DARESBURY LABORATORY

INFORMATION QUARTERLY for PROTEIN CRYSTALLOGRAPHY

An Informal Newsletter associated with Collaborative Computational Project No. 4
on Protein Crystallography

Number 11 June 1983

Contents

LOCATION AND REFINEMENT OF HEAVY ATOM SITES

Proceedings of a meeting held at Bristol University, 16-17 September 1982

Editor:	Pella Machin	Science & Engineering Research Council, Daresbury Laboratory, Daresbury, Warrington WA4 4AD, England
Deputy Editor for Imperial College:	Dr. Alan Wonacott	Imperial College of Science & Technology, The Blackett Laboratory, Prince Consort Road, London SW7 2BZ
Deputy Editor for Birkbeck College:	Dr. David Moss	Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX

Foreword

A CCP4 meeting was held at Bristol University on September 16 - 17 1983 on the Location and Refinement of Heavy Atom Sites. The meeting covered both the theory and current practice of the isomorphous replacement method with discussion of recent developments.

I would like to thank Alan Wonacott and David Moss for organizing the meeting and thank the speakers for both giving for formal talks and preparing papers for this newsletter. The final paper included here describes work relevant to this topic which was completed recently.

A summary of the discussion at the meeting will be included in the next newsletter.

Pella Machin

Solving Pattersons for Heavy Atom Sites

Eleanor Dodson, York University

Most protein crystal structures are solved using the measured Xray amplitudes for the lattice with phases generated from isomorphous replacement. Heavy atoms are enticed into the protein, the amplitudes for the modified crystal are remeasured, the heavy atom sites are located and from all this information, a phase for the Fp is derived. The theory is given in various books and papers - I would like to dwell a little on what can go wrong. If the observed differences between the Fp, the Fph(+) and Fph(-) are due to heavy atom substitution; and if these data sets can be sensibly scaled together, then it is possible to calculate a Patterson function which should show the vectors between all heavy atom pairs. Possible estimates of Fh^2 for this summation are listed here.

1. Isomorphous Differences. (Δ iso)

$$(k(Fph)\ \exp(Bs^2) - Fp)^2 \approx (Fh \cos(Aph - Ap))^2$$
$$= Fh^2/2 + Fh^2 \cos 2(Aph-Ap)/2$$

The second term has a random phase and will just generate background noise.

A Patterson using these differences should give a map with the expected vectors at about half weight.

2. Anomalous Differences. (Δ ano)

$$(Fph(+) - Fph(-))^2 \approx (2\ F''h \sin(Aph - Ap))^2$$
$$= 2\ F''h^2 - 2\ F''h\ 2\cos 2(Aph-Ap)$$

Again the second term generates noise only, and the Patterson should show peaks at half the $F''h^2$ height. Obviously these are much lower than those from an isomorphous Patterson.

3(a) $Fhle^2$

$$Fhle^2 \approx \Delta iso^2 + (F'h\ \Delta ano/(2f''h))^2$$

Phil Evans has given a detailed description of the derivation of Fhle. Fhle should theoretically equal Fh for most (h k l).

3(b) $Fhle^2$ - Bias Correction

The root mean square $<Fhle^2>$ is always an overestimate of $<Fh^2>$. If any observation D is distributed about its true value with an expected error ±E, then:

$$<D^2> = <Dtrue^2> + <E^2>$$

Phil is explaining in detail how to estimate this bias from the standard deviation of the observations. It expresses statistically the obvious truth, that the differences observed between two inaccurate measurements (usually the weak Fp and Fph) are unreliable.

It is worth remembering that Pattersons are effectively functions of their large terms only - the example illustrates this very clearly, and therefore if the largest terms are wrong the map will be useless.

This means that the anomalous difference Patterson is often not useful, despite the facts that the two observations have come from the same crystal and should therefore have the same systematic errors, and that there is no scaling problem. A few terms with large errors can disguise the true peaks completely. A bias correction here usually excludes all the differences between the weak amplitudes.

The isomorphous difference Patterson is usually better since while the Eiso is of the same order as Eano, the Diso are much bigger. This is always providing the differences ARE due to specific heavy atom substitution, and not just to changes in the protein structure. We must remember that even in a good derivative the Fh parameters rarely explain more than 50% of the observed differences. One way of judging the quality of the substitution is by inspection of $<\Delta iso>$ and $<\Delta ano>$ against $\sin\theta$ and mod(Fp). The size of the deltas should be independent of mod(Fp); and the ratio $<\Delta iso>/<\Delta ano>$ should be close to the theoretical value of Fh/F''h.

But even good Pattersons can be very difficult to interpret. In a complicated substitution, either because of high crystal symmetry or because there are many heavy atom sites, the vectors often overlap and diffraction ripples can blot out true peaks.

The moral ?? - measure your data as carefully as possible; try to get the same systematic errors in all data sets, (i.e. similar shaped crystals, mounted about the same axis, measure F+ and F- at about the same time). Relative Wilson plots give a fair estimate of the scale between protein and derivative

sets of intensities, and luckily Pattersons are not very sensitive to this scale. These parameters can be refined later along with the heavy atom coordinates before the isomorphous phase calculation is done. (Local scaling may help reduce the Patterson noise but its main use seems to be to let you spot crazy differences - whether you use it or not, devise some method for listing all your largest terms.)

And remember the good news - you really only need to solve one Patterson per protein. Once you have one set of heavy atom parameters with anomalous or centric data you can generate good enough protein phases for some of the amplitudes to let you find your next heavy atom sites from the difference Fourier.

References

Dodson, E.J., Evans, P.R. and French, S. (1975), Anon. Scattering, pp. 423-436, ed. S. Ramaseshan and S.C. Abrahams, Munksgaard.

Local Scaling of Heavy-Atom Derivative Data and the Solution of Heavy-Atom Difference Pattersons by Vector Search.

by Ian J. Tickle

Crystallography Dept., Birkbeck College, Malet Street, London, WC1E 7HX.

1. Local Scaling

Mathews and Czerwinski's local scaling technique (1) has been implemented on the SERC AS7000 and MRC VAX computers. Local scaling should only be performed on data which have already had a suitable overall scale applied; otherwise no conclusions can be made regarding the validity of the local scaling. This pre-scaling is done in the current implementation by determination and application of an overall anisotropic temperature factor. This minimises

$$\sum_{\underline{h}} w_{\underline{h}} (T_{\underline{h}} F_{1\underline{h}} - F_{2\underline{h}})^2$$

where $T_{\underline{h}} = K \exp(\underline{h}^T . \beta . \underline{h})$

K is the overall scale factor

β is the overall anisotropic thermal tensor,

(K and β to be found by least squares iteration.)

Then for local scaling minimise :

$$\sum_{\underline{h}'} w_{\underline{h}'} (T_{\underline{h}} F_{1\underline{h}'} - F_{2\underline{h}'})^2$$

where $\underline{h}'$ are reciprocal lattice points local to $\underline{h}$, within a sphere of predefined radius s_{max} centred on $\underline{h}$.

w is the weight = $(v_1 + v_2)^{-1} f$

v is the variance = $\sigma_c^2(F) + (0.5aF)^2$

f is an attenuation factor = $\exp (-\Delta s^2 / s^2_{max})$

Δs the distance in reciprocal space between $\underline{h}'$ and $\underline{h}$.

The significance of the local scaling is judged by comparing

$$\text{rms} (T_{\underline{h}} - 1) \quad \text{with} \quad \text{rms}\ \sigma(T_{\underline{h}})$$

as a function of h,k,l,s,F_1 and F_2.

2. Features of the Vector Search program.

a) It is completely space-group general : the user supplies the lattice type and general equivalent positions, and specifies the asymmetric units of the symmetry function and the space group.

b) The program takes the Patterson output by FFT, and will work with any number of Laue group asymmetric units (the more are supplied, the faster the program will run; usually half a unit cell is adequate.)

c) The Patterson is truncated at the "expected height of a single-weight vector" (this depends on some estimate of the number of heavy-atom sites expected) and at 0; after adding an estimated F_{000} contribution.

eg P(0,0,0) = 1000

estimated F_{000} contribution = 10 (0.5-1% of P(0,0,0))

expected number of major sites = 3/asym unit

= 24/unit cell in $P4_12_12$

estimated single weight vector = (1000+10)/24 = 42

d) Options of arithmetic mean function (equivalent to sum function), minimum function and harmonic mean function (reciprocal of arithmetic mean of reciprocals).

e) "Local search" feature - to be used with care (see results below). This involves a search of the Patterson within a predefined radius of the calculated position of each vector (based on the assumption that in low-resolution Pattersons peaks never appear exactly at the calculated position).

3. Procedure for automated vector search.

a) The program assumes each grid point in turn in the asymmetric unit of the "symmetry function" is a possible heavy-atom site. (The symmetry function has the symmetry of the space group convoluted with a function representing the equivalent origins; eg in $P2_1$ the symmetry function is a single section (x,0,z) with x and z = 0 to ½, since the equivalent origins are x and z = 0 or ½, y = anything.)

b) The program obtains the Patterson densities at the Harker vectors and combines them to produce a measure of fit (arithmetic/harmonic mean or minimum function according to the option selected).

c) The user chooses a site from the symmetry function and the program generates its general equivalent positions.

d) Steps a and b are repeated except that the space-group asymmetric unit is computed and the program combines peaks at cross-vectors into the measure of fit, as well as Harker vectors.

e) The user chooses another site from the new map, and repeats from c until no new sites can be found with $P > \sigma(P)$.

4. Results of local scaling and vector search with gamma-crystallin 5.5A FHLE Patterson, EtHgCl derivative.

The figure shows the symmetry mean functions after anisotropic and after anisotropic+local scaling. The space group is $P4_12_12$ and the asymmetric unit is x = 0 to 1/4, y = 0 to 1, z = 0 to 1/8. The Patterson was sampled at $d_{min}/4$. The anisotropic scaled map has noise peaks on the 2-fold axes which are partially eliminated in the local scaled map, with no loss of signal.

The table shows peak heights at the 5 sites for the arithmetic and harmonic mean functions, for various values of the local search

radius, demonstrating that the harmonic mean function (HMF) gives a better signal to noise ratio than the arithmetic mean function (AMF), and is also more sensitive to incorrect or slightly misplaced sites. However the sensitivity decreases as the search radius is increased beyond 1 grid unit.

Reference

1. Mathews, B.W. & Czerwinski, E.W., Acta Cryst.,(1975), A31, 480-7.

Table showing results of vector search on 7-site derivative.

