CCP4 NEWSLETTER ON PROTEIN CRYSTALLOGRAPHY

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

Number 37 **October 1999**

Contents

Editor: Peter Briggs

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

News from CCP4: October 1999

Peter Briggs, Martyn Winn, Sue Bailey, Alun Ashton, Sheila Peters, David Brown

---

1. Staff changes

Since the last newsletter there have been two new additions to the CCP4 staff based at DL.

In March we were joined by **David Brown**, who is the new administrative secretary at Daresbury. David was previously employed at the Daresbury Commercial Office, and has taken over some of the administrative workload for the CCP4 commercial liaison and workshop organisation.

More recently, at the start of July **Sheila Peters** joined us as a new programmer. Sheila was previously lecturing in theoretical and computational chemistry in Tiruchirapalli in India.

2. Workshops and Conferences

Following January's study weekend (see the report in the last newsletter), CCP4 has been involved in a variety of ways with workshops and conferences.

May saw the **5th European Workshop on Crystallography of Biological Macromolecules**, in the impressive surroundings of Villa Olmo on the shores of Lake Como, Italy. The workshop was organised by Phil Evans and Menico Rizzi (who managed to pack a lot of material into the four days!) and recieved financial support from CCP4 and a number of industrial sponsors. Also in May, Liz Potterton could be found demoing the CCP4 graphical user interface at the **ACA '99** meeting in Buffalo, NY.

In August CCP4 also had a strong presence at the **18th Congress of the IUCr**, Glasgow. Prior to the congress a small workshop was held which aimed to introduce people to the suite. The workshop was attended by around 60 people, mainly students and young postdocs. The morning was given over to general overview talks, and in the afternoon the meeting broke into smaller groups with more informal question-and-answer sessions based on particular topics. The workshop was organised by Martyn Winn and David Brown, and we would like to thank all those who gave up their time to participate - particularly the speakers (Martyn, Phil Evans, Eleanor Dodson, Garib Murshudov, Harry Powell). Serena Cooper has been kind enough to write a short summary of the workshop for this newsletter.

During the congress proper CCP4 provided sponsorship of a microsymposium on ``Problematics in Macromolecular Structures'', and also maintained a stand inside the commercial exhibition (manned at various times by varying numbers of people!). We would like to thank all those who visited the stand and we were heartened by the many positive comments that we recieved. As well as the core DL programmers (Martyn Winn, Alun Ashton, Peter Briggs) thanks must also go to Harry Powell, Garib Murshudov, Maria Turkenburg, Kevin Cowtan and Liz Potterton for doing their part (and taking the pressure off us!). CCP4 would also like to thank the groups who allowed us to use their structures in our display.

At the end of August CCP4 provided sponsorship of the annual **GALA meeting** held in Galashiels (Scotland), the **CCP4/BCA Summer School in Protein Crystallography** at St. Andrews University, and for the **International School of Crystallography in Biological Macromolecules** held in Barcelona in early September.

Finally looking ahead, preparations are continuing for next January's millenial CCP4 Study Weekend, on the subject of **"Low Resolution Phasing"**. (Pierre Rizkallah *et al* have contributed a "taster" article on low resolution phasing in this newsletter). For more information about the Study Weekend, see below.

3. Other Developments

3.1 Progress with ccp4i

The official release of version 1.0 of ccp4i, the CCP4 graphical interface, was in April of this year (followed shortly after by a patch release 1.0.1). Since then the interface has been demoed at a variety of meetings including Como, the ACA and the recent IUCr by Liz Potterton and Peter Briggs. Liz and Peter have also visited a number of academic venues to install and demonstrate the interface, and the response so far has been extremely positive. It is planned that as of the next release of the suite (see below), ccp4i will be part of the full CCP4 distribution.

For more information about the current status of the interface and future plans for it, see Liz Potterton's article.

3.2 Port to NT

CCP4 is now dedicated to providing a version of the suite running under Windows NT. Support for an NT version of the suite, which is expected to run exclusively through ccp4i on that platform, doesn't mean that we are withdrawing support for other platforms - so users shouldn't expect any change from their normal service... See Alun Ashton's article about these issues elsewhere in this newsletter.

3.3 News of CCP4 Release 3.6

The next version of the CCP4 suite (version 3.6) is planned sometime before the end of the year. This next release will contain a number of new programs as well as many developments and bug fixes for old favourites, and will also include the latest version of the graphical user interface. Several programs in this release have also been updated to allow the operation of Data Harvesting; see Martyn Winn's article in this issue for more information.

As always, once it is ready the new release will be advertised via the bulletin boards and newsgroups.

3.4 CCP4 Study Weekend 2000: Low Resolution Phasing

A meeting organised by the
Collaborative Computational Project in Macromolecular Crystallography (CCP4)
and Daresbury Laboratory
7-8th January 2000
The University of York

This study weekend will incorporate X-ray crystallography and electron microscopy in low resolution phasing and the two approaches will be shown to be complementary.

In X-ray crystallography there are many proteins for which phases cannot be obtained experimentally by using heavy atom derivatives, anomalous data or for which there is no suitable model for molecular replacement. We will discuss vario us techniques to obtain low-resolution phases including *ab initio* methods, using single crystal X-ray data, and experimental methods such as solvent contrast variation. The use of maximum entropy and other methods to extend these phases to higher r esolution will also be described. Structure determ-ination by electron microscopy from both 2D crystals and single particles will be covered. The combination of the two techniques will be demonstrated in both phasing of X-ray crystal structures by electro n microscopy and fitting of X-ray structures into electron microscopy maps.

The meeting will provide an introduction to the basic concepts and an opportunity to discover the most recent advances.

Invited speakers include:

D. Stuart (Oxford) M. Rossman (Purdue, USA)
P. Main (York) V. Lunin (I.M.P.B., Pushchino, Russia)
A. Urzhumtsev (Nancy, France) A. Podjarny (I.G.B.M.C, France)
M. Van Heel (Imperial, London) J. Miao (New York, USA)
E. Hewat (Grenoble, France) N. Grigorieff (Brandeis, USA)

How to apply:

For a web registration form and more details visit the CCP4 WWW home page at http://www.ccp4.ac.uk and follow the courses link.

Scientific Organisers: J.Wilson (York), H.Saibil (Birkbeck), J.Grimes (Oxford)

Organisers: Alun Ashton, David Brown, Pat Broadhurst

Further enquiries can be made to ccp4@dl.ac.uk (Tel: +44 (0) 1925 603528; Fax: +44 (0) 1925 603124; or to David Brown, Daresbury Laboratory, Daresbury, Warrington, WA4 4AD, U.K.).

The closing date for application is 12th November 1999. Applications **must** be received by this date

CCD Detector Installed on Multi-wavelength Station 9.5 at the SRS.

James Nicholson
Daresbury Laboratory, Warrington, WA4 4AD, UK (J.Nicholson@dl.ac.uk)

A 165 mm diameter MAR CCD detector was installed on station 9.5 (tunable wavelength) of the Synchrotron Radiation Source at Daresbury Laboratory during August 1999. The new detector replaces the 300 mm MAR image plate. The first users of the CCD were on 28th August 1999 and each group has benefited from the faster readout; six seconds compared with four minutes for the image plate. With average exposure times between three and five minutes on 9.5, the new detector enables collection of approximately twice as much data in the same time as previously. This means great improvements in the quantity and quality of data collected on the station, which is vital especially for Multi-wavelength Anomalous Dispersion (MAD) experiments.

Figure 1. The MAR CCD detector mounted on the optical bench of station 9.5 at the SRS.

As shown in Figure 1, the CCD detector overhangs the translation base by approximately 15 cm. This allows a much smaller minimum crystal-to-detector distance of 81 mm, compared with 230 mm previously. Consequently higher resolution data is now obtainable on station 9.5 (as outlined in Table 1).

Wavelength (Å)	***Achievable Resolution with the MAR CCD (Å)***	***Achievable Resolution with the MAR IP (Å)***
0.8	1.05	1.40
1.0	1.31	1.75
1.2	1.58	2.11
1.4	1.84	2.46

Table 1. Achievable resolution on 9.5 with the 165 mm CCD, compared with the 300 mm image plate detector. Note that the flux at a wavelength of 1.0 Å is approximately three times that at 0.8 Å and the flux at 1.4 Å is approximately five times that at 0.8 Å.

For more information about the CCD detector and station 9.5 in general visit URL:

http://www.dl.ac.uk/SRS/PX/9_5_manual/man.html

TCL/TK BASED PROGRAMS : CRYSTALLOGRAPHIC CALCULATOR

by

L.M. Urzhumtseva & A.G. Urzhumtsev

LCM^{3}B, UPRESA 7036 CNRS, B.P.239, Faculté des Sciences, Université Nancy I, 54506, Vandoeuvre-lès-Nancy, France. E-mail: sacha@lcm3b.u-nancy.fr

Structural crystallography needs a number of simple algebraic calculations in order to measure lengths, angles, define relations between molecules, transformations of the coordinates, etc. These calculation can be done by general means, e.g. with a usual calculator or with a computer using standard mathematical complexes. Another possibility is to have a specialised tool, tuned particularly to these calculations, easy to call and to use, fast to address.

A new program in the suite of Tcl/tk based programs (Urzhumtseva & Urzhumtsev, 1996-1999), named CALCRYS, has been developed. This program allows to work with vectors defined in real or in reciprocal three-dimensional space, expressed in Cartesian or crystallographic (fractional) co-ordinates. Unit cell can be defined also either in real or in reciprocal space. When the cell is defined, its parameters in the conjugate space and all metrical tensors are calculated automatically. The optional choice of the orthogonalisation agreement allows to establish automatically the orthogonalisation and deorthogonalisation matrices. This crystallographic information presents permanently at the screen (Fig. 1).

The list of available algebraic operations includes the calculation of a linear combination of vectors, of their scalar and vector product (taking into account the corresponding metrical tensors), product of a matrix by a vector or by another matrix, a linear combination of matrices, matrix inversion, matrix determinant etc. These operations are grouped in a menu displayed near the set of work windows containing the data with which these operations are executed.

One more field contains a number of "memory cell" which allow to save intermediate results (any mixture of numbers, vectors or matrices) and to use them later.

Information can be inserted into the working windows in several ways : copied from the crystallographic field, from the memory cells, directly typed or read from a file.

CALCRYS is programmed completely in Tcl/tl (Ousterhout, 1993) which gives a possibility to run the same script directly under Windows in PC, in SGI or in AlphaDEC computer.

The program can be used by researchers as well as by students and teachers of crystallography. It is available by request from the authors. Reported "bugs" are welcome.

The authors thank P. Allé and O. Louis for technical assistance and C. Lecomte for the interest to the work.

References.

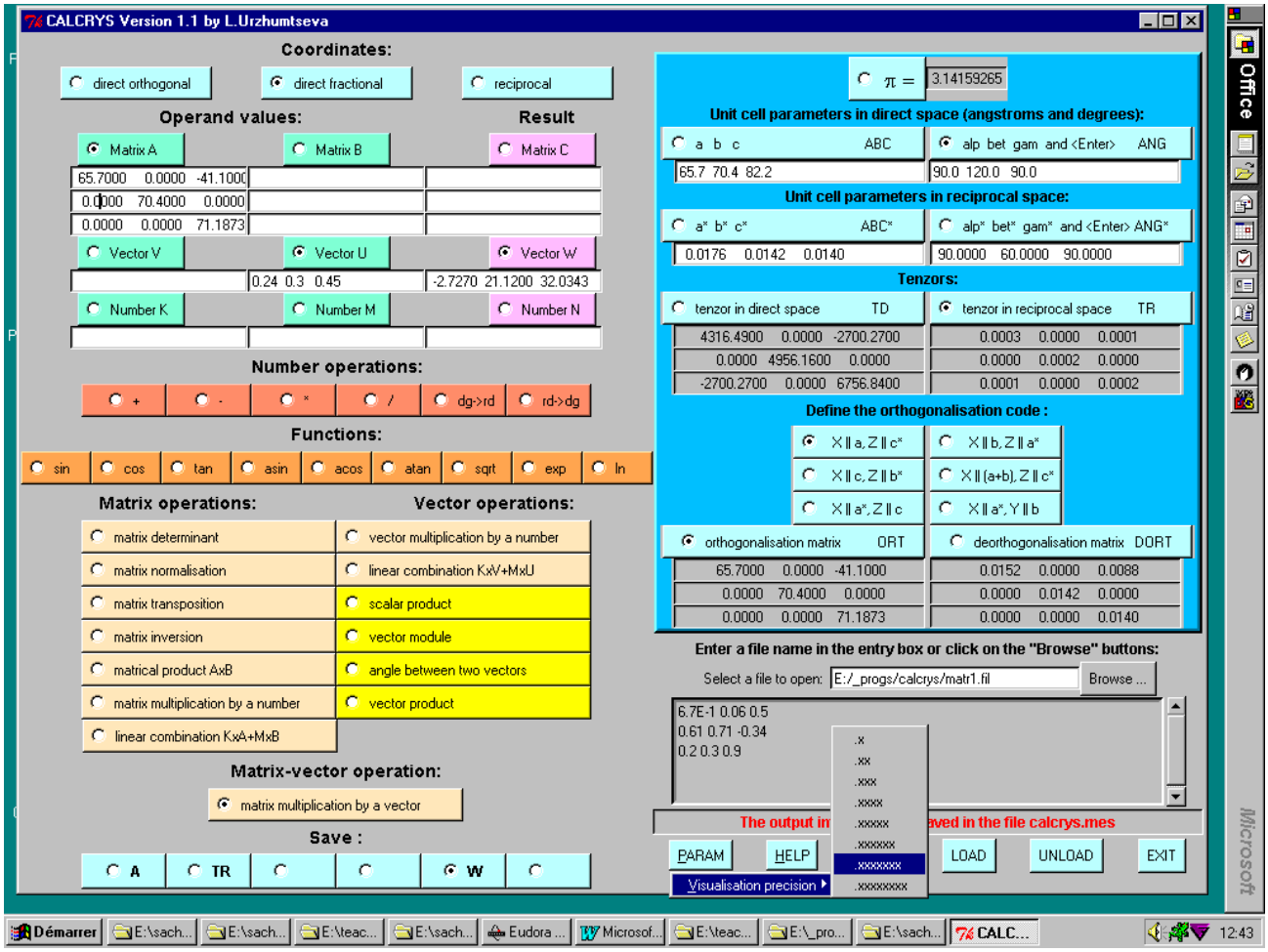
Ousterhout, J.K. (1993) *"Tcl and the Tk Toolkit".* Addison-Wesley Publishing Company.

