolution and contains a few heavier atoms - provides a method of *ab initio* structure solution. Frazão *et al.* (1995) were able to solve the structure of an unknown cytochrome c_6 in this way. The best sequence identity with a protein of known structure was only about 24%, so molecular replacement would have been difficult, but undoubtedly the structure could also have been solved - albeit at a higher cost in synchrotron beam-time - by MAD phasing. The iron and three sulfurs were located by automated Patterson interpretation (Sheldrick *et al.*, 1993) and the full structure was expanded from them by *peaklist optimisation* using synchrotron data collected to 1.1Å (although the 1.2 to 1.1Å shell was extremely weak) at the EMBL outstation in Hamburg. Fig. 4 shows the same region of the structure at different stages of the structure determination.

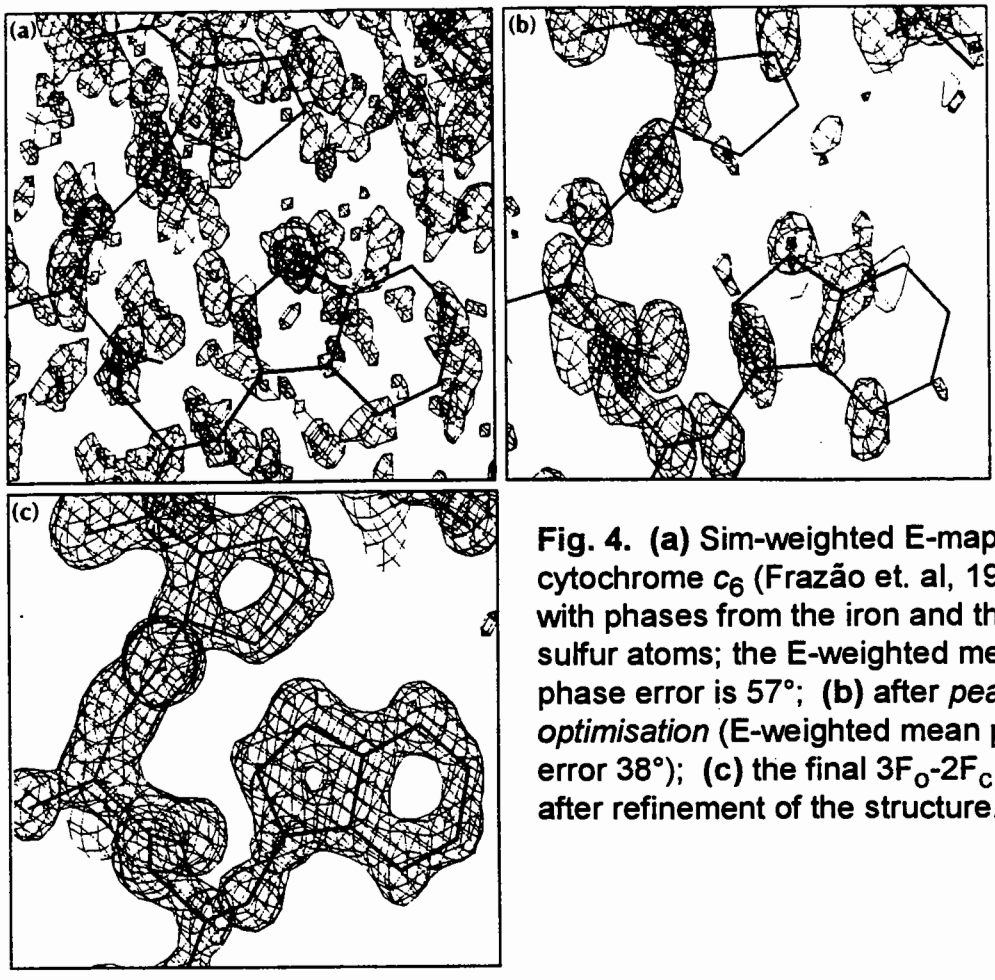


Fig. 4. (a) Sim-weighted E-map for a cytochrome c_6 (Frazão et. al, 1995) with phases from the iron and three sulfur atoms; the E-weighted mean phase error is 57°; **(b)** after *peaklist optimisation* (E-weighted mean phase error 38°); **(c)** the final $3F_o$-$2F_c$ map after refinement of the structure.

Iterative application of *peaklist optimisation* enables about 90% of the protein atoms to be identified from the peaklist alone without the need to examine any maps; this was however required to find the remaining atoms, which had high thermal displacement parameters or were disordered.

4. A real/reciprocal space recycling method for structure determination

Inspired by the *Shake & Bake* Philosophy described in the preceding lecture, Sheldrick & Gould (1995) turned the *peaklist optimisation* procedure into a full *ab initio* method for structure determination by the addition of the tangent formula in the reciprocal space stage. Their algorithm (Fig. 5) could start from random phases for a number of trials, or the initial phases could be generated by (a) a rotation search (to maximise $\Sigma E_c^2(E_o^2-1)$ for the largest E-values) for a known small fragment (a small piece of α-helix proved very effective) or (b) threefold Patterson superposition from vector triangles identified in the sharpened Patterson peaklist (to exploit the presence of heavier atoms such as sulfur or phosphorus). Since these two methods of generating slightly better than random starting phases are not able to position the origin of the space group, all calculations were performed on data expanded to the effective space group P1. Expansion to P1 may in any case increase the chances of this approach converging to the correct solution, but increases the computer time required.

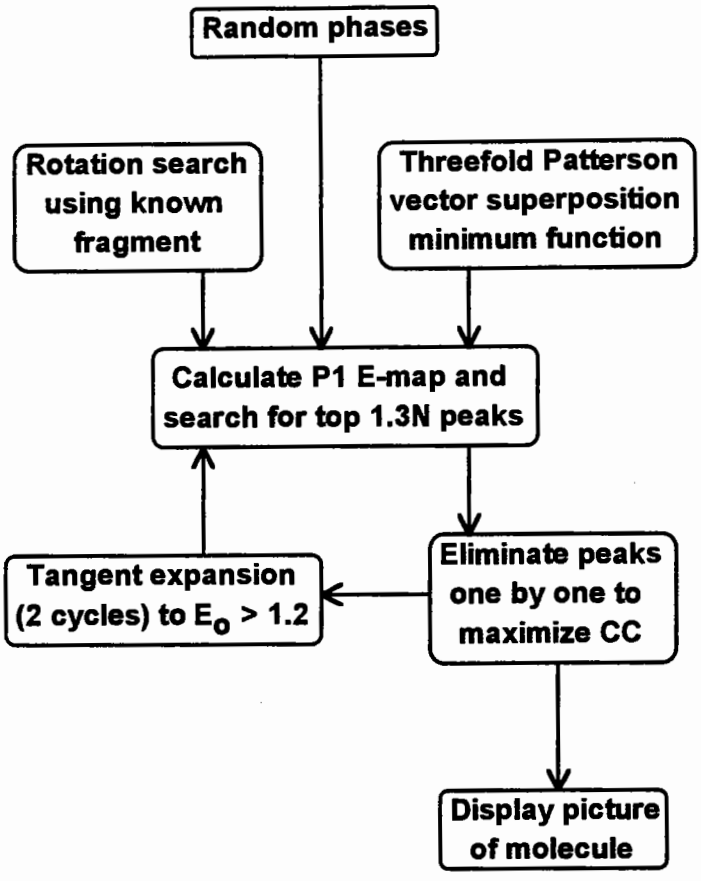


Fig. 5. Real/reciprocal space recycling starting from random or almost random phases as proposed by Sheldrick & Gould (1995). The correlation coefficient CC was calculated for all data expanded to the space group P1.

Tests showed that the *peaklist optimisation* was much more effective than simply accepting the top N peaks, but that it takes about the same CPU time as three structure factor calculations, and so is slower. Starting with slightly better than random phases from the rotation search or Patterson superposition map considerably increased the success rate of this approach. The method was successful in solving several structures with more than 200 atoms in the asymmetric unit, but proved very expensive in consumption of computer resources. The computer time required could be reduced considerably by calculating the correlation coefficient for only the largest E-values, for which structure factors were required anyway to provide initial phases for the tangent refinement. However the correlation coefficient proved much less effective when not applied to the full range of E-values. The solution was to divide the procedure into an *internal loop*, in which a specified number of peaks were eliminated so that $\Sigma E_c^2(E_o^2-1)$ remained as large as possible, alternating with tangent phase refinement, and an *external loop*, applied only for solutions with good values of CC (for all data), in which *peaklist optimisation* as described above was applied using all data so that the final structure was as complete as possible. The new procedure (which has somehow acquired the name *half-baked*) is illustrated in Fig. 6.

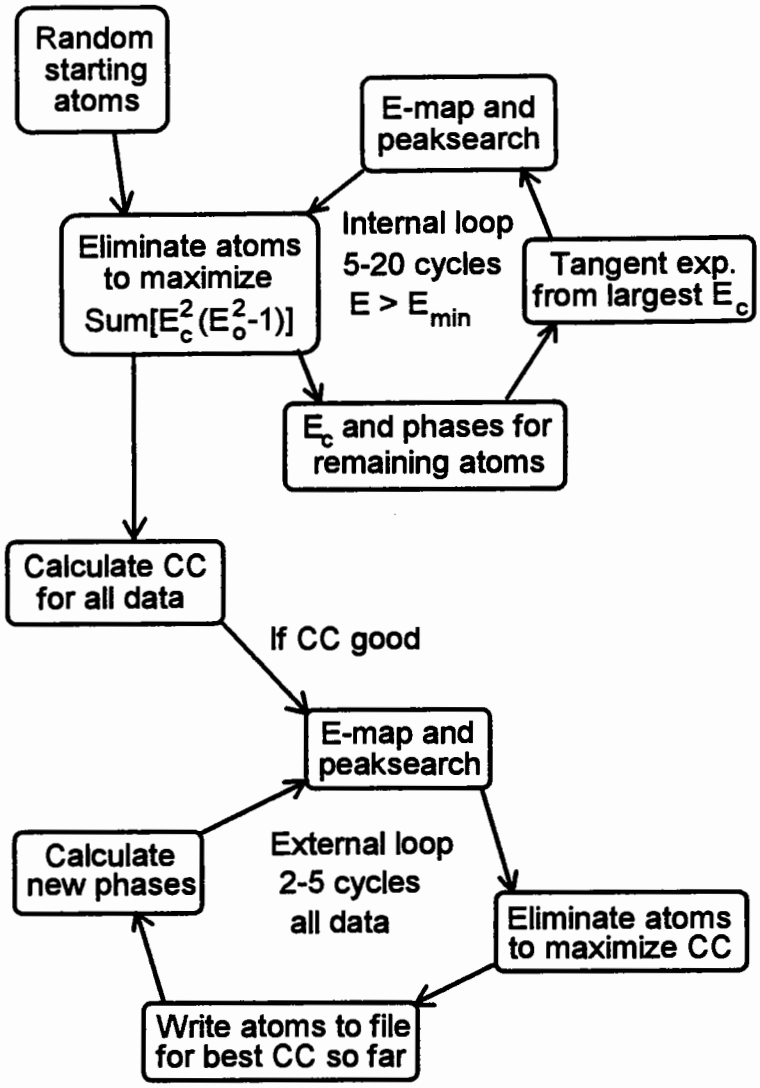


Fig. 6. The *half-baked* approach, as incorporated in SHELXD-97. It is repeated indefinitely, restarting from random atoms, until interrupted! It may be performed either applying the symmetry of the space group or after expanding all data to the effective space group P1 (in which case the starting phases may be generated by a rotation search). Usually not more than two tangent cycles are applied per internal loop cycle. E_{min} is normally chosen to be in the range 1.2 to 1.6.

In general, it appears to be computationally more efficient to expand the data to an effective space group of P1 for monoclinic structures; a larger percentage of trials lead to a solution, more than compensating for the increased cycle time. For higher symmetry it may be better to impose the full space group symmetry. It should also be possible to include twinning in the external loop; sometimes it is easier to guess the twin law than the space group, in which case the data could be expanded to P1.

The procedure described above is philosophically similar to *Shake & Bake,* but relative to *Shake & Bake* it does more of the work in real than in

reciprocal space. It appears to be roughly comparable in its ability to solve difficult structures. One structure solved at about the same time by both programs, but using two different synchrotron data-sets, is vancomycin, a glycopeptide antibiotic of crucial medical importance in the struggle against the evolution of antibiotic resistant bacteria. The unexpurgated solution obtained by Schäfer, Schneider & Sheldrick (1996) using the *half-baked* procedure is shown in Fig. 7. The data were 99.3% complete to 1.09Å, the edge of the image plate used for synchrotron data collection. Including solvent there are 313 atoms in the asymmetric unit in $P4_32_12$. 2000 trials with 8 cycles in the internal loop gave one solution; the CC of 75.5% was well separated from the rest (the next largest CC was 57.9%, for an incorrect solution). The CPU time used corresponded to a mere 4 VAX-years.

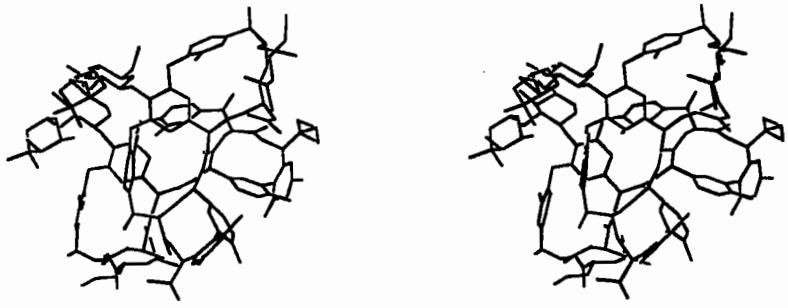


Fig. 7. Stereo view of the unedited *ab initio* solution of the structure of vancomycin. The two antibiotic molecules are almost complete, and form a tight unsymmetrical dimer.

6. The location of anomalous scatterers

In the implementation of the MAD method, a problem has arisen involving the location of the anomalous scatterers from anomalous ΔF or MAD F_A values. Patterson methods work well if there are only a few such atoms, but the complexity increases as the square of the number of atoms and becomes prohibitive, even for automated computer interpretation (Sheldrick *et al.*, 1993) when the number of anomalous scatterers is more than about 12. One would have expected that classical direct methods should be able to solve this problem, since they are capable of finding at least 100 equal atoms, and the anomalous scatterers are usually separated from one another by distances much greater than the limiting resolution of the reflection data, but in practice they invariably fail to locate say 20 independent selenium atoms. There seem to be several possible reasons for this unexpected problem.

(a) Both Patterson and direct methods work best with *complete* data. Missing centric and other reflections cause problems.

(b) The ΔF values represent lower limits on F_H (MAD F_A values should be better, at least in theory), so small ΔF values cannot be used in probability formulae such as those involving negative quartets.

(c) It is difficult to take $\sigma(\Delta F)$ or $\sigma(F_A)$ into account in conventional direct and Patterson methods, so the signal may get lost in the noise.

(d) The selenomethionines may be conformationally disordered.

Table 1. Crossword table for the second best solution from the Cu-Kα anomalous ΔF values for a HiPIP protein with two Fe_4S_4 clusters in the asymmetric unit (1.5Å data kindly donated by Hazel Holden & Gary Wesenberg, truncated to 2Å to make the test more difficult). The upper row gives the minimum distance between the atom defining the row and the atom defining the column, the lower row gives the corresponding Patterson superposition minimum function.

```
Try 89, CC(HA)=35.74%, PATFOM=39.67

Peak   x     y      z     self  cross-vectors
99.9 0.389 0.736 0.176 29.2
                       41.0

98.4 0.432 0.746 0.249 30.1  2.6
                       51.0 66.6

90.7 0.399 0.696 0.194 29.4  2.2  3.3
                        0.0 47.5 33.0

89.9 0.914 0.187 0.126 27.9 14.0 16.6 14.4
                       53.2 34.6 49.1 74.7

88.1 0.354 0.742 0.255 31.4  2.6  2.9  3.4 14.6
                       45.7 69.3 73.2 56.5 57.4

82.3 0.960 0.160 0.043 26.6 14.6 17.0 14.8  3.2 14.7
                       67.6 42.5 37.9 54.7 27.5 37.8

71.1 0.901 0.125 0.082 27.7 14.0 16.5 13.8  3.5 14.5  3.0
                       22.2 27.9 32.6 34.9 25.3 32.8 47.8

67.4 0.973 0.342 0.132 27.4 16.6 18.8 18.0  8.4 16.8  9.9 11.8
                       41.7  0.9 49.0 20.0  0.5 31.5  0.0  0.0

-------------------------------------------------------------------

46.8 0.966 0.143 0.145 27.6 16.4 18.9 16.5  3.1 16.8  3.0  3.1 10.4
                       38.3 34.2 43.8 19.7 22.5 25.3 26.7 45.8  0.0

41.3 0.500 0.749 0.286 28.8  5.1  2.7  5.3 19.1  5.4 19.6 18.9 21.3
                        0.0  4.5 46.5  0.7 14.7 21.6  5.8  2.8 22.7
```

Direct methods based on real/reciprocal space recycling have some advantages to offer that may help to overcome these problems. The number of anomalous scatterers N_H is usually known precisely; this information can be used in a very direct way. The elimination of atoms in turn to optimise the correlation coefficient CC, until exactly N_H atoms remain, does not require complete data. In addition CC incorporates weights based on the experimental sigmas. Finally, the Patterson function can still be used as an independent check, as shown in Table 1. The second best solution is illustrated; the Patterson superposition minimum function values clearly show that the atoms 1-7 and 9 correspond to the eight expected iron atoms. They form two Fe_4 clusters with Fe•••Fe distances of about 3Å. The PATFOM figure of merit is simply the mean of the Patterson superposition minimum function values for the top N_H atoms. The solution with the best PATFOM, but the second best CC, gave atoms 1-8 as the correct iron atoms.