		Search radius = 0			S.r. = 0.9		S.r. = 1.5 grid units	
occ	B	Site	HMF	AMF	HMF	AMF	HMF	AMF
.26	9	1	31	34	36	37	38	39
.29	9	2	31	34	37	38	41	41
.09	18	3	20	29	31	34	36	37
.20	6	4	10	27	18	31	33	36
.39	4	5	12	28	28	32	34	36
		σ	4	18	8	26	22	38

Sensitivity to incorrect site

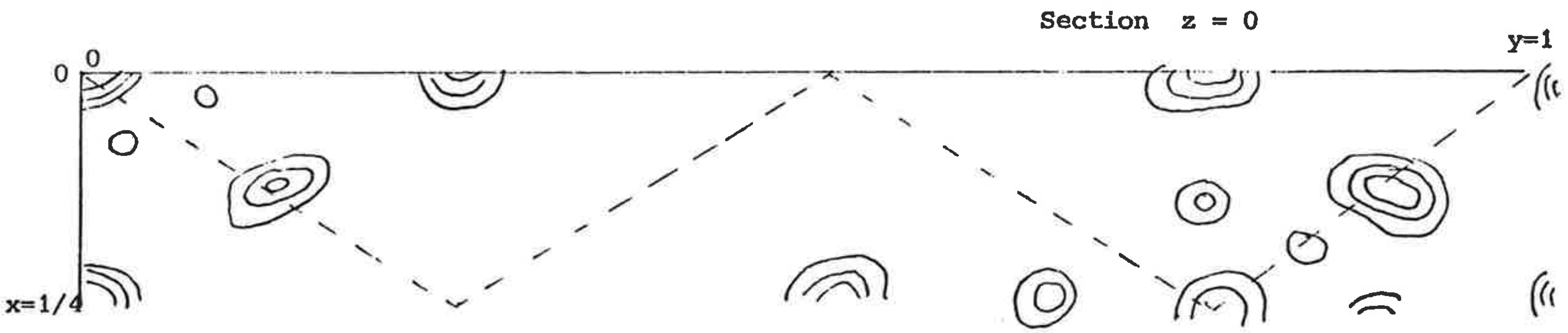
	S.r. = 0		S.r. = 0.9		S.r. = 1.5
Site	HMF	AMF	HMF	AMF	HMF
1	11	31	12	34	37
2	11	30	12	33	38
3	9	28	12	31	34
4	9	25	14	28	33
6*	3	17	10	21	24
5*	9	26	13	30	32
σ	4	18	6	24	25

*Site 6 was deliberately chosen to be inconsistent with the Patterson in place of site 5, which was not included in the calculation.

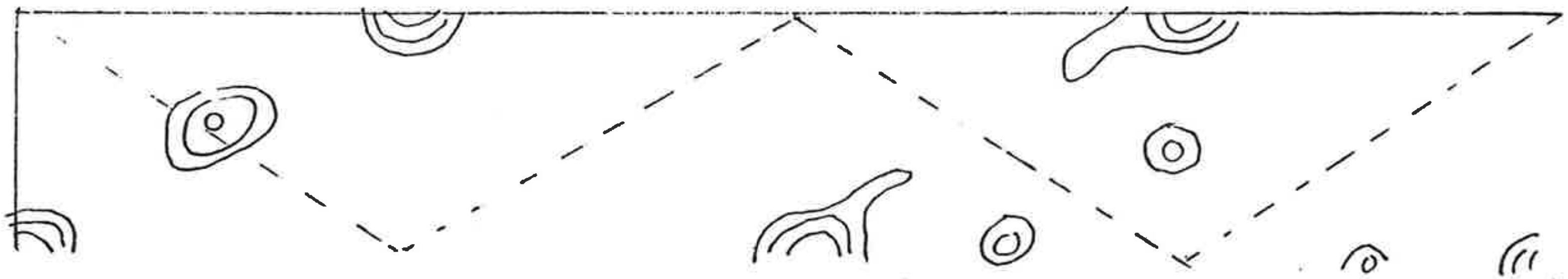
Sensitivity to slightly misplaced site

	S.r. = 0		S.r. = 1.5	
Site	HMF	AMF	HMF	AMF
1	15	32	38	39
2	23	31	41	40
3	18	30	38	38
4	7	25	33	35
5*	13	27	34	35
σ	4	18	25	39

*Site 5 moved 1.5 grid units from its correct position.



Symmetry mean functions (P4$_1$2$_1$2) after anisotropic (above) and anisotropic + local scaling (below)
Dashed lines indicate 2-fold axes (upon which there should be no signal peaks)



Use of Direct Methods in Locating Heavy Atoms

G L Taylor, Laboratory of Molecular Biophysics, Oxford

Most published papers on this subject are in the form of retrospective analyses [1,2,3,4], and useful though these are they point to the general conclusion that heavy atom constellations which can easily be solved using Patterson methods are solveable using Direct Methods (DM). Problems do arise however in the very cases where DM might be most useful, viz: multiple sites and high symmetry, where many solutions of equal probability exist.

That DM works at all may seem surprising when for example in the case of a single site, low space group symmetry heavy atom constellation we do not have a random distribution of scatterers. For Wilson's statistics to hold at least 10 randomly positioned atoms per unit cell are generally required [5].

A few ab initio studies have been successful [6,7,8], the most impressive being that of Shevitz et al. on tRNA. Therefore the method remains a useful adjunct to other methods and is certainly worth trying as the available 'black box' packages (MULTAN [8] and SHELX [9]) are computationally cheap and easy to use.

Below are listed some of the limitations and problems of the method.

Poor Estimates of F

Most previous studies have used isomorphous differences as estimates. F_{HLE}'s should in theory give better results, but this is not always the case [8]. Also observational errors in the moduli F_P and F_{PH} can lead to anomalously large differences. It therefore may be best to omit weak F_P's and F_{PH}'s.

Calculation of E's

Wilson statistics are not really valid, also the number of atoms in the unit cell (N) is probably unknown.

Series Termination Errors

Using low resolution data, only including high E's in DM as well as the fact that E's represent atoms as point scatterers.

How Many Heavy Atoms?

Not knowing N affects not only the calculation of E's, but also the probability expressions.

Whether phasing in 3-D using Sayre's equation (=Tangent formula for the case of one triplet):

$$\phi h = \phi k + \phi h\text{-}k$$

or in 2-D using the triple sign relationship:

$$s(Eh) = s(Ek)s(Eh\text{-}k)$$

the probability that a phase has a certain value (or sign) is:

$$P(\phi h) \propto N^{-\frac{1}{2}} |EhEkEh\text{-}k|$$

and since $< E^2 > = 1$ by definition, then $P \propto N^{-\frac{1}{2}}$, i.e. the probability is highly dependent on N [5]. [N.B. this essentially explains why whole proteins cannot be solved ab initio using DM] Therefore the larger N is, the broader is the probability distribution.

Wilson [4] has pointed out that this can be used to work to one's advantage: by overestimating N, more flexibility is allowed during the initial stages of phase propagation. Obviously N becomes very critical, but is worth experimenting with.

Discriminating Between Solutions

Often many equally probable solutions are produced and these must be discriminated between by extensive cross checking with Pattersons and with attempts at refining possible positions. It often happens that a solution is shifted from its true position because of the compound errors - sometimes too far away to refine. Navia and Sigler [3] overcame this problem by refining a constellation of heavy atoms constructed around a possible solution. Wilson [4] found when using MULTAN that the ABSFOM was the most sensitive amongst the figures of merit at pointing to the correct solution.

Advantages of DM may be summarised as below.

Ease of Use

SHELX and MULTAN are efficient, user friendly programs. SHELX contains an extremely powerful centrosymmetric package: for p starting reflections (where p is typically 12), phase propagation for all the 2-D reflections is attempted for each of the 2^P permutations of the starting set. Each is tested at an early stage for the efficacy of the propagation [9]. In a test which I ran on the Pt derivative of enodothia parasitica pepsin ($P2_1$) an essentially correct solution was obtained using 100 h01 reflections after omitting those with weak F_P or F_{PH}'s. The true solution was the only viable solution produced by SHELX and took 2 secs on the NAS 7000.

Real Space

Fourier maps are calculated using E's as coefficients and naturally represent real space.

Partial Structure Refinement

A possible solution can be cycled through Tangent refinement using the site to produce suitably weighted starting phases [3].

To sum up: DM remains a useful tool when used in a "combinatorial" approach with Patterson vector search methods and trial refinements.

References

1. Steitz, T.A., Acta Cryst. B24, 504-507 (1968)
2. Neidle, S. Acta Cryst. B29, 2645-2647 (1973)
3. Navia, M.A., Sigler, P.B. Acta Cryst. A30, 706-712 (1974)
4. Wilson, K.S. Acta Cryst. B34, 1599-1608 (1978)
5. Giaccovaco, C. Direct Methods in Crystallography, Academic Press.
6. Shevitz, R. et al. Science 177, 429-431 (1972)
7. Walkinshaw, M.D. et al. PNAS 77, 2400-2404 (1980)
8. Adams, et al. JMB 112, 183-197 (1977)
9. Germain, G., Main, P., Woolfson, M.M. Acta Cryst. A27, 368-376 (1971).
10. Sheldrick, G.M. SHELX Cambridge (1976)
11. Boeyans, J.C.A. Acta Cryst. A33, 863-864 (1977).

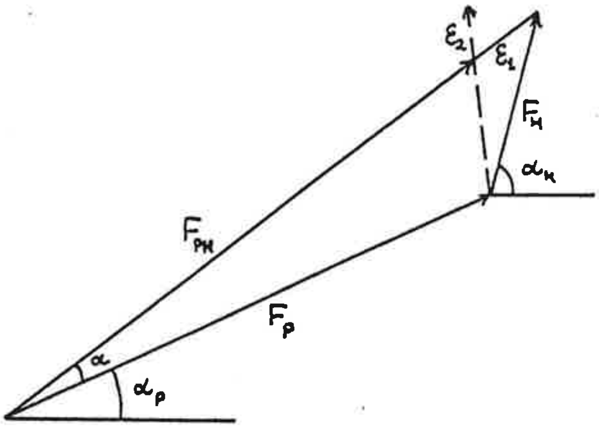
Schevitz, RW, Navia MA, Bantz, DA, Cornick, G, Rosa, JJ
Rosa MDH, + Sigler, PB (1972) Science 177 429-431

The refinement of heavy atom sites.

Phil Evans, MRC Laboratory of Molecular Biology, Cambridge

1. Introduction

All treatment of isomorphous replacement is centred on the phase triangle.



Two distinct methods of refinement have been used, based on minimizing different lack-of-closures in this triangle.

(a) 'Phase refinement'

This is very closely related to the phase calculation, and minimizes

$$\varepsilon_1 = |F_{PH}^{Obs}| - |F_{PH}^{Calc}|$$

$$= |F_{PH}^{Obs}| - |\underset{\sim}{F}_P + \underset{\sim}{F}_H|$$

In the phase calculation, we minimize $\Sigma\varepsilon_1^2$ summed over all derivatives for a given reflection, with respect to α_p, assuming $\underset{\sim}{F}_H$ is known from the known heavy atom parameters. In the phase refinement, we minimize $\Sigma\varepsilon_1^2$ summed over all reflections, with respect to the heavy atom parameters which define $\underset{\sim}{F}_H$, using the known most probable protein phase α_p.

In the classic alternate phase and refinement procedure, it has been assumed that $\underset{\sim}{F}p$ is independent of the heavy atom parameters, but this is not true if the derivative being refined was included in the phase calculation (see for example Blow and Matthews (1973)). This problem has been overcome by Gerard Bricogne, as discussed elsewhere.

(b) F_{HLE} refinement

This method is the one I shall mainly discuss here. We minimize $\Sigma\varepsilon_2^2$ with respect to the heavy atom parameters, where

$$\varepsilon_2 = |F_H^{Obs}|-|F_H^{Calc}|$$

As will be shown below, for acentric reflections, the anomalous difference of the derivative gives the angle α between F_p and F_{PH}, subject to an ambiguity, and hence gives an estimate of $|F_H^{Obs}|$. In the special case of centric reflections, α = 0 or 180°, so

$$\varepsilon_2 = |F_{PH} \mp F_P|- F_H^{Calc}$$

again with an ambiguity.

The lack of closures ε_1 and ε_2 are very similar for centric reflections, since

$$\varepsilon_1 = F_{PH} - |F_P{\pm}F_H^{Calc}|$$

Provided the correct choice of sign is made in each case, $\varepsilon_1 = \varepsilon_2$ for centric reflections, so refinement with centric reflections may equally well be cast in either form of refinement.

