DL/SCI/R33

MOLECULAR REPLACEMENT

Proceedings of the CCP4 Study Weekend,
31 January - 1 February, 1992

Compiled by
E.J. Dodson, S. Gover, and W. Wolf.

SERC

DARESBURY LABORATORY
Daresbury, Warrington WA4 4AD

MOLECULAR REPLACEMENT

Proceedings of the CCP4 Study Weekend
31 January - 1 February, 1992

Compiled by
E.J. Dodson, University of York
S. Gover, University of Oxford
and
W. Wolf, Daresbury Laboratory

SERC
DARESBURY LABORATORY
1992

CONTENTS

INTRODUCTION

It is now thirty years since the seminal paper of Michael Rossmann and David Blow on the detection of subunits within the crystallographic asymmetric unit (Acta Cryst 15 (1962) 24-31) and the launch of molecular replacement as a tool in the determination of multimeric protein structures and those for which a homologous or related structure has already been determined.

The first CCP4 Daresbury Study Weekend on molecular replacement in 1985 indicated the maturity of the methodology and aimed to deepen understanding of its theoretical basis and access its practical limitations. Since then a great deal of experience has been accumulated; and the rapidly increasing database of known structures, developments in genetic engineering and growing interest in investigating complexes of proteins with various cofactors have stimulated the development of molecular replacement techniques. The aim of this year's Study Weekend was to review this progress and illustrate the new methods with practical examples.

Several new approaches - a new strategy for rotation function calculations, PC-refinement and practical application of the T2 type translation function - have already have been implemented in computer programs. A Bayesian type statistical approach could well be in the pipeline. New envelope definition and symmetry averaging methods have recently been developed to improve phasing and these have been successfully applied to difficult structures.

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

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

Eleanor Dodson
Sheila Gover
Wojciech Wolf

November 1992

Detecting Structural Similarity from Protein Sequence Comparison

Geoffrey J. Barton

Laboratory of Molecular Biophysics, South Parks Road, Oxford OX1 3QU, UK.

Introduction

The first step in solving the phase problem by molecular replacement is to identify a suitable structure to use as a search object. If *ab initio* structure prediction techniques were able to provide an accurate three dimensional model of a protein from the amino acid sequence, then this would be a straightforward task. Fortunately for the protein crystallographer, *ab initio* methods are still some way from providing this ultimate solution! However, there are a number of powerful techniques available to detect similarity between a protein sequence and a protein of known three dimensional structure. Some methods improve the sensitivity and selectivity of conventional sequence alignment methods by incorporating secondary or tertiary structural information from the protein of known structure. However, it can be difficult to decide when any given method is indicating genuine structural similarity, or merely a spurious match. Accordingly, in this article, I first briefly review the range of available methods and show their relative success in identifying the globin fold. I then describe an analysis that identifies the limits of detection for a standard pairwise sequence comparison method. For more detailed information on current sequence/structure comparison algorithms see volume 183 (1990) of *Methods in Enzymology*.

Overview of Methods

The available comparison techniques may be loosely divided into five categories of increasing complexity and sensitivity. In practice there is a lot of overlap between the methods in categories 2-5.

1. Pairwise sequence comparison.
2. Pairwise sequence comparison with secondary/tertiary structure information.
3. Multiple alignment comparison with/without secondary structure information.
4. Flexible patterns and templates.
5. Environment specific weighting and optimal threading.

(6. Three dimensional structure prediction!)

Pairwise sequence comparison may be applied to any two protein sequences. A pair score matrix is chosen that assigns a weight to the alignment of all possible pairs of amino acids, (e.g. AA might score 10 whilst AK scores -5), aligning any residue with a gap is assigned a negative value. A dynamic programming algorithm

(e.g. see Needleman & Wunsch 1970) is normally used to find the best alignment of the two sequences including a consideration of insertions and deletions. Although robust, the standard pairwise alignment methods take no account of secondary or tertiary structural constraints, e.g. insertions and deletions generally do not occur in the core of the protein. If we know the structure of one of the proteins, the position of the gaps can be encoded in the comparison to avoid core secondary structures. This approach yields alignments that are more consistent with structural features than straightforward sequence-only methods (Barton & Sternberg 1987a).

Often, the protein with which we are searching can be unambiguously aligned to other members of its family. The resulting **multiple alignment** may then be used as a more sensitive probe of other family members. The sensitivity is improved, since positions important to the protein fold (e.g. buried hydrophobic residues) are given a higher weight, and variable regions are given a lower weight when aligning to a further sequence. As with pairwise comparison, this method may be combined with secondary structural information. An extension of the multiple alignment method is to abstract only the most highly conserved regions into a **flexible pattern** (Barton, 1990, Barton & Sternberg 1990) or **template** (Bashford *et al.*, 1987; Taylor, 1986) that encodes the conserved secondary structures and other important features. As I will show in the next section, flexible patterns can yield great sensitivity and selectivity.

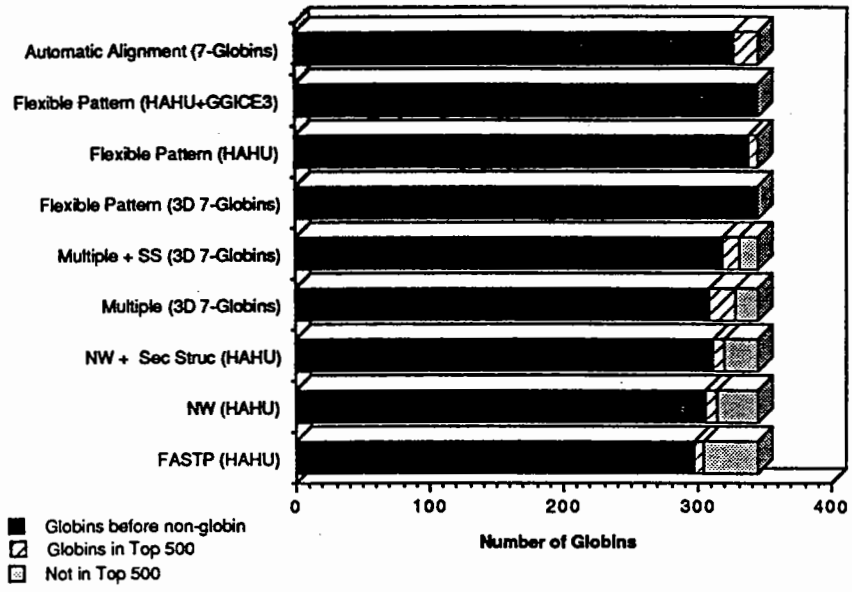
Environment specific weights are derived by studying the preferred environment, (i.e. exposed/buried, polar/non-polar contacts, secondary structure preference etc.) of each amino acid type, these weights are then applied to each residue in the protein of known three-dimensional structure to define a structural profile. Conventional dynamic programming is used to determine which proteins of unknown structure give the best alignment with the profile. Published accounts suggest, rather disappointingly, that environment weight methods give similar performance to conventional sequence comparison techniques (Bowie *et al.* 1991; Overington *et al.* 1992). **Pairwise potentials** take the idea of encoding the local environment a stage further. A residue-residue pair potential is derived from proteins of known structure, the sequence of unknown structure is then fitted to the core of the known structure and the lowest energy threading determined. This is a difficult optimization process that can not be solved by conventional dynamic programming techniques. However, preliminary results suggest this method has promise for detecting proteins that have similar folds, yet rather dissimilar sequences (e.g. two "Rossmann" beta/alpha/beta folds).

Evaluation of comparison methods

One approach to evaluating comparison methods is to select a well characterised protein family, then scan the entire sequence databank for members of that family. The globins provide a good test case with over 300 protein sequences known, and with representatives of diverse families with known crystal structures. Figure 1 illustrates the comparative success of different techniques for detecting the globin fold. In the databank scanned, there were 345 complete globin sequences, the query sequence or pattern was compared to all sequences (>6000) in the databank

Figure 1

Globin Scans (345 Whole Globins in Database)



and the resulting scores ranked. The results are presented as three values - number of globins before first non-globin, number of globins in top 500 scoring sequences, and number of globins not found in the top 500. These values give a measure of the selectivity and sensitivity of the different methods. Moving from the bottom of Figure 1, the methods get progressively better as more information is included in the scan. Starting with the simple, but fast FASTP algorithm (Lipman & Pearson, 1985), through Needleman & Wunsch (1970) (NW) , NW with secondary structural information, structurally derived multiple alignment and multiple alignment with secondary structure information, to flexible pattern, the methods improve. A flexible pattern derived from a single sequence and secondary structure does slightly worse than that derived from 7 structures, whilst adding a further sequence to the pattern recovers the sensitivity. Finally, deriving a pattern from an automatically determined alignment is slightly less specific than the structurally based alignment (see Barton & Sternberg, 1990 and Barton, 1990, for further details and discussion).

Guidelines for interpreting pairwise sequence alignments

Multiple alignment and flexible pattern comparisons provide good discrimination for the encoded protein fold. However, these techniques are not as frequently used as traditional methods. It is therefore important to have clear guidelines for interpreting conventional pairwise sequence alignments. For example, if I am given an alignment of two protein sequences that shows 32% identity, should I believe that the two proteins share similar folds? This question applies equally whether the proteins are of known or unknown structure. The following study was performed to establish general guidelines for the structural interpretation of sequence alignments.

Figure 2 illustrates the flow of the analysis. 477 chains from the Brookhaven PDB were grouped into two sets. Set 1 contained 89 unrelated protein chains selected from proteins known to have different folds. Set 2 comprised 57 distinct

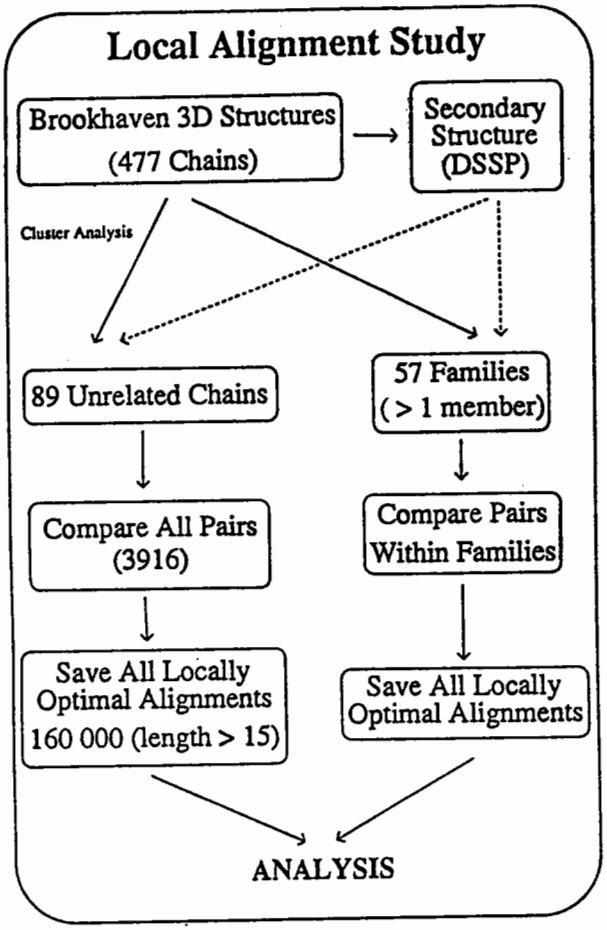


Figure 2

Figure 3

EXAMPLE LOCAL ALIGNMENT
SHOWING SECONDARY STRUCTURE

```
   CCCCCCCHHHHH HCHHHHHHHHHHHHHHHHH
45 RFKHLKTEAEMK ASEDLKKHGVTVLTALGA    74
   **     *     * * *** *   *   *
73 RFPEFRDENIRAVAIENLKKRGIDALVVIGG   103
   CCHHHHCHHHHHHHHHHHHHHCCCEEEEEEC

   (Sperm whale myoglobin and
    E.coli phosphofructokinase)

   Alignment length (including gaps) = 31
   Number of aligned positions       = 30
   TOTAL SCORE                       = 61
   PERCENTAGE IDENTITY               = 35.5
   PERCENTAGE ACCURACY               = 50.0
```

families with each family made up of at least two proteins. The secondary structure of all proteins was defined by the Kabsch & Sander (1983) program DSSP. The 89 unrelated protein sequences were then compared pairwise using a variant of the Smith-Waterman (1981) local similarity dynamic programming algorithm, scoring conservative substitutions with Dayhoff's matrix. This algorithm can locate *all* locally optimal alignments between two protein sequences and resulted in 160,000 alignments of length > 15 . For each alignment, the statistics shown in Figure 3 for two unrelated proteins were calculated . A similar process was also performed within each of the 57 families in Set 2.

The alignments obtained between the 89 unrelated proteins provide a set of scores and accuracies against which any sequence comparison may be measured. Figures 4a and 4b illustrate a plot of percentage accuracy against percentage identity for the unrelated and related proteins respectively. There is no correlation between the percentage identity and accuracy for proteins of unrelated structure. Indeed, local alignments of unrelated proteins can show up to 45% identity when optimally aligned, and many related pairs show less than this value! This would suggest that it is impossible to say whether or not our 32% identity alignment indicates structural similarity. Of course, the situation is not as bad as Figure 4

Figure 4a (Unrelated Proteins)

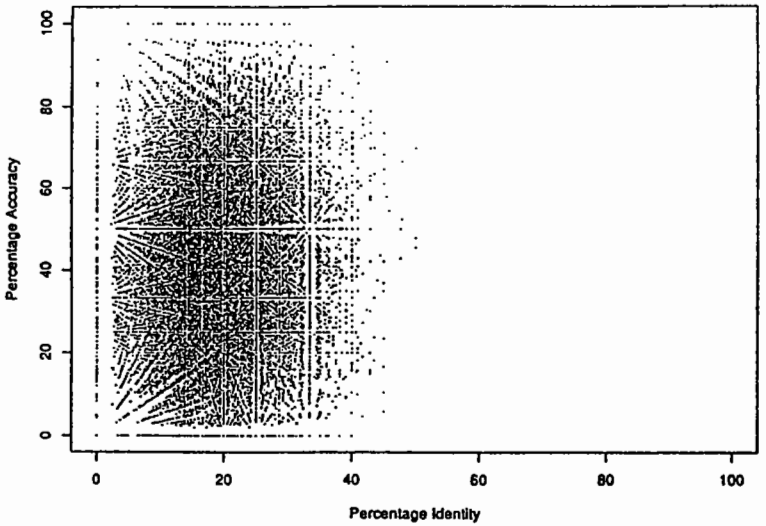


Figure 4b (Related Proteins)

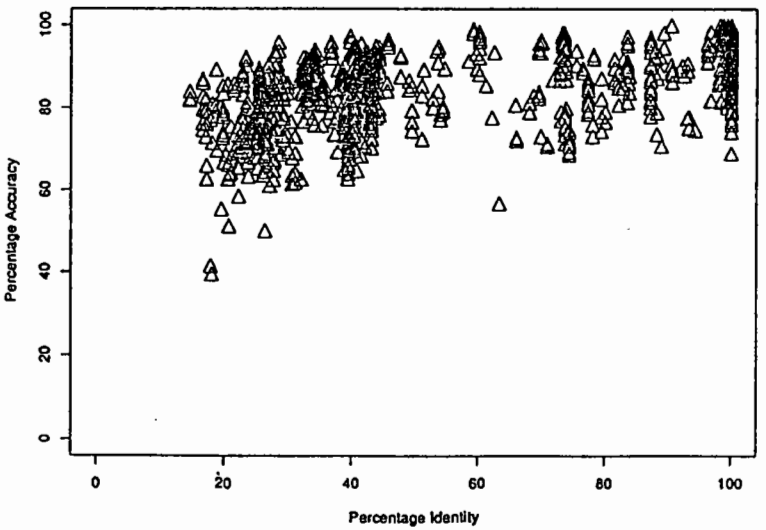


Figure 5a

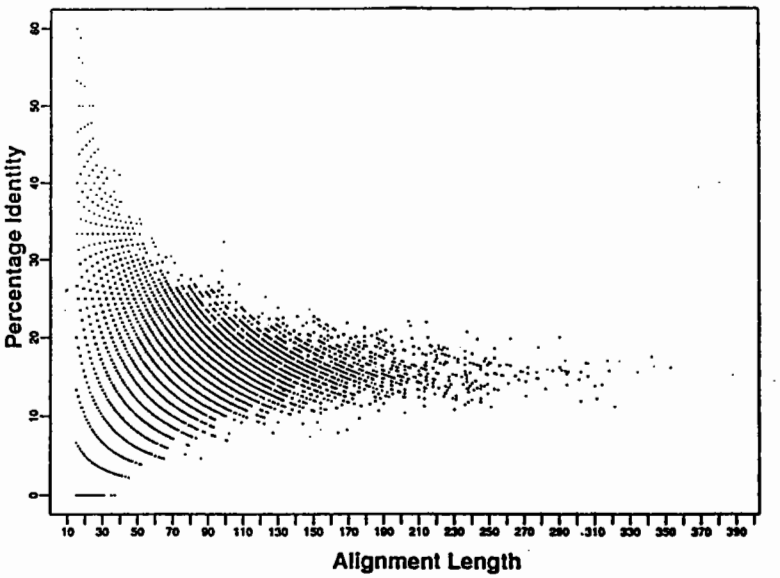


Figure 5b

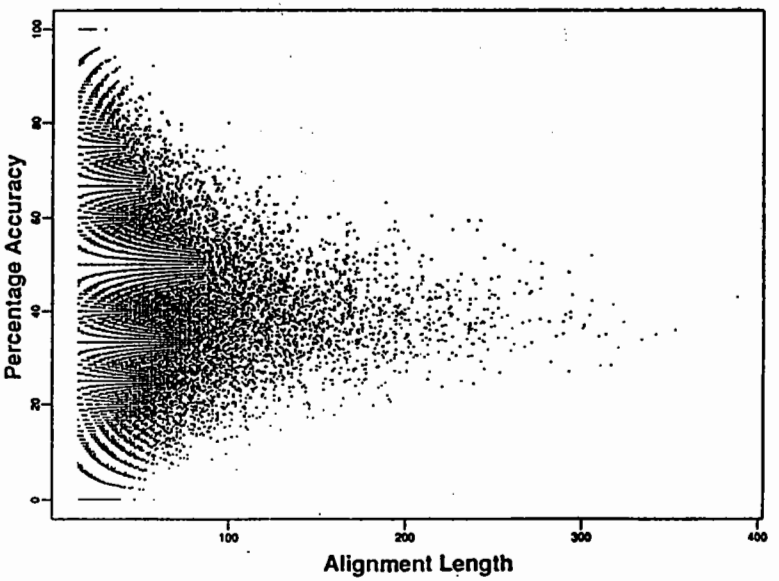


Figure 6a (Unrelated Proteins)

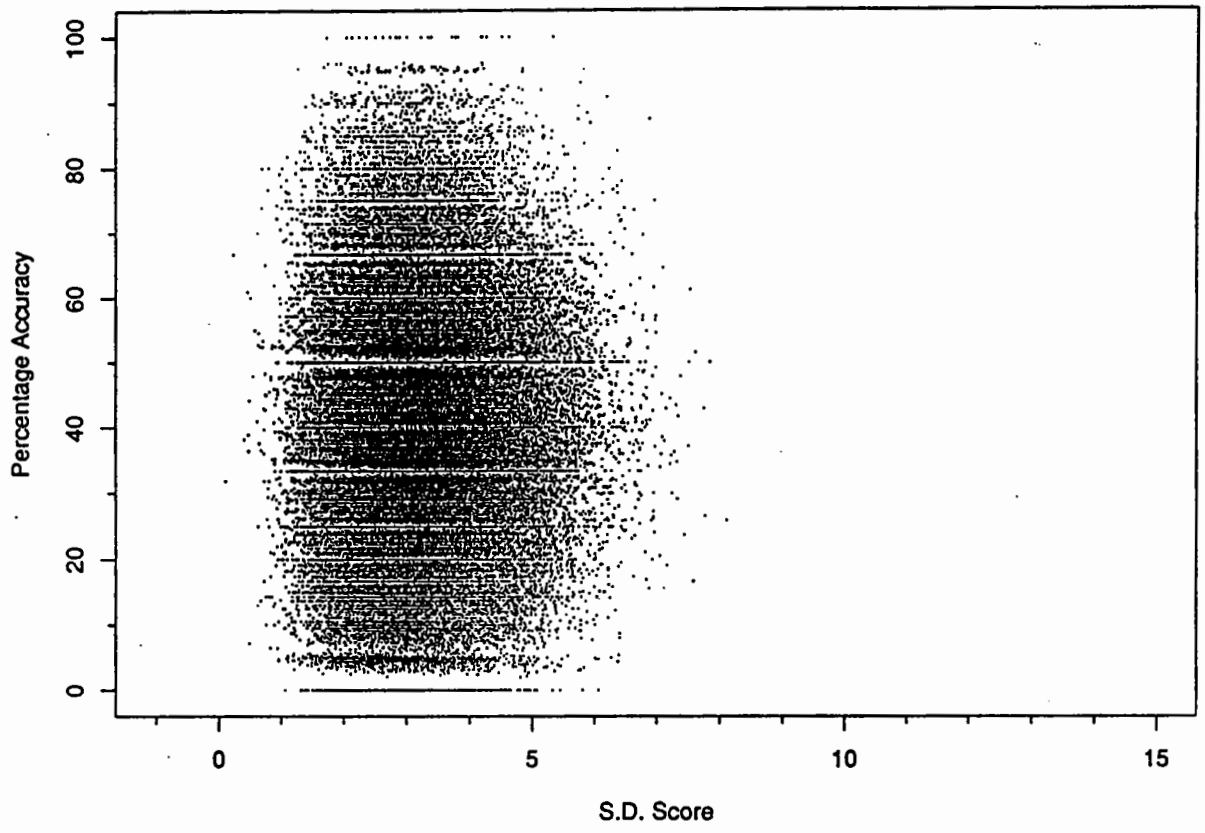
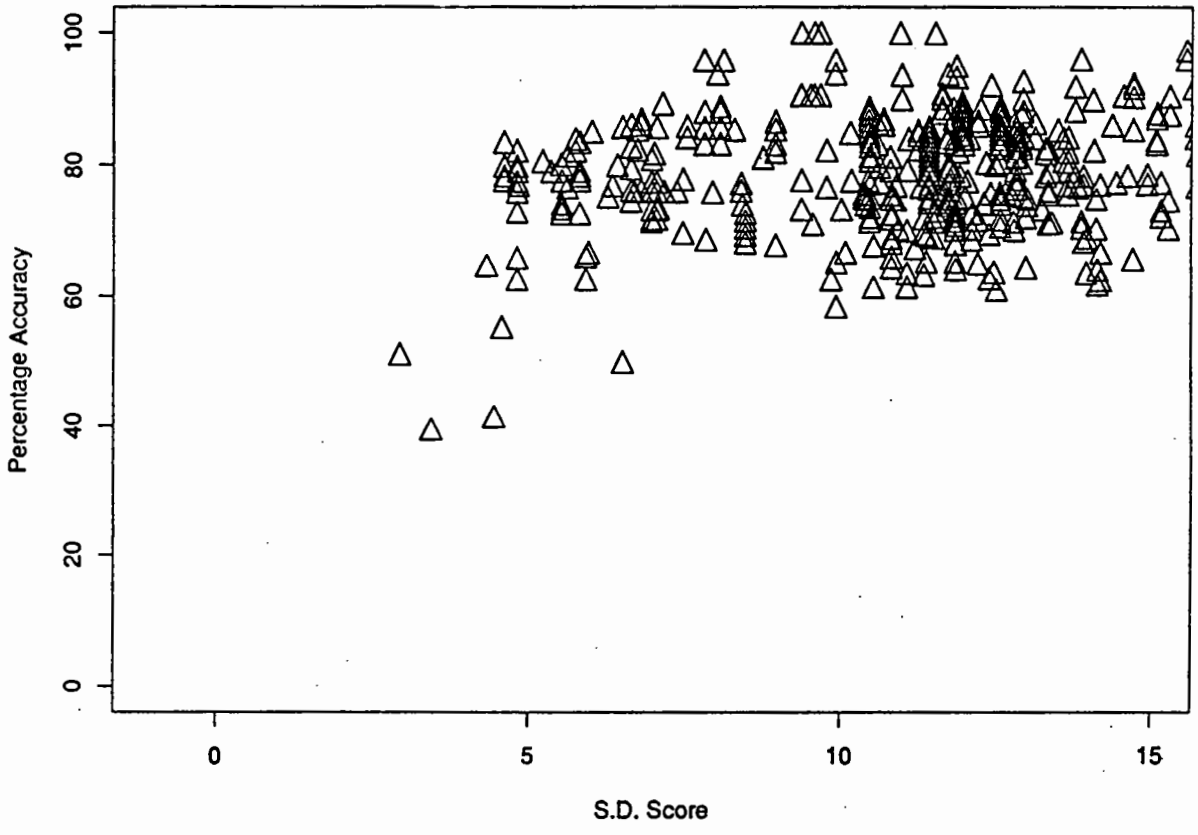


Figure 6b (Related Proteins)



suggests, since both percentage identity (Sander & Schneider, 1991) and percentage accuracy are strongly length dependent as is illustrated by Figures 5a and 5b. Figure 5a shows that 32% identity will be often seen by chance for alignments of < 100 residues, yet is comparatively rare above 100 residues. A typical protein sequence alignment of 150 residues that gives 32% identity over its entire length will, according to this plot, show that the two proteins share similar secondary structures. The data shown in Figure 5a and similar plots for other scoring strategies may be used to correct for the length dependency in the scoring scheme. Having corrected for the length effects, we can then compare the discriminating power of the different approaches. Table 1 shows the count of the number of protein pairs that we know share similar folds, yet give a corrected score lower than known unrelated pairs of proteins.

Table 1.

Length Corrected Scoring Scheme	Number
Percentage Identity	1280
Dayhoff Score	865
S.D. Score	846

Figure 7
Aligned Fragments of 2cts and 2paba

```
     *          *          *    *  *    **        * ****
     HHHHHCCCCCCCCHHHHHHHHHHCCCCCHHHHHHHHHHHHHCCCCCCCHHHHHHH
230  LYLTIHSDHEGGNVSAHTSHLVGSALSDPYLSFAAAMNGLAGPLHGLANQEVLV 283
  3  LMVKVLDAVRGSPAINVAVHVFRKAADDTWEPFASGKTSESGELHGLTTEEQFV  56
     EEEEEEECCCCEECCCCEEEEEECCCCCEEEEEEEECCCCCEECCCCCCCCCC
```

S.D. Score = 7.55
25.9% Identity
54 Residues

As one might expect, a scoring scheme that takes into account conservative substitutions (Dayhoff Score) performs somewhat better than the percentage identity scheme. The S.D. score is often calculated to estimate the similarity between protein sequences. This is done by first optimally aligning the two sequences, then shuffling the sequence orders, and re-aligning 100 or more times. The mean and standard deviation of the shuffled sequence alignment scores is then used to normalise the alignment score of the native sequences. S.D. scores also show a length dependency which is corrected for in Table 1. However, it is interesting to plot the raw S.D. scores against accuracy of alignment as shown in Figure 6a/b. It has long been known that S.D. scores higher than 5-6 are required to illustrate a genuine relatedness, or structural similarity (Barton & Sternberg, 1987b, Dayhoff, 1978), and FIgure 6 graphically illustrates this phenomena. The mean S.D. score for unrelated proteins is close to 3.0 and not 0.0 as it would be if protein sequence alignments were random. Some scores for the unrelated proteins are as high as 7.5 as for the example in Figure 7. Clearly, although these sequences give an S.D. score of 7.55, the value of 26% identity in 54 residues is insignificant according to Figure 5a.

Summary

Protein sequence comparison methods can be sensitive and selective tools for detecting proteins of known structure that share a similar fold to a protein undergoing crystallographic analysis. The analysis presented here, in particular Figures 5 and 6 should be useful when assessing the suitability of a protein as a search object for molecular replacement.

References

Barton, G. J., Meth. Enzymol. *183* (1990) 403-428.
Barton, G. J. & Sternberg, M. J. E., Prot. Eng. *1*, (1987a) 89-94.
Barton, G. J. & Sternberg, M. J. E., J. Mol. Biol. *198,* (1987b) 327-337.
Barton, G. J. & Sternberg, M. J. E., J. Mol. Biol. *212*, (1990) 389-402.
Bashford, D. Chothia, C. & Lesk, A. M., J. Mol. Biol. *196,* (1987) 199-216.
Kabsch, W. & Sander, C., Biopolymers *22*, (1983) 2577-2637.
Lipman, D. J. & Pearson, W. R., Science *227* (1985) 1435-1441.
Needleman, S. B. & Wunsch, J. Mol. Biol. *48*, (1970) 443-453.
Overington, J. *et al.* Prot. Sci. *1* (1992)
Sander, C. & Schneider, R., PROTEINS, *9*, (1991) 56-68.
Smith, T. F. & Waterman, M. S., J. Mol. Biol. (1981) 195-197.
Taylor, W. R., J. Mol. Biol. *188,* (1986) 233-258.

Lecture

Molecular Replacement

Paula Fitzgerald

Molecular Replacement in Macromolecular Crystallography

Definition of the Molecular Replacement Technique

Theory, Implementation and Limitations of:
- Rotation Function
- Positioning Techniques
  - Translation Function
  - Correlation Searches
  - Packing Analysis

Evaluation of the Correctness of a Potential Solution

Refinement of a Potential Solution

Some Illustrative Structures Phased by Molecular Replacement

Future Trends in Molecular Replacement Methodology

The Merlot Package

P. M. D. Fitzgerald (1988) *J. Appl. Crystallogr.* 21: 273-278

Attempt to rationalize and integrate existing programs
- Consistent conventions about angles
- Consistent data structures

Attempt to provide sufficient documentation and examples to make the technique accessible to the non-expert

Current major programs
- Crowther rotation function (Crowther)
- Lattman rotation function (Lattman)
- Crowther and Blow translation function (Lattman)
- R-value search (Fitzgerald)
- Packing function (Fitzgerald)
- Ward *et al.* R-Value minimization (Wishner and Ward)
- Glue to hold it all together (Fitzgerald)

Molecular Replacement

Molecular Replacement of the First Kind

The search for the orientation of non-crystallographic symmetry elements within the asymmetric unit

Rotation Function

Molecular Replacement of the Second Kind

The attempt to generate an initial phasing model for an unknown structure by orienting and positioning a molecule of known structure in the cell of the unknown

Rotation function
Translation function or some other positioning technique
Procedure for refinement of the orientation and position

Applications of Molecular Replacement of the First Kind

Used in combination with multiple isomorhpous replacement to improve refinement of heavy atom parameters

Rossmann	various dehydrogenase structures

Used to improve an electron density map by averaging around local symmetry elements

Bricogne	TMV
Rossmann	SBMV
Strandberg, Nordman	STNV
Harrison	TBSV
Wiley	Influenza virus hemaglutinin

Used for direct phasing

Rossmann	SBMV
Rayment	Polyoma virus

Rotation Function - Theory

M. G. Rossmann and D. M. Blow (1962) *Acta Crystallogr.* 15: 24-31

$$R = \int_{-\infty}^{\infty} P_2(x) * U(x) * P_1(C^*x)\, dV$$

where

$P_2(x)$	Patterson function of crystal 2 at real space point x
$P_1(x)$	Patterson function of crystal 1 at real space point x
C	Rotation matrix
U	Shape function
	1 inside some volume
	0 elsewhere

Rotation Function - Real Space

C. E. Nordman (1966) Transactions of the American Crystallographic Association, Volume 2

Search vector set can be modified in intuitively reasonable ways

Use weighted minimum function, rather than product function

Program Protein - Steigemann *et al.*

Program X-plor - Brünger

Rotation Function - Rossmann

M. G. Rossmann and D. M. Blow (1962) *Acta Crystallogr.* 15: 24-31

$$R = \Sigma_p \{F_1(p)^{**}2 * \Sigma_h [F_2(h)^{**}2 * G(h + C(t)^*p)]\}$$

where

$F_1(p)^{**}2$	Intensity for crystal 1 at reciprocal lattice point p
$F_2(h)^{**}2$	Intensity for crystal 2 at reciprocal lattice point h
C(t)	Transpose of rotation matrix C
G	An interference function; the Fourier transform of the shape function U

Completely general formulation; will work with any Patterson function
Extremely slow calculation
Limit length of calculation by using only top 10% of the data

Rotation Function - Lattman

E. E. Lattman and W. E. Love (1970) *Acta Crystallogr.* B26: 1854-1857

$$R = \Sigma_p [F_1(p)^{**}2 * C(t) * F_m(p)^{**}2]$$

where

$F_1(p)^{**}2$	Intensity for crystal 1 at reciprocal lattice point p
$F_m(p)^{**}2$	Intensity for crystal 2 at reciprocal lattice point p Replacing F_2 with F_m, the intensities for a crystal consisting of an isolated molecule in a large cell
C(t)	Transpose of the rotation matrix C

Not general, given the assumption that F_m is an isolated molecule
Requires intensities for F_m at reciprocal lattice point p, which is in general a non-integral reciprocal lattice point
Faster than Rossmann formulation, but still quite slow

Rotation Function - Crowther

R. A. Crowther (1972) in *The Molecular Replacement Method,* M. G. Rossman (editor), Gordon and Breach, New York, pp. 173-178

Expand Pattersons in terms of spherical Bessel functions
Problem then reduces to a series of two-dimentional Fourier series sums
Evaluate two-dimensional sums with a fast Fourier transform

Program is quite fast
Can use as much data as you like

Dimensioning limits maximum radius of integration for a given resolution of data
Space group symmetry limits fineness of grid on which the function can be evaluated

Rotation Function - Limitations

Problems with asymmetric molecules or asymmetric unit cells
Patterson vectors in the model may overlap with an adjacent origin in the Patterson of the unknown and give false peaks

Problems with large structures - Crowther program
Because of dimensioning limit on sphere in Patterson space over which function is evaluated

Problems with small structures - all programs
Low ratio of self-Patterson vectors to cross-Patterson vectors

Problems with poor models
Success depends absolutely on being able to accurately model the unknown structure

Translation Function - Theory

Vary the positions of two correctly oriented molecules in the crystal until the cross-Patterson vectors between them correlate with the Patterson of the unknown structure

The term "translation function" tends to be used to mean any technique used to determine the position of a properly oriented molecule

I use it specifically to mean those techniques, based in Patterson theory, whose evaluation involves a Fourier transform

Should be considered independently of positioning methods based on packing analysis or correlation searches

Translation Function - Crowther and Blow

R. A. Crowther and D. M. Blow (1967) *Acta Crystallogr.* 23: 554-558

$$T(t) = \Sigma_h [F_1(h)^{**}2 * F_{2i}(h) * F_{2j}^*(h) * \exp(-2\pi i h t)]$$

$F_1(h)^{**}2$	Intensity for crystal 1 at reciprocal lattice point h
$F_{2i}(h)$	Structure factor amplitude for molecule i of crystal 2 at reciprocal lattice point h
$F_{2j}^*(h)$	Complex conjugate of structure factor amplitude of molecule j of crystal 2 at reciprocal lattice point h
t	Translation vector

$T_1(t)$ replaces the term $F_1(h)^{**}2$ with $F_1(h)^{**}2 - \Sigma_{i=1,n} [F_{1i}(h)^{**}2]$

thus removing self-vectors from the Patterson function of crystal 1, which usually reduces noise in the search

Translation Function - Beurskens

P. T. Beurskens, R. O. Gould, H. J. Bruins Slot and W. P. Bosman (1987) *Z. Kristallographie* 179: 127-135

Review of published formalisms for the translation function, employing a common nomenclature

Emphasizes the theoretical continuum that exists between translation function formalisms and correlation searches

Testing of various formulations, concluding that:
- Intensity data should be sharpened
- Intensity data should be modified for origin removal
- Self-Patterson vectors should be subtracted
- All data, including weak data, should be used, especially when the search structure is a small fragment of the unknown

Translation Function - Langs

D. A. Langs (1975) *Acta Crystallogr.* A31: 543-550

Proposes a translation function formalism in which the structure factor terms for the model structure (intensity and phase) are replaced by terms with the phase information only

D. A. Langs (1985) *Acta Crystallogr.* A41: 305-308
D. A. Langs (1985) *Acta Crystallogr.* A41: 578-582
D. A. Langs (1987) *Acta Crystallogr.* A43: 733-734

Proposes modifications to the translation function formalism that reduce structure-dependent and structure-independent spurious maxima

Translation Function - Harada et al.

Y. Harada, A. Lifchitz, J. Berthou and P. Jolles (1981)
Acta Crystallogr. A37: 398-406

$$C(t) = \Sigma_i [x(i) * y(i)] / [\Sigma_i x(i)^{**}2 \; \Sigma_i y(i)^{**}2] 1/2$$

x(i)	= $F_1(h)^{**}2$	= Intensity for crystal 1 at reciprocal lattice point h
y(i)	= $F_2(h,t)^{**}2$	= Intensity for crystal 2 at reciprocal lattice point h given the translation vector t

Making a number of approximations, they end up with an expression that can be evaluated as a Fourier transform

A result of these approximations is that the formulation includes a term that is sensitive to allowed packing

Translation Function - Ruis and Miravitlles

J. Ruis and C. Miravitlles (1986) *Acta Crystallogr.* A42: 402-404

$$T(t) = \Sigma_h \Sigma_j \Sigma_{k>j} \{ [F_1(h)^{**}2 - \Sigma_{i,i-1,n} F_1(i)^{**}2] * F_2(j)F_2^*(k) * \exp(-2\pi i h r(kj))) \} * \exp(2\pi i h(kj)^* t)$$

$F_1(h)^{**}2$	Intensity for crystal 1 at reciprocal lattice point h
$F_2(j)$	Structure factor amplitude of molecule j of crystal 2
$F_2^*(k)$	Complex conjugate of structure factor amplitude of molecule k of crystal 2
r(kj)	r(k) - r(j), where r(k) and r(j) are the translation components of the symmetry operators that generated molecules k and j
h(kj)	h(R(k) - R(j)), where R(j) and R(k) are the rotational components of the symmetry operators that generated molecules k and j

The formulation allow the determination all vector components of the translation in a single calculation

Translation Function - Limitations

Extremely sensitive to the accuracy of the rotational angles

Sensitive to systematically missing data

Sensitive to pathological crystallographic situations
- Local symmetry parallel to crystallography symmetry
- Vectors patterns parallel to crystallographic symmetry

Some implementations (TrnSum in Merlot) require a lot of manual interpretation

Correlation Searches - Theory

Given: The structure and orientation of the contents of the unit cell are known

Then: Systemacally translate the molecule through the cell and calculate some measure of correctness at each sample point

Measures of correctness
- R-value - Program RvaMap in Merlot
- R-value - Many other programs by many other people
- Correlation coefficient - Program Brute (Fujinaga and Read)
- Correlation coefficient - Method of Harada

Correlation Searches - Fujinaga and Read

M. Fujinaga and R. J. Read (1987) *J. Appl. Crystallogr.* 20: 517-521

Use same correlation coefficient as described by Harada et al., but with the terms:

$x(i) = (F_1^{**}2 - \langle |F_2| \rangle^{**}2), \quad y(i) = (F_1^{**}2 - \langle |F_2| \rangle^{**}2)$

This formulation has the advantage of being independent of the relative scale of the observed and calculated data

Program Brute:
- Allows adjustment of rotation angles
- Allows search for correct position
- Allows searches with multiple types of molecules

Correlation Searches - Limitations

Functions under investigation (R-value or otherwise) tend to be very flat, with a sharp feature at the solution

Sampling frequency must be very fine so as not to miss the solution

Even with simplifying mathematics, every point evaluated requires a loop over all reflections, so these methods are computationally very expensive

These calculations become hopeless lengthy when there is more than one molecule in the asymmetric unit

Packing Analysis - Theory

Given: The orientation of a search molecule in the cell of the unknown crystal structure is known

Then: Use some measure of physical reasonableness to find the position of the search molecule in the cell of the unknown

Measures of physical reasonableness
- Require that molecules not interpenetrate
- Require that molecules fill space

Packing Analysis by Minimization of Bad Contacts

Systematically translate molecule through cell
- Calculate "all" intermolecular distances at each sample position

  Packing Function - G. Cohen and S.-W. Suh (unpublished)
  - Appoximate protein shape with a number of spheres
  - Calculate only intersphere distances

  Packing Function - R. Bott and R. Sarma (1976)
  - *J. Mol. Biol.* 106: 1037-1046
  - Define a criteria for bad contacts
  - Abandon evaluation of a sample translation when a user defined number of bad contacts had been encountered

  Program PakFun in Merlot
  - Uses a cubing algorithm to limit the number of distance calculations
  - Tries to encounter bad contacts first by sorting atoms on distance from molecular center

Packing Analysis - Hendrickson and Ward

W. A. Hendrickson and K. B. Ward (1976) *Acta Crystallogr.* A32: 778-780

Describe each molecule by a shape function M(x) which is
- 1 if x is intramolecular
- 0 if x is elsewhere

For each sample translation
- Define a grid in translation space and initialize all values to 0
- Translate the shape function
- Set the value at a grid point to 1
  - if M(x) is 1
  - if a symmetry mate of M(x) is 1
- Sum over the grid points to form T(R,t)

Investigate those points in the function where T(R,t) is large

Packing Analysis - Limitations

As with all positioning techniques, success depends on the accuracy of the input rotational information, although accuracy is not as critical in packing analysis as it is in translation function searches

Not an effective approach if only part of the unknown structure can be modeled

Calculations are slow, but new algorithms help a lot

Calculations are hopelessly slow when there is more than one independent molecule in the asymmetric unit

Packing analysis only provides a measure of allowed regions in translation space

Some other technique, usually a correlation search, must be used to determine the precise translation

Refinement of a Potential Solution - Theory

Limit of sensitivity of the rotation function is 3-5 degrees

Limit of sensitivity of the positioning techniques varies
- Correlation searches - moderate to high
- Translation function - moderate to high
- Packing analysis - low

Must improve on the fit before beginning refinement
- Six-dimensional problem with one molecule in the asymmetric unit
- Six*N-dimensional problem with N molecules in the asymmetric unit

Efficiency and accuracy of the approach employed matters

Refinement of a Potential Solution - Ward et al.

Local minimization of R-value

K. B. Ward, B. C. Wishner, E. E. Lattmann and W. E. Love (1975) *J. Mol. Biol.* 98: 161-177

Vary all independent angle and translation parameters
For each minimization cycle
- For each parameter
  - Calculate R for current value - delta
  - Calculate R for current value
  - Calculate R for current value + delta
- Adjust parameter values to correspond to R value minimum

Typically, with 4.0 Å protein data, one gets

Angle shifts of	0.0-3.0 degrees
Position shifts of	0.0-1.0Å
R value reduction of	2-10%

Refinement of a Potential Solution - Other Approaches

Program Corels - J. L. Sussman (1985) *Methods Enzymol.* 115: 271-303
- Group refinement in Corels is particularly nice, as you can increase the number of groups as needed
- Particularly good in cases where there is a molecular hinge that needs adjusting before refinement

ProFit(?) - Hendrickson
- Allows rigid body refinement in a least-squares formalism

X-plor - Brünger
- Allows rigid body refinement in a least-squares formalism

Evaluation of the Correctness of a Solution

Low residual
- Influenced by
  - quality of the data
  - quality of the model
  - any intensity screens applied to the data
- Differs with quality of the data and quality of the model
- Generally in the range 0.350-0.520

Non-interpenetration of molecules

Agreement with non-crystallography symmetry

Phasing power of the solution
- Remove part of the structure, particularly a heavy atom
- See if it returns in a difference map

Examples - HyHel5

S. Sheriff, E. W. Silverton, E. A. Padlan, G. H. Cohen, S. J. Smith-Gill, B. C. Finzel and D. R. Davies (1987) *PNAS* 84: 8075-8079

Unknown: anti-hen egg white lysozyme Fab complexed with lysozyme
Space group $P2_1$, 2 complexes per asymmetric unit

Probes: lysozyme
CL + CH1 domains of McPC603 Fab
VL + VH1 domains of McPC603 Fab without hypervariable residues

Results: All probes could be located, in both the rotation and translation functions, independently of the other two - remember that lysozyme represents only about 1/8 of the scattering matter in the cell
Brute was used in this determination to find the relative translations of the different probes.

Examples - Purothionin

M. M. Teeter, X.-q. Ma, U. Rao and M. D. Whitlow (1988) American Crystallographic Association Annual Meeting, Abstract T1

Unknown: Purothionin, space group I422

Probe: Crambin, or,
Crambin, amino acid susbstituted to represent purothionin and energy minimized

Results: Unmodified crambin gave a solution, but modifed crambin gave a significantly cleaner one

Selection of a Search Model

One needs to balance the accuracy of the search model with its completeness

The more structure you can accurately model, the higher the signal-to-noise ratio of the searches
The more inaccurate structure you include, the lower the signal-to-noise ratio of the searches

For protein models, one must decide whether or not to include

Side chains
High temperature factor regions
Sites of possible additions or deletions

Model Improvement - Program IntRef

T. O. Yeates and J. M. Rini (1990) *Acta Crystallogr.* **A46**: 352-359

Separate model into independent domains

Refine relative orientation of each domain simultaneously, using a least-squares approach

Need no knowledge of relative positions of the domains, but relative positions can be used if they are known

Particularly applicable to immunoglobulin structures

Model Improvement - Brünger

A. T. Brünger (1990) *Acta Crystallogr.* **A46**: 46-57

"Patterson" refinements of a large number of rotation function peaks
Calculate rotation function in real space

For each rotation function peak
Segment structure, use correlation coefficient to refine relative positions of the segments
Calculate translation function

Test case
Myoglobin with helices tilted by various angles
Can generate correct solution up to a 13° tilt

Molecular Replacement - Frontiers

Rotation function
Greater speed would be nice
Better approaches to large problems

Positionir.g techniques
Continuing stream of new approaches being published

Refinement
Refinement with dynamics (*a la* X-plor, Gromos) has revolutionized this part of the process

Model preparation
Mutation/energy minimization
Automated methods for determining incorrect portions of the model
Automated screening of multiple models

Fast Fourier Translation Functions.

by Ian J. Tickle, Birkbeck College, Malet Street, London WC1E 7HX.

1. Introduction

This paper describes the theory and some results obtained with the translation function programs TFSGEN and TFPART in the CCP4 suite. These programs are completely space-group general (provided standard settings in International Tables are used), and have recently been updated to accept keyworded input, to read general equivalent positions from the standard symmetry library, and òto use the new MTZ file format.

The only pre-requisite of these programs is that the orientation of the molecule whose position is to be determined is known accurately, typically by use of the Rotation Function (Crowther (1972), Navaza (1987, 1990)), preferably followed by rigid-body refinement (eg Yeates & Rini (1990), Brünger (1990)).

The TFSGEN program computes the Fourier transform of the translation function of the reference search molecule, using the complete set of intermolecular vectors between the reference molecule and all its equivalents related by crystallographic symmetry in the unit cell of the target structure. It uses a modified Crowther & Blow (1967) *T2* function, as originally suggested by Harada et al (1981). This was developed by Tickle (1985) and by Rius & Miravitlles (1986) to allow for subtraction of all intramolecular vectors (which are independent of the translation vectors and only add a noise contribution). This "full symmetry" translation function is most useful in high symmetry space groups, because the asymmetric unit of the translation function is largest for high-symmetry space groups (there are fewer alternative crystallographic origins).

The advantage of computing the Fourier transform over direct calculation in real space is of course that the Fast Fourier Transform (FFT) program can be used to effect the transformation, with a considerable saving in CPU time (as much as a factor of 10000 for a 3-dimensional calculation). This means that it is feasible to sample the rotation space finely around the putative solution and perform a large number of runs of the translation function to locate the solution in the 6-dimensional rotation/translation space, though to date this has not been performed automatically.

The original C&B *T2* function (which to the author's knowledge has never used to solve an unknown structure) was expressed in terms of vectors between local origins of symmetry-related molecules, so gave a peak for every pair of molecules when transformed into real space. The modified *T2* function uses a single translation vector of the reference molecule as the 3-dimensional variable (ie the vector required to shift the molecule to its correct position in the target structure), so that a single peak is obtained with improved signal/noise. The *T2* function is simply related to the C&B *T1* function: the *T1* function uses only the set of intermolecular vectors between a single pair of molecules, so the *T2* function is just the sum function of the individual *T1* functions.

The TFPART program is for use only in the case of multiple protomers in the crystallographic asymmetric unit, which are normally related by some non-crystallographic (local) symmetry operators; however the presence of non-crystallographic symmetry is not a pre-requisite, in principle it will work in the case of two or more unrelated molecules in the asymmetric unit, provided the orientation of each molecule whose position is to be determined is known.

TFPART uses the same translation function as TFSGEN, and calculates the contribution to the Fourier transform for the same reference molecule, but from the intermolecular vectors between that reference molecule and the set of molecules *not* related by crystallographic symmetry whose positions have already been determined, either by TFSGEN or by TFPART (Driessen et al (1991)).

For a full mathematical description of the *T1*, *T2* and other related translation functions, see Appendix 1 (case of crystallographic symmetry only) and Appendix 2 (case of crystallographic and non-crystallographic symmetry).

2. Applications of the crystallographic translation function.

A common application of the translation function is to resolve space-group ambiguities, not just in cases of enantiomorphic pairs that cannot be distinguished by systematic absences (eg $P4_12_12/P4_32_12$), but also cases where the space-group assignment is in doubt (eg $P2_12_12/P2_12_12_1$ where the c axis is short and so there may be insufficient $00l$ reflections to make a confident assignment).

For example Fig. 1 shows the *T2* function for a HIV1-proteinase inhibitor complex (space-group $P6_1$ or $P6_5$) which clearly identifies the space group as $P6_1$. The "ghost" peaks occur because some *T1* functions (for the 2-fold axes) are the same in the alternative space-groups, and also because the possible alternative origins for the *T1* functions depends on the symmetry of the rotation axis (for a 6-fold axis there is only 1 possible origin in the xy plane, for a 3-fold there are 3 alternatives, and for 2-fold there are 4).

Table 1 shows results of various translation functions for the hexagonal form of porcine pepsin (Cooper et al (1990)). I use this data as a test because the signal is very weak and it is therefore very sensitive to the various protocols.

In the *TO* function, no intramolecular vectors are subtracted; *TO/O* is the function proposed by Harada et al (1981). The *O* function is an overlap function which discriminates against solutions with bad packing. In fact the *T2* function already does the same thing because the short intramolecular vectors close to the origin are subtracted from the target Patterson, and this discriminates against solutions with short intermolecular vectors. I have noticed that even if the wrong Rotation Function solution is used, the *T2* function with intramolecular vector subtraction still tries to give solutions which pack well, so packing of a solution cannot necessarily be used as a criterion of correctness. Incorrect solutions of the translation function usually have very low signal/noise ($< 1\sigma$) relative to the second highest. The *TO/O* function requires 2 FFT's but normally does not give a better solution than *T2*.

Fig. 1. HIV1 proteinase / UK104561 inhibitor *T2* functions.

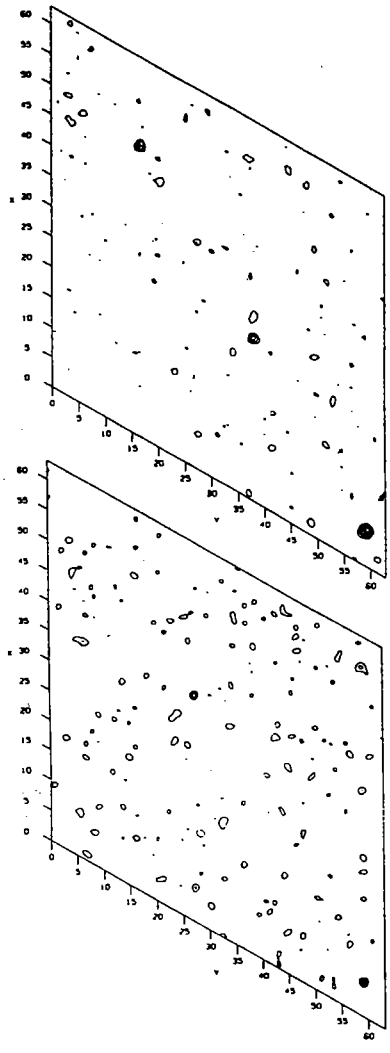


Fig. 2. Porcine pepsin *T2* function.

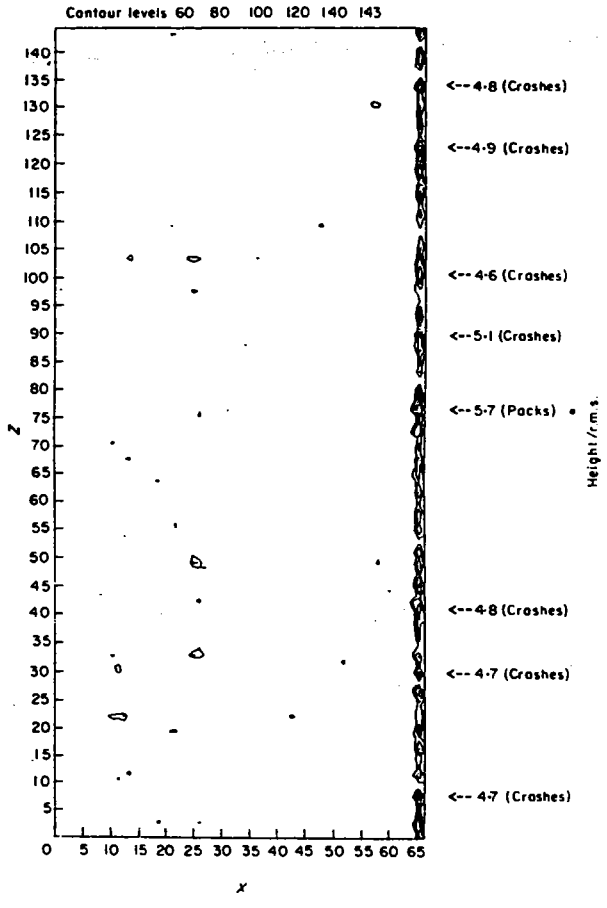


Fig.2 shows the *T2* function for porcine pepsin. The "streaking" along the 6-fold axis is a common feature of all translation functions that are computed as sum functions, and is a result of dominance by one or more of the component functions. In this case, one of the *T1* functions in the z=0 plane dominates, so the variation with z is smaller than that with x and y. Frequently one sees "streaks" along all 3 axes and these intersect at the solution peak (even if it is not the highest peak), so it is very easy to spot.

So far for all the structures that I know about that have been solved by the translation function programs, the highest peak has always been the correct solution, as judged by refining them. In cases where it has not worked there is therefore likely to be a problem with the accuracy of the Rotation Function solution.

Table 1: Translation functions.

Target: Porcine pepsin (M = 34kDa).
Space group: $P6_522$ (a = 67.4Å, c = 290.1Å)
Model: Penicillopepsin.
Resolution limits: 20-4Å

Function	Position	$\Delta S/\sigma$
TO	17	-1.11
T2	1	1.33
TO/O	1	0.88

Table 2: *T1* functions.

Axis	Position	$\Delta S/\sigma$
3_1 [001]	1	0.68
2 [010]	4	-0.13
2_1 [001]	2	-0.27
2 [120]	3	-0.51
6_1 [001]	7	-0.63
2 [210]	10	-1.85
2 [100]	33	-2.09
2 [110]	31	-2.16
2 [1$\bar{1}$0]	-	-3.29

The "Position" column shows the position of the correct peak in an ordered list produced by a peak search program; the "$\Delta S/\sigma$" column is the peak height of the correct solution relative to the highest noise peak in units of RMS translation function density.

Table 2 shows the *T1* functions for porcine pepsin. Only one of the *T1* functions produces a peak in the 1st position. The conclusion from these results is that it would have been very difficult to solve from *T1* functions as most of the peaks are buried in the noise.

3. Features of the TFSGEN and TFPART programs.

The most important features of the programs are: i) space-group generality, ii) speed and iii) good signal/noise.

i. The programs have now been tested with data in all crystal systems, and correct results obtained in all cases.

ii. Speed comparisons with the *R*-factor search program TSEARCH are shown in Tables 3 & 4. The FFT is several orders of magnitude faster than real space calculation. Also for porcine pepsin, which is a marginal case, the *R*-factor search gives lot of apparently better but wrong solutions; it probably has poor discrimination where the signal is weak, and application of resolution cutoffs or use of a finer (1Å) grid did not make any difference.

Table 3:
***T2* program timings.**

Res.lims.	Position	$\Delta S/\sigma$
20-4Å	1	1.33

Porcine pepsin 4Å data.
Timing for 1Å grid: 45 secs (TFSGEN) + 212 secs (FFT) = 4.3 mins total.
Least squares refinement (RESTRAIN) of this solution has given R = 19% to 2.3Å.

Table 4:
Luzzati *R1* program timings.

Res.lims.	$R1_{mean}$	$R1_{min}$	*R1*	Pos.	$\Delta S/\sigma$
20-4Å	64.9	62.3	62.6	17	-0.6
8-4Å	64.3	61.8	62.2	92	-0.7

Porcine pepsin 4Å data.
Timing for 2Å grid:
34 hrs (TSEARCH).
All timings were performed on a μVax-3000.

iii. Optimisation of $\Delta S/\sigma$ is particularly important for the non-crystallographic translation function where only a partial structure is used. The following recommendations are born of experience:

a) Use well-refined models.
b) Optimise the Rotation Function solution(s).
c) Use the modified *T2* function.
d) Subtract the transform of the intramolecular vector set and the intermolecular set between the previously determined subunits.
e) Use difference Wilson scaling in $(\sin\theta/\lambda)^3$ shells (Table 5).
f) Initially use all reflections (eg no amplitude or sigma cutoffs).
g) Use *E*'s instead of *F*'s (Tables 6 & 7).
h) Generally do not cut out data without very good reason; eg very low resolution (20Å) or poor high resolution data (eg > 3Å).

Table 5: Effect of shell scaling.

Porcine pepsin *T2* functions.

Nshells	Nrefls/ shell	Pos.	$\Delta S/\sigma$
1	3400	2	-0.34
20	170	1	1.33
68	50	1	1.12

Table 6:
Resolution limits & E's.

Porcine pepsin *T2* functions.

Res.lims.	*E*: Pos.	*E*: $\Delta S/\sigma$	*F*: Pos.	*F*: $\Delta S/\sigma$
20-4Å	1	1.33	1	1.25
20-5Å	3	-0.19	1	0.73
8-4Å	8	-0.53	7	-0.54
8-5Å	> 50	< -1.5	47	-1.38

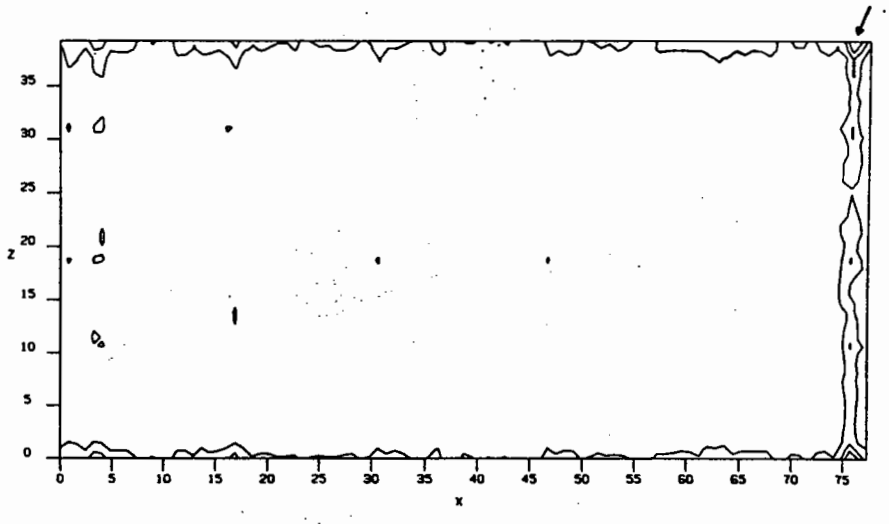
4. Applications of the non-crystallographic translation function.

Bovine eye-lens βB2-crystallin (H.P.C. Driessen *et al.*, Birkbeck).

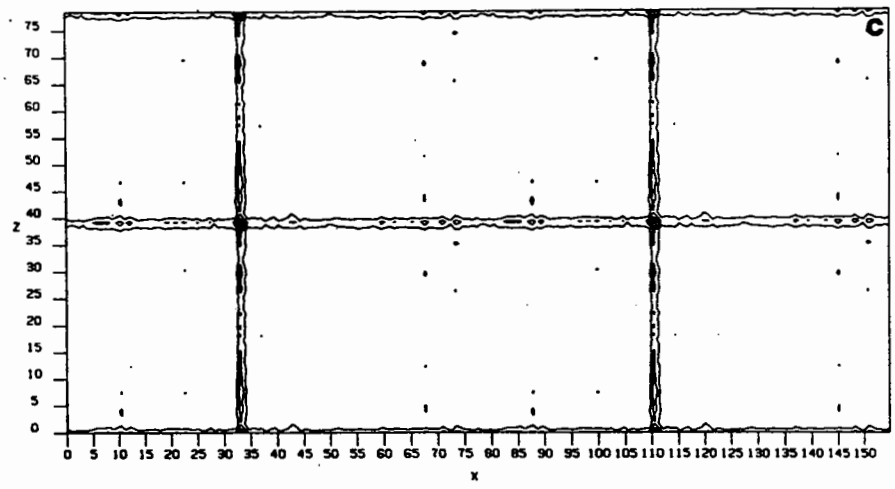
This has a tetramer in the asymmetric but a dimer of the *I*222 crystal form was used as a search model (Driessen *et al.* (1991)). Four *T2* functions were calculated (Table 8 and Fig. 3a,b,c,d). The choice of the first subunit is of course arbitrary; however

Fig. 3.
βB2-Crystallin *T2* functions.

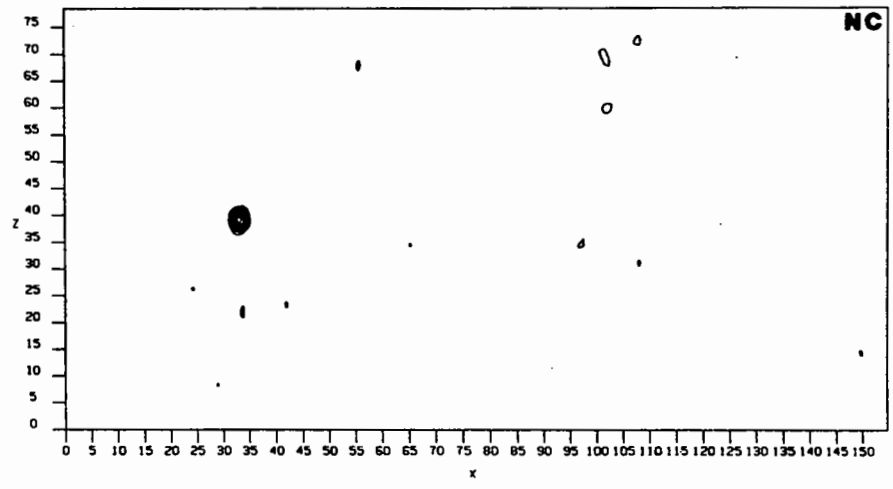
a) $T2_C$ for B dimer.



b) $T2_C$ for A dimer.



c) $T2_{NC}$ for A dimer.



d) $T2_{SUM}$ for A dimer.

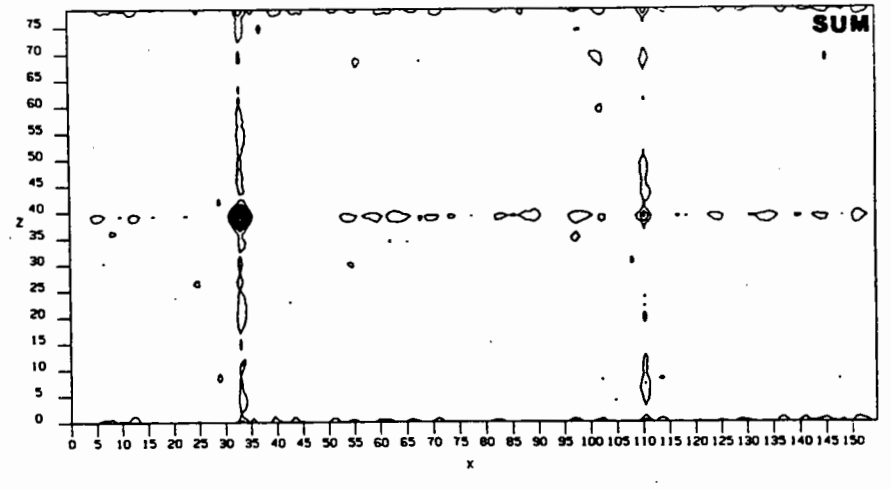


Table 7: Resolution limits & *E*'s.