Urzhumtseva, L.M., Urzhumtsev, A.G. (1996) "Tcl/Tk-based programs. I. CONFOR : user-friendly converter for crystallographic data files". *CCP4 Newsletter on Protein Crystallography,* **32b**, 41-43

Urzhumtseva, L.M., Urzhumtsev, A.G. (1997) "Tcl/Tk based programs. II. CONVROT: program to recalculate different rotation descriptions". *J.Appl. Cryst.,* **30**, 402-410

Urzhumtseva, L.M., Urzhumtsev, A.G. (1998) " Programs Tcl/Tk based crystallographic software : current state and new programs Tcl/tk interface ". *CCP4 Newsletter on Protein Crystallography,* **35**, 22-24

Urzhumtseva, L.M., Urzhumtsev, A.G. (1999) "Tcl/Tk based programs. III. CRITXPL: graphical analysis of the X-PLOR refinement log-files". *J.Appl. Cryst.,* **32**, 376-377



CCP4 Bulletin Boards: a short FAQ

Peter Briggs, CCP4

Amongst other services provided by CCP4 we currently maintain two e-mail discussion lists (bulletin boards), ccp4bb and ccp4-dev. This article is intended to answer general questions about the purpose and use of these facilities.

What is the purpose of ccp4bb@dl.ac.uk?

The main bulletin board ccp4bb is intended to host discussions about topics of general interest to macromolecular crystallographers, and provides a lively forum for discussions on a wide range of crystallographic topics. At the time of writing there are over 1,200 subscribed e-mail addresses, so it is a useful way to reach crystallographers all over the world.

Any crystallographic-related item is acceptable, and doesn't have to be directly related to CCP4. In the past questions have prompted (and provoked!) discussions covering all aspects of macromolecular crystallography, from experimental methods and strategy to advice on data processing, refinement and structure analysis.

We are always keen for people to post summaries of advice and information received from the discussions which they have initiated or participated in. This is especially useful since many of the responses do not come to the bulletin board, but instead directly to the person asking the question. This way everyone can benefit from your efforts!

The bulletin board can also be used to request information, and to inform people about job vacancies, new services and the availabilty of new or updated software.

What is ccp4-dev@dl.ac.uk?

ccp4-dev is the `developers bulletin board', and is aimed exclusively at program developers. Generally it is used for informing interested programmers about significant code changes in the developmental versions of CCP4 programs and libraries, and discussions related to this.

ccp4-dev is more specialised than ccp4bb, and would be pretty boring for most crystallographers without a direct interest in the programming issues.

What material shouldn't I send to the bulletin boards?

Inappropriate content for ccp4bb includes personal messages and abuse, and messages of an unrelated commercial nature. Basically any message that you send should be relevant to as many people reading the bb as possible - use your common sense! Equally, postings to ccp4-dev should be restricted to information or questions about software development.

Also, reports of program bugs (and fixes!) or specific questions about the CCP4 suite etc should be sent directly to the Daresbury staff at ccp4@dl.ac.uk, where we will do our utmost to answer promptly.

How do I subscribe to the bulletin boards?

Go to the web page at http://www.ccp4.ac.uk/ccp4bb.php, which has an automated facility for subscribing and unsubscribing to ccp4bb. Program developers should visit the CCP4 developers area at http://www.ccp4.ac.uk/dev/main.php for details of ccp4-dev. If you encounter any problems then e-mail us directly, describing your difficulty.

Once you have subscribed you will recieve messages directly from the bulletin board as they are posted. Being subscribed also allows you to post messages by mailing to the appropriate address (ccp4bb@dl.ac.uk for general messages, ccp4-dev@dl.ac.uk for messages about program development).

What steps are taken to stop junk mail ("spam")?

Different people have different definitions of what precisely constitutes junk mail, but unacceptable content here includes offensive and/or unrelated messages (particularly of the kind that has recently been filling up many of the crystallographic newsgroups).

Both the lists are moderated, which in this case means that only people subscribed to the list are able to post to it. This is done automatically by software which checks the sender's email address against the list of subscribed addresses, and suceeds in stopping practically all junk mail reaching the end users.

Sometimes this also prevents ``legitimate'' messages reaching the bb - the most common reason for this is posting from a different address to the one with which you are subscribed - so please check! We apologise for any inconvenience this might cause individuals, but believe that it is a small price to pay for keeping the bulletin board free of junk.

As regular subscribers will know, a very small amount of junk mail still gets through. This is because it has been posted by people who took the time to subscribe first. However this is minimal, and in such cases we take steps to remove the offending subscribers from the list at the earliest opportunity.

How can I get old messages from the bulletin board?

The automatic software which administers the bulletin boards also offers a facility to retrieve the old messages which have been posted, and information on how to do this can be found on the web at http://www.ccp4.ac.uk/ccp4bb.php.

The page also carries a link to a web-based archive maintained at Birbeck College London. Please note that this archive (and others like it) are not maintained by CCP4, and we make no promises of their reliability or availability.

(Martyn Winn has also summarised some of the recent discussions from the bb in his article elsewhere in this issue.)

What if I have problems with subscribing/unsubscribing/posting etc. to the bulletin boards?

Send an e-mail to ccp4@dl.ac.uk describing the problem and we will do our best to fix it.

September 1999

News on CCP4I

Liz Potterton August 1999

The first official release of the CCP4 graphical user interface, CCP4I 1.0, was announced in April this year. Since then something like 180 sites have downloaded the software from the York FTP site. It's hard to know how many of the 180 sites are using the Interface for real - I've received feedback from 20-30 sites - some of it even complimentary! So are all the others using it happily without problems or have they not really got into it yet? Please let me know!

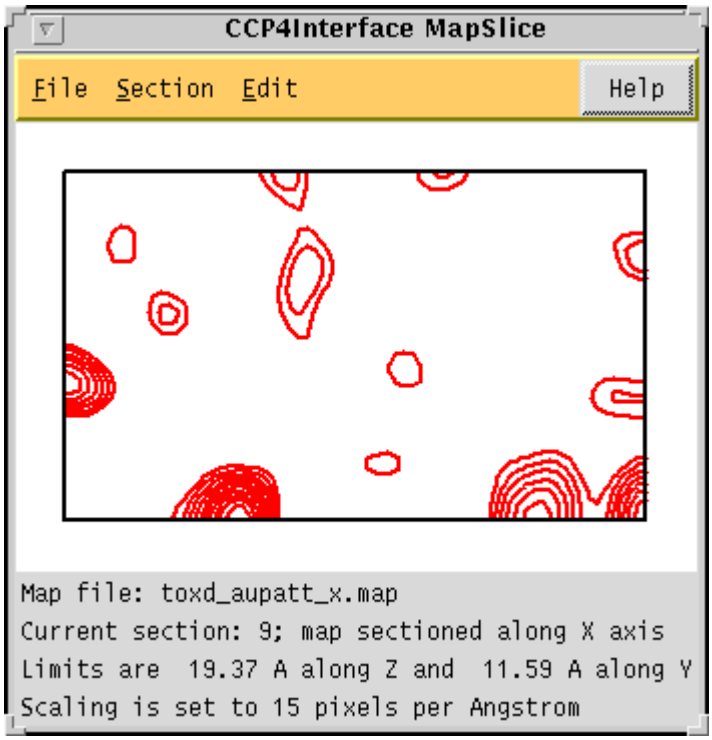
Peter Briggs and I have visited several crystallography labs in the UK to do demos and run mini-workshops to help people get started using the Interface and we would be very happy to do more so let us know if you would like us to visit you for a day. We've also done demos at conferences: the ACA in Buffalo, IUCr in Glasgow and the European Crystallography Workshop in Como.
The interface has been very useful for workshops where participants may be unfamiliar with CCP4 programs - we used it for the Refinement Workshop in York in December 1998 (prior to full release of the software) and at a recent International School in Barcelona.

Pete and I are now working on the enhancements to the Interface. Projects we are currently working on:

Interfaces to programs that got missed first time around - Sfall, Overlapmap, Detwin, Shelx and interfaces to some of the new programs entering the suite.

Enhancements to the **Loggraph** program particularly allowing the user to edit the graphs before printing them out for presentations or publications.

A **Mapviewer** which should replace NPO and xplot84driver for visualising map sections.

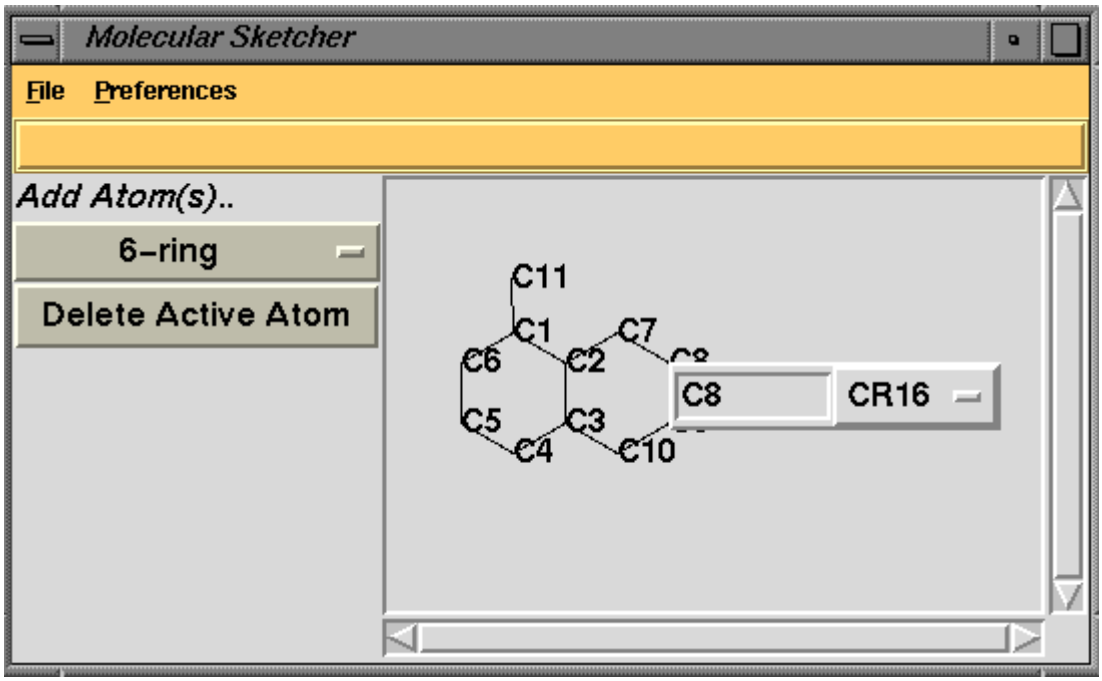


The "mapslice" utility is still under development and at present still lacks much of even the basic functionality required. Ultimately the aim is to include the following:

- View arbitrary sections along any axis
- Specify contour levels with either absolute or rms values
- Automatic selection of Harker sections
- Display of PDB files (either as peaks or as atoms)
- Printing sections either to a printer or file.

Comments or suggestions able other functions would also be useful at this early stage.

A **Molecular Sketcher** to simplify defining novel ligand molecules which need new entries in the refinement dictionary. This will go with the improvements to handling geometric constraints currently being worked on by Garib Mushudov and Alexei Vagin. An exampe of the Sketch window is shown below - the user has 'sketched' the structure and has clicked on the atom C8 to open a small window which enables editing of atom type and atom name. When the user has defined the molecule they will be able to automatically add the structure to their dictionary.



Porting the Interface to Windows NT. The basic Interface is now working on NT thanks largely to the Tcl/Tk language porting painlessly. There are a couple of more serious issues.

One problem I find working with the computer network setup at York is that the same file has different path names seen from Unix or NT machines (and you can get the same problem with an all-Unix system). I expect this is going to be a general problem as people will work in a 'mixed' environment. The interface already uses the idea of project directories and directory aliases and these need to be made a bit more flexible to cope with different machine domains.

The other possible issue with NT is that it does not support the rsh command (or anything similar) which the Interface uses to run remote jobs on Unix systems. The solution to this may be to run a 'CCP4I server' on the 'remote' machine. Or will NT users expect to always run jobs on their own machine? I would be interested to hear from anyone who has any ideas on what would be useful or possible solutions.

See the Installation Guide if you want to download the interface.

Release of MOSFLM 6.01

Harry Powell, MRC-LMB, Cambridge
The latest version of MOSFLM (version 6.01) was released in July. It contains a number of new features and numerous bug fixes.

The MOSFLM User Guide has now been made available in HTML format; a problem page and short FAQ are kept in my personal area on the LMB's web server. Errors and typos in these documents will be corrected as and when they become apparent.

The MOSFLM on-line help, which was previously only available while running the program, can now be accessed via a PERL script; comments would be received gratefully!

New features

1. Circle fitting; a set of points picked on e.g. an ice ring or a wax ring can be used to determine the direct beam position.
2. the DPS Autoindexing has been improved by inclusion of a least-squares fit of the predicted to found spot positions. This greatly improves the predictions. A by-product of this routine is that the central beam position can be refined as well.
3. The background measurement for the spot picking routine has been modified so that the measurement box is offset slightly for the tiled ADSC 2x2 CCD detector, to avoid the zero count region.
4. A user editable spot list has been implemented. This allows the user to both add and delete spots from the indexing list.
5. An option to write multiple MTZ files has been added. This is particularly useful if processing while collecting data.
6. The Mar 165mm CCD detector has been added to the list of supported formats.
7. The penalties in the DPS autoindexing have been normalized to a maximum of 999.