Table I summarizes the major differences between the various refinement methods.

Table I. Comparison of refinement methods.

<table>
<tr><th>3-dimensional F_{HLE} refinement</th><th>Centric refinement</th><th>Phase refinement</th><th>Gerard Bricogne's Phase refinement</th></tr>
<tr><td colspan="3">Uses reduced information</td><td>Uses all information</td></tr>
<tr><td colspan="2">Derivatives treated independently
Relative origins not determined</td><td colspan="2">Derivatives treated together
Relative origins determined</td></tr>
<tr><td>Fairly fast</td><td>Fast</td><td colspan="2">Slow</td></tr>
<tr><td>Dominated by centric data if present</td><td></td><td>Biased by common derivatives or common sites</td><td>Unbiased</td></tr>
<tr><td>Prerequisites:
Anomalous data</td><td></td><td colspan="2" rowspan="2">At least two derivatives, preferably independent</td></tr>
<tr><td colspan="2">Needs estimate of maximum F_H</td></tr>
</table>

2. F_{HLE} refinement

We wish to minimize

$$\Sigma w\varepsilon_2^2 = \Sigma w(|F_H^{Obs}|-|F_H^{Calc}|)^2$$

with respect to the parameters which define the heavy-atom derivative. Most of these parameters (position of sites, occupancy, temperature factor) affect F_H^{Calc}, but the relative scale and temperature factor between F_P and F_{PH} affect F_H^{Obs}. To do this refinement by conventional least-squares minimization, we need to:

(a) calculate F_H^{Obs}, rejecting those reflections for which the estimate is unreliable.

(b) calculate a weight = 1/variance (F_H^{Obs}).

(c) calculate the partial derivatives with respect to all the parameters.

Calculation of F_H and its variance

For centric reflections, the calculation of F_H is trivial.

$$F_{HLE} = F'_{PH} - F_P \qquad \text{lower estimate}$$

$$F_{HUE} = F'_{PH} + F_P \qquad \text{upper estimate}$$

where $F'_{PH} = K\exp(-Bs^2)F_{PH}$

The lower estimate F_{HLE} is usually the correct one, since for most reflections $F_H \ll F_P$, but reflections for which F_{HUE} is possible are best omitted.

For acentric reflections, an estimate for F_H may be calculated from the measured values of F_P, F_{PH} and the anomalous difference Δ. The derivations of the expressions are well known and are not given here (for reviews see Dodson (1976), Blundell and Johnson (1976) page 339).

The usual approximation is

$$F_H^2 \stackrel{\sim}{\sim} F_P^2 + F_{PH}^2 \mp 2\ F_P\ F_{PH} \sqrt{1 - t^2} \qquad (1)$$

$$\text{where } t = \sin\alpha = \frac{k\Delta}{2F_\rho}$$

$$k = \frac{f_H + f_H'}{f_H''} \quad \text{the anomalous ratio}$$

$$\Delta = F_{PH}^+ - F_{PH}^- \quad \text{the observed anomalous difference}$$

As in the centric case, there is an ambiguity of sign leading to two estimates F_{HLE} and F_{HUE} which must be resolved.

A simplified expression may be derived assuming the angle α is small

$$F_H^2 \stackrel{\sim}{\sim} \Delta_{iso}^2 + \left(\frac{k}{2}\right)^2 \Delta_{anom}^2 \qquad (2)$$

$$\text{where } \Delta_{iso} = F_{PH} - F_P$$

This a surprisingly good approximation, and is easier to think about: it points out the complementarity of the isomorphous difference and the anomalous difference.

However, a serious problem arises in the evaluation of F_{HLE} by either expression (1) or (2): the anomalous difference Δ is generally small compared with its error, and hence there is a systematic overestimation of Δ^2, which occurs in both expressions. The definition of variance may be rewritten

$$E(\Delta^2) = [E(\Delta)]^2 + \text{var}(\Delta) \qquad (3)$$

where E(Δ) denotes the expectation value of Δ (i.e. the mean).

In the worst case, if the true value of Δ is 0, the mean of Δ^2 from many measurements of Δ will be $<\Delta^2> = \sigma_\Delta{}^2$.

This may be generalized (Dodson, Evans and French (1975)) for a function $f(p_1,p_2,\ldots.p_n)$ of uncorrelated parameters p_i, with means $\bar{p}_i$ and variances $\sigma^2{}_i$

$$\mathrm{Var}[f(p_i)] \approx \sum_i \left(\frac{\partial f(\bar{p}_i)}{\partial p_i}\right)^2 \sigma_i^2 \qquad (4)$$

$$\mathrm{Bias}[f(p_i)] \approx \sum_i \tfrac{1}{2}\left(\frac{\partial^2 f(\bar{p}_i)}{\partial p^2}\right) \sigma_i^2 \qquad (5)$$

These expressions neglect higher order terms in a Taylor expansion.

If we have estimates of the errors of the observations, we can calculate the expected bias in F_{HLE} from equation (5), but note that this is a statistical correction, only correct on average over a large number of reflections. For instance if we consider the case of correcting Δ^2 for bias

$$\mathrm{Bias}\ (\Delta^2) = \sigma_\Delta^2$$

$$\Delta_{corr}^2 = \Delta^2 - \sigma_\Delta^2$$

but if $\Delta^2 < \sigma_\Delta^2$, a negative value of Δ_{corr}^2 is not sensible, and is perhaps better omitted.

(a) Derivation of bias and variance of F_{HLE} from expression (2).

$$I = F_{HLE}^2 = \Delta_{iso}^2 + \left(\frac{k}{2}\right)^2 \Delta_{anom}^2 \qquad (2)$$

from (4) and (5)

$$\mathrm{Var}(I) = \left(\frac{\partial I}{\partial \Delta_{iso}}\right)^2 \sigma_{iso}^2 + \left(\frac{\partial I}{\partial \Delta_{anom}}\right)^2 \sigma_{anom}^2$$

$$\mathrm{Bias}(F_{HLE}) = \tfrac{1}{2}\left(\frac{\partial^2 F_{HLE}}{\partial \Delta_{iso}^2}\right) \sigma_{iso}^2 + \left(\frac{\partial^2 F_{HLE}}{\partial \Delta_{anom}^2}\right) \sigma_{anom}^2$$

Now $\dfrac{\partial F}{\partial P} = \dfrac{1}{2F}\dfrac{\partial I}{\partial P}$ (6)

and
$$\frac{\partial^2 F}{\delta p^2} = -\frac{1}{2F^2}\frac{\partial F}{\partial p}\frac{\partial I}{\partial p} + \frac{1}{2F}\frac{\partial^2 I}{\partial p^2}$$

$$= -\frac{1}{4F^3}\left(\frac{\partial I}{\partial p}\right)^2 + \frac{1}{2F}\frac{\partial^2 I}{\partial p^2} \qquad (7)$$

So we need the first two derivatives of I with respect to Δ_{iso} and Δ_{anom}

$$\frac{\partial I}{\partial \Delta_{iso}} = 2\Delta_{iso} \quad ; \quad \frac{\partial^2 I}{\partial \Delta^2_{iso}} = 2$$

$$\frac{\partial I}{\partial \Delta_{anom}} = k^2\Delta_{anom}/2 \quad ; \quad \frac{\partial^2 I}{\partial \Delta^2_{anom}} = k^2/2$$

Hence $$\mathrm{Var}(F^2_{HLE}) = 4\left\{\Delta^2_{iso}\ \sigma^2_{iso} + \left(\frac{k}{2}\right)^4 \Delta^2_{anom}\ \sigma^2_{anom}\right\} \qquad (8)$$

$$\mathrm{Bias}(F_{HLE}) = \frac{1}{2F_{HLE}} \left\{\sigma^2_{iso}\ (1 - \Delta^2_{iso}/F^2_{HLE}) + (k/2)^2\ \sigma^2_{anom}\ [1 - \left(\frac{k\Delta_{anom}}{2F_{HLE}}\right)^2]\right\}$$

From Var (F^2_{HLE}), we can get

$$\sigma\ (F_H) = \frac{\sigma(F_H{}^2)}{2F_H} \qquad (9)$$

by equation (4), or use an alternative expression, which is probably better for small F_{HLE}:

writing $(\sigma\ (F) + F)^2 = \sigma\ (F^2) + F^2$.

then $$\sigma\ (F_H) = - F_H + \sqrt{F_H{}^2 + \sigma\ (F_H{}^2)} \qquad (10)$$

Equation (10) approximates to (9) when $F^2_H \gg \sigma(F^2_H)$

In the simple case of expression (2), it is possible to derive a more accurate expression for $\mathrm{Var}(F^2_{HLE})$ than that given by (4), assuming Δ_{iso} and Δ_{anom} are normally distributed, and including higher order terms.

$$\mathrm{Var}(F^2_{HLE}) = 4\ \Delta^2_{iso}\ \sigma^2_{iso} + 2\ \sigma^4_{iso}$$

$$+ \left(\frac{k}{2}\right)^4 \left\{4\ \Delta^2_{anom}\ \sigma^2_{anom} + 2\ \sigma^4_{anom}\right\} \qquad (11)$$

(b) Derivation of bias and variance of F_{HLE} from expression (1)

The derivation from expression (1) is more complicated: to quote Simon French "the differentiation is painful but seldom fatal."

$$I = F^2_{HLE} = F^2_P + F^2_{PH} - 2\ F_P\ F_{PH}\ q \qquad (1)$$

$$\text{where } q = \sqrt{1 - (k\Delta/2F_P)^2}$$

$$\frac{\partial}{\partial F_P}(F_P\ q) = 1/q$$

$$\frac{\partial}{\partial F_P}(1/q) = -\ (k\Delta/2)^2/(F_P\ q)^3$$

$$\text{So} \quad \frac{\partial I}{\partial F_P} = 2F_P - 2F_{PH}/q \qquad (12)$$

$$\frac{\partial^2 I}{\partial F_P{}^2} = 2 + \frac{F_{PH}\ k^2\ \Delta^2}{2(F_P\ q)^3} \qquad (13)$$

$$\frac{\partial I}{\partial F_{PH}} = 2F_{PH} - 2F_P\ q \qquad (14)$$

$$\frac{\partial^2 I}{\partial F^2_{PH}} = 2 \qquad (15)$$

$$\frac{\partial I}{\partial \Delta} = F_{PH}\ k^2 \Delta/2F_P\ q \qquad (16)$$

$$\frac{\partial^2 I}{\partial \Delta^2} = k^2\ F_{PH}/2F_P\ q^3 \qquad (17)$$

Hence $\mathrm{Var}(F^2_{HLE})$ and Bias (F_{HLE}) may be calculated, using (4) to (7) and (12) to (17).

The heavy atom refinement program which is currently in the Daresbury program suite calculates the bias of F_{HLE} from expression (1) as given above, but uses equations (10) and (11) for $\mathrm{Var}(F_{HLE})$,

substituting $\Delta^2_{anom} - \sigma^2_{anom}$ for Δ^2_{anom} in (11). The following checks are made for each reflection

(i) in (1) if $t > 1.0$ (i.e. $\sin \alpha > 1.0$), t is reset to 1.

(ii) of bias $(F_{HLE}) > F_{HLE}$ the reflection is rejected.

(iii) if $F_{HLE} > F_H$ max, the reflection is rejected.

(iv) of $F_{HUE} < F_H$ max, the reflection is rejected.

