CCP4 NEWSLETTER ON PROTEIN CRYSTALLOGRAPHY

Number 35. July 1998

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

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

Contents

Editor: Peter Briggs
Daresbury Laboratory, Daresbury, Warrington, WA4 4AD, UK

CCP4 News

Martyn Winn, Sue Bailey, Alun Ashton and Peter Briggs

Daresbury Laboratory,
Daresbury,
Warrington
WA4 4AD, U.K.
ccp4@dl.ac.uk

Staff changes

In October, we said goodbye to Adam Ralph who moved to the Institute of Virology, Glasgow. We wish him well, and thank him for his many contributions to CCP4.

In March, Peter Briggs joined the Daresbury staff. Peter moved from the University of East Anglia, where he obtained a PhD in Theoretical Physics.

Release 3.4

Version 3.4 of the CCP4 suite was released on May 5th. In addition to a number of bug fixes and enhancements to existing programs, this release included the following new programs:

- DMMULTI: multi-xtal density modification package (Kevin Cowtan)
- RANTAN: Direct Method module for the determination of heavy atom positions in a macro-molecule structure or to determine a small molecule structure (Yao Jia-xing).
- SFCHECK: Structure factor checking (Alexei Vagin).
- SFTOOLS: Reflection data file utility program (Bart Hazes).

There are also new versions of PROCHECK (3.5), REFMAC (3.3), RESTRAIN (4.6.9) and SCALA (2.4.2). There is a new script "duptree" for setting up multiple installations from a common source directory, and a preliminary installation procedure for the X-windows programs (at present, only for IRIX, OSF1 and LINUX). There has been significant progress on the linux/g77 installation, and most programs should now compile and run under g77.

Details of all the changes can be found in the CHANGES file in the top-level directory ($CCP4), and in $CCP4/html/CHANGESinV3_4.html.

On May 27th, there was a minor release 3.4.1 which included a few minor bug fixes to version 3.4. No distinction between 3.4 and 3.4.1 is made in program headers, etc. Subsequent problems will be reported on the Problems Page as usual.

Workshops

The 1998 CCP4 Study Weekend on databases and their use by crystallographers was held in Reading on the 9th/10th of January 1998. The meeting aimed to introduce practising macromolecular crystallographers to the wide range of database resources available. The areas covered included sequence alignment, properties of heavy atoms, tools for characterising non-bonded interactions, and tools for analysing models of macromolecules. In between the main talks, there was a demonstration by Liz Potterton of the new Graphical User Interface to the CCP4 Suite, and a demonstration of their latest products by the Cambridge Crystallographic Data Centre. A list of URLs related to the talks can be found at http://www.cryst.bbk.ac.uk/~ubcg09j/ccp4urls_1.html. Next year's Study Weekend will be at Sheffield and will be on Data Collection and Processing: more details in due course.

CCP4 and CCLRC Daresbury Laboratory hosted a workshop on Multiwavelength Anomalous Dispersion on 23 - 27 June. A combination of talks and hands-on practicals instructed participants on the planning and executing of a multiwavelength anomalous dispersion experiment up to the stage of obtaining an interpretable electron density map.

The next CCP4 workshop will be in Prague at the 18th European Crystallographic Meeting, and will be held prior to the main conference on Saturday August 15th. Introductory talks on the philosophy and use of the CCP4 Program Suite will be given in the morning, followed by practical sessions in the afternoon.

A joint CCP4-EBI workshop is to be held at the EBI (European Bioinformatics Institute, Hinxton, Cambridge) on 16th to 19th September, and is primarily aimed at software developers. There will be an introduction to the concept of automatic data harvesting during structure determination for deposition information. It is hoped to define and then accumulate the required information by "harvesting" output from existing packages. There will also be instruction on the use of the CCP4 libraries. These libraries are freely available, and it is hoped that their use by non-CCP4 software will aid standardisation of file formats and program usage. The workshop will also include an introduction to the new CCP4 working coordinate file format, CCIF, see below.

Future changes to the CCP4 working coordinate format

The macromolecular Crystallographic Information File (mmCIF) format was developed by a working group of the IUCr formed in 1990. It represents an extension of the CIF format used by small molecule crystallographers, and which is used for automatic submission to Acta Crystallographica C. mmCIF files are text files with a flexible format based around either <data_name> <data_value> pairs or a loop structure (works like a table). In particular, a wide variety of data items are supported (as defined in the mmCIF dictionary), and character data values may be lengthy and descriptive. This alleviates many of the restrictions of the traditional PDB format.

In view of the likely increasing importance of the mmCIF format, CCP4 intend to move an mmCIF-like format as the working format for coordinate data. Conversion programs will be provided to change between this working format and PDB, but the programs will no longer work directly with PDB files. Initially, we intend to use only a small subset of the full mmCIF format, which will mirror the current PDB format. Coordinate data files should not look too dissimilar from PDB files; in particular, the bulk of the file will remain as columns of atom data. As we gain experience with the format, and users become comfortable with it, we will probably increase the subset of mmCIF data items which can be used, thereby using making more use of the power of mmCIF. This extensibility is indeed one of the advantages of mmCIF over PDB.

The first version of the CCP4 suite to use the mmCIF-like format may appear towards the end of 1998. Releases before then will *not* include the new format. Precise details are still evolving, so watch this space. Background information on the full mmCIF format can be found at http://www.iucr.ac.uk/iucr-top/cif/mmcif/ndb/index.html. Details of the (still evolving!) CCP4 implementation can be found here.

Other developments

The Graphical User Interface to the CCP4 Suite is now reaching its final form, and it is hoped to make a public release later this year. Details can be found in Liz Potterton's article.

As mentioned above, the EBI Macromolecular Structure Database group are working on Data Harvesting, which involves the automatic collection of deposition information from programs used during structure determination (see here). CCP4 programs are currently being converted to output the information needed by the harvesting software.

Looking further into the future, we are investigating the new XML (Extensible Markup Language) as a way of handling complex information, e.g. more intelligent program documentation or program log files. An exmple of what can be done is provided by the Chemical Markup Language (CML).

And finally ... We have now settled on an official CCP4 logo. Many many thanks to all those who sent in suggestions, some of which were quite artistic. Our final choice is somewhat simpler, but we hope distinctive.

News of Progress on the CCP4 Graphical User Interface

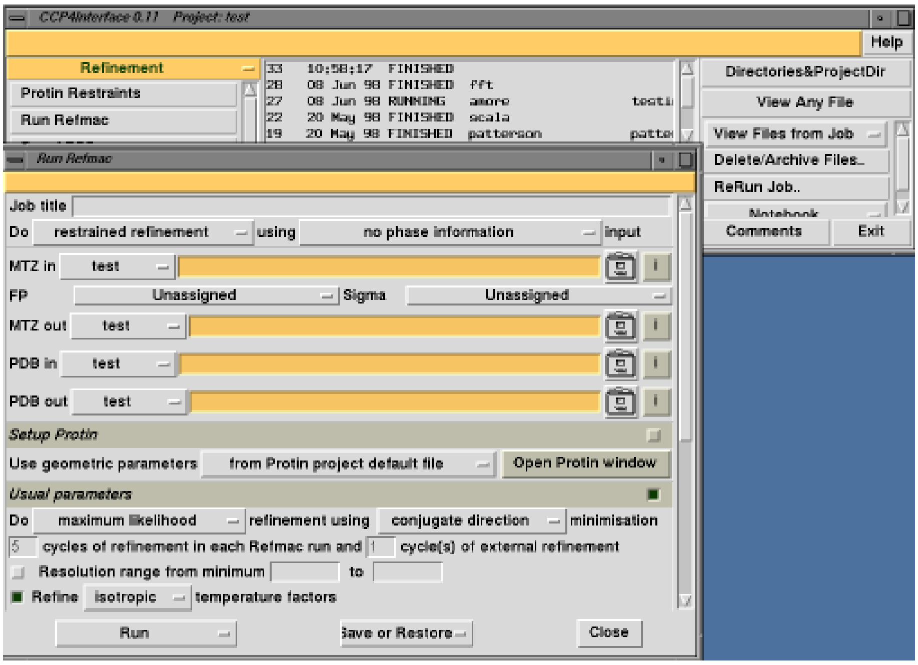
Liz Potterton (lizp@yorvic.york.ac.uk)
June 1998

What will it be called?

For a while now I have been calling the CCP4 graphical user interface **CCP4I** where the 'I' stands for 'Interface' - no one has objected so I think that name will stick.

What does it look like?

The Refmac interface in front of the main window:



```
33   10:58:17  FINISHED
28   08 Jun 98 FINISHED  fft
27   08 Jun 98 RUNNING   amore                 testi
22   20 May 98 FINISHED  scala
19   20 May 98 FINISHED  patterson             patte
```

Job title
Do restrained refinement using no phase information input
MTZ in test
FP Unassigned Sigma Unassigned
MTZ out test
PDB in test
PDB out test

Setup Protin
Use geometric parameters from Protin project default file

Usual parameters
Do maximum likelihood refinement using conjugate direction minimisation
5 cycles of refinement in each Refmac run and 1 cycle(s) of external refinement
□ Resolution range from minimum to
■ Refine isotropic temperature factors

When will we get it?

I would like to get out a **beta-test version** which will be generally available some time in June/July. We will publicise the release and how to get it on the CCP4 bulletin board and web pages.

The beta release should have all the basic functionality that is intended for the first full release but will not have interfaces to all the programs which we think are really necessary. Peter Briggs (the new CCP4 employee in Daresbury) and I will work on writing interfaces to some more popular programs before a full release with **CCP4 Version 4.0** at the end of 1998.

A few people in York and elsewhere are currently using a 'pre-release' version of CCP4I and have sent me lists of bugs/requests that will keep me busy for a a few days (understatement (-;). I'm also working on an interface to Amore which I would like to get into the beta-release because it would be very useful to get your feedback on that.

What will CCP4I do?

The main function of the first version of CCP4I will be to make running the popular programs easier. I've written about the basic design for a previous newsletter.

The access to programs is organised by task where a task usually corresponds to one main program but may also use other helper programs. For example the interface to FFT may also run the Mapmask program to extend the map and other utilities to convert the map file format to something appropriate for a graphics program.

CCP4I also has a simple project management system which keeps a database (purists will object that it is not a 'proper' database) of the jobs you have run and the input and output files and will automatically store the parameters you set so that it is easy to:

keep track of what you have done
re-run tasks - with the option to review and change the parameters used in the previous run
archive (i.e. save to a 'safe' directory) important files
clean-up after failed jobs by deleting all output files
use a simple on-line notebook to record your comments on a job

Which Programs are Interfaced?

Currently there are interfaces to the following tasks and programs:

Utilities

Convert HKL files to MTZ
Convert MTZ to HKL files
Merge MTZ files (Cad)

Data Reduction

Convert to multi-record MTZ files (Rotaprep)
Sort & merge MTZ files (Sortmtz)
Scale experimental intensities (Scala)
Convert Intensities to Fs (Truncate)

Isomorphous Replacement

Scale datasets (Scaleit & Fhscal)
Direct methods (Rantan)
Generate Patterson map (FFT, Vectors, Peakmax & NPO)
Heavy atom refinement (Mlphare)

Molecular Replacement

Amore (in progress)

Density Improvement

DM
Solomon
FFT

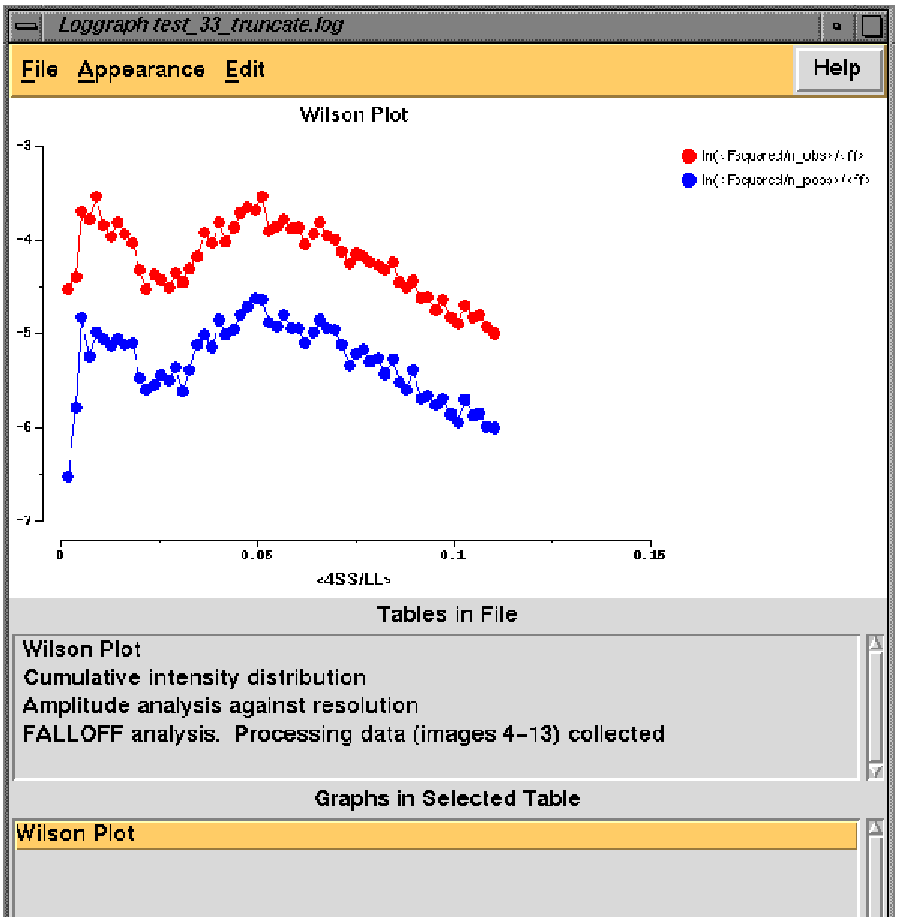
Refinement

Setup restraints (Protin)
Refinement (Refmac)
ARPP
FFT

A few more programs, particularly utility programs to handle map and MTZ files, will be added to this list before the first full release but we would prefer to make available what we have already and extend the list of supported programs in a follow-up release. Suggestions for what would be useful are welcome.

