DL/SCI/R26

IMPROVING PROTEIN PHASES

Proceedings of the Study Weekend held at Daresbury Laboratory
5 - 6 February, 1988

Compiled by S. Bailey, E. Dodson and S. Phillips

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

ISSN 0144-5677

IMPROVING PROTEIN PHASES

Proceedings of the Study Weekend organised by
Collaborative Computational Project No. 4: Protein Crystallography
5 - 6 February, 1988

Compiled by
S. Bailey, Daresbury Laboratory*
E. Dodson, University of York
and
S. Phillips, University of Leeds

SERC DARESBURY LABORATORY
1989

---

*Now at the University of Keele

PREFACE

The suggestion for a Study Weekend concerned with the practical problems of initial phase determination in protein crystallography was first made at a CCP4 committee meeting in 1987. It was pointed out that there are many people in the field, especially research students and younger post-doctoral fellows, who have not had to cope with this stage of structure analysis and therefore lack practical experience of the problems commonly encountered. In addition to this, the number of new structures under investigation has increased rapidly over the last few years due to advances both in molecular biology, and crystallographic instrumentation and computing. Since the Study Weekends have placed great emphasis on education of newcomers to the field, it seemed appropriate to invite crystallographers with direct experience of primary phasing to pass on their accumulated knowledge in a formal way. It also furnished an opportunity to review recent developments in computational methods for improving the experimentally derived phase information.

The meeting was jointly organised by CCP4 and Daresbury Laboratory, and attracted a large number of participants. The brunt of the organisation and running of the meeting was borne by Shirley Lowndes and David Brown, to whom we are very grateful. With such expert assistance, our task was reduced to a minimum and the weekend seemed to run itself.

We hope that the papers in these Proceedings represent a collection of accumulated experience not often recorded in black and white, but more usually passed on by word of mouth, and will be a valuable reference for crystallographers in their everyday battle with the phase problem.

Sue Bailey
Eleanor Dodson
Simon Phillips

November 1988

CONTENTS

PROBLEMS WITH ISOMORPHOUS REPLACEMENT PHASES

by
S.E.V. PHILLIPS
Astbury department of Biophysics,
University of Leeds, Leeds LS2 9JT.

1. INTRODUCTION

The determination of three-dimensional structures of proteins by X-ray crystallography is currently enjoying a boom in popularity. This is a result of a combination of technical advances in data collection and computing, and the advent of recombinant DNA techniques which allow large quantities of hitherto scarce proteins to be prepared, and modified by site-directed mutagenesis. Approximately 160 different protein structures were to be found in the Brookhaven Protein Data Bank as of the beginning of 1988, with perhaps an equal number of known structures not deposited. Almost without exception, new structures (ie. those where there is no known structure for a closely homologous molecule) are solved using isomorphous replacement to calculate the phases.

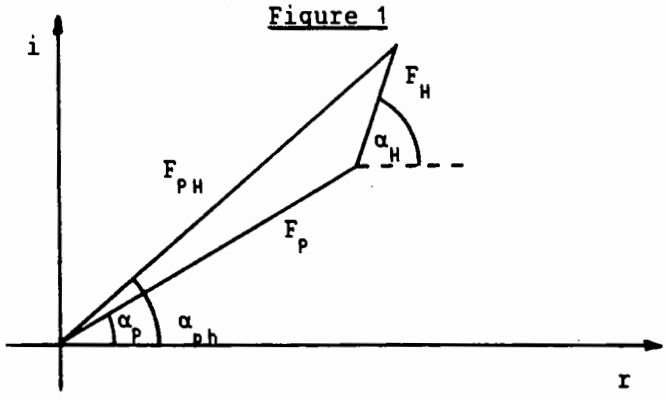
The production of good crystals of the protein, and preparation of useful heavy atom derivatives, have become the rate-determining steps in the process, since data collection, model building and refinement are now mostly routine. Good isomorphous derivatives are frequently difficult to find, and many studies founder at the stage of initial poorly phased electron density maps which are not interpretable in terms of an atomic model. There is a need for development of practical techniques to extract the best possible estimates of the phases in this situation, sufficient to allow the first model of the structure to be built. Once this has been achieved, the model can normally be improved by cycles of refinement and rebuilding.

Since the object of the exercise is to understand the biological functions of proteins, the experimental phase information is needed only to produce an initial map of interpretable quality. Fourier maps calculated with correct phases but incorrect coefficients (eg. all equal or even random) are generally interpretable, while those calculated with correct coefficients but random phases are not[1], so the solution of a protein structure rests on ability to produce a set of reasonably accurate phases to a resolution of about 3Å. A discussion of the problems commonly encountered in this process is given below, followed by a description of how some of them were overcome in the case of an antigen-antibody complex structure.

2. ISOMORPHOUS REPLACEMENT PHASES - PRACTICAL PROBLEMS

The method of isomorphous replacement was first used in determination of structures of alums in 1927[2], but found a place in protein crystallography in 1954[3] for phase determination in haemoglobin. Descriptions of the technique may be found in standard texts (eg. ref. 4) and a recent collection of reviews on diffraction methods for macromolecules[5], and the details will not be repeated here. The procedure depends on the phase triangles relating native and derivative structure factors for each hkl, an idealized version of which is shown in Fig. 1.

Figure 1



Idealized phase triangle relating the native protein structure factor F_P for a particular reflection to that of an isomorphous derivative F_{PH}, and the contribution of the heavy atom(s) F_H.

The only experimentally measurable quantities are the magnitudes $|F_P|$ and $|F_{PH}|$, with all phases, and the magnitude of the heavy atom contribution $|F_H|$ being unknown initially. In the event that the heavy atom contribution can be determined, for instance by determining the heavy atom structure and calculating its contribution, then the phase triangle can be closed (assuming no error). The phase α_P can then be determined for each hkl (with an ambiguity) from the cosine rule:

$$|F_{PH}|^2 = |F_P|^2 + |F_H|^2 + 2|F_P||F_H|\mathrm{Cos}(\alpha_P - \alpha_H)$$

Thus $\alpha_P = \alpha_H + \mathrm{Cos}^{-1}[(|F_{PH}|^2 - |F_P|^2 - |F_H|^2)/2|F_P||F_H|] = \alpha_H \pm \alpha'$

The ambiguity may be resolved if further phase triangles for the same hkl are available for other derivatives where the F_H vector is different from, and not collinear with, the first F_H. This will hold exactly in the error free case, if the contribution of the protein scattering to the derivative structure factor is equal to the native protein structure factor, ie. that the binding of the heavy atom to the protein causes no change in structure of the molecule, or its packing in the crystal lattice. This is the condition of isomorphism between native and derivative crystals, and rarely holds in practice, but it must be assumed in order to allow the construction of the vector F_{PH} from the vector sum of F_P and F_H.

Unfortunately, in the real case of experimentally measured $|F_P|$ and $|F_{PH}|$ there are experimental errors, and the isomorphism condition is unlikely to hold perfectly. This results in the phase triangles for different derivatives closing at different α_P values, or even not closing at all. The solutions for α_P are therefore not exact and must be treated using probabilistic methods, the usual approach being the formulation of Blow and Crick[6], which assumes all experimental measurement error resides in $|F_{PH}|$ and is Gaussian, as is the error in the calculated heavy atom contribution vector F_H. This allows calculation of a phase probability distribution for each hkl, and a figure of merit (m) which represents an estimate of the cosine of the phase error. Hendrickson and Lattman[7] showed how such phase probability distributions could be parametrized in a convenient way:

$$P(\alpha) = \mathrm{Exp}(K + A\cos\alpha + B\sin\alpha + C\cos 2\alpha + D\sin 2\alpha)$$

where K,A,B,C,D are constants

If phase probability data from different sources are expressed in this

way, they can readily be combined by summing the corresponding coefficients and normalizing to give a new joint probability distribution. These sources might be several isomorphous and/or anomalous derivatives, appropriately weighted data from partial atomic models or Fourier transforms of modified electron density maps etc.

The problem with isomorphous replacement phasing is a common one in experimental sciences - that of deriving useful information from small differences between large, and often inaccurate, values derived from experiment. To analyse where the major pitfalls lie, and where they might be avoided, it is instructive to examine each of the three vectors in the idealized phase triangle independently:

a) F_p

The magnitude $|F_p|$ is a basic experimental observation for the native protein crystal. It is therefore subject to experimental error, and phase calculations depend critically on its accuracy. Random error results from the statistics of the observations, which is worse for small or poorly diffracting crystals, but it can be correctly treated in the phasing calculations. Systematic error, on the other hand, is a much more serious problem, and can arise from a number of sources, eg. absorption, radiation damage, scaling of data from multiple crystals (especially if they are not all truly identical), and poor experimental technique and data reduction. The problem is that if not identified and corrected, it can distort phase calculations in unpredictable ways. Radiation damage can be monitored and corrections made using computer controlled diffractometers, either single counter or area detectors, if it corresponds simply to homogeneous destruction of the crystal, rather than radiation induced changes of structure. The latter case is difficult to handle, and most crystallographers prefer not to let data collection proceed past the point where the intensities drop below 85-90% of their initial values, to minimize errors in the correction. Use of new area detector systems allows more data to be taken per crystal, so should reduce the problem of scaling data from multiple crystals. The problem of absorption correction for area detector data is a major one in the development of software for these systems.

The moral is to collect the best data possible for the native crystal, especially considering these will be used in subsequent refinement of the structure once a initial model has been built. Careful attention to data quality, and application of corrections for systematic errors, will always pay dividends later. It should be noted that this includes a thorough investigation of the unit cell and space-group, and inspection of preliminary X-ray photographs to check for weak classes of spots, unusual symmetry etc. Failing to do this is bound to lead to trouble sooner or later.

The major error in construction of the phase triangles is probably lack of isomorphism. This can be discussed under the heading of F_p since it corresponds to a change of this vector in the phase triangle from its true value in the native crystal. The effect ranges from small structure changes around the sites of heavy atom substitution, which affect the quality of the multiple isomorphous replacement (MIR) map in that region, to large scale movements of the molecules in the lattice, either as shifts or rotations relative to the lattice symmetry. The problem was considered by Crick and Magdoff [8], who estimated its effects on the diffraction pattern. For instance, if, in a typical protein derivative crystal, all cell dimensions change by 0.5% from their native values, one might expect changes of 15% in intensities in the 3Å sphere. Cell changes can be monitored, of course, especially using diffractometers, and derivatives showing such changes must be regarded with suspicion. It is possible,

however, for all the molecules in the lattice to rotate by, say, 0.5°, which could cause similar intensity changes, and this would not be detected by monitoring cell parameters. Intensity changes due to lack of isomorphism, however, have the property of increasing with increasing resolution. Useful isomorphous changes normally lie in the range 10-25%, and should not rise too steeply with resolution. This problem may, therefore, be detected by examination of the behaviour of the fractional isomorphous differences with respect to resolution. The best cure for serious lack of isomorphism is to find a better derivative.

b) F_{PH}

The points noted above for F_P also apply to F_{PH}, but extra care needs to be taken if anomalous differences are to be utilized. In this case, radiation damage and absorption become even more important, since the differences to be measured are usually of the order of a few percent, and similar in magnitude to experimental errors. Radiation damage effects can be minimized by collecting anomalous pairs as close together in time as possible. Absorption, on the other hand, is more difficult to handle, and anomalous pairs should be collected as far as possible in a geometry where the absorption factors are similar for the + and - reflections. When collected carefully, such data can be very powerful, since there is no lack of isomorphism effect, and a surprising amount of information can be extracted, for instance in the resolved anomalous phasing technique[9].

In order to maximize the information available from a derivative, it is also necessary to tune the conditions of preparation of the derivative crystal. Time spent preparing the most isomorphous, yet well substituted, derivative crystal will be amply repaid later.

c) F_H

It may seem obvious, but the first question to consider here should be: "Is the heavy atom structure correct?". Heavy atom phasing and refinement programs produce many statistics, but these are often insensitive measures of the correctness of the heavy atom structure itself, and may look extremely convincing for sites that are incorrect, but related to the correct ones. The test of the proposed heavy atom arrangement is that it must explain the isomorphous (and anomalous) difference Patterson maps. Major peaks unnaccounted for in Pattersons are a danger sign, although it is usually not serious to find some expected vectors weak or displaced in the map.

A second, but related, point is that heavy atom parameters should be as accurate as possible, and great care should be taken with their refinement. This includes taking into account distortions of their apparent scattering factors caused by the displacement of solvent, and differing occupancies of equivalent sites in different crystals. The parameters also depend on accurate scaling of F_{PH} to F_P. The quality of the latter can be judged from inspection of the final estimates of the MIR phases to check for correlation with the phases of F_H.

3. PHASE IMPROVEMENT AND REFINEMENT

The result of a careful isomorphous phase calculation is a set of native phases, and figures of merit, which may be applied to $|F_P|$ to calculate an MIR electron density map. This may or may not be of interpretable quality, but any methods able to improve it could certainly reduce the time necessary to build the atomic model, or even make this possible rather than impossible.

Since the MIR data give a phase probability distribution, any other

information leading to an independent estimate of this distribution can readily be combined with it. This could consist of some prior knowledge of the joint probability distribution relating the phases and amplitudes of the reflections in the diffraction pattern, based on known or expected physical properties of protein crystals. The ultimate manifestation of the use of such distributions would be the complete ab-initio phasing of a protein structure, in a similar way to that now routine for small molecule crystal structures. (It must be noted, of course, that another source of information on the phase probability distribution could be obtained by adding data from another independent heavy atom derivative). Various methods are available for constructing these distributions, and they can be broadly considered in three categories:

a) Non-crystallographic symmetry: normally utilized by averaging the electron density in real space of equivalent molecules or subunits not related by crystallographic symmetry. It implicitly includes flattening of the electron density of solvent regions to a constant value as part of the process. The averaged map is then Fourier transformed, and the resulting phases used with the observed native structure amplitudes to calculate an improved electron density map, from which the cycle may be restarted. It may also be carried out in reciprocal space.

b) Other real space density modification techniques: the most frequently used being solvent flattening, ie. setting the density in the solvent filled interstices in the crystal lattice to a constant value in the electron density map. The density corresponding to the protein may also be modified using sharpening functions, searching for continuity etc.. The modified map is then Fourier transformed , and the resultant phases used to improve the current phase estimates.

c) Direct methods: using theoretically derived statistical relationships between phases and amplitudes. These methods have been outstandingly successful for small molecule crystallography, but have found less application for macromolecules. The maximum entropy approach would be categorized as a more recent development of these methods.

Discussions and applications of many of these methods appear in the following papers. The remainder of this paper consists of a description of a case where phase improvement by solvent flattening was essential to the solution of an important crystal structure.

4. PHASING THE Fab(D1.3)-LYSOZYME COMPLEX

During the course of a study of the structures of antigen-antibody complexes, initiated at the Institut Pasteur, crystals were prepared of an anti-lysozyme Fab-lysozyme complex. The antibody (D1.3) from which the Fab fragments were derived, originated from a hybridoma prepared from Balb/c mice immunized with hen-egg lysozyme (HEL). The crystals were grown from 15-20% PEG 8000 solutions, at pH 6.0 with 100mM phosphate buffer, using vapour diffusion. The space group was $P2_1$, with $\underline{a}$=55.7, $\underline{b}$=143.8, $\underline{c}$=49.1Å and β=120.5°, each asymmetric unit containing one Fab-HEL complex with a molecular weight of ca. 65,000. It was immediately apparent that there might be difficulty with phasing since the only centric zone in $P2_1$ is h0l, and these reflections constitute only a small proportion of the data, due to the short $\underline{a}$ and $\underline{c}$ axes.

The diffraction pattern extends to 2.8Å on precession films from conventional X-ray sources, and to about 2.2Å on films taken at the LURE synchrotron at Orsay. Preliminary precession photographs of various putative heavy atom derivative crystals revealed several promising cases, and 6Å data sets were collected for these, and a native crystal, on a 4-circle diffractometer. It was noted that the native crystal had $\underline{a}$=56.3Å,

while the derivatives had a ranging from 55.3 to 56.1Å, the other cell parameters being stable.

Isomorphous difference Patterson maps all showed similar features, regardless of the derivative, and even in cases where it was later discovered there was no specific binding. Since Patterson maps are determined chiefly by the largest 10% of the difference coefficients, the latter were printed out for inspection, after sorting into decreasing order of magnitude. The same reflections always gave the largest terms, and their magnitudes could be directly related to the length of the a axis, indicating that the majority of the Patterson features originated from lack of isomorphism rather than heavy atom vectors. The solution to the problem was to search for conditions that stabilized the unit cell to a=55.7± 0.2Å for native and derivative crystals, and to discard data from crystals falling outside this range. The Patterson maps were then interpretable, and several derivatives were solved. Heavy atom parameters refined using F_{HLE} (REFINE) and phased refinement (PHARE). Statistics are given in Table 1.

Table 1

Heavy atom phasing statistics at 6Å resolution for Fab-HEL Complex

Derivative	Conc (mM)	No of sites	R(native-derivative)	R(F_{HLE})
K_2PtCl_4	0.2	1	0.13	0.49
PHMBS	4.0	5	0.11	0.50
PHMBS	8.0	8	0.17	0.47
$K_3UO_2F_5$	0.1	7	0.12	0.44

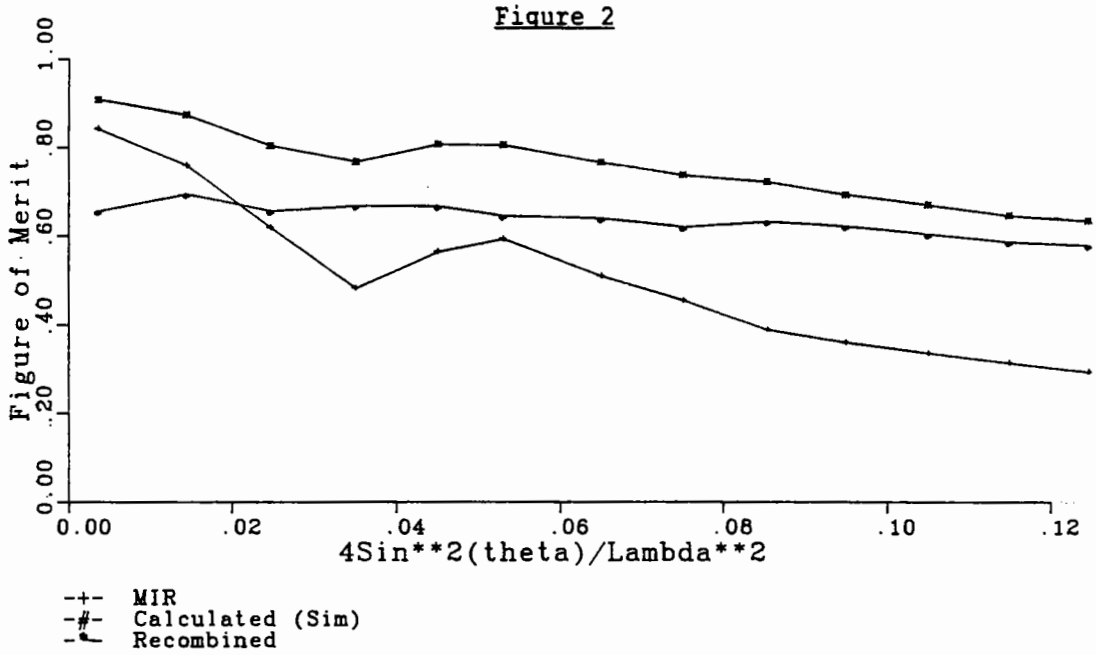
The overall figure of merit (<m>) for 1659 reflections was 0.75, using the CCP4 PHASE program (including the anomalous scattering contributions for all derivatives) with estimates of lacks of closure set to the observed values for the centric reflections. The value of <m> depends on these estimates, and tends to be a somewhat arbitrary measure of the quality of the phases in practice. The crucial test is the quality of the electron density map, which, in this case, was excellent. It was possible to fit models of lysozyme and another Fab of known structure (Fab-New) to it, using a PS300 graphics system, as 3 rigid bodies, corresponding to lysozyme, and the V and C regions of the Fab linked by a variable hinge[10].

Data collection was continued to 2.8Å resolution for the four derivatives in Table 1, but with the 4mM PHMBS replaced by 5mM PHMBS. Heavy atom parameters were refined first by F_{HLE}, and then by phased refinement, in the latter case always omitting the derivative to be refined from the phase calculations. Failure to observe this rule can lead to errors in the parameters[11] due to feedback. The phase information was poor, and resulted in underestimates for the occupancies of the heavy atom sites, so a final refinement pass included all derivatives in the phases and refined their parameters together. Unfortunately this tends to overestimate the occupancies, but proved to be a better compromise as judged by the quality of the resultant MIR phases.

A phase calculation to 2.8Å resolution gave <m>=0.47 for 15592 reflexions. Fig. 2 shows the behaviour of <m> with resolution for the observed phases

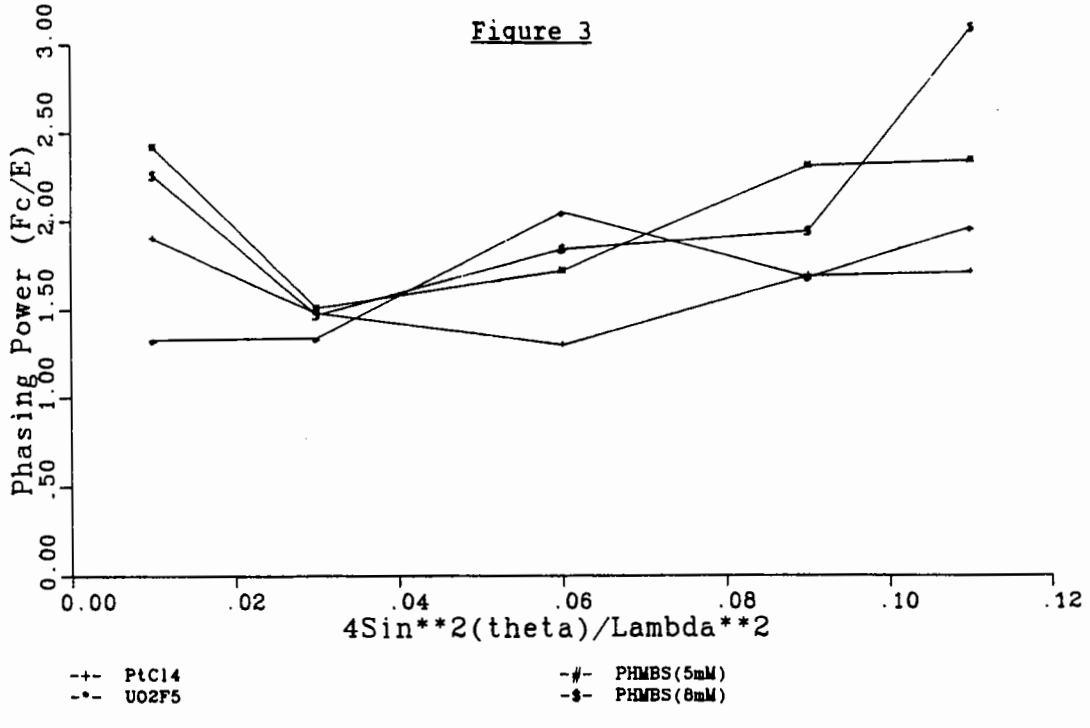
(labelled MIR), and Fig. 3 the phasing power of the derivatives. The latter is often taken as an indication of the usefulness of a derivative, but it depends on the occupancy assigned to the heavy atom sites, with overestimates producing an apparently superior phasing power. Fig. 3 suggests that the phasing is reasonable, even at high resolution, but this impression is misleading. The poor figure of merit can be increased by changing the parameters in the phase calculation, but the result in this case is always an uninterpretable electron density map.

Figure 2



Plot of figures of merit versus resolution for MIR phases, and the final solvent flattening phase refinement cycle calculated and combined phases.

Figure 3



Plot of phasing power (F_c/E) versus resolution for each derivative.

The MIR map was noisy, and appeared to be limited in resolution to 4-4.5Å. The problem was serious lack of isomorphism, and it was decided to try solvent flattening as a means of phase refinement, since the unit cell was estimated to contain 51% solvent. We generally followed the methods of Schevitz et al[12], although there were other papers in the literature at the time. Since we had a rough atomic model based on the 6Å map, a quick way to produce an envelope was to use these co-ordinates. A calculated map was made (using GENDEN) with all non-hydrogen atoms included with Biso=150Å^2. A contour drawn at 0.001e/Å^3 then enclosed a smooth envelope surrounding the protein, leaving 51% of the unit cell in the solvent region. A simple phase refinement cycle was then set up, using only CCP4 standard programs, plus one written to read the current and the envelope maps, and modify points in the former according to the following scheme:

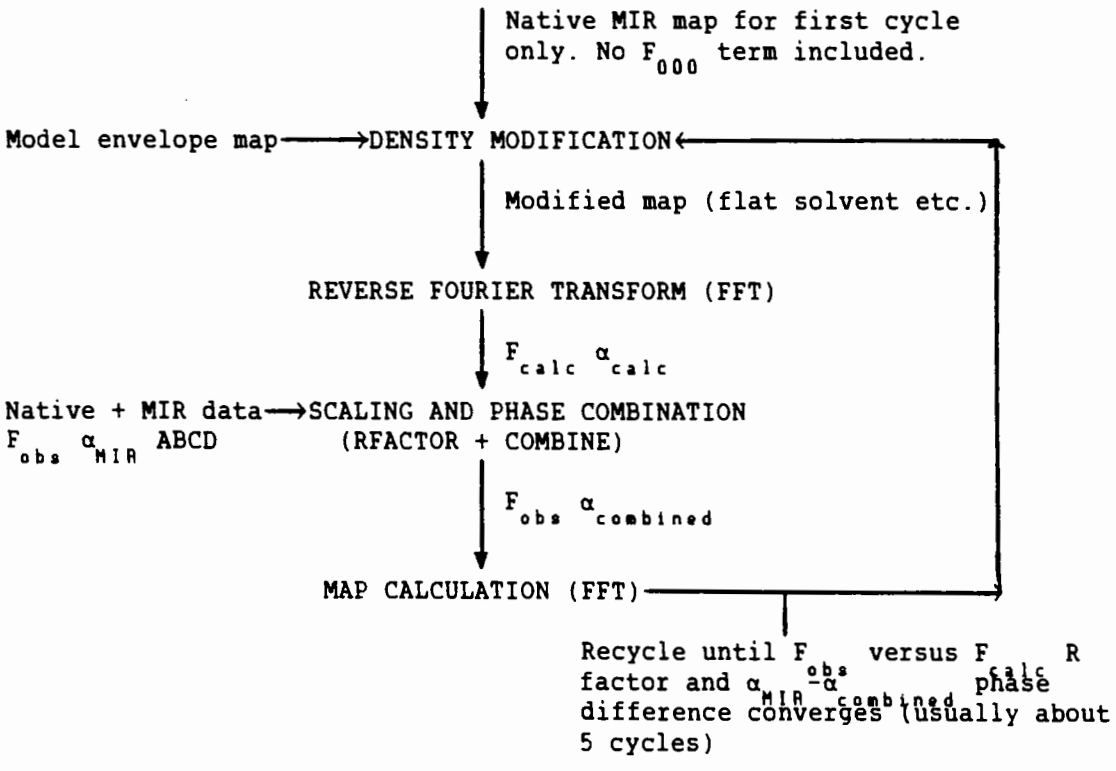
Inside protein envelope: $\rho = (\rho_{old} - \rho_{max}) \times 0.1 + \rho_{max}$ for $\rho_{old} > \rho_{max}$
(envelope $\rho > 0.001e/Å^3$) $\rho = (\rho_{old} - \rho_{min}) \times 0.1 + \rho_{min}$ for $\rho_{old} < \rho_{min}$
$\rho = \rho_{old}$ for $\rho_{max} > \rho_{old} > \rho_{min}$

Outside protein envelope: $\rho = \rho_{solv}$

where ρ_{old} = density in current map, ρ = density in new modified map, ρ_{max} and ρ_{min} are estimates of highest and lowest electron density values expected in the protein, ρ_{solv} is expected solvent density. Although the latter should strictly be non-zero, the best results were obtained with $\rho_{solv} = \rho_{min} = 0$. The basic cycle is shown in Figure 4.

Figure 4

Density Modification Cycle



The MIR phases were always retained since it was considered they would be useful if correctly weighted, although other workers have advocated removing them in the later stages. The envelope was fixed

throughout, and no attempt was made to update it in later cycles. The weighting of the α_{calc} contribution was the standard Sim weighting scheme provided in COMBINE.

An initial trial was made, starting from the 3Å MIR map, and cycling around the phase refinement to convergence at that resolution(6 cycles). The resulting map showed featureless solvent regions, but the protein density remained uninterpretable! A possible explanation of this observation is that the high resolution MIR phase information was so poor that the noise in the map overwhelmed the true density, so that the refinement was caught in a false minimum.

The second, ultimately successful, approach was to start the phase refinement with the 4Å resolution phases (<m>=0.61), which gave a good MIR map before modification. The course of this refinement was followed by observing the density in the region of the 76-94 disulphide bridge in the lysozyme molecule. Convergence was achieved after 6 cycles, with an average phase change from α_{MIR} of 26°. The map calculated from these phases was substantially improved, with new side chains appearing, and the disulphide bridge, which was not connected in the MIR map, lying in continuous density. This was then used as the basis for a cautious phase extension, increasing the resolution in narrow shells, and refining phases to convergence at each step before extending again (Table 2). The extension step was carried out by calculating the density modified map at one resolution, and then carrying out the reverse Fourier transform to calculate structure factors at slightly higher resolution, and including MIR phase information corresponding to the latter.

Table 2

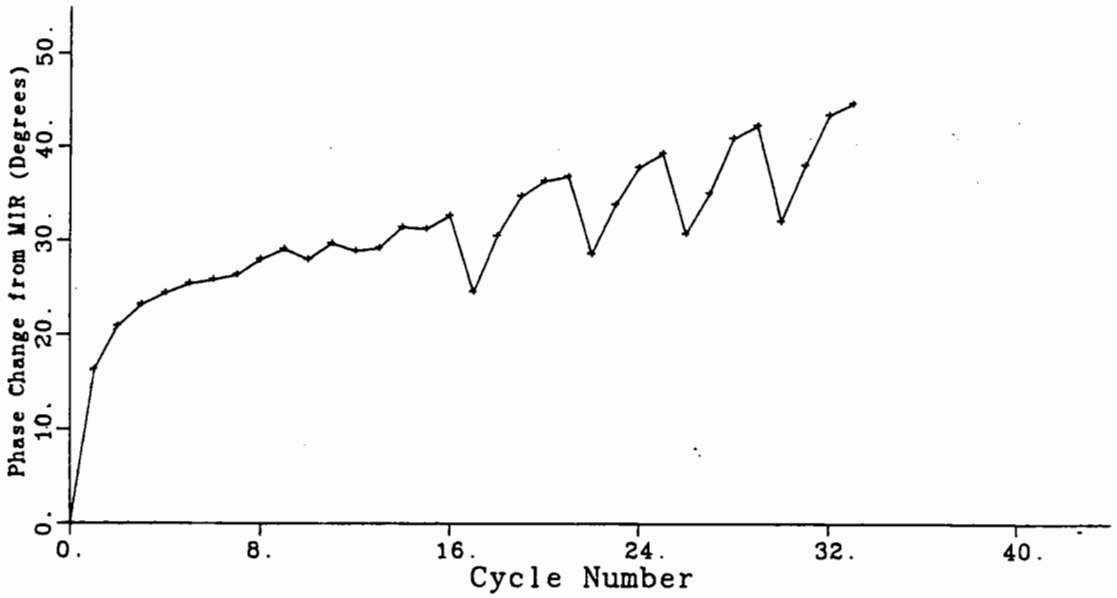
Course of Phase Extension and Refinement

Cycles	Resolution(Å)	Comments
1- 6	25-4.0	
7-11	25-3.7	
12-16	25-3.5	
17-21	25-3.3	α_{calc} given weight 0.4, 0.7 in first two cycles
22-25	25-3.1	" "
26-29	25-2.95	" "
30-33	25-2.80	" "

Fig. 5 shows the phase change of $\alpha_{combined}$ from α_{MIR} during the phase extension and refinement. The final figure of merit for the combined phases at cycle 33 was 0.73, with the statistics shown in Figure 2, but little importance should be attached to this value. The crucial

observation was that the 2.8Å combined map was now interpretable.

Figure 5



Course of the phase change from MIR phases during phase refinement. The minima at cycles 17,22,26 and 30 correspond to phase extension steps where α_{calc} was downweighted.

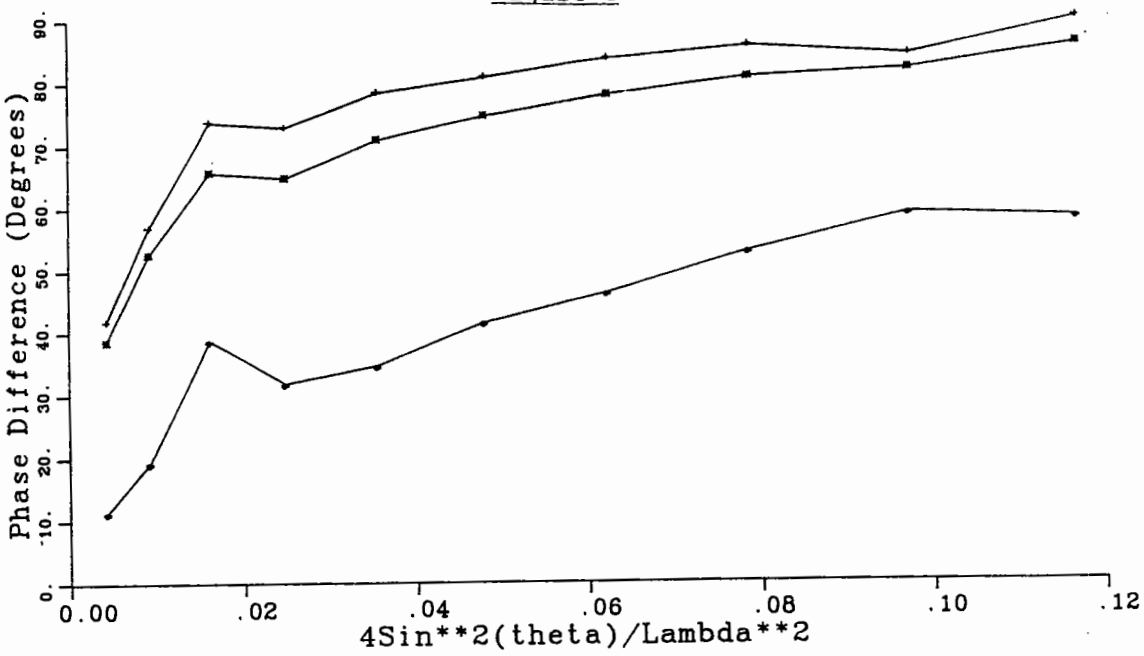
With the benefit of a knowledge of the approximate chain trace from the known lysozyme and Fab structures, all but 24 of the 562 residues in the complex were built into the cycle 33 map. This model was adequate to begin cycles of restrained least-squares refinement and model building to complete the atomic model[13]. The current R factor for all data from 10-2.8Å is 0.26.

Figs. 6 and 7 show the relationships of the MIR and solvent flattened phases to those calculated from the current atomic model, as a function of resolution and MIR figure of merit respectively. If the model phases are taken to be the best approximation to the true phases, it is clear that the phase refinement results in a significant improvement relative to the MIR phases.

At high resolution the MIR phases appear almost random relative to the model phases, with a mean difference of about 90°, but the refined phases are significantly better. The biggest improvements are in the middle resolution range, while the effect at low resolution is less marked, due to the good quality of the MIR phases to 6Å. The statistics with respect to figure of merit show little improvement for high m values, as expected, but phases with very low m, which are essentially random from MIR, have improved most significantly with phase refinement.

The striking observation is that such an apparently small improvement in the phases in numerical terms produces a huge difference in the subjective quality of the electron density map. The major result of this study is that solvent flattening is a straightforward technique, but which can make a major contribution to improving poor phases, and is probably a worthwhile procedure to apply systematically to all experimentally phased electron density maps.

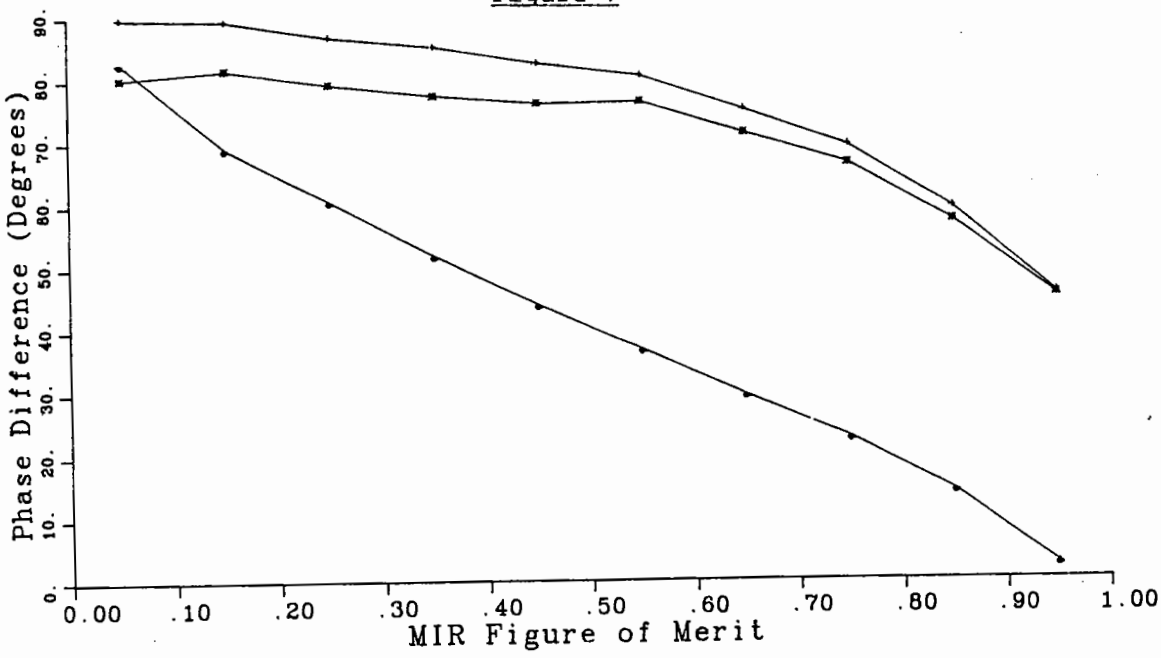
Figure 6



-+- MIR vs Calculated
-*- Solvent Flattened vs Calculated
-#- MIR vs Solvent Flattened

Plot of the average differences between the final model phases, and the MIR and refined phases versus resolution.

Figure 7



-+- MIR vs Calculated
-*- Flattened vs Calculated
-#- MIR vs Flattened

Plot of the average differences between the final model phases, and the MIR and refined phases versus MIR figure of merit.

---

5. ACKNOWLEDGEMENTS

I thank my colleagues on the D1.3 project; Adolfo Amit, Roy Mariuzza and Roberto Poljak. The work was supported in part by grants from EMBO, CNRS and the Institut Pasteur.

6. REFERENCES

1. G.N. Ramachandran and R. Srinivasan in "Fourier Methods in Crystallogaphy" (New York: Interscience, Wiley, 1970)

2. J.M. Cork, Phil. Mag. 4, (1927) 688.

3. D.W. Green, V.M. Ingram and M.F. Perutz, Proc. Roy. Soc. London, A225, (1954) 287.

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

5. Methods in Enzymology, Volume 115, (Orlando: Academic Press, 1985)

6. D.M. Blow and F.H.C. Crick, Acta Cryst., 12, (1959) 794.

7. W.A. Hendrickson and E.E. Lattman, Acta Cryst., B26, (1970) 136.

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

9. W.A. Hendrickson and M.M. Teeter, Nature(London) 290, (1981) 107.

10. A.G. Amit, R.A. Mariuzza, S.E.V. Phillips and R.J. Poljak, Nature(London) 313 (1985) 156.

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

12. R.W. Schevitz, A.D. Podjarny, M. Zwick, J.J. Hughes and P.B. Sigler, Acta Cryst., A37, (1981), 669.

13. A.G. Amit, R.A. Mariuzza, S.E.V. Phillips and R.J. Poljak, Science, 233, (1986), 747.

SOLVENT FLATTENING VS ADDITIONAL DERIVATIVE DATA IN THE IMPROVEMENT OF M.I.R. PHASES

by

A.G.W. LESLIE
Blackett Laboratory, Imperial College, London SW7 2BZ, UK

1. INTRODUCTION

In determining a new protein structure a stage is frequently reached where a partially interpretable m.i.r. phased electron density map has been obtained but it is not yet possible to follow the complete polypeptide backbone. In seeking to improve the interpretability of the electron density map the investigator is faced with two distinct alternatives: he can either search for additional derivatives, or he may resort to computational methods to improve the existing m.i.r. phases. Possible computational approaches include density modification (these might involve molecular averaging, simple solvent flattening or the potentially more powerful maximum entropy techniques) or phase combination with phases from a partial model (as described by David Rice in this volume).

In order to assess the relative merits of using additional derivative data vs a simple density modification procedure, a series of trial calculations have been made using data from the structure determination of chloramphenicol acetyltransferase (CAT).

The structure of CAT, a trimer of MW 3x25000 has recently been determined from an m.i.r. map calculated at 2.7Å resolution based on data from six isomorphous derivatives. The structure was subsequently refined using data to 1.75Å resolution, and the final model gave an R−factor of 19% for all data between 6 and 1.75Å resolution. The refined CAT structure and the isomorphous derivative data have been used as a trial system to evaluate the improvement in the initial m.i.r. phases as additional derivative datasets and a new native dataset were included in the phasing. This has been compared with the improvement in phases following the application of a solvent flattening procedure.

In order to obtain a quantitative estimate of the errors in the experimental phases a "perfect" phase set was calculated from the coordinates of the refined structure. As the R−factor is only 15.3% (6−2.7Å resolution data) the errors in these calculated phases are negligible compared to those of the experimental phases.

2. THE DERIVATIVES USED IN THE M.I.R. PHASING

Some details of the derivatives used in the structure solution of CAT are given in Table 1, in the order in which the derivative data were collected. The 5mM gold cyanide derivative gave the best phasing statistics, with a low Cullis R−factor and a high overall phasing power. By contrast the platinum and samarium derivatives are very poor, and the platinum data were only included to 4Å resolution. Both derivatives suffer badly from lack of isomorphism and in addition the samarium sites have relatively low occupancy. The iodinated substrate (PICM) derivative and the PHMB derivative are quite similar in phasing statistics, the lower scattering power of the iodines being at least partially compensated by the fact that this derivative was the most isomorphous.

Table 1. Derivatives used in the structure solution of CAT

Compound	R_{deriv}	R_c	Phasing Power	Sites
PICM	10.1	61.2	1.32	A, B
$PtCl_4$	13.8	82.6	0.88	C, C'
$SmNO_3$	9.1	91.1	0.55	D, E
2mM AuCN	13.2	58.1	2.34	F, F', G
5mM AuCN	17.0	51.3	2.92	F, F', G
PHMB	12.1	65.6	1.38	F, F'

Abbreviations: PICM para-iodo chloramphenicol, PHMB parahydroxy mercuribenzoate.

$$R_{deriv} = \frac{\Sigma\ |FPH - FNAT|}{\Sigma\ FNAT}$$

$$R_c = \frac{\Sigma\ |FPH_{calc} - FPH_{obs}|}{\Sigma\ FPH_{obs}}$$ for centric reflections (Cullis R factor)

$$\text{Phasing power} = \frac{FH_{r.m.s.}}{E_{r.m.s.}}$$

where FNAT, FPH, FH are the native, derivative and heavy atom structure factor amplitudes respectively, and E is the lack of closure.

Derivative sites less than 3Å apart are indicated by using a primed letter, eg C, C'.

It should be noted that the iodine, platinum, samarium and gold derivatives have completely independent sites, while the PHMB site is the same as the major site of the gold derivative.

Following collection of the 5mM gold cyanide derivative dataset, the native data were recollected, as the quality of the original native data had been affected by instability of the X-ray generator. (This was apparent from the merging R-factor of 7.2% on intensities compared with values of between 4% and 5% for the derivatives.) The new native dataset gave a merging R-factor of 4.4%, and gave the R_{deriv} values listed in Table 1 which are up to 2.8% lower than those derived from the original native data.

Table 2. Phase sets calculated at different stages of the CAT structure determination.

Phase Set	Derivatives	r.m.s. phase error	figure of merit
1	PICM(A)+ $PtCl_4$+ $SmNO_3$	81.3	0.60
2	PICM(A)+ $PtCl_4$+ 2mM AuCN(A)	69.1	0.73
3	PICM(A)+ $PtCl_4$+ 2mM AuCN(A) + 5mM AuCN(A)	65.7	0.75
4	New native + PICM + $PtCl_4$ + $SmNO_3$ + 2mM AuCN + 5mM AuCN	61.7	0.64
5	As (4) but including anomalous for PICM and AuCN	57.3	0.72

The inclusion of anomalous scattering data is indicated by (A) following the derivative.