Many small bugs which have been pointed out by users have been fixed.
Several users have had trouble running MOSFLM with modern PCs running RedHat Linux 6.0; these problems are almost all due to the incompatibility of the xdl_view libraries and the 24 - and 32- bpp colour graphics provided by the X-server distributed with RedHat Linux 6.0 (the same problem arises with other programs which use the libraries such as IPDISP). The fix is to reduce the colour depth on the PC to no more than 16 bpp.

The program is available via anonymous ftp from ftp.mrc-lmb.cam.ac.uk/pub/mosflm. Executables have been built on the following UNIX platforms:

```
Compaq Alphas:  Compaq Tru64 UNIX 4.0F
SGIs:           Irix 5.3, 6.2, 6.4, 6.5
Intel PC:       RedHat Linux 5.1 (should also run on 5.2,5.2 and 6.0)
PPC Macintosh:  LinuxPPC 1999 (will not run on Release 4)
```

Are You Sure You Know It All?

Summary of the IUCr CCP4 workshop

Serena Cooper, Manchester University

I was persuaded to attend the recent CCP4 workshop at the IUCr meeting in Glasgow. As a CCP4 user I was sceptical that I would learn anything new, however I was pleasantly surprised, or perhaps it was a reflection on how little I already knew!

The workshop began with a general overview and update of CCP4 by Martyn Winn. Although most of it was not new to me there were a few things I was unaware of, for example, map files could be dumped by mapdump or that you can now assign project and dataset names to different datasets within MTZ files. Phil Evans talked about SCALA and the best route for data reduction from various image processing packages, and the optimum ways to scale your data. He also gave a list of what to check in the output, in particular he stressed the importance of the normal probability plots. These can be obtained by using the keywords NORMPLOT and ANOMPLOT and viewed in Xmgr, although there was some debate to whether it would be more appropriate to include them in loggraph. Eleanor Dodson talked through using the GUI (despite a few technical problems) and certainly tempted me to try using it.

In the morning we were polled on what people wanted to have more specialist information on. Five discussion groups were set up for the afternoon session, these being for Refmac, MIR strategies, Mosflm, SCALA and the GUI. There was only time to attend two of these groups so I can't comment on the MIR, GUI or SCALA sessions.

Garib Murshudov gave examples of how to run Refmac, in particular using NCS. He also stressed the importance of including hydrogens in anisotropic calculations. He suggested a good way to get initial anisotropic values for co-factors or metal centres is to use the aniso keyword, which will calculate aniso cards for all atoms if there aren't any and then you can extract the relevant atoms. What most people wanted to know was what to check in the log files, and he gave a list of what he looks at: R; R*free;* σA estimations; Bσ's ; rms deviations in bonds, angles torsion angles and chiral volumes, the latter two being good indicators of where the model is poorly fitted.

Harry Powell was answering questions about Mosflm. It quickly became apparent that there were a lot of people who had either never used Mosflm or not within the last few years and were unaware of the recent improvements (and that there was an alternative to Denzo). He therefore gave an outline of how to run it and what to do if you have weak images, advising that the program works best using 200-500 spots. The session then moved on to general questions and there was a bit of a Denzo versus Mosflm discussion. From the tests Harry has been doing Mosflm is comparable speed wise and is no less capable at indexing. Mosflm has the added advantages of being free, with most detector types easily available and the support is helpful.

There were also opportunities throughout the day to have discussions with the programmers to sort out more individual queries. At the end of the workshop I felt that I had learnt something, been reassured that what I was doing was correct and that the programmers were approachable and always willing to help.

S.J.Cooper Tue Sep 21 13:40:45 BST 1999

Maximal Likelihood refinement. It works, but why? (Seminar notes).

by Vladimir Y.Lunin[a,b] & Alexander G.Urzhumtsev[a]

([a]LCM^{3}B, University of Nancy, France, e-mail: sacha@lcm3b.u-nancy.fr;

[b]IMPB RAS, Pushchino, Russia, e-mail: lunin@impb.psn.ru)

Recently the Crystallographic Laboratory (LCM^{3}B) of the University of Nancy, France, has organized a seminar on crystallographic methodology. One of the purposes of this seminar is to analyze some basic crystallographic approaches which are not always clear presented in the current literature. Such clarification becomes more and more important with increasing the number of scientists who does not have enough background intheoretical crystallography but rather uses it as a "black box" tools for their researches.

Last years (Bricogne & Irwin, 1996; Pannu & Read, 1996; Read, 1997; Murshudov et al., 1997; Pannu et al., 1998) the so called maximal likelihood refinement (MLR) had been proven as a useful tool, which significantly extends the possibilities of refinement. Nevertheless, the reasons for this success are not well explained in the literature. Another general problem is that in probabilistic or in statistical methods the authors of the papers quite rare clearly explain *which* are the random variables, *which* probability they are talking about and *which* statistical hypothesis is testing. An essay to get a more materialistic background for the MLR was a goal of the one of methodological seminars in Nancy. This paper contains a brief notes from this seminar.

1. Conventional refinement.

In order to have the situation maximally transparent, we do not consider below the problem in its most general form, but study simplified cases which still hold the main features of the problem. In particular, we suppose that the atomic coordinates only must be defined, while atomic scattering factors and temperature movement parameters are known precisely.

The goal of the conventional structure refinement may be formulated as follows:

Conventional refinement

The goal of the refinement is to find the atomic coordinates resulting in structure factors magnitudes which are as close as possible to the observed magnitudes.

When we say that atomic coordinates result in some structure factors (s.f. in what follows) we mean that there exists a s.f.-formula (*Appendix A1*) which allows to calculate s.f. provided the coordinates are known. Mathematically the conventional refinement goal may be formulated as a minimization problem, *e.g.* as follows

$$\sum_{\mathbf{h}} w(h)\left[F^{calc}\left(\mathbf{h};\{\mathbf{r}_j\}\right)-F^{obs}(\mathbf{h})\right]^2 \Rightarrow \text{ minimum} \quad , (1.1)$$

where the calculated magnitudes depend on atomic coordinates by means of (Appendix A1, (A1.1)) and $w(\mathbf{h})$ are some weights. Eventually, other criteria can be used as a measure of discrepancy between two data sets instead of the least-square criterion (1.1). The conventional R-factor or R-free factor may serve as a measure of the refinement success.

This goal seems to be quite reasonable when the model is complete and the s.f.-formula is precise, *i.e.* when the structure factors magnitudes calculated with the use of the full set of exact atomic coordinates are equal to the corresponding observed values. The goal is not so evident in some other cases when, for example, the observed magnitudes contain experimental errors or the model is incomplete (*e.g.*, solvent atoms are not included into the model). In such situations the s.f. magnitudes calculated with the exact coordinates of atoms are not equal, in general, to the observed magnitudes. So, the conventional refinement fits the calculated s.f. magnitudes to wrong values.

To reveal the differences in these two kinds of errors we consider them separately. The discussion how they may be combined may be found in (Pannu & Read, 1996).

2. Likelihood based refinement. I. Experimental errors.

Obviously, to take into account the experimental errors it is necessary to have some information about them. Sometimes this information may be introduced as a probability distribution for the errors values. For example, we can suppose that the errors e(**h**) in the magnitudes $F^{obs}(\mathbf{h})$ are independent random variables distributed in accordance with the Gaussian law with zero mean and the known standard deviation s(**h**):

$$P(\varepsilon) = \frac{1}{\sqrt{2\pi}\sigma(\mathbf{h})} \exp\left[-\frac{\varepsilon^2}{2\sigma^2(\mathbf{h})} \right]. \qquad (2.1)$$

The parameters s(**h**) characterize the instrument precision and are external for the refinement information. We will not discuss here how these values may be estimated in practice.

Supposing the values s(**h**) are known, we can estimate for every trial set of atomic coordinates $\{\mathbf{r}_j\}$ how high would be chances to obtain again in the experiment the same set of F^{obs}(h) values, would the quantities obtained in experiment differ from $F^{calc}(\mathbf{h};\{\mathbf{r}_j\})$ by errors e(**h**) only. In other words, we can calculate the probability that the measured magnitudes are just "the calculated ones plus random errors, distributed as (2.1)":

$$P\left\{ F^{calc}\left(\mathbf{h},\left\{\mathbf{r}_j\right\}\right) + \varepsilon(\mathbf{h}) = F^{obs}(\mathbf{h}) \right\} =$$

$$\prod_{\mathbf{h}\in S} \frac{1}{\sqrt{2\pi}\sigma(\mathbf{h})} \exp\left[-\frac{\left(F^{calc}\left(\mathbf{h};\left\{\mathbf{r}_j\right\}\right) - F^{obs}(\mathbf{h}) \right)^2}{2\sigma^2(\mathbf{h})} \right], \qquad (2.2)$$

where the product is calculated over all experimentally obtained magnitudes.

For different trial sets of atomic coordinates this probability is different. If it is too small, it is reasonable to reject the considered coordinates, as the differences between calculated and observed magnitudes fall outside our assumptions about possible errors. As a generalization of this idea, it is reasonable to consider as the best coordinates those which maximize this probability.

ML-refinement. I.

The goal of the refinement is to find the atomic coordinates $\{\mathbf{r}_j\}$ which maximize the probability to reproduce in the X-ray experiment the set of experimental values $\{F^{obs}(\mathbf{h})\}$ provided the experimentally obtained magnitudes differ from the calculated ones by the errors distributed as (2.1).

It must be emphasized that this idea is just a "common sense" rule. It can not be "proven" formally.

The suggested MLR-principle is nothing, but the maximal likelihood principle which is broadly used in the Mathematical Statistics. In the considered case the probability (2.2) is called as the likelihood corresponding to statistical hypothesis:

Hypothesis $H(\mathbf{r}_1,\ldots,\mathbf{r}_{Nfull})$

The experimentally determined $F^{obs}(\mathbf{h})$ values differ from $F^{calc}(\mathbf{h};\mathbf{r}_1,\ldots,\mathbf{r}_{Nfull})$ by random experimental errors $e(\mathbf{h})$. These errors are independent and distributed in accordance with the Gaussian law with zero mean and known standard deviations $s(\mathbf{h})$.

The maximization of (2.2) is equivalent to the minimization:

$$\sum_{\mathbf{h}} \frac{1}{\sigma^2(\mathbf{h})}\left[F^{calc}\left(\mathbf{h};\{\mathbf{r}_j\}\right) - F^{obs}(\mathbf{h})\right]^2 \Rightarrow \text{minimum} \quad . \quad (2.3)$$

We see that in the considered case the MLR refinement is formally equivalent to the conventional refinement (1.1) with appropriately chosen weights $w(\mathbf{h})=1/s^2(\mathbf{h})$. So the conventional refinement may be considered as a type of ML-refinement. Nevertheless a more sophisticated probabilistic modeling of the differences between calculated and observed magnitudes may result in penalty functions different from (2.3) .

3. Likelihood based refinement. II. Incomplete model.

We consider now another case when $F^{obs}(\mathbf{h})$ values are supposed to be measured precisely (or the errors may be considered as negligible ones), but the model is incomplete. We refer to this model as a partial one.

For the clarity we write the formulae below for nc/s reflections. The necessary corrections for centric reflections are straightforward.

The s.f. magnitudes calculated with the exact atomic coordinates of a partial model are not equal, in general, to the observed magnitudes, but differ from them in unknown quantities corresponding to the absent atoms. The problem could be overcome could we have a possibility to get from an experiment s.f. magnitudes corresponding to the partial structure

and adjust to them the values calculated with the partial model. As this is not possible in general, an alternative way could be to introduce some corrections into the measured magnitudes values compensating the absence of a part of atoms in the model. We will show below that ML refinement may be considered from this point of view too.

While the known true atomic position for a part of structure do not allow to reproduce the structure factor magnitudes $F_{full}(\mathbf{h})$ correctly, their are some chances to get these values if the absent atoms are added to the model with randomly chosen coordinates (*Appendix A2*). One can expect that it would be less chances to reproduce $F_{full}(\mathbf{h})$ correctly when the randomly chosen atomic positions are added to a wrong partial model than to the exact one. Therefore, it seems reasonable to consider the partial model coordinates which provide maximal chances to improve the model when generating the coordinates for absent atoms randomly as the best ones. Again, this idea is nothing but the maximal likelihood principle, which now is applied to a different type (in comparison with Sec.2) of statistical hypotheses.

We define now the goal of the refinement as the following one.

ML-refinement. II.

We look for the set of partial model coordinates which maximizes the chances to make calculated magnitudes equal to the observed ones when completing the partial model by N_{add} additional randomly placed atoms.

In a more formal way the goal may be formulated as the following: under the hypothesis that the experimental magnitudes may be reproduced by means of adding randomly of N_{add} atoms to the partial model, maximize the likelihood (*Appendix A3*) varying the atomic coordinates of the partial model.

It is worthy of noting that as many other statistical approaches, the maximal likelihood principle is just a "common sense" principle. It can not be "proven", and all its "good properties" reveal themselves "in mean", when it is applied regularly. In other words, the maximal likelihood principle works statistically and does not guaranty the correct choice when being applied to a single particular object.

It must be emphasized that the ML-refinement is not just a new penalty function. It changes the goal of refinement. We do not try any longer to fit the calculated magnitudes to the observed ones, but try to maximize chances for the further improvement of the model. As a consequence, the conventional R-factors (as well as R-free) may, in general, increase their values in the course of the likelihood maximization.