Loggraph

A new version of the program xloggraph which displays the graphs from CCP4 log files has been implemented using the Tcl/Tk graph plotting extension BLT. The new version is just called Loggraph and is based on the program written by Darren Spruce at the ESRF in Grenoble. Loggraph will read either CCP4 log files or a file with tabulated data and has options to change the appearance, title and axis labels on a plot and will produce colour or black and white Postscript files.



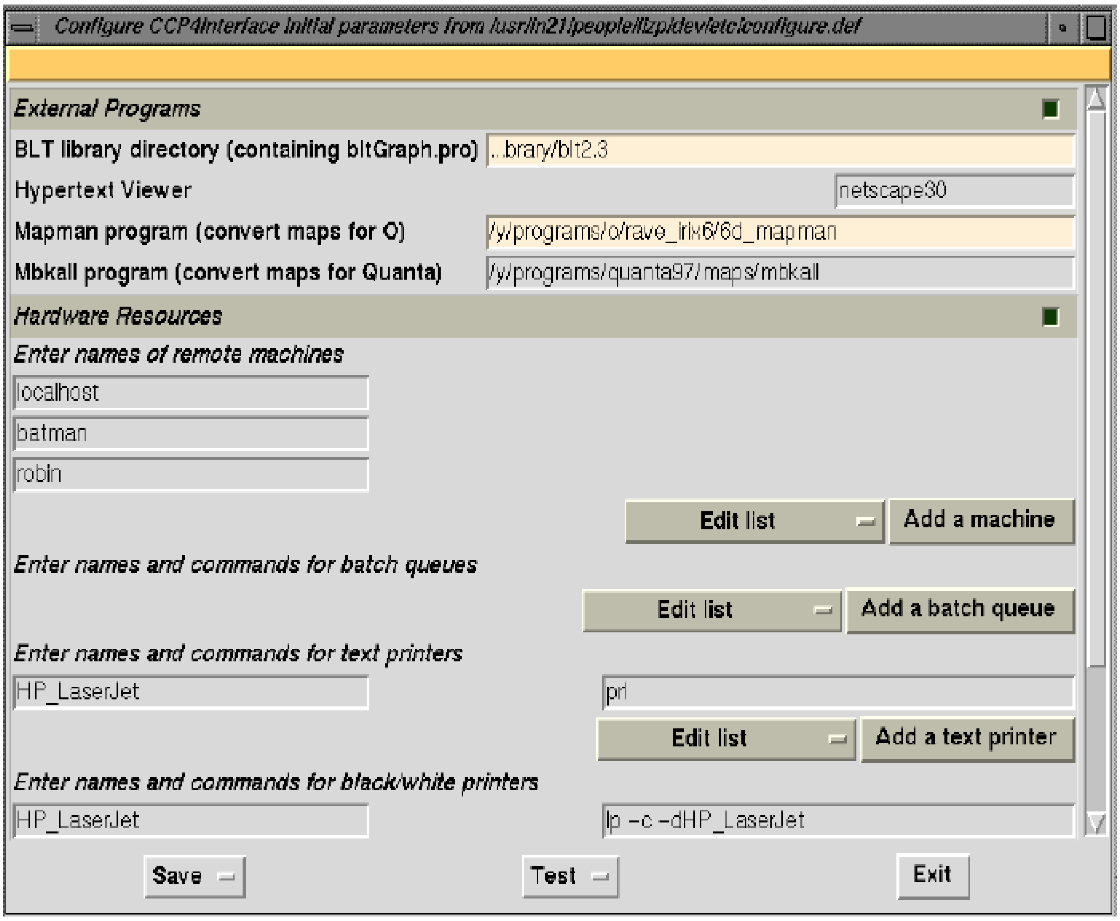
Wilson Plot
Cumulative intensity distribution
Amplitude analysis against resolution
FALLOFF analysis. Processing data (images 4-13) collected

Wilson Plot

Peter Briggs is going to add more functionality to to the Loggraph program - for example integrating it in with the analysis tools in Bart Hazes Sftools program could be very useful.

Installation

CCP4I requires Tcl/Tk and the graph plotting extension BLT. These are freely available and easy to install. CCP4I makes use of several non-CCP4 utilities, for example it uses Netscape (or any other browser) to view help text and will send jobs to other machines on your local network. To do this your local CCP4I installation must be configured appropriately. To simplify this there is now a graphical interface to the configure file (my thanks/apologies to early users of the system who have had to manage without this facility!).



Configure CCP4Interface Initial parameters from /usr/ln21/people/lizp/dev/etc/configure.def

External Programs

BLT library directory (containing bltGraph.pro) ..brary/blt2.3

Hypertext Viewer netscape30

Mapman program (convert maps for O) /y/programs/o/rave_irix6/6d_mapman

Mbkall program (convert maps for Quanta) /y/programs/quanta97/maps/mbkall

Hardware Resources

Enter names of remote machines

localhost

batman

robin

Enter names and commands for batch queues

Enter names and commands for text printers

HP_LaserJet prl

Enter names and commands for black/white printers

HP_LaserJet lp -c -dHP_LaserJet

What will be in future releases of CCP4I

Suggestions welcome but we will be particularly looking to implement tools to aid reviewing and analysing results so that crystallographers can quickly see how their structure solution is progressing and make the appropriate decisions about what to do next.

What colour will it be?

This is the tough one. We have a nice, distinctive gold and green colour scheme at the moment but it does not go well with the blue that has been chosen for the logo.

A Program to Detwin Merohedrally Twinned Data

Helena O. Taylor and Andrew G.W. Leslie
MRC Laboratory of Molecular Biology, Hills Rd., Cambridge CB2 2QH, U.K.

There have been a number of papers in the literature recently describing how to effectively "detwin" X-ray data collected from merohedrally twinned crystals (see Yeates, 1997, for a recent review). One of us (H.O.T.) has recently being attempting to solve a protein-DNA complex where the data show the classic signs of merohedral twinning (in the cumulative intensity distribution (N(z) test) tabulated in TRUNCATE). However, no program was available in the CCP4 suite to allow detwinning of the data.

A new program DETWIN has been written, which takes as input the MTZ file contained merged intensities written by SCALA, and writes out a new MTZ file, in the same format, in which the data have been detwinned. This file is then input to TRUNCATE in the usual way.

The detwinning is formally only possible for a twin fraction other than 0.5, and errors in the twin fraction will result in large errors in the detwinned intensities for twin fractions above 0.45. The twin fraction can be estimated from the N(z) statistics for the original data, or from DATAMAN (due to Gerard Kleywegt), or by running the program several times with different twin fractions and choosing the twin fraction that gives the best N(z) distribution or the smallest correlation between the intensities of reflections related by the twin operator (after an idea of Richard Henderson).

The program is very simple to run, but is not yet in its final form. In particular, it does not correct data with an anomalous signal. However, anyone wishing to try using it should send an E-mail to:

andrew@mrc-lmb.cam.ac.uk

Once the code is in a more satisfactory form it will be released as part of the CCP4 package.

A more complete description of experiences using the program will be given in a later Newsletter.

References

Yeates, T.O. in Methods in Enzymology, Vol 276, 344-358, 1997.

NEWS FROM THE UPPSALA SOFTWARE FACTORY - 9

Déjà-vu all over again

Gerard J. Kleywegt
Department of Molecular Biology
Biomedical Centre, Uppsala University
Uppsala - Sweden

While building a protein model into electron density, one often comes across features of the model that make one wonder: "*(where) have I seen this before?* ". At the level of the overall fold, there is plenty of software available nowadays that can help answer this question (DALI, DEJAVU, TOP, *etc.*). But when it comes to recognising smaller "motifs" (*e.g.* , a set of residues involved in binding a ligand or metal ion, or with seemingly "unusual" side chain-side chain interactions), answering the question "*has this been observed in any other protein structure?* " is not as simple.

At the 1995 CCP4 meeting, Peter Artymiuk described a program called ASSAM **[1] [2]** that could recognise spatial arrangements of side chains by comparing them to a database of protein structures. This provided the inspiration for the SPASM package **[3] [4] [5]** that contains programs for the recognition of arbitrary patterns or motifs in protein structures, interfaced with O **[6]** and other programs.

SPASM

SPASM is a program that can be used to recognise user-defined motifs in a database of protein structures (derived from the PDB). The user merely has to carve out those residues that (s)he is interested in (e.g., catalytic residues, a strange loop, ligand-binding residues, a weird Met-Trp interaction, a helix-turn-helix motif, etc. etc.; whatever is selected will be referred to as a "motif" from now on) and put them into a small PDB file. The program will read this file as well as its database, will prompt for values for a few parameters (the default values will do in most cases), and will subsequently find all instances of the motif in the proteins that are in the database. (The nitty-gritty and some of the bells and whistles are discussed in **[5]** and **[6]** .)

Besides simply listing the "hits", SPASM can also generate a macro file for use with O which, when executed, will automatically read the hits, apply the rotation-translation operator that superimposes the hits with the user's motif, and draw the hits. Thus, within five to ten minutes one obtains a visual answer to the original question: "*(where) has this motif been observed previously?* ".

If you find hits that display similarity to your own protein that extend beyond the matched motif (e.g., similar fold or domain), global superpositioning of the hits and your own model can be carried out by LSQMAN. An input file for LSQMAN that does this can be generated by SPASM as well, making this a very rapid process. Finally, an interface exists to the SBIN package of programs **[4] [7]** , that can be used to analyse superimposed structures to find similarities in their sequences. These, in turn, can be used to attempt "database mining" in sequence databases such as SWISS-PROT **[8]** , in the hope of identifying other proteins that might have the same fold, or share a common domain.

RIGOR

RIGOR is another program in the SPASM package that does in essence the opposite of SPASM. Where SPASM compares a user-defined motif to a database of protein structures, RIGOR looks for instances of a large number of predefined motifs in the user's model. Of course, the utility of this approach depends critically on the quality of the database. At present, it contains a few hand-crafted motifs, but the overwhelming majority has been generated automatically. These automatically generated motifs were extracted from proteins in the SPASM database, and consist mostly of sets of residues whose side chains cluster in space, or are all in close proximity to a hetero-entity. Just like SPASM, RIGOR is interfaced to O allowing for rapid visualisation of the results. Users are welcome to submit additional motifs for inclusion in future releases of the RIGOR database. Eventually, I hope to develop software that takes a more intelligent approach to detecting motifs that recur in several or many structures.

APPLICATIONS

Obviously, the SPASM package can be tremendously useful in the analysis of newly determined protein structures. The programs help crystallographers to make the most of their models, prior to publication and deposition. After all, nobody likes to see papers in which professional database scrutinisers (for want of a better word) announce that they have found an unexpected similarity between one's own protein (the structure determination of which may have taken you years) and some other protein that had been in the database for years.

In addition, SPASM can be used in comparative structural analysis, where one will typically be interested in finding all proteins that contain a certain arrangement of helices, strands, turns, and loops, or in all proteins that contain a certain constellation of residues or side chains. Other potential applications lie in the areas of protein design and engineering, and prediction of structure and function.

AVAILABILITY

The SPASM package contains the programs SPASM and RIGOR, as well as two programs to generate private databases for use with these programs (*e.g.* , with in-house structures that have not yet been released by the PDB). SPASM and friends (including databases and manuals) are available free of charge to academic users from *ftp://alpha2.bmc.uu.se/pub/gerard/spasm/* . Commercial users may contact GJK for more information (*gerard@xray.bmc.uu.se*). For more information about **O** , contact Alwyn Jones (*alwyn@xray.bmc.uu.se*). The O WWW site is at *http://imsb.au.dk/~mok/o/* , and the Uppsala Software Factory can be found at *http://alpha2.bmc.uu.se/usf/* .

REFERENCES

[1] Artymiuk, P.J., Poirrette, A.R., Grindley, H.M., Rice, D.W. and Willett, P. (1994). A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.* **243** , 327-344.

[2] Artymiuk, P.J., Poirrette, A.R., Rice, D.W. and Willett, P. (1995). Comparison of protein folds and sidechain clusters using algorithms from graph theory. *In* "From First Map to Final Model" (Bailey, S., Hubbard, R. and Waller, D.A., Eds.), pp. 71-81, SERC Daresbury Laboratory, Daresbury, U.K.

[3] Kleywegt, G.J. and Jones, T.A. (1998). Databases in protein crystallography. *Acta Cryst.* **D54** , in press. (A preprint of this paper is available at URL: http://alpha2.bmc.uu.se/~gerard/papers/databases.html.)