These last two checks are done against F_H max, the maximum expected heavy atom contribution, which maybe estimated as the largest heavy atom difference $F_{PH}-F_P$ in the appropriate resolution range. Check (iv) rejects reflections for which the upper estimate F_{HLE} is plausible, so the sign ambiguity in (1) is unresolved. Check (iii) mainly removes reflections with large spurious anomalous differences.

Weighting of reflections

The weight $1/\mathrm{Var}(F_{HLE})$ has two important consequences (see expression (11)).

(a) Centric reflections have much higher weights than acentric.

(b) Among acentric reflections, an F_H dominated by a large anomalous difference has a much lower weight than one dominated by the isomorphous difference.

i.e. the weighting scheme reduces the contribution of the anomalous measurements to the refinement.

3. General comments on heavy atom refinement

For heavy atom derivatives to be useful for phasing a protein, the most important thing is that the derivatives are reasonably isomorphous, and that all sites have been located and correctly placed. The exact values of the heavy atom parameters are much less important. The checks on the essential correctness are as follows:

(a) The Patterson map.

The Patterson is the only map derived directly from the observations, so it is important that it should be explained by the proposed solution, i.e. all substantial peaks on the Patterson should be accounted for, even if not all vectors appear in the map.

(b) Refinement statistics.

These are measures of agreement between an 'F_{Obs}' (F_H^{Obs} or F_{PH}^{Obs}) and the corresponding F_{Calc}. The agreement is often poor even in a usable derivative, and it is difficult to give rules for acceptable values of R-factors or correlation coefficients.

(c) Phasing statistics.

Correlation between α_P and α_H (which is always present) can indicate the sharpness of the probability distributions and errors in the derivative relative scale factor. The figure of merit is a measure of the sharpness of the probability distributions (i.e. precision, not accuracy), and is not a good indicator of the correctness of a structure.

(d) Double difference maps (residual difference).

Either basic method of refinement leads to a residual difference map with amplitudes $|F_{Obs}|-|F_{Calc}|$

for F_{HLE} refinement $(|F_H^{Obs}|-|F_H^{Calc}|)\exp(i\alpha_H)$

for phase refinement $(|F_{PH}^{Obs}|-|F_{PH}^{Calc}|)\exp(i\alpha_{PH})$

Ideally, these maps will show the difference between the true heavy atom structure and current model, i.e. it will show any sites which have been omitted. Note however that in 3-dimensional F_{HLE} refinement the refinement is weighted, while the residual difference map is not, and this leads to peaks on the positions of the sites included in the refinement.

(e) Cross-phased difference maps.

i.e. a map with amplitudes ($F_{PH}-F_P$), and phases αp calculated from one or more different derivatives. Because of phase correlations, these maps are liable to show 'ghost' peaks at the heavy atom sites of the derivatives included in the phasing, so derivatives with common sites must be considered carefully, checking the sites with the Patterson.

For checking derivatives, the centric and F_{HLE} methods which refine the derivatives independently are very useful. For the final parameters for phasing the protein, Gerard Bricogne's phase refinement is probably the method of choice, since this includes all the available information at once.

The different parameters defining a heavy atom derivative differ in how well they are determined by refinement. However, because of the close relationship between refinement and phase calculation, the parameters which are least well determined are the least important in the phasing (i.e. they have the least affect on F_H^{Calc}).

(a) Positions are easy to refine by any method (provided they are approximately right).

(b) Temperature factors are very uncertain.

One method of dealing with them is to refine the occupancy against shells of data on $\mathrm{Sin}^2\Theta/\lambda^2$, fixing the positions and setting the temperature factor = 0. The slope of ln(occ) against $\mathrm{Sin}^2\Theta/\lambda^2$ then gives the temperature factor. Alternatively such plots can be used to give an empirical atomic form-factor.

(c) Occupancies are highly correlated with temperature factors (occupancy is given by the intercept of the plot of ln(occ) versus $\mathrm{Sin}^2\Theta/\lambda^2$),and are also poorly determined.

(d) The relative scale and temperature factor are probably best determined by a final phase refinement, since this is closest to the phase calculation.

As a final warning, all derivatives are non-isomorphous. A great deal of effort and computer time can be invested in refining parameters without noticeable improvement in phasing. Refinement by different methods often leads to somewhat different parameters, which is mainly an indication of the uncertainty of isomorphous replacement.

References

1. Blow, D.M. and Matthews, B.W. (1973) Acta Cryst. A29, 56-62.

2. Blundell, T.L. and Johnson, L.N. (1976) Protein Crystallography, Academic Press (London).

3. Dodson, E.J. (1976) in Crystallographic Computing Techniques, ed. F.R. Ahmed, Munksgaard (Copenhagen), pages 259-268.

4. Dodson, E.J., Evans, P.R. and French, G.S. (1975) in Anomalous Scattering, ed. S. Ramaseshan and S.C. Abrahams, Munksgaard (Copenhagen), pages 423-436.

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

6. Kartha, G. and Parasarathy, R. (1965) Acta Cryst. 18, 745.

7. Matthews, B.W. (1966) Acta Cryst. 20, 230.

8. Singh, A.K. and Ramaseshan, S. (1966) Acta Cryst., 21, 279.

Use of Phases Calculated From Protein Atomic Positions

Guy Dodson, York University

Refinement of Heavy Atom Parameters

The phases determined by heavy atom isomorphous replacement have been traditionally used to analyse other heavy atom derivatives during the analysis of protein crystal structures. In the early stages of an analysis these phases can be seriously inaccurate and contain substantial systematic errors which can affect the refinement of heavy atom parameters - especially the occupany. These problems can be particularly acute for space groups such as R3, with no centric zones.

In an extension of the 2Zn insulin experimental phases from 2.8 to 1.9Å spacing, we used phases calculated from the atomic positions determined from the 2.8Å resolution map to refine the heavy atom parameters in the data sets between 2.8 and 1.9Å spacing. These phases were inaccurate (R ~ 48%), but they contained fewer systematic errors than those derived solely from the heavy atom derivatives. The refinement of the Cd sites in the Cd insulin derivatives and of the Pb sites in the Pb soaked derivatives converged smoothly. The occupancies in particular proved to be chemically sensible. Thus, as illustrated in Figure 1, the two axial Cd sites related by the local axis had each unit occupancy. In the Pb series the Pb site at the hexamer centre which always refined to a high occupancy with isomorphous phases, now refined to a value also near unity (Figure 2). The general Pb site (site 2 in Figure 2) occurs only in this particular derivative; it by contrast refines to a higher occupancy with the calculated phases showing that the weighted isomorphous phases contain more error than the calculated phase set.

In these calculations it was important either to select phases which, by the Sim criteria, were not likely to be wrong or to weight them with Sim weights.

The conclusion is that an inaccurate set of atomic parameters can give a very useful set of phases which can complement the usual techniques. With the increasing use of molecular replacement methods, it is worth emphasising that properly weighted phases derived from a reasonably good solution are likely to be sufficiently well determined to locate heavy atom positions in difference Fourier and bypass the difficulties sometimes (!) associated with Patterson functions. Thus, the production of experimental phases can be usefully accelerated.

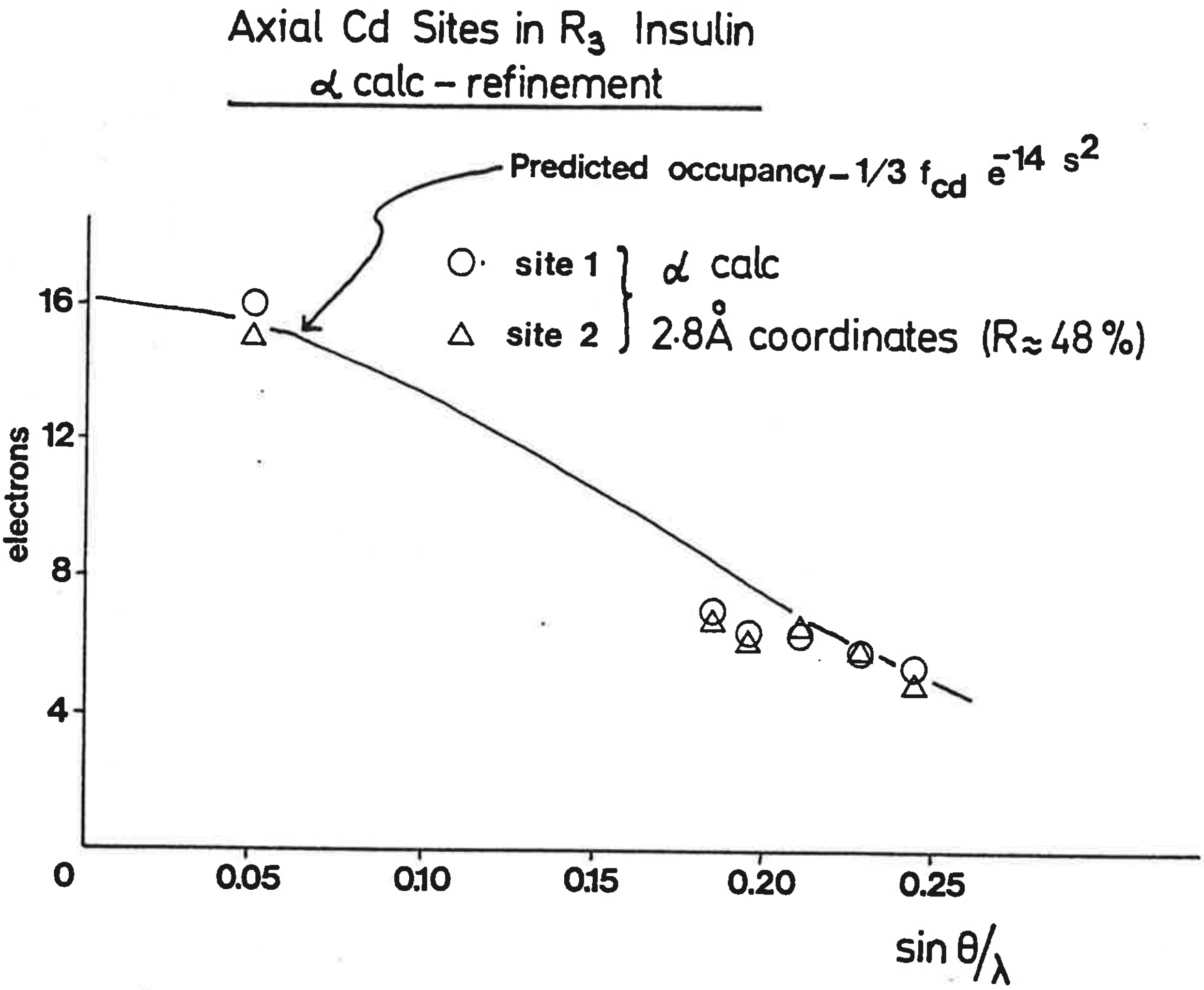
Analysis of Heavy Atom Interactions

There have been no detailed structural analyses of the structural changes in proteins associated with the binding and interactions of heavy atoms, and for very good reasons. However the protein-heavy atom structures are interesting, both from a chemical and structural point of view.

Iodination of tyrosine is a widely applied technique for labelling proteins. Although iodine is very bulky (not very different to a benzene ring), it often only partly reduces the reacted protein's potency. There are therefore good reasons to know more about how the addition of iodine affects the protein. Iodination of protein crystals has not usually produced useful derivatives, however - presumably because of the large steric affects of the iodine which destroy the isomorphism.

