CCP4 NEWSLETTER ON PROTEIN CRYSTALLOGRAPHY

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

Number 40 **March 2002**

Contents

Editors: Charles Ballard and Maeri Howard-Eales Daresbury Laboratory, Daresbury, Warrington, WA4 4AD, UK

News from CCP4: Spring 2002

Charles Ballard, Martyn Winn, Alun Ashton, Peter Briggs, Maeri Howard Eales, Pryank Patel

1. Staff changes

The newest member of the CCP4 staff, **Pryank Patel**, has left the buzz of London for the (hopefully!) slower pace of life at Daresbury Laboratory.

Pryank Patel started in January 2002 on a 3-year post as part of a major new project funded by the EU to expand the activities of the European Bioinformatics Institute and its collaborators. Pryank will be working on expanding the role of Data Harvesting in the CCP4 suite, easing structure deposition from CCP4 and enhancing the existing validation tools. Pryank recently finished an MSc course in Bioinformatics at Birkbeck College, before moving north to join the Daresbury team.

Also **Maeri Howard Eales**, who may be familiar to some of those that attended this year's Study Weekend, took over the reigns from the retired David Brown as the Administrative Assistant. As well as taking on the responsibilities that she has inherited from Dave, Maeri will be producing a smaller newsletter twice a year, developing the commercial web pages on the CCP4 website and using her experience from the Study Weekend by attending upcoming conferences, such as the ACA in Texas in May of this year, where she is looking forward to some good old fashioned American cooking!

2. Workshops and Conferences 2001/2002

It has been a busy year for CCP4 on the conference front as you can see by the amount of conferences on our schedule. This does not look as if it will slow down, as the summer months appear to have more than enough conferences to keep all of us busy. Highlights of conferences passed as well as ones upcoming for the summer can be found in more detail in the following paragraphs.

In July CCP4 was again represented at the American Crystallographic Association annual meeting in Los Angeles. Charles Ballard, Peter Briggs and MOSFLM developer Harry Powell were in attendance at the CCP4 exhibition stand, demonstrating the latest versions of the software to an unsuspecting American audience.

In particular, many people asked about using REFMAC5 and the graphical user interface CCP4i, and as always MOSFLM was a popular attraction. The new Windows NT port of the suite was used throughout for demonstrations and performed brilliantly.

CCP4 had a stand at the 20th European Crystollographic Meeting held in Krakow, Poland in August 2001. The stand was manned by Martyn Winn and Harry Powell, with guest appearances by Garib Murshudov, Alexei Vagin, Phil Evans and Eleanor Dodson. CCP4 software was also prominent in the microsymposia with talks on REFMAC by Garb Murshudov, on ACORN by Eleanor Dodson, and on BEAST by Randy Read.

The main meeting was followed by a satellite meeting entitled "Protein Crystallography beyond 2000: High Throughput and High Quality" which covered future directions of protein crystallography methodology. CCP4 was represented by talks from Eleanor Dodson, Garib Murshudov and Martyn Winn.

The beginning of the New Year was not a quiet time for those at CCP4 as The University of York was host to the annual CCP4 Study Weekend the 4th and 5th of January. This year's topic, High-Throughput Structure Determination, proved to be one of the most popular topics, attracting over 450 delegates to the two day seminar.

Hoping to extend the Study Weekend further afield than just to those that came to York, for the first time in Study Weekend's history, the talks were webcast live. Alun Ashton, Conference Organiser, said "We have tried to make the Study Weekend as accessible as possible to everyone and we felt that this was the next logical step. We researched into whether it was possible and we found that it was not as difficult to set up as we had thought. We hope to use streaming wherever possible in the future and will announce more on this on our website as it develops".

CCP4 would like to thank the scientific organisers - Robert Esnouf (Oxford), Dave Stuart (Oxford) and Keith Wilson (York) - as well as the speakers who braved the cold and ice of the North to present their talks. As usual, staff from Daresbury (Alun Ashton, Maeri Howard Eales, Pat Broadhurst, Sue Waller and Allison Mutch) made sure that registration was as painless as possible and that everyone got the help they needed at the conference. All of the proceedings from the weekend will be published in the Acta Cryst D later in the year.

The upcoming months will prove to be yet another busy one for the members of CCP4 that are attending a wide variety of conferences around the world. In March, we will be attending the most local of the conferences, the annual BCA Conference to be held at the University of Nottingham. Though CCP4 will not be taking a stand at the conference, a session on CCP4 will be chaired by Harry Powell. More information on this can be found at the CCP4 website.

Cowboy hats and boots will be out in force when CCP4 attend the ACA meeting in historic San Antonio, Texas. The meeting, held the 25th to the 30th of May, will be a chance for members of CCP4 to meet with their American users. In addition to this, CCP4 has been asked to host a workshop as an introduction to CCP4 during the meeting. Check the http://www.hwi.buffalo.edu/ACA.html ACA website for more details.

And not to be forgotten, CCP4 will be attending the IUCR meeting held in Geneva the 6th - 15th of August. Details on this are just being confirmed and once more information is available, it will be posted on the CCP4 website. However, you can check the http://www.iucr.org IUCr website for more comprehensive information on the conference.

3. Other News

Congratulations: A well deserved congratulations goes out to Eleanor Dodson who since May 2001 is known as Prof. Dodson. Eleanor plays an important role in the development of CCP4 and all of us at CCP4 wish her well in her new role.

New Editor: As mentioned in the last newsletter, Peter Briggs has stepped down as the Editor of this Newsletter and the role is now being filled by Charles Ballard. If you have any questions or suggestions about the newsletter, please feel free to contact me at **c.c.ballard@ccp4.ac.uk**

Developments with CCP4i

Peter Briggs
CCP4

Introduction

The last officially released version of the CCP4i graphical user interface was 1.2.7, which was included as part of CCP4 4.1, and since that release development of the interface project has continuted apace. This article outlines the changes that have happened so far, and the new features that people can expect to see in the next release.

Personnel Changes and Webpages

CCP4i was originally developed by Liz Potterton, who worked on the project from its inception in 1997. Following the last release, Liz relinquished control of CCP4i in order to concentrate on the CCP4 Molecular Graphics project (see the CCP4 3D Molecular Graphics webpages). Long-term users of the interface will agree that the development of CCP4i is a great achievement, and we wish Liz every success with the new project.

As a result Liz's move, in April 2001 responsibility for the development of the graphical user interface passed to the Daresbury CCP4 group, with Peter Briggs as the new CCP4i project manager.

The symbolic move of CCP4i to DL was also accompanied by new CCP4i webpages, which can be accessed via the main CCP4 website or else directly at http://www.ccp4.ac.uk/ccp4i_main.php. This page contains lots of useful links - including information on problems with CCP4i. Suggestions, bug reports and other comments about CCP4i should now be sent to the standard CCP4 address at ccp4@ccp4.ac.uk.

CCP4i Developers Workshop

As a part of handing over the reigns of the CCP4i project, in April Liz organised a one-day ``CCP4i Developers Workshop'' in York, with Liz and Peter acting as tutors. The aim of the workshop was to introduce programmers to the basics of writing task interfaces for CCP4i.

Notes from the workshop are available on the web, at http://www.ccp4.ac.uk/ccp4i/developers.html#workshop. Currently there are no plans to rerun the workshop, but this could change if there is sufficient interest - please let us know.

Changes for CCP4i 1.3

The next revision of CCP4i will be version 1.3, which will be included in the next release of the CCP4 suite sometime in early 2002. Many of the changes from 1.2.7 are relatively minor, basically fixing bugs and consolidating earlier changes. Also it has taken the Daresbury group a little while to get up to speed with the CCP4i project.

However there are still a number of significant developments planned for CCP4i 1.3, and these are detailed below.

New and Updated Interfaces

Since March 2001, new interfaces have been released for the CCP4 programs ANISOANL, TLSANL, and OASIS. These interfaces will now officially be incorporated into CCP4i. A new interface is also planned for the accessible surface area program AREAIMOL.

In addition the major new programs planned for inclusion in CCP4 4.2 will have corresponding interfaces. These include ACORN (*ab initio* procedure for the determination of protein structure at atomic resolution), BEAST (maximum-likelihood molecular replacement program) and WHAT_CHECK (the subset of protein verification tools from the WHAT IF program).

A number of other minor changes have been made to existing interfaces in an attempt to improve ease of use, for example the Scalepack2mtz and Dtrek2mtz tasks have been combined into a single task interface to import scaled data. Also there has been some reorganisation of the tasks and modules menus, to improve the access to relevant tasks in some of the modules.

New Utilities

It is intended that the MapSlicer application should become the default viewer for CCP4 map files in CCP4i 1.3. MapSlicer offers interactive display of contoured 2D sections through density maps, as well as can displaying map header information. A prototype version of MapSlicer was released in CCP4 4.1 (see the article in the previous newsletter) but this new version has been substantially rewritten to add extra features and to improve portability.

The utility for installing new task interfaces has also been substantially upgraded. The aim is to provide a robust mechanism for installing and tracking ``third-party'' interfaces - that is, interfaces provided for non-CCP4 software by its authors. (An example of this is the ARP/wARP interface written by Tassos Perrakis and previewed in an earlier article.)

For users, the new utility offers options to install, review and uninstall these interfaces quickly and easily. New interfaces can also be installed either ``locally'' (so only the person installing the task can use it) or ``publically'' (so the new task is available to all users on the system).

For developers there is a simple mechanism for version control and options to run external scripts to perform checks on the system before installing the task. It is also intended that the installer/uninstaller will run from the command line, so that it can be incorporated into Makefiles or installation scripts for other packages.

New Features

Some aspects of CCP4i have been altered with the aim of enhancing usability of the interface, for example:

- **Job Selection**
  The selection behaviour of the jobs database has been changed - now only the last selected job is highlighted. Mutliple jobs can still be selected using ``click-and-drag'', and by the use of the control and shift keys. This behaviour is more ``Windows''-like.

- **Long MTZ label Menus**
  An occassional complaint in the past has been that the menus of MTZ labels can get so long that the labels at the bottom of the list fall off the screen and are thus inaccessible. Long lists are now broken into multiple columns - the default length of each column is 25 items, but this number can be changed in the *System Administration* options.
- **Data Harvesting**
  In earlier versions of CCP4i the default setting has been to have Data Harvesting turned off. This has been changed so that by default harvesting information is always written to the current project directory. (For more information on Data Harvesting see the Data Harvesting documentation, which is part of the CCP4 html documentation.)

Documentation

A major change from 1.2 to 1.3 is the addition of inline ``doc-comments'' in the CCP4i source code. These comments can be extracted and turned into html documentation of the CCP4i code, and both the commented code and the extracted documentation will be included in the next release. We hope that this will be useful for external programmers wishing to make CCP4i work more easily with their programs.

Future Plans - beyond 1.3

The majority of changes planned for CCP4i 4.2 are intended to consolidate the existing interface. A number of longer-term projects are also envisaged:

- **MTZ viewer**
  The new MTZ libraries will impose a more formal hierarchical structure on reflection data stored in MTZ files, and it is intended to create a hierarchical viewer which will reflect this structure and make it easier to view and select datasets and columns.
- **Improvement tools for Data Harvesting and Validation**
  As part of a joint CCP4/EBI post it is intended to provide improved tools under CCP4i for Data Harvesting and structure validation, for example by offering an interface for reviewing harvesting files.
- **Extended ``Project Database'' and Automation**
  Currently CCP4i ``projects'' only store a history of the jobs run, with lists of input and output parameters and files for each task. By extending the definition of a project to include other data (for example, sequence information, molecular weight, number of molecules in the asymmetric unit and so on) it should be possible to speed up use of the interface for routine tasks by filling in many of the input fields automatically. This would also facilitate the automation of sets of tasks, something which is not currently possible in CCP4i.

Development of the CCP4 software library

Martyn Winn, Charles Ballard and Eugene Krissinel December 2001

Aims

The CCP4 software suite is based around a library of routines which cover common tasks, such as file opening, parsing keyworded input, reading and writing of standard data formats, applying symmetry operations, etc. Programs in the suite call these routines which, as well as saving the programmer some effort, ensure that the varied programs in the suite have a similar look-and-feel.

Over the past 12 months, there has been a major effort to re-write much of the CCP4 library. The aims are:

- To implement a better representation of the underlying data model. For example, Eugene Krissinel's mmdb library acts on a data structure which represents the various levels of structure of a protein model. The new MTZ library encapsulates the crystal/dataset heirarchy that is increasingly being used by programs.
- To maintain support for existing programs. In particular, the existing Fortran APIs will be maintined, although they will now often be only wrappers to functions in the new library. It is hoped that many existing programs will be migrated to using the new library directly.
- To provide support for scripting. It is possible to generate APIs for Python, Tcl and Perl automatically from the core C code. Thus, much of the standard CCP4 functionality wil be available to scripts used e.g. in ccp4i or the molecular graphics project.

This incremental approach, maintaining the existing suite while improving the underlying code, puts constraints on what is possible, but is considered more appropriate for a collaborative project like CCP4.

Major components

mmdb

The "mmdb" library is designed to assist CCP4 developers in working with coordinate files. The major source of coordinate information remains the PDB files, although more information is becoming available in mmCIF format. The "mmdb" library will work with both file formats plus an internal binary format portable between different platforms. At the level of the library's interface function, there is no difference in handling different formats.

The "mmdb" library provides various high-level tools for working with coordinate files, which include not only reading and writing, but also orthogonal-fractional coordinate transforms, generation of symmetry mates, editing the molecular structure and some others. More information can be found on the mmdb project pages (http://msd.ebi.ac.uk/~keb/cldoc/).

cmtzlib

In the new formulation, reflection data is still held in a table format, but columns of data are arranged in a heirarchical manner. From the top down, the heirarchy is:

File -> Crystal -> Dataset -> [Datalist] -> Column

A `Crystal' is essentially a single crystal form, a `Dataset' is a set of observations on a crystal, and a `Datalist' (not yet implemented) is a grouping of associated columns. Note that the `Project' used in Data Harvesting (Newsletter #37) is now simply an attribute of the crystal.

The MTZ file format has been extended slightly to record this heirarchy. I have written a C function library to read/write these extended MTZ files, and to manipulate a data structure representing the above data model. Some functions are derived from Jan-Pieter Abraham's solomon code, though they have been substantially altered (and therefore any problems are of my creation!).

One consequence of this formulation is that columns can now be identified in terms of their crystal or dataset, e.g.

```
LABIN  FP=NATIVE/FTOXD3 SIGFP=NATIVE/SIGFTOXD3
```

I have also written a Fortran API to the C library, which mimics the existing mtzlib.f. It should be possible to migrate existing Fortran programs to use this API with few/no changes. It is expected however, that future applications will use the core C functions which give better access to the data structure.

Some information on cmtzlib is available from http://www.ccp4.ac.uk/martyn/cmtz.html, but further developments will be integrated into the CCP4 library development.

cmaplib

Charles Ballard has written a C language library for the reading and writing of CCP4 format map files. A Fortran API mimics the existing maplib.f.

In a parallel development, there have been recent efforts to bring the MRC format for three-dimensional electron microscopy maps in to line with the closely-related CCP4 format map. For our part, we are considering suggested changes to the CCP4 map format which will be useful for EM applications. Some details can be found on the IIMS home page (http://msd.ebi.ac.uk/iims.html).

csymlib

The old implementation of symmetry held tabulated information in a manually-produced file symop.lib, together with other information distributed amongst routines in symlib.f (e.g. real space asymmetric unit limits in subroutine SETLIM). This set-up works in most cases, but was error-prone and difficult to maintain.

In the new formulation, symop.lib is replaced by another data file which is automatically generated. This is currently done using a short program which uses functions from sgtbx (part of the Computational Crystallography Toolbox, http://cctbx.sourceforge.net) . The new data file is more likely to be error-free, and is also more complete, in that many non-standard settings can be included easily. The new data file contains most quantities of

interest, and only a few pieces of tabulated data are retained in the code (e.g. specifications of centric and epsilon zones).

I have written a new C library of functions to manipulate this symmetry information. When a spacegroup is identified by its name, number or operators, all the information connected with that spacegroup is loaded into memory, where it can be accessed easily. Wrapper functions mimic the old symlib.f routines.

ccp4_unitcell

The CCP4 library contains many routines concerned with the unit cell, in particular to do with tranforming between orthogonal and fractional coordinates. I have written C functions to perform the basic manipulations, with Fortran-callable wrappers to mimic old routines in rwbrook.f and elsewhere.

core library

Charles Ballard has re-written most of the core code. The files concerned are:

```
library_file.c        )
library_err.c         ) replacing library.c
library_utils.c       )
ccp4_general.c       ) to give CCP4 look-and-feel
ccp4_program.c       )
ccp4_parser.c          parser functions from Pete Briggs
ccp4_diskio.c         replacing diskio.f
```

These functions mostly support other code, but may be of use to application developers.

Support for scripting languages

I have used SWIG to generate interfaces to the python, perl and tcl scripting languages. These interfaces exist as automatically-generated wrapper functions, for example the files

```
cmtzlib_python_wrap.c
cmtzlib_perl_wrap.c
cmtzlib_tcl_wrap.c
```

provide cmtzlib functions which can be called from python, perl and tcl respectively. These wrapper functions, together with the main library code, are compiled into shared libraries which can be loaded by the scripting language:

```
ccp4module.so         loaded by       from ccp4 import *
ccp4_pl.so            loaded by       package ccp4_pl;
ccp4_tcl.so           loaded by       load ./ccp4_tcl.so ccp4_tcl
```

Much of the library functionality can be made available to the scripting language with virtually no effort. Access to complex datatypes requires a little more effort, as does wrapping C++ code, and this will be provided as required.

Other features

The new CCP4 library can be used in a variety of ways, depending on the functionality required and the language chosen. Therefore, facilities will be provided to do partial compilation, so that only the functions appropriate to the application are provided in the library. In any case, care is being taken to establish appropriate namespaces, to avoid conflicts in any programming environment. Many of the existing Fortran routines are being

left unchanged, for example certain routines in ccplib.f and all of plot84driver.f. This is either because the routines are Fortran-specific, or simply because C equivalents have not been coded yet.

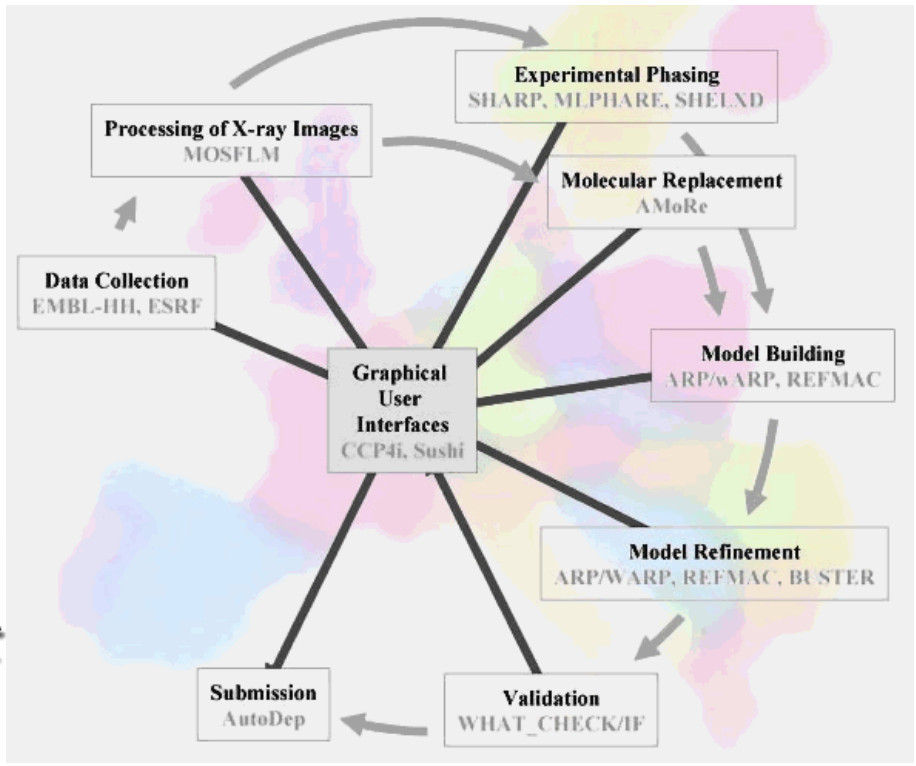
Acknowledgements

cmtzlib was originally based on Jan-Pieter Abrahams' functions in the program solomon. These have been substantially altered and added to, and so all problems are my creations, but I am grateful for the excellent starting point.

The formulation of this library has benefited from many discussions with Kevin Cowtan (who also provided some core functions), Eugene Krissinel, Airlie McCoy, and all members of the Daresbury team (Alun Ashton, Peter Briggs, Charles Ballard).

AUTOSTRUCT

Report by Alun Ashton, CCP4.



AUTOSTRUCT is the acronym for an EU supported project to bring together the efforts of many of the major European software developers for X-ray crystallography. The major objective is to make the procedures more effective, in terms of using more powerful algorithms to solve, refine and validate structures. The software should not only be as transparent as possible to the end user, but also assist in the case of user inexperience.

The project is barely a year old but much has already been achieved. DNA (DNA Not AUTOSTRUCT, http://www.dna.ac.uk) is an AUTOSTRUCT 'spin off' collaboration primarily between Daresbury Laboratory, MRC Laboratory of Molecular Biology, Cambridge and the ESRF. The primary aim is to facilitate communication between beamline data collection and control software and the data processing software. This will soon lead to real improvements for users on synchrotron beamlines. The ultimate aim of automating data collection has been helped by the adoption of the open XML standard as a communication protocol between software. This will undoubtedly enable any future expansions and reduce the time scale of implementing the procedure at other sites or beamlines.

Moving between packages such as SHARP, SHELX and CCP4 can occasionally be convoluted due to the evolution of individual standards for recording and reporting data formats (for example heavy atom positions). But now agreement on a common standard that will be implemented in all the mentioned packages is at an advanced stage. A route of easily accessing SHELXD within CCP4i is also under development.

Whether you derive your first map from direct methods, isomorphous replacement techniques or by molecular replacement (e.g. by the ever improving AMoRe package also receiving funding for developments under AUTOSTRUCT), automated model building and refinement makes life a whole lot easier. Undoubtedly one of the leaders in this field is the ARP/wARP procedure. As this package is heavily dependent on other programs such as REFMAC* a lot of work is being done to co-ordinate releases of these packages. Also ARP/wARP and CCP4 have collaborated to create an easily installable interface to run within CCP4i. The interface will exploit the full functionality of the ARP/wARP package and will be available soon with version 5.2 of ARP/wARP.

When thoughts turn towards structure validation and deposition you should probably also start thinking about the next project. AUTOSTRUCT can't help you choose or find the next project but making validation and deposition easier is a goal! In one recent AUTOSTRUCT meeting not only was agreement reached on distributing the popular WHAT_CHECK validation package with future releases of CCP4 but also an interface within CCP4i was developed. Regular readers of the CCP4 newsletter will already be familiar with the concepts of Data Harvesting and work is continuing to expand and facilitate this process to enable fast and easy structure deposition.

For more information on AUTOSTRUCT keep an eye on the web pages at www.autostruct.org. The project is sponsored under the EU program: Quality of Life and Management of Living Resources. The project co-ordinator is Prof. Keith Wilson, York Structural Biology Laboratory.

List of Partners:

1 YSBL, University of York, http://www.ysbl.york.ac.uk

2 Mosflm, LMB-MRC, http://www.mrc-lmb.cam.ac.uk/harry/mosflm

3 SHELX, Goettingen University, http://shelx.uni-ac.gwdg.de/SHELX/

4A EMBL Hamburg Outstation, http://www.embl-hamburg.de

4B EMBL European Bioinformatics Institute, http://www.ebi.ac.uk

5 AMoRe, CNRS, http://www.ccp4.ac.uk/ccp4/html/amore.html

6 WHAT_CHECK, http://www.cmbi.kun.nl/whatif

7 CCP4, http://www.ccp4.ac.uk

8 NKI, http://www.arp-warp.org

9 GlobalPhasing, http://www.globalphasing.com

The Clipper Project

Author: Kevin Cowtan, Department of Chemistry, University of York

There are currently two major pressures on crystallographic computing:

- Increased automation to increase throughput in line with genomics applications.
- Increased data complexity, as data from more sources is combined and carried through the whole calculation to solve more difficult problems.

The Clipper project is an initiative to address these pressures.

The aim of the project is to produce a set of object-oriented libraries for the organisation of crystallographic data and the performance of crystallographic computation. The libraries are designed as a framework for new crystallographic software, which will allow the full power of modern programming techniques to be exploited by the developer. This will lead to greater functionality from simpler code which will be easier to develop and debug.

Object oriented programming

The evolution of high level programing may be very imprecisely caricatured as follows:

- Early high-level languages (Fortran): There is no structure to data or code. Variables are not grouped in any way or specially associated with any code.
- Structured programming (C, Pascal, etc): Data structures are introduced, grouping related variables into compound objects, which can completely specify the state of some object.
- Object-oriented programming (Smalltalk, C++, etc): Data structures may now contain code, and now describe not only the state of the object, but also its behavior and interactions. (In the general case, all code becomes part of an object.)

Clipper is object-oriented. The main benefit of this approach is that code becomes much more reusable, since objects are self-contained, and may be reused, rewritten, or replaced without affecting other code. Additionally, the organisation of the code and data is generally much clearer.

Clipper objects

Clipper defines a wide range of objects. These fall into a number of groups, including:

- Crystallographic objects: : Cell, Spacegroup, Metric, R/T operator, etc.
- Data objects: : Reflection data, Crystallographic map, Non-crystallographic map.
- Method objects: : FFT arrays, Resolution functions, Conversion objects, Import/export objects

The one object type not addressed is the coordinate object: this is a substantial task and is addressed by the Dr Eugene Krissinel with the CCP4 'MMDB' project.

Some of the objects will be discussed in more detail:

Crystallographic objects:

These implements the fundamental properties of a crystal.

The Cell object:

> This object describes a unit cell. It holds the cell parameters, and derived information including coordinate conversion matrices and metrics. Any cell object may be used to convert coordinates between orthogonal and fractional forms, and calculate distances in real space and resolutions in reciprocal space.

The Spacegroup object:

> This object describes a spacegroup, using information from the 'cctbx' library of Dr Ralph Grosse-Kunstleve. It can be use to generate symmetry coordinates and reflections, phase shifts, centricity, asymmetric units, and so on.

Data objects:

These hold actual data. They are written as templates which can hold whatever type of data the developer requires.

The reflection data object:

> It is commonly necessary to store several related items of reflection data. Therefore this object is split into two parts; a parent object which holds a list of Miller indices and related data, and then several data objects which hold the actual data associated with each Miller index. The data objects can hold data of arbitrary types: these types will usually consist of several values. For example, a structure factor magnitude and its variance, or all four Hendrickson-Lattman coefficients, are usually held in a single data object.
>
> To the user, the data appears to cover the whole of reciprocal space, however in practice only an asymmetric unit is stored. Data is transformed about reciprocal spaces as required. When a new data type is defined, its behavior under transformation is also be defined so that this mapping can be performed.

The crystallographic map object:

> This object also implements crystallographic symmetry, and also cell repeat, in a manner which is transparent to the user. It may also hold arbitrary data types: common examples would include bits, real values, complex values, or orthogonal or fractional gradients.

The non-crystallographic map object:

> This object is used for map data which does not have symmetry or cell repeat, for example an NCS averaging mask.

Method objects:

These are used to provide additional functionality commonly required in crystallographic calculations. Examples include:

FFT map:

> This object holds data which may be represented in either real or reciprocal space. The data may be accessed in either form, and may be transformed between spaces as required.

The resolution function evaluator:

> This object creates an arbitrary function of position in reciprocal space, by optimising the parameters of some basis function in order to minimise some target

function. This is an extreme generalisation of the idea of 'resolution bins', and can be used for anything from $<|F|^2>_s$ to sigma-a and beyond.

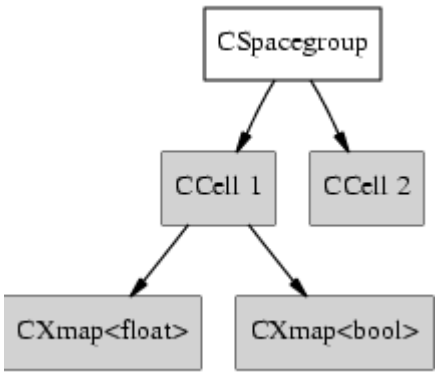
Import/export objects:

The only import/export object implemented so far is the MTZfile object. This is an object which stands as a proxy for an external file, and can transfer data between the Clipper objects and the file.

Data organisation

In order to handle data from multiple sources (and in particular multiple crystals, for phasing, multi-crystal averaging or refinement), the data must be well organised. Clipper provides an optional mechanism for data organisation by providing 'container' versions of each object. A container is an object which can contain other objects. Any object may contain any number of other objects of any type.

The data organisation can therefore be drawn as a tree-structure, with the top-level container as the root. For example, a Spacegroup might contain two cells, one of which contains two maps of different types, as follows:



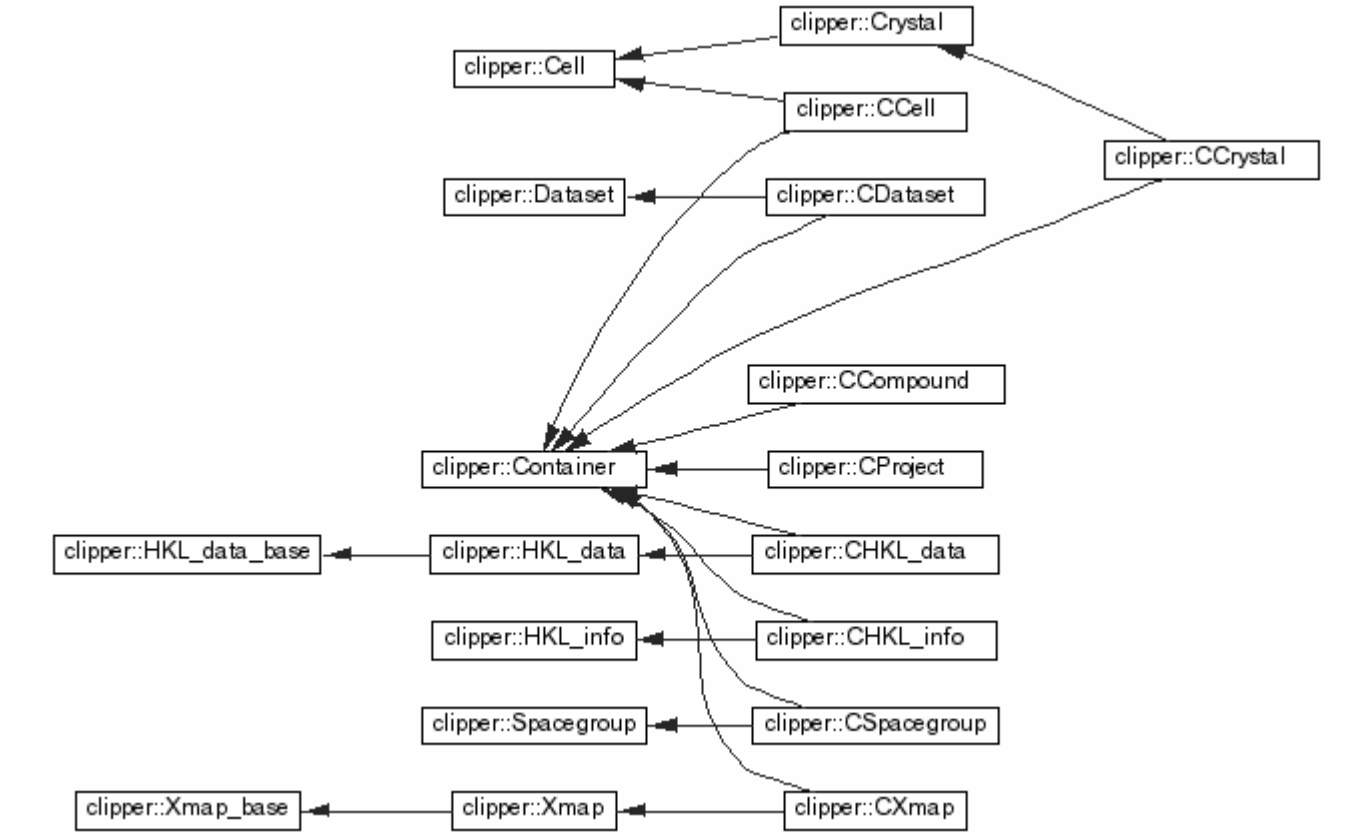
In practice this structure is most often used to describe the relationships between crystals, datasets, and reflection data.

Scripting

Automation of crystallographic tasks depends on being able to communicate between successive tasks, and by being able to execute control code to activate tasks and make protocol decisions. This functionality is provided through a scripting interface. A Python interface will be provided through the boost.python library. It is possible that interfaces to C and a range of scripting languages will be provided through other means. All the data and the full functionality of the methods will be available from the scripting layer, allowing full automation and full communication between tasks.

Eventually the individual programs should disappear, rather exposing their functionality directly to the scripting layer. At the same time, data will have to move from traditional files into a database, so that each task has immediate access to all the information currently available.

Partial class hierarchy (as of 31/10/2001)



Acknowledgments:

I would like to thank Ralph Grosse-Kunstleve, Airlie McCoy, Eugene Krissinel, Jan Zelinka and the CCP4 staff for their many and varied contributions to this effort. I also acknowledge the support of the Royal Society for this work.

Clipper stands for 'Cross-crystal Likelihood Phase Probability Estimation and Refinement', which is what I hope to use it for.

References:

- 'Clipper' code and documentation (K. Cowtan): http://www.ysbl.york.ac.uk/~cowtan/clipper/clipper.html
- 'mmdb' coordinate library (E. Krissinel): http://msd.ebi.ac.uk/~keb/cldoc/
- 'cctbx' crystallography toolbox (R. Grosse-Kunstleve): http://cctbx.sourceforge.net

FFFEAR

Author: Kevin Cowtan, Department of Chemistry, University of York

`fffear' is a package which searches for molecular fragments in poor quality electron density maps. It was inspired by the Uppsala `ESSENS' software (Kleywegt+Jones, 1997), but achieves greater speed and sensitivity through the use of Fast Fourier transforms and a mixed bag of mathematical and computational approaches (Cowtan, 1998). Currently, the main application is the detection of helices in poor electron density maps (5.0A or better), and the detection of beta strands in intermediate electron density maps (4.0A or better).

It is also possible to use electron density as a search model, allowing the location of NCS elements. Approximate matches may be refined, and translation searches may be performed using a single orientation.

The program has also been used to solve phased molecular replacement problems, and to locate missing NCS molecules, although this is not what it was designed for and the results may vary.

The program takes as input an mtz file containing the Fourier coefficients of the map to be searched, and a pdb file (in an arbitrary crystal cell) or map. A `fragment mask' is generated to cover the fragment density, and orientations and translations are searched to find those transformations which give a good fit between the fragment density and map density within the fragment mask.

The program has been highly optimised (Cowtan, 2001) using reciprocal-space rotations and grid-doubling FFT's, and crystallographic symmetry (Rossman+Arnold, 1993) giving 4-50 times speed improvement over the results published in 1998. The speed of the calculation is almost independent of the size of the model, thus the program may also be used for molecular replacement calculations where weak phases are available.

It is also possible to use the program to perform a naive Maximum Likelihood search by using a properly weighted target density and weighted mask. The target density and weighted mask are calculated by a massive search over matching fragments in the PDB. Currently only a 9-resdiue helix is available as a ML target.

Target function

The target function is a masked (weighted) mean squared difference between the search density and the current electron density map. As a result, the search function is sensitive to both peaks and voids in the search density. The search density and the map must however be carefully scaled; this is one of the most complex parts of the `fffear' program.

Let the target function be called $t(x)$. The fragment density is $p_f(x)$, and the corresponding fragment envelope is $e_f(x)$. Then the target function may be formed from the sum of the mean-squared difference in density between the offset fragment and the map:

$$\begin{aligned} t(x) &= \sum_y \epsilon_f(y) \left[\rho_f(y) - \rho(y-x)\right]^2 \\ &= \sum_y \epsilon_f(y)\rho_f^2(y) - 2\epsilon_f(y)\rho_f(y)\rho(y-x) + \epsilon_f(y)\rho^2(y-x) \end{aligned}$$

Note that in the expansion the first term is independent of *x* and so is only calculated once, whereas the second two terms are convolutions and may therefore be efficiently calculated in reciprocal space.

$$t(x) = \sum_y \epsilon_f(y)\rho_f^2(y) + \frac{1}{V}\mathcal{F}\left[\mathcal{F}^{-1}[\epsilon_f(x)]\mathcal{F}^{-1}[\rho^2(x)]^* - 2\mathcal{F}^{-1}[\epsilon_f(x)\rho_f(x)]\mathcal{F}^{-1}[\rho(x)]^*\right]$$

F represents the Fourier transform, F^{-1} the inverse Fourier transform, and * complex conjugation.

If the Fourier coefficients of the density and squared density are pre-calculated, then the translation function for a fragment in multiple orientations may be calculated by 3 Fast Fourier Transforms (FFTs) per orientation.

The Fourier transform of the search molecule and mask may also be pre-calculated on a fine grid in reciprocal space, and rotated values calculated by interpolation. By this means, the search is reduced to a single FFT per orientation.

References

http://www.ccp4.ac.uk/dist/html/fffear.html

Berman H. M., J. Westbrook, Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., Bourne P. E. (2000) Nucleic Acids Research 28, 235-242. The Protein Data Bank

K. Cowtan (1998), Acta Cryst. D54, 750-756. Modified phased translation functions and their application to molecular fragment location.

Cowtan K. (2001) Acta Cryst. D57, 1435-1444. Fast Fourier feature recognition

Kleywegt G. J., Jones T. A. (1997) Acta Cryst., D53, 179-185. Template convolution to enhance or detect structural features in macromolecular electron-density maps.

Rossman M. G., Arnold E. (1993) International Tables for Crystallography Volume C, Section 2.3: Patterson and molecular replacement techniques (Kluwer Academic Publishers).

ACORN - a flexible and efficient *ab initio* procedure to solve a protein structure when atomic resolution data is available

Yao Jia-xing

Department of Chemistry,
University of York,
Heslington,
York, YO10 5DD, U.K.
yao@ysbl.york.ac.uk

Introduction

A number of protein structures have been solved and/or refined at atomic resolution since strong X-ray sources from synchrotron radiation and modern data collection techniques are available. With atomic resolution data ACORN [1] can solve a protein structure from a small fragment as little as 5% or even 1% of of the scattering matter of the unit cell.

The starting fragment can be found from various sources according to the features of the structure to be determined such as single random atom, heavy atoms (Sulphur or heaver), Alpha Helices or motif from other structures in Protein Data Bank (PDB). Since the size of a fragment is very small it is easy to find out a motif from PDB which is growing larger quickly.

ACORN is divided into two parts: ACORN-MR and ACORN-PHASE. ACORN-MR is a fragment handling and generating a number of sets of initial phases with weights. ACORN-PHASE is a phase development and refinement procedure to obtain the best set of phases indicated by Correlation Coefficient (CC). Normally a model can be build up automatically from ACORN map by ARP/wARP in CCP4 or by QUANTA.

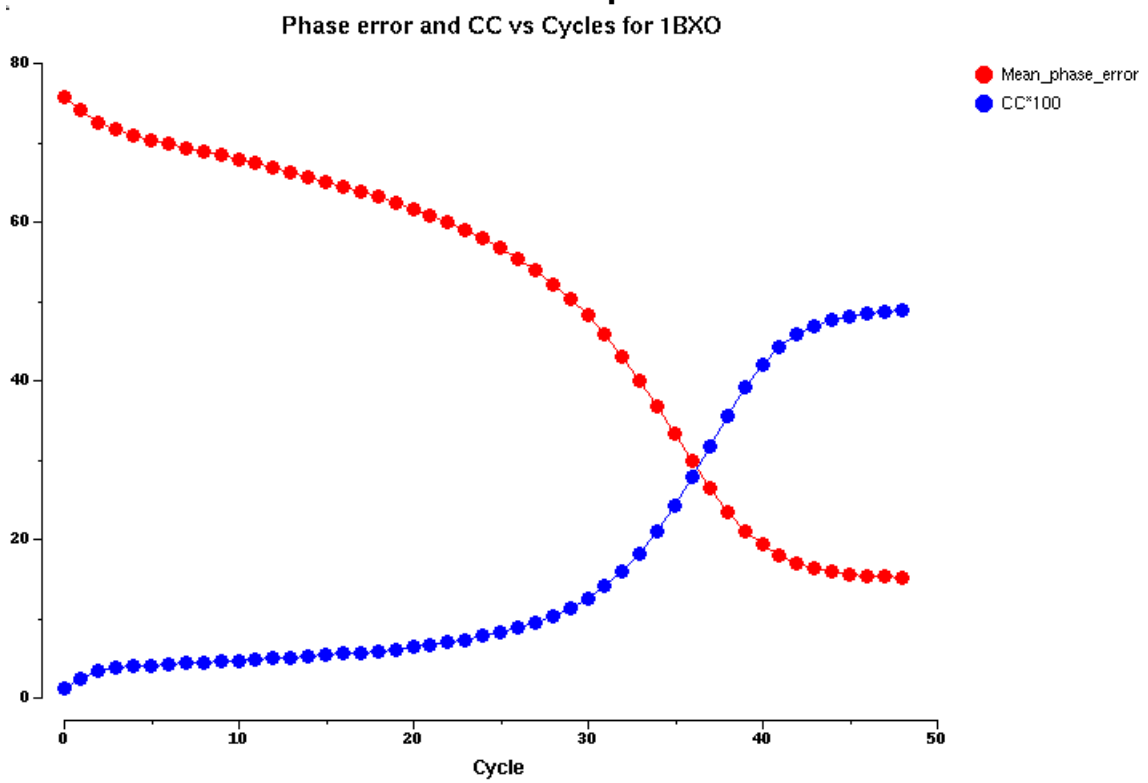
Reflection handling

All reflections are divided into three groups: strong, medium and weak. The strong and weak reflections are used in phase refinement procedures while the medium reflections are used only for calculation of CC as a figure of merit to indicate solutions.

Correlation Coefficient (CC)

Correlation coefficient is computed between the normalized structure factors and calculated ones from fragment or from modified map. CC for medium reflections (CC-medium) is used to indicate the quality of the phases computed from a modified density map while this map is calculated using the strong reflections only. Here is an example to show the relation between CC-medium and phase errors vs the cycles of DDM and the CC-medium does indicate the solution clearly.

CC-medium and phases error:



CC for all reflections (CC-all) is used by a molecular replacement method in ACORN-MR to indicate correct orientation and position. CC-all is also used to indicate correct position for single random atom searching.

ACORN-MR

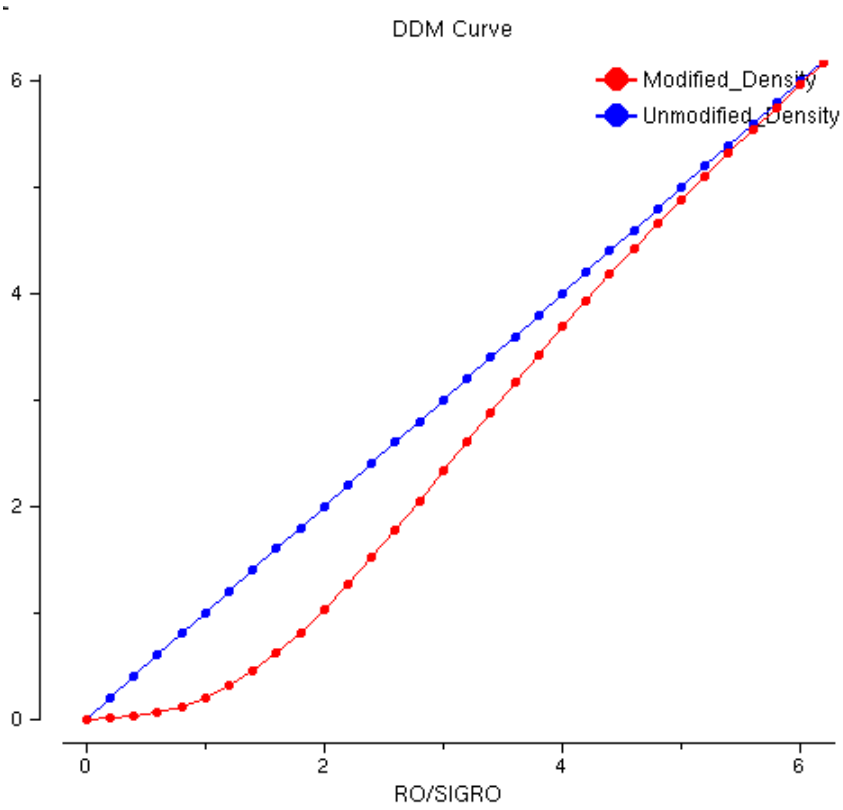
The first procedure in ACORN-MR is single random atom searching in a suitable region in unit cell according to the space group. ACORN-MR will generate a number of sets of single random atom and calculate CC-all for each set. Then all sets will be sorted on CC-all and provide maximum 1000 sets of initial phases with highest CC-all to ACORN-PHASE. If a structure contains some heavy atoms then some of the positions of the random atoms with highest CC-all may close to one of heavy atom positions. That will give a good enough initial phases, for example better than 80 degrees, to be refined by ACORN-PHASE to solve the structure.

Another procedure is a molecular replacement method of searching all possible sets of orientations and positions for the starting fragment. ACORN-MR will do the rotation function first and then translational function on the best solution of rotation function. The program will calculate CC-all for each set. Then all sets will be sorted on CC-all and the maximum 1000 sets of the initial phases and weights will be calculate with highest CC-all for ACORN-PHASE to refine. There are two kind of searching approaches: step by step searching and random searching. Normally random searching needs less computing time.

ACORN-PHASE

There are three procedures used for the phase refinement in ACORN-PHASE: Dynamic Density Modification (DDM), real space Sayre Equation Refinement (SER) and Patterson superposition (SUPP). For a default running only DDM is employed to refine the initial phases. DDM will calculate a map using a set of initial phases and weights of the strong

reflections and modify the map according to the ratio of map density (RO) over map standard deviation (SIGRO) and cut the top density at a level (CUTD) from 3*SIGRO to 15*SIGRO (default) according the number of cycles. A new set of structure factors will be obtained from the modified map by FFT and CC-medium will be calculated to check if a solution is reached. The DDM can modify not only E or F maps, but also Patterson or sharpened Patterson maps because the DDM will treat all maps as a map of ratio RO/SIGRO.



DDM-Curve:

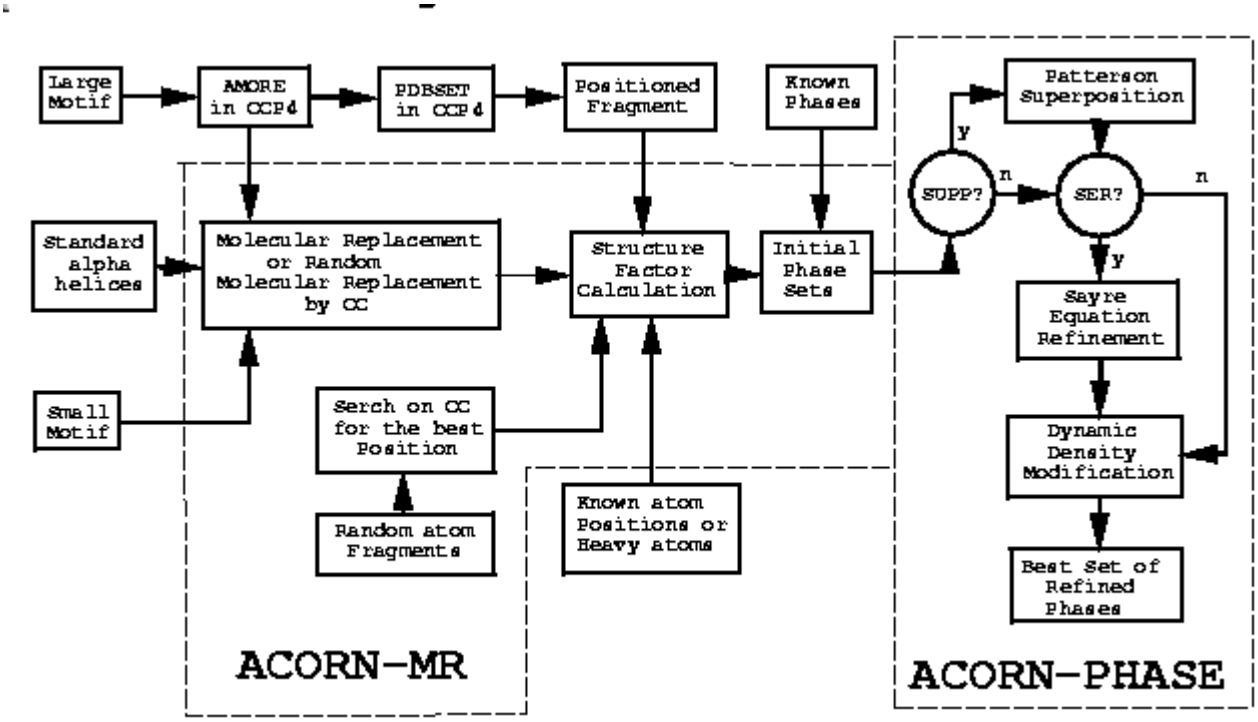
The DDM-Curve shows the approach of dynamic density modification that makes negative densities to zero and depresses the densities less than 1*SIGRO which mostly are noise and enhances the densities between 1*SIGRO and CUTD by cutting top densities in order to develop the new densities in protein region but outside the starting fragment region. Therefore the starting fragment including heavy atoms will not dominate the phase refinement process and the new densities in protein region will become more important.

In case of DDM alone did not reach the solution then SER should be introduced for a couple of cycles. SER uses a number of FFTs and inverse FFTs for all calculations to carry out the Sayre Equation refinement in real space. There are no phase relationships needed so that SER has no limit on the number of reflections to be used in the Sayre Equation. The purpose of using SER is to disturb the phase refinement process in DDM in order to reach global minimum other than local minimum.

SUPP calculates a half-sharpened Patterson and superposition the map by sum-function on the atom positions of the starting fragment. Then one cycle DDM will used to modify the map and a new set of phases will be calculated from the map. SUPP can be used only when the starting fragment is relatively large, say more than 10 atoms in it and normally the initial phase error can be reduce about 1 or 2 degrees.

Flow diagram for ACORN

ACORN:



Strategy to use ACORN

ACORN in CCP4 is a comprehensive program package which can handle all kind of starting fragments according to the features of the structure to be determined. The DDM in ACORN is a very powerful phase refinement procedure which can develop a complete structure from a fragment as little as 1% of whole structure. For example, ACORN can solve a structure of 1093 atoms from one Sulphur atom, a structure of 2130 atoms from one Calcium plus one Manganese and a structure of 4762 atoms from 9 Sulphur atoms or from a Heme group of 43 atoms with one Iron in it.

A lot of protein structures contain Sulphur atoms which can be located using anomalous scattering differences. Since the anomalous signal from Sulphur atoms is weak it has to be careful to collect the anomalous scattering data. But if the size of the structure is between 100 and 200 amino acids the first thing to try is the use of ACORN with random atom searching to solve the structure with native atomic resolution data before trying to collect the the anomalous scattering data. If atoms which are heaver than Sulphur atoms are in the structure the random atom searching in ACORN can solve even big structures, such as the Sel-Met proteins where Sulphur atoms are replaced by Selenium atoms.

As other direct method programs ACORN with random atom searching can be used to locate anomalous scatterers from SAD or MAD data, even the data is at low resolution, say 3-4 Angstrom because the anomalous scatterers are far apart from each other and can still be resolved at such resolution. Then the positions of the anomalous scatterers can be input to ACORN to solve the complete structure with atomic resolution data. Any known positions of anomalous scatterers or heavy atoms located by other methods can be used by ACORN directly.

A large part of protein structures contains alpha helices which have very similar configurations. Therefore a standard alpha helices in CCP4 fragment library is a very good starting fragment for ACORN to solve such protein structures. ACORN can pick up a part of the standard alpha helices with the size to suit the structure to be determined, for example 50 atoms which are 10 Alanines. The random searching MR approach in ACORN-MR is advised to obtain a correct orientation and position quickly. Another way to obtain a starting fragment is searching Protein Data Bank (PDB) or other data bank to find a small motif. For example the sequence searching by netblast can be used because the sequence of the structure is normally known. ACORN-MR will carry out the same procedure to find the correct orientation and position based on highest CC-all.

If a large motif can be found for the structure to be determined AMORE can be used alternately to obtain the correct orientation and position and then ACORN can be used to refine the initial phases from the positioned motif. But the size of the motif has to be large enough for AMORE to find the correct orientation and position.

A small molecular structure can be easily solved by random atom searching in ACORN even the structure contains only light (C, N and O) atoms since the data normally has higher than 1.0 Angstrom resolution.

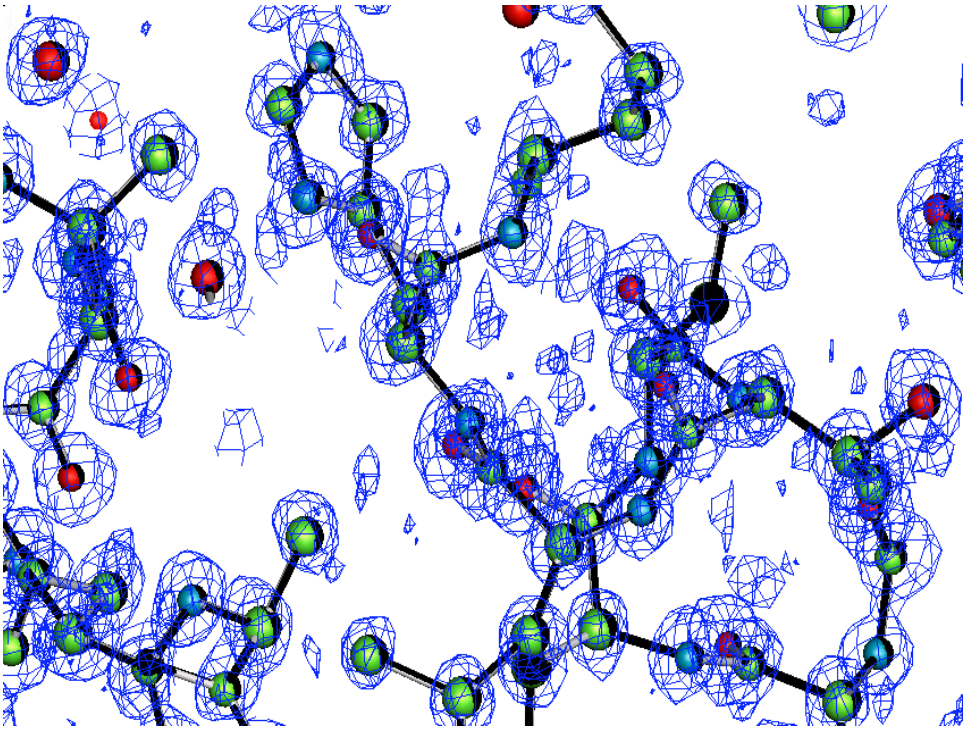
New protein structures solved by ACORN

ACORN has solved several protein structures and here are two examples which have been published:

1. A 19 kDa metalloproteinase [2] (pdb code 1eb6) with 1.0 Angstrom resolution data can be solved by ACORN in three ways:
    1. A Zinc atom was located by anomalous difference Patterson or by sharpened Patterson with native data. Then ACORN started from the Zinc position and solved the structure using only 168.9 seconds CPU time on SG computer.The whole process, starting from the processed and merged data and ending with a refined model, required less than 6 hours of computational time.
    2. ACORN solved the structure by random atom searching using 461.5 seconds CPU time that saved a lot of time to collect anomalous scatering data and locate the Zinc position.
    3. Since the structure contains alpha helices, a starting fragment of 50 atom standard alpha helices was used and the structure was solved by random moleculer replacement method in ACORN using 52844.6 seconds CPU time.
2. A 40 kDa homodimeric protein [3] (pdb code 1i4u) with 1.15 Angstrom resolution data was solved by ACORN with starting fragment of 12 Sulphur atoms using 1348.2 seconds CPU time. The Sulphur atoms were located by *SnB* and SHARP, but no further progress could be obtained before using ACORN.