[4] Kleywegt, G.J. (1998). Recognition of spatial motifs in protein structures. Submitted.
[5] The manuals for the SPASM programs are available at URL:
http://alpha2.bmc.uu.se/usf/s pasm.html.
[6] Jones, T.A., Zou, J.Y., Cowan, S.W. and Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr.* **A47** , 110-119.
[7] The manuals for the SBIN programs are available at URL:
http://alpha2.bmc.uu.se/usf/sb in.html.
[8] Bairoch, A. and Apweiler, R. (1997). The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucl. Acids Res.* **25** , 31-36.

CCP4 and Data Harvesting

K. Henrick

European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD

henrick@ebi.ac.uk

Introduction

The Macromolecular Structure Database group at the European Bioinformatics Institute was set up in 1996. The group members are Geoff Barton, Peter Keller, John Ionides and Kim Henrick with Janet Thornton as external advisor and a further member will join in October 1998. Its principal aim is to become the European partner in the deposition, archiving, and presentation of 3D services of macromolecular structural data, together with the Protein Data Bank at Brookhaven Laboratory and with the Nucleic Acid Database at Rutgers University. To achieve this aim we are working towards the design and production of a single, high quality, global structural data collection. We intended to generate a much richer database than that possible with the current PDB entries. This will require even more data to be deposited, and may make higher demands on the depositor, with respect to the quality and completeness of information required.

As part of the process of developing methods to make the process of data archiving relatively painless we have prototyped a data harvesting method with some of the CCP4 programs.

The proposed method takes the emphasis away from a web interface back to the research worker. The software used at each step in the process of the determination of a macromolecular structure by X-ray diffraction methods is to be modified such that a summary deposition file will be produced.

The research worker will be responsible for the management of these files, in the sense of making sure that they are in the correct file system location at the time of deposition. (The format of these files will be mmCIF, based on an extended mmCIF dictionary, but the researcher will not need to examine or edit them.) Deposition will then involve the transfer of the collection of files to the deposition site, where the data harvesting software will extract as much of the required information as possible. Missing information will have to be supplied by the depositor, but it is hoped that this will be kept to a minimum.

Harvest Files

The set of files will contain the same information that will be required for publication and for writing a thesis. The requested information will be used to characterise the reflection dataset, the refinement used and provide confidence levels for aspects of the final structure model. Information that is normally only derived from coordinates will not be requested in the deposition process. The deposition centre will use wherever possible the same procedures to derive the information that is currently calculated from coordinates by research workers.

There are three classes of data items.

(i) data that is currently contained in the output of the programs used. It is advantageous to generate this data in a deposition-ready form at the time the programs are run.

(ii) the final coordinates and structure factors which are currently stored and used in a variety of formats.

(iii) annotation and structure description files, for these items an EBI supplied tool will be used to allow a research worker to compile in-house most of the requested information to a harvest file.

In outline a program will write a file containing the following:

data_*phosphate_binding_protein[A197C_chromophore]*
_entry.id *phosphate_binding_protein*
_entry.data_set_id *A197C_chromophore*
_audit.creation_DayTime *'Wed Jul 16 12:58:55 1997'*

Where **entry.id** corresponds to the users ProjectID (protein) and **entry.data_set_id** corresponds to the users DataSetID

Every project has at least 1 dataset, but may have more than 1, as for example in the case of MIR structure solution. A ProjectID will be a research workers identifier for a data set that will be expected to be become a submitted Entry, i.e. a mutant is a new ProjectID, however, a heavy atom derivative will not be a new ProjectID but a different DataSetID.

The harvest snapshot files generated by running a particular CCP4 program will then contain the data items as mmCIF tags and value pairs that the particular program is capable of automatically generating. The identification of data items within a particular programs is a trivial task. For most CCP4 programs this closely corresponds to the data contained in the xloggraph tables. *[Trying to parse log files to get the requested deposition data items has been found to be very difficult as the output of the programs has varied and in some cases this is difficult to track from version to version and to understand the input parameters and the path used through the options available for each program.]*

To generate successfully a set of deposition files during the course of a structure determination, some minimal discipline will be required from the researcher. Essentially, this involves deciding on some name or code to identify the project, and sticking to it. This is necessary, to allow the data harvesting software to identify correctly and merge data from a number of different files. Where multiple datasets are involved, the user will also be required to identify each dataset in a consistent way.

The prototype CCP4 mechanism involves automatically placing the output from a program run into the directory (unix here),

```
$HOME/Deposition_Files/Project_Identifier/datasetname.program_name_function
```

In the above, "function" will be chosen by the program input keywords when the particular program is multifunctional. For example REFMAC can be used in several different modes, and MOSFLM can be run for selected sets of images for a particular dataset. No versions of these deposition files are to be kept and each output harvest file will be overwritten for each run of a program with a particular ProjectName, DataSetName, ProgramName and ProgramFunction.

The user requirement needed to produce these deposit files, will be to add two extra options to existing software. In the context of CCP4 programs, this will mean adding two extra keywords, i.e.

```
PROJECT  my_protein
DATASET  native
```

The presence of these keywords in command input will cause the program to generate the deposition files. The individual research worker is ultimately responsible for the management of their own files. The proposed system of consistently using keywords to flag every project and data set will require some data management organisation and discipline to be followed, however, if used this will make life easier for everyone concerned.

The software should then handle the details.

The **PROJECT** and **DATASET** instructions are required to differentiate between different cell dimensions, resolutions, etc. for different data collections, in particular where derivative structure factors and heavy atom sites and phasing information are deposited.

The additional keyword, **USECWD**, can be used to place the deposit file in the current working directory. This is required where a particular program is run on a machine where the research worker doesn't have a $HOME directory, i.e. at a synchrotron site, or an in-house generator/image plate system, or for example where part of the structure determination process is run on a different machine either within the same institution or between different institutions. The research worker will be responsible for collecting all the files into the one harvesting archive directory ready for deposition. This keyword can also be used for tests, and when a successful run is made the output harvest file can then be simply moved to the DepositFiles sub-directory.

At each stage in a protein structure determination the program steps are often re-run and the stages are repeated in an iterative cycle. Data sets are commonly discarded as experience is gained and better crystals that give better data are grown. Therefore the above two program instructions are only required for the final run of each program. However, it is not always known at the time of a particular run that it will be the final one, and in the case of a preliminary publication the results are not definitive. Therefore the instructions could be left in for every run of the programs, and will overwrite existing files that correspond to a particular program. (The accumulation of versions of the deposit files will be unmanagable.)

The CCP4 implimentation will involve a change to the MTZ library. The MTZ header will be modified to hold labels for each ProjectName and DataSetName and additional pointers for each column of information that will indicate which dataset the item belongs to. Then entry points to an MTZ file, such as MOSFLM (or ROTAPREP or F2MTZ), will demand values for the keywords **PROJECT** and **DATASET** which will be held in the header. Later procedures will also have mandatory values for these items but will take the values from the MTZ header where appropriate. The programs MTZUTILS, MTZDUMP and CAD will also be modified to manage the merging of reflection datasets. It is possible that the required items will also be placed in the MAP file header.

Sample Outputs

The CCP4 programs, MOSFLM, SCALA, TRUNCATE, MLPHARE, REFMAC and RESTRAIN have been modified to output harvest files and sample output files are available (these files are preliminary as the new mmCIF data names used here are under review):

http://www.ccp4.ac.uk/newsletters/newsletter35/mosflm.html
http://www.ccp4.ac.uk/newsletters/newsletter35/scala.html
http://www.ccp4.ac.uk/newsletters/newsletter35/truncate.html
http://www.ccp4.ac.uk/newsletters/newsletter35/mlphare.html
http://www.ccp4.ac.uk/newsletters/newsletter35/refmac.html
http://www.ccp4.ac.uk/newsletters/newsletter35/restrain.html

Timetable for Adopting Harvest Files

The EBI have created an Oracle database to represent a PDB entry and the mechanism for managing deposition. This system will be tested extensively to accept a variety of deposition files and mechanism. Before CCP4 or any other protein structure determination software can be modified and distributed, mechanisms that will be acceptable to research workers have to be finalised and tested. Towards this end a joint CCP4-EBI workshop will be held at the EBI in September (1998) to discuss such methods and to decide on the data to be harvested.

There has been support from other packages, including CNS, O, TNT, SHARP and SOLVE. It is hoped that within 2 years most new depositions to the central archive sites will be sending information automatically collected into a series of these snapshot files from any of the main crystallographic software suites.

The EBI harvest suggestion for deposition files labelled with the project and data set names is close to the proposed new image CIF header ideas and it is hoped to integrate the two approaches [see details on Crystallographic Binary File (CBF) and header information on The draft definitions]]

Potential use of MODELLER for Molecular Replacement Search Models

Max Paoli (max@cryst.bioc.cam.ac.uk)

The program Modeller has been employed to build homology models to be used in a molecular replacement search.

The serum glycoprotein haemopexin is made up by two domains connected by a flexible linker region. The N- and C-terminal domains contain internal repeats with strong sequence homology patterns which suggest that they have the same basic fold. Structure determination of the C-terminal domain revealed a beta propeller fold (Faber et al. 1995, Structure 3, 551-559) with four "blades" or modular beta sheets, corresponding to the four internal sequence repeats found in this domain, as well as in the other domain.

Structural studies on the whole molecule of haemopexin aimed at solving the structure by molecular replacemnt, using the model of the C-terminal domain to search for itself and for its mate, the N-terminal domain. While the first part of the search was straightforward, finding a solution for the N-terminal domain proved to be a very difficult task. A solution was found only using a chimeric construct made up of parts of structure of the C-terminal domain combined with parts of structure of the beta propeller domain of collagenase.

With a sequence identity of 23%, the N- and C-terminal domains of haemopexin show a lot of insertions and deletions of various lengths located between the modules or beta sheets. Curiously, some parts of the beta propeller structure of collagenase has a sequence which resembles more the haemopexin N-terminal domain than the haemopexin C-terminal domain does.

In order to check the possibility of using homology models of the N-terminal domain as search models for the molecular replacement calculations, the sequence of the N-terminal domain was threaded onto the structure of the C-terminal domain. Then five homology models were built using the program Modeller and then fed into the suite of programs AMORE-CCP4. These models were used without truncations, as given in the output of Modeller. The results show that the homology models give better and comparable solutions to the one obtained using the chimeric construct mentioned above.

These results are very encouraging indeed and prove that Modeller can produce homology models which are valid starting points for structure solution by molecular replacement searches. The home page for the program Modeller is http://guitar.rockefeller.edu/modeller/modeller.html.

Max Paoli

MOSFLM - Recent changes and future developments.

Andrew G. W. Leslie

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.

MOSFLM has undergone a number of improvements since the last CCP4 Newsletter described the implementation of the STRATEGY option in 1996. The algorithms employed in the STRATEGY option, which facilitates the design of an efficient data collection scheme, have been revised to avoid occasional problems with higher symmetry space groups that were present in the original implementation. The option is still most valuable when collecting data at a synchrotron, and particularly if multiple crystals are required to collect a complete data set. It proved invaluable when collecting data from crystals of Hepatitis B core particle at the ESRF, where a total of twenty crystals were used. Irrespective of how much data had already been collected, it only one to two minutes to determine what rotation range should be used for a new crystal in order to maximise data completeness. This includes auto-indexing the new image and running the strategy option. Contrary to a recent report in the literature (Ravelli, R.B.G., Sweet, R.M., Skinner, J.M., Duisenberg, A.J.M. and Kroon, J., J. Appl. Cryst., 30, 551-554, 1997) the facility to complete a partial dataset has always been part of the STRATEGY option.

The program can now process images from a wide variety of image plate and CCD based detectors. Recent additions have been the Mar 345 detector (both compressed and uncompressed formats), the Mar CCD detector (in use at the ESRF) and the ADSC Quantum 4 detector (soon to be installed on station 9.6 at Daresbury, and already in use at SSRL, CHESS, ALS and APS). These complement the existing formats for Mar, Raxis II, Raxis IV and Mac Science DIP scanners. The flexibility of the keyworded input allows a wide range of possible image formats to be accommodated without making any changes to the code (the SCANNER keyword). A previous limitation that the two-theta axis (for a swung-out detector) had to be parallel to the fast changing direction in the image has been removed, in order to deal with images collected at non-zero two-theta angles at the D2AM beamline at the ESRF and with the ADSC Quantum 4 detector.

A new feature introduced at the request of users is the ability to write a "savefile" containing the current processing parameters. This is designed to make it easier to set up a background processing job having auto-indexed and integrated a number of images interactively. The file contains keywords defining the current values of the processing parameters, including, for example, the name of the file containing the orientation matrix. It should only be necessary to add the appropriate PROCESS keyword defining the images to be processed when submitting a background job. The "savefile" can be written at any stage by using the SAVE keyword, or the option is provided by default when "Exiting" from an interactive session.

Several limitations present in previous versions have now been removed. These include the previous maximum reflection index of 255 (there is now no limit), and the maximum reflection width of ten images for partially recorded reflections (this is now 100 images). A significant number of minor bugs have also been corrected, and the output to the Xdl_view windows (when running interactively) has been improved. Non-standard FORTRAN code has been removed allowing the program to be compiled under Linux (at least on some systems), although some users have experienced problems when running the Linux version which are not yet understood.