βB2-crystallin *T*2 functions.

Res.lims.	*E*: $\Delta S/\sigma$	*F*: $\Delta S/\sigma$
∞-3.3Å	3.67	0.72
∞-4Å	4.01	1.43
∞-5Å	4.79	1.61
20-3.3Å	5.02	3.32
20-4Å	6.63	3.68
20-5Å	4.77	1.85
8-3.3Å	7.08	3.64
8-4Å	6.57	5.27
8-5Å	2.82	3.83

frequently one subunit gives a higher signal than the others, so it is probably better to choose this one first. The asymmetric unit of the crystallographic *T*2 function is 1/8 of the unit cell. The non-crystallographic *T*2 function discriminates between the alternative origins. The summed *T*2 function was unnecessary in this case, but can be useful if the search model is not so good.

Table 8:
βB2-crystallin translation functions.

Target: βB2-crystallin (*M* = 23kDa).
Space group: *C*222
Non-cryst.: 222
Model: βB2-crystallin (*I*222) dimer.
Res.lims.: 20-3.3Å

Function	Step	Pos.	$\Delta S/\sigma$
$T2_C$	Find B.	1	6.3
$T2_C$	Fix B; find A.	1	10.9
$T2_{NC}$	Fix B; find A.	1	18.1
$T2_{SUM}$	Fix B; find A.	1	22.1

Rat γE-crystallin (H.P.C. Driessen *et al.*, Birkbeck).

This has a dimer in the asymmetric unit. The space group is $P2_1$ so the crystallographic *T*2 functions are on one section (the origin along *y* is arbitrary for the first subunit). The non-crystallographic *T*2 function however is the whole unit cell, because the origin is fixed by the position of the first subunit. This is exactly analogous to the procedure for solving heavy-atom derivative Pattersons (Table 9).

Table 9:
Rat γE-crystallin translation functions.

Target: Rat γE-crystallin (*M* = 20kDa).
Space group: $P2_1$
Non-cryst.: 2 (Pseudo $P2_12_12_1$)
Model: Bovine γE-crystallin.
Res. lims.: 20-2.5Å

Function	Step	Pos.	$\Delta S/\sigma$
$T2_C$	Find A.	1	7.3
$T2_C$	Fix A; find B.	1	7.3
$T2_{NC}$	Fix A; find B.	1	11.9
$T2_{SUM}$	Fix A; find B.	1	12.0

Yeast proteinase *A* (C. Aguilar *et al.*, Birkbeck).

This is also a dimer in the asymmetric unit. The alternative space group ($I2_12_12_1$) did not give significant peak in the translation function (Table 10).

Table 10:
Yeast Proteinase *A* translation functions.

Target: Yeast Proteinase *A* (*M* = 41.5kDa).
Space group: *I*222
Non-cryst.: 2
Model: Porcine pepsin.
Res. lims.: 20-3.5Å

Function	Step	Pos.	$\Delta S/\sigma$
$T2_C$	Find B.	1	4.4
$T2_{NC}$	Fix A; find B.	1	12.5

$I2_12_12_1$: $\Delta S/\sigma = 0.12$

Human renin/CP85339 complex (V. Dhanaraj *et al.*, Birkbeck).

This has a dimer in the asymmetric unit (Table 11).

Table 11:
Human renin translation functions.

Target: Human renin/CP85339 complex (*M* = 40kDa).
Space group: $P2_13$
Non-cryst.: 2
Model: Porcine pepsin.
Res. lims.: 20-3Å

Function	Step	Pos.	$\Delta S/\sigma$
$T2_C$	Find A	1	6.6
$T2_{NC}$	Fix A; find B.	1	6.0

Mouse renin/CH66 complex (C. Dealwis *et al.*, Birkbeck).

This has a pseudo-222 tetramer in the asymmetric unit. The weak signal is greatly improved after refinement (Table 12). There were small rigid-body movements, but the improvement is mainly due to rebuilding of surface loops and sidechains.

Table 12:
Mouse renin translation functions.

Target: Mouse renin/CH66 complex (*M* = 40kDa).
Space group: $P2_1$
Non-cryst.: 222
Model: Porcine pepsin.
Res. lims.: 20-3Å

Function	Step	Pos.	$\Delta S/\sigma$	$\Delta S/\sigma^{\dagger}$
$T2_C$	Find A.	1	1.0	10.6
$T2_{NC}$	Fix A; find B.	1	2.0	26.2
$T2_{NC}$	Fix A,B; find C.	1	0.7	29.8
$T2_{NC}$	Fix A,B,C; find D.	1	1.9	44.9

†Note: last column shows value of $\Delta S/\sigma$ after SFLS.

Hen ovalbumin (A.G.W. Leslie *et al.*, MRC-LMB).

This example (Table 13; Stein *et al.* (1991)) demonstrates that the procedure works just as well in *P*1, with no crystallographic symmetry. There are two dimers related by a **b**/2 translation. The *T*2 function for the D subunit gives a very weak signal, possibly because it was difficult to discriminate in the rotation function between the very similar orientations of the B and D subunits.

Table 13:
Ovalbumin translation functions.

Target: Hen ovalbumin (*M* = 45kDa).
Space group: *P*1
Non-cryst.: 2+2
(pseudo **b**/2 translation)
Model: Plakalbumin.
Res. lims.: 8-4Å

Function	Step	Pos.	$\Delta S/\sigma$
$T2_{NC}$	Fix A; find B.	1	3.3
$T2_{NC}$	Fix A,B; find C.	1	6.5
$T2_{NC}$	Fix A,B,C; find D.	1	0.1

5. Appendices

In these notes I have attempted to show the main translation function expressions without going into details of derivations. Please consult the references for more detailed information. All the translation functions described here are based on the ideas of Crowther & Blow (1967). However their paper dealt only with the case of crystallographic symmetry; the relevant results are summarised in Appendix 1 below. An extension of their ideas to the case of non-crystallographic symmetry is then summarised in Appendix 2.

Notation for Appendices 1 and 2.

a,b,... Subscripts for subunits (if more than one) within the asymmetric unit.
$\mathbf{A}_j$ Rotation matrix component of the j'th crystallographic symmetry operator.
$\mathbf{d}_j$ Translation vector component of the j'th crystallographic symmetry operator.
E Normalised partial calculated structure factor for identity asymmetric unit.
Ec Normalised calculated structure factor.
E_j, E_k Normalised partial calculated structure factor for asymmetric units j & k.
E_m, E_n Normalised partial calculated structure factor for subunits m & n.
$|Eo|$ Normalised observed structure factor amplitude.
E_P Normalised partial structure factor for the set **P** of known subunits.
h Reciprocal lattice index vector; summations are over an asymmetric unit.
j,k Summation indices for asymmetric units within the primitive unit cell.
l Summation index for subunits within the primitive unit cell.
m,n Summation indices for subunits m & n within the asymmetric unit.
Pc Calculated Patterson function for the model structure.
Po Experimentally-derived Patterson function for the target structure.
P_{jk} Calculated intermolecular Patterson for the a.u.'s j & k of the search model.
P_{ll} Calculated intramolecular Patterson for the l'th subunit of the search model.
P The set of previously determined subunits (may be empty).
t Translation vector for the identity asymmetric unit.

$\mathbf{t}_j$ Translation vector (= $\mathbf{A}_j.\mathbf{t} + \mathbf{d}_j$) for the j'th asymmetric unit.
$\mathbf{t}_x$ Translation vector for subunit x in the identity asymmetric unit.
$\mathbf{t}_{jx}$ Translation vector for subunit x in the j'th asymmetric unit.
T_{jk} Pairwise translation function for asymmetric units j & k.
TO Full-symmetry translation function (Harada et al., 1981).
$T1_{jk}$ *T* function with subtraction of intramolecular vectors.
$T1_{jmkn}$ *T1* function for j'th a.u./m'th subunit & k'th a.u./n'th subunit.
T2 *TO* function with subtraction of intra- and known inter-molecular vectors.
$T2_{NC}$ Non-crystallographic component of the *T2* translation function.
u Vector in Patterson space.
U The remaining set of subunits (if any) whose positions are as yet unknown.
v Intermolecular vector (= $\mathbf{t}_k - \mathbf{t}_j$) between local origins of a.u.'s j & k.
Warning: C&B call the intermolecular vector **t**.
V Primitive unit cell volume.
x Subscript for subunit that is unknown but is currently being determined.

Appendix 1: The crystallographic translation functions.

Although, as will be seen, all the translation functions described here are computed via their Fourier coefficients, which are then transformed by FFT, the functions are most readily comprehended when expressed in real space.

C&B's *T* and *T1* functions are product functions of the experimentally-derived Patterson of the target structure with the calculated intermolecular Patterson for a structure consisting of just 2 correctly oriented molecules of the search model, regardless of the actual number in the unit cell. This is done for each crystallographic symmetry element (excluding lattice-centring translations) in the space group, and the relative orientation of each pair of molecules is determined by the respective symmetry element. Some manual effort is then required to interpret the results, particularly in high symmetry space groups.

The translation function *T* for the asymmetric units j and k, as a function of the intermolecular vector **v** is defined as:

$$T_{jk}(\mathbf{v}) = \int_V Po(\mathbf{u})\, P_{jk}(\mathbf{u},\mathbf{v})\, d\mathbf{u}$$

The translation function *T1* for asymmetric units j and k is the *T* function with all the intramolecular vectors subtracted:

$$T1_{jk}(\mathbf{v}) = \int_V (Po(\mathbf{u}) - \Sigma_l P_{ll}(\mathbf{u}))\, P_{jk}(\mathbf{u},\mathbf{v})\, d\mathbf{u}$$

Note the tacit assumption made here, that the intramolecular vector set of the target structure is the same as that of the model.

The full-symmetry translation function *TO* uses the calculated Patterson function for the total contents of the unit cell, but is most conveniently expressed in terms of the translation vector **t** for the identity asymmetric unit, because each pair of asymmetric

units is associated with a different intermolecular vector **v**:

$$TO(\mathbf{t}) = \int_V Po(\mathbf{u})\, Pc(\mathbf{u},\mathbf{t})\, d\mathbf{u}$$

Similarly the full-symmetry translation function T2 with intramolecular vectors subtracted is:

$$T2(\mathbf{t}) = \int_V (Po(\mathbf{u}) - \Sigma_I P_{II}(\mathbf{u}))\, (Pc(\mathbf{u},\mathbf{t}) - \Sigma_I P_{II}(\mathbf{u}))\, d\mathbf{u}$$

In reciprocal space, using normalised amplitudes, the *T1* and *T2* translation functions, which are the ones commonly used, become:

$$T1_{jk}(\mathbf{v}) = \Sigma_{\mathbf{h}} (|Eo(\mathbf{h})|^2 - \Sigma_I |E_I(\mathbf{h})|^2)\, \mathbf{Re}[E_j(\mathbf{h},\mathbf{t}_j)\, E_k^*(\mathbf{h},\mathbf{t}_k)]$$

$$T2(\mathbf{t}) = \Sigma_{\mathbf{h}} (|Eo(\mathbf{h})|^2 - \Sigma_I |E_I(\mathbf{h})|^2)\, (|Ec(\mathbf{h},\mathbf{t})|^2 - \Sigma_I |E_I(\mathbf{h})|^2)$$

$$= \Sigma_{\mathbf{h}} (|Eo(\mathbf{h})|^2 - \Sigma_I |E_I(\mathbf{h})|^2)\, 2\, \Sigma_j \Sigma_{k<j}\, \mathbf{Re}[E_j(\mathbf{h},\mathbf{t}_j)\, E_k^*(\mathbf{h},\mathbf{t}_k)]$$

Note that the *T2* function is just the sum of the individual *T1* functions. The normalised partial structure factor for the j'th asymmetric unit is related to the symmetry-equivalent structure factor for the identity by a phase shift:

$$E_j(\mathbf{h},\mathbf{t}_j) = E(\mathbf{h}.\mathbf{A}_j) \exp(i2\pi\mathbf{h}.\mathbf{t}_j)$$

$$= E(\mathbf{h}.\mathbf{A}_j) \exp(i2\pi\mathbf{h}.\mathbf{d}_j) \exp(i2\pi\mathbf{h}.\mathbf{A}_j.\mathbf{t})$$

This allows one to calculate structure factors from the model coordinates in a P_1 cell that has the same dimensions as the target structure, and produce a file containing the unique partial structure factors for each asymmetric unit.

Caveat - Because this operation is usually done by a program separate from the translation function, one must take care that the phase shifts due the symmetry translations $\mathbf{d}_j$ are applied once and once only, and that one remembers the order of the rotations $\mathbf{A}_j$! In practice the symmetry translations are best left to be performed by the translation function program, simply because in many cases there is a space-group ambiguity or uncertainty.

Finally, the Fourier coefficients of the *T1* and *T2* translation functions in the form required for computation, are as follows:

T1: Index vector = **h** Peak position = **v**

Structure factor = $(|Eo(\mathbf{h})|^2 - \Sigma_I |E(\mathbf{h}.\mathbf{A}_I)|^2)\, E(\mathbf{h}.\mathbf{A}_j)\, E^*(\mathbf{h}.\mathbf{A}_k)$

T2: Index vector = $\mathbf{h}.(\mathbf{A}_k - \mathbf{A}_j)$ Peak position = **t**

Structure factor =
$(|Eo(\mathbf{h})|^2 - \Sigma_I |E(\mathbf{h}.\mathbf{A}_I)|^2)\, \Sigma_j \Sigma_{k<j}\, E(\mathbf{h}.\mathbf{A}_j)\, E^*(\mathbf{h}.\mathbf{A}_k) \exp(i2\pi\mathbf{h}.(\mathbf{d}_j - \mathbf{d}_k))$

Appendix 2. The non-crystallographic translation functions.

The crystallographic translation functions can be readily generalised to give the non-crystallographic functions, with the difference that each subunit in the asymmetric unit is now associated with an independent translation vector. In real space:

$$T1_{jmkn}(\mathbf{v}) = \int_V (Po(\mathbf{u}) - \Sigma_l P_{ll}(\mathbf{u}))\, P_{jmkn}(\mathbf{u},\mathbf{v})\, d\mathbf{u}$$

$$T2(\mathbf{t}_a,\mathbf{t}_b,...) = \int_V (Po(\mathbf{u}) - \langle Pc(\mathbf{u})\rangle)\, (Pc(\mathbf{u},\mathbf{t}_a,\mathbf{t}_b,...) - \langle Pc(\mathbf{u})\rangle)\, d\mathbf{u}$$

Here, the total intramolecular Patterson $\Sigma_l P_{ll}(\mathbf{u})$ is replaced by the expectation of *Pc* with respect to the unknown translation vectors (see below). In reciprocal space:

$$T1_{jmkn}(\mathbf{v}) = \Sigma_h\, (|Eo(\mathbf{h})|^2 - \Sigma_l |E_l(\mathbf{h})|^2)\, \mathbf{Re}[E_m(\mathbf{h}.\mathbf{A}_j)\, E_n^*(\mathbf{h}.\mathbf{A}_k)\, \exp(i2\pi\mathbf{h}.\mathbf{v})]$$

$$T2(\mathbf{t}_a,\mathbf{t}_b,...) = \Sigma_h\, (|Eo(\mathbf{h})|^2 - \langle |Ec(\mathbf{h})|^2\rangle)\, (|Ec(\mathbf{h},\mathbf{t}_a,\mathbf{t}_b,...)|^2 - \langle |Ec(\mathbf{h})|^2\rangle)$$

As before the *T1* function uses only a pair of subunits at a time, so it is a function only of the intermolecular vector between that pair, and the FFT is 3-dimensional. However the *T2* function uses all pairs, so it is a function of all the translation vectors. This would require an FFT of dimension 3 times the number of subunits being searched for. This is not feasible, so the strategy adopted is to solve for the translation vectors in a stepwise fashion. In general at any stage of this procedure there will be a set **P** (which may be empty) of subunits whose translation vectors have been determined in previous steps, a translation vector $\mathbf{t}_x$ to be determined for the x'th subunit in the current step, and the set **U** of remaining subunits (if any) with unknown translation vectors. The total calculated structure factor, and the square of its amplitude is therefore:

$$Ec(\mathbf{h},\mathbf{t}_a,\mathbf{t}_b,...) = E_P(\mathbf{h}) + \Sigma_j E_{jx}(\mathbf{h},\mathbf{t}_{jx}) + \Sigma_j \Sigma_{m\in U} E_{jm}(\mathbf{h},\mathbf{t}_{jm})$$

and $$|Ec(\mathbf{h},\mathbf{t}_a,\mathbf{t}_b,...)|^2 = |E_P(\mathbf{h})|^2 + \Sigma_j \Sigma_{m\notin P} |E_{jm}(\mathbf{h},\mathbf{t}_{jm})|^2 +$$

$$2\,\mathbf{Re}[\Sigma_j \Sigma_{k<j} E_{jx}(\mathbf{h},\mathbf{t}_{jx})\, E_{kx}^*(\mathbf{h},\mathbf{t}_{kx})] + 2\,\mathbf{Re}[E_P(\mathbf{h})\, \Sigma_j E_{jx}^*(\mathbf{h},\mathbf{t}_{jx})]$$

Here the terms involving the **U** set have been replaced by their expectations because the translation vectors in **U** are unknown. The expectations of the cross-terms involving the **U** set with respect to the **t**'s in **U** are zero. The expectation of $|Ec|^2$ with respect to the translation vector $\mathbf{t}_x$ to be determined is given by the sum of the first 2 terms above. Thus the expectation $\langle Pc\rangle$, or $\langle |Ec|^2\rangle$ in reciprocal space, with respect to all the unknown translation vectors represents not just intramolecular vectors, but also known intermolecular vectors between subunits in set **P**. Therefore, subtracting this known contribution:

$$(|Ec(\mathbf{h},\mathbf{t}_a,\mathbf{t}_b,...)|^2 - \langle |Ec(\mathbf{h})|^2\rangle) =$$

$$2\,\mathbf{Re}[\Sigma_j \Sigma_{k<j} E_{jx}(\mathbf{h},\mathbf{t}_{jx})\, E_{kx}^*(\mathbf{h},\mathbf{t}_{kx})] + 2\,\mathbf{Re}[E_P(\mathbf{h})\, \Sigma_j E_{jx}^*(\mathbf{h},\mathbf{t}_{jx})]$$

The first term on the right here has already been encountered in the crystallographic translation function, and represents the contribution to the non-crystallographic translation function from the intermolecular vectors between subunit x and its crystallographic symmetry equivalents. The calculation of this term is therefore the same as before, except that the subtraction of intramolecular vectors from the target Patterson now also involves known intermolecular vectors.

The second term is new and represents the contribution from the intermolecular vectors between the known set **P** and the unknown subunit x and its equivalents. Expressing this in terms of the normalised partial structure factors for the identity asymmetric unit as before:

$$T2_{NC}(\mathbf{t}_x) = \Sigma_{\mathbf{h}} (|Eo(\mathbf{h})|^2 - \langle |Ec(\mathbf{h})|^2 \rangle) \, 2\,\mathbf{Re}[E_{\mathbf{P}}(\mathbf{h}) \Sigma_j E_x^*(\mathbf{h}.\mathbf{A}_j) \exp(-i2\pi \mathbf{h}.\mathbf{A}_j.\mathbf{t}_x)]$$

This is a Fourier series in $\mathbf{t}_x$ with the index vector $\mathbf{h}.\mathbf{A}_j$. Because the crystallographic and non-crystallographic contributions have different index vectors, it is convenient to perform the Fourier transforms separately and then sum them in real space.

6. References.

Brünger, A.T. (1990). *Acta Cryst.* **A46**, 46-57.

Cooper, J.B., Khan, G., Taylor, G., Tickle, I.J. and Blundell, T.L. (1990). *J. Mol. Biol.* **214**, 199-222.

Crowther, R.A. (1972). In *The Molecular Replacement Method* (Rossmann, M.G., ed.), pp.173-178, Gordon and Breach, New York.

Crowther, R.A. & Blow, D.M. (1967). *Acta Cryst.* **23**, 544-548.

Driessen, H.P.C., Bax, B., Slingsby, C., Lindley, P.F., Mahadevan, D., Moss, D.S. & Tickle, I.J. (1991). *Acta Cryst.* **B47**, 987-997.

Harada, Y., Lifchitz, A., Berthou, J. & Jolles, P. (1981). *Acta Cryst.* **A37**, 398-406.

Navaza, J. (1987). *Acta Cryst.* **A43**, 645-653.

Navaza, J. (1990). *Acta Cryst.* **A46**, 619-620.

Rius, J. & Miravitlles, C. (1986). *Acta Cryst.* **A42**, 402-404.

Stein, P.E., Leslie, A.G.W., Finch, J.T. & Carrell, R.W. (1991). *J. Mol. Biol.* **221**, 941-959.

Tickle, I.J. (1985). In *Proceedings of the Daresbury Study Weekend* (Machin, P.A., ed.), pp.22-26, SERC, Daresbury.

Yeates, T.O. and Rini, J.M. (1990). *Acta Cryst.* **A46**, 352-359.

MOLECULAR REPLACEMENT REAL SPACE AVERAGING

by

MICHAEL G. ROSSMANN*, ROBERT MCKENNA*, LIANG TONG*, DI XIA*,
JIN-BI DAI*, HAO WU*, HOK-KIN CHOI*, DAN MARINESCU†
AND ROBERT E. LYNCH†
*Department of Biological Sciences, Purdue University
West Lafayette, Indiana 47907, USA
†Departments of Computer Science and Mathematics, Purdue University
West Lafayette, Indiana 47907, USA

ABSTRACT

Structure determination of macromolecules often depends on phase improvement and phase extension by use of real space averaging of electron density related by noncrystallographic symmetry. Although techniques for such procedures have been described previously (1,2), modern computer architecture and experience with these methods have suggested changes and improvements.

Two unit cells are considered: (i) the p-cell corresponding to the actual crystal structure(s) being determined (there would be more than one of these if the molecule crystallizes in more than one crystal form) and (ii) the h-cell corresponding to the molecule in a standard orientation with respect to which the molecular symmetry axes are defined. Averaging can proceed entirely within the p-cell, referring to the h-cell only in as far as knowledge of the molecular symmetry is required. It is also possible to place the averaged molecule back into the h-cell, where it can be used to re-define the molecular envelope or for displaying a suitably chosen asymmetric unit of the molecule.

Techniques are discussed for automatically selecting a molecular envelope which is consistent with packing considerations within the p-cell and which retains the symmetry of the molecular point group. The electron density map to be averaged is divided into bricks for storage in virtual memory. Roughly as many bricks as there are noncrystallographic asymmetric units per crystallographic asymmetric unit need to be retained in memory at one time. This procedure minimizes paging problems and avoids double sorting. Use of 8-point interpolation permits storing the map at grid points separated by no more than $\frac{1}{2.5}$ of the resolution limit to obtain rapid convergence.

1. INTRODUCTION

Iterative phase improvement by means of electron density averaging of molecules related by noncrystallographic symmetry (for definition of this term see Rossmann (3) or Rossmann (4)) is now a frequent tool for phase improvement and phase extension to higher resolution. Whenever a molecule exists more than once either in the same unit cell or in different unit cells, then error in the molecular electron density distribution due to error in phasing can be reduced by averaging of the various molecular copies. The number of such copies, N, is referred to as the noncrystallographic redundancy. As the noncrystallographic symmetry is, by definition, only local (often pertaining to a particular molecular center), there are holes and gaps between the averaged density which presumably are solvent space between molecules. Thus, the electron density can be improved both by averaging electron density and by flattening the density between molecules. Phases calculated by Fourier back-transforming the improved density should be more accurate than the original phases. Hence, the observed structure amplitudes (suitably weighted) can be associated with the improved phases and a new and

improved map can be calculated. This, in turn, can again be averaged until convergence has been reached and the phases no longer change. In addition, the back-transformed map can be used to compute phases just beyond the extremity of the resolution of the terms used in the original map. The resultant amplitudes will not be zero because the map had been modified by averaging and solvent flattening. Thus, phases can be gradually extended and improved starting from a very low resolution approximation to the molecular structure. This procedure, sometimes referred to as molecular replacement averaging, was first implemented in reciprocal space (5-7) and more recently in real space (1,2,8).

Early examples of such a procedure for phase improvement are the structure determinations of α-chymotrypsin (9), lobster glyceraldehyde-3-phosphate dehydrogenase (10), hexokinase (11), tobacco mosaic virus disk protein (12,13), the influenza virus hemagglutinin spike (14), tomato bushy stunt virus (15) and southern bean mosaic virus (16). Early examples of phase extension, using real space electron density averaging, were the study of glyceraldehyde-3-phosphate dehydrogenase (17), hemocyanin (18) and human rhinovirus 14 (19). Since then, this method has been used in numerous virus structure determinations (e.g. 20-22) with the phase extension being initiated from ever lower resolution.

A popular computer program for real space averaging was written by Gerard Bricogne (8). Another program has been described by Johnson (2). Both programs were based on a double sorting procedure. Bricogne (1) had suggested that, with interpolation between grid points using linear polynomials, it was necessary to sample electron density at grid intervals finer than one-sixth of the resolution limit of the Fourier terms that were used in calculating the map. With the availability of more computer memory, it was possible to store much of the electron density, thus avoiding the time-consuming sorting operations (20,23). Simultaneously, the storage requirements could be drastically reduced by using interpolation with quadratic polynomials. While the latter required a little extra time, this was far less than would have been needed for sorting. Furthermore, it was found that Bricogne's estimate for the fineness of the map storage grid was too pessimistic, even for linear interpolation, which works well to about $\frac{1}{2.5}$ of the resolution limit of the map.

In addition to changes in strategy brought about by computers with larger memories, experience has been gained in program requirements of real space averaging for phase determination. These have been combined in a new program used recently in the structure determinations of ϕX174 (24), Sindbis core protein (25), Feline Parvovirus (Agbandje, Parrish and Rossmann, unpublished results) and Coxsackie B_3 virus (Bibler, Tong and Rossmann, unpublished results). Here we give a simple description of the program.

2. THE p- AND h-CELLS

It is useful to define two types of unit cells:

i. the "p-cell" is the unit cell of the unknown crystal structure and is associated with fractional coordinates $\underline{y}$ and lattice translations $\underline{a}_p$, $\underline{b}_p$, $\underline{c}_p$.
ii. the "h-cell" is the unit cell with respect to which the noncrystallographic axes of the molecule (or particle) are to be defined in a standard orientation and is associated with fractional coordinates $\underline{x}$ and lattice translations $\underline{a}_h$, $\underline{b}_h$, $\underline{c}_h$.

Since the averaged molecule is to be placed into all crystallographically related positions in the p-cell, it is essential to know the envelope which encloses a single molecule. Care must be taken that the envelopes[1] from neighboring molecules in the p-cell do not overlap. The remaining space between the limits of the envelopes of the variously placed molecules in the

[1] In this paper, "envelope" will be used to describe the external surface or boundary of a molecule, while "mask" will be used to denote the three-dimensional distribution of grid points that have been assigned to be within the molecular surface.

p-cell can be taken to be solvent and, hence, flattened, a useful physical assumption to help phase determination.

The h-cell must be chosen at least as large as the largest dimension of the molecule. In general, it is convenient to define the h-cell with $a_h = b_h = c_h$ and $\alpha = \beta = \gamma = 90°$, while placing the molecular center at $\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$. For example, if the molecule is a viral particle with icosahedral symmetry, the standard orientation can be defined by placing the twofold axes to correspond to the h-cell unit cell axes, a procedure which can be done in one of two ways. It will be necessary to know how the molecule (or particle) in the h-cell is related to the "reference" molecule in the p-cell. The known p-cell crystallographic symmetry then permits the complete construction of the p-cell structure from whatever is the current h-cell electron density representation of the molecule.

The h-cell is used to determine a molecule in the standard orientation by averaging all the noncrystallographic units in the p-cell. Noncrystallographic symmetry is true in general only locally, within the confines of a molecular envelope. Hence, while density within a specific molecule will tend to be reinforced by the averaging procedure, the density outside the molecular boundaries will tend to be diminished. Thus, by averaging into the h-cell, the molecular envelope is revealed automatically. Indeed, the greater the noncrystallographic symmetry, the greater will be the clarity of the molecular boundary. Hence, the averaged molecule in the h-cell can be used to define automatically a molecular mask in the p-cell.

Averaging into the h-cell is also useful to display the molecule in a standard orientation (i.e. obtaining the electron density distribution on skew planes). Thus, it is possible to display the molecule, for instance, with sections perpendicular to a molecular twofold axis, and to position accurately the molecular symmetry axes. From this it is then easy to define the limits of the molecular asymmetric unit. Hence, it is possible to save a great deal of computing time by evaluating the electron density in the h-cell only at those grid points within and immediately surrounding the noncrystallographic asymmetric unit.

3. COMBINING CRYSTALLOGRAPHIC AND NONCRYSTALLOGRAPHIC SYMMETRY

Transformations will now be described which relate noncrystallographically related positions distributed among several fragmented copies of the molecule in the asymmetric unit of the p-cell and between the p-cell and h-cell.

A. General Considerations

Let $\underline{Y}$ and $\underline{X}$ be Cartesian coordinates, with units of length, in the p- and h-cells which utilize the same origin as the fractional coordinates $\underline{y}$ and $\underline{x}$, respectively. Let $[\beta_p]$ and $[\alpha_h]$ be "orthogonalization" and "de-orthogonalization" matrices in the p- and h-cells, respectively (26). Then

$$\underline{Y} = [\beta_p]\underline{y} \quad \text{and} \quad \underline{x} = [\alpha_h]\underline{X} \quad , \qquad \text{(eq. 1)}$$

$$[\alpha_p] = [\beta_p]^{-1} \quad \text{and} \quad [\alpha_h] = [\beta_h]^{-1} \quad .$$

Thus, for instance, $[\alpha_h]$ denotes a matrix that transforms a Cartesian set of unit vectors to fractional distances along the unit cell vectors $\underline{a}_h$, $\underline{b}_h$, $\underline{c}_h$.

Let the Cartesian coordinates $\underline{Y}$ and $\underline{X}$ be related by the rotation matrix $[\omega]$ and the translation vector $\underline{D}$ such that

$$\underline{X} = [\omega]\underline{Y} + \underline{D} \quad . \qquad \text{(eq. 2)}$$

Therefore, from (1) and (2)

$$\underline{X} = [\omega][\beta_p]\underline{y} + \underline{D} \quad . \qquad \text{(eq. 3)}$$

Now if $[\omega]$ represents the rotational relationship between the "reference" molecule, m = 1, in the p-cell with respect to the h-cell, then from (eq. 3)

$$\underline{X} = [\omega][\beta_p]\underline{y}_{m=1} + \underline{D} \quad ,$$

where $\underline{y}_m$ refers to fractional coordinates of the m*th* molecule in the p-cell.

Assuming there is only one molecule per asymmetric unit in the p-cell, let the m*th* molecule in the p-cell be related to the reference molecule by the crystallographic rotational, $[T_m]$, and translational, $\underline{t}_m$, operators such that

$$\underline{y}_m = [T_m]\underline{y}_{m=1} + \underline{t}_m \quad . \qquad \text{(eq. 4)}$$

For convenience all translational components will be initially neglected in the further derivations below, but they will be reintroduced in the final stages. Hence, from (eq. 3) and (eq. 4)

$$\underline{X} = \{[\omega][\beta_p][T_m^{-1}]\}\underline{y}_m \quad . \qquad \text{(eq. 5)}$$

Further, if $\underline{X}_n$ refers to the n*th* subunit within the molecule in the h-cell, and if similarly $\underline{y}_{m,n}$ refers to the n*th* subunit within the m*th* molecule of the p-cell, then from (eq. 5)

$$\underline{X}_n = \{[\omega][\beta_p][T_m^{-1}]\}\underline{y}_{m,n} \quad . \qquad \text{(eq. 6)}$$

Finally, the rotation matrix, $[R_n]$, is used to define the relationship among the N (N = 2 for a dimer, 4 for a 222 tetramer, 60 for an icosahedral virus, etc.) noncrystallographic asymmetric units of the molecule within the h-cell. Then

$$\underline{X}_n = [R_n]\underline{X}_{n=1} \quad . \qquad \text{(eq. 7)}$$

B. Averaging Within the p-cell

Consider averaging the density at N noncrystallographically related points in the p-cell and replacing that density into the p-cell. By substituting for $\underline{X}_n$ and $\underline{X}_{n=1}$ in (eq. 7) using (eq. 6), it can be shown that

$$\underline{y}_{m,n} = [T_m][\alpha_p][\omega^{-1}][R_n][\omega][\beta_p][T_m^{-1}]\underline{y}_{m,n=1} + \underline{e}_{m,n} \quad , \qquad \text{(eq. 8)}$$

where $\underline{e}_{m,n}$ is the translational element. Note that this corresponds to the following sequence of transformations: (i) placing all the crystallographically related subunits into the reference orientation with $[T_m^{-1}]$, (ii) "orthogonalizing" the coordinates with $[\beta_p]$, (iii) rotating the coordinates into the h-cell with $[\omega]$, (iv) rotating these into the reference subunit of the molecule of the h-cell with $[R_n]$, (v) rotating these back into the p-cell with $[\omega^{-1}]$, (vi) "de-orthogonalizing" in the p-cell with $[\alpha_p]$ and (vii) placing these back into each of the M crystallographic asymmetric units of the p-cell with $[T_m]$.

The translational elements, $\underline{e}_{m,n}$, can now be evaluated. Let $\underline{s}_{p,m}$ be the fractional coordinates of the center (or some arbitrary position) of the m*th* molecule in the p-cell and,

hence, $\underline{s}_{p,m=1}$ denotes the molecular center position of the reference molecule in the p-cell. If $\underline{s}_{p,m}$ is at the intersection of the molecular rotation axes, then it will be the same for all n molecular asymmetric units. It therefore follows that

$$\underline{e}_{m,n} = \underline{s}_{p,m} - [E_{m,n}]\underline{s}_{p,m=1} \qquad \text{(eq. 9a)}$$

or

$$\underline{y}_{m,n} = [E_{m,n}]\underline{y}_{m,n=1} + (\underline{s}_{p,m} - [E_{m,n}]\underline{s}_{p,m=1}) \qquad \text{(eq. 9b)}$$

where

$$[E_{m,n}] = [T_m][\alpha_p][\omega^{-1}][R_n][\omega][\beta_p][T_m^{-1}] \quad .$$

Equation (9b) can be used to find all the N noncrystallographic asymmetric units within the crystallographic asymmetric unit of the p-cell. Thus, this equation is the essential equation for averaging the density in the p-cell and replacing it into the p-cell.

C. Averaging the p-cell and Placing the Results into the h-cell

Consider averaging the density at N noncrystallographically related points in the p-cell and placing that result into the h-cell. From (eq. 1), (eq. 6) and (eq. 7) it can be shown that

$$\underline{x}_{n=1} = [\alpha_h][R_n^{-1}]\{[\omega][\beta_p][T_m^{-1}]\}\underline{y}_{m,n} \quad . \qquad \text{(eq. 10)}$$

Since it is only necessary to place the reference molecule of the p-cell into the h-cell, it is sufficient to consider the case when m = 1, in which case $[T_m^{-1}]$ is the identity matrix [I]. It then follows, by inversion, that

$$\underline{y}_{m=1,n} = [\alpha_p][\omega^{-1}][R_n][\beta_h]\underline{x}_{n=1} \quad ,$$

which corresponds to (i) "orthogonalizing" the h-cell fractional coordinates with $[\beta_h]$, (ii) rotating into the *nth* noncrystallographic unit within the molecule using $[R_n]$, (iii) rotating into the p-cell with $[\omega^{-1}]$ and (iv) "de-orthogonalizing" into fractional p-cell coordinates with $[\alpha_p]$.

Now if $\underline{s}_h$ is the molecular center in the h-cell (usually $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$), then

$$\left.\begin{aligned} \underline{y}_{m=1,n} &= [E'_{m=1,n}]\underline{x} + (\underline{s}_{p,m=1} - [E'_{m=1,n}]\underline{s}_h) \\ \text{where} \\ E'_{m=1,n} &= [\alpha_p][\omega^{-1}][R_n][\beta_h] \quad . \end{aligned}\right\} \qquad \text{(eq. 11)}$$

Equation (11) determines the position of the N noncrystallographically related points $\underline{y}_{m=1,n}$ in the p-cell, whose average value is to be placed at $\underline{x}$ in the h-cell.

4. DEFINING THE MOLECULAR MASK IN THE p-CELL

The crystallographic asymmetric unit is likely to contain bits and pieces of molecules centered at various positions in the unit cell and neighboring unit cells. The number, M, of such molecules can be estimated by generating all centers, derived from the given position of

the center for the reference molecule, $\underline{s}_{p,n=1}$, and then determining whether a molecule of radius R_{out} would impinge on the crystallographic asymmetric unit within the defined boundaries. Here, R_{out} is a liberal estimate of the molecular radius. The corresponding rotation matrices $E_{m,n}$ and translation vectors $\underline{e}_{m,n}$ can then be computed from equations (8) and (9a).

It is then necessary to associate each grid point within the p-cell crystallographic asymmetric unit to a specific molecular center, or to solvent. Any grid point whose distance from all M centers is greater than R_{out} can immediately be designated as being in the solvent region. For other grid points, it is necessary to examine the corresponding h-cell density. Transfer of the electron density $\rho(\underline{x})$ from the h-cell to the p-cell is a way to obtain an initial structure. For the purpose of determining a suitable mask it is useful to evaluate a modified electron density $<\rho(\underline{x})>$ (see below) for the grid points immediately around $\underline{x}$ in the h-cell.

A parameter "CRIT" can be defined to establish the distribution of grid points that are within the molecular envelope. When the modified electron density $<\rho(\underline{x})>$ is less than CRIT, the corresponding grid point at $\underline{y}$ is taken to be in solvent. Otherwise, when $<\rho(\underline{x})>$ exceeds CRIT, the grid point at $\underline{y}$, is assigned to that molecule which has the largest $<\rho(\underline{x})>$. However, the grid point at $\underline{y}$ may be within the compass of more than one molecular center. In that case the grid point is assigned to that molecule which relates to the largest modified density $<\rho(\underline{x})>$ in the h-cell. It is useful to keep a record of the number of grid points where such a conflict occurs. If the percentage of such grid points with respect to the total number of grid points is large (say greater than 1%), it probably means that the value of CRIT has been chosen too low, or that the molecular boundary is still far from clear, or that the function used to define $<\rho(\underline{x})>$ was badly chosen. In the case of a virus, it can be useful to define an inner radius, R_{int}, inside which the density is assumed to belong to the nucleic acid core (rather than external solvent) if $<\rho(\underline{x})>$ is less than CRIT. Another, even smaller, radius, R_{core}, can be defined inside which the grid point will be assumed to correspond to nucleic acid irrespective of the value of $<\rho(\underline{x})>$ (Table 1; Fig. 1). Grid points outside the molecular envelope can be set to the average solvent or nucleic acid density, set to zero or left unchanged.

In most circumstances, Fourier summations to calculate electron density do not include the F(000) term. Hence, the sum of the positive density and negative density within the unit cell must be exactly opposite so that the total density is zero. Within the molecule the density will, therefore, tend to vary from near maximum positive (where there are atoms) to near minimum negative (between atoms separated by van der Waals distances). Thus, the sum of the density within a molecule is also likely to be close to zero. It follows that, in general, the solvent density between molecules will be close to zero. What, therefore, identifies grid points within the molecule is that their surrounding density is changing far more rapidly and to a far greater extent than those without. Thus, the modified electron density $<\rho(\underline{x})>$, used to determine the molecular envelope, can be defined as either the mean absolute density or the maximum absolute density of grid points within roughly a small radius around $\underline{x}$ ("the modifying radius"). Such a procedure has long been used by Wang (27) who uses the average of the positive density rather than considering absolute density. An equivalent procedure in reciprocal space has been described by Leslie (28). In defining a molecular envelope for the purpose of molecular averaging, it is important to maintain the resolution of the current density in assigning the mask boundaries. Otherwise, some grid points that should be inside the mask might be attributed to solvent or, on the other hand, the full complement of solvent density grid points may not be attained.

Another criterion for the molecular envelope is that it obeys the noncrystallographic point group symmetry. If the original h-cell electron density already possesses the molecular symmetry (e.g. icosahedral 532, 222, etc.), then the p-cell mask should also have that symmetry. However, where masks from different molecular centers conflict, the criterion based on the h-cell density $<\rho(\underline{x})>$, described above, may cause local errors in the correct molecular symmetry. Such errors can be corrected by re-imposing the noncrystallographic point group symmetry on the p-cell mask. This can be conveniently achieved by setting the

density at each grid point that was considered to be within the molecular envelope to a value of 100 and all other grid points to a density of zero. If, then, the resultant density is averaged using the same routine as is used for averaging the actual electron density of the molecule (see below), then, if the interpolated density is 100 at all noncrystallographically related points, the average density will remain 100. But, if the original grid point is near the edge of the mask, finding the density at symmetry related points may involve interpolation between density at level 100 and at level 0, giving an averaged density of less than 100. Hence, any grid point whose averaged density is below some criterion (e.g. LXTND) should be attributed to solvent (Table 2).

Table 1

ϕX174 - Generation of Mask in p-cell from h-cell Electron Density

A. p-cell:

	a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)	Space group
p-cell cell dimensions	305.6	360.8	299.5	90.0	92.89	90.0	$P2_1$
p-cell grid intervals	a/64	b/64	c/64				
p-cell asymmetric unit	$0 \le a \le 1$	$0 \le b \le \frac{1}{2}$	$0 \le c \le 1$				
Position of reference molecule	0.2505	0.2500	0.2505				

External radius R_{cut} = 170 Å, internal radius R_{int} = 122 Å, core radius R_{core} = 80 Å

Centers of particles which could impinge on the defined asymmetric unit:

Particle ID	x	y	z
A	-0.2505	-0.2500	-0.2505
B	-0.2505	-0.2500	0.7495
C (reference)	0.2505	0.2500	0.2505
D	-0.2505	0.7500	-0.2505
E	0.2505	0.2500	1.2505
F	-0.2505	0.7500	0.7495
G	0.7495	-0.2500	-0.2505
H	0.7495	-0.2500	0.7495
I	1.2505	0.2500	0.2505
J	0.7495	0.7500	-0.2505
K	1.2505	0.2500	1.2505
L	0.7495	0.7500	0.7495

Table 1 (continued)

B. h-cell:

	a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)	Space group
h-cell cell dimensions	645.1	645.1	645.1	90.0	90.0	90.0	P222
h-cell intervals	a/64	b/64	c/64				
h-cell stored	0≤a≤1	0≤b≤1	0≤c≤1				

Maximum and minimum h-cell density was 330 and 229 arbitrary units, respectively

CRIT, minimum density accepted as protein was 175 arbitrary units

Electron density modifying radius: 14.3 Å

Map resolution: 22 Å (from electron microscope reconstruction)

Particle center was at 0.5000 0.5000 0.5000

Particle orientation: twofold axes were parallel to h-cell axial directions and icosahedral symmetry generated by the following sequential operations (see Rossmann and Blow (26) for definition of polar coordinates κ, ψ, ϕ):

Operation	κ	ψ	ϕ
1	72.0	90.0	31.71747
2	180.0	108.0	58.28253
3	180.0	144.0	58.28253
4	120.0	54.73561	45.0000

C. Orientational relation between p- and h-cells:

$$[\omega]^{-1} = \begin{pmatrix} 0.9898 & -0.1361 & -0.0403 \\ 0.1290 & 0.9809 & -0.1455 \\ 0.0593 & 0.1387 & 0.9886 \end{pmatrix}$$

where

$\underline{X} = [\omega]\underline{Y}$

D. Mask creation in p-cell:

	Protein	Solvent	Acid	Total
Mean density	224	113	89	
Number of grid points	60,774	28,830	49,660	139,264
Percentage of cell volume	44	20	36	100

Number of grid points where there was conflict between molecules at different centers was 2779, corresponding to 2% of the cell volume.

The mask may be found to have single or adjacent grid points that have been assigned as solvent but yet are completely buried within the mask for a given molecule. This may be the result of a slightly poor value for CRIT, error in the h-cell density, or it might be a real pocket within the molecule. Whatever is the case, it is probably best to assign such grid points to the molecular mask which surrounds these points. In practice, it is possible to examine each solvent (or nucleic acid associated) grid point and determine the nature of the six nearest neighboring grid points. A criterion IFILL can be set such that if IFILL of the 6 grid points surrounding a grid point previously assigned to be solvent belong to one particular molecule, then the solvent grid point should be reassigned (IFILL = 6 or 5 are useful values) (Table 3).

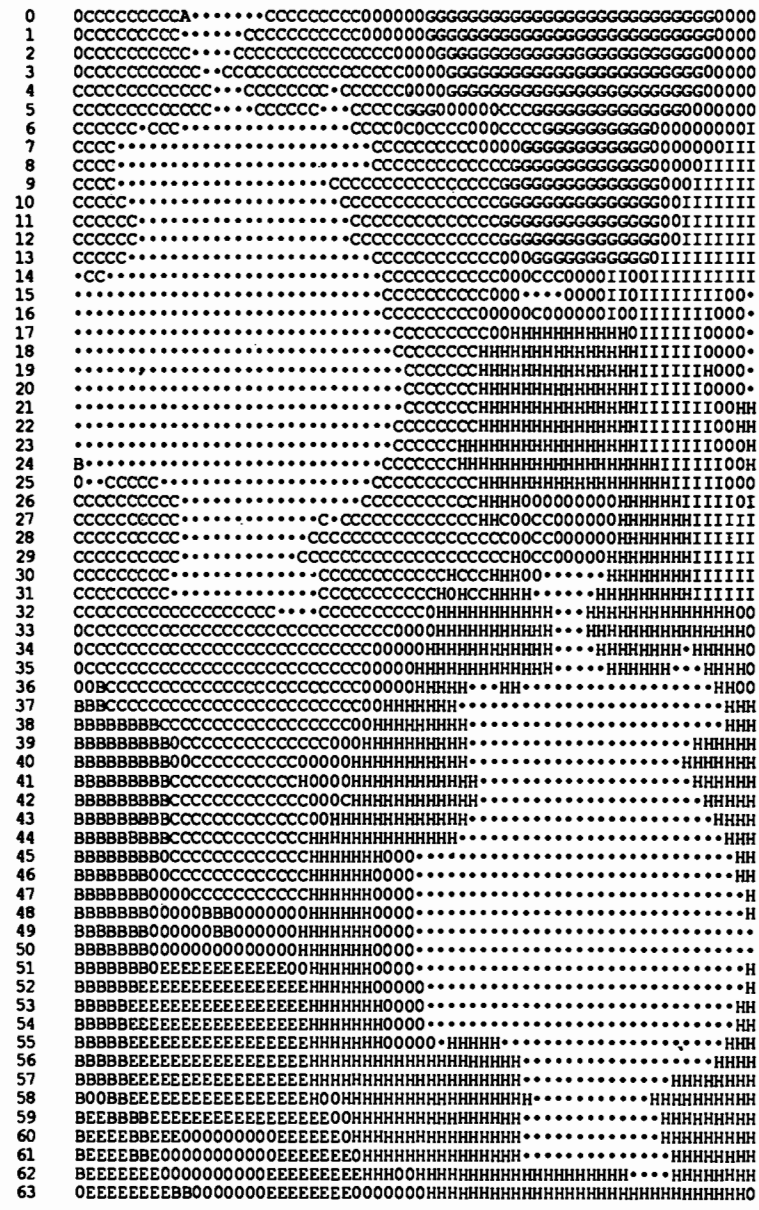


Figure 1. Section y = 0 of the p-cell crystallographic asymmetric unit for ϕX174 showing solvent regions (0), nucleic acid regions (•), the reference molecule mask (C) and other molecules that have components within the chosen asymmetric unit. Grid points associated with these molecules are labeled A, B, D, E, ... (see Table 1A).

Table 2

Enforcing Noncrystallographic Symmetry on the Mask

Average Mask Density	Number of Grid Points	r.m.s. Deviation of Mask Density
0-25	30,413	7.2
25-75	50,243	35.9
75-100	50,475	11.0
Overall	131,131	19.7

All grid points where the average was greater than LXTND = 50 were accepted to be inside the mask. This produced:

	Protein	Solvent	Nucleic Acid	Total
Number of grid points	68,548	50,808	19,908	139,264
Percentage of cell volume	49	37	14	100

Notes:
1. Results relate to mask created with conditions shown in Table 1.
2. Initially, the grid points were set to a density of 100 inside the mask and to 0 outside.

Table 3

The Number of Empty Holes in the Mask

Number of Surrounding Protein Grid Points	Maximum Number of Surrounding Grid Points Belonging to the Same Mask						Total
	1	2	3	4	5	6	
0	0	0	0	0	0	0	51020
1	7534						7534
2	955	4875					5830
3	58	1219	3326				4603
4	2	298	527	480			1307
5	0	24	220	72	37		353
6	0	2	37	20	10	0	69

Notes:
1. Results relate to mask created with conditions shown in Tables 1 and 2.
2. All solvent grid points which were surrounded by IFILL = 5 grid points belonging to a mask with the same molecular center were changed to belong to the appropriate mask (boxed in the table).

5. FINDING THE AVERAGED DENSITY IN THE p-CELL

Associated with each grid point in the p-cell asymmetric unit will be (i) the value of m ($1 \le m \le 26$) designating which molecular center is to be associated with that grid point (special values of m are 27 (solvent) or 28 (nucleic acid)) and (ii) the p-cell electron density at that point. A maximum value for m of 26 was chosen to permit easy representation of the mask in terms of the alphabet (Fig. 1).

The grid points within the asymmetric unit are then examined one at a time. If the grid point is solvent or nucleic acid, it can be (by specifying an input parameter) set to zero, to the mean solvent or nucleic acid density, or left unchanged. If the grid point is within the mask, it can be (by specifying another input parameter) set to zero, left unchanged, or averaged among the N noncrystallographically related equivalent positions belonging to molecule m. Leaving the density unchanged permits phase improvement due to solvent flattening alone.

Random access to map storage is a problem which can be solved by a double sorting procedure (1,2,8). With a virtual memory system, a different algorithm can be applied. Consider the map to be divided into bricks where each brick occupies about one page of disk memory. For the Cyber 205 computer that means each brick contained $16 \times 16 \times 64$ densities and one page contained two bricks. The N crystallographically equivalent points corresponding to the reference grid point will, in general, fall on N pages (or at worst some will be on page boundaries). All these N pages can be resident in memory at the same time. As the reference grid position moves systematically towards the edge of its brick, so the other equivalent grid points on the other pages also move towards the edge of their bricks. Only when the original grid point moves from one brick to another are the N equivalent grid points likely to move from one brick to another. By using sequentially reference grid points within a brick, then this reduces the number of times the other N equivalent positions fall outside the N bricks currently in memory.

The N noncrystallographically equivalent non-integral grid points can be computed from (eq. 9). Some of these will lie outside the crystallographic asymmetric unit. These will, therefore, have to be operated on by unit cell translations and crystallographic symmetry operations to bring them back into the asymmetric unit before the corresponding interpolated density can be calculated. This also can be a time-consuming calculation, but by a suitable combination of operations those points that need this kind of manipulation can be gathered into a single vector for rapid vector, rather than scalar, calculations.

It is useful to keep a record of the scatter among the electron densities that were averaged. Let $\underline{x}_i$ ($i = 1, 2, ..., N$) be a set of positions at which the electron densities are averaged. Then, the mean density to be stored is

$$\overline{\rho(\underline{x})} = \frac{\sum_{i=1}^{N} \rho(\underline{x}_i)}{N}$$

and the r.m.s. deviation from the mean is

$$\sigma(\rho) = \sqrt{\frac{\sum_i \left(\overline{\rho(\underline{x})} - \rho(\underline{x}_i)\right)^2}{N}}$$

The mean value of $\sigma(\rho)$, denoted by $\langle\sigma(\rho)\rangle$, is a useful criterion for refining the particle position and orientation (Table 4). Each rotational and translational parameter is varied one at a time and a search can be conducted for the values of the parameters which give the smallest $\langle\sigma(\rho)\rangle$. This process can be speeded up considerably by averaging into the h-cell and

considering only a limited volume of the molecule, as all parts of the molecule should give the same results for the best position and orientation.

6. AVERAGING INTO THE h-CELL

The procedure is very similar as for p-cell averaging, except that the rotation and translation matrices are given by (eq. 11). Furthermore, no mask is required as all the averaging into the h-cell (from p-cell electron density) can be with respect to the reference molecule centered at $\underline{s}_{p,m=1}$ in the p-cell. Each grid point is taken in turn in the h-cell. The electron density at any grid point that is further away from $\underline{s}_h$ than R_{out} is set to zero. Other grid point positions are expanded into the N equivalent positions in the p-cell surrounding $\underline{s}_{p,m=1}$. The interpolated density is then found, averaged over the N equivalent positions and stored at the original h-cell grid point in successive sections, in the same way as in the p-cell averaging. As in averaging within the p-cell, a record is kept of $\langle\sigma(\rho)\rangle$ as a function of $\langle\rho(\underline{x})\rangle$ (Table 4). In general, the local noncrystallographic symmetry is valid only within the molecule. Hence, the h-cell density will show the molecular envelope and can be used to recompute an improved p-cell density mask. The rate of build up of signal within the molecule should be roughly proportional to N, while the rate outside the molecule should be proportional to about $\sqrt{N}$.

Table 4

Mean r.m.s. Scatter Between Noncrystallographically Related Points
(Example taken from ϕX174 structure determination.)

	Density Derived from an Electron Microscopy Image at 25 Å Resolution		Density Derived from a 3.3 Å Crystal Structure	
$\langle\rho_8\rangle$	n	$\langle\sigma(\rho_8)\rangle$	n	$\langle\sigma(\rho_8)\rangle$
-375 to -325	1	44.7	0	0.0
-325 to -275	16	44.4	0	0.0
-275 to -225	22	39.5	41	31.4
-225 to -175	81	34.9	3,493	25.5
-175 to -125	299	34.7	65,049	20.5
-125 to -75	1,119	33.1	290,025	17.7
-75 to -25	16,617	34.7	661,386	15.0
-25 to 25	33,818	46.9	1,016,274	12.8
25 to 75	6,008	31.9	344,620	16.3
75 to 125	4,512	32.0	215,036	18.9
125 to 175	3,050	32.1	146,690	22.1
175 to 225	1,562	32.6	58,155	26.3
225 to 275	542	33.4	6,032	32.2
275 to 325	213	35.6	227	40.6
325 to 375	33	34.7	9	46.8

Notes:
$\langle\rho_8\rangle$ is mean density based on 8-point interpolation.
n is number of grid points with $\langle\rho_8\rangle$ in given range.
$\langle\sigma(\rho_8)\rangle$ is r.m.s. deviation from ρ_8 among noncrystallographic asymmetric points averaged over all points in the mask.

7. INTERPOLATION

Let the position at which the density is to be interpolated have the fractional grid coordinates Δx, Δy, Δz within the box of surrounding grid points. Let 0,0,0 be the point at $\Delta x = 0$, $\Delta y = 0$, $\Delta z = 0$. Other grid points will then be at 100, 010, 001, etc., with the point diagonally opposite the origin being at 111. Then

$$\rho_8(\Delta x,\Delta y,\Delta z) = \rho_{000} + \Delta x(A_{100} + \Delta y(A_{110} + \Delta z A_{111})) + \Delta y(A_{010} + \Delta z A_{011}) + \Delta z A_{001} \quad \text{(eq. 12)}$$

where

$$A_{100} = \rho_{100} - \rho_{000},$$
$$A_{010} = \rho_{010} - \rho_{000},$$
$$A_{001} = \rho_{001} - \rho_{000},$$
$$A_{110} = \rho_{000} + \rho_{110} - \rho_{100} - \rho_{010},$$
$$A_{011} = \rho_{000} + \rho_{011} - \rho_{010} - \rho_{001},$$
$$A_{101} = \rho_{000} + \rho_{101} - \rho_{001} - \rho_{100},$$
$$A_{111} = \rho_{100} + \rho_{010} + \rho_{001} + \rho_{111} - \rho_{000} - \rho_{101} - \rho_{011} - \rho_{110}.$$

In the structure determination of ϕX174 (24), convergence in just 3 or 4 cycles was achieved with this 8-point interpolant even when the grid spacing was as large as Resolution/2.5, although fewer cycles are required with smaller grid point intervals.

8. THE FAST, DOUBLE-INTERPOLATION, PROCEDURE

The single interpolation procedure described above requires the determination of the averaged electron density at each grid point within the crystallographic asymmetric unit that is within the mask. For each grid point, one interpolation is required at each of the N non-integral grid points whose densities are to be averaged. In this procedure, there is some wastage, because the new averaged density needs to be determined only within the noncrystallographic asymmetric unit, or $\left(\frac{1}{N}\right)th$ of the crystallographic asymmetric unit. Once this has been accomplished, all other grid point densities can be determined by folding them back into the noncrystallographic asymmetric unit. Thus, one interpolation takes the place of N interpolations for $\left(\frac{1}{N-1}\right)th$ of all the grid points within the mask. The disadvantage is that it will be necessary to interpolate a second time between the grid points within the designated noncrystallographic asymmetric unit to find the density corresponding to a grid point outside the noncrystallographic asymmetric unit. As every interpolation tends to degrade the density values to some extent, it will be necessary to use a somewhat finer grid for this, otherwise faster, procedure.

The first step in the fast double-interpolation procedure is to tag those grid points within the noncrystallographic asymmetric unit. This can be achieved by generating all N noncrystallographically related positions for a given grid point and folding these back into the defined crystallographic asymmetric unit. The three components of the coordinates $\underline{y}_{m,n}$ are then packed into a single word. The coordinates which give (say) the smallest number can be arbitrarily assigned as the noncrystallographic asymmetric unit. No matter where the initial position is, the N points which are generated will always bear the same relation. Thus, this provides a unique procedure for selecting the noncrystallographic asymmetric unit. All points that are outside this unit are tagged by turning on a specific bit associated with each grid point in the mask. In addition, the identity of the symmetry operator (one of NM) which relates a given grid point to a site within the noncrystallographic asymmetric unit is retained.

If the mask has not been tagged, then averaging proceeds for all grid points within the mask. If the mask has been tagged, averaging will be performed only at the untagged points (namely those within the noncrystallographic asymmetric unit). In a second pass over the mask, all the tagged grid points (those outside the noncrystallographic asymmetric unit) are folded back into the noncrystallographic asymmetric unit by using the previously identified appropriate symmetry operation. The averaged density can now be found by interpolation. Some unfavorable interpolation may occur at points that border the edges of the noncrystallographic asymmetric unit as the interpolation may involve densities that lie outside the noncrystallographic asymmetric unit where the average density has not yet been determined. In practice, a factor of six was achieved in computing speed but at a considerable loss of precision.

9. COMBINING DIFFERENT CRYSTAL FORMS

It frequently occurs that a molecule crystallizes in a variety of different crystal forms (e.g. hexokinase (11), influenza virus neuraminidase spike (29), histocompatibility antigen HLA (30) CD4 receptor (31) and Sindbis core protein (32,33)). It is then advantageous to average between the different crystal forms. This is best achieved by averaging each crystal form independently into a standard orientation in the h-cell (if the redundancy is N = 1 for a given crystal form, then this amounts to simply producing a skewed presentation of the p-cell in the h-cell environment). The different results, now all in the same h-cell orientation, can be averaged. Care, however, must be taken to put equal weight on each molecular copy.

With the h-cell density improved by averaging among different crystal forms, it can now be replaced into the different p-cells. These p-cells can then be back-transformed in the usual manner to obtain a better set of phases. These, in turn, can be associated with the observed structure amplitudes for each p-cell structure and the cycle can be repeated.

10. TEST EXAMPLES

The present program has been used in the structure determination of ϕX174 (24) and Sindbis virus core protein (33). It is being used to improve phases for monoclinic canine parvovirus as well as tetragonal canine parvovirus (34) in order to define a better envelope. The program has also been used in tests for assessing convergence rates in the presence of different grid step sizes with phosphoglucomutase diffraction data. These crystals belong to

Table 5

Convergence of Phase Refinement for Phosphoglucomutase

	Mean Correlation Coefficients	
Cycle	Grid Step = Resolution/3.0	Grid Step = Resolution/2.1
1	0.4675	0.4082
2	0.6553	0.5417
3	0.7232	0.5815
4	0.7571	0.6000
5	0.7791	0.6103
6	0.7941	0.6131
7	0.8051	0.6152
8	0.8130	0.6199
9	0.8191	0.6212
10	0.8247	0.6208
11	0.8299	0.6204

space group $P4_12_12$ and contain one dimer in the asymmetric unit. The mask was determined from the known molecular structure (35). Eleven cycles were performed on a 5 Å resolution map using grid spacing of (a) $\Delta a = \Delta b = 1.7$ Å and $\Delta c = 1.6$ Å and (b) $\Delta a = \Delta b = 2.4$ Å, $\Delta c = 2.4$ Å, corresponding to about $\frac{1}{3.0}$ and $\frac{1}{2.1}$ of the resolution. The results, using 8-point interpolation, are given in Table 5. While there was excellent convergence with the finer grid step size, the coarser grid size initially converged more slowly and eventually even slightly diverged. As the theoretical limit for interpolation in reciprocal space is one-half a reciprocal lattice point, it is indeed surprising that convergence was as good as it was for the very coarse grid size. Presumably in the initial stages of refinement phases for the lower resolution reflections converged, but later the phases of the higher resolution reflections, near the edge of resolution, diverged.

The program was written in FORTRAN for the Purdue University Cyber 205 computer. The 64-bit word lengths and relative lack of storage required dependence on special packing instructions, special vector processing instructions, and dependence on the virtual memory system. The program has now been adapted to an IBM RISC 6000 workstation model 540 where the code is in standard FORTRAN. The program has also been ported to an Intel-cube distributed memory multiprocessor network. Here each node has only a fairly small memory (16 Mbytes). The brick concept is particularly useful as each node can then work on modifying the density within a few bricks only. These require only a limited number of other bricks for density averaging. The current programs are available from the corresponding author.

12. ACKNOWLEDGMENTS

This manuscript is mostly a copy of a paper in press in Acta Crystallographica.

We are most grateful for much helpful advice from David Seaman of the Purdue University Computer Center at Purdue University. We thank Sharon Wilder and Helene Prongay for help in preparation of the manuscript. The work was supported by grants from the National Institutes of Health and the National Science Foundation.

REFERENCES

1. Bricogne, G., Acta Crystallogr. *A32*, (1976) 832-847.
2. Johnson, J. E., Acta Crystallogr. *B34*, (1978) 576-577.
3. Rossmann, M. G., The Molecular Replacement Method (1972) New York: Gordon and Breach.
4. Rossmann, M. G., Acta Crystallogr. *A46*, (1990) 73-82.
5. Rossmann, M. G. and Blow, D. M., Acta Crystallogr. *16*, (1963) 39-45.
6. Main, P., Acta Crystallogr. *23*, (1967) 50-54.
7. Crowther, R. A., Acta Crystallogr. *B25*, (1969) 2571-2580.
8. Bricogne, G., Acta Crystallogr. *A30*, (1974) 395-405.
9. Matthews, B. W., Sigler, P. B., Henderson, R. and Blow, D. M., Nature (London) *214*, (1967) 652-656.
10. Buehner, M., Ford, G. C., Moras, D., Olsen, K. W. and Rossmann, M. G., J. Mol. Biol. *82*, (1974) 563-585.
11. Fletterick, R. J. and Steitz, T. A., Acta Crystallogr. *A32*, (1976) 125-132.
12. Champness, J. N., Bloomer, A. C., Bricogne, G., Butler, P. J. G. and Klug, A., Nature (London) *259*, (1976) 20-24.
13. Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R. and Klug, A., Nature (London) *276*, (1978) 362-368.
14. Wilson, I. A., Skehel, J. J. and Wiley, D. C., Nature (London) *289*, (1981) 366-373.
15. Harrison, S. C., Olson, A. J., Schutt, C. E., Winkler, F. K. and Bricogne, G., Nature (London) *276*, (1978) 368-373.
16. Abad-Zapatero, C., Abdel-Meguid, S. S., Johnson, J. E., Leslie, A. G. W., Rayment, I., Rossmann, M. G., Suck, D. and Tsukihara, T., Nature (London) *286*, (1980) 33-39.