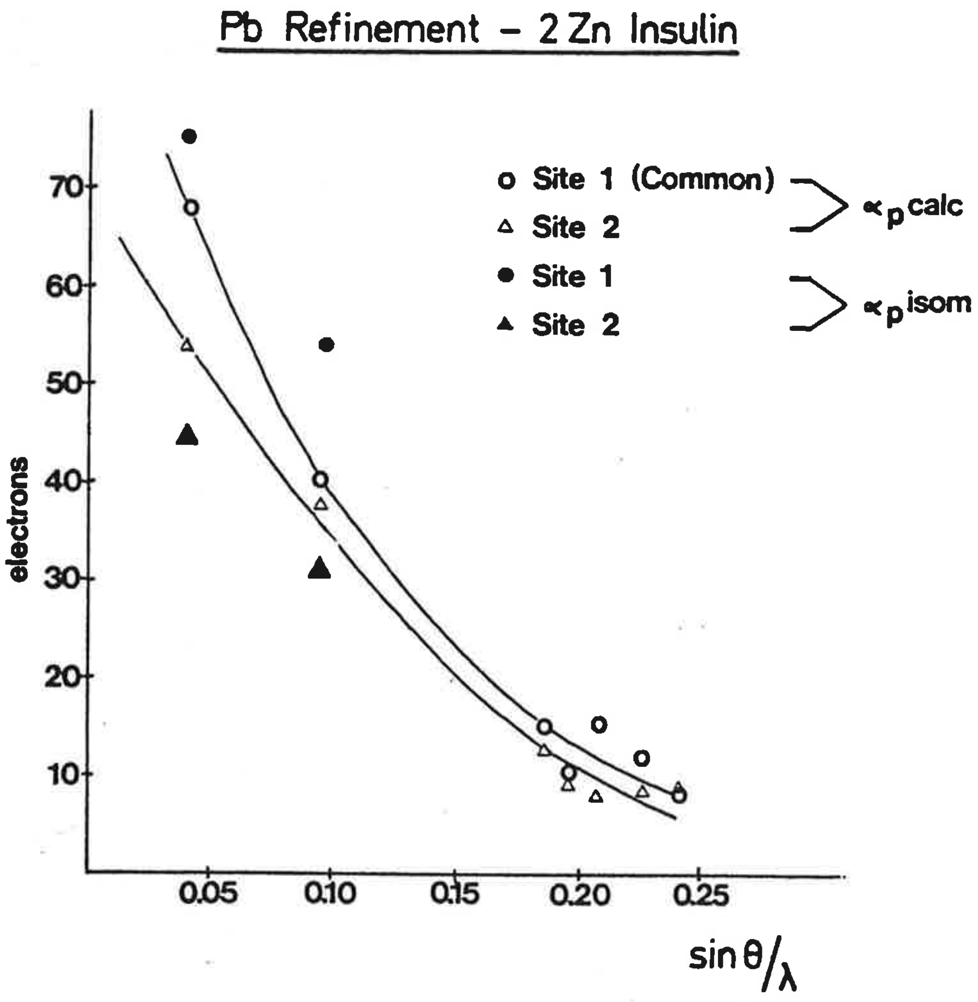
The second heavy atom derivative in barnase (a bacterial ribonuclease) was prepared by iodination of the crystals and we have begun a study on the structure. Analysis of the heavy atom positions showed that, for one site, all three molecules in the asymmetric unit appeared to have reacted equivalently. The iodinated derivative is being refined by fast Fourier least squares with 2.3Å spacing data. Thus, the positions of the iodines and the tyrosines can be determined with reasonable accuracy. Figure 3 illustrates the electron density at two of the iodinated tyrosines.

Inspection of the iodinated tyrosines shows that the surface tyrosines have reacted probably without much alteration of local structure. Fortunately for the use of the derivative in the phasing, only one edge of the majority substituted tyrosines is available for reaction. Where the packing around the tyrosine is closer however, the non-crystallographic equivalence is lost. Here the occupancy is reduced and there is evidence of movement in the surrounding protein. Detailed analysis of these must await the complete refinement of the iodinated enzyme - and the native enzyme for which the collection of high resolution data has just been completed.



Axial Cd Sites in R_3 Insulin
α calc – refinement
Predicted occupancy – 1/3 f_{cd} $e^{-14 s^2}$
○ site 1
△ site 2
α calc
2·8Å coordinates (R≈48%)
electrons
16
12
8
4
0
0.05
0.10
0.15
0.20
0.25
sin θ/λ

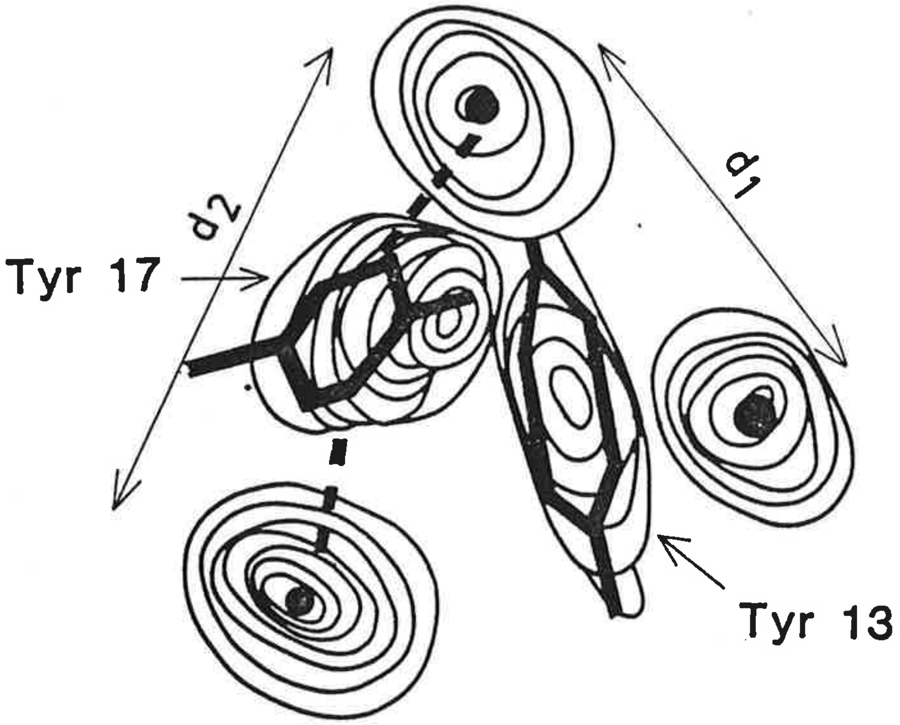
Fig 1



Pb Refinement - 2 Zn Insulin

Fig. 2

Molecule B (Barnase)



d_1 = sum of van der Vaals radii

d_2 = expected distance for diodinated tyrosine

Fig.3

An attempt to use the Bricogne heavy atom phasing-refinement procedure for 6-phosphogluconate dehydrogenase.

Margaret Adams, Laboratory of Molecular Biophysics, Oxford

6-phosphogluconate dehydrogenase (6-PGDH) is a dimeric protein of subunit molecular weight 50,000 (466 amino acid residues). It crystallises in $C222_1$ a = 72.72 Å, b = 148.15 Å, c = 102.91 Å with a single subunit per asymmetric unit. The starting point for the procedure was a nominally 2.6 Å map based on three heavy atom derivatives. The derivatives were independent : $KAu(CN)_2$ - 2 sites, $K_2Pt(CN)_4$ - 3 sites and $Pt(NH_3)_2Cl_2$ - 2 sites. There was a tendancy for heavy atoms to be in channels approximately parallel to the axes leading to serious heavy atom ripples. For the worst derivative, 50% of the data had $I \leqslant 3\sigma I$; generally this was true for 35% of the terms. Diffractometer data had been measured for the 6 Å sphere and oscillation camera data to higher resolution. Heavy atom parameters were refined independently for the two data sets using centric reflections only (there are three centric zones). The electron density map had three regions with ambiguous connections and it was unclear whether a sheet was all parallel or 'mixed'.

The film data from 6 - 2.6 Å was phased and refined by Gerard Bricogne as a 'one off' experiment. We did not go back and optimise the refinement after looking at the resulting electron density map. Doubtless we could have done better had we done so. Gerard started from occupancies half those determined in Oxford (as a result of a misunderstanding). The data set finally contained 8350 terms from 6 Å - 2.6 Å. The total observed data set contains 15800 terms.

The results can be viewed at two stages : (i) a comparison of the refinement parameters and (ii) a comparison of the electron density maps.

Refinement parameters

The occupancies returned quickly to those comparable to the Oxford refinement.

Detailed comparison was made with the centric refinements of (a) the film data to 2.6 Å and (b) the diffractometer data (∞ - 6 Å) with no common reflections. The positions and occupancies were closer to the low resolution than the high resolution refinement (see Table 1). Temperature factors were between 14 & 16 $Å^2$ compared with more improbable values with a tendancy to become negative found in the Oxford high resolution refinement. The overall scales varied by ∿ 2%. The ratio of fH/ε (Bricogne) was ∿ 2 compared with fH/E of ∿ 1 - 2 (Oxford refinement). The figure of merit was consequently much increased. (see Figure 1). At this stage the procedure looked promising.

Electron density maps

The basic problem is to define a way of deciding a map to be better, interpretability is the best criterion but the most obviously subjective. Four possible indicators are : (a) relative heights in the protein and solvent regions, (b) dominance of heavy atom ripples, (c) resolution of ambiguities in chain trace, (d) connectivity in regions not previously ambiguous. The first two are relatively more objective.

A technical problem arose. The omission of terms not phased in Bricogne's refinement led to an unrecognisable map. It was decided the original 'Oxford" F, α and m should be included where no 'Bricogne' term existed. There was, however, a discrepancy in weights since m was higher in the 'Bricogne' data set.

The map was compared with the original 'Oxford' map on the basis of the above four criteria. The noise in the solvent region, (a), was considerably worse, heavy atom ripples, (b), were worse. There was no improvement of the ambiguous regions, (c), connectivity in the previously unambiguous regions was also worse, (d). It seemed possible that using ε rather than E to define lack of closure might have increased the noise since terms with a large σ (I) tended to have high values of m. The map suffered from the incompatible weights of the two portions of the data : the 6 Å data had too low a weight since ε is always less than E (ε∿E/2 for these data).

Since the heavy atom parameters seemed to have improved, an attempt was made to use them. Phases were recalculated using a conventional Blow-Crick procedure and the 'Bricogne' derived parameters. As compared with the earlier phasing fH/E improved at low angle and became worse at high angle (as might be expected from including the measurement error for F_p in the refinement. The resulting electron density map was different from the original one but in no way better (nor probably worse). It was concluded that the Bricogne refinement was not the most promising way of improving the electron density map of 6-PGDH. None the less, the heavy atom parameters were probably a better fit to the data than those obtained by other refinements. A redefinition of error in the term and of m might have improved the method.

It is interesting to note that an interpretable electron density map (in which the sequence can be followed for all but the N-terminal 30 residues) has been obtained using essentially Bhat & Blow's electron density modification (Acta Cryst. A 38, 21-29, (1982)).