A typical E-map from ACORN

The following E-amp was calculated using the phases and weights from ACORN starting from Zinc position with 1eb6 atomic resolution data. It was quick to build the structure by ARP/wARP. The contour level of the E-map was 2*SIGRO. It is clear to see the E-map is very close to the final model.

E-map for 1eb6: 

REFERENCES

1. Foadi,J., Woolfson,M.M., Dodson,E.J., Wilson,K.S., Yao Jia-xing and Zheng Chao-de (2000) *Acta. Cryst.* **D56**, 1137-1147.
2. McAuley,K.E., Yao Jia-Xing, Dodson,E.J., Lehmbeck,J., Ostergaard,P.R., and Wilson,K.S. (2001) *Acta. Cryst.* **D57**, 1571-1578.
3. Gordon,E.J., Leonard,G.A., McSweeney,S. and Zagalsky,P.F. (2001) *Acta. Cryst.* **D56**, 1230-1237.

OASIS and Xe phasing: potential in high-throughput crystallography

Quan Hao
Cornell High Energy Synchrotron Source (CHESS), Cornell University, Ithaca, NY 14853, USA.
Correspondence e-mail: qh22@cornell.edu

The CCP4 supported program *OASIS* has been tested using one-wavelength anomalous scattering data collected from a xenon derivative of the lobster apocrustacyanin A1 protein [Cianci *et al.* (2001). The Xe atoms were located by the program *SAPI* and the absolute configuration was determined by the program ABS. The electron density map after *OASIS* and density modification clearly revealed the solvent boundary and the C_α trace. The test demonstrated that, by exploiting the anomalous signal at single wavelength, OASIS can be used to determine phases at moderate ($\sim$2.3 Å) macromolecular crystallographic resolution for a medium-size protein (3500 non-H atoms in the asymmetric unit). As the xenon derivatives can be obtained from native protein crystals using commercially available equipment in a relatively short time frame (a few hours), the method described in this paper may provide a good alternative to MAD or MIR phasing, in particular when high-throughput is desirable.

Keywords: OASIS; one-wavelength anomalous scattering; xenon derivative.

1. Introduction

In view of the mounting evidence that one-wavelength anomalous scattering (OAS or otherwise known as SAD) may be sufficient to solve protein structures (Hao, 2000), the OASIS program (Hao et al, 2000) was written to determine phases using one-wavelength anomalous scattering data instead of using additional multiwavelength diffraction data (MAD). This is of particular importance when protein crystals are sensitive to X-ray irradiation or the absorption edges of the anomalous scatterers, such as xenon and sulfur, are difficult to access. A number of minor changes in the new version of OASIS to be released by CCP4 include new keywords to allow resolution and sigma cutoff. The upper limit on the number of reflections has been increased to 150,000 (from 90,000).

In preparing samples for MAD or OAS phasing, the most favored approach is the incorporation of selenium into protein using seleno-methionine during the expression of the protein. However, this is only successful when the gene encoding the particular protein is known and an expression is established and when the substitution does not affect crystalline order. In cases where seleno-methionine substitution is not plausible an attractive method for preparing samples is to incorporate xenon gas into the crystal. Xenon is known to bind to hydrophobic pockets within proteins at modest pressure. The one-wavelength anomalous scattering data (courtesy of Dr Rizkallah and Professor Helliwell, see Table 1 for details) collected from a xenon derivative of the lobster apocrustacyanin A1 protein (Cianci *et al.*, 2001) was used to test the possibility of *ab initio* phasing.

Table 1
OAS data

Values in parentheses refer to the highest resolution shell. The abbreviation a.s.u. stands for asymmetric unit.

Space group	$P2_12_12_1$
Unit cell	*a* = 41.11 Å
	b = 79.81 Å
	c = 109.86 Å
Non-H atoms in a.s.u.	3505
Number of Xe sites in a.s.u.	3 major + 1 minor
Source	Daresbury SRS Station 7.2
Wavelength	λ= 2.045 Å
f'' (in electrons)	11.5
Resolution	64.5-2.3 Å
Unique reflections	16723
Completeness	99.7%
Redundancy	7.1
$I/\sigma(I)$	23.4 (12.3)
R_{sym}	7.3 (14.5)%

2. Locating the xenon sites

The Se anomalous scatterers for both structures were located by the conventional direct-methods program *SAPI* (Fan *et al.*, 1990; http://staff.chess.cornell.edu/~hao/sapi/sapi.html) using magnitudes of anomalous differences,

$$|\Delta F(H)| = ||F(H)| - |F(\bar{H})||$$

for reflections within 3.0 Å. The solution was selected by a default run of the program. The largest 416 normalized structure factors *E*'s were used in tangent formula phase refinement. The resultant electron density map produced a group of 3 highest peaks; there was a clear gap between this group and other peaks in terms of peak height. A Karle-recycle refinement (an option in SAPI) of these three sites yielded an additional minor site. The absolute configuration of these sites was determined by the program ABS (http://staff.chess.cornell.edu/~hao/abs/abs.html) based on the P_s-function method (Woolfson & Yao, 1994). These Xe sites agreed well with the published sites (Cianci *et al,* 2001) and formed the basis for the next phasing step.

3. OASIS and DM phasing

The *ab initio* phasing of the OAS data was implemented in the computer program *OASIS* (Hao *et al.*, 2000). All Friedel pairs (including centric reflections) were evaluated using *OASIS*. The script that was used to run OASIS is shown below:

```
#oasis.com
oasis HKLIN xea1_19_trn.mtz HKLOUT xea1_oasis.mtz << eof
TITLE DIRECT PHASING OF xea1 xenon OAS DATA
HCO XE 12
FIT
RES 2.3
LCE 7
ANO XE 11.5
```

```
POS XE    -0.62360 -0.52335 -0.50331  1  0.318
       XE    -0.87037 -0.55929 -0.99017  2  0.419
       XE    -0.65256 -0.76443 -0.87725  3  0.268
       XE     0.47078  0.20858  0.15206  4  0.080
LABIN F1=F SIGF1=SIGF F2=DANO SIGF2=SIGDANO
LABOUT F1=F SIGF1=SIGF PHI=PHIdp W=Wdp
END
eof
```

Density modification using the CCP4 program *DM* (Collaborative Computational Project, Number 4, 1994) was then applied to the resulting phase sets. Phase error analysis and figures of merit before and after *DM* are given in Table 2. The electron density maps after *OASIS* and density modification clearly revealed the solvent boundary. The C_α trace was clearly visible but there were a number of places where the electron density was broken. A correlation coefficient between the *OASIS* + *DM* phased map and the final refined structure was 0.57.

Table 2
Phase error analysis and figure of merit

Reflections were sorted in descending order of F_{obs} and cumulated into groups. Phase errors were calculated against the refined models (Cianci *et al.*, 2001, PDB reference 1h91) weighted by F_{obs}.

	Phase errors (°)	
Number of reflections	***OASIS***	***OASIS* + *DM***
3000	58.5	44.3
6000	59.5	47.9
9000	60.0	49.6
12000	60.9	51.2
15000	61.7	52.3
16723	62.1	52.9
Mean figure of merit	0.49	0.71

4. Discussion

Here we demonstrate that, by exploiting the anomalous signal at single wavelength, OASIS can be used to determine phases at moderate (~2.3 Å) macromolecular crystallographic resolution for a medium-size protein. The total CPU time consumed by SAPI, ABS and OASIS was about 3 minutes on an Alpha XP10000 workstation. As the xenon derivatives can be obtained from native protein crystals using commercially available equipment in a relatively short time frame (a few hours), the method described in this paper may provide an attractive alternative to MAD or MIR phasing, in particular when high-throughput is desirable.

Acknowledgments

I would like to thank Dr P J Rizkallah and Professor J R Helliwell for making available the apocrustacyanin A1 data and valuable discussions.

References

Cianci, M., Rizkallah, P.J., Olczak, A., Raftery, J., Chayen, N.E., Zagalsky., P.F. & Helliwell, J.R. (2001). *Acta Cryst.* D**57**. 1219-1229.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760-763.
Fan, H. F., Hao, Q., Gu, Y. X., Qian, J. Z., Zheng, C. D. & Ke, H. (1990). *Acta Cryst.* A**46**, 935-939.
Hao, Q., Gu, Y. X., Zheng, C. D. & Fan, H. F. (2000). *J. Appl. Cryst.* **33**, 980-981.
Woolfson, M. M. & Yao, J. X. (1994). *Acta Cryst.* D**50**, 7-10.

The Cambridge Structural Database System – from crystallographic data to protein-ligand applications

Stephen J. Maginn, on behalf of the staff of the Cambridge Crystallographic Data Centre (CCDC)

CCDC, 12 Union Road, Cambridge CB2 1EZ, UK
maginn@ccdc.cam.ac.uk

The Cambridge Structural Database (CSD) System is a well-known and widely used resource in structural chemistry. The Cambridge Crystallographic Data Centre (CCDC), which collates and makes the CSD System available, has recently been exploring the value and application of knowledge implicit within the database – knowledge about molecular conformation and about intermolecular interactions.

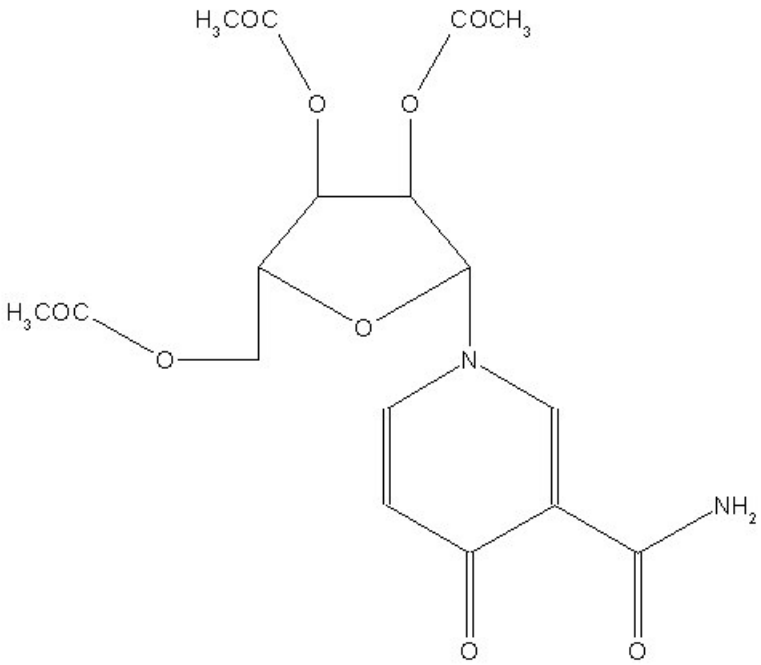
The Cambridge Structural Database (CSD)

Currently (forthcoming October 2001 release) containing 245392 entries, the CSD is the world's repository for small molecule organic and metal-organic crystal structures. There is a strict definition of the "turf" covered by the CSD with respect to the Protein DataBank (PDB) – structures with less than 1000 atoms in the asymmetric unit go into the CSD. CCDC has deposition arrangements with a host of journal publishers, whereby structures going through the publication process have their crystallographic data deposited at Cambridge, and once publication occurs, they are added to the database. An increasing number of unpublished structures, labelled as Private Communications, are now also included, and submission of these is encouraged, although all go through the same rigorous checking procedures as structures intended for publication.

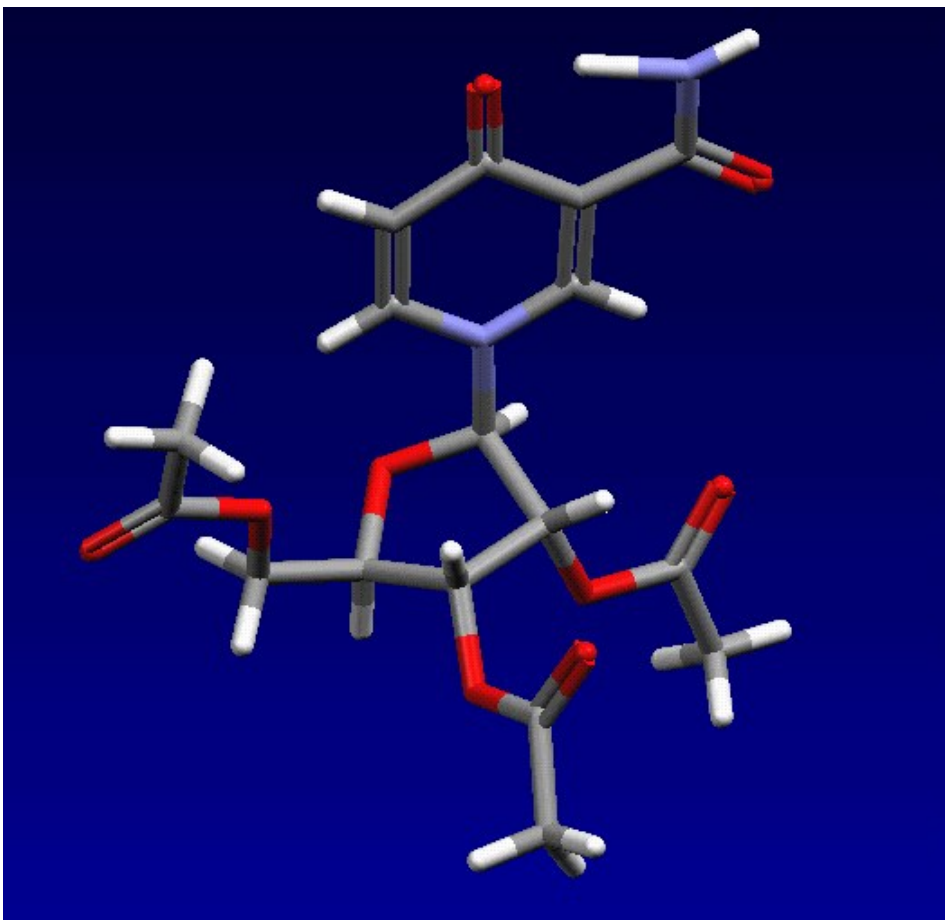
The information stored in the CSD for each entry can be considered in three classes. Firstly, there is the text-based (and sometimes numeric) information, containing the bibliography (i.e. full literature reference, where appropriate), chemical names and formulae, some experimental information about the crystal structure determination procedure, and any other information that may be available (e.g. compound's use, colour and shape of crystals, etc. etc.). Secondly, there is chemical connectivity information in the form of a 2D structural diagram – it is this that forms the basis of much of the sophisticated search mechanisms for the CSD System (see later). Thirdly, there is the crystallographic information, consisting of unit cell dimensions and space group, and atomic coordinates (these are available for the vast majority of entries, although not all). It is in this third category where the true value of the Database lies.

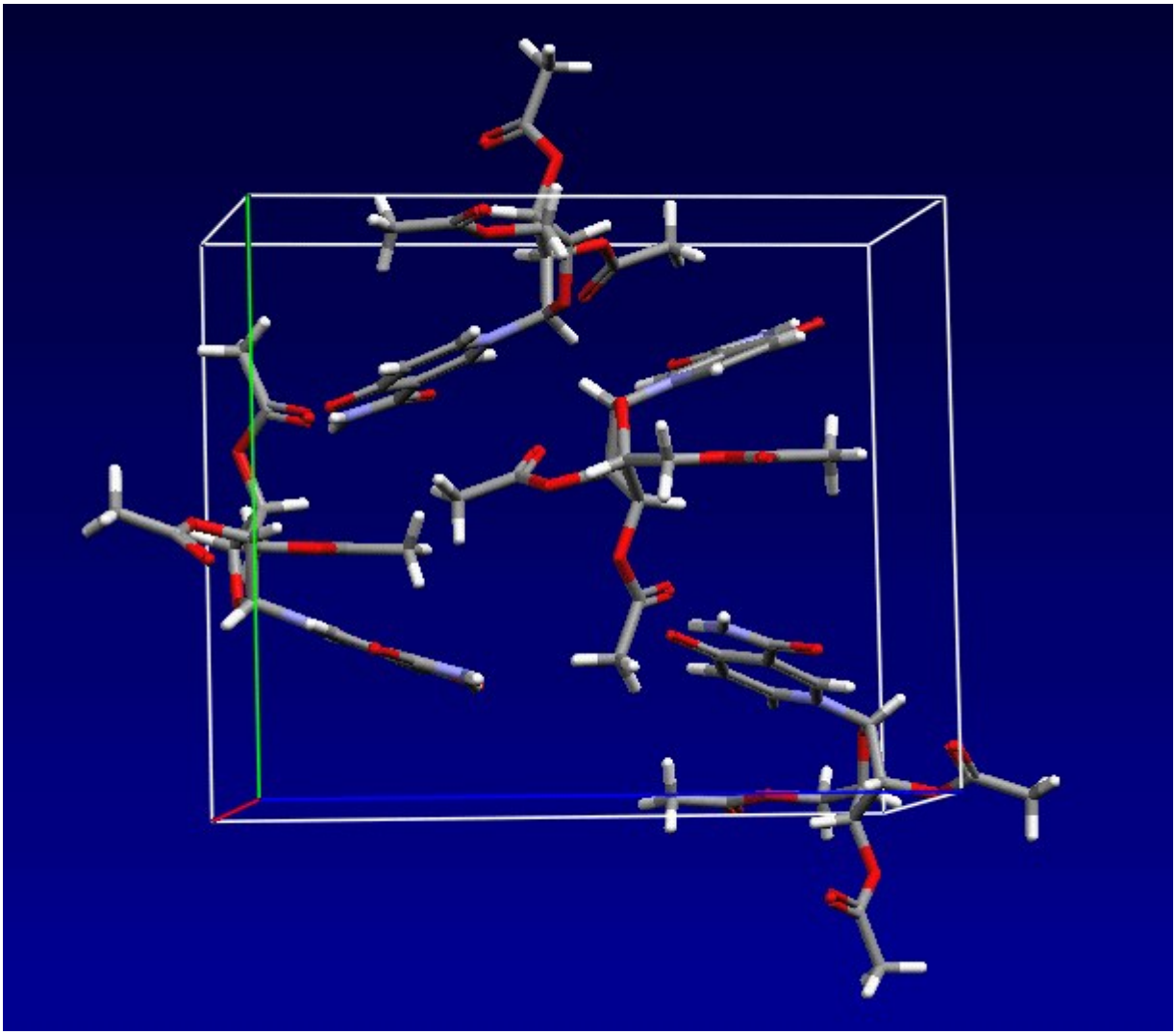
All database entries are identified uniquely by a six-letter reference code, or "refcode", which may be followed by two digits if the entry is a member of a family of entries. Fig. 1 shows some of the information stored for the entry BASYOJ, for example.

Fig.1: Cambridge Structural Database contents: bibliographic, 2D structure, 3D coordinates and packing



BASYOJ
4-Oxonicotinamide-1-(1'-beta-D-2',3',5'-tri-O-acetyl-ribofuranoside)
Source: Rothmannia longiflora
C17 H20 N2 O9
G. Bringmann, M. Ochse, K. Wolf, J. Kraus, K. Peters, E-M. Peters, M. Herderich, L. Ake, F. Tayman
Phytochemistry 51 (1999), p271



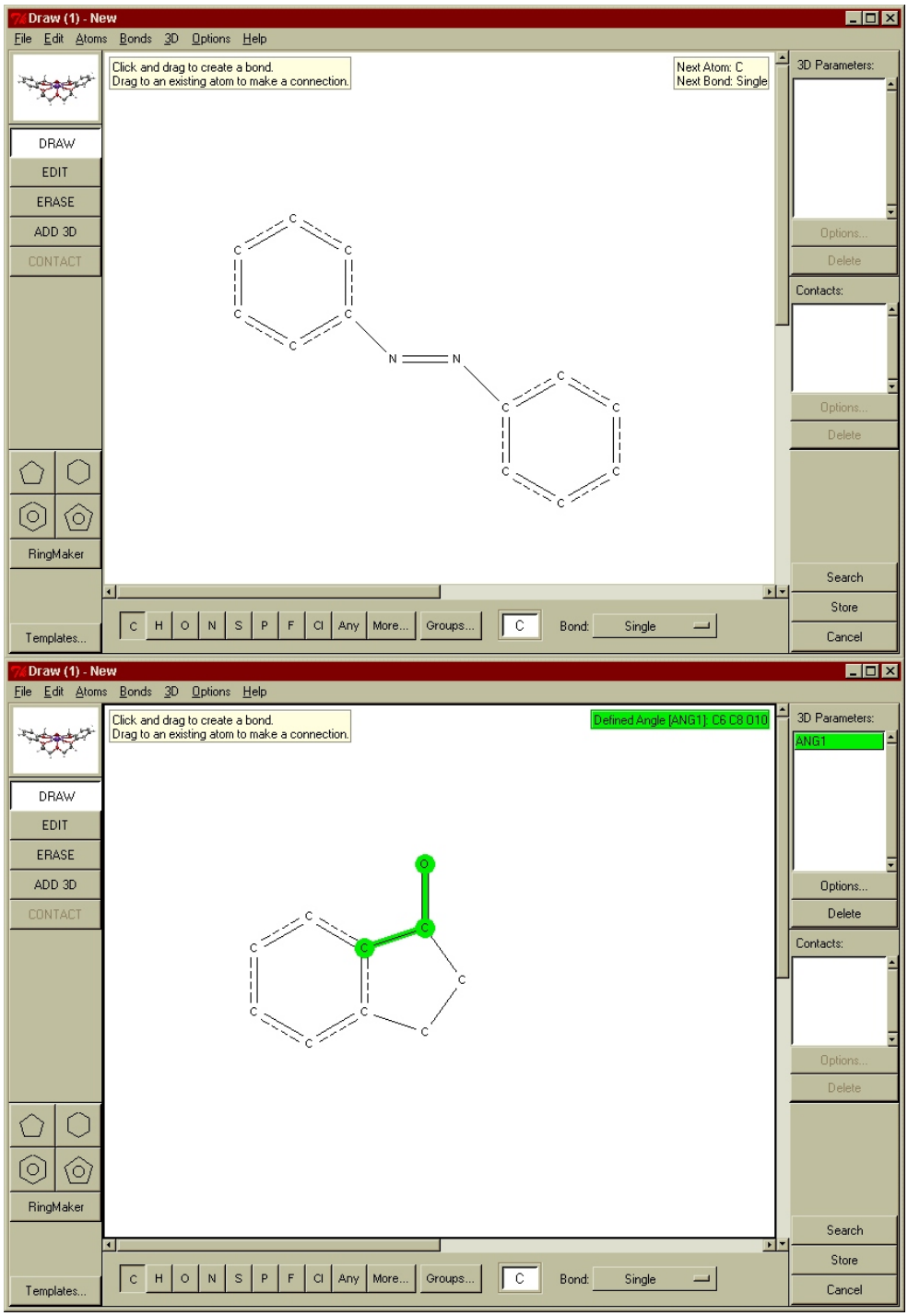


CSD Search and Analysis Software

The CSD is not provided to users in isolation. It comes with a package of software, designed for search of the Database and analysis of results, which can be used to extract knowledge about molecular conformation or intermolecular interactions.

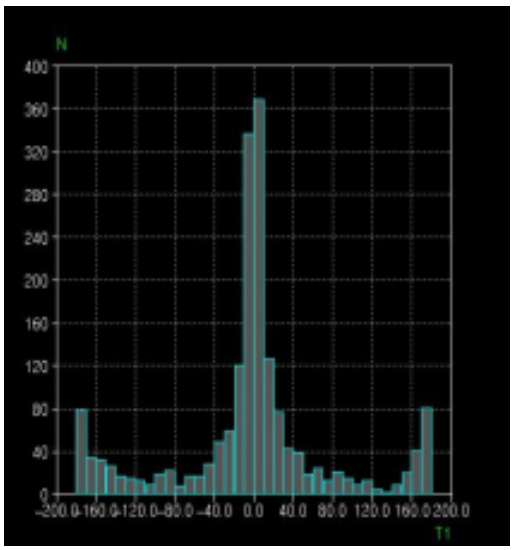
For many years, the main search interface to the CSD was the QUEST software. This remains available (see below), although the program has now been replaced by a much more modern interface known as ConQuest (see below for availability). All fields of data within the CSD are searchable using the ConQuest software, although the search functions used to extract geometrical knowledge are those based on the 2D structural diagram. Database search queries may be constructed to search for user-defined molecular fragments, within which molecular and intermolecular geometries may be tabulated and used to further constrain the search if required (see Fig. 2).

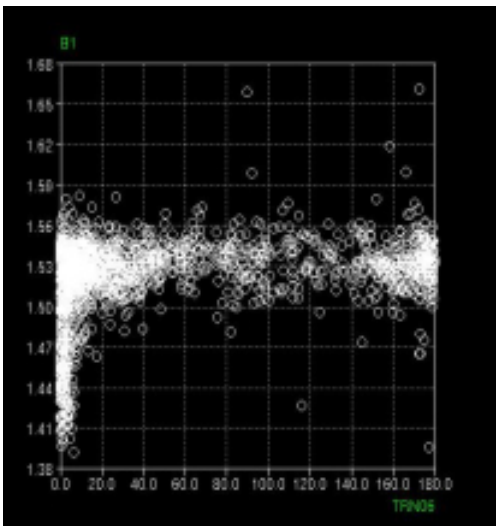
Fig. 2: A couple of examples of 2D fragment searches in ConQuest



Once geometries have been tabulated, the information may be exported to Excel for analysis, or to the program VISTA, which also forms part of the CSD System. Trends in molecular geometry or intermolecular interactions for compounds or molecular fragments of interest may then be discerned and correlated (Fig. 3).

Fig. 3: Correlation of O=C-C-O torsion angles with the central C-C bond length. When there is a hydroxyl group involved, an intramolecular H-bond is set up (as is tautomerism) and the C-C bond shortens. Plots produced using VISTA.





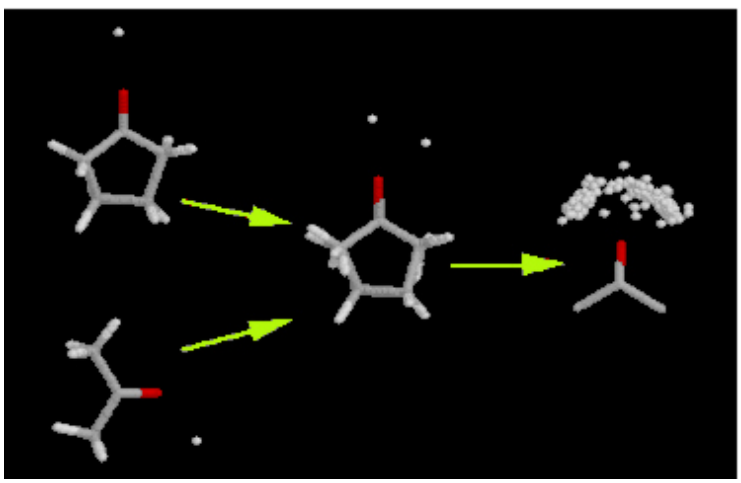
Knowledge-Based Libraries

CCDC has decided to expand the capabilities of the CSD System by including two knowledge-based libraries. These encapsulate much of the information on intermolecular interactions and on molecular geometry found within the CSD, which has been pre-extracted and is presented to the user in an easy to visualise form. One of these libraries, IsoStar, has been available for some years, whereas the other, Mogul, is in development with a view to release in 2002.

IsoStar:

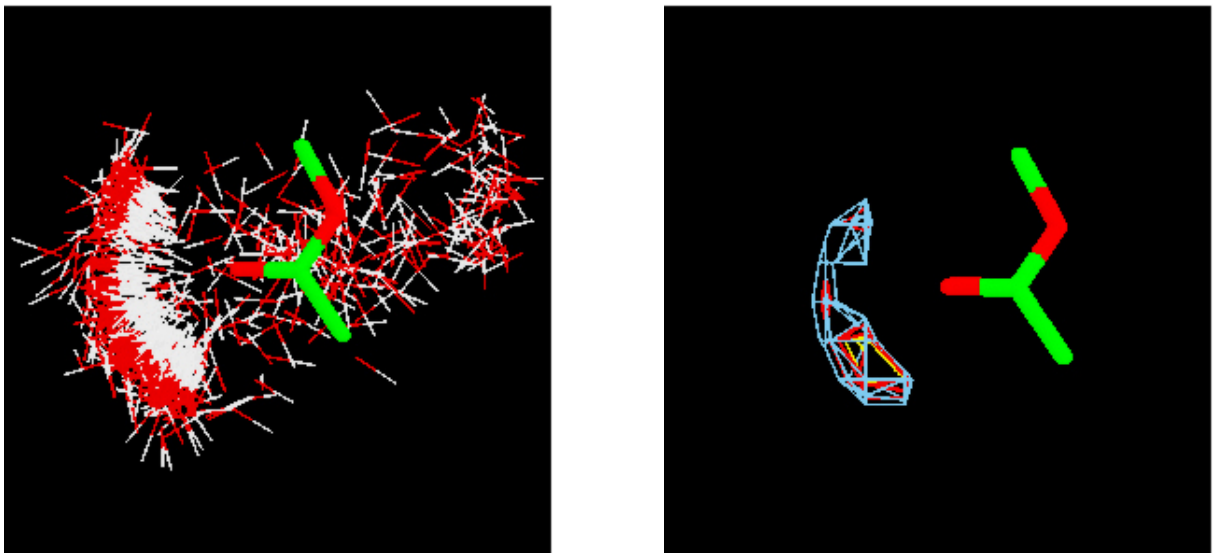
IsoStar **[1]** is a library of the intermolecular interactions found within the CSD and some from the PDB (see later). Searches of the CSD have been carried out for over 12000 particular intermolecular interactions. One of the participating fragments in each interaction has been designated a “central” group, the other an “interacting” group. The central groups are then superimposed and normalised for all the hits in each search, to produce “scatterplots” in which the interacting groups are arrayed around the central group, thereby displaying preferred geometries of interaction. It is possible in IsoStar to link back from individual data points in a scatterplot to the CSD entry that gave rise to that point.

Fig. 4: Methodology of IsoStar scatterplot production



IsoStar scatterplots can sometimes appear confusing, when there are thousands of “hits” contributing to them. The scatterplot can therefore be displayed as a contour plot, where the contours are values of density of interactions per unit volume of space. These may be scaled differently, by dividing by the expected density of interactions should they have been randomly distributed, such that they become probability or “propensity” densities, and can be directly compared (and combined – see later) quantitatively as well as qualitatively (see Fig. 5).

Fig. 5: The scatterplot for hydroxyl interactions with an ester link, and its corresponding contour plot, revealing the preference for interaction with the carbonyl oxygen and its lone pair directionality.



IsoStar also contains information on intermolecular interactions extracted from the PDB, in the form of over 3000 scatterplots – with certain restrictions. The PDB interactions within IsoStar are between a ligand and its host protein, and only from structures determined at better than 2.5Å resolution. There are therefore nowhere near as much data contained within the PDB-derived plots as in the CSD-derived plots, and hydrogen atom positions are missing, leading to some loss of directional information. Nevertheless, this enables some comparisons between plots derived from CSD and PDB to be made, and some conclusions to be drawn about how appropriate it may be to use small molecule crystal data to model protein-ligand interactions (Fig. 6).

Fig. 6: Scatterplots for O-H and N-H interactions with a COO- central group, derived from the CSD (left) and PDB (right). Lone pair directionality is visible in both.

IsoStar also contains the results of some calculations of interaction energies for certain systems, produced using intermolecular perturbation theory **[2]**, and model molecules.

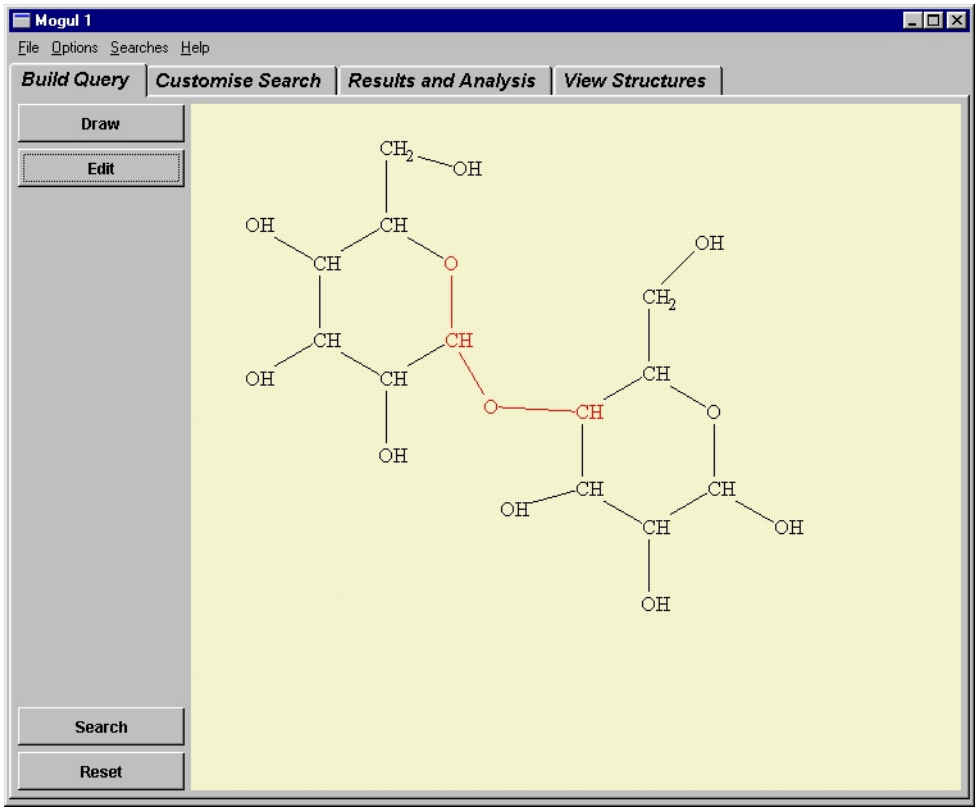
IsoStar scatterplots and contour plots are displayed in a customised version of Rasmol, whereas the IsoStar framework itself is browser-based.

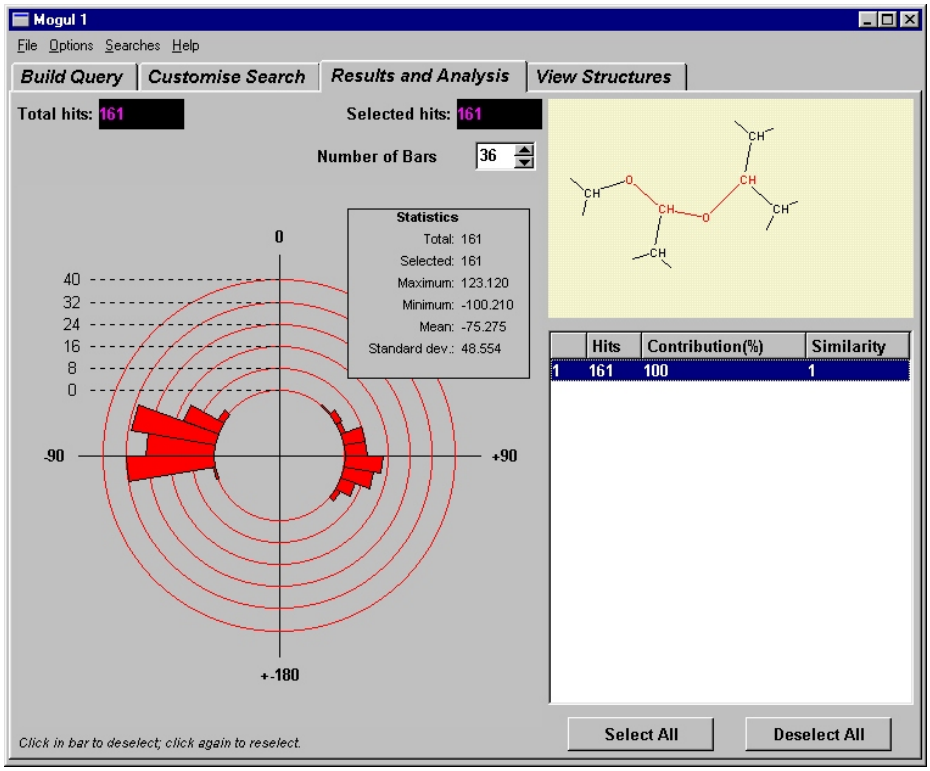
Mogul:

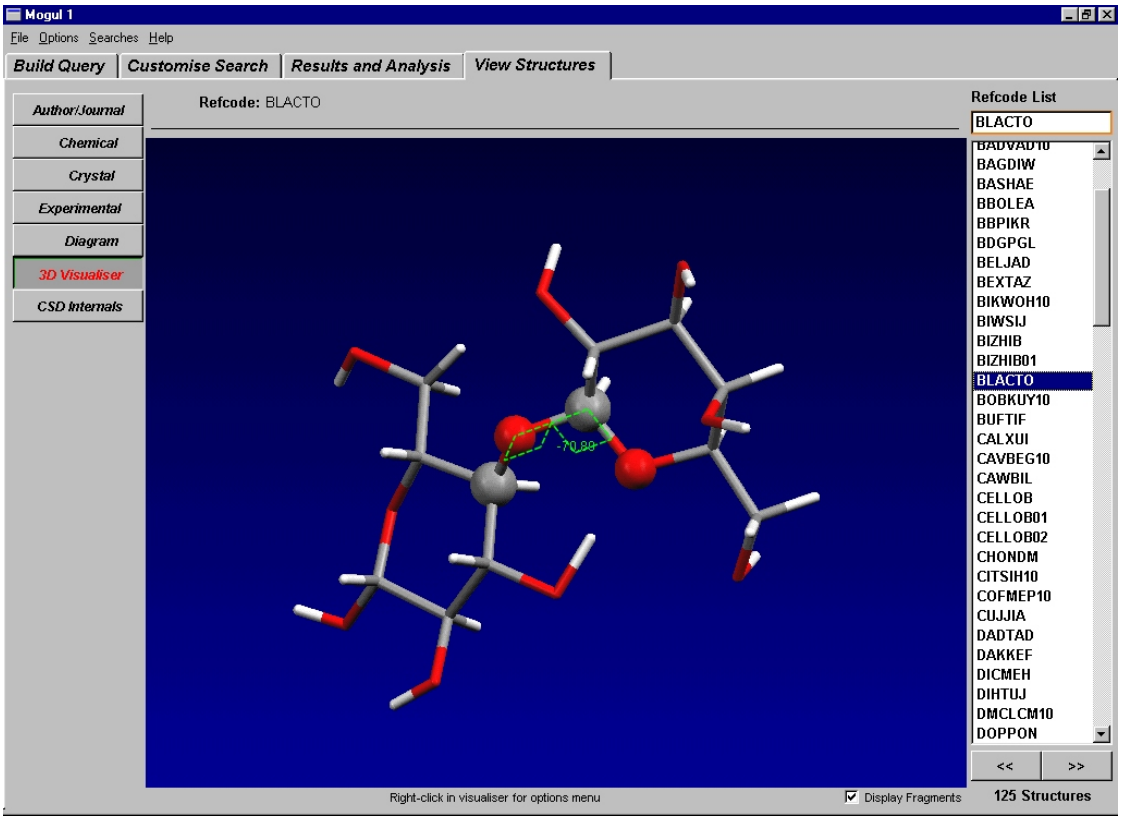
Currently in development at CCDC, Mogul embodies the molecular geometrical information from the CSD into three constituent libraries – one of bond lengths, one of valence angles, and one of torsion angles (excepting ring torsions). Values of these parameters within molecules are heavily influenced by the chemical environments of the constituent atoms, and therefore atoms of the same element and hybridisation need to be more clearly defined than in traditional atom-typing arrangements. Atoms in Mogul are therefore exactly defined by going out to 2 bonded atoms away from each constituent atom, considering atom and bond types for each. When searches for an exact match with the studied parameter, to this level of definition, are carried out, one often finds that there is not enough data (i.e. not enough exact matches, or "hits"), even throughout the quarter of a million CSD entries, to be statistically significant enough to draw firm conclusions. However, the Mogul libraries have a hierarchical tree structure, so that the strict atom definitions can be relaxed, and one can move up the "branches" of the tree structure, in order to obtain enough information.

Mogul is expected to be ready for release as part of the CSD System in 2002.

Fig. 7: Use of Mogul to find the torsional distribution about a link between rings in a sugar molecule.







IsoStar and Mogul therefore represent a catalogue of knowledge, extracted from the CSD (and in IsoStar's case also the PDB); such knowledge may also be extracted by judicious use of the CSD and its accompanying software. A similar procedure may be followed for extraction of conformational and interactional knowledge from the PDB, for protein-ligand complexes, using Relibase**[3]**. All this information is not merely of interest - it is genuinely useful in approaching real-world problems.

Applications of Crystallographic Knowledge

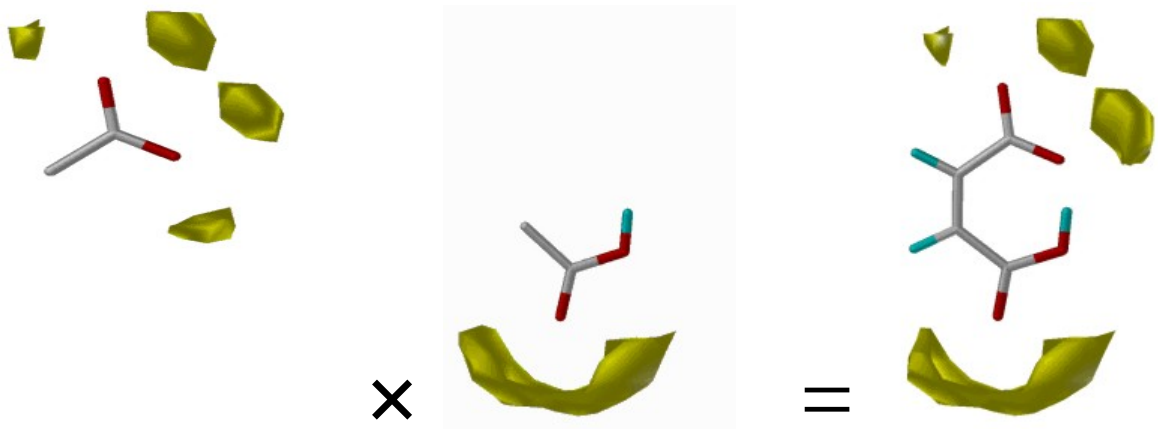
It is possible to envisage many applications of both forms of crystallographic knowledge referred to above. There are obvious examples such as the use of conformational information in conformational search programs, or to compare with and/or validate the results of molecular modelling calculations. Crystallographic information may be able to feed back into the latter in the form of enhancements and alterations to molecular mechanics forcefields. There are other, less obvious applications though, in "bootstrapping" crystal structure determinations - for example, it may be possible to use the information in a more automated way to help protein threading in poor resolution electron density maps (similar to the use of Ramachandran plots), or to determine potential ligand conformations in binding sites where the density may be inconclusive. In small molecule structures, validation of new structures and structure solution from powder diffraction (where crystallographic information is added to a model to constrain the number of variable parameters and therefore give a better chance of structure solution from the limited data available) are but two potential applications. CCDC is involved in collaborations to explore some of these areas, but the applications which are perhaps best developed up to now, because of commercial interests, are those involving protein-ligand docking studies. Two examples are shown below, each using a different methodology of exploiting information extracted from the CSD (and PDB).

SuperStar

The program SuperStar **[4,5,6]** has been developed as a wholly knowledge-based approach to determining where particular organic functional groups (known as "probe groups") like to sit in macromolecular binding sites. A study of a particular binding site with a range of probe groups may therefore produce the necessary information required for the creation of a pharmacophore or as the first stage in an ab initio ligand design procedure.

SuperStar depends entirely upon the propensity plots (i.e. the scaled contour plots) contained within IsoStar. The program derives the identity of organic functional groups on the protein structure which intrude upon the surface of the binding site, retrieves the relevant plots from IsoStar where the protein's group is the central group and the user-selected probe group is the interacting group, and combines them (by multiplication, as we are considering propensities / probabilities) to produce a grand, 3D binding map within the protein binding site for the probe group.

Fig. 8: Combination of IsoStar plots in SuperStar



The earliest versions of SuperStar have used only the IsoStar plots extracted from the CSD as their source. A new version is about to be released, however, which offers the choice of using the PDB-derived IsoStar plots instead; they are somewhat less well-defined as they contain less data, but it may be considered more appropriate to use them than CSD-derived plots in certain cases. SuperStar also has a limited capability for handling active site flexibility in that rotation of hydroxyl groups are considered.

This wholly knowledge-based approach employed by SuperStar offers huge advantages in terms of speed over a more traditional, energy-based approach to deriving this information in that there are no long-winded calculations. The academic papers **[4,5,6]** contain much in the way of validation of the method.

GOLD

GOLD stands for Genetic Optimisation for Ligand Docking **[7,8]**. Unlike SuperStar, GOLD is not a purely knowledge-based program - it depends upon a forcefield to perform the core of its function, which is the derivation of preferred binding modes for particular ligands within particular binding sites. A genetic algorithm is the means used to optimise this - hence the program's name. However, the forcefield is enhanced by the use of crystallographic information in two key ways; firstly, ligand torsion angles can be restricted to CSD-observed ranges of values using a cruder version of the torsion library in Mogul; secondly, there is directional information from IsoStar, particularly concerning hydrogen bonding, hard-coded into the forcefield. Thus GOLD may be considered to be something of a hybrid between energy-based and knowledge-based methods. The methodology has been greeted with some acclaim in comparative studies **[9]**.

Originally created as a 3-way collaboration between the University of Sheffield (Dept. of Information Studies), former GlaxoWellcome and CCDC, GOLD has been available for some time now and is constantly developing - in deed, CCDC has recently entered into collaboration with Astex Technology to facilitate onward development. GOLD's uses are in understanding how molecules of known activity may bind to their target proteins, in the absence of crystallographic information, and in so-called "virtual screening" studies, where comparisons with a training set of experimental results may be used to draw inferences about likely activity of compounds within a combinatorial library.

Conclusion

Although it is early days in the exploitation of crystallographic knowledge, derived both from the CSD and PDB, it is clear that the potential applications are many and varied and the potential benefits huge. Evidence seems to show that rather than replacing energy-based means of modelling systems, these methodologies will enhance them in a symbiotic way.

Availabilities

The Cambridge Structural Database, accessible via search programs ConQuest or QUEST, together with programs VISTA and PLUTO and the knowledge-based library IsoStar, is available free of charge to UK academics on the EPSRC-funded, Daresbury-based Chemical Database Service (CDS). For further details, see the CDS website.

Alternatively, the CSD System can be obtained for an annual subscription fee direct from CCDC. GOLD and SuperStar are also available commercially. See the CCDC website for details.

References

[1] I.J. Bruno, J.C. Cole, J.P.M. Lommerse, R.S. Rowland, R. Taylor, M.L. Verdonk; *Journal of Computer Aided Molecular Design* 11 (1997), p525-537

[2] I.C. Hayes, A.J. Stone; *J. Mol. Phys.* 53 (1983), p83-105

[3] http://relibase.ccdc.cam.ac.uk , http://relibase.ebi.ac.uk , http://relibase.rutgers.edu

[4] M.L. Verdonk, J.C. Cole, R. Taylor; *J. Mol. Biol.* 289 (1999), p1093-1108

[5] M.L. Verdonk, J.C. Cole, P. Watson, V. Gillet, P. Willett; *J. Mol. Biol.* 307 (2001), p841-859

[6] D.R. Boer, J. Kroon, J.C. Cole, B. Smith, M.L. Verdonk; *J. Mol. Biol.* (submitted)

[7] G. Jones, P. Willett, R.C. Glen; J. Mol. Biol. 245 (1995), p43-53

[8] G. Jones, P. Willett, R.C. Glen, A.R. Leach, R. Taylor; J. Mol. Biol. 267 (1997), p727-748

[9] C. Bissantz, G. Folkers, D. Rognan; J. Med. Chem. 43 (2000), p4759-4767

mcps: contour-grayscale rendering of CCP4 map sections

Nicholas M. Glykos
Crystallography Group, IMBB, FORTH,
PO Box 1527, 71110 Heraklion, Crete, Greece.
Tel ++30-(0)81-394429, Fax ++30-(0)81-394351
glykos@crystal2.imbb.forth.gr
http://origin.imbb.forth.gr/software/

March 2001

Program description

The program **mcps** will plot a section (or a stack of sections) from a CCP4 map file using both a (dithered monochrome) grayscale representation and contour lines. This may be useful in cases where the electron or potential density distribution has "high valleys" or "low peaks" that make simple contour-line plots confusing, as illustrated in the following figure :

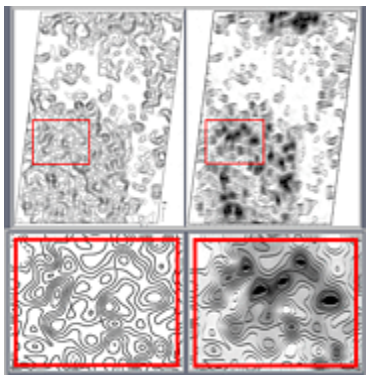


Figure 1: Comparison between the output from **npo**-**pltdev** (left panel) and **mcps** (right panel) for an 8Å potential density projection map of a large multiprotein complex.

The program comes with a (hopefully) sensible set of default values which should allow you to give it a try with just **mcps my.map** (to plot the first section of the map), or **mcps -first 3 my.map** to plot the **3**rd section.

The output (postscript) file is always written to a file whose name is constructed by appending the suffix ".ps" to the name of the input CCP4 map file.

User interface and program defaults.

All interaction with the program is through command-line arguments. This brings **mcps** closer to the unix tradition than the CCP4 tradition and has two important consequences. The first is that it is almost impossible to remember what is this week's name for the flag that changes the contouring level. This is somewhat remedied through the inclusion (with the distribution) of a formatted manual page. The second consequence is that the program must have a sensible set of default values (because very few of us would be prepared to type three lines worth of command-line flags to make it run). The program comes with the

following defaults : the grayscale gradient will be plotted starting from 1.0 sigma below the mean (white), to 3.0 sigma above the mean (black). Contour lines will be plotted starting from 0.50 sigma below the mean and then every 0.50 sigma. All contour lines that correspond to density higher that 3.0 sigma above the mean (and are, thus, on a black background) will be drawn white. With these defaults the emphasis is placed at the low density regions (which are the most difficult to follow with the traditional contouring methods). The high density areas will resemble a reverse-contrast **npo** plot as shown below for a map with well-resolved high-density peaks.

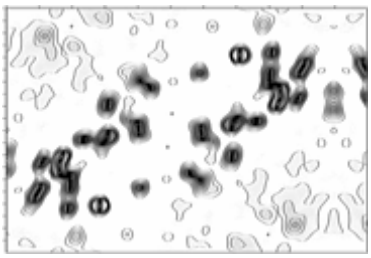


Figure 2: Output from **mcps** for a section from a 1.8Å electron density map.

Because the emphasis is placed on the low density areas of the map, **mcps** will not perform well with outstandingly strong features such as the origin peak of Patterson functions.

Examples

To plot the first section of a CCP4 map file with the name *myfile.map* using both contours and a grayscale representation, give

mcps myfile.map

To plot the 6th section of the map with tick marks every 0.250 fractional units, give

mcps -first 6 -ticks 0.250 myfile.map

To make plot the first section of the map without axes and tick-marks, and with the contrast reversed :

mcps -noaxes -reverse myfile.map

To produce an image sampled at 600dpi (instead of the default of 300dpi) :

mcps -scale 0.50 -resol 2.0 myfile.map

Bugs and ``features''.

mcps will not re-calculate any of the map statistics (like mean, minimum, maximum, rms deviation, etc.). These are all taken from the map header, and if they are not correct, neither will the plot be.

mcps will always produce a postscript file which at the default magnification will have (depending of the unit cell dimensions) a width of 7 inches or a height of 10 inches. Because this whole area is sampled at the default resolution of 300 dpi, the resulting

postscript files (although bitmapped monochrome) will be quite large (of the order of MBytes) and the procedure of calculating them is much slower than for normal contour plots.

mcps produces a bitmapped dithered monochrome image at a default resolution of 300dpi. The problem with this is, that while the result will look quite good when printed, the usual pre-viewing methods (and even ghostscript) will produce a rather loosy approximation to it. One of the possible work-arounds is to actually produce an intermediate file at the correct resolution, which you then display at a reduced magnification. If, for example, you have ImageMagick on your machine, you can try something like display -density 300 -geometry 50% myccp4.map.ps

Program availability

mcps is free software and is immediately available for dowload *via* http://origin.imbb.forth.gr/software/. The distribution includes documentation (html, manual page) and precompiled executables suitable for Irix, Linux, OSF and Solaris. When the author's load level permit, the source code of the program will be made human-readable and released.