Table 2. Crambin test, internal loop searching for 3 disulfide bonds, external loop expanding to full structure. The 0.92Å low-temperature data were collected and provided by Håkon Hope.

```
1625 E-sig(E) > 1.500 used to generate 77607 unique TPR
Try 19,  CC(HA) = 19.03%,   PATFOM = 13.80

Peak    x      y      z     self cross-vectors
99.9 0.3019 0.1253 0.1020  19.2
                           15.6

96.7 0.2571 0.0783 0.1028  22.4  2.0
                           18.4 14.3

96.7 0.3914 0.1707 0.4511  13.0  8.6  9.6
                           18.2 13.8 18.4

93.7 0.4373 0.1292 0.4262  11.1  9.1 10.3  2.1
                           16.3 35.5 12.2 11.3

90.7 0.0794 0.2353 0.0483  11.5  9.4  7.9 15.5 16.9
                            0.9 17.0 15.6 10.4 12.9

85.7 0.1098 0.3147 0.0591  13.1  8.6  7.5 14.6 16.0  1.9
                           13.2 11.5 10.3 16.1  7.7  0.0

Peaklist optimization cycle 1  CC=30.05% for 41 atoms
Peaks: 99 97 97 93 92 88 15 15 -14 14 -13 -12 -12 -12 -12

Peaklist optimization cycle 2  CC=47.75% for 108 atoms
Peaks: 99 95 93 92 85 81 34 -34 34 34 34 -33 33 33 33 33

Peaklist optimization cycle 3  CC=70.62% for 240 atoms
Peaks: 99 95 91 90 80 77 36 36 36 35 35 35 34 34 34 33 33

Peaklist optimization cycle 4  CC=81.57% for 354 atoms
Peaks: 99 96 92 88 74 73 37 37 37 36 36 35 35 35 35 35 34
Fragments: 310 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

It is possible to combine the search for a specific number of heavier atoms from the native data in the internal loop with expansion to the full structure in the outer loop, as illustrated by the crambin test in Table 2. In this case the three disulfide bridges can be identified by their distances of about 2Å; only solutions containing three disulfide bonds were expanded further by the program. The external loop of *peaklist optimisation* leads to the essentially full structure in 4 cycles with a convincing CC (values greater than 70% are invariably correct). A minus sign in the list of peak heights indicates that that peak was rejected in the elimination procedure. The final line shows that there is a connected fragment of 310 atoms, plus a number of well-defined water molecules that do not bond to other atoms.

7. Conclusions and future prospects

In reciprocal space, the phase refinement algorithm enables the structure to escape from local minima in search of a global minimum, and shows some similarities to *simulated annealing*. This is achieved by the correlation between reflection phases and intensities, and the dominant role of reflections with large E-values. In real space, the powerful constraint of an atomic model (the key to the success of *Shake & Bake* ?) provides detail that may have been lost in the reciprocal space stage, and forces convergence towards a physically reasonable solution.

To extend the method to lower resolution, density modification may not prove sufficiently incisive as a replacement for peak-picking. More promising is the method used in ARP (Lamzin & Wilson, 1993) to fill density with atoms. Alternatively, instead of using individual atoms, typical groups of 3-5 atoms (e.g. peptide units) could be fitted to the density (given a fast computer !).

It looks as though real/reciprocal space recycling has the potential to overcome the current difficulties in the location of a large number of anomalous scatterers from noisy MAD data, but this needs further testing on real data.

I am grateful to the Fonds der Chemischen Industrie for support. Figs. 1, 2 and 5 are reproduced from Sheldrick & Gould (1995), Fig. 4 from Frazão *et al.* (1995) and Fig. 7 from Schäfer *et al.* (1996), with permission of the respective publishers.

References

Frazão, C., Soares, C.M., Carrondo, M.A., Pohl, E., Dauter, Z., Wilson, K.S., Hervás, M., Navarro, J.A., De la Rosa, M.A. & Sheldrick, G.M. (1995). *Structure* **3**, 1159-1169.

Fujinaga, M. & Read, R.J. (1987). *J. Appl. Cryst.* **20**, 517-521.

Lamzin, V.S. & Wilson. K.S. (1993). *Acta Cryst.* **D49**, 129-147.

Miller, R., DeTitta, G.T., Jones, R., Langs, D.A, Weeks, C.M. & Hauptman, H.A. (1993). *Science* **259**, 1430-1433.

Miller, R., Gallo, S.M., Khalak, H.G. & Weeks, C.M. (1994). *J. Appl. Cryst.* **27**, 613-621.

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

Schäfer, M., Schneider, T.R. & Sheldrick, G.M. (1996). *Structure* **4**, 1509-1515.

Sheldrick, G.M. (1982). In *Crystallographic Computing*, edited by D. Sayre, pp. 506-514. Oxford: Clarendon Press.

Sheldrick, G.M. (1990). *Acta Cryst.* **A46**, 467-473.

Sheldrick, G.M., Dauter, Z., Wilson, K.S., Hope, H. & Sieker, L.C. (1993). *Acta Cryst.*, **D49**, 18-23.

Sheldrick, G.M. & Gould, R.O. (1995). *Acta Cryst.* **B51**, 423-431.

Maximum-Entropy Methods and the Bayesian Programme

G. Bricogne
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH England and
LURE, Bâtiment 209D, F-91405 Orsay Cedex, France
gb10@mrc-lmb.cam.ac.uk
http://Lagrange.mrc-lmb.cam.ac.uk

0. Introduction.

The success of direct methods programs at providing a quasi-automatic solution to the phase problem for small molecules has over the years dimmed the perception of the basic inference processes involved in such crystal structure determinations. Greater awareness of this sequence of inference steps has persisted in the macromolecular field, where the dialogue between numerical computation and human decision is still part of the daily experience of most crystallographers. The final step of turning the determination of macromolecular crystal structures itself into a purely computational and automatic process is therefore likely to involve – and even require – that a common basis be found for all phase determination methods used in these two fields. The purpose of this article is to present an overview of one such unifying scheme, the *Bayesian programme* formulated some years ago by the writer [1,2,3], partial implementations of which have given several encouraging results [4-14] along the way to a full implementation. Special attention is paid here to those areas where this viewpoint is having a practical impact on real applications in macromolecular crystallography. Its application to *ab initio* phasing at typical macromolecular resolutions will require in addition the incorporation of stereochemical information into structure factor statistics [15].

1. Motivation.

Bayesian concepts and methods are ideally suited to the "management" of crystal structure determination from diffraction data [16]. Indeed, the latter is fundamentally a sequence of steps aimed at gradually reducing the ambiguity created by the loss of phase information. Each step involves the formulation of a range of hypotheses designed to "fill in" the missing information in one way or another, followed by the testing of these hypotheses against the available diffraction data and also against prior chemical knowledge – either as such or after it has been converted into statistical correlations between structure factors.

The work published in [1] was a first step towards this goal, within the restricted context of direct phase determination. Its purpose was to urge a return to the fundamental problem of calculating joint probability distributions of structure factors and to find methods better suited to the macromolecular field which would increase the accuracy and the sensitivity of probabilistic phase

indications. Shortcomings of conventional direct methods were identified and shown to be related to the use of *uniform* distributions of random atomic positions, and of the associated *Edgeworth series* as an approximation to the joint distribution of structure factors. They were overcome by using instead *maximum-entropy* distributions [17] for the unknown positions of random atoms, and the associated *saddlepoint approximation* to the joint probability distribution of structure factors.

The scope of the Bayesian analysis was then enlarged to include other crystallographic methods, particularly those used in the macromolecular field (isomorphous substitution, anomalous scattering, molecular replacement, non-crystallographic symmetry averaging and solvent flattening) whose conventional formulations all involve some form of statistical treatment when they have to deal with the representation of incomplete knowledge – e.g. non-isomorphism in heavy-atom derivatives, or missing atoms in a partial model. These statistical treatments are as a rule rather simplistic in comparison with those on which direct methods are based, and yet it is through the resulting "phase probability densities" that these methods pool their abilities to determine phases. It is therefore no exaggeration to say that macromolecular phasing techniques have so far communicated with each other through their weakest component. These shortcomings were addressed by extending the initial framework into a "multichannel formalism" [2] which made possible the effective construction of a wide range of flexible statistical models involving mixtures of randomly positioned scatterers distributed with varying degrees of non-uniformity. Such models were precisely the hitherto missing devices for optimally describing any kind of phase uncertainty, and it was proposed that the corresponding likelihood functions should be used as a universal tool for consulting data in all conventional macromolecular phasing and refinement methods.

Finally, numerous other classes of situations occur in macromolecular crystallography where the existence of ambiguities is inadequately handled, either by taking centroids of multimodal distributions (as in the Blow & Crick [18] treatment of strongly bimodal phase indications), or by trying to apply iterative map improvement techniques from a single choice of starting point – irrespective of how uncertain that starting point may be (as in solvent flattening from a single choice of molecular boundaries) – creating the risk of potentially disastrous biases. Here again the Bayesian view point leads to a much needed general mechanism for dealing appropriately with ambiguities, which was missing in conventional methods.

2. The Bayesian viewpoint.

The key concept in the Bayesian view of crystal structure determination is the notion of *missing information*, or *ambiguity*, in the current situation. Typical instances of missing information encountered in macromolecular crystallography include uncertain molecular boundaries, inconclusive rotation or translation searches, strong bimodality in SIR phase probability indications, and of course the lack of any phase indications whatsoever for some reflexions.

The techniques of Bayesian inference can then be brought into action *whenever an item of missing phase information can be shown to influence the expected joint probability distribution of structure factor amplitudes*. In every such case, some of the missing information can be retrieved from structure factor amplitudes by the following procedure:

(1) generating an ensemble of hypotheses $\{\mathcal{H}_J\}_{j\in\mathcal{S}}$ forming a "representative sample" $\mathcal{S}$ of all possibilities left open by the current state of ambiguity (to avoid any bias), and assigning to each of them a prior probability $P^{prior}(\mathcal{H}_J)$ from knowledge available outside the diffraction measurements;

(2) constructing the jpd $P(\mathbf{F} \mid \mathcal{H}_J)$ of the observable structure factors $\mathbf{F}$ conditional on each hypothesis $\mathcal{H}_J$; then integrating the phases out to get the marginal cpd of amplitudes $P(|\mathbf{F}| \mid \mathcal{H}_J)$;

(3) forming the likelihood of each $\mathcal{H}_J$ from the observed data as $\Lambda(\mathcal{H}_J \mid |\mathbf{F}|^{obs}) = P(|\mathbf{F}|^{obs} \mid \mathcal{H}_J)$, and using Bayes's theorem to obtain the posterior probability of each hypothesis as

$$P^{post}(\mathcal{H}_J \mid |\mathbf{F}|^{obs}) = \frac{P^{prior}(\mathcal{H}_J)\,\Lambda(\mathcal{H}_J \mid |\mathbf{F}|^{obs})}{\sum_K P^{prior}(\mathcal{H}_K)\,\Lambda(\mathcal{H}_K \mid |\mathbf{F}|^{obs})} \qquad (2.1).$$

The basic computational mechanism in Bayesian crystal structure determination is therefore :

$$(\mathcal{H}) \text{ -----> } P(\mathbf{F} \mid \mathcal{H}) \text{ -----> } P(|\mathbf{F}| \mid \mathcal{H}) \text{ -----> } \Lambda(\mathcal{H}) = P(|\mathbf{F}|^{obs} \mid \mathcal{H}) \qquad (2.2)$$

i.e. the conversion of a hypothesis $\mathcal{H}$ into a likelihood function (*via* a jpd of structure factors) for testing that hypothesis against the available structure factor amplitude data. It is this mechanism which was analysed in [2] for a class of hypotheses wide enough to accommodate all conventional phasing and refinement techniques.

3. Basic computational processes.

The techniques involved in implementing the scheme just described fall naturally into three main categories. The first is concerned with the design of efficient sampling strategies for the generation of diverse hypotheses and with the book-keeping of that diversity (§3.1). The second deals with the analytical and numerical aspects of deriving joint probability distributions of structure factors (§3.2), then conditional distribution of amplitudes and likelihood functions (§3.3); these methods are mathematically intensive and can only be outlined here. The third and last category addresses the problem of assessing how much of the initially missing phase information can actually be retrieved from the statistical scores obtained after evaluating all hypotheses in the sample (§3.4).

3.1. Generation Of Diversity: factor permutation.

The generation of a "representative sample" of hypotheses to specify some of the currently missing information may involve a variety of "factors" for which multiple choices remain possible: the unrestricted assignment of trial phase values to totally unphased structure factor amplitudes, or the trial selection of modes if bimodal SIR phase indications pre-exist; choices of plausible molecular boundaries, or of possible redefinitions of an existing boundary; trial placements (i.e.

orientations and positions) of plausible molecular substructures; trial definitions of non-crystallographic symmetry elements and/or of the geometric transformations relating multiple crystal forms; and so on. All these factors have in common an *ability to influence* the expected distribution of the structure factor amplitudes attached to the crystal structure(s) under investigation, and thus a *testability* against observations of these amplitudes.

Because of the very large number of unknown factors, any scheme for factor permutation has to be *hierarchical*, which leads unavoidably to a sequential strategy similar to the exploration of moves in a computer chess-playing program. This can be represented by means of a *phasing tree*, each *level* of which corresponds to a *ply* in a chess game. Each node of the tree is a "factor hypothesis", and the early ruling out of some of these hypotheses is reflected by the *pruning* of the tree.

Call *basis set* at level ℓ, denoted H_ℓ, the set of unique reflexions **h** to which trial phases will be assigned. The hierarchical structure of the search implies that these are *nested*, i.e. that the basis set grows by concatenating successive increments of new reflexions:

$$H_1 = \{ \mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{m_1} \}$$

$$H_2 = H_1 \cup \{ \mathbf{h}_{m_1+1}, \mathbf{h}_{m_1+2} ..., \mathbf{h}_{m_2} \}$$

.....

$$H_\ell = H_{\ell-1} \cup \{ \mathbf{h}_{m_{\ell-1}+1}, \mathbf{h}_{m_{\ell-1}+2}, ..., \mathbf{h}_{m_\ell} \}$$

The simplest way to generate phase hypotheses for the reflexions contained in an increment of the basis set is to consider a regular grid of points around each phase circle, e.g. at 45+90k degrees (k=0,...,3) giving a "quadrant permutation" long used in MULTAN [19], or a more general grid. Much more powerful methods exist, based on error-correcting codes, for sampling several phases simultaneously. They are described elsewhere in [3,20].

For mixed factors involving both phases and other factor types, well-known techniques for designing optimal sampling schemes can be used [21,22]. Incomplete factorial designs, first introduced into crystallography by Carter & Carter [23] for the design of crystallisation experiments, were used successfully as general permutation designs for mixed factors involving phases and binary choices of molecular envelope attributes [12,13,14].

3.2. Expression of a phase hypothesis.

This is the mechanism through which the viability of a phase hypothesis, i.e. its structural realisability, is measured. According to the general scheme of §2 two main quantities need to be evaluated, or at least approximated:

(a) the probability that given phased structure factor values $\{ F_{\mathbf{h}_1}, F_{\mathbf{h}_2}, ..., F_{\mathbf{h}_m} \}$ attached to a basis set $H = \{ \mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_m \}$ belong to a chemically valid structure; this function of the F's is called their *joint probability distribution* and is denoted $P (F_{\mathbf{h}_1}, F_{\mathbf{h}_2}, ..., F_{\mathbf{h}_m})$ or $P(\mathbf{F}_H)$ for short;

(b) the probability distribution of other structure factor values for a set $K = \{ k_1, k_2, ..., k_n \}$ of non-basis reflexions over all possible valid structures compatible with the given phased structure factor values in the basis set H ; this is called the *conditional probability distribution* (cpd) of $\mathbf{F}_K$ given $\mathbf{F}_H$, denoted $P(F_{k_1}, F_{k_2}, ..., F_{k_n} | F_{h_1}, F_{h_2}, ..., F_{h_m})$ or $P(\mathbf{F}_K | \mathbf{F}_H)$ for short,.

All these probabilities can be calculated, to a consistent degree of accuracy in the whole range of applications, by the *maximum-entropy method.* The main features of this calculation, described elsewhere [1,2,24] in more detail, may be summarised as follows.

Let ν be a node of the phasing tree at level ℓ, where the basis set H_ℓ has m reflexions, and let $\mathbf{F}^{(\nu)} = (F^{(\nu)}_{\mathbf{h}_1}, F^{(\nu)}_{\mathbf{h}_2}, \ldots, F^{(\nu)}_{\mathbf{h}_m})$ be the vector of phased structure factor values describing the hypothesis attached to this node. $\mathbf{F}^{(\nu)}$ has n real components, where n is the number of centric plus twice the number of acentric reflexions in H_ℓ

If N identical atoms are thrown at random, independently, with probability density m(x) in the asymmetric unit D, then the *saddlepoint approximation* [1,2,24] to the prior probability of $\mathbf{F}^{(\nu)}$ is given by:

$$\mathcal{P}^{SP}(\mathbf{F}^{(\nu)}) = \frac{e^{NS(\nu)}}{\sqrt{\det(2\pi N \mathbf{Q}^{(\nu)})}} \qquad \text{with} \qquad S(\nu) = \mathcal{S}_m(q_\nu^{ME}) \tag{3.1a}$$

where $\mathcal{S}_m(q)$ denotes the relative entropy functional

$$\mathcal{S}_m(q) = - \int_D q(x) \log\left[\frac{q(x)}{m(x)}\right] d^3x \ , \tag{3.1b}$$

while q_ν^{ME} denotes the unique distribution compatible with the data in $\mathbf{F}^{(\nu)}$ which maximises this relative entropy. Matrix $\mathbf{Q}^{(\nu)}$ is the covariance matrix between the trigonometric structure factor contributions to the components of $\mathbf{F}^{(\nu)}$ when the random atoms are distributed with density q_ν^{ME}; it is calculated by structure factor algebra [1,2,24].

3.3. Assay of phase hypotheses.

To measure the "strength of binding to the data" of a phase hypothesis, we will try and assess to what extent that hypothesis is able to guess some characteristics of the distribution of data it has not yet seen – an idea which bears some similarity to that of cross-validation.

For this purpose we rely on maximum-entropy extrapolation as a prediction mechanism for structure factors: besides reproducing the amplitudes and phases $\mathbf{F}^{(\nu)}$ attached to node ν of the phasing tree for reflexions in the basis set H_ℓ, the maximum-entropy distribution q_ν^{ME} also gives rise to Fourier coefficients $F^{ME}_{\nu,\mathbf{k}}$ with non-negligible amplitude for many non-basis reflexions, i.e.

for $\mathbf{k}$ in the complement K_ℓ of H_ℓ. This phenomenon is known as *maximum-entropy extrapolation*. It is the Bayesian equivalent of the tangent formula of conventional direct methods (see §3.4 of [1] for a more detailed discussion). Intuitively, the maximum-entropy extrapolate $F^{ME}_{\nu,\mathbf{k}}$ is that value of $F_{\nu,\mathbf{k}}$ which can be fitted "for free" once the basis-set constraints $\mathbf{F}^{(\nu)}$ have been fitted.

The conditional distribution $\mathcal{P}(\mathbf{F}_{K_\ell} \mid \mathbf{F}_{H_\ell} = \mathbf{F}_{\nu,H_\ell})$ is affected by this extrapolation, since we may write:

$$\mathcal{P}(\mathbf{F}_K \mid \mathbf{F}_H) \approx \frac{1}{2\pi\sqrt{\det \mathbf{Q}^{ME}_{KK}}} \exp\left\{-\frac{1}{2}\left[\mathbf{F}_K - \mathbf{F}^{ME}_K\right]^T \left[\mathbf{Q}^{ME}_{KK}\right]^{-1} \left[\mathbf{F}_K - \mathbf{F}^{ME}_K\right]\right\} \qquad (3.2)$$

and so will the (marginal) conditional distribution of amplitudes $\mathcal{P}(|\mathbf{F}_{K_\ell}| \mid \mathbf{F}_{H_\ell})$ obtained by integrating over the phases present in $\mathbf{F}_K$. We will therefore obtain a prediction of the distribution of $|\mathbf{F}_{K_\ell}|$ which depends on the phases attached to node ν. The *likelihood* Λ of the trial phases $\Phi(\nu)$ in $\mathbf{F}_{\nu,H_\ell}$ is the conditional probability of the *observed* values for the amplitudes $|\mathbf{F}_{K_\ell}|^{obs}$:

$$\Lambda(\varphi^{(\nu)}_{\mathbf{h}_1},\ldots,\varphi^{(\nu)}_{\mathbf{h}_m}) = \mathcal{P}\left(|F_{\mathbf{k}_1}|^{obs},\ldots|F_{\mathbf{k}_n}|^{obs} \;\middle|\; |F_{\mathbf{h}_1}|^{obs} e^{i\varphi^{(\nu)}_{\mathbf{h}_1}},\ldots,|F_{\mathbf{h}_m}|^{obs} e^{i\varphi^{(\nu)}_{\mathbf{h}_m}}\right) \qquad (3.3).$$

It is often convenient to consider the log-likelihood $L = \log \Lambda$. Let $(\mathcal{H}_0)$ be the null hypothesis that the atoms are uniformly distributed, and $(\mathcal{H}_1)$ be the alternative hypothesis that they are distributed according to q^{ME}_ν. These two hypotheses can be tested against each other by calculating the log-likelihood gain:

$$LLG(\nu) = \log \frac{\mathcal{P}\left(|F_{\mathbf{k}}| = |F_{\mathbf{k}}|^{obs} \text{ for } \mathbf{k}\in K \;\middle|\; \langle F_{\mathbf{h}}\rangle = |F_{\mathbf{h}}|^{obs} \exp(i\varphi^{(\nu)}_{\mathbf{h}}) \text{ for } \mathbf{h}\in H\right)}{\mathcal{P}\left(|F_{\mathbf{k}}| = |F_{\mathbf{k}}|^{obs} \text{ for } \mathbf{k}\in K \;\middle|\; \langle F_{\mathbf{h}}\rangle = 0 \text{ for } \mathbf{h}\in H\right)} \qquad (3.4).$$