COMPARISON OF HEAVY ATOM PARAMETERS ON VARIOUS REFINEMENTS

Derivative	$K_2Pt(CN)_4$			$KAu(CN)_2$		$Pt(NH_3)_2Cl_2$	
Site	1	2	3	1	2	1	2
Occupancy (electrons)							
Bricogne	83.6	85.1	57.7	91.4	71.0	101.3	71.3
Oxford high resol^n	89.5	26.1	44.2	42.4	69.7	93.3	34.7
Oxford low resol^n	81.8	62.3	72.7	117.8	58.8	102.9	71.3
Temperature factor							
Bricogne	16.3	15.6	15.3	15.6	15.4	14.3	15.1
Oxford high resol^n	1.0*	20.0	7.5	1.0*	1.0*	1.0*	20.0
Oxford low resol^n (not refined)	(15.0)	(15.0)	(15.0)	(15.0)	(15.0)	(15.0)	(15.0)
Difference in occupancy							
Bricogne-Oxford 'high'	-5.9	+59.0	+13.5	+49.0	+1.3	+8.0	+36.6
Bricogne-Oxford 'low'	+1.8	+22.8	-15.0	-26.4	+12.2	-1.6	0
Oxford 'high'-Oxford'low'	+7.7	-36.2	-28.5	-75.4	-10.9	-9.6	-36.6
Difference in position (Å)							
Bricogne-Oxford 'high'	.38	.95	2.04	.72	.59	.92	.43
Bricogne-Oxford 'low'	1.50	.71	.73	.37	.33	.39	.75
Oxford 'high-Oxford' low'	.86	1.47	2.26	.66	.58	1.17	.91

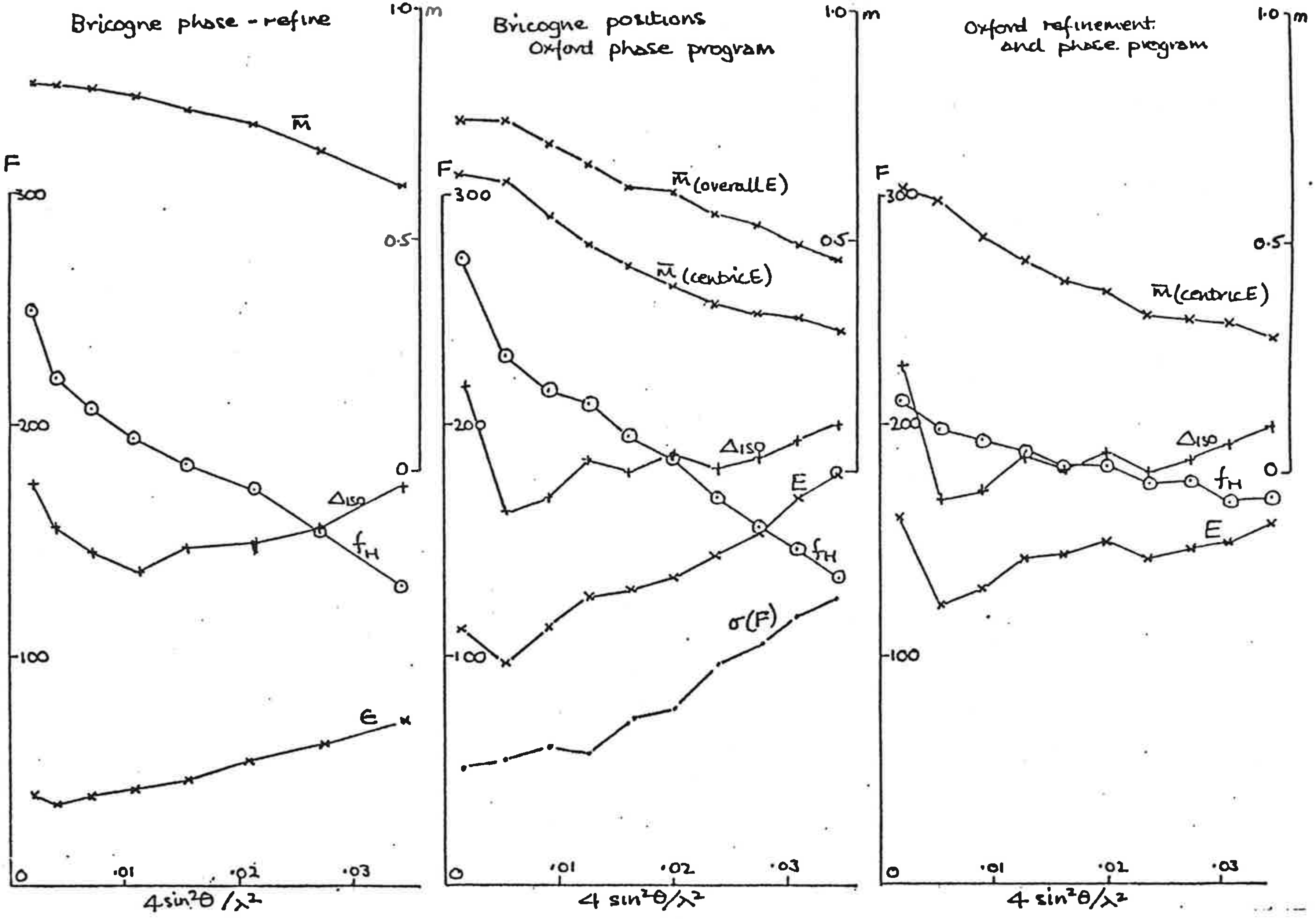
"Bricogne" - G Bricogne phase refinement on 8339 terms- film data
6 Å > d > 2.6 Å

Oxford high resol^n - Centric refinement on film data
∞ - 6 Å very incomplete
6 Å - 2.6 Å 'complete'

Oxford low resol^n - Centric refinement on diffractometer data ∞ - 6 Å.
B not refined.

Relative scalefactors between refinements ∿ 2%.

PHASE PROGRAM $K_2Pt(CN)_4$ DERIVATIVE



Peaks and Holes at Heavy Atom Sites

Anne Bloomer, MRC Laboratory of Molecular Biology, Cambridge

A protein electron-density map shows peaks or holes at a site of heavy atom substitution if the protein phases are biased towards, or away from, those calculated for this heavy atom. Ideally the histogram of the number of reflections versus $|\alpha_P - \alpha_H|$ (modulo 180°) should be flat and the protein map featureless at all sites of substitution. In practice, the histogram is usually slightly concave, even after removing any bias due to centric data. Provided that the peaks at each end of the distribution are approximately the same, this does not usually require special treatment. However, if the two ends of the distribution are grossly unequal, unwanted peaks or holes will appear in a protein map (and conversely, holes or peaks in a double difference map). The final parameters of the heavy atom(s) must be changed from their refined values, in order to eliminate such features. It is shown here that the most effective change is that made to a derivative scale factor.

Positional parameters of a heavy atom are only rarely difficult to refine. Any errors here show as pairs of closely adjacent peaks and holes, whose relative disposition shows the direction and magnitude of the error. The scale factor, which affects all sites of one derivative and the site occupancies have the most direct effect on any bias of the phases. Temperature factors, whether used in isotropic scaling of derivative to native or used as individual atom parameters, are so closely correlated with the scale factor and occupancy respectively that they are not considered separately here. They should be investigated by means of refinement in shells of resolution.

Any inadequacy in the protein phase determined by isomorphous replacement will effectively add a vector $\underline{F'}$ onto the true value of $\underline{F_P}$. The component of $\underline{F'}$ parallel to $\underline{f_H}$ determines the error in the electron density map which accumulates at the heavy atom position. Using the nomenclature of the phase triangle shown in Fig. 1:

$$\underline{F'} = \underline{F_P}\text{ (apparent)} - \underline{F_P}\text{ (true)}$$

$$\text{ERROR} = \text{component of } \underline{F'} \text{ parallel to } \underline{f_H}$$

$$\sim F' \operatorname{Sin}\phi$$

$$= 2\, F_p . \operatorname{Sin}\phi . \operatorname{Sin}\delta/_2 .$$

$$\sim F_p\,(\operatorname{Cos}\phi - \operatorname{Cos}\phi')$$

This approximation is valid whatever the cause of the error giving rise to F' and whatever the value of the angle ϕ. Two extreme cases are now considered.

Scale Factor

If k_{PH} is over-estimated by a fraction Δ, then the triangle with side F_{PH} is replaced by one with $F_{PH}\ (1+\Delta)$. Thus:

$$F_{PH}^{\ 2} = F_P^{\ 2} + f_H^{\ 2} - 2F_p.f_H\mathrm{Cos}\phi$$

$$(1+\Delta)^2F_{PH}^{\ 2} = F_P^{\ 2} + f_H^{\ 2} - 2F_p.f_H\mathrm{Cos}\phi'$$

Therefore

$$F_p(\mathrm{Cos}\phi-\mathrm{Cos}\phi') = (2\Delta F_{PH}^{\ 2} + \Delta^2F_{PH}^{\ 2})/2f_H$$

Therefore

$$\mathrm{ERROR} \sim \frac{F_{PH}^{\ 2}.\Delta}{f_H}\ (1 + \Delta/2)$$

This quantity is positive definite, irrespective of whether the angle ϕ is acute or obtuse; i.e. lengthening the side F_{PH} of a triangle always increases the opposite angle ϕ. Thus for every reflection HKL an over-estimate (under-estimate) of the scale k_{PH} gives a positive (negative) ERROR and thus a peak (hole) in ρ protein.

Occupancy

If the occupancy is over-estimated by a fraction Δ then the triangle with side f_H is replaced by one with the longer $f_H\ (1 + \Delta)$. This may increase or decrease ϕ depending upon the geometry of the phase triangle. Thus:

$$F_{PH}^{\ 2} = F_P^{\ 2} + f_H^{\ 2} - 2\ F_P.f_H.\cos\phi \qquad (1)$$

$$= F_P^{\ 2} + f_H^{\ 2}(1 + \Delta)^2 - 2\ F_P.f_H.(1 + \Delta)\ \cos\phi' \qquad (2)$$

Thus:

$$2\ F_P.f_H.(1 + \Delta)\ \cos\phi = (F_P^{\ 2} + f_H^{\ 2} - F_{PH}^{\ 2})(1 + \Delta)$$

$$2\ F_P.f_H.(1 + \Delta)\ \cos\phi' = (F_P^{\ 2} + f_H^{\ 2}(1 + \Delta)^2 - F_{PH}^{\ 2})$$

Therefore:

$$2\ F_P . f_H .(1 + \Delta)(\cos\phi - \cos\phi') = \Delta(F_P{}^2 + f_H{}^2 - F_{PH}{}^2) - \Delta(2f_H{}^2 + \Delta f_H{}^2)$$

$$F_P(\cos\phi - \cos\phi') = \frac{\Delta(F_P{}^2 - f_H{}^2 - F_{PH}{}^2) - \Delta^2 f_H{}^2}{2\ f_H\ (1 + \Delta)}$$

$$\text{ERROR} = \frac{\Delta\ [\ F_P{}^2 - F_{PH}{}^2 - (1 + \Delta)\ f_H{}^2\ ]}{2\ f_H\ (1 + \Delta)} \qquad (3)$$

$$= \frac{\Delta\ [\ f_H{}^2 - 2\ F_{PH} . f_H .\ \cos\gamma - (1 + \Delta)\ f_H{}^2\ ]}{2\ f_H\ (1 + \Delta)}$$

$$= \frac{-\Delta}{2\ f_H\ (1 + \Delta)}\ [\ \Delta\ f_H{}^2 + 2\ F_{PH} . f_H\ \cos\gamma\ ]$$

$$= \frac{-\Delta}{2\ (1 + \Delta)}\ [\ \Delta\ f_H + 2\ F_{PH}\ \cos\gamma\ ]$$

The sign of this error varies with the geometry of the phase triangle via the dependence of cosγ. Considering the average over all HKLS, and applying Wilson statistics to (3) we have:

$$<\text{ERROR}>_{hkl} = \frac{\Delta}{2\ (1 + \Delta)\ <f_H>}\ [<F_P{}^2 - F_{PH}{}^2> - (1 + \Delta) < f_H{}^2 >]$$

$$= \frac{-\Delta\ (2 + \Delta)\ <f_H{}^2>}{(1 + \Delta)\ <f_H>}$$

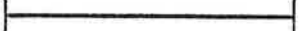
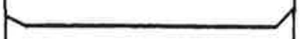
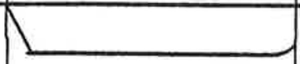
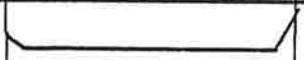
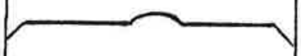
$$= \frac{-\Delta\ <f_H{}^2>}{<f_H>}\ (2 - \frac{\Delta}{1 + \Delta})$$

This expression is negative definite and thus, on average over all HKLS an over-estimate (under-estimate) of the occupancy and thus of f_H gives a negative (positive) ERROR and thus a hole (peak) in ρ protein.