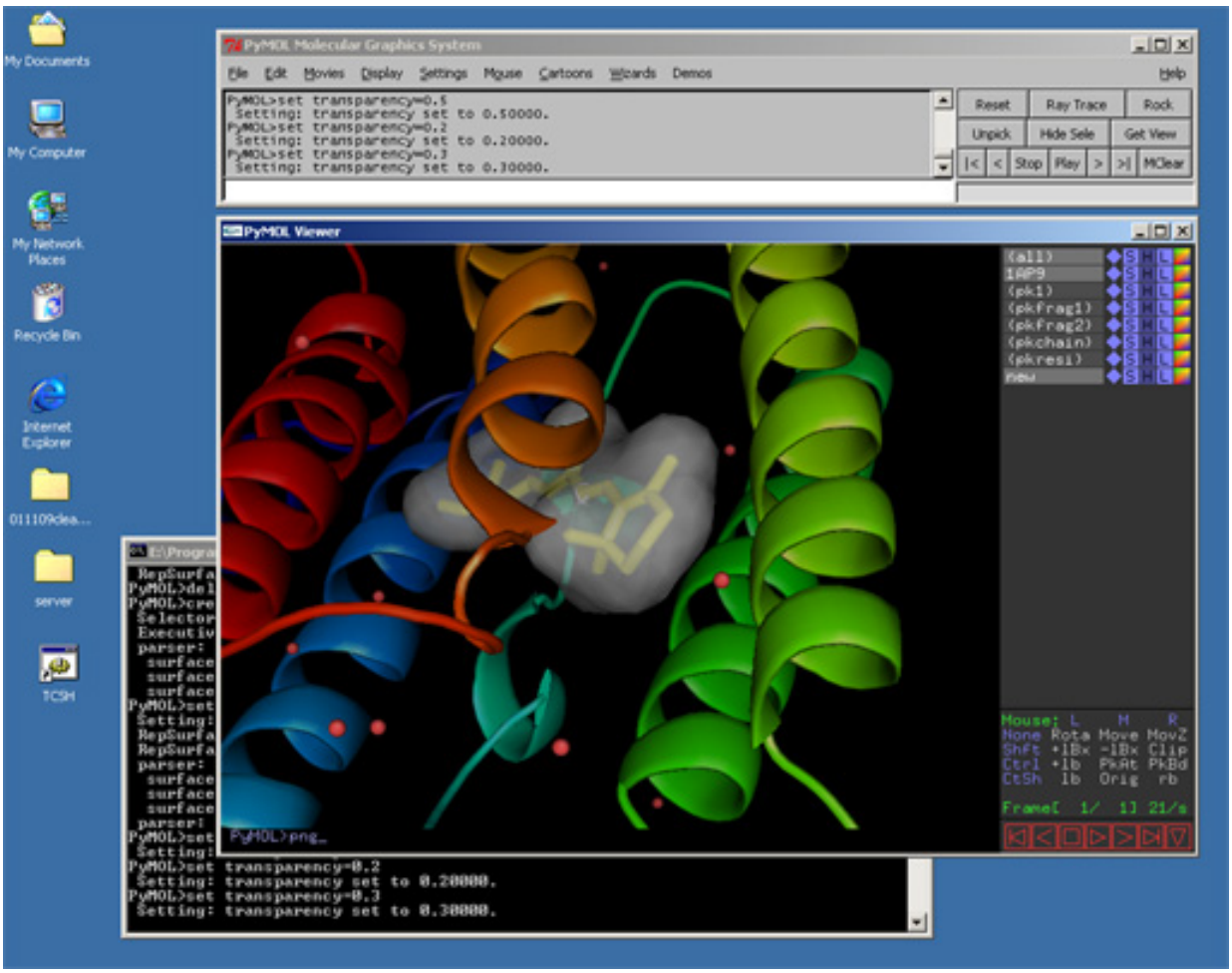
NMG, March 2001

PyMOL: An Open-Source Molecular Graphics Tool

Warren L. DeLano, Ph.D.
DeLano Scientific
San Carlos, California
USA
warren@delanoscientific.com

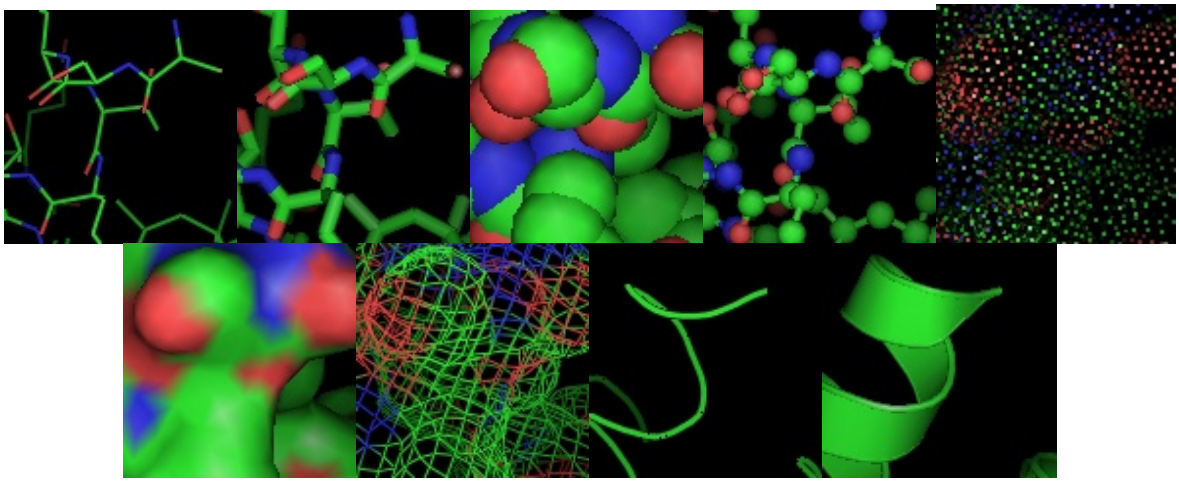
Introduction

PyMOL is a free cross-platform molecular graphics system made possible through recent advances in hardware [1] , internet [2] , and software development technology [3] . PyMOL provides most of the capabilities and performance of traditional molecular graphics packages written in C or Fortran [4] . However, its integrated Python interpreter endows it with features and expandability unmatched by any traditional package. PyMOL has been released under a completely unrestrictive open-source software license [5] so that all scientists and software developers can freely adopt PyMOL and then distribute derivative works based on it without cost or limitation.



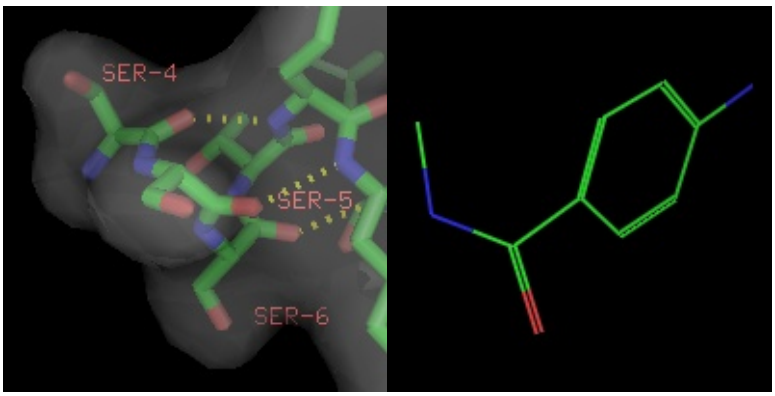
Molecular Graphics Representations

PyMOL supports most of the common representations for macromolecular structures: wire bonds, cylinders, spheres, ball-and-stick, dot surfaces, solid surfaces, wire mesh surfaces, backbone ribbons, and cartoon ribbons which are comparable to those generated by Molscript [6].



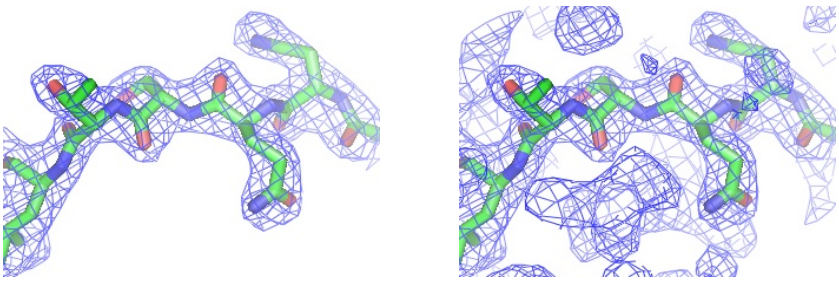
PyMOL Molecular Representations

Labels can be displayed for atoms, and dashed bonds can be used to indicate hydrogen bonding interactions and distances. Surfaces can be transparent, and molecules can be loaded from PDB files as well as several other common file formats.



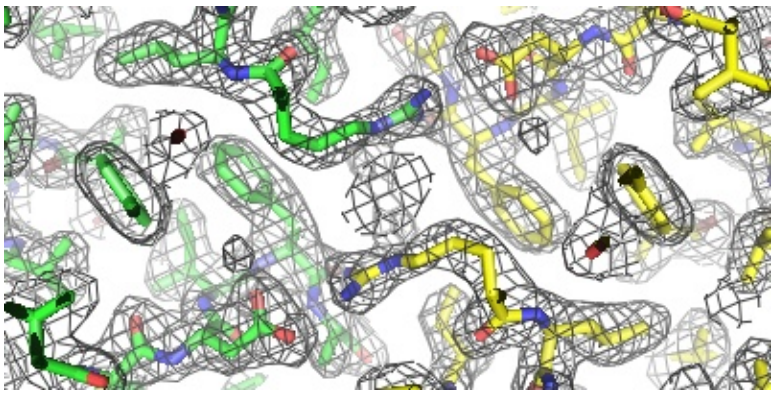
Crystallographic Visualization

PyMOL reads CCP4 and X-PLOR [7] map files and can display multiple arbitrary bricks of electron density within each map. PyMOL also has the ability to "carve" out electron density around any selection of atoms to create figures which show only localized electron density.



This example shows the clarity of "carved" electron density compared with a standard cartesian brick.

Provided that a structure has been loaded from a PDB file with correct unit cell and space group parameters [8] , PyMOL can generate symmetry-related molecules. This was made possible from R.W. Grosse-Kunstleve's generous contribution of the SgLite package [9] . At present, symmetry-related molecules are treated as independent objects, not as virtual images of the original object.

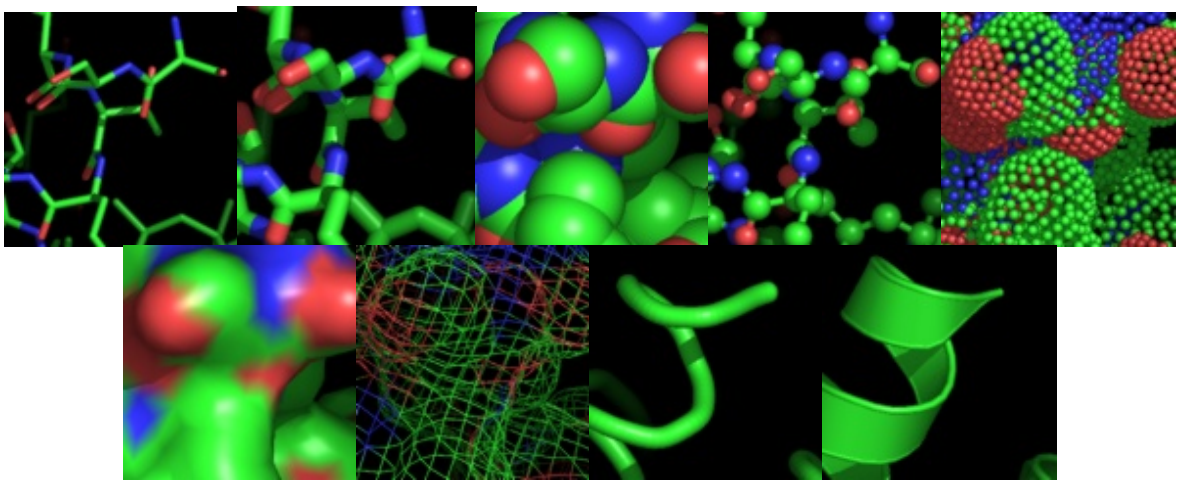


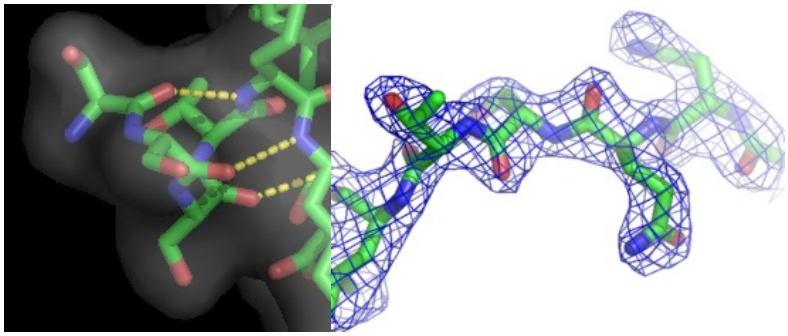
Electron density and models shown about a two-fold axis of symmetry.

PyMOL contains a rudimentary electron density "wizard" (written in Python) which can be used to quickly navigate through one or more electron density maps surrounding an atomic model. This wizard enables one to move and regenerate multiple meshes merely by CTRL-middle-clicking on the atom to center.

Publication Quality Figure Generation

A powerful time-saving feature found in PyMOL is its integrated ray-tracing engine. With a single mouse-click, any view displayed in the program can be immediately converted into a publication quality figure, complete with lighting, specular reflections, and shadows. This obviates the need for time-consuming efforts involving command-line programs such as Molscript [6] , Raster3D [4] , and ImageMagick [10] . With PyMOL, what you construct in real-time 3D is exactly what you will get when you ray-trace (except for labels), and you can import the resulting images directly into Microsoft PowerPoint.





Publication quality images generating using PyMOL alone (compare with OpenGL images above).

Although PyMOL's built-in ray-tracer is quite good, PovRay [11] support has recently been added, and it can now be used as a replacement renderer for generation of the highest quality images. Both OpenGL and ray-traced images can be output from PyMOL using standard Portable Network Graphics (PNG) files.

PyMOL's native cartoon ribbons are very similar to Molscript's, but if genuine Molscript output is required, PyMOL can read Molscript output in Raster3D format to perform rendering. The advantage of using PyMOL instead of Raster3D is that it allows users to orient and combine multiple objects in real-time 3D prior to rendering. The quality of the resulting images is comparable using either approach.

Molscript cartoon ribbons rendered with Raster3D

Molscript cartoon ribbons rendered with PyMOL

PyMOL cartoon ribbons rendered with PyMOL

PyMOL cartoon ribbons rendered with PovRay

Atom Selections

PyMOL supports multiple atom selection syntaxes which permit concise atom specifications: "C/143/CA", "C/PHE/", or "*/CA", as well as extended algebraic expressions resembling those found in X-PLOR and CNS [12]: "(byres ((resi 125 and chain A) around 5))". For most operations in PyMOL (such as coloring, zooming, or changing representations), atom selections can be used interchangeably with molecular objects. Selections can also be defined using the mouse in several different ways, including using a rectangular "lasso" around visible atoms. All atom selections can be visualized directly in the 3D window, which makes them easy to understand and verify.

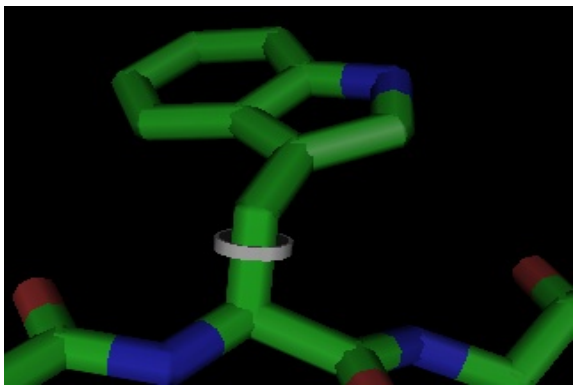
Selections are shown using colored dots over atoms.

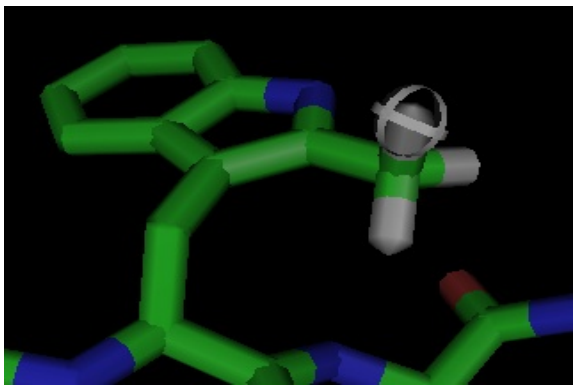
Molecular Editing

PyMOL supports molecular editing at several different levels. At the simplest level, it allows the user to create new objects out of atom selections which may span any number of other objects. Thus, removing atoms from an object or combining separate objects is trivial. Bonds can then be formed between atoms in the combined object.

PyMOL also supports conformational editing within objects using an intuitive click-and-drag interface (when in "editing" mode). For example, to rotate about a given bond torsion, one merely has to CTRL-right-click on a bond and then CTRL-left-click to drag atoms on either side of the bond. Bonds, angles, torsions, and positions are all editable using this kind of click-and-drag operation.

PyMOL also enables rudimentary chemical editing of structures on an atomic basis using mouse clicks and CTRL-key combinations to delete, replace, or grow new molecular structures. For example, to make tyrosine out of phenylalanine, one would simply CTRL-middle-click on the para-hydrogen and type CTRL-O to replace it with a hydroxyl.





Molecular editing uses graphical bond and atom "batons" which direct the effects of subsequent mouse clicks and CTRL-key combinations. Here a tryptophan is methylated.

Given PyMOL's existing editing capabilities, a full featured molecular modeling tool could now be implemented on top of PyMOL using only the Python language, without any C or C++ coding. The Python based mutagenesis "wizard" included with PyMOL proves this concept and shows one way such a builder might operate. A crystallography-oriented model building tool is a top development priority, and PyMOL should become suitable for this task in mid-2002. Some ambitious users are already using PyMOL to carry out limited conformational changes late in refinement.

Animations

PyMOL was designed from the ground-up to accommodate multiple atomic coordinates for each atom. Thus, it is possible to load trajectories and conformational ensembles directly into PyMOL for dynamic visualization. These can be viewed and rendered using all of the built-in representations, and there is even a facility present for constructing movies as a programmable sequence of molecular states.

PyMOL can then output images of these states as a series of numbered files for assembly into QuickTime or AVI movies. Ray-traced animations can also be previewed with PyMOL by rendering all frames into memory and then paging through them at machine speed.

Batch Processing and Command-Line Only Mode

PyMOL does not require a graphical user interface in order to run. With a simple command line switch, the program can be launched in command-line mode, which resembles a Python interpreter possessing all of PyMOL's built-in rendering and editing functions. One example use of this feature would be to farm out rendering of a molecular animation to a cluster of Linux workstations for parallel processing.

PyMOL support two related control languages: Python and the PyMOL command language. The PyMOL command language is merely a series of Python function calls with implicit quoted arguments and implicit parentheses. This syntax makes the power of Python directly accessible to non-programmers, and it provides a familiar feel for experienced users of crystallographic software. Because the PyMOL scripting language falls back on Python for evaluation, it can be thought of as a superset of the Python language.

Documentation

Although documentation on PyMOL's features is currently incomplete, a 129+ page user's manual can be downloaded from the PyMOL web site and will help in getting started. PyMOL also supports online help within the program and contains command and argument auto-completion and inference logic to make command-line usage tractable even for novices.

Software Development Features

By default, PyMOL starts up like a typical stand-alone molecular graphics program. However, command line options can be used to change this behavior. While PyMOL ships with a simple Tcl/Tk-based "external" graphical user interface (GUI) and a primitive OpenGL-based "internal" GUI, all such features can be disabled and replaced with new components from an external package. Other applications can simply utilize PyMOL as a molecular display window and provide their own external menus, windows, dialog boxes, and controls. These can be constructed using toolkits such as Tkinter (Tcl/Tk) [13], wxPython (wxWindows) [14], Qt [15], MS-Windows [16], or Mac OSX [17].

PyMOL also allows developers to add additional geometries into the 3D viewing environment using either PyOpenGL [18] (via Callback Objects) or PyMOL's Complied Graphics Objects (CGOs), which represent streams of OpenGL-like drawing commands. The advantage of CGOs is that geometries specified using them can automatically be rendered efficiently in OpenGL using lines and triangles and simultaneously conveyed to the ray-tracer as analytical spheres and cylinders. PyMOL uses CGOs internally for displaying cartoon ribbons and for emulating Raster3D.

So long as external applications are developed using PyMOL's Python-based application programming interface (API), it should be possible for packages to maintain compatibility even as each package evolves independently. PyMOL is thus an attractive platform for any kind of molecular software development project which requires molecular visualization.

A Volitional Approach to Software Funding

PyMOL is an independent software development project of DeLano Scientific [19], a sole proprietorship based in San Carlos, California, USA. Because PyMOL has been given away for free, DeLano Scientific has a microscopic budget, and nearly all PyMOL development has been performed using home equipment and uncompensated personal time. Unlike comparable academic open-source software development efforts, DeLano Scientific receives no direct support from university or government sources for PyMOL.

Instead, funding for PyMOL development comes directly from the voluntary contributions of PyMOL users and developers who are asked to directly support PyMOL out of their own self-interest. Nearly all biomolecular scientists will benefit from the existence of a powerful and universally available graphics program, so everyone has a reason to contribute. While donations or participation are not legally or ethically required to use PyMOL, they encourage development and are what will make future versions possible.

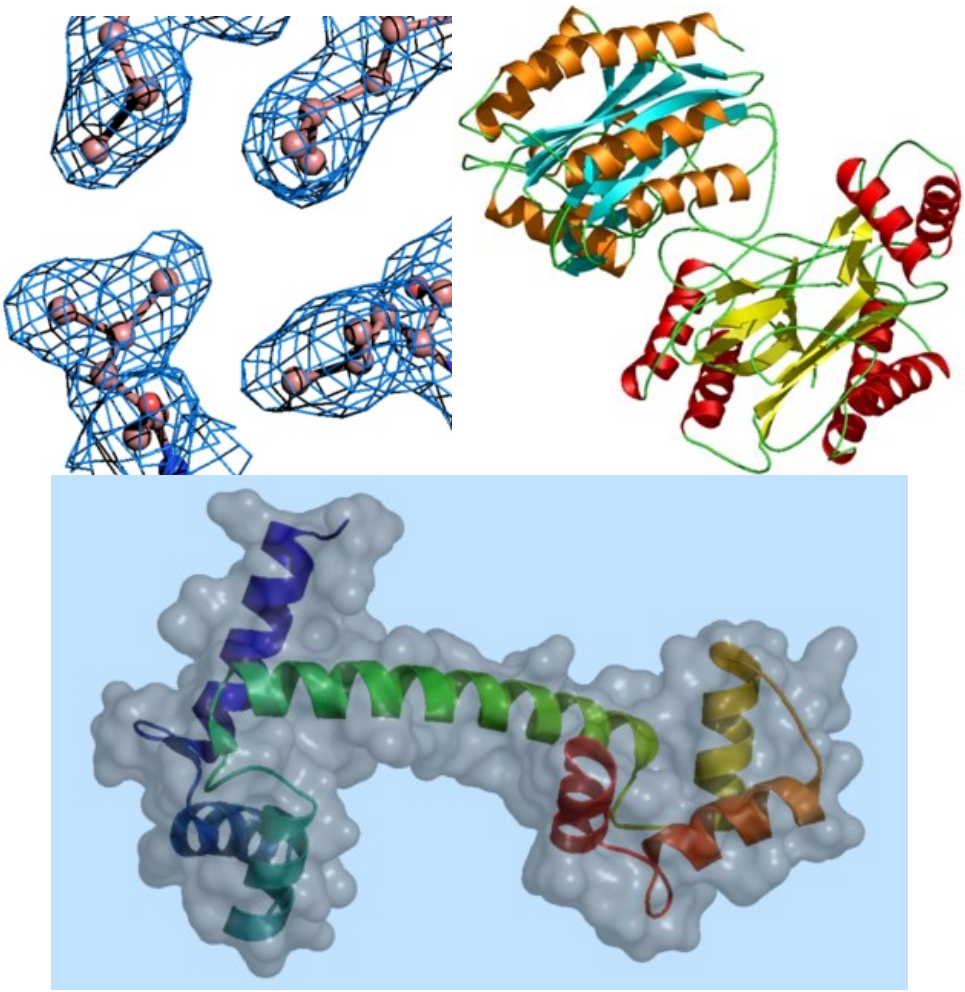
Future Outlook

The outlook for PyMOL is very bright now that robust molecular graphics capabilities are present within the program, and since hundreds of scientists, if not thousands, have already started using PyMOL in their research [20] . Even though PyMOL's C source code was developed extremely rapidly and does not meet professional coding standards, most of the functionality in PyMOL is exposed through a documented Python API. This breakthrough for molecular graphics enables facile reuse of PyMOL source code in many different contexts, without requiring an understanding of it.

Future efforts will continue on a number of fronts, from improving documentation, to adding important new features and polishing the programming interface. There are PyMOL-related efforts ongoing in crystallography, computational chemistry, molecular modeling, simulation, genomics (threading/homology modeling), and education. Eventually PyMOL will become part of an open full-featured molecular computing environment [21].

Obtaining PyMOL

PyMOL can be downloaded for free via the internet at http://www.pymol.org . The web site contains a variety of other useful information, such as a copy of the manual, information about the PyMOL mailing list, recent news, and related links. PyMOL currently runs on a variety of common platforms: Windows, Linux, IRIX, Mac OSX, and Tru64 Unix. Binaries are usually available for Windows and Linux.



Example figures which can be generated in just a few minutes using PyMOL.

Notes

1. Thanks to nVidia's fast and low-cost 3D graphics chips, which support Windows, Macintosh, and Linux http://www.nvidia.com. Graphics cards with nVidia chips can cost less than $100 but outperform common Silicon Graphics hardware for crystallography tasks. Complete molecular graphics workstations can now be built for about $600 (or for about $1000 if one needs stereo graphics).

2. Thanks to the SourceForge open-source development infrastructure, based around the concurrent versioning system (CVS). SourceForge reduces software distribution costs to near zero, and has greatly assisted open-source software development http://www.sourceforge.net .

3. Thanks to the languages Python http://www.python.org, C, and OpenGL http://www.opengl.org, and to various software development strategies inspired by the Extreme Programming (XP) approach: low complexity, minimalist implementation, continuous integration, rigorous testing, and routine user participation http://www.extremeprogramming.org.

4. Examples include: Grasp http://trantor.bioc.columbia.edu/grasp, MidasPlus http://www.cgl.ucsf.edu/Outreach/midasplus/index.html, O http://xray.bmc.uu.se/~alwyn , Molscript http://www.avatar.se , Raster3D http://www.bmsc.washington.edu/raster3d/raster3d.html, and MacroModel http://www.schrodinger.com/Products/macromodel.html .

5. PyMOL is released under the "Python" version 1.5.2 license. In essence, PyMOL source code can be used in any way and in any context so long as DeLano Scientific's copyright notices are not removed. See the PyMOL distribution for license details.

6. Molscript, a commercial package available at http://www.avatar.se, has set the standard for molecular graphics for many years. PyMOL comes very close to reproducing Molscript-quality cartoon ribbons in a free and open-source context.

7. X-PLOR home page http://atb.csb.yale.edu/xplor

8. Information on the PDB file format can be found at http://www.rcsb.org/pdb/info.html.

9. An updated version of SgLite is now available as part of the computational crystallography toolbox (cctbx) http://cctbx.sourceforge.net.

10. Imagemagick is an essential image conversion and display tool http://www.imagemagick.org.

11. Persistence of Vision home page http://www.povray.org

12. The Crystallography and NMR System home page http://cns.csb.yale.edu

13. Tkinter http://www.python.org/topics/tkinter is a Python implementation of the Tcl/Tk API http://tcl.activestate.com

14. wxPython http://www.wxpython.org is a Python implementation of the wxWindows API http://www.wxwindows.org.

15. Qt is a cross-platform Windowing library subject to certain commercial restrictions http://www.trolltech.com.

16. The Microsoft Windows environment http://www.microsoft.com.

17. The Mac OSX environment http://www.apple.com/macosx

18. PyOpenGL is a Python implementation of the OpenGL API http://sf.net/projects/pyopengl.

19. Please note the important PyMOL legal disclaimer: WARREN LYFORD DELANO AND DELANO SCIENTIFIC DISCLAIM ALL WARRANTIES WITH REGARD TO THIS SOFTWARE, INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS. IN NO EVENT SHALL WARREN LYFORD DELANO OR DELANO SCIENTIFIC BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.

20. According to SourceForge statistics in late November, 2001, PyMOL and related files had been downloaded over 12,000 times and generated over 84,000 web site hits since April, 2000.

21. The FreeMOL project aims to create an open alternative to commercial packages like Sybyl (Accelrys), Insight II (Accelrys), Cerius2 (Accelrys), and the Molecular Operating Environment (Chemical Computing Group) http://www.bioinformatics.org/freemol .

Binary Integer Programming and its Use for Envelope Determination

By

Vladimir Y. Lunin[1,2], Alexandre Urzhumtsev[3,†] & Alexander Bockmayr[2]

[1] Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 140292 Russia

[2] LORIA, UMR 7503, Faculté des Sciences, Université Henri Poincaré, Nancy I, 54506 Vandoeuvre-les-Nancy, France; bockmayr@loria.fr

[3] LCM3B, UMR 7036 CNRS, Faculté des Sciences, Université Henri Poincaré, Nancy I, 54506 Vandoeuvre-les-Nancy, France; sacha@lcm3b.uhp-nancy.fr

† to whom correspondence must be sent

Abstract

The density values are linked to the observed magnitudes and unknown phases by a system of non-linear equations. When the object of search is a binary envelope rather than a continuous function of the electron density distribution, these equations can be replaced by a system of linear inequalities with respect to binary unknowns and powerful tools of integer linear programming may be applied to solve the phase problem. This novel approach was tested with calculated and experimental data for a known protein structure.

1. Introduction

Binary Integer Programming (BIP in what follows) is an approach to solve a system of linear inequalities in binary unknowns (0 or 1 in what follows). Integer programming has been studied in mathematics, computer science, and operations research for more than 40 years (see for example Johnson *et al.,* 2000 and Bockmayr & Kasper, 1998, for a review). It has been successfully applied to solve a huge number of large-scale combinatorial problems. The general form of an *integer linear programming problem* is

$$\max \{ \mathbf{c}^T\mathbf{x} \mid A\mathbf{x} \bullet \mathbf{b}, \mathbf{x} \in \mathbf{Z}^n \} \quad (1.1)$$

with a real matrix A of a dimension m by n, and vectors $\mathbf{c} \in \mathbf{R}^n$, $\mathbf{b} \in \mathbf{R}^m$, $\mathbf{c}^T\mathbf{x}$ being the scalar product of the vectors **c** and **x**. If the system $A\mathbf{x} \bullet \mathbf{b}$ includes the constraints $\mathbf{0} \bullet \mathbf{x} \bullet \mathbf{1}$, we get a *binary integer linear programming problem (BIP).* A vector $\mathbf{x^*}$ in $\mathbf{Z}^n$ with $A\mathbf{x^*} \bullet \mathbf{b}$ is called a *feasible solution.* If moreover, $\mathbf{c}^T\mathbf{x^*} = \max \{ \mathbf{c}^T\mathbf{x} \mid A\mathbf{x} \bullet \mathbf{b}, \mathbf{x} \in \mathbf{Z}^n \}$, then $\mathbf{x^*}$ is called an *optimal solution* and $\mathbf{c}^T\mathbf{x^*}$ the optimal value.

In our phasing technique, we have developed an approach that combines a general strategy of low resolution phasing developed recently by (Lunin *et al*., 2000) with a local search procedure for the solution of BIP problems (Walser, 1997, 1998). The general strategy suggests to generate a large number of phase sets and to filter them by some criterion in order to select a relatively small portion of probable solutions. While some of the probable solutions can be quite wrong, the ensemble of the selected phases set is, as a rule, mostly populated by the phase sets which are close enough to the correct solution. Therefore averaging the probable solutions gives a phase set of a high quality. The major problems of this approach are to identify a good selection criterion and to propose an efficient and fast search strategy. The efficiency of the search may be increased if some local refinement is applied to the generated random sets. The local search procedure

realised in the program WSATOIP (Walser, 1997, 1998) allows one to perform this refinement when working with binary integer programming problems. The procedure begins the search with some randomly generated start values for the binary unknowns and then tries to improve the solution locally. In general, the optimisation does not result in the exact solution but in an 'improved' one, compared to the starting point.

2. Binary integer programming and crystallographic objects

Crystallographic problems are usually formulated either in terms of real electron density values or in terms of complex structure factors. These two sets of variables are linked by a linear transformation (Fourier Transform) if the electron density values in all points of the unit cell and the full (infinite) set of structure factors are considered. In practice, when the density is calculated in a grid with a relatively small number of divisions along the unit cell axes, and a small number of reflections is used, these formulae need to be corrected. A possible way to introduce these corrections is to replace the equalities which link density values and structure factors by linear inequalities.

Additionally, the magnitudes of these complex structure factors are supposed to be known from the experiment while the phases are the subject of search. This introduces non-linearity into the problem. Nevertheless, for centric reflections where the phase may take only one of two possible values, phase uncertainty may be represented by a binary variable. The equations that link this variable and the electron density values with the magnitudes are still linear. In order to get a similar situation for acentric reflections, an approximation can be used. The phase of the corresponding structure factors is restricted to one of four possible values $\pm\frac{p}{4}$, $\pm\frac{3p}{4}$ instead of an arbitrary value between *0* and *2*p; this allows one to code the phase uncertainty by two additional binary variables, which are linked also linearly to the density values.

As a rule, macromolecular crystallographers do not need the exact density values but the position and the shape of the region where the density values lie above a certain level, *i.e.,* a binary function representing this region : the molecular envelope, the trace of the polypeptide chain etc. Replacing the object of search by a binary mask has two important consequences. On the one hand, the restriction of the density values to 0 or 1 may enormously reduce the number of possible solutions of the phase problem. On the other hand, the equations connecting the search density values with the experimental structure factors are no longer strictly valid and require a correction.

3. Grid density function and grid structure factors

Let M_1, M_2, M_3 be the number of divisions along the unit cell axes (supposed to be consistent with the symmetry). Let $\mathbf{M} = diag(M_1, M_2, M_3)$ stand for the diagonal matrix with the diagonal formed by M_1, M_2, M_3, Π is the set of all grid points in the unit cell and $|\mathbf{M}| = M_1 M_2 M_3$ is the total number of these points:

$$\Pi = \left\{ \mathbf{j} = (j_1, j_2, j_3)^{\mathrm{T}} : j_1, j_2, j_3 \text{ are integers; } 0 \le j_1 < M_1; 0 \le j_2 < M_2; 0 \le j_3 < M_3 \right\}. \quad (3.1)$$

We introduce the *grid electron density function* $\{\rho^g(\mathbf{j})\}$ as the set of values of the density distribution at the grid points:

$$\rho^g(\mathbf{j}) = \rho\left(\frac{j_1}{M_1}, \frac{j_2}{M_2}, \frac{j_3}{M_3}\right) = \rho\left(\mathbf{M}^{-1}\mathbf{j}\right), \qquad \mathbf{j} \in \Pi, \quad (3.2)$$

and define the *grid structure factors* by the Inverse Discrete Fourier Transform (IDFT):

$$\mathbf{F}^g(\mathbf{h}) = \frac{1}{|\mathbf{M}|}\sum_{\mathbf{j}\in\Pi} \rho^g(\mathbf{j})\exp\left[2\pi i(\mathbf{h},\mathbf{M}^{-1}\mathbf{j})\right], \quad \mathbf{h}\in\Pi. \tag{3.3}$$

The Discrete Fourier Transform (DFT) may restore the grid density function unambiguously from the grid structure factors:

$$\rho^g(\mathbf{j}) = \sum_{\mathbf{h}\in\Pi} \mathbf{F}^g(\mathbf{h})\exp\left[-2\pi i(\mathbf{h},\mathbf{M}^{-1}\mathbf{j})\right], \quad \mathbf{j}\in\Pi, \tag{3.4}$$

but the values of the density distribution in the intermediate points cannot be retrieved. These grid structure factors are linked with the usual structure factors

$$\mathbf{F}(\mathbf{h}) = V_{cell}\int_V \rho(\mathbf{x})\exp[2\pi i(\mathbf{h},\mathbf{x})]d\mathbf{x}, \quad \mathbf{h}\in\mathbf{Z}^3. \tag{3.5}$$

by the formula (Ten Eyck, 1973):

$$V_{cell}\mathbf{F}^g(\mathbf{h}) = \mathbf{F}(\mathbf{h}) + \sum_{\substack{\mathbf{k}\in\mathbf{Z}^3\\ \mathbf{k}\neq\mathbf{0}}} \mathbf{F}(\mathbf{h}+\mathbf{M}\mathbf{k}). \tag{3.6}$$

If a Fourier synthesis of a finite resolution $d_{\min}$ is calculated at a grid whose step length is less than $d_{\min}/2$, then all structure factors in the sum on the right-hand side of (3.6) are supposed to be zero.

4. The phase problem as a binary programming problem

The main goal of this section is to derive linear inequalities that allow one to define the grid electron density values $\{\rho^g(\mathbf{j})\}$ provided the structure factor magnitudes $\{F(\mathbf{h})\}$ are known. Using formulae (3.3) and (3.6), one can write down a system of linear equations defining the values of the grid function $\{\rho^g(\mathbf{j})\}$ in the form

$$\begin{aligned}
\sum_{\mathbf{j}\in\Pi}\cos\left[2\pi(\mathbf{h},\mathbf{M}^{-1}\mathbf{j})\right]\rho^g(\mathbf{j}) &= \frac{|\mathbf{M}|}{V_{cell}}F(\mathbf{h})\cos\varphi(\mathbf{h}) + \mathrm{Re}\,\mathbf{R}(\mathbf{h})\\
\sum_{\mathbf{j}\in\Pi}\sin\left[2\pi(\mathbf{h},\mathbf{M}^{-1}\mathbf{j})\right]\rho^g(\mathbf{j}) &= \frac{|\mathbf{M}|}{V_{cell}}F(\mathbf{h})\sin\varphi(\mathbf{h}) + \mathrm{Im}\,\mathbf{R}(\mathbf{h})
\end{aligned}, \quad \mathbf{h}\in\Pi \tag{4.1}$$

where

$$\mathbf{R}(\mathbf{h}) = \frac{|M|}{V_{cell}}\sum_{\substack{\mathbf{k}\in\mathbf{Z}^3\\ \mathbf{k}\neq\mathbf{0}}}\mathbf{F}(\mathbf{h}+\mathbf{M}\mathbf{k}).$$

These equations link the unknown grid density values linearly with the real and imaginary parts, $F(\mathbf{h})\cos\varphi(\mathbf{h})$ and $F(\mathbf{h})\sin\varphi(\mathbf{h})$, of the structure factors (if both the magnitudes and phases are supposed to be known). However, if not only the density values but also the phases are considered to be unknown, then the equations become non-linear because the phases enter as an argument of trigonometric functions.

The value of $\mathbf{R}(\mathbf{h})$, which depends on magnitudes and phases of all structure factors is generally unknown. Therefore, the equations (4.1) cannot be written in the precise form. In general, the expression $\mathbf{R}(\mathbf{h})$ cannot be neglected if one of the indices is close to $M_1/2$, $M_2/2$, $M_3/2$. At the same time, it can be estimated by the sum of the structure factor magnitudes in the following way:

$$|\mathbf{R}(\mathbf{h})| \le \bar{e}_1(\mathbf{h}) = \frac{|M|}{V_{cell}}\sum_{\substack{\mathbf{k}\in\mathbf{Z}^3\\ \mathbf{k}\neq\mathbf{0}}}F(\mathbf{h}+\mathbf{M}\mathbf{k}). \tag{4.2}$$

As a consequence, the equations (4.1) may be replaced by a system of inequalities that restrict the density values in a weaker form, but do not require the knowledge of all structure factors

$$\begin{aligned} -e_1(\mathbf{h}) &\le \sum_{\mathbf{j}\in\Pi} \cos\left[2\pi(\mathbf{h},\mathbf{M}^{-1}\mathbf{j})\right] \rho^{g}(\mathbf{j}) - \frac{|\mathbf{M}|}{V_{cell}} F(\mathbf{h})\cos\varphi(\mathbf{h}) \le e_1(\mathbf{h}) \\ -e_1(\mathbf{h}) &\le \sum_{\mathbf{j}\in\Pi} \sin\left[2\pi(\mathbf{h},\mathbf{M}^{-1}\mathbf{j})\right] \rho^{g}(\mathbf{j}) - \frac{|\mathbf{M}|}{V_{cell}} F(\mathbf{h})\sin\varphi(\mathbf{h}) \le e_1(\mathbf{h}) \end{aligned}, \quad \mathbf{h}\in\Pi$$

. (4.3)

The inequalities (4.3) contain the phase values $\varphi(\mathbf{h})$ which cannot be determined directly in an X-ray experiment and which are the object of our search. The phases enter the inequalities in a non-linear manner. However, if the reflection **h** is centric then only two values of the phase, $\psi(\mathbf{h})$ or $\psi(\mathbf{h})+\pi$, with ψ being known, are possible, and (4.3) may be written as

$$\begin{aligned} -e_1(\mathbf{h}) &\le \sum_{\mathbf{j}\in\Pi} \cos\left[2\pi(\mathbf{h},\mathbf{M}^{-1}\mathbf{j})\right] \rho^{g}(\mathbf{j}) - a(\mathbf{h})\frac{|\mathbf{M}|}{V_{cell}} F(\mathbf{h})\cos\psi(\mathbf{h}) \le e_1(\mathbf{h}) \\ -e_1(\mathbf{h}) &\le \sum_{\mathbf{j}\in\Pi} \sin\left[2\pi(\mathbf{h},\mathbf{M}^{-1}\mathbf{j})\right] \rho^{g}(\mathbf{j}) - a(\mathbf{h})\frac{|\mathbf{M}|}{V_{cell}} F(\mathbf{h})\sin\psi(\mathbf{h}) \le e_1(\mathbf{h}) \end{aligned},$$

for centric **h**, (4.4)

Here, the phase ambiguity is represented by a new unknown $a(\mathbf{h})$, which takes one of the two values 1 or –1 and which enters the inequalities in a linear way. The inequalities (4.4) become linear with respect to $\{\rho^{g}(\mathbf{j})\}$ and $\{a(\mathbf{h})\}$ provided the structure factor magnitudes $\{F(\mathbf{h})\}$ are known.

For acentric reflections, an approximation can be done such that the phase $\varphi(\mathbf{h})$ can take only one of four values: $\pm\frac{\pi}{4}$, $\pm\frac{3\pi}{4}$. Under this hypothesis, the inequalities (4.3) become

$$\begin{aligned} -e_1(\mathbf{h}) - e_2(\mathbf{h}) &\le \sum_{\mathbf{j}\in\Pi} \cos\left[2\pi(\mathbf{h},\mathbf{M}^{-1}\mathbf{j})\right] \rho^{g}(\mathbf{j}) - a(\mathbf{h})\frac{|\mathbf{M}|}{V_{cell}} F(\mathbf{h})\frac{\sqrt{2}}{2} \le e_1(\mathbf{h}) + e_2(\mathbf{h}) \\ -e_1(\mathbf{h}) - e_2(\mathbf{h}) &\le \sum_{\mathbf{j}\in\Pi} \sin\left[2\pi(\mathbf{h},\mathbf{M}^{-1}\mathbf{j})\right] \rho^{g}(\mathbf{j}) - b(\mathbf{h})\frac{|\mathbf{M}|}{V_{cell}} F(\mathbf{h})\frac{\sqrt{2}}{2} \le e_1(\mathbf{h}) + e_2(\mathbf{h}) \end{aligned},$$

for acentric **h** (4.5)

where the unknowns $a(\mathbf{h})$ and $b(\mathbf{h})$ take one of the two values 1 or –1, and enter the inequalities in a linear way. Here, $e_2(\mathbf{h})$ reflects the error introduced by the sampling of the phase value and can be estimated by

$$e_2(\mathbf{h}) \le \bar{e}_2(\mathbf{h}) = \frac{\sqrt{2}}{2}\frac{|\mathbf{M}|}{V_{cell}} F(\mathbf{h}) \quad . \tag{4.6}$$

As a result we get a system of linear inequalities (4.5) where the unknowns are the values of the electron density at the grid points $\{\rho^{g}(\mathbf{j})\}$, and where the additional variables $a(\mathbf{h})$ and $b(\mathbf{h})$ represent the phase ambiguity. These inequalities are weaker than the initial equations, but they reduce the phase problem to linear integer programming, while initially the phase problem is essentially non-linear.

5. Solution of the BIP phase problem

One of the main difficulties in representing the phase problem as a BIP problem is that the X-ray experiment provides magnitudes $\{F^{obs}(\mathbf{h})\}$ corresponding to a real electron density and not to a binary function approximating it. Nevertheless, tests show that (see Section 6.1), at low and middle resolution, the correlation between the observed structure factor magnitudes and those calculated from binary envelopes may be high enough even for coarse grids. The inequalities may now be written as

$$\begin{aligned} -\mathrm{e}_{\mathbf{h}} - c_{\mathbf{h}}^{R} &\le \sum_{\mathbf{j}\in\Pi} a_{\mathbf{j}}^{R} z_{\mathbf{j}} + b_{\mathbf{h}}^{R} y_{\mathbf{h}}^{R} \le -c_{\mathbf{h}}^{R} + \mathrm{e}_{\mathbf{h}} \\ -\mathrm{e}_{\mathbf{h}} - c_{\mathbf{h}}^{I} &\le \sum_{\mathbf{j}\in\Pi} a_{\mathbf{j}}^{I} z_{\mathbf{j}} + b_{\mathbf{h}}^{I} y_{\mathbf{h}}^{I} \le -c_{\mathbf{h}}^{I} + \mathrm{e}_{\mathbf{h}} \end{aligned}, \qquad \mathbf{h}\in\Pi, \tag{5.1}$$

where $\{z_{\mathbf{j}}\}_{\mathbf{j}\in\Pi}$, $\{y_{\mathbf{h}}^{R}, y_{\mathbf{h}}^{I}\}_{\mathbf{h}\in\Pi}$ are unknown binary variables, which take 0 or 1 values only;

$$y_{\mathbf{h}}^{R} = \frac{\mathrm{a}(\mathbf{h})+1}{2}, \quad y_{\mathbf{h}}^{I} = \frac{\mathrm{b}(\mathbf{h})+1}{2}, \quad \left(y_{\mathbf{h}}^{R} = y_{\mathbf{h}}^{I} \quad \text{for centric reflections}\right), \tag{5.2}$$

$$a_{\mathbf{j}}^{R} = \cos\left[2\pi\left(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j}\right)\right], \quad a_{\mathbf{j}}^{I} = \sin\left[2\pi\left(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j}\right)\right], \tag{5.3}$$

$$b_{\mathbf{h}}^{R} = -2\kappa F(\mathbf{h})\cos\psi(\mathbf{h}), \quad b_{\mathbf{h}}^{I} = -2\kappa F(\mathbf{h})\sin\psi(\mathbf{h}), \quad \text{for the centric case,} \tag{5.4}$$

$$b_{\mathbf{h}}^{R} = -2\kappa F(\mathbf{h})\frac{\sqrt{2}}{2}, \quad b_{\mathbf{h}}^{I} = -2\kappa F(\mathbf{h})\frac{\sqrt{2}}{2}, \quad \text{for the acentric case,} \tag{5.5}$$

$$c_{\mathbf{h}}^{R} = -\kappa F(\mathbf{h})\cos\psi(\mathbf{h}), \quad c_{\mathbf{h}}^{I} = -\kappa F(\mathbf{h})\sin\psi(\mathbf{h}), \quad \text{for the centric case,} \tag{5.6}$$

$$c_{\mathbf{h}}^{R} = -\kappa F(\mathbf{h})\frac{\sqrt{2}}{2}, \quad c_{\mathbf{h}}^{I} = -\kappa F(\mathbf{h})\frac{\sqrt{2}}{2}, \quad \text{for the acentric case.} \tag{5.7}$$

κ is the optimal scale factor which reduces the observed magnitudes to a 'binary function scale', and the gap $\mathrm{e}_{\mathbf{h}}$ reflects three kinds of errors, namely grid sampling errors $\mathrm{e}_1(\mathbf{h})$, phase sampling errors $\mathrm{e}_2(\mathbf{h})$, and errors due to replacing the real density distribution by a binary function.

It should be noted that in space groups different from P1 some variables $\{z_{\mathbf{j}}\}_{\mathbf{j}\in\Pi}$ are linked by the crystallographic symmetry and therefore a set of independent variables must be chosen before solving the system (5.1).

In our tests to solve the phase problem, we used an approach that combines local search for the solution of BIP problems (Walser, 1997, 1998) with a general strategy of low resolution phasing developed recently by (Lunin *et al*., 2000). First, a set of random initial assignments of values to the binary variables is generated. From every initial assignment, one tries to find a feasible solution of (5.1) by local flips of the binary variables. This is done by the procedure WSATOIP (Walser, 1997, 1998). At each run, the program will try to minimise a *residual*, which is defined on the base of (5.1) as

$$R = \sum_{\mathbf{h}} \left\{ r\left(\sum_{\mathbf{j}} a_{\mathbf{j}}^{R} z_{\mathbf{j}} + b_{\mathbf{h}}^{R} y_{\mathbf{h}}^{R}; c_{\mathbf{h}}^{R}, \mathrm{e}_{\mathbf{h}} \right) + r\left(\sum_{\mathbf{j}} a_{\mathbf{j}}^{I} z_{\mathbf{j}} + b_{\mathbf{h}}^{I} y_{\mathbf{h}}^{I}; c_{\mathbf{h}}^{I}, \mathrm{e}_{\mathbf{h}} \right) \right\}. \tag{5.8}$$

Here

$$r(x;q,\varepsilon)=\begin{cases}0 & \text{if} \quad -\varepsilon+q\le x\le q+\varepsilon\\ x-(q+\varepsilon) & \text{if} \quad x>q+\varepsilon\\ (q-\varepsilon)-x & \text{if} \quad x<q-\varepsilon\end{cases} \tag{5.9}$$

so that $r(x;q,\varepsilon)=0$ if the inequality $-\varepsilon+q\le x\le q+\varepsilon$ is satisfied, and $r(x;q,\varepsilon)$ grows linearly with *x* otherwise (see Fig.1). The program stops if the residual has been reduced to 0 (*i.e.* a feasible solution has been found) or if a given maximal number of flips N_{flip} has been reached. So the result of a particular run is not always a feasible solution, but a final assignment where the initial residue has been reduced.

For every final assignment, the phases corresponding to the binary function $\{z_j^{fin}\}_{j\in\Pi}$ are calculated and used together with the observed magnitudes to obtain Fourier syntheses. It must be noted that the possible phase solutions found by this procedure may correspond to different choices of the origin and enantiomer. Therefore, the calculated syntheses are first aligned according to permitted origin and enantiomer choices (Lunin & Lunina, 1996). Then they are averaged to produce a single phase set. In this, way a centroid phase value and an individual figure of merit are defined for every reflection (Lunin *et al*., 2000).

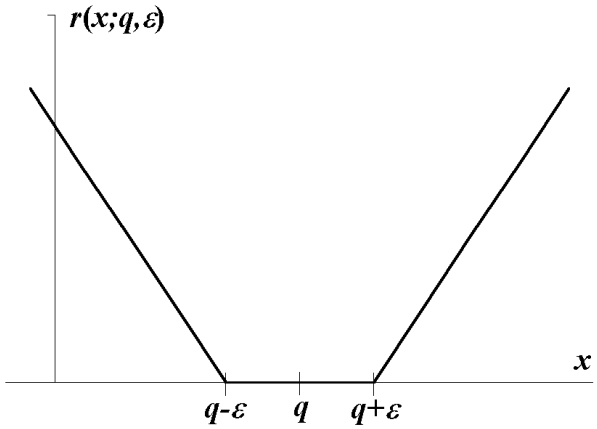


Fig.1. ***The penalty function used for the solution of BIP problem***.

6. Computer tests

The tests were performed with the Protein G data (Derrick & Wigley, 1984). This small protein (61 residues) contains one α-helix and one β-sheet. The protein was crystallised in the space group $P2_12_12_1$ with the unit cell dimensions 34.9×40.3×42.2 Å. The complete low resolution set of experimental diffraction magnitudes was available. The phases calculated from the refined atomic model were considered as the exact ones. The binary density was calculated at the grid N_{grid} x N_{grid} x N_{grid} with N_{var} independent density variables. In all phasing tests, N_{runs} starting randomly generated phases sets were optimised by WSATOIP, N_{flips} flips of binary variables were allowed for every of these runs. Some of the runs ($N_{R=0}$) converged to a set of variables giving the zero value of the criterion R. In any case, all results of the minimisation were analysed by a clustering

procedure, giving each time a small number N_{clust} of phase clusters; one of these clusters (composed from N_{solut} phase sets) always corresponded to the correct solution of the problem. The calculations were done on a Pentium III / 500 PC.

6.1. Binary approximations of Fourier syntheses

The goal of the first series of tests was to check how well small-grid binary functions approximate magnitudes and phases of structure factors. To get a binary approximation for the chosen grid, the Fourier synthesis $\{\rho^{g}(\mathbf{j})\}$ was calculated using the observed magnitudes and the exact phases. The binary approximation values $\{\rho^{bin}(\mathbf{j})\}$ were set to 1 (molecular envelope) for the given number *K* of points with highest synthesis values, and to 0 otherwise. The quality of the approximation depends on this parameter *K* and special tests were performed to determine the optimal *K* values for different grids. It was found that the optimal ratio of the value *K* to the full number of grid points (the one which maximises the correlation) depends on the synthesis resolution and decreases when the resolution increases (Table 1). The grid structure factors $\{F^{bin}(\mathbf{h})\exp[i\varphi^{bin}(\mathbf{h})]\}$ were then calculated and their magnitudes and phases were compared to the true ones (Table 1). This test demonstrated that even at surprisingly small grids, a binary envelope may provide low-resolution phases of a reasonable quality.

Table 1. ***Approximation of the observed structure factors by values calculated from binary maps***

The correlation coefficients

$$C_F = \sum_{\mathbf{h}} \left(F^{bin}(\mathbf{h}) - \left\langle F^{bin} \right\rangle\right)\left(F^{obs}(\mathbf{h}) - \left\langle F^{obs} \right\rangle\right) \Big/ \sqrt{\sum_{\mathbf{h}} \left(F^{bin}(\mathbf{h}) - \left\langle F^{bin} \right\rangle\right)^2 \sum_{\mathbf{h}} \left(F^{obs}(\mathbf{h}) - \left\langle F^{obs} \right\rangle\right)^2}$$

and $$C_\varphi = \sum_{\mathbf{h}} F^{obs}(\mathbf{h})^2 \cos\left(\varphi^{bin}(\mathbf{h}) - \varphi^{exact}(\mathbf{h})\right) \Big/ \sum_{\mathbf{h}} F^{obs}(\mathbf{h})^2$$

are represented for different resolution zones. The molecular volume defines the number *K* of non-zero grid values. It was adapted for every grid to get the maximal correlation coefficient.

Grid (*K* = mol.vol., %)	C_F / C_φ : Resolution range (Å) (Number of independent reflections)				
	16-• (15)	12-• (28)	8-• (85)	5-• (305)	4-• (580)
6*6*6 (50)	0.32 / 0.93	0.39 / 0.74	-	-	-
8*8*8 (35)	0.88 / 0.98	0.92 / 0.94	0.0 / 0.80	-	-
10*10*10 (30)	0.68 / 0.98	0.73 / 0.96	0.68 / 0.90	-	-
16*16*16 (20)	0.91 / 0.99	0.79 / 0.99	0.69 / 0.94	0.62 / 0.87	0.03 / 0.81

6.2. Resolving the phase ambiguity for binary functions

The goal of this test was to study to what extent the condition '0 or 1' allows one to reduce the phase ambiguity. An idealised situation was considered, where the exact magnitudes of the real and imaginary parts of the binary structure factors

$$A(\mathbf{h}) = \left|F^{bin}(\mathbf{h})\cos\varphi^{bin}(\mathbf{h})\right|, \quad B(\mathbf{h}) = \left|F^{bin}(\mathbf{h})\sin\varphi^{bin}(\mathbf{h})\right|, \tag{6.1}$$

were supposed to be known. In this case, the grid values satisfy the equations

$$\begin{aligned} &\sum_{\mathbf{j}\in\Pi} \cos\left[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})\right]\rho^{g}(\mathbf{j}) = a(\mathbf{h})A(\mathbf{h}) \\ &\sum_{\mathbf{j}\in\Pi} \sin\left[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})\right]\rho^{g}(\mathbf{j}) = b(\mathbf{h})B(\mathbf{h}), \end{aligned} \qquad \mathbf{h}\in\Pi \tag{6.2}$$

where the unknowns $a(\mathbf{h})$ and $b(\mathbf{h})$ take one of the values 1 or -1.

