DL/SCI/R32

# ISOMORPHOUS REPLACEMENT AND ANOMALOUS SCATTERING

Proceedings of the CCP4 Study Weekend
25-26 January 1991

Compiled by W. Wolf, P.R. Evans and A.G.W. Leslie

Science and Engineering Research Council
DARESBURY LABORATORY
Warrington WA4 4AD, U.K.

# ISOMORPHOUS REPLACEMENT AND ANOMALOUS SCATTERING

Proceedings of the CCP4 Study Weekend
25-26 January 1991

Compiled by
W. Wolf, Daresbury Laboratory
and
P.R. Evans and A.G.W. Leslie, MRC Cambridge

SCIENCE & ENGINEERING RESEARCH COUNCIL
DARESBURY LABORATORY
1991

# CONTENTS

# INTRODUCTION

The method of isomorphous replacement has been around for so long that it sometimes seems that everything about it is understood. It remains, however, the only method of ab initio phasing for proteins, at least for the time being, and we should not be complacent about how good our methods are. The "traditional" methods of heavy-atom parameter refinement and phase calculation are not the best that can be done, and several contributions to this meeting show that we can improve the statistical treatment of errors.

The use of information from the anomalous scattering of heavy-atoms has always been part of the isomorphous replacement method as applied to macromolecules, but in recent years measurements of the anomalous scattering at different wavelengths ("MAD" phasing) have been used to solve several structures, and this may be becoming the method of choice in some cases. To extract phases from such data requires even greater care about errors than for conventional MIR phasing.

A great deal of experience has been built up over the years in the practical use of isomorphous replacement, and this meeting allowed a summing-up of the "state-of-the-art".

The meeting was organized and supported by the SERC Collaborative Computational Project in Protein Crystallography (CCP4) at Daresbury Laboratory. We wish to thank the invited speakers for their considerable efforts in making the meeting a success and their co-operation in the preparation of these proceedings.

We thank the Daresbury Laboratory and its Director, Professor A.J. Leadbetter, for the provision of organisational help and support, for both the meeting and in the publication of the proceedings. In particular we thank Shirley Lowndes, David Brown and Pauline Shallcross for their great assistance in the planning and organisation of the Study Weekend. In addition the proceedings owe much to the efforts of Geoff Berry and his staff.

Phil Evans
Andrew Leslie
Wojciech Wolf

November 1991

# The Chemistry of Heavy Atom Attachment

J. Drenth
Lab. Chemical Physics, Nijenborgh 16
9747 AG Groningen, The Netherlands

## Introduction

Since the introduction of the isomorphous replacement method for the X-ray structure determination of proteins by Green, Ingram and Perutz (1954), the field has grown quantitatively in number of heavy atom reagents and in number of proteins to which the method has been applied. However qualitatively not much has changed. It is still basically an empirical method and very often many dozens of reagents are tried before a few suitable ones have been found.

## The chemical reaction

We are interested in the following chemical reaction:

$$\text{heavy atom reagent} + \text{protein} \rightleftarrows \text{protein derivative}$$

Usually we do have already some chemical and physico-chemical information about the protein e.g. its amino acid sequence or composition, its biological substrate, the isoelectric point. This can be used and should be used as much as possible to put some rational into the experiment.

The vessel in which the reaction occurs is the protein crystal because the preferred method is soaking the protein crystal in the solution of the reagent. The composition of this solution is identical to the mother liquor, often with a slight increase of precipitant concentration. Co-crystallisation is not frequently used, because there is the risk that crystals will not grow. The soaking procedure depends on the existence of relatively wide pores in the crystal, wide enough to allow the reagent to diffuse into the crystal and to reach the reactive sites on the surface of all protein molecules in the crystal. An extremely high excess of reagent is commonly used. I shall give an example.

Let the protein have a molecular weight of 40 000 and the crystal a size of $0.5 \times 0.5 \times 0.5$ mm$^3$. This crystal contains approximately 2 nMol of protein. If the crystal is soaked in 1 ml solution with a reagent concentration of 10 mM, the amount of reagent in the solution is $10^4$ nMol, an enormous excess in molarity of reagent with respect to protein. However the equilibrium is not determined by the total amount of the reagent and the protein, but by their concentrations. If the protein crystal is regarded as a concentrated solution its concentration can be calculated as 15 mM. This is of the same order of magnitude as the reagent concentration and - if the binding is not very strong - the occupancy of the protein binding site does not reach 100 %. It is then tempting to increase the reagent concentration. However the danger is then that the reagent will react with more sites and the chances of non-isomorphism or even crystal degradation are high.

The soaking time varies between hours and months. A minimum time is required to reach the equilibrium of the reaction and this is determined by a number of factors. First of all by the diffusion of the reagent through the pores in the crystal and this depends on the relative size of the pores and the reagent. Secondly a slight conformational change in the protein may be required for a snugly fitting of the reagent in its binding site and finally there is the chemical reaction itself. Sometimes it is an advantage to use a short soaking time, eg. if the protein molecule presents a great many binding sites to the reagent. If some of them are slow binding sites and others are fast , then if the crystal is soaked for a short time only the fast binding sites react and the chances of maintaining the quality of the crystal are higher. The longer times are simply the result of neglection of the soaked crystals for several weeks or months, but sometimes it is a necessity to soak them for such a long time if the reagent or the protein changes while standing and a suitable reagent or a reactive site on the protein develops in the course of weeks or months. E.g. Pt-compounds can gradually change their ligands. If $K_2PtCl_4$ is kept in an ammonium sulfate solution $[PtCl_4]^{2-}$ can exchange $Cl^-$ for $NH_3$ and $[Pt(NH_3)Cl_3]^-$ or $[Pt(NH_3)_2Cl_2]$ or even $[Pt(NH_3)_4]^{2+}$, which has an opposite charge, can be formed with a concomittant change in reactivity. Protein modification with time is caused by e.g. deamidation of Asn or Gln, or oxidation of SH-groups. This changes the overall charge of the protein and may influence the affinity for charged reagents. The solution can also change slowly, e.g. if an ammonium sulfate solution looses ammonia. Occasionally it

has even been observed that prolonged X-ray radiation can cause the migration of the heavy atom to a different site.

The X-ray intensity change caused by the heavy atoms

Crick and Magdoff (1956) have calculated the intensity change to be expected from the attachment of a heavy atom to a protein:

$$\frac{<\Delta I>}{\overline{I_P}} = \sqrt{2}\sqrt{\frac{\overline{I_H}}{\overline{I_P}}}$$

$<\Delta I>$ is the r.m.s. change in intensity

$\overline{I_P}$ is the average intensity of the protein reflections

$\overline{I_H}$ is the average intensity for the heavy atom alone.

The $\overline{I_P}$- values can be obtained with Wilson statistics:

$$\bar{I} = \sum_i f_i^2$$

Suppose a crystal has one protein molecule in the unit cell and one binding site for a reagent molecule with one mercury atom. $f_i$ is 80 for a mercury atom and on the average 7 for a protein atom. The relative intensity change can then be calculated as a function of the molecular weight (fig. 1). If we estimate conservatively that the intensities can be measured with an accuracy of 10%, then the practical limit for $\frac{<\Delta I>}{\overline{I_P}} = 0.10$. From the curve in fig.1 we derive that the maximum molecular weight is 360 000 if one mercury atom per protein molecule is attached with full occupancy and 90 000 for half occupancy.

The conclusions are rather obvious:

1. The isomorphous replacement method can be applied successfully for the structure determination of even large protein molecules and this has been shown experimentally.
2. The accuracy of the intensity data should be as high as possible for the structure determination of large size protein molecules.

For extremely large structures, such as ribosomes, clusters of heavy atoms in a single molecule have been used, but they will not be discussed here.

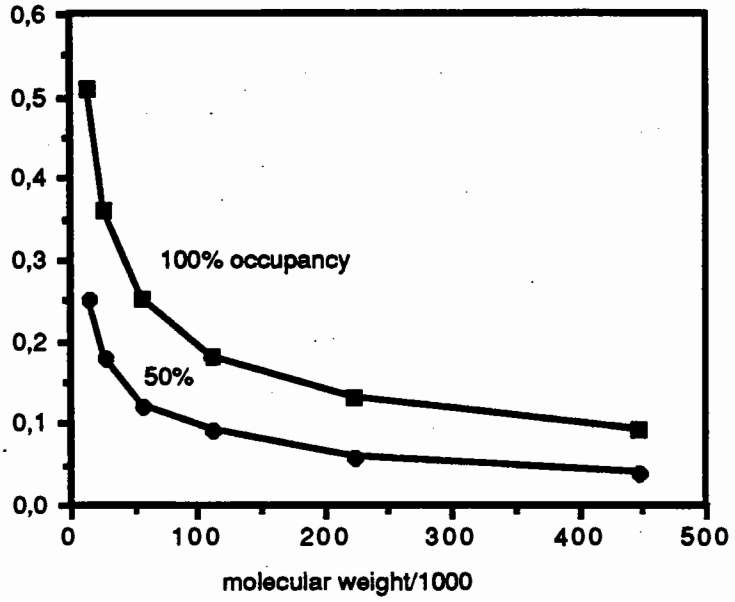



Fig.1. The relative change in intensity caused by 1 Hg atom

Site of attachment

Although the search for a suitable heavy atom reagent is an empirical process, one should employ all available chemical and biochemical properties of the protein. If the protein contains a free SH-group, it is obvious that mercurials should be tried. But if SH-groups are absent mercurials also have a chance. The classical example is myoglobin where PCMS is bound between two myoglobin molecules with the Hg-atom close to a histidine and the sulfonate group near polar residues (Watson et al., 1964). His residues are frequently found as ligands to heavy atoms. But the pH should not be too low because it is the neutral histidine side chain which acts as the ligand.

For proteins containing $Ca^{2+}$ or $Mg^{2+}$-ions it should always be tried to replace the metal ion by a heavier one, notably rare earth ions. If $Ca^{2+}$ is replaced by e.g. $Sm^{3}$, only the difference in electrons is added, that is 41 electrons and not the full number of 59 electrons. But anomalous scattering helps because it is strong for these heavy atoms. The radius of the ions is important because the cavity containing the metal ion is least disturbed if the diameter of the introduced metal ions is close to the diameter of the $Ca^{2+}$- or $Mg^{2+}$-ion. In Table 1 one can see that $Ba^{2+}$ and $Pb^{2+}$ are much larger than $Ca^{2+}$ and are no good replacements for $Ca^{2+}$. It is clear that the flexibility of the binding site also plays a role here. $Ca^{2+}$ can best be replaced

by one of the first rare earth ions because their radius is close to that of $Ca^{2+}$. The radius becomes smaller for the higher elements.

Table 1
The radius of some ions for 6-coordination

| | $Ca^{2+}$ | $Mg^{2+}$ | $Ba^{2+}$ | $Pb^{2+}$ |
|---|---|---|---|---|
| electrons | 18 | 10 | 54 | 80 |
| radius in pm | 114 | 86 | 149 | 133 |

| | $La^{3+}$ | $Er^{3+}$ | $Tm^{3+}$ | $Yb^{3+}$ | $Pr^{3+}$ | $Sm^{3+}$ | $Eu^{3+}$ | $Gd^{3+}$ | $Dy^{3+}$ |
|---|---|---|---|---|---|---|---|---|---|
| electrons | 54 | 56 | 59 | 60 | 61 | 63 | 65 | 66 | 67 |
| radius in pm | 117 | 113 | 110 | 109 | 108 | 105 | 103 | 102 | 101 |

Sometimes heavy atom derivatives of the biological substrate or co-factor are used, but this is less frequently done than one would expect. Examples are:

- Heavy atom derivatives of nucleotides.
- For hexokinase the inhibitor o-iodobenzoyl glucosamine, which binds at the glucose site, has been used ( Fletterick et al., 1975).
- For carbonic anhydrase acetoxymercuri sulphanylamide ( Fridborg et al., 1967).
- For $D^5$-3-ketosteroid isomerase the mercurial 4-acetoxymercuri estradiol, which binds at the steroid binding site (Westbrook et al., 1984).
- For chymotrypsin p-iodomercuribenzene sulphonyl fluoride (Sigler et al., 1966).

One should not neglect the pH of the solution. I mentioned already His, which is a better ligand at higher pH-values. The pH is also important in the binding of charged reagents e.g. $HgI_4^{2-}$, $Au(CN)^{2-}$, $PtCl_6^{2-}$ etc. At higher pH-values, where the protein has a higher negative charge, these reagents react less readily with the protein. This is an advantage if the reaction is too

strong at lower pH-values when too many sites react and non-isomorphism occurs. And if on the other hand, these negative ions do not bind at all or only poorly, the pH should be lowered a little. This is of course only possible if the protein crystal does allow these changes.

Heavy atom salts which are easily hydrolyzed cannot be used in the alcaline pH region. E.g. $UO_2^{2+}$- and $Sm^{3+}$-salts and several others.
If the medium is a water/organic solvent mixture, electrostatic forces are stronger because of the lower dielectric constant and the binding of ionic compounds is stronger. However the organic solvent might be a chelating agent for the heavy atom; this is e.g. true for 2-methyl-2,4-pentanediol (MPD).

If one has no patience, or if only a small number of crystals are available for soaking experiments, the crystals may be soaked in a cocktail of heavy atom compounds, e.g. in a mixture of 5 or 6 of them.

Modification of the native protein

*Chemical modification*

If straightforward soaking does not result in any useful complex, the situation is not completely hopeless. One can try to modify the protein by covalently attaching a heavy atom containing reagent or a potential heavy atom binder, e.g.
p-iodophenylisothiocyanate, or p-iodophenylisocyanate. They react with the $\varepsilon$-$NH_2$ of Lys side chains to form a thiourea or urea derivative, at least at sufficiently high pH.

Another chemical reaction is the iodination of Tyr side chains. They can take up a maximum of three iodine atoms. Iodine has a reasonable number of electrons: 53 but, although it has been applied successfully in a number of cases, it is not a very popular method. The reason is probably reaction with other groups in the protein. In the Dodson group at York University the reagent N-iodosuccinamide has been uesd with some success.

*Genetic modification*

Genetic engineering has opened up new areas for protein modification. For our purpose this is in the first place the introduction of a Cys residue by replacing another amino acid residue. Of course this replacement helps only if the Cys residue is not oxidized very readily, which is a potential danger. Therefore mutants should be prepared and purified in the presence of antioxidants. The new Cys residue should of course not disturb the protein structure, and it should be accessable. This, one does not know beforehand. The best one can do is to replace a residue in a very polar region of the amino acid sequence, which hopefully is a loop at the surface of the molecule. This Cys introduction has been successfully realized in a fragment of colicin A ( Tucker et al., 1989). Five mutants were produced but only one of them gave useful derivatives; the others gave poor crystals or bad derivatives. Elsewhere in this report K. Nagai describes interesting results obtained with the RNA-binding domain of the U1 small nuclear ribonucleoprotein A.

Another biological modification is incorporation of selenomethionine in place of methionine and solving the structure by the multiple wavelength anomalous diffraction technique (MAD).

Some problems and their solution

*Increased radiation damage*

This is caused by radical formation and subsequent chemical reactions. To slow down this process one can try to lower the temperature of the crystal in the X-ray beam. Even a few degrees lower, e.g. a temperature of + 5 °C helps.

*The insolubility of phosphates*

Phosphate buffers are often used for protein crystallisation and soaking. However some heavy metal phosphates are insoluble, eg. those of the rare earths and of uranyl. The phosphate buffer should then be replaced by another suitable buffer, which is usually an organic one.

*Ammonium sulfate*

Ammonium sulfate is the most popular precipitating agent. However it can prevent the binding of heavy metals in two ways. First of all it is in equilibrium with ammonia. Especially at somewhat higher pH values the ammonia concentration in the solution is appreciable and this can act as a ligand for the heavy ion, which might prevent binding to the protein. The solution of this problem is to replace the ammonium sulfate by another salt, e.g. Li- or Cs-sulfate or K- or Na-phosphate, or by PEG. The other problem with ammonium sulfate is the high ionic strenghs. This weakens electrostatic interaction and in this way can prevent the binding of a heavy ion. The solution is to change from ammonium sulfate to PEG.

References

Crick, F.H.C. and Magdoff, B.S. Acta Crystallogr. 9 (1956) 901-908.

Fletterick, R.J., Bates, D.J. and Steitz, T.A., Proc. Nat. Acad. Sci. USA 72 (1975) 38-42.

Fridborg, K., Kannan, K.K., Liljas, A., Lundin, J., Strandberg, B., Strandberg, R., Tilander, B. and Wirén, G. J. Mol. Biol. 25 (1967) 505-516.

Green, D., Ingram, V. and Perutz, M.F. Proc. Roy. Soc. A225 (1954) 287-307.

Sigler, P.B., Jeffery, B.A., Matthews, B.W. and Blow, D.M. J. Mol. Biol. 15 (1966) 175-192.

Tucker, A.D., Baty, D ., Parker, M.W., Pattus, F., Lazdunski, C. and Tsernoglou, D. Prot. Eng. 2 (1989) 399-405.

Watson, H.C., Kendrew, J.C. and Stryer, L. J. Mol. Biol. 8 (1964) 166-169.

Westbrook, E.M., Piro, O.E. and Sigler, P.B. J. Biol. Chem. 259 (1984) 9096-9103.

# Heavy Atom Derivative Screening

A.G.W. Leslie
MRC Laboratory of Molecular Biology
Hills Road
Cambridge CB2 2QH

## Introduction

The purpose of this paper is to give a general introduction to the initial stages of isomorphous replacement, namely the preparation of heavy atom derivative crystals and their evaluation in terms of the degree of heavy atom substitution and the extent of non-isomorphism. The first part of the paper deals with the choice of heavy atom compound and methods for preparing derivatised crystals, while the second part concentrates on the characterisation of putative derivatives.

## 1) Preparation of heavy atom derivative crystals

### a) Choice of compound

As with protein crystallisation, there are few ground rules for determining which heavy atom compounds are likely to provide good derivatives in any particular case, and therefore an essentially empirical approach is usually adopted. However, there are a few points which are generally valid and worth considering when an initial choice of compounds is being made.

Mercury is by far the most popular derivative, simply because it has been used successfully in a large number of structure determinations. This is largely because of the specific nature of the covalent modification of cysteine residues by mercury (Drenth, these proceedings). There are also a large number of mercury compounds available which vary in size, charge and reactivity (see, for example, Chapter 8 of Blundell & Johnson, 1976) . In cases where more than one potential binding site is present, this allows discrimination between the available sites and the possibility of obtaining more than one useful derivative. It should also be remembered that not all mercury derivatives bind covalently: $HgI_4$ for example sometimes binds in hydrophobic pockets.

After mercury, platinum is probably the most widely used derivative. This is again partly due to the availability of different compounds with different net charge (eg $PtCl_4^{2-}$ and $Pt(NH_3)_4^{2+}$ ) and with ligands which are less likely to be replaced by protein

ligands (eg $Pt(CN)_4^{2-}$ ) and which are therefore more likely to interact ionically rather than covalently.

Metal binding proteins present another case where a systematic approach is possible. In several cases (thermolysin, carboxypeptidase, carbonic anhydrase, α-amylase) it has been possible to replace bound Ca or Zn with heavier atoms, and members of the lanthanide series have been particularly useful. This approach is not always successful, as it may be difficult to replace the bound metal without denaturing the protein.

The use of heavy atom labelled substrates or substrate analogues (particularly iodinated analogues) has also been used, although in this case there is the danger that substrate binding may induce a conformational change in the protein resulting in significant non-isomorphism. This problem can sometimes be avoided by taking the protein-substrate (or analogue) complex as the "native" structure rather than the apo-enzyme.

**b) Procedures for preparing derivative crystals**

Soaking is the simplest and most common procedure for preparing derivatives. Crystals are simply transferred to a solution of the heavy atom compound made up in the normal mother liquor (possibly with a slightly higher concentration of the precipitant) and left to soak for a few hours or days (see below for parameters of the soaking experiment).

Under some circumstances it can be preferable to crystallise the derivatised protein rather than soak existing crystals. The principle advantage of this approach is that it allows a greater control over the stoichiometry of heavy atom binding than is possible using soaking. It is therefore most commonly used when dealing with covalently bound derivatives (usually mercury) and when soaking results in crystal damage (cracking or loss of birefringence). By limiting the stoichiometry so that only a subset of the reactive groups (generally thiols) are modified, it may be possible to obtain a useful derivative. The structure determination of Glutaminyl-tRNA synthetase complexed with $tRNA^{Gln}$ (Rould et al., 1989) provides an excellent recent example of this approach. The synthetase contains 10 cysteine residues, and all attempts to prepare mercury derivatives by soaking resulted in crystal cracking. Using co-crystallisation and stoichiometries between 3:1 and 5:1 heavy-atom:protein, 3 useful mercury derivatives were obtained, with binding at between 5 and 7 cysteines.

The disadvantage of co-crystallisation is that the derivatised protein may not crystallise isomorphously with the native protein. However, with the advent of MAD using the L absorption edge of the heavy atom, this requirement is no longer necessary, as the structure of the derivatised form of the protein can be solved (Hendrickson, these proceedings). There are also several examples when the derivatised crystals actually diffract significantly better than the native. A recent example is the small B2 subunit of ribonucleotide reductase (Norlund et al., 1989), where a mercury bound form of the enzyme crystallised in the same space group but with quite different cell parameters to the native enzyme. The mercury form contained one dimer rather than two in the asymmetric unit, and diffracted to 2.1Å resolution rather than 2.9Å, and it was the structure of this modified form that was solved. It should be realised, however, that an improvement of this magnitude is unusual.

A completely different approach that is gaining in popularity is the use of genetic engineering techniques to introduce a potential heavy atom binding site (usually by introducing a cysteine). This approach has been used successfully in a number of recent structure determinations, and will be described in more detail by K. Nagai (these proceedings).

### c) Parameters of the derivative preparation

### Concentration of the reagent

In considering the optimum concentration a distinction has to be made between derivatives that covalently modify the protein and those for which the interaction is primarily ionic or hydrophobic. For the former, very dilute solutions, typically 10-100μM are adequate. The use of higher concentrations is more likely to increase unwanted non-specific interactions, and some workers recommend back-soaking derivatised crystals in mother liquor to minimise non-specific binding. For many of the available mercury compounds the maximum concentration is dictated by their very limited solubility in aqueous solutions, and the use of saturated solutions in such cases is commonplace. For non-covalently binding derivatives the range of concentrations is typically 0.5-20mM, although concentrations at least as high as 200mM have been used, for example in the presence of high salt concentrations or chelating agents such as citrate. In general, the optimum concentration is that which results in the greatest degree of heavy atom substitution with the minimum deterioration in the "quality" of the crystal (generally in terms of the limit of diffraction) and a

minimum of non-specific binding (which only introduces changes in intensity at low resolution). This can only be determined by trial and error. A useful practical approach is to start at fairly high concentrations (5-10mM) and then, if necessary, reduce the concentration by successive factors of 4-5 until conditions are found that produce the desired differences in intensity with a minimal change in the quality of diffraction. The concentration can also be used to produce different relative occupancies at two or more sites, which will yield additional phase information (note that reducing all site occupancies by the same factor does *not* provide additional phase information). Even when this is possible is not normally as useful as finding a new derivative and so this approach is not commonly employed.

## Length of soak

There are a number of studies in the literature where the diffusion of small molecules into protein crystals has been monitored using X-ray diffraction techniques. In one of the earliest, Wyckoff and co-workers (Wyckoff et al., 1967) measured the diffusion of ammonium sulphate into crystals of RNAse-S and observed a half-time of 90secs. However, the half-time for the replacement of bound iodo-uridine phosphate by uridine phosphate was eleven hours. Similar experiments on phosphorylase *b* yielded half-times of about 1 minute for phosphate to 10-12 minutes for maltoheptose (Hajdu et al., 1987). The binding of mercury acetate to a crystal of lysozyme has been followed with a FAST area detector, and the intensity changes are essentially complete after 30 minutes (A.J. Wonacott, personal communication). A similar time scale, namely 0.5-2 hours, is observed for the diffusion of NAD, a much larger molecule than most heavy atom derivatives, into crystals of GAPDH, which can be monitored optically as it produces a change in birefringence (A.J. Wonacott, personal communication).

Compared to the other data, the results for uridine phosphate binding to RNAse-S look somewhat anomalous, but this should serve as a warning that there may be other factors involved in addition to simple diffusion. In this instance, the authors suggest that local depletion due to tight binding may provide an effective lowering of the diffusion rate, but it seems unlikely that this is the sole cause of the much longer half-time. There are certainly other instances where a soak period of several days has produced larger intensity differences than a 16 or 24 hour soak. In some cases this can be ascribed to chemical reactions other than a simple modification of the protein. A good example is the case of $PtCl_4^{2-}$

in ammonium sulphate, where the chlorine ligands are gradually replaced by ammonia over a time period of 1-2 days. In other cases the explanation is less obvious, and unfortunately such evidence is rarely documented or reported in the literature.

In any event, it is very important to characterise the time dependance of the level of substitution and to ensure that the derivative data are collected on crystals that have been prepared under identical conditions. While it may be possible to use the length of soak to control the degree of substitution, this carries the risk of introducing variability between crystals due to differences in crystal size. It is probably better to achieve the same result, if possible, by varying the concentration of the reagant and using a sufficiently long soak time to ensure that the system reaches equilibrium.

In practice, the great majority of derivatives are prepared with soak times of between 1 and 6 days.

### pH and temperature

The same derivative compound may bind to different or additional sites if the pH of the soaking solution is changed. For example, $PtCl_4^{2-}$ binds to a methionine at pH 5.5 but a further site near a histidine is occupied at higher pH (Wyckoff et al., 1967). Different rhenium derivatives of catabolite activator protein have been obtained at pH6.4 and pH8 ((McKay & Steitz, 1981). The stability of the crystals may well be a limiting factor in the application of this approach. Cooling to 4° C has been observed to have a significant effect on the rate of reaction and can therefore be used to control the extent of substitution (Blundell & Johnson, 1976).

### d) Practical Points

Use a low power polarising microscope to observe crystals when transferring them to the soaking solution. Watch out for signs of damage (cracking, dissolving, loss of birefringence).

Use small crystals for the initial soaking trials as they may well crack or dissolve immediately.

Use crystals of a reasonable size for actual data collection, even when screening for optimum soaking conditions, so that the heavy atom signal is not lost in the random errors of measurement. In addition, it has been observed that while small crystals may

survive a given set of soaking conditions, larger crystals will crack under the same conditions. There are also instances when soaking results in the crystals cracking almost immediately but then subsequently reannealing, over a period of hours or days, to give well ordered crystals suitable for data collection.

Many workers have a "favourite" derivative. By all means try this first, but be prepared to try a great many others.

Use freshly prepared heavy atom solutions as far as possible, particularly for $PtCl_4^{2-}$ in ammonium sulphate. Similarly, *cis*-$(NH_3)_2Cl_2Pt$ reacts with $H_2O$ to give the more reactive complex of *cis*-$(NH_3)_2(OH)_2Pt$.

For mercurial derivatives in particular, it is important that the reactive thiols of the protein are not modified in a way that would hinder derivative binding. The use of freshly prepared protein is desirable, or else the crystals can be soaked in a suitable reducing agent (DTT or β-mercaptoethanol) to reduce all accessible SH groups. If this is done it is *essential* that the crystals are washed thoroughly to remove the reducing agent prior to soaking in the derivative solution.

In direct contradiction of the last point, there are at least two recorded cases where mercury derivatives could *only* be prepared in the presence of reducing agent. In one case (McKay & Steitz, 1981) the derivative was 1mM methylmercuri-β-mercaptoethanol while in the second the solution contained 1mM PCMBS, 0.85mM β-mercaptoethanol (Katz et al., 1985). In the latter case, the omission of β-mercaptoethanol led to the crystals disintegrating in a matter of minutes. It seems clear that the role of the reducing agent is in buffering the reactivity of the mercury derivative.

Some derivative crystals are unstable and can loose isomorphism over a period of several days after soaking. This effect can be minimised by reducing the concentration of reagent, but in such cases it is advisable to collect data as soon as possible after the soak.

While there are examples of structures that have been solved using a single derivative, particularly in combination with density modification procedures, this is the exception rather than the rule. It is therefore important to resist the temptation to spend too long computationally massaging a poor set of phases; the time is better

spent back in the lab finding a new derivative ! Many workers have screened over 50 compounds in the search for good derivatives.

## 2) How to detect heavy atom binding

### Optical and density tests

Some derivatives will cause a colour change in the crystal - $PtCl_4^{2-}$ results in a faint straw-like colour. With coloured heavy atom solutions, the crystal should be more strongly coloured than the heavy atom solution if specific binding has occurred. Heavy atom binding will also increase the density of the crystals. If a solution (eg sucrose) is prepared in which the native crystals just float, then derivative crystals should sink (Sigler & Blow, 1965)

### Diffraction experiments

Diffraction data are ultimately required to evaluate how useful a derivative will be. Its usefulness can be characterised by three criteria:

a) Effect on quality of diffraction
b) Isomorphism
c) Degree of substitution

### a) Effect on quality of diffraction

This is the most straightforward criterion to evaluate, as a single "still" (stationary crystal) exposure will reveal both the strength of diffraction and give a good indication of the crystal quality in terms of mosaicity and other types of disorder, which manifest themselves as split spots, unusual spot shapes or streaks (ideally two stills $90^o$ apart in phi should be examined, as some types of disorder are not apparent from a single exposure). Significant loss of diffraction or disorder suggest that a lower concentration of reagent should be used. If this results in no substitution, then it may be worthwhile accepting a lower resolution limit for the derivative, on the grounds that low resolution phases are better than none at all. Derivative crystals are commonly more sensitive to radiation damage than the native protein. This generally only becomes apparent once data collection has begun, but it should be accepted at the outset that more crystals will probably be required for derivative data collection than for the native data.

## b) Isomorphism

A necessary, but not sufficient condition for isomorphism is that the unit cell parameters are unchanged. An accurate determination of the cell parameters provides a rapid and convenient way of checking for possible non-isomorphism, and has the added advantage that the crystal orientation is determined at the same time. Using film methods this can be done using two stills photographs 90° apart in phi (spindle axis rotation) , plus a small angle oscillation photograph if auto-indexing methods are to be used to determine the orientation. Using area detectors, two small wedges of data preferably 90° apart are required for accurate cell dimensions. Crick & Magdoff (1956) have shown that the average fractional change in intensity resulting from cell dimension changes of 0.5% will be 15% at 3Å resolution. Cell parameter changes significantly larger than this will seriously limit the effective phasing power of a derivative at high resolution.

Unfortunately, it is common for the effects of heavy atom binding to be localised to a region of the protein that is not directly involved in lattice contacts, and significant non-isomorphism can arise without any detectable change in cell parameters. In severe cases this can be identified by examining the fractional changes in intensity as a function of resolution, as discussed below.

## c) Degree of substitution

In contrast to the other two criteria, the degree of substitution can *only* be sensibly judged once a statistically meaningful sample of derivative intensities are available. As it will probably be necessary to screen many derivative compounds, each under a range of conditions, it is worthwhile considering the best strategy for obtaining these intensities.

Classically this has been done using precession photography, comparing native and derivative films by eye and looking particularly for "reversals" - that is, two reflections whose relative intensities are reversed between the two photographs. This avoids problems due to the relative strengths of the two exposures and differences in spot shape. The disadvantage of this method is that it is time consuming: a relatively long exposure is required for the precession photograph and the crystal orientation has to be accurately set. In addition, because of the long exposure times, typically only data to 4-5Å resolution are available. The

evaluation itself is qualitative rather than quantitative, unless the films are densitometered and processed.

Some of these objections can be overcome by taking a single oscillation photograph rather than a precession photograph. Because of the relative efficiency of unscreened oscillation photography the exposure can be significantly shorter and data to high resolution are available. Providing that a complete native dataset is available, it is unnecessary to orient the crystal accurately. Thus the combination of two still photographs and an oscillation photograph can provide accurate cell dimensions, a good estimate of the quality of diffraction and sufficient data to estimate the degree of substitution and provide an upper estimate of non-isomorphism.

An even more efficient approach is possible using a 2-D area detector. The strategy is essentially the same as that already outlined, using two segments of reciprocal space to provide both cell parameters and intensity data. In most cases this can be done within a few hours and requires only a minimum of operator intervention. The choice can then be made about collection of a complete 3-D dataset.

The degree of substitution is usually evaluated by calculating the mean fractional difference in structure factor amplitude after the native and derivative data have been put on a common scale. (CCP4 programs ANSC, ANISOSC or RFACTOR).

$$\text{Delta } F/F = \Sigma \mid F_{nat} - F_{deriv} \mid / \Sigma F_{nat}$$

In the absence of errors in the data, delta F will *decrease* monotonically with increasing resolution for true isomorphous differences, but will *increase* with resolution if the differences are due to non-isomorphism.

The question of the expected magnitude of the differences has been addressed theoretically by Crick & Magdoff (1956) who derived an expression for the mean fractional change in intensity as:

$$\text{delta } I/I = \gamma \, (f_H/f_P) \, \sqrt{(N_H/N_P)}$$

where $\gamma = 2$ for centric data and $\sqrt{2}$ for acentric

$f_H$ is the heavy atom scattering factor

$f_P$ is protein atomic scattering factor

$N_H$ is the number of heavy atoms per protein molecule

$N_P$ is the number of protein atoms.

Providing the differences are small, this can be expressed more usefully as:

$$\text{delta } F/F = \gamma/2\ (f_H/f_P)\ \surd\ (N_H/N_P)$$

This is plotted for a variety of molecular weight proteins in Fig.1. This figure clearly shows that for larger molecular weights, data of the highest quality are required if the heavy atom signal is to be significantly above the noise level.

Some empirical results for several derivatives of chloramphenicol acetyltransferase (CAT) (Leslie, 1990) are presented in Fig. 2 which shows the variation in delta F/F with resolution. A number of conclusions can be drawn from this data:

i) Delta F/F always increases with resolution because of random errors in the data. This fact alone cannot therefore be used as an indicator of non-isomorphism, but a dramatic increase at higher resolution (as seen for $PtCl_4^{2-}$ and $Sm(NO)_3$) does indicate significant non-isomorphism. (The $PtCl_4^{2-}$ derivative only provided useful phase information to 4Å resolution). It is worth noting that in the present example this would *not* have been

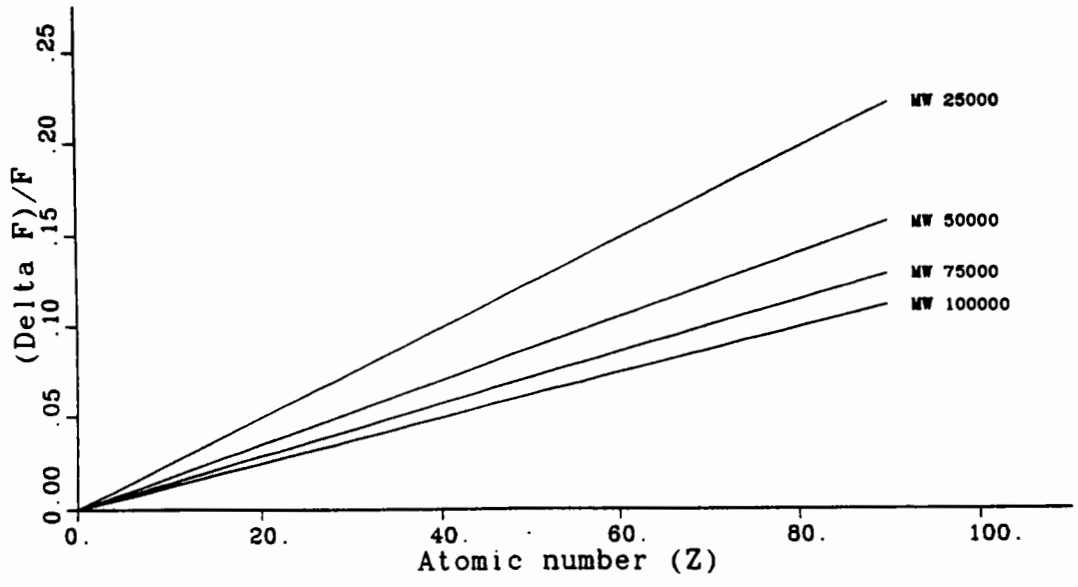



Fig 1. The theoretical mean fractional difference in structure factor amplitude (delta F/F) for proteins of different molecular weight as a function of the heavy atom atomic number (Z). These values assume all protein atoms scatter as nitrogen and a single fully-occupied heavy atom site per molecule. (Crick and Magdoff, 1956)

detected if only data to 3Å resolution were available. Unreasonably large values of delta F/F may also indicate non-isomorphism, but as the number of binding sites is unknown at this stage this is not a very useful criterion in most cases.

ii) Large values of delta F/F do *not* necessarily indicate a good derivative. Thus $Au(CN)_4$ and $PtCl_4^{2-}$ have similar values of delta F/F but the gold derivative is far superior in phasing power, because it is more isomorphous.

iii) Similarly, small values of delta F/F do not necessarily mean that the derivative does not provide useful phase information (compare the iodinated substrate PICM and $Sm(NO)_3$).

iv) The peak in delta F/F at about 7Å resolution arises because of a dip in $F_{nat}$ at this resolution rather than a decrease in delta F. In principle, this peak should be more pronounced if the differences

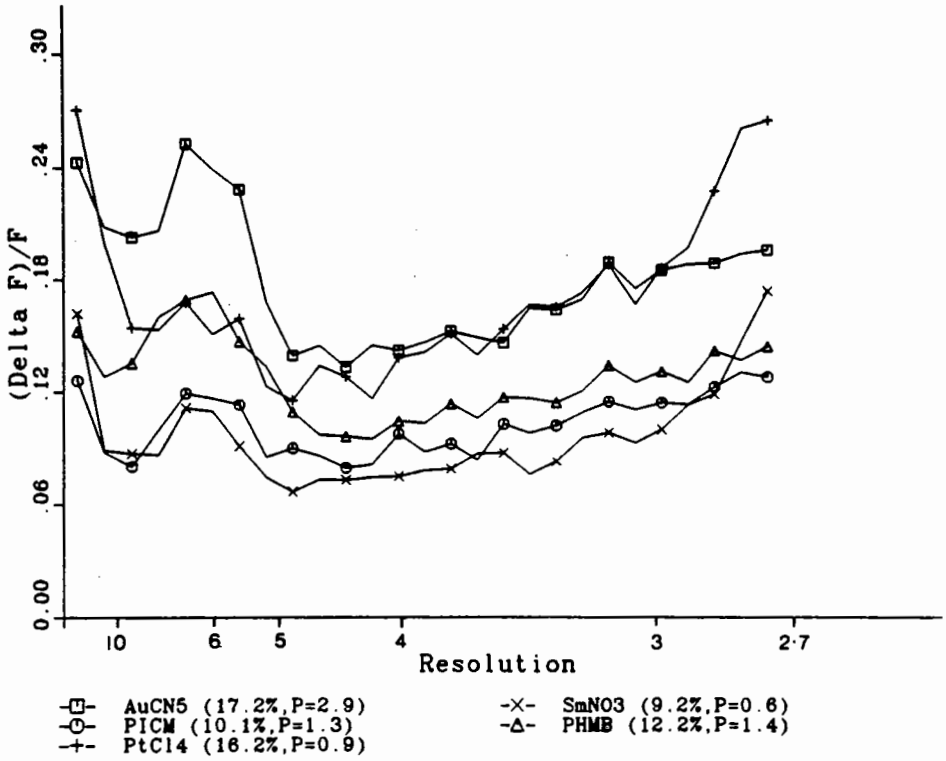



Fig. 2 Mean fractional difference in structure factor amplitude as a function of resolution for several derivatives of chloramphenicol acetyltransferase (Leslie, 1990). The molecular weight in the asymmetric unit is 25000. The derivatives were gold cyanide (AuCN5), para-iodo-chloramphenicol (PICM), platinum chloride (PtCl4), samarium nitrate (SmNO3) and PHMB. The values given in parentheses for each derivative are the overall delta F/F and the final phasing power of each derivative (to give an indication of derivative quality).

are due to heavy atom substitution rather than non-isomorphism. This may be an indicator of the extent of non-isomorphism at this resolution.

This analysis shows that a relatively small wedge of data will reveal the data quality, the size of the differences and, less reliably, the degree of non-isomorphism. It is important, when evaluating the magnitude of the differences, to have a firmly established base-line derived from the comparison of a small (independent) wedge of **native** data with the full native dataset. Even with data of high quality, a value of delta F/F of 5-8% is typical when comparing native with native in this way. Values significantly above this may make it difficult to detect single site substitution of the lighter derivatives such as iodine and the lanthanides. Values of delta F/F significantly greater than those expected on the basis of the measurement errors in F may also be indicative of variability amongst native crystals. It is of crucial importance to establish that the native data is reproducible prior to screening for potential derivatives. Additional information on the degree of substitution can be obtained by examining the anomalous differences, which has the advantage of being independant of the degree of isomorphism but is subject to much larger experimental errors. In this case the data from centric zones (where there is no anomalous signal) can be used to establish a base-line for comparison. Unfortunately at present there is no CCP4 program available to do this analysis routinely.

If this preliminary analysis looks promising, then full 3-D data collection is warranted. It is undoubtedly the case that a much less ambiguous indication of the quality of a derivative can be gained by calculating a difference Patterson than is possible by an analysis of delta F/F, even if only low resolution data is available. This is clearly demonstrated in the case of the CAT data (Fig. 3), where the difference betwen the platinum and gold derivatives is apparent even at 6Å resolution. Similarly the iodinated substrate gives a much cleaner Patterson than the samarium derivative.

When an area detector is available it is therefore worthwhile collecting a 3-D dataset for any derivative that looks at all promising. The resolution of this dataset can be chosen so that, if at all possible, a complete dataset can be collected overnight from a single crystal. The resulting Pattersons can then be used to decide if a full high-resolution 3-D data collection is desirable, when the emphasis should be on the quality of the data rather than limiting the collection time.

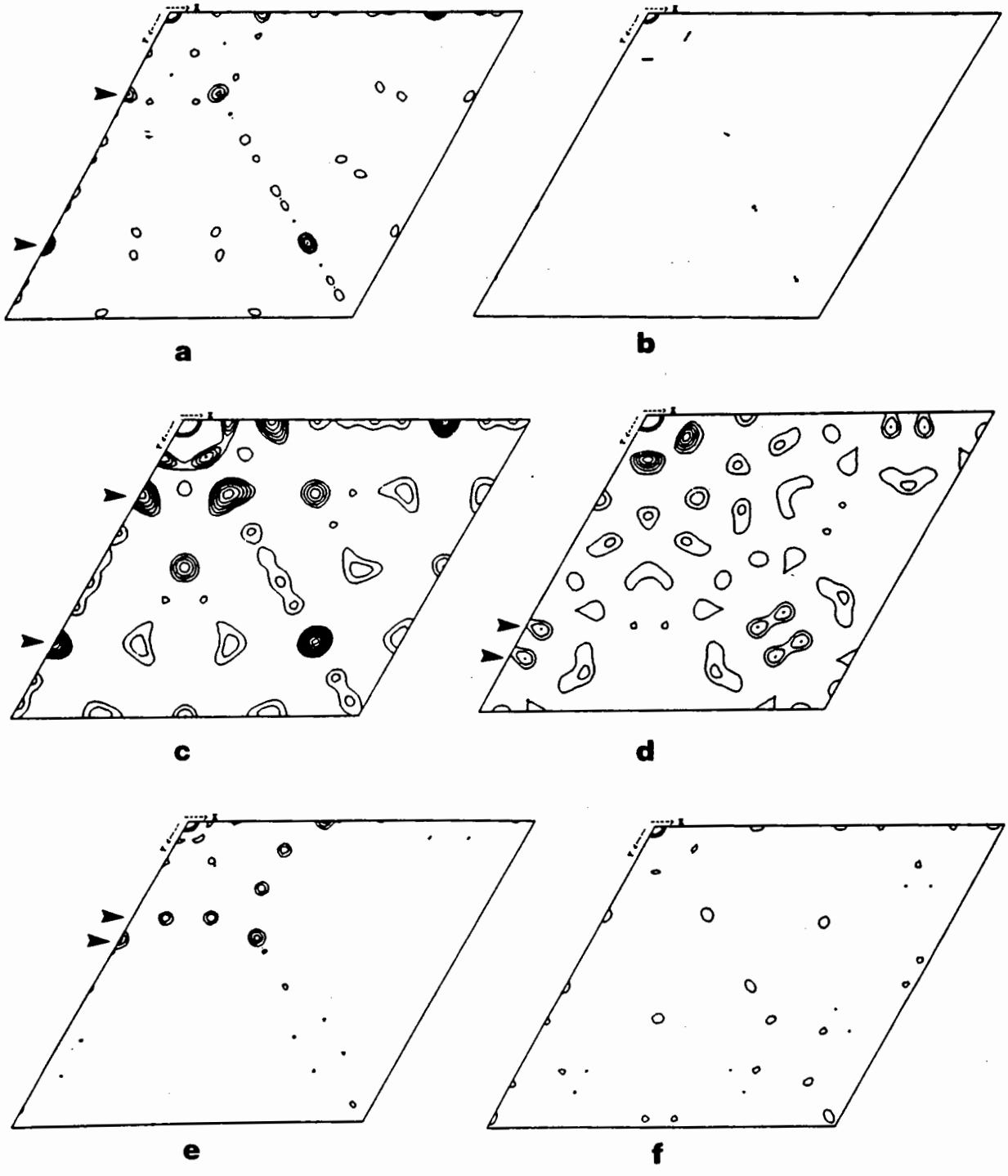



Fig 3. The Harker section (z=0) of the 3-dimensional difference Pattersons of CAT derivatives: (a) 2.7Å resolution gold cyanide, showing two clear sites (b) 2.7Å resolution platinum chloride; no sites are visible (c) 6Å resolution gold cyanide; although noisier, the two sites are still very clear (d) 6Å resolution platinum chloride; in this case the double site, which was lost in the 2.7Å Patterson because of non-isomorphism, is visible (labelled A and B) (e) 2.7Å PICM; two sites can be seen (f) 2.7Å samarium nitrate; no clear indication of sites.

## Acknowledgements

I would like to thank all my colleagues at LMB, and in particular Phil Evans, Jade Li and Peter Brick (Imperial College) for their help and advice in preparing this manuscript.

## References

Blundell, T.L. and Johnson, L.N. *Protein Crystallography*, chapter 8, Academic press, London (1976)

Crick, F.H.C. and Magdoff, B.S. Acta Cryst. **9**, 901-908 (1956)

Hajdu, J., Acharya, K.R., Stuart, D.I., McLaughlin, P.J., Barford, D., Oikonomakos, N.G., Klein, H. and Johnson, L.N. EMBO J. **6**, 539-546 (1987)

Katz, B.A., Ollis, D. and Wyckoff, H.W. J. Mol. Biol. **184**, 311-318 (1985)

Leslie, A.G.W. J. Mol. Biol. **213**, 167-186 (1990)

McKay, B.M. and Steitz, T.A. Nature **290**, 744-749 (1981)

Norlund, P., Uhlin, U., Westergren, C., Joelsen, T., Sjoberg, B-M and Eklund, H. FEBS Let **258**, 251-254 (1989)

Rould, M.A., Perona, J.J., Soll, D. and Steitz, T.A. Science **246**, 1135-1142 (1989)

Sigler, P.B. and Blow, D.M. J. Mol. Biol. **14**, 640-644 (1965)

Wyckoff, H.W., Doscher, M., Tsernoglou, D., Inagami, T., Johnson, L.N., Hardman, K.D., Allewell, N.M., Kelly, D.M. and Richards, F.M. J. Mol. Biol. **27**, 563-578 (1967)

## Heavy Atom Location using SHELXS-90

George M. Sheldrick, Institut für Anorganische Chemie,
Universität Göttingen, D-3400 Göttingen, Germany

### Abstract

The algorithms employed for automatic location of heavy atoms via the Patterson function in the "small-molecule" program SHELXS-90 are discussed, and are illustrated by applications of this program to heavy atom derivatives of macromolecules.

### Introduction

As increasing numbers of small-molecule crystallographers move into macromolecules, small-molecule programs are being increasingly (mis)used to solve macromolecular problems such as the location of heavy-atom sites. This generally involves using isomorphous or anomalous $|\Delta F|$ and $\sigma(\Delta F)$ instead of $|F|$ and $\sigma(F)$, hoping that the approximation involved will simply add random noise. Careful scaling of the derivative and native data, pruning of statistically unreasonable $\Delta F$'s, and good estimated standard deviations are essential to the success of this approach. I shall concentrate on the application of a new small-molecule program, SHELXS-90, which can be used for Direct or Patterson structure solution by changing one instruction (TREF or PATT). It will also recognize macromolecular $\Delta F$-data (from cell volume and contents) and set appropriate defaults. For both Direct and Patterson Methods the $\Delta F$-data are first merged to create a unique set and then normalised to give $E$-values.

### Direct Methods

Although small-molecule Direct Methods are able to solve small proteins from native data alone if unusually high-resolution data are available (say 1·1 Å; both SAYTAN and SHELXS-90 solve Avian Pancreatic Polypeptide with 302 unique atoms excluding H and solvent), the assumption of equal *resolved* atoms will generally require derivative $\Delta F$-data. Unfortunately there

are two fundamental difficulties with the application of Direct methods to $\Delta F$-data. The first is that the negative quartets:

$$\phi_h + \phi_k + \phi_l + \phi_m \simeq \pi$$

(if $E_h$, $E_k$, $E_l$ and $E_m$ are very large and $E_{h+k}$, $E_{h+l}$ and $E_{h+m}$ are very small) are useless, because the $|\Delta F|$ values represent lower bounds on their true values, and so are unsuitable for identifiying the very small $E$-values. On the other hand they do serve to identify correctly the *largest* $E$-values, and so the old triplet formula works well. The second problem is the estimation of probabilities for the triplet formula for use in figures of merit: what should replace the $1/N$ term (where $N$ is the number of atoms per cell) when $\Delta F$-data are used? Most of the advances in Direct Methods in the last 15 years exploit either the weak reflections or more sophisticated formulae for probability distributions, and so are wasted on $\Delta F$-data. The bottom line is that modern Direct Methods programs are no more successful on $\Delta F$-data than their predecessors. Nevertheless, Direct Methods will tend to perform better in space groups with (a) translation symmetry, (b) a fixed rather than floating origin and (c) no special positions, e.g. $P2_12_12_1$ but not C2 or R3. An advantage of Patterson methods operating in real space is that potential atoms on special positions (a common false solution for Direct Methods on $\Delta F$-data) can be rejected at an earlier stage.

**Patterson Interpretation**

At least a quarter of organic and inorganic structures are solved by the heavy atom method, and the preparation of heavy atom derivatives is almost obligatory for proteins if no suitable search model is available for molecular replacement. Despite this, there is a striking lack of user-friendly computer programs, general for all space groups and types of problem, for the automatic location of heavy atoms, in sharp contrast to the situation in small-molecule direct methods. This seems to be simply an historical accident; the necessary theory has been well understood for at least 40 years, and the computer resources required are modest.

My first attempt at automating the location of heavy atoms by interpretation of the Patterson function was, not unnaturally, simply a computerised version of how I used to solve Pattersons by hand. Since

this method was distributed widely in the form of the public-domain program SHELXS-86, I shall begin by describing it briefly, and then explain why I have abandoned it in favour of a completely different approach in SHELXS-90.

### The SHELXS-86 Approach to Patterson Interpretation

A major problem in the automation of Patterson methods is the storage of the full three-dimensional Patterson map, which can easily cause conflicts between hardware independence and efficiency, two essential ingredients of user-friendly software. This problem was solved in SHELXS-86 by calculating a rather sharp Patterson (based on coefficients $[E^3F]^{1/2}$), and generating a sorted list of peaks which was used for most but not all of the subsequent calculations. As a final check, the Patterson function was recalculated from the reflection data at selected points only. The use of this peak-list made the program very efficient and made it easy to incorporate space group symmetry in a general way; however the Patterson function contains ca. $N^2$ peaks, where $N$ is the number of heavy atoms, so peak overlap frequently caused problems.

The peak-list was analysed to find sets of atomic coordinates **x** for which all Harker vectors (i.e. vectors between an atom and all its symmetry equivalents) were present in the peak-list, to within a tolerance determined by the resolution of the data. If this condition was not fulfilled for at least one of the heavy atoms, the approach was doomed to failure. In the space group $P\bar{1}$, every peak was considered to be a potential 2**x** vector. In P1, just one "atom" was generated, at the origin.

The peak-list was used again to find a figure of merit for each potential atom generated as described above. First a list of coordinates **x'** was found for which every cross-vector with **x** was present in the peak-list; all cross-vectors and Harker vectors involving pairs of "atoms" **x'** were then calculated, and compared with the peak-list to estimate the figure of merit. The potential atom **x** with the best figure of merit was then used as a basis for the remaining calculations. If required, one or more extra atoms were added from the **x'** list. This is desirable in low-symmetry space groups and essential in P1. Alternatively the user could input one or more trial atoms himself to bypass the above part of the procedure. This was often necessary when several different potential atoms had acceptable figures of

merit and the program chose the wrong one. All Patterson peaks **X** were combined with the heavy atom(s) and their symmetry equivalents **x*** to generate further possible atomic positions **x"** = **x*** ± **X**. The list of the **x"** which arose most frequently was optimised and used with the heavy atom(s) **x** to generate a "crossword table" for chemical interpretation by the crystallographer. This valuable feature has been retained in SHELXS-90 and so will be discussed in detail below. Alternatively the user could opt for automatic assignment of scattering factor types to the heavy atoms and partial structure expansion based on *E*-Fourier recycling to find the rest of the structure.

With luck the fully automatic approach led to spectacular holes in one - in one cubic case the three unique hydrogen atoms were located automatically in the first job as well as all the heavier atoms - but failure at any stage of the procedure led to the production of nonsense for all subsequent stages. Experience using SHELXS-86 for small-molecule heavy atom problems showed that the fully automated procedure was successful at least half the time, and that about half the remaining structures could be solved with a little manual help (i.e. choosing a different heavy atom from the list of candidates printed out by the program).

With hindsight the basic weaknesses of this method were the reliance on the peak-list derived directly from the Patterson function, and the difficulty in choosing the "correct" heavy atom on which to base the remaining calculations; however the latter problem could have been overcome, at the cost of computer time, by a "multisolution" approach. It was particularly difficult to decide between (pseudo-)special sites (e.g. the origin) and general sites for this first heavy atom. In high symmetry space groups the number of Patterson peaks, and hence potential overlap, could be large, and it was also more likely that at least one Harker vector was missing from the peak-list for each of the heavy atoms, resulting in no solution.

### The SHELXS-90 Approach - Vector Superposition Minimum Function

The Patterson vector superposition minimum function approach employed in SHELXS-90 was suggested in the early 1950's, and a review of the early literature may be found in chapter 11 of Buerger's (1959) book "Vector Space", where the method is referred to as the "vector shift" method. At the time it did not prove very effective at solving unknown

structures, probably because it was usually used in projection, and because multiple vector superpositions were attempted rather than the single superposition used in SHELXS-90. The approach was revived by Richardson & Jacobson (1987) who showed that in three dimensions, computer analysis of a single vector superposition could be used to solve quite complex problems. A similar general approach is adopted in SHELXS-90, but (unlike Richardson & Jacobson) all operations are carried out in real space, and no attempt is made to average the different solutions (images) which are found.

The vector superposition approach can be demonstrated effectively in two dimensions by the use of overhead transparencies. Figure 1 shows a demonstration structure consisting of five atoms with one symmetry element (a horizontal mirror plane). The Patterson function may be regarded as a vector map, so we can generate it by placing each atom in turn at the origin, and marking the positions of the other atoms on a superimposed overhead transparency. This Patterson function (Figure 2) has retained the mirror plane and also possesses a centre of symmetry, since for each interatomic vector A->B there must be an inverse vector B->A. This way of generating the Patterson makes it clear that it contains *N* (here 5) images of the structure, plus *N* inverted images (because of the inversion centre). We can use the reverse procedure to deconvolute the Patterson. If we superimpose two copies of the Patterson, displaced from each other by one of the vectors in the Patterson (Figure 3), then two of the images will coincide exactly, as must two of the inverted images.

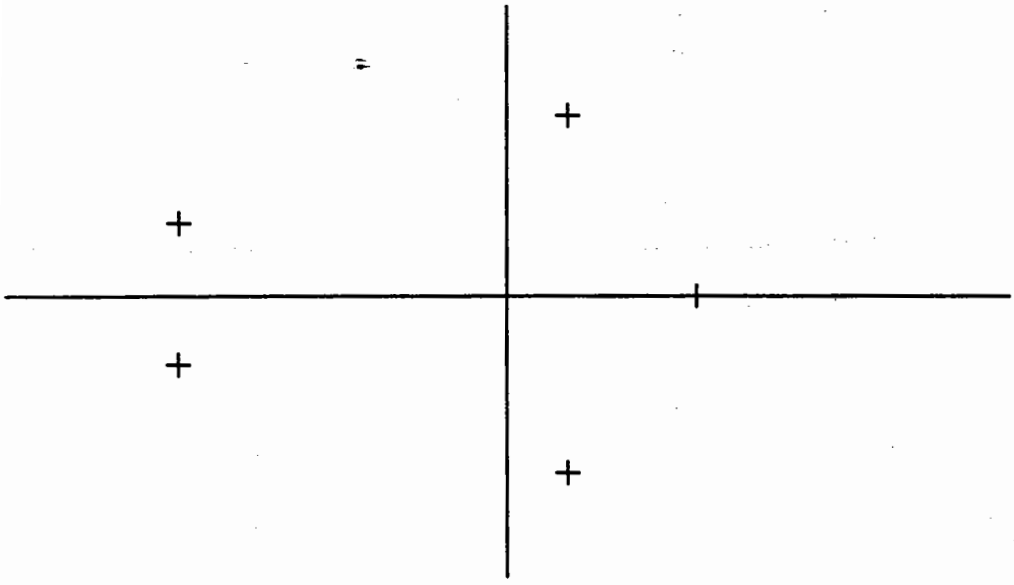


Fig. 1. Two-dimensional demonstration structure with m-symmetry

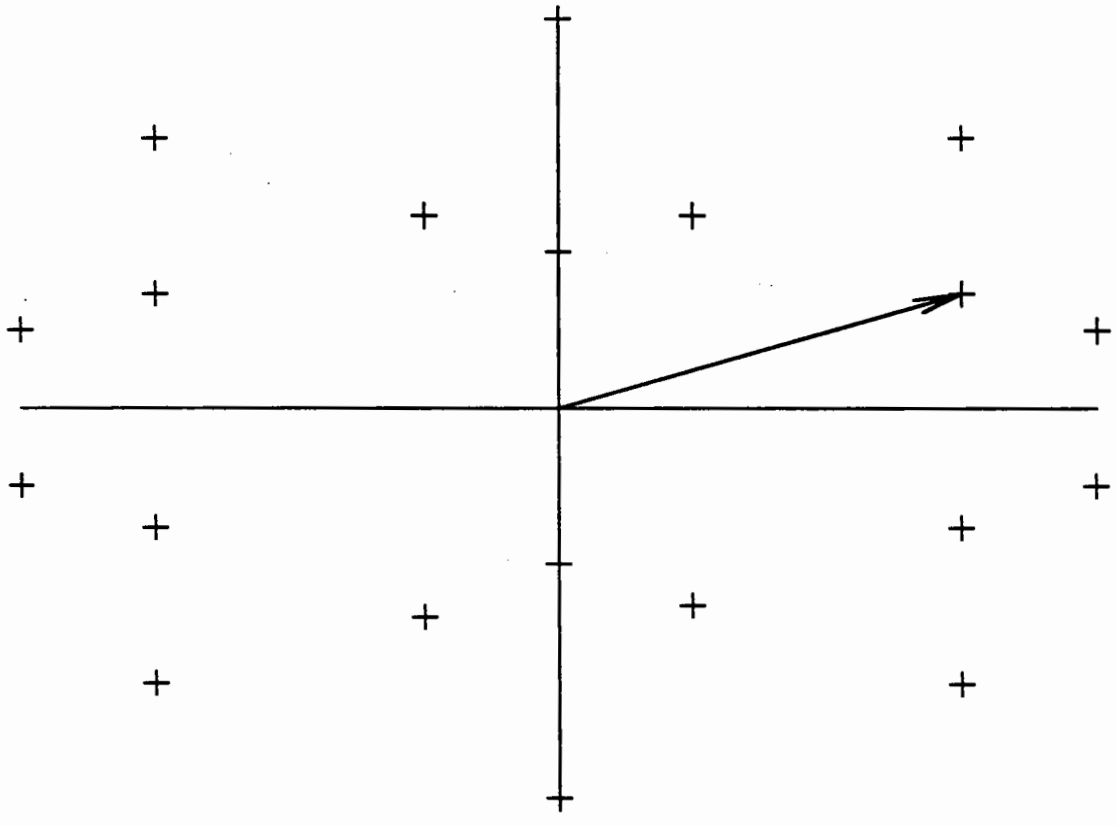


Fig. 2. The Patterson function of the structure shown in Fig.1. The m-symmetry has been retained and an inversion centre added. A suitable superposition vector is indicated (and used to generate Fig.3).

