DL/SCI/R35

FROM FIRST MAP TO FINAL MODEL

Proceedings of the CCP4 Study Weekend,
6-7 January 1994

Compiled by
S. Bailey, R. Hubbard and D. Waller

EPSRC

DARESBURY LABORATORY
Daresbury, Warrington WA4 4AD

FROM FIRST MAP TO FINAL MODEL

Proceedings of the CCP4 Study Weekend
6-9 January 1994

Compiled by:

Susan Bailey, Daresbury Laboratory
Rod Hubbard, University of York
David Waller, University of Leeds

EPSRC

Daresbury Laboratory
1994

CONTENTS

INTRODUCTION

This year's CCP4 Study Weekend was different in three important ways. First the venue was moved to Chester College, second it wasn't held on the usual Friday and Saturday at the end of January and third the topic From First Map to Final Model was something of a departure from previous meetings which have tended to concentrate on a single small area of protein crystallography.

The move to Chester College was a result of the popularity of the meeting, with over 300 participants, coupled to our requirements for top quality audiovisual facilities. This year the college could only accomodate us on the Thursday and Friday before their winter term began.

When Working Group 2 sat down to discuss possible topics for the meeting the subject of refinement was mentioned and after giving some thought to the matter we decided that a wider survey of the whole matter of taking your first electron density map, building a model into it, refining the model and considering errors that may arise in the model would make a useful topic for a meeting. Of course this topic touches on a whole range of matters which have been covered by previous meetings but there have been a number of advances notably in the software available to perform these tasks. In addition CCP4's brief includes the education of inform young crystallographers coming into the field and this area had not been covered for some time.

The meeting was organised and supported by the SERC Collaborative Computational Project in Protein Crystallography (CCP4) and the EC Human, Capital and Mobility Scheme. We wish to thank the invited speakers for their presentations and for their cooperation in providing manuscripts for 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 Val Matthews and Cheryl Stonier who ensured that the meeting ran smoothly. In addition the proceedings owe much to the efforts of Mel Davies and his staff.

Sue Bailey
Rod Hubbard
David Waller

June 1994

ACKNOWLEDGMENTS

CCP4 would like to thank the EC Human Capital and Mobility Scheme for the provision of funding allowing 43 young European scientists to attend the meeting.

CCP4 would also like to thank the following companies for their financial contributions to the CCP4 project in the year 1993. This support was an essential contribution to the costs of the 1994 CCP4 meeting.

Abbott Laboratories
Amgen Incorporated
Banyu Pharmaceutical Company Limited
Ciba-Geigy Limited
Farmitalia Carlo Erba
Glaxo UK Ltd.
Hoescht AG
Hoffmann - La Roche and Co.
Kabi Pharmacia
M.M. Dow
Pfizer Ltd.
Sandoz Pharma AG
Schering
Sumitomo Chemical Company Limited
Symbicom AB
Syntex (USA) Incorporated
Vertex Pharmaceuticals
The Wellcome Foundation Limited
Zeneca Pharaceuticals

Making the First Trace with O

T.A.Jones[1] & M.Kjeldgaard[2]

1. Department of Molecular Biology,
BMC, Box590,
S-55124 Uppsala, Sweden

2. Department of Chemistry,
University of Aarhus,
DK-8000 Aarhus, Denmark.

Introduction

The process of building a model from the initial electron density map is an important part of solving an x-ray structure. Up to this point, the diffraction data has been treated in a systematic fashion by computer programs. Reflections have been collected, merged, and scaled, and perhaps density modification/averaging methods may have been employed. Some judgement may have been used to reject outlier reflections, but all reflections will have been subjected to the same rejection criteria. Likewise, the creation of a mask for averaging may have introduced some subjectivity into the process but each single grid point in the Fourier maps will have been treated by the same algorithm. In the model building stage, however, the scientist no longer uses this systematic approach, but starts to incorporate his or her *interpretation* of the experimental results, trying to make chemical and biological sense of the shapeless blobs of electron density displayed on the graphics workstation. In this stage, errors are easily introduced into the structure solution process (Brändén & Jones, 1990) and such errors are not easy to get rid of again.

The O program system does not take the subjectivity out of map interpretation. Rather, it supplies a set of tools to carry out a systematic way of working with the electron density. This strategy, Figure 1, is based on the observation that 3-dimensional visualization is complicated and computer graphics tools are needed to show different representations of the data, and to manage ideas and theories that suddenly appear in the mind. In our experience, such ideas often disappear as quickly as they appear. Thus,

what is needed is a three-dimensional *scratchpad*, where the researcher can maintain ideas of the folding and other features of the macromolecule. We have previously described some aspects of this system (Jones *et al.*, 1990), in particular the generation of main chain atoms from Cα guide points using the fragment approach of Jones & Thirup (1986), the use of side chain rotamers, real-space optimization of a rough model to the density, and a residue based index of fit between model and electron density map (Jones *et al.*, 1991). In this paper, we describe in more detail some of the problems associated with map interpretation, how we work with skeletons, and how we decide where the sequence matches the density.

Pass	Description
1	Determine molecular boundary
2	Determine fold. Associate sequence with density. Solve the structure.
3	Assign Ca positions.
4	Build main chain using **Lego_auto_mc**. Fix and monitor mistakes with **Lego_ca** and **Pep_flip**.
5	Generate side chains with **Lego_auto_sc**, and **RSR_rotamer.** Check the fit of each residue. Use **Pep_flip, RSC_fit** and **RS_fit** to monitor and correct mistakes. Try to locate and correct out-of-register errors.

Figure 1. Steps in Building a Model in an Electron Density Map.

In O, we make use of a simplified representation of the electron density as a starting point for the 3-dimensional scratchpad. As well as giving a suitable data structure to manipulate, this representation allows one to get an overview of the electron density map that is not possible in the usual contour representation. We employ the skeletonization algorithm of Greer (1974) with an extension that allows us to associate different levels with the skeletons. Each level is characterized in O by a separate colour, and each colour can be given a different meaning by the crystallographer. At the outset, the skeleton atoms are classified as 'main chain' or 'side chain'. Of course, these categories are not necessarily correct, but are simply the programs' best guess based on the connectivity of the skeleton. O allows the user to manipulate the skeleton, i.e. to change the level classification (colour), the connectivity, and positions of the individual skeleton atoms. The program provides options for filtering and improving the skeleton automagically. The user can place pieces of 3-D text as reminders and labels in the

space of the map that can later aid in the task of deciding where the sequence matches the density. Tools are provided to compare the real sequence with guesses that are made based on the shape of the density.

Thus, the first step in map interpretation involves working with the skeleton. Having to deal with residues, sidechains, atoms and what-not is actually only an added complexity at the early stages of map interpretation. It slows you down, and removes your attention from the primary goal: to trace the polypeptide chain. In the following we will expand on the first-map-to-final-model process, using the fundamentalist approach of **O**. We will introduce the process in a number of passes through the skeleton. We will not describe the program itself; a basic knowledge of **O** is assumed.

Passes 1 and 2: Editing of Skeletonized Maps

The initial skeleton is calculated from an electron density map with the `bones` program. Two parameters control the appearance and connectivity of the resulting skeleton, *base* and *sigma*. Typical values of these parameters are 1.25 and 1.0 times the standard deviation of the map, respectively, but several combinations should be tried. If too many connected skeleton atoms are seen, *base* should be increased, and if too few connected skeleton atoms, the *base* level should be reduced. Users with alternative skeletonization algorithms need only adopt the final data structures, Figure 2, that are needed by **O**.

datablock name	type	size	description
_atom_xyz	r	3n	coordinates of skeleton atoms
_atom_bone	i	n	level of skeleton atom
_connectivity	i	m	connectivity of skeleton atoms
_atom_colour	i	n	colour code of skeleton atom
_residue_name	c	1	residue name of skeleton, usually 'BONES'
_residue_type	c	1	residue type of skeleton, usually 'SKL'
_residue_pointers	i	2	pointer to first and last residue
_atom_visible	i	n	visibility status of skeleton atoms

Figure 2. Data structure of the skeletonized density, assuming n skeleton atoms and m connections. All skeleton atoms belong to just one enormous residue.

Any number of different objects can be made from the skeleton. We advocate the use of at least two objects, one showing the current main chain trace over a large volume of space (maybe coloured to show different degrees of confidence), and the

second containing all main chain and side chain levels within a smaller volume. The first object would be used to get overview information, while the second would be the working object where actual changes are made to the skeleton. Suitable macros for the creation of these objects are shown in Figure 3 and would result in the creation of objects **50a** and **20a**. Since the screen menu in **O** is adjustable, we would recommend the use of the one shown in Figure 4. The complete set of commands associated with the **Bones** menu is shown in Figure 5.

The *first pass* is to determine the molecular boundary. This can be difficult sometimes, especially if chains form dimers etc. The most likely mistake is the possibility of moving into the 'wrong' molecule. This may be related to the 'correct' molecule by crystallographic or non-crystallographic symmetry. In **O**, we use the **Symmetry_object** command to create an object containing the symmetry mates of the current trace object. During the first pass, we use **Bond_break** to disconnect our molecule from copies related by symmetry. If the asymmetric unit contains non-crystallographic symmetry, a mask can be built while looking at the skeleton atoms using a macro to zap points close to an identified skeleton atom (Kleywegt & Jones, 1994) . If the non-crystallographic symmetry operators are known, they can be applied to copies of the trace or mask object. The improved mask can also be used for additional steps of non-crystallographic symmetry averaging.

(a)

```
bone_setup anol 20a 20.    3 2 1 ;
message Id an atom
centre_id wait_id
bone_draw
```

(b)

```
bone_setup anol 50a 50.    1 3 4  ;
message Id an atom
centre_id wait_id
bone_draw
```

Figure 3. **O** macros to display skeletons. (a) Draw all bones atoms of type 3, 2 and 1 (main chain, side chain, and working estimate of the trace) within 20 Å of an ID'ed atom. This creates an atomic object called `20a`, choosing skeleton atoms, from bones molecule `anol`. (b). Draw skeleton atoms of type 1, 3 and 4 (level 4 is here used to designate neighboring molecules), within a 50Å sphere. This generates the 'overview' object, called `50a`.

Having determined the molecular boundary, we move on to *pass two*, where we attempt to recognize and correct local errors in the skeleton. By editing the skeleton, we refine our impression about the folding and appearance of the protein. Redefining a segment of 'side chain' level bones to 'main chain' will change the appearance in the overview object, and vice versa. We start by looking at the overview object. We try to recognize a strand or helix and follow along the trace (looking at the contoured density and making any changes to the detailed skeleton object) until you don't know what to do, at which point we redefine the skeleton level to indicate our path through space. Then after looking at the overview, we try to recognize a new strand or helix and repeat the process. Secondary structures are a great help in deciding where the side chains are or should be pointing — this can help when the density fades. When in doubt, carry along in the direction of the secondary structure. In several instances, the density (and hence the skeleton) will appear continuous were you don't expect it, so watch out for: hydrogen bonds between strands, interacting side chains that make the density continuous, and di-sulphide bridges. Make notes using **O**'s ODL graphics language (Figure 5), and don't forget to save backup copies of your edited skeleton's connectivity rather often! You may want to backtrack and restart at an earlier stage. It is also advisable to save regularly after every change and to backup the complete user database every few hours.

During the second pass, we are trying to recognize local protein-like features in the skeleton (of course, anything spotted while contemplating the skeleton in pass 1 would have been worked on, and corrections made). We start to apply our more detailed knowledge about protein structures in general, and *this protein in particular*. Does it have a ligand? Does it have special or glycosylated side chains? Does it have sequence homology with a structure which has already been solved? Blatant mistakes have been made ignoring facts of this kind. Can one recognize super-secondary structural units, such as βαβ's, TIM folds etc. It could be useful to interrogate the PDB, for example using the Dejavu program (Kleywegt & Jones, 1994). Try to recognize domains, and separate folding units.

Use colour to paint your trace according to any scheme that may be helpful to you and your co-workers.

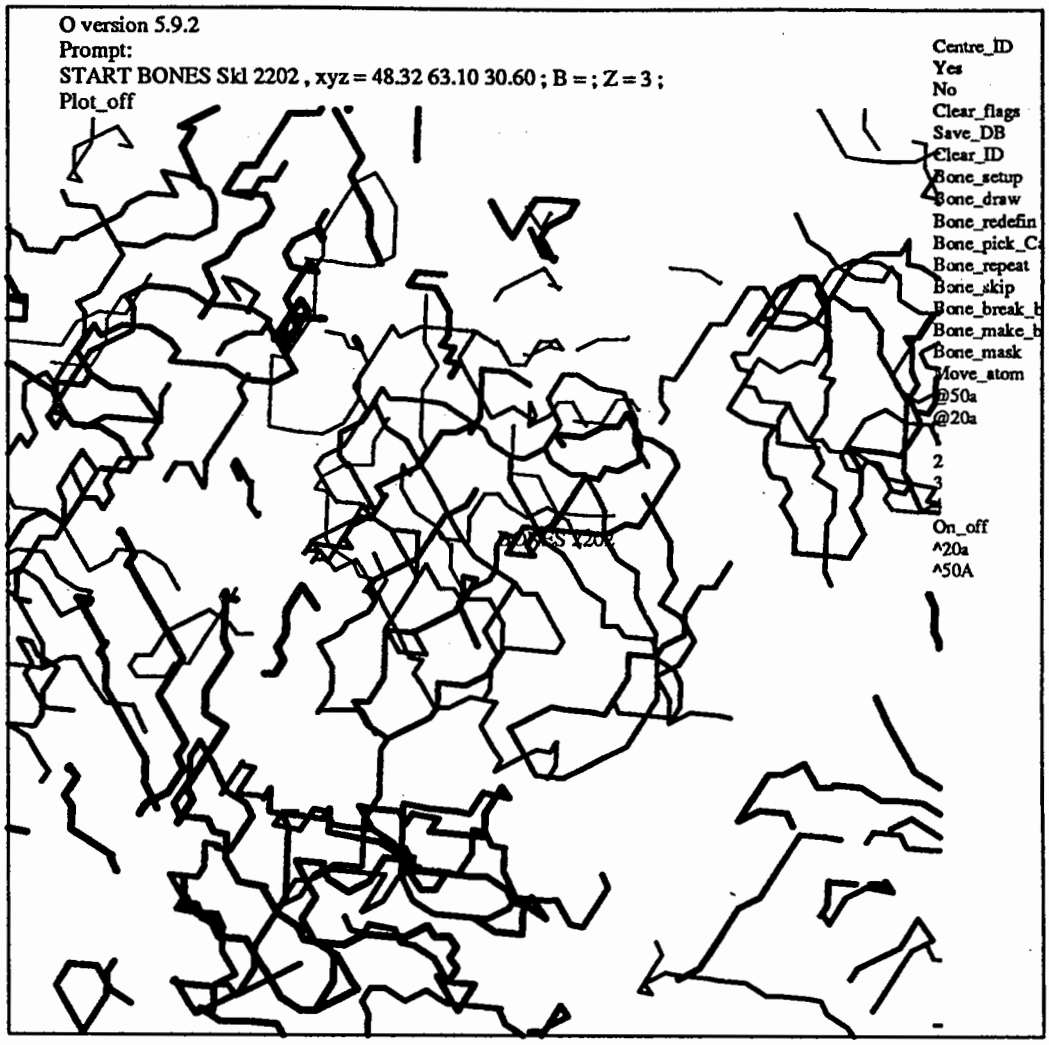


Figure 4. Example of the display screen when working with a skeleton. The un-edited overview object is shown. In the centre, one recognizes a single molecule; the neighboring molecules are also seen. At the bottom of the central molecule, a fake connection through a side chain density to another molecule is seen. At the right side, the menu has been customized to working with skeletons. The two macro's from figure 3 can be called by clicking at their names. The numbers 1–4 are also present on the screen menu, for easy redefinition of bones levels with **Bone_redefine**. Present on the screen menu is also **Move_atom**. The skeleton depicted is from P2 myelin (Jones *et al.*, 1988).

A commonly made mistake is reversing the chain direction in parts of the structure. This evidently leads to further mistakes, when the structural segments are connected. In maps calculated from high resolution data and good phases (~2.5Å or better), each secondary structural type has it's own fingerprint in the density that gives hints about

the local chain direction. Most obviously, the α-helix resembles a Christmas tree, when viewed with the N-terminal end down, and the C-terminal end up. In β-strands, the side chains have the characteristic up-down-up-down branching pattern, and the distances between carbonyl oxygen bumps in the density, and side chain branches, give information on the local chain direction. In β-sheets, subtle differences in the relative directions of side chains and carbonyl bumps on adjacent strands hint parallel vs. antiparallel sheet structure.

Gradually, sequence information can be included as one attempts to tie stretches of skeleton to pieces of the sequence. Often large aromatic sidechains such as tryptophans will stick out like beacons in the density. One can also make use of heavy atom binding sites from the derivatives that bind to characteristic residue types. Other information may come from studies on which residues make up the active site or other characteristic features. If the protein contains several domains, one needs to determine the relative order of domains along the sequence. Try to identify tertiary structure motifs, but watch out for variations from the 'classic'! More information generates more restraints on the connectivity of the segments. All observations can be recorded using the 3-D text facility of **O**, and/or using colours to mark the skeleton.

```
begin Notes
text 98.7 -13.3 36.7 Thr-94
text 98 -10 34 Gly-95
text 96 -11 27 Met-99
text 95 -13 32 Gln-98
text 99 -13 31.7 Ala-96
text 97 -15.3 32.7 Ala-97
end_object
```

Figure 5. Example of a 'Notes' file, using the ODL descriptor language. The file is edited by a text editor, and must be redrawn with the **Draw** command every time it has been changed.

At the end of pass two, the folding hypothesis should be evaluated: does it look like a protein? There are several items on the checklist:

- Long stretches of secondary structure are good. Beta turns in interior of protein are bad.
- Globularity of proteins: polar sidechains in the interior of the protein, that are not involved in polar (or salt bridge) interactions are bad. Hydrogen bonds are good: in high resolution structures only 10% of main chain N and O in the interior are not involved in H-bond interactions (Thornton *et al.*, this book). Hydrophobic clusters

are good. Aromatic clusters in the interior are very good, look for 'propeller' configuration.
- Interaction of non-local side chains: does it make chemical sense? Polar against nonpolar is bad.
- Distribution of particular side chains: proline residues are expected mostly at the surface, and mostly in loops and turn structure. Non-polar residues in the interior, polar residues on the surface.
- Does it look like your protein? Be merciless, you should be able to explain biology here.
  - ◊ Compare the structure with related sequences. Could you build one of the homologous sequences into your structure, and still make sense? Insertions/deletions almost always occur in loop regions.
  - ◊ Are the active site residues the ones you expect?
  - ◊ Expected di-sulphide pairings?
- Expect the unexpected!

Did your structure score OK? Congratulations, you may have solved it!

Notice, that in pass two, we have for the first time in the structure solution process included information that does not stem from the crystallographic experiment *per se*. We have used results from chemistry (on the chemical structure of peptides and side chains), we have used information from other protein crystallographic experiments (on the structure of other proteins, of α-helices and β-strands, and of general folds), and we have used information from biology (the sequence and perhaps function and cofactors of the protein). All this additional information is gradually included, in the order of increasing information content: chemistry → crystallography → biology.

Pass 3: Assign Cα positions from skeleton

At some stage in the process, it will be necessary to build at least a part of a protein molecule. This may not necessarily be part of the unknown structure but could be just a a poly-alanine structure that would be available for experimenting. Actually deciding where the sequence of your structure matches the bumps in the electron density is the hardest part of the building process, for even the most experienced crystallographer. Up to now, the tools available for this have been purely qualitative.

At the most basic level, the crystallographer scans the sequence and compares it against the size, shape and position of the observed electron density bumps. **O** offers some advance on the qualitative placement of the sequence, namely the **Slider** commands (Figure 6). The estimated local sequence is compared to the actual sequence

via a scoring matrix that has been earlier loaded into the program. The user's estimate usually looks like the one letter amino acid code but could actually be something else. For example, a classification based on the observed length of density protruding from the main chain could use the characters *s, m & l* for short, medium and long residues, respectively, with a suitable set of scores for the 20 amino acids. The crystallographer may like to study the scoring matrix to understand the basis for the score before it is actually used. The estimated match to the sequence can be stored in the O database and associated with a piece of structure. This allows one to see the guess and likely sequences close to the position of the density. It may be helpful when deciding one's guess to use an option, **Slider_lego**, that each of the twenty amino acids on the screen together with their preferred rotamers. This option requires that you have built at least a poly-alanine model through the density.

• guess	Enter a guess of the sequence, a comparison to the real sequence is made via a scoring matrix. The results can be associated with a poly-alanine chain and stored in the database. The top scores can be seen, displayed on the poly-alanine chain
• show	Show where the current named guesses are located in space
• display	Retrieve and display a past guess that has been stored in the database
• combine	Combine any number of guesses. The spacing between guesses must be specified as a maximum loop size, thus acting as an important constraint.
• lego	See how each of the possible 20 amino acids looks in the density. Change rotamers if applicable.
• calc	Calculate how each amino acid scores in a poly-alanine chain. Put score into slider data structures. Currently under development by Jin-yu Zou & T.A.Jones

Figure 6. Slider commands in O.

To be sure of getting the correct sequence as the top fit, requires that one searches for a long stretch of consecutive residues. Increasing the length of the stretch, increases the likelihood of an error due to missing insertions or deletions. One therefore works with a number of smaller connected pieces, using the **Slider_loop** command and specifying a missing 'loop' between each guess. Still errors will be made in making the correct assignment of the sequence to the density and they can be due to a number of things:

1. Directionality of the chain.
2. The quality of the density depends on where it is in the structure so that a tryptophan density could look like a glycine. Fortunately, a glycine density cannot look like a tryptophan.

3. Phase errors.
4. Limited resolution.

Just as large aromatic residues can often be clear markers for assignment of the sequence, it is equally important that glycine residues show no extraneous electron density. Unfortunately, glycines are often in loops that are usually external and often correspond to regions with the worse density. Lucky is the crystallographer who has lots of glycine residues in the middle of some solid density.

The crystallographer must beware of other sources of possible error. Very often, for example, glycosylation sites may be unknown. N-linked sites can often be guessed before hand, but O-linked sites are usually totally unexpected. In one case, one of us (TAJ) spent a considerable amount of effort trying to get an indole ring into such a site, resulting in an initial model that was out of register by a few residues. The structure may also contain unexpected ligands that show up with very strong electron density. The peptide ligand of HLA is a particularly good example (Bjorkman *et al.*, 1987) and in our own work the binding of the fatty acid to P2 myelin had very strong density, producing spaghetti in the first skeletons (Jones *et al.*, 1988).

The **Trace** commands, Figure 7, have been developed to speed up the assignment of Cα positions and to detect possible errors in the skeleton. These commands aim to make a skeleton look more like a Cα trace. They work by applying a set of filters to an input skeleton and generating a new skeleton with modified connectivity. For very good quality maps, the **Trace** commands can be used to more or less automatically build large parts of a structure. In a test on P2 myelin, **Trace** generates without user intervention 127 skeleton atoms, of which 115 can be fitted to the 131 Cα atoms in P2 with an r.m.s. deviation of 1.19Å. The commands can also be used with a skeleton that has been been manually edited to remove errors and to place branch points at likely Cα atoms. We also plan to make interactive versions of some of these filters.

At present, there are two ways to make a Cα trace. The most direct method uses a skeleton and the **Bone_pick_Ca** to position each Cα atom. The coordinates of the next identified atom are then associated with the Ca atom of the next residue. Other options (**Bone_repeat** and **Bone_skip**) allow some flexibility to make corrections and jump over regions.

• setup	
• sphere	remove connectivity outside some defined radius. This command reduces the number of skeleton atoms that have to be considered by other trace commands and, therefore, speeds them up.
• prune	keep the connectivity just between those atoms that have at least 3 branch points. The series of linked lines that may have existed between branched atoms are replaced by a single line. Unless side chain atoms are also involved in a branch, this command removes all side chain atoms from the connectivity. The new connectivity may contain some very long lines, if branch points were missing.
• fill_Ca	place Ca atoms along the trace at suitable spacings. This command adds back likely Ca atoms after a pruning operation. The resulting connectivity will have the added atoms roughly 3.8Å apart.
More filters are being planned:	
• loose_ends	add non-side chain loose ends back into the connectivity.
• carbonyls	remove likely carbonyl oxygens. This command will only be useful in higher resolution, well phased electron density maps.
• YASSPA	identify & optimize a/b units
• Deja-vu	search for similar structures in the Protein Data Base using only skeleton atoms (Kleywegt & Jones, 1994)
• wrong_SC	identify possible side chain branching errors
• merge_Ca	merge the skeleton Ca trace into the Ca's of a molecule

Figure 7. Trace commands.

The second method uses the **Baton** commands to build a poly-alanine chain. A di-peptide can be moved around, pivoting on the Ca atom of the 1st residue. When the user is satisfied, the main-chain atoms of the 2nd residue in the di-peptide are merged into the coordinates of the new structure and the di-peptide jumps forward ready for the next residue. One is free to step forwards or backwards through the sequence. The starting orientation of the di-peptide is controlled by a mode switch that has four settings. When set to 'alpha', the growing chain is compared to a perfect α-helix, and the peptide appears in the optimum position to fit a penta-peptide α-helix. In 'beta' mode, the fit is made to a perfect β-strand. In '2ry' mode, the program tries both 'alpha' and 'beta' and chooses the secondary element most similar to the previously built portion of the structure. The 'bones' mode, follows an existing skeleton, using the previously built portion of the structure to decide where to place the di-peptide. The aim of the **Baton** commands is to build the bits of structure that were left over by the **Trace** commands.

Pass 4 and 5 and ...

After making the Cα trace, the hard part is done. The complete coordinates can now be generated from main and side chain databases as we have described previously (Jones *et al.*, 1991). In *pass 4*, we generate the backbone atoms with **Lego_auto_mc**. This step is extremely rapid, and surprisingly accurate, and thus alleviates the need for manipulating residues in the initial model building step. This is the reason that we advocate spending as much time as possible getting an accurate Cα trace from the start. While the main chain autobuild is taking place, the group of 20 best fitting penta-peptides from the database are displayed on the graphics screen. Some fanning will be noticed in loop regions, whereas the clustering should be rather tight in secondary structural regions. Anomalies in such regions should be fixed at this stage, using the **Lego_ca** command. The side chains are added in their most common rotamer conformation with **Lego_auto_sc**. Real space refinement of rotamers into the density may be used, or the user may chose to position each rotamer by hand in *pass 5*. Whichever method is chosen, we have strongly promoted the use of rotamers to build the first complete structure. An analysis of the sidechain rotamers shows that among structures of extremely high resolution shows that sidechains tend to cluster even *more* tightly around the preferred conformation than in structures of more modest resolution (J. Thornton *et al.*, this book). Therefore, choosing one of the top rotamer conformations is as good a choice as any, and there should be a very good reason for choosing something different.

In *pass 5*, each residue has to be checked for how well it fits the density. If the user decides to interactively place rotamers in the density, then **Lego_side** is first used to decide on the rotamer and then **Move_zone** (double clicking on the Cα atom to force it as the pivot point) to optimize the fit. The **Pep_Flip** command should be used at this stage (and during every refinement rebuilding cycles), to point out possible erroneous conformations of the peptide planes. The command **RSC_fit** allows one to locate sidechains with unusual conformations (if only rotamers are used, this should not be needed). At the end of this pass, **RS_fit** should be used to check which residue are still fitting poorly to the density.

At this stage it is still necessary to look for errors in the trace, which remains just a hypothesis. When the complete molecule has been constructed, the interactions between symmetry related molecules can be checked for inconsistencies. Finally, refinement can start to the highest resolution possible. During each rebuild, it is necessary to evaluate

the model, especially for out-of-register errors. Hopefully there will be few, but keep the experimental map handy, it is the ultimate source of truth.

References.

1. Bjorkman, P.J., Saper, M.A., Samraoui, B., Bennett, W.S., Strominger, J.L. & Wiley, D.C. (1987). "Structure of the human class I histocompatibility antigen HLA–2". *Nature* **329**, 506–512.
2. Brändén, C.-I. & Jones, T. A. (1990). "Between Objectivity and Subjectivity". *Nature* **343**, 687–689.
3. Greer, J. (1974). "Three-dimensional Pattern Recognition: An Approach to Automated Interpretation of Electron Density Maps of Proteins". *J.Mol.Biol.* **82**, 279–288
4. Jones, T. A. & Thirup, S. (1986). "Using known substructures in protein model building and crystallography". *EMBO J.* **5**, 819–822.
5. Jones, T. A., Bergfors, T., Sedzik, J., & Unge, T. (1988). "The Three-dimensional Structure of P2 Myelin Protein".*EMBO J.* **7**, 1597–1604.
6. Jones, T. A., Cowan, S., Zou, J.-Y. & Kjeldgaard, M. (1991). "Improved Methods for Building Protein Models in Electron Density Maps and the Location of Errors in these Models". *Acta Cryst.* **A47**, 110–119.
7. Kleywegt, G. & Jones, T.A. (1994), "Halloween ... Masks and Bones". *this book.*

A semi-automated map fitting procedure

T.J.Oldfield
York University
Heslington
York YO1 5DD

The initial fitting of coordinates to a SIR, MIR or MAD map can be a difficult and frustrating process that is very sensitive to the quality of the initial phases, and hence the interpretablility of the map. In 1974 J. Greer described an algorithm to carry out data reduction of the electron density to a single line, now generally known as skeletonisation of electron density or just "bones". This has greatly aided the process of building the initial coordinates, but requires much manual manipulation of the skeleton for interpretation. The skeletonisation algorithm is generally used as a stand alone program provided with the program O (Jones & Kjeldgaard 1991) and therefore the parameterisation of the calculation is a matter of educated guess work. The resulting skeleton provides a guide for building the model coordinates.

The aim of this work is to provide the crystallographer with an integrated package that provides as many visual clues as possible and auto fitting the atoms which have obvious positions. The skeletonisation algorithm has be implemented from the original paper, (J Greer, 1974) but with some minor changes to give improvements in the connectivity and speed. This algorithm has been included in a user interface that provides information to aid in the fitting of CA atoms, and the encoding of some program intelligence to determine the next residue position. Finally, the program can automatically generate N,C,O and CB atoms from the built CA coordinates to speed up the placing of entire residues into the density.
The map fitting interface has been implemented the in program QUANTA (MSI).

Calculation of skeleton of electron density

The skeletonisation of density is a three dimensional data reduction algorithm that removes points from the electron density using the rules:-

1) Remove all points below a threshold
2) Remove edge points so that....
    the electron density connectivity is retained and
    the electron density chain length is not reduced.

Step 2 is repeated until no more points can be removed from the map.

Mathematically this is simple to implement, but difficult to make fast enough to provide an interactive display. With careful consideration of memory and the use of short cuts in the calculation it has been possible to carry out the calculation on a 12Å radius sphere of electron density in under a second (Indigo R4000). This allows the user to change the initial cut off and gradient of the algorithm for each section of the map they are working on. It is therefore possible to change a simple user dial in the program and watch the effect on the electron density skeleton. This particularly useful when the chain traverses the core of the protein where there is "good" density, then forms a loop at the surface, where lower density levels must be used to observe chain connectivity. The algorithm results in a more

more interpretable. In particular, the treatment of neighbouring points in the map during the map point removal has been changed.

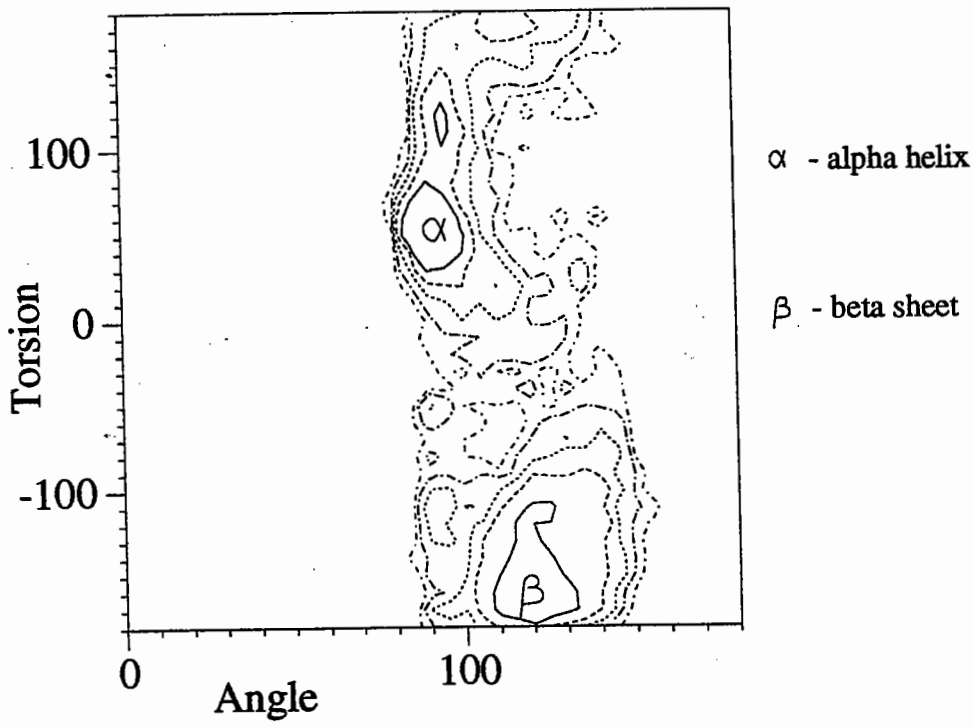
A fast skeleton smoothing function is included to change the remaining map points, to a smoother line of skeleton points. A fast variable branch trimming function allows the user to clean the skeleton of short branches depending on their length. The crystallographer can change the length of branches to be trimmed, so that for initial map fitting, only the longer branches are left, so that the larger residues flag the positions of the CA atoms, and the C=O branches are snipped off.

The result is a simple skeleton interface that allows the user to change the data reduction parameters and branch trimming interactively so that the he/she can fine tune the algorithm for the section of map they are fitting. For example, when fitting to the core of the protein it is necessary to use a "start" value in the skeletonisation routine that is larger than 1.2 sigma, while for the loop regions it may be necessary to use a start value of less than 1.0 sigma. The interactive control of the skeletonisation of a electron density map makes it a useful tool for the building of coordinates into variable quality electron density.

The CA conformation map

It has been shown that a probability surface can be generated if the pseudo torsion from each set of four CA atoms in a protein is plotted against the pseudo opening angle derived from three CA atoms (either subset of same 4 CA atoms) for a well refined set of proteins from the protein data bank (Oldfield & Hubbard - in press). The allowed regions on this plot are restricted. Different areas on the plot provide information on the local conformation of the protein backbone, such as alpha helix, beta sheet, and turn structures. It is therefore possible to use this empirically derived probability surface to direct the fitting of CA atoms to electron density. This probability surface using only CA atoms is shown in the figure.

The program provides this empirical probability surface contoured to highlight various conformations common in proteins. A pointer on this map indicates the current value of the angle and torsion, and its position can indicate the occurrence of helix or extended conformations. Hence, it is immediately obvious when the CA atoms are being fitted to helix, or B-strand conformations, or when impossible conformations are being generated, for example, when the CA atoms are fitted to side chains. This also has the advantage of indicating when the user is generating extensive right handed helices into their SIR (or even MIR maps) and hence that the wrong heavy atom solution has been used to phase the map. This can obviously save a great deal of time in the building process. An added advantage of the CA conformation map is that the user can pick a point from this map, and so place the current CA atom into this conformation, for example, an alpha helix.



Placing of new atoms and segments

The placing of the first atom of a segment is defined by picking a point in the skeletonised density. The most obvious point is at the most interpretable section in the map. The user is then provided with the next CA atom to fit to the map at a pseudo bond length of 3.8Å from the previous atom using the visual clues provided. At any stage the growing chain can be reversed and built in the opposite direction until both ends reach uninterpretable regions. New segments of chain can be fitted to each length of interpretable map. The current atom to be placed can be moved to the desired position by either picking a point in the skeleton, changing the pseudo angle and torsion, asking the program to auto fit the atom, or by picking a point on the CA conformation map as described in the previous section.

Marking all the next possible paths.

The crystallographer is provided with a set of markers that can be toggled on and off to indicate all points on the skeleton that are 3.8Å±0.3Å from the last CA atom fitted. This indicates all possible positions of the next CA atom regardless of the connectivity.

The auto fit of atoms

The program has a simple intelligence that will attempt to place the next CA atom to the best piece of the skeleton. This is currently based on the following rules:-
1) The set of conformation which is linked to the previous point by continuous skeleton.
2) The CA geometry with respect to the previous CA atoms fitted, weighted as a function of the CA conformation map.
3) The analysis of branches at/near the set of points 3.8Å from the last CA atom.

The auto fit of the next atom is always carried out when the user asks for the next CA atom in the sequence, so that the new atom is hopefully in the best density. The user can overide this initial guess by any of the inputs of new positions previously defined, or they can ask the program to auto fit the atom again. The proportion of correct solutions that the program finds is dependent on the quality of the map, and can never be as good as an experienced crystallographer, but the algorithm can generally place the next atom correctly on average 40% of the time for an "average" map. It is invisaged that this intelligence can be improved by recursive searches along the connected density, and by allowing analysis across disconnected density.

Building other atoms from the CA positions

An algorithm has also been added that allows the generation of the coordinates N, CB, C and O atoms from the CA atoms in conformations that are normally found in well refined proteins (Oldfield & Hubbard in press). The procedure uses a fragment matching method which finds the best conformation match in a sub set of well refined protein structure for all combinations of the positions of five connected CA atoms built into density. The program returns the generated polyalanine chain which can be used for further building into the electron density.

Summary

This map fitting algorithm allows display of "bones", electron density, and various visual clues to aid in the initial fitting of coordinates to MIR/SIR/MAD maps. The aim is to provide tools for the crystallographer to reduce the time taken in this essentially tedious process.

This work was supported by Glaxo Group Research, Molecular Simulations Inc, and the SERC.

References

Greer J. "Three-dimensional Pattern Recognition : An Approach to Automated Interpretation of Electron density Maps of Proteins." J.Mol. Biol. (1974) 82, 279-301

Jones T.A. and Kjeldgaard M. (1991) "O Manual", University of Uppsla, Sweden.

MSI - 16 New England Executive Park, Burlington, MA, 01803-5297 USA.

THE RIGHT MODEL?

Lessons from building a structure of 6PGDH starting from a wrongly connected model with an incorrect sequence

Margaret J. Adams & Sheila Gover

Data sets discussed in this paper have been collected, and calculations performed, by Grant H. Ellis, Katherine C.M. Pelly, Christopher Phillips, Richard W. Pickersgill & Donald O'N. Somers.

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

It is often difficult to decide whether a medium resolution electron density map is potentially interpretable when the chain trace is not immediately obvious. It is similarly difficult to decide, if a structure determination is not complete, whether the partial structure is correct or not. This paper attempts to address these problems with reference to our experience with 6-phosphogluconate dehydrogenase.

A The structure determination

6-phosphogluconate dehydrogenase (6PGDH) from sheep liver has been solved and refined using data to 2.0Å resolution. The molecule is a dimer with 482 residues per subunit. It crystallises in $C222_1$ with one monomer in the asymmetric unit. The structure determination used 2 heavy atom derivatives, $KAu(CN)_2$ and $K_2Pt(CN)_4$, which had phasing power greater than 2.0 for all (acentric and centric) data to 2.8Å and 5.0Å respectively. The final solution was obtained using Wang solvent flattening of the 2.8Å electron density map and rebuilding from several combined maps with incomplete models.

The subunit has two major domains: a dinucleotide binding fold, and a helical domain which contains most of the substrate binding and catalytic residues. The C-terminal 50 residues form a tail which penetrates through the second subunit of the dimer. After refinement using XPLOR, the final R factor for all 35031 reflections observed between 20Å and 1.99Å is 19.8%, corresponding to a mean coordinate error of 0.25-0.30Å. The mainchain dihedral angles of all residues but one are in the allowed regions of the ϕ,ψ plot. There are 346

waters included in the final model. Comparisons in this paper refer to the model refined at 2.5Å with an R factor for all data between 20Å and 2.5Å of 18.5% (Adams *et al.*, 1991).

The 6Å structure which was published in 1977 (Adams *et al.*) used data collected on a Hilger 4-circle diffractometer. The native data and data for 3 heavy atom derivatives were subsequently collected on film to 2.6Å or 2.8Å resolution using the rotation method and a conventional source. Most α-helices were apparent in the resulting electron density map but it proved impossible to follow the chain. The method of Bhat and Blow (1982) was chosen for map modification. The amino acid sequence (Carne & Walker, 1983) was used as an aid to following the chain and, in turn, one ambiguity of peptide alignment was decided on the basis of the modified electron density map. A chain trace was identified (Adams *et al.*, 1983), but considerable reliance was placed on the sequence in deciding the path to be followed in difficult parts of the map.

Data were then collected on film to 1.99Å and 2.1Å resolution using the Daresbury synchrotron source: the 1.99Å data set was collected using Vee cassettes at λ=1.488Å; the 2.1Å set used flat plates and λ=1.054Å. The structure proved difficult to refine and, in several places, the combined map indicated there might be errors in both the mainchain trace and sidechain identification. Refinement with PROLSQ failed to improve the agreement below an R factor of 0.36 and this was only achieved with unrealistic temperature factors. Efforts to correct the structure at this stage failed. When we aligned the amino acid sequence with the gene sequence of the *E. coli* enzyme (Nasoff *et al.*, 1984), we also became aware that the cyanogen bromide peptides composing the sequence might be wrongly assembled. Attempts to recognise the correct sequence in any combined or solvent flattened electron density maps available at the time failed.

The cDNA sequence of the sheep enzyme was therefore determined (Somers *et al.*, 1992); it demonstrated that 100 residues in 3 peptides placed in the sequence at 10 - 68, 69 - 81 and 82 - 102 should have been placed as 361 - 419, 420 - 433 and 340 - 360 respectively. The structure was finally solved after collecting new data sets to 2.8Å resolution on the Nicolet area detector for the native enzyme and the $KAu(CN)_2$ and $K_2Pt(CN)_4$ derivatives. The lessons which may be learnt from the models used in the structure determination are discussed in this paper.

B The correct and incorrect models

Scheme I notes some differences between the correct structure and the incorrect interpretation. Several secondary structure elements were missed and some helices and sheet strands had been interpreted as running in the wrong direction. The N-terminus was incorrectly identified, so the chain trace began after a disordered region (301 - 315 in the solved structure) which had not been visible in the electron density maps.

Scheme I
Differences between "right" and "wrong" structures

		right	*wrong*
Residues		469 *	433
Helix	number of residues (number of helices)	264 (20)	157 (15)
direction:			
	same		11
	opposite		4
Sheet	number of residues (number of strands)	34 (8)	22 (6)
direction:			
	same		3
	opposite		3

*2.5Å model contained 469 residues; the final 2.0Å model contains 473 residues.

The suitability of several models as starting point for the solution will be considered. The "wrong" models were generated from the original incorrect interpretation after different refinement protocols (models **w1** - **w5**). The "broken apart model" was made by fitting an experimental sequence to a combined map which used the Bhat/Blow modified phases and the 2.8Å model **w1**. Breaks were made where the map was most obviously discontinuous and sidechains were curtailed to fit density. The "right models" are the result of interpreting a Wang solvent flattened map which used new data collected on the area detector (**r1**) and of subsequent combined maps (models **r2** and **r3**). Models **r1** and **r2** were built with experimental sequences which fitted the electron density maps; model **r3** had the correct sequence. None of these "right" models are refined.

C The right map

The most useful response to an uninterpretable map or a model which will not refine is to collect better data or data from new and better derivatives. The effect of new data collection for 6PGDH is seen in table I where the heavy atom occupancies are compared for four different data collections of the gold derivative.

Table I
Gold derivative data sets

data set	number of crystals	site	occupancy	Δxyz	total no. electrons	phasing power
conventional 2.8Å area detector	3	1 2	0.92 0.39	- -	103	1.61
synchrotron 2.8Å film	1	1 2	0.61 0.31	0.5Å 0.65Å	72*	1.15
conventional 2.8Å film	many	1 2	0.20 0.33	0.8Å 0.6Å	42	1.01
conventional 6.0Å diffractometer	1	1 2	0.56 0.28	0.4Å 0.8Å	66	1.57

* two additional low occupancy sites were also included, accounting for a further 28 electrons

The phasing power quoted is that determined from the centric reflections alone after centric refinement. The highest individual heavy atom occupancies, largest total scattering contribution and the best phasing power correlate with the collection using the most sensitive detection system. This permitted use of a small number of crystals to collect a complete data set with high significance and little radiation damage. While loss of heavy atom from site 1 may occur on severe radiation damage, it is more likely that the real heavy atom occupancy for all data collections is represented by that determined for the area detector data. Even so, the phasing power of this, the best, derivative is not particularly high and the signal from 100 electrons is small for phasing an asymmetric unit which contains approximately 500 residues.

Electron density maps were calculated: (1) using isomorphous and anomalous differences from the area detector gold derivative data alone; (2) after Wang solvent flattening starting from the same heavy atom data; (3) after solvent flattening using area detector data from the gold and the, less isomorphous, platinum derivatives; and (4) after solvent flattening using all old and new phase information. Comparison of density corresponding to the well ordered long helix, α–h, with the 2.5Å model shows map 2 to be the most readily interpretable in this area. The SIRAS map (map 1) contoured at the same level (2σ) was discontinuous. Map 3 was an improvement over map 2 only near the gold sites and map 4 had little information not present in map 2. All maps were much more readily interpretable than the earlier modified or combined maps using the earlier, more radiation damaged, data.

D The right model to fit to the map beginning from a "wrong" model

1 Can the "wrong" model or reliable parts of it be refined to the "right" data?

Table II demonstrates the effect of refining a grossly incorrect model using either PROLSQ or XPLOR. Two different PROLSQ refinements are shown. The first, which generated the model **w2**, used the full but incorrect sequence and refined to the flat plate (λ=1.05Å) synchrotron native data set. The second refinement, yielding model **w3**, was attempted after the probable errors in the sequence were known and only those residues which were a good fit to a combined electron density map were included; again the λ =1.05Å native data were used. The first refined model deviated rather more from the correct structure at a Cα level than its starting model and the number of residues in correlated chains decreased. Only three long (>25 residue) consistent chains were found, one in the wrong sense, and they included only 101 residues whereas the starting model had 147 residues in long chains which followed the correct path. The second refinement broke up the model entirely, as is evidenced by the fact that no chain of 25 residues in length was consistent with the correct structure. The resulting electron density map was uninterpretable. This might be anticipated as the starting model is much too far from the true structure to be within the linear range for the least squares procedure.

Table II
Differences between coordinate sets: Cα comparisons

All compared with 2.5Å refined structure which was built with 469 residues
Cα positions compared if less than 3.5Å apart, chain number changed if gap in Cα numbering, comparison chain changed if individual chain sequences broken.

model	number of residues	number in common	number of chains	ΔCα (Å)	n chains with >25 residues ; nresidues	n chains with same / opposite sense
"wrong models"						
crayII 3.0Å refine* (**w1**)	434	311	31	1.38	4 147	3,1
PGDH3 2.1Å refine (**w2**)	434	298	34	1.39	3 101	2,1
PROLSQ on 'good' fit to map (**w3**)¥	346	247	48	1.30	0 0	
guessed sequence, failed XPLOR input (**w4**)†	436	321	31	1.30	4 132	3,1
guessed sequence, failed XPLOR output (**w5**)§	436	301	45	1.40	3 93	3,0
"broken apart model"						
pgdhrebuild** (**b1**)	443	319	30	1.29	4 169	3,1
"right" models						
fit to first Wang map (**r1**)	414	368	17	0.98	7 292	7,0
final experimental sequence (**r2**)	446	428	4	0.74	2 396	2,0
final build (**r3**)	466	464	1	0.52	1 464	1,0

* Combined phases after this cycle using synchrotron data = GEMCOMB
¥ Resulting map did not have continuous density,
§ Geometry improved and marginal improvement in agreement with X-ray data; heat and cool steps gave no improvement
† Used combined phases resulting from this refinement
** Used GEMCOMB phases

A simulated annealing (XPLOR) run was attempted using a model, **w4**, which, while it substantially retained the incorrectly connected sequence and connectivity, had been modified to fit the sidechain density better. The starting point for this refinement can be seen (table II) to fit the correct structure a little better than any of the previous models. The initial improvement of model geometry and energy minimisation was consistent with a slightly better fit to the diffraction data but simulated annealing gave no improvement and the output model, **w5**, is seen to be less consistent with the correct structure. Only three 25 residue chains with the same trace as the correct structure remained; it may be significant that the chain which was lost had been interpreted in the wrong direction. The smaller number of Cα's which were close enough to those of the true structure to be compared had moved further away from the true positions. It is apparent that these grossly wrong structures could not be refined usefully even to give a starting point for interpretation of a new electron density map.

Table III compares the extent to which agreement with the diffraction data signals that a model is incorrect and shows the differences between the phases of the "wrong" or "right" models and α_{calc} after the 2.5Å refinement (which was against area detector data).

The first 2.1Å refinement (model **w2**) gave only a small improvement in agreement; the R factor of 0.36 using λ=1.05Å data from 8 to 2.1Å had been achieved for this incorrect model at the expense of loss of regular secondary structure and with imposition of improbable and rapidly changing temperature factors for mainchain atoms. The refined phases have drifted slightly further from those for the correct model. The table shows there is little correlation between the effect on phases and on R factor for any pair of "wrong" models. While the energy minimisation and simulated annealing served to improve structure factor agreement, the phase set **w5** is significantly worse than that for the input model. The improvement of the R factor may not be entirely helpful in that it gives some cause for optimism about a structure which remains grossly wrong.

Table III
R factors and phase analysis for different F_{calc} sets,
taking the area detector data set and 2.5Å model phases as standard

Structure factors are compared between 8 and 2.8Å: 12414 reflections
Model B factors are 20 or 22 for all atoms

F_{calc} set	model	R factor	phase analysis	
			centric frac.same	acentric Δα (deg)
"wrong models"				
CrayII 3.0Å refine	**w1**	0.508	0.605	77.3
PGDH3 2.1Å refine	**w2**	0.474	0.615	77.3
PROLSQ on 'good' fit to map	**w3**	0.512	0.607	76.1
guessed sequence, failed XPLORinput	**w4**	0.522	0.623	74.8
guessed sequence, failed XPLOR output	**w5**	0.406	0.620	78.2
"broken apart model"				
pgdhrebuild	**b1**	0.536	0.623	74.5
"right" models				
fit to first Wang map	**r1**	0.509	0.677	63.6
final experimental sequence	**r2**	0.448	0.748	53.5
final build	**r3**	0.398	0.814	42.5
first XPLOR output	**r4**	0.268	0.872	27.9

Experimental and combined phase sets

F set	model built	figure of merit	phase analysis	
			centric frac.same	acentric Δα (deg)
Initial isomorphous		0.529	0.671	72.0
GEMCOMB	**w0, b1**	0.595	0.661	69.8
New isomorphous		0.513	0.734	64.3
Wang	**r1**	(0.824)	0.779	62.6

Table III also analyses certain experimental and combined phase sets. The improvement which was obtained from a new data collection is apparent: the new isomorphous phase set, although it does not give a connected density map without solvent flattening, is better than any "wrong" combined phase set. The dependence on a single derivative is apparent in the relatively large phase difference for the acentric terms. The figure of merit is seen not to be a sensitive indicator at this stage.

2 What is the best starting point from a "wrong" model?

Both tables II and III refer to a "broken apart" model. This model was generated from the phase set 'GEMCOMB' which resulted from minimal refinement of the original model. The model built had 'experimental sidechains' and no attempt was made to build chains through gaps in the density. While $\mathbf{F_{calc}}$ from the rebuilt model **b1** agreed poorly with the correct terms, this proved a useful initial model for interpreting the new solvent flattened map, primarily because there was no anticipation that the Cα trace represented the correct structure to any great extent. It can be seen in table II that it did contain more Cα atoms in long correct chains than any wrong model generated by refinement.

E How can the broken and partly incorrect model be used to interpret an imperfect map?