The equations (6.2) have a unique solution for any choice of right-hand side values (given by the Fourier transform of these values). So if the grid function is allowed to take any real values, the known magnitudes $\{A(\mathbf{h}), B(\mathbf{h})\}$ do not define the solution uniquely. Any permutation of the signs of $a(\mathbf{h})$ and $b(\mathbf{h})$ will result in a solution of (6.2) possessing the same magnitudes $\{A(\mathbf{h}), B(\mathbf{h})\}$. It may be expected that this is not the case if binary restrictions are added for the unknowns $\{r^{g}(\mathbf{j})\}$:

$$r^{g}(\mathbf{j}) = \{0 \text{ or } 1\} \quad . \tag{6.3}$$

Now an arbitrary choice of the signs of $a(\mathbf{h})$ and $b(\mathbf{h})$ may result in a solution of (6.2) which does not satisfy the condition (6.3). So the binary restrictions may reduce significantly the freedom of choosing the signs and thus may solve the phase problem (or, at least, reduce the phase ambiguity).

Table 2 shows the results of the corresponding tests. It can be noted that in all cases the procedure managed to get the correct solution. For the smallest grid this solution has characteristics similar to those of another, false phase set. Note also the computational difficulties for the largest tested grid .

Table 2. Test results for the resolution of the phase problem

Tests 1-3 were done with known magnitudes of the real and imaginary parts of the structure factors; tests 4-5 were done with known magnitudes of the structure factors, calculated from binary envelopes. For the notation of the columns, see Section 6.

Test N°	N_{grid}	N_{var}	% of '1'	N_{flips}	CPU/run	N_{runs}	$N_{R=0}$	N_{solut}	N_{clust}
1	6	108	50	50,000	2 min	100	73	37	2
2	8	128	35	250,000	30 min	100	80	80	1
3	10	250	30	10,000,000	70 hrs	5	3	3	1
4	6	108	50	50,000	2 min	100	0	47	2
5	8	128	35	250,000	30 min	100	0	19	1

7.3. Known magnitudes for binary envelopes

In a more realistic situation, the estimates (6.1) may be available for centric reflections only. For acentric reflections, only the value $\sqrt{A(\mathbf{h})^2 + B(\mathbf{h})^2}$ of the magnitude of the complex structure factor may be assumed to be known. The goal of the next test series was to study how such uncertainty affects the solution. It was supposed in these tests that the magnitudes $\{F^{bin}(\mathbf{h})\}$ of the binary stucture factors are known exactly, while the magnitudes of their real and imaginary parts were estimated by

$$\tilde{A}(\mathbf{h}) = \frac{\sqrt{2}}{2} F^{bin}(\mathbf{h}) \quad , \quad \tilde{B}(\mathbf{h}) = \frac{\sqrt{2}}{2} F^{bin}(\mathbf{h}) \quad . \tag{6.4}$$

A ‘gap’ was introduced into the equations (6.2) to take into account the errors caused by this approximation

$$\begin{aligned} -0.5\tilde{A}(\mathbf{h}) \le \sum_{\mathbf{j}\in\Pi} \cos\left[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})\right] r^{g}(\mathbf{j}) - a(\mathbf{h})\tilde{A}(\mathbf{h}) \le 0.5\tilde{A}(\mathbf{h}) \\ -0.5\tilde{B}(\mathbf{h}) \le \sum_{\mathbf{j}\in\Pi} \sin\left[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})\right] r^{g}(\mathbf{j}) - b(\mathbf{h})\tilde{B}(\mathbf{h}) \le 0.5\tilde{B}(\mathbf{h}) \end{aligned} \quad , \quad \mathbf{h} \in \Pi \quad . \tag{6.5}$$

Due to these approximations, we cannot expect any longer that the true solution will satisfy (6.5), so the goal was to make the residual value (3.1) as small as possible.

The results of the numerical tests are presented in Table 2 (tests 4 and 5). In test 4, the cluster for the correct solution was slightly smaller (47 phase sets) than the second cluster (49 sets) corresponding to a false phase set. Averaging all the 100 solutions for Test 5 gave the phases with a map correlation coefficient 0.95 with respect to the exact binary phases.

6.4. The use of observed magnitudes

When working with real objects, binary magnitudes are not known and must be estimated somehow. In the following test, the set of observed magnitudes was used to estimate the binary ones. The grid 8*8*8 was chosen for this test as it allows one to solve BIP problems in a reasonable time using the existing software. On the other hand, the approximation of the binary structure factors magnitudes using the observed ones is poor at this grid size. This may significantly influence the results. In order to get more reliable results, BIP methods applicable to larger grids are necessary. The gap in the inequalities (6.5) was reduced to 25% of the estimated $F^{bin}(\mathbf{h})$ value for acentric structure factors, and to 20% for the centric ones. After 100 runs of WSATOIP with random initial assignments, the obtained solutions were aligned and averaged. The found average solution revealed essential features of the 12Å resolution synthesis and had the map correlation coefficient (Lunin & Woolfson, 1993) equal to 0.74 with respect to the exact phases. Fragments of the obtained synthesis overlapped with the atomic model for Protein G are shown in Fig.2.

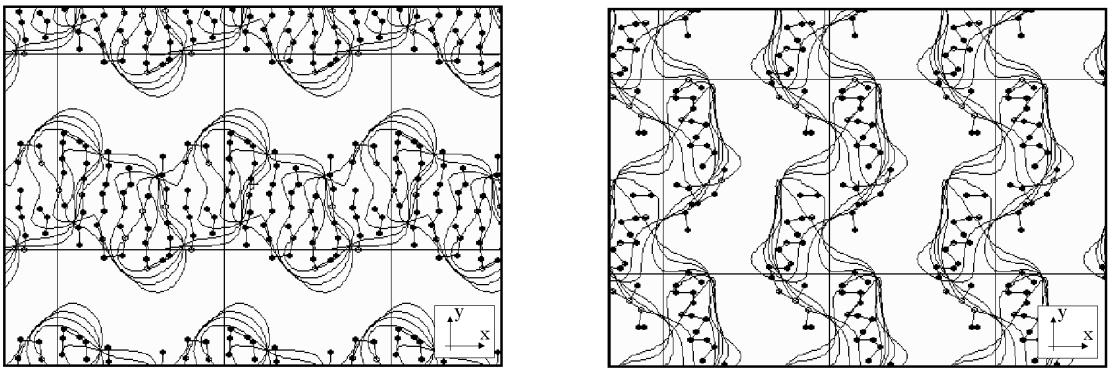


Fig.2. Fragments of BIP-phased Fourier synthesis superimposed with C_α atoms of the model for Protein G: a) projection of the slice z=-2 : 2/40 containing b-sheets; b) projection of the slice z=6:14/40 containing a-helices. The shown contour isolates 35% of the unit cell volume (0.4σ•cutoff level).

7. Conclusions

The theoretical part of this work shows how the crystallographic phase problem can be reduced to the solution of a system of linear inequalities in binary variables. The practical tests with simulated and experimental protein data illustrate the high potential of this new approach. Crystallographic images found from such phasing can be used for further phase improvement or as an important complementary tool for other techniques like molecular replacement. In order to get images of a higher quality, further work on integer programming methods and their application in crystallography is in currently in progress.

Acknowledgements

The work was supported by RFBR grants 00-04-48175 and 01-07-90317 and by the CPER Lorraine 2000-06. Programs O (Jones *et al*., 1991) and CAN (Vernoslova & Lunin, 1993) were used to prepare map illustrations. The authors thank L. Torlay for the computing assistance.

References

Bockmayr, A. & Kasper, T. (1998). *INFORMS J. Computing* **10**, 287-300.

Derrick, J.P. & Wigley, D.B. (1994). *J.Mol.Biol.* **243**, 906-918.

Johnson, E. L., Nemhauser, G. L. & Savelsbergh, M. W. P. (2000). *INFORMS J. Computing* **12**, 2-23.

Jones, T.A., Zou, J.Y., Cowan, S.W. & Kjeldgaard, M. (1991). *Acta Cryst*. A**47**, 110-119.

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

Lunin, V.Y. & Lunina, N.L. (1996). *Acta Cryst.* **A52**, 365-368.

Lunin V.Y., Lunina N.L., Petrova T.E., Skovoroda T.P., Urzhumtsev A.G. & Podjarny A.D. (2000). *Acta Cryst*. D**56**, 1223-1232.

Sayre, D. (1951) *Acta Cryst*., **4**, 362-367

Ten Eyck, L.F. (1973). Acta Cryst. A**29**, 183-191.

Vernoslova, E. A. & Lunin, V. Y. (1993). *J. Appl. Cryst*. 26, 291-294.

Walser, J.P. (1997). In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Innovative Applications of Artificial Intelligence Conference, AAAI 97*, IAAI 97, July 27-31, 1997, Providence, Rhode Island. AAAI Press / The MIT Press, 269-274

Walser, J.P. (1998). In *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference, AAAI 98*, IAAI 98, July 26-30, 1998, Madison, Wisconsin, USA. AAAI Press / The MIT Press, 373-379.

Bulk solvent correction for yet unsolved structures

By A. Fokine & A. Urzhumtsev

Laboratory of Crystallography and Modelling of Mineral and Biological Materials, UPRESA 7036 CNRS, University Henri Poincaré, Nancy I, 54506 Vandoeuvre-lès-Nancy, France

e-mail : fokine@lcm3b.uhp-nancy.fr

Abstract

The bulk solvent correction becomes a routine tool for macromolecular structure determination because the low resolution diffraction data contain an important structural information but cannot be used without such correction. The most reliable of the existing solvent correction methods is the flat solvent model. However, this method can be applied only when an atomic model is already placed in the unit cell; this is necessary in order to estimate two key parameters of the bulk solvent model, k_{sol} and B_{sol} . The statistical analysis of values of these parameters for already resolved structures shows that their fluctuation is relatively weak; as a rule the outliers correspond to incorrectly determined parameters. Therefore, the mean values of k_{sol} = 0.35 e/Å^3 and B_{sol} = 46 Å^2 may be used when refined values cannot be obtained thus extending essentially the limits of the application of the flat bulk solvent model. In particular, such modelling allows to increase drastically the signal in the translation search in molecular replacement.

1. Introduction

The macromolecular crystals contain a large part of disordered (bulk) solvent whose contribution to low resolution reflections is quite significant. An atomic model of a macromolecule without the contribution of the bulk solvent cannot describe these low-resolution diffraction data correctly. On another hand, these data are important to avoid map distortion (Podjarny *et al*., 1981; Urzhumtsev, 1991) and to refine efficiently and correctly the atomic macromolecular model (Kostrewa, 1997). It can be thought also that these data can greatly improve the resolution of the translation problem in Molecular Replacement (MR) method because they are insensitive to reasonable errors in the atomic positions and in the model orientation.

Among several methods allowing to estimate structure factors of the bulk solvent (for a review see Jiang & Brünger 1994; Badger 1997), the flat solvent model (Phillips, 1980; Jiang & Brünger 1994) has been proven to be of a superior quality with respect to others (Jiang & Brünger 1994; Kostrewa, 1997).

In this model the binary function **M** (solvent mask) is introduced which is equal to 1 inside the solvent region and to 0 outside. The structure factors of the bulk solvent are calculated as the scaled Fourier transform • of this function :

$$\mathbf{F_{solv}}(k_{sol},B_{sol}) = k_{sol}\ \exp(-B_{sol}\sin^2(\theta)/\lambda^2)\ \bullet\ (\mathbf{M}) \qquad (1)$$

The unknown parameters k_{sol} and B_{sol} of the bulk solvent are chosen from the best fit of total calculated structure factor $\mathbf{F_{total}}$ to experimental data :

$$\mathbf{G}(k_{sol},B_{sol}) = S\ [\ |\mathbf{F_{obs}}| - |\mathbf{F_{solv}}(k_{sol},B_{sol}) + \mathbf{F_{atoms}}|\ ]^2 \ \text{à min} \qquad (2)$$

where $\mathbf{F}_{atoms}$ are the structure factors calculated from the ordered atoms. Therefore, the knowledge of an atomic model of macromolecule already placed in the crystal is necessary to estimate k_{sol} and B_{sol} , the key parameters of the method.

2. Statistical analysis of the bulk solvent parameters

In order to study the variability of the values of k_{sol} and B_{sol} we analysed their distribution for the structures deposited in the Protein Data Bank (Bernstein *et al*., 1977). The corresponding models have been selected using the provided software (3DB Browser; **http://pdb-browsers.ebi.ac.uk/pdb-bin/pdbmain**). The obtained distribution (Fig. 1) shows that for the most of structures the parameter k_{sol} varies between 0.3 and 0.4 e/Å^3 and B_{sol} varies between 20 and 70 Å^2. The mean values are equal to k^*_{sol} = 0.35 e/Å^3 and B^*_{sol} = 46 Å^2, and the dispersion are 0.03 e/Å^3 and 17 Å^2 , respectively (this statistic was calculated for the models with 0< k_{sol} <0.6 e/Å^3 and 0< B_{sol} <100 Å^2).

A detailed study has been carried out for some outliers with the experimental data available in PDB in order to find the reason for such unusual values of the scale parameters. In all cases the deposited parameters have been found to be incorrect, and the optimal values obtained by us with the complete data set using the systematic search were in the limits reported above.

For such small variation of the scale parameters, the corresponding variation of the structure factors $\mathbf{F}_{solv}(k_{sol},B_{sol})$ is also relatively weak suggesting that the mean values k^*_{sol} and B^*_{sol} can be used when the refined values of the parameters can not be obtained. In particular, they can be used for the molecular replacement when low resolution data are used as it is discussed below. Another application is a map improvement when only a molecular envelope is known (Fokine & Urzhumtsev, 2001).

It should be noted that the distribution of k_{sol} and B_{sol} is quite different from that obtained for similar parameters of the exponential scaling model (Glykos & Kokkinidis, 2000). This can be explained by a more poor quality of this latter model, specially at a middle resolution (Urzhumtsev & Podjarny, 1995a), and by less clear physical meaning of the parameters of the exponential model.

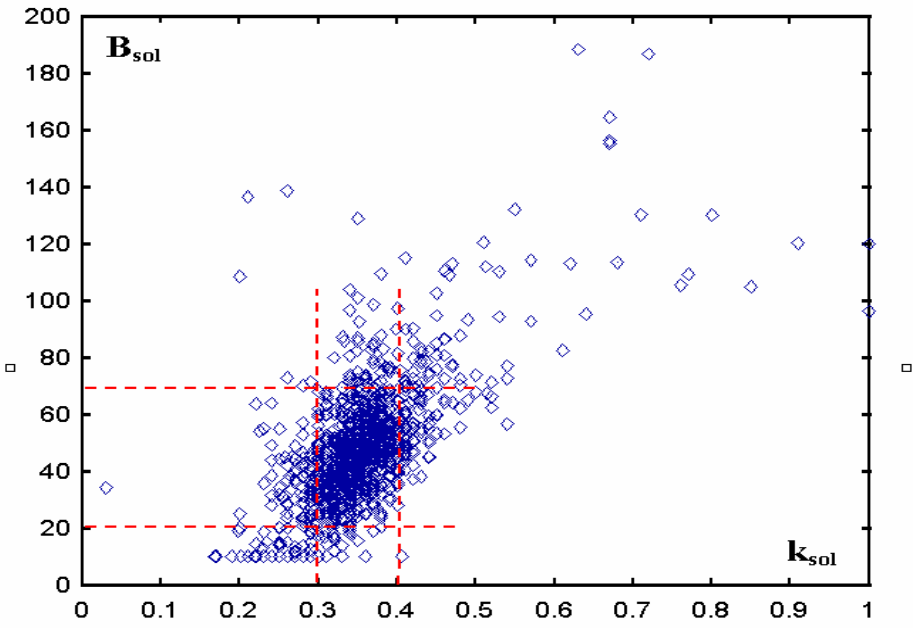


Fig. 1. Distribution of values of parameters k_{sol} and B_{sol} of the flat bulk solvent model for the refined structures deposited in PDB. Each rhomb corresponds to one structure.

3. Bulk solvent correction and fast translation search

It has been shown by Urzhumtsev & Podjarny (1995b) that low resolution reflections being less sensitive to model imperfection (including the errors in its orientation) can be extremely useful for the solution of the translation problem. However, standard molecular replacement protocols, except those by Glykos & Kokkinidis (2001) where the exponential model is used, do not use reflections with the resolution lower than 10-15 Å because they are strongly influenced by the bulk solvent.

For the translation search, the bulk solvent correction eventually can be done at every position of the search model (while the obtained k_{sol} and B_{sol} can be completely unreasonable for wrong positions). Unfortunately, such way of solvent correction cannot be included into fast translation algorithms (Navaza, 1994; Navaza & Vernoslova, 1995) making its practical application inefficient. However, the following observations can be done :

a) a) For the positions in the unit cell where the search model does not overlap with its symmetrically related copies, the mask of the region occupied by all molecules can be calculated as a junction of masks of individual molecules related by symmetries; as a consequence, the structure factors of such total molecular envelope can be rapidly recalculated from the structure factors of the envelope of a single model;

b) b) if the structure factors of the envelope of a single model are preliminary scaled by k^*_{sol} and B^*_{sol}, such total structure factors summarised over the symmetries give a good estimation of the bulk solvent structure factors;

c) c) for all non-overlapping model positions, such scaled structure factors from the envelope being added to the structure factors from the atomic model and expanded over all symmetries are structure factors from the whole content of the unit cell;

d) d) as a conclusion, a fast FFT-based translation search (Navaza, 1994; Navaza & Vernoslova, 1995) done using bulk-solvent-corrected structure factors instead of the values calculated directly from the atomic model allows to compare correctly the magnitudes of all reflections including those at low resolution for all non-overlapping positions.

Naturally, spurious peaks in the translation function are eventually possible for the positions where the models overlap; however, these spurious peaks will be eliminated by the packing criterion and will not appear in the final list anyhow.

4. Tests protocols

A good approximation for the molecular envelope can be available from a more or less complete atomic model. This is the case when NMR models are used as the search models for the molecular replacement. Several such cases reported as most difficult (for a corresponding review see Chen *et al*., 2000; Chen, 2001) were chosen to test the suggested approach of the improvement of the translation function (Table 1).

All test calculations were done with experimental data, and the orientation of the search models was supposed to be known (it can be noted that typical errors of about 5° in model orientation practically did not influence the searches when low resolution reflections were included). All translation searches were made with CNS (Brünger *et al*., 1998) using the fast translation function (Navaza & Vernoslova, 1995). The translation search parameters were taken without any optimisation (Chen *at al*., 2000); complete NMR models were taken as they are in the PDB; the B-factors for all atoms of the search models were assigned to be equal to 20 $Å^2$. In each test, a single NMR model was used for the translation search.

Table 1. Test structures : summary information

Protein name (reference)	PDB ID / NMR ID	Space group and unit cell parameters a,b,c (in Å)	Percentage of the solvent in the unit cell
Human interleukin-4 (Müller *et al*.,1995)	1hik 1bcn	$P4_12_12$ 92.1, 92.1, 46.4	63
P53 Tetramerization Domain (Mittl *et al*., 1998)	1aie 1pet	P422 45.5, 45.5, 32.2	53
Corn Hageman Factor Inhibitor (Behnke *et al*. 1998)	1bea 1bip	$P4_22_12$ 57.12, 57.12, 80.24	49

5. Results of improved translation searches

Figure 2 shows the results of the translation searches performed with and without low resolution data, with and without the bulk solvent correction using three experimental data sets. Each diagram shows the results of the translation search at a given resolution shell. The top diagrams show the peaks obtained in the translation search without any bulk solvent correction; the down diagrams show the peaks obtained at the same conditions when the bulk solvent correction was taken into account as suggested above. The height of each peak is shown in percents to the height of the first peak of the corresponding search, and the correct solution is indicated in red. It may be reminded that the total computation time for both type of the translation function was the same due to the fast correction procedure described in the previous section.

For human interleukin-4 (Müller *et al*., 1995), the translation search performed at the standard resolution of 4-15 Å without solvent correction gave the solution as the second peak. Including of all available reflections with the resolution lower than 15 Å brought the correct peak to the first position. Bulk solvent correction increased the contrast of the signal drastically, specially when low resolution data were included (Fig. 2a).

For p53 tetramerization domain (Mittl *et al*., 1998), the translation search without solvent correction at standard 4-15 Å resolution gave the correct solution hidden in noise and the search at 3-15 Å resolution gave it slightly higher in the list. With the solvent correction, the peak for the solution became the first with the best contrast at 4 Å even when no more low resolution data are available (Fig. 2b).

CHFI (Behnke *et al*., 1998) was reported as the worst case among all NMR-based searches (Chen *et al*., 2001). The multiple rotation function (Urzhumtseva & Urzhumtsev, 2001) allowed to find the orientation of the search model quite unambiguously and precisely and it was supposed to be known for the translation search. Without the bulk solvent correction, the solution did not appear among 10 highest peaks neither at the resolution 5-15 Å nor at 4-15 Å. When all available magnitudes with the resolution lower than 5 Å were used, the correct peak was the 7^{th} in height. At the same time, with the bulk solvent correction, this peak became the first one for the resolution lower than 5 Å while the contrast is not so high as for two previous cases (Fig. 2c).

Therefore, it can be concluded that the bulk solvent correction using the flat solvent model improves drastically the translation function, quite differently from the correction by the exponential model where no significant improvement has been observed (Glykos & Kokkinidis, 2001).

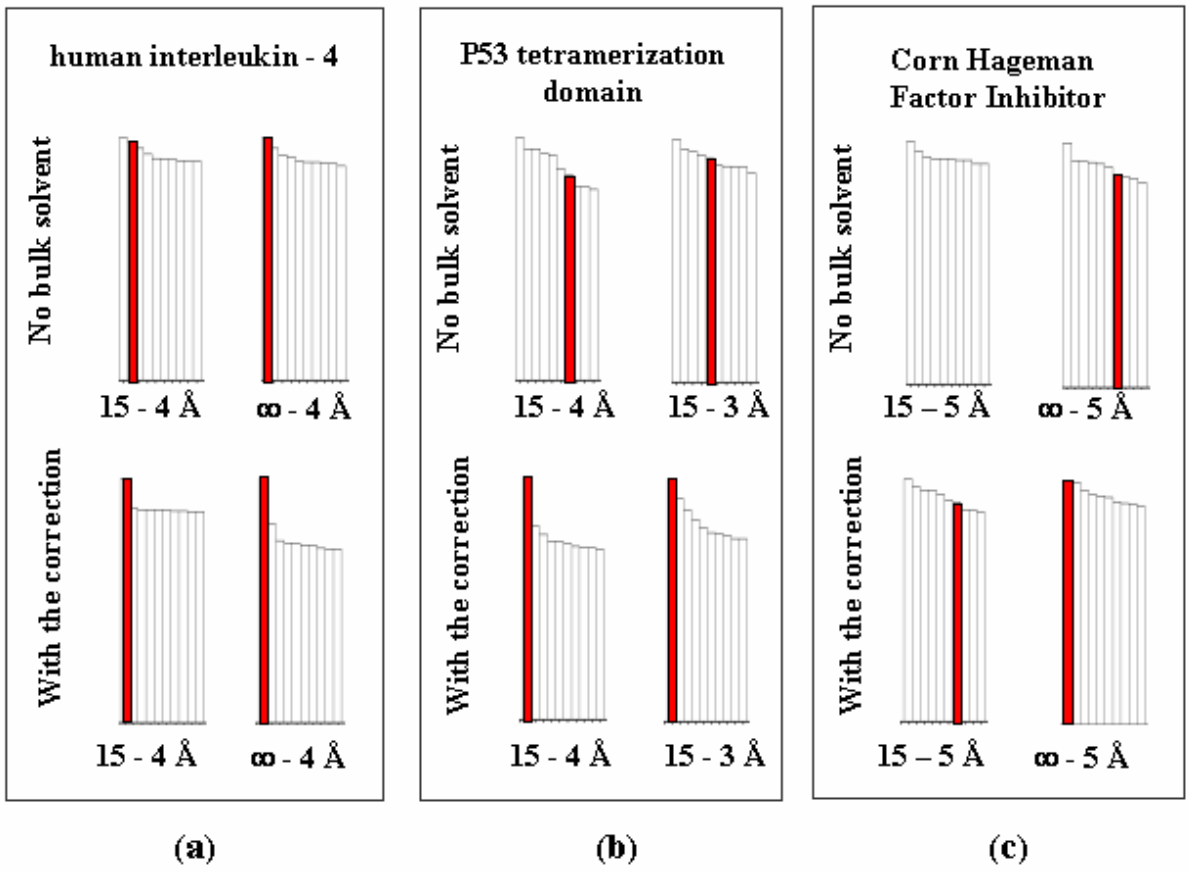


Fig. 2 Relative heights of peaks of the translation function
Each column represents a peak of the translation function and has a height in percents to the value of the first (highest) peak obtained at the same translation search. For each resolution range ten highest peaks corresponding to searches without solvent correction are shown at the top diagram and ten highest peaks corresponding to searches with such correction are shown at the bottom diagram. Peaks corresponding to the correct solution are shown in red. The figures represent the results of the searches for :

a) a) human interleukin-4 (Müller *et al*., 1995),
b) b) p53 tetramerization domain (Mittl *et al*., 1998),
c) c) Corn Hageman Factor Inhibitor (Behnke *et al*., 1998).

6. Physical meaning and possible values of the parameters of the flat solvent model

It has been discussed many times, that the parameter k_{sol} describes the mean electron density of crystallisation solution. Kostrewa indicated (1997) that the electron density of pure water is 0.33 e/Å^3, the density of 4M ammonium sulphate is 0.41 e/Å^3, so normally the value of k_{sol} should vary between these limits which corresponds well to the distribution found from the PDB analysis (Section 2).

It is clear that the parameter B_{sol} describes the sharpness of the solvent density at its border but his physical meaning has not been discussed previously. The larger is B_{sol} the deeper the electron density of the solvent penetrates to the macromolecular region and therefore very large values of B_{sol} are meaningless. On the other hand, the distance interval on which the electron density of the solvent decreases to zero should be at least larger than the radius of the solvent molecule (1.4 Å).

In fact, we have found that the optimal value of this parameter corresponds to the mostly flat electron density distribution at the border between the solvent and molecular

regions (details will be published elsewhere; manuscript in preparation). Again, these values agree well with the distribution found statistically.

It can be noted that a non optimal choice of parameters does not allow to fit equally well all calculated data to the experimental values and usually leaves elevated R-factor for lowest resolution reflections. In most of cases, such wrong choice can be avoided either by a systematic search or by a local search for k_{sol} and B_{sol} as it is realised in CNS (Brünger *et al*., 1998) but starting from k^*_{sol} and B^*_{sol}, differently from the currently existing procedure.

7. Conclusions

The distribution of values of the bulk solvent parameters k_{sol} and B_{sol} for crystallographic structures deposited in Protein Data Bank shows that their correct values vary in relatively small limits around k^*_{sol} = 0.35 e/Å^3 and B^*_{sol} = 46 Å^2. These limits and corresponding mean values have a reasonable physical interpretation; k_{sol} corresponds to the means electron density of the solvent and the optimal value of B_{sol} provides with the smooth and flat transition of the electron density between the solvent and molecular regions.

For a known atomic model in the unit cell, the optimal values of the bulk solvent parameters can be found either by systematic or by a local search; in the latter case, the start from k^*_{sol} and B^*_{sol} allows to avoid a wrong answer.

When the standard procedure can not be applied to obtain the optimal values of the parameters, for example when an atomic model in the unit cell is not known yet, the mean values k^*_{sol} and B^*_{sol} can be used instead of the optimal values.

In particular, this latter allows to include the bulk solvent correction using the flat solvent model into fast calculation of the translation function. The use of low resolution reflections with such bulk solvent correction improves drastically the signal in the translation search.

References

Badger, J. (1997). *Methods Enzymology*. 277, 344-352.

Behnke, C.A., Yee, V.C., Le Trong, I., Pedersen, L.C., Stenkamp, R.E., Kim, S.-S., Reeck, G.R. & Teller, D.C. (1998). *Biochemistry*, **37**, 15277-15288.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E. Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535-542.

Brünger, A.T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst*., D**54**, 905-921.

Chen, Y.W., Dodson, E.J. & Kleywegt, G.J. (2000). *Structure*, **8**, 213-220.

Chen, Y.W., (2001). *Acta Cryst.* D**57**, 1457-1461.

Fokine, A. & Urzhumtsev, A. (2001). *CCP4 Newslett*. **39**, 71-78.

Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst*. D**56**, 1070-1072.

Glykos, N. M. & Kokkinidis, M. (2001). *Acta Cryst.* D**57**, 1462-1473

Jiang, J.-S. & Brünger, A. T. (1994). *J. Mol. Biol*. **243**, 100-115.

Kostrewa, D. (1997). *CCP4 Newslett*. **34**, 9-22.

Mittl, P., Chène, P. & Grütter, M.G. (1998). *Acta Cryst*. D**54**, 86-89.

Müller, T., Oehlenschläger, F. & Buehner, M. (1995) *J.Mol.Biol*., **247**, 360-372.

Navaza, J. (1994). *Acta Cryst.* A**50**, 157-163.

Navaza, J. & Vernoslova, E. (1995). *Acta Cryst*., A**51**, 445-449.

Phillips, S. E. V. (1980). *J.Mol.Biol*., **142**, 531-554.

Podjarny, A. D., Schevitz, R. W. & Sigler, P. B. (1981). *Acta Cryst.* A**37**, 662-668.
Urzhumtsev, A. G. (1991). *Acta Cryst*. A**47**, 794-801.
Urzhumtsev, A.G. & Podjarny, A.D. (1995a) *CCP4 Newslett.* **31**, 12-16.
Urzhumtsev, A.G., Podjarny, A.D. (1995b) *Acta Cryst.,* **D51**, 888-895
Urzhumtseva, L. & Urzhumtsev, A.G. (2001) *CCP4 Newslett.* **39**,79-85

Search of the optimal strategy for refinement of atomic models

P. Afonine[§,*], V.Y. Lunin[#,*] & A. Urzhumtsev[*]

[$] *Centre Charles Hermite, LORIA, Villers-lès-Nancy, 54602 France*

[#] *IMPB, Russian Academy of Sciences, Pushchino, 142290, Moscow Region, Russia*

[*] *LCM3B, UPRESA 7036 CNRS, Université Henri Poincaré, Nancy 1, B.P. 239, Faculté des Sciences, Vandoeuvre-lès-Nancy, 54506 France*

e-mail: afonine@lcm3b.uhp-nancy.fr

1. 1. Introduction

Recently it has been shown (Afonine *et al*, 2001; Lunin *et al*., 2002) that the approximation of the maximum likelihood criterion (ML) by a quadratic functional (Lunin & Urzhumtsev, 1999) allows to understand the features of the ML refinement and its advantages with respect to the traditional least-squares (LS) refinement. In this latter, the magnitudes $\{F_s^{calc}\}_{s\in S}$ of structure factors calculated from the current atomic model are fitted to the observed structure factor magnitudes $\{F_s^{obs}\}_{s\in S}$ by minimisation of

$$Q_{LSQ} = \sum_{s\in S} w_s \left(F_s^{calc} - F_s^{obs}\right)^2 \quad , \qquad (1)$$

The weights $\{w_s\}_{s\in S}$ may reflect the accuracy of the observed magnitudes or some other effects, but most frequently the unit weights are used.

In the procedure that is usually referenced as ML-refinement the minimised criterion is the negative logarithm of the likelihood, the model-dependent part of which may be presented as

$$Q_{ML} = \sum_{s\in S} \Psi\left(F_s^{calc}; F_s^{obs}, a_s, b_s\right) \;\Rightarrow\; \min \qquad (2)$$

with

$$\Psi = \begin{cases} \Psi_a = \dfrac{a_s^2 \left(F_s^{calc}\right)^2}{e_s b_s} - \ln\left(I_0\left(\dfrac{2 a_s F_s^{calc} F_s^{obs}}{e_s b_s} \right)\right) & \text{for acentric reflections} \\ \Psi_c = \dfrac{a_s^2 \left(F_s^{calc}\right)^2}{2 e_s b_s} - \ln\left(\cosh\left(\dfrac{a_s F_s^{calc} F_s^{obs}}{e_s b_s} \right)\right) & \text{for centric reflections} \end{cases} . \qquad (3)$$

For every reflection, its parameter e_s depends on the reflection indexes and particular space group and the statistical parameters a_s and b_s, being the functions of the resolution, reflect the precision of the atomic parameters and the completeness of the model (see for example Lunin & Urzhumtsev, 1984; Read, 1986; Lunin & Skovoroda, 1997; Skovoroda & Lunin, 2000).

The approximation of the criterion (2,3) by a quadratic functional means its substitution by a functional

$$\tilde{Q}_{ML} = \sum_{\mathbf{s}\in S} w_{\mathbf{s}}^{*}\left(F_{\mathbf{s}}^{calc} - F_{\mathbf{s}}^{*}\right)^2 \quad . \qquad (4)$$

where the target values $F_{\mathbf{s}}^{*}$ are no longer the observed magnitudes and the non-unit weights $w_{\mathbf{s}}^{*}$ are crucial for a successful refinement. The minimisation of this function we will call LS*-refinement.

Previously (Lunin *et al*., 2002) we have discussed that $F_{\mathbf{s}}^{*}$ and $w_{\mathbf{s}}^{*}$ in (4) may be represented as

$$F_{\mathbf{s}}^{*} = \frac{\sqrt{\mathrm{e}_{\mathbf{s}}\mathrm{b}_{\mathbf{s}}}}{\mathrm{a}_{\mathbf{s}}}\,\mathrm{m}\left(\frac{F_{\mathbf{s}}^{obs}}{\sqrt{\mathrm{e}_{\mathbf{s}}\mathrm{b}_{\mathbf{s}}}}\right), \quad w_{\mathbf{s}}^{*} = c_{\mathbf{s}}\,\frac{\mathrm{a}_{\mathbf{s}}^{2}}{\mathrm{e}_{\mathbf{s}}\mathrm{b}_{\mathbf{s}}}\,\mathrm{n}\left(\frac{F_{\mathbf{s}}^{obs}}{\sqrt{\mathrm{e}_{\mathbf{s}}\mathrm{b}_{\mathbf{s}}}}\right), \qquad (5)$$

where *m(p)* and *n(p)* are some functions defined in Lunin *et al*. (2002) and whose behaviour explains the features of the ML refinement.

Formula (5) shows that the parameters $\mathrm{a}_{\mathbf{s}}$ and $\mathrm{b}_{\mathbf{s}}$, play the key role in the estimation of $F_{\mathbf{s}}^{*}$ and $w_{\mathbf{s}}^{*}$ and therefore in the whole refinement. In this article we discuss the best choice of $\mathrm{a}_{\mathbf{s}}$ and $\mathrm{b}_{\mathbf{s}}$.

2. 2. Estimation of $\mathrm{a}_{\mathbf{s}}$ and $\mathrm{b}_{\mathbf{s}}$

Several approaches can be suggested to estimate the parameters $\mathrm{a}_{\mathbf{s}}$ and $\mathrm{b}_{\mathbf{s}}$. If there exists some probabilistic hypothesis about irremovable errors in the atomic model (for example, about a missing part of the model) then for several particular cases these parameters may be calculated explicitly (Urzhumtsev *et al*., 1996). In particular, in the case of an incomplete model, if the absent atoms are supposed to be distributed uniformly in the unit cell, these parameters may be calculated as

$$\mathrm{a}_{\mathbf{s}} = 1 \quad \text{and} \quad \mathrm{b}_{\mathbf{s}} = \sum_{k=M+1}^{N} f_k^{2}(s), \qquad (6)$$

where $f_k(\mathbf{s})$ are atomic scattering factors of the absent atoms. It should be noted that in practice the exact number of missed atoms and their scattering factors can be known only approximately (for example, it is difficult to know the exact number of missed ordered solvent molecules).

Another way is to use likelihood-based estimates of these parameters when comparing the observed structure factor magnitudes with the ones corresponding to a starting atomic model (Lunin & Urzhumtsev, 1984; Read, 1986). It is important to note that the test set reflections (Brünger, 1992) only should be used (Lunin & Skovoroda, 1995; Skovoroda & Lunin, 2000). Eventually, these estimates can be recalculated iteratively during refinement.

These different ways to estimate $\mathrm{a}_{\mathbf{s}}$ and $\mathrm{b}_{\mathbf{s}}$ have been tested by comparison of LS-, ML- and various LS*-refinement approaches in order to suggest the best refinement strategy.

3. 3. Models and programs used for tests

Similarly to the previous work (Afonine *et al*., 2001; Lunin *et al*., 2002), the tests were carried out with CNS complex (Brünger *et al*., 1998) using the model of Fab fragment of monoclonal antibody (Fokine *et al.*, 2000) which consists of 439 amino acid residues and 213 water molecules, 3593 atoms in total. The crystal belongs to the space group $P2_12_12_1$ with the unit cell parameters a = 72.24 Å, b = 72.01 Å, c = 86.99 Å, one molecule per asymmetric unit.

For test purposes the values of F_{obs} at 2.2 Å resolution were simulated by the corresponding values calculated from the complete exact model and were used for all refinements. The errors in the atomic co-ordinates were introduced randomly and independently. Incomplete models were obtained by random deletion of atoms, both from the macromolecule and from the solvent.

4. 4. Choice of a_s and b_s

Several refinement strategies based on different choice of F_s^* and w_s^* through different estimation of a_s and b_s have been compared.

First of all, the parameters a_s and b_s have been calculated using the technique described previously (Lunin & Skovoroda, 1995; Skovoroda & Lunin, 2000) through the comparison of the $\{F_s^{obs}\}_{s\in S}$ magnitudes with the structure factors $\{F_s^{calc}\}_{s\in S}$ calculated from the starting model. These values were kept for the whole refinement process consisted of 800 cycles.

Secondly, the same method of the estimation of a_s and b_s has been applied but their values were recalculated every 400 or 200 refinement cycles, depending on the test.

Alternatively, the refinement was carried out using the estimations (6). In these tests the exact number of missed atoms and their scattering factors were supposed to be known.

Finally, the refinement was carried out with the mixed parameter values, a_s = 1 for all reflection as in (6) and b_s estimated from the comparison of $\{F_s^{obs}\}_{s\in S}$ with $\{F_s^{calc}\}_{s\in S}$.

The start models with the mean coordinate errors of 0.5 and 0.7 Å respectively and with 0.5% and 3.0% of incompleteness were optimised using LS*-criterion (4). For comparison, corresponding LS- and ML-refinements were also done. The results of these tests are shown in Table 1. It can be remarked that, as it has been discussed (Afonine *et al*., 2001; Lunin *et al*., 2002), even a small quantity of absent atoms can already strongly influence on the quality of the refined model.

Table 1. Mean coordinate errors in the model after refinement using different criteria. Starting models have mean coordinate errors of D_{st}. The incompleteness D_{abs} of the models of 0.5% and 3.0% correspond to 18 and 108 atoms deleted, respectively. The number of cycles indicates the frequency with which the parameters of the corresponding criterion were recalculated (the frequency of parameters updating is not definitely known for ML). a_F and b_F stand for the parameters estimated from the magnitude comparison and b_C stands for values calculated from (6). The final coordinate errors shown in italic indicate the cases where this error is higher than the starting error. The numbers in bold indicate the best refinement protocol for the given model.

criterion		LS* a_F,b_F			LS* a=1,b_F			LS* a=1,b_C	LS	ML
No of cycles		1*800	2*400	4*200	1*800	2*400	4*200	1*800	1*800	800*1?
D_{st}	D_{abs}	final error								
0.5Å	0.5%	0.320	0.140	**0.103**	0.358	0.156	0.127	0.111	0.212	0.108
	3.0%	0.453	0.345	0.397	0.475	0.353	0.311	**0.247**	0.375	0.305
0.7Å	0.5%	*0.784*	0.636	0.468	0.633	0.491	0.388	**0.284**	0.397	0.353
	3.0%	*0.803*	*0.711*	0.592	0.700	0.599	0.527	**0.404**	0.530	0.537

5. Influence of errors in the estimation of b_s

The LS*-refinement with the parameters estimated through (6) gives systematically better results in comparison with other known strategies and was chosen as the best one for further tests. The estimation of b_s in (6) depends on the number of missed atoms, on their type and on their temperature factors. The influence of possible errors in the estimation of these parameters from these 3 sources on the results of refinement has been studied.

First of all, the missed atoms were simulated by oxygens or by carbons with temperature factors as they were in the corresponding deleted atoms. No significant influence of such modification of the atomic type has been found (Table 2).

Table 2. Mean coordinate errors after LS*-refinement with the estimations (6) for different type assigned to missed atoms; CNO stands for the exact (mixed) type of atoms. Starting models have mean coordinate errors of D_{st} (in Å). D_{abs} is incompleteness of the models in percents; the number in parenthesis is the corresponding number of deleted atoms. The final coordinate errors shown in italic indicate the cases where this error is higher than the starting error.

D_{st}	D_{abs} Type	0.5 (18)	1.0 (36)	3.0 (108)	5.0 (180)	7.0 (252)	9.0 (325)
0.5 Å	CNO	0.105	0.133	0.256	0.343	0.447	*0.513*
	O	0.111	0.138	0.247	0.357	0.450	*0.521*
	C	0.113	0.136	0.256	0.343	0.439	0.499
0.7 Å	CNO	0.289	0.321	0.422	0.498	0.579	0.649
	O	0.284	0.278	0.404	0.468	0.598	0.645
	C	0.285	0.334	0.425	0.494	0.609	0.656

Table 3. Mean coordinate errors for different values <B> of the mean temperature factor assigned to missed atoms. Starting models have mean coordinate errors of D_{st} (in Å). D_{abs} is incompleteness of the models; the number in parenthesis is the corresponding number of deleted atoms. The final coordinate errors shown in italic indicate the cases where this error is higher than the starting error.

<B>, Å²	D_{st} D_{abs}=	0.5 (18)	1.0 (36)	3.0 (108)	5.0 (180)	7.0 (252)	9.0 (325)
5	0.5 Å	0.085	0.129	0.281	0.386	*0.528*	*0.582*
15		0.083	0.120	0.259	0.342	0.455	*0.508*
25		0.109	0.144	0.258	0.347	0.440	*0.506*
35		0.144	0.167	0.272	0.353	0.449	0.483
45		0.170	0.207	0.290	0.380	0.470	*0.507*
5	0.7 Å	0.178	0.233	0.378	0.508	0.626	0.693
15		0.264	0.274	0.377	0.474	0.565	0.610
25		0.304	0.356	0.431	0.522	0.605	0.655
35		0.374	0.432	0.494	0.595	0.677	*0.703*
45		0.517	0.581	0.599	*0.719*	*0.774*	*0.781*

To study the influence of the estimated temperature factor on the minimisation process, the known values of B-factors of missed atoms (following the results of previous test, all these atoms were assigned to be carbons) were considered to be equal to the same value which varied from 5 to 80 $Å^2$ in a series of runs while the mean value of the temperature factor for the deleted atoms varied in the limits 27-29 $Å^2$. Table 3 shows that the variation of the estimated temperature factors of missing atoms by ±15 $Å^2$ around the mean values does not seriously affect the quality of the refined model.

Finally, the influence of a wrong estimation of the number of missed atoms has been studied. For this purpose the start model with 5.0% (180 atoms) of deleted atoms and introduced error of 0.5 Å was generated. Different estimations of the number of missing atoms were used to get the b_s values and corresponding F_s^* and w_s^*. The error in this number of order of at least 25% practically did not influence the final coordinate errors.

6. 6. Conclusions

The quadratic approximation of the maximum-likelihood-based criterion allows to understand better the features of the ML-based refinement and its advantages. Even more, this approximation allows to choose a better refinement strategy and to build its new quadratic functional the minimisation of which leads to better models that those obtained both by traditional LS and ML-based refinement.

In this quadratic functional, the corresponding target values F_s^* and the weights w_s^* are calculated using formulas (6) for the parameters a_s and b_s of the variable part of the likelihood function (2,3). These formulas allow to get such estimation for the ideally refined model without knowing directly its parameters and therefore to build the quadratic approximation of (2,3) at the point of its minimum and thus to improve the refinement criterion.

These estimations of a_s and b_s are quite insensible to the choice of the type of atoms supposed to be missed, to their mean B-factor estimation and to the estimation of the number of such missed atoms making such new refinement strategy quite robust.

Acknowledgment

The work was supported partially by RFBR grants 00-04-48175 and 01-07-90317, by CNRS, UHP and Region Lorraine through financial support. The authors thank C. Lecomte and E. Dodson for their interest to the project.

References

Afonine, P., Lunin, V.Y. & Urzhumtsev, A.G. (2001). *CCP4 Newsletter on Protein Crystallography*, **39**, 52-56.

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

Brünger, A.T., Adams, P.D., Clore, G.M., DeLabo, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N.S., Read,

R.J., Rice, L.M., Simonson, T. & Warren, G.L. (1998) *Acta Cryst.* D**54**, 905-921.

Fokine, A.V., Afonine, P.V., Mikhailova, I.Yu., Tsygannik, I.N., Mareeva, T.Yu., Nesmeyanov, V.A., Pangborn, W., Li, N., Duax, W., Siszak, E., Pletnev, V.Z. (2000). *Rus. J. Bioorgan. Chem.*, **26**, 512-519.

Lunin, V.Y., Afonine, P.V., Urzhumtsev, A. (2002). *Acta Cryst.*, A, in press.

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

Lunin, V.Y. & Urzhumtsev, A. (1984). *Acta Cryst.*, **A40**, 269-277

Lunin, V.Y. & Urzhumtsev, A. (1999). *CCP4 Newsletter on Protein Crystallography*, **37**, 14-28.

Skovoroda, T.P. & Lunin, V.Y. (2000). Crystallography Reports **45**, part. 2, 195-198.

Urzhumtsev, A.G., Skovoroda, T.P. & Lunin, V.Y (1996). *J.Appl.Cryst*., 29, 741-744.

Metal Coordination Groups in Proteins Some Comments on Geometry, Constitution and B-values

Marjorie Harding

Structural Biochemistry Group, Institute of Cell and Molecular Biology,

Michael Swann Building, University of Edinburgh, Edinburgh EH9 3JR

Metals are found in a wide variety of proteins where they may have important functional (usually catalytic) or structural roles. In a large proportion of structures the coordination group around the metal atom is made up of donor groups (carboxylate, imidazole, etc) from several amino-acid side chains, usually, but not necessarily, within one polypeptide chain; water molecules or other small molecules incorporated in the crystal as inhibitors, substrate analogues, cofactors etc. may also participate in the coordination group. In iron proteins, haem groups and clusters like Fe_4S_4 are also common.

The aim of recent work here (Harding, 1999, 2000, 2001) has been to provide information on metal coordination geometry which could be of use to protein crystallographers determining structures - at the stage of interpreting an electron density map, or in the restrained refinement of structures where the data is of limited resolution, or in the validation of structures. The work included a systematic extraction of geometric data from the Protein Data Bank (PDB, Bernstein et al., 1977; Berman et al., 2000) for metalloproteins of six selected metals. This note describes some further work on these metal coordination groups, which although not geometrical, might also be of relevance in protein structure determination - a) a systematic description and listing of metal coordination groups in terms of constituent amino-acid donor groups and their relative positions in the amino-acid sequence of the polypeptide chain, and b) a brief comparison of reported B values for the metal atom and the donor atoms in some of these groups. Much of this information has been assembled in a website on metal coordination groups. This includes links to some other websites relevant to metal coordination chemistry in proteins; among these is metalloscripps, which contains extensive geometrical information and tools for manipulation. The present descriptions and listings take no account of biological function or other properties of the metal sites; Degtyarenko (2000) describes an approach in terms of 'bioinorganic motifs' which does take function into account, and gives information on the available databases relevant to the field.

1. Geometry around the metal atom

An analysis was recently made of the geometry of metal-ligand interactions in metalloproteins (Harding, 2001) based on protein structures reported in the PDB. It dealt with Ca, Mg, Mn, Fe, Cu and Zn, by far the commonest metals in the PDB. The analysis started with accurate structural information derived by diffraction methods (resolution < 0.9Å) for appropriate small molecule complexes of these metals. The information was extracted from the Cambridge Structural Database (CSD, Allen and Kennard, 1993), and used to prepare a set of 'target' values for distances between each metal and each type of donor atom. The agreement between these target values and the values actually observed in metalloprotein structures in the PDB was then assessed, a) for all structures determined

with resolution < 1.6Å, and b) for a representative set of structures determined with resolution < 2.8Å (July 1999 release in both cases). With some very small adjustments the target distances were shown to be good and they can therefore be recommended for use in interpretation of electron density maps, in restrained refinement, or in validation of metalloprotein structures. Also available are fuller details of the metal coordination geometry, listed for each protein in a representative set (<2.8Å resolution, no two proteins with more than 30% sequence identity, Feb. 2001 release of PDB at present); an example of the information given is shown in Table 1. [Note too that the preferred geometry of metal in relation to donor groups such as carboxylate , imidazole, is described by Harding (1999), and that target distances and comments on the geometry of Na and K in protein structures are available on request from the author and will be published in due course.]

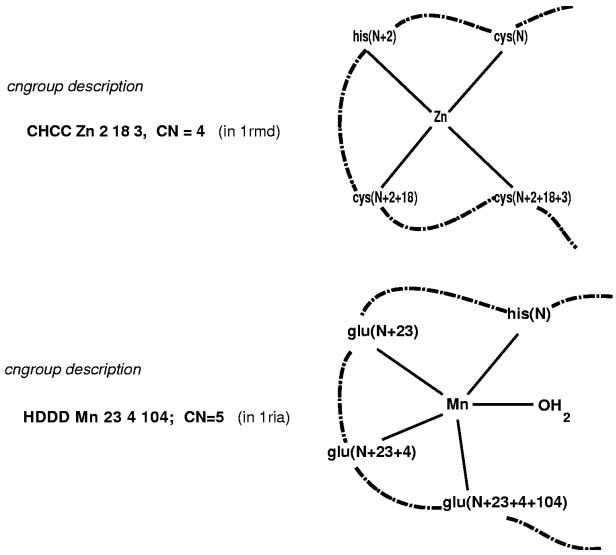
Table 1 Example of information provided on metal coordination geometry in one metalloprotein

1ctt	resolution = 2.20; total no. atoms = 2286; no. metal atoms = 1						
metal no.	coordn sphere	donor	dist (A)	dif from target	occ. product	B metal	B donor
1 ZN 296							
	1	ND1 HIS 102	2.02	0.02	1.0	22.4	26.3
	1	SG CYS 129	2.42	0.13	1.0	22.4	23.6
	1	SG CYS 132	2.11	-0.18	1.0	22.4	19.4
	1	O HOH 700	1.84	-0.25	1.0	22.4	13.7
	cngroup is HCC with CN 4 Zn, sequence diffs 27 3						
COORDINATION NUMBER: 4							
	nearest description of shape - tetrahedral (r.m.s. devn from tet 7.7, r.m.s. devn from sqp 41.1 degrees)						

2. Constitution of metal coordination groups and relation to sequence

It may be useful to describe and categorise all these metal coordination groups in terms of the constituent donor groups and their relative positions in the amino-acid sequence of the polypeptide chain, and some attempts have been made to do this. Software which had already been developed for the analysis of the geometry of metal ligand interactions in metalloprotein structures needed only small further extensions to allow the present study.

The description of a coordination group, or *cngroup*, used here includes the nature of the amino-acid donor groups, their sequence and separation (number of residues apart) in the protein chain(s), together with the metal coordination number (which includes non-protein donors, water molecules and other ligands). This information can be summarised as in the two examples below, where it is assumed that N is a residue number, cysteine coordinates through the thiolate sulphur, and histidine through imidazole nitrogen.



(Note that this takes no account of the nature of the other amino-acids in the protein chain, between those which contain donor groups, whereas sequence comparisons normally do.) Such *cngroup* sequences have been generated for Ca, Mg, Mn , Cu and Zn proteins.

Methods, procedures: The basis for generating the *cngroup* information is the program MP (Harding 2001, Acta Cryst D57,) which reads a PDB file, extracts coordinates and occupancy of each metal atom, and those of all atoms within 3.6 Å of the metal atom. Using target distances for each metal-donor atom combination it identifies atoms as donors if they lie within (target distance + tolerance) of the metal, and lists the amino-acids and residue numbers to which they belong, and the coordination number - as already illustrated in Table 1. It also evaluates sig, the r.m.s. deviation of the observed distances from the target values, and an alternative coordination number which would be found if an alternative, larger value of tolerance were used; these two, together with the resolution of the structure determination are useful in assessing the accuracy with which the *cngroup* has been identified. Metal coordination groups in which any atoms are disordered or have occupancy less than 0.7, are omitted.

Lists of PDB codes were obtained from the Jena Image Library search facility . (This gives a more complete listing than the PDB-3D Browser which searches in HET group names; some HET group names are odd, for example OC5 for Ca(OH2)52+, and in such cases no Ca atom is detected by the Browser.) From these lists protein and protein-nucleic acid complexes were selected, with structures determined by diffraction to a resolution $\leq$ 2.8 Å, and the program MP run, for all the structures available in the RCSB release of Feb2001. Additional smaller programs then gave the information on *cngroup* descriptions for the full lists or for selections from them. One such selection is a 'representative set' which excludes any structure which has more than 30% sequence identity with any other in the set; this used a culled PDB file 'cullpdb_pc30_res3.0_...'.

The list of *cngroup* descriptions is sorted in alphabetical order and gives sequence, metal, separation of coordinating residues in protein sequence, number of coordinating groups, pdbcode, 3 indicators of reliability, and the name of the metal and the first coordinated amino-acid in the chain. One letter amino-acid codes are used to specify the donor groups, and O indicates main chain carboxyl oxygen as donor. A carboxylate group (D or E) is always treated as one donor group, whether it is mono- or bidentate; at lower resolutions the distinction is not reliable, and particularly in Zn complexes, intermediate states are possible. The apaprent mono- or bi-dentate status is indicated at the end of the record.

The presence of water molecules or donor atoms from non-protein molecules is indicated and an alternative output option includes these within the *cngroup* before sorting into sequence order .

Concerning cngroup definition: An atom is identified here as a donor when its distance from the metal atom is within (target distance + tolerance). The target distances have been carefully established using appropriate small molecule compounds in the CSD and checking against high resolution protein structures (Harding 1999, 2000, 2001). Errors in determination of atom positions, especially in low resolution structures might result in incorrect decisions on whether or not an atom is within the metal coordination group. For this reason structures determined at resolutions less good that 2.8 Å are not included at all. The tolerance was set at 0.75 Å after examining the distribution of (observed - target) distances. When the resolution is <1.8 Å there should be no 'wrong decisions' about whether an atom is within the metal coordination group; when the resolution is poorer, but still < 2.8 Å, some 'wrong decisions' will inevitably be made, but their number should be well under 5% of the whole. The three indicators given for each *cngroup* are provided to show more about the reliability of these decisions: i) the resolution, ii) the number of additional donor atoms which would have been found if the tolerance had been 0.95A, and iii) the r.m.s. deviation of observed from target distance in the *cngroup*.