4. How to calculated the likelihood value.

In order to realize the goal of the ML-refinement we must have a possibility to calculate the likelihood value provided coordinates of the partial model are known. The usual way includes the following steps:

a) derive the joint probability distribution (j.p.d.) for magnitudes and phases of the calculated structure factors;

b) derive the marginal probability distribution for the calculated s.f. magnitudes by means of the integrating of the j.p.d. over the phases;

c) obtain the likelihood value by replacing the F^{calc}(**h**) with F^{obs}(**h**) in the formula for the magnitude probability distribution.

4.1. Joint probability distribution of real and imaginary parts of a structure factor.

The simplest way to get j.p.d is based on The Central Limit Theorem of the Theory of Probabilities. This theorem states that the sum of independent (or "slightly" dependent) random variables is distributed in accordance with the Gaussian distribution. This means that we know in advance the shape of distribution and all what must be defined is a small number of the distribution parameters.

We consider first one nc/s structure factor and study the j.p.d. of its real and imaginary parts. In the general case the two-dimensional Gaussian distribution is defined by five parameters, namely mean values and dispersions corresponding to every of the two coordinates and the correlation coefficient between these coordinates. However, in our case it is defined, in fact, by only one parameter *b* (Srinivasan & Parthasarathy, 1976) which is equal to $N_{add}*g(h)$:

$$P(A,B)=\frac{1}{\pi\beta}\exp\left[-\frac{\left(A-A_{part}\right)^2+\left(B-B_{part}\right)^2}{\beta}\right]. \quad (4.1)$$

Here *A* and *B* are real and imaginary parts of a structure factor corresponding to the mixed model (the known partial model plus random atoms), they both are random variables. A_{part} and B_{part} corresponds to the partial model, they are some defined values and not random variables! The parameter *b* is defined as $<A_{add}^2>=<B_{add}^2>=b/2$, A_{add} and B_{add} correspond to the additional atoms and <...> means the expected value of the corresponding random variable.

4.2. Joint probability distribution of magnitude and phase of a structure factor.

The j.p.d. of the magnitude and phase can be obtained simply by rewriting in "polar coordinates" the Gaussian distribution derived above:

$$P(F,\varphi)=\frac{F}{\pi\beta}\exp\left[-\frac{F^2+F_{part}^2}{\beta}\right]\exp\left[\frac{2FF_{part}}{\beta}\cos\left(\varphi-\varphi_{part}\right)\right]. \quad (4.2)$$

Here $F\exp[ij\ ]$ is a random structure factor corresponding to the mixed model, and $F_{part}\exp[ij_{\ part}]$ is those corresponding to the partial model; the latter is some fixed (not random) value.

4.3. Marginal distribution for s.f. magnitude. The likelihood.

Now we can derive a marginal distribution for the magnitude by integration of j.p.d. "magnitude-phase" over phases.

$$P(F) = \int_0^{2\pi} P(F,\varphi)d\varphi = \frac{2F}{\beta}\exp\left[-\frac{F^2 + F_{part}^2}{\beta}\right] I_0\left(\frac{2FF_{part}}{\beta}\right) \qquad (4.3)$$

and calculate the marginal likelihood $L=P\{F = F^{obs}\}$, provided a single experimental observation is taken into account

$$L = \frac{2F^{obs}}{\beta}\exp\left[-\frac{\left(F^{obs}\right)^2 + F_{part}^2}{\beta}\right] I_0\left(\frac{2F^{obs}F_{part}}{\beta}\right). \qquad (4.4)$$

Here $I_0(t)$ stands for the modified Bessel function of the zero order (*Appendix A4*).

In practice we have, obviously, observed magnitude values for many reflections. Therefore, the likelihood must be calculated using all these observations. The Central Limit Theorem allows to get j.p.d. of all magnitudes and phases, corresponding to the mixed model. This distribution is defined by the mean values and the full set of second order moments calculated for the real and imaginary parts of the structure factors. Nevertheless, it is not possible, in general, to perform in a close form the integration over the phases to get marginal distribution for s.f. magnitudes. A possible way is to neglect the correlations between different s.f. and to consider different complex structure factors as independent random variables ("diagonal approximation"). In this case the probability distribution for a set of s.f. magnitudes is just the product of distributions corresponding to individual reflections and the likelihood is

$$L = L\left(\mathbf{r}_1,\ldots,\mathbf{r}_{N_{part}}\right) =$$

$$\prod_{\mathbf{h}\in S}\frac{2F^{obs}(\mathbf{h})}{\beta(\mathbf{h})}\exp\left[-\frac{\left(F^{obs}(\mathbf{h})\right)^2 + F_{part}^2(\mathbf{h})}{\beta(\mathbf{h})}\right] I_0\left(\frac{2F^{obs}(\mathbf{h})F_{part}(\mathbf{h})}{\beta(\mathbf{h})}\right) \qquad (4.5)$$

We remind that $F_{part}(\mathbf{h})$ here depends on the model coordinates $\{\mathbf{r}_j\}$, while $F^{obs}(\mathbf{h})$ and $b(\mathbf{h})$ are known values.

5. Likelihood maximization

As the logarithm is a monotonically growing function, any function and its logarithm have minima or maxima simultaneously. This means that in our case we can replace the likelihood maximization by the maximization of its logarithm which computationally is much more convenient

$$\ln L\left(\mathbf{r}_1,\ldots,\mathbf{r}_{N_{add}}\right)=\sum_{\mathbf{h}\in S}\frac{2F^{obs}(\mathbf{h})-\left(F^{obs}(\mathbf{h})\right)^2}{\beta(\mathbf{h})}$$

$$-\sum_{\mathbf{h}\in S}\left\{\frac{F_{part}^2\left(\mathbf{h};\{\mathbf{r}_j\}\right)}{\beta(\mathbf{h})}-\ln I_0\left(\frac{2F^{obs}(\mathbf{h})}{\beta(\mathbf{h})}F_{part}\left(\mathbf{h};\{\mathbf{r}_j\}\right)\right)\right\}. \quad (5.1)$$

Additionally, it is convenient to skip the first sum which is independent on the partial model coordinates and to maximize the so called Logarithm Likelihood Gain (LLG) value

$$LLG=-\sum_{\mathbf{h}\in S}\left\{\frac{F_{part}^2\left(\mathbf{h};\{\mathbf{r}_j\}\right)}{\beta(\mathbf{h})}-\ln I_0\left(\frac{2F^{obs}(\mathbf{h})}{\beta(\mathbf{h})}F_{part}\left(\mathbf{h};\{\mathbf{r}_j\}\right)\right)\right\}. \quad (5.2)$$

The maximization of the LLG is equivalent to the minimization of

$$R_{ML}\left(\mathbf{r}_1,\ldots,\mathbf{r}_{N_{part}}\right)=$$

$$\sum_{\mathbf{h}\in S}\left\{\frac{F_{part}^2\left(\mathbf{h};\{\mathbf{r}_j\}\right)}{\beta(\mathbf{h})}-\ln I_0\left(\frac{2F^{obs}(\mathbf{h})}{\beta(\mathbf{h})}F_{part}\left(\mathbf{h};\{\mathbf{r}_j\}\right)\right)\right\}, \quad (5.3)$$

so ML-refinement is nothing but minimization of R_{ML} residual and all relevant minimization methods may be used.

6. Analysis of R_{ML} residual.

Let's analyze more thoroughly a particular item in the last sum as a function depending on $F_{part}(\mathbf{h})$:

$$\psi_{\mathbf{h}}(F)=\frac{F^2}{\beta}-\ln I_0\left(\frac{2F^{obs}}{\beta}F\right). \quad (6.1)$$

For large F values I_0 grows near exponentially, so the second term grows near linearly and the whole $y_{\mathbf{h}}(F)$ function grows as a quadratic function. For small F values we have

$$\psi_{\mathbf{h}}(F)\approx\frac{1}{\beta}\left[1-\frac{\left(F^{obs}\right)^2}{\beta}\right]F^2. \quad (6.2)$$

This shows that the function $y_{\mathbf{h}}(F)$ has different behavior depending on the ratio of $(F^{obs})^2$ to b .

For relatively small F^{obs} (the ratio is smaller than 1) the minimum of $y_{\mathbf{h}}(F)$ is attained for F=0. This means that in order to minimize the R_{ML}, the calculated s.f. magnitudes $F_{part}(\mathbf{h};\{\mathbf{r}_j\})$ for relatively week reflections must be fit to zero. If F^{obs} are relatively large (the ratio $(F^{obs})^2/b$ is larger than 1), then the function $y_{\mathbf{h}}(F)$ has nonzero minimum F^* and in order to minimize R_{ML} the $F_{part}(\mathbf{h};\{\mathbf{r}_j\})$ values in the corresponding members in the sum must be fit to $F^*(\mathbf{h})$.

In other words, the ML refinement is similar to the minimization of a simplified residual

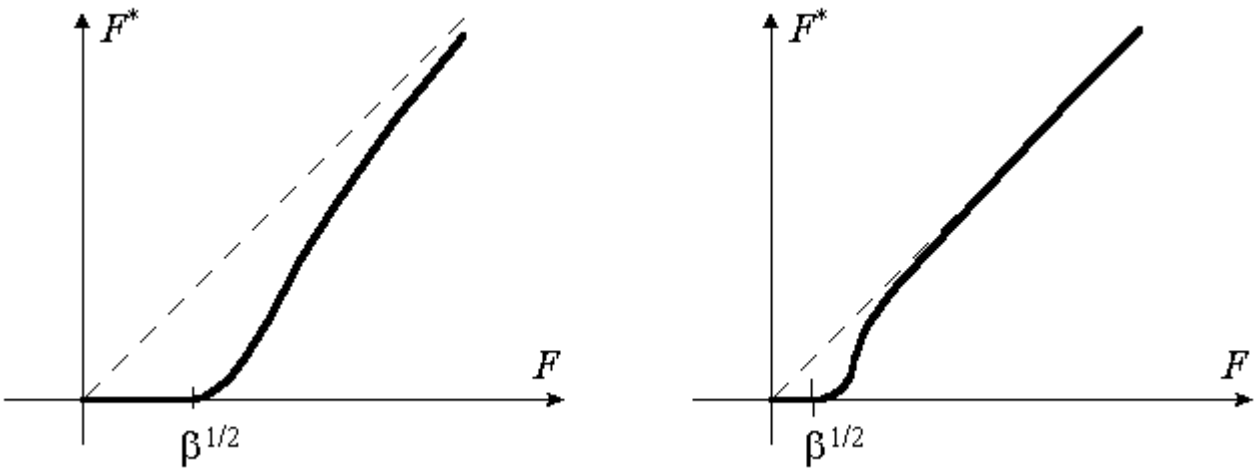
$$R_{simpl} = R_{simpl}\left(\mathbf{r}_1,\ldots,\mathbf{r}_{N_{part}}\right) = \sum_{\mathbf{h}\in S'}\left[F^{calc}\left(\mathbf{h},\{\mathbf{r}_j\}\right) - F^*(\mathbf{h})\right]^2 \qquad (6.3)$$

where $F^*(\mathbf{h})$ may be thought as modified observed values.

The modified magnitudes $F^*(\mathbf{h})$ are equal to zero for week reflections (with $(F^{obs})^2<b$). For relatively strong reflections the modified value $F^*(\mathbf{h})$ are found from the condition

$$\psi_{\mathbf{h}}(F) = \frac{F^2}{\beta(\mathbf{h})} - \ln I_0\left(\frac{2F^{obs}(\mathbf{h})}{\beta(\mathbf{h})}F\right) \Rightarrow \text{minimum}, \qquad (6.4)$$

and have nonzero values. The dependence $F^*(F^{obs})$ is shown for different b at the figure below (see *Appendix 5* for details). It is worthy of noting that the modified value $F^*(\mathbf{h})$ is always less than the experimental one.



It must be noted that the scale of the modification of the observed values depends on the value of *b* parameter which reflects the completeness of the current model. If (in the considered situation) the number of atoms absent in the model is big, then *b* value is high and many modified values *F**(**h**) are taken as zero ones, while other may be significantly less than F^{obs}(**h**) values. If the number of absent atoms (and *b*) is small, then deviations *F**(**h**) from F^{obs}(**h**) are small and ML-refinement coincides with the conventional one.

7. Discussion.

Some additional issues are worthy of noting.

1. It was shown that the ML-refinement may be considered as an attempt to fit the calculated (from an incomplete model) structure factors magnitudes to some modified experimental magnitudes. The modification consists in reduction of values of structure factor magnitudes; week magnitudes becomes zeros but are not excluded from the refinement. The cut-off level and the degree of reduction depend on the value of the parameter *b* , in other words, on the number of absent atoms in the considered case. If the relative number of absent atoms is relatively small, the modified *F**(**h**) values are close to F^{obs}(**h**) and ML-refinement is reduced to the conventional refinement.

2. In the considered case the j.p.d. for magnitude and phase of s.f. depends on one parameter *b* . In a more general situation it may depend on two parameters *a* and *b*

$$P(A,B)=\frac{1}{\pi\beta}\exp\left[-\frac{\left(A-\alpha A_{part}\right)^2+\left(B-\alpha B_{part}\right)^2}{\beta}\right]$$

which reflect the accuracy of s.f.-formula and must be defined in advance. The procedure based on the maximization of a marginal likelihood corresponding to the test set of reflections may be used for these purposes (Lunin & Skovoroda, 1995; Read, 1997). It was discussed also (Lunin & Skovoroda, 1995) that this distribution is valid for many cases of errors and therefore the considerations done above are valid in many other applications.

3. The simplified residual may be written in a more close to R_{ML} form if the weights are applied which are equal to $y_{\mathbf{h}}''(F^*)$.

4. To introduce stereochemical or energy restraints into the refinement the usual penalty functions may be added to R_{ML} residual.

This work was partially supported (VYL) by RFBR grant 97-04-48319 and by CNRS Fellowship.

8. References.

Bricogne, G., Irwin, J. (1996) Macromolecular Refinement: Proceedingof the CCP4 Study Weekend, E.Dodson, M.Moore, A.Ralph & S.Bailey, eds., pp.85-92. Warrington : Daresbury Laboratory.

