CCP4 NEWSLETTER ON PROTEIN CRYSTALLOGRAPHY

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

Number 41 **Fall 2002**

Contents

Editors: Charles Ballard and Maeri Howard-Eales

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

News from CCP4: Autumn 2002

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

---

1. CCP4 study weekend

The 2003 CCP4 study weekend is on "Experimental Phasing" and takes place on the 3rd and 4th of January at York University. The scientific organisers are Neil McDonald (Cancer Research UK) and Airlie McCoy (Cambridge, UK).

Invited speakers include:
Gérard Bricogne(Global Phasing, UK)
Ditlev Brodersen (MRC, UK)
Zbigniew Dauter (Brookhaven, USA)
Phil Evans(MRC, UK)
Eleanor Dodson(York, UK)
Gwyndaf Evans(Global Phasing, UK)
Joe Ferrara (Rigaku/Molecular Structure Corp,USA)
Elspeth Garman (Oxford, UK)
Ana González (Stanford, USA)
Ralf Grosse-Kuntsleve (Berkeley)
Simon Parsons,(Edinburgh, UK)
Michael Quillin (Oregon, USA)
Raimond Ravelli (EMBL, France)
Randy Read (MRC, UK)
Gabby Rudenko (Texas, USA)
Thomas Schneider (Göttingen, Germany)
Bi-Cheng Wang (Georgia, USA)

Again CCP4 will be running the highly regarded “Introduction to CCP4” mini-workshop on the Friday morning 9-10:30. The topics to be covered will include

- General introduction to CCP4
- CCP4 Road Maps
- Where to find help on CCP4
- What's New in CCP4?

Also, there will be a Protein Crystallography Specialist Users Group meeting on the Thursday afternoon, starting at 14:30.

The Proceeedings of the 2002 CCP4 Study Weekend on "High-throughput Structure Determination" are available in *Acta Crystallographica Section D* Volume 58, Part 11 (November 2002).

2. Workshops and Conferences

The CCP4 conference roadshow for 2002 had stands at the *XIX Congress and General Assembly of the IUCr* in Geneva, Switzerland, and at the *American Crystallographic Association annual meeting* in San Antonio, Texas.

CCP4 staff with the PDB at Geneva

In August CCP4 staff Charles Ballard and Peter Briggs along with MOSFLM expert Harry Powell (now a regular fixture at these events) travelled to the beautiful city of Geneva for the XIX IUCr congress. As usual we manned the CCP4 stand in the commercial exhibition and offered demos and tutorials of the software to an unsuspecting public. (We were also pleased to see the ab initio phasing program ACORN getting plenty of exposure, both in Michael Wolfson's opening lecture and Yao Jia-xing's talk in the scientific sessions.) We enjoyed the conference and the city (especially its amazingly efficient public transport!) and we'd like to thank all those people who took the time to come and talk to us - it's always a pleasure to meet our users face-to-face and hear your comments.

Prior to the ACA an “Introduction to CCP4” one-day workshop was held. This was a great success, being attended by more than 80 people. CCP4 would like to extend it's thanks to the speakers, Harry Powell, Roberto Steiner and our own Alun Ashton. Following this success there will be a one-day workshop on July 26thorganised in conjunction with the 2003 Annual Meeting of the ACA at Cincinnati, July 26-31.

Data processing and scaling with Mosflm and CCP4

ACA Annual Meeting, Cincinnati, 26 July, Organiser Harry Powell

The package distributed by CCP4 includes programs for all aspects of protein crystallography up to the model building stage. Following a short introduction to the suite, this workshop will concentrate on integration and scaling datasets with Mosflm and SCALA. Basic processing of routine datasets will be followed by detailed analysis of the programs' output and treatment of more demanding experiments. While new users of the programs may benefit, this is aimed at protein crystallographers with some experience of data processing who may want to develop their skill in using these programs. There will be presentations by CCP4 staff and developers as well as the opportunity for informal group discussions during the workshop.

3. Release of CCP4 4.2

On April 30 CCP4 release version 4.2 of the suite. This release included the follwing new applications:

ACORN
ab initio procedure for the determination of protein structure at atomic resolution (*Yao Jia-Xing*)

BEAST
Brute-force molecular replacement with Ensemble Average Statistics, Maximum likelihood-based molecular replacement (*Randy Read*)

PROFESSS
determination of NCS operators from heavy atoms (*Kevin Cowtan*)

ROTAMER
list amino acids whose side chain torsion angles deviate from Richardson's Penultimate Rotamer Library (*Dirk Kostrewa*)

ASTEXVIEWER
Java application for viewing proteins, ligands and electron density maps (*Mike Hartshorn*)

plus others.... Also new versions of REFMAC5 (5.1.19), MOSFLM (6.2.1), DYNDOM(1.2), FFFEAR(1.9), MOLREP(7.3) and SCALA (3.1.4).

Updates to the graphical user interface CCP4i:

- New and updated tasks, including:
  - New interfaces to: ACORN, BEAST, OASIS, ANISOANL, TLSANL, PHISTATS, PROFESSS
  - New stand-alone task in Coordinate Utilities to import and edit a protein sequence.
  - An explicit "Structure Factors for Deposition" task to encourage you to deposit your data!
- ARP/wARP: stand-alone interface has been withdrawn - get the latest ARP/wARP suite and CCP4i interface from http://www.arp-warp.org/
- Map sections can be viewed with MAPSLICER, and this can be set as the default viewer in the Preferences.
- Configuration: The task lists can be customised using the "Edit Modules File" task under the "System Admin" menu.
- Install New Task: substantially revised version allows install, uninstall and export of new tasks.

Other highlights in 4.2 include:

- Compliation for Mac OSX using Darwin configure switch.
- Compilation of the LAPACK linear algebra package is now the default as REFMAC, SCALA and BEAST now depend on it. In extremis, there is a --disable-lapack configure switch.
- New library pxxml.f for writing XML files. These routines are used by ALMN, MATTHEWS_COEF and PEAKMAX.

Plus many, many minor changes....

More recently version **4.2.1** was released on July 15 and 4.2.2 on November 25. Binaries are available for IRIX, linux (intel), OSF, SunOS, Mac OSX and Windows.

4. Other news

Graeme Winter, who previously worked on the updating of the Mosflm GUI and server, has joined Daresbury Laboratory as part of the e-HTPX project. His brief is to work on the automation of protein structural solution.

The aim of the e-HTPX project is to unify the procedures of protein structure determination into a single all encompassing interface from which users can initiate, plan, direct and document their experiment either locally or remotely from a desktop computer. More information is available at www.e-htpx.ac.uk.

Developments with CCP4i: October 2002

Peter Briggs, *Pryank Patel, Alun Ashton, Charles Ballard, Liz Potterton*, Maria Turkenburg*, Martyn Winn*

CCP4, Daresbury Laboratory, Warrington WA4 4AD, UK
**Structural Biology Laboratory, Department of Chemistry, University of York Y010 5YW, UK*

Introduction

CCP4i is the CCP4 graphical user interface. The last officially released version of the interface was 1.3.8, included as part of CCP4 4.2.1. This article outlines the major changes in 1.3.8 against the previous version of CCP4i, and looks ahead to some of the future developments planned for the next release and beyond.

Changes in CCP4i 1.3.8

Many of the new and updated features in the current version of CCP4i were previewed in the previous newsletter (issue 40 March 2002). As well as a number of relatively minor changes to fix bugs and consolidate earlier changes, there were a number of new and updated task interfaces, reflecting changes and additions to the suite in release 4.2.

New Interfaces

These include interfaces for CCP4 programs ANISOANL, TLSANL and OASIS, as well as interfaces for the major new programs in 4.2: ACORN (ab initio procedure for the determination of protein structure at atomic resolution), BEAST (maximum-likelihood molecular replacement program), PROFESSS (determination of NCS operators from heavy atom substructure) and ROTAMER (comparison of atomic coordinates against Richardson's Penultimate Rotamer Library).

In addition, the GET_PROT interface (now renamed SAPHIRE) is a CCP4i-only application which allows the user to download and edit protein sequence files accessed either from the EBI or from a local file. An interface for WHAT_CHECK (the subset of protein verification tools from the WHAT IF program) is also included although the program itself is not yet distributed with the CCP4 suite. (See http://www.cmbi.kun.nl/whatif/ for more information on WHAT IF.)

Major changes have been made to the SCALA interface, to accommodate changes in the handling of datasets. 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 (for example the addition of a ``Validation and Deposition'' module) to improve access to relevant tasks at various stages of the structure solution process.

New and Updated Utilities

MapSlicer

MapSlicer offers interactive display of contoured 2D sections through CCP4-format density maps. MapSlicer has been significantly improved between CCP4 4.1 and 4.2 (including a substantial redesign of its own user interface) and is now built as a standard part of the default for many platforms. As well as allowing the user to flip between different sections and map axes, the program also allows the display of ``slabs'' of multiple sections, and the ability to go directly to Harker sections.

Although not strictly CCP4i, MapSlicer uses many components from the interface and so maintains a similiar ``look and feel''. Also, it is now possible for the user to set their preferences to make MapSlicer the default viewer for maps when accessed via the ``View Files from Job'' menu on the main CCP4i window.

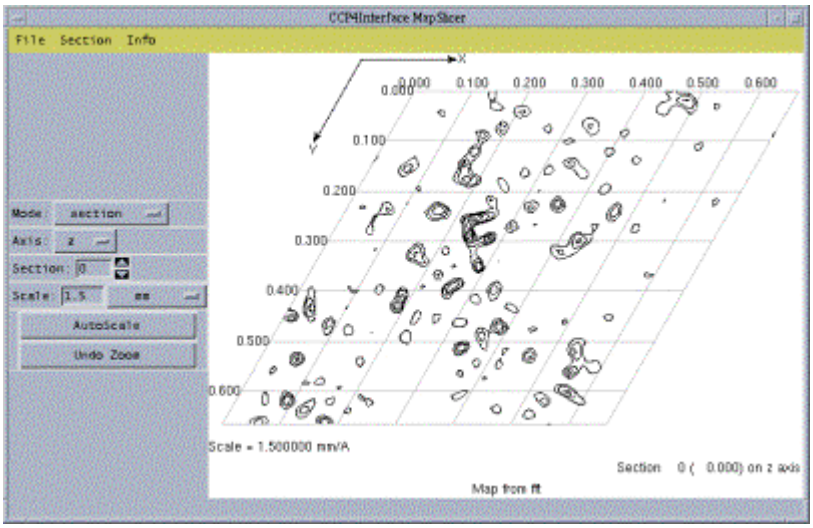


Figure 1: screenshot of MapSlicer

The range of features offered by MapSlicer is still relatively modest, for example displays are still only black and white. A number of possible improvements are envisaged, including a built in peaksearch algorithm, and display of coordinates - for example heavy atom positions or peaks corresponding to calculated Harker vectors.

AstexViewer

AstexViewer is a Java application written by Mike Hartshorn to display density maps and protein-ligand complexes. One way of using the viewer is to embed it as an applet in a webpage and then view it using your favourite browser. CCP4i now includes a task interface which will generate these pages automatically and launch a browser to view them.

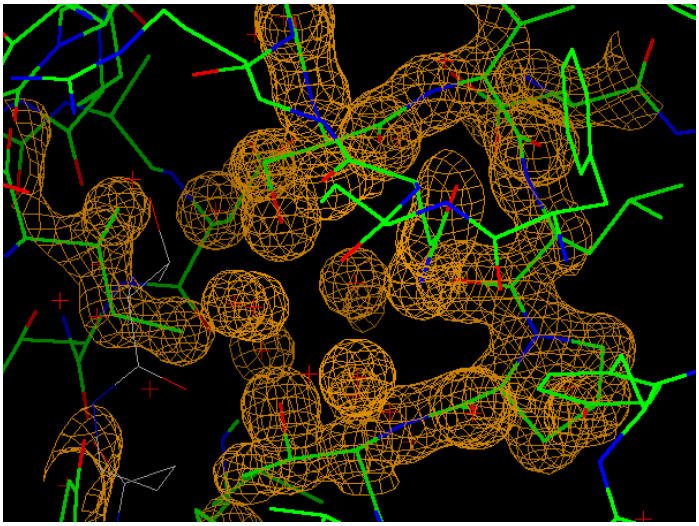


Figure 2: gif output from the AstexViewer, showing a protein structure inside a density map

Task Installer

The Task Installer utility for installing new task interfaces has been substantially upgraded, and aims to provide a robust mechanism for installing and tracking ``third-party'' interfaces - that is, interfaces provided for non-CCP4 software by the authors of that software. An example of this is the CCP4i interface for ARP/wARP, written by Tassos Perrakis and distributed with the latest version of the ARP/wARP suite (version 6.0 - see http://www.arp-warp.org for more details).

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 possible to access the ``install'' and ``uninstaller'' functions from the command line, via the *ccp4ish -install* and *ccp4ish -uninstall* options, which allows it to be incorporated into Makefiles or installation scripts for other packages.

The task installer can be accessed from the main CCP4i window via the ``System Administration->Install Tasks'' option.

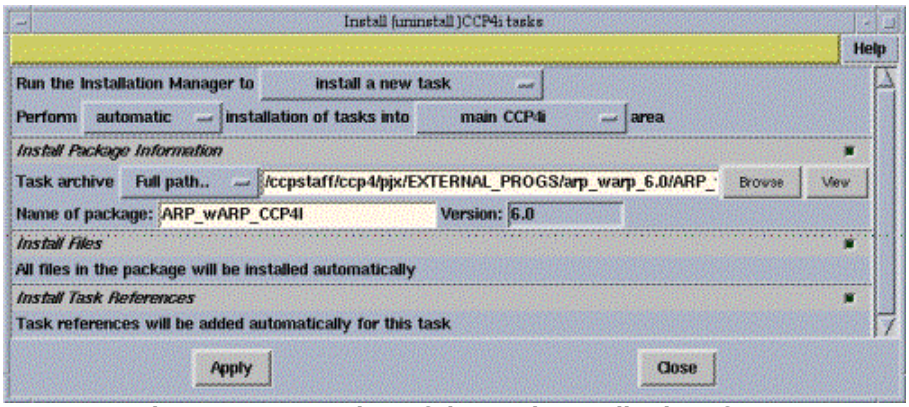


Figure 3: screenshot of the Task Installer interface

Automating Tasks

Two trial initiatives for automating tasks within CCP4i are included in 1.3.8, both of which involve passing information via XML files.

- **Parameter passing in Molecular Replacement**

  Background: before running the Molrep task it is useful to know certain details such as the number of monomers expected in the unit cell, and the existance of pseudotranslation vectors. These can be determined using the ``Cell Content Analysis'' and ``Analyse MR Data'' tasks respectively from the Molecular Replacement module, but in each case the data must be manually transfered from the log files to the Molrep interface.

  In this pilot project, both the ``Cell Content Analysis'' and ``Analyse MR Data'' tasks generate XML output files recording key information. These files are generated by calls to the CCP4 XML-writing library PXXML within the MATTHEWS_COEFF and PEAKMAX programs, activated using the XMLOUT keyword in each.

  The Molrep task can then automatically check for the existence of these output files and use the information in them to fill out the appropriate fields in the task interface for the user, reducing manual handling and typographical errors.

  Reading of the XML files is performed using a set of utility functions built upon xml.tcl 1.9 and sgml.tcl 1.7, both included in CCP4i 1.3.8. (Note that the XML output of ``Cell Content Analysis'' has to be processed, with the number of monomers used being estimated as that number giving a solvent fraction closest to 50% of the unit cell.)

  This option is turned off by default in CCP4i 1.3.8. Users wishing to try it can switch on the functionality in the ``XML Output'' folder in their ``Preferences'' (accessed from the menu on the RHS of the main window).

- **CAD AutoReindexing**

  Background: occasionally when merging together MTZ files using CAD, it is possible that some will have different indexing conventions to the others leading to errors when combining them.

  The CAD task now includes the option to ``Automatically check and enforce consistent indexing between files''. With this option selected each MTZ file is checked for consistent indexing against the first file, which is used as a reference. Files which are differently indexed are then reindexed prior to being merged. This mechanism cuts down on the overhead of manually diagnosing and correcting such cases when they arise.

  This option uses XML passing within a single task rather than between tasks, as in the previous example. In this case ALMN is used to diagnose whether reindexing is required between two files, and if so then which reindexing operator to use. This information is written to a temporary XML file using the same mechanism as before, and is read from within the script using Tcl XML utilties.

In both these examples it would also have been possible to pass the same information using other mechanisms, for example by processing the log files directly using a variant on ``grep'' and other Unix-type cutting-and-sorting tools. However such methods are usually overly-complicated and prone to being easily broken by even small changes to log file formats. In contrast XML parameter passing is far simpler, more robust and easily extensible.

Core Documentation

As of CCP4i 1.3.8 the interface source code includes inline ``doc-comments'', which are extracted and turned into html documentation of the code. Both the commented code and the extracted documentation are included in the current release, and will be useful for any programmers wishing to make CCP4i work more easily with their programs.

Future Developments

A number of longer-term projects are also envisaged:

- **MTZ Viewer**

  The MTZ files impose a formal hierarchical structure on the reflection data they store, of the form ``Crystal->Dataset->Column'' (see Martyn Winn's article ``Development of the CCP4 software library'' in the newsletter 40, March 2002, for more explanation). In the future it should be possible to write programs which exploit this hierarchy for automate selection of data columns based only on a crystal or dataset name.

  The MTZ viewer will display the crystal/dataset/column structure as a hierarchy or ``tree'', making it easier to visualise. Initially it is intended that the viewer should also act as a selection tool for datasets and columns, although ultimately it could also be used as an interface to perform CAD-like operations, for example to merge or split datasets.

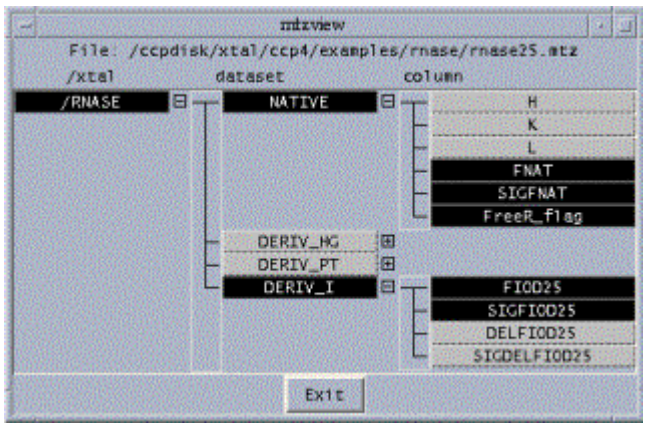


Figure 4: screenshot of prototype MTZ Hierarchical Viewer

- **Project History Database**

  The Project History Database is one of the most useful features of CCP4i above the ability to run the programs, as it tracks the jobs and associated datafiles which have been run within each project and allows the data from various parts of the structure

determination to be accessed quickly. (For an overview see the article on ``Using CCP4i as a Project Management Tool'' in newsletter 38, April 2000.)

There are however a number of limitations in the current implementation, which spring from the fact that the code for handling the information in the project history database is embedded within the main CCP4i process. It is intended therefore to separate this component - the ``project history database handler'' or db handler for short - i nto an independent server process which can talk to the main CCP4i process via sockets.

In the short term this should not affect users at all; it will make a number of possibilities more feasible in the future. External packages such as MOSFLM would be able to interact with the database independently of CCP4i and leave records of jobs that it ran. The db handler could one day use a different database backend, for example a mySQL database, and interact with other databases storing different information, for example laboratory information management systems (LIMS).

Socket communications are also good for transfering information across networks, and so the db handler could run on a different computer to that running CCP4i. This would facilitate the transfer of CCP4i to a distributed computing environment, such as that envisioned in The Grid (for more information about Grid technologies see for example the Global Grid Forum website at http://www.gridforum.org/).

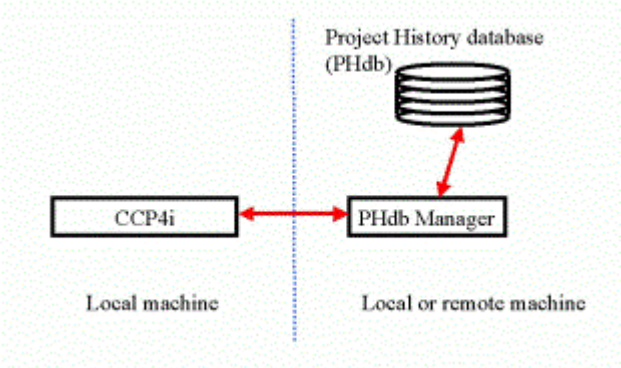


Figure 5: schematic for database interactions usin the db handler

- **Python Run Scripts**

  Recent developments suggest that the Python scripting language is rapidly gaining ground as the language of choice for writing crystallographic computing applications. It is already being used in a number of successful high-profile projects, and future CCP4 programs are likely to involve a significant Python scripting component. (For more information about Python see e.g. http://www.python.org/).

  At present CCP4i can run scripts directly only if they are written in Tcl. Extending the interface to allow run scripts written in Python enables CCP4i to keep up with these new developments as the new Python-based applications begin to emerge.

- **Managing Harvesting Files**

  Currently, harvesting files are created in mmCIF format by a number of CCP4 programs, and store certain information about the dataset being processed which

can be used later in the deposition process (see the CCP4 Harvesting documentation for more information).

Currently there is no way of tracking and validating these files during the structure solution process. A Harvesting File Manager is now under development which will allow users to track and validate the harvesting files, with the ultimate purpose being to make the deposition process simpler and faster.

- **Saphire extensions**

  The SAPHIRE task interface was designed to bring protein sequence information into CCP4 at an earlier stage of structure solution. Currently, protein sequences can be downloaded via this interface in FASTA format and edited if necessary before being saved locally. Extensions are being carried out to redesign the layout of this task and also to provide a graphical interface to running a local copy of CLUSTALW. (Note that CLUSTALW is not distributed by CCP4.)

Acknowledgements

CCP4i was originally developed by Liz Potterton, and Liz contributed the in-line documentation and the python run script functionality.

Pryank Patel is developing the GET_PROT/SAPHIRE application and the Data Harvesting Management tool. The BEAST interface was developed by Anne Baker with contributions from Peter Briggs. The ACORN and WHAT_CHECK interfaces were developed by Maria Turkenburg with the assistance of others including Yao Jia-xing, Eleanor Dodson, Gert Vriend, Liz Potterton and Peter Briggs. Other new interfaces were developed by Peter Briggs and Martyn Winn.

The automation projects were implemented by Martyn Winn, Alun Ashton and Peter Briggs.

Peter Briggs is responsible for the MapSlicer, Task Installer, MTZ Viewer and Project History Database developments. CCP4i is now maintained and developed by the DL CCP4 staff and other fixes and developments are due to them. Please send questions, requests and bug reports to us at ccp4@ccp4.ac.uk.

Development of the CCP4 software library II

Martyn Winn, Charles Ballard, Peter Briggs and Eugene Krissinel November 2002

What it's all about

Behind the scenes, the CCP4 software suite is undergoing a major overhaul, with the strategic aim of bringing the suite into line with the new high-throughput era. As part of this, the CCP4 software library, upon which most of the CCP4 programs depend, is being substantially re-written, and new libraries are being added. These new libraries will support new applications written in a modern object-orientated fashion. They will also act as computational modules for use in a scripting environment - particularly pertinant to automation efforts.

But **Don't Panic!**. All your favourite programs, and others, will continue to be supported. However, with the new framework in place, you will increasingly find new tools, more automation and better data management.

A preliminary account of the new software library was given in Newsletter 40. Therefore, this article will concentrate on recent progress and plans for the next CCP4 release.

What's happening with the core library

MMDB

The MMDB C++ library is designed to assist CCP4 developers in working with coordinate data, as obtained from PDB or mmCIF files. MMDB provides various high-level tools for working with coordinate files, which include not only reading and writing, but also orthogonal-fractional coordinate transformations, generation of symmetry mates, editing the molecular structure and others. Recently added functionality includes:

- A bond manager for adding or removing bonds to a selection of atoms
- The ability to add user-defined data to any level of the coordinate hierarchy
- XML support

A Fortran-callable interface, built on MMDB, replicates the functionality of the old rwbrook.f library file. MMDB also provides model handling for the CCP4 Molecular Graphics project (E. Potterton, S. McNicholas, E. Krissinel, K. Cowtan and M. Noble, *Acta Cryst.* **D58**, 1955-1957 (2002)).

More information can be found on the MMDB project pages (http://msd.ebi.ac.uk/~keb/cldoc/).

CMTZ

The CMTZ library implements the hierarchical view of reflection data:

```
File -> Crystal -> Dataset -> Column -> Reflection data
```

A `Crystal' is essentially a single crystal form, while a `Dataset' is a set of observations on a crystal. 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. CMTZ is a C function library to read/write these extended MTZ files, and to manipulate a data structure representing the above data model. A Fortran-callable interface to CMTZ replicates the functionality of the old mtzlib.f library file.

Older MTZ files will lack the Crystal level of the hierarchy. The new library will assume that each project consists of a single crystal, unless different cell dimensions (recorded for each dataset) indicate the presence of different crystals. As with the earlier introduction of dataset information, it is important that the user establish the correct data model at an early stage, for example by the correct labelling of datasets in MOSFLM. Given a correct data model, software downstream can infer appropriate relationships and thus work in a more automated manner.

The CMTZ library can now work in two modes, one which holds all reflection data in memory, and one which leaves the reflection data on disk for sequential processing. The latter is the traditional method used by Fortran programs, but the former is likely to be more useful for newer applications. The mode can be selected by an environment variable CMTZ_IN_MEMORY.

Recent work has concentrated on testing the Fortran interface and ensuring robust support for existing programs. When this work is completed, our attention will turn to providing new and improved tools. An early target is MTZ file handling in ccp4i which is currently done by interpreting MTZDUMP output. The tcl interface to the new library enables direct access to the MTZ data structure. This is both more robust and allows more advanced graphical manipulation of MTZ files.

CMAP

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. This work is essentially complete.

CSYM

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 syminfo.lib 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).

The new CCP4 library contains C 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. Recently added functionality includes:

- Addition of Cheshire cells to syminfo.lib.
- Reverse lookup by a list of operators, for example as obtained from the header of an MTZ file.

Other library functions

The new CCP4 library also contains a number of other functions which give the traditional look-and-feel of CCP4 programs, for example for parsing CCP4-style keyworded input and for writing the CCP4 banner at the top of the log file. There are also various utility functions which return date, time, program name, user name, etc. This functionality is now available to C-level programs as well as Fortran programs.

The library also retains various Fortran subroutine libraries where conversion is not appropriate or has not yet been attempted, e.g. certain routines in ccplib.f and all of plot84driver.f

Plans for CCP4 5.0

The next major release of CCP4 will include the new libraries as an integral part of the suite. In addition, it is hoped to include Kevin Cowtan's Clipper library and the FFTw Fourier transform library. For the average user, there should be few visible changes. There will be a few additional applications based on the new libraries, for example some coordinate manipulation programs based on the MMDB library. MTZ files will gain the CRYSTAL level of the reflection data model. There will be a graphical viewer for MTZ files which highlights the hierarchical nature of the data.

On the other hand, for developers CCP4 5.0 will provide a more powerful environment for writing applications and complex tasks. The new libraries, together with Clipper and FFTw, provide functionality for writing applications in C++, C or Fortran. In addition, much of this functionality is available to scripts written in python, tcl or perl using the SWIG-generated programming interfaces. Makefiles to be distributed with CCP4 allow the generation of shared libraries which then form loadable modules for a scripting environment.

At the moment, ccp4i executes a job as a separate process running a wish script. From CCP4 5.0, ccp4i will also be able to execute python scripts, with job parameters being saved in the database as usual. With the object-orientated capabilities of python, this allows the creation of more sophisticated, data-orientated tasks within the familiar user environment of ccp4i.

Acknowledgements

The formulation of this library has benefited from many discussions with Kevin Cowtan, York (who also provided some core functions). Alun Ashton (Daresbury) has helped with the Windows port of the library. Nick Sauter (Lawrence Berkeley NL) has given useful feedback on CMTZ. Phil Evans and Eleanor Dodson have tested the Fortran interface.

Daresbury Biology and Medicine College

Note from Gareth Jones, Biology and Medicine Head, Daresbury Laboratory, on the reorganisation of Synchrotron Radiation Department

There are many reasons why a reorganisation of staff and stations within the Synchrotron Radiation Department makes good sense at this particular moment in time. A new Director of SR, John Helliwell, with an expanded role, the Quinquennial Review and the changes in the roles of the Research Councils with regard to access to beamtime and funding for facility developments, and the ground-swell of staff and usersâ€™ opinion for a greater focus on science in light of the fact that the SRS now has a fixed number of years to run. The challenges for the staff and users are clear. Maximising science output from the SRS within heavy budgetary constraints particularly in light of the cost of improving the reliability of the SRS in the run up to Diamond. The resounding message from BioMed staff has been YES to using the new college system as a vehicle to create a more efficient science based culture of helpfulness among college members to the benefit of the biology and medical SR communities.

The new BioMed College will be responsible for a similar portfolio of stations as the old Life Sciences Programmes: all the protein crystallography stations, station 2.1 from non-crystalline diffraction and stations CD12, IR11 and beamline 13 of the old VUV-IR facility group. This is of course for administrative reasons and doesn't mean that biology or medical users are prevented from using the stations of other colleges. In fact, cross-college interactions are strongly encouraged so that we retain the interdisciplinary nature of our science.

What will be new within the college, will be a greater awareness among staff of the skills that reside within BioMed. We have already staged “awareness sessions” and BioMed College strategy away day in early October. The output from these will form the basis of BioMed College business plan. It is clear that we can't do everything we would like to in the department as a whole, we simply do not have the staff time and funding. In the new regime after consultation with users, the SRS will run with a reduced portfolio of stations, but as most of BioMed stations are either new or oversubscribed (mostly both) there is no respite for staff here. This is a healthy state to be in, but our stations are used at a very high user group turnover rate and therefore require much staff time.

Instead, the BioMed plan will have scientific focus to optimise its output. We hope to have a health mix of purely in-house research project and collaborations with external user groups. BioMed is open to collaborative projects and these should be discussed with individual facility scientists. I believe that most visiting scientist are totally unaware of the breadth and depth of expertise in BioMed which stretches way past structural biology to infrared microscopy, electron microscopy, molecular and cell biology and even virology.

The common goal for the BioMed College and our users is to produce the highest quality science output in the most efficient manner.