A few metal atoms in the *cngroup* listings have coordination numbers lower than would normally be expected (i.e. <5 for Ca, <4 for Mg, Mn, Fe, Zn, <3 for Cu) . Usually this is the result of failure to identify a donor group such as a water molecule, in the electron density map, but in a few cases it could be the result of a shortcoming in the software, which does not (yet) detect when the metal atom is coordinated to a donor group in a neighbouring asymmetric unit of the crystal.

Lists generated and their possible uses: The lists now include all proteins containing any of the metals Ca, Mg, Mn, Cu, Zn, whose structures have been determined at resolution $\leq$ 2.8 Å. They include many repeat entries where the same metalloprotein molecule occurs more than once in the crystal asymmetric unit, as well as occurrences of the same *cngroup* sequence in very closely related proteins such as mutants. They are sorted so that identical *cngroups* in different proteins appear next to each other. From these lists summary lists are also provided, which group together all protein structures which have the same *cngroup* sequence . Lists are also provided for a representative set of proteins (no two having sequence identity > 30%), and the summary lists for these - Table 2 is an example showing the form of the summary list.

Table 2 Example of parts of summary lists for representative Ca and Zn proteins.

DDDN	Ca	88	2	1	.	.	.	.	6	1	1alv
DDDOD	Ca	2	2	2	5	.	.	.	6	1	2scp
DDDOE	Ca	2	2	2	5	.	.	.	6	1	1cdl
DDDOE	Ca	2	2	2	5	.	.	.	6	4	1acc 1sra 1vrk 2pvb
DDEOOD	Ca	2	7	34	3	10	.	.	6	1	1acc
DDND	Ca	2	5	1	.	.	.	.	6	1	2por
DDNNOE	Ca	2	2	.	2	5	.	.	6	1	2cdl
DDNOOE	Ca	2	2	2	2	3	.	.	6	1	1cdl
DDNOE	Ca	2	2	2	5	.	.	.	5	1	1cdl
DDNOE	Ca	2	2	2	5	.	.	.	6	2	1rec 1vrk

CCCC	Zn	3	7	6	.	.	.	.	5	1	1zme
CCCC	Zn	3	7	7	.	.	.	.	5	1	1hwt
CCCC	Zn	3	14	3	.	.	.	.	4	1	1dsz
CCCC	Zn	3	17	3	.	.	.	.	4	3	1dcq 1rmd 1zin
CCCC	Zn	3	22	3	.	.	.	.	4	1	1zbd
HCC	Zn	27	3	.	.	.	.	.	4	1	1ctt
HCCC	Zn	11	1	10	.	.	.	.	4	1	1btk
HCCC	Zn	13	10	3	.	.	.	.	4	1	1gpc
HCCC	Zn	30	3	16	.	.	.	.	4	1	1ptq

These lists may have a variety of uses. It would be quick and easy to check whether in a newly determined metalloprotein structure the constitution of the metal coordination group is the same as that in a known structure, or structures. Here we plan to make geometrical comparisons after selecting proteins which have the same sequence of coordinating amino-acids in the *cngroup*, or closely related sequences, even though the overall sequences of the proteins are very different; polypeptide chain conformations in the chelating loops and nearby regions of the structure will then be compared, either by examining sequences of torsion angles, or by graphical superposition. Several sets of *cngroup* descriptions are easily recognised as familiar motifs, e.g. Ca proteins with EF hands, some Zn fingers, etc., which suggests that these listings may have some use in classifying metalloprotein structures. A long term aim of this project is to establish a representative set (in constitution and geometry) of metal coordination groups in proteins; this would have considerable overlap with the list of metal coordination groups in a representative set of proteins, but not be identical to it. This aim bears some relationship, but is also complementary to work nearing completion at University College London (M.W.MacArthur - private communication). The project there, which is in a much more advanced stage than this one, involves building a library of structural motifs (of metal coordination sites) including 3 dimensional aspects. These motifs, which can be used as templates or probes for a systematic classification of sites, include the positions of donor atoms of constituent amino-acids relative to the metal, regardless of the positions in which the amino-acids occur in the polypeptide sequence.

Some statistics of the numbers of structures and sequences found are given in Table 3 . (A few of the results or sequences are trivial, e.g. the numerous examples of $Mg(OH_2)_n^{2+}$ cations with no protein donor groups. There are also examples where groups are separately reported although the difference is small - e.g. a difference in coordination number due to different numbers of water molecules located near the metal.)

Table 3 Numbers of structures and cngroups

all metalloproteins		Ca	Mg	Mn	Cu	Zn
No. of structures		1081	546	207	183	782
No. of *cngroups* including repeats within asymmetric unit		2412	1113	571	434	1560
No. of *cngroups* occurring only once in different proteins		381	146	102	67	301
No. of *cngroups* occurring	2 times	69	41	24	16	54

in different proteins						
	3 times	49	23	10	4	18
	4 times	23	10	5	3	8
	5 times	87	30	11	10	65

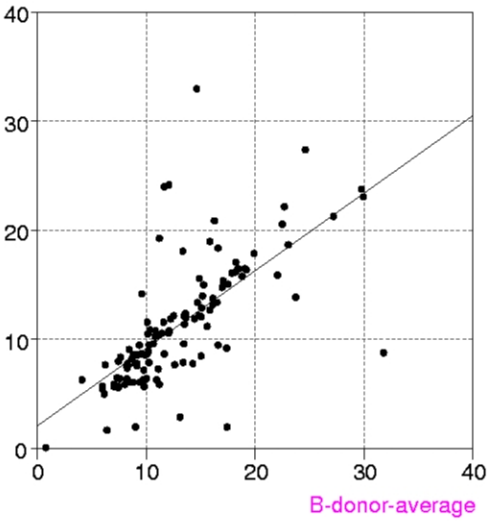
representative set of metalloproteins		Ca	Mg	Mn	Cu	Zn
No. of structures		121	127	21	22	116
No. of *cngroups* including repeats within asymmetric unit		289	271	53	52	26
No. of *cngroups* occurring only once in different proteins		184	78	29	31	183
No. of *cngroups* occurring in different proteins	2 times	10	7	0	0	8
	3 times	1	5	1	2	3
	4 times	2	4	0	1	1
	>5 times	0	5	0	0	1

3. B-values in metal coordination groups

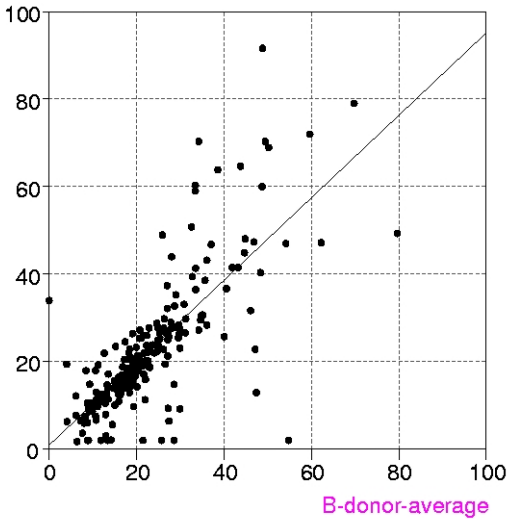
In addition to extracting geometrical information from PDB files the program MP extracted B values for metal atoms and donor atoms - see Table 1 for an example. Consideration of the value of B for the metal atom relative to the average B for the donor atoms might be helpful in identifying a metal atom, e.g an unexpectedly large B value for the metal relative to the surrounding donor atoms, could suggest that a metal of lower atomic number is actually present, or that the metal site is only partially occupied. Checking of the relative values might be useful in structure validation, but of course it would be essential to know what restraints had been applied to B values in refinement. A survey was made here of reported B values of donors relative to metal atoms, but without knowledge of the restraints applied.

Procedures : Metalloproteins in the PDB up to July 1999 containing any of the metals Ca, Mg, Mn, Fe, Mn, Cu, Zn, have been examined in two groups, a) all structures with resolution $\leq$ 1.6 Å and b) 'representative macromolecules' with resolution up to 2.8 Å. Coordination groups where the metal atom occupancy or any donor atom occupancy is less than 1.0 were excluded. The results were displayed in a spreadsheet (VISTA of the CSD system), and included, for each metal coordination group, Bmetal, the mean, minimum and maximum values of Bdonor, as well as the coordination number, resolution, etc. Bdmean is the average value of B for all the donor atoms in the first coordination sphere around the metal atom, in this case those within (target distance + 0.5 A).

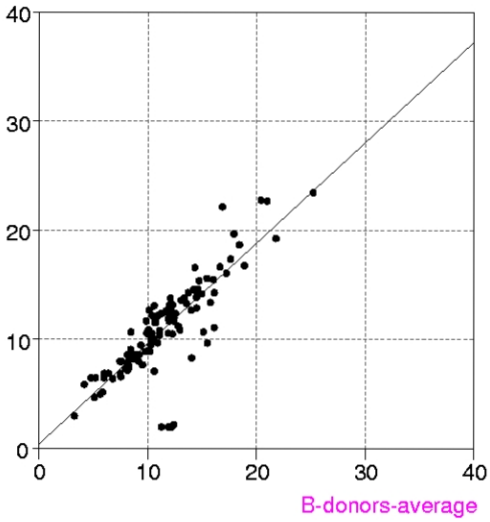
B- Ca



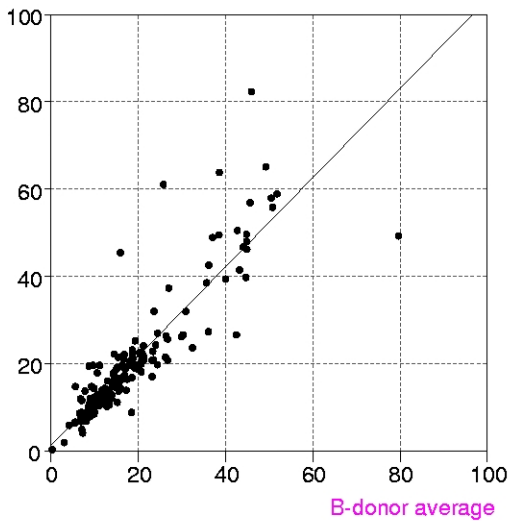
B- Ca



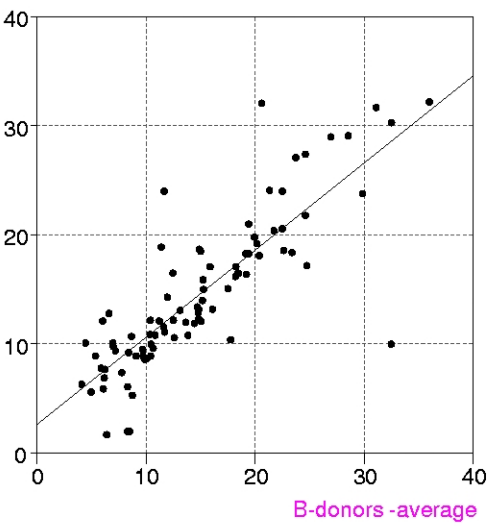
B- Fe



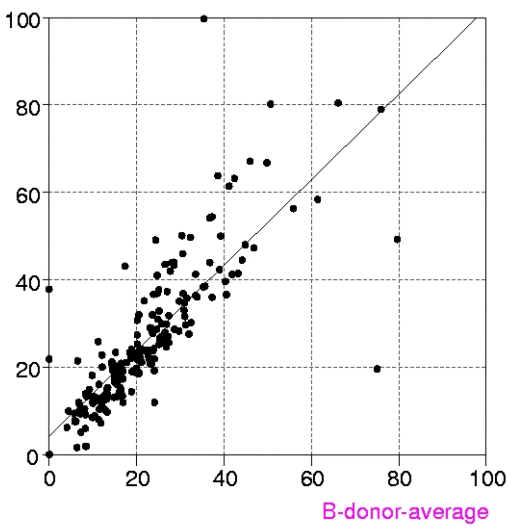
B- Fe



B- Zn



B- Zn



Results, Comments: The figures show the distributions of the reported values of Bmetal and Bdmean for Ca, Fe and Zn; on the left are results for all metalloproteins with resolution $\leq$ 1.6 Å, on the right for the representative set of proteins with resolution $\leq$ 2.8 Å. Mg, Mn and Cu show similar trends but there are fewer observations. It is clear that for most Bdmean is similar to or slightly greater than Bmetal , as expected, and as generally found in 'small molecule' complexes. In the structure determinations at poorer resolution the scatter is greater, and the B values can be much larger. No correlation with coordination number was found. There are some very marked outliers in the distributions. Examination of some of the outliers showed that the refinements were done by several different frequently used programs. Bmetal=2.00 is curiously common among the outliers - perhaps a minimum value set by one of the programs , but it seems physically unreasonable in most cases. Outliers are much less common with Fe than with Ca or Zn (or Mg, Mn or Cu). This might be because, in the Fe structures, there is rarely an ambiguity about which metal is present, or an occupancy other than 1.00, whereas such uncertainties are more possible in Ca and Zn structures; alternatively it might be because Fe is most commonly bound within fairly rigid groups like haem or Fe_4S_4, while Ca or Zn atoms are sometimes more flexibly bound. These are very speculative suggestions. Further investigation of the outliers in these distributions has not been made here, but is desirable, and should start with a check on what restraints have been applied in the refinements. In all new structure determinations it is recommended that examination of Bmetal /Bdmean should be a useful step in validation.

4. Summary

The website metal coordination groups in proteins provides target values for metal-donor atom distances in metalloproteins for the six metals most commonly found in metalloproteins, together with fuller geometrical information on the geometry of coordination groups in a representative set of proteins. It also provides information on constitution of metal coordination groups in terms of coordinating amino-acids and their relative positions in the polypeptide chain sequence - for all metalloproteins containing Ca, Mg, Mn, Cu or Zn in the PDB (to 2.8 Å resolution, February 2001 release at present). The B values for metal and donor atoms, extracted from PDB files at the same time as the geometrical information, mostly follow expected patterns; these B values should be examined in structure validation, taking information on restraints into account. In future work the inclusion of some other metals will be considered, as well as extended geometrical comparisons of whole coordination groups of similar constitution.

I am very grateful to Professor Malcolm Walkinshaw, University of Edinburgh, for access to computing facilities, to Dr. Paul Taylor for computational support, and to both for helpful discussions.

References

F.H.Allen & O.Kennard, 1993, Chem.Des.Autom.News, 8,1 & 31-37.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. & Bourne,P.E. (2000) Nucleic Acids Research, 28, 235-242.

Bernstein, F.C., Koetzle, T.F., Williams,G.J., Meyer,E.E., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. & Tasumi,M. (1977) J. Mol. Biol. 112, 535.

Degtyarenko,K. (2000) Bioinformatics Review 16, 851-864.

Harding, M.M. (1999) Acta Cryst. D55, 1432-1443.

Harding, M.M. (2000) Acta Cryst. D56, 857-867.

Harding, M.M. (2001) Acta Cryst. D57, 401-411.

X-RAY ABSORPTION IN 2D PROTEIN CRYSTALS

José R. Brandão Neto

Laboratório Nacional de Luz Síncrotron – CBME - CPR
Caixa Postal 6192 - Campinas - SP - CEP 13084-971
brandao@webcom.com

Abstract. X-ray diffraction is a technique that allows access to 3D structure of crystalline materials. In the cases in which it is possible to obtain protein crystals, diffraction studies are performed in order to obtain their 3D structures. As data collection techniques evolve, it is necessary to include more details to correctly model the experiment. At the Protein Crystallography Beamline at LNLS, the Brazilian National Synchrotron Source, X-ray experiments are performed at beam energies which enhance the absorption by the crystal, and decrease data quality introducing systematic errors in the recorded intensities that emerge from crystals. It is highly desired to establish a protocol that enables absorption correction for diffraction experiments. Absorption correction can be calculated by direct integration of a scalar function throughout the crystal volume, and the definition of the scalar function remains as the key issue in this kind of problem. This work presents the results for an automatic procedure to obtain the integrand function, as well as the main qualitative aspects of absorption effects, simulated in 2D protein crystals that present physical properties similar to the real crystals used in experiments.

1. INTRODUCTION

Proteins are the basic units of living organisms and can be extracted from natural materials (e.g. seed grinding), or more recently by expressing them in microorganisms. It can take months to years to establish the process, which results in protein solutions and is the key to obtain protein crystals.

The protein crystallography beamline at LNLS is running since 1998 (Polikarpov, Oliva et al, 1998), allowing researchers to collect X-ray diffraction data from protein crystals, in the range between 8keV and 12keV. Protein crystals (Fig. 1) are grown in crystallization experiments performed with concentrated protein solutions.

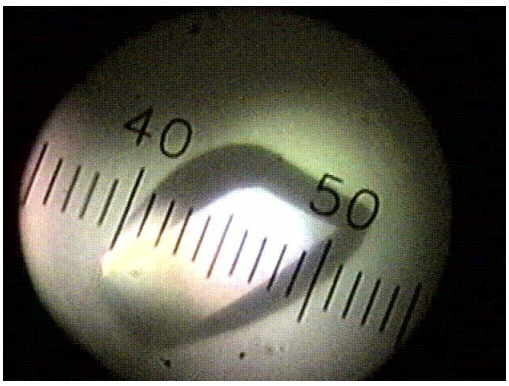


Figure 1 – Hen egg white lysozyme crystal (0.03mm/div).

Protein crystals are difficult to grow and sometimes are not found in big number, therefore it is necessary to collect the best data sets from the fewest experiments. X-ray diffraction experiments with protein crystals can be refined by including absorption correction schemes for the X ray absorption in the crystal.

Comparing with inorganic molecules crystals, organic macromolecular crystals present much less X-ray absorption, and this is the reason why at first X-ray absorption wasn't taken into account, though there always there were absorption correction studies (de Meulener and Tompa, 1965; de Titta, 1984; Maslen, 1995). The correction for this effect is important for protein crystallography when it is necessary to observe anomalous differences in diffraction patterns, in order to obtain phases for 'ab initio' structure determination.

3. X-RAY ABSORPTION

This section states the X ray absorption interactions, and places its importance inside the protein crystallography area. Firstly it is presented an overview, followed by the absorption aspects important in diffraction experiments and the description of the evaluation method.

3.1 Overview

The attenuation in the intensity of a monochromatic radiation beam that crosses an isotropic and homogeneous material follows the relation:

$$I = I_0 e^{-\mu T} \quad (1)$$

where I_0 is the initial intensity, I, the final intensity, m , the linear absorption coefficient, and T, is the distance traveled by the radiation while crossing the material. This expression holds for crystalline solids when subjected to X ray radiation in the range used in crystallographic experiments (Maslen, 1995).

From all the processes that can reduce the intensity of X-rays that traverse the materials, it must be highlighted the photoelectric effect, radiation scattering and extinction. Photoelectric absorption is the result from the interaction between the incident radiation and the atoms of the material, causing photons to be completely absorbed and electrons to be ejected from the atoms. Scattering happens in two ways: elastic scattering (Rayleigh) or inelastic scattering (Compton). Extinction can be modeled as the result from the elastic scattering added to the crystalline organization of the materials, which can increase or decrease the radiation intensity leaving the material.

Assuming that absorption is an additive phenomenon, it is possible to use the following relation for the linear absorption coefficient calculation (Maslen, 1995):

$$\mu = \frac{1}{V_c} \sum_{n=1}^{N_n} \sigma_n \quad (2)$$

where V_c is the volume of one crystallographic unit cell of the material, N_n is the number of n-type absorbing centers in this cell, and s_n is the cross section for the n-th contribution.

The transmission coefficient *TR* for a volume *V* of material that is traversed by X-rays is given by (Maslen, 1995):

$$TR = \frac{1}{V}\int e^{-\mu T} dV \quad (3)$$

where T is the path traversed by the radiation coming from outside the material and which leaves the material starting from a point inside the integrating volume. It is convenient to decompose T in two paths: the input path leading to the interior point and the output path from this point on. m is the linear absorption coefficient.

3.2 Protein Crystals and X-ray Absorption

Typical protein crystals are 0.1mm to 0.3mm big in their longest direction, reaching up to 1 mm long, and they can be seen as composed roughly by 60% of water (Blundell and Johnson, 1994). In Fig. 2 it is possible to see examples of crystalline morphology.

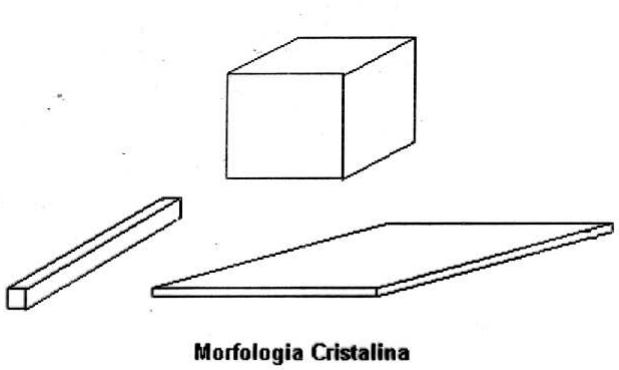


Figure 2 – Simplified crystalline morphology.

There are three basic types of protein crystals: rhomboedral crystals, plates and needles, and the longest linear dimension is taken as the reference to classify them. For this work, plates and needles are the most interesting for it is the difference in their dimensions that enhances the absorption effects, as it will be seen below.

Table 1 lists the values of m and for 1/m - the absorption length - for water, protein solutions and for air, when interacting with 1.4Å radiation.

Table 1. Absorption coefficients for materials at 1.4Å.

	m	1/m
Air	8.25×10^{-4} mm^{-1}	1211 mm
Water	0.75×10^{-1} mm^{-1}	1.33 mm
Protein Solution	1.0×10^{-1} mm^{-1}	1.0 mm

The absorption length tells one the path length that radiation should traverse inside the material until its intensity is reduced by 1/3 of its original intensity. X-rays for the protein crystals' diffraction experiments are highly absorbed as they traverse paths in the range of 1 mm inside protein solutions. Since typical protein crystals' sizes are about 0.1 mm, the absorption length indicates that attenuation shouldn't be neglected for these crystals.

If protein crystals' dimensions were similar in every direction, the effects of absorption differences would be small. Needle and plate crystals are very common, and this intrinsic characteristic may introduce differences in the diffracted intensities that are geometry-dependent, and it is interesting to calculate how much does the absorption effect affect the outgoing radiation.

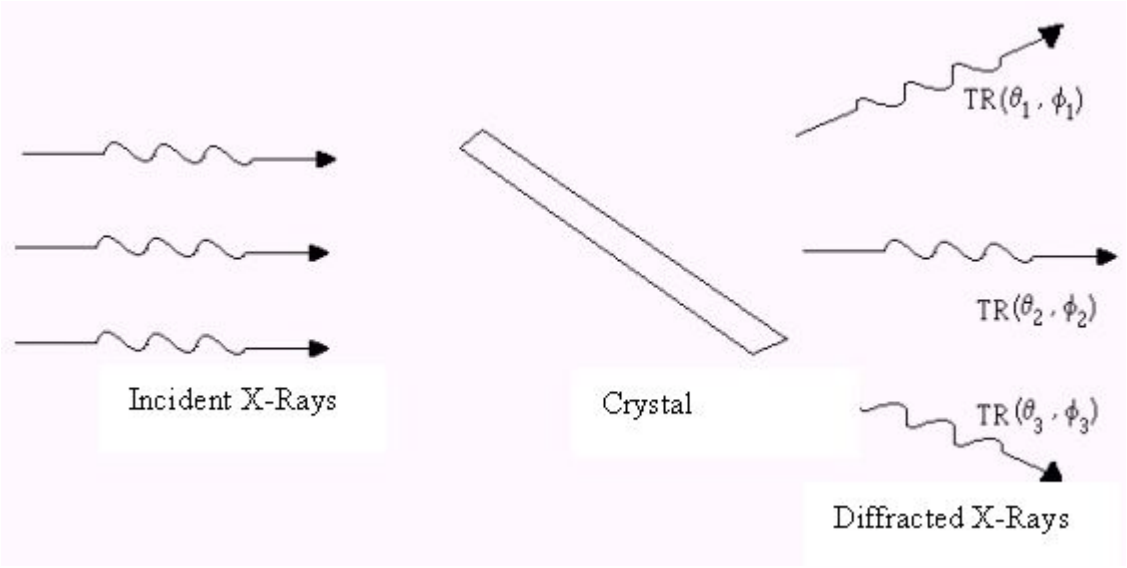


Figure 3 –Attenuation is geometry-dependent.

Figure 3 shows qualitatively that the diffracted radiation leaves the crystal with attenuation that depends on the direction (q ,f). The crystal transmission coefficient, TR, is a function that depends on these variables. The angular dependence can be analyzed in terms of the crystal geometry and its orientation in relation to the incident radiation, and is calculated by the integral given in Eq. 3.

In the case of anomalous diffraction, the anomalous difference can be estimated for Bijvoet pairs (Hendrickson & Ogata, 1997) according to the following relation:

$$D_A = \frac{\Delta F_{\pm}}{|F|} \approx \left(\frac{N_A}{2N_T}\right)^{\frac{1}{2}} \frac{2f_A''}{Z_{eff}} \quad (4)$$

where F is the measured structure factor, N_T is the total number of atoms (excluding hydrogen atoms), N_A, is the number of anomalous scatterers, f_A, is the anomalous scattering factor for the anomalous scatterers and Z_{eff}, is the effective scattering factor for all the light atoms.

In proteins Z_{eff}=6.7 e^- and N_T=7.8N_{res}, where N_{res} is the number of residues in the protein. For the HEWL protein (**H**en **E**gg **W**hite **L**ysozyme), the anomalous difference from sulfur atoms that are present in its structure is 0,8% under 1.4Å radiation, and if cesium atoms were placed in the structure, the anomalous difference would be 6%.

Assuming that the relation between structure factors F and measured intensities I is

$$\frac{\Delta F_{\pm}}{|F|} = 2\frac{\Delta I_{\pm}}{|I|} \quad (5)$$

it is possible to conclude that the differences in the measured intensities would be 0,4% e 3% for sulfur and cesium respectively.

Taking into account that good anomalous differences experimental signals might be essential to obtain phases in ‘ab initio’ structure determination, these anomalous differences values will be used as standards in evaluating the necessity for absorption correction.

3.3 Absorption Coefficient Evaluation

For the sake of simplicity, this work will present the results for 2D crystals, without loss of generality, and the main features of absorption differences will be highlighted.

According to Eq. 3, the transmission coefficient is calculated by an integral, which can be approximated by (Abramowitz & Stegun, 1972, Carnaham, 1969):

$$TR \approx \frac{1}{V} \sum_{i,j,k} w_i w_j w_k e^{-\mu T_{ijk}} \quad (6)$$

where *TR* is the transmission factor, *V* is the volume of the crystal, w_i are gaussian weights, m is the linear absorption coefficient, and T_{ijk} is the path traversed by the radiation when passing through the ijk-th point of the integration grid.

The absorption coefficient is calculated by:

$$A = 1 - TR \quad (7)$$

For the 2D case, Eq. 6 can be written as:

$$TR \approx \frac{1}{S} \sum_{i,j} w_i w_j e^{-\mu T_{ij}} \quad (8)$$

where *S* is the area of the crystal, instead of the volume, and m and T_{ij} are analogous to the ones in Eq. 6

4. SIMULATION AND RESULTS

Once the problem is stated, it is possible to simulate some representative cases, using typical constraints found in protein crystallography. In Fig. 4 it is possible to see the input radiation direction, the crystal and the output radiation direction.

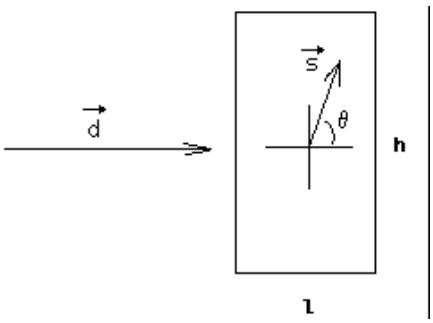


Figure 4- Geometrical description of the simulation experiment.

The simulations were performed in order to highlight the absorption differences for different crystal geometries. In Fig. 4 it is shown a crystal with length **l** and height **h**, subject to radiation coming from the direction of $\vec{a}$ and leaving the crystal in the direction of $\vec{s}$. The crystal is fully immersed in X-rays. Since this is a 2D problem, the output direction can be regarded as a function of the output angle, θ. This situation is equivalent to a rectangular cross-section cylinder.

In the simulations, the input radiation was kept fixed, and it was made a scan for the output directions, resulting in a plot of absorption versus output angle for each of the tested crystal dimensions. To validate the procedure adopted for the rectangular 2D crystals, the algorithm was tested for a crystal with dimensions comparable to the dimensions of tabulated absorption factors for cylinders (Maslen 1995).

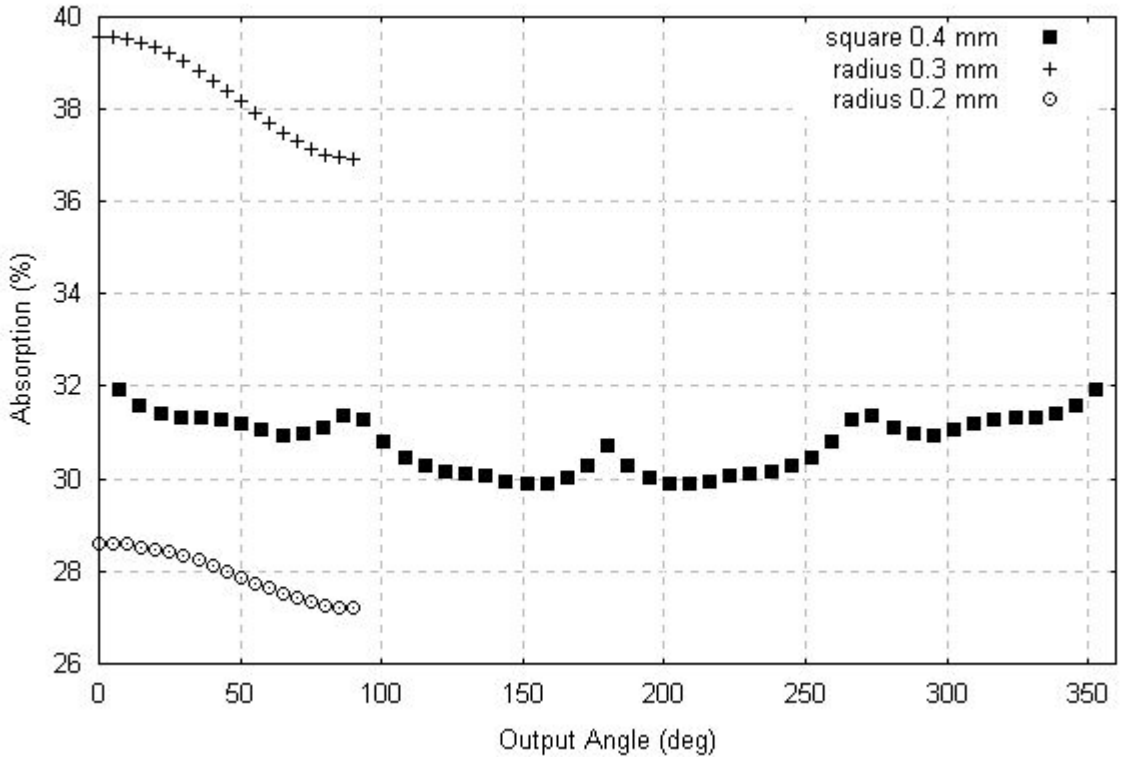


Figure 5 – Comparison between square and cylinders.

Figure 5 shows the comparison between a 0.4 mm x 0.4 mm square and a 0.2 mm radius cylinder and a 0.3 mm radius cylinder, as presented by Maslen (1995). It is very clear that the simulated square sample absorbs more than the smaller cross-section cylinder and less than the bigger cross-section cylinder, which agrees with qualitative expectations.

In order to analyze the behavior of the absorption in crystals, it was chosen to keep the height fixed at 0.3 mm and vary the length from 0.3 mm down to 0.01 mm, which is a typical range of sizes for protein crystals. These results are summarized in Fig. 6.

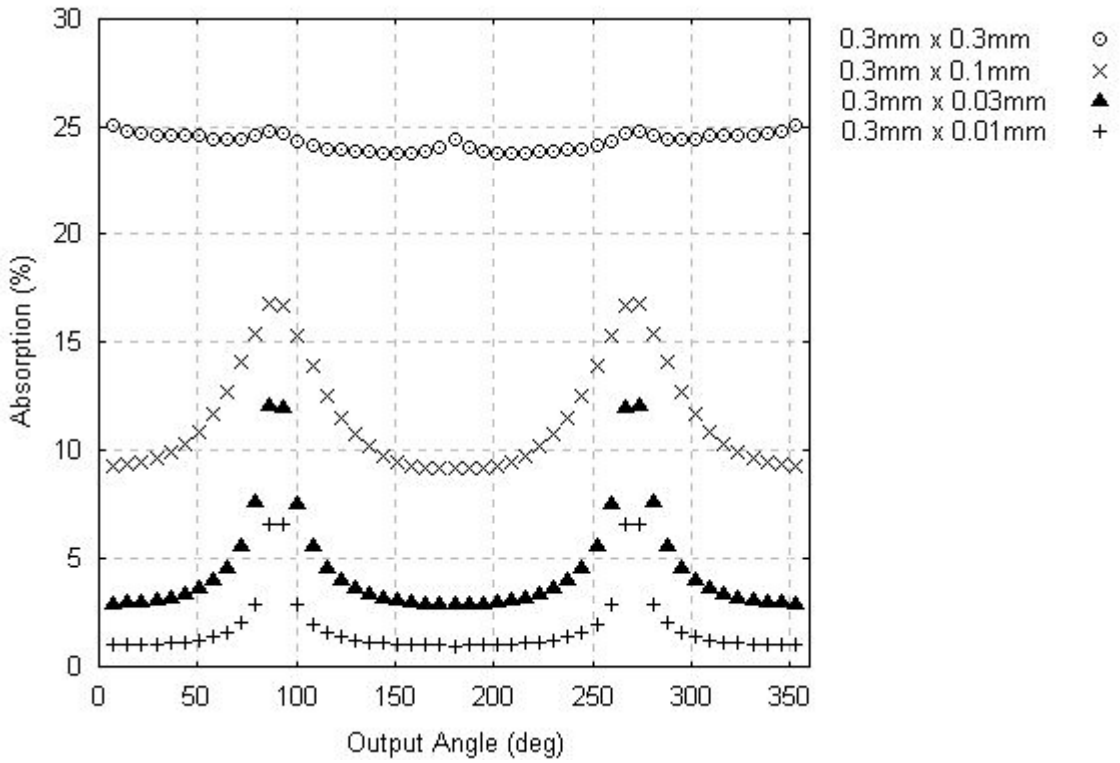


Figure 6 – Absorption for crystals with 0.3 mm height.

The maximum absorption differences seen range from 2% to 10% For smaller crystals, it were performed two simulations with 0.1 mm height. They test for absolute lower limits of absorption.

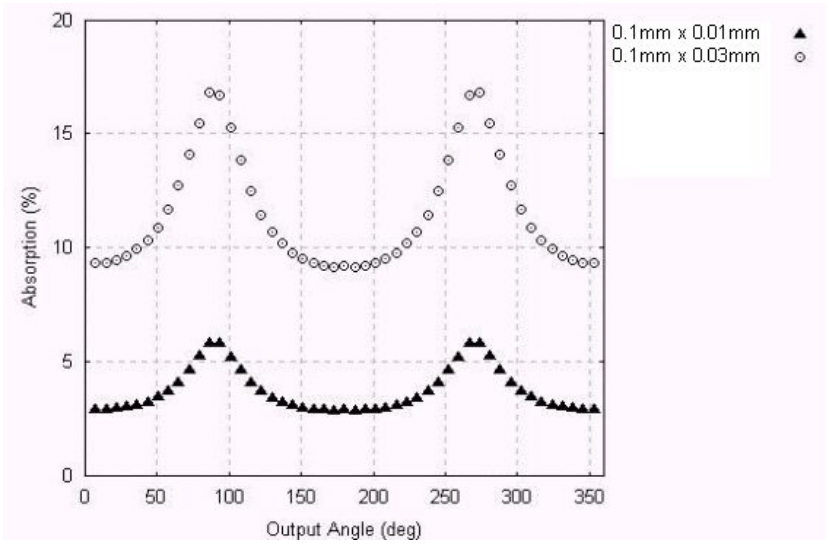


Figure 7 – Absorption for crystals with 0.1 mm height.

It is seen that the smaller crystals absorb much less X-rays - it should be noticed that the absolute minimum absorption achieved is around 3%. But the maximum absorption differences range from 3% to 8%.

5. COMMENTS AND CONCLUSIONS

Regarding crystals in which X-rays traverse different paths, there may arise big absorption differences for different output angles. In the cases studied, the largest difference is near 15%, which is far bigger than the anomalous differences estimated in section 3.2, which is between 1% and 4%. So it remains as an utterly important issue to reduce these disparities. Presently this problem is overcome by collecting redundant data from different output angles, so averaging lessens the absorption difference effects. But absorption introduces a systematic error, which can be eliminated if the geometric constraints of the experiments are known or estimated.

From the results obtained it is clear that care should be taken when choosing the crystal orientation during a data collection, in order to minimize the absorption difference effect. If this is done, the averaging method will be an effective manner to overcome this systematic error. For the cases that require a crystal scan in the maximum absorption difference region, it is strongly suggested that absorption correction is applied to the data, in order to eliminate the systematic error introduced by absorption.

The natural extension of this work is to prepare the simulation for the real 3D crystals, and then to apply the corrections to diffraction data from these crystals.

Acknowledgements

The author is deeply indebted to Dr Igor Polikarpov for fostering the investigations in absorption correction for protein crystallography.

Thanks to the free software initiatives worldwide, which made possible the utilization of Perl and GNUplot for computing and presenting the graphical results of this work.

To Adriana for her support, and for her patience.

REFERENCES

Abramowitz, M., Stegun, I.A., 1972, Handbook of Mathematical Functions, Dover Publications, Inc., New York, 9th printing.

Blundell, T., Johnson, L.N., 1994, Protein Crystallography, 4th printing, Academic Press Inc., London.

Carnaham, B., Luther, H.A., Wilkes, J.O., 1969, Applied Numerical Methods, John Wiley & Sons, New York.

de Meulener, J., Tompa, H., 1965, The Absorption Correction in Crystal Structure Analysis, Acta Cryst., vol. 19, p. 1014.

de Titta, G.T., 1985, Absorb: An Absorption Correction Program for Crystals enclosed in Capillaries with Trapped Mother Liquor, J. Appl. Cryst., vol. 18, p. 75.

Hendrickson, W.A., Ogata, C.M., 1997, Phase Determination from Multiwavelength Anomalous Diffraction, in Carter Jr., C.W., Sweet, R.M. (eds.), Methods in Enzymology, vol.276, part A, p.494, Academic Press.

Maslen, E.N., 1995, X-Ray Absorption, in Wilson, A.J.C. (ed.), International Tables for Crystallography, vol. C, p.520, Kluwer Academic Publishers.

Polikarpov, I., Oliva, G., Castellano, E. E., Garratt, R. C., Arruda, P., Leite, A. & Craievich, A., 1998, Nucl Instrum Methods A, 405: 159-164.

Recent CCP4BB Discussions

Maria Turkenburg (mgwt@ysbl.york.ac.uk)
December 2001

To make things much easier for both the users of the bulletin board and us writing this newsletter, *members who ask questions or instigate discussions on the board are now asked (urged!) to post a summary of all the reactions received*, whether on or off the board.

For each subject below, the original question is given in italics, followed by a summary of the responses sent to CCP4BB (together with some additional material). For the sake of clarity and brevity, I have paraphrased the responses, and all inaccuracies are therefore mine. To avoid misrepresenting people's opinions or causing embarrassment, I have tried not to identify anyone involved. Those that are interested in the full discussion can view the original messages on the CCP4 Bulletin Board Archive.

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

Thanks to all the users who are now dutifully posting summaries. Also I would like to thank Eleanor Dodson for her corrections and additions.

MOSFLM

MOSFLM features fairly heavily on the CCP4 Bulletin Board, both for crystallographically related queries and for problems related to installation on various computers. Most questions are answered very quickly by Harry Powell, who is also most happy to answer questions put to him directly.

MOSFLM, XDS, DENZO - conversion of crystal missetting angles

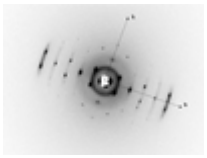
Also: how to deal with low resolution diffraction and partially recorded reflections

(February 2001)

I have a ~6Å dataset which I can index in XDS but so far not in MOSFLM or DENZO. Is there a simple way to convert the crystal missetting angles as given in XDS to the conventions used in MOSFLM or DENZO? I would like to try integrating the data in MOSFLM and in DENZO as well. The images are weak, one of the axes is almost perfectly aligned with the spindle axis and, as written, the resolution poor. XDS probably succeeds because it can use more frames than MOSFLM (6.1) or DENZO.

XDS2MOS

> Richard Kahn (Grenoble) has written such a program (XDS2MOS). It produces a MOSFLM indexing matrix from GLOREF.LP. It needs a modification to use IDXREF.LP as input in case GLOREF does not work, which is the case with very low resolution data.



Experience is based on a dataset of crystals from low density lipoproteins (LDL). Resolution between 28 and 15Å, unit cell: 200, 400, 400Å, C2. This is certainly an extreme case, but it shows the limits clearly. Image from Lunin *et al.*, Acta Cryst. D57, January 2001, 108-121 (click on thumb-nail to enlarge).

XDS can use your whole dataset for indexing if you like. The point is, that XDS constructs 3D profiles already in the indexing steps and everything is done in batch which is a great advantage if you need a lot of images for indexing.

MOSFLM

While there is a limit on the number of images you can use in MOSFLM for indexing, this number is large enough (it's 20 for all indexing options, as from version 6.11) to provide a sufficient sampling of reciprocal space to successfully index most datasets. The images don't have to be adjacent to each other.

Weak images, too, often don't seem to bother it. It is certainly possible to index images using just 10 spots, all worse than 4Å.

The only thing that does seem to matter is the beam position, but that seems to apply to all programs. The old MOSFLM indexing algorithm (the one you get when you say "no" when asked whether to use DPS), seems a bit more robust about the beam centre, but then you need strong images and certainly include spots from several images widely separated in phi. Harry adds: "You need to know the beam position to within half the minimum spot separation for any autoindexing to work or the indexing will be incorrect even if it *seems* to work".

Also be very aware that the x and y convention is switched between some programs. Where MOSFLM uses (x,y), DENZO and D*TREK use (y,x). Harry adds: "There's still a jiffy program around which will do the conversion from DENZO to MOSFLM indexing (from the days before the "new-style" indexing); see ftp://ftp.mrc-lmb.cam.ac.uk/pub/pre/denzo2mosflm.f".

MOSFLM has (had?) a problem with the integration of reflections extending over many images. Work on this is in progress, and the code for it is robust enough to be used cautiously. Keep in mind that SCALA (see Appendix 1: Partially recorded reflections) has options to deal with these results effectively.

DENZO/HKL2000

Denzo uses only one image and HKL2000 can index on multiple images. Several people tried to index very low resolution data by HKL2000 and DENZO but it did not work.

Summary from the enquirer:
Good knowledge of the beam center was pointed out as important for a successful indexing. The (x,y) convensions are program specific.
I used the modified version of Richard Kahns program XDS2MOSFLM to get a MOSFLM orientation matrix. The CCP4 ROTGEN program could easily do the conversion between MOSFLM and DENZO. The data could be integrated in XDS and MOSFLM though DENZO (vers. linux_1.96.5) had problems fitting parameters probably because it only uses one image at a time. I briefly tried D*TREK (7.0) for indexing but giving it a fair chance to succeed remains.

MOSFLM - ignore overlap??

(November 2001)

Does anybody happen to know if there is a MOSFLM keyword to ignore overlap?

Try looking at the SEPARATION keywords. If you don't have the mosflm.hlp files handy, check out synopsis.cgi which does a simple-minded markup of the help file. I'd guess (without examining your images) something like

SEPARATION x y CLOSE
(where x and y are the spot separation in x and y) might help. However, as the help file says:

```
**** IT MUST BE REALISED THAT THIS WILL LEAD TO SOME DETERIORATION IN DATA
QUALITY. IT IS FAR BETTER TO USE A SMALLER ROTATION ANGLE OR BETTER
COLLIMATION TO REDUCE THE NUMBER OF OVERLAPS IF THIS IS POSSIBLE ****
```

Data processing

Data processing - indexing problems

(March 2001)

I have a dataset at 3Å resolution, synchrotron source, 1 degree frames, 180 degrees; frozen crystal, reasonable mosaicity. DENZO table looks like this:

```
---------------------------------------------------------------------

Lattice          Metric tensor    Best cell (symmetrized)
              distortion index    Best cell (without symmetry restrains)

primitive cubic         14.85%  88.62 151.10  89.21  90.00  86.54  89.81
                               109.64 109.64 109.64  90.00  90.00  90.00

I centred cubic         26.41% 121.89 174.91 175.47  43.38 110.35 110.11
                               157.42 157.42 157.42  90.00  90.00  90.00

F centred cubic         26.61% 198.76 193.90 194.36  77.78  53.58  54.03
                               195.68 195.68 195.68  90.00  90.00  90.00

primitive rhombohedral  14.09% 174.91 151.10 175.47  30.56  40.71  30.44
                               167.16 167.16 167.16  33.90  33.90  33.90
                               105.25 105.25 471.14  90.00  90.00 120.00

primitive hexagonal     12.17%  88.62  89.21 151.10  90.00  90.19  93.46
                                88.91  88.91 151.10  90.00  90.00 120.00

primitive tetragonal     1.44%  88.62  89.21 151.10  90.00  90.19  93.46
                                88.91  88.91 151.10  90.00  90.00  90.00

I centred tetragonal     9.71%  88.62  89.21 325.58  75.07  75.38  93.46
                                88.91  88.91 325.58  90.00  90.00  90.00

primitive orthorhombic   1.43%  88.62  89.21 151.10  90.00  90.19  93.46
                                88.62  89.21 151.10  90.00  90.00  90.00

C centred orthorhombic   0.18% 121.89 129.49 151.10  89.87  90.14  89.62
                               121.89 129.49 151.10  90.00  90.00  90.00
```

```
I centred orthorhombic   9.71%  88.62  89.21 325.58  75.07  75.38  93.46
                                 88.62  89.21 325.58  90.00  90.00  90.00

F centred orthorhombic   9.45% 121.89 129.49 325.58  89.73  68.15  89.62
                               121.89 129.49 325.58  90.00  90.00  90.00

primitive monoclinic     0.08%  88.62 151.10  89.21  90.00  93.46  89.81
                                 88.62 151.10  89.21  90.00  93.46  90.00

C centred monoclinic     0.17% 121.89 129.49 151.10  89.87  90.14  89.62
                               121.89 129.49 151.10  90.00  90.14  90.00

primitive triclinic      0.00%  88.62  89.21 151.10  90.00  89.81  86.54

autoindex unit cell             88.62  89.21 151.10  90.00  89.81  86.54
------------------------------------------------------------------------
```

Indexing in primitive tetragonal or orthorhombic fails. C centred orthorhombic, monoclinic, and triclinic all work nicely, and indexing and integration are apparently OK. The problems starts when scaling the data: all possibilities, except for triclinic, produce unreasonable results (using 'default parameters'). Ridiculously high chi squares (50 or more!) in the first round, Rmerges over 50%, huge rejection files (half the data!). In the following rounds, chi squares 'drop' to 2 or so, but the rejection files grow even bigger, and Rmerges are stuck. Well, it must be triclinic...but the refined cell is the following: 89.211 89.214 150.819 90.000 89.989 87.153 ----> a=b, alpha=beta=90. Furthermore, assuming a 50% water content, I would have 10 molecules in the cell; a bit unlikely, and a real molecular replacement nightmare. On the other hand, the low resolution diffraction limit could hint at loose packing and high water content, and things might not be that bad. Any ideas or suggestions? Where should one look for possible problems or mistakes (before scaling)? I must confess my unease with symmetry, maybe the DENZO table is showing me something that I cannot see.

Summary from the enquirer: Here is a summary of the tips I received last week regarding my data processing troubles (reminder: indexing and integration show center orthorombic, or monoclinic, but scaling goes awry except for P1 with two 90 degree angles). Despite being specific for my problem, some of them might turn useful for the inexperienced crystallographer.

- P1 90/90 can actually be true, it's the internal symmetry that defines crystal system and space group
- try a different program (D*TREK, XDS, MOSFLM)
- unique axes could have been mixed up, swap them around
- wrong beam position can lead to misindexing by one
- scale in P1, make MTZ and use HKLVIEW to look for symmetry; it will still be visible even if your indexing is 1 out whereas the merging statistics will be completely destroyed..
- calculate self-rotation function, and check for symmetry
- crystal may be twinned
- a few alternative lattices were also suggested; this may be checked through maXus Structure Analysis Software

Dodgy indexing, or dodgy mosaicity

(February 2001)

I'm trying to index a data set with DENZO. The problem is with mosaicity - it looks about 1 but at this value it misses out some low resolution spots and seems to overfit for the number of spots at high resolution. Thought it might be misindexed but did a direct beam shot and the beam values look about right. It manages to autoindex it fine and the chi values are also fine.

First the obvious: Try indexing with one of the other programs and compare the results.

- MOSFLM
- dTREK
- DPS
- Wolfgang Kabsch who has some information on the program XDS....

Then a similar experience: This reminds me of my (large) crystal when I measured a high resolution data set at the synchrotron when we did not have enough time for the low-resolution scan. Although the final mosaicity from SCALEPACK postrefinement was lower than the input in DENZO (in the low resolution range not all spots were detected) the mosaicity seemed to be higher. The rest (autoindexing, chi^2, ...) of the scaling was very smooth. I made a compromize, taking a mosaicity slightly higher than the postrefined from SCALEPACK for a second DENZO-SCALEPACK run. The data-set was fine anyway up to high resolution! I don't know exactly the reasons for this behaviour. Maybe the crystal cracked a bit during freezing, or a small part of twinning (although merohedral twinning is not possible in this space group)...

Then some practical advice: Try to look at the background profile of some of those spots. DENZO may refuse them if the background is too steep for it etc. CAUTION with this: playing with these parameters may spoil your data processing.

The mosaicity parameter in DENZO can be compared to the Lemon-Larson peak integration limits for (small molecule) diffractometer scans. The main volume of the peak is integrated and the 'tails' (in phi rotation for oscillation photos) are excluded. The mosaicity chosen by DENZO/SCALEPACK generally results in the best I/s for the reflection, with only a fraction (<1% ?) of the total integrated intensity going to the tails. This is not generally noticeable, except at low resolution where the 'tails' have sufficient intensity to be observable in the oscillation photo. Other reasons for observing unindexed reflections at low resolution include TDS (thermal diffuse scattering). Increasing the DENZO/SCALEPACK mosaicity parameter a 'little bit' is I believe a common practice, and should not severely affect the data quality (I/s). While using incorrect error models (in SCALEPACK) is probably a more harmful and common practice to avoid.

What you describe is perfectly normal behavior (for DENZO, anyway). I understand that it is difficult to model some reflections at low angle that are spread out over many frames. They are often ignored. Use as many frames as you want for integration of your data (HKL2000, or denzo_3d). This gives you a very good estimate for the mosaicity right during integration. If you are just using DENZO, integrate your data, then scale them and re-integrate using the mosaicity value that SCALEPACK gives you (add about 0.2). Keep in mind that the mosaicity can change depending on crystal orientation, radiation damage, etc., that's why it is best to refine the mosaicity during integration in the first place.

In any case, there will be reflections that won't be "predicted", for various reasons. The most common is that they don't belong to the main lattice (freezing artifacts, satellite crystals) or they come from a different crystal altogether (salt). Furthermore, mosaicity, or the sum of parameters that most integration programs call "mosaicity", seems to be resolution dependent. As far as I know, no program can model this in a satisfactory way. Don't worry about the few unpredicted reflections. It seems that your data processing is just fine.

In DENZO, mosaicity is defined as the smallest angle through which the crystal can rotate about an axis or combination of axes while a reflection is still observed. From this definition, we can extend our logic for single crystal oscillation photography... delta phi = (smallest reciprocal cell constant / d*min)(180/pi) - mosaicity mosaicity should be smaller than the first term so that delta phi remains a positive quantity. If crystal is highly mosaic, oscillation angle(delta phi) should have been very small while collecting the data. Or else at the time of processing one has to select a shell of reflections to start indexing from equation mentioned above.

Rsym and Rmerge, what are the differences?

(May 2001)

I was hoping I could get some clarification on the difference between Rsym and Rmerge. Does the Rsym represent the differences in the symmetry-related reflections on a single image?
or,
Does it represent the differences in the symmetry-related reflections on a single crystal? If it is the latter, what about low and high resolution data collections? Do you report an Rmerge because you are comparing 2 different data sets?

Summary from the enquirer: Rsym and Rmerge are often used interchangeably. But sometimes they are not. You need to check the documentation of the particular program that is giving you numbers or the definition in the paper you are reading or what the person you are talking to defines it/them as. Sometimes Rsym is within an image (*i.e.* MOSFLM Rsym) and sometimes from reflections within a crystal. Rmerge usually includes these definitions of Rsym plus any other sources of reflections. The general consensus seems to be that Rmerge is between datasets (only from different crystals?). It is still not clear to me when you collect 2 datasets (with differing parameters, for example high and low resolution) on the same crystal if you should report a Rmerge or Rsym. From the responses it seems the general standard is that you would still report it as an Rsym.

Two references might shed more light:

- M. S. Weiss and R. Hilgenfeld (1997) J.Appl.Cryst.30, 203-205. On the use of the merging R factor as a quality indicator for X-ray data
- M. S. Weiss (2001) J.Appl.Cryst.34,130-135. Global indicators of X-ray data quality

Water rings, ice rings

(October 2001)

What is the best way to deal with water rings? I seem to remember it was possible to exclude the relevant resolution ranges in SCALA, but I can't find the keyword anymore. Or

should I exclude the resolution ranges in Refmac? Or, perhaps, should I not exclude anything at all because the modern procedures (maximum likelihood etc) will take better care of it than I could anyway?

Summary from the enquirer: XDS (latest version) has a (nice) option of excluding resolution bins. This way you can always decide for yourself what to exclude and not have a "black box" tool do it for you. Reminder: it is MOSFLM in which you can and always could exclude resolution bins and SCALA could never do this (*e.g.* `RESOLUTION 15.0 1.5 EXCLUDE 3.79 3.63 EXCLUDE 2.29 2.22 EXCLUDE 1.92 1.90`). Guess my memory was wrong here. Opinions are divided as to whether to remove data from the ice-rings or whether it is better to keep the information. Some people claim that maps including the data from the ice-rings looked better than using data with ice-rings removed. Perhaps for refinement the data without ice rings is the best, because the refinement programs will not include missing reflections in the target. For map calculation, the dataset with water rings may be best, because a bad estimate for a reflection is better than setting it to zero. So I have integrated the data both ways (obviously the statistics without the rings are better) and will try refining and map calculation with both datasets and compare the results. By the way, the data is quite redundant, overall multiplicity 6.0, so really bad outliers should be taken care of.