Lunin, V.Yu. & Skovoroda, T.P.(1995). *Acta Cryst*. A**51**, 880-887.

Murshudov, G.N., Vagin, A.A. & Dodson, E.J. (1997). *Acta Cryst*. D**53**, 240-255.

Pannu, N.S., Murshudov, G.N., Dodson, E.J. & Read, R.J. (1998) *Acta Cryst*. D**54**, 1285-1294

Pannu, N.S. & Read, R.J. (1996). *Acta Cryst*. A**52**, 659-668.

Read, R.J. (1997) In *Methods in Enzymology*, Academic Press, San Diego., C.W.Carter, Jr., R.M.Sweet, eds., 277, part B, 110-128.

Srinivasan, R. & Parthasaraty, S. (1976). *Some Statistical Applications in X-ray Crystallography.* Oxford: Pergamon Press.

Appendixes.

A1. Structure factors formula

We suppose that the structure factors are connected with the atomic coordinates by means the s.f.-formula

$$F^{calc}(\mathbf{h})\exp\left[i\varphi^{calc}(\mathbf{h})\right]=\sum_j f_j(\mathbf{h})\exp\left[2\pi i(\mathbf{h},\mathbf{r}_j)\right], \qquad \text{(A1.1)}$$

here

$\{\mathbf{r}_j\}$ are atomic coordinates; **h** is a reciprocal space vector; $f_j(\mathbf{h})$ are known functions which include both scattering and temperature factors and atomic occupancies; $\sum_j$ is calculated over all atoms included into the model.

In our consideration the values of the magnitudes and phases of calculated s.f. are functions which depend only on the coordinates of the atoms included into the model.

A2. Randomly generated additional atoms

We consider as the object of the refinement a current partial model which includes N_{part} atoms only while N_{add} atoms (solvent atoms *e.g.*) are absent:

$N_{full}=N_{part}+N_{add}$;

$$\vec{F}_{full}^{calc}(\mathbf{h}) = \vec{F}_{part}^{calc}\left(\mathbf{h};\mathbf{r}_1,\ldots,\mathbf{r}_{N_{part}}\right) + \vec{F}_{add}^{calc}\left(\mathbf{h};\mathbf{u}_1,\ldots,\mathbf{u}_{N_{add}}\right), \qquad \text{(A2.1)}$$

where

$$\vec{F}(\mathbf{h}) = F(\mathbf{h})\exp\left[i\varphi(\mathbf{h})\right]; \qquad \text{(A2.2)}$$

$\mathbf{r}_1,\ldots,\mathbf{r}_{Npart}$ are coordinates of atoms which are included into the current model (these coordinates are known approximately), while $\mathbf{u}_1,\ldots,\mathbf{u}_{Nadd}$ are totally unknown coordinates of the atoms absent in the model.

We consider here the coordinates of additional atoms $\mathbf{u}_1,\ldots\mathbf{u}_{Nadd}$ as random variables distributed uniformly in the unit cell and suppose that all atoms are of the same type and have the same scattering factor g(h). So

$$F_{add}^{calc}(\mathbf{h})\exp\left[\varphi_{add}^{calc}(\mathbf{h})\right] = N_{add}\,g(h)\sum_{j=1}^{N_{add}}\exp\left[2\pi i\left(\mathbf{h},\mathbf{u}_j\right)\right], \qquad \text{(A2.3)}$$

where $\{\mathbf{u}_j\}$ are the coordinates of additional atoms. (More sophisticated hypotheses may be considered in a similar way which take into account an extra information about possible positions of additional atoms.)

In such the case in the identity

$$\vec{F}_{full}^{calc}(\mathbf{h}) = \vec{F}_{part}^{calc}(\mathbf{h}) + \vec{F}_{add}^{calc}(\mathbf{h}); \qquad \text{(A2.4)}$$

$\vec{F}_{add}^{calc}(\mathbf{h})$ is a random (complex) variable;

$\vec{F}_{part}^{calc}(\mathbf{h})$ is a determined (non-random) value which depends on the partial model coordinates $\mathbf{r}_1,\ldots,\mathbf{r}_{Npart}$.

The “mixed model” structure factor $\vec{F}_{full}^{calc}(\mathbf{h})$ is now a random variable (as it includes a random part $\vec{F}_{add}^{calc}(\mathbf{h})$), but its probability distribution is different for different sets of coordinates $\mathbf{r}_1,\ldots,\mathbf{r}_{Npart}$ of the partial model.

A3. The likelihood corresponding to the partial model

The likelihood value corresponding to some statistical hypothesis may be considered as the probability to reproduce the experimentally observed data under this hypothesis (to be mathematically correct, when the observations correspond to continuous random variables it is necessary to speak about probability distribution function values rather then probabilities).

In the studied case we can formulate the statistical hypotheses as follows:

Statistical hypothesis $H(\{\mathbf{r}_1,\ldots,\mathbf{r}_{Npart}\})$

The correct structure may be obtained by adding of N_{add} atoms with randomly generated coordinates $\mathbf{u}_1,\ldots,\mathbf{u}_{Nadd}$ to the N_{part} atoms with the known positions $\mathbf{r}_1,\ldots,\mathbf{r}_{Npart}$.

Every of these hypothesis is specified by the set of the partial model coordinates $\{\mathbf{r}_j\}$. As a consequence, the problem of the choice of the particular values of these coordinates may be formulated as the problem of the choice of the hypothesis which is the most consistent with the experimental data.

Under the hypothesis *H*, the values of calculated (with the use of s.f.-formula A3) s.f. magnitudes are random variables (as they depend on random $\{\mathbf{u}_j\}$) and we can speak about the probability for these variables to have some particular values.

For every hypotheses H, we can define the probability of the calculated s.f. magnitudes to be equal to the observed ones:

$L(H)$=Probability$\{F^{calc}(\mathbf{h}) = F^{obs}(\mathbf{h})$ for **h** from S$\}$,

where *S* is the set of experimentally measured intensities. This value $L(H)$ is called the likelihood corresponding to this hypothesis. As the hypothesis is specified by the partial model coordinates, the likelihood value $L(H)$ is the function depending on the model coordinates:

$L(H)=L(\mathbf{r}_1,\ldots,\mathbf{r}_{Npart}) =$

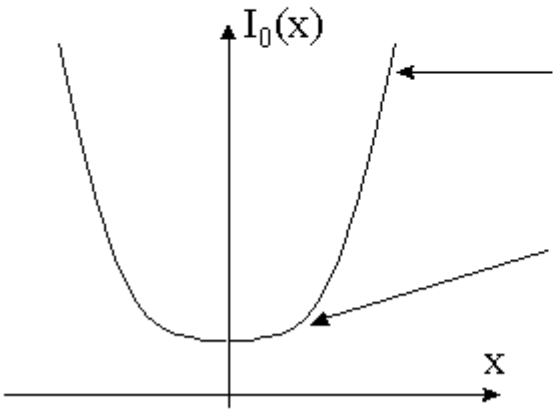
Probability$_{\{\mathbf{u}\}}\{F^{calc}(\mathbf{h};\{\mathbf{r}_j\}+\{\mathbf{u}_j\}) = F^{obs}(\mathbf{h})$ for **h** from S$\}$,

where the probability appears due to random coordinates $\{\mathbf{u}_j\}$.

Naturally, this probability varies with different hypotheses (i.e., for different partial model coordinates) and the maximal likelihood principle suggests to accept the hypothesis which possesses of the maximal likelihood value, i.e. to accept values of the partial model coordinates which maximize the $L(\mathbf{r}_1,\ldots,\mathbf{r}_{Npart})$.

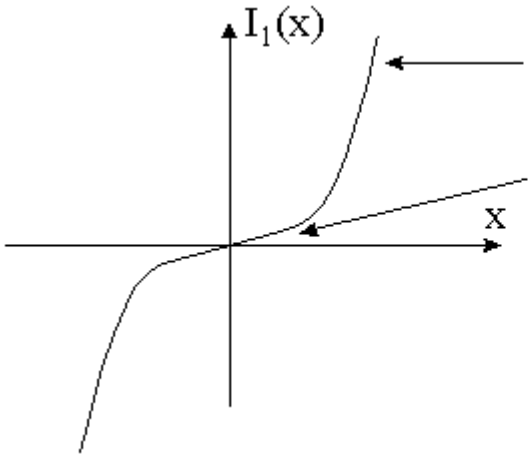
A4. Modified Bessel functions

$$I_0(x) = \frac{1}{\pi}\int_0^{\pi} \exp[x\cos\theta]d\theta$$



$$I_0(x) = \frac{\exp[x]}{\sqrt{2\pi x}}\left\{1+\frac{1}{8x}+...\right\} \quad \text{for large x}$$

$$I_0(x) = 1+\frac{1}{4}x^2+... \quad \text{for small x}$$



$$I_1(x) = \frac{\exp[x]}{\sqrt{2\pi x}}\left\{1-\frac{3}{8x}+...\right\} \quad \text{for large } x > 0$$

$$I_1(x) = \frac{1}{2}x+\frac{1}{16}x^3+... \quad \text{for small } x$$

A5. Modified observed magnitudes

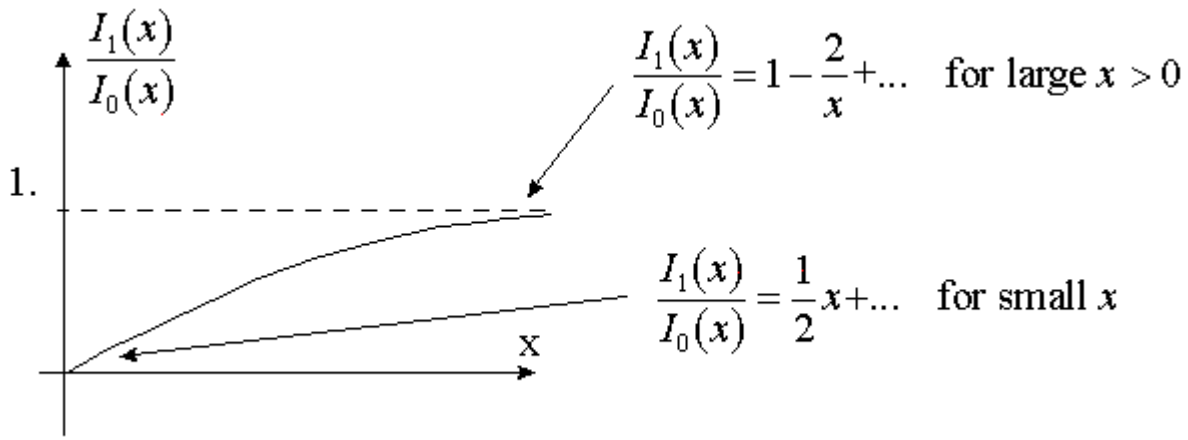
The modified values $F^*(\mathbf{h})$ may be found from the condition

$$\psi'_{\mathbf{h}}\left(F^*\right) = 0 \,, \qquad \text{(A5.1)}$$

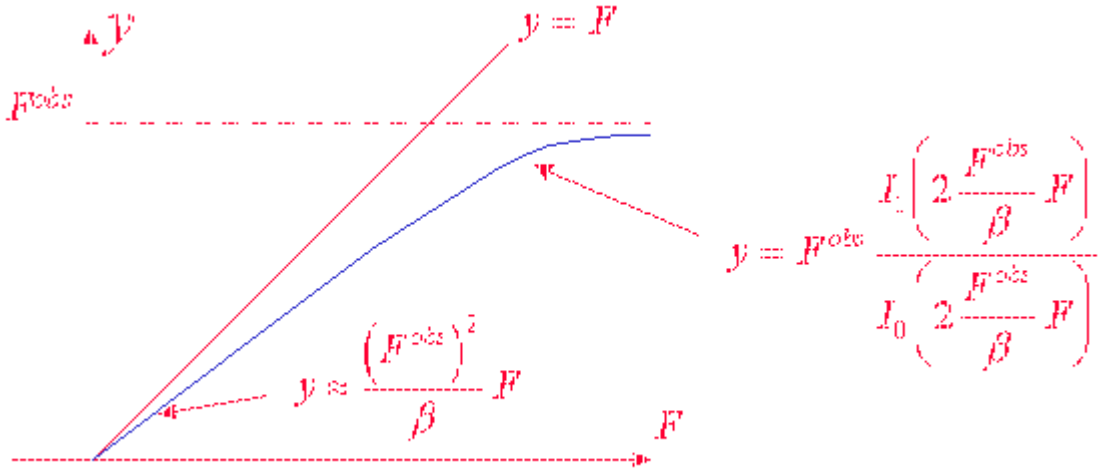
i.e. from the equation

$$F^* = F^{obs}\frac{I_1\left(2\dfrac{F^{obs}}{\beta}F^*\right)}{I_0\left(2\dfrac{F^{obs}}{\beta}F^*\right)} \,. \qquad \text{(A5.2)}$$

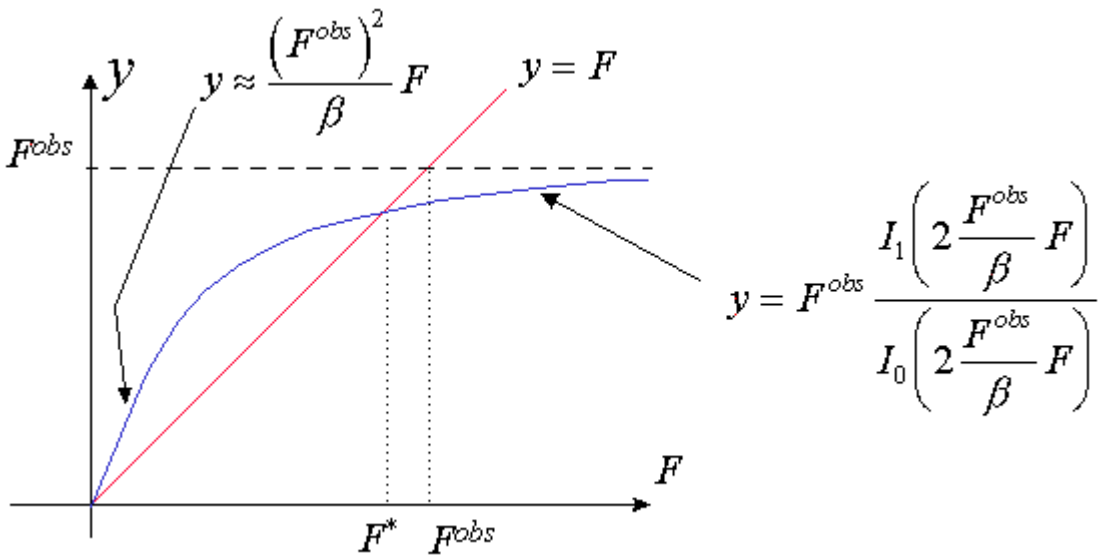
The ratio of the first and zero order modified Bessel functions has the following properties:



So, depending on the value $(F^{obs})^2/b$ two cases can exist:



The case $(F^{obs})^2 < b$. The equation y ’(F)=0 has the only solution F^*=0.