This quantity will be largest when the phase assumptions attached to node ν lead one to expect deviations from Wilson statistics for the unphased amplitudes $|F_{\mathbf{k}}|$, $\mathbf{k}\in K$, that most closely match those present in the distribution of the actual measurements $|F_{\mathbf{k}}|^{obs}$. It is therefore a quantitative measure of the degree of corroboration by the unphased data of the phase assumptions attached to ν.

The most fundamental likelihood function is that of Rice[25], derived as the marginal distribution for the amplitude R of an offset 2D Gaussian:

$$\mathcal{R}(r,R,\Sigma) = \frac{R}{\Sigma} \exp\left(-\frac{r^2+R^2}{2\Sigma}\right) I_0\left(\frac{rR}{\Sigma}\right) \qquad (3.5a)$$

in which r is the offset length and Σ the variance parameter. Typically r is the modulus of an F^{calc}, R is the observed modulus $|F|^{obs}$, and the variance parameter Σ is derived e.g. by Wilson statistics to represent the statistical dispersion of a contribution from random atoms.

It is straightforward, starting from a 1D rather than 2D Gaussian, to obtain the centric equivalent of the Rice distribution as:

$$C(r, R, \Sigma) = \sqrt{\frac{2}{\pi\Sigma}} \exp\left(-\frac{r^2+R^2}{2\Sigma}\right) \cosh\left(\frac{rR}{\Sigma}\right) \qquad (3.5b).$$

The derivation of both likelihood functions assumes that all phases or signs over which integration is carried out have equal probability. In the acentric case it is further assumed that the original 2D Gaussian is isotropic. The first assumption is invalid when prior phase information is available from MIR or MAD, while the second is violated when the distribution of random atoms is strongly non-uniform or obeys non-crystallographic symmetries. Fortunately both of these generalisations can be carried out simultaneously by defining the *elliptic* Rice distribution (although Rice never considered it) required for a "phased LLG", which was derived in [26].

As a decision criterion likelihood enjoys certain optimality properties both in conventional statistics (Neyman-Pearson theorem) and of course in Bayesian statistics through its role in Bayes's theorem (2.1). In both settings, likelihood evaluation or optimisation is the common mechanism through which the Bayesian structure determination process interrogates the observed data for the purpose of testing or refining hypotheses. The examples given in sections 4 to 7 provide an ample illustration of the confluence of a wide range of hitherto distinct detection and refinement operations into a single calculation, namely computing value and derivatives for a suitable likelihood function.

Whenever likelihood does not totally dominate prior knowledge, the full force of Bayes's theorem should be invoked. Using Bayes's theorem in the form (2.1) together with expression (3.1a,b) for the prior probability of a set of trial structure factor values, the *a-posteriori* probability $P^{post}(\nu)$ of the phase hypothesis attached to node ν of the phasing tree may be evaluated by first computing the Bayesian score:

$$B(\nu) = N\,S(\nu) - \frac{1}{2}\log\det(2\pi Q(\nu)) + LLG(\nu) \qquad (3.6)$$

(where N is the number of independent atoms in the asymmetric unit) and exponentiating it, then normalising this collection of numbers over a suitable collection of nodes ν.

3.4. Analysis of phase hypotheses.

In the course of the tree-directed search described above, a subpopulation of nodes with high scores will progressively be selected. Rather than consider this list of instances as the end product of the phasing process, we want to do some data reduction and relate the property of achieving these high scores to the property of having the right values for some critical phases or combinations of phases (or other factors).

Since the initial sampling plan according to which the nodes were generated (§3.1) will normally be based on some efficient sampling design, this is a typical setting in which to call upon standard techniques for analysing the results of "designed experiments" [27-30].

Phases, however, are special factors because of their periodic character. As a result any phase-dependent score function, for instance the Bayesian score $B = B(\varphi_1, ..., \varphi_m)$, is a *periodic* function, with an m-dimensional period lattice generated by translations of 2π along the coordinate axis of each of the m basis-set phases. It may therefore be written

$$B(\varphi_1, ..., \varphi_m) = \sum_{j_1} ... \sum_{j_m} C_{j_1 ... j_m} e^{i(j_1\varphi_1 + ... + j_m\varphi_m)} .$$

We are looking for those $C_{j_1 ... j_m}$ which are significantly higher in the successful nodes that in the general population, i.e. higher than one would expect in the absence of any trends. The issues of optimal sampling and of statistical analysis of node scores therefore belong to the realm of multidimensional Fourier analysis [3,20] . For factors other than phases, which may not be periodic, conventional techniques such as multiple linear regression [27] may be used.

Standard methods are available for assessing the level of significance of the results of a multiple regression analysis of experimental scores [27]. The simplest of them, namely the application of Student's t-test to the determination of a single sign, was described in §2.2.4 of [3] and was used [31] in the solution of a powder structure.

4. Overview of selected applications.

In the Bayesian picture, the privileged role played by the Rice likelihood functions in consulting the experimental observations leads naturally to delineating three main "regimes" in the process of structure determination, corresponding to three distinct approximation regimes for these functions: a "Patterson correlation regime", a "transition regime" and a "Fourier correlation regime". A number of examples pertaining to each regime will be worked out in some detail in the forthcoming sections. It will be a recurring observation that the Bayesian analysis turns out to yield improvements of these methods which had not (or had only just) arisen within their own theoretical framework, at the same time as providing an automatic unification of these improvements within a common computational mechanism.

4.0. Approximation regimes for the Rice log-likelihood functions.

We will begin with a few basic results on Wilson statistics and associated notation. For **h** acentric, F(**h**) is distributed as a 2D Gaussian centred at (0,0) with variance

$$\Sigma_a(\mathbf{h}) = \tfrac{1}{2} |G_{\mathbf{h}}| \sigma_2(\mathbf{h}) \qquad (4.0a)$$

along each component, while for **h** centric it is distributed as a 1D Gaussian centred at 0 with variance

$$\Sigma_c(\mathbf{h}) = |G_{\mathbf{h}}|\,\sigma_2(\mathbf{h}) \tag{4.0b}$$

where

$$\sigma_2(\mathbf{h}) = \sum_{\text{cell}} \left[f_j(\mathbf{h})\right]^2 . \tag{4.0c}$$

It will also be recalled, for later use, that standard normalised structure factor amplitudes $|E(\mathbf{h})|$ are defined by

$$|E(\mathbf{h})|^2 = \frac{|F(\mathbf{h})|^2}{2\Sigma_a(\mathbf{h})} \qquad \text{for } \mathbf{h} \text{ acentric} \tag{4.0d}$$

and

$$|E(\mathbf{h})|^2 = \frac{|F(\mathbf{h})|^2}{\Sigma_c(\mathbf{h})} \qquad \text{for } \mathbf{h} \text{ centric} \tag{4.0e}$$

so that

$$\langle\, |E(\mathbf{h})|^2 \,\rangle = 1 \qquad \text{for all } \mathbf{h} \tag{4.0f}.$$

Define the acentric and centric Rice log-likelihood functions L_a and L_c as the logarithms of (3.5a) and (3.5b) using variances Σ_a and Σ_c respectively, and introduce the following shorthand for the Sim figures of merit:

$$z_a = \frac{rR}{\Sigma_a} , \qquad z_c = \frac{rR}{\Sigma_c} , \qquad m_a(z_a) = \frac{I_1(z_a)}{I_0(z_a)} , \qquad m_c(z_c) = \tanh z_c \tag{4.1}$$

where I_0 and I_1 are the modified Bessel functions of order 0 and 1 respectively. It is then straightforward to obtain the following partial derivatives:

$$\frac{\partial L_a}{\partial r} = \frac{1}{\Sigma}\left(m_a(z_a)\,R - r\right) \tag{4.2a}$$

$$\frac{\partial^2 L_a}{\partial r^2} = \frac{1}{\Sigma}\left[\frac{R^2}{\Sigma}\left(1 - \frac{m_a(z_a)}{z_a} - m_a(z_a)^2\right) - 1\right] \tag{4.2b}$$

$$\frac{\partial L_a}{\partial \Sigma} = \frac{1}{\Sigma}\left[\frac{r^2 + R^2 - 2\,m_a(z_a)\,r\,R}{2\Sigma} - 1\right] \tag{4.2c}$$

and

$$\frac{\partial L_c}{\partial r} = \frac{1}{\Sigma}\left(m_c(z_c)\,R - r\right) \tag{4.3a}$$

$$\frac{\partial^2 L_c}{\partial r^2} = \frac{1}{\Sigma}\left[\frac{R^2}{\Sigma}\left(1 - m_c(z_c)^2\right) - 1\right] \tag{4.3b}$$

$$\frac{\partial L_c}{\partial \Sigma} = \frac{1}{2\Sigma}\left[\frac{r^2 + R^2 - 2\,m_c(z_c)\,r\,R}{\Sigma} - 1\right] \tag{4.3c}$$

which will be used to build Taylor expansions of L_a and L_c for small variations of r and Σ.

When the quantity z in (4.1) becomes large, the asymptotic formulae

$$I_0(z) \approx \frac{e^z}{\sqrt{2\pi z}} \qquad \text{and} \qquad \cosh z \approx \frac{1}{2} e^z \tag{4.4a,b}$$

give rise to the following approximations:

$$\mathcal{R}(r, R, \Sigma) \approx \sqrt{\frac{R}{r}} \sqrt{\frac{1}{2\pi\Sigma}} \exp\left(-\frac{(R-r)^2}{2\Sigma}\right) \tag{4.5a}$$

$$C(r, R, \Sigma) \approx \sqrt{\frac{1}{2\pi\Sigma}} \exp\left(-\frac{(R-r)^2}{2\Sigma}\right) \tag{4.5b}.$$

4.1. The Patterson correlation regime.

In detection problems the standard situation is that a null hypothesis with vanishing offsets r and some initial variances Σ is to be compared to alternative hypotheses with non-zero offsets δr and lower variances $\Sigma - \delta\Sigma$, where $\delta\Sigma$ is related to $\langle (\delta r)^2 \rangle$ through Wilson statistics.

For $r = 0$ we have $m_a(z_a) = m_c(z_c) = 0$ so that the first-order derivatives (4.2a) and (4.3a) vanish. Using the limiting value $\left.\frac{m_a(z_a)}{z_a}\right|_{z_a=0} = \frac{1}{2}$ and the basic relations (4.0a-f) we get the local Taylor expansions :

$$\delta L_a = L_a(\delta r, R, \Sigma_a - \delta\Sigma_a) - L_a(0, R, \Sigma_a) = \frac{1}{\Sigma_a}\left(|E|^2 - 1\right)\left(\tfrac{1}{2}(\delta r)^2 - \delta\Sigma_a\right) \tag{4.6a}$$

$$\delta L_c = L_c(\delta r, R, \Sigma_c - \delta\Sigma_c) - L_c(0, R, \Sigma_c) = \frac{1}{\Sigma_c}\left(|E|^2 - 1\right)\left(\tfrac{1}{2}(\delta r)^2 - \tfrac{1}{2}\delta\Sigma_c\right) \tag{4.6b}$$

When summed over all reflexions these quantities will be shown to give rise to correlation functions between the origin-removed $|E|^2$-based Patterson function for the observed data and the origin-removed Patterson for the partial structure whose presence is to be detected.

Once a partial structure has been detected, subsequent searches for more partial structures can be carried out with the benefit of the phase information generated by the first: one has then entered the "transition regime" (§4.2 below).

4.2. The transition regime.

The transition regime applies when the offsets r are no longer zero but the quantities z in (4.1) are small (a few units at most). The first-order derivatives (4.2a) and (4.3a) no longer vanish and hence create in the Taylor expansion of L_a or L_c an incipient Fourier-like sensitivity to the existing phase information. However the second-order terms remain substantial and continue contributing Patterson-like features. The simple proportionality relations between $\frac{\partial^2 L}{\partial r^2}$ and $\frac{\partial L}{\partial \Sigma}$ for

$r=0$ which give rise to (4.6a,b) cease to hold: the $|E|^2$ associated to these two types of partial derivatives are *renormalised* differently under the available phase information.

This is the "middle game" situation, where phase information is beginning to emerge but remains highly ambiguous. A large fraction of the observed intensities has to be accounted for through variance rather than expectation, reflecting the large amount of missing phase information. A useful measure of missing phase information at a reflexion $\mathbf{h}$ is the renormalised $|E|^2$ associated to $\frac{\partial L}{\partial \Sigma}$, which may be written

$$\left|E_{\mathbf{h}}^{\text{renorm}}\right|^2 = \left[r^2 + R^2 - 2\, m_a(z_a)\, r\, R\right] / 2\Sigma_a(\mathbf{h}) \qquad \text{for } \mathbf{h} \text{ acentric} \qquad (4.7a)$$

$$= \left[r^2 + R^2 - 2\, m_c(z_c)\, r\, R\right] / \Sigma_c(\mathbf{h}) \qquad \text{for } \mathbf{h} \text{ centric} \qquad (4.7b).$$

This quantity was considered by Nixon & North [32]. Reflexions $\mathbf{h}$ for which it is the largest are those where phase information is "most sorely missing", and hence are the best candidates for phase permutation (§3.1), or for the permutation of any other factor capable of leading to a lower renormalised $|E|^2$. This selection procedure was used successfully in the statistical phasing of Tryptophanyl-tRNA synthetase [12,13,14].

4.3. The Fourier correlation regime.

This regime is reached when most of the quantities z in (4.1) have appreciable values (several units). According to (4.5a,b) the likelihood function becomes approximately that of a least-squares residual on amplitudes, but with the important feature that *the variances may still come mostly from the structure factor statistics* rather than from the observational errors.

Current structure refinement protocols (PROLSQ [33], TNT [34] and XPLOR [35]) are still based on a least-squares residual, hence ignore this extra source of variance. As a result, they suffer from well-known problems of model bias. Maximum-likelihood structure refinement is proposed in §7.2 as a superior alternative to least-squares refinement, and this claim is supported by encouraging test results.

4.4. Introduction to the examples.

The selection of illustrations given in the forthcoming sections comprises a varied range of applications taken from both macromolecular and direct methods techniques. In the latter, the Bayesian viewpoint brings simplicity and clarity to what has traditionally been a thicket of formulae, and enlarges both their scope and their effectiveness. In the former, it also provides a thread of continuity between hitherto distinct techniques and reveals numerous possibilities of substantial improvements. Most importantly, this survey demonstrates the extent to which all existing techniques are subsumed within a unique computational protocol which needs no longer be aware of the multitude of specialisations through which it has so far been approximated.

5. Application to detection problems.

5.1. Heavy-atom detection in a structure.

In the standard use of the method for small molecules the "heavy atom" is first detected by examination of the Patterson, taking advantage of the fact that such an atom is localised and poses no rotation problem. One then switches over to Sim's formula [36] which can be expected to hold rather accurately since the 'light' atoms making up the rest of the structure are distributed uniformly enough for Wilson's statistics to be obeyed (small-molecule crystals are usually close-packed – there is no solvent – and the exclusion of light atoms by the heavy atom can be neglected in most cases).

The statistical treatment of the same problem shows readily that the optimal detection function under the "Patterson regime" approximation [26,37] has the form of a Patterson correlation (PC) function [38] between the origin-removed $|E|^2$–based Patterson for the whole structure and the heavy-atom origin-removed Patterson.

This approximation is valid only if the light atoms are distributed so as to give rise to Wilson statistics; it may be known that this is not the case, e.g. in crystals of zeolites or other small structures containing cavities, and of course in macromolecular crystals. A full LLG evaluation has no difficulty in remaining an optimal detection criterion in this case, while the PC coefficient will fail to do so. Furthermore the statistical variances used can be incremented so as to reflect measurement errors, while there is no natural way of doing so when calculating the PC coefficient.

5.2. Heavy-atom detection in the MIR and MAD methods.

The statistical theory of heavy-atom parameter refinement, and the SHARP program implementing it, are dealt with in [39]. However the problem of statistical heavy-atom detection is best illustrated by reference to analytical formulae for SIR likelihood functions derived in [40].

Statistical detection begins with a maximum-likelihood estimation of scale factors and of non-isomorphism $\sigma_2^{n\text{-}iso}$ under the null hypothesis that all discrepancies between $|F^P|$ and $|F^{PH}|$ are caused by non-isomorphism. The alternative hypotheses assume instead the presence of a "nascent" heavy atom at x^H whose occupancy increases from 0 while $\sigma_2^{n\text{-}iso}$ decreases by the corresponding amount as in §5.1.2. Approximating $LLG(x^H)$ by a second order Taylor expansion [26] yields a detection criterion of the form of a Patterson correlation function involving quantities which can be written as $\left((\Delta E)^2_{iso} - 1\right)$ and may be recognised as the normalised (sharpened) and origin-removed versions of coefficients advocated by Kalyanaraman & Srinivasan [41] as being the best ones from which to compute a difference-Patterson function for the determination of heavy-atom positions using isomorphous data. Preliminary tests in SHARP have shown that this criterion is indeed very promising for weak isomorphous signals, and a generalisation of the Rice distibution is being developed for detecting anomalous scatterers from MAD data sets.

5.3. Fragment detection in the molecular replacement method.

Instead of a heavy atom as in §5.1 we now have a known fragment described in a reference position and orientation by a density ρ^M with transform F^M. If ρ^M is rotated by $\mathbf{R}$ and translated by $\mathbf{t}$ to give the copy of the fragment lying in the chosen asymmetric unit, then statistical detection and placement of the fragment will proceed by calculating the log-likelihood gain

$$\mathrm{LLG}(\mathbf{R},\mathbf{t}) = \log \frac{\mathcal{P}\left(\left|F_{\mathbf{h}}\right| = \left|F_{\mathbf{h}}\right|^{\mathrm{obs}} \text{ for all } \mathbf{h} \;\middle|\; (\mathcal{H}_1[\mathbf{R},\mathbf{t}]) \right)}{\mathcal{P}\left(\left|F_{\mathbf{h}}\right| = \left|F_{\mathbf{h}}\right|^{\mathrm{obs}} \text{ for all } \mathbf{h} \;\middle|\; (\mathcal{H}_0) \right)} \tag{5.1}$$

where $(\mathcal{H}_0)$ denotes the null hypothesis that all atoms (including those of the fragment) are uniformly distributed in the asymmetric unit while $(\mathcal{H}_1[\mathbf{R},\mathbf{t}])$ denotes the alternative hypothesis that a subset of atoms is assembled into the known fragment and placed in the asymmetric unit with orientation $\mathbf{R}$ at position $\mathbf{t}$, and the rest are distributed at random.

The methods used in §5.1 then carry over to the present situation and yield [26,42] a detection criterion consisting of two terms. The first term depends only on the rotational placement $\mathbf{R}$ and is a PC-based rotation function in which a sum of point-group symmetry-related copies of the origin-removed self-Patterson of the rotated fragment is being correlated with the origin-removed sharpened self-Patterson of the whole structure. The second term, considered for a fixed value of the rotational component $\mathbf{R}$ of the placement, gives rise to a PC-based translation function, expressed as a Fourier series. The fact that the log-likelihood gain (which is an optimal criterion by the Neyman-Pearson theorem is based on E's provides a final explanation to the long-standing observations by Ian Tickle [42] that E-based (sharpened) translation functions always give better results than F-based (unsharpened) ones.

It is therefore clear in this case too that even the most approximate implementation of the statistical detection approach yields better criteria than the most sophisticated ones available so far, and suggests non-trivial improvements of the existing methodology which had not yet arisen within this methodology itself.

5.4. Detection of non-uniformity from variance modulation.

In small-molecule direct methods the phase determination process is often primed by means of so-called Σ_1 relations, which give immediate estimates of certain phases belonging to a subclass of reflexions (see e.g. [43]). Giacovazzo [44,45] and Pavelčík [46,47] have found a connection between these relations and Harker sections in Patterson functions.

It was shown in [26] that Σ_1 relations are related to the sensitivity of the Rice log-likelihood to its variance parameter and to the modulation of the latter by the non-uniformity of atomic distributions. Thus Σ_1 relations are a purely variance-based method for detecting non-uniform distribution of *all* atoms, by the same statistical technique (likelihood-based hypothesis testing) which was used earlier for detecting the non-uniform distribution of *one* heavy atom in a background of light-atoms.

6. Application to completion problems.

6.1. Detection of further heavy atoms or fragments.

All detection problems in sections 5.1 to 5.3 were examined under the assumption that no phase information existed (there is no basis set), so that the ordinary Rice likelihood functions (3.18a,b) apply. If we now assume that some external phase information has become available, the elliptic Rice likelihood function [26] should be used instead. If this phase information is strong and unimodal, then the LLG is essentially an electron density correlation function (a "phased rotation/translation function") as discussed in §4.3. If not, the LLG will possess features intermediate between those of a Patterson correlation coefficient and of a density correlation coefficient, giving rise to what might be called *partially phased* rotation and translation functions.

The log-likelihood gradient maps used in SHARP (§7.1 and [39]) are based on this idea. They have proved highly successful in revealing fine details of heavy-atom substitution such as minor sites, anisotropy of thermal disorder, and split sites caused by multiple conformations.

6.3. Maximum-entropy solvent flattening.

When a great deal of phase information is specified in H and the prior prejudice m(x) is non-uniform, the expressions above are of little help but numerical computation can proceed unimpeded. The first successful use of this phase extension procedure was reported in [11]. It showed on the cytidine deaminase structure that maximum-entropy solvent flattening (MESF) using the solvent mask as a non-uniform prior prejudice provided a better method of phase extension than did ordinary solvent flattening [48]. An interesting aspect of this work lies in the protocol used to build the maximum-entropy distribution when the basis-set phase information consists of MIR phases and is therefore tainted with noise. The preferred way of fitting such constraints would be to aim for the maximum Bayesian score (3.6), but this would require knowing the value N of the number of atoms in the structure, a quantity which is difficult to define at non-atomic resolution [1]. Instead, the fitting the noisy constraints was allowed to proceed only as long as the LLG outside the basis set kept increasing, and was halted when the LLG reached a maximum. In this way, only that part of the noisy constraints was fitted which contains the 'signal', i.e. which improves the predictive power of the statistical model. This is the familiar idea of cross-validation, and in fact this procedure carries out something akin to an estimation of the "effective N" by cross-validation, as well as cross-validated density modification by exponential modelling.