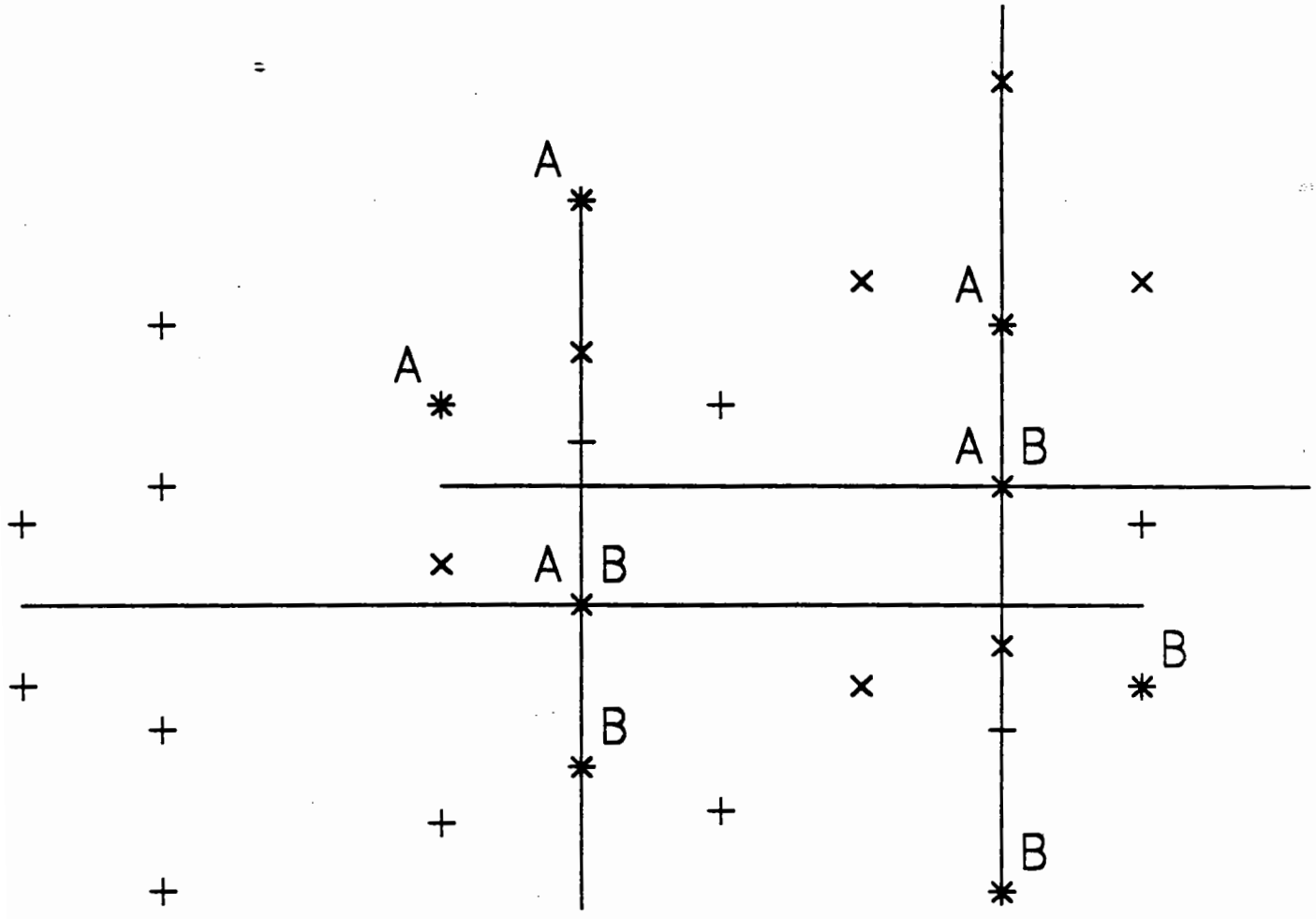



**Fig. 3. The superposition map (*) consists of a true image B and an inverted image A. The mirror has gone. but we may locate A or B by finding a group of atoms with m-symmetry**

Thus by forming the "superposition minimum function", which possesses peaks only where peaks are present in both maps, we get a much cleaner map which only contains ca. $2N$ peaks rather than the ca. $N^2$ of the original Patterson. Note that this map has retained the centre of symmetry but lost the mirror plane; in three dimensions all space group symmetry is lost but an inversion centre is always present. Note also that the points corresponding to the origins of the two Pattersons must belong both to the image and to the inverse image. If a multiple vector were used for the superposition (there are none in this simple demonstration) then in general a multiple image ($4N$ or more images) would result.

In the special case when a **2x** vector in a centrosymmetric space group is used for the superposition, the image of the structure will coincide with the inverse image and the superposition map is (in theory) the true structure, with its origin midway between the origins of the two Patterson functions. In all other cases the image is shifted by an unknown vector from its true origin. The problem thus reduces to the (relatively tractable) one of finding an origin shift which gives as many atoms as possible which are related to one another by the symmetry operations of the space group. In the case of the demonstration structure, the image (or its inverse) can readily be located by moving the transparency vertically until a group of atoms are found with a horizontal mirror plane of symmetry. We can then isolate our structure of five atoms (one of which happens to lie on a special position on the plane) by discarding the remaining three atoms, for which there are no symmetry equivalents. The procedure used in SHELXS-90 may therefore be summarised:

1. Select one peak from the sharpened Patterson as a superposition vector.

2. Calculate and then peaksearch the superposition minimum function based on this vector.

3. Find possible origin shifts of this map which give the best agreement with the space group symmetry.

4. For each acceptable origin shift, accept atoms for which most symmetry equivalents are present in the peaklist, attempt to assign atomic numbers based on average peak heights, and display as a "crossword table".

Each of these four stages will now be considered in more detail; they may be repeated for several different superposition vectors in the same job.

**1. Selection of a Suitable Superposition Vector**

As in SHELXS-86, the asymmetric unit of the Patterson is calculated in SHELXS-90 with coefficients $[E^3F]^{1/2}$ on a grid determined by the resolution of the data, and the coordinates of the strongest peaks are found by parabolic interpolation. We find that these coefficients give appropriate sharpening whilst the contribution from $F$ serves to damp out series termination effects, assuming that the data have been measured out to the effective resolution limit.

It is essential that the vector chosen for superposition corresponds to a real heavy-atom to heavy-atom vector. For simple structures this condition will be fulfilled by all the strongest peaks. We find it useful to require the vector chosen to have a length greater than some preset minimum, say 1.8 Å for small molecules and 8 Å for protein $\Delta F$-data; parallel aromatic groups or repeated secondary structure patterns may give rise to strong peaks at shorter distances. If the length or height of a suitable vector can be predicted, this information should be taken into account; for example there will often be P-P vectors at about 6 Å in polynucleotides.

Whereas a single (and hence weaker) heavy-atom to heavy-atom vector should give a superposition map containing only one image and its inverse, stronger vectors may be multiple vectors which will give rise to multiple images in the superposition map, making its interpretation more difficult. If the vector overlap is only approximate, further errors are introduced. Thus peaks which are higher than expected should usually be avoided, especially if they have (pseudo-)special coordinates. If there is only one heavy atom in a non-centrosymmetric space group, a Harker vector has to be used; otherwise a general vector is preferred because overlap tends to be more severe in the Harker sections. In difficult cases several different starting vectors should be tried and the results compared.

## 2. Calculation of the Vector Superposition Minimum Function

After choosing a suitable superposition vector **U** the sharpened Patterson is calculated with its origin shifted to **-U/2** (by applying a phase shift to the coefficients) and again with its origin shifted to **+U/2.** At each point of the common grid, the lowest of the two Patterson values is chosen; no grid point interpolation is required. The resulting "superposition minimum function" is then peak-searched in the usual way (on the fly, holding the last three layers in memory). The effective "space group" of this function is $P\bar{1}$ (or $C\bar{1}$ etc. if the lattice is centred); theoretically it should correspond to one or more images of the structure, plus their inverses.

## 3. Origin Location

The procedure used to find potential origin shifts depends on the space group. In the space group P1, this stage is skipped, and the strongest peaks of the superposition function are used directly to construct the "crossword table". Hand interpretation of this table may enable the separation of the image from its inverse.

In all centrosymmetric space groups, all peak positions, and all midpoints between two peaks, are potential origins (the space group setting with the origin on an inversion centre is always employed). The corresponding shifts are weighted according to the product of the two peak heights; similar indications are combined and the shift coordinates averaged. A list of the most strongly indicated shifts is passed on to the symmetry test stage.

All other space groups require algebraic analysis of 2 (or 3) symmetry operators and the corresponding peak positions. For example in the space group $P2_1$ we must find two peaks in the superposition map which differ by 0.5 in *y*:

$$( x + dx, y + dy, z + dz ) = ( x_1, y_1, z_1 )$$

$$( -x + dx, y + 0.5 + dy, -z + dz ) = ( x_2, y_1 + 0.5, z_2 )$$

Thus $dx = ( x_1 + x_2 ) / 2$ and $dz = ( z_1 + z_2 ) / 2$; $dy$ is not required.

As in the centrosymmetric case, equivalent indications are combined and a list of the strongest shift indications is passed on to the symmetry test.

For each potential origin shift, a symmetry test is performed. Atoms for which (say) more than 70% of the expected symmetry equivalents can be found in the shifted superposition peak list are retained for use in the crossword table, and contribute to the figure of merit SYMFOM which measures the consistency with which the peak positions and heights satisfy the space group symmetry. High symmetry space groups also enable local averaging (over symmetry equivalents) of the atomic coordinates and atomic numbers (derived from the peak heights).

**4. The "Crossword Table".**

This table is designed to summarise the information contained in the Patterson function relevant to a particular set of trial atoms. It enables the crystallographer to apply his chemical knowledge to the recognition of the correct solution without complications caused by symmetry operations. The "crossword table" takes the following general form:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $Name_1$ | $At_1$ | $x_1$ | $y_1$ | $z_1$ | $s_1$ | $d_{11}$ | | | |
| | | | | | | $P_{11}$ | * | | |
| $Name_2$ | $At_2$ | $x_2$ | $y_2$ | $z_2$ | $s_2$ | $d_{22}$ | $d_{21}$ | | |
| | | | | | | $P_{22}$ | $P_{21}$ | * | |
| $Name_3$ | $At_3$ | $x_3$ | $y_3$ | $z_3$ | $s_3$ | $d_{33}$ | $d_{31}$ | $d_{32}$ | |
| | | | | | | $P_{33}$ | $P_{31}$ | $P_{32}$ | * |
| | etc. | | | | | | | | |

Where $At_i$ is the estimated (relative) atomic number of atom i which has coordinates $x_i$, $y_i$, $z_i$. $s_i$ is a factor which is unity for an atom on a general position and less than unity for a special position; it is equal to the reciprocal of the number of general equivalent positions which coalesce to give the special position in question. $d_{ij}$ is the minimum distance between an atom i and the atom j or any of its symmetry equivalents (taking lattice translations into account as well). $P_{ij}$ is the minimum value of the sharpened Patterson function at any of the vectors between atom i and atom j or its symmetry equivalents (i.e. the same vectors which are used to

find $d_{ij}$). These Patterson function values are recalculated at the corresponding points in Patterson space directly from the reflection intensity data, since the Patterson is not stored. It should be noted that the first column of *d,P*-pairs ($d_{ii}$, $P_{ii}$) involves the *self*-vectors between atom i and its symmetry equivalents; the remaining pairs ($d_{ij}$, $P_{ij}$) involve *cross*-vectors. This arrangement facilitates annotation by hand; if the asterisk at the end of a row is replaced by the name assigned to that atom, it will be seen that this name also heads the column corresponding to the same atom.

Interpretation of this crossword table thus requires assigning element types to atoms so that the minimum distances make chemical sense, and the relative values of the Patterson minimum function are not less than the product of two atomic numbers involved (they may of course accidentally be greater). If an atom is involved in several small or zero Patterson minimum function values involving other strongly indicated atoms, it may be spurious and should be eliminated. In some cases (pseudosymmetry or space group P1) a multiple image may be resolved by repeated elimination of atoms in this way, until the most strongly linked (large $P_{ij}$) set remains. The following examples should illustrate the interpretation of the crossword table.

**Figures of Merit**

Three separate figures of merit are calculated for each solution (crossword table). All three are expressed as percentages and should be large for a good solution. SYMFOM measures how well the solution obeys the symmetry of the space group and PATFOM the internal consistency of the Patterson minimum function values in the crossword table. The correlation coefficient (Fujinaga & Read, 1987) is often used in molecular replacement. SYMFOM and PATFOM are scaled to 99·9 for each superposition vector, but the correlation coefficient is useful for comparing solutions for different superposition vectors. To sort the solutions for a given vector, the three figures of merit are combined to give CFOM.

## Examples

The input for the first example consisted of a file BARNAU.HKL containing *h*, *k*, *l*, $\Delta F$ and $\sigma(\Delta F)$ in FORMAT (3I4, 2F8.2) plus the following free format instruction file BARNAU.INS:

```
TITL Au iso in P3(2) (E.Dodson)
CELL 1.54178 58.97 58.97 81.58 90 90 120
LATT -1
SYMM -Y, X-Y, .66667+Z
SYMM -X+Y, -X, .33333+Z
SFAC N AU
UNIT 120 9
PATT
HKLF 3
```

The program is started with the command: SHELXS BARNAU
and produces a listing file BARNAU.LST from which the following has been extracted:

```
Patterson superposition on vector 1     0.1822  0.2053  0.6653

  360 Superposition peaks employed, maximum height  79.2
  and minimum height  11.5  on atomic number scale

Heavy-Atom Location for  Au iso in P3(2) (E.Dodson)

  5076 reflections used for structure factor sums


Solution 1  CFOM = 13.2  PATFOM = 99.9  Corr.Coeff.= 36.4  SYMFOM = 99.9

Shift to be added to superposition coordinates:  0.3723  0.6413  0.0000


Name  At.No.   x       y       z    s.o.f.  Minimum distances / PATSMF
                                                        (self first)
AU1   84.4  0.1318 0.0495 0.5458  1.000  29.64
                                          51.7

AU2   80.9  0.2832 0.5398 0.6667  1.000  29.45   27.48
                                          85.5    67.5

AU3   47.8 -0.2261 0.3629 0.6308  1.000  28.90   35.00 35.45
                                          42.6    36.5  45.5

AU4   21.7  0.0246 0.0134 0.6418  1.000  27.28    9.61 37.98 30.80
                                          12.4     0.4  14.1   0.0

AU5   21.0 -0.3769 0.3885 0.7866  1.000  28.57   25.83 27.49 16.01 36.30
                                           1.5     0.0  32.6   0.0   1.3
   .. etc. ..
```

This is a very clean derivative (Barnase data kindly donated by Eleanor Dodson) with three gold sites. Inspection of the isomorphous crossword table shows that only the first three suggested sites have suitable values for the Patterson Minimum function for the self-vectors (first column) and cross-vectors (triangular table). Each entry consists of a minimum distance, under which is given the Patterson minimum function based on the same vectors. The remaining suggested atoms are "noise". Direct Methods and Direct and Patterson Methods on the anomalous differences also give the same three sites and no others.

The second example (also run with default settings) involved the *native* data for Crambin (see next page).

An interesting feature of this table is the three distances (involving atoms 1-5 and 9) in the range 2.02 to 2.11 Å; a disulphide bridge in a protein should have a length of about 2.06 Å. These are also the six atoms which, taking a global view, give the most consistent (largest) set of Patterson minimum function values $P_{ij}$, though there is one zero value (between atoms 4 and 9) and $P_{ii}$ for S4 is only 1.1 However S...S vectors are so close to the noise level in protein Pattersons that such fluctuations are scarcely surprising. These six atoms were used in a routine SHELXS partial structure expansion using the commands "TEXP 1000 6" and "FMAP 10", which produced a tangent formula phase extension followed by iterative *E*-Fourier recycling. A peaksearch of the resulting final *E*-map revealed all but 14 of the unique protein non-hydrogen atoms (plus some of the solvent atoms); it should be noted that since Crambin is a mixture of two compounds which differ at two residues, two side chains are effectively disordered.

It must be emphasised that this unexpectedly simple solution of the phase problem for a small protein (all the calculations were performed on a personal computer!) was only possible because excellent quality very high-resolution low-temperature data were available (courtesy of Prof. Håkon Hope). A certain amount of luck was also involved; the program had automatically selected a vector corresponding to one of the three S-S bonds for the superposition. However it would not have been difficult to select this vector from the Patterson peak-list by hand. In multisolution mode (PATT 20), superposition vectors 1, 4 and 13 gave all six sulphurs; two other vectors gave 5 of them.

Patterson vector superposition minimum function for Crambin 0.9A low-T in P2(1)

Patt. sup. on vector 1  0.9538  0.0425  0.0250  Height 23.  Length 2.12

Maximum = 38.12, minimum = -109.93  highest memory used = 15524 /318011  disk mode

500 Superposition peaks employed, maximum height 18.4
and minimum height 2.5 on atomic number scale

Heavy-Atom Location for Crambin 0.9A low-T in P2(1)

19982 reflections used for structure factor sums

Solution 1 CFOM = 2.18 PATFOM = 99.9 Corr. Coeff. = 14.8 SYMFOM = 99.9

Shift to be added to superposition coordinates: 0.4148 0.0000 0.4401

| Name | At.No. | x | y | z | s.o.f. | Minimum distances / PATSMF (self first) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S1 | 17.1 | 0.8932 | 0.0208 | 0.4592 | 1.000 | 12.82 | | | | | | | | |
| | | | | | | 10.6 | | | | | | | | |
| S2 | 12.8 | 0.8004 | -0.0241 | 0.0984 | 1.000 | 18.49 | 8.90 | | | | | | | |
| | | | | | | 7.3 | 9.3 | | | | | | | |
| S3 | 11.9 | 0.9361 | -0.0211 | 0.4212 | 1.000 | 11.16 | 2.10 | 9.04 | | | | | | |
| | | | | | | 7.3 | 9.5 | 31.1 | | | | | | |
| S4 | 11.7 | 0.5768 | 0.0819 | 0.0477 | 1.000 | 11.35 | 15.80 | 9.38 | 16.89 | | | | | |
| | | | | | | 1.1 | 7.5 | 13.0 | 12.4 | | | | | |
| S5 | 10.9 | 0.7571 | -0.0775 | 0.1016 | 1.000 | 18.49 | 9.84 | 2.02 | 10.21 | 8.00 | | | | |
| | | | | | | 19.2 | 16.3 | 6.5 | 8.1 | 15.1 | | | | |
| S6 | 10.7 | 0.8975 | 0.0993 | 0.2539 | 1.000 | 16.56 | 4.81 | 5.71 | 4.61 | 13.82 | 7.39 | | | |
| | | | | | | 0.5 | 6.5 | 1.7 | 8.8 | 0.0 | 10.0 | | | |
| O7 | 9.3 | 0.8161 | 0.0187 | -0.1343 | 1.000 | 18.49 | 9.64 | 5.30 | 11.14 | 10.67 | 6.08 | 9.37 | | |
| | | | | | | 0.0 | 10.9 | 0.0 | 6.8 | 1.0 | 4.8 | 0.0 | | |
| O8 | 8.4 | 1.0450 | 0.0196 | 0.6901 | 1.000 | 13.05 | 8.01 | 11.52 | 7.46 | 18.32 | 11.89 | 8.21 | 10.16 | |
| | | | | | | 0.9 | 2.9 | 0.0 | 5.4 | 0.0 | 0.0 | 0.0 | 4.8 | |
| O9 | 8.1 | 0.6090 | 0.1713 | 0.0526 | 1.000 | 13.02 | 14.91 | 8.65 | 16.00 | 2.11 | 7.66 | 12.62 | 9.87 | 16.49 |
| | | | | | | 19.0 | 16.3 | 8.5 | 3.7 | 0.0 | 5.0 | 3.6 | 0.0 | 0.0 |
| O10 | 7.6 | 0.8389 | -0.2104 | 0.0658 | 1.000 | 16.36 | 10.00 | 3.86 | 9.50 | 11.98 | 4.22 | 7.48 | 6.22 | 8.75 | 11.73 |
| | | | | | | 4.4 | 11.7 | 0.0 | 3.4 | 0.0 | 6.3 | 0.0 | 3.4 | 1.5 | 1.2 |

## Conclusions

The direct location of all the sulphur atoms will clearly never be a suitable method for routine solution of protein structures, but on the other hand this example forcefully demonstrates the power of the method for the location of heavier atoms. SHELXS-90 is able to solve all the structures in a test bank of about 50 small-molecule heavy-atom problems; in some cases a little manual intervention is required or there are multiple solutions because the Patterson is genuinely ambiguous. It failed to solve one known and several unknown derivatives (mainly bromine) of polynucleotides, and fails on two protein isomorphous datasets which both have multiple partially occupied sites in high symmetry space groups. For a few other heavy-atom derivatives of proteins it disagreed about some of the more speculative minor sites. However it is easy to use and is general for all space groups and computers (given a FORTRAN Compiler!).

I am particularly grateful to E. Dodson, P.R. Evans, R. Hilgenfeld, H. Hope, A. Leslie, A. Sielecki, I. Tickle, N. Walker and many others for supplying me with suitable test data.

## References

M.J. Buerger (1959). *Vector Space.* New York: Wiley.

M. Fujinaga & R.J. Read (1987). *J. Appl. Cryst.* **20**, 517-521.

J.W. Richardson & R.A. Jacobson (1987). *Patterson & Pattersons,* eds. J.P.Glusker, B.K. Patterson & M. Rossi. Oxford: I.U.Cr. and O.U.P., pp. 310-317.

# Locating heavy atom sites by automatic Patterson search - GROPAT

by

**Yvonne Jones and David Stuart**
Laboratory of Molecular Biophysics, Rex Richards Building,
South Parks Road, Oxford. OX1 3QU

## Introduction.

In this paper we will discuss the interpretation of difference Patterson maps. General points will be raised followed by details of one Patterson search method (GROPAT), a 'toolkit' style approach which we are developing. Finally, some of the points will be illustrated by examples from the structure determination of Tumour Necrosis Factor.

Before any phase information for a protein may be determined by the method of isomorphous replacement the position(s) of the heavy atom site(s) in the unit cell must be correctly identified for at least one heavy atom derivative. Classically the key tools comprise an isomorphous difference Patterson map and a copy of the International Tables for Crystallography. The difference Patterson function is based on the squares of the differences in amplitudes between two sets of data. Given native and heavy atom derivative data the difference Patterson map consists of peaks corresponding to the vectors between the heavy atom sites. The basics of the interpretation of Patterson functions are described in (1). A Harker section comprises a plane in the Patterson map in which vectors between equivalent atoms related by a particular crystallographic symmetry operator must appear. A clear mean fractional isomorphous difference, when it arises from one or two major heavy atom sites per asymmetric unit, in a low symmetry space group, should produce a few strong peaks on the Harker sections of the difference Patterson map. These self vectors can be trivially interpreted to yield the heavy atom positions. Given more than one site the solution should be confirmed by the presence of peaks at general positions in the difference Patterson map corresponding to the cross vectors between the crystallographically independent sites. For many space groups these cross vectors are vital in defining the relative origin of the different heavy atom sites. For instance, in space group $P2_1$ there are 4 alternative origins in the xz plane and any number in the y direction.

In working with difference Patterson maps several points should be standard practice:

i) rejection of outliers in the data (eg reflections for which the difference in amplitude is greater than 3 x mean difference). Since a difference Patterson map is calculated using the square of the differences between pairs of numbers a few rogue differences can have a disasterous effect. When the differences are small

it is important to scale the data carefully, for instance heavy atom substitution may anisotropically disorder the crystal (2).

ii) origin removal, by subtracting $<F^2>$ (ie the mean square structure factor) from each term contributing to the Patterson the value at the origin of the function becomes zero. Large ripples around the origin generated from the origin peak by series termination errors may otherwise obscure genuine features in the map. The ripples are caused by the Fourier transform of the reciprocal space sampling envelope convoluted with the Patterson function. Hence these series termination effects are the same for each peak and are essentially the same as those observed in an ordinary Fourier synthesis. However, given a moderate number of heavy atom sites in a moderately high symmetry space group, the Harker peaks in the Patterson difference map will be only a few percent of the height of the origin peak and thus all too easily swamped in the associated ripples.

iii) always plot the Patterson peaks predicted by a given solution on the actual Patterson synthesis.

For the less obvious cases interpretation may be aided by:

i) comparison of difference Patterson maps calculated for independent resolution ranges (eg 10 to 7Å and 7 to 5Å) or data sets. This can serve to highlight the significant peaks.

ii) sharpening the observed structure factor amplitudes (by applying a negative B factor to artificially counter the observed fall off in the mean structure factor amplitude with resolution). This literally sharpens the Patterson peaks.

iii) extra information from self-rotation functions (the orientation of non crystallographic symmetry axes) and/or the low resolution (eg 15 to 7Å) native Patterson (simple translations between molecules in the asymmetric unit). A non-crystallographic rotation axis will give rise to a non-crystallographic Harker section in the Patterson map (normal to the symmetry axis and passing through the origin). Inspection of such planes (for example by simply orientating the difference Patterson map appropriately using FRODO on a graphics workstation) will highlight peaks corresponding to vectors between any heavy atom sites which conform to the non-crystallographic symmetry; but note that some or all of the heavy atom sites may not obey this symmetry.

Given a non-trivially interpretable difference Patterson map one should objectively assess the possible reasons for the difficulty. For most useful cases the observed isomorphous differences should exceed those expected on the basis of random noise. Also the derivative should actually be isomorphous (the cell dimensions should show little change compared to those of the native and the magnitude of the mean differences should not increase with resolution). It is

important to keep in mind the difference between complexity (often arising from high crystallographic and/or non-crystallographic symmetry) which may obscure real information in a difference Patterson map and straightforward noisiness indicative of a poor quality derivative which is unlikely to provide adequate phasing power and should probably be consigned to the bin. The calculation of a simple correlation function on the peaks above $3\sigma$ between difference Patterson maps calculated for independent resolution ranges or data sets serves as a good measure of their useful information content (see below).

As the true complexity of the problem increases automatic Patterson search procedures become increasingly useful and ultimately, for most crystallographers, essential.

## GROPAT - a toolkit for Patterson map interpretation

GROPAT comprises a flexible set of programs which we wrote (and continue to develop) to aid in the interpretation of difference Patterson maps. Many of the features simply correspond to automated versions of procedures which would be employed manually to solve a less complex difference Patterson map. However, rather than select peaks on the Harker sections as candidate self vectors, all possible heavy atom positions are explored in a grid search. This search generates an exhaustive list of putative single heavy atom sites (default 200 sites), ranked in terms of the agreement of their predicted peaks with the actual difference Patterson map. This list then forms the basis from which to develop a unique (if necessary multiple site) solution.

The criteria on which the putative single sites are ranked, and all further judgements of 'goodness of fit' decided, are central to the toolkit. They comprise of a combination of six separate measures. These are based on the values of the actual difference Patterson function at the map pixels close to the predicted vectors (the peak value of the Patterson function for the pixels in the close vicinity is selected as the observed pixel value at the predicted peak):

i) mean of pixel values at predicted peaks

ii) product of pixel values at predicted peaks

iii) minimum pixel value at predicted peaks

iv) root mean square of pixel values at predicted peaks

v) mean probability for pixel values at predicted peaks

vi) product of probabilities for pixel values at predicted peaks

Several of the above represent well established ways of quantifying 'goodness of fit', however, the last two measures (based on probabilities) are nonstandard.

**Figure 1.** The GROPAT toolkit for Patterson map interpretation.

**GROPAT1**
gridsearch to generate ranked list of possible single sites
either judged on self peaks or given one or more fixed sites on self and cross peaks

**DUP**
to check for essentially duplicate solutions in one list
or to find duplicate solutions between several lists

**PATREFINE**
to refine each of the solutions in the list by a finer
grid search to optimise agreement with the Patterson map

**GROPAIR**
to check all pairwise combinations of sites in a list on crossvectors,
if necessary (eg monoclinic spacegroups) includes
grid search (eg on y) to find relative origin

**CLUSTER**
to search through single site list
for certain patterns of non-crystallograpically related sites

**PREPAT**
to plot predicted Patterson vectors for selected sites(s)
applicable at each of the above stages

An ideal height for a true difference Patterson peak ($P_{ideal}$) and the amount of deviation ie error to be reasonably expected from this ideal height ($\sigma$) may be used to give a measure of the probability that any particular pixel position corresponds to an inter-site vector by

$$\exp[-(|P_{ideal} - P_{obs}|)^2 / \sigma^2$$

where $P_{obs}$ is the value of the actual difference Patterson function at this pixel. This provides a powerful gauge of the apparent reasonableness of a solution by drawing on further information derived from a 'global' assessment of the difference Patterson map. Two pieces of information are required:

i) the apparent 'error' level in the map $\sigma$ (may be based on the statistics quoted for the original map calculation)

ii) an estimate for the height expected for a single peak $P_{ideal}$ (based on inspection of the Patterson)

The six measures are self-normalising. The mean value ($M_{mean}$) for each of the six (taken over many positions in the grid search) and the corresponding mean deviation ($Dev_{mean}$) are used to give

$$\text{normalised measure} = (\text{measure} - M_{mean}) / Dev_{mean} .$$

The normalised measures are combined either as sums or products as appropriate for independent or dependent quantities to yield an overall 'probability' that the proposed site(s) is/are a true solution.

For GROPAT the general philosophy is to consider initially a large number of ranked possible solutions and to gradually winnow out the false solutions by applying increasingly stringent filters. Thus lists of ranked solutions may be operated on by successive programs. If desired some provision for a locked search incorporating non-crystallographic symmetry exists but rather than impose such rigorous constraints on the solutions at an early stage we prefer the more gradual route of introducing restraints, particularly in view of the frequent departure of heavy atom sites from non-crystallographic symmetry. The current toolkit is represented in figure 1.

## TNF - a case study

Tumour necrosis factor (TNF) is an important polypeptide mediator of inflammation and the cellular immune response. Mature human TNF is an unglycosylated protein of 157 amino acids (relative mass 17350 Da). In solution three such TNF subunits are tightly associated to form the biologically active trimer.

The structure of TNF was determined from crystals of space group $P3_121$ with unit cell dimensions a=b=165.9Å, c=93.1Å (3). The crystallographic asymmetric unit contains two independent TNF trimers (ie six subunits) and the fractional volume of solvent in the crystal is 65%. The crystals diffract to a resolution

limit of 2.9Å. Self-rotation functions failed to reveal the two non-crystallographic threefold axes. In retrospect this may be understood as a consequence of the combination of a high symmetry (trigonal) space group, the nature of the TNF trimer and the somewhat special directions of the threefolds (4). A range of standard heavy atoms were tried. Also compounds with internal threefold symmetry, for example $K_2HgI_4$ in excess KI (5) and WAC (6) (a gift from Prof. R. Huber), were tested in the hope of restricting binding to positions on the non-crystallographic threefolds. With hindsight the $K_2HgI_4$ almost worked according to plan but was a poor, multiple site, derivative (see below). A potential target for limiting the number of sites to one per subunit was presented by the presence of a single disulphide bond. This disulphide was reduced and reformed in the presence of mercury acetate $Hg(Ac)_2$ in accordance with the method of reference (7). Native and derivative data were scaled using an established seven parameter, anisotropic B factor method (2) and a number showed encouraging isomorphous differences (table 1).

**Table 1.** Putative heavy atom derivatives for TNF.

| Deriv. - | R merge‡ (all data) | Av. iso. diff. (oo to 5Å) | Corr. 3$\sigma$ * 10-7 v 7-5Å | Corr.† 3$\sigma$ 10-7Å | Useful Deriv. |
|---|---|---|---|---|---|
| $K_2Pt(CN)_4$ | 8.0 (3.4Å) | 16.1 | 11.8% | 43.6% | Yes |
| WAC | 5.4 (5.0Å) | 11.8 | 3.9% | - | No |
| $K_2HgI_4$ | 9.5 (5Å) | 27.5 | 27.0% | - | Yes |
| $K_2PtCl_4$ | 9.7 (3.4Å) | 11.2 | 12.1% | - | No |
| $KAu(CN)_2$ | 10.3 (3.4Å) | 18.5 | 9.7% | - | No |
| $Hg(Ac)_2$ | 7.3 (3.5Å) | 19.1 | 31.9% | 99.7% | Yes |
| Noise | - | - | 4.8% | 0.0% | - |
| Perfect | - | - | 83.0% | - | - |

‡

$$Rmerge = \frac{\sum_j \sum_h |I_{h,j} - < I_h >|}{\sum_j \sum_h < I_h >}$$

* correlation between Pattersons based on 10-7Å and 7-5Å data.

† correlation between Pattersons based on independent 10-7Å data sets.

With six molecules per asymmetric unit in a $P3_121$ cell even a single heavy atom binding site per subunit results in over 1000 peaks at close to noise level

in the difference Patterson map. Thus visual inspection of the difference Patterson maps proved of little efficacy in assessing whether a heavy atom compound had yielded an isomorphous derivative (figure 2). To distinguish between difference Patterson maps which consisted purely of random noise peaks and those containing actual information a simple correlation coefficient

$$\frac{\sum_u D_u^1 \times D_u^2}{\sum_u |D_u^1 \times D_u^2|}$$

(where $D_u^1$ and $D_u^2$ are the values at corresponding pixels u in maps 1 and 2 and the summations are over all pixels above a given threshold, typically $3\sigma$) was calculated between pairs of origin removed difference Patterson maps (4). Difference Patterson maps based on independent data for the same derivative should show a high level of agreement. As may be seen from table 1 the assessment of information content provided by this method tallies well with the eventual usefulness of the putative derivative. Overall the analysis correctly highlighted $Hg(Ac)_2$ as the most promising derivative. Ultimately this derivative provided much the best phase information.

Further analysis of the difference Patterson maps was undertaken using the Patterson search techniques incorporated in GROPAT, which was developed for the purpose. Origin removed, sharpened Patterson maps were used, generally for the resolution range 20 to 6Å. A grid search over the asymmetric unit first established the top 200 positions for single sites judged to generate Harker peaks in best agreement with the actual Patterson map. These sites were then refined individually on a finer grid to optimise agreement with the Patterson function. All possible pairwise combinations of these 200 sites were then assessed in terms of the agreement of their crosspeaks with the actual Patterson and the resulting best 200 pairs were listed. This analysis was carried out for the six putative derivatives listed in table 1. The $Hg(Ac)_2$ derivative immediately stood out. It yielded a clear pattern of interconnected pairwise solutions with all possible combinations of the single site solutions originally ranked 3, 57, 65, 67 and 124 grouped at the top of the pairs list. The match between observed and predicted Patterson map peaks is illustrated in figure 2b. Fourier difference maps based on phases from these 5 sites revealed a sixth site. Sites 3, 65 and 67 obeyed threefold symmetry forming an equilateral triangle with sides of about 11Å. Sites 57 and 124 with the addition of the new site formed a second similar but independently orientated triangle. The $Hg(Ac)_2$ had indeed bound to each subunit at the disulphide.

Difference Fouriers revealed $K_2HgI_4$ and $K_2Pt(CN)_4$ as poorer quality multiple site derivatives. In retrospect the major site in the $K_2Pt(CN)_4$ derivative (on a crystallographic twofold) and one of the minor sites, had been picked up by GROPAT and ranked 4 and 117 in the single site list. For the $K_2HgI_4$ derivative GROPAT had been successful in detecting three of the strongest sites which it ranked 1, 18 and 78. Of these 18 was identical to the major (twofold) site of the $K_2Pt(CN)_4$ derivative. 1 and 78 were indeed sites on the trimer threefolds with 1 corresponding to site 117 in the $K_2Pt(CN)_4$ derivative. However the poor

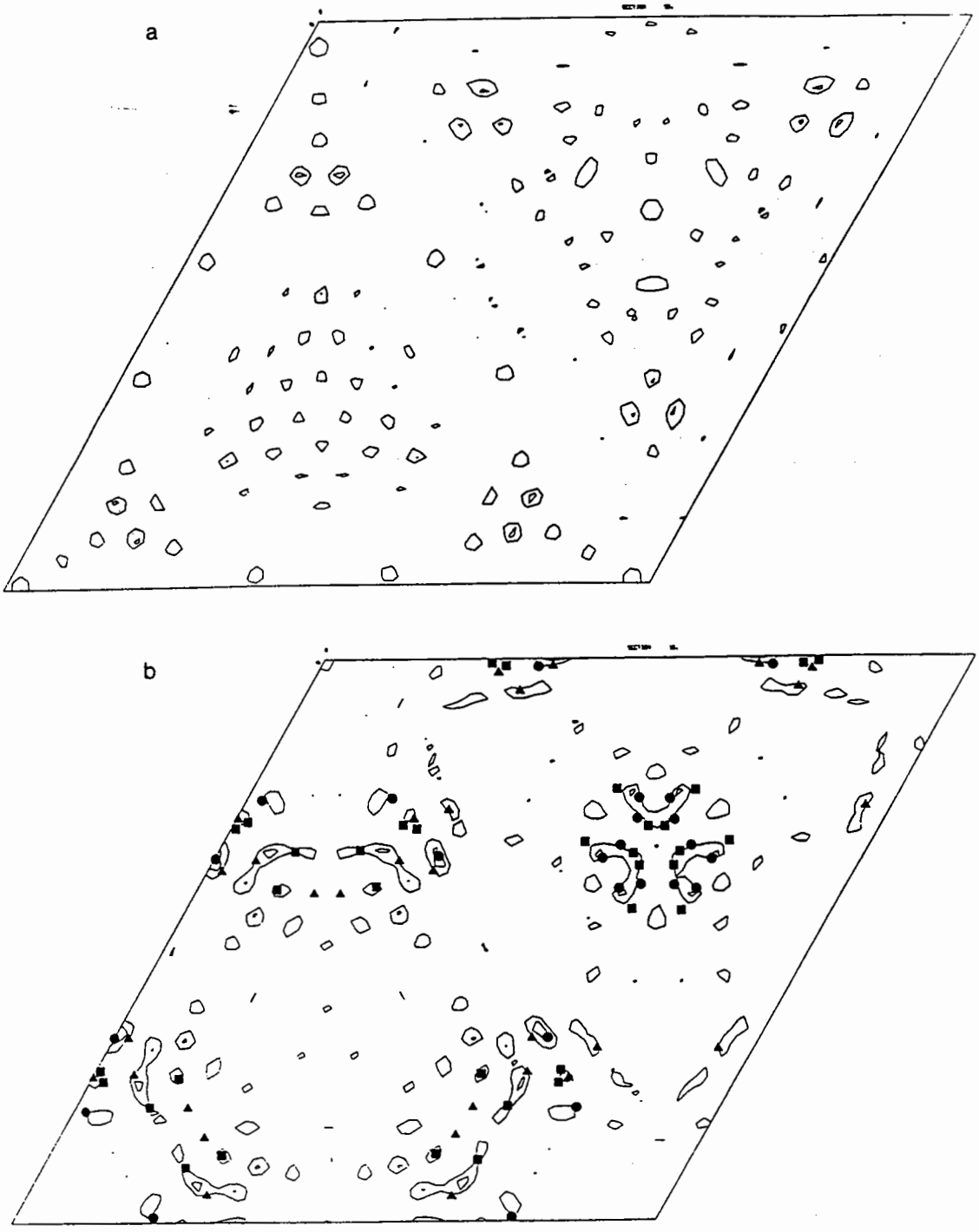



**Figure 2.**

a) 1/3 w Harker section of a difference Patterson map (10-7Å contoured at about $1.5\sigma$) calculated for the $KAu(CN)_2$ putative derivative. This heavy atom proved to be useless (the differences in the structure factor amplitudes were due to non specific binding occuring at many sites).

b) 1/3 w Harker section of a difference Patterson map (10-6Å contoured at about $1.5\sigma$) calculated for the $Hg(Ac)_2$ derivative. Peaks predicted by the five sites found by GROPAT are superimposed. Triangles equal Harker peaks, circles equal crossvectors and squares equal crossvectors actually centred on neighbouring map sections (+ or - 1/48 w).

quality of the $K_2HgI_4$ and $K_2Pt(CN)_4$ derivatives, combined with their complete failure to conform to the non-crystallographic symmetry and the binding at crystallographically special positions had led us to place little confidence in the original analyses of the difference Patterson maps.

## References


1. Stout G.H. and Jensen L.H. (1968). X-ray structure determination. The Macmillan Company, London.

2. Stuart D.I., Levine M., Muirhead H. and Stammers D.K. (1979) J. Mol. Biol. **134** 109-142

3. Jones E.Y., Stuart D.I. and Walker N.P.C. (1989) Nature **338** 225-228

4. Jones E.Y., Walker N.P.C. and Stuart D.I. (1991) Acta. Cryst. Section A. In press.

5. Petsko G.A. (1985) Meth. Enzym. **114** 147-156

6. Ladenstein R., Bacher A. and Huber R. (1987) J. Mol. Biol. **195** 751-753

7. Ely, Girling, Schiffer, Cunningham and Edmundson (1973) Biochemistry **12** 4233

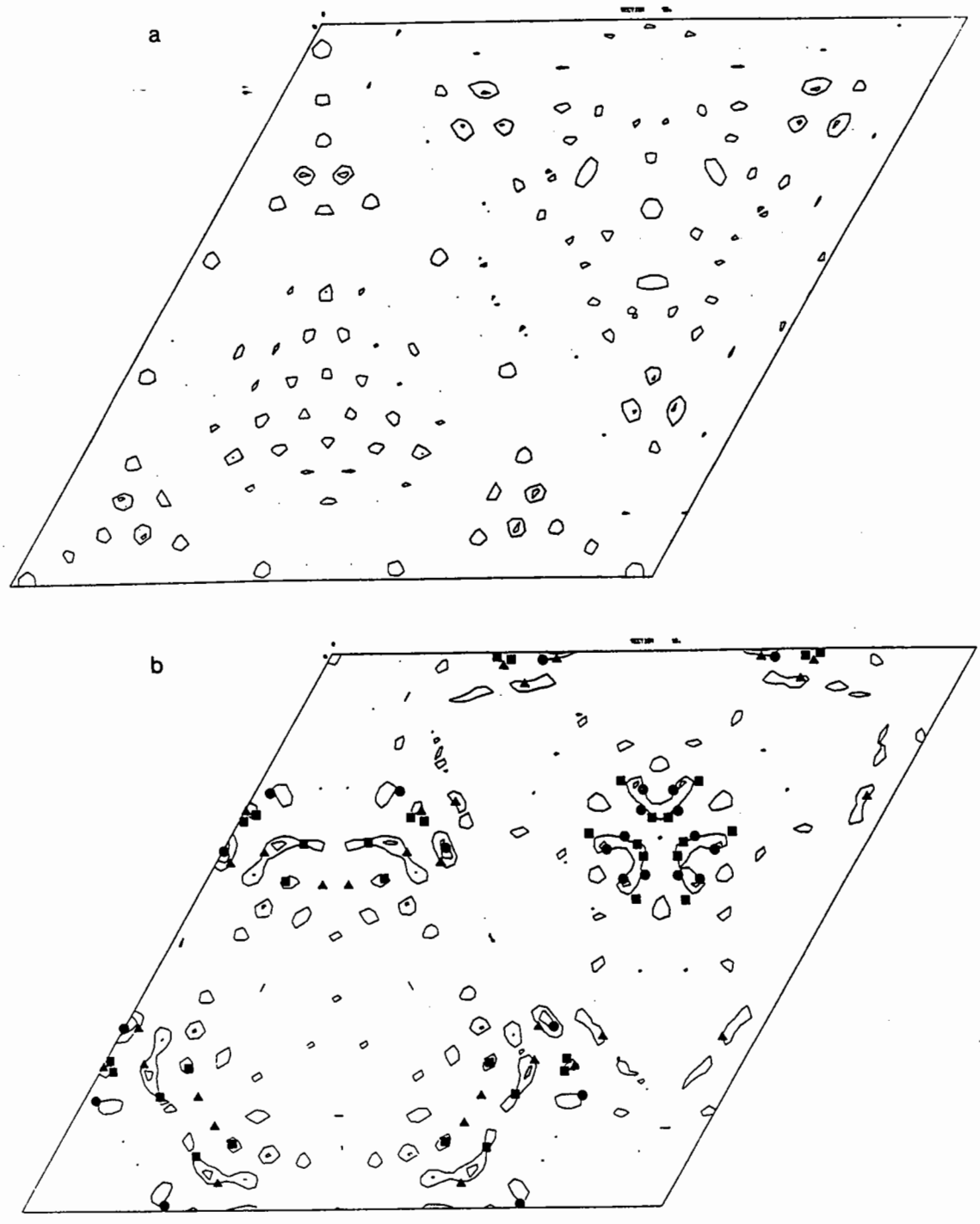



**Figure 2.**

a) 1/3 w Harker section of a difference Patterson map (10-7Å contoured at about $1.5\sigma$) calculated for the $KAu(CN)_2$ putative derivative. This heavy atom proved to be useless (the differences in the structure factor amplitudes were due to non specific binding occuring at many sites).

b) 1/3 w Harker section of a difference Patterson map (10-6Å contoured at about $1.5\sigma$) calculated for the $Hg(Ac)_2$ derivative. Peaks predicted by the five sites found by GROPAT are superimposed. Triangles equal Harker peaks, circles equal crossvectors and squares equal crossvectors actually centred on neighbouring map sections (+ or - 1/48 w).

# Refinement of heavy-atom parameters and isomorphous phasing

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

In other contributions to this meeting, a number of novel treatments of isomorphous replacement data are discussed. In contrast, this paper presents a brief overview of what might be called "traditional" methods of phasing and heavy-atom parameter refinement, methods which have been used with great success for many years, but at least some of which should be superceded by improved methods.

To put the refinement & phasing into context, in order to solve a structure by isomorphous replacement we need to :-

(a) identify all sites in each derivative, including minor sites, with no false sites
(b) get best parameters describing the heavy-atom model
(c) get error estimates (primarily from non-isomorphism)
(d) calculate best structure factors (phases)

A typical procedure might be:-

1) initial solution of 1 derivative
2) initial refinement of 1 derivative
3) phase calculation from 1 derivative: 2 sets from alternative hands of solution, *xyz* & *-x-y-z* (using anomalous data), plus SIR without anomalous (3 phase sets altogether)
4) calculate cross-phase difference Fouriers for other derivatives, $\Delta F_2 \exp(i \alpha_1)$. Use most-probable phases (best phases if no anomalous). Choose hand from which of the phase sets gives the best signal-to-noise.
5) check solution against Patterson: plot predicted vectors on Patterson map.
6) refine all derivatives
7) calculate phases
8) calculate difference & residual Fouriers to find minor sites ($\Delta F$ & $\Delta\Delta F$). Go to 5
9) get error estimates from phase calculation: final phasing
10) examine map
11) phase improvement: solvent flattening, averaging
12) solve structure

Of these steps, this paper is concerned with the refinement of heavy-atom parameters and with the calculation of phases: phasing will be discussed first.

## Phasing

Phasing may be considered geometrically by the Harker construction, involving the solution of simultaneous triangles. In the absence of errors, this leads exactly to a pair of possible solutions, and the ambiguity is resolved by a second derivative (or by anomalous scattering) which produces another pair of triangles sharing one edge. With real data, the phase triangle does not in general close even at the correct phase, and different derivatives give conflicting phase information, so we need a statistical treatment of phase *probability* to combine the phase information from different derivatives, including the information from anomalous scattering. To calculate phases, we need to know:-

(a) $|F_P|$ native amplitude

(b) $|F_{PH}|_j$ derivative amplitude scaled to native (and anomalous difference $\Delta|F_{PHj}| = |F_{PHj+}| - |F_{PHj-}|$ )

(c) a model for the difference $\mathbf{F_{Hj}}$ (dependent on the heavy atom parameters)

The phase triangles finally determined by the phasing process are distorted from their true (& unknown) shape by three types of errors

1) errors in amplitudes $|F_P|$, $k|F_{PH}|$
2) errors in the heavy-atom model $\mathbf{F_H}$ (non-isomorphism)
3) errors in protein phase $\alpha_p$

where the phase errors (3) arise from the first two errors.

Ideally, we should determine the optimum $\mathbf{F_P}$ consistent with all observed amplitudes, simultaneously optimizing the heavy-atom parameters and the phases. This is difficult, but new methods may allow this (see contributions to these proceedings by Bricogne and Otwinowski). Traditional methods separate phasing from refinement, and take some simple approximations to the errors: the major error is non-isomorphism for which there is no good general model, so simple methods probably work no worse than more elaborate methods.

We then assume:

1) errors in amplitudes are random & known, $\sigma(F_P)$, $\sigma(F_{PH})$
2) derivatives are *isomorphous*, ie
   (a) $\mathbf{F_P}$ and $\mathbf{F_{PH}}$ are sampled on the same lattice
   (b) the difference between derivative & native may be modelled as a small number of discrete sites, characterized by the heavy-atom parameters $q_i$, $B_i$, & $\mathbf{r}_i$ for the i'th site of the j'th derivative

$$\mathbf{F_{Hj}} = \Sigma_i\, q_i\, f_i\, \exp(-B_i\, s^2)\ \exp(2\pi i\, \mathbf{h}.\mathbf{r}_i)$$

$q_i$ occupancy

$f_i$ atom form-factor (known)

$B_i$ temperature factor
$r_i$ position

With these assumptions, a suitable probability distribution can be set up for the complex native structure factor, $p(\mathbf{F_P} \mid F_P, F_{PHj}, \text{parameters}_{ij})$. Because the non-isomorphous errors are much larger than measurement errors, it is valid to make the simple approximation (Blow & Crick, 1957) to consider the phase probability only as a function of phase angle, rather than considering all values in the complex plane, and to put all the lack of closure error $\varepsilon$ into $F_{PH}$ (see figure 1). The probability is expressed as a normal distribution of width $E_j$

$$p(\alpha) = N \exp( - \varepsilon_j^2(\alpha) / 2E_j)$$

or for all derivatives j

$$p(\alpha) = N \Pi_j \exp( - \varepsilon_j^2(\alpha) / 2E_j)$$

where N is a normalization factor, $\varepsilon(\alpha)$ is the lack of closure of the triangle along $F_{PH}$ at protein phase angle $\alpha$, and $E_j$ is the standard deviation of the distribution.

$$\varepsilon(\alpha) = | F_{PH}|_{obs} - | F_{PH}|_{calc}$$
$$= | F_{PH}|_{obs} - | \mathbf{F_P}(\alpha) + \mathbf{F_H} |$$

This needs an estimate of the width of the normal distribution $E_j$, in addition to the other parameters for the derivative.

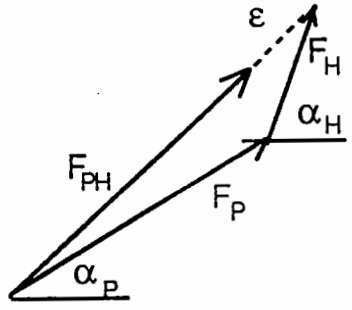



Figure 1
The phase triangle: $\varepsilon$ is the lack of closure for this value of $\alpha_p$

A similar probability expression may be written for the anomalous scattering component

$$p'(\alpha) = N' \exp( - \varepsilon'^2_j(\alpha) / 2E'_j)$$
$$\varepsilon'(\alpha) = \Delta F_{ano\ Obs} - \Delta F_{ano\ Calc}$$

It is better to separate the contributions to the overall probability distribution from isomorphous replacement and anomalous scattering information, because the errors associated with the anomalous component $E'_j$ are different to those for the isomorphous component $E_j$. They are smaller because there is no error due to non-isomorphism for the anomalous difference. However, the anomalous signal is also smaller than the isomorphous signal.

Each piece of phase information (isomorphous or anomalous) gives a bimodal $p(\alpha)$, but the joint distribution should in principle resolve the ambiguity. In practise, the different pieces of phase information often conflict. Also, if $|F_H|$ is small, there is little information.

Although, this formulation of $p(\alpha)$ is crude, it generally works reasonably well. One problem is that the relative weights between different derivatives is not set well. If the phase circles do not cross, the probability is everywhere low, but it can still lead to a sharp joint probability.

**How to use the phase probability**

We may derive 2 sorts of phase angle from $p(\alpha)$

1) Most probable phase, from $\max(p(\alpha))$
   This is the phase angle to use for cross-difference Fouriers and for refinement
2) Centroid ("best") phase
   The native map with the least square error is that calculated with coefficients given by the centroid of the probability distribution (Blow & Crick, 1959)

$$\mathbf{F_{best}} = m\,|F_P|\exp(i\,\alpha_{best})$$

$$= \int |F_P| \exp(i\alpha)\, p(\alpha)\, d\alpha \,/ \int p(\alpha)\, d\alpha$$

This is the phase (& amplitude) for the final map. Note that the "figure-of-merit" m is a measure of the sharpness of the phase probability (precision not error). If the errors E are underestimated, the figure of merit will be higher.

**Practical phasing**

1) How to get error estimates
   (a) set isomorphous E = rms(ε) for centric reflections: for acentric reflections rms(ε) is an underestimate of the error. E probably should be different for centric & acentric reflections, but it is difficult to estimate from acentrics (see paper by Read for better ways of estimating error)
   (b) set anomalous E' = rms(ε') for most-probable phase: this is not a perfect estimate, but is the best available

Many programs ( eg the CCP4 programs Phase & Phare) attempt to separate the error into isomorphism and data errors

Total error $\langle\varepsilon^2\rangle = \langle\sigma^2\rangle + \langle\delta^2\rangle$

where $\langle\sigma^2\rangle$ is the known data error, and $\langle\delta^2\rangle$ is the error contribution from lack of isomorphism, then $\langle\delta^2\rangle$ may be estimated from

$$\langle\delta^2\rangle = \langle\varepsilon^2\rangle - \langle\sigma^2\rangle$$

In general, the contribution from non-isomorphism is much larger than the contribution from data errors.

For initial phasing (before rms($\varepsilon$) is available), a useful estimate of E can be obtained from the rms residual in refinement, and and the anomalous $\varepsilon'$ guessed at 1/3 $\varepsilon$. Use these for 1st phase run, get new values, then rerun.

2) Checks on quality of phasing

(a) the value of mean m (figure-of-merit) is a poor indicator. The figure-of-merit measures the sharpness of the probability distribution, not its accuracy. In published structures, the figure-of-merit has often been artificially reduced by underestimation of E.

(b) $\alpha_P$ should be uncorrelated with $\alpha_H$: this can be demonstrated by the distribution of $|\alpha_P - \alpha_H|$ being flat (though in practice it never is), and by $\langle|\alpha_P - \alpha_H|\rangle$ being = 90°. Correlation arises particularly from errors in the derivative scale factor & shows bias in the phase. This gives peaks or holes at heavy-atom positions in the final map

(c) The 'Phasing power' = rms($|F_H|_{calc}$) / rms($\varepsilon$) is a rough estimate of signal-to-noise ratio for each derivative.

(d) The quality of the map - this is the most important criterion

## Refinement

A number of methods have been used to refine the parameters defining the model for the difference between native & derivative, ie that define $\mathbf{F_H}$ for the phasing process. As pointed out by Bricogne in his contribution to this meeting, all standard methods violate the basic principles of least-squares, a process which allows optimization of parameters defining a model by minimizing the sum of weighted squared differences between observed and calculated values of a quantity. The problem arises because it is not possible to write equations for the calculated value of any observed quantity ($|F_P|$ or $|F_{PH}|$ or better, the intensities) independent of other observed quantities, or to make the weights strictly independent of the parameters. All standard methods are thus statistically suspect, though they may well give useful results, and work mostly by limiting the data used to a subset for which the statistical bias is small, eg centric reflections,

or reflections which are well-phased by other derivatives. The methods described by Bricogne & by Otwinowski in this volume should be better behaved.

Three categories of method may be considered:-

1) *Centric refinement:* restricting consideration to centric reflections only is a special case of the other methods. Where it is appropriate (see below), this is the simplest & best method.
2) *Using phases:* provided that more than one derivative is available, parameters may be refined using precalculated phases, either in alternate cycles of phasing and refinement, or simultaneously.
3) *Phaseless amplitude methods:* these methods do not use phases, and treat each derivative independently, so may be use even when only one derivative is available.

Phasing & refinement are correlated, because in the phase triangle we can minimize $\varepsilon$ by either changing phase $\alpha_p$ or $\mathbf{F_H}(p_i)$ (ie the heavy-atom parameters): the phase triangle is squashy. This correlation is :-
(a) *bad*: phases are biased to account for errors
(b) *good*: parameters which are ill-determined in refinement are less important in phasing

For centric reflections, the phase triangle is forced to be a straight line, so parameter errors cannot easily be traded for phase errors, hence the value of centric refinement.

**Heavy-atom parameters**

These are the parameters which define the model for the derivative.

(a) relative scale & temperature factor, to put the derivative data on the same scale as the native

$$F_{PH}' = k_{rel} \exp(-B_{rel} \sin^2\theta / \lambda^2) \, |F_{PH}|$$

(b) site parameters

$q_i$ occupancy
$B_i$ temperature factor (or anisotropic $\beta$)
$\mathbf{r}_i$ position

$$\mathbf{F_{Hj}} = \Sigma_i \, q_i \, f_i \, \exp(-B_i \, s^2) \, \exp(2\pi i \, \mathbf{h}.\mathbf{r}_i)$$

where $f_i$ is the known atom form-factor

## Centric refinement

It is legitimate to use a subset of reflections for refinement, provided:-
a) there are a large number of observations/parameter. This condition is generally met, since we have thousands of reflection measurements to determine a few dozen (or fewer) parameters.
b) the selection is uncorrelated with $F_H$, otherwise parameters may be biased (particularly occupancy)

In the centric case, phased & phaseless refinements become essentially equivalent. The functions minimized are (see below)

**Phased**

$R_1 = \Sigma_h\ w_h\ (\ |\ F_{PH}|_{obs}\ -\ k_{rel}\ |\ \mathbf{F_P + F_H}|\ )^2$

where the sign (phase) of $\mathbf{F_p}$ is assumed known

**Unphased**

$R_2 = \Sigma_h\ w\ (\ |F_H|_{obs}\ -\ |F_H|_{calc})^2$

where $|F_H|_{obs} = |\ |F_{PH}| \pm |F_P|\ |$

In the unphased refinement, the sign ambiguity is equivalent to the relative phase of $F_P$ & $F_{PH}$. For most reflections, $F_P$ & $F_{PH}$ have the same sign, since $|\ F_H\ | << |\ F_P\ |$ , but "cross-over" terms must be dealt with. The two possibilies (LE = lower estimate, UE = upper estimate) are:-

$F_{HLE} = |\ |\ F_{PH}\ | - |\ F_P|\ |$

$F_{HUE} = |\ |\ F_{PH}\ | + |\ F_P|\ |$

Then for refinement, we can either:-
(a) exclude reflections for which $F_{HUE} < F_{H\ max}$
or
(b) use $F_{HLE}$ or $F_{HUE}$ nearest to $|\ F_H\ |_{calc}$ This is equivalent to phasing with a single derivative.