Twinning, indexing, re-indexing

Indexing Relationship Table

(March 2001)

I'm currently looking for a table that lists all possible indexing relationships between two different data sets of the same crystal form if the true space group symmetry is lower than the lattice symmetry (i.e. true space group P3, lattice point group 3barm). I don't need this only for my special case (where I think I've got all possibilities), but I believe this should be of general interest to all crystallographers who have to get consistent data sets from the same crystal form (i.e. all searches by trying different soaking conditions). Of course, the first thing I did was to look into the International Tables A,B,C, but surprisingly, I didn't find such a table (or I have eggs on my eyes). Do you know about such a table and could tell me and the CCP4BB the reference?

Summary from the enquirer: I've received several pointers to tables with possible reindexing relationships. Many of them were lying directly in front of me! Here are the pointers:

- $CHTML/reindexing.html
- XDS indexing routine lists reindexing possibilities
- the HKL manual deals with them in its scalepack scenarios
- It's in the special Acta D issue on data collection and processing, Dauter (1999), Acta Cryst. D55, 1703-1717

I222 to $P2_12_12_1$

(April 2001)

I have a question not directly related to CCP4 but may be interesting to most crystallographers. We have a protein crystallized in I222 space group. The structure was

solved by MIR with one molecule per asymmetric unit. Recently we crystallized the same protein in a very similar condition, but the space group is P212121. The unit cell dimensions of the P212121 cell are almost identical to those of I222. So the only difference is that the reflections with h+k+l=2n+1 are now present!
We thought this is an easy problem that we just need to solve the structure by molecular replacement methods. But we did not find obvious solutions. The chance that a protein packs differently but resulting in exactly the same size of unit cell should be rare! So is it possible that there are two crystals and one is mis-indexed by one, so the combination of the two I222 gives a diffraction pattern of P222? Has any one dealed with this type of problem before, changing of space group but not unit cell dimensions? What is the explanation?
By the way, the Rsym is quite low (around 5%).

Summary from the enquirer: The first suggestion was to check if h+k+l=2n+1 reflections in P212121 cell are weak and do the native Patterson to see if there is a peak closed to (0.5, 0.5, 0.5). This is to find out if we have a pseudo-I-centered cell. In our case h+k+l=2n+1 reflections are not weak; they have an average F of about 10% smaller than that of 2n reflections. However, we indeed see a strong native Patterson peak at (0.5, 0.5, 0.5) with an ellipsoid shape but not a perfect sphere like what we observed for the I222 case. So it is likely that our P222 cell has a pseudo-I symmetry but (x, y, z) is not translated exactly to (x+0.5, y+0.5, z+0.5). See the paper describing 3cel: Stahlberg J, Divne C, Koivula A, Piens K, Claeyssens M, Teeri TT, Jones TA, Activity studies and crystal structures of catalytically deficient mutants of cellobiohydrolase I from Trichoderma reesei. J Mol Biol 1996 Nov 29;264(2):337-49. A similar example of changing from a I222 (1cel) to a P21212 (3cel) with the same unit cell dimension was shown. In this case, the single molecule in I222 is located at (x, y, z) and the two molecules in P21212 are located at (x, y, z-0.25) and (x-0.5*, y-0.5*, z-0.75). 0.5* indicates a value closed to 0.5. The translation between the two molecules in P21212 cell is (0.46, 0.5, 0.5), so it is transformed from I to P. In our case, although h+k+l=2n+1 are strong, all the 2n+1 reflections in the axes are absent. So we thought the new cell is a P212121. But it turns out that the correct space group is P21212. It is a little bit more complicated for us, since the 2-fold is along b-axis, so we have to move it to the c-axis. The result is similar to the cel case that the two molecules in P21212 is separated by (0.49, 0.52, 0.50) and the origin of I222 cell is moved to (0, 0, 0.25) in the new but the similar packing P21212 cell.

Indexing in I222

(May 2001)

Denzo proposed the highest symmetry lattice as I centered orthorhombic with the skew parameter 0.18%. This gives space groups I222 or I212121. The predictions in I222 never meet the spots. The unit cell parameters match almost exactly those of a similar structure of the same protein in the same space group. So it looks likely. But the misfit is BIIIIIG (30 degrees in orientation of reciprocal space rows or more, different pattern, even the spacing a bit bigger). Anyone has any ideas what to do? (Except indexing in P1, which is possible, and searching for symmetry)

Summary from the enquirer: It was a trivial error. I have never tried to force indexing under the proposed space group, *i.e.* without cryst. rot and unit cell. Once the autoindexing is done with I222 forced it works of course.

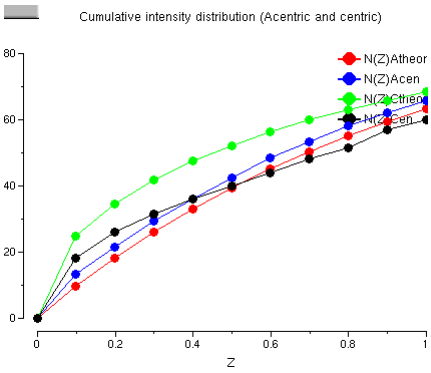
Tetragonal Twinning & Detwin

(September 2001)

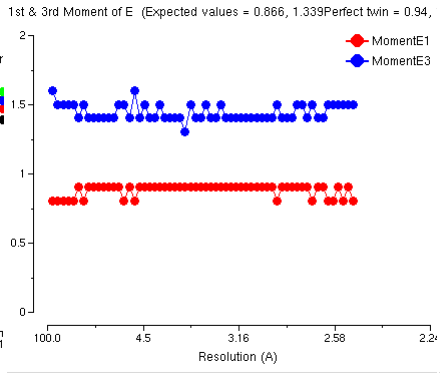
I have a dataset that I scales equally well in P4 and P422. In order to resolve this ambiguity I looked at the P4 scaled data in HKLVIEW and found mirror planes in all the right places suggesting that the Laue class was 4/m mm, therefore P422. All the moments and intensity statistics in SCALA/TRUNCATE look fine when the data is scaled in P422, but not quite as good in P4. Just for the hell of it, I ran the DETWIN program on the P4 scaled data, and DETWIN reckons my data is pretty much a perfect twin. So... if the true space group is P422, and you put P4 data through DETWIN, will it appear twinned (as 'perfectly' twinned P4 data can appear to be P422...)? The UCLA twinning server indicates that my data is not perfectly twinned when tested in P422... so now I'm getting two conflicting results and I'm confused... (basically... is my data twinned or not!!!)

The enquirer kindly provided plots from TRUNCATE (click on thumb-nails to enlarge):

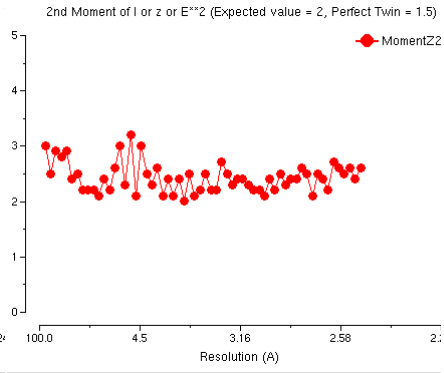
P4



cumulative intensity distribution

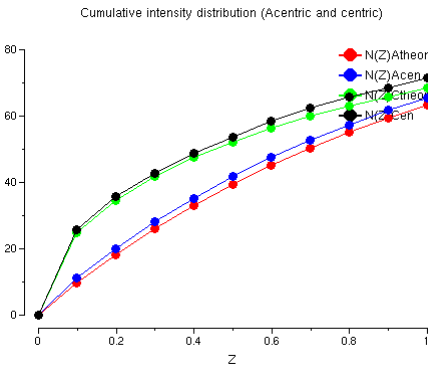


1st&3rd moments

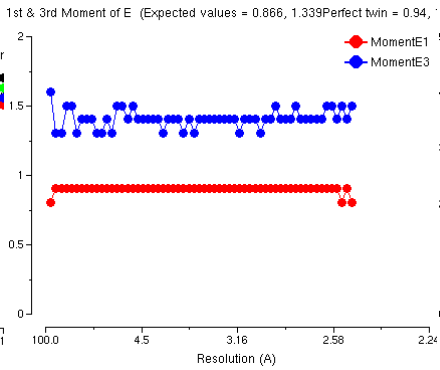


2nd moment

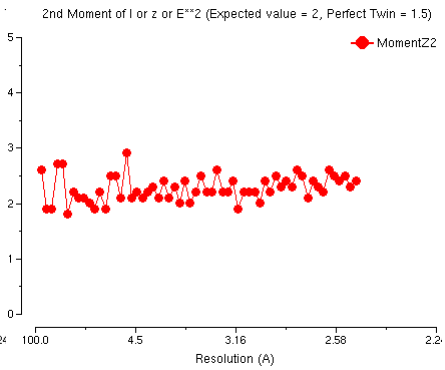
P422



cumulative intensity distribution



1st&3rd moments



2nd moment

Also, have a look at his http://student.cryst.bbk.ac.uk/~ebrig02/twin/ on this.

Perhaps packing considerations can help you out with your twinning problem: In P4 there are 4 a.u. per unit cell, in P422 it would be 8 a.u./unit cell. If the true space group was P4 and you have a perfect twin, and assume you have one protein molecule per asym. unit, then when you calculate Matthews parameters for both P4 and P422, they would look alright for P4 and one molecule, but for P422 you would obtain a reasonable Matthews parameter only for 0.5 molecules per a.u. The other way around: the wrong assumption of P422 caused by perfect twinning means that the lattice is too small to accommodate the number of molecules required by this space group. Think this was what made Luecke et al. suspicious about the possibility of twinning in the case of bacteriorhodopsin (Luecke, H., Richter, H.T., and Lanyi, J.K. (1998). Proton transfer pathways in bacteriorhodopsin at 2.3 angstrom resolution. Science 280, 1934-1937.) It would become a bit more difficult when the true space group is P4, and you have 2 molecules in the asym. unit, connected by two-fold NCS. Then you obtain normal Matthews parameter for the true space group and 2 mol. per asym. unit, but also for the wrong sp.gr. P422 with 1 mol. per a.u. However, if you are lucky and the NCS axes do not run parallel to the crystallographic axes, you should then be able to differentiate between NCS and pseudo-crystallographic two-fold axes (caused by perfect twinning) by examination of the self rotation function. The self rotation peaks of data processed in P4 should be at kappa 180, omega 90, and phi _exactly_ at 0°,45°,90° etc. only in the case of perfect twinning. If they are off 0°, then it is NCS and thus not perfect twinning.

Note from the enquirer: Unfortunately, I am not that lucky. I have 2mols per asu in P422 (therefore 4 in P4) - everything SHOULD fit. My NCS two fold does run parallel to my crystallographic axes, as I have rather nice looking pseudo-translation peaks on my native patterson...
The DETWIN program indicated a near perfect twin for the P4 scaled data. As there are no twinning operators for P422, I could not use DETWIN on this data.
The UCLA twinning server allows you to detect presence of a perfect twin using your higher space group (for me, P422). The perfect twin test gave a resounding "NO, you are not twinned!". However, the partial twin test using P4 data gave a "yes, you are greater than 45% twinned" answer. Which is right?!

A piece of wisdom: one should always go for the highest symmetry that gives consistent results. If the true symmetry is P4, you might be looking at twice as many molecules in the asymmetric unit, with an 'accidental' packing that looks like P422. To distinguish between them, you might want to do rigid body refinement of the P422 derived model in P4 (using the appropriate 422 symmetry operator to complete the contents of the P4 asymmetric unit), and then observe how far apart the two are. If there are genuine differences, go for the lower sp. gr. However, Rigid Body refinement only tells you about gross errors in positioning the molecules. This might not be significant. So you might have to go further and do a full refinement in both sp. gr. and observe particularly the side chains near interfaces that make lattice contacts. A few of these differences would force a lower symmetry (P4), but if you assume the higher symmetry (P422) you would not notice in the statistics, always taking into account the degree of difference (the resolution obviously has a great deal of impact on the significance of the differences). 'Accidental' packing that looks like a higher sp.gr. usually gives a slightly odd N(z) plot in TRUNCATE, where the observed graphs are to the left of the theoretical ones. If they are to the right of the theoretical graphs, especially in the bottom left corner, then you should suspect twinning.

The solution (?): Following on from my problems regarding tetragonal twinning and some ambiguity between P4 (twinned) and P422 (non-twinned), we took an un-scaled MTZ file from a solved/published structure from our college that was solved in P422 (4/m mm). This

integrated MTZ file was in P4. We then re-indexed this in P422 and repeated SCALA/TRUNCATE/DETWIN on both P4 and P422 datasets. Both my data and the solved data scale equally well in P4 and P422 (sensible stats, very few rejections...) The P4 centric intensity distribution was also a little odd, whereas the P422 looks fine. All the various moments in P4 and P422 indicated that the data was not twinned. Detwin also indicated that, in P4, this data was an almost perfect twin. The UCLA perfect twinning test for P422 indicated "no twin", but the partial test in P4 indicated almost perfect twin.
As this structure has been solved to about 2.8A, it is fairly safe to assume that it is not twinned...
When data scales equally well in both higher and lower space groups, provided that there are NO indications of twinning in both intensity distribution AND moments, then is it safe to assume that it isn't twinned, and it IS the higher symmetry, despite the fact that Twinning tests indicate that the lower symmetry is almost perfectly twinned? (making us believe that the higher symmetry is an artifact of the merohedral twinning).... (!) Therefore, for near perfect twinning, should one pay more attention to the UCLA "perfect Twinning Test" than other tests designed for partial twinning?

More thoughts: The 2-fold NCS parallel to your 4- or 2-fold crystallographic NCS can cause systematic weakening of some sets of reflections while strengthening others (depending on whether the pseudo-translationally related molecules scatter in phase or the other way around). This would result in more weak and more strong reflections with fewer "average intensity" reflections. Just the opposite of twinning where you see fewer weak or strong reflections. Your cumulative intensity distribution plot (the first one) shows such a pattern for the centrics (black line) which rise quickly (many weak reflections), flattens off, and then (with a bit of fantasy) rises again at the end. However, in all cases the line remains below the theoretical line (green) which doesn't make sense. You also don't see an effect in the acentrics or the P422 curve. Perhaps it is just bad statistics since you won't have that many centric reflections in P4 (only the HK0 plane). Correct me if I'm wrong, but I thought P422 couldn't form merohedral twins as the unit cell morphology has the same P422 symmetry as its content (unless your c axis happens to be the same as a and b). For P4 you can have twinning. Is it possible that the UCLA server with the "higher space group option" is comparing twin-related reflections in this situation rather than intensity distributions? If so then of course your P4 processed data suggests 50% twinning. Based on your TRUNCATE data I would suggest to go ahead and assume that things are ok unless you run into a brick wall somewhere. Your parallel NCS and crystallographic symmetry may turn out to be a greater problem than the perceived twinning.

Twinning problems (again....)

The problem is that we appear to be getting twinned crystals, but that neither TRUNCATE nor the twinning server shows this up. We have tetragonal crystals, apparent space group P41212 or P43212. The crystals show 100% incorporation of Se by mass-spec and the fluorescence scan shows a Se edge. We collected Se-SAD data sets at the peak wavelength for five crystals, all diffracting to 2.8 - 3.0Å. The data was processed with MOSFLM. Parts of the SCALA and TRUNCATE logfiles for one are reproduced below. As you can see, the anomalous R merge is lower than the normal R merge, indicating (as I understand it) that there is little or no anomalous signal. This (as I also understand it) indicates twinning and the twinning cancels out any anomalous signal. The truncate output, though, clearly indicates an untwinned crystal.

N	1/resol^2	dmax	Run1	AllRun
1	0.0128	8.85	0.079	0.079
2	0.0255	6.26	0.082	0.082
3	0.0383	5.11	0.082	0.082
4	0.0511	4.42	0.075	0.075
5	0.0639	3.96	0.076	0.076
6	0.0766	3.61	0.078	0.078
7	0.0894	3.34	0.086	0.086
8	0.1022	3.13	0.095	0.095
9	0.1149	2.95	0.111	0.111
10	0.1275	2.80	0.137	0.137
Overall			0.082	0.082

N	1/d^2	Dmin(A)	Rfac	Rfull	Rcum	Ranom	Nanom	Av_I	SIGMA	I/sigma
1	0.0128	8.85	0.079	0.060	0.079	0.058	291	18431.	2824.4	6.5
2	0.0255	6.26	0.082	0.064	0.081	0.048	647	11072.	1877.5	5.9
3	0.0383	5.11	0.082	0.061	0.081	0.052	877	6824.	1198.3	5.7
4	0.0511	4.42	0.075	0.060	0.079	0.043	1084	8979.	1406.5	6.4
5	0.0639	3.96	0.076	0.059	0.079	0.039	1241	6881.	1032.3	6.7
6	0.0766	3.61	0.078	0.061	0.078	0.040	1383	4751.	740.5	6.4
7	0.0894	3.34	0.086	0.066	0.079	0.043	1521	2825.	453.7	6.2
8	0.1022	3.13	0.095	0.074	0.080	0.048	1648	1529.	259.4	5.9
9	0.1149	2.95	0.111	0.087	0.081	0.054	1744	954.	188.3	5.1
10	0.1275	2.80	0.137	0.111	0.082	0.061	1826	565.	125.0	4.5
Overall:			0.082	0.064	0.082	0.046	12262	4432.	955.9	4.6

Cumulative intensity distribution (Acentric and centric)

Z	N(Z)Atheor	N(Z)Acen	N(Z)Ctheor	N(Z)Cen
0.0	0.0	0.0	0.0	0.0
0.1	9.5	9.8	24.8	25.8
0.2	18.1	18.8	34.5	35.5
0.3	25.9	26.9	41.6	41.7
0.4	33.0	34.0	47.3	47.2
0.5	39.3	39.9	52.1	52.6
0.6	45.1	45.8	56.1	56.5
0.7	50.3	51.2	59.7	59.4
0.8	55.1	56.0	62.9	61.9
0.9	59.3	60.1	65.7	64.8
1.0	63.2	64.0	68.3	67.6

The questions are these:

1. *Are the crystals twinned, or is there another explanation?*
2. *If so, why doesn't truncate or the twinning server show this?*
3. *Can any useful info about the twinning be gained from the above two questions?*
4. *Is there any other way of showing the twinning, without the need to collect anomalous data (because otherwise it is going to be a hard slog screening to find untwinned crystals)?*
5. *Back to finding another crystal form?*

Summary from the enquirer: The overwhelming majority were of the opinion that the data were not twinned and that Rano doesn't need to be greater than Rmerge for there to be a signal. Ranom is the differences between Mn(I+) and Mn(I-) and will decrease as you increase multiplicity and get better data. But Rmerge reflects the scatter about a mean and

usually increases with multiplicity - that is why it is a pretty useless measure of data quality. A suggestion is the use of XPREP to check the data. This was actually run on the first set of data that we collected while we were at the ESRF and it indicated that the data were around 40-50%. This is where the idea originally got into my head. Initially I discounted this result because everything else looked OK. But since I haven't been able to solve the structure with either SOLVE, SnB or SHELXD, I was beginning to think that maybe XPREP was correct. Can someone tell me where to get hold of XPREP? Is it only available through Bruker? Hot off the press: the XPREP analysis will also be available in SCALA in the new year. Somebody pointed out that "A trivial (if unpleasant) possible explanation--the Se-Met residues are all disordered". This is something I had considered but rejected on the account that there are (meant to be) 10 Se atoms in the a.s.u. The Rmerge is quite high especially in the low resolution bins. This I had noted (and also the rather low I/sigI) which was part of the reason I think something funny is going on with the data. TRUNCATE is for general cases of merohedral twinning. You can have a variety of other nasty artefacts like hemihedral twinning and whatever. You could be able to see funny effects in TRUNCATE output in the table listing h/k/l odd/even intensities. If odd intensities are less or more than evens that is usually bad news. Hemihedral twinning can be seen by careful examination of the diffraction images, as double spots in higher resolution with preference along specific lattice directions. With this suggestion, came an example from experience: We had 3 years of that. A P21 disguised as C2221 which was hemihedrally twinned P21 at the end (or so I like to think). What worked was getting actually another protein... If the protein xtallises and shows some non-standard (merohedral) twinning (which is usually due to a high-symmetry shape of the molecule) I think it usually means that you have two separate protein species that interconvert during crystallisation and can both be incorporated to the lattice, since the difference is small. In MutS, which is an ATPase, adding ADP together with cutting 53-c-terminal residues did the trick. This may be an important clue. The protein involved is mistargetted by mutants that make the protein temperature sensitive. These switch at around 30°C. So even at 15°C there will probably still be some population of both forms - enough to screw everything up maybe.

B-factor

B-factor and resolution

(January 2001)

Does any one know if there is any correlation between the overall B-factor of a structure in relation to its resolution? Are there any publications on this topic? Also is there any correlation between the extent of disorder in a structure and the R-factor/Rfree?

As usual, the B-factor stirs up some controversy.

The first reaction to the question was: Well, I had a quick look at the data stored in QDB (gjk, acta cryst d52, 842-857) which shows that for 435 structures the correlation coefficient between resolution and average B is only 0.06, *i.e.* insignificant. The only non-trivial correlate (using a 0.2 cut-off) is the percentage of secondary structure (makes sort of sense) with cc=0.20. In my other large-scale test, mentioned a couple of weeks ago, I found that essentially all temperature-factor-related statistics are "incorrectly" correlated with measures of model accuracy (*e.g.*, higher average B tends to be accompanied by higher accuracy!). Average B is very strongly correlated with

completeness on the other hand. I suspect that problems with data and/or restraints (rather than physics) are a major determinant of the temperature factors we calculate for our models.

Then there was a call to repeat this B-value (Debye Waller factor) analysis with structures determined from data better than, say, 1.7Å. It is believed that B-values are kind of fudge factors at resolution lower than maybe 2.5Å, whereas at higher resolution they indeed make sense, since the restraints are practically downweighted by the X-ray term.

Armed with a quote by Eleanor which was a reaction to a ccp4bb query on 26/27 October 1998:

```
> 3. What's the significance of the atomic B-factors when you have a low
> resolution data, for example, 3.0Å; or 3.5Å.
 Very very little - common sense indicates that if the data peters out at that
resolution the overall B must be 50 or greater..
But depending on scaling procedure it can be seriously under-estimated - there
are several structures in the PDB with swathes of negative Bfactors!
```

another reader enumerates how the *average* B-factor may be 'normalized' (or corrupted, this reader might have called it) during the course of structure determination:

1. When putting the data on an absolute scale, a B-factor as well as scale factor is applied, to make the average B 0 or 20 or some ideal value (however note the default behaviour of TRUNCATE is to apply the scale but NOT the B-factor, so some intervention is required to corrupt the B-factor at this stage). For isomorphous phase determination a B-factor must be applied to bring all data sets to the same scale, but it should be applied to the derivatives not the native.
2. When making maps to build the model, a negative B-factor (sharpening) is often applied to enhance high-resolution details. This is well and good, but the final model should not be refined against this "sharpened" data, but against the original data.
3. During refinement of low-resolution structures, the problem of fixing scale and B-factors for protein and solvent models may be somewhat underdetermined, especially when the solvent model is the same as the protein model (Babinet-type approach used in REFMAC, see Kostrewa's article in the September 1997 CCP4 newsletter, and earlier work *e.g.* Fraser *et al.* 1978), and an arbitrary choice of some parameter can make the process more robust. From the refmac documentation:

   ```
   SCALe      LSSC      FIXBulk      SCBulk      <scbulk>      BBULk      <bbulk>
   [Lower resolution structures may not have sufficient data to find sensible
   overall scales and B values for both the BULK and the protein component.
   It can help to fix these.]
   ```

Suggestions/recommendations from this reader:

- I have the impression that using a mask-based solvent correction as in CNS the B-factors for solvent and protein can be well determined at 3 or 4Å resolution. This could be tested by writing out F-part and F-model and scaling them against the data with ICOEFL, which prints some statistics about the correlation between terms.
- The correlation of resolution limit with *minimum* B-factor is probably better than with *average* B -factor. There are many examples of high-resolution structures with disordered loops; the contribution from the disordered parts would drop out at low resolution and the resolution limit would be determined by the best-ordered parts of the structure.

- I recommend a new REMARK card for deposited coordinate files which would indicate whether the final atomic B's are refined against original data in an attempt to determine absolute B's, or whether the overall B is arbitrary and atomic B's should only be used to see which parts of the structure are relatively well- or dis-ordered.

Another reader suggests that the low correlation between B-factor and resolution may be partly due to the following: small crystals collected on an in-house source might diffract only to 3.0Å while still being well ordered (*i.e.* low B-fators). From a large crystal using synchrotron radiation you may be able to reach 2.0Å even though it has higher B-factors.

The first reader reacts:To be sure, if factors like size of the crystal and synchrotron source were far more important than B-factor in determining resolution, the CC might be negligible. But I think the opposite is the case. First of all I have a gut feeling that if my lousy crystal diffracts to only 3Å, dropping the B-overall to 10 would give a greater improvement than making the crystal 10x bigger or going to the hottest synchrotron in the world. (Unfortunately dropping the B-overall is the most difficult approach to take, unless we find a better crystal form.)Slightly more quantitatively, say B-overall for the structures range from 10-70. At 2Å, and if I haven't dropped a factor of 2 somewhere, that makes a 1,808x difference in intensity. Say scattered intensity is proportional to the number of ordered electrons in the beam. Going from an 0.1 mm crystal to a 1 mm crystal would give 1000 times the intensity, but I wouldn't expect such a dramatic improvement in resolution, partly because much of the background is from scattering by the crystal, and would increase nearly in proportion. Also that intensity is spread out over a bigger spot, so peak intensity is increased by a smaller factor. Going to a smaller unit cell makes the average spot intensity greater because that total scattering is divided between fewer reflections. But the variation in unit cell volume for the majority of protein crystals is probably less than 100-fold.
Perhaps the hottest synchrotron in the world has 1800 times the brilliance of an x-ray tube, but I doubt if the signal/noise is better by that factor. So I doubt if any of these factors is great enough to completely overwhelm the effect of crystal order in Gerard's statistics. but maybe taken altogether? and with other factors I haven't thought of? As was pointed out: not all crystallographers use the same criterion for reporting resolution of a crystal, which would add further jitter to the relationship. A number of people indicated that low resolution B-overall shouldn't be taken literally. That was actually my main point, then I wanted to ask "can we do better?" or should we acknowledge that fact in a REMARK that will warn the non-crystallographer against using the B-factor as a criterion of structure quality when comparing low-resolution structures? (OK- maybe non-crystallographers pay no attention to B-factors and even less to REMARK statements).

Summary from the enquirer: To summarize, many of you believe that there is a (good) correlation between the overall B-factor and the resolution cutoff. But then Gerard's statistics showed otherwise. Some of you attributed this observation to the correlation being masked by effects of experimental limitations.

Anisotropic ellipsoids

(March 2001)

According to many textbooks the first three of the thermal parameters U11 U22 U33 U12 U13 and U23 describe the displacements along the perpendicular principal axis of the ellipsoid and the latter three give the orientation of the principal axes with respect to the

unit cell axes. However, I can't find anywhere how U12 U13 and U23 (apparently as direction cosini) exactly describe the orientation of the ellipsoid, say in a cartesian system. Any hint is appreciated (but don't suggest to try to follow the ortep code)...

Summary from the enquirer: First of all, my question was based on the false assumption that U11, U22 and U33 are the components along the principal axes of the ellipsoid. The text on page 533 of Glusker et al. "Crystal structure analysis for chemists and biologists" led me to that conclusion, although the example on page 536 indicates that things are not as simple as that. U11, U22 and U33 are the <u2> values along the reciprocal cell axes a*, b* and c*, respectively (*e.g.* Drenth, page 94). The principal axes of the thermal ellipsoid can be obtained from the U values via a principal axes transformation. This is described *e.g.* in Giacovazzo et al., p. 75 ff. and 148 (don't rely on the index), in the ORTEP manual, International Tables Vol.II p.327, and $CLIBS/rwbrook.f

For the full summary, including equations, see the CCP4BB archive version of this posting.

Movies and other picturesque queries

Structural Transition

(January 2001)

Is there a program that can make a movie of a protein structural transition, given a "start" and an "end" conformation of the same protein? We have determined two very different structures of one protein domain, and would like to present the structural transition in a reasonable way. Going from one structure to the other may involve unfolding part of the protein and refold it. Such a big conformational change is difficult to model, therefore, a program with some level of automation would be really helpful.

- You can do that with LSQMAN, see: http://xray.bmc.uu.se/usf/mol_morph.html. For an alternative, see: http://bioinfo.mbb.yale.edu/MolMovDB/.
- I recently did a fairly complex cartesian-space interpolation between multiple structures with different numbers and types of atoms using OpenDX. This may be of use if, for example, you find that a covalently-bound oxygen is replaced by a crystallographic water and you want to animate the change. We also animated movements in crystallographic waters. The process was tedious, but could be done. I suspect lsqman is an easier solution in cases where you are only interested in conformational changes of a single structure rather than chemical changes. Another alternative is to use Ron Elber's method of finding paths of minimum energy on the potential surface by minimizing an unusual action functional. You can specify starting and ending states and in initial guess for the path (often a line). This method takes you a step beyond LSQMAN in that an empirical forcefield is used. You should be able to get the code from Ron through the NCRR at Cornell.

Summary from the enquirer: The morph server at Yale seems to be easier to use. However, I had some trouble getting results, probably due to the fact that some serious unfolding is involved in my case. The authors have been notified about the problem and hopefully they are trying to fix it. Haven't tried other programs yet. As a word of caution: this kind of "movie" will need more justification as to its biological relevance. Our purpose of making such a movie is just to show the magnitude of the structural changes.

Digital Imaging of Crystals

(May 2001)

I would like to purchase a system to record images of crystals electronically. If anyone has come up with a relatively cheap method of doing this, I would be grateful if they could share their experiences. I guess the cheapest way is to stick a digital camera on your microscope - we already have the adaptor for a regular SLR camera. However, I would also like to hear about other, perhaps more sophisticated solutions.
Then, after a few days, this was added: *In the light of some of the responses I should have qualified it by saying I wanted a system that gave me an instant result. I didn't want to record a whole tray automatically, just the ones with crystals. Neither did I have a requirement for sophisticated annotation features. I just wanted to be able to transfer the images easily to a PC.*

Summary:

- Olympus is offering a rather sophisticated solution for a digital camera. You can catch that signal 'live' via the analog output of the camera at low resolution (around 600x400) and you can also take stills in high resolution 2048x1536. You have to buy the camera (~2000 Euro) the frame grabber for the PC (~500 Euro) - the PC obviously - and some software from Olympus (which IS necessary to combine the live and still-high-quality capabilities) which is another ~1500 Euro.
  The alternative we chose (again from Olympus) was to buy from them a JVC camera for ~1800 Euro for live image and use the frame grabber to save images. The quality of that is not outstanding - by any means - but good enough even for publication in small size - *i.e.* single column Acta D. Some free-ware framegrabbers (*e.g.* IrfanView) have capabilities for time lapse photography. Together with a real 'cold-light' source it can be fun and educational to take pictures of crystals growing.
  Another solution is the Pixera cameras which have some cheaper models which are fine. You can buy these from Olympus as well or directly from Pixera. Olympus will be slightly more expensive, but then they gurantee that the whole boogie works.
- A much cheaper option would be to use a flat-bed scanner (no need to spend more than 50-100GBP; if you want to scan 35mm slides as well you can buy an adaptor for many scanners for an extra 30-50 GBP) to scan photographs of your crystals taken with your film SLR. Of course, you'd still have the running costs of film, and delays in processing etc... A reaction to this suggestion:
  If I might respectfully disagree here, flat bed scanners are often extremely poor negative/slide scanners. They are especially atrocious for slides. Much better to get a slide/negative scanner (HP, Canon, Nikon, Minolta, Poloroid all make respectible models), *e.g.* the HP Photosmart S20 gets good reviews. There's a fair amount of www info out there on the "digital darkroom" if you want to go that route.
  The response was: No need to be respectful about it - I haven't tried the slide/negative adaptors so can't make any comment about their quality! However, I note that the US price of the Photosmart S20 is $499, which is rather higher than the cost of the slide adaptors I suggested. You pays your money and you takes your choice...
- We bought a Nikon Coolpix950 last year with an adaptor (sold by Nikon) to our Nikon microscope and we are very happy with its performance. It records the pictures on a flashram card which can easily and fast be transferred to a computer with an USB port. This is much cheaper than special high-end digital cameras for microscopes but my feeling is that it is more than enough for our purposes, with the

additional advantage that it can be used as a normal digital camera as well if you want to document something in the lab. We also use it for PAGE gels etc.

- There are several alternatives:
  - CrystalScore from Diversified Scientific, Inc. is one option. They have an automated stage and can take one complete set of pics from a crystal plate.
  - Emerald Biostructures also sells a good digital camera for a microscope, and a notebook system for recording and annotation the images (note from mgwt: but their website isn't really in English...).

  The basic issues are what are you going to do with the images. Do you want to save them all, or just one or two from a crystallization run, or time-elapsed images. The easiest thing to do is get a good digital camera for the microscope, take the image, and use photoshop, or some other application like it to modify and store the image. Good digital images are about 1MB in size, with enough resolution to zoom in after the image is collected. If you are talking about saving an entire set of images from a crystal plate, it's more complicated, since you have to worry about where the drop is, the zoom level, focussing, etc.

- I have a video camera (#700) attached to my microscope (Leica) which is attached to a Matrox video card (#800) on a PC. The system works reasonably well, and I can capture images to put into Powerpoint presentations, and also for archiving crystallization tray results. The system is about 4 years old. I think video cameras cost about the same, but video capture cards have come down in price a lot. I am told that the quality of the picture I get in the monitor is pretty good and much better than the system set up at the MRC (Cambridge). I got my information about video capture from the microscope representatives when I bought my microscope. They are of course interested in selling the most expensive high quality system, but if pressed they will offer cheaper alternatives. This is what I did. The risk I had was the unknown quality of the captured image when I bought the hardware. But I think it is pretty good for almost everything I want to use the images for.
- We bought a Pixera camera about 3 years ago....primarily because it was so afforable (~$1200 at the time which was quite good then). We still use it, but the old adage is definately true: you get what you pay for. It is slow and the quality is pretty good at low magnification (on the scope) for "macroscopic" objects, but when you get down to the level of most typical protein crystals (100 microns or less), it doesn't do such a fabulous job. Also, it's purely a manual setup - no options for auto-scanning trays or dropping all of the images into a database or anything like that. I can forward a representative image if you're at all interested..... On the flip side, I know several people who have bought the digital microscope cameras from Kodak - there the quality is much higher, but I understand that it is also much more difficult to use - the images are stored on the camera until you manually download them to a computer. The Pixera at least works through a card that you plug into your computer and images are dropped directly to disk.
- Birdwatchers have been doing something analogous for a while - taking digital pictures from the optics of their (rather high quality) telescopes ("digiscoping"). With digiscoping, often the simple expedient of putting the digital camera up to the eyepiece and taking the picture will work. Some tinkering with focus is sometimes necessary. The digital camera's picture review facility makes life easier. See: digiscoping as an example. The pictures are surprisingly high quality. I am guessing that the same approach will work with microscopes as with telescopes since the optical designs are closely related.

- What we did is similar to what you have, but instead of taking the images with an "off the shelf" digital camera, we purchased a ccd chip, a focusing lense and an electronic board. After assembling the components, we mounted it on a C-mount. We connected the output terminals to a computer and to a small TV. The TV is used for oberserving the crystals and the computer is used for capturing and storage. You can use the computer for observing as well and not need the TV of course. We also connected a printer to the TV so that a low quality hard copy can be printed without going through the computer.
-

Our students found that you can take any digital camara (ie one meant for photographing scenery on vacation), hold it just so over the microscope eyepiece, and shoot quite nice pictures. If you make a little cardboard adaptor tube that fits over the eyepiece, its even easier. The preview thingy on the back of the camara is crucial. The attached pic was taken with a Canon PowerShot A5 (click to enlarge).

- I am very happy with our Olympus AX70 Digital microscopy system. It has the Olympix 2000 digital camera on it, and DIC optics. I admit, maybe it was a leetle bit pricey....
  I would suggest also getting the lowest power objective available - sometimes I grow crystals that are too big to photograph! For crystals grown under oil, you might wish to purchase an inverted scope.
- If you're tending toward the high end, I suggest looking into a robotic microscope stage and crystal tray manipulator so you can give it a stack of 12 trays and have it take a picture of each well at 0, 1, 6, 12, 24 hr, and daily thereafter; without the necessity of some human coming into the cold room and breathing moist air all over the lenses. Then if you solve the structure from the coffin-shaped crystal in well C5 of tray 7, you can go back and make a time-lapse movie of the growth of that crystal to show in you powerpoint presentation. And get Emerald or Hampton to mass-produce the system and sell it for under $10k so we can all get one.
  On the low end you can get adaptors to put an inexpensive ccd video camera on the same port used by the film camera, and something like Connectix "Quick clip" device to grab video or still images from the video stream. Resolution is lousy, but if you zoom in till the crystal fills the view it's not that bad. Pixera has a digital video system with the same functionality but refresh rate is much slower than video making it difficult to focus (at least on slow PC's).
-

I use a Nikon system. An adapter arm fits between the lens and binoculars. You can then place a threaded mount on top of the adapter and screw on your digital camera. I use the Nikon Coolpix 990 which runs about $1000. But you can use any digital with a threaded lens mount.
Attached a picture of crystals that I took with the system (click to enlarge).

- Our cheap trick is to use the little ccd camera that SG gave away with indies a few years ago.

Image production

(June 2001)

I'm preparing some color images for structure paper submission. However, there is still not satisfactory solution for producing images of required resolution. Any image cropped by snapshot on the SGI work stations only has a resolution of 72 dpi, making it unrealistic for further processing or direct submission. Trial-and-error photography of these images displayed on the screen using the best film-loaded or digital camaras suffers a lot from the over-saturation of local white regions and the white margins of imgview or imgworks, and terrible distortion of the image by the screen. Could any person give me some tips about this issue?

Some more detail was added after a few days: *Thanks for those who have responded to my problem. Before I could report a summary about this issue here, I wish I could have a chance to go into some specific details about my problems. GRASP: Grasp would produce nice .ps files. However, for purpose of further processing, such as for labeling charged residues, I have to outport it from SGI to Adobe Photoshop in PC. Since I don't know any img-format (from .ps to .tiff, for example) conversion programs in SGI, I have to use 'snapshot'. Suggestions of using gimp or imagemagick for format conversion have led us to download the programs. But the installation of gimp failed. It complained that 'the gtk-config script installed by GTK could not be found', although we've installed the glib-1.2.8.tar.gz (obtained from The GIMP Toolkit) beforehand. Please look at http://www.ccp4.ac.uk/newsletters/newsletter40/gimpproblem.log and I wish somebody could help me out with this. Imagemagick seems to need more other things. Stereo-pair electron density map superimposed with structure model: O is exhaustedly used for model building, but for image production Turbo-Frodo seems to achieve more brilliant color and much better ball-and-stick model, and is able to produce stereo pairs (although they will crossover at the middle). Sometimes the feature of Van der Vaals surface presentation in Turbo could be a simple reason why using it. In such case, snapshot seems to be the only way of catching the images. I've been advised to use Bobscript and am lucky enough to obtain it today (I wish it will help soon). But still, is there any other program producing good VDW surface images? Program Molscript and Raster3D have been running on our SGIs. Stereo-pairs production by Raster3d needs tiff library. However, we have real trouble in installing the tiff library. This make the production of stereo pairs with Raster3d impossible. We're using SGI O2 (Irix 6.5 operating system). We've downloaded the tiff software from TIFF Software (file: tiff-v3.5.6-beta.tar.gz), but failed in compilation (there seem to be many errors, see http://www.ccp4.ac.uk/newsletters/newsletter40/tiffproblem.log). Have I got the right thing, please?*
It seems that if these two problems could be solved, I would be able to find my way out.

Summary:

- Molscript (for ribbon diagrams and amino acid residues
- Bobscript (for map files)
- Conscript
- Grasp
- Rasmol
- Chime
- Molmol
- VMD

- Setor
- SPOCK
- MidasPlus
- RIBBONS
- PyMOL (for maps in XPLOR format)
- Swiss-PdbViewer
- POV-Ray
- ImageMagick (for image manipulation tools)

Some tips and hints:

- Many of these can generate raster3D files which can then be made into several file formats including tiff. Possibly the best thing to do, is to increase the size of the file in raster3D so that it can be resized without altering picture quality.
- XV has the pesky feature that it only outputs as many pixels as displayed. BUT: In the XV box "save" dialog box, there is a checkbox labeled "use normal size" or something like that. If you check this box the full size will be saved, even if it is too large to fit on screen fullsize.
- If you have the opportunity to use Photoshop or almost any other Image Processing Program, you should be able to import a postscript file as "generic eps" and then you're also able to increase the dpi's as you need it.
  You might also try the following (have all your images in the directory snapshots):
- ```
  #!/bin/tcsh -f
  ```
-
- ```
  set files=`ls snapshots/*`
  ```
-
- ```
  foreach image ($files)
  ```
- ```
     makemovie -o ${image}_mod.rgb -c jpeg -f qt -r 1 -s 800,600  ${image}
  ```
- ```
  @ counter ++
  end
  ```
  You can also increase the size (-s 800,600) to something what you might need, but remember the resolution is still 72 dpi, so in order to get more dpi's you'll have to shrink the modified image with *e.g.* Photoshop, Gimp, Showcase.

- More links with extras:
  - http://toolbox.sgi.com/TasteOfDT/public/freeware/
  - http://pov4grasp.free.fr/
  - http://freeware.sgi.com/index-by-alpha.html
  - ftp://ftp.x.org/
  - http://www.imagemagick.org/www/archives.html, which also has links to some or all of the following:
  - http://ftp.nluug.nl/graphics/ImageMagick/binaries/home.html
  - http://www.planetmirror.com/pub/imagemagick/binaries/
  - http://www.ccl.net/cca/software/X-WINDOW/ImageMagick/binaries/index.shtml
- You can generate stereo pairs with RASTER3D without the tiff library. You can generate two views rotated around y by 3-6 degrees and then paste the two views together (ImageMagick montage or Photoshop). You will need to use Normal3d to generate a raster3d file with an identity transformation matrix in the orientation you want. Should look something like:
  ```

- 1 0 0 0
- 0 1 0 0
- 0 0 1 0

  0 0 0 1

  somewhere near the top. Use this raster3d file to generate the first image. Then edit the file so that the transformation matrix looks like:

  ```
  1 0 0.1 0
  0 1 0 0
  0.1 0 1 0
  0 0 0 1
  ```

  and render again (with different output name of course). Then crop both images leaving about 5-10 pixels of background around the edges. Join the two images along one of the vertical edges and you've got a stereo image. Join them one way and you have wall-eye stereo, join them the other way and you have cross-eye stereo.

How to control the size of .ps files in NPO

(June 2001)

Here I have a chance to ask for help to a problem with the old command-line ccp4 version at a brandnew era of ccp4i. Year ago I produced some patterson maps for heavy atom harker peaks in P212121 space group. The ps file is fine when printed out, but when viewed with xpsview, the top portion is missing! Today I try to convert it to pdf format using acrobat distiller, the top portion in the pdf file is still missing. I'm thinking of aligning different sections up so as to have a nice view of the heavy atom sites. Of cause, printing the images out and then scanning them back to computer is a way out. But it is really a clumsy one. I'm attaching the input file and one of the problematic ps file here, wish anybody could give me a shortcut. In fact, I could not find a control line to specify the image site in npo. Would it be possible for me to get the control by modification of the ASCII ps file?

Here are the NPO script (http://www.ccp4.ac.uk/newsletters/newsletter40/nposcript.com) and the NPO postscript file http://www.ccp4.ac.uk/newsletters/newsletter40/nposize-jun2001.bin (which may be saved and viewed with your favourite postscriptviewer).

Most suggestions are directly related to editing the ps file:

1. The bounding box in the postscript file is wrong. If you have ghostscript up-and-running do a

   ```
   gs -sDEVICE=bbox nposize-jun2001.bin
   ```

   to get the right bounding box and change the corresponding entry in the postscript file (%%BoundingBox: and %%PageBoundingBox:).

2. Replace the line that says

   ```
   %%BoundingBox: 0 0 365 800
   ```

   with the followin three lines

   ```
   %%Orientation: Portrait
   ```

```
%%DocumentMedia: A4 596 842
%%BoundingBox: 18 18 578 824
```

3. Insert a scale command in the postscript file at the end of the postscript file header. *E.g.* to scale both x and y by 0.5 add "0.5 0.5 scale"

```
4.  %%EndProlog
5.  %%Page: 1 1
6.  %%PageBoundingBox: 0 0 365 800
    0.5 0.5 scale
```

7. Have you tried using ghostview instead of xpsview? The file looks fine when I look at it in ghostview. I think xpsview is broken with respect to large bounding boxes. Distiller may be too. Failing that, add the line:

   ```
   0.5 0.5 scale
   ```

   as line 14 of the ps file.

8. I used Illustrator running under classic running under osx and the file opened without a hitch. I exported it to a jpg file in rgb color mode with standard compression. Hence it is not your file but your programs that are to blame.

(note from mgwt: I tried ghostview, presumably the same version as in suggestion 4 as I'm in the same lab, and for me it did **not** display the top part. What **did** work for me, was to use xv.)

Movies for powerpoint

(July 2001)

I thought someone had recently enquired about how to make movies of rotating structures for importing into PowerPoint. I've searched the ccp4 archives but can't find the Q/A. So, what programs do people use to do this?

- The easiest way is using software like (Gifsicle) or comparable software to combine several ...gif files into one animated gif. You have to manually rotate your structure say 5-10? each time and save a .gif image afterwards for input into Gifsicle. Alternatively, you can use video grabbing software. I've tried a few demos available online, but the result usually is not better than with animated gifs. Powerpoint takes .avi videos, e.g. movies made with a CCD camera are easy to integrate into a presentation. If you find molecular graphics software that saves .avi files, please let me know!
- I would suggest something which can be done with a script, so you don´t have to convert images by hand. The general outline:
  1. Make Images of your molecule
  2. Render them
  3. Convert them to a movie-format (MPEG, Quicktime, DivX, AVI, ...). You could also use a animated gif, but the quality isn't that good.

  Scriptable Programs for image generation might be:

  - Molscript
  - Molmol

- ICM
- WebLabViewer
- InsightII(?)
- Rasmol (no publication image quality)
- .....

Renderers:

- PovRay
- Raster3D
- Renderman
- .....

Converters:
There are several commercial applications, which can produce movies from single images.

- ppmtompeg (?)
- ImageMagick & MPEG
- Quicktime-encoder (Apple ?)

I would write a script which does the rotation for you and writes out and renders the imagefiles and concatenates them to a movie. So you only need to prepare the Inputfile for your Renderer (say Molscript) and afterwards change the rotation/translation-matrix stepwise to get the desired effect. Of course you can do without rendering, but it will look better.

- 'Ribbons' for the SGI will make a series of images with small rotations or translations between each one. These are written as RGB files. A jiffy program gets them into a suitable form for input into 'makemovie' on the SGI again, which will give you a Quicktime format movie. You should be able to play this in Powerpoint, or convert it to an AVI with the movie conversion tool of your choice (SmatVid for windows?)
- Tony Crofts recently prepared the same sequence as an animated gif and an avi movie, from the original chime (~rasmol) script for his cytochrome bc1 web page:
  - http://www.life.uiuc.edu/crofts/bc-complex_site/etp-model_annotated.html
  - http://www.life.uiuc.edu/crofts/bc-complex_site/etp-model_annotated.avi
  - http://www.life.uiuc.edu/crofts/bc1_in_chime/etp_model_struc.html

  The gif is 4 MB as compared to almost 7 for the avi, and the gif starts playing while it is still loading, so seems much faster, at least for a web page. The CHIME presentation is just the pdb files and some script, so less than 1 MB in this case, but I don't know if you could run chime inside powerpoint. I'm sure Tony would be glad to tell you how he made the avi.

- You can try Molray that was developed by Mark Harris at our lab. Molray is a web interface to ray-tracing program pov-ray to generate still images or movies from O plot files and other sources. You can even run it on our server. It is very easy to make movies if you just need simple rotation and translation and it can export MPEG,QuickTime movie or animated GIF.
- We also use Mark Harris's molray for general rotations, but we've used a couple other methods which produce good output with relatively little effort.

0. There is a version of molscript available that outputs in povray format. Adding a header to the outputted povray file to rotate it is relatively simple. The output is a series of images that can be seamed together with Gifsicle for animated gifs or moviemaker on the sgi's for quicktime output. We then take either form and convert it to avi with Quicktime pro or just keep it in quicktime. I've been playing with the idea of making a script to add this header automatically, but haven't gotten around to it. Is there interest in this? It shouldn't be a very big project.
1. To make movies that "morph" we use LSQMAN. LSQMAN will output a series of images between two positions. We then take these images and seem them together in the same way as above.
2. If you want to put two movies side by side (a favorite of my PI) you can use "montage" from the ImageMagick suite. We have a few tricks that make the movies play a bit better and I use a couple of very simple perl scripts to avoid some ugly command lines, but it isn't too difficult (you are a bit limited with this format since not all movie formats like movies which are not perfect squares).
3. If none of the above work we can kick out a series of stills from ribbons, molscript/bobscript or even O and seem them together, but thus far this has not been used for anything serious.

- I create individual frames in MidasPlus running on SGI (use a simple script to render picture and rotate repetitively), stitch the individual frames into a QuickTime movie using MediaConvert on SGI (various options available for compression and file formats) and then ftp to my Mac..
- PyMOL is simply the best molecular movie-making solution available. Nothing else even comes close in terms of ease and capabilities. [Disclaimer: I am the author and may have a biased view.] The program was written specifically for generating movies and includes
    0. Real-time OpenGL graphics for proteins & maps
    1. The ability to read Molscript's Raster 3D ribbon input files
    2. A built-in ray-tracer which gives you WYSWIG rendering
    3. A powerful movie description language
    4. Support for multiple structures/coordinates within a movie
    5. Previewing of dynamic movies in 3D (OpenGL)
    6. Previewing of raytraced images in 2D (from memory)
    7. PNG format export
    8. A built-in Python scripting language to automate conversion and compression

  Plus, it is free and unrestricted open-source -- my gift to the field. Here is an example 5 line script for creating a ray-traced movie:

```
mset 1 x120
util.mrock 1,120,15
set ray_trace_frames=1
set cache_frames=0
mpng mov
```

  It outputs 120 PNG files for generating a 4 second, 30 fps movie. If you don't want ray tracing, then leave off the "set" commands.You can obtain the latest version at Sourceforge. I recommend using it under Windows with an nVidia card, but it works under Linux, Tru64, IRIX, and soon OSX as well. I have yet to find an open-source unix solution that produces the same quality AVI files as Adobe Premiere, thus I tend to use PyMOL to do the rendering, Imagemagick to convert to TGA, and

Adobe Premiere for the final compression (Cinepak codec at 99%). Alternatively you can use Imagemagick to batch convert to a format that SGI's media tools can read.

- On an SGI you can just use the following command (after you have made your rgb files)