1 The starting chain trace

Use of the coordinates of the broken model could nowadays be simplified by employing the various facilities in 'O' - particularly the data base comparisons in LEGO-CA. At the time these were not available and this may have made interpretation slower. The strategy, which probably follows much of O's implied strategy, would remain the same. Three types of misinterpretation will be discussed; and the reasons why they may be common and the effect they may have on a combined map will be considered.

a A helix slid by an atom

The beginning of helix α-h is immediately preceded by a sharp turn with the sequence GAG; the density was thin with no reinforcement by the position of β-carbons. The initial turns of this helix were built with Cα an atom misplaced. A 3.5Å map using the correct $\mathbf{F_{calc}}$ makes it clear that in the absence of phase error there is no difficulty in defining Cα positions for this helix. The solvent flattened map based on new derivative data in this area allowed the model to be corrected (model **r1**) and the resulting experimental sequence, used in a succession of phase combination cycles, yielded the correct structure. The combined map built with the 'slid' sequence (phase set GEMCOMB) showed the density to be broken between the mainchain and the sidechain. Refinement of this model resulted in some loss of secondary structure. The effect of the wrong model on the phase set was such that neither the correct nor the incorrect model was a good fit to the map.

b A helix built in the wrong direction

Helix α-k (residues 262 - 273) was close to the less substituted gold site and, possibly because ripples from this site deformed the density, was built in the wrong direction. The best fit had the characteristics of a 3_{10} helix rather than an α-helix. The >C=O group in a 3_{10} helix points outwards and after one sidechain (residue *n*) had been placed in sidechain density, the model built in the wrong sense was unwound. This allowed the >C=O group of

residue *(n-2)* to be placed in the density belonging to Cβ of residue *(n+2)*, which points out in the direction anticipated for the 3_{10} mainchain. The mainchain build was further distorted so that the *(n+2)* >C=O density could be used for the *(n-2)* Cα-Cβ bond. Combined maps using coordinates from the wrong sense helix had density which became progressively less helical and more broken up. Changing the sense improved all succeeding combined maps. Even at 3.0Å, some >C=O bulges are visible in helices; the 'Christmas tree' effect of sidechains is apparent at a lower resolution than 3.0Å in an experimentally phased map. The absence of any β-carbon density, and broken mainchain density, in a combined map may suggest that the helix is built in the wrong sense.

c Incorrectly connected β-sheet

A poor 3Å map contains cross-connections in a sheet where density is continuous through the hydrogen bonds rather than through mainchain. Interpretation can prove particularly difficult near the edge of a relatively short sheet where there is higher than average thermal movement. The incorrect model had a U-turn across a hydrogen bond giving 3 anti-parallel sheet strands instead of parallel strands. Again, a correctly phased 3.5Å map is readily interpretable. The solvent flattened map in a rather mobile region is often artificially flattened, but the appearance of rows of sidechain density on alternate sides of the sheet and running across it should make it possible to decide the connectivity and should prevent one from building an incorrect U-turn. Although the >C=O density is less obvious than in a helix, a 3Å solvent flattened or combined map will give clues as to the sheet direction. The combined map calculated with the incorrectly interpreted U-turn gave broken density both for the model mainchain and for the correct mainchain. The correct connectivity and sheet direction was established by three phase combinations leading from model **r1** to model **r2**.

Table III shows that the phase errors in the phase set used for the combined map with errors of the types discussed above (GEMCOMB) are slightly less than those of the initial isomorphous phase set or those of the "wrong" models. However, despite the improved figure of merit, they are worse than the new isomorphous phase set generated from the area detector data. This combined phase set enabled the starting "broken" model to be built, but was not good enough for 'boot-strapping' to the structure.

2 The structure with an experimental sequence

The models **r1**, and **r2** were both built without making any attempt to follow the actual sequence. Combined maps used only those atoms which were within the density. The atoms of the model **r1** used in the first combined map corresponded to some 78% of the subunit. Although successive cycles contained an increasing number of correctly positioned Cα atoms, one subsequent phase combination used only 70% of the subunit density. The phase combination which resulted in model **r2** included 90% of the expected density; some 90% of the Cα positions were within 0.75Å of the true positions and it can be seen in table III that the phase errors are correspondingly reduced. The phase errors in all "right" models are lower than

those for any "wrong" model. Combined maps allowed better models to be built in successive cycles. The starting isomorphous phases were themselves slightly better than those of model **r1**. When model **r2** had been reached, it was possible to use a combined phase set where all residues were cut at Cβ (only 55% of the structure) which allowed an unbiased estimate of the sidechain.

F Fitting the sequence

It is easy to recognise patterns of large sidechains in a good high resolution map. The SLIDER facility in 'O' will allow pattern recognition in a slightly less beautiful map. If neither of these methods gives an alignment which seems sensible, what alternative routes are available?

It is apparent from the previous section that a succession of cautious builds to combined maps will give improved models and phase agreement (see the sequence **r1**, **r2** in tables II and III). At the stage of the model **r2**, attempts were made to fit the correct sequence. At this point there were still 4 different chains and the connections between them had not been made. Almost all the elements of secondary structure had been recognised and the β–α–β domain and the helical domain were apparent. Two helices of more than 30 residues had been built. Although a secondary structure prediction which was ultimately shown to be 60-70% correct was available, neither the long helices nor the correct pattern of the β–α–β domain could be found.

Figure 1

```
cDNA sequence                              A Q A D I A L I G L A V M G Q N L I L N
secondary structure - refined model              -beta-A--     ------alpha-a----
secondary structure prediction             H H H H H H H H H B B B B B B B B B
secondary structure - pre-refinement          B B B B B b       H H - H H H H H H
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

 M N D H G F V V C A F N R T V S K V D D F L A N E A K G T K V L G A H S L E E M
------         -beta-B         ----alpha-b----                  betaC      --alpha
    T T B B B B B B       B B B B H H H H H H H H                   H H H H H H H H
H          B B B B b       h h H H H H H H                 b B B b     H H H H
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

 V S K L K K P R R I I L L V K A G Q A V D N F I E K L V P L L D I G D I I I D G
-c---         --beta-D---         ------alpha-d--------              --beta-E---
H H H H H T   H B B B B B B H H   H H H H H H H H H H H H     B B B B B B B T T
H H H h         b B B B B b     h - H H H H H H H H H H H       B B B B B B
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~


~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

 G D D G A G H F V K M V H N G I E Y G D M Q L I C E A Y H L M K D V L G L G H K
 betaH   ----------------------alpha-h-----------------------------          ----
 T T T T T     H H H H H                H H H H H H H H H H H H H H H H H H H H
           h H H H H H H H H H H H H H H H H H H H H H H H H H H H H h
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

 E M A K A F E E W N K T E L D S F L I E I T A S I L K F Q D A D
-----alpha-i-- ---               ------alpha-j--------
 H H H H H H H H H H H H H H H H H H H H H H H H   H H H H H H H H
 h H H H H H H                 H H H H H H h h h h
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

Figure 1 demonstrates, using two examples, why the prediction could not be used. The actual sequence, the predicted secondary structure and the observed secondary structure, both as defined before any refinement and after the 2.5Å refinement, are shown for residues 1-100 (in the coenzyme binding domain) and 174-245 (lead into and beginning of the helical domain). The prediction for the sheet strand β-A and helix α-a is helix - sheet instead of sheet - helix. This impeded recognition of the dinucleotide binding fold. While some 19 residues of α-h are predicted as helix, the long predicted helix begins more than half way through helix α-h and continues through α-i and α-j. The breaks between secondary structural elements are crucial if the prediction is to be used for alignment and even with a 70% correct prediction, the break points are not reliable enough to recognise particular patterns.

The residues in helix were used to align the sequence on the basis of a matrix of hydrophobicity, length and 'likeness' (table IV, over). The likeness parameter (0 or 1) does not commute and is based on the probable appearance in an electron density map; thus a proline residue may be mistakenly built as a serine, but it is much less likely that a serine is built as a proline. The fit of experimental sequence to actual sequence may be scored by normalised summation of the likeness of corresponding residues and the difference in their hydrophobicity and length parameters. Figure 2 shows the starting positions giving the best (lowest) scores for each attribute, and the result of combining scores with unit weight, for the experimental build **r2** to the two long helices. The computer program, written in Fortran, is available on request.

Figure 2
Alignment statistics for helices α-h and α-n

sequence 179 - 207: experimental α-h: GHMVQCSHSALHWGCQRLICHAYLLFNNT cDNA α-h: GHFVKMVHNGIEYGDMQLICEAYHLMKDV								
order	hydrophobicity		length		likeness		overall	
	position	score	position	score	position	score	position	score
1	439	0.414	179	0.586	179	0.345	179	0.517
2	179	0.621	48	1.379	200	0.552	418	1.023
3	451	0.655	43	1.483	372	0.552	48	1.046
sequence 317 - 354: experimental α-n: AAQQKLAASLFVSMLLSYAQGHAILQTA cDNA α-n: SFLEDIRKALYASKIISYAQGFMLLRQA								
order	hydrophobicity		length		likeness		overall	
	position	score	position	score	position	score	position	score
1	343	0.536	317	1.036	317	0.321	317	0.655
2	317	0.607	191	1.464	73	0.536	3	0.941
3	408	0.643	3	1.500	112	0.536	415	1.012

At the stage of build **r2,** it was possible to assign each helix correctly using the matrix. The fit of the residues which were not in secondary structure could then be achieved without difficulty and the final connections made. The model **r3** can be seen (table III) to be close to the final model with much smaller phase errors. The build was only 0.5Å in error over 464 of the 473 residues finally fitted; the R factor, however, would not have discriminated between this model and model **w5**. The first XPLOR output for the right structure is seen to have an R factor of 26.8 and acentric phase errors of 28°. This model continued to refine by simulated annealing to the maximum resolution with little need for manual intervention except to place solvent.

Table IV
Alignment parameters

inside-outside on scale 0=in - 3=out
length on scale 0 - 7

residue in sequence	inside-outside	length	likeness ACDEFGHIKLMNPQRSTVWY
ala A	1	1	#----G---------S----
cys C	1	3	-#D------L-N---S----
asp D	3	4	-C#E-----L-N-Q------
glu E	3	5	--D#----K-MN-Q------
phe F	0	6	----#-H----------WY
gly G	1	0	A----#----------A---
his H	2	6	----F-#---M-------WY
ile I	0	4	-------#-L------TV--
lys K	3	5	---E----#-M--QR-----
leu L	0	4	-CD----I-#-N----TV--
met M	1	5	---E--H-K-#N-Q------
asn N	2	4	-CDE-----L-#-Q------
pro P	2	4	----------#--S----
gln Q	2	5	--DE----K-MN-#------
arg R	3	7	--------K-----#---W-
ser S	2	2	AC----------#T---
thr T	2	3	-------I-------S#V--
val V	0	3	-------I-L------T#--
trp W	0	7	----F-H-------R---#Y
tyr Y	1	6	----F-H----------W#

How soon could the matrix have been used in defined secondary structure? Figure 3 indicates the 'likeness' count for the helix α-h in the final experimental build r2, the fit to the solvent flattened map r1, the broken model b1 and the wrong sequence w2. The decrease in the proportion of 'like' residues is apparent and it is unlikely that a reasonable confidence match would have been estimated with fit r1 where the residues his, phe (180,181) were built as ala, ala. It may be advisable to delay fitting sequence until the build is reasonably reliable to avoid incorrect positioning of sequence.

Figure 3
α-h for various coordinate sets

underline indicates "likeness"

```
VGDDGAGHFVKMVHNGIEYGDMQLICEAYHLMK DVL      true   (36 residues)
WAKDAAGHMVQCSHSALHWGCQRLICHAYLLFN NT       r2
AAKRAAAAAVQLLHAALHFGWRRASCHLRLLFG LLD      r1
AAIVS--GL-FALAAVAAMGQRHFVALAHGKVL DAS      b1
--KVGTG-E-PCCDWVGDDGAGHFVKMVHN-GIEY-G      w2
175       185       195       205
 -βH ------------α-h------------

VGDDGAGHFVKMVHNGIEYGDMQLICEAYHLMK DVL      true
WAKDAAGHMVQCSHSALHWGCQRLICHAYLLFN NT       r2     (23 like)

VGDDGAGHFVKMVHNGIEYGDMQLICEAYHLMK DVL      true
AAKRAAAAAVQLLHAALHFGWRRASCHLRLLFG LLD      r1     (16 like)

VGDDGAGHFVKMVHNGIEYGDMQLICEAYHLMK DVL      true
AAIVS--GL-FALAAVAAMGQRHFVALAHGKVL DAS      b1      (8 like)