When it is valid, centric refinement is probably the best method. It is particular valid:-
1) for dihedral point groups, ie 222, 422, 32, 312, 622, 23, 432, but not in polar spacegroups. In dihedral point groups, there are at least 2 centric zones.
2) when there are not too many sites. The method is perhaps not suitable for complex derivatives (heavy-atom clusters), or for poorly resolved sites.

## Phased methods

(a) *alternate phase & refinement:* this is the most commonly used method. The residual minimized is

$R = \Sigma_h\ w_h\ (\ |F_{PH}|_{obs}\ -\ |F_{PH}|_{calc})^2$

$$= \Sigma_h \; w_h \; ( \; | \; F_{PH}|_{obs} \; - k_{rel} \; | \; \mathbf{F_P} + \mathbf{F_H} | \; )^2$$

where $\mathbf{F_P} = \; | \; F_p \; |_{obs} \; \exp(i\alpha_{most\ probable})$

Note that the most-probable phase should be used, not the centroid phase. The traditional method treated only $\mathbf{F_H}$ as a function of the heavy-atom parameters. This has been extended by Bricogne to allow for the dependence of $\mathbf{F_P}$ on the parameters, but he evaluated the dependency only at the current phase estimate. Neglect of this dependency leads to problems, particularly where derivatives have common sites.

Phases derived from external sources have also been used with success: these include phases calculated from preliminary models of the structure, and isomorphous phases modified by solvent flattening or averaging. In these cases, the feedback bias of the method is reduced or eliminated.

(b) Sygusch (1977, see also He & Carter, 1989) described a simultaneous optimization of parameters & phases, using a diagonal matrix approximation to the least-squares matrix corresponding to the phases.

(c) Better methods related to this formulation are described in the papers by Otwinowski and Bricogne in this volume. These methods should supercede existing methods.

**Amplitude methods**

(a) *"$F_{HLE}$" method*: this uses the anomalous signal to estimate $F_H$ for acentric reflections (Dodson,1976)

$$F_H^2 \quad = F_P^2 + F_{PH}^2 \; -+ \; 2F_P \; F_{PH} \; \sqrt{[1-(k_{ano}\Delta_{ano}/2F_P)^2]}$$
$$\approx \Delta_{iso}^2 \; + \Delta_{ano}^2$$

The sign ambiguity is similar to the centric case, giving $F_{HLE}$ & $F_{HUE}$ (lower & upper estimates). The residual minimized is:-

$$R_2 \quad = \Sigma_h \; w_h \; ( \; |F_H|_{obs} \; - \; |F_H|_{calc})^2$$
$$w_h \quad = \; 1 \; / \; Var(|F_H|_{obs})$$

But $F_H^2$ and hence $|F_H|$ is not an observed quantity, and is biased, particularly because $\Delta_{ano}^2$ is biased. The bias may be estimated and subtracted (Dodson, Evans & French, 1975), most simply from

$$\langle \Delta^2 \rangle \; = \; \langle \Delta \rangle^2 + Var(\Delta)$$

The weighted refinement is dominated by centric reflections if present, since the estimate of $F_H$ is much more accurate for centrics. This method is only suitable if good anomalous scattering data are available.

(b) *Heavy:* the program Heavy implements a reciprocal-space version of refinement to an origin-removed Patterson (Terwilliger & Eisenberg, 1983). This is an unbiased single-derivative refinement which requires no anomalous scattering data. The residual minimized is:-

$$R = \Sigma_h w_h \{ \underbrace{[(F_{PH}-F_P)^2-\langle(F_{PH}-F_P)^2\rangle]}_{\textit{observed Patterson}} - c \underbrace{[ F_H^2{}_{calc} - \langle F_H^2{}_{calc}\rangle ]}_{\textit{calculated Patterson}} \}^2$$

where c = 1/2 for an acentric reflection
= 1 for a centric reflection

(c) *Patterson space refinement:* this is discussed by Tickle (this volume).

**Comparison of different refinements**

*Separate (SIR)*

Advantages:

Can refine 1st derivative or SIR
No feedback between common sites
Eliminates wrong sites even if present in other derivatives

Disadvantages:

Can't refine relative origin in polar spacegroups

*Centric*

Advantages:

Separates (mostly) interdependence of phase and heavy-atom parameters

Disadvantages:

Uses fewer observations

*Alternate phase & refine*

Advantages:

Uses all derivatives (all information)
Refines relative origin
Gives good values for relative scale

Disadvantages:

Parameters for common sites biased
Needs at least 2 derivatives

*Fhle*

Advantages:
Uses all data for single derivatives

Disadvantages:
Needs good anomalous difference data

*Terwilliger/Eisenberg (Heavy)*

Advantages:
Claimed to be unbiased single derivative refinement
Doesn't need (or use) anomalous data

Disadvantages:
Single derivative

*New methods*

Not yet proven for many cases, but the method of Otwinowski seems to give good results.

## Practical refinement

1) Choose method: for 1st derivative use centric refinement or $F_{hle}$ or Heavy

2) Heavy & Patterson space (program Vecref) methods are claimed to be good at removing spurious sites

3) How do you know if a derivative is valid? (Does it refine?)

Agreement residuals (R-factors for "$F_{obs}$" versus "$F_{calc}$") are nearly always poor. Correlation coefficient between $|F_H|_{obs}$ & $|F_H|_{calc}$ (from $F_{HLE}$ refinement, acentric or centric) is a better indicator (any value >0.2 is promising!)

4) Although it is good to try different methods, it is a mistake to spend too much time massaging heavy-atom parameters. The correctness of the phasing is limited ultimately by the degree of non-isomorphism, not by the refinement program.

## References

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

Dodson,E.J., in Crystallographic Computing Techniques, Ahmed, F.R. (ed), Munksgaard, Copenhagen (1976)

Dodson, E.J, Evans,P.R. & French,S. in Anomalous Scattering, Ramaseshan,S. & Abrahams,S.C. (eds.), pp 423-436, Munksgaard, Copenhagen (1975)

He, X.M. & Carter, D.C. Acta Cryst., A45, 308-308 (1989)

Sygusch, J. Acta Cryst., A33, 512-518 (1977)

Terwilliger, T.C. & Eisenberg,D. Acta Cryst., A39, 813-817

Watenpaugh, K.D., Methods in Enzymology, 115, 3-15 (1985)

# A Maximum-Likelihood Theory of Heavy-atom Parameter Refinement in the Isomorphous Replacement Method.

G. Bricogne

L.U.R.E., Université Paris-Sud, 91405 Orsay, France

and

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England.

## *0. Introduction.*

The problem of refining heavy-atom parameters from acentric reflexions in the MIR method remains a troublesome issue in macromolecular crystallography. The conventional approach to this problem was originally conceived (Dickerson, Kendrew & Strandberg, 1961) as a straightforward adaptation of the least-squares method previously used on centric data by Hart (1961): the "most probable" or the "best" estimates of the phases, as defined by Blow & Crick (1959), were simply made to play a rôle analogous to that of the signs of centric reflexions. Blow & Matthews (1973) 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 which were being refined. An alternative refinement scheme was devised, under the name of "$F_{HLE}$ method" (Dodson, 1976), in which the use of acentric phase estimates was avoided altogether. A common feature of both procedures was that the phase information used in the refinement had to be *purposely impoverished*, in order to make it truly independent of the parameters to be refined and thus avoid bias. Sygusch (1977) recognized that these restrictions could be lifted 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 of 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.

In previous papers (Bricogne, 1982, 1984b), a further analysis of the problem was carried out, and a solution was proposed which partially overcame the aforementioned difficulties. The main idea was that structure factor estimates for acentric reflexions are *implicit functions* of the parameters which are being refined. This dependence was expressed analytically by means of the implicit function theorem, and 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. In this way, an important source of bias is removed without sacrificing any of the available phase information, and no further parameters needed to be refined. A conspicuous limitation of this approach, however, was its inability to deal with the single isomorphous replacement (SIR) case.

Extensive tests of this method (Bricogne, 1984-85, unpublished) in the course of the structure determination of an Fab-lysozyme complex (Amit, Mariuzza, Phillips & Poljak, 1986) showed that many previously observed pathologies, such as the rapid divergence of the site occupancies of good derivatives, had indeed been cured by this analysis. However, slower instabilities were observed which resulted in divergent behaviour of the estimates for the *lack of isomorphism* of the various derivatives. In retrospect, the cause was clear: the refinement process was still violating one of the main tenets of the least-squares method, namely the requirement that *the weights used should be kept fixed* as if they were part of the observed data. Since the method of least-squares is a particular case of the maximum-likelihood method when errors are normally distributed with fixed (co)variances, it was clear that the problem of properly estimating the lack-of isomorphism parameters (which do alter the weights used in calculating the residual) demanded a full-fledged maximum-likelihood rather than a least-squares treatment. This conclusion, together with the natural rôle of

likelihood as a heuristic criterion envisaged in Bricogne(1984a), was the main motivation which led to formulating the synthetic approach presented in Bricogne (1988a,b).

The maximum-likelihood theory of MIR and SIR presented below was obtained in mid-1985. In the mean time other authors, namely Z. Otwinowski and R. Read, have independently advocated the use of maximum-likelihood methods in dealing with the refinement of heavy-atom parameters, and have taken steps to implement it. Their approaches, however, are mostly computational ones, and make no attempt at providing a systematic analytical formulation of the statistical problem involved. In what follows, on the contrary, the emphasis will be on first deriving the relevant likelihood functions in closed form, at the level of the diagonal approximation, then on examining how MIR and SIR information interacts with direct phasing methods.

*1. Acentric MIR and SIR likelihood functions.*

Given a fixed acentric reflexion $\mathbf{h}$, let $re^{i\varphi}$ denote the (complex-valued) native structure factor $F^P(\mathbf{h})$, let $f_j e^{i\alpha_j}$ denote the heavy-atom contribution $F_j^H(\mathbf{h})$, and let $r_j$ denote the structure factor amplitude $\left|F_j^{PH}(\mathbf{h})\right|$ for the $j^{th}$ isomorphous derivative. In the case of perfect isomorphism, the Harker construction gives the relation :

$$r_j = \left( r^2 + 2rf_j\cos\theta_j + f_j^2 \right)^{\frac{1}{2}} \qquad \text{with} \qquad \theta_j = \alpha_j - \varphi \qquad (1).$$

The existence of lattice-preserving non-isomorphism may now be modelled as a uniform distribution of random "clutter atoms" (Bricogne, 1988b, §3.1, §3.2). This makes $F_j^{PH}(\mathbf{h})$ into a random complex number whose distribution is a two-dimensional Gaussian centred at $\langle F_j^{PH} \rangle = re^{i\varphi} + f_j e^{i\alpha_j}$ with a variance $\Sigma_j^a$ related to the clutter atom model by the Wilson relation (Wilson, 1949, 1950) :

$$\Sigma_j^a = \frac{1}{2}\varepsilon(\mathbf{h}_j)\,\sigma_2(\mathbf{h}_j) \qquad (2)$$

where $\varepsilon(\mathbf{h}_j)$ is the statistical weight of $\mathbf{h}_j$ and where $\sigma_2(\mathbf{h}_j)$ is the sum of the squared scattering factors summed over all the clutter atoms in the unit cell. It should be noted that the population of clutter atoms may comprise not only ordinary atom-like features (with roughly Gaussian form factors, decreasing with resolution), but also the type of residual features associated with random disturbances of the positions of some atoms in the native structure (with form factors best represented by spherically averaged Hermite functions of degree 1, whose magnitude would increase rather than decrease with resolution within the resolution range considered). Integrating with respect to the phase of $F_j^{PH}$ gives the following conditional probability distribution for its amplitude $R_j$ :

$$\mathcal{P}(R_j \mid r, \varphi, f_j, \alpha_j) = \mathcal{R}(r_j(r, \varphi, f_j, \alpha_j), R_j, \Sigma_j^a) \qquad (3)$$

where $\mathcal{R}$ denotes the Rice distribution (Rice, 1944, 1945):

$$\mathcal{R}(r, R, \Sigma) = \frac{R}{\Sigma} \exp\left(-\frac{r^2+R^2}{2\Sigma}\right) I_0\left(\frac{rR}{\Sigma}\right) \qquad (4).$$

If there are M distinct isomorphous derivatives, the conditional joint probability distribution of the M structure factor amplitudes for these compounds is

$$\mathcal{P}(R_1, ..., R_M \mid r, \varphi, f_1, \alpha_1, ..., f_M, \alpha_M) = \prod_{j=1}^{M} \mathcal{R}(r_j(r, \varphi, f_j, \alpha_j), R_j, \Sigma_j^a) \qquad (5).$$

The likelihood of the current heavy-atom model is then obtained by integrating over all possible values of the unknown phase $\varphi$ of the native structure factor and substituting the observed amplitudes for the $R_j$ (which will still be denoted $R_j$ to keep the notation simple) into the resulting expression :

$$\Lambda_{MIR}^a(f_1, \alpha_1, ..., f_M, \alpha_M) = \int_0^{2\pi} \prod_{j=1}^{M} \mathcal{R}(r_j(r, \varphi, f_j, \alpha_j), R_j, \Sigma_j^a) P(\varphi)\, d\varphi \qquad (6).$$

Here $P(\varphi)$ contains any phase information available for $F^P$ from sources other than MIR (e.g. from direct phasing), and is uniform if no such information exists. The integration can be carried out analytically, starting from (4), by invoking the generating relation

$$e^{z\cos\theta} = \sum_{m=1}^{\infty} I_m(z)\, e^{im\theta} \tag{7}$$

and the addition theorem for $I_0$:

$$I_0\left(\sqrt{x^2+2xy\cos\theta+y^2}\right) = \sum_{n=1}^{\infty} I_n(x)\, I_n(y)\, e^{in\theta} \tag{8}$$

to expand the integrand in (6) into a trigonometric series with respect to $\varphi$. With the notation $\mathbf{m}=(m_1, m_2, ..., m_M)$, $\mathbf{n}=(n_1, n_2, ..., n_M)$, $\alpha=(\alpha_1, \alpha_2, ..., \alpha_M)$, the likelihood may be written

$$\Lambda^a_{MIR} = 2\pi \prod_{j=1}^{M} \left(\frac{R_j}{\Sigma_j^a}\right) \exp\left[ - \sum_{j=1}^{M} \frac{r^2+f_j^2+R_j^2}{2\Sigma_j^a} \right] \times \sum_{p=-\infty}^{\infty} c_p \sum_{(\mathbf{m},\mathbf{n})\in J_p} (-1)^{\mathbf{m}}\, I_{\mathbf{m},\mathbf{n}}\, e^{i(\mathbf{m}+\mathbf{n})\cdot\alpha} \tag{9}$$

where $J_p$ denotes the set of those $(\mathbf{m}, \mathbf{n})$ such that $\sum_{j=1}^{M} m_j + n_j = p$, $c_p$ denotes the $p^{th}$ Fourier coefficient of $P(\varphi)$, and

$$(-1)^{\mathbf{m}}\, I_{\mathbf{m},\mathbf{n}} = \prod_{j=1}^{M} (-1)^{m_j} I_{m_j}\left(\frac{r f_j}{\Sigma_j^a}\right) I_{n_j}\left(\frac{r R_j}{\Sigma_j^a}\right) I_{n_j}\left(\frac{f_j R_j}{\Sigma_j^a}\right) \tag{10}.$$

If $P(\varphi)$ is uniform, only the terms for $p = 0$ are present, with $c_0 = 1$; if not, the extra terms combine the phase information contained in $P(\varphi)$ with that available from MIR. If furthermore there is a single isomorphous derivative (M = 1) the likelihood simplifies to

$$\Lambda^a_{SIR} = 2\pi \frac{R_1}{\Sigma_1^a} \exp\left[ - \frac{r^2+f_1^2+R_1^2}{2\Sigma_1^a} \right] I_0\left(\frac{r f_1}{\Sigma_1^a}\right) I_0\left(\frac{r R_1}{\Sigma_1^a}\right) I_0\left(\frac{f_1 R_1}{\Sigma_1^a}\right) \times \left[ 1 + 2\sum_{m=1}^{\infty} (-1)^m \tau_m\left(\frac{r f_1}{\Sigma_1^a}\right) \tau_m\left(\frac{r R_1}{\Sigma_1^a}\right) \tau_m\left(\frac{f_1 R_1}{\Sigma_1^a}\right) \right] \tag{11}$$

where $\tau_m(z) = \dfrac{I_m(z)}{I_0(z)}$ .

*2. Centric MIR and SIR likelihood functions.*

It is a straightforward matter to derive the corresponding likelihood expressions for a *centric* reflexion. Relation (1) still holds, but since $\theta_j = 0$ or $\pi$ the quantity $\cos\theta_j$ is simply a sign $s_j$. Assuming a uniform distribution of clutter atoms, $F_j^{PH}(\mathbf{h})$ becomes a random real number distributed around its expectation value $\langle F_j^{PH} \rangle = r e^{i\varphi} + f_j e^{i\alpha_j}$ as a one-dimensional Gaussian with variance $\Sigma_j^c$ given by :

$$\Sigma_j^c = \varepsilon(\mathbf{h}_j)\,\sigma_2(\mathbf{h}_j) \tag{12}$$

Summing over two possible values of the sign of $F_j^{PH}$ gives the conditional probability distribution for its amplitude $R_j$ in the form :

$$\mathcal{P}(R_j \mid r, \varphi, f_j, \alpha_j) = C(r_j(r, \varphi, f_j, \alpha_j), R_j, \Sigma_j^c) \tag{13}$$

where $C$ denotes the centric version of the Rice distribution :

$$C(r, R, \Sigma) = \sqrt{\frac{2}{\pi\Sigma}}\,\exp\left(-\frac{r^2+R^2}{2\Sigma}\right)\cosh\left(\frac{rR}{\Sigma}\right) \tag{14}.$$

Assuming as before that there are M distinct isomorphous derivatives, the likelihood of the current heavy-atom model is then :

$$\Lambda_{MIR}^c(f_1, \alpha_1, ..., f_M, \alpha_M) = \sum_{\varphi} \prod_{j=1}^{M} C(r_j(r, \varphi, f_j, \alpha_j), R_j, \Sigma_j^c)\,P(\varphi) \tag{15}$$

where the sum is over the two allowed values of the native phase $\varphi$. The centric equivalent of (7) may be written :

$$e^{zs} = \cosh z + s\sinh z = \sum_{m=0,1} C_m(z)\,s^m \quad \text{with } C_0 = \cosh \text{ and } C_1 = \sinh \tag{16}$$

and that of (8) then reads :

$$C_0(x + sy) = \sum_{m=0,1} C_m(x)\,C_m(y)\,s^m \tag{17}.$$

The integrand in (15) may then be rewritten as a dyadic expansion which, when summed over the two possible values of the native phase $\varphi$, yields the centric equivalent of (9) :

$$\Lambda^{c}_{MIR} = \prod_{j=1}^{M} \sqrt{\frac{2}{\pi\Sigma^{c}_{j}}} \exp\left[ - \sum_{j=1}^{M} \frac{r_j^2 + f_j^2 + R_j^2}{2\Sigma^{c}_{j}} \right] \times \sum_{p=0,1} c_p \sum_{(\mathbf{m},\mathbf{n}) \in K_p} (-1)^{\mathbf{m}} C_{\mathbf{m},\mathbf{n}} \, e^{i(\mathbf{m}+\mathbf{n})\cdot\mathbf{s}} \qquad (18)$$

where $K_p$ ($p = 0, 1$) denotes the set of those $(\mathbf{m}, \mathbf{n})$ such that $\sum_{j=1}^{M} m_j + n_j \equiv p$ modulo 2, $c_p$ denotes the $p^{th}$ (dyadic) Fourier coefficient of $P(\varphi)$, and

$$(-1)^{\mathbf{m}} C_{\mathbf{m},\mathbf{n}} = \prod_{j=1}^{M} (-1)^{m_j} C_{m_j}\left(\frac{r f_j}{\Sigma^{c}_{j}}\right) C_{n_j}\left(\frac{r R_j}{\Sigma^{c}_{j}}\right) C_{n_j}\left(\frac{f_j R_j}{\Sigma^{c}_{j}}\right) \qquad (19),$$

while

$$\mathbf{s} = (s_1, s_2, \ldots, s_M) \qquad \text{with} \qquad s_j = \cos\theta_j \qquad (20).$$

For a uniform $P(\varphi)$, only the term for $p = 0$ is present, with $c_0 = 1$ ; if not, the extra term combines the phase information contained in $P(\varphi)$ with that available from MIR . In particular, the SIR expression for uniform P which corresponds to (11) is given by

$$\Lambda^{c}_{SIR} = \sqrt{\frac{2}{\pi\Sigma^{c}_{1}}} \exp\left[ - \frac{r^2 + f_1^2 + R_1^2}{2\Sigma^{c}_{1}} \right] \cosh\left(\frac{r f_1}{\Sigma^{c}_{1}}\right) \cosh\left(\frac{r R_1}{\Sigma^{c}_{1}}\right) \cosh\left(\frac{f_1 R_1}{\Sigma^{c}_{1}}\right) \times \left[ 1 - \tanh\left(\frac{r f_1}{\Sigma^{c}_{1}}\right) \tanh\left(\frac{r R_1}{\Sigma^{c}_{1}}\right) \tanh\left(\frac{f_1 R_1}{\Sigma^{c}_{1}}\right) \right] \qquad (21).$$

*3. Maximum-likelihood refinement of heavy-atom parameters.*

The problem of MIR or SIR parameter refinement from data containing any mixture of acentric and centric reflexions may now be viewed as that of maximising (by a Newton or quasi-Newton method) the appropriate likelihood expression with respect to

(1) the heavy-atom parameters (site occupancies, site temperature factors, and atomic coordinates), which enter *via* $f_1, \alpha_1, \ldots, f_M, \alpha_M$ ;

(2) = the overall scale and temperature factors of the derivative data with respect to the native data (or to absolute scale), which enter *via* the "observed" values $R_j^{obs}$ substituted for the $R_j$ ;

(3) the lack-of-isomorphism parameters $\Sigma_j^a$ and $\Sigma_j^c$ for the various derivatives, which are themselves described parametrically by a few global parameters : the number of clutter atoms and their scattering factors.

In this way all sources of bias are removed, since no value of the native phase $\varphi$ is treated in a privileged manner ; externally available phase information can be introduced *via* $P(\varphi)$ to help the refinement ; the lack-of-isomorphism parameters are refined along with the other parameters ; and the method applies to the SIR as well as to the MIR case, using all the data available, both centric and acentric. Furthermore, the rather crude statistical modelling of non-isomorphism used here (and in the standard approach!) can be improved by using separate and non-uniform distributions for different types of "clutter" atoms (§1), which would provide useful *ME residual maps* for the location of structural disturbances and/or of minor sites.

*4. Discussion.*

It now seems safe to conclude that the demon which has plagued SIR/MIR heavy-atom parameter refinement for the past 25 years has finally been exorcised. Only the task of implementation remains!

This maximum-likelihood re-formulation of the MIR and SIR methods has been extended to the case of *anomalous scatterers*, to the problem of *detecting* (rather than just refining) isomorphous substituents, and to the use of intensity data from *powders* and *fibres* which present extra difficulties because of the *overlap problem*. These results are being published elsewhere.

## *References*


AMIT, A.G., MARIUZZA, R.A., PHILLIPS, S.E.V. & POLJAK, R.J. (1986). *Science*, **233**, 747-753.

BLOW, D.M. & CRICK, F.H.C. (1959). *Acta Cryst.* **12**, 794-802.

BLOW, D.M. & MATTHEWS, B.W. (1973). *Acta Cryst.* A**29**, 56-62.

BRICOGNE, G. (1982). In *Computational Crystallography*, edited by D. SAYRE, pp. 223-230. New York: Oxford University Press.

BRICOGNE, G. (1984a). *Acta Cryst.* A**40**, 410-445.

BRICOGNE, G. (1984b). In *Methods and Applications in Crystallographic Computing*, edited by S.R. HALL & T. ASHIDA, pp. 141-151. Oxford : Clarendon Press.

BRICOGNE, G. (1988a). In *Crystallographic Computing 4* , edited by N.W. ISAACS & M.R. TAYLOR, pp. 60-79. New York : Oxford Univ. Press..

BRICOGNE, G. (1988b). *Acta Cryst.* A**44**, 517-545.

DICKERSON, R.E., KENDREW, J.C. & STRANDBERG, B.E. (1961). 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.

DODSON, E.J.(1976). In *Crystallographic Computing Techniques*, pp.259-268. Edited by F.R. AHMED, K. HUML and B. SEDLACEK. Copenhagen: Munksgaard.

HART, R.G. (1961). *Acta Cryst.* **14**, 1194-1195.

RICE, S.O. (1944, 1945). *Bell System Tech. J.* **23**, 283-332 (parts I and II) ; **24**, 46-156 (parts III and IV). Reprinted in *Selected Papers on Noise and Stochastic Processes* (1954), edited by N. WAX, pp. 133-294. New York : Dover Publications.

WILSON, A.J.C. (1949). *Acta Cryst.* **2**, 318-321.

WILSON, A.J.C. (1950). *Acta Cryst.* **3**, 258-261.

# Dealing with imperfect isomorphism in multiple isomorphous replacement*

Randy J. Read
Department of Medical Microbiology & Infectious Diseases
University of Alberta, Edmonton, Alberta T6G 2H7, Canada

Multiple isomorphous replacement (MIR) is the major technique for obtaining the phase information necessary to solve new protein crystal structures. The optimal use of information from MIR experiments requires a good understanding of the sources of error, and of how they propagate into the phases. The chief sources of error are errors in the amplitude measurements, imperfect isomorphism, and errors in the heavy atom model. Their effects are examined, and an improved treatment for non-isomorphism is proposed. In addition, a method to combine properly the information from several derivatives is reviewed. Tests using calculated data with simulated errors give results consistent with the probability distributions.

## 1. Introduction

The familiar phase problem of crystal structure determination arises because the diffraction experiment measures only the intensity (or amplitude) of the structure factor, but not its phase. In order to reconstruct the electron density in the crystal, by taking the Fourier transform of the structure factors, we need estimates of the phases. Multiple isomorphous replacement (MIR) is still the major technique used to obtain this phase information for new protein structures.

In the MIR method, a small number of heavy atoms is added to the structure in a native protein crystal to give a heavy atom derivative. In the best case, the structure of the protein and its packing in the crystal are unchanged, and the derivative crystal is said to be perfectly isomorphous with the native crystal. The derivative structure factor is then the vector sum of the protein and heavy atom structure factors. If we have a model of the heavy atoms in the crystal, we can estimate the heavy atom contribution to the structure factor and draw inferences about which values for the protein phase would be consistent with the observed amplitudes.

If there were no errors in a single isomorphous replacement (SIR) experiment, two discrete values of the native phase would be consistent with the measured diffraction data and the heavy atom substitution model. Two independent derivatives would give an unambiguous phase determination. This ideal case is illustrated by the Harker (1956) construction in Figure 1. However, the various sources of uncertainty in the experiment introduce uncertainty into the possible phases; instead of discrete possibilities for the protein phase, we obtain a probability distribution. The optimal use of phase information requires accurate estimation of the phase probability distribution, which in turn requires a good understanding of the propagation of errors in the experiment. Phase probabilities are needed to obtain optimal electron density maps (Blow & Crick, 1959) and to combine independent sources of phase information (Rossmann & Blow, 1961; Hendrickson & Lattman, 1970).

There are three major sources of uncertainty in an MIR phase determination. The first is error in measuring and scaling the diffraction data for native and derivative crystals. In other terms, $F_o$ has errors as an estimate of F, and $F_{jo}$ as an estimate of $F_j$. (Terms and notation are summarized in Table 1, along with the relationships among the variables.) The second is errors in the heavy atom model, *i.e.* in $\mathbf{H}_{jc}$ as an estimate of $\mathbf{H}_j$. The third is imperfect isomorphism, or non-isomorphism, which affects how well the equation $\mathbf{F}_{pj} = \mathbf{F}$ is satisfied.

Blow & Crick (1959) showed for the SIR case that, to a reasonable approximation, errors in the intensity measurements and in the heavy atom model can be lumped together and treated as Gaussian errors in $F_{jo}$, the measured derivative structure factor amplitude. Terwilliger & Eisenberg (1987) performed a more detailed analysis of error propagation, including the effects of non-isomorphism. They pointed out that the sources of error must be divided into two

---

* This contribution is a revised version of a chapter written for Crystallographic Computing 5, to be published in connection with the 1990 Crystallographic Computing School held in Bischenberg, France.

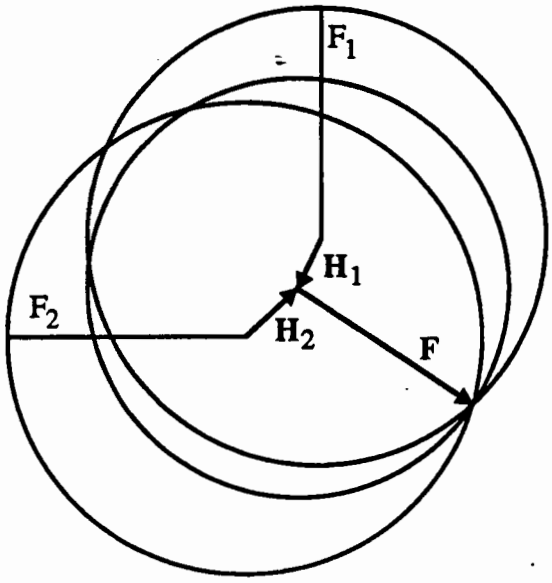



**Figure 1.** A Harker (1956) construction for two derivatives. The intersections of the circles show possible solutions to the equations $\mathbf{F}_1=\mathbf{F}+\mathbf{H}_1$ and $\mathbf{F}_2=\mathbf{F}+\mathbf{H}_2$, consistent with the amplitudes and the structure factor contributions from the heavy atoms.

**Table 1: Terms and Notation**

| | |
|---|---|
| $\langle x \rangle$ | = expected value, or probability-weighted average, of x |
| $p(x,y\|z)$ | = joint probability distribution of x and y, conditional on z |
| $\mathbf{F}$ | = true structure factor for native protein<br>$= F\exp(i\alpha)$ |
| $F_o$ | = observed structure factor amplitude for native protein |
| $\sigma$ | = standard deviation of $F_o$ as an estimate of F |
| $\mathbf{F}_{pj}$ | = true structure factor for protein part of derivative j<br>= **F** in case of perfect isomorphism |
| $D_j$ | = multiplication factor giving component of **F** correlated to $\mathbf{F}_{pj}$<br>= 1 in case of perfect isomorphism |
| $\mathbf{H}_j$ | = true structure factor for heavy atom part of derivative j |
| $\mathbf{H}_{jc}$ | = calculated structure factor for model of heavy atoms in derivative j<br>$= H_{jc}\exp(i\alpha_{jc})$ |
| $\mathbf{F}_j$ | = true structure factor for derivative j<br>$= \mathbf{F}_{pj}+\mathbf{H}_j$ |
| $F_{jo}$ | = observed structure factor amplitude for derivative j |
| $\sigma_j$ | = standard deviation of $F_{jo}$ as an estimate of $F_j$ |
| $\mathbf{F}_{jc}$ | = structure factor for derivative j that would be calculated knowing the true native **F**<br>$= D_j\mathbf{F}+\mathbf{H}_{jc}$ |
| $F_{jc}$ | $= \|\mathbf{F}_{jc}\|$<br>$= [D_j^2F^2+H_{jc}^2+2D_jF\,H_{jc}\cos(\alpha-\alpha_{jc})]^{1/2}$ |
| $F_{jco}$ | = value of $F_{jc}$ calculated using $F_o$ for F |
| $\Delta_j$ | $= \mathbf{F}_j-\mathbf{F}_{jc}$ |
| $\varepsilon$ | = expected intensity factor for reciprocal lattice zone |
| $\sigma_{\Delta j}^2$ | $= \langle\|\Delta_j\|^2/\varepsilon\rangle$ |
| m | = expected value of the cosine of the phase error<br>$= \langle\cos(\alpha-\alpha_c)\rangle$, where $\alpha_c$ is the centroid phase, defined by $\langle\exp(i\alpha)\rangle = m\exp(i\alpha_c)$ |

categories that affect centric and non-centric reflections differently. Measurement errors affect only the amplitudes for both centric and non-centric reflections. However, errors from the heavy atom model and from non-isomorphism are distributed in the complex plane for non-centric reflections. It is still possible to treat the error as residing in $F_{jo}$, but for non-centric reflections the contribution to the variance from errors in the complex plane (errors in $\mathbf{H}_{jc}$ and $\mathbf{F}_{pj}$) must be divided by two. Green (1979) also derived a probability distribution taking explicit account of the possibility of non-isomorphism, but his treatment covered only the SIR case, and he made the unusual assumption that the measurement errors for non-centric reflections are errors in the complex plane.

In most work, including that of Blow & Crick (1959) and Terwilliger & Eisenberg (1987), it is assumed that MIR probabilities can be obtained by multiplying together a series of SIR distributions. But as Raiz & Andreeva (1970) and Einstein (1977) have noted, this gives too great emphasis to the measurement of $F_o$, which is included in each SIR distribution. Einstein gave a clear account of how to combine the information from several derivatives. However, he treated only the non-centric case and did not include the effects of non-isomorphism.

I will adopt a similar strategy to Einstein (1977) to combine the information from several derivatives. However, the approximations will be more similar to those used by Terwilliger & Eisenberg (1987). A new element to this work is in the treatment of non-isomorphism. Terwilliger & Eisenberg recognized, but neglected, the component of the error from non-isomorphism that is negatively correlated to the protein structure factor. This component of the error is covered by probability relationships derived for structure factors from related structures (Read, 1990).

## 2. Theory

a) Overall joint probability distribution

The first major goal will be to derive an overall joint conditional probability distribution, $p(F,\alpha,F_o,(F_{jo})_{j=1,N}|(\mathbf{H}_{jc})_{j=1,N})$, from which other distributions of interest can be derived by standard manipulations. We start with a general statement of the multiplication law for probabilities in equation (1).

$$p(a,b,...,m,n|z) = p(a|z)p(b|a,z)...p(n|a,b,...,m,z). \quad (1)$$

A very general expression for the desired distribution could be given in a form similar to equation (1). However, we can simplify this by remembering that, if x and y are statistically independent, p(x|y)=p(x). In principle, derivatives that are related by, for instance, shared heavy atom sites could be treated by allowing for statistical dependence, but we will assume that all N derivatives are independent.

$$p(F,\alpha,F_o,(F_{jo})_{j=1,N}|(\mathbf{H}_{jc})_{j=1,N}) = p(F,\alpha,F_o)\prod_{j=1}^{N} p(F_{jo}|F,\alpha,\mathbf{H}_{jc}). \quad (2)$$

i) Measurement error

The distribution for the native structure factor, $p(F,\alpha,F_o)$, will have different expressions in the centric and non-centric cases, as shown by equations (3). (The subscript c will be used for probability distributions that apply to centric reflections only, and n for non-centric distributions. Normalization factors will be given as K, the factor required for the integral over the variables of the joint distribution to be unity.)

$$p_c(F,\alpha,F_o) = K\exp(-\frac{F^2}{2\varepsilon\Sigma_N})\exp(-\frac{(F-F_o)^2}{2\sigma^2}). \quad (3a)$$

$$p_n(F,\alpha,F_o) = KF\exp(-\frac{F^2}{\varepsilon\Sigma_N})\exp(-\frac{(F-F_o)^2}{2\sigma^2}). \quad (3b)$$

We are assuming a Gaussian measurement error for the amplitude F. The parameter $\varepsilon\Sigma_N$ is the variance characterizing the Wilson (1949) distributions, and $\varepsilon$ is the expected intensity factor that arises from crystal symmetry (see, *e.g.*, Stewart & Karle, 1976). For useful diffraction data, the measurement error will be much smaller than the expected range of the amplitudes; so $\sigma^2 << \varepsilon\Sigma_N$. If $F_o > 3\sigma$, say, the Wilson distributions will be relatively constant over the range of F for which the second exponential in equations (3) has significant values. Therefore, equation (4) is a reasonable approximation for moderately-strong centric and non-centric reflections.

$$p(F,\alpha,F_o) = K\exp(-\frac{(F-F_o)^2}{2\sigma^2}). \qquad (4)$$

Weak reflections tend to be poorly phased, so errors in their distributions are probably tolerable. In any event, since the approximation of a Gaussian error distribution is particularly poor for weak structure factors, even equations (3) would be inaccurate.

Similarly, we will assume a Gaussian error distribution in the measurement of the derivative amplitude, given by equation (5).

$$p(F_{jo}|F_j) = K\exp(-\frac{(F_j-F_{jo})^2}{2\sigma_j^2}). \qquad (5)$$

ii) Conditional probability of the derivative amplitude

We approach the conditional probability distribution $p(F_j|F,\alpha,\mathbf{H}_{jc})$ by considering two crystals, the first being the true derivative crystal and the second a hypothetical derivative crystal. The hypothetical crystal contains all the atoms in the native crystal, with identical fractional coordinates and B-factors, plus the heavy atoms in the heavy atom model. Although we label some of the differences between the two crystals as non-isomorphism and others as errors in the heavy atom model, they all consist of differences in fractional coordinates and scattering factors. (Note that an incomplete heavy atom model can be considered to contain the missing atoms, but with a zero scattering factor.) Recently, I have shown (Read, 1990) that the combination of such differences between related structures gives rise to probability relationships similar to those for simpler cases (Luzzati, 1952; Sim, 1959; Woolfson, 1956; Srinivasan & Ramachandran, 1965).

Non-isomorphism can result from several types of difference between two crystals. There could be random or systematic changes of the fractional coordinates for the protein atoms. Systematic changes could be the result of rotation or translation, or of changes in cell dimensions (Crick & Magdoff, 1956). In addition, there could be perturbations of the B-factors of the protein atoms. No doubt pathological cases are possible, but numerical tests (results not shown) indicate that the structure factor distributions are remarkably similar, whether random or systematic changes are involved. The distribution of possible values for $\mathbf{F}_{pj}$, the structure factor contribution from the protein part of the derivative crystal, is approximated by a Gaussian centered on $D_j\mathbf{F}$. For centric reflections, the Gaussian is one-dimensional, and it is two-dimensional for non-centric reflections. $D_j$ is a resolution-dependent factor that depends on the differences between the two structures. In the simple case of coordinate differences drawn from a single Gaussian distribution, $D_j$ is equivalent to the effect of an overall B-factor (Luzzati, 1952; Read, 1990). It is usually less than one, so the error due to non-isomorphism will usually be negatively correlated with F.

For errors in the heavy atom model, we can incorporate the factor equivalent to $D_j$ into $\mathbf{H}_{jc}$. In fact, as shown previously (Read, 1990), the effect of an error in the coordinates for an atom can be modeled fairly well by an increase in its B-factor. So errors and missing atoms in the heavy atom model should lead approximately to a Gaussian distribution of $\mathbf{H}_j$ centered on $\mathbf{H}_{jc}$.

$\mathbf{F}_{jc}$ ($=D_j\mathbf{F}+\mathbf{H}_{jc}$) thus has two independent Gaussian error contributions as an estimate of $\mathbf{F}_j$ ($=\mathbf{F}_{pj}+\mathbf{H}_j$). If we denote the combined error as $\Delta_j$, and the combined variance as $\varepsilon\sigma_{\Delta j}^2$, the

probability distributions for $\Delta_j$ are also Gaussian, as shown in equations (6). The distribution for the non-centric case is illustrated schematically in Figure (2).

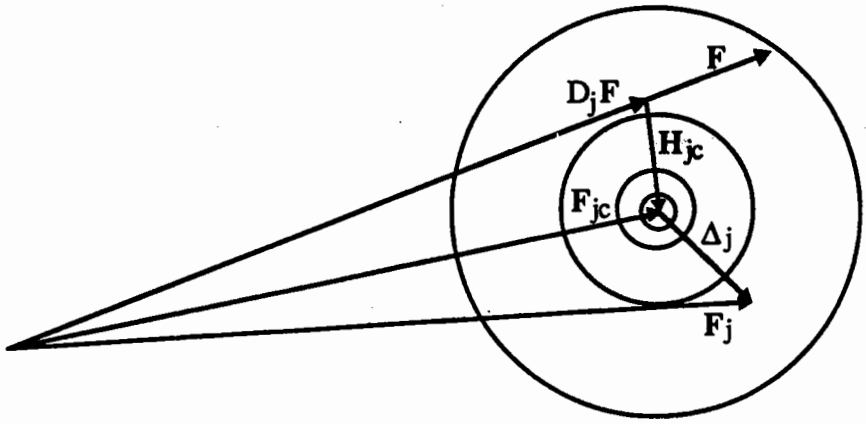



**Figure 2.** Schematic representation of the probability distribution $p(\Delta_j)=p(\mathbf{F}_j|F,\alpha,\mathbf{H}_{jc})$ for the non-centric case. The circles represent a two-dimensional Gaussian distribution with variance $\varepsilon\sigma_{\Delta j}^2$.

$$p_c(\Delta_j|F,\alpha,\mathbf{H}_{jc}) = K\exp(-\frac{|\mathbf{F}_{jc}-\mathbf{F}_j|^2}{2\varepsilon\sigma_{\Delta j}^2}), \tag{6a}$$

$$p_n(\Delta_j|F,\alpha,\mathbf{H}_{jc}) = K\exp(-\frac{|\mathbf{F}_{jc}-\mathbf{F}_j|^2}{\varepsilon\sigma_{\Delta j}^2}). \tag{6b}$$

The error in the amplitude measurement is independent of $\Delta_j$, so the distributions in equations (5) and (6) are multiplied to get equations (7).

$$p_c(F_{jo},\Delta_j|F,\alpha,\mathbf{H}_{jc}) = K\exp(-\frac{\Delta_j^2}{2\varepsilon\sigma_{\Delta j}^2}-\frac{(F_j-F_{jo})^2}{2\sigma_j^2}), \tag{7a}$$

$$p_n(F_{jo},\Delta_j|F,\alpha,\mathbf{H}_{jc}) = K\exp(-\frac{|\Delta_j|^2}{\varepsilon\sigma_{\Delta j}^2}-\frac{(F_j-F_{jo})^2}{2\sigma_j^2}). \tag{7b}$$

The expressions required for equation (2) are obtained by integrating out $\Delta_j$. This is fairly straightforward for the centric case. Strictly, $F_j=|\mathbf{F}_{jc}+\Delta_j|$, but substituting $F_{jc}+\Delta_j$ introduces errors only for weak amplitudes of the order of $\sigma_{\Delta j}$. Integrating over $\Delta_j$ gives equation (8).

$$p_c(F_{jo}|F,\alpha,\mathbf{H}_{jc}) = K\exp(-\frac{(F_{jc}-F_{jo})^2}{2\sigma_j^2+2\varepsilon\sigma_{\Delta j}^2}). \tag{8}$$

For the non-centric case, $F_j$ can be approximated by a second order Taylor expansion, and the result is integrated to give equation (9).

$$p_n(F_{jo}|F,\alpha,\mathbf{H}_{jc}) = K\exp(-\frac{(F_{jc}-F_{jo})^2}{2\sigma_j^2+\varepsilon\sigma_{\Delta j}^2}). \tag{9}$$

Equation (8) differs from equation (9) only by the factor of 2 applied to $\varepsilon\sigma_{\Delta j}^2$. All data can thus be handled by equation (10), in which $\varepsilon'$ equals either $2\varepsilon$ for centric or $\varepsilon$ for non-centric reflections.

$$p(F_{jo}|F,\alpha,\mathbf{H}_{jc}) = K\exp(-\frac{(F_{jc}-F_{jo})^2}{2\sigma_j^2+\varepsilon'\sigma_{\Delta j}^2}). \tag{10}$$

With equation (10), the components necessary for the overall joint distribution of equation (2) have all been assembled. The result is given in equation (11).

$$p(F,\alpha,F_o,(F_{jo})_{j=1,N}|(\mathbf{H}_{jc})_{j=1,N}) = K\exp(-\frac{(F-F_o)^2}{2\sigma^2} - \sum_{j=1}^{N}\frac{(F_{jc}-F_{jo})^2}{2\sigma_j^2+\varepsilon'\sigma_{\Delta j}^2}). \quad (11)$$

b) Conditional probability of the native structure factor

The probability distribution for the native structure factor (amplitude and phase), conditional on the observed amplitudes and heavy atom model, is obtained from equation (11) by fixing the values of the observations and renormalizing to give equation (12).

$$p(F,\alpha|F_o,(F_{jo},\mathbf{H}_{jc})_{j=1,N}) = K\exp(-\frac{(F-F_o)^2}{2\sigma^2} - \sum_{j=1}^{N}\frac{(F_{jc}-F_{jo})^2}{2\sigma_j^2+\varepsilon'\sigma_{\Delta j}^2}). \quad (12)$$

Equation (12) can be interpreted as combining information from several independent sources.

$$p(F,\alpha|F_o,(F_{jo},\mathbf{H}_{jc})_{j=1,N}) = p(F,\alpha|F_o)\prod_{j=1}^{N}p(F,\alpha|F_{jo},\mathbf{H}_{jc}), \text{ where} \quad (13a)$$

$$p(F,\alpha|F_{jo},\mathbf{H}_{jc}) = K\exp(-\frac{(F_{jc}-F_{jo})^2}{2\sigma_j^2+\varepsilon'\sigma_{\Delta j}^2}). \quad (13b)$$

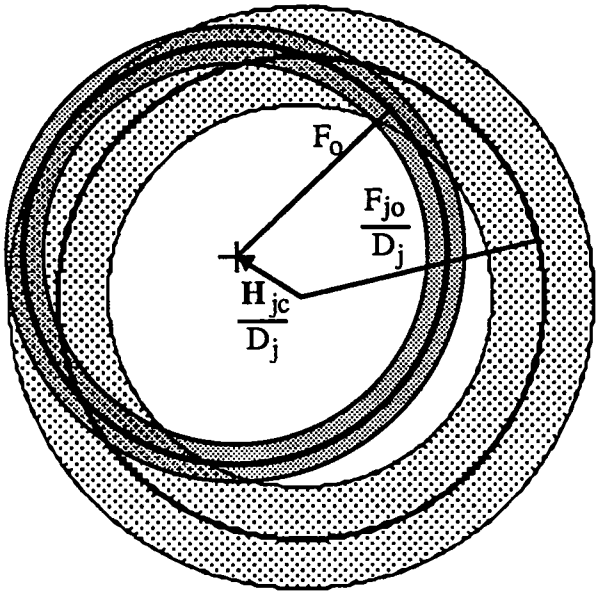



**Figure 3.** Schematic representation of the combination of information from native and derivative data. The shaded regions indicate high probability for either $p(F,\alpha|F_o)$ (dark shading) or $p(F,\alpha|F_{jo},\mathbf{H}_{jc})$ (light shading). The combined probability distribution $p(F,\alpha|F_o,F_{jo},\mathbf{H}_{jc})$ is obtained by multiplying together the two independent distributions.

Equation (13b) is derived similarly to equation (10). Equation (13a) embodies the approach used by Einstein (1977) to construct an MIR probability distribution. Figure (3), which is similar to figure 1(b) from Einstein (1977), gives a schematic picture of the combination of probability distributions from native data and a single derivative. Basically, Figure (3) can be seen as a Harker (1956) construction in which the circles are smeared out to represent the effect of uncertainty. Further derivatives would be included by multiplying in their probability distributions. With such a picture, it is easier to see how the information from several derivatives is combined without overemphasizing the native amplitude measurement.

Following the arguments of Blow & Crick (1959), the rms error in the electron density map could be minimized by using the probability distribution in equation (12) to estimate $\langle F\exp(i\alpha)\rangle$. However, errors in the phase are much more important than errors in the amplitude, so it is reasonable to follow conventional practice and integrate out F from equation (12) to obtain the MIR probability distribution for the phase. The integral is approximated

using a second-order Taylor expansion of F about $F_o$, integrating, ignoring some terms that vary relatively little with $\alpha$, and assuming that $H_{jc}\sin(\alpha-\alpha_{jc})<F_{jco}$. The result is given in equations (14), where all of the sums are over the N derivatives.

$$p(\alpha|F_o,(F_{jo},H_{jc})_{j=1,N}) = K\exp\left(+\frac{(\Sigma\lambda_j)^2}{\frac{1}{2\sigma^2}+\Sigma\kappa_j} - \Sigma\frac{(F_{jco}-F_{jo})^2}{2\sigma_j^2+\varepsilon'\sigma_{\Delta j}^2}\right), \text{ where} \tag{14a}$$

$$\lambda_j = \frac{\pm D_j(F_{jco}-F_{jo})}{2\sigma_j^2+\varepsilon'\sigma_{\Delta j}^2}, \text{ and} \tag{14b}$$

$$\kappa_j = \frac{D_j^2}{2\sigma_j^2+\varepsilon'\sigma_{\Delta j}^2}. \tag{14c}$$

The sign for $\lambda_j$ is usually positive but is negative for cross-overs, *i.e.* when $[D_jF_o+H_{jc}\cos(\alpha-\alpha_{jc})]$ is negative. The result in equations (14) is reasonably similar to the result obtained by Einstein (1977), differing primarily because it includes the factor $D_j$ and omits some non-exponential terms.

The phase probability distribution in equations (14) is a complicated function of the phase, as the phase enters into each value of $F_{jco}$ (see Table 1). It will be necessary to use numerical integration to evaluate the expected values of any functions of the phase, including the phase itself.

c) Joint distribution of the observed amplitudes

To make use of the phase probability distribution in equations (14), estimates are required for all the parameters on which it depends, from the variance $\sigma_{\Delta j}^2$ to the heavy atom coordinates, B-factors and occupancies. The most generally applicable method to estimate the parameters of probability distributions is the maximum likelihood method. To apply this method, we need the joint distribution function of all the observed values. The optimal set of parameters will be those that maximize the likelihood of having made the set of observations. If we assume that the observations for each hkl are independent, the likelihood function is given by the product (over all reflections) of the joint probability distributions of the observed amplitudes for each reflection.

This joint distribution is obtained by integrating out F and $\alpha$ from equation (11). Since the expression in equation (11) differs from that in equation (12) only by the normalization constant, the integration over F is carried out as for equation (14). The joint distribution of the amplitudes is then given by equation (15), where the integration over $\alpha$ must be carried out numerically.

$$p(F_o,(F_{jo})_{j=1,N}|(H_{jc})_{j=1,N}) = K\int_0^{2\pi} d\alpha \exp\left(+\frac{(\Sigma\lambda_j)^2}{\frac{1}{2\sigma^2}+\Sigma\kappa_j} - \Sigma\frac{(F_{jco}-F_{jo})^2}{2\sigma_j^2+\varepsilon'\sigma_{\Delta j}^2}\right). \tag{15}$$

d) The SIR case

In the SIR case, equation (14) simplifies considerably, giving equation (16).

$$p(\alpha|F_o,F_{jo},H_{jc}) = K\exp\left(-\frac{(F_{jco}-F_{jo})^2}{2\sigma_j^2+\varepsilon'\sigma_{\Delta j}^2+2D_j^2\sigma^2}\right). \tag{16}$$

If the factors $D_j$ and $\varepsilon$ are both equal to one, this is equivalent to equations (13) and (16) of Terwilliger & Eisenberg (1987). However, they assumed that the MIR case can be handled by multiplying together several SIR distributions, which differs considerably from equation (14). The significance of this difference will be examined below.

## 3. Some numerical tests

To derive equations (14) and (15), a number of approximations have been made. Some were necessary because the expressions cannot be integrated analytically, while others were convenient or simplified the results. To assess whether the approximations are reasonable, three numerical tests have been performed. The first is intended to test the distributions when non-isomorphism provides a major source of uncertainty. The second tests the case in which measurement error predominates. The third tests the ability of an existing MIR phase refinement program to deal with the errors in test data sets. For these tests, native data were calculated to 2Å resolution from the structure of *Streptomyces griseus* trypsin (SGT, Read & James, 1988), space group $C222_1$.

a) Test 1: non-isomorphism

Data were computed for two hypothetical derivatives. The first, a mercury derivative with 3 Hg sites (B=30Å$^2$ for each Hg atom), was seriously non-isomorphous. Non-isomorphism was simulated by changing the cell dimensions (a and b were increased by 1%, and c decreased by 1%), rotating the protein by 0.6° about z, and translating it so that its center moved by 0.2Å in x, y and z. With these changes, $D_j$ varies from near 1 at low resolution to about 0.8 at 2Å. Further uncertainty was added by using only 2 of the 3 Hg sites in the heavy atom model, and by adding Gaussian errors to the amplitudes [with $\sigma(\sigma_j)=20+0.025F(F_j)$]. The second derivative, a zinc derivative with 2 Zn sites (B=25 or 35Å$^2$), was perfectly isomorphous. However, the second Zn site was omitted from the heavy atom model, and measurement errors were added ($\sigma_j=20+0.02F_j$). Table 2 presents R-factors that give an impression of the size of the errors introduced in the data.

**Table 2: R-factors* for Test 1**

| Comparison | Hg derivative | Zn derivative |
|---|---|---|
| F *vs.* $F_{pj}$ | 0.271 | 0.0 |
| $F_{pj}$ *vs.* $F_j$ | 0.276 | 0.085 |
| F *vs.* $F_j$ | 0.382 | 0.085 |
| $F_j$ *vs.* $F_{jo}$ | 0.059 | 0.057 |
| $F_o$ *vs.* $F_{jo}$ | 0.392 | 0.122 |
| F *vs.* $F_o$ | 0.061 | |

* R-factor for $F_a$ *vs.* $F_b$ is defined as $\frac{\Sigma|F_a-F_b|}{\Sigma F_a}$

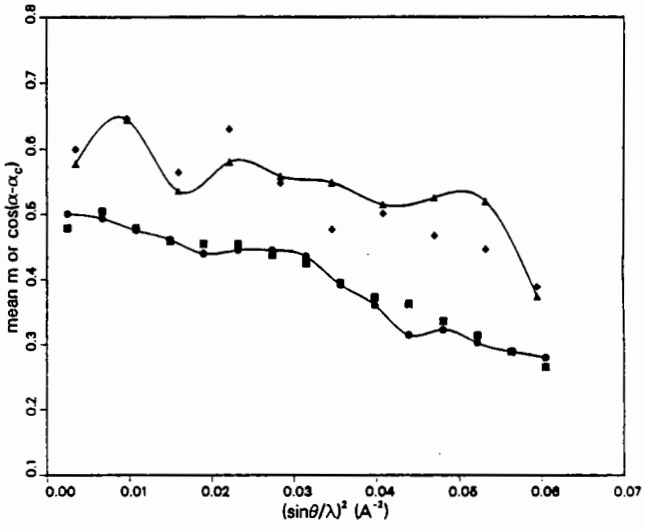



**Figure 4.** Comparison, for the test 1 data, of the mean cosine of the phase error (non-centric data: circles; centric data: triangles) with the mean figure of merit (non-centric: squares; centric: diamonds). Equation (14) was used to compute centroid phases and figures of merit.

Phase probabilities were computed for each reflection using equation (14), and the centroid phase and figure of merit (Blow & Crick, 1959) were evaluated. If the phase probabilities are correct, the mean figure of merit should be equal to the mean cosine of the error in the centroid

phase. Figure 4 shows a comparison of these quantities as a function of resolution. There is good agreement for both centric and non-centric data. Equally good agreement was obtained when the data were divided into bins according to the predicted figure of merit (results not shown).

To test the importance of the factor $D_j$, the phase calculations were repeated with $D_j$ set to 1. Surprisingly, the reduction in phase accuracy was marginal, except for the most intense reflections; there was an overall decrease from 0.391 to 0.389 in the mean cosine of the phase error. Apparently, when $D_j$ differs significantly from 1 the variances are so large that the phase probability distributions are broad and relatively insensitive to the amplitudes. On the other hand, preliminary tests indicate that $D_j$ must be allowed to vary from 1 to get good results in the maximum likelihood estimation of the variances. Neglecting $D_j$ is similar to assuming a Sim (1959) distribution for a structural model with coordinate errors. It has been shown for such cases that the assumption of a Sim distribution leads to underestimates of the variances (Read, 1986).

b) Test 2: measurement errors

To test the practical difference between equation (14) and the one derived by multiplying together the SIR distributions from equation (16), data were generated for 5 different single-site Zn derivatives (B=35Å$^2$). The derivatives were all perfectly isomorphous, and the heavy atom models were perfect. The only errors were in the amplitude observations [$\sigma(\sigma_j)$=25+0.025F($F_j$)]. However, these errors were larger on average than the amplitude changes from the heavy atom substitution; the R-factors between F and $F_j$ were all 0.052, while the R-factors between the true and "observed" amplitudes were all 0.072. The heavy atom signal, then, was largely hidden in random observational noise, with the R-factor between $F_o$ and $F_{jo}$ being 0.117.

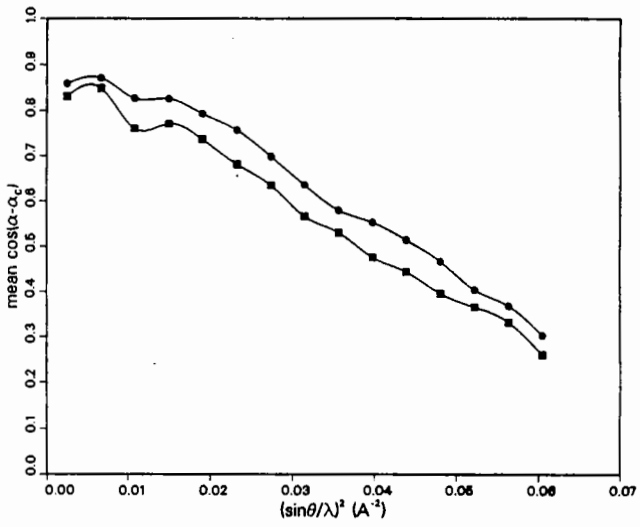



**Figure 5.** Comparison of phase accuracy (mean cosine of the phase error) using two methods to evaluate the phases for the test 2 data set. Centroid phases were computed either using equation (14) (circles) or by multiplying together the SIR distributions from equation (16) (squares).

As in the first test, the mean cosine of the phase error agrees very well with the mean figure of merit. It is more interesting to see (Figure 5) that there is a substantial reduction in the phase accuracy when the phase probabilities are evaluated by multiplying together 5 SIR probability distributions. The mean cosine of the phase error drops from 0.574 to 0.517.

c) Test 3: heavy atom refinement

Otwinowski (1990, 1991) has implemented maximum-likelihood refinement within the program PHARE (Bricogne, 1982). This program is based on a probability model similar to that of Blow and Crick (1959). The chief differences with the probability model proposed here are in the combination of independent derivatives, and in the treatment of the variances for

centric and non-centric reflections. The factor $D_j$ is not included explicitly, but in a sense is modelled by the scale and B-factors that are refined for each derivative.

**Table 3: Results from Otwinowski's PHARE**

| | Non-centric | | Centric | |
|---|---|---|---|---|
| | mean cos(Δα) | mean m | mean cos(Δα) | mean m |
| Test 1: | | | | |
| Equation (14) | 0.371 | 0.378 | 0.538 | 0.527 |
| PHARE | 0.343 | 0.341 | 0.469 | 0.586 |
| Test 2: | | | | |
| Equation (14) | 0.542 | 0.546 | 0.818 | 0.821 |
| 5×SIR | 0.485 | 0.462 | 0.765 | 0.762 |
| PHARE | 0.483 | 0.284 | 0.765 | 0.600 |

The results from running PHARE on the two test data sets are summarized in Table 3. For comparison, the results obtained above are also included. PHARE performed reasonably well on the test 1 data, better on the non-centric than on the centric data. One might expect an improvement in the figures of merit for centric data with a better treatment of the variances. The refined scale and B-factors mimicked the resolution dependence of $D_j$ fairly well, and the refinement of the heavy atom parameters was very stable (results not shown). For the test 2 data, the phases were only as accurate as those determined by multiplying together the 5 SIR distributions; in addition, the figures of merit were significantly underestimated. This suggests that a significant improvement could be achieved by the proper treatment of the combination of derivative information.