Handling Reflection Data using the Clipper libraries

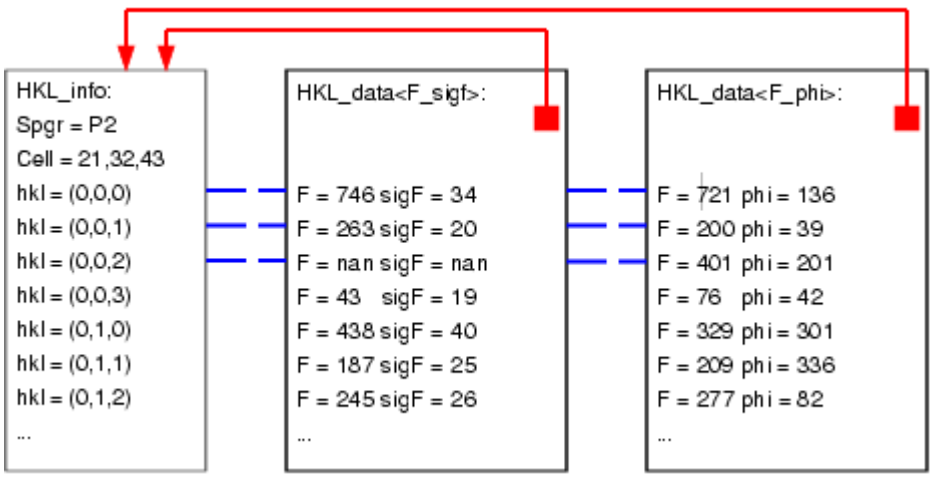
Kevin D. Cowtan, Department of Chemistry, University of York, YO10 5YW

This paper provides a brief introduction to the storage and manipulation of reflection data in the Clipper C++ libraries for crystallographic computation. The library is designed to make all common calculations quick and convenient. In particular, symmetry is handled automatically without any additional code. The library is extensible to new data types which were not available at the compilation of the libraries.

Understanding HKL_info and HKL_data<>

Storage of reciprocal space data in Clipper involved two objects: HKL_info and HKL_data<>. The first stores a list of reflections (HKLs), and the second stores a list of data of some type for each reflection.

Why is this division made? Because often, many types of data will be stored for the same reflection list, and duplicating the list of reflections for each data list would be wasteful. The relationship of an HKL_info object and several HKL_data<> objects is shown below.



These objects handle all the crystallographic symmetry operations associated with storing data in reciprocal space. To the programmer, the data appears to fill a sphere of reciprocal space. However, only a unique set of data is stored, and any changes to a data values are automatically reflected in the symmetry and Friedel related reflections.

The HKL_info object

The HKL_info object stores a list of reflections, transformed into a standard reciprocal space asymmetric unit. For example:

- **P1**: data is stored for l>0 or (l=0 and (h>0 or (h=0 and k>=0)))
- **P2**: data is stored for k>=0 and (l>0 or (l=0 and h>=0))
- **P212121**: data is stored for h>=0 and k>=0 and l>=0

However, the programmer should never need to know which asymmetric unit is being used, as data in other parts of reciprocal space are generated automatically.

The HKL_info object also stores a lookup table, which is used to rapidly determine the location in the list of a given HKL. It also stores lookup tables for the reflection class (HKL_class) and resolution (invresolsq).

The reflection list depends upon the spacegroup (which determines the reciprocal asymmetric unit and systematic absences), the cell parameters, and the resolution limit. Therefore to initialise an HKL_info object, objects of types Spacegroup, Cell and Resolution are passed to the constructor or initialiser.

Reflection classes

The 'class' of a reflection is a function of its HKL and the spacegroup, and describes whether it is centric, systematically absent, and what its allowed phase and symmetry enhancement factor or multiplicity are (i.e. epsilon). The class of a reflection is described by the class clipper::HKL_class.

Reflection resolution

The resolution of a reflection is a function of its HKL and the unit cell parameters. It is generally handled in terms of the inverse resolution squared, or 'invresolsq'. This is equal to 4 $\sin^2$theta/lambda^2.

The HKL_data<> object

The data object is not much more than an array of data, but it has a number of special methods to make crystallographic operations on that data more convenient. It is defined as a template class, where the template type is the type of crystallographic information to be stored in the object. It simply stores a list of data of that type, and a pointer to the parent HKL_info object, which defined the HKL for each element in the list. Additionally a pointer to a Cell object is stored, which may optionally be used for the case where different data comes from slight different unit cells (e.g. RT and frozen data). Therefore to initialise an HKL_data object, an HKL_info object and optionally a Cell object are passed to the constructor or initialiser.

Data types typically include several values. Examples include measured X-ray intensity and its standard deviation (I_sigI), structure factor magnitude and phase (F_sigF), and Hendrickson-Lattmann coefficients (ABCD). Data types are derived from a base type (Datatype_base), and should override all the methods of that type. This will allow the data to be automatically transformed about reciprocal space, and imported or exported to a file, as required.

Methods are provided to access the data by index or by HKL. Any transformations which must be applied to the data in obtaining a symmetry or Friedel related value are applied automatically.

In order to use the class efficiently, some important difficulties must be borne in mind.

A problem arises when we wish to apply some transformation to the values stored in a data list. In this case, we must access every unique value in the asymmetric unit once and once only, applying the desired transformation. Only then will the entire data list have been transformed correctly.

A second problem arises if we want to access the stored value of the data at some position in reciprocal space, for example to expand to a lower spacegroup. Then it is necessary to search through all the symmetry operators, applying each one in turn to the desired HKL to find the operator which brings the HKL into the stored asymmetric unit, with a Friedel inversion if necessary. Clearly this can be time consuming, especially if there are many symmetry operators.

Both these problems are addressed by the use of HKL reference types. These come in two forms:

- index-like references (clipper::HKL_info::HKL_reference_index)
- coordinate-like references (clipper::HKL_info::HKL_reference_coord)

The index-like reference behaves like an index: it stores a reference to an HKL_info object and a position in the reflection list. It is used to loop over all the values in the reflection list, using the HKL_info::first(), and HKL_reference_index::last() and HKL_reference_index::next() methods. The HKL corresponding to the index, its resolution, and reflection class can be returned at any point.

The coordinate-like reference behaves like an HKL coordinate: it stores a reference to an HKL_info and an HKL. However to enhance performance it also stores the index corresponding to that HKL, and the number of the symmetry operator used to get back into the stored asymmetric unit, along with a flag to signify Friedel inversion. Since reflections are usually accessed systematically, the next HKL used will commonly require the same symmetry operator, and so that operator is tried first. Methods are provided for incrementing and decrementing along the h, k, and l directions.

The differences between the index-like and coordinate-like reference types can be summarised as follows:

- index-like types can only refer to the position of a stored datum, i.e. reflection in this list.
- coordinate-like types can refer to any possible position, and therefore also store the symmetry transformations required to get back to the stored data.

HKL reference types may be shared between any data lists which have the same reflection list. It is the responsibility of the programmer to ensure this restriction is obeyed.

HKL_data code fragments

Importing and exporting HKL_data

To import a datalist we need an HKL_info object to hold the reflection list and an HKL_data object to hold the actual data. We also need MTZdataset and MTZcrystal objects to hold the additional information which will be returned from the MTZ file. Then we create an MTZfile object, open it onto a file, and read the reflection list and data.

```
clipper::HKL_info myhkl;            // define objects
clipper::MTZdataset myset;
clipper::MTZcrystal myxtl;
clipper::HKL_data<clipper::data32::F_phi> fphidata( myhkl, myxtl );
```

```
clipper::MTZfile mtzin;
mtzin.open_read( "my.mtz" );        // open new file
mtzin.import_hkl_info( myhkl );     // read sg, cell, reso, hkls
mtzin.import_hkl_data( fphidata, myset, myxtl, "native/CuKa/[FCAL PHICAL]" );
mtzin.close_read();
```

The same process is more elegantly achieved using containers, allowing all the information read from the reflection file to become part of a hierarchy. CMTZcrystal and CMTZdataset objects will be inserted in the hierarchy between the HKL_info and HKL_data objects.

```
clipper::CSpacegroup myspgr();      // define objects
clipper::CCell mycell( myspgr );
clipper::CResolution myreso( mycell );
clipper::CHKL_info myhkl( myreso );
clipper::CHKL_data<clipper::data32::F_phi> fphidata( myhkl );

clipper::MTZfile mtzin;
mtzin.open_read( "my.mtz" );        // open new file
mtzin.import_hkl_info( myhkl );    // read sg, cell, reso, hkls
mtzin.import_chkl_data( fphidata, "native/CuKa/[FCAL PHICAL]" );
mtzin.close_read();
```

Expanding reflection data to a lower symmetry

To expand a list of data to a lower symmetry, we need two reflection lists, one for each spacegroup; and two datalists, one for each reflection list. The lower symmetry list is then filled by looping over all the reflections in that list an requesting the value from the other list for that HKL.

```
clipper::HKL_info oldhkl( .... );
clipper::HKL_data<clipper::data32::f_phi> olddata(oldhkl);
// ---- fill the objects here ----
clipper::HKL_info newhkl( Spacegroup( Spgr_descr( 1 ) ),
                          oldhkl.cell(), oldhkl.resolution() );
clipper::HKL_data<clipper::data32::f_phi> newdata(oldhkl);
HKL_info::HKL_reference_index ih;
for ( ih = newhkl.first(); !ih.last; ih.next() ) {
  newdata[ih] = olddata[ih.hkl()];
}
```

Note that the '.hkl' is vital, as we want the data with the corresponding hkl, not the data from the corresponding position in the list. If efficiency is paramount, using an HKL_reference_coord to access the old list will save some searches over the symmetry operators:

```
clipper::HKL_info::HKL_reference_index ih;
clipper::HKL_info::HKL_reference_coord ik( oldhkl );
for ( ih = newhkl.first(); !ih.last; ih.next() ) {
  ik.set_hkl( ih.hkl() );
  newdata[ih] = olddata[ik];
}
```

Applying simple operations to a whole data list

While it is simple to loop through a reflection list and apply some transformation on the data, some simple operations have been automated using built-in C++ arithmetic

operators for data of specific types, logical operators for data or any type, comparison operators for flags, and function 'Computation operators' for more complex operations.

Arithmetic operators.

Arithmetic operators are defined for the addition, subtraction, and scaling of map coefficients (i.e. HKL_data<datatypes::F_phi>), and for the addition and scaling of Hendrickson-Lattmann coefficients (class HKL_data<datatypes::ABCD>). Thus, to add two lists of map coefficients, the following code is required:

```
clipper::HKL_data<clipper::data32::F_phi> fphi1, fphi2, fphi3;
// ---- set data here ----
fphi3 = fphi2 + fphi1;
```

The columns are added using vector addition. If any values in either list are missing, then the result is missing. Subtraction is similar. Multiplication by a scalar scales the magnitude of every non-missing element in the list.

Logical operators.

Standard C/C++ bitwise logical operators (&, |, ^, !) may be applied to any data list. For each data in the list, the value 'true' will be returned if the data is not missing, or false if it is missing. The result of the operation is a new data list of type HKL_data<Flag_bool>, containing the results of the logical operation. This may be used in further logical operations, or may be used as a mask to eliminate data from a list using the HKL_data::mask() method.

Comparison operators.

Comparison operators (==, !=, >, <, >=, <=) may be applied to a data lists of flags (i.e. HKL_data<datatypes::Flag>), to compare the values in the list with a single integer. This is commonly used in the handling of Free-R test sets. The result is a list of HKL_data<Flag_bool>, where the value of the flag for each reflection is the result of the compraison of the frag for that reflection and this given integer. So, for example, to make a list of data containing only the values for which the test set is numbered 18 or greater, use the following code:

```
clipper::HKL_data<clipper::data32::F_sigF> fsigf, fsigftest;
clipper::HKL_data<clipper::data32::Flag> flag;
// ---- set data here ----
fsigftest = fsigf;
fsigftest.mask( flag >= 18 );
```

Computation operators

Computation operators handle more complex crystallographic tasks, and will be discussed in more detail.

To use a computation operator, call the compute() method of the destination datalist. This method must be supplied with one or two source datalists, and a computation operator. This is an object which performs the computation for an individual reflection, and is usually constructed on the fly.

Some computation operators simply convert a datalist of one type to a datalist of another type. For example, you can convert a phase and weight to Hendrickson Lattmann coefficients. (Of course C and D will be 0, because a phase and weight can only describe a unimodal distribution).

```
clipper::HKL_data<clipper::data32::Phi_fom> myphifom;
// ---- set data here ----
clipper::HKL_data<clipper::data32::ABCD> myabcd;
myabcd.compute( myphifom,
  clipper::data32::Compute_abcd_from_phifom() );
```

Some computation operators take data from two datalists. For example, you can calculate map coefficient (magnitude and phase) from a set of observed magnitudes and a phase and weight:

```
  clipper::HKL_data<clipper::data32::F_sigF> myfsig;
  clipper::HKL_data<clipper::data32::Phi_fom> myphifom;
  // ---- set data here ----
  clipper::HKL_data<clipper::data32::F_phi> myfphi;
  myfphi.compute( myfsig, myphifom,
    clipper::data32::Compute_fphi_from_fsigf_phifom() );
```

Computation operators may operate on a datalist itself, and can also take parameters. These parameters are passed to the constructor of the computation operator. For example, to apply a scale factor of 2.0 and and U-value of 0.5 to a list of reflections, use the following code:

```
  clipper::HKL_data<clipper::data32::F_sigF> myfsig;
  // ---- set data here ----
  myfsig.compute( myfsig,
    clipper::data32::Compute_scale_u<data32::F_sigF>( 2.0, 0.5 ) );
```

Computation operators are provided for performing addition, subtraction and scaling of map coefficients (i.e. F_phi), addition and computation of Hendrickson Lattmann coefficients, and for scaling data. It is fairly simple to define new computation operators, see core/hkl_convops.h.

Defining a new reflection datatype

Several data types are defined in the file hkl_datatypes.h . Defining a new type proceeds as follows:

1. Define a struct containing the data which needs to be stored for each reflection. A default constructor should be supplied which initialises all the data to NaN for floats, or an illegal value for ints (e.g. -1 for Free-R flag).
2. Defined a member function `void friedel()' which chages the values of the data to the values of the Friedel opposite of the data. (e.g. a magnitude is unchanged, a phase will be negated).
3. Defined a member function `void shift_phase(const float)' which chages the values of the data to the value of a symmetry equivalent with the given phase shift from the original. (e.g. a magnitude is unchanged, a phase will have the shift added to it).
4. Define a member function `static const string type()' for that struct which returns a `type name' string for this type. This is used to identify the data type and to infer column names in an mtz file.

For example, an F_phi group are defined as follows:

```
struct F_phi
{
  float f,phi;
  F_phi() { f=phi=Nan(); }
  static const String type() { return "F_phi"; }
  void friedel() { phi=-phi; }
  void shift_phase(const ftype dphi) { phi+=dphi; }
  const bool missing() const { return (isnan(f) || isnan(phi)); }
};
```

The datalist types are constructed from the individual data type by a template class.

If you need to store your new datatype in an MTZ file, you must also define an MTZ_iotype by derivation from clipper::MTZ_iotype_base, create a static instance of the new MTZ_iotype, and add it to the mtz_iotype_registry.

Further reading

See the Clipper documentation at http://www.ysbl.york.ac.uk/~cowtan/clipper/clipper.html.

Atomic displacement in incomplete models caused by optimisation of crystallographic criteria

P.V. Afonine

LCM3B, UMR 7036 CNRS, Université Henri Poincaré, Nancy 1, B.P. 239, Faculté des Sciences, Vandoeuvre-lès-Nancy, 54506 France, and

Centre Charles Hermite, LORIA, Villers-lès-Nancy, 54602 France

Abstract

It is known that incompleteness of the atomic model can seriously affect its refinement. In particularly, the minimisation of a crystallographic criterion (least-squares or maximum-likelihood) shifts the atoms of an incomplete model from their exact positions. The more incomplete model, the larger the mean atomic displacement. This article studies individual atomic displacements in such a model.

Introduction

The basic goal of a crystallographic refinement is to obtain a model that is consistent as much as possible with the experimental diffraction data. For example, the conventional least-squares refinement fits structure factor modules calculated from the model to the experimental values. This goal is justified when one deals with a complete model which practically is never the case. Even at late stages of refinement some fragments with high B factors, some solvent molecules and often the bulk solvent are not taken into account. For such models, the structure factor magnitudes calculated from the exact model are different from observed amplitudes even in an ideal case without experimental errors. As a consequence, in the test case when initially the atoms of an incomplete model are placed correctly, the minimisation of the least-squares criterion without stereochemical restraints shifts them from their correct positions (Afonine *et al*, 2001; Lunin *et al*, 2002).

This negative effect can be reduced if the maximum-likelihood approach is used (Pannu & Read, 1996; Bricogne & Irwin, 1996; Murshudov *et al.*, 1997; Adams *et al*., 1997) which takes into account the absent part of the model statistically. Nevertheless, even in this case the mean atomic displacement can be quite large (Lunin *et al*, 2002).

A series of tests was carried out in order to study the distribution of an atomic displacement over the model.

Numerical tests

The tests were carried out with CNS complex (Brünger *et al*., 1998) using the structure of Fab fragment of monoclonal antibody (Fokine *et al*., 2000). This molecule crystallises in space group $P2_12_12_1$ with the unit cell parameters a = 72.24Å, b = 72.01Å, c = 86.99Å and one Fab molecule per asymmetric unit. The full model includes 439 amino acid residues and 213 water molecules. The observed structure factors were simulated by the corresponding values calculated from the complete exact model in order to exclude experimental errors from the analysis. The standard least-squares criterion *LS* was used in the tests. The minimisation of this criterion was performed till the convergence independently at the resolution d_{min}>2.2Å (the resolution at which the model was constructed). For comparison, the second series of tests was done at d_{min}>1.3Å. Two

incomplete starting models were generated by random deletion of approximately 3 and 20 % of atoms, both macromolecular and water oxygens. In all tests, the atoms of such incomplete models initially were placed at their correct positions.

For each atom of an incomplete model two distances were calculated:
1) the distance between this atom in the starting model and the nearest removed atom;
2) a similar distance in the model after minimisation.

Results and discussion

Fig. 1 shows the distribution of distances between each atom of an incomplete model and the former position of the closest deleted atom before and after minimisation of the *LS* criterion. The maximal shift corresponds to the atoms situated in the sphere of approximately 2.4 Å radius around the deleted atom. These are atoms covalently bonded to it or located within the van der Waals distance. Most of such atoms shift towards the positions of deleted atoms (points below the straight diagonal line; final distance is smaller that the initial distance). For the atoms situated far away from the deleted atom, the shifts are much smaller and less regular, both to and from the deleted atom.

The separated bars at the left side of Fig. 1 correspond to specific pairs of bonded atoms one of which was deleted. For example, the bar 1 corresponds to the double bond atoms C=O where either C or O atom was deleted, the bar 2 is for the C–N bond, etc. The bar 6 corresponds to the C–S bond.

In the case of the 3%-incomplete model (Fig. 1a), the tendency of atoms to move toward the place, previously occupied by a deleted atom, is seen much better in comparison with the case of 20%-incomplete model (Fig 1b). A possible reason may be that in the latter case it is more difficult to ‘choose’ the direction of its shift because of a large number of ‘holes’ in the structure. It can be concluded that the crystallographic criterion taken alone, without stereochemical restraints, shifts the atoms of an incomplete model mostly toward the positions of deleted atoms trying to compensate their absence. Stereochemical restraints when used allow reduction of such a displacement by cancelling differently oriented shifts from linked atoms.

In the second series of tests, when the minimisation of the criterion was carried out at a higher resolution of 1.3 Å, the behaviour of atoms of the partial model was the same (Fig. 1c and 1d).

References

Adams, P.D., Pannu, N.S., Read, R.J., & Brünger, A.T. (1997). *Proc.Natl.Acad.Sci.USA.,* **94**, 5018-5023.

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

Bricogne, G. & Irwin, J. (1996). *Proceedings of the CCP4 Study Weekend*, 85-92.

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

Lunin, V.Y., Afonine P.V. & Urzhumtsev, A. (2002). *Acta Cryst.* A**58**, 270-282.

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

Pannu, N.S. & Read, R.J. (1996). *Acta Cryst.*, A**52**, 695-668.

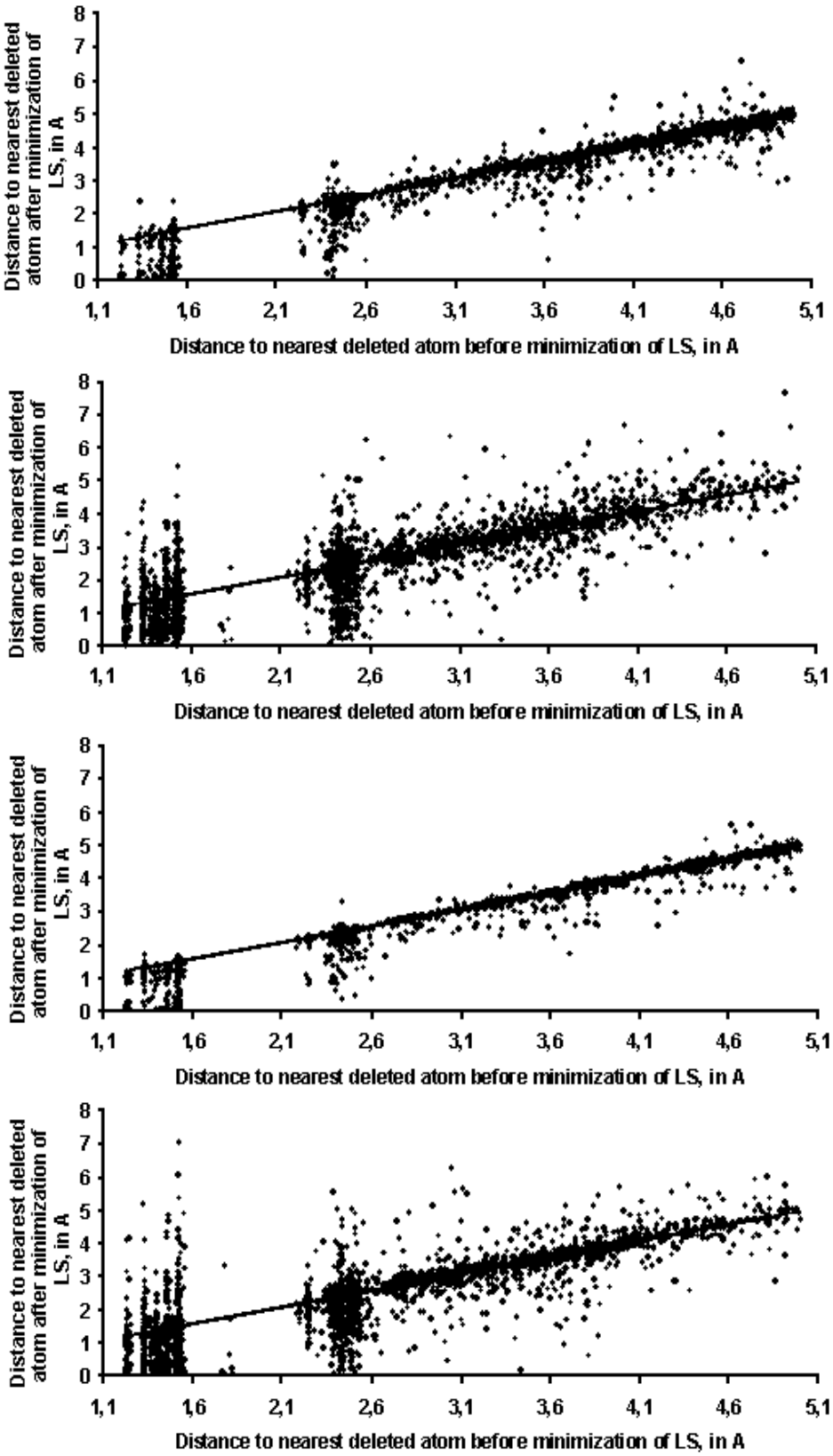


Fig. 1. Distribution of distances between each atom of an incomplete model and the closest deleted atom, before and after minimisation of *LS*. The results for two models with different incompleteness are shown: (a) approximately 3% of atoms are deleted randomly, and (b) approximately 20% of atoms are deleted randomly. *LS* was minimised at the resolution of 2.2Å. The numbered arrows correspond to the pairs of bonded atoms one of which was deleted, namely 1 is for C=O bond, 2 is for C-N bond, 3 is for atoms of side chains, 4 is for C^{α}-N bond, 5 is for C^{α}-C or C^{α}-C^{β} or C^{β}-C^{γ} bonds, 6 is for C-S bonds and 7 is for C^{γ}-S^{δ} (Met) or C^{ε}-S^{δ} (Met) or C^{β}-S^{γ} (Cys), N–O (for N or O atoms of main chain) bonds. (c) and (d) are the same as (a) and (b), respectively, but shown for the refinement carried out at the resolution of 1.3 Å.

Modelling of bond electron density by Gaussian scatters at subatomic resolution

P. Afonine [1,2], V. Pichon-Pesme[1], N. Muzet[1], C. Jelsch[1], C. Lecomte[1] & A. Urzhumtsev[1]

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

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

Abstract

At resolutions of 1-2 Å and lower, traditional for macromolecular crystallography, the electron density of the crystal unit cell is modelled by a sum of contributions from individual atoms (plus the bulk solvent). The contribution of each atom is calculated as a spherical or an elliptic function centred at the atomic position. Such models are not adequate at a subatomic resolution, higher than 1 Å, as proven by residual Fourier maps. In particular, these maps show extra density peaks between the bonded atoms. A modelling of this density by placing extra 'gaussian' electrons at the corresponding positions allows the improvement of crystallographic criteria. This new model allows also the determination of some important characteristics of the crystal such as critical points of the electron density distribution.

1. Introduction

Crystal models play several roles. First, they give a possibility to describe the content of a crystal by a finite number of parameters, usually with a clear physical meaning. Secondly, these models allow the determination of properties which are not available directly from the experiment. For example, model parameters can be used to calculate the electron density distribution while the diffraction experiment gives only a limited set of its Fourier coefficients and allows thus the calculation of Fourier syntheses only at a limited resolution. Naturally, such a model should be quite precise and detailed.

In macromolecular crystallography, for many years the usual resolution of the experimental data was about 2-3 Å. At such a resolution, the models, composed from isotropic atoms and refined using crystallographic restraints or constraints, give a good description of experimental diffraction data. During last years, many macromolecular structures were reported at atomic resolution, about 1 Å (for a review, see Dauter *et al.*, 1997; Longhi *et al.*, 1998). Such data require more detailed models. In particular, anisotropic thermal displacement parameters become necessary. An introduction of extra parameters is justified by increased number of structure factor magnitudes as confirmed by R-free (Brünger, 1992) calculations. Recently, several cases were reported where macromolecular crystals diffracted to the resolution higher than 0.9 Å. An example is aldose reductase (Lamour *et al*., 1999) crystals of which diffract to 0.64 Å even when the size of this protein is quite large, 315 amino acid residues. At such a subatomic resolution, the spherical electron density model even with anisotropic thermal parameters is already

inadequate. In particular, difference maps show significant density peaks around the atoms, essentially at the interatomic bonds (Housset *et al*., 1994; Lamzin *et al*., 1999). On the other hand, a large amount of data could allow the detailed study of physical properties of electron density of the crystal.

In small molecules crystallography such subatomic resolution is usual and the problem was solved by introduction of multipolar models (Stewart, 1969; Hansen & Coppens, 1978). Here the density of each atom is modelled by a function:

$$\rho_{atom}(\mathbf{r}) = \rho_{core}(\mathbf{r}) + P_{val}\kappa'^3\rho_{val}(\kappa\mathbf{r}) + \sum_{l=0}^{l_{max}}\kappa'^3 R_l(\kappa'\mathbf{r})\cdot\sum_{m=-l}^{l} P_{lm}y_{lm}(\theta,\varphi)$$

where the total atomic electron density (r_{atom}) is decomposed into three terms corresponding to the core electrons, valence electrons and non-spherical part of the valence electron distribution as a multipole density (see Hansen & Coppens, 1978, for details and notation). The number of parameters for each atom grows with the number of included spherical harmonics. For traditional models, a multipolar atom, different from a hydrogen, can have up to 28 parameters (18 multipolar, 3 positional, 6 temperature parameters, 1 occupancy). Recently, Pichon-Pesme *et al.* (1995) and Jelsch *et al*. (1998) demonstrated the transferability of such models to macromolecular crystals when a large enough number of structure factor magnitudes allows this (in other words, when the resolution of the diffraction data set is high enough). Such models have been successfully refined for crystals of crambin (Jelsch *et al*., 1999), for a scorpion toxin (Housset *et al*., 2000) and for an aldose reductase complex (Guillot *et al*., 2000), respective resolution of these crystals is 0.54, 0.9, and 0.65 Å.

Some other protein crystals, for example those of other complexes of aldose reductase, diffract to slightly lower resolution, about 0.9 Å. Use of multipolar atoms and refinement of hydrogen atoms reduces the ratio $N_{data}/N_{parameters}$ to the limits when such type of a model can be hardly used. The major risk is, as in any modelling, to overfit the data introducing too many parameters in the model.

In order to answer this problem, a project was started to develop a simplified molecular model which is more detailed than a usual model of anisotropic atoms but contains less parameters than a multipolar model. This model can be used when the amount of experimental data is not sufficient enough to use multipolar models. Additionally, such a model would accelerate the calculations of crystallographic values (structure factors, electron density, and electrostatic potential) which are quite time consuming when multipolar models are employed for macromolecules.

2. Model of bond electrons

$$\Delta\rho_{def} = V^{-1}\sum_{\mathbf{h}}\left(k^{-1}F_{obs}(\mathbf{h})e^{i\varphi_{best}} - F_{sph}(\mathbf{h})e^{i\varphi_{sph}}\right)e^{-2\pi i\mathbf{hr}}$$

As traditionally for macromolecular crystals, incorrect parts of an available model are highlighted in difference density maps. In particular, so called ‘deformation density maps’ can be calculated as the Fourier series

Here V – unit cell volume, k – scale factor, $F_{obs}(\mathbf{h})$ – observed structure factor amplitudes, $\varphi_{best}(\mathbf{h})$ – best phases available, $F_{sph}(\mathbf{h})$ and $\varphi_{sph}(\mathbf{h})$ – structure factors amplitudes and phases obtained from spherical atomic model. These maps show an excess and a missed part in the available model by negative and positive peaks whose size does not allow seeing them at lower resolution. Such maps show the redistribution of the electron density due to formation of interatomic bonds and other interactions. Among various ‘deformations of density’, large peaks are seen most clearly on the bonds (Fig. 1A). This extra density, which represents reorganisation of electrons to bonding, seems to be the major need for a necessary model modification.

Previously, several attempts have been done (Ewald & Höln, 1936a, 1936b); Brill, 1960; Hellner, 1977; Dietrich & Scheringer, 1978; Scheringer, 1980; Pietsch, 1981; Pietsch & Unger, 1981; Pietsch, 1985; Pietsch *et al*, 1986) to model this density by placing there an additional scattering matter, a kind of pseudo atom. All these attempts were done with small-molecular crystals with a small number of diffraction intensities. On contrary, macromolecular crystals have a very large number of reflections at such a subatomic resolution. This allows an easy application of the R-free methodology (Brünger, 1992) which has been shown as a powerful tool to indicate the data overfitting. First checks (Cetin *et al*., 2000) showed a feasibility of this modelling.