3. THE M.I.R. PHASES

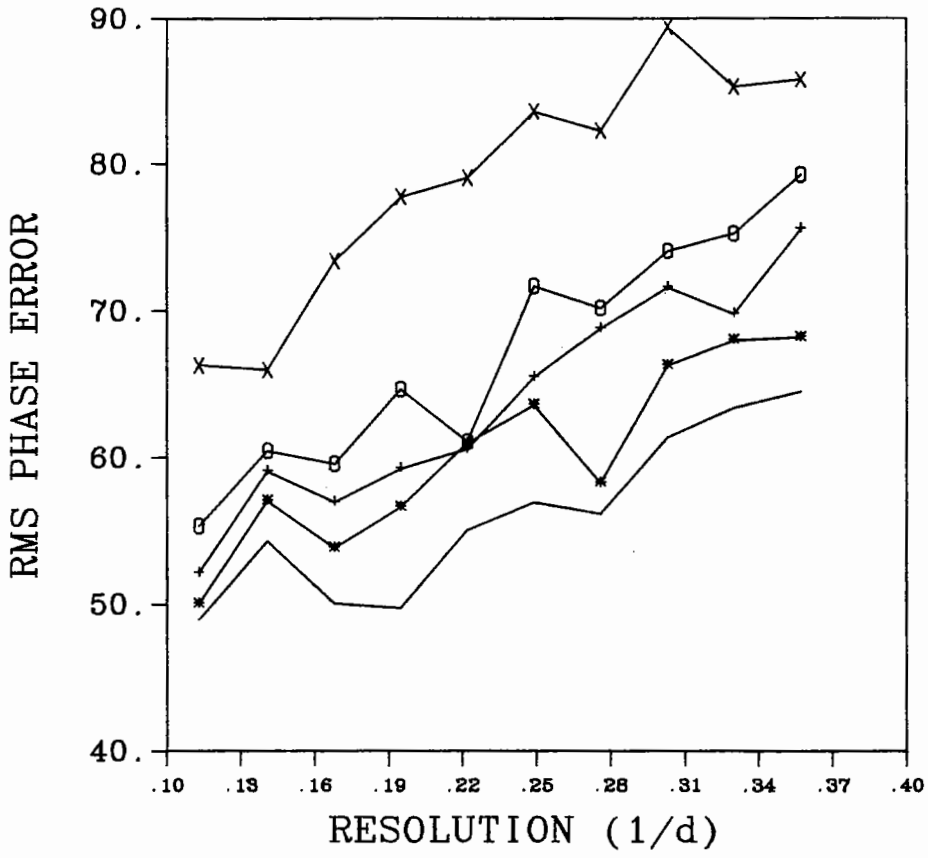
In order to investigate the effect on the initial m.i.r. phases of including additional derivative data, phases were calculated using five different combinations of native and derivative data (See table 2). These are the same phases that were calculated at different stages in the structure determination. The resulting m.i.r. phases were compared with the "perfect" model phases as a function of resolution and figure of merit (Fig 1). A number of conclusions can be drawn from these results.

(i) M.i.r. phases are very inaccurate. Even for the final phase set, from which the structure was solved, there is no resolution bin where the r.m.s. phase error falls below 50°.

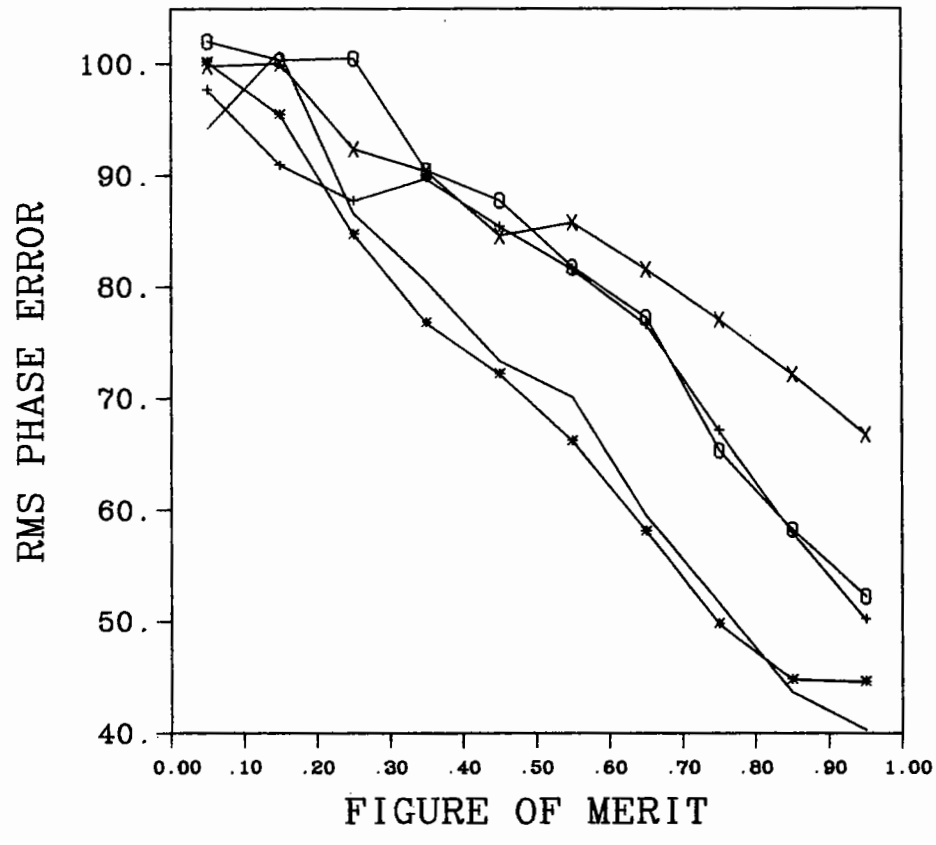
(ii) As expected, the m.i.r. phases improve steadily as additional derivative data are included, and the introduction of the new native dataset was particularly effective in improving the phasing.

(iii) Comparing the fourth and fifth sets of phases the power of anomalous scattering is very apparent (this was recognised at the time by the significantly lower noise level in the solvent region of the electron density map phased using anomalous data).

Points (ii) and (iii) argue very strongly for taking the greatest possible care in the collection and measurement of the X−ray data, particularly since the anomalous signal is typically of the same order of magnitude as the error in the corresponding structure factor amplitude. Recent developments in film processing software and the advent of commercially available area detectors should help in taking full advantage of the power of anomalous scattering in phase determination. From Figure 1b it is apparent that, as expected, there is a good correlation between the figure of merit of a reflection and its phase error. However it is clear that this relationship is not a simple one, as the phase error for reflections with the same figure of merit varies by over twenty degrees for different phase sets. The relationship between the figure of merit and the mean phase error, $\Delta\Phi$ = acos (m) is obeyed quite closely for the final m.i.r phases, but would give a serious underestimate of the phase error for the earlier phase sets.



(a)



(b)

Fig. 1. The r.m.s. phase error as a function of resolution (a) and figure of merit (b) for the phases described in Table 2. The symbols used to denote the different sets of phases are: set 1 (X); set 2 (O); set 3 (+), set 4 (*), set 5 (no symbol).

Table 3. The effect of individual derivatives on the m.i.r. phases

	r.m.s. phase error
Including all derivatives	57.3
OMIT:	
PICM (A)	64.6
$PtCl_4$	57.6
Sm NO_3	58.0
2mM AuCN (A)	57.0
5mM AuCN (A)	60.4
both AuCN (A)	66.3
PHMB	59.0
INCLUDING ONLY:	
PICM (A)	73.5
5mM AuCN (A)	70.5
PICM (A) +5mM AuCN (A)	59.7
Following s.i.r. parameter refinement	
PICM (A) only	74.2
5mM AuCN (A) only	71.2
PICM (A) + 5mM AuCN (A)	60.0

In order to assess the contribution of individual derivatives to the final phases, each derivative in turn (and both gold derivatives) were removed from the phasing and the resulting phases compared with the model phases. The results are given in Table 3, and again a number of interesting conclusions can be drawn:—

(i) The highly non—isomorphous platinum and samarium derivatives contribute very little to the phasing, as would be expected. However it is important to note that the inclusion of these derivatives does not give rise to an increase in the r.m.s. phase error, so that even poor derivatives can safely be included in the phasing providing that their contribution is weighted appropriately.

(ii) As expected from the phasing statistics given in Table 1, removal of both gold derivatives causes the greatest deterioration in the m.i.r. phases. By contrast, removal of the 2mM gold derivative actually reduces the phase error. A possible explanation is that both gold derivatives contain essentially the same phase information (although the "signal" is certainly greater in the 5mM derivative), and by including both derivatives too much weight is being placed on this derivative relative to the other derivatives.

(iii) Although the PICM and PHMB derivatives have similar phasing statistics (Table 1), removing the PICM derivative has a much greater effect than removal of the PHMB derivative. This demonstrates the advantage of having derivatives with completely independent sites, for although the PHMB derivative gives good phasing statistics it suffers from sharing a common site with the gold derivative, thereby greatly reducing its effective contribution to the overall phasing. (It should be pointed out that it is not the inclusion of anomalous scattering that makes the PICM derivative so much more effective, since repeating these calculations excluding the anomalous contribution gave very similar results.)

Table 3 also presents the results of calculating single derivative phases for the PICM and gold cyanide derivatives, and for the combination of these two. The s.i.r. + anomalous phases have very high r.m.s. errors, even for the gold derivative which, on the basis of its statistics (Table 1), is an unusually powerful derivative. However the combination of these two yields phases with an r.m.s. error only 2.4° worse than the final m.i.r. phases, emphasizing once again that these two derivatives are making the greatest contribution to the overall phasing.

4. THE EFFECT OF HEAVY ATOM PARAMETERS

The quality of the isomorphous phases will obviously depend on the success of the heavy atom parameter refinement, and this can present difficulties, particularly if only one or two derivatives are available. In order to obtain a quantitative estimate of the dependence of the phase error on the values of the heavy atom parameters, these parameters were refined completely independently for the PICM derivative alone and for the 5mM gold derivative alone. A phase refinement program (PHARE) was used, and the anomalous data were included in the phase calculation. The s.i.r. refinement led to changes in occupancy of up to 28% for the PICM derivative and 15% for the gold derivative, but only small shifts in position. Even so, the resulting s.i.r. + anomalous phases were only slightly worse than those calculated with the "optimum" values derived from parameter refinement using all six derivatives (Table 3), and the two derivatives combined gave an r.m.s. phase error only 0.3° worse, suggesting that the phases are not critically dependent upon the heavy atom parameters. (It should be noted however that only the correct sites were included during the s.i.r. heavy atom parameter refinement, and it is well known that one of the major difficulties in s.i.r. refinement is the effective elimination of incorrect sites). As a final test the parameters of all six derivatives were refined against the "perfect" model phases. This resulted in even more dramatic changes in site occupancies and thermal factors, but the resulting phases showed only 1.8° reduction in the r.m.s. phase error, supporting the idea that the phases are relatively insensitive to errors in the heavy atom parameters, although again it should be noted that no incorrect sites were included in the heavy atom model.

5. THE RELATIONSHIP BETWEEN R.M.S. PHASE ERROR AND MAP INTERPRETABILITY

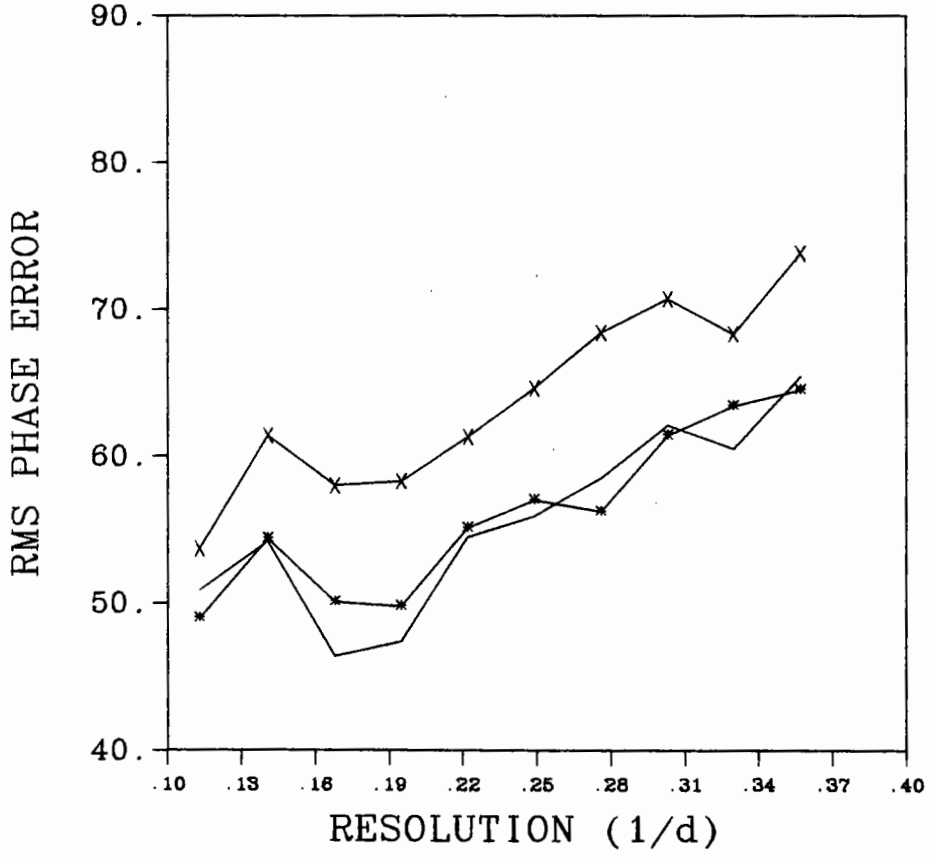
While it is difficult to generalise about the point at which an m.i.r. phased electron density map becomes interpretable, a review of published structures reveals that in those cases where the final m.i.r. phases have been compared with phases derived from a refined model an error of around sixty degrees is typical (but note that the mean phase difference rather than the r.m.s. value is often quoted). In the case of the CAT structure a number of different maps based on different sets of phases were examined. The initial map, based on phases with an r.m.s. error of 81°, m= 0.6 was completely uninterpretable. The map based on the second and third phases sets (r.m.s. error 69°, m= 0.73 and 66°, m= 0.75 respectively) were partially interpretable. In particular the strands of the β−sheet were quite clear, but the connectivity between strands, and the molecular boundary between the subunits of the trimer were rather ambiguous. Perhaps surprisingly, the alpha−helices were also very difficult to recognise in these maps. By contrast the map calculated using the final m.i.r. phases (r.m.s. error 57°, m= 0.72) was of a very high quality and the polypeptide chain could be followed without difficulty. The one point of ambiguity in strand connection was easily resolved from the known sequence. One conclusion that can be drawn from these results is that, at least in the case of the CAT structure at 2.7Å, the improvement in m.i.r. phases from an r.m.s. error of 65° to 57° had a dramatic effect on the interpretability of the electron density map, and a phase error of around sixty degrees seems to represent a "break−point" between a map in which the secondary structure can be recognised (β−sheet more easily than α helices) but the connectivity is ambiguous and a map which is readily interpretable. This "break−point" is pertinent to the discussion of the improvement afforded by solvent flattening methods.

6. PHASE IMPROVEMENT BY SOLVENT FLATTENING

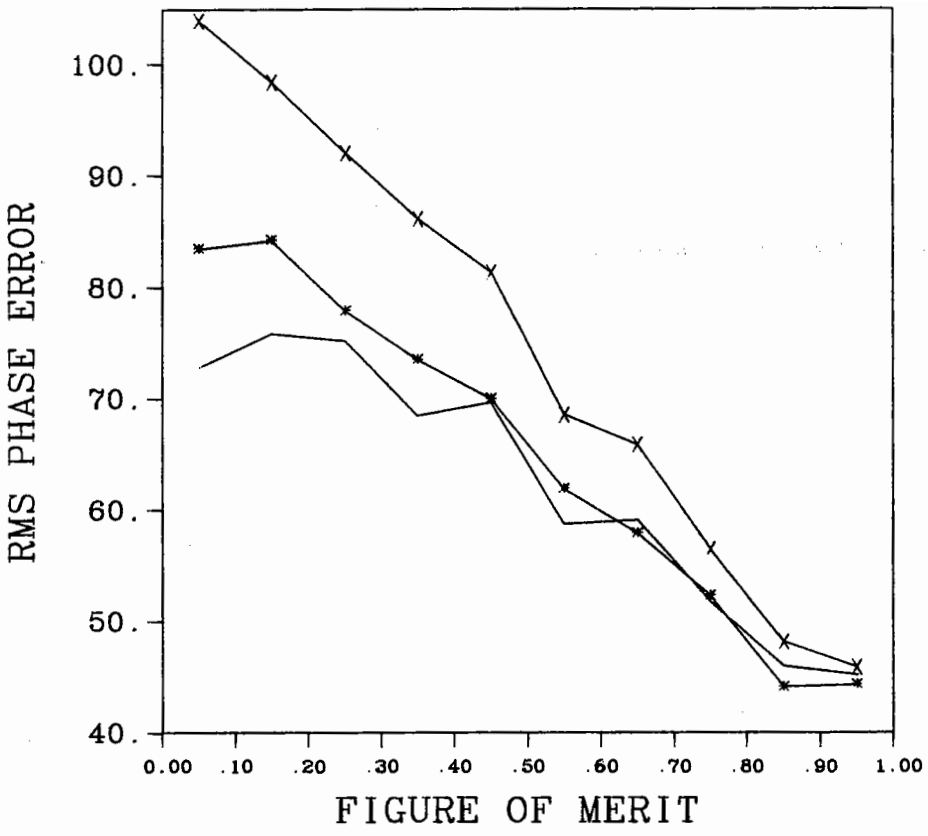
The concept of using solvent flattening to improve isomorphous replacement phases is not particularly new. Rossmann and co−workers applied the technique to lobster GAPDH [1] and the necessary programs were described by Bricogne [2] and were used in the structure determination of TMV coat protein discs and *B. stearothermophilus* GAPDH. However in these applications solvent flattening was only used in conjunction with molecular averaging.

One of the earlier successes in using solvent flattening alone was in the structure solution of fMet−tRNA by Sigler and colleagues [3], but the method has only become widely applied since the work of B.C. Wang [4]. In particular, Wang's algorithm for automatically determining the molecular envelope has greatly simplified the routine application of the method, and if the envelope is determined in reciprocal space rather than real space [5] the whole procedure is very inexpensive in computational resources.

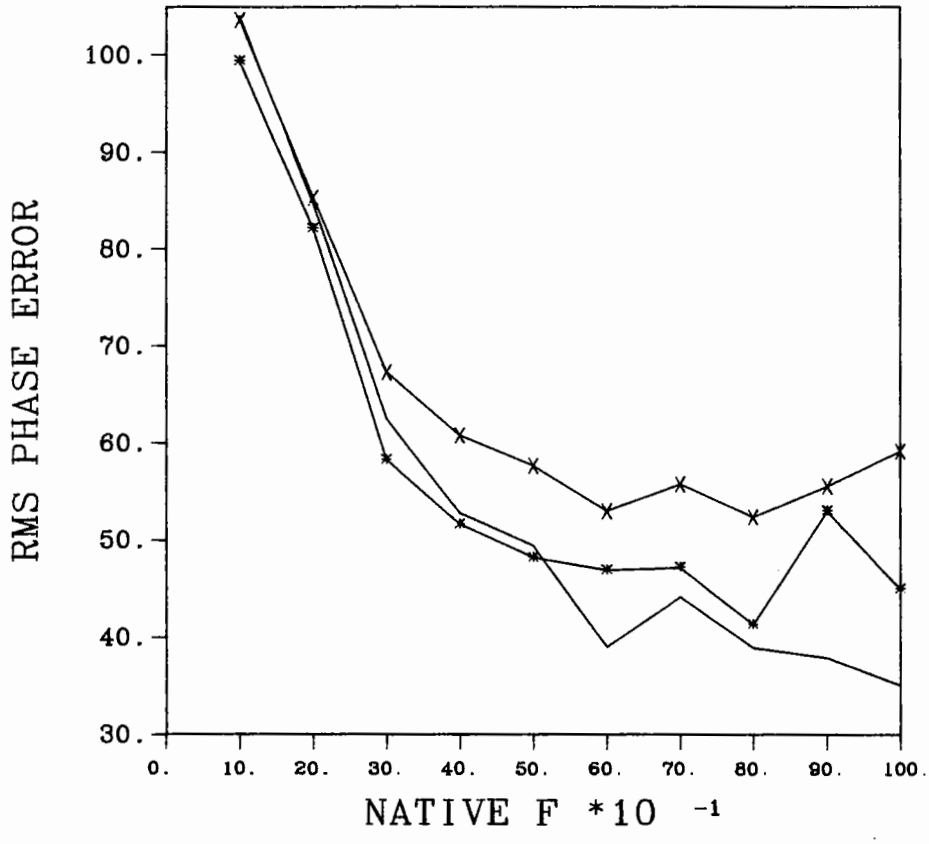
The solvent flattening procedure was initially applied to the m.i.r. phases that gave a partially interpretable electron density map. These phases were calculated using the derivatives described as phase set 3 in Table 2, but slightly modified lack of closures (derived from the centric reflections) which resulted in a figure of merit of 0.67 (rather than 0.75) and an r.m.s. phase error of 65.3° rather than 65.7°. The envelope was determined using the reciprocal space Wang algorithm with an averaging radius of 10Å and assuming a solvent content of 50% (for CAT V_m= 2.54Å^3/dalton suggesting a solvent content of 55%). The same envelope was used for each cycle of the procedure. The ratio of the mean electron density inside and outside the molecular envelope was adjusted artificially by adding a constant density to all grid points so that only 3% of the grid points inside the protein region had negative densities. These negative densities were then truncated to zero. Three cycles of solvent flattening were performed, and on each cycle the phases obtained by Fourier inversion of the solvent flattened and negative−protein truncated map were combined with the m.i.r. phases. The final combined phases gave an overall figure of merit of 0.85, and the r.m.s. phase error had been reduced to 56.8°, 0.5° lower than the final m.i.r. phases (Table 2). The improvement in the phases as a function of resolution, initial figure of merit, and native structure factor amplitude is shown in Figure 2. Figure 2a demonstrates that the phases are improved over the entire resolution range in a very similar way to the improvement obtained by using the new native data and the additional (PHMB) derivative (The r.m.s. error in the solvent flattened phases is slightly lower in the resolution range 5−6Å but this may not be significant). The same general conclusion can be drawn from the results presented in Figure 2b, although in this case the improvement for reflections with a very low figure of merit (< 0.2) is substantially greater for solvent flattening. The solvent flattened phases are also significantly better for the larger structure factors (Figure 2c), and as this includes the strongest 12% of the data this should have a significant effect on the quality of the final map.



(a)



(b)



(c)

Fig. 2. A comparison of the r.m.s. phase error for the initial (starting) m.i.r. phases (X), these phases following three cycles of solvent flattening (no symbols) and the best m.i.r. phases (Set 5 in Table 2) (*), as a function of resolution (a), figure of merit of the initial m.i.r. phases (b) and the native structure factor amplitude (arbitrary scale) (c).

6.1 The effect of the averaging radius

The value assigned to the averaging radius used in the determination of the molecular envelope [4] primarily affects the detailed appearance of the surface of the envelope rather than changing its gross morphology. A smaller averaging radius (say 6Å) gives a more convoluted surface which, in principle, should be able to follow the true structure of the protein surface more readily, but has the disadvantage that it tends to introduce cavities inside the protein, and may also increase the risk of cutting off external protein loops if their density is rather weak in the initial m.i.r. map. A larger averaging radius (say 10Å) may therefore be more appropriate, especially if the quality of the initial map is rather poor. In order to assess the effect of using different averaging radii in determining the molecular envelope the phase errors following three cycles of refinement were calculated using averaging radii of 10, 8 and 6Å, and finally using a "perfect" envelope derived from the model co-ordinates (Table 4). The results indicate that the choice of averaging radius is not critical, and even the perfect envelope gives only marginally greater improvement in the phases.

Table 4. The effect of different envelopes on the phase error after solvent flattening. Initial phase error 65.3°, solvent content 50%.

Averaging radius used to determine envelope (Å)	r.m.s. phase error after 3 cycles
10	56.8
8	56.9
6	56.9
Perfect envelope generated from model	56.1

In each case the envelope was determined on the first cycle and not updated on succeeding cycles.

6.2 The effect of solvent content

The power of solvent flattening as a phasing technique depends on the ratio of the molecular volume to the unit cell volume, and hence on the solvent content used in the determination of the molecular envelope. The method has been most successful when the solvent content is high (say 70%). However the results on the CAT system show that a significant improvement in phases can be obtained for a solvent content of only 50%. As an extreme case, the solvent content was set to 25% in the determination of the envelope. Three cycles of refinement reduced the r.m.s. phase error from 65.3° to 60.5° (compared to 56.8° for 50% solvent). This suggests that even when the solvent content is very low, solvent flattening can still give a worthwhile improvement in m.i.r. phases.

6.3 The effect of truncating large negative protein densities

In the version of the software used in these tests the effective F(000) term to be added to all grid points in the map prior to truncating negative densities within the molecular envelope was determined by specifying the desired ratio of the mean electron density inside and outside the molecular envelope after adding in the constant term. In theory this ratio should be equal to the calculated ratio of the mean protein and solvent electron densities, the former having the value 0.433 e/A^3 and the latter depending on the crystallisation medium (0.33e/Å^3 for water). However this approach is valid only if all the low resolution terms are included in the m.i.r. map, a condition rarely met in protein crystallography. In practice the ratio is adjusted to give the desired level of truncation of negative protein electron densities. The dependence of the resulting phases on the extent to which the protein density was truncated was investigated by choosing values for the

protein to solvent density ratio which resulted in between 3% and 49% of the protein grid points having their density truncated to zero. The results are presented in Table 5, and demonstrate that a significant improvement in the phases results from truncating about 44% of the protein density rather than 3%.

The true F(000) contribution (0.375e/Å^3) should have been 10 on the scale adopted in Table 5, showing that the best results were obtained by truncating rather more of the protein density than is justified theoretically, although the omission of low resolution terms (beyond 10Å) from the electron density map may affect the optimum level for truncation.

Table 5. The effect of truncating negative protein densities. Initial phase error 65.3°, averaging radius 10Å, solvent content 50%.

F(000) term added to all densities (arbitrary scale)	Percentage of grid points within protein envelope where density was truncated (mean over 3 cycles)	r.m.s. phase error after 3 cycles
25	3%	56.8
11	24%	55.0
4	44%	54.1
2	49%	54.3

In view of the undesirable nature of the sharp cut off introduced by truncating the negative densities to zero, which will inevitably give rise to noise features in the high-resolution terms when the modified electron density map is Fourier transformed, the alternative approach of attenuating the negative densities proposed by Schevitz et al.[3] was also tested. Using an attenuation factor of 0.1 for the negative densities gave exactly the same r.m.s. phase error following 3 cycles of solvent flattening as simple truncation (for 44% grid points modified) and even at the highest resolution (2.7Å) there was no indication that the phases were superior. It should be mentioned, however, that a fourth cycle did result in an improvement of 0.2° for the "attenuated" phases relative to the "truncated" phases, but it is clear that, at this resolution, the truncation of the negative densities does not cause serious errors.

6.4 Effect of errors in the initial phase set

Three cycles of the solvent flattening procedure were applied to a number of different starting phase sets, with r.m.s. phases errors between 57.3° and 73.5°. The resulting improvement in the phases is given in Table 6. It is noteworthy that the final r.m.s. phase error for the gold derivative is below 60°, suggesting that the structure could have been solved on the basis of this derivative alone. Starting with the iodinated substrate phases the final r.m.s. error is significantly higher, and in view of the experiences with the original m.i.r. phases it seems unlikely that this map could have been fully interpreted. Even when the final m.i.r. phases were used as an initial phase set the improvement in the solvent flattened phases was substantial. This result emphasises the point that the solvent flattening is doing more than simply resolving the ambiguity inherent in phases determined using a single isomorphous derivative. The final m.i.r. phases were based on six derivatives, three of which included anomalous scattering in the phasing, and therefore only very few reflections would have a simple bimodal probability distribution.

Table 6. The results of applying solvent flattening to different starting phases. In all cases the envelope was determined using an averaging radius of 10Å and a solvent content of 50%. For the first example only 3% of the protein density was truncated, but for the remainder between 30 and 50% was truncated.

Starting phases (see Table 2)	figure of merit	initial r.m.s. phase error	phase error after 3 cycles
Phase set 3 (with modified lack of closures)	0.67	65.3	56.8
Phase set 5	0.72	57.3	49.0
5mM AuCN (A)	0.49	70.5	59.2
PICM (A)	0.35	73.5	64.6

(A) denotes inclusion of anomalous scattering in phase determination.

7. CONCLUSIONS

The primary conclusion of these experiments is that solvent flattening provides a very powerful, straightforward and computationally inexpensive way of improving isomorphous replacement phases, in full agreement with the results obtained by B.C. Wang from a similar experiment using data from the Bence Jones dimer [4]. The analysis shows that the improvement in phases due to solvent flattening is very similar to that obtained using additional (or higher quality) isomorphous derivative data, but with the added advantage that the phases for the strongest structure factors show a significantly greater improvement. (With isomorphous replacement methods the larger absolute error of the large structure factors limits the accuracy with which they can be phased.)

The molecular envelope determined using the Wang algorithm appears to be very satisfactory, as the use of a "perfect" envelope generated from the model gave little additional reduction in the r.m.s. phase error (Table 4). In these calculations the choice of averaging radius used in the determination of the envelope has very little effect on the final phase error, so the success of the method does not depend critically on this parameter.

The degree to which the negative density within the envelope is truncated or attenuated has a more dramatic effect, and the results obtained here (Table 5) suggest that truncation or attenuation of the electron density at between 30% an 50% of the grid points within the protein envelope provides the best final phases. Experience with other systems is required before drawing any definite guidelines on the best strategy to be adopted, as this may well depend on other parameters such as the resolution of the electron density maps, the quality of the initial phases, and the inclusion or omission of very low resolution terms.

The results obtained using different sets of initial phases (Table 6) suggest that, in terms of the reduction in the r.m.s. phase error, the method is almost as powerful when applied to a set of "good" m.i.r. phases as when used with s.i.r. + anomalous phases, although the improvement in the interpretability of the map will be more striking in the latter case. This refutes the suggestion that the method is more suitably applied to s.i.r. rather than m.i.r. phases. The earlier m.i.r. phase tests (Table 3) clearly indicate that the inclusion of a "poor" derivative such as the samarium still improves the m.i.r. phases, rather than degrading them. (The same result is obtained if the PICM s.i.r. + anomalous phases and the PICM + $SmNO_3$ phases are compared, i.e. a poor derivative when added to a single good derivative still improves the phases). However it is crucial that the derivatives are given the appropriate relative weights in the phasing, so the

estimates of the lack of closures (E) are critical. This can present difficulties, particularly if phase refinement is used, as the lack of closures are frequently underestimated. If there are any centric zones, the simplest approach is to calculate the lack of closure for the centric data only, and use these values when calculating acentric phases. Although the centric lack of closures will probably be overestimates of the true lack of closure for acentric reflections [6], this will at least ensure that the correct relative weights are applied.

The only element of caution that is required when applying solvent flattening arises because the automatic molecular envelope determination may result in poorly defined surface loops being excluded from the molecular envelope. The corresponding density will therefore be set to zero at the map modification step. For this reason it is advisable to examine both the initial m.i.r. map and the solvent flattened map when attempting to follow the polypeptide backbone of an unknown structure. In all other respects solvent flattening appears to be highly beneficial to the quality of the m.i.r. phases, suggesting that it should perhaps be used routinely prior to interpreting m.i.r. phased electron density maps, rather than reserving it as a method of last resort.

I would like to thank Peter Brick, Alan Wonacott and David Blow for many useful, critical discussions of the results presented in this paper.

8. REFERENCES

1. Argos, P., Ford, G.C. & Rossmann, M.G. (1975). Acta Cryst. A31, 499−506.

2. Bricogne, G. (1976). Acta Cryst. A32, 832−847.

3. Schevitz, R. W.,Podjarny, A.D., Zwick, M., Hughes, J.J. & Sigler, P.B. (1981). Acta Cryst. A37, 669−677.

4. Wang, B.C. (1985). In Methods in Enzymology, Vol. 115: Diffraction Methods for Biological Macromolecules, edited by H. Wyckoff, C.H.W. Hirs & S.N.Timasheff. New York: Academic Press.

5. Leslie, A.G.W. (1987). Acta Cryst. A43, 134−136.

6. Terwilliger, T.C. & Eisenberg, D. (1987). Acta Cryst. A43, 6−13.

A Reciprocal Space Algorithm For Calculating Molecular Envelope Using The Algorithm Of B.C. Wang

Andrew G. W. Leslie, Blackett Laboratory, Imperial College, London SW7 2BZ.

A suite of programs designed to improve the quality of protein electron density maps has recently been developed and distributed by B.C. Wang and colleagues (Wang, 1985). The basis of their method is to use the electron density map to determine a molecular envelope and then to set the electron density in the solvent region to a constant value (solvent flattening) and apply a positivity constraint to the electron density in the protein region. The modified electron density map is Fourier transformed, and the resulting phases combined with the original m.i.r. (or s.i.r.) phase information. The combined phases are then used to calculate a new electron density map, and the whole procedure is repeated iteratively until there is no further improvement in the quality of the electron density.

Huber's group have successfully used the solvent flattening option in the structure determination of Human alpha-1 proteinase inhibitor (Loebermann et al., 1985) the photo-synthetic reaction centre (Deisenhofer et al., 1984) and a light harvesting biliprotein (Schirmer et al., 1985), all at 3A resolution, and similar but less dramatic results have been obtained by Wang and colleagues in the structure determination of cytochrome c5 at 2.5A resolution (Carter et al., 1985). As expected, the method is most powerful when the solvent content of the crystals is high (70% for the structures from Huber's laboratory).

The concept of using solvent flattening to improve isomorphous replacement phases is not new, and all the necessary programs are available in Bricognes molecular averaging package. Sigler and colleagues in Chicago (Schevitz et al., 1981) used the same approach to produce a dramatic improvement in the electron density map of fMet tRNA at 4A resolution (also 70% solvent). What is novel about Wang's approach is the algorithm that he uses to determine the molecular envelope from the original electron density map. Instead of relying on visual inspection of the map (usually using an interactive graphics display), Wang's procedure has the advantage of being fully automatic. The first step in this procedure is to calculate an "averaged" map from the starting m.i.r. map, by replacing the electron density at each grid point by the weighted average of the electron density at all surrounding grid points within a sphere of radius "R".

The weighting function used is:

$$w(i) = 1 - r(i) / R \qquad \text{for } \rho(i) > 0$$

$$= 0 \qquad \text{for } \rho(i) < 0$$

where rho(i) is the electron density at grid point "i", at a distance r(i) from the centre of the sphere. It is important to realise that because negative densities are ignored (ie given a weight of zero), the result is NOT the same as simply calculating a map at low resolution. The second step is to compute a histogram of the electron densities in the resulting averaged map, and to choose a "solvent level" so that the number of grid points with density less than this solvent level corresonds to the expected solvent content of the crystal. (The solvent content can be estimated using the formula given by Matthews (1968) based on the unit cell contents and the protein molecular weight.) All grid points in the averaged map with a density less than the solvent level are then considered to be in the solvent, while the remainder define the protein.

The optimum value of the averaging radius "R" depends primarily on the resolution of the map and to a lesser extent on its quality (ie the noise level in the solvent region). Typically a value between 8A and 10A is used to average a 3A resolution m.i.r. map.

The calculation of the average map can be extremely expensive in c.p.u. time, particularly since Wang's distributed programs require that the calculation is done in space group P1. As an example, chloramphenicol acetyl transferase (CAT) crystallises in space group R32 with equivalent hexagonal cell parameters a = 107.6A, c = 123.4A. A 3A resolution map calculated on a 1.1A grid was averaged using a radius R = 10A; this calculation required 35 hours c.p.u. time on a VAX 11/750.

The calculation can be made very much faster by using reciprocal space methods based on the Fast Fourier Transform. Wang's averaging procedure in real space is directly equivalent to convoluting the truncated m.i.r. map (ie the m.i.r. map with all negative electron density values set to zero) with the weighting function w(r) given by:

$$w(r) = 1 - r / R \qquad r < R$$

$$= 0 \qquad r > R$$

This may be written as:

$$\rho_{av}(i,j,k) = \rho_{tr}(i,j,k) \hat{} w(r)$$

where rhoav is the averaged map, rhotr is the truncated map and " ^ " denotes convolution.

From the convolution theorem it follows that

$$FT[rhoav(i,j,k)] = FT[rhotr(i,j,k)] * FT[w(r)]$$

where FT[] denotes the Fourier transform. The Fourier transform of the truncated map is readily calculated using the standard FFT program package, and it can be shown that the Fourier transform of w(r) is given by:

$$g(s) = FT[w(r)] = Y(uR) - Z(uR)$$

where:

$$s = 2 * \sin(\theta) / \lambda$$

$$u = 2 * \pi * s$$

$$Y(x) = 3(\sin(x) - x*\cos(x)) / x^3$$

$$Z(x) = 3(2*x*\sin(x)-(x^2-2)\cos(x)-2) / x^4$$

(See James (1948) p466 for a similar example).

Thus to compute the averaged map, the structure factors obtained by back-transforming the truncated map are multiplied by the function g(s) and the modified coefficients are used to calculate a new map which will be identical to that produced by averaging in real space. In the case of CAT, the c.p.u. time was reduced from 35 hours to 40 minutes, even though the calculation was performed in space group P1.

The function g(s) is plotted in Figure 1 for a radius R = 10A. It is similar in form to the transform of a sphere (which would correspond to the weighting function w = 1 for r < R, w = 0 for r > R) but falls off rather less rapidly. The function is less than 0.001 for Bragg spacings less than 5A, and therefore Fourier terms corresponding to spacings less than 5A will make no significant contribution to the averaged map.

The averaging procedure can easily be modified to use different weighting functions w(r), providing that the Fourier transform g(s) can be calculated analytically. Tests using the function:

$$w = 1 - (r/R)^2$$

gave very similar results to the original weighting function, suggesting that the averaged map, and hence the molecular envelope, is rather insensitive to the precise form of the weighting function.

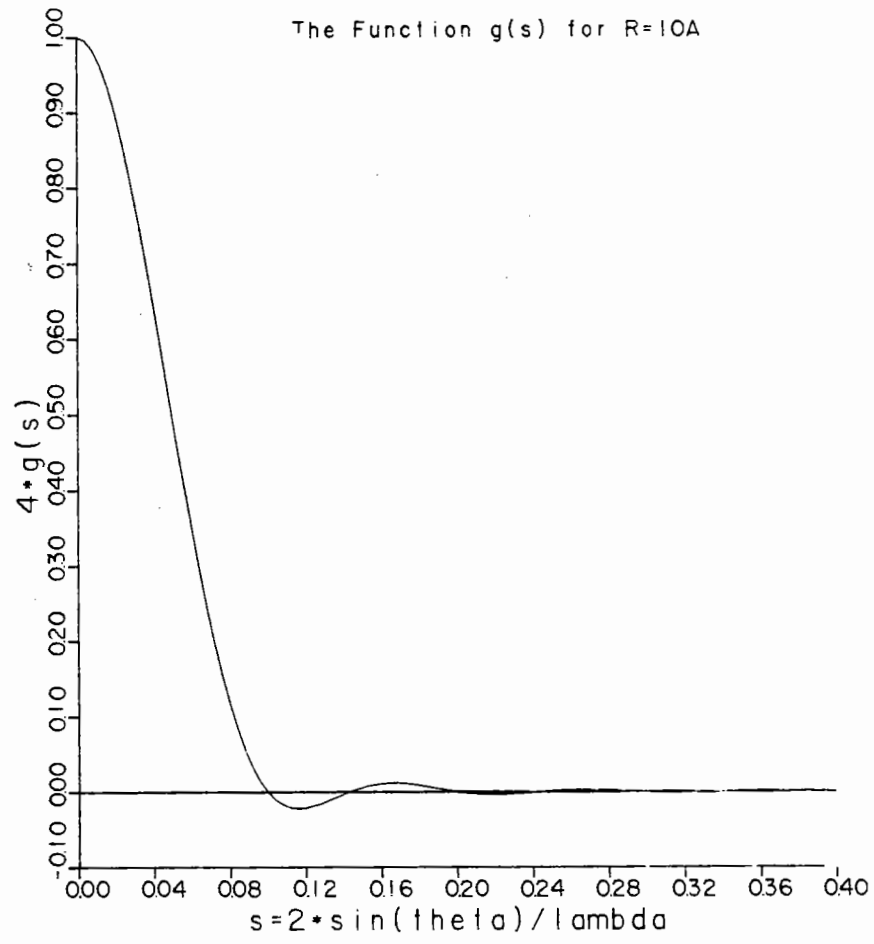


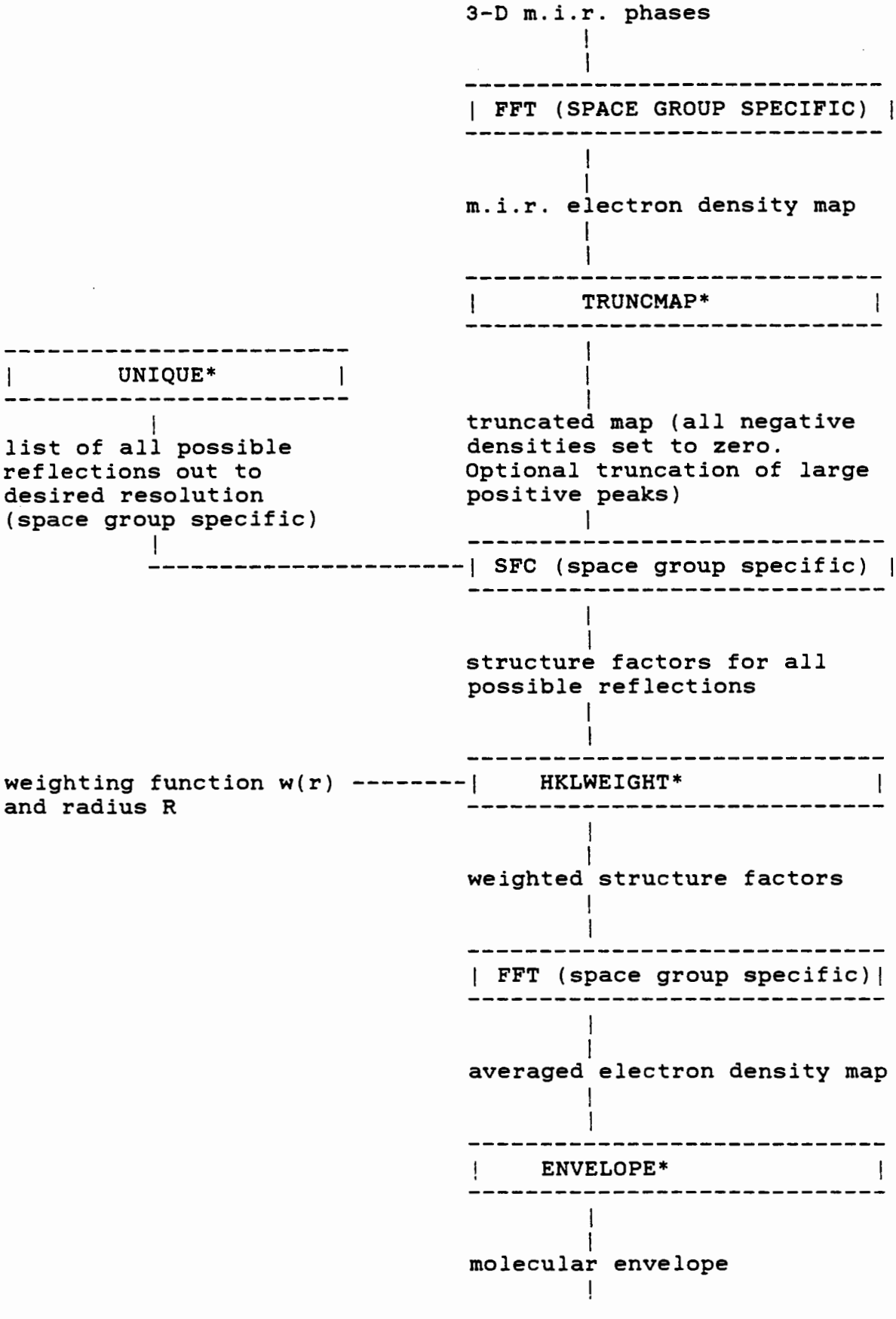
Figure 1

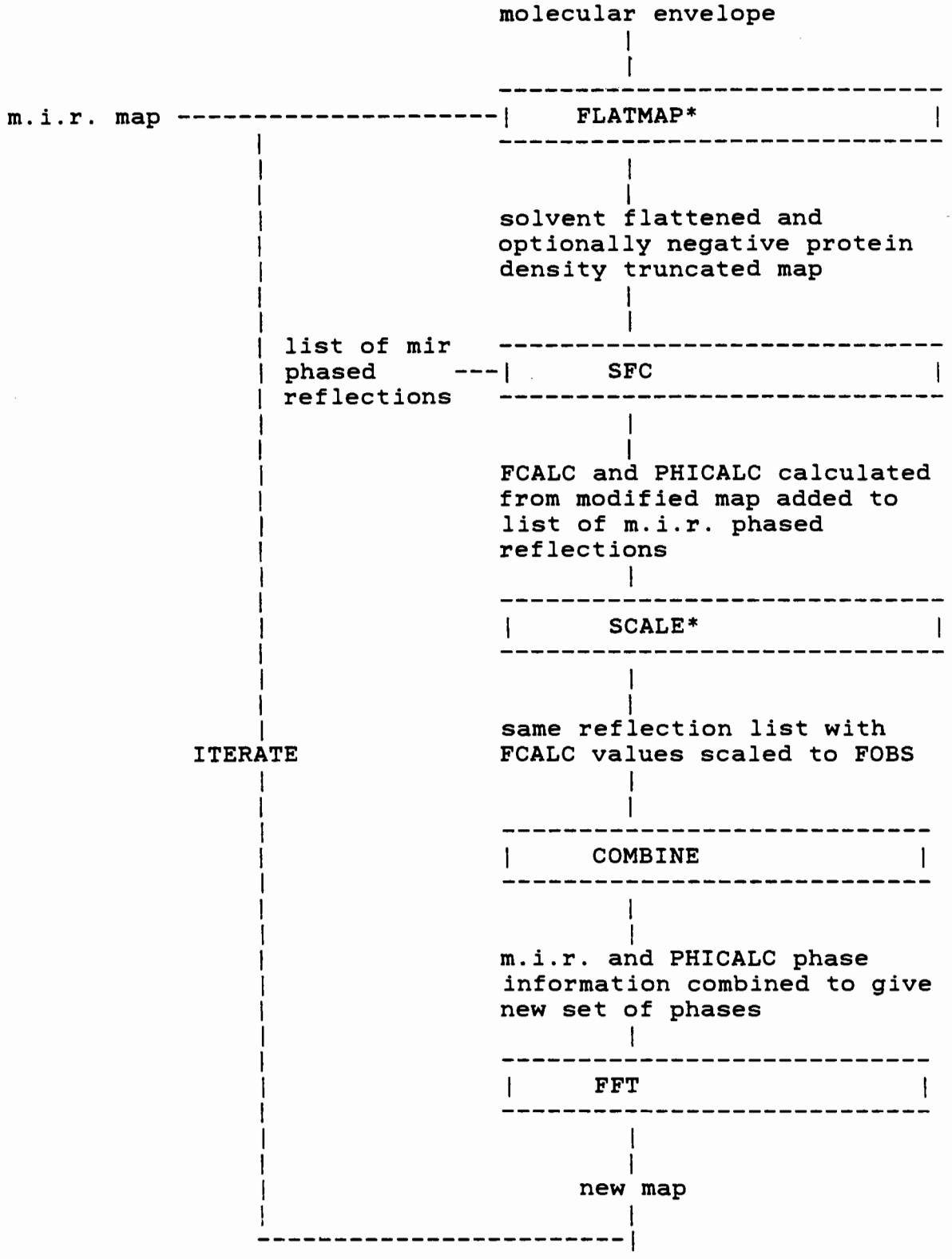
Three practical points are worthy of mention:

1) When using the reciprocal space algorithm, it is essential that all possible structure factors are calculated from the back-transform of the truncated map and included in the calculation of the averaged map. In particular, all low resolution terms must be used, even if those terms were not included in the calculation of the original m.i.r. map. There must be NO low resolution cuttoff in either calculation.