17. Argos, P., Ford, G. C. and Rossmann, M. G., Acta Crystallogr. *A31*, (1975) 499-506.
18. Gaykema, W. P. J., Hol, W. G. J., Vereijken, J. M., Soeter, N. M., Bak, H. J. and Beintema, J. J., Nature (London) *309*, (1984) 23-29.
19. Rossmann, M. G., Arnold, E., Erickson, J. W., Frankenberger, E. A., Griffith, J. P., Hecht, H. J., Johnson, J. E., Kamer, G., Luo, M., Mosser, A. G., Rueckert, R. R., Sherry, B. and Vriend, G., Nature (London) *317*, (1985) 145-153.
20. Hogle, J. M., Chow, M. and Filman, D. J., Science *229*, (1985) 1358-1365.
21. Hosur, M. V., Schmidt, T., Tucker, R. C., Johnson, J. E., Gallagher, T. M., Selling, B. H. and Rueckert, R. R., Proteins, *2*, (1987) 167-176.
22. Tsao, J., Chapman, M. S., Agbandje, M., Keller, W., Smith, K., Wu, H., Luo, M., Smith, T. J., Rossmann, M. G., Compans, R. W. and Parrish, C. R., Science *251*, (1991) 1456-1464.
23. Luo, M., Vriend, G., Kamer, G. and Rossmann, M. G., Acta Crystallogr. *B45*, (1989) 85-92.
24. McKenna, R., Xia, D., Willingmann, P., Ilag, L. L., Krishnaswamy, S., Rossmann, M. G., Olson, N. H., Baker, T. S. and Incardona, N. L. (1991). Manuscript in preparation.
25. Choi, H. K., Tong, L., Minor, W., Dumas, P., Boege, U., Rossmann, M. G. and Wengler, G. (1991). Manuscript in preparation.
26. Rossmann, M. G. and Blow, D. M., Acta Crystallogr. *15*, (1962) 24-31.
27. Wang, B. C., Meth. Enzymol. *115*, (1985) 90-112.
28. Leslie, A. G. W., Acta Crystallogr. *A43*, (1987) 134-135.
29. Varghese, J. N., Laver, W. G. and Colman, P. M., Nature (London) *303*, (1983) 35-40.
30. Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. and Wiley, D. C., Nature (London) *329*, (1987) 506-512.
31. Wang, J., Yan, Y., Garrett, T. P. J., Liu, J., Rodgers, D. W., Garlick, R. L., Tarr, G. E., Husain, Y., Reinherz, E. L. and Harrison, S. C., Nature (London) *348*, (1990) 411-418.
32. Boege, U., Cygler, M., Wengler, G., Dumas, P., Tsao, J., Luo, M., Smith, T. J. and Rossmann, M. G., J. Mol. Biol. *208*, (1989) 79-82.
33. Tong, L., Choi, H. K. and Rossmann, M. G. (1991). Manuscript in preparation.
34. Wu, H. and Rossmann, M. G. (1991). Unpublished results.
35. Lin, Z., Konno, M., Abad-Zapatero, C., Wierenga, R., Murthy, M. R. N., Ray, W. J., Jr. and Rossmann, M. G., J. Biol. Chem. *261*, (1986) 264-274.

Molecular replacement with X-PLOR: PC-refinement and free R value.

Axel T. Brünger

The Howard Hughes Medical Institute and
Department of Molecular Biophysics and Biochemistry,
Yale University,
New Haven, CT 06511

1 Introduction

In macromolecular crystallography, the initial determination of phases by molecular replacement (MR)[1, 2, 3, 4, 5] is often attempted if the structure of a similar or homologous macromolecule is known ("search model"). MR involves the placement (i.e., rotation and translation) of the search model in the unit cell of the target crystal in order to obtain the best agreement between calculated and observed diffraction data. The optimally placed search model is used to obtain initial phases for crystallographic structure fitting and refinement.

If there is one molecule in the crystallographic asymmetric unit, then three positional and three angular parameters fully describe the placement of the search model in the unit cell of the target crystal. This six-dimensional search can be reduced to a sequence of a three-dimensional angular search using a "rotation function" followed by a three-dimensional positional search using a "translation function". This procedure assumes that the highest peak of the rotation function yields the correct orientation. Examples are known where this is not true. Due to advances in computer technology, multi-dimensional search strategies with more than three parameters are no longer beyond available computational resources anymore.

In this paper we will address two common problems in molecular re-

placement: the failure of the rotation and/or translation functions due to an inaccurate atomic model, and the discrimination between correct and incorrect molecular replacement solutions.

1.1 *PC*-refinement

MR can be viewed as a nonlinear optimization problem with the aim of maximizing the standard linear correlation coefficient *PC*[6], given by

$$PC(\Omega,T) = \frac{< |E_{obs}|^2|E_m(\Omega,T)|^2 - < |E_{obs}|^2 >< |E_m(\Omega,T|^2 > >}{\sqrt{< |E_{obs}|^4 - < |E_{obs}|^2 >^2 >< |E_m(\Omega,T)|^4 - < |E_m(\Omega,T)|^2 >^2 >}}. \quad (1)$$

The symbols <> denote an averaging over the set of observed reflections. E_{obs} denote the normalized observed structure factors and $E_m(\Omega,T)$ denote the normalized structure factors of the search model oriented according to the rotation matrix Ω and positioned according to the translation vector T. This particular correlation coefficient between observed and computed diffraction data is a means for minimizing phase error[7]. Instead of maximizing *PC*, it is equivalent to minimize a function of the type

$$E_{xray} = c(1 - PC). \quad (2)$$

Refinement is carried out against the negative correlation coefficient *PC* since minimization algorithms normally locate minima as opposed to maxima; a minimum of $-PC$ corresponds to a maximum of *PC* which is what one is really aiming at.

Recently MR has been generalized by introducing additional parameters into *PC* , i.e.,

$$PC(p,\Omega,T) = \frac{< |E_{obs}|^2|E_m(p,\Omega,T)|^2 - < |E_{obs}|^2 >< |E_m(p,\Omega,T)|^2 > >}{\sqrt{< |E_{obs}|^4 - < |E_{obs}|^2 >^2 >< |E_m(p,\Omega,T)|^4 - < |E_m(p,\Omega,T)|^2 >^2 >}}. \quad (3)$$

where p denote atomic coordinates, rigid body coordinates, occupancies, temperature factors, or other generalized coordinates[8]. If a single copy of the search model is oriented according to Ω and placed in a triclinic unit cell identical in geometry to that of the crystal, expression Eq. 3 becomes independent of the

translation T, and thus can be computed after the rotation function prior to the translation function.

Analytic derivatives of PC or E_{xray} with respect to p can be computed by application of the chain rule. It is then possible to minimize E_{xray} as a function of the parameters p. In addition, E_{xray} may be combined with a geometric or empirical energy function, thus defining a hybrid energy function[9]. The refinement of a molecular replacement model prior to translation searches has been termed PC-refinement.

1.2 Generalized molecular replacement strategy

Since the mathematical details of the procedure are published elsewhere[8], here we will just summarize the main points. First, a conventional rotation function is evaluated. A large number of orientations corresponding to the highest peaks of the rotation function are selected. Here one makes the *ad hoc* assumption that the correct orientation is among this selected subset. Then a small number of parameters p are introduced that describe the most dominant differences that are expected to occur between the crystal structure and the search model. For example, this can involve the orientations and positions of large rigid groups of atoms (secondary structural elements or protein domains). For n rigid bodies this requires $6n - 3$ parameters where three parameters specifying the arbitrary overall position have been subtracted. Next, the most important step of the strategy follows, which consists of refinements of p against the negative correlation coefficient PC for each selected orientation. The major difference to standard least-squares refinement is that PC-refinement is carried out in a triclinic P_1 cell, that is, without crystallographic symmetry. It should be noted that a large number of the PC-refinements are actually carried out for incorrect orientations of the search model; only the PC-refinements starting close to the correct orientation are expected to yield a relatively large correlation coefficient after PC-refinement. Furthermore, one expects the search model to become more accurate in the latter cases. A necessary but not sufficient condition for the correct solution of the crystal structure is that PC assumes a maximum. One thus selects another subset of orientations that have produced the largest correlation coefficients after

PC-refinement. The final step of the strategy consists of translation functions using the *PC*-refined search models for this subset of orientations. The net result of the strategy is to reduce the number of possible orientations to be checked by subsequent translation functions and to improve the accuracy of the search model ***prior*** to the translation function[8, 10].

1.3 Applications

A number of computer studies were carried out[8] in order to evaluate the radius of convergence of *PC*-refinement. A search model of crambin with a 2 Å backbone atomic r.m.s. difference from the crystal structure that was obtained from an NMR structure determination using simulated data[11] failed to provide the correct orientation when using a rotation function or a six-dimensional search. *PC*-refinements of the search model in several orientations resulted in the identification of the correct orientation by showing the lowest value of hybrid energy function. In an application to myoglobin it was shown that rigid-group *PC*-refinement of the orientations of the eight α-helices has an approximate radius of convergence of 13°. The crystal structure of myoglobin was made inaccurate by tilting the eight α-helices artificially by 13° around independent axes. This search model failed to provide the correct orientation when using a conventional rotation function or a six-dimensional search. *PC*-refinements uniquely determined the correct overall orientation of the myoglobin search model by returning the α-helices to their original orientations.

The generalized molecular replacement strategy was successful in obtaining phases for several previously unknown crystal structures. The structures include a monoclonal antibody Fab fragment[12, 13] with bound hapten which crystallized in space group $P6_522$, a Fab fragment (26-10) with a bound digoxin molecule which crystallized in space group $P2_1$ with two molecules in the asymmetric unit[14, 15], a complex of an anti-angiotension II Fab with bound peptide which crystallized in space group $P4_1$[16], a quadruple mutant (V66L/ G79S/ G88V/ L108V) of staphylococcal nuclease which crystallized in space group $P6_522$[17], and the active form of the P21 oncogene protein complexed with a GTP analog which crystallized in space group $P2_1$ with four molecules in

the asymmetric unit[18]. In the case of the Fabs, *PC*-refinements of 24 parameters consisting of the orientations and positions of the constant and variable domains of the heavy and the light chain were carried out, whereas in the other two cases *PC*-refinements of 3 parameters consisting of the overall orientation of the search model were carried out. Previous attempts to solve the structures with multiple isomorphous replacement or conventional molecular replacement methods[19] had failed. While one cannot completely exclude the possibility that the structures could have been solved solely by conventional molecular replacement methods, the results suggest that they would have been very difficult cases. In retrospect, all crystals have in common is that the inaccuracy of the model in combination with the presence of non-crystallographic symmetry or the high crystal symmetry of the space group made it difficult to identify the correct orientation of the molecule(s) or single domains by conventional rotation functions.

PC-refinement acts as a filter of the rotation function. If successful, it can discriminate between correct and incorrect orientations of the search model. The two highest peaks of the rotation function for the 26-10 Fab are incorrect (Fig. 1). *PC*-refinement was carried out at two different resolution ranges for 150 of the highest peaks of the rotation function (Fig. 1). Two significant peaks emerged (No. 4 and No. 89) which correspond to the two molecules in the asymmetric unit. In both cases, *PC*-refinement modified certain interdomain angles between 10° and 20°; the elbow-angle increased by about 10°. The root-mean-square difference between the original model and the *PC*-refined model was 6.3 Å when the constant domains of the Fab were fitted.

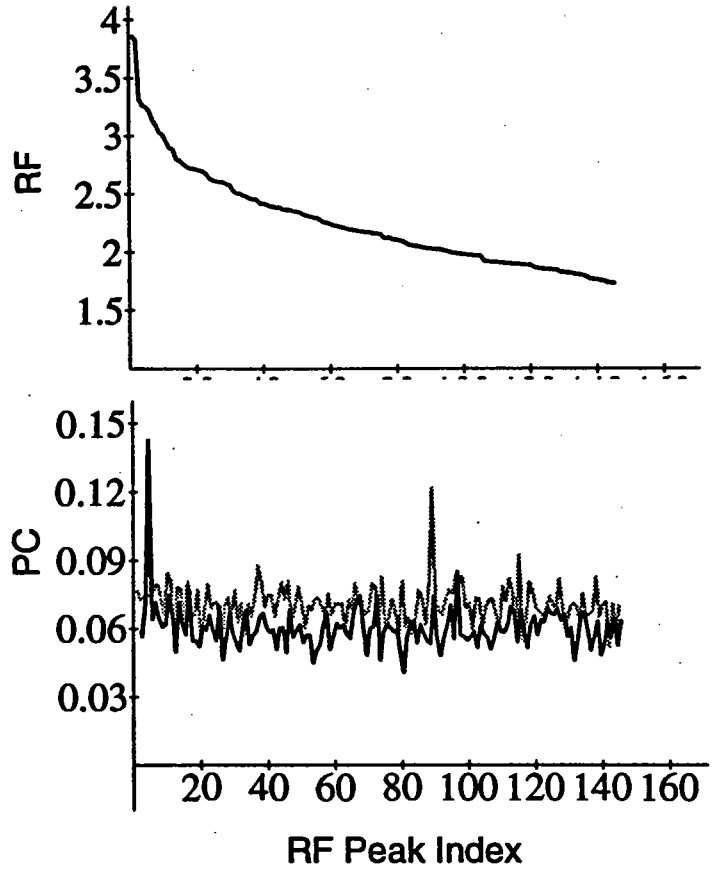


Fig. 1 *Rigid-body PC-refinements for each orientation of the HyHel-5 search model*[29] *corresponding to the 150 highest peaks of the rotation function against the 26-10 data*[14, 15]. *"RF Peak Index" represents a numbering scheme of the peaks of the rotation function, e.g. "1" corresponds to the highest peak, "50" corresponds to the fiftieth highest peak. Shown are the values of the rotation function (RF) and the correlation coefficients (PC) after refinement of the overall orientation and the orientational and positional parameters of the four domains (V_H,V_L,C_H1,C_L) of the Fab at 15–3.5 Å resolution (black line) and at 15–4 Å resolution (grey line). Details of the rotation function and PC-refinement are described elsewhere*[15].

As Fig. 1 shows, the convergence of *PC*-refinement is resolution-dependent:

one of the orientations is obtained at 15–4 Å resolution, the other one at 15–3.5 Å. This can be understood by plotting *PC* as a function of the elbow angle (shown for ANO2 in Fig. 2). The correct elbow angle emerges as the global maximum and its location is resolution-independent. There are a number of local maxima in which *PC*-refinement can get "trapped" (Fig. 2). The situation gets worse if one considers more than just one parameter, e.g., the orientations and positions of the four Fab domains. The location of the local maxima and thus the convergence of *PC*-refinement is resolution-dependent. The local maxima correspond to out-of-register superpositions of the β-sheet pattern of the variable Fab domains.

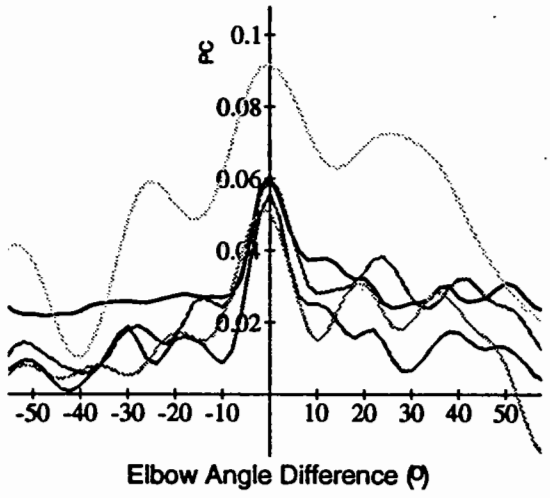


Fig. 2 *PC as a function of the elbow-angle difference between an artificially modified ANO2 structure and the correct ANO2 structure with elbow angle* 153.6°[12, 13]. *The elbow angle is defined as the angle between the pseudo 2-fold axes of symmetry of the V_L-V_H and C_L-C_H1 domain pairs of the Fab. The elbow-angle was modified by rotating the variable domains around an axis connecting the two linker regions of the Fab. PC was evaluated at 15–2.5 (darkest line), 15–4, 15–6, 15–8, 15–11 (lightest line) Å resolution.*

PC-refinement improves the significance of the translation function by

making the search model more accurate. In the case of the 26-10 Fab the *PC*-refined model shows a single significant peak, whereas the correctly oriented but unrefined search model produces a noisy translation function (not shown). The R-factors for the five structures that were solved by generalized molecular replacement were generally in the forties at 8–3 Å resolution and a single round of SA-refinement (SA) produced R-factors in the middle twenties.

2 Assessment of the correctness of atomic models: the free R value

Often molecular replacement produces more than one possible solution. The usual criterion to distinguish between a correct and an incorrect solution is the R value of the molecular replacement solution after some positional refinement. However, it is possible to overfit or "misfit" the diffraction data: an incorrect model can be refined to fairly good R values as several recent examples have shown[20]. Recently we proposed a reliable and unbiased indicator of the accuracy of such models[21]. In analogy to testing statistical models by cross-validation[22, 23] we defined a statistical quantity (R_T^{free}) that measures the agreement between observed and computed structure factor amplitudes for a "test" set of reflections that is omitted in the modelling and refinement process. As examples show, there is a high correlation between R_T^{free} and the accuracy of the atomic model phases. This is useful since experimental phase information is usually inaccurate, incomplete, or unavailable.

R_T^{free} reflects the information content of the atomic model. Suppose both the atomic model and diffraction data are perfect, resulting in $R = 0$. Refinement against A as opposed to all data will not change the atomic model and, thus, $R_T^{free} = 0$. Suppose the data contain small errors and an atomic model is overfit to a very low R value by introducing a large number of free parameters. As the noise is independent among different reflections, overfitting against A will not bias R_T^{free}. A similar argument applies to the case of partially incomplete or incorrect atomic models where the agreement with the diffraction data is improved by fitting noise.

The enhanced sensitivity of R_T^{free} with respect to model errors was illustrated in ref. [21] where the correct[24] and incorrect[25] crystal structures of the plant ribulose-1,5-biphosphate carboxylase oxygenase (RuBisCO) were compared (data and coordinates were kindly provided by Dr. D. Eisenberg). While the R difference between the correct and incorrect model was only 4% for comparable geometry, the R_T^{free} difference was 13%, suggesting that the incorrect model had been overfit.

The free R value approach can be used to address a number of fundamental questions in macromolecular crystallography. This is illustrated for the crystal structure of penicillopepsin from *Penicillium janthinellum*[27, 26] for which diffraction data and coordinates were kindly provided by Drs. M.N.James and A. Sieleki. As an independent assessment of the quality of the atomic model we made use of multiple isomorphous replacement (MIR) phases at 6–2.8Å resolution; these phases were of exceptional quality with an overall figure of merit of 0.9. Experimental phase information is normally less accurate, incomplete or missing.

The information content of a random distribution of scatterers is obviously minimal, although it can be refined to a very low R value (Fig. 3) against the penicillopepsin diffraction data at 1.8 Å resolution; R_T^{free} stays at 54% which is close to the random limit of 59% for an acentric space group[28]. Unrestrained refinement with a model consisting of the same scatterers starting at the positions of the non-hydrogen protein atoms yields R_T^{free}=43% (Fig. 3). Thus, R_T^{free} can distinguish between a distribution of scatterers that is close to the crystal structure and a random distribution, both of which can be refined to a very low R. Inclusion of chemical restraints increases R somewhat while greatly decreasing both R_T^{free} and $\overline{|\Delta\Phi|}$, thus improving the information content of the model (Fig. 3). Inclusion of ordered water molecules lowers R, R_T^{free}, and $\overline{|\Delta\Phi|}$ (Fig. 3). Refinement of randomly placed scatterers in the bulk solvent region of the crystal lowers R while *increasing* both R_T^{free} and $\overline{|\Delta\Phi|}$, thus decreasing the information content.

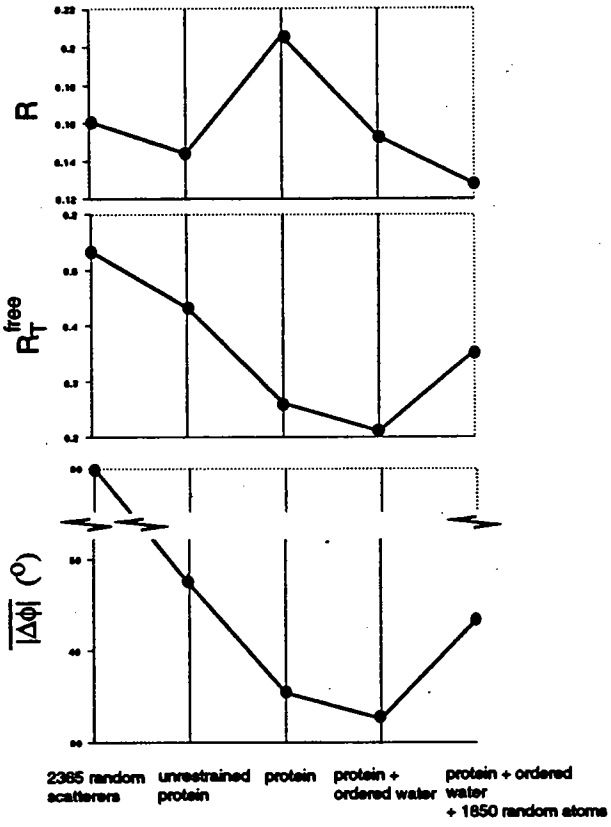


Fig. 3 *"2365 random scatterers" consists of 2365 oxygen atoms with a reduced van der Waals radius of 1.57 Å randomly placed in the asymmetric unit of the crystal. "unrestrained protein" consists of the same scatterers placed near the non-hydrogen positions of the protein portion of the penicillopepsin structure. "protein" is protein portion of the penicillopepsin structure refined with chemical restraints. "protein+ord.water" includes additional 314 ordered water molecules. "protein+ord.water+1850 random atoms" includes additional 1850 oxygen atoms randomly placed in the bulk solvent region. T was obtained by a 10% random selection.* $\overline{|\Delta\Phi|}$ *is the figure-of-merit weighted mean phase difference between model phases and the most probable MIR phases at 6–2.8 Å resolution. Each refinement consisted of a two iterations of SA-refinement and restrained B-factor refinement.*

R_T^{free} represents a reliable and unbiased parameter by which to evaluate the information content of a model produced by X-ray crystallography. In particular, it will distinguish between correct and incorrect molecular replacement

solutions. However, the approach cannot provide any information about *how* to reach a correct molecular replacement solution. We suggest to routinely monitor R_T^{free} during the complete course of modelling of and refinement against crystallographic diffraction data. Any significant increase of R_T^{free} or a stagnation of R_T^{free} might indicate a possible problem.

References

[1] Hoppe, W. (1957). Die Faltmolekülmethode – eine neue Methode zur Bestimmung der Kristallstruktur bei ganz oder teilweise bekannter Molekülstruktur. *Acta Cryst.* **10**, 750–751.

[2] Rossmann, M.G., & Blow, D.M. (1962). The detection of sub-units within the crystallographic asymmetric unit. *Acta Cryst.* **15**, 24–31.

[3] Huber, R. (1965). Die automatisierte Faltmolekülmethode. *Acta Cryst.* **A19**, 353-356.

[4] Rossmann, M.G. (1972), Ed. *The Molecular Replacement Method* (International Science Review No. 13, Gordon & Breach, New York).

[5] Lattman, E.E. (1985). Use of the rotation and translation functions. *Methods Enzymol.* **115**, 55–77.

[6] Fujinaga, M. & Read, R.J. (1987). Experiences with a new translation-function program. *J. Appl. Cryst.* **20**, 517–521.

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

[8] Brünger, A.T. (1990). Extension of molecular replacement: A new search strategy based on Patterson correlation refinement. *Acta Cryst.* **A46**, 46–57.

[9] Brünger,A.T. (1991). Simulated annealing in crystallography. *Ann. Rev. Phys. Chem.* **42**, 197–223.

[10] Yeates, T.O., Rini, J.M. (1990). Intensity-based domain refinement of oriented but unpositioned molecular replacement models. *Acta Cryst.* **A46**, 352–359.

[11] Brünger, A.T., Clore, G.M., Gronenborn, A.M., Karplus, M (1986). Three-dimensional structures of proteins determined by molecular dynamics with interproton distance restraints: Application to crambin. *Proc. Natl. Acad. Sci. USA.* **83**, 3801–3805.

[12] Leahy, D.J., Hynes, T.R., McConnell, H.M., Fox, R.O. (1988). Crystallization of an anti-2,2,6,6-tetramethyl-1-piperidinyloxy-dinitrophenyl monoclonal antibody Fab fragment with and without bound hapten. *J. Mol. Biol.* **203**, 829–830.

[13] Brünger, A.T., Leahy, D.J., Hynes, T.R., FOX, R.O. (1991). The 2.9 Å resolution structure of an anti-dinitrophenyl-spin-label monoclonal antibody Fab fragment with bound hapten. *J. Mol. Biol.* , **221** , 239–256 (1991).

[14] Strong, R.K. (1990). Ph.D. thesis, Harvard University.

[15] Brünger, A.T. (1991). Solution of a Fab (26-10) /digoxin complex by generalized molecular replacement. *Acta Cryst.* **A47**, 195–204.

[16] Garcia, K.C., Ronco, P.,Verroust, P.J., Amzel, L.M. (1989). Crystallization and preliminary X-ray diffraction data of an anti-angiotensin II Fab and of the peptide-Fab complex. *J. Biol. Chem.* **264**, 20463–20466.

[17] Loll, P.J., Meeker, A.K., Shortle, D., Pease, M., Lattman, E.E. (1988). Crystallization and preliminary X-ray analysis of a quadruple mutant of staphylococcal nuclease. *J. Biol. Chem.* **263**, 18190–18192.

[18] Brünger, A. T., Milburn, M. V., Tong, L., de Vos, A. M., Jancarik, J., Yamaizumi, Z., Nishimura, S., Ohtsuka, E., Kim, S.-H. (1990). Crystal Structure of an Active Form of *ras* Protein, a Complex of GTP Analog and c-H-*ras* P21 Catalytic Domain. *Proc. Natl. Acad. Sci. USA* 87:4849–53.

[19] Fitzgerald, P. (1988). MERLOT, an integrated package of computer programs for the determination of crystal structures by molecular replacement. *J. Appl. Cryst.* **21** 273–278.

[20] Brändén, C. I. and Jones, A. *Nature* **343**, 687–689 (1990).

[21] Brünger, A T. *Nature*, **355**, 472–474 (1992).

[22] Mosteller,F., Tukey,J.W. *Data Analysis and Regression. A Second Course in Statistics.* (Addison-Wesley, Reading, MA., 1977).

[23] Efron, B., Tibshirani, R. *Science* **253**, 390–395 (1991).

[24] Curmi, P.A.M., Schreuder, H., Cascio, D., Sweet, R.M., Eisenberg, D., *J. Biol. Chem*, in press (1991).

[25] Chapman, M.S., Suh, S.W., Curmi, P.M.G., Cascio, D., Smith, W.W., Eisenberg, D. *Science* **241**, 71–74 (1988).

[26] Hsu, I-N., Delbare, L.T.J., James, M.N.G., Hofmann, T. *Nature* **266**, 140–145 (1977).

[27] James, M.N.G., Sielecki, A.R. *J. Mol. Biol.* **163**, 299–361 (1983).

[28] Stout, G. H., Jensen, L. H. In *X-ray Structure Determination, A Practical Guide*, (John Wiley & Sons, New York, 1989).

[29] Sheriff, S., Silverton, E.W., Padlan, E.A., Cohen, G.H., Smith-Gill,S.J., Finzel, B., Davies, D.R. (1987). Three-dimensional structure of an antibody-antigen complex. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 8075–8079.

A Statistical Formulation of the Molecular Replacement and Molecular Averaging Methods

G. Bricogne
MRC Laboratory of Molecular Biology, Cambridge, UK,
and LURE, Batiment 209D, 91405 Orsay, France.

This talk is intended to provide an informal preview (hurriedly written up) of my on-going work on an enhancement of molecular replacement and molecular averaging techniques whose principle was outlined as part of a "Bayesian programme" (Bricogne, 1988) aimed at integrating all phasing methods within a common statistical framework.

0. Need for a statistical formulation.

At first sight there isn't one. Molecular Replacement proceeds by (i) orienting a known fragment by self-Patterson superposition (rotation function), (ii) registering the oriented fragment by cross-Patterson superposition (translation function), and (iii) completing the structure by means of phase combination and difference maps. Molecular Averaging proceeds by (i) possibly orienting non-crystallographic symmetry (ncs) axes by self-Patterson superposition (rotation function), but more often obtaining both rotational and translational ncs elements by examination of heavy-atom positions in heavy-atom derivatives, (ii) obtaining a starting envelope from the first averaging or sometimes from electron micrographs, and (iii) iterating the sequence of (a) averaging, (b) phase combination with possibly a small degree of extension, and (c) revision of the envelope and/or of the ncs elements.

The main tools are therefore the Fourier methods involved in the calculation of **Patterson superposition** functions; **Wilson statistics**, which intervene in the scaling of the amplitudes from the partial or averaged structure to the observed amplitudes and in the derivation of phase information by means of Sim's formula; and the standard difference map procedures affording an **interface to refinement**.

The limitations of the present methodology arise from the limitations of each of these tools. These may be briefly characterised as follows. (1) The **Patterson functions** available for macromolecular structures (a) do not have atomic resolution, (b) are distorted by missing data, and (c) are unweighted superpositions and hence cannot take account of measurement errors. (2) **Wilson statistics** assume that the 'rest' of the structure consists of a uniform distribution of independent random scatterers in the asymmetric unit, which is clearly inadequate since (a) in molecular replacement the known fragment tends to exclude the unknown atoms, and (b) in molecular averaging much of the missing structure will obey the ncs and hence give rise to strongly non-Wilsonian intensity distributions. (3) The current **interface to refinement** is not well equipped to refine a modelled fragment while a substantial unmodelled part exists.

In this talk I will show how an improvement of the statistical model for joint distributions of structure factors along the lines which I indicated earlier (Bricogne, 1988) affords a solution to all three categories of problems simultaneously.

1. Molecular Replacement.

1.1. Prototype : the "heavy-atom" method.

The "heavy-atom" method is undoubtedly the direct ancestor of molecular replacement, and I would argue that many of the problems mentioned above are a result of having uncritically carried over procedures and approximations which are justifiable in the former but questionable in the latter.

1.1.1. Basis of the method.

The "heavy atom" is first detected by examination of the Patterson, taking advantage of the fact that such an atom is localised and that there is no rotation problem. One then switches over to Sim's formula, which can be expected to hold rather accurately since the 'light' atoms making up the rest of the structure are distributed uniformly enough for Wilson's statistics to be obeyed (small-molecule crystals are usually close-packed – there is no solvent – and the exclusion of light atoms by the heavy atom can be neglected in most cases).

1.1.2. Statistical treatment.

This is simple and will help introduce the notation. Writing the operation of an element g of the space group G as

$$S_g(\mathbf{x}) = \mathbf{R}_g \mathbf{x} + \mathbf{t}_g \tag{1.1}$$

the contribution $\Xi(\mathbf{h}, \mathbf{x})$ of a point atom of unit scattering factor placed at $\mathbf{x}$ to the structure factor at $\mathbf{h}$ may be written :

$$\Xi(\mathbf{h}, \mathbf{x}) = \frac{1}{|G_{\mathbf{x}}|} \sum_{g \in G} e^{2\pi i \mathbf{h} \cdot S_g(\mathbf{x})} \tag{1.2}$$

where $G_{\mathbf{x}}$ is the isotropy subgroup of $\mathbf{x}$ and $|G_{\mathbf{x}}|$ denotes the number of its elements. Labeling quantities belonging to light and heavy atoms by superscripts L and H respectively we may write at each $\mathbf{h}$

$$F(\mathbf{h}) = F^H(\mathbf{h}, \mathbf{x}^H) + F^L(\mathbf{h}) \tag{1.3}$$

with

$$F^H(\mathbf{h}, \mathbf{x}^H) = f^H(\mathbf{h})\, \Xi(\mathbf{h}, \mathbf{x}^H) \tag{1.4}$$

and

$$F^L(\mathbf{h}) = \sum_{j \in J^L} f_j^L(\mathbf{h})\, \Xi(\mathbf{h}, \mathbf{x}_j^L) \tag{1.5}$$

where $\mathbf{x}^H$ is the known or assumed position of the heavy atom.

The $F^L(\mathbf{h})$ obey Wilson's statistics, i.e. are distributed (1) as a 2D Gaussian centered at (0,0) with variance $\Sigma_a^L(\mathbf{h}) = \frac{1}{2} |G_{\mathbf{h}}|\, \sigma_2^L(\mathbf{h})$ along each component for $\mathbf{h}$ acentric, and (2) as a 1D Gaussian centered at 0 with variance $\Sigma_c^L(\mathbf{h}) = |G_{\mathbf{h}}|\, \sigma_2^L(\mathbf{h})$ for $\mathbf{h}$ centric, where

$$\sigma_2^L(\mathbf{h}) = \sum_{\text{cell}} \left[f_j^L(\mathbf{h})\right]^2 \qquad (1.6).$$

Therefore $F(\mathbf{h})$ is distributed as a similar Gaussian with its centre shifted away from the origin to $F^H(\mathbf{h}, \mathbf{x}^H)$.

The origin-shifted Gaussian distribution of $F(\mathbf{h})$ may be integrated over all possible phases (for acentric **h**) or summed over the two possible signs (for centric **h**) to yield the conditional probability distribution for the modulus $|F_\mathbf{h}|$:

$$\mathcal{P}\left(|F_\mathbf{h}| = |F_\mathbf{h}|^{obs} \;\middle|\; F^H(\mathbf{h}) = F^H(\mathbf{h}, \mathbf{x}^H) \right)$$

$$= \mathcal{R}\left(|F^H(\mathbf{h}, \mathbf{x}^H)|, |F_\mathbf{h}|^{obs}, \Sigma_a^L(\mathbf{h}) \right) \quad \text{for } \mathbf{h} \text{ acentric} \qquad (1.7a)$$

$$= C\left(|F^H(\mathbf{h}, \mathbf{x}^H)|, |F_\mathbf{h}|^{obs}, \Sigma_c^L(\mathbf{h}) \right) \quad \text{for } \mathbf{h} \text{ centric} \qquad (1.7b)$$

where $\mathcal{R}$ and C denote the Rice distributions :

$$\mathcal{R}(r, R, \Sigma) = \frac{R}{\Sigma} \exp\left(-\frac{r^2+R^2}{2\Sigma}\right) I_0\left(\frac{rR}{\Sigma}\right) \qquad (1.8a)$$

and

$$C(r, R, \Sigma) = \sqrt{\frac{2}{\pi\Sigma}} \exp\left(-\frac{r^2+R^2}{2\Sigma}\right) \cosh\left(\frac{rR}{\Sigma}\right). \qquad (1.8b)$$

1.1.3. Statistical detection of the heavy atom.

The statistical detection of the heavy atom works by comparing two hypotheses:

$$(\mathcal{H}_0) \qquad : \mathcal{P}(|F_\mathbf{h}|) = \mathcal{R}(0, |F_\mathbf{h}|, \Sigma_a(\mathbf{h})) \quad \text{for all } \mathbf{h} \text{ acentric} \qquad (1.9a)$$

$$= C(0, |F_\mathbf{h}|, \Sigma_c(\mathbf{h})) \quad \text{for all } \mathbf{h} \text{ centric} \qquad (1.9b)$$

where Σ_a and Σ_c include the scattering power of the heavy atom *via*

$$\sigma_2(\mathbf{h}) = \sigma_2^L(\mathbf{h}) + |G|\left[f^H(\mathbf{h})\right]^2 \qquad (1.10)$$

$$(\mathcal{H}_1[\mathbf{x}^H]) : \mathcal{P}(|F_\mathbf{h}|) = \mathcal{R}(|F^H(\mathbf{h}, \mathbf{x}^H)|, |F_\mathbf{h}|, \Sigma_a^L(\mathbf{h})) \quad \text{for all } \mathbf{h} \text{ acentric} \qquad (1.11a)$$

$$= C(|F^H(\mathbf{h}, \mathbf{x}^H)|, |F_\mathbf{h}|, \Sigma_c^L(\mathbf{h})) \quad \text{for all } \mathbf{h} \text{ centric} \qquad (1.11b)$$

by means of the *log-likelihood gain* :

$$\text{LLG}(\mathbf{x}^H) = \sum_\mathbf{h} \log \frac{\mathcal{P}\left(|F_\mathbf{h}| = |F_\mathbf{h}|^{obs} \;\middle|\; (\mathcal{H}_0) \right)}{\mathcal{P}\left(|F_\mathbf{h}| = |F_\mathbf{h}|^{obs} \;\middle|\; (\mathcal{H}_1[\mathbf{x}^H]) \right)} \qquad (1.12).$$

In order to examine this quantity more closely let us denote by $\mathcal{O}_a$ (resp. $\mathcal{O}_c$) the list of acentric (resp. centric) reflexions which are unique with respect to symmetry and Friedel equivalence. Then

$$\mathrm{LLG}\ (\mathbf{x}^H) = \sum_{\mathbf{h}\in O_a} \left[\log \frac{\Sigma_a(\mathbf{h})}{\Sigma_a^L(\mathbf{h})} - \tfrac{1}{2}\left|F_{\mathbf{h}}^{obs}\right|^2 \left(\frac{1}{\Sigma_a(\mathbf{h})} - \frac{1}{\Sigma_a^L(\mathbf{h})}\right) - \tfrac{1}{2}\frac{\left|F_{\mathbf{h}}^H(\mathbf{x}^H)\right|^2}{\Sigma_a^L(\mathbf{h})} \right]$$

$$+ \sum_{\mathbf{h}\in O_c} \left[\tfrac{1}{2}\log \frac{\Sigma_c(\mathbf{h})}{\Sigma_c^L(\mathbf{h})} - \tfrac{1}{2}\left|F_{\mathbf{h}}^{obs}\right|^2 \left(\frac{1}{\Sigma_c(\mathbf{h})} - \frac{1}{\Sigma_c^L(\mathbf{h})}\right) - \tfrac{1}{2}\frac{\left|F_{\mathbf{h}}^H(\mathbf{x}^H)\right|^2}{\Sigma_c^L(\mathbf{h})} \right]$$

$$+ \sum_{\mathbf{h}\in O_a} \log I_0\left(\frac{\left|F_{\mathbf{h}}^H(\mathbf{x}^H)\right|\left|F_{\mathbf{h}}^{obs}\right|}{\Sigma_a^L(\mathbf{h})}\right) + \sum_{\mathbf{h}\in O_c} \log\cosh\left(\frac{\left|F_{\mathbf{h}}^H(\mathbf{x}^H)\right|\left|F_{\mathbf{h}}^{obs}\right|}{\Sigma_a^L(\mathbf{h})}\right) \qquad (1.13).$$

1.1.4. Quadratic approximation and Patterson correlation.

To obtain a quadratic approximation to (1.13) valid when the heavy atom represents only a small fraction of the scattering power of the total structure (so that $\Sigma^L \approx \Sigma$), recall that

$$\log\ I_0(z) \approx \tfrac{1}{4}z^2 \qquad (1.14a)$$

and

$$\log\cosh(z) \approx \tfrac{1}{2}z^2 \qquad (1.14b).$$

Recall also that under symmetry and Friedel expansion an acentric $\mathbf{h}$ has $2\frac{|G|}{|G_{\mathbf{h}}|}$ equivalents, while a centric $\mathbf{h}$ has $\frac{|G|}{|G_{\mathbf{h}}|}$ equivalents. It is then straightforward to derive the approximation :

$$\mathrm{LLG}\ (\mathbf{x}^H) \approx \frac{1}{2|G|} \sum_{\text{all } \mathbf{h}} \left(\left|E_{\mathbf{h}}^{obs}\right|^2 - 1\right) \frac{\left|F_{\mathbf{h}}^H(\mathbf{x}^H)\right|^2}{\sigma_2(\mathbf{h})} \qquad (1.15)$$

where

$$\left|E_{\mathbf{h}}^{obs}\right|^2 = \frac{\left|F_{\mathbf{h}}^{obs}\right|^2}{|G_{\mathbf{h}}|\,\sigma_2(\mathbf{h})} \qquad (1.16)$$

is the squared normalised amplitude.

This shows very neatly the form of a Patterson correlation (PC) function, calculated in reciprocal space *via* Parseval's theorem, between the origin-removed $|E|^2$–based Patterson for the whole structure and the heavy-atom Patterson. However this correspondence is valid only in the limiting case where the heavy atom contributes a negligibly small part to the total scattering power. This is not usually the case, i.e. in general Σ^L differs substantially from Σ : then (1.13) remains valid and optimal, provided the likelihoods in the numerator and denominator of (1.12) are separately maximised with respect to overall scale and temperature factor in two distinct maximum-likelihood normalisation operations corresponding to the two distinct hypotheses (see Bricogne, 1991, 1993). The quadratic approximation (1.15) then has an extra constant term, and it will be shown elsewhere that the $|E|^2$ values involved in forming the Patterson function are derived by normalising with respect to the light atoms (i.e. with respect to σ_2^L) the 'renormalised' amplitudes

$$\left|F_{\mathbf{h}}^{\text{renorm}}\right|^2 = \left|F_{\mathbf{h}}^{\text{obs}}\right|^2 + \left|F_{\mathbf{h}}^{H}(\mathbf{x}^H)\right|^2 - 2\, m_{\text{Sim}} \left|F_{\mathbf{h}}^{\text{obs}}\right| \left|F_{\mathbf{h}}^{H}(\mathbf{x}^H)\right| \qquad (1.17)$$

where m_{Sim} is the Sim figure of merit derived from the heavy atom. The expectation value of the corresponding $|E|^2$ is then exactly 1, which ensures perfect origin removal in the Patterson function. This accurate renormalisation has the disadvantage of depending on $\mathbf{x}^H$, but this dependence can be removed by replacing $\left|F_{\mathbf{h}}^{H}(\mathbf{x}^H)\right|$ by its rms expectation; this then yields a best compromise renormalisation which gives optimal coefficients for a fast search of the heavy atom (see §1.1.6 below).

1.1.5. Further advantages of LLG over PC.

(0) The correspondence (1.15) is valid only if light atoms are distributed so as to give rise to Wilson statistics; it may be known that this is not the case, e.g. in crystals of zeolites or other small structures containing cavities, and of course in macromolecular crystals. The LLG has no difficulty in remaining an optimal detection criterion in this case, while the PC coefficient will fail to do so.

(1) The statistical variances Σ_a and Σ_c can be incremented so as to reflect measurement errors, while there is no natural way to do so in calculating the PC coefficient.

(2) The LLG is still correct if data are missing.

(3) The scene is already set to define the conditional probability distribution of $\left|F_{\mathbf{h}}\right| e^{i\varphi_{\mathbf{h}}}$ for $\mathbf{x}^H$ fixed, and obtain Sim's formulae for $P(\varphi_{\mathbf{h}})$ and the centroid $\langle F_{\mathbf{h}} \rangle$.

(4) Significance tests can be applied to make the detection of the heavy atom a quantitative affair.

(5) The analysis given above shows that LLG maximisation affords a means of refining the parameters of the heavy atom in the presence of randomly positioned light atoms.

(6) Partial phase information available from an external source can be incorporated at the stage where one integrates over acentric phases and sums over centric signs. If this phase information is point-sharp, then the LLG is essentially an electron density correlation function. In all intermediate cases, the LLG will possess features intermediate between those of a Patterson correlation coefficient and of a density correlation coefficient.

In summary, the sensitivity of the detection procedure can be increased to the maximum level achievable by using the LLG instead of previously used criteria; and in simple cases where the latter can be expected to be good, the analytical expression of the LLG does show close similarity with them, or suggests sensible adaptations of these criteria which had not yet arisen within their own theoretical framework.

1.1.6. Fourier series representation of the LLG.

Combining (1.4) and (1.15) in the simple case of a single heavy atom in the asymmetric unit we get in the general case

$$\mathrm{LLG}\ (\mathbf{x}^H) \approx \mathrm{const} + \frac{1}{2|G|} \sum_{\text{all } \mathbf{h}} \left(\left| E_{\mathbf{h}}^{\text{renorm}} \right|^2 - 1 \right) \frac{\left(f_{\mathbf{h}}^H\right)^2}{\sigma_2^L(\mathbf{h})} \left| \Xi(\mathbf{h}, \mathbf{x}^H) \right|^2 \qquad (1.18).$$

Now there are two different ways of rewriting $\left|\Xi(\mathbf{h}, \mathbf{x})\right|^2$:

(1) in reciprocal space, by invoking Bertaut's linearisation formula (assuming that $\mathbf{x}^H$ is in general position) :

$$\left|\Xi(\mathbf{h}, \mathbf{x}^H)\right|^2 = \sum_{g \in G} e^{2\pi i \mathbf{h} \cdot \mathbf{t}_g} \Xi(\mathbf{h} - \mathbf{R}_g^T \mathbf{h}, \mathbf{x}^H) \qquad (1.19a)$$

which can be directly sustituted into (1.18) to yield a Fourier series in which the data at $\mathbf{h}$ are used to form the coefficients at $\mathbf{h} - \mathbf{R}_g^T \mathbf{h}$ for each $g \in G$;

(2) in real space, by invoking the dual identity :

$$\left|\Xi(\mathbf{h}, \mathbf{x}^H)\right|^2 = \sum_{g \in G} \Xi(\mathbf{h}, \mathbf{x}^H - S_g(\mathbf{x}^H)) \qquad (1.19b)$$

whose substitution into (1.18) casts the log-likelihood gain into the form :

$$\mathrm{LLG}\ (\mathbf{x}^H) \approx \mathrm{const} + \frac{1}{2|G|} \sum_{g \in G} \mathrm{Patt}\ (\mathbf{x}^H - S_g(\mathbf{x}^H)) \qquad (1.20).$$

which can be recognised as a Buerger implication function calculated from a Patterson synthesis based on coefficients $\left(\left| E_{\mathbf{h}}^{\text{renorm}} \right|^2 - 1 \right)$, hence origin-removed, weighted by $\frac{\left(f_{\mathbf{h}}^H\right)^2}{\sigma_2^L(\mathbf{h})}$.

Quite generally the logical equivalence between (1.19a) and (1.19b) may be viewed as the basis for the connections between direct methods and Patterson methods which have been investigated recently by several researchers, notably Giacovazzo and coworkers, and Pavlecik.

It is straightforward to derive by the same method approximate expressions for the log-likelihood gain which allow one to search simultaneously for several heavy atoms, either as Fourier series or as multiple implication functions.

1.2. Molecular Replacement proper.

This method would be better called the "Molecular Placement method", since it consists in *placing* a known fragment in an unknown structure rather than in replacing something by something else.

1.2.1. Basic assumptions and relations.

Instead of a heavy atom we have a known fragment described in a reference position and orientation by a density ρ^M with transform F^M. If ρ^M is rotated by $\mathbf{R}$ and translated by $\mathbf{t}$ to give the copy of the fragment lying in the chosen asymmetric unit, then the density for the known partial structure in the crystal may be written :

$$\rho^{par} = \sum_{g \in G} \tau_{S_g(t)} (R_g R)^{\#} \rho^M \qquad (1.21)$$

where the translate of a function by a vector $\mathbf{a}$ and its image under a rotation $\mathbf{R}$ are defined as usual by

$$(\tau_a f)(x) = f(x-a) \qquad \text{and} \qquad (R^{\#} f)(x) = f(R^{-1}x) \qquad (1.22).$$

The partial structure factors which play the same role here as $F_h^H(x^H)$ in §1.1 are therefore :

$$F_h^{par}(R, t) = \sum_{g \in G} e^{2\pi i h \cdot S_g(t)} F^M \left[(R_g R)^T h \right] \qquad (1.23)$$

and the corresponding squared amplitudes may be written :

$$\left| F_h^{par}(R, t) \right|^2 = \sum_{g \in G} \left| F^M \left[(R_g R)^T h \right] \right|^2 + \sum_{\substack{g, g' \in G \\ g \neq g'}} F^M \left[(R_g R)^T h \right] \overline{F^M \left[(R_g R)^T h \right]} e^{2\pi i h \cdot (t_g - t_{g'})} \times e^{2\pi i (R_g^T h - R_{g'}^T h) \cdot t} \qquad (1.24).$$

1.2.2. Statistical treatment.

The statistical detection and placement of the fragment will proceed by calculating the log-likelihood gain

$$LLG(R, t) = \log \frac{\mathcal{P}\left(|F_h| = |F_h|^{obs} \text{ for all } h \mid (\mathcal{H}_0) \right)}{\mathcal{P}\left(|F_h| = |F_h|^{obs} \text{ for all } h \mid (\mathcal{H}_1[R, t]) \right)} \qquad (1.25)$$

where $(\mathcal{H}_0)$ denotes the null hypothesis that all atoms are uniformly distributed in the asymmetric unit while $(\mathcal{H}_1[R, t])$ denotes the alternative hypothesis that a subset of atoms is assembled into the known fragment and placed in the asymmetric unit with orientation $\mathbf{R}$ at position $\mathbf{t}$, and the rest are distributed at random.

When the same assumptions are fulfilled as in the heavy-atom method (but they never are! see below §1.2.3) the joint probabilities in the numerator and denominator of (1.25) become products of probabilities associated to each reflexion :

$$LLG(R, t) \approx \sum_h \log \frac{\mathcal{P}\left(|F_h| = |F_h|^{obs} \mid (\mathcal{H}_0) \right)}{\mathcal{P}\left(|F_h| = |F_h|^{obs} \mid (\mathcal{H}_1[R, t]) \right)} \qquad (1.26).$$

The methods used in §1.1.4 and §1.1.6 then carry over to the present situation with $\left| F_h^H(x^H) \right|^2$ replaced by $\left| F_h^{par}(R, t) \right|^2$ and the log-likelihood gain is approximately :

$$\mathrm{LLG}\ (\mathbf{R},\mathbf{t}) \approx \mathrm{const} + \frac{1}{2|G|} \sum_{\text{all } \mathbf{h}} \frac{\left(\left|E_{\mathbf{h}}^{\mathrm{renorm}}\right|^2 - 1\right)}{\sigma_2^L(\mathbf{h})} \left\{ \sum_{g \in G} \left| F^M\left[(\mathbf{R}_g\mathbf{R})^T \mathbf{h}\right] \right|^2 \right\}$$

$$+ \frac{1}{2|G|} \sum_{\text{all } \mathbf{h}} \frac{\left(\left|E_{\mathbf{h}}^{\mathrm{renorm}}\right|^2 - 1\right)}{\sigma_2^L(\mathbf{h})} \left\{ \sum_{\substack{g,g' \in G \\ g \neq g'}} F^M\left[(\mathbf{R}_g\mathbf{R})^T \mathbf{h}\right] \overline{F^M\left[(\mathbf{R}_g\mathbf{R})^T \mathbf{h}\right]} \right.$$

$$\left. \times\ e^{2\pi i \mathbf{h}\cdot(\mathbf{t}_g - \mathbf{t}_{g'})}\ e^{2\pi i (\mathbf{R}_g^T\mathbf{h} - \mathbf{R}_{g'}^T\mathbf{h})\cdot\mathbf{t}} \right\} \qquad (1.27)$$

where σ_2^L is defined as a sum over all atoms not belonging to the fragment. As before the amplitudes $\left|E_{\mathbf{h}}^{\mathrm{renorm}}\right|$ should in principle be calculated for each placement $(\mathbf{R}, \mathbf{t})$ after a maximum-likelihood normalisation of the observed amplitudes incorporating the presence of the fragment (see Bricogne, 1993), although compromise values suitable for all placements may be obtained so as to give (1.27) the form of a Fourier series.

The first term in (1.27) depends only on the rotational placement $\mathbf{R}$ and can be recognised as a PC-based rotation function in which a sum of point-group symmetry-related copies of the self-Patterson of the rotated fragment is being correlated with the origin-removed self-Patterson of the whole structure. A similar function is used in XPLOR (Brünger, 1990) but without point-group symmetrisation of the fragment's self-Patterson, so that the correlation coefficients obtained have low values (also, the calculation of E's in XPLOR does not use renormalisation and ignores the statistical weights of reflexions). So far, test calculations using symmetry-expanded self-Pattersons for the fragment have remained inconclusive (David Stuart, these proceedings) but I would venture to suggest that the problem lies in the necessity to use renormalised E's instead of ordinary E's, or indeed of F's in the standard method.

The second term in (1.27) considered for a fixed value of the rotational component $\mathbf{R}$ of the placement, is a PC-based translation function, expressed as a Fourier series in which the data at $\mathbf{h}$ are used to form the coefficients at $\mathbf{R}_g^T\mathbf{h} - \mathbf{R}_{g'}^T\mathbf{h}$ so that the argument of the series can be $\mathbf{t}$ itself (Harada, Lifchitz, Berthou & Jollès, 1981) ; the indifference of this function to the component of $\mathbf{t}$ transversal to all the $\mathbf{R}_g^T\mathbf{h} - \mathbf{R}_{g'}^T\mathbf{h}$ is then seen to be equivalent to the freedom of choice of an origin permissible for the space group G . Once again, a similar function is used by XPLOR but with E values which take no account of the need for renormalisation nor of the statistical weight of reflexions. The fact that the log-likelihood gain (which is an optimal criterion by the Neyman-Pearson theorem) is based on E's provides a final explanation to the long-standing observations by Ian Tickle that E-based translation functions always give better results that F-based ones.

It is therefore clear that even the most approximate implementation of the statistical approach proposed in Bricogne (1988) yields better criteria than the most sophisticated ones available so far,

and suggests non-trivial improvements of the existing methodology which had not yet arisen within this methodology itself.

1.2.3. Advantages of LLG over PC.

The simplification of LLG to a Patterson superposition function has been obtained at the cost of several approximations, some of which are either unjustified or detrimental.

(1) The simplification of (1.25) to (1.26) and the use of Wilson statistics for the structure factor contributions from the "rest" of the atoms amount to assuming that the atoms not belonging to the fragment are uniformly distributed in the asymmetric unit. This is plainly incorrect since these atoms will tend to be excluded from the volume occupied by the fragment, and may also be excluded from a solvent region of known shape. I have derived a multivariate generalisation of the Rice distributions (1.8a,b) which allows the calculation of the log-likelihood gain according to (1.25) rather than (1.26) and can use any non-uniform distribution for the non-fragment atoms and the solvent atoms, thereby overcoming this difficulty.

(2) The normalisation needed to make the renormalised E values in (1.27) independent of the placement $(\mathbf{R}, \mathbf{t})$ has to be carried out as a compromise between all possible placements, and hence cannot be as good as if it were carried out separately for each assumed placement. The LLG calculated pointwise for a suitable sample of placements $(\mathbf{R}, \mathbf{t})$ would therefore be more sensitive than the PC criterion (1.27), and the maximum-likelihood normalisation used to derive the placement-sensitive renormalised E's would confer to this LLG a greater accuracy than the scale-insensitive PC coefficient used by Fujinaga & Read (1987) in their BRUTE program.

(3) The Fourier series (1.27) is vulnerable to missing data and to measurement errors. By contrast the Bayesian method uses maximum-entropy distributions for random atoms which can, in part at least, remove some of the distortions associated with systematically missing data. Furthermore it is possible to increment the variance parameters Σ to reflect measurement errors.

(4) No provision is made for the fact that ρ^M may only be *homologous* and not *identical* to a part of the unknown structure. This can be mended by using the "multichannel formalism" described in Bricogne (1988) to introduce a distribution of "clutter atoms" with scattering factors defined so that they may represent difference features associated to local atomic displacement, as well as missing or supplementary atoms. The distribution of these clutter atoms can be made non-uniform to reflect prior knowledge that some regions are expected to be less well conserved than others. The effect of the clutter atoms on the statistical model is to increment the variance parameters Σ in a resolution-dependent way.

(5) In the standard molecular replacement method there is no weighting in the translational search to reflect possible inaccuracies in the results of the rotational search. In the statistical method it is possible to increment the variance parameters Σ by an extra term which increases as a function of resolution (as was first proposed in an other context by Rossmann & Blow, 1961) and which will downweight high resolution terms. This is likely to increase the radius of convergence of the classical two-stage search strategy.

(6) Any external phase information available can be incorporated into the calculation of likelihoods at the stage of acentric phase integration or centric sign summation. If this phase information is sharply peaked, i.e.if all reflexions have high figures of merit, it is straightforward to show (using the asymptotic behaviour of I_0 and cosh) that the LLG is now closely related to an E-map correlation coefficient rather than to a Patterson correlation coefficient. In intermediate cases, therefore, the LLG will afford what might be called *partially phased* rotation and translation functions.

1.2.4. Scope of future developments.

(1) I am implementing the use of the log-likelihood gain as a general search method for placing known fragments in unknown macromolecular crystals. Besides the expression (1.27), which lends itself to fast Fourier calculations in each of the two stages of the classical search, I am planning to use a simultaneous rotation and translation search in which finer and finer discrete sample grids of all possible rotations are used for $\mathbf{R}$ and the translation-sensitive part of (1.27) is used as a translation function, with renormalised E's calculated as a compromise over all $\mathbf{t}$'s but for that particular $\mathbf{R}$, and with variances Σ incremented as a function of resolution in accordance with the coarseness of the rotational sampling. There are well developped (if little known) techniques for optimally sampling the rotation group by discrete grids. As peaks begin to appear in this 6D search, their position can be found more accurately by using the exact expression (1.25), with a separate data normalisation for each grid point. Subsequent stages of the search on finer grids may then be limited to the vicinity of these peaks, thus breaking away from a full 6D search. The working hypothesis is that, besides the intrinsic superiority of the LLG as a search criterion, this *hierarchical* search strategy ought to overcome some of the limitations of the present two-stage strategy, in which the rotational search is first undertaken without any translational information, and the translational search is then carried out with the handicap of possible errors in the previously obtained rotation. This work will be described in detail elsewhere.

(2) Once $\mathbf{R}$ and $\mathbf{t}$ have been determined the conditional distribution of the F's for the entire structure can be obtained. Classically this leads to Sim's formula but the validity of the latter again depends on that of Wilson statistics. The techniques described in Bricogne (1988) will give better conditional distributions, incorporating *in advance* any known non-uniformity in the distribution of the residual atoms. Looking at the derivatives of the LLG with respect to the parameters of the fragment model will allow the refinement of these parameters in the presence of randomly placed residual scatterers in a way which will ensure that the latter are not wiped out. This is particularly relevant to the interface with refinement mentioned in §0, as present algorithms do display an irrepressible propensity to eliminate density features not explicitly accounted for in the atomic model. I have advocated (Bricogne, 1991) that the LLG should be used for refinement purpose instead of the current residual, since it is able – unlike the latter – to downweight the amplitudes constraints according to the uncertainty on the associated phases and according to the incompleteness of the

model. The gradient maps of the LLG would then be the natural statistically-based couterparts of difference maps.

(3) If the gains in sensitivity of detection from (1) and in efficiency of recycling from (2) turn out to be as substantial as I expect them to be, it would become conceivable to attack the *ab initio* determination of protein structures by systematically searching for super-secondary fragments of 20 to 30 amino-acids, for which it may be possible to compile a library similar to the library of short fragments used by Jones & Thirup (1986) to assist map interpretation. If this does not work without some initial phase information, then the direct phasing method based on entropy maximisation and likelihood ranking which I have described elsewhere (Bricogne, 1993) might afford a means of producing phase information *ab initio* in order to increase the detection sensitivity of strategy (1) above a critical threshold.

2. Molecular Averaging.

Here the problems are : (1) find the ncs rotations; (2) find the ncs translations; (3) find an initial envelope; (4) find some initial phases to prime the phasing process. Not all of them are independent.

The "prototype" of this situation (in the sense of §1.1) is the detection of crystallographic symmetry. In that case symmetry means total (perfect) correlation between $F(\mathbf{h})$ and $F(\mathbf{R}_g^T\mathbf{h})$. It is interesting to see how this comes about within statistics. By Bertaut's linearisation formula the covariance matrix between the real and imaginary parts of $F(\mathbf{h})$ and those of $F(\mathbf{R}_g^T\mathbf{h})$ is a 4×4 matrix with rank 2 (for $\mathbf{h}$ acentric) or a 2×2 matrix with rank 1 (for $\mathbf{h}$ centric), as its second half is equal to its first half multiplied by a phase factor. Other symmetry-related phenomena which can thus be captured in a statistical form are *centric character* (one variance becomes zero) and *specialness* (for an allowed reflexion the variances are multiplied by the order of the isotropy subgroup; for a forbidden one the variances are zero). The symmetry of the intensity pattern alone does not always determine the space group uniquely, but intensity statistics (i.e. in effect a LLG test) do enable the resolution of the remaining ambiguities.

In the case of non-crystallographic symmetries, the extra symmetry induced in the intensity pattern is imperfect (because it is sampled at points of the reciprocal lattice, which is not invariant under ncs rotations) hence gives rise to correlation coefficients which are less than unity but are exactly calculable; and systematic reinforcements or attenuations associated to these symmetries are more elusive that in the crystallographic case but do nevertheless occur.

A quantitative treatment of these imperfect symmetry effects requires extending intensity statistics to include non-crystallographic symmetries. Early attempts at characterising what were called "hypercentric" distributions included only translational repeats, not general rotational+translational repeats. I have shown how to treat the general case (Bricogne, 1988, Appendix) by generalising the Bertaut linearisation formula so as to incorporate non-crystallographic symmetries.

Let U denote the envelope of the reference monomer (i.e. that on which both non-crystallographic and crystallographic symmetries act as the identity transformation), which is first

repeated by m ncs operations $\mathbf{x} \rightarrow \mathbf{C}_k \mathbf{x} + \mathbf{d}_k$ ($k \in K$, $K = \{1, ..., m\}$) before each of its images is repeated by the space group operations. The trigonometric contribution corresponding to a position $\mathbf{x} \in U$ (assumed, for simplicity, to be a general position with respect to all symmetry operations) is then

$$\Xi^{nc}(\mathbf{h}, \mathbf{x}) = \sum_{g \in G} \sum_{k \in K} e^{2\pi i \mathbf{h} \cdot [\mathbf{R}_g \mathbf{C}_k \mathbf{x} + \mathbf{R}_g \mathbf{d}_k + \mathbf{t}_g]} \qquad (2.1).$$

For variance and covariance calculations we have to calculate the expectation values of products of the form $\Xi^{nc}(\mathbf{h}, \mathbf{x}) \times \bar{\Xi}^{nc}(\mathbf{h}', \mathbf{x})$ when the random position $\mathbf{x}$ is distributed (say uniformly, to start with) in U. It is straightforward to obtain the expression :

$$\langle \Xi^{nc}_{\mathbf{h}} \times \bar{\Xi}^{nc}_{\mathbf{h}'} \rangle = \sum_{g \in G} \sum_{k \in K} \sum_{g' \in G} \sum_{k' \in K} e^{2\pi i (\mathbf{h} \cdot \mathbf{t}_g - \mathbf{h}' \cdot \mathbf{t}_{g'})}$$
$$\times e^{2\pi i [(\mathbf{R}_g^T \mathbf{h}) \cdot \mathbf{d}_k - (\mathbf{R}_{g'}^T \mathbf{h}') \cdot \mathbf{d}_{k'}]} \mathcal{G} [\mathbf{C}_k^T \mathbf{R}_g^T \mathbf{h} - \mathbf{C}_{k'}^T \mathbf{R}_{g'}^T \mathbf{h}'] \qquad (2.2)$$

where $\mathcal{G} = \frac{1}{\mathrm{vol}(U)} \overline{\mathcal{F}}[\chi_U]$ is the interference function, i.e. the normalised transform of the indicator function χ_U of U.

This expression shows that the complex correlation coefficient between structure factors belonging to reflexions whose orbits under the combination of local and global symmetries contain a pair of points closer than the spacing between integral lattice points can have a modulus arbitrarily close to unity. These correlations produce not only (for $\mathbf{h} \neq \mathbf{h}'$) the *approximate symmetry in reciprocal space* whose detection is normally attempted by means of the self-rotation function, but also (for $\mathbf{h} = \mathbf{h}'$) *intensity modulations* related to the non-crystallographic rotations and to the associated translations (up to those available "for free" to fix the origin). More specifically, *spikes* of normalised intensity will be found along directions in reciprocal space for which Miller indices $\mathbf{h}$ satisfy the relation

$$\mathbf{C}_k^T \mathbf{R}_g^T \mathbf{h} \approx \mathbf{C}_{k'}^T \mathbf{R}_{g'}^T \mathbf{h} \quad \text{for some } g, g' \in G \text{ and } k, k' \in K \qquad (2.3)$$

and the variation of normalised intensity *along the directions of the spikes* is related to the ncs translations. This is a generalisation of what happens in the prototype case of crystallographic symmetry, e.g. for a screw axis : the direction of the rotation axis corresponds to a spike since some of the reflexions found along it are systematically stronger than average; and the modulation of intensity along that direction (i.e. the possible systematic absences) reveal the nature of the associated non-primitive translation. By judiciously exploiting these predicted intensity modulations, it may be possible in some cases to characterise the ncs rotations directly, without recourse to a self-rotation function, and to determine the associated translations without heavy-atom derivatives.