VGDDGAGHFVKMVHNGIEYGDMQLICEAYHLMK DVL      true
--KVGTG-E-PCCDWVGDDGAGHFVKMVHN-GIEY-G      w2      (6 like)
```

G Some conclusions and suggestions

. Be cautious when relying upon an amino acid sequence, especially if deduced by peptide overlap; look for confirmation from other species.

. Collect as complete and reliable ıative and derivative data as possible; if the data will refine to higher occupancy, the phasing power will be greater.

. Do not be tempted to refine too early, or to inappropriately high resolution; a structure with an incorrect chain trace is likely to refine to a false minimum which is further from the true structure than its starting point. Beware also reduction in the R factor at the expense of secondary structure; the unrestrained behaviour of neighbouring atom temperature factors may be instructive as these should be correlated and appropriate to the situation within the molecule.

. Never lose sight of the original isomorphous map; use combined phases with minimally-biased Fourier coefficients.

. Be prepared to break connectivity and curtail sidechains so there is no build across or into gaps in the electron density; also omit unsure regions and look for their return in a subsequent combined map. Partial structures with an 'experimental' sequence can be used to reach a correct structure by a series of phase combinations, even if the initial map has considerable ambiguity and lack of connectivity and as liitle as 55% of the scattering matter is present. Sidechain density will improve to the correct shape if the degree of truncation is adjusted to the limit of the map after each combination cycle. Thin or broken density can indicate mis-register in a helix, a reversal of helix direction or cross-sheet connections. If built carefully, the right structure will emerge from a succession of combined maps with no refinement other than regularisation.

References

M.J. Adams, S. Gover, R. Leaback, C. Phillips & D.O'N. Somers (1991). *Acta Cryst.* **B47**, 817 - 820.

M.J. Adams, I.G. Archibald, C.E. Bugg, A. Carne, S. Gover, J.R. Helliwell & R.W. Pickersgill (1983). *EMBO J.* **2**, 1009 -1014.

M.J. Adams, J.R. Helliwell & C.E. Bugg (1977). *J. Mol. Biol.* **112**, 183 - 197.

T.N. Bhat & D.M. Blow (1982). *Acta Cryst.* **A38**, 21 - 29.

A. Carne & J.E. Walker (1983). *J. Biol. Chem.*, **258**, 12895 - 12906.

M.S. Nasoff, H.V. Baker II & R.E. Wolf Jr. (1984). *Gene* **27**, 253 - 264.

D. O'N. Somers, S. Medd, J.E. Walker & M.J. Adams (1992). *Biochem. J.*, **288**, 1061 - 1067.

Model Bias and Phase Combination

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

Because the phase is more important to the appearance of an electron density map than the measured amplitude, some degree of model bias is inevitable in model-phased maps. Through an understanding of the relationship between the two structure factors involved, the true one and the one calculated from a model, we can predict the systematic bias component of the model-phased structure factor. Knowing this, we can devise map coefficients that reduce model bias in electron density maps. With a few more assumptions, model bias can also be reduced in maps computed using phase information combined from several sources. The computation of map coefficients to reduce model bias is performed by the program SIGMAA, which also computes model phase probabilities and carries out phase combination. A special form of bias, the "refinement bias" that results from refining too many parameters against too few observations, cannot be corrected by these weighting schemes, but its influence can be reduced by appropriate refinement strategies.

1. Introduction

The intensities of X-ray diffraction spots measured from a crystal give us only the amplitudes, not the phases, of the diffracted waves. To reconstruct a map of the electron density in the crystal, we require both. This is usually referred to as the phase problem of crystallography. In fact (as discussed in section 2), the unknown phases are actually much more important to the appearance of the electron density map than the measured amplitudes.

To use model phase information optimally, we need an estimate of its reliability, specifically the probability that various values of the phase angle are true. To get such an estimate, we have to start first with the relationship between the structure factor (amplitude and phase) of the model and that of the true crystal structure (section 3). The phase probability distribution can then be obtained and used, for instance, to provide a figure-of-merit weighting that minimizes the rms error from the true electron density (section 4).

Even with figure-of-merit weighting, the use of model phases introduces a bias that causes the map to resemble the model. The systematic bias component of model-phased map coefficients can be predicted, allowing the derivation of map coefficients that correct for systematic bias. This results in electron density maps with reduced model bias. By making a few simple assumptions, we can also correct for bias when different sources of phase information are combined. Map coefficients that reduce model bias are discussed in section 5.

Finally, the refinement of a model against the observed amplitudes allows a certain amount of overfitting of the data, which leads to an extra "refinement bias". Some strategies to reduce the severity of this problem are described in section 6.

2. Model bias: importance of phase

Dramatic illustrations of the importance of the phase have been published. Ramachandran and Srinivasan (1961) calculated an electron density map using phases from one structure and amplitudes from another. In this map there are peaks at the positions of the atoms from the structure that contributed the phase information, but not for the structure that contributed the amplitudes. Similar calculations with 2-dimensional Fourier transforms of photographs show that the phases of an alarm clock completely overwhelm the amplitudes of Walter Cronkite (Oppenheim, 1981).

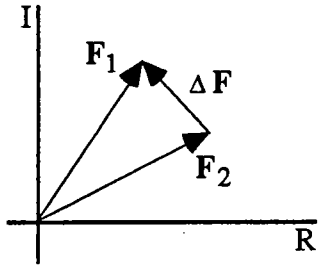
These examples, though dramatic, are not really representative of the normal situation, where the structure contributing the phases (the atomic model of a crystal structure) is partially

or even nearly correct. Nonetheless, model phases always contribute bias, so that the resulting map bears too strong a resemblance to the model.

Parseval's theorem

The importance of the phase can be understood most easily in terms of Parseval's theorem, a result that is important to the understanding of many aspects of the Fourier transform and its use in crystallography. Parseval's theorem states that the mean square value of the variable on one side of a Fourier transform is proportional to the mean square value of the variable on the other side. Since the Fourier transform is additive, Parseval's theorem also applies to sums or differences.

If ρ_1 and ρ_2 are, for instance, the true electron density and the electron density of the model, Parseval's theorem tells us that the rms error in the electron density is proportional to the rms error in the structure factor, considered as a vector in the complex plane (Fig. 1).



$$\langle \rho^2 \rangle = \frac{1}{V^2} \sum_{\text{all h}} |\mathbf{F(h)}|^2$$

$$\langle (\rho_1 - \rho_2)^2 \rangle = \frac{1}{V^2} \sum_{\text{all h}} |\mathbf{F}_1\mathbf{(h)} - \mathbf{F}_2\mathbf{(h)}|^2$$

Figure 1. Vector difference in complex plane.

This understanding of error in electron density maps explains why the phase determines the appearance of an electron density map much more than the amplitude. As illustrated in Fig. 2, a random choice of phase (from a uniform distribution of all possible phases) will generally give a larger error in the complex plane than a random choice of amplitude (from a Wilson distribution of amplitudes; Wilson, 1949).

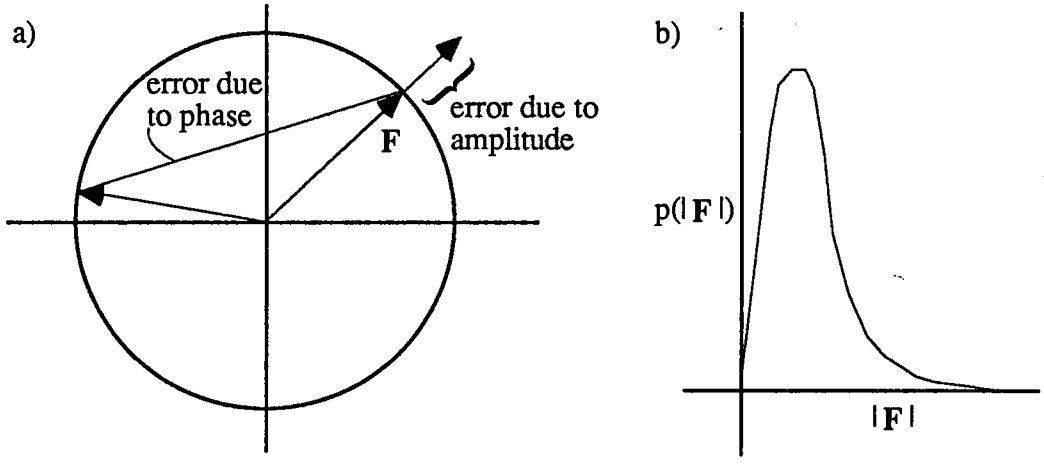


Figure 2. Schematic illustrations of: a) the relative errors introduced by a random choice of phase or a random choice of amplitude; b) Wilson distribution of amplitudes.

3. Structure factor probability relationships

To use model phase information optimally, we need to know the probability distribution for the true phase (or, equivalently, the distribution of the error in the model phase). This distribution can be obtained by first working out the probability distribution for the true structure factor (or the distribution of the vector difference between the model and true structure

factors). Then, by fixing the known value of the structure factor amplitude and renormalizing, we obtain the phase probability distribution.

We will begin with a short discussion of the central limit theorem, which will be invoked to justify the use of Gaussian distributions for structure factors. A number of Gaussian structure factor distributions are related to one another, differing in the amount of information available about the structure and in the assumed form of errors in the model. We will build from the Wilson distribution, which applies when none of the atomic positions is known, through to a distribution that applies when there are a variety of sources of error in an atomic model.

An implicit assumption in this approach is that all structure factors are independent of one another, even though we know from the success of direct methods that they are not. While there is certainly much valuable information to be obtained by considering higher order collections of structure factors (Bricogne, 1993), the phase probabilities for single structure factors turn out to be very useful in practice, and are also inexpensive to compute.

The Central Limit Theorem

When there are a number of sources of error, it is often found that the overall error follows a Gaussian distribution, regardless of the individual distributions of the various sources of error. The central limit theorem lays out the very general conditions under which the Gaussian distribution provides a good approximation. When these conditions are satisfied, it is often possible to sidestep the laborious manipulations required to work out the probability distribution of a random variable.

In order for the central limit theorem to apply, two important conditions must be satisfied. First, there must be a sufficient number of independent random variables in the sum. Second, none of the variables may dominate the errors. If these conditions are satisfied, regardless (within some very general limits) of the form of the probability distributions of the individual random variables, the probability distribution of the sum will closely approach a Gaussian distribution. Two parameters, the centroid (mean) and the variance, are required to specify a Gaussian distribution. The centroid of the distribution of the sum is the sum of the centroids of the individual distributions, and the variance is the sum of the individual variances.

$S = \sum_j x_j$, where x_j are independent random variables

$$\langle S \rangle = \sum_j \langle x_j \rangle \text{, and } \sigma^2(S) = \sum_j \sigma^2(x_j)$$

$$p(S) \approx \frac{1}{\sqrt{2\pi\sigma^2(S)}} \exp\left(-\frac{(S-\langle S \rangle)^2}{2\sigma^2(S)}\right)$$

Wilson and Sim distributions in P1

For the Wilson (1949) distribution, we assume that the atoms in a crystal structure in space group P1 are scattered randomly and independently through the unit cell. In fact, it is sufficient to make the much less restrictive assumption that the atoms are placed randomly with respect to the Bragg planes defined by the Miller indices. The assumption of independence is somewhat more problematic since there are restrictions on the distances between atoms, and large volumes of protein crystals are occupied by disordered solvent; for higher order relationships, the lack of independence becomes important (Bricogne, 1993). For the purposes of simpler relationships between the calculated and true structure factors for a single *hkl*, the lack of complete independence does not seem to create serious problems.

The contribution of each atom to the structure factor will have a phase varying randomly from 0 to 2π. The overall structure factor can then be considered to be the result of a random walk in the complex plane (Fig. 3). In terms of the central limit theorem, the structure factor is the sum of the independent atomic contributions, each of which has a distribution defined as a

circle in the complex plane centered on the origin, with a radius of f_j. The centroid of this atomic distribution is at the origin, and its variance is f_j^2. The probability distribution of the structure factor that is the sum of these contributions is thus a two-dimensional Gaussian, referred to as a Wilson distribution (Wilson, 1949); its centroid is at the origin, and its variance is the sum of the individual variances.

$$\mathbf{F} = \sum_{j=1}^{N} f_j \exp\left(2\pi i \mathbf{h}\cdot\mathbf{x}_j\right)$$

$$\langle \mathbf{F} \rangle = \sum_{j=1}^{N} \left\langle f_j \exp\left(2\pi i \mathbf{h}\cdot\mathbf{x}_j\right)\right\rangle = 0$$

$$\left\langle \left|\mathbf{F} - \langle \mathbf{F} \rangle\right|^2 \right\rangle = \sum_{j=1}^{N} \left\langle \left| f_j \exp\left(2\pi i \mathbf{h}\cdot\mathbf{x}_j\right)\right|^2 \right\rangle = \sum_{j=1}^{N} f_j^{\,2} = \Sigma_N$$

The Sim distribution (Sim, 1959), which is relevant when the positions of some of the atoms are known, has a very similar basis, except that the structure factor is now considered to arise from a random walk starting from the position of the structure factor corresponding to the known part, $\mathbf{F}_P$ (Fig. 3). Atoms with known positions do not contribute to the variance, while the atoms with unknown positions (the "Q" atoms) each contribute f_j^2, as in the Wilson distribution. The variance is referred to as Σ_Q.

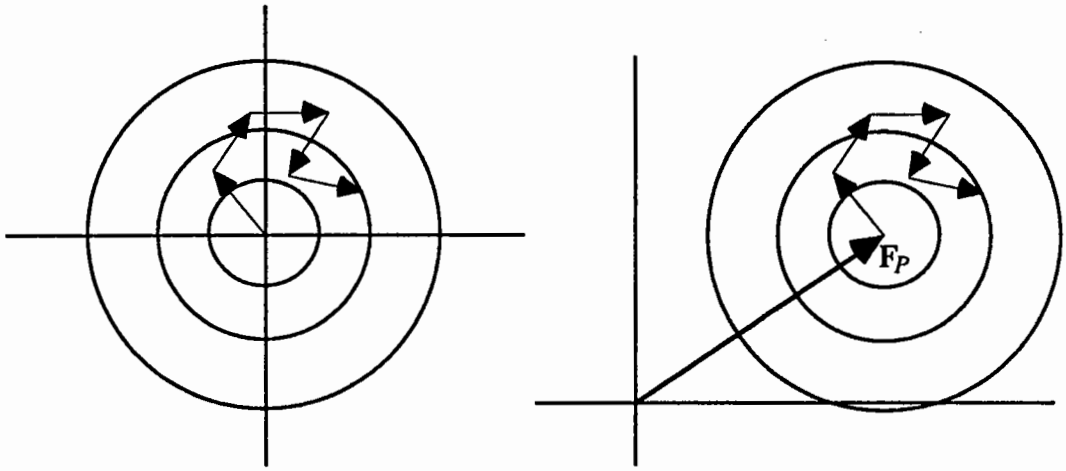


Figure 3. Schematic illustrations of the Wilson (left) and Sim (right) structure factor probability distributions for space group P1.

The Wilson (1949) and Woolfson (1956) distributions for space group P$\bar{1}$ are obtained similarly, except that the random walks are along a line, and the resultant Gaussian distributions are 1-dimensional. (The Woolfson distribution is the centric equivalent to the Sim distribution.) For more complicated space groups, it is a reasonable approximation to treat acentric reflections as following the P1 distribution, and centric reflections as following the P$\bar{1}$ distribution. However, for any zone of the reciprocal lattice in which symmetry related atoms are constrained to scatter in phase, the variances must be multiplied by the expected intensity factor for the zone, because the symmetry related contributions are no longer independent.

Probability distributions for variable coordinate errors

In the Sim distribution, an atom is considered to be either exactly known or completely unknown in its position. These are extreme cases, since there will be varying degrees of uncertainty in the positions of various atoms in a model. The treatment can be generalized by allowing a probability distribution of coordinate errors for each atom. In this case, the centroid for the individual atomic contribution to the structure factor will no longer be obtained by

multiplying by either zero or one. Averaged over the circle corresponding to possible phase errors, the centroid will generally be reduced in magnitude, as illustrated in Fig. 4. In fact, averaging to obtain the centroid is equivalent to weighting the atomic scattering contribution by the Fourier transform of the coordinate error probability distribution, d_j. By the convolution theorem, this in turn is equivalent to convoluting the atomic density with the coordinate error distribution. Intuitively, the atom is smeared over all of its possible positions. The weighting factor, d_j, is thus analogous to the thermal motion term in the structure factor expression.

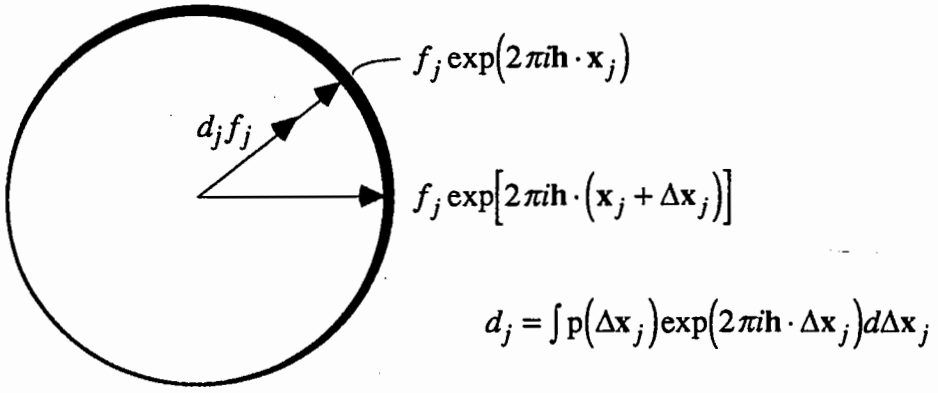


Figure 4. Centroid of the structure factor contribution from a single atom. The probability of a phase for the contribution is indicated by the thickness of the line.

The variances for the individual atomic contributions will vary in magnitude but, if there are a sufficient number of independent sources of error, we can apply the central limit theorem again and assume that the probability distribution for the structure factor will be a Gaussian centered on $\sum d_j f_j \exp(2\pi i \mathbf{h} \cdot \mathbf{x}_j)$ (Fig. 5). If the coordinate error distribution is Gaussian, and if each atom in the model is subject to the same errors, the resulting structure factor probability distribution is the Luzzati (1952) distribution. In this special case, d_j=D for all atoms, and D is the Fourier transform of a Gaussian, which behaves like the application of an overall B-factor.

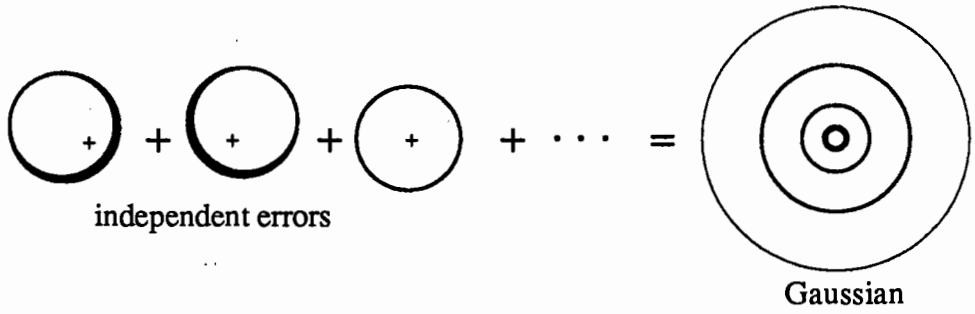


Figure 5. Schematic illustration of the combination of errors in the calculated structure factor. Each small cross indicates the position of the centroid of an atomic contribution; the possible errors arising from using the centroid are vectors from the cross to points on the circle.

General treatment of the structure factor distribution

The Wilson, Sim, Luzzati and variable-error distributions have very similar forms, because they are all Gaussians arising from the application of the central limit theorem. The central limit theorem is valid under many circumstances, and even when there are errors in position, scattering factor and B-factor, as well as missing atoms, a similar distribution still applies. As long as these sources of error are independent, the true structure factor will have a Gaussian distribution centered on $D\mathbf{F}_C$ (Fig. 6), where D now includes effects of all sources of error, as well as compensating for errors in the overall scale and B-factor (Read, 1990).

$$p(\mathbf{F};\mathbf{F}_C) = \frac{1}{\pi\varepsilon\sigma_\Delta^2}\exp\left(-\frac{|\mathbf{F}-D\mathbf{F}_C|^2}{\varepsilon\sigma_\Delta^2}\right)$$ in the acentric case, where

$\sigma_\Delta^2 = \Sigma_N - D^2\Sigma_P$, ε is the expected intensity factor,

and Σ_P is the Wilson variance for the model.

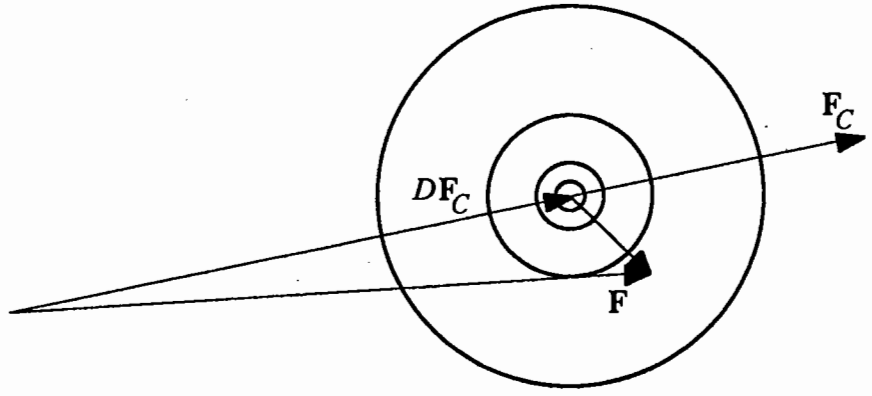


Figure 6. Schematic illustration of the general structure factor distribution relevant in the case of any set of independent random errors.

For centric reflections, the variance is distributed along a line, so the probability distribution is a one-dimensional Gaussian.

$$p(\mathbf{F};\mathbf{F}_C) = \frac{1}{\sqrt{2\pi\varepsilon\sigma_\Delta^2}}\exp\left(-\frac{|\mathbf{F}-D\mathbf{F}_C|^2}{2\varepsilon\sigma_\Delta^2}\right)$$

Estimating σ_A

Srinivasan (1966) showed that the Sim and Luzzati distributions could be combined into a single distribution that had a particularly elegant form when expressed in terms of normalized structure factors, or E values. This functional form still applies to the general distribution reflecting a variety of sources of error; the only difference is the interpretation placed on the parameters (Read, 1990). If we replace $\mathbf{F}$ and $\mathbf{F}_C$ with the corresponding E values, a single parameter σ_A plays the role of D, and σ_Δ^2 reduces to $(1-\sigma_A^2)$. By normalizing the structure factors one also eliminates overall scale and B-factor effects. To characterize the structure factor probability distribution, σ_A must be estimated as a function of resolution. This parameter must be deduced from $|\mathbf{F}_O|$ and $|\mathbf{F}_C|$, since the phase (thus the phase difference) is unknown.

A general approach to estimating parameters for probability distributions is to maximize a likelihood function. The likelihood function is the overall joint probability of making the entire set of observations, as a function of the desired parameters. The parameters that maximize the probability of making the set of observations are the most consistent with the data.

In this case, we maximize the joint probability of making the set of observations of $|\mathbf{F}_O|$. If we assume that the structure factors are all independent, the joint probability distribution is the product of all the individual distributions. This assumption is not completely justified in theory, but the results are fairly accurate in practice.

$$L = \prod_{\mathbf{h}} p(|\mathbf{F}_O|;|\mathbf{F}_C|)$$

The required probability distribution, $p(|F_O|;|F_C|)$, is derived from $p(\mathbf{F};\mathbf{F}_C)$ by integrating over all possible phase differences and neglecting the errors in $|F_O|$ as a measure of $|F|$. The form of this distribution differs for centric and acentric reflections, and is given in other publications (Read, 1986, 1990). It is more convenient to deal with a sum than a product, so the log likelihood function is maximized instead. In the program SIGMAA, reciprocal space is divided into spherical shells, and a value of the parameter σ_A is estimated for each resolution shell. Details of the algorithm can be found in Read (1986).

4. Figure of merit weighting for model phases

Blow and Crick (1959) and Sim (1959) showed that the electron density map with the least rms error is calculated from centroid structure factors. This conclusion follows from Parseval's theorem, since the centroid structure factor (its probability-weighted average value, or expected value) minimizes the rms error of the structure factor in the complex plane (in the acentric case). Since the structure factor distribution $p(\mathbf{F};\mathbf{F}_C)$ is symmetrical about $\mathbf{F}_C$, the expected value of $\mathbf{F}$ will have the calculated phase, but the averaging over different possible phases around the phase circle will reduce its magnitude if there is any uncertainty in the phase value (Fig. 7). We treat the reduction in magnitude by applying a weighting factor called the figure of merit, m, which is equivalent to the expected value of the cosine of the phase error.

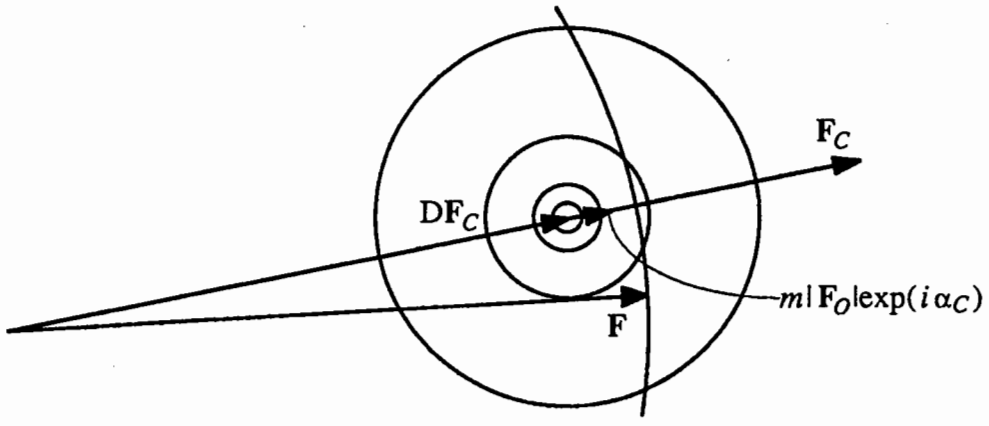


Figure 7. Figure of merit weighted model-phased structure factor, obtained as the probability-weighted average over all possible phases.

5. Map coefficients to reduce model bias

Model bias in figure-of-merit weighted maps

A figure of merit weighted map, calculated with coefficients $m|\mathbf{F}_O|\exp(i\alpha_C)$, has the least rms error from the true map. In a quantitative sense, then, it is the best map. However, such a map will suffer from model bias; if its purpose is to allow us to detect and repair errors in the model, that is a serious qualitative defect. Fortunately, it is possible to predict the systematic errors leading to model bias, and make some correction for them.

Main (1979) dealt with this problem in the case of a perfect partial structure. Since the relationships among structure factors are the same in the general case of a partial structure with various errors, once $D\mathbf{F}_C$ is substituted for $\mathbf{F}_C$, all that is required to apply his results more generally is a change of variables (Read, 1986, 1990).

In Main's (1979) approach, the cosine law is used to introduce the cosine of the phase error, which is converted into a figure of merit by taking expected values. Some manipulations (see Main, 1979 or Read, 1986 for the details) allow us to solve for the figure-of-merit weighted map coefficient, which is approximated as a linear combination of the true structure factor and the model structure factor. Finally, we can solve for an approximation to the true

structure factor, giving map coefficients from which the systematic model bias component has been removed.

$$m|\mathbf{F}_O|\exp(i\alpha_C) = \frac{\mathbf{F}}{2} + \frac{D\mathbf{F}_C}{2} + \textit{noise terms}$$

$$\mathbf{F} \approx (2m|\mathbf{F}_O| - D|\mathbf{F}_C|)\exp(i\alpha_C)$$

Examples of such maps, showing the qualitative effect on interpretability, can be found in Read (1986).

Model bias in combined phase maps

When model phase information is combined with, for instance, MIR phase information, there will still be model bias to the extent that the model influences the final phases. However, it is not, in general, appropriate to continue to use the same map coefficients to reduce model bias, because some phases could be almost completely determined by the MIR phase information. It makes much more sense to have map coefficients that reduce to the coefficients appropriate for either model or MIR phases, in the extreme cases where there is only one source of phase information, and that vary smoothly between those extremes.

Map coefficients that satisfy these criteria, even if they are not rigorously derived, are implemented in the program SIGMAA, and the resulting maps are reasonably successful in reducing model bias. Two assumptions are made: 1) the model bias component in the figure of merit weighted map coefficient, $m_{com}|\mathbf{F}_O|\exp(i\alpha_{com})$, is proportional to the influence that the model phase has had on the combined phase; and 2) the relative influence of a source of phase information can be measured by the information content H (Guiasu, 1977) of the phase probability distribution. The first assumption corresponds to saying that the figure-of-merit weighted map coefficient is a linear combination of the MIR and model phase cases.

MIR: $m_{MIR}|\mathbf{F}_O|\exp(i\alpha_{MIR}) \approx \mathbf{F}$

Model: $m_C|\mathbf{F}_O|\exp(i\alpha_C) \approx \frac{\mathbf{F}}{2} + \frac{D\mathbf{F}_C}{2}$

Combined: $m_{com}|\mathbf{F}_O|\exp(i\alpha_{com}) \approx (1-\frac{w}{2})\mathbf{F} + \frac{w}{2}D\mathbf{F}_C$

where $w = \dfrac{H_C}{H_C + H_{MIR}}$

and $H = \int_0^{2\pi} p(\alpha)\ln\dfrac{p(\alpha)}{p_0(\alpha)}d\alpha;\ p_0(\alpha) = \dfrac{1}{2\pi}$

When we solve for an approximation to the true **F**, we obtain the following expression, which can be seen to reduce appropriately when w is 0 (no model influence) or 1 (no MIR influence).

$$\mathbf{F} \approx \frac{2m|\mathbf{F}_O|\exp(i\alpha_{com}) - wD\mathbf{F}_C}{2 - w}$$

Features of the program SIGMAA

The computer program SIGMAA (Read, 1986) has been developed to implement the results described here. Model phase probabilities can optionally be combined with phase information from other sources such as MIR, using Hendrickson-Lattman coefficients (Hendrickson and Lattman, 1970), and four different types of map coefficient can be

produced. Apart from the two types of coefficient discussed above, there are also two types of difference map coefficient:

1) Model-phased difference map: $(m|\mathbf{F}_O| - D|\mathbf{F}_C|)\exp(i\alpha_C)$

2) General difference map: $m_{com}|\mathbf{F}_O|\exp(i\alpha_{com}) - D\mathbf{F}_C$

The general difference map, it should be noted, uses a vector difference between the figure-of-merit weighted combined phase coefficient (the "best" estimate of the true structure factor), and the calculated structure factor. When additional phase information is available, it should provide a clearer picture of the errors in the model.

6. Refinement bias

The structure factor probabilities discussed above depend on the atoms having independent errors (or at least a sufficient number of groups of atoms having independent errors). Unfortunately, this assumption breaks down when a structure is refined against the observed diffraction data. Few protein crystals diffract to sufficiently high resolution to provide a large number of observations for every refinable parameter. The refinement problem is, therefore, not sufficiently over-determined, so it is possible to overfit the data. If there is an error in the model that is outside the range of convergence of the refinement method, it is possible to introduce compensating errors in the rest of the structure to give a better, and misleading, agreement in the amplitudes. As a result, the phase accuracy (hence the weighting factors m and D) are overestimated, and model bias is poorly removed. Because simulated annealing is a more effective minimizer (Brünger *et al.*, 1987), it is also more effective at locating local minima, so structures refined by simulated annealing probably tend to suffer more severely from refinement bias.

There is another interpretation to the problem of refinement bias. As Silva and Rossmann (1985) point out, minimizing the rms difference between the amplitudes $|\mathbf{F}_O|$ and $|\mathbf{F}_C|$ is equivalent (by Parseval's theorem) to minimizing the difference between the model electron density and the density corresponding to the map coefficients $|\mathbf{F}_O|\exp(i\alpha_C)$; a lower residual is obtained either by making the model look more like the true structure, or by making the model-phased map look more like the model through the introduction of systematic phase errors.

It is possible to deal with refinement bias in a number of ways. First, the effect can be reduced by placing less weight on the agreement of structure factor amplitudes. Anecdotal evidence suggests that the problem is less serious, in structures refined using X-PLOR (Brünger *et al.*, 1987), when the Engh and Huber (1991) parameter set is used for the energy terms. In this new parameter set, the deviations from standard geometry are much more strictly restrained so, in effect, the pressure on the agreement of structure factor amplitudes is reduced.

If errors are suspected in certain parts of the structure, "omit refinement", in which the questionable parts are omitted from the model, can be a very effective way to eliminate refinement bias in those regions (*e.g.* James *et al.*, 1980; Hodel *et al.*, 1992).

If MIR phases are available, combined phase maps tend to suffer less from refinement bias, depending on the extent to which the MIR phases influence the combined phases. Finally, it is always a good idea to refer occasionally to the original MIR map, which cannot suffer at all from model bias or refinement bias.

Acknowledgments

Bart Hazes and Allan Sharp helped greatly by pointing out where the presentation could be clarified. Of course, I am responsible for any remaining lack of clarity. This work was supported by the Medical Research Council of Canada, a Scholarship from the Alberta Heritage Foundation for Medical Research and, in part, by an International Research Scholar award from the Howard Hughes Medical Institute.

References

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

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

Brünger, A.T., Kuriyan, J. and Karplus, M. (1987). Science **235**: 458-460.

Engh, R.A. and Huber, R. (1991). Acta Cryst. **A47**: 392-400.

Guiasu, S. (1977). *Information Theory With Applications*, McGraw-Hill, London.

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

Hodel, A., Kim, S.-H. and Brünger, A. T. (1992). Acta Cryst. **A48**: 851-858.

James, M.N.G., Sielecki, A.R., Brayer, G.D., Delbaere, L.T.J. and Bauer, C.-A. (1980). J. Mol. Biol. **144**: 43-88.

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

Main, P. (1979). Acta Cryst. **A35**: 779-785.

Oppenheim, A.V. (1981). Proc. IEEE **69**: 529-541.

Ramachandran, G.N. and Srinivasan, R. (1961). Nature **190**: 159-161.

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

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

Silva, A.M. and Rossmann, M.G. (1985). Acta Cryst. **B41**: 147-157.

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

Srinivasan, R. (1966). Acta Cryst. **20**: 143-144.

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

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

Entropy Maximization, Permutation and Likelihood Scoring Methods for Improving Macromolecular Electron Density Maps

C. W. Carter, Jr.
Department of Biochemistry and Biophysics, CB # 7260
University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7260

Entropy maximization to maximum likelihood constrained jointly by the best available experimental phases and by a sufficiently good envelope (**M**aximum **E**ntropy **S**olvent **F**lattening; MESF) can bring about substantial, model-independent map improvement, even at medium (3.1 Å) resolution (Xiang, Carter, Bricogne & Gilmore, 1993). Moreover, in the structure determination of *Bacillus stearothermophilus* tryptophanyl-tRNA synthetase (TrprS; Mr ~ 37,000; Doublié, 1993; Doublié, Xiang, Gilmore, Bricogne & Carter, 1994) MESF was used more actively, to determine unknown phases directly. In that work, critical but missing information about the correct phase and molecular envelope constraints was recovered from the native amplitudes and the incorrect phases by the Bayesian strategy of hypothesis permutation and likelihood scoring, using data to only 3.0 Å resolution. The resulting structure of TrpRS documents the successful application of maximum entropy methods to the solution of an unknown structure. The effectiveness of permutation, scoring and statistical inference in a real, non-trivial case also provides a paradigm for more ambitious direct phase determination for macromolecular structures as outlined by Bricogne (Bricogne, 1984; Bricogne, 1988a; Bricogne, 1991b; Bricogne, 1993). At the same time, this paradigm already represents a powerful way to improve electron density maps based on conventional sources of phase information.

Bricogne (Bricogne, 1988b) has given an especially accessible overall description that does not rely heavily on difficult constructs necessary for mathematical proof. There are also strong heuristic connections between MESF and conventional density modification algorithms. Nevertheless, the strategies and procedures are different enough to warrant a detailed introductory review of how entropy maximization, permutation and likelihood scoring are used in the Bayesian approach. To see how the paradigm works, the basic concepts at work in the computer program, MICE (Bricogne & Gilmore, 1990), will be illustrated with results of calculations for a simulated, one-dimensional example for which all maps and structure factors can be plotted. This introduction will be followed by examples of the use of MICE for map improvement.

1. The paradigm: a one-dimensional example.

Direct methods use conditional probability to exploit the complex statistical coupling that exists between conspicuously strong structure factors. The example given by Bricogne (Bricogne, 1984) involving a single structure factor is extended here for a one dimensional problem to introduce the new nomenclature. To provide chemical sense, a linear triatomic molecule, $HgCl_2$, was placed along the X axis of a crystal in space group P_1, the electron density and structure factors ({F}) shown below were generated using standard programs (Daresbury Laboratory, 1990). Unitary structure factors ({U}) were obtained using default normalization options in MITHRIL (Gilmore, 1984; Gilmore & Brown, 1988). All subsequent calculations were carried out with MICE, and the axial electron density and structure factors were taken to represent a one-dimensional case by virtue of the projection theorem.

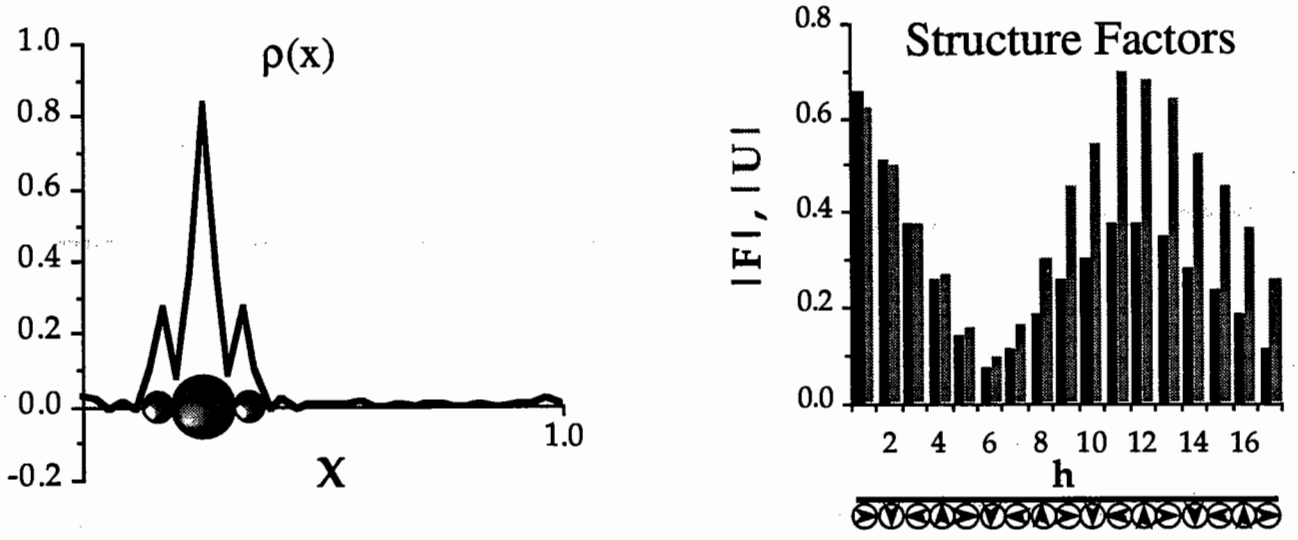


Figure 1. The (one-dimensional) electron density of a linear triatomic molecule, together with observed and unitary structure factors derived by Fourier transformation of a one-dimensional crystal with unit cell dimension a=1.0. Phases are indicated by the dials below the **h** axis.

The method of conditional probability seeks to describe the distribution of atoms *on the condition* that the phase and amplitude are known for some reflections. Here, reflection [h] = [1] was arbitrarily assigned a phase of 90 degrees, since it is the strongest low resolution reflection in the dataset. The optimal probability under this constraint is given by a normalized distribution, $qME(x) = \frac{1}{C} \exp\{\zeta_{[1]}\, e^{-2\pi i x}\}$, obtained by exponentiating the product,

$\omega = \zeta_{[1]}\, e^{-2\pi i x \bullet 1}$, where ζ_1 is an adjustable, complex parameter chosen such that the inverse transform of q(x) agrees in amplitude and phase with the constraining structure factor vector, $\{U_{[1]}\}$, and the constant C is chosen so that the integrated probability along the unit cell is unity. Bricogne, 1984) proved (in three dimensions) that $q^{ME}(\mathbf{x})$ simultaneously maximized the conditional probability for the distribution of atomic positions with respect to a uniform distribution m(**x**), *and* the integrated value of the relative entropy of the two distributions,

$S = -\int q(\mathbf{x}) \log [q(\mathbf{x})/m(\mathbf{x})]\ d^3\mathbf{x}$, subject to the constraints of known structure factors. An initial guess for qME(x), obtained by setting $\zeta^0 = U_{[1]}$, is illustrated in Figure 2, together with the one-term Fourier series, ω^0, for $U_{[1]}$.

This one-dimensional example has the following properties:

The probability distribution, q(x), and its structure factors

- Since it is a normalized probability distribution, its integral over the unit cell is 1.0.
- Its inverse Fourier transform gives *unitary* structure factors, $\{U_{calc}\}$.
- The real part of the function $\omega^0 = \zeta^0_{[1]}\, e^{-2\pi i x}$, is a cosine wave and hence has both positive and negative values. However, $q^0_{[1]}(x)$ is positive everywhere, with a peak at the maximum of ω. It is therefore a better probability estimate for finding atoms in the structure than is the "electron density map" calculated as the one-term Fourier series $\omega = U_{[1]}\, e^{-2\pi i x}$.
- The amplitude and phase of the parameter, $\zeta_{[1]}$, must be fitted to satisfy the constraint.

One algorithm for finding the appropriate value for $\zeta_{[1]}$ is to start with $\zeta_{[1]}$ equal to the known structure factor $U_{[1]}$ and evaluate successively the difference Fourier map with coefficient $\{U^1_{obs} - U^1_{calc}\}$, dividing by q(x) at each point, and then back transforming to get the shift vector, $\Delta\zeta_1$. This process of iterating real (exponentiation) and reciprocal (U_{obs}) space constraints is called "exponential modeling". Other methods can be used to solve the maximum entropy

equations (Bricogne, 1993), but exponential modeling provides a heuristic parallel between constrained entropy maximization and more familiar density modification algorithms.

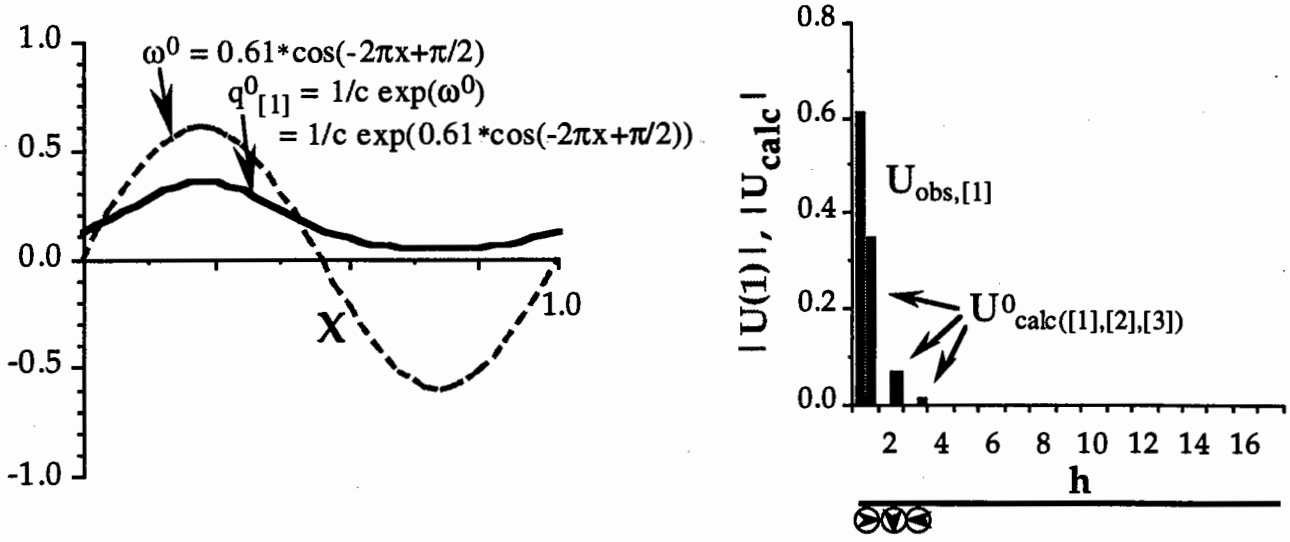


Figure 2. The fundamental density modification involved in expressing the conditional probability for the distribution of atoms given a phase of $\pi/2$ for the [1] reflection. The light grey trace is the one-term Fourier synthesis, $\omega(x)$, the black trace is the probability density, $q(x)$, obtained by exponentiating $\omega(x)$. Structure factors for both are indicated to the right as in Figure 1. Structure factors for the [2] and [3] reflections are extrapolations resulting from the exponentiation.

Maximum entropy extrapolation

- Exponentiation modifies the density. Thus, although the transform of $\omega(x)$ has a single nonzero structure factor, $\zeta_{[1]}$, that of $q(x)$ has richer spectrum of structure factors, the "extrapolated" structure factors being precisely related to the density modification. In this example, even the first iteration extrapolates significant structure factors for reflections [2] and [3] with the correct phases. The parameter, $\zeta^{ME}_{[1]}$, having been found, the distribution qME(x) has achieved the desired properties of an exponential (hence, maximum entropy) model whose Fourier coefficients agree in amplitude and phase with the constraint, $U_{[1]}$ (Figure 3).

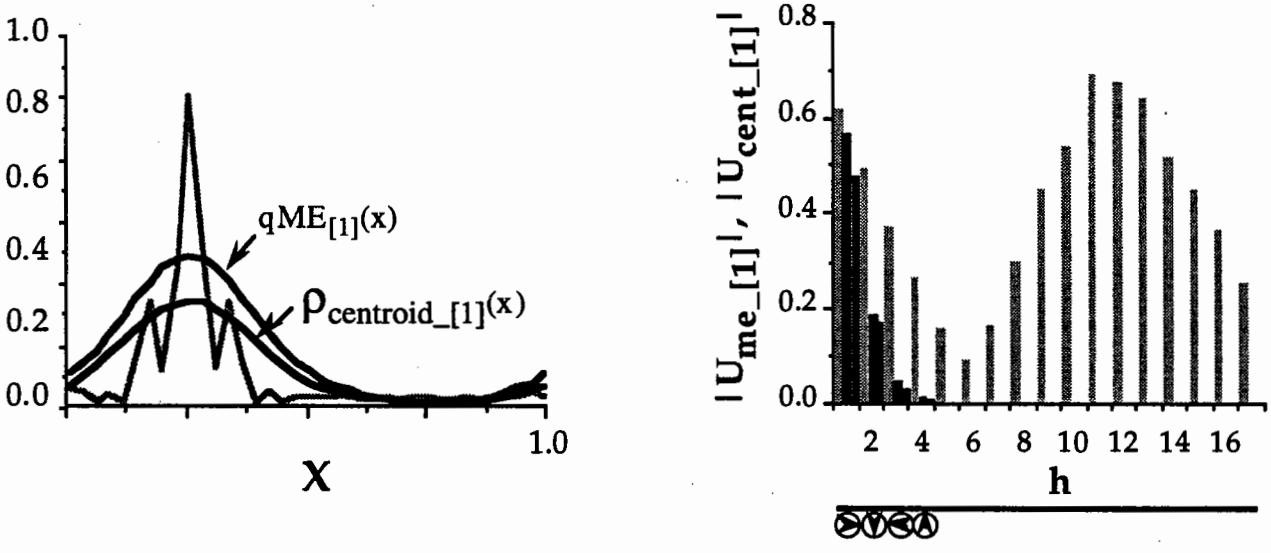


Figure 3. The constrained maximum entropy distribution, $q^{ME}_{[1]}(x)$ and the centroid map, $\rho_{centroid\_[1]}(x)$ and their structure factors. The Fourier coefficient in the exponent of the exponential model, qME[1](x), has been adjusted so that the amplitude and phase of $U_{me\_[1]}$ agree with the "basis set" U_{obs}. Coefficients of the centroid map consist of Sim-weighted amplitudes and calculated phases, ϕ_{me}.

The centroid electron density map

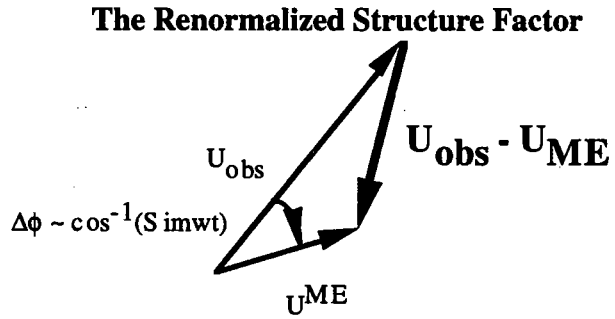
- The "extrapolated" Fourier coefficients have an immediate interpretation in the sense that the conditional probability for all structure factors outside the constraint, $U_{[1]}$, is a multivariate Gaussian centered around these values. Among the extrapolated structure factor amplitudes, some will agree better, others worse with the observed (and unphased) data. Reflections with better agreement between $|U_{obs}|$ and $|U_{calc}|$ are more reliable extrapolations than those with poorer agreement, and the uncertainty spread around the extrapolated structure factor values can be expressed quantitatively by "Sim-like" weights,

$$Wt_{sim} = \tanh(\frac{N}{\varepsilon}|U^i_{obs}||U^i_{calc}|).$$

- The observed amplitudes are weighted and combined with the extrapolated phases to provide an estimate for the electron density, $\rho_{centroid\_[1]}(x)$, implied by the probability $qME_{[1]}(x)$. This map clearly provides an adequate representation of the low-resolution structure.

Completing the structure by phase permutation and likelihood scoring

The pattern of maximum entropy extrapolation from the [1] reflection is restricted to the first few orders. Specifying the one phase is insufficiently constraining to reach the strong reflections evident at higher resolution. Under these circumstances the only way to complete the structure determination is to supplement the information provided by the one reflection. Phases can be determined for the higher order terms by selecting from among the unphased reflections those with the greatest potential to seed the unpopulated region of reciprocal space with the power to extend phases by maximum entropy extrapolation. In keeping with the Bayesian paradigm, these "vanguard" reflections are identified by examining the pattern of "renormalized" structure factor amplitudes (Bricogne, 1992) estimated from the amplitudes $|U_{obs}|$ and $|U_{ME}|$ and making use of the Sim-like weight to estimate the phase difference (Figure 4).



$$|U_{obs} - U_{ME}|^2 \sim |U_{obs}|^2 + |U_{ME}|^2 - 2|U_{obs}||U_{ME}|\,Sim\,wt$$

Figure 4. Weak or incorrect extrapolation gives rise to large "renormalized" structure factor amplitudes outside the basis set. Comparing renormalized amplitudes reveals their relative importance.

The list of renormalized structure factor amplitudes indicates that the largest of these is that for [h]=[11]. Although in general there is a maximum entropy phase for these reflections (because the spectrum of qME(x) can be evaluated anywhere) it is folly to follow that indication because its Sim-weight is essentially zero. Thus, the phase must be *permuted,* generating a "phasing tree" with a separate branch for each distinct hypothesis about the phase. In this example, the [11] was given phases of 90, 180, 270, and 360 degrees, and separate exponential models were fitted for each hypothesis (Figure 5).

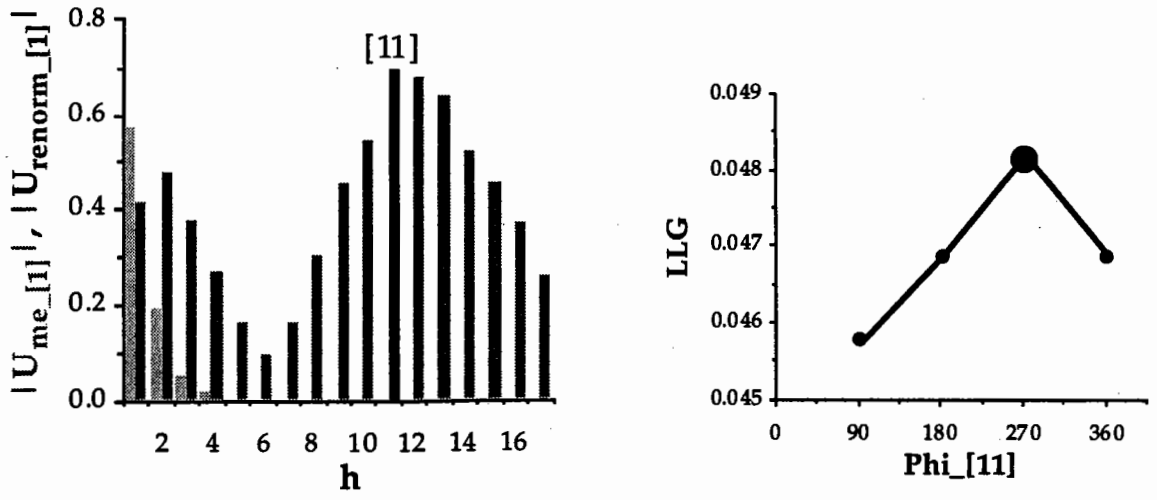


Figure 5. Finding the biggest missing piece. A suitable reflection is chosen for phase permutation by examining the pattern of "renormalized" structure factor amplitudes. Its phase is determined by permutation and likelihood scoring. The log-likelihood gain score is a maximum when the trial phase is 270°.

Likelihood

The overall agreement between extrapolated and observed structure factor amplitudes is quantified by evaluating the likelihood. This statistic is proportional to the probability of measuring the experimental data when the random atoms obey the distribution qME(x). That condition corresponds to the hypothesis, H_1, that the exponential modeling parameters have the values, $\{\zeta_i^{ME}, i=1,11\}$. The likelihood is more useful when compared to the probability of measuring the same data under the null hypothesis, H_0, that the random atoms are uniformly distributed in the cell and hence that the data follow Wilson statistics.

Likelihoods are evaluated by integrating the conditional probability for the unphased structure factors over all possible values of their phases, giving the conditional probability of their amplitudes. Summing over all reflections and taking the difference between the respective logarithms gives the global log-likelihood gain (LLG; Bricogne, 1988b; Bricogne & Gilmore, 1990):

$$L(U^K) = \sum_{\mathbf{k}\ \text{acentric}\in K} \{\log I_0[(2N/\varepsilon_{\mathbf{k}})|U_{\mathbf{k}}^{obs}||U_{\mathbf{k}}^{ME}|] - N/\varepsilon_{\mathbf{k}}|U_{\mathbf{k}}^{ME}|^2\}$$

$$L(U^K) = \sum_{\mathbf{k}\ \text{centric}\in K} \{\log \cosh[(N/\varepsilon_{\mathbf{k}})|U_{\mathbf{k}}^{obs}||U_{\mathbf{k}}^{ME}|] - N/2\varepsilon_{\mathbf{k}}|U_{\mathbf{k}}^{ME}|^2\} \qquad [II]$$

The LLG is also sometimes called the "support" for the hypothesis, H_1, relative to H_0 (Edwards, 1972). It is a cross-validation statistic, sharing with the free R-value (Brunger, 1993; Brunger, 1992a) the fact that it is evaluated using reflection data that were not included in the fitting of the exponential model.

Scoring basis set assumptions and the Neyman-Pearson Theorem

The ratio of likelihoods for qME(x) and the uniform distribution of atoms describes the improvement obtained when known values are assigned by hypothesis to some structure factors, referred to as a "basis set". It reflects the ability of the corresponding hypothesis to assign a high probability to measurements outside the basis set before actually knowing about them. Thus, it provides a way to score competing phase permutation experiments from the relative increase they bring to the LLG.

In the present example the same data are used in all four cases, so the LLG can be considered as a function of competing hypotheses, $\{H_1^i,$ i=90, 180, 270, and 360 degrees$\}$ on the phase of $U_{[11]}$. The LLG has a privileged status among statistics on the basis set phase choices by virtue of the Neyman-Pearson theorem, which states that it is a "most powerful" indicator of the relative correctness of the model, in the sense that it is minimally vulnerable to statistical errors of the second kind (acceptance of the null hypothesis when it should be rejected) at any given level of exposure to errors of the first kind (rejection of the null hypothesis when it should be affirmed) (Bricogne, 1991a). In practice, it is a superior statistic for comparing hypotheses about the structure. Previous tests have demonstrated the unparalled ability of the LLG to identify correct phase sets from among a large number generated for a small protein structure by conventional direct methods (Gilmore, Henderson & Bricogne, 1991).

The overriding importance of the LLG in this context is that it provides a way to convert hypotheses about phases of some reflections into *calculable* scores that depend on the measured amplitudes of all the others. It can therefore be used to make statistical inferences about the phases. Here, the LLG assumes a maximum value when the phase of [11] is equal to 270 degrees. This is, in fact, the correct phase, when the [1] reflection has a value of 90 degrees.

The centroid map based on the $q_{ME\_[1],\,[11]}(x)$ probability distribution is remarkably close to the target map (Figure 6). The only features above the significance level are three peaks at or greater than 2.5σ, which correspond to the atomic positions of the chlorine and mercury atoms.

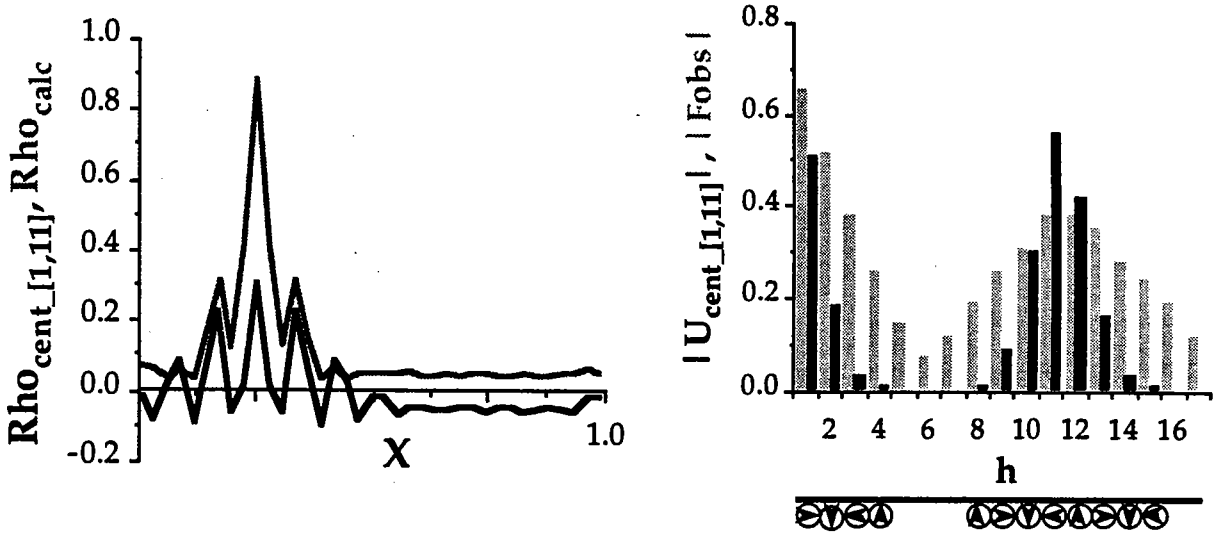


Figure 6. Maximum entropy extrapolation from two basis set reflections is sufficient to solve the structure. The centroid map (black) displays the essential features of the "target" map. Phases of extrapolated reflections on either side of the [11] reflection are correct.

Maximum entropy extrapolation from the two reflections, indicated by the (Sim-weighted) centroid structure factor amplitudes on the right in the above figure, shows that the [11] reflection has "recruited" significant contributions from four or five additional reflections. The mean absolute phase error for orders 2 through 16 is 2 degrees, so all extrapolated reflections have almost exactly the correct phases. Thus, the extrapolation now correctly accounts for nearly all the strong diffraction in the dataset.

A connection with traditional direct methods.

A key feature of this paradigm is the ability to identify and deal effectively with the most likely sources of missing information at any node, thereby providing a reliable "agenda" for building up correct phase information in a prioritized, stepwise fashion. The central notion in this prioritization is the renormalized structure factor amplitude, which provides an estimate for the size of the vector difference between the observed structure factor and that obtained by maximum entropy extrapolation (Bricogne, 1992). Reflections for which the observed structure factor, $|U_{obs}|$, is strong but the extrapolated $|U_{ME}|$ is weak represent the most significant of the

measured data which the current hypotheses are essentially powerless to anticipate. For this reason, they are particularly good reflections to add to the basis set by permutation. In this sense, their use is analogous to choosing reflections to permute from those with the strongest normalized structure factors in a direct phase determination for a small molecule (Germain, Main & Woolfson, 1970). However, since the conditional probability distribution in this Bayesian paradigm has been updated to reflect the information already contained in the basis set, the list of *renormalized* structure factors provides an improved selection criterion for phase permutation.

2. The MICE program

This one-dimensional example provides the necessary terminology to understand the use of the MICE program in a macromolecular crystal structure determination. Normally in such a situation phase information is already available from a variety of sources and is not generated *ab initio* . Bricogne has shown that each of these sources of phase information can be combined in an optimal way by including their impact on the joint probability for the distribution of atoms (Bricogne, 1991b; Bricogne, 1988a). At any stage of phase determination, the symmetry-unique non-origin reflections are divided into two sets. The *basis set* , {H}, consists of reflections for which reliable phase assumptions are available, and for which associated parameters, $\{\zeta_1^{ME}\}$, are to be fitted. The complementary set {K} includes *non-basis* reflections for which only unphased or poorly phased amplitudes are available and for which it is hoped that maximum entropy extrapolation will provide improved phases.

In favorable cases maximum entropy solvent flattening alone, representing the first stage of the paradigm outlined in section 1, can substantially improve macromolecular electron density maps, greatly easing the interpretation and providing a substantial measure of security against incorrect models. It also turns out that when the extrapolation is inadequate to solve the structure, the second half of the paradigm described in section 1, namely phase permutation and likelihood scoring, provides a practical way to complete the structure determination by enlarging the basis set with quite reliable, directly phased reflections.

MICE uses exponential modeling, summarized in Figure 7, to determine values for the set of adjustable parameters, $\{\xi_i^{ME}\}$, associated with the basis set reflections:

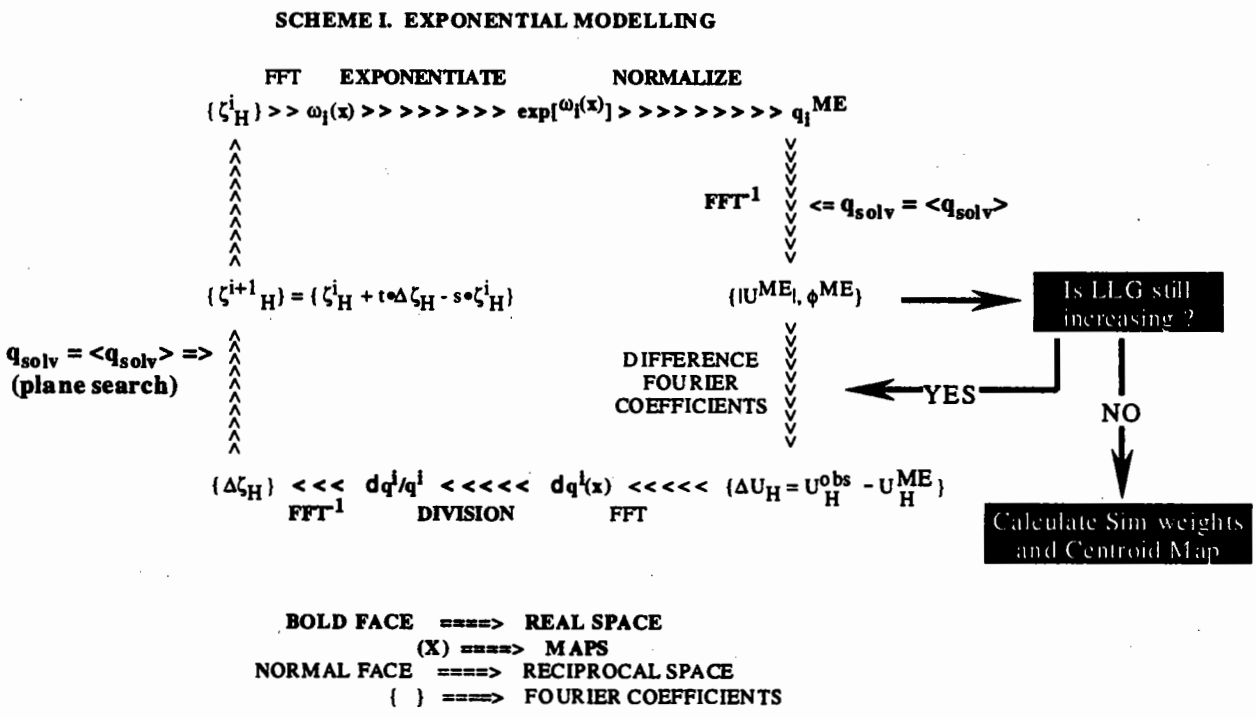


Figure 7. Schematic of the exponential modeling and solvent flattening algorithm in MICE.

Maximum entropy solvent flattening.

As suggested in section 1, optimal evaluation of both the *a priori* probability and the likelihood of a hypothesis in which some phase values and a molecular boundary are specified is intimately related to the numerical process of constrained entropy maximization. If m(**x**) denotes the uniform probability distribution within the protein region, the goal is to construct an exponential model, $q^{ME}(\mathbf{x})$, for the distribution of atoms:

$$q^{ME}(\mathbf{x}) = \frac{m(\mathbf{x})}{Z(\zeta)} \exp \sum_{\mathbf{h} \in \{H\}} \zeta_{\mathbf{h}} e^{-2\pi i \mathbf{h} \cdot \mathbf{x}} \qquad [I]$$

where the exponent is a Fourier synthesis over the basis set reflections whose coefficients are the exponential modeling parameters, $\zeta_{\mathbf{h}}$. Structure factors from a macromolecular crystal are strongly attenuated by the presence of solvent atoms, which reduce the contrast at low resolution (Bricogne, 1988a; Carter, Crumley, Coleman, Hage & Bricogne, 1990; Bricogne, 1991b). Because the solvent atoms have very different properties from those inside the molecular envelope the distribution m(**x**) in [I] cannot adequately represent their effect. The appropriate algorithms involved in solving the maximum-entropy equations under the constraint of solvent flatness have been described (Bricogne, 1988a; Bricogne, 1991b) and in principle, treatment of the solvent atoms requires a second, maximum entropy "channel" to impose the envelope constraint in advance. However, it is possible to approximate the effect of such a second channel by averaging the solvent regions on each cycle of fitting the exponential model to the basis set structure factor constraints, as indicated in the scheme above. We have used this approximation quite successfully (Xiang, Carter, Bricogne & Gilmore, 1993). Although weaker than the algorithm described for multichannel entropy maximization (Bricogne, 1988a), it was shown with both simulated and experimental data to be an effective approximation.

3. Running MICE

Mice reads one essential binary file, prepared by MITHRIL and containing the data (treated by MICE as normalized structure factor amplitudes) and supplementary optional input files, including:

- an ascii file with phases from which the basis set is chosen
- an ascii file with phase probability coefficients for phased likelihood and phase recombination
- a binary file with a full unit cell of the envelope map
- a binary file with a full unit cell of a reference or target map

Output files created by MICE include a binary file containing the exponential modeling parameters for all nodes, together with optional output files, including:

- log and abbreviated summary files
- an ascii file with Uobs, U^{ME}, Sim-weights, phase probability coefficients and renormalized structure factor amplitudes
- binary files with a full unit cell of either the q^{ME} or centroid maps

Data and Constraints

- **Resolution of the data.** Maps calculated internally by MICE must be sampled on a sufficiently fine grid to avoid aliasing errors in estimating Fourier coefficients at the resolution limit of the reflections used in calculating the likelihood. Both the grid of the internal maps and the maximum resolution for reflections contributing to the LLG can be specified by the user. Some care must be taken in specifying these parameters because the envelope maps must currently be calculated externally to MICE on a pre-determined grid which must then match that used inside MICE. For this reason, it is useful to first run one cycle of MICE without an envelope, specifying a resolution limit with the DMIN parameter and letting MICE determine its default grid. That grid can then be used to calculate the envelope.

• **Chosing a basis set.** Much work remains to be done in this respect. We have worked with a number of different problems, including both simulated and experimental datasets from known and unknown structures. These fall into two general categories, according to the quality of the experimental phases and the correctness of the envelope. Cases in which that information is sufficient to produce an interpretable map through the application of conventional solvent flattening seem to respond best if the basis set is chosen as we did for cytidine deaminase (Xiang, Carter, Bricogne & Gilmore, 1993), by choosing reflections whose MIR(AS) figure of merit, calculated with the program MLPHARE, is greater than about 0.7. In general, those reflections have a mean phase error of around 30°, and for CDA they amounted to nearly 35% of the total number of phased reflections to 3.1Å.

When the phase and envelope constraints are insufficient to generate an interpretable map in this manner we have found it necessary and useful to identify reflections (both within and outside the basis set) for which the renormalized strucure factor amplitude is large and permute them 6 to 10 at a time, depending on the total number of bits of phase information.

• **Chosing an envelope.** Generation of a suitable envelope from good phases can usually be done using the algorithms of Wang (Wang, 1985) and Leslie (Leslie, 1988). However, we have found that such envelopes often benefit from manual editing (Minor, 1992). The envelope must be formatted for use by MICE. To minimize disk usage, envelope files can be compressed or stored as ascii files representing the boundary (Xiang, unpublished) and regenerated when needed.

Scaling and normalization

The most important practical task in setting up to run MICE is to find an appropriate scale for the observed data. The importance of having a nearly correct scale is illustrated in Figure 8. If the data and hence the U_{000}^{obs} are too weak (a), relative to absolute scale, then the fitting process will stop prematurely, before the full effect of the constraints can be realized. On the other hand, if the data are too strong, then the exponentiation will truncate real features in the distribution of atoms (c) and fitting the constraints will involve distorting the resulting distribution. In either case, the LLG statistic will be lower than that for the correct scale, and this provides a crude, but effective way to establish an approximate scale for the data (Xiang, Carter, Bricogne & Gilmore, 1993). So far, we have not used any sharpening or other normalization, relying only on a linear scale factor to prepare the amplitude data for input to MICE, although this needs to be tested.

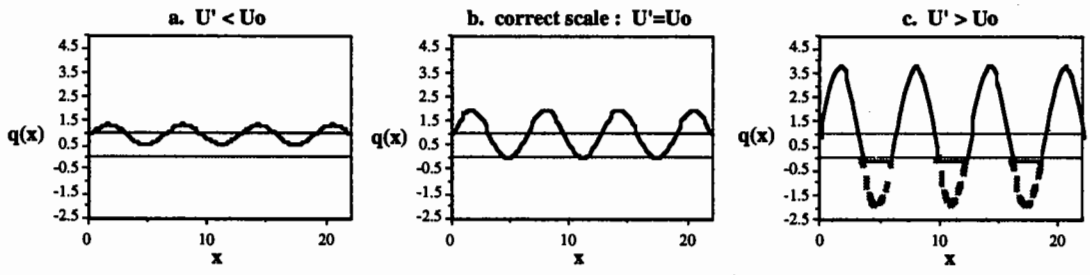


Figure 8. Problems associated with using incorrect absolute scales for the amplitude data.

Although maximum likelihood algorithms have been developed for determining the correct scale for the data (Bricogne, 1988a; Bricogne, 1993) these have not yet been incorporated into MICE. Amplitude data must be prepared for input to MICE using the normalization routine in MITHRIL. For protein datasets we bypass the calculation of normalized structure factors and produce directly a file of scaled |Fobs| values. A semi-empirical scaling step works effectively in most cases. This procedure arose from studies on simulated datasets (Xiang, Carter, Bricogne & Gilmore, 1993) of the effect of the scale factor on the LLG of the first cycle without masking by an envelope. It involves choosing an input scale factor which produces a ratio of ~1.2 between the $\langle U_{obs} \rangle$ and $\langle U_{calc} \rangle$ values on the first cycle. The behavoir of the LLG is somewhat unpredictable when using an envelope, as is that observed at low (<6.0 Å) resolution.

Use of existing phase probability distributions - the phased likelihood.

Experimental phase information, provided by the four Hendrickson-Lattman coefficients (Hendrickson & Lattman, 1972), can be incorporated into calculations of the LLG and the Sim-weights (Bricogne, 1991b). Use of this "phased likelihood" can be useful in two ways when MICE is used for map improvement. First, it endows the LLG with some additional discrimination when used either as a stopping criterion or when comparing competing permutation hypotheses by combining the experimental phase probabilities with the Sim-like probabilities associated with maximum entropy extrapolation. Second, the implicit recombination of phase probabilities can be useful in resolving ambiguities intrinsic in isomorphous phase probabilities, leading to improved figure of merit weighting of centroid maps.

Monitoring how MICE behaves

Several aspects of this process deserve special comment, either because they are counterintuitive or because they require special attention.

- **Overfitting and stopping criteria.** There are an equal number of observed data and adjustable parameters for fitting the exponential model. It follows that values of the $\{\zeta_i^{ME},\}$ can be found that bring $\{U_{calc}\}$ arbitrarily close to $\{U_{obs}\}$. If $|U_{obs}|$ is an experimental observation with a standard deviation, σ, one should discontinue fitting at the point where $\{|U_{obs}| - |U_{calc}|\}$ becomes small compared to σ to avoid fitting the noise in the data. That stopping point is provided, in principle, by a reduced χ^2 statistic related to $\{|U_{obs}| - |U_{calc}|\}/\sigma$. In practice, the reduced χ^2 is hard to know exactly, due to several uncertainties, including the true values for the observed σ's. A natural alternative to using the reduced χ^2 statistic as a stopping criterion is to stop at maximum log-likelihood gain.

- **The reduced χ^2 statistic and "p".** Because of uncertainties in the estimation of both measurement errors, σh, and the expected dispersion of structure factor amplitudes around their mean value, estimated by the Σ parameters (Bricogne & Gilmore, 1990), the relative contributions of these two variance parameters to the weighted residual used by MICE as a reduced χ^2 statistic involves an additional parameter, **p**:

$$C = \frac{1}{2} \sum_{h \in H_{acen}} \frac{|U_h - U_h^*|^2}{p\varepsilon_h\Sigma_a + \sigma^2_h} + \frac{1}{2} \sum_{h \in H_{cen}} \frac{|U_h - U_h^*|^2}{p\varepsilon_h\Sigma_c + \sigma^2_h}$$

This parameter must be adjusted empirically. A useful rule of thumb is to set **p** such that the starting reduced χ^2 value is at about 8.

- **Statistics other than the LLG.** At the beginning of the fitting process the initial probability distribution is nearly flat and the $\{U^{ME}\}$ are all quite small, relative to the target $\{U^{obs}\}$. As fitting proceeds, $\{U^{ME}\}$ values gradually grow in magnitude throughout the dataset, with the phases of basis set reflections constrained to their input values. Extrapolation develops gradually. This phenomenon leads to what can be counter-intuitive behavior of both the entropy and the crystallographic R-factor between $|U_{obs}|$ and $|U_{ME}|$. The entropy actually decreases, rather than increasing, as the parameters are fitted. The maximum entropy property arises because qME(**x**) represents the minimum entropy loss that must be incurred in order to reproduce the constraints.

In MICE the observed and calculated structure factor amplitudes are never scaled together, and the R-factor behaves quite differently from the conventional R-factor. Typically, it starts off at a value of 0.8 - 0.9. As the $\{U_{ME}\}$ grow in magnitude, this value decreases gradually for both basis set and extrapolated reflections. The R-factor is nevertheless a useful indicator when evaluated at maximum LLG. Depending on the size of the basis set, its final value ranges from 0.4 to 0.2 or better for basis set reflections. A rule of thumb has developed, both in Glasgow and in Chapel Hill, that for initial runs with limited basis sets, the final value for basis set reflections is often around 0.3 at maximum LLG. In cases where a large basis set with good experimental

phases is used, the R-factor can be reduced to an almost arbitrarily low value, and in the final stages of model-independent map improvement, Shibin Xiang in my laboratory has achieved values around 0.15, corresponding closely to that achieved when refining a complete atomic model (Xiang, unpublished).

The LLG itself is an *extensive* quantity that depends strongly on the number and identity of reflections over which it is calculated. For this reason it can be compared only between equivalent nodes, for which it is evaluated using the same reflections. In order to compare results using different numbers of basis set reflections, Xiang (unpublished) has introduced an *intensive* statistic analogous to the crystallographic R-factor. This "likelihood factor" evaluates the average absolute difference between the each contribution to the likelihood and that which would obtain if $|U_{obs}|$ were equal to $|U_{ME}|$, relative to the average contribution itself. This statistic starts out with a value around 1.0 and decreases to varying degrees, depending on the quality of the constraints. Empirically, a value at or below 0.7 indicates that the map should be readily interpretable over most of the unit cell.

Updating the constraints

During map improvement calculations we found it useful to update the both the phase and envelope constraints, as is the case with conventional solvent flattening (Wang, 1985; Podjarny, Bhat & Zwick, 1987). Once a problem has come within the radius of convergence, ie when the envelope is essentially correct, and the initial basis set phases have been improved by phase combination the map can often be refined further, using all or nearly all the reflections in the basis set and allowing the constraint phases to approach the maximum entropy phases on successive cycles of exponential modeling.

Noncrystallographic symmetry

There is no facility currently implemented in MICE for dealing with noncrystallographic symmetry averaging. However, the presence of local symmetries can be exploited, in principle, by alternating cycles of exponential modeling with averaging by an external set of programs.

4. Permutation designs and sampling

Phasing trees

The essence of the iterative Bayesian paradigm (Bricogne, 1988a; Bricogne, 1993) consists of two stages. The first is to define as accurately as possible the information content represented by the current basis set of phase and envelope choices. This is done by constructing qME(x), its transform, $\{U^{ME}_{h,k}\}$, and the Sim-like weights, and then assessing whether or not the centroid map is sufficient to solve the structure. The second is to identify discrete and manageable increments of missing (and/or incorrect) information most likely to improve the centroid map and search for useful indications of the correct choices for these new increments. Bricogne uses the term "phasing tree" to describe the process, and the term "node" to describe each distinct set of new phase and envelope choices. Each node requires a complete entropy maximization.

Each node represents a different hypothesis about the basis set phases and the solvent boundary. In the context of that hypothesis, entropy maximization plays two roles. First, it yields the numerical value of the smallest entropy loss which has to be accepted in order to provide an exponential model that fits the hypothesis attached to each node. That entropy loss is related to the *a priori* probability of the corresponding hypothesis, by virtue of the link with the saddlepoint method established in Bricogne (1984). Second, and most importantly, the interaction it creates between the basis and non-basis reflections through the phenomenon of maximum-entropy extrapolation for non-basis reflections $\mathbf{k}$ in $\{K\}$ depends on the phase and envelope information assumed in the hypothesis. It thus provides, via the likelihood statistic, a sensitive "score" by which competing hypotheses about phases and envelopes can be ranked.

At each new node, the procedure consists of (a) generating multiple hypotheses about the missing information so as to form a representative sample of all possible alternatives (*permutation*); (2) evaluating the degree of corroboration of each hypothesis by the observed data, as measured by its likelihood (*scoring*); (c) combining the *a priori* probability of each hypothesis with its likelihood, by means of Bayes's theorem, to obtain its *a posteriori*

probability; (d) making decisions, on the basis of these *a posteriori* probabilities, about which hypotheses should be rejected and which should be expanded further (*statistical inference*).

Identifying the "most influential" reflections

Strong unphased or incorrectly phased reflections which are inaccessible to maximum-entropy extrapolation from the phased ones are identified from a sorted list of the renormalized structure factor amplitudes. Generally, these occur roughly in order of increasing resolution, and this trend is worth verifying, as it is often preferable to permute phases for a lower resolution reflection with a smaller renormalized amplitude than for stronger reflections at higher resolution. Occasionally, we have found reflections for which the MIRAS probability distribution indicates an incorrect phase with a high figure of merit, and which consequently have a large renormalized structure factor (Doublié, Xiang, Gilmore, Bricogne & Carter, 1994). These phase errors can be corrected by permutation methods.

Permuting features of the molecular envelope

In the course of the successful application of this process to phase determination, we found that it was equally effective to permute hypotheses about possible modifications of the molecular envelope in the same way (Bricogne, 1988a, §2.3). In the structure determination of TrpRS permutation was carried out for six different binary choices regarding the calculation and description of the molecular envelope. The LLG score also provided statistically significant indications that ultimately proved to be correct in guiding the progressive editing of the unknown molecular envelope.

Multidimensionality and factorial design

The related phase and envelope definition problems are quintessential factorial situations. Simultaneous permutation of phases for different reflections by factorial design rapidly generates an unwieldy number of different combinations, and the total number of experiments necessary to exhaustively test all possibilities quickly becomes impossible. It is therefore essential to *sample* the complete factorial design as efficiently as possible, consistent with retaining the crucial information.

A major difference between applications aimed at map improvement and the *ab initio* problem is that in the former case there is usually sufficient phase information available to bypass the most serious "branching" problems (Bricogne, 1984; Bricogne, 1993). Branching arises whenever the LLG is more sensitive to combinations of phases than it is to the values of individual phases themselves, and is a severe problem in *ab initio* applications. The relative freedom from branching behavior in map improvement applications simplifies the task of sampling because it is less critical that the factorial designs preserve the higher-order interactions between reflections. This simplification also facilitates simultaneous permutation of more bits of phase information with a given number of nodes.

- **Covering and substantialization.** It is both unfeasible and unnecessary to examine all possible combinations in a factorial phase permutation design. In fact, with a given number of nodes it is preferable, within certain limits, to permute more reflections at the same time than to sample more finely a design with fewer reflections. By sampling larger volumes of phase space, the additional reflections in the design provide a larger "dynamic range" for the LLG . Bricogne has presented an elegant analysis of this sampling problem, emphasizing the importance of the covering radius (the maximum distance between sampling points and the remaining points of the complete factorial design) and substantialization (the maximum distance between the correct node in the complete factorial and the nearest sampling point) properties of sampling designs (Bricogne, 1993). Designs based on error correcting codes take advantage of the periodic nature of the phase angle to provide designs of extraordinary efficiency. Unfortunately, most of these involve between 256 and 4096 different nodes, an unworkable number, given current computation speeds. Incomplete factorial designs (Carter & Carter, 1979; Carter, 1990; Carter, 1992) were introduced to provide high sampling efficiencies for designs with any desired number of nodes at the expense of sacrificing interactions of higher order than two. Designs using 16, 24, 60, and 100 nodes have been generated and used in our own work to test for 7, 10-11, 12, and 14 bits of phase information. Examples of these designs and their use are given below and in (Doublié, Xiang, Gilmore, Bricogne & Carter, 1994). In these designs centric reflections are treated at two levels, plus and minus, and acentric reflections are tested at the quadrant permutations, 45, 135, 225, and 315 degrees.

• **Statistical analysis and significance testing.** It is important to note that the goal of a permutation experiment is not to determine which node is the best of the group, but rather to infer from the ensemble of results which choice is correct for each bit of phase information. As the illustration in section 1 suggests, the LLG achieves a maximum value for the correct constraint phase, other things being equal. However, in a factorial design the other permuted phases are not equal, and identification of the correct phase becomes a matter of statistical inference. This is done by analyzing the functional dependence of the LLG on the phase choices. For magic lattice designs based on coding theory, this can be done in a very powerful way using multidimensional Fourier analysis (Bricogne, 1993). For incomplete factorial designs we use multiple regression least squares analysis of linear models of the LLG as functions of the individual phase bits (Wilkinson, Hill & Vang, 1992). Each acentric reflection is given two degrees of freedom, corresponding to its real and imaginary components. An essential source of the power of factorial designs is that for each binary choice they provide a comparison between averages of one half of the experiments with the other half. This effect can be seen graphically from the histograms in Figure 9. LLG scores from all 24 experiments are grouped four different ways, each according to the treatment of a particular phase bit. The average improvement in LLG score for one choice versus the other is evaluated and compared to the residual error by a t-test to determine its statistical significance (Table 1). Acentric phases can be estimated as the inverse tangent of the real and imaginary coefficients.

Table 1. Statistics for the Regression Model for a Phase Permutation Experiment

Calculated values for the dependent variable, the LLG, were obtained from a linear model in the ten phase bit choices for two acentric and six centric reflections permuted simultaneously according to an incomplete factorial design with 24 nodes. The variables are indicated in column 1, where I refers to the imaginary component of reflection [5 3 8] and S refers to the sign of the five centric reflections. Various models were considered by stepwise multiple regression using the program SYSTAT (Wilkinson, 1992). The best model, from which phases were determined, was taken to be that with the highest squared multiple correlation coefficient and the lowest F-ratio p-value from the analysis of variance, and is shown below. The coefficients of this model are shown in column 2 and their standard errors in column 3. Student's t value is shown for each coefficient in column 4 and its probability under the null hypothesis in column 5. The constant value is equal to the mean value of the LLG for the 24 nodes. The squared multiple correlation coefficient final model was 98%, attributing all but 2% of the variation in LLG score among the nodes to the phase choices made for each node in the respective basis sets. The absence of any indication for the real component of the [5 3 8] implies that its phase is close to 90°. No significant indications were obtained or for either of two additional acentric reflections, which were also permuted together with those shown here.

REGRESSION STATISTICS

VARIABLE	COEFFICIENT	STD ERROR	t	P(2 TAIL)	Phase
CONSTANT	3634.958	2.674	.14E+04	.10E-14	
I538	17.484	3.176	5.505	.39E-04	90
S600	-29.229	2.782	-10.505	.75E-08	180
S3015	-29.862	2.887	-10.342	.94E-08	315
S3021	-13.833	2.777	-4.981	.11E-03	225
S606	-28.208	3.331	-8.469	.17E-06	180
S3318	-36.148	2.717	-13.303	.20E-09	180

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	115687.113	6	19281.185	112.336	.100397E-11
RESIDUAL	2917.845	17	171.638		

Figure 9

Phase Determination by Factorial Design

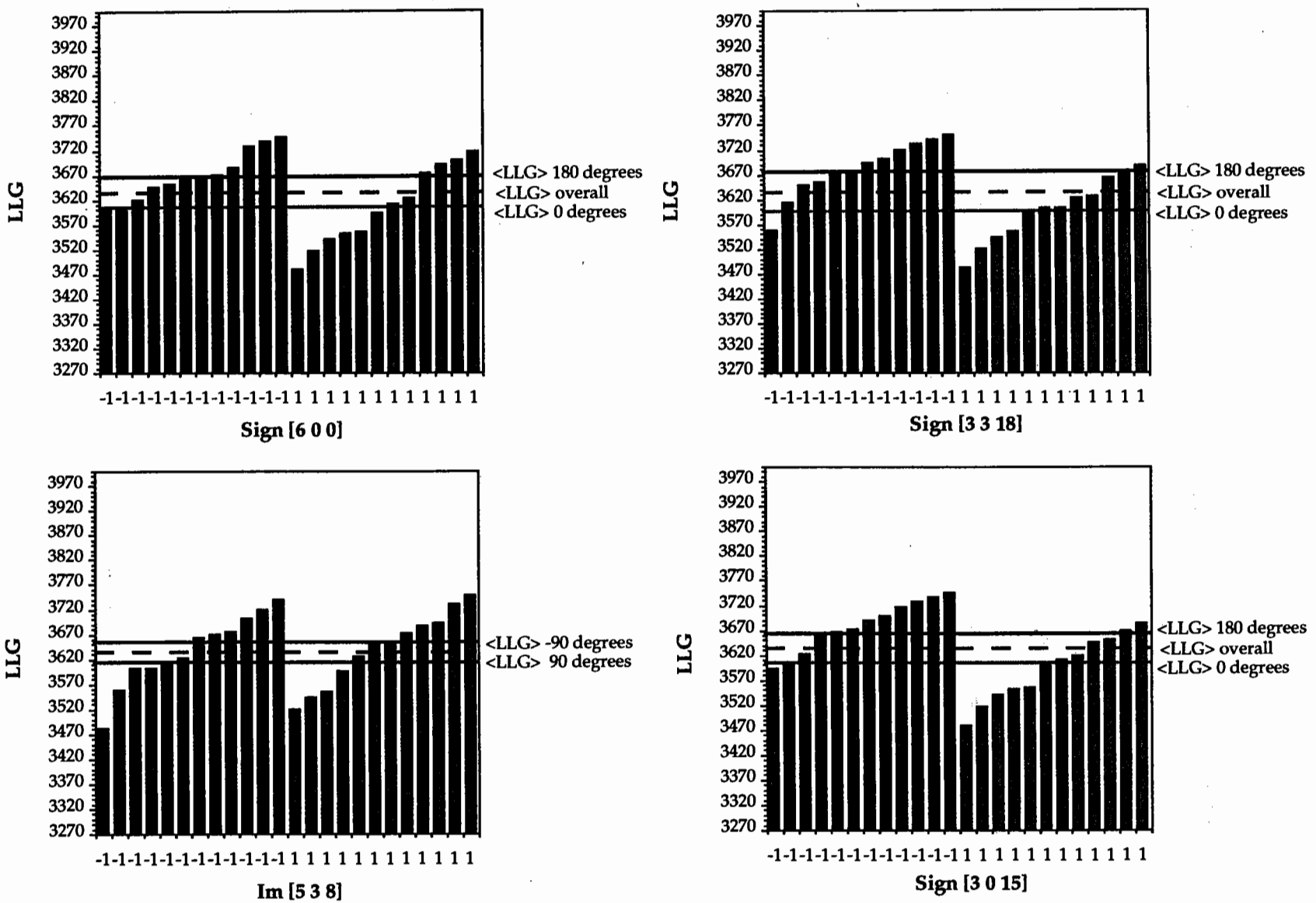


Figure 9. Factorial design uses averaging to enhance the power of likelihood scoring. The average difference between groups of reflections treated with the plus and minus signs for a given reflection is often a significant indicator of the correct phase choice.

5. Applications.

We have used MICE in three different contexts.

Map refinement

As a density modification algorithm (Xiang, Carter, Bricogne & Gilmore, 1993) maximum entropy solvent flattening is superior to conventional solvent flattening (Wang, 1985). Density in the centroid map is modified minimally from that specified by the experimental basis set phases, as expected from the heuristic "maximally non-committal" description of the maximum entropy method. Nonetheless, the centroid electron density map is nearly indistinguishable from the $\{2F_{obs} - F_{calc}\}$ map based on a refined model. If the envelope is correct and if the experimental phases are good enought to lie within a radius of convergence, (which we do not yet know how to define before the fact), then essentially the entire data set can be used as the basis set. Weighting the constraint reflections by their figure of merit improves the radius of convergence. Moreover, under these circumstances the maximum entropy phases for the basis set reflections are closer to the correct phases than are the experimental phases. Map refinement can therefore proceed via phase shifts in the basis set. Convergence of this process can bring the unscaled crystallographic R-factor as low as 0.15 for all reflections, rivalling that achieved by refining atomic coordinates (Xiang, unpublished).

Map refinement by MESF is preferable in two respects to refining model coordinates as a way to improve electron density maps. First, since the crystallographic R-factor is a function of the amplitudes only, coordinate refinement via its derivatives tends to shift the atomic positions in ways that preserve the current phases, whereas MESF constrained by a large basis set actually adjusts the phases to fit the measured amplitudes. Second, map improvement is achieved before investing in a possibly incorrect interpretation that would bias all subsequent refinement. MESF should thus be of general interest for most routine problems.

Supplementing the basis set with directly determined phases

Although very seductive, uncritical applications of MESF can be derailed by serious errors in the phases. The real space bias of an incorrect model has a reciprocal space counterpart in the biases due to incorrect phases. We have encountered two different examples. In solving the TrpRS structure (Doublié, Xiang, Gilmore, Bricogne & Carter, 1994) we had to deal simultaneously with (1) a serious lack of isomorphism in the heavy-atom derivatives, resulting in large starting phase errors, and (2) an initially poorly known molecular envelope. Because the constraints - both phases and envelope - were insufficiently well-determined at the outset, maximum-entropy solvent flattening as previously applied, though helpful, was insufficient to solve the structure. The strongly extrapolated phases tended to reinforce errors in the constraints, and efforts to extend the basis set by including them lead quickly to a catastrophic decrease in the LLG. This deadlock was broken by subordinating the MESF process to a permutation process designed to simultaneously sample the space of possible electron-density and envelope modifications represented by a number of critical reflections with unavailable or unreliable MIRAS phases.

- **A modest number of new phases can bring a significant improvement.** It should be emphasized that because we were able to identify and permute hypotheses about the most critical missing information, the Bayesian paradigm did not require unrealistic amounts of computing time by current standards. We were lead directly to a correct definition of the molecular envelope after a very modest number of reflections had been phased. A total of 25 centric and 7 acentric reflections (39 bits of phase information), mostly at low resolution, were phased directly by evaluating 94 nodes, each taking roughly 20 hours on a DECstation 5000/200. One of these was actually a basis set reflection whose MIRAS phase was incorrect, despite having a high figure of merit.

The most effective strategy involved permuting the largest practical number of bits of phase information simultaneously. For 24 node designs, this meant permuting 10-11 phase bits at a

time. Use of the recursive, tree-directed algorithm for choosing reflections to permute allowed newly recruited phases to have an appropriate impact on the selection of new reflections that had become critical as a result of recruiting them into the basis set. At one point, it became clear that redefinition of the envelope could be handled more effectively by permuting features of the envelope directly, rather than depending on new phases to do the job. Three ambiguities in the envelope were resolved by envelope permutation. Using the corrected envelope we were able with MESF to improve the native phases sufficiently to identify positions for nine of ten selenium atoms in a difference Fourier map for selenomethionine-substituted TrpRS, providing the necessary *isomorphous* phases to solve the structure. Seven additional reflections were phased by permutation and likelihood scoring after the selenomethionine derivative had been incorporated. The resulting electron density map was nearly indistinquishable from the final $\{|2F_{obs} - F_{calc}|, \phi_{calc}\}$ map from the refined structure. This work was the first successful application of maximum entropy phasing methods (Bricogne, 1988a; Bricogne, 1988b; Bricogne, 1991b; Bricogne, 1993) to the solution of an unknown macromolecular crystal structure for which previously available methods were insufficiently powerful.

• **Sampling and statistical inference.** Multiple regression least squares analysis of the LLG approximated as a linear function of the hypotheses produced significantly better results than could be obtained by chosing the "best" node from the design. The incomplete factorial designs test only about $1/(2\sqrt{N})$ of the possible permutations, so it is very unlikely that any of the nodes tested actually represent correct choices for all the permuted quantities. Extracting statistically significant phase and/or envelope choices by procedures illustrated in Table 1 and Figure 9 exploits the power of averaging, for example, 12 nodes for each choice to infer which of the *unsampled* points actually represents the correct node. With TrpRS, Student's t-tests provided significant, and useful indications for 92% of the permuted reflections and for all six hypotheses regarding the molecular envelope (Doublié, Xiang, Gilmore, Bricogne & Carter, 1994). Where it has been possible to compare phases determined this way to those from a refined model, centric phases were correct, and permuted acentric reflections were phased with a mean phase error of 26°.

• **An example using molecular replacement phases.** More recently, we encountered an unexpected difficulty in solving a very closely related crystal form of TrpRS by molecular replacement. This work is still unfinished but a preliminary account is instructive. There are two different tetragonal crystal forms of TrprS. Both are grown in potassium phosphate but one is obtained from the other by stabilizastion in ammonium sulfate, which causes a 6% increase along the long c axis and a compensating 3% decrease along a and b. This change increased the diffraction limits from 2.8 Å to 1.7 Å and was associated with a chemical transformation of the ligand (Carter, Doublié & Coleman, 1994). For reasons related to the biochemistry we wanted to solve the form stabilized in the native phosphate mother liquor. The molecular replacement solution was unusually difficult and unconvincing; a fragment containing only 65% of the structure having a cleaner solution with rigid-body refinement statistics almost indistinguishable from those obtained for the intact molecule. Molecular replacement searches did not detect the missing 120-residue fragment, however, so the oriented/translated monomer was therefore refined by simulated annealing (Brunger, 1992b.) The resulting R-factor was 20%, and the model fitted nicely to the 2Fo-Fc density essentially everywhere.

We were reluctant to accept this solution because there was no density for a ligand, which radiolabeling studies had previously demonstrated to be present in the crystals, and which was different from the ligand present in the known structure. Moreover, when we used the best third of the refined Fc phases as constraints, together with an envelope generated from the refined model the R-factor at maximum LLG was an unusually high 0.42, compared to 0.32 for a control experiment with the known structure. More disturbing was the fact that about half of the 30 largest renormalized structure factor amplitudes were from the basis set, a result we had never seen before. We therefore tried permuting the phases of basis set reflections from the list of large renormalized structure factor amplitudes. The results in Table 1 and Figure 9 were taken from this work (Yin, unpublished). Using the phases from this permutation, we could visualize some of the density expected in the active site. The extent of error in the refined structure is not currently known, but it is significant that phases from the unrefined molecular

replacement model could be fitted to a higher LLG and a lower R-factor, and left many fewer basis set reflections among the list of the strongest renormalized structure factors.

Summary

Entropy maximization, hypothesis permutation and likelihood scoring represent new tools that can have considerable impact on the technology for solving macromolecular crystal structures. By itself, entropy maximization (as MESF) is a powerful density modification technique. However, other approaches to density modification also offer considerable map improvement in some cases (SQUASH: Zhang & Main, 1990; Cowtan, 1991; Cowtan & Main, 1993; Zhang, 1993; Skeletonization, PRISM: Wilson & Agard, 1993). It is therefore worth emphasizing that the Bayesian paradigm is the only method that involves a proper implementation of conditional probability distributions, and is therefore the only method capable of hypothesis permutation and significance testing. The capacity to convert hypotheses about phases into statistically significant scores represents an entirely new and powerful tool for macromolecular structure determination.

Acknowledgments. The MICE program was developed and written by Chris Gilmore and his associates in the Chemistry department of Glasgow University together with Gérard Bricogne. I am grateful to both of these collaborators for their continual advice and encouragment.

REFERENCES

Bricogne, G. Acta Cryst, *A40* (1984) 410-445.

Bricogne, G. Acta Cryst, *A44* (1988a) 517-545.

Bricogne, G. in *Crystallographic Computing 4*, edited by N. W. Isaacs & M. R. Taylor Oxford, UK: IUCR and Oxford University Press (1988b) 60-79

Bricogne, G. in *Direct Methods of Solving Crystal Structures*, edited by H. Schenk New York: Plenum Press (1991a)

Bricogne, G. . in *Crystallographic Computing V*, edited by D. Moras, A. D. Podjarny & J. C. Thierry Oxford: Clarendon Press (1991b) 257-297

Bricogne, G. . in *Molecular Replacement*, edited by E. J. Dodson, S. Gover & W. Wolf Daresbury, Warrington, UK: Science and Engineering Research Council, Daresbury Laboratory (1992) 62-75

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

Bricogne, G. & Gilmore, C. J. Acta Cryst, *A46* (1990) 248-297.

Brunger, A. Nature, *355* (1992a) 472-474.

Brunger, A. T. (1992b). X-PLOR. Yale University.

Brunger, A. T. Acta Cryst, *D49* (1993) 24-36.

Carter, C. W., Jr. METHODS: A Companion to Methods in Enzymology, (1990) 12-24.

Carter, C. W., Jr. in *Crystallization of Proteins and Nucleic Acids: A Practical Approach*, edited by A. Ducruix & R. Giegé Oxford: IRL Press (1992) 47-71

Carter, C. W., Jr. & Carter, C. W. J. Biol. Chem, *254* (1979) 12219-12223.

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

Carter, C. W., Jr., Doublié, S. & Coleman, D. E. J. Mol. Biol, *In press* (1994) .

Cowtan, K. in *Direct Methods of Solving Crystal Structures*, edited by H. Schenk New York: Plenum Press (1991) 421-424

Cowtan, K. & Main, P. Acta Cryst, *D49* (1993) 148-157.

Daresbury Laboratory (1990). CCP4. Daresbury Laboratory.

Doublié, S. Ph.D., University of North Carolina at Chapel Hill (1993)

Doublié, S., Xiang, S., Gilmore, C. J., Bricogne, G. & Carter, C. W. J. Acta Cryst, *A50* (1994) 164-182.

Edwards, A. W. F. Likelihood, Cambridge: Cambridge University Press (1972)

Germain, G., Main, P. & Woolfson, M. M. Acta Cryst, *B26* (1970) 274-285.

Gilmore, C. J. J. Appl. Cryst, *17* (1984) 42-46.

Gilmore, C. J. & Brown, S. R. J. Appl. Cryst, *22* (1988) 571-572.

Gilmore, C. J., Henderson, A. N. & Bricogne, G. Acta Cryst, *A47* (1991) 842-846.

Hendrickson, W. A. & Lattman, E. E. Acta Cryst., *B26* (1972) 136.

Leslie, A. in *Improving Protein Phases*, edited by S. Bailey, E. Dodson & S. Phillips Warrington, UK: SERC Daresbury Laboratory (1988)

Minor, W. MAP_CCP4: a Program for Editing Molecular Envelopes Purdue University (1992)

Podjarny, A. D., Bhat, T. N. & Zwick, M. Ann. Rev. Biophys. Biophys. Chem, *16* (1987) 351-373.

Wang, B. C. Meth. Enzymol, *115* (1985) 90-112.

Wilkinson, L., Hill, M. & Vang, E. (1992). Systat: Statistics, version 5.2 Edition. SYSTAT, Inc.

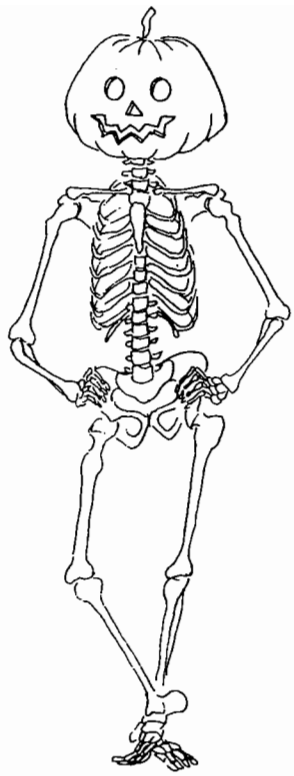
Wilson, C. & Agard, D. A. Acta Cryst, *A49* (1993) 97-104.

Xiang, S., Carter, C. W., Jr., Bricogne, G. & Gilmore, C. J. Acta Cryst, *D49* (1993) 193-212.

Zhang, K. Acta Cryst, *D49* (1993) 213-222.

Zhang, K. Y. J. & Main, P. Acta Cryst, *A46* (1990) 377-381.

Halloween ... Masks and Bones



Gerard J. Kleywegt & T. Alwyn Jones,
Department of Molecular Biology,
Biomedical Centre,
Uppsala University,
Box 590,
S-751 24 Uppsala,
SWEDEN.

Introduction.

Two years ago, one of us (TAJ) reported on a set of programs called "A" for use in single-crystal, non-crystallographic symmetry (NCS) electron-density averaging [1]. Here, we wish to report some recent extensions to and improvements of this software package (now yclept "RAVE" [2]). Subsequently, we shall discuss molecular envelopes, or masks, in some more detail. Finally, we will outline a new method of "recycling" existing protein structures in the process of building an initial model in an MIR map.

RAVE.

RAVE is a set of density-modification programs which can be used for solvent flattening and for real-space single- and multiple-crystal, single and multiple domain (NCS) electron-density averaging [2] (iterative skeletonisation has also been implemented, but hitherto the results obtained with it have been rather underwhelming). The program suite is an extension of the previous A package [1]; some programs have been modified, extended and/or improved, and some new tools have been added. RAVE uses standard CCP4 programs [3] for structure factor calculations, scaling, phase combination, map calculation *etc.*, and produces masks and maps which can be displayed with O [4]. For a discussion of some of these issues, see [1]; for references to the original literature about averaging, see [1] and [2]. At present, RAVE comprises the following programs:

° **NCS6D:** a program to obtain NCS-operators through a brute-force, six-dimensional rotation/translation search. The program attempts to find the operator which gives the highest correlation coefficient between the density at a set of atomic positions (either genuine or "bones" atoms), and the density at the same positions after application of the operator.

° **IMP:** a program to improve approximate NCS operators. IMP tries to maximise the correlation coefficient between density inside a mask and density within that mask after application of the operator. Usually, the program is able to fully automatically improve imprecise operators significantly.

° **AVE:** a program for averaging and/or expanding density using a mask and a set of NCS-operators.

° **MAVE:** combines the functionality of IMP and AVE for the case of multiple-crystal (NCS) averaging: it can be used to average density, to expand it back into the (asymmetric units of the) unit cells, and to improve inter-crystal operators.

° **MAMA:** a program for generating, manipulating, improving and combining masks (molecular envelopes).

° **DATAMAN:** a program for manipulating reflection data. It is needed to apply Wilson scaling to the various datasets involved in multiple-crystal (NCS) averaging.

° **COMAP:** a program to combine the averaged maps in multiple-crystal (NCS) averaging, prior to expansion into the individual unit cells.

° **COMDEM:** a program to combine multiple-domain averaged and expanded maps in order to calculate and set the background density level.

° **MAPMAN:** a map-manipulation program. It combines the functionality of several older, stand-alone programs (mappage, bones, swapbytes, format conversion) and adds other functions (peak-picking, map normalisation, map combination, map statistics, *etc.*). It is used to convert CCP4 maps into a format suitable for O, and to convert maps into masks or the other way around.

RAVE is, in our opinion, easy to use, non-arcane and conceptually simple: there are no skewing operators; except for spacegroup symmetry, all operators are in Cartesian space; there is no limit on the number of crystal forms that can be averaged; there is no dummy P1 cell in multiple-crystal averaging; the programs are spacegroup-general; there is always only one mask (even in multiple-crystal averaging, though not in multiple-domain averaging, where there is one mask per domain); proper and improper symmetry are usually treated in the same way (namely, as improper symmetry); generation, manipulation and improvement of masks is simple. The fact that RAVE is interfaced with O and CCP4 is another forte. There are example Unix C-shell scripts available for running the programs, as well as full documentation for all programs and a complete worked example *cum* tutorial (including all files needed to re-work the example). In its first year, RAVE has already been successfully used for tracing and/or rebuilding a few dozen structures (including one based on electron microscopy data), the results of which are expected to start make their way into the literature this year.

Masks.

When one is averaging density, as opposed to flattening solvent, it is important to have a high-quality mask (molecular envelope). A mask [1] is a "logical print" of the molecular envelope on a grid where a grid point is set to "1" ("true") if it is part of the molecular envelope, and to "0" ("false") if it is not. Masks are stored as ASCII files which contain information about the unit cell and the grid, plus the information needed to reconstruct the mask inside the grid. Masks can be generated, edited and manipulated in myriad ways, either with O [4] or with MAMA. Using these programs, one can fairly easily and rapidly generate a high-quality mask, *i.e.* a mask which satisfies the following criteria:

° no cavities inside the mask (unless they are used to mask out heavy-atom sites);

° no "blobs" of mask points which are not connected to the bulk of the mask;

° most or all atoms covered by the mask;

° no sharp extrusions or invaginations of the mask surface (it should be smooth and "curvaceous");

° some room between the surface of the mask and the borders of its grid;

° as little overlap as possible with (non-crystallographic) symmetry-related copies of itself.

Mask generation.

Depending on the stage of the structure determination, there is a plethora of methods available for generating an initial mask:

°**from scratch**: one may generate an empty mask with MAMA and subsequently "fill it in" inside O, for example, as one goes along editing the skeletonised density (bones).

°**from one or more points**: MAMA contains options to generate spherical or cubic masks around a user-defined point in space. Optionally, two or more of such masks can be combined (*vide infra*).

°**from bones atoms**: one may feed a skeleton into MAMA, associating a certain radius with each bones atom, and use it to generate a mask around it.

°**from atoms**: once one has obtained a (partial) model, one can use MAMA to generate a mask around a set of atoms (in PDB format) by associating a radius with each of them. If one wants to use "real" van der Waals radii, one may use our cavity program VOIDOO [5] to generate the mask.

°**from other masks**: one may combine existing masks in several ways (*vide infra*).

°**from an old mask**: if one solves a mutant structure, and still has the old mask of the wild-type structure, it can easily be transformed with MAMA to cover the new structure, even if the grids and spacegroups are different.

°**from a map**: MAPMAN contains an option to convert a map into a mask, by setting all points to "1" for which the density exceeds a certain threshold value, and to "0" otherwise.

Different masks can be combined in various ways with MAMA. First, there is a UNITE option, which creates a new mask that is the unison of two other masks. Second, there are a number of logical operations which can be applied to masks: NOT, AND, OR, BUTNOT (*i.e.*, "mask_1 AND (NOT mask_2)") and XOR (exclusive OR). In this way, one may combine two monomer masks into a dimer mask, account for atoms which moved during refinement (without having to re-generate a mask), *etc.*

As one uses data to higher resolution, or if one wants to remove overlap from a mask in multiple crystal forms, one needs to be able to transform a mask onto a new grid and even into a different unit cell and/or spacegroup. MAMA contains options to do all this. In addition, the program will make sure that the volume of the mask before and after the transformation is virtually identical (usually, to within 0.5 %).

Mask improvement.

Previously, mask improvement had to be carried out exclusively with the mask-editing commands in O. With the advent of MAMA, much mask-editing time can be saved. For example, MAMA contains an option (FILL_VOIDS) to automatically fill any cavities that may exist inside a crude initial mask. Another option (ISLAND_ERASE) automatically removes

"droplets" of mask points which are unconnected to the bulk of the mask. One may also use the ATOM_FIT option to check if all atoms are covered by the mask, given a certain radius around each atom. In order to prevent problems with the density interpolation during map expansion, it is important that there be some space between the surface of the mask and the borders of its grid. MAMA contains options both to check if such problems exist, and to remedy them.

To get rid of sharp extrusions and invaginations of the mask surface, MAMA contains options which either slice some points off the surface, or add some points to it. The SMOOTH option adds all points to the mask which are not currently in it and which have at least a user-defined number of neighbour points which are in the mask (this removes small invaginations); the EXPAND option does the same by adding all non-mask points to the mask which have at least one neighbour which is in the mask (this effectively adds one layer of points to the mask). Similarly, the CUT option will remove all points from the mask which have at least a user-defined number of neighbours outside the mask (this removes small extrusions from the mask); the CONTRACT option does the same, using a threshold of one neighbour (this effectively removes one layer of mask points, much like a cheese slicer does).

MAMA can also be used to investigate and remove mask overlap. This is done by expanding the mask into the asymmetric unit of the unit cell and keeping track of either *which* or *how many* NCS-operators project a mask point onto a certain point in the asymmetric unit. By "averaging" back onto the mask, one obtains an "overlap map" which contains information about the degree of overlap due to each mask point. Subsequently, all overlap-inducing points may be removed from the mask. However, this operation is bound to remove too much from the mask. Therefore, one may "trim" the mask instead. This entails slicing bits from the surface of the mask in areas give rise to overlap. Repeating this operation a few times usually removes most of the overlap.

Nowadays, we only use O to assess the overall shape of the mask, and to edit in or out small areas around certain atoms. This can be done conveniently with the help of two small O macros, one to "mask" an atom and one to "unmask" an atom:

```
! mask_atom.omac - MASK a small area around an atom
message 'Click on atom to MASK'
! switch all points within 3 Å "on"
mask_set 1 1 3.0 ; ; ;
mask_on wait_id no
mask_contour on_off
bell message done

! unmask_atom.omac - UNMASK a small area around an atom
message 'Click on atom to UNMASK ...'
! switch all points within 1 Å "off"
mask_set 0 1 1.0 ; ; ;
mask_on wait_id no
mask_contour on_off
bell message done
```

Spin-offs.

After MAMA had been written, it turned out that the program had two unintended, but useful, additional capabilities:

° It can be used to calculate the "shape similarity" of two molecules. First, one needs to align the two molecules (*e.g.*, using least-squares superpositioning) and to calculate a mask around each (using identical grids *etc.*). The SIMILARITY option in MAMA will then calculate the shape-similarity index (SI), which is defined as: $SI = N_{1\&2} / SQRT (N_1 * N_2)$, where N_1 and N_2 are the number of points in the mask of molecule 1 and 2, respectively, and $N_{1\&2}$ is the number of points their masks have in common.

° It can be used to visualise (and measure) tunnels, clefts and other cavities which are connected to "the outside world" and which can, therefore, not be handled by our cavity program VOIDOO [5]. It turns out that appropriate combinations of options in MAMA (EXPAND, CONTRACT, AND, BUTNOT, *etc.*) elegantly enable emulation of Delaney's cavity-detection algorithm [6].

Molecular theft.

"Recycling" old protein structures is an established procedure in crystallography. Of course, Molecular Replacement springs to mind immediately as the technique which relies on the availability of a presumably similar structure to solve another. Also, during model building and rebuilding, anyone who uses the LEGO commands in O [4] is recycling parts of previously solved, high-resolution protein structures. This begs the question: "Can we use parts of solved structures already while tracing the map in MIR and SIR studies ?" Recently, we have begun to explore the possibility of recycling (motifs of) secondary structure elements (SSEs), in effect creating a "synthesis" of isomorphous and molecular replacement, and some of the results are encouraging. The procedure requires skeletonised density (bones) and a -soon to be published- program called DEJAVU [7].

DEJAVU is a program we wrote during 1992 and 1993; it can be used to recognise structural similarities between protein structures (which becomes all the more important as the number of solved structures increases). It can be used in two different "modes": (a) looking for proteins which contain a particular, user-defined structural motif that occurs in one's structure, or (b) looking for proteins which have "many" SSEs in similar orientations as (some of) those in the user's protein. DEJAVU compares the SSEs in the user's protein to those listed in a database (derived from the PDB) that contains the SSEs that occur in ~1,600 (X-ray) protein structures. Each SSE is characterised by its type (ALPHA or BETA), the number of residues it contains and the Cartesian coordinates of the Cα atoms of the first and the last residue. The latter are used to calculate the length of each SSE as well as its direction vector.

The program uses a constrained, recursive depth-first combinatorial search algorithm [8] to match SSEs of the user's protein with those of database proteins. The constraints are used to prune unpromising branches of the search tree as early as possible. Constraints are imposed for each SSE (type, number of residues, length) and for the assembly of SSEs (mutual distances, cosines of the angles between their direction vectors). Optionally, a neighbour-connectivity constraint and a sequence-directionality constraint may be imposed.

Using DEJAVU, one may be able to "steal" (or, at least, detect) similar motifs of SSEs from

other structures. The whole idea is based on the observation that, while one is editing the skeleton in the MIR map, one often begins to discern strands and/or helices. There are, however, two problems: (a) connecting loops or wholes stretches of the structure may be "invisible" in the map, and (b) at the resolution typical of initial MIR maps (~3 - 3.5 Å), one will usually not be able to tell which end of the SSE is N-terminal and which end is C-terminal. In other words: one knows little or nothing about the direction and connectivity of the SSEs. The solution is simply to use DEJAVU to do the hard work, but with the neighbour-connectivity and sequence-directionality constraints switched off.

Modus operandi.

The first thing one needs to do is to delineate as many SSEs as possible (and as accurately as possible). Until appropriate commands have been implemented in O, this means finding the Cartesian coordinates of two points which one judges to be the termini of the SSE, and "guestimating" the number of residues in the SSE (this may be very rough). These need to be put into a small ASCII file, which may look as follows:

```
MOL     bone
NOTE    manually generated from bones of P2 myelin
PDB     /nfs/taj/gerard/progs/secs/test/bone.pdb
! approx. strand 120 -> 112
BETA    'B1'  '1'  '9' 9 62.43 61.51 44.14  43.15 51.56 30.90
! approx. strand 108 -> 100
BETA    'B2' '11' '19' 9 47.74 50.65 28.25  63.35 70.56 38.84
! ... et cetera ...
```

Each line contains the SSE type (ALPHA or BETA), the "name" of the SSE (B1, *etc.*), the (dummy) names of the first and last residue, the estimated number of residues and the Cartesian coordinates of the Cα atoms of the first and last residue.

If one runs DEJAVU, and tells the program that one does a "bones search", the program will suggest suitable values for all its parameters. The output is a list of "hits", *i.e.* proteins which contain similar SSEs in similar orientations, and an O macro. The latter is the most interesting. It contains a set of O instructions which, for each "hit", will:

° print some information about the hit;

° read the PDB file of the protein;

° create an O datablock containing the best rotation-translation operator found by DEJAVU to align the SSEs of the hit with those identified in the skeleton;

° apply this operator to the protein;

° draw a Cα-trace of the protein after the coordinate transformation, in which matched SSEs are coloured red and the rest of the structure is coloured blue.

Results.

We have tested the method using a skeleton of P2 myelin protein [9, 10], a lipid-binding protein containing 131 residues. We manually delineated seven SSEs, six strands and one helix, and used this as input to DEJAVU. The program (after consuming less than six seconds of CPU time on a DEC Alpha/OSF1 to investigate all ~1,600 database proteins !) comes up

with seven hits (PDB codes: 1IFB, 1ALB, 2HMB, 1IFC, 1MDC, 1OPA and 1OPB; all of these are lipid-binding proteins). All hits are reasonable: most of them are "off-spring" of P2 myelin protein (*i.e.*, lipid-binding proteins solved by Molecular Replacement using P2 -or a structure which in turn was solved using P2- as the search model), and all of them are also found by DEJAVU when the program is fed the actual SSEs and the actual coordinates, and, even more important, the operators are very similar in both cases. In the case of P2, the resulting hits and their alignments are of such impressive quality, that one would immediately abandon the tracing of the MIR map, and use one of the hits (after appropriate mutations and rebuilding) as the initial model for refinement ! (Of course, this wouldn't have worked at the time when P2 was solved, since none of its off-spring was in the PDB yet !)

A second test was carried out using the MIR map of *Candida antarctica* lipase B [11]. In this case, ten SSEs were delineated in the skeleton (six strands and four helices). Unfortunately (but realistically), there are no proteins in the database which are as similar to this lipase as the previous seven lipid-binding proteins are to P2 myelin protein. DEJAVU now comes up with widely different numbers of hits, depending critically on the choice of parameters. It's a promising fact that both acetylcholinesterase (1ACE) and another lipase (1THG) are among the hits. When using full sets of SSEs and actual coordinates, 125 residues of 1ACE can be aligned with lipase B (RMS distance on Cα atoms 1.87 Å), and 121 residues of 1THG (1.91 Å). In both cases, DEJAVU matches seven of the ten SSEs, with an RMS distance of the centroids of corresponding SSEs of ~3.6 Å. Unfortunately, the operators obtained in a conventional search and those obtained in this bones search are completely different ... One can find combinations of parameter values which result in the correct operator for 1ACE, but in that case 1THG disappears and two dozen new hits come up. Another problem is the occurrence of "false hits" and "poor hits"; for instance, leucine aminopeptidase (1BPN) shows up as a more promising hit than 1ACE and 1THG (seven SSEs out of ten aligned with an RMS centroid distance of ~2.8 Å), even though only 40 residues can be aligned with an RMS Cα-distance of ~2.3 Å if one uses the full structures. The major reason why the results are so much worse for lipase B is probably the fact that there are no proteins in the database which are very similar (in the case of P2, there were several structures to which 90 to 100 % of the P2 residues could be aligned with an RMS distance on Cα atoms of ~0.7 to 1.5 Å; in the case of lipase B, there are two structures to which only ~40 % of the residues can be aligned with an RMS distance on Cα atoms of ~1.9 Å). Other possible causes are inaccurate delineation of SSEs in the skeleton (if, say, only the first half of an SSE is visible, then the *apparent* centroid position will differ significantly from the actual one) and the fact that perhaps too few SSEs were delineated (only 10 out of 27 possible; in the case of P2, 7 out of 13 SSEs were used).

In conclusion: the possibility of using the skeleton to scan a database for proteins containing a similar subset of SSEs in a similar spatial arrangement exists, but more work is needed to investigate how to use the method optimally. In addition, some new commands will have to be implemented in O to make the process of delineating the SSEs in the skeleton easier. Finally, we'll have to think about what to do with the results. For instance, one could simply use the best hit as a guide for tracing the chain (this is the most conservative alternative, which requires no new O commands). Alternatively, one could envision replacing bones atoms in a matched

SSE by new atoms, derived from the matched SSE of the hit protein (perhaps combined with rigid-body optimisation of the fit between the fragment and the MIR map).

Availability of the software.

RAVE, DEJAVU and VOIDOO have been implemented on Evans & Sutherland, Silicon Graphics and DEC Alpha/OSF1 workstations. Additional implementations of RAVE are maintained by other people (Alliant, Convex *etc.*). Academic O users who have signed a licence agreement for O, may freely download the software, the manuals and the RAVE tutorial from the Uppsala ftp-server. Others may contact TAJ for further information (E-mail: "alwyn@xray.bmc.uu.se").

Acknowledgment.

This work was supported by the Swedish Natural Science Research Council and Uppsala University. During part of the implementation of VOIDOO and DEJAVU, GJK was supported by a post-doctoral research fellowship from the Netherlands Organisation for Scientific Research (NWO). We are grateful to Christina Divne for drawing the cartoon on the title page of this paper. We would like to thank Prof. Gérard Bricogne for contributing subroutines for Wilson scaling of datasets from different crystal forms. All complaints, suggestions and bug reports from RAVE users the world over are gratefully acknowledged.

References.

[1] Jones, T.A. In *"Molecular Replacement"* (E.J. Dodson, S. Glover & W. Wolf, Eds.), SERC Daresbury Laboratory (1992) 91.

[2] Kleywegt, G.J. and Jones, T.A. *"Convenient single and multiple-crystal real-space averaging of macromolecular electron-density maps"*, to be published (1994).

[3] SERC Daresbury Laboratory. *"CCP4. A Suite of Programs for Protein Crystallography"*, SERC Daresbury Laboratory, Warrington, England (1986).

[4] Jones, T.A., Zou, J.Y., Cowan, S.W. and Kjeldgaard, M. Acta Cryst., *A47* (1991) 110.

[5] Kleywegt, G.J. and Jones, T.A. *"Detection, delineation, measurement and display of cavities in macromolecular structures"*, Acta Cryst., *D50* (1994) in the press.

[6] Delaney, J.S. J. Mol. Graphics, *10* (1992) 174.

[7] Kleywegt, G.J. and Jones, T.A. *"Detecting similarities in protein structures"*, to be published (1994).

[8] Kleywegt, G.J., Vuister, G.W., Padilla, A., Knegtel, R.M.A., Boelens, R. and Kaptein, R. J. Magn. Reson. (Series B), *102* (1993) 166.

[9] Jones, T.A., Bergfors, T., Sedzik, J. and Unge, T. EMBO J., *7* (1988) 1597.

[10] Cowan, S.W., Newcomer, M.E. and Jones, T.A. J. Mol. Biol., *230*, 1225.

[11] Uppenberg, J., Hansen, M.T., Pathar, S. and Jones, T.A. *"The sequence, crystal structure and refinement of two crystal forms of Lipase B from Candida antarctica"*, submitted for publication (1994).

Use of the free R-Factor as a guide in parameter optimisation for density modification

by

Jonathan Grimes and Dave Stuart
Laboratory of Molecular Biophysics
Rex Richards Building, South Parks Road, Oxford, OX1 3QU, U.K.

The free R-factor

A key stage in most macromolecular crystal structure analyses is the improvement of the phases of experimentally measured structure factor amplitudes by the modification of a real space model in order to improve the fit in reciprocal space to the observed data. A recurrent problem in this process has arisen from the difficulty of establishing whether the modifications to the model are providing a genuine improvement in the phase angles or simply over-fitting the structure factor amplitudes. Brünger (1993) has described a simple measure to address this problem, the 'free R-factor'. In his protocol a selection of measured structure factor amplitudes are excluded from the optimisation step and thereby become available for use as monitors of the process – if the model modification improves the agreement between structure factor amplitudes derived from the model with these 'closet' data then we may assume that the phase components of the structure factors are similarly improved. Brünger has implemented the method in the program XPLOR (Brünger, 1992) which is designed to optimise atomic models. We have used essentially the same technique at an earlier stage, where an electron density map is available for modification. Since we began our work a closely similar approach has been reported by Baker *et al.* (1993). At the outset we expected that this technique might be useful in providing objective answers to some questions that tend to come up over and over again when crystallographers modify their electron density maps. These issues cover several areas:–

- Envelope definition. In particular can one define a value for the percentage solvent content that will produce the lowest final mean phase error?
- Phase combination. Given that we often have SIR or MIR phase estimates, albeit poor, should they be used in phase combination with phases derived from the averaged map during the cyclic phase refinement, or thrown away after calculation of the first map?
- Weighting scheme. Is there an optimum weighting scheme for the structure factors used in electron density map calculation, that gives reliable and rapid convergence to the best set of phase angles?

- The use of unobserved data. Should one substitute $|F|_{calcs}$ (derived from inversion of the modified electron density map) for unmeasured $|F|_{obs}$ in the FFTs, and if so how should one weight these substituted data?

- Phase extension. Is it best to "throw away" the phases at high resolution, and phase extend from a lower resolution, where the phases are better, simply using the non-crystallographic redundancy and solvent regions as phase constraints?

Note that these answers are obtained at a cost, the free R-factor method entails throwing away some hard won data from the optimisation process and and this will in itself weaken the optimisation. We address below the effect of loss of data on the phase improvement process – we would argue that the free R-factor is best seen as a method of designing optimal protocols and perhaps once it has fulfilled this role all the available data should be used to obtain the definitive electron density map.

Implementation

Our implementation is directly analogous to the free R-factor in XPLOR (Brünger, 1993). N% of the observed data are randomly selected (4% in our test cases) and flagged. Throughout cycles of map modification and inversion the flagged data are treated as unobserved. As the modification process develops, estimates of them (as seen in the $|F|_{calcs}$) are produced and it is the agreement between the flagged observations and the corresponding $|F|_{calcs}$ which is measured by the free R-factor equation. This equation is the usual R-factor equation, but calculated using only the flagged subset of reflections (h', k', l').

$$R^{free} = \frac{\sum ||F|_{obs}(h', k', l') - k\,|F|_{calc}(h', k', l')|}{\sum |F|_{obs}(h', k', l')} \quad (1)$$

The free R-factor thus monitors how well the density modification process replicates unobserved data and provides us with an objective assessment of the success of the method. Our implementation is centred on the general purpose averaging/map modification programme GAP (Grimes and Stuart, unpublished) which is flanked by a number of CCP4 programmes (principally FFT, SFALL and SIGMAA) and locally written programmes that use MTZ format data files (for selection of free R-factor reflections, data merging and scaling, monitoring phase changes, etc) and will not be described in detail here. The purpose of this paper is to investigate whether the free R-factor can be used as a sensitive and reliable indicator of phase error. To this end we have performed a number of tests using a case study structure, Bluetongue virus VP7 (BTV VP7), which has been solved recently in our laboratory.

Brief outline of structure solution

BTV VP7 is a trimer of 114kD, and in our monoclinic crystal form there are 2 trimers in the asymmetric unit. A single gold cyanide derivative data set was collected to 3.5Å spacings, with one heavy atom site per subunit, but unfortunately with no useful anomalous

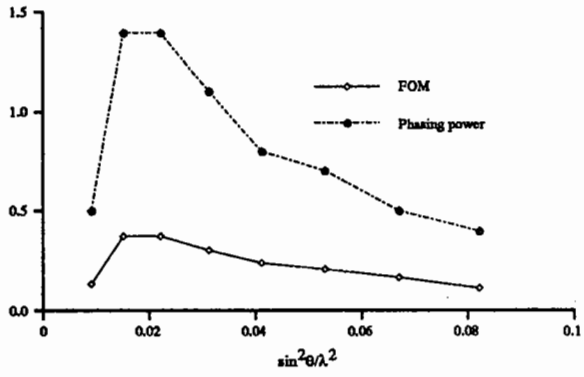


Figure 1: BTV SIR phases: figure of merit and phasing power with respect to resolution.

scattering signal. The overall average figure of merit was 0.2 to 3.5Å and the phasing power was similarly unimpressive (see figure 1). The heavy atom sites were solved using GROPAT (Jones and Stuart, 1991), refined in MLPHARE (Otwinowski, 1991) and the SIR map was then cyclicly solvent flattened using GAP. A solvent content of 50% was used and the solvent envelope automatically redetermined every 4 cycles using a procedure developed from that of B. C. Wang (1985). The molecular unit envelope was defined by a logical .and. procedure of the final "Wang" type envelope, extended to cover the 2 non-crystallographic trimers, and 2 sets of planes, each defined as (i) normal to and halfway along vectors between heavy atom sites in the molecular unit and their nearest crystallographic partners, and (ii) between the centres of the constellation of heavy atom sites in each non-crystallographic trimer. This seemed sensible, because the arrangement of the heavy atom sites conformed to the proper 3-fold molecular symmetry and to our assumption of the relationship of each trimer with its neighbour. Cyclic averaging and solvent flattening were carried out at 3.5Å resolution (now with a solvent content of 40%), with the solvent envelope being redetermined and the non-crystallographic operators being refined a number of times. A very crude model was then built into the 3.5Å map (there was no ambiguity over the chain trace, but no attempt was made to ensure the correct alignment of the amino acid sequence with electron density map and over $\frac{1}{3}$ of the model was out of register by 3 amino acids) and then refined in XPLOR using strict non-crystallographic constraints, initially at 3.5Å, then to 3Å and finally using all the available data to Bragg spacings of 2.6Å. This gave a model with an R-factor of 28% to 2.6Å. Cyclic averaging and solvent flattening were then resumed at 2.6Å, and after 20 or so cycles the process had converged. The reciprocal space R-factor was 12% and the correlation coefficient was 97%. After a round of rebuilding that involved substantial changes that were nonetheless trivial to perform, the model was refined in XPLOR, using strict non-crystallographic constraints, to an R-factor of 19.3%. In order to obtain a 'true' set of phases for our tests of the free R-factor, a $2|\boldsymbol{F}|_{\mathrm{obs}} - |\boldsymbol{F}|_{\mathrm{calcs}}$ electron density map was calculated using the XPLOR phases, solvent flattened, 6-fold averaged and back transformed. The linear correlation coefficient between the 6 copies of the electron density was 96% and we are confident that the derived phases are indeed highly accurate. In the

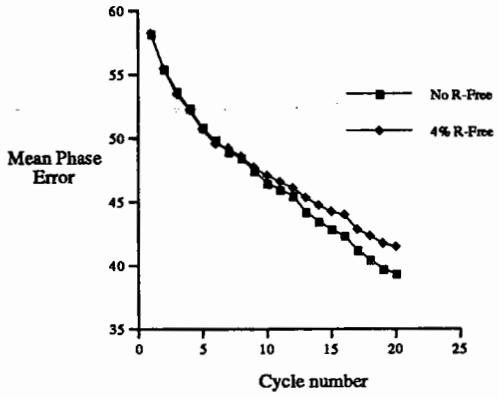


Figure 2: The change in mean phase error during cyclic averaging with and without the use of the free R-factor, showing the minimal effect on convergence.

tests that will be described below, the SIR/solvent flattened map at 3.5Å was used as a starting point (unless otherwise stated) and the molecular unit envelope was derived from the refined coordinates. This is therefore a somewhat idealised case where one has an optimised molecular unit envelope, but provides consistency between tests. The tests then consisted of twenty cycles of 6-fold averaging and solvent flattening, carried out using the optimised non-crystallographic symmetry operators derived from the final XPLOR refinement cycles.

Does the free R-factor harm convergence?

The first point that has to be addressed concerning the use of the free R-factor, is to what extent does it harm the convergence of phase refinement? Two tests were done, one using all the data, and the other with a random 4% of the data excluded. This 4% was used to calculate the free R-factor. Figure 2 shows that cutting out 4% of the data has only a small effect on convergence. There is a 3° increase in the final mean phase error when the data are removed for free R-factor calculation, with the free R-factor run having a 42° mean phase error. As expected one gets the best results using all the data, but there is no reason why, once the free R-factor has allowed the optimisation of the phase refinement procedure, the process should not be repeated using all the data in order to derive the best possible set of phases.

Optimisation of the percentage solvent content

Several runs were done, using solvent envelopes calculated with varying fractions of the unit cell designated as solvent, to see whether the free R-factor could give any indication of the optimum value to choose (i.e. that giving the lowest difference from the "true" phases). As can be seen from figure 3 the change in the free R-factor mirrors very closely the change the mean phase error as the solvent content is varied. However the conventional averaging R-factor is rather insensitive to the choice of solvent content and, as expected,

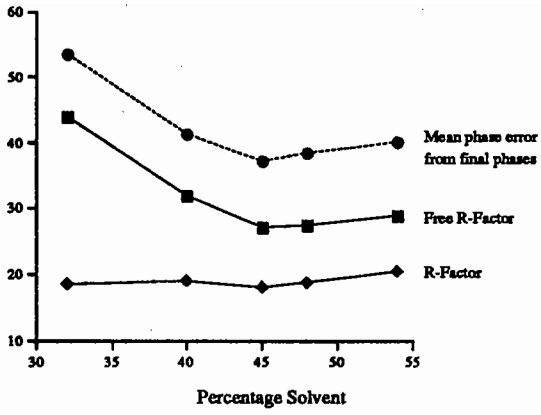


Figure 3: The change in mean phase difference, free R-factor and R-factor after 20 cycles of averaging/solvent flattening when using solvent envelopes calculated with various solvent contents.

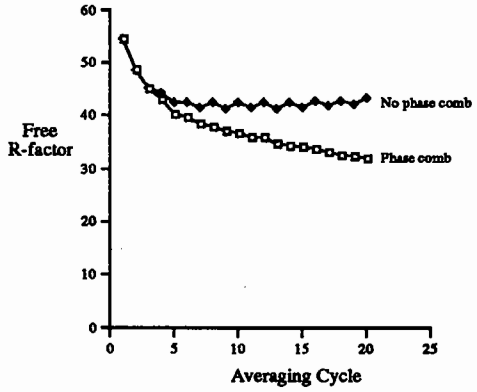


Figure 4: The difference in free R-factor between tests using phase combination and no phase combination.

is misleading, favouring an over-conservative estimate of the solvent content. In fact reducing the percentage solvent content produces a marked loss of phasing power, which leads to an increase in mean phase error and increase in the free R-factor. Although there seems to be little drop in phasing power as the solvent content is increased substantially beyond 45% of the unit cell this is partly artifactual since in our test protocol the averaged density is "folded down" into the asymmetric unit of the solvent flattened map using the accurate molecular unit envelope, which will "rescue" some protein pixels. Nonetheless it is clear that the free R-factor can provide a useful guide to the most appropriate choice of solvent content.

Phase combination?

Given the poor quality of the starting phases, we were interested to see whether the use of phase combination (between the SIR and averaged phases) produced lower phase errors than simply using the averaged phases during cyclic averaging/solvent flattening, and whether changes in the average phase error were mirrored in the free R-factor. Good SIR or MIR phases will obviously provide a useful restraint on the averaged phases during refinement, but it is less clear what effect our poor SIR phases (overall average figure of merit of 0.2) will have on the process; whether they will provide a useful phase restraint or slow down convergence. Two runs, one with and one without phase combination were done. For the phase combination test run Rayment weighting (see description of weighting schemes given below) was used for the averaged phases and these were then combined with the SIR phases using SIGMAA. A joint weighting scheme (Rayment*Sim) was used for the control case in an attempt to mimic as closely as possible the weights in the phase combination run (however it will be seen from the following section that this is not ideal). After 20 cycles the final mean phase errors for combined and uncombined runs were 41.4° and 47.1° and this is reflected well by the free R-factor (fig 4). Clearly even poor SIR phases provide a very useful restraint on the averaging procedure especially in the initial stages of phase refinement where they considerably speed convergence. For the final stages of averaging SIR/MIR phases will receive very little weight and will therefore have little effect. Thus after 40 more cycles the free R-factors for combined and uncombined were 23.1% and 22.4% respectively, with both having a mean phase error of 32.4°. Note that for neither run had convergence been achieved within the first 20 cycles.

Amplitude weighting

How should the Fourier coefficients be weighted? We have tested five schemes.

- Sim
- Sim*Rayment
- Rayment
- (Rayment)^2
- Unit

Sim weighting is defined as :

$$W = \begin{cases} \tanh X & \text{for centric reflections} \\ \frac{I_1(X)}{I_0(X)} & \text{for acentric reflections} \end{cases} \qquad (2)$$

where $X = 2\,|\boldsymbol{F}|_o\,|\boldsymbol{F}|_c \,/ \left(\frac{\sum_N (|\boldsymbol{F}|_o^2 - |\boldsymbol{F}|_c^2)^2}{N} \right)^{\frac{1}{2}}$ and I_0 and I_1 are zero and first-order modified Bessel functions (Sim, 1960).

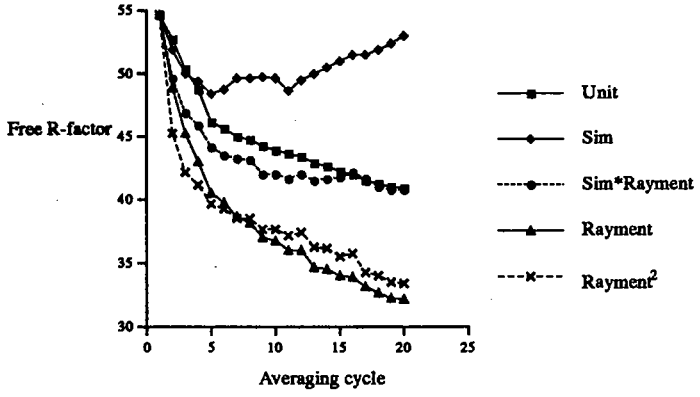


Figure 5: The mean phase errors for tests carried out using different weighting schemes.

Rayment weights are defined (Rayment, 1983) as

$$\exp \frac{-||\boldsymbol{F}|_c - |\boldsymbol{F}|_o|}{|\boldsymbol{F}|_o} \tag{3}$$

Phase combination was not used to avoid the introduction of further weighting in SIGMAA between the SIR and averaged phases. As can be seen from figure 5, in our VP7 test case the free R-factor indicates that Rayment and (Rayment)^2 weights are superior to either Sim, joint or unit weights. The free R-factor was lower for Rayment weighting and produced the lowest phase errors, 41.4° for simple Rayment weights and 40.7° for (Rayment)^2, which interestingly seems to speed up convergence. The final mean phase errors for the joint, unit and Sim weight test runs were, respectively, 42.7°, 48.1°and 51.3°.

Investigations by Rayment (1983) suggested that his scheme produced good convergence, but found that on convergence, switching to unit weights then gave more accurate phases. This is the strategy usually employed by our research group.

Missing amplitudes and how to weight them

An issue that has been addressed by Nordman (1980), Rossmann (1990) and Rayment (1983), amongst others, is one of missing data and the problems that it will cause on convergence of phase refinement. If, as is often the case in the highest and lowest resolution shells, a large proportion of the data are not measured, maps calculated at this resolution will have series termination errors due to the discontinuity of the Fourier transform in reciprocal space, which can cause severe problems to the convergence of phase refinement.

As a matter of course, we use a procedure where these missing data are filled in by values determined from map inversion during cyclic averaging. The problem then arises of how these data should be weighted, since Rayment weights can not be calculated due to the lack of observed data. Rayment suggests a weighting of about 0.9 for the included $|\boldsymbol{F}|_{\mathbf{calcs}}$. Previously we have usually applied an arbitrary overall mean weight of 0.7. However a more objective way of defining weights is now possible by using those reflections involved in the free R-factor calculation. Weights are calculated for these data in the normal

way and these weights are then averaged within resolution shells. This average weight can then be applied to the unobserved data in that particular shell. Typical values for Rayment weights for $|F|_{obs}$ and $|F|_{calcs}$ after cycle 1 are 0.61 and 0.60 respectively, for 25517 observed and 3438 "unobserved" data, and after 20 cycles these values increased to 0.8 and 0.71.

Phase extension or not?

Another question that is of interest is, given poor starting phases, will the final phase errors be lower if one discards phases at higher resolution and extends slowly from the better determined lower resolution phases, or if one starts phase refinement at the resolution limit of the derivative phasing?

In the test case of VP7, phase extension was carried out from 5.0Å to 3.5Å simply using the non-crystallographic redundancy and solvent flattening as phase constraints. The SIR phases were not used to restrain the averaged phases during phase extension. The phases were extended in step sizes of $\frac{1}{500}$Å. This very gentle phase extension process gave phases that were no better than those obtained starting directly at 3.5Å with the SIR phases (after phase extension the mean phase error was 33.5° and the free R-factor was 22.9%, compared to a mean phase error of 32.4° and free R-factor of 23.1% obtained by direct refinement at 3.5Å).

However in many cases phase extension has been the key to successful phase refinement and is the only way forward if no phase information is available. The present results with VP7 may reflect the considerable non-crystallographic redundancy, which provides sufficient power to drive close to random phases to their correct values, and may not be generally applicable. An important point is that with reduced symmetry redundancy (or even none; i.e. for simple solvent flattening) the free R-factor should allow a check on the success of the strategy adopted.

A further point worth mentioning, is that in the case of VP7 and also BEV (Symth, Tate and Stuart, unpublished) crude models were built at low resolution, and then refined in XPLOR to higher resolution. In both cases strict non-crystallographic symmetry constraints were used. In the case of VP7 the resolution was extended from 3.5Å to 2.6Å in 2 steps. Similarly for BEV, with 60-fold redundancy, a model was built into a superb 4.3Å map (essentially no noise) which was able to provide phases immediately to 3.0Å. With this type of strategy we are effectively using a whole mass of extra chemical information to allow the phase extension process to be speeded up enormously.

Phase perturbation

An idea we are developing that seems to be of value is the use of phase perturbation in phase refinement. It is somewhat analogous to "simulated annealing" in XPLOR (Brünger, 1992) and involves perturbing the phase angle for each reflection by some value. The perturbation is chosen by randomly consulting a Maxwellian distribution with width dependant on the weight of the particular reflection. These new phases are then

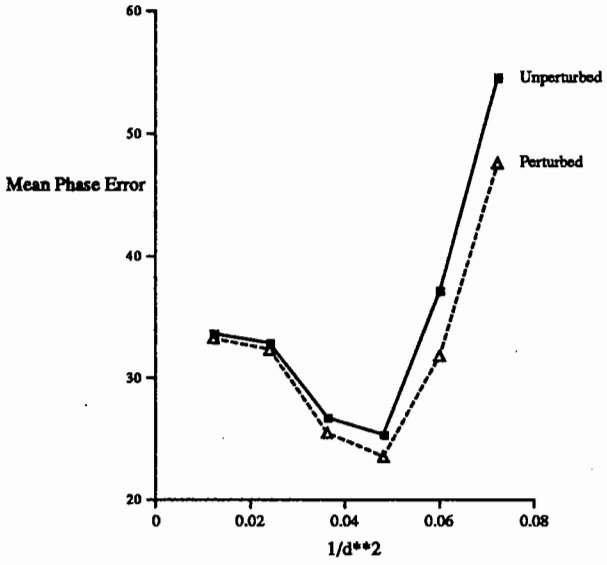


Figure 6: The differences in final mean phase error between tests using phase perturbation and a control.

refined before repeating the perturbation process. "Annealing" is achieved by gradually contracting the width of the distributions.

For our test case phase convergence is already very good and so there is little to gain, but even here phase perturbation does seem to help, especially in improving the less well phased high resolution reflections. The free R-factor for the test run with annealed phase perturbation was 21.0% as opposed to 23.1% for the control case with a mean phase error of 29.7° as against 32.4° (see figure 6).

Conclusion

In all of our test cases of phase refinement strategies the free R-factor provides a reliable and objective monitor of convergence and phase error.

We expect that most of the results obtained from our tests will be relevant to *de novo* structure determinations, but it should be noted that the test case was idealised in that the molecular unit envelopes, within which the averaging is done, were accurately defined, as were the non-crystallographic symmetry operators. The effect of these felicities will have been to render the refinement process more robust than it would be in a real structure determination. Nevertheless the free R-factor is a sensitive indicator of phase error and, for those issues we have addressed, has allowed optimisations which have produced significant improvements in the rate of convergence of the refinement and in the final mean phase error. In real cases these issues may prove to be vital for successful structure determination.

References

Baker, D., Bystroff, C., Fletterick, R. J. and Agard, D. A. (1993). PRISM: Topologically constrained phase refinement for macromolecular crystallography. *Acta Cryst.* **D49**, 429–439.

Brünger, A. T. (1992). *XPLOR version 3.1.* The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, C.T., U.S.A.

Brünger, A. T. (1993). Assessment of phase error by cross validation: the free R value. Methods and application. *Acta Cryst.* **D49**, 24–36.

Jones, E. Y. and Stuart, D. I. (1991). Locating heavy atom sites by automatic Patterson search - GROPAT. In *Isomorphous replacement and anomalous scattering*, pages 39–48, Daresbury Laboratory, Warrington, UK.

Nordman, C. E. (1980). Procedures for detection and idealisation of non-crystallographic symmetry with application to phase refinement of the Satellite Tobacco Necrosis virus structure. *Acta Cryst.* **A36**, 747–754.

Otwinowski, Z. (1991). Maximum likelihood refinement of heavy atom parameters. In *Isomorphous replacement and anomalous scattering*, pages 80–86, Daresbury Laboratory, Warrington, UK.

Rayment, I. (1983). Molecular replacement method at low resolution: optimum strategy and intrinsic limitations as determined by calculations on icosahedral virus models. *Acta Cryst.* **A39**, 102–116.

Rossmann, M. G. (1990). The Molecular Replacement Method. *Acta Cryst.* **A46**, 73–82.

Sim, G. A. (1960). A note on the heavy-atom method. *Acta Cryst.* **13**, 511–512.

Wang, B. C. (1985). Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol.* **115**, 90–111.

From a poor MIRAS-map to a final structure : importance of 2-fold averaging in the structure determination of thiolase

Magali Mathieu

Rik K. Wierenga

EMBL
Postfach 10.2209
D-69012 Heidelberg
GERMANY

Introduction

Thiolases are ubiquitous enzymes involved in various different metabolic pathways. We are interested in yeast (*S.cerevisiae*) peroxisomal thiolase, an enzyme of the peroxisomal β-oxidation pathway. This project is a collaboration with Dr. Kunau (Ruhr University, Bochum, Germany). The structure determination is interesting because not much is known about the structure-function relationship of this class of enzymes ; for example, the thiolase sequence is not homologous to any sequence of a protein whose crystal structure is known. In addition the structure determination of thiolase might give insight into the import mechanism of proteins synthesized in the cytosol but active in the peroxisomes.

Data collection

Data were collected on an Enraf-Nonius FAST area detector and on a MAR image plate area detector. Several heavy atoms were tested, but only those leading to derivatives are indicated in table 1, along with the native datasets collected.

Table 1 : Data collection statistics

	collected on	Max. resol.	nb. of obs.	nb. of unique refl.	Rmerge (Int)	Completeness
native	Image Plate	2.77 A	46818	17367	7.9 %	91.7 %
native	Fast	2.67 A	46765	17667	5.1 %	74.9 %
$KAuCl_4$ (1)	Image Plate	2.71 A	70097	19972	6.8 %	86.5 %
$KAuCl_4$ (2)	Image Plate	3.02 A	52960	10097	8.1 %	60 %
$KAu(CN)_2$	Image Plate	2.71 A	78419	16704	7.7 %	72.5 %
$KAuI_4$	Fast	3.23 A	22237	10343	9.1 %	76.1 %
$K_3UO_2F_5$	Image Plate	4.02 A	12250	5108	12.4 %	69.3 %
$(NH_4)_2U_2O_7$	Fast	3.10 A	57949	11041	8.9 %	73.1 %
native (at 100K)	Fast	2.4 A	63078	27248	6.1 %	84.1%

Thiolase is active as a dimer, and crystallizes with a dimer per asymmetric unit in the space group $P2_12_12_1$, using the microdialysis method (1). Crystals were stabilised for storage and data collection with a buffer containing 200 mM DTT. The 2.7 Å native datasets and the heavy atom datasets (table 1) were collected at room temperature. Using cryo-cooling methods, a high resolution (2.4 Å) dataset was also collected. The cell dimensions are, at room temperature 71.78 Å, 93.72 Å, 120.45 Å, and at 100K 71.00 Å, 92.10 Å, 117.70 Å. The solvent content is 46 % and 43 % respectively.

Use of non-crystallographic symmetry

The first native dataset, collected to 2.7 Å, was used with the GLRF package (2) to find the orientation and position of the local two-fold axis of the dimer.

We first used the self-rotation option, for a kappa of 180°. In this option, the self-rotation function is calculated as originally proposed by Rossmann and Blow (3). Different resolution shells were tried (10 to 4Å, 10 to 7Å and 7 to 4Å), different protein radii (25, 30 and 40 Å) and different cut-off values for the amplitudes of the reflections. One consistent peak was found in all those searches. This peak was most of the time the first peak after the origin peak and was of 1.8 to 2.2 sigma above the background, depending on the parameters used.

The first derivative was obtained by soaking a crystal in a mother solution containing 1 mM $KAu(CN)_2$. The heavy atom sites were found by visual inspection of the Harker sections of the Patterson map. Sites and occupancies were refined using the CCP4 program mlphare (4). That derivative gave a phased dataset until 4 Å, with figures of merit (centric and acentric respectively) of 0.57 and 0.37.

We used the SIRAS phases for the calculation of a phased translation function (option tfun=2) of the GLRF package (5). The function searches for the molecular centre by calculating the quality of the superposition of the original map around a position S, on the map rotated by the local two-fold

axis when centered at position S. This calculation is done for each S in the asymmetric unit. All grid points within a radius of 25 Å of position S are used for calculating the values of the overlap function. Maximum overlap occurs when position S coincides with the molecular centre (6).

In our case however, the quality of the heavy atom information from the SIRAS map was not sufficient to find a significant peak.

Another derivative was eventually found, $KAuCl_4$, which did bind to the same sites as observed for $KAu(CN)_2$, but with higher occupancy. Combining the two derivatives (MIRAS) resulted in a figure of merit of 0.63 and 0.41 to 4 Å, and 0.52 and 0.34 to 3 Å. We used again option tfun2 of GLRF, to 5, 4, and 3Å, with datasets phased to the corresponding resolution, for each enantiomeric form of the heavy atom sites. We did not have any result to 5 Å, but, despite the low figures of merit, we could see a consistent peak to 4 and 3 Å for one of the enantiomeric forms.

This position of the two-fold axis was then checked visually by looking at a map of a unit cell centered on the predicted molecular centre. Some features in the map indeed agreed with the presence of a local two-fold axis.

Having found the orientation and the position of the local two-fold axis, we optimised them using programs from the "O" package (7). We first created a mask by manual editing of the bones, followed by using the program bones_to_mask ; it should be noted that the O-package contains now better and easier ways of creating a mask. The program a_rt_improve was then used to improve the orientation and position of the 2-fold axis. The shifts were quite significant :

old orientation = 78° 63° 180° ; translation : 76.47 -77.95 118.46 (in Å)
new orientation = 76.6° 62.3° 180° ; translation : 73.95 -79.77 115.14 (in Å)

Subsequently, the new orientation/position of the local two-fold axis was used in averaging procedures, using O and CCP4 programs, to improve the map (table 2). The averaging procedure consisted of 10 cycles of averaging, map inversion and calculation of a new Fo, α_c map.

Table 2 : Correlation coefficient between subunit-1 and subunit-2 in the MIRAS map obtained with KAuCl4 and KAu(CN)2

before optimization of the 2-fold axis	0.274
after optimization	0.360
after averaging around the 2-fold axis	0.761

The improved phases obtained were then used in a difference Fourier procedure to screen through all heavy atom datasets already collected, in order to try to find out if these datasets could be used as additional heavy atom derivatives. Three weak derivatives were found by that method : $KAuI_4$, $K_3UO_2F_5$ and $(NH_4)_2U_2O_7$. It is interesting to note that all gold derivatives occupy the same two sites, related by the local 2-fold axis. The uranyl sites are situated rather far away from the 2-fold axis of the dimer, but two of them, present in both derivatives, are related by the 2-fold axis. The final MIRAS-statistics (6 derivatives) are summarised in table 3.

Table 3: Refinement statistics (to 3.1 Å)

Figure of merit (to 3.1 Å) : 0.4580 (acentric), 0.5994 (centric)

	Soaking time	Concentration	Rdifference	nb. of sites	Rcullis (centric data)	Phasing power centr./acentr.
$KAuCl_4$ (1)	2 nights	1 mM	11.1 %	2	0.75	1.2 / 0.9
$KAuCl_4$ (2)	1 night	1 mM	11.4 %	2	0.81	0.8 / 0.5
$KAu(CN)_2$	3 nights	1 mM	9.8 %	2	0.80	0.8 / 0.6
$KAuI_4$	2 nights	1 mM	20 %	2	0.97	0.6 / 0.5
$K_3UO_2F_5$	1 night	2 mM	20.1 %	2	0.86 (to 4Å)	0.7 / 0.5 (to 4Å)
$(NH_4)_2U_2O_7$	1 week	1 mM	26.6 %	3	0.94	0.5 / 0.4

Averaging

The map obtained from these 6 derivatives had better boundaries between solvent and protein. It permitted us to make a better mask of the dimer, which was used in new averaging cycles, after another improvement of the orientation and position of the two-fold axis. The phase shifts between the MIRAS-phases and the phases from the averaged map are shown in table 4.

Table 4:Phase shift between the MIRAS and the averaged phases (as a function of the figure of merit of the MIRAS phases)

	Figure of merit	number of reflections	average absolute phase shift
	0.0 - 0.1	1749	87.628
	0.1 - 0.2	1731	86.455
	0.2 - 0.3	1485	78.346
	0.3 - 0.4	1301	71.266
	0.4 - 0.5	1226	68.045
	0.5 - 0.6	1252	60.524
	0.6 - 0.7	1247	56.624
	0.7 - 0.8	1420	49.750
	0.8 - 0.9	1423	41.096
	0.9 - 1.0	1077	30.447
total :	0.469	13911	64.784

It can be noted that the phase difference is almost random for reflections with a low figure of merit, while it progressively diminishes as the figure of merit increases.

We could then start building a polyalanine model into the electron density, at 3.1 Å resolution. This incomplete polyalanine model was refined with TNT (8) at 3.1Å with the "real space" option, and then used in phase combination with the sigmaa program (9) of the CCP4 suite. This method really helped by emphasizing the wrong segments of the model, which could then be corrected. The process of phase combination using map phases and model phases was repeated several times, starting always with the same map (the averaged MIRAS map), until it was possible to see a part of the sequence of thiolase, with the help of the "slider" options of O. With that method, we were able to build around 80% of the sequence.

We then decided to try another averaging procedure, using the DEMON package(10). It involved :

- another slight improvement of the 2-fold axis, using the AVGSYS (11) suite

- a new definition of the mask, based on a correlation map between the subunits : the cut-off of that correlation map was chosen such as nearly covering the model available at the time ; that corresponded to a correlation coefficient of roughly 23 %

- a new weighting scheme, namely use of Sim weights (12) instead of sigmaa weights

- Structure factor completion : in the absence of a measured structure factor, the calculated structure factor was used

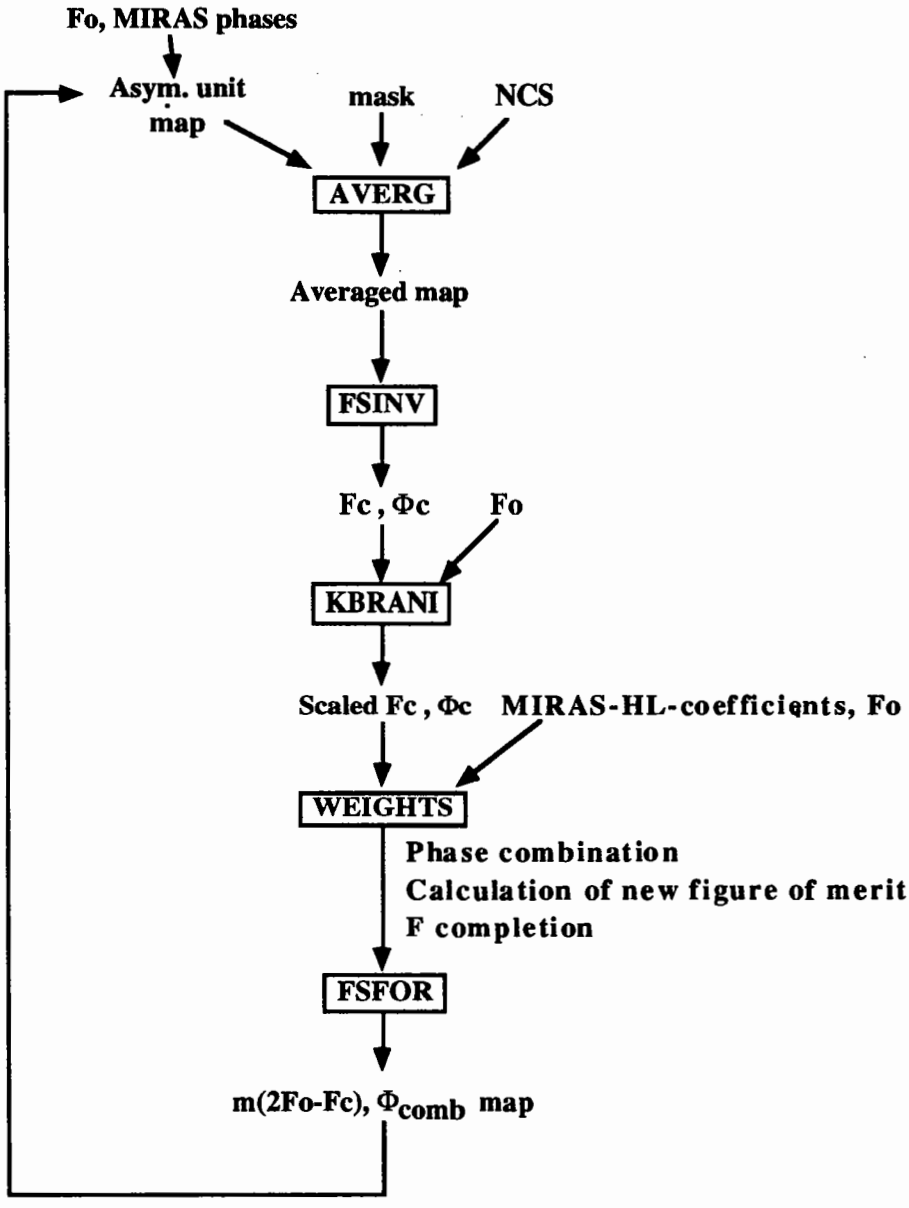
- and of course different programs.

In that protocol, we also tested by visual inspection of the map whether phase combination produced better phases, and also whether we should use 2Fo-Fc or Fo maps. We subjectively decided in favor of phase combination and 2Fo-Fc maps.

This procedure, described in figure 1 has resulted in a complete tracing of the chain of the two monomers.

It should be noted that the phase combination of the MIRAS-phases and the calculated phases by the program "weights" of the DEMON package is done by adding up the Hendrickson-Lattman coefficients of the MIRAS-phases (as provided by mlphare) and those of the calculated phases, as calculated from the discrepancy between Fo and Fc (for example, using Sim-weights). From the final Hendrickson-Lattman coefficients a new figure of merit is calculated, which is then used in the map calculation.

Figure 1 : Averaging procedure using the DEMON package
(only the main steps are indicated ; program names are written in boxes)



Discussion

The presented protocol allowed us to determine a complete chain tracing of the thiolase fold at 3.1 Å resolution.

We think that the following aspects of that protocol have been of particular importance in this structure determination :

- Extracting the maximum of information from the heavy atom derivatives, by using the improved phases, obtained after averaging, in a careful difference Fourier analysis of all datasets. In addition we always used the anomalous signal.

- careful definition of the boundaries of the mask, which had to take into account the putative presence of the missing fragments.

- using "optimal" weights in the phase combination protocol.

The final model, obtained at 3.1 Å resolution, was transferred to the "cryo-cooling cell" and is now being refined. We refined it first with X-PLOR (13), using Non-Crystallographic Symmetry restraints. It was quickly evident that two loops (consisting of approximately 30 residues each) were behaving strangely. The refinement was continued, omitting these loops in both monomers. The subsequent omit and difference maps showed marked differences between these loops in the two monomers. Further cycles of refinement were done, to build progressively those loops. At the moment, refinement of the structure is going on at 2.4 Å, without applying non-crystallographic symmetry restraints, using TNT.

Acknowledgments : we thank Drs. Noble and Read for discussions and Dr. Read for the implementation of the DEMON package.

(1) J.Ph.Zeelen, R.K.Wierenga,R.Erdmann and W.-H.Kunau (1990), J.Mol.Biol. **215**, 211-213
(2) L.Tong and M.G.Rossmann (1990), Acta Cryst. **A46**, 783-792
(3) M.G.Rossmann and D.M.Blow (1962), Acta Cryst. **15**, 24-31
(4) Z.Otwinowski (1991), In "*Daresbury Study Weekend Proceedings*" (W.Wolf, P.Evans and A.G.W.Leslie, ed.), pp.80-86, SERC, Daresbury
(5) L.Tong, H.Choi, W.Minor and M.G.Rossmann (1992), Acta Cryst. **A48** 430-442
(6) P.Argos and M.G.Rossmann (1980), In "*Theory and practice of direct methods in crystallography*" (M.F.C.Ladd and R.A.Palmer, ed.), pp.361-417, Plenum Press, New York and London
(7) T.A.Jones, J.-Y.Zou, S.W.Cowan and M.Kjeldgaard (1991), Acta Cryst., **A47**, 110-119
(8) D.E.Tronrud (1992), Acta Cryst. **A48**, 912-916
(9) R.J.Read (1986), Acta Cryst. **A42**, 140-149
(10) F.M.D.Vellieux, J.Hajdu, C.L.M.J.Verlinde, H.Groendijk, R.J.Read, T.J.Greenhough, J.W.Campbell, K.H.Kalk, J.A.Littlechild, H.C.Watson and W.G.J.Hol (1993), Proc.Natl.Acad.Sci.USA **90**, 2355-2359
(11) J.L.Smith (1993), Purdue University, W.Lafayette, USA, AVGSYS package
(12) G.A.Sim (1959), Acta Cryst., **12**, 813-815
(13) A.T.Brünger, J.Kuriyan, and M.Karplus (1987), Science **235**, 458-460

Protein Models: Past Progress and Future Possibilities

Lyle H. Jensen
Department of Biological Structure and the Department of Biochemistry
University of Washington, Seattle, WA 98195 USA

Since the earliest days of structure determination based on X-ray diffraction data, adjusting the general parameters of crystal models has been essential in deriving the most reliable results from a set of data. Ingenious methods have been devised for refining model parameters, and an extensive literature had developed covering all aspects of the various methods and their application. Space will permit us to trace only the main methods of refining small structural models before going on to treat the developments and status of refining protein models and to consider some future possibilities.

Past Progress

F_o series

In 1915, only three years after the discovery of X-ray diffraction by crystals, W.H. Bragg suggested that the electron density $\rho(x,y,z)$ in a crystal could be represented by a Fourier series (W.L. Bragg, 1962). Such a series is given by the equation:

$$\rho(x,y,z) = 1/V \sum_h \sum_k \sum_l |F_{hkl}| \cos 2\pi(hx + ky + lz - \alpha_{hkl})$$

where V is the volume of the unit cell, $|F_{hkl}|$ is the magnitude of the structure factor h,k,l and α_{hkl} is its phase. The summation is over all reflections h,k,l. But in practice there was a problem: only the intensities I_{hkl} are observed--the phases are ordinarily lost on recording the data. Until recent years, therefore, phases for small structures were usually calculated for an assumed trial model. If the assumed model was sufficiently close to the actual structure, the Fourier map calculated by use of F_{obs} and α_{calc} would lead to a better model than the initial one. Successive cycles of F_o maps moved atoms successively closer to their true positions, i.e., the model was refining. When no further changes occurred in the atomic parameters, and thus the phases, the refinement was said to have converged. Refinement by F_o maps provides a direct view of the electron density, $\rho(x,y,z)$ in real space.

Least squares

The method of least squares goes back to Legendre (1806) and was first applied to crystallographic refinement by E.W. Hughes (1941). As applied in this field, the principle asserts that the best values of the atomic parameters minimize the sum of the weighted squares of the differences between the oberved and calculated structure factors, $\sum w(|F_o| - |kF_c|)^2$, k being the reciprocal of the scale factor of F_o and must be applied to F_c within each refinement cycle. The method can be applied only when the number of observations exceeds the number of parameters. In combination with various subsidiary conditions such as bond lengths and angles, the method is at the heart of numerous computer progams that have been developed to refine macromolecular models.

(F_o-F_c) series

In a definitive paper in Acta Crystallography, W. Cochran (1951) showed the power of (F_o-F_c) series in refining crystal models. The following personal account will show how the application of (F_o-F_c) maps in refining a small structural model in projection would serve years later to play a key role in successfully refining a protein by the classical methods developed for small molecule structures.

Having accepted a position in 1949 in one of the basic science departments of the University of Washington School of Medicine in Seattle, I decided to begin with small molecules of medical or biological importance, in part because of the limited computational facilities available at the time, and in part to familiarize myself with every step of the structure determining process. The first of these was isonicotonic acid hydrazide ($C_5H_4NCONHNH_2$) which, along with dihydrostreptomycin, was proving so effective in treating tuberculosis. The crystals were beautiful hexagonal needles, a fraction of a millimeter in cross section, optimal in size for the Weisenberg camera used to collect data for the three principal zones: hk0,h0l,0kl. The crystals proved to be orthorhombic, a=11.22Å, b=14.74Å, c=3.84Å, space group $P2_12_12_1$. Thus, projections of the structure on each unit cell face are centrosymmetric with phase angles either 0 or π, and the short c axis insures resolution of all atoms in projection along that axis. The projection along c, i.e., on 001, was easily solved from the Patterson projection and Bragg-Lipson graphs for a few low order hk0 reflections. The x and y parameters were refined by a series of F_o maps projected on (001), Fig.1c.

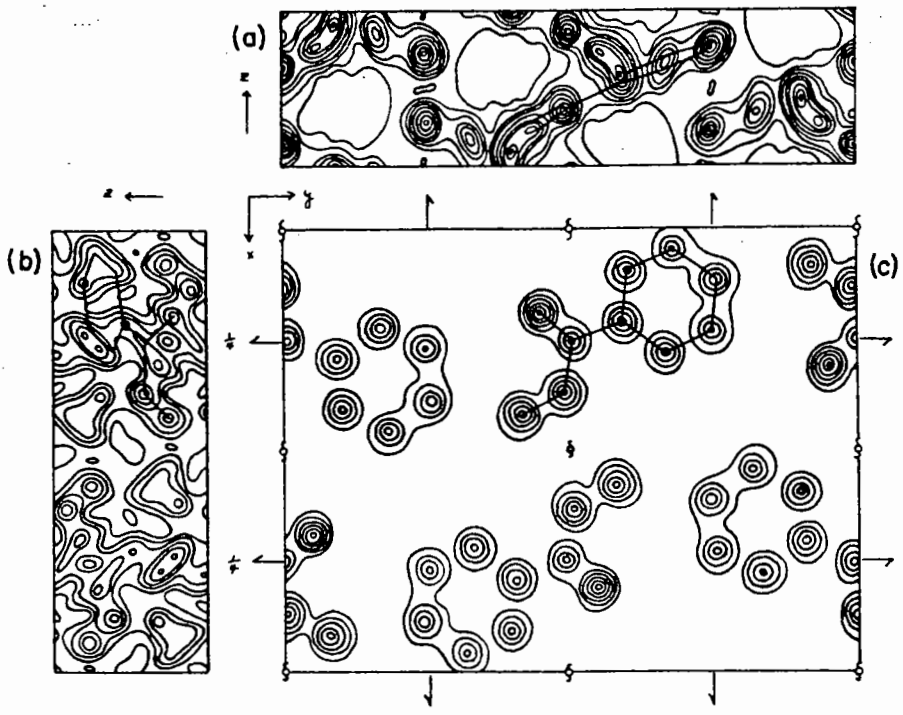


Fig.1. Projection of $\rho(x,y,z)$: (a) on (100), (b) on (010), (c) on (001); contours at even integral values of eÅ^{-2}, zero contour omitted. Reproduced by permission, Am Chem. Soc.

The F_o projections on (010) and (100) for refining the z parameters confronted me with a problem: some of the atoms in the (010) projection were unresolved and the rest only poorly so, Fig. 1b, and the projection on (100) was even worse, all atoms being unresolved in pairs, Fig. 1a. Thus no reliable z coordinates could be determined from F_o maps. An attempt was then made to adjust the approximate z parameters by least squares, but the time required was prohibitive--two refinement cycles for only ten parameters required three months to calculate, in part because a new desk calculator was making random errors due to a slightly bent shaft, and to make matters worse, the parameter shift were unreliable.

In checking my reprint file at the time, I came across the paper by W. Cochran cited above. From the underlining of certain points in the reprint, it was clear that I had read parts of it, now I read it from the standpoint of necessity! Since I had a set of Beevers-Lipson strips, it was easy to calculate an (F_o-F_c) map, the first one being the projection on (001) with all atoms resolved. I was surprised and pleased with the result: additional shift in the x and y coordinates were obvious, differences from the assumed B values were clearly evident as were the hydrogen atoms. Even the shifts in positions for pairs of unresolved atoms in projection on (010) and (100) were readily evaluated by fixing one atom and moving the other by an arbitrary amount, then noting the result in the following (F_o-F_c) map. Refining the isonicotinic acid hydrazide model was completed by three additional maps on (001) and seven each on (010) and (100), the z parameters being the average of the values for each of the atoms from the final projections on (010) and (100). The results were published in the Journal of the American Chemical Society (Jensen, 1954). The power and sensitivity of (F_o-F_c) maps, even for two-dimensional data, was truly impressive, and the resulting bond lengths and angles would be quite acceptable, even by today's standards.

By the mid-1950's, as we moved on to three-dimension data and programming for the electronic computers, all structures determined in my laboratory were refined by (F_o-F_c) maps. I went back to least squares in the early 1960's only after fast computers became available along with the Busing-Levy full matrix least squares program.

Refinement of Proteins

The development of the multiple isomorphous replacement (MIR) method of phase determination in the 1950's led to solving the myoglobin structure, first reported at 6Å resolution (Kendrew et al., 1958). Since data could be collected to 1.4Å resolution, an effort was eventually mounted to refine myoglobin in the best tradition of classical crystal structure determination, initially by F_o maps and by the method of least squares. Despite the fact that the myoglobin data extended to relatively high resolution, many atoms remained unresolved in the F_o maps, and others were only poorly resolved so that it was difficult to determine reliable positional parameters, and little could be done about thermal parameters. The method of least squares was also applied, but storage requirements for a full matrix treatment were beyond the capacity of even the largest computer of the time, leading to the use of the diagonal approximation. Neglect of the off-diagonal terms often leads to unreliable coordinate shifts, particularly in crystals belonging to a non-orthogonal axial system such as the monoclinic myoglobin crystals. D.C. Phillips (1966) has summarized in considerable detail the extensive efforts to refine myoglobin, but it was clear that

only limited progress was achieved, and the idea of refining proteins faded with the passing years.

In the mid-1960's studies were initiated in our laboratory on small metalloproteins, and in July of 1968 one of these, rubredoxin from *C. pasteurianum* (54 aa, M_r 6000; Lovenberg and Sobel, 1965), was solved at 3Å resolutions by use of anomalous scattering to resolve the phase ambiguity of a single HgI_4^{2-} derivative. Subsequently and quite independently the structure was solved by the same method from a UO_2^{2+} derivative. Resolution was first extended to 2.5Å with phases determined from both derivatives, including anomalous scattering (Herriott et al., 1970), then to 2.0Å resolution early in 1970. Although the chemical sequence had not been completed, a tentative sequence could be read from the 2Å MIR phased map.

