

CCP4 NEWSLETTER ON PROTEIN CRYSTALLOGRAPHY

An informal Newsletter associated with the BBSRC Collaborative
Computational Project No. 4 on Protein Crystallography.

Number 38

April 2000

Contents

1. **News from CCP4**
Peter Briggs, Martyn Winn, Sue Bailey, Alun Ashton, David Brown, Charles Ballard
2. **Report on the CCP4 Summer School at St. Andrews**
Jim Naismith
3. **Comparison of SURFACE and AREAIMOL for accessible area calculations**
Peter Briggs
4. **Report on the CCP4 Study Weekend, York, January 2000**
Julie Wilson
5. **Experiences from my CCP4 scholarship 2000**
S. Ravichandran
6. **Summaries of recent discussions on ccp4bb**
Maria Turkenburg
7. **Gaussian Likelihoods in real and reciprocal space**
Kevin Cowtan
8. **Modelling of Disordered Solvent in Macromolecular Crystals**
A. Urzhumtsev
9. **CCP4i as a Project Management Tool**
Peter Briggs
10. **mmCIF in the CCP4 Suite**
Martyn Winn

Editor: Peter Briggs

Daresbury Laboratory, Daresbury,
Warrington, WA4 4AD, UK

NOTE: The CCP4 Newsletter is not a formal publication and permission to refer to or quote from the articles reproduced here must be referred to the authors.

News from CCP4: April 2000

[Peter Briggs](#), Martyn Winn, Sue Bailey, Alun Ashton, David Brown, Charles Ballard

1. Staff changes

There have been two changes to the CCP4 staff since the last newsletter. At the end of October 1999 we said farewell to Sheila Peters, who left Daresbury for the more temperate climates of Manchester. Sheila was only with us for a short time, and we wish her well in her new job.

However, as one door shuts so another one opens ... at the beginning of March the Daresbury staff was joined by Dr Charles Ballard. Charles was working as a Computational Chemist in Japan prior to taking up his current post here as a scientific programmer.

2. Workshops and Conferences

BCA Summer School at St Andrews: the last summer school took place in August 1999 and was organised by Gerry Taylor and Jim Naismith. The week-long intensive course on Protein Crystallography was aimed at students and young postdocs, and was funded in part by CCP4. For details see the report by Jim Naismith in this newsletter. The next BCA summer school will be in Bristol, this September - see http://www.bch.bris.ac.uk/pxss/sum_school.html for more details.

Developers seminars: December 2000 saw the first CCP4 Developers Seminar, organised by Garib Murshudov, which was aimed at program developers and focused on developments in methodology. Although the meeting was deliberately small and low-key, it is hoped that this will evolve in a regular series of meetings to encourage discussion about various issues. More information to follow.

Study Weekend 2000: This years CCP4 study weekend was held at the Univeristy of York and the topic - Low Resolution Phasing - proved as popular as ever with some 350 delegates attending. CCP4 would like to thank especially the scientific organisers - Julie Wilson (York), Jonathan Grimes (Oxford) and Helen Saibil (Birbeck) - and of course all the speakers. The proceedings from the study weekend will be published later this year in *Acta Cryst. D*. For more details of this year's goings-on, see the [report by Julie Wilson](#), elsewhere in this issue.

The next study weekend (on January 5/6th 2001) will also be held at York. The topic will be "Molecular Replacement and its relatives". The organisers are Jim Naismith (St-Andrews) and Kevin Cowtan (York). If you have any ideas for speakers or topics to be covered they would be delighted to hear from you (so I'm told!). Their email addresses are naismith@st-andrews.ac.uk and cowtan@ysbl.york.ac.uk. Further announcements will follow so keep an eye out on the CCP4 Bulletin Board.

Finally, the CCP4 staff will also be making trips to conferences and workshops this summer - look out for us at the **ACA meeting in St. Paul Minnesota** in July (see <http://nexus.hwi.buffalo.edu/ACA/ACA-Annual/StPaul/StPaul.html>), and the **ECM 19 meeting in Nancy** in August (see <http://www.lcm3b.u-nancy.fr/ecm19/>).

3. New Release 4.0

Mid-January saw the release of version 4.0 of the CCP4 suite, followed shortly after by the patch release 4.0.1. As always the patch release is intended only to fix minor bugs in the release, and if you are already using 4.0 without any problems then we don't recommend you bothering to upgrade.

The major changes from 3.5 to 4.0 are:

- **CCP4i - CCP4 graphical user interface** is now an integral part of the suite, and will no longer be distributed separately.
- **Automatic Data Harvesting:** the programs SCALA, TRUNCATE, MLPHARE, REFMAC and RESTRAIN now produce data harvesting files for ultimate structure deposition. Details of the harvesting concept can be found in Martyn Winn's article in the newsletter 37.

There are also a number of new programs:

- **MOLREP:** automated program for molecular replacement, from Alexei Vagin
- **OASIS:** program for breaking phase ambiguity in OAS or SIR, from Quan Hao
- **RWCONTENTS:** promoted from unsupported.
- **SC:** program to analyse shape complementarity of molecular interfaces
- **T_SHIFT:** part of ARP/wARP suite v5.0.

Other highlights include:

- Major new versions of **AMORE, RASMOL, SFCHECK, ARP_WARP** (renamed from **ARPP**)
- New library libccif from Peter Keller for reading/writing **mmCIF** files, held in \$CCP4/lib/ccif (see Martyn Winn's article in this newsletter)
- Other new libraries **harvlib** (for data harvesting) and **cciflib** (plays a role similar to rwbrotok for mmCIF files)
- Updated **xdl_view** library
- More general documentation, see \$CHTML/INDEX.html

Another new development is the limited provision of precompiled binaries for CCP4 4.0. These are only available for IRIX (32- and 64-bit versions, prepared on IRIX6.5 R10k) and alpha (prepared on Digital UNIX V4.0F). These must be downloaded **in addition** (not instead of) the normal CCP4 distribution, and you should read the enclosed BINARY.readme file.

As always, details of all the changes can be found in the CHANGES file in the top-level directory (\$CCP4), and in \$CCP4/html/CHANGESinV4_0.html. We also urge people to check the CCP4 Problems Pages before reporting any bugs (with fixes, if possible!) to ccp4@dl.ac.uk.

4. Other News

As the start of CCP4's 21st anniversary year, January 2000 saw the award of the first CCP4 "travel scholarships", providing money for crystallographers (mainly from developing countries) to attend the annual CCP4 Study Weekend and then to visit crystallography labs around the UK. The scheme is seen as a way of encouraging further international collaboration in the Protein Crystallography community.

This year three scholarships were awarded, to Mr S. Ravichandran (Saha Institute of Nuclear Physics, Calcutta), Dr Tao Jiang (Institute of Biophysics, Academia Sinica, Beijing) and Professor Li-wen Niu (University of Science & Technology of China, Hefei). It is expected that the scheme will run for a number of years.

We are grateful to Ravi for recounting some of his very positive experiences this year in his article, elsewhere in this newsletter.

Report on the CCP4 Summer School at St. Andrews

Organisers: Garry Taylor and Jim Naismith, University of St. Andrews

The participants arrived on St. Andrews on Sunday the 23rd August (except a few latecomers from a minster city) for a week long intensive course on Protein Crystallography. In the end 26 students from all over the UK attended the School, whittled down from over 40 applicants. The work started 9.00am Monday with Garry Taylor welcoming the students to the new Biomolecular Sciences Building in St. Andrews. To squeeze everything in, lectures came in 1 1/2 hour blocks, with a 1/2 hour coffee break or lunch between each block. Steve Wood (Southampton) and Phil Jackson (Perseptive) dealt in some depth with protein expression, analysis and crystallisation. This was well received by the students as majority of them are struggling with the molecular biology rather than X-ray analysis in their projects. At the ice breaker on Monday evening, 10 minutes to tell it like it is, some students had taken heart from that days lectures.

Elsbeth Garman (Oxford), stole the show with her 'how to do it' guide to data collection and cryocrystallography. Rumours that she was starting an alternative and complimentary crystallography were unfounded when she confessed that the acupuncture needles were for moving crystals not stabbing bacteria or SF9 cells. A couple of students though they had some merit in supervisor control. The University provided a PC classroom which enabled all the students to actually play around with MOSFLM (thanks to Martin Noble and Andrew Leslie). The St. Andrews computers groaned under the strain but it was good that each student got a chance to try to process some data. Jim Naismith tried to explain to the students what CCP4 is and why they should support it. Liz Potterton (CCP4, York) got the new CCP4 GUI running and the students got a chance to play around with it. It was interesting that many of the students had no experience with computer operating systems except Windows. Tuesday evening everyone unwound by going ten pin bowling, smashing skittles over seemed a good way to escape from reciprocal space.

Garry Taylor popped up again to tell them what structure factors are, how to calculate and solve Pattersons and how to find non-crystallographic symmetry. Ian Tickle (Birbeck) told us in depth how we refine heavy atoms, some pitfalls and some tips. He explained how the derivatives lead to phases and the role of probability in deciding what phase to use. Bill Hunter (Dundee) appeared and urged everyone to go MAD. Many of the students came away energised to try their hand at this increasingly routine technique. Wednesday evening, subtitled at home with Struther Arnott was the conference dinner. Principal Arnott, (the Vice-chancellor of the University to those south of the Tweed) hosted a champagne reception at his house and acted as a tour guide. He had arranged a visit to the University Chapel (which dates from the 15th Century) and gave an amusing account of the University's turbulent history since its foundation in 1410. At the School dinner, he spoke about his own experience as a student in diffraction which involved sleeping inside the computer at night to check the valves kept working.

On Thursday morning, Garry Taylor delved into averaging and phase extension. Rupert Russell (St. Andrews) ran a practical on molecular replacement, again one student per computer. The final day Susan Crennell (Bath) gave an example of and ran a practical on

electron density fitting. The students had to split into teams for this, as St. Andrews did not have one graphics workstation per student (funding agencies take note). For some students this was their first glimpse of electron density in all its glory. In the afternoon, Jim Naismith talked about refinement, some pitfalls and some recipes. The light hearted tone was continued and expanded upon by Pete Artymuik (Sheffield) who discussed validation, deposition and what a structure actually tells you. He produced some pretty incriminating photographic evidence of style deficiency at Birbeck in the early 80's.

A survey of the students revealed all found the course very useful and informative. All the students found it very hard going and a lot of material to cope with. Many expressed a desire to spend another week. The residences were thought to be comfortable and the lecture theatre good.

Garry and Jim are very grateful to all the tutors and lectures. They pulled out all the stops preparing detailed handouts and giving well thought out lectures. All lectures were highly rated by the students. Behind the scenes Margaret Wilson (St. Andrews) provided outstanding secretarial support, photocopying, typing and phoning.

The School costs money to run. Nearly all students were completely financed including travel costs. The bulk of the financial support comes from CCP4. The BBSRC contributed generously also and the MRC paid for their students to attend. The School was fortunate to have had generous support from Perkin Elmer Biosystems, MAR, Brucker, Pfizer, Molecular Structure Corporation, Unilever and Zeneca. The University of St. Andrews was generous with its in kind donations of resources. The next school in St. Andrews is 2001. The next CCP4 summer school is in Bristol in 2000.

Jim Naismith

Comparison of SURFACE and AREAIMOL for accessible surface area calculations

Peter Briggs
CCP4

1. Introduction

The CCP4 program suite contains two programs for calculating the solvent accessible surface area (ASA) of macromolecules: SURFACE and AREAIMOL. The two programs use different algorithms for the area calculations and offer different ranges of functionality. In particular, AREAIMOL has recently undergone a number of changes which have enhanced its functionality and usability.

This article examines and compares the two programs.

2. Definition of Solvent Accessible Surface Area

The concept of the solvent accessible surface of a protein molecule was originally introduced by Lee and Richards (1971), as a way of quantifying hydrophobic burial. The solvent accessible area (ASA) describes the area over which contact between protein and solvent can occur.

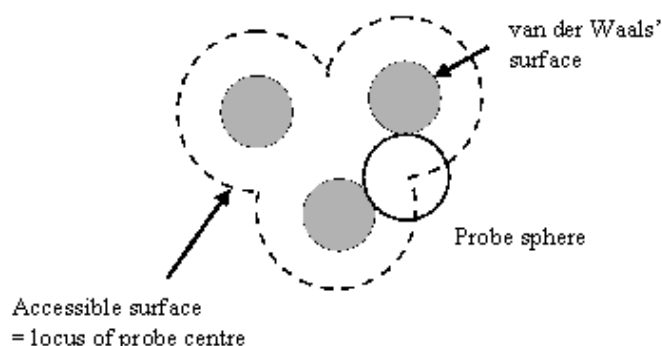


Figure 1: accessible surface of a molecule, defined as the locus of the centre of a solvent molecule as it rolls over the Van der Waals surface of the protein.

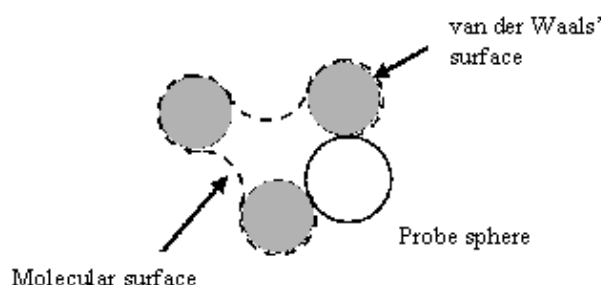


Figure 2: molecular surface of a molecule, defined as the locus of the inward-facing probe sphere.

The solvent accessible surface is defined as the locus of the centre of a probe sphere (representing the solvent molecule) as it rolls over the Van der Waals surface of the protein (illustrated for a simple case in figure 1). It is important to realise that this is different to the molecular surface, which is defined as the locus of the inward-facing probe sphere (Richards 1977) (illustrated in figure 2).

The original motivation for the calculation of the accessible surface was the study of the protein-folding problem and hydrophobicity. The size of the solvent accessible surface area buried in an interaction between protein units can be used to discriminate between crystal packing and a functional protein-protein interaction (used for example in the EBI's Quaternary Structure File Server (PQS), see <http://pqs.ebi.ac.uk/pqs-doc/pqs-doc.shtml>).

3. Comparison of the Programs

There are two main differences between AREAIMOL and SURFACE which can impact on the estimates of ASA which they produce. The first is that they use different algorithms to calculate ASA; this is examined in section 3.1 below. The second difference is the choice of van der Waals radii assigned to non-hydrogen atoms in the calculations, and this is examined in section 3.2.

Other differences are more to do with the ranges of functionality offered by the programs, and are discussed briefly in section 3.3. These differences shouldn't have an effect on the ASA estimates.

Finally, it is worth mentioning a couple of the similarities between the two programs. Both programs have parameters which control the precision of the area calculation (ZSTEP in SURFACE, PNTDEN in AREAIMOL). In either case, higher precision means relatively longer running times (though this is not a significant overhead in practice). Both programs also have a PROBE keyword, used to define the radius of the probe (solvent) molecule. It is usual to assume that accessibility to water is being assessed, and so in these tests the standard radius of 1.4 Å was used (the default for both programs).

3.1. Comparison of the algorithms

In all cases the solvent accessible surface can be obtained by drawing a sphere around each atomic position with a radius equal to the van der Waals' radius of the atom plus the radius of the probe sphere (referred to as the *expanded atom radius*). The union of the expanded atoms is then the solvent-excluded volume.

The original Lee and Richards algorithm for calculating ASA, implemented in the program SURFACE, effectively "slices" the expanded atom volume into 2-dimensional cross-sections along a single direction. Each slice contributes a surface area equal to the perimeter of the cross-section multiplied by the spacing between the slices. In SURFACE the keyword ZSTEP controls the fineness of the slices (values vary between 0.1 for most accurate, to 0.5 being the coarsest - default is 0.25).

The AREAIMOL program uses a different method, after Shrake and Rupley (1973). Dots are placed on the expanded atom surfaces, with a fixed number of dots per unit area. Dots which are not inside any other expanded sphere are considered to be accessible to solvent. The number of accessible dots multiplied by the density of the dots gives the accessible area on each atom. The keyword PNTDEN defines the number of points generated per square Angstrom on the surface (values go from the default of 1 to 100).

A number of simple tests were performed to assess the algorithms and compare their performance, as implemented in the two programs.

1. Calculations of the surface area of a single sphere (i.e. single atom), and the surface area for a pair of spheres separated by a distance less than the sum of their expanded radii (i.e. two close-contacting atoms).

The values obtained from the programs for different values of ZSTEP/PNTDEN are given in tables 1 and 2, and in each case are compared with the analytical values, which can be calculated directly.

SURFACE		AREAIMOL	
ZSTEP	ASA	PNTDEN	ASA
0.5	116.9	1	117.0
0.25	121.7	10	116.9
0.1	120.8	100	116.89
<i>Theoretical</i>	120.76	<i>Theoretical</i>	116.90

Table 1: ASA (in Å²) for a single nitrogen atom, calculated by SURFACE and AREAIMOL and compared with the exact theoretical values. The different values are due to SURFACE using an expanded radius of (1.70+1.4) Å and AREAIMOL using (1.65+1.4) Å.

SURFACE			AREAIMOL		
ZSTEP	A1	A2	PNTDEN	A1	A2
0.5	90.5	82.9	1	89.0	86.0
0.25	95.9	76.9	10	89.8	85.1
0.1	94.7	78.0	100	89.46	85.29
<i>Theoretical</i>	94.73	78.03	<i>Theoretical</i>	89.51	85.29

Table 2: comparing the output of SURFACE and AREAIMOL with the theoretical value for the ASA (in Å²) of a pair of atoms, oxygen (atom 1) and nitrogen (atom 2), separated by distance 3.15 Å. "A1" and "A2" refer to the ASA assigned to the two atoms individually; the total ASA would be the sum.

- Calculations of the ASA of single residues and a single chain. Here analytical values are not available, so the estimates from each program are compared directly, again using different values of ZSTEP/PNTDEN, with atomic coordinates taken from ribonuclease from *Streptomyces aureofaciens* (RNase Sa) (Sevcik et al, 1996). Note that SURFACE was modified to use the same set of van der Waals' radii as AREAIMOL.

	SURFACE		AREAIMOL	
	ZSTEP	ASA	PNTDEN	ASA
a) VAL2 residue in RNase Sa	0.5	265.5	1	261.0
	0.25	264.3	10	264.5
	0.1	265.0	100	264.8
b) SER3 residue in RNase Sa	0.5	233.5	1	236.0
	0.25	231.9	10	233.3
	0.1	234.5	100	234.5
c) chain "A" in RNase Sa	0.5	5593.0	1	5587.0
	0.25	5593.6	10	5565.5
	0.1	5571.6	100	5570.2

- Table 3:** comparing the estimates of ASA (in Å²) output by SURFACE and AREAIMOL for a) VAL2 residue in RNase Sa, b) SER3 residue in RNase Sa, c) chain A of RNase Sa (containing 751 atoms).