2) Unless there is a version of the back-transform program (SFC) for the correct space group (in which case the program will generate a list of all reflections by default) it will be necessary to provide the program with a list of all possible reflections out to the desired resolution limit. This list is generated by program UNIQUE.

3) The step which truncates the m.i.r. map can also be used to eliminate large positive peaks in the map which could otherwise distort the local molecular boundary. Such peaks can arise from several sources, such as ripples around heavy atom positions, build-up of errors on crystallographic symmetry axes or the presence of metal ions in the protein structure.


3-D m.i.r. phases
FFT (SPACE GROUP SPECIFIC)
m.i.r. electron density map
TRUNCMAP*
UNIQUE*
truncated map (all negative densities set to zero. Optional truncation of large positive peaks)
list of all possible reflections out to desired resolution (space group specific)
SFC (space group specific)
structure factors for all possible reflections
weighting function w(r) and radius R
HKLWEIGHT*
weighted structure factors
FFT (space group specific)
averaged electron density map
ENVELOPE*
molecular envelope



A * denotes programs not currently available in the CCP program package.

Wang's package has been modified to incorporate the reciprocal-space map averaging algorithm and the standard CCP map format and LCF data format. The sequence of steps required to perform one cycle of solvent flattening is given above. Usually the envelope is only determined once, from the original m.i.r. map, so this step need not be repeated in subsequent cycles.

ENVELOPE and FLATMAP are modified versions of Wang's programs ENVELP.FOR and DSFLT.FOR. These are the only programs that have been retained from Wang's distributed package. The changes allow reciprocal space calculation of the envelope and the use of standard CCP map and data file formats. SCALE is the Cambridge program SCALENEW.FOR. UNIQUE, TRUNCMAP, HKLWEIGHT have been written by the author, from whom copies may be obtained.

REFERENCES

Carter, D.C., Melis, K.A., O'Donnell, S.E., Burgess, B.K., Furey, W.F. Jr., Wang, B.C., and Stout, C.D. (1985) J. Mol. Biol. (184) 279-295

Deisenhofer, J., Epp, O., Miki, K., Huber, R. and Michel, H. (1984) J. Mol. Biol. (180) 385-398

James, R.W. (1948) The Crystalline State Vol. II : The Optical Principles of The Diffraction of X-rays, G. Bell and Sons Ltd., London.

Loebermann, H., Tokuoka, R., Deisenhofer, J. and Huber, R. (1984) J. Mol. Biol. (177) 531-556.

Matthews, B.W. (1968) J. Mol. Biol. (33) 491-497.

Schevitz, R.W., Podjarny, A.D., Zwick, M., Hughes, J.J. and Sigler, P.B. (1981) Acta Cryst. (A37) 669-677.

Schirmer, T., Bode, W., Huber, R., Sidler, W. and Zuber, H. (1985) J. Mol. Biol. (184) 257-277.

Wang, B.C. (1985) in Diffraction Methods for Biological Macromolecules (Wyckoff, H., ed.) Academic Press, New York.

PRACTICAL PROBLEMS OF ISOMORPHOUS REPLACEMENT

by

D.M. BLOW, K. HENRICK and A. VRIELINK

Blackett Laboratory, Imperial College, London SW7 2BZ.

1. INTRODUCTION

This paper is about the difficulties in using the isomorphous replacement method to solve a protein structure. It will be illustrated with two examples from structures being studied in our group at Imperial College; glucose isomerase which has been solved by Kim Henrick, and which raised no overwhelming difficulty in the isomorphous replacement stage; and cholesterol oxidase, being studied by Alice Vrielink, and which exemplifies some of the difficulties which can occur.

In applying the isomorphous replacement method to a protein, various crystals soaked with heavy atoms are surveyed, and those showing appropriate differences of diffracted intensity are chosen for three-dimensional data collection. As the data from each heavy atom derivative is processed, a difference Patterson is calculated. Sooner or later, a difference Patterson will be found which shows some peaks which are outstandingly above the background. These outstanding peaks are interpreted in terms of a small number of heavy atom sites. The sites are used to compute approximate phases using the single isomorphous replacement method. These phases are used to compute difference electron density maps for the other sets of heavy atom data, allowing heavy atom sites to be assigned for these derivatives. We are then ready to refine the heavy atom parameters and proceed to solve the structure.

In practice, the procedure is usually not straightforward. The main obstacle arises from the difficulties of the single isomorphous replacement method. Particular difficulties arise when the same site is occupied in several different heavy atom derivatives. The problem of determining the occupancy at this site interacts with the correct assignment of an overall scale factor to the observed structure amplitudes.

2. THE SINGLE ISOMORPHOUS REPLACEMENT METHOD

Even if there were no errors, the single isomorphous replacement method (SIR) does not determine the phase angles. In a simple case where the arrangement of heavy atoms is centrosymmetric, all that the SIR method can do is to determine the cosine of the phase angle. That is, the real part of the structure factor is determined; the imaginary part is not. Using these real structure factors, the computed map has a centre of symmetry imposed upon it.

When there is a less symmetric arrangement of heavy atoms, there will always be partial centres of symmetry in the heavy atom structure, for example, half-way between two heavy atom sites. This partial symmetry is carried by the phase angles derived by the single isomorphous replacement method, and it leads to ghost peaks in the computed maps. These spurious peaks are an inevitable consequence of using the SIR method[1].

In the multiple isomorphous replacement method (MIR), the phase angles from different derivatives should confirm each other, and this provides a method of refinement of the parameters. No such refinement is possible in the SIR method, and only statistical criteria are available to refine parameters[2].

A serious problem concerns the overall scale factor of the observed structure amplitudes of the heavy atom derivatives. The usual procedure is to use the Wilson method to establish a scale factor bringing the two sets of observed structure amplitudes to the same scale. We are expecting that the derivative will include some heavy atoms, and in their pioneering work Green, Ingram and Perutz[3] showed that the presence of one mercury atom in the haemoglobin asymmetric unit increased the mean structure amplitude by a few percentage. In the CCP4 suite, the programme RFACTOR includes this factor as an arbitrary factor, specified by the user, traditionally set to 1.05, which is applied to the structure amplitudes. Various methods are available to estimate the scale factor but in practice they define it only to 2−3% accuracy.

This overall scale factor has a direct effect on the computed phase angles. Figure 1 shows how, in a typical case, a 20% change in scale factor can change the indicated SIR phase through 180 degrees. The effect of this change of scale factor is to add a component to F_{best}, and this component is always exactly in phase with the heavy atom component. It is therefore equivalent to adding an image of the heavy atom to the protein structure factor. In the example chosen in fig 1, a 5% change in scale factor is equivalent to adding or subtracting a fully occupied heavy atom to the electron density (Table 1).

Not surprisingly a frequent feature of SIR electron density maps or difference maps is the existence of large peaks or holes at the SIR heavy atom sites.

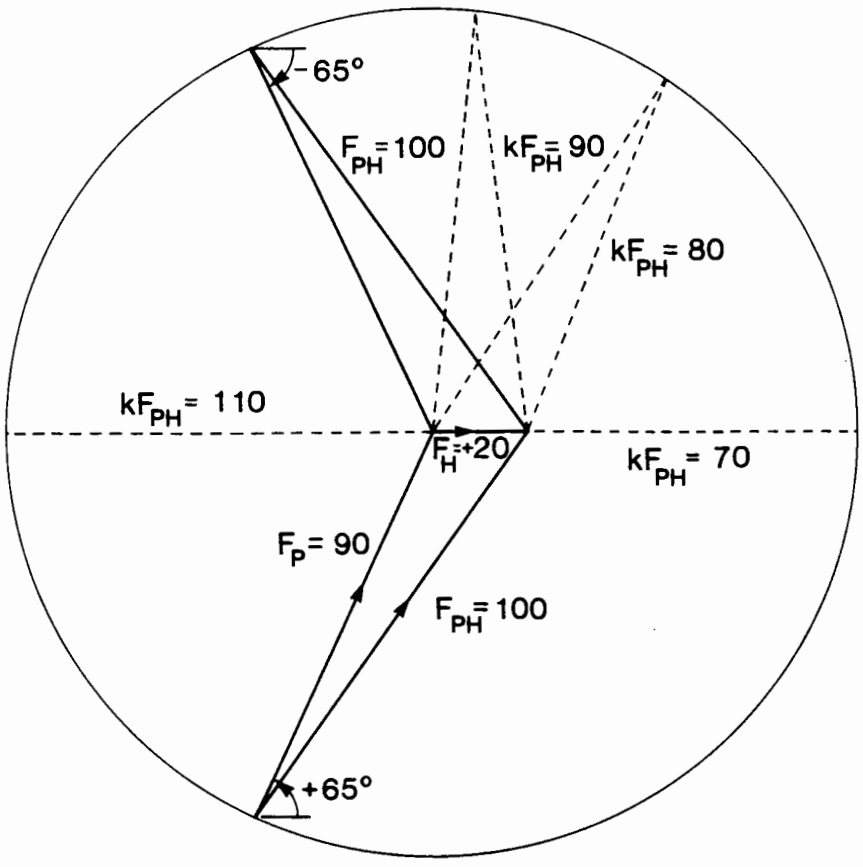


Fig.1. Harker construction for the case $|F_P|$= 90 el, $|F_{PH}|$= 100 el, F_H= + 20 el, giving α_{sir} = ± 65°. If a scale factor k is applied to the structure factor of the heavy atom compound, the resulting α_{sir} changes. The figure shows the cases k= 0.7, 0.8, 0.9, 1.0, 1.1, with α_{sir} varying from 0° to 180°.

Table 1 The "best" estimate of the protein structure factor indicated by single isomorphous replacement in the cases illustrated in Fig.1. F_{best} is calculated as $|F_P|\cos\alpha_{sir}$. In the last column, the value x is calculated from $F_{best} = |F_P|\cos\alpha_1 + xF_H$, where α_1 is the value of α_{sir} when the scale factor $k = 1.0$. Change of the scale factor k is equivalent to adding a component of F_H to F_{best}. In this case a 5% change in the scale factor is enough to change x by 1 unit. An increase of k by 5% will add a fully occupied heavy atom H to the electron density map based on the "best" phases.

k	$\cos\alpha_{sir}$	F_{best}	x
0.7	−1.00	−90	−6.4
0.8	−0.58	−52	−4.5
0.9	−0.03	−3	−2.1
1.0	+0.42	+38	0
1.1	+1.00	+90	+2.6

3. IMPROVEMENT OF THE SIR METHOD

Most of the difficulties with the SIR method vanish in the case of centrosymmetric reflections, whose signs are determined almost without ambiguity (except for experimental error). The heavy atom parameters of lysozyme were refined entirely on the centro−symmetric reflexions. In discussion of this paper Dr E.J. Dodson suggested that scaling problems are most satisfactorily solved by the use of centric reflexions[4].

The contrasting approach takes the view that noise in difference maps is reduced when very large numbers of terms are included, and uses the probabilistic approach of the "best" electron density map for non−centrosymmetric phases[5]. Where anomalous scattering data are available, the phase estimate of SIR can be improved, and electron density difference maps are calculated by appropriate combinations of isomorphous and anomalous scattering differences[6].

These two approaches give some help in getting over the problems of the SIR method, but they do not always remove them. A more general approach is to use data from other isomorphous derivatives to help refine the parameters. The crucial difficulty here is that the parameters are linked to each other through the phase calculation. Refinement of heavy atom parameters, based on calculated phases using these same parameters, is hopelessly sluggish[7]. Bricogne[8] has analysed this problem, but it is not yet resolved in practice.

4. GLUCOSE ISOMERASE

A new structure solved these methods (by K.H.[9]) illustrates how the procedure sometimes causes few difficulties. The trigonal form of glucose isomerase (space group $P3_121$, a = 106 A, c= 154 A)[10] contains two molecular subunits in the crystal asymmetric unit, related by a two−fold axis whose direction was determined using the rotation function. The local symmetry was helpful in resolving doubts about the heavy atom sites. Centrosymmetric $h0l$ precession photographs were used to search for suitable heavy atom derivatives and four were found (Table 2), referred to as Pb, Au, Hg, and Pt. Three dimensional data were collected for these derivatives.

Table 2

Cholesterol oxidase: Derivatives and phase refinement parameters

	$Pb(OAc)_2$	$KAuI_4$	K_2HgI_4	K_2PtCI_4
Resolution (Å)	2.7−10.0	2.7−10.0	2.7−10.0	3.0−10.0
R_{deriv} (%)*	14.6	21.9	23.6	22.1
No. metal sites	2	3	10	16
Phasing power	1.69	2.17	2.01	1.56

* Derivative R−factor: $R_{deriv} = \frac{\Sigma_h | F_{deriv}(h) - F_{nat}(h) |}{\Sigma_h | F_{nat}(h) |}$

Fortunately, the first difference Patterson computed was the easiest to solve − the Pb derivative. This Patterson presented special features because both z coordinates are small and both x coordinates are similar, but the two sites were correctly assigned.

The Au derivative is essentially a single cluster of sites. It gives a simple difference Patterson, but one whose solution would always have caused worries, because the Patterson peak comes at a special position with $v=0$, $w=1/3$. The Pb−phased SIR difference map shows the Au cluster clearly.

The Hg derivative gives a cluster at the same site as Au, but the difference Patterson is extremely complicated because there are several other major sites. The Pb−phased SIR difference map shows two pairs of peaks related by the local two−fold axis, which also relates the two Pb peaks.

All the heavy atoms are strong anomalous scatterers in 0.88A radiation (station 9.6). The anomalous scattering from the Pb derivative was used to compute a "combined" difference map[6] from SIR and anomalous phase information. These maps were calculated in both possible space groups $P3_121$ and $P3_221$. In $P3_121$, there were very clear peaks, but none in the $P3_221$ map. The Hg SIR phases and anomalous data were used to compute a combined difference map for the Pb derivatives. The Pb sites came up in both maps, but much higher in $P3_121$. No new Pb sites were revealed.

The Pt derivatives could only be interpreted when the Pb, Au and Hg derivatives were all used in the multiple isormorphous replacement method. It then showed 16 sites exceeding 5 standard deviations in the difference map, including 3 pairs of sites related by the local two−fold axes.

In this case the interpretation of the difference maps was done entirely by the use of a peak searching programme which lists the position of the highest peaks. It was not necessary to inspect the individual map sections. The clusters of peaks Hg1 and Au1 were not studied in detail at this stage. Recently calculated phases from the refined protein structure were used to compute a difference map for the Hg derivative. When the peak was displayed on the computer graphics system, using FRODO, very clear density for an HgI_3^- group was revealed (Fig 2).

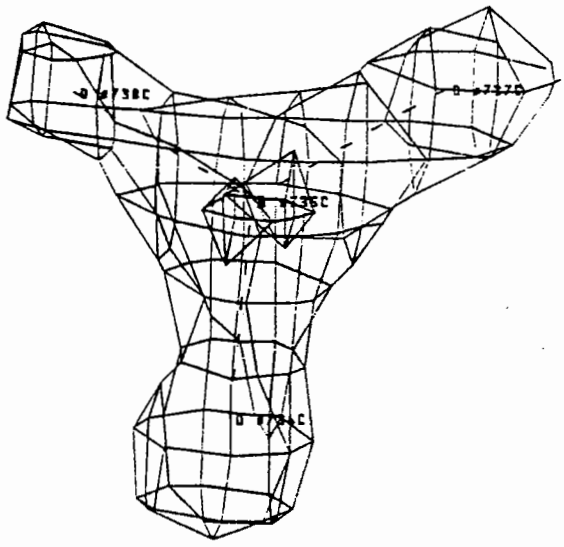


Fig.2 Difference density at the cluster of peaks Hg1 in glucose isomerase difference map

5. CHOLESTEROL OXIDASE

The application of isomorphous replacement to cholesterol oxidase (orthorhombic, a= 68A, b= 85A, c= 88A) has proved troublesome. A.V. discovered four derivatives (Table 3) referred to as PCMB, Hg, Au, and Pt. They all gave large differences, the Pt derivative significantly more than the others. With the radiation used, only PCMB and Hg were useful anomalous scatterers.

Fortunately, once again, the first difference Patterson map (PCMB) was readily interpretable, indicating a single mercury site (site A). The next derivative (Hg) gave a difference Patterson map showing the same Harker peaks as for PCMB, but considerably weaker, and other peaks suggesting a second mercury site (site B). A difference map using the PCMB SIR phases showed a large peak at site A and a weaker peak at site B. Because of the difficulties of the SIR method, there was no real evidence about the relative occupancy of the two sites. The site B peak was very convincing because of its shape (Fig.3).

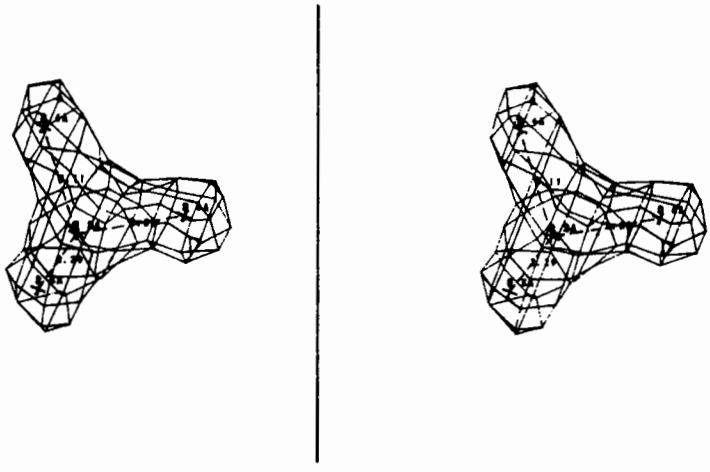


Fig.3. Difference density at site B in cholesterol oxidase difference map.

The Au and Pt derivatives both gave uninterpretable difference Patterson maps. Difference electron density maps, whether calculated by SIR, or by using both the PCMB and Hg derivatives showed strong peaks at site A, but the difference Pattersons did not indicate any substitution at this site. Careful study of the maps from centrosymmetric projections showed peaks at Site A in all cases. Although there appeared to be two good derivatives, and at least one further derivative giving reasonable differences, the relative occupancies at sites A and B could not be determined.

Data for a new derivative (Os) has now been processed. The difference Patterson shows that this is the first derivative which does not substitute at Site A, binding at two different sites C and D. The Table summarises the peak heights for the four derivatives PCMB, Hg, Au, and Pt in difference Fourier maps using different sets of SIR and MIR phases. The preliminary SIR phases for Os gave large positive peaks at sites C and D for all four derivatives: this is clearly an artefact due to incorrect scaling. In just the same way the PCMB SIR phases, and the MIR phases using PCMB, Hg and Pt only had all indicated positive occupancy at Site A for the Os derivative.

Table 3. Summaries of peak heights at the various heavy atom sites of cholesterol oxidase, in difference Fouriers phased in different ways.

The entries in the table show peak heights in arbitrary units, (stated as a number of standard deviations in brackets). Peak heights are underlined when the peak appears at a site not used in phasing. A question mark indicates an underlined peak whose significance is questionable. The final column indicates whether heavy atom substitution at the site is confirmed.

		Phases				
		PCMB SIR (site A)	PCMB HgI_4 $PtCl_4$ MIR+anom	OsO_4 SIR (Sites C+D)	PCMB HgI_4 OsO_4 $Au(CN)_2$	
Derivative	Site		(Sites A+B)		MIR+anom	
PCMB	A	–	–	100(17)	100(71)	Yes
	B				6(4)	No
	C			89(16)	6(4)	No
	D			56(10)	7(5)	No
HgI_4	A	100(57)	100(61)	72(19)	100(50)	Yes
	B	15(9)	45(27)	49(13)	59(29)	Yes
	C			123(33)	10(5)	No
	D			85(23)	11(6)	No
$Au(CN)_2$	A	100(59)	100(39)	32(6)?	100(40)	Yes?
	B		23(9)		23(9)	No
	C			151(26)	20(8)	No
	D			100(17)	17(7)	No
	Au	20(12)	77(30)	58(10)	75(30)	Yes
$PtCl_4$	A	–	100(25)		100(26)	No
	B		29(7)		32(8)	No
	C			134(20)	22(6)	No
	D			100(15)	36(9)	No
	Pt		67(17)	31(5)?	38(9)?	Yes?

Using the Os derivative, which does not bias the interpretation at the mercury sites A and B, there is clear evidence that Hg binds at both A and B, while Pt does not bind at either. Au binds at its own site but the Os phases suggest that it may bind weakly at site A.

The difficulties are still acute because the only derivatives which have a good phasing power beyong 4A are PCMB and Hg, both of which bind strongly at site A. The other heavy atoms are all bound in sites with high temperature factors, and nothing can prevent PCMB and Hg from dominating the phasing. The higher resolution reflexions are phased almost by single isomorphous replacement.

The last column in Table 3 indicates which sites have been confirmed. The proposed Pt site is still doubtful, giving a peak which is only just significant in the Os−phased map, and a somewhat stronger peak at the same site using MIR and anomalous scattering based on all the other derivatives.

6. CONCLUSION

The case of glucose isomerase shows that the procedure can run smoothly. However, the standard procedure routinely gives positive peaks in difference maps at the sites of heavy atoms used in the phasing. It is important to improve the overall scaling procedure to eliminate this bias. In particular, when running RFACTOR, users are advised to set the arbitrary scale factor, which is applied to the derivative F's after Wilson scaling, to 1.0.

The cholesterol oxidase structure poses these problems in an acute form because nearly all heavy atom derivatives, and particularly those which are useful at higher resolution, bind at one particular site.

The best remedy, obviously, is to find another useful isomorphous derivative, substituting at a different site. Other possibilities are

- to give more weight to anomalous scattering measurements which tend to resolve the ambiguity of single isomorphous replacement phase angles;

- to take the PCMB derivative as "parent", considering the native enzyme to have a negative heavy atom at site A (in common with the other less useful derivatives), and the Hg derivative to have a single major substitution at site B. Thus two good derivatives can be treated as substitutions at different sites.

1. D.M. Blow and M.G. Rossmann, Acta Cryst. 14, (1963) 1195.
2. M.G. Rossmann, Acta Cryst. 13, (1960) 221.
3. D.W. Green, V.M. Ingram and M.F. Perutz, Proc. Roy. Soc. A 225 (1954) 287.
4. E.J. Dodson in Crystallographic Computing Techniques (ed. F.R. Ahmed). (Copenhagen: Munksgaard, 1976) pp 259−268.
5. D.M. Blow and F.H.C. Crick, Acta Cryst. 12, (1959) 794.
6. B.W. Matthews, Acta Cryst. 20, (1966) 230.
7. D.M. Blow and B.W. Matthews, Acta Cryst. A29, (1973) 56.
8. G. Bricogne, in Computational Crystallography (ed. D. Sayre) Clarendon Press, Oxford (1982) pp 223−230.
9. J. Akins, P. Brick, H.B. Jones, N. Hirayama, P−C Shaw and D.M. Blow, Biochim. et Biophys. Acta 874 (1986) 375.
10. K. Henrick, D.M. Blow, H.L. Carrell and J.P. Glusker, Protein Engineering 1, (1987) 467.

ITERATIVE MOLECULAR AVERAGING AND PHASE REFINEMENT OF TWO HLA-A2 CRYSTAL FORMS

by

M. A. Saper[+], P. J. Bjorkman[+*], and D. C. Wiley[+]
[+]Howard Hughes Medical Institute and
Department of Biochemistry and Molecular Biology,
Harvard University, 7 Divinity Ave, Cambridge, MA 02138, USA

[*]Current address: Department of Medical Microbiology,
Stanford University, Stanford, CA 94305, USA

1. INTRODUCTION

The human histocompatibility (or leukocyte) antigen, HLA, is a cell surface glycoprotein found on virtually all cells and identifies that cell as self or foreign to the immune system. Moreover, cytotoxic T lymphocytes can kill virus-infected cells by recognizing self HLA in complex with foreign peptide antigens (e.g. from viral proteins). The structure of papain-cleaved HLA-A2, described at 3.5Å resolution [1,2], shows that the highly polymorphic α_1 and α_2 domains form a deep, narrow cleft predicted to be the peptide binding and T cell receptor recognition site.

This paper presents details of iterative molecular averaging of two HLA crystal forms that enabled us to accurately trace the polypeptide chain and align the amino acid sequence.

2. CRYSTALLOGRAPHIC BACKGROUND

HLA-A2 crystallizes in two different, but highly related, crystal forms: monoclinic and orthorhombic [3]. Table 1 summarizes crystal data and structure solution status as of the time the averaging experiments were carried out. Maps in the monoclinic crystal form phased by multiple isomorphous phases and solvent-flattened by the Wang procedure [4] revealed the locations of the α_3 and β_2-microglobulin domains predicted to fold as immunoglobulin-like constant domains. Model building, phase combination, and solvent-flattening improved the monoclinic maps allowing regions of secondary structure in the α_1 and α_2 domains to be defined but an unambiguous trace and sequence alignment could not be made.

3. GOALS

Since native data and one derivative dataset to 3.5Å resolution were available from an orthorhombic crystal form, we chose to iteratively average the two electron density maps and calculate an 'improved map' for completely tracing the polypeptide chain, positioning side chains according to the known amino acid sequence, and locating regions of functional interest. We also hoped to derive phases for the orthorhombic data that would be sufficient for difference Fourier analysis of peptides soaked into the crystals.

Table 1 - Crystallographic summary of HLA-A2 crystal forms preceeding density averaging experiments [1,3]. This table summarizes data and results obtained by Bjorkman, P.J., Samraoui, B., Bennett, W., and Wiley, D.C.

HLA-A2 Crystal Summary

	Monoclinic	Orthorhombic
Crystallization Conditions	10-15% polyethylene glycol 8000 pH 6.2 0.3 × 0.3 × 0.05 mm	10-15% polyethylene glycol 8000 pH 6.5 0.3 × 0.3 × 0.1 mm
Space Group	$P2_1$	$P2_12_12_1$
Cell Dimensions	a = 60.35Å b = 80.40Å c = 56.49Å β = 120.42°	a = 60.2Å b = 80.4Å c = 112.2Å
Native Data	97% complete to 2.7Å	82% complete to 3.5Å 51% from 3.5 to 2.7Å
Existing Map	3.0Å MIR map (5 derivatives and anomalous); phase combined with current model and then solvent flattened (Wang method).	3.5Å SIR map (1 derivative) and solvent flattened (Wang method).
Existing Model	80% of backbone 50% of side chains (mostly α_3 and β_2)	

4. MOLECULAR AVERAGING

The theory and applications of iterative molecular averaging and phase refinement have been clearly described in the previous talk by G. Bricogne and in a recent review article by Podjarny et al. [5]. Since the solution of the first virus structures [6,7], molecular averaging of oligomers that crystallize with noncrystallographic symmetry has become common practice. Usually the averaging is combined with solvent flattening to further improve the map.

The technique as first described by Bricogne [8] can be directly applied to other instances of structural similarity such as averaging the electron density of molecules that crystallize in two different crystal forms. Two such examples are described below.

Loebermann et al. [9] crystallized the human α_1-proteinase inhibitor in two space groups: tetragonal and hexagonal. M.i.r. phases to 3.0Å resolution (m = 0.54) derived from 6 derivatives were available for the tetragonal form; s.i.r. phases (m = 0.39) for the hexagonal form. The transformation relating one crystal form to the other was determined with a real space rotation and translation search with the Munich program package PROTEIN [10,11]. The envelopes for averaging were interpretted manually from solvent-flattened maps. Cyclic averaging started with the m.i.r. and s.i.r. maps. After each cycle, new phases were calculated from the averaged map and either phase combined with the m.i.r./s.i.r. phases or used directly with Sim weighting. The final map after 6 cycles (R=0.191, Table 2) allowed an unambiguous chain trace to be made.

	R-factors		
Cycle	Hexagonal	Tetragonal	Map averaged
1	0.348	0.340	2Fo - Fc; phase combination m.i.r. and Fc
2	0.332	0.311	2Fo - Fc; phase combination m.i.r. and Fc
3	0.314	0.292	2Fo - Fc; Sim-weighting
4	0.258	0.228	2Fo - Fc; Sim-weighting
5	0.232	0.205	2Fo - Fc; Sim-weighting
Final	0.191		2Fo - Fc; Sim-weighting

Table 2 - Summary of cyclic averaging of tetragonal and hexagonal forms of human α_1-proteinase inhibitor (data from Table 7, ref. 9).

Varghese et al. [12] applied a similar algorithm in the solution of influenza neuraminadase. Most of the averaging was actually done in reciprocal space. A map from the Tokyo strain (space group I422, 1 subunit/a.u.) was calculated from 3.1Å m.i.r. phases (m=0.48) and solvent-flattened. The similar RI/5+ strain ($P4_12_12$) contained 2 subunits in the asymmetric unit. The molecular electron density from the Tokyo strain was used in a R-factor translation search of the RI/5+ data and then positioned into that cell. Calculated phases from the molecular replacement solution were used to locate the single heavy atom site in RI/5+. Noncrystallographic symmetry averaging and phase extension to 2.9Å in RI/5+ gave an R=0.245. The electron density corresponding to the molecule was then transferred back into the Tokyo cell, Fourier transformed to yield structure factors, and the phases combined with the m.i.r. probabilities. The entire procedure was then repeated. The final R for the RI/5+ strain was 0.246 at 2.9Å.

5. METHODS

The overall scheme of the HLA-A2 averaging is presented in fig. 1. Parallel calculations were done in each space group. Each point within the monoclinic molecular envelope was averaged with the corresponding electron density value interpolated from the orthorhombic map. Grid points in the original map outside of the envelope were solvent-flattened by their mean value. The averaged map was expanded to an entire unit cell and transformed by FFT to obtain calculated structure factors. These phases were then combined with the m.i.r. phase probability, a new map calculated, and the entire procedure repeated. A similar series of calculations were done simultaneously in the orthorhombic space group.

5.1 Determination of Relative Orientation

Molecules of HLA were predicted to pack similarly in both forms based on space group and unit cell similarities [3]. To define the exact transformation between the molecules from each space group, a partial model, derived from fitting the monoclinic m.i.r. solvent flattened map, was used in a 6-dimensional rotation and translation real space search of the orthorhombic s.i.r. solvent-flattened map. One major solution was found which was later refined with the real space refinement option of FRODO [13]. The transformation was essentially a translation to a different origin; the rotation matrix corresponded to only a 1.25° rotation.

To confirm this result, electron density points from the monoclinic map (Fo coefficients, phases calculated from partial model and combined with m.i.r., then solvent-flattened by Wang procedure) and within the presumed molecular boundary were alternatively rotated and translated in the orthorhombic map (Fo, s.i.r. phases, solvent-flattened) using the SEARCH routine of PROTEIN [11]. A three-dimensional translation search, that used the 200 highest density points from the monoclinic map, gave a single solution with a peak heght of six sigma (see fig. 2). Refining this solution with all monoclinic grid points above one sigma (6900 points) gave a result similar to that derived from the partial model search described above.

The correct transformation was confirmed by comparing the coordinates of a heavy atom platinum site common to both crystal forms.

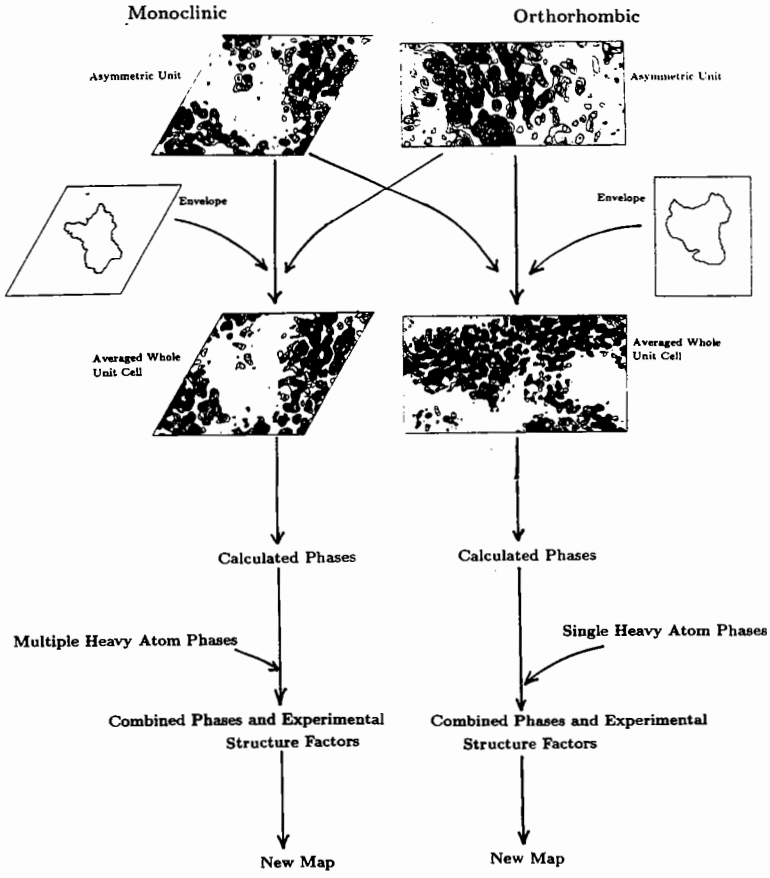


Fig. 1 - Overall scheme of averaging the monoclinic and orthorhombic crystal forms of HLA-A2. In all cases the new maps calculated had 2Fo-Fc coefficients weighted by figure of merit, or Sim-weight.

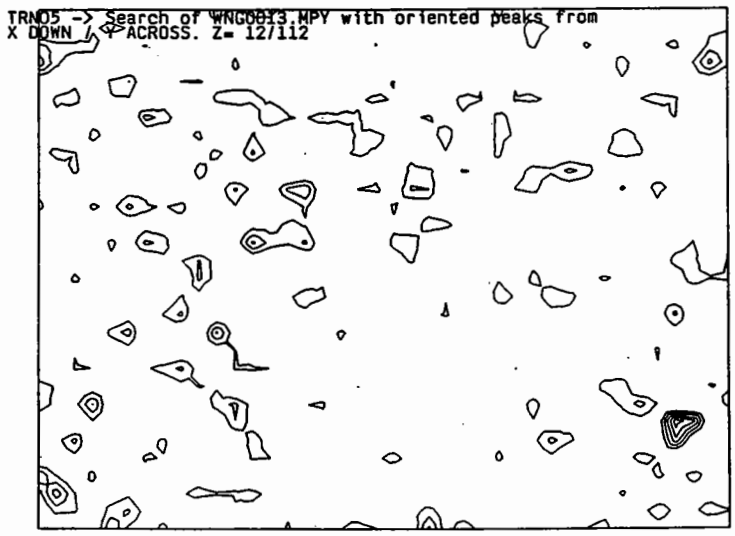


Fig. 2 - Section from real space translation search of orthorhombic 3.5Å s.i.r. and solvent-flattened map using 200 highest peaks from monoclinic map. First contour is at one sigma, with one sigma intervals. Section is z=0.11, x down, y across.

5.2 Envelope determination

Two programs were used to generate envelopes suitable for the iterative averaging procedures. Both programs used model coordinates to define the extents of the molecular volume. The program ENVATOM (based on a similar program by S. Harrison) simply set all grid points within a specified radius (typically 5Å) of each atom as TRUE (or 'inside'). All grid points within a similar distance of symmetry-related atoms (generated with program SYMBUMP)

were set FALSE (or 'solvent'). An envelope determined in the orthorhombic frame could be transferred to the monoclinic frame with programs from the Bricogne suite [8]: GENERATE mode 3 and RECNV1 (similar to RECNV3).

In later stages of averaging, a more precise envelope was needed that did not overlap with neighboring, symmetry-related molecules. The program ENVTOM, derived from a program originally written by Tom Garrett, searches around each grid point to see which atom within a specified radius is closest. If the atom is not a symmetry-related atom then this grid point is set TRUE ('inside'). This program effectively defines the envelope boundary precisely between two neighboring molecules, without overlap. The program can also be used to prepare labelled envelopes for monomers related by noncrystallographic symmetry. Typically, coordinates of the current HLA model in the orthorhombic frame were transformed to the monoclinic frame. Symmetry-related molecules in the orthorhombic space group, as well as symmetry-related atoms in the monoclinic cell transformed back into the orthorhombic cell, were used by ENVTOM to constrain the extent of the molecular envelope. A parallel manipulation was done to calculate the monoclinic envelope. Final envelopes contained about 35% solvent.

5.3 Software

A detailed flowchart of the software used in one cycle of the averaging calculations is shown for the orthorhombic case in Fig 3. Most software was from Bricogne's original suite of programs [8]. For the monoclinic case, we modified GENERATE to include a Q2 matrix that permitted easy skewing into a nonorthogonal frame (grid 2 map).

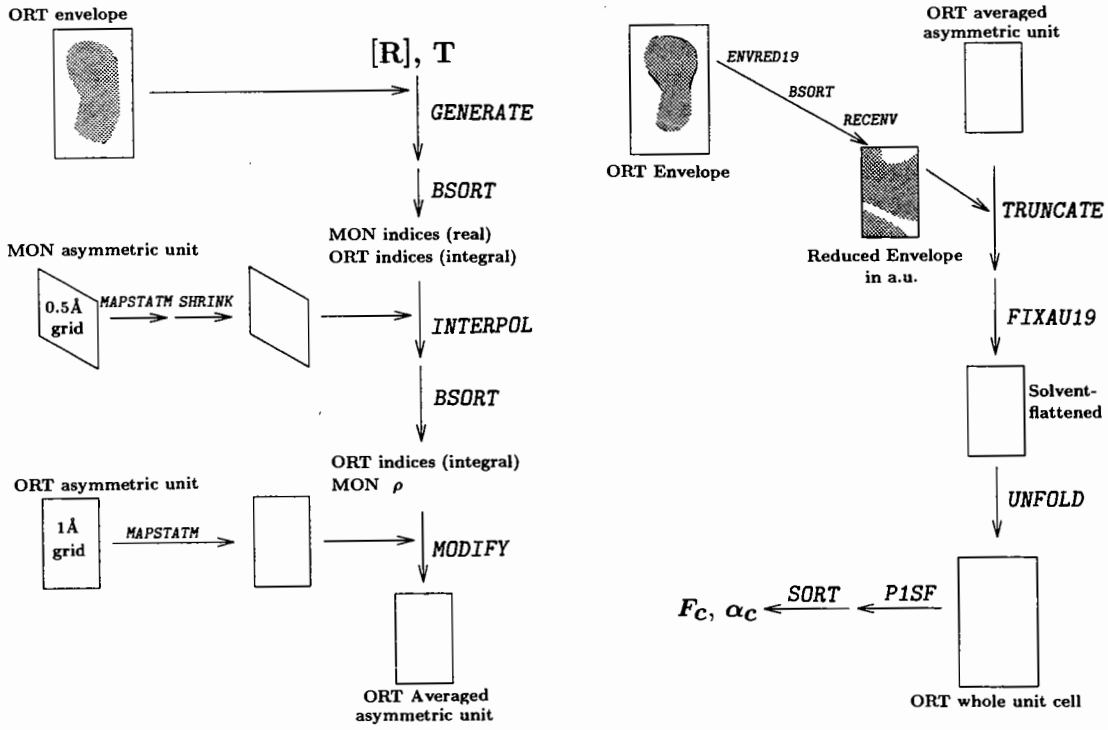


Fig. 3 - Steps in a typical cycle of averaging in the orthorhombic frame. Programs (slanted type) are mainly from Bricogne suite [8]. UNFOLD is a general program to manipulate subsets of maps. P1SF is a FFT structure routine from the PROTEIN package [10,11].

Mode 5 of GENERATE utilizes an envelope to define which points are averaged and which are considered solvent for truncation. All manipulations are done in the map's asymmetric unit which later can be transformed to yield structure factors by a space-group specific FFT routine or (as in our case) expanded to an entire unit cell and transformed with a P1 FFT structure factor program. After a series of mode 5 averaging cycles were completed, the final averaged map is calculated by mode 1 of GENERATE and skews one map and averages it to the other, without an envelope or solvent-flattening.

When averaging two maps together it is, of course, important that they both be on the same relative scale. We did this simply by prescaling the asymmetric unit maps from each space group to have the same standard deviation. More recently this has been found to introduce approximately 10% scaling error. This may be due to statistical differences between all of the points in the asymmetric unit and the set of points actually averaged. Additionally interpolation from a skewed, fine grid map (as in monoclinic) might also alter the density distribution. This has been corrected by implementing an additional pass in the program MODIFY to examine the two lists of values being averaged and determine a scale factor between them. The second pass applies this factor and averages the grid points.

6. RESULTS

6.1 First experiment

The first series of averaging calculations used maps calculated to 3.5Å resolution: orthorhombic (ORT) - Fo, s.i.r. phases, and solvent-flattened with Wang procedure; monoclinic (MON) - Fo, m.i.r. phases combined with phases calculated from partial model, and solvent-flattened. The envelope was calculated with ENVATOM from a preliminary, though incorrect, trace of the entire molecule that accounted for most of the electron density. For each of the first four cycles, calculated phases from the averaged map were combined with the respective m.i.r. or s.i.r. phases and used with m*(2Fo-Fc) coefficients to calculate a new map for the next cycle. In the last two cycles the calculated phases were used directly with Sim-weighted, 2Fo-Fc coefficients. A program originally written by Jim Remington was used for phase combination [11]; maps were calculated with PROTEIN.

To follow convergence of the procedure R-factors between observed amplitudes and those calculated from averaged maps were examined as well as phase changes from previous and first cycles. The final R-factors after 6 cycles were: ORT 20.6%, MON 22.2%. Rms phase change from the starting phases was ORT 79° and MON 68°. Phase change from the penultimate cycle was about 20° suggesting that the procedure had not yet converged.

A final map calculated with mode 1 of GENERATE showed substantial connectivity and side chain density. A new polyalanine model was built connecting all strands of α_1 and α_2 previously seen in other maps. In addition, new regions of helical density were resolved. Examination of mini-maps revealed strong density spanning a helix and a strand which defined the positions of Cys101 and Cys164, a disulphide in α_2. This was crucial in aligning the α_1 and α_2 sequence with the polyalanine trace.

Errors due to the crude envelope appeared in several places. During alignment of the sequence with the trace it was apparent that about 9 residues were missing between 12 and 20. The original MON map revealed that density corresponding to the loop between the first and second strands in α_1 had been truncated by the envelope. Also, no density was resolved connecting domains α_1 and α_2. We suspected that residues 86 - 93 formed a loop extending beyond the current envelope. By superimposing a contoured envelope during interactive refitting, it was obvious when regions of the map were affected by an inaccurate envelope.

6.2 Second Averaging

A new envelope had to include not only the current model, but also regions of suspected structure: residues 86-93, the suspected carbohydrate region, and the unknown or 'extra' density between the two helices of α_1 and α_2 [1]. To do this we positioned dummy atoms in the region of the extra density and the suspected position of 86-93 so that these areas would be placed inside the envelope by the ENVTOM program. The new envelopes contained 67% (ORT) or 80% (MON) of the unique grid points assigned as protein. Solvent-flattening therefore would not affect any more than 33% of the asymmetric unit.