## Conclusions

This analysis of the propagation of errors in the isomorphous replacement method combines elements from the treatments by Terwilliger & Eisenberg (1987) and by Einstein (1977). A new feature is the incorporation of the factor $D_j$, which accounts for the component of the error from non-isomorphism that is negatively correlated with the native structure factor. The results of numerical tests are consistent with the probability distributions that have been derived. As yet, no comparisons have been made with the distributions derived by Einstein (1977). He retained non-exponential terms that have been eliminated by the approximations employed here; it will be interesting to see if such terms would increase the accuracy significantly.

It is proposed that the maximum likelihood method be used to estimate the parameters in the probability distributions. In essence, Terwilliger & Eisenberg (1987) proposed estimating the lack-of-closure variance by maximum likelihood, using the phase probabilities to compute the expected value of the mean-square lack-of-closure errors. Otwinowski (1990, 1991) has shown that such a procedure gives a dramatic improvement in both the behaviour and the results of heavy-atom refinement, compared to the earlier methods using the centroid or most probable phases. Tests performed on data with known errors show that his approach gives good, but not yet optimal, results. A maximum likelihood analysis applied to the probability distributions derived here should lead to refinement procedures that are even more robust and reliable in the presence of non-isomorphism.

## Acknowledgements

Trevor N. Hart provided much-appreciated advice on "torturing" the integrals. The author is an Alberta Heritage Foundation for Medical Research Scholar.

## References


Blow, D.M. and Crick, F.H.C. Acta Cryst. **12**: 794-802 (1959).

Bricogne, G. **In**: Computational Crystallography (Sayre, D., ed.), Oxford University Press, Oxford, pp. 223-230 (1982).

Crick, F.H.C. and Magdoff, B.S. Acta Cryst. **9**: 901-908 (1956).

Einstein, J.R. Acta Cryst. **A33**: 75-85 (1977).

Green, E.A. Acta Cryst. **A35**: 351-359 (1979).

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

Hendrickson, W.A. and Lattman, E.E. Acta Cryst. **B26**: 136-143 (1970).

Luzzati, V. Acta Cryst. **5**: 802-810 (1952).

Otwinowski, Z. ACA 1990 annual meeting, New Orleans, abstract no. C04 (1990).

Otwinowski, Z. This volume (1991).

Raiz, V.Sh. and Andreeva, N.S. Sov. Phys. Crystallogr. **15**: 206-210. Translated from Kristallografiya **15**: 246-251 (1970).

Read, R.J. Acta Cryst. **A42**: 140-149 (1986).

Read, R.J. Acta Cryst. **A46**: 900-912 (1990).

Read, R.J. and James, M.N.G. J. Mol. Biol. **200**: 523-551 (1988).

Rossmann, M.G. and Blow, D.M. Acta Cryst. **14**: 641-647 (1961).

Sim, G.A. Acta Cryst. **12**: 813-815 (1959).

Srinivasan, R. and Ramachandran, G.N. Acta Cryst. **19**: 1008-1014 (1965).

Stewart, J.M. and Karle, J. Acta Cryst. **A32**: 1005-1007 (1976).

Terwilliger, T.C. and Eisenberg, D. Acta Cryst. **A43**: 6-13 (1987).

Wilson, A.J.C. Acta Cryst. **2**: 318-321 (1949).

Woolfson, M.M. Acta Cryst. **9**: 804-810 (1956).

# MAXIMUM LIKELIHOOD REFINEMENT OF HEAVY ATOM PARAMETERS

Zbyszek Otwinowski

Howard Hughes Medical Institute
and Department of Molecular Biophysics and Biochemistry,
Yale University, New Haven, CT 06514, USA

## Introduction

Analysis of isomorphous replacement (MIR) data has two major steps. The first is refinement of the heavy atom substitution parameters. The second is application of the results of the first step to calculate probabilities of all possible phases and from that distribution the centroid phase and figure of merit.

It was recognized in the early days of protein crystallography that in the second step of MIR analysis it is preferable to consider simultaneously all possible phases rather than only the most likely phase[1]. The calculated value of the complex structure factor is a weighted average, with the weights being equal to the probability of each phase being correct. In contrast, estimates of the heavy atom parameters were obtained by a least squares procedure that considered only one phase for each reflection. Such refinement, under most circumstances, produces highly inaccurate (biased) estimates of heavy atom substitution parameters.

The novelty of the approach presented here is in applying the phase probability as a weight in heavy atom refinement as well as during the structure factor calculation. The program ML-PHARE (**M**aximum **L**ikelihood **PHA**se **RE**finement) corrects the inconsistency between the refinement and the structure factor calculation steps, resulting in an unbiased refinement[2,3,4].

## Likelihood Function

Behind every refinement procedure there is an explicit or implicit probabilistic model. The likelihood $\mathcal{L}(\overline{F}_p, F_o, F_{d_1}, F_{d_2}, ..., \overline{F}_{h_1}, \overline{F}_{h_2} ...)$ is a function of complex parent structure factors $\overline{F}_p$, observed parent structure-factor amplitudes $F_o$, observed derivative structure-factor amplitudes $F_{d_1}, F_{d_2}, ...$, heavy atom structure factors $\overline{F}_{h_1}, \overline{F}_{h_2}, ...$, error estimates, scale factors etc..

The observed intensities have known gaussian errors due to counting statistics, radiation damage, etc.; these errors typically are approximated by gaussian errors in the observed amplitude $F_o$:

$$\mathcal{L}(F_o, \overline{F}_p) = e^{-\frac{|F_o - F_p|^2}{2\sigma_o^2}}$$

where $\sigma_o$ represents errors of the measurement; all errors being combined in one error estimate. This approximation is deficient for weak reflections with intensity about or below its measurement error.

Observed isomorphous differences are affected by measurement errors of scattering amplitudes and by lack of isomorphism - change in the macromolecular structure that is not properly modeled. Lack of isomorphism can be modeled by a a gaussian function of complex structure factors:

$$\mathcal{L}(\overline{F}_p, \overline{F}_d) = e^{-\frac{|\overline{F}_d - (\overline{F}_p + \overline{F}_h)|^2}{2\sigma_i^2}}$$

where $\sigma_i^2$ represents resolution dependent magnitude of lack of isomorphism. The estimate of the lack-of-isomorphism error should be function of reflection centricity and multiplicity. This dependency is often ignored; consequences of that are discussed by Read[3].

If one tries to determine the heavy atom structure, one should integrate likelihood over all possible values of parent and derivative complex structure factors. If amplitude errors are small, then the dominant component will be the likelihood value at the most likely amplitude value - the observed one. If the phase is well determined, then the likelihood integration can be restricted to the most likely phase value. If one makes the preceeding simplification for all structure factors, finding the maximum of likelihood becomes equivalent to finding the minimum of $\chi^2$.

The likelihood function of individual reflections should be multiplied to generate the global likelihood function. This function is maximized with respect to all adjustable parameters during the refinement procedure.

The maximum likelihood method, unlike minimum $\chi^2$, can also fit the $\sigma_i$ values simultaneously with refinement of other parameters, but this has not yet been implemented.

### Blow and Crick Likelihood Function

In the current version of program ML-PHARE only the simple form of likelihood defined by Blow and Crick has been implemented. This assumes that errors in measurements of parent amplitudes are smaller than the combination of errors in lack-of-isomorphism

and in the measurements of heavy atom derivative amplitudes. This likelihood function also assumes that once a particular choice of parent phase has been made the derivative phase is well determined. The latter assumption is valid if the lack-of-isomorphism errors are smaller than parent amplitudes. The dependence of lack-of-isomorphism errors on centricity is ignored. These assumptions work quite well in a majority of cases.

Dickerson[5] proposed minimizing the 'lack-of-closure' residual $\left||F_o e^{i\phi} + \overline{F}_h| - F_d\right|^2$. The related form of likelihood function for a single derivative is:

$$\mathcal{L}(\phi, \overline{F}_h) = e^{-\frac{\left||F_o e^{i\phi} + \overline{F}_h| - F_d\right|^2}{2\sigma_i^2}}$$

This formula assumes that errors of the parent structure factors are ignored or added to errors of the derivative structure factor.

The Blow and Crick formulation was implemented here due to its simplicity and compatability with existing program PHARE[6]. Program ML-PHARE can be further upgraded with a more accurate likelihood function[2,3,7,8] within the current program structure.

## Least Squares vs Maximum Likelihood

Any statistical analysis of the data starts with a likelihood model which describes the probability of the outcome of an experiment for any particular set of data. Often, instead of likelihood, a $\chi^2$ statistic is used. The $\chi^2$ is derived from a Taylor expansion of likelihood around its maximum. The proper procedure should involve the expansion around the maximum of the product of the individual *hkl* likelihoods; however, this would require, first finding the maximum of the global likelihood function. In the simplification that is generally used, $\chi^2$ is derived from the expansion of likelihood around its individual *hkl* maxima. This procedure depends upon the likelihood function being well approximated by a gaussian function, which is frequently invalid in crystallography. This departure from the assumptions behind the minimum $\chi^2$ method is the cause of the significant bias encountered in heavy atom refinement.

## Bias in Least Squares Refinement

The most frequent departure from gaussian approximation to likelihood happens when the parent phase is poorly defined. The range of likely parent phases is well described by the figure of merit - the expected value of cosine between the true value and the 'best' estimate of the phase. If the figure of merit is close to one, than the minimum $\chi^2$ method works well with low bias. This is the rationale behind some methods used to improve least squares refinement procedures. One such method uses only

reflections with high figure of merit, another restricts refinement to centric reflections only. The latter approach helps as centric reflections have, on average, higher figure of merit than non-centric ones. It is worth stressing that the restricting refinement to centric reflections does not eliminate bias. Refinement with low figure of merit centric reflections is as biased as with acentric reflections of the same figure of merit.

It is easier to analyze bias in the refinement of only a single derivative or a comparable case of an MIR analysis with only one good derivative. Because parent phase is unknown, only the magnitude of $F_h$ can be refined against observed differences $|F_o - F_d|$. Least square refinement that minimizes lack of closure will make the average magnitude of $F_h$ close to the average of $|F_o - F_d|$. Any experimental or lack-of-isomorphism errors will, on average, enlarge observed differences, resulting in refined occupancies that are too high. Under some circumstances the occupancy refinement bias can even be negative for non-centric refinement. The program PHARE has the option to calculate least square lack of closure residual based on either the most likely or the centroid ('best') phase. The bias varies dependent on this choice. To understand the bias produced by the program PHARE, one should be aware of the program's automatic switch to centroid phase for reflections with only a single derivative without anomalous scattering data. Generally, only heavy atom occupancies and lack-of-closure estimates are severly biased; atomic positions, on the other hand, are quite accurate. One can think of atomic temperature factors as descriptors of resolution-dependent occupancies. Because bias in the occupancy refinement has strong resolution dependence, least square refined temperature factors can be very wrong.

The reason behind the overall success of least square refinement is that refined phases (but not amplitudes) of a heavy atom constellation are almost unbiased. The errors in occupancy parameters produce an incorrect difference between heavy atom and parent phase, however in most cases the effect is moderate.

### Estimating Lack of Closure Error

Knowing scattering amplitudes of the heavy atoms constellation one can calculate parent phases. The uncertainty in the value of the parent phase comes from experimental errors and lack of isomorphism. The magnitude of the measurement error is usually known from counting statistics and other known sources of experimental error. On the other hand, experimental data do not provide direct estimates of lack-of-isomorphism errors, they have to be estimated during the heavy atom refinement process.

The traditional approach was to assume that the calculated mean square lack of closure is a sum of two components, the square of the experimental error and the square of the lack of isomorphism. By this method, estimates of the lack of isomorphism are obtained from reversing the preceeding assumption. This procedure has been implemented in a number of programs. Terwilliger and Eisenberg[9] noticed that lack of closure should be summed over all possible parent phases, weighted by the probability of

each phase being correct. This is equivalent to the maximum likelihood estimate of the lack-of-closure error if isomorphous differences are on average much smaller than the parent structure factors.

The traditional calculation of lack of isomorphism could underestimate it by a large factor. This resulted in a refinement with highly inflated figures of merit and phasing power statistics. For this reason these important statistics were treated suspiciously, as they lost discriminatory power. With maximum likelihood refinement these statistics can be used to properly estimate the quality of derivatives. One has to deflate one's expectation (in terms of the above statistics) for what represents a good derivative, as the results of ML-PHARE refinement are cosmetically worse than the results of PHARE refinement. The maximum likelihood method refines phasing power to a lower value than the least square method. This decrease produces more realistic the figure-of-merit values and results in maps of higher quality.

### Refining a Single Derivative

Failure of the program PHARE to refine occupancy correctly is due to a particularly strong bias in the refinement of the heavy atom parameters when only one derivative is used. One can only fit the modulus of the calculated heavy atom constellation's scattering factor to the modulus of the observed difference between the parent and the derivative structure factor. For this reason even wrong heavy atom constellation will reduce the lack of closure, as defined in PHARE.

It was recognized before that such least squares refinement is biased[10]. In the SIR refinement case, it is possible to calculate exactly how much bias is introduced and correct it. There are both reciprocal space and real space (Patterson function) versions of such refinements[11].

The Shortcomings of the above method reside in assuming uniform weights for different reflections, and more importantly, in not being able to use cross-phasing information between derivatives.

### Refining Multiple and Anisotropic Sites

The refinement is only as good as the model being refined. Very often the heavy atom sites are disordered and/or clustered. This happens frequently with soaked-in heavy atoms that bind on the surface of a protein. Anisotropic refinement can be very useful in cases of heavy atom structures that are more complicated than a sum of separated isotropic gaussian peaks.

Anisotropic refinement can indicate the need for splitting one site into two; the program ML-PHARE also prints the coordinates of such possibly split sites. Splitting a single site into two closely placed ones adds many highly correlated parameters to the

refinement, thus making it intristically less stable. The high stability of ML-PHARE refinement makes it easier to judge if splitting the sites is actually helpful or not.

## Summary

The program PHARE was modified to include the uncertainty of the protein phase. The lack-of-closure residual is now weighted by the probability of the phase being correct. This was previously applied to the estimate of the lack-of-isomorphism errors[9], but is now extended to include the residual used in the refinement. The program PHARE was also modified to facilitate splitting the complex sites into closely separated ones. With these modifications, the original heavy-atom refinement is significantly improved.

## Acknowledgements

This research was supported in part by grants from the USPHS (GM 22324 and GM 15225 to P. Sigler).

## REFERENCES


1. Blow, D.M. and Crick, F.H.C. (1959). **Acta Cryst. 12**, 794-802.

2. Bricogne, G. Presented at meeting and in the proceedings of the meeting.

3. Read, R.J. Presented at meeting and in the proceedings of the meeting.

4. Otwinowski, Z., (1989). Ph D thesis, University of Chicago, Chicago Illinois.

5. Dickerson, R.F., Kendrew, J.C. and Standberg, B.E. (1961). Appendix by Hart R.G. **Acta Cryst. 14**, 1188-1195.

6. Program PHARE (originally under a different name) was written by M. Rossmann in 1967. It was later modified by A. Wonacott, M. Adams, R. Schevitz, J. E. Ladner and G. Bricogne among others.

7. Raiz, V.Sh. and Andreeva, N.S. (1970). **Sov. Phys. Crystallogr. 15**, 206-210. Translated from **Kristallografiya (1970). 15**, 246-251

8. Einstein, J.R. (1976). **Acta Cryst. A33**, 75-85.

9. Terwilliger, T.C. and Eisenberg, D. (1987). **Acta Cryst. A43**, 6-13.

10. Dodson, E.J. in "Crystallographic Computing Techniques", (1975) pp 259-268. Ahmed, F.R. (editor), Munksgaard, Copenhagen.

11. Terwilliger, T.C. and Eisenberg, D. (1983). **Acta Cryst. A39**, 813-817.

# Refinement of single isomorphous replacement heavy-atom parameters in Patterson vs reciprocal space.

by Ian J Tickle, Department of Crystallography
Birkbeck College, University of London

## Introduction

This paper presents a novel method for refinement of heavy-atom parameters in single isomorphous derivatives. The refinement is done in vector space against the experimental heavy-atom Patterson. It turns out that the new method possesses a number of advantages over the traditional reciprocal space method, for example, better discrimination against false minima, and greater radius of convergence.

The experimental heavy-atom Patterson may be either the isomorphous difference ($\Delta_{ISO}$), or where reliable anomalous data are available, the combined difference ($F_{HLE}$) Patterson. In practice it is found that the $\Delta_{ISO}$ Patterson actually gives better results, probably because the high relative errors in the anomalous differences outweigh any theoretical advantage. In reciprocal space, except in the case of centric reflections, the isomorphous differences are not unbiased estimates of the heavy-atom structure amplitudes, so that if no anomalous data are available, reciprocal space refinement is theoretically justified only when the amplitude data is limited to the centric zones. This is obviously a cause of difficulty in space groups without centric zones, such as R3. Even in those cases where reliable anomalous data are available, it is frequently found that the centric zone reflections are essential for a correct refinement solution; unfortunately some data collection strategies lead to partial or even complete omission of the centric zone data.

It is therefore evident that a method which is not dependent on the presence of either anomalous data or centric zones has advantages over the traditional method. However it is necessary to justify the use of the $\Delta_{ISO}$ Patterson. Phillips (1966) analysed the $\Delta_{ISO}$ Patterson:

$$\Delta_{ISO}^2 = (F_{PH}-F_P)^2$$
$$= 4F_P^2 \sin^4\left(\frac{\alpha_P-\alpha_{PH}}{2}\right) + F_H^2 \cos^2(\alpha_{PH}-\alpha_H) - 4F_PF_H \sin^2\left(\frac{\alpha_P-\alpha_{PH}}{2}\right)\cos(\alpha_{PH}-\alpha_H)$$

Here, the $F_P^2$ term will produce the native Patterson, but at considerably reduced weight because the phase angle difference $\alpha_P$-$\alpha_{PH}$ will be small provided the average $F_H$ is small compared with the average $F_P$ and $F_{PH}$. The $F^2_H$ term will give the heavy-atom Patterson weighted by the square of the cosine which will average to approximately 0.5 because $\alpha_{PH}$ and $\alpha_H$ are nearly uncorrelated. The $F_PF_H$ term will produce a Patterson of protein-heavy atom cross-vectors which will again be weighted down by the sine term, but also will have its sign changed at random by the cosine term, and will therefore behave as random noise on the systematic heavy-atom Patterson (Dodson & Vijayan, 1971).

Thus although individual $\Delta^2_{ISO}$ terms are certainly not half of $F^2_H$, the effect of the Fourier transform is to produce a Patterson with peaks in the same place as the

heavy-atom Patterson but at half the height, with some random noise superposed. Provided the minimisation of the sum of squares of differences between the experimental and calculated Pattersons is done only for points near the vector peaks, the accumulation of noise will not be too great. This is precisely why the reciprocal space refinement method has such poor discrimination and convergence properties; each amplitude term represents mostly noise and very little signal, because the heavy-atom Patterson constitutes relatively few vectors in mostly empty space.

**Theory of the vector space refinement method**

The crucial feature of the method is therefore to consider only those grid points in the Patterson density which lie within the calculated peaks. For two atoms with electron density functions $\rho_j(\mathbf{r})$ and $\rho_k(\mathbf{r})$, the calculated Patterson peak profile for the cross-vector is the convolution of the electron density functions:

$$P_{jk}(\boldsymbol{u}) = \int_{\infty} \rho_j(\boldsymbol{r}).\rho_k(\boldsymbol{u}-\boldsymbol{r}).d\boldsymbol{r}$$

where **u** is a point in Patterson space.

(The integration limits are infinite because we are considering only two isolated atoms, not a crystal). The electron density functions are related to the occupancy n, scattering factor f(s) and thermal parameter B (assumed isotropic) by the Fourier transform:

$$\rho(\boldsymbol{r}) = n\int_{\infty} f(s).\exp(-Bs^2).\exp(i2\pi \boldsymbol{H}.\boldsymbol{r}).d\boldsymbol{H}$$

where **H** is the scattering vector and $s = \sin\theta/\lambda = |\mathbf{H}|/2$. Therefore the Patterson density is the transform of the product of the transforms:

$$P_{jk}(\boldsymbol{u}) = n_j n_k \int_{s<s_{max}} f_j(s).f_k(s).\exp(-(B_j+B_k)s^2).\exp(i2\pi \boldsymbol{H}.\boldsymbol{u}).d\boldsymbol{H}$$

In practice the observed Patterson can only be computed with data of finite resolution ($s < s_{max}$) so the same is done for the calculated Patterson. Since the peak profile is assumed to be isotropic, it is convenient to write it in terms of spherical polar coordinates and integrate out the angles, leaving the radial coordinate, s:

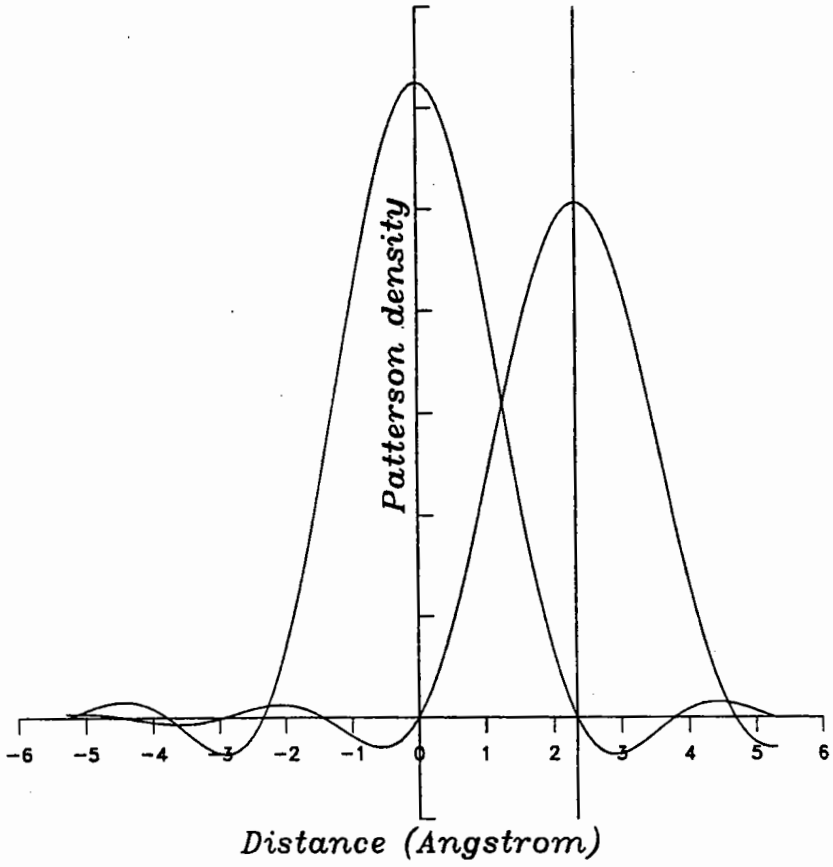
$$P_{jk}(u) = \frac{8n_j n_k}{u}\int_0^{s_{max}} f_j(s).f_k(s).\exp(-(B_j+B_k)s^2).\sin(4\pi us).s.ds$$

Here, u is the distance from the calculated peak position. The case u=0, the peak maximum, has to be treated specially using:

$$\left.\frac{\sin(4\pi us)}{u}\right|_{\lim u\to 0} = 4\pi s$$

The scattering factors are conveniently approximated as sums of Gaussians (Lee and Pakes, 1969). Computations of the integral and its derivatives with respect to the occupancies, thermal parameters and atomic coordinates, which are required for the least-squares minimisation, are readily performed with adequate precision by 8-point Gaussian quadrature.

Typical Patterson peak profiles (for Hg-Hg vectors, $d_{min}$ = 3Å, B = 10Å$^2$, occupancies = 1.0 and 0.8) separated by a distance equal to the first zero of the function are shown in the Figure below.




The distance of separation shown is approximately $d_{min}/\sqrt{2}$ and is roughly the expected radius of convergence of the refinement. This should be compared with the corresponding expected radius of convergence for reciprocal space refinement of $d_{min}/4$. This improvement is borne out by trial calculations.

It is convenient to define the peak radius as the radius of the first zero, so that all the calculated density values used are positive. In the first version of the refinement program, the zero was found with the Newton-Raphson formula:

$$u_{n+1} = u_n - \frac{P}{dP/du} \qquad u_0 = d_{min}/\sqrt{2}$$

Unfortunately in some circumstances (low resolution and large B values) the peak profile can have a minimum before the first zero, and the simple Newton-Raphson method fails catastrophically. The solution in the second and current version was to locate the first minimum in the square of the peak profile; this will locate either the zero or the minimum in the profile, whichever comes first. Using steepest-descent minimisation:

$$u_{n+1} = u_n - \frac{P.dP/du}{(dP/du)^2 + P.d^2P/du^2}$$

In the least squares refinement the function minimised is:

$$\sum_{u \in P} w(u).(P_o(u) + F_{000}^2/V_c - \sum_j P_j(u))^2$$

where the outer summation is over all points within the calculated Patterson peaks (omitting the origin peak), and the weight w is a function only of the coordinate vector **u**, the resolution $d_{min}$, and the point group symmetry:

$$w(u) = \left(\sum_k 3(\sin(g_k(u)) - g_k(u).\cos(g_k(u)))/g_k(u)^3\right)^{-1}$$
$$g_k(u) = |u - S_k.u| / d_{min}$$

$S_k$ is the k'th point group symmetry operator (space group with centre of symmetry, minus translations). This arises because the variance of an electron density or Patterson function is higher than average on or near symmetry elements, the distance dependence being a function of the resolution. In reciprocal space refinement there are usually several weighting parameters to be specified, often somewhat arbitrarily, by the user; here the maximum resolution is the only user-specified parameter, and this is fairly well-defined.

In the expression to be minimised, account has to be taken of the fact that the observed Patterson $P_o(\mathbf{u})$ does not include the $F_{000}$ term ( = $\Sigma_j n_j f_j(0)$ ). This is treated as part of the calculated Patterson density, since it depends on the occupancies. The $\Sigma_j P_j(\mathbf{u})$ term allows for the possibility of overlapping peaks in the calculated Patterson.

Scaling

Two further issues need to be considered: the derivative to native scaling, and scaling of the isomorphous difference Patterson. In reciprocal space refinement the $F_{PH}/F_P$ scale factor can be refined along with the atomic parameters; in vector space it is in theory possible to do the same, but rather cumbersome (it would be necessary to refine simultaneously against the $F_P^2$, $F_{PH}^2$ and $F_P.F_{PH}$ Pattersons). Effort was therefore invested into devising an improved derivative scaling program; existing programs only produce an approximation which is then refined. Kraut et al. (1962) published a formula for $F_{PH}/F_P$ scaling which does not rely on anomalous data, and which surprisingly has been little used. The author has improved on this, and so it is probably worth describing in detail.

Kraut's formula is based on equating the Patterson origins:

$$k_s^2 \sum F_{PH}^2 = \sum F_P^2 + \sum F_H^2$$

where $k_s$ is the scale factor to be determined. In practice independent scale factors are assigned to equi-volume shells in reciprocal space to allow for variations due to thermal parameter differences etc, and the scale applied to individual $F_{PH}$'s determined by smoothing and interpolating the shell scale factors. For acentric reflections the mean square $F_H$ is expected to be twice the mean square isomorphous difference, to a good approximation:

$$\sum_a F_H^2 = 2\sum_a (k_s F_{PH} - F_P)^2$$

Elimination of the unknown $\Sigma F_H^2$ gives a quadratic in $k_s$ which is solved to give Kraut's formula:

$$k_s = \frac{2\sum F_P F_{PH} - ( 4(\sum F_P F_{PH})^2 - 3\sum F_P^2 \sum F_{PH}^2 )^{1/2}}{\sum F_{PH}^2}$$

However Kraut et al. did not indicate that this applies only to acentric reflections, and that centric reflections must be excluded. For high symmetry space-groups at low resolution there will be a large proportion of centric reflections, and since they actually give a more accurate estimate of the mean square $F_H^2$:

$$\sum_c F_H^2 = \sum_c (k_s F_{PH} - F_P)^2$$

it would seem absurd to exclude them. Another factor ought to be taken into account: the sum of squares of isomorphous differences is biased positively due to random errors of measurement (Dodson, Evans and French, 1974). Therefore:

$$\sum F_H^2 = \sum m\left((k_s F_{PH} - F_P)^2 - (k_s^2\sigma_{PH}^2 + \sigma_P^2)\right)$$

where m=1 for centric, 2 for acentric. Using this to solve for $k_s$:

$$k_s = \frac{\sum ((m+1)F_P^2 - m\sigma_P^2)}{\sum mF_P F_{PH} + \left((\sum mF_P F_{PH})^2 - \sum ((m+1)F_P^2 - m\sigma_P^2).\sum ((m-1)F_{PH}^2 - m\sigma_{PH}^2)\right)^{1/2}}$$

Care must be taken that reliable estimates for $\sigma_P$ and $\sigma_{PH}$ are used; unfortunately some data processing programs do not calculate $\sigma(F)$ correctly when calculating $F=I^{1/2}$ for small I. The correct formula is:

$$\sigma(F) = (F^2 + \sigma(F^2))^{1/2} - F$$

For large F, this gives the commonly used formula $\sigma(F)=\sigma(F^2)/2F$ whereas in the limit of F=0, it becomes $\sigma(F)=(\sigma(F^2))^{1/2}$.

Lastly, factors which affect the scaling of the Patterson have to be considered: a correction for incomplete data should be applied, so that the Patterson values are divided by the completeness fraction. The centricity factor, m, which multiplies $\Sigma\Delta_{ISO}^2$ to get $\Sigma F_H^2$ is dealt with empirically by multiplying the Patterson values by:

$$\overline{m} = \frac{\sum m(\Delta^2_{ISO} - \sigma^2_P - \sigma^2_{PH})}{\sum (\Delta^2_{ISO} - \sigma^2_P - \sigma^2_{PH})}$$

This produces a weighted mean value of m according to the relative number and magnitude of centric and acentric reflections. The observed Patterson is computed with the FFT program using $(\Delta^2_{ISO} - \sigma^2_P - \sigma^2_{PH})$ as the Fourier term.

**Results**

Various tests, with both real and simulated data were performed, comparing the new scaling and refinement programs (FHSCAL and VECREF) with existing programs (RFACTOR, PHARE, REFINE, HEAVY) in the CCP4 suite, and the results are shown in Tables 1, 2 and 3.

---

**TABLE 1 - Xylose isomerase 2 site Pb derivative at 3.5Å**

1. Refinement of $F_{PH}/F_P$ overall scale and B

| Program | Scale | B |
|---|---|---|
| FHSCAL | 0.985 ± 0.008 | 9.1 ± 0.6 |
| RFACTOR | 0.968 | 8.8 |
| REFINE ($F_{HLE}$) | 0.943 ± 0.003 | 8.7 ± 0.2 |
| REFINE (centric) | 0.944 ± 0.006 | 8.2 ± 0.6 |
| PHARE (4 derivs) | 0.960 ± 0.003 | 8.0 ± 0.2 |
| HEAVY* | 0.977 ± 0.003 | 9.0 ± 0.2 |

2. Refinement of occupancies (with atomic B's fixed)

| Program | Occ(1) | Occ(2) | R(%) |
|---|---|---|---|
| VECREF | 1.94 ± 0.07 | 1.96 ± 0.07 | 39 |
| REFINE ($F_{HLE}$) | -0.50 ± 0.04 | 2.36 ± 0.03 | 72 |
| REFINE (centric) | 2.19 ± 0.06 | 1.87 ± 0.06 | 57 |
| PHARE (4 derivs) | 2.21 ± 0.02 | 2.03 ± 0.02 | 64 |
| HEAVY* | 1.88 ± 0.06 | 1.90 ± 0.06 | (not calculated) |

*Version of HEAVY modified to use LCF format, and also corrected for an error which grossly underestimated standard deviations (Cruickshank, 1965).

Table 1 shows the results for a 2 site Pb derivative of xylose isomerase at 3.5Å (space group = $P3_121$, completeness = 0.93). The $F_{PH}/F_P$ scale factor varies significantly (about 4%) between the various methods, but the overall relative B-factor is surprisingly consistent. In the refinement of the occupancies, the reciprocal space refinement with the full 3-D data by the $F_{HLE}$ method produced a negative occupancy for the major site given by PHARE (by phase refinement with 3 other derivatives), and the coordinates (not shown) for this site diverged. Otherwise there was good agreement for the coordinates; it is generally observed that it is much easier to obtain reliable coordinates than reliable occupancies or B-factors. In fact in this and the other tests, the B-factors were kept fixed at the same values, otherwise their high correlation with the occupancies would make accurate comparisons impossible. One interesting point about this derivative, and the author was not aware of this at the time, is that xylose isomerase is a dimer and the Pb sites are fully occupied, and so the occupancies would be expected to be the same. In fact the occupancies obtained by VECREF are almost too good for the standard deviations quoted, so this result may be fortuitous.

Table 2 shows the results for scaling and refinement of a PCMBS (Hg) 2 site derivative (space group = $P2_1$, completeness = 0.90) of MetJ apo-repressor at 3.0Å. The data had already been scaled by the RFACTOR program, and there is considerable disagreement (7%) between the programs over the scale factor. Similarly the refinement programs do not agree on the occupancy of the second site. The vector-space R-factor obtained for this derivative was very low (18%). Of course the 'R-factors' obtained by the different methods cannot be compared directly since they are defined in different ways. For the vector space method:

$$R = \left( \frac{\sum_{u \in P} w(u).(P_o(u) - P_c(u))^2}{\sum_{u \in P} w(u).P_o^2(u)} \right)^{1/2}$$

where the summation is over all points within calculated peaks, excluding the origin peak, and $P_c(u)$ is the calculated Patterson allowing for overlapping peaks and the $F_{000}$ term.

As a further test of the discriminating power of the refinement methods 7 extra wrong sites were added and all sites refined. These extra sites share vectors in common with the correct sites. In the vector space refinement all but two occupancies of the wrong sites refined to zero, and these two were very small; in the reciprocal space refinements the wrong occupancies refined to rather larger values both positive and negative. In addition the vector space R-factors are more discriminating: the initial occupancies gave R=211% which reduced to 35% on refinement and then 18% on removal of the two wrong low occupancy sites; for the reciprocal space refinement (REFINE) the initial R-factor was 78% reducing to 45% on refinement and then to 41% on removal of all the wrong sites. This indicates that the vector space method discriminates against incorrect low occupancy sites, whereas the reciprocal space method tends to encourage the addition of dubious minor sites which are only partially consistent with the Patterson.

**TABLE 2 - MetJ apo-repressor 2 site PCMBS (Hg) derivative at 3.0Å**

1. Refinement of $F_{PH}/F_P$ overall scale and B

| Program | Scale | B |
|---|---|---|
| FHSCAL | 1.07 ± 0.01 | 0.4 ± 0.8 |
| RFACTOR | 1.00 | 0.0 |
| REFINE ($F_{HLE}$) | 1.029 ± 0.003 | 0 (not refined) |
| REFINE ($F_{HLE}$) | 1.07 ± 0.01 | 3.4 ± 0.4 |
| HEAVY | 1.07 ± 0.01 | 0.3 ± 0.7 |

2. Refinement of occupancies (with atomic B's fixed)

| Program | Occ(1) | Occ(2) | R(%) |
|---|---|---|---|
| VECREF | 1.72 ± 0.05 | 1.57 ± 0.07 | 18 |
| REFINE ($F_{HLE}$) | 1.73 ± 0.02 | 1.29 ± 0.02 | 41 |
| HEAVY | 1.62 ± 0.04 | 1.47 ± 0.05 | (not calculated) |

3. Extra wrong sites test

| Atom | Occ | X | Y | Z | B | Occ(VECREF) | Occ(REFINE) | Occ (HEAVY) |
|---|---|---|---|---|---|---|---|---|
| Hg 1 | 0.7 | .12 | 0 | .47 | 7 | 1.72 | 1.83 | 1.61 |
| Hg 2 | 0.7 | .42 | .048 | .80 | 25 | 1.58 | 1.23 | 1.48 |
| Hg 3 | 0.7 | .42 | .048 | .30 | 25 | 0.00 | 0.32 | 0.15 |
| Hg 4 | 0.7 | .42 | .548 | .80 | 25 | 0.08 | -0.29 | 0.09 |
| Hg 5 | 0.7 | .42 | .548 | .30 | 25 | 0.09 | 0.37 | 0.15 |
| Hg 6 | 0.7 | .92 | .048 | .80 | 25 | 0.00 | 0.34 | 0.11 |
| Hg 7 | 0.7 | .92 | .048 | .30 | 25 | 0.00 | 0.21 | 0.01 |
| Hg 8 | 0.7 | .92 | .548 | .80 | 25 | 0.00 | 0.11 | 0.24 |
| Hg 9 | 0.7 | .92 | .548 | .30 | 25 | 0.00 | -0.47 | 0.12 |

---

With real data it is obviously not possible to say what are the true values of the refined parameters, and so a test was set up using simulated data. This used the native MetJ amplitudes, random protein phases, the observed lack-of-closure errors, and heavy atom structure factors calculated from the parameters obtained by reciprocal space refinement. Thus simulated derivative amplitudes were used with real derivative standard deviations and native amplitudes and standard deviations, to calculate the derivative scale factor and refine the atomic parameters. The result was again considerable disagreement (8%) in the scale factor between Kraut scaling and reciprocal space refinement, but now it can be seen that the error is only 1.3% for Kraut scaling, but 6.8% for reciprocal space refinement (REFINE). In the occupancy refinement, the maximum error was only 2% for vector space refinement, but 8% for reciprocal space refinement. As a final test of the power of discrimination of the methods, the parameters were refined in the wrong space group (P2). In the vector space refinement the R-factor increased from 19% to 74% and the standard deviations increased 4-fold. In contrast, in the reciprocal space refinement (REFINE) the occupancies were not significantly different, the standard deviations were unchanged, and the R-factor only increased from 45% to 53%.

**TABLE 3 - Simulated MetJ-Hg data**

Overall scale = 1, overall B = 0, random protein phase, Gaussian random errors on $F_P$, $F_{PH}$, $F_H$, using observed standard deviations and RMS lack-of-closure. Observed atomic coordinates and B's. Occupancies = 1.7, 1.3.

1. Refinement of $F_{PH}/F_P$ overall scale and B

| Program | Scale | B |
|---|---|---|
| FHSCAL | 1.013 ± 0.006 | 0.6 ± 0.3 |
| RFACTOR | 0.946 | 1.0 |
| REFINE ($F_{HLE}$) | 0.932 ± 0.006 | 0.4 ± 0.4 |
| HEAVY | 0.992 ± 0.010 | 0.2 ± 0.7 |

2. Refinement of occupancies

| Program | Occ(1) | Occ(2) | R(%) |
|---|---|---|---|
| VECREF | 1.73 ± 0.05 | 1.31 ± 0.06 | 19 |
| REFINE ($F_{HLE}$) | 1.57 ± 0.02 | 1.19 ± 0.03 | 45 |
| HEAVY | 1.60 ± 0.05 | 1.21 ± 0.06 | (not calculated) |

3. Refinement of occupancies in wrong space group (P2)

| Program | Occ(1) | Occ(2) | R(%) |
|---|---|---|---|
| VECREF | 0.90 ± 0.19 | 0.99 ± 0.23 | 74 |
| REFINE ($F_{HLE}$) | 1.51 ± 0.02 | 1.20 ± 0.03 | 53 |
| HEAVY | 0.76 ± 0.10 | 0.93 ± 0.12 | (not calculated) |

---

**Acknowledgements**

I would like to thank Drs K Henrick and S E V Phillips for allowing me to use their data for test purposes.

**References**

Cruickshank, D.W.J. in 'Computing Methods in Crystallography', Ed. J.S. Rollett. Oxford: Pergamon (1965) Chapter 14.

Dodson, E.J., Evans, P.R. and French, S. in 'Anomalous Scattering', Ed. S. Ramaseshan and S.C. Abrahams. Copenhagen: Munksgaard (1974), Chapter 7.

Dodson, E.J. and Vijayan, M. Acta Cryst. B27 (1971) 2402-2411.

Kraut, J., Sieker, L.C., High, D.F. and Freer, S.T. Proc. Natl. Acad. Sci. USA 48 (1962) 1417-1424.

Lee, J.D.and Pakes H.W. Acta Cryst. A25 (1969) 712-713.

Phillips, D.C. in 'Advances in Structure Research by Diffraction Methods', Ed. R. Brill and R. Mason. New York and London : Interscience (1966).

# MULTIWAVELENGTH ANOMALOUS DIFFRACTION ANALYSIS OF A LARGE PROTEIN

Janet L. Smith, Eugene J. Zaluzec, Jean-Pierre Wery *
Dept. of Biological Sciences, Purdue University
West Lafayette, Indiana 47907 USA

and Yoshinori Satow
Faculty of Pharmaceutical Sciences, University of Tokyo
Bunkyo-ku, Tokyo, Japan

## INTRODUCTION

Multiwavelength anomalous diffraction (MAD) is a newly developed method for solving the crystallographic phase problem. It has recently been highly successful in the determination of several macromolecular structures. Although the potential for phase determination from diffraction measurements at multiple wavelengths has been understood for many years, application to biological macromolecules awaited the availability of tunable synchrotron radiation as well as an understanding of absorption edges of anomalous scatterers in biological molecules. The reader is referred to recent reviews (1,2) for summaries of the method, and to a publication on the lamprey hemoglobin test problem (3) for a more complete outline of its application. Here we describe a MAD analysis in progress for a larger macromolecule.

Total atomic scattering is the sum of normal scattering and the real and imaginary components of anomalous scattering ($f = f^\circ + f' + if''$). Differential anomalous scattering by a subset of atoms in the crystal gives rise to differences in total scattering intensity at different incident wavelengths. In favorable cases, these differences can be used to extract phase information directly from measured intensities. The MAD signal is optimized by precise tuning of the incident X-rays to wavelengths producing extreme values of f' and f'', which occur at absorption edges. The MAD formalism is exact, so direct phase determination to the diffraction limit of the crystals is possible in principle. But, as in all experimental science, principle differs somewhat from practice.

The MAD phase equation, as proposed by Karle (4) and modified by Hendrickson (5), is given below.

$$|^\lambda F_{obs}|^2 = |^\circ F_T|^2 + a(\lambda)\,|^\circ F_A|^2 + b(\lambda)\,|^\circ F_T|\,|^\circ F_A| \cos(^\circ\phi_T - ^\circ\phi_A) + c(\lambda)\,|^\circ F_T|\,|^\circ F_A| \sin(^\circ\phi_T - ^\circ\phi_A)$$

where $a(\lambda) = (f'^2_\lambda + f''^2_\lambda)/f^{\circ 2}$, $b(\lambda) = 2(f'_\lambda/f^\circ)$ and $c(\lambda) = \pm 2(f''_\lambda/f^\circ)$ (sign dependent on Bijvoet mate). This formulation is especially powerful because all wavelength dependence [$a(\lambda)$, $b(\lambda)$, $c(\lambda)$] is separated from structure dependence ($|^\circ F_T|$, $|^\circ F_A|$, $^\circ\phi_T$-$^\circ\phi_A$). Structure factors pertinent to MAD phasing are illustrated in Fig. 1.

---

* current address: Eli Lilly & Co., Indianapolis, IN, USA

The steps in a structure determination by MAD (3) are 1) measurement of the absorption spectrum from the anomalous scatterer *in situ* and calculation of anomalous scattering factors from absorption data, 2) measurement of $|^{\lambda}F_{obs}|$ at multiple wavelengths selected to maximize differences in anomalous scattering, 3) derivation of the quantities $|^{\circ}F_T|$, $|^{\circ}F_A|$ and $^{\circ}\phi_T$-$^{\circ}\phi_A$ by least squares fit of the multiple measurements to the phase equation given above, 4) conventional determination of the partial structure of the anomalous scatterers from the magnitudes $|^{\circ}F_A|$, 5) computation of the desired phase $^{\circ}\phi_T$ from the calculated $^{\circ}\phi_A$ of the partial structure and the phase difference $^{\circ}\phi_T$-$^{\circ}\phi_A$, and 6) Fourier synthesis of an image of the structure from $|^{\circ}F_T|$,$^{\circ}\phi_T$.

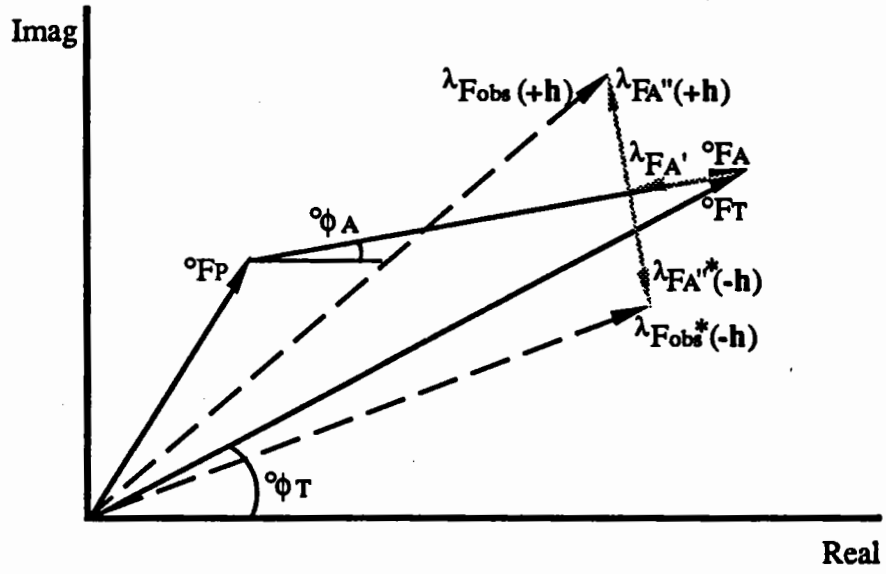



**Fig. 1. Argand diagram of structure amplitudes and phases pertinent to MAD phase determination for a hypothetical reflection at a single wavelength.** Structure factors preceded by ° pertain to normal scattering only: $^{\circ}F_P$ for the partial structure that does not scatter anomalously, $^{\circ}F_A$ for the partial structure that does scatter anomalously and $^{\circ}F_T$ for the total structure. $^{\lambda}F'_A$ and $^{\lambda}F''_A$ are, respectively, the real and imaginary components of the total structure factor due to a single type of anomalous scatterer. Their magnitudes can be expressed as $(f'_{\lambda}/f^{\circ})|^{\circ}F_A|$ for $|^{\lambda}F'_A|$ and $(f''_{\lambda}/f^{\circ})|^{\circ}F_A|$ for $|^{\lambda}F''_A|$. The illustration magnifies the anomalous contribution for clarity. A typical low-angle reflection for the ATase example of this paper would have $|^{\circ}F_P|$:$|^{\circ}F_A|$:$|^{\lambda}F'_A|$:$|^{\lambda}F''_A|$ of 100:26:7:4.

## MAD AND GLUTAMINE PRPP AMIDOTRANSFERASE

The successful applications of MAD to date have been crystal structures with less than 30,000 daltons of macromolecule in the crystallographic asymmetric unit. This paper describes a MAD analysis for a larger problem, having 200,000 daltons in the asymmetric unit. Glutamine phosphoribosylpyrophosphate amidotransferase (ATase) catalyzes the first and committed step of purine biosynthesis. It is a tetrameric molecule of identical 50,000-dalton subunits (6). The ATase used in this experiment is isolated from *B. subtilis* and has one $Fe_4S_4$ cluster of uncertain function in each subunit. The $Fe_4S_4$ cluster imparts to the enzyme an oxygen lability, which appears related to cluster function. Spectroscopic studies indicate that the ATase cluster is of the 4-Fe ferredoxin type (7). The Fe atoms in the cluster are the anomalous scatterers for this MAD analysis.

The magnitude of the MAD signal for the ATase experiment was estimated based on the concentration of Fe in the ATase subunit. The signal is customarily calculated at zero scattering angle where $f^\circ$ = # electrons and no assumptions about relative B parameters are required (8). Since the Fe anomalous scatterers in ATase are expected to be arranged in tetrahedral clusters with Fe-Fe distances of ~2.8Å, a strong enhancement of anomalous scattering is expected for data at scattering angles too low to resolve individual Fe's. The maximum signal enhancement is $\sqrt{N}$ where N is the number of anomalous scatterers in the cluster. Diffraction ratios for the estimated MAD signal are given in Table 1. The cluster effect is shown graphically in Fig. 2. The anomalous scattering due to S in the cluster, both inorganic sulfide and cysteine, does not increase the diffraction ratios more than 0.2% and was not considered further in our calculations.

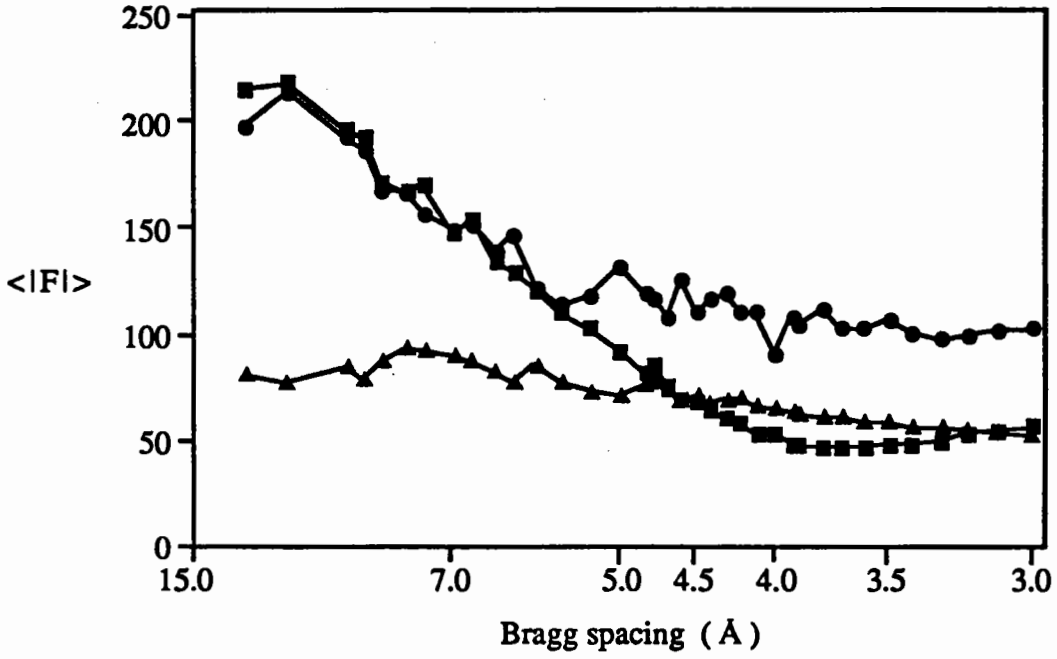



**Fig. 2. Mean structure amplitude as a function of d-spacing for 16 Fe atoms in the ATase asymmetric unit.** The three curves are: ● experimental $|^\circ F_A|$ from MAD phasing; ■ $|F_{calc}|$ from a model of four tetrahedral Fe clusters (Fe-Fe = 2.8Å, B = 15Å$^2$); ▲ $|F_{calc}|$ from a model of 16 Fe atoms (Fe-Fe > 10Å, B = 15Å$^2$). The enhancement of scattering at low-angles due to the cluster is clear, as is a slight diminution of the cluster transform between ~4.5Å and ~3.2Å. While the "observed" $|^\circ F_A|$'s can be fit well to theory within ~5.5Å spacings, noise dominates at higher angles.

We expected to have little trouble detecting the MAD signal from data with Bragg spacings out to ~5.5Å, where the maximum signal is 5.0-5.8% of |F|. Beyond this point, the signal is halved and detection becomes more difficult. Crystal quality, counting statistics, detector accuracy and correction for systematic errors become critical. However, MAD signals at the 2-3% level have been measured accurately for other problems, demonstrating that detection is possible in careful experiments.

## THE MAD EXPERIMENT

ATase was crystallized by batch methods in an inert-atmosphere chamber due to its oxygen lability. Two related crystal forms grew under identical conditions. The predominant form is of space group $P2_1$ with a = 158.8Å, b = 75.7Å, c = 94.1Å and $\beta$ = 91.4°. The minor form is of space group $P2_12_12_1$ with a = 160.0Å, b = 74.1Å and c = 185.0Å. The two crystal forms can be distinguished only by their diffraction patterns. Both forms have one tetrameric molecule in the crystallographic asymmetric unit. Crystals for the MAD experiment were inert-mounted and sealed in capillaries before travel to the synchrotron source.

**TABLE 1**
Diffraction Ratios for the ATase MAD Experiment

| | Observed ($d_{min}$ = 5Å) | | | Estimated (Zero scattering angle) | | | Scattering Factors ($e^-$) | |
|---|---|---|---|---|---|---|---|---|
| Wave-length | 1.7425Å | 1.7390Å | 1.5000Å | 1.7425Å | 1.7390Å | 1.5000Å | f' | f" |
| 1.7425Å | 0.060 (0.053) | 0.028 | 0.053 | 0.033 (0.000) | 0.011 | 0.050 | -7.864 | 2.304 |
| 1.7390Å | | 0.071 (0.062) | 0.046 | | 0.058 (0.000) | 0.038 | -6.257 | 4.053 |
| 1.5000Å | | | 0.053 (0.037) | | | 0.044 (0.000) | -0.897 | 3.065 |

Diffraction ratios are as described in refs. 1 and 8. Values on the diagonal are for Bijvoet differences (centrics in parentheses), off the diagonal are for dispersive differences.

The MAD experiment was designed to measure data to Bragg spacings of 3.0Å from three wavelengths near the Fe K-absorption edge. The 3Å limit was imposed by beamtime constraints and by detector and unit-cell size (Fig. 3); ATase crystals diffract synchrotron radiation to near 2Å spacings. Data were measured by the oscillation method on beamline 14A (9) at the Photon Factory (storage ring = 2.5 GeV, 270-150 mA, injection every 12 hours). Crystals were positioned by a vertically oriented four-circle diffractometer with an imaging-plate cassette on the detector arm, which was set at 0° 2$\theta$. The $\omega$ axis was used for data collection. Fuji imaging plates (type HR3) were read out on a home-built scanner (10). Prior to data collection, a fluorescence scan through the Fe K-absorption edge was measured from a crystal to choose optimal wavelengths for data collection and to provide data for calculation of anomalous scattering factors. The inflection point of the absorption edge (1.7425Å), the first maximum above the edge (1.7390Å) and a remote high-energy point (1.5000Å) were chosen for data collection. Scattering factors were determined as described for the lamprey hemoglobin test problem (3) and are shown in Table 1.

Data were measured so that the individual observations contributing to the phase of each reflection were recorded from the same crystal, at nearly the same time, in nearly the same absorbing geometry. It proved cumbersome and time-consuming to align crystals with the unique monoclinic axis parallel to the oscillation axis, so most data were collected from unaligned crystals. After measurement of two or three successive oscillation images, their Friedel images were collected by inverse geometry ($-\chi, \phi+\pi$). This procedure was then

repeated at the other wavelengths for the same oscillation range. Each wavelength change required optimization of beamline optics and thus occupied 2-3 min. Exposure times were dose-dependent and ranged from 35 sec to 4 min with oscillation angles of 1.15° to 2.65°. Imaging plates were scanned within 4 hours following exposure. Crystal decay was monitored by occasional display of scanned images, although this slowed the scanning process significantly. A minimum of 7 min was required to scan and erase each imaging plate without display of the scanned image. The speed of data collection was limited more by the scanning procedure than by any other factor.

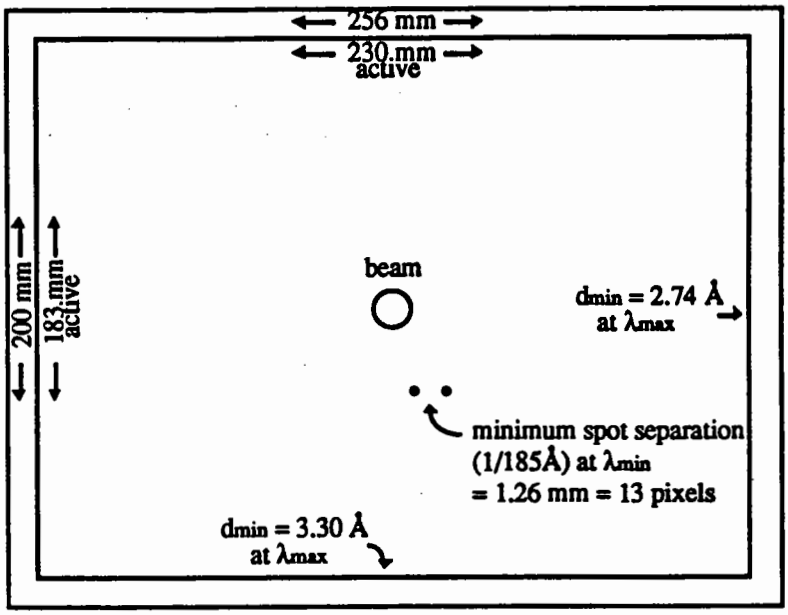



**Fig. 3. Parameters in design of the MAD experiment.** Experimental design was based on a crystal-to-detector distance of 155 mm, $\lambda_{min}$ of 1.50Å and $\lambda_{max}$ of 1.74Å. The oscillation axis was vertical. Scanned images had 2560 x 2000, 100-μ, 16-bit pixels. The logarithmic output from the scanner had 12 bits of precision with a dynamic range of 1 to 9978 when converted to counts.

A total of 560 oscillation images were recorded from 19 crystals. Crystal quality and lifetime in the X-ray beam were highly variable. The final data sets derive from 461 images from eight monoclinic crystals and one orthorhombic crystal.

## SCALING AND MERGING OF MAD DATA

Scanned images were reduced to integrated intensities and scaled with the DENZO programs run on a microVAX II workstation (Z. Otwinowski, personal communication). Since neither fiducials nor direct beam were marked on the images and the reflections were closely spaced, some interactive processing and graphical display were required for each image to ensure that the initial beam position was accurate enough for automatic parameter refinement.

Because the MAD signal from ATase crystals is weak (Table 1), only fully recorded reflections were used for phasing. Reflection widths were estimated to be <0.2° based on previous area-detector experiments, and an overlap of 0.15° between adjacent images was taken to ensure that all reflections would be fully recorded. Mosaicity values for each image were estimated during data processing from a histogram of mean intensity vs. predicted reflection position within the oscillation range. Unfortunately, the crystals used for the MAD experiment

had mosaicities estimated between 0.25° and 0.85° (average = 0.48°). The high mosaicity significantly reduced the number of fully recorded reflections and, consequently, the number of reflections for which there were enough multiple measurements for MAD phase determination. The loss of multiples arises from the fact that Ewald's sphere is larger for the edge-remote wavelength (1.5000Å) than for the near-edge wavelengths (1.7425Å and 1.7390Å). Many reflections that were fully recorded at one energy were partially recorded on adjacent oscillation images at the other energy. Although the merged monoclinic data set contains intensities for 90% of the unique data to 3.0Å spacings, only 52% of all unique data have enough measurements for MAD phasing. The problem of high mosaicity has since been solved by a change in crystal stabilizing solution. MAD analysis has not been carried out for the orthorhombic data, which represent only about 40% of the unique data to 3.0Å spacings. $R_{sym}$ for these data was 4.8% to 6.1%, dependent on wavelength. Further discussion pertains only to the monoclinic data.

Integrated intensities for the crystal with the broadest sampling of reciprocal space were merged into a unique set for use as a scaling standard. Separate unmerged data sets were generated for each orientation of each crystal in order not to intermingle multiwavelength measurements for any single reflection that had been taken from different crystals or asymmetric units. For six crystals, data were recorded from a single orientation only, but data were measured in three orientations from one crystal and in two orientations from another. Integrated intensities for each wavelength of each orientation of each crystal (a total of 33 sets) were scaled separately to the scaling standard using the DENZO scaling program (linear scale and B factors for each oscillation image). Scaling statistics are summarized in Table 2. The multiple measurements taken from the same crystal at nearly the same time and orientation were collected together for phase determination (reducing the number of data sets from 33 to 11). $|^{\lambda}F_{obs}|$ from the three wavelengths were scaled together within each data set independent of the others and without merging redundancies. The resulting observed anomalous and dispersive MAD signal is compared to expected values in Table 1. At this point, inter-crystal scale factors were computed by comparing all data from all crystals. Using the DENZO scaling program, $\langle|^{\lambda}F_{obs}|\rangle$ values were scaled together by refining a single scale factor and B factor for each crystal relative to the others. Scale factors ranged from 0.996 to 1.094 and B factors from 0.00 $Å^2$ to 1.34 $Å^2$. The inter-crystal $R_{sym}$ on $\langle|^{\lambda}F_{obs}|\rangle$ improved from 4.2% to 3.5% as a result of this scaling.