Expression (2.2) also shows that the second-order moments from which all joint probability distributions of structure factors will be built, and all likelihood functions will be derived by integrating out the phases, depend not only on the assumed ncs elements but also on the assumed

envelope U . For the purpose of using such likelihood functions to validate or refine these assumptions, it would be desirable to have an automatic procedure for deriving one of these classes of assumptions from the other, or at least to check their mutual consistency. If an envlope is assumed, then there are various packing functions which will help rule out those ncs elements which would give rise to unacceptable clashes. In the reverse direction, i.e. in order to derive an initial "default" envelope from the specification of a set of ncs elements, a procedure first used in two dimensions by Carter et al. (1990) can be generalised to deal with any proper ncs defined by a group of local symmetries around a given or assumed centre Ω . In brief, it consists in saying that a point x belongs to U if all its images under the local ncs operations are closer to Ω than to any space-group equivalent Ω_g of Ω or to a lattice translate of such an Ω_g . The advantage of this first guess for U is that it is a convex (Voronoi) polyhedron, whose images under space group operations can touch each other but do not overlap, and that it fills up as much of the asymmetric unit as possible under these restrictions (hence is a "most conservative convex approximation" to the true envelope). The fact that it is implicitly defined in terms of the local symmetry group and of the position of Ω makes it an attractive choice for an initial likelihood-based search for these elements. Prior to an actual likelihood calculation, examination of the volume of U and of its connectivity when considered with its neighbours would already help eliminate non-sensical arrangements.

Once initial ncs elements and a starting envelope have been found in this way, the joint probability distribution of any collection of structure factors with strong normalised amplitudes can be set up. Note that the normalisation in question is carried out on the basis of the generalised intensity statistics derived from (2.2); this downweights those amplitudes whose strength may be due to *symmetry alone*, so that those whose strength is due to *density variation within U* may be identified and their interactions exploited. These joint distributions are singular since they are concentrated in the allowed subspace of structure factor space defined by the ncs conditions (i.e. the eigenspace for eigenvalue 1 of Crowther's **H** matrix). Furthermore they are non-uniform in that subspace and give strong preferential indications for starting phase sets.

Conditional probability distributions of structure factors, given assumed phases for those attached to the reflexions of a "basis-set", can then be built by the maximum-entropy method to propagate phase information outside the basis set. Phases are extended by this procedure much faster than by simply back-transforming a symmetry-averaged and solvent-flattened map. Furthermore, these conditional distributions can to a large extent impose the ncs constraints *in advance* to the extended phases, yielding centroid maps which will need much less symmetry filtering than those produced by Sim's formula (which ignores all statistical correlations between the phases of distinct reflexions).

Possible ambiguities in the phase extension (i.e. pseudo-symmetry and/or strong multimodality of the conditional probability as a function of the non-basis phases) can be handled routinely within the tree-directed search strategy (Bricogne, 1984) now proven effective in a variety of contexts for phase extension and *ab initio* phasing (Bricogne & Gilmore, 1990 ; Bricogne, 1993).

3. Conclusion.

I hope the reader will, on the basis of the evidence presented above, feel inclined to share my conviction that a substantial portion of the future of the Molecular (Re)Placement and Molecular Averaging methods will be dominated by statistical developments, and that a considerable strengthening of the current versions of these methods can be expected to result from the general adoption of this new point of view.

References.

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

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

BRICOGNE, G. (1991). *Acta Cryst.* A**47**, 803-829.

BRICOGNE, G. (1993). *Acta Cryst.* D**49**, 37-60.

BRICOGNE, G. & GILMORE, C.J. (1990). *Acta Cryst.* A**46**, 284-297.

BRÜNGER, A.T. (1990). *Acta Cryst.* A**46**, 46-57.

CARTER, C.W.Jr., CRUMLEY, K.V., COLEMAN, D.E., HAGE, F. & BRICOGNE, G. (1990). *Acta Cryst.* A**46**, 57-68.

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

HARADA, Y., LIFCHITZ, A., BERTHOU, J. & JOLLÈS, P. (1981). *Acta Cryst.* A**37**, 398-406.

JONES, T.A. & THIRUP, S. (1986). *EMBO J.* **5**, 819-822.

ROSSMANN, M.G. & BLOW, D.M. (1961). *Acta Cryst.* **14**, 641-647.

AUTOMATED REFINEMENT PROCEDURE

Victor S. Lamzin and Keith S. Wilson

European Molecular Biology Laboratory (EMBL), c/o DESY, Notkestrasse 85, D-2000 Hamburg 52, Germany.

In crystallography the experimental observables are the X-ray diffraction intensities, and hence structure factor amplitudes. From these the three-dimensional structure must be deduced. Knowledge of the amplitudes is unfortunately insufficient for direct calculation of the electron density. Figure 1 shows a flow chart of a crystal structure determination. For small molecules the phase problem is solved by either direct or Patterson methods through programs such as SHELX (Sheldrick, 1986) which assume data to atomic resolution (1.2-1.0 Å or better). The resulting atomic coordinates are refined automatically by least-squares minimisation of the residuals between observed and calculated amplitudes (or better intensities) with difference Fourier syntheses to update the model.

For proteins the situation is severely complicated by the lack of atomic resolution data. The phase problem is experimentally solved by multiple isomorphous replacement (MIR) or by molecular replacement (MR) using a similar known structure. There are usually large errors in the phases from this initial model. In addition at resolutions poorer than about 2.5 Å the number of parameters actually exceeds the observations even with isotropic atomic temperature factors. The observations are normally increased in addition to the X-ray data by including a set of stereochemical or energy restraints based on the known structures of small molecule models. However the radius of convergence of conventional least-squares is at best 1/3 of the resolution and the program cannot move atoms into new features of the map substantially different from the current positions. Incorporation of molecular dynamics aimed to extend the radius of convergence, e.g. the X-PLOR package (Brünger, 1988) allowing atoms to cross local barriers in the least-squares minimisation. However atoms still cannot move through other atoms and dynamics refinement requires a lot of computer time. Thus protein refinement remains a tedious and lengthy procedure with extensive iterative improvement of the current model and tedious manual rebuilding using computer graphics.

In summary restrained least-squares refinement does not provide an automatic way of refining protein structures at resolution lower than atomic to a final model. We propose an automated refinement procedure (ARP) for proteins, as indicated in Figure 1. ARP is comparable to the iterative least-squares/Fourier refinement of small molecule structures.

AUTOMATED REFINEMENT

There are a lot of regions which need to be substantially corrected in the initial protein model from molecular replacement, or built into an isomorphous density synthesis and preliminarily refined with constraints. A simple example for one-dimensional model of 4 atoms is shown in Figure 2, top. How do we pass from the wrong to the correct model ? Usually such corrections require a large amount of time, especially for big proteins, in identifying such pieces of structure and correcting them. Two possible refinement schemes, given a wrong model are shown schematically in Figure 2. Least-squares minimisation alone cannot find the global minimum in such a case.

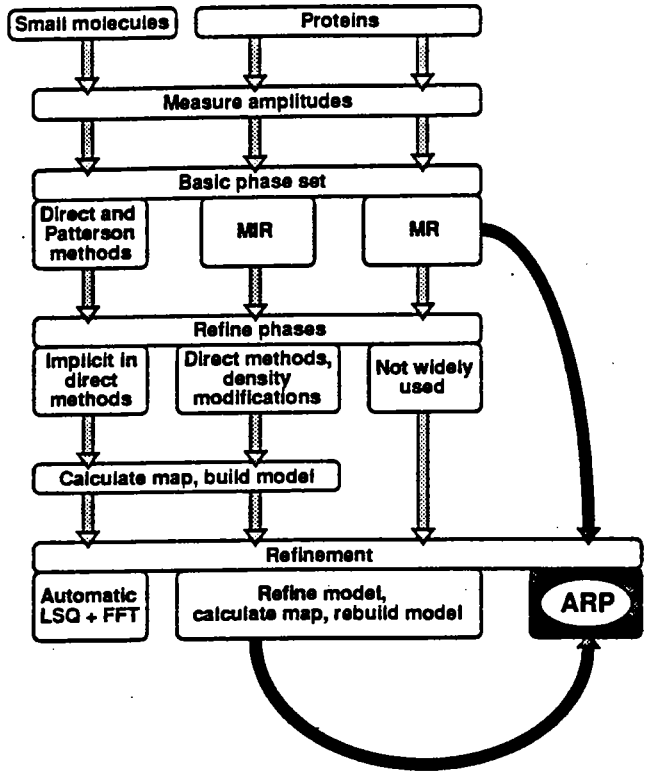


Fig. 1. A highly simplified flow chart of the steps in small molecule and protein crystallographic analyses.

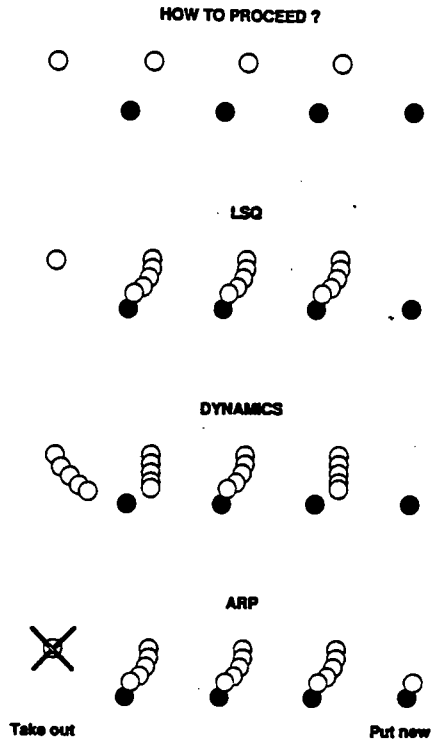


Fig. 2. A schematic representation of refinement of one-dimensional model. The white circles correspond to the initial model and the black ones to the final model (see text).

There are three steps in ARP. Firstly all atoms in the initial model are reset to atoms of the same type. This is not in fact essential and was applied here for convenience. Secondly the positional and thermal atomic parameters of the relabelled ARP model are refined by unrestrained least-squares using X-ray data using from the start data from the complete resolution range. Thirdly the ARP model is updated by: 1) Calculation of $(3F_o-2F_c)$ and $(2F_o-2F_c)$ electron density maps from the present model. 2) Rejection of a few atoms if they are in low $(3F_o-2F_c)$ density. The percentage rejected depends on the resolution. 3) Addition of new atoms found in positive difference density. The percentage of atoms to be added depends on the resolution, on the percentage rejected and the number of atoms in the initial model. In the examples described here the initial model was incomplete, without waters. Best results were obtained if the number of atoms was constantly increased during the refinement to a final value approximately 120 % the number of expected protein atoms. Usually 10 % of the excess atoms corresponded to real water molecules bound to the protein and 10 % to imitate pseudo solvent flattening. The crystallographic R-factor and values of high moments of the maps can be used as criteria of convergence. In addition the ARP model can be visually inspected at any point during the procedure, to ensure it continues to resemble the protein in terms of stereochemistry. Relabelling of the ARP model to the correct protein atoms can be carried out using computer graphics, or automatically. All steps in ARP can be iterated in a completely automated manner until convergence is achieved.

ARP, Figure 2, updates the model by removing atoms in weak density in the present electron density and placing new atoms in roughly the correct position from the difference density, followed by unrestrained least-squares optimisation of their parameters. This needs X-ray data to a resolution where the atomic positions can be approximately estimated from the density, ideally at least 2.0 Å. The criteria by which atoms can be rejected from or added to the model are many and varied. We have used very simple, but apparently effective, methods in these first applications. Much more elegant schemes can be envisaged for the future.

For unrestrained refinement of the model and calculation of density maps the SFKH, PROLSQ (Hendrickson & Konnert, 1981), RSTATS and FFT programs (CCP4, 1979) were used. Analysis of electron density maps, rejection of the atoms located in low density, addition of new atoms according to difference density map, resetting of the atom name and keeping the model within the asymmetric unit were performed by the ARP program.

APPLICATIONS

We here describe the application of ARP to four proteins, Table 1. For all four, data were collected on an imaging plate scanner built in-house, using synchrotron radiation from beam line X31 at EMBL Hamburg. For each the data were more than 90 % complete overall, more than 98 % at low resolution, and more than 66% of the theoretical unique intensities were greater than 3 σ in the outer resolution shell. We define the initial model as that input to ARP, the ARP model as that refined during ARP and the final model as that obtained by conventional restrained least-squares refinement. The maps calculated from these models are described as initial, ARP and final maps respectively and similar terminology is used for atoms and for phases calculated from the models.

1) Apo Formate Dehydrogenase

The crystal structure of NAD-dependent apo formate dehydrogenase (FDH) was solved by molecular replacement using the refined holo model (Lamzin *et.al.*, 1992). The model was refined by CORELS in resolution range from 8.0 to 4.0 Å using each of the four FDH domains as a separate rigid body and the resulting model was used as the initial model for the development of the automated refinement procedure (ARP). All data in the range 10.0 to 1.8 Å were used from the start. The r.m.s. deviation from the initial model after rigid body refinement to the final model was about 0.5 Å for CA atoms and about 0.9 Å for all protein atoms. The starting R-factor was 42.6 %. 24 cycles of ARP were run. Each consisted of 1 cycle of unrestrained refinement in the whole resolution range, followed by removal of the 0.3 % atoms in the weakest ($3F_o$-$2F_c$) density and the addition of the 1.0 % peaks in the strongest ($2F_o$-$2F_c$) density. The R factor fell to 13.8 %.

TABLE 1. Characteristics of the crystals of the four proteins refined using ARP.

Protein	Source	Space group	Asymmetric unit content	Resolution	Reference
Apo Formate Dehydrogenase	*Pseudomonas* sp.101	$P2_1$ a=110.5, b=54.5, c=70.3 Å, β=101.9°	2 x 393 residues	1.8 Å	Lamzin *et.al.*, to be published
Narbonin	*Vicia narbonesis* L	$P2_1$ a=46.9, b=75.5, c=50.9 Å, β=120.5°	about 33 kDa	1.8 Å	Hennig *et.al.*, 1990
Trypsin-like Proteinase	*Fusrium oxysporum*	$P2_12_12_1$ a=40.4, b=51.5, c=69.3 Å	190 residues	1.4 Å	Dauter *et.al.*, to be published
Inorganic Pyrophosphatase	*Saccharomyces cerevisiae*	$P2_12_12_1$ a=116.5, b=106.3, c=56.1 Å	2 x 281 residues	2.4 Å	Chirgadze *et.al.*, 1991

TABLE 2. Formate dehydrogenase. A comparison of conventional restrained refinement (Lamzin *et al.*, to be published) and ARP procedures. For both the initial model is from molecular replacement using holo FDH followed by 3 cycles of refinement with four rigid bodies (see text).

Methods used	R factor (%) 10.0 - 1.8 Å	r.m.s. deviation in bond length (Å)	Real time
51 cycles of PROLSQ, 1 cycle of X-PLOR, plus much manual rebuilding	16.9	0.022	2 months
Automated Refinement Procedure, automatic rebuilding (90 % of the model), manual rebuilding (10 % of the model), 10 cycles of PROLSQ	17.5	0.021	1 week

Most ARP atoms were located rather close to initial protein atoms and the interatomic distances in the ARP model were similar to the corresponding distances in the protein. 90 % of the new protein model was automatically constructed using a rebuilding program. The remaining 10 % of the structure corresponded to regions with large shifts from the initial model were rebuilt manually from the ARP model and ($3F_o$-$2F_c$) density. The new protein model was subjected to restrained refinement using PROLSQ, mainly to improve the stereochemistry. After 10 cycles the R factor dropped to 17.3 % and the r.m.s. deviation in bond lengths to 0.021 Å.

A comparison of ARP and conventional restrained refinement is shown in Table 2. The conventional refinement of apo FDH (51 cycles of PROLSQ, XPLOR and a lot of manual intervention) took about 2 months (Lamzin *et al.*, to be published). In contrast a model was obtained with only 2 days for ARP plus 3-4 days to rebuild the 10 % of the structure which could not be automatically assigned and to tidy up the stereochemistry. The ARP model is essentially identical to that from standard methods. An example is shown in Figure 3 with ARP electron density. There is a movement of a large loop between the initial and the final model, with a systematic shift of about 2 Å. The ARP atoms do indeed look protein-like (left). The ARP density is certainly good enough to allow the correct model to be built. The same region with both the ARP and the final models is on the right. This indicates that ARP converges on the correct solution with a model similar to the final one.

Thus application of ARP to FDH at 1.8 Å resolution gave a density map with parameters almost identical to the final map. The resulting model was refined by 10 cycles of restrained refinement and proved to be nearly identical to the final model from conventional methods. In real time use of ARP for apo FDH made the refinement approximately 10 times faster.

2) Narbonin

This example shows the application of ARP to a partial model without complete amino acid sequence information, Table 1. A Narbonin model was built into a 2.2 Å resolution MIR map and preliminarily refined with PROLSQ to an R-factor of 29.9 % to 1.8 Å by Michael Hennig and coworkers (to be published). This partial model had several breaks in the polypeptide chain and contained 85 % of the protein atoms. The density maps did not clearly show which regions of the model should be corrected or how the missing regions should be built, especially difficult in the absence of chemical sequence.

6 steps of ARP were carried out. In each step several cycles of unrestrained refinement of x, y, z and B parameters for each atom were carried out with all data in the resolution range 10.0 to 1.8 Å followed by updating of the model by rejection of the 2-10 % atoms in lowest ($3F_o$-$2F_c$) density and addition of the 10-15 % in highest ($2F_o$-$2F_c$) density. The total number of atoms was gradually increased. The R factor dropped to 14.9 %. The initial protein model

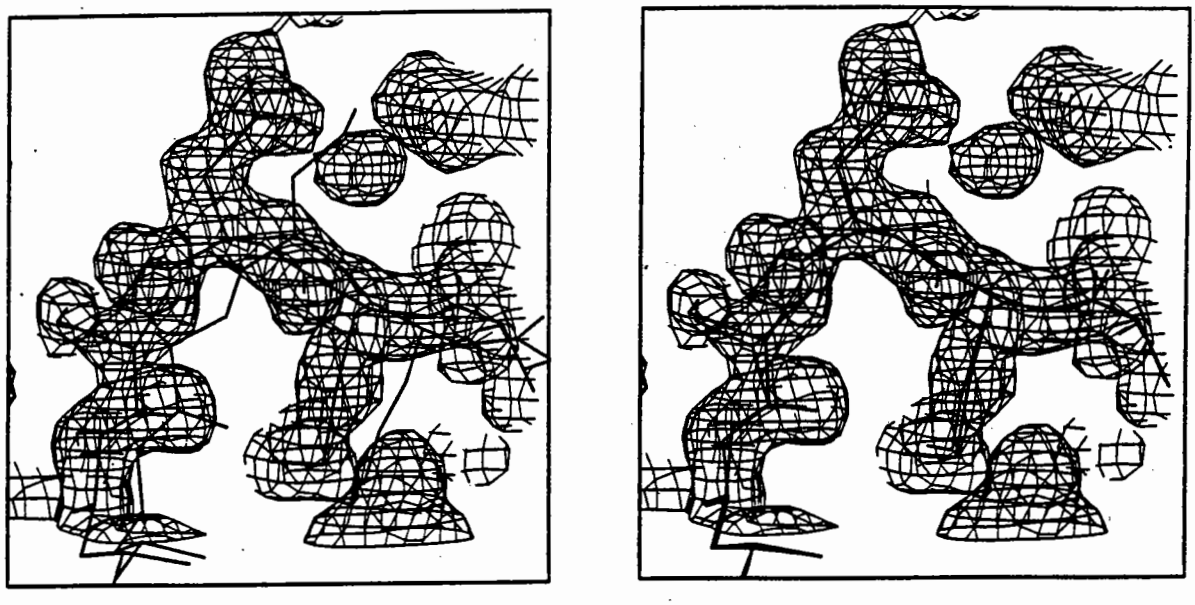


Fig. 3. Formate dehydrogenase. Approximately 2 Å shift of the loop from the initial to the final model. The ($3F_o$-$2F_c$) electron density shown is from the ARP model contoured 1 σ above mean density. Left - Initial model and ARP atoms connected according to protein interatomic distances. Right - ARP and final models.

was corrected using the ARP electron density and the resulting model was refined using PROLSQ with restraints to an R factor of 20.5 %. ARP was applied for a second time. The model was rebuilt in several parts and refined with restraints to an R factor of 16.9 %.

The application of ARP to narbonin at 1.8 Å resolution resulted in substantial improvement of the density especially in places where the initial map was poor. The positions of ARP atoms generally corresponded to real atoms and the model was rebuilt from the new features in the ARP map. ARP was applied for a second time starting from the improved protein model and gave further improvement. The second ARP model was successfully rebuilt and refined by Michael Hennig.

3) Trypsin-like Proteinase

The refinement of a trypsin-like proteinase, Table 1, is an example of ARP at high resolution starting from a poor initial model. This model was obtained by MR using a bacterial proteinase with 50 % identity in primary structure and a slightly different number of residues. After 3 cycles of low resolution rigid body refinement the R-factor was 50.0 % in the resolution range from 10.0 to 1.4 Å. Each step of ARP consisted of 1 cycle of unrestrained refinement in the whole resolution range, rejection of the 1.0 % worst atoms and addition of 1.5 % new peaks found in the difference density. After 50 cycles the R factor dropped to 18.9 %.

Figure 4 shows the place where the polypeptide chain moves about 4 Å from the initial model. There is a dramatic improvement compared to the initial density. The ARP model resembles a protein model and the quality of the ARP density is good enough to build the correct model. About 50 % of the new protein model was built automatically and the remaining atoms were corrected on the basis of the ARP model and density map. The model has been refined with restraints to an R factor of 13.7 %, which dropped to 12.4 % when hydrogen contributions were included. In real time the refinement of this protein took less than 2 weeks starting from the MR solution.

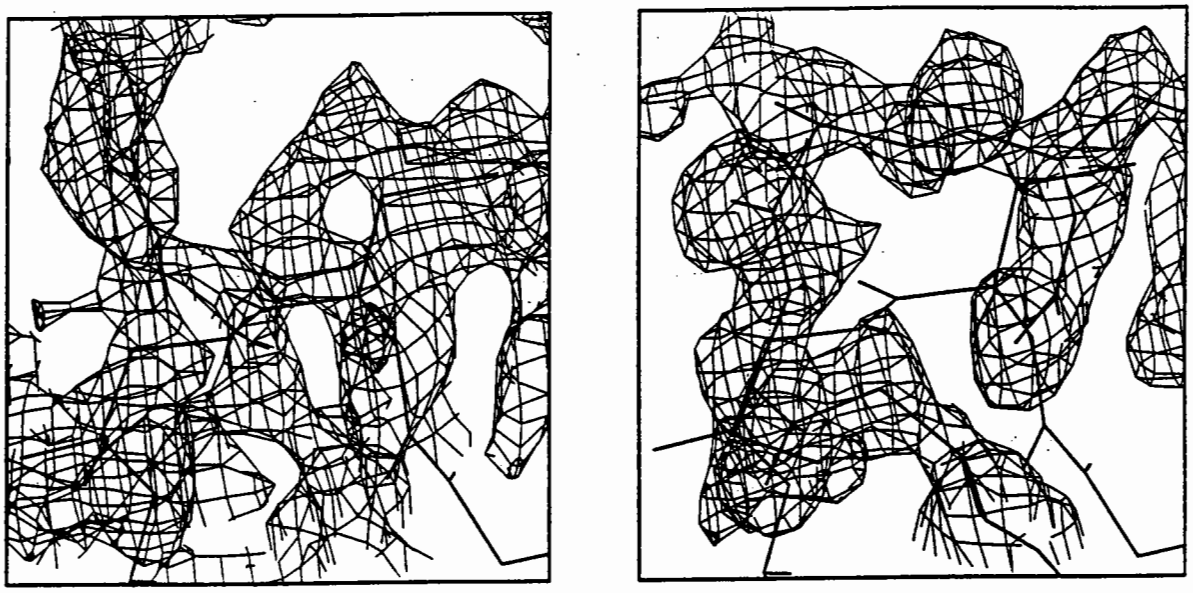


Fig. 4. Trypsin-like proteinase. The polypeptide chain moves from the initial model. Left - initial model and initial density. Right - the same place with ARP density and both the ARP and initial models. ARP atoms are connected according to protein interatomic distances. The ($3F_o$-$2F_c$) electron densities are contoured 0.8 σ above mean density.

4) Pyrophosphatase

The refinement of pyrophosphatase (PP), Table 1, is an example of ARP at medium resolution. The MR model of PP, preliminarily refined by PROLSQ to an R-factor of 27.3 % in the resolution range from 7.0 to 2.4 Å (E. Harutyunyan and coworkers, to be published), was used as the initial model for ARP. 50 cycles of ARP were carried out. Each involved 1 cycle of unrestrained refinement followed by the removal of 9 "weakest" atoms, i.e. 0.2 % of the total number of atoms, and addition of 9 new atoms. The R factor dropped to 12.6 %. The total number of atoms was kept constant because the limited resolution does not allow the introduction of a large number of waters to the model. ARP effectively resulted in solvent flattening in spite of the fact that the molecular boundaries were not specially set. The distribution of the new electron density is considerably better according to histogram matching criteria. However at 2.4 Å resolution the ARP model was not sufficiently good to allow automatic reconstruction of the protein from the ARP model to proceed with confidence. Nevertheless the ARP map is considerably better and less noisy than the initial map. New features and connectivities appeared in several places where the initial map was relatively poor. Building of the new protein model according to the ARP electron density has been carried out and the R factor is currently 19.0 %.

CONCLUSIONS

ARP has been successfully applied to four proteins. It is clearly more powerful when high (better than 2.0 Å) data are available, but nevertheless gives definite improvement, at least in the density map, even at 2.4 Å. The better the initial model, the better the result, at least in the present implementation. ARP resembles the use of alternating cycles of least-squares and difference Fourier syntheses in small molecule crystallography where atomic resolution data are available. The FFT is essential if the calculations for proteins are to be carried out in a tractable time.

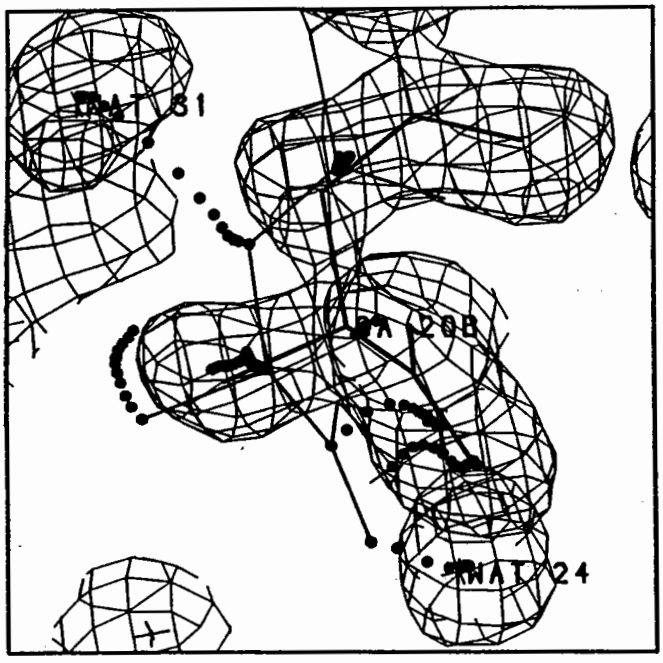


Fig.5. A region of FDH model near Ala20. The ($3F_o$-$2F_c$) electron density after ARP is contoured at 1 σ above the mean density. Thin line corresponds to initial model and thick line to final model. Points indicate intermediate positions of ARP atoms during refinement.

ARP can produce a corrected model and thus improve the calculated density by introducing atoms in "new" places. An example of how a set of atoms in the initial model is continuously changed by ARP to the new set of atoms is shown in Figure 5 for Ala20 in FDH. The ARP density is presented and the points indicate intermediate positions of the atoms during ARP. CB in the initial model moved out of density and was removed when the model was updated. CA in the initial model moved to the true position of the CB atom. The true place for the CA atom was occupied by a new atom picked up in positive difference density after several cycles of ARP. The initial main chain N and O atoms moved to what were in reality water positions during ARP and their places were taken up by C and CA atoms from neighbouring residues. The trajectories shown in Figure 5 are generally not linear. In ARP the atoms do not move in a smooth and continous manner as is more typical for conventional least-squares. There are changes of direction in the trajectories, which can be quite sharp in ARP especially after the model has been updated.

It is difficult to give an accurate assessment of the "radius of convergence" of ARP. Both the completeness of the initial model and how similar it is to the final model will affect the ability of ARP to converge on the correct structure. In FDH at 1.8 Å resolution a MR model including all the protein atoms with a root mean square deviation in CA positions about 0.5 Å compared to the final model was easily refined using ARP. This remains true if the model is mostly correct and about 85 % complete, but still has several uninterpreted or even incorrect regions, as for narbonin. At 1.4 Å resolution even in the case of poor initial model, as for the trypsin-like proteinase, ARP easily resulted in a correct solution.

ARP can refine a model made up of a subset of atoms most of which are in essentially the correct position, Figure 1. The initial set of atoms comes from a protein model and the initial map is calculated with phases from this model. The initial protein phases are not used again. The refinement imposes atomicity and real protein structure, arising implicitly from the high resolution data coupled with the reasonable starting model.

An extended version of this report has been submitted for publication in Acta Cryst. D (1993).

REFERENCES

Brünger, A.T. (1988). *X-PLOR Manual,* Version 1.5, Yale University.

CCP4 (1979). The SERC (UK) Collaborative Computing Project No. 4: a Suite of Programs for Protein Crystallography, distributed from Daresbury Laboratory, Warrington WA4 4AD, UK.

Chirgadze, N.Y., Kuranova, I.P., Nevskaya, N.A., Teplyakov, A.V., Wilson, K.S., Strokopytov, B.V., Harutyunyan, E.H. and Höhne, W. (1991) Kristallografiya (USSR) 36, 128-132.

Hendrickson, W.A. & Konnert, J.H. (1981). In *Biomolecular Structure Conformation, Function and Evolution* (Srinivasan, R. ed.) vol.1, pp.43-57 Pergamon Press, Oxford.

Hennig, M., Schlesier, B., Pfeffer, S. and Höhne, W.E. (1990) *J.Mol.Biol.* 215, 339-340.

Lamzin, V.S., Aleshin, A.E., Strokopytov, B.V., Yukhnevich, M.G., Popov, V.O., Harutyunyan, E.H. and Wilson, K.S. (1992) *Eur J. Biochem. in press.*

Sheldrick, G.S. (1986) SHELX: a program for crystal structure determination, Univ. of Göttingen, Germany.

From the molecular replacement solution to the refined structure

E J Dodson, Department of Chemistry, University of York
York, YO1 5DD, UK

The previous papers have shown that the path of molecular replacement often does not run smoothly, and when an apparent solution has been found we still face the challenge of proving it approximately correct. The new structure is unlikely to match the homologous model completely; it may have sequence differences, have crystallised in a different form, have different chemistry, and these differences will probably be reflected in the molecule's fold. There is also a possibility that some error has been made in the application of the technique!

I will outline a few verification criteria which can help decide whether a solution is correct.

1) Verify the result by applying the rotation and translation to the model and repeating the calculations with this model. The result should have rotation angles of (0.0,0.0,0.0) and a translation vector of (0.0,0.0,0.0).

2) Check the crystal contacts. Bad crystal contacts can rule out a solution most convincingly. If the symmetry equivalent molecules overlap, then there is something seriously wrong.

There are a variety of ways of checking contacts; graphics packages should do this, or there are various programs which tabulate bad contacts. The refinement packages such as PROTIN and XPLOR list unacceptable contacts, and programs such as DISTANG or CONTACT list all contacts within set limits.

If there are many bad contacts and no reasonable explanation abandon the solution. DO NOT try to " refine" your way out of trouble - refinement procedures often "cure" the bad contacts without reaching a correct solution.

However the absence of bad contacts does not guarantee a correct solution. Crystals with a high solvent content have minimal packing restrictions, and in crystals with low solvent content, the molecular fold may have been forced to change considerably to accommodate crystal packing.

3) R factor for model. The initial R factor obtained for the "correctly" positioned molecule is not a good criteria: our initial R factors always seem to be in the region of 60% whether a solution is essentially correct or not.

4) Rigid body refinement. We believe that if rigid body refinement either of the whole molecule or of a few well defined domains improves the initial R factor significantly giving reductions of 2–3%, the solution is worth pursuing.

There are several ways of doing this: packages such as MERLOT and Jorge Navaza's include it, XPLOR has an option for rigid body, and CORELS is another possibility. Zygmunt Derewenda suggested a simple technique of doing several loosely restrained cycles of conventional refinement and then fitting the core of the starting model on to the shifted coordinates and starting

again. This can be done with the whole molecule (he was very successful with haemoglobin which is a particularly rigid protein), or with separate domains (Derewenda, 1989). It works well when there are small corrections to be made to the rotational parameters; less well or not at all when the whole molecule needs translating, or when there is considerable change between the model and the new structure (this may well be the case when the model is based on a unrefined solution itself).

5) Chemical verification. It is always worth looking for chemical clues that you are on the right track. York examples include:

a) An insulin mutant crystallised in cubic insulin, space group $P2_13$, where the correct solution positioned the molecule with a histidine pointing towards the 3 fold axis, within bonding distance of density which proved to be due to an unexpected zinc atom (Turkenburg, 1992).

b) Another insulin mutant which was meant to crystallise as a monomer. However the solution positioned the molecule with the dimer face along a two fold axis, in the same conformation as the native structure (Xiao Bing, unpublished material).

6) Consistency between different crystal forms. If there are two forms of the molecule, related by a known rotation, the solutions mapping the model onto both the forms must be related by the same rotation (Bi et al, 1983; Swift et al, 1991).

7) Heavy atom information. In my experience if you have some isomorphous phases, and an orientation from the rotation function the PHASED TRANSLATION function always works. This defines the translation which gives the best overlap between the corrected orientated model density, and the isomorphous map (Read, 1988; CCP4 documentation, 1992). And if the solution has any truth in it, the phases it generates should be good enough to phase a Fph -Fp difference fourier and show the heavy atom sites which should lie in chemically sensible places.

But beware - I used to think this was an infallible proof, but our experience has shown that if five of the molecular replacement parameters are correct, and one is out by some fraction of a unit cell, then a considerable percentage of the reflections will be phased correctly, enough to show heavy atom positions and disulphides in the expected places.

8) Most important - behaviour during refinement. The 1990 Study Weekend addressed the problem of detecting inadequacies in refinement and the criteria outlined by David Blow must be satisfied (Blow, 1990).

Detection of errors

a) It is often quite difficult to satisfy oneself that a molecular replacement solution is correct, and there have been several cases of incorrect structures being published. Simon Phillips discusses the problems in the structure determination of turkey egg lysozyme in the 1990 Study Weekend book (Phillips et al, 1990). Problems can arise because there is not enough X-ray data for sensible refinement. You can obtain a molecular replacement solution with

4Å data but you cannot refine a structure. Since the questions which these structures are meant to elucidate are usually more subtle than those asked of a new structure, where the fold itself is of great interest, we must take even greater care to refine them sensibly. To do this it is usually necessary to acquire high resolution data.

b) The original structure will probably have had good geometry, so the useful geometric checks which detect errors in isomorphous structures will not be possible during the early stages of refinement.

c) What clues can we look for to detect errors? In two cases worked on in York, there were errors in the first solution. One is an orthorhombic insulin done by Urszula Derewenda, where the molecular replacement solution was in error in that the z translation used was about 0.25 of a unit cell away from the true value. The appearance of the refinement is described in the Daresbury Study Weekend book (Derewenda et al, 1990). It refined to R factor of 26% at 2.0Å but the density was broken along the main chain, and the helix became distorted. New data gave a correct solution, and refinement proceeded smoothly. The temperature factors indicated that there was a problem; the internal alpha helix where we know B factors are usually significantly lower than average did not show this, and the root mean square difference in B factors within each residue was suspiciously high (this phenomenon is of course reflecting the breaks in the chain). In Taka amylase the original solution fitted well for the larger domain but the small domain had rotated by 5.9° in the $P2_12_12_1$ structure. Again this part of the structure did not give clean 2Fo-Fc maps, and it was not clear how to correct it (Swift et al, 1991).

References:

Bi, R.-C., Cutfield, S.M., Dodson, E.J., Dodson, G.G., Giordano, F., Reynolds, C.D. and Tolley, S.P. (1983) Acta Cryst. B39, 90-98.

Blow, D. Proceedings of the CCP4 Study Weekend, Daresbury Laboratory, 26-27 January, 1990 (compiled by K Henrick, D S Moss and I J Tickle) pp 115-117.

Derewenda, U., Dodson, E.J., Dodson, G.G., Hodgkin, D.C and Swift, H.J. Proceedings of the CCP4 Study Weekend, Daresbury Laboratory, 26-27 January, 1990 (compiled by K Henrick, D S Moss and I J Tickle) pp 103-113.

Derewenda, Z.S. (1989) Acta Cryst, A45, 227-234.

Phillips, S.E.V., Somers, W.S., Bhat, T.N. and Parsons, M.R.. Proceedings of the CCP4 Study Weekend, Daresbury Laboratory, 26-27 January, 1990 (compiled by K Henrick, D S Moss and I J Tickle) pp 63-72.

Read, R.J. and Schierbeek, A.J. (1988) J. Appl. Cryst. 22, 490-495.

Swift, H.J., Brady, R.L., Derewenda, Z.S., Dodson, E.J., Dodson, G.G., Turkenburg, J.P. and Wilkinson, A.J. (1991) Acta Cryst, B47, 535-544.

Turkenburg, J., DPhil Thesis, University of York, 1992.

AMoRe: A NEW PACKAGE FOR MOLECULAR REPLACEMENT

Jorge **Navaza**
UPR 180 CNRS, Lab. de Physique, Faculté de Pharmacie.
92290 Châtenay Malabry, France.

The two main steps in the Molecular Replacement Method are the orientation and the translation searches: a known molecule or molecular fragment has to be positioned in the unit cell of the unknown crystal structure. This six-dimensional search is seldom done in practice because of computing time limitations. The usual approach consists in performing the rotation and translation searches separately. Different techniques have been proposed and the practical recipes to perform the actual calculations have been extensively discussed. Most of these techniques are now implemented in available packages.

Many strategies have been developed to enhance the significance level of rotation function (RF) and translation function (TF) potential solutions. When the correct position is among the top peaks of the combined RF and TF outputs, the molecular replacement problem is solved in a few number of steps. Otherwise the current implementation of the available software requires a good deal of manual intervention. In difficult cases the procedure is stopped after a certain number of unsuccessful trials, and eventually restarted with another, hopefully better, search model. The situation becomes more complicated when several molecules or molecular fragments have to be positioned in the asymmetric unit.

Although easy problems are most welcome, we will be mainly concerned with difficult "solvable" ones (this qualification can only be done a posteriori). We may attempt to define a solvable problem as one which could have been solved by a six-dimensional search. This assumes: acceptable quality of the diffraction data of the target crystal; enough accuracy of the search models.

A new strategy has been developed. It is based on the observation that, under the above assumptions, RF calculations yield maps where the correct peaks are above about 50% of their maximum values. This result is the cumulated experience using the ROTING program with Fab structures (Alzari & Navaza, 1991), medium-size molecular structures (Navaza, de Rango & Sarrazin, 1991) and other biologic molecules (Saludjian, Prangé, Navaza, Menez, Guilloteau, & Ducruix, 1992). Then, instead of "improving" the search model in order to promote the correct peak to the top of the RF output, we explore all the retained orientations (sometimes of the order of some hundreds) in a fast, automatic way. Perhaps the only new physical idea behind the procedure is the enhancement of rotation peaks by skipping low angular resolution contributions. However, on mathematical and numerical grounds, the package is based on new algorithms and conception; indeed, the automation requires something more than a simple collage of existing procedures.

The most time consuming and tedious part of available softwares concerns manipulation of coordinates, in order to calculate structure factors. The latter have to be computed for each orientation of the search model and at each

iteration during fast rigid-body refinements. We have thus limited its use to a strict minimum: atomic coordinates are used only once to compute the Fourier coefficients corresponding to the model electronic densities of the search molecules or molecular fragments. Subsequent structure factor calculations are performed by simple interpolation: for a given rotation matrix $\mathbf{R}(\alpha,\beta,\gamma)$ and fractional translation vector $\mathbf{x}$, the calculated Fourier coefficient $F(\mathbf{H})$ in the crystal frame, in terms of those of the search model $f(\mathbf{h})$ calculated in a big (ideally infinite) box, is given by

$$F(\mathbf{H}) = \sum_s f[\mathbf{H}\mathbf{M}_s\mathbf{D}\mathbf{R}(\alpha,\beta,\gamma)\mathbf{O}] \exp(2\pi i \mathbf{H}\mathbf{t}_s) \exp(2\pi i \mathbf{H}\mathbf{M}_s\mathbf{x}) \ . \qquad (1)$$

where $\mathbf{M}_s|\mathbf{t}_s$ stands for the symmetry operators, $\mathbf{O}$ and $\mathbf{D}$ for the orthogonalization and deorthogonalization matrices, respectively. We notice that the Fourier transform of the rotated model electron density $f(\mathbf{h})$ must be calculated at the reciprocal vector

$$\mathbf{h} = \mathbf{H}\mathbf{M}_s\mathbf{P}\mathbf{R}(\alpha,\beta,\gamma)\mathbf{O} \ , \qquad (2)$$

which does not in general lie on the Bravais lattice of the model box. Therefore, its computation requires a linear interpolation on $\mathbf{h}$.

The **A**utomatic-**Mo**lecular-**Re**placement package here presented consists of three main programs: ROTING,TRAING and FITING. Two other ones, SORTING and TABLING cast the input data into a suitable representation. TABLING further produces the arrays of Fourier coefficients of the model densities corresponding to the search molecules or molecular fragments. Formula (1) is one of the corner-stones of **AMoRe**, as it allows for a fast and automatic multiple exploration.

The other corner-stone of the package is the ROTING program which calculates rotation functions. The formulation is essentially that of Crowther, but uses numerical integration instead of Fourier-Bessel expansions in the radial variable (Crowther, 1972; Navaza, 1987). Also, important numerical instabilities were removed by using a new algorithm to calculate the reduced rotation matrices $d^{\ell}_{mm'}$ (Navaza, 1990). As a consequence, more accurate results are obtained and there are no limitations concerning the ratio (outer radius of integration)/(data resolution) . The fast rotation function takes the factorized form

$$RF(\alpha,\beta,\gamma) = \sum_{\ell=0}^{\infty} \sum_{m,m'=-\ell}^{\ell} C^{\ell}_{mm'} \, d^{\ell}_{mm'}(\beta) \, e^{-i(m\alpha+m'\gamma)} \ . \qquad (3)$$

The $C^{\ell}_{mm'}$ depend on the intensities of both crystals and on the definition of the spherical domain of integration, but not on angular variables. They are given by the expresion (a=inner radius, b=outer radius),

$$C^{\ell}_{mm'} = (3/4\pi)\,(b^3-a^3)^{-1} \int_a^b c^{(1)}_{\ell m}(r)\, c^{(2)}_{\ell m}(r)^*\, r^2\, dr\,, \qquad (4)$$

in terms of the radial functions

$$c_{\ell m}(r) = 4\pi \sum_{\mathbf{H}} |F_{\mathbf{H}}|^2\, (-i)^{\ell}\, j_{\ell}(2\pi h r)\, Y_{\ell m}(\hat{\mathbf{H}})^*\,. \qquad (5)$$

j_{ℓ} and $Y_{\ell m}$ stand for the spherical Bessel function and the spherical harmonic of order ℓ, respectively, and $\hat{\mathbf{H}} = \mathbf{H}/h$ denotes the angular part of vector **H**. The code computes (4) using a Gauss-Legendre quadrature formula

$$C^{\ell}_{mm'} = \sum_{n=1}^{N} c^{(1)}_{\ell m}(r_n)\, c^{(2)}_{\ell m'}(r_n)^*\, w_n\,, \qquad (6)$$

where r_n denotes the n^{th} integration abscissa and w_n its corresponding weight. ROTING offers the possibility of using the information of the self-rotation function, to compute locked rotations.

The translation functions so far implemented are Crowther & Blow's overlap function (1967) and a simplified version of the full symmetry phased translation function developed by Bentley (1992). Soon, we hope to include a correlation function using FFT's written by Sarrazin. In all cases the output of the TRAING program is the correlation coefficient and the R-factor corresponding to the highest peaks of the TF employed. All the available options include the possibility of fixing a certain number of positioned fragments. In this case the unit cell of the TF is the whole crystal cell. Otherwise it is the Cheshire cell (Hirshfeld, 1967), which results in a substantial gain in computing time. Given the data, the program choses the pertinent cell.

Finally, the program FITING performs fast rigid-body refinement along the lines proposed by Huber and Schneider (1985), in an improved version described by Castellano, Oliva and Navaza (1992). If there are N search models, the target Fourier coefficient is the sum of N contributions like (1). Assuming that the individual model Fourier transforms have been set to a common scale, the function minimized by FITING is

$$Q = \sum_{\mathbf{H}} \{|F^{ob}(\mathbf{H})| \exp(-B|\mathbf{H}|^2) - \lambda |\sum_{n=1}^{N} F_n[\alpha_n,\beta_n,\gamma_n,x_n,y_n,z_n](\mathbf{H})| \}^2 . \qquad (7)$$

The minimization is alternately performed with respect to the positional parameters of each search model, keeping the others fixed. At each step the unique scale factor λ and the overall temperature factor B are chosen so as to minimize (7).

AMoRe has been succesfully employed in a number of difficult cases, where standard packages had failed. In particular, the antigen-antibody complex of Guinea-fowl Lysozyme and Fab F9.13, which crystallizes in $P2_1$ with two complex molecules in the asymmetric unit, was solved using as search models the Lysozime, the constant and the variable parts of an Fab (which amounts to roughly one sixth of the asymmetric unit content !). Some solutions were near the bottom of the list of peaks proposed by ROTING. Other examples are a Lipase and PGK, solved during a memorable nineteenth hole played in York, after the CCP4 Meeting.

Saludjian played an important role in the gestation of this package, convincing the author of its utility. We were also encouraged by the observation that anytime that a procedure has been automatised, not only the straightforward problems became trivial, but also new problems, so far considered as untractable, could be addressed. **AMoRe**, name proposed by Alzari, has not still attained a high degree of automation: a certain number of decisions have to be taken by the user, to avoid unnecessary computations.

References:

ALZARI, P. & NAVAZA, J. (1991). Computing School in Crystallography (Bischenberg, 29 Julliet-4 Aout 1990). "On the Use of the Fast Rotation Function".

BENTLET, G.A. & HOUDUSSE, A. (1992). Acta Cryst. A**48**, in press.

CASTELLANO, E., OLIVA, G. & NAVAZA, J. (1992). J. Appl. Cryst. **25**, 281-284.

CROWTHER, R.A. (1972). The Molecular Replacement, edited by M.G.

CROWTHER, R.A. & BLOW, D.M. (1967). Acta Cryst. **23**, 544-548.

HIRSHFELD, D. (1968). Acta Cryst. **24**, 301.

HUBER, R. & SCHNEIDER, M. (1985). J. Appl. Cryst. **18**, 165-169.

NAVAZA, J. (1987). Acta Cryst. A**43**, 645-653.

NAVAZA, J. (1990). Acta Cryst. A**46**, 619-620.

NAVAZA, J., de RANGO, C. & SARRAZIN, M. (1991). XIII European Crystallographic Meeting, Trieste, Italie, 26-30 Août 1991. "Molecular Replacement Methods applied to Medium-Size Molecular Structures".

SALUDJIAN, P., PRANGE, R., NAVAZA, J., MENEZ, R., GUILLOTEAU, J.P. & DUCRUIX, A. (1992). Acta Cryst. A**48**. "Structure Determination of a Dimeric Form of Erabutoxin-b, Crystallised from a Thiocyanate Solution". In press.

a, yaap, asap, @#*?
A set of averaging programs.

T.Alwyn Jones,
Department of Molecular Biology,
Biomedical Centre, Box-590,
S-75124 Uppsala, Sweden

Introduction

During the 1960's it was realized that if a macromolecule could be crystallised with more than one copy in the asymmetric unit, then it should be possible to derive useful phasing information from this multiplicity. Most of the early theoretical developments aimed at a reciprocal space formulation of the problem (Rossmann & Blow, 1963, 1964; Crowther, 1967, 1969; Main ,1967). This 'mind-set' was, with hindsight, probably due to the great success achieved by Rossmann & Blow (1962) in their reciprocal space formulation of a search function to locate the orientation of non-crystallographic symmetry (NCS) elements (this was essentially a reciprocal space formulation of a Patterson sum function).

Despite these theoretical developments, the first use of non-crystallographic symmetry averaging was made in real space: on chymotrypsin at 2Å resolution (Matthews *et al.*, 1967) and on haemoglobin at 5Å resolution (Muirhead *et al.*, 1967). The first time that averaging made the difference between solving or not solving a structure occurred a few years later with the work on GAPDH (Buehner *et al.*, 1974); again in real space. Attempts continued, however, on a reciprocal space formulation (Jack, 1973).

Although Main (1967) had mentioned *en passant* that working in reciprocal or real space should be the same, the theoretical basis for a real space formulation and, even more importantly, a practical implementation was derived by Bricogne (1974, 1976). His implementation on the computers of the time was elegant and efficient, and made use of a sorting trick to construct the averaged structure. These programs rapidly led to the structure determination of the first virus structures; the TMV disk protein complex (Bloomer *et al.*, 1978) and the spherical virus TBSV (Harrison *et al.*, 1978). The package was widely distributed and has been used to solve many structures. A similar algorithm was independently derived by Johnson (1978), and used by Rossmann and co-workers in their phenomenal virus work.

The amount of memory available on the average computer increased dramatically during the 1980's. Alternative algorithms have been developed that hold the complete map in core and that do not rely on a sorting process (Nordman, 1980; Smith *et al.*, 1983; Hogle *et al.*, 1986). These algorithms use a more sophisticated interpolation method that allows the use of coarser grid spacings in the electron density maps.

So why did I want to write a new program? In my work on STNV with Lars Liljas, we used Bricogne's computer programs for phase-refinement, first to solve the

structure (60 fold averaging of an asymmetric unit containing 12000 residues) and then to refine it to a resolution of 2.5Å (Jones & Liljas, 1984). Others in our laboratory used the programs to solve two rubisco structures (Schneider *et al.*, 1986; Andersson *et al.*, 1989), the R2 subunit of ribonucleotide reductase (Nordlund *et al.*, 1990), bacteriophage MS2 (Valegård *et al.*, 1990). However (to my embarrassment) when trying to solve the structure of the cellulase CBH2, I wanted to carry out 2-fold averaging of the MIR map and spent a week trying to do it but without success (although I managed to solve the structure anyway, Rouvinen *et al.*, 1990). I never figured out what went wrong but perhaps there was a 'special' space group specific feature somewhere that I didn't know about. Whatever the reason, because more than half of the structures we work on crystallise with NCS, it persuaded me that I should write my own program. As a prime goal, this new program should be easy to use. It should also be coupled to my graphics program, O, not only for interactive functionality and ease of use, but also for error checking.

Overview of the averaging process.

The reader is encouraged to read Bricogne's papers. The problem is illustrated schematically in Figure 1.

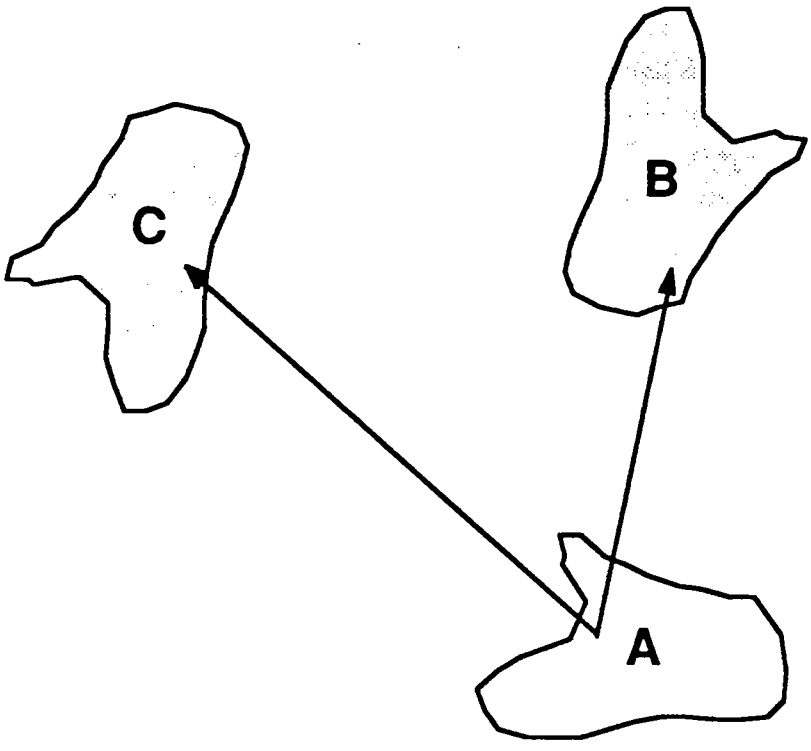


Figure 1. Three NCS related molecules, enclosed by a 2-D mask, with their operators defined relative to the standard molecule A.

Before carrying out averaging, one needs all of the following:

1. An electron density map covering at least the asymmetric unit of the crystal cell.
This map may be obtained from experimental phases or from a molecular replacement solution. In the current program, the map must be made by the CCP4 program package.

2. A mask that covers the non-crystallographic asymmetric unit.
This is illustrated by the bound curve in Figure 1 but, in fact, the mask is a three dimensional matrix of 0's and 1's. All pixel points that lie within the volume defining the NCS a.u. have the value 1, those outside have the value 0. It is important that the mask fits the NCS asymmetric unit as closely as possible, and that there is no overlap between NCS or crystallographic copies of the mask. However, it is also true that too small a mask will result in truncation of density.

The electron density map and the mask must have the same coordinate system and the same separation between grid points. Normally the mask contains many fewer points then the asymmetric unit of electron density. The limits of the mask do not have to be contained within the electron density map.

The mask file is formatted and first contains four lines of header information to define the origin in grid units, the extent, the number of grid units along each cell edge, and the unit cell constants. The rest of the file is a matrix written out with the Fortran statements

```
Write (lun,10) (((mask(i,j,k), i=1,nx), j=1,ny), k=1,nz)
10 format (40i2)
```

where `nx, ny, nz` are the extents along the edges of the mask. Using the Unix `compress` program saves a lot of space on a mask file.

Masks can be produced from a PDB file of coordinates or from an O skeleton. They can then be edited in O using the `Mask` menu of commands. This is almost always necessary since an initial mask will either contain too many holes (if one uses too small an atomic radius), or may be too big and fat (with too large a radius) so that it would overlap with either an NCS or crystallographically related mask.

It frequently happens that one may change the grid spacing of a map during an averaging experiment. Therefore a program exists to modify a mask accordingly.

3. A set of operators that describe the NCS.
These are illustrated by the arrows in Figure 1. Each operator relates the NCS a.u. to another copy of itself. The directionality of the transformation is very important. For A and B in Figure 1,for example, it is the operation needed to bring A (the NCS a.u.) onto B.

The operators are 12 floating point numbers that describe a 3x3 rotation matrix and a translation matrix. The numbers would usually be obtained from O and correspond to the O date block `.lsq_rt_<user_text>` produced by the LSQ menu. For an operator R(12), the transformation is as follows:

new_x = R1 x + R4 y + R7 z + R10
new_y = R2 x + R5 y + R8 z + R11
new_z = R3 x + R6 y + R9 z + R12

where (x,y,z) are coordinates in the orthogonal axis system used in O (and in the PDB file format).

In the stand alone programs, each operator must exist as a separate O source file produced with the O `write` command. Thus, an operator file might look like:

```
vega [2]# cat a_to_b.o
.LSQ_RT_A_TO_B              R          12 (3f10.5)
   0.38837   0.08298   0.91776
   0.11473   0.98383  -0.13750
  -0.91434   0.15869   0.37257
  64.76807 -34.23807  -6.04810
vega [3]#
```

How the operators are first produced depends on the problem at hand. They may originate from comparing heavy atom sites, by building the same part of a molecule in each NCS a.u., or for a molecular replacement solution via the O LSQ menu. A program exists to optimize an approximate NCS operator; it will not work if the operator is unknown.

The averaging program does not differentiate between proper and improper symmetry.

Once these three items are available, the electron density can be averaged.

Programs.

Table 1 lists the stand-alone programs and the O commands that make up the averaging system.

Table 1. List of stand-alone programs and options in O

A	O
bones_mask	bones_mask
pdb_mask	
mask_overlap	mask_read
mask_new_grid	mask_write
	mask_setup
rt_improve	mask_on
average	mask_off
expand	@mask_wire_frame

mappage	*av_read_map*[1]
mappage_mask	*av_write_map*
	av_setup_map
corr_coeff	*av_average*
	av_expand
map_add	*av_rt_improve*
	mask_bones[11]
bones	*mask_pdb*
	mask_overlap
	mask_wire
	mask_fill
	mask_expand
	mask_contract
	mask_trim
	mask_logical
	mask_not

[1] *Someday options in O.*

[11] *Any day now options in O.*

Some of the more important programs and commands will now be described in more detail.

Creating a Mask from Coordinates.

The easiest way to make a mask file is from a set of coordinates. The program `pdb_mask` converts a standard Brookhaven Protein Data Bank file of coordinates into a mask file. The user needs to specify the unit cell constants and an O NCS operator file. The latter allows one to place the mask over a different 'standard' NCS a.u.

Although the program is meant to be run interactively, the following script shows the input

```
vega [2]# cat pdb_mask
/usr/people/alwyn/a/bin/es_pdb_mask << EOF-input
m12a.pdb
3.
91.8 99.5 56.5 90. 90. 90.
100 110 60
unit_rt.o
```

```
EOF-input
vega [3]#
```

The first line is the file of coordinates, the second is a radius to be placed around each atom so that all grid points within this distance are set to be in the mask. Line 3 contains the unit cell constants. Line 4 supplies the number of grid points along each axis. In this example, we have diffraction data to 2.7Å, so we have chosen a spacing of 0.9Å between each grid point. This spacing, one third the resolution, is coarser than that recommended by Bricogne. Line 5 is the name of the O-style file containing the NCS operator, in this case the unit matrix.

The program generates the mask within a parallelopiped containing the atoms. Choosing too low an atomic radius may result in a mask containing holes in the protein interior. Too large a value may cause overlap with NCS or crystal symmetry related molecules. In either case the mask will need editing (see below).

Creating a Mask from Skeleton Atoms.

A skeletonised electron density can be used to create an initial mask. The O command `bones_mask` generates a set of coordinates and radii for subsets of skeleton atoms. Each skeleton atom has a status code associated with it. In O this code is used to decide what skeleton atoms are to be drawn and to decide their colour. The codes are used by the user to decide likely main chain and side chain atoms. During map interpretation the user usually re-assigns main chain atoms according to his/her folding hypothesis. The `bones_mask` command allows one to decide which status codes to use for defining the mask and what radius to associate with each code. If the skeleton has the name `ano1` in O, for example, then a new datablock vector `ano1_mask` is generated in the database. This can be written to a file with the O `write` command and converted to a mask with the `bones_mask` program. This program is very similar to `pdb_mask` but the first line of input defines the name of this file instead of a PDB file.

Converting a Mask into a Map File.

It may sometimes be advantageous to view the mask as a contoured object. The program `mappage_mask` converts a mask file into a binary dataset that can be contoured with the O `Map` commands. The program prompts for the mask file name and then the map file name. Points within the mask are set to 100., outside to 0.; the map should therefore be contoured at a value of 99. This program will be made redundant when the O command `mask_wire` becomes available, Table 1.

Contoured objects generated in O from a map file can be transformed by the NCS operators. This can be a useful check of both the mask and of the electron density files.

Editing an existing Mask.

Once a mask file has been generated, it can be edited in O with some of the `Mask` menu of commands. The use of the existing commands is illustrated by the following

example (user input is high-lighted):

```
O > mask_read
Msk> File name of mask to be read: m12a.mask
Msk>  Reading Mask.
Msk>  Mask read OK.
Msk>  Number of points in mask 33221
O  > mask_set
Msk> Mask value to be set : 1
Msk> Mask value displayed : 0
Msk> Radius in which pixels are set ([1.0]):
Msk> Radius in which pixels are displayed ([10.0]):
O  > mask_on
O  > mask_off
O  > mask_write
Msk> File name of mask to be written: m12a_edit.mask
O >
```

The `mask_read` and `mask_write` commands read and write out mask files. Only one mask file can be stored at a time (at present).

The `mask_setup` command set parameters used with the `mask_on` and `mask_off` commands that control what will be displayed and how the mask is to be edited. In the above example, both `mask_on` and `mask_off` will display non-mask grid points within 10Å of a moving atom as dots. When editing is set 'on' by the `mask_on` command, points 1.0Å from the moving atom will be set as lying within the mask. This will have the affect of erasing the dots drawn on the display. The setup values used here are, therefore, suitable for detecting holes in the mask and for removing them.

The commands `mask_on` and `mask_off` start the display of the mask. Both commands require the user to identify an atom that can be moved around in space. Dots will appear that represent grid points either within or outside the mask depending on the values set with `mask_setup`. If `mask_off` is active, then as the atom is moved around, the mask object is updated but the mask is not changed. When `mask_on` is active, the mask gets changed and grid points within the specified radius are reset, according to the `mask_setup` values. Activating `mask_on`, deactivates `mask_off` and *vice versa*. When either command is active, `mask_setup` can be activated to change the current values.

At present, there is no command to contour a mask that has been read into O. The macro `mask_wire` is supplied and makes a map file that is then contoured. Experience suggests that it is advantageous to display (or have available) both a chicken wire, contour representation and a dot representation of the mask.

More mask commands are planned for this summer (1992). We mention them since they should be available before this manuscript:

`mask_wire` contours the mask.

`mask_fill`	removes enclosed volumes in the mask.
`mask_expand`	expands the mask by a fixed percentage.
`mask_contract`	contracts the mask by a fixed percentage
`mask_trim`	trims the surface layer of the mask, close to a point in space.
`mask_logical`	applies logical operations between masks.
`mask_not`	applies a logical not operation to a mask.
`mask_bones`	replaces the stand-alone program `bones_mask`
`mask_pdb`	replaces the stand alone program `pdb_mask`
`mask_overlap`	replaces the stand alone program `mask_overlap` (mentioned below).

Multiple masks will be allowed.

Changing the Grid Parameters of a Mask.

During the course of a cyclic averaging experiment, it may be desirable to increase the resolution of the diffraction data. This usually requires a change in the sampling of the electron density file produced. Since the mask and the electron density file must be on the same grid, a new mask must be constructed. The program `mask_new_grid` allows one to take an existing mask and create a new mask from it that has different sampling. During the editing of an existing mask, it may be necessary to increase the extent of the mask i.e. increase the number of grid points along one or maybe all three directions. This can also be done with `mask_new_grid`.

If the grid spacing is to be changed, then the transformed index to the new mask pixel will be non-integral. Therefore, the 8 nearest neighbour mask pixels are set on. This will usually result in an expansion of the mask volume. If the user wants to just change the origin or extent, then the transformation is to an integral pixel value and no expansion occurs.

Improving an NCS Operator.

The operators relating NCS units are usually approximations that need improving. This can be done with the program `rt_improve`. As well as needing an approximate operator, the program also requires a mask. At present the program will find an improvement in the translational or the rotational component of the operator (the rotation is about the centre of gravity of the mask). The user must also specify a suitable step size. The program is interactive and either translation or rotation searches can be specified. Table 2 illustrates how it corrected a large deliberate error in the NCS operator relating molecules A and B of P2 myelin protein.

Translation and rotational scans are shown in Figure 2 to illustrate the accuracy that can be achieved.

Table 2. Correcting a deliberate error in an NCS Operator with `rt_improve`
'Rotate' implies an orientation search about the centre of gravity of the transformed mask. 'Translate' implies a translational search. The correlation coefficient is computed using all of the grid points within the mask.