Comments and Analysis

There seems to be a generally held opinion that "surface dot" methods such as that used in AREAIMOL are somehow inherently less accurate than those using the original Lee and Richards method (see for example Connolly (1996)).

This is not supported by these results. In the case of single/pairs of atoms, the ASA estimates from the two programs (obtained using the highest values of ZSTEP/PNTDEN) differ from the appropriate theoretical values by less than 0.1% in each case. In the larger examples the estimates of the ASA from SURFACE and AREAIMOL also differ from each other by less than 0.1%.

In fact this is much smaller than the variation due to using different values of PNTDEN/ZSTEP - in the worst cases, around 3-6% for different values of ZSTEP in SURFACE and 1.4% for different values of PNTDEN in AREAIMOL.

Finally, it is interesting to examine the extra time associated with running the programs using higher values for ZSTEP/PNTDEN. Table 4 gives a summary of the elapsed times for SURFACE and AREAIMOL when calculating ASA for chain A of RNase Sa:

SURFACE		AREAIMOL	
ZSTEP	Elapsed time (s)	PNTDEN	Elapsed time (s)
0.5	4.0	1	0.5
0.25	7.2	10	1.4
0.1	17.2	100	11.0

Table 4: elapsed times running SURFACE and AREAIMOL for ASA of chain A of RNase Sa, on an SG Origin 2000 with IRIX 6.5.

Using the highest values of ZSTEP/PNTDEN significantly increases the running time relative to using lower values - but compared to the time taken to run (for example) a refinement job, these times are negligible. It would certainly be realistic to use the programs with ZSTEP=0.1/PNTDEN=100 as the defaults.

3.2. Effects of using different van der Waals' radii (SURFACE)

Hydrogen atoms are not considered individually in the calculations (because it is not usual for them to be included in coordinate files describing the protein atom positions). The van der Waals' radii used for non-hydrogen atoms are therefore modified to account for the implicit presence of hydrogens. The choice of van der Waals' radii used in the calculations would also be expected to have an effect on the estimates of ASA.

SURFACE has two sets of radii taken from the literature: the first from Lee and Richards' original paper, and the second from Chothia (1975). The values are summarised in the table below:

Lee and Richards

Main-chain alpha carbon	1.70 Å
Main-chain carbonyl oxygen	1.52 Å
Main-chain amide NH group	1.55 Å

Chothia

Oxygen	1.40 Å
Trigonal nitrogen	1.65 Å
Tetrahedral nitrogen	1.50 Å

Main-chain carbonyl carbon	1.80 Å	Tetrahedral carbon	1.87 Å
All side-chain atoms and groups	1.80 Å	Trigonal carbon	1.76 Å
		Sulphur	1.85 Å
		Water	1.40 Å

The Lee and Richards values are the default in SURFACE, and are selected using the VDWR RICH keywords; Chothia's values are selected using the VDWR CHC keywords. (AREAIMOL uses a simpler scheme where the same van der Waals' radius is assigned to all carbons, all oxygens and all nitrogens, regardless of the number of associated hydrogens.)

To investigate the effects of using different sets of van der Waals' radii, ASA was calculated using SURFACE with either VDWR RICH or VDWR CHC, for the residues and molecule used in the previous examples. The results are given in table 5, along with previous calculations using the same set of radii as AREAIMOL. ZSTEP was set to 0.1 in all the calculations.

	a) VAL2 residue	b) SER3 residue	c) Chain A
SURFACE with VDWR RICH	272.8	233.7	5591.6
SURFACE with VDWR CHC	263.0	225.7	5570.2
SURFACE with values from AREAIMOL	265.0	234.5	5571.6
AREAIMOL	264.8	234.5	5570.2

Table 5: effect of using different van der Waals' radii on the ASA (in Å²) estimated by SURFACE in the case of a) VAL2 residue in RNase Sa, b) SER3 residue in RNase Sa, c) chain A of RNase Sa. Values from AREAIMOL are also given for comparison.

Comments and Analysis

In the examples considered, the percentage differences in ASA due to using different sets of van der Waals radii seem to decrease with increasing numbers of atoms considered - for the two residues considered the range of values differed by 3-4%, for the single molecule they differed by less than 1%. Interestingly, these are smaller variations than those due to changing ZSTEP (see the results in section 3.1).

An effect of different van der Waals' radii which was not considered here was the impact on the ASA of an individual atoms within a residue or molecule. An atom with a small ASA under one scheme might have zero ASA under the other (i.e. being buried completely by its neighbours). Whether this is important depends very much on the situation under investigation.

3.3. Differences in functionality ranges

Ultimately the choice of program to use will be most likely be dictated by the problem the user is looking at.

SURFACE offers a powerful method for selecting which atoms are included in the calculations. The user is able to specify which atoms are to have their area calculated,

which are to be included (so that they can "obscure" accessible area on other atoms, without their own area being considered), and which are to be excluded entirely (and thus play no part at all in the calculations). The option to use different sets of van der Waals radii has already been discussed in section 3.2 above. The output of SURFACE is used to prepare input for the VOLUME program.

On the downside, SURFACE (or rather, its users) suffers from an anachronistic input procedure and the program output is somewhat limited. It is also unable to directly examine crystallographic contacts or area differences.

AREAIMOL offers a very different range of functions. Most importantly, it is able to examine the effects of intermolecular and crystallographic contacts on the ASA, and can be used to look at ASA differences directly in a number of different situations. It also offers a fairly comprehensive analysis of the ASA values, breaking them down by residue, chain and molecule. However there are equivalents to SURFACE's atom selection options, and a much simpler assignment of van der Waals radii is employed.

The functions offered by each program are summarised in table 6.

Function	SURFACE	AREAIMOL
Area on individual atoms?	yes	yes
Atom selection?	yes	no
Different VDWR radii?	yes	no
Intermolecular/crystallographic contacts?	no	yes
ASA differences?	no	yes
Analysis of output?	no	yes
Prepares input for VOLUME?	yes	no

Table 6: Summary of functions offered by SURFACE and AREAIMOL (see main text for fuller descriptions).

4. Conclusions and comments

A small number of tests have been performed to examine the estimates of ASA obtained from the CCP4 programs AREAIMOL and SURFACE. Two important factors which are held to affect ASA estimates are 1) the different algorithms used by the programs, and 2) the choice of modified van der Waals radii used in the calculations (see Connolly (1996)).

From these examples it seems that both algorithms are capable of giving good estimates of the ASA, and give good agreement with each other. It also appears that better estimates are obtained using higher values for the "precision parameters" ZSTEP (in SURFACE) and PNTDEN (in AREAIMOL), as would be expected. It also seems that different choices of van der Waals' radii have a small effect on the ASA estimates, of a similar order of magnitude to the effect of varying the precision parameters.

The test cases chosen are far from comprehensive. Chothia examined a much larger number of cases (15 proteins) and remarked that "the values of residue accessible surface areas [using the Lee and Richards algorithm] were similar to those found by Shrake and Rupley, though they used slightly different van der Waals' radii and averaged

over a number of different residue confirmations." The results obtained here seem consistent with this conclusion.

Finally, the aim of this article was not to "prove" that one method or program is "better" than the other. Unsurprisingly, both programs seem to perform best when using the highest precision settings on ZSTEP/PNTDEN, and there seems little to choose between the two algorithms. Ultimately the choice of program will be determined by the type of problem being examined, based on the list of program functionalities in section 3.3.

5. The Future

AREAIMOL is still undergoing development, and at present the plan is to include the desirable features from SURFACE (which is no longer under active development) which are currently missing - for example, atom selection options and more powerful Van der Waals assignments.

Beyond that the possibilities are to extend the range of analyses offered by the program. One example is the inclusion (as of CCP4 4.0) of the facility to search for "isolated surfaces", which identifies surfaces which are completely enclosed inside the molecule (i.e. closed cavities at least as big as the probe sphere). Such cavities will contribute to the total accessible area of the molecule using the algorithms described in section 3.1, but in practice should not be counted as part of the external accessible surface. They may also be of interest in their own right, as features of the protein structure. Other possibilities under consideration include more comprehensive analyses of the location of area differences and buried atoms, or trying to identify features such as tunnels.

As always comments and feedback are welcome. If you have any suggestions for improvements or future developments then please address them to me at p.j.briggs@dl.ac.uk.

References

Connolly, M.L. (1996) "Molecular Surfaces: A Review", <http://www.netsci.org/Science/Compchem/feature14.html>

Chothia, C. (1975) *Nature* **254** 304-308

Lee, B. and F.M. Richards (1971) *J. Mol. Biol.* **55** 379-400

Richards, F.M. (1977) *Annu. Rev. Biophys. Bioeng.* **6** 151-176

Sevcik, J., Dauter, Z., Lamzin, V.S. and Wilson, K.S. (1996) *Acta Cryst. D* **52** 327-344

Shrake, A. and Rupley, J.A. (1973) *J. Mol. Biol.* **79** 351-371

Report on the CCP4 Study Weekend on Low Resolution Phasing

Julie Wilson

*CCP4 Study Weekend 2000 on Low Resolution Phasing
University of York (UK), 7-8th January 2000*

Organisers: Julie Wilson (York), Jonathan Grimes (Oxford) and Helen Saibil (Birbeck)

The meeting began with an overview of the standard tools used in small molecule crystallography and their relevance to macromolecular structure solution. Chris Gilmore described the limitations of these methods as the number of atoms increases and the resolution of the data is reduced, but showed how the situation can be simplified, for example by the use of "globs" of density rather than atoms, in order to use the direct methods approach. Chris also talked about the use of electron microscopy images and maximum entropy in providing low resolution phases. This led nicely into Helen Saibil's talk on cryo-electron microscopy, in which she introduced the different methods, i.e. 2D crystals, 1D helical assemblies and single particle EM. Helen described the various stages associated with achieving a 3D image from 2D projections and finished with a video showing how even at 30Å the dramatic changes as GroEL binds either ATP or ADP can be seen.

We then moved on to a series of talks on the ab initio phasing of macromolecular crystal structures. Vladimir Lunin stressed the importance of good selection criteria in choosing the best phase set from a large population of random starting phase sets. He discussed selection procedures and concluded that, although as yet no infallible criteria was known, they can imply that certain "sets" of phase sets are more likely. Thus the alignment and clustering of electron density maps to produce such sets can lead to a determination of the molecular envelope after averaging. Alexandre Urzhumtsev continued with a review of constraints which can be applied to the electron density. He discussed the use of electron density histograms and connectivity theory as well as constraints given by packing considerations and known topological properties. Alberto Podjarny explained how a small number of large Gaussian spheres, generated randomly in the "Few Atoms Method", can be used to calculate structure factors whose magnitudes can be compared with the experimental magnitudes. After clustering, a generalized likelihood criteria is used to determine the best cluster which can then be averaged to provide a low resolution mask. The resolution can be increased gradually by generating smaller spheres within the mask obtained at each stage.

We then returned to electron microscopy with an account of the single particle crystallography of 50S ribosomal subunits given by Elena Orlova. She explained how the multiple orientations in this method are aligned and averaged before the "common lines" in the 2D images can be used to reconstruct the 3D image. The structure has been solved to between 9 and 5Å by various groups, a resolution which Elena showed allows the fitting into the density of those parts for which X-ray crystal structures are available, for example the Principle Interface Protrusion and Protein L9. Vinzenz Unger followed with the use of 2D crystals in the solution of the membrane proteins, connexions, which form channels to allow small proteins through the membrane. The 5-10Å electron microscopy maps clearly

show separated helices at either end from which a first model based on the sequence could be built. Vinzenz explained how information on which residues are likely to have access to the channel, be exposed to the lipid bilayer or adjacent to other helices was used. Mutant studies have been used to test and improve the model. Niko Grigorieff then gave further details of the problems involved, particularly in the alignment of noisy data when no "perfect" model is available for comparison. He described the program FREALIGN, which maximises a likelihood function rather than using a correlation coefficient to align the images. The correlation between the refinement of two half data sets allows an independent assessment of the quality and provides a resolution cut-off.

The second day of the meeting began with a combination of the techniques of X-ray crystallography and electron microscopy as Steve Fuller described the use of X-ray structures in the interpretation of EM maps. As well as fitting X-ray structures into EM maps, Steve showed how, in the case of large viruses, a very low resolution X-ray structure (~30Å) can be used as a model in the early stages of an electron microscopy solution. Bill Shepard followed with the use of anomalous scattering to generate contrast variation. MASC requires a single crystal rather than a change of solvent to determine the molecular envelope.

We then heard how it is possible to collect the very low resolution reflections, so vital in *ab initio* phasing, in-house. Gwyndath Evans described how, if the experiment is set up carefully, these reflections can be measured accurately. Then Jian-wei Miao dispensed with the phase problem completely by over-sampling the diffraction pattern in the case of non-crystalline particles and suggested the same approach could be used for crystals.

Then it was back to the molecular envelope and its use in phase extension and improvement. Pietro Roversi showed that the Fermi-Dirac distribution could be used to provide a continuous envelope rather than binary map for use in solvent flattening or the refinement of incomplete structures. Peter Main followed with a method to extend the phases provided by the envelope. In order to increase the resolution of an electron density map, the right amount of detail must be added to the correct place in the map and Peter showed that the wavelet transform gives some control over this. He showed that 10Å starting phases could be extended as far as 6-7Å.

Dave Stuart described the use of averaging both between different crystal forms and through non-crystallographic symmetry. He also showed how known sub-units can be fitted together into EM maps for use in the molecular replacement solution of large X-ray structures. This began the theme for the afternoon's talks. The problems, particularly of scale, encountered when fitting X-ray structures into EM maps were discussed by Alan Roseman. In the case of receptors with sugars, Michael Rossmann showed how difference maps were used to confirm the orientation of the receptors and finally Elizabeth Hewat warned of the additional difficulties involved when fitting into disordered maps.

The meeting ended with a discussion session led by Michael Rossmann. The recent advances in electron microscopy have been spectacular and it was agreed that the future would lead to further combination the technique with X-ray crystallography.

Experiences from my CCP4 scholarship 2000

S. Ravichandran, January 2000

At the beginning of this millennium, CCP4 announced a new bursary scheme providing scholarships for young crystallographers, mainly from the developing countries, to attend the annual study-weekend program and also to visit several protein crystallography (PX) laboratories across the UK. Being one of the recipients of this year's scholarship award, I am glad here to come forward and share my experiences with others.

On 15th of September 1999, the Chairman of CCP4, Prof. N.W. Isaacs, offered me the first study-weekend scholarship for the year 2000. I owe a lot of thanks to the selectors, who are senior protein crystallographers in India (and also in the UK), for nominating me. It was my first-ever scholarship, or for that matter even for anyone, traveling from India to attend such a (CCP4-sponsored) crystallographic meeting in the UK. Like me, two other young crystallographers were selected from China and were also offered this scholarship.

The scholarship included all financial expenses like traveling, and daily allowances for a period of two weeks. Also, as a part of this scholarship I was asked to select some PX groups (of my own choice) across the UK for visits to discuss scientific matters. I owe a special thanks to Mr. David Brown, the CCP4 executive assistant, who helped me in arranging financial and official matters. He also supplied me with a bunch of e-mail addresses of protein crystallographers in the UK who had shown interest in hosting me. So, all my official correspondence with the CCP4 office (at Daresbury) and my host laboratories across the UK were made through e-mails.

I was much delighted to see the positive and swift response from everyone, i.e., Prof. E.J. Dodson (York University), Prof. N.W. Isaacs (Glasgow University), Dr. W.N. Hunter (Dundee University), Dr. E.F. Garman (Oxford University), Prof. R. Acharya (Bath University) and Prof. R. Read (Cambridge University). I had no problems in communicating with all the above hosts and therefore I was able to fix my schedule well in advance. They had also arranged my accommodation, the day's programme, etc. well in advance and so I could gain one or two extra days in some of these places. I utilized this time by visiting a few other places like Edinburgh, St. Andrews, Daresbury and London. In this regard, I would like to give my special thanks to Dr. J. Naismith and Prof. G. Taylor at the University of St. Andrews, and Dr. P.J. Rizkallah and Dr. S. Bailey at Daresbury Laboratory.

The topic for this year's study-weekend programme was "Low resolution Phasing" and the meeting was held on January 7-8, 2000 at the University of York. Although my flight to London was delayed from the scheduled time by 14 hours due to poor climatic conditions at New Delhi, I managed to reach London on the evening of Wednesday, 5th Jan. at 19:15 pm. I was happy (and also amazed) to see my host, Prof. E.J. Dodson, waiting for me at York railway station at quarter past midnight. In York, I was put up at the Goodricke College.

During the day on 6th of January, I visited the Chemistry department (the York Structural Biology group), where I had many scientific discussions with researchers like Dr. G. Murshudov, Dr. J. Turkenburg, Prof. E.J. Dodson, Dr. S. Lewis and Dr. S. Roberts. I was

lucky to be invited for lunch at my host's sister-in-law's place, near the University. On the next day (7th Jan), the conference started at 11-00 am at the University Central Hall. At first, the Chairman briefed about the scholarship scheme and called the three nominated names for this year. This was followed by the two introductory talks by Prof. C. Gilmore on Direct methods and Prof. H. Saibil on the subject of Cryo-E.M. The rest of the scientific sessions were about the various interesting topics like the *ab-initio* phasing and the Few Atom Method (FAM), Cryo-E.M. and its applications in (virus) crystallography and phase extension techniques. More details about the programme is presented elsewhere. Some of my photo collections on this meeting can be seen here <http://www-bsg.saha.ernet.in/uk-visit.html>

After the CCP4 meeting, I went to Glasgow University and visited the protein crystallography group at the Chemistry department. Here, my host was Dr. Jeremy Beauchamp, in the absence of my original host Prof. N.W. Isaacs. I had scientific discussions with people like Dr. A. Freer, Dr. A. Laphorn, Dr. P. Emsley and some young researchers there. I also visited the Theoretical Crystallography group and had discussions with Dr. K. Anderson. In Glasgow I was given an opportunity to speak (informally) about the kind of research work I am doing in India.

On January 12th, my train to Dundee started at 16-15 pm and I arrived at 18-00 hrs. Despite the weather being cold there (1 ° C), I was given a warm welcome by Dr. W.N. Hunter. I was accommodated in a hotel at Dundee and on the next day I visited the Biochemistry department (Wellcome Trust building) of Dundee University. Here again, I met several researchers, viz. Dr. Van Aalten, Dr. Charlie, Mr. Magnus, Prof. D. Liley etc. and talked about their scientific activities. The next day, Bill (Dr. Hunter) took me to the University of St. Andrews, where I met Dr. J. Naismith and Prof. G. Taylor. The next day again, I had a chance to present my scientific research report in the form of an informal seminar at the department. This was followed by a dinner party in the evening, arranged by the graduate students of Dr. Bill Hunter.