## DETERMINATION OF THE Fe PARTIAL STRUCTURE

Data were placed on an approximately absolute scale by careful calculation of the contents of the crystallographic unit cell and reasonable assumptions about temperature factors (20$Å^2$ for protein and estimated total ordered solvent and 150$Å^2$ for the solvent continuum). The 11 unmerged data sets from different crystals or orientations were phased by least-squares fit of multiple observations to the phase equation (5). To solve the Fe partial structure, phased data ($|^{\circ}F_T|$, $|^{\circ}F_A|$, $^{\circ}\phi_T$-$^{\circ}\phi_A$) were merged without further scaling (11). Statistics from comparison of redundant, phased reflections indicated that the phasing was probably okay to ~5.2Å spacings ($R_{merg}$ = .36 for $|^{\circ}F_A|$, rms $\Delta(^{\circ}\phi_T$-$^{\circ}\phi_A)$ = 38°), but that phasing for data at 3.0Å spacings was not at all reliable ($R_{merg}$ = .49 for $|^{\circ}F_A|$, rms $\Delta(^{\circ}\phi_T$-$^{\circ}\phi_A)$ = 77° for the 3.2-3.0Å shell). The noisiness of the data beyond ~5.2Å can also be seen in Fig. 2, where $|^{\circ}F_A|$ from MAD phasing is compared to $|F_{calc}|$. The MAD signal for 3.0Å data was expected to be weak relative to that at 5.2Å, where individual Fe atoms in each cluster are not resolved, but we had anticipated more precise results.

Initial analysis, in both Patterson and real space, was limited to 5.5Å data and then extended. A MAD Patterson map was computed from a selected subset of the $|^{\circ}F_A|$'s. Even at

**TABLE 2**
Scaling Statistics for ATase MAD Data

| Wavelength | Crystal | Total Observations | Outliers Rejected | Unique Reflections | Mean $I/\sigma_I$ | Rsym[1] Intra-crystal | Rsym[1] Inter-crystal[2] |
|---|---|---|---|---|---|---|---|
| 1.7425Å | | | | | | | |
| | AT15 | 10,330 | 0 | 4,206 | 21.8 | 0.044 | 0.063 |
| | AT19 | 13,716 | 0 | 5,668 | 23.4 | 0.038 | 0.068 |
| | AT20 | 73,443 | 4 | 17,753 | 32.8 | 0.039 | 0.046 |
| | AT22 | 12,442 | 0 | 3,827 | 33.0 | 0.058 | 0.062 |
| | AT23 | 16,984 | 0 | 7,219 | 25.7 | 0.057 | 0.057 |
| | AT24 | 80,361 | 7 | 10,413 | 16.5 | 0.097 | 0.073 |
| | AT26 | 47,446 | 13 | 11,658 | 22.8 | 0.058 | 0.049 |
| | AT31 | 16,824 | 0 | 3,894 | 23.7 | 0.083 | 0.072 |
| | Total | 271,546 | 24 | 64,638 (36,761 merged) | | | 0.059 |
| 1.7390Å | | | | | | | |
| | AT15 | 8,279 | 0 | 3,440 | 23.5 | 0.046 | 0.064 |
| | AT19 | 13,745 | 0 | 5,703 | 24.0 | 0.039 | 0.066 |
| | AT20 | 71,901 | 6 | 17,565 | 32.2 | 0.044 | 0.047 |
| | AT22 | 4,249 | 0 | 1,553 | 37.3 | 0.054 | 0.061 |
| | AT23 | 17,233 | 2 | 6,735 | 25.7 | 0.069 | 0.060 |
| | AT24 | 79,072 | 18 | 10,390 | 16.4 | 0.096 | 0.072 |
| | AT26 | 45,124 | 10 | 11,688 | 23.9 | 0.053 | 0.049 |
| | AT31 | 17,089 | 4 | 4,214 | 23.7 | 0.083 | 0.076 |
| | Total | 256,692 | 40 | 61,288 (35,922 merged) | | | 0.059 |
| 1.5000Å | | | | | | | |
| | AT15 | 10,223 | 0 | 4,234 | 29.7 | 0.035 | 0.049 |
| | AT19 | 12,887 | 0 | 5,350 | 30.9 | 0.033 | 0.047 |
| | AT20 | 58,679 | 10 | 14,496 | 38.6 | 0.035 | 0.041 |
| | AT22 | 6,735 | 12 | 2,334 | 42.4 | 0.050 | 0.050 |
| | AT23 | 13,993 | 0 | 6,128 | 29.6 | 0.051 | 0.053 |
| | AT24 | 77,568 | 2 | 9,679 | 24.1 | 0.055 | 0.048 |
| | AT26 | 50,283 | 8 | 11,861 | 29.1 | 0.048 | 0.040 |
| | AT31 | 18,461 | 0 | 4,102 | 34.3 | 0.048 | 0.042 |
| | Total | 248,829 | 32 | 58,184 (34,801 merged) | | | 0.045 |

[1] $R_{sym} = \sum_h \sum_i |I_{hi} - \langle I \rangle_h| / \sum_h \sum_i I(h)_i$. [2] Inter-crystal $R_{sym}$ for information only; no inter-crystal, single-wavelength scale factors were applied to the data.

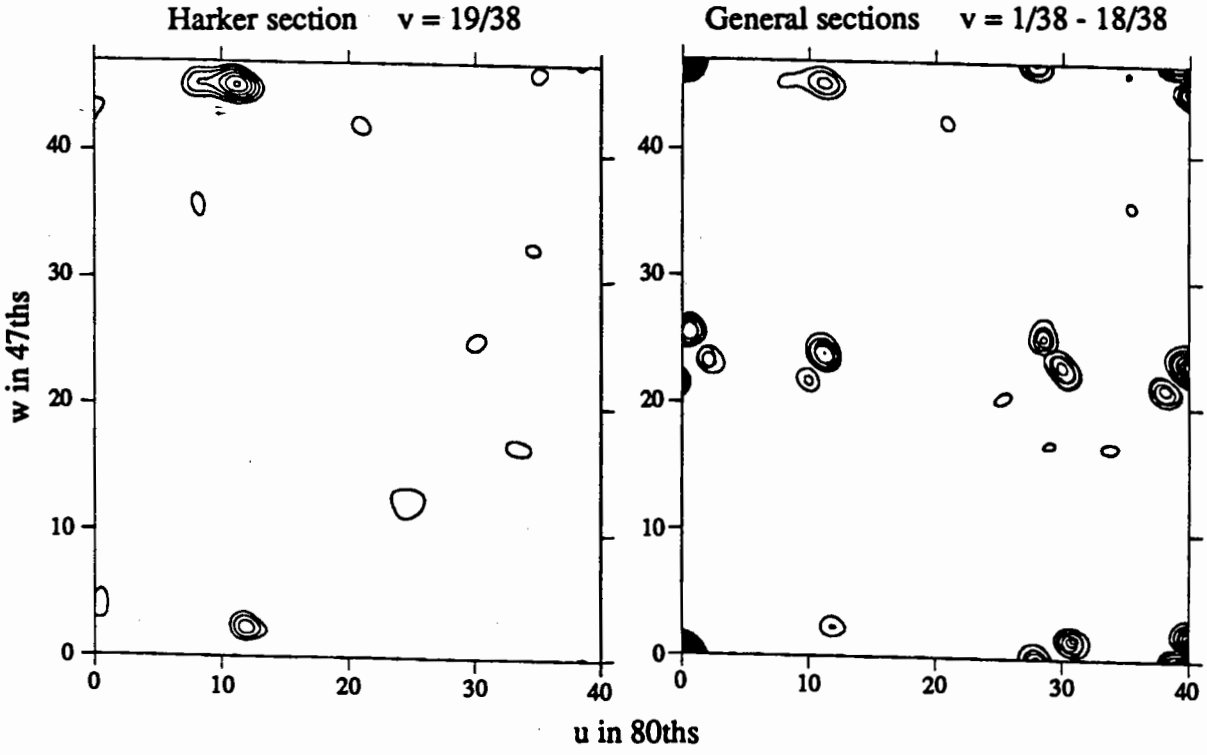



**Fig. 4. Patterson map at 5.5Å resolution based on selected $|°F_A|$'s.** Four Harker vectors are expected in the P2/m map; three of these are overlapped because the molecular axes are nearly parallel to the crystal axes. The twelve general peaks are the largest in the map, aside from the origin and the Harker peaks. The contour interval is on an arbitrary scale; the lowest level has been omitted.

5.5Å resolution, it is obvious that the $|°F_A|$'s include considerable error. An upper cut on $|°F_A|$ of 125% of the theoretical maximum (#Fe's x 26e⁻/Fe) yielded the cleanest map. More realistic cuts of 75-100% of the theoretical maximum resulted in much noisier maps. The 5.5Å Patterson map (Fig. 4) was easily interpretable for the cluster structure. The four cluster sites were related by D2 symmetry, as expected for the tetrameric ATase molecule. A model with one Fe atom of occupancy 4 for each of the four independent clusters refined to R = 0.262 for 2083 selected $|°F_A|$'s to 5.5Å spacings (3). The symmetry of the cluster sites was also in excellent agreement with the results of a self-rotation function (12,13).

Atomic positions for individual Fe atoms within each cluster were determined by fitting (14) a tetrahedral model (Fe-Fe = 2.75Å) to electron density for each of the four crystallographically independent $Fe_4S_4$ clusters. Amplitudes for this map were selected $|°F_A|$'s from MAD phasing to 3.0Å spacings and phases were calculated from the refined low-resolution model with each cluster as a single Fe of occupancy 4. Since the map contained no imposed knowledge of noncrystallographic symmetry, agreement of the fit with the molecular symmetry was used to assess the reliability of the result. When superimposed (15) by the known molecular symmetry operator, these individual Fe atomic sites had an rms deviation of 0.62Å, clearly showing agreement with the local symmetry. The individual-Fe model was refined in the program PROLSQ (16) against selected $|°F_A|$'s within 5.5Å spacings to an R factor of 0.260 with rms deviation of 0.013Å of intracluster Fe-Fe distances from the 2.75Å target, rms deviation from noncrystallographic symmetry of 0.045Å, and mean B of 14Å$^2$ with no deviation of B within a cluster. Attempts to refine this model against $|°F_A|$'s to 3.0Å spacings resulted in an R factor of 0.387 and unrealistic B values, again illustrating the problems with the phasing result from the high-angle data.

## PROTEIN PHASING AND PHASE REFINEMENT

We plan to produce an interpretable electron density map at 3.0Å resolution by the familiar process of phase refinement and extension by fourfold noncrystallographic symmetry averaging, solvent flattening, and combination of density-modified with experimental phases (17). ABCD coefficients for MAD phases (18) are being used to facilitate the phase combination. This formulation may also alleviate the problem of too few multiple measurements to phase many reflections. Phase probabilities (ABCD's) were calculated for all single measurements using the known Fe partial structure, with the assumption that $|^{\circ}F_T|$ can be represented by $\langle|^{\lambda}F_{obs}|\rangle$ for unphased reflections. ABCD's were then combined for all crystals and asymmetric units, and an initial electron density map was calculated at 5.5Å resolution.

The first test of the electron density map was to determine the correct enantiomer for the Fe partial structure. This does not come directly from the MAD phasing procedure, but rather by comparison of electron densities. Phases $^{\circ}\phi_T$ calculated from the correct and incorrect enantiomer for the Fe partial structure differ by $2^{\circ}\phi_A$; one phase set should produce a map with an image of ATase, the other a map without physical meaning. Correlation of electron density related by the molecular symmetry was used to distinguish the correct Fe enantiomer in this case. The correlation coefficient of the D2 molecular symmetry operator for grid points within 35Å of the molecular center, excluding the Fe clusters themselves, was 0.323 for the correct enantiomer and 0.086 for the incorrect enantiomer. Thus, the map was of sufficient quality to clearly distinguish the Fe enantiomer. Phase refinement (JLS and W.A. Hendrickson, unpublished) by fourfold averaging and solvent flattening is currently underway.

## GENERAL FEATURES OF MAD FOR LARGER STRUCTURES

The ATase MAD experiment thus far has been successful, if somewhat difficult. Considering the large mosaicity of the ATase crystals used in our MAD experiment, the data were measured with high precision. The MAD signal was readily detected for data to ~5.2Å. However, the weaker signal at higher angles, where Fe atoms scatter individually and not as clusters, was not accurately measured, although we were able to locate individual Fe atoms within each cluster with selected $|^{\circ}F_A|$'s to 3.0Å spacings. The success of the MAD experiment in this case hinges on accurate signal detection.

Signal detection has been the limiting factor in other MAD experiments as well. The average MAD signal for all reported examples of MAD phasing for macromolecule crystals is ~4% of |F| (1), quite similar to the signal from ATase crystals. Our results and those of others clearly show that such weak MAD signals are readily detectable with modern data collection techniques. However, the experiment must be carried out far more carefully than a typical isomorphous replacement experiment with a signal on the order of 10-20% of |F|. Film is not a suitable detector for MAD signals of this magnitude, for example. The data collection challenge is particularly acute for larger protein structures, which often produce more weakly diffracting crystals. The precision of data measurement often tolerated for larger structures is insufficient for detection of the relatively weak MAD signal for problems such as ATase.

The ATase MAD experiment, as well as nearly all others that have been reported, utilized the K absorption edge of the anomalous scatterer. K edges are limited to maximum magnitudes of about 9 $e^-$ for f' and about 4 $e^-$ for f". Thus, the only way to increase the MAD signal for K-edge scatterers is to increase their number in the macromolecule. A critical component of structure determination by MAD is the determination of the partial structure of anomalous scatterers from the $|^{\circ}F_A|$'s. This would be rather complicated if the number of anomalous

scattering centers were greater than 20 or so. Patterson or direct methods can both be used to solve the anomalous-scatterer partial structure. Direct methods is the more likely of the two to succeed with large numbers of scatterers or with space groups of high symmetry. However, the test of stereochemical reasonableness to assess a direct methods solution, which is commonly used in organic small molecule crystallography, is not applicable in most MAD cases because there is little *a priori* knowledge of the spatial relationship between anomalous scattering centers. This is a particular problem for larger proteins in the case of MAD via Se in biologically introduced selenomethionine (5,11). The typical concentration of methionine in proteins (1 in 59 residues) would give a MAD signal of ~4% of |F|. ATase has a slightly higher concentration (1 in ~36) for a signal of ~5%. Our Fe MAD experiment shows that this signal could be detected. However, solving the partial structure of 52 Se scatterers in the asymmetric unit, even with knowledge of the noncrystallographic symmetry, would be a nontrivial problem.

One way to get around this problem is to exploit the $L_{III}$ edges of metals such as Pb, Hg, Pt and Au. Theoretical $L_{III}$ edge features are typically two to three times larger than are those for theoretical K edges. Interaction of these elements with proteins is relatively well understood because of their extensive use as isomorphous derivatives. Finally, the edges occur in the range of 0.9Å to 1.1Å, which is very favorable for crystallographic data collection. Two cases of MAD phasing based on $L_{III}$ edges have been reported (19,20). Neither of these, however, was an optimized MAD experiment in which one wavelength was tuned to the inflection point of the edge (sharp minimum of f' ). Enhancement above theoretical values for f" may also be realized if sharp features due to the particular chemical environment of the scatterer exist for $L_{III}$ edges, which have been little studied for these elements in biological macromolecules. An example of the reduction in complexity of the anomalous scatterer partial structure relative to selenomethionine is illustrated by the ATase problem. A MAD signal equivalent to that from 52 selenomethionines would be generated by only eight Hg atoms in ATase and would present a far more tractable partial-structure problem.

Determination of larger protein structures by MAD is quite feasible, as this work shows. Although the experiments are challenging, the ability to determine phases directly from observed amplitudes promises both to speed the crystallography and to improve the quality of the result for biological macromolecules in general.

## ACKNOWLEDGEMENT

This work was supported by grant DK-42303 from the U.S. Public Health Service and in part by the Lucille P. Markey Foundation.

## REFERENCES

1. Smith, J.L. Curr. Opinion Struc. Biol. 1, in press (1991)
2. Fourme, R. and Hendrickson, W.A. in Synchrotron Radiation and Biophysics (ed. Hasnain, S.S.) 156-175 (Ellis Horwood Limited, Chichester, 1990)
3. Hendrickson, W.A., Smith, J.L., Phizackerley, R.P. and Merritt, E.A. Proteins: Struct., Funct., Genet. 4, 77-88 (1988)
4. Karle, J. Int. J. Quantum Chem.: Quantum Biol. Symp. 7, 357-367 (1980)
5. Hendrickson, W.A. Trans. Amer. Crystallogr. Assn. 21, 11-21 (1985)
6. Wong, J.Y., Bernlohr, D.A., Turnbough, C.L. and Switzer, R.L. Biochemistry 20, 5669-5674 (1981)
7. Averill, B.A., Dwivedi, A., Debrunner, P., Vollmer, S.J., Wong, J.Y. and Switzer, R.L. J. Biol. Chem. 255, 6007-6010 (1980)

8. Hendrickson, W.A. and Teeter, M.M. Nature (London) 290, 107-113 (1981)
9. Satow, Y. in Methods and Applications in Crystallographic Computing (eds. Hall, S.R. and Ashida, T.) 56-64 (Clarendon Press, Oxford, 1984)
10. Amemiya, Y., Satow, Y., Matsushita, T., Chikawa, J., Wakabayashi, K. and Miyahara, J. in Synchrotron Radiation in Chemistry and Biology II (ed. Mandelkow, E.) 121-144 (Springer-Verlag, Berlin, 1988)
11. Yang, W., Hendrickson, W.A., Crouch, R.J. and Satow, Y. Science (Washington, D.C.) 249, 1398-1405 (1990)
12. Fitzgerald, P.M.D. J. Appl. Cryst. 21, 273-278 (1988)
13. Crowther, R.A. in The Molecular Replacement Method (ed. Rossmann, M.G.) 173-178 (Gordon and Breach, New York, 1972)
14. Jones, T.A. J. Appl. Crystallogr. 11, 268-272 (1978)
15. Hendrickson, W.A. Acta Crystallogr. sect. A35, 158-163 (1979)
16. Hendrickson, W.A. and Konnert, J.H. in Computing in Crystallography (eds. Diamond, R., Ramaseshan, S. and Venkatesan, K.) 13.01-13.23 (Indian Academy of Sciences, Bangalore, 1980)
17. Bailey, S., Dodson, E. and Phillips, S. (eds.) Improving Protein Phases, (SERC Daresbury Laboratory, Warrington, UK, 1989)
18. Pähler, A., Smith, J.L. and Hendrickson, W.A. Acta Crystallogr. sect. A46, 537-540 (1990)
19. Kahn, R., Fourme, R., Bosshard, R., Chiadmi, M., Risler, J.L., Dideberg, O. and Wery, J.P. FEBS Lett. 179, 133-137 (1985)
20. Ryu, S.-E., Kwong, P.D., Truneh, A., Porter, T.G., Arthos, J., Rosenberg, M., Dai, X., Xuong, N., Axel, R., Sweet, R.W. and Hendrickson, W.A. Nature (London) 348, 419-426 (1990)

# HEAVY ATOM REFINEMENT AGAINST SOLVENT-FLATTENED AND LOCAL-SYMMETRY AVERAGED PHASES.

V. Cura*, A. D. Podjarny*, S. Khrishnaswamy*, B. Rees*, J. M. Rondeau*§, F. Tete*§, L. Mourey*, J. P. Samama* and D. Moras*.

* Laboratoire de Cristallographie Biologique, IBMC du CNRS, 15 rue Descartes, 67084 Strasbourg, and § Biostructure, Parc d'Innovation, 67400 Illkirch,France.

**Table of contents**



## INTRODUCTION

The refinement of heavy atom parameters, as seen in previous chapters,relies on the proper estimate of either the heavy atom amplitude or the protein phase. The first option is adequate if enough centric reflections or good quality anomalous dispersion data are available; in other cases, it is necessary to estimate the protein phase independently of the derivative under refinement (Dodson, 1976, and references therein). This task is reasonably simple if several derivatives of proper quality are available. However, it is often the case that no independent phase estimation is possible, either because of uneven derivatives, common sites or simply because only one derivative is available.

For these cases, it becomes crucial to decouple the direct connection between the protein phase and the heavy atom parameters. We have seen in previous chapters by Otwinowski and Bricogne a maximum likelihood approach to this problem, where new probability distribution functions are obtained using the same data. In this chapter, we will explore another possibility; that of modifying the phases via the introduction of new information about the electron density distribution.

These phase refinement procedures, which rely on the solvent content or on the existence of a local symmetry axis, can improve significantly the quality of the electron density map ( Podjarny et al, 1987, and references therein) . They have the additional property, crucial in this context, of rendering the set of phases under refinement more independent of heavy atom parameters.

A heavy atom refinement based on this concept has been applied with success to solve the structure of the Glutaminyl tRNA-synthetase complex ( Rould et al, 1989). We have applied the same idea to several cases in our laboratory,and also tested it in calculated cases.A detailed discussion of the results of this work follows.

## ALGORITHMS

Figure 1 shows a flowchart of the algorithm used in this work. A first set of heavy atom parameters is obtained from heavy atom amplitude refinement against centric differences, performed independently for each derivative(A). This set of parameters leads to a native MIR phase determination, where only the RMS lack of isomorphism is refined(B). The map obtained from this set of phases is subjected to a density modification procedure, where either solvent flattening or local symmetry averaging is used to improve the quality of the map. This density modification procedure includes a mask determination step. It is cycled until the phases do not change more than a preset value, and the resulting phases are then used to perform a phase refinement of heavy atom parameters where the difference between calculated and observed derivative amplitudes is minimized independently for each derivative(C). This new set of parameters is then used to recalculate a set of MIR native phases(B), which in turn is used to calculate a map which is subjected to the density modification procedure. The outer cycle is iterated until the heavy atom parameters converge.

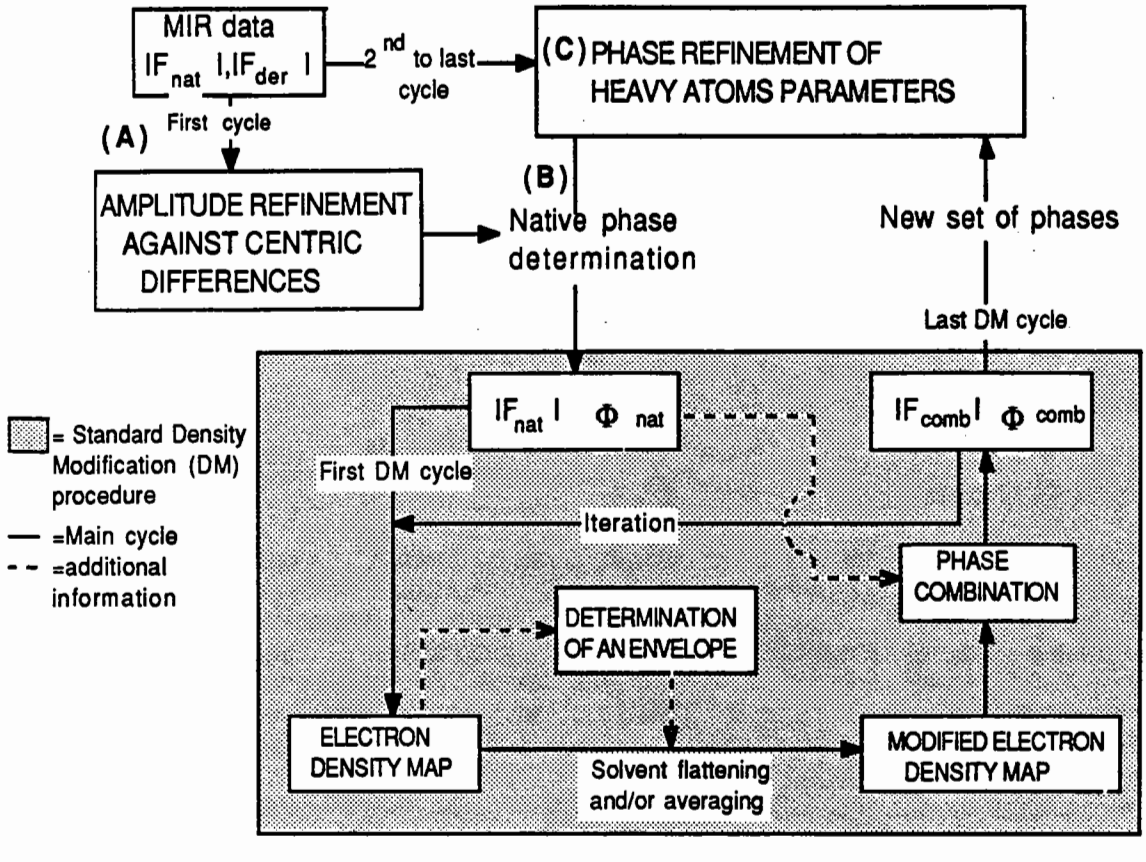



**Figure 1: Flowchart of combined heavy atom refinement/density modification procedure.**

This procedure has been implemented with programs of the CCP4 suite (Daresbury Laboratory, UK), such as REFINE2 for the heavy atom refinement, PHARE for the heavy atom phasing, and RFACTOR, TRUNCMAP, ENVELOPE, FLATMAP and FFT for the solvent flattening procedure. The local averaging was performed using the RMOL suite of programs , provided by B. Rees.

## TEST CASES

In order to check whether this procedure acts as a refinement of heavy atom parameters, several test cases were set up. The data for these cases were obtained from cardiotoxin (MW =6715 Daltons), which crystallizes in space group $P6_1$, a=b=73.9 Å, c=59.0 Å. The data were collected to a resolution of 3 Å, and the structure was solved by the MIR method (Rees et al, 1990). The test cases and the results are detailed below. Although the asymmetric unit contains a

dimer, the density was modified by solvent flattening only.

**1) Calculated case**

| | |
|---|---|
| $F_{nat}$ data : | From atomic positions with errors. |
| $F_{der}$ data : | Fnat + $F_{heavy}$ ( from two Pt atoms). |
| Starting positions: | True positions + generated errors |
| Starting occupations: | From centric data refinement. |
| Protein mask: | From model phases. 60% of volume. |
| Phase refinement: | Against density modified phases; Fom>0.8. |

**2) Observed case.**

| | |
|---|---|
| $F_{nat}$ data, $F_{der}$ data : | Measured data. |
| Starting positions: | True positions + generated errors |
| Starting occupations: | From centric data refinement. |
| Protein mask: | From SIR phases. 80% of volume. |
| Phase refinement: | Against density modified phases; Fom>0.8. |

**Results**

| CASE | | Occupations and positions of Platine1 and Platine2 | | | RESULTS | |
|---|---|---|---|---|---|---|
| | | initial | final | exact | initial | final |
| MODEL FNAT WITH ERRORS<br>MODEL ENVELOPE<br>SMALL POSITIONAL ERRORS<br>INITIAL CENTRIC REFINEMENT<br>3 Å RESOLUTION<br>SOLVENT VOLUME 40%<br>5 REFINEMENT CYCLES | Occ | 0,51 0,53 | 0,59 0,70 | 0,63 0,75 | < Delta Phi > | |
| | X | 0,450 0,530 | 0,469 0,545 | 0,469 0,545 | 73,21 | 51,16 |
| | Y | 0,490 0,350 | 0,501 0,370 | 0,501 0,370 | MAP CORREL. | |
| | Z | -0,098 -0,188 | -0,098 -0,188 | -0,098 -0,188 | 0,52 | 0,76 |
| FOBS<br>AUTOMATIC ENVELOPE (CCP4)<br>SMALL POSITIONAL ERRORS<br>INITIAL CENTRIC REFINEMENT<br>3 Å RESOLUTION<br>SOLVENT VOLUME 20%<br>AVERAGING RADIUS R=8Å<br>4 REFINEMENT CYCLES | Occ | 0,44 0,46 | 0,63 0,66 | 0,63 0,75 | < Delta Phi> | |
| | X | 0,450 0,530 | 0,475 0,540 | 0,469 0,545 | 78,77 | 69,33 |
| | Y | 0,490 0,350 | 0,503 0,371 | 0,501 0,370 | MAP CORREL. | |
| | Z | -0,098 -0,188 | -0,098 -0,188 | -0,098 -0,188 | 0,51 | 0,63 |

**Table 1. Results of test case refinements.**

Table 1 shows the result of these tests. The quality of the final phases is measured by the mean phase difference with true phases < Delta Phi> and that of the final map by its correlation with the true map, given by

Corr $(\rho_1,\rho_2)= \Sigma (F^2.fom.\cos(\phi_1-\phi_2))/ \Sigma (F^2.fom)$.

Figure 2 shows the evolution of heavy atom parameters as a function of refinement cycles. Note that the true occupancies are indicated, and that in both cases the refined ones are lower.In both cases, heavy atom parameters tend toward the true values. The error of SIR phases diminishes, and the correlation of the SIR map with the true map increases. **Therefore, in both cases the method acts as a refinement procedure.**

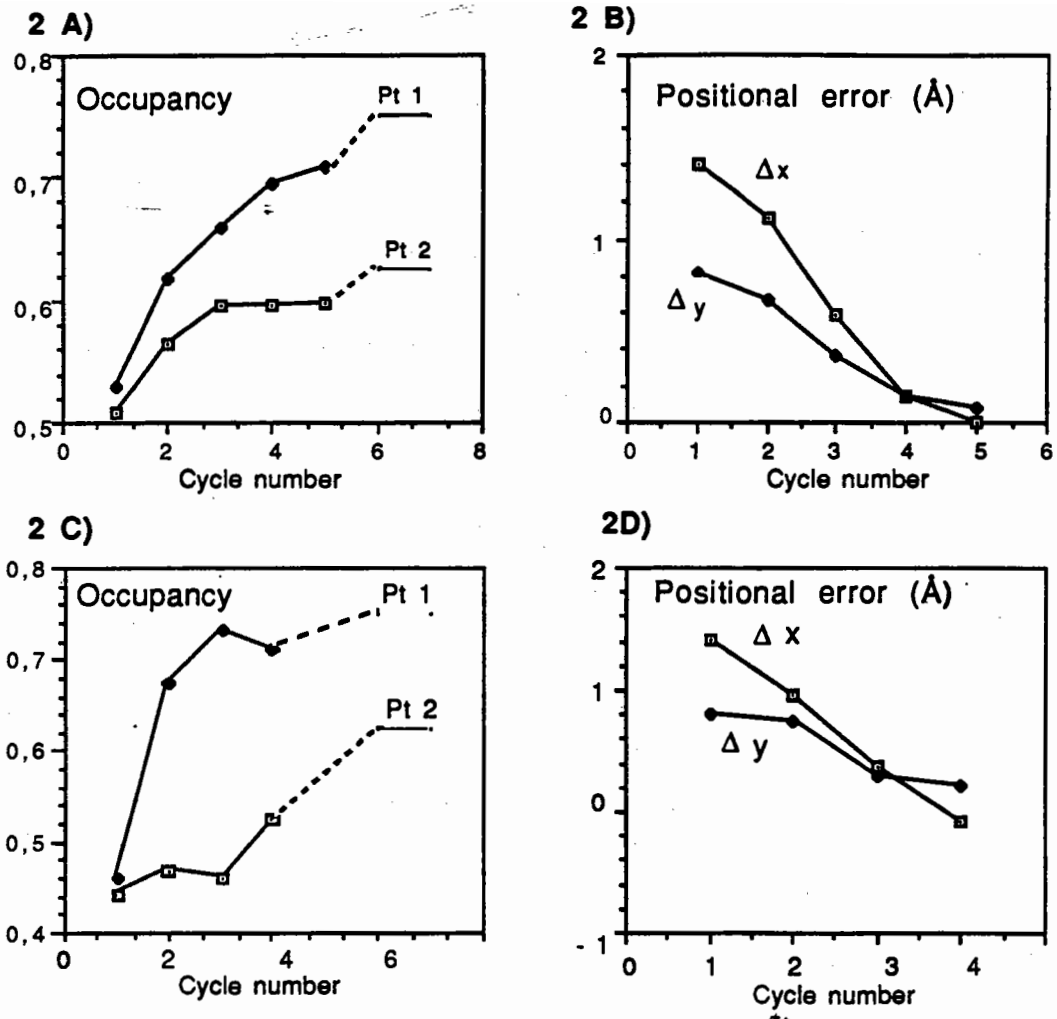



**Figure 2.** Evolution of occupancies (2A) and positional error (2B) as a function of cycle number.for the calculated cardiotoxin case Same results for the occupancies (2C) and positional errors(2D) for the observed case. Note the true occupancies marked as Pt1 and Pt2.

**PRACTICAL APPLICATIONS**

Since the proposed refinement method imposes extra constraints to the heavy atom parameters, it is possible that when used in addition to standard refinement methods it will improve their quality. This might be particulary useful for difficult cases where the electron density map is in the limit of interpretability, and each additional gain in phase quality is important. The following text describes some of cases where this method has been applied at the Laboratoire de Cristallographie Biologique, IBMC du CNRS, Strasbourg.

**1) The case of Aspartyl tRNA-Synthetase complex from yeast.**

This complex crystallizes in space group $P2_12_12$, a=210.25, b=146.17, c=85.13 Å, with one synthetase dimer (MW=125 Kdaltons) and two tRNA molecules (MW=25 Kdaltons) per asymmetric unit. Currently, an atomic model is being built (Ruff et al, in press) in a map calculated with phases obtained by MIR plus solvent flattening and local symmetry averaging(the 'current' phase set).

Heavy atom initial phasing

Data from a native crystal to 2.7 Å and from three heavy atom derivatives ( Hg to 4 Å, Au to 6 Å and Sm to 3.5 Å) were collected. Heavy atom sites were located by difference Patterson and cross difference Fouriers. Heavy atom parameters were refined against centric differences between 15 and 5 Å, with statistics of medium quality ( Table 2) . During the determination of native - derivative scales an algorithm developed by P. Dumas ( personal communication) based on the analysis of difference Patterson origin peaks was used.

| | Hg derivative 15-5 Å. 6 sites | AU derivative. 15-6 Å.Two sites. | Sm derivative. 15-5Å.Two sites. |
|---|---|---|---|
| R-Factor | 51% | 59%. | 66%. |
| Gradient | 59% | 31% | 13% |
| Correlation | 51%. | 44%. | 22%. |
| Phas.power | 2.49 | 1.74 | 1.51 |

**Table 2.** Centric refinement statistics as obtained from the program REFINE 2 ( For a definition of gradient and correlation, see Dodson ,1976).

These parameters were used to phase 10562 reflections between 15 and 5 Å, without cutoffs, with an overall FOM of 0.55.The data of the three derivatives were used to calculate a map with terms from 15 to 6 Å. Although this map showed regions of connected density **it was not clearly interpretable.** It should be noted that the correlation of this map with a similar one calculated with current phases is **very high (78.6%).**

Density modification and heavy atom iterative refinement

A density modification procedure by solvent flattening using an automatic mask determination (averaging radius = 15 Å, solvent volume = 30%) was started. The map showed clearly increased contrast which marked the molecular region. However, detailed analysis showed that the continuity had not really improved. Indeed, the correlation of this map with a similar one calculated with current phases is lower (**70.5%**) than in the MIR case.

At this point, a 5 cycle procedure of iterative density modification and heavy atom refinement with terms from 15 to 5Å was engaged, the final result and its comparison with the starting point being shown in Figure 3.

**Figure 3** . Comparison of initial 6 Å MIR map (A) and final map after density modification and heavy atom iterative refinement (B). H signals the position of a tRNA acceptor stem.

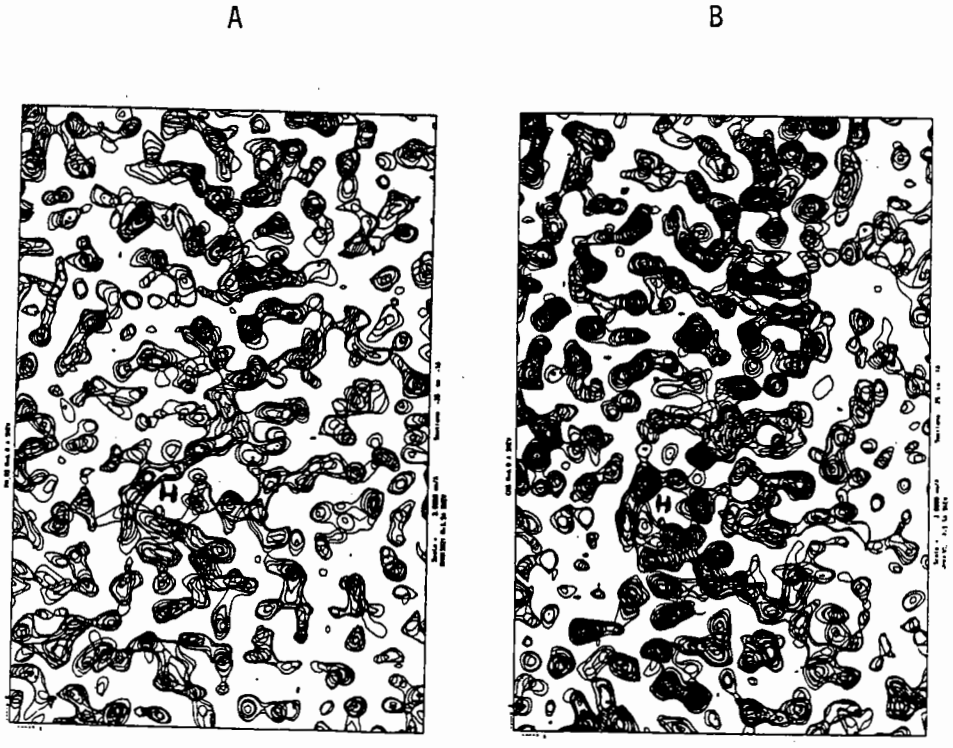


Figure 4 shows the evolution of the occupancy(Fig. 4.A), and of the phase difference with current phases (Fig. 4.B) . The final MIR figure of merit is 0.6.

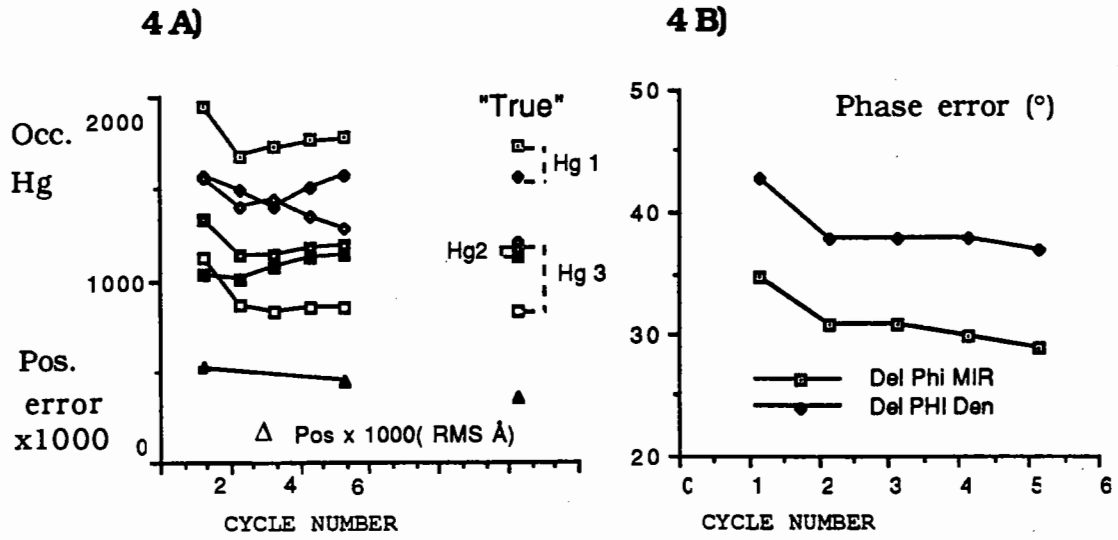



**Figure 4.** Evolution of mercury occupancy (4 A) and phase error of MIR and density modified phases (4 B) as function of cycle number. There are 6 Hg atoms, 3 in each of the two monomers. The 3 pairs are identified as Hg 1, Hg 2 and Hg 3. The "true" values are obtained by refinement against current phases. The positional error is measured by comparing the Hg positions in the two monomers.

The results can be summarized as follows:

1) The occupancies tend toward their true values and they tend to be the same for the same sites in different monomers ( Fig. 4A).

2) During the cycles, the phase and map quality improves both for the MIR and the DENMOD maps. For example, the correlation of the DENMOD maps with the current one improves from 70 to 74%, and that of the MIR maps improves from 78 to 82%. The mean error in MIR phases diminished from 36 to 28°.

Even when this improvement is clear in terms of statistics, the effect on map interpretability is small. However, the improved quality of MIR phases allowed the calculation of a better mask, using the following parameters: averaging radius = 10 Å, solvent volume = 40%.

The density modification procedure gave in this case a map with a correlation of 78% with the current one. The maps were recalculated at 4 Å, and this procedure was followed by a phase extension to 3.5 Å which eventually led to the current model.

It is interesting to note that density modified maps have better contrast than MIR maps (Fig. 3) even when their correlation with the correct map is poorer. This fact is probably linked to the error distribution as a function of amplitude, as shown in Fig. 5 by the correlation with the current map. It is clear that while in MIR maps the quality of the signal, as measured by the correlation, decreases with Fobs, in density modified maps it increases and becomes larger for the highest Fobs.

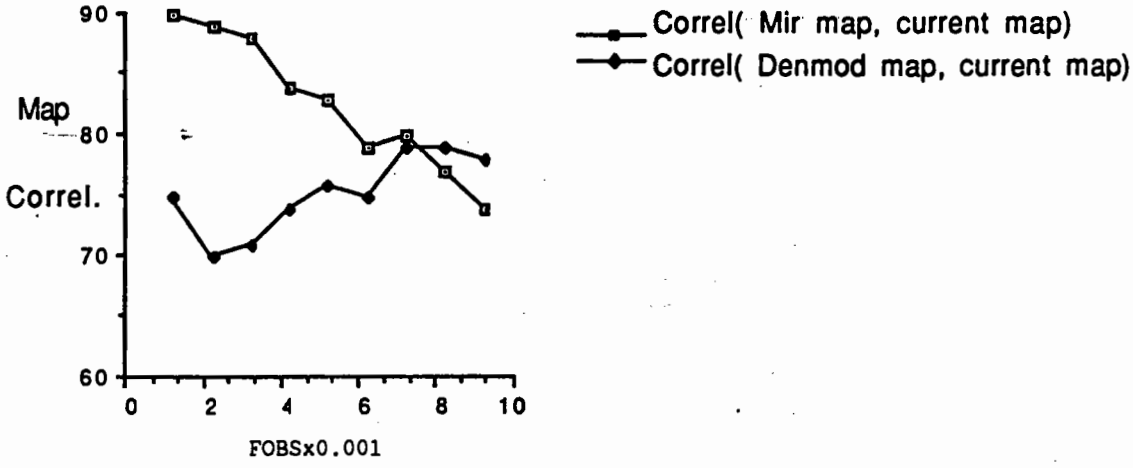



**FIGURE 5** . Distribution of phase error for MIR and density modified phases , measured as map correlation with current map, as a function of observed amplitude .

**2) The case of Aldose Reductase.**

Aldose reductase is an enzyme of molecular weight 35 Kdaltons which crystallizes in space group P1 with 4 copies per asymmetric unit, local symmetry 222. It is involved in diabetes complications. The molecule crystallizes also in the tetragonal space group $P4_12_12$, with one monomer per asymmetric unit. A mercury derivative was used and three sites per monomer were found, both in the triclinic and in the tetragonal space group. The structure has been solved and refined in both crystal forms ( F. Tete, J. M. Rondeau, A. Podjarny and D. Moras, manuscript in preparation).

Since the space group P1 has no centric zones, the heavy atom occupancies were determined using the program HEAVY (Terwilliger and Heisenberg, 1983). This led to a SIR map at 3.5 Å resolution. Accurate envelopes were determined from a map weighted by local-symmetry correlation (Rees et al, 1990). An averaging procedure was done using the local 222 symmetry and **the heavy atom parameters were refined against the resulting ("averaged") phases.**

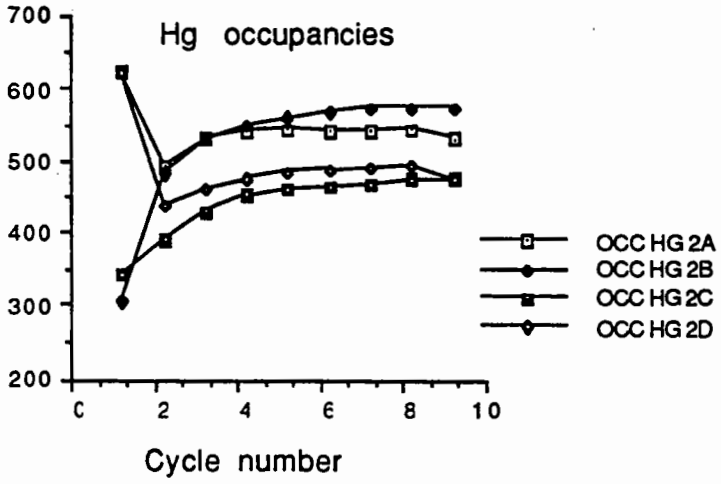



**Figure 6.** Evolution of Hg occupancies of local symmetry related sites during cycles of local symmetry averaging and heavy atom iterative refinement.

Figure 6 shows the evolution of one set of occupancies during this procedure.The effect on the averaged maps of the iterative refinement was small but significant (Fig. 7).

**Figure 7A**

3.5 Å map of aldose reductase obtained by averaging,starting from heavy atom parameters obtained by refinement by HEAVY.

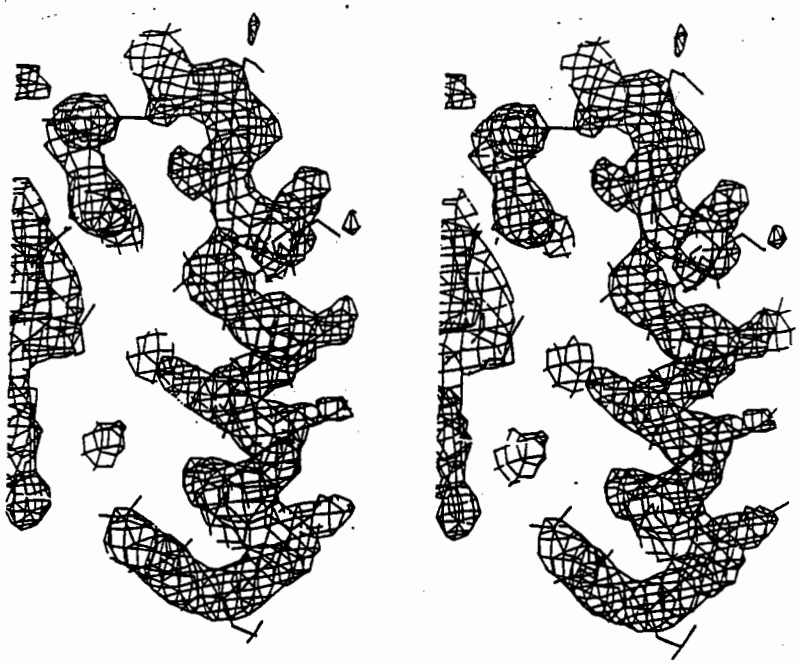
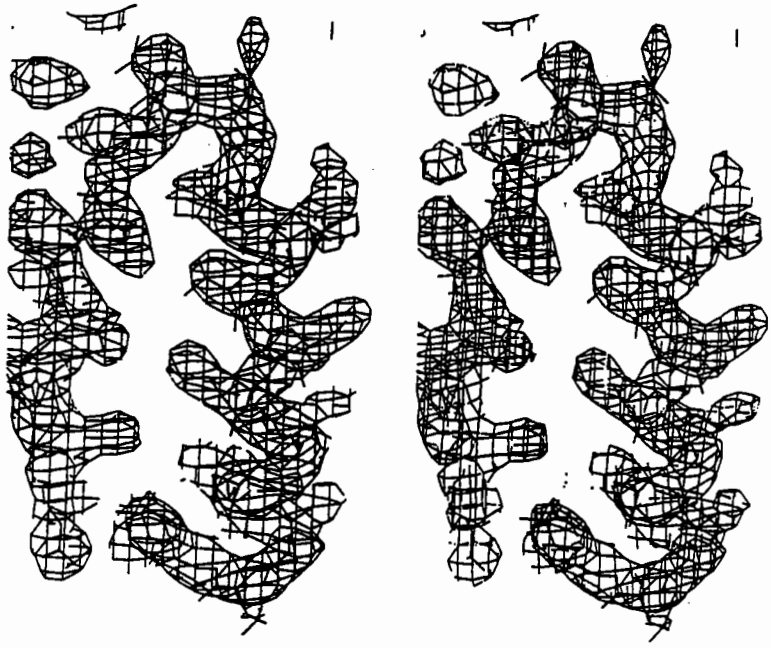


**Figure 7B**

3.5 Å map of aldose reductase obtained by averaging,starting from heavy atom parameters obtained by iterative refinement against averaged phases.



Phases from 3.5 to 2.7 Å resolution were determined using a phase extension procedure, during which the lower-resolution phases were refined. The resulting map is clearly interpretable and a model has ben built and refined to an R-factor of 24.6% at 2.5 Å resolution. Taking the last averaged phases as "true phases", a phase error can be estimated before and after refinement against averaged phases. This shows a very slight improvement of the SIR phase error from 46 to 44°, consistent with the fact that the maps are very similar.

**3) The Antithrombin III (AT III) case.**

ATIII is a protein of MW of 58000 Daltons involved in the coagulation process. It crystallizes in space group P43212 with two molecules per asymmetric unit. The phase problem was solved by molecular replacement using a search molecule of α1-antitrypsin (Mourey et al, 1990). A first model of ATIII was built and refined based on this solution, and the corresponding phases were used to locate 5 Pt heavy atom sites in Fourier difference maps. Heavy atom parameters were refined against centric reflections.The corresponding SIR map was subject to solvent flattening with a mask obtained from the molecular model which occupied 30% of the unit cell volume.

Heavy atom parameters were refined at the end of each density modification cycle (Fig. 8.1). The phase error after density modification, as measured against phases from the current model, changed from 75° to 73°(Fig.8.2). The last set of phases was the starting point for an averaging process which diminished the phase error to 65° (Fig. 8.2, point A) and the averaged phases were used to refine the heavy atom parameters (Fig. 8.1 point A). The resulting occupancies showed that the refinement against density modified phases did improve the overall quality of the occupancies even when the effect on final phases is small. As a check that the final occupancies(Fig 8.1, point A) are independent of the density modification procedure, the averaging procedure was repeated without the previous density modification step, giving a very close result (point B).

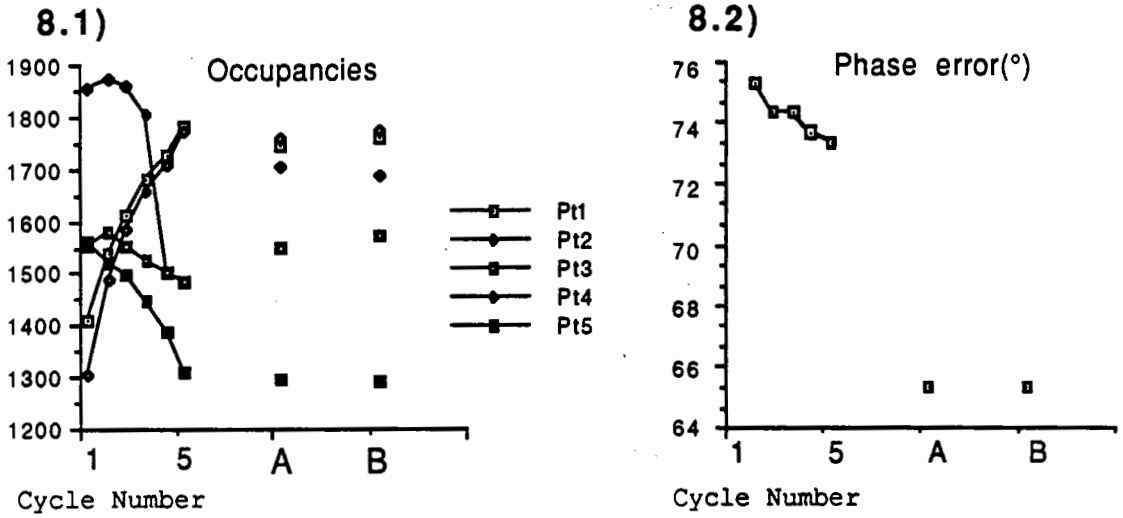



**Figure 8.** Effect of heavy atom re-refinement on heavy atom occupancies (1) and on phase error (2) after density modification (points 1-5). Note the improvement of occupancies (as compared with point A) for Pt1, Pt4 and Pt5. Pt3 remains stable, while the correction of Pt2 creates an error of similar magnitude but opposite sign. The effect of local symmetry averaging is shown by points A (starting from density modified phases) and B ( starting from heavy atom phases)

This overall process, after a last density modification step, led to a phase set at 4.5 Å resolution, which was extended to 3.2 Å resolution using averaging and solvent flattening. The map at 3.2 Å showed clear differences with the original model from molecular replacement. The model was rebuilt from this indications and final refinement is under way.

## CONCLUSIONS

1) Additional refinement of heavy atom parameters increases parameter accuracy and diminishes phase error. However, the size of this decrease might be too small to be significant.

2) Increased phase accuracy can lead to a better molecular envelope and to a more efficient density modification procedure.

**In this sense, the additional refinement of heavy atom parameters can and should be considered as part of a larger process were both the heavy atom parameters and the density modification parameters are varied.**

## REFERENCES

Dodson,E. Cryst. Computing Techniques, Munskgaard, 259-268. (1976)
Mourey, L. , Samama, J.P., Delarue, M., Choay, J., Lormeau, J.C., Petitou M. , and Moras D. Biochimie, 72 , 599-608,(1990).
Podjarny, A.D., Bhat,T.N. and Zwick,M. Ann. Rev. Biophys. Biophys. Chem, 16, 351-373(1987)
Rees, B., Bilwes, A., Samama, J.P.and Moras,D.,J. Mol. Biol., 214, 281-297 (1990)
Rould, M.A., Perona, J.J.,Soll, A. and Steitz, T.A., Science 246,1135-1141 (1989)
Ruff, M, Krishnaswamy, S., Boeglin, M., Poterszman, A. ,Mitschler, Podjarny, A., Rees, B., Thierry J.C. and Moras,D., Science, in press.
Terwilliger, T.C. and Heisenberg, D. Acta Cryst. A39 , 813-817 (1983)

# The Structure Determination of Galactose Oxidase by Multiple Isomorphous Replacement with Anomalous Scattering

Nobutoshi Ito

Department of Biochemistry & Molecular Biology
University of Leeds

Isomorphous replacement with anomalous scattering has been the most successful method for solving the phase problem in the determination of *ab initio* structures in X-ray protein crystallography. Despite the importance of the quality of derivatives, there have been few reports where details such as binding site and effect on the protein structure have been investigated after the structure of the protein was solved.

The structure of galactose oxidase has been solved by this method and refined to 1.7Å [1]. The three derivatives used in the study have different binding sites from one another. Relatively high resolution derivative data (2.4–2.2Å) have made it possible to investigate the nature of the heavy-atom binding. In this report, the phase determination will be described in detail as well as the quality of each derivative as a case study.

## 1. Progress of the Phase Determination

Galactose oxidase (GOase; EC 1.1.3.9) is an extracellular enzyme from the fungus *Dactylium dendroides* and consists of a single polypeptide chain of 68kD with a copper ion as a sole cofactor [2]. The primary sequence was not known before the MIRAS map was obtained in this study. GOase was crystallised from 0.80M acetate buffer (pH~4.5) with ammonium sulphate as precipitant. The space group is $C2$ with unit cell parameters $a = 98.0$, $b = 89.4$, $c = 86.7$Å, $\beta = 117.8°$. There is one protein molecule per asymmetric unit, and the solvent content is 50%.

The structure was solved from the 2.5Å MIRAS map and refined to 1.7Å (Fig. 1). Heavy-atom derivatives were prepared by soaking the native crystals in various solutions of heavy-atom compounds. 0.1M acetate buffer (pH4.5) with 25% PEG6000 was used as the soaking buffer to avoid inhibition of heavy-atom binding by ammonium ion.

All the data sets used here were collected with Xentronics/Siemens area detector mounted on a Rigaku rotating anode generator. The data frames were integrated by XDS [3], and all other crystallographic calculations were done with the CCP4 programme suite. Statistics for the derivatives are shown in Table 1.

**$K_2PtCl_4$** The first derivative found to be useful was $K_2PtCl_4$. Both 3Å isomorphous and anomalous difference Patterson maps show a clear single peak on $v = 0$ plane, which is the Harker section of the space group. The peak indicates the presence of a single platinum site per asymmetric unit. Since the origin is arbitrary in $y$ direction in $C2$, the

| Compound | $K_2PtCl_4$ | $H_2IrCl_6$ | $Pb(NO_3)_2$ |
|---|---|---|---|
| Concentration | 10mM | 2mM | 200mM |
| Soaking time | 1 day | 1 day | 13 days |
| No. of binding sites | 1 | 3 | 10 |
| $\Delta_{ISO}$ (20–2.5Å) | 16.1% | 14.9% | 13.9% |
| $k_{EMP}$ (20–2.5Å) | 4.2 | 4.9 | 3.9 |
| $F_{HLE}$ refinement (20–3.0Å, centric zone only) | | | |
| $R_{CENTRIC}$ | 54% | 45% | 56% |
| Correlation | 0.32 | 0.62 | 0.38 |
| Gradient | 0.22 | 0.53 | 0.26 |
| SIRAS phase (20–2.5Å) | | | |
| Figure of merit | 0.29 | 0.43 | 0.30 |
| Average phase error† | 64.6° | 52.6° | 63.1° |

**Table 1** Statistics of the heavy-atom derivatives.
† Difference from the final model phases.

$y$ coordinate of the platinum site was defined as zero. Thus the arrangement of the platinum atoms in the crystal is centrosymmetric, so that the problem of enantiomorphism ("the choice of hand") is avoided. The correct enantiomorph for subsequent derivatives could be determined directly from difference Fourier maps phased from isomorphous and anomalous data for the platinum derivative.

**$H_2IrCl_6$** After $K_2PtCl_4$, $H_2IrCl_6$ was also found to be useful. A cluster of peaks can be clearly seen in the isomorphous difference Patterson map, and similar but weaker features can be seen in the anomalous difference Patterson map. With the help of a difference Fourier map with phases calculated from the platinum derivative ($\alpha_P$(Pt)), two major and one minor sites clustering as a group were deduced from these peaks.

The phases recalculated with the two derivatives ($\alpha_P$(Pt,Ir)) have a much higher figure of merit than $\alpha_P$(Pt) or $\alpha_P$(Ir). The 6Å MIRAS map clearly shows the solvent regions in the crystal lattice. In the 3Å map, although the solvent regions are less flat, the molecular envelope is recognisable.

**Problem of a mirror image** One serious problem with the two derivatives is that the three iridium atoms have almost the same $y$ coordinate as the platinum (i.e. zero; Fig. 1). This implies that the iridium atoms are arranged in a pseudo-centrosymmetrical manner, and that their $F_H$ is almost collinear with that of the platinum. Consequently the phase ambiguity is resolved almost solely by the anomalous scattering. If the anomalous scattering was too weak for a protein of 68kD, the MIRAS map would be a mixture of the true electron density with its mirror image across $y=0$ plane.