3. Preparation of the test model

More detailed tests were done using the leu-enkephalin peptide (Wiest *et al.*, 1994). This pentapeptide (Tyr^1-Gly^2-Gly^3-Phe^4-Leu^5) crystallises in space group $P2_12_12_1$ with unit cell parameters a = 10.851 Å, b = 13.095 Å, c = 21.192 Å and Z = 4. The diffraction data were collected to 0.43 Å. However, the tests described below were done at a lower resolution, d > 0.56 Å (sinq/l < 0.89 $Å^{-1}$), beyond which the completeness of data is insufficient (Sheldrick, 1990), see Fig. 2. At the resolution of 0.56 Å the number of independent reflections is 8707 and completeness for the last high-resolution shell is higher than 60 %. The enkephalin model contains 43 non-hydrogen and 43 hydrogen atoms. A refined multipolar model (M1) was obtained (Wiest *et al.*, 1994) using the program MOLLY (Hansen & Coppens, 1978). This refinement was done without separation of the data set into work and test (*R*-free) subsets.

Because the *R*-free values were important for our tests (one of the indicators of modelling progress) and in order to estimate the quality of previously refined model (M1) in terms of 'R-free', a usual procedure was applied (Brünger, 1993). First, a test set was chosen, composing 20% of the total amount of reflections belonging to 0.56-11.0 Å resolution. This selection was done randomly and uniformly in several resolution shells. Random and independent errors were introduced into atomic coordinates of M1 model by performing molecular dynamics simulation at 2000K with subsequent energy minimisation using CNS (Brünger *at al*, 1998). Such procedure removed some of the 'memory' of previous refinement towards test set and gave the model differing from M1 by ~0.06 Å shift in coordinates. Then, this model was refined using only the 80% rest of reflections (work set) by the program MOPRO (Guillot *et al*., 2001) at the exactly same conditions as previously, reported by Wiest *et al* (1994). The desired R / R_{free}-factor statistics for the model obtained (M2) is summarised in Table 1.

4. Test of bond electron models

All the subsequent tests were done with the program suite SHELX (Sheldrick & Schneider, 1997). No sterochemical constraints were used in these tests.

First of all, the conventional anisotropic refinement of enkephalin was done at 0.56-11.0 Å resolution in order to have reference values for such type of models at this resolution (model A1 in Table 2). It is not surprising that the results of standard anisotropic refinement (R = 9.08 %, R_{free} = 9.74 %) are worse than those after multipolar refinement (R = 7.90 %, R_{free} = 8.63 %; M2 model in Table 1). This tendency is true in every resolution shell (Fig. 3).

For the first test with bond electron (BE in what follows) models, standard hydrogen atoms were used. These atoms were placed in the middle of interatomic bonds (as in Fig. 1B). The optimal starting values of occupancies for BEs were found as 0.5, agreeing with previous tests (Cetin *et al*., 2000) and the starting values of isotropic temperature factors

were taken equal to the average temperature parameters of the neighbouring atoms. Different set of parameters were refined and various refinement strategies were tested. In many cases, including the case of anisotropic BEs, the *R*-free criterion increased or refinement was unstable (details of these numerous tests will be discussed elsewhere). The best improvement in both *R* and *R*-free values, to 8.43 and 9.16 % (model BE1 in Table 2), was obtained when refining the following parameters:

- for each non-hydrogen atom of the model: *coordinates, anisotropic temperature factor and occupancy*;
- for each hydrogen atom of the model placed according to ideal stereochemistry: *isotropic temperature factor (while occupancy is fixed equal to 1)*;
- for each bond electron (BE): *isotropic temperature factor and occupancy*.

It is important to note that this refinement decreased the occupancies of the enkephalin atoms showing the redistribution of the density to newly placed BEs.

The number of BEs is approximately equal to the number of bonds (some extra BEs can be also introduced for lone pair electrons) or, in other words, is approximately equal to the total number of atoms in the molecule. Therefore, the number of additional parameters in BE-model can be estimated as 2 parameters per atom, resulting in 12 parameters per atom in total.

The model was improved when the BEs, taken previously as hydrogens, were replaced by Gaussian scatterers defined from DFT calculations by program SIESTA (Sanchez-Portal *et al.*, 1997) as follows. First of all, the theoretical deformation density maps were obtained for all types of residues as the difference between the exact electron density distribution calculated by quantum-chemical methods and that calculated from the model using standard scattering factors for neutral atoms. For these calculations, each idealised single residue was taken as an isolated molecule in gas phase. The peaks at the bonds were approximated by a single-gaussian function, and the corresponding distance between the peak centre and the neighbouring atoms was calculated. Then, such gaussian peaks were placed in the covalent bond positions of enkephalin. Similarly to the previous case, the starting values of isotropic temperature factors were taken equal to the average temperature parameters of the neighbouring atoms. The refinement of this model decreased the *R* and *R*-free factors to 8.12 and 8.76 % (model BE2 in Table 2). The behaviour of *R*-factors as a function of resolution is shown in Fig. 3.

Summarising, it should be noted that the proposed BE-model:

- is better then the best anisotropic model;
- requires about 12 parameters per atom, in comparison with 10 parameters for the anisotropic model and 28 for a multipolar model;
- gives values of overall *R*- and R_{free}-factors close to that for the refined multipolar model.

5. Further validation of bond electrons models

A lower value of the *R*-free factor is a good proof of a higher quality of BE-models in comparison with classical anisotropic models. However, more tests were done to confirm the quality and physical meaning of the BE-model.

First, the rigid-bond test (Hirshfeld, 1976) showed that BEs do not cause unphysical perturbations in the thermal parameters of atoms. A refinement even without rigid-bond constraints gave the rigid-bond criterion at the same level as for the multipolar model (Fig. 4). An inclusion of the rigid-bond criterion into refinement (not shown) improved the

similarity of the projections of thermal ellipsoids on the bond without increasing the *R* and *R*-free values.

The second check consisted in verification of a predicting power of such BE-models. Multipolar models allow the calculation of accurate electron density. In small-molecule crystallography, these electron density maps are called 'experimental' because the models used for their generation are obtained from the experimental data. One of the important features of this distribution is critical points, where the gradient of the density is equal to zero (Bader, 1990). In particular, these points allow the characterisation of atomic interactions. Each such point is characterised by three numbers which are the eigenvalues of the normal matrix of the electron density calculated at this critical point (the matrix of second derivatives with respect to three coordinates). These values allow one to characterise the type of covalent bonds. A special role is played by density Laplacian which is the sum of the three eigenvalues.

In order to obtain the parameters of critical points for BE-models, a special program was written allowing a very fast calculation of the electron density map, exact maps of the gradient of electron density and that of the density Laplacian directly from the BE model. Electron density maps calculated from the best available isotropic and anisotropic models did not reproduce the critical points derived from the multipolar model. On the contrary, the BE2 model gave both the correct position of the critical points and, for most of bonds, the eigenvalues were very close to those obtained from the multipolar model (Table 3). A few cases with incorrect eigenvalues, for example that for the C=O bonds, prove that for some types of bonds the parameters of BEs, including their position, should be improved in further studies.

6. Conclusions and discussions

Several independent criteria confirmed that modelling of the density on the interatomic bonds plays the major role in the improvement of anisotropic models on the way to multipolar models. The simplest way to improve an anisotropic model is its completion by appropriate gaussian scatters for bond electrons. The available software does not allow the complete refinement of parameters of these electrons as it seems to be necessary. Nevertheless, by refining only the shape and not the position of these BEs, *R* and *R*-free factors can be significantly decreased in comparison with anisotropic models. Moreover, our tests demonstrated that the refinement of BE-models gives the values of *R*- and R_{free}-factors similar to those for multipolar models.

These models reproduce the experimental structure factor values better thus allowing the calculation of improved Fourier syntheses using the experimental structure factors. In addition, such models provide electron density maps that cannot be obtained using isotropic or anisotropic models. The maps reproduce important features of the electron density distribution, for example, its critical points.

A slight decreasing in occupancies of C, N and O atoms in the refined BE-model indicates the tendency to conserve the total charge of the model without electroneutrality constraint; this confirms physical meaning of such a modelling.

The BE-models are described by a smaller number of parameters than multipolar models (approximately 2 times less) and therefore can be used at lower resolution when the number of experimental structure factor magnitudes is smaller. In addition, gaussian scattering factors of all atoms of the model (a 5-gaussian approximation of atomic scattering factor works well at least up to the resolution of 0.25 Å, see International Crystallographic Tables, 1998) should allow a very fast and direct calculation of the exact maps of the gradient of electron density and the maps of the density Laplacian.

The refinement of BE-models does not require any special refinement programs. Nevertheless, more sophisticated software would allow the use of special criteria and refinement strategies.

The work is in progress.

Acknowledgment

The authors thank C. Katan, M. Souhassou, N.-E. Ghermani for useful discussions of high-resolution features of crystals, I. Uson and G. Sheldrick for their help with the use of SHELX program suite and V. Lunin for fruitful discussions. The authors are participants of the GdR 2417 CNRS.

References

Bader, R. W. F. (1990). *Atoms in Molecules. A Quantum Theory.* Oxford University Press.
Brill, R. (1960). *Acta Cryst.*, **13**, 275-276
Brünger, A. T. (1992). *Nature*, **355**, 472-475
Brünger, A. T. (1993). *Acta Cryst.*, D**49**, 24-36
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.
Cetin, A., Pichon-Pesme, V., & Urzhumtsev, A. (2000). *University of H. Poincaré, Nancy 1, LCM3B; internal report*.
Dauter, Z., Lamzin, V. S. & Wilson, K.S. (1997). *Curr. Opin. Struct. Biol.*, **7**, 681-688
Dietrich, H. & Scheringer, C. (1978). *Acta Cryst.*, B**34**, 54-63
Ewald, P. P. & Höln, H. (1936a). *Ann. Phys.*, Lpz.[5], **25**, 281
Ewald, P. P. & Höln, H. (1936b). *Ann. Phys.*, Lpz.[5], **26**, 173
Guillot, B., Jelsch, C., Muzet, N., Lecomte, C., Howard, E., Chevrier, B., Mitschler, A., Podjarny, A., Cousson, A., Sanishvili, R. & Joachimiak, A. (2000). *Acta Cryst.*, A**56** (Supplement), s199
Guillot, B., Viry, L., Guillot, R., Lecomte, C. & Jelsch, C. (2001). *J. Appl. Cryst.*, **34**, 214-223
Hansen, N. K. & Coppens, P. (1978). *Acta. Cryst.*, A**34**, 909-921
Hellner, E. (1977). *Acta Cryst.*, B**33**, 3813-3816
Hirshfeld, F. L. (1976). *Acta Cryst.*, A**32**, 239-244
Housset, D., Habersetzer-Rochat, C., Astier, J.-P. & Fontecilla-Camps, J. C. (1994). *J.Mol.Biol.,* **238**, 88-103
Housset, D., Benabicha, F., Pichon-Pesme, V., Jelsch, C., Maierhofer, A., David, S., Fontecilla-Camps, J. C. & Lecomte, C. (2000). *Acta Cryst.*, D**56**, 151-160
Jelsch, C., Pichon-Pesme, V., Lecomte, C. & Aubry, A. (1998). *Acta Cryst.* D**54**, 1306-1318
Jelsch, C., Teeter, M. M., Lamzin, V., Pichon-Pesme, V., Blessing, R. H. & Lecomte, C. (2000). *PNAS*, **97,** no. 7, 3171-3176
Lamour, V., Barth, P., Rogniaux, H., Poterszman, A., Howard, E., Mitschler, A., Van Dorsselaer, A., Podjarny, A. & Moras, D. (1999). *Acta Cryst.* D**55**, 721-723
Lamzin, V., Morris, R.J., Dauter, Z., Wilson, K.S. & Teeter, M. M. (1999). *J.Biol.Chem.*, **274,** 20753-20755.
Longhi, S., Czjzek, M. & Cambillau, C. (1998). *Curr. Opin. Struct. Biol.*, **8**, 730-737
Pichon-Pesme, V., Lecomte, C. & Lachekar, H. (1995). *J. Phys. Chem.,* **99**, 6242-6250.
Pietsch, U. (1981). *Phys. Stat. Sol., (b)* **103**, 93
Pietsch, U. & Unger, K. (1981). *Phys. Stat. Sol., (b)* **104**, 253
Pietsch, U. (1985). *Phys. Stat. Sol., (a)* **87**, 151
Pietsch, U., Tsirelson, V. G. & Ozerov, R. P. (1986). *Phys. Stat. Sol., (b)* **138**, 47-52
Sanchez-Portal, D., Ordejon, P., Artacho, E. & Soler, J.M. (1997). *Int. J. Quant. Chem*., **65**, 453.
Scheringer, C. (1980). *Acta Cryst.*, A**36**, 205-210
Sheldrick, G. M. (1990). *Acta Cryst.*, A**46**, 467-473
Sheldrick, G. M. & Schneider, T. R. (1997). *Methods in Enzymology*, **277**, 319-343
Stewart, R.F. (1969). *J. Chem. Phys.*, **51**, 4569
Wiest, R., Pichon-Pesme, V., Bénard, M. & Lecomte, C. (1994). *J. Phys. Chem.*, **98**, 1351-1362

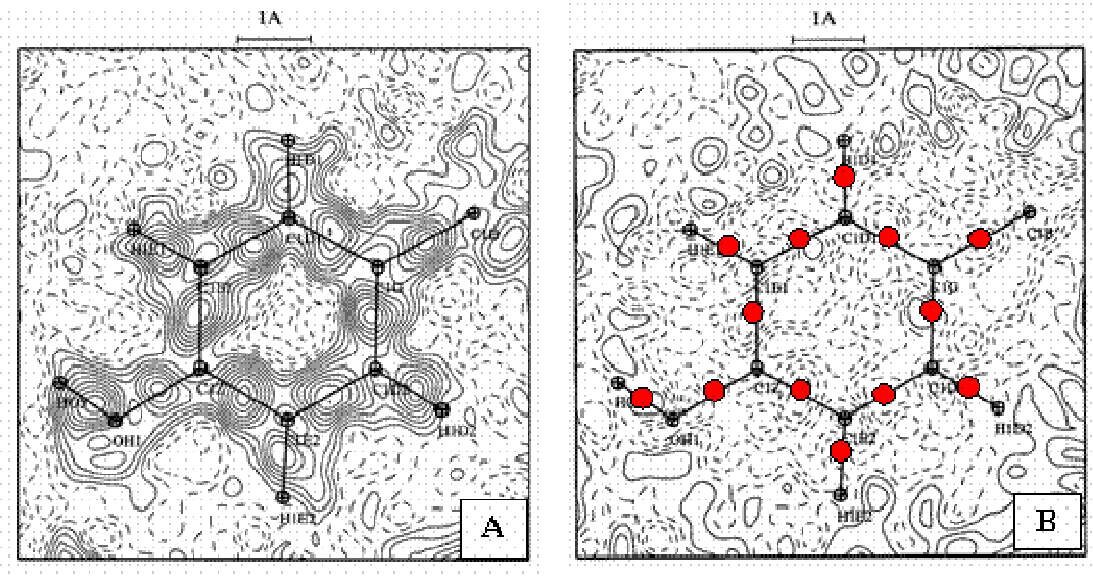


Fig. 1. (A) The model of spherical atoms and corresponding deformation density map showing the extra electron density peaks at interatomic bond distances. Such peaks are subject to modelling at resolution higher than 0.9 Å. **(B)** The same model completed by 'bond electrons' (red circles) placed between chemically bonded atoms and difference Fourier map after refinement of this BE-model. The maps calculated at 0.45 Å resolution show the Tyr-side chain of the YGG polypeptide (Cetin *et al.*, 2000).

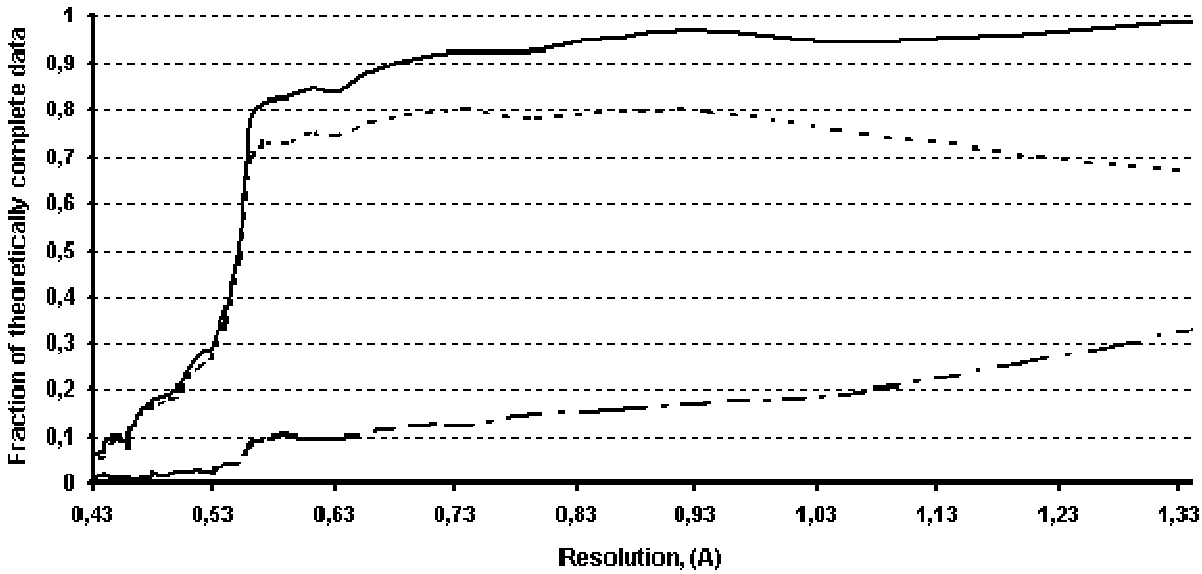


Fig. 2. Fraction of theoretically complete data: dashed line for acentric, dashed-dotted for centric reflections accordingly, bold line for total completeness.

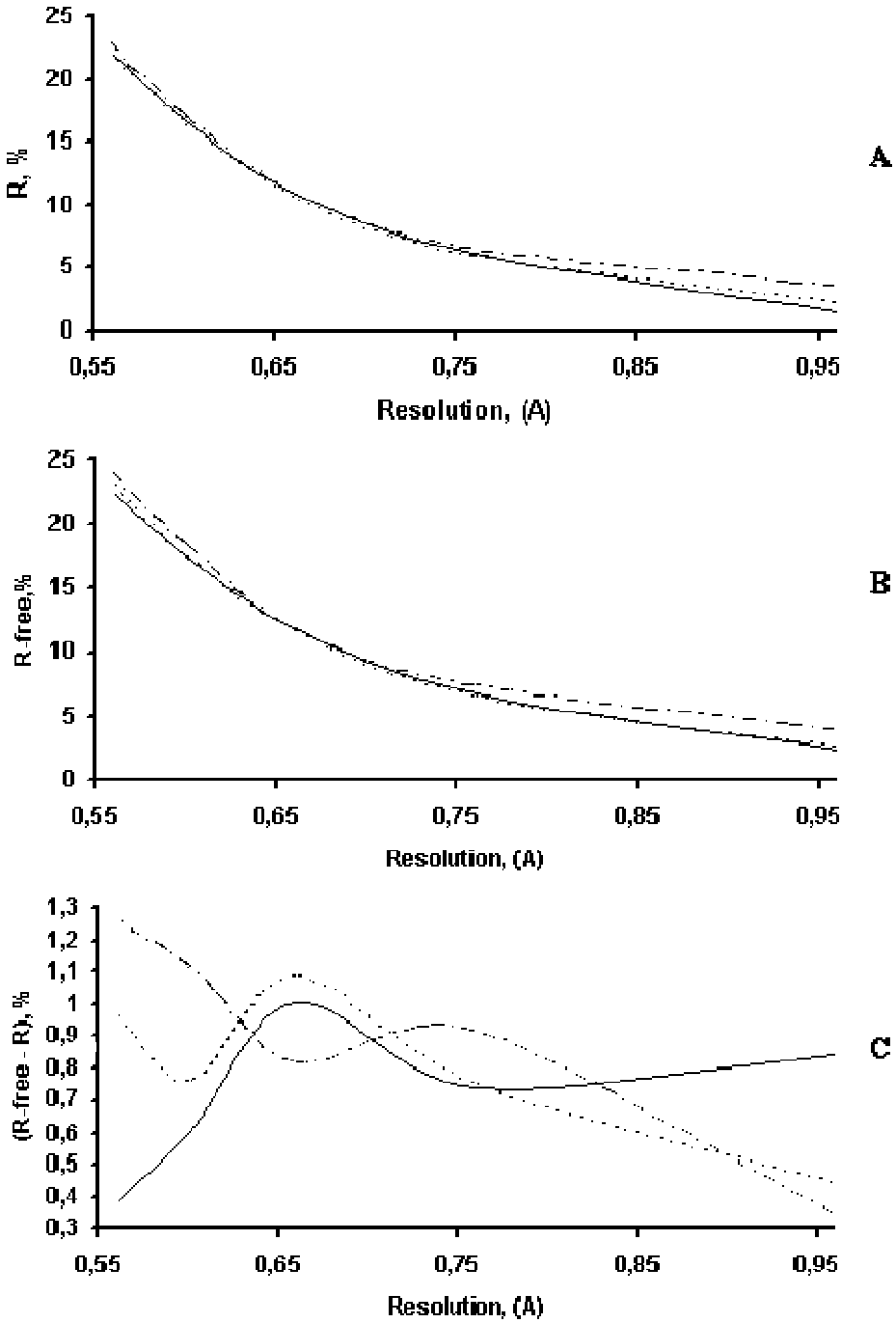


Fig. 3. Graphics A, B and C show respectively R-, R_{free}- and (R_{free}-R)-factors as a function of resolution. Solid line: multipole model; dotted line: BE-model; dashed-dotted line: standard anisotropic model.

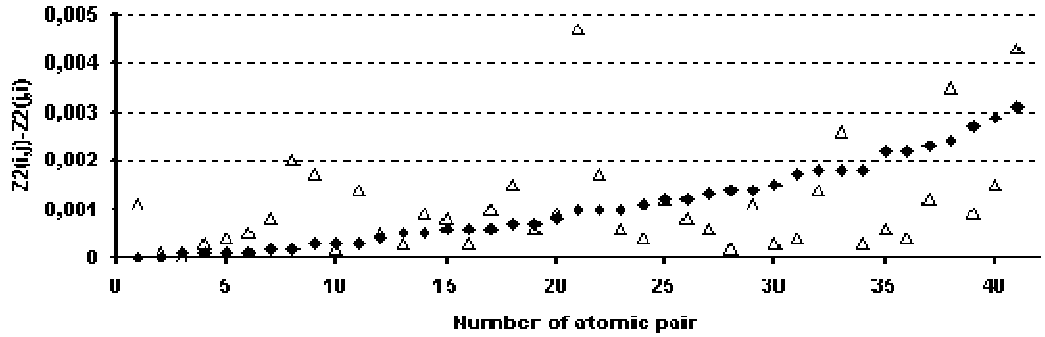


Fig. 4. Results of Hirshfeld's rigid bond test. For every covalently bonded pair of atoms *i* and *j*, the discrepancy of the mean square vibration amplitudes $Z^2(i,j)-Z^2(j,i)$ along their mutual bond is presented as a function of index number for each bond. The bonds are given in increasing order of the discrepancy for the multipole model. Filled rhombs are for BE- and triangles are for previously refined multipolar model. Mean value is equal to 0.001 $Å^2$ for both cases, corresponding to the commonly acceptable limit.

Table 1. Retrieved *R* / R_{free}-factor statistics for previously refined multipolar model of enkephalin. M1 is the initial multipolar model (Wiest *et al*., 1994), and M2 is the same model with recovered R_{free} statistics (see text, section 3, for details). Statistics is given for all reflections with $I > 0$.

Models (resolution 0.56 -11.0 Å)	*R* (F), %	R_{free} (F), %	(R_{free} - *R*), %	Number of reflections		
				work set	test set	total
M1	8.09	-	-	8707	-	8707
M2	7.90	8.63	0.73	6894	1813	8707

Table 2. Refinement statistics for different models

Model composition		Parameters *)				*R* (*I*>0), % and data / parameters	R_{free} (*I*>0), %	(R_{free} - *R*), %
		U_{aniso}	B_{iso}	*Q*	*XYZ*			
Model A1	→ non-H atoms of the model → H-atoms of the model	+ -	- +	+ fixed =1.0	+ -	9.08 6894 / 474	9.74	0.66
Model BE1	→ non-H atoms of the model → H-atoms of the model ® *BE (bond electrons) taken as H-atoms*	+ - -	- + +	+ fixed =1.0 +	+ - -	8.43 6894 / 658	9.16	0.73
Model BE2	→ non-H atoms of the model → H-atoms of the model → BE (bond electrons) taken from *ab initio* calculations	+ - -	- + +	+ fixed =1.0 +	+ - -	8.12 6894 / 658	8.76	0.64

*) 1) U_{aniso} and B_{iso} – anisotropic and isotropic displacement parameters, respectively; *Q* – occupancy coefficient;

XYZ – atomic coordinates

2) The sign '+' or '-' means that the corresponding parameter was refined or fixed, respectively

Table 3. Selected examples of parameters for bond critical points obtained from BE2-model (Table 2)

Atomic pair, A-B	R_A (Å)	R_B (Å)	ρ ($eÅ^{-3}$)	λ_1 ($eÅ^{-5}$)	λ_2($eÅ^{-5}$)	λ_3($eÅ^{-5}$)	∇^2($eÅ^{-5}$)	η
C1-C2	0.774	0.762	1.60	-8.8	-8.6	5.7	-11.7	1.54
C1-N1	0.621	0.881	1.59	-8.9	-8.8	12.3	-5.5	0.73
C25-C26	0.775	0.747	1.56	-8.9	-8.8	5.2	-12.5	1.72
C23-C24	0.781	0.756	1.53	-9.7	-9.5	4.5	-14.2	2.14
C24-C25	0.768	0.757	1.53	-8.4	-7.9	5.8	-10.5	1.44

R – distance from the critical point to the neighbouring atom

ρ – value of electron density at the critical point

λ_1, λ_2, λ_3 – eigenvalues of the normal matrix of the electron density calculated in this critical point (the matrix of second derivatives with respect to three coordinates)

∇^2 – Laplacian

$\eta = |\lambda_1| / \lambda_3$

Bulk-solvent correction for use with the CCP4 version of *AMoRe*

Guido Capitani[1] and Andrei Fokine[2]
[1]Dept. of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland
[1]LCM3B, UMR 7036 CNRS, Faculté des Sciences, Université Henry Poincaré, Nancy I, 54506 Vandoeuvre-lés-Nancy, France
Email: capitani@bioc.unizh.ch

1. Introduction

Low-resolution reflections, being less sensitive to model imperfections, are known to be very useful (Urzhumtsev & Podjarny, 1995a, Fokine & Urzhumtsev, 2002a) for solving the translation problem in molecular replacement (Rossmann, 1972). In protein crystallography, traditional molecular replacement protocols, however, exclude reflections of resolution lower than 10 or 15 Å, because they are strongly influenced by the contribution of the bulk solvent in the crystals. As a consequence, at low resolution, comparison of structure factors calculated from the atomic model with experimental values is not reliable. In order to obviate this problem, a bulk solvent correction has to be introduced in the translation search procedure. Two different bulk solvent correction approaches are available (Kostrewa, 1997), the exponential scaling model (Moews & Kretsinger, 1975) and the mask model (Phillips, 1980; Jiang & Brünger, 1994), -both originally developed for macromolecular refinement purposes-. The exponential scaling model, based on Babinet's principle, is of simple implementation and is now applied to calculate a bulk solvent correction in the molecular replacement programs MOLREP (Vagin & Teplyakov, 1997), QS (Glykos & Kokkinidis, 2000) and BEAST (Read, 2001). The assumptions this approach is based on are, however, only true at resolutions below ~ 15 Å (Urzhumtsev & Podjarny, 1995b) and its performance was shown to be inferior to that of the mask model (Kostrewa, 1997). This latter method involves explicit calculation of an envelope around the protein model, creating a 'solvent mask'. The solvent region delimited by this solvent mask is then filled with bulk solvent electron density, and structure factors for this density are calculated and vectorially added to those derived from the protein model. Fokine and Urzhumtsev (2001) suggested a way to employ the mask method to calculate accurate bulk solvent correction for fast translation searches in molecular replacement. The corresponding program BULK (Fokine, Capitani, Grütter & Urzhumtsev, 2002) can be used with the standalone version of *AMoRe* (Navaza, 1994; Navaza & Vernoslova, 1995), with the CCP4 version of *AMoRe* (CCP4, 1994) and with CNS (Brünger *et al.,* 1998).

2. Description of the BULK program and of its usage

A typical molecular replacement run with *AMoRe* involves three main steps: the calculation of a structure factor (SF) table from the search model (often referred to as 'tabling step'), the cross-rotation function search (' roting step') and the fast translation search (' traing step'). The table below summarizes those steps and their input and output files:

INPUT	STEP	OUTPUT
Search model (PDB file)	Tabling	SF table (' search tab') 'tabbed model'* Tabling step log file
Experimental amplitudes SF table (' search. tab')	Roting	List of cross-rotation function peaks
Experimental amplitudes SF table (' search tab') List of cross-rotation function peaks	Traing	List of translation function peaks

*see text for the meaning of 'tabbed model'

The program BULK introduces another step (' bulking') in this procedure and calculates a bulk-solvent corrected structure factor table to be used in the 'traing' step. The scheme is then modified as follows:

INPUT	STEP	OUTPUT
Search model (PDB file)	Tabling	SF table (' search tab') 'tabbed model' Tabling step log file
Experimental amplitudes SF table (' search. tab')	Roting	List of cross-rotation function peaks
SF table (' search. tab') Search model Tabling step log file	Bulking	**Corrected** SF table (search tabs)
Experimental amplitudes **Corrected** SF table (search tabs) List of cross-rotation function peaks	Traing	List of translation function peaks

The 'bulking' procedure encompasses various computational tasks, which are to be carried out after the *AMoRe* 'tabling' step, where the search model is placed in a large rectangular box (the box dimensions are determined automatically) with its centre of mass at the origin and its principal inertia axes parallel to the box axes. A model rotated and translated in this way can be referred to as a 'tabbed' model. Then, structure factors $\mathbf{F}_{mP1}(\mathbf{h})$ from this model are calculated by *AMoRe.* 'BULK' goes then through the following steps:

1. It computes a molecular envelope from the model coordinates, which are placed at the origin exactly as in the *AMoRe* 'tabling' step;
2. It calculates $\mathbf{F}_{eP1}(\mathbf{h})$**,** the Fourier coefficients for this envelope, using the same box as in the *AMoRe* 'tabling' step;
3. It obtains the corresponding solvent structure factors as

$$\mathbf{F}_{sP1}(\mathbf{h}) = [-\tilde{k}_{sol} \exp(-\tilde{B}_{sol} h^2/4)]\, \mathbf{F}_{eP1}(\mathbf{h})$$

   using $\tilde{k}_{sol}$ and $\tilde{B}_{sol}$ values defined in the input file;

4. It carries out the sum $\mathbf{F}_{corrP1}(\mathbf{h}) = \mathbf{F}_{mP1}(\mathbf{h}) + \mathbf{F}_{sP1}(\mathbf{h})$**,** where $\mathbf{F}_{mP1}(\mathbf{h})$ are the model structure factors previously calculated by the *AMoRe* 'tabling' step. The corrected structure factors $\mathbf{F}_{corrP1}(\mathbf{h})$ can be then used in the *AMoRe 'traing'* (and optionally 'fiting') step.