6.4. Hypothesis permutation in the 'middle game' of structure determination.

The crystal structure determination of the tetragonal form of *Bacillus Stearothermophilus* tryptophanyl-tRNA synthetase (TrpRS) provided the first application to the determination of an unknown macromolecuar structure of the full Bayesian scheme (§2) for inferring missing phase information.

Unlike the case of cytidine deaminase on which the MESF procedure was first tested (§6.3), the case of TrpRS was marred by serious non-isomorphism in the heavy-atom derivatives, resulting in large starting-phase errors and hence in a poor definition of the molecular envelope. MESF proved unable to produce better maps from such an unfavourable starting point: instead it led to a severe deterioration of the maps, accompanied by a dramatic decrease of the LLG statistic as phases were extended from about 5.0Å resolution to the 2.9Å limit of the diffraction data. It was therefore necessary to somehow improve simultaneously the quality of the starting phases and the correctness of the molecular envelope - a task whose circularity from the conventional standpoint made it at first sight as impossible as lifting oneself by one's bootstraps.

The deadlock was broken by a straightforward application of the exploratory process described in §2. The "most sorely missing information" was found to be associated with strong reflexions having large renormalised $|E|^2$ values (§4.2) which were initially unphased and were inaccessible by maximum-entropy extrapolation from the phased ones. Their phases were permuted, at first on their own, then together with permuted hypotheses concerning possible modifications of the molecular envelope. All permutations were carried out by using incomplete factorial designs [23]. For each such specification of the new phases and of the envelope the MESF process was applied, the maximal value reached by the LLG statistic was noted, and these scores were subsequently analysed by multiple-regression least-squares. Student t-tests were performed to assess significance, and turned out to provide reliable indications for most of the phases of 28 strong reflexions and for the six binary choices of envelope attributes involved in the permutations. The resulting phase improvement made it possible to assign positions, hitherto unobtainable, for nine of the ten selenium atoms in an isomorphous difference Fourier map for SeMet-substituted TrpRS. Further phase permutation continued to produce improved maps from the pooled MIR phase information and played a critical role in solving the structure [12]. This is the first practical demonstration of the effectiveness of the Bayesian approach at a typical macromolecular resolution [14].

The use of the renormalised $|E|^2$ value as a criterion for choosing candidates for phase permutation bears an interesting relationship to the expression giving the mean-square noise level in a centroid map as a function of the distribution of figures of merit [49,50,51]. Indeed $\left|E_{\mathbf{h}}^{\text{renorm}}\right|^2$ is closely analogous to the quantity $|F_{\mathbf{h}}|^2 \times (1-m_{\mathbf{h}}^2)$, where $m_{\mathbf{h}}$ is the figure of merit [18], which gives the contribution of $\mathbf{h}$ to the overall noise in the centroid map. Permuting the phases of reflexions having the largest $\left|E_{\mathbf{h}}^{\text{renorm}}\right|^2$ is therefore the fastest way to "remove heat from the system". Acentric reflexions of this type fall into two distinct categories, according to the character of their phase probability densities: (1) those for which it is *flat*, in which case phase permutation has to proceed in the same way as in *ab initio* phasing; (2) those for which it is *bimodal*, for which it

is preferable to use *mode permutation* which boils down to a simple binary choice. In the latter case simultaneous choices of modes for several reflexions may be sampled efficiently by invoking the combinatorial techniques described in [20] These multiple choices may then be evaluated by means of the elliptic Rice likelihood [26] which measures the extent to which the phases extrapolated from each combination of binary choices of modes in the basis set agree with one of the modes for each second neighbourhood reflexion.

6.6 Outlook.

If the expected gains in sensitivity of detection (§5.3), in efficiency of recycling (§6.2) and in effectiveness of completion (§6.4-5) brought about by the full implementation of the Bayesian approach actually materialise, it would become conceivable to attack the *ab initio* determination of protein structures by systematically searching for super-secondary fragments of 20 to 30 amino-acids, for which it may be possible to compile a library similar to the library of short fragments used for assisting map interpretation by Jones & Thirup [52]. If this does not work without some startup phase information, then an initial round of phase permutation may afford a means of building up enough phase information *ab initio* in order to increase the detection sensitivity for such fragments above a critical threshold.

This line of development is connected to the more radical approach presented in [15] where the stereochemical information pertaining to such libraries of fragments is incorporated directly into the structure factor statistics from which joint probabilities are built.

7. Application to refinement problems.

7.1. Maximum-Likelihood heavy-atom refinement (SHARP).

This topic is treated in a separate contribution to this Volume. The main difference between least-squares (LS) and maximum-likelihood (ML) parameter refinement resides in the fact that one integrates the partial derivatives around the native phase circle, just as one integrates the structure factor in the Blow & Crick method [18]. This removes the bias previously introduced by phase "estimates" [53] which were particularly questionable in the SIR method. A more subtle difference is that this integration is carried out also in the radial direction to deal with the measurement errors on native amplitudes, and with the absence of a native measurement in the MAD method. Last but not least, the ML method also allows the refinement of parameters describing the lack of isomorphism and hence influencing the weighting of observations; such a refinement is impossible within the least-squares method where weights are necessarily assumed to be fixed.

The rapidly growing list of SHARP successes and the remarkable sensitivity of its LLG gradient maps demonstrate that the full implementation of the ML method for heavy-atom parameter refinement was well worth the extra effort it required.

7.2. Maximum-Likelihood structure refinement.

The Bayesian viewpoint has long suggested that structure refinement should be carried out by maximising the LLG rather than by minimising the conventional least-squares residual [2,7,37]. Here again, only the ML method can take into account the uncertainty of the phases associated to model incompleteness and imperfection by suitably downweighting the corresponding amplitude constraints. It was predicted [37] that ML refinement would allow the refinement of an incomplete model by using the structure factor statistics of randomly distributed scatterers to represent the effects of the missing atoms, in such a way that the latter would not be wiped out; and that the final LLG gradient map would then provide indications about the location of these missing atoms.

These predictions have now been confirmed by actual tests. Bricogne & Irwin have used BUSTER and TNT on a test data set for crambin [54] suffering from both model imperfection and model incompleteness, and compared the results of LS and ML refinements from these data [55]. In these conditions ML refinement clearly outperformed LS refinement, giving a mean-square distance to the correct positions of 0.176 (ML) instead of 0.415 (LS). Furthermore the final LLG gradient map produced by the ML method showed highly significant, correct connected features for the missing part (40%) of the molecule, while the final LS difference map showed no such features. This enhances the possibilities of "bootstrapping" from an otherwise unpromising molecular replacement starting point to a complete structure. Essentially the same behaviour was observed at 2.0Å resolution, and with experimental rather than calculated data.

Other prototypes for ML structure refinement have been built and tested by Read [56] (using XPLOR and an intensity-based LLG) and by Morshudov [57] (using PROLSQ and the Rice LLG). The BUSTER+TNT prototype has the advantage of being able to use external phase information by means of the elliptic Rice function (see [26]), as well as prior information about non-uniformity in the distribution of the missing atoms in incomplete models. It also allows the ML refinement of an incomplete model to be carried out in conjunction with phase permutation or phase refinement for those strong amplitudes which are most poorly phased by that model, i.e. have the largest renormalised $|E|$'s ; or in conjunction with maximum-entropy updating of the distribution of random atoms, initially taken as essentially featureless within the given envelope. Using the method of joint quadratic models of entropy and LLG described in [1] before and after refinement of the incomplete model produced updated ME distributions showing the missing structure in its entirety, demonstrating clearly the advantage of carrying out ML refinement within the integrated statistical framework provided by BUSTER. This ME "after-burner" establishes a seamless continuity between the middle game of structure determination and the end game of structure refinement.

8. Conclusion.

As shown by the current applications, ranging from *ab initio* phasing through structure completion to structure refinement, all aspects of the determination of crystal structures are

inextricably linked to a common body of statistical methods and concepts. The goal of the "Bayesian programme" is to invite a comprehensive implementation of these methods into an integrated software system for users of crystallographic phasing techniques in the macromolecular field.

An appreciation of the benefits which can be expected to follow from this systematic approach can be obtained by recalling that, a very short while ago, it would have seemed insane to even consider using experimental heavy-atom phases in macromolecular structure refinement: heavy-atom phasing still produced severely biased results, creating numerous carbuncles in electron-density maps; and structure refinement suffered from the biases of the least-squares method which would have further amplified the ugly effects of these phase errors. Today, ML heavy-atom refinement with SHARP delivers safe experimental phase probability densities; and ML structure refinement using for example BUSTER+TNT with the elliptic Rice likelihood function can now safely exploit this phase information to widen the domain of convergence of the refinement.

Finally it should be clear that the statistical analysis at stage 0 of the basic procedure in §2, which allows one to identify the "most sorely missing phase information" for the purpose of seeking to obtain it *computationally*, could equally well be invoked dynamically for deciding how to obtain it *experimentally* in the most effective way. Such a procedure of "Phasing on the beamline" would extend the field of use of statistical methods into the realm of experimental strategy.

Acknowledgements.

I thank my colleagues in the BUSTER Development Group (John Irwin, Eric de La Fortelle and Pietro Roversi) for their numerous contributions towards the realisation of the Bayesian programme.

This research was supported in part by a Tage Erlander Guest Professorship from the Swedish Natural Sciences Research Council (NFR) in 1992-93, by an International Research Scholars award from the Howard Hughes Medical Institute since 1993, and by grants from Pfizer Inc. and Glaxo-Wellcome PLC since 1996.

References.

[1] G. Bricogne, *Acta Cryst.* **A40**, 410-445 (1984).
[2] G. Bricogne, *Acta Cryst.* **A44**, 517-545 (1988).
[3] G. Bricogne, *Acta Cryst.* **D49**, 37-60 (1993).
[4] G. Bricogne and C.J. Gilmore, *Acta Cryst.* **A46**, 284-297 (1990).
[5] C.J. Gilmore, G. Bricogne, and C. Bannister, *Acta Cryst.* **A46**, 297-308 (1990)..
[6] C.J. Gilmore and G. Bricogne, this Volume.
[7] G. Bricogne, *Acta Cryst.* **A47**, 803-829 (1991).
[8] C.J. Gilmore, K. Henderson, and G. Bricogne, *Acta Cryst.* **A47**, 830-841 (1991).
[9] C.J. Gilmore, A.N. Henderson, and G. Bricogne,. *Acta Cryst.* **A47**, 842-846 (1991).
[10] W. Dong, T. Baird, J.R. Fryer, C.J. Gilmore, D.D. MacNicol, G. Bricogne, D.J. Smith, M.A. O'Keefe and S. Hovmöller, *Nature* **355**, 605-609 (1992).
[11] S. Xiang, C.W. Carter Jr., G. Bricogne and C.J. Gilmore (1993). *Acta Cryst.* **D49**, 193-212.
[12] S. Doublié, S. Xiang, C.J. Gilmore, G. Bricogne and C.W. Carter Jr. *Acta Cryst.* **A50**, 164-182 (1994).
[13] S. Doublié, G. Bricogne, C.J. Gilmore and C.W. Carter Jr. (1995). *Structure* **3**, 17-31.
[14] C.W. Carter Jr. (1995). *Structure* **3**, 147-150.
[15] G. Bricogne, *Methods in Enzymology* **277**, in the press.
[16] S. French, *Acta Cryst.* **A34**, 728-738 (1978).

[17] E.T. Jaynes, *Papers on Probability, Statistics and Statistical Physics*, edited by R.D. Rosenkrantz, D. Reidel Publishing Co, Dordrecht (1983).
[18] D.M. Blow and F.H.C. Crick, *Acta Cryst.* **12**, 794-802 (1959).
[19] G. Germain and M.M. Woolfson, *Acta Cryst.* B**24**, 91-96 (1968).
[20] G. Bricogne, *Methods in Enzymology*, **276**, 424-448.
[21] W.G. Cochran and G.M. Cox, *Experimental Designs*, 2nd edition, John Wiley and Sons, New York (1957).
[22] A.C. Atkinson and A.N. Donev, *Optimum Experimental Designs*, Clarendon Press, Oxford (1992).
[23] C.W. Carter Jr. and C.W. Carter, *J. Biol. Chem.* **254**, 12219-12223 (1979).
[24] G. Bricogne, In *Maximum Entropy in Action*, edited by B. Buck and V.A. Macaulay, 187-216, Oxford University Press, Oxford (1991).
[25] S.O. Rice, *Bell System Tech. J.* **23**, 283-332 (parts I and II) ; **24**, 46-156 (parts III and IV) (1944, 1945). Reprinted in *Selected Papers on Noise and Stochastic Processes*, ed. by N. WAX, 133-294, Dover Publ., New York (1954).
[26] G. Bricogne, *Methods in Enzymology*, **276**, 361-423.
[27] R.H. Myers, *Classical and Modern Regression with Applications*, 2nd edition. PWS-KENT, Boston (1986).
[28] R. Mead, *The design of experiments*, Cambridge University Press, Cambridge (1988).
[29] J.A. John and M.H. Quenouille, *Experiments: Design and Analysis*, Griffin, London (1977).
[30] P.W.M. John, *Statistical Design and Analysis of Experiments*, Macmillan, New York (1971).
[31] K. Shankland, C.J. Gilmore, G. Bricogne, and H. Hashizume, *Acta Cryst.* A**49**, 493-501 (1993).
[32] P.E. Nixon and A.C.T. North, *Acta Cryst.* A**32**, 325-333 (1976).
[33] J.H. Konnert and W.A. Hendrickson, *Acta Cryst.* A**36**, 344-349 (1980).
[34] D.E. Tronrud, L.F. Ten Eyck, and B.W. Matthews, *Acta Cryst.* A**43**, 489-501 (1987).
[35] A.T. Brünger, J. Kuriyan, and M. Karplus, *Science* **235**, 458-460 (1987).
[36] G.A. Sim, *Acta Cryst.* **12**, 813-815 (1959).
[37] G. Bricogne, In *The Molecular Replacement Method. Proceedings of the CCP4 Study Weekend 31 January – 1st February 1992*, edited by W. Wolf, E.J. Dodson and S. Gover, 62-75, Daresbury Laboratory, Warrington (1992).
[38] M. Fujinaga and R.J. Read, *J. Appl. Cryst.* **20**, 517-521 (1987).
[39] E. de La Fortelle and G. Bricogne, *Methods in Enzymology*, **276**, 472-494.
[40] G. Bricogne, In *Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P.R. Evans and A.G.W. Leslie, 60-68. Daresbury Laboratory, Warrington (1991).
[41] A.R. Kalyanaraman and R. Srinivasan, *Z. Kristallogr.* **126**, 262-??? (1968).
[42] I. Tickle, In *Molecular Replacement. Proceedings of the CCP4 Study Weekend 15-16 February 1985*, edited by P.A. Machin, 22-26. Daresbury Laboratory, Warrington (1985).
[43] M.F.C. Ladd and R.A. Palmer, *Theory and Practice of Direct Methods in Crystallography*, Plenum Press, New York (1980).
[44] G. Ardito, G. Cascarano, C. Giacovazzo, and M. Luic, *Z. Kristallogr.* **172**, 25-34 (1985).
[45] G. Cascarano, C. Giacovazzo, M. Luic, A. Pifferi, and R. Spagna, *Z. Kristallogr.* **179**, 113-125 (1987).
[46] F. Pavelčík, *J. Appl. Cryst.* **22**, 181-182 (1989).
[47] F. Pavelčík, *Acta Cryst.* A**46**, 869-870 (1990).
[48] B.C. Wang, In *Methods in Enzymology*, Vol. 115, *Diffraction Methods for Biological Macromolecules*, edited by H. Wyckoff, C.W. Hirs and S.N. Timasheff, 90-112. Academic Press, New York (1985).
[49] R.E. Dickerson, J.C. Kendrew, and B.E. Strandberg, In *Computing Methods and the Phase Problem in X-ray Crystal Analysis*, edited by R. Pepinsky, J.M. Robertson and J.C. Speakman, 236-251, Pergamon Press, Oxford (1961).
[50] R.E. Dickerson, J.C. Kendrew, and B.E. Strandberg, *Acta Cryst.* **14**, 1188-1195 (1961).
[51] K.C. Holmes, In *Computing Methods in Crystallography*, edited by J.S. Rollett, 183-203. Pergamon Press, Oxford (1965).
[52] T.A. Jones and S. Thirup, *EMBO J.* **5**, 819-822 (1986).
[53] D.M. Blow and B.W. Matthews, *Acta Cryst.* A**29**, 56-62 (1973).
[54] W.A. Hendrickson and M.M. Teeter, *Nature* **290**, 107-109.
[55] G. Bricogne and J.J. Irwin, in *Macromolecular Refinement*, edited by M. Moore and E.J. Dodson, 85-92. Daresbury Laboratory, Warrington (1996).
[56] N.S. Pannu and R.J. Read, in *Macromolecular Refinement*, edited by M. Moore and E.J. Dodson, 75-84. Daresbury Laboratory, Warrington (1996).
[57] G.N. Morshudov, E.J. Dodson, and A.A. Vagin, in *Macromolecular Refinement*, edited by M. Moore and E.J. Dodson, 93-104. Daresbury Laboratory, Warrington (1996).

Holographic Methods in X-ray Crystallography.*

Abraham Szöke, Hanna Szöke
Lawrence Livermore National Laboratory, Livermore, CA 94550
and
John R. Somoza
Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, CA 94143

Abstract

The holographic method for finding the electron density of macromolecules is based on the expansion of the electron density into Gaussian basis functions. The computer program searches for the optimum set of such basis functions in physical space. Therefore it is capable of changing the phases of the structure factors in reciprocal space. The technique makes consistent use of real and reciprocal space information to produce electron density maps. It enforces positivity of the recovered electron density and makes effective use of prior knowledge about the electron density, such as that of a solvent region or of a partial structure. In this paper we summarize the theory underlying the holographic method and describe how we deal with multiple isomorphous replacement (MIR) data, multiple anomalous dispersion (MAD) data and knowledge of non-crystallographic symmetry. We discuss the convergence properties and the limiting accuracy of the method. We illustrate its power for synthetic problems and we apply the method to experimentally measured MIR data from kinesin, a recently solved motor protein domain.

1. Introduction

The most important limitation on the power of X-ray crystallography is that, as a consequence of Bragg's law, the electron density of a crystal cannot be fully recovered from its diffraction pattern alone. This is the well known phase problem. Many years ago the eminent mathematician Lánczos (1961) emphasized that no mathematical trickery can remedy lack of information. The holographic method does not attempt to circumvent Lánczos' dictum. Our principal claim is that the holographic method is a clear, simple and effective way of using all available information simultaneously, consistently, explicitly and sometimes even optimally.

First, we have to explain why our method is called holographic. Let us start from the analogy of an X-ray diffraction pattern and a hologram. We assume, maybe artificially, that we know the electron density in part of the unit cell of a crystal. This is the situation in molecular replacement and also during the solution of crystal structures. The complex amplitude of the wave diffracted from the known part can then be calculated and identified as a holographic reference wave. Similarly, the wave diffracted from the unknown part of the unit cell is analogous

to an object wave in holography. The pattern of intensities observed in X-ray diffraction from a crystal is then analogous to a recorded hologram. It contains the sum of the intensities of the waves scattered from the known and the unknown parts of the electron density of the crystal and also their interference. The interference term contains phase information that can in turn be used to find the unknown part of the electron density. In the language of holography the unknown wave can be reconstructed and its source can be found. We show below that the reconstruction reduces to a standard inverse problem, similar to those encountered in image processing. Our algorithm, built on the above observations, searches in real space for an electron density that minimizes the deviation of the magnitudes of the calculated structure factors from the measured ones. We have found that the ubiquitous holographic "dual image" also appears in X-ray crystallography, in fact it is equivalent to the phase problem of crystallography. Under favorable conditions, additional information can eliminate the dual image.