The case $(F^{obs})^2 > b$. The equation y ’(F)=0 has a nonzero solution F^*.

It is easy to see that in this case $F^* < F^{obs}$.

Circular Dichroism Spectroscopy and X-ray Crystallography: A Dynamic Duo

B. A. Wallace

Centre for Membrane Structure and Dynamics, Daresbury Laboratory

and

Dept. of Crystallography, Birkbeck College, University of London

At the recent Daresbury PX Users meeting, the complementary natures of circular dichroism (CD) spectroscopy and X-ray crystallography were discussed. While it is clear that CD is no substitute for X-ray data, it can provide both dynamic and static structural information and information that can be of assistance in producing crystals.

What you can learn from CD Spectroscopy (while you are waiting for crystals to grow):

CD spectroscopy is a useful technique for examining protein structures in solution under different conditions and for examining both their dynamics and folding pathways. The former types of experiments can be important for establishing the "physiological relevance" of the conditions used, say, for crystallisation. The latter can be important for examining the conformational effects of ligand binding or complex formation with another macromolecule. In addition, the processes associated with protein folding can be monitored on a fast time scale.

Secondary and tertiary structures can be monitored by the peptide transitions in the far UV (~190-220 nm) and by the aromatic side chain transitions in the near UV (~270-290 nm), respectively. Other chromophores in the visible region (for coloured proteins) may be valuable for establishing the fidelity of tertiary and quaternary folding.

In advance of obtaining a crystal structure, CD can be useful in conjunction with modelling efforts: it can provide a good test of the secondary structural content of any model produced, and can be used to determine the class of molecule. For crystallography, it may be used as an assay of whether a particular structure will be a reasonable model for molecular replacement.

CD can also be used to examine conformational changes associated with ligand or drug binding, thereby aiding in elucidating the mechanism of action of enzymes and receptors. These changes can be quantitated by CD since shifts in conformation involving as few as 10 amino acids may be readily detectable.

How CD Spectroscopy can help you to grow (or improve) crystals:

CD spectroscopy may be useful for establishing protein integrity and conditions for complex formation:

Ligand binding studies can allow titration of complex formation, thereby assuring a single species under the crystallisation conditions (as opposed to a mixture of proteins with and without ligand bound). It can also be useful for monitoring complex formation with other macromolecules.

Examination of the effects of different pHs, detergents, solvents or other additives can be a guide to the integrity of the protein and "physiological relevance" of the crystallisation conditions.

The secondary structures and thermal stability of molecular biology constructs created to produce domain structures for crystallisation can be assayed, as can the fidelity of refolding of proteins initially expressed as inclusion bodies or in other denatured forms.

Finally, thermal denaturation studies by CD (melting curves) can be used to establish means of enhancing the stability of the protein by addition of various types of additives to the crystallisation mixtures.

The Centre for Protein and Membrane Structure and Dynamics (CPMSD):

Earlier this year the BBSRC created the CPMSD at Daresbury as one of six Structural Biology Centres throughout the U.K. This centre is focused on CD spectroscopy, both conventional and synchrotron-based. As well as providing a state-of-the-art facility for SRCD data collection, the Centre will be a "virtual centre" for CD and its interpretation, where "visitors" will have access to an Expert Panel of advisors, a user-friendly website for data analyses, an interactive bulletin-board, and BioMOO-type virtual meetings to discuss current techniques and methods for analysis of data. The SRCD at the Centre will be a unique facility world-wide with capabilities for low wavelength, high resolution and very fast stopped-flow measurements.

The annual U.K. CD Spectroscopy Meeting will be held on 21 October 1999 at the Daresbury Laboratory. Registration is free and open to any interested person. If you wish to attend, please contact: p.b.broadhurst@dl.ac.uk prior to 15 October.

More information on CD spectroscopy and the CPMSD can be obtained at: http://www.dl.ac.uk/SRS/VUV-IR/CD/cpmsd.html

CCP4 served as you like it

A general overview of CCP4 portability

Alun Ashton, CCP4 Daresbury

A.W.Ashton@dl.ac.uk

CCP4 on IRIX

Silicon Graphics machines' running the IRIX operating system are still the most popular platform for crystallographers world wide. The current version of IRIX is 6.5 and for a change this version should run on all SGI machines. Having said that, rumor has it most machines require a new hard drive since 6.5 is very big. Many users have expressed concerns about needing IRIX 6.5 to ensure Y2K compliance. A visit to the SGI web pages on www.sgi.com seems to imply that all the major IRIX versions from 5.3 onwards can be made Y2K compliant by a series of free patches (see http://www.sgi.com/tech/year2000/o_s.html or contact your SGI representative).

As far as CCP4 is concerned the main software suite (version 3.5.1), the X-windows programs and the GUI will all compile and work well with IRIX (I may regret saying that!). However, if you do not compile the suite with the default options i.e. running configure with the irix system option, there may be problems. Compiling the suite as 64 bit (on IRIX 6.+, configure irix64) can sometimes be more problematic especially with the X-windows programs. Users of IRIX 5.3 should also refer to the CCP4 problems page as the compilers also have a problem with naming conventions recently adopted by some CCP4 libraries.

If you're buying a new SGI be sure to check you will also get the compilers - they will be needed to install/compile CCP4. There are reports of successes in compiling CCP4 with the GNU Freeware compilers but you would save time and effort by getting the SGI compilers.

CCP4 on LINUX

LINUX is now maturing from a trendy operating system to becoming an operating system of choice. At the highest level the battles between LINUX distributions are continuing. RedHat has long dominated the market after displacing the slackware distributions in the mid 90's. But now the LINUX trend setters/puritans are turning to other distributions - the fear being of another 'Microsoft' dominating the market.

As far as CCP4 is concerned it doesn't matter. The main problem we have as users is the reliability of the compilers. The gcc, g77, egcs GNU compilers are now the accepted way to compile CCP4 software on LINUX but as this is a freeware project, releases are often prone to undesirable features that will cause problems in compiling a large suite such as CCP4.

The CCP4 LINUX page is updated to reflect the changes needed when new version of compilers come out (apologies that this is not done more regularly). Currently the main change that is not in version 3.5.1 of CCP4 is the use of the g2c.h header file instead of the f2c.h file. This is a problem if you install the freeware version of RedHat 6.0.

Some long-standing problems with LINUX are

1. temporary files are not deleted automatically
2. the Xfree86 server must be in 16bit mode for some of the CCP4 X-windows programs, and mosflm to work (you will know what I mean if you've tried it!)
3. with the current version of the GUI, tcl/tk must be recompiled otherwise an undesirable feature of the GUI will appear in that it will never know if a job has finished!

CCP4 on OSF1

I have little experience with the Alpha operating systems and compilers, but I do know some high level users and developers do use it and have very few problems. Recently a problem was reported with optimization on one of the programs and this is now documented in the problems page.

One 'small' outstanding problem is with loggraph. Loggraph, a replacement for xloggraph, is used by the GUI and needs the graphical extensions in a package known as BLT. To the best of my knowledge BLT does not work correctly on OSF1 and so neither does loggraph.

CCP4 on SunOs

Martyn Winn here at Daresbury has a SUN on his desk and he can happily compile and use all the CCP4 programs and the GUI. But Martyn is very 'careful with his money' so he hasn't got any of SUNs distributed compilers. He chose to use the GNU compilers instead. I don't have any current information on the SUN compilers but from Martyn's experiences we do know its possible to get the suite to work!

CCP4 on HP-UX

The compiler on HP-UX machines is possibly the most unforgiving compiler on a UNIX operating system. But as there are users who use HP machines as servers and others as workstations, so the suite is tested on HP-UX. To date I know the suite will compile and run under HP-UX but there are some problems with the x-windows programs.

CCP4 on VMS

VMS is still a popular operating system with some well established UK and worldwide research groups. Because of this we at CCP4 get a lot of 'feedback' and fixes that ensure the suite continues to compile and work on VMS.

CCP4 on other UNIX OS

There are/have been many other hardware manufacturers who have their own particular variation of UNIX and whenever possible CCP4 has worked with users of these systems to get a working suite. For a current list of supported platforms see the configure script.

CCP4 on Windows NT™

The increasing power and popularity of the Microsoft operating systems has inevitably meant that crystallographers see the potential for exploiting their desktop machines. For a

detailed progress report and discussion on the approach taken to get CCP4 working on NT please see the web pages in the developers area.

As a summary of these pages I can tell you that progress on the port is currently going to schedule. The boundaries between the stages have become a little less defined as testing the CCP4 libraries has lead to several programs being compiled. Some other good news is that the tcl/tk and BLT versions for NT have been tested and with some alterations both the GUI and loggraph work well. If the work can be continued at its current pace we are looking at a release date for ***CCP4nt*** sometime in the New Year.

Dessert

You want more? Well without you there can be no more! So whenever you find new information or even if you have more upto date information than I have put here please let us know at ccp4@dl.ac.uk. We look forward to hearing from you.

Beamline 14 at the SRS Daresbury Laboratory

Dr. E. Duke

CLRC Daresbury Laboratory, Warrington, Cheshire. WA4 4AD

Beamline 14 is a new beamline being built on the Synchrotron Radiation Source (SRS) at Daresbury Laboratory. The beamline will have 2 stations for protein crystallography. One of the stations will be of high intensity and will be able to access the selenium k-edge at 0.979Å. The other station will operate at higher wavelengths in order to gain the most benefit from the multipole wiggler insertion device.

The SRS at Daresbury was the first ever storage ring dedicated to the production of synchrotron radiation, it is what is termed a second-generation source. In reality what this means is that there are no long straight sections in the machine where insertion devices can be positioned with ease. However there was a desire to gain experience with insertion devices both from the machine point of view and the beamlines, as working with insertion devices requires different techniques especially when it comes to dealing with the high heatloads produced by the photon beams. This desire fitted neatly with a demand for more high intensity protein crystallography stations on the SRS; thus the concept of Beamline 14 came into being.

Given the parameters of the SRS and the desire for high flux in the hard X-ray region a multipole wiggler (MPW) was best suited to our requirements. The accelerator physicists carried out various experiments on the SRS to determine the minimum electron beam height before the lifetime of the stored beam was seriously compromised. This then dictated the minimum gap of the multipole wiggler. It was found that a gap of 20mm was possible. Calculations were then carried out to determine the magnet parameters. Two stations on the beamline were planned so good flux off-axis was an important criterion as well as that in the centre. For this reason a 2.0T magnet with 9 full poles plus two half-poles was chosen. The other possibility was a 1.8T magnet with 11 poles which gives 4.3×10^{13} photon/s/mrad at 300mA into 0.1% bandwidth at the centre of the fan. However there is a significant deterioration in flux going out to 6.5mrad. The 2.0T, 9 pole magnet gives marginally less flux at the centre (4.2×10^{13} photon/s/mrad at 300mA into 0.1% bandwidth) but the reduction in flux out towards 6.5mrad horizontally is much less.

Having established the magnet design the next stage was to design the beamline starting with the position that the 2 stations would take on the fan of radiation from the MPW. As the flux from the MPW drops by roughly an order of magnitude between the centre of the fan and 6.5mrad it is important to position the two stations as close to the centre as possible. This poses a problem: there needs to be space between the stations so that, for example, the monochromator mounting assembly does not put a shadow onto the beam for the adjacent station. A novel solution was found: have the beams for the two stations cross over each other. This allows the dead space on the outside of the radiation fan to be used for all the cooling and bending mechanisms. Calculations were done to establish the best horizontal ranges for the two stations. The best beam is at the centre of the fan. However if one station were set symmetrically about the centre then the second station would have significantly less flux. A compromise was found: one station was slightly offset from the centre of the fan bringing the second stations closer to the centre. The gain in beam quality for the second station more than compensates for the slight loss in flux for the first station. Therefore Station 14.2, the most intense station, takes beam from +3mrad to –1mrad and Station 14.1 from –1.5mrad to –4.5mrad.

In order for the beamline to be built quickly and have it commissioned for users as soon as possible simplicity of design was important. Therefore the decision was made to use an optical configuration similar to that used on 2 existing protein crystallography stations on the SRS, Stations 9.6 and 7.2. Thus the two stations on Beamline 14 have vertical focussing by a cylindrical mirror and monochromation and horizontal focussing with a single bounce monochromator. The effect of the single bounce monochromator, to deflect the beam out sideways, also allows the two stations to be fitted into the limited space available for the beamline.