The presence of the mirror image was confirmed when the data from the apoenzyme (copper-removed enzyme) were collected. The difference Fourier map with $\alpha_P$(Pt,Ir) shows a very strong peak at $(x, y, z)$ corresponding to the copper, as well as a significant

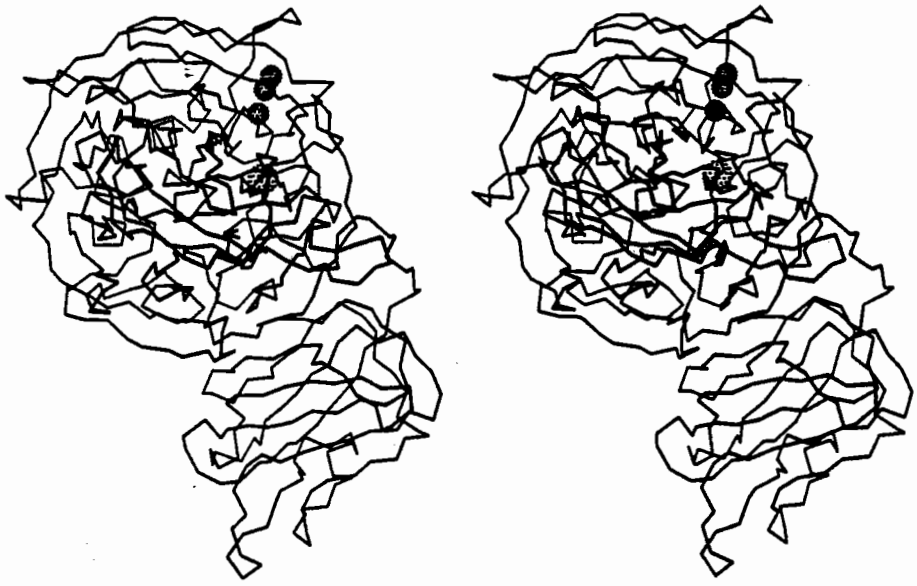


**Fig. 1** A stereo view of the Cα trace of GOase, looking down $x$ ($a^*$) axis. $y$ ($b$) axis is horizontal. The larger sphere indicates the platinum binding site, and the smaller ones for the iridium sites.

peak at $(x, -y, z)$. Another derivative with a heavy-atom having $y \neq 0$ was necessary to solve this problem. Obviously the apoprotein itself, which could be seen as a derivative with the copper of negative occupancy, was an immediate candidate, but the copper proved too light to phase this relatively large protein.

**$Pb(NO_3)_2$** When an apoenzyme crystal was soaked in 100mM $Pb(NO_3)_2$ solution in the hope of replacing the copper with a lead ion, lead ions were found to bind relatively weakly to the protein at several positions other than the copper site. Subsequently a native crystal was soaked in 200mM $Pb(NO_3)_2$, which was the highest concentration achievable without damaging the crystal.

The difference Fourier map indicates the presence of 10 lead sites, many of which are not close to $y = 0$ plane. The final MIRAS phases ($\alpha_P$(Pt,Ir,Pb)) have an average figure of merit of 0.58 at 20–2.5Å. Although the improvement relative to $\alpha_P$(Pt,Ir) of the figure of merit and, indeed, phasing error is small (0.06 and 4°, respectively), the ratio between the true copper peak and the mirror image in the difference Fourier map for apoenzyme increased about 70% with $\alpha_P$(Pt,Ir,Pb). The mirror image was almost at noise level, suggesting the phase ambiguity had been resolved by the third derivative.

After 10 cycles of solvent flattening (SF) [4], the figure of merit had improved to 0.80, and the map was of excellent quality. Table 2 summarises the progress of the phase determination.[1]

[1]The difference of FOM and $\Delta\alpha$ in Table 1 & 2 is due to the difference of atomic parameters used. The SIRAS phases in Table 1 are calculated with those used in the final MIRAS (Pt+Ir+Pb) whereas Pt and Ir in Table 2 are values obtained during the progress of the phase determination.

| Derivatives | Height of copper peaks | | | | | FOM | $\Delta\alpha$ |
|---|---|---|---|---|---|---|---|
| | True | | Mirror | | Ratio | | |
| Pt | 503.9 | (24.0) | 264.4 | (12.6) | 1.9 | 0.32 | 65° |
| Ir | 827.8 | (30.1) | 277.7 | (10.1) | 3.0 | 0.41 | 54° |
| Pt+Ir | 1175.9 | (36.3) | 301.5 | ( 9.3) | 3.9 | 0.52 | 47° |
| Pt+Ir+Pb | 1381.6 | (38.9) | 210.5 | ( 5.9) | 6.6 | 0.58 | 43° |
| Pt+Ir+Pb+SF | 2012.3 | (43.4) | 42.1 | ( 0.9) | 47.8 | 0.80 | 32° |
| Model | 2472.3 | (43.3) | 117.2 | ( 2.1) | 21.1 | | |

Numbers in ( ) are intensity divided by RMSD of the map
$\Delta\alpha$ : Average phase difference from the model phase (acentric zone)

**Table 2** Peak height in 2.5Å difference Fourier maps for apo-GOase.

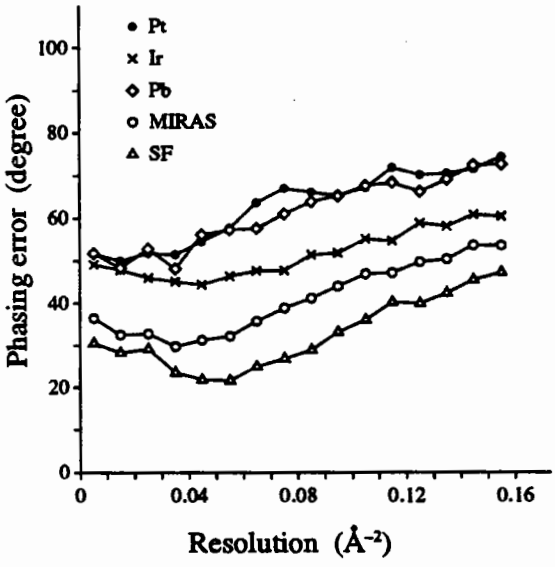



**Fig. 2** Phase differences from the model phase (acentric zone).

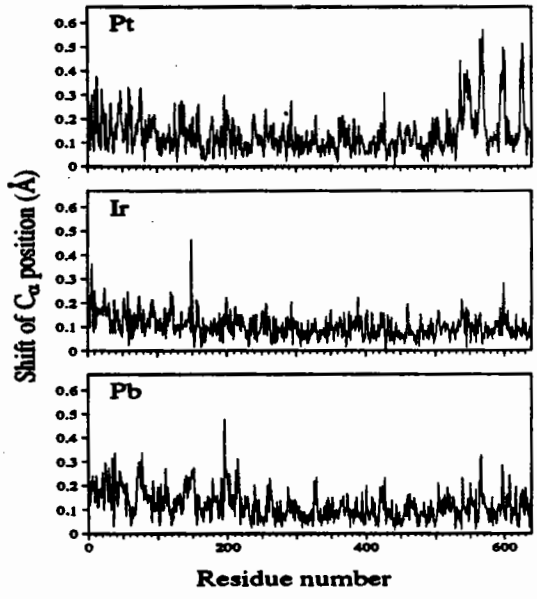



**Fig. 3** Shift of Cα atoms in the derivatives.

## 2. Nature of Heavy-Atom Binding

All statistics in Table 1 indicate the superiority of $H_2IrCl_6$ to $K_2PtCl_4$ and $Pb(NO_3)_2$. This is confirmed when the experimental phases are compared with the model phase calculated from the refined protein structure (Fig. 2). SIRAS phases for the iridium derivative are much better than those for the other two. It would be interesting to understand the reasons underlying the differences between the derivatives at molecular level.

After the protein structure had been solved and refined to 1.7Å, the heavy-atom derivatives were reinvestigated. Each derivative has been refined as an independent structure. The binding sites of the heavy-atoms, and the effects of the binding are described below on the basis of the refined derivative structures.

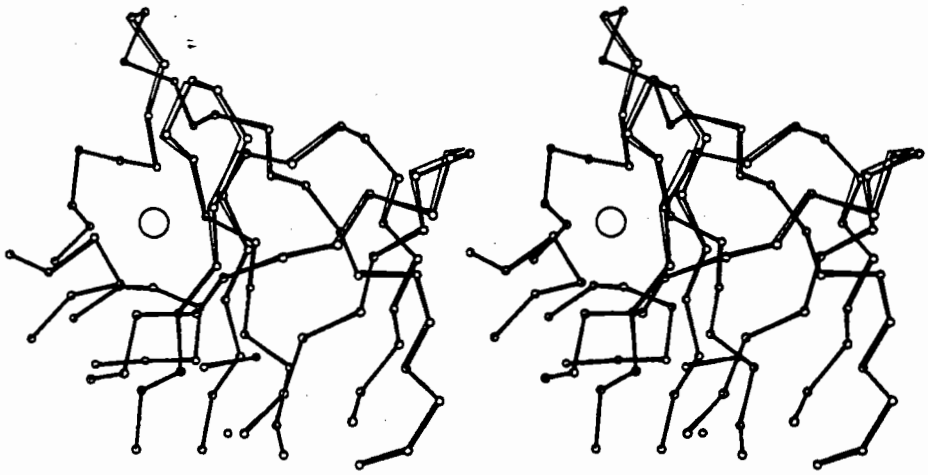


**Fig. 4** A stereo view of the Cα trace around the platinum binding site. The derivative is shown in the ball & stick, and the native in the thin lines. The large circle is the platinum binding site.

**$K_2PtCl_4$** The platinum atom binds covalently to the $S_\delta$ of Met623, which is a fairly common binding site for the compound. The difference Fourier map with the refined native phase shows a strong density whose shape is an inflated square (like a square cushion). The distortion from a sphere appears to be due to the ligands coordinated to the platinum. In the original compound, $K_2PtCl_4$, the platinum atom has four ligated chlorine atoms in a square planar coordination. One of them is obviously replaced with the sulphur, and two chlorine atoms next to the sulphur can be seen in the refined map of the derivative. The fourth site at the *trans* position from the sulphur has much lower density than the two chlorine sites, and yet too much to be due to the platinum alone. It is difficult to determine the exact species of this ligand, although there are several possibilities; residual ammonium from crystallisation medium, acetate ion in the buffer, or a water molecule. In any case, it is likely that the binding of the sulphur in the thioether, which has a very strong *trans* effect [5], weakened the bond between the platinum and the *trans* chlorine and contributed to the replacement of the ligand.

The platinum binding site is deep in the cleft between two domains of the protein, and the binding of the platinate complex, causes conformational change in the protein. Four β-strand–turn–β-strand motifs in the smaller domain are "pushed out" by the intrusion of the platinate complex (Fig. 4). Shifts of Cα positions in the derivatives relative to the native structure are shown in Fig. 3 as a function of residue number. Here the four regions are seen as four peaks at residues 530–639. These shifts are much larger than those in the iridium and lead derivatives.

The large change in the protein structure, i.e. lack of isomorphism, reduces the quality of this derivative, despite the full occupancy of the site.

**$H_2IrCl_6$** The iridium cluster binds in a small pocket on the protein surface. The pocket has Arg589 at the bottom and Asp379 and Glu479 as the wall, suggesting the interactions between the cluster and protein are electrostatic. The binding of the iridium cluster causes hardly any change on the protein structure, except for Lys393 whose side

chain, invisible in the native structure due to high B-factors or disorder, is seen as some density next to the cluster.

When the structure was refined with iridium atoms at the two major sites, the presence of several ligands became evident in the Fo–Fc map. The ligands, which must be responsible for the formation of the cluster, are probably chlorides. Each of the two iridium atoms has six ligands in octahedral coordination, two of which are shared between them and act as bridging ligands. The coordination of the third iridium atom of the minor site is less clear. At least two of the six ligands of the second site are coordinated to the third. Although the third iridium may also have six ligands, it is impossible to confirm this because of its lower occupancy and/or higher B-factor.

The presence of the ligands, which was not considered in the phase calculations, does not seem to have disturbed the quality of the derivative. This may be explained by the octahedral coordination, where the extra electron density of the ligands is distributed in a roughly isotropic way. The apparent occupancies and B-factors of the iridium atoms might have compensated for it to some extent.

The high degree of isomorphism, together with the formation of a "heavy" cluster, makes this derivative the best of the three.

**$Pb(NO_3)_2$** In all cases except one, the lead ions bind to aspartate or glutamate side chains on the protein surface. The low pH of the buffer is probably the reason for the weak binding of the cation. During the phase calculation, nine lead atoms had been identified at such sites.

The binding site of the lead with the highest occupancy is somewhat different from the other sites. It is surrounded by a loop and an adjacent strand of the protein. Four main chain (Lys29, Thr37, Asn34, and Ala141) and one side chain (Asp32) oxygens form the first shell of ligands (2.0–2.4Å) with side chains of Thr37 and Glu142 in the second (~2.6Å). In the native structure this site is occupied by another external ligand, which is currently assigned as a sodium ion, the most abundant cation in the solution. One practical lesson which may be learned here is that the location of this lead site helped the interpretation of the native map; this part of the protein has a relatively high B-factor, and the presence of the metal ion had not been perfectly clear.

Unlike the platinum derivative, no significant structural change can be seen in the derivative structure (Fig. 3). There are several possible reasons for the relatively poor quality of the derivative;

1. *Presence of minor sites.* During the refinement of the derivative structure, several minor binding sites were found. Some water molecules with stronger density than in the native structure are also possible minor sites.
2. *Anisotropy of the binding.* Many of the lead sites have rather distorted electron density peaks. This is probably due to the less specific nature of the binding and was not represented properly by isotropic B-factors in the heavy-atom parameter refinement.
3. *Insufficient isomorphous change.* This is the most serious problem for the derivative. Since all the sites are weakly occupied, $F_H$ is significantly smaller than that of the other derivatives (Fig. 5). The small $F_H$ gives reduced phasing power as well

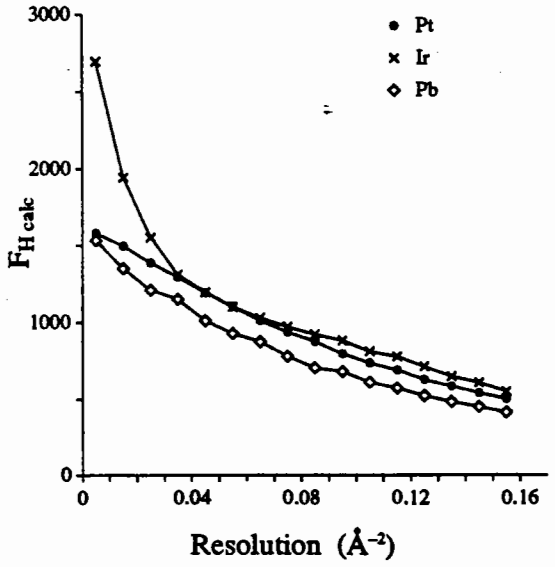



**Fig. 5** Heavy-atom structure factor amplitude (acentric zone).

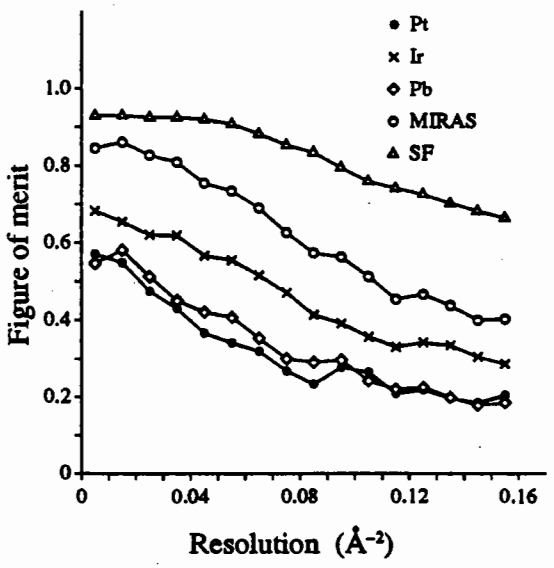



**Fig. 6** Average figure of merit (acentric zone).

as the difficulties in the refinement.

## 3. Reliability of Statistics

It would be very useful if one could assess the quality of derivatives during the course of the phase determination. There are a number of criteria proposed to be useful in the estimation of the quality of derivatives and the phases derived from them.

**Figure of merit** The comparison between Fig. 2 and 6 indicates that the figures of merit were qualitatively a good criterion of the accuracy of phases, i.e., (a) superiority of the iridium derivative (b) significant improvement by MIRAS and SF over the whole range of resolution. Quantitatively, however, they were greatly underestimated, taking $< m > = \cos\theta$ where $\theta$ is the mean phasing error. This is due to our conservative estimation of the lack of closure, where the centric $E$ value was used. The figure of merit for the combined phase between MIRAS and SF shows better agreement with the phase error.

As mentioned earlier, the lead derivative did not change the figure of merit very much, even though it resolved the phase ambiguity. Figure of merit may not be very sensitive to such systematic bias.

**Phasing power** Fig. 7 shows the phasing power of the three derivatives. The superiority of the iridium derivative is again evident. The fact that the platinum and lead derivatives retain phasing power larger than one even at high resolution explains the significant improvement of the phase by these derivatives, although they are far less powerful than the iridium.

**Isomorphous difference and lack of closure** It has been suggested that the lack of isomorphism can be detected as a rapid increase of isomorphous difference at higher

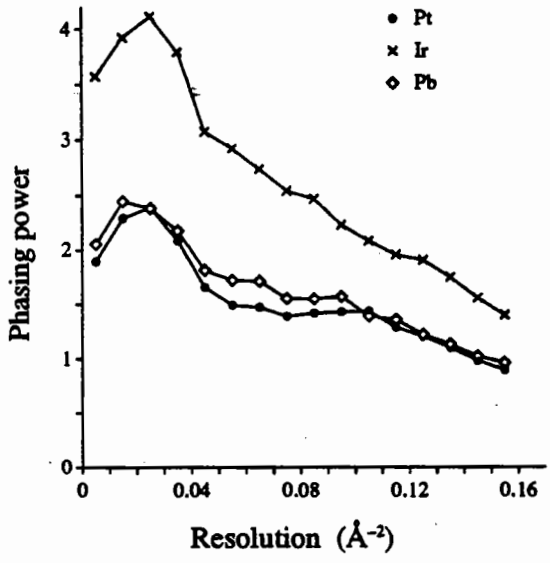



**Fig. 7** Phasing power defined as $F_{Hcalc}/E$ (acentric zone).

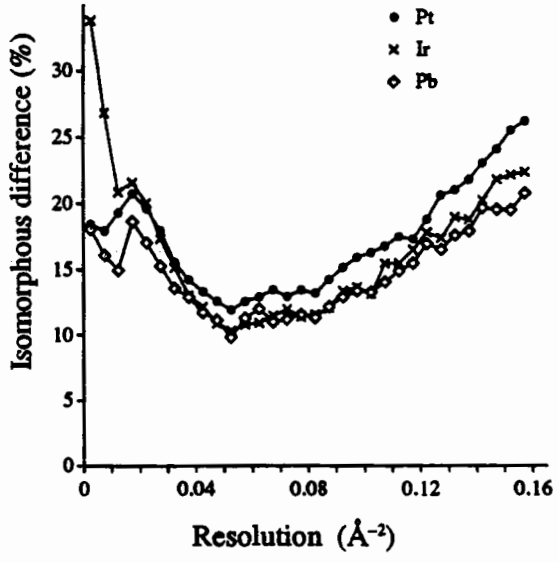



**Fig. 8** Isomorphous differences for the derivatives.

resolution. However, the plot of the isomorphous change against resolution shows no significant differences among the three derivatives (Fig. 8). This may not be surprising considering that, even though the platinum derivative is not as isomorphous as the iridium one, it still improves the phase throughout the resolution range, i.e. 20–2.5Å. Its lack of isomorphism is probably not severe enough to make the derivative totally useless, and is not detectable in the plot.

While the isomorphous difference failed to suggest the lack of isomorphism in the platinum derivative, the lack of closure ($E$ value in the phase calculation) clearly shows it (Fig. 9). However, one should be aware that the lack of closure obtained in the phase calculation is likely to be biased in favor of the most powerful derivative. In fact, "theoretical" lack of closure derived from the refined structures ($|F_{PcalcNAT} - F_{PcalcDER}|$) shows much smaller difference between the iridium and lead derivatives (Fig. 10) though $E$ value represents general tendency reasonably well.

## 4. Conclusion

Good derivatives should have large phasing power. In other words, it should fulfil two conditions; large isomorphous change and small lack of closure. The detailed analysis of the three derivatives has revealed one case where both conditions are met ($H_2IrCl_6$), and two cases where only one is met ($K_2PtCl_4$ and $Pb(NO_3)_2$).

Statistics available during the phase determination seem to be good indicators of the progress, at least qualitatively. Used with care, they can provide useful information.

Finally the confirmation of the heavy-atom binding site can offer extra information for the interpretation of the native structure itself, which would add to the reliability of the model.

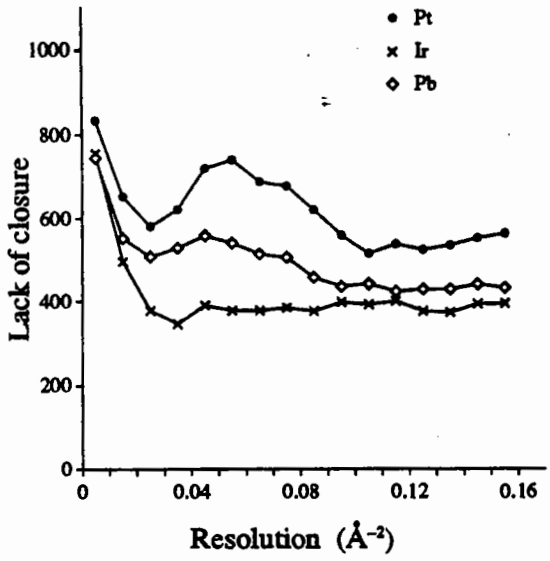



**Fig. 9** Observed lack of closure. $E$ value from PHASE is used (acentric zone).

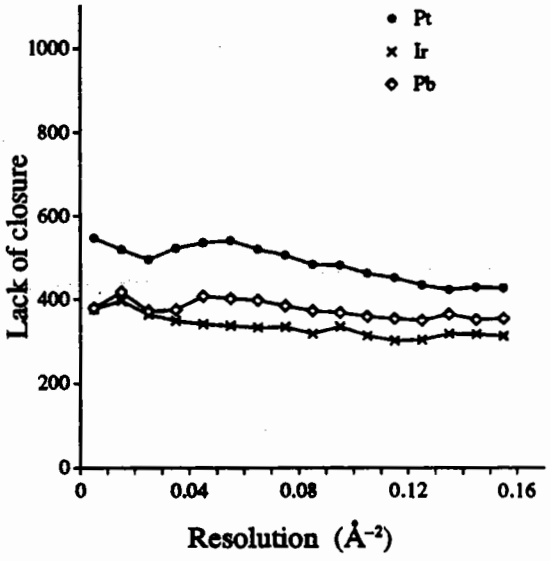



**Fig. 10** Calculated lack of closure ($|F_{PcalcNAT} - F_{PcalcDER}|$, acentric zone).


## Acknowledgement

I thank Dr Simon E. V. Phillips for useful discussions as well as for his help during the original structure determination.


## References


[1] Ito,N., Phillips,S.E.V., Stevens,C., Ogel,Z.B., McPherson,M.J., Keen,J.N., Yadav,-K.D.S., and Knowles,P.F. *Nature* **350**, 87–90 (1991).

[2] Kosman,D.J. in *Copper Proteins and Copper Enzymes* Vol. 1, 1–26 (R.Lontie, ed.) CRC Press, Boca Raton Publishers, Boca Raton (1984).

[3] Kabsch,W. *J. appl. Crystallogr.* **21**, 916–924 (1988).

[4] Wang,B.C. *Meth. Enzym.* **115**, 90–12 (1985).

[5] Tobe,M.N. in *Comprehensive Coordination Chemistry* Vol. 1, Wilkinson,G. Ed., Pergamon Books (1987).

Theory and Practice in the use of Heavy Atom Substitution.
E.J. Dodson

My brief is to re-examine the type of heavy atom substitution in some solved protein structures.

The theory of isomorphous replacement assumes that the heavy atoms bind without causing any perturbation in the protein structure, but this is often wishful thinking. I will describe what conformational changes have occurred in several proteins and then try to pinpoint some features of the observed data which could help us screen for useful heavy atom derivatives.

To recap briefly on the theory:

Terminology used:

- Fp - structure factor of protein
- Fph - structure factor of protein with heavy atom
- Fh - structure factor of heavy atom
- |F| - amplitude
- f - scattering factor
- n atoms. number of protein atoms
- n.h.a. - number of heavy atoms

1) Wilson statistics suggest that

$$\langle I_P(\theta)\rangle = \sum_{i=1}^{n\ atoms} f_i^2(\theta)$$

and r.m.s. $\langle F_P(\theta)\rangle = \sqrt{\langle I_P(\theta)\rangle}$

Then

$$\langle I_{PH}(\theta)\rangle = \underbrace{\sum_{i=1}^{n\ atoms} f_i^2(\theta)}_{(protein)} + \underbrace{\sum_{i=1}^{n.h.a.} f_H^2(\theta)}_{(heavy\ atoms)}$$

This means that there is a significant contribution to the average intensity from even a single site heavy atom. If the "protein" consisted of 1000 nitrogen atoms, $I_P(\theta=o)$ would equal 49000 and the contribution added for a single Hg atom would be 6400, almost 10%. (This means that the assumption usually made for the initial scaling that $\langle F_P\rangle$ should equal $\langle F_{ph}\rangle$ is inaccurate, but since this scaling is only used to calculate difference Pattersons, which are insensitive to errors in scale it is not serious. During the heavy atom refinement it is easy to correct this scaling.)

2) The rms difference between Fph and Fp should be

$$\langle |F_{PH}-F_P|\rangle \approx \langle |F_H \cos\phi r|\rangle$$

when $\phi r$ is the random angle between vectors Fp and $F_{PH}$

and therefore

$$\text{r.m.s.}\ \langle |F_{PH}-F_P|\rangle \quad (=Diso) \approx \frac{\langle F_H\rangle}{\sqrt{2}}$$

This difference should fall off with resolution in the same way as the scattering curve for that heavy atom. But the effect of errors in data measurement will mean that Diso is consistently overestimated. This overestimate will be proportional to the expected error in both Fp and Fph, and becomes serious if

either data set is relatively inaccurate (ie weak data) or there are different systematic errors on the two data sets or there is very little heavy atom substitution (Dodson, 1975).

$R_{iso}$ is defined as $\langle F_{PH}-F_P\rangle/\langle F_P\rangle$

which means it should approximately equal $\frac{\langle F_H\rangle}{\sqrt{2}\langle F_P\rangle}$

So we would expect Riso to increase with resolution as f (h) / f (protein) does. (About 40% by 2.5Å resolution for a Hg.)

If the derivative is to be useful Riso should be more than Rmerge for each data set, but less than 63% - the theoretical value for a random match!

Non-isomorphism means that both Diso and Riso increase sharply with resolution.

3) Considering anomalous differences

$\langle|F_{PH}(+)-F_{PH}(-)|\rangle$ (=Dano) ≈ $\langle|2F_H''\sin\phi r|\rangle$

and therefore r.m.s. $\langle(F_{PH}(+)-F_{PH}(-)\rangle \approx \sqrt{2}\langle F_H''\rangle$

The ratio of $\frac{\langle F_H''\rangle}{\langle F_H\rangle} = f_H''/f_H$ which is known as Kemp

So we can test the accuracy of both Diso and Dano by inspecting

$\frac{\langle(F_{PH}-F_P)\rangle}{\langle F_{PH}(+)-F_{PH}(-)\rangle}$ (known as Kemp)

But the anomalous differences are relatively smaller than the isomorphous ones so the estimates are more seriously distorted by errors (Dodson,1975).

The three proteins I will use as illustrations are listed in Table 1.

A fragment of the Gyrase B protein has been solved by Dale Wigley and Gideon Davies; mucolipase by Zygmunt Derewenda, and Ribonuclease_Sa by Josef Sevcik (Wigley, 1991; Brady, 1991; Sevcik, 1990).

The crystallographic details are summarised in Table 1.

Table 1

| Compound | Cell dimensions | No. residues | Space group | Reactive groups | % solvent |
|---|---|---|---|---|---|
| Gyrase-b | 89.2 143.1 79.8 | 392 | C2221 | Cys(2) | 60 |
| Mucolipase | 71.6 75.0 55.0 | 269 | P212121 | Met(9) | 42 |
| Ribonuclease | 64.9 78.3 38.9 | 96 * 2 | P212121 | His(13) | 37 |

(Ribonuclease_Sa has two molecules in the asymmetric unit.)

## Data Collection

| Data | Method of collection | Resln | Rsym | Number of measurements | | | |
|---|---|---|---|---|---|---|---|
| | | | | Total | Unique | Completeness % | Ano. Data |
| Gyrase-b | | | | | | | |
| Native | film | 2.5Å | 7.9 | 40769 | 15725 | | No |
| Thimersal | film | 2.5Å | 5.7 | 42395 | 13596 | 86.3 | Yes |
| K2PtCl4 | diff. | 6.0Å | - | 1205 | 1205 | 75.3 | No |
| NaAuCl4 | diff. | 6.0Å | - | 1229 | 1229 | 85.5<br>87.2 | No |
| Mucolipase | | | | | | | |
| Native | Xentronics | 1.9Å | 8.1 | 132555 | 24101 | | No |
| Iodine | Xentronics | 3.1Å | 7.9 | 20077 | 5577 | | Yes |
| Ribo-nucleaseSa | | | | | | | |
| native | film | 1.8Å | 5.6 | 85496 | 17202 | 96.0 | Yes |
| K2PtCl4 | film | 2.5Å | 5.7 | 29631 | 6871 | 94.0 | Yes |
| Iodine | film | 2.5Å | 7.5 | 26846 | 6839 | 95.0 | |

film - rotation camera, diff. Rigaku diffractometer, Xentronics - area detector

## Heavy Atom Substitution

| Derivative | No. of sites | Reactive residue | Riso | Rcullis | Phasing power |
|---|---|---|---|---|---|
| Gyrase-b | | | | | |
| Thimersal | 2 full | 56Cys<br>268Cys | 21.9 | 54.0 | 1.75 1.21 |
| K2PtCl4 | 2 full | 166Met<br>297Cys | 26.9 | 64.7 | 1.39 |
| NaAuCl4 | 1 full | 200Glu | 20.1 | 84.5 | 1.20 |
| Mucolipase | | | | | |
| Iodine | 7 full | Tyr20,60<br>13,158 | 27.4 | 72.0 | 1.32 1.11 |
| Ribo-nuclease_Sa | | | | | |
| K2PtCl4 | 3 full<br>2 half | 53HisA<br>85HisA<br>53HisB<br>85HisB (conf. 1)<br>85HisB (conf. 2) | 24.1 | 48.1 | 2.1 |
| Iodine | 6? | 49TyrA?<br>54GluA?<br>85HisA?<br>59TyrB?<br>84AspB? | 28.7 | 82.0 | 0.80 0.63 |

Rcullis and Phasing power are taken from the program PHARE_ML, an extension of PHARE incorporating the maximum likelihood method described by Zbysek Otwinowsti in this book. The program is part of the CCP4 suite at Daresbury (Otwinowsti, 1991).

1). Gyrase B protein fragment

Various heavy atom derivatives were prepared for the gyrase which were targeted at substitution at the cysteine, Met or histidine residues. Those which changed the appearance of the crystals were pursued and the concentration of the heavy atom salts increased until the crystals showed signs of deterioration. Three derivatives were screened to 6A on the Rigaku diffractometer, and all showed useful and different substitution. Crystals were prepared for high resolution data collection using the synchrotron at Daresbury. The Wiggler disaster intervened which meant this synchrotron was out of action for some time, and after a nerve-wracking week while the crystals visibly decayed Roger Fourme who is in charge of scheduling the Lure synchrotron made time available. Data was collected from the survivors; the native and the thimersal derivative using 1.47Å wavelength radiation. The thimersal derivative crystals were carefully aligned to record the anomalous pairs on the same film; the native crystals were misset by 10 degrees to minimise the size of the blind region.

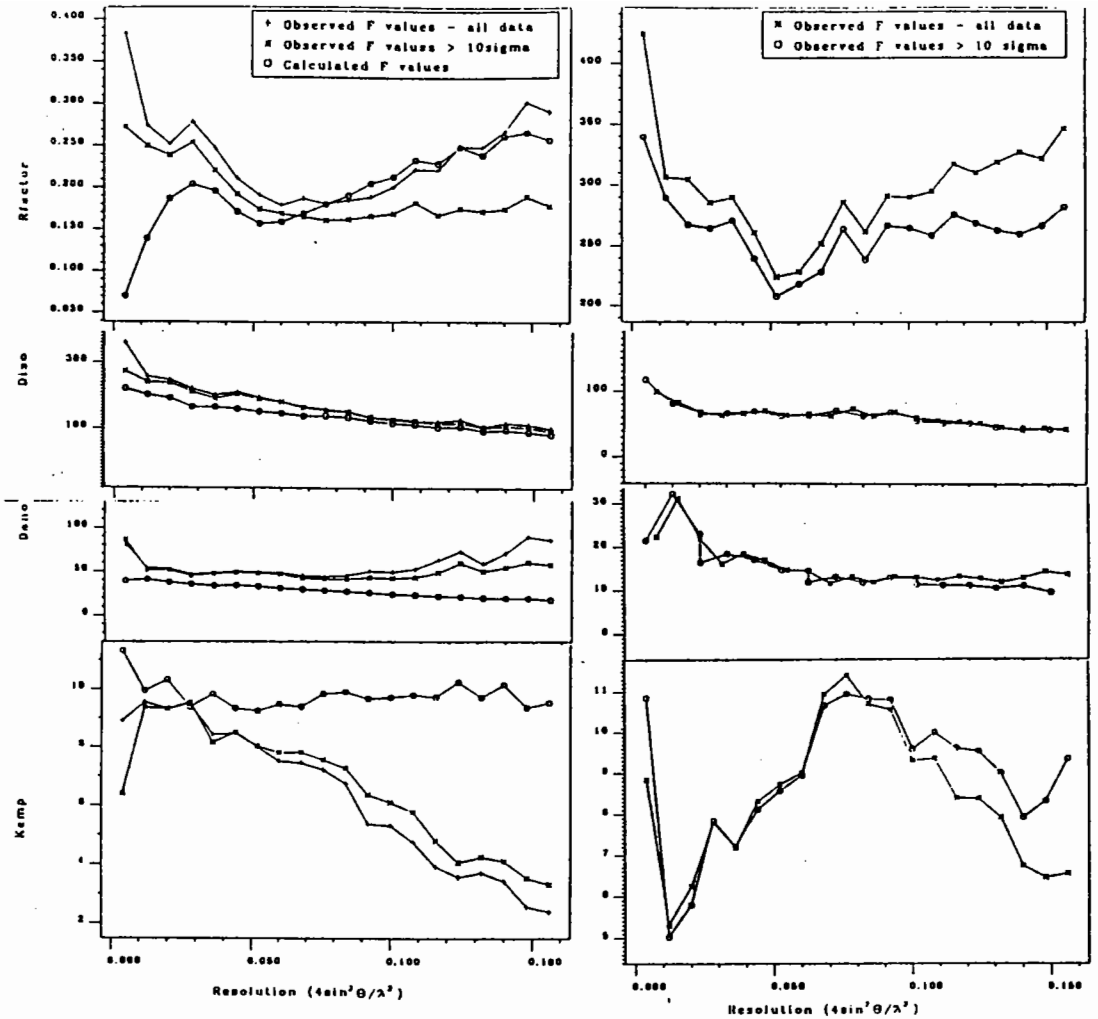
This data was processed and scaled, and Dale and Gideon calculated a map phased on the isomorphous and anomalous differences of the thimersol derivative to 2.5Å and using the 6Å data from the other two derivatives which had been screened on the diffractometer. This map was solvent flattened and proved to be interpretable.(?)

The protein structure has now been refined to an R factor of 17.8% against the 2.5Å native data. Phases from this model have been used to calculate maps to inspect the thimersal binding. The Hgs had reacted with the two cysteines as expected. The position of the 268Cys is virtually unchanged while there have been small perturbations in the neighbourhood of the 56Cys. Refinement of a model including HG against the thimersal data show both the residues 56Cys and 57Lys have moved over 1Å. The rms deviation for all the atoms is 0.192Å. We had hoped to be able to see details of the Hg ethyl , but in the 2.5Å maps the Hg peak is far too large to allow any detail of the ethyl group to be seen.

2). Mucolipase.

Mucolipase was solved by a combination of multiple isomorphous phasing, and molecular replacement. The first two derivatives had sites in special positions which gave very little phase information for some classes of reflections. Marek Brzozowski reacted N-iodo succinamide anhydride with the protein to produce a multi site iodine derivative which is very isomorphous to the native protein.N-bromo succinamide anhydride has been used for protein labelling and cleavage, and it was known to react preferentially at neutral pH with tryptophan at the CD position and tyrosine at the CE positions. At higher concentrations it will react with cysteine, arginine, histidine and other residues (Spande, 1970).

The compound is soluble, and easy to handle. The experiment was extremely successful. Mucolipase contains 4 tryptophans and 15 tyrosines. The distances between the iodine sites and the refined native protein coordinates are given in Table 2. The theoretical C-I bond length is 2.0Å so we felt these established that there is a high degree of isomorphism.




Gyrase B Ribonuclease-Sa

Riso = <FPH-FP>/<FP>; Diso = r.m.s. <|FPH-FP|>; Dano = <|FPH(+)-FPH(-)|>; Kemp = 2Diso/Dano
For this derivative I have calculated Fp, Fph(+) and Fph(-) from the refined coordinates and included statistics for these to give an "ideal" set. The tables are also given for all observed data, and data excluding all terms where an observation was less than 10 sigma. The over-estimation of Diso and Dano is not so bad when these terms are excluded, but they are still considerably greater than the theoretical values. This illustrates the effect of error bias. It also explains why Kemp falls off so much faster than theory would suggest. The errors have a worse effect proportionally on the smaller absolute values of Dano than on Diso.
Note: a) The theoretical R factor is distorted at low resolution by the huge overestimate of Fc which is typical for protein structure factors when the water molecules are excluded.

anomalous difference should exceed about 4 times the rms difference and this gives a useful criterion for excluding terms from Pattersons.

I feel the lessons to be learnt from this analysis are very old but still bear repeating.

1) Do careful and intelligent biochemistry to optimise substitution while preserving isomorphism.

2) Collect your data to an order of completeness and accuracy much greater than that required for refinement.

3) Screen your data with neurotic care!

John Kendrew is quoted as saying, "If I can't interpret my Pattersons in 10 minutes I throw them away and make another derivative." If this is not an option you can follow, take heart - many proteins have been solved with weakly substituted and unsatisfactory derivatives.

References:

Brady, R.L., Brzozowski, A.M., Derewenda, Z.S., Dodson, E.J., Dodson, G.G., Tolley, S,P., Turkenburg, J., Christiansen, L., Huge-Jensen, B., Norskov, L., Thim, L. and Menge, U. Nature, **343**, No 6260, pp 767-770 (1990).

Dodson, E.J. Proceedings of the Study Weekend held at Daresbury Laboratory (eds S.Bailey, E.Dodson and S.Phillips). pp73-87 (1988).

Dodson, E.J. Crystallographic Computing Techniques (ed F.R Ahmed), Munksgaard, pp 259-286 (1975).

Dodson, E.J., Evans, P.R. and French, S. Anomalous Scattering (eds S.Ramaseshan and S.C Abrahams) Munksgaard, pp 423-436 (1975).

Otwinowsti, Z. Proceedings of CCP4 Study Weekend (eds P Evans and A Leslie) (1991).

Sevcik, J., Dodson, G.G. and Dodson, E.J. Acta Cryst, **B47**, 240-253 (1991).

Spande, T.F., Witkop, B., Dogani, Y. and Patchornik, A. Advances in Protein Chemistry, **24**, pp1580-193 (1970).

Wigley, D.B., Davies, G.J., Dodson, E.J., Maxwell, A. and Dodson, G.G. Nature, **351**, 624-629 (1991).

Table 2

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| CE1 | 20TYR | I 6 | 2.36 | CD1 | 55TRP | I 3 | 2.12 |
| CE1 | 60TYR | I 7 | 2.12 | CE2 | 60TYR | I 5 | 1.86 |
| CE2 | 133TYR | I 4 | 2.30 | SG | 153CYS | I 2 | 1.90 |
| CE2 | 158TYR | I 1 | 2.12 | | | | |

3). Ribonuclease_Sa.

This protein was crystallised in Bratislava and brought to York by Josef Sevcik. He prepared heavy atom derivatives by diffusion methods using K2PtCl4, and I2,KI and C2H2O4. The crystals seemed robust, and the cell dimensions did not change significantly with heavy atom substitution. The phases were calculated using both anomalous and isomorphous measurements from both derivatives. The protein coordinates were fitted into the isomorphously phased and solvent flattened map.

The coordinates have been refined to an R factor of 17.2% against the 1.8Å native data. Phases calculated from these coordinates were used to generate maps to inspect the Pt and Iodine substitution. The Pt has reacted with all four histidine residues, but in every case the histidine residue has flipped, breaking hydrogen bonds from His NE2 to allow the Pt to react there. In addition 85HISB now has two conformations. This derivative has also been refined to some extent, and the r.m.s difference between the 2 sets of coordinates is 0.23.

Inspection of the iodine derivative showed a much messier pattern of substitution. We have not tried to model this carefully but the maps show extensive movement of the protein in the neighbourhood of TYR81 and HIS85 in both molecules.

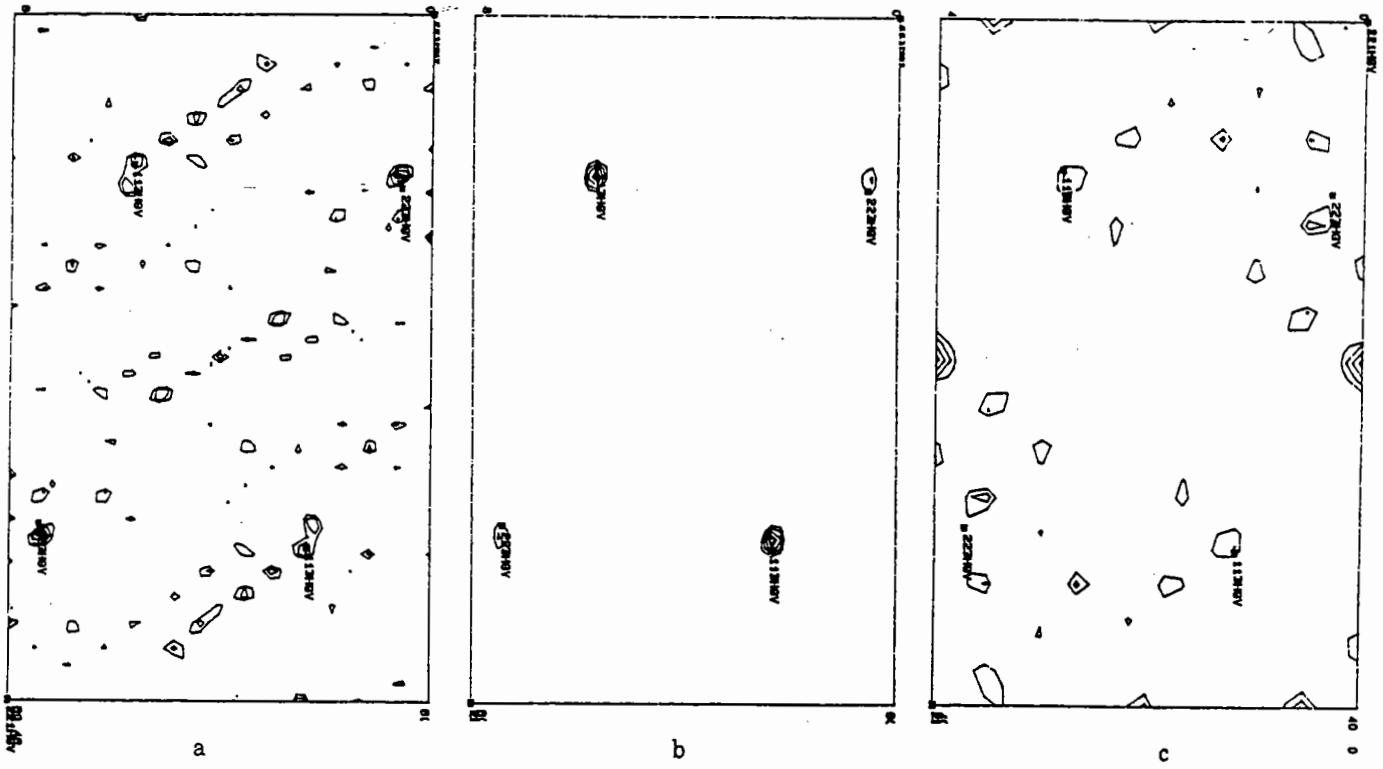
Previous phase analysis had already shown that the contribution of the iodine to the phase calculation was quite unimportant (Dodson, 1988).

It is easy to analyse derivatives after the structure is solved, but it would be much more useful to find clues to which derivatives will prove useful before hand.

Figure ? shows plots which help evaluate the derivatives. I have given details for the best and worst of the derivatives, Gyrase B thimersal, and the ribonucleaseSa iodine.

All this information is now included in the output of ANISOSC, a program for analysing and scaling data sets..

Since isomorphous phasing uses the isomorphous DIFFERENCES between 2 measured amplitudes these can be better (or worse) than the Rmerge for the data sets may suggest. If each measurement has the same systematic error; for example anomalous pairs being collected from the same film; native and derivative crystals mounted about the same axes, the results may be better than we have any right to expect. Of course all this information is encoded in the isomorphous and anomalous Pattersons. If both are clear and easy to solve it is reasonable to assume that the derivative will prove useful. If the isomorphous Patterson is clear but the anomalous one is not, the derivative is probably useful but the anomalous measurements may not be accurate enough to give reliable phases. This will produce <Dano> averages which increase with resolution. If the anomalous Patterson is clear but the isomorphous is not the derivative may not be truly isomorphous but it contains a precisely substituted heavy atom. (Such a derivative could be a good candidate for Multiple anomalous dispersion measurements.). However to get a good Patterson it is always necessary to exclude spurious differences. Ian Tickle has pointed out that it is unlikely that any isomorphous or




Gyrase - b Harker Sections

a. 2.5Å ($F_{PH(+)}$-$F_{PH(-)}$) Anomalous Difference Patterson

b. 2.5Å ($F_{PH}$-$F_H$) Isomorphous Difference Patterson

c. 6Å ($F_{PH(+)}$-$F_{PH(-)}$) Anomalous Difference Patterson

Gyrase B Analysis

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| + | 139 | 133 | 128 | 100 | 110 | 75 | 180* | 101 | 86 | 68 | 74 | 69 | 52 | 51 | 51 | 48 | 47 | 57 | 53 |
| * | 64 | 123 | 122 | 93 | 101 | 73 | 176* | 99 | 73 | 65 | 72 | 72 | 55 | 51 | 48 | 42 | 44 | 45 | 45 |
| o | 99 | 80 | 90 | 79 | 83 | 70 | 68 | 76 | 63 | 54 | 56 | 53 | 52 | 52 | 52 | 43 | 47 | 38 | 37 |

Maximum values of Diso in the above ranges.

Note the *!

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| + | 21 | 17 | 16 | 25 | 18 | 24 | 34 | 67* | 25 | 28 | 35 | 35 | 39 | 58 | 27 | 39 | 31 | 32 | 35 |
| * | 19 | 16 | 15 | 24 | 17 | 21 | 31 | 65* | 25 | 28 | 31 | 15 | 16 | 22 | 20 | 17 | 19 | 15 | 22 |
| o | 14 | 14 | 15 | 14 | 11 | 14 | 12 | 11 | 13 | 10 | 10 | 10 | 9 | 8 | 8 | 8 | 7 | 7 | 7 |

Maximum values of Dano in the above ranges.

Note the *. This single bad term causes the ripple visible in the Anomalous Patterson.

The values for all data are more seriously distorted than those when data < 10 sigma is excluded.

The statistics of intensity distribution (intensively studied for direct methods) suggests that Maximum F values are usually about 4 times RMS F. It is important to monitor terms larger than this and exclude them from Pattersons.

There is a very bad error in Diso and Dano in the 7th bin, due to a single crazy reflection. This term was not excluded from the Anomalous Patterson, and produces the ripples visible in it.

# Native non-isomorphism in the structure determination of Heat Labile Enterotoxin (LT) from *E. coli*

Titia K. Sixma, Sylvia E. Pronk, Anke C. Terwisscha van Scheltinga, Angel Aguirre*, Kor. H. Kalk, Gerrit Vriend‡, Wim G.J. Hol

BIOSON Research Institute, Dept of Chemistry, University of Groningen, Nijenborgh 16, 9747 AG Groningen, The Netherlands
*On leave from: Dept of Analytical and Physical Chemistry, University of Oviedo, 33006 Oviedo, Spain.
‡ Present address: EMBL Heidelberg, BRD.

In the case study presented in this paper the focal point is MIR structure determination when even the native crystals show individual differences. We will discuss successively the problem, a possible and an actual solution, and some potential explanations for the phenomenon, which can be derived in hindsight.

The object of our investigations is Heat Labile Enterotoxin (LT) from *E. coli*. This protein, a close relative of Cholera Toxin (> 80% sequence identity), is the causative agent of a severe diarrheal disease, which can be lethal to small children, and a major cause of traveller's diarrhea in developing countries. It has an $AB_5$ subunit structure with an enzymatic A subunit of 240 amino acids and 5 identical B subunits of 103 amino acids each, which are involved in membrane binding. Each subunit contains one internal disulfide bond (9-86 in B, 187-199 in A). For activity the A subunit needs to be proteolytically 'nicked' at position 192 or 194 and reduced into two separate subunits A1 and A2 (For references concerning LT and CT see refs. 1-3).

LT was overexpressed and purified from *E. coli* as described in 3). Crystals were grown from one batch of protein for 4 years reproducibly by liquid-liquid diffusion from 5 µl protein solution (9 mg/ml) containing 0.3-0.5 M KF in TEA (100 mM Tris, 1 mM EDTA and 0.02% azide at pH 7.50) with 5 µl TEA containing 3-15 % PEG 6000 as a second layer. All crystals were transferred to a standard mother liquor solution containing 0.175 M KF, 10% PEG 6000, in TEA. Crystals have space group $P2_12_12_1$ and cell dimensions a=119.2 Å, b=98.2 Å and c=64.8 Å. They usually diffract very well and generally give excellent data. A large number (72 different compounds) of heavy atoms has been checked by precession photographs and only a limited number of these showed obvious differences to native. Data sets, collected on a number of promising derivatives had generally good internal statistics (table 1), but no heavy atom sites could be identified. Extensive searches were done using difference Patterson inspection, direct methods as well as vector search methods. The latter were applied by correlating peaks in Harker sections as well as by using the local fivefold symmetry of the B subunits. This symmetry was known to be present in the data from a self rotation function and heavy atom sites on the B-subunits were expected to obey this symmetry.

Several native data sets were collected and it was noted that the native to native R-factors were quite high even after local scaling (table 2). These data sets were collected on different types of detectors and this fact could partially explain the high R-factors. However, upon careful inspection, some differences were also observed among native precession photographs, ranging from quite small to fairly large changes. Extensive checks showed that these were not due to different batches of PEG in the buffers nor to any other obvious difference in the crystallization conditions or mounting buffers, although there was an increase in the frequency and variation of the differences with the ageing of the protein.

Apparently the native crystals showed some sort of intrinsic non-isomorphism. Therefore all combinations of heavy atom derivatives and native data sets were examined, in an attempt to find optimal combinations with minimal protein non-isomorphism, but this did not lead to

Table 1: Data collected on LT from 1984-1989

| | | | X-ray source | Detector | Resol. (Å) | Rsym (%)* |
|---|---|---|---|---|---|---|
| a | Native § | A | Rot.An. | Film | 2.3 | 6.1 (F) |
| | $CdCl_2$ | | Rot.An. | Film | 2.3 | 7.1 (F) |
| | $NaReO_4$ | | Rot.An. | Film | 2.3 | 6.0 (F) |
| | Native | B | S.Tube | FAST | 2.3 | 3.8 (F) |
| | $HgCl_2$ | | S.Tube | FAST | 5.0 | 5.0 (F) |
| | $HgCl_2$ | | S.Tube | FAST | 3.3 | |
| | $K_2PtI_6$ | | S.Tube | FAST | 5.0 | 4.2 (F) |
| | $NaReO_4$ | | S.Tube | FAST | 5.0 | 3.7 (F) |
| | $CdCl_2$ | | S.Tube | FAST | 5.0 | 5.5 (F) |
| | $KAuCl_4$ | | Synch. | Film | 4.0 | 5.1 (I) |
| | $Na_2IrCl_6$ | | Synch. | Film | 4.0 | 4.4 (I) |
| | Native | C | S.Tube | CAD4 | 4.7 | 3.4 (F) |
| | $K_2PtCl_4$ | | Rot.An. | FAST | 3.8 | 2.8 (F) |
| | $HgCl_2$ | | Rot.An. | FAST | 3.4 | 5.0 (F) |
| | Merc.Chr. | | Rot.An. | FAST | 5.0 | 6.5 (F) |
| | Native | D | Rot.An. | FAST | 6.5 | 2.6 (I) |
| | $ErNO_3$ | | Rot.An. | FAST | 3.4 | 4.2 (I) |
| | $SmNO_3$ | | Rot.An. | FAST | 3.5 | 4.0 (I) |
| | $UO_2Ac_2$ | | Rot.An. | FAST | 3.5 | 4.7 (I) |
| b | Native | E | Rot.An. | FAST# | 2.5 | 4.2 (I) |
| | $K_2PtCl_4$ | | Rot.An. | FAST# | 3.8 | 3.8 (I) |
| | Native | F | Synch. | Image Plate | 3.2 | 7.5 (I) |
| | $KAuCl_4$ | | Synch. | Image Plate | 3.2 | 7.4 (I) |
| | Native | 1 | Rot.An. | FAST# | 3.1 | 3.4 (I) |
| | $K_2PtCl_4$ | | Rot.An. | FAST# | 3.1 | 4.9 (I) |
| | $HgCl_2$ | | Rot.An. | FAST# | 3.5 | 7.6 (I) |

a) Data collected on old batch of protein, no Patterson's solved.

b) Multiple data sets collected on one crystal. The procedure was used 3 times, and succesfull in all cases. The main disadvantage for further use was the large difference between the image plate and FAST native data sets. The relatively high image plate R-factor may be partially caused by the fact that this was one of the last crystals from the old batch of protein.

All synchrotron data were collected in Hamburg on the EMBL beamlines. Data collection took place at Cu Kα wavelength except for the image plate data take at 0.96 Å. Data processing: Film data with the Groningen version of the Munich film package (native A, $CdCl_2$, $NaReO_4$)(8,9) or the Purdue package, including profile fitting (native A and all other film data sets)(10); FAST data with Madnes (11) and those data indicated with # with the XDS profile fitting option (12). Image Plate with Mosco, adapted for image plates (13). Rot. An. Elliot GX21 rotating anode; S.Tube : Philips PW 2213/20 sealed tube. Image Plate: Hendricx Image Plate at Hamburg. CAD4: Enraf Nonius single crystal single counter diffractometer. FAST: Enraf Nonius TV area diffractometer.

$$* \; R_{sym} = \frac{\sum_{hkl}\sum_{j}\left|I_{hkl,j} - \bar{I}_{hkl}\right|}{\sum_{hkl}\sum_{j}\bar{I}_{hkl}}$$

§ $R_{sym}$ given for 2.3 Å data, processed with Munich Package. Reprocessing with Purdue package to 4.0 Å gave an $R_{sym}$ of 5.1%.

convincing heavy atom sites. A merged native data set, prepared with equal weights on the different data sets did not lead to a solution of the difference Pattersons either (table 2).

### Multiple data sets from a single crystal

Finally a solution to the problem of non-isomorphism was found by using a single crystal multiple times (table 1b). This procedure is a modification of that used by McKay and coworkers in a case where the crystals were severely twinned. To minimize the influence of twinning they cut the (very long) crystals in 2 pieces to allow collection of multiple data sets from the same initial crystal. The LT crystals were not sufficiently large to cut into pieces, but they are quite stable in the X-ray beam, especially at medium resolution.

Therefore a native data set was collected from one crystal, which was subsequently soaked in a heavy atom solution, remounted and exposed to collect a second data set of this same crystal. This procedure resulted in the first interpretable difference Pattersons, indicating that the crystal individuality had indeed been a major problem. The $K_2PtCl_4$ derivative that was tried with this procedure on the FAST slipped during the second data collection, but correct heavy atom sites could be found with a combination of direct methods and vector search. The two data sets collected in Hamburg from one single crystal gave a clear first $KAuCl_4$ site that could be found from three correlating Harker peaks within the 10 highest peaks of the difference Patterson, while the anomalous difference Patterson was also interpretable. When the procedure was repeated on the FAST even three reasonable quality data sets were obtained from one single crystal. The difference Pattersons of the third data set from this crystal could not be solved directly, but sites could be identified by difference Fourier techniques.

Table 2: Representative R-factors[#] between data sets.

| | | A Film | B FAST | C CAD4 | M9 merge | M5 merge | 2 FAST |
|---|---|---|---|---|---|---|---|
| B | FAST | 11.8 | - | 9.1 | .. | .. | .. |
| C | CAD4 | 12.4 | 9.1 | - | x | x | .. |
| D | Low | 9.7 | .. | .. | x | x | .. |
| E | FAST | .. | .. | .. | 7.6 | .. | .. |
| F | I.P. | 12.9 | 10.7 | 10.3 | 12.1 | .. | .. |
| Ir | Film | 8.7 | 13.6 | 11.9 | x | 8.4 | .. |
| Cd | Film | 16.5 | 18.1 | .. | 17.2 | .. | .. |
| Hg | FAST | 15.9 | 15.2 | .. | .. | 16.4 | 19.7 |
| Cd | FAST | 12.4 | .. | .. | 9.8 | .. | .. |
| Au | Film | 17.8 | .. | 21.0 | 19.7 | 20.5 | 17.8 |
| Pt | FAST | 13.8 | .. | 12.9 | 11.9 | 13.1 | 14.9 |
| Hg | FAST | 18.7 | 19.7 | 18.1 | 17.0 | 17.4 | .. |
| Er | FAST | .. | .. | .. | 18.8 | 21.7 | 21.8 |
| Sm | FAST | .. | .. | .. | 28.3 | 28.6 | .. |

These R-factors are the relevant R-factors, as used for Patterson and difference Fourier calculations, after local scaling and outlier rejection. R-factors on all data are generally 1-1.5% higher. In most cases various other strategies and parameters for the scaling procedures have also been tried. A-E & 2 are native data sets (see table 1&3). The $Na_2IrCl_6$ data have been used as a separate native data set, since precession photographs, taken after data collection did not show differences with native. M5 is a merged data set of natives A,C & D. M9 combines these three with the $Na_2IrCl_6$ data.

Data in the right hand column were calculated for the $F_{PH}$-$F_P$ difference Fouriers using phases from the model refined with respect to the 2.3 Å FAST data set labelled "native 2" in table 3.

$$\#\ R_{native} = \frac{\sum|F_2 - F_1|}{\sum F_1}$$

x means the data forms part of the merged set.

.. means that the R-factor has not been determined or saved.