```
#!/bin/sh

name=

end=`expr $2 + 1`
num=$1
while [ $num != $end ]
do
        name="$name $num.rgb"
        num=`expr $num + 1`
done

echo $name

makemovie -o $3 -c jpeg -f qt -r 10 $name
```

this will generate a QT Movie, but be careful:
If you're using Powerpoint on a Wintel machine, don't expect Powerpoint to manage your QT movies, you'll need mpeg or avi; running Powerpoint on a Mac is no problem !!!

- VideoMach is a not-great-but-good-enough tool for assembling together other movie files, gif files, animated gif files or whatever and outputing a decent variety of formats. It has a free 30 day trial and a single copy costs 50$. If you really find the 50$ too much you can reinstall every 30 days ... but that is a 50$ well spent! http://www.gromada.com/
  How to get the 'frames' I guess it will always be a matter of taste. We mostly use Bobscript/Raster3d to get RGB files, then ImageMagick (freeware) to get an animated GIF (you can preview that in Netscape). Then I ftp the GIF in my notebook and convert it to mpeg or avi with VideoMach.
  Note that PowerPoint 2000 will play animated GIFs but it is truly pathetically slow. The same gif in the same computer plays great using Netscape though ... yet another example of enlighted programming from Microsoft. Talking about it, in PowerPoint when you import a movie or an animated GIF it shows in 'true' resolution while editing (*i.e.* a 'file' pixel takes a 'screen' pixel) and when going to 'full screen'/'presentation' mode it scales it up (*i.e.* resamples the image with some sort of undocumented dithering technique), which most of the time goes unnoticed but in fact makes the image quality a bit worse. Does anybody know a trick (other than not to use PowerPoint) to ovecome this?

Stereo figure from molscript

(August 2001)

Is there an easy way to generate a stereo figure from molscript? I couldn't find anything in the documentation.

Summary from the enquirer:

- Just use the same input file and add "rotate y -6.0" to the transformation to generate the right hand figure. I've also used -8° myself. You can also rotate the left image by +3.0 and the right image by -3.0 relative to the orientation you have worked so hard to obtain.Alternatively the Bobscript distribution includes, or at least has included, a jiffy script to do this automatically.
- Create an r3d output file with molscript. Render the r3d file with the stereo option in raster3d.

How to generate postscript files, and how to achieve the correct resolution

(January 2001)

I am using Bobscript to generate image files with electron densitiy maps. Is it possible to save them in postcript format? If this is not possible what is the best way to submit rgb files to publication??

Bobscript outputs postscript files by default (without any flags), *i.e.* bobscript &lgt; input.inp > output.ps.

Also, you can put labels etc. within Bobscript itself; no need to take it elsewhere for that purpose. The area command on top of the file can set the exact size of the output for printing or including in any documents. To help with this, there is a grid (in the O distribution): edit.ps. Usage: print or copy onto an overhead, overlay on your plot and read off postscript coordinates. Displaying a file with `ghostview' or `gv' and reading the mouse coordinates is another easy way to determine PostScript coordinates.

The preferred format(s) is (are) in most cases explicitely mentioned in the instructions for authors. Most journals will like TIFF and EPS. On a related issue: If a journal requests 400 dpi (dots per inch) pictures and you plan the reproduction (print) size to be *i.e.* 8x4 inches, that means that you need 3200 dots on x and 1600 dots (pixels) on y. So if you make an RGB or TIFF file make sure it is 3200x1600 pixels in the first place. Importing a standard 'render' output of 1200x1200 pixels and then 'set resolution to 400 dpi' in Photoshop is not nearly a cure for good quality pictures …..... and, talking about photoshop: Do not forget that TheGimp is out there!

'Hardware' (and some Software)

Oils and cryo-protection

(January 2001)

This started off as a question about low-temperature data collection: *How do you collect a low-temperature dataset with a deoxyhemoglobin crystal without exposing the crystal to atmospheric air?*

The discussion evolved into one about oils used for cryo-protection.

Summary from a helpful bulletin board member:

It is clear from the responses that oil is no panacea, but it seems to work very well in many cases. We've had good luck so far, but organic solvents in the drop may pose problems. We do see diffuse scattering due to Si, but not enough to be concerned. Some suggest drying the oil as an aid in removing the water layer on the surface of the crystal. We suspect technique is very important here, and oil composition less important. We tried a silcone-based diffusion pump oil from Dow (750). It is thermally stable and claims to be radiation and oxidation resistant. References:

- S. Parkin and H. Hope, J. Appl. Cryst. (1998) pp945-953. Section 2.1 of this paper recommends Paratone-N, possibly saturated with water. Recommends against Si- or F-containing oils due to higher scattering power. Half the xtals they have tried survive oil treatment. Main problems are mechanical strength, loss of water by xtal resulting in cracks, or difficulty removing water layer. They are advocates of quick-dunk cryoprotection when oil does not work.
- H. Hope, Annu. Rev. Biophys. Chem. 1990 19:107-126. More details of oil/cryo handling (covering hanging drop with oil and dragging xtal through oil-water phase, wicking etc.)
- Riboldi-Tunnicliffe and Hilgenfeld, J. Appl. Cryst. (1999) 32, 1003-1005
- "The structures of deoxy human haemoglobin and the mutant Hb Tyrosine a42->His at 120K" Tame and Vallone, Acta Cryst D56, 805-811. It is possible to protect the crystals from oxygen using dithionite, at least long enough to cryo-cool them.

Then some accounts from users, both positive and negative:

Using oil is an excellent method and has been used for many years by small molecule crystallographers for freezing extremely air-sensitive crystals. I've used it successfully with macromolecular crystals too. I've used a perfluoropolyether oil for this (used to be Riedel-de-Hahn RS3000, but this hasn't been manufactured for many years. I haven't needed any since '95 so haven't looked into it seriously, but new sources have been discussed on this BB in the last year or so). For the small molecule case, it works by providing a physical barrier - the amount of oxygen that can diffuse through the oil is actually quite small. Also, something I didn't mention before - most air-sensitive compounds are actually sensitive towards hydrolysis, so it isn't the oxygen that reacts directly with them. Water, of course, is not terribly soluble in perfluoropolyethers. However, nothing which isn't pfpe is soluble in pfpe oils. For macromolecules, it stops evaporation of water from the crystal, giving you time to cool to create a vitreous phase. But the migration of oxygen through the oil is also limited, so that helps too.

We have used MO (mineral oil) only occasionally and with indifferent results. that is, sometimes we get useful freezing but never better diffraction. We purchased a 'panjelly kit' and tried their suggested protocols. Nothing (including lysozyme) diffracted any better than we had obtained by conventional means and in no way did we find any help annealing crystals. Add to this that the stuff does not perform well in the cold room we let it languish on the shelf for some months.

I tried 3 different oils and their mixtures - all successful so far and now I always use it by default. The first oil was the machine oil from the workshop, the latest - Paratone N. No special preparations were required.

Our laboratory has used oil, in place of a cryprotectant, for cubic lipid phase bacteriorhodopsin crystals successfully...

We've tried oil once so far, on crystals of a rather large protein-DNA complex grown from Ammonium sulphate. At room T, they diffracted to 13Å, and frozen in propane, 13Å, but the ones we tried in oil didn't yield a single spot (at a synchrotron). (And we did have help from someone who swears by oil). Now granted, these crystals seem to be useless no matter what we do, but oil-freezing certainly didn't improve things!

The oil method has worked very well with four different crystals in my hands, and it is now the first thing I try. It decreased mosaicity with regard to other cryos in one case, and proved essential in freezing one extremely fragile crystal without damaging it. The other advantage I find is that you do not need a artificial mother liquor. I have also had one crystal that it did not work with, so it is not always a sure thing. I have a feeling that in that latter case it may have had something to due with high solvent content. Briefly the technique I employ is as follows (for hanging drops):

1. cover the drop on the coverslip with a small amount of oil (20-40 ul). When I first read of this technique, I was eager to try it a a troublesome crystal and actually used fresh vacuum pump oil. it worked like a charm, and I have used it since with no trouble.
2. with a loop, fish the crystal out. I like to use a loop smaller than the crystal (spoon it). I get less of the mother liquor sticking to the crystal/loop that way. I also find that it is not to difficult to ge rid of any risidual mother liquor by passing the crystal back and forth through the mother liquor/oil interface. I had trouble with this and loops big enough to hold the entire crystal. The oil "glues" the crystal to the loop.
3. plunge in liguid N2 or freeze in a stream. I usually plunge myself.

Oils are great. We use perfluoropolyether, paratone-N, and 75:25 or 50:50 paratone-N:mineral oil. At least in one case where 100% paratone-N cracks the crystals, the 75:25 mixture worked.

I frequently use oils when using high salt precipitants as the phase difference traps the salt in the crystal and stops diffusion between the cryo protectant and the crystal.
I have found it usually works for most high salt crystals and some PEG grown crystals as well.
The problem is the oils diffract and give diffraction rings. I have always found parrafin oil (Hampton) works fine. It gives rings at ~4 and 2.3Å so a normal data set has only two rings. The rings are usually quite small so I don't loose much data. If it wasn't for the rings I'd use oil as first choise as it usually works first time and therefore saves time fiddeling with cryoconditions. Recently I got three to work from: 4M NaFormate, 2.5M A.S. and 24%PEG grown crystals

System backup devices

(February 2001)

How does everybody out there do their SYSTEM backups (SGI)? This question is related to a little discussion a few weeks ago on LINUX backups. Right now I am not doing any. scary ..... so I figure I'll have to buy something.

- *CD is out of the question, I guess. I figure I would need quite a few CDs to backup our 6 Gbyte hard discs. Plus some other machine to do the backup on, since on the fly burning seems somewhat risky? CDs work great for our MAR datasets, though.*

- *DVD is not quite there yet, as I understand. Also, would it be synchrotron trip compatible? Prob. not? Would it have to be? I guess CDROM is good enough for that. Would DVD hold a full backup? Don't think so (only 4.8 Gbyte).*
- *Leaves us with tape. I thought of some DAT DDS-3 system as a compromise for now. DDS-4 is a bit spendy still. same for Exabyte Mammoth (or so ...)*

Summary from the enquirer:

The *s denote the number of people mentioning the respective devices:

general

- One user found that tapes (in general) are not reliable for long term storage, but should be OK for backups (*)
- One user pointed out transtec for good prices on DDS-4 systems. This company has websites in various countries (check out the bottom of the first page). (*)

about different media

- DLT tape is highly praised, albeit expensive (***)
- DAT (at least the older ones) was reported to be problematic from the hardware side (break-down of tape drive and frequent cleaning required) (**) Seems like other people have been content with DDS-3 and DDS-4 (**)
- An Ecrix VXA (Ecrix) system was reported to be relatively cheap (as compared to Mammoth) and very dependable (*)
- One user is in the process of switching everything over to a backup server, doing all the backups via ethernet on an array of 80Gbyte hard discs. Thanks for re-emphasizing the difference between backup and archiving! (*)

Dry shipper container

(March 2001)

We recently purchased a Taylor-Wharton LN2 dry shipper dewar (cp-100) but are having trouble getting the outer shipping container that houses the dewar. Is there, by any chance, another company that makes these containers?

Summary from the enquirer: The company no longer makes the hard plastic outer container; this has been replaced with a much cheaper, somewhat reusable cardboard container. I suppose this is why our local representatives got nowhere with T&W when trying to ask for the plastic box.

Crystal growing cabinets and crystallisation incubators

(May 2001)

Does anyone have references/makers of crystal growing cabinets capable of covering a temperature range of 4-40 degrees C? Also, does anyone have any experience to report using the Mini-T product from Diversified Scientific Inc.?

Not long after that, a similar question: *Slightly off the topic but can you recommend crystallization incubators for 0 - 50 (90) degrees C, 50 to 100 Litres?*

Summary from enquirer1: Several refrigerated incubators (Revco BOD, Fisher Precision, EJS Systems, Inc.) have had reported temperature problems. At least one group has gone to the trouble of making its own temperature programmable crystallisation boxes (Personal crystallisation boxes) which might be available semi-commercially. The consensus (3/10 replies) appears to be that the Hampton M6 incubators covering a range of 4-60 degrees are the most dependable. The down side is these can only hold 6 Linbro trays.

Summary from enquirer2:

- We have an incubator made by VWR, model 1525, part 9120833. It spans the temperature range you mention, however I am not sure of the volume. If you would like me to measure it in order to calculate the volume, please let me know. Website: VWR Scientific Products.
  This is not a cooled incubator even if the company produces such machines.
- I bought two such incubators from

  Molecular Dimensions,
  61-63 Dudley Street,
  Luton, Beds LU2 0NP,
  Tel: 01582 481884
  Fax: 01582 481895
  Their Web site: Molecular Dimensions Ltd

  These incubators are good, reliable and vibration free as well.

- We have incubators manufactured by ehret. Good price and with suspended motors (vibration free). Website: http://manuf.labworld-online.com/ehret/.

Replating anodes

(August 2001)

We have a target from a Rigaku rotating anode generator where the copper is badly etched, so we can't use it. Has anyone ever had a target replated? Anyone done it themselves?

Summary from the enquirer: I asked about repairing a damaged target in our Rigaku rotating anode X-ray generator, and received many helpful responses. First I should clarify the problem. The target in question has a deep groove, probably caused by a combination of a cooling problem and having the bias set too high. From the responses, the standard dogma is polish, machine, or replace (re-cup), depending on the severity of the damage. There were several suggestions as to how copper might be added, sputtering and electroplating, but no one reported actually trying these methods. There was one suggestion that the target may not be pure copper, and if true, then adding metal by electroplating is not possible (I don't know about sputtering). I've got a call in to MSC/Rigaku and I'll see what info and prices I can get from them. Regardless, I think I will try to electroplate the one I have. It is too badly grooved to polish or machine, so what have I got to lose? I'll report back on how it goes.

Physical models

(May 2001)

Some people in my group seem to vaguely recall a way to have a plastic or rubber space-filled model made from a pdb file for ornamental or display purposes. Does anyone know of a company that does this type of thing?

- A company called Z Corporation makes a machine that casts models of a hardened resin from VRML files.
- I don't know if it's what you're thinking of, but the visualization group at SDSC has a laminated object manufacturing facility that can construct a model of a solvent accessible surface from layers of paper. The result of the process is a model that looks and feels like it's carved out of wood. They also have a similar machine that can produce plastic models that can be translucent or opaque, but I'm less sure of how that works. See NPACI & SDSC Visualization Lab and Tele-Manufacturing Facility Research Project for more details and some photos and explanation. They definitely don't run a mass production facility, nor a novelty factory, since it's a fairly expensive machine to run, but if you've got a "genuine scientific reason" for wanting such a model, they might be willing to make one as a one-off.

Dynamic light scattering

Interpreting DLS - discrete dimer vs. random assembly

(January 2001)

I'm running dynamic light scattering (DynaPro99) and am wondering how to interpret what I'm looking at. If any experts out there, I'd appreciate any input.
Scenario:
I have a protein where the active form and a previous xtal form both are homo-dimers (45kDa monomer). Previous xtal conditions were not screened for DLS. I observe the protein as a sharp monomer DLS peak in the storage buffer and as a BROAD DLS peak centered around 500kDa in the previously successful xtallization conditions (this is the same protein that gave xtals, but 2 months later). I can decrease the precipitant concentration to a point where I find a slightly-less-broad DLS peak centered around 100kDa... which could correspond to the dimer... or to the average MW of a random distribution of monomers and small-ish agregates. My thinking is that it's the small-ish aggregate option. My thinking is that if it were the active dimer form, the distribution would be just as sharp as the monomer distribution in storage buffer. I'm wondering if anyone has any rule of thumb about how sharp a peak needs to be to call the solution homogeneous?

Summary from the enquirer:

Concensus answers:

- Look at the errors: if they are large, redo the experiment.
    - baseline should be 1+/-0.003
    - count rate should be steady
    - SOS error should be less than 5

- Make sure the routine to exclude bad data points is enabled!!! To do this: (At the pull down menu) Tools -> Settings -> Data filtering Protein Solutions recommended these limits:
- `Over SOS Error:  not needed, coupled to rest`
- `Under baseline limit:   0.98`
- `Over baseline limit:    1.02`
- `Under Amplitude limit:  0.01`
- `Ignor 1st # coeff:      4`

  `Truncate at channel #:  120`

  To make these as default, set them with **no** data set open. If you set new limits with a data set open, the limits will apply to *only* the open data set. All previously taken data sets will contain the old limits... unless you reset the limits and run.... (At the pull down menu) Analysis -> Recalculate All for each data set.

- Look at the polydispersion index: This is the percentage obtained by dividing the dispersity of your peak (how broad it is) by the hydrodynamic radius. If it is less than 0.1 your solution is monodisperse. Some people went as high as 0.15 to say it is monodisperse. Above 30% is a polydisperse solution.
- Look at the bi-modal distribution: If the polydispersion index is larger than 0.1, the bi-modal distribution can sometimes tell you if there is a high MW aggregate that is contributing to the scattering. The % of each component is listed.
- Notes on applying info data to crystallisation:
  - The peak will always come to a higher MW than the true MW, unless your protein is a perfect sphere, due to unaccounted for additional rotation friction.
  - A protein does have higher chances to crystallise if it is monodisperse (which is not saying it *will* crystallise), but a low level polydispersity (2% or less) of aggregation in most cases did not make a difference.
  - Linked to Habel *et al.*, Acta D57, 254-259: On several occasions it has been possible to crystallise solutions with a Cp/Rh (polydispersion index) of 20-25%.

N.B.: contrary to the information in one of the postings in this discussion, Protein Solutions Inc no longer provides a message board on its website.

Filters for DLS measurements

(October 2001)

Which filter size do you normally use to prepare protein solutions for dynamic light scattering measurements? Is it really necessary to take 0.02 micrometer filters as recommended by ProteinSolutions and found in many papers, or are 0.2 or 0.1 micrometer filters also reliable?
In our lab some people made good experiences with 0.2 micron filters. In one case good DLS data (monodisperse solution) and excellent and reproducible crystals afterwards were obtained. However after filtering the same protein solution with 0.02 micron filters the protein was apparently away. At least no DLS signal could be detected any more. Normally this observation itself could be interpretated as an indication of aggregation, but the crystallization results do not support this idea.
So is it generally legitimate to swap to 0.2 micron filters if 0.02 micron filters catch away the protein? How are your experiences?

Summary from the enquirer - experiences from others:

- I was using 0.2 um spin-filters (I think they came from Eppendorf) when I ran out of the 0.02 um filters, and the results were alright. Normally, we are using the MicroFilters from Hampton without loss of protein. I would assume that if you loose protein, it's because the protein sticks to the membrane, in which case centrifugation might be a better way to get rid of aggregates in the first place.
- I have used 0.2 micron filters for DLS with no apparent problems.One of the original papers describing this method (Methods in Enzymology Vol. 276 p.157 by Ferre-D'Amare and Burley) use 200 Angstrom pore size ... Also, you should probably check your filtered sample another way to make sure that your protein is being trapped and nothing weirder is going on. Run a gel orUV-Vis spectra ??
- I regularly use a 0.1um filter to filter protein prior to DLS. I use the centrifugal filter from Millipore as this has no dead volume it doesn't waste your precious protein.
- I've found the smaller size filters (0.02) more difficult to use reliably. They seem to break or leak easily under modest pressure. In some cases, however, they seem to be necessary. I always try a larger filter first. With some practice you can recover most of the protein only slightly diluted and filter again with a finer mesh if need be.
- I normally start with 0.2 micron filters and work my way down if needed. As you indicated, sometimes you catch away all the protein which is indicative of lots of small aggregates. You will not be able to make good measurements from these whether you filter or not. You are wrong however to assume that aggregates and crystal gowth are incompatible!
- The folks at Prot.-Sol. say you can sometimes get away without filtering if you spin the sample first.
- We never filter our samples prior to DLS. Instead we centrifuge them in a benchtop centrifuge at maximum speed (the same as we would treat any sample prior to crystallisation). We were actually shown this by the Protein Solutions rep who visited our lab. We have had no problem analysing our samples after treating them this way.
- The golden question here is the size of your protein. If it is too big to go through 0.02 micron filters, you won't see a DLS signal. There is no protein in the solution. The protein is in the filter. Here I do not mean the unit-cell dimensions or contents of the asymmetric unit, but really truly monodisperse particle sizes in solution. That's where the answer is. Find out the true aggregation state of your protein, then you know what size filter cut-off to use.

RedHat7*

(June 2001)

I don't know how many others of you received a similar email from redhat, informing us that the compilers shipped with redhat7.1 were broken - so what's news?! Anyhow I have updated my Redhat _7.0_ system to use the new compiler and the news is .... better.

1. *compiling the suite with the standard options (including optimisation level O2) - the compiled code still does not work.*
2. *compiling the WHOLE suite (not just progs) with the compiler optimisation level O0 - the suite does work!!! Well it compiles and $CEXAM/unix/runnable/run-all works (apart from hbond for some strange reason).*

the way I did this was:

a. *up2date my Linux redhat 7* system to the new compilers*
b. *download and unpack the ccp4 package (remember to check the ccp4 problems pages for fixes to some programs)*
c. *edit and source ccp4.setup as usual*
d. *run configure --with-your-options linux as usual (shared lib not tested)*
e. *edit $CCP4/config.status - change as below:*
f. `FOPTIM="-O" COPTIM="-O"`
g.
h. *to*
i.
`FOPTIM="-O0" COPTIM="-O0"`
and re-run the config.status script (this in turn will re-run configure with the altered options)
j. *make and install as usual.*

I am assuming that the values the programs produce are sensible - I'm just pleased they didn't crash...
some system info:
```
ccp4h 2:57pm /runnable>45% rpm -q gcc glibc
gcc-2.96-85
glibc-2.2-12
```
If anybody has any other/similar/more experiences please let me know. In turn I will let Kevin know and maybe he can update his excellent summary page.

Summary from the enquirer:
I now have the following system:

1. redhat 7.0.
2. upgraded compiler from Redhat (using up2date) to gcc-2.96.85 (and g77 etc)
3. installed gcc(etc) 3.0 in /usr/local/bin - this was straight-forward

following this I had a clean distribution of CCP4 4.1.1. I configured as
.../configure --with-x linux
then edited $CCP4/config.status and changed
FC to /usr/local/bin/f77 and
CC to /usr/local/bin/gcc
ran config.status (NB no change to optimization level)
make and make install of the suite.....
ran the $CEXAM/unix/runnable/run-all script.....
AND IT ALL WORKED! again I hasten to add that the programs ran and didn't crash (so I'm assuming they gave sensible answeres).....
so in summary this probably isn't the best way to have your linux box set up but it does at least give a working compiled version of CCP4 for redhat 7* boxes.
I hope this helps. If the demand is really there I will make a web page (or Kevin might update his) with detailed instructions. Though my real current recomendation is - stick with RedHat 6.2.

A while later, Kevin adds to this:
Just to set the record straight, I would like to state here and now that the problems people have been having on RedHat 7.1 and 7.2, and other linux distros, are not the fault of the compilers shipped with 7.1 or 7.2. (either 2.96rh, or 3.0).
There are some assumptions in some ccp4 code, which are technically invalid with the F77 spec, but have been traditionally incorrectly implemented in the majority of Fortran

compilers. g77 is exceptional in interpreting the specifications correctly.
To compile ccp4 on Linux, simply add

-fno-automatic
to the XFFLAGS in all the makefiles.
The resulting code will give sensible results using the example refmac scripts. (Of course there may be smaller problems not picked up by this test, if so we now have a chance of finding them.)
To summarise: Redhat did good. We didn't.

Chemical discussions

Selenomethionine

Selenomethionine oxidation during RP-HPLC

(May 2001)

I'd like to purify a small disulphide-rich protein containing selenomethionine for MAD on a C8 reverse phase column. The buffers I normally use contain 0.1% TFA and Acetonitrile and are purged with helium, but the disulphide bonds in my protein don't allow me to use a reducing agent such as DTT or beta-mercaptoethanol. If anyone has had to deal with a similar case before, could they please let me know whether the selenomethionine became chemically modified during this purification step.

Summary from the enquirer:

- Why not do the structure with oxidised SeMet protein? Oxidised selenium gives a stronger MAD signal than reduced selenium. The problem is when you get mixed oxidation states, because then you don't get any absorption peak at all. That's why most people add DTT to their crystallisation buffers. You could try crystallising in the presence of an oxidising agent instead. If your problem is not that huge (*i.e.* relatively few seleniums [20-ish], relatively small protein [30 kDa-ish]), then you can probably even get away with using the remote wavelength, where there is no absorption peak and a relatively weak anomalous signal. But then you have to collect your data properly.
- If your protein's folded and disulphided, and you then add DTT, does it unfold? Because if not, it's often okay to reduce the SeMet only just before you freeze the crystal, because the SeMet oxidation is reversible. So, you purify and crystalize the thing without DTT, and in the minutes or hours before mounting, you add the DTT.
- Have you considered using the sulphur anomalous scattering to solve the phases? May be possible with sufficient resolution (*e.g.* using ACORN).

This summary raised a few issues:

1. Pardon me but do not get that. The Se has a absorption edge, no matter what. The stronger 'oxidized' edge is probably an electronic effect at the XANES creating an additional component seen as a white line feature (the peak above the edge jump level). A different chemical environment shifts the edges (up energy when oxidized, few eV) and may lead to a superposition and thus broadening of the edge. The no signal theory I do not understand?

2. I also do not quite understand what exact species the 'oxidized Selenium' or what oxidized Se-Met actually is. Does anyone have some insight into that?
3. I don't understand the rather cavalier attitude towards Se-Met oxidation. For one of the proteins I worked on, Apo A-I, oxidation of the (sulfur) methionines DRAMATICALLY alters the physicochemical and biological properties of the protein. The reason is obvious....Met is a hydrophobic side chain; oxidized Met is VERY hydrophilic (it is basically like the universal solvent DMSO) and has a strong, permanent dipole moment. I would imagine that the properties of most proteins would be altered if you suddenly stuck a (delta+)Se--O(delta-) bond in the middle of the hydrophobic core. I think care to prevent Se-Met oxidation is called for.

Answers to these questions:

1. Of course selenium still has an edge, but that reply stated "mixed oxidation states", "you don't get" and "peak", not "there is no" and "edge". In my experience the peak (the anomalous signal) is far more important for MAD than any dispersive signal (which is very small anyway if the edge is no good), because it allows you to solve the substructure. In fact, I go for the peak and remote, and consider the peak a luxury. Yes, the edge gets broadened, and so similarly the peak gets flattened, especially because the oxidised peak sits on top of the reduced edge. At least, its net effect changes that way, but that's what the anomalous signal is after all. Have a look at:
    1. AJ Sharff, E Koronakis, B Luisi and V.Koronakis, Acta Cryst. (2000). D56, 785-788, Oxidation of selenomethionine: some MADness in the method!
    2. Smith, J. L. & Thompson, A. (1998). Reactivity of selenomethionine - dents in the magic bullet? Structure 6, 815-819, through PubMed.
2. Mono-oxygenated (I believe):

```
3.          CH3
4.          |
5.          Se=O
6.          |
7.          CH2
8.          |
9.         ...
```

10. We recently phased a difficult protein with 1 Se per 90 residues using MAD. Origionally, we were unable to see a signal or find the Se positions. We subsequently pushed the Se to the oxidative state with the addition of HOOH. On our second trip to the synchrotron, we got a great adsorption spectrum and have now found all the seleniums, phased and are now model building. The oxidation of the selenium appears to be the key. Besides, oxidation should result in the removal of electrons from the Se.
11. We have collected MAD data on a fully (naturally) oxidised and fully reduced (using DTT) selenomethionine protein, and found that they are isomorphous except for the bound water molecules to the oxidised selenium atoms (Thomazeau et al., Acta Cryst.D57,1337-1340).
    So my idea about the question is that again, it depends on the protein.

BR's lecture on Se-Met and X-ray absorption

(May 2001)

The following was a posting in reaction to the previous discussion (about Selenomethionine oxidation during RP-HPLC). It is reproduced almost exactly as it was posted, with an addendum/erratum from the author at the end.

I got flamed for Borhani's message - don't worry I can take it - and received a few comments that make me wonder whether we use the same language here in terms of X-ray absorption. X-ray absorption is a lot less mystical than crystallization, so even at the risk of appearing redundant/boring/condescending you name it I shall briefly summarize for the more biologically inclined (admitting that I simplify as I feel it's permissible without being flat out wrong; if something is absolutely stupid or incomprehensible please tell me; textbook references at the end).

A bound electron can absorb a photon and leave its original energy level (orbit). The atomic level (quantum number n) it originates from is used to name the edge - K (1) L (2) M (3) etc. The lower (tighter bound) the level and the more protons in the nucleus (heavier the element), the higher the absorption edge energy.

Then the question is what happens to the electron. Assuming a free atom for now, absorbing at or above the binding energy the electron can take off into the vacuum and turn into a photoelectron (more about condensed state below), or at slightly lower energies, it can jump into unoccupied higher levels (states) of the atom (if the electron kicks out another electron from a higher occupied level, we have a secondary Auger electron but due to their low energy - except for line broadening - the Auger processes are of no relevance for us here).

The superposition of all the discrete possible lower energy resonance transitions in the series plus the phototransitions at the series limit create each absorption edge. The sum (integration) of the closely spaced and life-time broadened transitions at the series limit gives an arctangent curve (sigmoid shape) for the basic absorption edge. The sharp, saw-tooth curve in theoretical absorption cross section calculations results from assuming sharp photoelectric transitions. The most prevalent code I know and use to calculate absorption coefficients/edge energies is Don Cromer's FPRIME (note from mgwt: the best link I could find is a PDF file: GSASmanual.pdf).

In case of high transition probabilities, some of the pre-edge resonance transitions can be rather high, and give rise to stronger absorption. These pre-edge features are also called white lines, because some of the old dudes (like those who wrote all these nice F-66 CCP4 programs for you) used film to record absorption: Less X-rays on film due to absorption in the sample means less blackening on the negative (*i.e.*, a white line at that energy). White line resoncances obey dipole selection rules, and their intensity depends on transition probabilities and initial and empty state density. K-edges have weaker white lines (s->np transitions) as do L1 edges (n=2,l=0,j=1/2 2s -> nd, n>2) which have 'K- or S-character' due to l=0 compared to l=1 for L2 (n=2,l=1, j=1/2) and L3 (j=3/2) edges.

The L3 edge is at the lowest energy of the L series and twice as high as L2,1 due to the transition from the 4 2p3/2 states, at L2 (few keV higher energy) there is usually also less intensity from the ring (above critcal energy).

It appears that the white line features are what some call 'peak', so when they talk about 'disappearing peak' they may mean a smaller white line, not the whole edge disappearing. Btw, that white line region at the low energy of the edge is called the XANES (Xray Absorption Near Edge Structure).

Now to finally sort XAS out, we need to consider condensed matter. A bit more delicate, but it will become clearer (harharhar). First, on its way out of the atom, an above-edge energy photoelectron can bounce off the neighbouring atoms. If there is a distinct near range order - like in a let's say octahedral environment - the resonance absorption cross section oscillates in a decaying way with a period distinct (reciprocal, as you guessed) to the distances in the coordination shell geometry in the environment. The amplitude envelope of these periodically extending EXAFS wiggles tells you about the nature of neighbouring atoms - the heavier the more 'wavy' the envelope becomes.

So, if you have a rapidly decaying EXAFS (Extended Xray Absorption Fine Structure) you know that you have light atoms and/or inhomogenous environment around your anomalous atoms - which does not mean much: Unfortunately, detailed EXAFS analysis requires much better scans than we usually do and the difference between a Se atom in solvent and in the protein environment is not all that big. Well-defined metals in active sites (plastocyanine, cytochrome c oxidase, laccase etc) can have in fact an interpretable EXAFS. It naturally also kinda works in solid matter, but deconvolution is occasionally overdone (30 data points 25 parameters - sounds familar to the low res victims, doesn't it?).

On top of this, if in a chemical environment outer electrons get stripped (oxidation, delocalization etc) the remaining electrons feel more of the nuclear charge thus more energy required thus upshift of the edge features (someone got confused about that apparently). Shifts range in few to a few 10 eVs, and you nearly always need a reference spectrum to determine absolute values (think monocromator slew for example - which is one reason why it is not a bad idea to move the crystal (energy) from the same side to the peak as you did in your scan).

The condensed state environment also allows due to symmetry violations (think Jahn-Teller) additional transitions in the pre-edge region that where verboten before, plus allows additional band levels to become occupied by photoelectrons. This means that larger white line features often appear. The same holds for any new bound or localized states, like in oxides, which become now available compared to the free atom case we described in the beginning.

All of the above to varying degree is the reason why a) the oxidized Se-Met spectrum is upshifted, b) the white line in the solid Selenate sample shown in the Structure6:815 paper is so huge, and less high for oxidized Se-Met in protein.

Now let us consider what happens in an inhomogenous environment:

First, each Se that is present will absorb. There is no absorption quenching or any funky similar stuff. If it's there, it will contribute to signal. Chemically different species will add, and we will obtain a sum of the partial spectra. This means that the white line features can become less sharp, as will the whole edge. But: After the edge, the total signal will be the same - *i.e.*, if your Se's remain in periodic positions - oxidized or not - up-edge (remote) anomalous data can give a decent signal/map (but less white line - or 'peak' contributions to f'' of perhaps a few tenth to a few e-). For the f' (inflection wavelength), we have a worse scenario: the slope of the edge and/or peak becomes flatter - thus the derivative

(equivalent to the Kramers-Kronig transform) is much smaller and your dispersive gain from sitting on f'max (inflection point) suffers drastically. If the anomalous scatteres are all over the place then signal but of course no map. Backsoaking of heavy metals derivatives for anomalous data collection is thus advisible.

Based on above, I cannot rationally explain how one can have no signal (I mean no edge, not only no 'peak', or white line), then oxidize the same material, and get an absorption scan. Sounds like some trans-substantiation. Most likely I did not understand the story right.

For more details on the L-edge and white line superposition stuff you can glance over the mini-intro (II, III) in that Physical Refuse article: http://www-structure.llnl.gov/pdf/moment_collapse.pdf.

Details: Agarwal, X-ray spectroscopy, Springer, chapter 7

Another interesting point: If one measures above the edge past the white lines, very good monochromaticity (low bandwidth) actually is not necessary and rather a waste of beam. A beam with a bandwidth not exceeding the mosaic spread of the crystal would allow a really fast (or good signal/noise) data collection about 100 eV above the edge. The anomalous signal is nothing to sneer about there and the gain for SAS in data quality could be tremendous. Any thoughts on that from the SAS gang? I mean going from 1 to 10 eV bandwith does contribute to spot size (below) but ~10 times the bang should do something for the data!

Problem to be solved: The beam fans out to about ~30 mrad at 0.1% bandwidth at 1.0Å after the monochromator and needs to be refocussed: 0.1% 12.35eV bandwidth at 12.35 keV (1.000Å) 0.9005 to 1.0005Å on Si 111 (d 001 = 5.43), d(111)= 3.14 (Pi, funny, isn't it?) lambda/2d=sintheta I get delta theta of .923 deg = 1.8 deg 2theta = 30 mrad which is ugly.

Less excessively, let's say at 0.8 deg or so this might actually be useful. Lots of partials though.

Any thoughts on feasibility? Knowledge of the absolute of the fs is actually unimportant here btw.

Addendum/erratum:

Of course I stuck my foot into my mouth on this one - the calculation of the spread is wrong 1.0 - 0.0005 is NOT .9005 dah...I had already a bad feeling - thx to Pierre for actually reading my blurb and finding that mistake. Consider that even the Cu natural line width at 8keV is 2.6 eV which has no practical effect on point spread.....

Bart pointed out that 3rd generation sources fry the crystal dead anyhow so why bother - that is true, I was admittedly more thinking along maximizing weaker sources like small Compton sources (a electron bunch is 'wiggled' by a laser) which, using the broader bandwidth, may begin to compare well to synchrotron sources.

Correction of the same calculation of 0.1% bandwidth spread at 1Å (1.0005 to 0.9995 <-!) leads to 0.02 deg (0.4 mrad) in 2theta which is negligible as it should be. So even wider bandwidth ranges would be possible for flux gain in the scenario I described. Problem solved.

Selenomet from O and REFMAC5

(June 2001)

I am using Refmac5 through the CCP4I-4.1.1 interface and am wondering whether selenomethionine (MSE) is being recognized properly on coordinate file input. To start, I mutated MET to MSE in 'O' and here is representative output of the relevant part of the coordinate file from 'O':

```
ATOM    320  N   MSE X  41      29.824  31.488  35.626  1.00 11.19   7
ATOM    321  CA  MSE X  41      29.652  32.610  36.538  1.00 11.64   6
ATOM    322  CB  MSE X  41      28.225  33.094  36.510  1.00 12.01   6
ATOM    323  CG  MSE X  41      27.852  33.686  35.170  1.00 14.13   6
ATOM    324  SE  MSE X  41      28.681  35.384  34.700  1.00 20.00  34
ATOM    325  CE  MSE X  41      27.259  36.407  35.447  1.00 17.00   6
ATOM    326  C   MSE X  41      30.038  32.246  37.955  1.00 11.06   6
ATOM    327  O   MSE X  41      30.707  33.006  38.648  1.00 10.75   8
```

Note "34" for SE in the last column of the fifth row. Using the above coordinate file from 'O' as input, here is the resulting relevant part of the output file after refinement with Refmac5:

```
ATOM    629  N   MSE X  41      29.826  31.486  35.628  1.00
11.15           N
ATOM    631  CA  MSE X  41      29.653  32.611  36.538  1.00
11.57           C
ATOM    633  CB  MSE X  41      28.227  33.091  36.512  1.00
11.90           C
ATOM    636  CG  MSE X  41      27.851  33.685  35.175  1.00
14.00           C
ATOM    639  SE  MSE X  41      28.681  35.387  34.704  1.00
5.03          S
ATOM    640  CE  MSE X  41      27.244  36.414  35.435  1.00
16.74           C
ATOM    644  C   MSE X  41      30.036  32.247  37.952  1.00
10.98           C
ATOM    645  O   MSE X  41      30.706  33.002  38.641  1.00
10.84           O
```

From the above, Refmac5 appears to be interpreting the SE atom (34 electrons) as sulfur (16 electrons) (I guess also giving the unexpectedly low B value for Se).

Summary from the enquirer: Refmac must read "SE" starting at either position 13 or position 77 in an ATOM (or HETATM) record. On the other hand, 'O' outputs coordinate files with SE aligned with the standard amino acid atom identifiers (starting in position 14) and atomic number for the element in position 69 or 70 (for one digit or two-digit atomic numbers, respectively). So, the first ATOM record below will not be read properly by Refmac5; editing this line to either of the formats in the second and third lines will work:

```
ATOM    198  SE  MSE X  26      16.208  48.882  45.142  1.00 13.97  16
ATOM    198 SE   MSE X  26      16.208  48.882  45.142  1.00 13.97  16
ATOM    198  SE  MSE X  26      16.208  48.882  45.142  1.00 13.97          SE
```

Glycerol - bad or good?

(May 2001)

I have a not crystallographic computing related but still interesting question. As often in protein crystallization, firm and validated information is rare and thus in this case I am happy to solicit also opinion and anecdotal evidence. Glycerol is used to protect proteins while being stored frozen. This is a particular issue for any high-throughput operations,

where the protein cannot be processed immediately and needs to be stored in aliquots until machine time becomes available. Now, the question is, how high a price will you have to pay later in crystallization success rate if you do not dialyze the glycerol out? I.e, what is the overall statistical chance that it is harmful vs. not? In particular, has reduced diffraction quality (vs.non-gycerol) been observed? I clearly understand that some proteins do crystallize fine with glycerol as additive, and we have it also in CRYSTOOL, but as a principal component in the protein stock, at lets say 10%, what's the effect? Does anybody have hard numbers (or some statistics) on that or at least more than single case evidence for the one or the other? Electronic web research in Medline and inspec did not provide a lead. Manual search in J. Crystal Growth (1889-90) where we hoped to find an article presented at the first ICCBM conference in 88 was negative. Please let us know if you can help with any information, references or leads.

Summary from the enquirer: Probably more bad than good. If you don't need it, don't have it in. If you need it for stability, don't worry (actually, you are free to worry). There are no crystallization data on possible substitutes for glycerol either. Glycerol may be useful as a retardant when things grow too fast (problem also often seen in nanodrops?) Snap freezing sounds interesting. Anyone else use that? It would be a good idea to use the robots for a systematic study. Ok, I will.

For full discussion, see http://www-structure.llnl.gov/copc.htm.

Monovalent cations

(April 2001)

I am seeking guidance on interpreting e-density that appears to arise from monovalent cations -- how to differentiate, e.g., Na+, K+, NH4+. Any relevant references and/or programs would be appreciated.

Summary from the enquirer: It goes without saying that resolution of data is critical in differentiating possibilites (my data are to 1.5Å).

1. Different number of electrons for different metals will obviously give different local electron densities (possibly can also make use of differences in anomalous signals for metal sites).
2. Look at USF program XPAND -- "w ater scrutinizer" option -- which checks for possibilities other than HOH. I think using the approach of M Nayal & E Di Cera, JMB 256, 228-234 (1996) -- this paper describes the calculation of valence for metal-to-ligand interactions [the valence calculation quantitates distances for all metal-to-ligand interactions for a putative metal site, and relates these to expected distances and coordination numbers for potential ions; e.g., (K+) should have longer metal-to-ligand bonds, on average, than (Na+)-to-ligand]. The Di Cera valence analysis has been developed by the author into the WASP program (WAter Screening Program) -- see http://www.biochem.wustl.edu/~enrico/wasp.htm or contact enrico@caesar.wustl.edu
3. Valence bond calculation methods are being implemented by George Sheldrick (I'm not sure if part of ShelX) -- suggested to contact Dr. Sheldrick directly to inquire.
4. Some general references were also suggested:
    - SJ Cooper .. WN Hunter (1996) Structure 4: 1303-1315 The crystal structure of a class II fructose-1,6-bisphosphate aldolase shows a novel binuclear

metal-binding site embedded in a familiar fold [includes table comparing Metal-N, Metal-O distances [Metal = K(+), Zn(++)] from CCDC]
    - CA Bonagura .. TL Poulos (1999) Biochemistry 38: 5538-5545 The effects of an engineered cation site on the structure, activity, and EPR properties of cytochrome c peroxidase
    - S Rhee .. DR Davies (1996) Biochemistry 35: 4211-4221 Exchange of K+ or Cs+ for Na+ induces long-range changes in the thre-dimensional structure of the tryptophan synthase a2b2 complex [tabulates Na(+)-O, K(+)-O and Cs(+)-O distances in tryp synthase]
5. Bond-valence calculations can be done using a different equation than that used in the above-mentioned Di Cera paper:
    - Brese, N. E. & O'Keeffe, M. 1991. Acta Cryst. B47: 192-197. Bond-Valence Parameters for Solids.
    - Carugo, O., Djinovic, K. & Rizzi, M. 1993. Comparison of the Co-ordinative Behaviour of Calcium (II) and Magnesium (II) from Crystallographic Data. J. Chem. Soc. Dalton Trans. 2127-2135.

    An example using this alternate equation was kindly provided:

```
vij=exp[-(dij-Rij)/b]

vij-bond valence for bond between i and j
dij-bond length between i and j
Rij-bond-valence parameter
[K+ 2.13 for O, 1.99 for F, 2.52 for Cl
Na+ 1.80 for O, 1.677 for F, 2.15 for Cl]
b-"universal" constant b=0.37

V=vij(1)+vij(2)+...

V-valence of the metal centre


dij (O-Me)      vij (Na+)
2.78            0.07
2.47            0.16
2.30            0.26
2.58            0.12
2.30            0.26
2.25            0.30

V=1.17 (in my case I was sure that it cannot
be Ca2+ - I checked anomalous signal)
```

Atoms used for anomalous dispersion (a survey)

(April 2001)

We have recently solved the structure of the PDZ1 domain of Na+/H+ exchanger regulatory factor using the dispersive signal from the LIII edge of Mercury (see Webster at al. (2001) Acta Cryst D57, 714-716 and our J.Mol.Biol.308, 963-973 (2001) paper). We were unable to obtain satisfactory expression of our protein from selenomethionine auxotrophs and only obtained a single mercury derivative in spite of an extensive heavy atom screen from which the SIR phases were insufficient to solve the structure. In the end then, we decided to try a MAD experiment using our lone Mercury derivative and obtained a beautiful anomalous signal at three different wavelengths on beamline F2

at CHESS. An analysis of our data with SOLVE yielded excellent phases and a model consisting of over 80% of the protein was built by ARP/WARP in the first electron density map calculated with the new phases.

I was wondering whether anybody had done a survey of elements other than Selenium that have been successfully used for structure determination with MAD, since it seems that a lot of time can be saved if even a single, suitable heavy-atom derivative of a protein can be obtained for such an experiment. I know that there are plenty of tables of wavelengths and dispersive differences for different elements, but I would be very interested to see if anybody had compiled statistics for which elements had actually worked for MAD structure determinations. Such a survey might beneficially bias our choice of which heavy-atoms are worth screening first, especially if the biological labelling of proteins is not an option due to time constraints or technical problems at the level of expression etc.

Summary from the enquirer: It seems that there hasn't really been a comprehensive review of this for some time now. I was pointed to an article in Synchrotron Radiation News Vol 8 No 3, pp 13-18 (1995) written by Craig Ogata and Wayne Hendrickson, and a later article from 1999 also by Wayne Hendrickson (J. Synchrotron Rad. 6, 845-851). People at Daresbury have found Xenon at high pressure to be an excellent choice, their results for this work on the structure of crustacyanin is Cianci et al. Acta D.57,1219-1229. Note that sulphur (sulfur if you celebrate July 4th) has a useful anomalous signal at around 2.0Å and work using this method will be published in a forthcoming paper. It is commented that 3 wavelength experiments are often unnecessary and that the anomalous signal from a single atom of *e.g.* iron or zinc per protein molecule can be enough for structure determination with MAD. Also advocated is the use of elements that have a significant anomalous signal close to the copper K-alpha wavelength and therefore do not require a trip to the synchrotron. Even mercury has 7.7 anomalous electrons at 1.54Å and it was suggested that we might possibly have been able to solve our PDZ structure in-house. A protein using Xenon at 1.54Å, with 4 atoms per 47 kDa molecule (another plug for Xe there), has just been solved.

A whole slew of elements (Fe, Co, Zn, Se, Br, Rb, Ta, W, Re, Os, Ir, Pt, Au, Hg, Tl, Pb, U) was listed, with which success has been had on the beamline I9ID at the APS (Argonne Il.). It was pointed out that Se-Met has become a very popular choice due to the very high success rate that it has for phasing. The number of Se atoms generally increases with the size of the protein and there is no disturbance of the crystals by soaking as is required for traditional heavy-atom labeling. My own experience with Se-Met has led me to ...

WEBSTER'S LAWS OF METHIONINE DISTRIBUTION
"The probability of a methionine residue occurring in a protein is inversely proportional to my desire to solve the structure of that protein" "The probability of finding a methionine residue at any given point in my protein is directly proportional to the conformational flexibility of my protein at that point"
Please don't flame me or bombard me with your "selenomethionine has changed my life" stories, I know it works very well, but I just haven't been very lucky with it so far! A third article was mentioned: C. Ogata (1998) "MAD phasing grows up" Nat Struct Biol Synchrotron suppl, 638-640. Somebody mentioned they did a survey of the elements used for MAD a few years ago (but did you publish the survey?) and also cited many of the elements in the list above. Another made the excellent suggestion of having specific phasing records included in the PDB database format. This would make the compilation of the kind of statistics that I was after, effectively automatic, since users would be able to compile their own surveys directly from the database itself. How about it RCSB?

It was pointed out that you can do MAD with any element that has an absorption edge within the energy range of the most commonly used beamlines (7000 - 15000 eV) and that L-edges like the one that we used in our PDZ structure determination, often give better results than K-edges. Along with mercury, gold and lead are recommended as good candidates. Reservations are expressed about using platinum which tends to yield many poorly occupied sites and a resulting poor signal. Also recommended: Lanthanides for their excellent signal with the caveat that they may be harder to get to bind to your protein (apparently they substitute for Ca very well in Ca binding proteins). Tantalum bromide has been used for very large cells (didn't they use this for the ribosome?). And again the recommendation for trying high pressure derivatization using Xe and NaBr.

Beryllium Fluoride-ADP

(September 2001)

I'm looking to purchase Beryllium fluoride to use in combination with ADP as a transition state analogue of ATP. Sofar my searches in catalogues (on-line and on plain paper, Sigma, Aldrich, Fluka) yielded nothing. Does anyone have experience in these matters? Do we have to make it ourselves?

Summary from the enquirer:

1. Do you really want to work with this? It is VERY Toxic! Enquirer: Well, I must say I'll think it over again after all the warnings. Maybe Aluminium Fluoride is a good alternative.
2. BeF2 can be purchased. Three companies have been mentioned to me, Alfa-Aesar (Germany), Interchim (France) and Strem (US).
3. BeF2 more often is made by adding proper quantities of BeCl2 and KF (or NaF). The extra KCl (or NaCl) in the drop should not worry us! Enquirer: OK, fair enough. At least I won't have at least 5 grams of BeF2 left in our chemical storage after this experiment!

Mercury Phenyl Glyoxal

(October 2001)

We are currently investigating the possibilities of covalently bonding heavy-atoms to specific residues types using modified reagents, for use in structure determination through MIR, etc...
We have a review citing the use of mercury phenyl glyoxal as an arginine specific reagent, and have found the recipe for it on the web, but our collaborators reckon that the reaction conditions for it are nowhere near strong enough to force mercury onto a phenyl ring... They have used this recipe, and using NMR, have discovered that all you get back at the end is the phenyl glyoxal that you started with...
Other than Don Wiley's work (which did use Phenyl glyoxal, but it did not bind to the Arg residues), has anybody...

1. *got a decent prep for Hg-phenyl glyoxal that they KNOW works*
2. *actually solved a structure using Hg-phenyl glyoxal AND seen it bound to arginines*

Summary from the enquirer: It seems to be the general consensus that the prep for mercury phenyl glyoxal on the Metazoa.com website is wrong. Other alternatives have been suggested and we'll let you know if they work when it happens... The mercury phenyl glyoxal as reported in Wilson et al Nature, 289, pp386, 1981, was in fact not Mercury phenyl glyoxal, and the heavy atom sites bound were due to residual mercury in the compound. The two structures that claim the use of Mercury phenyl glyoxal are Haemagglutinin (2HMG)(from Don Wiley, reference above) and Galactose binding protein 2(GBP). None of the mercury sites are anywhere near an Arg. It does seem like mercury phenyl glyoxal is a bit of a myth.

Various

XYZ-limits and real space asymmetric units

(November 2000 and January 2001)

I use :

```
.....etc
#-------------------- crystallographic project data ----------------------
# the unit cell dimensions
set cell = ( 140.080  140.080  271.630  90.00  90.00  90.00 )
# spcgrp
set spacegroup = ( p43212 )
# spcgrp no. in symm
set symm = ( 96 )
# FFTSYMMETRY
set sfsg = ( p43212 )
# fftgrid GRID
# set grid = ( SAMPLE 3 ) does not work for SFALL
set grid = ( 128 128 512 ) <--- note : nx=2n, ny=nx, nz=8n (n=64) as per FFT
instructions
# the asymm unit box for SFALL/FFT
# set xyzlim = (ASU) does not work for EXTEND
# set xyzlim = ( 0 1 0 1 0 0.25 ) for p212121
# set xyzlim = ( 0 1  0 1  0 0.33333 ) for p31
# set xyzlim = ( 0 1  0 1  0 0.166667 ) for p61
set xyzlim =   ( 0 1  0 1  0 0.125 ) <--- per instructions and tables
# -----------------------------------------------------------
....etc
```

REFMAC, EXTEND, and FFT run fine. ARP finally gets mad at me as follows:

```
 Map limits Z   0  64 <== This is incorrect   Recommended   0 256
```

What is different in ARP compared to the other programs that take my grid input? Do I need xyzlimits in (0 128 0 128 0 64) format and not fractional? Do I need the FULL cell in the xyzlimits? But this is not P1 as in general cases?

Quick-and-dirty answer:
Not that I really know how it works, but here's an absolutely filthy fix that seems to work, at least for the spacegroups which I tried. If you specify GRID SAMPLE in FFT, then there's a line in the log file that says "Map limits in grid points on xyz" and some numbers. Add 1 (one) to those numbers and give them to EXTEND as grid, then ARP is happy.

Slightly more sophisticated answer:
For P43212 (96) you'll need the following asymmetric units

```
FFT    0. 1.  0. 1.  0. 0.125
SFALL  0. 1.  0. 1.  0. 0.125
ARP    0. 0.5 0. 0.5 0. 0.5
```

There are consistent inconsistencies with real space asymmetric units between the various programs.

Another solution: The easiest is simply to define the AU limits in fractional coords in MAPMASK ... just take care in trigonal and cubic sg's to use 0.334 instead of 0.333333 and 0.0834 instead of 0.08333333. I actually think that only ARP is 'inconsistent', but it's very polite on telling you what it really needs as AU limits.

Another response: ARP is not alone: I can think of other 'inconsistencies' too. Compare the asymmetric units FFT is using with the ones SFALL requires:

Spacegroup	FFT			SFALL		
	X	Y	Z	X	Y	Z
P21212	0. 1.	0. 0.25	0. 1.	0. 1.	0. 1.	0. 1.
C2221	0. 0.5	0. 0.25	0. 1.	0. 1.	0. 1.	0. 1.
C222	0. 0.5	0. 0.25	0. 1.	0. 1.	0. 1.	0. 1.
F222	0. 0.5	0. 0.25	0. 1.	0. 1.	0. 1.	0. 0.25
I222	0. 0.5	0. 0.25	0. 1.	0. 1.	0. 1.	0. 1.
P4212	0. 1.	0. 0.25	0. 1.	0. 1.	0. 1.	0. 1.
P4122	0. 1.	0. 1.	0. 1.	0. 1.	0. 1.	0. 0.125
P4322	0. 1.	0. 1.	0. 1.	0. 1.	0. 1.	0. 0.125
I422	0. 0.5	0. 0.25	0. 1.	0. 1.	0. 1.	0. 1.
P3	0. 1.	0. 1.	0. 1.	0. 0.67	0. 0.67	0. 1.
P622	0. 1.	0. 1.	0. 1.	0. 0.67	0. 0.67	0. 1.
P6322	0. 1.	0. 1.	0. 1.	0. 0.67	0. 0.67	0. 1.
F23	0. 0.5	0. 0.25	0. 1.	0. 1.	0. 1.	0. 0.25
I23	0. 0.5	0. 0.25	0. 1.	0. 1.	0. 1.	0. 1.
F432	0. 0.5	0. 0.25	0. 1.	0. 1.	0. 1.	0. 0.25
F4132	0. 0.5	0. 0.25	0. 1.	0. 1.	0. 1.	0. 0.25
I432	0. 0.5	0. 0.25	0. 1.	0. 1.	0. 1.	0. 1.

(I'm not totally sure about these different requirements: some of them are probably due to the differences in spacegroup-specific routines FFT and SFALL use. But it gives you an idea.)
So sometimes you need to put a MAPMASK/EXTEND step between. Anyway, once you know the various asymmetric unit definitions you can easily extend your maps.
Then an extensive list for ARP, just to be complete:
If you use these limits in MAPMASK before ARP/wARP it should work.

```
1 : 0 1 0 1 0 1
2 : 0 1 0 1 0 0.5
3 : 0 1 0 1 0 0.5
4 : 0 1 0 0.5 0 1
5 : 0 0.5 0 1 0 0.5
16 : 0 0.5 0 1 0 0.5
17 : 0 0.5 0 1 0 0.5
18 : 0 1 0 0.25 0 1
19 : 0 1 0 1 0 0.25
20 : 0 0.5 0 0.5 0 0.5
21 : 0 0.25 0 0.5 0 1
22 : 0 1 0 0.25 0 0.25
```

```
23 : 0 0.5 0 0.5 0 0.5
24 : 0 0.5 0 0.5 0 0.5
75 : 0 0.5 0 0.5 0 1
76 : 0 0.5 0 0.5 0 1
77 : 0 0.5 0 0.5 0 1
78 : 0 0.5 0 0.5 0 1
79 : 0 0.5 0 0.5 0 0.5
80 : 0 1 0 0.5 0 0.25
89 : 0 0.5 0 0.5 0 0.5
90 : 0 0.5 0 0.5 0 0.5
91 : 0 1 0 1 0 0.125
92 : 0 0.5 0 0.5 0 0.5
93 : 0 1 0 0.5 0 0.25
94 : 0 0.5 0 0.5 0 0.5
95 : 0 1 0 1 0 0.125
96 : 0 0.5 0 0.5 0 0.5
97 : 0 0.5 0 0.5 0 0.25
98 : 0 1 0 0.25 0 0.25
143 : 0 1 0 1 0 1
144 : 0 1 0 1 0 0.334
145 : 0 1 0 1 0 0.334
146 : 0 0.334 0 0.334 0 1
149 : 0 1 0 1 0 0.5
150 : 0 1 0 1 0 0.5
151 : 0 1 0 1 0 0.167
152 : 0 1 0 1 0 0.167
153 : 0 1 0 1 0 0.167
154 : 0 1 0 1 0 0.167
155 : 0 0.334 0 0.334 0 0.5
168 : 0 1 0 0.5 0 1
169 : 0 1 0 1 0 0.167
170 : 0 1 0 1 0 0.167
171 : 0 1 0 0.5 0 0.334
172 : 0 1 0 0.5 0 0.334
173 : 0 1 0 1 0 0.5
177 : 0 1 0 0.5 0 0.5
178 : 0 1 0 1 0 0.0834
179 : 0 1 0 1 0 0.0834
180 : 0 1 0 0.5 0 0.167
181 : 0 1 0 0.5 0 0.167
182 : 0 1 0 1 0 0.25
195 : 0 1 0 1 0 0.5
196 : 0 1 0 0.5 0 0.5
197 : 0 0.5 0 0.5 0 1
198 : 0 0.5 0 0.5 0 1
199 : 0 0.5 0 0.5 0 1
207 : 0 0.5 0 0.5 0 1
208 : 0 1 0 0.5 0 0.5
209 : 0 0.5 0 0.5 0 0.5
210 : 0 0.5 0 0.75 0 0.667
211 : 0 0.25 0 0.75 0 0.667
212 : 0 1 0 1 0 1
213 : 0 1 0 1 0 1
214 : 0 0.667 0 0.75 0 1
```

CCP4i comes to the rescue: Conventions - who needs them! As you have noticed the ARP asymm units and the CCP4 ones are not always the same. Both are in fact correct, but different. The GUI script adds an extra stage to move the P43212 CCP4 map to the ARP map. Or you could use FFTBIG XYZ Y X Z to get the whole P1 map then trim it back to ARP requirements.

Bring on libraries: The ASU should be determined for most purposes by a library call to symlib.f:SETLIM. (MAPMASK and many, but not all, other programs do this). FFT should not be used without a compelling reason. FFTBIG should be consistent with MAPMASK, otherwise it needs fixing. SFALL is more difficult. There are two main reasons for inconsistencies:

3. The libraries were written when maps had to be ordered in such a way as to make an out-of-core fft practical. There are other problems - the ASU contains duplicate grid points on special sections for most spacegroups, and whole duplicate volumes in some.
4. Some programs don't use the libraries at all.

The solution to (1) is maps which understand their own symmetry and better ASU definitions. The solution to (2) is to shoot any programmer who doesn't use them. These things are in hand for future software, through the provision of new libraries and superior firepower. On the whole existing software will, I'm afraid, have to remain hidden under a GUI.

Added to this: For information, the documentation for the CCP4 symmetry s/r library SYMLIB ($CHTML/symlib.html on your local system) has an appendix with the asu limits for both real and reciprocal space (go and have a look!) It might be useful if the list of Arp asu's was carried in the Arp documentation?
The real space limits in the SYMLIB document are those which will be used by any program which calls the CCP4 library routine SETLIM (reciprocal space limits are from PGNLAU) - unfortunately not all CCP4 programs use this routine, which is where the inconsistencies start to arise within the suite.
I'm all for Kevin's methods for dealing with programmers who don't use libraries (btw I hope his reference to "superior firepower" actually meant bigger and better computers...). In the meantime it would be useful for us if people could highlight the specific inconsistencies so that we could start to address the problem at source.

Another question a few months later: *With reference to the list(s) as described/tabulated above: It appears that at least for #20 SFALL does well using the FFT grid and does not need P1 expansion - am I interpreting this list wrong? Is there yet a final, authoritative compilation of settings somewhere?*

```
#  Spacegroup      FFT                              SFALL ARP
#                X      Y       Z           X       Y       Z          X      Y
Z
#    C2221    0. 0.5 0. 0.25 0. 1.       0. 1.   0. 1.   0. 1.      0. 0.5 0.
0.5 0. 0.5
```

Here is the extract for SFALL limits from the documentation.
X1 and X2 are always set to 0 to NX1-1; 0 to NX2-1;
BUT by far the best way to run sfall is to precede it with MAPMASK to generate a "whole cell" map and use the inverse FFT in P1. The other cells are archaic remnants of the days when we were seriously short of memory and it mad sense to work with the smallest possible map volume.