The second round of averaging started from the same maps as the first experiment. Three cycles of averaging done as before with phase combination were followed by seven cycles of averaging with Sim-weighted coeffients. Fig. 4 shows R-factor and phase changes over the 10 cycles. The rms phase difference between the last and penultimate cycle was 9.2° (ORT) and 8.4° (MON). The missing density between 86 and 93 (see above) appeared in this averaged map (see fig. 6 below).

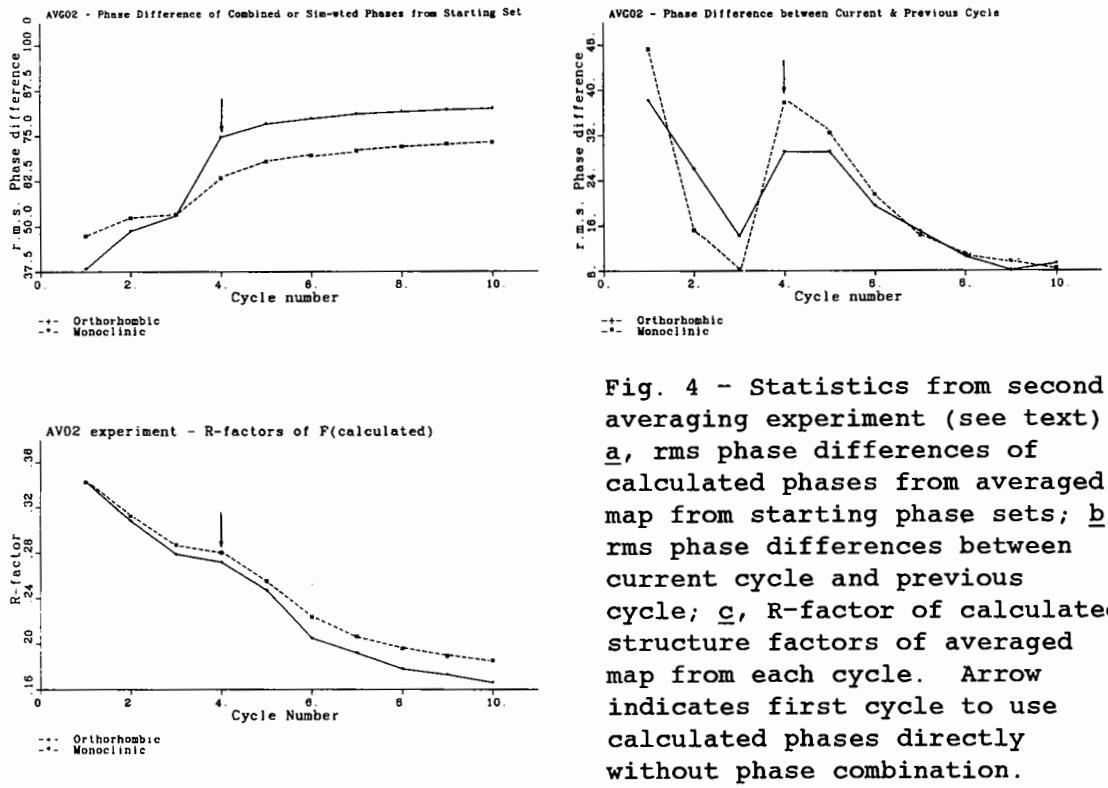
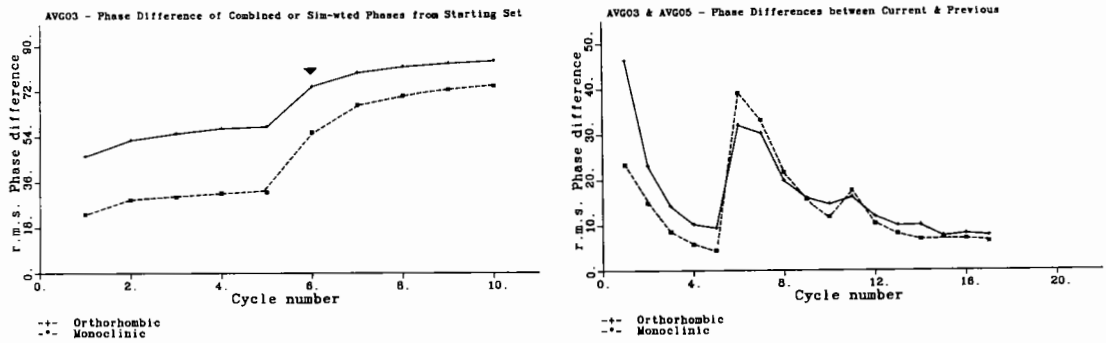


Fig. 4 - Statistics from second averaging experiment (see text). a, rms phase differences of calculated phases from averaged map from starting phase sets; b, rms phase differences between current cycle and previous cycle; c, R-factor of calculated structure factors of averaged map from each cycle. Arrow indicates first cycle to use calculated phases directly without phase combination.

6.3 Third Experiment

The above averaging experiments may have been biased by starting maps that had partial model contributions. To test the effects of this we repeated the entire averaging procedure but started with maps calculated solely with m.i.r. or s.i.r. best phases and m*Fo coefficients.

For the first 5 cycles, the calculated phases from inversion of the averaged map were combined with m.i.r. or s.i.r. phase probabilities to calculate m*(2Fo-Fc)maps . This was followed by 5 cycles with Sim-weighted, 2Fo-Fc coefficients and calculated phases. Similar changes in R-factor and phase changes were seen except that even after 10 cycles the procedure had yet to converge (rms phase change from last to next-to-last cycle: ORT 15^{o}, MON 12^{o}, fig. 5). The final averaged map from this experiment was used to refit the side chains for the published structure at 3.5Å [1]. At least 80% of the side chains could be oriented with confidence.



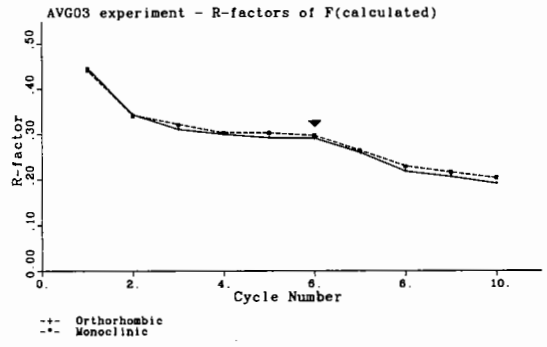


Fig. 5 - Statistics from third averaging experiment (see text). a, rms phase differences of calculated phases from averaged map from starting m.i.r. (MON) or s.i.r. (ORT) best phases (first 10 cycles only); b, rms phase differences between current cycle and previous cycle (for all cycles); c, R-factor of calculated structure factors from averaged map for each cycle (first 10 cycles only). Arrow indicates first cycle to use calculated phases directly without phase combination.

Before running more cycles to allow the preceeding experiment to converge, we redetermined the transformation relating the two crystal forms by comparing HLA models refined in both space groups at 3.5Å with the program CORELS [14]. Models oriented by this transformation differed from the original by less than 0.5Å. We averaged an additional 7 cycles starting from the final maps above. The R-factor further decreased: ORT - 0.187 to 0.159 and MON - 0.203 to 0.179. The phase difference between the final and next-to-last cycle was 7.7° (ORT) and 6.3° (MON) (fig. 5). The rms phase change from the starting m.i.r. or s.i.r. best phase was 85.6° (ORT) and 79.0° (MON). An example of a side chain resolved in the averaged map from this experiment but not in any of the others was Glu89 shown in fig. 6.

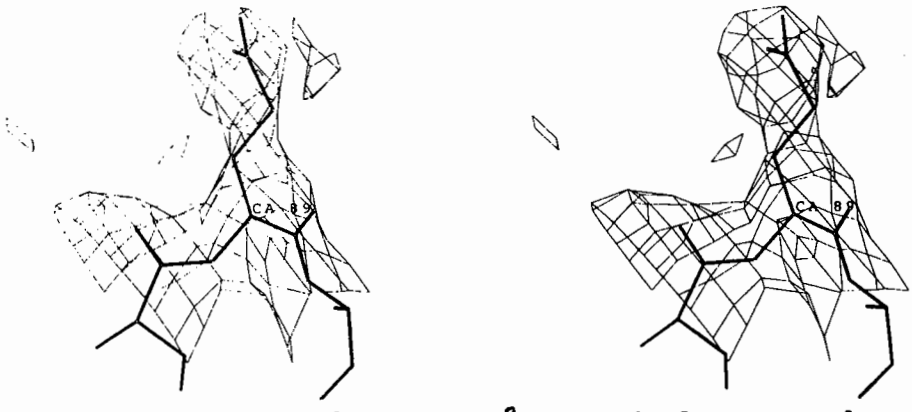


Fig. 6 - Glu89 in final averaged map. 3.5Å iteratively averaged map (contour level 0.9 sigma) using MON m.i.r. and ORT s.i.r. starting maps. The starting maps alone were uninterpretable in this region.

7. SUMMARY

The final experiment demonstrates that given two very poor (and uninterpretable) 3.5Å s.i.r. or m.i.r. maps and knowledge of the molecular boundary, we generated, by iterative density averaging, an improved map. From this we successfully traced the backbone and correctly positioned most of the amino acid side chains.

8. ACKNOWLEDGEMENTS

We thank Tom Garrett, Doug Freymann, and Mike Blum for helpful advice.

9. REFERENCES

1. P.J. Bjorkman, M.A. Saper, B. Samraoui W.S. Bennett, J.L. Strominger, and D.C. Wiley, Nature, 329, (1987) 506.
2. P.J. Bjorkman, M.A. Saper, B. Samraoui W.S. Bennett, J.L. Strominger, and D.C. Wiley, Nature, 329, (1987) 512.
3. P.J. Bjorkman, J.L. Strominger, and D.C. Wiley, J. Mol. Biol. 186, (1985) 205.
4. B.C. Wang, Methods Enzymol. 115, (1985) 90.
5. A.D. Podjarny, T.N. Bhat, and M. Zwick, Ann. Rev. Biophys. Biophys. Chem. 16, (1987) 351.
6. F.K. Winkler, C.E. Schutt, S.C. Harrison, and G. Bricogne, Nature, 265, (1977) 509.
7. J.N. Champness, A.C. Bloomer, G. Bricogne, P.J.G. Butler, and A. Klug, Nature, 259, (1976) 20.
8. G. Bricogne, Acta. Cryst. A32, (1976) 832.

9. H. Löebermann, R. Tokuoka, J. Deisenhofer, and R. Huber, J. Mol. Biol. 177, (1984) 531.
10. W. Steigemann, Ph.D. dissertaion, Technische Universität München, (1974).
11. S. Remington, G. Wiegand, and R. Huber, J. Mol. Biol. 158, (1982) 111.
12. J.N. Varghese, W.G. Laver, and P. M. Colman, Nature, 303, (1983) 35.
13. T.A. Jones in Computational Crystallography (ed. D. Sayre) (Oxford: Oxford University, 1982) 303.
14. J.L. Sussman in Methods and Applications in Crystallographic Computing (eds. S.R. Hall and T. Ashida) (Oxford: Clarnedon, 1984) 206.

DETERMINATION OF VIRUS STRUCTURES BY THE USE OF MOLECULAR REPLACEMENT DENSITY AVERAGING

by

MICHAEL G. ROSSMANN
Department of Biological Sciences, Purdue University,
West Lafayette, Indiana 47907, USA

Accurate phases have been determined for a number of icosahedral virus crystal structures by extending phase information from poor low resolution (≃8 Å) estimates to high resolution (≃3 Å). Strategies are discussed for successful phase extension.

1. INTRODUCTION

The following demonstration of the reciprocal-space equivalence to real-space averaging and iterative cycling used in phase improvement and extension is based on papers by Main and Rossmann [1] and Arnold and Rossmann [2].

Let there be N noncrystallographically related identical structures in the crystallographic asymmetric unit. Then their averaged density is given by

$$\rho_{avg}(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \rho(\mathbf{x}_n)$$

where

$$\mathbf{x}_n = [C_n]\mathbf{x}_1 + \mathbf{d}_n \quad ,$$

and $[C_n]$ and $\mathbf{d}_n$ are the nth noncrystallographic rotation and translation elements.

Since the electron density $\rho(\mathbf{x})$ can be expressed as a Fourier series with the structure factors F_h as coefficients, it follows, by substitution for $\rho(\mathbf{x}_1)$, that

$$\rho_{avg}(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{V} \sum_{h} F_h \exp(-2\pi i \mathbf{h}\cdot\mathbf{x}_n) \quad .$$

By back-transformation of the averaged density we have

$$F_p = \int_U \rho_{avg}(\mathbf{x}) \exp(2\pi i \mathbf{p}\cdot\mathbf{x}) d\mathbf{x}$$

where U is the volume of that part of the cell that has been averaged and where the rest of the cell density has been set to zero. Now by substitution of ρ_{avg} it follows that

$$\mathbf{F}_p = \frac{1}{NV} \sum_h \mathbf{F}_h \sum_{n=1}^{N} \exp(-2\pi i \mathbf{h} \cdot \mathbf{d}_n) \int_U \exp\{2\pi i(-\mathbf{h}[C_n] + p) \cdot \mathbf{x}\} d\mathbf{x}$$

If we now define

$$G_{hpn} = \frac{1}{U} \int_U \exp\{2\pi i(p - \mathbf{h}[C_n]) \cdot \mathbf{x}\} d\mathbf{x} \qquad (1)$$

and

$$T_{hpn} = \exp(-2\pi i \mathbf{h} \cdot \mathbf{d}_n) \quad ,$$

then

$$\mathbf{F}_p = \frac{U}{NV} \sum_h \mathbf{F}_h \sum_{n=1}^{N} G_{hpn} \cdot T_{hn} \quad ,$$

On further simplification by putting

$$\mathbf{a}_{hp} = \sum_{n=1}^{N} G_{hpn} \cdot T_{hn} \quad ,$$

then

$$\mathbf{F}_p = \frac{U}{NV} \sum_h \mathbf{F}_h \, \mathbf{a}_{hp} \quad . \qquad (2)$$

These are the molecular replacement equations first derived by Main and Rossmann [1] and again, in a similar form, by Crowther [3]. The coefficients $\mathbf{a}_{hp}$ depend only on a knowledge of the noncrystallographic operators $[C_n]$ and $\mathbf{d}_n$ and the shape of the volume U within which the local noncrystallographic operators are true. The equations show that real-space averaging and cycling [4-7] is equivalent to substitution of the old phases and observed amplitudes on the right-hand side and obtaining a new and improved set of calculated structure factors on the left-hand side, ready for the next cycle.

2. IMPLICATIONS OF THE MOLECULAR REPLACEMENT EQUATIONS TO PHASE DETERMINATION

2.1 Phase extension in small steps

The G function (eqn. 1) is perhaps better known in its application to the rotation function [8]. It has values close to unity whenever

$$p - h[C_n] \simeq 0$$

while for other combinations of p and h, G is very small and such terms can, therefore, be neglected. Hence, the important terms in the molecular replacement equations are those for which F_h has about the same resolution as F_p. Thus, the determination of F_p at the edge of the resolution for which phases are known will be inaccurate, since about half the terms will have to be omitted from each equation where there is no knowledge of the phase. In general

$$F_p = P + Q \qquad (3)$$

where

$$P = \sum_h F_h \, a_{hp}$$

for those terms where there is a current phase estimate and

$$Q = \sum_h F_h \, a_{hp}$$

for those terms where there is no phase estimate. Computation of F_p for a term outside the current resolution limit will decrease the number of terms that can be used in calculating P_h and increase that component Q_h which must be neglected. For that reason, phase extension at a given time must be limited. In general it is useful to limit phase extension for those terms in the first large positive loop of the G function. Hence, $HR < 0.7$ where H is the reciprocal distance between the point p (at which F_p is being determined) and any interaction with a known phase at h. That is, $H = |p - h|$, and R is the radius of the volume U. If extension is to go over n reciprocal lattice units of length $\frac{1}{a}$, then

$$\frac{n}{a} R < 0.7$$

or

$$n < \frac{0.7a}{R} .$$

For example, $a \simeq 450$ Å and R = 150 Å for human rhinovirus 14 (HRV14) cubic crystals [9,10]. Hence, $n < 2.1$ or phase extension should proceed over less than about $2a^*$ units at a time.

The consequences of too rapid phase extension can be catastrophic as reasonable phase solutions to the molecular replacement equations at differing resolutions may be inconsistent. This gives rise to "bumps" in a plot of correlation coefficients versus resolution [10]. Different solutions of the molecular replacement equations could occur due to:

(i) Choice of different origin - although the envelope definition relative to the crystallographic symmetry axis in general makes this impossible.

(ii) Choice of different hand to the solution, unless the distribution of molecular envelopes is itself lacking a center of symmetry, in which case the coefficients a_{hp} (eqn. 3) are themselves noncentric. Similarly the orientation of noncrystallographic symmetry elements may be noncentric.

(iii) Opposite "Babinet" selection - that is, positive density is negative and vice versa. The basic differentiation of these two solutions is derived from the positive (as opposed to negative) value of F(000) and the low resolution terms that interact with F(000).

2.2 Inclusion of calculated terms where there are no observations

If a reflection is unobserved, it should be included in the computation of the next electron density map as a weighted F_c term instead of being omitted. From eqn. (3) it is easy to see that

$$F_p = P_{obs} + Q_{calc}$$

where P is taken over the observed terms and Q over the unobserved terms on substituting the calculated phase angles. Clearly the error is less by using Q_{calc} as an estimate of the unobserved terms rather than by omitting it entirely. As the number of unobserved terms increases

$$F_p \rightarrow Q_{calc} \quad ,$$

implying that, in the limiting case, $F_p = Q_{calc}$ which means no change in F_p at all; that is, no phase improvement takes place.

The weights which we have found useful in computing a new map are $\omega = \omega_S \omega_R$, where ω_S are Sim weights [11,12] and ω_R are Rayment weights [13]. For those terms where only a calculated value is avaialable, it is possible to use either the mean weight for all observed terms in a given resolution range or a weight equal to the mean correlation coefficient in that resolution range.

2.3 The power of phase determination

Arnold and Rossmann [2] have shown that the power, P, of phase extension and phase improvement can be expressed as

$$P \propto \frac{\sqrt{Nf}}{R(\frac{U}{V})}$$

where N is the noncrystallographic redundancy, f is the fraction of observed terms in the resolution shell, (U/V) is the fraction of volume within which noncrystallographic symmetry is true, and R is an estimate of accuracy of the structure amplitude, such as R_{merge}. Hence, by putting N = 1, the power is based only on solvent flattening in the volume (V - U).

This formula assumes that there is no error in:
a) the rotational parameters that determine the particle orientation,
b) the translational parameters that determine the particle position,
c) the definition of the boundary of the molecular envelope.
These quantities can normally be defined with sufficient accuracy prior to phase extension and refinement to produce high quality phases. Viral envelopes chosen in the structure determinations at Purdue University are generally defined by spheres limiting the external and internal radii of the protein coat and tangent planes where an overgenerous definition of the spherical volumes causes overlap between particles. While we have not had to redefine envelopes during phase extension, this has been done, for instance, in the determination of the structure of the influenza virus hemagglutinin spike [14] and the histocompatability antigen II [15].

2.4 Progress in refinement

The F_c values are scaled in resolution shells to the observed values. Necessarily the external values are smaller in amplitude (because of omission of those terms outside the current resolution limit) and the scale factor has to be appropriately adjusted. The correlation coefficient (introduced by Bricogne and Harrison in the work on tomato bushy stunt virus [16]) is, thus, a good measure of progress and is defined as

$$C = \frac{\sum(|F_o| - \langle F_o \rangle)(|F_c| - \langle |F_c| \rangle)}{\{\sum(F_o - \langle F_o \rangle)^2 \sum(F_c - \langle F_c \rangle)^2\}^{\frac{1}{2}}}$$

where $\langle F_o \rangle$ and $\langle F_c \rangle$ are mean F_o and F_c in a given resolution range. The *R*-factor

$$\frac{\sum|(|F_o| - k|F_c|)|}{\sum|F_o|}$$

can also be used but is more intimately associated with a knowledge of the scale factor k and is less sensitive. It is also useful to observe the mean phase shift from cycle to cycle. The correlation coefficient necessarily decreases rapidly at the edge of resolution where essentially half the terms are missing from the molecular replacement equations. We frequently attempt to refine the phases at a given resolution until $C > 0.5$ in the outermost resolution shell. An example of these coefficients after the final cycle of

TABLE 1. Final molecular replacement results at 3.08 Å resolution.

Mean Resolution Range	*R*-factor	Correlation Coefficient	Number of Reflections	Relative Local Scale Factor
15.18	17.8	0.89	6,831	0.030
10.10	14.9	0.92	10,354	0.029
8.09	13.9	0.93	12,548	0.029
6.94	13.8	0.93	14,330	0.029
6.17	13.8	0.94	16,002	0.029
5.62	14.3	0.93	17,775	0.029
5.19	14.3	0.93	19,290	0.029
4.84	14.6	0.93	20,786	0.029
4.56	15.0	0.92	22,102	0.030
4.32	15.8	0.91	22,962	0.030
4.11	16.8	0.90	23,892	0.031
3.94	17.8	0.88	24,237	0.032
3.78	18.7	0.87	25,033	0.033
3.64	20.0	0.85	25,382	0.034
3.51	21.5	0.82	25,457	0.036
3.40	23.4	0.78	25,582	0.038
3.30	25.5	0.73	25,527	0.040
3.20	27.3	0.69	24,430	0.042
3.12	30.3	0.59	14,274	0.048
3.04	40.0	0.35	11,543	0.086

phase extension for HRV14 is given in Table 1.

The quality of the final map is usually quite exceptional. Presumably this is because the noncrystallographic symmetry is almost perfect, whereas the assumptions involved in phase determination by multiple isomorphous replacement are mostly crude approximations (shape of heavy atom substitutions and general defects described as "lack of isomorphism"). A model built to the 3 Å resolution map of HRV14 gave an *R*-factor of 30% without any refinement because it was possible to interpret the electron density so easily (Table 2).

2.5 Computations

The map which was calculated for the purpose of averaging was not calculated on a fine grid (one-fifth of the resolution) in order to permit linear interpolation [6]. Rather it was calculated on a grid of between one-half and one-third of the resolution. Quadratic interpolation [17] was used to determine electron density at non-integral lattice points. This greatly reduced the number of electron densities that had to be stored. In turn, that meant we could store the whole map (or at least one-third of the map for Mengo virus [18]) for fast random access. Thus, we could avoid time consuming sorting [7,19]. This procedure was suggested to us by Jim Hogle and David Filman [20] and was applied in the latter stages of the Mengo virus determination where storage and sorting were a real problem on account of the 60-fold redundancy.

3. ACKNOWLEDGMENTS

I have been fortunate in having many very talented postdoctoral fellows and

TABLE 2. Randomly selected sample of reflections comprising one-fifteenth of observed amplitudes between 10.0 and 3.0 Å resolution.

Resolution Range (Å)	*R*	Number of Reflections
10.0 - 8.44	37.4	537
8.44 - 7.43	38.5	660
7.43 - 6.72	39.7	682
6.72 - 6.18	36.5	775
6.18 - 5.75	35.7	819
5.75 - 5.40	33.1	842
5.40 - 5.11	29.3	948
5.11 - 4.86	26.6	938
4.86 - 4.64	25.3	1,073
4.64 - 4.45	25.7	1,121
4.45 - 4.28	26.7	1,121
4.28 - 4.13	26.3	1,083
4.13 - 4.00	27.5	1,237
4.00 - 3.87	27.4	1,161
3.87 - 3.76	28.7	1,187
3.76 - 3.66	29.8	1,280
3.66 - 3.56	30.1	1,314
3.56 - 3.47	31.2	1,307
3.47 - 3.39	31.7	1,264
3.39 - 3.32	33.0	1,200
3.32 - 3.25	35.2	1,268
3.25 - 3.18	34.3	1,268
3.18 - 3.12	36.2	1,221
3.12 - 3.06	38.4	1,217
3.06 - 3.00	37.9	878
Overall	31.6	26,401

graduate students who have participated in developing the techniques described here. In particular, I would like to mention Eddy Arnold, Greg Kamer, Ming Luo and Gert Vriend. The work was supported by grants from the National Science Foundation and the National Institutes of Health.

REFERENCES

1. P. Main and M. G. Rossmann, Acta Crystallogr. 21, (1966) 67-72.
2. E. Arnold and M. G. Rossmann, Proc. Natl. Acad. Sci. U.S. 83, (1986) 5489-5493.
3. R. A. Crowther, Acta Crystallogr. B25, (1969) 2571-2580.
4. M. Buehner, G. C. Ford, D. Moras, K. W. Olsen and M. G. Rossmann, J. Mol. Biol. 82, (1974) 563-585.
5. P. Argos, G. C. Ford and M. G. Rossmann, Acta Crystallogr. A31 (1975) 499-506.
6. G. Bricogne, Acta Crystallogr. A30, (1974) 395-405.
7. G. Bricogne, Acta Crystallogr. A32, (1976) 832-847.
8. M. G. Rossmann and D. M. Blow, Acta Crystallogr. 15, (1962) 24-31.
9. M. G. Rossmann, E. Arnold, J. W. Erickson, E. A. Frankenberger, J. P. Griffith, H. J. Hecht, J. E. Johnson, G. Kamer, M. Luo, A. G. Mosser, R. R. Rueckert, B. Sherry and G. Vriend, Nature (London) 317, (1985) 145-153.
10. E. Arnold, G. Vriend, M. Luo, J. P. Griffith, G. Kamer, J. W. Erickson, J. E. Johnson and M. G. Rossmann, Acta Crystallogr. A43, (1987) 346-

361.
11. G. A. Sim, Acta Crystallogr. 12, (1959) 813-815.
12. G. A. Sim, Acta Crystallogr. 13, (1960) 511-512.
13. I. Rayment, Acta Crystallogr. A39, (1983) 102-116.
14. I. A. Wilson, J. J. Skehel and D. C. Wiley, Nature (London) 289, (1981) 366-73.
15. P. J. Bjorkman, M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger and D. C. Wiley, Nature (London) 329, (1987) 506-512.
16. S. C. Harrison, A. J. Olson, C. E. Schutt, F. K. Winkler and G. Bricogne, Nature (London) 276, (1978) 368-373.
17. C. E. Nordman, Acta Crystallogr. A36, (1980) 747-754.
18. M. Luo, G. Vriend, G. Kamer, I. Minor, E. Arnold, M. G. Rossmann, U. Boege, D. G. Scraba, G. M. Duke and A. C. Palmenberg, Science 235, (1987) 182-191.
19. J. E. Johnson, Acta Crystallogr. B34, (1978) 576-577.
20. J. M. Hogle, M. Chow and D. J. Filman, Science 229, (1985) 1358-1365.

HISTOGRAM MATCHING AS A DENSITY MODIFICATION TECHNIQUE FOR PHASE REFINEMENT AND EXTENSION OF PROTEIN MOLECULES

by

Kam Y J Zhang and Peter Main
Department of Physics, University of York,
Heslington, York YO1 5DD, England.

1. INTRODUCTION

Phase refinement and extension are clearly important in the determination of macromolecular structures. The initial phases, most commonly available from multiple isomorphous replacement (MIR), may not be good enough to give an interpretable map or may only be available at low resolution, even if high resolution data are available for the native crystal. Under these circumstances, we would like to be able to refine and extend the MIR phases to produce an interpretable map at the full resolution of the native data. This may be carried out either in real space (density modification) or reciprocal space (direct method) or a combination of the two. For a recent review of density modification methods, see ref. [6].

We describe here a density modification technique, known as histogram matching, which is already used in image processing. It has been applied to the known structure of R3 2Zn insulin [2] where it achieves both phase refinement and extension at modest computing cost. We compare the results with those obtained from the following methods:

(i) Agarwal and Isaacs [1] described a dummy atom structure refinement method. This consists of placing dummy atoms to satisfy the density in an approximate map and then refining the positional and thermal parameters by least squares to minimise the discrepancy between observed and calculated magnitudes.

(ii) Direct phase extension and refinement have been achieved by Sayre [8] using Sayre's equation. This was applied successfully to the structure of rubredoxin [9]. Later, the same method was applied to insulin and reported by Agarwal and Isaacs [1].

(iii) Another reciprocal space method is the maximum determinant rule of Tsoucaris [10]. The application of this to the structure of insulin is reported in ref. [7].

(iv) The method of solvent flattening was described by Wang [11] as a means of solving the phase ambiguity of single isomorphous replacement or single anomalous scattering data. It is now widely used in the refinement of MIR phases and we report here its application to 2Zn insulin.

As was mentioned above, histogram matching is a standard technique in image processing. See, for example, ref. [3], chapter 6. It may be regarded as a generalisation of solvent flattening and is also related to the "phase correction" technique of Hoppe and Gassmann [5]. For any discrete image, a histogram of density values can be obtained. In favourable cases, this can be compared with the histogram expected of a good image and used as a measure of the quality of the image. Furthermore, the test image may be improved by adjusting its density in a systematic way to make its histogram match the correct histogram. To apply this to X-ray crytallography, we need to be able to predict the density histogram of the electron density map of an unknown structure and show that this differs from the histogram of a MIR map. We then modify the density values of the approximate map to match the correct histogram and calculate new phases from the modified map. This will form one cycle of an iterative process of map improvement.

2. METHOD

2.1 Histogram matching.

The density histogram of an electron density map is the probability distribution of electron density values at the grid points at which the map is evaluated. It provides a global description of the map in which all spatial information is lost. The process of histogram matching transforms the electron density distribution into that expected of a good map. This may be done as follows:

a) Compute the histogram of the map to be modified and obtain the expected histogram at the same resolution. The latter may be taken from a similar structure or calculated from a formula (see later).

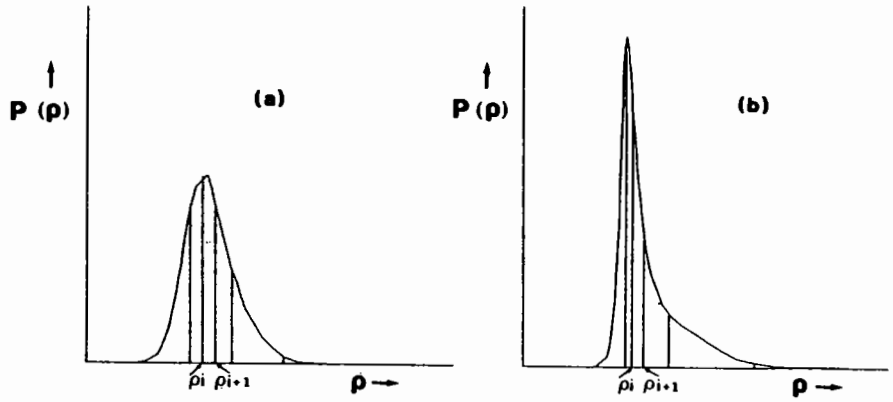


Fig. 1 (a) Electron density histogram from an approximate map. The area is divided in equal smaller areas with boundaries at ρ_i, i=1,n.
(b) Expected histogram divided into equal areas as in (a) with boundaries at ρ_i', i=1,n. This gives corresponding density values ρ_i and ρ_i', i=1,n in the two maps.

b) Divide the two histograms into equal areas as in Fig. 1. This gives corresponding density values ρ_i and ρ_i', i=1,n in the two histograms. We have used a value of about 250 for n and this is quite satisfactory.

c) From these corresponding values, calculate scale factors a_i and shifts b_i which map ρ onto ρ' within the interval ρ_i to ρ_i' as

$$\rho_i' = a_i\rho_i + b_i \qquad (1)$$

i.e. $$a_i = \frac{\rho'_{i+1} - \rho'_i}{\rho_{i+1} - \rho_i} \quad \text{and} \quad b_i = \frac{\rho_{i+1}\rho'_i - \rho'_{i+1}\rho_i}{\rho_{i+1} - \rho_i}$$

Note that $b_i>0$ shifts the histogram to the right, while $b_i<0$ shifts it left. Also, $a_i<1$ narrows the histogram while increasing its height to keep the area constant.

d) Alter the density according to equation (1), using the appropriate values of a_i and b_i for each range of ρ. This results in a new map which has the same electron density distribution as the expected one. This operation incidentally applies a maximum and a minimum to the electron density, imposes the correct mean and variance and defines the entropy of the new map.

2.2. Phase refinement and extension.

The iterative procedure of map improvement which we have used is a combination of histogram matching with the solvent flattening technique. Starting from an approximate map calculated from MIR phases:

a) Determine the molecular envelope.
b) Set the density within the solvent region to a constant.
c) Obtain the expected histogram at the desired resolution. For phase extension, this will be at a higher resolution than the present map.
d) Modify the density within the molecular envelope to match the expected histogram.
e) Calculate structure factors from the modified map and calculated their Sim weights.
f) Combine the new phases with the original MIR phases, taking their weights into account. Extended phases and weights are accepted at their calculated values.
g) Calculate a new map and repeat from a) till the process has converged.

3. EXPERIMENTAL RESULTS

3.1 Density histograms of actual protein maps.

The density histograms of a small number of protein structures were examined and they were all found to behave in the same way. They were independent of the grid size on which the map was calculated, provided this was fine enough to give a good representation of the underlying continuous density. They were also independent of the structure itself. This means that a histogram for a known structure could be used to predict the histogram for an unknown structure. However, the histogram depended upon resolution and also the overall temperature factor.

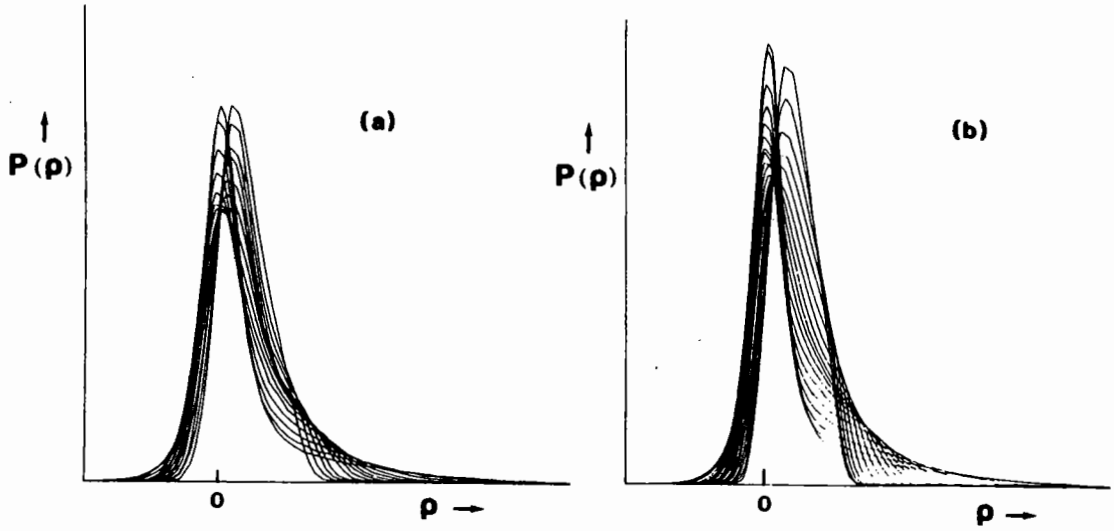


Fig. 2 (a) Electron density histogram of 2Zn pig insulin at resolutions ranging from 1.5Å to 4.1Å obtained from maps given by refined atomic coordinates.
(b) As in (a), but the histogram are for haemoglobin [4].
The high-resolution maps give rise to the high, narrow peak near $\rho(x)=0$. The corresponding peak for low resolution maps is much lower and broader.

Fig. 2 shows the histograms from two different proteins at a range of resolutions. The densities used are from within the molecular envelope only. The volume of the envelope was standardised to an average of 10Å^3 for the protein atoms (including hydrogen) to ensure the same ratio of atomic volume to background in all cases. As the resolution decreases, Fig. 2 shows that the peak of the histogram lowers to a minimum at about 3.0Å and then rises again. As the peak rises, its maximum moves towards higher density and the peak becomes broader. This large peak comes from the low density in the space between atoms. With decreasing resolution, the space becomes smaller and

eventually disappears. Low density then becomes even more scarce with increasing atomic overlap, pushing the mode of the distribution to higher values. The atoms in the map give rise to the long, low tail of the histogram stretching out to high density values. This tail contracts as the resolution decreases and the atoms become less sharp.

There is nothing in this explanation of histogram behaviour that depends upon the details of the structure. The shape of the histogram depends only upon the fact that the density consists of atoms at certain characteristic distances apart with a particular local stereochemistry. This will be true for all polypeptide structures. Because of this, the histograms used in the present work were all taken from maps calculated from the refined atomic coordinates of 2Zn pig insulin. These will differ very little from histograms taken from other similar source.

In order to make use of histogram information, approximate electron density maps should have histograms which differ from those expected. This is seen in Fig. 3 where the histogram of the 1.9Å maps of 2Zn pig insulin calculated from refined atomic coordinates and from the isomorphous phases are compared. There is a considerable difference between them, indicating the possibility of map improvement by the process of histogram matching.

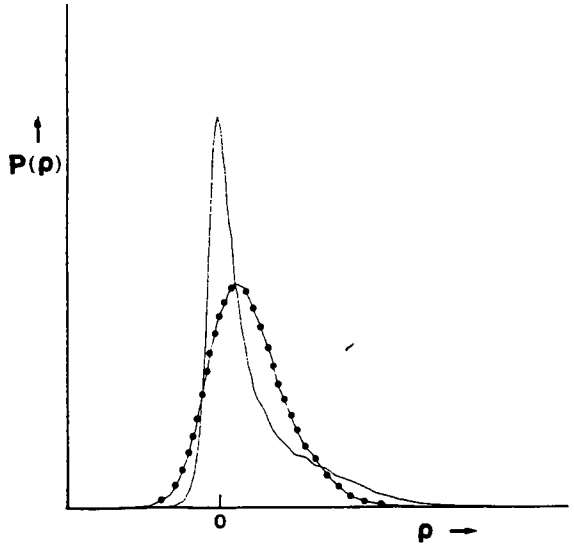


Fig. 3 Density histograms of 2Zn pig insulin at 1.9Å resolution.
-.-.- map from MIR phases.
——— map from refined atomic coordinates.

3.2 Phase refinement and extension.

The procedure outlined in section 2.2 was tested on the 1.9Å MIR map of 2Zn pig insulin. It is already of good quality, for the structure was obtained from a earlier MIR map at 2.8Å resolution. However, the same data have been used in tests of other methods, as mentioned in the introduction, allowing us to compare results. The space group is R3 with 2 Zn atoms in the cell and 806 non-hydrogen atoms belonging to the protein in the asymmetric unit.

The 1.9Å MIR phases were refined using the procedure outlined in section 2.2 and convergence was reached after five iterations. The phases were then extended from 1.9Å to 1.5Å in four stages, increasing the resolution by 0.1Å at each stage. Five iterations were required for convergence each time. The results are shown in table 1. The two measures of quality used are the mean phase error over all reflexions within the indicated resolution range and the correlation coefficient between the map and that obtained from the known atomic coordinates at the same resolution.

Table 1. Result of map improvement by histogram matching for 2Zn insulin.

resolution (Å)	no. of reflexions.	mean phase error (°) 1.9Å phases	extended phases	all phases	correlation coefficient
1.9	6537	59.9	-	59.9	0.668
1.9	6537	47.0	-	47.0	0.778
1.5	13287	42.3	54.8	48.4	0.803

The correlation coefficient is calculated from

$$\frac{\overline{\rho(x)*\rho'(x)} - \overline{\rho(x)}*\overline{\rho'(x)}}{\text{SQRT}(\overline{\rho(x)**2} - \overline{\rho(x)}**2)*\text{SQRT}(\overline{\rho(x)**2} - \overline{\rho(x)}**2)}$$

where $\rho(x)$ and $\rho'(x)$ represent the density values in the different maps and $\overline{\rho(x)}$ is the mean value calculated over the whole map. Weights are applied in the calculation of the maps, but no weights are used in the calculation of mean phase error.

The correlation coefficient indicates a clear improvement in the quality of the map. In terms of the phases, the MIR phases refine from an initial mean error of 60° to 47°, then continue to improve to 42° during the phase extension. The newly extended phases are determined to a better accuracy than the original MIR phases.

Further tests were carried out to see if additional improvement could be obtained by using sharpened F's, instead of the F_{obs} used so far. The most accurate phases were obtained using F's from which the overall temperature factor had been removed, i.e. F's corresponding to stationary atoms. The results of this phase refinement and extension are set out in Table 2. Compared with the previous results, they show a mean improvement of 2° for the original phases and 3° for the extended phases.

Table 2. Result of map improvement for 2Zn insulin using sharpened F's.

		mean phase error (°)			
resolution (Å)	no. of reflexions	1.9Å phases	extended phases	all phases	correlation coefficient
1.9	6537	59.9	-	59.9	0.589
1.9	6537	46.3	-	46.3	0.728
1.8	7657	44.2	57.0	46.1	0.746
1.7	9130	42.6	55.4	46.2	0.754
1.6	10946	41.3	52.4	44.9	0.757
1.5	13287	40.1	51.8	45.9	0.756

It can be seen from Table 2 that the 1.9Å phases continue to improve as more structure factors are included in the map. This indicates the possibility of further improvement by repeating the phase extension starting from the more highly refined 1.9Å phases. Table 3 shows the results of this. Additional improvement is obtained, though it is not worth the doubling of computer time it entails.

Table 3. Repeat of phase extension starting from refined 1.9Å phases.

		mean phase error (°)			
resolution (Å)	no. of reflexions	1.9Å phases	extended phases	all phases	correlation coefficient
1.9	6537	40.1	-	40.1	0.771
1.5	13287	38.8	49.9	44.4	0.773

4. DISCUSSION

It was mentioned previously that the density histogram depends upon the overall temperature factor. However, the best phases are obtained after removing the effects of temperature from the F's. Thus, if magnitudes sharpened in this way are always used, it becomes unnecessary to change the

histogram according to the temperature factor of different structures. This will simplify the application of the method.

It is satisfying to note that the 1.9Å phases continue to improve as more magnitudes are added to the map. It is also satisfying to find that the extended phases are more accurate than the original MIR phases. A comparison of phase errors with those obtained by the methods mentioned in section 1 is shown in Table 4. The solvent flattening was carried out by the present authors using a volume of 30% of the cell for the solvent. All other results were obtained by the authors referenced. Table 5 gives a closer comparison with the dummy-atom refinement method and Table 6 details the comparison with the maximum determinant results.

Table 4. Comparison of mean phase errors (in degrees) for 2Zn insulin obtained by different methods.

reso-lution	MIR	*dummy atom refinement	Sayre's equation	maximum determinant	solvent flattening	histogram matching
1.9	60	65	52	49	45	39
1.5	-	70	55+	52	52	44

* the dummy atom refinement was started from 3.0Å MIR phases.
\+ only 10000 phases were included in this mean error.

Table 5. Mean phase error in degrees of the strongest reflexions. The MIR phases are at 1.9Å resolution, the remainder are all at 1.5Å.

no. of strongest reflexions	MIR phases	dummy atom refinement	solvent flattening	histogram matching
250	31	27	18	16
500	33	32	20	17
1000	37	39	23	20
2000	46	47	30	24

Table 6. Mean phase error in degrees as a function of E-value.

resolution (Å)	number of reflexions	|E|	MIR phases	maximum determinant	solvent flattening	histogram matching
1.9	604	>1.5	57	27	29	20
1.5	1147	>1.5		28	31	22
1.9	2408	>1.0	54	33	32	25
1.5	5020	>1.0		38	40	31
1.9	6522	>0.1	60	49	45	39
1.5	13281	>0.1		52	52	44

The histogram matching method (which incorporates solvent flattening) produces the best results. In addition, it requires much less computing time than all the other methods considered except solvent flattening. This is because most of the computation is the calculation of maps and structure factors.

We have demonstrated that the density histogram of an electron density map contains information which can be exploited in a process of map improvement at high resolution. When combined with solvent flattening, we have a method which restricts electron density values over the whole cell instead of just the solvent region or the molecule. Tests of the method at lower resolution were not as satisfactory as those already described. The initial isomorphous phases refined very well but, upon phase extension, the error in the new phases rapidly became too large. Work is now in progress to improve this. As was pointed out previously, the density histogram discards all positional

information. Although the histogram is unique for any particular map, vastly different maps can have identical histograms. This makes histogram matching inherently less powerful than solvent flattening since, in the latter method, positional information is always available if the molecular envelope is known.

Histogram matching, as applied here, suffers from the same defects as solvent flattening in that molecular density outside the envelope is strongly suppressed. To combat this, we are experimenting with a new technique of determining the envelope. Also, false density inside the envelope tends to remain. However, in our tests on insulin it was observed that much false density was suppressed and new, correct density appeared, resulting in a genuine improvement of the map. This may be judged from Fig. 4 which shows the same section from each of three maps - the original 1.9Å isomorphous map, the 1.5Å map obtained from histogram matching and the 1.5Å map given by refined atomic coordinates. The molecular boundary is superimposed on the latter map.