Summary

The distributions of the value of the angle ϕ (i.e. $|\alpha_P - \alpha_H|$) as typically observed, for acentric data, are shown in the table. Errors in the scale factor give a bias arising from every reflection whereas errors in the occupancy give rise to an opposite bias, only when averaged over all reflections under conditions where Wilson's statistics are valid.

TABLE

Distribution	Histogram	Diagnosis
Ideal		
Normal		Usually OK
Positive bias		Scale too high. (Occupancy too low)
Negative bias		Scale too low. (Occupancy too high)
Convex		Rarely observed. Occupancies wrong (A.J. Wonacott, verbal report at the CCP discussion).

0° 180°

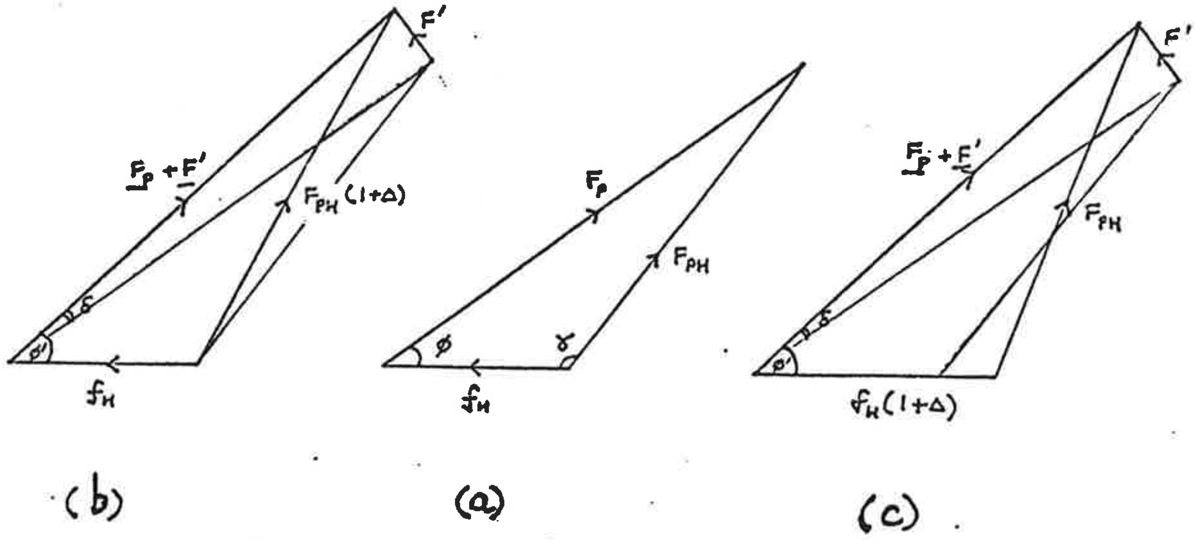


Figure 1. The phase triangle: (a) Idealised case using true amplitudes, (b) Scale factor for F_{PH} increased by $(1+\Delta)$, (c) Occupancy of heavy atom site(s) increased by $(1+\Delta)$. In this case Ø' may be greater then or less than Ø , whereas in (b) Ø' is always the larger angle.

HEAVY ATOM PARAMETER REFINEMENT

A.G.W. LESLIE, IMPERIAL COLLEGE, LONDON

1) OBJECTIVE

A new algorithm has recently been introduced by Gerard Bricogne to refine heavy atom parameters using acentric reflections and so-called "phase refinement" (Bricogne, 1982). At present there are no published results suggesting that this new algorithm is in practice superior to the conventional "phase refinement" algorithm, and in at least two cases (the orthorhombic crystal form of Glyceraldehyde 3-phosphate Dehydrogenase being worked on at Imperial and 6-phospho-gluconate dehydrogenase (M.Adams and coworkers, Oxford), heavy-atom parameter refinement using the new algorithm has not produced significantly improved MIR phases as judged by the resulting electron density maps.

In order to make an objective assessment of the Bricogne algorithm, and to investigate the effect of non-isomorphism in the derivative data, a series of test refinements have been conducted using model data. Although the relevance of results obtained using model data to real problems can sometimes be called into question, in this particular instance it seems reasonable to suggest that if the algorithm is to be generally useful it must at least prove effective in a trial using model data, when all the sources of error are under direct control.

2) GENERATION OF THE MODEL DATA

All tests were conducted on a model of the orthorhombic crystal form of Glyceraldehyde 3-phosphate dehydrogenase which crystallises in space group $P2_12_12$ with a tetramer of total molecular weight 140,000 in the crystallographic asymmetric unit. A set of coordinates derived from the the 2.7Å structure determination of the monoclinic crystal form were used to generate perfect native data, using the program GENED to generate the model electron density and Fourier transforming this density to produce a set of native structure factors. Ideal derivative data were calculated by vectorial addition of calculated heavy atom and native structure factors. Data were prepared for a mercury and a platinum derivative. For each derivative there were a total of four heavy atom sites per tetramer, two fully occupied and two half occupied. The heavy atoms were positioned at actual heavy atom sites determined for this crystal form of GAPDH.

In order to produce "non-isomorphous" derivative data, a new set of native structure factors was calculated corresponding to a structure in which the entire tetramer had been rotated by one degree in the unit cell. This rotation led to a maximum shift in atomic position of 0.65Å and a mean shift of 0.28Å with respect to the unrotated molecule. This "non-isomorphous" native data was combined

with the calculated heavy atom structure factors for the mercury sites to produce a "non-isomorphous" mercury derivative dataset. The "non-isomorphous" platinum derivative data were produced in the same way, but the sense of the rotation applied to the molecule when generating the native data was reversed.

Finally, random errors were applied to all the datasets. The errors were generated using the Gaussian random number generator in the NAG library (GØ5CBF) with standard deviations derived from an analysis of an actual GAPDH native data set. This analysis showed that there was a gradual increase in the standard deviation as a function of |F|. The standard deviations were independant of resolution except in the smallest |F| bin, where a significant increase with increasing resolution was apparent. This feature was incorporated in generating the model errors.

3) STARTING PARAMETERS

The same set of starting parameters was used in all the refinement tests. Each heavy atom was peturbed by 1.ØÅ from its true position. Starting occupancies are listed in Table 1. Derivative scale factors were set to the ideal values.

4) REFINEMENT TESTS

All test refinements were carried out using the program PHARE on the NAS machine at Daresbury. Data in the resolution range 2ØÅ to 6Å were included, in order to avoid excessive use of computing time. Each test was carried out twice, once using the conventional algorithm and once using the Bricogne algorithm. The refined parameters were the heavy atom positions and occupancies and the derivative scale and overall temperature factors (a total of 36 parameters). Phases were calculated during every cycle of refinement, and the lack of closure values were also updated on every cycle. The following tests were performed:

a) Using isomorphous derivative data.

b) Using non-isomorphous derivative data.

c) Using isomorphous platinum data and non-isomorphous mercury data.

The following points are pertinent to the refinements.

(i) Refining with isomorphous data

The Bricogne algorithm depends for its success on a predominantly unimodal phase distribution. In order to achieve this in practice, a figure of merit cutoff is applied to reflections included in the refinement. This cutoff was set to Ø.5 for this test, and this explains the difference in the number of reflections included in the refinement using the different

algorithms, as no such criterion was applied to the conventional refinement. Five cycles of refinement were carried out in each case.

(ii) Refining with non-isomorphous derivative data
In this test, a figure of merit cutoff of Ø.8 was applied to both the Bricogne and the conventional refinements. A significantly lower cutoff (Ø.5) gave extremely poor convergence for both types of refinement.
A total of 12 cycles of refinement were more than sufficient for convergence.

(iii) Refining with isomorphous platinum and non-isomorphous mercury derivatives.
The figure of merit cutoff was again Ø.8. Seven cycles of refinement produced convergence.

The main results of these tests are summarised in Table 1.

5) ASSESSMENT OF FINAL PARAMETER VALUES

There are several ways of assessing the quality of the final refined parameters. Most simply, the deviations of the refined parameters from their true values can be examined. However, it is difficult to determine from this comparison alone how the final MIR phases will be affected.Therefore DIR phases were calculated using each set of refined parameters, and these phases were compared with the true native phases. In each case the mean, rms and weighted rms phase differences were calculated. The results are presented in Table 2. It is apparent that both the refined parameters and the DIR phases are insignificantly different for the two algorithms employed. A direct comparison of the two sets of DIR phases obtained for the non-isomorphous test (case (ii)) gave mean, rms and weighted rms phase differences of 6.1,14.9 and Ø.4 degrees.

Another criterion on which the refinements can be compared is the speed of convergence. Indeed, one of the principle drawbacks of the conventional phase-refinement method is its poor rate of convergence (Blow and Matthews, 1973). This has been ascribed to the failure to allow for the correlation between the heavy atom parameters and the phases used in the refinement. Because this correlation is explicitly accounted for in the Bricogne algorithm, one might expect that this algorithm would produce significantly faster convergence. The rate of convergence of two representative parameters is illustrated in Figures 1 and 2. It is apparent that the rate of convergence is slightly faster using the Bricogne algorithm in the case of the isomorphous derivative data, but when using non-isomorphous derivative data even this slight advantage is lost.

Finally, Fourier maps were calculated using the different sets of DIR phases, and the heavy atom positions were examined for evidence of large peaks or troughs in the electron density. However, none of the phase sets produced either a peak or a trough that was above the electron density maxima and minima for the protein, and it was therefore not possible to differentiate between the quality of different sets of phases on this basis.

6) ADDITIONAL REFINEMENT TESTS

In order to provide additional criteria on which to assess the Bricogne algorithm, two further tests were contrived. In the first, an extra fifth site was added to each derivative at full occupancy and with the same atomic coordinates (ie a common site). It has often been found in practice that the occupancy of common sites is overestimated. In the second test, a false site was added to the starting parameters of each derivative, at half occupancy. The occupancy of this false site should refine to zero.

Each test was conducted with both isomorphous and non-isomorphous derivative data, using the conventional algorithm and the Bricogne algorithm. The results are presented in Table 3. Again, there is no suggestion that the Bricogne algorithm is superior, although it is perhaps suprising that the occupancy of the common site in the first test is not in fact over-estimated.

7) CONCLUSIONS

Apart from a marginal gain in the rate of convergence with isomorphous derivative data (a gain which is achieved at significantly greater computational cost because the Bricogne algorithm requires the full normal matrix), the Bricogne algorithm did not produce significantly better refined heavy atom parameters based on any of the four criteria applied (parameter values, DIR phases, rate of convergence and peaks or troughs at heavy atom positions). It is conceivable that under a different set of conditions the algorithm may perform rather better, but the test results suggest that this method cannot in general be relied upon to provide refined parameters which are superior to those obtained using conventional phase refinement.