On Saturday (Jan. 15th) morning, I boarded the train for my longest journey across the country (Dundee to Oxford). That evening, I got a warm reception from my host (Dr. E.F. Garman) at Oxford railway station. Then she took me for dinner at her house where I thoroughly enjoyed the company of her family. Later I went to my accommodation, the University Guest house (or the Club house) near the science area. In Oxford, I visited the Laboratory of Molecular Biophysics (LMB) at the Rex Richard building and had talks with the Head of the Department (Prof. L. Johnson) and other faculty members. I had a detailed discussion with my host Dr. E.F. Garman, regarding several aspects about her expertise subject area - ``cryocrystallography". On the same day, one of our senior colleagues (Dr. M. Ghosh) who is doing her research in the same building, took me for dinner at her residence.

The next day I received a surprise invitation from Dr. P.J. Rizkallah, the seminar secretary of the protein crystallography group at Daresbury laboratory. My rail tickets to Daresbury (and also the return-ticket to Bath) were delivered to my address (in my host's name) in Oxford. Though my visit to Daresbury was short, I was able to spend my time usefully, visiting the SRS Daresbury synchrotron stations for PX. Also, I had the chance to meet the CCP4 staff again, viz. Dr. P. Briggs, Dr. M.D. Winn, Dr. A.W. Ashton, Dr. S. Bailey and Mr. D. Brown. After my talk in Daresbury on Wednesday (Jan. 19th), I proceeded to Bath University.

In Bath I received a very kind welcome from my host, Prof. R. Acharya. I was put up at the Polden Court University guest house. I visited the Biology and Biochemistry department at the South Building there. I am thankful to my host for spending an hour of his valuable time in driving me through the city to show me the historic spots. I had my laboratory visit during the noon and in the evening I (once again) gave an informal presentation of my research work. Later that evening, I was invited to my host's place for dinner. Early on Friday morning (Jan. 21st), I started towards Cambridge from Bath. My host at Cambridge (Prof. R. Read) was expecting my visit at the Hematology department of the Wellcome Trust building, which is located near Addenbrooke's hospital. I had a short discussion with Dr. B. Hazes and others there. Later in the evening, I visited the Biochemistry department of the University, located in Tennis Court Road. I am thankful to Dr. V. Dhanaraj (who is at Prof. T.L. Blundell's laboratory) for spending his time with me there. I also had a brief discussion with Prof. B. Johnson.

For the last three days of my stay I was held up in London, as my return flight was scheduled for Monday night (Jan. 24th). Because it was the weekend, I wasn't able to visit any scientific institutions. Instead, I spent those days sight-seeing.

Though my journey was bit tiring and hectic, I was able to thoroughly utilize my time and enjoy my visit during that short fortnight in the UK. I would like to convey my thanks to all those whom I met during this period, especially the scholarship organizers (CCP4) and all my hosts, who gave their kind cooperation and spent their valuable time on my behalf.

S. Ravichandran

Crystallography & Molecular Biology Division,
Saha Institute of Nuclear Physics,
Calcutta - 700 064,
INDIA.

e-mail:- ravi@cmb2.saha.ernet.in

Recent CCP4BB Discussions

Maria Turkenburg (mgwt@york.ac.uk)
March 2000

Back by popular demand

In the October 1999 Newsletter, Martyn Winn started what will, hopefully, become a trend in keeping track of interesting discussions on the CCP4BB. To make things much easier for both the users of the bulletin board and the track-keeper, *members who ask questions or instigate discussions on the board are now asked (urged!) to post a summary of all the reactions received*, whether on or off the board.

The introduction to the October 1999 version also goes for this article:

For each subject below, the original question is given in italics, followed by a summary of the responses sent to CCP4BB (together with some additional material). For the sake of clarity and brevity, I have paraphrased the responses, and all inaccuracies are therefore mine. To avoid misrepresenting people's opinions or causing embarrassment, I have not identified anyone involved: those that are interested in the full discussion can view the original messages (see the CCP4BB web pages on how to do this).

These summaries are not complete, since many responses go directly to the person asking the question. While we understand the reasons for this, we would encourage people to share their knowledge on CCP4BB, and also would be happy to see summaries produced by the original questioner. While CCP4BB is obviously alive and well, we think there is still some way to go before the level of traffic becomes inconvenient.

We, or at least I, would like to thank Sasha Urzhumtsev and Xavier Gomis Rueth for their various attempts to get the summarising off the ground. It seems to have paid off. And thanks to all the users who are now dutifully posting summaries. Finally I would like to thank Eleanor Dodson for her corrections and additions.

Subjects covered in this newsletter's offering, are:

['dm'](#)

- Averaging
- Masking

- Twining

- Playing with Symmetry

 - Local Scaling in P1
 - CAD in R32

- Unexpected and artificial values

 - Temperature Factors and the Wilson Plot
 - Anisotropic Scaling to get at 'true' Bfactors
 - Artificial Bfactors for NMR Structures
 - What to Expect of Correlation Coefficients in AMoRe
 - Predicting the Number of Reflections
 - FOMs in MLPHARE for MAD
 - REFMAC for Partial Poly-Ala

Various

AMoRe and CCP4 Asymmetric Unit
Distinguish Calcium from Magnesium
Predicted Rfree-R Difference
Packing Density
f' and f''

'dm'

Averaging

For a case of complicated NCS-symmetry, namely six copies of a protomer grouped as three dimers A (consisting of A1 and A2), B and C, the following symmetry operators are known:

inter-dimer: A → B
 B → C

intra-dimer: A1 → A2 (from subunit 1 of dimer A to subunit 2 of dimer 1)
 B1 → B2
 C1 → C2

where the operators consist of a 3x3 matrix plus a vector. The question is: is this a representative set of operators for successful averaging?

The following fundamental principle should be applied:

YOU ONLY NEED ONE MASK FOR 'dm'.
IF YOU HAVE MORE THAN ONE MASK YOU ARE DOING SOMETHING WRONG

(unless you are an expert and are solving a very specific problem, in which case you know all about it). What you need to do is supply one mask, which covers A1, and the operators which map A1-A1 (i.e. the identity), A1-A2, A1-B1, A1-B2, A1-C1, A1-C2.

If your dimer rotations are self-inverse, then your mask could equally well cover A instead of A1. If all six NCS operators form a closed group, then you could use a hexamer mask.

How to generate the symmetry operator A1-B2 if you just know the operators A-B and B1-B2?

$(A1-B2) = (B1-B2) \times (A-B)$, i.e. multiply the matrices, take the right order! As all symmetry operators X consist of a 3x3 matrix M AND a vector t, for the multiplication you have to multiply the augmented 4x4 matrices:

$$X = M + t \Rightarrow \begin{pmatrix} m_{11} & m_{12} & m_{13} & t_1 \\ m_{21} & m_{22} & m_{23} & t_2 \\ m_{31} & m_{32} & m_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$X(b1) = \{M(a1-b1) + t(a1-b1)\} X(a1) \Rightarrow \begin{pmatrix} m(a1-b1)_{11} & m(a1-b1)_{12} & m(a1-b1)_{13} & t_1 \\ m(a1-b1)_{21} & m(a1-b1)_{22} & m(a1-b1)_{23} & t_2 \\ m(a1-b1)_{31} & m(a1-b1)_{32} & m(a1-b1)_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$X(b2) = \{M(b1-b2) + t(b1-b2)\} X(b1) \Rightarrow \begin{pmatrix} m(b1-b2)_{11} & m(b1-b2)_{12} & m(b1-b2)_{13} & t_1 \\ m(b1-b2)_{21} & m(b1-b2)_{22} & m(b1-b2)_{23} & t_2 \\ m(b1-b2)_{31} & m(b1-b2)_{32} & m(b1-b2)_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

so

$$X(b2) = \{M(b1-b2) + t(b1-b2)\} \{M(a1-b1) + t(a1-b1)\} X(a1)$$

But it is much easier to use LSQKAB (*please use your local version of the CCP4 Program Documentation to view this*) to fit the coordinates of A1 onto those of B2, etc.

Masking

Is there a way to harness the averaging to generate a better map from which to derive the solvent mask; effectively a single run of maskless averaging?

A dimer mask (proper NCS) made using MAPROT and MAPMASK turned out dismal. On the other hand, the one from 'dm', assumed to be calculated on the unaveraged map, is much better. An NCS mask can be simply made by putting an atom of 40Å radius at the centroid of the heavy atom position. In tests this is about as good as a monomer mask.

This stimulated a lively discussion on the use of the phrase "maskless averaging", after which the question was rephrased as *How to improve the original 'dm' solvent mask?*

There are several ways of using averaging and the answers need to be summarised differently for each application. It is assumed you know the NCS operators which map the "master" molecule onto all other copies in the asymmetric unit.

1. To average an existing map using known NCS operators - this is Eleanor's definition of maskless averaging.

If you are only interested in averaging a MAP there is no reason, except for speed, to use a mask. Those parts of the map which obey the NCS symmetry will be improved by averaging, whilst those parts which do not, will deteriorate.

If the NCS operators are "proper" or "form a closed group" such that there is a NCS-defined pure rotation, the whole complex will be improved. The density for molecule 1 will be overlapped and summed with that of molecule 2, that of molecule 2 overlapped with that of molecule 3, until the complete rotation is done.

If the NCS operators are "improper", only the master molecule will be improved.

Hopefully this will also generate a better molecular boundary and thus a better initial solvent flattening mask.

2. The NCS operators can be refined to improve the correlation between different copies of the electron density. This refinement will be done using density within some masked region but this need not cover the whole molecule. This restricted mask can be obtained in various ways. The simplest is to put a single "atom" at the centre of the molecule, then generate a mask around that "atom" assigning it a large VDW radius.

The centre of mass can be decided either by inspecting the map, or possibly, if the NCS was determined by fitting heavy atoms within each molecule, it could be the centre of the heavy atom cluster. If the map is very noisy, it may be sensible to only use the strong density to refine the NCS operators.

3. However once the NCS operators have been refined, it is sensible to use density modification to improve the phases. This is done in the following way
 - a. The maps are averaged over the volume of a single molecule; *i.e.* a mask is needed. If no mask is supplied the program 'dm' endeavours to determine one, but this procedure may not be as effective as can be determined from visual inspection of an averaged map.
 - b. The asymmetric unit of the crystal is reconstructed with the improved density in place (and flattened density everywhere outside the mask).
 - c. Structure factors are calculated by transforming the density to give modified phases.
 - d. The modified phases are combined in some way with the starting set (this step will be seriously hampered if the low resolution data are missing).

And the whole process cycles round, with the option of phase extension as well as phase refinement.

N.B.

New versions are available of the map manipulation utilities MAPMASK and MAPROT. The changes dramatically simplify the process of cutting out a region of density for use as a molecular replacement search model, *e.g.* for locating NCS or multi-crystal averaging operators.

As a result, map cutting may be performed by supplying MAPROT with a work map and mask (WRKIN+MSKIN), leaving the unit-cell map (MAPIN) blank, and giving a single operator (which may be unitary) along with GRID and CELL for the cutout map. The cut density is written to the CUTOUT file.

The new maputils are available by ftp as follows:

```
> ftp ftp.yorvic.york.ac.uk
login: anonymous
password: type_your_full_email_address_here
ftp > cd pub/ccp4
ftp > get maputils.tar.gz
ftp > quit
> gunzip maputils.tar.gz
> tar xvf maputils.tar
> cd maputils/maprot_
> makemaprot
> cd ../mapmask__
> makemapmask
```

Twining

(January 2000)

- a. *What conditions have people successfully used to overcome twinning in crystallization?*
- b. *What MIR structures have been solved using (de)twinned data (besides cephalosporin synthase)?*
- c. *Have any structures been solved by MAD using detwinned data?*
- d. *Has anybody experience in detwinning pseudomorphedral twinned data?*

Thanks to the original enquirer for the summary:

- DETWIN (*please use your local version of the CCP4 Program Documentation to view this*).
New (CCP4 Suite version 4.0) features: DETWIN now prints twinning tests for a range of twinning fractions.
 - The Yeates plot of $\langle H \rangle = \langle (I_{tw1} - I_{tw2}) / (I_{tw1} + I_{tw2}) \rangle$.
The estimate of the twinning fraction is given by $1/2 - \langle H \rangle$.
 - The Britten plot: essentially the number of negative intensities generated by twinning fractions ranging from 0 to 0.5.
 - Moments of $\langle E \rangle^2$; these have characteristically different values for twinned and proper data (the same plot is given in TRUNCATE; *please use your local version of the CCP4 Program Documentation to view this*).
 - Correlation coefficients between I_1 and I_2 after detwinning. Ideally these should be zero, but the correlation coefficient can be distorted when there is NCS aligned with a possible twinning axis.

A specific twinning fraction only needs to be given if you want to write detwinned data to HKLOUT.

- Detwinning program of Rams (S. Ramaswamy) in Uppsala, referred to in cephalosporin synthase paper, Nature 394 (1998), 805-809

Playing with symmetry

Local Scaling in P1

(January 2000)

In a case where 4 complete 360 degree rotations of data were collected for a high symmetry (spacegroup 96) crystal, scaling all the data together resulted in high R values, presumably due to absorption effects. Since the object was to measure a weak anomalous signal at a relatively long wavelength, an idea would be to take the scaled, unmerged data in P1 and apply a sliding box type of local scaling, averaging all the Bijvoet pairs down to the asymmetric unit in the process. The Matthews & Czerwinski protocol (Local Scaling: A Method to Reduce Systematic Errors in Isomorphous Replacement and Anomalous Scattering Measurements (1975), Acta Cryst A31, 480-487) can not be used, since their derivation assumes a comparison of only two quantities (e.g. F_+ and F_- or F_{nati} and F_{deri}). Here, it would be desirable to average over all 16 of the symmetry-related reflections. Questions:

1. *Does this sound nuts? It seems to me that if one chooses a sufficiently large box, the simultaneous estimation of even a large number of local scale factors might remain robust.*
2. *Can anyone recommend a program to do this? Or toss some snippets of code my way?*

Some thoughts on the first, and a firm answer to the second question:

1. I don't think going to P1 will help you - the improvement in R factors is purely cosmetic. Did you look at Rmerge statistics [Nat Struct Biol 4, 269 (1997)]? These should reveal that the high redundancy actually helps the quality. I don't quite understand why absorption effects should be so detrimental - I twice

collected data at the Fe edge and they scaled beautifully. Could there be a problem with mis-indexing due to origin offset? If you use XDS/XSCALE you have the option to only use Friedels that are a given max number of frames apart for calculating the anomalous signal.

2. SCALA (*please use your local version of the CCP4 Program Documentation to view this*) has a number of useful options to do this. The latest version (ftp://ftp.mrc-lmb.cam.ac.uk/pub/pre/scala_2.7.1.tar.gz or CCP4 release 4.0) has a spherical harmonic scale parameterisation which should work well in this case. It will probably only work with data integrated with MOSFLM, since e.g. SCALEPACK declines to reveal the essential geometrical information in its output.

Then two references to sliding box scaling, the 'original' way:

J Appl Cryst 30: 176 (1997)
Met Enzymol Vol 276, pp. 461-472.

CAD in R32

(September 1999)

I am running CAD on a data set in the spacegroup R32 and am wondering why cad does what it does. Input reflections (4 -2 6) and (4 -1 -7) are transformed to (2 2 6) and (3 1 -7), respectively. How are these reflections equivalent? The matrix that seems to perform this transformation, is not a symmetry operator for R32. Nor is the transpose, which would presumably cater for real space. Can anyone tell me what asymmetric unit CCP4 uses for R32? And why?

CAD is correct, and those pairs of reflections are indeed equivalent.

Remember that if you generate a symmetry equivalent position in real space using this expression:

$$\begin{aligned} [X_j] &= (S_{j11} \ S_{j12} \ S_{j13} \ St_1)[X_1] \\ [X_j] &= (S_{j21} \ S_{j22} \ S_{j23} \ St_2)[X_1] \\ [X_j] &= (S_{j31} \ S_{j32} \ S_{j33} \ St_3)[Z_1] \\ &= (0 \ 0 \ 0 \ 1)[1] \end{aligned}$$

an equivalent reflection in reciprocal space is generated by:

$$[H_k \ K_k \ L_k] = [H_1 \ K_1 \ L_1] \begin{pmatrix} S_{k11} & S_{k12} & S_{k13} \\ S_{k21} & S_{k22} & S_{k23} \\ S_{k31} & S_{k32} & S_{k33} \end{pmatrix}$$

For R32 a symmetry matrix: -Y, X-Y, Z can be represented as:

$$\begin{aligned} [X_2] &= (0 \ -1 \ 0 \ 0)[X_1] \\ [Y_2] &= (1 \ -1 \ 0 \ 0)[X_1] \\ [Z_2] &= (0 \ 0 \ 1 \ 0)[Z_1] \\ &= (0 \ 0 \ 0 \ 1)[1] \end{aligned}$$

The equivalent reflection is

$$[H_k \ K_k \ L_k] = [H_1 \ K_1 \ L_1] \begin{pmatrix} 0 & -1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = [K_1, -H_1 - K_1, L_1]$$

For R32 the complete set of reciprocal space equivalents are:

$$[h, k, l] \ [k, -h-k, l] \ [-h-k, h, l] \ [k, h, -l] \ [-h-k, k, -l] \ [h, -h-k, -l]$$

and the equivalent $[-h, -k, -l]$ sets

Symmetry operator 5 gives (4,-2,6) equivalent to (-2,-2,-6) and negating all signs gives (2,2,6).

Symmetry operator 5 gives (4,-1,-7) equivalent to (-3,-1, 7) and negating all signs gives (3,1,-7).
(-h-k) is sometimes written as i, and people refer to the four-index Miller-Bravais symbols (h k (-h-k) l):
(4 -2 6) becomes (4 -2 -2 6); (4 -1 -7) becomes (4 -1 -3 -7).

Unexpected and artificial values

Temperature Factors and the Wilson Plot

(January 2000)

In a straightforward REFMAC refinement of a structure with MAD phasing, with data good to 2.4Å, TRUNCATED and UNIQUEified as should be, unexpected Bfactors ranging from 40 to 130 (average 65) appeared. The Wilson B value was 56. I guess I have no real specific questions except to wonder if this is telling me something about the quality of either my data, my model, or my crystal. I have been unable to find a reference for the acceptable values for a Wilson B factor. Would anyone suspect that I ran TRUNCATE improperly? Any thoughts would be appreciated.