CCP4 has generously provided support for further development of the MOSFLM software. The current priority is to implement the FFT based auto-indexing procedure developed in Michael Rossmann's group (Steller, I., Bolotovsky, R. and Rossmann, M.G., J. Appl. Cryst. 30, 1036-1040, 1998). This code is in the public domain, and is currently in use at CHESS and should provide a more robust auto-indexing procedure. It is hoped that this option will be available within the next six months, and should greatly improve the general usefulness of the program. Other plans include allowing the use of partials extending over several images in the post-refinement. At present, this is restricted to reflections present on two images, which can cause problems when processing data from crystals with large unit cell parameters (necessitating the use of small oscillation angles) and with a large mosaic spread (equal to or greater than the width of two oscillations). Finally, it would be an advantage to use reflection spot positions, in addition to the post-refinement data, in the refinement of crystal cell and orientation parameters. This is particularly true when processing low resolution data (less than 3.5E resolution) where post-refinement does not lead to very precise values.

The source code and executables for SGI and Dec Alpha machines for the latest version (5.51) are available by anonymous ftp:

```
ftp ftp.mrc-lmb.cam.ac.uk
cd pub/mosflm
```

The README file in this directory contains instructions on installation and a listing of known bugs with workarounds where possible.

Maximum Entropy and CCP4

Chris Gilmore
Theoretical Crystallography Group
University of Glasgow

The maximum entropy (ME) phasing method has still not entered mainstream protein crystallography despite its successes [1-6]. In part this is undoubtedly due to a lack of a general-purpose computer program freely distributed within the protein crystallography community. CCP4 is therefore funding a one year programming post in which the MICE computer program developed by Chris Gilmore, Gerard Bricogne and Charles Carter will be adapted by Chris Gilmore and his group at Glasgow to become part of theCCP4 suite. It is proposed to include:

- Solvent flattening.
- Non-crystallograpic symmetry
- Partial structures.
- Low resolution envelopes.

How will it be used?

It will be possible to compute a maximum entropy centroid map at any time where Fourier coefficients and some measure of their reliability are available are available. The user selects the reflections using acceptance criteria based on the agreement between observed and calculated structure factors and figures of merit from phasing procedures, and these are used as the starting point of a constrained entropy maximisation in which the amplitudes, their associated phases and any other crystallographic knowledge as listed above are used as constraints. In cases where traditional maps cannot quite be interpreted, this alone may push the structure over the edge into the realm of interpretability. The ME methodology is very good at dealing with low resolution data and situations where there are considerable measurement errors.

Strong reflection for which phase ambiguities exist at this point can be phased using phase permutation methods based on error correcting codes or incomplete factorial designs coupled with entropy maximation using the concepts of a phasing tree. This was used to great effect in solving the TrpRS structure. This step can also include the resolution of MIR/SIR phase ambiguities for subsets of strong reflections which are highly correlated with each other.

Ab initio phasing to determine heavy atom sites may be possible in favourable cases.

The User Interface

The program has a graphical user interface based on Tcl and the Tk toolkit and is fully integrated into CCP4 and its file structures. Wei Dong is carrying out this work at Glasgow University and we hope to have a beta release ready for distribution by Christmas. It is intended that the user need have no great understanding of the ME formalism, and that the software is simple to use.

References

1. 'Entropy Maximisation Constrained by Solvent Flatness: A New Method for Macromolecular Phase Extension and Map Improvement', S.Xiang, C.W.Carter Jr., G.Bricogne, and C.J.Gilmore, *Acta Cryst.* (1993), **D49,** 193-212.
2. 'Overcoming Non-Isomorphism by Phase Permutation with Likelihood Scoring: Solution of the TrpRS Crystal Structure' S.Doublié, S.Xiang, C.J.Gilmore, G.Bricogne, & C.W.Carter, *Acta Cryst,* (1994), **A50,** 164-182.
3. 'Entropy Maximisation Constrained by Solvent Flatness: Macromolecular Phase Extension and Refinement' C.W.Carter, S.Xiang, S.Doublié, G.Bricogne & C.J.Gilmore, *Acta Cryst,* (1993), **A49**, 48.
4. ‘Overcoming Non-Isomorphism by Phase Permutation and Likelihood Scoring: Solution of the TrpRS Crystal Structure’ S. Doublié, S.Xiang, C.J.Gilmore, G.Bricogne & C.W.Carter Jnr. *Acta Cryst*. (1994), **A50**, 164-182.
5. 'Maximum Entropy and Bayesian Statistics in Crystallography' C.J.Gilmore, *Acta Cryst.* (1996) **A52**, 561-589
6. ‘A Multisolution Method of Phase Determination by Combined Maximisation of Entropy and Likelihood. VI The Use of Error-Correcting Codes as a Source of Phase Permutation and their Application to the Phase Problem in Powder, Electron and Macromolecular Crystallography’ C.J.Gilmore, W.Dong & G.Bricogne (1998), *Acta Cryst*. A, accepted for publication.

Tcl/Tk based crystallographic software : current state and new programs

L.M. Urzhumtseva#* and A.G. Urzhumtsev§*
#UPR de Biologie Structurale, IGBMC, Parc d'innovation, B.P. 163, Illkirch, France
§LCM3B, Université Nancy 1, Faculté des Sciences, 54506, Vandoeuvre-lès-Nancy, France
**IMPB, RAS, Pushchino, Moscow Region, 142292, Russia*
e-mail : sacha@lcm3b.u-nancy.fr

Abstract

A series of interactive crystallographic programs, aimed to facilitate some routine crystallographic jobs, is in the course of development. These programs are based on the Tcl/Tk language (Osterhaut, 1993). Some of these programs have been reported earlier (CONFOR, CONVROT). A new program, CRITXPL, which presents the results of the X-PLOR refinement (Brünger, 1992) in the form of a number of plots, is available.

I. Introduction

The current state of crystallographic work shows a discrepancy between the complexity of existing methods and their routine applications. This gap can be reduced by user-friendly programs using the rich graphic possibilities of modern computers. Another problem where such programs are important is crystallographic data processing. In our series of programs, which fill some "holes" in crystallographic software the choice for the Tcl/Tk macrolanguage (Osterhaut, 1993) has been done because of its accessibility (free of charge) and the compatibility of the scripts with different types of computers (SGI, DEC Alpha, IBM PC, …).

The first programs were chosen to answer some routine needs in protein crystallography. The historically first program, CONFOR, changes the formats of standard crystallographic data files. The second, CONVROT, helps in the transitions between different rotation and orthogonalisation conventions, which is important, in particular, for the molecular replacement analysis. The third program, CRITXPL, allows to present a long X-PLOR output file after model refinement in a graphic form and to manipulate this presentation. Some details of these programs are discussed below.

II. CRITXPL

Refinement of an atomic model using X-PLOR (Brünger, 1992) is currently carried out in a large number of macromolecular crystallography laboratories. The main results of the refinement are a model and the corresponding log-file which shows the process of refinement, giving crucial information on the success or failure of the process and proposes necessary modifications, for example, of the weights for different criteria. This log-file is a long list which contains a large amount of tables with criteria values, showing their variation during the refinement. The analysis of these tables takes significant time and effort.

The program CRITXPL has been developed to give a graphical representation of the variation of the minimisation criteria during an X-PLOR refinement. The program can be run either after X-PLOR or "in parallel" to it. In the latter case, the graphics can be easily updated during the refinement. To start the data processing, the name of an X-PLOR log-file can be chosen through the menu. This file can eventually contain several "minimisation processes". A process is defined as a sequence of X-PLOR "refinement frames" with increasing cycle number and with the same list of criteria. To easily identify the process to be displayed, a special line of comments can be optionally used in the X-PLOR script. For every process, CRITXPL shows, as a function of the cycle number, the plot of the gradient, of the temperature, of the R- and R-free factors, and, optionally, the plots of other criteria used. The latest values of the criteria are displayed numerically. The full list of all criteria applied in the analysed process is given by a set of buttons which can be used to remove, to replace or to show a plot of a chosen criterion. A maximum of eight criteria can be shown simultaneously. Any combination of them can be chosen and changed at any moment. The processes can be displayed either one by one or together. The latter option is useful for molecular dynamics; in this case, the criteria values both in the first and the last points of the chosen processes are displayed. The plots can be updated during an X-PLOR run, can be shown in a larger scale, and the point coordinates (the cycle number and the criterion value) can be obtained using a "mouse".

III. CONVROT

The need to describe the orientation of a known model is obvious for crystallographers. Possible descriptions by 3 (or 4) parameters differ by:

a) the choice of the rotated object (body or coordinate system),

b) the definition of the positive direction of the rotation (clockwise or counter-clockwise),

c) the type of angles (Eulerian or polar),

d) the choice of parameters inside a given type of angles,

e) the orthogonalisation convention for a given crystal.

In protein crystallography, at least about 20 of these variants are used by different programs. Unfortunately, this variety (which reflects the point of view of different authors) does not allow an easy understanding for users of crystallographic packages and a straightforward comparison of results. Actually, what the majority of crystallographers needs is simply to have *a rotation matrix, R, which should be applied to the coordinates of the model* to place it in the unit cell as the rotation function suggests. However, these matrices are not always given by the authors.

The program CONVROT (Urzhumtseva & Urzhumtsev, 1997) can convert the rotation description from any of the widely used molecular replacement systems to any other. All rotation conventions are considered from a *common* point of view (rotation matrices which are applied to the body's coordinates) which facilitates the transfer of results between the systems. It should be noted that all rotation conventions are defined in a Cartesian coordinate system, which for a given crystal can be introduced in various ways. The knowledge of the orthogonalisation agreement is also important when the crystal has symmetry because the corresponding operator(s) are naturally defined in crystallographic coordinates. The program can recalculate rotation parameters when the orthogonalisation convention is changed. In order to facilitate the analysis of the rotation function and the comparison of results of different molecular packages, CONVROT can provide the full list of rotation parameters symmetrically related to given ones, if such symmetries are defined.

The computational part of the program is written in standard FORTRAN. The program can be used in two different modes. Firstly, it can be run in a usual command-line shell such as UNIX or VMS. The input of the control data and the output of the results are organised in the form of a dialogue, and no special documentation is necessary in order to use the program. CONVROT can be run on any type of computer.

Secondly, a user-friendly environment for this program is developed with the use of the Tcl/Tk macrolanguage (Ousterhout, 1993). In this case all control data can be defined using menu and the order of data definition is arbitrary. The rotation parameters and the symmetry operators can be either typed in the corresponding windows or chosen by clicking the name of the corresponding file in a list. The output information is saved in a disk file. The program has an extended help for rotation and orthogonalisation conventions. This help includes definitions of the parameters, corresponding matrices, formulae for parameter recalculation and necessary references.

IV. CONFOR

The variety of the formats for main crystallographic data (structure factors, maps, atomic coordinates) creates a problem of data exchange between different applications. The program CONFOR (Urzhumtseva and Urzhumtsev, 1996) was planned to be a user-friendly converter of such files. The basic effort has been concentrated in the development of the part corresponding to the conversion of map files while the parts corresponding to the structure factors and coordinates have only few basic options for a time being. The program can read a density file, check the header information and save the map, optionally only partially and/or in different axes orientation, in a new format. It is also possible to change the variable part of the header, e.g., unit cell parameters; this can be important when preparing the map files for some versions of graphic display programs which have limitations in the unit cell size they can handle.

V. Technical information

The computational part of programs is written in Fortran 77, and the interactive graphics part is written in Tcl/Tk (Osterhaut, 1993). CONFOR uses the CCP4 library and some FRODO/O subroutines with the kind permission of their authors. The programs can be ran at a SGI computer and an Alpha DEC station under UNIX or at IBM PC under LINUX with the Tcl/Tk libraries installed. The program can be also run at IBM PC or Macintosh with a X-terminal simulator like *exodus*, *xwin* etc. CRITXPL accepts log-file produced by the 3.1x, 3.8x X-PLOR versions as well as by CNS.

No special documentation is necessary due to the menu-based character of the program. On-line help is available. The programs source code and the Tcl/Tk script are available on request from the authors. The corresponding e-mail address is sacha@lcm3b.u-nancy.fr . The authors thank Dr. A.Podjarny for his reading the manuscript and correction the language.

References

Brünger, A.T. (1992) *X-PLOR. A System for X-ray Crystallography and NMR,* Version 3.1. Yale University, Connecticut, USA

Osterhout, J.K. (1993) *Tcl and Tk Toolkit.* Reading, MA : Addison-Wesley Publishing Company

Urzhumtseva, L.M., Urzhumtsev, A.G. (1996) "Programs based on the Tcl/Tk interface. I. CONFOR - program to reformat 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.,* 402-410

On the density modification at very low resolution

A.G. Urzhumtsev§* ,A.D. Podjarny# and V.Y. Lunin*
§LCM3B, Université Nancy 1, Faculté des Sciences, 54506, Vandoeuvre-lès-Nancy, France
**IMPB, RAS, Pushchino, Moscow Region, 142292, Russia*
#UPR de Biologie Structurale, IGBMC, Parc d'innovation, B.P. 163, Illkirch, France
e-mail : sacha@lcm3b.u-nancy.fr

Abstract

The current techniques for *ab initio* macromolecular phasing are applicable at very low resolution and need to be followed by some phase extension procedures in order to improve the obtained image. Density modification methods can be used for this image improvement. Since they are usually applied at medium and high resolution (d < approx. 6 Å), a preliminary analysis of the applicability of these methods at very low resolution becomes necessary and is reported here.