Operation	CorrelationCoefficient
Start	.004
Rotate 2°	.035
Translate 0.4Å	.061
	.065
Translate 0.2Å	.067
Rotate 2°	.098
	.118
	.124
Rotate 1°	.124
Translate 0.4Å	.197
	.247
	.279
Rotate 2°	.402
	.500
	.545
Translate 0.4Å	.643
Translate 0.2Å	.653
Rotate 2°	.665
Rotate 1°	.671

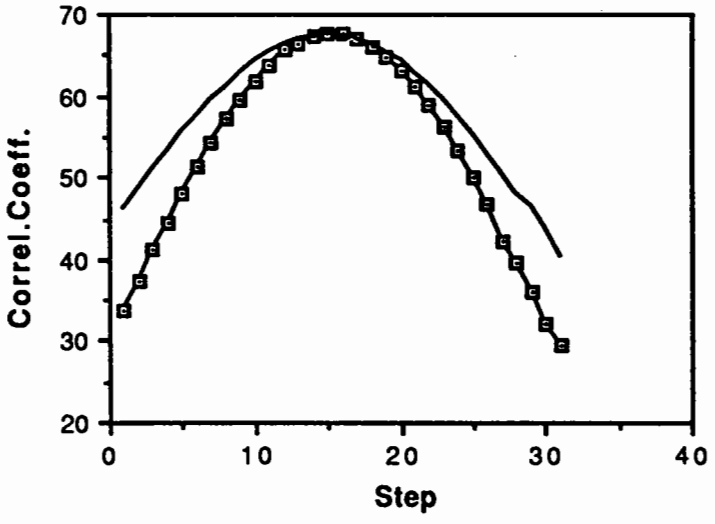


Figure 2. The effect of rotational and translational scans about a correct NCS operator. The translation step, marked by squares, corresponds to shifts of 0.1Å; the rotation to shifts of 0.5°.

It should be noted that for this protein which has a radius of ~17Å, a grid point spacing of 0.9Å subtends an angle of ~3° from the centre of the molecule. The accuracy we obtain, therefore, is better then one third of the grid spacing.

Averaging.

The following script illustrates the use of the program `average`:

```
/usr/people/alwyn/a/bin/es_average <<EOF-average
cyc1.map
m12a_edit.mask
average_1.E
y
symop.o
unit_rt.o
a_to_b.o
a_to_c.o

EOF-average
```

Line 1 defines the input CCP4 map and must correspond to at least the crystallographic asymmetric unit.
Line 2 defines the mask file
Line 3 the name of the output CCP4 style electron density map
Line 4 is an answer to the question that asks whether the averaged map should be expanded (see below).
Line 5 is an O style datablock of the crystallographic symmetry operators.
Lines 6-9 are the NCS operators in O style datablock files. The final blank line terminates input. In this example, there are 3 molecules making up the NCS a.u. as illustrated in Figure 1. Note that the unit operator needs to be described. File `a_to_b.o` contains the optimised operator that moves a point within molecule A to the equivalent position in molecule B. File `a_to_c.o` contains the optimised operator that moves a point within molecule A to the equivalent position in molecule C.

Once the input files have been read, the program consists of 2 parts. The first part produces an averaged map, the second expands this into the volume of the input map, making use of the NCS and crystallographic operators. The averaging algorithm is as follows:

for each pixel in the mask
    zero the averaged electron density array at this pixel P
    for each NCS operator

```
        apply current operator to orthogonal coordinate of pixel P
        convert to a new pixel coordinate
        get electron density at this non-integral pixel position by interpolation
        add to averaged electron density array at pixel P
    end
end
```

The averaged density array is the same size as the mask array, with the same origin, extent, and grid separations. Note that the coordinate of the NCS transformed pixel will not normally correspond to an integral grid point. The value of the electron density at this position must be obtained by interpolation. This program uses a routine written by Wayne Hendrickson and Janet Smith that uses cubic spline interpolation. To date, we have worked with masks and maps where the grid spacing is one third of the resolution of the diffraction data. This is coarser then the value recommended by Bricogne and although I have not made a careful comparison of the effects of changing the grid spacing, one third the resolution seems to suffice.

If one wishes to carry out cyclic phase refinement, it is necessary to re-construct the crystallographic asymmetric unit from the NCS a.u. I call this process expansion and the algorithm is as follows:

```
zero input map
for each grid point in the mask
    for each NCS operator
        transform grid point to NCS non-integral grid point
            for each neighbouring grid Q point to the non-integral grid point
                transform back to the NCS a.u to give non-integral grid point
                interpolate electron density in averaged map
                store in all grid points of input map associated by crystallographic
                symmetry to Q
            end
        end
    end
end
```

The background (non-mask) electron density grid points are set to a value so that the average value in the volume is 0.0 The output map is in CCP4 style.

The complete averaging and reconstruction requires two interpolations, one for the averaging step and one for the expansion.

If the mask occupies too large a volume, it may overlap with another NCS a.u. or a crystallographically related one. This does not result in a build up of density at these points of overlap. The first value to be set is retained. But note that it could be the

wrong one. The program `mask_overlap` can be used to check for mask grid points that overlap.

A number of other programs complete the current system:

`expand` : carries out the second part of the averaging program, expanding the NCS a.u. into a defined volume, normally the asymmetric unit or the complete unit cell. By separating this from the average program, the user can define one set of NCS operators to carry out the averaging, and another set for the expansion

`mask_overlap`: can be used to check for mask grid points that overlap. This produces an expanded CCP4 style map where the density takes on the value 0 (outside the mask), 1 (non-overlapped mask pixel point), 2, 3, 4 ... (overlapped mask pixel points). This map can, of course, be bricked with `mappage` and viewed on the graphics.

`corr_coef` : calculates the correlation coefficent between 2 NCS units. This requires a map, a mask and an NCS operator.

`map_add` : adds or subtracts two maps. The user actually defines some value so that new_map = map1 + value*map2 for each grid point.

`mappage` : creates a bricked electron density file for use with the Map menu in O. This program can read in maps from X-plor, CCP4, FFT of Tenn-Eyck, and PROTEIN (on the Vax, only). For Steigemann's most recent Unix version of PROTEIN one can convert his map to an FFT style map and read that one. This program stores each density point (normally represented as a real number by most program systems) as a single byte, positive integer. A set of scaling values, `prod` and `plus`, can be provided by the user, or can be obtained by the program. They apply the following transformation to the input density:

```
new_value = old_value*prod+ plus
```

If set manually, care must be taken to prevent round off errors and to ensure a large enough dynamic range for contouring. For example, if the standard deviation in an MIR map is 0.15 and one is not interested in displaying negative values, then defining prod=100. and plus=0, allows one to contour at values between 0. and 2.55. Using a prod=1. or 10. would cause truncation problems.

`bones` : skeletonizes an electron density for use with O. One needs to specify a base level at which to start skeletonising and a step constant. Normally we use a step constant equal to the standard deviation of the map. It may be necessary to experiment with the base level, but we usually start at 1.25-1.5 σ. The user also defines the name to be associated with the skeleton when viewed with O.

These programs allow you to do some complicated things, e.g. apply a set of NCS operators within one region, do not in some other region, and apply a different set of NCS operators in a third region. At present, they do not allow one to combine maps from different cells (this will be changed).

Some Examples of Using the Program

P2 Myelin

This structure was solved by isomorphous replacement methods (Jones *et al.*, 1988). It has 3 molecules in the asymmetric unit (but not related by a 3-fold rotation axis), and has been used as the test molecule for the algorithms. We solved the structure with 2 derivatives, but they had identical sites. The data was collected on an area detector and the anomalous derivative measurements were good enough to give an extremely clear map (considering). The structure was traced and built without any averaging. The refinement was made with X-plor to a resolution of 2.7Å to give an R-factor of 16% with a conservative number of added solvent molecules (Cowan, Newcomer & Jones, in preparation). Despite the use of our new tools for the location of peptide errors, use of rotamers etc, I was still somewhat unhappy with the final models, especially of the bound, internal ligand and its interaction with protein residues. At the same time I carried out an averaging experiment on the experimental map.

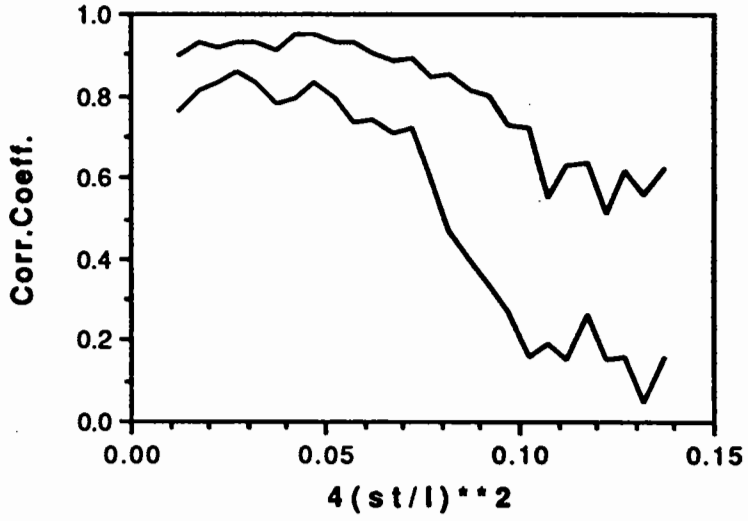


Figure 3. Correlation coefficients for P2 Myelin maps. The bottom curve is for the first averaged map; the top curve for the map after 5 cycles of phase refinement by 3-fold averaging.

Starting from the coordinate set M12 (the final X-plor model), a mask was generated with a grid spacing of 0.9Å and then edited with O to remove internal holes,

and to smooth out some of the outer surface produced by the presence of water molecules in the model. The NCS operators were obtained by least squares comparison of the Cα coordinates of the three chains. The correlation coefficient relating F_{obs} and F_{calc} obtained by Fourier transforming the first averaged map was .70 and is shown as a function of resolution in Figure 3. The R-factor for the averaged MIR map was 36.8%. Five cycles of phase refinement were carried out, using $2|F_{obs}|-|F_{calc}|$ amplitudes and calculated phases for the Fourier; we did not phase combine with the experimental phases. The correlation coefficient increased to .80 and the R-factor decreased to 23.8%. The resulting averaged map was superb, much more clearly showing the ligand and even water molecules inside the β-barrel. The model was re-built, all waters removed except the internal waters that obey the NCS symmetry and this model refined with tight NCS restraints using our local version of Hendrickson's Prolsq program (Kleywegt & Jones, maybe published sometime). This has an R-factor of 26%.

Rubsico from Synechococcus PCC6301

This work has been carried out by Janet Newman. In this structure the whole L8S8 rubisco molecule occurs in the crystallographic asymmetric unit. The structure was solved by molecular replacement. Because of the size of the asymmetric unit, an initial refinement to 2.2Å resolution was unsatisfactory and produced a rather distorted model. This was used for phase calculation to produce a map, to produce a mask which was edited with O, and to produce the NCS operators. Three cycles of phase refinement with eight fold averaging produced a spectacular map, clearly showing many water molecules and raised the correlation coefficient from 0.78 to 0.93. After rebuilding and the addition of solvent, the refinement has been continued, maintaining strict NCSymmetry to give an R-factor of 23%.

The Complex of Fc-Protein G.

We have recently solved the structure of the immunoglobulin Fc fragment in complex with the C2 fragment of protein G at 3.2Å resolution (Eriksson, Uhlen & Jones, to be published). One cycle of averaging with phases calculated from the Fc search fragment produced a map in which the protein G could be completely traced. Protein G has been solved by 2D-NMR (Gronenborn *et al.*, 1991) and their model could be placed in the density as a rigid body.

Up to now I have not spent a lot of time optimizing the speed of the program. Lars Liljas and his co-workers are commencing the high resolution refinement of the bacteriophage MS2 and so I will probably look for some improvements in efficiency in the coming months.

References.

Andersson,I., Knight,S., Schneider,G., Linqvist,Y., Brändén,C.-I. & Lorimer,G.H. (1989) Nature 337, 229.

Bloomer,A.C., Champness,J.N., Bricogne,G., Staden,R., & Klug,A. (1978) Nature 276, 362.

Bricogne,G. (1974) Acta Cryst. A30, 395.

Bricogne,G. (1976) Acta Cryst. A32, 832.

Buehner,M., Ford,G.C., Moras,D., Olsen,K.W., & Rossmann,M.G. (1974) J. Mol. Biol. 82, 563.

Crowther,R.A. (1967) Acta Cryst. 22, 758.

Crowther,R.A. (1969) Acta Cryst. B25, 2571.

Gronenborn,A.M., Filpula,D.R., Essig,N.Z., Achari,A., Whitlow,M., Wingfield,P.T., Clore,G.M. (1991) Science 253, 657.

Harrison,S.C., Olson,A.J., Schutt,C.E., Winkler,F.K. & Bricogne,G. (1978) Nature 276, 368.

Hogle,J.M., Chow,M. & Filman,D.J. (1986) in Crystallography in Molecular Biology. Plenum Publishing Corp. New York, pp 281.

Jack,A. (1973) Acta Cryst. A29, 545.

Johnson,J.E. (1978) Acta Cryst. B34, 576.

Jones,T.A. & Liljas,L. (1984) J. Mol. Biol. 177, 735.

Jones,T.A., Bergfors,T., Unge,T. & Sedzik,J. (1988) EMBO J., 7, 1597

Main, P. (1967) Acta Cryst. 23, 50.

Matthews,B.W., Sigler,P.B., Henderson,R. &Blow,D (1967) Nature 214, 652.

Muirhead,H., Cox,J.M., Mazzarella,C., & Perutz,M. (1967) J.Mol. Biol. 28, 117.

Nordlund,P., Sjöberg,B.M. & Eklund,H. (1990) Nature 345, 593.

Nordman,C.E. (1980) Acta Cryst. A36, 747.

Rossmann,M.G.& Blow,D. (1962) Acta Cryst. 15, 24.

Rossmann,M.G.& Blow,D. (1963) Acta Cryst. 16, 39.

Rossmann,M.G & Blow,D. (1964) Acta Cryst. 17, 1474.

Rouvinen,J., Bergfors,T., Teeri,T., Knowles,J. & Jones,T.A. (1990) Science 249, 380.

Schneider,G., Lindqvist,Y., Brändén,C.-I., & Lorimer,G.H. (1986) EMBO J., 5, 3409.

SMith,J.L., Hendrickson,W.A. & Addison,A.W. (1983) Nature 303, 86.

Valegård,K., Liljas,L., Fridborg,K. & Unge,T. (1990) Nature 345, 36.

Experiences with Molecular Replacement in the Case of Antithrombin III: the Combination of Different Starting Models and Exploration of the Power of a "Cross Translation Function"

Herman A. Schreuder, Bijtske de Boer, Titia K. Sixma,
Alexei V. Tepliakov*, Angel Aguirre and Wim G.J. Hol
BIOSON Research Institute, University of Groningen
Nijenborgh 4, 9747AG Groningen, the Netherlands
*European Molecular Biology Laboratory, Hamburg Outstation
Notkestrasse 85, 2000 Hamburg 52, G.D.R.

Introduction

Antithrombin III, further referred to as antithrombin, is a key anti-coagulant protein in human blood. It inhibits serine proteases of the blood coagulation cascade such as factor IXa, Xa and XIa, and thrombin. Antithrombin is activated by heparin; the association constant between antithrombin and thrombin increases 1000-fold in the presence of heparin. Knowledge of the three-dimensional structure of antithrombin and of the antithrombin-heparin complex would assist the design of chemical analogs of heparin, which could potentially be useful as antithrombotic drugs (Mourey et al., 1990).

Antithrombin is a member of the serpin (serine proteinase inhibitor) superfamily of proteins. Serpins have long and flexible reactive site loops, in contrast to the short and fixed loops present in other families of proteinase inhibitors. The first serpin crystal structure to be reported was the structure of proteolytically nicked α_1-proteinase inhibitor (also called antitrypsin) by Loebermann et al. (1984). The most striking feature of this structure was that Met358 (the P1 residue) and Ser359 (the P1′ residue) are separated by 67 Å, and that they are located at opposite sides of the molecule. These residues are covalently linked before proteolytic cleavage. The crystal structure of ovalbumin, an intact, non-inhibitory member of the serpin superfamily by Stein et al. (1990), revealed that an unprecedented conformational change occurs in serpins upon proteolytic cleavage. The "reactive site" loop in intact ovalbumin is exposed and adopts an α-helical conformation, while the reactive site loop in cleaved α_1-proteinase inhibitor is a ß-strand, and is part of the central ß-sheet of the protein. These conformational differences are shown in Figure 1.

Crystals of human antithrombin

Human antithrombin crystallizes in spacegroup $P2_1$. The cell dimensions and ß-angle vary somewhat between the different crystals. The celldimensions initially determined with an Enraf-Nonius FAST area detector were: a=89.8 Å, b=100.8 Å, c=70.0 Å and ß=106°. These cell dimensions were used in the initial studies. The asymmetric unit contains two molecules of 58 kD each. V_M is 2.63 $Å^3$/dalton and the crystals contain 53% solvent. The crystals are small (0.1*0.15*0.2 mm^3) and diffract to about 3.5 Å resolution. The dataset used for molecular replacement was collected on film at the Daresbury synchrotron.

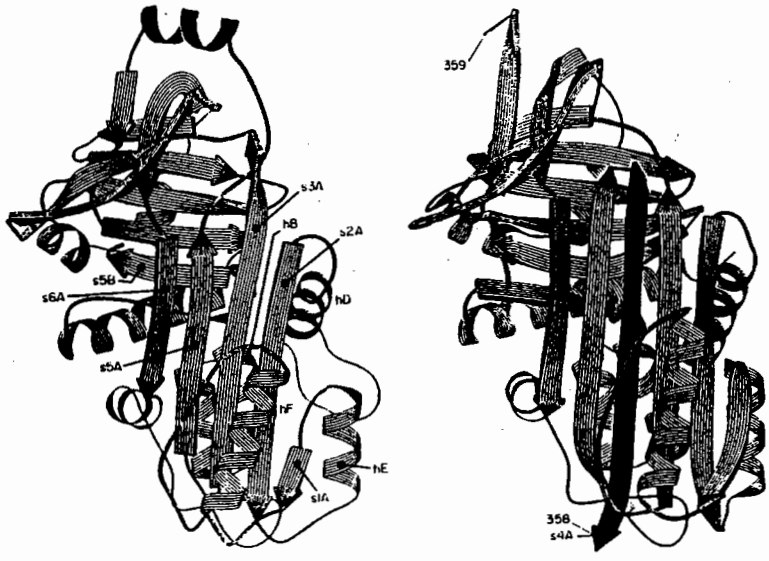


Figure 1:
Ribbon plots of left: ovalbumin and right: α_1-proteinase inhibitor. The reactive site loop (indicated in black) is exposed and has an α-helical conformation in ovalbumin, while it is part of the central ß-sheet in α_1-proteinase inhibitor. (Taken with permission from Stein and Chothia, (1991).

Analysis of dissolved crystals showed that the crystals contain biologically active and therefore intact antithrombin.

Solution of the molecular replacement problem

The structure of antithrombin crystals was solved by molecular replacement using two different models: (i) the structure of intact ovalbumin as determined at 1.95 Å resolution by Stein et al. (1990) and (ii) the structure of cleaved α_1-proteinase inhibitor as solved at 3.0 Å resolution by Loeberman et al. (1984). Ovalbumin shares 31% sequence identity with antithrombin, and α_1-proteinase inhibitor shares 28% sequence identity with antithrombin.

The Crowther (1972) fast rotation function failed to give a clear solution for the ovalbumin model, in spite of varying numerous parameters, as discussed e.g. by Schierbeek et al. (1985). Rotation functions in combination with Patterson correlation (PC) refinement (Brünger, 1990; 1991), however, did give clear solutions. The strategy to position two molecules in the asymmetric unit was as follows:

1. Find the orientation of molecule 1
2. Find the position of molecule 1 in the XZ plane
   (the spacegroup of human antithrombin crystals is polar, hence the Y coordinate can by choosen arbitrarily)
3. Find the orientation of molecule 2
4. Find the position of molecule 2 in the XZ plane
5. Find the relative Y position of molecule 2 with respect to molecule 1

The ovalbumin model gave a clear solution with a Patterson correlation coefficient of 0.077. The subsequent translation search in the XZ plane gave a clear maximum correlation coefficient of 0.113, which is 4.2σ above the mean. However, no signal was detected for the second molecule in the asymmetric unit.

A rotation search with the α_1-proteinase model gave, again, a clear single solution with an orientation different from the orientation of the ovalbumin model, and a maximum patterson correlation coefficient of 0.111. The subsequent translation search resulted in a maximum correlation coefficient of 0.122, which is 5.2σ above mean. Finally, the translation search to find the relative Y position of molecule 2 with respect to molecule 1 gave a maximum correlation coefficient of 0.184. Subsequently, rigid body refinement with each molecule divided into two parts (N-terminus to P1 residue and P1′ residue to C-terminus) yielded an R-factor of 52.3% (48.6% using FAST data, collected at a later stage, using better crystals; see Table 1). These R-factors are high, but the signals were clear and a translation search using the BRUTE program (Fujinaga and Read, 1987) gave the same results.

The molecular replacement study showed that the crystals contain two types of molecules: type 1 is similar to intact ovalbumin, and type 2 is similar to cleaved α_1-proteinase inhibitor. Subsequent analysis of dissolved crystals by gelelectrophoresis showed two closely spaced bands of equal intensity. The surprising conclusion of these observations is that the crystals contain one intact antithrombin molecule and one cleaved molecule, even though the initial crystallization solution only contained intact molecules. The difference in conformation between the two molecules is apparently so large that two different models were neccessary to find them.

Could the structure have been solved using only one search model?

Since the solution is known, it is possible to analyze in hindsight whether the two different molecules in the asymmetric unit could have been found using a single search model. Also, during the course of the project the quality of the crystals and the data gradually increased, allowing us to test the influence of the quality and completeness of the data on the molecular replacement process. Statistics about the quality and completeness of the three available datasets are listed in Table 1.

Table 1: Statistics of the three data used in the test.

Detector	Location	number of crystals	resolution	R-sym	completeness
Film	Daresbury	11	4.0 Å	19.7%	75.6%
FAST	Groningen	1	3.5 Å	10.1%	61.0%
Image Plate	Hamburg	1	3.5 Å	7.4%	94.9%

Reasons for the poorer statistics of the film dataset could be that the crystals used were smaller than the crystals used to

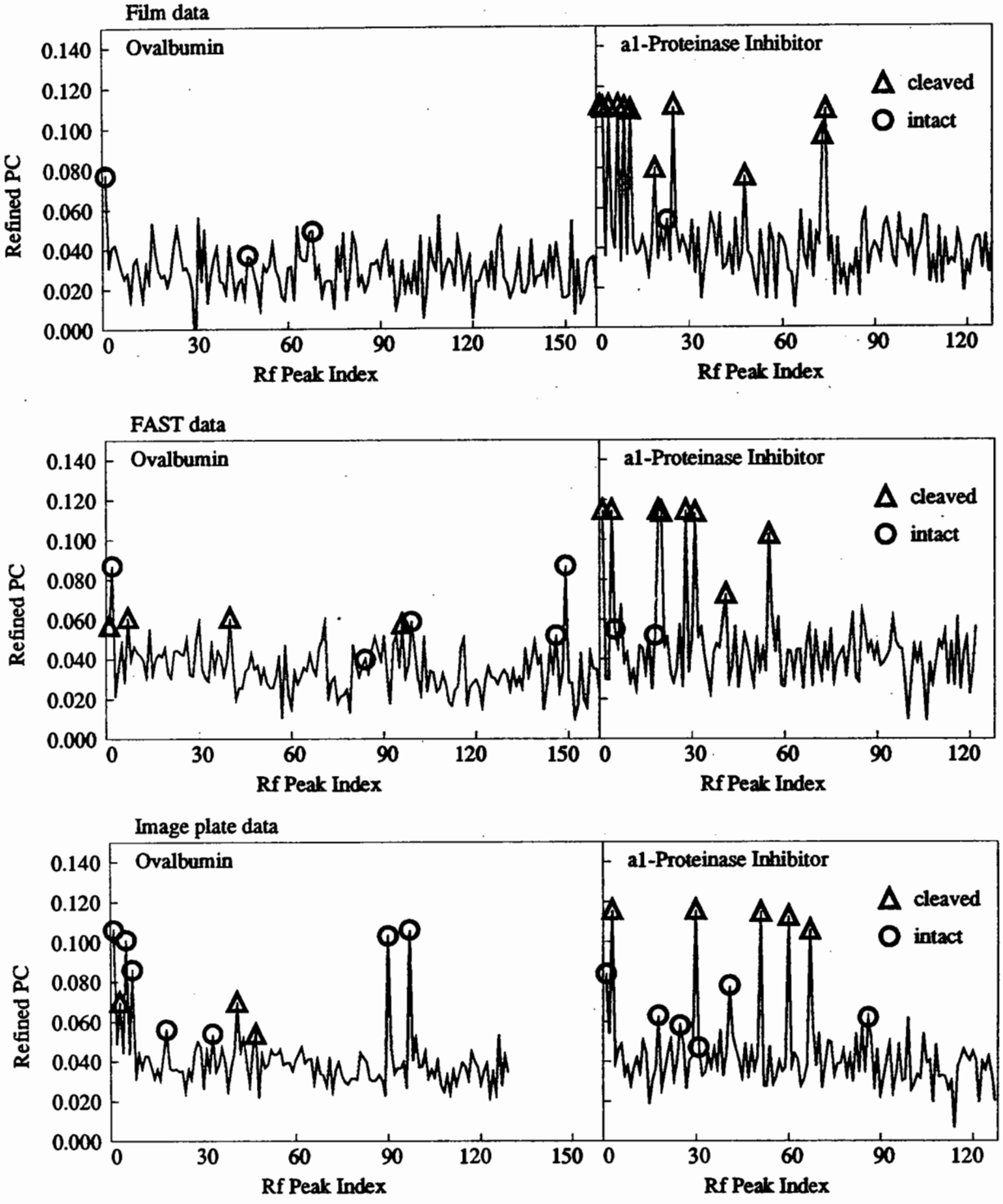


Figure 2:
Patterson correlation coefficients of the 100-150 highest rotation function peaks after PC-refinement. Left panels: using the ovalbumin model; right panels: using the α_1-proteinase inhibitor model. From top to bottom: calculations with the film dataset, the FAST dataset and the image plate dataset, respectively. Circles mark solutions differing less than 5° from the orientation of the intact antithrombin molecule, triangles mark solutions differing less than 5° from the orientation of the cleaved antithrombin molecule.

collect the other two datasets, and that data from 11 crystals with possibly slightly different cell dimensions were merged.

The effect of the different models and datasets was used to evaluate the power of the PC-refinement procedure. To get an idea of the signal to noise ratio's, we will show the results of PC-refinement of the highest 100-150 solutions of the rotation function. But before showing it we will shortly explain the protocol which is as follows: First a rotation function is calculated, using the real space patterson search method of Huber (1985), as implemented in X-PLOR. Peaks in the rotation function differing less then 5°-10° (depending on input parameters), are clustered and the 100-150 highest peaks are written to a list file. Each solution of this list file is subsequently subjected to PC-refinement (Brünger, 1990, 1991). Often different peaks of the list file will converge to the same final solution after this PC refinement.

The procedure described above was done with the models of ovalbumin and the α_1-proteinase inhibitor for all three datasets. The results are shown in Figure 2. For all three datasets, the intact ovalbumin model gave a clear solution for the intact antithrombin molecule, and the model of cleaved α_1-proteinase inhitor gave a clear solution for the cleaved antithrombin molecule. However, the "cross solution" (finding the cleaved antithrombin molecule with an intact model and vice versa) is not always clear. For the film dataset and the FAST dataset, the "cross solution" is only as high as many other noise peaks and it is doubtful if we would have been able to find the second molecule with these data sets. However, the "cross solution" stands out above the noise using the better and much more complete image plate data. We like to emphasize that these results can by no means be used as an evaluation of different synchrotrons or detector systems, since the quality of the crystals was very different in the three cases: very poor for the Daresbury film dataset, poor for the FAST dataset and better, but still poor for the Hamburg image plate dataset.

The effect of the completeness of the data was analyzed by deleting from the image plate dataset all reflections not present in the FAST dataset and repeating the procedure. The result, as indicated in Figure 2a, is that the "cross solution" is clear for the α_1-proteinase inhibitor model, but not for the ovalbumin model. As a last test, we checked whether the signal could be improved by dividing the search model for PC-refinement into six fragments instead of a single rigid body. The purpose of this test was to see if this six-fragment PC-refinement would be able to induce changes in the model, similar to the changes occuring in serpins upon proteolytic cleavage. The result was somewhat disappointing, in that the Patterson correlation coefficient of the correct solution improved by 0.02, but that the correlation coefficient of the noise peaks increased by 0.03, actually decreasing the signal-to-noise level. The six-fragment approach apparently allows, in this case, the PC-refinement to better "optimize" wrong solutions. Translation searches were always successful once the correct orientation was found (data not shown).

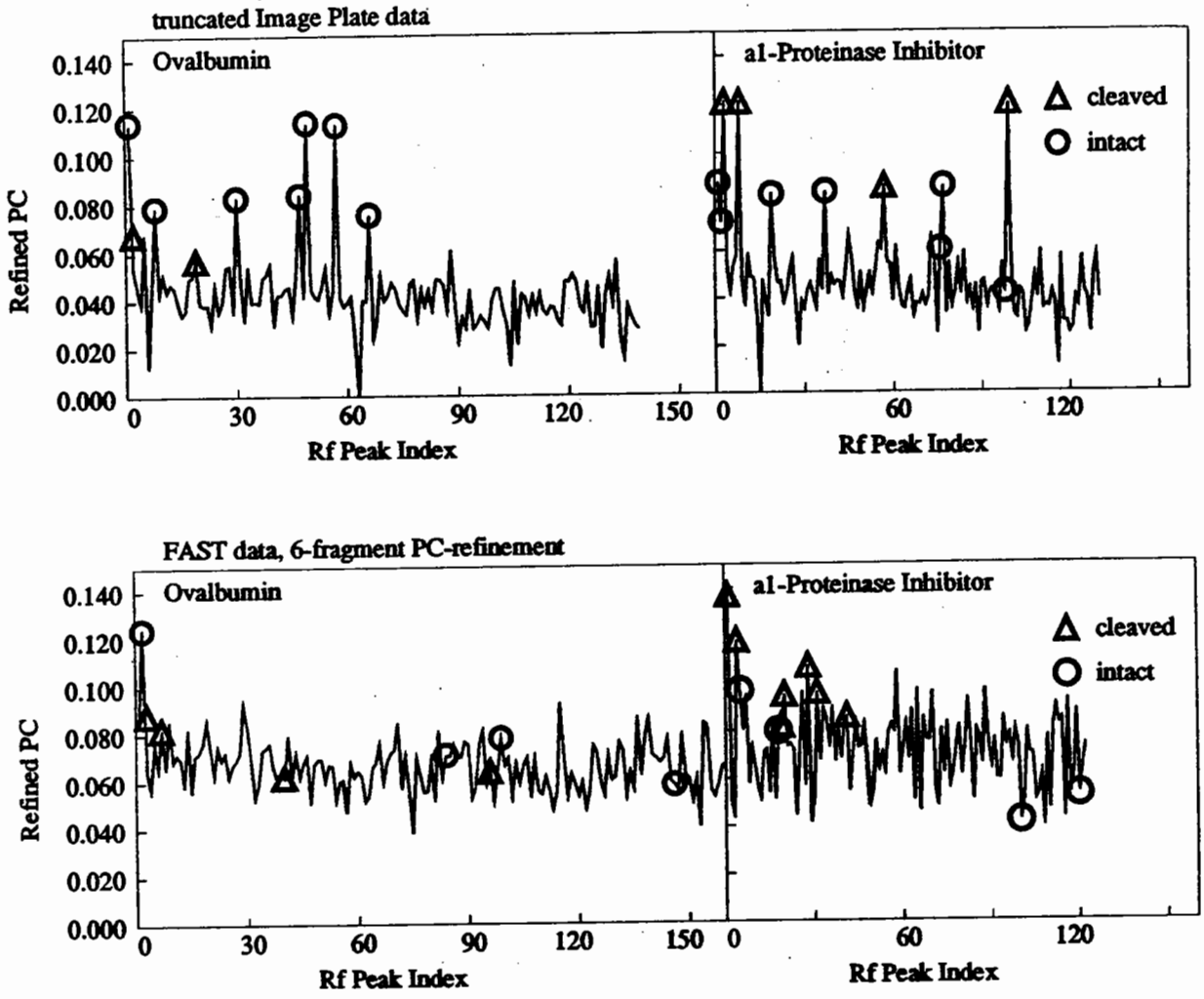


Figure 3:
Patterson correlation coefficients of the 100-150 highest rotation function peaks after PC-refinement. Left panels: using the ovalbumin model; right panels: using the α_1-proteinase inhibitor model. Top calculations with the image plate dataset with all reflection not present in the FAST dataset deleted; bottom: calculations with the fast dataset and PC-refinement using six fragments instead of single rigid bodies. Circles mark solutions differing less than 5° from the orientation of the intact antithrombin molecule, triangles mark solutions differing less than 5° from the orientation of the cleaved antithrombin molecule.

Conclusions

Our crystals of human antithrombin contain one intact, and one cleaved molecule. To find two molecules of the same protein with different conformations, it was neccessary to use two different search models. Identifying the correct peak of the rotation function was problematic if a cleaved model was used to find an uncleaved molecule and vice versa. However, the correct solution was always there. It is apparently possible to obtain the correct molecular replacement solutions, by either using accurate models, and then quite incomplete datasets are allowed, or by using accurate and complete data, and then a more deviating model is acceptable.

Acknowledgements

We thank Dr. Rein Dijkema, Dr. John Mulders and Dr. Henri Theunissen for providing us with protein. We also like to thank Dr. Colin Nave at the Daresbury synchrotron and Cor Kalk in Groningen for help with data collection. This research has been sponsored by Organon International BV.

References

Brünger, A. T. Acta Cryst. *A46* (1990) 46

Brünger A. T. Acta Cryst. *A47* (1991) 195

Crowther, R. A. The molecular replacement method (Rossmann, M. G., ed.), pp. 713 (1972), Gordon & Breech, New York.

Fujinaga, M. and Read, R. J. J. Appl. Cryst. *20* (1987) 517

Huber, R. Molecular Replacement, Proceedings of the Daresbury Study Weekend, pp. 58, Daresbury Laboratory, Daresbury, U.K.

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

Mourey, L., Samama, J. P., Delarue, M., Choay, J., Lormeau, J. C., Petitou, M. and Moras, D. Biochimie *72* (1990) 599

Schierbeek, A. J., Renetseder, R., Dijkstra, B. W. and Hol, W. G. J. "Molecular replacement, Proceedings of the Daresbury Study weekend" (1985) 16

Stein, P.E. and Chothia, C. J. Mol. Biol. *221* (1991) 615

Stein, P.E., Leslie, A.G.W., Finch, J.T., Turnell, W.G., McLaughlin, P.J. and Carell, R.W. Nature (London) *347* (1990) 99

Appendix: A "Cross Translation Function"

During the molecular replacement search, it occurred to us that once the position of the first molecule is known, the position of the second molecule can be found in one single and elegant step, using the following, slightly modified Crowther and Blow (1967) coefficients:

$$T_{cross}(x2) = \sum_{h} I_{obs}\ F_{M1}\ F^{*}_{M2}\ \exp[-2\pi i(h.x2)]$$

M1: oriented and positioned molecule 1, including molecules, related by crystallographic symmetry. This is the "known Molecule"

M2: oriented, but not positioned molecule 2, which is not related by crystallographic symmetry to molecule 1. Indeed, it can be a molecule which bears no relationship whatsoever to molecule 1. Symmetry related partners of molecule 2 are not included in the calculation of structure factors. This is the "unknown" molecule.

X2 at the maximum of T_{cross} directly gives the translation vector to be applied to M2. T_{cross} can obviously be made more powerful by removal of self vectors by subtracting the proper terms from I_{obs} as pointed out by Crowther and Blow (1967). Moreover, we would like to point out that T_{cross} was, prior to our derivation, independently arrived at by Driessen et al. (1991).

As a first test, we used the oriented and positioned ovalbumin model as "known" molecule, and the oriented, but not positioned α_1-proteinase inhibitor model as "unknown" molecule and calculated the cross translation function on an 1.0 Å grid. The correct translation vector appeared as a peak of 6.9σ above mean, while the highest noise peak was only 5.3σ above mean.

As a second test, we used the molecular replacement search of $P2_12_12_1$ crystals of heat labile enterotoxin from *E. coli*. Enterotoxins are important agents in diarrhoeal diseases in man. The structure of heat labile enterotoxin, which has more than 80% sequence identity with cholera toxin, has been solved by MIR methods (Sixma et al., 1991). The space group of the crystals used was $P2_12_12_1$. The toxin consists of a complex between one A subunit (M_r~27k) and five B subunits (M_r~11.6k each) which form a tight pentamer. The C-terminal 20 residues of the A subunit are closely interacting with the B_5 unit, while the the remaining part of the A subunit has few contacts with the five B subunits and can change its relative orientations with respect to this pentamer. (Sixma et al., 1992a; 1992b).

Besides the $P2_12_12_1$ crystals, also crystals with spacegroup $P2_1$ were obtained. These crystals have two AB_5 complexes in the asymmetric unit and the structure was solved at 3.5 Å resolution by molecular replacement with X-PLOR (Brünger, 1990; 1991), using the structure of the $P2_12_12_1$ crystals as a model (Sixma et al., in preparation).

Table 1: Effect of leaving out parts of the model

a: applying $P2_1$ symmetry to the "known" model

"known" model	"unknown" model	correct peak(σ)	highest error peak (σ)
AB_5-I	AB_5-II	20.9	6.6
AB_5-I	B_5-II	15.4	5.6
AB_5-I	A-II	11.6	5.4
B_5-I	B_5-II	14.6	6.5
B_5-I	A-II	10.4	5.5
A-I	A-II	9.2	7.6
B_5-I+B_5-II	A-I	13.9	5.8
B_5-I+B_5-II	A-II	13.1	5.3
AB_5-I+B_5-II	A-II	13.5	5.6

b: without applying $P2_1$ symmetry to the "known" molecule

"known" model	"unknown" model	correct peak(σ)	highest error peak (σ)
AB_5-I	AB_5-II	16.1	6.1
AB_5-I	B_5-II	14.3	7.2
AB_5-I	A-II	9.9	6.6
B_5-I	B_5-II	12.4	7.0
B_5-I	A-II	9.5	7.3
A-I	A-II	absent	8.7
B_5-I+B_5-II	A-I	10.8	5.7
B_5-I+B_5-II	A-II	10.3	5.3
AB_5-I+B_5-II	AB_5-II	11.8	5.6

AB_5-I: first AB_5 complex in the asymmetric unit
AB_5-II: second AB_5 complex in the asymmetric unit

The two oriented and positioned AB_5 complexes in the $P2_1$ crystals were used for more elaborate tests of the "cross translation function", to see how much of the "known" and the "unknown" model can be left out and still produce a clear signal, and also to test how large a misorientation can be tolerated before the signal deteriorates.

Effect of leaving out parts of the model

The "cross translation function" was calculated several times with the A subunit or the B_5 unit left out of the "known" and/or the "unknown" model. The same runs were done without applying crystallographic ($P2_1$) symmetry to the "known" molecule. The results are summarized in Table 1.

In all but one case (A-I versus A-II without applying crystallographic symmetry), the correct peak is the highest peak in the cross-translation function. This means that a clear signal can be obtained by using models that represent only a small portion of the unit cell contents. E.g. for A-I versus A-II by applying $P2_1$ symmetry to the "known" molecule, the "known" model represents only 16% of the asymmetric unit and of the total unit

cell contents, and the "unknown" model represents 16% of the asymmetric unit and only 8% of the total unit cell.

Effect of misorientation

The correctly oriented AB_5-I model was used as "known" molecule, and the oriented AB_5-II model was rotated around an arbitrary axis from 0° to 10° in steps of 1°. For each step, the cross-translation function was calculated. The results indicated that a misorientation as large as 8° can be tolerated before the highest error peak becomes larger than the correct peak (data not shown).

References

Brünger, A. T. Acta Cryst. *A46* (1990) 46

Brünger A. T. Acta Cryst. *A47* (1991) 195

Crowther, R. A. and Blow, D. M. Acta Cryst. *23* (1967) 544

Driessen, H. P. C., Bax, B., Slingsby, C., Lindley, P. F., Mahadevan, D., Moss, D. S. and Tickle, I. Acta Cryst. B*47* (1991) 987

Sixma, T. K., Pronk, S. E., Kalk, K. H., Wartna, E. S., van Zanten, B. A. M., Witholt, B. and Hol, W. G. J. Nature (London) *351* (1991) 371

Sixma, T. K., Terwisscha van Scheltinga, A. C., Kalk, K. H., Zhou, K., Wartna, E. S. and Hol, W. G. J. FEBS Lett. *297* (1992) 179.

Sixma, T. K., Pronk, S. E., Kalk, K. H., van Zanten, B. A. M., Berghuis, A. M. and Hol, W. G. J. Nature (London) *352* (1992) 561

Molecular Replacement Studies of α-momorcharin

Jingshan Ren*, Yaoping Wang†, Yicheng Dong† and David I Stuart*.

* Laboratory of Molecular Biophysics, Rex Richards Building, South Parks Road, Oxford. OX1 3QU.

† Institute of Biophysics, Academia Sinica, Beijing, 100101, China

1. Introduction

Ribosome-inactivating proteins (RIPs) are a large family of proteins that have been classified into two types[1]. Type I RIPs are single chained, such as trichosanthin (TCS) from *Trichosanthes kirilowii*. Type II RIPs consist of two chains, such as ricin from *Ricinus communis*. The A chain has RIP activities and the B chain is a galactose-specific lectin responsible for binding the whole molecule to the target cell. Both types of plant RIPs inhibit protein synthesis of eukaryotic ribosomes by hydrolytically cleaving the N-glycosidic bond of a specific adenine residue in a highly conserved, single-stranded loop of 28S rRNA. Type I RIPs are homologous with the A chain of type II proteins, containing several invariant and highly conserved residues. The crystal structures of TCS and ricin have been reported[2,3]. TCS and the A chain of ricin (RCA) have very similar three dimensional structures.

α-momorcharin (αMMC) is a type I RIP, extracted from the seeds of *Momordica charantia*. The amino acid sequence deduced from cDNA contains 263 residues, of which 63% are identical with TCS and 34% with RCA[4]. The crystals of αMMC belong to space group R3 with cell dimensions a=b=131.3Å , c=40.2Å in the hexagonal system. There are nine molecules in the unit cell; one per asymmetric unit. The solvent content is about 53% by volume. A set of data, essentially complete to a d_{min} of 2.0 Å , was collected using 3 crystals on a Xentronics Area Detector with a very high mean redundency of 7.2. Two models, (i) the full set of TCS atomic coordinates and (ii) the C_α coordinates of RCA have been used for the molecular replacement studies of αMMC discussed in this paper. Both of them are preliminary unrefined structures. There are 234 residues in TCS model and 267 C_α atoms in RCA model. 212 C_α positions can be matched between the two models with an r.m.s deviation of 1.53 Å .

2. Determination of molecular orientation

2.1 Cross rotation

The TCS structure was chosen as a search model for the molecular replacement calculation. The TCS molecule is roughly wedge-shaped, about 55 Å long, 40 Å wide and 25-35 Å thick. The cross-rotation search was carried out using the real-space Patterson search method as employed in the programe X-PLOR[5] using

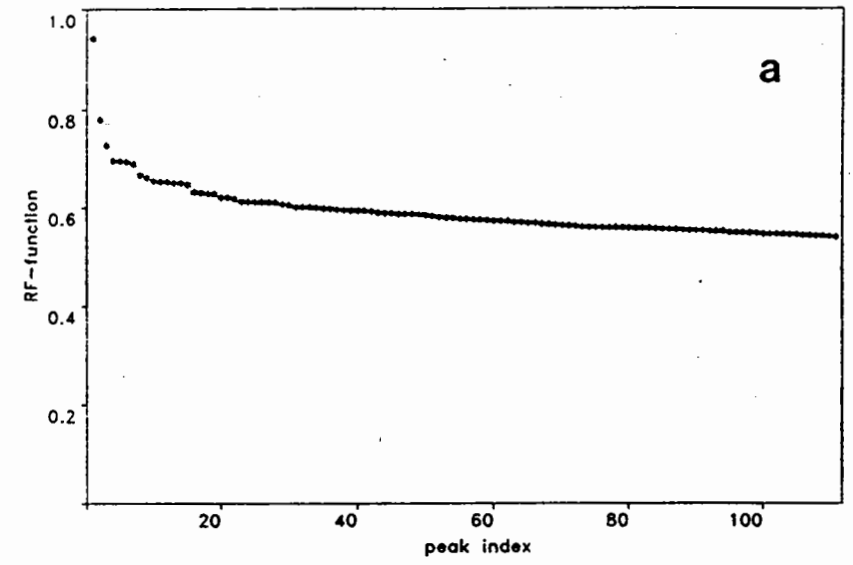
data between d spacings of 15.0 and 4.0 Å . The search-model Patterson maps were calculated by placing the search model into an orthogonal cell of P1 symmetry with a=b=c=120 Å , slightly larger than twice the maximum dimension of the model. An overall temperature factor of 20 Å ^{2} was used. A fine grid size of 0.4 Å was set for evaluating the structure factors and FFT of the squared amplitudes. All the positive model Patterson vectors with length between 5-24 Å were selected for calculation. A Lattmann grid[10] with limits θ_+ =0-360°, θ_2 =0-180° and θ_- =0-120° was searched with a step size of $\Delta\theta_2$ =2.5°. The 6000 highest peaks of rotation function (RF) were analyzed. These were then collected into clusters of width approximately 10° and the 111 resultant peaks were input into PC refinement[5]. The height of the selected RF peaks is plotted against their index in Fig.1a. The highest RF peak, with orientation of α =54.2°, β =27.5° and γ =318.9°, is 6.4σ above the mean and 2.4σ above the highest pseudo-peak.

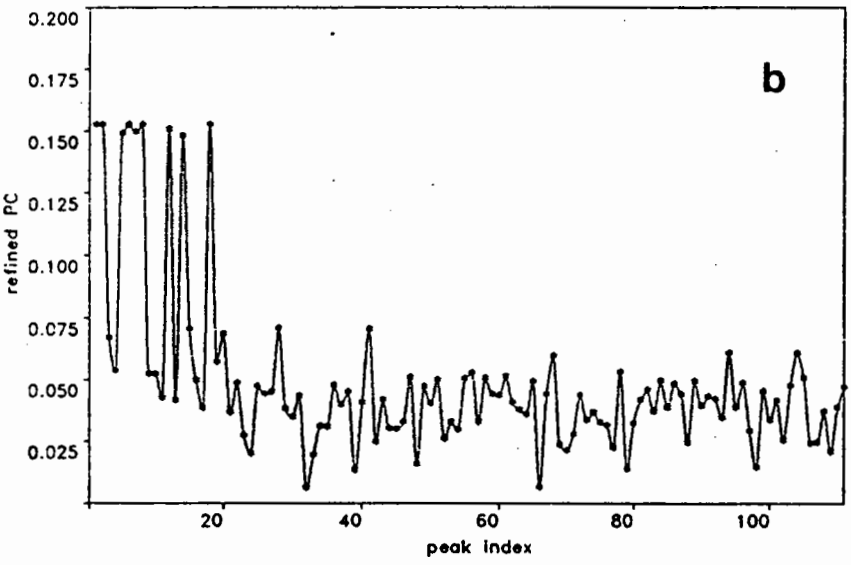
2.2 PC refinement

The orientation parameters of the selected RF peaks were refined using the rigid-body PC refinement. 20 steps were performed for each peak using the data from 15.0 to 4.0 Å . After the refinement nine of the 111 selected peaks converged to the same orientation as the first peak with a maximum value for the correlation coefficent of 15.2% (Fig.1b). The orientation of the first peak remained essentially unchanged, indicating that the solution of the rotation function was rather accurate. The most significant improvement occured at peak 14 for which the three Eulerian angles changed by $\Delta\alpha$ = 23.2°, $\Delta\beta$ = 3.9° and $\Delta\gamma$ = −26.4° to the correct orientation. The correlation coefficient of this peak increased from 4.8% to 14.8%. In general the rotation function result could be inaccurate for a variety of reasons, such as the size of the search step, the truncation of Patterson vectors etc. Obviously, any errors will be carried into the determination of translation parameters, and may prevent the detection of the correct solution. Owing a large radius of convergence, PC refinement is a poweful and very effective method for the correction of rotation parameters.

2.3 The effect of the Patterson vector cutoff

If a molecule is markedly non-spherical then any integration radius chosen for a reciprocal space rotation function is likely to either exclude large number of intramolecular vectors or include a significant number of intermolecular vectors. Thus the rotation function might be very sensitive to the integration radius. Usually several runs with different radius are needed to get correct solution. For αMMC , using the real-space rotation function of X-PLOR, the correct orientation always appears to be the first peak over a wide range of outer vector cut-offs from 12 to 36 Å , changing the length of the selected Patterson vectors only alters the height of the true peak relative to the spurious ones. At 12 Å , the true peak is 3.4σ above the mean of the RF peaks and 0.37σ above the highest pseudo-peak. The best solution is obtained at 32 Å , the true peak is 7.5σ above the mean and 2.9σ above the highest pseudo-peak. However the searches at 36 Å and 40 Å were done using the lower resolution data shell of 15.0-6.0 Å in





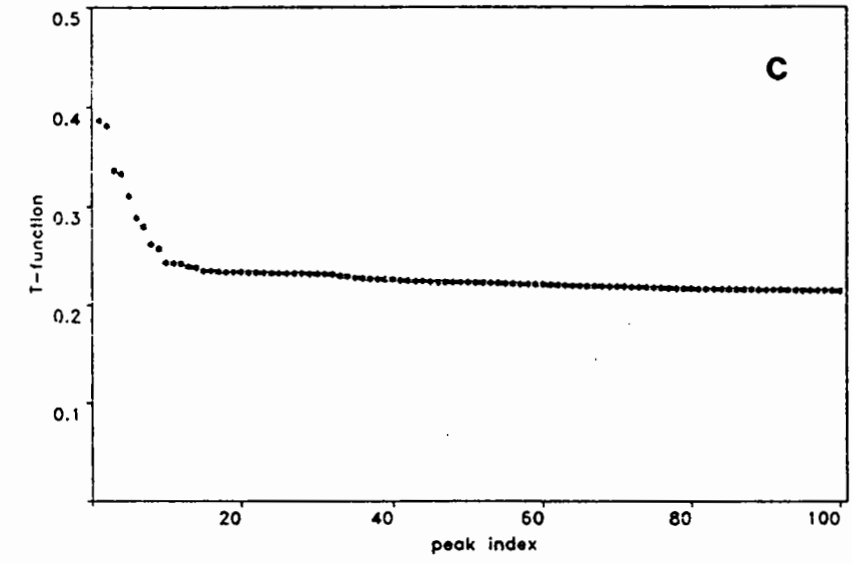


Fig.1. (a) Rotation search at 15.0-4.0 Å resolution with Patterson vector cutoff 5-24 Å . Shown are the RF values of the 111 selected peaks vs their index. (b) Rigid body PC refinement for the selected RF peaks. The 9 highest peaks correspond to the same orientation. (c) Translation search at 15.0-4.0 Å . Peaks 4, 6, and 7 correspond to the same position as the first peak with deviations on x and y of less than 1.0 Å , and 2, 3, 5, and 8 are symmetry related with the first one.

order to save CPU time. When the Patterson vectors between 5-10 Å or 5-40 Å are chosen there are still several peaks with errors of orientation within the radius of convergence of the PC rigid-body refinement, the highest 'correct' peaks are the second highest in the list of selected RF peaks, and less than half a sigma below the highest peaks. This is extraordinary because only a very small portion of the model Patterson vectors are between 5 and 10 Å long (121 compared to 9437 at 32 Å).

3. Determination of molecular position

For space group R3 of the αMMC crystals, both the 3-fold and 3_1 screw axes are along the direction of the c-edge of the cell. The origin along the z axis is arbitrary. Therefore, the translation search can be reduced to a two dimensional asymmetric unit of a= $0 - 2/3, b = 0 - 2/3$. A correlation coefficient grid search using the program X-PLOR was done with a step size of 1.0 Å . Data from 15.0 to 4.0 Å were included. This yielded a significant peak with a maxium value of the "translation function" (linear correlation coefficient between F^2_{obs} and F^2_{calc}) of 0.385, that is 9.7σ above the mean and 5.6σ above the highest noise peak. The initial R-factor was 53% to 2.0 Å . In addition there are also several peaks above the highest noise one, corresponding to either symmetry related or closely similar positions to the first peak (Fig.1c). Subsequently the rotated and translated model was subjected to rigid-body refinement, 30 steps at 10.0-6.0 Å (starting R-factor 46.3%, final R-factor 44.0%) and then at 10.0-4.0 Å , dropping the R-factor by 1% at 10.0-4.0 Å (final R-factor 44.8%). This produced a 1.0° rotation and 0.25 Å translation of the model.

4. Refinement

The TCS model used for the previous studies had been built into a medium resolution MIR map using an incorrect sequence which has 10 extra residues between residue 69 and 70, and 21 residues missing near the C-terminal end, one α-helix was missed and most side chains were misplaced in the model. However the overall fold of the model is essentially correct. To continue the αMMC refinement we deleted all side chain atoms beyoned C_β in the rigid-body refined TCS model to creat a TCS 'alanine' model (TCS_ala). In addition, we constructed an alternative model from the C_α coordinates of RCA by automatically generating main chain and C_β atoms using the program CALPHA (R. Esnouf, unpublished results). This model was then rotated and translated to have a maxmium overlap with the TCS_ala model. These two models were then refined independently using X-PLOR[6], first by molecular dynamics with simulated annealing, and then by individual temperature factor refinement to give R-factors of 39% for both models. To avoid spuriously low R values, all data from 8.0 to 2.0 Å were used thougthout the refinement and the B-factors were rather tightly restrained. Electron density maps calculated from the two refined models showed continuous and well defined density for most of the main chain backbone of the molecule. However, in some regions where the deviation was larger between the two models or deletions or insertions occurred in the sequence, the density was mainly along

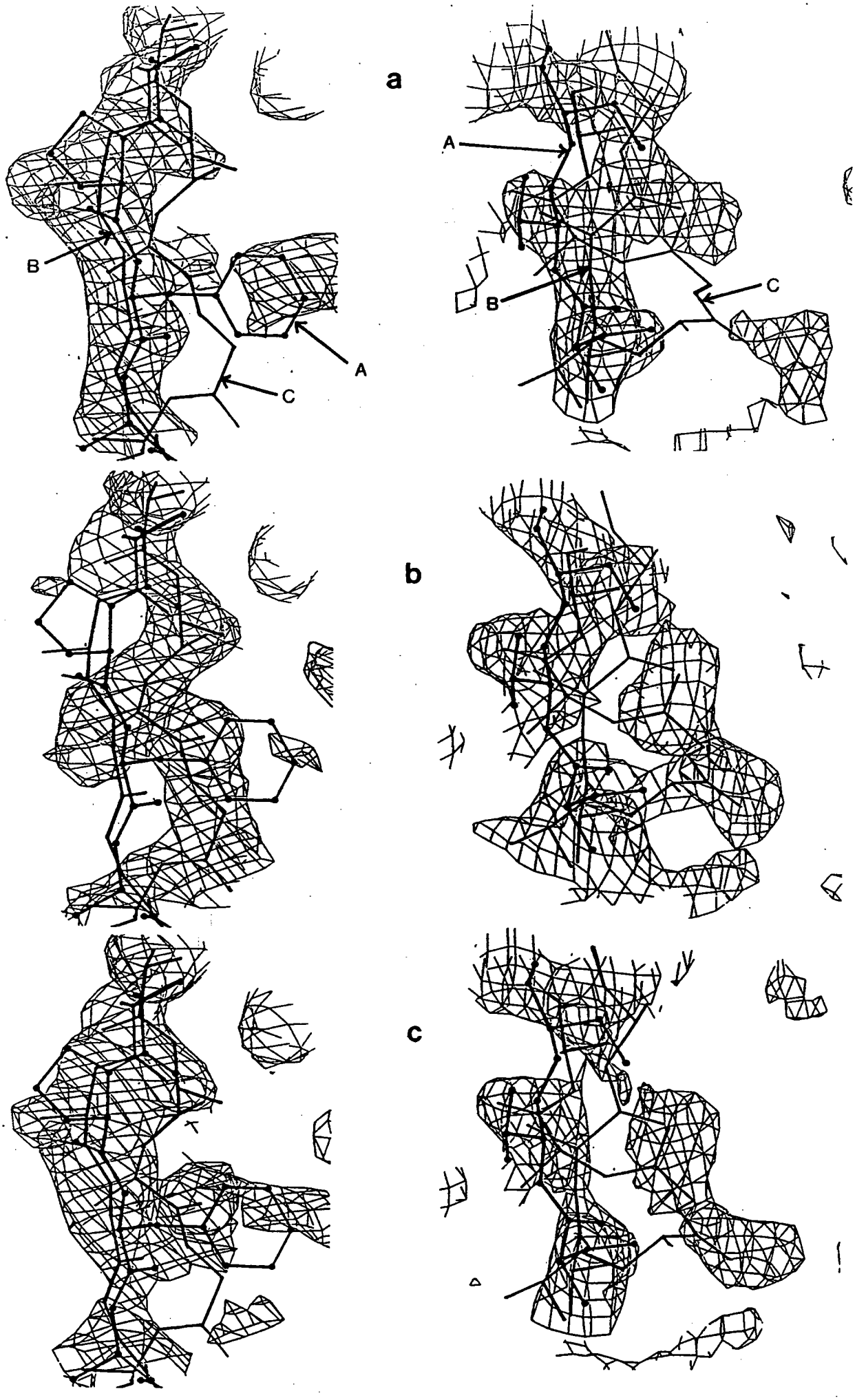

a
A
B
C
A
B
C
b
c

Fig.2. Electron density maps calculated using 3 different sets of F_cs and phases at regions of residues 26-28 (left column) and 40-42 (right column) of αMMC. A, B, and C indicate the final refined αMMC, refined TCS_ala and RCA_ala chains respectively. The maps calculated (a) using TCS_ala model and (b) using RCA_ala model show the well defined density for the correspoding chains. (c) The electron density calculated using the averaged F_cs and phases is mainly along the refined αMMC chain.

the corresponding phasing model, making intepretation difficult. Therefore we averaged the structure factors calculated from the two refined models by taking a vector mean. The density map calculated using the averaged complex structure factors was markedly improved (Fig.2), and allowed us to fit the first 50 side chains unambiguously. For the remaining part of the molecule only the backbone was rebuilt as it was difficult to be confident of the aligment of the amino acid sequence with the electron density map. This rebuilt model was then used in the second round of refinement. After 4 rounds of refinement and model rebuilding,

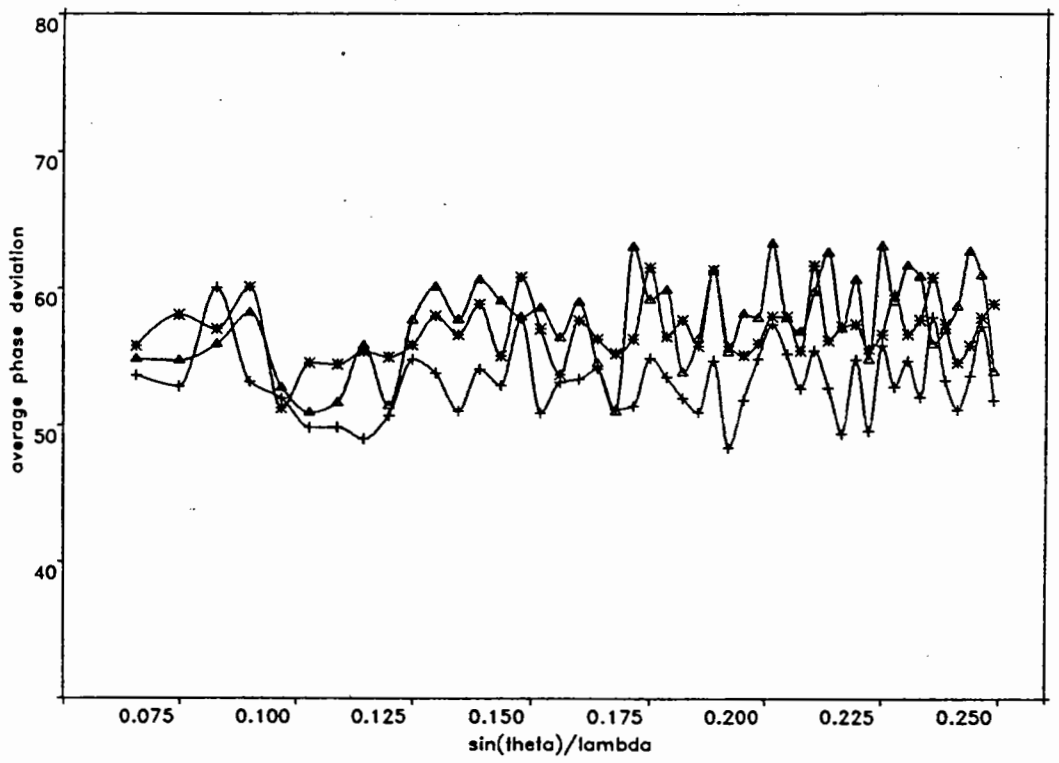


Fig.3. Average deviations between the phases calculated from the final refined αMMC and refined TCS_ala model ($\triangle$), and refined RCA_ala model ($*$), and the averaged phases ($+$). The calculation was done by dividing the reflections into resolution shells of equal volume, and then averaging the deviations of the equivalent reflections from two different data set over each shell.

a total of 246 residues were correctly allocated. The last 17 residues in the sequence cannot be seen in the density map. Since (i) the density for residue 246 is well defined and very neat and (ii) the 17 residues are not confirmed by peptide

sequencing of the natural protein, and (iii) the DNA sequences of genes encoding other plant RIPs contain a carboxyl-terminal extension that is not associated with the mature proteins[7,8], we conclude that the last 17 residues in the reported sequence of αMMC are likely to be an extension of the cDNA not relevent to the mature protein. The structure of αMMC has now been refined to an R-factor of 0.18 for all data from 8.0 to 2.0 Å with an r.m.s deviation of bond lengths from ideality of 0.013 Å . Whereas with most successful cases of molecular replacement, where the deviation between the backbone atoms of the search model and the unknown structure is less than 1 Å , the final refined αMMC structure has an r.m.s. deviation of 1.28 Å for 224 equivalent C_α atoms in TCS and 1.38 Å for 232 equivalent C_α atoms in RCA.

It is interesting to see how much the phases were improved by the averaging. We compared the phases calculated from the two refined polyalanine models and the averaged ones with those from the final refined αMMC structure. The average deviations are 58.1° and 57.1° between the phases from the final model and that from TCS_ala and RCA_ala respectively, while it is 53.2° between the final and the averaged phases. It is obvious that an improvement of only a few degrees can produce a much better density map when the initial phases are poor. As more crystal and NMR structures of macromolecules are determined, the molecular replacement method is likely become the most common technique for structure determination. Our experience suggests that more search models, if available, are useful for the initial phase calculation, especially when the search models have a substantial deviation from the target structure. This has also been shown in the structure determinations of food and mouth disease virus[9] and cricket paralysis virus (D.Logan, E.Fry and D.I.Stuart, unpublished results), where a combined model gave a better solution than any individual one.

Acknowledgement

We thank Dr. Hin-Wing Yeung and his colleagues for providing us αMMC sequence before the publication and Dr. Yvonne Jones for her help with data collection and helpful discussion. We thank the OCMS for support. DIS is a member of the OCMS

References

1. Barbieri, L. and Stirpe, F. FEBS Lett. *195*, (1986), 1-8.

2. Pan, K., Lin, Y., Fu, Z., Zhou, K., Cai, Z., Chen, Z., Zhang, Y., Dong, Y., Wu, S., Ma, X., Wang, Y., Chen, S., Wang, J., Zhang, X., Ni, C., Zhang, Z., Xia, Z., Fan, Z. and Tian, G. Sci. Sin. Ser. B. *30*, (1987), 386-395.