```
MAPMASK mapin asymm_unit.map mapout whole-cell.map
XYZLIM 0 0.999 0.999 0 0.999
AXIS Z X Y
END
```

then

```
sfall hklin asymm_unit.mtz hklout asymm_unit+FC.mtz mapin whole-cell.map
SFSG P1
MODE SFCALC MAPIN HKLIN
LABI FP=... SIGFP=...
LABO FC=FC_map PHIC=PHIC_map
END
```

sfall checks symmetry from the mtz file and outputs a list of h k l FP SIGFP ... FC PHIC for the asymm unit only.
If you have no HKLIN you must also give SYMM and RESO but otherwise resist the temptation to add any key words - the programs are meant to be able to sort themselves out.
Here is the extract from the sfall document:
Limits for axes for the various space groups (these are the same as those used as defaults in FFT):

- In space group P1, P21 and P21212a, 'b' is taken as the unique axis.
- In space group P21212a, 1/4 is subtracted from the X and Y values of the equivalent positions given in International Tables.

	X1	X2	X3	Range of X3	Axis order
P1	Z	X	Y	0 to Y	Z X Y
P21	Z	X	Y	0 to Y/2-1	Z X Y
P21212a	Z	X	Y	0 to Y/4	Z X Y
P212121	X	Y	Z	0 to Z/4	Y X Z
P4122	X	Y	Z	0 to Z/8	Y X Z
P41212	X	Y	Z	0 to Z/8	Y X Z
P4322	X	Y	Z	0 to Z/8	Y X Z
P43212	X	Y	Z	0 to Z/8	Y X Z
P31	X	Y	Z	0 to Z/3-1	Y X Z
P32	X	Y	Z	0 to Z/3-1	Y X Z
P3	X	Y	Z	0 to Z-1	Y X Z
R3	X	Y	Z	0 to Z/3-1	Y X Z
P3121	X	Y	Z	0 to Z/6	Y X Z
P3221	X	Y	Z	0 to Z/6	Y X Z
P61	X	Y	Z	0 to Z/6-1	Y X Z
P65	X	Y	Z	0 to Z/6-1	Y X Z

Limits for arp are embodied in this code: I guess someone should tabulate it nicely.

```
      PARAMETER (ROUND=0.00001, ROUND2=2.0*ROUND)
      PARAMETER (ONE=1.0+ROUND,HALF=0.5+ROUND,THRD=1./3.+ROUND,
     $     TWTD=2./3.+ROUND,SIXT=1./6.+ROUND,THRQ=0.75+ROUND,
     $     QUAR=0.25+ROUND,EIGH=0.125+ROUND,TWLT=1./12.+ROUND)
      PARAMETER (ONEL=ONE-ROUND2,HALFL=HALF-ROUND2,THRDL=THRD-ROUND2,
     $     SIXTL=SIXT-ROUND2,QUARL=QUAR-ROUND2)

C  asulim contains maximum limit on x,y,z: the box is always assumed to
C     start at 0,0,0
C
C  Space group numbers
      DATA NSPGRP/
     $   1,   2,   3,    4,   5,  10,  16,   17,  18,1018,  19,   20,
     $  21,  22,  23,   24,  47,  65,  69,   71,  75,  76,  77,   78,
     $  79,  80,  83,   87,  89,  90,  91,   92,  93,  94,  95,   96,
     $  97,  98, 123,  139, 143, 144, 145,  146, 147, 148, 149,  150,
     $ 151, 152, 153,  154, 155, 162, 164,  166, 168, 169, 170,  171,
     $ 172, 173, 175,  177, 178, 179, 180,  181, 182, 191, 195,  196,
```

```
     $ 197, 198, 199,  200, 202, 204, 207,  208, 209, 210, 211,  212,
     $ 213, 214, 221,  225, 229/
C
      DATA ((ASULIM(II,JJ),II=1,3),JJ=1,73)/
C        1:  P1          2:  P-1         3:  P2            4:  P21
     $  ONE,ONE,ONE,   ONE,ONE,HALF,   ONE,ONE,HALF,   ONE,HALF,ONE,
CCP4 $ ONEL,ONEL,ONEL, ONEL,HALF,ONEL, HALF,ONEL,ONEL, ONEL,HALFL,ONEL,
C        5:  C2          10:  P2/m       16:  P222         17:  P2221
     $ HALF,ONE,HALF,   half,half,onel,HALF,ONE,HALF,  HALF,ONE,HALF,
CCP4 $ HALF,HALFL,ONEL, HALF,HALF,ONEL,HALF,HALF,ONEL, HALF,HALF,ONEL,
C       18: P21212    1018: P21212      19: P212121       20:C2221
     $  ONE,QUAR,ONE,  onel,quar,onel, ONE,ONE,QUAR,  HALF,HALF,HALF,
CCP4 $ ONEL,QUAR,ONEL, ONEL,QUAR,ONEL, ONEL,ONEL,QUAR, HALF,QUAR,ONEL,
C       21:  C222       22:  F222       23:  I222         24: I212121
     $ QUAR,HALF,ONE,  ONE,QUAR,QUAR,  HALF,HALF,HALF, HALF,HALF,HALF,
CCP4 $ HALF,QUAR,ONEL, QUAR,QUAR,ONEL, HALF,QUAR,ONE, HALF,QUAR,ONEL,
C       47:  Pmmm       65:  Cmmm       69:  Fmmm         71:  Immm
     $ half,half,half, half,quar,half, quar,quar,half, half,quar,half,
CCP4 $ HALF,HALF,HALF, HALF,QUAR,HALF, QUAR,QUAR,HALF, HALF,QUAR,HALF,
C       75:  P4         76:  P41        77:  P42          78:  P43
     $ HALF,HALF,ONE, HALF,HALF,ONE, HALF,HALF,ONE, HALF,HALF,ONE,
CCP4 $ HALF,HALF,ONEL,ONEL,ONEL,QUARL, HALF,ONEL,HALFL,ONEL,ONEL,QUARL,
C       79:  I4         80:  I41        83:  P4/m         87:  I4/m
     $ HALF,HALF,HALF, ONE,HALF,QUAR,  half,half,half, half,half,quar,
CCP4 $ HALF,HALF,HALF,HALF,ONEL,QUARL, HALF,HALF,HALF, HALF,HALF,QUAR,
C       89: P422        90: P4212       91: P4122         92: P41212
     $ HALF,HALF,HALF, HALF,HALF,HALF, ONE,ONE,EIGH,   HALF,HALF,HALF,
CCP4 $ HALF,HALF,HALF, HALF,HALF,HALF, ONEL,ONEL,EIGH, ONEL,ONEL,EIGH,
C       93: P4222       94: P42212      95: P4322         96: P43212
     $  ONE,HALF,QUAR, HALF,HALF,HALF, ONE,ONE,EIGH,   HALF,HALF,HALF,
CCP4 $ HALF,ONEL,QUAR, HALF,HALF,HALF, ONEL,ONEL,EIGH, ONEL,ONEL,EIGH,
C       97: I422        98: I4122      123: P4/mmm       139: I4/mmm
     $  HALF,HALF,QUAR, ONE,QUAR,QUAR, half,half,half,  half,half,quar,
CCP4 $ HALF,HALF,QUAR, HALF,ONEL,EIGH, HALF,HALF,HALF,  HALF,HALF,QUAR,
C      143:  P3        144:  P31       145: P32          146:  R3
     $ ONE,ONE,ONE,   ONE,ONE,THRD,   ONE,ONE,THRD,    THRD,THRD,ONE,
CCP4 $ TWTD,TWTD,ONEL,ONEL,ONEL,THRDL,ONEL,ONEL,THRDL, TWTD,TWTD,THRDL,
C      147:  P-3       148:  R-3       149: P312         150:  P321
     $ twtd,twtd,half, twtd,twtd,sixt, ONE,ONE,HALF,   ONE,ONE,HALF,
CCP4 $ TWTD,TWTD,HALF, TWTD,TWTD,SIXT, TWTD,TWTD,HALF, TWTD,TWTD,HALF,
C      151: P3112      152: P3121      153: P3212        154: P3221
     $ ONE,ONE,SIXT,    ONE,ONE,SIXT,   ONE,ONE,SIXT,   ONE,ONE,SIXT,
CCP4 $ ONEL, ONEL,SIXT, ONEL,ONEL,SIXT, ONEL,ONEL,SIXT, ONEL,ONEL,SIXT,
C      155: R32        162:  P-31m     164: P-3m1
     $ THRD,THRD,HALF, twtd,half,half, twtd,thrd, one,
CCP4 $ TWTD,TWTD,SIXT, TWTD,HALF,HALF, TWTD,THRD, ONE,
C      166:  R-3m       168:  P6
     $ twtd,twtd,sixt,  ONE,HALF,ONE,
CCP4 $ TWTD,TWTD,SIXT, TWTD,HALF,ONEL,
C      169:  P61       170:  P65       171:  P62         172:  P64
     $  ONE,ONE,SIXT,  ONE,ONE,SIXT,   ONE,HALF,THRD,  ONE,HALF,THRD,
CCP4 $ ONEL,ONEL,SIXTL,ONEL,ONEL,SIXTL,ONEL,ONEL,THRDL,ONEL,ONEL,THRDL,
C      173:  P63       175:  P6/m      177: P622         178: P6122
     $  ONE,ONE,HALF,  twtd,twtd,half,  ONE,HALF,HALF, ONE,ONE,TWLT,
CCP4 $ TWTD,TWTD,HALFL, TWTD,TWTD,HALF,TWTD,HALF,HALF, ONEL,ONEL,TWLT,
C      179: P6522      180: P6222      181: P6422        182: P6322
     $ ONE,ONE,TWLT,   ONE,HALF,SIXT,  ONE,HALF,SIXT,  ONE,ONE,QUAR,
CCP4 $ ONEL,ONEL,TWLT, ONEL,ONEL,SIXT, ONEL,ONEL,SIXT, TWTD,TWTD,QUAR,
C      191: P6/mmm     195: P23        196: F23          197: I23
     $ twtd,thrd,half,  ONE,ONE,HALF,  ONE,HALF,HALF,  HALF,HALF,ONE/
CCP4 $ TWTD,THRD,HALF, ONEL,ONEL,HALF, QUAR,QUAR,ONEL, ONEL,ONEL,HALF/
      DATA ((ASULIM(II,JJ),II=1,3),JJ=74,NUMSGP)/
C      198: P213       199: I213       200: Pm-3         202: Fm-3
     $ HALF,HALF,ONE,  HALF,HALF,ONE,  half,half,half, half,half,quar,
```

```
CCP4 $ HALF,HALF,ONEL, HALF,HALF,HALF, HALF,HALF,HALF, HALF,HALF,QUAR,
C      204: Im-3       207: P432       208: P4232        209: F432
     $ half,half,half, HALF,HALF,ONE,  ONE,HALF,HALF,  HALF,HALF,HALF,
CCP4 $ HALF,HALF,HALF, ONEL,HALF,HALF, HALF,ONEL,QUAR, HALF,HALF,HALF,
C      210: F4132      211: I432       212: P4332        213: P4132
     $  HALF,THRQ,TWTD, QUAR,THRQ,TWTD, ONE,ONE,ONE,   ONE,ONE,ONE,
CCP4 $ HALF,ONEL,EIGH, HALF,HALF,QUAR, ONEL,ONEL,EIGH, ONEL,ONEL,EIGH,
C      214: I4132      221: Pm-3m      225: Fm-3m        229: Im-3m
     $ half,onel,eigh, half,half,half, half,quar,quar, half,half,quar/
CCP4 $ HALF,ONEL,EIGH, HALF,HALF,HALF, HALF,QUAR,QUAR, HALF,HALF,QUAR/
C
O
```

Contour levels

(September 2001)

Given a 'effective resolution' of the data, at what contour one has to examine the mixed fourier synthesis map (2Fo-Fc & Fo-Fc)? Is there any relation between resolution, completeness and contour?

Here's the plainly-practical answer:

The simple answer is that you want to contour at a level that gives the clearest view of the density. I normally contour 2Fo-Fc & Fo-Fc at 1 and 3 sigma, respectively, but in some parts of the map you want to lower the contour level. Your density will most likely not be equally strong throughout the map due to variations in B-factor or missing strong low resolution terms, so you'll have to adjust contour levels. If you lower the contour level too much you'll be blinded by all the noise features. Just use your eyes to tell what works best.

Then came a posting with some remarks, which sparked off a deep discussion about validity and statistics:

- The sigma level of an Fo-Fc is meaningless. In the early stages (poor and incomplete model), a 2-sigma feature may be genuine, whereas near the end of the refinement process (when the difference map is hopefully flat except for noise) even a 5-sigma peak need not be.
- If you cut out maps around your molecule and then use them in O, the "sigma level" is recalculated by O. As a consequence, this level will almost always be lower than the sigma level in the asymmetric unit, and features will show up at deceptive levels (*e.g.*, a "2-sigma" peak may be just a 1-sigma noise feature). [This problem does not occur when you use the good old map_* commands.]
- You want to be careful with going to too low a contour level. For an example of the dire consequences that can have, see Nature Structural Biology 8 (8), pp. 663-664 (2001) (you will need your nsb password to read the actual text) ... [If you need convincing, check the real-space fit and the map for 1F83 (chain B and C) at the Uppsala Electron-Density Server]

A practical summary of this discussion:

I think in practice everybody is doing the same thing. You DO look at 5 sigma peaks just as you look at outliers in the Ramachandran plot, too close contacts etc. Not because these must be errors, but because they are suspicious and you want to visually make sure that they are not errors or to fix them. Yes, a five sigma peak in a "perfect" map with very

low rms in terms of e/A**3 is meaningless but since the rms of the map is based on statistics, these will be extremely rare. What I would recommend is to use a peaksearch of the Fo-Fc map and look at the peaks sorted by peak hight. Starting at the strongest peak work your way down until you have had a whole row of peaks that you feel are not telling you anything. Important: don't forget to look at the biggest negative peaks. A -5 sigma peak is just as suspicious as a +5 sigma peak.
Wrt to putting a water in any >3sigma peak, this should clearly not be done. Interpreting density means you want to find a CHEMICALLY PLAUSIBLE explanation for any density feature. For a water that means that it should have at least one decent hydrogen bond with the protein and no too close contacts. You should always judge both density and "geometric/energetic sensibility" and if your density is poor you want to give more "mental weight" to the geometry. In many cases you will end up in the situation where you have the feeling that the difference density does indicate a problem but you can't figure out how to interprete it in which case you better leave it alone. Really the perfect model doesn't exist, you just want to get as close to it as possible given your data quality. Try never to overinterpret your density since others that look at the structure without seeing the density will blindly believe what you have built even if they shouldn't.

Please have a look at a http://www.ccp4.ac.uk/newsletters/newsletter40/contourdiscussionsummary.html and some of its follow-ups. At least one **very** practical point can be found there: At what level should one contour a difference map ? Well, one trick that may be useful is to leave out a well-defined atom (*e.g.* a carbonyl oxygen) in the map calculation and adjust the Fo-Fc contour level until that density looks just as good as the 2Fo-Fc density for the same (missing) atom. Then you know that well-ordered entities with ten-or-so electrons should have similar density features in both maps. This is completely general. When it comes to water molecules in particular, obviously one should use other criteria as well (plausible hydrogen-bonding partners, refine to reasonable B-factors, possess acceptable 2Fo-Fc density after refinement).

Then a purely statistical approach:

Speaking from a statistical point of view, a couple of points are worth making on the subject of calculating the standard uncertainty (SU) of the electron density (or difference density). Programs actually calculate the RMS deviation from the mean of the electron density. The question is, under what conditions is this an unbiased estimate of the SU ? - this is really what we are interested in if we want to judge the significance of peaks (or troughs) in the density. The answer is that the following conditions should apply:

10. The sample of density points used must be independent, for example 2 or more of the points used should not be related by the space-group symmetry. This is pretty self-evident, nevertheless most programs which purport to compute the RMS as an estimate of the SU violate this condition! FFT does it correctly since by default it always computes exactly one asymmetric unit (or should do!). However when you "extend" the a.u. to cover the volume of interest, the chances are that some points will have symmetry mates in the extended map. The correct procedure would be to simply use the value of the RMS originally computed by FFT.
11. The sample of density points used must either be the entire population or a random sample of it. Again the same argument as above applies here: it is clearly not valid to use the RMS value for a selected non-random portion of the a.u. as an unbiased estimate of the SU.

12. The sample of density points used must truly represent the "noise". The computed density will almost always include some of the "signal" we are looking for (of course this will always be true in a Fo or 2Fo-Fc map, and true for a Fo-Fc map except at completion of the structure). Therefore ideally the points containing signal+noise should be excluded from the calculation - for difference maps this can be done by using only the linear portion of the normal probability plot close to the origin to estimate the SU, and excluding the curved portion which should mostly represent the signal (assuming of course that the noise really does have a normal distribution). This method only really works well if the map is mostly noise with a small amount of signal - so it can only be used for difference maps.

How much effect these corrections will have on the estimate of the SU will depend obviously on the ratio of the volume of the map used to that of the a.u., and the ratio of the number of points containing some signal to those containing only noise.

Then, to round it off, some sound theoretical background:

The remarks below may still be of some help in pointing out that basic statistics cannot be ignored, even by those who do not love them, in the discussion of this question. It would seem that the 'central' concept behind this discussion is the Central Limit Theorem. If the lack of fit between Fo and Fc is randomly distributed without any trends nor correlations, the Fo-Fc map will be made up of white noise, *i.e.* its values will be normally distributed, so that the probability of a 5-sigma deviation will be less than 10**(-6). If the number of data is so vast that there are of the order of 10**6 independent data items or more, then a 5-sigma peak can occur by chance and hence be considered as noise. In more commonplace cases, however, the probability of a 5-sigma peak occurring by chance would be quite low, and therefore such a peak would be highly significant, as stated before.
GK's counterexample, with which AL disagreed, does seem rather contrived. If 10 times the sigma of the Fo-Fc map were to be considered as noise by some criterion, then the same criterion should lead one to conclude that the data have been grossly overfitted in the first place.

REJECT in SCALEPACK2MTZ

(January 2001)

I wanted to exclude a few reflections from my data-file using the REJECT flag in "scalepack2mtz". However, the reflections are kept in the output file. What can I do?

Here a summary of useful hints to the REJECT problem in SCALEPACK2MTZ:

13. This was indeed a bug. It has now been fixed in the CCP4 Suite version of the program.
14. There is no other CCP4 program to exclude selected reflections after processing (for some good reasons).
15. Use SFTOOLS with the following input:

```
16.        SELECT index h = 1
17.        SELECT index k = 10
18.        SELECT index l = 10
19.        SELECT INVERT
20.        PURGE
      YES
```

Using the following awk-script then gives the expected result which can easily be included into an input command file for SFTOOLS:

```
   awk '$7=="30.0000" {printf"SELECT index h = %3s\nSELECT index k =
%3s\nSELECT index l= %3s\nSELECT INVERT\nPURGE\nYES\n",$1,$2,$3}' fft.log
```

Real space difference map

(January 2001)

I'd like to compute a difference map, problem is that one dataset is in C2221 and the other in P63. I guess there's no way to do a difference fourier (Fo-Fo). But it should work in real space. How do I calculate a real space difference map?

21. You need to calculate maps in both space groups. Then mask the density for each with a mask from the model:

    NCSMASK XYZIN model.pdb MSKOUT model.msk

22. Then you will need to convert them to the same grid. MAPROT will do that - it is a bit complicated but there is an example.
23. Then MAPMASK or OVERLAPMAP can be used to "add" the maps applying a scale of -1.0 to one. MAPMAN can also be used for this procedure.

An alternative to step 1 and 2 may be to use MAVE option IMP, after which MAVE or MAPMASK should be used to cut out density inside a mask and "skew" it into position in the second cell.

Non-proline cis-peptide

(April 2001)

when I refine my structure, I can definitely see a cis peptide bond between proline and histidine. (it is very obvious from the 2fo-fc.map at R=17.1% and Rf=19.1%). This is a non-proline cis peptide because it is formed by CO of proline and NH of histidine. I am using CNS to refine the structure, and I changed the name of proline and defined the bond and dihedral parameters in the toppar files for this peptide bond, but it seems not successful in the map. It has not put the N atom to the density it should be. Does anyone have experiences on the refinement of non-proline cis peptide bond, or know how to deal with it? And where can I find the bond and dihedral parameters for non-proline cis peptide bond?

Summary from the enquirer:

24. Many people suggested using REFMAC5 becasue it can do it automatically.
25. Some people gave the toppar parameters to handle the situation, and some people even kindly provided their toppar files.

Even though this is not a CNS Newsletter, the various answers give food for thought, so a transcription is presented here:

- The most common answer is (with variations in parameters, as indicated):
  - Create a new file cis_peptide.param, similar to one you can get out of the script cis_peptide.inp (see below) and read this parameter file in refinement ".inp" file, as follows:

    {* parameter files *}
    {===>} parameter_infile_1="CNS_TOPPAR:protein_rep.param";
    {===>} parameter_infile_2="CNS_TOPPAR:water_rep.param";
    {===>} parameter_infile_3="CNS_TOPPAR:ion.param";
    {===>} parameter_infile_4="cis_peptide.param";

    The param file would look like this:

```
parameter
   dihedral
      (name ca and resid $res1) (name c and resid $res1)
      (name n and resid $res2) (name ca and resid $res2)
      1250. 2 180.
end
```

This defines a cis-peptide between residues $res1 and $res2.

  - A value of "5." instead of "1250." was suggested by one user.
  - Another suggestion is to put the lines from the cis_peptide.param file directly into the file 'refine.inp', in the following position:

```
§ ____________________________________________________
§
§    structure @&structure_infile end
§    coordinates @&coordinate_infile
§
§  end if
§
§ <<<< put statement here!
§
§
§  xray
§
§    @CNS_XTALLIB:spacegroup.lib (sg=&sg;
§                                 sgparam=$sgparam;)
§
§
   ____________________________________________________
```

  - For two molecules in the ASU, the param would be:

```
§ parameter
§  dihedral ( name ca and segid "A" and resid $res1 )
§           ( name  c and segid "A" and resid $res1 )
§           ( name  n and segid "A" and resid $res2 )
§           ( name ca and segid "A" and resid $res2 )
§           1250.0 1 180.0
§ end
§
§ parameter
§  dihedral ( name ca and segid "B" and resid $res1 )
§           ( name  c and segid "B" and resid $res1 )
§           ( name  n and segid "B" and resid $res2 )
§           ( name ca and segid "B" and resid $res2 )
```

```
     §            1250.0 1 180.0
        end
```

Please note the value of "1" instead of "2" (is this significant?).

- For refinement in CNS you have to have an extra parameter file which looks something like:

```
o   parameter
o       angl (name CA and resid $res1) (name C  and resid $res1)
o            (name N  and resid $res2)
o       485.856     119.700
o
o       angl (name C  and resid $res1) (name N  and resid $res2)
o            (name CA and resid $res2)
o        599.823    127.800
o
o       angl (name O  and resid $res1) (name C  and resid $res1)
o            (name N  and resid $res2)
o       759.150     120.600
o
o
o       dihe (name CA and resid $res1) (name C  and resid $res1)
o            (name N  and resid $res2) (name CA and resid $res2)
o       1250.0      2    180.0
    end
```

- If you're using TOPPAR/protein.top and TOPPAR/protein_rep.param files, the easiest way may be to modify the protein_rep.param as below (add last 4 lines). You don't need to define a special residue for cis-pept.

```
o   { very tight/rigid dihedrals }
o   dihe X    C    NH1  X     $kdih_rigid   1     0.0 ! omega torsion angle and
    ARG...
o   dihe CH1E C    N    CH1E  $kdih_rigid   2   180.0 ! allow cis PRO
o   dihe CH2E C    N    CH1E  $kdih_rigid   2   180.0
o   dihe CH2G C    N    CH1E  $kdih_rigid   2   180.0
o
o   dihe CH1E C    NH1  CH1E  $kdih_rigid   2   180.0 ! allow cis Pept
o   dihe CH2G C    NH1  CH1E  $kdih_rigid   2   180.0 !
o   dihe CH1E C    NH1  CH2G  $kdih_rigid   2   180.0 ! cis GLY
    dihe CH2G C    NH1  CH2G  $kdih_rigid   2   180.0 !
```

Large beta-angle in C2

(May 2001)

DENZO suggested a C2 cell with a = 143 b = 63 c = 94 beta = 130. Did anybody else observe such a large beta angle before in a protein crystal with a monoclinic cell?

- C2 cells often have this crazy angle. If you do an HKLVIEW plot of the *hnl* layers you can often see a more "sensible" set of reciprocal lattice vectors, with beta nearer 90 degrees, but they will require that you use a non-standard space group such as I2. In other words: the beta isn't biologically relevant, but it serves to predict all your spots.
- There are about 3 dozen entries in the PDB with space group C2 and a beta angle > 130 degrees. The highest beta angle of all PDB entries occurs for 1SPG:

```
CRYST1   89.600   75.600   69.700  90.00 141.90  90.00 C 1 2 1       4
```

However, a look at the WHATIF output for a few of these suggests there may be spacegroup problems or pseudo-symmetry for several of these entries.

Reflection vanishing act

(February 2001)

Yesterday I realized that I lost about half of my reflections in SHARP. Today I am looking for half of my reflections after converting CNS to mtz using f2mtz. It appears that every second reflection is simply missing??? Here is my script:

```
f2mtz \
        hklin hla.hkl \
        hklout  cbs_hla.extern.mtz \
        < f2mtz.log
CELL 144.524  144.524  108.161   90.000   90.000  120.000
SYMM P31
FORMT '(1X,3F6.0,6f10.3)'
LABO H K L FP SIGFP HLA HLB HLC HLD FOMcns
CTYPO H H H G L A A A A W
END
eof
```

First summary from the enquirer: It's neither the different AU nor the Friedel pairs nor the multiple line output from CNS but seems to be a read "feature" in F2MTZ. Anyway, including a blank line at every second line in my "free-format" data set gives me my complete "MTZ" data set.

But... this needed an update:

My previous summary was a little bit too early. The real bug was in the input script which was reading one item more than previously declared. F2MTZ thus kept reading also the next line and obviously gets troubles at the next line. This way, every second line was missing one variable while every second but one line was disappearing. So: Declare as many variables as you want to read - obviously.

Structure family

(May 2001)

I have resolved a new structure recently. How can I know whether it belongs to a new family or a family which have existed in SCOP?

The suggestions are visiting the following sites:

- http://www.ebi.ac.uk/dali/
- http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml

Related sites and servers:

- http://www.biochem.ucl.ac.uk/bsm/cath/
- http://cl.sdsc.edu/ce.html
- http://www2.ebi.ac.uk/dali/fssp/fssp.html
- http://www3.ebi.ac.uk/tops/

- http://bioinfo1.mbfys.lu.se/
- http://portray.bmc.uu.se/cgi-bin/dennis/dejavu.pl

(see: http://xray.bmc.uu.se/embo/structdb/links.html)

Stereo net

(May 2001)

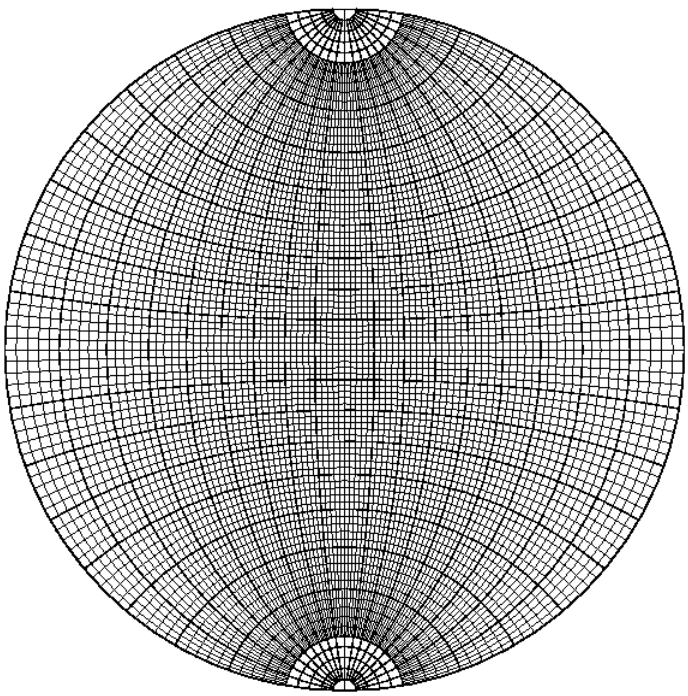
In the distant past I remember using a stereo net to measure the angle between different self-rotation peaks. Can anyone suggest where/how to get hold of one again?

- Don't know, but there was a program ROTANG in the BIOMOL package that given two rotation operations would give you the difference between them. It also did conversions between the common rotation specification definitions. I still use it in cases like this and could give you the source code.
- $CCP4/doc/stnet.doc
- `# this generates a stereographic net to overlay on the plot`
-
- `stnet plot ./net.plo`
  `pltdev -i ./net.plo -o ./net.ps`

  I _think_ this is intended for measuring *distance* between self-rotation peaks.
  If you need a net to measure *positions* of self-rotation peaks: I've got a postscript file that can be overlayed onto a POLARRFN plot.

- If you really want to do it the old-fashioned way, there's a Wulff net Postscript file on the CCP4 ftp server: Wulff net.
-

  Now available here (click on thumb-nails to get full-size net). If you have an automatic way of loading .ps files from the web, this is probably a bit better.



GETAX

(June 2001)

Why does getax complain "map not EXACTLY one cell" and how to fix it?

Summary from the enquirer: Mapmask run with either

ttt. explicit grid limits: 0 Xlim-1 0 ylim-1 0 zlim-1
uuu. XYZLIM CELL

I tried both and now getax runs.

How to combine phases from various sources

(June 2001)

I have many datasets for a protein from various sources, including MAD, SIRAS and MIR from different derivatives. Some of them are not isomorphous. I am just wondering whether there is any way by which I could refine and phase all of these derivatives in one single run of MLPHARE? (One problem is that I can't define different "natives" for different datasets, which I believe is necessary). If I can't do that, what's the best way to combine all of those phases from various sources? I know sigmaa can combine two sets of MIR phases. Is there any other program which can do this? and anything I ought to know for optimizing phase combination?

Summary from the enquirer:

- The simple easy method that works is just to write out the H-L coefficients for each individual refinement. Then simply add them up using SFTOOLS. The HL coefficients (phases) are very robust and non-isomorphism bothers them very little.
- You can try to run SOLVE, using the combine script !
- I don't know if it will be easy for you to install SHARP in your computer but I guess this is one of the best programs available for phasing. You will just need heavy atom coordinates from each dataset and the datasets themselves. It will take a while to combine phases from all datasets but I am almost sure they (phases) will be reliable.
- If you believe you need different native datasets to combine with your various derivatives, then you can't come up with one set of phases. Try to work with different subsets that each are sufficiently isomorphous. If one subset gives phases of adequate quality your problem is solved. If not, you could consider multiple crystal averaging across the maps derived from the various data subsets.
- Use dm_multi after sperately phasing your 'natives'. Since most are more or less isomorphous, start from a unity matrix and let it refine.
- In the end you have to choose a master data set which you want to phase, and phase that.
  So the way I would proceed:
  1. First make sure all our sites are as close to the origin as possible - that minimises the effect of cell differences.
  2. Then use native1 with derivatives 1H1, 1H2, etc, to refine the 1H1, 1H2, ... sites and get ISOE1 ANOE1.
  3. Use native2 with derivatives 2H1, 2H2, etc, to refine the 2H1, 2H2, ... sites. You want to use these sites, but get better estimates of ISOE2. (ANOE2 wont change..

4. So I would do one or two cycles of refinement of each of 2H1, 2H2, etc against native1, just to get the ISOEs and maybe let shift the coordinates a bit but not the occupancies.
5. Then do a final phasing run with all the derivatives v native1. In one case where we had awful non-isomorphism we could only get useful information to quite low resolution for the second set.

An alternative is to just add the HLA1 HLB1.. to HLA2 HLB2.. in Sigmaa but that takes no account of weighting the non-isomorphism.
Another way would be to use the two sets for multi-crystal averaging. See http://www.ysbl.york.ac.uk/~cowtan for a lecture where he has some discussion of this.

Molecular Replacement with Zn2+ as anchoring point

(July 2001)

Our protein contains two zinc ions for which we are able to pick up the signal. However, the phasing power is too low to solve the structure. With MR we also failed because the search model is less than half of the molecule with about 30% sequence identity but also containing two zincs.
Is it possible to use the zinc ions as an anchoring point and rotate the search model around this axis?! Which program will do so?

Summary from the enquirer:

- Hmmm - you lose the chance to use FFT search functions then. Best to verify your solutions by checking the Zn positions are consistent.
- Run SOLVE, RESOLUTION_STEPS 3 from let's say 20 to 4Å. Then run molrep in combination with those phases.
- Ask renaud.morales@ibs.fr at IBS. We published a paper concerning such an operation. Rotation about an axis defined by 2 Fe sites in that case. Note that you must place the model in two opposite positions, *i.e.* + direction and - direction, and carry out the search. Monitoring is with Rfree, and you eliminate all solutions with bad packing.
- In the old XPLOR/CNS you could specify your rotations explicitly and I guess that is still true. You'll have to find out the direction of your rotation and then rotate around it in let's say 2.5 degree steps. An alternative is to generate the set of models yourself and use each one of them for a translation search in for instance amore. This will require a bit of scripting but could be more sensitive. The real question is whether a successful solution of this problem is going to give you an interpretable map. At < 30% identity and only 50% of the full structure I think that's going to be tough.
- Did you try soaking a crystal in EDTA to remove the Zn, and generate a set of isomorphous differences that way? It's worked for me. And if you do the rotation search, don't forget you need to consider the two cases of ZnA from your model superimposed with Zn site 1 and site 2 from your data.
- If you want, I can try EPMR followed by SHAKE&WARP to salvage a weak solution. we rebuilt complete structures from less than 50%, but admittedly from reasonable models and decent data.
- I had the same problem with a structure I am working on right now. I can see two Ca2+ ions in the map but the solution does not give good phases to solve the whole structure. I agree with some of the others: use Se-Met or other heavy atom methods to solve it.
- If your data is of fairly high resolution and quality, try direct methods. In particular, try the new OASIS program in CCP4.

- Have you already tried to run MOLREP with your starting phases? There's another program BRUTEPTF which sounds interesting. NYSGRC
- If you could compute an electron density map with the Zn-derived phases and the map is just good enough for you to identify the molecule boundaries, you could try to get the approximate center-of-gravity of your own molecule. In this way, you get three anchoring points and would, together with the two zincs and the c-o-g of the model (could be easily found by a simple run of MOLEMAN2), be able to determine the RT matrix by Site2RT in RAVE.
- Since you have only < 50% of a model, with 30% seq identity, MR is going to be tough... Probably (or most likely...?) Randy Read's BEAST program will give you better results than conventional MR programs, since it uses maximum likelihood theory in its MR functions. At a workshop in Como last June, Randy presented some promising figures on test cases, which were outstanding compared to results from AMoRe, especially in non-trivial cases. Ask Randy for program and details.
- We have found BEAST from Randy Read to find solutions when we all but lost hope! If you want to go down the random search method, a simple automated method would be to work out the rotations and translation to take your zincs to lie along one axis (put a couple of points along this line and AMORE will do this during its centre of mass calculations). Then simply apply an incremental rotation around the axis (5 degrees around z if you do AMoRe), *i.e.* cycle = 0, cycle_now=cycle + 1, ot_now=cycle_now x 5, in LSQKAB apply rot_now to your centred model (which has had the same centre of mass centering rot and trans as the zinc atoms in a line applied). Then add another LSQKAB which would return the zinc atom line back to the correct position, this will move your model to a point in the cell rotated around the two zinc atoms. You can then test the model by some criteria (I would think packing first), then test it by calculating R-factors or correlation coeffs. The longest part of the whole business will be writing the script. My experience is that it almost never works (partly because I only do stuff like this when things are hopeless) but it's very satisfying to get the script running. I would use BEAST or do phased translation searches.

Rfree vs resolution (complete with graph!)

(August 2001)

I think I have pestered you already once with this question. Where was this elusive graph published of:

statistical expectation value of rfree (or so) vs resolution?

It looked somewhat like the Cruickshank Rfree vs DPI plot if I understand the rumors correctly....
A freeR of 20 for a 3.5Å structure is probably as unlikely as a freeR of 29 for a 1.2Å structure and both warrant some explanantion.... And just to heat up the flames: I think freeR was probably the single most significant contribution to put an end to DreamWorks crystallography....almost.

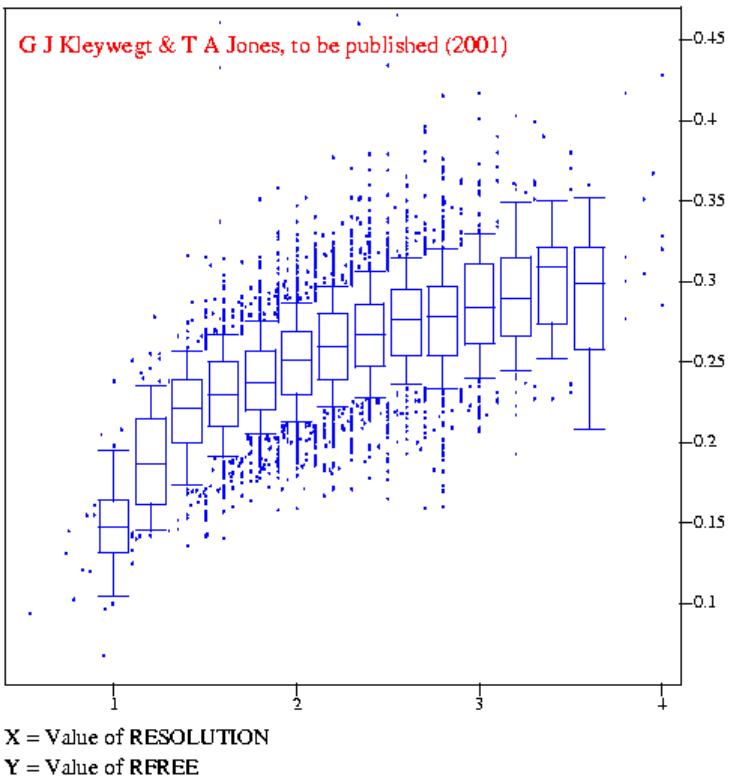
I suspect you may be refering to one of these two papers:

- Tickle, I.J., Laskowski, R.A. & Moss, D.S. (1998). Rfree and the Rfree ratio. I. Derivation of expected values of cross-validation residuals used in macromolecular least-squares refinement. Acta Crystallogr. D54, 547-557 (find the PDF version at Acta D).

- Tickle, I.J., Laskowski, R.A. & Moss, D.S. (2000). Rfree and the Rfree ratio. II. Calculation of the expected values and variances of cross-validation statistics in macromolecular least-squares refinement. Acta Crystallogr. D56, 442-450.

Less likely (but a gripping yarn nonetheless ;-): Kleywegt, G.J. & Brunger, A.T. (1996). Checking your imagination: applications of the free R value. Structure 4, 897-904.

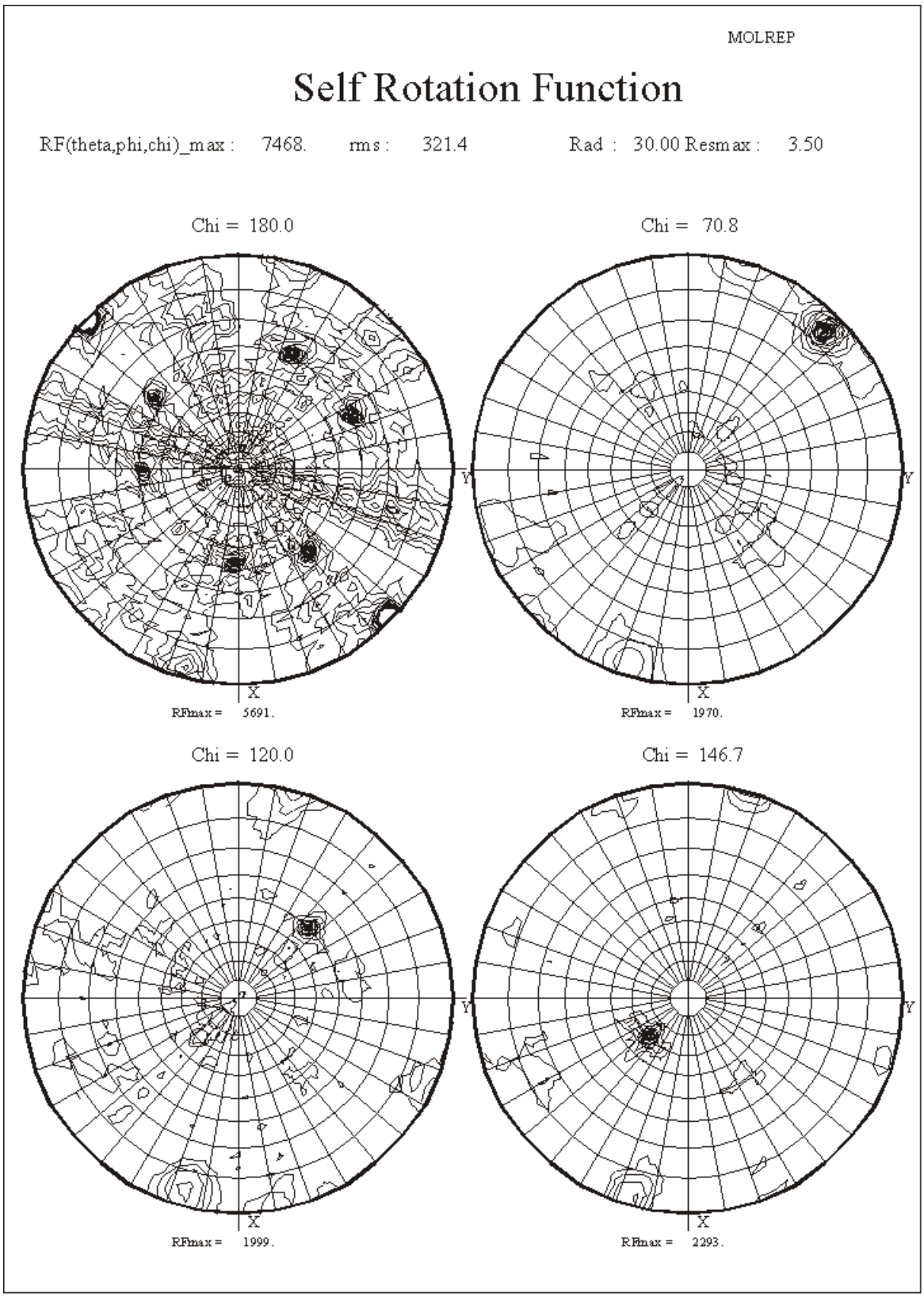
Find attached a plot of rfree versus resolution based on ~6500 PDB entries. This is a box plot - in every resolution bin (from the bottom up) the 10th percentile, 25th, 50th (*i.e.* median), 75th and 90th percentile are indicated; outliers below 10 and above 90 are shown individually as small specks. The linear correlation coefficient is +0.56 (figure from: gjk & ta jones, to be published (2001, if I have time)). Your remarks about certain Rfree warranting explanation is absolutely true. This is also borne out by the attached plot.



Trouble interpreting self-rotation

(August 2001)

I have a triclinic crystal of which structure can (hopefully...) be solved by molecular replacement. According to matthews-coef, the protein is possibly a hexamer(3.0) or an octamer(2.3). Gel filtration specified it's a hexamer. The problem is that I can't imagine the NCS point group by looking at the self-rotation map. It has strong peaks at chi=180, 146.7, 119, and 70.9. I could find 8 peaks at chi=180. The map is attached. Click on the thumbnail to enlarge. Can someone help me to understand the possible spatial arrangement of this multimeric protein? I have one more question: According to Schroder et al. (Structure, 2000, 8(6):605), they created 3,600 search models from the interpretation of self-ratation map and solved the structure by MR. How can one create such large numbers of probes?



Your MOLREP self-rotation map would suggest to me that you have two tetramers with 222 point group symmetry in your triclinic cell. These are related by a two-fold NCS axis (MAIN DYAD) located at phi=45, omega=90 for which the rotation function value is higher (RFmax = 5691) than other dyads peaks (Chi=180), which relate dimers in the tetramers and also odd peaks at Chi=70.8, Chi=120 and Chi=146.7 (RFmax about 2000) which relate dimers of the two tetramers which are not related by the MAIN DYAD.

Yes we have generated all 36 hundred decamers in the above paper using relatively simple C-shell script and a bunch of CCP4 programs. I can send you this script and a script to grep and analyse the solutions if you are interested.

Note on CCP4BB 'rules'

At the end of a long discussion on what a CCP4BB posting should or should not contain (including flaming, spelling and grammatical errors and anonimity), the following information was sent from CCP4:

Order in which to approach a problem with a CCP4 program (or related queries):

- RTFM
  including CCP4 manual, HTML docs and tutorials (in CCP4i etc), newsletters, study weekend proceedings.....
- ask your more experienced labmates and maybe even your supervisor, if you dare
- read the source code (and isn't it great that this is possible?) J may the force be with you
- ask the BB

The procedure outlined is an excellent one. I would only make a small addition - **you can also email CCP4 staff at DL** - but if it's a question on general usage etc. we prefer these go to CCP4BB as they are of a more general interest (the address for CCP4/dl staff is below. As far as abuse goes it's unfortunate, as someone has shown, how easy it would be to send such messages anonymously - but I'm sure the 1900 people subscribed to CCP4BB appreciate the near open forum and also have better things to do!

Here is what every new user of the CCP4BB is greeted with:

```
     *** Welcome to the general CCP4 bulletin board 'ccp4bb' ***

If you wish to send messages to this bulletin board then send them to
ccp4bb@dl.ac.uk. Any crystallographic related item is acceptable, not
necessarily directly related to CCP4, for example: problems, job adverts
and requests for information.

Unacceptable content includes personal messages and abuse, and messages of
an unrelated commercial nature.

To prevent abuse of the mailing list only members of the list are able to
post to it. This is done by checking the email of the sender against the
email addresses of the members of the list; please check that you are
sending messages from the same address with which you have subscribed.

CCP4 reserve the right to remove addresses from the list without notice if
they have persistent delivery problems.

To unsubscribe from the mailing list send the message
      unsubscribe ccp4bb
to majordomo@dl.ac.uk. Any requests about the lists, for example for help,
should also be sent to majordomo.

Etiquette:
1. Always write messages in plain ASCII: attachments and/or encryption are
 not appropriate to this forum
2. Please always add a short but descriptive Subject line
3. Please post a summary of the replies you receive to ccp4bb, so that
```

```
  others may benefit

 More information about CCP4 can be found at
       http://www.dl.ac.uk/CCP/CCP4/main.html
```

Announcements

HIC-Update

(January 2001)

HIC-Up, the Hetero-compound Information Centre - Uppsala, has been updated and now contains information on 2,971 hetero-entities that have been taken from the PDB (up from 2,640 in July, 2000).

The URL for HIC-Up is: http://xray.bmc.uu.se/hicup.

(September 2001)

HIC-Up, the Hetero-compound Information Centre - Uppsala, has been updated and now contains information on 3,296 hetero-entities that have been taken from the PDB. For URL, see above.

RAVE (MAPMAN, etc.) for LINUX

(January 2001)

In Uppsala: if you use the "run"-script, just type things like "run mapman" etc. on your Linux box.
Elsewhere: you can download RAVE for Linux from xray.bmc.uu.se, directory pub/gerard/rave, file rave_linux.tar.Z (or the individual programs from directory rave_linux). Check Uppsala Software Factory and http://xray.bmc.uu.se/usf/ftp.html for help with download.

CCP4 v4.1

(30 January 2001)

```
    ############################################################
    #                                                          #
    #                    The CCP4 SUITE                          #
    #                                                          #
    #               -Computer Programs for                       #
    #            Macromolecular Crystallography                  #
    #                                                            #
    #                      VERSION 4.1                           #
    #                                                          #
    ############################################################


          ------ OUT NOW ! ------
```

Further details on obtaining the Suite can be found on the http://www.ccp4.ac.uk.

CCP4 v4.1.1

(2 March 2001)

The Daresbury ftp server has been updated to patch release 4.1.1. Relative to 4.1, this release contains some fixes to problems discovered in 4.1. If you have successfully installed 4.1 and none of these problems is relevant to you, then there is probably no point in updating.

If you want/need to update, then there is a global patch file provided, but note that this will not patch any binary files *e.g.* images or .class files) - otherwise it should be safe to take individual files.

MOSFLM - release of version 6.11

(March 2001)

I have put a new version of Mosflm on the ftp://ftp.mrc-lmb.cam.ac.uk/pub/moslfm/ver611 (can also be accessed through http://www.mrc-lmb.cam.ac.uk/harry/).

MOLREP 7.0

(March 2001)

New version of MOLREP (MOLecular REPlacement program) (7.0) is now available (beta release) from ALExei.
Or use york's ftp

```
ftp ftp.ysbl.york.ac.uk
login anonymous
cd pub/alexei
get molrep7.tar.gz
```

After gunzipping and untarring follow instructions in README.

ACORN in CCP4

(March 2001)

A test CCP4 version of ACORN is available now. ACORN is a flexible and efficient *ab initio* procedure to solve a protein structure when atomic resolution data is available and has already solved at least 4 protein structures with the size from 125 to 350 amino-acid residues.
To obtain the program:

```
ftp ftp.ysbl.york.ac.uk
login: anonymous
password:your full email address
ftp > cd pub/yao
ftp > get acorn.f
ftp > quit
```

More tutorials for SFTOOLS etc.

(March 2001)

In case there are users of my old web pages that haven't found the new site, here is the http://eagle.mmid.med.ualberta.ca/. Other tutorials of interest (Data collection, and heavy atom binding) can be found on the http://eagle.mmid.med.ualberta.ca/highlights.html page.

Honorary Doctorate for Eleanor

(April 2001)

As you can (or cannot - depending on how good your Swedish is) read in the attached newspaper clipping from last Saturday's "Upsala Nya Tidning", Eleanor Dodson will receive an honorary doctorate from Uppsala University this Spring. Congratulations, Eleanor!

UNT, 2001-04-07

Peter Wallenberg hedersdoktor

UPPSALA UNIVERSITET. Direktör Peter Wallenberg är en av flera hedersdoktorer som utsetts vid Uppsala universitets teknisk-naturvetenskapliga fakultet.

Det blir vårpromotion 1 juni också för biokemisten Eleanor Dodson, Storbritannien – samt professorerna Peter Roepstorff, Danmark, Dan-Olof Riska, Helsingfors, Jack F Kirsh, USA och Franz Grieser, Australien.

Peter Wallenberg

(July 2001)

It is our pleasure to announce that Eleanor Dodson has been promoted to personal chair at the University of York. Since the end of May 2001 she has changed from 'Mrs. Dodson' through 'Dr. Dodson' (from the beginning of June 2001, as announced earlier on this bb by DVD), to 'Prof. Dodson' from now. But thankfully she will always be Eleanor!

Our warmest congratulations on what many of us who know her, have felt was due a long time ago, and what many others have thought was true already anyway (judging by the mail she receives).

cctbx - Computational Crystallography Toolbox

(May 2001)

```
-----------------------------------------------------
          First general release of the

      Computational Crystallography Toolbox

          http://cctbx.sourceforge.net/
-----------------------------------------------------
```

AutoDep 3.0 at EBI

(May 2001)

```
***  Announcement of AutoDep Version 3.0 at EBI ****
```

A new version of AutoDep at EBI for PDB Submissions, will become available on Tuesday 8th May at the EBI, via URL autodep.ebi.ac.uk. This version is designed to support Harvest/Deposition file information from CCP4 (SCALA, TRUNCATE, REFMAC) and from the CNS programme suite. The information from Harvest files can be merged, resulting in quicker deposition. For some further details see http://autodep.ebi.ac.uk/autodep-doc/html/v3.html.

New Version of PDB mode for Emacs

(June 2001)

pdb-mode is a major mode for the GNU-Emacs/XEmacs editors, providing editing functions of relevance to Protein DataBank (PDB) formatted files. This includes simple ways of selecting groups of atoms and changing attributes such as B-factor, occupancy, residue number, chain ID, SEGID etc.

PyMOL v0.56 (+ Windows Installer)

(July 2001)

PyMOL v0.56 has been released at http://pymol.sourceforge.net/.

Updated Tcl/Tk/BLT on CCP4 ftp server

(September 2001)

After recent (entirely justified!) complaints about the pre-built Tcl/Tk/BLT executables on the CCP4 ftp server, I have rebuilt the IRIX and OSF1 binaries and also added a new webpage to the relevant ftp directory. Please ignore this message if you are already happily using Tcl/Tk 8.3 and BLT 2.4 on your system, since the source code has NOT been updated. Otherwise you can pick up the packages by accessing

ftp://ccp4a.dl.ac.uk/pub/ccp4/tcltk/README.html

and following the appropriate link (alternatively you can connect directly via anonymous ftp to the server).