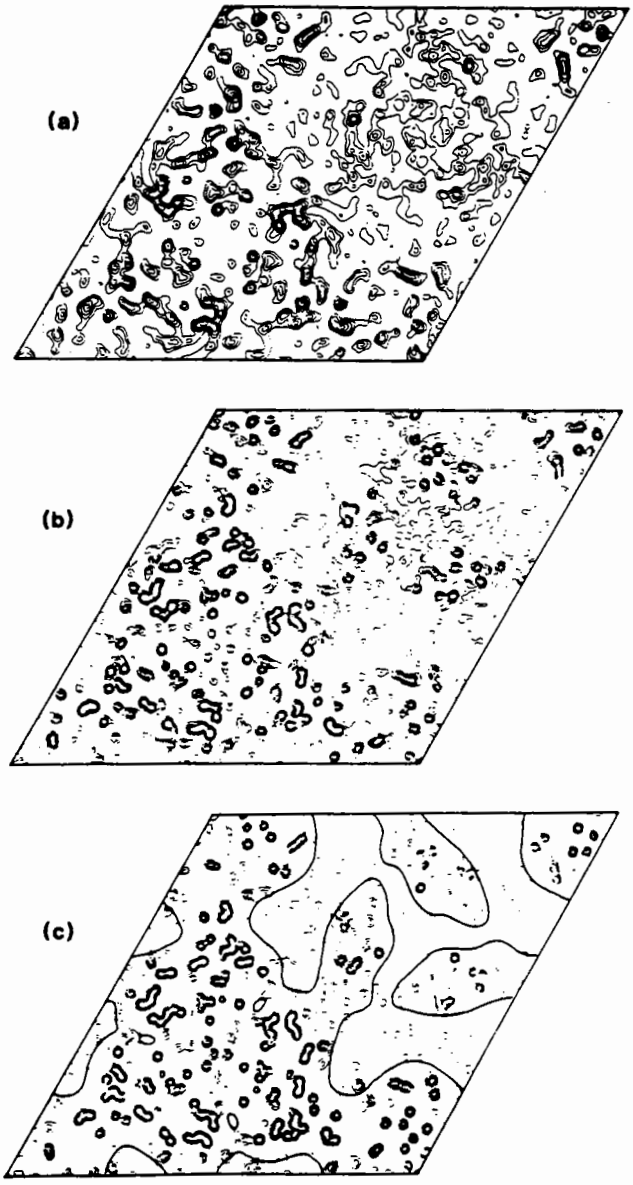


Fig. 4 (a) Section of insulin map calculated from 1.9Å MIR phases.
(b) Same section calculated at 1.5Å with phases from histogram matching.
(c) Same section obtained at 1.5Å resolution from atomic coordinates.

5. ACKNOWLEDGEMENTS

We wish to thank Prof. G. Dodson for kindly supplying the 2Zn insulin data and atomic coordinates. We are also indebted to Mrs. E. Dodson for the use of computer programs and helpful discussions. One of us (KYJZ) is very grateful to the Dodsons for the use of their laboratory facilities.

6. REFERENCES

1) Agarwal, R. C. and Isaacs, N. W. (1977) *Proc. Natl. Acad. Sci. USA*, 74, 2835-2839.
2) Baker, E. N., Blundell, T. L., Cutfield, J. F., Cutfield, S. M., Dodson, E. J., Dodson, G. G., Hodgkin, D. C., Hubbard, R. E., Isaacs, N. W., Reynolds, C. D., Sakabe, N. and Vijayan, M. (1985) *Phil. Trans. Roy. Soc.*
3) Castleman, K. R. (1979) "*Digital Image Processing*", Prentice-Hall, New Jersey.
4) Derewenda, Z. S., Dodson, E. J., Dodson, G. G., and Bizozowski, A. M. (1981) *Acta Cryst.*, A37, 407-413.
5) Hoppe, W. and Gassmann, J. (1968) *Acta Cryst.*, B24, 97-107.
6) Podjarny, A. D., Bhat, T. N. and Zwick, M. (1987) *Ann. Rev. Biophys. Chem.*, 16, 351-373.
7) de Rango, C., Maugen, Y., Tsoucaris, G., Dodson, E. J., Dodson, G. G. and Taylor, D. J. (1985) *Acta Cryst.* A41, 3-17.
8) Sayre, D. (1972) *Acta Cryst.*, A28, 210-212.
9) Sayre, D. (1974) *Acta Cryst.*, A30, 180-184.
10) Tsoucaris, G. (1970) *Acta Cryst.*, A26, 492-499.
11) Wang, B. C. (1985) *Methods in Enzymology*, 115, 90-112.

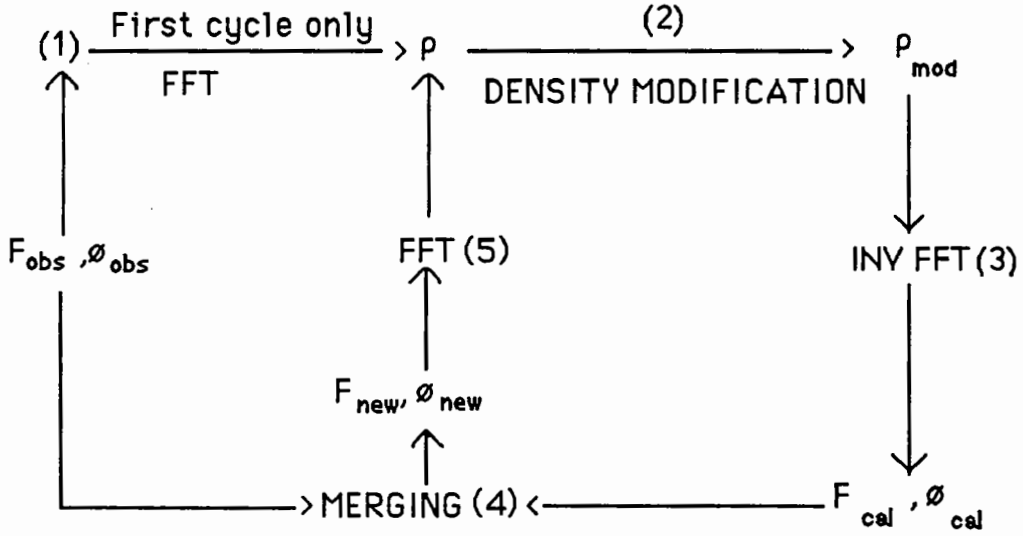
IMPROVING PROTEIN PHASES IN REAL SPACE

by

A. D. Podjarny
IBMC. 15, Rue Descartes , 67084 Strasbourg, France.

1. DEFINITION OF THE PROBLEM

Real space phasing methods are aimed at improving the agreement of a density map with a set of physically meaningful constraints.Density maps must agree simultaneously with experimental data such as the observed amplitudes and the MIR phase distribution and with physical constraints based on a-priori knowledge of the characteristics of the density function. These physical constraints include positivity; high-resolution atomicity; boundedness; uniformity of solvent regions; continuity of the bio-polymer chain; and known non-crystallographic symmetry of the density distribution. A starting map to be processed by real space methods, such as an MIR map, is usually obtained from, and thus will fully agree with, the experimental data. If the map also agrees with the set of physical constraints, there is no room for real space techniques, and map interpretation should be attempted. It is more often the case, however, that the map does not fully agree with all the physical constraints. In this case there is room to improve the agreement by real space techniques. To impose the physical constraints on an experimental map, a "Density Modification (DM)" iterative algorithm has been proposed [1,2]. It alternates real and reciprocal space operations, and the procedure is outlined in the following diagram:



The steps are:

1) The electron density map ρ is calculated by Fast Fourier transform from the experimental data F_{obs} , ϕ_{obs} .
2) A modified electron density ρ_{mod} is obtained from the current map by the application of a known physical constraint,as discussed below.
3) The map ρ_{mod} is inverted by Fast Fourier transform to calculate F_{cal}, ϕ_{cal} .
4) The information from the calculated SFs is merged with the experimental SFs to produce a new set of SFs, F_{new} ,ϕ_{new}, by one of the merging methods discussed below.
5) A density map is calculated from these merged SFs. The map is either interpreted or used as input to step (2).

The method thus consist of two substantive steps, density modification and structure factor merging, and two Fourier transform steps (forward and inverse) per cycle.The phases are judged convergent if the phase difference (before and after modification, before and after merging, between ϕ_{cal} of consecutive cycles, or between ϕ_{new} of consecutive cycles) is less than an arbitrary threshold. Such phase convergence only means that the new phase set has incorporated the DM restraints, and does not imply that the set is correct.

2. DENSITY MODIFICATION

The density map is modified by applying one or several of the following physical constraints. In general, the phasing power of a constraint increases with the number of density points it affects and the magnitude of the change it imposes on the density value. The following density modification techniques obviously do not exhaust all possible physical constraints that can be applied to the density.

2.1. Positivity.

To impose positivity, negative density regions are deleted or attenuated [3]. Proper implementation of this constraint requires a good estimate of the value of F_{000}, which can be obtained by adjusting the observed solvent level to a physically maningful value. The attenuation of negative regions may cause the appearance of excessively high positive peaks; this problem is solved by positive density truncation.

2.2 Atomicity.

This constraint can be applied when very high-resolution data are available and series termination errors are negligible.Sayre [4] observed that at atomic resolution and for equal gaussian atoms an atomic density is roughly proportional to its square,i.e. $\rho_{mod}=K\ \rho^2$. Squaring, however, sometimes produces increasingly large densities. To solve this problem, the "3-2" rule was proposed [5], in which a density function normalized to a maximum of 1 is modified first by the imposition of positivity and then by $\rho_{mod}=3\rho^2-2\ \rho^3$.

2.3 Solvent flatness.

The existence of a uniform solvent region implies strong constraints on the structure factor phases [6],as shown by the following relation between SFs [7,8]:

$$F_h = (1/V).\ \sum_k F_k \int_U e^{(2\pi i\,(h-k)r)} d^3r$$

h,k= reciprocal vectors,r = real space vector,U= Molecular volume,V= Unit cell volume.

The smaller and more detailed the molecular volume, the larger the number of SFs related by this equation. The following procedures have been used for defining the molecular boundary (roughly in chronological order) :

a) Hand digitalization of a minimap with the aid of a graphic tablet[9,10]. This procedure showed the potential of the method and encouraged further developments.
b) Definition of the molecular volume as regions of linked high density [11]. This method allowed for the first time the automatic and fast calculation of a molecular envelope.
c) Definition of the molecular volume as regions of high mean density [12]. The volume is obtained by replacing every point of the map by a weighted average over all neighbours whithin a sphere of given radius, and choosing the points above a given level. The map average is equivalent to a convolution between the map and a radially-weighted sphere. This convolution can be performed by reciprocal space multiplication [13,14,15].
d) The molecular volume is defined by the expectation that densities inside this volume have greater excursions, both positive and negative, from the mean value than densities in the solvent region [16].
e) Definition of the molecular volume by large diffracting elements at low resolution. These elements are gaussian spheres positioned by low-resolution translation searches [17].

2.4 Map continuity and use of a partial model.

An important property of any density map corresponding to a biological macromolecule is that,at medium resolution, the density displays single-chain connectivity, corresponding to the single-chain stereochemistry of proteins and nucleic acids. This feature is essential for the interpretation of a map that does not display atomic resolution, but it is quite difficult to implement without a molecular model. Bhat and Blow [11] have successfully applied this constraint in a crystal where a partial model was known and could be used to define connectivity. The method was used to include both the missing atoms in the ordered domain and the density corresponding to the disordered domain in tyrosyl tRNA synthetase, and a more readily interpretable map was obtained.

3. MERGING OF CALCULATED AND OBSERVED SF'S

After the modification step, an inverse Fourier transform generates calculated structure factors and phases (F_{cal},$\emptyset_{cal}$) from the modified map.

This information is now combined with the observed amplitudes and phases (F_{obs},ϕ_{obs}) to obtain new values (F_{new},ϕ_{new}) and synthesize a new improved map. The following merging techniques have been used:

3.1 Replacing the phases.

The first merging procedure simply used the observed amplitude and the calculated phase [1,2]. This approach ignored the experimental and calculated phase probability distributions as well as the calculated amplitudes, and produced maps biased towards the calculated phases.

3.2 Merging the amplitudes and replacing the phases.

To diminish the bias of simple phase replacement, the coefficients $F_{new}=(2F_{obs}-F_{cal})$, $\phi_{new}=\phi_{cal}$ were used [5,18,19]. Main [20] analyzed the characteristics of the resulting map in detail. Read [21] has suggested the expression $F_{new}=(2mF_{obs}-DF_{par})$, $\phi_{new}=\phi_{cal}$, where $F_{par}\exp(i\phi_{cal})$ is obtained from a partial model with errors, m is a figure of merit dependent on F_{obs} and F_{par}, and D is a function of the coordinate error. In this approach still only the centroid and figure of merit of the original phase distribution are utilized, i.e., full use is not made of the experimental phase distribution. This is especially relevant to cases in which the original phase distribution is multimodal and is thus not derived simply from the centroid and the figure of merit.

3.3 Merging the phases.

To overcome the problems of multimodal phase probability distributions and excessive bias toward the calculated phase,the experimental and calculated phase probability distributions are multiplied and the merged phase, ϕ_{mer}, is defined as the centroid of the product distribution. The phase probability distribution for the calculated phase, ϕ_{cal}, is obtained following Sim [22]. This approach has been used successfully in various density modification phase improvement and extension schemes [9,10,11,12,23,24].

3.4 Merging both the amplitudes and the phases.

Podjarny at al [17] have found that the coefficients $F_{new}= 2F_{obs}-F_{cal}$, $\phi_{new} = \phi_{mer}$ led to better results than merging only the phases. Rice [25] found a similar result while combining experimental and model SFs. Stuart and Artymiuk [26] have proposed a coefficient of the form $F_{new} = f_{mer}(F_{obs}+ Q_{mer}(F_{obs}-F_{cal}))$,$\phi_{new}=\phi_{mer}$, where Q_{mer} is a function of the figure or merit f_{mer}. Zelwer [27] used a merging technique in which the difference in structure factors introduced by the density modification is treated as a heavy atom contribution.

A new map is computed using the merged Fourier coefficients and the entire procedure is iterated to obtain convergence. Optimally, the final map agrees better with the physical constraints because of the modification step, and therefore the phase error is reduced. The merging step is needed to restore, at least partially, the experimental information; this step reduces the bias introduced by the modification.

4. RESULTS AND NEW DEVELOPMENTS.

The real space algorithm described above has become a well established method for improving a poor MIR map. A list of results until 1986 can be found in Tulinsky [28], Wang [12] and Podjarny et al[29], and some of the latest applications in which improvement of the electron density maps were reported are :

a) β-lactamase at 2.5 Å [30];
b) Yeast Enolase at 2.8 Å [31], with an overall figure of merit (FOM) improvement from 0.63 to 0.80 ;
c) Cardiotoxin at 3.0 Å [32].FOM improved from 0.45 to 0.78;
d) Aconitase at 3.0 Å [33] . Final FOM 0.76;
e) Bovine neurophysin II dipeptide amide complex at 3 Å [34]and
f) Methionyl-tRNA-synthetase at 2.5 Å [27].Final FOM: 0.89.

To improve the power of solvent flattening, it has been succesfully combined with entropy maximization by Prince et al [35] for phase extension of SIR phases from bovine prothrombin and bull testis calmodulin to 2.4 and 3.0 Å respectively, obtaining clearly interpretable maps.

However, the technique has so far not been powerful enough to solve the phase problem ab initio. To solve this problem, Bhat [36,37] has proposed the "consistent electron density" technique. This algorithm adds an extra step after density modification in order to stringently filter the noise coming from the original phase data. In this step, the electron density map is replaced by a complete "omit" map. The map is divided in elements, and the density in every element is recalculated from the original amplitudes and phases from the other elements only. This algorithm produces a map where wrong peaks are not echoed, and correct peaks are echoed with weights of 0.5. This technique has produced 4 Å phases for creatine kinase which are correct enough to phase a heavy atom derivative difference map. It also solved,using only 3.0 Å data,a small molecule (42 atoms) which had defied classical direct methods [38].

Another method of ab initio phasing has been employed for the low resolution determination of the structure of the Aspartyl tRNA-aspartyl tRNA synthetase complex from yeast [17].The method combines initial phase determination by low resolution pseudo-atom searches, phase refinement by density modification and least squares refinement of pseudo-atom models. It alternates between least squares refinement and density modification, and proceeds in zones of increasing resolution. The final result was a 15 Å resolution model of 140 spheres, 70 corresponding to the synthetase molecule and 70 corresponding to the two tRNA molecules. During phase refinement by density modification, a major problem was posed by the large noise regions associated with low resolution structure factors of high amplitude and wrong phase.

In order to correct wrong phases at very low resolution in a single algorithm, the constraint of maximum entropy has been imposed for solving the problem of phase refinement at 30 Å resolution[39,40]. The algorithm combines a maximum-entropy approach, a binary modelization of the electron density, a refinement of the proposed map against the observed amplitudes and solvent-flattening outside a molecular envelope. The

algorithm has been applied to the data of the complex of Aspartyl-tRNA and Aspartyl-tRNA synthetase [17]. These data included both neutron and X-ray diffraction amplitudes. Reference phases were calculated from the low resolution model of 140 spheres. Three different cases were tested: 1) calculated X-ray amplitudes and phases from a partial model, where one tRNA molecule was not included; 2) mixed X-ray observed and calculated amplitudes and phases from a partial model, as in (1), and 3) observed neutron amplitudes and phases from a very approximate model of five gaussian spheres. This model has been derived from ab-initio translation searches with spheres. The change of correlation with the map calculated with phases from the 140 spheres model at 30 Å resolution was used as a measure of correctness. Upon application of the algorithm, this correlation changed from 62 to 99% in case 1, from 60 to 79% in case 2 and from 72 to 89% in case 3. In all cases, the method was succesful in correcting large phase errors, deleting noise regions and producing the correct low resolution molecular image.

This algorithm was also used for phase refinement and extension from 15 to 10 Å resolution in the case of the Fab fragment from human cryoimmunoglobulin IgG1 Hil [41]. The starting phases were calculated from a NbCl derivative and were of acceptable quality only to 15 Å resolution. A SIR electron density map was uninterpretable. Upon application of this algorithm, the phase change from the original SIR phases was 83°. The final electron density maps clearly showed a molecular envelope corresponding to one Fab molecule.

5.CONCLUSIONS

It is quite clear that DM has the power to improve a MIR map over a wide range of resolutions when a suitable protein boundary can be found. Since DM techniques are not geared toward overcoming large errors introduced by almost random initial phases, their application has been limited to cases in which tentative phases have been determined, most commonly with MIR methods. However, new developments suggest the possibility of ab initio solutions to the phase problem in macromolecular crystallography. The work of Bhat [36,37,38] suggests that the replacement of the density map is by a composite omit map is powerful enough to force the procedure to converge to the correct solution even if the starting set of phases is random. The work of Podjarny et al [17] shows that large phase errors at low resolution can be corrected by a combination of density modification, low resolution models and least squares refinement of calculated versus observed amplitudes. Navaza [39], Podjarny et al [40] and Alzari et al [41] showed that these same constraints could be used at low resolution in a single algorithm developed whithin the maximum entropy framework. A similar combination of constraints is effective at higher resolution, as shown by Prince et al [35]. These developments point towards the possibility that MIR will eventually be replaced as the main phasing method for difficult cases where it is not easily applicable.

6. ACKNOWLEDGMENTS.

The author thanks the organizers of the " Improving Protein Phases " study weekend for the invitation to present this paper and for providing an outstanding opportunity for scientific discussion. He also thanks P. Alzari, T.N. Bhat, M.S. Chapman, D. M. Collins, D.R Davies, W.

A. Hendrickson, D. Moras, J. Navaza, B. Rees, M.G. Rossmann, P.B. Sigler, R.W. Schevitz, J.C. Thierry, B.C. Wang, E. Westbrook and M. Zwick for their encouragement and useful discussions.

7. REFERENCES

1. W. Hoppe and J. Gassman, Acta Cryst. B24, (1968) 97.
2. A. N. Barrett and M. Zwick, Acta Cryst. A27, (1971) 6.
3. G. Kartha, Acta Cryst. A25, (1969) S87.
4. D. Sayre, Acta Cryst. 5, (1952) 60.
5. D.M. Collins,M.D. Brice, T.F.M. La Cour and M.J. Legg, in Crystallographic Computing Techniques, edited by F. R. Ahmed, K. Huml and B. Sedlacek, (Copenhagen:Munskgaard,1976) 330.
6. G. Bricogne. Acta Cryst. A30, (1974) 395.
7. W.A. Hendrickson, in Structural Aspects of Biomolecules, edited by R. Srinivasan and V. Pattabhi, (New Delhi:Macmillan, 1981) 31.
8. E. Arnold and M.G. Rossmann PNAS 83, (1986) 5489.
9. W.A. Hendrickson,G.L. Klippenstein and K.B. Ward,PNAS 72,(1975) 2160.
10. R.W. Schevitz,A.D. Podjarny, M. Zwick, J.J. Hughes and P.B. Sigler, Acta Cryst. A37, (1981) 669.
11. T.N. Bhat and D.M. Blow, Acta Cryst. A38, (1982) 21.
12. B.C. Wang, (1985), Methods in Enzymology 115, (1985) 90.
13. E. Westbrook, private Communication.
14. A.D.Podjarny,J.L.Sussman,T.N. Bhat, E.M. Westbrook,M.Harel,A.Yonath and M. Shoham, Acta Cryst. A40 Supplement, (1984) C-14.
15. A.G.W. Leslie, Collaborative Computational Project Number 4:Protein Crystallography Quarterly Newsletter 17, (1986),1. (Internal Publication,Daresbury Laboratory)
16. R. Reynolds, S.J.Remington., L.H. Weaver, R.G. Fisher, W.F.Anderson, H.L. Ammon and B. W. Mathews, Acta Cryst. B41, (1985) 139.
17. A.D.Podjarny,B. Rees, J.C. Thierry, J. Cavarelli, J.C. Jesior, M. Roth, A.Lewitt-Bentley, R.Kahn, B.Lorber,J.P. Ebel,R. Giege and D. Moras,J.Biomol.Struct. and Dynamics 5, (1987) 187.
18. M.Zwick,D.Bantz and J. Hughes, Ultramicroscopy 1, (1976) 275.
19. N.V.Raghavan and A. Tulinsky,Acta Cryst. B35, (1979) 1776.
20. P. Main, Acta Cryst. A35, (1979) 779.
21. R.J. Read, Acta Cryst. A42, (1986) 140.
22. G.A. Sim, Acta Cryst. 12, (1959) 813.
23. C. Keith,D. Feldman, S. Deganello, J. Glick, K. Ward,E. Oliver Jones and P.B.Sigler, JBC 256, (1981) 8602.
24. E.Westbrook,O.Piro and P.B. Sigler, JBC 259, (1984) 9096.
25. D.W.Rice, Acta Cryst. A37, (1981) 491.
26. D. Stuart and P. Artymiuk, Acta Cryst. A40, (1985) 713.
27. C. Zelwer, private communication.
28. A. Tulinsky, Methods in Enzymology 115, (1985) 77.
29. A.D. Podjarny, T.N. Bhat and M. Zwick,Ann. Rev. Biophys. Biophys. Chem.16, (1987) 351.
30. O. Herzberg and J. Moult, Acta Cryst. A43 Supplement, (1987) C-25.
31. L. Lebioda and B. Stec, Acta Cryst. A43 Supplement, (1987) C-27.
32. B. Rees, J.P. Samama, J.C. Thierry,M. Gilibert,J. Fischer, H. Schweitz, M. Lazdunski and D. Moras, PNAS 84, (1987) 3132.
33. A.H. Robbins and C.D. Stout, Acta Cryst. A43 Supplement,

(1987) C-28.
34. J.P. Rose, D.S.C. Yang, W. Furey, C.S. Yoo, M.Sax , E. Breslow and B.C.Wang, Acta Cryst. A43 Supplement, (1987) C-30.
35. E.Prince, L.Sjolin and R. Alenjlung,Acta Cryst. A44,(1988) 216.
36. T.N. Bhat, Annual Meeting of the ACA, Lexington, Kentucky, (1984), Section Q.
37. T.N. Bhat, Annual Meeting of the ACA, Stanford, California,(1985) Section H.
38. T.N. Bhat, private communication.
39. J. Navaza, Acta Cryst A41, (1985),232.
40. A.D. Podjarny, D.Moras, J. Navaza and P. Alzari.Submitted for publication, Acta Cryst. A.
41. P. Alzari, J. Navaza, A.D Podjarny and R. Poljak, Crystallographic Computing School, Adelaide, Australia, August 1987.

Improving Electron Density Maps by Density Modification

by:

Eleanor Dodson

Department of Chemistry, University of York, Heslington, York

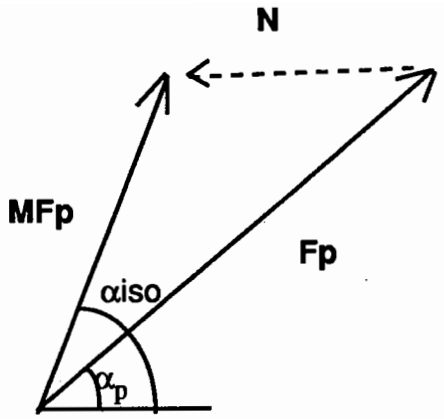
The atomic coordinates of macromolecules are usually obtained by building a skeleton structure into error ridden electron density maps, either phased from isomorphous addition of heavy atoms, or from a model structure fitted by molecular replacement. The errors are mostly due to imperfect phasing; the intensity measurements are comparatively good. We attempt to screen out phasing error by weighting each amplitude with a Figure of Merit, M, which is less than or equal to one. This is based on the quality of the phase.

The effect of these inadequacies is to cloud the image of the unknown structure in various ways. Certainly the true density will be damped by at best a factor of (Σ(MFp) / Σ(Fp)).

Before discussing methods for improving the image let us draw the vector diagram for a structure amplitude **F(hkl).**

Fp is the true structure amplitude, magnitude Fp, and phase αp.

MFp is the structure amplitude used in the electron density map, with magnitude MFp, and phase α(estimate). (In isomorphous phasing M is the figure of merit, and α(estimate) is αiso.)



Then **N = MFp - Fp.**

To write it as a fourier summation

$$\rho(\text{iso}) = \Sigma \mathbf{MFp} \quad \exp(-2\,i\,\pi \mathbf{H.R})$$

$$= \Sigma(\mathbf{Fp} + \mathbf{N}) \exp(-2\,i\,\pi \mathbf{H.R})$$

$$= \Sigma \mathbf{Fp} \exp(-2\,i\,\pi \mathbf{H.R}) + \Sigma \mathbf{N} \exp(-2\,i\,\pi \mathbf{H.R})$$

$$= \text{TRUE MAP} + \text{NOISE MAP.}$$

and the isomorphous map is equal to the sum of the true protein map plus a "noise" map.

If it is possible to recognise any part of the noise contribution this can be removed from the map, and an inverse transform of this density should give structure amplitudes **M'Fp** which will be nearer **Fp**, and will produce a better map.

The simplest visible effects of noise are

1. Negative density anywhere,

2. Positive density in solvent regions

3. Differences between chemically similar molecules in the assymmetric unit.

What improvements can be made?

1. Correcting the negative density is easy, but this gives only a small contribution .to the noise map.

2. Flattening density in the solvent regions helps according to the percentage of solvent in the crystal. The solvent boundary can be generated automatically by the density averaging algorithm suggested by B.C.Wang (6) . A.L.Leslie (2) showed how to modify the structure factors to generate the same solvent mask from reciprocal space which is much faster. After the protein map has been modified its inverse transform generates a modified set of intensities and phases, which are combined with the original phases to give better phases and a better map, always providing the density modification was sensible. (See Appendix for details of procedure.)

3. Averaging density for similar molecules includes solvent flattening, and also modifies the protein density.

Peter Main and Yong Jian Zhang(7) have shown that it is possible to modify the positive density to give a theoretically sensible distribution and that this too will improve the phases.

Of course for stages 2. and 3. it is necessary to have a protein boundary, and for density averaging it is also necessary to be able to assign which part of the density belongs to each molecule and the rigid body transformations between them.

In York last year we used the solvent flattening technique alone to improve two unsolved isomorphous maps, lactoferrin(1) and ribonucleaseSA(5). The 3Å lactoferrin isomorphous map had been studied by experienced crystallographers and they had not been able to trace the 700 residue long chain. After solvent flattening and extending the phases from 3.1Å resolution to 2.8Å the map had improved sufficiently to allow most of the residues to be assigned. And yet when they looked back to the isomorphous map there were very few obvious differences. The psychology of interpreting maps is not really understood; it seems that if too many small decisions are ambiguous the crystallographer loses confidence. Studying equivalent sections of the isomorphous map, the phase refined map and the phase extended map showed just how subtle the changes are at each stage.

However the lactoferrin structure has not yet been fully refined, so it seemed better to do an analysis of the results for ribonucleaseSA where the R factor is now 18%. Rather than compare phases before and after density modification and phase refinement, I tabulated the correlation coefficients for various maps with the Fcalc map based on these final refined protein coordinates. The correlation coefficient will reflect the vector error for each reflection, whereas a phase error of 180° for a very small amplitude will have no effect on the map at all. The correlation coefficient of the whole map and that for each residue was calculated, and plotted for molecule A.

The definition of the correlation coefficient is:

$$\text{correlation coefficient} = \frac{\text{ave}(\text{rho1}*\text{rho2}) - \text{ave}(\text{rho1})*\text{ave}(\text{rho2})}{\sqrt{(\text{ave}(\text{rho1}^2)-\text{ave}(\text{rho1})^2)}*\sqrt{(\text{ave}(\text{rho2}^2)-\text{ave}(\text{rho2})^2)}}$$

The crystallographic details for RibonucleaseSA are:

CELL 64.9 78.32 38.79 90 90 90

SPACEGROUP P212121

Two molecules per assymmetric unit, each of 96 residues

Solvent fraction 0.48

R factor for 1832 atoms, 1600 protein, 203 waters 18.12%

DATA COLLECTION

Native - Synchroton 1.9Å 17202 reflections merging R 0.056

PtCl4 - CuKα- Rotation Camera 2.5Å 6923 reflections merging R 0.057

Iodine - CuKα- Rotation Camera 2.5Å 6925 reflections merging R 0.075

Low resolution diffractometer data.

MAPS

Isomorphous map to 2.5Å phased on PtCl4 and Iodine _ figure of merit 0.67

Isomorphous map to 2.5Å phased on PtCl4 _ figure of merit 0.54

The isomorphous phases were calculated to 2.5Å from the two derivatives, using the anomolous pairs. The data was collected on film, and was incomplete at low resolution so 6Å diffractometer data was merged with the film data for each set. This set of phases gave the first map which was used for phase refinement at 2.5Å, and then as a basis for phase extension to 1.9Å. Although there were two molecules in the assymmetric unit, it was not possible to get accurate parameters for the transformation matrix to fit one onto the other until the backbone of the molecules had been built, so density averaging was not used to improve the phases. (This is often the case, unless the heavy atom coordinates bind to each molecule at equivalent residues, helping to fix the transformation matrix accurately.)

However the correlation against the averaged density is given for reference. B.C. Wang suggested using solvent flattening with single isomorphous phases, so I examined the results obtained from several starting sets of phases. We now know that the Iodine was a very poor derivative, and in fact these results show we would have done better to have omitted it altogether! The first question to examine was to assess the quality of the mask. The ideal density generated from "FC" was correlated with the "FC" map truncated by the various masks derived from different sets of isomorphous phases. Obviously a perfect mask would have given correlation = 1.0. The results are given for the whole map and for each of the two molecules for both the main chain and side chain density.

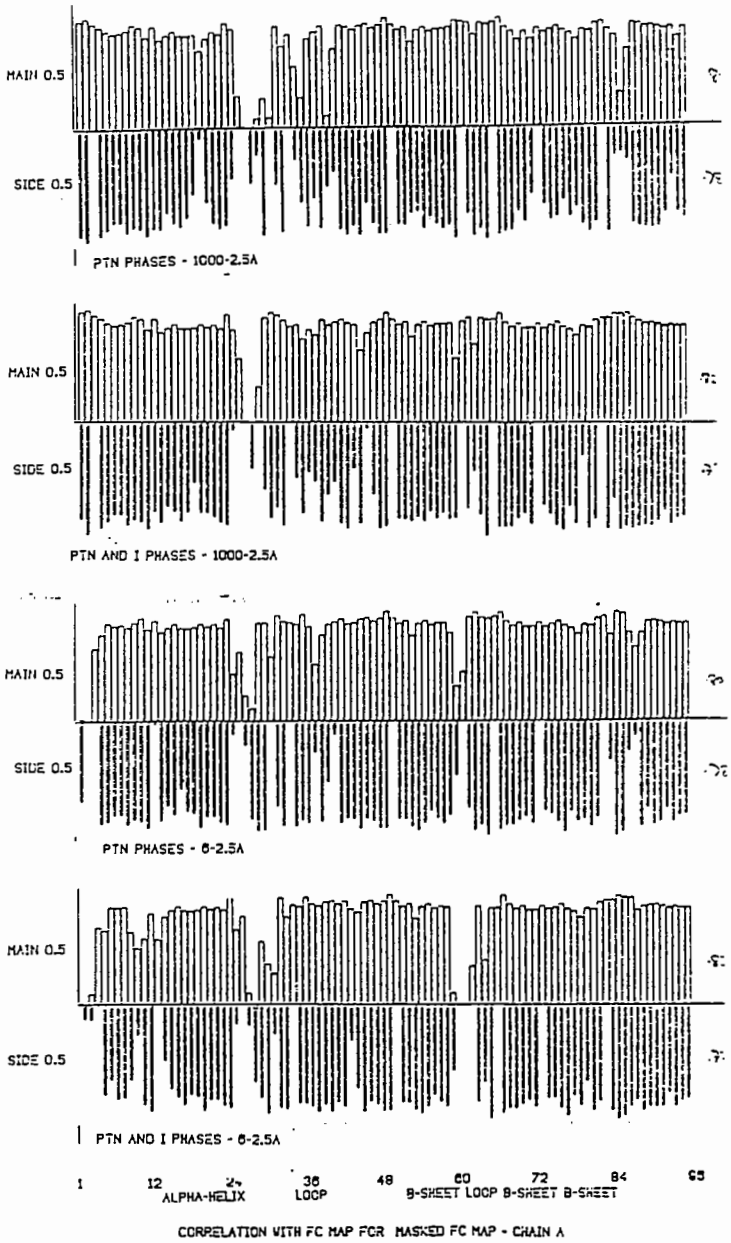


Figure 1. Correlation of ideal density with masked ideal density plotted for residues 1 to 96; chain A RibonucleaseSA.

a) Mask derived from isomorphous phases from ptn - resolution 1000-2.5A

b) Mask derived from isomorphous phases from ptn/iod - resolution 1000-2.5A

c) Mask derived from isomorphous phases from ptn - resolution 6-2.5A

d) Mask derived from isomorphous phases from ptn/iod - resolution 6-2.5A

Correlations for

1. Ideal density v. Masked Fcalc map

Masks from
1) Isomorphous map based on ptn data (1000-2.5Å)
2) Isomorphous map based on ptn/iod data (1000-2.5Å)
3) Isomorphous map based on ptn data (6-2.5Å)
4) Isomorphous map based on ptn/iod data (6-2.5Å)

TOTAL	A(main chain	-side chain)	B(main chain	-side chain)
1) 0.77	0.80	0.78	0.73	0.73
2) 0.78	0.82	0.77	0.75	0.75
3) 0.77	0.80	0.78	0.73	0.73
4) 0.78	0.82	0.77	0.75	0.75

(see Fig. 1 which shows the residues which were truncated by the mask) Note that all four masks gave almost identical overall correlation coefficients. The poor iodine derivative in the (6 - 2.5A) range has generated a mask which has destroyed the density at several places along the chain. The effect of including the low resolution data terms has been to give a more uniform correlation along the whole chain.

The solvent flattening procedure was carried out for the four starting sets and correlations are listed here for two of them against the following maps.

a) Ideal density v. Isomorphous Map (1000-2.5Å) after masking

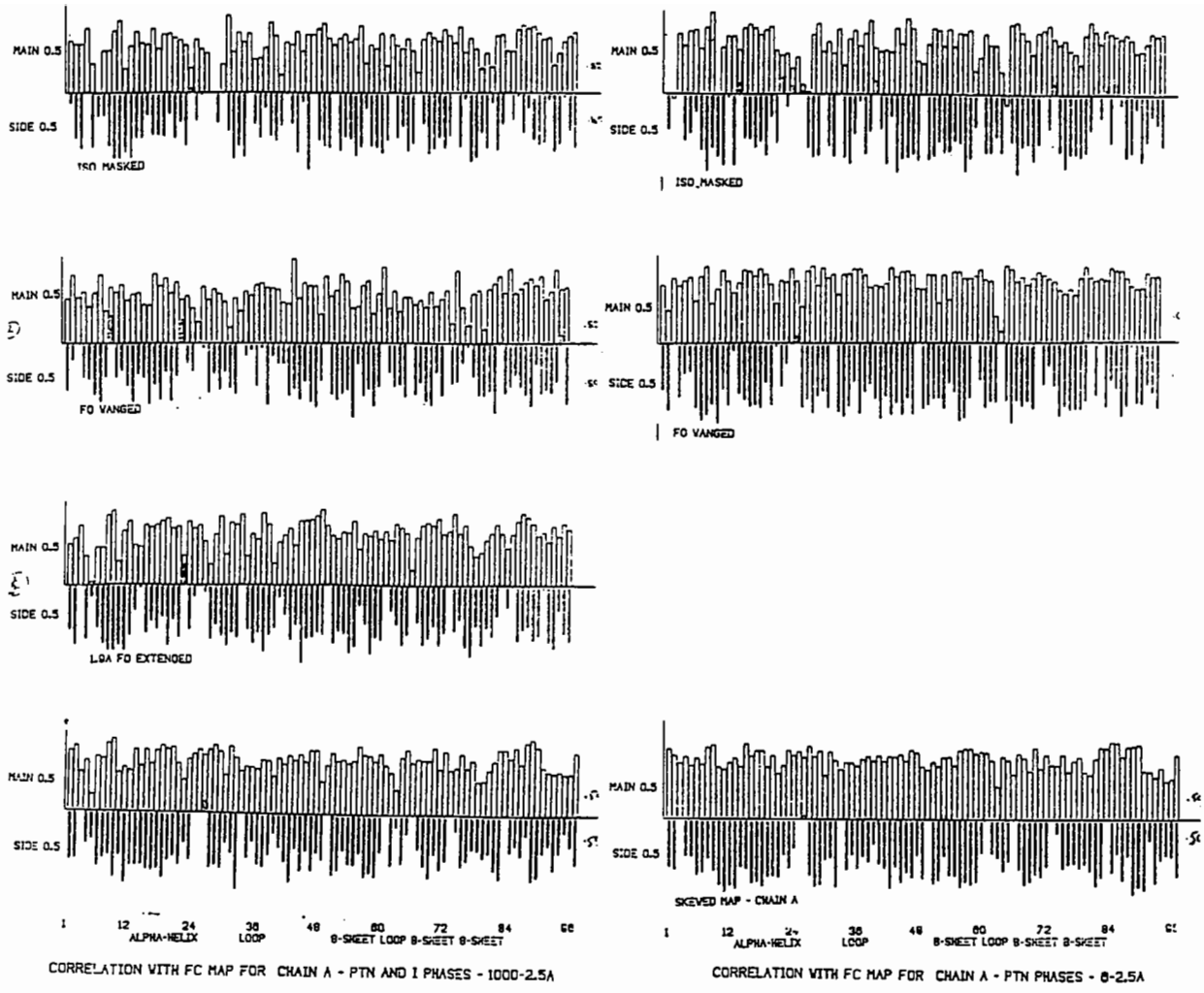


Figure 2. Correlation of ideal density with various protein maps; plotted for residues 1 to 96; chain A RibonucleaseSA.

Set 1 (left hand column) Mask derived from isomorphous phases from ptn/iod - resolution 1000-2.5A

Set 2 (right hand column) Mask derived from isomorphous phases from ptn - resolution 6-2.5A

a) Masked map from iomorphous phases.

b) Map after solvent flattening and phase recombination.

c) Map after solvent flattening and phase recombination, and phase extension to 1.9A. (only done once)

d) Averaged density for molecules A and B from the isomorphous map.

b) Ideal density v. "Wanged" Fobs Map (1000-2.5Å) after four cycles of density modification and phase combination.

c) Ideal density v. "Wanged" Fobs Map (1000-1.9Å) after four cycles of density modification and phase combination at 2.5Å followed by phase extension in 0.1Å steps.

d) Ideal density v. Isomorphous 2.5Å map averaged for the two molecules (output of SKEWPLANES)

Masks and phases from

1) Isomorphous map based on Pt data (1000-2.5Å)

	TOTAL	A(main chain-side chain)		B(main chain-side chain)	
a)	0.49	0.57	0.53	0.52	0.49
b)	0.51	0.58	0.52	0.58	0.52
c)	0.39	-	-	-	-
d)	-	0.59	0.50		

2) Isomorphous map based on Pt/I data (1000-2.5Å)

	TOTAL	A(main chain-side chain)		B(main chain-side chain)	
a)	0.45	0.52	0.48	0.46	0.44
b)	0.44	0.55	0.50	0.51	0.46
c)	0.39	-	-	-	-
d)	-	0.54	0.45		

Points to note are:

1) The correlation coefficients show that the overall changes in map quality are small but real.

2) The error in the masks which truncated the B molecule more severely than the A molecule were corrected in a very encouraging way - there was more improvement seen for the B molecule`residues.

3) The phases derived from the PtCl4 derivative data and its anomolous pairs alone gave both a better starting set of phases for refinement than those based on both PtCl4 and the not-quite-isomorphous iodine derivative data, and also gave better refined phases.

4) In this case phase extension from 2.5Å to 1.9Å gave little if any improvement in the map. The fall in the overall correlation is unfair, since it reflects the increased "peakiness" of the maps at 1.9Å

5) Density averaging of the isomorphous map improved the main chain but reduced the quality of the sidechain density.

APPENDIX: THE PROCEDURE OF WANGING

B.C.Wang suggested a technique which allowed the protein boundary to be assigned automatically by computer. This is the essential and most difficult part of any solvent flattening procedure. He substituted at each grid point the average value of the electron density of the original map taken over a sphere centred on that point, then chose a limiting value of ρ which allowed him to assign the required percentage of grid points to solvent when the value of ρ was below this cutoff. This gave him a solvent mask which could be used to truncate the original map.

His averaged map looked rather like a very low resolution isomorphous map. Even a poor map (providing it is not randomly phased!) can generate a reasonable mask.

The truncated map is used to generate new structure factors and phases. (Remember if there had been no change to the isomorphous map the sructure factors generated will be equal to MFp, phase α(iso))

These new amplitudes and phases are combined with the original set, using some system such as the SIGMAA weighting scheme described by Randy Read(3), and a new protein map using the modified values of M and α is calculated.

The procedure can then be recycled through the solvent flattening, structure factor generation, phase combination and map calculation, using the same mask till the phases stop changing - usually about four cycles suffice. Then the mask can be recalculated from the new protein map and the refinement procedure continue. There is little advantage in doing this more than two or three times.

The slow part of this procedure in the original WANG program suite was the density averaging to produce the mask. However Andrew Leslie(2) pointed out that this step could be done in reciprocal space - modified structure factors are derived from the isomorphous map, with appropriate weighting and these are used to calculate the averaged map.

I have polished his programs a little; the structure factor routine will now

a) truncate the map

b) apply the weighting appropriate for a given sphere of averaging,

c) generate the unique set of amplitudes for the structure's spacegroup

(ie it is no longer necessary to work in the P1 space group)

Stages are

1. Calculate isomorphous map in correct spacegroup (use FFT program)

2. Read isomorphous map, set negative regions equal zero, truncate high peaks if desired, and calculate a set of weighted "FCs" appropriate to the "averaging" radius chosen. (ref B.C.Wang, A.Leslie) (using SFRF programs)

3. Calculate "averaged" map in correct spacegroup from these "FCs"(use FFT program)

4. Generate envelope - only data needed is the percentage of solvent, reads a map and outputs a mask. (Use a program from the B.C.WANG suite ENVELOP1 modified to read our map format.)

The following four steps are recycled till phases converge.

5. Flatten the isomorphous map outside the mask, reads the map and the mask and outputs a modified map. (Use a program from the B.C.WANG suite -FLATMAP1- modified to read our map format.)

6. Use flattened map to generate new F and phase. Read map and the intensity data file containing h k l Fobs α(iso) and the Hendrickson-Lattmann coefficients to be used for phase combination. Outputs h k l Fobs ... plus "Fc αc".

These will be similar to the input MFobs α(iso) and the amount they change will be a function of the changes imposed on the input map by masking. (Use SFRK program)

7. COMBINE new phases with isomorphous phases

We use Randy Read's SIGMAA program modified to read an LCF intensity file, but any phase combination program will do.

8. Calculate new Fp map with these combined phases

SECOND CYCLE - return to flatten this map with the same mask. Usually four to six cycles produce convergence

REFERENCES

1. Anderson.B.F., Baker.H.M., Dodson.E.J., Norris.G.G., Rumball.S.V., Waters.J.M., and Baker.E.N. P.N.A.S.(USA) (1987) 84, pp 1769-1773.

2. Leslie,A.G.W.(1987). Protein crystallography Newsletter.

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

4. Program Suite for Protein Crystallography (CCP4) SERC Daresbury Laboratory,WARRINGTON. U.K.

5. Sevcik.J.,Dodson.E.J., Dodson.G.G., and Zelinka.J.(19867) pp 33-46 in Metabolism of Nucleic Acids, including Gene Manipulation. Slovak Academy of Science. BRATISLAVA. Czeckoslovakia.

6. Wang B.C. (1985) Methods in Enzymology, Vol. 115, Diffraction Methods for Biological Macromolecules, edited by H.W. Wyckoff, C.H.W. Hirs and S>N> Timasheff, pp 90-111 Academic Press,Inc.