Thanks to the original enquirer for the summary:

The general consensus was that I have nothing to worry about. Several people pointed out that most temperature factors for protein structures are underestimated leading to the bias that "good" structures should have average temperature factors in the 20-30 range. A couple of people asked about the solvent content of the crystal. As some of you guessed, this is high. I have a V_m of 4.0 yielding an approximate solvent content of 69 %. This explains some of the thermal motion. It was also suggested to check the native as well as the SeMet data. The Wilson B values are high for all data sets that I have, however the native is the lowest. I was also advised to examine the Wilson plot to look for typical features of the dip around 5.5Å and the peak around 4Å. In this respect, my Wilson plot looks fine. One last suggestion was to compare the results to refinement with another program. I will give this a try.
A selection of a few comments I received:

As far as I can tell from your post your data seem to be alright. A B of 56 is quite high but your data extend to 2.4Å resolution only, so that is consistent. A model B being slightly higher than that is also ok, since your model is still incomplete. The only thing you need to watch is whether your model B keeps increasing from one round of REFMAC refinement to the next. If it does you should set the overall value back to the 56 before a new round.

I believe that average B factors given for structures are more often too low than too high. If the data peters out at 3Å it is very difficult to get any sensible estimate of a Wilson plot gradient, but it is most likely that a true Av_B should be in the range 80-100, rather than 10 as is sometimes given! REFMAC values should reflect the gradient of the Wilson plot, and they do seem to do so quite well in your case.

Seems fine to me, the low B factors (~20) were normal when people only worked on crystals that diffracted strongly on weak x-ray sources. Synchrotrons and mirror optics allow the measurement of weaker diffracting crystals (with higher B-factors).

I found that different crystals within one batch might yield quite different B-factors (45-75 avg). But they didn't really make any difference to the biological interpretation.

Anisotropic Scaling to get at 'true' Bfactors

(September 1999)

For a 2.3Å structure, using strict NCS and group Bfactors (2 per residue) in XPLOR, with anisotropic scaling, improves the Rfactors but the Bfactors start to behave strangely. They increase, and some main chain atoms have higher Bfactors than the side chain atoms in the same residue. This is found in the Protein Data Bank, too. Am I the only one worried about it?

The general consensus is that striving for low Bfactors is a thing of the past. High Bfactors can be very 'true' Bfactors. Relatively low side chain Bfactors would make sense if they are buried, but connected to an exposed piece of main chain which would then have higher Bfactors. The PDB is not a reliable source of information about Bfactors, since the scaling algorithms in XPLOR inevitably mean that most low resolution structures have totally unrealistic Bfactors. XPLOR scales Fobs rather than Fcalc, which, especially at lower resolution and at the beginning of a refinement procedure, could cause problems. It would be sensible to start refinement not with an average B of 20 (as output by many model building programs), but with an average B that represents the Wilson plot as closely as possible. Also, to apply Bfactor corrections sensibly before each round of refinement. There may be valid reasons for adjusting the Bfactor of the Fobs, for making maps or Patterson or rotation function searches, sharpening the data to accentuate the high resolution details, but the final atomic Bfactors should be refined against an uncorrected dataset.

With regards to the use of NCS, most people now accept that relaxing it, with care, is good crystallographic practice, and could have a beneficial effect on Bfactors in refinement.

Artificial Bfactors for NMR Structures

(October 1999)

Can anyone tell me what program I can use to compute artificial temperature factors for a NMR structure based on rmsd of the average structure?

Have a look at:

Wilmanns & Nilges, Acta Cryst (1996) D52:973-982. Molecular Replacement with NMR Models Using Distance-Derived Pseudo B Factors. Script available from Jovine Luca.

and also at:

Molecular Replacement and NMR models - what gives? and the corresponding script through FTP.

What to Expect of Correlation Coefficients in AMoRe

(September 1999)

The correlation coefficient in AMoRe's FITFUN behaves unexpectedly in an otherwise fairly straightforward molecular replacement case. The starting cc in FITFUN is much lower than the one that came out of the rotation and translation

steps, while there is also an unexpected Bfactor correction. Interestingly, this behaviour is not seen when using E instead of F. Following a request to do so, more specific information was supplied: the data extends to 4Å only, are very strong to 7Å and trail off quite steeply. Estimating the Bfactor from the Wilson plot is therefore not straightforward, but put at approximately 50Å². The average Bfactor of the search model is 44Å².

If you calculate the correlation between Fobs and Fcalc you will always get a positive outcome because they are correlated in the sense that both decrease with resolution. If you use E values you do not see this since E values do not decrease with resolution, as you noted. In your case the effects may have been more severe, perhaps due to too large B-values for the search model?

Predicting the Number of Reflections

(January 2000)

What is the easiest way to calculate the theoretical number of reflections in a particular resolution shell given that all cell dimensions are known?

Quick and dirty

Get volume of reciprocal lattice unit, and divide volume of spherical shell by this.

More precise

Run UNIQUE to generate a reflection list within the resolution range required and see how many sets are generated.

FOMs in MLPHARE for MAD

(September 1999)

Does anyone have a feel for whether MLPHARE will over estimate the FOM from its MAD analysis? Secondly, if you include the native and three wavelengths, given the sites are all the same XYZ and differ only slightly in AOCC and OCC, how do we judge the FOM?

It most definitely will if you refine the xyz and occupancy for all wavelengths at once. There is an implicit assumption in the algorithm that each "derivative" is independent, which is obviously not true when a "derivative" is taken as another wavelength, with all differences arising from the same atom sites. The more data sets at different wavelengths with the same sites, the fatter the FOMs get.

The best way to minimise this problem is to refine the coordinates and temperature factors against the largest set of dispersive differences for the centric reflections only (assuming you have such reflections). After that refinement fix the xyz and B for all wavelengths, and the ISOE and occupancies for that pair of dispersive differences. Then estimate the ISOE and occupancies for all other pairs of dispersive differences.

Only then with all XYZ, B, OCC and ISOEs fixed, start refining the anomalous occupancies for all wavelengths. There is a good deal of anecdotal evidence that many wavelengths actually gives worse maps than 2 or 3. This is probably not because the phases are worse, but because the FOMS are overestimated.

Be careful about including the native in the MAD phasing. I wouldn't put it in as the native, but as a derivative (with negative OCC and 0 AOCC). If you put it in as the native, the error estimates in the differences with all the other data sets may be bigger than your actual signal, unless the latter is truly gigantic. In that case the FOMS do reflect the quality of the phases (both would be kind of bad).

REFMAC for Partial Poly-Ala

(January 2000)

The problem we encountered is with a partial poly-Ala trace and poor MIR phases to 3Å. We employed example 9 in the REFMAC documentation, described as "very bad model, rms error 2Å". The resultant model is drastically distorted with some Ca-Ca distances over 6Å long. The question is which protocol should be followed with an initial Ca trace at 3Å?

Thanks to the original enquirer for the summary:

A straight answer to the question about the distortion, was found in the input to PROTIN: the start of the polypeptide chain needed to be dealt with slightly differently (NTERM 1 rather than NTERM 195).

The question also sparked a discussion about the use of all data in refinement, and the use of torsion angle refinement to decrease the number of parameters to refine. In this case, the data really only extended to 3Å, and careful testing of REFMAC and CNS produced very similar and equally interpretable maps.

Maximum-Likelihood refinement using experimental phases (as in REFMAC) can deal with poor/partial models correctly, *if* the phases have reliable FOMs. Including all data works best providing you are estimating errors from the Rfree set. Initially the weights assigned to the outer resolution bins are very small, but they still contribute significantly to the important overall scaling parameters. The amount of bias is often a function of the amount of missing data. By default REFMAC substitutes these reflections as Dfc instead of the 2mFo-dFc used for observed data, which usually gives better maps than omitting them altogether (that is tantamount to setting them to zero), but any such term will introduce bias. If you have many missing reflections you may have a problem.

Do not expect miracles, and keep an eye on the theory: ML is a minimisation CRITERION, and NOT a minimisation method. One can compare it with the least-square criterion. It works in the same way except that instead of fitting calculated magnitudes to the observed ones, it suggests to fit them to some MODIFIED VALUES, and these modifications are estimated through the model quality and completeness (see, for example, Maximal Likelihood refinement. It works, but why? in the October 1999 CCP4 Newsletter). If the model is incomplete, it is wrong to fit the FC only to experimental amplitudes (as is the case in the LS refinement) and ML attempts to introduce corrections. These corrections are estimated in resolution shells. If your model does not fit at all your data at a given resolution shell, the ML puts the "corrected experimental magnitudes" in this shell near to zero. As the model improves, the weighting will increase. This is similar to the "good old scheme" of slowly increasing the resolution to include in the refinement, but less subjective.

Various

AMoRe and CCP4 Asymmetric Unit

(January 2000)

Is there a program to transfer AMoRe solutions into the CCP4 asymmetric unit (i.e. just redefining rotation and translation from the AMoRe output)?

Any MR program is simply going to find a CRYSTALLOGRAPHIC solution, and not care about building a sensible model. The following procedure would take care of this:

- a. Always shift the first molecule to have translations between $-1/2$ and $1/2$. Assuming you apply the solution to the coordinate file OUTPUT from the TABFUN run, which has its centre of mass at (0,0,0), proceed with:

```
b.  pdbset xyzin tabfun_out.pdb xyzout soln1.pdb
c.  cell A B C alpha beta gamma (new cell)
d.  rota euler ALPHA BETA GAMMA
e.  shift frac Tx Ty Tz
f.  CHAIN A
    end
```

- g. Then for further molecules you may well want a symmetry equivalent of the AMoRe solution to build up a tetramer or something. Easiest way:

- o Generate all molecules from the AMoRe solutions with different CHAIN IDs and make one coordinate file (all_solns.pdb).
- o Run:

```
o  distang xyzin all_solns.pdb
o  SYMM whatever
o  RADI CA 4
o  dist VDW
o  END
```
- o That will give you contacts between molecules, complete with symmetry codes: e.g. 2 -1 0 1. There are various choices, but once you see that maybe there are better contacts between molecule 1 and molecule 2 using symmetry 2 -1 0 1, you would generate a new version of molecule 2:

```
o  pdbset xyzin soln2.pdb xyzout soln2a.pdb
    SYMGEN -X-1, Y+1/2, -z+1 (symmetry 2 -1 0 1 in spacegroup P21)
```
- o Then rebuild the all_solns.pdb with the new soln2a.pdb and run distang again.

You are aiming to get a nice compact molecule with good contacts produced by symmetry X,Y,Z generated by particular symmetry operators. Can be messy but it saves a lot of time on the graphics later!

Distinguish Ca^{2+} from Mg^{2+}

(December 1999)

I am having troubles in identifying two metal ions in a structure. I used Mg^{2+} in sample preparation and Ca^{2+} in crystallization. The coordination geometry, refinement statistics (both Rfactors and Bfactors) and maps do not resolve the ambiguity.

I would like to know about the experts' opinions. The question in my mind is whether the bonding distance with the coordinating oxygens is a discriminator ($\sim 2.3\text{\AA}$ in my case)? Also how much the B-factor can tell? (in my case, $\text{Mg}^{2+} \sim 12$, $\text{Ca}^{2+} \sim 29$, average of the molecule ~ 37).

There are a few things you can do to distinguish between Mg^{2+} and Ca^{2+} :

Crystallographically

- h. Do you still have the raw data? If yes, even a little anomalous data will decide this question. Don't merge the Friedel pairs and calculate an anomalous difference Fourier (CCP4 FFT with $\text{DANO}=\text{D_nat}$ $\text{PHI}=\text{PHIC}$, i.e. coefficients $(F^+ - F^-) \exp[i(\text{phimodel} - \pi/2)]$). You should see very clear peaks for Ca and none for Mg. This is because Ca has an f'' of 1.286 electrons at CuK α radiation, whereas magnesium has only an $f''=0.177$. Everyone collects some anomalous data - but if you run SCALEPACK with ANOM NO, you can lose it. If you always set ANOM YES, you still get a merged $\langle l \rangle$ for all hkl and -h-k-l pairs but the output also preserves the anomalous differences where observed. This is the default for SCALA.
- i. You should be including the low resolution data if you have it (i.e., 20 - 30 \AA) and this will allow an accurate bulk solvent correction (this is really just general advice).
- j. Ca^{2+} has 8 more electrons than Mg^{2+} which should give you a much higher electron density (contour at, say, ~ 4 rmsd(ρ)): do you see the well ordered sulfurs and your metal, only?). But with this you can only identify Ca^{2+} if its occupancy is close to unity. If the occupancies are close to unity, a falsely placed Mg^{2+} instead of a real Ca^{2+} would have a very low B-factor and vice versa. Another way of looking at this, is: The change in B-factor is mopping up the difference. You should check the B-factor of the protein atoms that are the ligands. If the atoms have B-factors around 12, then the Mg^{2+} appears to be appropriate. If they are around 29, then the Ca^{2+} is likely the answer. Yet another contributor was surprised that difference maps weren't good indicators at 2.2 \AA : "At 1.9 \AA we observed respective peaks or troughs in the F_o-F_c maps when too light or too heavy a cation was used in the model, even though the B-factors were soaking up a lot of the error."
- k. If you still have crystals, soak one with Mn^{2+} instead of Mg^{2+} : Mn^{2+} is a very good substitute for Mg^{2+} but has 13 more electrons. If you calculate an $(F_o(\text{Mn}^{2+}) - F_o(\text{unknown}))$ electron density map, you should see a clear signal if $\text{unknown}=\text{Mg}^{2+}$ and a weak signal (if any) if $\text{unknown}=\text{Ca}^{2+}$

Metal-Ligand Geometry

- l. Looking at bond distances only, is rather risky; the refinement programs often have "hidden" restraints which can distort your geometry. For instance most programs

- apply a VDW repulsion unless you specifically request that it be turned off. However, if you have been careful, some or all of the following should help:
- m. Mg^{2+} should be always octahedrally coordinated and an average Mg^{2+} to O distance of 2.1Å, and Ca^{2+} has preferably seven or eight ligand atoms with an average Ca^{2+} to O distance of 2.4Å. These statements were challenged by another contributor: It is dangerous to say "always" in any scientific discussion. We have refined a 1.25Å structure containing a Ca^{2+} coordinated by six ligands and a Mg^{2+} coordinated by five ligands, consistent with our ICP Atomic Emission Spectroscopy measurements. A search of the database at <http://metallo.scripps.edu/current/raw.html> turned up 43, 54 and 269 respective matches for Mg^{2+} coordinated respectively by 4, 5 or 6 ligands, while 102, 160, 369, 445 and 114 respective matches were made for Ca^{2+} coordinated by 4, 5, 6, 7 or 8 ligands. Clearly, there is considerable variability in metal ion coordination geometries.
 - n. Both metals are "hard" metals which like "hard" ligands, that is oxygen, only (nitrogen is a very rare exception, sulfur shouldn't appear). There are two excellent reviews about metals in proteins: Glusker, J. P. (1991) Advances in Protein Chemistry, Vol. 42, 1-75, and Harding, M. M. (1999) Acta Crystallographica, Vol. D55, 1432-1443.
 - o. You have observed an octahedral ligand sphere (indicative for Mg^{2+}) with an average metal-to-ligand distance of 2.3Å (indicative for Ca^{2+}). However, be cautious! Your metal-ligand distances are the result after refinement usually with geometrical restraints, in this case van der Waals radii. In X-PLOR and CNS, the van der Waals radii of Mg^{2+} and Ca^{2+} are by far too large resulting in too large metal-to-ligand distances (which could be an explanation for 2.3Å for Mg^{2+} -to-oxygen distance)! I use sigma of 0.8552 for Mg^{2+} which gives an energy minimum for Mg^{2+} -to-O at 2.08Å, and 1.4254 for Ca^{2+} which gives an energy minimum for Ca^{2+} -to-O at 2.4Å (look at the formula for the energy minimum, insert the value for oxygen and solve for the "best" value for the metal). So, please, before you judge refined metal-to-ligand distances, check the "ideal" geometry parameter of the refinement program of your choice!

Predicted Rfree-R Difference

Can someone post the reference to the paper dealing with the predicted Rfree-R difference as a function of data resolution again?

Tickle, I.J., Laskowski, R.A. & Moss, D.S. (1998) Acta Cryst D54, 547-557.

Packing Density

Can anybody tell me how to calculate the packing density of residues in protein? Is there any standard programme for that?

See Applications for Volume and Packing Calculations (Yale). This will be very helpful, because so many related programmes are available in this site. Furthermore, try CCP4's AREAIMOL (*please use your local version of the CCP4 Program Documentation to view this*), and Columbia's GRASP.

f' and f''

(October 1999)

I am looking for a tool for calculating f-prime and f-doubleprime from X-ray fluorescence spectra for MAD-datasets. Does anyone know a good one?

- CCP4's CROSSEC (*please use your local version of the CCP4 Program Documentation to view this*)
- CHOOCH
- Crystallographic Computing Services - Scattering factors
- Anomalous Scattering Coefficients

Gaussian Likelihoods in real and reciprocal space

Dr. Kevin Cowtan
Structural Biology Laboratory, University of York, UK

1. Introduction

Terwilliger (1999) describes a new likelihood-based framework for employing structural information which was previously exploited by means of conventional density modification calculations. The statistical basis of Conventional density modification is dubious at best (Cowtan & Main, 1996), despite the introduction of 'fudge-factors' which attempt to correct for the worst sources of bias (Abrahams 1997, Cowtan & Main 1999). Terwilliger's treatment provides a better foundation for the integration of density-modification-like information with those elements of structure solution which have already received a more thorough statistical treatment.

In this paper a similar treatment is considered as part of a general consideration of the relationships between likelihood functions calculated for structure factors and electron density maps. This provides some new insights into likelihood-based phase improvement.

Suppose the likelihood of a set of structure factors can be represented by the product of likelihoods for the individual structure factors, and that these likelihoods in turn may be represented by Gaussians centred on some expected value for that structure factor. The log-likelihood is then the log of a general n-dimensional Gaussian, which is a sum of quadratic terms in the complex structure factors. The quadratic distribution may be represented in turn by a general quadratic distribution for the log-likelihood based on the electron density across the whole map.