The theoretical ground for the above calculations is discussed in (Fokine, Capitani, Grütter & Urzhumtsev, 2002). Here we just point out that the procedure uses constant solvent scaling and B-factor parameters ($\tilde{k}_{sol}$, $\tilde{B}_{sol}$) as defined by an input file. Mean values for these parameters (0.35 e/Å^3 and 46 Å^2 , respectively) were derived by Fokine & Urzhumtsev (2002b) through a statistical analysis of well-refined structures deposited in the Protein Data Bank (Bernstein *et al.,* 1977). If for a certain crystal the buffer density differs markedly from standard values, the user can accordingly change the default $\tilde{k}_{sol}$ and $\tilde{B}_{sol}$input parameters. Another point worth mentioning is that 'BULK' can be used only in cases when one entity in the asymmetric unit is searched for (this can be also a tetramer or an octamer, provided that a corresponding tetrameric or octameric search model, respectively, is available). On the contrary, if the molecular replacement problem involves, for instance, locating two independent monomers, the current approach cannot be used because it would produce unreliable structure factors (Fokine, Capitani, Grütter & Urzhumtsev, 2002).

To install and run BULK one first has to copy the 'bulk' directory (as extracted from the BULK distribution tar file) into the *AMoRe* working directory, then to compile it with the command:

```
./bulk/make_bulk
```

This creates two executables, prep_bulk and bulking.

The only input file needed, (an example is provided in the distribution), is prep_bulk.inp, with the following contents:

search.pdb	name of the file with model coordinates (the same as used for the 'tabling'step)
search.tab	name of the file of structure factors from the *AMoRe*'tabling' step
tab.log	name of the log file of the *AMoRe* 'tabling' step
search.tabs	name of the file with bulk-solvent-corrected structure factors (this file is created by the 'bulking' procedure)
bulking.inp	name of an intermediate control file to be input to 'bulking'; this file is created by running 'prep_bulk'
0.35	value of $\tilde{k}_{sol}$(scaling parameter for solvent electron density, in e/Å^3)
50.0	value of $\tilde{B}_{sol}$(temperature factor for solvent, in Å^2)
1.0	value of the solvent radius used for the solvent mask calculation

To run the procedure, one issues the two following shell commands:

```
prep_bulk < prep_bulk.inp
bulking < bulking.inp
```

'prep_bulk' creates the file 'bulking.inp', containing the information needed by 'bulking' to calculate the actual bulk-solvent correction. The output file 'search.tabs', created by 'bulking', contains the bulk-solvent-corrected structure factors and can be used for the translation search instead of 'search.tab'. Importantly, the corrected structure factor file should not be used in the 'roting' step, since low-resolution reflections contribute rather negatively to the rotation function search results.

3. A test case

To evaluate the performance of BULK in a typical CCP4 *AMoRe* run, a test case was carried out based on a structure solved very recently at the University of Zürich. The crystals of that protein (an enzyme involved in apoptosis) diffract well and a good quality dataset (space group $P2_12_12_1$, resolution 15.0-1.8 Å, completeness 96.0%, Rsym 8.1%) could be collected at a synchrotron source. The protein is a heterotetramer composed by two • and two • subunits, and it was solved by molecular replacement with CCP4 *AMoRe* using a related tetrameric protein (PDB code 1CP3) as a search model. For the test case, a more distant search model (PDB code 1F9E) was employed (rmsd 1.6 Å for 144 common C• atoms considering one • and one • subunit) and its subunits, corresponding to 37 % of the model, were deleted.

A conventional rotation search was then carried out in the range 8-4 Å and the top 30 rotation function solutions were listed. BULK was then used to calculate a corrected structure factor table. Translation tests were performed in the ranges 10-4 Å and 15-4 Å, both with and without bulk solvent correction. The top 30 solutions in each case, sorted by correlation coefficient, are reported below (•,• and • are the eulerian rotation angles for each solution, Tx, Ty, Tz the fractional translations, cc, Rf and cc-I the amplitude-based correlation coefficient, the R-factor and the intensity-based correlation coefficient, respectively):

A. 10-4 Å WITHOUT bulk solvent correction

	•	•	•	Tx	Ty	Tz	cc	Rf	cc-I
SOLUTIONTF1	**1**	**139.59**	**56.87**	**225.89**	**0.1691**	**0.4390**	**0.2361**	**9.2**	**55.6**	**9.5**
SOLUTIONTF1	1	134.00	59.87	50.32	0.1338	0.4407	0.4377	6.0	57.2	8.3
SOLUTIONTF1	1	72.06	41.69	133.77	0.3280	0.2580	0.0577	5.4	57.4	6.2
SOLUTIONTF1	1	145.89	45.72	44.62	0.4242	0.1225	0.0964	5.2	57.3	5.1
SOLUTIONTF1	1	88.83	90.00	60.70	0.2029	0.1829	0.1960	5.2	58.8	5.4
SOLUTIONTF1	1	68.57	36.76	316.05	0.1218	0.4297	0.3087	5.2	57.6	4.5
SOLUTIONTF1	1	60.70	42.44	244.79	0.3353	0.3195	0.1074	4.6	56.8	5.0
SOLUTIONTF1	1	111.50	36.91	56.00	0.2603	0.3188	0.1500	4.5	56.9	5.1
SOLUTIONTF1	1	113.50	46.00	58.30	0.2170	0.0728	0.2099	4.2	57.7	5.1
SOLUTIONTF1	1	86.00	90.00	18.47	0.4114	0.9958	0.0000	4.2	61.3	3.9
SOLUTIONTF1	1	95.24	90.00	69.66	0.4683	0.2917	0.1845	4.0	57.8	3.6
SOLUTIONTF1	1	43.81	38.01	48.64	0.2352	0.2299	0.0202	4.0	56.9	3.0
SOLUTIONTF1	1	116.65	35.64	233.53	0.0059	0.0018	0.4482	3.9	57.3	5.4
SOLUTIONTF1	1	87.08	26.20	306.42	0.3940	0.3887	0.1388	3.9	57.8	4.6
SOLUTIONTF1	1	56.96	81.80	194.07	0.1389	0.0040	0.1650	3.9	57.4	3.3
SOLUTIONTF1	1	54.21	51.00	39.16	0.3014	0.3577	0.3644	3.8	57.4	2.7
SOLUTIONTF1	1	15.39	70.84	128.24	0.0188	0.0999	0.4365	3.7	58.2	3.5
SOLUTIONTF1	1	158.40	86.17	268.98	0.4258	0.3079	0.0769	3.5	57.8	2.7
SOLUTIONTF1	1	91.00	29.96	121.50	0.3952	0.3952	0.1324	3.5	57.7	4.5
SOLUTIONTF1	1	83.68	80.82	311.33	0.1649	0.1792	0.0221	3.5	57.6	4.9
SOLUTIONTF1	1	76.30	36.00	311.71	0.3069	0.2979	0.0116	3.5	58.0	4.1
SOLUTIONTF1	1	57.79	87.04	284.00	0.3811	0.2135	0.4233	3.5	57.6	2.7
SOLUTIONTF1	1	94.96	70.55	251.19	0.1450	0.1805	0.2657	3.4	57.5	3.6
SOLUTIONTF1	1	80.54	34.85	309.13	0.0546	0.0800	0.2310	3.4	58.0	4.0

```
SOLUTIONTF1 1 51.00  35.70 221.50 0.2660 0.1106 0.0859 3.2 57.5 2.8
SOLUTIONTF1 1 92.23  90.00 197.89 0.3074 0.0060 0.0000 3.1 63.3 3.5
SOLUTIONTF1 1 11.86  64.00 129.40 0.3105 0.4471 0.4815 3.1 57.2 3.7
SOLUTIONTF1 1 151.00 41.61 314.00 0.3379 0.4513 0.2173 2.9 58.2 2.6
SOLUTIONTF1 1 43.60  81.50 13.16  0.1041 0.3664 0.3687 2.8 57.7 3.3
SOLUTIONTF1 1 86.40  86.41 247.14 0.2704 0.4661 0.2848 2.5 57.5 2.7
```

This first case represents a 'classical' resolution range for a translation search run. Sorting by correlation coefficient identifies one slightly emerging solution but the signal is weak. Further analysis, including comparison with the newly solved structure, shows that the top-ranking solution (in blue, with cyan background) is correct, whereas the second is wrong.

B. 10-4 Å WITH bulk solvent correction

	•	•	•	Tx	Ty	Tz	cc	Rf	cc-I
SOLUTIONTF1	**1**	**134.00**	**59.87**	**50.32**	**0.1824**	**0.4440**	**0.2365**	**11.2**	**54.9**	**13.4**
SOLUTIONTF1	**1**	**145.89**	**45.72**	**44.62**	**0.1943**	**0.4757**	**0.2176**	**9.9**	**54.5**	**9.7**
SOLUTIONTF1	1	72.06	41.69	133.77	0.2079	0.1879	0.4616	9.1	55.3	9.2
SOLUTIONTF1	1	113.50	46.00	58.30	0.2201	0.0987	0.2040	8.5	55.4	9.8
SOLUTIONTF1	1	111.50	36.91	56.00	0.4681	0.3272	0.3476	8.3	55.0	8.1
SOLUTIONTF1	1	88.83	90.00	60.70	0.2028	0.1860	0.1941	8.3	57.2	8.1
SOLUTIONTF1	1	68.57	36.76	316.05	0.2626	0.0711	0.4542	8.1	55.3	8.5
SOLUTIONTF1	1	57.79	87.04	284.00	0.3830	0.2146	0.4251	7.8	55.3	6.9
SOLUTIONTF1	1	15.39	70.84	128.24	0.3441	0.9989	0.4286	7.8	55.4	7.5
SOLUTIONTF1	1	76.30	36.00	311.71	0.3643	0.4896	0.2877	7.7	55.5	8.8
SOLUTIONTF1	1	158.40	86.17	268.98	0.1451	0.4607	0.0658	7.5	55.5	7.7
SOLUTIONTF1	1	95.24	90.00	69.66	0.3852	0.0560	0.3756	7.4	55.2	7.0
SOLUTIONTF1	1	116.65	35.64	233.53	0.0096	0.4709	0.4455	7.3	55.1	8.0
SOLUTIONTF1	1	80.54	34.85	309.13	0.0566	0.0790	0.2314	7.3	55.9	8.8
SOLUTIONTF1	1	56.96	81.80	194.07	0.4703	0.3603	0.1054	7.3	55.1	6.9
SOLUTIONTF1	1	94.96	70.55	251.19	0.1271	0.4514	0.3930	7.1	55.6	6.8
SOLUTIONTF1	1	60.70	42.44	244.79	0.4287	0.3201	0.0630	7.1	54.6	7.4
SOLUTIONTF1	1	54.21	51.00	39.16	0.3420	0.2334	0.0730	7.1	55.5	7.3
SOLUTIONTF1	1	87.08	26.20	306.42	0.2921	0.0411	0.4443	6.9	55.5	7.4
SOLUTIONTF1	1	83.68	80.82	311.33	0.2528	0.4874	0.1895	6.9	56.3	9.3
SOLUTIONTF1	1	11.86	64.00	129.40	0.3803	0.4433	0.1825	6.9	55.7	8.0
SOLUTIONTF1	1	86.00	90.00	18.47	0.4117	0.9946	0.0000	6.8	59.9	5.9
SOLUTIONTF1	1	43.81	38.01	48.64	0.2616	0.0488	0.3946	6.7	55.4	6.0
SOLUTIONTF1	1	51.00	35.70	221.50	0.4401	0.1680	0.0470	6.4	55.4	6.2
SOLUTIONTF1	1	92.23	90.00	197.89	0.4567	0.0409	0.2660	6.3	58.8	5.5
SOLUTIONTF1	1	86.40	86.41	247.14	0.1486	0.4503	0.2280	6.3	56.2	7.4
SOLUTIONTF1	1	43.60	81.50	13.16	0.1480	0.4897	0.2492	6.3	56.3	6.4
SOLUTIONTF1	1	91.00	29.96	121.50	0.2939	0.2384	0.4492	6.1	55.8	6.3
SOLUTIONTF1	1	151.00	41.61	314.00	0.0900	0.2623	0.3265	5.7	56.2	6.2

In this run the resolution range is the same as before, but the first solution (also same as before) now exhibits a much higher correlation coefficient. By comparing the first to the second ranking solution (both in blue), it appears that they are both correct and equivalent (same translations and rotations, except for a ~ 180° difference in •, which is readily explained by the two-fold symmetry of the search model). Interestingly, also in run A) a similar situation is observed, but the value for Tz in the second solution differs from that of the first and turns out to be wrong by further analysis. Comparison of runs A) and B) shows then that using the bulk solvent correction was instrumental for obtaining correct translation parameters for the second solution.

C. 15-4 Å WITH bulk solvent correction

		•	•	•	Tx	Ty	Tz	cc	Rf	cc-I
SOLUTIONTF1	**1**	**139.59**	**56.87**	**225.89**	**0.1716**	**0.4378**	**0.2365**	**16.4**	**53.5**	**16.8**
SOLUTIONTF1	**1**	**134.00**	**59.87**	**50.32**	**0.1846**	**0.4415**	**0.2346**	**12.7**	**54.8**	**14.2**
SOLUTIONTF1	1	145.89	45.72	44.62	0.1950	0.4740	0.2176	11.1	54.7	9.7
SOLUTIONTF1	1	113.50	46.00	58.30	0.3360	0.3380	0.3054	9.0	55.9	9.7
SOLUTIONTF1	1	68.57	36.76	316.05	0.3924	0.4734	0.2250	8.9	56.4	9.0
SOLUTIONTF1	1	111.50	36.91	56.00	0.0206	0.0148	0.4481	8.8	55.6	8.9
SOLUTIONTF1	1	88.83	90.00	60.70	0.2027	0.1848	0.1937	8.8	57.6	8.3
SOLUTIONTF1	1	72.06	41.69	133.77	0.1529	0.3820	0.4211	8.8	55.9	9.9
SOLUTIONTF1	1	158.40	86.17	268.98	0.4053	0.4080	0.0285	8.4	55.6	6.8
SOLUTIONTF1	1	60.70	42.44	244.79	0.3510	0.2223	0.1524	7.9	55.4	6.7
SOLUTIONTF1	1	76.30	36.00	311.71	0.3651	0.4889	0.2886	7.8	56.2	7.4
SOLUTIONTF1	1	116.65	35.64	233.53	0.0059	0.0035	0.4482	7.7	55.9	8.7
SOLUTIONTF1	1	94.96	70.55	251.19	0.0268	0.2212	0.2433	7.6	55.8	6.5
SOLUTIONTF1	1	83.68	80.82	311.33	0.4264	0.0023	0.3584	7.2	56.7	8.0
SOLUTIONTF1	1	80.54	34.85	309.13	0.0563	0.0787	0.2313	7.2	56.4	7.3
SOLUTIONTF1	1	15.39	70.84	128.24	0.2261	0.1199	0.4634	7.2	56.2	6.8
SOLUTIONTF1	1	11.86	64.00	129.40	0.1529	0.1712	0.1759	7.0	55.5	6.6
SOLUTIONTF1	1	43.81	38.01	48.64	0.4538	0.1625	0.0477	6.9	56.1	6.5
SOLUTIONTF1	1	56.96	81.80	194.07	0.0170	0.3098	0.1236	6.7	55.9	5.8
SOLUTIONTF1	1	43.60	81.50	13.16	0.1484	0.4888	0.2491	6.7	56.8	6.1
SOLUTIONTF1	1	54.21	51.00	39.16	0.4464	0.3624	0.2688	6.6	55.8	4.6
SOLUTIONTF1	1	95.24	90.00	69.66	0.3267	0.4105	0.0458	6.5	56.4	6.6
SOLUTIONTF1	1	87.08	26.20	306.42	0.2932	0.0408	0.4435	6.5	56.2	6.5
SOLUTIONTF1	1	86.00	90.00	18.47	0.4130	0.9969	0.2583	6.5	58.3	5.7
SOLUTIONTF1	1	57.79	87.04	284.00	0.2481	0.2693	0.4858	6.3	56.0	6.3
SOLUTIONTF1	1	51.00	35.70	221.50	0.4421	0.1697	0.0471	6.3	55.9	5.9
SOLUTIONTF1	1	86.40	86.41	247.14	0.1754	0.2091	0.4515	6.1	56.4	6.6
SOLUTIONTF1	1	151.00	41.61	314.00	0.0314	0.2624	0.1190	5.8	56.4	5.7
SOLUTIONTF1	1	92.23	90.00	197.89	0.3987	1.0000	0.2551	5.8	59.5	4.8
SOLUTIONTF1	1	91.00	29.96	121.50	0.0376	0.3296	0.3309	5.7	56.6	6.7

In run C) the full available resolution range was used with bulk solvent correction, the results are similar to those of run B), but with higher correlation coefficients (also cc-I) for the first-and second-ranking solution (both in blue).

D. 15-4 Å WITHOUT bulk solvent correction

		•	•	•	Tx	Ty	Tz	cc	Rf	cc-I
SOLUTIONTF1	1	139.59	56.87	225.89	0.1678	0.4373	0.2365	9.7	58.3	8.3
SOLUTIONTF1	1	134.00	59.87	50.32	0.1305	0.4409	**0.4365**	7.3	59.4	7.1
SOLUTIONTF1	1	145.89	45.72	44.62	0.2279	0.2650	0.3167	6.2	59.8	5.7
SOLUTIONTF1	1	111.50	36.91	56.00	0.3655	0.2451	0.0275	6.1	59.9	6.1
SOLUTIONTF1	1	113.50	46.00	58.30	0.1324	0.2200	0.0288	5.7	59.7	6.0
SOLUTIONTF1	1	68.57	36.76	316.05	0.4704	0.4331	0.3752	5.6	60.0	5.1
SOLUTIONTF1	1	94.96	70.55	251.19	0.0260	0.2183	0.4526	5.3	59.9	4.7
SOLUTIONTF1	1	72.06	41.69	133.77	0.4068	0.4374	0.1118	5.0	59.9	5.5
SOLUTIONTF1	1	158.40	86.17	268.98	0.4706	0.1000	0.3846	4.9	60.3	4.8
SOLUTIONTF1	1	88.83	90.00	60.70	0.3382	0.3900	0.1923	4.8	62.2	4.1
SOLUTIONTF1	1	60.70	42.44	244.79	0.0529	0.2115	0.3682	4.7	60.2	5.4
SOLUTIONTF1	1	56.96	81.80	194.07	0.1636	0.0649	0.4190	4.6	59.4	3.9
SOLUTIONTF1	1	87.08	26.20	306.42	0.2059	0.2900	0.2788	4.3	60.4	3.5
SOLUTIONTF1	1	83.68	80.82	311.33	0.4454	0.1092	0.0475	4.3	60.6	4.3
SOLUTIONTF1	1	43.81	38.01	48.64	0.4706	0.2500	0.2596	4.3	59.5	3.1
SOLUTIONTF1	1	80.54	34.85	309.13	0.2072	0.2827	0.3633	4.0	60.9	3.8
SOLUTIONTF1	1	76.30	36.00	311.71	0.0452	0.1494	0.0303	4.0	60.7	5.2
SOLUTIONTF1	1	116.65	35.64	233.53	0.2206	0.2600	0.1058	3.9	60.3	4.5
SOLUTIONTF1	1	43.60	81.50	13.16	0.1034	0.3654	0.3704	3.8	60.0	4.0
SOLUTIONTF1	1	11.86	64.00	129.40	0.2546	0.1097	0.2098	3.7	60.4	3.2
SOLUTIONTF1	1	95.24	90.00	69.66	0.3196	0.2313	0.4455	3.6	60.6	2.9
SOLUTIONTF1	1	92.23	90.00	197.89	0.0335	0.3088	0.2327	3.6	63.4	3.0
SOLUTIONTF1	1	15.39	70.84	128.24	0.0600	0.0896	0.1717	3.6	60.8	3.3
SOLUTIONTF1	1	91.00	29.96	121.50	0.2794	0.3400	0.4038	3.5	60.8	3.6
SOLUTIONTF1	1	86.40	86.41	247.14	0.2928	0.2094	0.4518	3.4	60.6	2.8
SOLUTIONTF1	1	51.00	35.70	221.50	0.3235	0.2900	0.3269	3.4	60.5	2.9
SOLUTIONTF1	1	86.00	90.00	18.47	0.1664	0.3998	0.0288	3.2	62.2	3.0
SOLUTIONTF1	1	57.79	87.04	284.00	0.1324	0.2700	0.4904	3.1	60.2	2.8
SOLUTIONTF1	1	54.21	51.00	39.16	0.4559	0.3600	0.3654	2.9	60.3	1.7
SOLUTIONTF1	1	151.00	41.61	314.00	0.0987	0.3222	0.0750	2.5	60.2	3.3

Run D) shows the importance of applying the bulk solvent correction when using low-resolution reflections down to 15 Å: the results are in fact much poorer compared to those of run C) and similar to those of run A). Also in this case the Tz translation is wrong for the second-ranking solution (in blue).

4. Future developments

BULK has been tested with CCP4 up to version 4.2.1. It is at the moment distributed upon request by the authors at the following addresses: fokine@lcm3b.uhp-nancy.fr, sacha@lcm3b.uhp-nancy.fr and capitani@bioc.unizh.ch. Future developments will involve the integration of BULK into the CCP4 distribution.

Acknowledgements

The authors are grateful to Profs. Alexandre Urzhumtsev and Markus G. Grütter for support and encouragement. G. C. wishes to acknowledge the Baugarten Foundation, Zürich, as well as Andreas Schweizer and Christophe Briand. A. F. acknowledges Lorraine Regional administration, Pole "Intelligence Logiciels" CPER-Lorraine and GdR 2417 CNRS for the financial support.

References

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. 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., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.,* **D54,** 905-921.

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

Fokine, A., Capitani, G., Grütter, M. G. & Urzhumtsev A. (2002). *J. Appl. Cryst.,* submitted.

Fokine, A. & Urzhumtsev A. (2001). *CCP4 Newsletter,* **40.**

Fokine, A. & Urzhumtsev A. (2002a). *Acta Cryst.,* **A58,** 72-74.

Fokine, A. & Urzhumtsev, A. (2002b). *Acta Cryst.* **D58,** 1387-92.

Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst.* **D56,** 1070-1072.

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

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

Moews, P. C. & Kretsinger, R. H. (1975). *J. Mol. Biol.,* **91,** 201-225.

Navaza, J. (1994). *Acta Cryst.* **A50,** 157-163.

Navaza, J. & Vernoslova, E. (1995). *Acta Cryst.,* **A51,** 445-449.

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

Read, R. J. (1997). *Acta Cryst.* **D57,** 1373-1382.

Rossmann, M. G., *ed.* (1972). *The Molecular Replacement Method,* Gordon & Breach; New York, London, Paris.

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

Urzhumtsev, A. & Podjarny, A. D. (1995b). *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography,* **31,** 12-16.

Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30,** 1022-1025.

Variation of solvent density and low-resolution *ab initio* phasing

Andrei Fokine

LCM3B, UMR 7036 CNRS, Faculté des Sciences, Université Henry Poincaré, Nancy I, 54506 Vandoeuvre-lés-Nancy, France
e-mail: fokine@lcm3b.uhp-nancy.fr

Abstract

The bulk solvent plays a key role in the direct phasing at low resolution. The connectivity-based phasing method has been tested for the crystal of the $tRNA^{Asp}$-aspartyl-tRNA synthetase complex. The neutron diffraction data have been measured for different solvent contrast masking different components of the crystal. The obtained low resolution images are perfectly correlated with the solvent level.

1. Introduction

Low-resolution Fourier syntheses provide a very useful information about molecular packing and the shape of molecules. Such low-resolution images can be obtained by direct phasing of a single set of experimental modules (Lunin *et al.*, 2000a). The information supplied by *ab initio* phasing methods is inestimable when conventional phasing techniques do not work. The low resolution phases can be used as a starting point for phase extension. Additionally, the knowledge of molecular position can facilitate the solution of molecular replacement problem. Here the *ab initio* phasing method based on topological properties of density maps is considered. This method employs the expected connectivity of a Fourier synthesis as an additional information to obtain phases (Lunin *et al.,* 2000b). Among the known low-resolution direct phasing techniques this method is considered to be the most promising. There are several examples when this method allowed to determine molecular positions, molecular shapes (Lunin *et al.*, 2000a; Lunin *et al.*, 2000b) and even secondary structure elements (Lunin *et al.,* 2002).

Macromolecular crystals contain a large part of the bulk solvent whose contribution to low-resolution reflections is very important. At low resolution the assumption of a uniform scattering density distribution in the solvent region is held well and the value of solvent scattering density can be estimated from the composition of crystallization solution. In X-ray structural analysis the electron density of the bulk solvent varies in the small range 0.33 – 0.41 e/Å^3 (Kostreva, 1997). The solvent electron density is close to the average electron density of a typical protein 0.43 e/Å^3 (Kostreva, 1997; Carter *et al.,* 1990) and less than that of DNA / RNA. In neutron diffraction experiment, the solvent scattering density can be varied in a large range by changing the ratio of D_2O / H_2O in the mother liquor. The solvent scattering density can be made lower equal or greater than the average scattering density of a protein or DNA / RNA. Such possibility of choice of an appropriate solvent scattering density is the basis of the method of contrast variation (Jacrot, 1976).

The connectivity properties of low-resolution Fourier syntheses essentially depend on the solvent contribution. Therefore, the results of direct phasing which uses these

properties as the selection criterion depend on the solvent scattering density. The purpose of the current work was to study the role of the bulk solvent in the connectivity-based direct phasing.

In this paper we present the results of *ab initio* phasing of three sets of neutron diffraction data from the same crystal of the $tRNA^{Asp}$-aspartyl-tRNA synthetase complex (Moras *et al.,* 1983) measured for different solvent contrast masking different components of the crystal. The obtained low resolution images are perfectly correlated with the solvent scattering density level.

2. Phasing method

The basic idea of the connectivity based phasing method (Lunin *et al.,* 2000b) consists in the observation that topological properties of high density regions of the Fourier syntheses are different for properly phased syntheses and for those calculated with random phases.

Let $\rho(\mathbf{r})$ be a Fourier syntheses calculated on a finite grid and N be a number of grid points in the unit cell. The high density region $\rho(\rho_\alpha)$, corresponding to the relative volume α is defined as a set of αN grid points of highest density. For a correctly phased low-resolution synthesis this region would be composed of a small number of isolated 'blobs' corresponding to independent molecules if the cut-off level α is chosen appropriately. The number of these blobs is usually equal to the number of molecules in the unit cell. The blob volumes (measured in the number of grid points) must be equal between themselves if all molecules are linked by crystallographic symmetries and must be approximately equal if non-crystallographic symmetry is present. On the other hand, randomly phased syntheses are likely to show infinite merged regions or a large number of 'drops'.

The phasing procedure consists in following steps.

1) A large number of random phase sets are generated. The phases are generated with the uniform distribution at the beginning of the procedure or in accordance with a known phase distribution if this information is already available.
2) For every generated phase set the Fourier synthesis is calculated using the experimental structure factor modules.
3) The high density regions of each calculated synthesis are subjected to connectivity analysis in order to determine the number of separated connected components in the unit cell and to calculate their volumes.
4) If high density regions consist of a desired number of components, the corresponding phase set is considered as admissible and is stored, otherwise the phase set is rejected.
5) After a reasonable number (about one hundred) of admissible phase sets have been selected, they are averaged in order to produce the corresponding 'best' phases $\varphi^{best}(\mathbf{h})$ and figures of merit $m(\mathbf{h})$, which reflect the spread of the admissible phase sets

$$m(\mathbf{h})\exp\left[i\varphi^{best}(\mathbf{h})\right] = \frac{1}{M}\sum_{j=1}^{M}\exp\left[i\varphi^{j}(h)\right] . \qquad (1)$$

Here *M* is the number of selected phase sets and $\varphi^{j}(\mathbf{h})$ is the value of phase of the structure factor with the index $\mathbf{h}$ in the j-th selected phase set. It should be noted that the optimal alignments of the phase sets in accordance with the permitted origin shifts (Lunin & Lunina, 1996) must be performed before averaging.

The phasing method is described in more details in Lunin *et al.* (2000b).

3. Test object

Cubic form of tRNA^{Asp}-aspartyl-tRNA synthetase complex

The crystals of the cubic form of the tRNA^{Asp}-aspartyl-tRNA synthetase complex belong to the space group I432 (48 asymmetric unit / unit cell) with a unit cell parameter of 354 Å. The asymmetric unit contains one protein homodimer and two molecules of tRNA. The crystals contain 82% of the bulk solvent. The protein and tRNA molecules occupy 14 and 4 % of the unit cell respectively. The structure was solved by molecular replacement (Urzhumtsev *et al.,* 1994) using 15-8 Å resolution X-ray diffraction data and high resolution (3 Å) model of the complex obtained from the orthorhombic crystal (Ruff *et al.,* 1991).

Neutron diffraction data

Three sets of neutron diffraction data, all complete at low resolution (from infinity to 24 Å), were measured for the same crystal of the complex with different concentrations of D_2O in the mother liquor (Moras *et al.,* 1983).

The first data set was measured without D_2O in the mother liquor which corresponds to a very low solvent scattering density. This data set corresponds to the full complex molecule.

For the second data set the concentration of D_2O was chosen so that the solvent scattering density was equal to the average density of the protein. Thus the protein was masked by the solvent and only tRNA molecules gave a signal in diffraction.

The third data set was measured with the solvent scattering density matching the average density of tRNA, therefore this data set corresponds to the protein molecules only.

These three data sets were used for the direct phasing.

4. *Ab initio* phasing of experimental neutron data sets

It should be noted that for the direct phasing described below we used only a general information (known *a priory*) such as the number of molecules in the unit cell and the relative unit cell volumes occupied by molecules.

Phasing of the data set corresponding to the full complex molecules

The crystals contain one densely packed dimer of tRNA^{Asp}-aspartyl-tRNA synthetase complex in the asymmetric unit. Since the space group I432 has 48 symmetry operations, it is natural to expect that a low-resolution synthesis with correct phases would show 48 blobs of equal volumes corresponding to the complex molecules.