A practical algorithm for X-ray crystallography was developed by recognizing that the holographic method can use fast Fourier transforms and a conjugate gradient optimizer, that is capable of incorporating various constraints. The result was a suite of computer programs (*EDEN*, for Electron DENsity) for the solution of crystallographic problems of current interest. The main solver program runs in *PlogP* time, where *P* is the total number of resolution elements in the unit cell. Work stations (IBM 6000, HP 9000, SGI Iris, or equivalents) are adequate for treating realistic problems. Our progress was documented in five published papers [Szöke (1993, paper II), Maalouf *et al.*, (1993, paper III) and Somoza *et al.* (1995, paper IV), Szöke, Szöke & Somoza (1997, paper V) also Béran & Szöke (1995)]. *EDEN* is available free of charge to qualified collaborators. Please contact H. S. by e-mail at szoke2@llnl.gov.

2. Brief summary of the theory.

Unfortunately, our language is different from the standard one used by crystallographers. Nevertheless, we attempt to use accepted crystallographic notation wherever possible. More precise definitions can be found in paper II and in Appendix A of paper V.

The electron density in the unit cell of a crystal is divided into a known and an unknown part. The structure factors of the known part are denoted by $R(\mathbf{h})$. They are given by

$$R(\mathbf{h}) = \int_{\text{unit cell}} \rho_{known}(\mathbf{r}) \exp(2\pi i \mathbf{h}\cdot\mathcal{F}\mathbf{r})\, d\mathbf{r}. \qquad (2.1)$$

where $\rho_{known}(\mathbf{r})$ is the electron density of the known part and $\exp(2\pi i \mathbf{h}\cdot\mathcal{F}\mathbf{r})$ is just a fancy notation for $\exp[2\pi i(hx + ky + lz)]$.

Now we do something new. We make a three dimensional grid by dividing the unit cell into P_a, P_b, P_c equal parts along the crystallographic axes **a**, **b**, **c** respectively. The grid points are denoted by $\mathbf{r}_p$; $p = 1, \ldots, P$ where $P = P_a \cdot P_b \cdot P_c$. The

unknown part of the electron density is described as a sum of Gaussian blobs of equal widths, centered on the grid points. Each Gaussian basis function is assumed to contain an unknown number of electrons, n(p):

$$\rho_{unknown}(\mathbf{r}) \approx \frac{1}{(\pi\eta\Delta r^2)^{3/2}} \sum_{p=1}^{P} n(p) \exp\left[\frac{-|\mathbf{r}-\mathbf{r}_p|^2}{\eta\Delta r^2}\right], \qquad (2.2)$$

where Δr is the mean grid spacing and η determines the width of the Gaussians relative to the grid spacing. If the grid spacing is sufficiently fine, the electron density of the unknown part of the molecule can be well approximated by such a superposition of Gaussians. When (2.2) is extended periodically over the repetitions of the unit cell, the structure factors of the unknown part, $O(\mathbf{h})$, can be expressed as

$$O(\mathbf{h}) = \exp\left[-\eta(\pi\Delta r\,|\,\mathcal{F}^T\mathbf{h}\,|)^2\right] \sum_{p=1}^{P} n(p) \exp\left(2\pi i\mathbf{h}\cdot\mathcal{F}\mathbf{r}_p\right). \qquad (2.3)$$

The notation $R(\mathbf{h})$ for the structure factors of the known part and $O(\mathbf{h})$ for those of the unknown part of the structure is adopted from holographic theory, where $R(\mathbf{h})$ and $O(\mathbf{h})$ denote the reference and object wave, respectively. The squares of the absolute magnitudes of the structure factors of the crystal, $|F(\mathbf{h})|^2$, then satisfy the equations

$$|F(\mathbf{h})|^2 = |R(\mathbf{h}) + O(\mathbf{h})|^2 = |R(\mathbf{h})|^2 + R(\mathbf{h})\,O^*(\mathbf{h}) + R^*(\mathbf{h})\,O(\mathbf{h}) + |O(\mathbf{h})|^2. \qquad (2.4)$$

As promised, the measured intensities, that are proportional to $|F(\mathbf{h})|^2$, are the sum of the diffracted intensity of the known part, $|R(\mathbf{h})|^2$, the diffracted intensity of the unknown part, $|O(\mathbf{h})|^2$ and the interference terms, $R(\mathbf{h})\,O^*(\mathbf{h}) + R^*(\mathbf{h})\,O(\mathbf{h})$.

When the representation of the unknown density is substituted from (2.3), equation (2.4) becomes a set of quadratic equations in the unknowns, n(p). The number of equations, N_h, is usually not equal to the number of unknowns, P. The equations may contain inconsistent information, e.g. due to experimental errors, or lack of isomorphism in MIR, or incomplete non-crystallographic symmetry. The equations are also ill conditioned and therefore their solutions are extremely sensitive to noise in the data. Under these conditions the equations may have many solutions or no solution at all. One way mathematicians deal with these problems is by minimizing a cost function that measures the discrepancy between the two sides of (2.4). We define such a cost function as

$$f_{eden} = \frac{1}{2} \sum_{\mathbf{h}} w'(\mathbf{h})^2 \left[\,|R'(\mathbf{h}) + O(\mathbf{h})| - |F'(\mathbf{h})|\,\right]^2, \qquad (2.5)$$

where $R'(\mathbf{h})$ and $F'(\mathbf{h})$ are smeared (apodized) versions of $R(\mathbf{h})$ and $F(\mathbf{h})$ and $w'(\mathbf{h})^2$ are weights. Unless the structure factors are appropriately smeared (apodized), the Gaussian basis functions in (2.3) are not able to fit the experimental data. This shows up as a vicious(!) numerical instability in the solution. Let us be even more blunt: such numerical instabilities are inherent to inverse problems, of which the solution of crystal structures is an example; they do not depend on the

representation chosen. They cause arbitrariness in the structures at the high resolution end of the data. In other words, no matter how well you measure your diffraction intensities, at some resolution your structure depends almost entirely on the structure you postulate during refinement. We attempt to make such arbitrariness explicit and eliminate it if possible.

The summation in the cost function (2.5) includes only available experimental data, i.e. we do not include values of R(**h**) for which the corresponding F(**h**) are missing. Thus the cost function (2.5) does not make unwarranted assumptions about unobserved reflections: their values are indeterminate, as they should be. Therefore truncation errors of Fourier inversions are absent.

The solution of equation (2.5) is not unique; this is an expression of the well known crystallographic phase problem. A simple geometric representation of this lack of uniqueness is shown in Fig. 1. R(**h**) is the vector representing the structure factor of the known part of the electron density. The circle around the origin has the radius |F(**h**)|. Since the phase of |F(**h**)| is unknown, any O(**h**) that connects the tip of R(**h**) to any point on the circle with radius |F(**h**)| satisfies equation (2.4). However, additional information in the form of constraints reduces the arbitrariness of the solution. The unweighted difference Fourier solution is $O_2(\mathbf{h})$. If the "correct" solution is $O_1(\mathbf{h})$, the "dual image" is represented by $O_3(\mathbf{h})$.

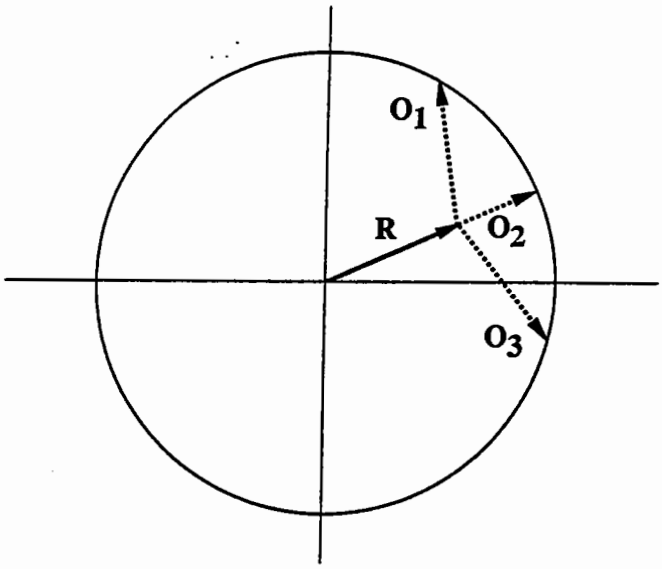


Figure 1. Geometric representation of equation (2.4) in the complex plane for acentric reflections.

A second type of information on the crystal structure is (possibly imperfect) knowledge of the electron density in parts of the unit cell. For example, part of the molecule may be very similar to another molecule whose structure is known. As another example, the solvent volume has a featureless electron density at a well known value. Such knowledge can be incorporated into *EDEN* as a "target" density, expressed in terms of the amplitudes of the basis functions used in the main program. They will be denoted by $n(p)_{target}$. We introduce a corresponding cost function

$$f_{space} = \frac{1}{2}\lambda_{space} P \sum_{p=1}^{P} \tilde{w}_p^2 \{n(p) - n(p)_{target}\}^2. \qquad (2.6)$$

The overall relative weight, λ_{space}, and the individual weights at each point, $\tilde{w}_p^2 \leq 1$, express the "strength of our belief" in the correctness of the target density: the weights may be used to emphasize or de-emphasize different regions of the unit cell, while the overall weight determines the relative importance of deviations in space

vs. deviations in reciprocal space. In the presence of a target density, the actual cost function used in the computer program is the sum of f_{eden} (2.5) and f_{space} (2.6):

$$f_{total} = f_{eden} + f_{space}. \tag{2.7}$$

It can be seen from Eq. (2.3) that the structure factors of the unknown part can be calculated by fast Fourier transforms followed by a scalar multiplication. The gradients of the cost function can be calculated similarly. This leads to a fast (PlogP) algorithm and to the ability to minimize Eq. (2.7) without ever calculating or saving PxP matrices. In *EDEN* the cost function is minimized using a conjugate gradient algorithm that is very efficient in the presence of non linear constraints.

A basic constraint, non-negativity of the electron density, is incorporated directly into the conjugate gradient optimizer by stipulating that all elements of the solution vector, n(p), be non-negative. The constraint can be used in two different ways: in "correction mode" n(p) is bounded from below by the negative of the known initial electron density and in "completion mode" the added density itself is non-negative everywhere. Additional information leads to additional terms in the cost function (2.7).

The representation of the unknown density, presented in Eq. (2.2), uses an overcomplete set of Gaussian basis functions that are not orthogonal to each other. Mathematicians have done extensive research on such non-orthogonal, redundant basis sets: they are called frames. Excellent discussions can be found in a book by Daubechies (1992) and in a review by Heil & Walnut (1989). Some of their important results are summarized below. The mathematicians assure us that electron densities can be approximated well by such representations, if the electron density does not vary too wildly. Restated in technical language, the requirements are that the diffraction pattern and the basis set should have similar intrinsic resolutions, and that the grid spacing should be about twice as fine as required by the corresponding Nyquist criterion. In our algorithm this is achieved by the appropriate choices of η and Δr in Eq. (2.2). Although a given electron density can be represented by several different sets of coefficients, the algorithm used by *EDEN* is mathematically stable.

We tested the accuracy of our representation by placing a single Gaussian blob onto an arbitrary point on the grid and trying to represent it by our Gaussians on the grid. We found that if we use a simple grid the maximum phase error was 47° and the corresponding amplitude error was 28%. Using a body centered grid the maximum phase error was 26° and the corresponding amplitude error was 20%. The phase error for a complicated molecule that is uniformly distributed in the unit cell is expected to be less.

In our next set of tests we recovered parts of a model of Thaumatin, a protein with 207 residues, with no noise or solvent. Without a solvent mask 70 consecutive residues could be recovered. We used the values Δr = 1.8Å and η = 0.28 for this test. When a hard solvent mask that covered half the unit cell was imposed, as many as 160 consecutive residues out of the 207 (or 77%) were found essentially perfectly. The recovered electron density was within 10% of that of the original model and the phases were accurate to 25°. In the next test we used Δr = 1.4Å and η = 0.75 that

corresponds to an input resolution of 2.0Å. The phase accuracy of the recovery is very good, better than 10°. The convergence of the algorithm without a solvent mask is also better: in fact 90 out of 207 residues (i.e. 43%) were recovered essentially perfectly. With a solvent mask or a solvent target function we got perfect recovery only up to 120 residues. This is a respectable 58%, but it is less good then using the previous values of Δr and η. We do not understand the reasons for the difference.

Some additional tests show the power of the positivity constraint in model problems that have no noise or solvent. *EDEN* solved the model of Staphylococcal Nuclease at 3Å resolution using a low resolution (≈ 6Å) solvent target that covered 61% of the unit cell and no other information. A similar result was reported by Bricogne (1993) using a very sophisticated algorithm in reciprocal space. We recall that Béran & Szöke (1995) found that the phases of the structure factors of a model protein could be recovered completely when the electron density was given in a little more than half the unit cell. The above results seem to contradict recent conclusions of Millane (1996). We interpret Millane's conclusion as establishing an upper limit to the additional information needed to solve the crystal structure. In our opinion he does not exclude the possibility of solving the structure with less information.

3. Multiple Isomorphous Replacement and Multiple Anomalous Dispersion.

3.1. *Derivation of equations for MIR* .

Crystal structures can be solved by multiple isomorphous replacement (MIR) if the only change in crystal structure is the addition of heavy atoms. MIR methods require that individual data sets be taken for each derivative and that the positions of the heavy atoms and their occupancies be found by Patterson or direct methods. Conventional MIR methods then proceed to find the phases of the native protein. Very often the resulting phase set does not give electron density maps that are easily interpretable. This is the stage where the holographic method can be of advantage. In principle, the holographic method is equivalent to the conventional method of finding the phases of the structure factors of the native protein. In practice, the convergence of *EDEN* for ab initio phasing is worse than that of traditional programs (*PHASES* or *MLPHARE*). On the other hand, we expect that a consistent use all known constraints should improve the attainable accuracy of the solution. In a test case using real data, *EDEN* resulted in a clear improvement over conventional methods.

From a mathematical point of view, heavy atom derivatives (as well as anomalous dispersion) increase the number of independent equations with respect to the unknowns. With a sufficient number of derivatives the phase problem should therefore be solvable. The relevant equations in *EDEN* are simple generalizations of (2.1) - (2.5). The unknown density (2.2) is that of the native protein. So is the structure factor, $O(\mathbf{h})$, of equation (2.3). Suppose M + 1 sets of diffraction amplitudes have been measured: one for the native and one each for the M derivatives. Suppose also that the positions of the heavy atoms and their occupancies were found, using Patterson or direct methods. The calculated structure

factors for the heavy atoms then belong to the known part of the structures. For the m'th derivative they will be designated $R_m(\mathbf{h})$. The measured structure factors of the m'th derivative will be denoted $|F_m(\mathbf{h})|$. Then equation (2.4) can be generalized to the set

$$|F_m(\mathbf{h})|^2 = |R_m(\mathbf{h}) + O(\mathbf{h})|^2 = |R_m(\mathbf{h})|^2 + R_m(\mathbf{h})\,O^*(\mathbf{h}) + R_m^*(\mathbf{h})\,O(\mathbf{h}) + |O(\mathbf{h})|^2, \quad (3.1)$$

where m = 0, ... , M (m=0 designating the native protein). It can be demonstrated that, Eqs. (3.1) are equivalent to traditional MIR algorithms (the simple minded ones, see e.g. Giacovazzo 1992). Equations (3.1) are solved by minimizing a cost function that is analogous to (2.5)

$$f_{eden} = \frac{1}{2} \sum_{m=0}^{M} \lambda_m \sum_{\mathbf{h}} w'(\mathbf{h})^2 \,[\,|F_m(\mathbf{h})| - |R'_m(\mathbf{h}) + O(\mathbf{h})|\,]^2. \quad (3.2)$$

In equation (3.2) we introduced weights, λ_m, that can express the reliability or quality of the measurements of each derivative.

3.2. *Equations for MAD.*

Multiple anomalous dispersion (MAD) can be treated very similarly to MIR. As MAD data sets are taken on a single crystal, the basic assumption of isomorphism is correct; the main problem with the method is usually the low signal to noise ratio. The fundamental assumption in *EDEN*'s treatment of MAD is that the structure amplitudes of the unknown part (that will be called the native) have no anomalous dispersion, i.e. f'' for all the unknown atoms is zero and their f' is independent of X-ray energy. In other words, the anomalously scattering atoms are always considered to be "heavy atoms". We will start from the point where the anomalously scattering (heavy) atoms have been found by Patterson methods or by direct methods and their structure factors, including the anomalous part, have been calculated. In a P1 crystal the $h \geq 0$ data set can now be treated exactly as a derivative in MIR. In P1 symmetry the $h \leq 0$ reflections are an independent data set. The easiest way to use them in *EDEN* is to create a "flipped" data set by negating all the indices of the reflections, $\mathbf{h} \rightarrow -\mathbf{h}$, at the same time flipping the signs of the phases of the heavy atoms and declaring this new data set to be a separate derivative. Friedel's relations apply to the structure factors of the native because that part of the structure has no anomalous dispersion. Therefore the unknowns in this "flipped" data set are the same as those for the $h \geq 0$ data set. In higher symmetry similar considerations apply. Equations analogous to (3.1) are now defined and solved by minimizing the cost function that is analogous to (3.2). It is clear that such data sets can be used together with MIR data. The only difficulty one might encounter in this procedure is that the anomalous data sets are weighted too heavily. The procedure is able to solve problems in which there is no native data set, by setting λ_o to zero.

3.3. *MIR results using data from Kinesin.*

To test the effectiveness of the MIR algorithms using real data, we studied the protein kinesin. Kinesin is a microtubule "motor" protein that functions in intracellular transport and chromosome movement. The data that were used for

our tests were collected from a 349-residue piece of the protein that encompasses the motor head. The structure of the kinesin head domain was solved by Kull et al. (1996). The original MIR maps were fairly poor, suggesting that they may be improved using the holographic method. Native data to 1.8 Å were available for this protein, as well as data collected from two derivatives, one containing one iodine atom and one containing three mercury atoms. The data for each derivative extended to 2.5Å.

The *EDEN* implementation of the MIR algorithm suffers from convergence problems if the starting phases are too far from the correct solution. To circumvent this problem, we started from the *MLPHARE* estimate of the phases. The native data were placed on an absolute scale with a Wilson plot program from *CCP4* using data between 3.0 Å and 1.8 Å resolution. The derivatives were scaled to the native dataset and the total number of electrons in the unit cell was estimated.

Our first step was to check the occupancies and positions of the heavy atoms. To do this, we worked at a resolution of 3.0 Å. The initial MIR phase set was used to prepare a corresponding electron density map. *EDEN* was run at 3.0 Å resolution in correction mode, and the resulting electron density maps were visually inspected to see if there were either peaks or holes at the heavy atom positions. Ideally, there should be no evidence of the heavy atoms in the resulting native electron density. If they do, either the occupancies of the heavy atoms or the scaling are wrong. By repeatedly running *EDEN* and inspecting the results we adjusted the occupancies of the heavy atoms and made slight adjustments to the relative scaling of the derivative and the native datasets.

Preliminary *EDEN* runs were done at 3.0 Å resolution. The results were quite encouraging. At this point the isomorphism of the derivatives was checked. In order to do that we started from the MIR map and ran it in correction mode against the measured structure factors of the native alone. This way the program is not constrained by any of the derivatives. The same procedure was done with each one of the derivatives. Pairwise comparisons of the results should reveal lack of isomorphism and local distortions around the heavy atoms. We found that, within our ability to detect differences, the two derivatives of kinesin were isomorphous with the native.

The next step was to obtain an estimate of the solvent envelope. This was done by apodizing the output of the previous 3.0Å MIR run to 7.0Å. The appropriate *EDEN* utility was used to select the 50% of the grid points with lowest electron density. These were used as the solvent region, and assigned a target electron density of 0.33 electrons/Å^3.

Two *EDEN* runs were done at 3.0 Å resolution using the solvent target and the two derivatives with λ_{space} = 0.003 and 0.01. The results were very encouraging, and we used the same solvent target to do a full *EDEN* run at 2.0 Å resolution. The resulting electron density map was compared with that obtained from the original phases derived from *MLPHARE*, and with a *DM* modified map (Cowtan & Main, 1993). The fully refined kinesin structure was used as a guide for comparing the maps. The *EDEN* map was comparable to the *DM* map everywhere and in some places it was clearly better.

The reader may have noticed that the kinesin work involved a large number of preliminary runs. This is typical of the holographic method; whereas a particular *EDEN* run takes only 1-3 hours on a standard workstation, many such runs are needed to establish optimal values of critical parameters (λ's, smearing factors and scaling factors). In particular, Eq. 2.4 shows that data and models must be carefully scaled in order that the cost function minimization be meaningful.

4. Non crystallographic symmetry.

Non crystallographic symmetry (NCS) is treated in a manner similar to previous sections. We use a real space cost function that "encourages" the symmetry but does not enforce it. Although our method has similarities to well established and successful methods of NCS there are also differences. Some of these differences are advantageous, at least in theory. First, the exact knowledge of the molecular envelope is not critical. Second, the non crystallographic constraint is "soft" and its strength can be varied. Third, we do not interpolate in reciprocal space; instead we use an expansion into basis functions in physical space. However, this is not an important distinction from other methods. These properties of the method allow the determination of the goodness of the symmetry from the data alone. One should also be able to find out if there are differences in the monomers that are related by non crystallographic symmetry. The main disadvantage of the method is that it uses basis functions on a grid and therefore it has limited accuracy. The mathematical derivation of the cost function has been reported in paper V. The NCS option has not been tested on any realistic problem yet.