Furthermore, the whole procedure may be reversed. Thus, a likelihood distribution in real or reciprocal space may be represented in the other space by the following series of steps:

Real space	$\xrightarrow{\log}$	Real space	\xrightarrow{FT}	Reciprocal space	$\xrightarrow{\exp}$	Reciprocal space
Gaussian	$\xleftarrow{\exp}$	Quadratic	\xleftarrow{FT}	Quadratic	$\xleftarrow{\log}$	Gaussian

2. Electron density log-likelihood expression

Consider an ensemble of sets structure factors, which may be represented by a mean structure factor $F_0(h)$. Let the distribution of values of $F(h)$ about $F_0(h)$ be represented by a general two-dimensional Gaussian

$$L_{sf}(F(h)) = \exp \left(- \frac{\text{Re}[F(h) - F_0(h)]^2}{2v_r(h)} - \frac{\text{Im}[F(h) - F_0(h)]^2}{2v_i(h)} - \frac{2\text{Re}[F(h) - F_0(h)]\text{Im}[F(h) - F_0(h)]}{2v_{ri}(h)} \right)$$

where $v_r(h)$ and $v_i(h)$ and $v_{ri}(h)$, and are the Gaussian coefficients obtained by inverting the variance-covariance coefficients for the real and imaginary parts of the reflection..

Let $D(h) = F(h) - F_0(h)$, $D_A(h) = \text{Re}[D(h)]$, $D_B(h) = \text{Im}[D(h)]$, then:

$$P_{sf}(D(h)) = \exp \left(-\frac{D_A(h)^2}{2v_r(h)} - \frac{D_B(h)^2}{2v_i(h)} - \frac{2D_A(h)D_B(h)}{2v_{ri}(h)} \right) \quad (1)$$

Assuming the reflections are independent, the conditional probability of a unique set of structure factors given a particular model for the magnitude and phase errors may be calculated from the product of the individual likelihoods:

$$L_{sf}(D) = \prod_{h \text{ unique}} \exp \left(-\frac{D_A(h)^2}{2v_r(h)} - \frac{D_B(h)^2}{2v_i(h)} - \frac{2D_A(h)D_B(h)}{2v_{ri}(h)} \right) \quad (2)$$

If we take the logarithm of this expression we obtain a log-likelihood, which is a summation over the log-likelihoods for individual reflections::

$$\begin{aligned} LL_{sf}(D) &= \sum_{h \text{ unique}} -\frac{D_A(h)^2}{2v_r(h)} - \frac{D_B(h)^2}{2v_i(h)} - \frac{D_A(h)D_B(h)}{v_{ri}(h)} \\ &= 1/2 \sum_{h \pm \text{unique}} D(h)^2 W_+(h) + D(h)D(-h)W_-(h) \end{aligned}$$

where

$$\begin{aligned} W_+(h) &= \frac{1}{2v_r(h)} - \frac{1}{2v_i(h)} - \frac{2}{2v_{ri}(h)} \\ W_-(h) &= \frac{1}{2v_r(h)} + \frac{1}{2v_i(h)} \end{aligned}$$

and the summation is now over the unique reflections and their Friedel opposites.

Substituting $d(x) = \sum_k D(h) \exp(2\pi i h \cdot x)$, $w_{\pm}(x) = \sum_h W_{\pm}(h) \exp(2\pi i h \cdot x)$:

$$LL_{sf}(d) = c \int \int d(y)d(z)w_+(y+z) + w_-(y-z)dydz \quad (3)$$

This expression gives the change in log-likelihood as the density moves away from the likelihood maximum, which in turn is the Fourier transform of the reciprocal space maximum.

How does the log-likelihood of the map vary as the density at a single point in the map is varied? For a map in spacegroup P1, set $d(y)=0$ for $y \neq x$ (for other space groups all symmetry related densities must be included).

$$LL_{sf}(d(x)) = c d(x)^2 [w_+(2x) + w_-(0)] \quad (4)$$

Note that the density variance, which is inversely proportional to the quadratic coefficient, varies over a cell whose extent is half that of the crystal cell. Suppose that there is a large uncertainty in the magnitude of one reflection, but that its phase is well known. Then the corresponding uncertainty in the electron density will be large wherever the electron density wave corresponding to that reflection has a peak or a trough. Thus the features in the variance map have half the spacing of the features in the electron density map.

3. Structure factor log-likelihood expression

Consider an ensemble of density maps which may be represented by a mean density ρ_0 and a variance σ_ρ at each of N_{var} points distributed throughout the cell, where N_{var} is the effective number of parameters required to represent the map. If the conditional probabilities for values at each sample point are independent, then the likelihood of any particular map $\rho(\mathbf{x})$ is given by:

$$L_{map}(\rho) = \prod_{\mathbf{x}} \exp\left(-\frac{(\rho(\mathbf{x}) - \rho_0(\mathbf{x}))^2}{2\sigma_\rho(\mathbf{x})^2}\right) \quad (5)$$

If we take the log of this function, we obtain the log-likelihood of the map:

$$LL_{map}(\rho) = \sum_{\mathbf{x}} -\frac{(\rho(\mathbf{x}) - \rho_0(\mathbf{x}))^2}{2\sigma_\rho(\mathbf{x})^2} \quad (6)$$

This may be generalised to the continuous electron density by integrating over the cell and re-normalising to correct for the effective number of parameters:

$$LL_{map}(\rho) = \frac{N_{var}}{V} \int -\frac{(\rho(\mathbf{x}) - \rho_0(\mathbf{x}))^2}{2\sigma_\rho(\mathbf{x})^2} d\mathbf{x} \quad (7)$$

Let $d(\mathbf{x}) = -(\rho(\mathbf{x}) - \rho_0(\mathbf{x}))$ and $v(\mathbf{x}) = 1/\sigma_\rho(\mathbf{x})^2$. Then

$$LL_{map}(d) = N_{var}/2V \int v(\mathbf{x})d(\mathbf{x})^2 d\mathbf{x} \quad (8)$$

Replacing $d(x)$ and $v(x)$ with their Fourier transforms $D(h)$ and $V(h)$ we obtain:

$$LL_{map}(D) = c \sum_{\mathbf{k}} \sum_{\mathbf{l}} D(\mathbf{k})D(\mathbf{l})V(-\mathbf{k} - \mathbf{l}) \quad (9)$$

How does LL_{map} as we vary a single reflection, i.e. ignoring inter-reflection correlations? $k \neq \pm h$

Set $D(k)=0$ for $k \neq \pm h$, since the correlation between a reflection and its Friedel opposite may not be ignored. (For spacegroups other than P1, symmetry equivalent reflections should also be included). Then:

$$LL_{map}(D(h)) = c \sum_{\mathbf{k}=\pm h} \sum_{\mathbf{l}=\pm h} D(\mathbf{k})D(\mathbf{l})V(-\mathbf{k} - \mathbf{l}) \quad (10)$$

Again this expression describes the change in log-likelihood as we move away from the maximum, given by the Fourier transform of the real space maximum.

Let $D(\mathbf{h}) = D_A(\mathbf{h}) + iD_B(\mathbf{h})$, $V(\mathbf{h}) = V_A(\mathbf{h}) + iV_B(\mathbf{h})$ and expand:

$$LL_{map}(D(\mathbf{h})) = 2cD_A(\mathbf{h})^2[V_A(0) + V_A(-2\mathbf{h})] + 2cD_B(\mathbf{h})^2[V_A(0) - V_A(-2\mathbf{h})] + 4cD_A(\mathbf{h})D_B(\mathbf{h})V_B(-2\mathbf{h})$$

For a general spacegroup, the expression is as follows:

$$LL_{map}(D(\mathbf{h})) = c \sum_{\pm\mathbf{h}_i} \sum_{\pm\mathbf{h}_j} 2D_A(\mathbf{h}_i)D_A(\mathbf{h}_j)[V_A(\mathbf{h}_i - \mathbf{h}_j) + V_A(-\mathbf{h}_i - \mathbf{h}_j)] + 2D_B(\mathbf{h}_i)D_B(\mathbf{h}_j)[V_A(\mathbf{h}_i - \mathbf{h}_j) - V_A(-\mathbf{h}_i - \mathbf{h}_j)] + 4D_A(\mathbf{h}_i)D_B(\mathbf{h}_j)V_B(-\mathbf{h}_i - \mathbf{h}_j)$$

where the summations are over the symmetry equivalent reflections and their opposites.

The terms in $D_A(\mathbf{h})^2$ and $D_B(\mathbf{h})^2$ are equivalent to the curvature expressions given by Terwilliger (1999) for a phase of 0 or $\pi/2$. The additional cross term allows the semi-major axes of the quadratic function to be oriented in an arbitrary direction in the Argand diagram. This term could easily be incorporated into Terwilliger's expression.

By using the full expression in equation (9) the correlations between reflections may also be calculated.

4. Discussion

The obvious application for this mathematics is phase improvement through real-space constraints, as demonstrated by Terwilliger. The new electron density log-likelihood expressions allow a variance to be calculated for the electron density at any position in the map. This variance may be used in Terwilliger's expressions to determine the probability that a map coordinate is in protein or solvent. It may also be used in non-crystallographic symmetry averaging to give an appropriate weight to the information from other copies of the molecule (with an additional term to account for non-isomorphism between the NCS related molecules).

The function $V(h)$ in equation (10), used in calculating structure factor log-likelihoods, is related to the G-function of Hendrickson and Lattman (1970). Since all the minus-log-likelihood curvatures $v(x)$ in real space are positive, $V(h)$ must be dominated by a large origin term, which contributes to every structure factor variance. $V(h)$ must also be limited in extent, since it is the Fourier transform of the low-resolution mask-outline. The use of a continuously varying probability instead of the traditional binary mask by Terwilliger will further limit the extent of $V(h)$, but this is clearly better than including the meaningless high-resolution ripples introduced by a binary mask.

One notable difference between the approach described here and that of Terwilliger is that the density probability distribution is approximated here before changing from real to reciprocal space. As a result, no Taylor series expansion is required: the resulting

distribution is correct across the whole of the Argand diagram. The disadvantage of this approach is that some information about the probability distribution of density values is discarded. For example, a bimodal probability distribution for the value of the electron density at some point in the map will be approximated by a single broad distribution. The practical implications of these changes have yet to be investigated. However, for the simple application of solvent flattening, the differences between a fixed width Gaussian plus constant and variable width Gaussian model are likely to be small.

It is now possible to ask what information can be gained from a solvent flatness constraint, by considering the variation in log-likelihood across the whole of the Argand diagram for each reflection. Suppose real space probability distributions are constructed to represent flat density in the solvent region and completely unknown density in the protein. The probability distribution for the electron density at any position in the map is therefore a Gaussian centred on the solvent density, whose width increases with the probability that the density belongs to the protein.

Ignoring cross terms between reflections, the likelihood distribution for each structure factor from this information has its maximum at the origin of the Argand diagram, since the Fourier transform of the flat solvent density is zero for all non-origin reflections. Therefore, in the case of unobserved reflections for which both magnitude and phase are unknown, this information alone cannot predict non-zero values for unobserved reflections. In the case of reflections whose magnitude is known, the Gaussian probability centred on the origin can give rise to at best a symmetrical bimodal phase distribution (i.e. in terms of Hendrickson Lattman coefficients, $A=B=0$). Combination of this probability distribution with an experimental phase probability distribution can provide some phase improvement.

Incorporation of the inter-reflection cross terms in the probability function however has the potential to dramatically increase the power of the approach for the following reasons. Suppose one reflection is known in magnitude and phase. Then, neighbouring reflections will be more likely to take values which suppress the contribution of the first structure factor in solvent regions, and as a result add to it in protein regions. The implicit phase relationships in solvent flattening are expressed explicitly in this manner. The calculation may be achieved by constructing a 2-D Gaussian model in the Argand diagram for the experimental phase probabilities on the basis of the structure factor magnitude, its standard deviation, and the Hendrickson-Lattman coefficients. The log of this distribution gives a (block diagonal) matrix of quadratic coefficients, which may be summed with the (full matrix) quadratic coefficients from the map variance information. The combined matrix may then be inverted to determine the shift to reach the likelihood maximum, and the variances of all the parameters. The calculation required the inversion of a matrix whose dimension is the number of reflections, but such a calculation is probably within the range of high-end workstations for small to medium sized structures. Alternatively, some form of sparse matrix approach may well be practical.

Terwilliger's approach of calculating the local curvature around the centroid map calculated from the experimental data provides an alternative and computationally efficient means of avoiding this lengthy calculation for the purposes of finding the likelihood maxima, as long as the shifts to the structure factors remain small or the curvatures are recalculated appropriately. However the incorporation of cross terms may improve the error estimation for the resulting phases.

The formulation described here could be modified to achieve similar efficiency by calculating as many terms as are desired of the curvature once (since it is constant in this

formulation), and then using a gradient/curvature search to find the log likelihood maximum.

A further development of the ideas described here would be to drop the assumption of independence between density values in the map. However it is hard to see where density covariance information might come from except when calculated from experimental phase information. This offers the possibility of completely representing a Gaussian model for every structure factor and phase by a set of variances and covariances in real space, however the applications of such an approach are not obvious.

5. Acknowledgements

This work was funded by the BBSRC, grant number 87/B09875

6. References

1. Abrahams J. P. (1997) Bias reduction in phase refinement by modified interference functions: Introducing the gamma correction *Acta Cryst*, **D53**, pp. 371-376
2. Cowtan K. D., Main P. (1996) Phase combination and cross validation in iterated density modification calculations. *Acta. Cryst.*, **D52**, pp. 43-48
3. Cowtan K. D. (1999) Error Estimation and Bias Correction in Phase Improvement Calculations. *Acta. Cryst.*, **D55**, pp. 1555-1567
4. Hendrickson, W. A. and Lattman, E. E. (1970) *Acta Cryst* **B26**, pp. 136-143.
5. Terwilliger, T. C. (1999) Reciprocal-space solvent flattening *Acta Cryst*, **D55**, pp. 1863-1871

MODELLING OF DISORDERED SOLVENT IN MACROMOLECULAR CRYSTALS

By A. Urzhumtsev

Laboratory of Crystallography and Modeling of Mineral and Biological Materials

University Henri Poincaré, Nancy 1, Vandoeuvre-les-Nancy, France

sacha@lcm3b.u-nancy.fr

1. Atomic models and structure factors

One of the main goals of structural crystallography is, given a crystal, to propose a model which will be chemically consistent and which will explain experimentally measured structure factor magnitudes. Currently, atomic models of spherical (isotropic) or elliptical (anisotropic) individual atoms are accepted as good at conventional resolutions, when the high resolution limit of the data set vary between roughly 1 and 5 Å. When the resolution of a data set is higher than 1 Å, more complicated models such as multipolar ones (Hansen & Coppens, 1978; Lecomte, 1998) are necessary. At the opposite end, at resolutions 5-10 Å and lower, atomic models are useless because they have too many parameters with respect to a very limited number of experimentally measured magnitudes. At the same time, at such resolutions there is another, more important problem. Starting roughly at 7-8 Å, these models being used alone do not explain diffraction data (see, for example, one of the pioneering works by Phillips, 1980).

A reason for such behaviour is the incompleteness of atomic models which miss the contribution of the bulk solvent :

$$F_{\text{obs}} \exp\{i\phi_{\text{obs}}\} = F_{\text{mol}} \exp\{i\phi_{\text{mol}}\} + F_{\text{solv}} \exp\{i\phi_{\text{solv}}\} \quad (1)$$

This resolution limit of 7-8 Å shows that the solvent features are of this size (therefore they are not important at higher resolutions) and that in order to use corresponding diffraction data a model for the bulk solvent should be introduced. A review of possible approaches for such modelling is proposed below; it follows the talk given at the IUCr-18 (Urzhumtsev, 1999). A part of this presentation was inspired by the comparative analyses done by Jiang & Brünger (1994) and by Kostrewa (1997). The current review differs from the former by more schematised presentation and from the latter by including larger material. Other known reviews on the solvent modelling are those by Tronrud (1997) and Badger (1997).

2. Contribution of the bulk solvent : what to do with?

Currently, macromolecular crystallographers use solvent modelling in two extreme cases : for well ordered (crystallographic) solvent molecules which are modelled exactly in the same way as the macromolecule itself and for completely disordered molecules which are modelled by a continuum function. These models are used mainly during atomic model refinement. An intermediate case of partially ordered solvent molecules can occur as it is the case for water channels. In this case also, as for "crystallographic" molecules, an atomic model can be produced (Podjarny *et al.*, 1997) similarly to multicopy models of flexible chains (Burling & Brünger, 1994). On contrary, such (or other) models for the case

of the bulk solvent are not yet developed and some approaches to do so are analysed below. It is important to note that this modelling is important at any stage of crystallographic studies when molecular models appear and not only for a refinement of atomic models.

Different approaches are possible to deal with the problem of the contribution of the bulk solvent to diffraction data or to corresponding crystallographic image:

1. Exclusion of low resolution data from the calculations;
2. Use of information on density distribution in the solvent region for map improvement;
3. Estimation of the solvent contribution statistically;
4. Explicit modelling of the bulk solvent;
5. Use of a specificity of the solvent diffraction for phasing.

The approaches 1-3 and 5 are discussed briefly in this Section, and the rest of the paper deals with different explicit models of the bulk solvent.

Exclusion of low resolution data

Until 90th most of refinements were done excluding reflections of the resolution lower than 6-7 Å. This solved also another, technical problem of low resolution : measuring these reflections. Later an importance of low resolution data for the refinement (see, for example, Kostrewa, 1997) and for the map quality (Urzhumtsev, 1991) became accepted by the crystallographic community and currently this approach is no longer used. However, the same idea is exploited now for molecular replacement where the lower resolution limit used is about 10-15 Å. While such exclusion is not important (and seems to be even useful) for the rotation search, low resolution data can be extremely useful for the solution of the translation problem as it is shown by Urzhumtsev & Podjarny (1995a). A search with low resolution data is less sensitive to eventual orientation errors in the model. Even more, the possibility of low resolution direct phasing (see, for example, Lunin *et al.*, 2000) can give a complementary way to solve the translation and in some case, the rotation problem.

Solvent flattening or relevant techniques

A known information on the density distribution in the solvent region can be used in order to improve available electron density maps. In most of cases, such distribution is supposed to be flat, and a number of solvent flattening techniques exist now being originated in its modern form by Bricogne (1974). This information can be used alone or together with other informations, for example, with a known histogram for the electron density distribution inside the molecular envelope (Zhang & Main, 1990).

It is important to remind that if the "observed" phases ϕ_{obs} would be available the map calculated with them will not have a flat density in the solvent regions because of a limited resolution (Fourier truncation effects) and a usual absence of some reflections, specially those of very low resolution. In other words, flattening procedures look for some *wrong* phases which nevertheless produce a more interpretable image. Such effect was analysed by Brazhnikov *et al.* (1993) when the same quality of an improved map was obtained by a simple reconstruction and addition to the synthesis of few low resolution reflections or by iterative application of the solvent flattening procedure (Wang, 1985).