Reflections in the resolution range • -45 Å were used (37 reflections) for phasing and phases were generated with the uniform distribution. The selection criterion was formulated as follows: the high density region occupying 5% of the unit cell (• = 0.05) must be composed of 48 connected blobs of equal volumes. 100 selected variants were stored after about 89000 generations. The selected variants were averaged to produce the 'best' phases $j^{best}(\mathbf{h})$ and the figures of merit $m(\mathbf{h})$. The *ab initio* phased synthesis calculated at 45 Å is shown in Fig. 1. From this synthesis, the position of the complex molecules in the unit cell can be determined.

Phasing of the data set corresponding to tRNA molecules

The asymmetric unit contains 2 tRNA molecules related by non-crystallographic symmetry and separated at the surface of the complex. It is natural to expect that the low-resolution synthesis with correct phases would show 48 blobs of equal volumes corresponding to the

first tRNA molecule and 48 blobs of equal volumes corresponding to the second tRNA molecule. The volumes of blobs corresponding to tRNA molecules related by non-crystallographic symmetry must be approximately equal.

As previously, reflections in the resolution range • -45 Å were used for phasing. Phases were generated with the uniform distribution. The selection criterion was formulated as follows: the high density region occupying 3% of the unit cell (• = 0.03) must be composed of 48 connected regions of equal volumes corresponding to the first tRNA molecule and 48 connected regions of equal volumes corresponding to the second tRNA molecule; the ratio of volumes of the connected regions corresponding to different tRNA molecules must be higher than 0.7. It was *a priory* known that the tRNA molecules occupy 4% of the unit cell therefore the slightly higher cut-off level of 3% was used. About 270000 random phase sets were generated, 100 sets from them satisfied the selection criterion. The corresponding *ab initio* phased synthesis is shown in Fig. 2. From this synthesis the positions of the tRNA molecules in the unit cell can be determined unambiguously.

Phasing of the data set corresponding to the protein molecules

Similarly to two previous cases, reflections in the resolution range • -45 Å were used and phases were generated with the uniform distribution. The selection criterion was formulated as follows: the high density region occupying 5% of the unit cell (• = 0.05) must be composed of 48 isolated blobs of equal volumes. 100 selected variants were stored after about 50000 generations. The *ab initio* phased synthesis (Fig. 3) shows clearly the position of the protein homodimer.

5. Conclusion

The current study shows the robustness of the connectivity-based phasing method which was capable to determine unambiguously the position of both tRNA molecules and that of protein homodimer in the $tRNA^{Asp}$-aspartyl-tRNA synthetase complex.

The results confirm that the bulk solvent plays the key role in the connectivity-based direct phasing. The phasing of three data set measured from the same crystal led to completely different images depending on the solvent scattering density.

The author thanks A. Urzhumtsev, V. Lunin, N. Lunina, E. Chabriere and P. Afonine for useful discussions.

The work was done in the frame of the pole “Intellegence Logiciels” CPER-Lorraine and in collaboration with CCH, Nancy, and was supported by the grant of “Region Lorraine”. The author is a member of GdR 2417 CNRS.

References

Carter C.W., Jr., Crumbley, K.V., Coleman, D.E., Hage, F. & Bricogne, G. (1990). *Acta Cryst*. A**46**, 57-68.
Jacrot, B. (1976). *Rep. Prog. Phys.* **39**, 911
Kostrewa, D. (1997). *CCP4 Newslett*. **34**, 9-22.
Lunin, V.Y. & Lunina, N.L. (1996). *Acta Cryst.,* A**52,** 365-368.
Lunin, V.Y., Lunina, N.L., Petrova, T.E., Skovoroda, T.P., Urzhumtsev, A.G. & Podjarny A.D. (2000[a]). *Acta Cryst*. D**56**, 1223-1232.
Lunin, V.Y., Lunina, N.L., Urzhumtsev, A.G. (2000[b]). *Acta Cryst.,* **A56,** 375-382.
Lunin, V., Lunina, N., Podjarny, A., Bockmayr, Urzhumtsev, A. (2002) *Z. Kristall.,* in press.
Moras, D., Lorber. B., Romby, P., Ebel. J.P., Giege, R., Lewit-Bentley, A., Roth, M.J. (1983) *Biomol. Struct. Dyn*., **1**, 209-223
Ruff, M., Krishnaswamy, S., Boeglin, M., Poterszman, A., Mitschler, A., Podjarny, A., Rees, B., Thierry J.-C. & Moras D. (1991) *Science,* **252**, 1682-1689.
Urzhumtsev, A.G., Podjarny, A.D. & Navaza, J. (1994*). CCP4 Newslett*. **30**, 29-36.

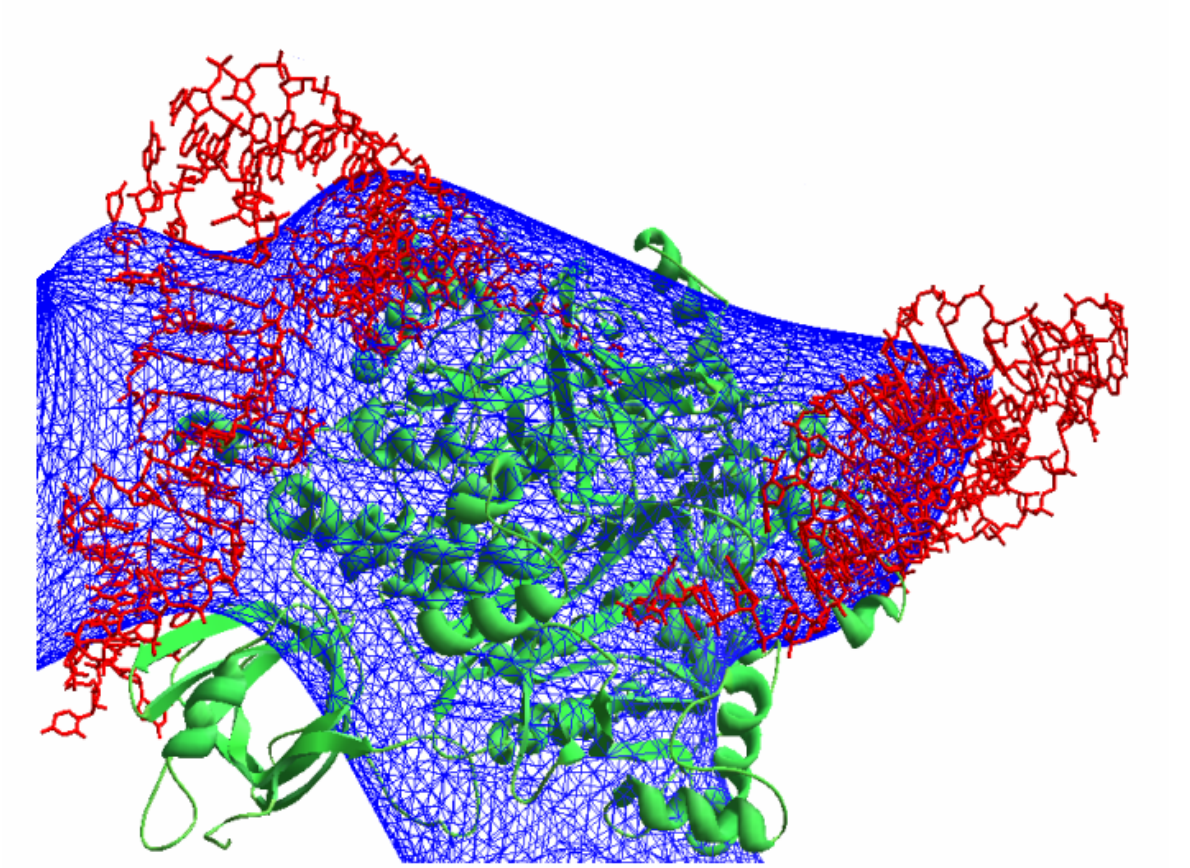


Fig. 1 Synthesis calculated at 45 Å resolution (37 reflections) using experimental modules corresponding to the full complex molecule and phases obtained *ab initio.*
Protein molecule is shown in green and tRNA molecules are shown in red.

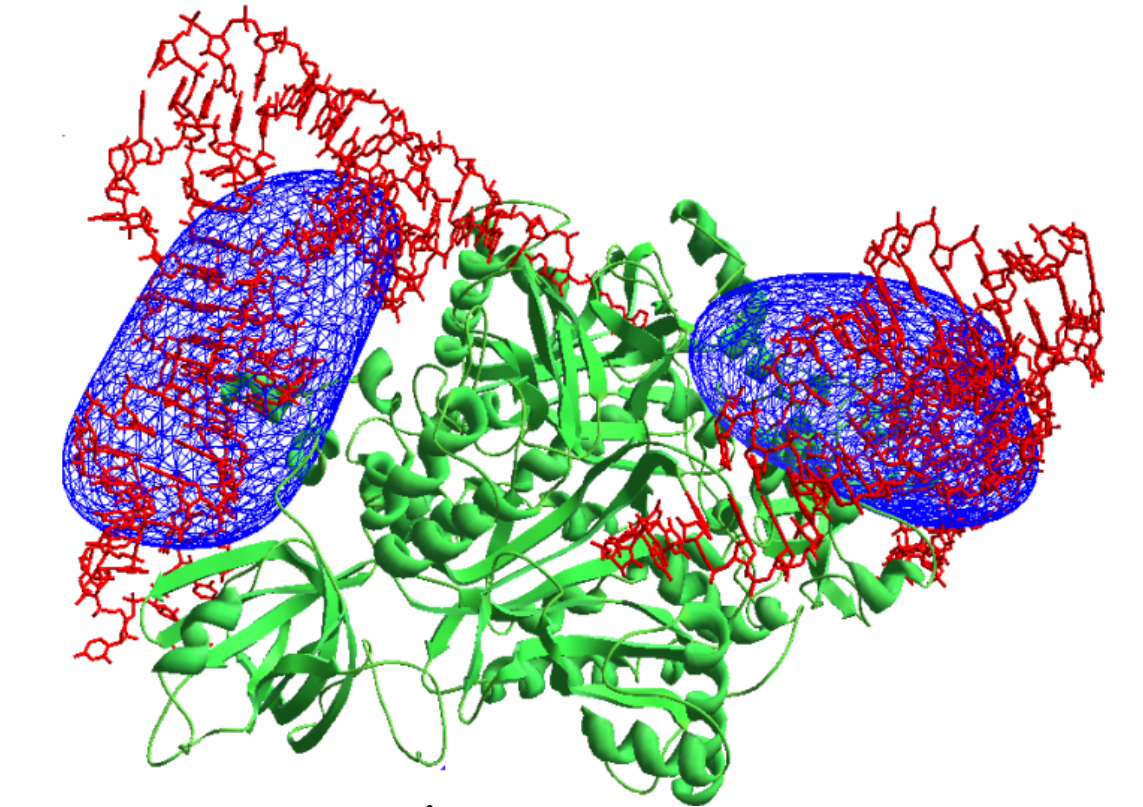


Fig. 2 Synthesis calculated at 45 Å resolution (37 reflections) using experimental modules corresponding to the tRNA molecules and phases obtained *ab initio.*
Protein molecule is shown in green and tRNA molecules are shown in red.

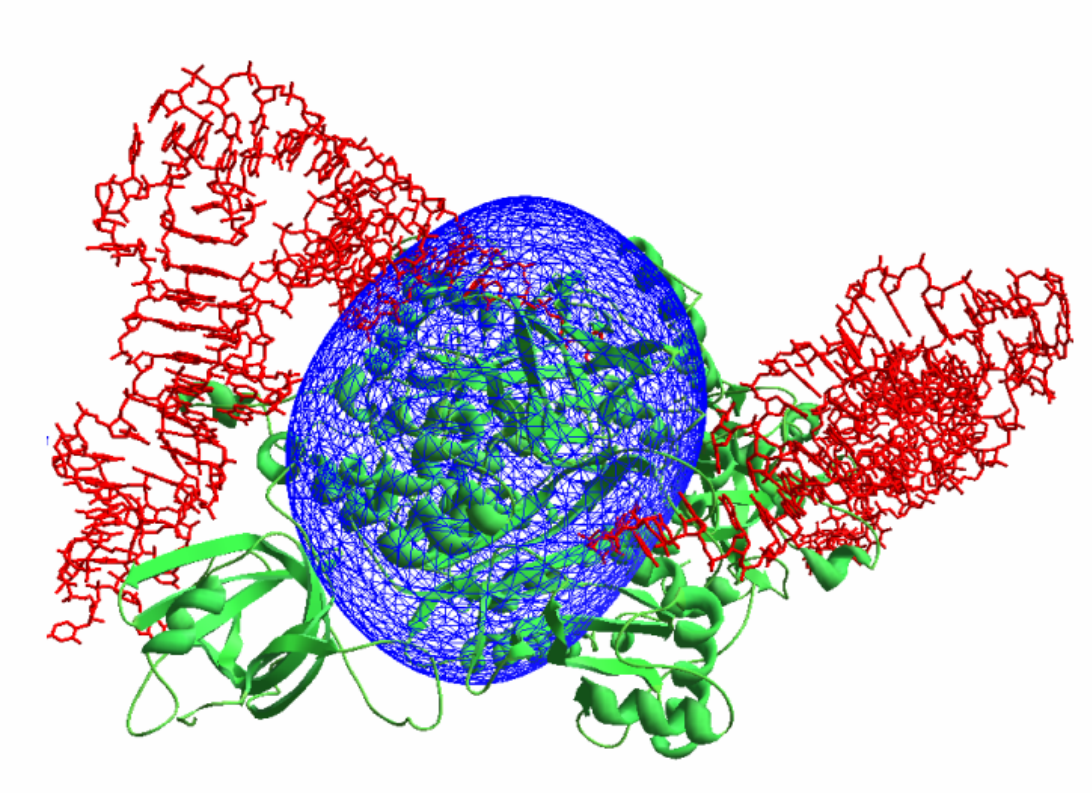


Fig. 3 Synthesis calculated at 45 Å resolution (37 reflections) using experimental modules corresponding to the protein molecules and phases obtained *ab initio.*
Protein molecule is shown in green and tRNA molecules are shown in red.

Retrieval of lost reflections in high resolution Fourier syntheses by a ‘soft’ solvent flattening.

by Natalia L. Lunina[1], Vladimir Y. Lunin[1] & Alberto D. Podjarny[2]

[1]Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow region, 142290 Russia; lunina@impb.psn.ru .
[2]UPR de Biologie Structurale, IGBMC, BP 163, 67404 Illkirch, CU Strasbourg, France, podjarny@igbmc.u-strasbg.fr.

1. Abstract.

Accurate studies of a high resolution Fourier synthesis require the full set of structure factors to be used when calculating the synthesis. Structure factors with unknown phase or even with unknown amplitude may be restored with a reasonable accuracy through density modification methods. For the case of Aldose Reductase, measured at 0.9 A resolution, a special type of the solvent flattening was tested for restoring about 20 000 (10% of the full set) structure factors in 0.9Å resolution zone. This flattening is based on the connectivity analysis of the Fourier synthesis map and is applied to small ‘drops’ only.

2. Model-free structure factors retrieval.

The quality of Fourier syntheses maps depend both on the accuracy of magnitude and phase values of structure factors and on the completeness of the set of structure factors used to calculate the map. The impact of a relatively small number of lost reflections on the quality of low resolution syntheses was demonstrated in a number of papers and different ways to restore these lost values were discussed (Podjarny, Schevitz & Sigler, 1981; Lunin, 1988; Lunin & Skovoroda, 1991; Urzhumtsev, 1991). Figs. 1-3 give some examples of such restoring.

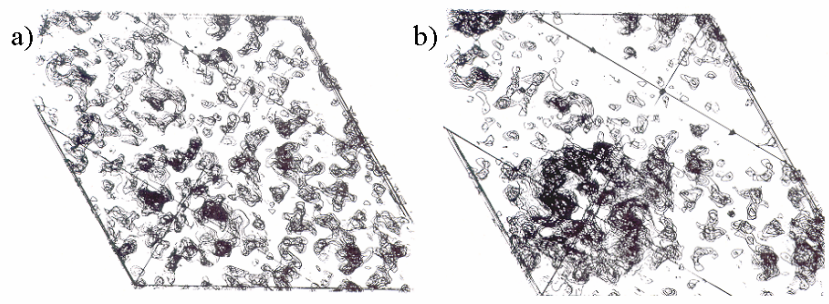


Fig.1. Comparison of ten sections of the 4.5Å map of yeast tRNA before (a) and after (b) the inclusion of 28 low-resolution terms whose phases were determined by matrical methods (Podjarny, Schevitz & Sigler (1981), Acta Cryst. A**37**, 662-668).

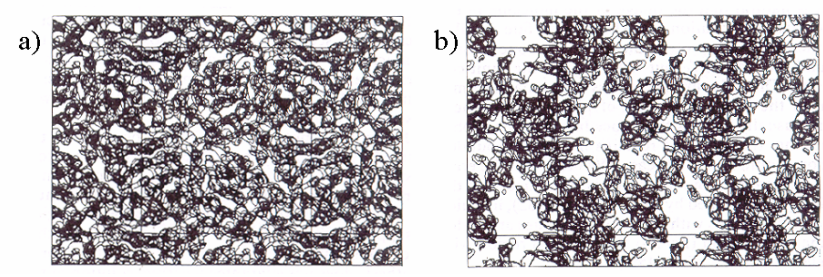


Fig. 2. (a) Initial (SIR) 8Å map for the elongation factor G and (b) the map with additional 29 low resolution reflections restored by the Double Step Filtration method (Urzhumtsev (1991), Acta Cryst. A**47**, 794-801.

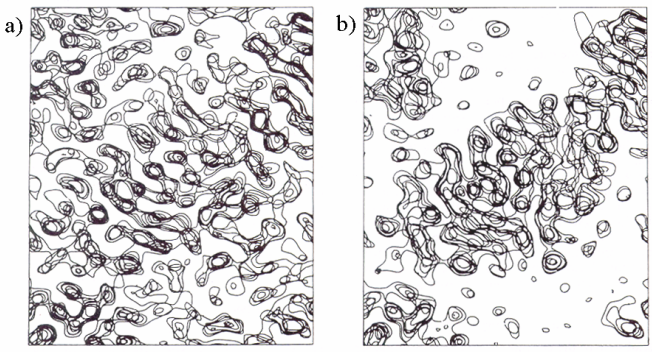


Fig. 3. (a) Initial 4 Å map for -crystallin IIIb and (b) the map with additional structure factors restored from electron-density histogram (Lunin & Skovoroda (1991), Acta Cryst., A**47**, 45-52.

The importance of complete data sets is not restricted to low resolution only, but is also true at subatomic resolution, where the electron density distribution in itself becomes the subject for studies. The missing reflections can be important when investigating these fine features at low contour levels. At the same time, the positions of peaks in high resolution Fourier syntheses are less sensible to missing reflections and may be found correctly even when some structure factors are absent (see Fig.5, 6 for comparison).

The retrieval of the lost structure factors may consist in either determination of their phases (if the reflection magnitude was measured, but was not phased) or restoring both the magnitude and the phase of structure factor (if the reflection was not measured). We discuss below the possibility of restoring both magnitudes and phases of the lost structure factors in the 'nominal' resolution zone. The extrapolation of data to higher resolution zones (Karle & Hauptman, 1964; Langs, 1998; Xu & Hauptman, 2000) is out of the scope of these notes. In high resolution density studies an atomic model of the studied object usually exists and might be used to calculated values of the lost structure factors. Nevertheless, such structure factors are highly biased towards the model used and do not usually show new features of the real electron density in a crystal, which can differ from the atomic model density. These circumstances require the developing of 'model-free' methods of calculation of structure factors not-determined in an experiment. Density modification methods (for a review, see *e.g*. Podjarny, Rees & Urzhumtsev, 1996, *e.g*.) provide ways to do this.

3. 'Soft' solvent flattening

The outlines of the procedure used are general for iterative density modification methods. Each cycle of the procedure consists of the following steps:

- the synthesis is calculated with the current set of structure factors (the weighted MAD-phased synthesis is used at the first step of the procedure);

- the modification of the synthesis is performed;
- new values of structure factors are calculated from the modified synthesis;
- a new set of values of structure factors is obtained combining calculated phases with the observed magnitudes (if these latter are known), and taking both the calculated magnitude and phase for reflections with unknown magnitudes.

The last step is different from the usual one for unmeasured reflections, for which the whole calculated structure factor is used to update the current values.

In our tests we applied the density modification based on new type of density flattening in the solvent region. The solvent density flattening belongs to the oldest methods of density modification (Bricogne, 1974; Wang, 1985) and is one of the most frequently used tools for phase improvement. Nevertheless, the usual goal of its application is phase refinement and only rarely magnitude restoring. The other feature of the usual approach is that the all density in the assigned solvent region is flattened. At the same time there may exist real density features in this region, which are not interpreted yet. Such density might be removed from the maps as a result of the 'total' solvent flattening. The procedure discussed below uses a softer type of density modification in the solvent region. It is based on the observation that the small drops in the maps represent usually noise, while real structural features are represented by more extended regions. The modification discussed consists in reducing of density corresponding to sharp narrow peaks, while the larger 'blobs' of a density in the solvent region are left unchanged. Such procedure combines the features of traditional density modification methods with the connectivity based phasing (Lunin, Lunina & Urzhumtsev, 1999, 2000).

Every step of density modification is defined by:

- a chosen cut-off level ρ_{crit} in the Fourier synthesis;
- a mask of the molecular region (*i.e*., every point in the unit cell is assigned either to the molecular region or to the solvent region);
- a limit size $N_{\min}$ of drops to be 'erased'.

First, the set of the points in the unit cell with the synthesis values $\rho(\mathbf{r}) \geq \rho_{crit}$ is analyzed. The goal of the analysis is to find the number and sizes of connected isolated components in this set. If a component has no common points with the molecular region it is considered as a solvent 'drop'. If the size of this drop is small enough (smaller than $N_{\min}$) the density values for all points in this drop are replaced by ρ_{crit}. All other points in the unit cell keep their previous values.

The parameters of the modification are generally updated from step to step.

4. Test object.

The tests were performed with a high resolution data set for aldose reductase. The crystals of human aldose reductase belong to space group $P2_1$ and have unit cell parameters *a*=49.97, b=67.14, c=48.02 Å, β=92.2^o. There is one molecule per asymmetric unit cell with the molecular weigh about 36 kDa. The crystals diffract to 0.66 Å (Lamour *et al*., 1999; Howard *et al*., 2000, Sanishvili et al, paper in preparation) and the structure was refined using SHELX (Howard et al, paper submitted) allowing to collect MAD data to 0.9 Å resolution. This data was phased using SHARP (D'Allantonia et al, paper in preparation). Such high resolution allows

starting density distribution studies (Guillot *et al*., 2000), which were carried before mostly with small molecules. Our tests were done at 0.9Å resolution, starting from MAD-phased data set.

All the theoretically possible reflections in 0.9Å resolution zone were divided in three sets. Set I consisted of reflections with the measured magnitude and the phase determined by MAD-method. Set II consisted of reflections whose magnitudes were measured, but the phases were not determined. Set III consisted of reflections with unmeasured magnitudes. The last was composed mostly from the very low resolution reflections and reflections of the highest resolution shell (see Table 1).

To calculate a Fourier synthesis, which represents correctly the electron density, it was necessary:

- to refine the phases in Set I;
- to determine the phases in Set II;
- to restore both magnitudes and phases in Set III.

In the first half of our test for reflections in Set II the goal was changed for restoring both magnitudes and phases. Then, the restored magnitudes for Set II were replaced by the observed ones, and the phases only were refined further. The reason for this was that at the first cycles of structure factors improvement the quality of the newly defined phases is very poor (see Fig.4.) and the use of real magnitudes values deteriorated the process. The restored magnitudes were significantly less than corresponding experimental values and this might be considered as a kind of weighting of such reflections.

Table 1. Distribution of reflections in resolution shells.

Resolution shell	Set I: MAD-phased	Set II: measured, unphased	Set III: unmeasured
49.28-4.01	2476	148	31
4.01-2.84	4454	248	0
2.84-2.32	5723	344	0
2.32-2.01	6803	356	0
2.01-1.80	7707	408	1
1.80-1.64	8413	485	0
1.64-1.52	9019	706	0
1.52-1.42	9835	589	0
1.42-1.34	10455	606	1
1.34-1.27	11058	706	0
1.27-1.21	11596	708	0
1.21-1.16	12162	734	0
1.16-1.12	12624	786	0
1.12-1.08	13085	817	0
1.08-1.04	13579	885	0
1.04-1.01	14031	888	0
1.01-0.98	14446	978	0
0.98-0.95	14719	1072	94
0.95-0.92	14582	1272	474
0.92-0.90	10292	1348	5082
Total	207 059	14 084	5 683

5. The choice of parameters of the modification.

The same mask of the molecular region was used during the whole test. It was constructed as the joint region formed by spheres with the radius 0.8Å centred at the atomic centres. The influence of the sphere radius on the progress in structure factors restoring will be discussed elsewhere.

For every cut-off level it is possible to estimate the mean number of points in the 'blobs' corresponding to the recognized water molecules. Approximately half of this value was used to set the parameter $N_{\min}$, which defines the minimal allowed size for the drops in the solvent region.

In the total, 1000 steps of iterative structure factors improvement were performed. The procedure was divided into several series of 50-100 steps in average. In every series, the cut-off level r_{crit} decreased step by step varying in the range from 1.5 – 1.2 'sigmas' in first series to 2.2 – 0.45 in last ones.

6. Test results

The structure factors calculated from the atomic model cannot be considered as the final goal of the restoring because one of the purposes of high resolution density studies is to find the difference between the real density distribution and that calculated from the model. Nevertheless, these values may serve as a good reference values when restoring structure factors. The increase of the correlation of restored structure factors (of the Set II) with the model structure factors is shown at Fig. 4 for different resolution shells. This figure shows that the quality of restored structure factors significantly grows in the course of modification.

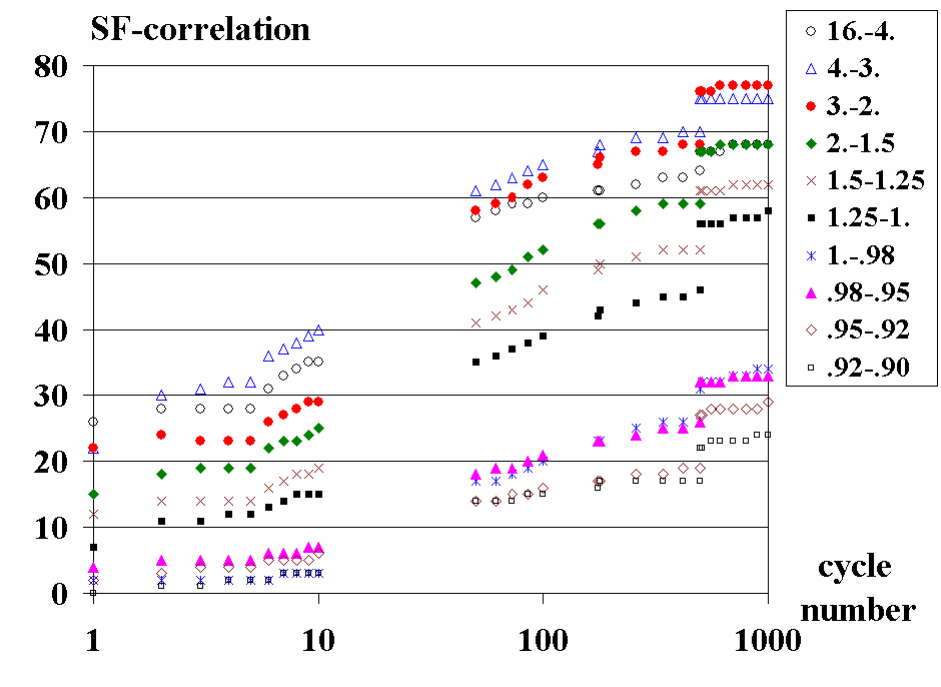


Fig. 4. The growth of accuracy of the restored structure factors for the Set II in the course of the restoring is shown for different shells in the reciprocal space. The shown SF-correlation was calculated as

$$C_{SF} = \frac{\sum_{s} F_s^{mod} F_s^{calc} \cos\left(\varphi_s^{mod} - \varphi_s^{calc}\right)}{\sqrt{\sum_{s} \left(F_s^{mod}\right)^2 \sum_{s} \left(F_s^{calc}\right)^2}} * 100\%$$

where $\left(F_s^{mod}, \varphi_s^{mod}\right)$ are the structure factors calculated from the atomic model and $\left(F_s^{calc}, \varphi_s^{calc}\right)$ are calculated from the modified density maps.

Sections of Fourier syntheses maps before and after modification are shown in Figs. 5 and 6. In contrast to high cut-off values (Fig.5), the low cut-off maps are sensible to the completeness of the set of structure factors and to the phases accuracy (Fig.6).

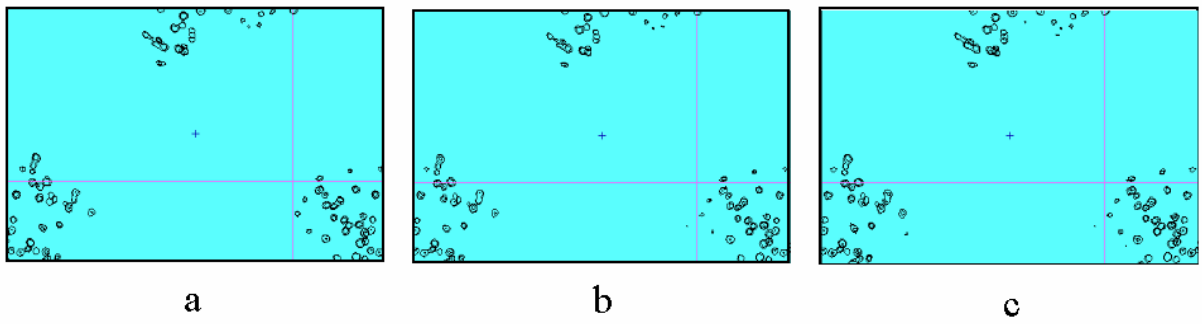


Fig. 5. Aldose reductase Fourier syntheses of 0.9 Å resolution, '3-sigma'-cutoff. a) MAD-phased synthesis, 10% of reflections are absent; b) refined phases, 10% of reflections are still absent; c) refined phases, the restored values of structure factors are added for 10% of reflections absent in (a) and (b). The difference between the syntheses is extremely small (see Fig.6 for comparison).