The optical configuration chosen for the stations tends to lead to operation at a single wavelength - both Station 9.6 and Station 7.2 operate as fixed wavelength stations (0.87Å and 1.488Å respectively). However it is possible to provide a degree of tunability around certain pre-determined wavelengths and this is what is planned on Station 14.2. Both stations will operate at two wavelengths given by 2 different monochromators. For Station 14.2 these wavelengths will be 0.97Å and 1.2Å and for Station 14.1, 1.2Å and 1.5Å. The flux output from the multipole wiggler increases with wavelength in this wavelength region. However given the current demand for data collection facilities at the Se k-edge (0.979Å) it seemed sensible to provide this facility on Station 14.2. Even with slightly less flux we estimate that the flux provided at 0.97Å will be in the region of 20 times greater than what is available on Station 9.5. It should also be possible to provide tunability around other absorption edges. It is intended to look at the operational issues pertaining to this during the commissioning of the beamline.

In order to allow the wavelength to be changed between the 2 wavelengths of operation (eg. 1.2Å and 1.5Å on Station 14.1) a new mount has been developed which allows the 2 monochromators to be stacked in a "double-decker" arrangement. This means that the wavelength change can be effected by a vertical translation rather than the lengthy process of letting the monochromator vessel up to atmosphere, changing the monochromator followed by pumping back down again.

The heat-load within the X-ray beam on beamline 14 is great. Therefore all of the beamline components which could be struck by the beam will be water-cooled. This includes the mirror and monochromators and all the slits.

The two stations will be equipped with state-of-the-art detectors. Station 14.2 will have a PX210 made by Oxford Instruments and an ADSC Quantum 4R will be mounted on Station 14.1. The PX210 is a 3x3 CCD array with an active area of 210mm x 210mm. The full-frame readout time is 1.8s; this is reduced to 0.5s if 2x2 binning is used. This fast readout time will be of benefit for data collection in fine phi-slicing mode. The Quantum 4R is an updated version of the detector currently on Station 9.6. Here the active area of the 2x2 CCD array is 180mm x 180mm and the full frame "slow" readout is 9s. This time drops to 3s with the full frame fast readout. Initially the ADSC Quantum 4R will be on Station 14.2 as delivery of the PX210 is not expected until January 2000. Once the PX210 is commissioned on Station 14.2 the ADSC Quantum 4 will be installed on Station 14.1. The rotation axes on both stations will be an updated version of the rotation axis currently on Station 7.2 that was designed at the EMBL in Hamburg.

Commissioning has started on Beamline 14 and is progressing well. The first users are scheduled on Station 14.2 in late November. However prior to that users will be invited to come and collect data on Station 14.2. Station 14.1 will follow on after Station 14.2 with first scheduled users in early February.

Recent ccp4bb discussions

Martyn Winn

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

This article is an attempt to summarise some of the discussions recently held on the CCP4 Bulletin Board. Many of these discussions reflect common concerns of protein crystallographers, and should be of general interest. For each subject below, the original question is given in italics, followed by my 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 www.ccp4.ac.uk/ccp4bb.php 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 questionner. While ccp4bb is obviously alive and well, we think there is still some way to go before the level of traffic becomes inconvenient.

Bulk solvent correction in REFMAC and CNS

(May and July 1999)

> *We started our initial refinement using CNS (ver. 0.5). Bulk solvent correction was applied for resolution range 15.00 to 1.9 A. The value of R and R-free was very reasonable (R=21.0 %; Rfree=24.5 %).*
>
> *[With REFMAC] ... the 'bulk solvent correction' results were not as good as CNS. The R-free value is also higher (26.4%) and the R - R-free difference is also higher when we use REFMAC.*
>
> *May I know why this difference occurs with REFMAC? If I use SCALE TYPE BULK (taking FPART and PHIPART from CNS results), will it be a good idea?*

CNS makes a bulk solvent correction by masking the protein, filling the remaining bulk solvent region with a constant electron density and then making an FFT (with a B-factor applied) to generate FPart and PHIPart for the solvent. In contrast, REFMAC uses the exponential scaling model, based on the Babinet principle which states the Fourier transform of the solvent mask is related to the Fourier transform of the protein mask by a 180 deg phase shift. In the limit of low resolution, this implies that the FT of the solvent region is directly proportional to the (known) FT of the protein region apart from a 180 deg phase shift. However, the bulk solvent often makes a significant contribution up to 4

Angstrom or so, and the Babinet principle is not applicable at this resolution (an extra Bfactor is used to down-weight the contribution in this region).

Hence, the bulk solvent correction in REFMAC is not always very accurate. Users can check the graph of <Fobs> against <Fc> to see if there is a problem. If there is, then it may be better to include the CNS bulk solvent correction in REFMAC. This is done by assigning FPART and PHIPART on the LABIN line, in which case SCALE TYPE SIMPLE should be specified. However, FPART and PHIPART will need to be recalculated whenever there is a significant change in the model.

See for example Dirk Kostrewa, CCP4 NEWSLETTER 34, September 1997.

N.B. An improved bulk solvent correction will be included in a future version of REFMAC.

Cheshire cell in AMORE

(May 1999)

> *It is mentioned in the AMoRe manual that the asymmetric unit that needs to be searched for the translation function (1-body) is the Cheshire cell while for n-body translation it is the whole cell.*
>
> *Can someone enlighten me as to what constitutes the Cheshire cell -Please*

In terms of the translation function for Molecular Replacement, the Cheshire cell is the minimum volume which will allow a unique solution. For the first molecule it will be the cell which covers a volume from one possible origin to the next, e.g. for P212121 the Cheshire cell is 0-0.5,0-0.5,0-0.5. For P1, the Cheshire cell consists of a single point (i.e. all points in the space are equivalent), which means that in P1 the translation function for 1 mol./a.u. is already solved before you start! If you are searching for the NMOLth molecule of a set, the Cheshire cell will now be the whole primitive volume. You have assigned the origin by choosing the position of the first molecule, and the other molecules will have to be positioned relative to that choice.

A table of the Cheshire cells for all 230 space groups is available in Hirshfeld (1968) Acta Cryst A24 301-311.

How to choose same test set for related data sets

(May 1999)

> *... our situation is that we have several data sets with different ligands bound to the same protein, all in the same spacegroup and essentially the same unit cell. After solving one of these structures by molecular replacement, we intend to use that model as the starting model for the rest of the datasets, (and then of course look for the ligand in each). My understanding is that we should choose the same reflections to be the test set for all datasets in order to maintain true cross-validation in the later refinements.*
>
> *My question is: how do we go about ensuring that the same reflections are chosen for the test set in all cases? We are processing the data with Mosflm/Scala, and*

doing refinement with X-PLOR, so a strategy that works either with mtz files or X-PLOR cv files would be fine.

CAD can be used to transfer a single test set between datasets, e.g.

```
cad hklin1 in.mtz hklin2 old.mtz \
    hklout out.mtz <<eof
LABI FILE 1 ALL
LABI FILE 2 E1=FreeR_flag
END
eof
```

transfers a reference free-R column contained in file old.mtz to the file for the current data set in.mtz. uniqueify with the -f option can be used to complete the test set to higher resolution if necessary. One suggestion was to create a dummy data set to really high resolution, assign free_R flags, and then use that as a reference for all subsequent data. This way you don't have to worry about what to do if you collect a higher resolution data set than your chosen reference set.

One word of caution: when you have a significant change in cell dimensions and/or your molecules roll around in the various forms (as you often get when co-crystallising/changing from room temp to cryo/etc.) then your freeR reflections won't really be free any more.

Applying NCS restraints to B factors

(June 1999)

I am refining a structure with six molecules in the asymmetric unit. The data set is only to 2.8 A resolution. I thought I was done with it, having R-free at about 23% and R-work at about 18%. But when I checked average B-factors, here is what I saw:

Chain name	Atoms	Bave	Bsdv	Bmin	Bmax
M1A	2894	33.57	18.77	2.00	100.00
M1B	2894	42.50	19.15	2.00	100.00
M2A	2894	** 58.96 **	24.67	4.87	100.00
M2B	2894	** 56.48 **	25.61	2.00	100.00
M3A	2894	34.52	21.01	2.00	100.00
M3B	2894	40.98	21.16	2.00	100.00
WAT	80	31.76	12.33	5.44	69.18

Two copies of my molecule have alarmingly high average B-factors.

Refinement has been done primarily in X-plor, with one round of refinement in CNS. Strong ncs restraints have been used throughout.

How should I check whether these high B-factors indicate a problem with chains m2a and m2b, or whether they simply have high B-factors? Would rigid-body refinement be appropriate? Or annealed omit maps? Or should I remove the ncs restraints on those two copies?

Respondents supported the use of strong NCS restraints on B factors in general, but accepted that there were some genuine cases where NCS didn't apply. Some reported similar cases where strict NCS wasn't appropriate, whereas one person felt that people

were too ready to give up on NCS and that it usually gave better results. It clearly depends on the specific case in question.

Many felt that having uniformly higher B factors for a chain did indicate a problem, probably inappropriate NCS although possibly something more severe such as mistracing. However, if the higher B values were only in loop regions, then strict NCS may be valid, with only local deviations. Other suggestions of things to check included inspecting the quality of the fit to the electron density, and comparing the refinement to one without NCS restraints.

Superpose and rotation angle calculation

(July 1999)

> *1. What is the easiest way to superimpose two structures and get the rotation and translation vectors?*
>
> *2. Is there a program that can calculate rotation angle between two domains in a protein?*

Suggestions for the first question included LSQKAB in the CCP4 suite, LSQMAN (part of the dejavu package) run via the O macro align2.omac, pdbfit routine in XtalView, and programs TOP and MAPS from Guoguang Lu (the former is now in the CCP4 suite as the program TOPP). In O, two structures can be superimposed using lsq_explicit with 3-4 equivalent residues in each molecule, and then lsq_improve is used to improve the fit for all CA's. You can then retrieve the rot-trans matrix by

```
write .LSQ_RT_foo ; ;
```

if you named your alignment "foo".

Regarding the second question, if the two domains are structurally very similar, then the situation is the same as for question one. If they are not, then the answer depends on how the angles are defined, e.g. as the angles between the principle axes of inertia.

Implementation of Data Harvesting in the CCP4 Suite

Martyn Winn

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

Introduction

The Data Harvesting paradigm pioneered by Kim Henrick at the European Bioinformatics Institute (EBI) has been under development for a couple of years, and will soon be in operation in users' labs. Background information can be found on the EBI-MSD web, and in two earlier Newsletter articles which give an overview, and a report on the September 1998 Joint CCP4/EBI Software Developers and Data Harvesting Workshop. Briefly, Data Harvesting means that software used in structure solution outputs to a deposition file details of the method used and results obtained, for example heavy atom sites used in phasing. By the time the user is ready to deposit the model coordinates, there should be a collection of files holding details of how the model was obtained. These files can be sent directly to the deposition centre, thereby by-passing much of the manual processing needed by AutoDep.

The EBI plan to be in a position to accept harvest files in Autumn 1999. Meanwhile, changes are being made to CCP4, MOSFLM and other common programs to produce harvest files. In this article, I will describe the relevant changes to CCP4.

Definition and application of datasets

Every deposition file should have associated in-house tags that identify the "Project Name" and "Dataset Name", where the Project Name is the working equivalent to what will become a PDB ID code (or in mmCIF terms the _entry.id) and the Dataset Name is the particular dataset within the project (either X-ray diffraction structure factors or NMR experimentally determined data) that is being used (_diffrn.id in mmCIF). For each program that writes out a deposition file, it is possible to specify the Project and Dataset names using the program keywords PNAME and DNAME. In principle, however, the Project and Dataset names should be considered attributes of the dataset being used, and be specified once only for that dataset. The Project and Dataset names would then be inherited from the dataset by each program in turn.

This has been implemented in CCP4 by adding information on Project and Dataset names to the header of the MTZ file. In a merged MTZ file, datasets are held as one or more data columns. In addition to the label and type attributes, each column now has an extra attribute specifying to which dataset it belongs. A list of all datasets included in the file, with the corresponding Project and Dataset names, is held separately in the MTZ header.

The code changes necessary to manipulate this information were included in CCP4 release 3.5. Ideally, dataset information should be added to the MTZ file at the beginning, e.g. in MOSFLM, but this information can be added at any time, most conveniently with the program CAD. Once the information is in the MTZ file, it can be checked by running mtzdmp which shows all the MTZ header information (go on, try it!), including the list of datasets:

```
 * Number of Datasets =   4

 * Dataset ID, protein name, dataset name:

        1 TOXD
          NATIVE
        2 TOXD
          DERIV_AU
        3 TOXD
          DERIV_MM
        4 TOXD
          DERIV_I
```

and the datasets which each column corresponds to:

```
 * Column Labels :

H K L FTOXD3 SIGFTOXD3 ANAU20 SIGANAU20 FAU20 SIGFAU20 FMM11 SIGFMM11 FI100
SIGFI100 FreeR_flag

 * Column Types :

H H H F Q D Q F Q F Q F Q I


 * Associated datasets :

   1   1   1   1   1   2   2   2   2   3   3   4   4   1
```

In CCP4, columns to be used are selected from the MTZ file by the LABIN keyword; for example, the command

```
LABIN FP=FMM11 SIGFP=SIGFMM11
```

tells the program to use the 10th and 11th columns. In addition, the program now also knows that these columns are from the 3rd dataset, with Project Name TOXD and Dataset Name DERIV_MM.

Unmerged or multi-record MTZ files are treated slightly differently. In this case, a particular column may correspond to several datasets, distinguished by different batch numbers. Datasets are therefore attached to batches rather than columns, and a pointer to the relevant dataset is held in the batch header.

As an aside, classifying MTZ columns according to dataset has other uses. Previously, it was assumed that columns existed as independent entities, but this is clearly not the case, for example F(+) and F(-) columns, or F and sigmaF columns. Some programs now use dataset information to check for certain dependencies, for example the program REINDEX may need to swap F(+) and F(-) columns and therefore needs to identify which F(+) column goes with which F(-) column.