Table 3: Phasing statistics of MIR analysis used for LT structure determination

| Derivative | Unique reflections | Complete (%) | Rsym * (%) | Rnative 1 # (%) | $R_{cullis}$‡ (%) | Phasing power§ |
|---|---|---|---|---|---|---|
| Native 1 | 11979 | 83.4 | 3.4 | - | | |
| Native 2 | 29769 | 85.8 | 6.2 | 6.2 | | |
| $K_2PtCl_4$ | 12282 | 85.6 | 4.6 | 14.7 | 69.8 | 1.45 |
| $KAuCl_4$ | 13177 | 91.8 | 3.0 | 14.5 | 75.8 | 1.33 |
| $K_2OsO_4$ | 13382 | 91.8 | 3.6 | 14.3 | 70.6 | 1.57 |

$$^{*}\ R_{sym} = \frac{\sum_{hkl}\sum_{j}\left|I_{hkl,j} - \bar{I}_{hkl}\right|}{\sum_{hkl}\sum_{j}\bar{I}_{hkl}} \qquad ^{\#}R_{native} = \frac{\sum|F_2 - F_1|}{\sum F_1} \qquad ^{\ddagger}R_{cullis} = \frac{\sum\left||F_{PH}| - |\bar{F}_P + \bar{F}_{H(calc)}|\right|}{\sum\left||F_{PH}| - |F_P|\right|}$$

$$^{\S}\ \text{Phasing Power} = \frac{\text{mean value of heavy atom structure amplitude}}{\text{residual lack of closing error}}$$

However, a new batch of protein was found solved the problems in a different manner since data sets collected on crystals from this fresh protein did not show the large individual differences of the older batch. A heavy atom data set and a native collected from different crystals now resulted in an easily interpretable Patterson. As can be seen in table 3, two native data sets collected on this new batch of protein have a mutual R-factor of 6.2%, indicating a reasonable degree of isomorphism and these crystals were used for the structure determination (1).

## Analysis of problems

To analyze the problems in the LT structure determination in more detail it is useful to make a distinction between different types of errors in the value of $|F_{PH}|$ which are crucial for the success or failure of the MIR procedure. For example:

$$F_{PH} = F_P + F_H + \Delta_P + \varepsilon_H + \delta \qquad (1)$$

Where:

$F_{PH}$ = structure factors of heavy atom derivative
$F_P$ = structure factors of native
$F_H$ = structure factors of heavy atom scatterer
$\Delta_P$ = non-isomorphism in protein
$\varepsilon_H$ = non-isomorphism due to heavy atom binding
$\delta$ = measurement errors

When $F_{PH}$ is described as above, it is clear that the importance of the errors is dependent on their relative size compared to the true heavy atom contribution. In the final, successful MIR analysis of LT the R-factor between derivatives and native data set was in the order of 14-15% (table 3), which is not very high and the number of sites was quite small for a protein with 755 amino acids. The final MIR map was not very good and the structure solution was only possible after extensive density averaging of the five B subunits (1).

## Native non-isomorphism

To analyse the changes in the native protein a comparison was made between the old (6 years) and the fresh batch of LT by gel electrophoresis. No significant changes were seen on SDS-PAGE gels, but on iso-electric focussing gels the differences were dramatic (fig 1). The fresh batch of protein showed a set of bands with a pI around pH=8.45, which agrees well with a calculated theoretical value for the AB5 complex (pI= 8.5). The old batch, however, showed two sets of bands, one around pH=6.3 and one at pH=9.5 (running off the gel shown in figure 1). These values agree well with the calculated values for the separate A (pI = 6.3) and B (pI=9.1) subunits. Under the conditions of the IEF experiment the complex falls apart, indicating that the dissociation constant between A and B has changed. Surprisingly the A2 subunit seems to remain bound to A1, since A1 alone is expected to have a higher pI at pH=7.0. It could be that the B5 pentamer also loses its integrity but since B monomers have the same pI as the pentamer this would not be visible on an IEF gel.

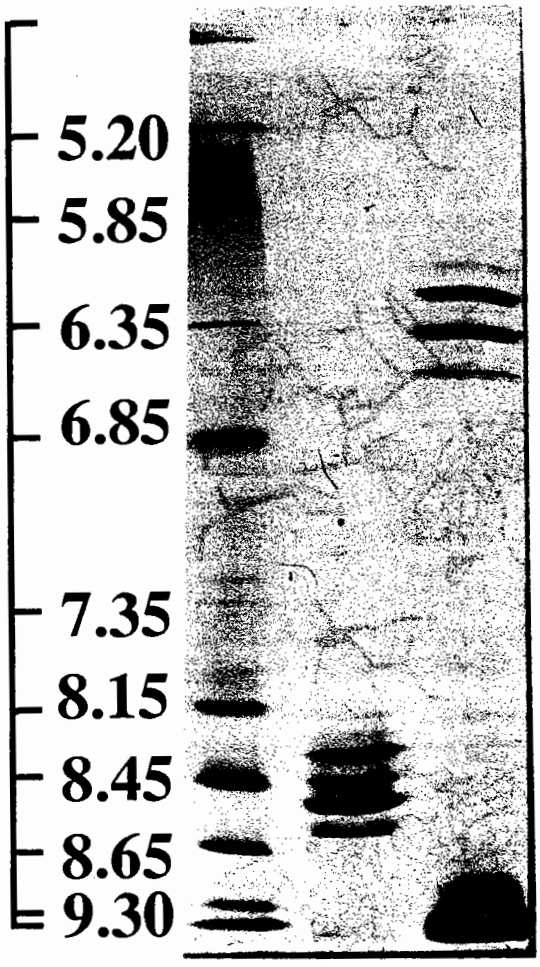



Fig 1. Iso electric focussing gel (pH:3-9) of LT old and fresh batches, run on Pharmacia PHAST system. Bands from left to right: 1) markers, 2) fresh LT, 3) aged LT.

The IEF gel also shows that both the old and the new batch have multiple bands with slightly varying pI, indicating a heterogeneity of charge. A possible cause for this is deamidation of asparagines. An isoelectro focussing titration curve of the protein seems to agree with this conclusion, since the heterogeneity in charge is not visible below pH=3. Such a deamidation has been observed in other proteins and was thought to prevent the crystallisation of Cholera Toxin (4). The LT crystals still grow when a limited heterogeneity is present such as in the 'new' batch of protein.

Another change in the native protein can be found from an analysis of $F_{obs}$-$F_{obs}$ Fouriers of old heavy atom derivatives ($HgCl_2$ and $K_2PtCl_4$, table 2) phased with the refined model phases. In these electron density maps the largest negative peak was at position A:Trp 174 (fig 2a,b). Apparently this Trp has variable positions in different data sets, although the density is good in native 1 (fig 2c). The position of this Trp is close to the junction of A1 and A2, a site of local disorder in the present model. It could be that the native non-isomorphism, the $\Delta_P$ in equation [1], may be related to different positions of Trp A:174. Possibly there is an intrinsic propensity of the A subunit for a conformational change after reduction of the disulfide bridge and nicking of a peptide bond around residue 193. There are, however, no hard facts available to support this suggestion.

To learn more details of the native non-isomorphism a refinement of the molecular model was carried out with the old native data set 'A' by TNT (5), which yielded an R-factor of 18.1% at a resolution of 2.3 Å. The overall rms shift was .32 Å for Cα and

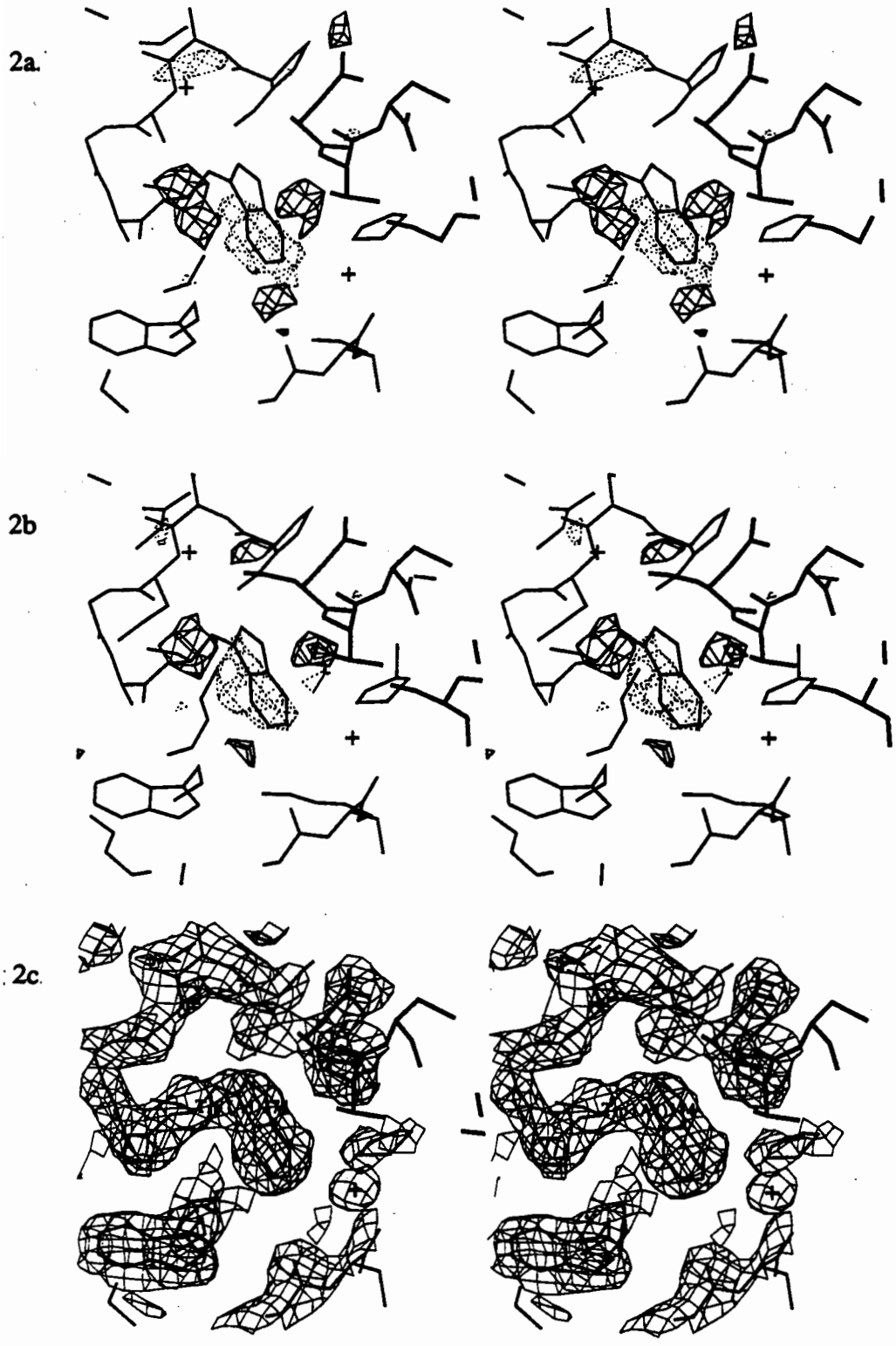



Fig 2. Trp 174 in the A1 subunit. A1 subunit in thin lines, A2 subunit in thick lines

a) {Fo (Pt)-Fo (native 2)} Fourier, phases and SIGMAA weights from Fo(native 2)-Fc for refined model. Negative density shown with dashed, positive with continuous lines, contoured at 4 sigma.

b) {Fo (Hg)-Fo (native 2)} Fourier, contoured as in a)

c) {2m|Fo(native 2)|-D|Fc|} electron density with SIGMAA weights of refined model, contoured at 1 sigma. It can be seen that a Thr in the A2 subunit (in thick lines) is not in density in this model, showing part of the area where the flexible A1/A2 junction is located.

.47 Å for all atoms. Shifts were larger for surface loops, but no concerted shifts of entire subunits, domains or folding units of the structure could be found. From this refinement no firm conclusions could be derived for the reasons of native non-isomorphism, but since this data set 'A' was derived from 7 crystals, it may be presenting a mixture of states. None of the other data sets was considered to be of sufficient resolution and quality to be used for refinement, and further knowledge about changes in the protein structure have to await more high resolution data sets.

## Characteristics of the derivatives

None of the three useful heavy atom derivatives conformed to the pentamer symmetry of the B subunits. The osmium and gold derivative contained only one ($K_2OsO_4$) or two ($KAuCl_4$) sites, located in the A subunit, while the $K_2PtCl_4$ derivative contains 4 out of 5 symmetry related sites plus one site in the A subunit. The occupancy of the fivefold related sites in the latter derivative varies widely by a factor of 4. This variation in occupancy may be caused by the oxidation of methionines. Fig 3 shows part of a Sigmaa (6) weighted $\{F_{obs}$(native A)-$F_c\}$ difference map near Met 37 in B#2. Clearly this methionine has been oxidized to a sulfoxide (7) in native A. Since B:Met 37 is one of the Pt binding sites such an oxidation has an effect on the occupancy of the Pt sites and may be the reason that an incomplete heavy atom pentamer was found. The apparent difference in oxidation rate of the methionines can, however, not be explained straightforwardly. There is no clear relation between the accessibility of the sulphur in the five different B subunits and either crystal contacts or interactions with the A subunit, but more subtle relations may exist and induce the observed variations.

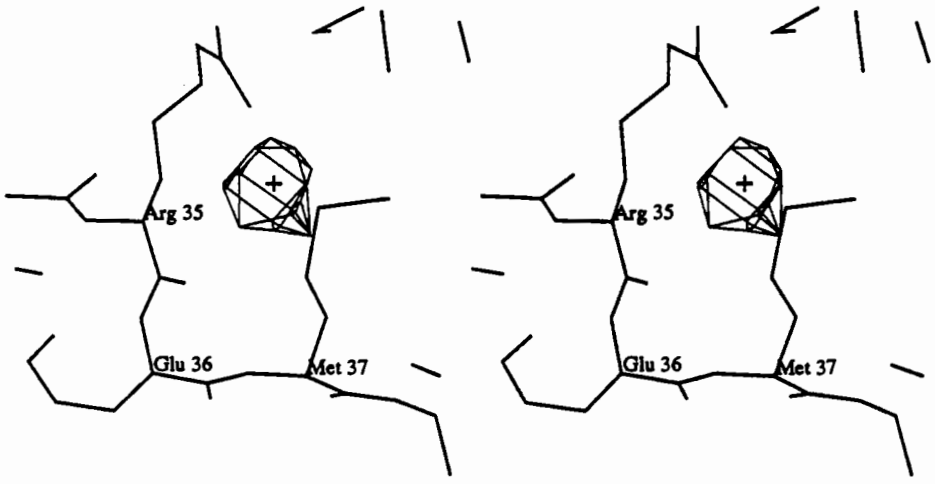



Fig 3. Oxidized methionine B:37 in difference Fourier {mIFo(A)I-DIFcI}, phases from refined 2.3 Å model, native data set 'A'. This is the fifth highest peak at 4.3 sigma, contoured at 3 sigma

Analysis by difference Fouriers (calculated with phases from the refined model) of some of the unused heavy atom derivatives (table 2) showed that although heavy metals were bound the occupancies were very low. In the $HgCl_2$ derivative three out of five equivalent sites were found. These sites are not close to any methionine, nor is there any clear correlation with the presence or absence of crystal contacts or A/B interaction.

Reexamination of the erbium and samarium derivative data sets revealed clear divalent and trivalent ion binding sites at functionally interesting locations, as will be described elsewhere. The erbium derivative has cell dimensions close to the native but some rigid-body movement of the subunits were found after refinement. The same is true for the samarium derivative which moreover has clearly different cell dimensions (a=118.8 Å, b=97.5 Å, c=65.1 Å). This

explains why these data had large native-derivative R-factors (table 2) and did not provide phase information.

## Concluding remarks

The structure solution of LT has long been delayed due to the lack of phase information. Not many heavy atom compounds bind to LT and those that did gave relatively small changes, in some cases possibly due to the oxidation of methionines. Therefore the observed native non-isomorphism presented large problems. Some possible changes that were seen in the protein are deamidation of asparagines, changes in the A/B interaction and some changes at the A1/A2 interaction site. The use of a fresh batch of protein provided the solution of these problems. This structure determination suggests that:

i) The use of DTT throughout might have lead to better platinum derivatives by avoiding methionine oxidation;

ii) When large batches of protein are available (and we were most generously provided with hundreds of milligrams of LT by prof B. Witholt and colleagues from the Biochemical Laboratory, University of Groningen) it is crucial to check protein quality and in particular the 'iso-electric purity' at regular intervals;

iii) The use of a single crystal for collection of a native plus one or more derivative data sets was shown to be a possible way of overcoming the non-isomorphism problems encountered in the structure determination of heat labile Enterotoxin and may be of more general use in other cases.

## Acknowledgements

We thank Hillie Groendijk, Ellen Wartna and Ben van Zanten for help in the crystallization and heavy atom derivative searches, Kyriakos Petrakos, Zbigniew Dauter, Christian Betzel and Keith Wilson for help during our stays at the EMBL in Hamburg, plus everyone from the Groningen protein crystallography group who came along on those trips for their help in the data collection. Finally Jaap Kingma and Bernard Witholt are gratefully acknowledged for providing us with the protein. This research was supported by the Netherlands Foundation for Chemical Research (SON) with financial aid from the Netherlands Organisation for Scientific Research (NWO) and by the Plan Regional de Investigacion de Asturias.

## Literature

1 Sixma T. K., Pronk S.E., Kalk K.H., Wartna E.S., van Zanten B.A.M.,Witholt B., and Hol W.G.J., Nature, (1991) in press
2 Sixma, T. K., Dauter Z., and Hol W.G.J., High resolution structure of heat labile enterotoxin at 1.95 Å., in preparation.
3 Pronk S. E., Hofstra H., Groendijk H., Kingma J., Swarte M.B.A., Dorner F., Drenth J., Hol W.G.J., Witholt B. , J. Biol. Chem., 260 (1985) 13580-13584
4 Spangler B.D. and Westbrook E.M., Biochemistry 28 (1989) 1333-1340
5 Tronrud D.E., Ten Eyck L.F. , Matthews B.W., Acta Cryst. A43 (1987) 489-501
6 Read R.J., Acta Cryst. A42 (1985) 140-149
7 Creighton T., "Proteins: structure and Molecular Properties", Freeman and Co. New York, (1984), p 20.
8 Schwager P., Bartels K., Jones A. J., Appl. Cryst. 8 (1975) 275-280
9 Wierenga R.K., Ph.D. Thesis, University of Groningen, The Netherlands (1978)
10 Rossmann M.G., Leslie A.G.W., Abdel-Meguid S.S, Tsukihara T., J. Appl. Cryst. 12 (1979) 570-581
11 Messerschmidt A., Pflugrath J.W., J. Appl. Cryst. 20 (1987) 306-315
12 Kabsch W., J. Appl. Cryst. 21 (1988) 916-924
13 Machin P.A., Wonacott A.J., Moss D., Daresbury Labnews 10 (1983) 3-9

# Phase determination using mercury derivatives of engineered cysteine mutants

Kiyoshi Nagai, Phil R. Evans, Jade Li & Christopher Oubridge

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, ENGLAND

## (1) Introduction

Green et al. (1954) first demonstrated that attachment of mercury atoms to haemoglobin crystals is sufficient to cause intensity changes in the diffracted X-rays from which phases can be calculated. Conventionally heavy atom derivatives are prepared by soaking crystals in solutions of heavy atom compounds, which penetrate into the crystals and bind to certain sites in the protein either covalently or ionically. For accurate determination of phases it is desirable to prepare several derivatives having heavy atoms bound to different sites. This is normally achieved by using heavy atom compounds of different chemical properties or changing concentration and soaking time. Because of problems with non-isomorphisms this is one of the rate-limiting steps in macromolecular structural detemination.

Techniques of site-directed mutagenesis allow any amino acid residues in the protein to be replaced by any other residues and one useful application of this technique to X-ray crystallography is the introduction of cysteine residues to which mercury atoms can be attached. This method is superior to the conventional soaking method since heavy atom can be directed to predetermined positions in the protein. Therefore derivatives with heavy atoms bound to completely independent sites can be prepared and this allows phases to be determined more accurately. Heavy atom sites can also be used as landmarks in chain tracing. This is particularly useful when the electron density map is of low quality or has some disordered regions.

Dao-Pin et al. (1987) were the first to report the attachment of mercury atoms to engineered cysteine residues in protein crystals and determined phasing parameters of some heavy atom derivatives. Stock et al. (1989) applied this method to the structural determination of the CheY protein. Ser is structurally anologous to Cys and the Ser→Cys mutation is therefore unlikely to cause large structural changes. Hence they introduced cysteine residues at each of the positions occupied by Ser in the wild type and attempted to crystallise these mutants. Three mutants gave only amorphous precipitate and one crystallised in a different crystal form. Only one mutant out of five crystallised in the form isomorphous to the wild type crystal, and soaking these crystals in a solution of ethylmercury chloride yielded a good heavy atom derivative. In the structural determination of γδ resolvase, Hatfull et al. (1989) made 13 cysteine mutants of which four crystallised in the form isomorpous to the wild type. Two good derivatives were obtained by soaking those in a solution of ethylmercury . They used the sequence alignment of eight related proteins and chose positions where the amino acid residue is poorly conserved and a charged residue is found at least in one of the related proteins. Such positions are likely to be on the surface of the proteins and amino acid replacement of these sites are unlikely to cause large structural changes.

Recently we have solved the structure of the U1 A small nuclear ribonucleoprotein using the methylmercury derivatives of four single cysteine mutants. In an attempt to prepare derivatives we made ten mutants of which six crystallised in the form isomorphous to the wild type. We report the structural environment of the heavy atom sites and speculate why some mutant crystals were isomorphous and others were not.

We describe some important points to be considered in choosing residues to be mutated to cysteine.

## (2) Crystal of the U1 A small nuclear ribonucleoprotein

The N-terminal domain of the U1 A ribonucleoprotein containing 95 residues was crystallised in the space group of I4132. A unit cell of this space group contains 48 asymmetric units and judging by the size of the unit cell (unit cell edge of 148Å) and the molecular weight of the protein (MW = 11kDal) each asymmetric unit was expected to contain at least two molecules. Interpretation of the Patterson map would be extremely difficult if more than one heavy atom were attached to each molecule. Since this protein does not contain any cysteine residues, preparation of single cysteine mutants by site-direct mutagenesis was the only way to ensure attachment of one heavy atom per molecule. This protein belongs to a family of over 20 proteins normally referred to as the RNP type RNA binding protein family (Bandzeulis et al. 1987). The method of Hatfull et al. (1987) seems very reasonable but the sequence alignment of proteins in the RNP family is not very reliable because of low overall sequence homology. Hence we considered only four sequences including the N- and C-terminal domains the U1 A protein and the closely related U2 B" protein (Fig 1). We chose mainly uncharged polar amino acids since these are likely to be on the surface of the protein and the replacement with uncharged cysteines is less likely to alter electrostatic interactions that could affect packing of crystal. We also mutated Ser, which is structurally similar to cysteine, and some charged amino residues.

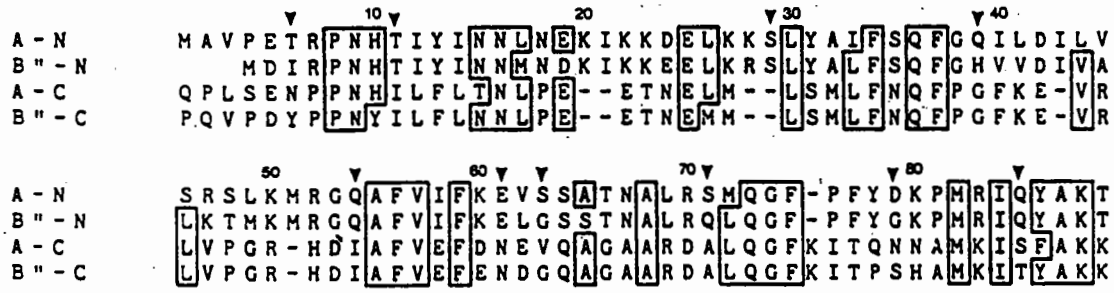



Fig 1. Amino acid sequence alignment of the N- and C-terminal RNP domains of the U1 A and U2 B" proteins (adapted from Sillekens et al.(1987). ∀ indicates positions where residues are mutated to cysteine.

## Methods

Methylmercury chloride was obtained from Strem Chemicals, Inc. Newburyport, MA 01950 USA (Cat. No. 125404-S). 3 gm of methylmercury chloride (12 mmoles, waxy flakes) were stirred with 1.7 gm silver nitrate (10 mmoles) in 100 ml water in a 250 ml round bottom flask with a ground glass stopper for 24 hrs in the dark at room temperature. After removing the precipitated silver chloride and unreacted methylmercury chloride using a 0.45 μm pore Nalgene filter, the faintly yellow solution of methylmercury nitrate was obtained. All steps were carried out in the fumehood wearing gloves.

Single cysteine mutants of the U1 A protein were prepared and treated with 5 mM dithiothreitol for one hour in 0.1 M phosphate buffer pH 6.9 at 4°C and gel-filtered on a Sephadex G-25 (fine) column against 50 mM phosphate buffer pH 7.0. The Sephadex column should be sufficiently large to ensure complete removal of DTT. Fractions containing protein were pooled and a 1.5 fold molar excess methylmercury nitrate was added immediately. After one hour of incubation on ice the solution was dialysed against 50 mM NaCl overnight to remove unreacted methylmercury. Crystallisation was carried out by microdialysis against approximately 1.7 M Na/K phosphate buffer between pH 5.0 and 5.5. Crystallisation of the unreacted form of the U1 A mutant

protein was carried out by microdialysis against degassed buffer containing 5 mM dithiothreitol. Cystals normally appear within a few days and grow to maximum size within a week. Since heavy atom-free crystals of the Ser-71→Cys mutant were larger than those of the methylmercury form, the heavy atom-free crystals were soaked in 2 M Na/K phosphate buffer pH 5.5 containing 0.25 mM methylmercury nitrate overnight and unreacted methylmercury was soaked out in the same buffer with no mercury.

X-ray data were collected on a Nonius FAST TV diffractometer using a rotating anode X-ray source. Crystals were rotated around the [110] axis so that Friedel pairs were measured simultanuously and therefore small anomalous differences can be determined accurately.

Table 1 Crystals of single cysteine mutants

| Mutant | Derivative | Space group (morphology) | Size (mm) |
|---|---|---|---|
| Ser-29→Cys | native | I4132 | 0.7x0.7x0.4 |
| | methylmercury | lozenge (birefringent) | 0.5x0.3x0.3 |
| Gln-39→Cys | native | I4132 | 0.6x0.6x0.3 |
| | methylmercury | I4132 | 0.6x0.6x0.3 |
| Gln-54→Cys | methylmercury | I4132 | 0.8x0.8x0.4 |
| Glu-61→Cys | native | I4132 | small |
| | pMB | I4132 | small |
| | methylmercury | I4132 | 0.6x0.6x0.3 |
| Ser-71→Cys | native | I4132 | 0.4x0.4x0.2 |
| | methylmercury | I4132 | 0.3x0.3x0.15 |
| Gln-85→Cys | native | I4132 | 0.8x0.8x0.4 |
| | methylmercury | I4132 | 1.0x1.0x0.5 |
| Thr-6→Cys | native | no precipitate | |
| | methylmercury | no precipitate | |
| Thr-11→Cys | native | no precipitate | |
| | methylmercury | birefringent crystals | 0.1x0.1x0.1 |
| Ser-63→Cys | methylmercury | amorphous precipitate | |
| Asp-79→Cys | native | amorphous precipitate | |
| | pMB | no precipitate | |
| | methylmercury | no precipitate | |

## Crystal forms of the mutants

No crystal was obtained for the Thr-6→Cys, Thr-11→Cys, Ser-63→Cys and Asp-79→Cys mutants (Table 1). These mutations affected the solubility of the protein substantially and addition of higher concentration of phosphate buffer yielded only amorphous precipitate. The methylmercury derivative of the Thr-11→Cys mutant produced small crystals. These crystals are birefringent and belong to a non-cubic space group. The native form of the Ser-29 mutant was crystallised in the I4132 space

group but crystals of the methylmercury derivative were birefringent and also morphologically different from the cubic crystals. Both the native and methylmercury derivatives of the Gln-39→Cys, Glu-61→Cys, Ser-71→Cys and Gln-85→Cys mutants were crystallised in the space group I4132. In addition to the methylmercury derivative we crystallised the p-hydroxylmercuribenzoate derivative of the Glu-61 mutants to mimic the negatively charged Glu side chain. The native crystal of the Ser-71 mutant was larger than that of the methylmercury derivative so we soaked the native crystals in 2.5 mM methylmercury solution over night.

**Phase determination**

We collected diffraction data of the methylmercury derivatives of four mutants: the Gln-39→Cys, Glu-61→Cys, Ser-71→Cys and Gln-85→Cys mutants. Since the heavy atom-free form of the Gln-85→Cys mutant was much larger than the wild type crystals, this was used as the native instead of the wild-type protein. The heavy atom positions of the Gln-39→Cys mutant was determined by the direct method option of ShelX-86 (Sheldrick, 1986). The solution for the Gln-39→Cys derivatives had two heavy atom positions per asymmetric unit and calculation of the Patterson vectors using these two heavy atom positions accounted for most Patterson peaks. Heavy atom coordinates (x, y, z) and its negative counterparts (-x, -y, -z) give rise to the the same isomorphous Patterson map and therefore phases were calculated for both sets of heavy atom coordinates including anomalous differences . The heavy atom positions of other derivatives were determined by difference Fourier synthesis using these two sets of phases. For all three derivatives only one set of phases gave prominent peaks corresponding to heavy atom positions, and the correct hand of the heavy-atom solution was unambiguously chosen. Multiple isomorphous replacement (MIR) phases were calculated using all four derivatives. The crystallographic data are summerised in Table 2. The MIR map was readily interpretable, and solvent flattening made only small improvement to the quality of the map (see Fig 3 in Nagai et al., 1990). Model building was carried out from both termini as well as four heavy atom sites using programme O (Jones et al., 1991).

Table 2 Crystal data

| Derivatives | Resolusion (Å) | No of unique reflections | Mean redundancy | $R_{merge}$ | No of heavy atom sites | $\langle F_H \rangle / E$ |
|---|---|---|---|---|---|---|
| native | 2.8 | 7,067 | 7.7 | 0.090 | - | - |
| S71C + MeHg | 3.5 | 3,509 | 5.0 | 0.101 | 4 | 1.7 |
| Q85C + MeHg | 2.8 | 6,913 | 8.0 | 0.082 | 4 | 1.2 |
| Q39C + MeHg | 2.8 | 6,674 | 7.2 | 0.077 | 2 | 1.4 |
| E61C + MeHg | 3.0 | 5.774 | 7.7 | 0.087 | 2 | 1.3 |

R merge = $\Sigma$ |Ij -<I>|/$\Sigma$<I> where Ij is an individual measurement of intensity and <I> is the mean intensity for this reflection.

<FH>/E, phasing power, (r.m.s. calculated FH)/(r.m.s. error), for centric reflections only

## Discussion

Mutations giving rise to non-isomorphous crystals or no crystals

(1) Thr-6→Cys mutation

The N-terminus of this protein is disordered and Thr-6 is the first residue which we could locate in the electron density map. From the structure it is not obvious why this mutant did not crystallise under the same condition.

(2) Thr-11→Cys mutation

The Thr-11 side chain does not seem to be involved in structurally important hydrogen bond nor direct subunit contact. Only amorphous precipitate was obtained for the heavy atom-free form and the methylmercury derivative crystallised in a different crystal form (birefringent).

(3) Asp-79→Cys mutation

This residue was chosen because charged Asp is likely to be on the surface of the protein and the residues at this position are variable in the related proteins (Fig. 1). This residue is located at the molecular contact between subunits related by the crystallographic dyad axis. As shown in Fig. 2 Asp-79 forms a salt-bridge with Arg-70 of an adjacent subunit. This salt-bridge must be an important crystal contacts since this mutant remains soluble even at much higher concentration of the precipitant.

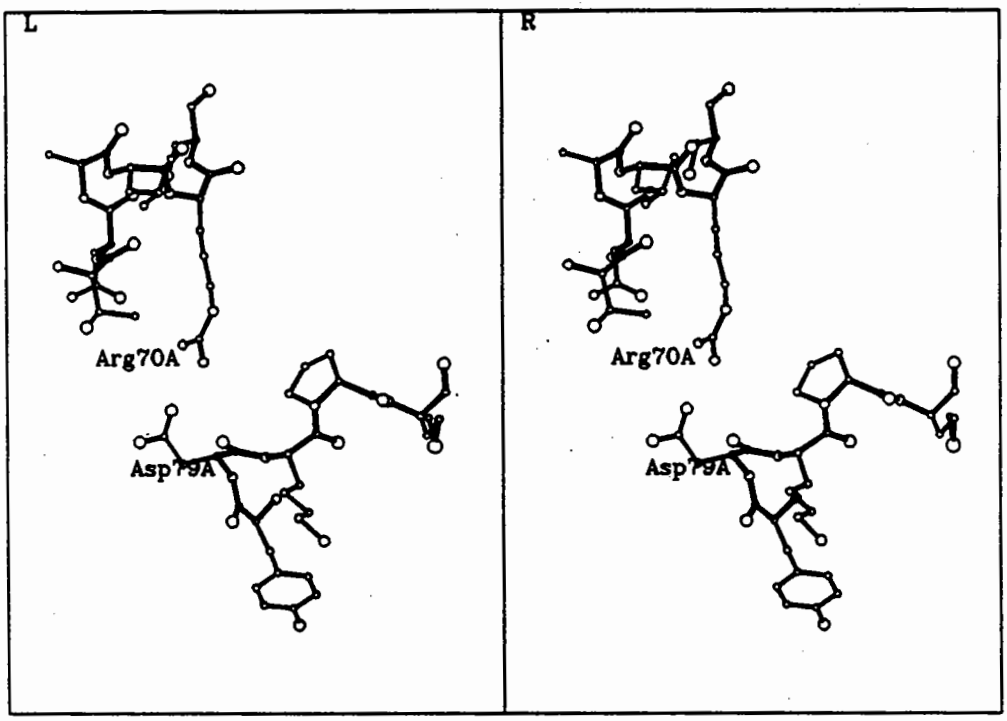



Fig. 2. Salt-bridge between Arg-70 and Asp-79 in the U1 A protein crystal

(4) Ser-63→Cys mutation

This residue is at the area of contact with an adjacent molecule related by a non-crystallographic dyad axis. Only amorphous precipitates were obtained for the methylmercury form of this mutant. The reason why this mutation prevented crystallisation is not obvious, but there may be water mediated hydrogen bonds with the adjacent subunit.

**Mutants crystallising in the isomorphous form.**

(1) Ser-29→Cys mutation

Ser-29 is in the region where two residues are deleted in the C-terminal RNP domains of the U1 A and U2 B" protein (Fig 1). As shown in Fig. 3 Ser-29 is part of the A helix and lies very close to the equivalent residue of an adjacent molecule related by a non-crystallographic dyad axis. The $O_\gamma$ atoms of these two residues are 3.7Å apart and these two Ser residues are unlikely to form a hydrogen bond. The replacement of this

residue with structurally similar Cys should not affect the structure significantly and in fact the native form was crystallised in the isomorphous form. There seems to be no room for methylmercury to be accommodated unless surrounding amino acids move away. Crystals of the methylmercury derivative are not isomorphous with those of the wild type crystal. If Cys-29 is accessible to methylmercury in the native crystal and the lattice forces are strong enough to keep the crystal intact upon binding of methylmercury to Cys-29, it might be possible to obatain an isomorphous crystal by soaking.

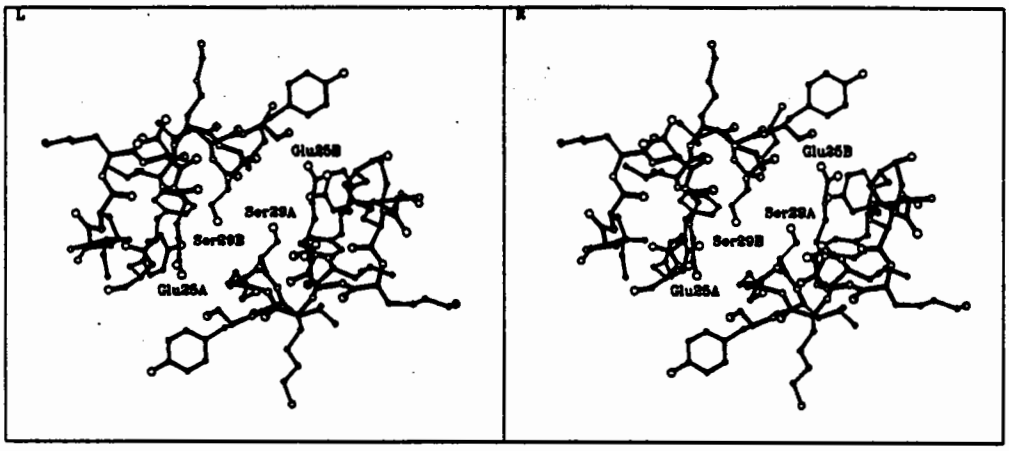



Fig. 3. Ser-29 at crystal contact near non-crystallographic dyad axis

(2) Gln-54→Cys mutation
This residue was chosen because it is hydrophilic and likely to be on the surface of the protein. Gln-54 is an external residue on the four stranded β sheet and forms hydrogen bond with phenol oxygen of Tyr-13 on the adjacent β strand. The methylmercury derivative of this mutant crystallised in the isomorphous form but we did not characterise these crystals crystallographically.
(3) Gln-39→Cys mutation
As shown in Fig. 4 Gln-39 is near a non-crystallographic dyad axis, but is not directly involved in crystal contacts. Although the movement of the methylmercury attached to Cys-39 does not seem to be restricted, there is only a unique mercury position at each site (Table 2). The heavy atom positions of this derivative were determined by the Shel-X86 programme (Sheldrick, 1985).
(4) Glu-61→Cys mutant
Glu-61 is at the crystal contact near a non-crystallographic dyad axis (Fig.4). Lys-60 and Glu-61 residues from two adjacent subunits are clustered around the non-crystallographic dyad axis. It is rather surprising that replacement of closely spaced negatively charged residues by a neutral residue near the crystal contact did not alter crystal packing. Methylmercury attached to Cys-61 has a unique conformation with the two mercury atoms attached to the two subunits only 4 Å apart.
(5) Ser-71→Cys mutant
Ser-71 is near a non-crystallographic dyad axis (Fig.5), but Ser-71 residues are not in direct contact with a neighbouring subunit. This subunit interface consists of many residues and the Ser-71→Cys mutation or further attachment of methylmercury is

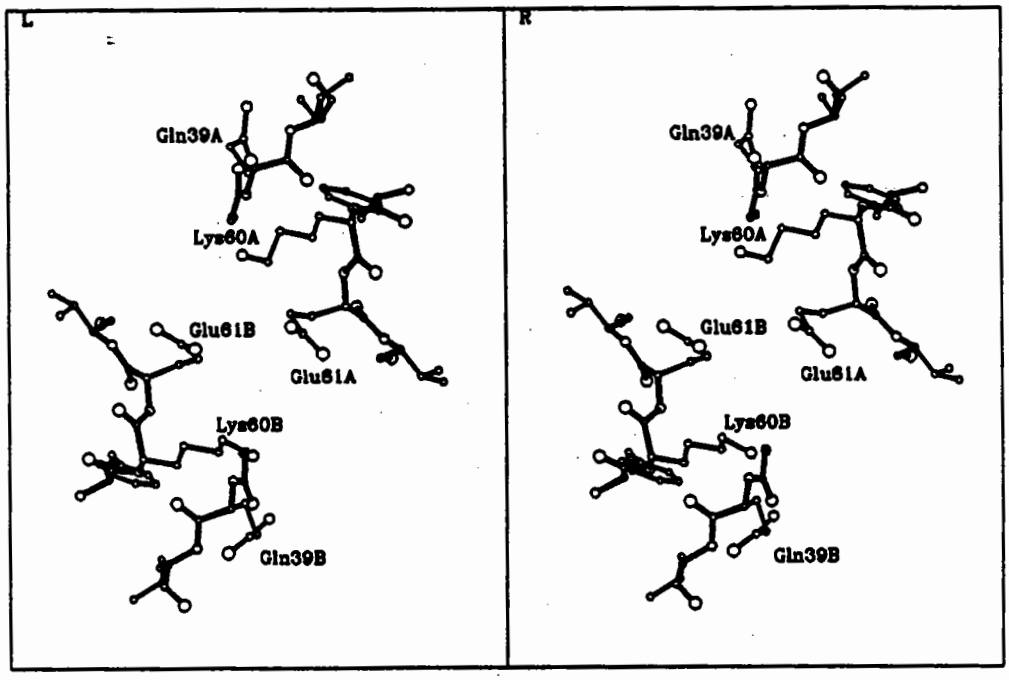



Fig. 4. Gln-39 and Glu-61 at crystal contact

unlikely to affect the subunit interface. Both the native and methylmercury forms were indeed crystallised in the isomorphous form to the wild type. Methylmercury attached to Cys-71 has two conformations with different occupancies.

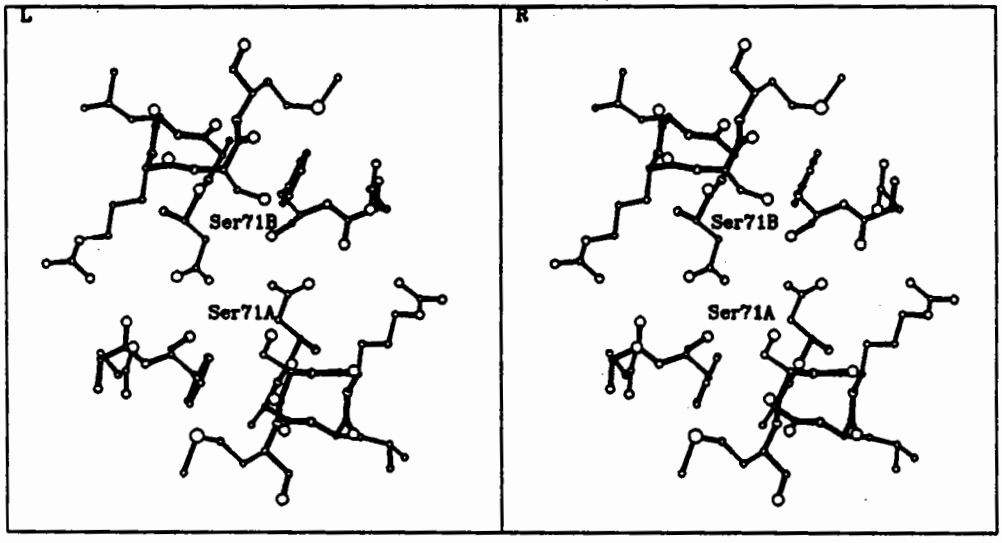



Fig. 5 Ser-71 at non-crystallographic dyad axis

(6) Gln-85→Cys mutation

Gln-85 is on the surface of the four-stranded β sheet and is not involved in subunit contacts. Both the native and methylmercury forms of this mutant were crystallised in the isomorphous form. Crystals of both forms are much larger than the wild type crystals. Methylmercury attached to this residue has two conformations with approximately equal occupancies.

## What residue to choose

Mutations in the protein interior often affect the structure and stability of proteins. We therefore mutated hydrophilic residues which are likely to be on the surface of the protein. We thought replacement of charged residues such as Arg, Lys, Glu or Asp with Cys might affect crystal packing via electrostatic interactions. The effect of mutations of charged residues on crystal packing may not be serious in high salt unless it is involved in inter-molecular salt-bridge. We have made several Hb mutants in which charged residues are replaced by neutral or by residues of opposite charge but most of them resulted in isomorphous crystals (Tame and Nagai, unpublished results). Gln is a uncharged hydrophilic residue and most of our Gln→Cys mutants gave good isomorphous derivatives. Ser and Cys are structurally similar and the Ser→Cys mutation is unlikely to perturb protein structure. However, when a Cys residue has reacted with mercury compounds it is no longer structurally similar to Ser, so there is no reason why Ser is the best residue to choose. Moreover Ser is often buried and forms a internal hydrogen bond, so the attachment of mercury compound can cause structural disturbances.

## Soaking or cocrystallisation?

Heavy atom derivatives can be prepared either by soaking crystals in a solution of heavy atom compounds or crystallisation of protein that has been reacted with heavy atom compounds. It is generally believed that the former is more likely to result in better isomorphous derivatives because lattice force prevent rearrangement of molecules within the crystal lattice upon attachment of heavy atom compounds (Blundell & Johnson, 1976). But this methods may results in poor incorporation of heavy atom compounds since restricted mobility of protein molecules by lattice forces in the crystal may limit the accessibility of heavy atom compounds to cysteine residues. Reaction of protein with heavy atom compounds in solution ensures good incorporation, but structural disturbance caused by attachment of heavy atom compounds may result in poor isomorphism, although crystallisation of the methylmercury forms of most U1 A cysteine mutants resulted in useful isomorphous derivatives (Table 1).

Evolutionary selection has probably eliminated cysteine residues on the surface of protein molecules which could form intermolecular disulphide bridges in non-reducing environment. Engineered cysteine residues have not undergone such evolutionary selection and therefore intermolecular disulphide-bridges may be formed even the presence of reducing reagents such as β-mercaptoethanol and dithiothreitol, and this may prevent crystallisation. It must be noted that these reducing reagents oxidize in the air and can promote disulphide-bridge formation. If crystallisation takes a long time it is essential to set up crystallisation under nitrogen. Co-crystallisation is preferable in this respect since irreversible attachment of heavy atom compounds will prevent disulphide-bridge formation during crystallisation. In the case of the U1 A protein crystals grew to maximum size within a week and good crystals were obtained for the unreacted form. It is therefore advantageous to crystallise Cys mutants both in the native and mercury forms.

## Acknowledgements

We thank Drs. Max Perutz, Andrew Leslie and Paul McLaughlin for critical reading of the manuscript, Dr. Timm Jessen for his contributions to the U1 A protein project. This work was supported by the MRC and grants from the HFSP and NIH.

## References


Dao-Pin, S., Alber, T., Bell, J. A., Weaver, L. & Mathews, B. W. (1987) **Protein Engineering 1**, 115-123

Hatfull, G. F., Sanderson, M. R., Freemont, P. S., Raccuis, P.R., Grindley, N. S. F. Grindley & Steitz, T. A. (1989) **J. Mol. Biol. 208**, 661-667

Stock, A. M., Mottonen, J. M., Stock, J. B., Schutt, C. E. (1989) **Nature 337**, 745-749

Sheldrick, G. M. in **Crystallographic Computing 3** (eds Sheldrick, G. M., Krüger, C. & Goddard, R.) 175-189 (Oxford University Press, Oxford, 1985)

Sillekens, P. T. G., Habets, W. J., Beijer, R. P. & van Venrooij, W. J. (1987) **EMBO J. 6** 3841-3848

Green,D. W., Ingram,V. M. and Perutz, M. F. (1954) **Proc. Roy. Soc., A225**, 287-307

Bandziulis, R., Swanson, M. S. & Dreyfuss, G. (1989) **Genes Dev. 3**, 431-437

Nagai, K., Oubridge, C, Jessen, T.-H., Li, J. & Evans, P. R. (1990) **Nature 348**, 515-520

Blundell, T. L. and Johnson, L. N. (1976) **Protein Crystallography**, Academic Press London,

Jones, T. A., Zou, J. Y. , Cowan, S. W. & Kjeldgaard, M. (1991) **Acta Crystallogr. A47**, 110-119 (1991)

# ESTABLISHMENT OF A HEAVY-ATOM DATABANK FOR PROTEIN STRUCTURES

David Carvin*, Suhail A. Islam+, Michael J. E. Sternberg+

and

Thomas L. Blundell*

*Department of Crystallography,
Birkbeck College,
Malet Street,
London WC1E 7HX

+Biomolecular Modelling Laboratory,
Imperial Cancer Research Fund Laboratories,
Lincoln Inns Feilds,
London WC2A 3PX

## ABSTRACT

The method of isomorphous replacement in protein crystallography involves the preparation of heavy-atom derivatives either by reaction of the protein with a heavy-atom followed by crystallisation or, more usually, by the reaction of the native crystals with a solution of the heavy atom reagent. 'Heavy-atoms' include iodine, second and third row transition metals, lanthanides and actinides, although lighter atoms have also been used successfully. A retrospective analysis of the sites of heavy-atom reaction with the crystalline protein provides valuable information on the chemistry of heavy-atom protein interactions. We have now brought these structural data together with the conditions of preparation in the form of a protein heavy-atom structure databank. This is a computer-based archival file system containing information on the preparation and interaction of heavy-atom compounds with crystalline proteins. The data include soak conditions, heavy-atom coordinates and literature citations. Arrangements for regular periodic updates and public distribution of the databank are under discussion. The data contain information of value not only to protein crystallographers in the preparation of heavy-atom derivatives for use in isomorphous replacement, but also to others with interests in heavy-atom interactions with proteins at the molecular level.

## INTRODUCTION

In order to reconstruct the image of the electron density of a crystalline molecule during X-ray analysis, it is necessary to determine both the amplitudes and phases of the diffracted waves. Although the amplitudes can be derived from the intensities of the diffraction pattern, the phases cannot be derived directly.

For most protein crystallographic analyses the phases have been estimated using the method of multiple isomorphous replacement [Green et. al., 1951] ; Blundell and Johnson, 1976]. Although multiple anomalous dispersion [Hendickson et. al., 1985] and molecular replacement [Rossman, 1972] will play an increasingly important role in X-ray analysis and NMR [Wuthrich et. al., 1984] will become applicable to larger proteins as new 3D and 4D methods [Bax et. al., 1991] are developed, it is likely that multiple isomorphous replacement will continue to play an important role in future work on the determination of protein structures.

There are two approaches to the preparation of heavy-atom derivatives of protein crystals [Blundell and Johnson, 1976]. A systematic approach is to carry out a specific chemical reaction with the

protein. The modified protein is then purified and characterised before crystallisation. The reaction sites are often sulphydryls of cysteinyl side chains, which provide good covalent links. The replacement of integral metal atoms, prosthetic groups or inhibitors with a labelled analogue also provides a useful route. Unfortunately, the preparation of specifically labelled proteins becomes increasingly difficult as the size of the protein increases. In any case there are few general methods. Furthermore, chemical modification of the protein often leads to crystallisation in a form that is not isomorphous with the crystals of the native protein. For these reasons, trial and error methods have proved more useful, although the heavy atom substitution patterns tend to be complex with sites frequently only partially occupied. Often the specificity is determined by entropic factors. Thus, where sites exist between molecules in the crystal lattice or between several different side chains brought together by the tertiary structure of the protein, side chains with no strong affinity for the metal may be involved in interactions. Such sites cannot be foreseen. Nevertheless, further knowledge of the conditions for optimal reaction, given knowledge of the sequence, is certainly a prerequisite for making some progress in producing better derivatives.

Over the past three decades a vast quantity of information has accumulated in the literature relating to heavy-atom compounds used for the method of isomorphous replacement. Blake [1968] reviewed the data available and suggested some generalisations. These were extended in a comprehensive review of protein heavy-atom derivatives [Blundell and Johnson, 1976 ; Blundell and Jenkins, 1977], which described a systematic analysis of the dependence of reactivity on the protein side chain identity, nature of the reagent, pH, concentration, buffer etc. However, since then the data available have increased many fold, but no comprehensive review has been reported, although there have been reports on some specialist areas. One consequence of this has been that information about protein heavy-atom interactions has not been available in a format that could be used for systematic computer-based analysis.

Over the past years we have collected, through analysis of the literature and by correspondence with protein crystallographers, information on the preparation and characteristion of heavy-atom derivatives of proteins. We have assembled the information in the form of a databank, in which the coordinate data for the heavy-atom positions is compatible with the crystallographic data in the Brookhaven Protein Databank [Bernstein et. al., 1977]. In this brief report we outline our approach. The databank contains a wealth of information which is providing the basis for further, more detailed analyses of heavy-atom binding to proteins. This will be published elsewhere.

## COMPILATION OF DATA

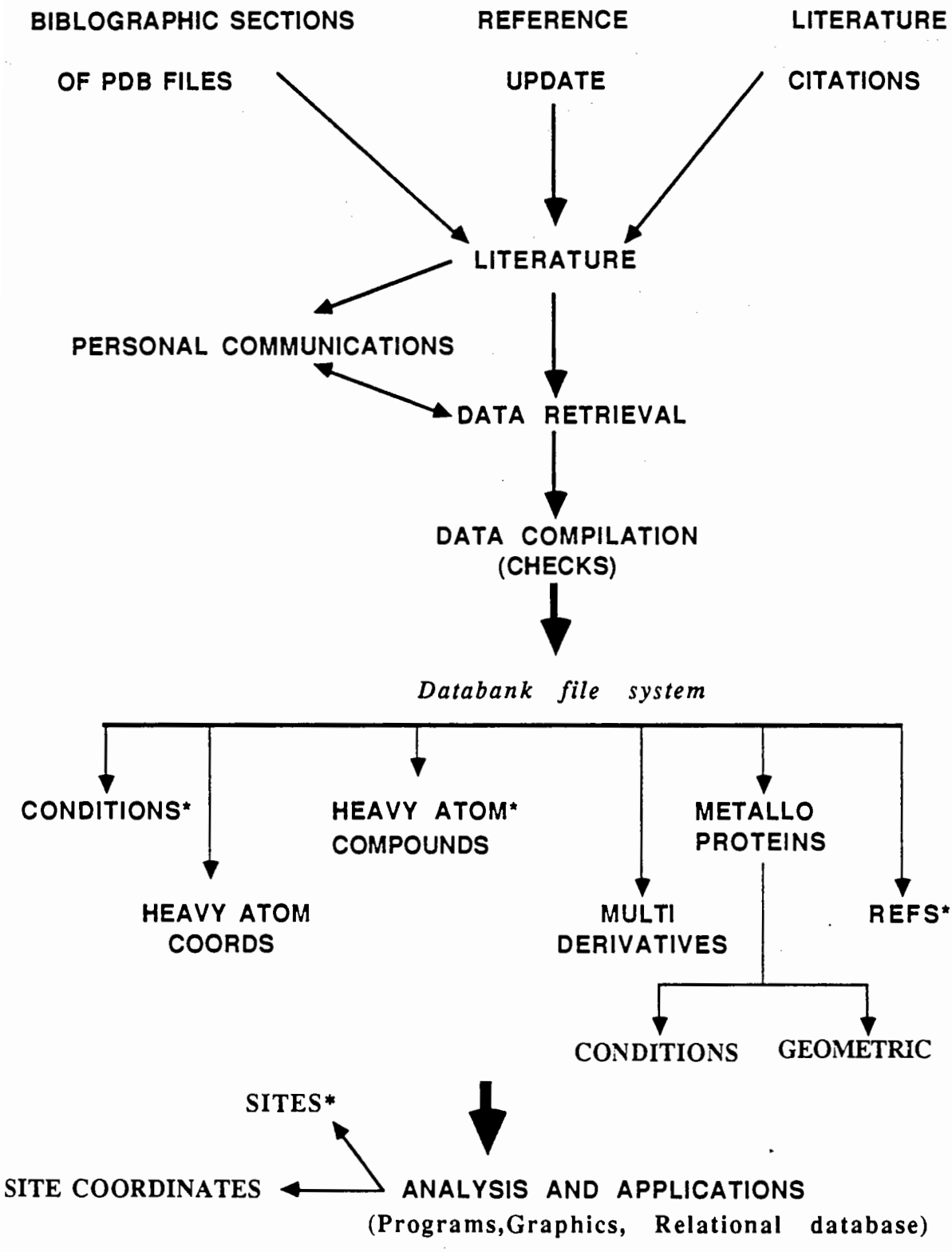
Some of the data was obtained from an exhaustive search of the literature. Although the positions of the heavy-atoms in the crystal unit cell are always determined during the method of isomorphous replacement, the atomic coordinates of the protein itself are determined at a much later stage. Unfortunately, many crystallographers do not make a retrospective analysis of the heavy-atom positions with respect to the protein structure and, of those who do, many do not publish the results of their study. As a result much of this is unpublished and it was necessary in most cases to supplement the published information by personal communications direct with the relevant authors. Thus, the comprehensiveness of the databank is dependent not only on information published, but also the willingness of the authors to obtain and supply details by personal communication. A list of authors who have kindly supplied us with unpublished information from their studies is given. We are very grateful to the many crystallographer's who responded so generously to our queries.

## HEAVY-ATOM DATABANK

The Heavy-Atom Databank is a computer-based archival file system. The purpose is to collect, collate, systematically format, deposit and distribute information concerning the interactions of heavy atoms with macromolecular assemblies. The information relates not only to proteins whose atomic coordinates have been deposited in the Brookhaven Protein Databank, but also to other proteins whose structures have yet to be deposited. The databank consists of individual file systems where each file system is categorised according to the nature of the deposited information that it holds. The general scheme for the collation and catagorisation of the databank is shown in Figure 1. The main file system categories are:

1] **CONDITIONS** (details of heavy-atom soak conditions)
2] **HEAVY ATOM COORDINATES** (atomic coordinates of heavy atoms)
3] **HEAVY ATOM COMPOUNDS** (chemical details of compounds used).
4] **REFS** (bibliographic references)
5] **MULTIDERIVATIVES** (heavy-atom soak conditions involving several heavy-atom compounds to form a single heavy-atom-protein complex)
6] **METALLO PROTEINS** (geometrical details of native metal-atom binding sites)
7] **SITES** (geometrical details of heavy atom sites)

# FIGURE 1
# PROCEDURE FOR COMPILATION OF DATA




* This indicates that the data are also stored within the ORACLE relational database

8] **SITE COORDINATES** (atomic coordinates for entire binding site i.e. protein residues making contact with heavy atom)

File systems 1-6 essentially contain raw data whereas 7-8 comprises data derived from our analysis of the heavy atom binding sites. A brief description of the general contents of file systems 1-6 and and an example of the file formatting are described below.