I. Introduction

While direct methods for small molecules have become an usual crystallographic tool, the solution of the phase problem for large macromolecules needs extra information like heavy atom derivatives or anomalous dispersion. During recent years some progress in the development of *ab initio* phasing methods at very low resolution has been reported (see, for example, Lunin, Urzhumtsev and Skovoroda, 1990; Subbiah, 1993; Harris, 1995; Volkman et al., 1995; Lunin et al., 1995, Urzhumtsev & Podjarny, 1995a, Andersson & Hovmöller, 1996; see also the review by Podjarny & Urzhumtsev, 1997). In particular, these methods have obtained the first crystallographic image for the 50S ribosome particle from *Thermus thermophilus* (Volkman et al., 1995; Urzhumtsev, Vernoslova and Podjarny, 1996). However, the resolution which can be obtained today by these methods is not high enough. A possible way to get a higher resolution image is to develop further the methods like FAM (Lunin et al., 1998; its development and application to the T50S ribosome data allowed an increase in the resolution from 100 Å initially to about 30 Å). Another way is to create or to modify existing methods for image improvement. These methods have been developed to be applied at "usual" resolution (2-5 Å) therefore at very low resolution some basic hypotheses may be not fulfilled. A special analysis should be done in order to show how these methods should be updated, if possible, to be applied at very low resolution data.

One of the basic groups of methods for image improvement is density modification (see, for example, Podjarny, Rees and Urzhumtsev, 1997). The usual iteration procedure of modification of the density distribution function, depending either on the density value in a given point or on the coordinates of this point, can be eventually applied at any resolution. The simplest information which can be used is the flat density in the solvent region (Bricogne, 1974). The following basic questions should be analysed :

1. is such flattening applicable at low resolution (i.e., are structure factors recalculated at the same resolution close enough to the correct values) ?

2. can such procedures extend the phases at low resolution (do phases of structure factors at higher resolution have any structural information) ?
3. can the phase information be improved only by refining the envelope border (flat envelope) or are density values important ?
4. if the procedure is applicable, which are the optimal parameters ?

II. Computational experiment

Test calculations have been carried out using model data calculated from a density modulated envelope for the 50S ribosomal particle (Berkovich-Yellin, Wittmann & Yonath, 1990) arbitrarily placed in the experimentally observed unit cell (space group $P4_32_12$; a = b = 496Å, c = 196Å) as described before in Lunin et al. (1995). Since no molecular model was used in the procedure, the contribution of the disordered solvent (Urzhumtsev & Podjarny, 1995b) was not taken into consideration. No experimental error was introduced.

The following procedure has been applied :

1. calculate a density distribution at a given resolution (60, 40 or 30 Å);
2. define a molecular envelope as a set of unit cell points with the density value above a given threshold; the threshold was defined to leave a given percentage of the unit cell volume above this level;
3. flatten the density distribution below the threshold keeping the density above the threshold ("soft modification"); alternatively, at the test for using flat envelopes, the density above the threshold was also flattened ("hard modification");
4. calculate structure factors form the modified density distribution;
5. compare calculated structure factors with the exact values

We considered that "good" structure factors at the starting resolution indicate the applicability of such flat solvent hypothesis (or flat envelope hypothesis), and that "good" phases at higher resolution indicate the possibility to use such procedure for image improvement. The limit of "goodness of added phases" was taken as 60 degrees (mean cosine value is 0.5) following Lunin & Woolfson (1993). Note that no iterations were done, all numbers are shown for structure factors obtained after a single density modification.

III. Table explanation

The procedure was applied for the maps of the resolution of 60, 40, 30 Å. Tables 1-3 give the results of the comparison of structure factors. Different columns correspond to different cut-off level values varied from 10 to 90% (percentage of the volume above the threshold value).

The amplitude correlation was calculated for F_{obs} (simulated data set) and F_{calc} (after density modification) around their mean values.

The weighted phase correlation was calculated as the mean value of cos($\delta\phi$) weighted by F_{obs}**2, which corresponds to the closeness of the maps calculated with F_{obs} and with exact or calculated phases.

Cells are coloured accordingly to the criteria value; for the phase difference the limit value is 60 degrees. The shell of reciprocal space corresponding to the resolution slightly higher than the starting one is analysed in more detail and is given at the bottom of every table.

Table 1.1. Starting density at 60 Å; "hard" modification

Table 1.1a. Corr (Fcalc, Fobs), in percentage; resolution *vs* cut-off level

N	**Dmin**	**Dmax**	**refl**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**
1	80.0	500.	40	91	96	92	87	83				
2	60.	80.	49	83	88	89	84	74				
3	50.	60.	52	29	16	-2	-2	5				
4	45.	50.	48	1	-21	11	20	-3				
5	40.	45.	74	7	-6	-15	0	4				
6	35.	40.	118	- 4	11	-5	-13	-18				
7	30.	35.	211	- 3	-13	-12	-7	3				
8	25.	30.	383	10	5	6	3	7				
9	20.	25.	876	10	14	6	9	7				
3.1	55.	60.	22	36	5	-1	7	1				
3.2	50.	55.	30	-7	0	-5	-5	12				

Table 1.1b. Weighted phase correlation, in percentage; resolution *vs* cut-off level

N	**Dmin**	**Dmax**	**refl**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**
1	80.0	500.	40	98	99	97	92	90				
2	60.	80.	49	96	97	96	93	85				
2	50.	60.	52	38	55	46	17	-33				
2	45.	50.	48	29	3	-13	-42	-44				
2	40.	45.	74	51	40	11	-17	-46				
2	35.	40.	118	2	23	15	-7	-13				
2	30.	35.	211	27	21	1	-18	-30				
2	25.	30.	383	21	11	-7	-26	-23				
2	20.	25.	876	-5	0	-10	-3	12				
2	55.	60.	22	40	56	55	20	-55				
2	50.	55.	30	16	52	39	16	-16				

Table 1.1c. Mean $\delta\phi$, in degrees

N	**Dmin**	**Dmax**	**refl**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**
1	80.0	500.	40	27	26	20	36	42				
2	60.	80.	49	27	23	31	35	41				
3	50.	60.	52	75	60	65	85	105				
4	45.	50.	48	78	79	88	86	95				
5	40.	45.	74	66	65	76	95	115				
6	35.	40.	118	83	83	83	96	98				
7	30.	35.	211	78	82	94	101	101				
8	25.	30.	383	81	85	94	100	99				
9	20.	25.	876	91	92	96	90	85				
3.1	55.	60.	22	62	43	49	81	121				
3.2	50.	55.	30	85	72	77	87	93				

Table 1.2. Starting density at 60 Å; "soft" modification

Table 1.2a. Corr (Fcalc, Fobs), in percentage; resolution *vs* cut-off level

N	Dmin	Dmax	refl	10	20	30	40	50	60	70	80	90
1	80.0	500.	40	68	87	94	98	99	100	100	100	100
2	60.	80.	49	76	86	91	94	96	98	99	100	100
3	50.	60.	52	-7	4	14	28	33	33	33	31	25
4	45.	50.	48	18	20	15	11	12	15	19	22	22
5	40.	45.	74	-7	-3	5	4	-5	-5	0	1	4
6	35.	40.	118	4	6	3	2	3	1	0	-2	-2
7	30.	35.	211	-11	25	8	27	24	10	27	6	7
8	25.	30.	383	6	23	17	13	1	3	14	18	12
9	20.	25.	876	20	9	16	16	14	14	18	17	12
3.1	55.	60.	22		0	3	10	12	12	12	12	10
3.2	50.	55.	30		7	17	32	38	36	34	34	28

Table 1.2b. Weighted phase correlation, in percentage; resolution *vs* cut-off level

N	Dmin	Dmax	refl	10	20	30	40	50	60	70	80	90
1	80.0	500.	40	92	97	99	100	100	100	100	100	100
2	60.	80.	49	88	95	98	99	99	100	100	100	100
3	50.	60.	52	7	24	42	54	59	60	61	61	54
4	45.	50.	48	37	47	52	50	48	47	47	44	40
5	40.	45.	74	-10	20	35	40	33	25	21	20	19
6	35.	40.	118	-3	13	30	32	28	27	25	19	11
7	30.	35.	211	1	25	42	24	15	21	4	5	1
8	25.	30.	383	-5	27	34	18	1	-7	-9	-9	-5
9	20.	25.	876	11	1	-4	-15	-15	-10	-4	-3	2
3.1	55.	60.	22		49	64	69	70	71	71	71	70
3.2	50.	55.	30		4	17	29	35	36	37	37	35

Table 1.2c. Mean $\delta\phi$, in degrees

N	Dmin	Dmax	refl	10	20	30	40	50	60	70	80	90
1	80.0	500.	40	27	26	21	20	12	2	1	1	1
2	60.	80.	49	32	25	22	18	16	12	10	3	1
3	50.	60.	52	87	79	67	62	55	53	51	59	72
4	45.	50.	48	80	69	69	69	64	66	69	67	67
5	40.	45.	74	96	79	63	59	68	70	73	79	84
6	35.	40.	118	88	89	82	77	78	79	83	83	85
7	30.	35.	211	90	83	71	91	90	87	90	92	93
8	25.	30.	383	92	78	80	83	88	94	94	94	96
9	20.	25.	876	86	88	96	97	93	91	89	88	89
3.1	55.	60.	22		58	42	38	36	35	34	41	48
3.2	50.	55.	30		94	85	80	69	66	64	70	80

Table 2.1. Starting density at 40 Å; "hard" modification

Table 2.1a. Corr (Fcalc, Fobs), in percentage; resolution *vs* cut-off level

N	Dmin	Dmax	refl	10	20	30	40	50	60	70	80	90
1	80.0	500.	40	91	98	92	89					
2	60.	80.	49	93	89	87	81					
3	50.	60.	52	88	92	79	71					
4	45.	50.	48	63	65	53	55					
5	40.	45.	74	59	72	60	64					
6	35.	40.	118	16	14	5	19					
7	30.	35.	211	10	3	6	17					
8	25.	30.	383	24	2	9	28					
9	20.	25.	876	8	15	15	9					
6.1	37.	40.	66	9	2	15	15					
6.2	35.	37.	52	24	25	-1	23					

Table 2.1b. Weighted phase correlation, in percentage; resolution *vs* cut-off level

N	Dmin	Dmax	refl	10	20	30	40	50	60	70	80	90
1	80.0	500.	40	99	99	96	94					
2	60.	80.	49	98	98	96	95					
3	50.	60.	52	95	96	93	89					
4	45.	50.	48	91	93	90	85					
5	40.	45.	74	86	96	90	85					
6	35.	40.	118	54	44	2	-50					
7	30.	35.	211	49	16	-22	-55					
8	25.	30.	383	29	15	-22	-36					
9	20.	25.	876	9	-18	-30	-21					
6.1	37.	40.	66	48	53	8	-45					
6.2	35.	37.	52	61	32	-8	-55					

Table 2.1c. Mean $\delta\phi$, in degrees

N	Dmin	Dmax	refl	10	20	30	40	50	60	70	80	90
1	80.0	500.	40	23	15	21	43					
2	60.	80.	49	27	22	28	29					
3	50.	60.	52	29	21	22	34					
4	45.	50.	48	30	24	35	42					
5	40.	45.	74	38	26	37	38					
6	35.	40.	118	65	74	88	115					
7	30.	35.	211	65	85	100	114					
8	25.	30.	383	79	84	99	107					
9	20.	25.	876	87	94	106	99					
6.1	37.	40.	66	63	72	81	113					
6.2	35.	37.	52	67	75	96	117					

Table 2.2. Starting density at 40 Å; "soft" modification

Table 2.2a. Corr (Fcalc, Fobs), in percentage; resolution *vs* cut-off level

N	Dmin	Dmax	refl	10	20	30	40	50	60	70	80	90
1	80.0	500.	40	76	91	97	99	99	100	100	100	100
2	60.	80.	49	85	95	98	99	99	100	100	100	100
3	50.	60.	52	72	90	97	99	99	100	100	100	100
4	45.	50.	48	51	77	90	95	96	97	98	99	99
5	40.	45.	74	41	68	83	90	94	96	98	99	99
6	35.	40.	118	34	37	40	40	40	40	39	36	29
7	30.	35.	211	18	24	34	35	34	32	29	28	31
8	25.	30.	383	19	36	42	38	35	34	32	32	34
9	20.	25.	876	20	14	14	14	12	11	11	12	13
6.1	37.	40.	66	26	29	36	39	38	35	31	24	13
6.2	35.	37.	52	42	47	49	49	50	52	53	54	55

Table 2.2b. Weighted phase correlation, in percentage; resolution *vs* cut-off level

N	Dmin	Dmax	refl	10	20	30	40	50	60	70	80	90
1	80.0	500.	40	95	99	100	100	100	100	100	100	100
2	60.	80.	49	95	98	100	100	100	100	100	100	100
3	50.	60.	52	79	94	99	100	100	100	100	100	100
4	45.	50.	48	78	94	98	99	100	100	100	100	100
5	40.	45.	74	69	89	96	98	99	99	100	100	100
6	35.	40.	118	57	69	77	80	81	79	76	72	66
7	30.	35.	211	51	66	70	70	70	68	66	62	56
8	25.	30.	383	31	50	52	46	43	42	40	36	31
9	20.	25.	876	30	28	1	-18	-23	-24	-25	-25	-24
6.1	37.	40.	66	53	68	77	81	81	79	74	69	61
6.2	35.	37.	52	61	70	77	80	80	79	78	76	72