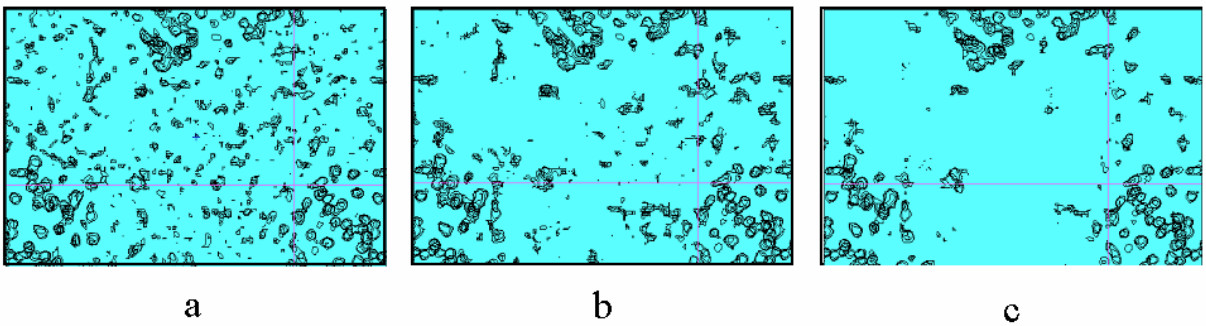


Fig. 6. Aldose reductase Fourier syntheses of 0.9 Å resolution, '1-sigma'-cutoff. a) MAD-phased synthesis, 10% of reflections are absent; b) refined phases, 10% of reflections are still absent; c) refined phases, the restored values of structure factors are added for 10% of reflections absent in (a) and (b).

Fig. 7 shows the decreasing of the number of drops in the solvent region from the MAD-phased synthesis to the final one.

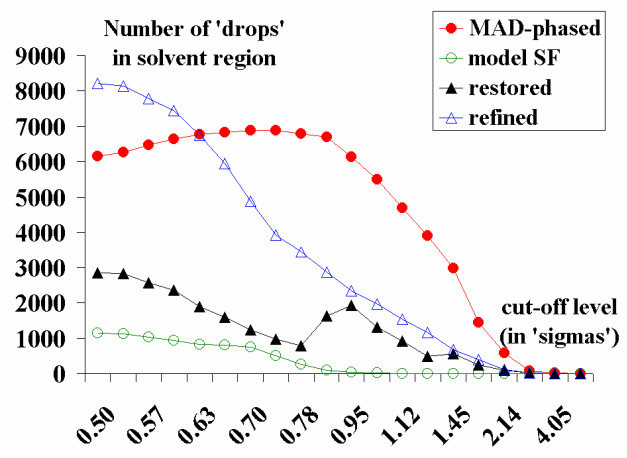


Fig. 7. The number of isolated components in the solvent region for different Fourier syntheses of 0.9 Å resolution: the start MAD-phased synthesis; the synthesis calculated with structure factors obtained from the atomic model (model SF); the synthesis with refined phases, but 10% of reflections absent (refined); the synthesis with all reflection restored and refined (restored).

One of the possible measures of progress in the structure factors restoring is the growth of mean value of restored structure factor magnitudes in comparison with the mean value of the measured magnitudes in the corresponding resolution shell (Fig. 8).

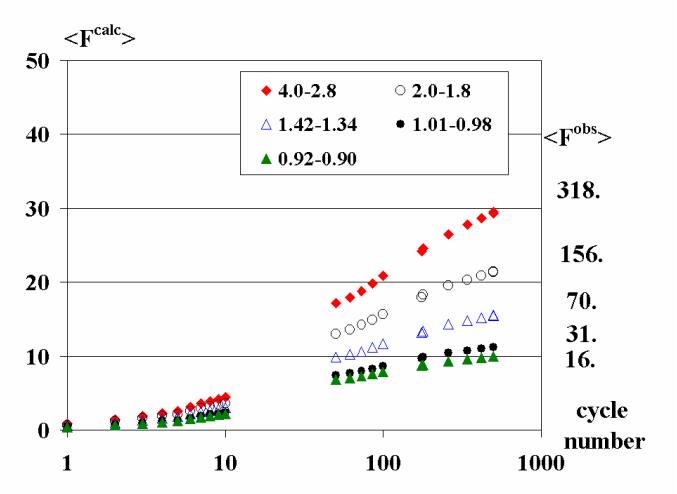


Fig. 8. The growth of mean values of calculated from the modified syntheses magnitudes in the course of the restoring for the Set II (several resolution shells only are shown. The mean values of the observed magnitudes in the considered shells are shown at the right.

The correlation of restored magnitudes with the measured ones (if they are known) may serve as control value too (will be discussed elsewhere).

7. Conclusions

The test results show that a density modification allows restoring a large number of unmeasured or unphased structure factors. These structure factors improve the image of density distribution map and may be very important in high resolution density studies. Obviously, more advanced density modification techniques may be used as well to improve the power of the method.

8. Acknowledgments

This work was supported by CNRS-RAS collaborative program and RFBR grant 00-04-048175. The authors thank Alexandre Urzhumtsev and Claude Lecomte for valuable discussions and possibility to use LCM3B resources. We thank also Andrzej Joachimiak, Ruslan Sanishvili, for their help in collecting the Aldose Reductase data, and Thomas Schneider, Fabio D'Allantonia and George Sheldrick for their collaboration in refinement and MAD phasing.

References

Bricogne, G. (1974). Geometric Sources of Redundancy in Intensity Data and Their Use for Phase Determination *Acta Cryst*, A**30**, 395-405.

Guillot, B., Jelsch, C., Muzet, N., Lecomte, C., Howard, E., Chevrier, B., Mitschler, A., Podjarny, A., Cousson, A., Sanishvili, R. & Joachimiak, A. (2000). Multipolar refinement of aldose reductase at subatomic resolution. *Acta Cryst*. A**56** (Supplement), s199.

Howard, E. I., Cachau, R., Mitschler, A., Barth, P., Chevrier, B., Lamour, V., Joachimiak, A., Sanishvili, R., Van Zandt, M., Moras, D. & Podjarny, A. (2000). Crystallization of

Aldose Reductase leading to Single Wavelength (0.66 Å) and MAD (0.9 Å) subatomic resolution studies. *Acta Cryst*. A***56*** (Supplement), s57.

Karle, J. & Hauptman, H. (1964). Positivity, Point Atoms, and Pattersons. *Acta Cryst*. **17** , 392-396.

Lamour, V., Barth, P., Rogniaux, H., Poterszman, A., Howard, E., Mitschler, A., Van Dorsselaer, A., Podjarny, A. & Moras, D. (1999).Production of crystals of human aldose reductase with very high resolution diffraction. Acta Cryst. D55, 721-723.

Langs, D.A. (1998). Reinvestigation of the Use of Patterson Maps to Extrapolate Data to Higher Resolution. *Acta Cryst*. A**54**, 44-48.

Lunin, V. Yu. (1988). Use of the information on electron density distribution in macromolecules. *Acta Cryst*. A**44**, 144-150.

Lunin, V. Yu. & Skovoroda, T. P. (1991). Frequency-restrained structure-factor refinement. I. Histogram simulation. *Acta Cryst*. A**47**,45-52.

Lunin, V. Y., Lunina, N. L. & Urzhumtsev, A. G. (1999). Seminvariant density decomposition and connectivity analysis and their application to very low resolution macromolecular phasing. *Acta Cryst*. A**55**, 916-925.

Lunin, V. Y., Lunina, N. L. & Urzhumtsev, A. G. (2000). Connectivity properties of high-density regions and ab initio phasing at low resolution. *Acta Cryst*. A**56**, 375-382.

Podjarny, A.D., Schevitz, R.W. & Sigler, P.B. (1981). Phasing Low-Resolution Macromolecular Structure Factors by Matrical Direct Methods., *Acta Cryst*. A**37**, 662-668.

Podjarny, A.D., Rees, B. & Urzhumtsev, A.G. (1996). Density Modification in X-Ray Crystallography. In "*Methods in Molecular Biology"*, **56**, 205-226.

Urzhumtsev, A. G. (1991). Low-resolution phases: influence on SIR syntheses and retrieval with double-step filtration. *Acta Cryst*. A**47**, 794-801.

Wang, B.C. Resolution of phase ambiguity (1985). *Methods Enzymol*., **115**, 90-112.

Xu, H. & Hauptman, H. A. (2000). On the extrapolation of the magnitudes |E| of the normalized structure factors E. *Acta Cryst.* A**56**, 284-287.

Recent CCP4BB Discussions

Maria Turkenburg (mgwt@ysbl.york.ac.uk)
October 2002

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 try not to identify anyone involved: those that are interested in the full discussion can view the original messages (see the CCP4BB 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 inquirer. 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.

Refmac vs. ...

Refmac *vs.* ...

(January 2002)

I refined my structure using Refmac5 and CNS using the same set of Rtest reflections. Always, Refmac5 gave a lower R-factor compared to CNS using max likelyhood refinement. Has anybody else noticed this? Why would this occur?

I am sure we would like to say that this was because refmac is better - and of course it is.. BUT the R factor within a few decimal places you get is very much a function of your scaling algorithm and at low resolution the two can differ quite a lot.... Just for comparison - do you have resolution? did you use bulk solvent? TLS? etc etc..
Usually good indicator is behaviour. *I.e.* difference between initial and final values. To confirm the above: scalings are different.

At that point a question was added to this thread:

I am very interested in this comparison too. Can anybody give me some details about the comparison between Refmac5 and CNS? I have never used Refmac5 before. Can refmac5 do simulated annealing?

Can some-one comment on how valuable simulated annealing is at various resolutions? In examples at better than 2.5Å we usually seem to finish up doing just about the same amount of rebuilding as you do after a ML run.
At resolutions better than ~2.0Å there is not doubt in my personal (and slightly biased, see also Warren/PyMol) opinion that ARP/wARP with Refmac(5) will outperform simulated annealing in CNS (or Refmac-on-its-own). It takes (significantly) longer, but no longer than a few hours that can be used for coffee/beer/sleep/more-hard-work (use the last option carefully, it can be bad for you). Most of the rebuilding can be eliminated by doing the autobuilding which will create a very good model (nearly perfect really) for the bits that it does build - lets say 80-95% of the structure. You have to do the rest on your own. And as a reminder there IS in ARP/wARP 5.1 a script/program to build and real-space refine side chains as well - most people seem to ignore that (there are at least two ActaD papers that explicitely state that 'ARP/wARP does not build side chains' !!!).
I am not 100% convinced of what to say for 2.2 - 3.2Å. There have definitely been cases that ARP/wARP with Refmac has produced at that resolution range (even at 3.2Å, but with high solvent content) MAPS that were far better than various tries with CNS and simulated annealing. Especially when starting from incomplete models. HOWEVER, CNS produced the BETTER model ! (ARP/wARP will throw away atoms it does not like so the model is BAD !) Thus I would suggest to get the map with ARP/wARP - REFMAC and the model either from refmac(5) alone (for resolutions not much worse than ~2.7Å) or from CNS (especially if resolution is worse than ~2.7Å).
Initially, I was very enthusiastic when simulated annealing was introduced in macromolecular refinement. My enthusiasm quickly turned into skepticism, when in several projects with resolutions ranging from 1.7 to 2.5Å I looked both at the resulting structures and the electron density maps: I had to go through the whole structure again to correct all these little errors (side chain and main chain) that simulated annealing obviously has introduced; the electron density maps showed a lot of model bias, *i.e.* too-good-to-be-true density around the "refined" model, and virtually flat density for the missing parts. Thus, simulated annealing heavily over-fitted my structures, which was presumably the reason for the introduction of the concept of the Free-R factor and the use of multiple simulated annealing rounds with averaged electron density maps. I also tried the combination of torsion angle dynamics with maximum likelihood target, and it seems to give good composite-omit-maps, but apart from that, I don't see any real advantage over careful inspection of electron density (omit) maps and fitting by hand. Some people think, that it might still be very useful when the resolution is low (say, around 3Å), and if your model is a beta-strand with many short side chains (Ser, Thr, Val, Asx) that have to flip around simultaneously along with a slight shift in the main chain. But at that resolution, model bias is a very serious problem and it will be difficult to judge the results. At high resolution, simulated annealing should work very well, but there are more efficient algorithms available (ARP/wARP, for instance).
To summarize, my personal verdict is that simulated annealing in macromolecular refinement is a heavily overestimated technique.
Since there seems to be little defense of CNS in this debate, I thought I'd give my two cents worth. First, there is no doubt that both ARP/wARP and REFMAC are superb programs. However, many statements in this debate are of questionable validity. First, CNS (and earlier, XPLOR) were developed to address difficult refinement problems at low to moderate resolution (*i.e.* when data are very limited). So, to claim that simulated annealing is best suited for higher resolution refinements is untrue and misses the declared purpose of the approach - to minimize the probability of a poor initial model being trapped in a local minimum when there are not enough data to guide the refinement to the global minimum. This does not mean that you will never have to manually correct the resulting model, or to adjust the starting temperature for the refinement. It does mean that

you don't have to rely solely on manual rebuilding to correct potentially major errors in an initial model that has been built into poor maps. The value of simulated annealing for this purpose has been extensively studied and documented in the literature. In addition, the combination of torsion angle refinement and simulated annealing optimization reduces the dimensionality of the search space approximately 10-fold, vastly decreasing the risk of overfitting by eliminating bond angles and lengths as refineable parameters. If you have data in the 3.5 to 2.5 Å range, you have no hope of fitting these model parameters meaningfully, no matter how tightly they are restrained, and it is proper refinement practice to not fit them at all.

Secondly, the Free R was not introduced as a stopgap measure to correct for problems in simulated annealing refinement. It was introduced because reporting a conventional R value as an indicator of model quality is at best naive and at worst willfully deceptive. Again, this has been extensively documented in the literature.

In summary, a direct comparison of REFMAC-ARP/wARP and CNS is complex, and should include a more careful consideration of the type of refinement being performed, the amount of data available, and the reliability of the initial model. A collection of anecdotes about how these programs compare in a handful of refinements does not add meaningfully to a proper assessment of their worth.

A historical perspective... Simulated annealling was quite the rage in the late 1980's. At that time, most electron density maps were fit by someone who had never fit an electron density map (*i.e.* a student or post doc doing their very first map). The map-fitting programs had neither rotamer libraries, nor fragments of main chain, nor any database automation that could be used as a tool. Thus, I believe many coordinates initially were fit rather badly without regard to stereochemistry. We needed a good refinement program with a large radius of convergence to get atoms into the right position. X-PLOR fulfilled that need.

Nowadays, with the tools provided by fitting programs and the knowledge of our predecessors, fewer mistakes are made in the early interpretation stages and those that are made are often discovered quickly. Thus, there is less need for simulated annealing to get out of false minimums.

I'm still favouring CNS for conventional maximum likelihood refinement: it is very fast and produces both excellent models and electron density maps. And it has the better bulk solvent correction due to the missing solvent B-factor in the mask approach of REFMAC.

While one can continue work on the relative merits of CNS *vs.* REFMAC, I think I will focus on what I miss in ARP/wARP-REFMAC. ARP/wARP is terrible at using NCS. If you are working with multimeric proteins and at medium resolutions around 2.5Å, CNS - strict NCS works miracles. While I can definitely use REFMAC to refine, '`dm`' to average, and manually build one model and generate the next etc.. would it not be great if ARP/wARP could do it. Few people will dispute that 4 fold averaging at 2.5Å produces better maps than no averaging at ~2.0Å. CNS, simulated annealing and water picking do an absoltely wonderful job with strict NCS.

I would disagree only partially. ARP/wARP does not do a terrible job with NCS. It does NOTHING with it, it just disregards it. Otherwise you are right! It would be great to use NCS!

Since torsional MD became stylish I have been unable to make it run properly, since I didn't have whatever parameter file it wanted for haem. Somehow I managed to survive without it. And from the same contributor: I am not entirely satisfied with NCS under CNS. You have a choice between strict NCS (constraints) or NCS restraints. If you want to switch from one to the other, you have to build a new coordinates file and then re-run generate. This is a nuisance. And in both cases, the NCS is constrained/restrained by the NCS matrix, which itself is not refined. To do that, you have to build the whole multimer and put it through rigid body fitting. Why not allow the NCS matrix to be refined with

either/or constraints & restraints? And to bring in a dark horse, SHELX has a different approach to NCS which makes a lot of sense to me. NCS restraints are applied not to the position as defined by the NCS matrix, but by extra restraints on torsional angles. This is particularly appropriate for structures with hinges or other localized differences between monomers.
I pretty much shudder at my late 80's low R-factors which I annealed. I am sure CD will do me one day (but hey I deposited the data). I know that the real question everyone is dying to have answered is: "How many levels of recursion can you embed in a CNS or RefMac command file?" Simulated annealing looks very attractive from a theoretical as well as practical point of view so I tried it many times using default protocols and all kinds of variations. I never was happy about the result. I kept trying because it looked like it should work but still without any real success stories. These were 2.2 to 2.8 Angstrom resolution structures. My feeling is that simulated annealing always causes damage to your model in addition to improvements. The more accurate the model, the less improvements (simulated annealing) can be made but there is still the damage. I still think, that TNT (5F) is one of the most elegant and transparent refinement packages if it only had a good implementation of the maximum likelihood target (the one from Navraj Pannu works in principle but it didn't write out sigmaa-weighted electron density maps; BUSTER/TNT is an attractive alternative, but it is still in the development phase).
From my personal experience and from working with other CNX users on their problems comparing performance of the different programs is a valuable and productive way to learn when and how these programs can be applied. In fact, X-PLOR/CNS/CNX and REFMAC-ARP/wARP can be quite complimentary to each other depending on model quality, the amount and accuracy of X-ray data available and type of refinement. As has already been stated, ARP/wARP works the best with high resolution data, I have seen CNS/CNX do an excellent job with data in the (3.5-2.2) Å range, particularly when a starting protein model is relatively crude. In many cases loops can move more than 2Å to their correct positions and side-chains are correctly placed into the density. Another advantage that X-PLOR/CNS/CNX provides is the ability to write/merge various task files in order to make your own refinement protocols. To summarize, a direct comparison of REFMAC-ARP/wARP and X-PLOR/CNS/CNX should not focus on which program is better or worse but rather how one can benefit from using many programs with different data.

A related question:

I'd be interested in people's humble opinions on their experiences with nucleic acids in particular. Many of the backbone torsion angles aren't well defined in moderate-resolution maps (2.5ish), but they do matter. Especially for non-canonical structures, I've wondered if we're really dealing with them properly.

The word is that XPLOR/CNS is far better than REFMAC on refining nucleic acids. REFMAC (.. and PROLSQ back then) make the sugar backbone 'funny' according to some, the bases non-planar according to others. Partially true - I think.
Well, IMVHO (V stands for Very) "partially" is the good adverb here. It was not the case for me: Refmac (at that time it was Refmac4/Protin) did a very good job with my DNA structure.
As far as I can see, it is always wrong to restrain ring-puckers in sugar rings. This also answers a question posed a few weeks later: *Is there any way of restraining the sugar puckers of nucleic acid residues during refinement in REFMAC (I am using version 5.1.09 with TLS refinement)?*

Then, a long time coming...:

To enhance the (bio)diversity of this fight: how about SHELXL for high-resolution data?

And, again, a number of weeks later by someone else:

Has anyone had experiences of refining ultra-high (better that 1.0Å) resolution structures with refmac5 and SHELXL? How did they compare?

But no answers to those questions...

A question about not-so-high-resolution and the performance of SHELX:

I was able to fit the AA into the solvent-flipped density map from CNS - everything looks just wonderful except for one five-residue loop. However, SHELXL refinement does not like this model, returning R(free) around 50% at 1.8Å resolution. Any suggestions, anyone? Thanks very much! BTW, does this have anything to do with these two facts: 1) the space group is I432; and 2) the model is being refined against Se-Met data without merging Friedel mates?

SHELXL isn't very happy at 1.8Å - why not use REFMAC or CNS - both faster and more appropriate at this resolution?
I feel I must take exception -- at least, with the first part of that statement: SHELXL does a rather good job even at 3Å, particularly with geometries. But I can't argue with the 'faster' thing... lest this degenerate into a flame war ;-) As to why SHELXL craps out... hard to tell, it could be anything: mistakes in the ins file (are you sure your HKLF is correct), wrong formatting of the hkl colums...

Refinement weights

(January 2002)

While we're on the subject of refinement, with REFMAC, CNS, TNT, and SHELX, what do people do to adjust the relative weights. In CNS I typically refine with "wa=-1", which sets up a reasonable relative weight between the geometry terms and the xray terms for an incomplete model. Towards the end of refinement, I double or triple the value found with "wa=-1". I've found that with "wa=-1" various programs like WHATCHECK and PROCHECK find that the restraints are too tight. When I increase wa, then both the R(cryst) and R(free) decrease a bit; at some point R(free) doesn't drop anymore, even if R(cryst) does, so that's a reasonable place to stop. I also look at the rmsd on bond lengths (0.01 or lower as a target), and on bond angles (1.5 or lower as a target). In SHELX, I leave the weight as 0.2, and never mess with it, even with small molecule structures. The sigma's of the intensities seem to be correct from various packages (DENZO, bioteX, MOSFLM), since the small molecule structures gives GOF's close to 1.0 without the additional fudge (weight adjustment) in SHELX. Since I don't know what the systematic errors are in the model or the data, I don't believe that the GOF's should be close to 1.0, unless you know you have excellent data and an excellent model. I haven't used REFMAC or TNT in awhile, so I'm not sure how the weighting schemes are adjusted.

In my experience I get very good results in all different refinement programs when I set the relative weights such that the final rmsd for bond lengths is about 0.012-0.015Å and the rmsd for bond angles is around 1.5-2.0 degrees. Regarding the doubling of the wa in CNS: if you look at the scalenbulk module, which determines that weight, the refined wa is

divided by a factor of 2. This factor was introduced later (as Paul Adams told me), but I always remove it, because otherwise CNS refines the model with, in my experience, too tightly restrained geometry. In REFMAC use the command "WEIG EXPE MATR" followed by a value that has to be lower for tighter geometry. I usually end up with values around "0.5".
I suspect this is the factor two that goes back to the early days of Rfree, when many people playing around with WA in Xplor found that taking 1/3 or 1/2 of the WA value recommended by Xplor itself tended to give the lowest Rfree values (using simulated-annealing refinement). I would like to caution against letting "experience" or "validation programs" talking you into increasing WA mindlessly. Unless you have very high resolution data, you should restrain your geometry tightly (how tightly? ask Ms R Free). Dictionaries have target values and ESDs for bond lengths, angles, etc., but you should remember that these ESDs are calculated for a population of very-high-resolution small-molecule structures. There is *no reason whatsoever* for you to expect the combined bond lengths and angles in your 4Å, 3Å, or even 2Å model to reproduce the ESDs of that very special population. Of course, you can always relax the weight on the stereochemical restraints and obtain just about *any* ESD you like. *However*, you should remember that "where freedom is given, liberties are taken" - the extra slack you are cutting the geometry is likely to be distributed randomly (or rather, in such a way as to enable the largest drop in value of the refinement's target function). If Rfree tells you that this is a great thing to do - fine. if not, leave WA alone. Some pointers for the youngsters and/or novices:

- Kleywegt, G.J. and Jones, T.A. (1995). Where freedom is given, liberties are taken. Structure 3, 535-540.
- Kleywegt, G.J. and Brunger, A.T. (1996). Checking your imagination: applications of the free R value. Structure 4, 897-904. (and references therein)
- http://xray.bmc.uu.se/gerard/gmrp/gmrp.html
- http://xray.bmc.uu.se/usf/whatif.html
- http://xray.bmc.uu.se/gerard/embo2001/modval/index.html

Refmac FOM

(January 2002)

*Some rtfrm (r for refmac) yields FWT and PHWT are amplitude and phase for weighted "2Fo-Fc" map (2mFo-DFcalc) Sofar so good. Nice map. Not as nice as Shake&wARP, but it takes about a factor of 10**2 less time ;-). more rtfrm: FOM = - The "figure of merit" for this reflection. How do/can I use this particular FOM in a map? FP*FOM PHIC is quite a biased map. Where did m's buddy D go? Rtfp (paper) ? Also (Brent as well raised the question), qtfm (q for quoting): "Rebuilding into these 2mFo-DFcalc and mFo-DFcalc maps seems to be easier than using classic nFO-(n-1)FC and difference maps, consistent with the established technique for SigmaA style maps. One advantage here is that since the m and D values are based on the Free set of reflections they are less biased than the values obtained by the CCP4 version of SIGMAA after refinement". Ok I can see that for the first (no previous rebuild) map - but then in further cycles, haven't you actually taken info from your crossvalidation set and (real space) refined against it?*

Well - I only use FOM if I want to use the PHIC for some other purpose than rebuilding - for instance you hope now you can find your anomalous scatterers as markers for the building.. You would do a map DANO, PHIC , FOM. Or if you had a putative derivative : You could look at the difference map: Fph-Fp, PHIC, FOM. Otherwise I don't think it is very

useful. Guy likes to look at Fo FOM PHIC maps - and sometimes at lower resolutions they are better - estimating D at 3Å from incomplete data can be tricky!
The 2mFo-DFc coefficients do not include the Free R reflections; those terms are set = D*Fc, so there should not be corruption of the crossvalidation set.. That IS in the paper - maybe not in the fm.. But if you insist on using nFO-(n-1)FC maps, yes; the cross validation set will be corrupted, although there is an option in FFT to exclude them, not that I think many people use it!
As indicated in the previous reaction, the refmac map used D*Fc terms for Rfree reflections. This will make the purists happy, but I wonder if it is the best thing to do. Really, how big is the risk of introducing model bias by a clumsy human trying to best fit a piece of model into the density. We are not talking about clever software adjusting thousands of parameters to "force" a fit between model and observations. Perhaps lavishly adding waters to any positive density feature would introduce some bias but I hope us humans aren't THAT clumsy. The disadvantage of using D*Fc is that it doesn't contain ANY information about how to modify the current model to resemble the real structure more closely. It just contributes a weighted down echo of the Fcalc density. This may make the density a bit more beautiful and give the suggestion that your model fits the density better than it really does. So instead of model bias we create mind bias. Just imagine the silly example that you select 100% of reflections for the Free set leading to the use of D*Fc for all reflections. The map would probably look very good. My gut feeling is that the real purists should leave out the Rfree reflections completely and the practicalists would include them just like the working set reflection. The D*Fc option seems half-hearthed at best.

And, to confirm this:

Come on, guys - this is an ancient (non-)issue. Back in 1996, two "Rfree-cionados" observed:
"It is desirable to include all diffraction data when calculating electron-density maps to avoid truncation errors. In principle, the use of real-space refinement techniques could introduce some bias towards the test reflections, but the seriousness of this effect has not been demonstrated. Moreover, if simulated annealing is used throughout the refinement, any model bias is likely to be removed during subsequent refinement." I'm still not aware of any convincing demonstration (and not even of an unconvincing one, for that matter) of said effect, neither when it comes to manual rebuilding or real-space refinement. I suspect it's a red herring (and a sophistic one at that) :-)

The monomer libraries

Modified amino acids in Refmac5

(January 2002)

The problem is as follows...I am refining structures with refmac5 which contain modified (acylated) cysteine residues. Needless to say that no corresponding library entries exist in the standard set. I can "easily" make the library entries for the residues so that the correct geometry is retained, but refmac does not seem to handle these residues as being part of the chain, i.e. the C-N distances on both side of such a residue increase during refinement. Apparently, the program breaks my protein chain on both sides of the modified amino acid to start with. It also gives this error for each "broken" peptide bond in the beginning:

```
INFO:   connection is found (not be used) dist= 1.416 ideal_dist= 1.329
            ch:AA res:  88  LEU at:C    -->    ch:Aa res:  89  ACC at:N
```

And, of course, complains about the VDW outliers later on... I tried reading the fantastic manual, but was not able to extract the relevant info. I started from the cysteine residue entry, adding the atoms of the acyl group at the SG and so on but did not touch any of the main chain atoms there.

There are two ways for dealing with modified amino acid residues:

1. Using LINK record. You create dictionary for added group. Then use link between this group and amino acid it is attached to. You can just run refmac5 with MAKE LINK yes and it will create link for you if groups are close to each other. Then you can modify or keep your description of link. Description of link can have info about bond angle, bond length, deletion of atoms, addition of atoms etc. I like this option more as you keep your original residue name.
2. Create dictionary entry for the modified amino acid and declare it as L-peptide. This option works with refmac 5.1 (which is available from york's ftp ftp://ftp.ysbl.york.ac.uk/pub/garib/).

Using new library in Refmac5

(March 2002)

I am having problems refining my ligand in Refmac5 in ccp4.1.1i. I have 3 monomers and 3 CoA molecules in the pdb file. The monomers by themselves refined just fine, but I want to refine with the ligand. I don't want to use the refmac5 library for CoA, so I created my own from pdb using Sketcher. My question is, which library do I give refmac for LIB_IN? I am still refining protein, so I need mon_lib_prot.cif, but I don't want refmac to use its own CoA library in mon_lib_1.cif and instead use COA_mon_lib.cif that came out of Sketcher. I tried specifying this library when setting up refmac, but I get the error:

```
  Number of atoms    :    5238
  Number of residues :     672
  Number of chains   :       6
  I am reading library. Please wait.
                mon_lib.cif
  WARNING : COA      : program can not match library description....
                           program will create new complete description
  WARNING : COA      : default angle value will used
                  chem_type : P   -O2  -C     120.000 (P3  -O3' -C3' )
  WARNING : COA      : default angle value will used
                  chem_type : C   -CT  -NR5   109.500 (C2' -C1' -N9  )
.......... etc.
```

Which to me looks like it doesn't know what COA is. By the way, this new library works, I tried loading it in Sketcher and it correctly lists COA as the only non-polymer ligand with correct geometry and atom names. I guess a more general question would be how to make refmac use its default library for peptide, but user-specified library for ligand? Any help would be greatly appreciated.

If you give LIBIN and some of the monomer names coincide with the library, the program will take your monomer description. Reason for WARNING and creating new dictionary could be that your coordinates may not be ideal. To force the program to use the dictionary description without checking validity you can use

MAKE CHECK none

in the command line. Or using interface in the SETUP RESTRAINTS section you can specify 'do not check anything'. Then it should work OK.

LINK statement

(April 2002)

I am trying to link two sugar residues using the edit restraints in pdb file using the gui. I am getting a boolean error. The MODRES works fine. Is there actual example of this (LINK statement) somewhere? (I have gone through the RTFM but unable to find ...). Also, on the same issue, wouldn't it be greater if the CCP4 has examples of

1. *files to create a link between amino acid and sugar*
2. *files to create a link between two non-standard ligands*
3. *how to fix the conformation of a sugar say in 4C1 etc.*

I think these are very standard and will provide novices a greater appreciation of what can be achieved using Refmac5.
If you run Refmac with MAKE LINK YES command then it creates all necessary and unnecessary links. You remove unnecesary links and decided whatever you need. You are right that we should have kind of howto. I am planning, but have not done yet.