## DATA FILE SYSTEMS

**Conditions Data File** contains conditions for preparation of heavy-atoms derivatives, information on the composition and concentration of the heavy-atom solution used in the experiments. This includes details of the chemical compound, precipitant, buffer, additives, pH, time of soak and source of protein. Additional techniques employed (i.e. variations in temperature, stabilisation of the crystal by cross linking, mutagenesis of the primary structure) are shown, as are the side chains of the protein involved at each heavy-atom binding site.

```
HEAVY     ATOM CONDITIONS DATA
REFS      20.00
BCOD      2AZA
PNAM      AZURIN (ALCALIGENES DENITRIFICANS)
 PNAM
HCOD      GHB
HCON      1.74
PPPT      AMMONIUM SULPHATE
 PPPT
PCON      -8.,75.,
BUFF      PHOSPHATE
 BUFF
BCON      0.10000
 ADDT
 ADDT
ACON      0.00000
SOAK      504.00000
SPEC
PHPH      6.00000
TMAU      +3,2 MOLECULES
BIND      1A0.41/RES 42 CO:TRP 118 +3:
BIND      1B
BIND      2A0.11/RES 42 CO +3:TRP 118:
BIND      2B
BIND      3A0.06/RES 37 CO +3:TRP 118:
BIND      3B
```

**Heavy-Atom Coordinates Data File** contains the atomic coordinates and associated data as derived from the primary literature or as provided by personal communications.

```
HEADER     2AZA00 ELECTRON TRANSPORT PROTEIN(CUPROPROTEIN)
COMPND     AZURIN (OXIDIZED)
REMARK     HEAVY-ATOM COORDINATES ORTHOGONALISED USING
REMARK     PROTEIN SCALE MATRIX
CRYST1    75.000    74.200    99.500  90.00  90.00  90.00 C 2 2 21       16
HET    GHB              POTASSIUM DICYANO AURATE (I)
FORMUL   1  GHB      K1,AU1,C2,N2
FORMUL   2  GHB      AU1,[C1,N1]2 -1
HET    UXC              URANIUM (VI) OXYACETATE
FORMUL   1  UXC
FORMUL   2  UXC      [U1,O2]1 +2
HET    HDU              DICHLORODIAMINO MERCURATE (II)
FORMUL   1  HDU      HG1,N2,H6,CL2
FORMUL   2  HDU      HG1,CL1 +1 : HG1,CL2 -0
HET    HAZ              THIOMERSAL, ETHYL MERCURY THIOSALICYLA
FORMUL   1  HAZ      C9,H9,O2,NA1,S1,HG1
FORMUL    2   HAZ       [C6,H4]1,[C1,O2]1,[(S1),(HG1),(C1,H2),(C1,H3)]1  -
SCALE1       .013333  0.000000  0.000000        0.00000
SCALE2      0.000000   .013477  0.000000        0.00000
SCALE3      0.000000  0.000000   .010050        0.00000
HETATM    1 AU1  GHB     1      10.875   9.794   5.274
HETATM    2 AU2  GHB     2      -6.450   8.904  24.677
HETATM    3 HG1  HAZ     3      -1.350  11.872   9.453
HETATM    4 HG1  HDU     4       1.425  11.798  11.443
```

Heavy-Atom Compound Data File containing physical and chemical characteristics of each chemical compound that has proved successful in past protein crystallographic analyses. This includes the IUPAC name, trivial name, molecular formula, oxidation state, solution chemistry and stereochemistry. To assist analysis an "in house" three character alphabet code was developed to designate the heavy atom compound (i.e. PEN = $K_2PtCl_4$).

```
HCOD      PEN
CNAM      POTASSIUM TETRACHLORO PLATINATE (II)
TNAM      POTASSIUM CHLOROPLATINITE
FORM      K2,PT1,CL4
OXYN      2  D8
STER      SPL
SOLC      PT1,CL4  -2
INFO
COOR      4
```

Reference Data File contains literature citations :- author(s), title, journal name, year of publication, volume number, first and last page number.

```
REFS      20.0
BCOD      2AZA
AUT1      NORRIS, G.E., ANDERSON, B.F., AND
AUT2      BAKER, E.N.
TIT1      STRUCTURE OF AZURIN FROM ALCALIGENES
```

```
TIT2        DENITRIFICANS AT 2.5    RESOLUTION.
JOUR        J. MOL. BIOL. (1983) 165, 501 - 521
YEAR        1983
VOLU        165
PAGF        501
PAGT        521
COMM        !
```

**Multiderivative Data File** includes details of the composition and concentration of the two or more heavy-atom solutions used in making double and more complex derivatives.

```
REFS        1076.00
BCOD        3GRS
PNAM        GLUTATHIONE REDUCTASE (OXIDISED, HUMAN)
HCOD        GHG
HCON        2.0
SOAK        370.0
HCOD        HAB
HCON        0.05
SOAK        19.2
PPPT        AMMONIUM SULPHATE
PCON        2.0
BUFF        DIPOTASSIUM HYDROGEN PHOSPHATE
BCON        0.1
ADDT
ACON
PHPH        7.0
SPEC
TMAU        2 IDENTICAL SUBUNITS PER MOLE: 1/2 MOLE IN AU +64:
BIND        1 A
BIND        1 B
```

**Metalloprotein Data File** is divided into two subsections, namely, conditions and geometry. The conditions file contains details of type, quantity, geometry and function of the native metal(s) present. The procedure for native-metal substitution, the composition and concentration of the derivatising solution. The interatomic distances and angles between the substituted heavy-atom and protein environment ligands are recorded.

```
REFS        210.1
BCOD        3PCY
PNAM        PLASTOCYANIN (MERCURY SUBSTITUTED)
NMET        COPPER
OMET        NONE
FUNC        MECHANISTIC (ELECTRON TRANSFER)
GEOM        DISTORTED TETRAHEDRAL
PLIG        HIS 37:HIS 87:CYS 84:MET 92:
PLIG
ASSC        ?
```

```
DSSC        ?
SUBS        YES
PROS        CRYSTALLINE
MEBA        NONE
MEBC        MERCURIC ACETATE
HCOD        HDP
HCON        50.0
PPPT        AMMONIUM SULPHATE
PCON        3.3
BUFF        SODIUM PHOSPHATE
BCON        0.1
ADDT
ACON
SOAK        23.0
SPEC        DIRECT DIFFUSSION:RETURNED TO FRESH STABILISER
SPEC        SOLUTION TO REMOVE EXCESS MERCURY  REAGENT:
SPEC        SLIGHT ENLARGEMENT COORDS POLYHEDRON TO
SPEC        ACCOMMODATE MERCURY ATOM.
PHPH        6.0
TMAU        MONOMER
PD01        HG - HIS 37 (ND) - 2.34:
PD02        DOUBLE DERIVATIVE DATAHG - HIS 87 (ND) - 2.36:
  .
PD10
PA01        HIS 37 (ND) - HG - HIS 87 (ND) - 100:
PA02        HIS 37 (ND) - HG - CYS 84 (SG) - 133:
  .
PA10
RA01        CHURCH, W.B., GUSS, J.M., POTTER, J.J., AND FREEMAN,
RA01        H.C.
RJ01        J. BIOL. CHEM. (1986) 261, 234 - 237.
```

The second metalloprotein file describes the geometry of coordination and includes interatomic distances and angles between the native metal and its protein ligands.

```
REFS        210.0
BCOD        1PCY
PNAM        PLASTOCYANIN (POPULUS NIGRA)
NMET        COPPER
OMET        NONE
FUNC        MECHANISTIC (ELECTRON TRANSFER)
GEOM        DISTORTED TETRAHEDRAL
PD01        CU - HIS 37 (ND) - 2.04:
PD02        CU - HIS 87 (ND) - 2.10:
PD03        CU - CYS 84 (SG) - 2.13:
PD04        CU - MET 92 (SD) - 2.90:
PA01        HIS 37 (ND) - CU - HIS 87 (ND) - 97:
PA02        HIS 37 (ND) - CU - CYS 84 (SG) - 132:
PA03        HIS 37 (ND) - CU - MET 92 (SD) - 85:
PA04        HIS 87 (ND) - CU - CYS 84 (SG) - 123:
PA05        HIS 87 (ND) - CU - MET 92 (SD) - 103:
PA06        CYS 84 (ND) - CU - MET 92 (SD) - 108:
DP01        CU - HIS 37 (ND) - CYS 84 (SG) - MET 92 (SD) - 0.72:
DP02        CU - HIS 37 (ND) - CYS 84 (SG) - HIS 87 (ND) - 0.34:
RA01        CHURCH, W.B., GUSS, J.M., POTTER, J.J., AND
```

RA01 FREEMAN H.C.
RJ01 J. BIOL. CHEM. (1986) 261, 234 - 237.

## CONTENTS OF THE DATA BANK

The quantity of data available and its computer requirements as on 31st December 1990 are summarised below. Information is available for 375 distinct proteins, of which 174 are deposited in the Brookhaven Protein Databank. Thirty-three of these proteins could not be used in the full analysis due to errors present in their respective SCALE matrix data.

Approximately 8% of the 5,500 heavy-atom coordinates were derived from multiderivative 'soaks'. The heavy-atom were obtained from 274 different chemical compounds or iodine, configured into simple inorganic salts, cluster compounds, or organometallic compounds.

Heavy-atoms can interact with proteins in several distinct chemical forms, for example, as pure ions (i.e. $Cd^{2+}$) or as coordinated compounds (i.e. $PtCl_4^{2-}$). Except for a few iodomercury compounds (i.e. $K_2HgI_4$) only the coordinates of the heavy-atom are available. This requires the remaining ligand(s) to be 'built-in' before an analysis of the binding site can be attempted. The Cambridge Crystallographic Databank [Allen et al., 1983] is used to obtain structural information for all the heavy-atom compounds.

**Summary of the contents of the heavy-atom data base as on 31st December 1990.**

| | | Computer requirements (Mb) |
|---|---|---|
| No. of distinct proteins | 374 | - |
| Total No. of heavy-atom sites | >5,500 | - |
| No of 'bad' proteins | 33 | - |
| Total No. of heavy-atom sites which have had protein coordinates deposited in the PDB | >2,500 | 0.3 |
| No. of Multiderivatives | 43 | - |
| Total No. of chemical comounds | 274 | 0.1 |
| No. of different elements | 37 | - |
| No. of metalloproteins | 137 | - |
| No. of soak conditions files | >1,300 | 0.6 |
| No. of references | >850 | 0.6 |
| Geometric information on sites | - | 3.5 |
| Coordinates for general sites | - | 8.0 |
| Total computer requirement | - | 13.0 |

## ANALYSIS

Computer programs have been developed to create, check and analyse the Databank. In addition the relational database ORACLE [ORACLE corporation] is being used to aid analysis. The principal programs carried out the following:

(a) creation, maintenance and check of the Databank

(b) generation of the heavy-atom environment i.e. atomic coordinates for the protein and solvent interacting with the heavy-atom (using predefined criteria for interatomic interactions). This is done using symmetry operators so that the heavy-atom coordinates are appropriate to the asymmetric unit of the crystallographic cell used and all interactions are identified for the protein coordinates deposited in the Brookhaven Protein Databank.

(c) generation of coordinates of the non heavy-atoms of the heavy-atom agents used where they are not available from the X-ray analysis. Often the coordinates for a complete ligand can be obtained from the Cambridge Small Molecule Databank. In many cases, however, information on the orientation or configuration of a reagent had to be inferred from the protein environment.

(d) preparation of data suitable for generation of relational database tables. A number of the file systems [see Figure 1] have been tabulated and placed in the relational database, Oracle [ORACLE corporation]. The tabulated data can be made suitable for incorporation into most database systems.

Molecular graphics are also used to check results. All the heavy-atom binding sites have been generated and analysis is at an advanced stage. The results will be published elsewhere.

All data and programs reside on a DEC VAX 8700. Molecular graphics is performed on an IRIS 4D/70 GT using the Quanta software [Polygen corporation].

## CONCLUSION

The establishment of the Databank for protein heavy-atom interactions was made possible by a rigorous search of the literature and by many personal communications. Even so the data are rather incomplete in many places. This is presumably because the main

objective of the protein crystallographer is to solve the three-dimensional structure and understand the function. Unfortunately isomorphous replacement is just a means of re-establishing the lost phases and is considered of little interest by most crystallographer's once the "model" is established. A consequence of this is that there is no requirement by journals to provide information about the heavy-atom sites. We suggest there should be. It would be of great assistance if all future publications presenting isomorphous replacement data could incorporate as much information as possible on heavy-atom derivatives. There should also be more encouragement to authors to submit such information with the protein coordinates to the Brookhaven Databank.

The Databank contains information of value not only to protein crystallographers in the preparation of heavy-atom derivatives but also to others with interests in heavy-atom interactions with protein at the molecular level. We are now beginning a detailed analysis of the data available in order to draw out further general trends and rules that may be helpful in understanding heavy-atom protein interactions. Furthermore, we are carrying out some sample analyses to see how much protein rearrangement occurs on metal binding. Our analyses at present assume there is none. A proper analysis would involve through analysis of the heavy-atom derivative data.

Arrangements for regular updates and public distribution of the Databank are under discussion. We hope to arrange for the distribution with the CCP4 suite of computer programs for protein crystallography. We are also pleased to make them available to the Brookehaven Data Bank, when the necessary funding and personnel are available. Copies of the file systems are maintained at Birkbeck College and at the Imperial Cancer Research Fund Laboratories London.

All correspondence and requests for further information should be addressed to S. A. Islam (email address: S_ISLAM@UK.AC.ICRF).

## ACKNOWLEGEMENTS

The authors acknowledge with gratitude the many colleagues who supplied unpublished information/data to be incorporated into the Databank.

PERSONAL COMMUNICANTS

| | | |
|---|---|---|
| ADAMS, M. | KE, H. | RICHARDSON, J.S. |
| AMZEL, L.M. | KNOX, J | RICHMOND, T.J. |

BAKER, E.N.
BALLY, R.
BANASZAK, L.J.
BARFORD, D.
BLOOMER, A.
BODE, W.
BORISOVA, S.N.
BRICK, P.
BUGG, C.E.
DAVIES, D.R.
DELBAERE, B.T.J.
DIJKSTRA, B.W.
DODSON, E.
DRISSEN, H.
EVANS, P.
GLUSKER, J.P.
GRAVES, B.J.
HARRISON, P.M.
HEINEMANN, U.
HERZBERG, O.
HILGENFELD, R.
HOL, W.G.J.
HOLMGREN, A.
JANSONIUS, J.N.
JURNAK, F.
LA COUR, T.
LAWERENCE, M.C.
LEBIODA, L.
LESLIE, A.G.
LINDQVIST, Y.
MC PHERSON, A.
MATHEWS, B.W.
MATHEWS, F.S.
MARVIN, D.A.
MILLER, M.
MITSUI, Y.
MONACO, N.L.
MUIRHEAD, H.
MUTTON, J.M.
O'HAHA, B.
OFFNER, C.
OLLIS, D.
PABO, C.O.
PETRATOS, K.
PRIESTLE, J.
RAFFERTY,J.
REEKE, G.N.
REES, B.
RESHETNIKOVA, L.
RICE, D.W.
ROBERTUS, J.
ROSSMAN, M.G.
ROULD, M.A.
SACK, J.S.
SAENGER, W.
SCHULZ, G.E.
SHOHAM, M.
SIELECKI, A.
SOWADSKI, J.M.
SPRANG, S.
STOUT, C.D.
SZEBENYI, D.M.E.
TAINER, J.
TSUKIHARA, T.
TUCKER, A.
UNGE, T.
WANG, J-H.
YONATH, A.

## REFERENCES


Allen, F.H.,Kennard,O.,Taylor,R., Acc. Chem. Res. 64,146-153 (1983).
Bax, R. Annu. Rev. Biophys. Biophys. Chem. 20 (1991) 29-63.
Bernstein, F.C. et al., J. Mol. Biol. 112 (1977) 534-552.
Blake, C.C.F. Adv. Protein Chem. 22 (1968) 59-120.
Blundell, T.L. and Johnson, L.M. Protein Crystallography (1976) Academic Press, New York.
Blundell, T.L. and Jenkins, J. Chem. Soc. Rev. [Lond.] 6 (1977) 139-171.
Green, D.W. et al., Proc. Roy. Soc. A225 (1951) 287-307.
Hendrickson, W.A., Smith, J.L. and Sheriff, S. Methods in Enzymology 115 (1985) 41-55 [Eds: Wyckoff, H.W., Hirs, C.H.W. and Timasheff, S.M.] Academic Press New York.
Rossman, M.G. [Ed.] The Molecular Replacement Method, Int. Sci. Rev. Ser., 13 (1972) Gordon Breach, New York.
Wuthrich, K., Billetyer, M., and Braun, W.J. J. Mol. Biol. 180 (1984) 715-740.

# HEAVY ATOM STUDIES AT EMBL HAMBURG

Zbigniew DAUTER

European Molecular Biology Laboratory (EMBL), c/o DESY, Notkestrasse 85,
D-2000 Hamburg 52, Germany.

In Hamburg there are two beam lines available for protein crystallography, X11 and X31, both taking radiation from the storage ring DORIS. The X11 beam line consists of a triangular bent germanium monochromator, a mirror of 6 flat segments aligned on a bendable bench, several pairs of vertical and horizontal slits and a collimator equipped with its own pairs of slits and ionisation chambers at both ends. The X31 beam line contains a double channel-cut silicon monochromator, segmented toroidal mirror, pairs of slits and again a vacumised collimator with slits and ionisation chambers. On both lines the detector is placed on a cradle, which allows its position to be optimised in order to obtain maximum intensity through the collimator. On X11 any change of wavelength requires movement of the whole of the line, including mirrors and the cradle with the detector, around the monochromator axis and thorough beam optimisation. In contrast, change of wavelength at X31 is easy, as the doubly diffracted beam from the channel-cut monochromator is in principle parallel to the incoming beam and requires only slight optimisation of the mirror and cradle positions. This procedure is computer-controlled and takes only a few minutes. Routinely most of heavy-atom derivative data are thus collected on the X31 beam line with the wavelength adjusted to maximise the anomalous signal of the appropriate metal or ion. The bandpass of the monochromatised radiation on X11 is relatively wide, but on X31 is only about 5 keV ($\delta\lambda/\lambda = 0.0003$).

All examples described below are based on the diffraction data measured on the Image Plate Scanner developed in house by J. Hendrix and A. Lentfer and used routinely at EMBL Hamburg for synchrotron (and sealed-tube) data collection. The data were processed with programs from the MOSFLM package slightly modified for use with the scanner [1].

The scaling and merging of intensities and conversion to amplitudes was done using the programs ROTAVATA/AGROVATA/TRUNCATE from the CCP4 suite [2]. Programs from the same package were also used for all further calculations. For sake of saving time and avoiding the blind region problem, we very rarely align crystals accurately along the symmetry axes for anomalous data collection. Often two data sets are collected from the same crystal, one to high resolution with longer exposures, the second low resolution with short exposures to cover the strong reflections overloaded on the scanner in the other data set.

**d-UTPase** [3]

This enzyme has the molecular weight of about 16,000 (152 residues) and crystallizes in the space group R3 with a = b = 86.2, c = 62.3 Å (hexagonal setting). There is one cysteine in the sequence and soaking in ethyl mercury phosphate produced a good heavy-atom derivative with a single site. $K_2PtCl_4$ soaked into native crystals gave a single-site platinum derivative. The native data set was collected on X11 to a resolution of 1.9 Å using a wavelength of 1.0 Å. The Hg derivative data were collected on X31 to 2.0 Å with a wavelength of 0.95 Å, on the shorter side of mercury $L_{III}$ absorption edge (1.009 Å) and similarly the wavelength selected for the Pt derivative data was 1.05 Å (platinum $L_{III}$ edge is at 1.07 Å) for 2.1 Å data set. The intensities merged with agreement factors for symmetry equivalent reflections $R_{merge}$ of 6.9, 5.9 and 5.7 % for the three sets of data. Table 1 gives some statistics of the anomalous contribution within each of the two derivatives. As can be seen form the extreme values of $\Delta_{an}$, the platinum derivative set contains some abnormally large differences, resulting probably from a few reflections lying close to blind region.

Table 1.
Statistics of derivative data sets.

| Data | Resolution (Å) | Refl. | Av. F | Av. $\Delta_{an}$ | Av. $\sigma\,\Delta_{an}$ | Range of $\Delta_{an}$ |
|---|---|---|---|---|---|---|
| Native | 1.9 | 13,640 | 3,056 | | | |
| Hg | 2.0 | 11,737 | 3,330 | -8 | 208 | -2,009 : +1,815 |
| Pt | 2.1 | 9,908 | 3,583 | -26 | 244 | -4,570 : +3,843 |

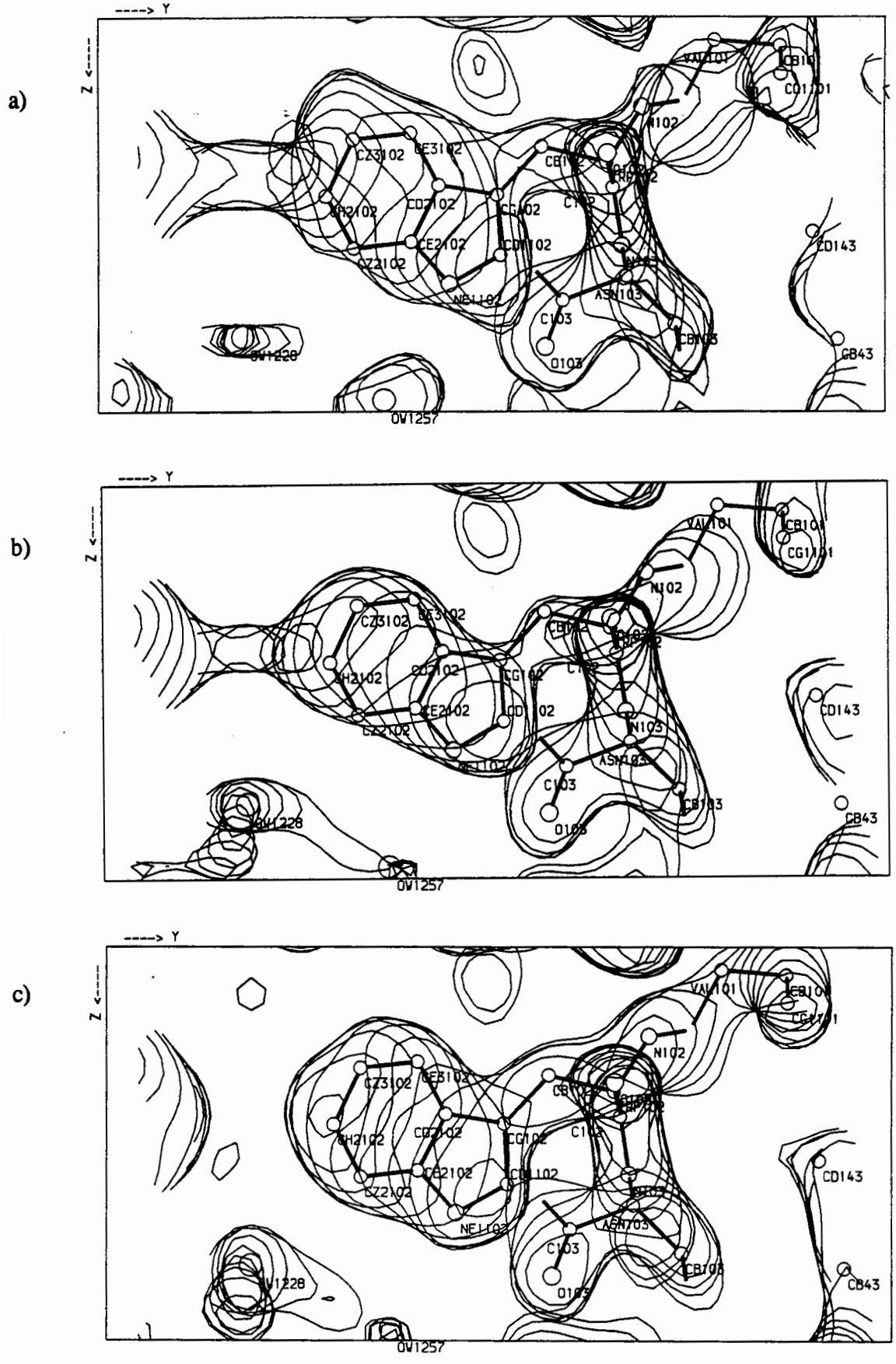



Fig. 2. The region of the electron density maps around the Trp102 residue of d-UTPase at 2.2 Å resolution: (a) initial map with experimental phases, (b) solvent-flattened map (c) final $2F_o$-$F_c$ map for the refined model.. All contours at 1 σ.

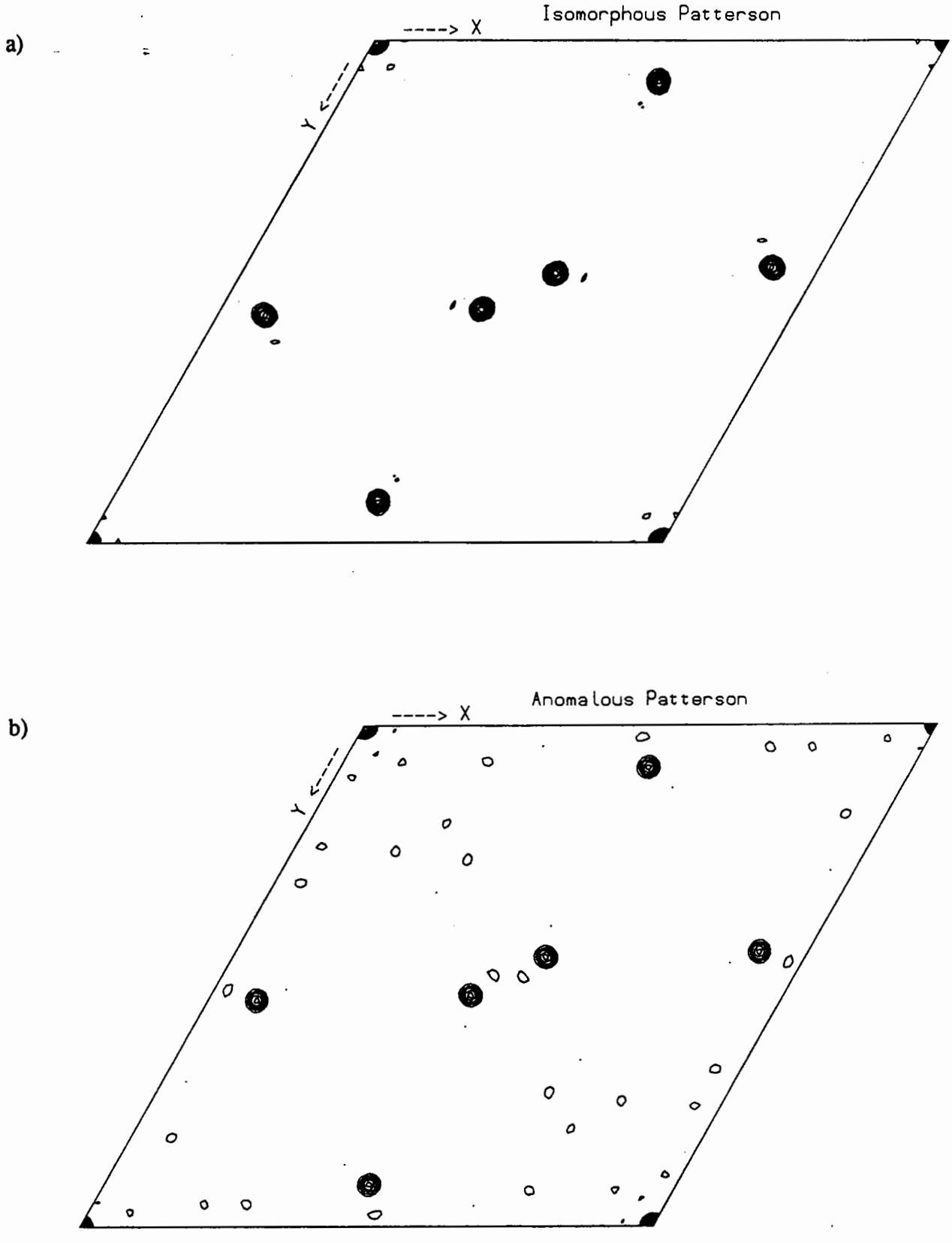



Fig. 1. (a) Isomorphous difference and (b) anomalous difference Patterson syntheses for the Hg derivative of d-UTPase calculated with data between 3 and 10 Å

The initial map resulting from the above phasing procedure was subjected to 6 cycles of a solvent flattening procedure [4,5], which increased the figure of merit from 0.66 to 0.79. The resulting map was very clear and allowed the unambiguous tracing of 136 residues with their side chains. The d-UTPase structure was refined by a restrained least-squares procedure [6] combined with FRODO [7] graphics sessions to an R = 14.5 % at 1.9 Å. The final model consists of 136 amino acid residues and 189 water molecules. The 15 C-terminal residues not traced in the initial map are also not visible in the final maps, in spite of being present in the crystal of d-UTPase [8]; they are completely disordered.

The representative region of the three electron density maps, (a) with experimental phases, (b) with phases after the solvent-flattening procedure and (c) with phases corresponding to the final, refined model is shown in Figure 2. The comparison of the three phase sets is summarised in Figure 3 which shows the average difference of phases as a function of resolution. The overall average difference between the initial and final phases is 41°, and is only 27° if the resolution is restricted to 3 Å.

**Narbonin** [9]

Narbonin is a seed storage protein with molecular weight of about 33,000. It crystallizes in the space group $P2_1$ with cell dimensons a = 46.9, b = 75.5, c = 50.9 Å, $\beta$ = 120.5°. The data were collected for the native crystals to 1.8 Å on the X11 beam line and for several derivatives on synchrotron beam lines X11 and X31, or on the sealed tube source, Table 2.

Table 2.
Summary of the data collected for Narbonin and its derivatives.

| Crystal | X-ray source | Resolution (Å) | $\lambda$ (Å) | $R_{merge}$ (%) | $F_{del}$ (%) | $K_{anom}$ |
|---|---|---|---|---|---|---|
| Native | X11 | 1.8 | 0.96 | 5.6 | | |
| $Hg(CH_3COO)_2$ | X11 | 2.1 | 0.95 | 5.3 | 26 | 6.3 |
| $Hg(CH_3COO)_2$ | X31 | 3.8 | 1.01 | 3.7 | 23 | - |
| $UO_2(NO_3)_2$ | X11 | 4.0 | 0.96 | 4.8 | 30 | - |
| $K_2PtCl_4$ | Cu $K_\alpha$ | 2.2 | 1.54 | 6.1 | 24 | 4.3 |
| $SmCl_3$ | Cu $K_\alpha$ | 2.2 | 1.54 | 4.3 | 13 | 4.4 |
| $SmCl_3$ | X31 | 3.9 | 1.01 | 4.2 | 13 | 4.4 |

The amplitudes were scaled to the native set and gave average isomorphous differences $F_{del}$ and $K_{anom}$ parameter of 15.4 % and 6.15 for the Hg derivative and 15.0 % and 6.58 for the Pt derivative. The Harker sections at w = 0 of the isomorphous difference and anomalous Patterson syntheses for the Hg data are reproduced in Figure 1. It corresponds to a single-site derivative and peak heights are about 1/5 and 1/7 of the origin in those the respective syntheses. A similar but slightly lower peak appeared in Patterson functions for the Pt data although in different positions. As found out later, Pt binds to methionine, not cysteine as Hg. The mutual relation of the Pt and Hg coordinates was found from a Fourier synthesis based on differences between the Pt derivative and native amplitudes and phases derived from Hg derivative, which gave a clear maximum in agreement with the Pt Patterson functions. A this stage the correct enentiomorph was selected, as inversion of Hg coordinates led to a cross-phased synthesis with no meaningful peaks for Pt. The map phased on the Hg derivative showed clear contrast between the protein and solvent regions only for the correctly chosen enantiomorph.

The phasing procedure based on two derivatives gave a phase set with an overall figure of merit of 0.70 at 2.2 Å based on 10108 reflections. It can be stressed that in space group R3 there are no centric zones and therefore no reflections with restricted phases.

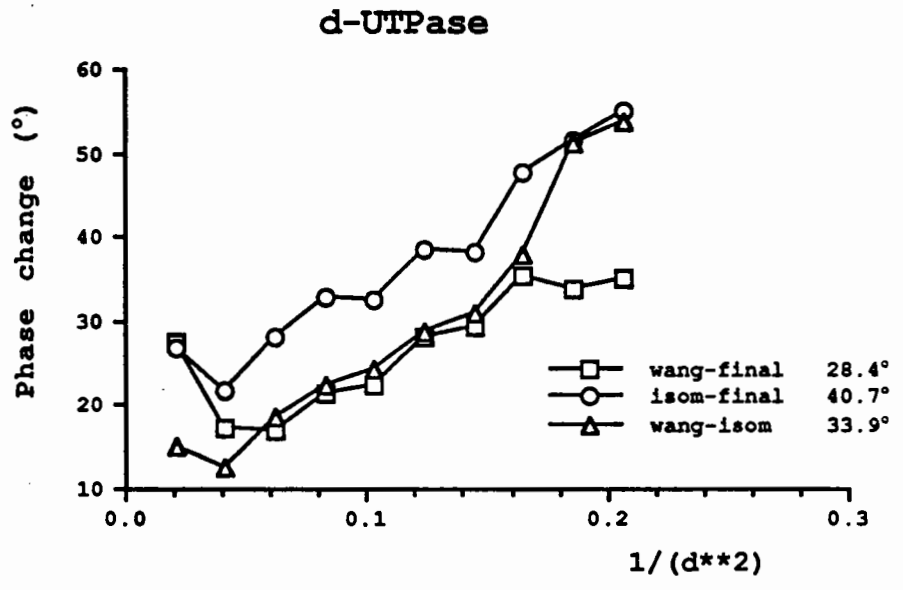



Fig. 3. The average difference between the experimental, solvent-flattened and final phase sets of d-UTPase as a function of resolution.

The sequence of this protein is not known at present and several heavy atom derivatives were used to give a high quality initial map. The phasing power of these derivatives ranged from 1.0 (Sm, one site) to 5.1 (Hg, four sites). All six derivatives were used for phasing and gave an overall figure of merit of 71 %. Four cycles of solvent flattening [4,5] improved the map quality and raised the figure of merit to 83 %.
The resulting map showed clear continuity of the electron density and allowed the tracing of 70 % of the polypeptide chain using the FRODO [7] program. After a few iterations of restrained refinement [6] and graphics sessions the complete protein chain was built into the density with the majority of side chains identified.

**Pyrophosphatase** [10]

The pyrophosphatase project was started during a period of rebuilding of the DORIS storage ring and no synchrotron radiation was at that time available in Hamburg. The X-ray data were therefore collected using the Image Plate scanner mounted on a sealed tube source with $CuK_\alpha$ radiation. The enzyme has 175 residues in a single chain and crystallises in the space group R32 with two subunits in the asymmetrirc unit. The cell dimensions are a = b = 110.4, c = 154.9 Å. The native and mercury derivative (ethyl mercury thiosalicylate, EMTS) data were collected to 2.7 Å resolution. The $R_{merge}$ values for the native data was 5.6 % and for the Hg data 7.5 %. Scaling to the native data gave an $F_{del}$ of 27 % and $K_{anom}$ of 10.7 for the Hg derivative.

Figure 5 shows two sections, w = 0 and 0.113, of the isomorphous and anomalous Pattersons. The presence of the very strong peak at 1/3, 2/3, 0.116 suggests that the packing of two molecules in the asymmetric unit is highly pseudosymmetric and that they are related by a translation along z-axis of 0.450 (1/3 + 0.116). Two major and one minor Hg sites were identified and phasing based on those sites gave an overall figure of merit of 64 %. Subsequent application of solvent flattening [4,5] increased the figure of merit to 78 % and produced a map which is at present being interpreted.

The Harker sections (v = 1/2) of the isomorphous difference and anomalous Patterson syntheses for the first mercury acetate derivative collected on the X11 beam line are presented in Figure 4. Three strongly bound and one weak Hg sites can be identified from these maps.

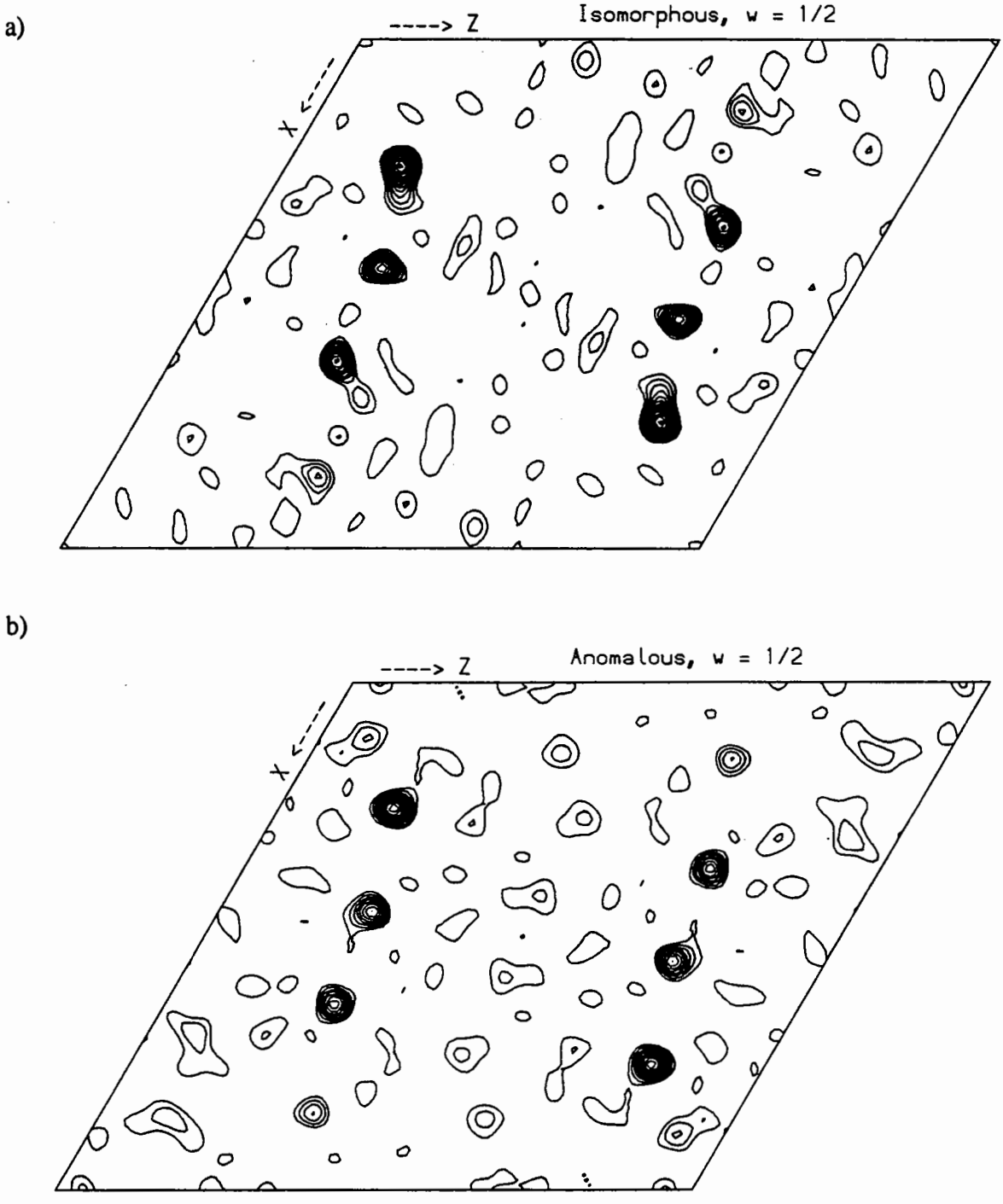



Fig. 4. The isomorphous (a) and anomalous (b) difference Patterson syntheses for the mercury acetate derivative of narbonin with data between 2.5 and 8 Å

## REFERENCES


[1] Leslie, A.G.W., Brick, P. and Wonacott, A.J. (1986) CCP4 News 18, 33-39.

[2] CCP4 (1979). The SERC (UK) Collaborative Computing Project No. 4: a Suite of programs for Protein Crystallography, distributed from Daresbury Laboratory, Warrington WA4 4AD, UK.

[3] d-UTPase from *E.coli*, Cooperation of Z.D. and K.S. Wilson (EMBL) with E.Cedergren-Zeppezauer (Stockholm) and G. Larsson and P.O. Nyman (Lund).

[4] Wang, B.C. (1985) Methods Enzymol., 115, 90-112.

[5] Leslie, A.G.W. (1987) Acta Crystallogr. A43, 134-136.

[6] Konnert, J.H. and Hendrickson, W.A. (1980) Acta Crystallogr. A34, 791-809.

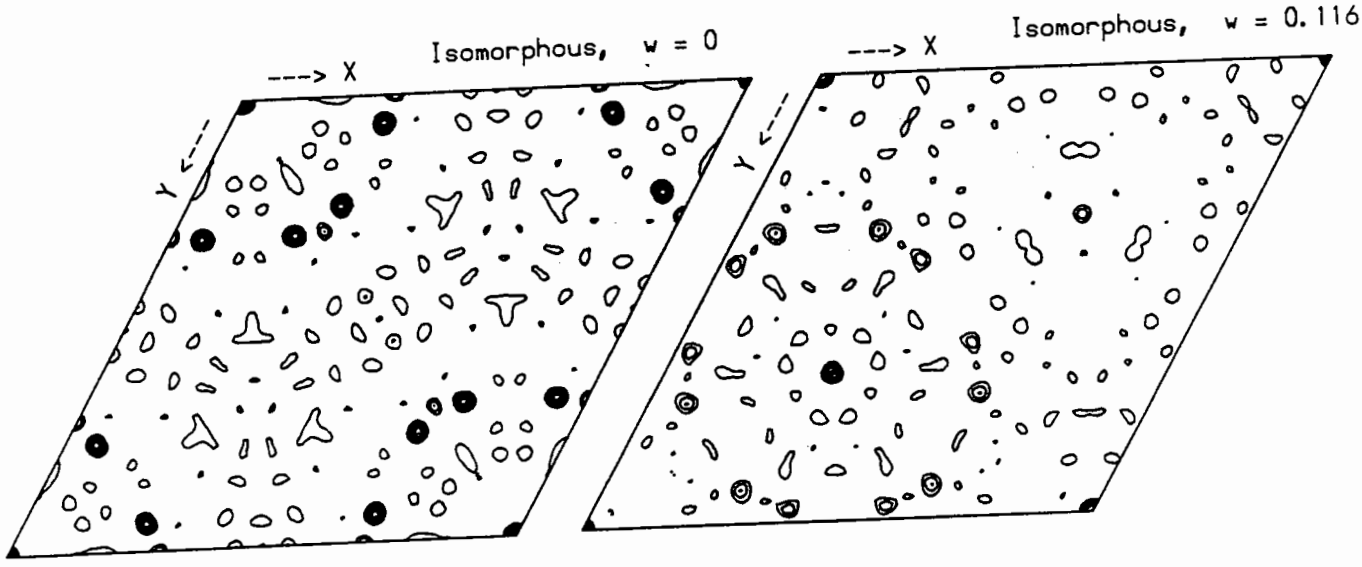
[7] Jones, T.A. (1978) J. Appl. Crystallogr. 11, 268-272.

[8] Confirmed by the sequencing of the dissolved crystalline material by G. Larsson.

[9] The protein isolated form seeds of *Vicia narbonensis* L. Cooperation of M. Hennig (EMBL) with B. Schlesier (Gatersleben).

[10] Inorganic pyrophosphatase from *E coli*. Cooperation of Z.D. with E. Harutyunyan (Moscow).

a)




b)

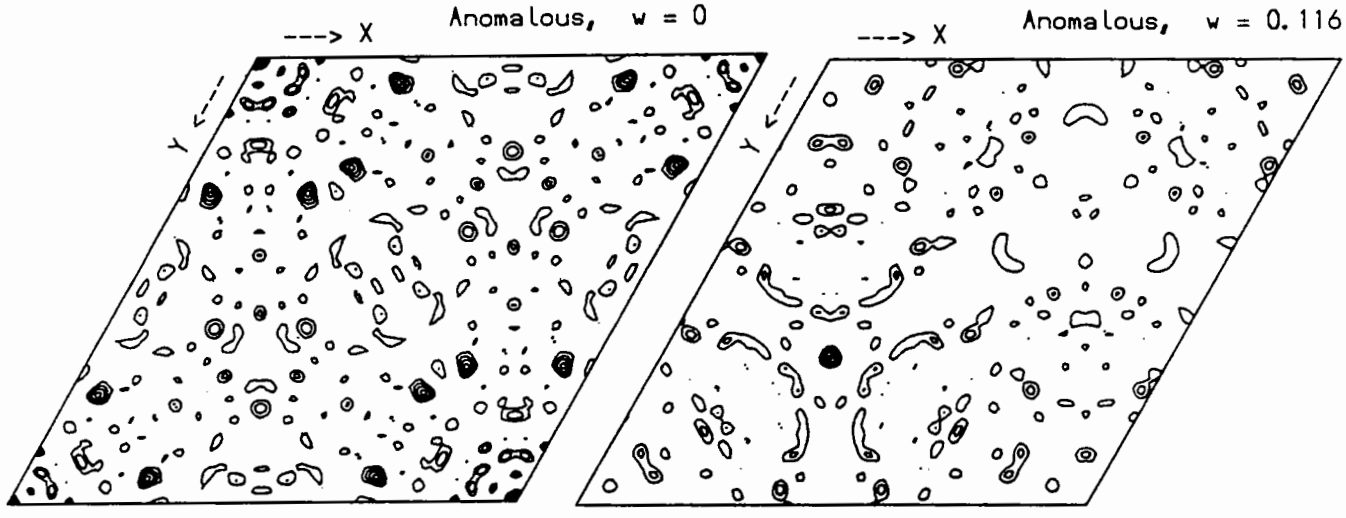



Fig. 5 Two sections, w = 0 and w = 0.116, of Patterson maps for the Hg derivative of inorganic pyrophosphatase; (a) isomorphous and (b) anomalous difference Patterson synthesis.

# LIST OF DELEGATES

Abeysinghe, I S B
Department of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield S10 2TN.

Achari, A
Genex Corporation. 16020 Industrial Drive, Gaithersburg, Maryland 20877, USA.

Adams, M J
Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.

Adams, P D
Department of Biochemistry, University of Edinburgh, Hugh Robson Building, George Square, Edinburgh EH8 9XD.

Aevarsson, A
Department of Molecular Biophysics, Chemical Centre, University of Lund, PO Box 124, S-221 00 Lund, Sweden.

Armstrong, S
Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT.

Artymiuk, P
Department of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield S10 2TN.

Bailey, S
Daresbury Laboratory, Warrington WA4 4AD.

Baker, P J
Department of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield S10 2TN.

Barrett, T E
Department of Chemistry, University of York, Heslington, York YO1 5DD.

Basak, A K
Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.

Bax, B
Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.

Bentley, G
Unite d'Immunologie Structurale, Institut Pasteur, 25 rue de Dr Roux, 75724 Paris, France.

Betzel, C
EMBL, c/o DESY, Notkestrasse 85, 2000 Hamburg 52, Germany.

Bewley, M
Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT.

Bloomer, A C
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH.

Blow, D M
Blackett Laboratory, Imperial College, London SW7 2BZ.

Bocskei, Z
Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT.

Boys, C W G
Department of Biochemistry, University of Edinburgh, Hugh Robson Building, Edinburgh EH8 9NX.

Brick, P
Blackett Laboratory, Imperial College, Prince Consort Road, London SW7 2BZ.

Bricogne, G
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH.

Britton, K L
Department of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield S10 2TN.

Brocklebank, I P
Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT.

Broutin, I
ICSN, Avenue de la tenasse, 91198 Gif/Yvette, France.

Brown, D G
CRC Biomolecular Structure Unit, Institute of Cancer Research, Block F, Cotswold Road, Sutton, Surrey SM2 5NG.

Brownlie, P
Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.

Buchanan, S
MRC Laboratory of Molecular Biology, PNAC, Hills Road, Cambridge CB2 2QH.

Fabiane, S M
Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.

Flower, D
Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT.

Ford, G C
Krebs Institute, University of Sheffield, Western Bank, Sheffield S10 2TN.

Fortier, S
Department of Chemistry, Queen's University, Kingston, Ontario, Canada.

Franken, S M
Max Planck Institute for Medical Research, Jahn Strasse 29, D-6900 Heidelberg, Germany.

Frazao, C
CTQB R da Quinta Grande 6, Apartado 126, 2780 Oeiras, Portugal.

Freemont, P S
Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX.

Frey, M
Laboratoire de Cristallographie et de Cristallogenese, Lab d'Ingeniere des Proteines, CEN.G 85 X, 38041 Grenoble cedex, France.

Ghosh, M
Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.

Goldberg, J D
Blackett Laboratory, Imperial College, Prince Consort Road, London SW7 2BZ.

Gonzalez, A
Daresbury Laboratory, Warrington WA4 4AD.

Gordon, E J
Department of Biochemistry, University of Edinburgh, Hugh Robson Building, George Square, Edinburgh EH8 9XD.

Gorman, M A
Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX.

Gouet, P
Centre d'Etudes Nucleaires de Grenoble, DBMS/DSV/BS, BP85 X 38041 Grenoble, France.

Gover, S
Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.

Green, T P
Department of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield S10 2TN.

Greenough, T J
Department of Physics, University of Keele, Keele, Staffs ST5 5BG.

Grimes, J M
Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.

Guthrie, N
Department of Chemistry, University of Glasgow, Glasgow G12 8QQ.

Harding, M M
Department of Chemistry, University of Liverpool, PO Box 147, Liverpool L69 3BX.

Harlos, K
Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.

Harris, D
Department of Chemistry, University of Glasgow, Glasgow G12 8QQ.

Harris, G W
Department of Protein Engineering, AFRC, Institute of Food Research, Shinfield Road, Reading RG2 9AT

Harrop, S J
Department of Structural Chemistry, University of Manchester, Manchester M13 9PL.

Hasnain, S S
Daresbury Laboratory, Warrington WA4 4AD.

Helliwell, J R
Department of Chemistry, University of Manchester, Manchester M13 9PL.

Hemmings, A M
Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.

Hendrickson, W A
Department of Biochemistry and Molecular Biophysics, Columbia University, New York 10706, USA.

Henrick, K
IRC for Protein Engineering, MRC Centre, Hills Road, Cambridge CB2 2QH.

Higgins, T
Department of Chemistry, University College, Galway, Ireland.

Hilgenfeld, R
Central Research G864, Hoechst AG, POB 800320, D-6230 Frankfurt-80, Germany

Holden, P
Daresbury Laboratory, Warrington WA4 4AD.

Hunter, W N
Department of Chemistry, University of Manchester, Manchester M13 9PL.

Husain, J
Department of Crystallography,Birkbeck College, Malet Street, London WC1E 7HX

Isaacs, N
Department of Chemistry, University of Glasgow, Glasgow G12 8QQ.

Ito, N
Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT.

Jaeger, J
Department of Structural Biology, Biozentrum, Universitat Basel, Klingelbergstr. 70, CH-4056 Basel, Switzerland.

Jeffrey, P
Bristol Myers Squibb Pharmaceutical Research Institute, PO Box 4000, Princeton, New Jersey 08543-4000, USA.

Jensen, K
Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT.

Jhoti, H
Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.

Jiang, J
Department of Chemistry, University of York, Heslington, York YO1 5DD.

Jones, E Y
Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.

Jones, A
Department of Molecular Biology, University of Uppsala Biomedical Centre, Box 590, S-751 Uppsala, Sweden.

Kahn, R
LURE, Bâtiment 209D, F-91405 Orsay Cedex, France

Keep, N H
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH.

Knight, S D
Department of Molecular Biology, SLV, Uppsala Biomedical Centre, Box 590, S-75124 Uppsala, Sweden.

Knossow, M
Laboratoire de Biologie Physichimique, Bat 433, Universite Paris Sud, 91405 Orsay Cedex, France.

Kokkinidis, M
Institute of Molecular Biology and Biotechnology, PO Box 1527, GR-71110 Iraklion, Crete, Greece.

Lambert, R
Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.

Langdon, G M
Department of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield S10 2TN.

Lange, G
Department of Chemistry, University of York, Heslington, York YO1 5DD.

Lantwin, C B
Max Planck Institute for Medical Research, Jahn Str 29, 6900 Heidelberg, Germany.

Laughlan, G M
Department of Protein Crystallography, Institute of Chemistry, University of Glasgow, Glasgow G12 8QQ.

Lea, S M
Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.

Leadbetter, A J
Daresbury Laboratory, Warrington WA4 4AD.

Leslie, A G W
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH.

L'Hermite, G
LURE, Bat 209D, Centre Universitaire, 91405 Orsay Cedex, France.

Li, J
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH.

Li de la Sierra, I M
LURE, Bat 209D, Centre Universitaire, 91405 Orsay Cedex, France.

Littlechild, J A
Department of Biochemistry, University of Bristol, School of Medical Sciences,

Littlejohn, A
Department of Chemistry, University of Glasgow, Glasgow G12 8QQ

Logan, D T
Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.

Louie, G V
Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.

Luisi, B
MRC Virology Unit, Institute of Virology, Church Street, Glasgow G11 5JR.

Martinez, C
LCCMB-CNRS, Faculte de Medecine, Secteur Nord, Bld Pierre Dramard, 13326 Marseille Cedex 15, France.

McDonald, N
Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.

McLaughlin, P
Laboratory of Molecular Biology, MRC, Hills Road, Cambridge CB2 2QH.

McMichael, P
Department of Physics ISAT, Liverpool Polytechnic, Byrom Street, Liverpool L3 3AF.

Moody, P C E
Department of Chemistry, University of York, Heslington, York YO1 5DD.

Moore, M H
Department of Chemistry, University of York, Heslington, York YO1 5DD.

Morais Cabral, J H
Department of Biochemistry, University of Edinburgh, Hugh Robson Building, George Square, Edinburgh EH8 9XD.

Moss, D
Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.

Mowbray, S
Department of Molecular Biology, University of Uppsala Biomedical Centre, Box 590, S-751 Uppsala, Sweden.

Mueller, A
Institut fur Kristallographie, c/o Professor Saenger, Freie Universitat Berlin, Takustrasse 6, D-1000 Berlin 33, Germany.

Murphy, L
Daresbury Laboratory, Warrington WA4 4AD.

Murray-Rust, J
Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.

Muskett, F W
Department of Biochemistry, University of Edinburgh, Hugh Robson Building, George Square, Edinburgh EH8 9XD.

Nagai, K
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH.

Naismith, J H
Department of Chemistry, University of Manchester, Manchester M13 9PL.

Nave, C
Daresbury Laboratory, Warrington WA4 4AD.

Neil, T K
Department of Chemistry, University of York, Heslington, York YO1 5DD.

Neu, M
Daresbury Laboratory, Warrington WA4 4AD.

Newman, M
ICRF Unit, Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.

North, A C T
Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT.

Nunn, R
Department of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield S10 2TN.

Nunn, C M
University Chemical Laboratory, University of Cambridge, Lensfield Road, Cambridge CB2 1EW.

Ogg, D J
Symbicom AB, Glunten, S-751 83 Uppsala, Sweden.

O'Hara, B
Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.

Oliva, G
Laboratorio de Cristalografia de Proteinas, Departamento de Fisica e Ciencia dos Materiais, IFQSC-USP, Caixa Postal 369, Sao Carlos, Brazil.

Onesti, S
Blackett Laboratory, Imperial College, Prince Consort Road, London SW7 2BZ.

Otwinowski, Z
Dept of Molecular Biophysics and Biochemistry, Yale University, 260 Whitney Avenue/JWG 402, New Haven, CT 06511, USA.

Papiz, M
Daresbury Laboratory, Warrington WA4 4AD.

Pebay-Peyroula, E
Institut Laue-Langevin, BP 156 X, 38042 Grenoble Cedex, France.

Phillips, C
Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.

Phillips, K
Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT.

Phillips, S E V
Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT.

Pickersgill, R
AFRC Reading Laboratory, Shinfield, Reading RG2 9AT.

Podjarny, A D
IBMC, 15 Rue Descartes, 67084 Strasbourg, France.

Powell, H R
Institute of Food Research, AFRC, Shinfield, Reading.

Priestle, J
K-681.5.43, CIBA-GEIGY AG, CH-4002 Basel, Switzerland.

Prince, S M
Biophysics Research, ISAT, Liverpool Polytechnic, Byrom Street, Liverpool L3 3AF.

Rafferty, J
Department of Biochemistry and Biology, University of Leeds, Leeds, LS2 9JT

Raftery, J
Department of Structural Chemistry, University of Manchester, Manchester M13 9PL.

Rao, Z
Sir William Dun School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE

Rawas, A
Department of Chemistry, University of Bristol, University Walk, Bristol BS8 1TD.

Read, R
Department of Medical Microbiology & Infectious Diseases, 1-41 Medical Sciences Bldg, University of Alberta, Edmonton, Alberta T6G 2H7, Canada.

Reid, A J
ISAT, Department of Physics, Liverpool Polytechnic, Byrom Street, Liverpool L3 3AF.

Ren, J
Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.

Rice, D W
Department of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield S10 2TN.

Rizkallah, P
Daresbury Laboratory, Warrington WA4 4AD.

Roth, M
Centre d'Etudes Nucleaires, L.I.P./L.C.C.P., BP 85X, 38041 Grenoble cedex, France.

Sanderson, M R
CRC Biomolecular Structure Unit, Institute of Cancer Research, Block F, Cotswold Road, Sutton, Surrey SM2 5NG.

Sawyer, L
Department of Biochemistry, University of Edinburgh, George Square, Edinburgh EH8 9XD.

Schiering, N S
Max Planck Institut fur Medizinische Forschung, Abt Biophysik, 29 Jahn Str, D-6900 Heidelberg, Germany.

Schirmer, T
Biozentrum, Universitat Basel, Klingelbergstr. 70, CH-4056 Basel, Switzerland.

Schreuder, H
Laboratory for Chemical Physics, University of Groningen, Nijenborgh 16, 9747 AG Groningen, Netherlands.

Sery, A
Laboratoire de Cristallographie et de Cristallogenese, Lab d'Ingeniere des Proteines, CEN.G 85 X, 38041 Grenoble cedex, France.

Sharff, A J
Laboratory of Molecular Biophysics, University of Oxford, Rex Richards Building, South Parks Road, Oxford OX1 3QU.

Shaw, A
Department of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield S10 2TN.

Sheldrick, G M
Institut fur Anorganische Chemie, Tammannstr 4, D-3400 Gottingen, Germany.

Shrive, A K
Department of Physics, University of Keele, Keele, Staffs ST5 5BG.

Simpson, A A
Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.

Sixma, T
Department of Chemistry, University of Groningen, Nijenborgh 16, 9747 AG Groningen, The Netherlands.

Skarzynski, T
Blackett Laboratory, Imperial College, Prince Consort Road, London SW7 2BZ.

Skelly, J V
CRC Biomolecular Structure Unit, Institute of Cancer Research, Block F, Cotswold Road, Sutton, Surrey SM2 5NG.

Smith, J L
Department of Biological Sciences, Purdue University, West Lafayette, IN 49707, USA.

Stein, P
Department of Haematology, MRC Centre, Hills Road, Cambridge CB2 2QH.

Stillman, T J
Department of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield S10 2TN.

Stoll, V S
Max Planck Institut fur Biophysik, 29 Jahn Str, D-6900 Heidelberg, Germany.

Strathdee, S
Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT.

Su, X-D
Department of Molecular Genetics, Medical Nobel Institute, Karolinska Institutet, Box 60400, 104 01 Stockholm, Sweden.

Svensson, L A
Department of Molecular Biophysics, Chemical Centre, University of Lund, Box 124, S-221 00 Lund, Sweden.

Tame, J
Department of Chemistry, University of York, Heslington, York YO1 5DD.

Thayer, M M
Department of Chemistry-0654, University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093 0654, USA.

Thierry, J-C
IBMC, 15 Rue Rene Descartes, 67084 Strasbourg Cedex, France.

Thompson, A W
Daresbury Laboratory, Warrington WA4 4AD.

Tickle, I J
Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.

Tierney, S G T
Department of Chemistry, University of Glasgow, Glasgow G12 8QQ.

Turkenburg, J
Department of Chemistry, University of York, Heslington, York YO1 5DD.

Turner, M
Department of Biochemistry, University of Edinburgh, Hugh Robson Building, George Square, Edinburgh EH8 9XD.

van Tilbeurgh, H
Laboratoire de Cristallographie, Fac de Medecine, Bvd Pierre Damard, 13326 Marseille Cedex 15, France.

Veerapaneni, N
Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX.

Vellieux, F M D
Department of Chemical Physics, State University of Groningen, Nijenborgh 16, 9747 AG Groningen, The Netherlands.

Vidgren, J
Department of Molecular Biophysics, Chemical Centre, University of Lund, PO Box 124, S-22100 Lund, Sweden.

Walker, N
BASF Ag, Carl Bosch strasse 38, D-6700 Ludwigshafen, Germany.

Walkinshaw, M
Preclinical Research, Sandoz Pharma AG, 4002 Basel, Switzerland.

Waller, D A
Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT.

Walsh, M
Department of Chemistry, University College, Galway, Ireland.

Wan, T
Biophysics Section, King's College London, 26-29 Drury Lane, London WC2B 5RL.

Watson, H C
Department of Biochemistry, University of Bristol, School of Medical Sciences, Bristol BS8 1TD.

Webster, G D
CRC Biomolecular Structure Unit, Institute of Cancer Research, Block F, Cotswold Road, Sutton, Surrey SM2 5NG.

Weisgerber, S
Department of Structural Chemistry, University of Manchester, Manchester M13 9PL.

White, S
Department of Microbiology and Immunology, Box 3020, Duke Medical Center, Durham, North Carolina 27710, USA.

Whittingham, J L
Department of Chemistry, University of York, Helsington, York YO1 5DD.

Wigley, D B
Department of Chemistry, University of York, Heslington, York YO1 5DD.

Wilson, K S
EMBL, c/o DESY, Notkestrasse 85, 2000 Hamburg 52, Germany.

Wolf, W
Daresbury Laboratory, Warrington WA4 4AD.

Wonacott, A
Glaxo Group Research Ltd, Protein Structure Group, Greenford Road, Greenford, Middlesex UB6 0HE.

Xiao, B
Department of Chemistry, University of York, Heslington, York YO1 5DD.

Young, F
Department of Chemistry, University of Glasgow, Glasgow G12 8QQ.

Young, R
Department of Biomolecular Science, King's College London, 26 Drury Lane, London WC2B 5RL.

9101-spare