5. Summary and Discussion.

We would like to address the question: how similar and how different is the holographic method from other, well established methods of X-ray crystallography? In other words, why do we bother and should you bother?

Our approach can be described as a real space method, based on an expansion of the electron density in basis functions and on a search for the number of electrons in each one of the basis functions. It is one more step removed from reciprocal space methods then other density modification methods. Note that we almost never refer to the phases of the structure factors.

In its simplest form, the holographic method can be used to complete a partly known structure. If there are no external constraints, the electron density maps obtained using the holographic method are very similar to traditional Fo - Fc and 2Fo - Fc maps. Traditional Fourier maps are actually marginally more accurate, because the holographic method is limited in its accuracy by the (incomplete) basis function expansion. However, if there are known constraints that must be satisfied by the electron density, the holographic method is able to use that information to recover electrons more accurately than traditional Fourier methods.

The fact that the electron density is always positive is an important constraint; positivity is always enforced in *EDEN*. In addition, often the electron density is known in part of the unit cell, either because the solvent region is known, or because a partial structure has been placed in the unit cell. *EDEN* is able to use the localized nature of the known electron density in real space: it can constrain it in some part of the unit cell and not in other parts. It can also use the known electron density as a mild constraint. Therefore errors in the "known" part can be both detected and corrected.

We have shown that MIR, MAD, and NCS information can be incorporated into the holographic method. Using simulated heavy atom data, we have explored in some detail the convergence of our algorithm and the uniqueness of the solution it supplies. These simulations show that the holographic method does not converge as well as traditional reciprocal space methods, even though the equations are mathematically equivalent. However, once the electron density is within the radius of convergence of the correct minimum, the holographic method quickly and accurately finds the correct structure. Given these findings, we propose that conventional methods should be used to identify an initial MIR solution, and that the holographic method should then be able to improve that solution. We have made use of this strategy to determine the structure of the protein kinesin, using experimental MIR data. An initial structure of kinesin was identified using the program *MLPHARE*. Using *EDEN* to optimize this solution led to a clear improvement in the resulting electron density maps.

On the theoretical side, we scrupulously differentiate between lack of information and tacitly assumed information. For example we consistently avoid the use of Fourier back transforms. In usual practice, unknown structure factors are given zero value as opposed to keeping them unknown. Similarly, in the presence of non crystallographic symmetry, some formulations implicitly assume that the electron density is featureless outside the symmetry related regions. We try to live by Lánczos' dictum: use all the available information and no more. In principle, given a sufficient amount of information it is possible to recover the crystal perfectly. However, different algorithms may have very different convergence properties and may have very different sensitivity to imperfections in the data. In our opinion, this last point alone is sufficiently important to justify the development of new methods for crystallographic computations.

Finally, we want to give a preview of coming attractions. In the near future we intend to extend the variety of real space target functions. We will then be able to treat molecular replacement and resolution extension (usually called phase extension) conveniently. Following David & Subbiah (1994), we will also try to solve proteins ab initio at low resolution using our algorithm. We have already found that the holographic method converges well in the presence of low resolution information. At high resolution we will incorporate atomicity, i.e. our own version of Sayre's equation. Both the low resolution and the high resolution information can be cast in a very transparent and clean form. In reciprocal space, we will use $1/\sigma^2$ weights for the measured structure factors, in order to take into account inaccuracies of measurements. We will also incorporate some statistical information: The probability distribution of the measured structure factors can be

estimated more accurately than given by the single number σ^2. Similarly, the probability distribution of the unmeasured structure factors can be estimated. If the number of missing atoms is known, the probability distribution of the missing part $O(\mathbf{h})$ can be estimated. The most exciting (and difficult) development is the incorporation of some chemical knowledge into the holographic method. It has the following ingredients: Kleywegt has shown how to fit small parts of proteins into the electron density. We intend to use a variant of his method. Also Fortier has shown how to find extremal points of the electron density and how to connect them to find the protein backbone. We also intend to use a variant of her method. Finally, we intend to put in a mock-electrostatic force in order to improve the electron density and to impose chemical constraints on it. We hope that our efforts will be a beginning of "automated" crystallographic refinement.

References

* Part of this work was performed under the auspices of the US. Department of Energy under Contract No. W-7405-ENG.-48.

Béran, P. & Szöke, A. (1995). *Acta Cryst.* A.**51**, 20-27.

Bricogne, G. (1993). *Acta Cryst.* D**49**, 37-60.

Bricogne, G. (1992). *International Tables for Crystallography. Vol. B.* edited by U. Shmueli. Dordrecht: Kluwer Academic Publishers.

Cowtan, K. D. & Main, P. (1993). *Acta Cryst.* D**49**, 148-157.

Daubechies, I. (1992). *Ten Lectures on Wavelets.* Philadelphia, PA: SIAM.

David, P. R. & Subbiah, S. (1994). *Acta Cryst.* D**50**, 132-138.

Giacovazzo, C. (1992). editor, *Fundamentals of Crystallography.* Oxford: IUCr, Oxford University Press.

Heil, C. E. & Walnut, D. F. (1989). *SIAM Review* **31**, 628-666.

Kull, F. J., Sablin, E. P., Lau, R., Fletterick, R. J. & Vale, R. D. (1996). *Nature* **380**, 550-555.

Lánczos, C. (1961). *Linear Differential Operators.* London: Van Nostrand.

Maalouf, G. J., Hoch, J. C., Stern, A. S., Szöke, H. & Szöke, A. (1993). *Acta Cryst.* A**49**, 866-871.

Millane, R. P. (1996). *J. Opt. Soc. Am.* A**13**, 725-734.

Somoza, J. R., Szöke, H., Goodman, D. M., Béran, P., Truckses, D., Kim, S-H. & Szöke, A. (1995). *Acta Cryst.* A.**51**, 691-708.

Szöke, A. (1993). *Acta Cryst.* A**49**, 853-866.

Szöke, A., Szöke, H. & Somoza, J. R. (1997). *Acta Cryst.* A.**53**, To be published.

The world according to wARP:
improvement and extension of crystallographic phases

Anastassis Perrakis[1], Titia K. Sixma[1], Keith S. Wilson [2] and Victor S. Lamzin[3]

1. Netherlands Cancer Institute (NKI), Department of Molecular Carcinogenesis, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
2. Protein Structure Group, Dept. of Chemistry, University of York, Heslington, York YO1 5DD, UK
3. European Molecular Biology Laboratory (EMBL) Hamburg, c/o DESY, Notkestrasse 85, 22603 Hamburg, Germany

Abstract

We have developed procedures for the improvement of crystallographic phases resulting either from the position of a heavy atom within the native molecule, or from a multiple isomorphous replacement experiment.

In the first case the position of a heavy atom as located from native Patterson maps is used as a starting model for least squares or maximum likelihood refinement and iterative model updating in an ARP procedure. Automatic update and completion of the model by ARP, results to maps of excellent quality. Furthermore, the atomic positions of the final ARP model are very accurate and can be used to initiate automatic model building techniques, currently under development.

For the second case, the best initial map is used to construct a number of dummy free atom models which are subjected to ARP refinement. Averaging of the phase sets calculated from the refined models and weighting of structure factors by their similarity to an average vector, results in a phase set that improves and extends the initial phases if the native data set has sufficiently high resolution (beyond ~ 2.4 Å). This procedure allows shortening of the time-consuming step of model building in a lot of crystallographic structure solutions.

NOTE: The ARP program is freely available as part of the CCP4 package. C-shell scripts and the actual averaging program, are available to run wARP. They perform the dummy model building, ARP refinements and final averaging in an automated manner. They are also capable to split jobs in a 'parallel' manner to different processors which can be located in different computers over a network, thus minimizing the actual required run time to the one needed for a single ARP job - provided that enough processors are available. The scripts are tested on several Irix 5.3 based clusters, but should be straight-forward to adapt for usage with any Unix based system. A WWW ARP/wARP home page is now available, at http://den.nki.nl/~perrakis/arp.html from where the complete ARP/wARP package can be obtained. A mailing list is also open for questions and discussion for ARP/wARP usage. To subscribe, simply do it through the WWW page or send a mail with one line 'subscribe arp-users' to majordomo@linde.nki.nl.

Outline of the (w)ARP method

ARP from single heavy atom model

One single or a few heavy atoms located from the native Patterson synthesis are used as an initial model. Starting from these atoms, a model consisting of only oxygen atoms is slowly created by ARP. This model consists of free atoms that are not subjected to any kind of restraint. One ARP refinement cycle has two parts: (1) unrestrained least-squares minimization or maximum likelihood refinement in reciprocal space, to properly match calculated to observed structure factor amplitudes and (2) substantial modification of the current atomic dummy model in real space, using ARP [1,2]. For the unrestrained refinement step, C-shell scripts have been constructed to employ most currently available programs in the procedure. Standard protocols include PROLSQ [3] and REFMAC [4] from the CCP4 [5] suite. ARP, after each reciprocal space refinement cycle, updates the model mimicking human intervention between refinement cycles. It removes atoms based on the density in the $3F_o$-$2F_c$ Fourier synthesis and adds atoms in significant density in the F_o-F_c Fourier synthesis, provided that they are bonded to existing atoms. After several such cycles of ARP, the atoms that are added gradually constitute a model that resembles the protein to a great extend.

From ARP to wARP

The procedure described above requires data to very high resolution to be available and a heavy atom present in the native protein. In most crystallographic projects, however, this is not the case. Since it is very hard if at all possible to provide an *ab initio* solution to the phase problem in such cases, our effort has been concentrated on improving phases that are available by experimental techniques. Such phase information can be very inaccurate and means of improvement will speed up the efficiency and the quality of model building. ARP needs high resolution data to converge to global minimum during refinement. If such data are not available, the refinement will most likely not converge and inaccuracies are introduced to the 'final' model. With wARP we try to overcome this problem by the weighted averaging of structure factors from individual models.

ARP from MIR maps: wARP

The first step in the wARP procedure is the creation of moderately different free atom models in the best available map. The procedure for building a 'dummy model' is then invoked as described in the ARP manual. Briefly, starting from a small set of dummy atoms placed anywhere in the protein region, a model is slowly expanded by the stepwise addition of atoms that are in bonding distances with existing atoms and in significant density in the electron density map exists for their placement. Six such models are created, using slightly different ARP building protocols, which are used for all subsequent steps. Next, these models are subjected to ARP refinement. Due to the limited amount of diffraction data, they will presumably at the end contain different errors, which by the averaging procedure will be canceled out.

Structure factors are calculated for all models after refinement and scaled to observed amplitudes. A vector average of the calculated structure factors from the different refined models is then calculated. The phase of the vector average is remarkably better than those calculated from any of the individual runs. Subsequently, a weighting scheme is applied to enhance the overall quality of phases, depending on the variance of the individual structure factors around the average.

Examples

ARP from single heavy atom model

Rubredoxin

Rubredoxin is a small protein of 51 residues, the first protein to be refined to atomic resolution [6]. It contains a Fe atom which is coordinated by 4 Cys residues. The position of the Fe atom and the 4 sulfurs of the cysteines side chains can be located from the native Patterson map, if data better than 1.5 Å resolution are available. A high resolution data set of rubredoxin (0.92 Å) was used. In all cases that lower resolution is quoted, that means that the data were simply truncated at that resolution limit.

The starting model for the ARP refinement procedure was initially the Fe atom with the coordinates calculated from the native Patterson map. After 80 cycles of least squares refinement, or 30 cycles of maximum likelihood refinement a complete model was available. The map correlation coefficient [7] improved from only 26 % to more than 90 % in both cases. The lowest resolution at which the method works, starting from the Fe atom alone, is 1.1 Å. However, if the positions for the four sulfur atoms are included, the method can work with 1.4 Å data, in other words with less than one third of observed reflections at 0.92 Å. In that case, the correlation coefficient with the final map is 96%, because the correct atomic types are used for the four sulfur atoms. If we try to use data to 1.5 Å resolution, there is a small increase in correlation coefficient to 45%, but after that no improvement could be achieved. Protocols involving the use of E-maps and the wARP averaging described below are being tested to extend the use of method to much lower maximal resolution. It is of interest to note, that the atomic positions of atoms placed by ARP are these of atoms in the final model, with slight variation, Figure 1. It would be thus feasible to use them to initiate automatic model building techniques to minimize the amount of time spent in traditional model building and the errors introduced by this procedure. Characteristic parts of the maps before and after the ARP procedure are shown in Figures 2 and 3.

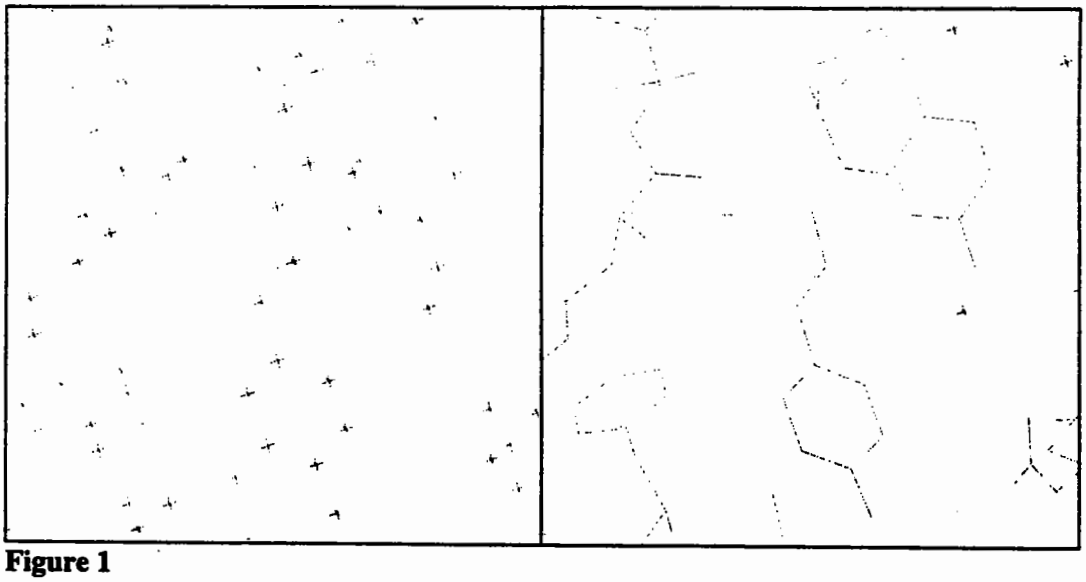


Figure 1

Positions of the ARP atoms (left) and of the atoms in the final model (right), for a representative region of the protein. You are welcome to try the 'join the correct dots' game in the left panel of the figure ...

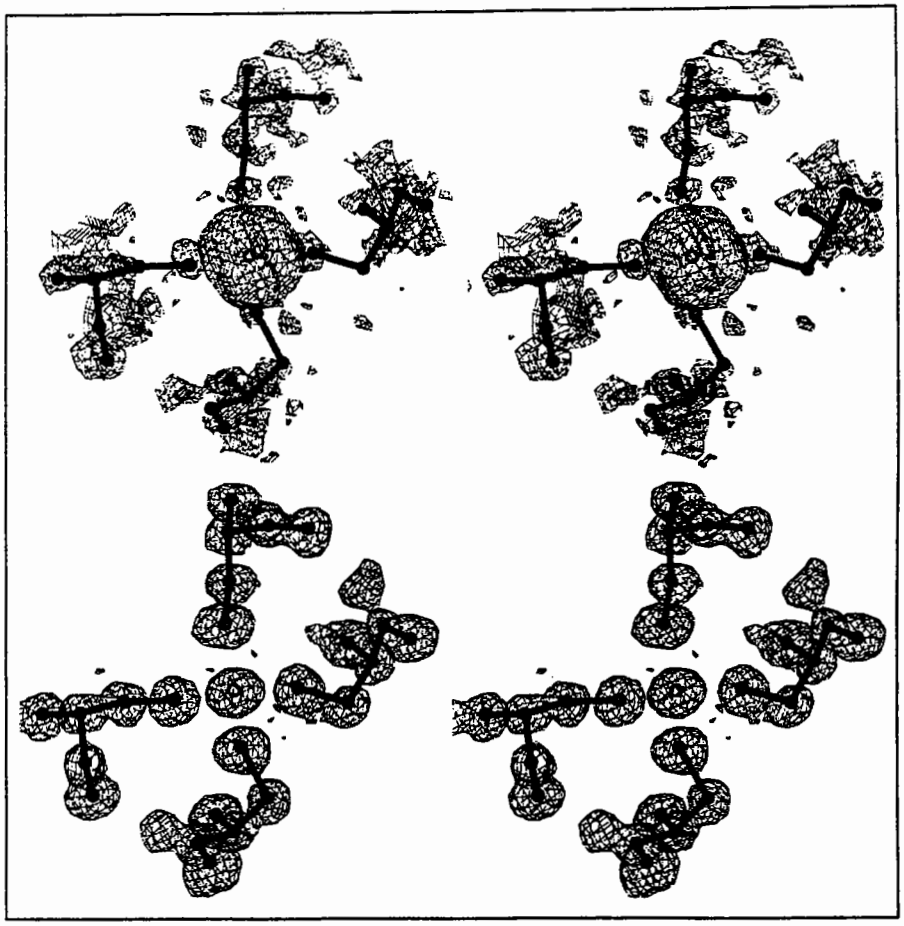


Figure 2

Stereo figures of the area of the map around the starting Fe atom. In the initial map (top) resulting from the phases calculating from Fe atom position alone, a big bolb of density is representing the ion. Although there is density for the four sulfurs of the cysteines coordinating the Fe ion, it is hardly interpretable. After ARP the atomic positions are clearly visible and the map is of excellent quality.

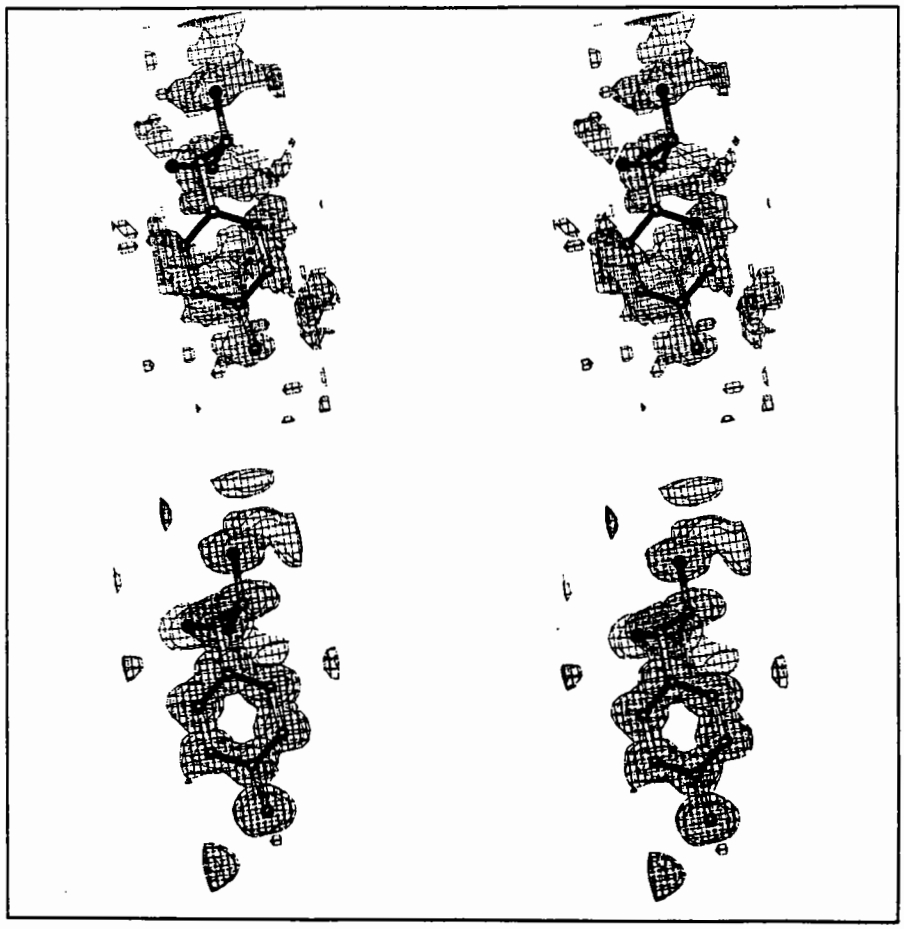


Figure 3

Stereo figures of one area of the map far from the starting Fe atom. In the initial map (top), although density for some of the Tyr atoms is present, the Tyr residue is in practice not recognizable. After ARP refinement the Tyr main and side chains are clearly recognizable.

ARP from MIR maps: wARP

Leishmanolysin

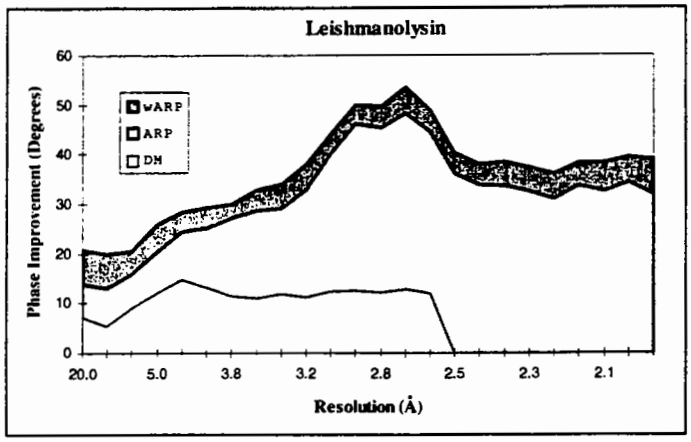


Figure 4

Phase improvement provided by DM, ARP and wARP in resolution shells is shown for phase sets from MIR, optimal solvent flattening, the best single ARP run and the wARP phase combination, for leishmanolysin. Map correlation coefficient for electron density maps resulting from these datasets were 42.7 %, 66.6 %, 88.3 % and 92.0 % for MIR, DM, ARP and wARP maps respectively.

The structure of the *Leishmania* virus coat protein (Leishmanolysin, PSP) was solved with a complicated protocol involving the use of SIRAS phases for two different crystal forms, averaging between those, solvent flattening and density skeletonization (unpublished data were kindly provided by Dr. Peter Metcalf). For the wARP test one set of SIRAS phases was used, which extends to a resolution of 3.0 Å. These phases were determined for the first crystal form for which native data extending to 2.5 Å were used for solvent flattening and phase extension with the DM program [16], CCP4. This density modified map was used to build the initial models with ARP. The ARP unrestrained refinement was performed against a higher resolution native data set from a frozen crystal (2.0 Å). REFMAC maximum likelihood minimization was used with ARP. All of these models gave maps of dramatically better quality than the solvent flattened map. Here the power of ARP procedure itself is large because the resolution of the native data is good. The wARP procedure resulted in a small but significant additional improvement. Statistics on phase improvement are in Figure 4 and a representative part of the map at Figure 5.

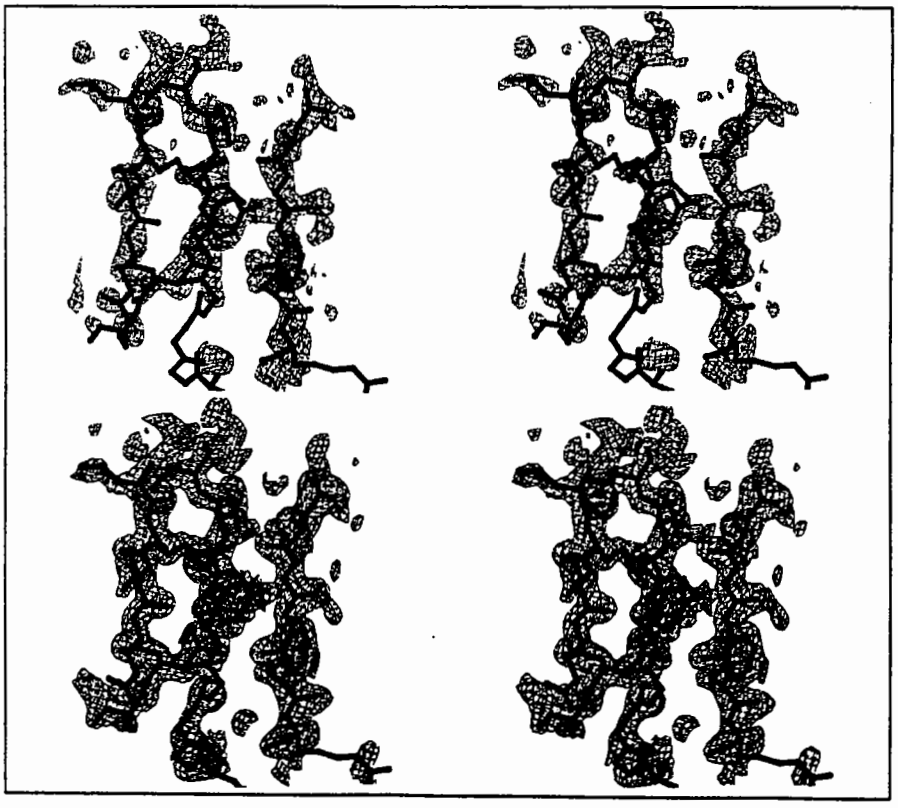


Figure 5

Representative regions of the solvent flattened (a,c) and equivalent wARP averaged maps (b,d) for Leishmanolysin, shown in stereo.

Chitinase A

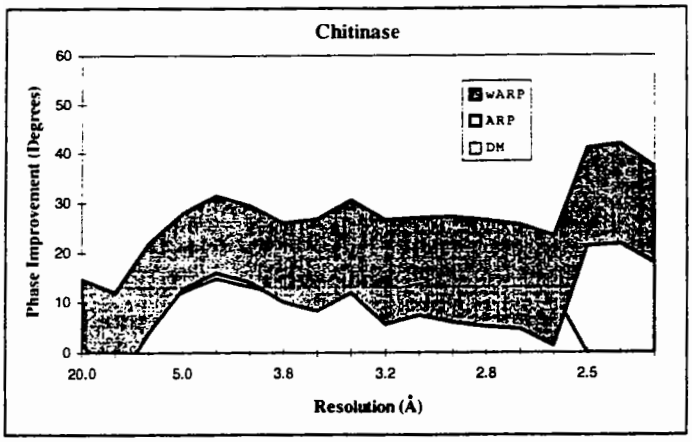


Figure 6

Phase improvement provided by DM, ARP and wARP is shown for phase sets from MIR, optimal solvent flattening, the best single ARP run and the wARP phase combination, for chitinase. The map correlation coefficient for electron density maps resulting from these datasets were 42.7 %, 66.6 %, 88.3 % and 92.0 % for MIR, DM, ARP and wARP maps respectively

The chitinase A structure from *Serratia marsescens* (ChiA) was initially solved by MIRAS [8]; with one only derivative contributing to better resolution than 5.0 Å [ref]. The MIRAS map (2.5 Å) was solvent flattened with the procedures implemented in the PHASES package [15]. Model building was not straightforward and much time was spent in tracing the protein chain. In the wARP procedure the solvent flattened map was used to initiate building of dummy models. PROLSQ least squares minimization against the native 2.3 Å data was used with ARP. Refinement of the models resulted in crystallographic R factors ranging between 20.1 % and 22.4 %. All of the ARP refined models gave phases same or worse than the phases already available by solvent flattening, due to the limited resolution of the native data, Figure 1a. At that case, where limited resolution if the data prevent convergence of the refinement, the wARP averaging procedure results in a much further improved map, comparable to the improvement achievable with higher resolution data. The phase improvement in resolution shells, for all phase sets, is analytically shown in Figure 6. A characteristic region of the map is shown in Figure 7.

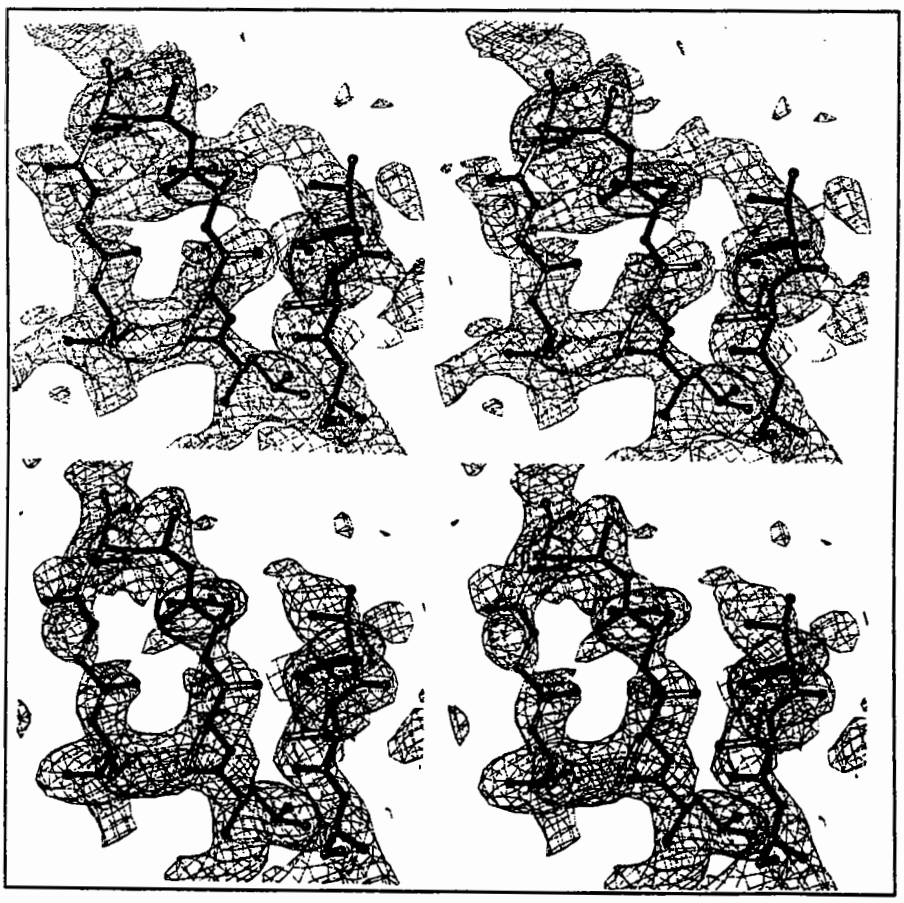


Figure 7

Representative regions of the solvent flattened (a,c) and equivalent wARP averaged maps (b,d) for Chitinase A, shown in stereo.

Applicability and requirements

ARP for ab initio structure solution: Capabilities and limitations

Ab initio methods in protein crystallography have only recently been successfully applied, the most characteristic examples being the structure solution of crambin [9] by direct methods and cytochrome c6 [10] by Patterson expansion methods. The limitation for succesful application of *ab initio* methods is the resolution of the diffraction data. Allthouh our current example, rubredoxin, can be solved easily by any relevant procedure if atomic resolution data are available, these methods fail if data worse than ~ 1.2 Å are available. With ARP we managed to produce an excellent map and an atomic model, with only 1.4 Å data, ie with essentially ~ 60 % of the reflections. Many more proteins diffract to resolution around 1.5 Å than 1.2 Å, according to the data on projects recently collected at EMBL Hamburg synchrotron X-rays facilities. Furthermore, we believe that we will be able to extend that limit in the near future, possibly with the application of wARP averaging.

Resolution requirements and use of different refinement methods for wARP

In contrast to most density modification methods the wARP procedure is extremely sensitive to the resolution of observed data in the native dataset. This is due to the limitations of the unrestrained refinement step, which requires that the observations/parameters ratio is more than 1.5 for convergence to a minimum. It is crucial to realise, that the real limitation can not be expressed solely in resolution terms, but better as observations/parameters ratio, which is largely dependent on solvent content. Thus, for a crystal with high solvent content 2.5 Å data will be sufficient while for a crystal with low solvent content data to 2.0 Å resolution must be available. Obviously the collected data must be of good quality, as can be judged by R_{merge}, $I/\sigma(I)$, and completeness. The success of refinement can be easily assessed by the crystallographic R factor.

Our experience shows that if the ratio of the number of reflections in the dataset to refined atomic parameters (four parameters per atom, x,y,z,B) is more than 2.0 (resolution ~ 2.0 Å) then use of maximum likelihood refinement as implemented in REFMAC can be used very effectively, as shown in Leishmanolysin. If the observations to parameters ratio drops below 2.0 traditional least squares refinement as implemented in PROLSQ produce better results, as shown for ChiA. When the observations to parameters ratio drops below 1.5 the method does not work.