Table 2.2c. Mean $\delta\phi$, in degrees

N	Dmin	Dmax	refl	10	20	30	40	50	60	70	80	90
1	80.0	500.	40	34	20	15	7	3	3	1	1	1
2	60.	80.	49	35	21	17	10	9	7	4	3	2
3	50.	60.	52	48	31	18	11	9	8	7	7	1
4	45.	50.	48	48	26	20	10	8	6	5	5	1
5	40.	45.	74	52	38	26	15	12	10	7	5	3
6	35.	40.	118	71	63	58	57	54	53	54	57	66
7	30.	35.	211	64	59	54	52	55	55	57	60	65
8	25.	30.	383	78	72	68	69	74	75	77	78	82
9	20.	25.	876	78	82	93	101	102	102	101	101	100
6.1	37.	40.	66	71	62	59	57	51	49	53	56	73
6.2	35.	37.	52	70	64	56	56	57	58	57	59	58

Table 3.1. Starting density at 30 Å; "soft" modification

Table 3.1a. Corr (Fcalc, Fobs), in percentage; resolution *vs* cut-off level

N	Dmin	Dmax	refl	5	10	20	30	40	50	60	70	80	90
1	80.0	500.	40	73	85	96	99	100	100	100	100	100	100
2	60.	80.	49	71	87	97	99	99	100	100	100	100	100
3	50.	60.	52	49	76	92	97	98	99	99	100	100	100
4	45.	50.	48	50	69	87	95	96	97	98	99	99	100
5	40.	45.	74	51	68	89	96	97	98	99	99	99	100
6	35.	40.	118	67	77	91	96	98	98	99	99	100	100
7	30.	35.	211	65	76	87	92	95	97	98	99	99	100
8	25.	30.	383	41	48	56	58	57	57	55	50	40	22
9	20.	25.	876	25	34	36	31	28	26	25	24	22	21
8.1	27.	30.	199	45	52	59	63	62	61	58	52	38	19
8.2	25.	27.	184	31	39	49	48	47	47	48	47	42	36
9.1	22.	25.	427	21	31	35	29	25	25	25	26	25	24
9.2	21.	22.	213	14	22	18	14	12	9	6	4	0	-2

Table 3.1b. Weighted phase correlation, in percentage; resolution *vs* cut-off level

N	Dmin	Dmax	refl	5	10	20	30	40	50	60	70	80	90
1	80.0	500.	40	93	97	100	100	100	100	100	100	100	100
2	60.	80.	49	87	95	99	100	100	100	100	100	100	100
3	50.	60.	52	78	90	98	100	100	100	100	100	100	100
4	45.	50.	48	72	88	97	99	100	100	100	100	100	100
5	40.	45.	74	84	89	96	99	99	100	100	100	100	100
6	35.	40.	118	85	91	97	99	99	100	100	100	100	100
7	30.	35.	211	87	93	97	98	99	100	100	100	100	100
8	25.	30.	383	67	73	81	85	85	84	84	83	81	82
9	20.	25.	876	53	64	67	60	54	50	47	43	40	36
8.1	27.	30.	199	71	77	83	87	87	87	86	85	84	86
8.2	25.	27.	184	58	64	76	79	79	78	77	76	73	66
9.1	22.	25.	427	56	67	74	72	68	67	66	64	62	58
9.2	21.	22.	213	48	58	53	35	27	21	15	10	6	4

Table 3.1c. Mean $\delta\phi$, in degrees

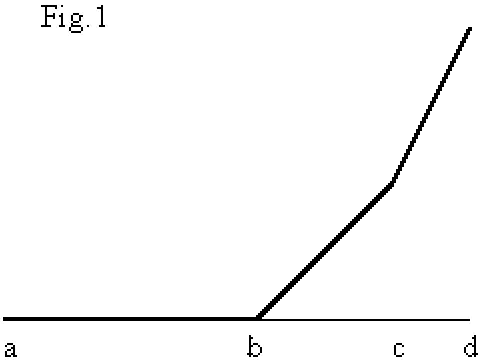
N	Dmin	Dmax	refl	5	10	20	30	40	50	60	70	80	90
1	80.0	500.	40	35	25	20	7	1	1	1	1	1	0
2	60.	80.	49	44	35	25	7	5	2	2	2	1	1
3	50.	60.	52	48	40	24	10	8	7	7	4	4	4
4	45.	50.	48	51	39	24	13	10	9	8	8	5	4
5	40.	45.	74	43	33	24	15	10	8	7	6	4	2
6	35.	40.	118	50	36	21	11	9	8	7	5	1	2
7	30.	35.	211	35	38	21	17	14	10	9	6	4	3
8	25.	30.	383	57	53	44	43	43	43	44	46	49	54
9	20.	25.	876	66	60	60	66	69	71	72	73	73	75
8.1	27.	30.	199	53	52	42	42	42	42	43	45	48	54
8.2	25.	27.	184	62	55	46	43	43	44	46	47	50	53
9.1	22.	25.	427	63	55	50	53	55	58	59	60	60	64
9.2	21.	22.	213	67	63	65	76	81	82	83	85	86	88

IV. Results and discussion

The following observations can be made :

1. The structure factors, calculated at the same resolution as the starting synthesis, are close enough to the correct values, which means that the hypothesis on the flat solvent can be used at such resolution.
2. For the structure factors calculated in higher resolution shells, the calculated amplitudes are not so good. However, in many cases the phases in the closest resolution shell have sufficiently high quality to be used to improve the starting image.
3. The results obtained by the "soft" modification, i.e. when the highest density distribution values are not changed (keeping the density "inside the envelope") are much less sensitive to the threshold level. Note that a similar observation was found for the molecular replacement searches with an envelope (Urzhumtsev & Podjarny, 1995a).
4. The hard modification is significantly less useful at 40 Å than at 60 Å. It means that at such (and higher) resolution the phase extension cannot be achieved by a simple refinement of the envelope border and needs the knowledge of density distribution values.
5. It is interesting to note that in several cases (for example, Tables 2.2) a "wave effect" can be observed : when starting from the 40Å -resolution map, the amplitudes and the phases in the resolution shell 35-37 Å are better than those in the resolution shell 37-40 Å. However, it could be a purely statistical effect.
6. The density modification became more efficient when increasing the resolution of the starting synthesis.
7. Optimal cut-off level does not necessarily correspond to the correct molecular volume (in this case it was about 35 percent).

Fig.1



As was discussed by Lunin (1988), the basic idea which is behind most of density modification procedures and which defines the form of the density modification function is the closeness of the electron density histograms. In our case, the corresponding histograms have been calculated at different resolution from 20 to 90 Å. We have no possibility to discuss here the details of this analysis and can only mention that the density modification function corresponding to the optimal density modification from 90 to 20 Å resolution maps can be nicely approximated by a function schematically presented in Figure 1 : at the interval (a,b) it corresponds to the solvent flattening, at the interval (b,c) it "keeps" the density values and the interval (c,d) the function needs to sharpen the highest density values. In general, this function supports the idea of "soft" modification. The application of histogram-fitted density modification will be discussed elsewhere.

References

Andersson, K.M., Hovmöller, S. (1996) *Acta Cryst.*, **D52**, 1174-1180

Berkovich-Yellin, Z., Wittmann, H.G., Yonath, A. (1990) *Acta Cryst.*, **B46**, 637-643

Bricogne, G. (1974) *Acta Cryst.*, **A30**, 395-405

Harris, G.W. (1995) *Acta Cryst.*, **D51**, 695-702

Lunin, V.Yu. (1988) *Acta Cryst.*, **A44**, 144-150

Lunin, V.Yu., Urzhumtsev, A.G., Skovoroda, T.A. (1990) *Acta Cryst.*, **A46**, 540-544

Lunin, V.Yu., Woolfson, M.M. (1993) *Acta Cryst.*, **D49**, 530-533

Lunin, V.Yu., Lunina, N.L., Petrova, T.E., Vernoslova, E.A., Urzhumtsev, A.G., Podjarny, A.D. (1995) *Acta Cryst.,* **D51,** 896-903

Lunin, V.Yu., Lunina, N.L., Petrova, T.E., Urzhumtsev, A.G., Podjarny, A.D. (1998). *Acta Cryst.,* **D53,** in press

Podjarny, A.D., Rees, B., Urzhumtsev, A.G. (1996) "Density modification". In "Methods in Molecular Biology", **56** : "Crystallographic Methods and Protocols", C.Jones, B.Milloy, M.R.Sanderson, eds., Totowa, New Jersey : Humana Press, 205-226

Podjarny, A.D., Urzhumtsev, A.G. (1997) "Low resolution phasing". In *Methods in Enzymology*, Academic Press, San Diego., C.W.Carter, Jr., R.M.Sweet, eds. **276A**, 641-658

Subbiah, S. (1993) *Acta Cryst.,* **D49**, 108-119

Urzhumtsev, A.G., Podjarny, A.D. (1995a) *Acta Cryst.,* **D51**, 888-895

Urzhumtsev, A.G. & Podjarny, A.D. (1995b) *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography,* **32**, 12-16.

Urzhumtsev, A.G., Vernoslova, E.A., Podjarny, A.D. (1996) *Acta Cryst.,* **D52**, 1092-1097

Volkmann, N., Schlunzen, F., Urzhumtsev, A.G., Vernoslova, E.A., Podjarny, A.D., Roth, M., Pebay-Peyroula, E., Berkovitch-Yellin, Z., Zaytsev-Bashan, A. & Yonath, A. (1995) *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography,* **32**, 23-32.

MIR: An Automated Program For Isomorphous Replacement

A.Vagin (1), A.Teplyakov (2) and M.Isupov (3)
(1) Department of Chemistry, University of York, Heslington, York, UK
(2) European Molecular Biology Laboratory, Hamburg, Germany
(3) Departments of Chemistry and Biological Sciences, University of Exeter, Exeter, UK

MIR is an automated program for heavy atom/anomalous scatterers location and subsequent phasing of multiple isomorphous and/or anomalous X-ray data. It is based on the translation function approach for heavy atom (HA) location as implemented in the program TRAHALO [1] which forms part of the program MIR.

In the first step, HA search is performed for each derivative separately. A full-symmetry translation function (TF) [2] is calculated using a one-atom probe model. This gives a primary list of HA selected on the basis of the highest phasing power (PP). For each of these HA sites the TF search for the second site is performed using a model consisting of a one-atom probe and one atom in the fixed position. The sites thus found are checked against the primary list and a pair of HA with the highest PP is considered to be the best solution. Note that the pair of HA will have a common coordinate origin and hand. The HA search goes on until the addition of a new atom does not increase the PP.

In the second step, cross-difference Fourier syntheses are calculated to verify the HA sites located in the first step and to find additional sites. Thus, phases from derivative 1 are used to locate HA sites in derivative 2 which are then used for HA location in derivative 1. If these HA-1-NEW sites coincide (at least partially) with the original set of sites for derivative 1, the sites used for phase calculation (derivative 2) are considered to be correct. This procedure helps to detect some additional sites which were missed in the first step and at the same time to avoid incorporation of a large number of false sites.

Third step is a density modification (solvent flattening) to produce the final phases. Additionally, MIR can use external phases, e.g. the molecular replacement phases.

Anisotropic scaling and correction of the experimental data have been introduced in MIR. Although anisotropic scaling of derivatives data is relatively common, the native data could be anisotropic as well, which could adversely affect the results of HA location. The work is in progress to determine whether anisotropic correction of the native data can improve the HA search results.

MIR is a part of program suite BLANC [3], which contains programs to convert data from MTZ [4] and CIF [5] input data formats to BLANC format. The program is available free as part of BLANC from AV (alexei@yorvic.york.ac.uk).

Test

Isomorphous data for hevamine, crystallised in the space group P212121, a = 52.3, b = 57.7, c = 82.1 Å, one 30 kDa molecule per asymmetric unit [6].

The structure has been solved with four derivatives,

- trimethyllead acetate (TLA),
- 2,4-dichloromercuri-6-nitrophenol (CMNP),
- 4-diacetamido-phenyl-mercuric acetate (APMA) and
- silver nitrate (AgNO3).

The heavy atom search has been performed at 4 Å resolution with low resolution cut-off at 10 Å (Boff = 400 Å^2) [1].

Step one: Translation function HA search for each derivative (diagonal terms in Table 1). Solutions for TLA (one site), AgNO3 (both sites) and CMNP (the major site) were essentially correct. Two out of four sites wele located in APMA.

Step two: Difference Fourier search for each derivative using phases from another derivative. CMNP is considered to be the best derivative. Phases calculated using the CMNP sites and new TLA sites are characterised by the FOM of 0.62. These combined phases are used to locate minor sites in the other two derivatives. Finally, all four HA were located in APMA. The FOM of combined phases from all four derivatives was 0.69.

Step three: Five cycles of solvent flattening produce phases with the FOM of 0.92.

Acknowledgements

We thank Anke Terwisscha van Scheltinga for providing the experimental data for testing the program and Garib Murshudov for useful discussions.
The work was supported by the EU BIOTECH grant BIO4CT-96-0189 to A.V. and by the grant from the Chemical and Pharmaceutical Directorate of the BBSRC to M.I.