In fact, the most interpretable macromolecular map would be calculated rather with the coefficients

$$F_{\text{mod}} \exp\{i\phi_{\text{mod}}\} = F_{\text{obs}} \exp\{i\phi_{\text{obs}}\} - F_{\text{solv}} \exp\{i\phi_{\text{solv}}\} \quad (2)$$

and not with

$$F_{\text{obs}} \exp\{i\phi_{\text{mod}}\} \quad (3)$$

where the phases ϕ_{mod} are found by some phase improvement technique. Naturally, in practice when the phases ϕ_{obs} are unknown, the best known their approximation ψ_{obs} should be used:

$$F_{\text{obs}} \exp\{i\psi_{\text{obs}}\} - F_{\text{solv}} \exp\{i\phi_{\text{solv}}\} \quad (4)$$

Maps with the coefficients (4) could be used *before* an atomic model is known in order to facilitate its construction and the map interpretation. Therefore, the first problem of the solvent modelling is:

Problem 1: how to estimate the bulk solvent contribution *before* an atomic model is known?

The goal here is to improve a map in order to build an atomic model of the molecule

Statistical approach

The equality (1) can be treated in a such way that a complete model of a crystal consists of a macromolecular model plus a number of atoms, ordered or not, which are not yet identified and not yet taken into account:

$$F_{\text{obs}} \exp\{i\phi_{\text{obs}}\} = F_{\text{mol}} \exp\{i\phi_{\text{mol}}\} + F_{\text{abs}} \exp\{i\phi_{\text{abs}}\} \quad (5)$$

Under some statistical hypotheses on possible distribution of these absent atoms, the mean value for correcting structure factors $F_{\text{abs}} \exp\{i\phi_{\text{abs}}\}$ can be estimated and therefore the best macromolecular model (the most probable hypothesis on values of atomic parameters) can be chosen using maximum likelihood approach (Bricogne & Irwin, 1996; Pannu & Read, 1996; Read, 1997; Murshudov *et al.*, 1997; Pannu *et al.*, 1998; some comments on this subject can be found in Lunin & Urzhumtsev, 1999). One can accept a hypothesis that the missed atoms are, at the first approximation, randomly and uniformly distributed in the unit cell or another, better hypothesis, for example that they are distributed only in the solvent region, and, maybe, not uniformly. In any case, when refining an atomic model with the maximum likelihood approach, model structure factors F_{mod} are fitted to the values

$$\langle |F_{\text{obs}} \exp\{i\phi_{\text{obs}}\} - F_{\text{abs}} \exp\{i\phi_{\text{abs}}\}| \rangle \quad (6)$$

averaged over all possible positions of missed atoms (in the current study, supposed to be in the solvent region) and no longer to the experimental values F_{obs} .

In such approach the knowledge of an approximate atomic model for the macromolecule is crucial and the problem differs from the Problem 1 discussed above:

Problem 2: how to estimate the bulk solvent contribution when an atomic model is known?

The goal here is to obtain the best possible atomic model of the whole crystal which can be used in order to answer different physical, chemical or biological questions.

Phasing methods based on specific solvent diffraction

In contrast to previous, "passive" treatment of the solvent, an "active" way of doing this is possible where a variation of the solvent density allows to get phase values (contrast variation methods; Bragg & Perutz, 1952; Roth, 1987, 1991; Carter *et al.*, 1990). Alternatively, the phases can be obtained from the same solvent but diffracting anomalously (MASK method by Fourme *et al.*, 1994). These methods need to be analysed independently and will not be discussed here.

3. Determination of the solvent region.

Molecular and solvent regions

The most direct way, supplying with most interesting information but not easy to realise is to build an explicit model of solvent. There are several suggested approaches discussed below.

In general, a model of the bulk solvent can be built both in real and in reciprocal space. In reciprocal space, it is a set of solvent structure factors, usually with the same indices as for the set of experimentally measured data while eventually it can be more complete or go to slightly higher resolution. In real space, a solvent model is usually a density distribution. This distribution can be either ideal or approximate at a given resolution.

In order to build a bulk solvent model in real space, one needs to define two objects :

- a) solvent region and
- b) density distribution inside this region.

Solvent region usually is considered as a part of the unit cell complementary to the molecular region (or molecular mask). Sometimes, a thin shell between them is initially excluded from both and then filled with more complicated procedures (Jiang & Brünger, 1994). Traditionally, the molecular region is represented by a binary function $M(\mathbf{r})$ calculated in a chosen grid. Correspondingly to the two problems formulated in the previous section, two different situations are possible when an atomic model is known or not.

In the first case, the molecular mask $M(\mathbf{r})$ is traditionally defined as a conjunction of spheres of a given radius centred in the atomic positions. This mask is final, exact (within the limits of the atomic positional errors, their radius etc).

In the second case, the molecular mask can be defined from an electron density distribution $\rho_d(\mathbf{r})$ calculated at the resolution d . Molecular (solvent) region defined on the base of a density distribution does not have any longer an absolute meaning as $M(\mathbf{r})$. Since all details with the resolution higher than d is absent in the map, the best mask image also cannot have high resolution details. The best possible envelope $M_d(\mathbf{r})$ at the

resolution d can be imagined, for example, as a result of calculation of structure factors from $M(\mathbf{r})$, suppression of high resolution reflections, calculation of a new function with the rest of structure factors and, finally, of a selection of points with highest values of this new Fourier synthesis.

Naturally, when neither molecular mask $M(\mathbf{r})$ nor an atomic models are known, the ideal $M_d(\mathbf{r})$ cannot be calculated and an approximation $m_d(\mathbf{r})$ to it can be defined as a set of (grid) points with

$$m_d(\mathbf{r}) = \{\mathbf{r} : \rho_d(\mathbf{r}) \geq \rho^*\} \quad (7)$$

where ρ^* is chosen such that the volume of the region $m_d(\mathbf{r})$ is roughly equal to the molecular volume. Clearly, the result strongly depends on the resolution and on the quality of the initial density $\rho_d(\mathbf{r})$ and, in particular, on errors in structure factors and on absence of some of them (Fig. 1). In many cases this procedure is not good because it gives a multidomain regions that does not agree with the idea to have a molecular mask as a single domain.

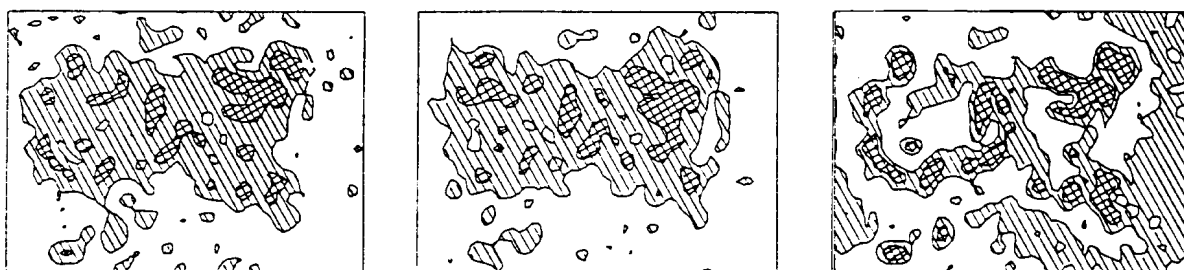


Fig. 1. Influence of the phase errors and the exclusion of low resolution data on the quality of molecular envelope; test model (Urzhumtsev, 1991). ideal map at 6Å resolution, about 2500 reflections used; external contour selects a half of the unit cell (left); SIR phased map with the same set of reflections (centre); the same structure factors as for the left figure except all 29 reflections with the resolution lower than 30Å excluded from the map calculation (right)

Map features and automatic mask determination

In order to determine the molecular mask from a density distribution, some key density features which allow to distinguish the macromolecular and solvent regions should be formulated. For every grid point where the density is calculated, these features should allow to make a choice whether the point belongs to the molecular region or not, or, in more soft way, with which probability it belongs to the molecular region. In other words, a density transformation

$$\rho_d(\mathbf{r}) \rightarrow p(\mathbf{r}) \quad (8)$$

should be defined which for every point \mathbf{r} assigns a probability value $p(\mathbf{r})$ instead of the Fourier synthesis value $\rho_d(\mathbf{r})$. In the simplest case (7) discussed above, this is a binary function

$$p(\mathbf{r}) = P[\rho_d(\mathbf{r})] = \begin{cases} 1, & \text{if } \rho_d(\mathbf{r}) \geq \rho^* \\ 0, & \text{if } \rho_d(\mathbf{r}) < \rho^* \end{cases} \quad (9)$$

which depends on the value ρ in a given point \mathbf{r} . In the example of the density modification function suggested by Wang (1985) it is:

$$\mathbf{p}(\mathbf{r}) = \mathbf{p}[\rho_d(\mathbf{r})] = \begin{cases} \kappa[\rho_d(\mathbf{r}) - \rho^*], & \text{if } \rho_d(\mathbf{r}) \geq \rho^* \\ 0, & \text{if } \rho_d(\mathbf{r}) < \rho^* \end{cases} \quad (10)$$

where κ is a normalising factor. This step is logically executed in direct space and can be considered as a corresponding filtration procedure for $\rho_d(\mathbf{r})$.

For noisy maps which is usually the case, such procedure is not sufficiently good and does not always give a single domain regions suggesting further treatment of this information. For example, for the binary selection (9), the molecular region can be considered as a zone of highest concentration of the selected points. If the resulted function $\rho(\mathbf{r})$ is considered as a probability, then the molecular region can be defined as a zone where a whole sphere of a given radius likely belongs to the molecule. Both reasoning can be realised through a local averaging of the transformed function $\rho(\mathbf{r})$:

$$\mathbf{p}(\mathbf{r}) \rightarrow \tau(\mathbf{r}) = \langle \mathbf{p}(\mathbf{r}) \rangle_{\text{points around } \mathbf{r}} \quad (11)$$

An understanding and implementation of this second step allowed to create an automated approach for envelope determination (Westbrook *et al.*, 1984; Wang, 1985; Urzhumtsev, 1985; Jones *et al.*, 1991). The result of averaging varies but not strongly with different weighting functions used during averaging; on contrary, the correct choice of the averaging radius is more important (Urzhumtsev, 1991). This step is easier to be done through structure factors calculation (Leslie, 1987; Lunin, in Urzhumtsev, 1985) and can be considered as a filtration procedure in reciprocal space.

Now, with a correct choice of the filtration parameters, a selection of highest probability points in $\tau(\mathbf{r})$:

$$\mathbf{m}(\mathbf{r}) = \{\mathbf{r} : \tau(\mathbf{r}) \geq \tau^*\} \quad (12)$$

constructs a region, which consists of a single domain per molecule and has quite low probability of errors.

In spite of importance of the second step of averaging, it seems that it is the first step of density filtration (9-10) which plays the crucial role through the information used for the selection of points. Highest density values used by Wang (1985) and Westbrook *et al.* (1984) is one example of such information (Fig. 2). It was noted by Urzhumtsev *et al.* (1989) that due to the truncation effects the points with lowest density values at a synthesis of a finite resolution also indicate the molecular region (Fig. 2). A similar observation was used by Jones *et al.* (1991) who noted that the points inside the molecular region correspond to highest local density fluctuation; in this case the density filtration function depends not only on a density value in a given point like in (9) or (10) but on value in several points.

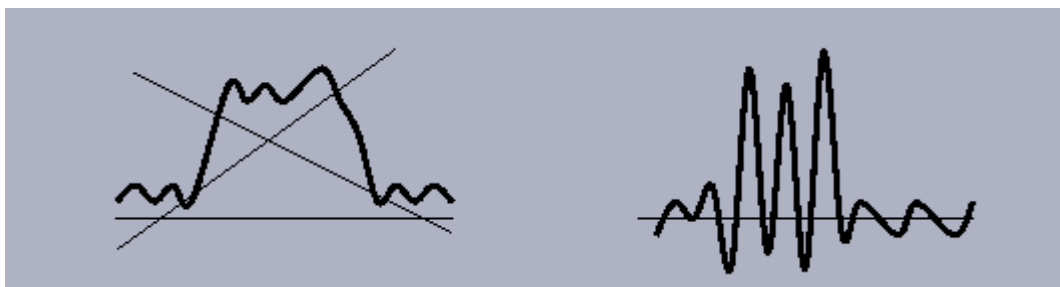


Fig. 2. Schematic one-dimensional representation of the distribution of the values of the Fourier synthesis calculated at a limited resolution. In many cases, molecular region contains not only highest density values but also lowest density values and has points with highest local density fluctuation.

For cases with a known non crystallographic symmetry, some automated procedures were suggested (Rees *et al.*, 1990) based on the similarity of the density values for the symmetrically related points differentiating them from the points of solvent for which this symmetry is not applicable.

4. Explicit modelling of the solvent contribution.

Flat envelope. Mask method.

When the solvent region is determined, the next step is to construct a density distribution inside this region. This problem was addressed several times from the beginning of 50ths (Wrinch, 1950; Bragg & Perutz, 1952; Langridge *et al.*, 1960; Fraser *et al.*, 1965, 1978; O'Brien & MacEwan, 1970) for several particular problems. Later Phillips (1980) published a clear evidence of a discrepancy between experimental and calculated structure factors magnitudes for a macromolecular case, indicated resolution limits for this and proposed a scheme of a correction for the bulk solvent which with some variations is used until nowadays:

- knowing an atomic model, calculate a binary molecular mask (1 inside molecular region and 0 outside it);
- calculate structure factors $F_{\text{env}}(\mathbf{s}) \exp\{i\phi_{\text{env}}(\mathbf{s})\}$ for this flat molecular mask; following the Babinet principle, structure factors for a flat solvent region are opposite to $F_{\text{env}}(\mathbf{s}) \exp\{i\phi_{\text{env}}(\mathbf{s})\}$;
- estimate solvent contribution as weighted (scaled) values obtained at the previous step

$$F_{\text{solv}}(\mathbf{s}) \exp\{i\phi_{\text{solv}}(\mathbf{s})\} = -\lambda(|\mathbf{s}|) F_{\text{env}}(\mathbf{s}) \exp\{i\phi_{\text{env}}(\mathbf{s})\} \quad (13)$$

- calculate corrected structure factors for the whole crystal

$$F_{\text{calc}}(\mathbf{s}) \exp\{i\phi_{\text{calc}}(\mathbf{s})\} = F_{\text{mod}}(\mathbf{s}) \exp\{i\phi_{\text{mod}}(\mathbf{s})\} - \lambda(|\mathbf{s}|) F_{\text{env}}(\mathbf{s}) \exp\{i\phi_{\text{env}}(\mathbf{s})\} \quad (14)$$

The basic hypothesis of this scheme is a flat density distribution in the solvent region. Here, a macromolecular atomic model is supposed to be known and $\lambda(|\mathbf{s}|)$ is a function, usually gaussian, of the resolution:

$$\lambda(|\mathbf{s}|) = \kappa_{\text{solv}} \exp\{-B_{\text{solv}} s^2/4\} \quad (15)$$

Its parameters (κ_{solv} and B_{solv} in this particular example) are chosen from the best fit of $F_{\text{calc}}(\mathbf{s})$ to $F_{\text{obs}}(\mathbf{s})$ for a given fixed macromolecular model. This approximation may be not always efficient. An example is the case of aldose reductase (Rondeau *et al.*, 1992) for which the mean $\lambda(|\mathbf{s}|)$ value varies rather as a sigmoid and not as a gaussian function (Fig. 3). On the other hand, the approximation (15) is important only up to the resolution of about 5 Å above which the solvent contribution is negligible; in this sense, the case of aldose reductase does not really contradict the approximation. Another possibility for $\lambda(|\mathbf{s}|)$ is a local scaling discussed, for example, by Tronrud (1997).

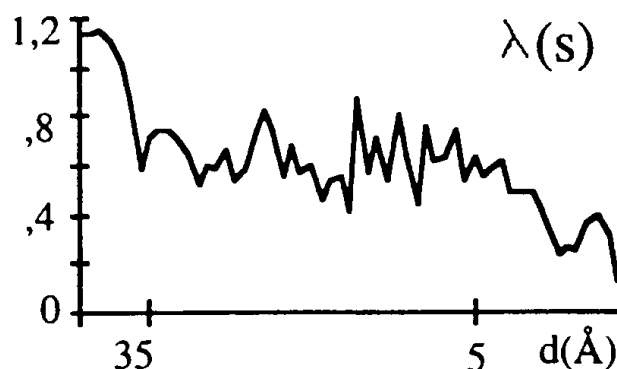


Fig.3. Optimal scaling coefficient $\lambda(|\mathbf{s}|)$ for the envelope structure factors as a function of the resolution.

Among the variations introduced into the initial scheme, a special procedure to assign the density values for the points of a narrow shell between the molecular and solvent regions (Jiang & Brünger, 1994) is of high importance (Kostrewa, 1997).

While the method, realised for example through X-PLOR (Brünger, 1992) and CNS (Brünger *et al.*, 1998), allows to reduce dramatically the discrepancy between observed and calculated low resolution structure factors magnitudes, a detailed analysis by Jiang & Brünger (1994) allowed them to conclude that "the best solvent models results in R-factors significantly higher than one might expect". Therefore, two questions arise :

- if the method is not optimal anyhow, whether some its simplifications can be done ?
- can the method be improved (for example, by more complicated procedures) ?

Flat envelope. Exponential scaling model.

A simplification of the previously described method was suggested following the argumentation that the bulk solvent correction is important only at low resolution where density both in macromolecular and in solvent region can be considered as flat. In this case, structure factors calculated for two binary (flat) functions, complementary in the unit cell, are proportional each to other (see, for example, Langridge *et al.*, 1960 or Tronrud, 1997):

$$\begin{aligned}
 F_{\text{calc}}(\mathbf{s}) \exp\{i\phi_{\text{calc}}(\mathbf{s})\} &= F_{\text{mod}}(\mathbf{s}) \exp\{i\phi_{\text{mod}}(\mathbf{s})\} + F_{\text{solv}}(\mathbf{s}) \exp\{i\phi_{\text{solv}}(\mathbf{s})\} \\
 &\approx F_{\text{env}}(\mathbf{s}) \exp\{i\phi_{\text{env}}(\mathbf{s})\} - \lambda(|\mathbf{s}|) F_{\text{env}}(\mathbf{s}) \exp\{i\phi_{\text{env}}(\mathbf{s})\} \\
 &= F_{\text{env}}(\mathbf{s}) \exp\{i\phi_{\text{env}}(\mathbf{s})\} \{1 - \lambda(|\mathbf{s}|)\} \\
 &\approx F_{\text{mod}}(\mathbf{s}) \exp\{i\phi_{\text{mod}}(\mathbf{s})\} \{1 - \lambda(|\mathbf{s}|)\}
 \end{aligned} \tag{16}$$

($F_{\text{mod}}(\mathbf{s})$ and $F_{\text{env}}(\mathbf{s})$ are supposed to be quite close).