The idea of refining proteins was never far from my thoughts as the rubredoxin study progressed. Indeed, the experience of many years before in refining unresolved atoms in projection by means of (F_o-F_c) maps strongly suggested their use in an attempt to refine rubredoxin. Accordingly, I estimated coordinates for 401 of the 424 atoms in the protein and 23 water sites. The crystallographic residual $R(=\sum||F_o|-|F_c||/\sum|F_o|)$ for the model with an overall isotropic B of 12Å^2 was 0.372 for the 5027 F_o in the 1.5Å data set ($0.03 < \sin\theta/\lambda < 0.325$). In the first ($F_o$-$F_c$) map, coordinate shifts were immediately obvious and readily estimated from the gradient/curvature for each atom. It was also obvious from the positive electron density in the central part of the molecule and negative density for surface residues that B was less the 12Å^2 well within the molecule and much greater than 12Å^2 at the periphery.

Relevant data for the first three (F_o-F_c) refinement cycles are recorded in Table 1. Note that after the first cycle, the overall isotropic B of 12Å^2 was retained in order to limit the number of

Table 1
Refinement data for rubredoxin, first three cycles
5027 reflections in the range $0.03 < \sin\theta/\lambda < 0.325$

Coordinates	R	No. Protein Atoms	No. Water O Atoms	B(Å)^2
Initial F_c	0.372	401	23	12Å^2
From 1st ΔFmap	0.321	411	22	12
from 2nd ΔFmap	0.289	417	22	*9,12,15,20
From 3rd ΔFmap	0.262	418	22	*7,9,12,15, 20, 25, 30

*By groups of atoms, see text.

added parameters. After the second cycle four groups of isotropic B's (9-20Å^2) were introduced to account for the variation of thermal motion within the molecule, and after the third cycle three additional groups with B values ranging from 7-30Å^2. It was to be expected that B's would vary within the molecule, but the extreme range, 7-30Å^2 was a surprise.

Since I had critically checked details in the first three (F_o-F_c) maps and determined all atomic shifts manually, I was certain that the refinement was genuine. My associate, K.D. Watenpaugh, who was deeply involved in the computational side of crystallography and who had done all of the machine calculation thus far, completed the refinement against the 1.5Å data by a fourth (F_o-F_c) cycle and four block diagonal least squares cycles with a final R=0.126. The early results were first reported at the Ottawa ACA Meeting in the summer of 1970 (Watenpaugh et al., 1970), more fully at the Cold Spring Harbor Symposium in 1971 (Watenpaugh et. al., 1972), and in detail in Acta Crystallographica (Watenpaugh, et al., 1973).

Over the next several years, other investigators refined models of somewhat larger proteins with results similar to those reported for rubredoxin (Carter et al., 1974; Deisenhofer and Steigemann, 1975). In order to reduce the time necessary to refine large proteins, efforts were initiated to automate the procedure, and since only limited data can be collected for most protein crystals, it is essential to supplement the diffraction data with other information, such as stereochemical data, in order to improve the conditioning of the equations. Table 2 lists four widely distributed refinement programs that incorporate such information in addition to the X-ray data. These programs are efficient, have been extensively used and thoroughly checked.

Table 2
Four Widely Distributed Refinement Programs

Program	Comments	Reference
CORELS (COnstrained-REstrained Least-Squares)	Least squares on rigid groups (constrained) while imposing restraints on intergroup distances	Sussman et al., 1977 Sussman, 1985
PROLSQ (PROtein Least SQuares)	Simultaneously minimizes x-ray residuals while imposing numerous stereochemical restraints	Hendrickson, 1981 Hendrickson, 1985 (see also Waser,1963)
EREF	Simultaneously minimizes x-ray residuals and the potential energy	Jack and Levitt, 1978 Deisenhofer et al., 1985
X-PLOR	Molecular dynamics followed by simultaneously minimizing x-ray residuals and the potential energy. With mol. dyn. turned off, it is essentially EREF as programmed Jack and Levitt	Brünger, et al., 1987 Brünger, 1992b

Other programs have been developed that incorporate both real space and reciprocal space restraints such as density modification (including solvent flatening), regions of known structure, histogram matching, individual structure factors in Sayr's equation, non-crystallographic symmetry (Main,

1990; Cowtan and Main, 1993), and protein-like topology (Wilson and Agard, 1993; Baker et al., 1993). Other methods including entropy maximization (Xiang et al., 1993) and maximum likelihood (Read, 1993) have also been proposed to improve protein models.

Refinement and R Values

The Crystallographic R factor defined above has been used since the early days of crystal structure analysis, and it is still the standard index to express the agreement between the observed and calculated structure factors (for discussion see Stout and Jensen, 1989). As an overall index, however, it is insensitive to local error in macromolecular models. Nevertheless, it is useful as an indicator of the quality of the initial model, for following the progress of refinement, and for providing a basis for estimating the accuracy of the final coordinates.

A.J.C. Wilson (1950) has shown that if the requisite number of approximately equal atoms are distributed at random in an acentric unit cell, the most probable value of R is 0.59, but somewhat greater if one or more centric zones are present (e.g. structures conforming to the symmetry of space group $P2_12_12_1$). Thus R for a model derived from an experimentally phased F_o map would be expected to be no more than 0.45-0.50, but models with somewhat higher values have been refined successfully. In the case of rubredoxin, we began from a much lower initial R (0.372) which decreased smoothly as the refinement progressed. In contrast R for most protein refinements decreases more slowly and often holds up, indicating the model has lodged in a false minimum. Examining various kinds of F_o maps (e.g. omit maps which delete selected residues), (F_o-F_c) maps, and ($2F_o$-F_c) maps, then rebuilding bad regions will usually enable the refinement to continue, but for difficult models numerous rebuilding sessions may be required during the course of the refinement.

In the early stages of refinement, one should fix an overall B for the protein in order to minimize the number of parameters (except to correct model errors or omissions) until R reaches ~0.30. At that point one is normally justified in adding individual atom B values to allow for differences in atomic mobility within the protein molecule. If the B's remain physically realistic in the following cycles, one may reasonably add water atoms at the peak positions in the solvent near the protein surface, based on ($2F_o$-F_c) and (F_o-F_c) maps. For good data, R should end up below 0.20, and in favorable cases it may fall below 0.15.

Brünger (1992a) has proposed an R(designated as the free R or R_{free}) that is calculated on a subset of the reflections rather than the whole data set. The reflections in this subset are not included in the data used in the course of the refinement and therefore can exert no driving force to reduce R. Accordingly, R_{free} for these reflections is a more reliable indicator of the progress of the refinement than the conventional R.

Brändén and Jones (1990) have defined a real space residual, $R_{rs}=\sum |\rho_o - \rho_c| / \sum |\rho_o + \rho_c|$ that is sensitive to local errors, but it requires both an F_o and an F_c map. It can be calculated over any desired local region of the map, but in practice its use has been limited to a residue by residue check of models.

Finally, an (F_o-F_c) map calculated at the end of refinement is itself a sensitive indicator of possible errors wherever they may occur in a model. Such maps can simply be assessed visually for significant differences between ρ_o and ρ_c.

Future Possibilities

Since the first protein structures were solved more than 30 years ago, developments in the field have been truly remarkable. Not only has an impressive array of refinement programs been developed to assure our passage from first maps to final models, but the whole field, from growing crystals to collecting data and processing it, has also advanced. Technical developments have revolutionized virtually every aspect of our research efforts, and we can expect such developments to continue.

Synchrotron radiation sources as provided at various national facilities have played a key role in collecting diffraction data, particularly for poorly diffracting crystals of large proteins and viruses. We can expect the upgraded beam lines at existing facilities and the new sources now being built, including improved detectors and the use of shorter wave lengths, to provide data of substantially improved quality.

In recent years procedures have been developed for collecting data from protein crystals maintained at liquid nitrogen temperatures. Although the technique cannot be applied to all crystals, it may be more widely applicable than once thought. The reduction in thermal motion at low temperatures is an obvious advantage, but possibly more important, protein crystals generally suffer far less radiation damage at liquid nitrogen termperature than at room temperature (for review see Hope, 1990).

Very recently techniques have been perfected to flash cool protein crystals and mount them without the traditional capillary and adhering solution. This leads to a reduction in absorption and ensures a more accurate absorption correction. The technical advances cited here all point to the possibility for much improved data which will inevitably lead to better first maps, more trouble-free refinements, and more reliable models.

While looking forward to technical advances, we should critically review present practices. In particular, the general practice of omitting the aqueous continuum from our models and cutting off the low order reflections (d spacings greater than some value in the range 7-10Å) should be reexamined (see S.E.V. Phillips, 1980; Blake et al., 1983). Although the continuum contributes primarily to the low order data, including it in the model will affect the interpretation of Fourier maps in the region of the protein/solution interface.

To test the effects of the solvent continuum, accurate intensities for all low order data for selected proteins should be collected and the continuum included in the model at the expected value based on its composition. The transition region in the solution near the surface of the protein has long been recognized as a difficult region. Here we often see electron density peaks, representing

water molecules hydrogen bonded to the protein, with peak densities as great as those of surface protein atoms. Beyond these peaks, discrete water peaks of lower density often appear and fade into the continuum at increasing distance from the protein surface. How best to merge the discrete peaks into the disordered continuum is still an open question.

In order to determine the discrete water sites, it is advantageous to calculate the Fourier maps [F_o, (F_o-F_c), ($2F_o$-F_c)] on as nearly as possible an absolute scale. This means that the F_o data should be properly scaled and that the F_{∞} and ΔF_{∞} terms must be included in the calculation. It then remains to select the electron density level above which a peak will be accepted as a discrete site. An objective choice can be made in terms of the standard deviation σ in $\rho(x,y,z)$. Thus if σ is realistic, a peak density of 2.6σ means that of one-hundred such peaks, ninety-nine would represent solvent structure. For a reasonably good map, this would be a relatively small peak and presumably would represent a partially occupied water site (Jensen, 1990).

An interesting use of (F_o-F_c) maps has been to model regions of a structure that cannot be readily fit by individual atomic sites. Structure factors are calculated as the sum of part of the structure and the (F_o-F_c) map for the rest of the structure. These F_c's become the input to the next (F_o-F_c) map. Badger and Caspar (1991) have studied the solvent structure in cubic insulin by this method. In twelve refinement cycles, R decreased from ~ 0.20 to 0.06, a dramatic drop. Nevertheless, the method must be viewed with caution, because (F_o-F_c) maps will introduce peaks, both positive and negative, representing experimental errors and deficiencies in the protein model itself, that will persist in the final map (Jensen, 1991).

In preparing this paper, it has been a rewarding experience to review the developments in both the theory and practice of crystal structure determination based on x-ray diffraction data. Indeed, it is the creative ideas of the past that enable us to view in such surprising detail the complexity of protein molecules and to catch a glimpse of how they function. But when I consider the vast array of these incredible molecules, I feel that we have only begun to scratch the surface of what lies before us. In this sense the future possibilities are virtually unlimited.

References

Badger, J. and D.L.D. Caspar (1991) Proc. Natl. Acad. Sci. USA 88, 622-626.

Baker, D., C. Bystroff, R.J. Flettrick and D.A. Agard (1993) Acta Crystallogr. D49, 424-434.

Blake, C.C.F., W.C.A. Pulford, and P.J. Artymiuk (1983) J. Mol. Biol. 167, 693-723.

Bragg, W.L. (1962), in "Fifty Years of X-ray Diffraction", P.P. Ewald, Ed. p. 124, N.V.A. Oosthoek's, Utrecht.

Brändén, C.-I. and T.A. Jones (1990) Nature (London) 343, 687-689.

Brünger, A.T., J. Kuriyan and M. Karplus (1987) Science 235, 458-460.

Brünger, A.T. (1992a) Nature (London) 355, 472-474.

Brünger, A.T. (1992b) X-PLOR Version 3.0, Yale University, New Haven, USA.

Carter, Jr., C.W., J. Kraut, S.T. Freer, Ng.H. Xuong, R.A. Alden, and R.G. Bartsch (1974) J. Biol. Chem. 249, 4212-4225.

Cochran, W. (1951) Acta Crystallogr. 4, 408-411.
Cowtan, K.D. and P. Main (1993) Acta Crystallog. D49, 148-157.
Deisenhofer, J., S.J. Remington, and W. Steigemann (1985), in "Methods of Enzymology", H.W. Wyckoff, C.H. W. Hirs, and S.N. Timasheff, Eds., Vol. 115, pp. 313-323, Academic Press, New York.
Deisenhofer, J. and W. Steigemann (1975) Acta Crystallogr. B31, 238-250.
Herriott, J.R., L.C. Sieker, L.H. Jensen, and W. Lovenberg (1970) J. Mol. Biol. 50, 391-406.
Hendrickson, W.A. (1981), in "Refinement of Protein Structures", P.A. Manchin, J.W. Campbell, and M. Elder, Eds., pp. 1- Daresbury Laboratory, Daresbury, Warrington, England.
Hendrickson, W.A. (1985), in "Methods of Enzymology," H.W. Wyckoff., C.H.W. Hirs, and S.N. Timasheff, Eds., Vol. 115, pp. 252-270, Academic Press, New York.
Hope, H. (1990) Annu. Rev. Biophysics and Biophysical Chem. 19, 107-126.
Hughes, E.W. (1941) J. Am. Chem. Soc. 63, 1737-1752.
Jack, A. and M. Levitt (1978) Acta Crystallogr. A34, 931-935.
Jensen, L.H. (1954) J. Am. Chem. Soc. 76, 4663-4667.
Jensen, L.H. (1990) Acta Crystallogr. B46, 650-653.
Jensen, L.H. (1991) Amer. Cryst. Assn. Abstr. p.35, Univ. of Toledo, Toledo.
Kendrew, J.C., G. Bodo, H.M. Dintzis, G. Parrish, H.Wykoff and D.C. Phillips (1958) Nature 181, 662-666 .
Legendre, A.M. (1806), cited by E. Whittaker and G. Robinson (1944), "The Calculus of Observation", Fourth Ed., p. 210, Blackie & Son, Ltd., London.
Lovenberg, W. and B.E. Sobel. (1965) Proc. Nat. Acad. Sci.(Washington) 54, 193-199.
Luzzati, P.V. (1952) Acta Crystallogr. 5, 802-810.
Main, P. (1990) Acta Crystallogr. A46, 372-377.
Phillips, D.C. (1966), in "Advancement in Structure Research by Diffraction Methods", R. Brill and R. Mason, Eds., Vol. 2, pp. 112-127, Interscience, New York-London.
Phillips, S.E.V. (1980) J. Mol. Biol. 142, 531-554.
Read, R. J. (1993), personal communication.
Stout, G.H. and L.H. Jensen (1989) X-ray Structure Determination, Second Edition, John Wiley & Sons, Inc. New York, pp. 229-230, 317-318, 372-377, 387-389.
Sussman, J.L., S.R. Holbrook, G.M. Church, and S.H. Kim (1977) Acta Crystallogr. A33, 800-804.
Sussman, J.L. (1985), in Methods of Enzymology, H.W. Wyckoff, C.H.W. Hirs, and S.N. Timasheff, Eds., Vol. 115, pp. 271-303, Academic Press, New York.
Waser, J. (1963) Acta Crystallogr. 16, 1091-1094.
Watenpaugh, K.D., L.C. Sieker. J.R. Herriott, and L.H. Jensen (1970) Amer. Cryst. Assn. Abstr., p. 44, Carelton Univ., Ottawa.
Watenpaugh, K.D, L. C. Sieker, J.R. Herriott, and L.H. Jensen (1972), in Cold Spring Harbor Symposium on Quantitative Biology XXXVI, pp. 359-367.
Watenpaugh, K.D., L.C. Sieker, J.R. Herriott ,and L.H. Jensen (1973) Acta Crystallogr. B29, 943-956.
Wilson, A.J.C. (1950) Acta Crystallogr. 3, 397-398.
Wilson, C. and D.A. Agard (1993) Acta Crystallogr. A49, 97-104.
Xiang, S., C.W. Carter, Jr., G. Bricogne, and C.J. Gilmore (1993) Acta Crystallogr. D49, 193-212.

Using the MIRmodelmask procedure to improve map interpretability and reduce model bias in the structure determination of the HIV-1 RT/DNA/Fab complex

Jianping Ding and Edward Arnold

Center for Advanced Biotechnology and Medicine (CABM) and Rutgers University Chemistry Department, 679 Hoes Lane, Piscataway, NJ 08854-5638, USA

Introduction

In protein crystallography, the multiple isomorphous replacement (MIR) method is widely used for the initial phase determination in the solution of new protein structures (Green *et al.*, 1954; Watenpaugh, 1985). However, due to errors in the measurement of diffraction intensities and uncertainty in the determination of heavy-atom parameters, the initial MIR phases usually contain errors, yielding electron density maps of limited accuracy and resolution that may be ambiguous or difficult to interpret in some regions of molecules. Since protein crystals frequently have a high solvent content, the technique of solvent flattening is commonly applied to improve the quality of initial MIR phases. In the conventional solvent flattening procedure, the solvent boundary between protein molecule and solvent is determined by assigning regions with relatively low electron density as solvent regions (Wang, 1985; Fenderson *et al.*, 1990). The density in those solvent regions is set to a constant value. The resulting boundary can be applied as a solvent mask to constrain the refinement of the heavy atom parameters and MIR phases (Rould *et al.*, 1989; Cura *et al.*, 1992; Rould *et al.*, 1992). Iterations of the solvent mask determination, solvent leveling, and phase refinement can significantly improve the phase quality and map interpretability. Nevertheless, the automatic solvent leveling procedure indiscriminately flattens out all regions with lower electron density, including density that may correspond to amino acid side chains and regions of the macromolecule with weaker electron density, which may correspond to conformationally mobile regions. On the other hand, in solving a new protein structure, if the diffraction data for native and heavy atom derivatives have reasonable quality and completeness to a moderate or high resolution, the electron density maps calculated from initial MIR phases are usually sufficient to permit backbone tracing and model building for the majority of the protein molecule. As soon as a partial atomic model is available, the calculated phases from the partial atomic model, in turn, can be combined with MIR phases to improve the accuracy of experimental phases. The electron density maps computed from these combined phases are used for further model building and structure refinement. However, the incorporation of model phases introduces bias from the atomic model. Even though various omit-map calculation techniques have been developed to detect and reduce the model bias in protein structures (Bhat & Cohen, 1984; Hodel *et al.*, 1992), it is difficult to detect and eliminate in the later stages of atomic model refinement if the quality and resolution of diffraction data and phases are limited. This paper reports a procedure for efficiently combining the partial model phases with MIR phases called the "MIRmodelmask" procedure. This procedure has been used successfully to improve the map interpretability and reduce the model bias in the structure determination of HIV-1 reverse transcriptase complexed with a 19-mer/18-mer double-stranded DNA template-primer and an Fab fragment of a monoclonal antibody (HIV-1 RT/DNA/Fab). The structure of the HIV-1 RT/DNA/Fab complex has been reported at 3.0 Å resolution (Jacobo-Molina *et al.*, 1993).

The MIRmodelmask Procedure

Conventional solvent flattening procedures can improve the MIR phase quality but may also lose phase information corresponding to some side chains and the regions of a protein molecule with weak electron density. Combination of MIR phases with those computed from a partial atomic model can enhance the phase information so as to result in improvement of electron density in regions that were initially uninterpretable. However, the input atomic model at the same time biases the density maps and the structure refinement. In order to exploit the phase information derived from a partial atomic model but reduce the model bias, we have developed a

procedure that is outlined in Fig. 1. In the structure determination of the HIV-1 RT/DNA/Fab complex, based on the initial MIR phased electron density maps, we were able to trace many secondary structural elements of the HIV-1 RT p66/p51 heterodimer. The iterative ordinary solvent flattening and phase refinement improved the initial MIR phases and resolved the electron density for the majority of the polypeptide chains. However, the density was still very weak for most of the side chains and the density was poor or discontinuous for some portions of the protein molecule. Therefore, a variety of approaches was tried for combining the initial partial model structure with the MIR phases in order to improve the quality and interpretability of electron density maps. The most successful strategy is the technique that we have designated as the MIRmodelmask procedure. The calculated partial atomic model phases are first combined with the initial MIR phases ("MIRmodel phases"). Iterations of solvent mask determination, solvent leveling, and phase refinement using the combined MIRmodel phases as starting phases lead to generation of a solvent mask which has information from both MIR and atomic model phases ("MIRmodelmask"). The phases corresponding to this initial MIRmodelmask refinement are then used as a constraint to refine the heavy atom parameters and the pure MIR phases. Further phase extension is carried out as usual (see the flowchart in Fig. 1). This procedure yields electron density maps with significantly improved quality in the connecting regions and for the side chains where there were weaker electron densities in the maps calculated from MIR phases with the ordinary average solvent flattening. An iterative process of model building and phase combination using the MIRmodelmask procedure was performed in conjunction with further atomic model refinement. The partial atomic model used in phase combination should include only the atoms of the residues without ambiguity. Depending on the quality of the partial atomic model, the relative weight of the MIR phases and the model phases can been varied. During the solvent mask determination, solvent flattening, and phase refinement, the parameters, for example, solvent content, negative density truncation, and sphere radius, can be varied to achieve better results. Use of Brunger's "free R-factor" for evaluation of the optimal parameters to use for this and other density modification procedures could be particularly advantageous (Brunger, 1992). This procedure was implemented using programs from the PHASES package (Furey & Swaminathan, 1990). In this study the parameters used for solvent mask generation during solvent flattening at 3.0 Å resolution were: solvent fraction = 60-75%, empirical constant S=0.06-0.086, and radius=8 Å (Furey & Swaminathan, 1990).

Background on HIV-1 reverse transcriptase

The reverse transcriptase (RT) of the human immunodeficiency virus type 1 (HIV-1) is the essential enzyme responsible for the catalytic conversion of the single-stranded viral RNA genome into a double-stranded DNA which is integrated into host cell chromosomes. Infection by HIV-1 eventually causes the deadly disease, acquired immunodeficiency syndrome (AIDS). HIV-1 RT is a potential therapeutic target of many inhibitors against AIDS. Nucleoside analog inhibitors, such as 3'-azido-3'deoxythymidine (AZT), dideoxyinosine (ddI), and dideoxycytidine (ddC), are clinically effective drugs and have been widely used for treating HIV-1 infections (De Clercq, 1992; Larder, 1993). Nevertheless, their effectiveness is limited by serious side effects which include toxicity and the rapid emergence of drug-resistant strains of virus (St. Clair *et al.*, 1991; Richman, 1993; Schinazi, 1993). Nonnucleoside inhibitors, for example, the TIBO derivatives (Pauwels *et al.*, 1990), nevirapine (Merluzzi *et al.*, 1990), pyridinones (Goldman *et al.*, 1991), BHAP derivatives (Romero *et al.*, 1991), TSAO derivatives (Balzarini *et al.*, 1992), and α-APA (Pauwels *et al.*, 1993) are highly potent and very specific inhibitors of HIV-1 RT. However, strains of virus that contain mutations and confer drug-resistance also arise rapidly (Nunberg *et al.*, 1991; Mellors *et al.*, 1992; Richman, 1993). Understanding the three-dimensional structure of HIV-1 RT should provide a basis for better understanding of structure-function relationships, mechanisms of drug inhibition, and of antiviral resistance to HIV-1 RT. This understanding could lead to development of new and improved drugs for the treatment of AIDS.

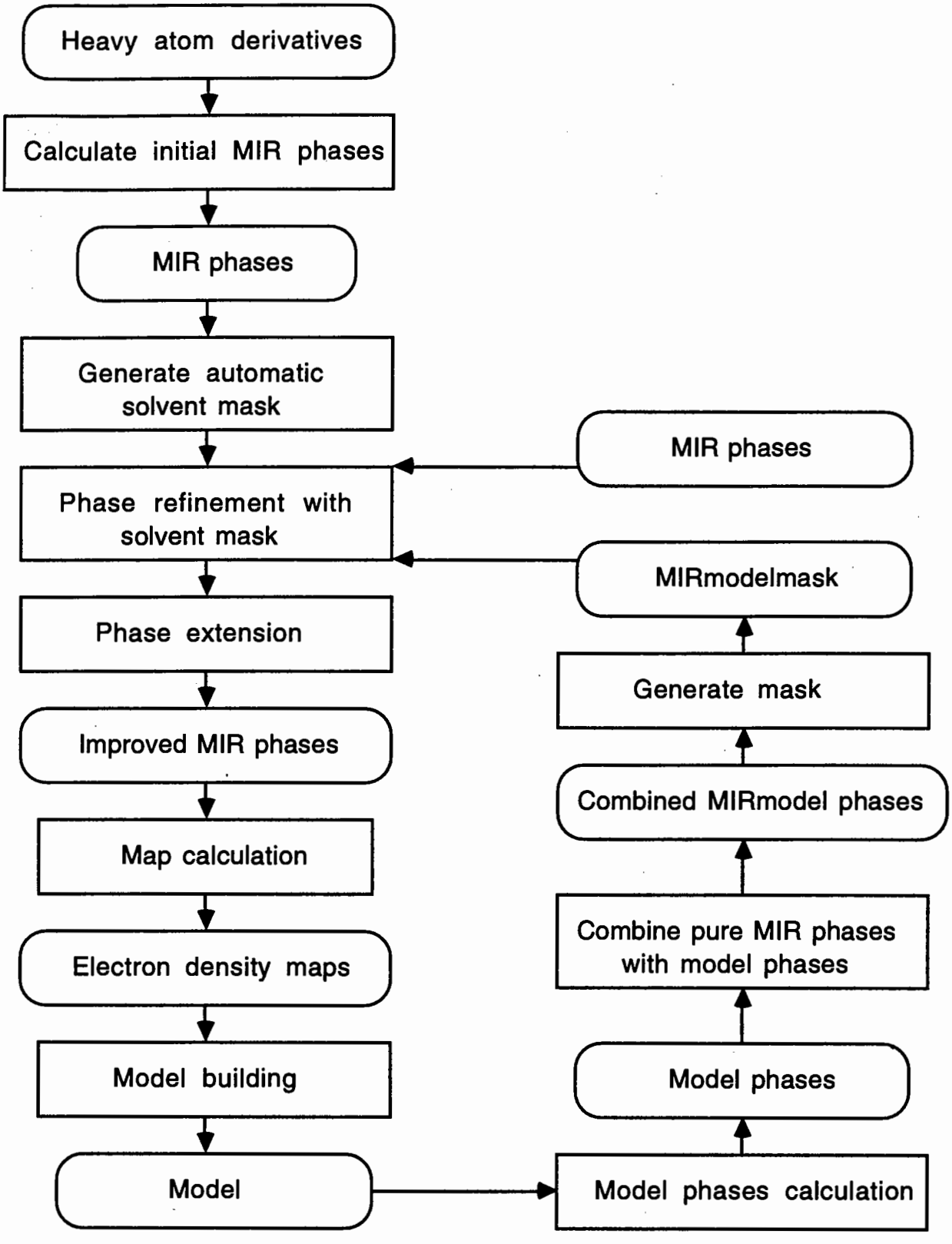


Figure 1. Flow chart for the MIRmodelmask procedure. Details are discussed in the text.

Structure Determination of the HIV-1 RT/DNA/Fab Complex

A variety of experiments were carried out to search and optimize crystallization conditions, including protein engineering aimed at changing the specific amino acids on the surface, and forming complexes with antibody fragments and with synthetic nucleic acids that mimic template-primer substrates. High quality crystals were obtained by cocrystallizing the HIV-1 RT p66/p51 heterodimer, a double-stranded DNA template-primer, and a monoclonal antibody Fab fragment (Fab28). These HIV-1 RT/DNA/Fab crystals diffract X-rays to 2.8 Å resolution at the Cornell High Energy Synchrotron Source (CHESS) (Jacobo-Molina *et al.*, 1991; Arnold *et al.*, 1992; Jacobo-Molina *et al.*, 1993), belong to the space group $P3_212$, and contain one complex per asymmetric unit. The asymmetric unit has a molecular mass of 180 kDa, corresponding to a specific volume of V_M=5.03 $Å^3$/Da (which is within the normal range for proteins (Matthews, 1968)), and a solvent content of 76% (assuming the standard partial specific volume for protein and DNA is 0.74 ml/g).

Low resolution X-ray diffraction data (resolution limits 5.9 to 6.5 Å) were collected using the San Diego Multiwire electronics area detector system (Xuong *et al.*, 1985). The merged data have completeness of around 80% and overall R_{merge} factors of 5-10 % on intensity with $I/\sigma(I)\geq 2$. High resolution X-ray diffraction data were measured at the CHESS F1 beamline using a rotation camera and were processed, merged and scaled together from multiple crystals using a modified version of the Purdue oscillation film processing package (MTOPS, Kamer and Arnold, unpublished). All datasets from both area detector and CHESS were scaled together using local scaling by resolution ranges (Matthews & Czerwinski, 1975) where the CHESS native dataset was used as reference. The native dataset was merged from 186 image plates and/or films exposed from 36 crystals with an R_{merge} (I) = 0.13, and has 76908 unique reflections ($I\geq 3\sigma(I)$) to 2.8 Å resolution with a completeness of 88.1%.

Three-dimensional solutions for heavy-atom isomorphous difference Patterson syntheses were determined using both minimum and sum functions in the program VMAP (Arnold *et al.*, 1992) (Williams and Arnold, unpublished). Cross-phased difference Fourier syntheses clearly verified the major sites and revealed additional sites that were initially confirmed using cross-vector searches with known sites as input into VMAP. The relative occupancy parameters were on an arbitrary scale factor. The heavy atom scattering contributions were calculated from appropriate single atom scattering factors even for the heavy-atom cluster derivatives and a constant isotropic thermal parameter (B=20 $Å^2$) was used during heavy-atom refinement.

The initial atomic model was built using program O (Jones *et al.*, 1991) and based on a series of electron density maps calculated from the MIR phases with resolution limits between 3.5 and 4.0 Å. These MIR phased maps revealed most of the secondary structural elements of HIV-1 RT. Assignment of some of the connection loops was facilitated by the reported structure of HIV-1 RT complexed with nevirapine (Kohlstaedt *et al.*, 1992). RNase H modeling was initially based on the structure of the free HIV-1 RNase H domain (Davies II *et al.*, 1991), with subsequent adjustments of loops and side chain orientations. The electron density was prominent for the sugar-phosphate backbone of the DNA. There was good electron density corresponding to the primer terminus and the one base adenine overhang of the template strand. At the blunt end of the duplex, the electron density for the last two base pairs was relatively poor. Interpretation of the Fab fragment structure was initially guided by matching the four β-barrel domains of McPC603 (Satow *et al.*, 1986) with the electron density. Further side chain adjustment and residue assignment were fulfilled based on electron density maps and the amino acid sequence of Fab28 (Ferris *et al.*, unpublished). Calculation of MIR phases was extended to 3.0 Å resolution and the conventional solvent flattening procedure improved the MIR phase quality and map interpretability. The 3.0 Å MIR solvent flattened phased map further resolved the backbone structure and revealed some side chain information (Fig. 2a). Nevertheless, the electron density was still poor for some portions of the molecule, especially the β3-β4 connecting loop, β11b to β13 of p66, β11 to β14 of p51,

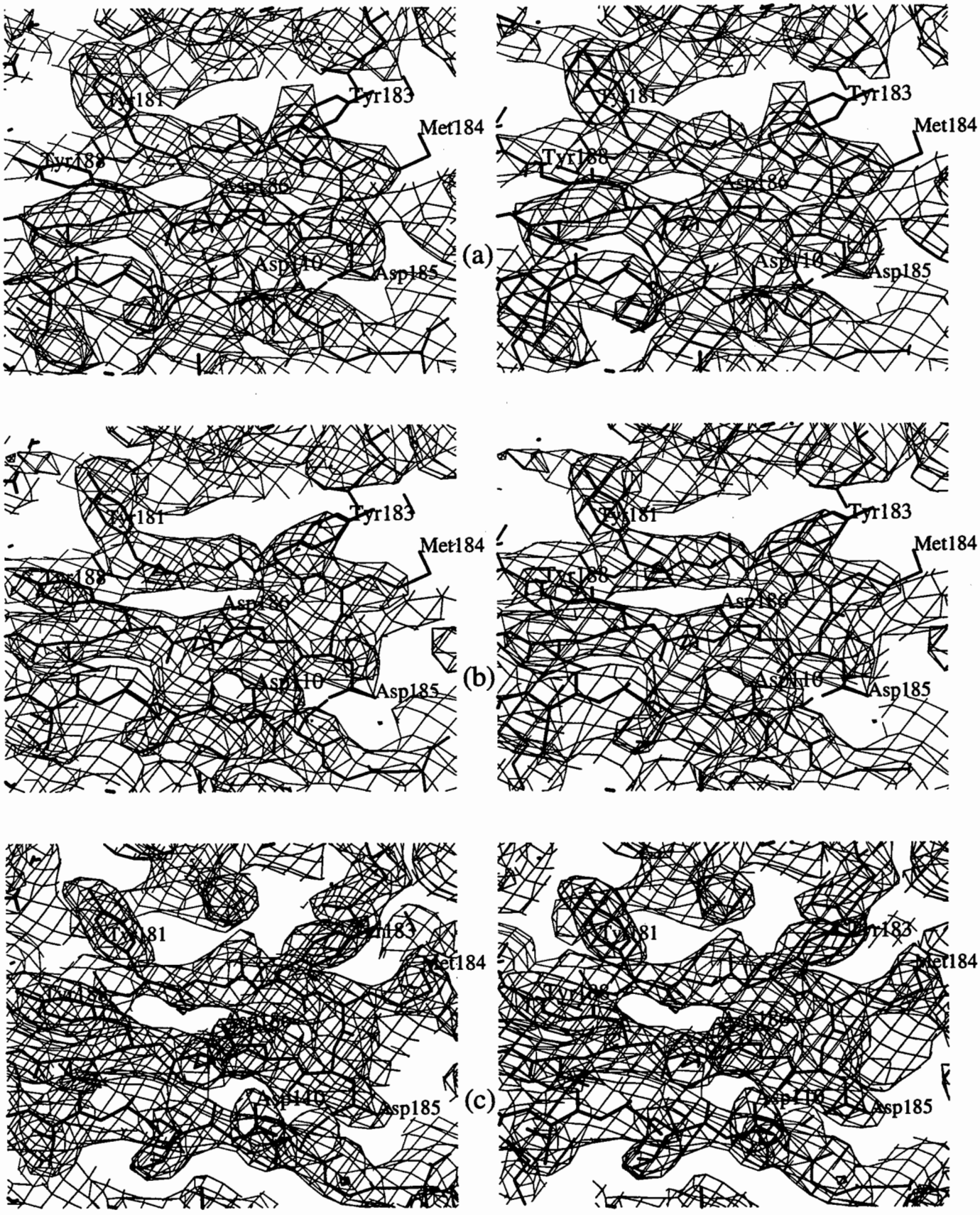

Tyr183
Met184
Asp185
(a)
(b)
(c)

Figure 2. Refined atomic model in a portion of electron density maps in the region near the polymerase active site showing the effects of various combination of the MIR phases with partial atomic model and the improvement of phases and map interpretability at different stages of structure determination. All maps were calculated using the same glycerol-soaked native data collected at CHESS, but with different phases, and were contoured at 1.0σ. (a) 3.0 Å MIR-phased map: the MIR phases were calculated from 12 glycerol-soaked heavy-atom derivatives and refined using 16 cycles of conventional solvent flattening and 12 cycles of phase extension. (b) 3.0 Å map: MIR phases were calculated from 12 glycerol-soaked heavy-atom derivatives and were refined using the MIRmodelmask procedure. The atomic model phases input for generating the solvent mask corresponded to a partial polyalanine model (6095 non-hydrogen atoms from protein and 674 non-hydrogen atoms from DNA) that was extracted from a refined partial model at 3.5 Å resolution (R=0.31). (c) 2.8 Å $2F_O$-F_C map: phases were calculated from a full atomic model (R=0.25 with 12180 non-hydrogen atoms).

and part of the thumb subdomains in both p66 and p51. In addition, correct positioning of the amino acid sequence into the electron density maps was difficult since the density corresponding to many side chains was missing. To make use of the initial partial atomic model, we tried different approaches to combine the atomic model phases with the MIR phases. Using the conventional phase combination method, the resulting maps exhibited clear indications of model bias. Wherever the atomic model was input, there was electron density that corresponded exactly to the input model. A variety of experiments indicated that the MIRmodelmask procedure is a very useful and successful technique to improve the map interpretability and reduce the model bias in the combination of MIR phases with partial atomic model phases. Fig. 2b is a 3.0 Å resolution electron density map calculated using MIR phases that were refined with the MIRmodelmask procedure. The atomic structure combined here was a partial polyalanine model that was extracted from a model that had been partly refined at 3.5 Å resolution. It can be clearly seen that the map quality is improved, and especially that the density is enhanced for many side chains. For example, the density for the side chains of Tyr183 and Tyr188 is evident in Fig. 2b (in comparison with the density map shown in Fig. 2a, which was calculated using the same MIR phases but refined otherwise with the conventional solvent flattening procedure). In the regions where no atomic model was introduced, moderate but obvious density was revealed at the two ends of the polypeptide chains and usually had a size corresponding to 1 to 2 residues. Further experiments were carried out using model phases calculated from partial atomic structures containing both main chains and side chains to combine with MIR phases using the MIRmodelmask procedure. In some places the resulting map showed side chain densities in different orientations from that of the side chains which were input. This is a good indication that the combined phases were less biased by the input atomic model than were phases generated using more conventional phase combination approaches. Iterations of model building, which included progressive side chain placement, residue insertion, and backbone adjustment, phase combination using MIRmodelmask procedure, and structure refinement enabled us to locate electron density for almost all amino acid residues and side chains that permitted us to complete the residue assignment in most of the p66/p51 heterodimer with confidence. Further structure refinement was performed using program X-PLOR (Brunger, 1993). The refinement of a full atomic model including 12180 non-hydrogen atoms has converged to an R-factor of 0.25 for 73,041 reflections (82.6% completeness to 2.8 Å) and a free R-factor of 0.38 (Brunger, 1992) for 3,867 reflections (4.4% completeness to 2.8 Å) between 15.0 and 2.8 Å resolution with $F \geq 3\sigma(F)$. A portion of the current $2F_O$-F_C electron density map in the region around the polymerase active site is also shown in Fig. 2c.

Brief description of the HIV-1 RT/DNA/Fab structure

The HIV-1 RT enzyme functions biologically in the form of a heterodimer consisting of two subunits of 66 kDa (p66) and 51 kDa (p51). The p66 subunit contains a polymerase domain that is composed of the N-terminal 440 amino acids and an RNase H domain that is composed of the

C-terminal 120 amino acids. The p51 subunit corresponds to the polymerase domain of the p66 subunit (reviewed in (Goff, 1990; Jacobo-Molina & Arnold, 1991)). The crystal structure of HIV-1 RT complexed with the nonnucleoside inhibitor nevirapine at 3.5 Å resolution (Kohlstaedt *et al.*, 1992) revealed the overall folding of the HIV-1 RT p66/p51 heterodimer, but was not able to identify the location of the side chains and assign the position of individual amino acid residues. The crystal structure of HIV-1 RT/DNA/Fab complex at 3.0 Å resolution (Arnold *et al.*, 1992; Jacobo-Molina *et al.*, 1993) not only described the amino acid residue assignment throughout the polypeptide chains but also elucidated numerous interesting features of the interactions between HIV-1 RT and the bound template-primer. In the crystal structure of HIV-1 RT/DNA/Fab, both polymerase domains of the p66 and p51 subunits are folded into four subdomains, named fingers, palm, thumb, and connection (Kohlstaedt *et al.*, 1992). Even though the overall folding of individual subdomains is similar, however, their spatial arrangements within the respective subunits are dramatically different, resembling that reported by Kohlstaedt *et al.* (1992). The structure of the template-primer is a hybrid which conforms to A-form DNA near the polymerase active site and B-form DNA towards the RNase H active site, with a significant bend of about 40-45° at the A-form/B-form junction that occurs in the vicinity of the αH and αI helices of the p66 thumb. The most numerous interactions between protein and nucleic acids occur with the amino acid residues of the p66 palm and thumb subdomains and the sugar-phosphate backbone of DNA. Highly conserved regions in the p66 palm near the polymerase active site include a β-hairpin (β12-β13 hairpin) that interacts with the primer strand (designated as the "primer grip") and a loop (β8-αE loop) that interacts with the template strand ("template grip"). These structural elements, together with two α-helices of the p66 thumb, act as a clamp to position the template-primer precisely relative to the polymerase active site. The 3'-hydroxyl of the primer terminus is close to the catalytically essential Asp110, Asp185, and Asp186 residues at the active site and is in a position for nucleophilic attack on the α-phosphate of an incoming nucleoside triphosphate. The structure of the HIV-1 RT/DNA/Fab complex described in Jacobo-Molina *et al.* (1993) is the first detailed structure of any polymerase with nucleic acid bound in a mode relevant for DNA polymerization.

Acknowledgments

We are very grateful to the other members of our laboratory who contributed to the structure determination of the HIV-1 RT/DNA/Fab complex, especially Arthur D. Clark, Jr., Alfredo Jacobo-Molina, Xiaode Lu, Raymond G. Nanni, Chris Tantillo, and Roger Williams. We also thank Stephen Hughes and the members of his laboratory for their collaborative efforts that made this work possible, Roger Williams and Birgit M. Roy for critical reading of the manuscript, and the staffs at CHESS and the NIH-NCI Cray Supercomputer Facility for access to crucial data collection and computational resources, and NIH, CABM, and Janssen Research Foundation for financial support.

References

Arnold, E., Jacobo-Molina, A., Nanni, R. G., Williams, R. L., Lu, X., Ding, J., Clark, A. D., Jr., Zhang, A., Ferris, A. L., Clark, P., Hizi, A. & Hughes, S. H. (1992). *Nature*, **357**, 85-89.

Balzarini, J., Perez-Perez, M.-J., San-Felix, A., Schols, D., Perno, C.-F., Vandamme, A.-M., Camarasa, M. J. & De Clercq, E. (1992). *Proc. Natl. Acad. Sci. USA*, **89**, 4392-4396.

Bhat, T. N. & Cohen, G. H. (1984). *J. Appl. Cryst.*, **17**, 244-248.

Brunger, A. T. (1992). *Nature*, **355**, 472-475.

Brunger, A. T. (1993). *X-PLOR: A system for X-ray crystallography and NMR. Version 3.1*, Yale University Press, New Haven and London.

Cura, V., Krishnaswamy, S. & Podjarny, A. D. (1992). *Acta Cryst.*, **A48**, 756-764.

Davies II, J. F., Hostomska, Z., Hostomsky, Z., Jordan, S. R. & Matthews, D. A. (1991). *Science*, **252**, 88-95.

De Clercq, E. (1992). *AIDS Res. Human Retroviruses*, **8**, 119-134.

Fenderson, F. F., Herriott, J. R. & Adman, E. T. (1990). *J. Appl. Cryst.*, **23**, 115-131.

Furey, W. & Swaminathan, S. (1990). *Am. Crystallogr. Assoc. Mtg. Abstr., Ser. 2*, **18**, 73.

Goff, S. P. (1990). *J. of Acquired Immune Deficiency Syndromes*, **3**, 817-831.

Goldman, M. E., Nunberg, J. H., O'Brien, J. A., Quintero, J. C., Schleif, W. A., Freund, K. F., Gaul, S. L., Saari, W. S., Wai, J. S., Hoffman, J. M., Anderson, P. S., Hupe, D. J., Emini, E. A. & Stern, A. M. (1991). *Proc. Natl. Acad. Sci. USA*, **88**, 6863-6867.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *Proc. R. Soc. London*, **Ser. A. 225**, 287-307.

Hodel, A., Kim, S.-H. & Brunger, A. T. (1992). *Acta Cryst.*, **A48**, 851-858.

Jacobo-Molina, A. & Arnold, E. (1991). *Biochemistry*, **30**, 6351-6361.

Jacobo-Molina, A., Clark, A. D., Jr., Williams, R. L., Nanni, R. G., Clark, P., Ferris, A. L., Hughes, S. H. & Arnold, E. (1991). *Proc. Natl. Acad. Sci. USA*, **88**, 10895-10899.

Jacobo-Molina, A., Ding, J., Nanni, R. G., Clark, A. D., Jr., Lu, X., Tantillo, C., Williams, R. L., Kamer, G., Ferris, A. L., Clark, P., Hizi, A., Hughes, S. H. & Arnold, E. (1993). *Proc. Natl. Acad. Sci. USA*, **90**, 6320-6324.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Crystallogr., Sect. A*, **A47**, 110-119.

Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A. & Steitz, T. A. (1992). *Science*, **256**, 1783-1790.

Larder, B. A. (1993). In *Reverse Transcriptase* (A. M. Skalka & S. P. Goff, ed.), pp. 205-222, Cold Spring Harbor Laboratory Press, Plainview, New York.

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

Matthews, B. W. & Czerwinski, E. W. (1975). *Acta Crystallogr., Sect. A*, **A31**, 480-487.

Mellors, J. W., Dutschman, G. E., Im, G.-J., Tramontano, E., Winkler, S. R. & Cheng, Y.-C. (1992). *Antiviral Research*, **17 S1**, 48.

Merluzzi, V. J., Hargrave, K. D., Labadia, M., Grozinger, K., Skoog, M., Wu, J. C., Shih, C.-K., Eckner, K., Hattox, S., Adams, J., Rosenthal, A. S., Faanes, R., Eckner, R. J., Koup, R. A. & Sullivan, J. L. (1990). *Science*, **250**, 1411-1413.

Nunberg, J. H., Schleif, W. A., Boots, E. J., O'Brien, J. A., Quintero, J. C., Hoffman, J. M., Emini, E. A. & Goldman, M. E. (1991). *J. Virol.*, **65**, 4887-4892.

Pauwels, R., Andries, K., Debyser, Z., Van Dale, P., Schols, D., Stoffels, P., De Vreese, K., Woestenborghs, R., Vandamme, A.-M., Janssen, C. G. M., Anne, J., Cauwenbergh, G., Desmyter, J., Heykants, J., Janssen, M. A. C., De Clercq, E. & Janssen, P. A. J. (1993). *Proc. Natl. Acad. Sci. USA*, **90**, 1711-1715.

Pauwels, R., Andries, K., Desmyter, J., Schols, D., Kukla, M. J., Breslin, H. J., Raeymaeckers, A., Van Gelder, J., Woestenborghs, R., Heykants, J., Schellekens, K., Janssen, M. A. C., De Clercq, E. & Janssen, P. A. J. (1990). *Nature*, **343**, 470-474.

Richman, D. D. (1993). *Antimicrob. Agents and Chemother.*, **37**, 1207-1213.

Romero, D. L., Busso, M., Tan, C.-K., Reusser, F., Palmer, J. R., Poppe, S. M., Aristoff, P. A., Downey, K. M., So, A. G., Resnick, L. & Tarpley, W. G. (1991). *Proc. Natl. Acad. Sci. U.S.A.*, **88**, 8806-8810.

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

Rould, M. A., Perona, J. J. & Steitz, T. A. (1992). *Acta Cryst.*, **A48**, 751-756.

Satow, Y., Cohen, G. H., Padlan, E. A. & Davies, D. R. (1986). *J. Mol. Biol.*, **190**, 593-604.

Schinazi, R. F. (1993). *Perspectives in Drug Discovery and Design*, **1**, 151-180.

St. Clair, M. H., Martin, J. L., Tudor-Williams, G., Bach, M. C., Vavro, C. L., King, D. M., Kellam, P., Kemp, S. D. & Larder, B. A. (1991). *Science*, **253**, 1557-1559.

Wang, B.-C. (1985). In *Methods Enzymol.* (Wyckoff, H.W., Hirs, C.H.W & Timasheff, S.N., ed.), **115**, pp. 90-112,

Watenpaugh, K. D. (1985). In *Methods Enzymol.* (Wyckoff, H.W., Hirs, C.H.W & Timasheff, S.N., ed.), **115**, pp. 3-14.

Xuong, N. H., Nielsen, C., Hamlin, R. & Anderson, D. (1985). *J. Appl. Crystallogr.*, **181**, 342-350.

Determination of the structure of *Bacillus subtilis* pectate lyase

Richard Pickersgill, Gillian Harris and John Jenkins

AFRC Institute of Food Research, Reading Laboratory,
Protein Engineering Department,
Reading, RG6 2EF, UK

1. INTRODUCTION

1.1 Pectate lyase

Pectate lyases (EC 4.2.2.2) are secreted by plant pathogens and cleave polygalacturonate, a major component of the plant cell wall. Pectate lyases are distinguished from the homologous pectin lyases by their specificity for demethylated pectin and their requirement for calcium for activity. The enzyme cleaves the α-glycosidic bond between O1 and C4 of α-1,4 linked galacturonsyl-residues by a β-elimination reaction that results in an unsaturated C4-C5 bond[1]. The *Bacillus subtilis* pectate lyase is a monomer of molecular weight 43,505 daltons and 399 amino-acid residues[2].

1.2 From first map to final model

The first electron density map of pectate lyase was calculated with protein phases calculated using eight heavy atom derivatives. This map showed that pectate lyase is a β-protein but the map could not be interpreted in terms of contiguous polypeptide chain. A striking improvement in the quality of the electron density map was achieved by using the program SQUASH to improve the phases. The all parallel β-sheet architecture of pectate lyase could be clearly seen in this map. The first model of this β-sheet domain was built with the newly determined sequence but three long aromatic-rich loops were not clearly defined in this map. The conformation of these long loops was defined by refining the model against high resolution data from a crystal soaked in calcium using the automated refinement procedure (ARP). The map from ARP was of exceptional quality, the active site calcium was clear and the conformation of the long loops could be defined. The final model has a crystallographic R-factor of 0.15 for all data in the range 10.0 through 1.8 Å.

2. THE FIRST ELECTRON DENSITY MAP

2.1 Crystallization

Crystals were grown in the absence of calcium as described previously[3]. Three types of crystal could be grown; types I and II diffracted X-rays well. Type I crystals belong to spacegroup $P2_1$ with a = 132.9 Å, b = 41.2 Å, c = 156.8 Å and β = 114.9°. These crystals have four molecules in the asymmetric unit. Type II crystals belong to spacegroup $P2_1$ with a = 55.0 Å, b = 88.1 Å, c = 50.2 Å and β = 109.0°. Type II crystals, which could be grown from 2-methyl-2,4-pentandiol

(MPD) or from polyethylene glycol (PEG) in the pH range 6.50 to 8.75, have one molecule in the asymmetric unit. Interpretation of difference Pattersons would be simpler for type II crystals and they were selected for the work described here. In total approximately 15 mg of crystallizable protein were used to determine the structure at 1.8Å.

2.2 Heavy atom derivatives

Both MPD and PEG grown crystals were used for the heavy atom soaks. Eight heavy atom reagents were found to give suitable changes in the structure factor amplitudes and the soak conditions are given Table 1.

Table 1 The derivatives used to calculate the first pectate lyase map

Data set	Soak conditions	Resolution (Å)	Rsym(I) (%)	mfid (%)	Phasing power
Native		2.35	7.2		
$SmCl_3$	1 mM, 12h	3.3	7.8	11.9	0.8
$SmAc_3$	25 mM, 5h	2.6	6.5	13.1	1.2
K_2PtCl_4	15 mM, 60h	3.6	10.1	23.6	0.8
PEDC	2 mM, 69d	3.6	6.2	20.0	1.1
$PbAc_2$	15 mM, 5h	3.6	5.5	16.2	1.2
TAMM	1 mM, 1.5h	3.6	7.3	17.0	0.6
Na_2IrCl_6*	30 mM, 3d	3.6	6.2	17.7	1.0
UO_2Ac_2*	16 mM, 1.5h	2.7	5.5	23.5	1.1

Heavy atom derivatives prepared using crystals grown from PEG are indicated by *, all other were derivatives prepared using crystals grown from MPD. PEDC is platinum ethylenediamine dichloride and TAMM is tetrakis acetoxymercury methane. Soak times are given in hours (h) or days (d).

Heavy atom data were collected using the XENTRONICS area detector and processed using XENGEN[4]. The difference Patterson was solved for the platinum ethylene diamine dichloride derivative. The other heavy atom positions were obtained from difference Fourier maps calculated with the platinum isomorphous and anomalous scattering phases. The derivatives fall into two families. The samarium, lead and uranyl derivatives share a principal site formed by aspartates 184, 223 and 227. The platinum, mercury and iridium derivatives share a principal site which is histidine 374. The phasing power is that output by the program by MLPHARE[5], unless stated the CCP4 program suite[6] was used for all calculations.

2.3 The 3.0 Å electron density map

The overall figure of merit of the mir map was 0.53 for data to 3.0 Å resolution. This map showed a more electron dense region corresponding to protein and a less dense region corresponding to solvent and revealed tubes of electron density with spacing consistent

with β-sheet (Fig. 1). However, this first map gave no clear indication of how the β-strands are connected.

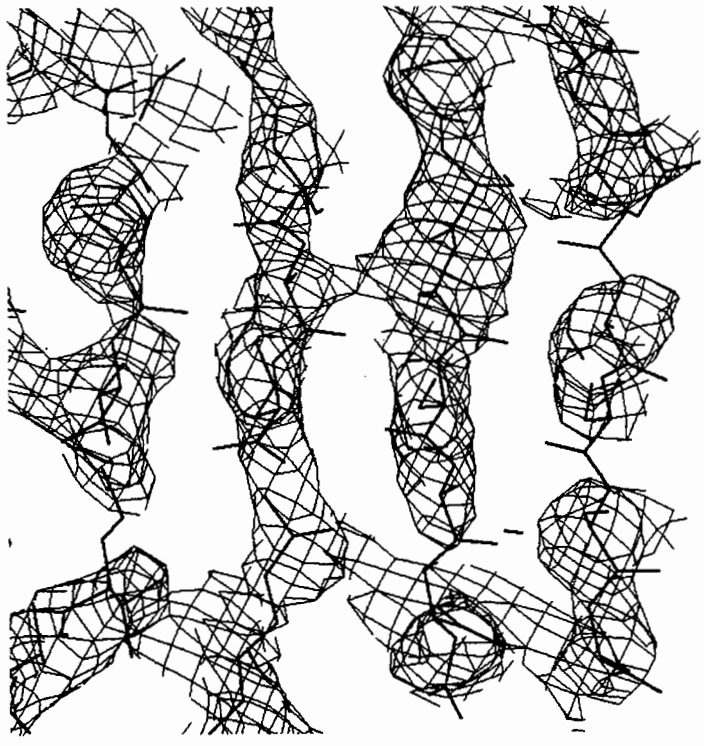


Figure 1 The first map of *Bacillus subtilis* pectate lyase calculated using heavy atom data to 3.0Å resolution shows tubes of electron density separated by 4.5 Å which indicates that this is a β-sheet, however the connectivity within the β-strands is poor. The map is contoured at 1σ.

3. PHASE IMPROVEMENT AND EXTENSION

3.1 Solvent flattening

Solvent flattening is now a standard method of improving protein phases. It can break the phase ambiguity in single isomorphous replacement and can critically improve multiple isomorphous replacement phases. We chose to use the program SQUASH[7]. We changed the input parameters and looked at the quality of the output electron density map which was calculated using combined phases. The quality of the electron density map was evaluated by examining the continuity of the β-strands and the connections between them and by checking that the heavy atoms were located at the protein/solvent boundary.

3.2 A striking improvement in map quality at 3.0 Å

A striking improvement in the quality of the electron density map was achieved when phases were refined at 4.0 Å before extension to 3.0 Å. An initial polyalanine model was built into the solvent-flattened map using the program O aided by the bones and lego options[8]. The main-chain connectivity was clear for 25% of the polypeptide chain and the direction of the chain was established by examining the direction of the side-chain density of the central two residues of a β-turn. The longest contiguous stretch of main-chain was for the carboxy-terminal part of the β-sheet domain and sequence guesses were made for this region. Tryptophan 310 and tyrosines 273 and 295

were clear in the map and were useful initial markers for the sequence fitting. Several other stretches of contiguous chain could be built but sequence could not be unambiguously assigned and these were built as polyalanine. The improvement in the protein phases can be seen by comparing the heavy atom and SQUASH phases with those calculated from the final model (Table 2) and by comparing the heavy atom (Fig. 1) and solvent-flattened maps (Fig. 2).

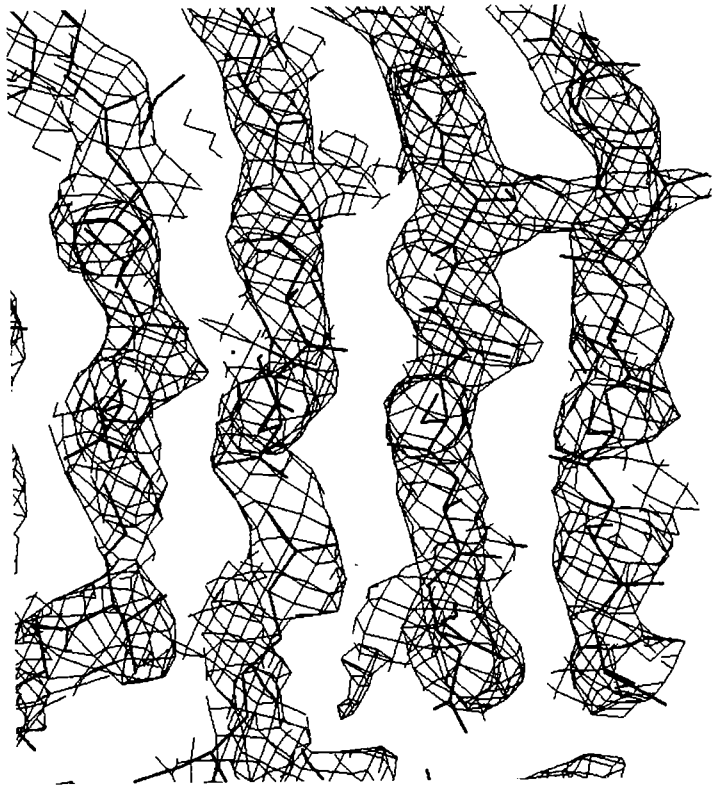


Figure 2 The solvent-flattened map calculated at 3.0 Å resolution using combined phases from SQUASH. This map has the same limits as the first map and is also contoured at 1σ. All four β–strands shown have improved connectivity compared to the heavy atom map, see Fig. 1.

The improvement in the quality of the solvent flattened map compared to the heavy atom map is also illustrated in Fig. 3.

Table 2 The SQUASH phases are closer to the final phases than are the heavy atom phases

	4.0Å	3.0Å
$< \lvert \alpha_{hatom} - \alpha_{final} \rvert >$	47°	61°
$< \lvert \alpha_{squash} - \alpha_{final} \rvert >$	38°	50°

3.3 Some practical suggestions for using SQUASH

(a) Use as many options as possible: solvent flattening, histogram matching, non-crystallographic symmetry and Sayre's equation.
(b) Refine phases to a resolution at which they are reasonable (mean figure of merit of 0.6 say), then extend the resolution to the limit of the heavy atom data.
(c) Include the low resolution terms for solvent flattening or the protein/solvent boundary will be poorly defined.
(d) Do not overestimate the solvent content of the crystal or the protein will be truncated, this is important if you are improving sir or mir phases but less important if you are improving molecular replacement phases since you know the connectivity, but try to use as much solvent as possible or the power of solvent-flattening will be

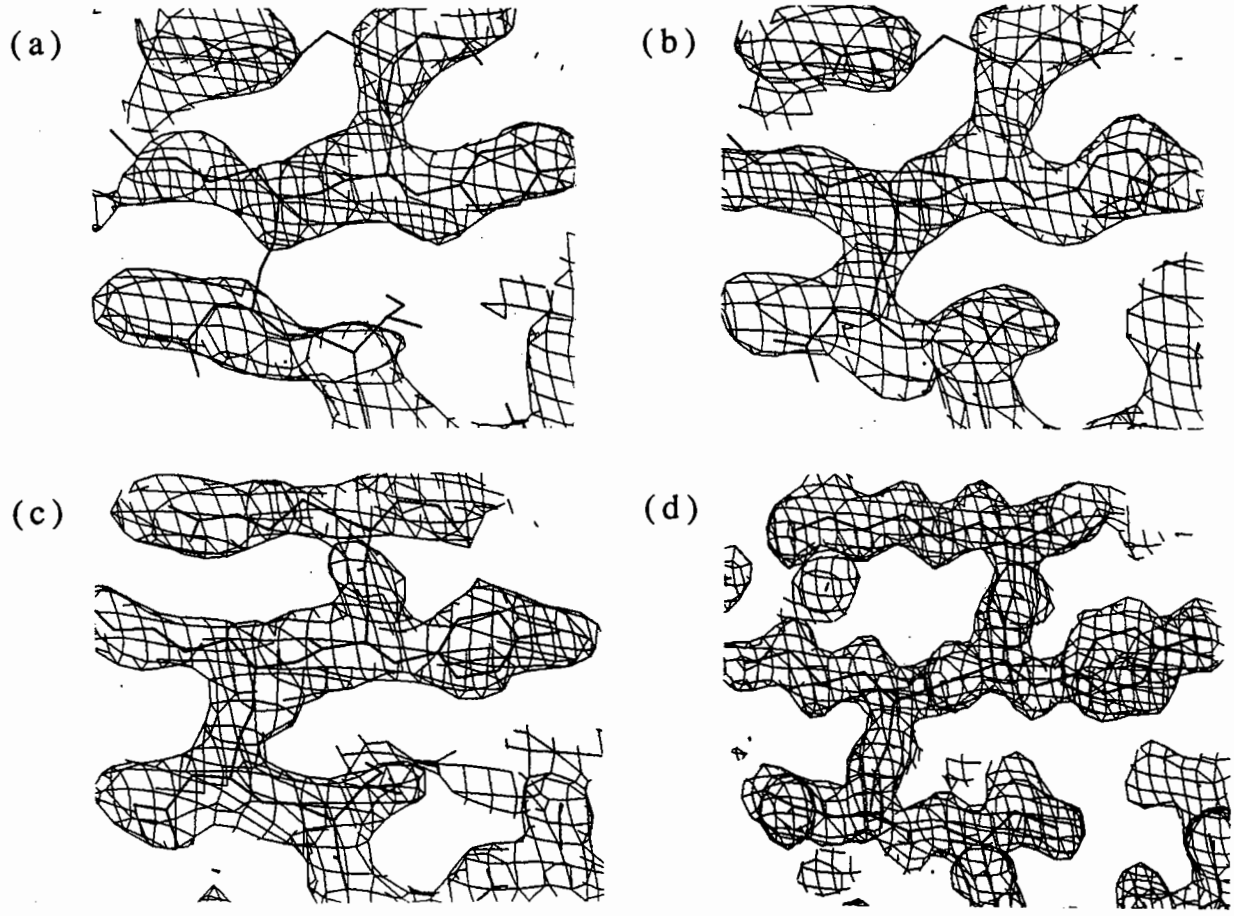


Figure 3 Key steps in the determination of the structure of pectate lyase (a) The first map calculated at 3.0 Å using heavy atom data (b) Solvent-flattened 3.0 Å map showing improved connectivity (c) Map calculated after cycle 7 of XPLOR simulated annealing refinement R = 0.31 at 2.5 Å (d) Final map after ARP refinement at 1.8 Å. Coordinates shown are intermediate for (a), (b) and (c) with R = 0.31 and final for (d) with R = 0.15. All maps are contoured at 1σ.

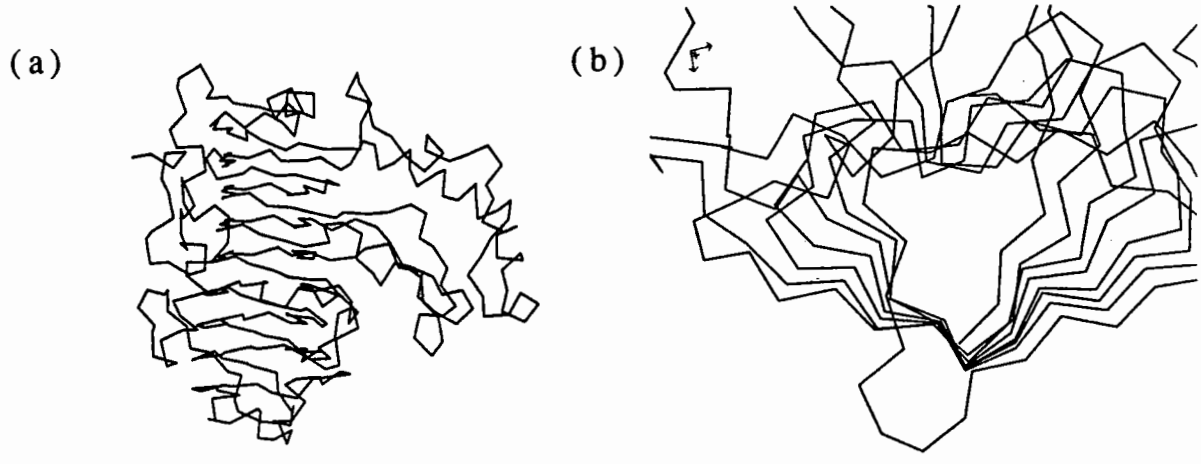


Figure 4 Alpha carbon representation of pectate lyase (a) showing the parallel β-helix and loop domains, (b) view down the axis of the parallel β-helix domain showing the unparalleled regularity of the structure.

reduced.
(e) Check the combined map for protein features, heavy atom location and against other prior knowledge.
(f) SQUASH is easy to use but not robust.
(g) As for all solvent flattening procedures the figure of merit output is inflated (the figure of merit output for pectate lyase was 0.85 when the true value was about 0.64).

4. MAP INTERPRETATION AND SIMULATED ANNEALING REFINEMENT

4.1 Some practical advice on building protein models

The most efficient way of building a protein model of a new structure is to build to a skeletonised electron density map using the model building tools bones and lego in program O[8] and then use a refinement program to sort out the stereochemistry. Simulated annealing refinement using XPLOR[9] has a larger radius of convergence than least squares refinement procedures and is the current method of choice. Simulated annealing in XPLOR assists but does not replace manual rebuilding. For instance, simulated annealing will not fix out-of-register errors.

Cycles of phase improvement involving building a partial model to the experimental or solvent-flattened map, calculating model phases and then combining the model phases with the experimental phases is now routine but care must be taken to avoid model bias. Phase combination should be used where possible to reduce model bias. We found σ_A weighted[10] 3-2 or 2-1 maps useful in reducing bias. Checks for bias include checking the model against a non-model-biased map, either the experimental or solvent-flattened map, and checking against prior knowledge such as the location of heavy atom sites and the distribution of side-chains.