Applicability of the averaging method

The averaging method we describe has also been succesfully used in our laboratory to combine maps obtained by different phasing techniques. We have used MIR phase sets determined for 'cold' and 'warm' native datasets, different solvent flattening protocols and partial model phase sets, to combine them with the wARP procedure. The resulting map appears to be of substantially better quality. Unfortunately, this project is still under refinement and we can not quote the exact phase improvement figures. Furtermore, it is not a usual case to obtain many phase sets, with different sources of errors. Also, other more standard and theoretically sound procedures are developed for standard phase combination. Thus, we will not treat it as a test case, allthough potential users that think this procedure might be applicable in their particular cases are encouraged to inquire after this possibility with us.

References

1. Lamzin, V.S. & Wilson, K.S. (1993) Automated refinement of protein models. *Acta Crystallogr.* **D49**, 129-147.

2. Lamzin, V.S. & Wilson, K.S. (1996) Automated refinement for protein crystallography. In *Methods Enzymol.: Macromolecular Crystallography.* (Carter, C.M. & Sweet, R.M. Eds.) **in the press**

3. Konnert, J.H. & Hendrickson, W.A. (1980) A restrained-parameter thermal-factor refinement procedure. *Acta Crystallogr.* **A36**, 344-350.

4. Murshudov, G., Vagin, A. and Dodson, E (1996) Application of maximum likelihood refinement. In *The refinement of protein structures* Proceedings of Daresbury study weekend

5. CCP4 (1994) Collaborative Computational Project Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr.* **D50**, 760-763.

6. Dauter, Z., Sieker, L.C. & Wilson, K.S. (1992) Refinement of rubredoxin from Desulfovibrio vulgaris at 1.0 Å with and without restraints. Acta Crystallogr. B48, 42-59.

7. Watenpaugh, K.D., Sieker, L.C. & Jensen, L.H. (1980) Crystallographic refinement of rubredoxin at 1.2 Å resolution. J. Mol. Biol. 138, 615-633.

8. Lunin, V.Y. & Woolfson, M.M. (1993) Mean phase error and the map correlation coefficient. *Acta Crystallogr.* **D49**, 530-533,

9. Cowtan, K. (1994), Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography, **31**, 34-38

10. Perrakis, A.,*et al* and Vorgias, C.E. (1994) Crystal structure of a bacterial chitinase at 2.3 Å resolution. *Structure* **2**, 1169-1180.

11. Furey, W & Swaminathan, S.(1990). PHASES - a program package for the processing and analysis of diffraction data from macromolecules. *American Crystallographic Association Meeting Abstracts,* **18**, 73

12. Weeks, C.M., Hauptman, H.A., Smith, G.D., Blessing, R.H., Teeter, M.M. & Miller, R. (1995) Crambin: a direct solution for 400-atom structure. Acta Crystallogr. D51, 33-38.

13. Frazao, C., Soares, C.M., Carrondo, M.A., Pohl, E., Dauter, Z., Wilson, K.S., Hervas, M., Navarro, J.A., De la Rose, M.A. & Sheldrick, G.M. (1995) *Ab initio* determination of the crystal structure of cytochrome *c*6 and comparison with plastocyanin. *Structure* **3**,

Experimental low resolution envelopes from solution scattering

D. I. Svergun

European Molecular Biology Laboratory, Hamburg Outstation, EMBL c/o DESY, Notkestrasse 85, D-22603 Hamburg, Germany, and Institute of Crystallography, Russian Academy of Sciences, Leninsky pr. 59, 117333 Moscow, Russia. E-mail: Svergun@EMBL-Hamburg.DE

Introduction

Solution scattering is one of the most effective methods for investigating low resolution structure of biopolymers and their complexes (Feigin & Svergun, 1987). The scattering intensity I(s) from a dilute monodisperse solution is proportional to the scattering from a single particle averaged over all orientations [here s denotes the modulus of the scattering vector **s**, $s=(4\pi/\lambda)\sin\theta$, λ is the wavelength, and 2θ the scattering angle]. Main advantage of solution scattering is the possibility to study the structure and structural dynamics of native particles in physiological solutions; its main disadvantage is the loss of information due to the chaotic orientation of particles.

Information content in solution scattering data is usually estimated with the Shannon sampling theorem (Shannon & Weaver, 1949). A scattering curve I(s) is the Fourier image of the spherically averaged Patterson function of the particle $P(r)=\langle P(\mathbf{r})\rangle$ which equals to zero beyond $r=D_{max}$, where D_{max} is the maximum particle size. I(s) is therefore an analytical function. The sampling theorem states that the number of parameters (Shannon channels) required to represent an analytical function on an interval $[s_{min}, s_{max}]$ is equal to $N_s= D_{max}(s_{max} - s_{min})/\pi$. In practice, solution scattering curves decay rapidly with s and they are normally recorded only at low (not better than 1 nm) resolution, so that the typical number of the Shannon channels does not exceed 10 to 15.

In keeping with the low resolution of the solution scattering studies, the data interpretation is usually performed in terms of homogeneous bodies. Homogeneous approximation reduces the number of free parameters N_p in the model and is well justified in X-ray studies of single component particles (*e.g.* proteins) in water solutions. In conventional modelling, however, the particle is represented by hundreds of spheres, so that $N_p \gg N_s$ thus making only trial and error approach possible. Below an *ab initio* method is presented which utilizes spherical harmonics to economically describe low resolution particle envelopes and to restore them from solution scattering curves. Examples of the application of the method are given and possibilities of the joint use of crystallographic and solution scattering data are discussed.

Theory

Granted that the information content in solution scattering is low, an *ab initio* shape determination procedure should require as few parameters as possible. Let us represent the particle envelope by a two dimensional angular function $r=F(\omega)$ describing the particle boundary in spherical coordinates (r, ω). This function is conveniently parameterized as

$$F(\omega) \approx F_L(\omega) = \sum_{l=0}^{L} \sum_{m=-l}^{l} f_{lm} Y_{lm}(\omega) \qquad (1)$$

where $Y_{lm}(\omega)$ are spherical harmonics, the multipole coefficients f_{lm} are complex numbers and the truncation value L defines the resolution of the representation. The particle density distribution in homogeneous approximation can be written as

$$\rho(\mathbf{r}) = \begin{cases} 1, & 0 \le r < F(\omega) - \Delta \\ [F(\omega) - r]/\Delta, & F(\omega) - \Delta < r \le F(\omega) \\ 0, & r > F(\omega) \end{cases} \qquad (2)$$

where Δ is the width of the particle-solvent interface which for dissolved macromolecules can be taken Δ=0.3 nm to account for the first hydration shell. The particle envelope is thus represented by $(L+1)^2$ numbers f_{lm} at a spatial resolution $\delta r \approx \pi R_0/(L+1)$, where R_0 is the radius of the equivalent sphere.

Solution scattering intensity is $I(s) = \langle I(\mathbf{s})\rangle_\Omega = \langle\{\mathcal{F}[\rho(\mathbf{r})]\}^2\rangle_\Omega$, where $\mathcal{F}$ denotes the Fourier transform, $\langle\rangle_\Omega$ stands for the average over the solid angle Ω in reciprocal space, and $\mathbf{s}=(s, \Omega)$ is the scattering vector. Expanding $\rho(\mathbf{r})$ in spherical harmonics

$$\rho(r) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \rho_{lm}(r) Y_{lm}(\omega) \qquad (3)$$

the scattering intensity is expressed as (Stuhrmann, 1970a)

$$I(s) = 2\pi^2 \sum_{l=0}^{\infty} \sum_{m=-l}^{l} |A_{lm}(s)|^2 \qquad (4)$$

where the partial amplitudes $A_{lm}(s)$ are the Hankel transforms from the radial functions

$$A_{lm}(s) = i^l \sqrt{2/\pi} \int_0^{\infty} \rho_{lm}(r) j_l(sr) r^2 dr \qquad (5)$$

and $j_l(sr)$ are the spherical Bessel functions.

Inserting (2-3) into (5) and using the power series expansion for $j_l(sr)$ a closed expression for the partial amplitudes *via* the f_{lm} coefficients is obtained allowing one to rapidly evaluate the scattering intensity (4) from a given envelope (Stuhrmann, 1970b; Svergun & Stuhrmann, 1991; Svergun, 1997). Using this approach, an algorithm for *ab initio* determination of the low resolution envelopes of biopolymers in solution from their experimental scattering curves is developed. Starting from a spherical shape (for which all coefficients but f_{00} are equal to zero), the f_{lm} coefficients are obtained which minimize the discrepancy between the experimental $[I_{exp}(s_k), k=1,...N]$ and calculated curves

$$\chi^2 = \sum_{k=1}^{N} \left\{ \left[I_{exp}(s_k) - I(s_k) \right] W(s_k) \right\}^2 / \sum_{k=1}^{N} \left[I_{exp}(s_k) W(s_k) \right]^2 \qquad (6)$$

with the weighting factor $W(s_k)= s_k^2[\sigma(s_k)/I_{exp}(s_k)]$, where $\sigma(s_k)$ is the standard deviation in the k-th point. Details of the shape determination algorithm are presented elsewhere (Svergun *et al.*, 1996; 1997a).

Uniqueness

A natural question arises whether the low resolution shape determination is unique, in other words, whether, in addition to the trivial case of an enantiomorphic envelope, different shapes exist at the same level of resolution (i.e. at the same L) yielding identical scattering curves. This problem was considered by Svergun *et al.* (1996) using computer simulations on model bodies described by the envelope functions exactly represented by a finite series (1) on spherical harmonics. Given the scattering intensity calculated from a model envelope, the particle shape was restored from this intensity with the above algorithm. Both error-free curves and those containing statistical noise were simulated in different angular intervals.

The results indicated that the shape restoration for error-free data is unique, even when using very limited ranges in the simulated curves. In the presence of errors, ambiguity of the shape determination depends on the relation between the number of model parameters N_p and that of the Shannon channels N_s. The shape restoration was found to be practically independent of the initial approximation and stable with respect to the random errors if $N_p \approx 1.5\ N_s$.

Experimental solution scattering curves cover usually about 10 to 15 Shannon channels thus allowing to use 15 to 20 variables in the shape description. The number of independent parameters in series (1) is equal to $N_p=(L+1)^2-6$ (here, the reduction by six variables is due to arbitrary rotations and displacements of the particle which do not alter the scattering curve). It means that in practice the multipole resolution up to L=4 can be used.

Practice

Practical implementation of the shape determination algorithm required several extensions to account for the deviations from the ideal model:

(i) When using raw X-ray scattering data, homogeneous approximation may not be valid in the outer parts of the scattering curves where the scattering from the inhomogeneities of the polypeptide chain can no longer be neglected, especially for proteins of low (less than 20kDa) molecular mass. This effect is taken into account as follows. From the inner part of the scattering curve (first three Shannon channels), the best fit three-axial ellipsoid is found. Scattering from the internal inhomogeneities $I_s(s)$ inside the ellipsoidal envelope is evaluated using the method of Svergun (1994), and this curve is subtracted from the experimental data so that the difference $I_{exp}(s)- I_s(s)$ at higher angles follows the asymptotic behavior s^{-4} according to the Porod's law for homogeneous particles (Feigin & Svergun, 1987).

(ii) The model envelope is represented by a finite set of harmonics, whereas real particles would require the infinite series. To reduce the truncation effect, the best fit ellipsoidal envelope is developed into spherical harmonics, and its the shape representation (1) is truncated at the same L value as that used in the shape determination (usually, L=4). The ratio $w(s)=I_L(s)/I_{el}(s)$ is calculated where $I_{el}(s)$ is the scattering curve from the ellipsoid, $I_L(s)$ from its truncated representation. The experimental intensity is then multiplied by this "ellipsoidal filter" w(s) and the resulting curve $J_{exp}(s)= w(s)[I_{exp}(s)- I_s(s)]$ enters the shape determination.

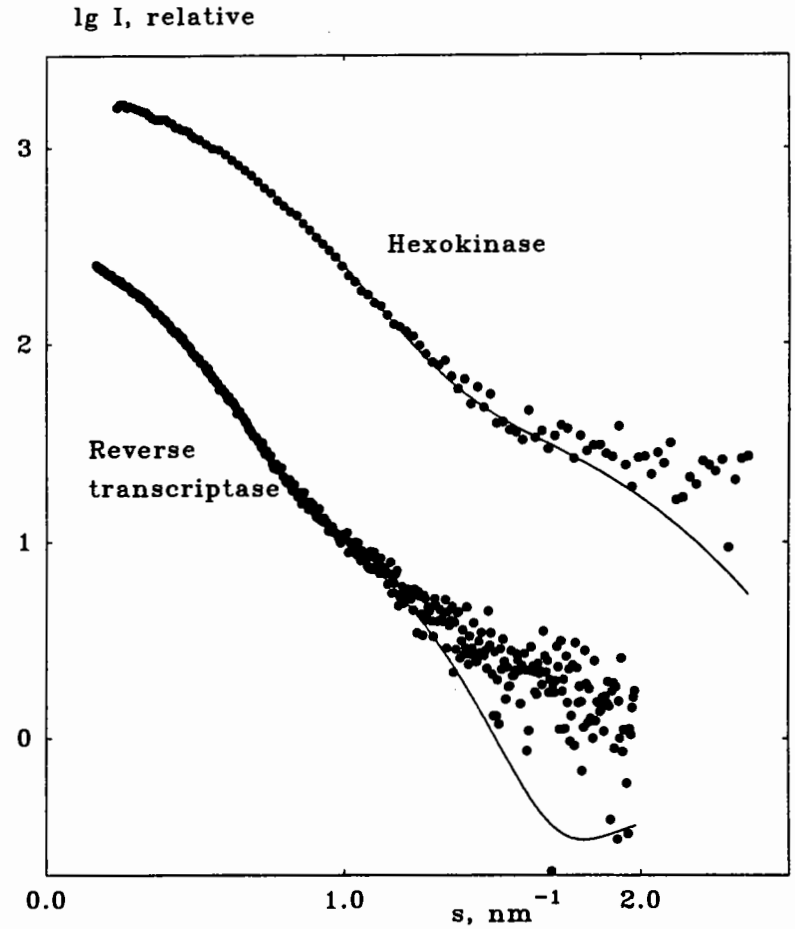


Figure 1. Shape determination of the hexokinase and HIV-1 reverse transcriptase. Dots: experimental X-ray scattering data, solid curves: scattering from the restored shapes.

(iii) When minimizing functional (6), the calculated intensity I(s) at each function evaluation is multiplied by the scaling factor

$$\mu = \sum_{k=1}^{N} J_{exp}(s_k) I(s_k) W^2(s_k) \Big/ \sum_{k=1}^{N} [I(s_k) W(s_k)]^2 \qquad (7)$$

which provides the currently best least squares fit to the experimental curve. The shape determination can therefore be directly applied to raw experimental data on a relative scale.

The *ab initio* shape determination program with the above extensions runs on IBM-PC and on major UNIX platforms (Svergun *et al.*, 1997a). Its implementation on a SUN Sparc-20ZX workstation is coupled with a three-dimensional rendering program ASSA allowing the user to monitor the process of the shape determination (Kozin, Volkov & Svergun, 1997).

The program has been tested on several proteins with known atomic structures in the crystal (X-ray solution scattering patterns were collected as parts of ongoing projects at the EMBL Outstation in Hamburg). Figs 1 and 2 present the shape determination of two proteins, monomeric hexokinase and HIV-1 reverse transcriptase (molecular masses 52 and 105 kDa, respectively). In both cases, particle envelopes up to L=4 (19 free parameters) were directly restored from the experimental data starting from a spherical initial approximation. The envelopes are displayed in Fig. 2 along with the atomic structures of the hexokinase (Bennett & Steitz, 1980), and of the reverse transcriptase (Wang *et al.*, 1994) deposited in the Protein Data Bank (Bernstein *et al.*, 1977), entries 1HKG and 3HVT,

respectively). As the orientation of the restored models is arbitrary, they and their enantiomorphs were rotated so as to minimize the deviation

$$R_\omega = \int \left[F_{cryst}(\omega) - F(\omega) \right]^2 d\omega \,/\, \int F^2_{cryst}(\omega) d\omega \qquad (8)$$

where $F_{cryst}(\omega)$ is the envelope function evaluated for the atomic structure at the same L using the program CRYSOL (Svergun, Barberato & Koch, 1995). As seen from the comparison, the *ab initio* restoration provides an adequate low resolution description of the protein envelopes. The R_ω factors are equal to 0.20 and 0.22 for the hexokinase and for the reverse transcriptase, respectively.

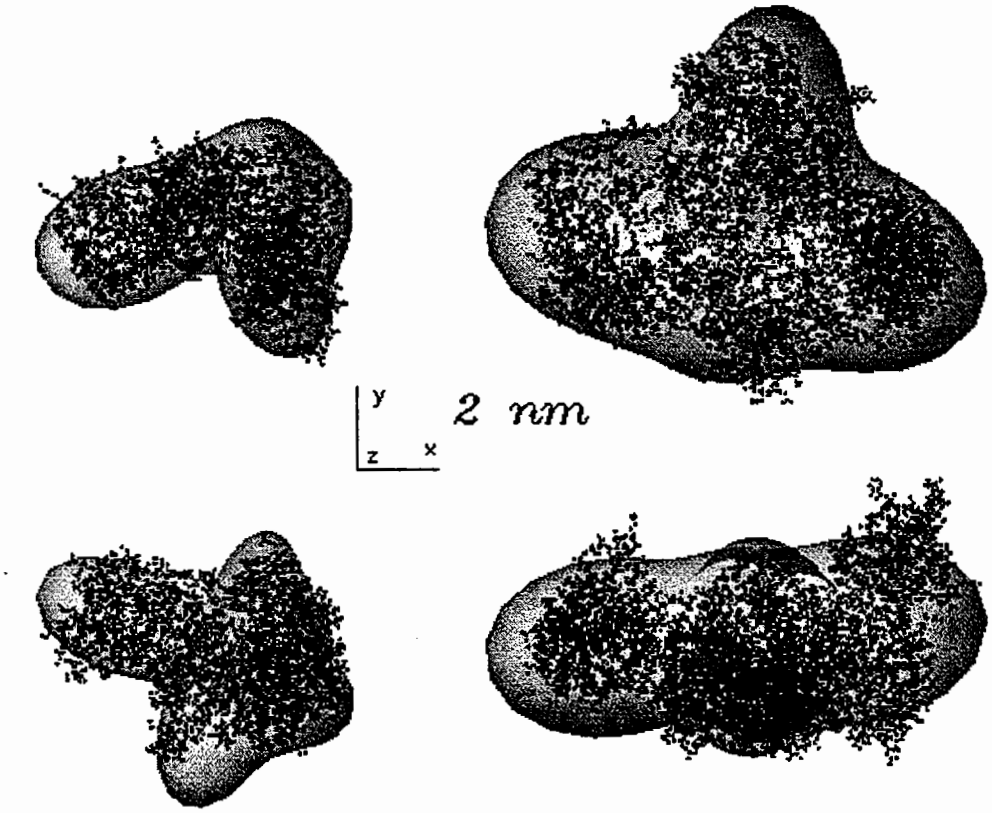


Figure 2. Comparison between the envelopes of the hexokinase (left) and reverse transcriptase (right) restored from solution scattering data (transparent solids) with corresponding crystallographic structures (dots). Bottom pictures are rotated 90^0 clockwise around X.

The shape determination program was also used to restore the envelopes of other proteins with known atomic structures (lysozyme, ribonucleotide reductase, pyruvate decarboxylase, enopyruvil transferase, *etc.*). In all these cases the restored shapes agreed well with the atomic structures, with the R_ω factors ranging from 0.10 to 0.25. Of course, the program is aimed at the shape determination of the proteins with unknown atomic structure; the above tests have been done to check the reliability of the method in real experiment.

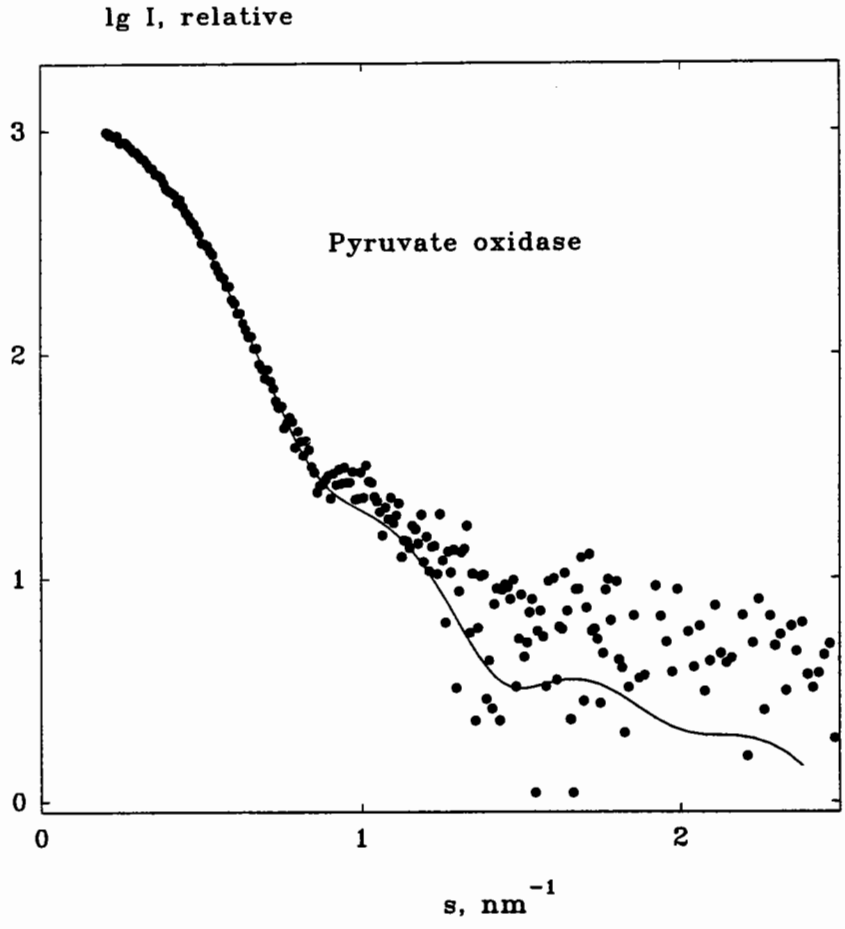


Figure 3. Shape determination of the pyruvate oxidase. Dots: experimental X-ray scattering data, solid curve: scattering from the restored shape.

Symmetry

Particle symmetry imposes restrictions on the multipole coefficients f_{lm} in series (1) and the information about the symmetry, if available, can improve the reliability of the *ab initio* shape restoration by reducing the number of parameters to be determined. Consider, for example, a homodimeric particle with a two fold symmetry axis along z. In this case, all f_{lm} coefficients with odd m vanish, and the particle shape at L=4 is described by 12 independent parameters instead of 19 for a non-symmetric case.

The higher the symmetry, the more multipole coefficients can be omitted, and this allows one to enhance the resolution of the restoration. Figs 3 and 4 present the shape determination of the homotetramer of pyruvate oxidase (molecular mass 260 kDa) assuming the 222 point symmetry. The multipole expansion up to L=6 for this symmetry group requires only 13 free parameters. The restored envelope displays a good agreement (R_ω=0.15) with the crystal structure (Muller & Schultz, 1993, PDB entry 1POW)

The quaternary structure of symmetric particles can also be restored in terms of the envelope function of the asymmetric unit. Thus, scattering from a symmetric homodimer is readily expressed *via* the shape of a monomer and the distance Δd between the monomers. The shape determination is performed as described above with a single additional parameter Δd. This approach has already been successfully used in practice (Schmidt *et al.*, 1995; Svergun *et al.*, 1997a).

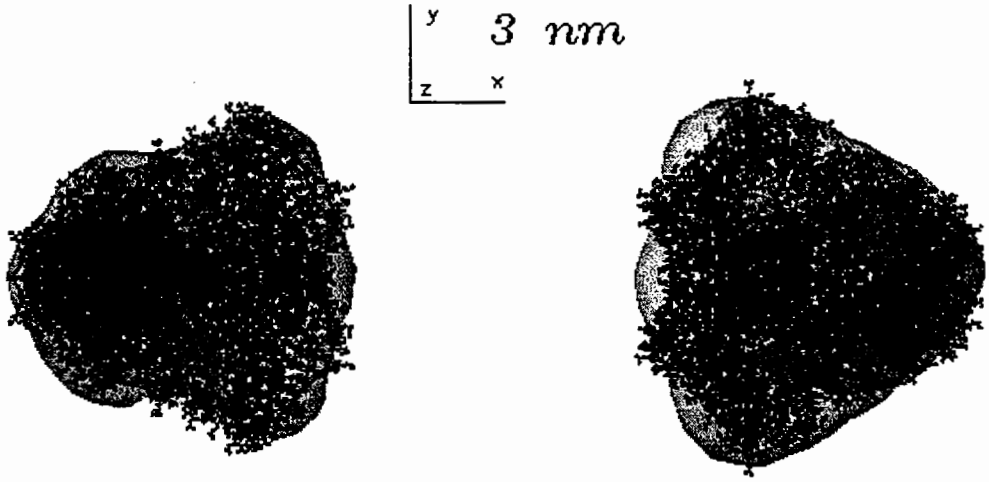


Figure 4. Comparison between the envelope of the tetrameric pyruvate oxidase restored from the solution scattering data assuming the 222 symmetry (transparent solid) and the crystallographic structure (dots). Right picture is rotated 90^0 clockwise around X.

Discussion

The first question to address is why is it at all possible to restore the three-dimensional envelope from a one-dimensional curve using more parameters than predicted by the theory? The answer is that the estimate of N_s reflects only one (and most often quoted) part of the sampling theorem. The other part says that full information about the entire analytical function is contained in any finite contiguous portion of it. An oversampled scattering curve measured with the angular increment much smaller than the sampling distance π/D_{max} can be analytically extrapolated beyond the experimental range (so-called superresolution). As experimental solution scattering curves are always heavily oversampled, they are able to provide more parameters than N_s.

Limitations of the model (1) used to describe the particle envelope should be mentioned. First, as $F(\omega)$ is assumed to be single-valued, complicated (*e.g.* U-like) shapes or those containing internal holes cannot be exactly represented. Second, omission of the higher harmonics with $l>L$ is compensated in the fitting procedure by the artificial enhancement of the lower ones. This effect is partially corrected by the above described ellipsoidal filtering and thus produces only marginal distortions for globular particles but can still be significant for anisometric structures because of a slow convergence of series (1). Remaining deviations between the restored envelopes and the crystal structures in Fig. 2 provide an idea on the magnitude of the truncation effect (it is worth noting that both proteins are rather anisometric, with the axial ratios of the approximating ellipsoid equal to 2.8 and 3.6 for the hexokinase and reverse transcriptase, respectively).

What is the relation between the solution scattering and crystallographic data? The latter clearly contain more information and provide much higher resolution. However, test runs of the shape determination using simulated reflections instead of solution scattering curves encountered difficulties because of a high multimodality of the goal function. The reason for the multimodality is that the crystallographic data, contrary to the solution scattering curves, are undersampled: separation between the reflections is twice the sampling distance required to describe the three-

dimensional scattering intensity as the Fourier image of the density in the unit cell (*e.g.* Baker, Krukowski & Agard, 1993). Solution scattering data provide therefore complementary information and their use can improve the efficiency of *ab initio* phasing procedures. Low resolution experimental envelopes can be positioned in the crystal cell using molecular replacement and further refined against both solution scattering and the crystallographic data.

Measurements in solution provide also a possibility to model the structure and structural transitions of complex macromolecules in solution by rigid body movements of their crystallographically known domains (subunits) so as to fit the experimental scattering from the complex (Svergun, 1991; 1994; 1997). Thus, in solution scattering study of the classical allosteric enzyme aspartate transcarbamylase (Svergun et al., 1997), the overall changes accompanying the T→R transition in solution were found to be about 50% larger than those in the crystal (Kantrowitz & Lipscomb, 1988). This approach is now being used in several ongoing projects at the EMBL Outstation in Hamburg to study multidomain proteins in solution.

Acknowledgments

The work was supported by the INTAS Grant No 93-645. The author thanks V.V.Volkov, M.B.Kozin and H.B.Stuhrmann for the help in program development, and M.H.J.Koch, C.Reißner, L.Goobar-Larsson and S.König for providing the experimental data.

References

Baker, D., Krukowski, A.E. and Agard, D.A. *Acta Cryst.*, D49 (1993) 186.
Bennett, W.S. Jr. and Steits, T.A. *J. Mol. Biol.* 140 (1980) 183.
Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F. Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi M. *J. Mol. Biol.*, 112 (1977) 535.
Feigin, L.A. and Svergun, D.I.. *Structure analysis by small-angle X-ray and neutron scattering.* New York: Plenum Press (1987).
Kantrowitz, E.R. and Lipscomb, W.N. *Science,* 241 (1988) 669.
Kozin, M.B., Volkov, V.V. and Svergun, D.I. *J. Appl. Cryst.*, 30 (1997), in press.
Muller, Y.A. and Schultz, G.E. *Science*, 259 (1993) 965.
Shannon, C.E. and Weaver, W. *The mathematical theory of communication.* University of Illinois Press, Urbana (1949).
Stuhrmann, H.B. *Acta Cryst.*, A26 (1970a) 297.
Stuhrmann, H.B. *Z. Phys. Chem. Frankfurt,* 72 (1970b) 177.
Schmidt, B., König, S., Svergun, D., Volkov, V., Fischer, G. and Koch, M.H.J. *FEBS Letters,* 372 (1995) 169.
Svergun, D.I. *J. Appl. Cryst.*, 24 (1991) 485.
Svergun, D.I. *Acta Cryst.*, A50, (1994) 391.
Svergun, D.I. *J. Appl. Cryst.*, 30 (1997), in press.