3. Montfort, W., Villafranca, J.E., Monzingo, A.F., Ernst, S.R., Katzin, B., Rutenber, E., Xuong, N.H., Hamlin, R. and Robertus, J.D. J. Biol. Chem. *262*, (1987), 5398-5403.

4. Ho, W.K.K., Liu, S.C., Shaw, P.C., Yeung, H.W., Ng, T.B. and Chan, W.Y.

Biochem. Biophys. Acta, *1088*, (1991), 311-314.

5. Brünger, A.T. Acta Cryst. *A46*, (1990), 46-57.

6. Brünger, A.T. J. Mol. Biol. *203*, (1988), 803-816.

7. Shaw, P.C., Yung, M.H., Zhu, R.H., Ho, W.K.K., Ng, T.B. and Yeung, H.W. Gene, *97*, (1991), 267-272.

8. Benatti, L., Nitti, G., Solinas, M., Valsasina, B., Vital, A., Ceriotti, A. and Soria, M.R. FEBS Lett. *291*, (1991), 285-288.

9. Acharya, K.R., Fry, E., Stuart, D.I., Fox, G., Rowlands, D. and Brown, F. Nature, *337*, (1989), 709-716.

10. Lattman, E.E. Acta Cryst. *15*. (1972) 24-31.

PHASE EXTENSION FROM A CRUDE MODEL. THE STRUCTURE DETERMINATION OF BACTERIOPHAGE MS2.

by

Lars Liljas and Karin Valegård

Department of Molecular Biology, BMC, Uppsala University, Box 590,
S-751 24 Uppsala, Sweden

1. INTRODUCTION

Bacteriophage MS2 is an icosahedral virus with 180 copies of a coat protein of 129 amino acids. It has T= 3 symmetry, which means that the asymmetric unit of the icosahedron contains three chemically identical, but structurally different subunits. MS2 virions have been crystallized in space group R32 (a=b=288.0 Å, c=653.0 Å). The asymmetric unit of the crystal contains 10 icosahedral asymmetric units, which corresponds to 30 coat protein subunits. The virions also contains one RNA molecule and a single copy of a different polypeptide, the A protein.

The redundancy of data in the presence of non-crystallographic symmetry can be used for phase refinement (1). The structure determinations of viruses as well as other macromolecules with several subunits in the asymmetric unit have been simplified by the procedures for real space averaging developed by Bricogne (2, 3). Non-crystallographic symmetry has also been used to extend the resolution of a model in small steps. In the structure determinations of rhinovirus (4) and poliovirus (5), isomorphous substitution was used to determine phases to 5 Å resolution, and the phases were successfully extended to 2.9 Å. In this paper we will describe some of our experiences from the structure determination of MS2 (6, 7) of initial phasing at low resolution and phase extension.

2. INITIAL PHASING

The data was collected on film using oscillation technique. The total number of independent observations between 35 and 2.8 Å resolution was 197000. The direction of a threefold and a twofold axis is fixed by the crystal symmetry and there is therefore

no ambiguities in the orientation of the particle. We constructed an initial model for the phasing from the structure of southern bean mosaic virus (SBMV), a plant T=3 virus. The SBMV protein subunit is a β-sandwich of two four-stranded antiparallel sheets, connected with loops of different lengths. The coordinates of the three subunits in the icosahedral asymmetric unit of SBMV were translated radially by 18 Å, placing then at the expected radius of the MS2 protein shell. All the connecting loops were removed resulting in a sandwich of 90 amino acids. From this model of the protein shell phases were calculated between 300 and 13 Å.

3. PHASE REFINEMENT AND EXTENSION

Phase refinement was done by cyclic averaging of electron density maps calculated with 2Fobs - Fcalc as amplitudes and current calculated phases (3). The envelop used was a spherical shell between 91 and 137 Å. The phase improvement was checked by the correlation between calculated and observed amplitudes. After convergence the phases were extended in steps of less than 1.5 reciprocal lattice points, and the phases were again refined. Initially the correlation coefficient reached 0.85 after several cycles. At 3.5 Å resolution the correlation coefficients were relatively low also after convergence, indicating that the procedure no longer gave correct phases (Fig. 1, lower curve).

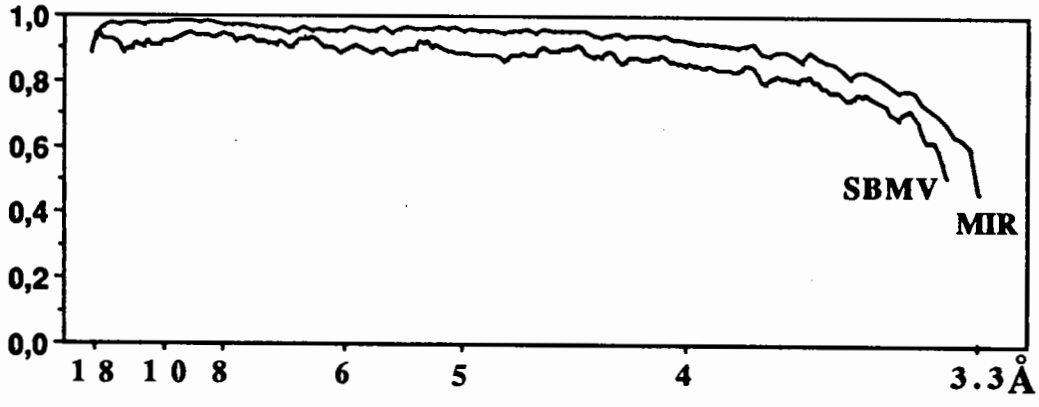


Fig. 1. Plot of the correlation coefficient as a function of resolution.

Using the phases obtained by the phase extension from the SBMV model, difference Fouriers were calculated for two heavy atom derivatives. Both derivative data sets included data to 6 Å resolution, and were very limited. Three peaks were obtained for one of the derivatives, and four peaks for the other, all peaks consistent with the

expected quasi-symmetry of a T=3 virus. However, the peaks were all negative, which indicated that the phases had systematic errors. The derivatives were used to calculate a electron density map to 8.8 Å resolution. The phases were refined by averaging of the map and the resolution of the model was extended in steps of less than 1.0 reciprocal lattice point. The phase extension was stopped at 3.3 Å resolution (Fig. 1, upper curve). The good quality of the phases was obvious already at low resolution, and allowed the definition of an improved envelop, which was used throughout the extensions. This envelop included a small portion of the protein which was excluded in the initial envelop. The map allowed a chain tracing already at 4.0 Å, and a model of the three independent MS2 coat protein chains was built at 3.3 Å resolution. The MS2 coat protein molecule forms tightly interacting dimers (Fig. 2). It has two layers: one β sheet with β meander topology is facing the RNA-containing interior, and the outer surface is formed by a hairpin loop and two helices, which are inserted in a groove in the other subunit of the dimer. The MS2 structure has a completely different topology and a very limited structural similarity to the SBMV model used in the initial phasing.

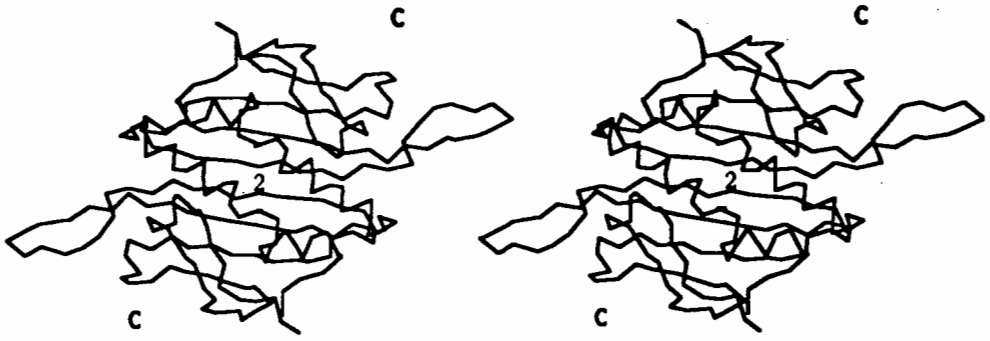


Fig. 2. Stereo picture of a Cα tracing of our final model of the MS2 coat protein dimer.

4. PHASE ANGLE COMPARISONS

The negative peaks in the difference Fouriers indicated that the first phasing attempt had resulted in the Babinet opposite of the true structure. To confirm this we made a two-dimensional histogram of the number of reflections with a certain combination of phase angles at the first and second phasing procedures. We assume that the second phase angle is essentially correct. This is reasonable considering the high quality of the final map. In addition we have made a phase angle comparison between the final

experimental (MIR) phases and phases obtained from the atomic model built in the 3.3 Å map and found that they are very similar. Fig. 3 a - c shows the contoured two-dimensional histogram for reflections between 15. and 5.2 Å resolution. Outside these limits the comparison indicates no significant correlation between the phase angles. As expected the majority of the reflections have a phase difference of 180 degrees, which means that the final phase set obtained from the SBMV model corresponds to the

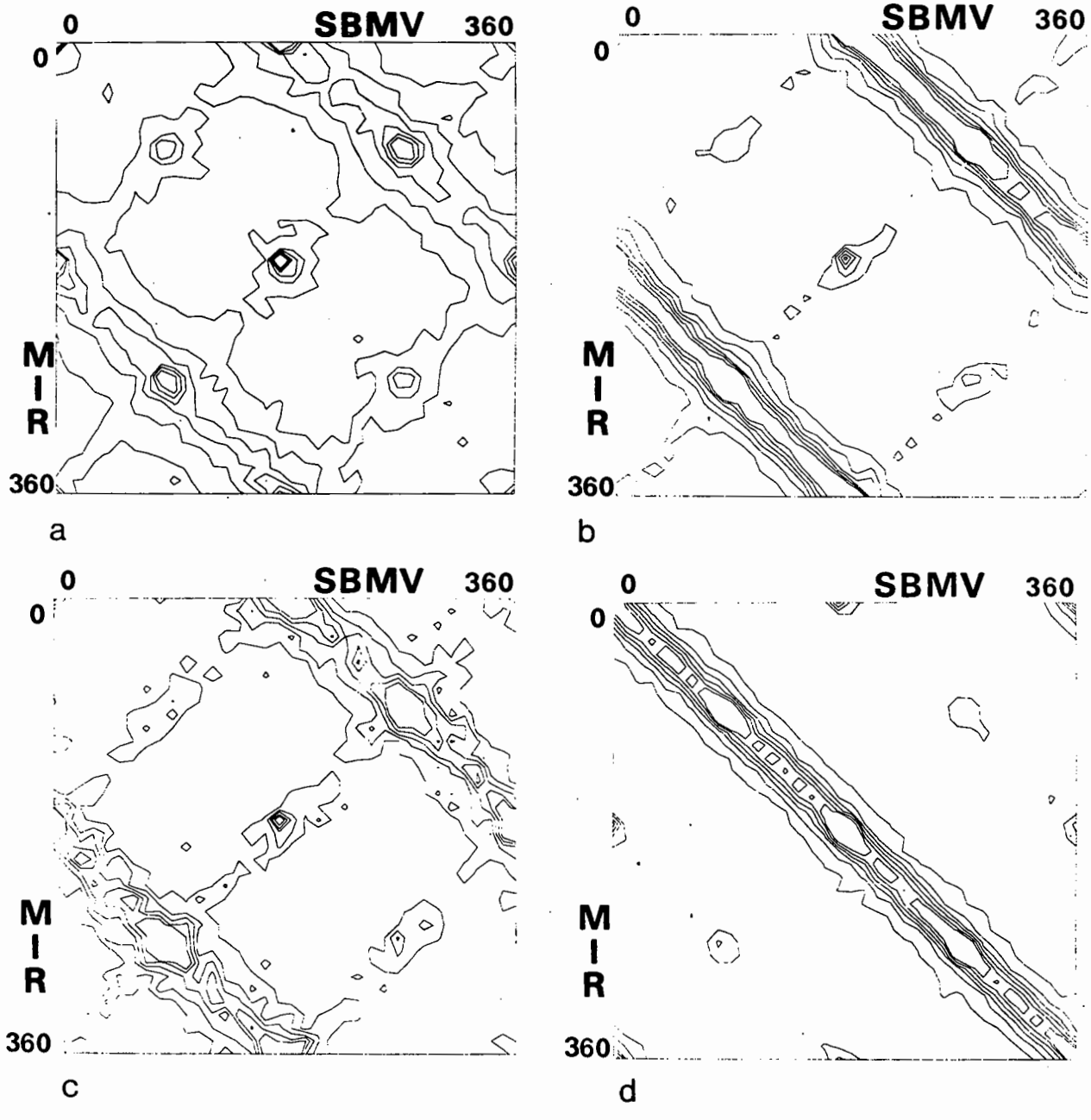


Fig. 3. Contoured two-dimensional plots showing the number of reflections with a certain combination of phase angles in the two phase sets obtained by extension from the SBMV model and from derivative phases.

Babinet opposite of the true structure. However, for a significant number of reflections, the phase relation is in stead $\alpha_2 = -\alpha_1$, corresponding to the wrong hand of the correct structure, or $\alpha_2 = 180-\alpha_1$, corresponding to the Babinet opposite with a different hand. Very few reflections (except the special case of reflections with $\alpha_1 = \alpha_2 = 180$) have the same phase in both the data sets. All these four solutions will be essentially equivalent in the averaging procedure, and the choice of solution will depend on the initial model. It can be noted that this mixture of solutions is present at all resolution intervals.

5. FACTORS OF IMPORTANCE FOR THE PHASE EXTENSION

Although the majority of the reflections obtained a phase angle corresponding to the Babinet opposite in the initial phasing, the quality of the phases was not good enough to allow an interpretation of the electron density map, even if it was contoured at negative density. Fig. 4 shows 5 Å thick sections of the electron density map obtained with various phase sets at 5 Å resolution. A comparison of Fig. 4b and c shows that although most of the strong features in the correct map are present in the first map, it is still very noisy. Further phase extension did not improve the quality, indicating that the phase errors prevented the procedure from converging to a single solution. A section of the first map, contoured at positive density (Fig. 4a) is only slightly more noisy than the same section of the map, contoured at negative density.

To test the influence of the envelop on the phase refinement, we repeated the phase extension starting with the SBMV model at 13 Å resolution, but using our final, improved envelop. The phase extension procedure gave in this case the correct structure in stead of the Babinet opposite. The quality of the phases was also much improved as seen by comparing fig. 3c with fig. 3d. The electron density map was also improved(Fig. 4d), but it was still more noisy than the map in Fig. 4c. This experiment shows that the choice of phase angle set in this case can be influenced by relatively subtle differences in the used procedure. The reason for this is probably that the initial phasing model was very poor, but the power of the phase refinement procedure was still sufficient to lead to a consistent set of phases. The quality of the final phase set is very dependent on the envelop, as seen by the considerable effect by the exclusion of a small volume of the structure in the initially used envelop.

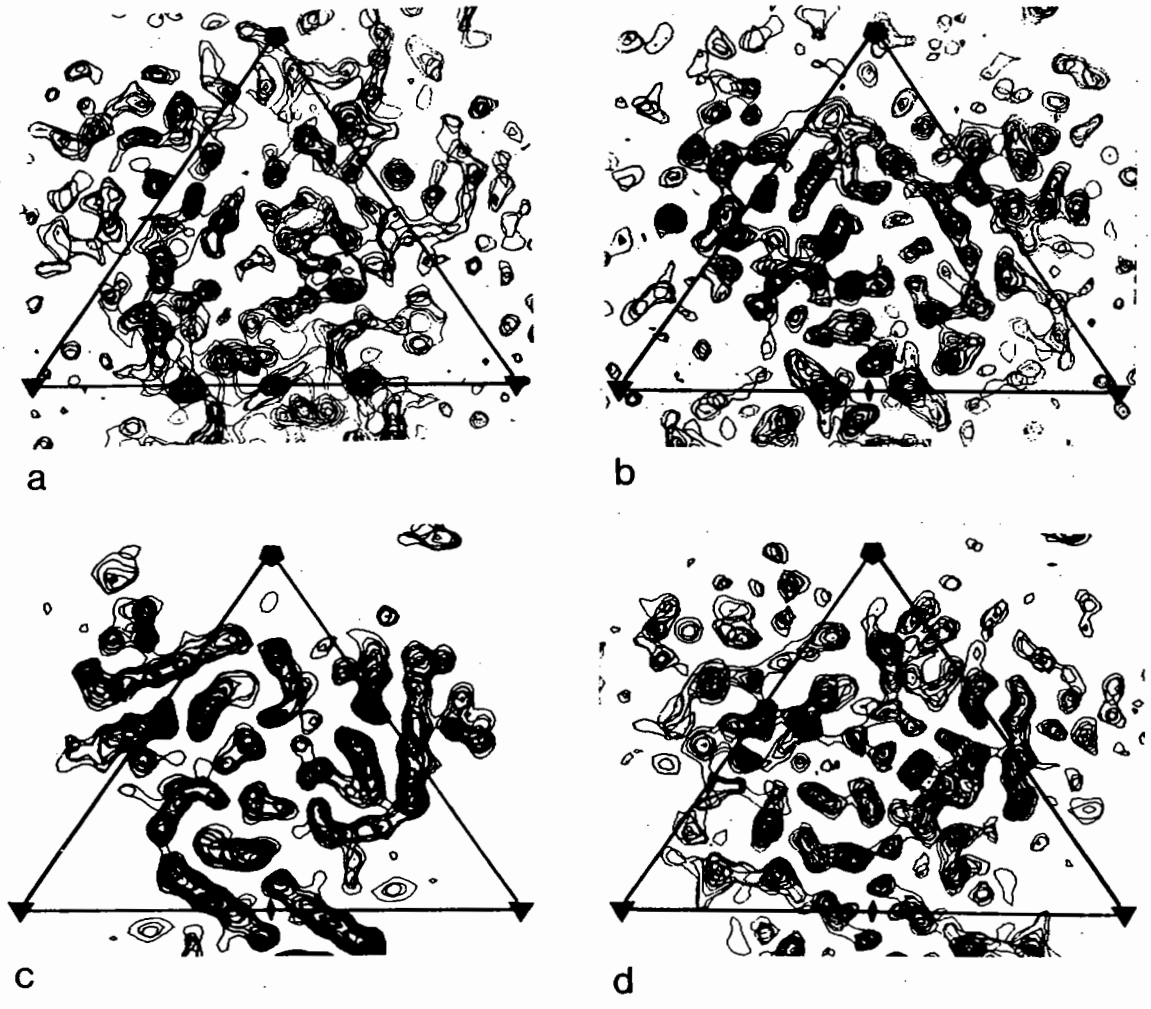


Fig. 4. 5 Å thick sections of the electron density map at 5.0 Å resolution. a) Map after phase extension from SBMV model, contoured at negative density. b) The same map contoured at positive density. c) Map after phase extension starting with isomorphous replacement phases. d) Map from the second phase extension from SBMV model, using improved envelop.

Similar procedures for phase extensions have more recently been used also to solve other virus structures. In the case of canine parvovirus, phase extension was done from a low resolution using a spherical shell of density. Also in this case, the Babinet opposite of the true structure was obtained, and interpretable maps were obtained after phase extension from a limited low resolution phase set obtained by isomorphous replacement (8). The structure of bacteriophage ΦX 174 was solved by phase extension, starting with phases from a low resolution model constructed from the plant virus CpMV, a virus which looks similar to ΦX 174 in electron micrographs, but has no structural similarity at the atomic level (9). Like in the case of MS2, an analysis of

the phase extension for the ΦX 174 data has shown that different phase sets, corresponding to the true structure and its Babinet opposite could be obtained depending on the choice of various parameters (10).

REFERENCES

1. M. G. Rossmann, D. M. Blow, *Acta Cryst.* **16**, 39- (1963).

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

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

4. M. G. Rossmann, *et al.*, *Nature* **317**, 145-153 (1985).

5. J. M. Hogle, M. Chow, D. J. Filman, *Science* **229**, 1358-1365 (1985).

6. K. Valegård, L. Liljas, K. Fridborg, T. Unge, *Nature* **345**, 36-41 (1990).

7. K. Valegård, L. Liljas, K. Fridborg, T. Unge, *Acta Cryst.* **B47**, 949-960 (1991).

8. J. Tsao, *et al.*, *Science* **251**, 1456-1464 (1991).

9. R. McKenna, *et al.*, *Nature* **355**, 137-143 (1992).

10. R. McKenna, D. Xia, P. Willingman, L. Ilag, M. G. Rossmann, *Acta Cryst.* in press, (1992).

The porins: structural homology established by molecular replacement.

Richard Pauptit, ICI Pharmaceuticals, UK.

In this molecular replacement application, there was no detectable sequence homology between the trial model and the target protein structures. It could therefore be considered a non-trivial case. Solving the rotation and translation problems has not actually led to useful phasing models. However, the successful location of the trial model in four unknown unit cells was taken as evidence that structural homology exists between the proteins studied (Pauptit *et al.*, 1991a: this reference is a collaborative effort between laboratories in Basel, Heidelberg and Freiburg and contains most of the work described here).

(i) Introduction and History

The target structures are porin, maltoporin and phosphoporin, all from *E.coli.* The porins are membrane proteins found abundantly in the outer membrane of Gram-negative bacteria. They have pores, allowing diffusion of small metabolites across the membrane. Porin allows non-specific diffusion of molecules up to an exclusion limit of about 600 Da. Maltoporin and phosphoporin show preferential diffusion rates for sugars and phosphorylated metabolites, respectively. Slight cation specificity is found for porin, while phosphoporin is somewhat anion specific. The molecules exist as very stable trimers. A detailed review is given by Jap & Walian (1990).

The *E.coli* porin crystallographic history is a long one. In Basel, the porin work is a collaboration between the groups of Juerg Rosenbusch, where all the biochemistry, biophysics, genetics, isolation, purification and crystallization are expertly carried out, and Hans Jansonius, where all the crystallographic analyses are performed. On the crystallographic side of the project, Mike Garavito, John Jenkins and Rolf Karlsson worked on porin before my involvement. Presently, Tilman Schirmer and Sandra Cowan are continuing this work. There was much excitement when porin was crystallized over a decade ago, opening the field of membrane protein crystallography (Garavito & Rosenbusch, 1980). Of several crystal morphologies, a tetragonal crystal form was selected for further study since it showed the best diffraction characteristics (Garavito *et al.*, 1983). Data collected in 1988 from a single native crystal and a single iridate derivative crystal on a Xentronics detector at MPI Heidelberg with Emil Pai are still the best data we have for the tetragonal crystals - vastly superior to multi-crystal synchrotron film data collected previously. Platinum and iridate derivatives were solved by Patterson and direct methods, but they showed centrosymmetric distributions of heavy atoms (figure 1), which implies that the enantiomer cannot be resolved. Any electron density calculated will contain the superposition of both protein enantiomers. We looked at the density anyway - it appeared random and uninterpretable. Having two centrosymmetric derivatives doesn't help much if the inversion center is the same - this is analogous to the case of two derivatives with the same site but different occupancies. Nonetheless, we tried hard to phase this crystal form. Careful anomalous dispersion measurements were collected on the FAST at Daresbury with the help of Pierre Rizkallah and Miroslav Papiz, but we were never able to extract an accurate anomalous signal. A number of chemical (by Malcolm Page, Basel) and genetic (by Robin Ghosh, Basel) modifications were constructed to include heavy atoms or heavy atom binding moieties, but these presented crystal quality problems. There is no inherent reason why the derivatives can only be centrosymmetric (just bad luck), so the search for derivatives continued without success. It was always thought that 6-fold symmetry averaging would improve even a poor starting set of phases in this crystal form, but the starting sets we produced were apparently too far removed from the truth.

It seemed obvious to me that an alternate crystal form might circumvent the centrosymmetric

TABLE I
Crystal Parameters

Porin	Space group	Cell parameters	V_M[a]	N[b]	Ref.
R. capsulatus:					
Porin	$R3$	95.3, 95.3, 146.8 Å 90, 90, 120°	4.1	1	Nestel *et al.* (1989)
Porin	$R3$	92.3, 92.3, 146.2 Å 90, 90, 120°	3.8	1	Kreusch *et al.* (1991)
E. coli:					
Porin	$P4_2$	154.6, 154.6, 171.0 Å 90, 90, 90°	4.6	6	Garavito *et al.* (1983)
Porin	$P321$	118.5, 118.5, 52.8 Å 90, 90, 120°	2.9	1	Pauptit *et al.* (1991)
Maltoporin	$C222_1$	130, 213, 216 Å 90, 90, 90°	5.2	3	Stauffer *et al.* (1990)
Phosphoporin	$P6_322$	121.0, 121.0, 111.1 Å 90, 90, 120°	3.2	1	Tucker *et al.* (1991)

[a] Volume to mass ratio (Å^3/Da).
[b] Number of monomers per asymmetric unit.

Figure 1

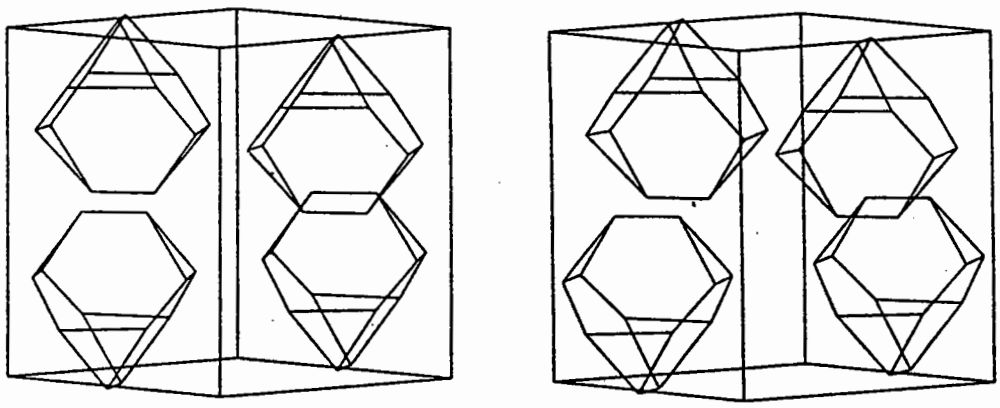


Figure 2

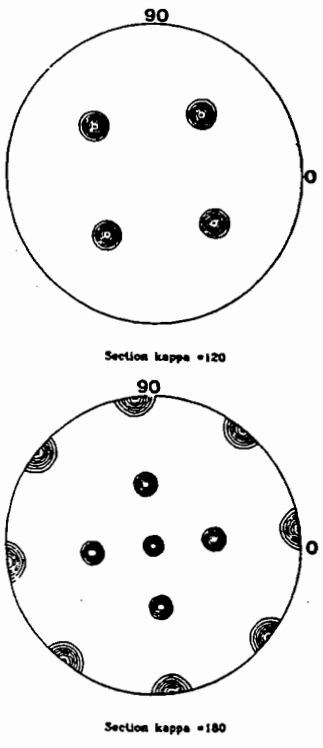
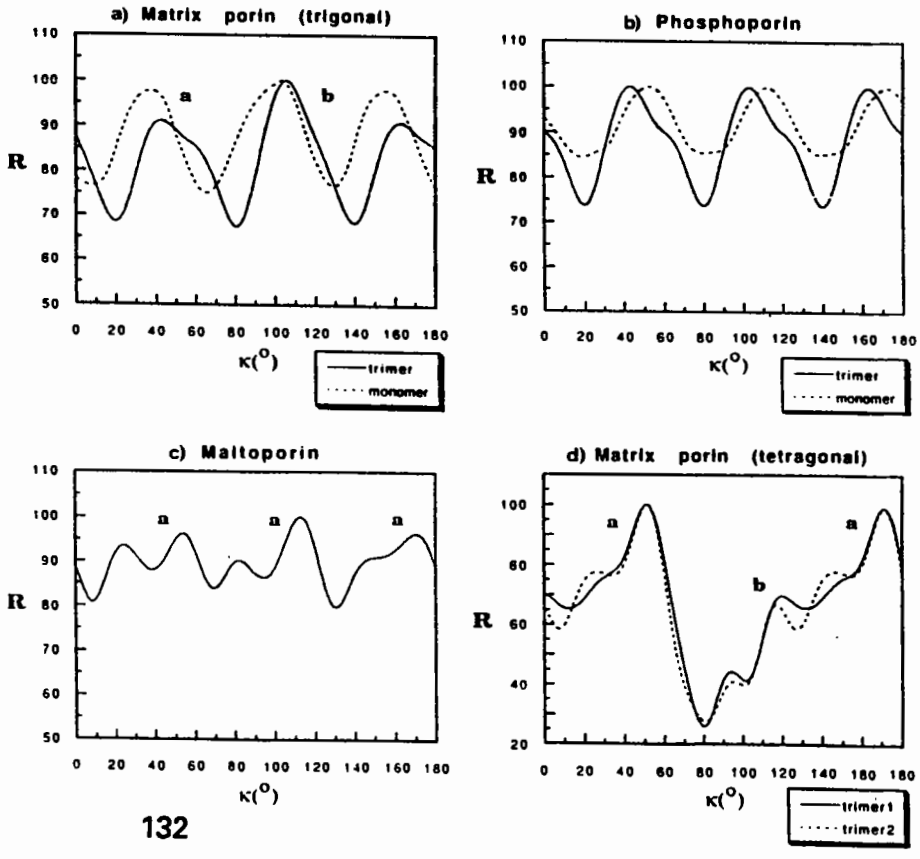


Figure 3



heavy atom problem. This idea received more attention once Freiburg offered competition. Attempts to reproduce the other morphologies observed in the early days (Garavito *et al.*, 1983) were not successful, but the trials led to a new trigonal crystal form of porin (Pauptit *et al.*, 1991b). Again derivatives were found containing centrosymmetric heavy atom distributions!!! At this stage, I left Basel. It wasn't long before Sandra Cowan (Basel) found a good derivative and solved the structure (manuscript submitted), confirming the molecular replacement analysis presented below.

Data from crystals of maltoporin (Stauffer *et al.*, 1990) and phosphoporin (hexagonal form: Tucker *et al.*, 1992; trigonal form: Steiert *et al.*, 1992) were also collected recently. No useful derivatives have been found to date for either protein. However, phosphoporin has 63% sequence identity with porin and forms isomorphous trigonal crystals, for which the structure was solved by Tilman Schirmer using molecular replacement and the refined porin structure as a trial model (Cowan *et al.*, manuscript submitted).

The trial model is porin from *Rhodobacter capsulatus,* for which, to our envy, the structure was rapidly solved using four good derivatives (Weiss *et al.*, 1989). The similarity of the *R.capsulatus* structure to low resolution electron crystallographic models of *E.coli* porins (Jap, 1989) was noticed. It seemed sensible to attempt molecular replacement, despite the lack of sequence homology. Crystal parameters of the various porins are reproduced in Table 1.

(ii) Molecular Replacement

We initially attempted the molecular replacement with crude trial models. At the first indication from Freiburg that porin was a 16-stranded antiparallel beta-barrel, we obtained encouraging results using a constructed barrel trimer model. The rotation function oriented the barrel trimers parallel to the heavy atom triangles in the tetragonal crystal form. We tried a model extracted from a stereo C-alpha plot (Weiss et al., 1990) before we were given the 3 Å model. When *R.capsulatus* porin was refined at 1.8 Å (Weiss *et al.*, 1991), Georg Schultz kindly provided the backbone coordinates. The quality of our results improved with each improvement of the trial model, which suggests a convergence of trial and target structures, supporting the putative similarity.

CCP4 programs were used. The trimer axis in the trial model was along *z*. Using POLARRFN, the polar angle convention conveniently allows us to visualize Omega as the tilt angle of the trimer axis away from *z*, and Kappa as the azimuthal rotation of the trimer about the new trimer axis. Structure factors were calculated in a large *P* 1 cell using program GENSFC. TFSGEN was used for the translation function. For use with tetragonal porin, TSEARCH was modified by Tilman Schirmer to carry out a 5-dimensional search (see below). The resolution range used was 10-6 Å, but other ranges gave equivalent results.

Let us consider the four target proteins in turn:

(a) Tetragonal porin

Although the heavy atom distribution was useless for phasing, it indicated the aggregation of trimers in the unit cell. The heavy atom structure consists of perfect tetrahedra of trimers (figure 1). The self-rotation function (figure 2) using native data is consistent with tetrahedral symmetry, showing 3-fold axes at 54° to 222 axes. The conclusion is that porin molecules pack as tetrahedra of trimers, as for phaseolin (Lawrence *et al.*, 1990). There are only two trimers in the asymmetric unit, so one 2-fold of the 222 system is crystallographic. Thus, the tetrahedron is centered on a crystallographic 2-fold (say at $x,y=1/2,0$) with arbitrary z - translation since the space group ($P4_2$) is polar.

The rotation function gave clear indications for the orientations of the trimer axes, agreeing

entirely with the self-rotation function and heavy-atom geometry. The variation of the rotation function with Kappa, *i.e.*, as the trimer is rotated about the new trimer axis defined by Omega and Phi, is shown in the Kappa-plots (figure 3). At Kappa=0°, the two trimers were placed so as to obey local 2-fold symmetry, hence the Kappa value at the correct peak applies to both trimers. Peaks recurr every 120°, as expected from the trimer local 3-fold. There is a strong set of peaks **a** and a weaker set **b**, displaced 60° in Kappa. These two possible orientations of the trimer pair correspond to two different aggregations, essentially one is upside-down with respect to the other. In one the smooth side of the barrel is closest to the tetrahedral center, in the other the rough (long loop) side of the barrel is closest to the tetrahedral center. High correlation for upside-down pseudo-solutions is not unexpected given the intrinsic symmetry of the barrel. A successful translation function would discriminate between these two possibilities. The only degree of freedom in the translation function is in fact the distance of the trimer from the tetrahedral center - everything else is fixed by local symmetry, for which we fortunately know both orientation *and position*.

TFSGEN was not successful. Not surprising, since the trial model (a polyalanine trimer) represented much less than half the scattering matter in the asymmetric unit. An R-factor search, in which the model was positioned at various incremental distances from the tetrahedral center, gave minima that varied with resolution, lending no confidence in the results. The position of the model along the local triad is, however, readily determined by packing considerations. Moving the trimer too close to the tetrahedral center causes clashes with others trimers in the tetrahedron, while moving it too far away causes clashes with the adjacent tetrahedron of trimers.

Successful computational positioning of the trimers could be carried out by Tilman Schirmer only after the trial model had been improved. The *R.capsulatus model* which had been successfully positioned in the trigonal unit cell and subjected to CORELS rigid body refinement was used. TSEARCH was modified to allow a 5-dimensional search: each trimer in the asymmetric unit has 3 translational degrees of freedom, but for one trimer the z-coordinate can be fixed because $P4_2$ is a polar space group. 5D-TSEARCH is not very fast, but the search could be limited to the vicinity of the expected solution. A minimum was obtained only for trimers oriented according to peaks **b** in the rotation function (figure 3). Thus the positions of the trimers could be established by an R-factor search, and it turns out that it is the rugged side of the trimers that face the tetrahedral centers (figure 4a is wrong! apologies...). The positions obtained give rise to good packing.

(b) Trigonal porin

Here, the asymmetric unit is a monomer, so the trimer axis must coincide with crystallographic symmetry. In $P321$, only the 3-fold axes at x,y =1/3,2/3 and 2/3,1/3 are suitable since the 3-fold at x,y= 0,0 is intersected by 2-fold axes. This places the trimer axes less than 70 Å apart. This led me to suggest that the protein surfaces that are exposed to the membrane interior form direct hydrophobic crystal contacts, something which has not been observed until now. Only limited orientations of the trimer around the trimer axis do not cause packing clashes with the adjacent trimer (figure 4b). Any *z*-translation of the trimer along the 3-fold is feasible. The *c*-dimension of the unit cell corresponds to the height of the porin trimer. Thus, the crystal packing will consist of adjacent columns of trimers.

The orientation search using a monomeric trial model positioned the trimer axis along *z*, (*i.e.*, Omega=0) and two sets of peaks, 60° apart, were found in the Kappa plot (figure 3). One of these orientations can pack if the trimer is placed on the 3-fold at x,y =1/3,2/3, while the other can pack only if the trimer is placed on the 3-fold at x,y =2/3,1/3, which would generate an upside-down trimer on the first 3-fold. So again, we have an upside-down pseudo-solution. These two packings are physically indistinguishable, differing only in indexing. With the current indexing, only one of these possibilities could lead to a successful translation

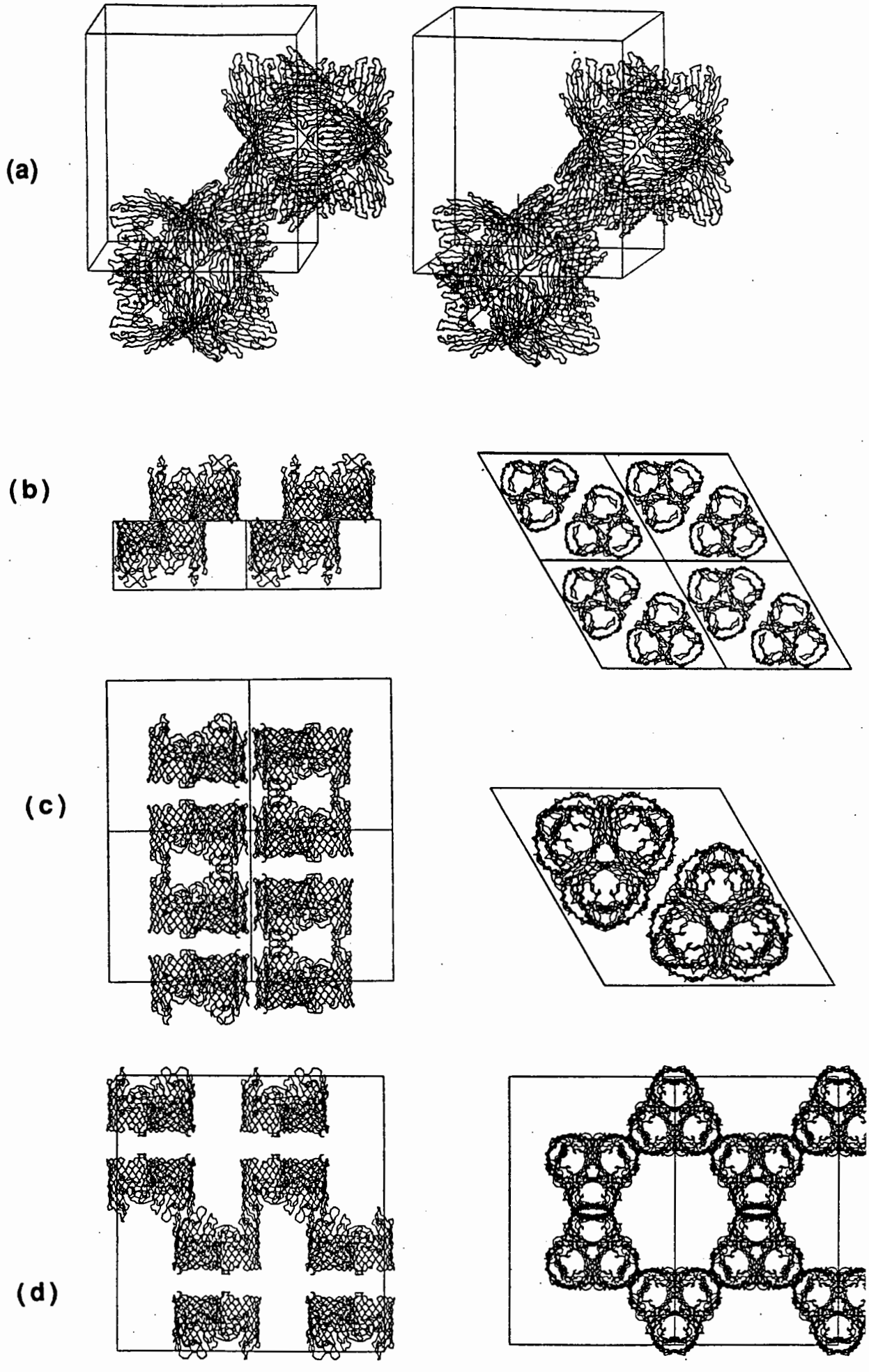


Figure 4

function.

Better discrimination and a stronger rotation function signal were obtained using the trimer as search model. The trimer constitutes the contents of 3 asymmetric units, yet the signal was enhanced over three times. If the oligomeric structure is conserved, it is justifiable (even recommended) to make use of as much of the oligomer in the search model as possible - then Patterson overlap corresponding to inter-subunit vectors is included in the rotation function signal.

The translation function confirmed that the trimer oriented according to peak **b** in the Kappa plot is centered on the 3-fold axis (the other orientation produced no translation function signal). The peak was sharp in the *xy*-plane but smeared out over many *z*-sections. This was rationalized through the columnar nature of the packing along *z*, reducing discrimination in the *z*-direction. Figure 4b shows two views of the packing, along *a* and down *c*.

(c) Phosphoporin

The hexagonal crystal form also has a monomer in the asymmetric unit, so again the trimer axes must coincide with the crystallographic 3-folds at *x,y* =1/3,2/3 and 2/3,1/3. These 3-fold axes are intersected by 2-fold axes perpendicular to *z* at *z*=1/4 and *z*=3/4. So unlike the trigonal form of porin, the *z*-translations of the trimers are restricted: the trimers must lie exactly between the 2-fold axes which relate them. The *c*-dimension corresponds to two trimer heights. The azimuthal (Kappa) rotation of the trimer about the 3-fold is again limited since the 3-folds are relatively close.

The rotation function indeed orients the trimer axes parallel to *z*. This time crystallographic symmetry imposes the 60° repeat found in the Kappa plot (figure 3). Again, better discrimination and a higher rotation function signal is obtained when the trimer is used as a search model. The translation function again gives a peak at the 3-fold which is sharp in the *xy*-plane and smeared out along *z*, presumably for the same reasons. Figure 4c shows 2 views of the packing, along the *ab* diagonal and down *c*.

The trigonal crystals of phosphoporin were grown following these molecular replacement analyses, but we expect they would produce similar results to the trigonal crystals of porin because they are isomorphous.

(d) Maltoporin

The cleanest molecular replacement results were obtained for maltoporin. This was a surprise, since maltoporin (420 aa) is considerably larger than the other *E.coli* porins (340 aa) and much larger than the trial model (300 aa). It was a welcome surprise, since it allowed us to include maltoporin in the family of structurally homologous porins.

There is a trimer in the asymmetric unit. The rotation function aligned the trimer axis closely parallel to *z*. There is a 60° repeat in the Kappa plot (figure 3) due to crystallographic symmetry. The highest peak was correct. The translation function presented three orthogonal streaks (parallel to the cell edges). The intersection of the streaks was the obvious correct solution, although it was not the highest peak in the map. The resultant packing is shown viewed down *a* and down *c* in figure 4d and agrees entirely with that predicted by Tilman Schirmer.

(3) Strong intensity distribution and beta-strand tilt

Jap and Walian (1990) presented an elegant way of extracting the beta-strand tilt angle from (in their case electron) diffraction data. We used this to obtain the tilt angles for the strands in all the porins studied. In each case, the tilt angle was 30-40°, providing further evidence for

the similarity between these proteins. Details of this analysis are described in Pauptit *et al.* (1991a), and will not be reproduced here since they do not concern molecular replacement directly.

(4) Comparison with solved structure

We are fortunate in that the structure of *E.coli* porin has now been refined at 2.5 Å, so we are able to compare the extent of similarity between the trial and target structure. The simplest way to do this is by visual inspection (figure 5). Although the topology of the 16-stranded beta-barrel is conserved, there are considerable differences in barrel height and loop structure. The loop structure contains helical segments in both trial and target structures, but at different locations in the sequence. The helical segments have been omitted from the figures. For use in phasing, the trial model must be placed very accurately in the target unit cell. It seems that the differences in structure prevent such an accurate placement.

(5) Conclusions and general remarks

(a) Molecular replacement was used to show that the porins have the same fold.
(b) Phases obtained were not good. This is probably because the trial model is not sufficiently representative of the target proteins, whereby the molecular replacement results become inaccurate. This appears contradictory to (a), but it is just a question of degree(s).
(c) Oligomers that comprise several asymmetric units can give better results than a protomer for the rotation function. When this happens, there is an independent indication that the oligomer is maintained.
(d) Predicted packing arrangements are extremely useful for lending confidence to molecular replacement results. Another good indication is that for a correct solution, the R-factor increases with resolution, whereas for a wrong solution there is no reason for such behaviour and the R-factor is random in all resolution shells. Consistency with self-rotations and heavy atom structures, if available, is also very useful.
(e) A refined trial model should give better results in molecular replacement even though the resolution of the analysis is limited. The analysis depends on an overlap of Patterson vectors, which depend on interatomic distances. The more accurate these are, the greater the relevance to physically real observed data, and the greater the Patterson overlap, whatever the resolution. In our analysis as well, improving the trial model made a lot of difference.
(f) When things don't work, it would not be unwise to try another source or another crystal form.

I sincerely thank all porin people and the organizers of the meeting.

Figures:

(1) Iridium centrosymmetric heavy atom structure in tetragonal porin. The heavy atoms form perfect tetrahedra, which are related by inversion centers between them.

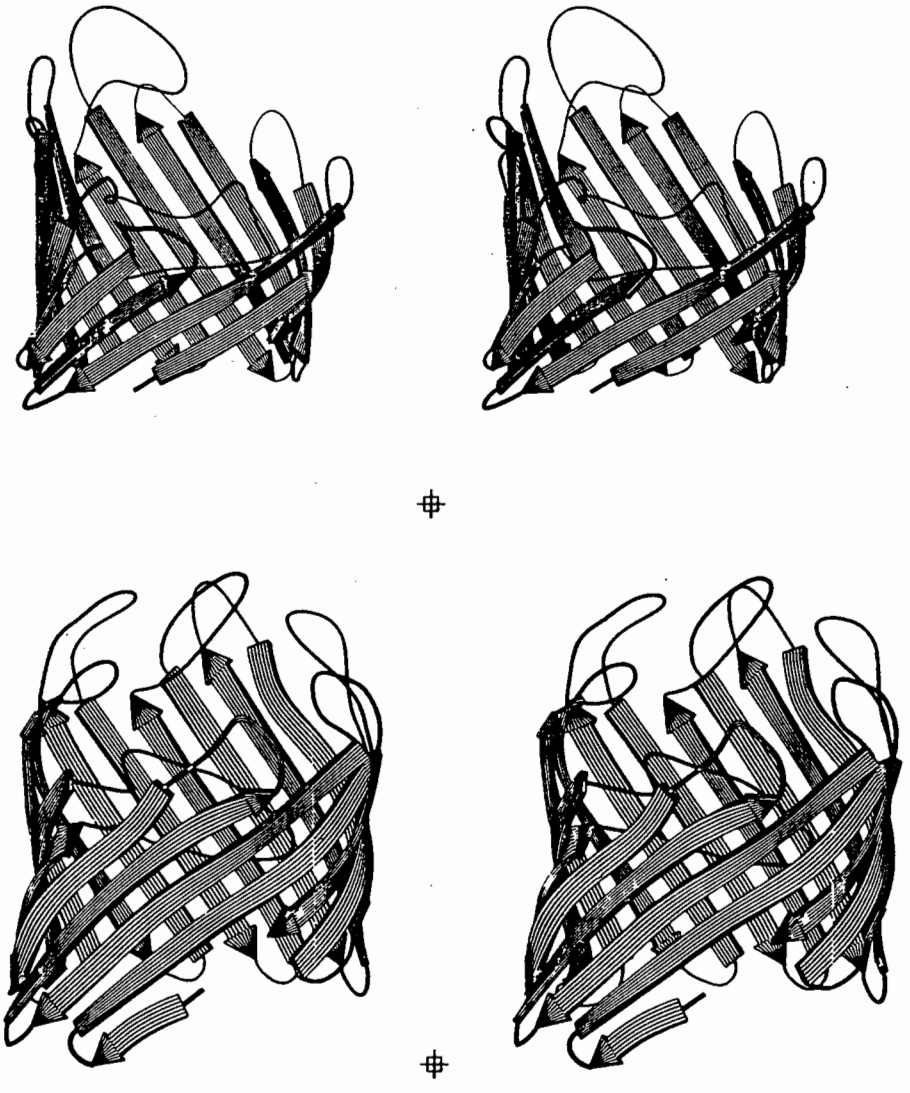
(2) Tetragonal porin self-rotation function: kappa=120° section and kappa=180° section. The local 3-fold and 2-fold axes are consistent with the tetragonal symmetry of the heavy atoms.

(3) "Kappa plots" for the four cross-rotation functions, normalized to a maximum of 100.

(4) Packing diagrams for (a) tetragonal porin (stereo), (b) trigonal porin, (c) phosphoporin and (d) maltoporin.

(5) RIBBON drawings of the trial model and the target structure. Helices in the loop structure have been omitted.

trial model



target structure

Figure 5

References:

Garavito,R.M. & Rosenbusch,J.P. (1980). *J.Cell.Biol.* **86**, 327-329.

Garavito,R.M., Jenkins,J.A., Jansonius,J.N, Karlsson,R. & Rosenbusch,J.P. (1983). *J.Mol.Biol.* **164**, 313-327.

Jap,B.K. (1989). *J.Mol.Biol.* **205**, 407-419.

Jap,B.K. & Walian,P.J. (1990) *Q.Rev.Biophys.* **23**,367-403.

Kreusch, A., Weiss,M.S., Welte,W., Weckesser,J. & Schulz,G.E. (1991) *J.Mol.Biol.* **217**, 9-10

Lawrence,M.C., Suzuki,E., Varghese,J.N., Davis,P.C., Van Donkelaar, A., Tulloch,P.A. & Colman,P.M. (1990). *EMBO J.* **9**, 9-15.

Nestel,U., Wacker,T., Woitzik,D., Weckesser,J., Kreutz,W. & Welte,W. (1989). *FEBS Letters* **242**, 405-408.

Pauptit,R.A., Schirmer, T., Jansonius,J.N., Rosenbusch, J.P., Parker, M.W., Tucker, A.D., Tsernoglou,D., Weiss,M.S. & Schulz, G.E. (1991a). "A common channel-forming motif in evolutionarily distant porins", *J. Struct.Biol.* **107**, 136-145.

Pauptit,R.A., Zhang,H., Rummel,G., Schirmer,T., Jansonius,J.N. & Rosenbusch,J.P. (1991b). *J.Mol.Biol.* **218**,505-507.

Stauffer,K.A., Page,M.G.P., Hardmeyer,A., Keller,T. & Pauptit, R.A. (1990). *J.Mol.Biol.* **211**, 297-299.

Steiert,M., Ghosh,R., Schirmer,T. & Rosenbusch,J.P. (1992). *Experientia* **48**, A22.

Tucker,A.D., Jackman,S., Parker,M.W. & Tsernoglou,D. (1991) *J.Mol.Biol*, **222**, 881-884.

Walian,P.J. & Jap,B.K. (1990). *J.Mol.Biol.* **215**, 429-438.

Weiss,M.S., Wacker,T., Nestel,U., Woitzik,D., Weckesser,J., Kreutz, W., Welte,W. & Schultz,G.E. (1989). *FEBS Letters* **256**,143-146.

Weiss,M.S., Wacker,T., Weckesser,J., Welte,W. & Schultz,G.E. (1990). *FEBS Letters* **267**, 268-272.

Weiss,M.S., Kreusch,A., Schiltz,E., Nestel,U.,Welte,W. ,Weckesser,J. & Schultz,G.E. (1991). *FEBS Letters* **280**,379-382.

Use of Molecular Replacement in the Structure Determination of α-Lactalbumins

K. Ravi Acharya
Department of Biochemistry, University of Bath,
Claverton Down, Bath BA2 7AY, England.

Introduction

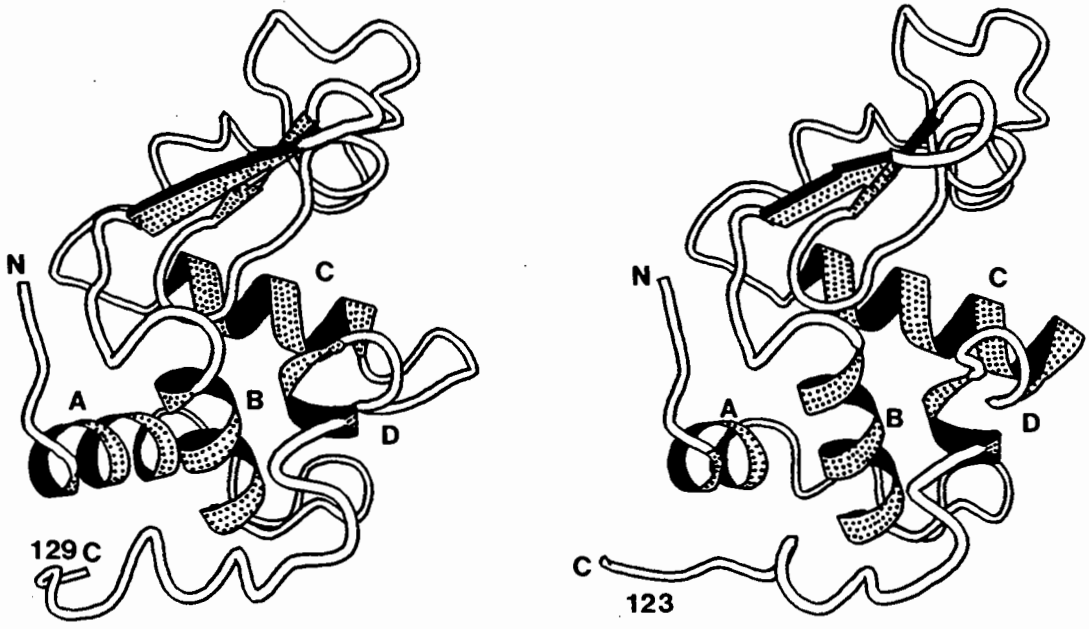
α-Lactalbumin (α-lac) is a globular protein secreted in the lactating mammary gland and has a molecular weight of 14,200. It regulates lactose biosynthesis by modulating the specificity of trans-golgi galactosyltranferase (GTase) (Hill *et al.*, 1975). Comparison of amino acid sequences (Brew *et al.*, 1970; Findlay & Brew, 1972), gene sequences (Dandekar & Qasba, 1981; Hall *et al.*, 1982) and the exon-intron organisation of the genes (Qasba & Safaya, 1984) firmly established that α-lac is homologous to C-type lysozymes having evolved by divergence from a common ancestor. α-Lac is a metalloprotein (Hiraoka *et al.*, 1980) having a tight Ca^{2+} binding site (apparant affinity constant as large as 10^6 - 10^9 M^{-1}) (Segawa & Sugai, 1983; Hamano *et al.*, 1986; Berliner & Johnson, 1988; Kronman, 1989) and differs from lysozyme in its biological role.

Crystal structure analysis of baboon milk α-Lactalbumin

X-ray crystallographic studies of the three-dimensional structure of α-lac were begun many years ago but have been frustrated, first by the difficulty of finding amenable crystals to study and, subsequently, by the difficulty of preparing isomorphous heavy-atom derivatives for use in the multiple-isomorphous-replacement method of x-ray structure analysis. Baboon milk α-lac, which has an amino-acid sequence closely similar to that of human α-lac (over 90 % sequence identity) was the first species found to give suitable crystals for x-ray analysis. These dimond shaped crystals belong to orthorhombic space group with unit-cell dimensions a=35.5 Å, b=69.1 Å , c=46.1 Å. The space group is $P2_12_12$ and there is one molecule of α-lac per asymmetric unit (Aschaffenberg *et al.*, 1979). Incorporation of mercury into one of the disulphide bridges of the crystalline protein gave a useful derivative at low resolution (4.5 Å). Studies at this resolution by Smith *et al.*, 1987), together with earlier predictions based upon comparisons of amino acid sequences left no doubt that class C lysozymes and α-lacs are homologous proteins. However, initial attempts at refinement using a model placed on the basis of low resolution studies failed and therefore a rather cautious approach was adopted as described below.

Combination of low resolution phases (to roughly 4.5 Å resolution) and the high resolution native data set (1.7 Å resolution) formed the master and the starting data set for refinement. A common orientation between hen egg-white lysozyme and α-lac (Smith *et al.*, 1987) was derived from the low resolution electron density map and applied to the hen egg-white co-ordinate set. For historic reasons, we chose to use hen egg-white lysozyme as a starting model, the sequence homology between this molecule and human α-lac (the species most closely related to baboon α-lac for which the sequence is known) was the same as between human lysozyme and human α-lac. The unmodified hen egg-white coordinates were fed into the CORELS program (Sussman, 1985). Constrained

refinement was performed with all the measured data in the stated resolution ranges. It was found at an early stage that the convergence was unsatisfactory in the absence of a solvent component in the refined model. This was acheived by a simple modification of the atomic form factors based on Babinet's principle (incorporated into CORELS program by A.G.W.Leslie, based on a modified form of the equation due to Fraser *et al.*, (1978). Treating hen egg-white lysozyme as a single rigid body, 6 cycles of refinement (infinity to 8 Å resolution) were performed. The crystallographic R-factor dropped from 0.546 to 0.455. At this stage, the model was divided into 10 pieces based on blocks of secondary structure in the lysozyme molecule. These were treated as seperate unlinked rigid bodies and 5 cycles of refinement were done using all data to 3 Å resolution. This lowered the R-value from 0.49 to 0.43. At this stage, the human α-lac sequence was incorporated and the model was fitted to the electron density in a $2/F_o/ - /F_c/$ map and the geometry was idealised. Further refinement was performed treating individual residues (main chain and side chain seperately) as different (but linked) CORELS groups. Isotropic temperature factors were refined for each such group. Refinement converged at an R-value of 0.32 on all data to 2.7 Å. At the end of the CORELS refinement, the overall quality of the Fourier map was greatly improved.



lysozyme

α-lactalbumin

Figure 1

Further refinement was carried out using the stereochemically restrained least squares refinement method using PROLSQ at 1.7 Å resolution (Hendrickson & Konnert, 1980). Simple solvent correction based on Babinet's principle was used to model bulk solvent throughout the refinement. The final crystallographic R-factor was 0.22 for 1141 protein atoms. In the final model, the root-mean-square deviation from ideality for bond distances was 0.015 Å and for angle distances was 0.027 Å. The refinement was carried out using the human α-lac sequence and 'omit maps' calculated during the course of the refinement indicated eight possible sequence changes in the baboon α-lac x-ray sequence. During the refinement, a tightly bound calcium ion and 150 water molecules were located (Stuart *et al.*, 1986; Acharya *et al.*, 1989). Figure 1 shows the tertiary structure of α-lac and hen egg-white lysozyme. An overall alignment of α-lac and hen egg-white lysozyme has 122 equivalent α-carabon atoms with a root-mean-square deviation of 1.85 Å . All large deviations occur in the loops (where all sequence deletions and insertions are found).

Crystal structure analysis of human milk α-Lactalbumin

Crystallisation of human α-lac was carried out using the protocol described by Fenna (1982). The crystals belong to orthorhombic space group $P2_12_12$, with unit cell dimensions a=33.6 Å, b=69.4 Å and c=47.3 Å (for baboon α-lac, $P2_12_12$ space group, a=35.5 Å, b=69.1 Å and c=46.1 Å). There are four molecules in the unit cell (i.e. one molecule/assymetric unit). X-ray intensity data to 1.7 Å were collected on a Nicolet Imaging Proportional Counter Area Detector (XENTRONICS) using a Rigaku RU-200 rotating anode X-ray generator source.

The crystals of baboon α-lac and human α-lac belong to the same space orthorhombic space group and appear almost isomorphous (mean fractional isomorphous difference = 0.28). First, all the human α-lac amino acid changes (Hall *et al.*, 1982) were modelled into baboon α-lac X-ray structure. To start with, a simple rigid body R-factor minimisation of this model was performed using Axel Brünger's crystallographic refinement program X-PLOR (Brünger *et al.*, 1977; Brünger, 1988; Brünger, 1989; Brünger *et al.*, 1989) implemented on a Convex 210 mini super computer. After 20 cycles of energy minimisation against the X-ray data, at various resolution shells, the crystallographic R-factor dropped to a minimum of 0.30 at 3.0 Å resolution. At this stage it was difficult to further improve the fit to the experimental data. The refinement was halted and a fresh start made using the MERLOT package of molecular replacement programs (Fitzgerald, 1988). The initial search was done using the baboon α-lac protein model (Acharya *et al.*, 1989) (excluding water molecules) with human α-lac amino acid sequence changes incorporated into the baboon α-lac structure. The Crowther fast rotation function procedure (Crowther, 1972) yielded one significant peak at [α=95.0°, β=90.0°, γ=85.0°], 1.7σ above any other peak in the map and 5.0σ above the mean of the map. No other peak in the map was greater than 65 % of the maximum peak. The Lattman rotation function (Lattman & Love, 1970) gave refined rotation angles of [α=95.0°, β=88.0°,γ=85.0°]. These angles from the rotation search were then used in the translation function (Crowther & Blow, 1967). The rotation function solution with the highest correlation peak gave a single, self consistent set of translation vectors on the Harker sections. The molecular replacement solution corresponds to a slight movement of the molecule in the cell, as will be discussed below.

The molecular replacement model coordinates were fed into the constrained- re-

strained least squares reciprocal space refinement program, CORELS (Sussman, 1985). In the refinement, reflections from 8.0 to 3.0 Å were used. 30 cycles of CORELS refinement lowered the R-factor from 0.44 to 0.40. The resultant structure was further refined using the refinement program X-PLOR (Brünger, 1989). Ten rounds of atomic positional refinement (initially with an overall B-factor and during later stages with individual B-factors) were performed. Initial refinement utilised only energy- crystallographic least squares minimisation and was similar in strategy to the procedure of Jack & Levitt (1978), but with the energy parameters of X-PLOR. For individual atomic B-factor refinement, the target standard deviations for bonded atoms and for atoms linked by one other atom were 1.0 and 1.5 and the actual values for the final model were 1.7 and 2.3 respectively. Following convergence, difference electron density maps with amplitude coefficients $2/F_o/-/F_c/$ and $/F_o/-/F_c/$ were used as guides for manual changes in the model. A bound calcium ion was identified at this stage as a strong feature in the electron density map and was included in the refinement. After each round of refinement, model checking was done using the program FRODO. The final refined model consists of the protein with 90 solvent molecules and one calcium ion. The refinement converged with an R-factor of 0.209 for 11,373 data [all the observed data for which $F_o \geq 2\sigma(F_o)$] in the resolution range 8-1.7 Å. The estimated average positional standard deviation will be about 0.15 Å , based on the statistical method of Luzzati (1952). The model has r.m.s. deviation from ideality of 0.013 Å in bond lengths and 2.9° in bond angle.

The overall structural features of human α-lac are similar to those of baboon α-lac (Acharya *et al.*, 1991). If the two molecules are expressed in equivalent crystallographic coordinate systems, the root-mean-square deviation between the two structures for all main chain atoms is 1.8 Å. The human α-lac structure may then be superimposed upon baboon α-lac by a 4° rotation and 0.9 Å translation along this rotation axis. The effect of this is to produce shifts of up to 3 Å in the X and Z directions of the unit cell which presumably accounts for the failure in the initial rigid body refinement. At this point the average deviation between the two structures for all main chain atoms is 0.4 Å and the r.m.s. deviation 0.67 Å. The solvent structure also has similarities, with several water molecules common to both structures (distance less than 1.0 Å).

From the above discussion it is clear that one could effectively use Molecular Replacement method in determining the 3D structures of homologous molecules and the hope is that it may be possible to derive the 3D structures of all the members of a family relatively easily once one or two have been analysed experimentally.

Acknowledgements

I would like to thank Dr. David Stuart (Laboratory of Molecular Biophysics, University of Oxford) and Dr. Nigel Walker (BASF AG, Germany) for their contribution to the α-lac project.

References

Acharya,K.R. *et al.*, (1991). J.Mol.Biol. **221**,571-581.

Acharya,K.R. *et al.*, (1989). J.Mol.Biol. **208**,99-127.

Aschaffenburg,R. *et al.*, (1979). J. Mol. Biol. **127**, 135-137.

Berliner,L.J. & Johnson,J.D. (1988) in: 'Calcium binding proteins-biological functions (Thompson,M.P. Ed.) vol. **2**, pp 79-116, CRC Press, Boca Raton, USA.

Brew,K. *et al.*, (1970). J.Biol.Chem. **245**,4570-4582.

Brünger,A.T. (1988). J.Mol.Biol. **203**,803-816.

Brünger,A.T. (1989). Acta Crystallogr. **A45**,42-50.

Brünger,A.T. *et al.*, (1987). Science, **235**,458-460.

Brünger,A.T. *et al.*, (1989). Acta Crystallogr. **A45**,50-61.