7. Yong Jian Zhang and Peter Main(1988) Submitted to Acta Cryst.

THE USE OF SOLVENT-FLATTENING PROCEDURES IN THE CRYSTAL STRUCTURE DETERMINATION OF QUINOPROTEIN METHYLAMINE DEHYDROGENASE

by

F.M.D. Vellieux, H. Groendijk, F. Huitema, M.B.A. Swarte, J. Drenth and W.G.J. Hol
Laboratory of Chemical Physics, University of Groningen
Nijenborgh 16, 9747 AG Groningen, The Netherlands

1. INTRODUCTION

Quinoproteins are enzymes which contain the recently discovered cofactor pyrrolo-quinoline quinone [1], or PQQ (fig. 1). The existence of PQQ as a cofactor was first shown with the enzyme methanol dehydrogenase [2,3]. More recently, De Beer et al. proposed that the cofactor of the enzyme methylamine dehydrogenase (or MADH) has a structure similar to that of PQQ [4]. In order to study the exact structure and function of this cofactor in catalysis by MADH, we have embarked upon a project aimed at determining the complete three-dimensional structure of this enzyme [5].

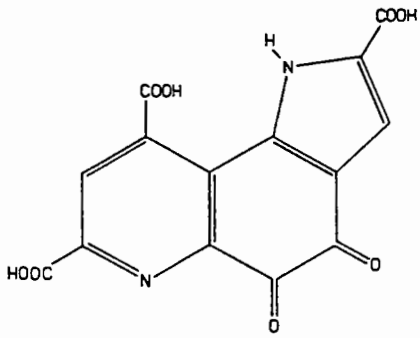
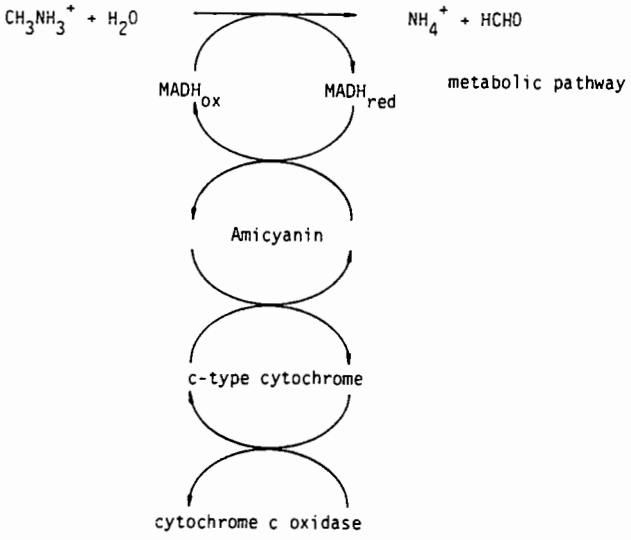


Fig. 1 PQQ, the cofactor of quinoproteins.

Methylamine dehydrogenase catalyzes the oxidative deamination of methylamine to formaldehyde and ammonia [6,7] according to:



The enzyme has a relative molecular mass of 123,500 [5]. It is a tetramer having an $\alpha_2\beta_2$ subunit structure, where the two subunit types have Mr 47,500 (heavy subunit) and 12,900 (light subunit) [5]. The PQQ is covalently bound at two sites in the light subunit [8,9].

2. STRUCTURE DETERMINATION PRIOR TO SOLVENT FLATTENING

2.1 Crystallization

Crystals of methylamine dehydrogenase were grown as described previously [5]. Large single crystals (ca. 0.8 * 0.8 * 0.9 mm^3) of MADH are found in the hanging drops after three weeks at 4°C. The space group of the crystals is $P3_121$, with a = 130.4 Å; b = 104.2 Å. They diffract to at least 1.9 Å.

2.2 Heavy atom derivatives

After screening more than thirty reagents, three heavy atom compounds were found to give suitable changes in the diffraction pattern of MADH crystals. The results of this heavy atom derivative search are given in table 1.

Table 1 The heavy atom derivatives used in the structure determination of MADH

Heavy atom reagent	Concentration	Soaking time
K_2PtI_6	saturated	72 hrs.
$Pt(en)Cl_2$	5 mM	48 hrs.
$UO_2(CH_3CO_2)_2$	2 mM	24 hrs.

2.3 Data collection and processing

All the data sets were collected using an Arndt-Wonacott rotation camera [10]. A native data set was collected with a rotating anode X-ray source, and three derivative data sets were collected on the X-31 beam line of the DESY-EMBL outstation in Hamburg. With the exception of the Uranyl acetate data set, where integrated intensities were obtained by the method of profile fitting [11] using the program OSC, all other data sets were processed with the program OSCIL [12], where the intensities are obtained by simple summation. Table 2 gives a summary of data collection and processing. Processing of data collected at the synchrotron was simpler than those obtained from the rotating anode, since the background level on these films is greatly reduced. Also, due to the very low beam divergence, reflections are very sharp and well separated from each other. Furthermore, the high flux and monochromaticity of the beam increase the resolution limits and extend the lifetime of MADH crystals. These beneficial effects can be seen from the low R_{sym} values for the derivative data sets, which must be compared to the higher value (ca. 10%) obtained for the native data.

Table 2 Data collection statistics

Data	#crystals	Scan range	#filmpacks collected	#filmpacks processed	resolution	R_{merge}	R_{sym}	completeness
Native	7	176.55°	114	72	2.8 Å	5.8%	10.02%	91.5%
K_2PtI_6	3	106.30°	72	46	2.4 Å	4.8%	5.01%	79.3%
$Pt(en)Cl_2$	2	79.00°	51	40	2.7 Å	8.6%	6.03%	83.3%
$UO_2(CH_3CO_2)_2$	2	104.50°	72	71	2.55 Å	9.78% (on $\|F\|^2$)	5.44%	82.0%

$$R_{merge} = \frac{\sum^{N} |F_i - \bar{F}|}{\sum F_i} * 100; \quad R_{sym} = \frac{\sum^{N} |F_i - \bar{F}|}{N * \bar{F}} * 100.$$

Table 3 MIRAS phase calculation statistics

		Resolution range (Å)								
		35.0-18.95	12.99	9.88	7.97	6.68	5.75	5.05	4.50	Total
	Figure of Merit	0.79	0.82	0.84	0.80	0.81	0.79	0.74	0.66	0.75
$UO_2(CH_3CO_2)_2$	Centric Cullis R	61.33	46.91	53.33	49.13	56.32	52.27	66.51	81.88	60.77
2 heavy atom sites	Phasing Power	1.76	2.08	1.78	1.90	2.04	1.94	1.47	1.07	1.56
K_2PtI_6	Centric Cullis R	54.74	57.85	46.90	44.40	55.49	58.87	78.61	86.61	62.35
3 heavy atom sites	Phasing Power	1.48	1.50	1.65	1.68	1.77	1.69	1.45	1.14	1.45
$Pt(en)Cl_2$	Centric Cullis R	71.99	66.22	53.02	55.89	51.44	47.31	57.03	81.33	59.94
5 heavy atom sites	Phasing Power	1.27	1.56	1.92	1.86	2.25	2.23	1.80	1.44	1.76

2.4 MIRAS heavy atom phase determination

Difference Patterson maps were calculated using data in the range 25.0 to 6.0 Å for each derivative. The heavy atom substitution pattern was solved independently for the platinum iodate - and uranyl acetate derivative using the VSFUN# vector search programs [13,14]. Heavy atom positions were obtained for the third derivative from a difference Fourier map calculated with DIRAS phases. Heavy atom refinement and phase calculation were carried out using the program PHARE which was kindly made available to us by Dr. G. Bricogne (see Table 3). Phasing statistics were found to deteriorate rapidly when attempts were made to increase the resolution beyond 4.5 Å (figure 2).

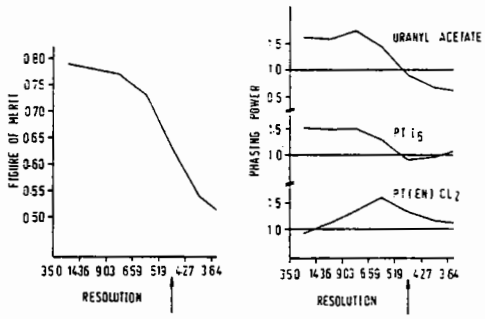


Fig. 2 Phasing to 3.5 Å resolution

2.5 The 4.5 Å electron density map

An electron density map was calculated at 4.5 Å resolution with best phases and figure of merit values [15] using the XRAY program suite [16]. The map was plotted as sections across the c axis on transparent plastic sheets. A few superimposed sections of this map are shown in figure 3.

Fig. 3 Slab of a few superimposed sections of the 4.5 Å electron density map.

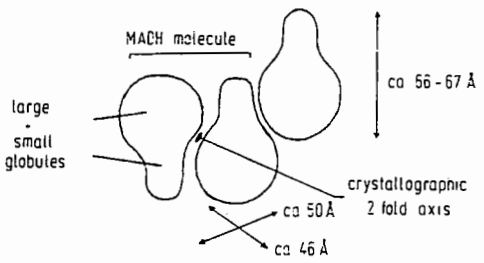


Fig. 4 Low resolution interpretation of the structure of MADH.

The map shows a very large volume of solvent in the unit cell, with a clear boundary between protein and solvent regions. In fact, interpretation of the map lead us to revise our ideas concerning the protein content in the asymmetric unit [5]. The asymmetric unit is seen to contain only half an enzyme molecule, i.e. an ($\alpha\beta$) dimer (fig. 4). Assuming that the partial specific volume for MADH is 0.74 $cm^3\ g^{-1}$, we obtain V_M = 4.14 $Å^3\ Da^{-1}$ [17]. This corresponds

to a value of ca. 70% for the fraction of the volume of the unit cell which is occupied by solvent.

The map shows the protein separating into subunit-like globules (fig. 4). Furthermore, within the protein regions, several secondary structure elements can be identified as surface loops or β sheet regions. However, the complete determination of the structure of MADH required higher resolution information. Since the solvent content of the crystals is quite high, we investigated possibilities to obtain such information using solvent flattening.

3. SOLVENT FLATTENING

Solvent flattening is only one member of the family of methods known as density modification methods. The technique has been employed to determine the structure of several macromolecules [see e.g. 18-21]. In the structure determination of a DNA - Eco RI endonuclease complex, solvent flattening was also used for phase extension from 3.5 to 3.0 Å [18,22]. However, this procedure of phase extension was not entirely successful (due to a large fraction, ca. 28%, of reflections which were unavailable). The electron density map was seen to improve sufficiently only after introduction of both calculated amplitudes and phases for the missing reflections [22]. A better test case to show the power of phase extension by solvent flattening would require phase extension starting from low resolution (at which a macromolecular structure cannot be solved, as judged from a correct chain tracing), to a resolution at which the main chain can be correctly traced, and using a more complete set of data. In this paper, we are reporting the results of such a test.

Figure 5 shows the solvent flattening cycle as programmed by B.C. Wang and coworkers [23] and subsequently modified by A. Leslie [24]. Different from the original method of B.C. Wang, solvent flattening was carried out without density truncation within the protein region [25].

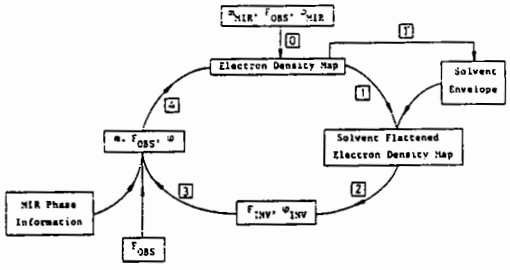


Fig. 5 The solvent flattening cycle.

4. SOLVENT FLATTENING PROCEDURES

Since solvent flattening and molecular averaging are essentially equivalent, we expect to have to follow similar strategies with both methods (see appendix). In particular, with phase extension, we will begin our procedure with starting estimates for the phases which are close to the correct values for these phases [26]. Phase extension is then carried out gradually, extending the resolution by a small volume in reciprocal space at a time. The choice of a suitable size for the phase extension step can be derived by considering the G function of Rossmann and Blow [27] for the molecule under consideration.

With MADH, we considered the current MIRAS phases to be reliable only to 4.5 Å resolution. The strategy we decided to follow was therefore first to improve the MIRAS phases by solvent flattening, then to extend the phases gradually to higher resolution. Throughout the whole procedure, a new mask was calculated at every cycle. We used a conservative value of 0.6 for the solvent content estimate during the calculation of the mask, with R_{mask} set to 10.0 Å.

4.1 Solvent flattening at constant resolution

In order to investigate the effects of phase combination with the original MIRAS phases we carried out solvent flattening cycles at 4.5 Å resolution using two different procedures (table 4), which we will refer to as procedure A and procedure B.

Table 4 Statistics of the solvent flattening procedures at 4.5 Å resolution

Procedures	A	B
Phase combination	NO	YES
No. of cycles	8	20
No. of reflections phased	6314	6332
$Rf_{cycle\ 1}$	0.299	0.327
<fom>	0.69	0.78
Rf_{final}	0.145	0.187
<fom>	0.76	0.84
Accumulated phase shift	37.2°	26.0°

With procedure A, no phase combination step is included, whereas with procedure B, a phase combination step of "solvent flattened" and original MIRAS phases is included by simple addition of Hendrickson-Lattman coefficients. In each case, the procedure was stopped when convergence had been reached. This was judged from the stabilization of the R-factor value, as well as the drop of the average phase change between successive cycles to a negligible value. It is clear from table 4 that procedure A converges much faster, and gives lower R-factors than procedure B.

Fig. 6 Electron density maps at 4.5 Å resolution. Detail of a loop region connecting two β-strands
A. MIRAS map
B. map calculated after procedure A i.e. no phase combination with MIRAS phases
C. map calculated after procedure B i.e. with phase combination of inverted and MIRAS phases
The continuous lines within the density represent an alpha carbon tracing obtained from a later map.

After applying both procedures, the density within the protein region has cleared. This is obvious from fig. 6, where the loop is seen to open up, when it appeared as a single bulge of density before solvent flattening. However, at this low resolution it is impossible to decide which of the two procedures gives the most readily interpretable map.

4.2 Solvent flattening and phase extension

At this stage, we assumed that the solvent flattening cycles carried out previously had improved phases sufficiently for phase extension. The resolution was gradually increased from a starting resolution of 4.5 Å to 3.5 Å. Again, two procedures were used (table 5).

Table 5 Statistics of the solvent flattening procedures during phase extension

Procedures	A	B
extension step phases from	map inversion	MIR phases
Phase combination	NO	YES
No. of cycles	37	170
No. of phase extension steps	14	17
New reflections per step	545	408
No. of reflections phased	13815	13261
$Rf_{cycle\ 1}$	0.184	0.226
<fom>	0.74	0.82
Rf_{final}	0.106	0.196
<fom>	0.86	0.83
Accumulated phase shift	39.6°	33.1°

In procedure A no phase combination is carried out (just as in the phase improvement steps at constant resolution), and, hence, the new phases during the phase extension step come from inversion of the solvent flattened map. With the other procedure (B) where phase combination is carried out, the starting estimates for the new phases in the phase extension step are the original MIRAS phases. Both procedures were stopped at convergence. The same behaviour is seen during phase extension than previously at constant resolution: procedure A gives faster convergence than procedure B, and gives also much lower R-factors (table 5). This behaviour is seen more clearly when plotting the R-factors as function of resolution (fig. 7).

Both figures 7A and 7B are typical of a successful phase extension procedure [28]. The results of the phase extension procedures can better be judged from the resulting electron density maps (fig. 8 and 9).

In both cases, we can see that the level of detail in the map has improved. The β-sheet regions are seen to resolve into individual strands, and gaps along the polypeptide chain are filled. Thus, phase extension using the technique of solvent flattening is seen to refine phases towards their "correct" value.

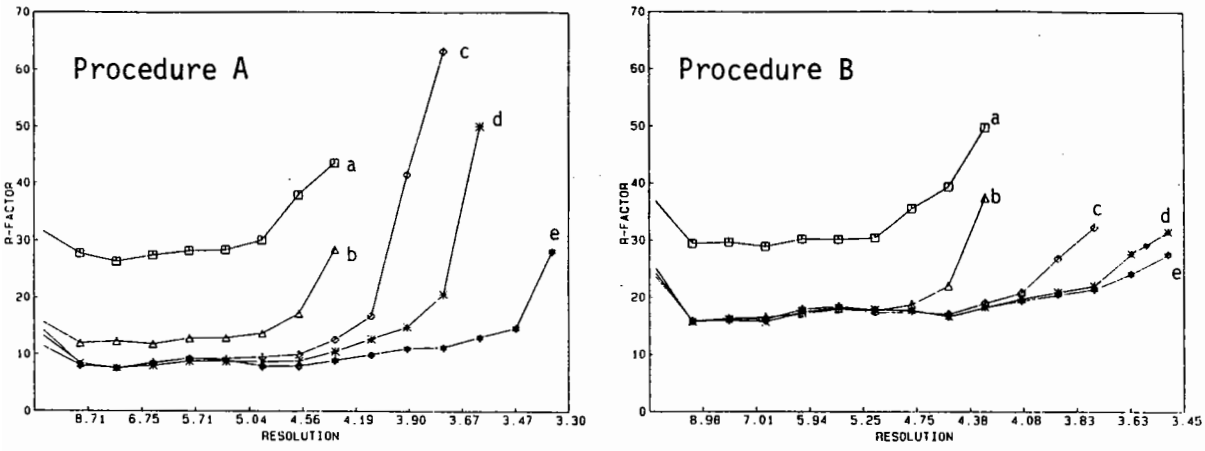


Fig. 7 Results of phase extension: R-factors as function of the resolution
a) after the initial solvent flattening cycle, 4.5 Å resolution
b) after the final cycle at 4.5 Å
c) after phase extension to 3.9 Å
d) after phase extension to 3.6 Å
e) at the end of the phase extension procedure

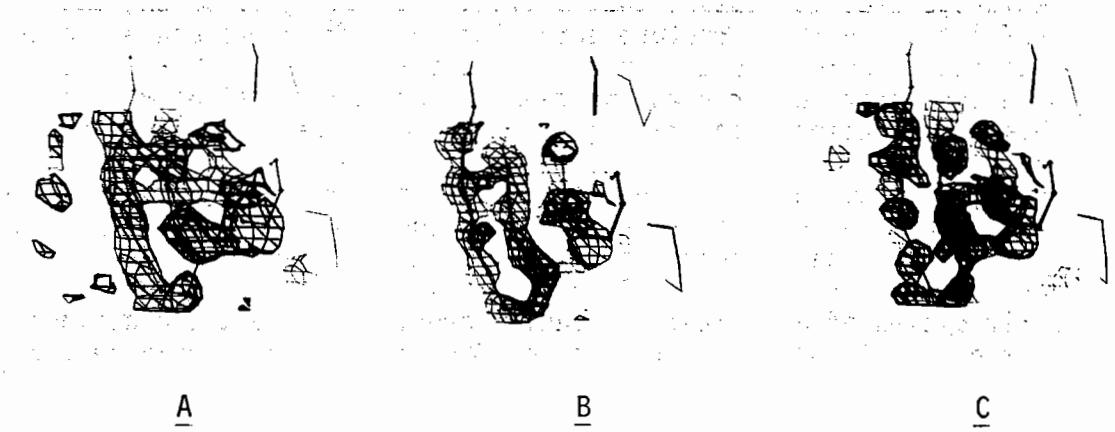
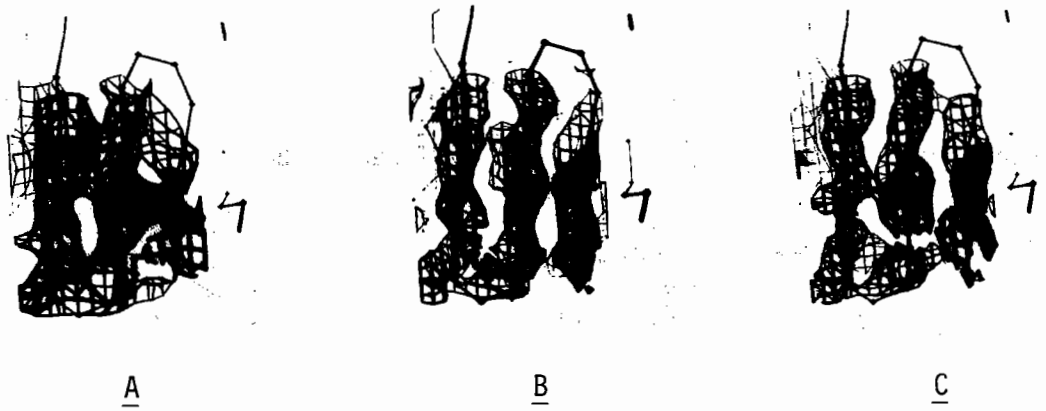


Fig. 8 Detail of a β-sheet region with a loop connecting adjacent strands
A. MIRAS map at 4.5 Å resolution
B. after phase extension from 4.5 to 3.5 Å with procedure A
C. after phase extension from 4.5 to 3.5 Å with procedure B

However, when comparing the map obtained after completion of B against that obtained from A, we immediately see that the former is far less continuous than the latter. In particular, we observe more breaks along the main chain in the map obtained after procedure B.

In the structure determination of MADH, procedure A was successfully applied, providing us with an electron density map in which the polypeptide chain within the large subunit (of Mr 47,500) could be traced rather straightforwardly.



A B C

Fig. 9 Three strands of a β-sheet

A. 4.5 Å MIRAS map
B. 3.5 Å map obtained after procedure A
C. 3.5 Å map obtained after procedure B

5. CONCLUSIONS

From the studies described above with quinoprotein methylamine dehydrogenase, it appears that solvent flattening is an extremely useful tool in the structure determination of macromolecules, even when the technique is used for phase extension. In our case, we have seen that better electron density maps are obtained when no phase combination with original heavy atom phases is performed.

This success with one of the procedures is seen due to a combination of factors. The starting phases are reasonably accurate, as judged from the original electron density map. As a result, the solvent masks were quite reliable. This is particularly important in the early cycles of a solvent flattening procedure, where the definition of the mask is usually not so good. Therefore, care should be taken in these stages to ensure that the masks do indeed correctly represent the solvent regions.

The lower success of phase combination with our work can be attributed to two main reasons. First, by combining phase information from two sources which are not independent, we are probably overestimating the quality of the resulting phases. Second, during phase extension, phase combination was carried out between more reliable solvent flattening phases, and heavy atom phase information of questionable quality. It is therefore expected that phase combination in such a case will always shift the phases back towards the less reliable heavy atom phases.

6. ACKNOWLEDGEMENTS

We wish to thank Dr. J. Frank and Prof. J.A. Duine (Lab. of Microbiology and Enzymology, Delft University of Technology) for their eagerness to provide us with very pure MADH samples which were used for crystallization. This research was carried out under the auspices of the Netherlands Foundation for Chemical Research (SON) with financial aid from the Netherlands Organisation for the Advancement of Pure Research (ZWO).

7. APPENDIX (F.M.D. Vellieux and W.G.J. Hol)

The restrictions on phases caused by the presence of solvent in the unit cell has already been discussed in reciprocal space by Rossmann and Arnold [29]. For the sake of completeness of this paper, the derivation of the molecular replacement equations will be given here.

Peter Main was the first to recognize the importance of the solvent contribution in protein crystals [30]. For a system composed of N identical subunits, he expressed the value of a structure factor $F_{\underline{p}}$ as:

$$F_{\underline{p}} = \sum_{n=1}^{N} \int_U \rho(\underline{x}_n) \exp(2\pi i \underline{p}\underline{x}_n) d\underline{x}_n + \int_{V-NU} \rho_s \exp(2\pi i \underline{p}\underline{x}) d\underline{x} \quad (1)$$

where U is the subunit volume (envelope), V is the asymmetric unit volume and ρ_s is the constant solvent density outside the molecular region. N is the number of identical subunits in the asymmetric unit and $\rho(\underline{x}_n)$ refers to the density within the subunit.

In the simple case where we have only one subunit in the asymmetric unit, we can rewrite (1) as:

$$F_{\underline{p}} = \int_U \rho(\underline{x}) \exp(-2\pi i[-\underline{p}\underline{x}]) d\underline{x} + \int_{V-U} \rho_s \exp(2\pi i \underline{p}\underline{x}) d\underline{x} \quad (2)$$

The second term in this equation represents a fixed contribution by the solvent to the complex structure factor $F_{\underline{p}}$, which we shall write as $F_{\underline{p}}^{s}$. Therefore:

$$F_{\underline{p}} = \int_U \rho(\underline{x}) \exp(-2\pi i[-\underline{p}\underline{x}]) dx + F_{\underline{p}}^{s} \quad (3)$$

We can now replace $\rho(\underline{x})$ by its Fourier expansion:

$$\rho(\underline{x}) = \frac{1}{V} \sum_{\underline{h}} F_{\underline{h}} \exp(-2\pi i \underline{h}\underline{x}) \quad (4)$$

With rearrangement, this gives:

$$F_{\underline{p}} = \frac{1}{V} \sum_{\underline{h}} F_{\underline{h}} \int_U \exp(-2\pi i[\underline{h}-\underline{p}]\underline{x}) d\underline{x} + F_{\underline{p}}^{s} \quad (5)$$

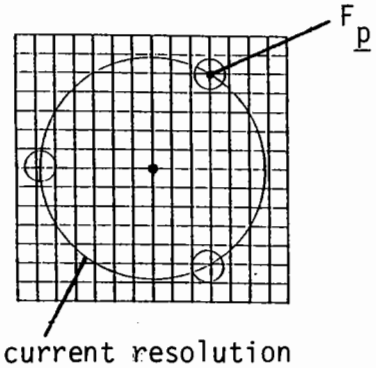
The integral given in (5) can be evaluated. It is the Fourier transform of the molecular envelope U:

$$\int_U \exp(-2\pi i[\underline{h}-\underline{p}]\underline{x}) d\underline{x} = U\, G_{\underline{h},-\underline{p}} \quad (6)$$

where G is the function described by Rossmann and Blow [27]. Substituting, we obtain:

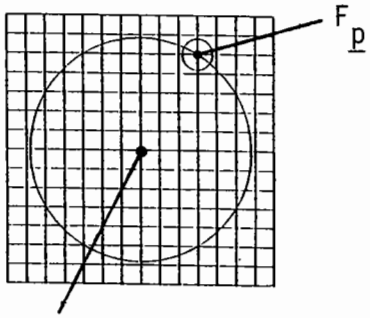
$$F_{\underline{p}} = \frac{U}{V} \sum_{\underline{h}} F_{\underline{h}}\, G_{\underline{h},-\underline{p}} + F_{\underline{p}}^{s} \quad (7)$$

Equation (7), previously given by Arnold and Rossmann [29], expresses the value of a single structure factor $F_{\underline{p}}$ as the weighted sum of all structure factors. The implications of equation (7) for use in macromolecular crystallography can best be seen on schematic diagrams (fig. 10).



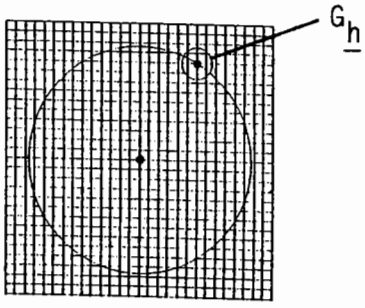
The smaller circles represent the region of reciprocal space where the G function for the 3 independent subunits takes significant values. We can see that the interference function has the effect of relating distant regions of reciprocal space due to the non crystallographic symmetry.

A. Non crystallographic 3 fold symmetry



Here, due to the absence of non crystallographic symmetry, there is no relation existing between distant regions of reciprocal space. Consequently, the phasing power is smaller.

B. Absence of non crystallographic symmetry: only one subunit in the asymmetric unit.



With the same subunit in the asymmetric unit, the G function is identical to that in B; however, the addition of solvent implies that the unit cell volume increases, and consequently the reciprocal lattice skrinks. Therefore, more points contribute to the value of $F_{\underline{p}}$, and the phasing power increases.

C. Higher solvent content.

Fig. 10 The importance of the G function for density modification methods.

8. REFERENCES

1. C. Hartmann and J.P. Klinman, Biofactors 1, (1988) 41.
2. S.A. Salisbury, H.S. Forrest, W.B.T. Cruise and O. Kennard, Nature 280, (1979) 843.
3. R. de Beer, D. van Ormondt, M.A. van Ast, R. Banen, J.A. Duine and J. Frank, J. Chem. Fys. 70, (1979) 4491.
4. R. de Beer, J.A. Duine, J. Frank and P.J. Large, Biochim. Biophys. Acta 622, (1980) 370.
5. F.M.D. Vellieux, J. Frank, M.B.A. Swarte, H. Groendijk, J.A. Duine, J. Drenth and W.G.J. Hol, Eur. J. Biochem. 154, (1986) 383.
6. R.R. Eady and P.J. Large, Biochem. J. 106, (1968) 245.
7. R.R. Eady and P.J. Large, Biochem. J. 123, (1971) 757.
8. Y. Ishii, T. Hase, Y. Fukumori, H. Matsubara and J. Tobari, J. Biochem. (Tokyo) 93, (1983) 107.
9. W.S. McIntire and J.T. Stults, Biochem. Biophys. Res. Commun. 141, (1986) 562.
10. U.W. Arndt, J.N. Champners, R.P. Phizackerley and A.J. Wonacott, J. Appl. Cryst. 6, (1973) 457.
11. M.G. Rossmann, J. Appl. Cryst. 12, (1979) 225.
12. P. Schwager, K. Bartels and A. Jones, J. Appl. Cryst. 8, (1975) 275.
13. C.E. Nordman, Acta Cryst. A28, (1972) 134.
14. P. Argos and M.G. Rossmann, Acta Cryst. A30, (1974) 672.
15. D.M. Blow and F.H.C. Crick, Acta Cryst. 12, (1959) 794.
16. J.M. Stewart (1976) The XRAY-system. Technical report TR-446 of the computer science center. University of Maryland, College Park, Maryland.
17. B.W. Matthews, J. Mol. Biol. 33, (1968) 491.
18. C.A. Frederick, J. Grable, M. Melia, C. Samudzi, L. Jen-Jacobson, B.C. Wang, P. Greene, H.W. Boyer and J.M. Rosenberg, Nature 309, (1984) 327.
19. J. Deisenhofer, O. Epp, K. Miki, R. Huber and H. Michel, J. Mol. Biol. 180, (1984) 385.
20. Z. Xia, N. Shamala, P.H. Bethge, L.W. Lim, H.D. Bellamy, N.H. Xuong, F. Lederer and F. Scott Mathews, P.N.A.S. 84, (1987) 2629.
21. O. Herzberg and J. Moult, Science 236, (1987) 694.
22. J.A. McClarin, C.A. Frederick, B.C. Wang, P. Greene, H.W. Boyer, J. Grable and J.M. Rosenberg, Science 234, (1986) 1526.
23. B.C. Wang, Methods Enzymol. 115, (1982) 90.
24. A. Leslie, Acta Cryst. A43, (1987) 134.
25. H.M. Holden, W.R. Rypniewski, J.H. Law and I. Rayment, EMBO Journal 6, (1987) 1565.
26. G. Bricogne, Acta Cryst. A30, (1974) 395.
27. M.G. Rossmann and D.M. Blow, Acta Cryst. 15, (1962) 24.
28. W.P.J. Gaykema, A. Volbeda and W.G.J. Hol, J. Mol. Biol. 187, (1986) 255.
29. E. Arnold and M.G. Rossmann, P.N.A.S. 83, (1986) 5489.
30. P. Main, Acta Cryst. 23, (1967) 50.

MAXIMUM ENTROPY ESTIMATES OF THE ELECTRON DENSITY

by

Jorge NAVAZA

Immunologie Structurale, Institut Pasteur, Paris and
E.R. 180 du C.N.R.S., 92290 Chatenay Malabry, France

1. INTRODUCTION

Most formulations of the phase problem that invoke the principle of maximum entropy as their fundamental basis, involve the constrained maximization of a local functional of the sought map, the configurational entropy. From a theoretical point of view they were proved to be particular cases (in the sense that the same equations can be recovered) of a more general formulation in which the sought map is defined as the expected value of the admissible maps (i.e. those maps that satisfy the a priori information on the electron density) computed with a maximum entropy probability distribution (Navaza, 1985; JN1 hereafter). Basically the estimated map may be constrained to take a constant value at the solvent region, satisfy the prior information of positivity and boundedness, and fit the observed Fourier coefficients of the true electron density.

As long as the phases of the observed structure factors are known (the convex problem), the equations that determine the sought map admit a unique solution, provided that no inconsistency in the data exists. Most of the numerical algorithms so far proposed to solve this problem should produce practically the same map. A certain amount of super-resolution can be expected, whose meaning and origin was already discussed (Navaza, 1986; JN2 hereafter). On the contrary, if the phases are not available (the non-convex problem), the final map strongly depends on the starting point of the calculations, as well as on the particular algorithm used. Therefore, a characterization of the resulting solution is highly desirable and most useful at a time when still great speculation about the phasing power of maximum entropy methods exists.

The algorithm described in JN1 to solve the non-convex problem allows for a complete local characterization of the solution, at the cost of a great computational effort. The procedure considers the configurational entropy as an implicit function of the unknown phases. The calculation of this function and its derivatives, from which the phases are iteratively updated, involves basically an accurate solution of the convex problem (Newton-Raphson

procedure) and the inversion of a positive definite Toeplitz matrix. It was thus possible to show that most local configurational entropies are not necessarily concave functions of the phases in a neighbourhood of the correct solution, i.e. the correct phases are not even close to a local maximum. It was also found that, in the case of non-centrosymmetric structures, Boltzmann's entropy $[-\rho\ell n(\rho)]$ very often admits a centrosymmetric solution which is locally a maximum, whereas other configurational entropies with built-in upper limits do not.

Particular although they are, these results are strong enough to question the utility of most maximum entropy methods, at least at resolutions less than atomic. Indeed, the calculations used experimental as well as model diffraction data extending to 2.5 Å of a small structure (Prostaglandin), and were performed for different configurational entropies: Boltzmann's and those given in JN1 (equations 31 and 48) with realistic upper bounds (JN2 and references therein). Therefore, either information at a higher resolution or more informative configurational entropies seems to be necessary for an ab-initio solution of the phase problem.

Two important results indicated the lines of present and future work - it was found that rather simple minimization algorithms can give accurate atomic maps if phases are supplied to 3.5 - 3 Å resolution and moduli to 1.5 Å, when using non-local constraints, as shown in Fig. 1-2 (JN2). The problem is then to estimate low resolution phases.

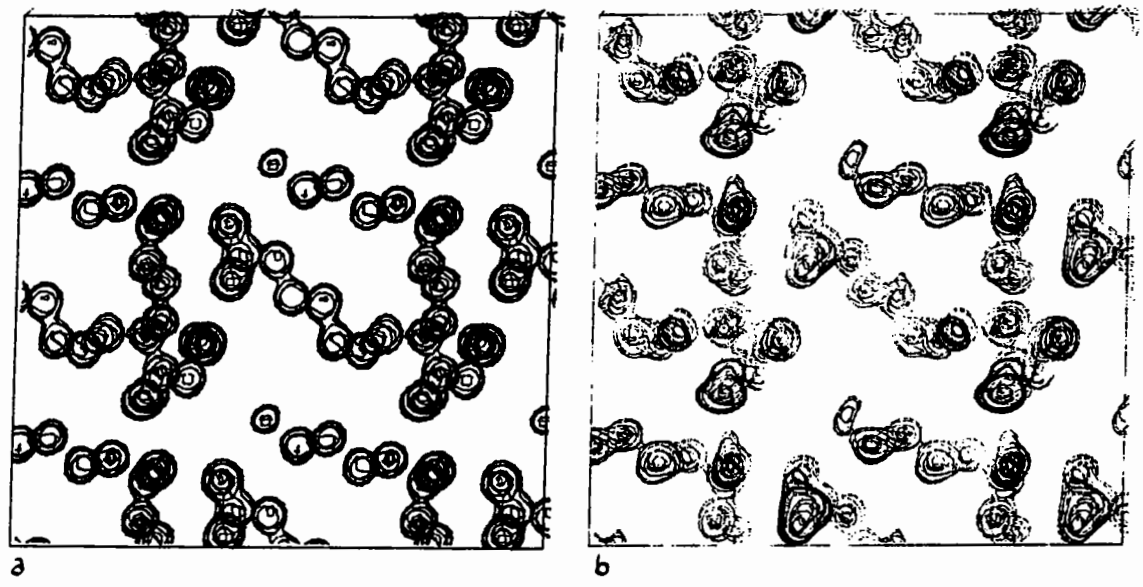


Fig.1. Prostaglandin. a) Fourier summation with observed amplitudes and model phases to 1 Å. b) Maximum entropy estimate of the electron density using model phases to 3.5 Å and observed amplitudes to 1 Å.

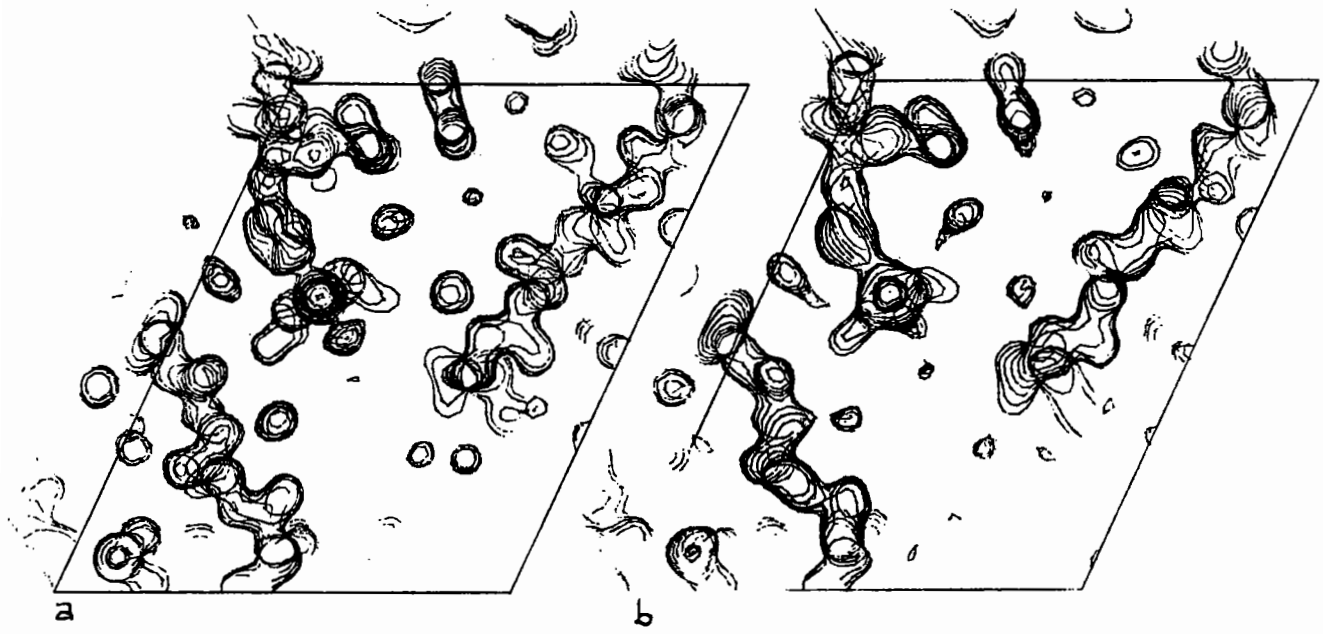


Fig. 2. Insulin. a) Fourier summation with observed amplitudes and model phases to 1.5 Å. b) Maximum entropy estimate of the electron density using model phases to 3 Å and observed amplitudes to 1.5 Å.

- At low resolution the information of atomicity is no longer valid, and has to be replaced by that of connectivity and binary model of the map (i.e. an almost two-level map). This can be achieved, within the formalism developed in JNl, by setting the upper limit of the admissible maps to an unrealistic low value. The convex problem becomes ill-conditioned and numerical instabilities appear because of the extremely severe constraints imposed on the sought map. It was however possible to obtain a 3 Å envelope of the Prostaglandin, as shown in fig. 3, from which the atoms were recovered by phase extension using non-local constraints.

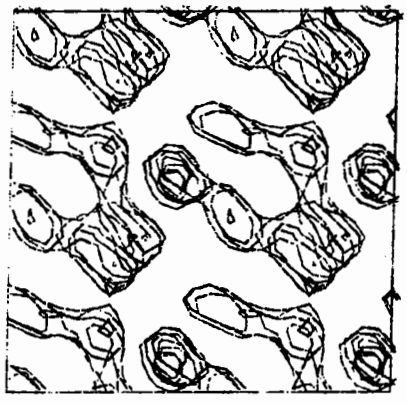


Fig. 3. Prostaglandin. a) Ab-initio maximum entropy estimate of the electron density using the observed amplitudes to 3 Å.

A less accurate numerical method was also successfully applied to model very low resolution electron densities (Podjarny, Moras, Navaza and Alzari, 1987; Alzari, Navaza, Podjarny and Poljak, 1987), but interaction with other classical methods was necessary to palliate for the lack of good programmes capable of satisfying the severe constraints.

The practical limitation of the method is that a rather accurate solution of the convex problem is needed, for any set of trial phases, in order to compute reliable phase shifts. To keep the dimension of the problem within reasonable limits, the existing programmes are being modified to exploit the space group symmetry via the Goedkoop matrices.

Fortunately enough, an almost analytical solution of the convex problem can be obtained when a tolerance in the fit of the observed amplitudes is allowed for. This is achieved by artificially extending the resolution of the diffraction data, making the hypothesis that the unobserved Fourier coefficients are zero. Only three fast Fourier transforms will suffice to exactly calculate the function to be minimized, its gradient and Hessian matrix. The above hypothesis is certainly a realistic one for small structures where high resolution data is currently available. In macromolecular structures, it may be a reasonable one too, since the extensively used methods of solvent flattening and density modification exactly correspond to a solution of the resulting minimization problem by a fixed-point algorithm. However, this formulation should give poorer results than those obtained with the preceding method where, according to the principles of information theory, unobserved information is not used.

The new equations can be readily applied to the traditional problems in crystal structure determination: ab-initio, phase refinement and phase extension.

2. ANALYTICAL SOLUTION OF THE CONVEX PROBLEM. CONNECTION WITH SOLVENT FLATTENING AND DENSITY MODIFICATION TECHNIQUES.

Notation:

$\rho(r)$	true electron density function
$F_{obs}(h) = \mid F_{obs}(h) \mid \exp[i\psi(h)]$	Fourier coeff. of $\rho(r)$
$m(r)$	admissible map
$F(h) = 1/V\int_{\Omega} m(r)\exp(2\pi ihr)dr$	Fourier coeff. or m(r). V is the volume of the unit cell Ω.

A	set of all admissible maps
P(m)	Probability law defined on the set A
$H(P) = -\int_A P(m) \ell n[P(m)] D_m$	entropy of P (summation over maps!)

We recall now the fundamental results of the theory (for details see JN1). Given the a priori information on the electron density function

$$\rho_{min}(r) < \rho(r) < \rho_{max}(r),$$

$$F_{obs}(h),\ h \in H,$$

the admissible set A is defined by

$$[m(r) \in A \text{ if } \rho_{min}(r) < m(r) < \rho_{max}(r)],$$

and the maximum entropy probability distribution of maps is (Z and u are normalization constants)

$$P_{me}(m) = 1/Z \exp [- \mu \Sigma_{h \in H} \lambda^*(h) F(h)] = (1/Z)\exp[-\mu/V \int_\Omega \chi(r)\, m(r)\, dr].$$

$\lambda(h)$ is the Lagrangian multiplier associated to $F_{obs}(h)$ and

$$\chi(r) = \Sigma_{h \in H} \lambda(h) \exp(-2\pi i h r).$$

The maximum entropy estimate of the electron density is the expected value of the admissible maps

$$\langle m(r) \rangle = \int_A P_{me}(m)\, m(r)\, Dm.$$

All expected values are explicit functions of the Lagrangian multipliers. In particular the configurational entropy S defined as the entropy of Pme,

$$S = H[Pme] = 1/V \int_\Omega s(\langle m \rangle)\, đr.$$

The Lagrangian multipliers are determined so as to satisfy the constaints (exact fitting of the data)

$$\langle F(h) \rangle = 1/V \int_\Omega \langle m(r) \rangle \exp(2\pi i h r)\, dr = F_{obs}(h),\ h \in H.$$

If a tolerance in the fit, controlled by the positive parameter t, is allowed for, and for given values of the trial phases corresponding to the observed amplitudes, the convex problem consists in minimizing the function

$$G = (1/2)\,\Sigma_{h\in H}\,|\,\langle F(h)\rangle - F_{obs}(h)\,|^{2} - tS$$

with respect to the Lagrangian multipliers λ. The conditions of minimum of G,

$$\langle F(h)\rangle - t\lambda(h) - F_{obs}(h) = 0,\ h\in H \qquad (1)$$

define the λ as implicit functions of the phases which, in turn, give the 'free energy' G as a function of the trial phases. The conditions of minimum of this function are

$$\text{phase}\,[\langle F(h)\rangle] = \text{phase}\,[F_{obs}(h)],\ h\in H \qquad (2)$$

When the prior information on the amplitudes is extended to the whole reciprocal space,

$$F_{obs}(h) = 0,\ h\notin H$$

the conditions of minimum (1) for the convex problem can be written as

$$\langle m(r)\rangle - t\,\chi(r) - \rho(r) = 0,\ r\in v, \qquad (3)$$

where

$$\rho(r) = \Sigma_{h\in H}\,|\,F_{obs}(h)\,|\,\exp[i\psi(h)]\exp(-2\pi ihr)$$

is the classical Fourier summation. For any given value of the positive parameter t, (3) defines <m> as a function of ρ, as shown in fig.4. This relationship is in fact the solution of the convex problem.