Because the $\lambda(|\mathbf{s}|)$ values usually vary between 0 and 1, this formula (16) illustrates the phenomenon that at low resolution observed structure factor magnitudes are lower than those calculated from the macromolecular model. The formula (16) suggests also two ways to diminish the discrepancy between mean values of magnitudes: either observed magnitudes $F_{\text{obs}}(\mathbf{s})$ should be corrected as

$$F_{\text{obs}} \rightarrow F_{\text{obs}} / (1 - \text{scale}) \quad (17)$$

before being compared with $F_{\text{mod}}(\mathbf{s})$, or the model structure factors should be multiplied by $\{1 - \lambda(|\mathbf{s}|)\}$. The latter can be also done by correction of atomic scattering factors (Fraser *et al.*, 1978).

The key question in this approach is whether such linearity between molecular and envelope structure factors holds up to a reasonable resolution. Some tests were done with the diffraction data for aldose reductase (Rondeau *et al.*, 1992). For this case of a well refined atomic model, solvent structure factors as complex numbers were estimated and compared with the molecular structure factors (Urzhumtsev & Podjarny, 1995b; Podjarny & Urzhumtsev, 1997). The results of such comparison (Fig. 4) can be interpreted so that these structure factors are approximately anti-collinear at the resolution lower than 12-15 Å, and are not correlated at higher resolutions. This shows that the scaling method has a limited range of application in comparison with the mask method what was found experimentally (Kostrewa, 1997).

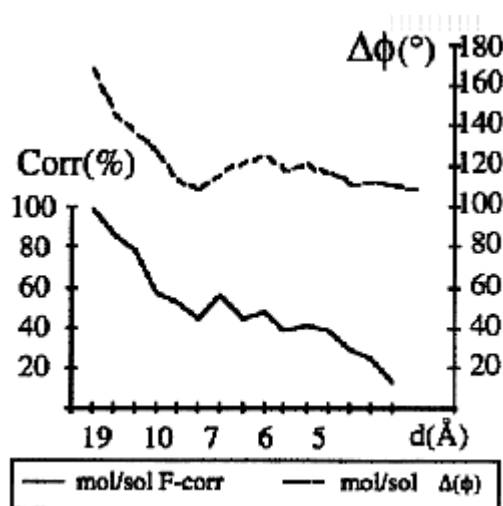


Fig.4. Comparison of model and solvent structure factors. Amplitude correlation and phase difference

Distance dependent solvent modelling

In order to improve the quality of solvent modelling obtained by the mask method, several efforts were undertaken. The first idea was to replace a flat solvent density by some more complicated distribution. As it is known, the solvent molecules are distributed in shells, and a logical step is to model the solvent density as a function dependent on the distance r to the molecular border (Schoenborn, 1988; Cheng & Schoenborn, 1990).

$$\rho_{\text{solv}}(\mathbf{r}) = \text{const} \rightarrow \rho_{\text{solv}}(\mathbf{r}) = \rho(r) \quad (18)$$

In this case, the knowledge of a molecular model is important in order to reproduce correctly the shells. While such idea is very attractive, an analysis done by Jiang & Brünger (1994) showed that the improvement obtained by this method is marginal in comparison with the mask method.

One of possible explanation is that in fact while the mean density in the shells correlates with the distance to the molecular border, the distribution of density in every shell is far from be uniform. In other words, the density in the points with the same distance to the molecular border is rather different and can depend on the shape of the border (cavity, channel, etc.; a nice illustration for this can be obtained in Levitt & Park, 1993). Such points can be distinguished if the molecular envelope $M_d(\mathbf{r})$ or its approximation $m_d(\mathbf{r})$ is calculated at several different resolutions d (Fig. 5; see also Section 3 for $M_d(\mathbf{r})$ calculation). Small cavities, after the map is recalculated without higher resolution reflections, become hidden, and corresponding points outside the envelope become inside it.

An additional advantage of such approach is that eventually it does not need a molecular atomic model. If molecular envelopes in the crystal are known at the resolutions d and below, this could be eventually enough to reproduce the solvent density distribution at the same resolution d (or, maybe, slightly lower).

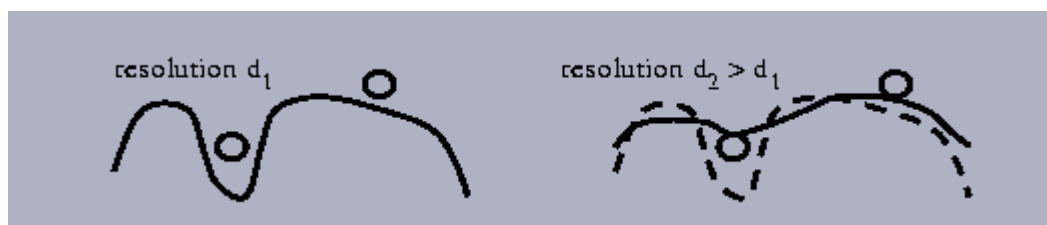


Fig. 5. Schematic presentation of the difference in the relative position of water molecules with respect to the molecular envelope calculated at different resolution.

2D density histograms

In order to use these ideas, some initial information on typical density distributions should be collected. This information can be obtained and reproduced using two-dimensional distance-dependent density histograms. If for a given crystal a molecular envelope can be calculated at several resolution d_n , $n = 1, \dots, N$, then for a density map $\rho_D(\mathbf{r})$ calculated at the resolution D and for every such envelope $M_n(\mathbf{r})$ a common distribution of (ρ, r) can be calculated where r is the distance from a given point \mathbf{r} to the molecular envelope and ρ is the density value in this point. Par analogy with usual density histograms $H_D(\rho)$ (Lunin, 1988), these common distributions can be considered as two-dimensional histograms $H_{D,d}(\rho, r)$.

This information can be used in the following way (Urzhumtsev & Podjarny, 1995b; Urzhumtsev *et al.*, 2000).

If a single envelope $M_d(\mathbf{r})$ and the corresponding 2D-histogram $H_{D,d}(\rho, r)$ are known, then for every point \mathbf{r} in the unit cell, first, its distance r^* to the envelope is calculated, second, a one-dimensional density distribution $h_{D,d}(\rho) = H_{D,d}(\rho, r^*)$ corresponding to this distance is extracted from $H_{D,d}(\rho, r)$ and, third, the density value in this point is assigned to be equal to the mean (or to the most probable) value of this distribution $h_{D,d}(\rho)$. The procedure is essentially the same as the procedure suggested by Schoenborn (1988) with the difference that it does not need to know an atomic model and reconstructs not the ideal solvent density distribution but its image at the resolution D .

However, when several envelopes $M_n(\mathbf{r})$ and histograms $H_{D,n}(\rho, r)$ are known for the resolutions d_n , $n = 1, \dots, N$, then for every point \mathbf{r} a set of one-dimensional distributions $h_{D,n}(\rho) = H_{D,n}(\rho, r_n^*)$ can be obtained that allows to identify better the position of this point with respect to the molecular surface as it is discussed above. These distributions can be multiplied (this is a very crude treatment of the information because naturally these distributions $h_{D,n}(\rho)$ are not independent) and, again, the mean (or the most probable) value of this product can be assigned as the density value in this point. It is easy to see that this approach can have larger applications and be used for the density reconstruction in the whole unit cell and not only for the solvent.

The first analysis done with the experimental data for aldose reductase (Rondeau *et al.*, 1992) showed that indeed such approach allows reasonably well to reconstruct the density but the quality of solvent models, in spite of more detailed density function, is not better either than the result obtained by the "standard" mask approach.

Atomic modelling

An idea to build an atomic model for the bulk solvent is attractive as a logical continuation of a models for "crystallographic" and for partially ordered water molecules (Podjarny *et al.*, 1997). Some variant was tested quite a long ago (O'Brien & MacEwan, 1970) when random atoms were placed inside the macromolecular envelope in order to calculate structure factors using the Babinet principle. However, a use of "bulk solvent atoms" together with the macromolecular atoms in general is not possible because it essentially increases the number of parameters and makes it too high with respect to the number of experimental data. Therefore, some more delicate approaches could be proposed in future.

Difference density approach

Several ideas to construct a density model are based on the analysis of the difference density which is not yet interpreted by the current macromolecular model. Badger & Caspar (1991) and Badger (1993) suggested to include iteratively the peaks in the solvent region of the difference map

$$(F_{\text{obs}} - F_{\text{calc}}) \exp(i\phi_{\text{calc}}) \quad (19)$$

into the solvent model and to suppress these peaks if necessary in following iterations. This idea reminds an approach for direct phasing described by Simonov (1976) with the difference that here the density is analysed only in the solvent region. An opposite, in certain sense, hypothesis that the large positive and negative peaks do not correspond to any features but exist due to important phase errors and should be removed from the map was used for solvent correction by Jiang & Brünger (1994). However, the same authors found both difference density approaches overfitting the model and giving quite marginal improvement in comparison with the mask method.

5. Conclusions

The mask approach is a simple method for solvent modelling which gives a reasonable estimation for bulk solvent contribution. However, this model is not yet sufficiently good and its application needs a knowledge of a macromolecular atomic model.

Currently, several problems of macromolecular crystallography are relevant to the problem of the bulk solvent modelling. First, the work at extra high resolutions with more structural

details and studies of molecular potentials needs *better* solvent models. Second, a work with crystals of not sufficient quality and an interpretation of corresponding density maps could be advanced if the solvent contribution can be estimated and removed from the maps *before an atomic model is known*. Finally, a *fast estimation* of the solvent contribution could be useful in the search procedures where many different molecular positions must be checked like in molecular replacement. These problems of solvent modelling are open for further researches.

Acknowledgements

The author thank D.Kostrewa, V.Lunin and A.Podjarny for many useful discussions on the subject and critical discussion of the manuscript and C.Lecomte for his interest to this work.

References

- Badger, J. (1993) Multiple hydration layers in cubic insulin crystals. *Biophys. J.* **65**, 1656-1659.
- Badger, J. (1997) Modeling and Refinement of Water Molecules and Disordered Solvent. In *Methods in Enzymology*, Academic Press, San Diego., C.W.Carter, Jr., R.M.Sweet, eds. **277B**, 344-352
- Badger, J. & Caspar, D.L.D. (1991) Water structure in cubic crystal. *Proc.Natl.Acad.Sci.USA* **88**, 622-626.
- Bragg, W.L. & Perutz, M.F. (1952). The External Form of the Haemoglobin Molecule. I. *Acta Cryst.* **5**, 277-283.
- Brazhnikov, E., Chirgadze, Yu., Aevansson, A., Svensson, A., & Lunin, V. (1993). The Similarity of Results Obtained by Solvent Flattening and Low-Resolution Phase Retrieval for the Improvement of Protein Electron Density Map. *3rd European Workshop of Biological Macromolecules*, Centro di Cultura Scientifica «A.Volta», Como (Italy), May 21-25, 1993. M4.
- Bricogne, G. (1974) Geometric Sources of Redundancy in Intensity Data and Their Use for Phase Determination. *Acta Cryst.* **A30**, 395-405.
- Bricogne, G. & Irwin, J. (1996) In *Macromolecular Refinement*. Proceeding of the CCP4 Study Weekend, E.Dodson, M.Moore, A.Ralph & S.Bailey, eds., pp.85-92. Warrington : Daresbury Laboratory.
- Brünger, A. T. (1992) *X-Plor, Version 3.1: A System for X-ray Crystallography and NMR*. Yale University Press, New Haven, CT
- Brünger, A.T., Adams, P.D., Clore, G.M., DeLago, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. & Warren, G.L. (1998) Crystallography & NMR System: A new Software Suite for Macromolecular Structure Determination. *Acta Cryst.*, **D54**, 905-921.
- Burling, F.T. & Brünger, A.T. (1994) Thermal motion and conformational disorder in protein crystals structures: Comparison of multi-conformer and time-averaging models. *Isr.J.Chem.*, **34**, 165-175.
- Carter C.W., Jr., Crumbley, K.V., Coleman, D.E., Hage, F. & Bricogne, G. (1990) Direct Phase Determination for the Molecular Envelope of Tryptophanyl-tRNA Synthetase from *Bacillus stearothermophilus* by X-ray Contrast Variation. *Acta Cryst.* **A46**, 57-68.
- Cheng, X.D. & Schoenborn, B.P. (1990) Hydration in protein crystals. A neutron diffraction analysis of carbonmonoxymyoglobin. *Acta Cryst.*, **B46**, 195-208
- Fraser, R.D.B., MacRae, T.P. & Miller, A. (1965) X-Ray diffraction Patterns of α -Fibrous Proteins. *J.Mol.Biol.*, **14**, 432-442

- Fraser, R.D.B., MacRae, T.P & Suzuki, E. (1978) An Improved Method for Calculating the Contribution of Solvent to the X-ray Diffraction Pattern of Biological Molecules. *J.Appl.Cryst.*, **11**, 693 -694
- Fourme, R., Shepard, W., Hermite, G.L. & Kahn, R. (1994) The Multiwavelength Anomalous Solvent Contrast Method (MASC) in Macromolecular Crystallography In "ACA Annual Meeting, June 25 - July 1, 1994", p.40. INFORUM, Atlanta, Georgia, Atlanta Convention Center..
- Hansen, N.K. & Coppens, P. (1978) Testing Aspherical Atom Refinement on Small-Molecule Data Sets. *Acta Cryst.*, **A34**, 909-921.
- Jiang, J.-S. & Brünger, A.T. (1994) Protein Hydration Observed by X-ray Diffraction. Solvation Properties of Penicillopepsin and Neuraminidase Crystal Structures. *J.Mol.Biol.*, **243**, 100-115
- Jones, E.Y., Walker, N.P. & Stuart, D.I. (1991) Methodology Employed for the Structure Determination of Tumour Necrosis Factor, a Case of High Non-Crystallographic Symmetry. *Acta Cryst.* **A47**, 753-770.
- Kostrewa, D. (1997) Bulk Solvent Correction : Practical Application and Effects in Reciprocal and Real Space. *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography*, **34**, 9-22.
- Langridge, R., Marvin, D.A., Seeds, W.E., Wilson, H.R., Hooper, C.W., Wilkins, M.H.F. & Hamilton, L.D. (1960) The Molecular Configuration of Deoxyribonucleic Acid II. Molecular Models and their Fourier Transforms *J.Mol.Biol.*, **2**, 38-64
- Lecomte, C. (1998) *NATO ASI and Euroconference*, NATO ASI Book, Kluwer Acad.Pub.Netherlands
- Leslie, A.G.V. (1987) A reciprocal-space method for calculating a molecular envelope using the algorithm of B.C.Wang. *Acta Cryst.* **A43**, 134-136.
- Levitt, M. & Park, B.P. (1993) Water: now you see it, now you don't. *Structure*, **1**, 223-226
- Lunin, V.Yu. (1988) Use of the Information on Electron Density Distribution in Macromolecules. *Acta Cryst.* **A44**, 144-150.
- Lunin, V.Y. & Urzhumtsev, A.G. (1999) Maximal Likelihood refinement. It works, but why ? *CCP4 Newsletter on Protein Crystallography*, **35**, 8, 13-28
- Lunin, V.Yu., Podjarny, A. & Urzhumtsev, A.G. (2000) Low resolution phasing : advances and perspectives. *Acta Cryst.*, **D56**, sent to editor
- Murshudov, G.N., Vagin, A.A. & Dodson, E.J. (1997) Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Cryst.* **D53**, 240-255.
- O'Brien, E.J. & MacEwan, A.W. (1970) Molecular and Crystal Structure of the Polynucleotide Complex: Polynosinic Acid plus Polydeoxycydic Acid. *J.Mol.Biol.*, **48**, 243-261
- Pannu, N.S. & Read, R.J. (1996) Improved Structure Refinement Through Maximum Likelihood. *Acta Cryst.* **A52**, 659-668.
- Pannu, N.S., Murshudov, G.N., Dodson, E.J. & Read, R.J. (1998) Incorporation of Prior Phase Information Strengthens Maximum-Likelihood Structure Refinement. *Acta Cryst.* **D54**, 1285-1294
- Phillips, S.E.V. (1980) Structure and Refinement of Oxymyoglobin at 1.6Å Resolution. *J.Mol.Biol.*, **142**, 531-554
- Podjarny, A.D. & Urzhumtsev, A.G. (1997) Low resolution phasing. In *Methods in Enzymology*, Academic Press, San Diego., C.W.Carter, Jr., R.M.Sweet, eds. **276A**, 641-658
- Podjarny, A.D., Howard, E.I., Urzhumtsev, A.G. & Grigera, J.R. (1997) A multicopy modeling of the water distribution in macromolecular crystals; towards the compatibility between diffraction and NMR data. *Proteins*, **28**, 303-312

- Read, R.J. (1997) Model phases : Probabilities and Bias. In *Methods in Enzymology*, Academic Press, San Diego., C.W.Carter, Jr., R.M.Sweet, eds., **277B**, 110-128.
- Rees, B., Bilwes, A., Samama, J.P. & Moras, D. (1990) Cardiotoxin V⁴ form *Naja mosambica mosambica*. The refined crystal structure. *J.Mol.Biol.*, **214**, 281-297
- Roth, M. (1987) Best Density Maps in Low-Resolution Crystallography with Contrast Variation. *Acta Cryst. A***43**, 780-787.
- Roth, M. (1991) Phasing at Low Resolution. In "*Crystallographic Computing 5*", Moras, D., Podjarny, A.D. and Thierry, J.-C., eds., p.229-248. Oxford University press.
- Rondeau, J.-M., Tete-Favier, F., Podjarny, A., Reymann, J.-M., Barth, P., Biellmann, J.-F. & Moras, D. (1992) Novel NADPH-binding domain revealed by the crystal structure of aldose reductase. *Nature*, **355**, 469-472.
- Schoenborn, B.P. (1988) Solvent effect in protein crystals. A neutron diffraction analysis of solvent and ion density. *J.Mol.Biol.*, **201**, 741
- Simonov, V.I. (1976) Phase refinement by the method of modification and Fourier transformation of an approximate electron density distributions. In : *Crystallographic Computing Techniques*, Ahmed, F.R., Huml, K., Sedlacek, B., eds. Copenhagen : Munksgaard, 138-143
- Tronrud D. (1997) TNT Refinement Package. In *Methods in Enzymology*, Academic Press, San Diego., C.W.Carter, Jr., R.M.Sweet, eds., **277B**, 306-319.
- Urzhumtsev, A.G. (1985). *The Use of Local Averaging to Analyze Macromolecular Images in the Electron Density Maps*. ONTI NCBI, USSR Acad. of Sci., Pushchino.
- Urzhumtsev, A.G. (1991) Low-Resolution Phases: Influence on SIR Syntheses and Retrieval with Double-Step Filtration. *Acta Cryst.*, **A47**, 794-801.
- Urzhumtsev, A.G. (1999) Solvent modelling at low resolution. *Collected Abstracts, XVIIIth IUCr Congress & General Assembly, 4-13 August 1999, Glasgow, Scotland*, M09.AA.003, 122
- Urzhumtsev, A.G., Lunin, V.Yu. & Luzyanina, T.B. (1989) Bounding a Molecule in a Noisy Synthesis. *Acta Cryst.*, **A45**, 34-39.
- Urzhumtsev, A.G. & Podjarny, A.D. (1995^a) On the Solution of the Molecular-Replacement Problem at Very Low Resolution: Application to Large Complexes. *Acta Cryst*, **D51**, 888-895.
- Urzhumtsev, A.G. & Podjarny, A.D. (1995^b) On the problem of solvent modeling in macromolecular crystals using diffraction data: 1. The low-resolution range. *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography*, **32**, 12-16.
- Urzhumtsev, A.G., Podjarny, A., Lunin, V.Yu. (2000) Density information and low resolution phasing. *Acta Cryst.*, **D56**, sent to editor
- Wrinch, D. (1950) Vector maps of hydrated protein crystals. *Acta Cryst.*, **3**, 475-476
- Wang, B.C. (1985) Resolution of phase ambiguity *Methods Enzymol.*, **115**, 90-112.
- Westbrook, E.W., Piro, O.E. & Sigler, P.B. (1984) The 6-Å crystal structure of delta5-3-ketosteroid isomerase *J.Biol.Chem.* **259**, 9096-9103.
- Zhang, K.Y.J. & Main, P. (1990) Histogram Mathing as a New Density Modification Technique for Phase Refinement and Extension of Protein Molecules. *Acta Cryst.* **A46**, 41-46.