Crowther,R.A. (1972). in: 'The Molecular Replacement Method' (Rossmann,M.G. Ed.) The fast rotation function. pp 174-178. Gordon & Breach, New York, USA.

Crowther,R.A. & Blow,D.M. (1967). Acta Crystallogr. **23**,544-548.

Dandekar,A.M. & Qasba,P.K. (1981). Proc.Natl.Acad.Sci. (USA). **78**,4853-4857.

Fenna,R.E. (1982). J.Mol.Biol. **161**,211-215.

Findlay,J.B.C. & Brew,K. (1972). Eur.J.Biochem. **27**,65-86.

Fitzgerald,P.M.D. (1988). J.Appl.Crystallogr. **21**,274-278.

Fraser,R.D.B. *et al.*, (1978). J.Appl.Cryst. **11**,693-694.

Hall,L. *et al.*, (1982). Nucleic Acids Res. **10**,3503-3515.

Hamano,M. *et al.*, (1986). J.Biochem. (Tokyo) **100**,1617-1622.

Hendrickson,W.A. & Konnert,J.H. (1980). In computing in Crystallography (Diamond,R. *et al.*, Eds), pp 13.01-13.23, Indian Academy of Sciences, Bangalore, India.

Hill,R.L. & Brew,K. (1975). Adv.Enzymol.Relat.Areas Mol.Biol. **43**,411-490.

Hiraoka,Y. *et al.*, (1980). Biochem. Biophys.Res.Commun. **95**,1098-1104.

Jack,A. & Levitt,M. (1978). Acta Crystallogr. **A34**,931-935.

Kronman,M.J. (1989). CRC Crit.Rev. in Biochem. and Mol.Biol. **24**,565-667.

Lattman,E.E. & Love,W.E. (1970). Acta Crystallogr. **B26**,1854-1857.

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

Qasba,P.K. & Safaya,S.K. (1984). Nature(London) **308**,377-380.

Segawa,T. & Sugai,S. (1983). J.Biochem. (Tokyo). **93**,1321-1328.

Smith,S.G. *et al.*, (1987). Biochem. J. **242**, 353-360.

Stuart,D.I. *et al.*, (1986). Nature(London) **324**,84-87.

Sussman,J.L. (1985). In methods in Enzymology (Wyckoff,H.W. *et al.*, Eds). **115**, pp 271-303, Academic Press, Florida, USA.

Molecular Replacement Studies of a Ternary Complex of an Allosteric Lactate Dehydrogenase from *Bacillus stearothermophilus*

Dale B. Wigley.
Dept. of Chemistry, York University, York YO1 5DD, U.K.

Introduction

The structure determination of a ternary complex of *B.stearothermophilus* L-lactate dehydrogenase (1) is an example of a complex crystal form containing eight molecules in the asymmetric unit. Because these eight subunits are in a special arrangement (i.e. two tetramers) the problem is simplified somewhat, though untangling the local symmetry remains a complicated business, particularly due to the symmetry of the packing of the tetramers which resembles that of a higher symmetry spacegroup. This example also illustrates the use of some of the special features of the MERLOT suite of programs (2) which are not available in many other molecular replacement programs.

Crystal Forms

At least five different crystal forms were grown under similar conditions from polyethylene glycol 6000, of which three were subjected to X-ray analysis (Table 1).

Table 1 - *B.stearothermophilus* LDH Crystals

Spacegroup	Cell Dimensions	Vm (Å^3/Da)	Subunits per asymm. unit	Diffraction Limit (Å)
orthorhombic $P2_12_12$	a = 86 Å b = 105 Å c = 136 Å	2.27	4 (135,000 Da)	3.0
monoclinic $P2_1$	a = 112 Å b = 85 Å c = 136 Å ß = 91°	2.40	8 (270,000 Da)	1.8
monoclinic $P2_1$	a = 84.9 Å b = 118.2 Å c = 135.5 Å ß = 96.07°	2.52	8 (270,000 Da)	1.8

Although more suitable from a crystallographic viewpoint, the orthorhombic crystals did not diffract as strongly as other crystal forms. Of the two monoclinic forms, the type V crystals were

only rarely obtained, so the type IV form was chosen for analysis. It is interesting to note the similarity in cell dimensions between the three crystal forms. The *h0l* zones of the monoclinic crystal forms also exhibit unexpected pseudo mm symmetry, and several other features of these photographs are worthy of comment:

1) the striking similarity between equivalent zones of different crystal forms and, to a lesser extent, between different zones of the same crystal form,

2) the tendency for 00*l* reflexions of even index to be stronger than those where *l* is not equal to 2*n* (suggesting an approximate 2_1 screw along *c*),

3) the non-primitive appearance of the reciprocal lattice as viewed down the *a* axis ($k + l = 2n$, strong) in the type IV crystals particularly for reflexions of low *l* index. This is indicative of a displacement of a half in *b* and in *c* between two of the molecules in the cell. The separation on *b* is apparently closer to a half than is the displacement along *c*, as the relationship, $k + l = 2n$, is maintained at higher values of *k* than of *l*.

These features of the low resolution diffraction patterns proved to be important hints which aided interpretation of the results of the rotation and translation functions.

Data collection

Two data sets were collected, both at the Synchrotron Radiation Source at Daresbury, U.K. The first was collected from a single crystal using the 0.88Å wavelength radiation available on the Wiggler station. The data were collected on film to 4.7Å resolution. The films were deliberately under-exposed to ensure that most of the strongest reflexions did not exceed the dynamic range of the film and thus were recorded accurately, improving signal to noise in the rotation and translation functions (both of which are based on Patterson functions, and are strongly dependent upon the reflexions of high intensity).

The second data set (to 2.5Å) was collected using 1.488Å wavelength and comprised data from three crystals.

Table 2 - Data analysis (4.7Å)

No. of photographs	18
No. of crystals	1
Resolution (Å)	50 - 4.7
R_{merge} (all data)	7.5 %
(data > I_{mean})	3.5 %
No. of measurements	24,036
No. of independent reflexions	12,874
Percentage possible data	91 %

Molecular packing

In order to facilitate the reader's understanding of this rather complex structure determination, the molecular packing is presented in figure 1. The rotation and translation function studies which support this model are presented in subsequent sections.

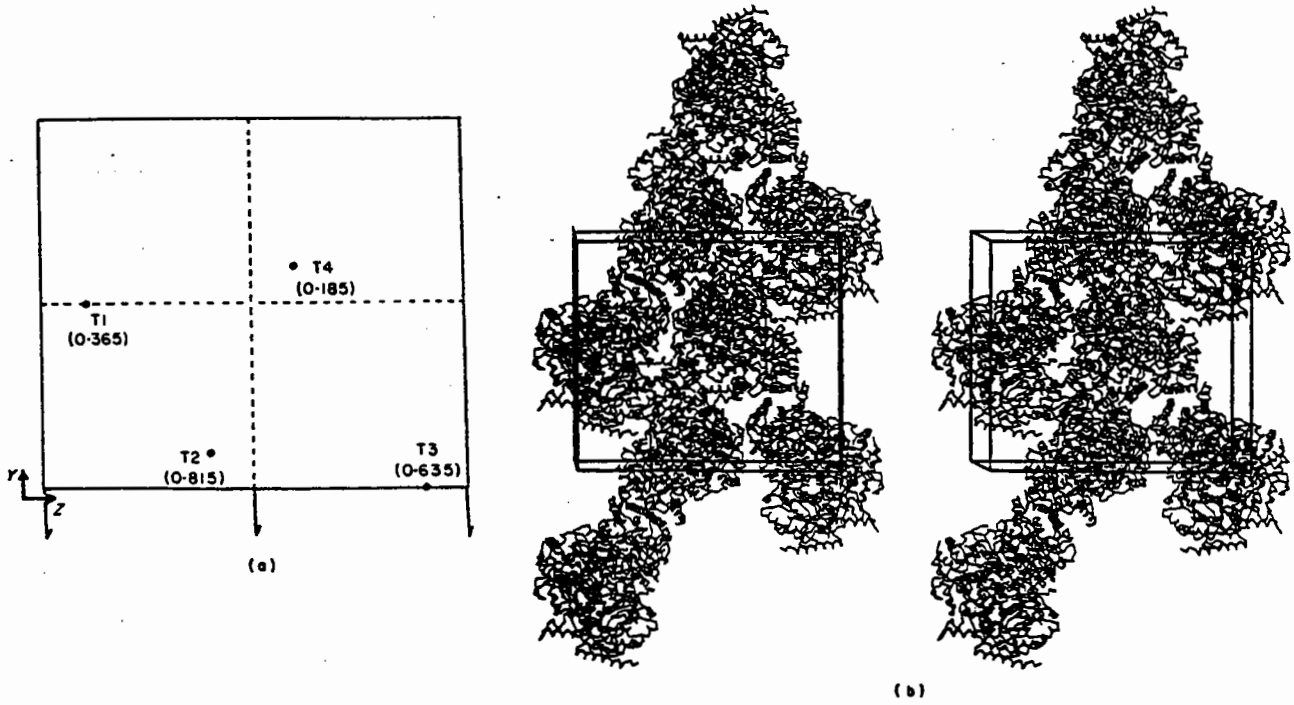


Figure 1 - The molecular packing in the type IV crystals, a) schematic diagram indicating the positions of the four LDH tetramers in the unit cell (labelled T1 to T4; the asymmetric unit comprises T1 and T2). Figures in parentheses refer to the relative heights of the molecules along the x axis, b) diagram illustrating the molecular orientations. The view is approximately the same as in a).

Molecular Orientations

Determination of the molecular orientation and positions was achieved using the 4.7Å data, which contained most of the very strong reflexions. Because the LDH tetramer shows 222 molecular symmetry, inspection of the $k = 180^{o}$ section of a self-rotation function map should reveal the positions of these molecular two-fold axes. With two tetramers in the asymmetric unit, six peaks corresponding to the three 2-folds of each tetramer were expected. These peaks should, of course, fall into two mutually orthogonal sets. In addition, it might be expected that certain other peaks in the self-rotation function would relate one tetramer to the other of the same crystallographic asymmetric unit. Such peaks need not result from rotations of 180°, but in view of the pseudo-orthorhombic symmetry of the crystals we might expect them to lie close to the $k = 180^{o}$ section, with ϕ close to 0° or 90° and ψ close to 90°.

Orthogonal axes (X, Y, Z) were defined with the crystal axis *c* along X, *a** along Y, and *b* along Z. All rotation function results are expressed relative to this axial system. A self-rotation function was calculated (using POLARRFN from the CCP4 suite) using the data with amplitudes greater than the mean, between 15 and 6Å spacing. A radius of integration of 25Å was chosen. The rotation function was sampled at 5° intervals on ϕ and ψ, while k was held at 180°. With two tetramers in the asymmetric unit in spacegroup $P2_1$, the theoretical height of any peak corresponding to an intramolecular two-fold axis should be 25% of that of the origin. Hence a rotation function was calculated and the map produced was contoured at every 5% from 15% (figure 2).

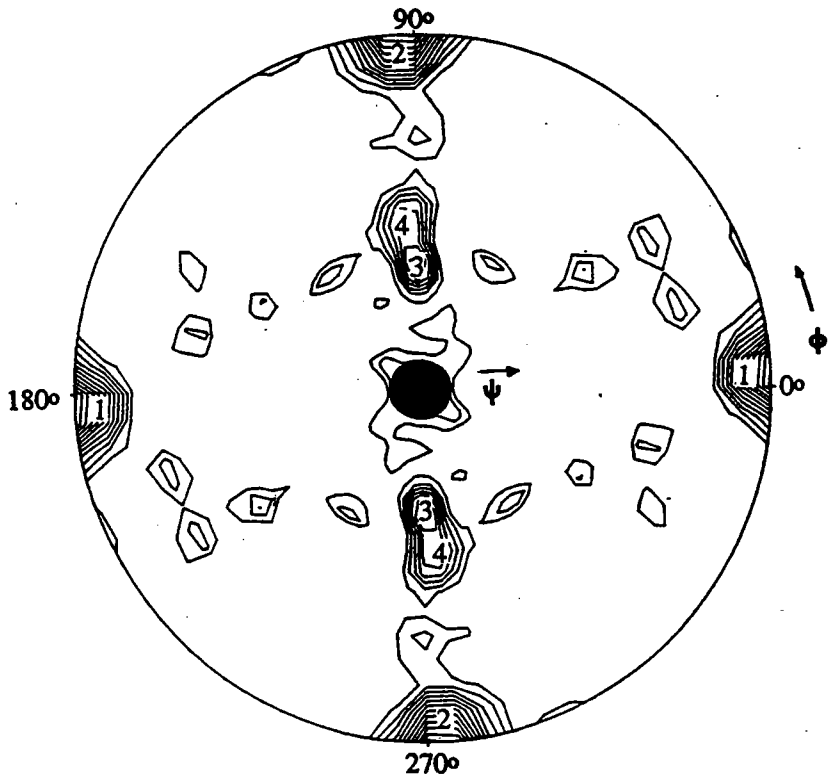


Figure 2. A 15–6 Å self-rotation function. Stereographic projection of rotations with $\kappa = 180°$

Several features were immediately apparent, the most obvious being the very large peaks at $\psi =$ 90°, $\phi = 5°$ and, by symmetry, 95° (peaks 1 and 2, height = 71% of origin). There were also two large peaks at $\psi = 40°$, $\phi = 90°$(peak 3, 50% of origin) and at $\psi = 50°$, $\phi = 95°$ (peak 4, 42% of origin). The heights of all four of these peaks were around twice that expected for a single molecular dyad axis. The pseudo-orthorhombic symmetry of the crystals is also evident in this self-rotation function, such that there is an approximate mirror plane on the $k = 180°$ section at $\phi \sim$ 90°. This symmetry is not expected for monoclinic spacegroups. This observation provides further strong evidence that at least the orientations of the two molecules of the asymmetric unit are related by approximate orthorhombic symmetry. This self-rotation function could be interpreted such that the two-fold axes of the two tetramers are in very similar positions, either placing the two molecules in very similar orientations, or the molecular two-folds coincident but different in each case. In either situation, self-rotation functions should reveal peaks which place one tetramer onto the other of the asymmetric unit. In the first instance, these rotations will be coincident with the molecular two-folds (or their symmetry equivalents), so cannot be distinguished from them. In the latter case, there must be rotations placing one tetramer in a given orientation onto another in a different orientation. Such rotations are not constrained to lie on the $k = 180°$section. Hence a search on k from 0 - 180° was conducted to search for any such peaks. The only significant peaks found in the rest of the self-rotation function were symmetry equivalents of peaks 1-4 on the $k =$ 180° section. These results suggested the two tetramers of the asymmetric unit to be in very similar orientations, an interpretation which seemed to be supported by cross-rotation function studies using the unrefined pig H_4 NAD-S-lactate coordinates from the Brookhaven databank. However, if this were indeed the case, then a native Patterson function should reveal a large peak corresponding to the translation vector between the two non-crystallographically related tetramers. No significant peak was found, even at very low resolution (25-15Å data).

The solution of the structure of the binary (enzyme/NADH/$FruP_2$) complex of *B.stearothermophilus* LDH (3) gave us access to a partly-refined (R-factor = 28%) set of coordinates. A model was constructed, based on the binary structure but in which the active site loop (which was not visible in the binary structure) was modelled in the "down" position observed in other ternary complexes of LDH. Structure factors were calculated for this *B.stearothermophilus* "binary" molecule placed in a large (160 x 160 x 160Å) P222 cell. These data were then used in a

cross-rotation function using the MERLOT suite of programs (2). The fast rotation function was employed using all of the data between 8 and 4.7Å spacing with $F/\sigma > 3$. The terms were modified for removal of the origin, and a radius of integration of 25Å was chosen. The whole of the asymmetric unit of rotation space was explored at 2.5° on alpha and 5° on beta and gamma. The known model was oriented with its molecular P, Q, and R axes aligned initially along X, Y, and Z, while the ternary crystal was oriented to place c, $a^* \cdot b$ along X, Y, and Z. The rotation function produced two very significant peaks:

Table 3 - Cross-rotation peaks (8 - 4.7 Å)

Peak	α	β	γ	Height	RMS
1	167.5	70.0	40.0	100.0	8.96
2	17.5	105.0	45.0	75.4	6.76
3*	167.5	70.0	185.0	48.3	4.32

* Height of the highest noise peak is shown for comparison.

The ROTSYM option within MERLOT was used to determine the relationship between these two peaks (and their symmetry equivalents). This analysis revealed two sets of peaks; those related to others by exact two-folds, and those related by approximate two-folds. This latter set again confirmed the pseudo-orthorhombic nature of the crystals.

Table 4 - Symmetry relationships between cross-rotation peaks

Peak	α	β	γ	
1A	167.5	70.0	40.0	Subunits 1-4 of tetramer 1 (T1)
1B	167.5	70.0	220.0	
1C	347.5	110.0	140.0	
1D	347.5	110.0	320.0	
2A	17.5	105.0	45.0	Subunits 1-4 of tetramer 2 (T2)
2B	17.5	105.0	225.0	
2C	197.5	75.0	135.0	
2D	197.5	75.0	315.0	

Table 5 - Relationship between the cross rotation and self rotation peaks

Rotation	φ	γ	κ	
1A - 1B	168	70	180	Subunit 1 onto the other 3 subunits of T1 (i.e. 2-folds of T1)
1A - 1C	274	53	180	
1A - 1D	55	44	180	
2A - 2B	198	75	180	Subunit 1 onto the other 3 subunits of T2 (i.e. 2-folds of T2)
2A - 2C	93	47	180	
2A - 2D	302	47	180	
1A - 2A	355	18	151	Subunit 1 of T1 onto subunits 1 - 4 of T2
1A - 2B	92	87	184	
1A - 2C	185	61	73	
1A - 2D	180	90	80	

These results prompted a re-examination of the self-rotation function calculated with data between 8 and 4.7Å resolution (figure 3).

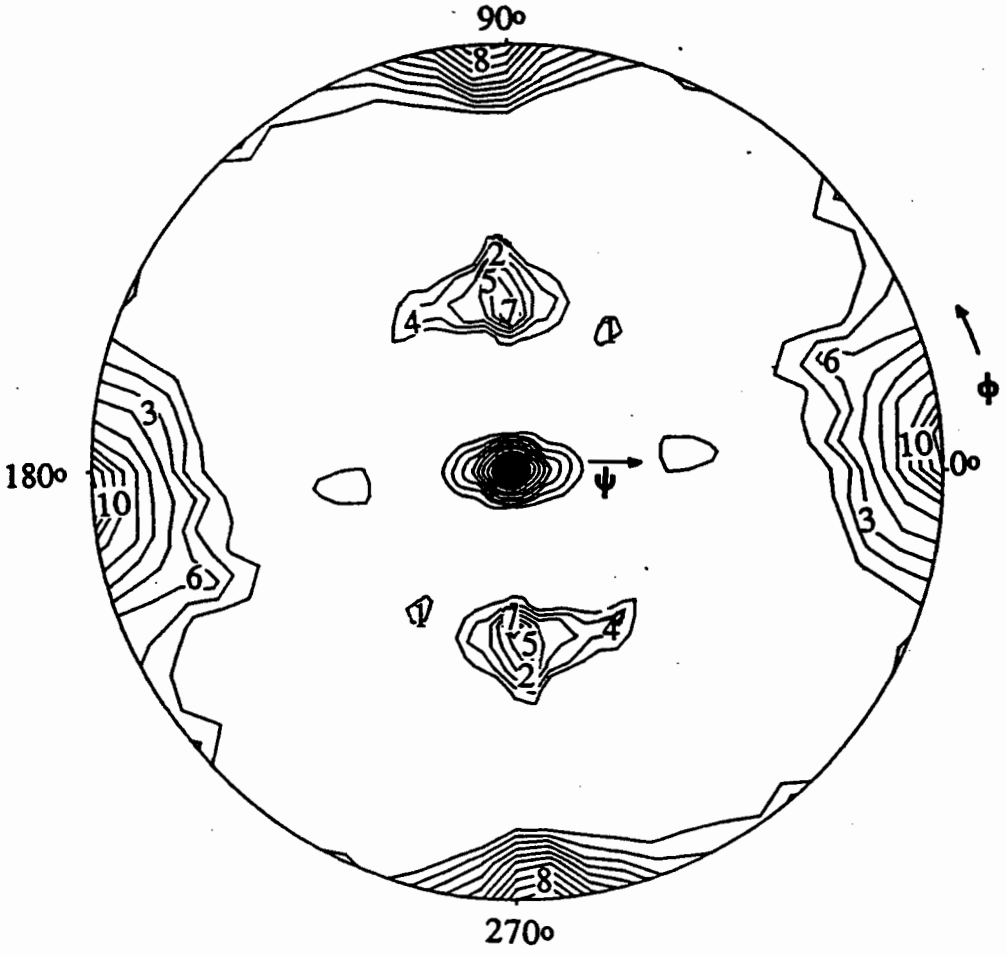


Figure 3. A 8–4·7 Å self-rotation function. Stereographic projection of rotations with $\kappa = 180°$

Originally these self-rotation results were discarded because the $k = 180°$ section seemed to show essentially the same features as the function using 15-6Å data, but the peaks were less sharp making interpretation rather difficult. Interpreted in the light of the new cross-rotation results, based on a set of refined coordinates, it was possible to identify a set of peaks consistent with these results (Table 5). From the set of peaks which were related by exact two-folds, it was possible to extract two orthogonal sets - each corresponding to an individual tetramer. In addition, those rotations which were not exact two-folds corresponded to rotations which placed one orthogonal set onto another. In this way, all of the peaks observed in the self-rotation function could be accounted for (Table 6). It appears that the lower resolution terms were dominated by the intermolecular dyad axes, swamping those from the intramolecular two-folds. The dominance of such "Klug peaks" has been noted in other rotation function studies (4).

Table 6 - Self-rotation peaks (8 - 4.7 Å)

Peak	ϕ	ψ	κ	Height	Local Symmetry	
1	55	44	180	20	-P-axis	T1
2	274	53	180	15	Q-axis	T1
3	168	70	180	22	R-axis	T1
4	302	47	180	38	-P-axis	T2
5	93	47	180	27	Q-axis	T2
6	198	75	180	22	-R-axis	T2
7	92	40	180	47	Peaks corresponding to rotations which place at least one tetramer onto another	
8	93	87	178	65		
9	175	19	150	20		
10	3	89	174	65		
11	95	50	132	20		
12	2	90	80	47		
13	85	50	49	19		
14	5	60	108	21		

A more accurate determination of the orientation of the two independent tetramers was obtained by Lattman's rotation function as implemented in MERLOT. A fine search was conducted around the peak positions on a 0.5°grid. It is important to note that the relative heights of the two peaks were more similar at this stage (Table 7), showing the small discrepancy in the peak heights obtained from the fast rotation function to be a consequence of the coarse sampling, rather than of real differences between the two tetramers.

Cross-rotation function studies were performed using refined dogfish M_4 NADH/oxamate LDH coordinates (Brookhaven databank) and with P, Q, and R-axis dimers of *B.stearothermophilus* binary complex LDH. All results confirmed those described above with the P222 (ie whole tetramer) structure factor set.

Table 7 - Cross-rotation peaks (fine search)

Peak	α	β	γ	Absolute Value	Relative Height
1	167.5	68.5	38.0	0.277×10^{15}	99.3
2	19.5	106.5	45.5	0.279×10^{15}	100.0

Molecular Positions

Of the rotation and translation functions, it is frequently the latter which is the most difficult. This is particularly true where there are two independent molecules in the asymmetric unit. For the $P2_1$ case, this results in a five dimensional search problem (the Y coordinate of one molecule fixes the origin along the unique axis). Searches based on R-factor minimisation procedures, in this particular case were shown to be unsuccessful even with the benefit of hindsight. The translation function of Crowther and Blow, on the other hand (5), uses a modified Patterson function to look for cross vectors between different molecules in the unit cell. This approach is eminently more suitable for the case where there are two crystallographically independent molecules. Hence, this

translation function was employed (available as TRNSUM within MERLOT). A variety of searches were carried out to look for vectors between both crystallographically related, and independent molecules. Searches between crystallographically related molecules could be restricted to the section where Y = 0.5, whereas those between independent tetramers were made over the entire unit cell. All searches employed a grid sampling of 1/100th of a cell edge. The results were very convincing (Table 8), and could be interpreted to define the molecular packing (figure 1).

Table 8 - Translation function

Search	X	Y	Z	RMS	Height
T3 -> T1	0.73	0.50	0.21	9.43	100
	0.40	0.50	0.81	3.49	37*
T4 -> T2	0.63	0.50	0.79	8.99	100
	0.93	0.50	0.10	3.62	40*
T2 -> T1	0.55	0.40	0.71	11.59	100
	0.28	0.01	0.14	4.72	41*
T3 -> T2	0.18	0.10	0.50	12.68	100
	0.57	0.77	0.25	4.73	37*

*The height of the highest noise peak for each search is given.
Calculations used data with $F > 3\sigma$ in resolution range 8 - 5.3 Å.

The molecules were correctly oriented, placed at their known positions in the cell, and structure factors calculated to 4.7Å. At this stage, the R-factor between the binary LDH model structure factors and the ternary LDH data was 44%. The molecular positions and orientations were refined by a least squares procedure (RMINIM in MERLOT) in 0.25° intervals on alpha, beta, and gamma and by 0.1Å on X, Y, and Z. This procedure resulted in very small overall shifts, but reduced the R-factor to 39%, showing the *B.stearothermophilus* binary structure to be a good model for the ternary complex, as might be expected.

Refinement

Electron density (2Fo-Fc, 2.5Å resolution) maps calculated at this stage were generally of good quality. Side chain densities were clearly distinguishable from that of the main chain (though atoms were not always in density at this stage), and density attributable to the NADH, oxamate, and $FruP_2$ was observed even though these molecules had been omitted from the calculations. Consequently, the maps did not appear to be biased by the starting model.

With over 20,000 non-hydrogen atoms in the asymmetric unit, the refinement problem was not trivial and required several approaches. Manual model building (which was potentially very time consuming for such a large asymmetric unit) was kept to a minimum and until the final stage of the refinement was only carried out upon one of the non-crystallographically related subunits, with the other subunits being produced by rotation and translation of the rebuilt subunit to the other positions.

The first stage involved a conventional positional refinement using the X-PLOR package (6). Strict 222 molecular symmetry was imposed on both tetramers at this stage of the refinement, and an initial overall B-value of 15.0 $Å^2$ was applied. Cycles of X-ray restrained energy minimisation refinement were repeated until convergence had been achieved. After a few cycles of individual B-

factor refinement the R-factor had fallen to 28.8%. At this stage the bound NADH and oxamate were included in the model. After a few cycles of least squares refinement, the refinement of the bound ligands had converged. However, the R-factor at this stage was still 24.4%, so refinement was continued using the simulated annealing mode of X-plor with non-crystallographic symmetry constraints maintained. Solvent molecules were also fitted at this stage. Further refinement employed a conventional least-squares minimisation procedure but maintained strict non-crystallographic symmetry restraints on main chain atoms, but less strict for side chains. The R-factor at this stage was 18.2%. The final stage of the refinement allowed a release of the non-crystallographic symmetry restraints, and required a further round of model building. The bound $FruP_2$ molecules were also included at this stage. Because of the problem of statistical disorder, two $FruP_2$ molecules were included at each site but the occupancies were set to 0.5. The R-factor of the final model using all data between 10 and 2.5Å resolution was 14.7%.

Molecular packing and pseudo-symmetry

Any proposed packing model must agree with the strong pseudo-symmetry features of the precession photographs described earlier. The approximate 2_1 screw along *c* is a consequence of the relationship of molecules 1 to 4, and of 2 to 3. At low resolution, the orientations of these pairs of molecules is very similar but more importantly, the translation between them is exactly a half on Z. The approximate screw axis along *c* only breaks down at higher resolution when the difference in orientation between the two molecules of a given pair can be distinguished. The approximate centring of the lattice, when viewed in projection down *a*, gives rise to the appearance of the non-primitive lattice effect observed in the 0*kl* zone. As predicted, the displacement of those molecules most closely related in X (eg T1 and T2) is closer to a half in the Y direction than along Z (Y = 0.5 $\pm$ 0.1, Z = 0.5 $\pm$ 0.2).

In view of the similarity between the precession photographs of the orthorhombic crystals and both of the monoclinic crystal forms, it seems likely that the molecular packing is similar in all three crystals. The observed pseudo-orthorhombic nature of the packing in the type IV crystals would support this, and it would be easy to obtain true orthorhombic symmetry by small changes in the molecular packing. Similarly it would be easy to obtain the type V monoclinic packing from the type IV by very small changes which would alter the definition of the unique axis, since there is an approximate screw axis along both *a* and *b*.

References

1) Wigley, D.B., Gamblin, S.J., Turkenburg, J.P., Dodson, E.J., Piontek, K., Muirhead, H., and Holbrook, J.J. (1992). The structure of a ternary complex of an allosteric lactate dehydrogenase from *Bacillus stearothermophilus* at 2.5Å resolution. *J. Mol. Biol.* **223**, 317-335.

2) Fitzgerald, P.M. (1988). MERLOT, an integrated package of computer programs for the determination of crystal structures by molecular replacement. *J. Appl. Crystallogr.* **21**, 274-8.

3) Johnson, J.E., Argos, P., and Rossmann, M.G. (1975). Rotation function studies of southern bean mosaic virus at 22Å resolution. *Acta Crystallogr.* **B31**, 2577-83.

4) Piontek, K., Chakrabarti, P., Schar, H-P., Rossmann, M.G., and Zuber, H. (1990). Structure determination and refinement of *Bacillus stearothermophilus* lactate dehydrogenase. *Proteins: Structure, Function, and Genetics* **7**, 74-92.

5) Crowther, R.A. and Blow, D.M. (1967). A method of positioning a known molecule in an unknown crystal structure. *Acta Crystallogr.* **23**, 544-8.

6) Brunger, A.T., Karplus, M., and Petsko, G.A. (1989). Crystallographic refinement by simulated annealing : application to crambin. *Acta Crystallogr.* **A45**, 50-61.

Some Applications of the Phased Translation Function using Calculated phases

G.A.Bentley, Unité d'Immunologie Structurale, Institut Pasteur,25 rue du Dr. Roux, Paris.

Introduction

The solution of the translation problem in molecular replacement may often require placing one molecule only in the asymmetric unit. It is not unusual, however, to be faced with the situation of placing more than one molecule or molecular fragment, each with respect to a common crystallographic origin. The correct choice of space group enantiomorph, which requires identification of the hand of a screw axis, is another problem to be resolved by a translation function. We discuss here the use of the Phased Translation Function (PTF) in each of these situations.

The PTF may be viewed as an image-seeking function. It requires two elements: an electron density map of the unknown structure, ρ_{obs}, (calculated from phases which may be derived by isomorphous replacement or calculated from partial structures) and a model for the unknown molecule or molecular fragment correctly oriented but arbitrarily placed in the unit cell (ρ_{model}). The electron density calculated from the oriented model is systematically moved over the electron density of the unknown structure (figure 1) and at each grid point, the product of the two electron density functions is taken:

$$S(t) = \int_V \rho_{obs}(x)\rho_{model}(x-t)dx \qquad (1)$$

where ρ_{obs} and ρ_{model} are the densities of the unknown crystal structure and the model structure, respectively.

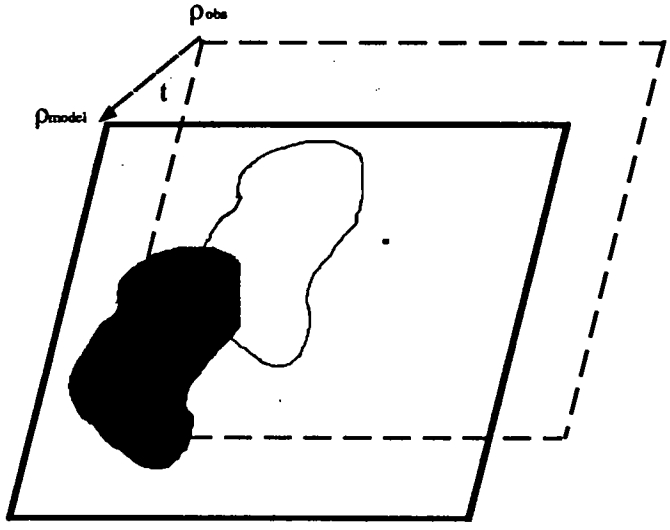


Figure 1

The principle of the phased translation function.

The position where the function has its maximum corresponds to the optimum superposition of the two electron densities and thus gives the translation required to place the oriented model correctly in the unit cell. While the calculation can be done in direct space, it may be more efficiently carried out in reciprocal space as a Fourier summation:

$$S(\mathbf{t}) = 1/V \, \Sigma_{\mathbf{h}} \, \mathbf{F}_{obs}(\mathbf{h})\mathbf{F}_{model}^{*}(\mathbf{h}) \exp(-2\pi i \mathbf{h}.\mathbf{t}) \qquad (2)$$

where $\mathbf{F}_{obs}(\mathbf{h})$ and $\mathbf{F}_{model}(\mathbf{h})$ are the structure factors of ρ_{obs} and ρ_{model} respectively. Since ρ_{model} contains one orientation only (i.e., has symmetry P1), the volume of integration, V, is the complete unit cell. The PTF has been described elsewhere, both for the case of isomorphous phases[1,2] and calculated phases from partial structures[3,4,5,6,12]. Here, the discussion will be restricted to the use of calculated phases.

(a) Finding a common origin for independent components of the asymmetric unit

Although this represents an intermediate and not the first step in solving the translation problem, we begin with this case since it will serve as a simpler introduction to the PTF. It is probably here, in fact, that the PTF is most useful. We shall use a complex formed between two Fab fragments, FabD1.3-FabE225 [7], to illustrate the application of the function. An Fab fragment is cleaved enzymatically from an immunoglobulin molecule and consists of two polypetide chains termed light (L) and heavy (H). Each chain folds into two separate domains which, due to sequence variabilty or conservation, are denoted as variable (V) and constant (C) respectively. In the quaternary structure of the Fab, the variable domains, V_H and V_L, and the constant domains, C_H1 and C_L, each associate as dimers in a way that generally varies little from one species of Fab to another. Since the variable and constant domains of each chain are connected by a flexible peptide link, the relative disposition between the variable and constant dimers, defined by the *elbow angle*, can differ widely between different Fabs. The search model for an Fab must therefore be divided into two independent parts: the variable dimer and the constant dimer. Consequently, the solution of FabD1.3-FabE225 complex by molecular replacement requires placing four independent components, i.e., two variable and two constant dimers, with respect to the same origin. These were first separately oriented and placed correctly with respect to a permitted crystallographic origin of the unit cell.

The space group of the Fab-Fab complex is $P2_1$; thus while the choice for both x and z is 0 or 1/2, y may take an arbitrary value. The task that remained was to place all four components with respect to a common origin. One method of achieving this is to apply the PTF using calculated phases from a partial structure. The partial structure in this example can be any one of the dimers which has been correctly oriented and placed in the unit cell. An electron density map calculated using phases derived from this component will thus contain weak density from the remaining structure. Since we calculate the PTF in reciprocal space we expand the observed

structure amplitudes phased by this component by symmetry to P1. Since the contribution from the phasing model would dominate the PTF, it must be removed either by using difference coefficients (Fobs(**h**)- Fcalc(**h**))exp[iα_{calc}] in place of $\mathbf{F}_{obs}(\mathbf{h})$ in equation (2), or by removing the electron density of the phasing model in an Fobs, αcalc Fourier synthesis by means of a molecular mask and then using the structure factors obtained by the inverse Fourier transform of the modified map. The position of any of the other components may then be found by using each correctly oriented component in turn as a search model. Structure factors from this second component are calculated in P1 by placing it in an arbitrary position in the unit cell; these serve as $\mathbf{F}_{model}(\mathbf{h})$ in equation (2). It is not obligatory, of course, that the search models be placed initially with respect to a possible crystallographic origin, but doing so can provide a useful check that a consistent result for a common origin has been obtained. A flow diagram indicating the necessary steps is shown in figure 2 for the particular case where the contribution from the phasing component is removed by applying a mask to the electron density.

The complex FabD1.3-FabE225 was solved by molecular replacement by first determining the orientation (using *ROTFUN* [8]) and translation (using *TFSGEN* [9]) of the two variable and the two constant dimers of the complex as separate components. Table 1 shows the results of the PTF in placing all four components relative to a common origin. Here, each correctly oriented and positioned dimer was used in turn to phase while the other three dimers were used successively as search models. Columns (a), (b) and (d) show results obtained by masking the electron denisty of the phasing model (figure 2) while column (c) gives results for the case where difference coefficients were used. Comparison of columns (a) and (b) shows clearly the advantage of using high resolution data. Although there is little difference between the use of difference coefficients and molecular masks at high resolution (columns (b) and (c)), there is a defintite advantage in using the latter if the calculation is made at low resolution (data not shown). Column (d) tabulates results obtained from the refined atomic coordinates for comparison. The number of dimer pairs giving a clear maximum is sufficient to give a clear solution as well as providing a generous cross-check for its consistency.

(b) Search for symmetry elements using the phased translation function

Here we discuss how the PTF can be used to place an oriented model correctly in the unit cell with respect to a crystallographic origin. (Note that in the previous section, the crystallographic origin had been already defined by first placing the phasing component.) The procedure described here follows closely that given by Doesburg and Beurskins[3]. To start with, the oriented model is placed at an arbitrary position in the unit cell and these coordinates alone are used to calculate phases (i.e., we do not use the space group symmetry) for the structure amplitudes that have been expanded by symmetry to P1 to serve as $\mathbf{F}_{obs}(\mathbf{h})$ in equation (2).

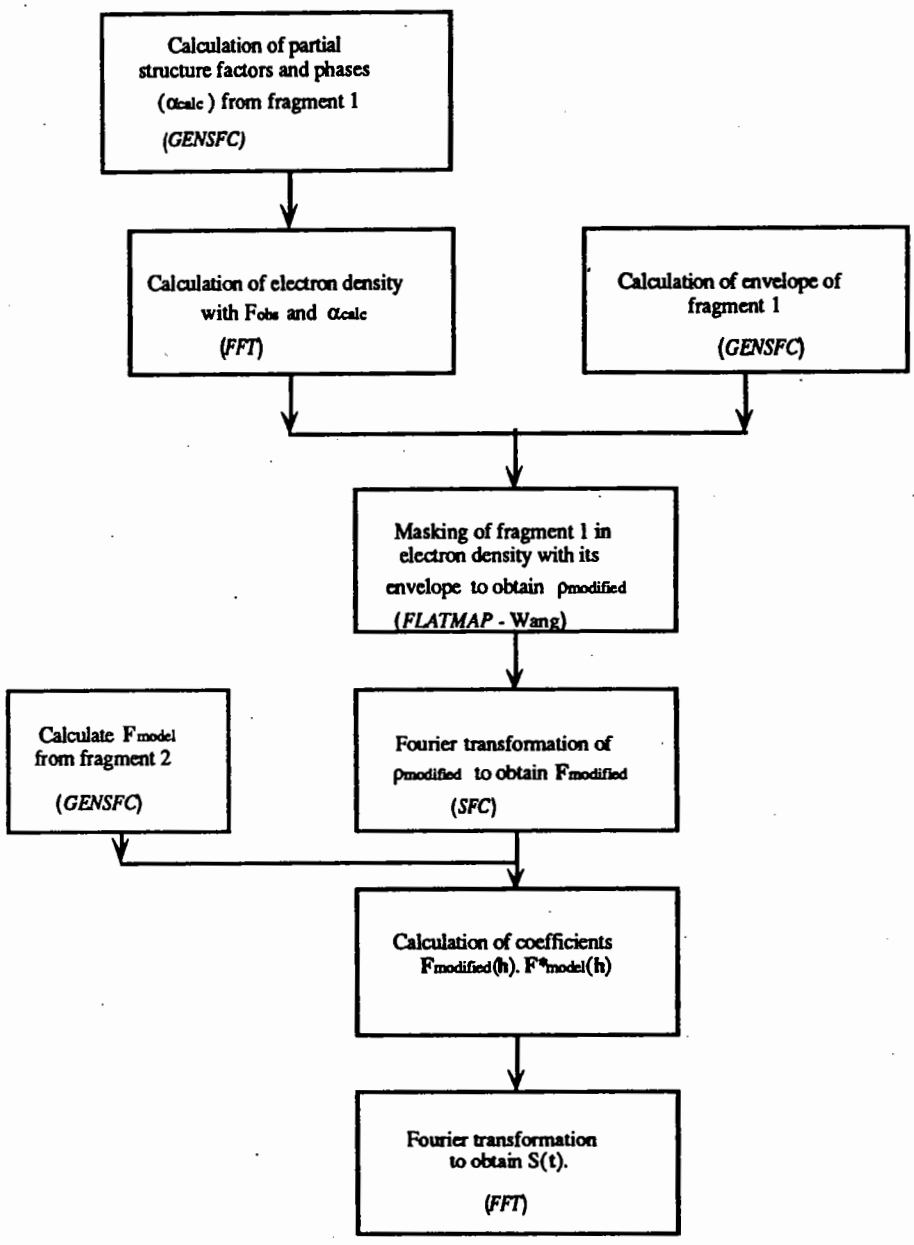


Figure 2

Flow diagram for PTF application using molecular masks to remove the phasing density. The programs used (CCP4 [11]) are given in parenthesis. The mask was generated by creating an electron density map from the atomic coordinates of the phasing model using the program *GENSFC*; every point where the electron denity was greater than zero was taken as being included inside the mask. Tests showed this way of masking the phasing model density to be perfectly adequate even though the procedure did not remove it entirely.The program *FLATMAP* was modified to supress the denisty inside rather than outside the envelope. A small program was added to form the product of the structure factors for the PTF (equation (2)) at the second to last stage.

Fourier maps calculated from these structure factors will have electron density for the symmetry-related molecules as well as the phasing molecule, but with a much lower signal than the latter. If the symmetry-related molecules can be located, the positions of the corresponding symmetry elements can be determined and the translation problem is solved.

Table 1

The search for a common origin for independent components: PTF of FabD1.3-FabE225 using partial calculated structure factors.

Phasing component	Search component	(a) 20 - 6Å		(b) 20 - 3.5Å		(c) 20. - 3.5Å		(d) 20 - 3.5Å	
D1.3v	E225v	5.6 (1.2)	1	14.0 (3.0)	1	12.8 (2.7)	1	18.3 (3.9)	1
	E225c	5.8 (1.4)	1	10.2 (2.3)	1	8.2 (1.7)	1	13.2 (3.0)	1
	D1.3c	5.3 (1.3)	1	10.3 (2.1)	1	9.7 (2.3)	1	13.9 (3.4)	1
D1.3c	E225v	4.5 (0.9)	2	8.2 (1.9)	1	6.8 (1.5)	1	15.5 (3.2)	1
	E225c	3.3 (0.7)	19	5.8 (1.3)	1	5.6 (1.3)	1	10.3 (2.2)	1
	D1.3v	4.6 (1.1)	1	8.7 (1.8)	1	7.7 (1.6)	1	15.9 (3.6)	1
E225v	E225c	4.0 (0.9)	5	5.5 (1.3)	1	5.5 (1.4)	1	11.8 (2.0)	1
	D1.3v	3.9 (0.9)	5	11.3 (2.4)	1	10.2 (2.3)	1	21.3 (4.4)	1
	D1.3c	3.8 (0.8)	6	6.2 (1.3)	1	5.9 (1.4)	1	13.2 (3.1)	1
E225c	E225v	3.0 (0.8)	68	8.6 (1.9)	1	7.3 (1.5)	1	14.1 (3.1)	1
	D1.3v	4.8 (1.0)	2	9.3 (2.1)	1	8.8 (1.9)	1	17.9 (3.9)	1
	D1.3c	2.8 (0.7)	69	5.1 (1.1)	1	4.5 (0.9)	3	12.1 (2.6)	1

D1.3v, D1.3c, E225v and E225c denote the variable and constant dimers of FabD1.3 and FabE225 respectively. Columns (a), (b) and (c) show results using the initial unrefined coordanates; in (d) the refined structure was used. The density from the phasing component was removed by means of the mask for (a), (b) and (d) and by means of a difference coefficients in (c). The resolution range of the data used for each set of trials is indicated at the head of each column. Peak heights are expressed in terms of the rms value of the translation function. The S/N is given in parenthesis, followed by the rank in height of the correct peak.

This is achieved by applying each symmetry operation in turn to the arbitrarily placed but correctly oriented model and using this as the search component; structure factors calculated (in P1) with the coordinates of this symmetry transformed molecule serve as $\mathbf{F}_{model}(\mathbf{h})$ in equation (2). The relationship between the position, **q**, of the maximum in the PTF calculated in this manner, and the translation, **t**, to be applied to the phasing model is illustrated in figure 3 and given by equation (3), where A_j is the rotation matrix of the j-th symmetry position with $\mathbf{s}_j$ the corresponding translation, and I is the unit matrix.

We shall illustrate the application of the PTF in placing a search model correctly with respect to a crystallographic origin using the structure of FvD1.3[10] (an immunoglobulin fragment consisting of a dimer of the variable domains, V_L and V_H). This molecule crystallises in the space group $P4_32_12$. Table 2 gives, for each symmetry position, the matrix $[A_j - I]$ and the corresponding "inverse" matrix to obtain the translation, **t**, from the peak in the PTF, **q**. Results for placing the FvD1.3 dimer in the unit cell are given in Table 3. From the "inverse" matrices given in Table 2 it is clear that at least two suitably chosen symmetry positions must be selected in order to obtain a complete solution for the translation since certain components of the translation vector remain undetermined for each operation taken separately. There is, of course, an advantage in exploiting all possibilities to provide a generous cross-checking from the redunancy. For the particular example given here, each component of **t** is estimated four times. The results presented in Table 3 were obtained by using difference coefficients to remove the contribution of the phasing model, but comparable results were found were molecular masks were employed.

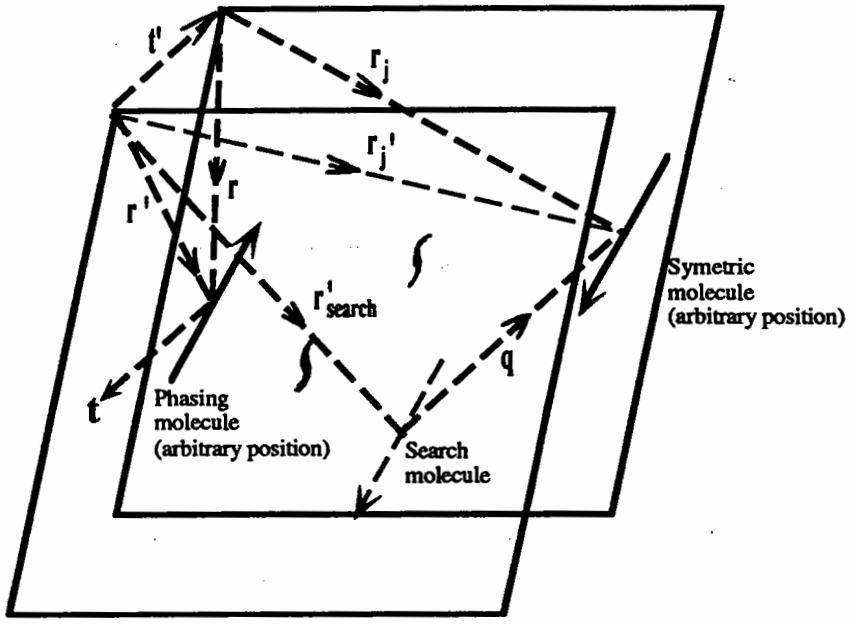


Figure 3

$$
\begin{aligned}
\mathbf{r}_j' &= \mathbf{t}' + \mathbf{r}_j \\
&= \mathbf{t}' + \mathbf{A}_j\mathbf{r} + \mathbf{s}_j \\
&= \mathbf{t}' + \mathbf{A}_j(\mathbf{r}' - \mathbf{t}') + \mathbf{s}_j \\
&= (\mathbf{A}_j\mathbf{r}' + \mathbf{s}_j) - (\mathbf{A}_j - \mathbf{I})\mathbf{t}'
\end{aligned}
$$

Alternatively $\mathbf{r}'_j = \mathbf{r}'_{search} + \mathbf{q}$

Since $\mathbf{r}'_{search} = \mathbf{A}_j\mathbf{r}' + \mathbf{s}_j$

then $\mathbf{q} = -(\mathbf{A}_j - \mathbf{I})\mathbf{t}'$

$= (\mathbf{A}_j - \mathbf{I})\mathbf{t}$ (3)

The result from *TFSGEN*, which uses information from all symmetry operations simultaneously is also compared in Table 2; the signal-to-noise ratio for the two methods is comparable in this case.

Calculations were also tried just using the V_L domain of the molecule, which corresponds to using only about 8% of the unit cell contents to calculate phases; these are given in Table 4. Here, with a lower phasing power of the model, we see a clear advantage in using a molecular mask to remove the contribution from the phasing model. Interestingly, we find also that unlike the case presented in Table 4, the preformance of *TFSGEN* is notably superior, even though the same observed structure amplitudes were used.

(c) Determination of space group enantiomorph

For the example of FvD1.3 we must make the correct choice between the enantiomorphic space groups $P4_12_12$ and $P4_32_12$. Four of the symmetry positions are in common between the two space groups while the other four, determined by the hand of the 4-fold screw axis, are

Table 2

Relationship between **t** and **q** for space group $P4_32_12$

$$\mathbf{q} = (A_j - I)\mathbf{t}$$

Symmetry positions	[A - I]			"inverse"		
(1) -x, -y, 1/2+z	-2	0	0	-1/2	0	0
	0	-2	0	0	-1/2	0
	0	0	0	0	0	0
(2) 1/2-y, 1/2+x, 3/4+z	-1	-1	0	-1/2	1/2	0
	1	-1	0	-1/2	-1/2	0
	0	0	0	0	0	0
(3) 1/2+y, 1/2-x, 1/4+z	-1	1	0	-1/2	-1/2	0
	-1	-1	0	1/2	-1/2	0
	0	0	0	0	0	0
(4)1/2-x, 1/2+y, 3/4-z	-2	0	0	-1/2	0	0
	0	0	0	0	0	0
	0	0	-2	0	0	-1/2
(5) 1/2+x, 1/2-y, 1/4-z	0	0	0	0	0	0
	0	-2	0	0	-1/2	0
	0	0	-2	0	0	-1/2
(6) y, x, -z	-1	1	0	0	0	0
	1	-1	0	0	0	0
	0	0	-2	0	0	1/2
(7) -y, -x, 1/2-z	-1	-1	0	0	0	0
	-1	-1	0	0	0	0
	0	0	-2	0	0	-1/2

Table 3

Phased translation function in reciprocal space for FvD1.3.

Equivalent position of search component in PTF	Position in PTF (q_x, q_y, q_z)	Translation vector (t_x, t_y, t_z)	peaks using difference cffs.
-x,-y,z+1/2	(0.792, 0.822, 0.000)	(0.604, 0.589, -)	12.5 (2.3)
1/2-y,1/2+x,3/4+z	(0.809, 0.014, 0.000)	(0.603, 0.589, -)	10.3 (1.8)
1/2+y,1/2-x,1/4+z	(-0.011, 0.807, 0.000)	(0.602, 0.591, -)	11.0 (2.3)
1/2-x,1/2+y,3/4-z	(0.795, 0.000, 0.394)	(0.603, - , 0.803)	11.4 (2.2)
1/2+x,1/2-y,1/4-z	(0.000, 0.824, 0.393)	(- , 0.588, 0.804)	10.5 (2.1)
y,x,z	(-0.011, 0.014, 0.396)	(- , - , 0.792)	10.4 (2.1)
-y,-x,1/2-z	(0. 808, 0.808, 0.398)	(- , - , 0.796)	11.2 (2.1)
TFSGEN		(0.602, 0.590, 0.802)	17.4 (1.9)

The translation vector was calculated using equation 3. The height of the correct peak is given in rms units (always the highest in this example) followed by the S/N ratio. Contributions from the phasing component was removed using difference coefficients and calculations were made using data in the resolution range 20 - 3.5 Å. The translation vector, **t**, is obtained from the peak in the PTF, **q**, using the transformations in Table 2; an additional translation of (0,0,0), (0,0,1/2), (1/2,1/2,0) or (1/2,1/2,1/2) may be required to place each result with respect to a common origin.

Table 4

Phased translation function in reciprocal space using only V_L of FvD1.3.

Equivalent position of search component in PTF	Translation	peaks using mask	peaks using difference cffs.
-x,-y,z+1/2	(0.604, 0.589, -)	1.02 #1	0.67 #50
1/2-y,1/2+x,3/4+z	(0.604, 0.586, -)	1.00 #1	0.60 #52
1/2+y,1/2-x,1/4+z	(0.603, 0.585, -)	1.10 #1	0.87 #8
1/2-x,1/2+y,3/4-z	(0.605, - , 0.802)	1.18 #1	0.69 #6
1/2+x,1/2-y,1/4-z	(- , 0.586, 0.809)	1.03 #1	0.70 #27
y,x,z	(- , - , 0.815)	0.82 #16	0.59 #76
-y,-x,1/2-z	(- , - , 0.815)	1.05 #1	0.73 #16
TFSGEN	(0.602, 0.588, 0.809)		1.56 #1

The translation vector was calculated using equation 3. Peak heights are expressed as S/N ratios followed by their rank in height. Note that *TFSGEN* uses difference coefficients. Calculations were made using data in the resolution range 20 - 3.5 Å.

different. If we choose symmetry positions (2) or (3) in Table 2, the translation function peak will occur on section z=0 for the correct choice of enantiomorph and on section z=1/2 for the incorrect choice; in each case the x and y coordinates and the peak height remain unchanged by the space group chosen for the calculation (i.e., the whole function is translated by 1/2 in the z direction). The correct choice of space group may also be obtained by using symmetry positions (4) and (5) in Table 2 , but here we have to check the compatibility of the z coordinate with that obtained from the symmetry positions (6) and (7).

Discussion

Some remarks should be made concerning the removal of the contribution of the phasing model to the translation function. If this is done by means of difference coefficients, one consideration is how to determine the scale factor between Fobs and Fcalc. Equating <Fobs> to <Fcalc> as a function of resolution gave the best S/N ratio, irrespective of the fraction of the unit cell contents used to phase the structure factors. Although this does not correspond to the correct scale factor for the complete structure, we have consistently found that this estimation gave the best results, even for the most extreme case given in Table 4 (8% of the unit cell). Where molecular envelopes are employed to remove the electron density of the phasing model, we have found that setting the density inside the envelope to a value close to zero (F(000) was set to zero) gave the best results, although the PTF proved to be rather insensitive to this parameter. It is generally better to use molecular envelopes rather than difference coefficients but the two procedures give similar results provided the phasing model is at least above a certain fraction of the unit cell contents (~15%) and higher resolution data are used (~3.5 Å) as shown in Table 1.

The PTF offers a convenient alternative to solving the translation, especially since use can be made of programs existing in the normal repertoire used by crystallographers (see figure 2).

The examples described here show the versatility of the method in approaching various aspects of the translation problem in molecular replacement.

References

1) COLMAN, P. M., FEHLHAMMER, H. & BARTELS, K. (1976). In *Crystallographic Computing Techniques*, edited by F. R. AHMED, K. HUML & B. SEDLACEK, pp. 248-258. Copenhagen: Munksgaard.

(2) READ, R. & SHIERBEEK, A. J. (1988). *J. Appl. Cryst.* **21**, 490-495.

(3) DOESBURG, H. M. & BEURSKENS, P.T. (1983). *Acta Cryst.* **A39**, 368-376.

(4) CYGLER, M. & ANDERSON, W. F. (1988b). *Acta Cryst.* **A44**, 300-308.

(5) CYGLER M. & DEROCHERS, M. (1989). *Acta Cryst.* **A45**, 563 - 572.

(6) DREISSEN, H.P.C., BAX, B., SLINGLBY, P.F., MAHADEVAN, (MOSS, D.M. & TICKLE, I.J. (1991). *Acta Cryst.* **B47**, 987-997.

(7) BENTLEY, G. A., BOULOT, G., RIOTTOT, M. M. & POLJAK, R. J. (1990). *Nature*. **348**, 254-257.

(8) NAVAZA, J. (1987). *Acta Cryst.* **A43**, 645-653.

(9) TICKLE, I. J. (1985). *In Molecular Replacement. Proceedings of the Daresbury Study Weekend 15-16 February 1985*, edited by P. A. Machin, pp 22-26. SERC Daresbury Laboratory.

(10) BHAT, T. N., BENTLEY, G. A., FISCHMANN, T. O., BOULOT, G. & POLJAK, R..J. (1990). *Nature*. **347**, 483-485.

(11) CCP4, (1979). The SERC (UK) Collaborative Computing Project no. 4; a suite of programs distributed by the Daresbury Laboratory, Warrington, WH4 4AD, UK.

(12) BENTLEY, G.A. & HOUDUSSE, A; (1992). *Acta Cryst.* **A48**. In the Press.

MOLECULAR REPLACEMENT STUDIES AT EMBL HAMBURG

Zbigniew DAUTER

European Molecular Biology Laboratory (EMBL), c/o DESY, Notkestrasse 85, D-2000 Hamburg 52, Germany.

The examples shown below are structures solved recently in EMBL Hamburg using data collected on synchrotron beam lines or a sealed tube source using an imaging plate scanner as a detector. The programs usually used were ALMN and SEARCH from the CCP4 suite (CCP4, 1979).

Several structures of proteinases from the subtilisin family (cooperation with NOVO-NORDISK in Copenhagen) have been solved by molecular replacement, as almost every mutant or complex crystallized in a different cell. These structures are similar and usually the rotation/translation problem was solved easily using the native subtilisin structure as a model. However at first the native savinase structure (Betzel et al., 1991) could not be solved, when the unrefined model of subtilisin BPN′ (Drenth et al., 1972), PDB data set 2SBT (Bernstein et al., 1977), was used. With the well refined model of subtilisin Carlsberg taken from the complex with eglin-c (McPhalen and James, 1988, PDB data set 2SEC) the solution presented no difficulty. The sequence homologies between savinase and two subtilisins are very similar as are their overall structures. This shows the importance of using refined structures as models for molecular replacement searches to ensure that most interatomic vectors are correct, in spite of using only relatively low resolution reflections in the procedure.

The structure of a mesophilic subtilisin, mesentericopeptidase, complexed with eglin-c has also been solved by molecular replacement (Dauter et al., 1991). The first attempt using as model the subtilisin Carlsberg : eglin-c complex gave a clear solution. The correct peak in

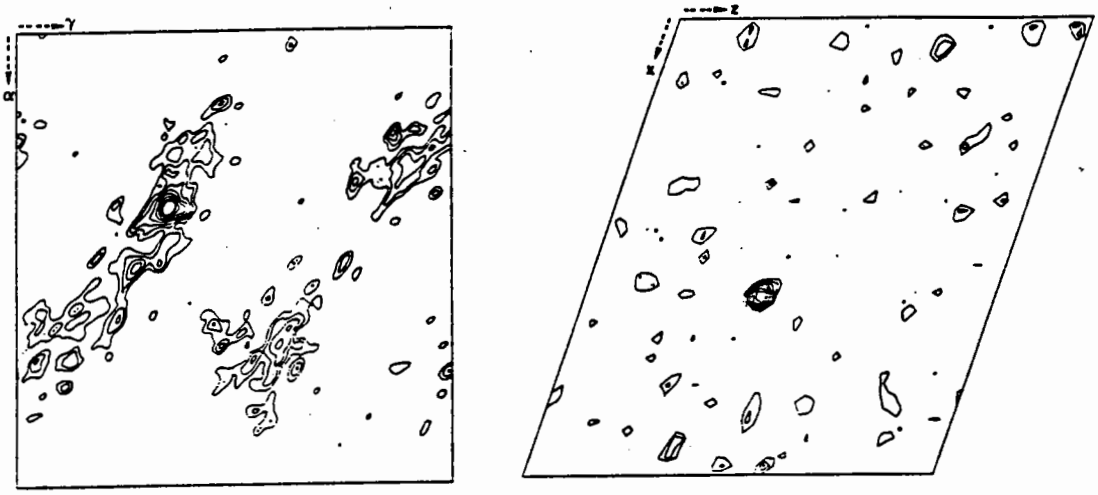


Fig. 1. Section $\beta = 50°$ of the rotation function, and the R-search map for eglin-c in its complex with mesentericopeptidase using difference amplitudes.

the rotation function map was at the 9 σ level, 2.5 times higher than the next peak. The R-factor search, only in the (x,z) plane for space group $P2_1$, produced the correct solution with R = 35.0 %, and a next lowest value of 38.7 %. This model was subjected to several cycles of restrained refinement using data to 2.5 and then 2.0 Å resolution to an R value of about 28 %. Inspection of the $(3F_o\text{-}2F_c)$ Fourier map showed good agreement of enzyme model with electron density, but poor density for most of the eglin-c molecule. This suggested that the relative orientation of the eglin molecule was to some extent different than in the subtilisin Carlsberg complex. Flexibility of eglin binding to subtilisins has previously been observed in its complexes with thermitase (Dauter et al., 1988, Gros et al., 1989)

This problem was overcome by independent rotation and translation of subtilisin and eglin molecules separately. The enzyme has 275 amino acids and the inhibitor 65. The use of the subtilisin Carlsberg molecule alone easily produced a solution equivalent to the previous one based on the complex. The next step involved prerefinement of the enzyme molecule by 6 cycles of restrained least-squares minimisation. The structure factors resulting from this model consisting of the enzyme molecule only, i.e. about 4/5 of the structure, were calculated and, after scaling down to about 80%, were subtracted from the observed amplitudes. The resulting set of difference amplitudes was subsequently used to establish the orientation and position of the eglin molecule in the cell. A clear solution was obtained (Fig. 1), with the peak in the rotation function at 10 σ, more than two times higher than any other, and the R-value search resulted in a single best value of 45.0 % with next highest of 46.9 %. The subtilisin and eglin molecules were then placed correctly in the mesentericopeptidase cell

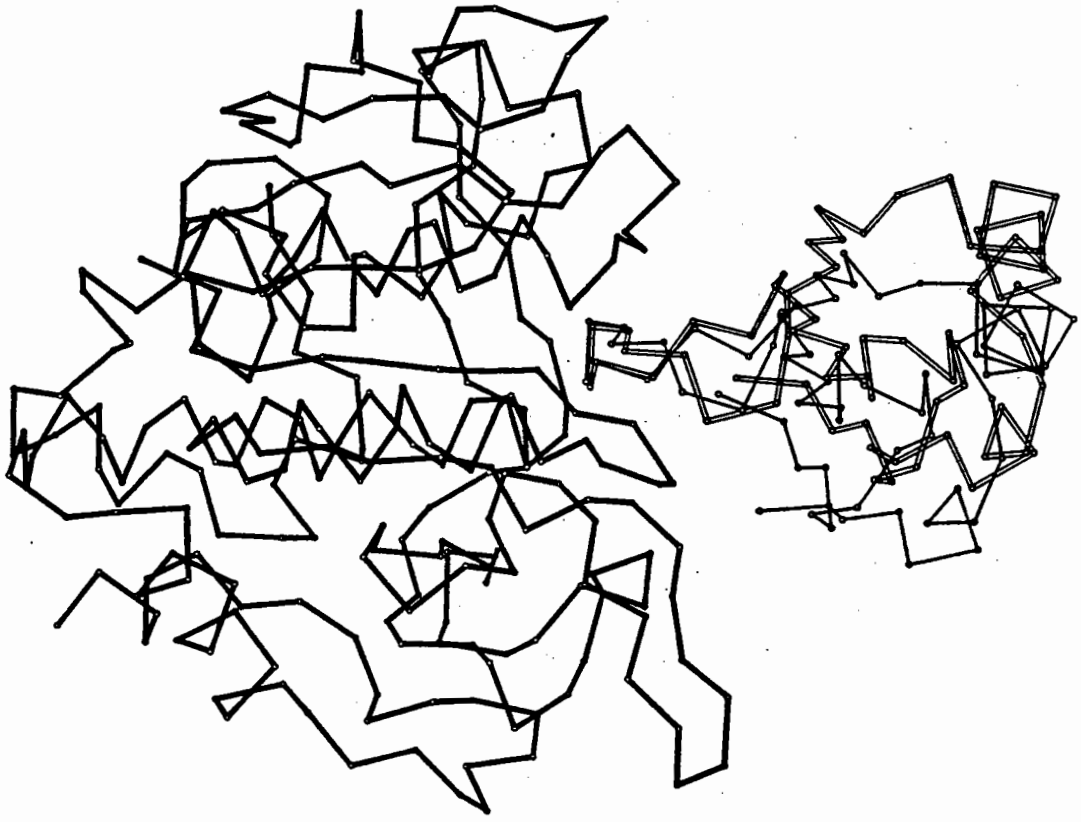


Fig. 2. Mesentericopeptidase molecule in complex with eglin-c (thin lines). The eglin molecule in complex with subtilisin Carlsberg is also shown (open lines).

and subsequent refinement of the model progressed without difficulty, giving a final R value of 15.1 % for all data to 2.0 Å. The eglin molecule within the complex appeared to be rotated 14° relative to the subtilisin molecule compared to the Carlsberg complex, Figure 2.