Acknowledgements. I would like to thank Dr. P. Brick and Dr. A.J. Wonacott for many useful discussions, and Dr. E. Dodson for suggesting the tests described in section 6.

References

Bricogne, G. in Computational Crystallography. D. Sayre, ed., Clarendon Press, 1982.

Blow, D.M. and Matthews, B.W. (1973) Acta Cryst. A29,p56-62.

TABLE HEADINGS

Table 1. Initial and refined heavy atom parameters for the three tests described in section 4. GB refers to results obtained using the Bricogne algorithm. Δr is the deviation between the true and refined heavy atom positions in Å, occ. is the heavy atom occupancy.

Table 2. Comparison of MIR phases calculated using different sets of refined heavy atom parameters. The comparison is always with the true (calculated) native phases.

Table 3. Additional refinement tests introducing a fifth site. For details see section 6.

FIGURE CAPTIONS

Figure 1. Occupancy shifts for the first mercury and first platinum sites as a function of cycle number. The shifts for the mercury site are negative, those for the platinum are positive. The full lines represent the refinement using the Bricogne algorithm, and the dotted lines using the conventional algorithm. The refinements were performed using isomorphous derivative data.

Figure 2. As Figure 1, but refining with the non-isomorphous derivative data. .

TABLE 1

			Ideal Occupancy	Starting Parameters		(I) Isomorphous Data				(II) Non-Isomorphous				(III) Mercury non-Isomorphous Platinum Isomorphous			
						G.B.		conventional		G.B.		conventional		G.B.		conventional	
	Derivative	Site		occ.	Δr	occ.	Δr	occ.	Δr	occ.	Δr	occ.	Δr	occ.	Δr	occ.	Δr
	Hg	1	1.0	1.5	1.0	1.009	.05	1.031	.04	1.194	.10	1.281	.22	1.144	.23	1.132	.12
	Hg	2	1.0	1.5	1.0	1.043	.02	1.045	.02	1.295	.31	1.305	.30	1.297	.22	1.280	.25
	Hg	3	0.5	0.75	1.0	0.506	.14	0.523	.16	0.587	.53	0.508	.51	0.609	.40	0.549	.36
	Hg	4	0.5	0.75	1.0	0.516	.15	0.523	.11	0.693	.45	0.748	.31	0.746	.36	0.674	.44
Hg Overall	Total excess occupancy rms positional shift			1.5	1.0	0.074	.11	0.122	.10	0.769	.38	0.842	.35	0.796	.31	0.635	.32
	Pt	1	1.0	0.5	1.0	1.036	.04	1.050	.03	1.230	.27	1.221	.18	1.067	.07	1.065	.06
	Pt	2	1.0	0.5	1.0	1.012	.02	1.014	.06	1.206	.13	1.259	.21	1.024	.07	1.041	.06
	Pt	3	0.5	0.5	1.0	0.508	.06	0.521	.09	0.640	.85	0.695	.77	0.540	.11	0.532	.10
	Pt	4	0.5	0.5	1.0	0.544	.06	0.545	.07	0.616	.52	0.674	.53	0.563	.14	0.575	.11
Pt Overall	Total excess occupancy rms positional shift			-1.0	1.0	0.100	.05	0.130	.07	0.692	.52	0.849	.49	0.194	.10	0.213	.09
All sites	Total excess occupancy rms positional shift			0.5	1.0	0.174	.08	0.252	.08	1.461	.46	1.691	.42	0.990	.23	0.848	.23
Number of reflections						3578		4282		1360		1078		2051		1752	

TABLE 2

	(i) Isomorphous Data			(ii) Non-isomorphous Data		(iii) Mercury non-isomorphous Platinum isomorphous	
	Starting Parameters	G.B.	conventional	G.B.	conventional	G.B.	conventional
mean phase difference	40.6	26.2	26.2	45.9	45.9	36.0	36.2
rms phase difference	60.3	42.6	42.6	65.5	65.8	54.9	55.1
weighted rms phase difference[1]	25.4	10.1	10.1	33.5	34.0	17.7	17.4
figure of merit	0.65	0.77	0.77	0.63	0.63	0.70	0.70

Footnote

[1] weight $= \frac{1}{(1 - m^2)}$

where m is the figure of merit.

TABLE 3

	Occupancy of the fifth site							
	Isomorphous data				Non-isomorphous data			
	G.B.		Conventional		G.B.		Conventional	
	Hg	Pt	Hg	Pt	Hg	Pt	Hg	Pt
(a) Common site, true occupancy 1.0	1.041	1.041	1.055	1.051	1.017	1.130	1.077	1.043
(b) False site, input occupancy 0.5	0.019	0.089	0.003	0.009	0.164	0.049	0.127	0.103

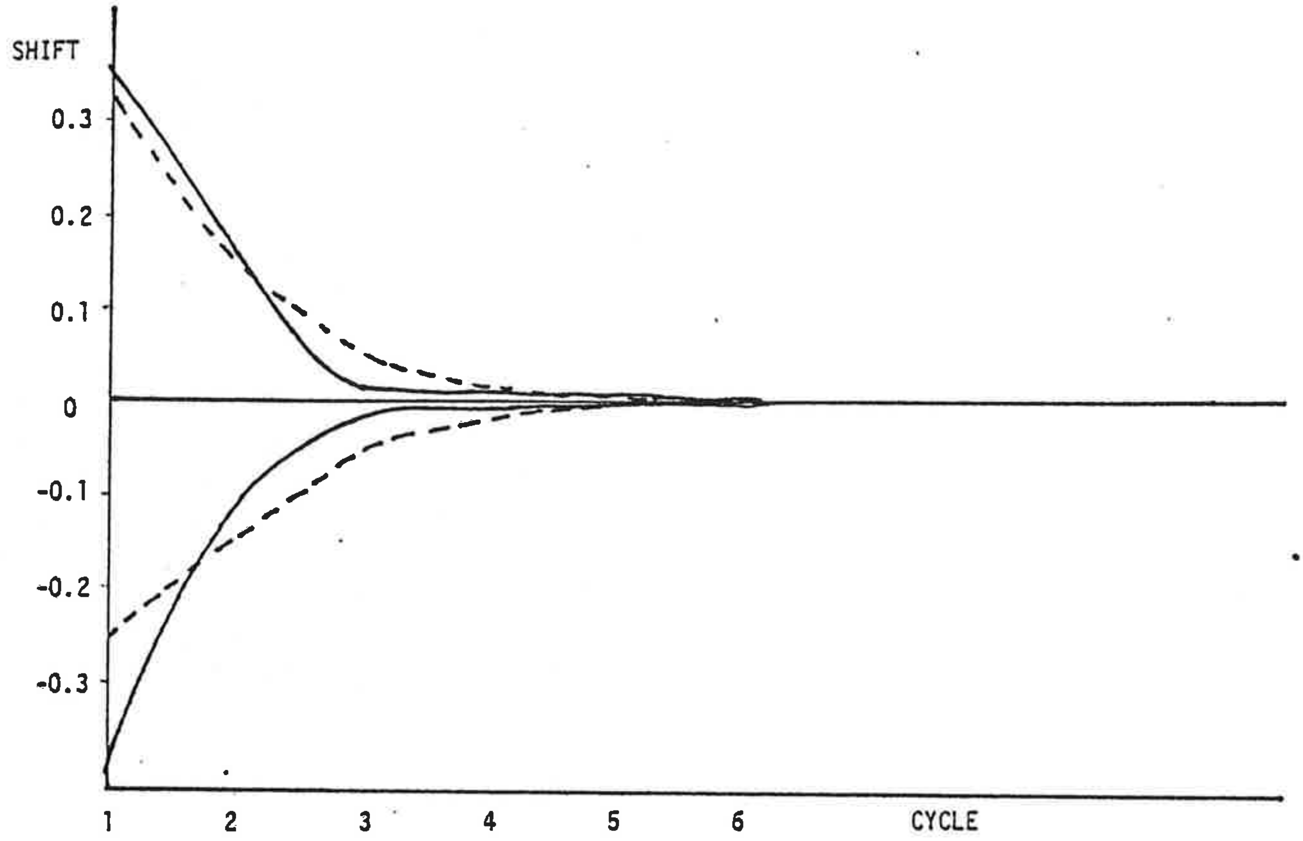

SHIFT
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
1
2
3
4
5
6
CYCLE

Fig.1

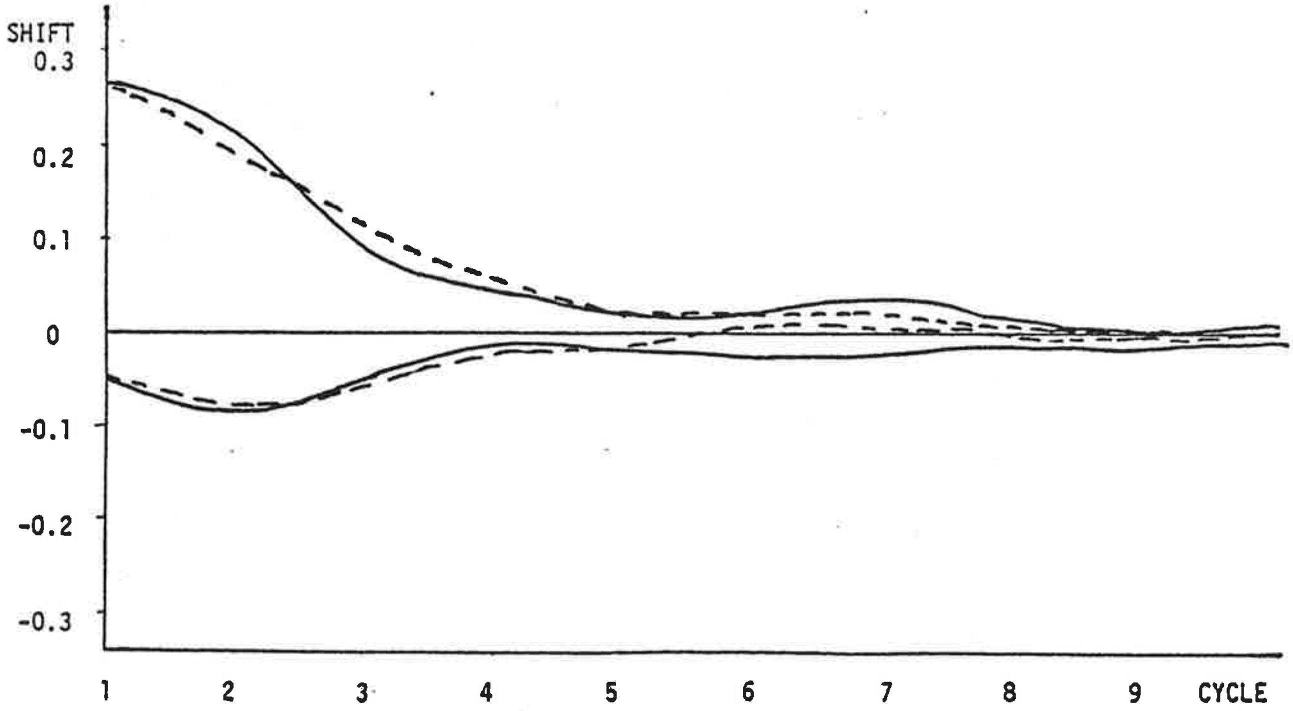

SHIFT
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
1
2
3
4
5
6
7
8
9
CYCLE

Fig.2