CCP4i as a Project Management Tool

Peter Briggs
CCP4

In previous newsletter articles on the CCP4 graphical user interface (CCP4i), the focus has very much been on using it as a friendlier way of interacting with the CCP4 programs.

This article looks at some of the other options offered by CCP4i, and how they can be used to help in managing structure solution projects.

Introduction

In this context, *project management* refers to managing data about the structure solution process - this data can be broken down into three different types:

1. **Data files** - e.g. reflection data (mtz), sets of coordinates (pdb), maps, log files
2. Information about which **programs** were used to generate the data files and the **order** in which they were run
3. Information about the **parameters** that were used in each program.

Together, 2 & 3 connect together the various data files in 1 and let you see the steps involved in arriving at a particular stage of the structure solution.

Obviously, having access to this kind of information could be useful, for example it can help to:

- Remind you what you did 6 months ago
- Keep some semblance of order if you are working on many projects
- Find out what your student has been doing - like go back three steps to see why the thing is stuck now

Using CCP4i for project management

CCP4i has a number of features and options which can help with project management. Some of the functionality is provided automatically, but be warned - it's not a panacea! Users still need to take responsibility for project management in order for it to work effectively.

With this in mind, the main options available in CCP4i which can help with project management are outlined below. For more details of specific options refer to the [interface documentation](#).

1. Project Directories

CCP4i was designed with the expectation that all data files relating to one crystallographic project will be in one directory. Since each project directory has its own ``Job database'' (see below), splitting your work between different project directories keeps things tidier,

and means that the solution process will be easier to follow for each project. In fact it is *vital* to do this if you are working on more than one project simultaneously.

(Project directories are set up and managed through the **Directories&ProjectDir** window. It is easy to create new project directories and to switch between them; for more information see documentation on [Directories & Project Directory](#).)

2. Job Database

Within each project, CCP4i keeps a database of the tasks that have been run. This "job database" is the core of CCP4i's project management tools.

The database is accessed through the central panel in the main window, which lists the tasks, the date when they were run, and their status (running, finished, failed, killed). Also visible is a one-line "title" comment, which is entered by the user via the **Title** box at the top of each task interface. It is important to fill in the title box for each job, so that different runs of the same task can be distinguished from each other. (The **Edit Job Info** option can be used to add or edit titles, if you forget them initially.)

The database also automatically records the names of input and output files and stores the parameters set by the user. Previously-run jobs are selected by clicking on them with the mouse (at which point they are highlighted in yellow). The user then has a number of options, which can be accessed from the [database menu](#) (see figure 1):

- **View Files From Job** - presents a list of input and output files associated with the job, which then can then be viewed. This also gives options to view the log file and bring up graphs in loggraph.
- **Delete Files** - this cleans up unwanted files, and lets you remove failed or unwanted jobs from the database.
- **Archive Files** - important files can be saved to a "safe" directory
- **ReRun Jobs** - with the option to review and change the parameters used in a previous run
- **Notebook** - a simple on-line notebook allows the user to record in detail any important comments about a particular job, for later reference.
- **Edit Job Info** - useful if you rename a file and want to update the information in the database, or if you want to change (or add) a title.

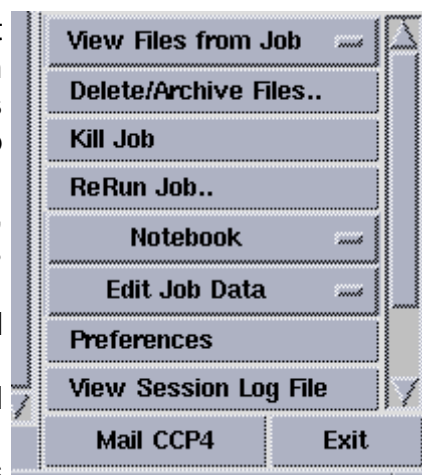


Figure 1: database menu on the right-hand-side of the main interface window.

An example of the **Notebook** facility is given in [figure 2](#); an example of the **Edit Job Info** window can be seen in [figure 3](#).

3. Reporting "external tasks"

Realistically, nobody will use CCP4i for every part of their structure solution - there is a large number of programs outside the CCP4 suite, and of course some programs within CCP4 aren't interfaced yet either. So there will inevitably be "holes" in the list of jobs, and places where CCP4i can't help you recover parameters and so on.

This is a big problem for a project management tool, and to try and deal with it CCP4i provides a **Report external task** option (under the **Edit Job Info** button). This lets the user add the details of a job which was run outside of the interface, namely:

- Title and program name (which appear in the list of jobs in the main window)
- Control file (e.g. script) and log file
- Lists of any input/output files

In this way a record of external tasks can also be added to the database for future reference. Next to splitting projects into separate directories *this is probably the single most important thing to do if you want to have a complete record of the structure solution process.*

4. CCP4i and Data Harvesting

As of version 4.0 of CCP4, **Data Harvesting** has been implemented in a number of CCP4 programs (SCALA, TRUNCATE, MLPHARE, REFMAC and RESTRAIN). Data harvesting automatically captures information about the structure solution process, so that this information can be used at the deposition stage. (see Martyn Winn's [article on harvesting](#) in newsletter 37 for more information). To enable data harvesting as default in CCP4i, go to the **Preferences** window and set the appropriate option.

Summary

If you don't remember anything else, remember to ...

- ... split separate projects up into different directories
- ... use the **Title** box in the task interfaces to distinguish between different runs of the same task
- ... use the **Report External Task** to add a record of runs of un-interfaced programs
- ... use the **Archive** facility to save important data

... with the resulting benefits of ...

- ... having a full record of exactly what you did
 - ... the ability to review log files and data files at click of a button
 - ... the possibility of repeating a job at the click of a button
-

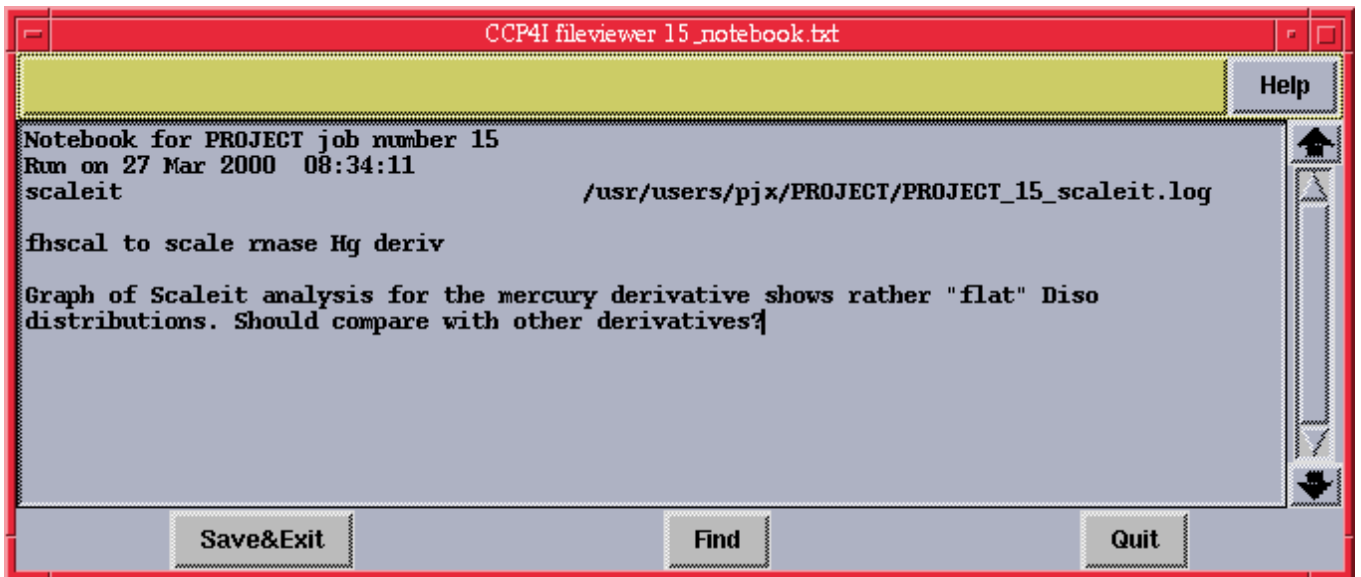


Figure 2: (above) Example of the "notebook" facility. Information about the job (title, date and so on) is automatically inserted at the top of the entry, and the user can add or edit their own comments below, for later reference.

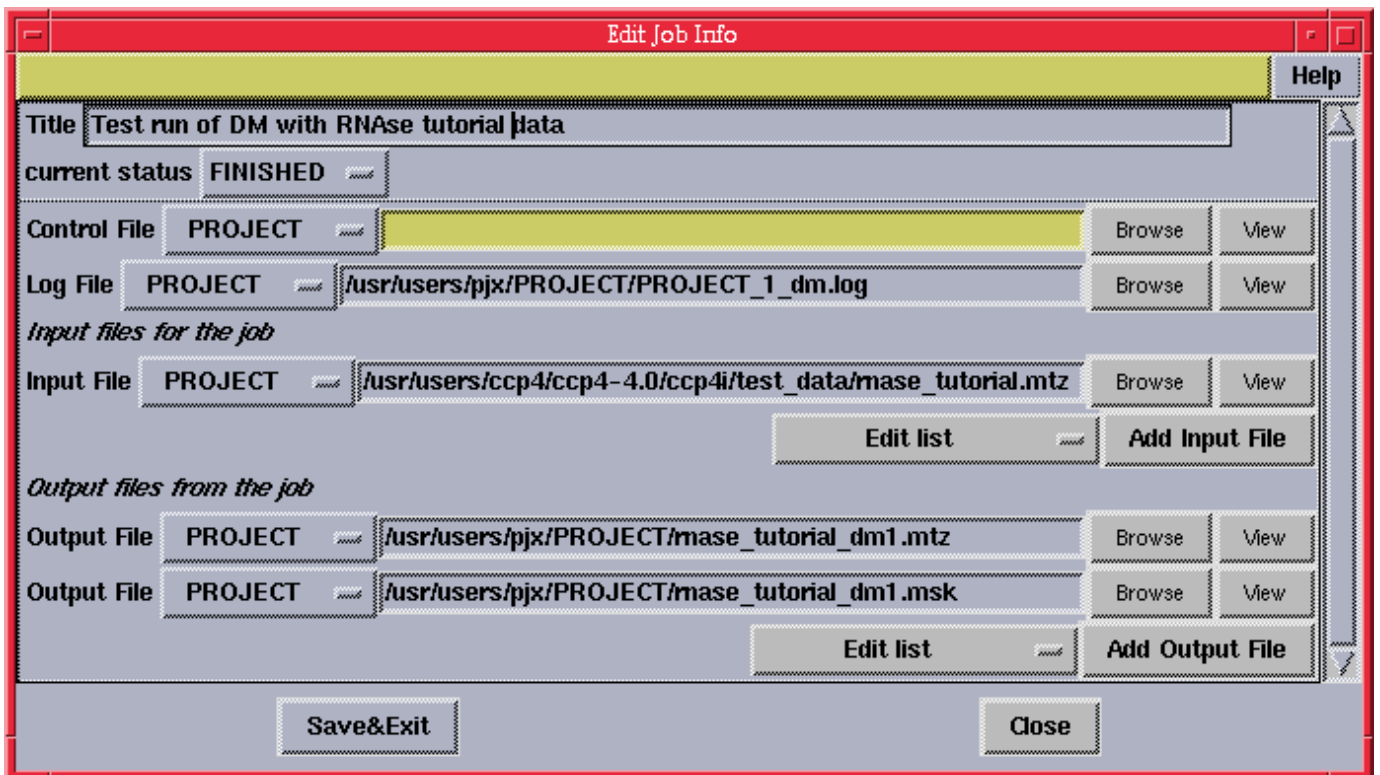


Figure 3: (above) Example of the "Edit Job Info" facility. This lets you view and if necessary change information (e.g. title, input/output files) relating to the job in question.

mmCIF in the CCP4 Suite

Martyn Winn

Daresbury Laboratory,
Daresbury,
Warrington
WA4 4AD, U.K.
m.d.winn@dl.ac.uk

Introduction

The macromolecular Crystallographic Information File (mmCIF) format was developed by a working group of the IUCr formed in 1990. It represents an extension of the CIF format used by small molecule crystallographers, and is the IUCr-recommended medium for electronic transfer of macromolecular crystallographic data. Consequently, mmCIF is likely to be encountered more and more by practising protein crystallographers, and already occurs in a number of contexts in the CCP4 suite.

The aim of this article is to outline the mmCIF resources currently available in CCP4, as well as plans for some future ones. Some of these resources represent extra functionality of the suite, while others are resources for the program developer. While the instances of mmCIF described below vary widely in purpose, they are connected by their use of the mmCIF format, and share the semantics implied by the mmCIF dictionary, and it therefore useful to consider them together.

Full details of the mmCIF format can be found on the mmCIF Home Page or one of its mirrors. Links to the various CCP4 resources are included below.

mmCIF dictionary

mmCIF files are text files with a flexible format based around either <data_name> <data_value> pairs or a loop structure (works like a table). A wide variety of data items are supported, and these are defined in the associated mmCIF dictionary. As well as listing the data items that may be included in an mmCIF file, the dictionary details attributes of each data item, such as type, allowed range, whether or not compulsory, and to which other data items it is related. Data items are grouped into categories.

The standard mmCIF dictionary is maintained by the IUCr. However, the dictionary is designed to be extensible, and local extensions are possible which may then be submitted for inclusion in the main dictionary.

An mmCIF dictionary is distributed with the CCP4 suite as \$CCP4/lib/data/cif_mm.dic, consisting of version 1.0.00 of the mmCIF dictionary together with extensions required for data harvesting. A binary symbol table representation (\$CCP4/lib/data/cif_mmdic.lib) of the dictionary is built during compilation of the suite, and it is in fact this that is used by the libccif library routines.

libccif: the core mmCIF library routines

Peter Keller's C language library of routines for reading and writing mmCIF files was included in release 4.0 of the CCP4 suite. The source files are held in \$CCP4/lib/ccif, and when compiled give the separate archive file libccif.a or the shared library libccif.so. These routines are used by the CCP4 library routines harvlib.f (used in data harvesting) and cciflib.f (see below). For those wishing to use libccif to read/write mmCIF files, there is a Fortran-callable interface, which is described in \$CCP4/doc/ccifdoc.ps.

Data Harvesting

Data Harvesting was introduced in CCP4 4.0, and the technique has been described in Newsletter 37. The implementation of Data Harvesting uses mmCIF files to store information from the programs SCALA, TRUNCATE, MLPHARE, REFMAC and RESTRAIN for future transfer to the deposition site. These files (which should not be edited!) can be found in directory \$HARVESTHOME/DepositFiles where \$HARVESTHOME defaults to the user's home directory.

mmCIF reflection files

MTZ reflection files can be converted to mmCIF format by the CCP4 program MTZ2VARIOUS. The output file may then be used for the deposition of structure factors to the PDB.

Rasmol 2.7

Version 2.7 of the popular molecular viewer Rasmol, which is included in the CCP4 distribution, will display molecules input from CIF or mmCIF format files (other formats are also supported). Some restrictions are imposed, for example the chain identifier (`_atom_site.label_asym_id`) is restricted to one character whereas there is no such restriction in the full mmCIF format. See the program documentation for more details.

mmCIF major mode for Emacs

The editor Emacs can be run in various so-called "major modes" which allow one to set colour schemes, key binding, etc. appropriately for a particular file type. The standard Emacs distribution provides major modes for HTML, Fortran and many others. In CCP4 4.0, a file cif.el is provided which defines a major mode for mmCIF (see the top of the file for how to load it). A simple colour scheme helps viewing of mmCIF files, while a "CIF" menu provides some extra functionality, for example finding the dictionary entry for a particular data item. I hope to extend the functionality of cif.el in future.

cciflib.f: application interface for coordinate handling

It has been proposed that mmCIF should be used by CCP4 programs as a working format for coordinate files, to replace the current use of PDB files. Just as CCP4 currently only use a subset of the full PDB format, so only a subset of the full mmCIF format would be used, nicknamed "ccif". As a step towards this, a set of Fortran routines have been written (which in turn call libccif routines) which provide an application interface for CCP4 programs. This effectively replaces the `rwbrook` routines currently used for PDB files.

These routines were included in CCP4 4.0, see the accompanying documentation. A number of CCP4 programs have been converted to use these routines, and some additional utilities are in development. Details can be found on the developers' web pages.

REFMAC

Version 5.0 of REFMAC (at the time of writing, not yet released) will include several new features, one of which is a completely new mechanism for handling geometric restraints. Restraint information for residues, cofactors, etc. is held in dictionary files `mon_lib_*.cif`, which is designed to be easily extensible to include new chemical species. REFMAC will also be able to read coordinate files in either PDB or mmCIF format.