References

[1] Vagin,A. & Teplyakov,A. (1998). Acta Cryst. D 54, 400-402.
[2] Vagin,A. & Teplyakov,A. (1997). J. Appl. Cryst. 30, 1022-1025
[3] Vagin,A., Murshudov,G. & Strokopytov,B. (1998). J. Appl. Cryst. 31, 98-102.
[4] CCP4 (1994). Acta Cryst. D 50, 760-763.
[5] Hall, S. (1991) Acta Cryst. A 47, 655-685.
[6] Terwisscha van Scheltinga,A.C, Kalk,K.H, Beintema,J.J. & Dijkstra,B.W. (1994). Structure 2, 1181-1189

Table 1

Derivatives used for phasing

	TLA	CMNP	APMA	AgNo3
TLA (1 site)				
Natom :	1 (1)	2 (1)	4 (1)	3 (0)
PP :	1.29	1.20	1.04	0.96
FOM :	0.53	0.45	0.40	0.38
CMNP (2 sites)				
Natom :	4 (1)	5 (1)	2 (1)	0
PP :	1.20	1.44	0.99	0
FOM :	0.46	0.40	0.31	0

APMA (4 sites)				
Natom :	1 (1)	1 (1)	9 (2)	2 (0)
PP :	1.06	1.11	1.16	1.02
FOM :	0.40	0.35	0.35	0.32
AgNo3 (2 sites)				
Natom :	2 (2)	0	0	2 (2)
PP :	1.24	0	0	1.22
FOM :	0.40	0	0	0.34

Natom

The number of HA found in this derivative in the first step using TRAHALO (diagonal terms) and in the second step using phases from other derivatives (non-diagonal terms). In brackets is the number of correctly located HA sites.

PP

Phasing power <FH/ Lack of closure>

FOM

Figure of merit

The Effect of Overall Anisotropic Scaling in Macromolecular Refinement

Garib N. Murshudov[1,2],Gideon J. Davies[1], Mikhael Isupov[3], Szymon Krzywda[4], Eleanor J. Dodson[1].*

[1]Chemistry Department, University of York, York, U.K.
[2]CLRC, Daresbury Laboratory, Warrington, Daresbury, U.K.
[3]Chemistry Department, University of Exeter, Exeter, U.K.
[4]Crystallography Department, Faculty of Chemistry, Adam Mickiewicz University, Poznan, Poland,
[*]e-mail garib@yorvic.york.ac.uk

1. Introduction

Parametrisation is still one of major problems of refinement programs. This includes parameters of individual atoms, molecules and the whole crystal structure. All parametrisations depend on available experimental data and the current stage of structural analysis. In this note we discuss the importance of overall anisotropic scaling during refinement and give several examples where it has improved refinement substantially.

It was noted by S.Gamblin (1996): ``in the absence of an overwhelming argument(such as cubic space group), it is always safest to assume that diffraction is anisotropic". This fact should be taken into account in refinement as well as in data collection strategy and data processing. Figure 1 shows that sometimes only at high resolution does the anisotropicity of data becomes apparent on the diffraction images. Anisotropicity of data might also cause problems in data collection. If one accidentally collects the first image in the direction of high thermal motion, one might not make the optimal decision for data collection. An image perpendicular to the first should also be collected to observe the true behaviour of the crystal.

The anisotropicity of the data should be taken into account during the data processing stage. This is possible with the CCP4 SCALA program, written by Phil Evans. In the absence of prior information about the contents of the crystal, refinement of the overall anisotropicity at the data processing stage will remain ill determined. In that case, any residual anisotropy should be taken into account at the refinement stage or alternatively refinement and data processing could be alternated. An even better approach would be simultaneous refinement and data processing, but this would only be possible if the structure had already been solved.

In the first ever paper on least-squares refinement of crystal structures, Hughes (1941) noted the existence of anisotropicity and described improved refinement behaviour by the introduction of anisotropic scale factors. It is surprising that in highly mobile and large structures, such as proteins, this fact has not been taken into account until recently.

2. Sources of anisotropicity

Several factors contribute to apparent atomic anisotropicity. The crystal itself (except in a cubic space group) is, in general, an anisotropic field, so it is to be expected that the data collected from it may exhibit overall anisotropicity. Freezing and/or addition of substrates will usually change the anisotropicity of the crystal, in general increasing it.

A second source of anisotropicity is the movement of whole molecules as a rigid bodies within the crystal lattice. This can be described by TLS parameters (20 more per molecule) which are independent of the crystal form (Schomaker and Trueblood 1968). The RESTRAIN program (Moss et al., 1996) is able to evaluate these, and the correction has been shown to be valuable in some situations.

A third source of anisotropicity is vibration along torsion angles. In principle this might be described by refining the torsion angles themselves, and estimating their displacement parameters. However there are problems, since these parameters are highly correlated and such refinement may be sensitive to small perturbations of one or several of these. It may be better to deduce the displacement parameters of the torsion angles from the individual anisotropic atomic U values.

To summarise, the observed atomic anisotropic **U** values can be written as:

$$\mathbf{U}_{\mathrm{atom;overall}} = \mathbf{U}_{\mathrm{crystal}} + \mathbf{U}_{\mathrm{TLS}} + \mathbf{U}_{\mathrm{torsion}} + \mathbf{U}_{\mathrm{atom}} \quad (1)$$

where $\mathbf{U}_{\mathrm{atom;overall}}$ is the overall anisotropic U value, $\mathbf{U}_{\mathrm{crystal}}$ is the contribution of crystal anisotropicity, $\mathbf{U}_{\mathrm{TLS}}$ that of model anisotropicity (TLS), $\mathbf{U}_{\mathrm{torsion}}$ that of motion about torsion angle and finally $\mathbf{U}_{\mathrm{atom}}$ that of the atomic anisotropicity along and across covalent bonds. Cruickshank (1956) noted that removing the $\mathbf{U}_{\mathrm{crystal}}$ made the refinement of individual anisotropic **U** values more stable, and it seems reasonable to apply these simple corrections to remove the modes related to $\mathbf{U}_{\mathrm{crystal}}$ and $\mathbf{U}_{\mathrm{TLS}}$.

Care should be taken in the refinement of these different contributions as they are highly correlated. To overcome this difficulty they could be refined at the different levels. I.e. first $\mathbf{U}_{\mathrm{crystal}}$, second $\mathbf{U}_{\mathrm{TLS}}$, third $\mathbf{U}_{\mathrm{torsion}}$ and finally $\mathbf{U}_{\mathrm{atom}}$, or alternatively refine $\mathbf{U}_{\mathrm{crystal}}$, $\mathbf{U}_{\mathrm{TLS}}$ and along internal degrees of freedom as described by Diamond (1990).

$\mathbf{U}_{\mathrm{crystal}}$ is in principle sum of two factors: 1) those remaining after data processing and 2) common mode from U_{TLS}.

Here effect of the anisotropic scaling only will be discussed. For refinement of individual atomic anisotropic thermal parameters see Murshudov et al. (1998)

3. Anisotropic Scaling

For anisotropic scaling, at each cycle of refinement the least-squares residual is used to derive overall parameters:

$$\sum(|F_o| - k(s)|F_c|)^2 \longrightarrow min \qquad (2)$$

where the scale factor $k = k_0\ e^{-h^T U^* h}$. U^* is symmetric reciprocal space anisotropic tensor. The space group puts constraints on the anisotropic **U** tensor. For example, cubic space groups do not have an overall anisotropic **U**. The space group $P4_2\ 2_1\ 2$ has 2 parameters and so on. In the implementation in the program REFMAC (Murshudov et al. 1997) this fact has been taken into account. As in this treatment anisotropic **U** is the difference between the observed and calculated structure factors, there is no need to use positive definite constraints. At each cycle of refinement, the program refines anisotropic scale factors and applies them to calculated ones. There is also an option to apply anisotropic scale to the observed structure factors. At this stage, application of this option is not recommended since it changes the observed structure factors. Thus, the calculated R-values would not be comparable with each other. If anisotropic **U** values would be applied to observed structure factors then:

$$R_k = \frac{\sum(|k_1(s)F_o| - |F_c|)^2}{\sum k_1^2(s)|F_o|^2} = \frac{\sum k_1(s)^2(|F_o| - \frac{1}{k_1(s)}|F_c|)}{\sum k_1^2(s)|F_o|^2} \qquad (3)$$

It is clear from this equation that at each cycle calculated R-value is in fact weighted R-value with weights $k_1^2(s)$, where $k_1(s) = [1/\ k(s)]$. If k(s) is refined at each cycle then behaviour of R and R_k could be different.

4. Examples of application

All the examples given here are structures which have previously been refined with isotropic scale factors. When refinement of anisotropic scale factors became available they were applied to different test cases. In all cases application of anisotropic scale factors not only improved R and R-free but also refinement that had apparently converged restarted. It is important to note, that in addition to R values, the geometric parameters of the model improved significantly and difference maps became cleaner.

Native Catalase at 1.5Å

Crystals of catalase from the bacterium *Mycrococcus lysodeikticus* are almost perfect. Data from these crystals have now been collected at 0.9Å resolution. Even in this case one can see that anisotropic scale factor improves R-value and free R-value (Table 1)

Catalase frozen at 1.96Å

Data from catalase soaked in peracetic acid solution were used in order to obtain the reaction intermediate. Data were collected, from a frozen crystal, using CuK_a radiation and the RAXIS II as detector. In this case it can be seen that effect of anisotropic scaling is much larger than in native room temperature data (Table 1).

Cellulase

Data for this enzyme were collected from frozen crystals in the home laboratory to 1.6Å resolution and the MIR structure determination was essentially trivial (Davies et al., 1998) The refinement, although straightforward, converged with unusually high values for both R and Rfree, both above 20%. It was only upon collection of atomic (0.9Å) resolution data that the anisotropic nature of the diffraction became easily apparent to the authors from inspection of the diffraction images. At this point, the anisotropic data scaling became available resulting in immediate reductions in R and R free of over 6% (Table 1).

Myoglobin

This case was one of the prime reasons for speeding up the implementation of anisotropic scale factor refinement. There were 10 different data sets of mutant and native myoglobins with different complexes. All data sets were collected from frozen crystals. Refinement with overall isotropic B values stuck with R 22%, R-free around 29%. Refinement of overall anisotropic scale factor immediately reduced R-value and free R-value (Table 1).

Oxoindolyl-L-alanyn complexed tryptophanase

The complex of tryptophanase from *P.vulgaris* with competetitve inhibitor Oxindolyl-L-alanine have been crystallised in the space group $P2_12_12$ with a=152.480 , b= 213.694 , c=63.518 which was different from holotryptophanase crystals. The structure has been solved by molecular replacement using the holotryptophanase coordinates. The conventional REFMAC refinement at 18-3 Å converged with R/R-free = 28.4/31.3%. Refinement with anisotropic scaling reduced R/R-free to 24.88/27.8. Moreover, the refinement went on to R/R-free=18.0/24.7 (Table 1)

5. Conclusions

Examples given here show that application of anisotropic scale factors improves the refinement. The current implementation is only a partial solution to the problem of anisotropicity and the general problem of scaling observed and calculated structure factors. A better solution would be to use likelihood functions which contain information about scale factors, model errors and partiality, the experimental uncertainty of observed structure factors, NCS, phase information and any other available information.

Another problem related to anisotropic scaling is the overall molecular motion (TLS) in the unit cell. Future development of REFMAC will incorporate this information. This can be achieved easily since the fast refinement of individual **U** values by FFT now is available.

In principle, the treatment of anisotropic **U** values should start from the data processing so that many factors contributing to anisotropic scale factor could be accounted for. Then the refinement protocol could be used to model the residual overall anisotropic scale factor.

Acknowledgements

We thank the members of the Structural Biology Laboratory,Department of Chemistry, University of York for testing the program. GNM is supported by BBSRC postdoctoral fellowship awarded to CCP4 (grant B05273), GJD by the Royal Society, EJD by the MRC, MI by the BBSRC.