A similar problem has been encountered with the solution of formate dehydrogenase (FDH) in the apo form using the holo enzyme as a model (Lamzin et al., 1992). The 2 x 43 kD enzyme is built up as a non-crystallographic dimer of subunits each consisting of two domains. The larger, catalytic, domain forms the dimer contacts and the smaller, coenzyme binding, domain lies at the periphery of the dimer. The relative orientation of the domains is different in the apo and holo form. As in the case of mesentericopeptidase, it was possible to obtain a solution of the rotation and translation problem using the whole subunit as a model, but subsequent refinement and rebuilding of the structure was difficult.

A better model was obtained by solving the molecular replacement problem first for the large domain, and then for the small domain separately using the difference set of amplitudes, as explained above for mesentericopeptidase. In this case the automated refinement procedure developed by V. Lamzin greatly facilitated rebuilding and refinement of the model obtained from molecular replacement (see Lamzin, this volume).

The glucose isomerase from *Streptomyces rubiginosus* crystallizes in space group I222, a = 93.9, b = 99.7 and c = 102.9 Å, with the tetramer of 4 x 40 kDa subunits positioned at the 222 symmetry site (Dauter et al., 1990). Two attempts to solve this structure were made. At first the unrefined PDB coordinate set 3XIA (Farber et al., 1987) of a similar tetrameric enzyme from *S. olivochromogenes* crystallizing in $P2_12_12$ cell with a = 99.2, b = 94.2 and c = 87.5 Å was used. These crystals are highly pseudosymmetric and the structure had been built in the I222 cell, ignoring the weak reflections with (h + k + l) odd. As the 222 site in the I222 cell lies at a special position, in this case the problem was limited to finding the orientation of the tetramer in the new cell, i.e. selecting one of six possible permutations of three molecular twofold axes. The rotation function gave a very clear answer. However subsequent refinement of the model proved very difficult. Substantial parts of the model did not agree with the electron density maps and the use of omit maps did not allow correct rebuilding of the wrong parts. This can be probably due to

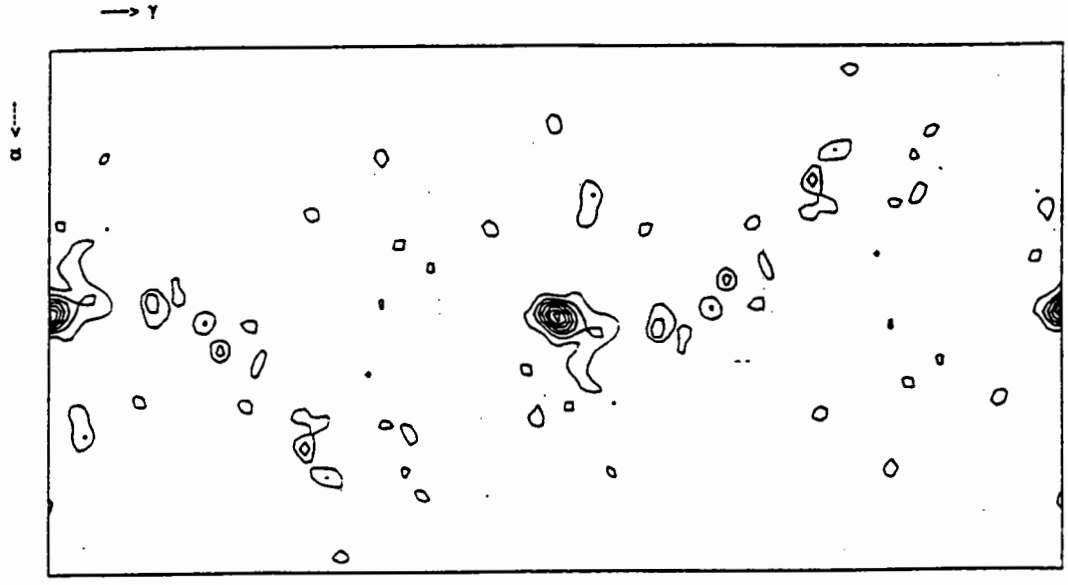


Fig. 3. Section $\beta = 90°$ of the glucose isomerase rotation function. The peak at $\alpha = 0$, $\beta = 90$ and $\gamma = 0°$ corresponds to a cyclic permutation of the tetramer axes.

the use of only half of the reflections in the 3XIA structure analysis and as a result wrongly tracing part of the polypeptide chain in the neighbouring subunits of the tight tetramer.

The second, successful attempt at the structure solution involved a different model, from *Arthrobacter*, (Henrick et al. 1989, PDB set 4XIA) crystallizing in the trigonal space group $P3_121$ and also consisting of a tetramer positioned on the 2-fold axis. The sequence homology for these two enzymes is lower than between the two *Streptomyces* enzymes. Nevertheless the model proved to be far better. The single subunit of the enzyme was rotated into the I222 cell, Fig. 3, and again only cyclic permutation of the axes was necessary. The refinement and updating of the resulting model proceeded smoothly to an R value of 14.1 % at 1.65 Å resolution. This example illustrates that it is possible to solve the molecular replacement problem with a model so different from the right structure that it makes the subsequent rebuilding and refinement very difficult. A degree of pseudosymmetry exacerbated the difficulties.

Another example illustrating the problem of adequacy (or rather inadequacy) of the models used in molecular replacement is the structure solution of bacterial trypsin (cooperation with NOVO-NORDISK, Rypniewski et al., to be published). This enzyme crystallizes in space group $P2_1$ with a = 33.4, b = 67.6, c = 39.8 Å, β = 107.6° and data have been recorded to 1.9 Å. The use of bovine trypsin (PDB set 1TPO, Bode et al., 1983) or trypsin from *Streptomyces griseus* (PDB set 1SGT, Read and James, 1988) led to a relatively clear solution of rotation and translation problem in both cases. Rebuilding and refining those models proved to be difficult, as both sequences differ substantially from each other and from the enzyme studied. Several insertions/deletions are necessary to align these three sequences. Eventually a composite model was created from carefully chosen parts of maps based on the two known structures, to match as well as possible the sequence of the protein to be solved. This model was then oriented and translated to the correct place and the structure rebuilt and refined to a final R-factor of 14.1 %.

The mutant of bovine pancreatic trypsin inhibitor, BPTI, is an example of a crystal structure containing two molecules in the asymmetric unit of the monoclinic $P2_1$ cell with a = 24.6, b = 41.7, c = 41.1 Å and β = 98.6°. The molecule of BPTI (Deisenhofer and Steigemann, 1975, PDB set 4PTI) was a good model for the molecular replacement searches. Solution of the rotation function and R-factor searches came up clearly in the corresponding maps, Fig. 4, in spite of the fact that one molecule constitutes only half of the content of the asymmetric unit. After orienting the two molecules, they were translated

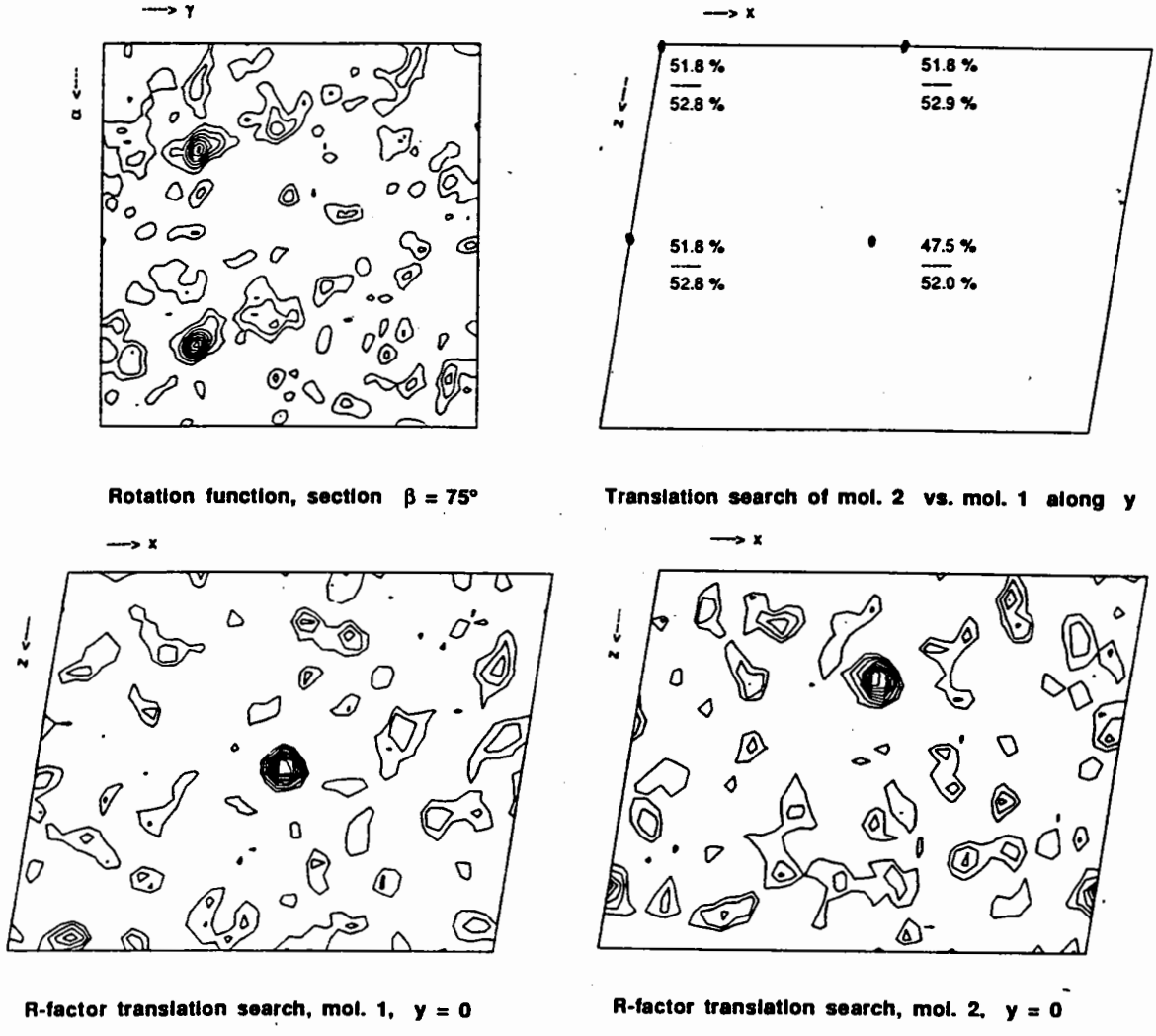


Fig. 4. Molecular replacement results for a BPTI mutant containing two molecules in the asymmetric unit. Both peaks on the rotation function map appeared on the same section.

individually to the correct place in the (x,z) plane. Their relative translation along the y-axis, taking into account four possible origin shifts, was found in a one-dimensional search giving a minimum R value of 47.5 % for the correct origin shift and of 51.8 % for the three other choices. The average R value was about 52.8 % for all four origins.

REFERENCES

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Mayer, E.F.Jr., Bryce, M.D., Rodgers, J.R., Kennard, O., Simanouchi, T. and Tasumi, M., *J.Mol.Biol.*, **112**, 535 (1977).

Betzel, C., Klupsch, S., Papendorf, G., Hastrup, S., Branner, S. and Wilson, K.S., *J.Mol.Biol.*, **223**, 427 (1991).

Bode, W., Walter, J. and Huber, R., *Acta Crystallogr.*, **B39**, 480 (1983).

CCP4, Collaborative computing project no. 4, Daresbury Laboratory, Warrington, England (1979).

Dauter, Z., Betzel, C., Genov, N., Pipon, N. and Wilson, K.S., *Acta Crystallogr.*, **B47**, 707 (1991).

Dauter, Z., Betzel, C., Höhne, W.E., Ingelman, M. and Wilson, K.S., *FEBS Lett.*, **236**, 171 (1988).

Dauter,Z., Terry, H., Witzel, H. An d Wilson, K.S., *Acta Crystallogr.*, **B46**, 833 (1990).

Drenth, J., Hol, W.G.J., Jansonius, J. and Koekoek, R., *Eur.J.Biochem.*, **26**, 177 (1972).

Farber, G.K., Petsko, G.A and Ringe, D., *Protein Eng.*, **1**, 459 (1987).

Gros, P., Betzel, C., Dauter, Z., Wilson, K.S. and Hol, W.G.J., *J.Mol.Biol.*, **210**, 347 (1989).

Henrick, K., Collyer, C.A. and Blow, D.M., *J.Mol.Biol.*, **208**, 129 (1989).

Lamzin, V.S., Aleshin, A.E., Strokopytov, B.V., Yukhnevich, M.G., Popov, V.O., Harutyunyan, E.H. and Wilson, K.S., *Eur J.Biochem.*, **206**, 441 (1992).

McPhalen, C.A. and James, M.N.G., *Biochemistry*, **27**, 6582 (1988).

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

Two Examples of Molecular Replacement with X-PLOR

Leo Brady & Jiang Jian-sheng

Department of Chemistry, University of York, York YO1 5DD

Brunger (1990) recently described a new search strategy for molecular replacement within the program package X-PLOR. The method entails "Patterson refinement" of large numbers of possible solutions to the rotation function, the target function for this refinement being combined with an empirical function describing geometric and non-bonded interactions (see paper by Brunger in these proceedings). The correct orientation is identified by a minimum value in the target function after refinement. We describe two examples where the application of this method has led to an improvement over a conventional molecular replacement strategy.

Example 1: B72.3, a chimaeric antibody Fab' fragment.

Our first example is the structure solution of a monoclonal antibody Fab' fragment. The antibody, B72.3, binds to a cell-surface glycoprotein found on tumour cells, and has many clinical therapeutic and diagnostic applications. The Fab' fragment is derived from an engineered chimaeric form of the antibody and contains both human and mouse sequence. As this antibody is the target of many grafting and "humanisation" experiments, the structure of the B72.3 Fab' fragment is of considerable interest.

Crystals of B72.3 are orthorhombic ($P2_12_12_1$) with a=67.3, b=93.2 and c=208.8 Å (Brady et al., 1991). Although diffraction data to 2.6 Å could be observed when using synchrotron radiation,

Table 1: Crystallographic data statistics for B72.3 Fab'

1) 3.5 Å Data Set		2) 3.1 Å Data Set	
Number of Observations	38,849	Number of Observations	57,781
Unique reflections	10,254	Unique reflections	15,644
Redundancy	3.8	Redundancy	3.7
87% complete to d_{min} =	5.0 Å	98% complete to d_{min} =	5.2 Å
64% " "	3.5 Å	84% " "	3.5 Å
		74% " "	3.1 Å
Number of crystals:	1	Number of crystals:	1
R_{merge} =	6.2%	R_{merge} =	6.8%

R_{merge} for each data set was calculated on intensities using all of the data in the CCP4 program AGROVATA.

acute radiation sensitivity of the crystals made data collection problematic. This problem was partly overcome by using the Weissenberg camera and image plate at the Photon factory in Japan. On separate trips two native data sets were collected: one to 3.5 Å spacing, and a second to 3.1 Å (Table 1). Both data sets were less than ideal: they were limited in resolution, incomplete due to the geometry of the Weissenberg camera, and the first data set was also collected from an imperfect (slightly twinned) crystal.

Estimates of solvent content and the unit cell volume suggested two Fab' molecules in the asymmetric unit. A native Patterson revealed a large peak (approximately 25% of origin peak height) at (0 0.5 0.125). This was interpreted as being due to having the two molecules lying in very similar orientations and separated by half a unit cell edge along y, and one-eighth along z. This relationship was also reflected in the distribution of intensities in the data, where reflections for which k+l/4=odd integer were either altogether absent or very weak in the data, a relationship which persisted to at least d_{min} = 4 Å. We therefore began our molecular replacement with two requirements: although we expected two molecules in the asymmetric unit there should be only one dominant orientation in the cross rotation function, while in the translation function we expected two solutions of similar intensity separated by the vector (0 0.5 0.125).

Sequence similarity searches with Fab fragments for which crystal structures were known showed B72.3 to have greatest sequence homology with HyHel-5 Fab (PDB file 2HFL.PDB - this is a complex between the Fab and lysozyme). The HyHel-5 structure also had the advantage of being refined to reasonable resolution (2.5 Å) and having an elbow angle close to the middle of the observed range (155°, in range 120-180°). Two search models were constructed from HyHel-5: in both all hypervariable loops were deleted, as was a loop in the constant domain which is unusual in HyHel-5. Model 1 retained the HyHel-5 sequence. In Model 2 we replaced the sequence with that for B72.3 wherever this differed form that of HyHel-5. New side-chains were initially placed in idealised conformations, then energy-minimised using CHARMM.

Cross-rotation searches using different combinations of each model with each of the data sets and integration radii of 8-3.5Å or 8-3.1Å were performed within X-PLOR (version 2.1). In each case the top 200 possible orientations were then subjected to the X-PLOR PC-refinement procedure. In the rigid-body refinement incorporated within the PC-refinement the Fab was firstly treated with 2 degrees of freedom (C and V domains) and then with 4 degrees of freedom (V_H, V_L, C_H1 and C_L domains treated as independent). In

each of the four cases (Table 2), as expected, a single orientation solution was dominant after PC-refinement - this often corresponded to the solution with the greatest peak height in the initial rotation function. The peak height, measured as the correlation coefficient of the PC-refinement, varies for each of these solutions. Closer examination of each of these solutions, which all gave orientation angles within 5° of one another, showed the rigid-body refinement procedure had altered the Fab elbow angle by differing amounts in each case (Table 2). The largest movement (18°, see Figure 1) was only achieved when using the higher resolution (3.1 Å) data in combination with Model 2, which had the correct protein sequence.

Table 2: Variation of r_{PC} (expressed as percentage) and Fab' elbow angle using both search models and data sets in the PC-refinement procedure.

	Model 1		Model 2	
	r_{PC}	Elbow angle	r_{PC}	Elbow angle
3.5 Å data	6.7%	4°	6.0%	5°
3.1 Å data	6.9%	14°	10.6%	18°

The models were then rotated by the angles corresponding to the solutions identified as 1 and 2 in Table 2 , and submitted to the translation function search within XPLOR using data 15.0 to 3.1Å. The non-crystallographic symmetry (NCS) evident in the native Patterson required two solutions related by the translational vector (0 0.5 0.125). Only Solution 2 gave two peaks (relative heights peak 1 =100.00, peak 2 = 90.83, next peak = 65.80) related by this vector -

(0.152 0.315 0.355) and (0.152 0.804 0.485) - Solution A

A traditional R-factor translation search performed using the CCP4 program TFSGEN did, however, produce possible solutions for both rotation solutions. For Solution 2 these were again the most intense two peaks (relative heights 100 and 95.1, next highest peak 65.9), and corresponded to the translation vectors

(0.143 0.312 0.355) and (0.143 0.805 0.485)

which are essentially the same as Solution A. The peaks from the Solution 1 oriented model, however, were lower in relative peak heights (66 and 68, highest peak 100) and corresponded to a different set of vectors

(0.43 0.42 0.285) and (0.43 0.92 0.410) - Solution B

Solution B was refined against the 3.5Å data using simulated annealing (SA) in XPLOR. The standard crystallographic R-factor (as calculated in X-PLOR on data greater than 2 sigma in the resolution range 15 to 3.5 Å) dropped from an initial value of 0.524 to 0.225 with root mean square deviations from ideality of bond lenths (rms-bonds) = 0.047Å and deviations of angles (rms-

angles) = 7.23°, but no non-crystallographic constraints were applied and the inappropriate weights in the refinement resulted in poor protein stereochemistry. However, when refined against the 3.1Å data and with strict NCS constraints applied, the R-factor rose to 0.315 with rms-bonds = 0.048Å and rms-angles = 7.49°. This solution was therefore believed incorrect.
Solution A, by contrast, could be refined readily with strict non-crystallographic symmetry restraints maintained, to eventually produce a correct final model with R-factor = 0.176 and far better stereochemistry (rms-bonds= 0.017Å and rms-angles= 3.73°). Full details of the refinement and final structure are presented in Brady *et al* (1992).

Discussion

Brunger *et al.* (1991) have previously demonstrated the effectiveness of the PC-refinement procedure in XPLOR in orienting multi-domain search models such as Fab fragments. The advantage of the X-PLOR procedure over a conventional cross-rotation search is that it enables the search to be carried out with a complete 4-domain model - rather than with separate domains as is often the case. As the rigid-body refinement within the PC-refinement procedure can be used to automatically adjust the elbow angle in a Fab prior to a translational search, this improved orientation leads to an increase in the sensitivity of the translation search (Brunger, 1990). The radius of convergence of the rigid-body refinement procedure has been estimated at 13° (Brunger *et al.*, 1991). In our

Figure 1: Overlay of Cα's of B72.3 (thick) on those of HyHel-5 (thin), on which the search model was based. The 18° variation in elbow angle, obtained in the PC-refinement procedure, is clearly shown.

case this convergence appears to be both data and model sensitive. For B72.3, a lower resolution and less complete data set failed to

adjust correctly the elbow angle of the search model. However, even when an improved data set was used, the original search model did not give the correct orientation without manual adjustment of the elbow angle. By 'optimising' our search model - in this case by correction of the sequence and energy minimisation of replaced sidechain conformations - we were able to obtain the correct orientation even though this required the rigid body refinement procedure to adjust the elbow angle by 18°. This occurred only when both the most complete data set and optimised model were used. Hence, in this case at least the radius of convergence of the PC-refinement procedure appears to extend well beyond the 13° limit previously proposed, provided that an appropriate search model is used.

Our experience in being able to refine the incorrect Solution B to a reasonably low R-factor demonstrates the caution that must be applied when using powerful refinement techniques such as simulated annealing with data of limited resolution. In this case the incorrect solution was easily identified because agreement could only be achieved at the expense of good protein stereochemistry and non-crystallographic symmetry. Other symptoms included the decreased agreement when the model was refined against the higher resolution and more complete data, and breaks in main chain continuity in density maps. That the solution could be refined at all may in part be due to the similarity of two of the components of the translation vector to their corresponding components of the correct solution. This solution was in any case considered less likely from the outset as both the rotation and translation maxima were of lower intensity than those for Solution A, although initial packing analyses of both solutions revealed acceptable intermolecular interactions.

2. X8DPI, a mutant monomeric insulin

Our second example is an engineered single-point mutant (A8Thr-> His) of human insulin from which the 5 C-terminal residues of the B-chain had been proteolytically removed. The resulting protein (X8DPI) is small (45 residues) and, unlike normal insulin, monomeric. Two crystal forms of the protein were obtained: a C2 form with 2 molecules in the asymmetric unit and isomorphous with other known DPI structures, and a second novel crystal form which was orthorhombic ($P2_12_12_1$) with a tiny unit cell (a=24.09, b=27.1, c=55.23Å) and one molecule/asymmetric unit. This latter form had proved problematic for solution by molecular replacement: repeated searches with various insulins as search models had produced a reasonably consistent although not altogether convincing solution by conventional means, but for

which the R-factor could not be refined beyond 28%. This was a difficult case for molecular replacement: the unit cell was small with very little solvent (18%), the protein is elliptical rather than globular in shape and the many insulin structures solved have shown a high degree of flexibility making the choice of a suitable search model difficult. In contrast to B72.3 Fab', however, in this case the low solvent content at least had the advantage that the crystals diffracted very well: the native data was 99% complete to 1.9Å resolution, with an R_{merge} on intensities of 5.1%.

Our starting point for molecular replacement within X-PLOR was the refined model for the C2 form of the X8DPI. This model, refined to R=14.9% with 1.7 Å data, contained two independent copies of X8DPI, which were treated as two separate search models for the orthorhombic crystal solution. Regions known from existing insulin structures to be highly variable were deleted: residues A1-A4 and B1-B4 (this represented nearly 20% of the structure). Cross-rotations with each of these models and using data from 8-3Å produced a list of possible orientations, the top 200 of each were submitted for PC-refinement. In this case no inter-domain flexibility could be assigned within the rigid-body refinement procedure. As high resolution data were available, the PC-refinement was performed with a variety of shells of data, although addition of high resolution data beyond 3 Å decreased the clarity of solutions. One of the two models gave a consistently higher solution (Figure 2), although this solution is closer to noise

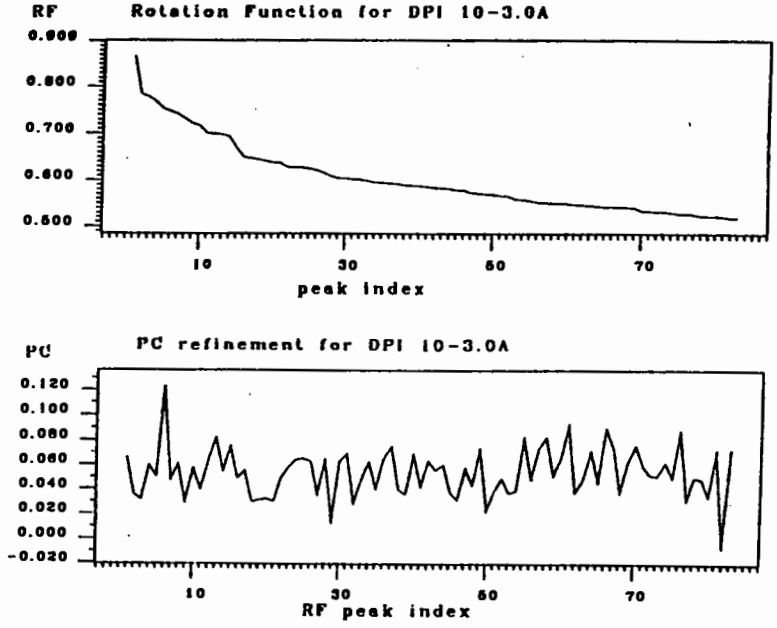


Figure 2: Cross-rotation solutions for X8DPI both before (above) and after (below) PC-refinement. The correct solution (6th in intensity in the original search) is clearly distinguished by the PC-refinement (see lower graph).

levels than was observed for the Fab case. This solution went on to perform well in the translation function, in which close examination of intermolecular contacts was used to validate the correct solution.

Refinement started with a high initial R-factor (55% on 8-1.9Å data) and did not improve at all on rigid-body refinement within the very closely packed unit cell. A single round of simulated annealing refinement saw a rapid drop in R-factor (to 30%) after which density for the missing N-terminal residues of the B-chain was obvious. These residues were manually built-in using FRODO, and refinement continued using both X-PLOR and PROLSQ. Judicious and cautious insertion of water residues, each selected carefully by examination of real-space density-figures-of-merit (DFOM, manuscript in preparation) after refinement, was necessary in order to refine the model to an R-factor of 0.195 (all data to1.9Å). The relatively high R-factor of the current model reflects disorder for the 4 N-terminal residues of the A-chain (nearly 10% of the protein residues), for which several conformations are apparent in the crystals. Further details will be reported in Brady et al (manuscript in preparation).

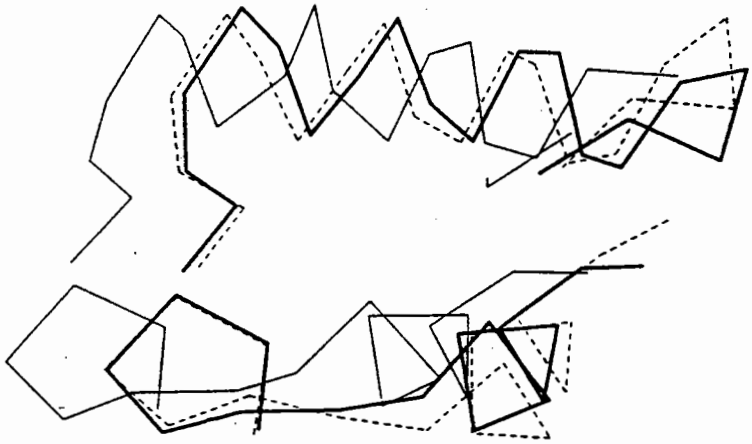


Figure 3: Cα plot showing the final solution of X8DPI (thick line) overlaid on the initial solution obtained with CCP4 programs (thin line) and the initial solution obtained with X-PLOR (dashed line). Both of the latter solutions have been rigid-body refined but are otherwise before any further refinement.

Discussion

A comparison of the final structure, obtained with X-PLOR, with that derived by conventional (CCP4 program suite) molecular replacement, shows the original solution to be essentially the same

(1 Å rms displacement of mainchain, Figure 3). With persistent and careful refinement this solution could most likely had been refined successfully. However, the X-PLOR derived solution appears to have been more successful for two reasons. Firstly, the initial orientation after PC-refinement is more accurate than that obtained with ALMN (see Figure 3). In this particular case the very dense packing in the crystal unit cell presents many false minima during refinement, and hence the more accurate the original orientation, the more likely is correct refinement. Secondly, the PC-refinement procedure provides a means of ranking cross-rotation search solutions. In this case this is crucial when there are many peaks of similar height. Most importantly, this further "filtering" step provides additional confidence that the solution arrived at is indeed the correct one. Our failure to pursue refinement of the original solution from ALMN further could largely be attributed to uncertainty in the correctness of this solution which was present in many cross-rotation solutions, but rarely the most prominent solution.

References

Brady, R.L., Edwards, D.J., Hubbard, R.E.,Jiang, J.-S., Lange, G., Roberts, S.M., Todd, R.J., Adair, J.R., Emtage, J.S., King, D.J., and Low, D.C. (1992). Crystal structure of a chimeric Fab' fragment of an antibody binding tumour cells. *J. Mol. Biol.* (in press).

Brady, R.L., Hubbard, R.E., King, D.J., Low, D.C., Roberts, S.M., Todd, R.J. (1991). Crystallization and preliminary X-ray diffraction study of a chimeric Fab' fragment of an antibody binding tumour cells. *J. Mol. Biol.* **219**, 603-604.

Brunger, A.T. (1990). Extension of molecular replacement: a new search strategy based on Patterson Correlation refinement. *Acta Cryst.* **A46**, 46-57.

Brunger, A.T., Leahy, D.J., Hynes, T.R., Fox, R.O. (1991). 2.9Å Resolution structure of an anti-dinitrophenyl-spin-label monoclonal antibody Fab fragment with bound hapten. *J. Mol. Biol.* **221**, 239-256.

General Discussion

Led by Wim Hol; and reported by Eleanor Dodson

At the end of the meeting there was a lively discussion where Wim posed several important questions. It seems worthwhile reporting them, and some of the answers. Please DO NOT treat this as definitive - I may have misunderstood speakers, and certainly have overlooked some contributions.

1) The power of density averaging

Many structures have only been solved because non crystallographic symmetry averaged has enhanced the signal, and reduced error. There are effectively more intensity observations per atomic parameter, so such map have a quality usually associated with higher resolution structures. Gerard Bricogne describes the theory in the 1970s, and there are many examples of structures solved using the technique.

2) Phase extension with density averaging

This is a very powerful technique used for solving virus structures. Its power is dependent on the number of copies of the molecule available for averaging, the accuracy of the description of the non-crystallographic symmetry and on the accuracy of the envelope.

The virus studies described here by Michael Rossmann and Lars Liljas show that with many copies, and with the extremely precise description of the non crystallographic symmetry required to generate the virus isododecahedron phase extension can proceed from low resolution (< 8Å) and produce an excellent high resolution map, even with a very approximate envelope. (Michael pointed out that if the envelope was centro symmetric, eg a spherical shell, the phases generated may belong to the Babinet lattice, so extra care is needed (see Lars Liljas' paper)).

Phase extension has also been used for non viruses; Wim Hol reported his work on Haemerythrin in the 1985 Proceedings, where he extended phases from 5Å with 6 fold averaging.

3) Phase refinement using density averaging and solvent flattening

This is usually used when there is some phase information available from isomorphous replacement. It works extremely well and can be a method for refining the non crystallographic symmetry parameters (see Daresbury Proceedings 1990 and 1991).

4) Completeness of data?

There was no general agreement about the necessity of this. There are several examples of structures where the molecular replacement technique was unsuccessful with partial data sets. Gideon Davies' paper describes the

disastrous effect of omitting the strong reflections from the data. Jorge Navaza says that his method for generating spherical harmonics is less perturbed by missing terms than that used in ALMN or MERLOT. Various tricks can help in special cases.

When there is well defined non crystallographic symmetry within the asymmetric unit, the locked rotation function which looks simultaneously for solutions compatible with this, screens out noise.

5) Does " Model bias" survive refinement?

This question presumes that there are sufficient intensity observations available to proceed to sensible refinement (see Eleanor Dodson's paper). If the molecular replacement solution is correct, Wim said errors in loops show up very clearly (see A. Mattevi's paper). Keith Wilson pointed out that if the solution is incorrect, there will not be much movement during refinement and you need to use more subtle arguments to detect the error. Gideon says: If there is bias how will you ever see it in a conventional map?

6) Does success depend on using a well refined model?

There are a lot of examples where molecular replacement has NOT succeeded with a poor model. Zbigniew Dauter describes one such case . XPLOR is able to adjust well described domains to give a solution (see R L Brady's paper). So far there has been little success with using models derived from NMR to solve crystal structure.

7) How big a fragment is required for a successful solution?

When the model is well conserved, one eighth of the asymmetric unit has been fitted. Dale Wigley could see one LDH molecule from a asymmetric unit containing 2 tetramers. Paula Fitzgerald referred to a paper by Sheriff where in a complex of 2 lysozymes and two FABs, he was able to detect one FAB domain. No-one so far has been able to fit a single helix. Chris Nordmann searched for helices in myoglobin and found the direction but not the rotational parameters using a perfect helix, and data to 1.5Å. Zbysek Otwinowski says "the larger the molecule the smaller the fragment needed". Gerard Bricogne pointed out that better statistical methods should increase the sensitivity. Zbysek Otwinowski recommends the correlation coefficient used in XPLOR.

8) How did a fragment is needed to give useful phasing to rebuild the missing bits?

Gerard thinks 15%. This estimate is given by analogy with the findings of small-molecule crystallographers, as exemplified by Paul Beurskens 's program DIRDIF. It may depend on a degree of data completeness normally unachievable for macromolecular structures. Eleanor Dodson has tried using the Phased Translation function (Read, 1988) to fit a second fragment using phases based on the first. This can work, but is probably more complicated than using Ian Tickle's procedure.

At this point the discussion was cut short by honking taxis waiting to whisk people away. Most of the references are covered in the text, but here is a short extra list.

I found it a most useful resume of the two days and Wim chaired it superbly.

Extra references:

Dauter, Z., Terry, H., Witzel, H. and Wilson, K.S. (1990) Acta Cryst, B46, 833-842.

Liang Tong & Rossmann, M.G. (1990) Acta Cryst, A46, 783-792.

Sheriff, S., Padlan, E.A., Cohen, G.H. and Davies, D.R. (1990) Acta Cryst B46, 418-425.

Mr I S B Abeysinghe
Department of Molecular Biology and Biotechnology
University of Sheffield
Western Bank
Sheffield S10 2TN

Dr J P Abrahams
MRC Laboratory of Molecular Biology
Hills Road
Cambridge CB2 2QH

Dr A Achari
Wellcome Research Laboratories
Langley Park
South Eden Park Road
Beckenham
Kent BR3 3BS

Dr R Acharya
Department of Biochemistry
University of Bath
Claverton Down
Bath BA2 7AY

Mr P D Adams
Department of Biochemistry
Hugh Robson Building
University of Edinburgh
Edinburgh EH8 9XD

Dr M J Adams
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Dr D Alexeev
Department of Biochemistry
Hugh Robson Building
University of Edinburgh
Edinburgh EH8 9XD

Dr P M Alzari
Unité D'Immunologie Structurale
Institut Pasteur
25 Rue du Dr Roux
75724 Paris Cedex 15
France

Dr I A Andersson
Laboratory of Molecular Biophysics
University of Oxford
South Parks Road
Oxford OX1 3QU

Mr A Antson
EMBL
C/o DESY
Notkestrasse 85
2000 Hamburg 52
Germany

Miss S Armstrong
Department of Biochemistry and Molecular Biology
University of Leeds
Leeds LS2 9JT

Dr B Arnoux
Laboratoire de Cristallochimie
ICSN-CNRS
91198 Gif sur Yvette Cedex
France

Dr P Artymiuk
Krebs Institute
Department of Molecular Biology and Biotechnology
University of Sheffield
Western Bank
Sheffield S10 2TN

Dr S Bailey
Department of Chemistry
University of Manchester
Manchester M13 9PL

Dr P J Baker
Krebs Institute
Department of Molecular Biology and Biotechnology
University of Sheffield
Western Bank
Sheffield S10 2TN

Dr A K Basak
Laboratory of Molecular Biophysics
Rex Richards Building
University of Oxford
South Parks Road
Oxford OX1 3QU

Dr G Bentley
Dept d'Immunologie
Institut Pasteur
25 rue du Dr Roux
75724 Paris
France

Ms M C Bewley
Department of Biochemistry and Molecular Biology
University of Leeds
Leeds LS2 9JT

Dr T Bizebard
Laboratoire de Biologie Physicochimique
Bâtiment 433
Université Paris Sud
91405 Orsay Cedex
France

Professor D Blow
Blackett Laboratory
Imperial College
London SW7 2BZ

Dr B Boys
Department of Biochemistry
University of Edinburgh
Hugh Robson Building
Edinburgh EH8 9XD

Dr G Barton
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Dr B Bax
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX

Mr H Berchtold
C/o Hoechst AG
HL III
G864
D-6230 Frankfurt/Main 80
Germany

Dr V Biou
EMBL
C/o ILL
BP 156
F-38042 Grenoble Cedex
France

Dr A C Bloomer
MRC Laboratory of Molecular Biology
Hills Road
Cambridge CB2 2QH

Professor T L Blundell
Birkbeck College
Malet Street
London WC1E 7HX

Dr L Brady
Department of Chemistry
University of York
Heslington
York YO1 5DD

Dr P Brick
Blackett Laboratory
Imperial College
Prince Consort Road
London SW7 2BZ

Dr G Bricogne
Laboratory of Molecular Biology
MRC
Hills Road
Cambridge CB2 2QH

Ms L Britton
Krebs Institute
University of Sheffield
Western Bank
Sheffield S10 2TN

Mr P Brownlie
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX

Professor A Brunger
Howard Hughes Medical Inst & Yale University
Department of Molecular Biophysics and Biochemistry
260 Whitney Avenue
New Haven
CT 06511
USA

Dr P K Bryant
The Wellcome Foundation Ltd
The Wellcome Research Laboratories
Langley Court
South Eden Park Road
Beckenham
Kent BR3 3BS

Mr A Buckle
Cambridge Centre for Protein Engineering
MRC
Hills Road
Cambridge CB2 2QH

Mr A D Cameron
Department of Chemistry
University of York
Heslington
York YO1 5DD

Dr J W Campbell
Daresbury Laboratory
Warrington WA4 4AD

Dr J Cavarelli
UPR de Biologie Structurale
15 Rue Rene Descartes
67084 Strasbourg
France

Mr A R Cervi
Department of Chemistry
University of Manchester
Manchester M13 9PL

Dr J N Champness
Wellcome Research Laboratories
B133
Park Langley
Beckenham
Kent BR3 3BS

Dr S Chaudhuri
Department of Chemistry
University of York
Heslington
York YO1 5DD

Mr G M T Cheetham
Department of Chemistry
University of Liverpool
PO Box 169
Liverpool L69 3BX

Mr Y W Chen
IRC
MRC Centre
Hills Road
Cambridge CB2 2QH

Mr R Chopra
Biophysics Group
Imperial College
Prince Consort Road
London SW7 2AZ

Dr M Churchill
Laboratory of Molecular Biology
MRC
Hills Road
Cambridge CB2 2QH

Dr A Cleasby
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Mr I J Clifton
Laboratory of Molecular Biophysics
University of Oxford
South Parks Road
Oxford OX1 3QU

Mr Y Corazza
Department of Biochemistry
University of Edinburgh
George Square
Edinburgh EH8 9XD

Dr S W Cowan
Biozentrum der Universität Basel
Department of Structural Biology
Klingelbergstrasse 70
CH-4056 Basel
Switzerland

Dr S J Crennell
Department of Biochemistry
University of Bath
Claverton Down
Bath BA2 7AY

Dr S Curry
AFRC Institute for Animal Health
Pirbright Laboratory
Ash Road
Pirbright
Woking
Surrey GU24 0NF

Mr P J Daly
Daresbury Laboratory
Warrington WA4 4AD

Dr Z Dauter
EMBL
C/o DESY
Notkestrasse 85
2000 Hamburg
Germany

Dr G Davies
Department of Chemistry
University of York
Heslington
York YO1 5DD

Dr C Davies
Department of Biochemistry
School of Medical Sciences
University of Bristol
Bristol BS8 1TD

Dr A De
Protein Structure Laboratory
Imperial Cancer Research Fund
44 Lincoln's Inn Fields
London WC2A 3PX

Dr V Dhanaraj
Birkbeck College
Malet Street
London WC1E 7HX

Professor O Dideberg
C/o CENG - DB MS.DIR
Avenue des Martyrs
BP 85X -38041
Grenoble Cedex
France

Ms E Dodson
Department of Chemistry
University of York
Heslington
York YO1 5DD

Professor G G Dodson
Department of Physics
University of York
Heslington
York YO1 5DD

Mr D Doyle
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX

Mr M Dreyer
Institut für Organische Chemie und Biochemie
Universität Frieburg
Albertstr 21
W-7800 Frieburg
Germany

Dr H Driessen
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX

Dr H Duggleby
Department of Chemistry
University of York
Heslington
York YO1 5DD

Miss E Duke
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Dr P Dumas
IBMC
Laboratoire de Cristallographie Biologique
15 Rue René Descartes
67084 Strasbourg Cedex
France

Ms K J Edwards
CRC Biomolecular Structure Unit
Institute of Cancer Research
Block F
15 Cotswold Road
Sutton
Surrey SN2 5NG

Mr J Emsley
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX

Mr R M Esnouf
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Dr P R Evans
MRC Laboratory of Molecular Biology
Hills Road
Cambridge CB2 2QH

Ms S M Fabiane
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX

Dr J C Fontecilla
DIEP/LCCP
CENG
BP 85X
F-38041 Grenoble Cedex
France

Dr S M Franken
Max Planck Institut fuer Med Forschung/Biophysik
Jahnstr 29
D-6900 Heidelberg
Germany

Dr P S Freemont
Protein Structure Lab
Imperial Cancer Research Fund
44 Lincoln's Inn Fields
London WC2A 3PX

Dr V Fülöp
Laboratory of Molecular Biophysics
University of Oxford
South Parks Road
Oxford OX1 3QU

Dr S Gamblin
Department of Biophysics and Molecular Biology
Harvard University
7 Divinity Avenue
Cambridge
MA
USA

Dr M Ghosh
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Dr P Fitzgerald
Merck Sharp and Dohme Research Laboratories
Room 203
PO Box 2000
Rahway
NJ 07065
USA

Dr G C Ford
Krebs Institute
University of Sheffield
Sheffield S10 2TN

Dr C Frazao
Centro Tecnologia Quimica e Biologica
Apartado 127
P-2780 Oeiras
Portugal

Dr E Fry
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Dr B Gallois
C/O Prof J R Helliwell
Department of Structural Chemistry
University of Manchester
Manchester M13 9PL

Dr E F Garman
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Mr N M Glykos
Department of Biochemistry and Molecular Biolog
University of Leeds
Leeds LS2 9JT

Mr J D Goldberg
Biophysics Group
Blackett Laboratory
Imperial College
Prince Consort Road
London SW7 2AZ

Dr B Golinelli-Pimpaneau
Laboratoire de Biologie Physicochimique
Bâtiment 433
Université Paris Sud
91405 Orsay Cedex
France

Ms A Gonzalez
Daresbury Laboratory
Warrington WA4 4AD

Ms E J Gordon
Department of Biochemistry
University of Edinburgh
Hugh Robson Building
George Square
Edinburgh EH8 9XD

Mr M A Gorman
Imperial Cancer Research Fund
44 Lincoln's Inn Fields
London WC2A 3PX

Mr P Gouet
CENG
BP 85X
38041 Grenoble Cedex
France

Dr S Gover
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Miss K Grabham
Krebs Institute
Department of Molecular Biology and Biotechnology
University of Sheffield
Western Bank
Sheffield S10 2UH

Ms S E Greasley
Department of Biochemistry and Molecular Biology
University of Leeds
Leeds LS2 9JT

Mr J M Grimes
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Dr C R Groom
Department of Biochemistry and Molecular Biology
University of Leeds
Leeds LS2 9JT

Dr J M Gulbis
Department of Chemistry
University of York
Heslington
York YO1 5DD

Mr N Guthrie
Lab 265
Department of Chemistry
University of Glasgow
Glasgow G12 8QQ

Dr J Hajdu
Laboratory of Molecular Biophysics
University of Oxford
South Parks Road
Oxford OX1 3QU

Mr K Håkansson
Department of Molecular Biophysics
Chemical Centre
PO Box 124
S-22100 Lund
Sweden

Dr Q Hao
Daresbury Laboratory
Warrington WA4 4AD

Dr K Harlos
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Dr G W Harris
AFRC Institute of Food Research
Reading Laboratory
Shinfield
Reading RG2 9AT

Ms D Harris
Department of Chemistry
University of Glasgow
Glasgow G12 8QQ

Mr S Harrop
Department of Chemistry
University of Manchester
Manchester M13 9PL

Professor E Harutyunyan
Institute of Crystallography
Academy of Scuences of the USSR
Leninsky pr 59
Moscow 117333
USSR

Professor S S Hasnain
Daresbury Laboratory
Warrington WA4 4AD

Dr P D Hempstead
C/o Dr P Artymiuk
Department of Molecular Biology and Biotechnology
University of Sheffield
Western Bank
Sheffield S10 2UH

Dr T Higgins
Department of Chemistry
University College
Galway
Ireland

Professor W G M Hol
Bioson Research Institute
Nijenborgh 4
9747 AG Groningen
The Netherlands

Dr P Holden
Department of Biochemistry
School of Medical Sciences
University of Bristol
Bristol BS8 1TD

Dr G Hope
MRC Institute of Virology
Church Street
Glasgow G11 5JR

Professor A Hordvik
IMR/University of Tromso
N-9000 Tromso
Norway

Ms A Houdusse
Institut Pasteur
Department d'Immunologie Structurale
28 rue de Dr Roux
75015 Paris
France

Dr W Hunter
Department of Chemistry
University of Manchester
Manchester M13 9PL

Dr J Husain
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX

Professor N Isaacs
Department of Chemistry
University of Glasgow
Glasgow G12 8QQ

Dr N Ito
Department of Biochemistry and Molecular Biology
University of Leeds
Leeds LS2 9JT

Dr P Jeffrey
Department of Macromolecular Crystallography
Bristol-Myers-Squibb Pharmaceutical Research Inst
PO Box 4000
Princeton
New Jersey 08543-4000
USA

Dr J Jenkins
AFRC Institute of Food Research
Reading Laboratory
Shinfield
Reading RG2 9AT

Dr H Jhoti
Glaxo Group Research
Protein Structures Group
Greenford Road
Greenford
Middlesex UB8 0HE

Dr J S Jiang
Department of Molecular Biophysics and Biochemistry
Yale University
260 Whitney Avenue
New Haven
CT 06511
USA

Mr J John
Department of Biochemistry
University of Bath
Claverton Down
Bath BA2 7AY

Dr Y Jones
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Prof A Jones
Department of Molecular Biology
Biomedical Centre
University of Uppsala
Box 590
S-75124 Uppsala
Sweden

Dr S Jordan
Department of Structural and Biophysical Chemistry
Glaxo Inc
5 Moore Drive
Research Triangle Park
North Carolina 27709
USA

Dr J Kallen
Sandoz Pharma AG
Bau 503/1208
CH-4002 Basel
Switzerland

Mr G J Keen
Department of Biophysics
King's College London
26-29 Drury Lane
London WC2B 5RL

Ms C Kisker
Institut für Kristallographie
Takustr 6
1000 Berlin 33
Germany

Dr S Knight
MRC Laboratory of Molecular Biology
Hills Road
Cambridge CB2 2QH

Professor M Kokkinidis
Institute of Molecular Biology and Biotechnology
PO Box 1527
GR-71110 Heraklion
Crete
Greece

Dr D Kostrewa
F Hoffman-La Roche Ltd
Pharmaceutical Research - New Technologies
CH-4002 Basel
Switzerland

Ms P R Kuser
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX

Mr J M Lally
Protein Structure Laboratory
Imperial Cancer Research Fund
44 Lincoln's Inn Fields
London WC2A 3PX

Dr V Lamzin
EMBL
c/o DESY
Notkestrasse 85
D-2000 Hamburg 52
Germany

Mr G M Langdon
Krebs Institute
University of Sheffield
Western Bank
Sheffield S10 2TN

Dr G Lange
Department of Chemistry
University of York
Heslington
York YO1 5DD

Dr B Langlois d'Estaintot
Laboratoire de Cristallographie
Université de Bordeaux I
351 Cours de la Libération
33405 Talence Cedex
France

Mr R L Larsen
Protein Crystallography Group
Department of Chemistry
University of Tromso
N-9000 Tromso
Norway

Mr G M Laughlan
Department of Protein Crystallography
Institute of Chemistry
University of Glasgow
Glasgow G12 8QQ

Professor W G Laver
John Curtin School of Medical Research
ANU
Canberra ACT 2601
Australia

Dr D M Lawson
Department of Chemistry
University of York
Heslington
York YO1 5DD

Ms S M Lea
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Dr G A Leonard
Department of Chemistry
University of Manchester
Manchester M13 9PL

Dr J Lescar
Dept d'Immunologie
Institut Pasteur
25 rue de Dr Roux
75724 Paris
France

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

Mr R J Lewis
Department of Chemistry
University of York
Heslington
York YO1 5DD

Dr J Li
MRC Laboratory of Molecular Biology
Hills Road
Cambridge CB2 2QH

Dr L Liljas
Department of Molecular Biology
Biomedical Centre
University of Uppsala
Box 590
S-751 24 Uppsala
Sweden

Dr P R Lindley
Daresbury Laboratory
Warrington WA4 4AD

Dr J Lisgarten
Institute voor Moleculaire Biologie
VUB
Paardenstraat 65
1640 Sint Genesius Rode
Brussels
Belgium

Dr J A Littlechild
Department of Chemistry
University of Exeter
Stocker Road
Exeter EX4 4QD

Dr A Littlejohn
Department of Chemistry
University of Glasgow
Glasgow G12 8QQ

Mr C D Livingstone
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Mr D T Logan
IBMC du CNRS
15 rue René Descartes
67084 Strasbourg Cedex
France

Dr B Luisi
MRC Virology Unit
Church Street
Glasgow G11 5JR

Dr T Lundqvist
MRC Laboratory of Molecular Biology
Hills Road
Cambridge CB2 2QH

Ms C Marks
Laboratory of Molecular Biology
MRC
Hills Road
Cambridge CB2 2QH

Mr A Mattevi
Bioson Research Institute
Nijenborgh 16
9747 AG Groningen
The Netherlands

Mr A S McAlpine
Department of Biochemistry
University of Edinburgh
George Square
Edinburgh EH8 9XD

Mr G McDermott
Department of Chemistry
University of Glasgow
Glasgow G12 8QQ

Dr P McLaughlin
MRC Laboratory of Molecular Biology
Hills Road
Cambridge CB1 3QH

Dr S McSweeney
Daresbury Laboratory
Warrington WA4 4AD

Dr P Metcalf
EMBL
Postfach 10.2209
D-6900 Heidelberg
Germany

Dr V Mikol
Sandoz Pharma AG
Bau 503/1208
CH-4002 Basel
Switzerland

Dr M Milburn
Department of Structural and Biophysical Chemistry
Glaxo Inc
5 Moore Drive
Research Triangle Park
North Carolina 27709
USA

Dr P C E Moody
Department of Chemistry
University of York
Heslington
York YO1 5DD

Dr M H Moore
Department of Chemistry
University of York
Heslington
York YO1 5DD

Mr J Morais Cabral
Department of Biochemistry
University of Edinburgh
Hugh Robson Building
George Square
Edinburgh EH8 9XD

Dr E Morgunova
EMBL
C/o DESY
Notkestrasse 85
2000 Hamburg
Germany

Dr D S Moss
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX

Mr A Mueller
C/o Professor Saenger
Institut fur Kristallographie
Takustrasse 6
W-1000 Berlin 33
Germany

Dr H Muirhead
Department of Biochemistry
University of Bristol
Bristol BS8 1TD

Mrs L M Murphy
Daresbury Laboratory
Warrington WA4 4AD

Dr J Murray-Rust
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX

Ms R S Nagasuma
C 52A
Department of Biochemistry
School of Medical Sciences
University of Bristol
Bristol BS8 1TD

Mr J H Naismith
Department of Chemistry
University of Manchester
Manchester M13 9PL

Dr J Navaza
Laboratoire de Physique
Centre Pharmaceutique
Rue J B Clement
99290 Chatenay-Malabry
Paris
France

Dr C Nave
Daresbury Laboratory
Warrington WA4 4AD

Dr M Neu
Daresbury Laboratory
Warrington WA4 4AD

Mr M E Noble
EMBL
Meyerhofstr 1
D-6900 Heidelberg
Germany

Professor A C T North
Department of Biochemistry and Molecular Biology
University of Leeds
Leeds LS2 9JT

Mrs R Nunn
Krebs Institute
Department of Molecular Biology and Biotechnology
University of Sheffield
Western Bank
Sheffield S10 2TN

Mr B O'Hara
Laboratory of Molecular Biology
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX

Dr V Oganesyan
Institute of Crystallography
Academy of Sciences of the USSR
Leninsky pr 59
Moscow 117333
USSR

Dr D Ogg
SYMBICOM AB
Glunten 751 83
Uppsala
Sweden

Dr S Onesti
Biophysics Group
Imperial College of Science & Technology
Prince Consort Road
London SW7 2AZ

Dr Z Otwinowski
Howard Hughes Medical Institute and Yale University
Department of Molecular Biophysics and Biochemistry
260 Whitney Avenue
New Haven
CT 06511
USA

Dr M Papiz
Daresbury Laboratory
Warrington WA4 4AD

Mr E Passalacqua
University of Bath
Claverton Down
Bath
Avon

Dr R A Pauptit
EMBL
Meyerhofstrasse 1
D-6900 Heidelberg
Germany

Ms D H Peapus
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX

Mr G Pflügl
Biozentrum
University of Basel
Klingelbergstrasse 70
CH-4056 Basel
Switzerland

Mr C Phillips
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Buildings
South Parks Road
Oxford OX1 3QU

Dr R P Phizackerley
Stanford Synchrotron Radiation Laboratory
PO Box 4349
Stanford University
Stanford
California 94-309-0210
USA

Dr R Pickersgill
AFRC Institute of Food Research
Reading Laboratory
Shinfield
Reading RG2 9AT

Dr K Polyakov
Institute of Crystallography
Academy of Sciences of the USSR
Leninsky pr 59
117333 Moscow
USSR

Mr J P Porter
Department of Biochemistry
School of Medical Sciences
University of Bristol
Bristol BS8 1TD

Dr J P Priestle
Ciba Geigy AG
Biotechnology K-681.5.43
CH-4002 Basel
Switzerland

Dr J Raftery
Department of Chemistry
University of Manchester
Manchester M13 9PL

Dr V Ramakrishnan
Laboratory of Molecular Biology
MRC
Hills Road
Cambridge CB2 2QH

Dr A Rawas
Department of Biochemistry
School of Medical Sciences
University of Bristol
University Walk
Bristol BS8 1TD

Dr B Rees
IBMC
Laboratoire de Cristallographie Biologique
15 Rue René Descartes
67084 Strasbourg Cedex
France

Dr D W Rice
Krebs Institute
Department of Molecular Biology and
Biotechnology
University of Sheffield
Western Bank
Sheffield S10 2TN

Dr H R Powell
AFRC Institute of Food Research
Reading Laboratory
Shinfield
Reading RG2 9AT

Dr J B Rafferty
Department of Biochemistry and Molecular Biology
University of Leeds
Leeds LS2 9JT

Mr J Rahuel
C/o Ciba Geigy AG
Biotechnology K-681.5.03
CH-4002 Basel
Switzerland

Dr Z Rao
Laboratory of Molecular Biology
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Mr M G Redshaw
Department of Chemistry
University of Manchester
Oxford Road
Manchester M13 9PL

Mr J Ren
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Dr P J Rizkallah
Daresbury Laboratory
Warrington WA4 4AD

Dr C Rojas
Department of Biochemistry and Molecular Biology
University of Leeds
Leeds LS2 9JT

Dr M J Romão
Max Planck Institut für Biochemie
AM Klopferspitz
8033 Martinsried
B-München
Germany

Professor M G Rossman
Department of Biological Sciences
Purdue University
1392 Lilly Hall of Life Sciences
West Lafayette
IN 47907-1392
USA

Mr P Rowland
Labaoratory of Molecular Biophysics
University of Oxford
Rex Richards Buildings
South Parks Road
Oxford OX1 3QU

Mr R B Russell
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Dr M R Sanderson
CRC Biomolecular Structure Unit
Block F
Institute of Cancer Research
15 Cotswold Road
Sutton
Surrey SM2 5NG

Mr U Sauer
EMBL
Meyerhofstr 1
D-6900 Heidelberg
Germany

Dr L Sawyer
Department of Biochemistry
University of Edinburgh
George Square
Edinburgh EH8 9XD

Mr H Schindelin
Institut für Kristallographie
Takustr 6
1000 Berlin 33
Germany

Dr T Schirmer
Biozentrum der Universität Basel
Department of Structural Biology
Klingelbergstrasse 70
CH-4056 Basel
Switzerland

Mr T Schneider
EMBL
C/o DESY
Notkestrasse 85
2000 Hamburg 52
Germany

Dr H A Schreuder
Bioson Research Institute
Nijenborg 4
9747 AG Groningen
The Netherlands

Mr E Schröder
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Dr J W R Schwabe
Laboratory of Molecular Biology
MRC
Hills Road
Cambridge CB2 2QH

Dr T Simon
Cambridge Centre for Protein Engineering
MRC Centre
Hills Road
Cambridge CB2 2QH

Mr A A Simpson
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX

Dr I M Sinning
Department of Molecular Biology
BMC
University of Uppsala
Box 590
S-75124 Uppsala
Sweden

Dr T Skarzynski
Blackett Laboratory
Imperial College
London SW7 2BZ

Dr J Skelly
CRC Biomolecular Structure Unit
Institute of Cancer Research
Block F
15 Cotswold Road
Sutton
Surrey SN2 5NG

Dr J M A Smith
Department of Molecular Biology and
Biotechnology
PO Box 594
University of Sheffield
Western Bank
Sheffield S10 2UH

Dr C F Snook
Institute of Cancer Research
15 Cotswold Road
Sutton
Surrey SM2 5NG

Dr M K Sohi
Department of Biophysics
King's College London
26-29 Drury Lane
London WC2B 5RL

Professor P Spadon
Department of Organic Chemistry
Padova University
Via Marzolo 1
35131 Padova
Italy

Mr N Spink
CRC Biomolecular Structure Unit
Block F
Institute of Cancer Research
15 Cotswold Road
Sutton
Surrey SN2 5NG

Dr W Steigemann
Max-Planck-Institut für Biochimie
Computer Centre
D-8033 Martinsried
Germany

Dr T J Stillman
Krebs Institute
Department of Molecular Biology and
Biotechnology
University of Sheffield
Western Bank
Sheffield S10 2TN

Dr R Strange
Daresbury Laboratory
Warrington WA4 4AD

Mr S Strathdee
Department of Biochemistry and Molecular Biology
University of Leeds
Leeds LS2 9JT

Dr D Stuart
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Dr C G Suresh
Department of Chemistry
University of York
Heslington
York YO1 5DD

Dr L A Svensson
Department of Molecular Biophysics
Chemical Centre
University of Lund
S-22100 Lund
Sweden

Dr J Tame
Department of Chemistry
University of York
Heslington
York YO1 5DD

Ms M M G Thunnissen
Department of Chemical Physics
University of Groningen
Nyenborg 16
9747 AG Groningen
The Netherlands

Mr H Tsuge
Japan Tobacco Inc
Life Science Research Laboratory
6-2 Umegaoka
Midori-ku
Yokohama
Kanagawa 227 Japan

Mr J Turkenburg
Department of Chemistry
University of York
Heslington
York YO1 5DD

Mr X Su
Department of Molecular Genetics
Krolinska Institutet
Box 60400
10401 Stockholm
Sweden

Dr B J Sutton
Department of Biophysics
King's College London
26-29 Drury Lane
London WC2B 5RL

Dr L Tabernero
Department of Macromolecular Crystallography
Bristol-Myers Squibb
PO Box 4000
Princeton
NJ 08543-4000
USA

Dr G L Taylor
Department of Biochemistry
University of Bath
Claverton Down
Bath BA2 7AY

Dr I J Tickle
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX

Mr A Tucker
ICI Pharmaceuticals
Chemistry Department
Alderley Park
Macclesfield
Cheshire

Mrs M G W Turkenburg
Department of Chemistry
University of York
Heslington
York YO1 5DD

Dr M Turner
Department of Biochemistry
University of Edinburgh
Hugh Robson Building
George Square
Edinburgh EH8 9XD

Dr A Urzhumtsev
IBMC
Laboratoratoire de Cristallographie Biologique
15 Rue René Descartes
67084 Strasbourg Cedex
France

Mr J Vidgren
Orion Pharmaceutica
Research Centre
Box 65
SF-02101 Espoo
Finland

Dr D A Waller
Department of Biochemistry and Molecular Biology
University of Leeds
Leeds LS2 9JT

Mr T Wan
Department of Biophysics
King's College London
26-29 Drury Lane
London WC2B 5RL

Miss M L Waugh
Krebs Institute
Department of Molecular Biology and
Biotechnology
University of Sheffield
Western Bank
Sheffield S10 2UH

Mr S A Weston
EMBL
Meyerhofstr 1
D-6900 Heidelberg
Germany

Mr M J Uppenberg
Department of Molecular Biology
Biomedical Centre
Box 590
75124 Uppsala
Sweden

Dr N Veerapaneni
Department of Crystallography
Birkbeck College
Malet Street
London WC1E 7HX

Dr S Wakatsuki
Laboratory of Molecular Biophysics
University of Oxford
South Parks Road
Oxford OX1 3QU

Mr M Walsh
Department of Chemistry
University College
Galway
Ireland

Dr H C Watson
Department of Biochemistry
School of Medical Sciences
The University
Bristol BS8 1TD

Ms S Weisgerber
Department of Chemistry
University of Manchester
Oxford Road
Manchester M13 9PL

Dr S W White
Department of Microbiology
Box 3020
Duke University Medical Centre
Durham
North Carolina 27710
USA

Miss J L Whittingham
Department of Chemistry
University of York
Heslington
York YO1 5DD

Dr R Wierenga
EMBL
Meyerhofstr 1
D-6900 Heidelberg
Germany

Dr D B Wigley
Department of Chemistry
University of York
Heslington
York YO1 5DD

Mr T Wilkinson
ICI Pharmaceuticals
Chemistry Department
Alderley Park
Macclesfield
Cheshire

Dr K S Wilson
EMBL
C/o DESY
Notkestrasse 85
2000 Hamburg 52
Germany

Dr W Wolf
Daresbury Laboratory
Warrington WA4 4AD

Dr A Wonacott
Glaxo Group Research
Protein Structure Group
Greenford Road
Greenford
Middlesex UB6 0HE

Ms K A Woods
Laboratory of Molecular Biophysics
University of Oxford
Rex Richards Building
South Parks Road
Oxford OX1 3QU

Dr B Xiao
Department of Chemistry
University of York
Heslington
York YO1 5DD

Dr S J Yewdall
Department of Molecular Biology and Biotechnolog
University of Sheffield
PO Box 594
Western Bank
Sheffield S10 2TN

Mr R Young
Department of Biophysics
King's College London
26-29 Drury Lane
London WC2B 5RL

Mr J Zou
Department of Molecular Biology
Uppsala University
BMC
Box 590
S-751 24 Uppsala
Sweden