4.2 XPLOR

Seven cycles of simulated annealing and model building to $3F_{obs} - 2F_{calc}\ exp(i\alpha_{calc})$ maps reduced the R-factor from 0.47 at 4.0 Å to 0.31 at 2.5 Å. Further rebuilding and simulated annealing failed to reduce the R-factor. At this stage 66% of the structure, the β-sheet domain, had reasonable density in the 3-2 map (Fig. 3c) but the three aromatic-rich long loops were poorly defined. These loops are of length 68, 25 and 23 amino-acid residues and the worst region was between the parallel β-helix domain and the domain formed by these loops. The definition of the amino and carboxy terminal extensions, of approximately 30 residues each, was intermediate.

At this point publication of the structure of the distantly homologous, 29% identity, pectate lyase PelC from Erwinia reassured us of the correctness of overall connectivity but could not help with structure of the loops since these were much shorter in PelC and had a significantly different structure. The term parallel β-helix was introduced to describe the coiled β-sheet supersecondary structure seen in PelC[11,12]. This should not be shortened to β-helix to avoid

confusion with structures of gramicidins.

Maps calculated using the combined phase gave no clear indication of the conformation of the loops, neither did omit maps nor simulated annealed omit maps[13]. Refining the positions and occupancies of multiple conformations in XPLOR did not reduce the R-factor. One region of the map that was particularly difficult to define was around the principal samarium site. This was expected to be the calcium binding site but no calcium was seen at this site in the native map. Calcium-binding might reasonably be expected to improve the order in this region and data for the pectate lyase-calcium complex were therefore collected.

4.3 High resolution pectate lyase-calcium data

Calcium is essential for activity but there was no evidence of calcium in the native map. A crystal was soaked in 5.0 mM calcium chloride in mother liquor overnight and data collected using the small MAR scanner on station 9.5 at the SRS Daresbury. These data, reduced using DENZO[14], were 99.8% complete to a resolution of 1.8 Å with a mean multiplicity of 4.2 and Rsym on intensity of 0.057. These data were used in XPLOR but the R-factor remained at 0.31, there was no evidence for calcium and the loops remained poorly defined.

4.4 Automated refinement procedure (APR)

The XPLOR model, the 1.8 Å resolution data and unrestrained-ARP[15] refinement gave an exceptionally clear map. The R-factor fell from 0.38 to 0.17 over 40 cycles against all data, 43,813 reflections, in the range 10.0 to 1.8 Å. ARP tore the long loops apart and broke any model bias present. The unrestrained model made no stereochemical sense in the incorrect regions but gave a clear indication of what was wrong with the structure. ARP is a fast and user friendly program. After rebuilding the loops, a second round of unrestrained ARP was used to find water molecules. The amino and carboxy terminal extensions and several other regions were rebuilt to the resulting map. Twenty cycles of restrained ARP (Fig. 3d) then resulted in an R-factor of 0.14.

5. THE FINAL MODEL

5.1 Final refinement

Restrain[16] with strict stereochemistry (bond length deviation 0.012Å) was used to refine the final model. The model consists of all 399 amino-acids, the active site calcium-ion and 300 water molecules and the R-factor is 0.15 for all data in the range 10.0 to 1.8Å.

5.2 The architecture of pectate lyase

The simple yet elegant polypeptide fold of the parallel β-helix domain of pectate lyase and the long irregular loops are shown in Fig. 4a. The regularity of the internal structure is especially striking (Fig. 4b) as is the existence of several asparagines within the hydrophobic core and the partially solvent exposed parallel β-sheet. The underlying structural theme responsible for forming the parallel β-helix is the α_L-bounded β-

strand. The three aromatic-rich long loops form a local globular structure of nearly 100 residues with 2 helices and 2 β-hairpins with its own hydrophobic core, forming only limited interactions with the rest of the molecule. It is unclear if this second "domain" could fold independently of the parallel β-helix. The calcium binding site of pectate lyase is between the two domains and is formed by three aspartates.

ACKNOWLEDGEMENT

We should like to thank our collaborators Professor Janine Robert-Boudouy and Dr. William Nasser for purifying and sequencing the *Bacillus subtilis* pectate lyase.

REFERENCES

1. Rombouts, F.M. and Pilnik, W. Economic Microbiology: Microbial Enzymes and Bioconversions *5*, Ed. Rose, A.H. Academic Press, New York (1980) 228
2. Nasser, W., Awadé. A.C., Reverchon, S. and Robert-Boudouy, J. FEBS Letts. *335* (1993) 319
3. Jenkins, J.A., Nasser, W., Scott, M., Pickersgill, R., Vignon, J.-C. and Robert-Boudouy, J. J. Mol. Biol. *228* (1992) 1255
4. Howard, A.J. *et al.* J. Appl. Crystallogr. *20* (1987) 383
5. Otwinowski, Z. Proceedings of the CCP4 Study Weekend (1991) 80
6. SERC Daresbury Laboratory CCP4, Warrington, England (1979)
7. Zhang, K.Y.J. Acta Cryst. *D49* (1993) 213
8. Jones, T.A., Zou, J.-Y., Cowan, S.W. and Kjeldgaard, M. Acta Cryst. *A47* (1991) 110
9. Brünger, A.T., Kuriyan, J. and Karplus, M. Science *235* (1987) 458
10. Reid, R.J. Acta Cryst. *A42* (1986) 140
11. Yoder, M.D., Keen, N.T. and Jurnak, F. Science 260 (1993) 1503
12. Yoder, M.D., Lietzke, S.E. and Jurnak, F. Structure 1 (1993) 241
13. Hodel, A., Kim, S.-H. and Brünger, A.T. Acta Cryst. *A48* (1992) 851
14. Otwinowski, Z. Personal Communication
15. Lamzin, V.S. and Wilson, K.S. Acta Cryst. *D49* (1993) 129
16. Driessen, H., Haneef, M.I.J., Harris, G.W., Howlin, B., Khan, G. and Moss, D.S. J. Appl. Crystallogr. *22* (1989) 510

Methods of Minimization and their Implications

D. E. Tronrud

Howard Hughes Medical Institute
and
Institute of Molecular Biology
University of Oregon
Eugene, OR 97403

Abstract

The process of refinement is a large problem in function minimization. To reduce the amount of computation the methods chosen to minimize the function incorporate a number of assumptions. When these assumptions break down special procedures must be used.

Most of these procedures are commonly known, such as rigid body refinement, but an understanding of the details of the methods themselves allows one to know when and what procedure to apply.

1 Introduction

After the crude model of a protein is constructed we enter the stage of refinement. The parameters of the model are altered to improve the agreement between the model and the experimental observations. If we construct a function which reflects the discrepancy between the these two, refinement becomes the minimization of this function.

All are familiar with fitting models to data in this fashion. Finding the "least-squares" line through a collection of points is the classic example. However line fitting is easy and refinement is hard — The difference lies in the relationship between the model and the experimental observations. This paper will discuss how the particulars of our structural model limits our ability to interpret diffraction data.

1.1 Method vs. Function

There is one distinction which must be clearly made, but is usually treated in an ambiguous fashion. This is the difference between the two choices to be made. First the function which describes the difference between the observations and the predictions of the model. The second is the choice of the method by which this function will be minimized.

There are several methods of minimization commonly used today. Most are described in detail below. Each of them can be used to minimize any function.

In crystallography three functions commonly used. They are the least-squares residual, the empirical energy function, and the correlation coefficient.

The least-squares residual function is

$$f(\mathbf{p}) = \sum_{i}^{\text{all data}} \frac{1}{\sigma(i)^2}(Q_o(i) - Q_c(i,\mathbf{p}))^2, \tag{1}$$

where $Q_o(i)$ and $\sigma(i)$ are the value and standard deviation for observation number i. $Q_c(i,\mathbf{p})$ is the model's prediction for observation i using the set of model parameters $\mathbf{p}$. The values of the parameters found by minimizing this function are those which have the smallest individual standard deviation, or the smallest probable error[4].

The justification for refining against an empirical energy function is the belief that the true protein structure should be at an energy minimum as well as a best fit to the crystallographic observations. While this is undoubtedly correct in the absence of errors in the measured intensities and energy parameters, an analysis of the effect of the presence of such errors has not been done. In practice, usually the parameters of the energy function are chosen in a fashion to allow the energy to mimic the least-squares residual.

Confusion can result if the value of such an "energy" function is interpreted as an energy.

The correlation coefficient is a different measure of the agreement between the model and the observations. In statistics it is used to judge whether there is any agreement at all. This makes it very sensitive to changes in the model when the agreement between the model and the observations is only barely detectable. The correlation coefficient is commonly used in the solution of rotation functions, but has not been used commonly in individual atom refinement.

To describe a refinement protocol it is not sufficient to state one or the other of these choices. One can not meaningfully state that a model was refined with "least-squares". Both the function and the method must be stated.

2 Minimization Methods

Function minimization methods fall on a continuum. The distinguishing characteristic is the amount of information about the function which must be explicitly calculated and supplied for the algorithm. All methods require the ability to calculate the value of the function given a particular set of values for the parameters of the model. There are methods which require only the function values (Simulated Annealing is such a method, it uses the gradient of the function only incidentally in generating new sets of parameters.). Some methods require gradient of the function as well. These methods, as a class, are called Gradient Descent methods.

The method of minimization which uses the gradient and all of the second derivative (or curvature) information is call the "Full-Matrix" method. The Full-Matrix method is quite powerful but the requirements of memory and computations for its implementation are beyond current computer technology except for small molecules and smaller proteins. Also, for reasons to be discussed, this algorithm can only be used when the model is very close to the minimum — closer than most "completely" refined protein models. For proteins, it has only been applied to cases where the molecule is small (< 1000 atoms) which diffract to high resolution and have previously been exhaustively refined with gradient descent methods.

2.1 The Full-Matrix Method

An analysis of the Full-Matrix method, and all gradient descent methods begins with the Taylor's series expansion of the function being minimized. For a generic function ($f(\mathbf{p})$) the Taylor's expansion is

$$f(\mathbf{p}) = f(\mathbf{p}_0) + \left|\frac{df(\mathbf{p})}{d\mathbf{p}}\right|^t_{\mathbf{p}=\mathbf{p}_0} (\mathbf{p}-\mathbf{p}_0) + \frac{1}{2}(\mathbf{p}-\mathbf{p}_0)^t \left|\frac{d^2 f(\mathbf{p})}{d\mathbf{p}^2}\right|_{\mathbf{p}=\mathbf{p}_0} (\mathbf{p}-\mathbf{p}_0) + \cdots, \tag{2}$$

where $\mathbf{p}_0$ is the current set of parameters of the model. In all cases the additional terms (represented by "$\cdots$") are ignored. This assumption has considerable consequences which will be discussed later.

We can change the nomenclature used in equation 2 to more closely match those in refinement by defining $\mathbf{p}_0$ to be the parameters of the current model and $\mathbf{s}$ to be a "shift vector" which we want to add to $\mathbf{p}_0$. $\mathbf{s}$ is equal to $\mathbf{p}-\mathbf{p}_0$. The new version of Equation 2 is

$$f(\mathbf{p}_0 + \mathbf{s}) = f(\mathbf{p}_0) + \left|\frac{df(\mathbf{p})}{d\mathbf{p}}\right|^t_{\mathbf{p}=\mathbf{p}_0} \mathbf{s} + \frac{1}{2}\mathbf{s}^t \left|\frac{d^2 f(\mathbf{p})}{d\mathbf{p}^2}\right|_{\mathbf{p}=\mathbf{p}_0} \mathbf{s} \tag{3}$$

and its derivative is

$$\left|\frac{df(\mathbf{p})}{d\mathbf{p}}\right|_{(\mathbf{p}=\mathbf{p}_0+\mathbf{s})} = \left|\frac{df(\mathbf{p})}{d\mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}_0} + \left|\frac{d^2 f(\mathbf{p})}{d\mathbf{p}^2}\right|^t_{\mathbf{p}=\mathbf{p}_0} \mathbf{s}. \tag{4}$$

Since the first and second derivatives can be calculated given any particular value for $\mathbf{p}_0$ this equation allows the gradient of the function to be calculated given any shift vector. In addition the equation can be inverted to allow the shift vector to be calculated given the gradient of the function.

At the minimum (or maximum) of a function all components of the gradient are zero. Therefore we should be able to calculate the shift vector between the current model ($\mathbf{p}_0$) and the minimum. The equation for this is simple —

$$\mathbf{s} = -\left|\frac{d^2 f(\mathbf{p})}{d\mathbf{p}^2}\right|^{-1}_{\mathbf{p}=\mathbf{p}_0} \left|\frac{df(\mathbf{p})}{d\mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}_0}. \tag{5}$$

The Full-Matrix method is to use this equation, evaluated with the current parameters, to calculate $\mathbf{s}$. $\mathbf{s}$ is then added to $\mathbf{p}_0$ to give the set of

parameters which cause the function to be minimal, and in the case of refinement the best fit to the observations.

This method sounds great. One calculates a single expression and the minimum is discovered. When fitting a line to a set of points this is exactly what is done. In refinement something is obviously different. The difference arrises from the "$\cdots$" which we choose to ignore. In the case of fitting a line to points the terms represented by "$\cdots$" in fact are zero. The truncated Taylor's series is exact and the shift vector is also exact. In refinement these terms are not equal to zero resulting in the shift vector giving only the approximate location of the minimum.

The quality of the estimate is limited by the size of the terms which are ignored. The terms of the Taylor's series have increasing powers of $\mathbf{s}$. The first term ignored is multiplied by $\mathbf{s}^3$ and the higher order terms are multiplied by ever higher powers. If $\mathbf{s}$ is small these higher order terms become quite small also. Therefore the closer $\mathbf{p}_0$ is to the minimum the better estimate $\mathbf{s}$ becomes.

The Full-Matrix method, and all the gradient descent methods which are derived from it, becomes a series of successive approximations. An initial guess for the parameters of the model ($\mathbf{p}_0$) is manufactured somehow. For the shift vector to actually give an improved set of parameters the guess must be sufficiently close to the minimum. The "sufficiently close" criteria is rather difficult to calculate exactly.

The distance from the minimum at which a minimization method breakdown is called the "radius of convergence". It is clear is that the Full-Matrix method is much more restrictive than the gradient descent methods, and the gradient descent methods are more restrictive than simulated annealing, Metropolis, and Monte Carlo methods. Basically the less information about the function calculated at a particular point the larger the radius of convergence will be.

The property of the Full Matrix method which compensates for its restricted radius of convergence is its "power of convergence". If the starting model is within the radius of the Full Matrix method that method will be able to bring the model to the minimum quicker than any other method.

2.1.1 The Normal Matrix

The aspect of the Full-Matrix minimization method which prevents it being used in common refinement is the difficulty in calculating the term

$$\left|\frac{d^2 f(\mathbf{p})}{d\mathbf{p}^2}\right|^{-1}_{\mathbf{p}=\mathbf{p}_0}. \tag{6}$$

This matrix written out in full is

$$\begin{pmatrix} \frac{\partial^2 f(\mathbf{p})}{\partial p_1^2} & \frac{\partial^2 f(\mathbf{p})}{\partial p_2 \partial p_1} & \cdots & \frac{\partial^2 f(\mathbf{p})}{\partial p_n \partial p_1} \\ \frac{\partial^2 f(\mathbf{p})}{\partial p_1 \partial p_2} & \frac{\partial^2 f(\mathbf{p})}{\partial p_2^2} & \cdots & \frac{\partial^2 f(\mathbf{p})}{\partial p_n \partial p_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{p})}{\partial p_1 \partial p_n} & \frac{\partial^2 f(\mathbf{p})}{\partial p_2 \partial p_n} & \cdots & \frac{\partial^2 f(\mathbf{p})}{\partial p_n^2} \end{pmatrix}^{-1}. \tag{7}$$

This matrix contains $n \times n$ elements, where n is the number of parameters in the model. In a typical case n will be on the order of 10,000. The number of elements in the second derivative matrix, often called the Normal matrix, would be 100,000,000. It would take a lot of computer time to calculate it, a lot of memory to store it, and a lot more computer time to invert it. The gradient descent methods make various assumptions about the importance of different parts of the Normal matrix to reduce these requirements.

To understand the relative importance of the different elements of the Normal matrix we need to understand the meanings of each part. The most important classification of the elements is the distinction between the elements on the diagonal and those off it. The elements on the diagonal are affected by a single parameter and are therefore somewhat easier to analyse. The off-diagonal elements are affected jointly by two parameters.

The information contained in the off-diagonal elements described how the effect on the function of changing parameter a is affected by changes in parameter b. In essence it is related to the correlation of the two parameters. If one considers the simple case where each parameter is varied in turn. Parameter a is moved to the value where the function in minimized. Then parameter b is changed. If the off-diagonal element for a and b is nonzero than parameter a will have to be readjusted, and the larger the value the greater the adjustment required.

The diagonal elements contain information about the affect of a parameter's value on its own affect on the function. This, of course, will always be

large. (If the diagonal element is zero than any value for that parameter will be equivalent: a property which is usually undesirable in a parameter.)

2.2 Sparce Matrix Method

One can examine the relationship between the parameters in the model to determine which pairs will have significant off-diagonal elements in the normal matrix. The pairs whose off-diagonal elements are predicted to be small can then be ignored. Such selective attention only pays off when the vast majority of elements are removed.

With some functions all the off-diagonal elements may be ignored where other functions do not allow any. One must treat functions on a case by case basis to determine which elements to use. An analysis of the residual function for x-ray diffraction shows that the size of the off-diagonal elements is related to the extent of electron density overlap of the two atoms. Since atoms are fairly compact all off-diagonal terms between parameters in atoms are negligible except for atoms bonded to one another, and the terms for those pairs are small. Since an atom has a large overlap with its own electrons the diagonal elements are very large compared to any off-diagonal ones.

The stereochemical restraints commonly used in protein refinement have a different pattern. Here the parameters of atoms connected by a bond distance or angle have strong correlation. Atoms not restrained to one another have no correlation at all. The off-diagonal terms which are nonzero are as significant as the diagonal ones.

This knowledge allows one to calculate the normal matrix as a sparse matrix, the vast majority of the off-diagonal elements are never calculated or even have computer memory allocated for their storage. The only elements calculated are the diagonal ones (including contributions from both the crystallographic and stereochemical restraints) and the off-diagonal elements for parameters from atoms directly connected by geometric restraints.

Even with the simplification of the normal matrix introduced by the sparse approximation the problem of inverting the matrix is difficult. There are a number of methods available for generating an approximation to the inverse of a sparse matrix. A discussion of these methods is beyond the scope of this paper. However it is important to note that each of them includes assumptions and approximations which should be understood when they are used.

The refinement program PROLSQ[2] uses the sparse matrix approximation to the normal matrix. PROLSQ inverts the matrix using a method called "Conjugate Gradient" which is unrelated to the Conjugate Gradient method used to minimize functions. It is a sign of confusion to state that X-PLOR[1] and PROLSQ both use the Conjugate Gradient method.

It is quite difficult to calculate the proper values for the elements of the normal matrix. To simplify these calculations Konnert and Hendrickson decided to implement all stereochemical restraints as distances. While this restructuring of the restraints does simplify the normal matrix it makes the restraints more difficult for the user to visualize and prevents the minimization method from seeing the true, underlying nature of the restraints.

While the minimization method used in PROLSQ is the most powerful of those used in large molecule refinement (and therefore the smallest radius of convergence) in practice it does not seem to work any better than simple the Conjugate Gradient method. Its limitations arise from the approximations made in the calculation of the normal matrix elements and the way the space matrix is inverted.

2.3 Diagonal Matrix

A further step in simplifying the normal matrix is made by ignoring all off-diagonal elements. The normal matrix becomes a diagonal matrix, which is inverted by simply inverting each diagonal element in turn. In essence working with the matrix has become a one-dimensional problem. Since any correlation between parameters has been assumed away the shift for a particular parameter can be calculated in isolation from the shifts of all other parameters. The Full Matrix equation 5 becomes

$$s_i = -\left|\frac{\partial f(\mathbf{p})}{\partial p_i}\right|_{\mathbf{p}=\mathbf{p}_0} \Big/ \left|\frac{\partial^2 f(\mathbf{p})}{\partial p_i^2}\right|_{\mathbf{p}=\mathbf{p}_0} . \tag{8}$$

2.4 Steepest Descent

A further simplification can be made if all the diagonal elements of the normal matrix have the same value. If this is true none of them need to be calculated at all. The average value can be estimated from the behaver of the function value as the parameters are shifted. The shift for a particular parameter is

simply

$$s_i = -\left|\frac{\partial f(\mathbf{p})}{\partial p_i}\right|_{\mathbf{p}=\mathbf{p}_0} \tag{9}$$

The Steepest Descent method is far from Full Matrix. However is has the advantage of a large radius of convergence. Since the gradient of a function points in the steepest direction up hill, the Steepest Descent method simply shifts the parameters in the steepest direction down hill. It is guaranteed to reach the local minimum, given enough time. Any method which actually divides by the second derivative is subject to problems in the curvature is negative, or worst yet zero. Near a minimum all second derivatives must be positive. Near a maximum they are all negative. As one moves away from the minimum the normal matrix elements tend toward zero. The curvature becomes zero at the inflection point which surrounds each local minimum. The Full Matrix becomes unstable somewhere between the minimum and the inflection point. The Diagonal Approximation method has a similar radius of convergence.

However, the Steepest Descent method simply moves the parameters to decrease the function value. It will move toward the minimum when the starting point is anywhere within the ridge of hills surrounding the minimum.

2.5 Conjugate Gradient

The Steepest Descent method is very robust. It will smoothly converge to the local minimum whatever the starting parameters are. However it will require a great deal of time to do so. One would like a method which would reach the minimum quicker.

The problem with Steepest Descent is that no information about the normal matrix is used to calculate the shifts to the parameters. Where ever the assumptions break down (the parameters have correlation and have different diagonal elements) the shifts generated will be inefficient.

Just as one can calculate an estimate for the slope of a function by looking at the function value at two nearby points, one can estimate the curvature of a function by looking at the change in the function's gradient at two similar points. This experiment is routinely performed in Steepest Descent refinement. The gradient is calculated, the parameters shifted a little, and the gradient calculated again. In Steepest Descent the two gradients are never

compared but if they were a bit of information about the normal matrix could be learned.

The Conjugate Gradient method[3] does just this. The analysis of Fletcher and Reeves showed that the Steepest Descent shift vector can be improved by adding a well defined fraction of the shift vector of the previous cycle. Each cycle essentially "learns" about one dimension of curvature in the n dimensional refinement space. Therefore after n cycles everything is known about the normal matrix and the minimum is found.

The shift vector for cycle $k+1$ using Conjugate Gradient is

$$\mathbf{s}_{k+1} = -\left|\frac{df(\mathbf{p})}{d\mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}_k} + \beta_{k+1}\mathbf{s}_k, \tag{10}$$

where β_{k+1} is the ratio of the length of the function's present gradient to its previous length. During the first cycle there is no previous cycle. The first cycle must be Steepest Descent.

The fundamental limitation of the Conjugate Gradient method is that it is guaranteed to reach the minimum in n cycles only if the Taylor's series does indeed terminate, as assumed in equation 3. If there are higher order terms, and there are in crystallographic refinement, then n cycles will only get the model nearer to the minimum. One should start over with a new run of n cycles to get the model even closer.

Even n cycles is a lot in crystallography. No one runs thousands of cycles of Conjugate Gradient refinement, nor can they be run with current software. The shifts become too small to be represented with the precision of current computers. Small shifts are not necessary unimportant ones. These small shifts add up to significant changes in the model, but we cannot calculate them.

2.6 Conjugate Direction

The Conjugate Gradient method is better than the Steepest Descent method because it uses some information about the normal matrix to improve the quality of the shift vector. It would seems reasonable to believe that the shift vector could be improved further if additional information were added. For instance, we can calculate the diagonal elements of the normal matrix directly, and quickly.

All this information is combined together in the Conjugate Direction method[5]. This method operates like the Conjugate Gradient method except it uses the shifts from the Diagonal Matrix method for its first cycle instead of the Steepest Descent method's. The shift vector in Conjugate Direction is

$$\mathbf{s}_{k+1} = -\left|\frac{df(\mathbf{p})}{d\mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}_k} \Big/ \left|\frac{d^2 f(\mathbf{p})}{d\mathbf{p}_i^2}\right|_{\mathbf{p}=\mathbf{p}_k} + \beta'_{k+1}\mathbf{s}_k \,, \tag{11}$$

where the trick is calculating β'_{k+1} correctly. This matter is discussed in detail in [5].

3 What Does This Mean?

You have now read more than you ever wanted to know about function minimization methods. The basic facts which have been presented are that the methods commonly used find difficult cases where there are correlation between parameters and where the diagonal elements for some parameters differ from the average. These two limitations are the source of most problems in refinement.

3.1 Rigid Body

If you have a model which is in error because of an overall rotation and translation, or even if only one domain has such an error, none of the refinement packages will be able to correct this automatically. To correct this error a concerted shift of a large number of atoms must be made. The indication that such a shift is required in located in the off-diagonal elements of the normal matrix, which has been discarded. For this problem to be corrected you will have to resort to rigid body refinement.

In rigid body refinement the molecule is defined to contain one or more groups within which the atoms cannot move relative to one another. Basically the parameters of the model are recast in a form in which there no longer are correlations between the parameters.

While it is unlikely that errors of this type will arise when solving a problem with the MIR method it is quite common in MR and even Molecular Substitution (Isomorphous inhibitor or mutant structures). Refinement in these latter cases should always be started with overall rigid body refinement

and refinement with each domain as a separate body. There have been a number of cases where the refinement has "hung up" in the mid to upper 20's (percent R-factor) where the problem was eventually traced to a small unguessed domain shift.

3.2 Separate XYZ and B

Whenever no part of the normal matrix is directly calculated, as in Steepest Descent and Conjugate Gradient, the method tends to minimize the function by shifting only those parameters which have large diagonal elements. Because the diagonal elements are larger for positional parameters (like x, y, and z) than the thermal factors, the B values will not be shifted to their correct values. This is why routinely these classes of parameters are refined in separate cycles. One first refines the positional parameters holding the B values fixed and then refines the thermal factors holding the positions of the atoms constant.

However because these parameters are correlated to one another it is difficult for both types of parameters to reach their optimal values. One must repeat the cycle many times for the parameters to settle down, more cycles than are usually done.

All parameters may be varied simultaneously when the diagonal elements of the normal matrix are explicitly included in the calculation of the shifts. This is one reason why Conjugate Direction refinement requires fewer cycles.

3.3 Heavy Atoms

When refining with the Conjugate Gradient method the assumption is made that all the diagonal elements are equal. For the positional parameters these elements usually are similar enough that the real differences can be accommodated with the usual number of cycles. However one factor which contributes to the magnitude of the diagonal elements is the number of electrons surrounding the atom. Heavy atoms such as iron, calcium, and chlorine have much larger diagonal elements and will be shifted much larger distances. In fact, with Conjugate Gradient they will be shifted too far.

This is why heavy atoms tend to oscillate in refinement. The amount of shift is determined based on the average atom, which is about the size of carbon, so the heavy atoms will be over shifted. The next cycle of refinement

will attempt to correct their position but again will over shift. The atoms will slip back and forth from cycle to cycle. This problem occurs for all parameters of the heavy atom, both positional and thermal.

The problem can easily be overlooked if only the overall statistics are monitored. The mean and rms shift may be very small even though one or two atoms continue to move a great deal in each cycle. One must monitor the atoms with the largest shifts and try to understand why they continue to shift.

A minimization method which uses at least the diagonal elements will correct the problem.

3.4 High B's

Another factor which contributes to the magnitude of the diagonal elements is the size of the B value of the atom. Atoms with high B values have smaller diagonal elements for all their parameters. These atoms will be under shifted in Conjugate Gradient refinement. Since usually atoms are created with small B values and are under shifted in each cycle they will never reach their correct location nor will their thermal parameters become as large as required by the observations.

Atoms with large B values almost certainly should have even larger ones. Again, a method which uses the diagonal elements will not exhibit this problem. However, most programs which do utilize the diagonal elements of the normal matrix calculate them with the approximation that only the element type of the atom is important. They ignore the contribution of the B value to these terms. This is done for historical reasons. In small molecule structures, where refinement was originally developed, there usually are no atoms with particularly large B values, and the assumption is good. The early protein refinement packages made the same assumption without reconsidering its validity.

Therefore some programs will underestimate the thermal factors even when using the diagonal elements. To learn if this is a problem in your refinement you must discover exactly how these matrix elements are calculated. Usually this is not an easy task. The program's source code must be examined.

3.5 Local Minima

The last problem which must be considered is that we can never reach the local minimum. Often it has been said that refinement was continued until convergence at the local minimum. Even in a perfect case, where our refinement residual was quadratic, both Conjugate Gradient and Conjugate Direction would require n cycles where n is at least four times the number of atoms in the model. No one has ever ran that many cycles.

This means that no one has ever been "trapped in a local minimum". They have never reached a local minimum.

4 Summary

The fact that we cannot include all the information about our residual function into our refinement results in some parameters of the model oscillating, other becoming stuck, and the requirement that we run many, many cycles. Until more powerful methods of minimization become available the crystallographer must be on guard.

References

[1] Brunger, A.T., Kuriyan, K. & Karplus, M. (1987). Crystallographic R factor refinement by molecular dynamics. *Science* **235**, 458–460.

[2] Hendrickson, W.A. & Konnert, J.H. (1980). In *Computing in Crystallography*, edited by Diamond, R., Ramaseshan, S. & Venkatesan, K., chapter 13, pages 13.01–13.25. Indian Academy of Sciences, Bangalore.

[3] Fletcher, R. & Reeves, C. (1964). Function minimization by conjugate gradients. *Computer Journal* **7**, 81–84.

[4] Mandel, John. *The Statistical Analysis of Experimental Data.* Dover Publications, Inc., New York, (1984).

[5] Tronrud, D.E. (1992). Conjugate-direction minimization – An improved method for the refinement of macromolecules. *Acta Crystallogr A* **48**, 912–916.

A Generalized Approach to the Fitting of Non-Peptide Electron Density

Paula M. D. Fitzgerald
Department of Biophysical Chemistry, Merck Research Laboratories
P. O. Box 2000 (Ry50-105), Rahway, New Jersey 07065 USA

Introduction

Proteins have been studied crystallographically in complex with non-peptides of various types (cofactors, prosthetic groups, inhibitors, substrates) for many years. However, the current wide-spread application of crystallographic techniques to structure-based drug design projects has greatly increased the number of these complexes being analysed. Each group involved in such a project develops its own way of solving the problems that are unique to fitting and refining non-peptide electron density, but descriptions of the details of the techniques employed seldom appear in print. There is a need for a generalized approach to fitting non-peptide density, one that is robust even in situations complicated by multiple, overlapping images of the non-peptide in the electron density. This paper outlines the evolution of my thinking on this subject, the result of a number of years spent studying inhibited complexes of HIV-1 protease. The methods outlined are valid for any type of non-peptide, but for simplicity the term inhibitor will be used to represent the many categories of non-peptides alluded to above.

HIV-1 protease often turns out to be a particularly challenging system for structural studies, a consequence of the symmetry of the enzyme. The active form of HIV-1 protease is a dimer, with the active site lying at the interface between the two monomers. The binding site for inhibitors is thus itself twofold symmetric (or nearly so). Some inhibitors bind to the enzyme in a single orientation, but it is not uncommon to find that an inhibitor binds in two, approximately twofold symmetric, orientations (1). The crystallographer is then left with the task of attempting to deconvolute the images corresponding to each of the independent orientations. This can be straightforward if the two images are not overlapped, but interpretation of the electron density can be extremely difficult if they are.

Our laboratory has determined the structures of more than 25 different inhibited complexes of HIV-1 protease to date; of these, approximately 60 percent show two superimposed inhibitor orientations, and roughly 20 percent have proved difficult to interpret with confidence. The approach outlined below has helped enormously with these difficult cases, and it is a reliable and simple approach to follow even in straightforward cases. What follows is a description of the steps used in generating an atomic model for the inhibitor, generating restraints for refinement from that model, and fitting the inhibitor density during the course of structure refinement. Some refinement quirks that have been encountered are described, and a case study, the interpretation of the density for an inhibitor with badly overlapped alternative orientations, is presented.

Generating a Model for the Inhibitor

Before the density for an inhibitor can be fit, an atomic model for the compound must be obtained. The model will be used as a source of atoms to fit to the density, and as a source of chemical restraints for the crystallographic refinement of the structure. In some cases, a model can be extracted from the small molecule structure literature, but more often the compound under study will be a new chemical entity, and there will be no complete structural information available.

The steps involved in the process of model generation depend on the tools used to perform the task. In this work the two dimensional chemical sketching program CHEMNOTE in the program package QUANTA (2) was employed to generate a two-dimensional drawing of the inhibitor. Stereochemistry was designated at this stage, and a user-defined atom numbering scheme was imposed (overwriting the default atom numbers assigned by the program). When one leaves CHEMNOTE and returns to QUANTA proper, the two dimensional chemical drawing is automatically converted into a three-dimensional representation of the molecule. This inital atomic model must then be minimized so as to generate a model with appropriate bond distances, angles, and so forth.

This work was begun using early versions of QUANTA, and there were consistent difficulties with the assignment of stereochemistry (problems that have been fixed in more recent versions of the program). Faced with these troubles, it was decided to use the QUANTA/CHEMNOTE tools for drawing the molecule, numbering it, and converting it to three dimensions, but to then transfer the process to program AMF (a Merck proprietary program that is an extension of the work described in (3) and (4)) to check and fix the stereochemistry, and then to idealize (but not minimize) the coordinates (the distinction being that idealization minimizes the deviation from ideal bond distances and angles, but does not attempt to minimize the energy of the molecule.) AMF is then used to output coordinates in Protein Data Bank format for input to program CHAIN (5) for graphical fitting.

Generating Restraints for Refinement

The refinement described here uses the PROTIN/PROLSQ suite of programs (6). PROTIN requires a file containing geometric restraints (bonds, angles, planes, chiral centers, torsion angles, and non-bonded distances) for each of the chemical entities being refined. (Note that when using PROTIN/PROLSQ, only geometric restraints need to be specified for refinement. If a program containing energy terms, such as X-PLOR, is to be used for refinement, the user must also generate energetic restraints.)

Program CONEXN (7), written by Pähler and Hendrickson, reads a set of coordinates for the new compound and generates a new entry in the Ideals file provided with the program suite. The user must specify which atoms are to be treated as chiral centers and which should be constrained to lie in a plane. CONEXN was used for many structure refinements in our laboratory, but after many successes, a difficulty was encountered with the refinement of an inhibitor that contained a non-aromatic ring. Eventually a bug in the program was located (the program was applying non-bonded contact restraints to atoms in such a ring); this

problem was fixed by adding a user specification of ring atoms, so that unwarranted non-bonded restraints are not applied.

This modified version of the program has been adapted for free-format, keyword driven input, and a detailed document has been written. The program is available from the author upon request by email (paula_fitzgerald@merck.com). It should be noted that the Ideals files read and written by this version of CONEXN may not be compatible with all versions of PROTIN/PROLSQ, as those programs have undergone divergent evolution over the years.

The calculation of intramolecular non-bonded restraints for the inhibitor is dependent on the conformation of the inhibitor. The user inputs a limiting radius, and each atom is compared to all others within a sphere of that radius. If the two atoms being compared are connected by bond through two intervening atoms, a non-bonded restraint is applied. I have adopted the practice of recalculating the restraints file towards the end of refinement, once the bound conformation of the inhibitor has been established, in order that the appropriate non-bonded restraints are applied.

Fitting the Density

Most often one is faced with the problem of fitting an inhibitor that contains a number of adjustable torsion angles. Occasionally the preferred conformation of the non-peptide is obvious, but much more often there are many degrees of freedom. When we fit proteins, we rely on our knowledge of previously determined structures to guide us in exploiting preferred conformations of main and side chains, but we seldom have an equivalent database of information for a new inhibitor (although a search of the structures in the Cambridge Crystallographic Data Base can be enlightening for determining preferred conformations of substructures of the inhibitor.)

The approach used in our studies of inhibited complexes of HIV-1 protease is to begin refinement with a model from a previous structure determination in the same space group (with the previous inhibitor, the solvent model, and any side chains modeled in alternative conformations removed). Each grand cycle consists of 10 cycles of PROLSQ refinement, followed by calculation of $2F_O$-F_C and F_O-F_C maps. The model is then adjusted graphically, and potential solvent atoms are added to the model by reference to a peak list calculated from the F_O-F_C map. In our earlier studies, the entire inhibitor would be fit to the difference density after 2 or 3 such grand cycles. The process is continued until the entire model (protein and inhibitor) appears to be optimized and the solvent model appears to be complete.

Experience with a number of structures demonstrated the potential drawbacks of this approach. It is quite painful to fit an inhibitor with 10-12 degrees of freedom by torsional adjustments only, as small torsional adjustments at one end of the molecule can have large structural consequences at the other end. This difficulty can be gotten around by breaking the inhibitor into fragments and fitting them independently, but this has the consequence of destroying the geometry of the inhibitor (although that will be restored during refinement). A more serious difficulty is that once the model is fit, the refinement becomes biased toward

that interpretation of the density, and it is easy to become locked into an initial erroneous interpretation.

Hence a different approach was adopted, one that does not introduce actual atomic coordinates (and restraints on those coordinates) for the inhibitor until much later in the process. This methodology is based on the ideas presented in the Automated Refinement Procedure of Lamzin and Wilson (8). It involves placing dummy atoms at local maxima in the difference electron density during every grand cycle of refinement. These peaks are treated as ordinary water in refinement, without taking any special cautions to relieve restraints against close non-bonded contacts. New dummy atoms are added at each cycle of refinement, until the point when no new peaks appear (this usually occurs between grand cycles 4 and 6).

At this point atomic coordinates for the inhibitor are fit to the most easily identified fragments of the inhibitor. Fitting these atoms is quite straightforward, as the dummy atoms refine to positions either near atomic positions in the final refined structure, or along atomic bonds. These fragments are enlarged in subsequent grand cycles until the entire inhibitor has been fit. This incremental fitting of the atomic model for the inhibitor greatly limits the degrees of freedom available for fitting each subsequent piece, which can aid enormously in the fitting of portions of the electron density that prove difficult to interpret. But the major advantage of the approach is the fitting aid provided by the dummy atoms, and the fact that the inhibitor is fit to the model at a much more advanced stage of refinement.

Quirks of Refinement

Several recurrent refinement quirks have been encountered during the course of refinement of this series of inhibited complexes. One is that refinement often destabilizes when the complete model for the inhibitor is added to the refinement. One can only guess at the theoretical reasons for this, but a practical fix is to turn off the refinement of chiral centers for the first few grand cycles of refinement after the inhibitor model is added. Once the entire inhibitor model has settled into the refinement, the chiral restraints can be turned back on and the refinement remains stable.

Another practical problem is the estimation of occupancy for each orientation of the inhibitor when there are two overlapped orientations present in the atomic model. This is handled by initially setting the occupancies of the two orientations to a fixed value of 0.5. After each grand cycle of refinement, the mean *B* factor for each orientation of the inhibitor is calculated. If the mean *B* factors are significantly different, the occupancy of the orientation with the lower mean *B* is raised by 0.05 (or less) and the occupancy of the orientation with the higher mean *B* is lowered by the same amount. This has the effect of making the refined mean *B* factors more similar to one another at the end of the next grand cycle of refinement; if a disparity in mean *B* factors still remains, the process is repeated in the next grand cycle. The largest difference in occupancy seen in our series of structures has been 0.65 and 0.35; 0.35 probably represents the lowest occupancy that will give an interpretable image in the electron density.

The Approach Applied to a Difficult Structure Interpretation

The course of the refinement of HIV-1 protease with the inhibitor L-735,489 will be described as an illustration of the methods outlined above. The inhibited complex crystallized in space group $P6_1$, a = 63.57 Å, c = 83.40 Å. The data were measured using a Siemens multiwire detector, and processed with the XENGEN (9) software; the processed data had an R_{merge} of 0.055 for the data from infinity to 1.95 Å (85.3% complete for data with I > sigma(I).)

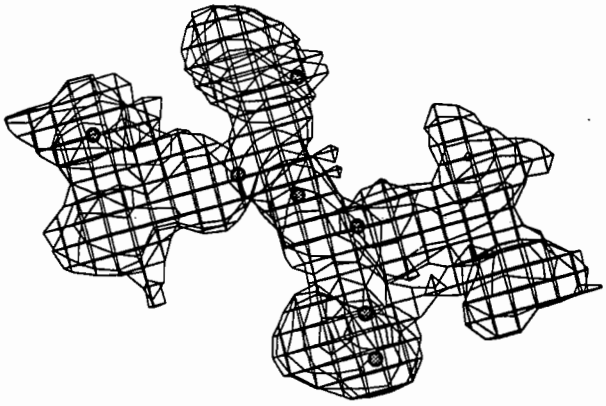
The initial difference density for this structure clearly indicated that there were two overlapping orientations of the inhibitor. The refinement of the structure was attempted twice before the current refinement protocol was adopted, using the data between 8.00 and 1.95Å, and both times ran aground with the inability to deconvolute the overlapping density for the P2 and P2' moieties of the inhibitor. The initial density did not lend itself to an interpretation that made both conformational and chemical sense, and once the entire inhibitor was fit, it became impossible to refine away from the bias introduced by the improperly fit inhibitor.

The course of refinement under the current approach is described below. This time the structure was refined against the data between 20.0 and 1.95 Å resolution, and the lower resolution data seemed to have a significant beneficial effect on the connectedness of the inhibitor density at all stages of refinement. The R-value for the final model was 0.184 and deviations from ideal bond distances were 0.018 Å.

The figures that follow show the $2F_O$-F_C electron density map obtained at the indicated grand cycle, contoured at the one sigma level. Dummy atoms corresponding to peaks in the F_O-F_C map (not shown) are indicated with gray circles; dummy atoms that have been carried through refinement from a previous grand cycle are indicated with black circles.

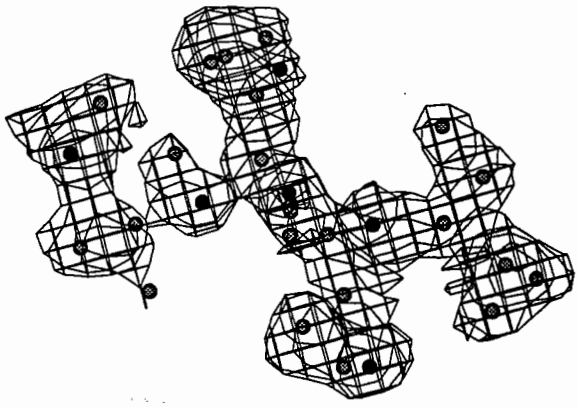
References

1. Fitzgerald, P.M.D., McKeever, B.M., VanMiddlesworth, J.F., Springer, J.P., *et al.* J. Biol. Chem, *265* (1990) 14209-14219
2. Quanta, Molecular Simulations, Burlington, MA
3. Gund, P., Andose, J.D., Rhodes, J.B. and Smith, G.M. Science, *208* (1980) 1425-1431
4. Smith, B.M., Hangauer, D.G., Andose, J.D., Bush, B.L., *et al.* Drug Inf. J., *18* (1984) 167-178
5. Sack, J.S. J. Mol. Graphics, *6* (1988) 224-225
6. Hendrickson, W.A. Methods Enzymol., *115* (1985) 252-270
7. Pähler, A. and Hendrickson, W.A. J. Appl. Crystallogr., *23* (1990) 218-221
8. Lamzin, V.S. and Wilson, K.S. Acta Cryst., *D49* (1993) 129-147
9. Howard, A.J., Gilliland, G.L., Finzel, B.C., Poulos, T.L., *et al.* J. Appl. Crystallogr., *20* (1987) 383-387



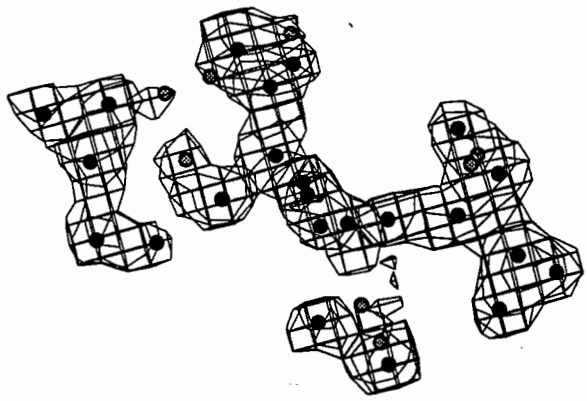
Cycle 1

The $2F_o$-F_c electron density after the first grand cycle of refinement. The circles filled with gray correspond to local maxima in the F_o-F_c electron density. These positions are assigned "dummy" atoms, which are refined as solvent (oxygen) atoms in the next grand cycle of refinement.



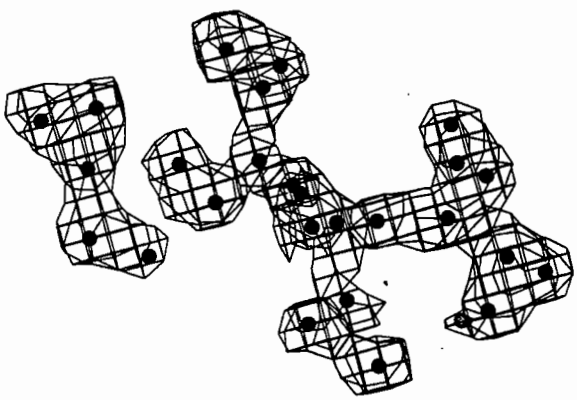
Cycle 2

The dummy atoms from Cycle 1 are now considered part of the model; they now are drawn as black circles. Grey circles corresponded to local maxima in the F_o-F_c electron density map calculated after the second grand cycle of refinement.



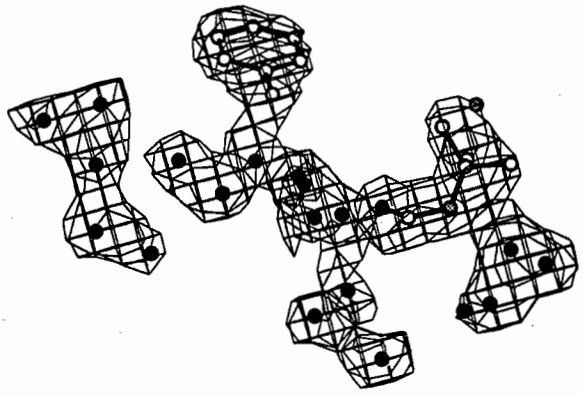
Cycle 3

The number of new dummy atoms that appear in each cycle is now diminishing. A number of dummy atoms are rejected from one cycle to another because they refine to an unreasonably high B factor.



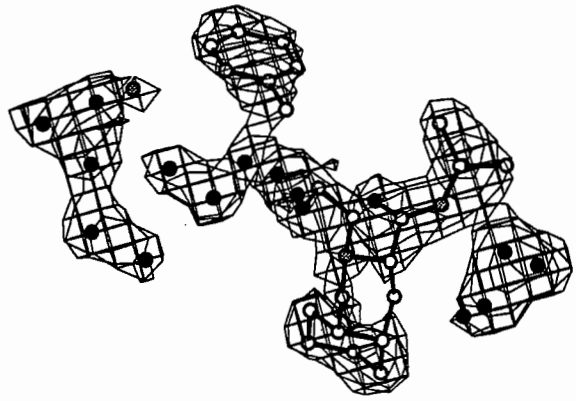
Cycle 4

The dummy atoms model for the inhibitor is now essentially complete.



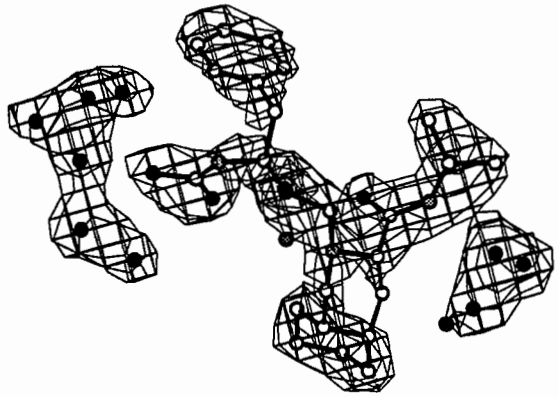
Cycle 5

The most easily recognized fragments in the inhibitor are now fit to the electron density, using the dummy atoms as guide points. The refined positions of the dummy atoms end up lying either very close to atomic positions, or along atomic bonds.



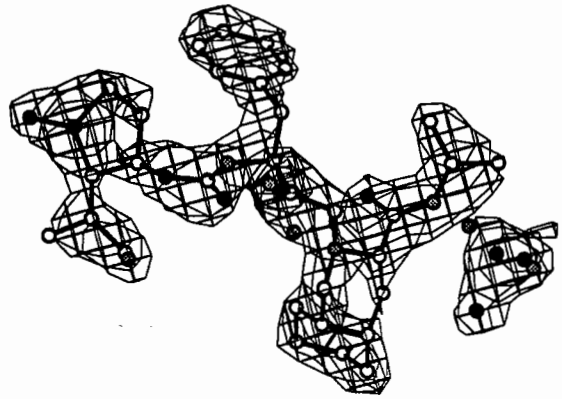
Cycle 6

The framents are enlarged, again using dummy atoms as guide to aid the fitting.



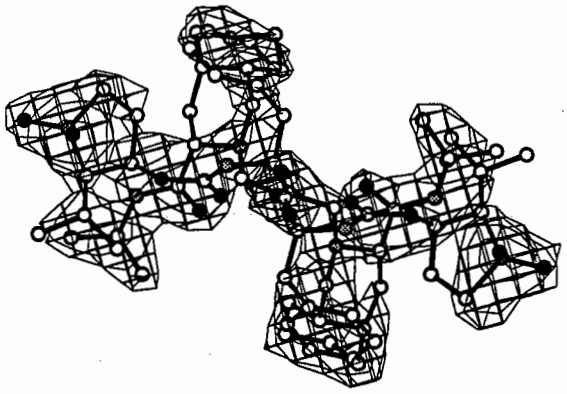
Cycle 7

The inhibitor is now missing only the N-terminal fragment. This fragment was left to the last to fit, as it had been very difficult to be certain of the interpretation of the electron density corresponding to this portion of the inhibitor.



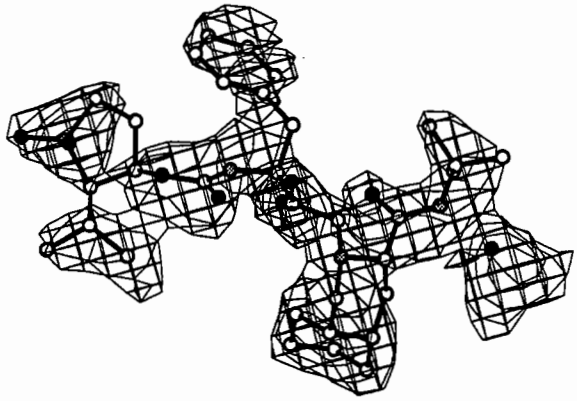
Cycle 8

The N-terminal fragment is added, and the inhibitor model is now complete. The dummy atoms remaining in the lower right correspond to atoms in the second orientation of the inhibitor, as do the dummy atoms in the center of the molecule.



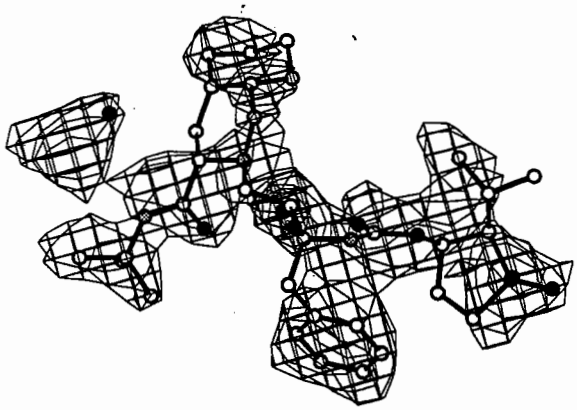
Cycle 11

Grand cycles 9 and 10 do not involve fitting to the inhibitor. In Cycle 11 the second orientation of the inhibitor is fit and the last of the dummy atoms are removed.



Cycle 14