Mon-lib problem when using ARP/WARP 6.0 (CCP4i)

(August 2002)

*I have a few "new" ligands in a structure that don't have the full discription in the distributed monomer libraries. I managed to creat the full discriptions and Refmac5 would take them, in the "Library" entry of CCP4i/Run Refmac5, and refined well. However, when I tried to use ARP/WARP 6.0 with the option "improvement of model by atoms update and refinement", where do I put the new *cif file I generated? The default lib of course failed when Refmac5 started. BTW, I am using CCP4i in version 4.2.1.*

Arp/wARP 6.0 can read a REFMAC library file but only in 'solvent building' mode ('refmac parameters/use user defined library') - btw, that's where you can also choose to use TLS, one of my favorites these days. Indeed it can not read a user defined library while "improvement of model by atoms update and refinement". The reason that we chose that is that when you use the above protocol we presume you have either a crude model to improve or a missing ligand that you want to see if it will appear or not. If the ligand is already modelled I would not use that protocol and would discourage people of using it, thus I chose not to give the option.

Treatment of oxidized Cys in Refmac

I am trying to refine a structure with some of the Cys residues oxidized with one/two oxygen atoms. I looked around and found that these modified Cys are referred to as CSX and CSW. Furthermore, there are minimal descriptions for CSX and CSW in CCP4 lib. I tried to refine my model with CSX, but Refmac complained that the lib is not complete. I tried to use only Oxygen of CSX in a different residue name (say CS1) leaving Cys residue as normal, use LINK between Cys-SG and O-CS1, generate a library for CS1, and then use Refmac to refine. Still, error messages are:

1. `CS1 : last atom of the tree absent`
2. `Subscript out of range on file line 12838... Attempt to access the 0-th element of variable s2_conn[****]`

My question is: what is a clean and easy way to describe and refine oxidized Cys residues?
In general there are three ways of using modified residues:

1. Use original residue and add modification on them and use MODRES record in the header. Modification should be defined in the library (user's library). It is good for one or two atom additions.
2. Use LINK record to add a group and define link in the dictionary. LINK can handle complicated chemistry. It is my favourite as it leaves original residue in the pdb.
3. Define residue in the dictionary and use it. If it is amino acid then it should have type L-peptide. For others like sugar, RNA/DNA there are types also.

Some of the residues in the library have minimum description. *I.e.* only list of atoms, bonds and bond orders. When Refmac sees them it creates complete description and asks user to check and if satisfied to use it.

What is the CCP4 (refmac5) equivalent of an omit map? - leading to: do lower Rfactors after TLS refinement reflect a better model?

(January 2002)

I wish to claim that I have a certain molecule present in my active site that co purified with my protein (GTP as it happens). My initial thought was I'll calculate an omit map, but my structure is refined in refmac5 with TLS and some alternative conformations so to go to CNS to do a simulated annealing omit map is quite a hassle and the Rfactors as a CNS model of my solution is much higher. Would people believe maps out of refmac5 after say 20 cycles omitting the coordinates of the ligand? Is this the nearest CCP4 comes to an omit map? The alternative I thought of is to go back to my original MR map (based on a structure with ligand removed) before any refinement. Any suggestions welcome.

Answers fall into three camps:

1. Shake up the coordinates a bit (with pdbset or moleman2), omit the ligand and run refmac for a few cycles.
2. Use OMIT (CCP4 programme)
3. Display some early unbiased map. Ideally refine the protein as far as you can without building in any ligand.

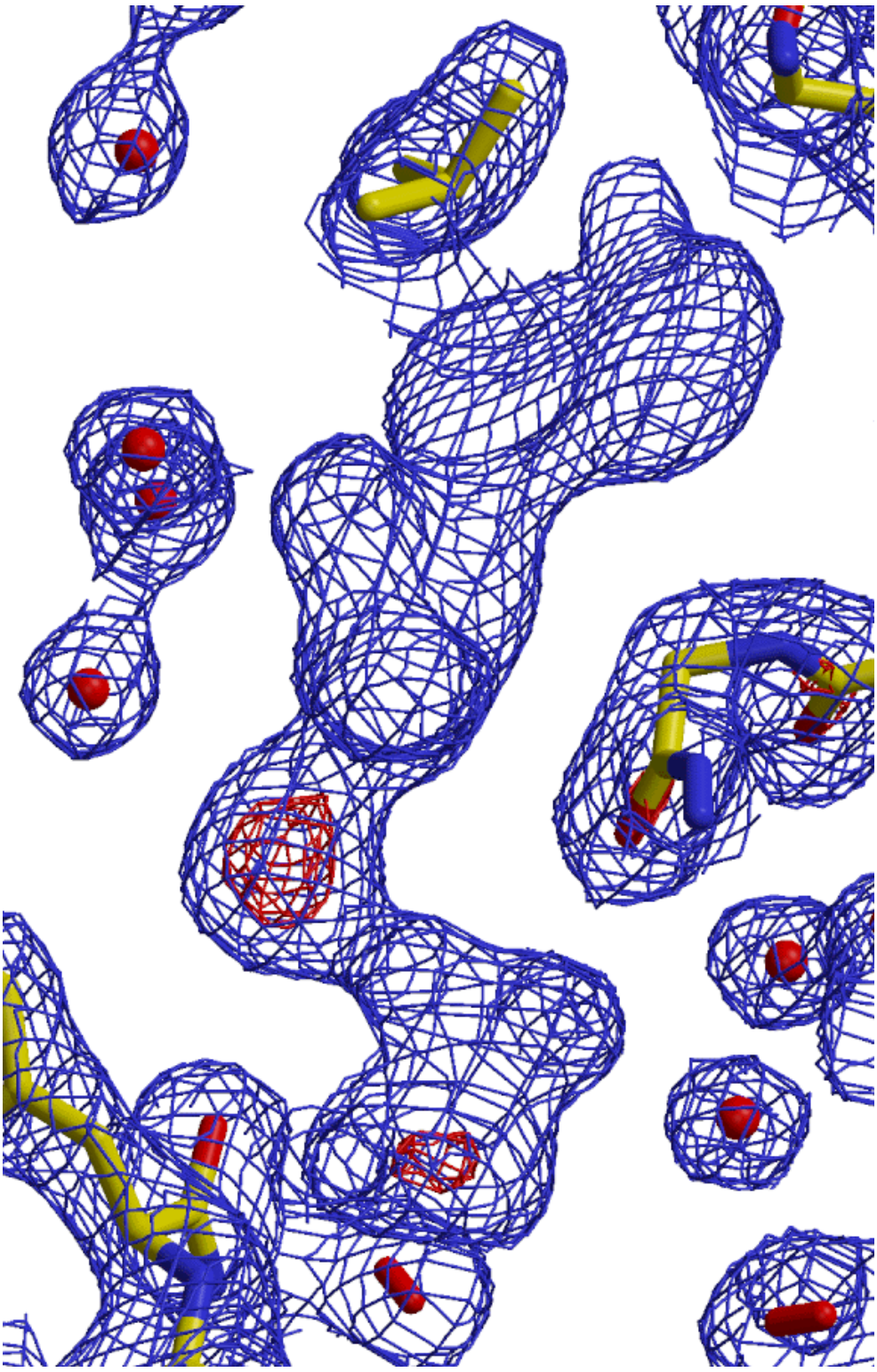
Oh and a suggestion to see if ARP/wARP puts atoms there (No guesses as to where that came from).

The last remark sparked the following:

That was NOT AT ALL what I suggested. I was referring to an idea/approach (which indeed uses ARP/wARP) which is described for example nicely in: 'Questions about the structure of the botulinum neurotoxin B light chain in complex with a target peptide' by Bernhard Rupp, Brent Segelke in Nature Structural Biology 8, 663 - 664 (01 Aug 2001) Seeing if ARP/wARP puts atoms in density is in general terms a bad idea and I would not recommend it to anybody that has the same problem.

Then an addendum to that from elsewhere:

I'd like to put that into different words - using the wARP model of course is not the idea, but to convince yourself of the power of wARP in reconstructing maps of ligands that are actually there



The remark about the different Rfactors from Refmac and CNS sparked another question:

*In pdb are a number of structures, where refmac TLS was used, (converted to Uij s) and these structures show *extraordinarily* low freeR and R for the resolution. So I wonder to what degree your higher rfactors of the CNS model reflect just that, and also, to what degree those low TLS rfactors reflect a truely better model?*

If you talk about coordinates then the TLS model doesn't need to be better than without TLS, even if the R-factors are significantly lower. However, if you consider B-factors as part of the model, and of course they are although our normal display programs don't visualize them, then it appears that the TLS models are better indeed, especially at medium and low resolution. The lower R-free also suggests better Fcalc and thus a better map and possibly less model bias. All of these could help to actually get better model coordinates as well.

Overlapping in Refmac

Overlapping ligands

(March 2002)

In refmac5, can one refine two different overlapping monomers - amino acid residues, ligands, metals etc.? Note that this is not the same as alternate conformations, in which the overlapping monomers are the same. I know how to do this in CNS/X-plor, but I wish to take advantage of TLS refinement.

If you give both groups partial occupancies, then the internal restraints for each residue/ligand are maintained, but there is no VDW clash. Similarly metal LINKs can include a partial_occup flag. So all you need is standard PDB format. What is harder are linked partial occupancies - *e.g.* LYSA + HOHsA, with appropriate VDW checks between them, and another conformation: LYSB and HOHsB.

Overlapping TLS groups in Refmac5 and TLSANAL?

(June 2002)

I wanted to try refinement of a multi-group TLS model as follows:

```
TLS group 1:          residues    1 2 3 4 5 6 7 8 9 10.....250
TLS group 2:          residues          4 5 6 7
TLS group 3:          residues                      110 111 112 113
```

and so on. What I am describing is a bulk libraration of the whole chain, superimposed on which is separation libration of smaller groups (turns, loops, individual sidechains). Refmac5 seems to deal with this just fine. The refinement is stable; R and Rfree are improved over a model with only a single TLS group. The Biso values in the output PDB file look plausible, and they track the Biso values from a single TLS model except for those residues contained in the additional TLS groups. However, when I try to run the resulting PDB file through TLSANL, I get the following error message:

```
*** ERROR: OVERLAPPING ISO/ANISO/TLS GROUPS:  1  6  ATOM:  N  A   8  102
```

So, two obvious questions:

1. *Does refmac5 in fact properly handle the case of overlapping TLS groups, as it seems to do?*
2. *Is there any reason why TLSANL cannot be made to process the result of such a multi-group TLS refinement?*

If it's just a matter of adding code to TLSANL I can have a look at it myself. If, however, there is some fundamental reason why this cannot work I'd rather hear about it before spending a lot of time on the problem. And if I've been mis-using Refmac5 then I'd better find that out also.

From the same inquirer:

The Biso values in the pdb file output by Refmac5 after TLS refinement are quite different from the Biso values output by TLSANL after converting the TLS description into individual ADP records starting from that very same Refmac5 pdb file. I realize that the Uij parameterization is only an approximation to the actual atomic scattering described by Biso+TLS, but I would expect the Uij parameterization to be chosen so that _something_ came out the same, and Biso (a.k.a. Beffective) is the logical quantity to hold constant.

- *What exactly is the quantity stored in the Biso field of ATOM records output by Refmac5 after TLS refinement?*
- *What is being fit (or minimized?) by TLSANL when it converts the Refmac5 output into a description of individual atomic ADPs?*
- *If I shouldn't expect Biso to be the same before and after running TLSANL, then what should I expect? Should it get larger? smaller? Does the direction or magnitude of change tell me anything about the nature of the TLS model?*

And then some more:

The light dawns. I think I understand after all. Please correct me if I'm on the wrong track. The Biso values in the Refmac5 are the raw values refined subsequent to the TLS model. They are independent of the displacements described by the TLS, which have not yet been applied (this was what I misunderstood originally). TLSANL converts the Biso and the TLS parameters jointly into a set of ADPs. Yes that's what the documentation says, but somehow I had the idea originally that Refmac had already combined them isotropically, and TLSANL was just correcting that to an anisotropic description. My other question (about overlapping TLS groups) still stands, however.

As to Biso: Exactly. What is called "residual B factors". TLSANL gives ANISOU which is U from TLS plus B added to diagonal. By default, the B in the ATOM line is 1/3 trace of the ANISOU line, but you can change this (ISOOUT keyword) - useful for comparing contributions. As to the 'overlapping TLS groups' question: I don't see any problem with this in principle, but this wasn't planned for. To be investigated later.

REFMAC *vs.* CNS SigmaA maps

(March 2002)

I have been looking at improving the current map I have made from MIR/MAD phases. Having traced 35% of the structure, I thought I'd give SigmaA a try. I did the following: REFMAC (5):

1. *Rigid body refinement using 25 chains (secondary structure)*
2. *Restrained refinement*
3. *Calculated and viewed the SigmaA weighted maps*

CNS:

1. *Rigid body refinement using 25 chains (secondary structure)*
2. *Simulated annealing*
3. *Calculated and viewed the SigmaA weighted maps*

My question is, why does REFMAC seem to have done such a good job of the mFo-Fc map, whereas in CNS the map looks like junk? I've compared the REFMAC mFo-Fc map with the original Fo map with MIR/MAD phases, looking at regions I suspected were secondary structure but didn't model in, and it suggests that there is little bias as these regions are improved. Could I have fallen into a trap, and CNS is giving me the right answer?
I am just wondering what you did with bulk solvent correction. Did you in/exclude it out in both cases, and what about the Babinet correction in Refmac, was that switched on or off. In such cases, would it help to include a bulk solvent mask based on the solvent flattening mask from the experimental phases?

And from the inquirer: Thanks for the response. I switched off bulk sovent correction in CNS as model completeness is so low (isn't that right, since the model is used to make the mask? - now I see your point about the solvent flipping mask). I've tried without bulk solvent scaling in REFMAC (I assume this is the same as correction - therefore the same reason for not including it). I thought that Babinet's correction was the same as Bulk solvent correction.

And again: Unfortunately the answer is pretty simple as to why the SigmaA CNS maps look so bad. It's coz the phases haven't been combined - the "SIGMAA" option just uses PHIC. Use the "COMBINE" option instead. Doesn't make much sense to me - but thanks to the person who pointed this out.

ARP/wARP Mode Solvent

(March 2002)

My question is addressed to those who have successfully used ARP/wARP for solvent building. Our model has R/Rfree of about 24.5/28.5 to 2.2Å resolution with no solvent built. Is it best to use experimental phases for 'mode solvent', i.e.

[Do you plan to use experimental phases as input (i.e. for mode warpNtrace or warp) (Y/N) ? Y]
Secondly, any suggestions on which REFMAC protocol is most appropriate in our case (shown below)? I would assume either 'R' or 'W' but was curious if anybody has previous experience with others.

```
----------------------------
  You can choose between the following REFMAC protocols:
    F   A fast protocol that works with good data.
    S   A considerably slower one which might work better in difficult cases.
    R   The slow protocol together with Rfree.
    P   Phased maximum likelihood refibement.
    O   The good old SFALL ...
    H   Optimised parameters for starting from heavy atoms alone.
    W   Optimised parameters for solvent building.
    A   Advanced mode for setting parameters manually.
  What is your choice  ? (F/S/R/P/O/H/W/A) W
----------------------------
```

The reactions are a little mixed:

Use experimental phases? Always -- unless your phases REALLY suck, I guess. Not using phases is like chucking half of your observations. (Well, maybe not EXACTLY half. But it's the idea that counts.) As to the Refmac protocol: P: phased refinement.
In principle, all prior information should be usefull during refinement. In practice, I have no clue if experimental phases will add much information when you are allready in the position that you can add waters (unless your phases are extremely good I guess). Besides that, one has to be sure that the HLC's are okay. Was this particular case an (M/S)(IR/AD) or MR model? I guess there are some people out there that have some answers based on experience in stead of 'gut-feelings' ... The fine manual says on this point:

```
/*----------------------------------------------------------------------*/
Do you plan to use experimental phases as input (i.e. for mode warpNtrace
or warp) (Y/N) ? Y
Amplitude (weighted) for initial map calculation: FSE1
Phase for initial map calculation: PHIDM_123p
FOM. Press  if amplitude is already weighted : FOMDM_123p

First the number of residues in the asymmetric unit is entered. Since we
want to start from experimental phases the answer was Y.

>> Answering with N
   means that you are interested in either starting from a
   molecular replacement solution, building the solvent of a refined
   structure or trying the ab initio option.  <<

/*----------------------------------------------------------------------*/
```

Protocol wise, I would choose W or write my own script so one can tweak around a bit more because (I_hate_automated_scripts option).

Refmac and prior phase information

(April 2002)

After we run Refmac with "prior phase information", what are the phases that we should use to calculate the final map ....should they be the usual PHWT or PHCOMB .... I have a protein in spacegroup p6522 with a very high solvent content .... I ran 'dm' for solvent flattening and the density improved considerably.....what is the best way of using this improved density after model building ....do I run Refmac with "prior phase information" and use PHCOMB or is that wrong ...

PHWT and PHCOMB are all COMBINED phases - combining the calculated phase from the model, and the experimental phase information. You should use FWT PHWT to get a 2mFoDFc map. The PHCOMB would only be used if you needed to use an overall phase for a heavy atom difference map or some non-standard purpose. Your "prior phase information" for REFMAC5 should be PHIDM FOMDM or the HL coefficints from DM. You may need to "blur" this phase information by scaling down the FOMs - try phase blur 0.7. Then again look at a map with FWT/PHWT.

R factors from Refmac and SFCHECK

(May 2002)

I have a question regarding Refmac5 and Sfcheck. I used Refmac5 with TLS refinement. After Refmac5, I run TLSANL to add TLS contributions to the isotropic B factors as well as to add "ANISOU" lines for anisotropic B factors. Please correct me if I'm wrong. My questions are:

1. *When I run SFCHECK on the output pdb of TLSANL, it calculates R factors ~2% higher than those from Refmac5 output. I think part of this is because SFCHECK does not use ANISOU records to calculate Rs. Is there a way to include anisotropic B factors when calculating R factors?*
2. *Are there other reasons that result in the difference in the R factors from Remac5 and Sfcheck? If I delete the ANISOU records from the output of TLSANL and use the resulting pdb as input pdb for another round of Refmac5 (thus no ansiotropic TLS contributions, just for checking R factors), I notice the starting R factors is only ~1% higher than those from previous Refmac (with TLS), but still ~1% lower than those from Sfcheck. One reason I can think of is that maybe Sfcheck uses a different way in scaling and bulk solvent correction. But it seems to me the difference shouldn't be so big.*
3. *What criteria do people use to judge whether to report anisotropic B factors? Only with very high resolution? I only have 1.9Å data, if I leave them out, what R factors should I report, those from Remac5 or those from Sfcheck?*

First off, your obsevrations are quite correct. I tried for an example here and got R factors:

```
Refmac + TLS                          0.17
SFCHECK                               0.24
Refmac with TLS-derived B's only      0.21
as previous, with bulk solvent off    0.24
   (SOLV NO - can't do this in GUI!)
```

I'm pretty sure that SFCHECK doesn't use ANISOU lines, so that explains a lot of the difference. Also, SFCHECK uses a very different scaling function. Refmac5 uses a mask generated solvent correction in addition to the Babinet-style correction. SFCHECK certainly doesn't have the former, and removing that (4th number above) account for the rest of the difference. Note that the 3rd and 4th numbers are without any refinement. Some refinement would lower the difference, as the model without TLS and bulk solvent correction will adapt to their absence. Re: your 3rd point. Certainly quote the R factor from Refmac5. That is the only program which uses your full model. R factors calculated from only a subset of your model parameters are not an accurate reflection of your model. For submission to PDB, you should submit TLS parameters (Refmac5 includes these in PDB header) not ANISOU lines, since this is your model. But you can discuss aniso U's provided they are clearly flagged as derived values rather than refined values.

Non-crystallographic symmetry

Nonx restraints on split residues (Refmac5)

How can I set up non-crystallographic restraints involving only one half of (a) residue(s) in dual conformation? In refmac5. Specifically, I need to restrain:

```
A 235(A)-239(A) -> B 235-239
A 235(B)-239(B) -> C 235-239
```

NCS restraints can only be put on atoms which conform to the following: Residue numbers and names must be the same, atom names must be the same and alternative codes must be the same. *I.e.* if you want to restrain residue A 235 alt A with corresponding B chain residue then they must have the same alt code.

NCS in Refmac5 - troublesome zinc

(July 2002)

Thanks for earlier advice on how to apply TLS parameters to my protein, which has lowered both the R and Rfree by 1 and 2%, respectively. I've got six molecules in my asymmetric unit (one TLS per molecule), and now I want to move on to applying both NCS restraints (which I've found, at least in CNS, to greatly improve stereochemistry and lower the Rfree) along with TLS restraints. Unfortunately, I've hit a wall.

The problem is everytime I include NCS restraints, the program chokes and gives me a 'problem with NCS' message. However, I've found that this problem is eliminated by getting rid of the catalytic zinc ion in my pdb file. Therefore, I don't think there is anything wrong with the NCS syntax - rather, there is some other problem lurking in the shadows. Any explanation for this phenomenon would be greatly appreciated.

Because my pdb is coming straight out of CNS, which I've been using for some time, there must be something I'm missing in my library file. I'm simply letting Refmac5 read in the old CNS pdb file and then create the appropriate .cif file. Specifically, do I need a special library file to describe the Zn+2 ion and its coordination sphere? The zn ligands are unusual - an Asp(2 bonds), Cys and His (1 bond) and a substrate ligand, which can have either a sulfur or a nitrogen. The zn appears in the pdb file as follows:

```
ATOM   5423 ZN+2 ZN2 Z   1      26.149  91.599  68.426  1.00 45.67      Z
```

I have tried other notations for the Zn in the pdb file (taken from deposited pdb files), but to no avail.

Then an unrelated question in the same email:

Also, does anyone know of any programs that can create rendered/anti-aliased figures that show the Richardson's contact dots? Xfit, although it displays them, doesn't seem to send them to Raster3D for rendering.

Zn as element is in the dictionary. If you want to add links then you can run refmac5 with
MAKE LINK YES
The program will then write all possible links to your pdb file and will create a CIF file. Then you can edit and use these links for restraints. With regards to the NCS problem: if you are using SGI and CCP4-4.2 could you please try Refmac from York's ftp site ftp://ftp.ysbl.york.ac.uk/pub/garib/. There is a slight problem with Refmac5 on SGI in CCP4-4.2. It is being dealt with.

Problem in MAKE_U_POSITIVE

Problem with MAKE_U_POSITIVE was reported four times this year, not all with an appropriate answer on CCP4BB.

1. *I am at the end of refinement of my structure. Things went great. then now that my R-factors are 15.5 and Rfree is 17.0 (data to 1.1Å resolution) refined with anisotropic b factors, I am suddenly getting*

   ```
   Problem in MAKE_U_POSITIVE -0.1387620
   ```

   How do I find out the offending atom(s) - or did something else go crazy. The geometry are all well behaved. I have 8 Zn+2 in my structure and their equivalent isotropic B's are positive.....

2. *I've a problem with refmac5. I use TLS refinement and after rebuilding model and start new refinement program stops work and give me information:*

   ```
   Problem in MAKE_U_POSITIVE  -59.93923
   ```

   Where is a problem? Is it a problem with my structure?

3. *I was using Refmac_5.0.36 on an alphaserver including TLS refinement.*
   A typical cycle of refinement was like this:

```
4.   ------------------------------------------------------------------------
5.   Overall                 : scale = 0.705, B = -0.594
6.   Babinet"s bulk solvent  : scale = 0.316, B = 200.000
7.   Partial structure 1     : scale = 0.710, B = 15.788
8.   Overallanisotropic scale factors
9.      B11 = -3.71 B22 = 9.88 B33 = -6.17 B12 = 0.00 B13 = 0.00 B23 = 0.00
10.
11.  Overall sigmaA parameters : sigmaA0 = 0.903, B_sigmaA = 4.751
12.  Babinet"s scale for sigmaA : scale = -0.001, B = 150.000
13.  SigmaA for partial structure 1: scale = 0.150, B = 0.001
14.  ------------------------------------------------------------------------
15.  ------------------------------------------------------------------------
16.  Overall R factor           = 0.2633
17.  Free R factor              = 0.3039
18.  Overall figure of merit    = 0.6529
    ------------------------------------------------------------
```

   Then we decided to upgrade to Refmac_5.1.19
   The same input files, the same alphaserver gave:

```
 Problem in MAKE_U_POSITIVE -1.4325634E-02
------------------------------------------------------------------------
Overall               : scale = 2.000, B = 0.000
Babinet"s bulk solvent: scale = 0.881,B= 50.000
Partial structure 1: scale = 0.350, B = 70.000
Overallanisotropic scale factors
   B11 = -3.73 B22 = 10.03 B33 = -6.30 B12 = 0.00 B13 = 0.00 B23 = 0.00
------------------------------------------------------------------------

------------------------------------------------------------------------
Overall R factor        = 0.2975
Free R factor           = 0.3439
Overall figure of merit = 0.5353
------------------------------------------------------------------------
```

   (and the Rms Delta for distances and angles slighty higher) And finally at the second case the final statistics went even higher. I tried with the same input files then on an SGI where Refmac_5.1.27 is installed. Here the first 2 cycles of TLS had

slighty lower R's and FOM(!) but the TLS values where different. At the 3rd TLS cycle started problems messages:

```
Problem in MAKE_U_POSITIVE -2.6608959E-02
Problem with atom in ELDEN 3140 (several times) and
Problem with atom in GRAD  3112 etc.
```

I would appreciate any help or hints. Moreover, if this is a compilation problem, is it possible and other CCP4.2 programs to produce such problems? (for example arp/warp interacts with refmac).

19. *I seem to have a problem in my refinement. I get a solution to my data using molecular replacement which seems very correct. When I refine I get the following error:*

    ```
    Problem in MAKE_U_POSITIVE -0.351249039
    ```

    I have done a similar procedure using the same model on different data, and have never seen this error. Can anyone tell me what it is? Is the problem in my data, model, bug in CCP4?

The first possible solution is:

Run the output PDB file through the PARVATI web server for analysis of the anisotropic refinement, a list of problematic atoms, and mug shots of offending residues.

The second query sparked the following:

Basically, you get this error message if an atom has an eigenvalue of U (anisotropic displacement parameter) lower than a minimum value. If this eigenvalue is in fact negative, then you have a non-physical U - a thermal ellipsoid that's disappeared up its own principal axis ...
Normally, this is just a warning. The program adds an isotropic component and continues. Subsequent cycles may or may not converge to physical values. Program TLSANL will generate U values and tell you which are non positive definite. However, -59 is very non-physical and suggests major problems. What TLS values do you get? I guess incomplete or incorrect model, in which case defer TLS until later.

TLS refinement

TLS refinement - which tls file to use in subsequent cycles

(June 2002)

I am working with data of protein: two chains (155 AA) in AU, space group 18. I am using Refmac5 with TLS refinement. I've got a question about it: At first stage of TLS refinement I use a file tls.in, in which I described information about my protein (number of chains, parts of protein to TLS refinement). Output file from refinement is tls.out, which gives me information about TLS tensors. What file I should use to do the next stage of refinement after rebuilding the model? tls.in or tls.out with tensors? When I use tls.in again, I obtained better R/Rfree factor than for tls.out

I usually use first tls.in with no T L S matrices. Usually convergence reaches in a few cycles. Reusing of previous cycle TLS parameters is a good idea if coordinates don't change.

TLS refinement - how to describe the TLS groups

(June 2002)

I can get Refmac to refine my structure just fine, but the TLS part of the refinement doesn't seem to work. About my project: My crystals contain 6 molecules per asymmetric unit. Each molecule is a heterodimer consisting of an alpha and beta subunit (molecule 1 = chain A + chain B, mol2 = chain C + chain D, etc., etc.). My current model has been fully refined in CNS with Rfree and Rcryst at 24.0 and 21.4%. Waters and ligands are included in the model. My data set is complete out to 2.4Å. The average B factor is ~55 Å^2. The protein N-terminal domains have not been modeled, and some side chains have not been modeled because they have no omit density (Is this important?). I want to see what happens if I add TLS parameters to each of the six molecules in the asymmetric unit. I have not included water or ligands in each TLS group (should I?). The starting TLS input file is as follows:

```
TLS mol 1
RANGE 'A 60.' 'A 360.' all
RANGE 'B 20.' 'B 360.' all

TLS mol2
RANGE 'C 60.' 'C 360.' all
RANGE 'D 20.' 'D 360.' all

TLS mol3
RANGE 'E 60.' 'E 360.' all
RANGE 'F 20.' 'F 360.' all
```

etc., etc., I run Refmac5 (5 cycles TLS refinement, 5 cycles restrained ML refinement), and I see no error messages during the initial TLS refinement (with B values fixed at 50 Å^2). Refmac spits out perfectly fine .pdb, ...tls and .lib files. However, during TLS refinement, refmac always picks the center of mass for each TLS group as 0,0,0 and assigns TLS values of 0 for all paramters. The refined .tls file for the six groups is as follows:

```
TLS mol 1
RANGE 'A 60.' 'A 360. 'all
RANGE 'B 20.' 'B 360. 'all
ORIGIN 0.000 0.000 0.000
T     0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
L     0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
S     0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

TLS mol2
RANGE 'C 60.' 'C 360. 'all
RANGE 'D 20.' 'D 360. 'all
ORIGIN 0.000 0.000 0.000
T     0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
L     0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
S     0.0000 0.0000 0.0000 0.0000 0.00000.0000 0.0000 0.0000
```

etc., etc., What am I doing wrong? It's probably something very simple, and related to the fact that the PDB file comes straight out of CNS. Any suggestions would be appreciated.

The answer is simple but not necessarily easy to spot:

If the centre of masses come out as 0,0,0 then that certainly means the program thinks there are no atoms in your TLS group. Which in turn means there is something wrong with the group definitions. The line should read:

```
RANGE 'A  60.' 'A 360.' ALL
```

Residue number must be I4, and ALL should be upper case. If you use the 'Create/Edit TLS File' task in the Refinement module of the GUI, then it will do this for you.

The last remark, however, is disputed by another user:

Nope, if you enter lowercase in the GUI you get lower case in the output!

TLS refinement and structure deposition

(July 2002)

My question has two parts:

1. *I am refining a structure that has 222 residues. The current R factore is 25% and the Rfree is 31%. When I divided my structure into three fragments (according to the gaps present) and enabled the TLS refinement the Rfactor as well as the Rfree dropped 5%. I did not observe a particular improvement of my maps afterwards and the shifts in the coordinates are minor. Is this reasonable?*
2. *If this is acceptable, what is the common praxis for including the TLS parameters when depositing structures with to the PDB?*

Ad 1: Yes. If TLS simply models the smearing of your electron density, then the R factors will improve while the density remains the same. If this modelling improves the determination of the average coordinates, and thus corrects model errors, then the maps can improve. Obviously, this is problem-dependent. Ad 2:The TLS parameters are included in the PDB header and also in the data harvesting file. The PDB should be able to process either, though it's probably safest to alert them to the presence of TLS.

TLS refinement - 'actual' and 'residual' Bfactors

I have used TLS refinement in Refmac .... as far as I remember the B-factors that we get after the refinement are 'residual' b-factors and not the 'actual' b-factors...these 'residual' b-factors are much less than those expected from the Wilson plot..... what are the 'actual' b-factors ....and how do we calculate them?