Harvesting from CCP4 programs

The current CCP4 release (3.5) thus handles datasets, but does not as yet write out deposition files. This is currently being implemented and will be included in the next release. The CCP4 programs affected are SCALA, TRUNCATE, MLPHARE, REFMAC and RESTRAIN. Provided a Project Name and a Dataset Name are specified (either explicitly or from the MTZ file) and provided the NOHARVEST keyword is not given, these programs will automatically produce a deposition file. This file will be written to

`$HARVESTHOME/DepositFiles/<`*projectname*`>/ <`*datasetname*`>.<`*programname*`>`

The environment variable $HARVESTHOME defaults to the user's home directory, but could be changed, for example, to a group project directory.

At the end of a project, the entire contents of the directory $HARVESTHOME/DepositFiles/*<projectname>* can be sent to the deposition centre for processing. Note that, because of the file-naming scheme, only the last run of a particular program with a particular dataset will be preserved, and it is the user's responsibility to ensure that this is the authoratative version. The USECWD keyword can be used to send deposit files from speculative runs to the local directory rather than the official project directory. This keyword can also be used when the program is being run on a machine without access to the directory $HARVESTHOME, in which case the user must transfer the deposition file afterwards.

In summary, the extra keywords associated with harvesting that will be included in most programs are:

PNAME
: Project Name. In most cases, this will be inherited from the MTZ file.

DNAME
: Dataset Name. In most cases, this will be inherited from the MTZ file.

PRIVATE
: Set the directory permissions to '700', i.e. read/write/execute for the user only (otherwise '755').

USECWD
: Write the deposit file to the current directory, rather than a subdirectory of $HARVESTHOME

RSIZE
: Maximum width of a row in the deposit file (default 80).

NOHARVEST
: Do not write out a deposit file; default is to do so provided Project and Dataset names are available.

There will inevitably have to be cooperation between members of a group working on the same project to ensure that all relevant deposition files are gathered together in the same directory, but such cooperation should occur anyway. At the time of deposition, there should be a resultant saving of time, as well increased reliability in the information submitted.

Deposition files

Deposition files are written in mmCIF format. The possible contents of an mmCIF file are described in a continually-evolving dictionary of allowed data items. Harvesting requires additional data items to those in the current standard dictionary, and an extended dictionary will be distributed by CCP4.

Example of deposition files

The distributed TOXD example dataset contains 4 datasets, all assigned to the Project Name "TOXD", and having the Dataset Names "NATIVE", "DERIV_AU", "DERIV_MM" and "DERIV_I" (see above). Running mlphare to phase the native dataset produces a file /home/mdw/DepositFiles/TOXD/NATIVE.mlphare where $HARVESTHOME has defaulted to my home directory. This file starts with information on when and how the file was created:

```
data_TOXD[NATIVE]
_entry.id TOXD
_diffrn.id NATIVE
_audit.creation_date 1999-07-08T11:19:51+01:00
_software.classification phasing
_software.contact_author 'Z.Otwinowski or E.Dodson'
_software.contact_author_email 'ccp4@dl.ac.uk, ccp4@yorvic.york.ac.uk'
_software.description
'maximum likelihood heavy atom refinement & phase calculation'
_software.name mlphare
_software.version CCP4_3.5
```

This is followed by details such as the cell dimensions and symmetry information, and then by a summary of the results, for example the figures of merit for the phases obtained:

```
loop_
_phasing_MIR_shell.d_res_high
_phasing_MIR_shell.d_res_low
_phasing_MIR_shell.reflns
_phasing_MIR_shell.fom
_phasing_MIR_shell.reflns_centric
_phasing_MIR_shell.fom_centric
_phasing_MIR_shell.reflns_acentric
_phasing_MIR_shell.fom_acentric
 9.56 15.00      61   0.484      41   0.553      20   0.343
 7.01  9.56      80   0.315      36   0.423      44   0.227
 5.54  7.01     120   0.351      45   0.502      75   0.261
 4.58  5.54     186   0.338      61   0.506     125   0.256
 3.90  4.58     255   0.327      68   0.484     187   0.270
 3.40  3.90     345   0.276      86   0.417     259   0.230
 3.01  3.40     430   0.271      90   0.446     340   0.225
 2.70  3.01     536   0.287     108   0.454     428   0.245
```

The deposit files should be easily readable, but they should not be altered - they represent an authentic record of the structure solution process.

Ab Initio Phasing of Crystallographic Data

Pierre Rizkallah, James Nicholson and Robert Kehoe
Daresbury Laboratory, Warrington, Cheshire, WA4 4AD, U.K.

e-mail addresses: p.j.rizkallah@dl.ac.uk, j.nicholson@dl.ac.uk, r.c.kehoe@dl.ac.uk

Background

It has long been the aim of crystallographers to be able to determine phases from first principles. This is now routine in small molecule crystallography, where the sample is usually long-lived in the X-ray beam, and can give measurable diffraction to a resolution better than 1.2Å. Indeed, the size of molecule that can be tackled with these Direct Methods is growing. When the 'Shake and Bake' approach was added to the armoury, the size limit for structure determination went up to around 600 atoms. But the resolution limitation holds fast, and although many interesting proteins are within the target size limits, their diffraction limit is still too poor for these methods.

Low-resolution reflections are very important in determining the molecular envelope, or the shape, of a protein. They can be used to supplement phase information from elsewhere (e.g. from SIR or single wavelength anomalous scattering). They can also provide a check for models of the solvent contribution in structure refinement. Many groups, including Gerard Bricogne's at Cambridge and the groups of Urzhumtsev and Lunin (Puschino and Strasbourg), are interested in pursuing these approaches to supplement phasing studies with SIR or single wavelength studies. Contact addresses, literature references and further information for potential users may be accessed at URL: http://www.dl.ac.uk/SRS/PX/lowres/lowres.html

Technical Issues

Low-resolution terms, down to 250Å, are now easily accessible using the EMBL camera installed on station 7.2 at the SRS. Station 9.5 may also be used for this purpose. Low order reflections can of course be collected using a laboratory-based source, provided care is taken to collimate the beam, reduce the scatter and align a small beamstop. However, Station 7.2 has many features, which make it easy to access these data points:

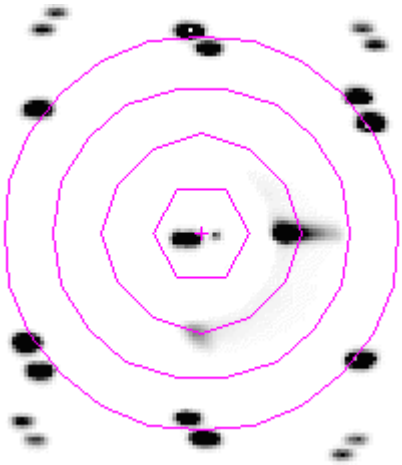
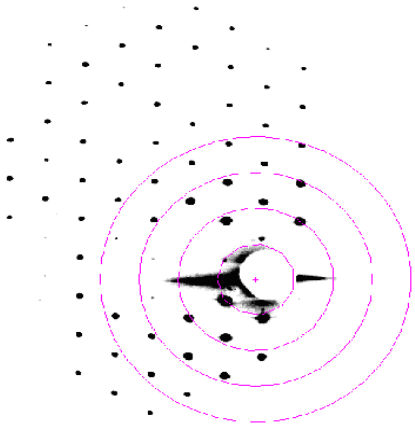
1. The SRS keeps the flux sufficiently high after narrow collimation (slits may need to be reduced to around 100 micron), so that experiment turnround would be reasonable.

2. The standard wavelength of 1.488Å on Station 7.2 gives some advantage over the popular 0.9Å elsewhere, due to the higher angle of diffraction, away from the primary beam. Longer wavelengths are also accessible, to expand the reciprocal lattice further.

3. The Hamburg PX camera, which features a long collimator, allows fine beam trimming for reduced divergence. It also has a long beamstop translation track, so

that the beamstop could be moved a significant distance away from the sample, making the obscured angle smaller.

4. The tuneability of Station 9.5 gives the added flexibility of repeating the low-resolution data collection on either of an absorption edge of one of the buffer components. Such data can be used for mask contrast measurements, and therefore 'experimental measurement' of the phases.

	
A composite malto-porin diffraction pattern, between 100 and 400 Å resolution.	Human ceruloplasmin diffraction in the range 200 to 50Å resolution.

Theoretical Methods

Over the last decade, a variety of approaches have been used to solve the phase problem directly, at low resolution. None of them have come into common usage, perhaps because they are still in need of development, but also perhaps due to the pressure on synchrotron facilities for a higher throughput, therefore ignoring the 'less important' low-resolution terms. Some of the methods are briefly outlined here:

1) The Few Atoms Method (FAM):

Lunin, Urzhumtsev et al.[1,2,3,4] have used a small number of large spheres to approximate the solvent part of the protein. These 'atoms' would be refined against 8Å data, in a fashion similar to real atoms, after substituting scattering factors of a sphere for the atomic scattering factors. This approach worked reasonably well for mask determination, but could not provide any extra phasing information up to 3Å, where atoms become interpretable. Non-crystallographic symmetry might be able to break this deadlock, but it is not universally available. The low resolution might also be too much of a challenge for the maximum likelihood methods of phase extension.

2) The Sphere of Atoms Approach (SA):

Harris[8] used a sphere of point scatterers (water molecules), placed at uniform intervals within the sphere surface, to maintain the atomicity of the simulation. The whole sphere would then be 'refined' by systematic translations along the 3 axes, while checking to exclude symmetry clashes, in a translation function analogy. Although this approach worked for some test cases, it was computationally expensive (back in 1994, using a

micro-Vax). It also had very poor discrimination towards the shape of the solvent region, which may be a cylinder or a dumbbell or any other complex shape. There was only one shape, that of a sphere, that was applied. This method too required the very low-resolution terms, and had the same limitation as above regarding phase extension to around 3.0Å.

3) Random Placement of Waters (RPW):

Subbiah[5,6,7] used a random distribution of water molecules as a starting point. He then moved each one in turn, by a random distance, in a random direction and then decided whether the move was 'Good' or 'Bad' according to a number of conditions. The total number of water molecules used was usually around 2/3 of the C-alpha atoms in the protein. This method appeared to work well, but there was an equal likelihood of the water molecules converging into solvent or protein mask. Methods were developed to distinguish between one and the other, and they seem to work well. Although this method could work with routine PX data sets, i.e. those lacking the low-resolution terms, its performance was enhanced by their presence.

4) Genetic Algorithms (GA):

This approach starts by defining spherical volume elements covering an arbitrary unit-cell, useful for solution scattering simulations (Chacón et al.[9]). See also URL http://srs.dl.ac.uk/fcis/fcis/dalai_ga . Then, each sphere is randomly assigned as occupied or unoccupied, by setting a bit in the binary string. A few permutations of the string are tested against the data for hierarchical ranking. 'Daughter' strings may be generated, by combining equally ranked strings. Other string manipulation tactics may be employed, e.g. dividing a string in half and then combining it with another half string. An elegant feature of this approach would emerge, if various starting combinations were farmed out to a computer network, to share out the load. The need to duplicate the data points would necessitate a smallish data set, well suited to the low-resolution exercise. Other work with this approach is also underway within the Bricogne team.

The Way Forward

The different approaches are sufficiently disparate in philosophy, but they are more further apart pragmatically. FAM is allied with the CCP4 package, SA is heavily dependent on it, although no optimised program was ever written. RPW evolved independently of CCP4, and further evolution is still possible. The 'Pantos' implementation of GA also evolved independently of CCP4, and could be easy to port, although that may have been done already in Cambridge.

More importantly, the principles behind these approaches have to be extended. As a first step it would be straightforward to envisage an amalgamation of the different methods into one program, brought up to date to exploit the current computer technology. Each method would be applied at a particular stage, and after an initial period of gathering experience, the optimum order would be quickly established.

Another scenario might follow this path: FAM/GA produced crude masks provide a starting envelope, which would be filled with matched spheres of point scatterers for SA applications. Then the spheres would be made smaller and their number increased to keep filling the surviving mask. This in turn would provide a starting set of waters for RPW runs, with an increasing number of waters until a reasonable fraction of the protein had been placed. Because FAM selects the protein region, there is no risk that the subsequent

SA and RWP runs would coalesce to the wrong region. At this stage, a round of ARP refinement might complete the phasing to the highest resolution of the data set. The resultant map then would be ready for all the manipulation techniques of DM, for instance, and an easy interpretation should become feasible.

References:

1. Urzhumtsev, A., Lunin, V. & Podjarny, A. (1997). Recent Advances in Phasing; Proceedings of the CCP4 Study Weekend, pp. 207-214. CCLRC Daresbury Laboratory.

2. A. G. Urzhumtsev, E. A. Vernoslova and A. D. Podjarny (1996); Approaches to Very Low Resolution Phasing of the Ribosome 50S particle from Thermus thermophilus by the Few-Atoms-Models and Molecular-Replacement Methods; Acta Cryst., D52, 1092-1097.

3. A. Urzhumtsev and A. Podjarny (1995); On the solution of the molecular-replacement problem at very low resolution: application to large complexes; Acta Cryst. D51, 888-895.

4. V. Yu Lunin, N. L. Lunina, T. E. Petrova, E. A. Vernoslova, A. G. Urzhumtsev and A. D. Podjarny (1995); On the ab initio solution of the phase problem for macromolecules at very low resolution: the few atoms model method; Acta Cryst. D51, 896-903.

5. Subbiah, S. (1991); Science, Vol 252, pp. 128-133.

6. Subbiah, S. (1993); Acta Cryst., D49, pp. 108-119.

7. David, P.R., & Subbiah, S. (1994); Acta Cryst., D50, pp. 132-138.

8. Gillian Harris, (1995); Acta Cryst., D51, pp. 695-702.

9. P. Chacón, F. Morán, J.F.Díaz, E. Pantos and J.M. Andreu, (1998); Low resolution structures of proteins in solution retrieved from X-ray scattering with a genetic algorithm Biophysical Journal, 74(6), 2760-2775.