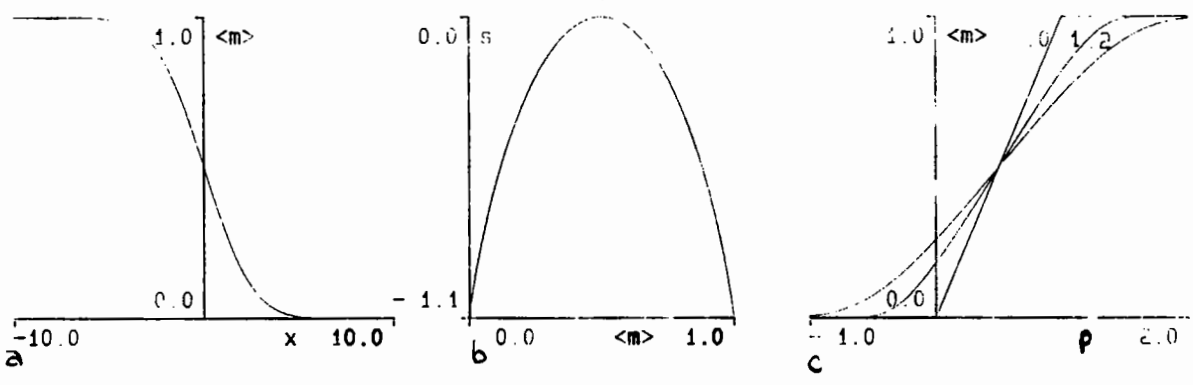


Fig. 4. Estimates. a) Maximum entropy estimate of the map <m> as a function of χ. b) Configurational entropy density s as a function of <m>. c) <m> as a function of ρ (eqn. 3) for different values of t.

A fixed-point algorithm aiming to satisfy (2), exactly corresponds to the density modification method of phase refinement and extension. However, it is well known that fixed-point algorithms, although expeditive, are less robust than any minimization procedure, in the sense that they get trapped in any stationary point close to the initial values of the variables.

The amount of computation needed per iteration on the phases is: one FFT to compute $\rho(r)$ from which we obtain $\chi(r)$, $\langle m(r) \rangle$, and the 'free energy' G. Another two FFT to obtain the Lagrangian multipliers from which the gradient and the Hessian of G are calculated.

3. REFERENCES

J. Navaza (1985) Acta Cryst. A41, 232

J. Navaza (1986) Acta Cryst. A42, 212

A.D. Podjarny, D. Moras, J. Navaza and P. Alzari (1988) Acta Cryst. A44, 545.

A.P. Alzari, J. Navaza, A.D. Podjarny and R. Poljak (1987) communication to the Congress of the IUC, Perth, Australia

Some Assorted Maximum Entropy Calculations

Richard K. Bryan
European Molecular Biology Laboratory,
Meyerhofstrasse 1, 6900 Heidelberg, West Germany.
Tel. 06221/387365. Earn/Bitnet BRYAN@EMBL

1. Maximum entropy.

The Maxent solution is that map which maximises the Shannon-Jaynes entropy over the set of maps agreeing with the experimental data (Gull & Daniell, 1978). The entropy (Jaynes, 1968) is defined on the suitably discretised density ρ as

$$S(\rho) = -\sum_j p_j \log p_j/m_j, \qquad p_j = \rho_j/\sum \rho, \tag{1}$$

where $\mathbf{m}$ is the normalised prior map, the estimate of the solution before the data are considered. $\mathbf{p} = \mathbf{m}$ has the global unconstrained entropy maximum. Agreement with the data, to within the noise level, is measured by an appropriate statistical test. All native and derivative data may be included with the correct weights by using the expression

$$C(\rho; I^n, I^{d_i}, F^p) = \sum_{\mathbf{h}} \Big\{ \quad w_{\mathbf{h}}^n \big(|F_{\mathbf{h}}|^2 - I_{\mathbf{h}}^n\big)^2 \tag{2a}$$

$$+ \sum_i w_{\mathbf{h}}^{d_i} \big(|F_{\mathbf{h}} + H_{\mathbf{h}}^{d_i}|^2 - I_{\mathbf{h}}^{d_i}\big)^2 \tag{2b}$$

$$+ w_{\mathbf{h}}^p |F_{\mathbf{h}} - F_{\mathbf{h}}^p|^2 \quad \Big\}, \tag{2c}$$

where $I_{\mathbf{h}}^n$ are the observed native intensities, $I_{\mathbf{h}}^{d_i}$ the observed intensities for the i^{th} derivative, weighted by $w_{\mathbf{h}}^n$ and $w_{\mathbf{h}}^{d_i}$ (usually inverse variances) respectively, $H_{\mathbf{h}}^{d_i}$ the transform of heavy atom contribution to the i^{th} derivative, and the $F_{\mathbf{h}}^p$ phased data, included to take account of reflections which can be phased reliably by conventional isomorphous replacement, such as some centrics when a single isomorphous derivative is used. To include anomalous difference data, a separate term is used in (2b) for each member of the Bijvoet pair. For a large number M of observations, agreement is achieved when $C \leq M$. This condition means that the differences between the observed and calculated quantities can be attributed solely to noise. Note that if the $F_{\mathbf{h}}$ fit the native data exactly, (2b) is the same as the expression of Hendrickson & Lattman (1970) for the phase probability (more correctly, likelihood) distribution.

One difficulty remains, as F_{000} cannot be measured directly, a problem which also occurs in, say, radio interferometry. The total density has a special status in the entropy expression, and unless constrained, $\sum \rho$ may increase without bound, and hence, for a given level of fluctuation on the map, the entropy as well. However, taking uniform $\mathbf{m}$ for simplicity,

$$S = \log \sum \rho - \frac{1}{\sum \rho} \sum \rho \log \rho, \qquad \text{so} \qquad \nabla S = \frac{1}{\sum \rho} \left(\frac{\sum \rho \log \rho}{\sum \rho} - \log \rho \right), \tag{3}$$

showing that it effectively measures the deviation of ρ from its weighted geometric mean. A fixed value A can be assigned to this background, or 'default', level, equivalent to using a modified entropy (Skilling & Bryan, 1984)

$$S = -\sum \rho \log \rho / eA. \tag{4}$$

In the following computations, the entropy maximisation algorithm (Skilling & Bryan, 1984), modified for non-convex problems (Bryan & Skilling, 1986), with an additionally enhanced procedure for finding descent directions of C with negative curvature, was used.

2. Applications.

It is obvious that the above formalism may be applied to many possible combinations of data and prior maps, ranging from the reduction of truncation error when reliable phases are already available *via* phase extension to the much harder *ab initio* phase problem, with a corresponding variation in the difficulty of calculating an appropriate solution due to possible ambiguities in the latter case. Structure factor data give a unique solution, and several examples of such calculations have been presented (*e.g.*, Bricogne, 1984, Wei, 1985, Wilkins & Stuart, 1986), showing the increase in map quality expected in any application of Maxent to linear problems (Gull & Daniell, 1978). We will first focus on the SIR problem, which is somewhat more tractable than the full *ab initio* problem.

SIR data are conventionally (Blow & Rossmann, 1961) used to compute the best structure factors (which, with perfect data, will be the average of the two possible ones), so if the possible phases are very different, the amplitude is weighting down strongly. The resulting map may sometimes be interpretable, or ameanable to improvement by solvent flattening or similar methods. If the data are used directly in the constraint function (2a and 2b), a suitably small value of C will only be obtained near either of the two 'most probable' phases, with the full, correct, native amplitude. Thus, in contrast to the usual use of SIR data, Maxent will select between the phases allowed by the data. In a sense, the difficulty of this problem is comparable with the small-molecule centrosymmetric problem (see also Gull *et. al.*, 1987), provided our data are sufficiently accurate. Centrics do pose a special problem if used directly in the intensity constraint function (2a), as there are two independent regions of good fit to the data, separated by a large peak, which often prevents the calculated phase changing sign during refinement. A complex structure factor can change its phase at constant amplitude, and no such problem occurs if only native intensity data are provided. SIR data, with (usually) two wells, are intermediate in character, as the barrier between is not usually so great as for centrics. Therefore, during the following calculations, centrics were either included with their correct sign, or omitted altogether.

As an example of this sort of calculation, a problem with a centrosymmetric heavy atom distribution will be considered (Bryan, 1988a). The 'best' map would be a superposition of the true map and its enantiomorph, and hence uninterpretable. Synthetic data to 3Å resolution for this problem were calculated for a 20 amino acid protein fragment in the asymmetric unit of a $P2_12_12_1$ unit cell, $a = b = 24$, $c = 64$Å, and a zinc derivative thereof (Bryan & Banner, 1987). Part of the density synthesised from the native structure factors is shown in fig. 1 d. The heavy atom parameters were refined in order to emulate the real situation, where they would be estimated initially from a difference Patterson map. All the heavy atom structure factors are now real, as are the best phases, so the best SIR map (fig. 1a) exhibits the *Pbca* symmetry of the heavy atom array. One must be careful when applying Maxent to this problem, in order that phases are chosen consistently for one enantiomorph. However, the mechanism is already available for introducing the data in a controlled way, by calculating a map from some of the data, and using this map as a prior when introducing more data. First, one uses the data with known phases, in this case the real centrics, plus one further reflection to which one can assign a phase to fix the enantiomorph. This was chosen to be the largest remaining reflection, which, as it happened, was the imaginary centric 012. Maxent is used to find a map fitting these data alone (fig. 1b), which is then used as a prior map $\mathbf{m}$ for a further calculation against all the data. This prior map is non-centrosymmetric, and gives predictions for the rest of the structure factors. Thus, the second part of the calculation is started off with the phases biased towards one particular enantiomorph. The resulting map (fig. 1c) is again very like the original, but with an average amplitude-weighted phase error of around 19°. The small amount of incorrect density is much weaker than the correct. Bryan & Banner (1987) contains calculations for the slightly easier case of a non-centrosymmetric heavy atom distribution, and also investigates the robustness of the method with respect to noise, and Bryan *et. al.* (1983) solves an unknown structure from fibre diffraction data with a single derivative.

So far, we have assumed that the heavy atom structure factors are known, *e.g.*, by solving the difference Patterson. Can we avoid this step? Perhaps, by computing two maps simultaneously, one, as before, representing the native density, and the other, the heavy atom density, whose transform, $H_{\mathbf{h}}$, replaces the previously fixed value in (2b). The total entropy of both maps is maximised, with the constraint ensuring that the transform of the first map agrees with the native data, and the sum of the transforms fits the derivative data. With perfect data, this method has been found to work in practice (Bryan, 1988a), recovering the heavy atom density at the correct location. It can also be shown (Bryan, 1988b) that such a method is related to the combined direct methods/isomorphous replacement formalism of Hauptman (1982). However, analysing both native and heavy atom maps from the point of view of *a priori* uniformity means that the very important knowledge that the heavy atom distribution consists of a few distinct atoms is ignored. Indeed, if the data are of sufficient quality for these procedures to be applicable, it is very likely that the difference Patterson can be solved for the heavy atom positions anyway, making the full two-map treatment unnecessarily complicated.

In assessing the use of Maxent in the *ab initio* phase problem, we need to know whether the solution really is likely to be correct, irrespective of the difficulty of actually computing it. Is the constraint of Maxent sufficient to compensate for the loss of phase information, or, in other words, is the correct solution close to *some* local entropy maximum when only native intensity constraints are used? This can be investigated in a simulated calculation (Bryan & Banner, 1987), where the correct phasing is already known, so that the calculation can be started near a suitable maximum, and hence the possible existence of other local maxima does not have to be examined. Here, we start from the atomic structure of scorpion neurotoxin 3 (Almassy *et. al.*, 1983), obtained from the Brookhaven database. This protein has 65 residues, and refined positions of an additional 72 solvent molecules are included. Firstly, from synthetic 3Å structure factors, a Maxent map

was calculated (fig. 2 a), using constraint (2c). The phases of the data were then forgotten, and the intensity contraint (2a) used. A path of monotonically increasing entropy at constant fit to the intensity data was followed from the initial (correctly phased) map until a maximum was attained (fig. 2b). We cannot be certain that this was the nearest maximum, however, only that the initial map was on a slope leading to this peak. The phases shifted by an amplitude weighted average of some 50°, and the density became broken, with a very inhomogeneous distribution of peak heights. Clearly, Maxent does not fully compensate for the loss of phase information, at least at this resolution.

This result suggests that the underlying assumptions are inadequate, *i.e.*, there is more to molecular structure than an *ab initio* random, uniform and independent distribution of atoms - which we knew anyway, but nevertheless remarkable progress in direct methods has been made using it. One or more of these assumptions must be dropped in order to make further progress, and I would suggest that the first candidate is independence, as some of the strongest model-building constraints we have are those of atomic bond lengths and angles. It may also explain the difficulty in extending direct methods to larger structures, as dependence of atomic positions becomes more and more significant. Moreover, the value of the entropy is unchanged if the pixels are shuffled in position, although the map would obviously change from the sort of smooth density expected into uninterpretable apparent noise. Although complete independence between adjacent pixels may be useful in some application, *e.g.*, spectroscopy or astronomy, where isolated points could occur in reconstructed maps, it is a disadvantage in crystallography as we know exactly what sort of density to expect. Such knowledge allows one almost immediately to classify a density as a protein density or not, but is very hard to find a function of the map which will do such classification for us.

3. Incorporating correlations via a second order prior.

We are thus naturally lead to consider the question of incorporating prior information, in terms of molecular structures, directly into the phasing process. Translation invariance means that in the first place we must use a flat initial model $\mathbf{m}$. Perhaps if we have low resolution phase information, solvent regions can be identified and flattened, or with poor phase information, some of the structure, such as the backbone, can be picked out visually. An atomic model of this part of the structure can be built, and used as $\mathbf{m}$ in a further Maxent calculation. However, this will not always be possible, particularly in the *ab initio* phase problem. The information we wish to use is in terms of atomic bond lengths and angles, or possibly lower resolution features, which means that it is in the form of positional correlations of densities. A mechanism for dealing with such information was proposed by Skilling (1986). The idea is to work in the space of N samples from the map, and to apply the entropy to the N-sample joint distribution. It was shown that this is equivalent to working in the 1-sample space, but with an effective prior which depended on the current map, and was illustrated with a simple example using the position-independent information that a star (point source) was present in the observed object.

Clearly, the prior information in molecular structures extends to very high orders of correlation, and will be very difficult to encode. We shall therefore start most simply, and illustrate the idea with 2-point correlations, although this is clearly of limited practical use, since its effect is not very different from manipulating the inner part of the Patterson, but it is applied in quite a different way.

Following Skilling (1986), the 2-sample entropy is defined on the 2-sample distribution p_{ij} as

$$S^{(2)} = -\sum_{ij} p_{ij} \log\bigl(p_{ij}/m_{ij}\bigr). \tag{5}$$

Since we want m_{ij} to represent the correlation of positions, it must be a function of $|i-j|$ only, and we set $p_{ij} = p_i p_j$ to obtain a representation of the solution as a 1-sample density. Hence

$$S^{(2)} = -\sum_{ij} p_i p_j \bigl(\log p_i p_j / m_{i-j}\bigr) \quad = \quad -2\sum_{i} p_i \bigl(\log p_i - \tfrac{1}{2}(p * \log m)_i\bigr), \tag{6}$$

and by comparison with (1), the expression

$$m_{\text{eff}} = \exp\bigl(\tfrac{1}{2} p * \log m\bigr) \tag{7}$$

can be identified as the 'effective prior' in the 1-sample space.

To test this method, a model molecule was constructed in one dimension on a 256 pixel array as groups of Gaussian atoms, with uniform spacing of 6 between atoms within a group, but no positional correlation between groups (fig. 3 a). Fourier coefficients to about the 40^{th} order are required to resolve such atoms classically. Using Maxent on the Fourier coefficients to the 18^{th} order and a flat prior gave fig. 3b, where the atoms are obviously not resolved, but the positions of the groups are revealed. The peaks in this map do not correspond to atomic positions. The second order prior was taken as the correlation function of a

regular array of such atoms, but set constant at distances from the centre greater than that of the second minimum (fig. 3c). The programming was fairly crude, in that the effective prior was recalculated once per iteration, rather than using the full derivatives of $S^{(2)}$, so that the existing 1-sample entropy maximisation program could be used with minimal changes. Fig. 3d shows the result, and fig. 3e the final effective prior. The solution consists of sets of correctly-spaced spikes within the envelope defined by the data. When many atoms are in contact, they strongly reinforce each other through the prior. The singles and doubles are much less affected, except where they are close to a larger group, when the relative positioning may even cause unfavourable interactions. Beyond the edges of groups, there are further ripples in the prior, which cause a smaller peak in the map, but only to the amount allowed by the data.

The same formalism can be applied to higher orders of correlation, but as yet even the required third-order prior has not been established; to do so may not be a trivial exercise.

References.

Almassy, R. J., Fontecilla-Camps, J. C., Suddath, F. L. & Bugg, C. E. (1983). Structure of variant-3 scorpion neurotoxin from centruroides sculpturatus ewing, refined at 1.8Å angstroms resolution.. *J. Mol. Biol.*, **170**, 497–522.

Blow, D. M., & Rossmann, M. G. (1961). The single isomorphous replacement method. *Acta Cryst.*, **14**, 1195–1202. Correction. *Acta Cryst.*, **15**, 1060.

Bricogne, G. (1984). Maximum Entropy and the Foundations of Direct Methods. *Acta Cryst.*, **A40**, 410–445.

Bryan, R. K. (1988a). The Maximum Entropy Method Applied to Intensity Data. *Scanning Microscopy Suppl.*, **In press**, –.

Bryan, R. K. (1988b). A Maximum Entropy Derivation of the Integrated Direct Methods -Isomorphous Replacement or -Anomalous Scattering Probability Distributions.. *Acta Cryst.*, **Submitted**, –.

Bryan, R. K. & Banner, D. W. (1987). Maximum entropy calculation of electron density with native and single isomorphous replacement data. *Acta Cryst.*, **A43**, 556–564.

Bryan, R. K., Bansal, M., Folkhard, W., Nave, C. & Marvin, D. A. (1983). Maximum entropy calculation of the electron density at 4Å resolution of Pf1 filamentous bacteriophage. *Proc. Natl. Acad. Sci. USA*, **80**, 4728–4731.

Bryan, R. K. & Skilling, J. (1986). Maximum entropy image reconstruction from phaseless fourier data. *Optica Acta*, **33**, 287–299.

Gull, S. F. & Daniell, G. J. (1978). Image reconstruction from incomplete and noisy data. *Nature*, **272**, 686–690.

Gull, S. F., Livesey, A. K. & Sivia, D. S. (1987). Maximum entropy solution of a small centrosymmetric crystal structure. *Acta Cryst.*, **A43**, 112–117.

Hauptman, H. (1982). On integrating the techniques of direct methods and isomorphous replacement I. The theoretical basis. *Acta Cryst.*, **A38**, 289–294.

Hendrickson, W. A. & Lattman, E. E. (1970). Representation of phase probability distributions for simplified combination of independent phase information. *Acta Cryst.*, **B26**, 136–143.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Trans.*, **SCC-4**, 227–241.

Skilling, J. (1986). Theory of Maximum Entropy Image Reconstruction. *In* Maximum Entropy and Bayesian Methods in Applied Statistics, Proceedings of the Fourth Maximum Entropy Workshop, University of Calgary, 1984, ed. James H. Justice, pp. 156–178. Cambridge: Cambridge University Press.

Skilling, J. & Bryan, R. K. (1984). Maximum entropy image reconstruction: general algorithm. *Mon. Not. R. astr. Soc.*, **211**, 111–124.

Wei, W. (1985). Application of the maximum entropy method to electron density determination. *J. Appl. Cryst.*, **18**, 442–445.

Wilkins, S. W. & Stuart, D. (1986). Statistical geometry. IV. Maximum-Entropy-Based Extension of Multiple Isomorphously Phased X-ray Data to 4 Å Resolution for α-Lactalbumin. *Acta Cryst.*, **A42**, 197–202.

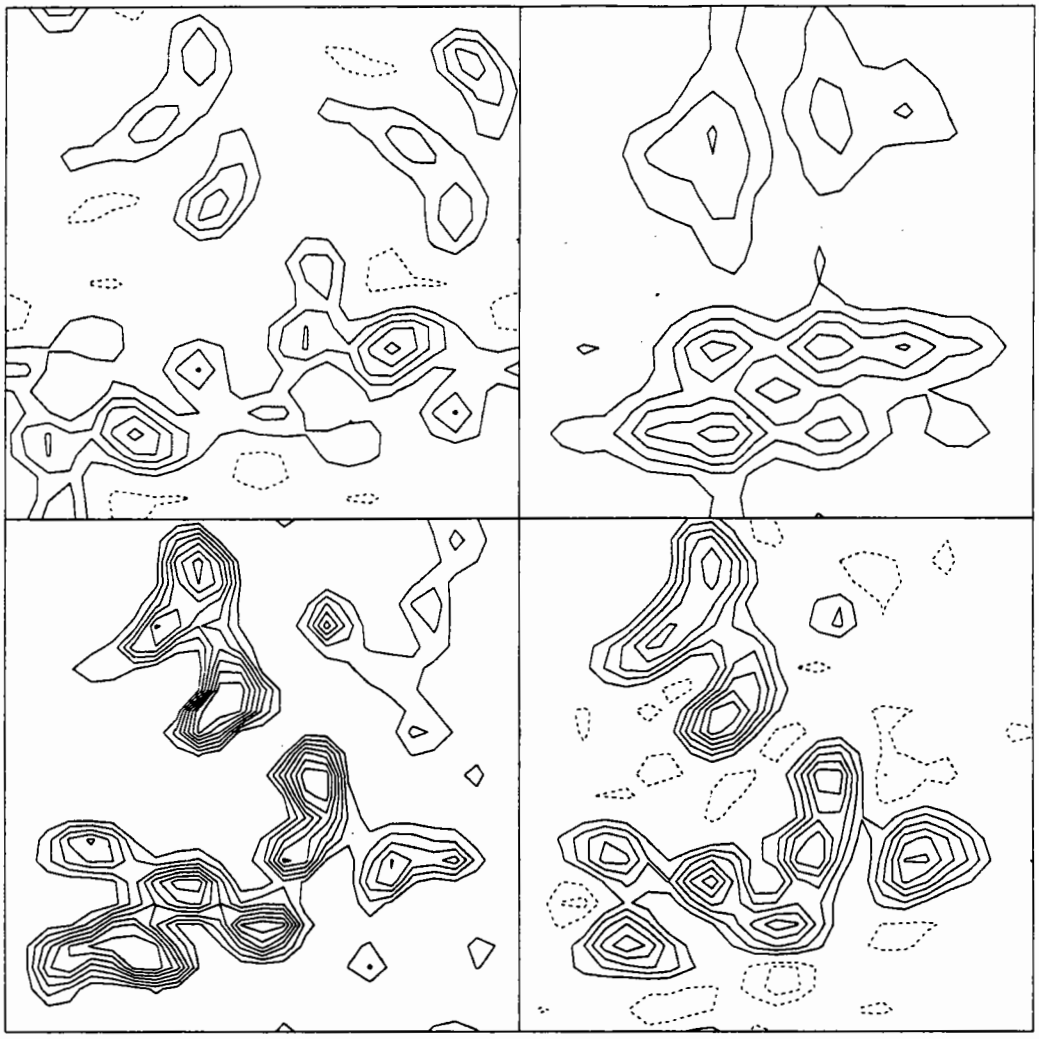


Figure 1. Density from SIR data with a centrosymmetric heavy atom. a. best map; b. Maxent map from real centrics and one enantiomorph defining structure factor; c. final Maxent map; d. Fourier map from correct structure factors.

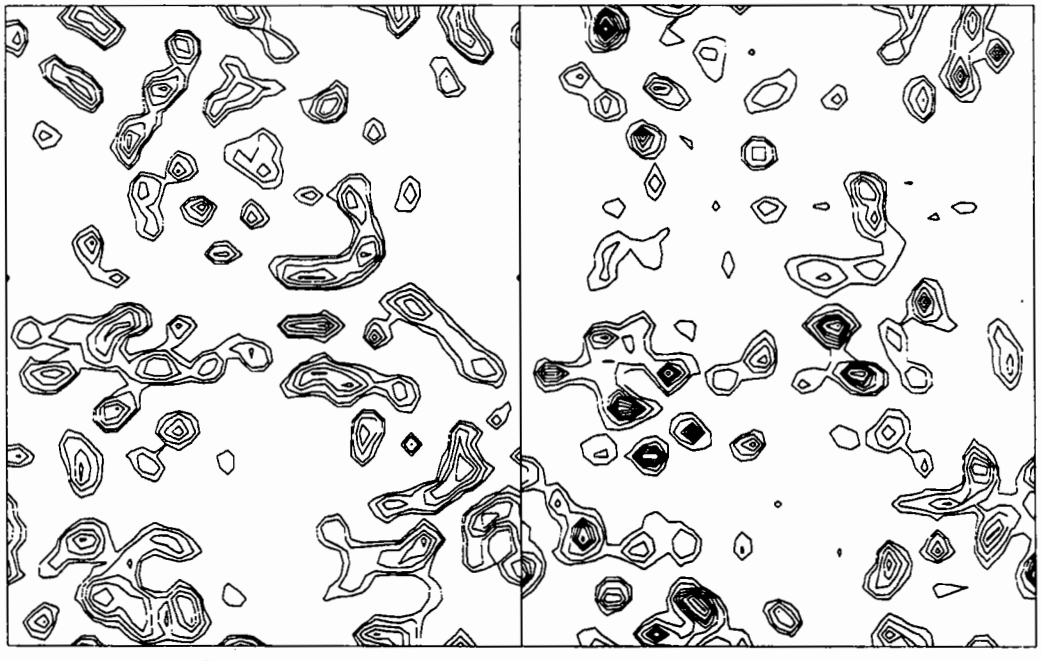


Figure 2. Sections of neurotoxin density. a. Maxent density from calculated structure factors. b. Maxent density from intensities, starting from a.

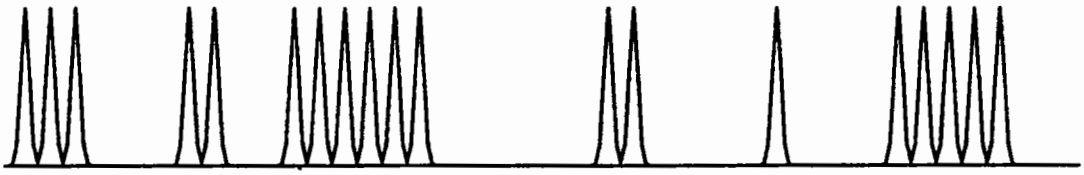
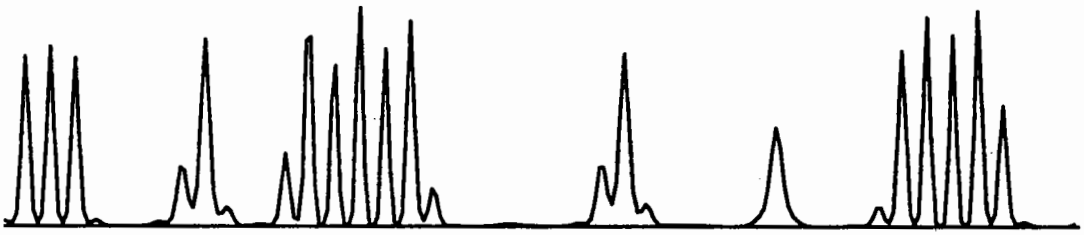


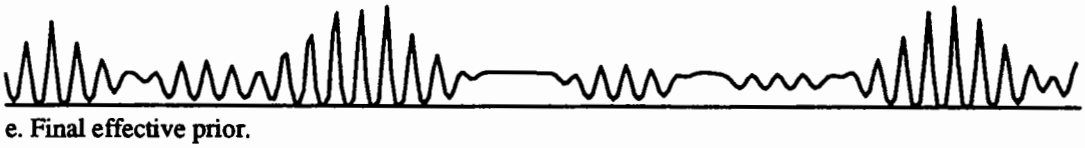
Figure 3. a. Original simulated density.

b. Maxent reconstruction from 18 lowest resolution structure factors.

c. Correlation function for second-order prior.



d. Reconstruction using second order prior in addition to 18 structure factors.



e. Final effective prior.

A PRACTICAL GUIDE TO THE USE OF PARTIAL STRUCTURAL PHASE COMBINATION

by

D.W. RICE*, B.F. ANDERSON‡, and E.N. BAKER‡

* Department of Biochemistry, Sheffield University, Sheffield S10 2TN U.K.

‡ Department of Chemistry and Biochemistry, Massey University, Palmerston North New Zealand.

1. INTRODUCTION

The procedure by which phases derived by isomorphous replacement can be improved by the combination of phase information from a gradually improving molecular model has now received widespread attention and use in the field of protein crystallography (1,2). The technique is most appropriate when dealing with structure determinations at medium resolution where the initial electron density map is inadequate for a complete interpretation of the molecular structure. The last few years have seen the methods of solvent flattening successfully applied in structure analysis to enhance initial isomorphous maps. However, in many situations, there still remains a clear need for further enhancements over and above those which can be obtained using these methods. In such cases the technique of partial structure phase combination offers greater potential for obtaining the map enhancements necessary for the structure solution. The purpose of this paper is to provide a practical guide to the use of this approach and particular emphasis will be placed on describing the optimum strategy for structure refinement incorporating this method of phase enhancement.

2. THE NEED FOR PHASE IMPROVEMENT

The aim of a protein structure determination is to be able to provide a reliable model of the atomic arrangement within the molecule. From this model deductions can then be made about the interactions which govern its stability and relate to its function. The reliability of these deductions is then directly related to the quality of the electron density map from which the model was built. The starting point for the models of most protein structures is usually obtained by the inspection of a medium resolution electron density map (3.0-2.5A), whose phases are calculated by the method of multiple isomorphous replacement (3). These phases are subject to errors which arise from a combination of a number of factors which may include non-isomorphism of the derivatives, low occupancy of the heavy atoms, poor heavy atom refinement and errors in the measurements of the X-ray intensity data. The normal analysis of these errors lead to the assignment of a figure of merit to each reflection (4), this factor expressing the precision of the phase determination. The result of these phase angle errors and the figure of merit weighting is to degrade the quality of the resultant electron density image with noise and to lower its apparent resolution. Therefore the interpretation of electron density maps based solely on isomorphous replacement phases is a procedure often fraught with difficulties. The resolution of the electron density image is usually blurred, features such as side chains are poorly defined and the bulges in the electron density associated with the carbonyl groups, which serve to guide the orientation of successive peptides, are absent. Moreover, the

electron density may be severely disturbed in the vicinity of the heavy atom binding sites and areas in the molecular structure where there is either high thermal motion or static disorder, are often characterized by weak and broken electron density leading to errors in chain connectivity, particularly when no amino acid sequence information is available.

The isomorphous phases themselves may be greatly improved in systems where the possession of non crystallographic symmetry allows techniques of molecular averaging to be applied. In other cases too, the application of solvent flattening procedures has undoubtedly produced very useful improvements to isomorphously phased electron density maps. Nevertheless the technique of partial structure phase combination has not been made redundant by these improvements. Further enhancements are invariably necessary to complete the description of the molecular structure. Whilst in some cases the quality of the solvent flattened map may be sufficiently good to provide an interpretation which can then be refined using high resolution data (2.0A data), this is neither universally true, since the maps may not be that good, nor possible since the high angle data may not be accessible.

3. REFINEMENT PROTOCOL INCORPORATING PARTIAL STRUCTURE PHASE COMBINATION

A flow chart illustrating a typical medium resolution refinement is illustrated in fig. 1. Each of the elements of the diagram is very straightforward but the practical aspects of a number of the steps will now be outlined.

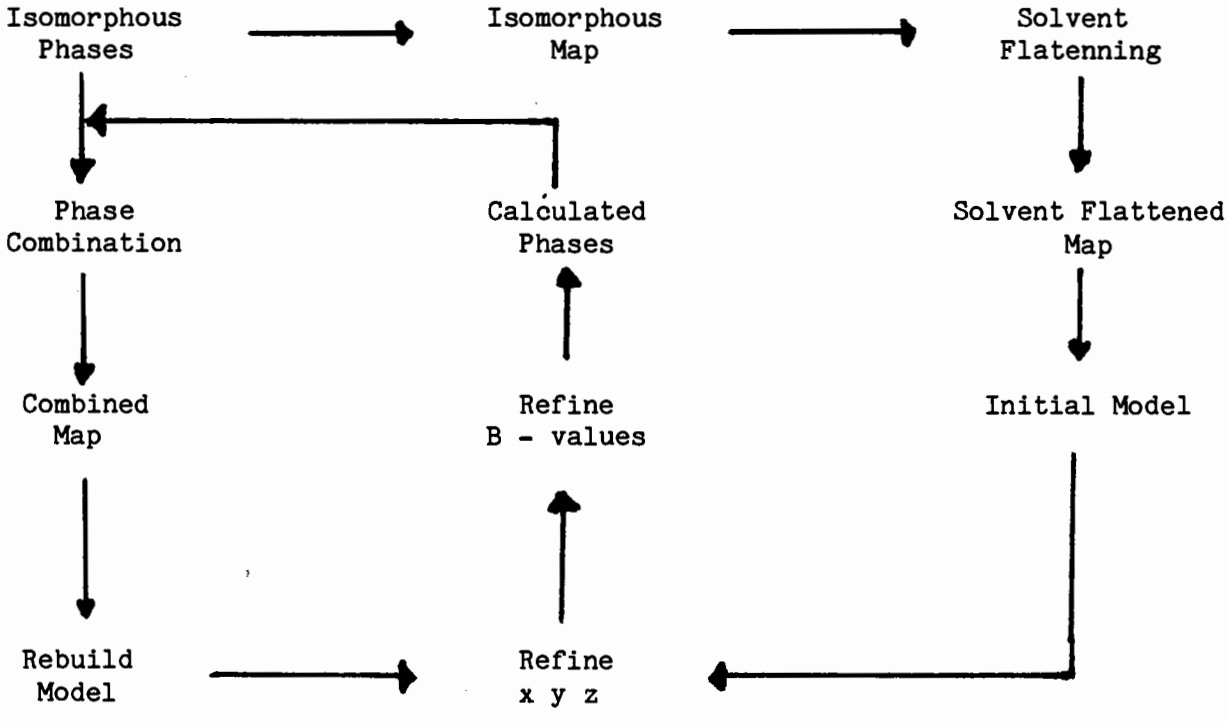


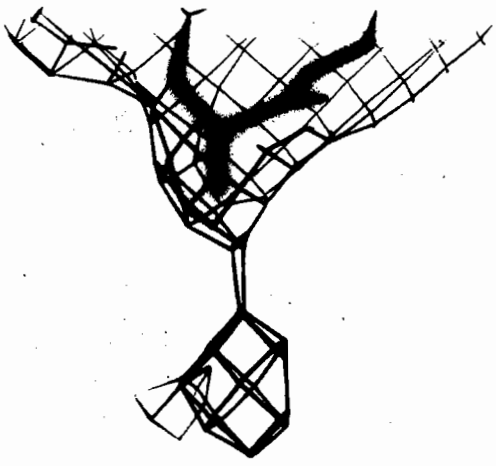
Figure 1 Flow diagram of a refinement protocol incorporating partial structure phase combination.

3.1 **The initial model**

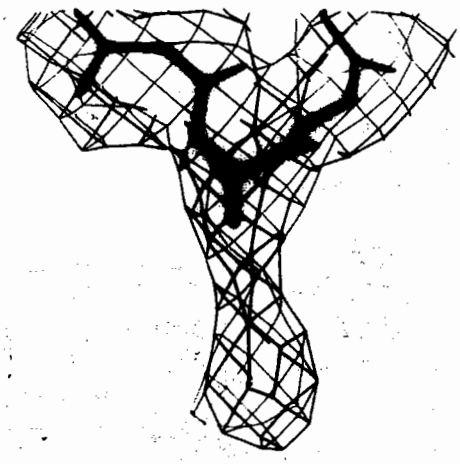
The obvious worry when dealing with the refinement of any structure, no matter at what resolution, is the extent to which the electron density maps used for the process of reinterpretation are biased towards the input model. If this bias is excessive, then errors in the model may go unnoticed with potentially disastrous consequences. Whilst this concern remains true for all refinements, it is particularly the case for refinements at medium resolution (d > 2.5A) where the parameters to observation ratio can lead to artificially low R factors. It is even more the case when dealing with a refinement where the initial model is based on a poor electron density map, and in this situation, the proportion of the correct structure, may be so low that excessive bias is guaranteed.

Figure 2 The gradual improvement in electron density around residue LYS19 in lactoferrin. Figure 2a depicts the situation in the initial solvent-flattened map; figure 2b and c at the stages COMB1 and COMB2 respectively (see Table 1 for details of the refinement of this structure).

a



b



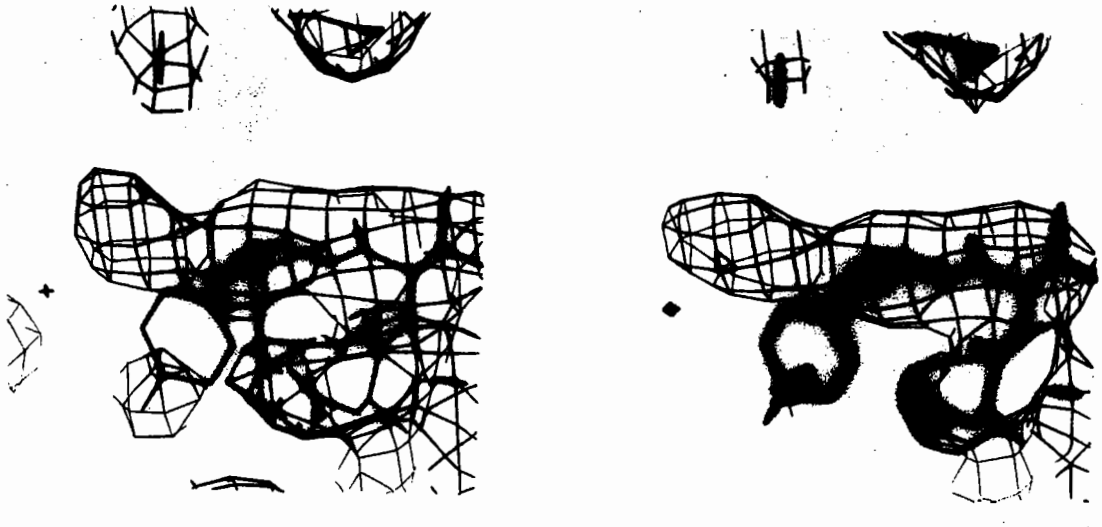
c

In adopting a strategy for refinement incorporating the calculated phases in conjunction with those derived by isomorphous replacement, the bias in the calculated component must be diluted. However, given the concern over bias, it is far safer to adopt a defensive strategy and allow the maps to improve gradually. In practical terms this means that the initial interpretation should be an underestimate of the electron density map rather than an over-optimistic interpretation. So, for example, where sidechains are clearly defined they can be safely included, but poorly defined sidechains should be included at a much reduced level. In practice this might mean a representation of a sidechain as alanine if some limited evidence for it is clear, but glycine if the definition is very poor. Whilst these statements can apply to regions of very regulate structure (alpha helix and beta sheet), they apply most strongly to the loops, and the initial interpretation of these parts of the structure, will often be poly ala/gly. The clear rationale behind this approach is to produce in successive rounds of the refinement/phase combination procedure, a gradually improving image of the molecule. At each stage, once a sidechain can clearly be seen, it is then safe to incorporate it. An example of this is show in fig. 2. This strategy is almost certainly preferable to one which is based on over-optimistic interpretation where, when electron density subsequently improves around the input sidechains, there remains a doubt as to whether what is seen is the truth or simply bias.

3.2 **Bias checks**

As well as adopting a cautious approach to the initial interpretation, another procedure which has been found to be exceedingly useful is that of incorporating deliberate mistakes into the model. Thus, the combined electron density map can be rapidly assessed for the degree of bias by scanning these regions alone. Ideally, the bias checks should fall into 2 distinct categories. The first of these involves well resolved regions of the electron density map, where sidechains can be unambiguously defined. Thus a well defined sidechain, often an aromatic residue, is deliberately misplaced from its electron density and is allowed to refine in this wrong position. The subsequent electron density map can then be examined to see if the density corresponds to the true or false position of the sidechain. An example of this is shown in fig. 3.

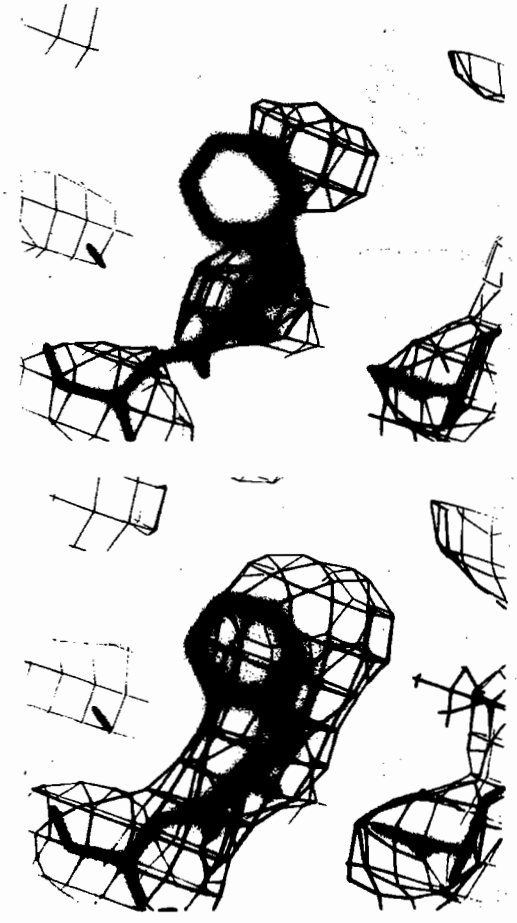
Figure 3 Bias checks in a well-phased region of the lactoferrin map is illustrated here. Figure 3a represents the deliberate misplacement of tyrosine 66 from its virtually unambiguous positions in the solvent-flattened map. Figure 3b then illustrates the electron density in the map based on COMB1 phases.



a b

However, the effects of bias are naturally likely to be low in well defined regions of the electron density map, since these are inherently encoded in the isomorphous phases. A class of bias checks should therefore also be included which address the problems of bias in more poorly defined regions. The key to this class, is to identify unconnected regions of density near to a piece of regular secondary structure. For example, the regular distribution of sidechains on an alpha helix fixes their relative positions. Thus, given one or two good sidechains nearby, an unconnected piece of density near an alpha carbon atom can often be assigned to a sidechain, even though the connection is broken. This sidechain can then be misplaced and its appearance in subsequent combine maps noted. An example of this type can be found in found in fig. 4.

Figure 4 Bias checking in a weakly phased region of the lactoferrin map is illustrated here. Figure 4a represents the density close to the alpha carbon of phenylalanine 557 in the initial map, whilst figure 4b shows the electron density in the subsequent combined map (COMB1). The sidechain can clearly be seen to be misplaced.



a

b

Whilst the bias that has been detected in combine maps is generally low, there can be no doubt that the introduction of bias checks, such as those described here, greatly increases the confidence with which map interpretations may be made. Although these areas of deliberately incorrect structure may offend the eye and undoubtedly lead to other local distortions in the structure, it is recommended that their correction is left until the latter stages of the refinement, since they can be fixed quite easily and their value as monitors of the reliability of the electron density maps, is immense.

3.3 **Progress of the refinement**

The phase combination procedure allows the structure to be refined in a controlled manner with a continuous check on the level of bias in the maps. As the refinement proceeds, and the R value decreases the interpretation of the structure is placed on a firmer footing, and the percentage of the atoms in the structure, gradually increases. The appearance of new information in the maps gradually increases, and the improvements to the structure are made in such a way that confident interpretations can be made. Table 1 illustrates the progress of the medium resolution refinement of lactoferrin which proceeded using partial structure phase combination technique.

Table 1

Progress of the lactoferrin refinement

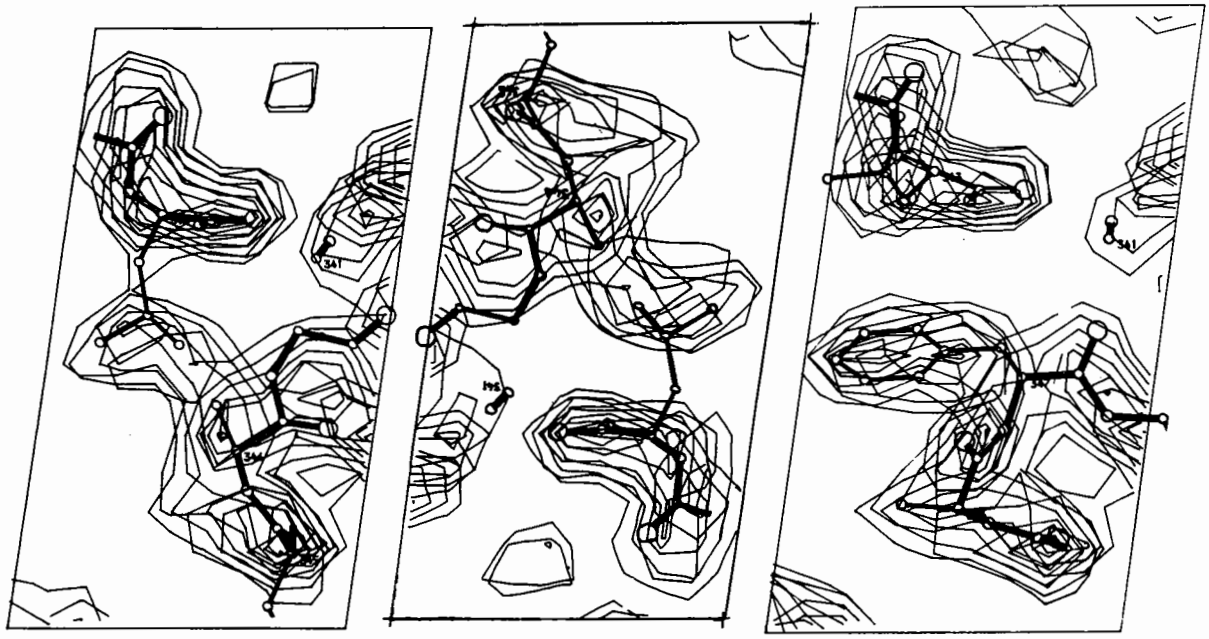
Phase Set	No. of Atoms (5380 in total)	R factor
ISO/WANG	-	45.9
COMB1	4690	33.0
COMB2	4929	26.6
COMB3	5013	25.5
COMB4	5126	24.9
COMB5	5164	20.4

3.4 **Relative weighting of the isomorphous and calculated phases**

The weighting scheme used in the work presented here has been described previously (1). The isomorphous phase probabilities were derived by the treatment of Blow and Crick (4) and the phase probabilities from the partial structure determined using the probability analysis of Sim (5). The estimate of the missing structure for use in Sim's formula was obtained empirically from the lack of closure between Iobs and Icalc in ranges of resolution following the manner of Bricogne (6). The combination of the isomorphous and partial structure phase information was then achieved by multiplying the individual phase probability functions together (7).