Svergun, D.I. and Stuhrmann, H.B. *Acta Cryst.*, A47 (1991) 736.
Svergun, D.I., Barberato, C. and Koch, M.H.J. *J. Appl. Crys.* 28, (1995) 768.
Svergun, D.I., Volkov V.V., Kozin M.B. and Stuhrmann H.B. *Acta Cryst.*, A52 (1996) 419.
Svergun, D.I., Volkov V.V., Kozin M.B., Stuhrmann H.B., Barberato C. and Koch M.H.J (1997). *J. Appl. Cryst.*, (1997a), in press.
Svergun, D.I., Barberato, C., Koch, M.H.J., Fetler L. and Vachette P. *Proteins*, 27 (1997b), in press.
Wang, J., Smerdon, S.J., Jaeger, J., Kohlstaedt, L.A., Friedman, J., Rice, P.A. and Steitz T.A. *Proc. Natl. Acad. Sci. USA,* 91 (1994) 7242.

LOW RESOLUTION CRYSTALLOGRAPHIC IMAGES

by A.Urzhumtsev#*, V.Lunin* and A.Podjarny#

#UPR de Biologie Structurale, IGBMC, B.P.163, 67404, Illkirch, c.u. de Strasbourg, France
**IMPB RAN, Puschino, Moscow region, 142292, Russia*

1. Introduction.

The definition of «low resolution» depends on the traditions of a specific laboratory and, first of all, on their typical subjects. In the case of small molecules it can be 3Å. In the case of typical proteins, it is rather about 6-8Å. Another meaning of the term «low resolution» is about 20-25Å, the limit below which X-ray data are quite often not collected.

This paper deals with the analysis of macromolecules, and the resolution below 6-8Å will be referred to as «low resolution» and the one below 20-25Å as «very low resolution» (VLR in what follows). It should be noted that these two limits define the resolution zone where the contribution of the bulk solvent is strong and uncorrelated to that from the macromolecule itself. At higher resolutions the contribution is negligible, and at lower resolutions it is strong but roughly proportional to the one of the macromolecule (Urzhumtsev & Podjarny, 1995).

Measuring the very low resolution X-ray data is technically difficult, and many research groups do not collect them. However, they carry information that can be useful. This paper discuss their importance for improving the molecular images as well as the possibilities of an independent use of these data.

2. Do very low resolution data have any information ?

The basic sources of a noise in macromolecular crystallographic images are *systematic* errors. While in a real case the synthesis is usually worse than expected it is much more difficult to obtain a noisy image in a test calculation. The mean value of *independent* phase errors can reach about 60-70° and the synthesis will still be quite good and close to the ideal image. However, it is easy to destroy an image by introducing *systematic* errors, for example, by error in heavy atoms parameters. Another example is missing of a part of the model, for example the solvent.

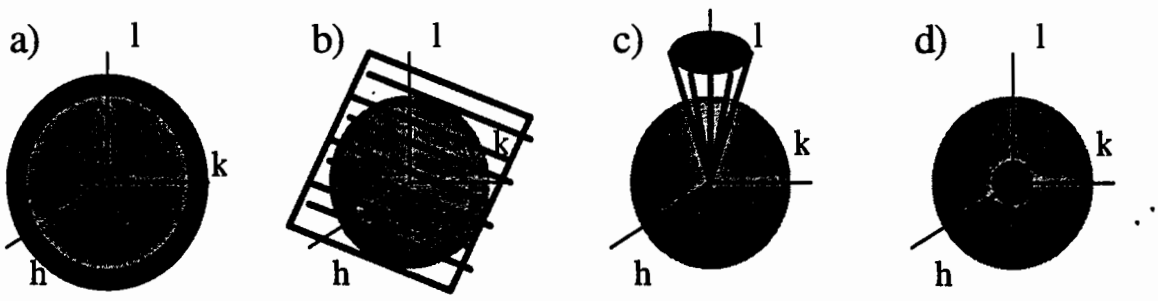


Fig. 1. Schematic presentation of different possibilities of systematic missing of X-ray amplitudes: a) standard resolution cut-off; b) in plane; c) along one axis; d) very low resolution data

The second possibility is *systematically missing* of reflections in the map calculations. Some examples are schematically shown in Fig. 1. Fig 1a corresponds to the *usual* high resolution cut-off. Fig 1b and 1c corresponds to relatively *rare cases* which nevertheless exist. Missing of a plane of reflections causes breaks in the density in the conjugate direction in real space (Lunin, 1991). A systematic absence of reflections along an axis can cause a complete loss of the molecular envelope (Urzhumtsev et al., 1989).

Fig. 1d corresponds to a *usual situation* when VLR data are excluded from the map calculation. They can be either not measured or measured but not phased. Usually they are only a small number of reflections but they are strong and removed systematically.

A phase extension procedure for phasing the VLR data applied by Podjarny et al. (1981) in the case of tRNA demonstrated a drastic improvement of the image. For calculated data (Urzhumtsev, 1991) it was clearly shown that the exclusion of only 1% of the data (29 reflections out from 2500) completely destroy the molecular image at 6Å resolution. In the case, for example, of SIR phase errors, the molecular envelope keeps its position but the electron density peaks are shifted. In the case of missing VLR reflections the effect is inverted: the envelope is lost but the peaks are at their places. This is natural because the exclusion of VLR terms should cause large scale modulations of the density in the unit cell.

The fact that the peaks are at the right place has important consequences. Firstly, when a map is calculated at high resolution, its peaks have a high contrast and such density modulation does not «hide» them completely; this has allowed crystallographers to ignore VLR data for a long period. Secondly, it gives a possibility of automatically determining the molecular envelope from such synthesis.

The knowledge of the envelope can be used to improve the molecular image. The phases of its structure factors can be used as a good approximation to the phase values of VLR reflections. If their amplitudes are available, simple adding them to the Fourier calculation can completely change the map (see Urzhumtsev, 1991, for an example of dractic improvement of the SIR image of the Elongation Factor G). Calculated amplitudes can be used to estimate the quality of the calculated phases and give corresponding weights for the Fourier coefficients through the comparison with the experimental ones.

3. How to use the information from very low resolution data ?

Therefore, VLR data do carry important information, first of all, on the shape of molecule. Such an information can be used in different cases (Podjarny & Urzhumtsev, 1997), for example:

- in density modification procedures for the image improvement;
- in the molecular replacement if the *internal* differences between two molecules are large;
- if diffraction data are not available at higher resolution;
- in the case of very large molecular complexes, like ribosome;

etc.

If VLR amplitudes have been measured, the determination of their phases by isomorphous replacement is difficult while not impossible (Podjarny & Urzhumtsev, 1997). In the case of viruses where practically all VLR reflections are centrosymmetric, a good approximation can be done by calculation of structure factors from a spherical shell. In the general case, a searching procedure based on some *a priori* knowledge of the density can be

applied to find these phases. For this procedure, it is needed to specify: a) the search model (parameters); b) the search space; c) the sampling procedure; d) the available data; e) the criteria of the search.

Sampling of the whole phase space

In the simplest case, the search is either systematic or random with a representative sampling of the whole phase space. In practical terms it means that the space should not be too large, i.e. a small number of reflections can be phased.

An information which can be used to identify the correct solution should be of general type, for example, the knowledge of the correct electron density histogram (Lunin, 1988; 1993). For any generated phase set, a map of a given resolution can be calculated with experimental amplitudes and these phases. A correlation of histograms, target and calculated, can be use as a search criterion.

Table 1 shows the distribution (histogram correlation, C_H, *vs* phase correlation, C_P) of phase sets of a typical search done with calculated data. Most of the generated phase sets have a poor value for both correlations. However, the converse is not true and the phase sets with highest value of the histogram correlation are not necessarily correct. The phase correlation distribution of these phase sets (columns with $C_H \geq 0.9$ in Table 1) shows two groups of phase sets. Some of them have a reasonable phase correlation value, while the others are far from the correct solution. Note also that there are a number of sets with high C_P but low histogram correlation value. The single phase set with the highest C_H value and also the highest C_P value is not statistically significant.

C_H	--->	0.7	-----	-----	-----	-----	0.8	-----	-----	-----	0.9	-----	-----	-----	--->	1.0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Λ
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	I
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	I
0	0	0	0	0	0	1	1	4	1	5	1	1	0	0	0	0.950
0												0	0	0	0	I
												0	0	0	0	0.875
												0	0	0	0	I
												0	0	0	0	I
											0	0	0	0	0	Λ
												0	0	0	0	I
												0	0	0	0	0.500
												0	0	0	0	I
												0	0	0	0	C_P

Table 1. Two-dimensional distribution of the generated variants of phase sets for the case of model data at 30Å resolution (29 reflections). The horizontal dimension corresponds to the histogram correlation, C_H, and the vertical one to the weighted phase correlation, C_P. The correct solution should be in the top right corner. Two major clusters are marked by a frame, the variants with highest C_P are indicated by inverted colours.

The behaviour of the C_P values for the sets with highest C_H can be generalised. The search criterion does not select for a single solution, but gives an indication of possible solutions. This solutions are not uniformly distributed with respect to the C_P. Further analysis shows that they appear in clusters (e.g., the two peaks in Table 1 correspond to two clusters); one of these clusters is close to the correct solution.

On the basis of these observations, the following procedure has been suggested to obtain *ab initio* the phases of the VLR reflections (Lunin, Urzhumtsev, Skovoroda, 1990):

a) generation of a large number of phase sets (e.g., one million for 30 phases);
b) calculation of an electron density map for every phase set and calculation of its histogram;
c) selection of the phase sets with highest histogram correlation as admissible ones;
d) after a sufficient number (e.g., one thousand) phase sets are selected, analysis of the distribution of these sets by some clustering procedure;
e) classification of the clusters according to their size in a 'cluster tree'; for every major cluster average the corresponding phase sets in order to obtain the mean phase values and their figures of merits;
f) calculate the corresponding weighted maps and choose, if possible, the correct one.

For the step (d), a proper distance between two phase sets should be defined taking into account different choices of the unit cell origin (Lunin & Lunina, 1996), density flipping and enantiomer.

The procedure was found quite robust in several applications both to the calculated and experimental data. In these cases about 30 reflections were successfully phased which gave images of reasonable quality. The limiting point was the computing time. In order to get finer details, it is necessary to go deeply in the cluster tree to smaller clusters and still have large enough number of phase sets with a high enough value of the criterion.

Another problems is that, unfortunately, while for the middle resolution maps a general method to obtain the corresponding histogram *a priori* has been suggested (Lunin & Skovoroda, 1991), no similar method was found for the very low resolution.

It is important to note that a similar behaviour of the selected phase variants has been observed when the criterion of the histogram closeness was replaced by the criterion of a compact globular envelope.

Simplest parametrisation of the phase space

In order to increase the number of phased reflections for the same level of computing power, the search model should be parametrised. A proper parametrisation should automatically avoid sampling of the «empty» regions of the phase space and the correct phase set should belong to the chosen subspace or, at least, be close enough to it. The number of parameters of every model should be small enough (at least, less that the number of data) in order to make the criterion of choice significant.

The simplest possible way of modelling a molecule is to replace it with a large gaussian sphere, which involves only four parameters (position and radius). Systematic R-factor search with such a model is a known approach to find the centre of the gravity of the molecule. It has been successfully applied in several cases, for example, by Podjarny et al. (1987).

A search with several (N) spheres can be tried but for N>2 it is computationally difficult. In this case, a random sampling can been applied, similarly to the one used for the histogram criterion. A number of test calculations have been carried out using the experimental data of the $tRNA^{Asp}$-RS complex (Giegé et al., 1980; Urzhumtsev et al., 1994).

First, several models of 5-7 large spheres were constructed manually which reproduced the low resolution (30-50Å) image of the complex with a high correlation (0.75-0.80) with the exact one. Then a large number of models, each composed of a small number (2-5) of spheres with randomly distributed centres was generated. Corresponding structure factors were calculated and compared with the correct values, using the amplitude correlation, C_F, as the search criterion. It was found that, similarly to the search with the histogram criterion, the phase sets corresponding to the models with highest C_F are grouped in a small number of clusters, one of which is quite close to the correct phase set. A typical distribution is shown in Table 2 and is schematised in Fig. 2. To check whether this type of the distribution of selected variants is related to the random sapling, two different 2-spheres searches, a random one and a systematic one, have been carried out exactly at the same conditions. The corresponding distribution were very similar.

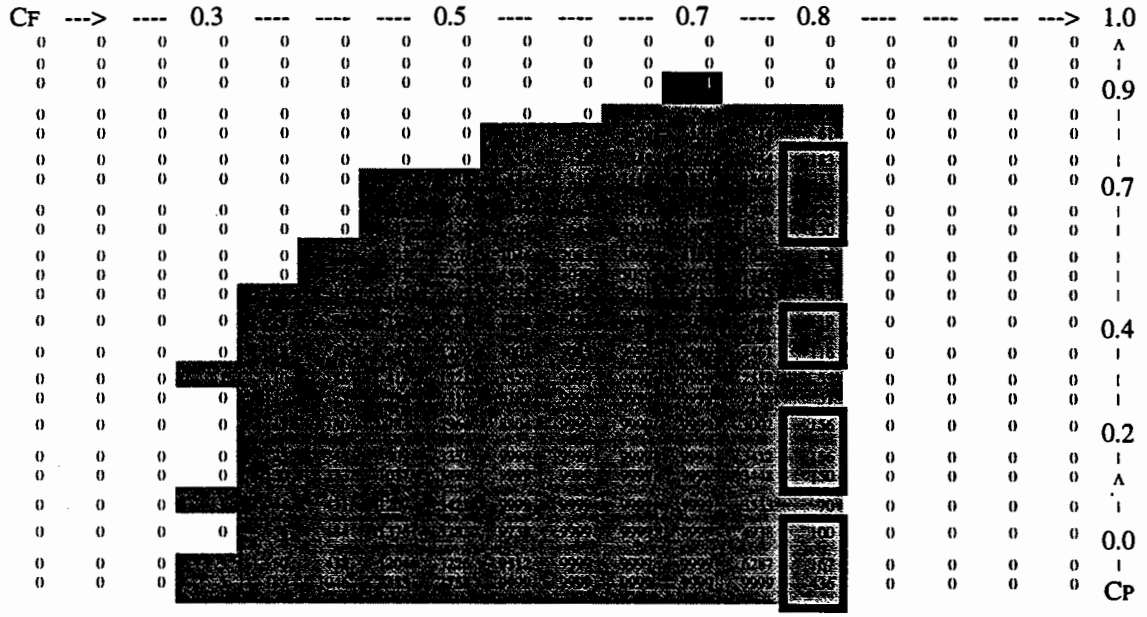


Table 2. Two-dimensional distribution of the FAM-generated variants of phase sets for the case of experimental data of the AspRS complex at 50Å resolution (31 reflections). The horizontal line corresponds to the amplitude correlation, C_F, and the vertical one to the weighted phase correlation, C_P. The correct solution should be in the top right corner. The major clusters are marked by a frame, the variant with highest C_P is indicated by inverted colours.

Several important observations should be noted:

1) the best phase set (C_P=0.9, C_F=0.7) does not correspond to the model with the highest amplitude correlation (C_F=0.8);
2) some of the phase sets with high amplitude correlation (C_F=0.8) are close to the correct phase set (C_P=0.7-0.8);
3) a phase set calculated from a model with high amplitude correlation (C_F=0.8) can belong to a cluster quite far (C_P=0.0) from the correct point;
4) averaging of phase sets inside the correct cluster produces a new phase set which is usually better that any individual solution.

A systematic procedure for this search, called FAM (Few Atoms Model), was proposed (Lunin et al., 1995) consisting of the following steps:

a) generation of a large number of simple pseudo-atomic models; every model consists of a the same small (2-10) number of large gaussian spheres; the co-ordinates of the centre of the spheres are distributed randomly in the unit cell;
b) structure factors calculation for every model;
c) comparison of the calculated amplitudes with the experimental ones and selection of the models with the highest CF;
d) merging of the selected phase sets by a clustering procedure;
e) analysis of the cluster tree; averaging of the phase sets inside every major cluster;
f) calculation of corresponding maps and identification, if possible, of the correct one.

This procedure was applied to several calculated and experimental data sets, giving good results. In particular, a 70Å-resolution crystallographic image (about 160 reflections) has been obtained for the 50S ribosomal particle (Volkmann et al., 1990) from *Thermus thermophilus* (T50S; Urzhumtsev et al., 1996). The FAM procedure is in the course of further development.

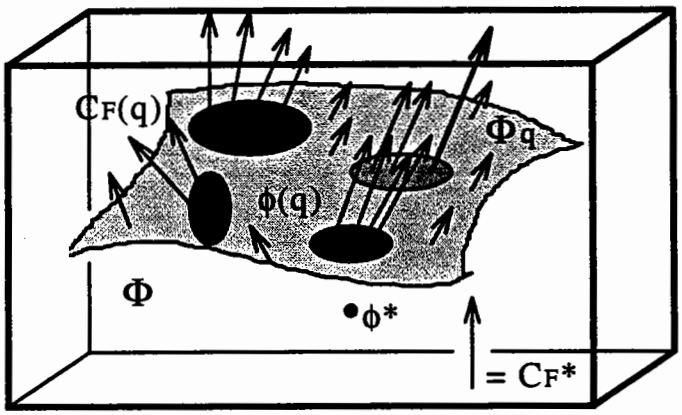


Fig. 2. Schematic presentation of the distribution of the phase sets in the FAM method. Every phase set is presented by a point in the phase space, Φ. Φ_q is a subspace of phase sets, $\phi(q)$, calculated from FAM models. The length of an arrow is proportional to the corresponding amplitude correlation, $C_F(q)$. The variants with $C_F > C_F^*$ are forming few clusters, one of them is close to the correct solution, ϕ^* (thick point).

Further parametrisation of the phase space

In the case were precise information about the three-dimensional molecular structure is available, the search space can be drastically reduced. This leads to the molecular replacement procedure (Rossmann, 1972), which reduces the dimension of this space to six, making possible a quasi-complete search. This procedure has been recently simplified (Navaza, 1994) giving automatically a list of possible positions and orientations of the model. In the case of good data and model, the correct solution corresponds to the maximum amplitude correlation. Alternative (wrong) variants have much lower correlation values, which allows to choose the solution easily. Otherwise, finding the answer is a difficult problem.

Molecular replacement is a standard technique, carried out usually at middle resolution (4-6Å) with an atomic model. At the VLR end the search model becomes a molecular envelope. If the search model is perfect, and the data are very accurate, a similar procedure with some important modifications (Urzhumtsev & Podjarny, 1995) brings the solution with reasonable contrast. In the case of less accurate data and an imperfect model

the contrast is much lower, as it was the case for the T50S particle (Urzhumtsev et al., 1996).

At very low resolution the imperfections of the model envelope can be important. For example, images reconstructed from electron microscopy can be compressed in one direction. When working with such models, molecular replacement puts the envelope either at its correct place (if possible) or into the solvent region but practically never at an intermediate position. This confirms a clustering behaviour of selected variants also for this case.

4. General conclusions

Several different low resolution phasing techniques which explore either the whole phase space or some specific subspace have been analysed. In all cases, the variants with best values of the search criterion are grouped in a small number of clusters which can be easily identified. One of these clusters is usually very close to the correct solution of the phase problem while others can be very far from it. It is important to note that the phase set with the best value of the criterion does not necessarily belong to this correct cluster. This observation explains, in particular, the problems with searches selecting a *single* solution. In general, this typical distribution of phase correlation *vs* search criterion has (by a peculiar coincidence) schematically the shape of the Strasbourg cathedral (Fig. 3; compare, for example, with Table 2). The top corresponds to the best variant which is impossible to identify by the available criteria, the floors correspond to the clusters, and the highest floor is the best cluster.

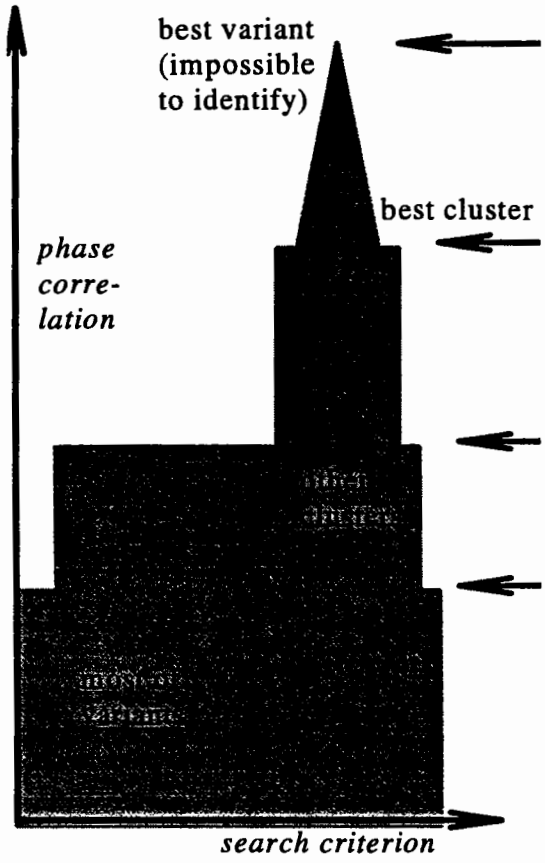


Fig. 3. Schematical profile of the Strasbourg cathedral.

As it was observed, the character of this distribution does not depend on the particular information and criterion used. For example, it can be noted in addition that the LAPS method developed by Volkmann (Volkmann et al., 1995) based on the Bricogne's maximum likelihood criterion found the solution for the T50S case also through a cluster oriented search.

All these observations indicate that at the very low resolution end the available information and search criteria are *weak* in the sense that in general they cannot indicate unambiguously the correct solution; additional information is necessary. At higher resolution, the same information and criteria, e.g., an atomic model and the amplitude correlation, can be *strong* enough to indicate a single solution. The particular low resolution

cases where the information is very accurate and the same criteria can unambiguously identify the right solution remain the exception rather than the rule.

The authors thank D.Moras for his support. This work was supported by the CNRS-RAS collaboration. AGU and ADP were supported by the Centre National de la Recherche Scientifique (CNRS) through the UPR 9004, by the Institut National de la Santé et de la Recherche Médicale, by the Centre Hospitalier Universitaire Régional. VYL was supported by RFBR grant 94-04-12844.

References

Giegé, R., Lorber, B., Ebel, J.-P., Thierry, J.-C. and Moras, D. Comp.Rend.Acad.Sci. (Paris), série D, *291* (1980) 393

Lunin, V.Yu. Acta Cryst., *A44* (1988) 144

Lunin, V.Yu. Dr.Sci.Theses, Institute of Crystallography RAS, Moscow (1991)

Lunin, V.Yu. Acta Cryst., *D49* (1993) 90

Lunin, V.Yu., Urzhumtsev, A.G. and Skovoroda, T.A. Acta Cryst., *A46* (1990) 540

Lunin, V.Yu. and Skovoroda, T.P. (1991) Acta Cryst., *A47* (1991) 45

Lunin, V.Yu., Lunina, N.L., Petrova, T.E., Vernoslova, E.A., Urzhumtsev, A.G. and Podjarny, A.D. Acta Cryst., *D51* (1995) 896

Lunin, V.Yu. and Lunina, N.L. Acta Cryst., *A52* (1990) 365

Navaza J. (1994) Acta Cryst., *A50* (1994) 157

Podjarny, A.D., Schevitz, R.W. and Sigler, P. Acta Cryst., *A37* (1981) 662

Podjarny, A.D., Rees, B., Thierry, J.-C., Cavarelli, J., Jesior, J.C., Roth, M., Lewitt-Bentley, A., Kahn, R., Lorber, B., Ebel, J.-P., Giegé, R. and Moras, D. J.Biomol.Struct.Dynam., *5* (1987) 187

Podjarny, A.D. and Urzhumtsev, A.G. In *Methods in Enzymology*, (1997) in press

Rossmann, M.G. «The Molecular Replacement Method». Gordon & Breach, New York (1972)

Urzhumtsev, A.G. (1991) Acta Cryst., *A47* (1991) 794

Urzhumtsev, A.G., Lunin, V.Yu. and Luzyanina, T.B. Acta Cryst., *A45* (1989) 34

Urzhumtsev, A.G., Podjarny, A.D. and Navaza, J. Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography, *30* (1994) 29

Urzhumtsev, A.G. and Podjarny, A.D. Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography, *32* (1995a) 12

Urzhumtsev, A.G. and Podjarny, A.D. Acta Cryst., *D51* (1995b) 888

Urzhumtsev, A.G., Vernoslova, E.A. and Podjarny, A.D. Acta Cryst., *D521* (1996) 1092

Volkmann, N., Hottenträger, S., Hansen, H.A.S., Zaytzev-Bashan, A., Sharon, R., Yonath, A. and Wittmann, H.G. (1990) J.Mol.Biol., *216* (1980) 239

Volkmann, N., Schlunzen, F., Urzhumtsev, A.G., Vernoslova, E.A., Podjarny, A.D., Roth, M., Pebay-Peyroula, E., Berkovitch-Yellin, Z., Zaytsev-Bashan, A. and Yonath, A. Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography, *32* (1995) 23