One orientation of the inhibitor is shown after the final cycle of refinement. The mean *B* factor for atoms in this orientation is 18.49. A solvent atom that is hydrogen bonded to a main-chain nitrogen in the inhibitor in the final model is also shown.



Cycle 14

The second orientation of the inhibitor, also shown after the final cycle of refinement. The mean *B* factor for atoms in this orientation is 20.11, so the occupancies for both orientations were left at their initial values of 0.5.

Analysis of Water Structure.

Hugh Savage.
Chemistry Dept.
Univeristy of York.
Heslington, York YO1 5DD.
U.K.

Water is an important component in the structure, stability and functioning of living organisms. Crystals of biological macromolecules contain between 25 and 70% (some cases ~90%) solvent, which makes a significant contribution to the scattering process in the x-ray and neutron diffraction analyses of such systems.

The derivation of solvent structure within crystalline hydrates, involves two main steps. Firstly, the electron (or neutron) densities of the solvent regions must be examined, and solvent positions assigned (modelled for example, as single sites or a grid), which are then refined. Secondly, the solvent structure has to be interpreted from the given set of refined sites. In this talk we will mainly concentrate on the first step, but will also highlight the level of atomic detail one can achieve in the analysis of solvent structure using very high resolution neutron data.

There are several problems related to size, resolution and disorder that prevent us from deriving all the useful structural information about the solvent within larger crystal hydrates. These include (a) attainable resolution, (b) modelling disorder, (c) side chain disorder, (d) overlap of peaks, (e) chemical identity of peaks and (f) estimating occupancies. In general, these factors become more problematical as the size of the hydrate increases, particularly when analysing large macromolecular structures, such as protein crystals in which large regions of disordered bulk-solvent often exist.

(a) **Resolution:** Figure 1 shows a region of solvent electron density at different resolution cut-offs of 1.0Å and 2.0Å for one set of x-ray data for coenzyme B12 crystals. At 2.0Å a six-membered H-bonded water ring is readily apparent. However, at 1.0Å, both a six-membered and a five-membered can be identified and also a third six-membered ring. On the left of the 2.0Å solvent density, three waters appear to be present (third site lies under one of the two shown), with distances of ~2.5Å between them. However at 1.0Å, there are four peaks apparent (again one lies below), and the region can be identified as an acetone molecule, which is readily confirmed in three other 1.0Å data sets. It may thus be dangerous to rely solely on automated solvent peak assignment programs that are often used, especially at lower resolutions of 1.5-3.0Å, without a careful analysis of the associated solvent geometries (H-bonds, van der Waals etc.). The analysis of substrates bound to macromolecules can also be mis-interpreted at lower resolutions.

(b) **Modelling disorder**: the interpretation of the more diffuse regions of solvent is another problem in solvent analysis and refinement. Figure 2 shows such a case for neutron solvent density around a hydroxyl group in coenzyme B12.

Figure 1

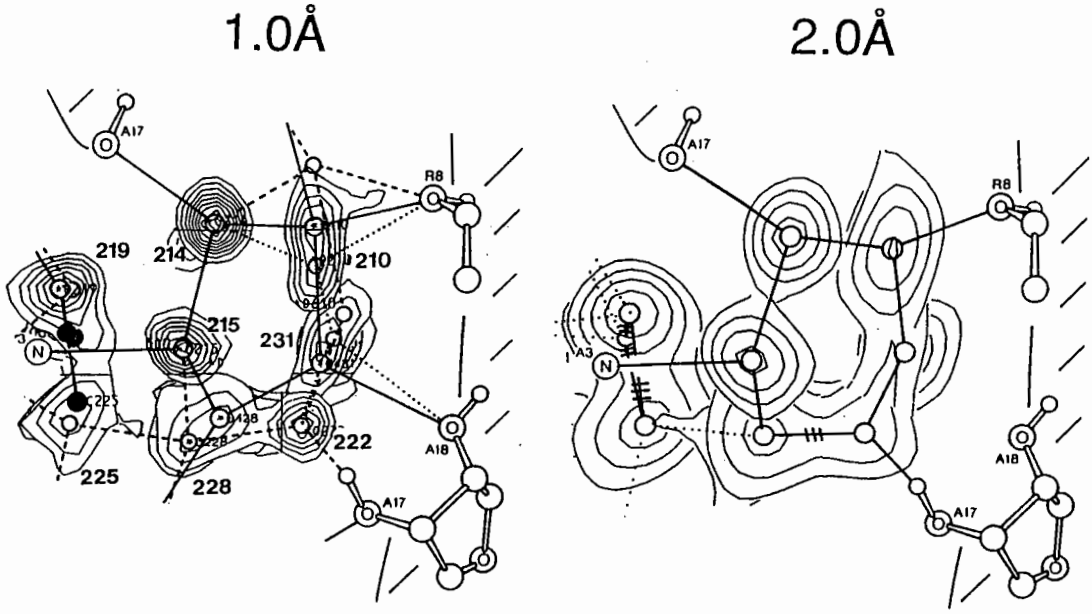
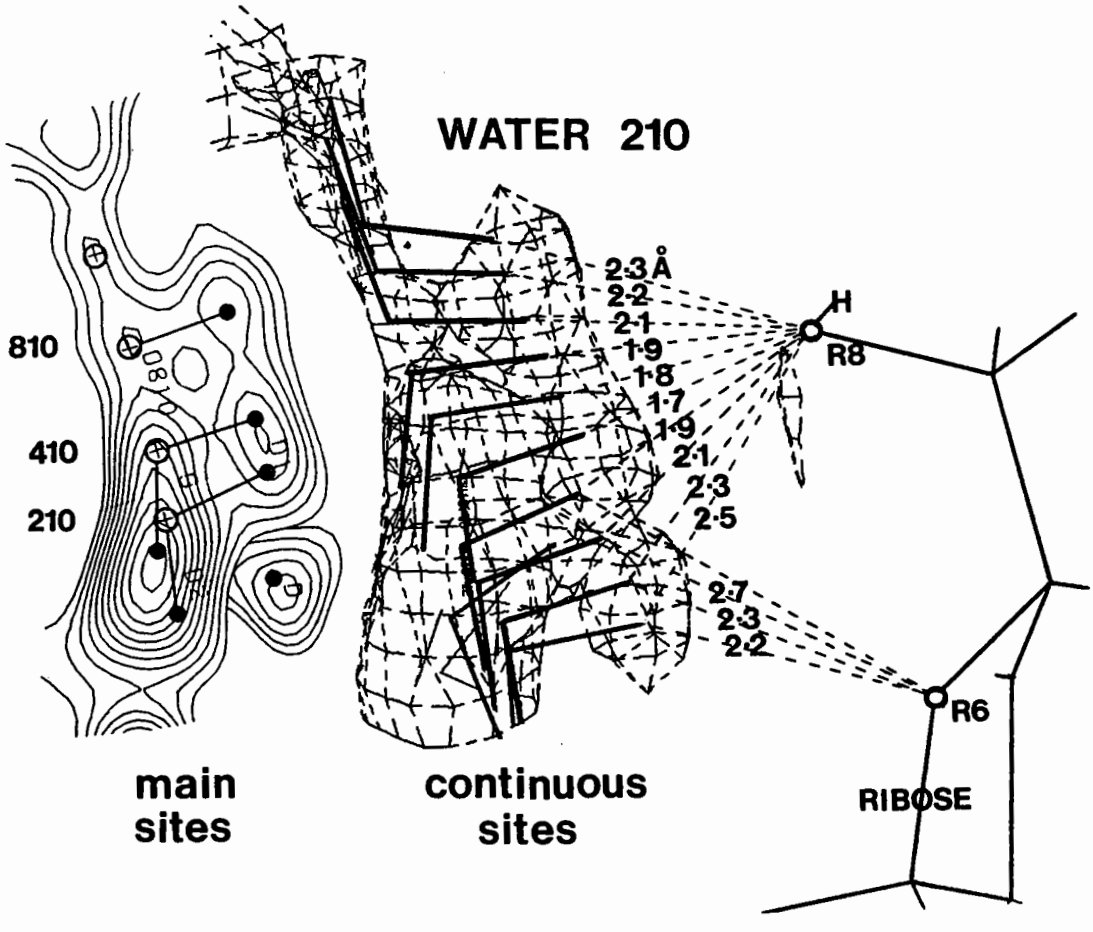


Figure 2

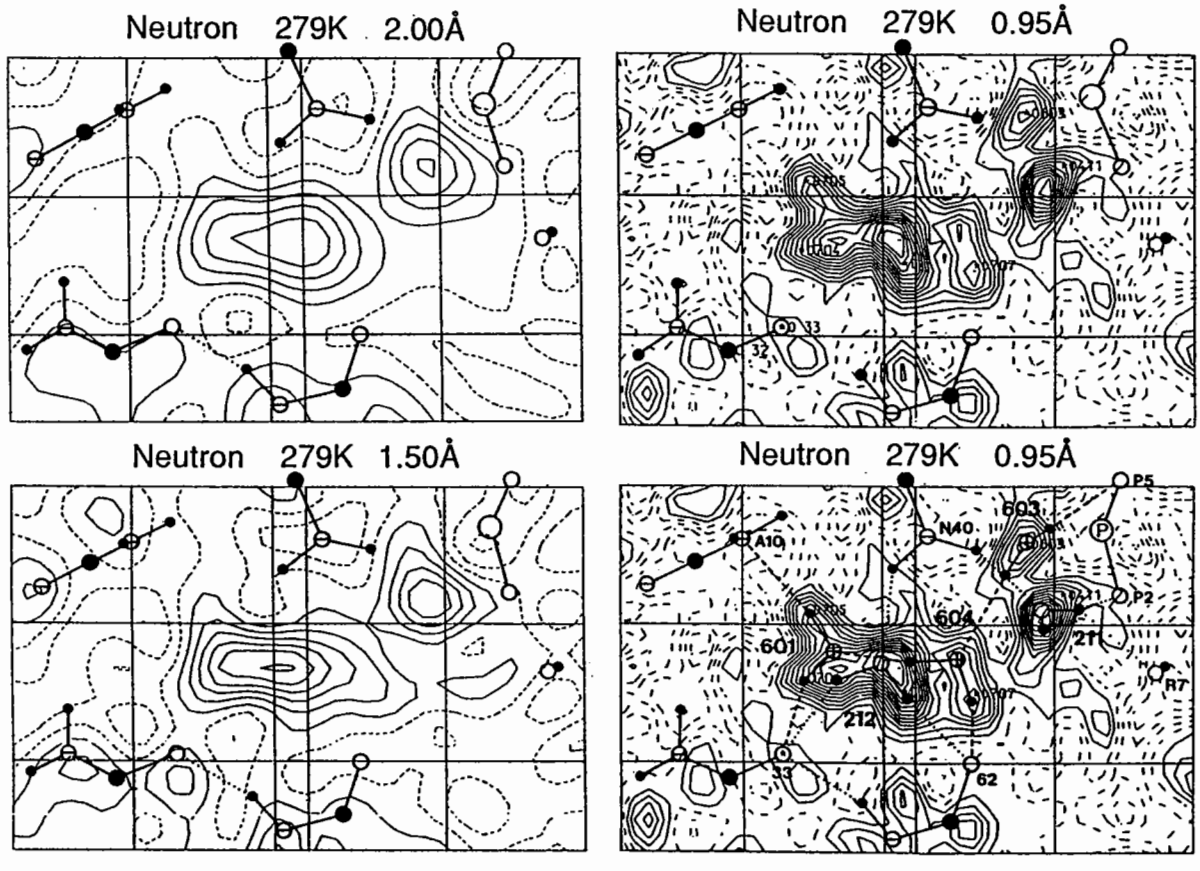


The hydroxyl group (R8) is shown on the right, while a chicken wire representation of the solvent density is shown in the middle and a contour representation of the same density is shown on the left. The elongated solvent density is continuous from the top to the bottom in figure 2. How does one model such density which represents a time-averaged dynamic assembly? Incorporation of appropriate Monte Carlo and molecular dynamics methods may be considered, but become very difficult for large complex systems whereby many different molecular conformations must be taken into account. Using a more "static" approach, such a solvent region can be modelled by first assigning "main" sites to the well-defined solvent density (left in figure 2), and then assigning "continuous" sites to the elongated and diffuse regions of density at intervals of about one third of the resolution of the data (middle of figure 2). For continuous bulk solvent, a three-dimensional grid is commonly used, with each point smeared out by a significantly high thermal-factor.

(c) **Disorder of side-groups:** disorder of the crystallising molecule increases significantly the degrees of complexity in the surrounding solvent structure and to the task of interpreting the H-bonded networks, since all cases must be accounted for. This is particularly the case in larger macromolecular hydrates for which high resolution (<1.0Å) analyses may reveal a number of alternate H-bonded water networks around each disordered group.

(d) **Overlap of peaks and resolution**: the overlap of solvent peaks at different resolutions may give rise to different interpretations of the structure (as in (a) above). This is of particular consequence in the interpretation of enzyme active sites and the binding of drugs and other substrates to macromolecules. Figure 3 shows neutron solvent density at three different resolutions of 2.0Å, 1.5Å and 0.95Å over a region of solvent density in coenzyme B12 crystals. At 2.0Å and 1.5Å resolutions, two or three water molecules appear to be present. However, at 0.95Å five water molecules are readily interpreted forming two well-defined water networks with good H-bonding geometries. Clearly, very high resolution data of better than 1.0Å provides the most unambiguous interpretation.

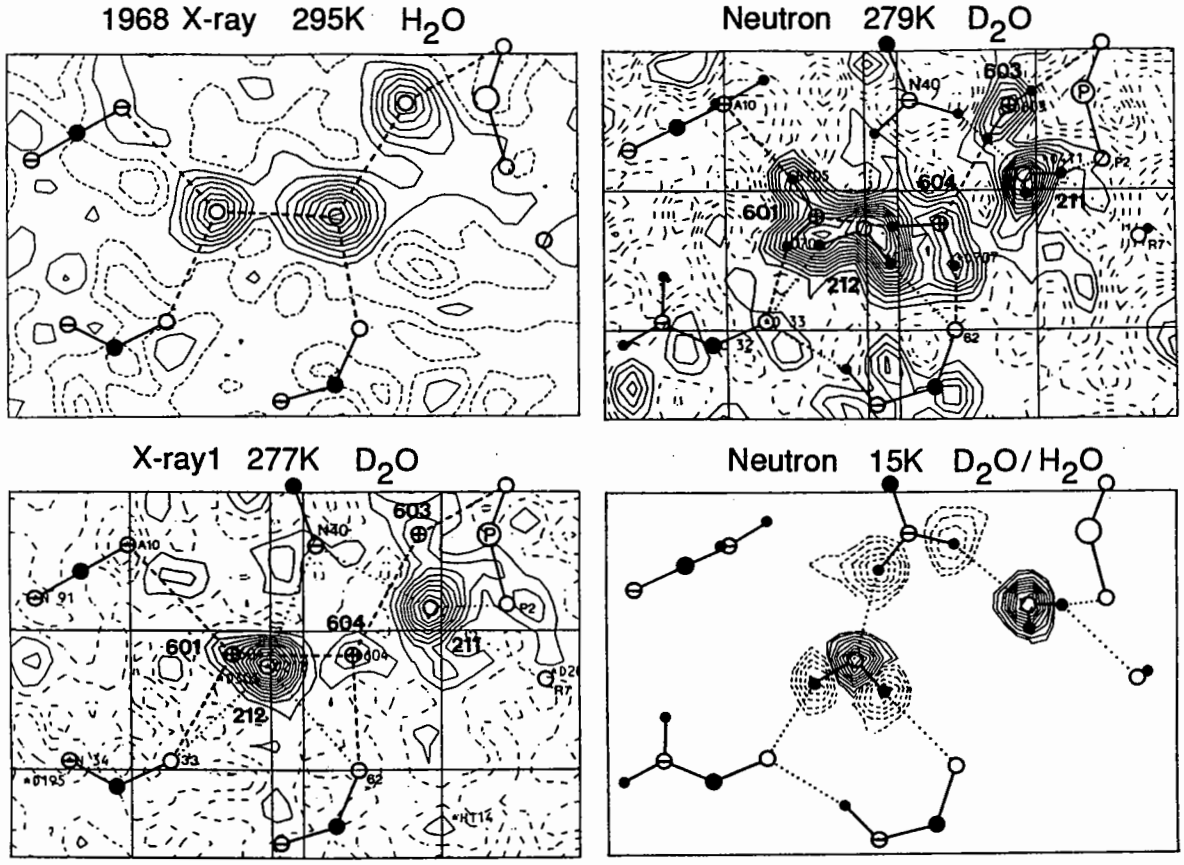
Figure 3



(e) **Chemical identity of peaks**: in the crystallisation of macromolecules, other solvent species are often present, as for example in protein crystallisations. The "foreign" species such as salt ions or organic molecules (eg acetone, ethanol) may only be partially occupied, thus care is needed in assigning, interpreting, and checking the geometries of the solvent structure within hydrate crystals.

(f) **Estimating occupancies:** this is a problematic area of solvent analysis. In least squares refinement, these two parameters are usually highly correlated and except for well-ordered solvent positions, it is difficult to obtain reliable values for these parameters in more disordered solvent regions. When well-ordered solvent positions are present, each site may have different occupancies in different data sets collected on the same hydrate system. Figure 4 shows an example of this for four different data-sets (2 x-ray and 2 neutron) collected on coenzyme B12 crystals.

Figure 4



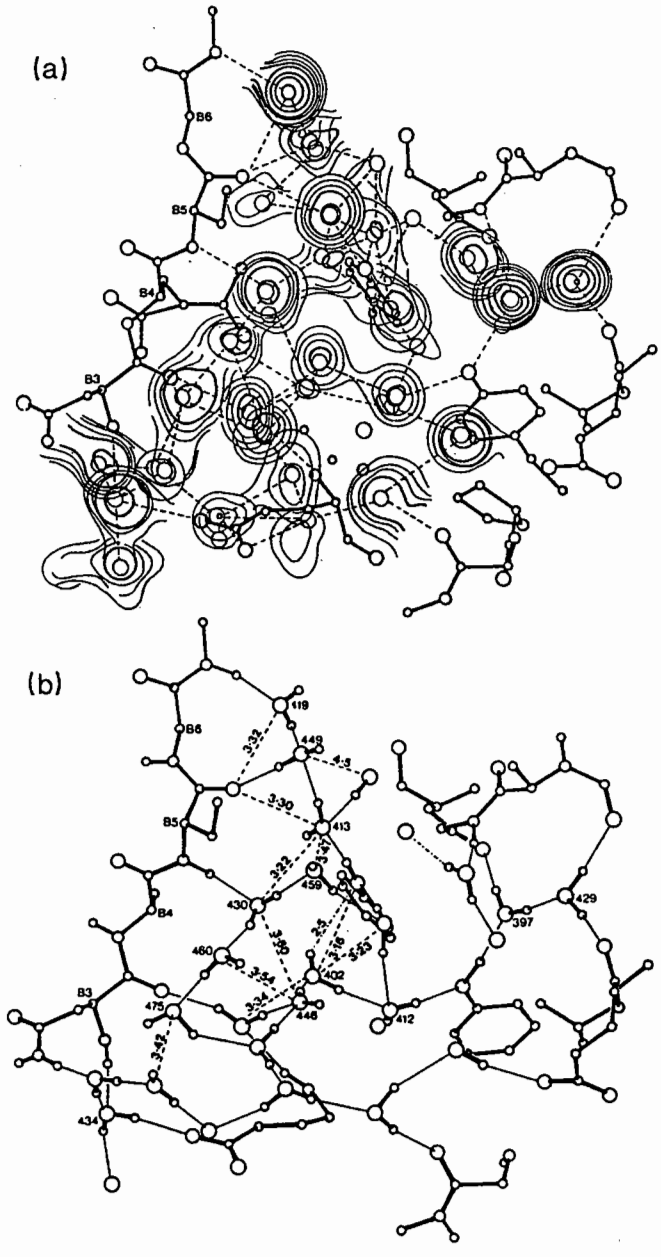
Between the four data-sets, two alternate water networks are present: network A (dotted line) and network B (dashed line). However, the occupancies of each network are different in the different data-sets as follows:

Data-set:	Occ of Network A	Occ of Network B
X-ray 295K:	0.0	1.0
X-ray 277K:	0.9	0.1
Neutron 279K:	0.6	0.4
Neutron 15K:	1.0	0.0

The best criteria for a detailed analysis and refinement of the solvent structure include the following: (1) The use of very high resolution data of 1.0Å or better. This criteria cannot be overstressed, although it is at present difficult to achieve for most macromolecular systems, however with advances in synchrotron data collection, it is greatly improving (K. Wilson, next article). (2) Use low temperatures of 10-100K. (3) Use neutron diffraction data as well as x-ray data: the former provides a very much higher signal for partially occupied hydrogens than x-rays can give. (4) Use hydrate systems of adequate size from which high resolution data can be obtained. (5) Collect several sets of data on the same system.

Once a set of solvent sites have been assigned and refined, the next stage is to interpret and assign the solvent structure. In most cases for macromolecules only x-ray data is available, from which there is usually no information about the solvent hydrogen positions. In addition, most assigned solvent sites are assumed to be water oxygen sites, unless the local stereochemical geometries suggest a non-water species (eg acetone at 1.0Å in figure 1). Given this initial solvent information as in the top part of figure 5 for 2Zn insulin at 1.5Å resolution (Baker et al, 1988), we must try to formulate water networks between the potential oxygen sites using the rather flexible geometrical restraints of for example, hydrogen bonding (O...O H-bonds between 2.6-3.5Å) and non-bonding (O...O non-bonded >3.2Å). One significant problem that arises in the use of these restraints is for solvent O...O distances of around 3.2-3.5Å. Are they non-bonded of H-bonded? This problem is also exasperated by the presence of significantly larger errors (0.1-0.5Å) in the solvent positions within macromolecular hydrates.

Figure 5

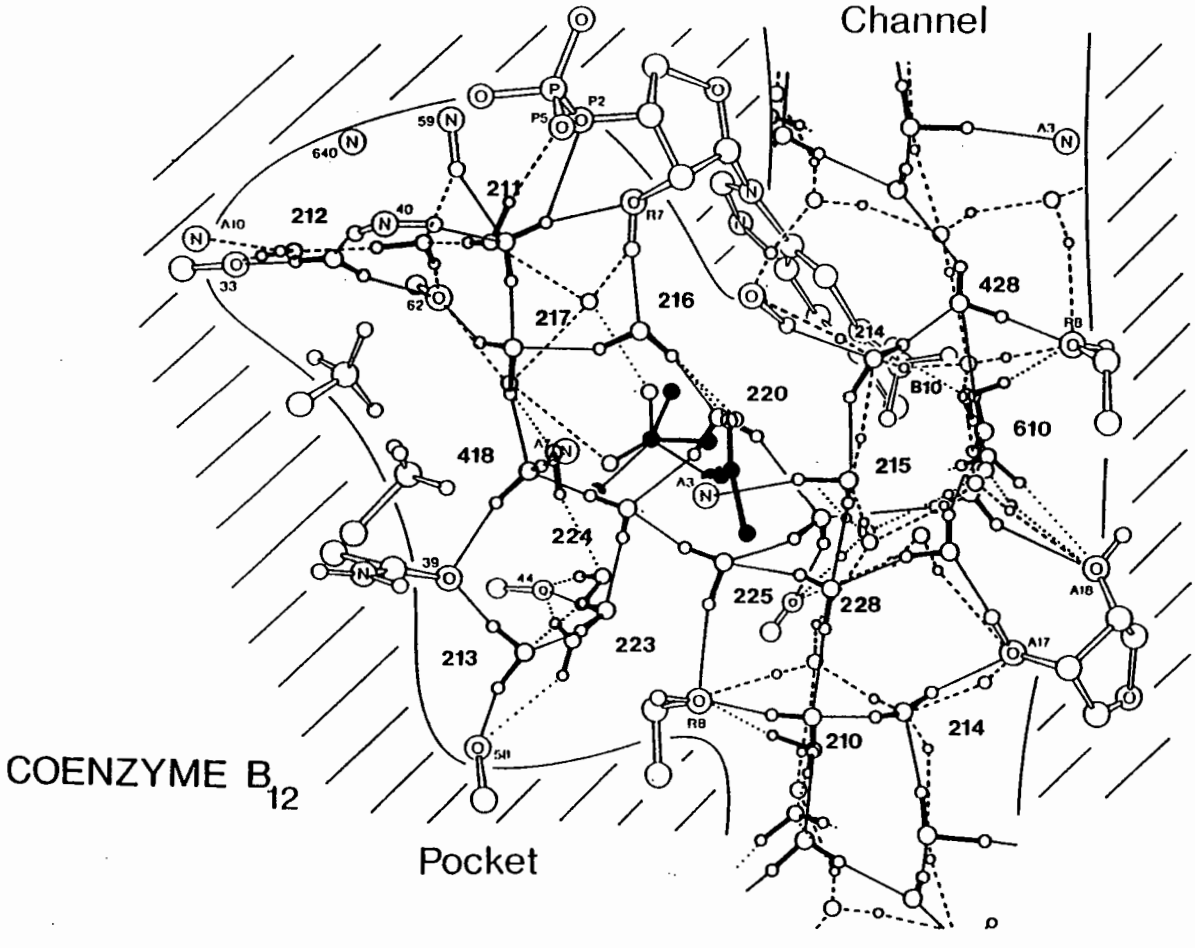


Using a set of four structural restraints (RR1-4; Savage and Finney, 1986)) between water oxygen and hydrogen atoms, it is possible to assign the water hydrogen positions and formulate water H-bonded networks between solvent oxygen sites. An example of this is given in the lower part of figure 5 for a solvent analysis of 2Zn insulin (solvent electron density shown in upper part of figure 5).

In figure 6, the complete analysis and refinement of the water structure is shown for the solvent in crystals of the macromolecule, vitamin B12 coenzyme (Bouquiere et al, 1994, 1993; Savage, 1986). One asymmetric cell is shown which contains a maximum of 17 water molecules, for which approximately 40 oxygen and 60 hydrogen sites were assigned. At least six different complete water network structures can be formulated from these sites that readily satisfy H-bonding and non-bonding geometrical restraints. This analysis used 5 different sets of data (3 x-ray, 2 neutron) of between 0.90Å and 1.20Å resolution and demonstrates the complexity and flexibility of solvent interactions and structure around bio-macromolecules. Detailed information about the interactions between water and apolar bio-groups, particularly with respect to the positions of the hydrogen atoms, can also be gained at very high resolutions using neutron diffraction. For example, several partial H-bonded clathrate cage water structures are seen to occur around methyl B10 of the coenzyme's mainly apolar benzimidazole group (below water 214 in figure 6) which protrudes into the channel. Also, water bridges are made around two methyl groups pointing into the left of the pocket: in the region of waters 212, 217, 418, 224 and 223.

Figure 6

The shaded areas represent the main areas of coenzyme B12 atoms. The solvent regions are divided in to a pocket region (left) and a channel (right), the latter extending throughout the crystal.



Given the right criteria, it is possible to extend the analyses of solvent in medium-sized macromolecular systems such as coenzyme B12, to larger systems such as small proteins. For example, x-ray synchrotron data to 1.2Å have been collected at Daresbury from γ-II crystallin eye-lens crystals, in which extensive water network structures are apparent. Other proteins include crambin crystals (neutron and x-ray data), wherein the solvent structure contains many alcohol molecules.

References :

Baker, E.N., Hodgkin, D.C. et al (1988). *Phil. Trans. Roy. Soc. Lon.*, **319**, 369-456.

Bouquiere, J.P., Finney, J.L., Lehmann, M.S., Lindley, P.F. and Savage, H.F.J. (1993). *Acta Cryst.* **B49**, 79-89.

Bouquiere, J.P., Finney, J.L. and Savage, H.F.J. (1994). *Acta Cryst.* **B**, in press.

Savage, H.F.J. (1986). *Biophys. J.*, **50**, 947-980.

Savage, H.F.J. and Finney, J.L.(1986) *Nature,* **322**, 717-720.

PROTEIN REFINEMENT AT ATOMIC RESOLUTION

Keith S. Wilson

European Molecular Biology Laboratory (EMBL), c/o DESY, Notkestrasse 85, D-22603 Hamburg, Germany.

This paper follows on directly from those of Victor Lamzin in the 1992 and of Zbigniew Dauter in the 1993 Daresbury Meeting Proceedings. The results described are entirely dependent on our collaboration with George Sheldrick and with many visitors who have brought crystals to the Hamburg Outstation.

The use of X-ray crystallography leads routinely to the determination of the structures of small molecules at atomic resolution. Atomic resolution requires that the measured data extend to about 1 Å or better and are essentially complete. In addition roughly 66 % of the intensities in the outer resolution shell should be greater than 3 $\sigma(I)$. Data to atomic resolution give advantages in both structure determination and in refinement. For small molecules such data allow automatic *ab initio* phase determination by either Patterson search or statistical direct methods. These approaches provide phases of sufficient quality to reveal the positions of most of the atoms in the structure. Furthermore refinement is normally straightforward.

Least-squares minimisation can in principle be carried out as soon as the number of observations exceeds the number of parameters, but given a degree of experimental error benefits from a ratio of 5:1 or more of observations to parameters. At a resolution of 1 Å there are about 5 observations for each atomic parameter when using a full anisotropic model for the structure, i.e. with nine parameters per atom: three positional and six representing the thermal ellipsoid. With high quality data, small molecule structures can be refined to final R factors of 1-2 % giving errors in atomic coordinates about 0.005-0.01 Å. Indeed with higher resolution data the model of well ordered crystals can be extended to include the deformation density from the spherical atom approximation, revealing the positions of the bonding electrons. Such studies will not be discussed further.

As the size of the unit cell increases so does the problem of collecting atomic resolution data as is shown in Table 1. The table is based on a typical small organic sample as standard.

TABLE 1

DATA COLLECTION AND UNIT CELL VOLUME

CELL DIMENSION Å	CELL VOLUME $Å^3$	RELATIVE INTENSITY	RELATIVE NO. REFLECTIONS	SAMPLE TYPE
5	125	8	1/8	SIMPLE SALT
10	1,000	1	1	SMALL ORGANIC
20	8,000	1/8	8	SUPRAMOLECULE
50	125,000	1/125	125	SMALL PROTEIN
100	1,000,000	1/1,000	1,000	MEDIUM PROTEIN
1,000	1,000,000,000	1/1,000,000	1,000,000	LARGE VIRUS

For supramolecular and macromolecular structures the problems of data collection to atomic resolution are clearly horrendous and it could be some time before we see a large virus at this resolution.

This has meant that many protein, and indeed supramolecular structures such as cyclodextrins, have been determined at less than atomic resolution, for proteins often in the range 3.0-1.5 Å. Successful approaches for solving the phase problem for such structures have been developed, such as isomorphous replacement and molecular replacement, providing experimental values for the initial phases. Refinement of structures against such data presents a second set of problems as is shown in Table 2. As can be seen from the table the parameters which can be refined vary considerably with the resolution of the data. Refinement at less than atomic resolution requires the use of either stereochemical restraints effectively increasing the number of observations (as in the approach established in the programme PROLSQ by Konnert and Hendrickson) or constraints where the number of parameters is limited by using rigid bodies of sets of atoms (as in the programme CORELS developed by Sussman).

TABLE 2

PARAMETERS AND OBSERVATIONS

Using Rubredoxin as an example.

RESOLUTION	NO. OF	OBSERVATIONS/PARAMETERS	
(Å)	REFLECTIONS	ISOTROPIC	ANISOTROPIC
0.92	27,000	13.5	6.0
1.0	21,000	10.5	4.7
1.2	12,200	5.1	2.7
1.5	6,200	3.1	1.4
2.0	2,600	1.3	0.6
2.5	1,300	0.6	0.3
3.0	800	0.4	0.2

This leads to problems in refinement as shown in Table 3. As the resolution becomes limited the observation/parameter ratio and the number of parameters which can be refined becomes less.

TABLE 3

NASTY EFFECTS OF LIMITED RESOLUTION

Resolution	Temperature	Distance	Increasing
(Å)	factors	error (Å)	Uncertainties
0.92	Aniso	0.04	None
1.0	Aniso	0.06	None
1.2	Aniso	0.10	Solvent
1.5	Iso/Aniso	0.15	Disorder
2.0	Iso	0.25	Double conformations
2.5	Iso ?	0.5	Side chain errors
3.0	None !	1.0	Mistakes almost certain
6.0	What ?	?	Probably Globular

For large unit cells there is an intrinsic problem that all the X-ray data are weak as shown in Table 1. There is in principle a practical solution to this problem, namely improving the counting statistics by measuring on a higher intensity source and/or using a more sensitive and efficient (2D) detector. For disordered structures, which clearly includes proteins with about 50 % solvent in the crystal, there is a second problem: the high resolution intensities are weak (or absent). There are potentially two solutions to this: (1) grow better crystals or (2) try cryogenic data collection if the disorder is truly thermal rather than statistical.

In EMBL Hamburg one of our major interests is the recording of accurate high resolution X-ray data to provide well refined parameters and ideally to allow direct statistical solution of the phase problem. Monochromatic synchrotron radiation coupled with the use of the imaging plate scanner developed and built by Jules Hendrix and Arno Lentfer were used in all of our studies. The project originated in data collection on β-cyclodextrin and similar sized structures as described last year by Zbigniew Dauter. We were interested to find out if atomic resolution data could be collected for at least a subset of small proteins by such techniques and in evaluating whether this would allow more detailed refinement and lead to more accurate structural parameters. As Zbigniew reported in 1992, atomic resolution data have now been collected on a number of proteins: the list has been extended to about 15 samples to date and several new proteins are shortly to be added. For all of these data have been collected to 1.2 Å resolution of better.

Refinement of each of these structures by stereochemically restrained least-squares minimisation (PROLSQ) is in most cases complete and the final R factors are 14-18 %. We are now refining these structures using conjugent gradient least-squares as implemented in the program SHELX92 (George Sheldrick, Göttingen) with the imposition of a weak external restraints library (essentially for the disordered parts of the structures) and with anisotropic atomic temperature factors. These structures are being used to build an improved library of stereochemical parameters for less well diffracting proteins.

What have we gained from these refinements ?

1. Firstly the final R factors have dropped, by about 6 % on average, to values in the range 8-12 %. That this represents a true improvement in the models, with the inclusion of extra parameters for anisotropic vibration, is confirmed by the value of R_{FREE}, which also drops by almost the same amount (Sheldrick and us, unpublished). The estimated errors in bond lengths drops to about 0.06 Å, corresponding to the value obtained from inversion on the least-squares matrix giving roughly 0.04 Å as the mean atomic coordinate error. This corresponds to the average error anticipated for the ordered part of the structure without stereochemical restraints, i.e. using X-ray data only.

2. The height of the features in the final F_o-F_c difference maps falls correspondingly to an r.m.s. density value of about 0.25 electrons per $Å^3$. This results from the satisfactory modelling of the anisotropic thermal motion. The features for this after isotropic refinement can also be up to 0.5 electrons per $Å^3$ and their inclusion in the model clearly leads to considerable improvement in the interpretability of the residual difference features such as solvent and hydrogen positions. These have comparable density values.

3. The positions of many hydrogen atoms can be directly determined from difference syntheses where they have been omitted from the model. For rubredoxin at 0.92 Å, more than 2/3 of the hydrogen atoms can be directly and automatically located using the SHELX93 peak search facility.

4. The lower residual difference density has allowed improved modelling of solvent with addition of many more water molecules. This story is not yet complete as is discussed below.

5. The high resolution structures have increased the complexity of the protein model itself. The percentage of residues with more than one conformation of the side chain is often 10 % or more. Thus for the tiny *Desulfovibrio vulgaris* rubredoxin with only 52 residues and a V_m of 1.8, 7 have more than one conformation. For our bacterial trypsin, 20 residues out of 180 are so disordered. For the majority of these residues two clear conformations can be seen, and satisfactorily modelled and refined in SHELX. For a small number of residues the side chains are truly disordered with no clear density for some atoms. Apart from the occasional chain termini this is rarely true for the main chain atoms.

Thus for well diffracting crystals of small proteins, it is clearly possible to collect and use atomic resolution data. The largest molecules for which we have 1 Å data to date are a bacterial and a fungal trypsin, both with molecular weights about 20 kDa. It is expected that this will be extended to 1 or 2 larger enzymes in the near future. This has already allowed extensive refinement of the structures with anisotropic models for atomic thermal motion. It has not yet led to a breakthrough in the success of direct methods for *ab initio* phasing. Only for rubredoxin with its FeS_4 cluster has George Sheldrick been able to produce a direct methods solution.

The availability of several atomic resolution data sets has produced a considerable increase in our computational needs. We have generally carried out the early part of the refinement with PROLSQ as implemented in the CCP4 package with Fast Fourier routines, in combination with the Automatic Refinement Procedure (ARP) developed by Victor Lamzin. The latter is of particular importance for the objective addition of water molecules to the model. It is possible to identify many more solvent molecules in the atomic resolution structures and it becomes an intractable problem to add these manually to the model using a graphics station. ARP has been considerably extended and improved to allow better distance constraints on the atoms added and a consideration of their sphericity. The later stages of refinement have used SHELX93 also in combination with ARP.

What are the residual problems in obtaining/dealing with such data ?

1. It is still not clear how optimally to model the solvent. We have to date used a simplistic model based on unit occupancy and some restrictions on the inter-water distance. From medium sized structures like cyclodextrins and vitamin B12 we know that there are overlapping sets of water networks in the solvent region. Even for medium size structures these are hard to model with certainty. For proteins one can only seek a pragmatic means of providing a reasonable model for the complex solvent volume in the crystal.

2. The use of cryogenic techniques during data collection gives the possibility of reducing the disorder in the crystal and recording data to higher resolution. Thomas Schneider in the outstation has recently collected data at 120 K on a bacterial trypsin to 0.96 Å of considerably

superior quality to those measured to 1.1 Å at room temperature. The analysis of these data will doubtless reveal much more detail in the structure and may well resolve some of the ambiguities in the room temperature structure. Nevertheless it is eventually the structure and function of the molecule at ambient cellular temperature which we wish to define. The behaviour of disordered parts of the protein and of the solvent may well be better defined at low temperatures but the function of the protein can depend critically on their flexibility.

REFERENCES

Dauter, Z., Sieker, L.C. & Wilson, K.S. (1992) 'Refinement of Rubredoxin from *Desulfovibrio vulgaris* at 1.0 Å with and without restraints'. *Acta crystallogr.* **B48**, 42-59.

Lamzin, V.S., & Wilson, K.S. (1993) 'Automated Refinement of Protein Models' *Acta Crystallogr.* **D49**, 129-147.

Sheldrick, G.M., Dauter, Z., Wilson, K.S., Hope, H. & Sieker, L.C. (1993) 'The Application of Direct Methods and Patterson Interpretation to High-Resolution Native Protein Data.' *Acta Crystallogr.* **D49**, 18-23.

Evaluation of protein coordinate data sets

Roman A Laskowski, Malcolm W MacArthur and Janet M Thornton

Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, UK.

Abstract. A number of different checks can be applied to a given protein structure's coordinates to evaluate their correctness. Here we describe the form these checks can take and mention various software packages that incorporate them. We then describe the latest improvements to our own structure-checking program, PROCHECK, in terms of factors that provide compatible measures of deviation from normality for a wide range of different properties.

1. Introduction

The end result of protein structure determination, whether experimental – by X-ray crystallography or NMR spectroscopy – or from computational modelling, is a model of the structure comprising: the coordinates of the atoms, their occupancies and temperature factors. In the last few years several groups, including our own at UCL, have been investigating what sorts of tests might be applied to the coordinates to provide some evaluation of their 'quality'.

Of course, some measures of quality have been around for a long time. In X-ray crystallography, for example, the two most widely used measures are the resolution, which provides a measure of the quality of the X-ray data, and the R-factor, which provides a measure of how well the model fits that data. Other indicators of quality include the Luzzati plot (Luzzati, 1952), the σ_A plot (Read, 1986), the residue R factor (Jones *et al.*, 1991) and the 'free R value', or R_T^{free}, of Brünger (1992).

All these measures require the experimental data to hand. Additional tests, however, can be applied to just the coordinates themselves, and it is these tests that are of interest here. The aim of the tests is to evaluate various aspects of the stucture, such as its geometry, stereochemistry, and other factors that come from what is known of protein structure. Much of this knowledge comes from systematic analysis of the existing structures in the Protein Data Bank, the PDB (Bernstein *et al.*, 1977). In essence the PDB provides a store of information on what is 'normal' for a protein structure from which quantitative measures of acceptable standards can be derived and against which new structure models can be compared.

2. Factors of interest

There are many aspects of protein structure that might be evaluated, and these will be described briefly below. Some tests have already been incorporated into various software packages, and these packages will be reviewed in section 4.

a. Covalent geometry

The first, and most obvious, check is that of the structure's covalent geometry; namely, its bond lengths and bond angles. On its own, however, this check is unlikely to provide a good test of a protein's quality for two reasons. The first is that bond lengths and angles are usually restrained during refinement; that is, 'ideal', or 'target', values are used as restraints to supplement the experimental data. The targets have a heavy influence on the structure's final bond lengths and bond angles, as has been shown by Laskowski *et al.* (1993*b*). Furthermore, it is always possible to achieve 'good' bond lengths and bond angles, at the expense of the experimental data, simply by an appropriate choice of weights during the refinement process.

The second problem is that we simply do not know what the 'correct' values of these parameters, and their standard deviations, are for proteins. Nor is it possible to determine them, despite having so many structures in the PDB, for various reasons including the problem of restraints (Laskowski *et al.*, 1993*b*). The best that can be done is to use small-molecule data. The structures of small molecules provide sufficient experimental data to allow them to be refined without restraints, and so give truer values for the bond lengths and angles. However, whether the same values are applicable to much larger structures such as proteins is not known.

The latest set of 'target' values, obtained from small-molecule structures, is that of Engh & Huber (1991). Dictionaries of these new target values are now available for the most commonly-used refinement programs (Priestle, 1993).

Although a check of the covalent geometry does not provide a good independent test of a structure's quality, it still needs to be done. This is not only to identify any oddities and outliers, which could signify problem regions in the structure, but also to ensure that the restraints on these parameters have not been deliberately relaxed as a means of reducing the final R-factor. And indeed, the standard deviations of the bond lengths and angles are usually quoted when a structure is deposited in the PDB (the values typically being 0.02Å for bond lengths and 2° for bond angles).

b. Planarity

Certain groups of atoms are expected to be planar, such as the aromatic rings of Phe, Tyr, Trp and His, and the end-groups of Arg, Asn, Asp, Gln, Glu. Again, many refinement packages already restrain these groups to lie within a plane, so the check is not an independent one. Nevertheless it is still worth doing to detect any severe distortions.

c. Dihedral (torsion) angles

The dihedral angles - ϕ, ψ and ω for the mainchain, and the different χ torsion angles of the sidechains - have preferred conformations. For example, the specific, 'restricted' distribution of the ϕ–ψ values on the Ramachandran plot has been known for a very long time (Ramachandran *et al.*, 1963). In high resolution structures it has been observed that not only do more ϕ–ψ values lie in the Ramachandran 'allowed' regions but they actually concentrate into 3 smaller 'core' regions within them (Morris *et al.*, 1992). The tightness of the clustering of the ϕ–ψ's in these core regions can provide quite a good, independent measure of 'quality' (Morris *et al.*, 1992).

The distributions of the other dihedral angles can also provide measures of 'quality', particularly if they have not been restrained during refinement.

d. Non-bonded interactions

There are two main checks that can be applied to non-bonded interactions. The first is whether any non-bonded atoms clash - *ie* whether they penetrate one anothers' van der Waals radii. Both intra- and inter-molecular contacts need to be considered here. For structures solved by X-ray crystallography, one also needs to consider contacts between symmetry-related molecules in the crystal lattice.

The second type of check takes into consideration which types of interaction are actually favoured, and which are not. In other words, the check involves looking at an atom's non-bonded neighbours, and at how they are distributed around it, and then comparing the observed distribution with distributions extracted from the existing structures in the PDB. The directionality of the interactions can be taken into account by looking at atom distributions around whole sidechains, or just fragments of sidechains, rather than around single atoms.

e. Fragment fitting

Short fragments, say three to five residues in length, will also tend to have only certain favoured conformations due to steric and other considerations. This has long been used for loop-fitting, both in modelling of protein structures and for tracing the mainchain through initial electron density maps (*eg* Jones *et al.*, 1991), where the appropriate conformation for a given sequence with end-point constraints is selected from a search through the existing structures in the PDB. It might also, however, provide an additional check of the protein's overall correctness.

f. Polarity

A number of studies have shown the importance of a model having the correct distribution of polar residues on the surface and in the interior. The first was the study of Novotný *et al.* (1988) who generated two deliberately mis-folded structures: the sequence of the α-helical hemerythrin was modelled onto the β-sheeted backbone of the immunoglobin VL domain, and the sequence of the latter was modelled onto the backbone of the former.

The study showed that in the correctly folded structures the exposed sidechain nonpolar surface was as small as possible.

A similar study was that of Baumann *et al.* (1989), in which it was found that improperly folded proteins can be distinguished from properly folded ones by measuring the polar fraction of sidechains on the protein surface and, independently, in the protein interior.

More recently, the same group has derived a set of atomic solvation preference parameters (Holm & Sander, 1992) which can distinguish between correctly and incorrectly folded models. As with the other methods here, they are really only applicable to the protein model as a whole, and are less useful at identifying incorrect regions within the model.

g. 3D profiles and threading

Although it is not possible to predict the 3D structure of a protein from its sequence, it is possible to measure the 'compatibility' between a sequence and a 3D structure. This is achieved by evaluating the preferences of appropriate residues in the sequence to be in the particular environment found in the structure. It has been shown that a given amino acid when correctly 'threaded' onto its actual 3D structure gives a higher score than when the threading is incorrect (*eg* through mis-registration), or the threading is onto an incorrect structure (Hendlich *et al.*, 1990; Lüthy *et al.*, 1992; Jones *et al.*, 1992). In essence, these methods assess a combination of accessibilities, secondary structure preferences and non-bonded interactions.

In the paper by Lüthy *et al.* (1992) this concept was extended to provide a measure of a structure's correctness, both overall and locally. For checking the local threading, a moving window of 21 residues was scanned along the sequence and the 3D profile calculated for each window. This managed to locate some of the faulty regions in the structures tested, though not all; because of the averaging involved, some regions, particularly at the termini were not detected.

h. Hydrogen-bonds

A check of the hydrogen bonds in a structure can provide another important test of the 'quality' of a structure. A recent study (McDonald & Thornton, 1994) has shown that as the resolution of a structure improves the percentage of buried unsatisfied donor and acceptor atoms in a protein's interior becomes very small.

i. Other factors

An analysis of the typical accessibility of each residue type can again provide a measure of what is 'normal' for each residue. A given structure can then be assessed against these measures of normality.

High temperature-factors are well-know to be associated with flexible and poorly-defined regions of a structure, particularly on the surface of the protein. Yet no systematic analysis has yet been performed on the variability of B-values in proteins in the PDB. So it is not yet possible to say whether a given structure, as a whole, has unusually high or unusually low B-values.

Other factors that need to be taken into account describe different properties of the environment in which atoms, sidechains or fragments find themselves - *eg* the type of secondary structure or the solvent accessibility. Such parameters can be used to assess misfolded proteins (Bowie *et al.*, 1991) and, via the 3D profiling method mentioned above, have recently been applied to crystal structures (Lüthy *et al.*, 1992).

j. NMR structures

NMR structures present additional problems in that many coordinate sets are involved, rather than just a single one, and conformational variability between sets of coordinates is much greater than in X-ray structures. Therefore parameters specific to NMR structures need to be derived (see for example MacArthur & Thornton, 1993). Also, standard measures both of the quality of the data used for structure determination and of the fit of the

model to that data (equivalent to the resolution and R-factor in crystallography) need to be agreed.

3. Testing a factor as a measure of 'quality'

The factors described above can provide measures of a protein's quality at either a local, residue-by-residue level, or for the protein as a whole.

The 'normal' value of a given factor, or its typical distribution can be determined by analysing the existing structures in the PDB. The analysis should preferably be confined to a set of non-homologous, well-refined, high-resolution structures.

There are then two common ways of testing the usefulness of the factor as a measure of quality. The first is to apply it to a set of structures in the PDB and to see how the factor varies as a function of resolution or R-factor. If it is well-correlated with one of these it suggests that the factor is indeed measuring the quality of the structure in some way.

The second way to test it is to apply it to one of the existing structures that are known to have errors in them, such those cited in Lüthy *et al.* (1992). Some of these erroneous models have been corrected since their original deposition, so a check of how the factor performs against the correct and incorrect structures is a useful guide to how good it is at detecting serious flaws. Other possible test cases are structures that have deliberately been made erroneous, such as the mis-folded structures of Novotný *et al.* (1988) and Holm & Sander (1992).

4. Review of error-checking software

As mentioned above, some of the tests just described have been incorporated into software packages aimed at evaluating protein coordinates, and these will be briefly reviewed here. Of course, many of the tests are already an integral part of the refinement programs used in refining protein structures, but these will not be included here.

a. Covalent geometry and dihedral angles

The simplest of the tests are the checks on a protein's covalent geometry and its dihedral angles. These checks are incorporated into several packages.

The first is PAP (Callahan *et al.*, 1990) which provides ϕ–ψ plots, both as 2D Ramachandran plots and along the sequence of the chain, distance diagonal plots of C^{α}–C^{α} interactions, and average mainchain and sidechain temperature factors. The package includes a number of other useful utilities.

The program GEOM (Cohen, 1993) analyses the distances, angles and dihedral angles of all mainchain and C^{β} atoms, showing the r.m.s. deviations from the mean observed values in decreasing order of the size of the deviation. It also includes an analysis of sidechain dihedrals, comparing them with the table of rotamers of Ponder & Richards (1987).

Checks of covalent geometry and dihedral angles are included in the much wider variety of tests incorporated in our own PROCHECK package (Laskowski *et al.*, 1993*a*), which are based on the analyses of Morris *et al.* (1992), and in the CHECK option of WHAT IF, the sophisticated and versatile molecular modelling and drug design program of Vriend (1990).

b. Non-bonded interactions

The simplest checks of non-bonded interactions are tests for clashes between non-bonded atoms. Such checks are included in WHAT IF and PROCHECK. The former program now also tests for clashes between symmetry-related molecules by making use of the crystallographic transformations in the PDB file to generate the surrounding molecules (Hooft *et al.*, 1993).

A slightly more sophisticated approach involves looking at the distributions of different non-bonded atom interactions. It is well known that different sidechains adopt different preferred distributions relative to one another in three-dimensions (Singh & Thornton, 1992).

Two programs exploit such distributions for checking protein quality. The first, ERRAT (Colovos & Yeates, 1993), analyses the relative frequencies of non-bonded interactions between C, N and O atoms, where O represents both oxygen and sulphur atoms. For a given distance cut-off (typically 3.5Å) the number of CC, CN, CO, NN, NO and OO interactions are computed for residues in a sliding window of 9 residues in length and are compared with the frequencies obtained from a high-resolution data set of existing structures. For each 9-residue window the difference between the observed and expected frequencies is evaluated using a Gaussian error function, and the resultant score can show mis-traced or mis-registered regions.

The second method is that of Vriend & Sander (1993), and is incorporated in the WHAT IF package. Here the distributions of different atom types around different sidechain fragments have been obtained from a high-resolution data set. To check a given structure, the distributions obtained from it are compared with the expected distributions from the high-resolution data by taking the product of the actual and expected distributions so that only contacts that occur in both the protein and the database contribute to the quality index. The resultant quality factor is a good measure of overall quality, though the method still needs to take into account crystal neighbours and water contacts.

c. Polarity

The work of Baumann *et al.* (1989) has already been mentioned in section 3*g* above, and their program for testing for mis-folding in proteins is called POL_DIAGNOSTICS_88.

5. PROCHECK

In this final section, we will discuss one of the latest enhancements to our own package, PROCHECK (Laskowski *et al.*, 1993*a*). PROCHECK uses a number of stereochemical parameters found by Morris *et al.* (1992) to be good measures of protein quality. It also uses the bond lengths and bond angles of Engh & Huber (1991).

The output of PROCHECK comprises a number of plots which give an at-a-glance view of possible problem areas in the structure, as well as several assessments of the protein's overall quality. A detailed residue-by-residue listing is also provided for further analysis and investigation of problem areas.

The program has now been incorporated in the CCP4 suite of programs, and so is widely available to the crystallographic community (CCP4 Manual v.2.2, SERC Daresbury Laboratory, 1993). The most up-to-date version is always available free of charge from the authors by sending an e-mail request to roman@bsm.bioc.ucl.ac.uk.

Latest amendments - *G*-factors

The PROCHECK package has undergone a number of improvements and additions since the version described in the original paper. The very latest addition represents an attempt at integrating the different measures of geometric quality into a single measure, or *G*-factor.

At first glance it is difficult to see how such an integration might be possible, as the existing measures appear far too disparate to be combined in any meaningful way.

The approach we have adopted is to use a consistent measure of how each parameter deviates from 'normality'. Take as an example the distribution of ϕ–ψ values on the Ramachandran plot. A separate distribution is first generated for each of the 20 residue types. The data set used comprises 163 non-homologous, high-resolution protein chains chosen from structures solved by X-ray crystallography to a resolution of 2.0Å or better and an *R*-factor no greater than 20%. No two of the 163 chains share a sequence homology greater than 35%, and atoms in the chains with zero occupancy are excluded from the analysis.

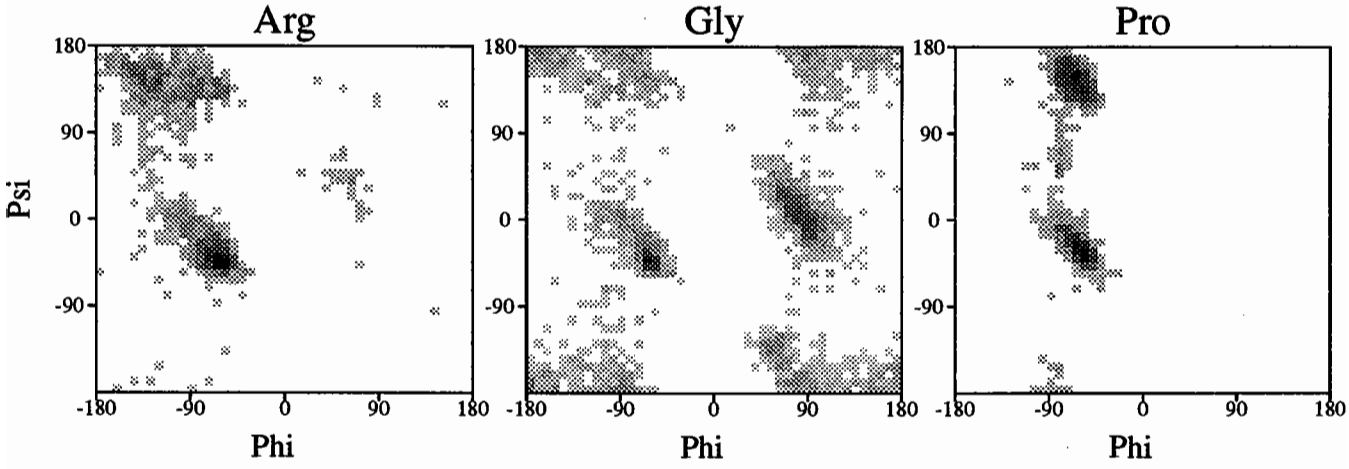


Figure 1. Three of the 20 Ramachandran plots derived from the non-homologous, high-resolution data set of 163 proteins. The plots shown are for Arg, Gly and Pro residue types. The darker the shading the more favourable the ϕ–ψ combinations.

Each Ramachandran plot is divided into 45 × 45 cells (Figure 1) and the numbers of observations in each cell stored. From the observations, it is possible to calculate the probability of a given residue type having a given ϕ–ψ combination.

However, rather than use the probabilities themselves, a log-odds score is calculated for each cell by taking the natural logarithm of the cell's probability. The rationale for doing this is akin to the conversion of probability density functions to potentials of mean force via the Boltzmann distribution in the method of Sippl (1990). Log-odds scores can be summed, rather than multiplied like probabilities, so taking meaningful averages becomes possible.

The 45 × 45 log-odds scores for each of the 20 Ramachandran plots can be applied to assess the ϕ–ψ values of a given protein structure, and so give a *G*-factor for these dihedrals, both for each residue and for the protein as a whole. Residues with low scores are in low-probability conformations. If the protein has many of these, then it suggests that something may be amiss with its overall geometry.

Practical considerations

Two practical aspects need to be considered when computing the log-odds scores. The first is when a given cell in one of the Ramachandran plots contains no observations (*ie* there is no such ϕ–ψ combination in the data set used to derive the distribution). Clearly, the probability calculated for the cell will be zero, and its corresponding log-odds score minus infinity.

To overcome this problem, the observations in adjoining cells are taken into account; the probability is calculated for a group of cells and then averaged over the number involved. The actual number of adjoining cells taken into account is such that they encompass at least 10 observations.

The second of the practical considerations concerns the combining of the log-odds scores together - for example, to calculate an average score for the protein as a whole. Different residue-types will have different Ramachandran plots, so the log-odds scores derived for each residue-type need to be normalised before they can be averaged over the whole protein.

To see why this is necessary, consider the difference between the Ramachandran plot for a Gly residue and that for any other residue (Figure 1). The Gly distribution is more spread out. Each cell will have a low probability and so return a low log-odds score. The other distributions exhibit significant clustering in the favourable regions. The cells in these regions will have a relatively high probability and return high log-odds scores. Thus each distribution will return a different range of log-odds scores depending on how the observations cluster together. This can be corrected for by normalising the scores returned by each distribution to make them comparable.

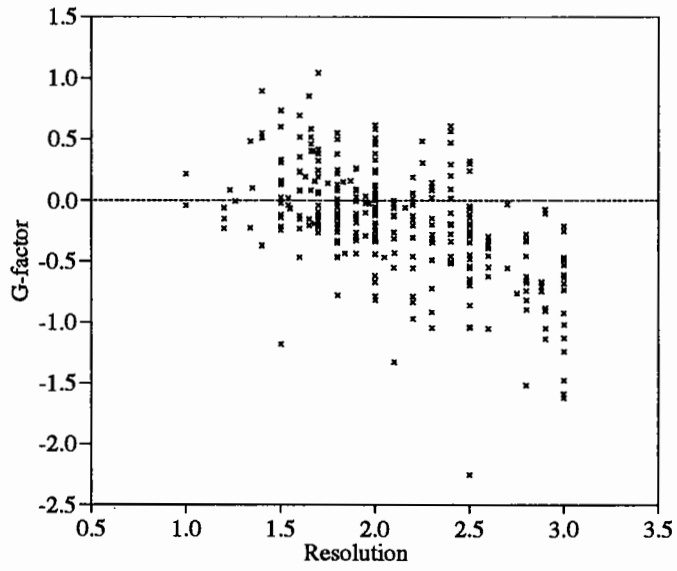


Figure 2. Plot of $G_{\phi-\psi}$-factor versus resolution for 295 non-homologous protein chains, showing the correlation between *G*-factor and resolution.

Empirical tests of the *G*-factors

Figure 2 shows the ϕ–ψ *G*-factors for 295 non-homologous protein chains plotted as a function of resolution (Figure 2). As can be seen, the $G_{\phi-\psi}$-factors show a tendency to be higher the better the resolution. In other words, as the resolution improves, and the quality of the structure gets better as a result of the better observational data, the $G_{\phi-\psi}$-factors

tend to increase also. This supports the idea that they may indeed provide as useful a measure of geometrical quality.

The same approach, when applied to χ_1–χ_2 distributions, gives the correlation shown in Figure 3 for $G_{\chi_1-\chi_2}$.

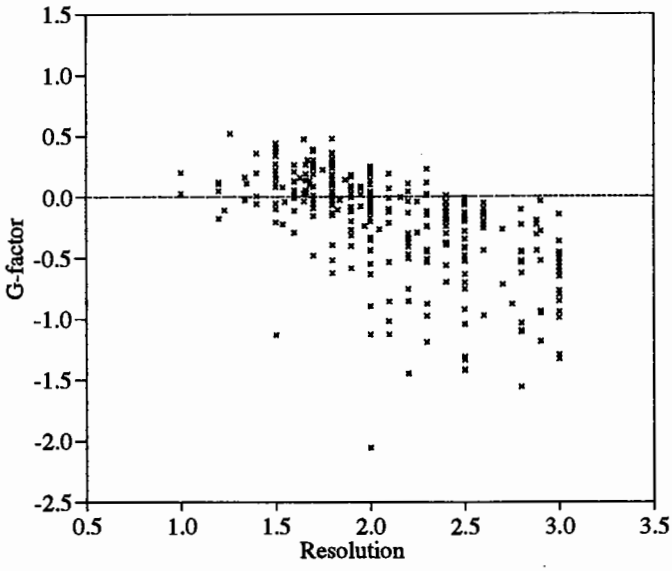


Figure 3. Plot of $G_{\chi_1-\chi_2}$-factor versus resolution for 295 non-homologous protein chains. As in Figure 2, there is a clear correlation between the *G*-factor and the resolution.

The ϕ–ψ and χ_1–χ_2 distributions are two-dimensional distributions, but the approach can of course also be applied in the one-dimensional case. Thus, for example, for residues that do not have a χ_2 torsion angle, their χ_1 distributions alone are used.

To date we have also used the distributions of χ_3, χ_4 and ω torsion angles. Putting all the *G*-factors from these different dihedral angle distributions together, it is possible to compute an overall *G*-factor, or G_{dih}, for a protein's dihedral angles.

Further application of the method has been made to compute a *G*-factor that covers a protein's covalent geometry, or G_{cov}. So far this has only involved the 5 main-chain bond lengths (C–N, C–O, C^{α}–C, C^{α} –C^{β}, N–C^{α}) and the 7 main-chain bond angles (C–N–C^{α}, C^{α}–C–N, C^{α}–C–O, C^{β}–C^{α}–C, N–C^{α}–C, N–C^{α}–C^{β}, O–C–N), but may eventually be extended to cover sidechain bond lengths and angles as well.

The approach can be equally well used for other properties, such as a G_{acc} for accessibilities, a G_{ss} for secondary structure preferences, a G_{nb} for non-bonded interactions, and so on.

Conclusion

A number of tests have recently been developed by several groups for evaluating the co-ordinates of protein structures. These tests use the information stored in the PDB to determine what represents 'normality' for protein structures against which to assess any new structures. A number of possible areas remain for investigation. One area that our group is currently investigating is the development of '*G*-factors' for various parameters using log-odds scores as measures of deviation from normality as a means of combining different measures together. These *G*-factors are being added to the PROCHECK package as and when they are developed.

References

Baumann G, Frömmel C & Sander C (1989). Polarity as a criterion in protein design. *Protein Engineering*, **2**, 329–334.

Bernstein F C, Koetzle T F, Williams G J B, Meyer E F Jr, Brice M D, Rodgers J R, Kennard O, Shimanouchi T & Tasumi M (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Bowie J U, Lüthy R & Eisenberg D (1991). A method to identify protein sequences that fold into a known 3-dimensional structure. *Science*, **253**, 164–170.

Brünger A T (1992). Free *R* value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, **355**, 472–475.

Callahan T, Gleason W B & Lybrand T P (1990). PAP: a protein analysis package. *J. Appl. Cryst.*, **23**, 434–436.

Cohen G H (1993). GEOM - a program to assess the reliability of a protein model. *J. Appl. Cryst.*, **26**, 495–496.

Colovos C & Yeates T O (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Science*, **2**, 1511–1519.

Engh R A & Huber R (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Cryst.*, **A47**, 392–400.

Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G & Sippl M J (1990). Identification of native protein folds amongst a large number of incorrect models - the calculation of low-energy conformations from potentials of mean force. *J. Mol. Biol.*, **216**, 167–180.

Holm L & Sander C (1992). Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.*, **225**, 93–105.

Hooft R W W, Sander C & Vriend G (1993). Reconstruction of symmetry related molecules from Brookhaven PDB files. *Preprint.*

Jones D T, Taylor W R & Thornton J M (1992). A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Jones T A, Zou J-Y, Cowan S W & Kjeldgaard M (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Cryst.*, **A47**, 110–119.

Laskowski R A, MacArthur M W, Moss D S & Thornton J M (1993*a*). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, **26**, 283–291.

Laskowski R A, Moss D S & Thornton J M (1993*b*). Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.*, **231**, 1049–1067.

Lüthy R, Bowie J U & Eisenberg D (1992). Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.

Luzzati P V (1952). Traitement statistique des erreurs dans la determination des structures cristallines. *Acta Cryst.*, **5**, 802–810.

MacArthur M W & Thornton J M (1993). Conformational analysis of protein structures derived from NMR data. *Proteins*, **17**, 232–251.

McDonald I K & Thornton J M (1994). Satisfying hydrogen-bonding potential in proteins. *J. Mol. Biol.*, submitted.

Morris A L, MacArthur M W, Hutchinson E G & Thornton J M (1992). Stereochemical quality of protein structure coordinates. *Proteins*, **12**, 345–364.

Novotný J, Rashin A A & Bruccoleri R E (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins*, **4**, 19–30.

Ponder J W & Richards F M (1987). Tertiary templates for proteins - use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, **193**, 775–791.

Priestle J P (1993). Standard stereochemical dictionaries for protein structure refinement and model building, in *CCP4 and ESF-EACBM Newsletter*, **29**, Daresbury Laboratory, Warrington, UK.

Ramachandran G N, Ramakrishnan C & Sasisekharan V (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7**, 95–99.

Read R J (1986). Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Cryst.*, **A42**, 140–149.

Singh J & Thornton J M (1992). *Atlas of Protein Side-Chain Interactions*, IRL Press, Oxford.

Sippl M J (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, **213**, 859–883.

Vriend G (1990). WHAT IF: a molecular modeling and drug design program. *J. Mol. Graphics*, **8**, 52–56.

Vriend G & Sander C (1993). Quality control of protein models: directional atomic contact analysis *J. Appl. Cryst.*, **26**, 47–60.

DL/SC1/R35
85 A4 pieces.
Yellow card cover
Stapled together as a book.