As well as the type of weighting scheme used in the phase combination, the type of synthesis utilised at each round of refinement may be varied. In the early stages when the R factor is high the most appropriate synthesis is a fourier synthesis based on coefficients $\mathbf{1}Fobs\mathbf{1}m.\alpha$ comb. However, as the R value falls to somewhere in the region of .3-.35 this synthesis is surpassed by one such as $2\mathbf{1}Fobs\mathbf{1}-\mathbf{1}Fcalc\mathbf{1}m.\ \alpha comb$ on a variant thereon (fig. 5). The reason for not using this synthesis straightaway, is the extra noise which appears due to the inclusion of the additional difference term when the R factor is poor. In practice, the optimum synthesis can be discerned by examining the various syntheses on the graphics, comparing the features around the bias check regions, and this empirical approach, is very convenient. The above weighting schemes certainly can be improved, and further discussion of these will be found in other papers in the proceedings.

Figure 5 A change in the type of synthesis used to compute the electron density map can have an effect not only on the level of bias but also on the quality of the new information brought back. Figure 5a shows a wrongly placed leucine sidechain discovered in the refinement of phosphoglycerate kinase in a combined map base on an Fobs synthesis (R = .3). The equivalent region based on a 2 Fobs - Fcalc synthesis with the same phases, is shown in figure 5b. The effects here appear subtle but more dramatic effects have been noted. Finally, figure 5c shows the correct interpretation which involves extensive re-modelling and replacement of the alanine residue at position 347 by a missing phenylalanine ring. The electron density indicated here is based on the subsequent refinement of the re-modelled structure (R = .21).

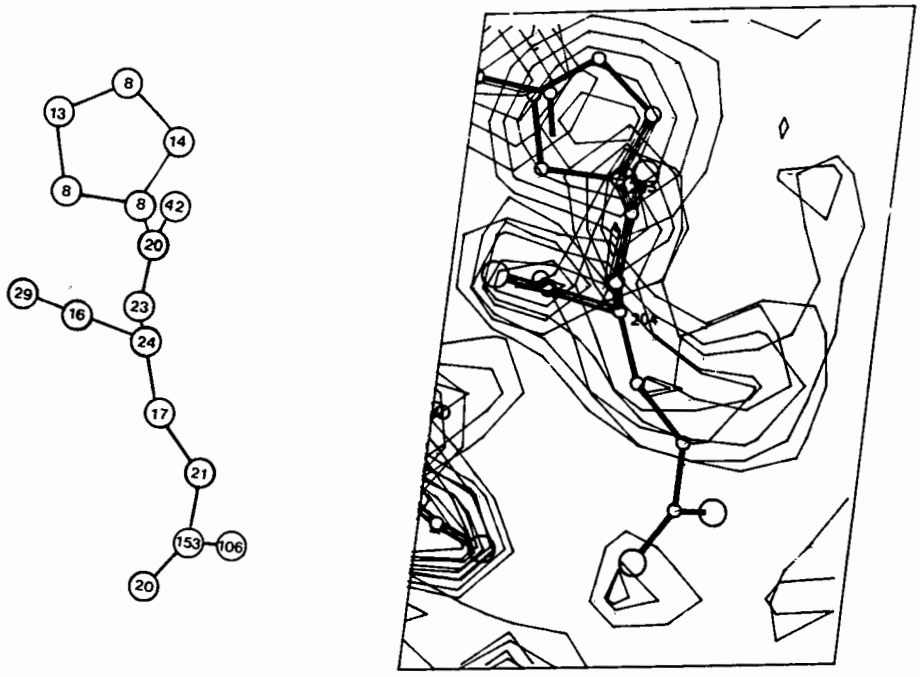


a b c

3.5 **Refinement of B values**

At medium resolution, the refinement of B values presents something of a problem. The parameters to observation ratio would certainly suggest that individual B values obtained by refinement procedures, should be interpreted cautiously to say the least, and may contain serious errors. Nevertheless, if the wrongly placed atoms in the structure could be refined to high B values, then their contribution to the calculated phases would decrease, further reducing the possible bias. In practice in the latter stages of refinement, even at 2.5A, it has proved possible to achieve some success with the refinement of B values, using them solely as a tool to reduce the bias, rather than as parameters with a realistic physical meaning. The procedure is one that should be approached with caution, but if bias checks have been incorporated, then the process may be monitored. For example, if the wrongly placed atoms of the bias check regions are seen to assume high thermal parameters in the refinements, then the procedure can be used fairly safely. However, if the wrongly placed atoms fail this test, then there is probably little value in proceeding with temperature factor refinement until a later stage. An example of the way in which the process can work is shown in fig. 6.

Figure 6 The temperative factors for an incorrectly placed glutamate residue and the corresponding electron density, is shown in figure 6a and b. Two atoms in the misplaced carboxyl group have attained high B factors effectively eliminating them from contributing to the calculated structure factors. The third atom of the group can be seen to have a normal temperative factor. This arises because it occupies a position in the structure which is though to contain a water molecule.



a b

4. CONCLUSIONS

The procedure of partial structure phase combinations has a great deal of potential in the refinement of structure at medium resolution. Used in conjunction with bias checking procedures a rapid assessment can be made of the status of the refinement and problems associated with refinement into false minima, can be avoided.

References

1. D.W. Rice, Acta.Cryst. (1981) A37, 491-500.

2. D. Stuart and P. Artymiuk, Acta.Cryst. (1985) A40, 713-716.

3. D.M. Blow and M.G. Rossmann, Acta.Cryst. (1961) 14, 1195-1202.

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

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

6. G. Bricogne, Acta.Cryst. (1976) A32, 832-847.

7. M.G. Rossmann and D.M. Blow, Acta.Cryst. (1961) 14, 631.

USE OF PHASE COMBINATION IN THE STRUCTURE ANALYSIS OF SERUM TRANSFERRIN

by

H. JHOTI
Laboratory of Molecular Biology, Department of Crystallography
Birkbeck College, Malet Street, London WC1E 7HX

1. INTRODUCTION

Serum Transferrin is a member of a family of proteins which also include Lactoferrin, Ovotransferrin and Melanotransferrin. These proteins are glycoproteins with a molecular weight of about 80kD and, with the possible exception of melanotransferrin, bind two Fe (III) ions per molecule. Their physiological functions vary and serum transferrin is the central iron-transport protein in invertebrates (see ref.1 for a review).

Phase combination techniques were used at two stages during the structure analysis of rabbit serum transferrin; initially, during the solvent flattening stage and later when a partial model of the protein had been built.

Crystals of rabbit serum transferrin have been grown in space group P43212 with cell dimensions of a=b=127.3(2)A, c=145.5(3)A which diffract to 3.0A and contain 68% solvent by volume. An M.I.R. map calculated using 3 derivatives clearly defined the molecular boundary but enabled relatively little secondary structure to be identified.

2. M.I.R. SOLVENT FLATTENING

Two different solvent flattening programs were employed in order to improve the phases.

2.1 Solvent flattening using DENMOD

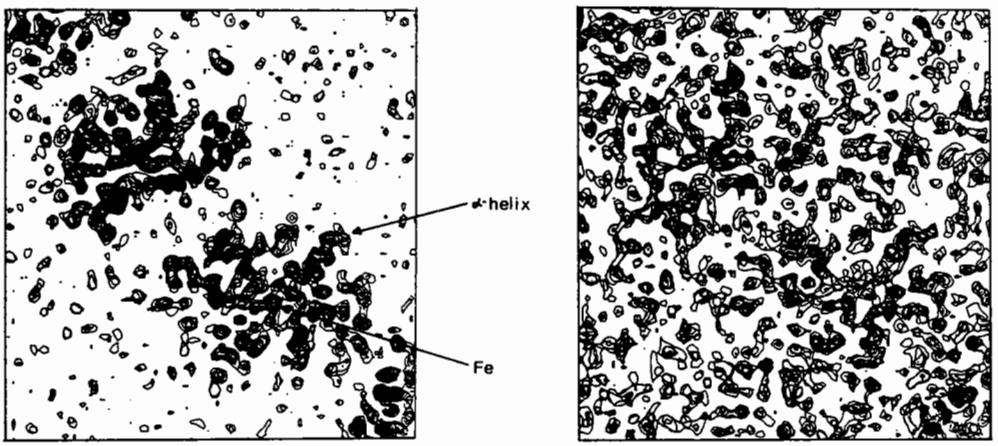
The program DENMOD (ref.2) was used which requires the molecular boundary to be set manually. Density outside the boundary was set to zero and negative density inside was truncated by multiplying by a factor of 0.1. Structure factors were then calculated from the modified map and a version of COMBINE (ref.3), modified by Ian Tickle, was used for the combination of the M.I.R. phases with those calculated from the modified map. After five cycles convergence was reached resulting in an overall phase shift from the M.I.R. phases of 36 degrees (Table 1).

The procedure resulted in a density map containing interpretable secondary structural elements such as alpha-helices and the positions of the two Fe atoms were confirmed (Fig.1).

TABLE 1 Comparison of the different solvent flattened procedures

	MEAN FOM	PHASE CHANGE FROM MIR (°)	No. of CYCLES
WANG R=8	0.76	33.71	4
WANG R=10	0.75	34.49	4
DENMOD	0.73	36.21	5

Fig.1 The effect of solvent flattening using DENMOD. Comparison of 6 sections of the MIR map before (right) and after (left) solvent flattening.



2.2 Solvent flattening using the Wang algorithm

The modified version (ref.4) of the Wang algorithm (ref.5) which automatically designates the molecular boundary was also employed. Solvent content was specified at 65% and averaging spheres of 8Å and 10Å were used. Convergence was reached after 4 cycles of density modification and phase combination resulting in a phase shift from the M.I.R. phases of 34 degrees in both cases (Table 1). There were relatively few differences between the Wang maps calculated with different radii nor indeed by artificially lowering the solvent content to 55% by volume. The 8.0Å radii map is representative and shows improvement which is comparable to the DENMOD map.

In general the DENMOD electron density map displayed greater clarity in the secondary structural elements than the Wang maps. However, there was more discontinuity present due to the omission of certain loops which were present in the Wang maps. This was due to the accidental excision of poorly-defined areas of density, such as surface loops, when manually designating the envelope. A comparison of the different molecular boundaries showed the Wang (R=8) boundary to be particularly convoluted as opposed to the more continuous DENMOD boundary (Fig.2).

Using a combination of the density modified maps an almost complete chain trace was possible and a model was built using the automated chain-tracing algorithm BONES and the fragment-fitting facility, both available with the graphics program FRODO version 6 (ref.6). The human serum transferrin sequence was used for the model interpretation since the rabbit sequence is as yet not completely defined; the rabbit sequence was incorporated into the model as it became available (ref.7).

2.3 The Model

The first model built (MODL1) consisted of about 3700 atoms which represented about two-thirds of the structure (the total number of atoms being about 5400). About 90% of the backbone was present and the tertiary structure of the protein was clearly defined (Fig.3). Serum transferrin has a bilobial structure with each lobe further divided into two dissimilar domains each of which consists of a β-sheet surrounded by helices and loops (ref.8).

3. REFINEMENT OF SERUM TRANSFERRIN

3.1 Phasing from a partial structure

In order to improve the density in the ill-defined regions the method of phase combination using the model as the partial structure was employed. Rice (ref.9) showed the potential of this kind of approach in the refinement of Phosphoglycerate kinase. Structure factors are calculated from the model and the calculated phases are combined with the M.I.R. phases using the modified version of COMBINE. These combined phases can be used together with the observed structure amplitudes to form Fourier coefficients in new Fourier synthesis.

Fig.2 A comparison of the different boundaries that were calculated. The DENMOD boundary is more continuous than the WANG boundaries.

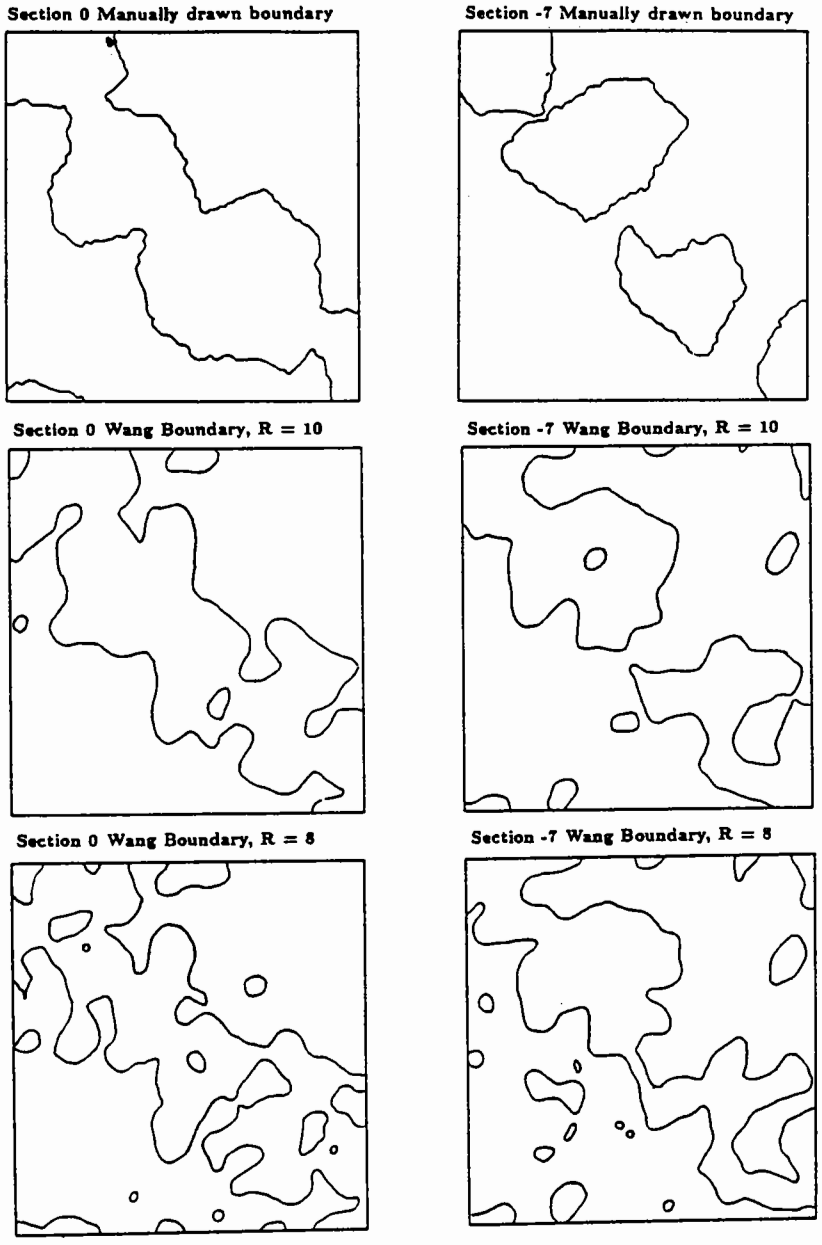
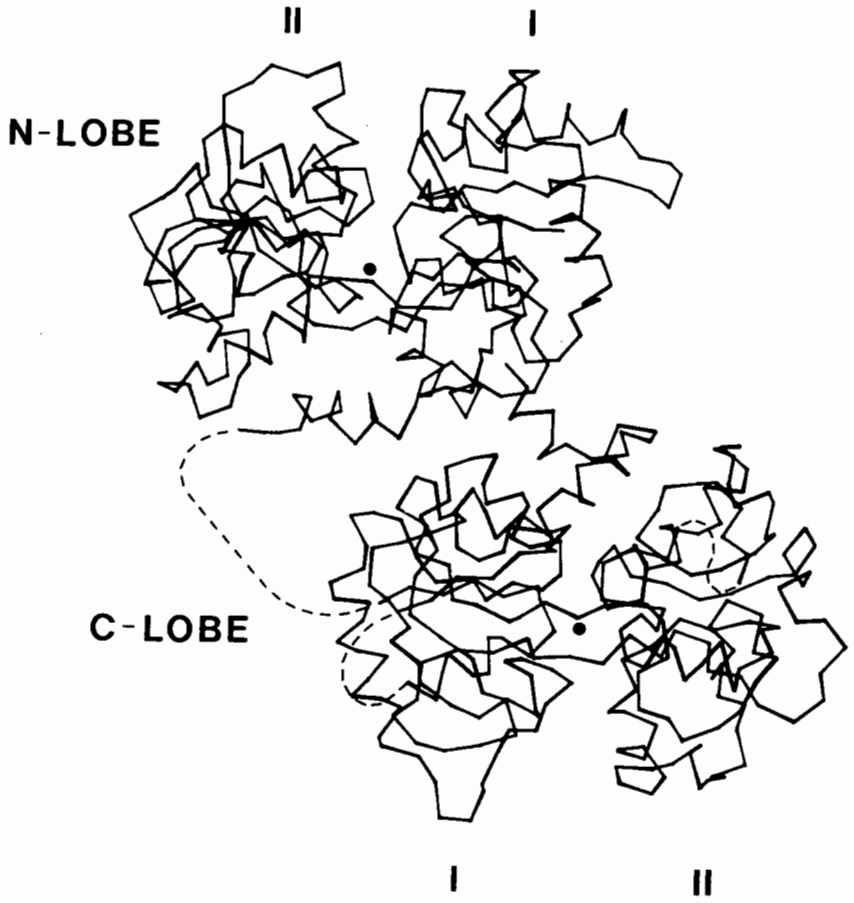


Fig.3 The bi-lobal structure of serum transferrin with each lobe further divided into two dissimilar domains, I and II. The iron-binding sites are located in the inter-domain clefts as indicated by the two dots. The dotted lines indicate regions which are at present ill-defined.



3.2 Rigid-body refinement

Using the least-squares refinement program RESTRAIN (ref.10) a rigid-body refinement of the model was performed prior to phase calculation but this had negligible effect in improving the R-factor or correlation coefficient. Phases were then calculated from the refined model and combined with the M.I.R. phases. Two maps were then calculated using different Fourier coefficients (Table 2), COMB1A was easier to interpret than COMB1B.

The COMB1A electron density map produced significant improvements and allowed 40 more residues, about half being alanines, to be inserted resulting in MODL2 (Table 2).

TABLE 2 Progress of the refinement of rabbit serum transferrin

	No. of ATOMS	R-factor after refinement	FOURIER COEFFICIENTS $Fo\phi_{COMB}$	$2Fo-Fo\phi_{CALC}$	$2Fo-Fc\phi_{COMB}$
MODL1	3700	49%	COMB1A	COMB1B	---
MODL2	4070	34%	COMB2A	COMB2B	COMB2C
MODL3	4855	34%	COMB3A	COMB3B	COMB3C

3.3 The effect of model-bias

The effect of model-bias was investigated by introducing deliberate errors into the model. Two side-chains which were well-defined by density were positioned out of the density and phases were then calculated from this model without prior refinement. The presence of density peaks around the wrongly positioned side-chains should give a qualitative analysis of model-bias in the structure. In both cases the wrongly positioned atoms did not give rise to significant density and the density for the original positions remained clear. A more rigorous analysis of model-bias could have been possible by refining individual temperature factors. However, the data to parameter ratio was a limiting factor.

3.4 Restrained least-squares refinement

MODL2 was then refined using the restrained least-squares refinement option in RESTRAIN. An overall temperature factor was employed and after 12 cycles the R-factor had been reduced from 49% to 34% for the data range 10-3.3 A. Phases were again calculated from the resulting refined model and used, in combination with the M.I.R. phases, to generate Fourier coefficients to calculate new maps (Table 2). Most improvements were observed in the COMB2C electron density map, however, MODL2 was rebuilt using all three electron density maps COMB2A, COMB2B and COMB2C. The resulting MODL3 consisted of 4855 atoms, an extra 70 residues. Surprisingly, the least-squares refinement of MODL3 was not successful in reducing the R-factor below 34% and the resulting maps COMB3A, COMB3B and COMB3C which used phase calculated from MODL3 showed limited improvements, vide infra.

4. S.I.R SOLVENT FLATTENING USING THE WANG ALGORITHM

The rationale behind combining phases calculated from the model with the experimental phases is such that the model is still only a partial structure with errors and so cannot solely provide the required phasing, it will also lead to reduction in model-bias. However, if the experimental phases are of poor quality leading to an M.I.R. map which is only partially interpretable (as in the case for rabbit transferrin) such a combination may be detrimental to the overall phasing. If there is an outstanding heavy atom derivative an alternative is to use solvent flattening on the S.I.R. map; this case corresponds to the initial concept of the Wang algorithm (ref.5). The resulting modified phases can then be used instead of the M.I.R. phases in the phase combination step.

Table 3 shows the effective phasing power of the three derivatives used in the calculation of the M.I.R. map. It is clear that the $HgCl_2$ derivative provides the best phasing power.

	← Resolution						
	14.74	8.51	6.59	5.57	4.91	4.45	
$HgCl_2$							
N_{ref}	510	1084	1389	1643	1840	2011	
F_H	210	203	192	181	172	162	
E	128	115	90	99	110	109	
UO_2Ac_2							
N_{ref}	526	1095	1393	1650	1854	2041	
F_H	201	182	167	157	148	142	
E	148	141	115	123	142	156	
$NaAuCl_4$							
N_{ref}	529	1071	1349	1619	1815	2008	2
F_H	155	146	139	132	125	118	
E	121	126	98	113	131	135	1
FOM	0.67	0.66	0.63	0.59	0.56	0.53	0.:

TABLE 3 The phasing statistics for the three derivatives used in the calulation of the M.I.R. map. The $HgCl_2$ derivative provides the best contribution to the overall phasing.

4.1 Calculation of solvent flattened S.I.R. phases

Phases were calculated based on the isomorphous and anomalous contributions from the $HgCl_2$ derivative and also the anomalous contribution from the two Fe atoms present in the protein. The mean figure of merit for this derivative was 0.50 compared with a value of 0.52 when all three derivatives were used. A comparison of the progress made during the solvent flattening of the M.I.R. and S.I.R. data can be seen in Table 4. In both cases the solvent level was set at 65%, a radius of 8Å was used and the molecular boundary was calculated only once. The final mean F.O.M. in both cases was about 0.76, however the absolute phase shift from the experimental was about 6° greater in the S.I.R. case.

	S.I.R.		M.I.R.	
CYCLE	MEAN FOM	PHASE CHANGE FROM S.I.R.	MEAN FOM	PHASE CHANGE FROM M.I.R.
1	0.69	33.05	0.70	26.46
2	0.74	36.76	0.74	30.53
3	0.76	38.74	0.75	32.42
4	0.77	39.90	0.76	33.71

TABLE 4 A comparison of using the Wang algorithm with S.I.R. and M.I.R. data. The final Mean Figures of Merit are similar in both cases.

The resulting solvent flattened S.I.R map was of a similar quality to the solvent flattened M.I.R. map indicating that the I.S.I.R., the iterative single isomorphous technique (ref.5), is at least as efficient as adding further heavy atom derivatives particularly if their phasing power is poor.

The S.I.R. solvent flattened phases were then combined with phases calculated from MODL3. A map was then calculated which used the same Fourier coefficients as the COMB3C map but by combining phases from MODL3 with S.I.R. solvent flattened phases rather than M.I.R. phases. This electron density map was compared with COMB3C and, generally, was of poorer quality suggesting that the M.I.R. phases should be used when employing this kind of approach in the refinement of a protein.

I acknowledge the support and assistance of the Transferrin group at Birkbeck College.

REFERENCES

1. Huebers, H.A. and Finch, C.A. Physiol. Revs. 67, (1987) 520-582

2. Bailey, S PhD Thesis, University of London (1987)

3. Bricogne, G. Acta Cryst., A32 (1976) 832-847

4. Leslie, A. Acta Cryst., A43, (1987) 134-136

5. Wang, B.C. in Methods in Enzymology (eds. Wyckoff, H.W., Hirs, C.H.W. and Timasheff, S.N.) 115, 90-112 (1985) (Academic Press Inc. London)

6. Jones, T.A. and Thirup, S. EMBO J., 5, (1986) 819-822

7. MacKenzie, H. and MacGillivray, R.T.A. (1988) Personnel Communication

8. Bailey, S., Evans, R.W., Garratt, R.C., Gorinsky, B., Hasnain, S.S., Jhoti, H., Lindley, P.F. and Sarra, R. (1988) In press.

9. Rice, D.W. Acta Cryst., A37 (1981) 491-500

10. Haneef, I., Moss, D.S., Stanford, M.J. and Borkakoti, N. Acta Cryst., (1985) A41 426-433

WEIGHTING IN PHASE COMBINATION

by

Ian J. TICKLE
Department of Crystallography, Birkbeck College, London.

1. INTRODUCTION

The aim of weighting the various parameters governing the probability density functions (p.d.f.'s) of amplitudes and phases obtained from different sources, such as isomorphous replacement and atomic (or other) model, is to obtain the "best" set of structure factors to use in a Fourier synthesis that will embody all the combined information.

For the isomorphous replacement information, we have amplitudes and both an algebraic form and the values of the coefficients for the phase p.d.f.'s (Hendrickson & Lattman [1]). This formalism is also ideally suited for expressing the p.d.f.'s of the "combined" phases, provided we have the correct values of the coefficients.

For the model-derived information, we also have amplitudes and an algebraic form for the phase p.d.f.'s (Sim [2,3]), but it will become apparent that the method by which the values of the coefficients arew obtained that is in common use does not have a sound theoretical basis, and needs to be re-examined.

In addition, we need to consider what we mean by "best" in this context; minimum r.m.s. error in the electron density (Blow & Crick [4]), or the minimum bias towards the model. Consequently, a re-examination of the algebraic form used for the amplitudes of the "combined" synthesis is necessary.

2. THE MODEL-DERIVED PHASE-PROBABILITY DENSITY

Sim's expression [2,3] for the model-derived phase p.d.f. in the general acentric case assumes i) a partial atomic model, P, and ii) error-free coordinates for the atoms in the model.

$$P(\phi) = \frac{\exp(X.\cos(\phi-\phi_P))}{2\pi I_0(X)} \qquad (1)$$

where

$$X = \frac{2F_N F_P}{\epsilon(\Sigma_N-\Sigma_P)} \qquad (2)$$

F_N = amplitude for complete structure,
F_P = amplitude for partial structure,
ϵ = symmetry factor
$\Sigma_N = \Sigma_N f^2$
$\Sigma_P = \Sigma_P f^2$
ϕ_P = phase for partial structure,
$I_n(X)$ = n'th order hyberbolic Bessel function.

Also define $\Sigma_Q = \Sigma_N - \Sigma_P$, where Q is the missing part of the structure. (Note that for the centric case we use Woolfson's p.d.f. [5]; I will not deal with the centric case explicitly, only mention it where the algebra differs from the acentric case.) The "Sim weight", or model-derived figure-of-merit, m, is given by

$$m.\exp(i\phi_P) = \int_0^{2\pi} P(\phi).\exp(i\phi).d\phi \qquad (3)$$

and

$$m = \langle\cos(\phi-\phi_P)\rangle \qquad (4)$$

$$= \frac{I_1(X)}{I_0(X)} \qquad (5)$$

For centric

$$m = \tanh(X/2)$$

The problem centres on the term $\Sigma_Q = \Sigma_N - \Sigma_P$. Currently various empirical estimates of Σ_Q are in use, based on the lack of agreement between either the observed and calculated amplitudes, or their squares, or the square of the amplitude of the difference structure factor. This lack of agreement has contributions from sources of error other than missing atoms, but we pretend that Sim's formulae apply in this case also. Three expressions currently in use are :

i) $\Sigma_Q = \overline{n(Fo-Fc)^2/\varepsilon}$
where n=2 for acentric, n=1 for centric, (cf Blundell & Johnson [6]).

ii) $\Sigma_Q = \overline{|Fo^2 - Fc^2|/\varepsilon}$
This is used in the currently distributed version of COMBINE.

iii) $\Sigma_Q = \overline{(Fo^2+Fc^2-2m.FoFc)/\varepsilon}$
(Nixon & North [7]). I have used this expression in a local version of COMBINE; however all three expressions are of dubious validity. Read [8] in fact compared $\cos(\phi_N-\phi_P)$ with m for a partial but correct structure and for a partial and incorrect structure where the correct phases were known and found that these expressions failed to give reliable estimates of phase probabilities.

3. THE EFFECT OF COORDINATE ERRORS

Srinivasan [9] showed that the algebraic forms for the phase p.d.f. are the same for a partial error-free structure and for a complete structure with coordinate errors, but that the values of the coefficients differ for a given discrepancy between Fo and Fc. The effect of the coordinate errors is to multiply the atomic scattering factors in the p.d.f.'s by $D = \langle\cos(2\pi\underline{\Delta r}.\underline{s})\rangle$ where $\underline{\Delta r}$ = coordinate error vector and $\underline{s}$ = scattering vector. For a normal distribution of coordinate errors this becomes $D = \exp(-\pi^3\langle\Delta r\rangle^2\sin^2\theta/\lambda^2)$. The Sim formula is modified by replacing F_P by $D.F^c_P$ where F^c_P is the amplitude for the partial structure with errors, and Σ_P is replaced by $D^2\Sigma_P$ so that now

$$X = \frac{2D.F_N F^C{}_P}{\varepsilon(\Sigma_N - D^2\Sigma_P)} \qquad (6)$$

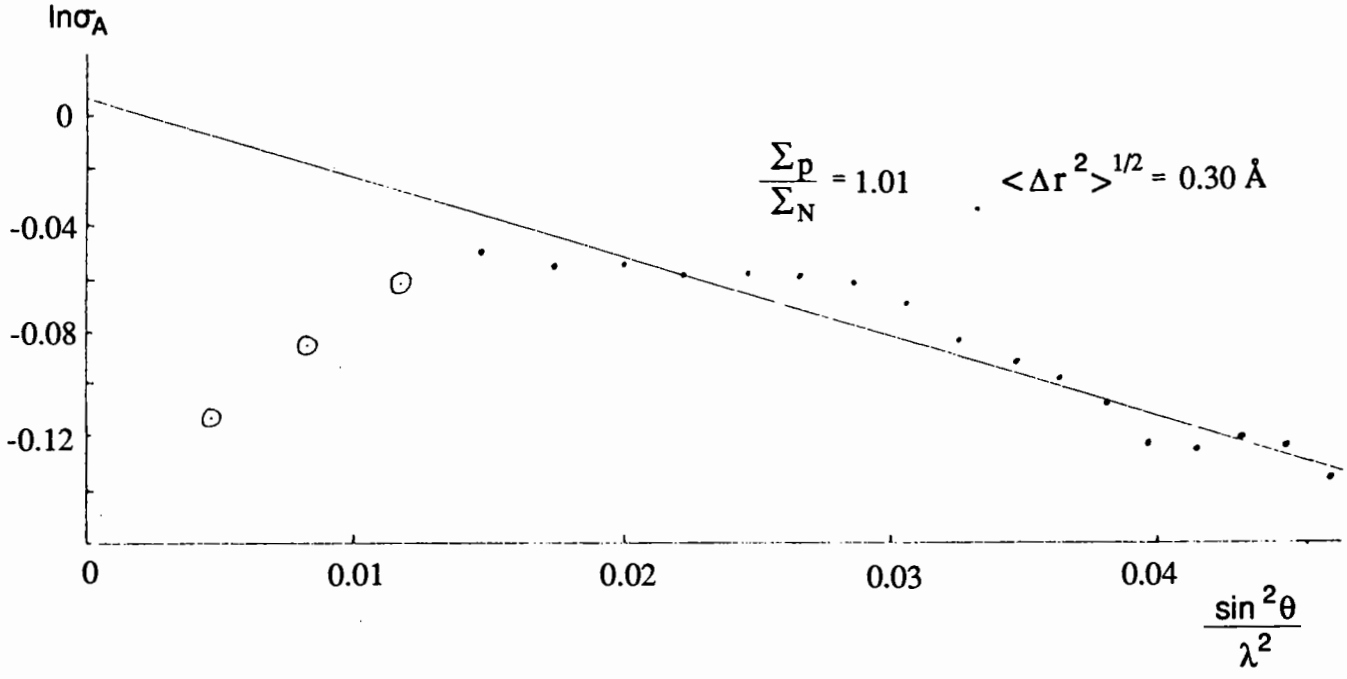
$$= \frac{2\sigma_A E_N E^C{}_P}{1-\sigma_A{}^2} \qquad (7)$$

where $\sigma_A = D(\Sigma_P/\Sigma_N)^{1/2}$ and the E's are normalised structure amplitudes. The new value for X goes into the original expressions (1), (4) and (5) for $P(\phi)$ and m. Setting D=1 (no errors) just gives back the original formulae.

Read [8] obtained σ_A by numerical methods. The r.m.s. Δr and Σ_P/Σ_N are obtained from a plot of $\ln\sigma_A$ versus $\sin^2\theta/\lambda^2$, since

$$\ln\sigma_A = -\pi^3\langle\Delta r\rangle^2\sin^2\theta/\lambda^2 + \tfrac{1}{2}\ln(\Sigma_P/\Sigma_N) \qquad (8)$$

A typical plot is shown below (the three circled points at low resolution were not used in the fit; they are probably due to contributions from atoms with relatively high temperature factors which violate the implicit assumption that Σ_P/Σ_N is independent of resolution).



4. THE UNBIASSED ESTIMATE OF THE MODEL-DERIVED STRUCTURE FACTOR

For a partial, error-free structure, the structure factor for the missing atoms is $\underline{F}_Q = \underline{F}_N - \underline{F}_P$. For the derivations that follow various expressions involving F_Q are required. Expressing the last equation in terms of amplitudes and phases :

$$F_Q\exp(i\phi_Q) = F_N\exp(i\phi_N) - F_P\exp(i\phi_P) \qquad (9)$$

Multiplying through by $\exp(-i\phi_P)$ and separating into real and imaginary components :

$$F_Q\cos(\phi_Q-\phi_P) = F_N\cos(\phi_N-\phi_P)-F_P \quad (10)$$
$$F_Q\sin(\phi_Q-\phi_P) = F_N\sin(\phi_N-\phi_P) \quad (11)$$

Squaring and adding (or equivalently using the cosine rule) :

$$F_Q^2 = F_N^2+F_P^2-2F_NF_P\cos(\phi_N-\phi_P) \quad (12)$$

Since all terms involving Q are naturally unknown, these equations have to be replaced by their expected values to be useful :

$$\langle F_Q\cos(\phi_Q-\phi_P)\rangle = m_PF_N-F_P \quad (13)$$
$$\langle F_Q^2\rangle = F_N^2+F_P^2-2m_PF_NF_P \quad (14)$$

where

$$m_P = \langle\cos(\phi_N-\phi_P)\rangle$$

Following Wilson [10], Σ_Q is obtained from $\langle F_Q^2\rangle/\varepsilon$ by averaging in shells in reciprocal space (indicated by subscript s). Σ_N and Σ_P are analogously related. Therefore

$$\overline{(\langle F_Q^2\rangle/\varepsilon)}_s = \overline{((F_N^2-F_P^2)/\varepsilon)}_s \quad (15)$$

and using (14), we obtain

$$\overline{(m_PF_NF_P/\varepsilon)}_s = \overline{(F_P^2/\varepsilon)}_s \quad (16)$$

The term to be used in the Fourier synthesis, which gives the minimum r.m.s. error in the electron density is $m_PF_N\exp(i\phi_P)$, but we need to express this synthesis in terms of the Fourier transforms of $\underline{F}_P$ and $\underline{F}_Q$, in order to have some idea of what peak positions and heights to expect in such a synthesis. The following derivation parallels that of Main [11], but it is useful to reproduce it here because of what follows later. Let

$$m_PF_N\exp(i\phi_P) = p.\underline{F}_P - q.\underline{F}_Q \quad (17)$$

where p and q are coefficients to be determined. To find p, multiply through by $\underline{F}_P^*$ and average in shells (the symmetry factor should be included). Since P and Q are totally independent of each other, having no atoms in common, the sum $\Sigma\underline{F}_Q.\underline{F}_P^*/\varepsilon$ will vanish provided sufficient terms are taken. Therefore

$$p = \frac{\overline{(m_PF_NF_P/\varepsilon)}_s}{\overline{(F_P^2/\varepsilon)}_s} = 1 \quad (18)$$

By a similar derivation

$$q = \frac{\overline{(m_PF_NF_Q\cos(\phi_Q-\phi_P)/\varepsilon)}_s}{\overline{(F_Q^2/\varepsilon)}_s} \quad (19)$$

or using the results (13), (15) and (16)

$$q = \frac{\overline{((m_P^2F_N^2-F_P^2)/\varepsilon)}_s}{\overline{((F_N^2-F_P^2)/\varepsilon)}_s} \quad (20)$$

To obtain the corresponding expressions for a partial structure with errors, it is only necessary to replace F_P by $D.F^C_P$ in (17) and (20). Main [12] showed that in theory for the error-free situation, p and q should be 1 and ½ respectively in the acentric case, while Read [8] showed that they should both be 1 in the centric case, and the following results (Table 1) for transferrin show that in practice q is somewhat less than ½ (p necessarily equals 1).

Table 1.	d_{min}	q(acentric)	q(centric)
	7.1	0.37	0.63
	5.6	0.34	0.53
	4.9	0.33	0.55
	4.5	0.31	0.49
	4.2	0.31	0.52
	3.9	0.30	0.51
	3.7	0.31	0.60
	3.5	0.34	0.52
	3.4	0.37	0.59
	3.3	0.37	0.53
	All	0.33	0.55

Setting p=1

$$m_P F_N \exp(i\phi_P) = \underline{F}_P + q.\underline{F}_Q \qquad (21)$$

hence

$$\underline{F}_N = \underline{F}_P + \underline{F}_Q$$

$$= (m_P F_N/q - (1-q)F_P/q)\exp(i\phi_P) \qquad (22)$$

To allow for coordinate errors now, it is only necessary to make the substitutions $D.F^C_P$ for F_P and ϕ^C_P for ϕ_P, so that

$$\underline{F}_N = (m_P F_N/q - (1-q)D.F^C_P/q)\exp(i\phi^C_P) \qquad (23)$$

In the case that $q=\frac{1}{2}$

$$\underline{F}_N = (2m_P F_N - D.F^C_P)\exp(i\phi^C_P)$$

which is the basis for the usual "2Fo-Fc" synthesis. For q=1/3 (approximately the value found for the transferrin data) we get

$$\underline{F}_N = (3m_P F_N - 2D.F^C_P)\exp(i\phi^C_P)$$

which is another variant on the same theme. The difference here is that the best coefficients can be deduced from the data itself, rather than being merely guessed at.

5. CONSIDERATIONS OF SCALING

It is normally necessary to apply a scale and temperature factor to obtain F_N from Fo; however the factor D automatically compensates for this, since

$$\sigma_A = D.\left\{\frac{\Sigma_P}{\Sigma_N}\right\}^{1/2} = D.\left\{\frac{\overline{(F^C{}_P{}^2/\epsilon)}_S}{\overline{(F_N{}^2/\epsilon)}_S}\right\}^{1/2}$$

The shell scale factors are therefore taken up in D, which then can no longer be interpreted purely in terms of coordinate error. However if reflections have been systematically omitted by intensity, the average E^2 will be biassed and this breaks down. For a difference Fourier it is important to have a correct relative scale (these characteristics are of course no different from those of ordinary Fouriers and difference Fouriers):

$$\underline{F}_N - \underline{F}_P = (2m_P F_N - (D+1)F^C{}_P)\exp(i\phi^C{}_P) \qquad (24)$$

Note that the scale factor determined by least squares is biassed, unless all parameters affecting the agreement are at a global minimum, so that the commonly used expression:

$$G(s) = \frac{\overline{(wFoF^C{}_P)}_S}{\overline{(wF^C{}_P{}^2)}_S}$$

obtained by minimising $\Sigma w(Fo-G.F^C{}_P)^2$ is invalid. Better is:

$$G(s) = \left\{\frac{\overline{(Fo^2/\epsilon)}_S}{\overline{(F^C{}_P{}^2/\epsilon)}_S}\right\}^{1/2} \qquad (25)$$

6. THE UNBIASSED ESTIMATE OF THE COMBINED STRUCTURE FACTOR

The "best" (i.e. minimum r.m.s. error in electron density) combined structure factor is $m_C F_N \exp(i\phi_C)$ where m_C and ϕ_C are the figure-of-merit and phase arising from combination and integration of the separate phase p.d.f.'s.

Expressing this again in the form

$$m_C F_N \exp(i\phi_C) = p.\underline{F}_P + q.\underline{F}_Q \qquad (26)$$

estimates of p and q can again be obtained from the data; in fact in this case no rigorous theoretical treatment is available. From (9)

$$F_Q\cos(\phi_Q-\phi_C) = F_N\cos(\phi_N-\phi_C) - F_P\cos(\phi_P-\phi_C) \qquad (27)$$

or

$$\langle F_Q\cos(\phi_Q-\phi_C)\rangle = m_C F_N - F_P\cos(\phi_P-\phi_C) \qquad (28)$$

so that now

$$p = \frac{\overline{(m_C F_N F_P \cos(\phi_C-\phi_P)/\epsilon)}_S}{\overline{(F_P{}^2/\epsilon)}_S} \qquad (29)$$

and

$$q = \frac{\overline{(m_C F_N F_Q \cos(\phi_C-\phi_Q)/\varepsilon)}_S}{\overline{(F_Q^2/\varepsilon)}_S} = \frac{\overline{(m_C F_N (m_C F_N - F_P \cos(\phi_C-\phi_P))/\varepsilon)}_S}{\overline{((F_N^2-F_P^2)/\varepsilon)}_S} \qquad (30)$$

Again, to obtain the expressions for a partial structure with errors, replace F_P by $D.F^C_P$ and ϕ_P by ϕ^C_P. The full expression for the unbiassed combined structure factor is therefore

$$\underline{F}_N = m_C F_N \exp(i\phi_C)/q - (p-q)D.F^C_P \exp(i\phi^C_P)/q \qquad (31)$$

For transferrin the values for p and q calculated from the data were around 1 and ½ for acentric and around p=q=1 for centric (Table 2). If these values were used in a Fourier synthesis the modified structure factor would be :

$$\underline{F}_N = 2m_C F_N \exp(i\phi_C) - D.F^C_P \exp(i\phi^C_P) \qquad (32)$$

for acentric, and

$$\underline{F}_N = m_C F_N \exp(i\phi_C) \qquad (33)$$

for centric. Note that in the general acentric case the resultant phase will be neither ϕ_C nor ϕ^C_P, but somewhere between the two.

Table 2.

d_{min}	acentric		centric	
	p	q	p	q
7.1	0.98	0.58	0.92	1.05
5.6	0.96	0.57	0.87	1.01
4.9	0.95	0.55	0.88	1.00
4.5	0.96	0.50	0.93	0.82
4.2	0.96	0.48	0.84	0.92
3.9	0.95	0.46	0.69	0.97
3.7	0.94	0.47	0.80	0.89
3.5	0.94	0.49	0.82	0.87
3.4	0.94	0.48	0.78	0.87
3.3	0.94	0.49	0.62	1.00
All	0.96	0.51	0.86	0.94

Stuart & Artymiuk [12] obtained estimates for q for individual reflections by making the assumption that the p.d.f. for the M.I.R. phase can be approximated by a Gaussian centred on the centroid phase with a standard deviation related in an ad hoc way to the figure-of-merit. In this derivation of p and q, the form of the p.d.f. of the M.I.R. phase is implicit in the values for m_C and ϕ_C used, and the formulae (29) and (30) above are based on the statistical properties of reciprocal space; it would therefore not be valid to attempt to calculate them on an individual reflection basis.

7. REFERENCES

1. W.A. Hendrickson & E.E. Lattman, Acta Cryst. B26, (1970) 136.
2. G.A. Sim, Acta Cryst. 12, (1959) 813.
3. G.A. Sim, Acta Cryst. 13, (1960) 511.
4. D.M. Blow & F.H.C. Crick, Acta Cryst. 12, (1959) 794.
5. M.M. Woolfson, Acta Cryst. 9, (1956) 804.
6. T.L. Blundell & L.N. Johnson, Protein Crystallography, London: Academic Press, (1976).
7. P.E. Nixon & A.C.T. North, Acta Cryst. A32, (1976) 325.
8. R.J. Read, Acta Cryst. A42, (1986) 140.
9. R. Srinivasan, Acta Cryst. 20, (1966) 143.
10. A.J.C. Wilson, Acta Cryst. 2, (1949) 318.
11. P. Main, Acta Cryst. A35, (1979) 779.
12. D. Stuart & P. Artymiuk, Acta Cryst. A40, (1984) 713.