References

CCP4. Collaborative Crystallographic Project, Number 4. (1994) *Acta Cryst.* **D50**, 760-763

Cruickshank, D.W. (1956) *Acta Cryst.* **9**, 747-753

Davies, G.J., Dauter, M., Brzozowski, A.M., Bjornvad, M.E., Andersen, K.V. & Schulein, M. (1998) *Biochemistry* **37**, 1926-1932

Diamond, R. (1990) *Acta Cryst.* **A46** 425-435

Gamblin, S.J. (1996) in *Macromolecular Refinement. Proceedings of the CCP4 Study Weekend* Ed. Dodson,E., Moore,M., Ralph,A., Bailey, S. 163-170

Hughes, E.W. (1941) *J. Am. Chem. Soc.* **63**, 1737-1752

Isupov et al., manuscript in preparation

Murshudov, G.N., Lebedev, A., Vagin, A.A., Wilson, K.S., Dodson, E.J. (1998) submitted to *Acta Cryst. D*

Murshudov, G.N, Melik-Adamyan, W.R., Grebenko, A.I., Barynin, V.V., Vagin, A.A., Vainshtein, B.K., Dauter, Z. & Wilson, K.S. (1992) *FEBS letters* **312**, 127-131

Murshudov, G.N., Vagin, A.A. & Dodson, E.J. (1997) *Acta Cryst.* **D53**, 240-253

Moss, D.S., Tickle, I.J., Theis, O. & Wostrack, A. in ``*Macromolecular refinement''* Proceedings of the CCP4 Study Weekend, 105-113, Ed. Dodson, E., Moore, M., Ralph, A. & Bailey, S. CCLRC, Daresbury Laboratory

Schomaker, V. & Trueblood, K.N. (1968) *Acta Cryst.* **B24** 63-76

Table 1: Effect of anisotropic scaling

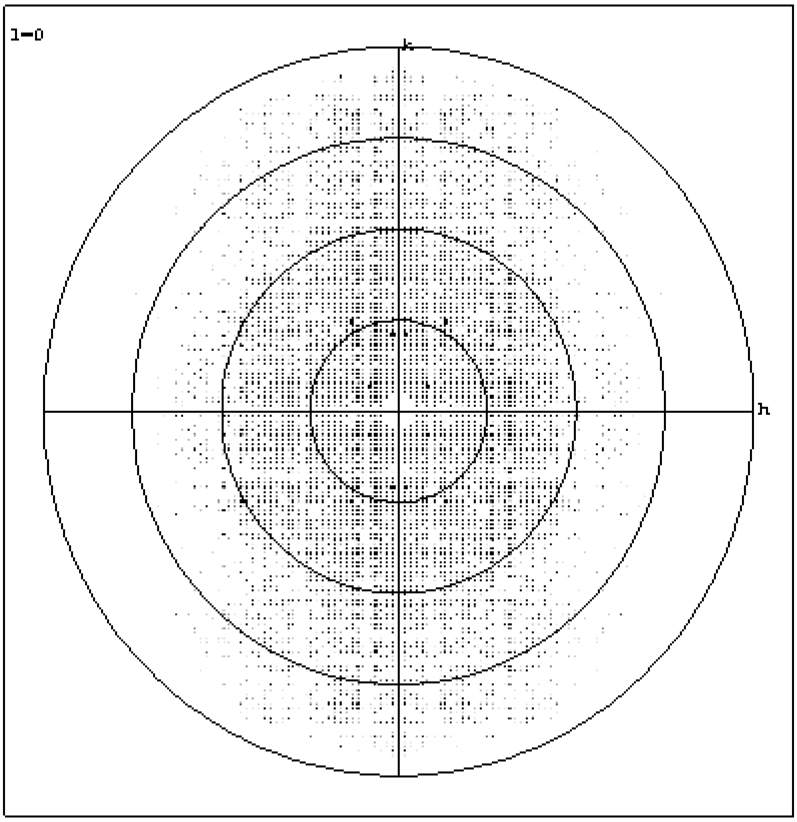
	MLC1	MLC2	CELL	MB	OIA
d(Å)	1.5	1.96	1.8	1.8	1.5
R/R-free(iso %)	11.7/14.0	17.3/22.6	20.2/25.1	22.1/28.8	28.4/31.3
R/R-free(aniso %)	11.6/13.9	15.1/20.6	14.3/18.0	20.6/27.0	18.0/24.7
B_{11}	-0.3	-3.4	9.1	5.6	-9.8
B_{22}	-0.3	-3.4	-4.2	1.0	-15.5
B_{33}	0.7	7.1	-4.8	-6.2	33.6
B_{12}	0.0	0.0	0.0	0.0	0.0
B_{13}	0.0	0.0	0.0	2.9	0.0
B_{23}	0.0	0.0	0.0	0.0	0.0

MLC1 - MLC native data collected at room temperature
MLC2 - MLC soaked in peracetic acid collected from frozen crystals
MB - Myoglobin collected from frozen crystals
OIA - Oxiindolyl-L-alanine complex of Tryptophanase
CELL - Cellulase

B_{11}, B_{22}, B_{33}, B_{12}, B_{13}, B_{23} are elements of anisotropic **B** tensor. $\mathbf{B} = 8\pi^2\mathbf{U}$

Figure 1:

a)



b)

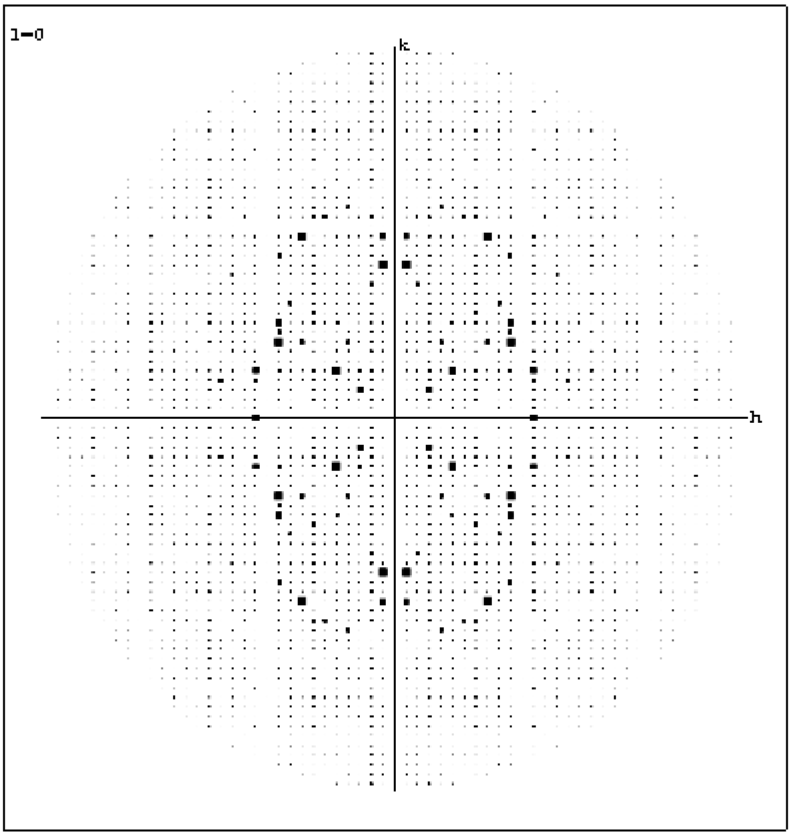


Figure 1. Cellulase data. a) 0.9A. circles starting from inside at 3.6, 1.8, 1.2 0.9A resolution b) 1.8A.

First Experience with Novel Microsource X-Ray Tube

A C Bloomer & U W Arndt
MRC Laboratory of Molecular Biology
Hills Road, Cambridge CB2 2QH, UK

Last month we took delivery of the first commercially-produced MicroSource X-ray generator from Bede Scientific Instruments Ltd (Durham, UK). This is a sealed tube, as distinct from the continuously pumped laboratory prototypes upon which development work has been done over the past several years. We present here the initial performance indications of this novel generator and some prospects for its future development. A more detailed report is being prepared for submission to J. Applied Crystallography.

Background

Macromolecular crystallographers seek X-ray beams where the highest number of photons illuminate the crystals whose diffraction they need to measure. Over the years, the most challenging crystals being studied tend to have smaller size and/or larger unit cells. A desirable X-ray source has high brightness and low cross-fire, combined with great stability and reliability. For most biological crystallographers, the early sealed tubes for laboratory production of X rays have long been overtaken by rotating anode generators, and more recently by centralised facilities for synchrotron radiation.

However, a new design of sealed X-ray tube for use in the laboratory specifically aims to meet the requirements in this field of crystallography for good collimation and utilisation of smaller crystals. Design features used in electron microscopes and similar devices have been used in the development of this low-power copper-target X-ray tube [1]. This has been optimised for use with focusing mirrors, chosen to maximise the solid angle of collection of the emitted X rays, thus achieving a high intensity at the sample. The additional requirement of small cross-fire in the beam imposes a geometrical constraint that the X-ray source should be very small in relation to the size of the sample. The present instrument, which has a typical source size of 20micron, is well-matched to samples of up to 300micron in size, positioned at 60cm from a specularly reflecting ellipsoidal mirror [1].

Comparison of different types of focusing mirrors for use with a microfocus X-ray tube [2] shows the advantages of ellipsoidal mirrors of very small diameter (typically less than 1mm). These can be manufactured by an electroforming technique. Data presented elsewhere [2] introduce the method of determining the performance of mirrors by their "insertion gain". This is the ratio of the intensity from an X-ray source delivered into a defined small area (comparable to the size of a specimen crystal) at the focus of the mirror when it is in position, to the intensity when the collimating mirror is removed. The expected insertion gain for a perfect system can be calculated for comparison with that achieved in practice [2].

Results from early prototype designs have led to the arrival on the market this year of a commercially-produced system.

Experimental Arrangement

The MicroSource tube which is now at LMB (Cambridge) has been operating at 25watts (50kV 0.5mA) and is mounted to enable its alignment to an early prototype of the MAR (18cm) diffractometer positioned so that the sample is at the optimum distance (60cm) from this source. A vacuum pipe covers much of this distance. Good diffraction data have been collected on this system from crystals of both lysozyme and also an antibody Fab fragment.

We have not yet completed a detailed comparative assessment of data obtained using this new installation with that from our standard laboratory rotating anode installation. However, we can already compare the characteristics of the incident X-ray beam at the position of a crystal sample. The rotating anode installation is a Nonius GX13 (big wheel) operating at 2,400watts (40kV 60mA) with a fine focusing cap (100micron nominal). This is used with a standard MAR (30cm) diffractometer equipped with double Franks mirrors and two pairs of slits to control the size and cross-fire of the beam at the crystal.

X-ray intensity is measured by an ionisation chamber positioned at or just behind the sample position. In some cases a small aperture (300micron) is carefully aligned on a goniometer head so that it is at the sample position. Measurements of the flux passing through this aperture are representative of the X-rays incident upon a crystal of size 300micron or less.

We take as our local standard of comparison the flux obtained with the MAR slits all set to 0.3mm and which passes through our 300micron aperture. Table I shows that the flux obtained from the GX13 tube running at 100* the power of the BEDE tube is less than 3 times greater. For use with larger crystals, the flux measurement (without the 300micron aperture) from the GX13 flux is 3.7 times greater, and if larger cross-fire can be tolerated (opening up all the MAR slits to 0.4mm) on bigger crystals the flux is 4.6 times greater from a GX13 than from the present BEDE tube operating at 1/100th of the power.

The closest comparison comes in the case where a reduced cross-fire is needed in the X-ray beam. When the MAR slits are all reduced to 0.2mm, the GX13 output exceeds that now obtainable from the BEDE tube by a factor of less than 2.

The expected life-time of this novel tube is several months, just as for other sealed X-ray tubes. Early indications of the present tubes are that the filaments survive for several weeks at least, but have not yet been tested for a complete life-time. The other question frequently raised about long-term usage concerns the rate of radiation damage to mirrors. Our calculations, which will be detailed elsewhere, indicate that the flux per unit area falling onto the small ellipsoidal mirror used with the BEDE tube is much less than that incident upon the first Franks mirror used with our rotating anode tube: the difference is a factor of about 5, within the range from 3 to 7 depending upon the exact geometry chosen for the ellipsoidal mirror. Experience with the Franks mirrors on our GX13 tube is that the first mirror usually needs replacing after about six months; by analogy one would expect the ellipsoidal mirrors to last for several times longer than this.

Future Prospects

Earlier ideas by one of us (UWA) of a rotating anode version of this novel generator, which would have increased the flux by a factor of about 5, have since been superseded by two other potential developments which both retain the great simplicity of the present MicroSource tube in comparison with other high brightness tubes.

a. Theoretical calculations indicate an insertion gain of 250-3000 from use of optimum ellipsoidal mirrors [2]. The gain presently achieved is only about 50, such that improvements in both the profile accuracy and the surface smoothness of mirrors can be expected soon to enable higher flux by a factor of 5-30. As the mirrors are readily inter-changeable, this can be a simple upgrade for any existing installations.

b. The maximum operating power of the BEDE tube is determined both by the size of the focal spot, and the efficiency of cooling of the target. We have operated the BEDE tube at 25 watts, taking a cautious view of the safe operating limits. Power loading of up to 100 watts should be readily achievable, with a consequent gain in flux by a factor of 4. This will depend both upon alternative methods of cooling the copper target within the tube [1] and upon further study of the effect of increased power loading upon the size of the source.

One major advantage of the MicroSource tube is its adaptability, with simple inter-change between different mirrors to match the beam to the needs of varying crystals. The ultimate limit to the flux at the sample depends upon which configuration is needed. However, the present system already delivers onto a small sample a well collimated beam of intensity comparable to that obtained from some existing synchrotron stations using bending magnets.

REFERENCES

[1] Arndt, U. W., Long, J. V. P., and Duncumb, P. (1998). A micro-focus X-ray tube with focusing collimators. J. Appl. Cryst.. In press.

[2] Arndt, U. W., Duncumb, P., Long, J. V. P., Pina, L., and Inneman, A. (1998). Focusing mirrors for use with microfocus X-ray tubes. J. Appl. Cryst.. In press.

Table I

Intensity measurements, using an ionisation chamber placed at the sample position, for two X-ray tubes, with and without a limiting aperture of 300micron. For details, see text.

Data:	**Normalised**		**Raw figures (arbitrary units)**	
300micron aperture in position:	**NO**	**YES**	**NO**	**YES**
BEDE tube, mirror ME-II	1.0	1.0	0.44	0.44
operating power of generator 25w				
GX-13, slits set to 0.2mm	1.9	1.9	0.83	0.84
0.3mm	3.7	2.7	1.62	1.21
0.4mm	4.6	2.9	2.02	1.28
operating power of generator 2400w				