Put TLSOUT and XYZOUT from Refmac through TLSANL. That will give you derived aniso U parameters and a choice of B - TLS contribution, residual or the sum. See ccp4i task "Analyse TSL parameters" Refmac gives you your model parameters - full B is a derived quantity.

TLS refinement: Fix B-factors?

(October 2002)

I am refining a structure with two protomers in the asymmetric unit; resolution range 25-2.0Å; hexagonal space group P6122; total number of reflections: 22 000; total number of atoms to refine: ~2800. I have been using TLS refinement in the last stage of refinement and got a nice drop in Rfree and Rcryst: before TLS: Rfree=19.3 Rcryst=17.3. after TLS

refinement: Rfree=18.2 Rcryst=16.6. I used two TLS groups for the two protomers in the asymmetic unit respectively, 5 cycles of TLS refinement+9 cycles of normal Refmac refinement. However I did not turn the "Fix B-factors" field on, though I entered Wilson B in the respective field. When going back and runnning the same refinement procedure as above, this time fixing B-factors to Wilson B I get: Rfree=18.5 Rcryst= 17.3. i.e little worse than above.

Since I am not very experienced in using TLS refinement my question is whether it is necessary to fix the B-factors. It makes more sense to me, but I would like to have some competent advice on that. If fixing is advisable: to what should I fix it (with this particular refinement problem).

Refmac does not fix B values. It sets initial value to whatever you give and then refines from that point. If you don't use individual atomic B value refinement then all B values will be equal not necessarily to that given by the user. If you use individual B refinement then the program will refine TLS and then residual B values. Check if you are using individual B refinement after TLS. That may make difference.
Assuming you are refining Bs after TLS, check that the refinement has converged in the second case. Since you are starting from uniform Bs, the refinement has further to go, so to speak. The value you choose for the starting point of Bs doesn't matter.
I have tried TLS-refinement both ways and after few structures trials I now always start with the B-factors fixed at Wilson plot B-value (as recommended above, but use a numerical value, NOT a keyword "Wilson B", as this doesn't work - yet) because it gives me better results. However, in case of B-factors fixed to one value for TLS, I run many more cycles of individual B-factors restrained refinement afterwards; it usually takes some 20 or 30 cycles before the refinement fully converges. I hope it will be the same for you.

Is it possible to write out Fc_solvent and Fc_TLS in REFMAC?

Hopefully I am not asking too much. Whether it is possible to write out the partial contributions from bulk solvent diffraction and/or TLS model in REFMAC? By this I mean the Fc's of the bulk solvent model and TLS model (They are added onto the atomic model for the final Fcal any way).

Garib says: At the moment it is not possible. It can be added. How should it be written out: after scaling or before scaling? Bulk solvent correction is not added to Fcalc. What you can do is to write out mask itself. If you specify in the command line MSKOUT then the program should write out mask file as CCP4 map file.

Nucleic acids and alternate conformations within a strand: Refmac5

(July 2002)

I am refining a nucleic acid, one strand of which has two consecutive nucleotides (backbone + ribose + base) that have two alternate conformers. I haven't been able to get Refmac5 to allow me to do this. If I focus on the first of the two, and break the chain so that it becomes the 3' end of a chain, rather than an internal nucleotide, I can get Refmac to refine the two conformations. However if I make two alternative 3' ends that are instead of one residue long, two residues long, Refmac no longer allows me to do this. Therefore I think I can safely conclude it is a problem with wanting to have more than one residue as an alternate conformer, rather than some other sort of syntax error. The error refmac reports in the cases where it doesn't work is:

```
ERROR: in chain SS residue: 120
         different residues have the same number
There is an error in the input coordinate file
At least one the chains has 2 residues with the same number
Check above to see error
```
Any suggestions for a workaround?

Summary from the inquirer:

What I found was that if I put in two sets of coordinates corresponding to the two conformers en bloc, (similar to the example below) Refmac only allowed me to put in one dual-conformation residue at a time. However if I instead put them pairwise for each atom, then it works. (I found it less tedious to add one residue at a time and let Refmac reorder the lines for me.)

```
ATOM    752  N   GLU A 104  0   18.756  12.225   0.940  1.00 15.86           N
ATOM    753  CA  GLU A 104  0   17.389  11.946   0.501  1.00 15.48           C
ATOM    754  C   GLU A 104  0   17.100  12.700  -0.791  1.00 16.38           C
ATOM    755  O   GLU A 104  0   17.978  13.384  -1.354  1.00 15.23           O
ATOM    756  CB  GLU A 104  0   16.485  12.310   1.683  1.00 15.10           C
ATOM    757  CG  GLU A 104  0   16.780  11.524   2.961  1.00 17.08           C
ATOM    758  CD  GLU A 104  0   17.887  12.087   3.839  1.00 19.16           C
ATOM    759  OE1 GLU A 104  0   18.146  11.459   4.918  1.00 23.11           O
ATOM    760  OE2 GLU A 104  0   18.513  13.136   3.555  1.00 18.75           O
ATOM    761  N  ALEU A 105  0   15.873  12.555  -1.280  0.50 14.44           N
ATOM    762  CA ALEU A 105  0   15.406  13.115  -2.517  0.50 15.47           C
ATOM    763  C  ALEU A 105  0   15.768  14.544  -2.859  0.50 15.35           C
ATOM    764  O  ALEU A 105  0   15.963  14.770  -4.062  0.50 17.22           O
ATOM    765  CB ALEU A 105  0   13.842  13.034  -2.542  0.50 15.44           C
ATOM    766  CG ALEU A 105  0   13.284  11.785  -3.214  0.50 14.92           C
ATOM    767  CD1ALEU A 105  0   11.782  11.945  -3.514  0.50 15.66           C
ATOM    768  CD2ALEU A 105  0   14.072  11.434  -4.470  0.50 14.39           C
ATOM    769  N  BLEU A 105  0   15.821  12.626  -1.167  0.50 14.96           N
ATOM    770  CA BLEU A 105  0   15.329  13.376  -2.325  0.50 15.94           C
ATOM    771  C  BLEU A 105  0   15.621  14.845  -2.029  0.50 15.00           C
ATOM    772  O  BLEU A 105  0   15.515  15.375  -0.907  0.50 12.33           O
ATOM    773  CB BLEU A 105  0   13.832  13.078  -2.478  0.50 17.67           C
ATOM    774  CG BLEU A 105  0   13.061  13.102  -3.782  0.50 18.74           C
ATOM    775  CD1BLEU A 105  0   13.549  12.096  -4.827  0.50 19.53           C
ATOM    776  CD2BLEU A 105  0   11.569  12.809  -3.496  0.50 18.49           C
ATOM    777  N  AGLY A 106  0   15.855  15.448  -1.884  0.50 16.13           N
ATOM    778  CA AGLY A 106  0   16.117  16.867  -2.119  0.50 15.31           C
ATOM    779  C  AGLY A 106  0   17.510  17.335  -1.714  0.50 17.06           C
ATOM    780  O  AGLY A 106  0   17.943  18.485  -1.641  0.50 15.64           O
ATOM    781  N  BGLY A 106  0   16.040  15.577  -3.068  0.50 15.29           N
ATOM    782  CA BGLY A 106  0   16.379  16.980  -2.902  0.50 15.26           C
ATOM    783  C  BGLY A 106  0   17.737  17.215  -2.246  0.50 16.69           C
ATOM    784  O  BGLY A 106  0   18.111  18.398  -2.293  0.50 15.89           O
```

REFMAC5 maximum likelihood refinement

(July 2002)

Under which circumstances would maximum likelihood target refinement in REFMAC5 produce a dramatic INCREASE in R (from 0.187 to 0.208) and Rfree (from 0.232 to 0.257) with 1.8Å data? This was not accompanied by improved geometry (reduced RMSs). Any idea where I should start looking for the underlying problem?

I am new to Refmac so I can hardly be considered and authority. But my understanding is the idea behind m.l. refinement is to do no more refinement than is justified, whereas minimizing a crystallographic residual (in the conventional manner) tends to over-refine the data. I've noticed both in CNS and in Refmac that conventional crystallographic residual

refinement always gives me lower R-factors than does m.l. refinement, all other things being equal. This does not mean that the refinement is better. The maps, and therefore the structure, will be less biased. Also I wouldn't call a 2% increase "dramatic". Differences in the methods for calculating a solvent mask, different libraries for scattering amplitudes, geometric parameters, etc. could also easily account for some if not all of the 2% difference. FWIW I am becoming convinced that Refmac works better than CNS for a final refinement, at least in my hands.

When you do ML refinement you "push" the calculated structure factor magnitudes NO LONGER to Fobs (as it is the case for LS refinement) BUT to some MODIFIED values F*. Essentially, the main "trick" behind ML is that it takes into account the missed part of your model. Instead of fitting 'Fmodel --> Fobs' you start to fit 'Fmodel + Fmissed --> Fobs' (Fmissed are estimated statistically from ML; such modification allows one also to take into account other imperfections in your current model). From this you see that your Fmodel theoretically can start to diverge from Fobs, especially if your model is imperfect , *e.g.* is incomplete enough (if Fmissed are significant). For some more detail, you can look at, for example: Lunin, Afonine & Urzhumtsev (2002) "Likelihood-based refinement. I. Irremovable model errors.". Acta Cryst., A58, 270-282.

That is hard question to answer.. Is there any indication something has "blown up"? *E.g.* there are serious clashes between rebuilt residues? You will need to scan the log file to see if this has been monitored. Or sometimes one accidently alters the resolution range or the number of reflections.. Or a different scaling algorithm could weight low resolution data differently, and give a change in overall R factor which is mostly due to changes at low resln..

Near the end of a refinement I would call a 2.5% decrease in Rfree a spectacular drop. Likewise I think it is fair to say that a 2.5% increase during refinement is dramatic and cause for worry. I just messed up a dual side-chain conformation by hand editing the PDB and Refmac was rightly unhappy. Scanning the log file often exposes such and other problems by direct warning messages, lists of bad geometry or strange behaviour of the refinement statistics. I also noted that the older Refmac5 version was often unstable when applying SCALE BULK as well as an explicit solvent model and I have seen cases where R and Rfree decrease nicely during TLS refinement but then increase; either both R & Rfree or just Rfree. In the latter case you have to think if you are giving the program too much flexibility to overrefine and may have to tighten up restraints on geometry or B-factors, or just use TLS parameters. If both R and Rfree go up significantly you hope you made an error but there have been occasions where this happened and I was rather certain of doing things right. It is true that the ML refinement target is not directly equivalent to improving the fit between Fc and Fobs, but increases by more than a few tenth of a % are worrisome. After all, when the model improves so should the fit between Fobs and Fcalc. Final advice: make sure you have the latest version of the program, check your output for clues of errors, read the manual to understand your options and if you have "an interesting case" contact Garib, he is great and so is Refmac (it did drop my R and R-free by 2.5% of an already highly refined model).

I do not know how maximum liklihood is implemented in REFMAC, but in CNS it is important that the Rfree test set be truly uncorrelated with the working set. I am not a guru, but my understanding is that (at least in CNS) the maximum liklihood target uses the Rfree test set to assign unbiased sigma A weights prior to refinement. I had a case where I missed a higher symmetry space group (used P21 instead of C2221), so the test set contained symmetry mates. The maximum liklihood target hung up during refinement, and only when I switched to the crystallographic redidual target could I get refinement back on track. I later discovered my mistake in space group, and by calculating a posteriori Rfree and test set after the fact and shaking up the structure, the maximum liklihood target behaved well. Having said this, I am concerned that your geometry did not improve. I

would have assumed that if it was a case of model biasing the refinement that the geometric component of the target would have at least "idealized" your geometry. I could be off base here in your case.

mmCIF dictionary to SMILES?

(August 2002)

Hopefully it's not a case of RTFM: is there a way to get all those incredibly handy Refmac ligand dictionaries into SMILES strings? I see a reference to a mysterious SMILES2DICT in $CHTML/intro_mon_lib.html, but it's not in $CBIN, and would go the wrong direction anyway.
Alternatively, are there any websites that serve the PDB ligands up as smiles?

I can't really answer the first part of your question, but the MSD website contains information about all the ligands present in the PDB. This includes both stereo and non-stereo SMILES. The URL was posted a while back on the bulletin board, but here it is again: http://www.ebi.ac.uk/msd-srv/chempdb/cgi-bin/cgi.pl.

Bulk solvent model in REFMAC

I have recently switched from CNS to using REFMAC for the final refinement/analysis of several structures. I have almost immediately come across two interesting cases where the Babinet Bulk Solvent model in REFMAC does some interesting things. In the case of a high resolution dimer structure (1.9Å), refined with 4 TLS domains, the freeR factor dropped so that is was lower than the working R. Why would this happen? In the second case, a moderate resolution structure (2.7Å) with five-fold NCS was refined with one TLS domain per chain. The B-factors dropped and a large fraction of them were pinned at 2, when the default Bbulk value of 200 was used. I have tried varying the Bbulk values to minimize the freeR, with the results below. Note that the test sets are the same in both CNS and REFMAC. This has raised several questions:

1. *Should the bulk solvent scale (ki) be interpreted to have a physical meaning as suggested by Glykos & Kokkinidis (Acta Cryst. (2000) D56, 1070-72), where they say that ki and Bbulk are not independent and Ki should be ~0.8 for most proteins? In this case why is ki=0.001 when the freeR is a minimum for the dimer structure?*
2. *How can the free R refine to a value lower than the working R, when its initial value is higher?*
3. *In the case of the low resolution pentamer structure with 5-fold NCS, is it reasonable to use such a low Bbulk? The ki for this case is in closer agreement with that suggested by Glykos & Kokkinidis. Certainly the refinements which result in many individual atomic B-factors being pinned at 2.00 are not correct. Note that normally REFMAC does not permit the Bbulk value to be set below 70.*

1.9A dimer:

	Bbulk	Overall	Working	Free R	Mean B	Babinet's scale
CNS	50.77		Rw=21.73	Rf=24.44	<B>=38.50	ki=0.340#
0813x60.	60	Ro=21.773	Rw=21.807	Rf=21.143	<B>=24.856	ki=0.169
0813x80.	80	Ro=21.694	Rw=21.729	Rf=21.040	<B>=21.860	ki=0.017
0813x100	100	Ro=21.695	Rw=21.731	Rf=21.028	<B>=21.461	ki=0.001
0813x120	120	Ro=21.702	Rw=21.740	Rf=20.999	<B>=21.019	ki=0.001<<
0813x150	150	Ro=21.741	Rw=21.780	Rf=21.002	<B>=20.637	ki=0.001
080515x	200	Ro=21.796	Rw=21.837	Rf=21.028	<B>=20.211	ki=0.001

2.7 A pentamer with five-fold NCS:

```
         Bbulk  Overall    Working   Free R    Mean B   Ki    #B=2
CNS       34.75            R=22.08   Rf=24.32  <B>=47.5 0.359 #CNS (ki=0.359)
080711x  200   Ro=23.322   R=23.232  Rf=25.000 <B>= 7.4 0.371 #(276)
0808125  125   Ro=23.174   R=23.086  Rf=24.795 <B>=11.3 0.465 #(25)
080616x  100   Ro=23.157   R=23.073  Rf=24.718 <B>=14.0 0.526 #(3)
080716x ~100.3 Ro=23.155   R=23.071  Rf=24.718 <B>=14.1 0.523 #(three B=2)
0808090   90   Ro=23.155   R=23.071  Rf=24.694 <B>=15.5 0.555 #(one)
0808075   75   Ro=23.161   R=23.080  Rf=24.671 <B>=18.4 0.608 ###
081316    60   Ro=23.174   R=23.094  Rf=24.651 <B>=22.5 0.670 ###
081314    50   Ro=23.187   R=23.108  Rf=24.643 <B>=25.5 0.718 #<<
081310    40   Ro=23.203   R=23.125  Rf=24.644 <B>=29.0 0.769 ###
```

B factors being pinned to 2 (the minimum allowed) is usually a sign that the TLS refinement has gone awry. Look at the TLS parameters. The size of L parameters depends on the size of the TLS group, but if you have diagonal L elements over 100 for molecule-sized groups then be suspicious ....
This can happen if the electron density is poorly defined, and/or the model is only partially built. In which case, restrict the TLS group to the well-defined sections, or refrain from TLS until the model has progressed.
The link with Bbulk is interesting and I don't have a simple rationalisation for it. But from a practical point of view, you might not need to fix Bbulk but can let it refine to an appropriate value.

Mosaicity - high and low

Problems with low mosaicity crystals

(March 2002)

*I would like to know whether anyone has the experience with low mosaicity crystals that give troublesome data sets (unreasonably high Rsym). I have seen quite a few cases and this happened again in our lab recently. The crystals diffracted to resolutions ranging from ~3Å to ~0.6Å, data collected at various beamlines in APS and CHESS with wavelength from ~1Å to ~0.62Å. In each case the diffraction pattern looked real clean with sharp and round spots. Seeing the diffraction image one would expect very good data. However, The final Rsym's are very high, starting from the lowest resolution bin (>10%). This problem remains even with reprocessing the data in P1 (Friedel pairs don't match!). These are regular protein/RNA crystals. A common character of them is that they all have low mosaicity (<0.3). Different crystals of the same sample, with larger mosaicity would give better statistics although the diffraction looked not as good (collected at the same beamline on the same trip). People at APS beamline 19 pointed out the vibration of the loop (hence the crystal) in the cold stream could be the reason for this. By aligning the pin and loop to the parallel direction of the cold stream (less vibration) this problem is reduced. Too much description of the problem, here is the question: how the vibration of the loop causes this problem? Strictly speaking, all the loops we are using are vibrating in the cold stream, and, compared to the cell constants the vibration are *large* in any case. The translational component of the vibration therefore has no effect. The fact this only happens to low mosaicity crystals suggests the angular movement of the vibration might be the culprit. I would like to see others' experience and understanding of this problem. And, giving data collected this way, is there a remedy to apply some correction?*

A summary and further clarification: First, a few people suggested to check misindexing, beam center position, twinning ... We are fairly confident these are not the problem, since those crystals are well characterized as data had been measured repeatedly before and

these possibilities were thouroughly checked, and at the time they were measureed, there were many other data sets also measured fine (unlikely beamline problem or beam centering).
A little more info for the data I can recall:

1. APS 14BMD, Q4; P422(36.2 36.2 74.1); Reso ~1Å; Rsym ~10, mosaicity .2-.3
2. APS 19-ID; SBC-2; same crystal as above; Reso <0.6AÅ Rsym ~10, mosaicity .21
3. CHESS A1(3/3/02); Q210; #1: P622(143.8 143.8 164.0); Reso ~2.5Å; Rsym ~19; mosaicity .29
4. CHESS A1(3/3/02); Q210; #2: P3121(105.9 105.9 182.1); Reso ~3Å; Rsym ~9; mosaicity .27

Note: Rsym's given are for low reso bins. And as said above, on the same trip we had data with similar crystal (different soak/mosaicity) with Rsym's ~4%. Regarding the loop vibration, some believe so while some don't. I am inclined to think this is the problem. A good illustration was given of how the vibration (angular vibration) could move the reflection in and out of the Ewald sphere. And this was backed up with an explanation of how the vibration of the loop would lead to such an angular movement. It was remarked that such a vibration would raise the apparent mosaicity (which contradicts the low mosaicity measured). This question was partly answered: the vibration is only in certain orientations thus does not have the same effect as the mosaic spread. Also, say, if the angular movement is about 10-20% of the mosaicity, then we don't really see smeared or more spots, while the effect on intensity measured might still be significant (explained below).
Others pointed out "Angular oscillation of the crystal (several Hz) would be time-averaged over the exposure and have similar effect as classic random distribution of microcrystal domains ..." My thought to this is that the angular oscillation is not homogeneous in all directions, instead it happens mainly along the cryo stream direction. Thus it will affect the spots on the Ewald sphere position perpendicular to this direction more (say 80-90% of time staying in diffraction condition), while those tangent to this direction might stay on Ewald sphere all the time! (smeared a little probably). 10%-20% of the mosaicity is already significant in this regard (easily cause 10% difference in intensity), and it will not show in diffraction pattren. And, since the augluar vibration is probably not big, only low mossaicity crystals can feel it. A few people also pointed out spindle problem, beam undulation, inefficiency in profile fitting all affect low mosaic crystal.

Data reduction for mosaic crystals

(June 2002)

I am working on a dataset where the crystal has a very high mosaicity - my estimate is ~2.0 or more. Hence, I have had to take 0.1deg frames during data collection. Denzo V. 1.97.2 does not seem to be scaling the data very well. Would anyone know if HKL2000 can do a good job or if there's any other package thats good with dealing with minislices. Also, any suggestions about how to treat and process minislices?

See http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00748.html

The discussion left some unresolved issues which were summarised thus:

If you DO have crystals with 2.0deg mosaicity:

1. *Am I alone thinking Raji did *not 'have had'* to take 0.1deg frames? Besides the good ideas offered for really processing the data, does fine phi slicing improve anything for *very mosaic* crystals?*
2. *Do people get good maps out of such datasets?*
3. *Have people solved non MR structures out of >1.5 deg mosaicity data?*
4. *What's at the end the minimal meaningful phi slice to get as function of mosaicity?*
5. *Any tricks people want to share with us for reducing mosaicity (annealing?) ?*

A few statements/remarks/issues from reactions that followed:

- http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00750.html
  - "Mosaicity is usually expressed as one number, when in fact it is anisotropic in a significant number of cases."
  - "Using microfocus beamline on the tip of a crystal circumvents the problem in two ways. The edge of a crystal must be less mosaic, at least in that case. The microfocus beamline helped by reducing the cross-fire, which gives a smaller spot cross-section."
  - "As for the ideal phi width that does not incur spatial overlaps in the x-y plane, well, here we have two schools of thought."
- http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00751.html
  - "I can see two reasons for fine slicing, one that says fine slicing helps with very low mosaicity crystals, and the other that says fine slicing is needed with high mosaicity."
- http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00752.html
  - "Once the phi-slice is smaller than the width of the rocking curve (*i.e.* all reflections are partials) there would be no improvement in signal/background which could be optimized by 3D profile fitting."
- http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00753.html
  - "summarise with a simple-minded formula:
    ```
    If (use pocket calculator)
    Mosaicity >~ arctan(Res/EMPCA)
    where:
    Res = Maximum Resolution
    EMPCA = Effective_Maximum_Primitive_Cell_Axis ;-)
    ```
    i.e. the primitive axis that will get to line up with the beam during data collection (*i.e.* in a C-lattice that's the diagonal!)
    Then the spots do physically overlap in reciprocal space and there's little or no point trying to avoid overlap by varying the data collection geometry ..."
  - "MOSFLM was not able to handle (...) cases till recently if mos > osc_angle /2 (...) ....BUT: It *DOES* work fine now (ccp4 4.2 version and higher)"
- http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00755.html
  - "There is no advantage to ever-thinner rotation increments per image. If the crystal mosaicity is less than 1 deg, then use 0.5 deg per image. Otherwise, use one-half of the crystal mosaicity. If this still results in spatial overlaps, reduce the rotation increment, but remember that for large crystal mosaicity values, reflections will be spatially inseparable no matter how small the rotation increment."
  - "As for helping Raji, I've looked at her images. The images exhibit some problems with crystal splitting or twinning and a mosaicity of well under 2

degrees, perhaps even as low as 0.3 degrees, so her idea of using 0.1 degree images seems helpful in this case."

- http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00758.html
  - "If your mosaicity + rotation width is larger than the theoretical allowed value (*i.e.* arctan(Res/EMPCA)) then you will still collect a lot of perfectly fine data. Even if lunes (...) overlap, the spots in those lunes often do not. The problem is also not global."
  - "If your apparent mosaicity is M (crystal mosaicity plus beam di/convergence) and your oscilation width is x*M then on average each reflection will accumulate (x+1)*M degrees of background. I would say an x of 0.5 to 2.0 is reasonable for standard data collection."

Twinning

Non-merohedral twinning

(April 2002)

I'm working on a dataset from a sample that is non-merohedrally twinned. I got the twinning operation and can process the reflections of the major component of the twin, but got a rather high Rint (11%). I thought that the two lattices would not overlap too much, but apparently they do. I'd like to detwin the data, and was wondering if somebody can point out programs that can handle NON-merohedral twinned data.

As far as I know there is no program that does the job for all possible cases. However, the procedure is relatively simple: The observed diffraction pattern is the overlay of two twin-related true diffraction patterns. Each observed reflection has a twin-related reflection, which can be calculated by the reciprocal twin-operator (in order to get the reciprocal twin-operator you need to transpose AND invert the real twin-matrix). If Ia and Ib are two OBSERVED twin-related reflections and I1 and I2 are their TRUE intensities, then is:

```
Ia = (1-k) * I1 + k * I2
```

and

```
Ib = k * I1 + (1-k) * I2
```

with the twin ratio k between 0 and 1.
Some easy arithmetics leads to the TRUE intensities:

```
I1 = [(1-k)Ia - k*Ib]/(1-2k)
```

and

```
I2 = [(1-k)Ib - k*Ia]/(1-2k)
```

Using these equations you can calculate the true intensities for each twin-pair of obeserved reflections. A problem is that normally, you do not know the twin ratio k. Unfortunately, k is a rather critical value for the untwinning. The best thing is to untwin the data with different values for k, and to calculate the |E^2-1|-statistics for each untwinned data-set. The best value for k gives you a value close to 0.736 in the |E^2-1|-statistics. Another problem is that while one reduces or removes the systematical errors from the twinning, one artifically creates some new systematical errors. These new errors are the larger the closer your twin ratio is to 0.5 (I think k&#lt;0.4, or k>0.6, respectively, is required). For k = 0.5 (*i.e.* a perfect twin) the whole thing does not work at all (division by zero).
Because of the serious systematical errors created by the untwinning, you should never REFINE against untwinned data. See as well T.R. Schneider et al., Acta Cryst. 2000 D56, 705-713. In this publication the authors report a case where untwinning of the data made a solution by direct methods possible.

If this is a molecular replacement problem I'd just go on with MR and structure refinement. The Rint is indeed a bit high but the discrepancies between Fobs and Fcalc are dominated by imperfections in the model not the data and I think you can get a perfectly acceptable model if data completeness and resolution are ok. This is NOT an excuse for sloppy data collection and it will be fatal during experimental MAD/MIR phasing, but if you have the data already and can't easily get better data from a non-twinned crystal then you have my blessing.
The alternative route would require that you can identify reflection overlaps between your twins. I don't know if there are programs available for general use to do this. Basically, I think you'll have to process the second lattice (in MOSFLM, just predict the spots from the first lattice and hand pick spots from the other lattice for indexing). The resulting MTZ files will have the spot position on the detector as well as the image number. You'll then have to write a small program, if none are available, to detect and reject reflections that overlap. If this is of general use I could consider adding such an option to SFTOOLS.
Either Bruker or Mar have some software that will index and process multiple lattices -- most likely what I'm thinking of is GEMINI from Bruker.
In case of a non-merohedral twin the reflections do not overlap perfectly. The only way is to untwin the data with the reduction software. And refine against the untwinned data. I can see basically two reasons, why your Rmerge is that high:

1. You included partially overlapping spots. Unfortunately there is (as far as I know) no working algorithm that integrates partially overlapping spots on the market. You have to reject them. If your completeness decreases too much you can try to get these reflections from the other domain.
2. The software is not able to determine the position of partially overlapping spots correctly. This leads to unsatisfactory refinement. Check whether or not some of your parameters are "running" away (orientation, distance, cell, ...). If yes fix them. Play with spot- and box-size.

B.T.W. such a structure will never be of the same quality as a single crystal. The best solution is to try to get single crystals. Or if you have surgeon skills try to cut your crystals.

Non-merohedral twinning - how do I determine if it is?

I need help with determining whether my crystal is non-merohedrally twinned or not. And, if my crystal is twinned, can I find a program that can detwin non-merohedrally twinned data? Here's what I find so far:

The crystals by themselves are not microscopically twinned. However, the diffraction pattern has some reflections adjacent to some main reflections. These reflections are NOT being picked up by the indexing routine. I am able to index, scale my data fine. The spacegroup is p212121 and the scaled data is very consistent with p212121. I'd like to point out that my a and b axis are almost close to one another (105 109). The R-merges are as low as I have seen with other related data sets and the completeness is excellent. I performed merohedral twinning tests and the crystals are untwinned according to the Yeates test.

I performed Molecular Replacement on my data and got good solutions for cross-rotation and translation searches. My maps don't look abnormal at all (although I am not experienced at all to determine what one can find out about twinning from a map). Is it possible to have completely non-overlapping twins and if so, can one ignore the 2nd lattice? Or, do I have some kind of crystal splitting?? Is there a way to distinguish between

splits and twins? Since, so far I am unable to decide what's going on, I'd like some suggestions as to how I can either establish or rule out non-merohedral twinning...

Is this synchrotron data or from a laboratory source? If the latter, what means of monochromatization? Are the extra reflections always just closer to the beam center than the regular reflections?
-- I am wondering if you could be seeing the Cu K beta reflections. Some people do not realize that focusing mirrors are not adequate for monochramatization, and still require Nickel filters. Multi-layer optics do a decent job of both focusing and monochromatization. This would not be relevant for synchrotron data. If this is either Cu K beta or non-merohedral twinning (satellite xtal? less likely if this happens for all xtals) I would say if the statistics are good the data are probably usable. The processing programs know where to find the primary reflections and do not look elsewhere (after indexing). The only problem would be if two spots coincide, and if the stats are OK this must not be a serious problem.
A number of possibilities:

1. Extra wavelengths (as suggested above) give extra spots in the radial direction (Bragg's law!). I find it unlikely to be K beta, the angular separation between the spots would increase dramatically for higher orders.
2. Two X ray sources! Each would subtend a different angle at the crystal and give diffraction spots at a different angle - again not likely but the angular separation would not increase for higher orders.
3. Two different unit cells co-existing. This is a strong possibility, especially for frozen crystals. The angular separation between related spots would increase with diffraction order.
4. A little partner crystal lined up slightly differently to the first. This would index on the same lattice with a different orientation.

In some cases, the processing software would treat close together reflections as one reflection anyway. It sounds that you don't really have a problem, though it would be nice to understand which of the above (if any) is the case and how the processing software treated it.

Perfect twinning

(July 2002)

I have a perfectly twinned dataset. At first, I determined that the spacegroup was P43212 and could find MR solution using Molrep. There were two molecules in the ASU, and the initial R-factor was about 0.5. After a cycle of rebuilding and refinement the R-factor dropped to 0.36. The density map calculated from the new model was good, and some omitted parts of the initial model also have clear densities. I found there were also many other densities outside the molecules. I could dock part of an other model in them, but the remaining parts would overlap on the existing molecules. I processed the data again in P4 spacegroup and found the data was perfectly twinned with a twinning fraction of 0.48. It is said that perfect twinned data is difficult to detwin. Does anybody have successful experience with perfect twinned data?

It is not difficult to untwin data of a perfect twin (k = 0.5), it is simply mathematically impossible. However, you should never REFINE against twinned data anyway. The detwinning is used mostly to find a solution for the phse problem, and since you already

solved that you don't need to detwin the data. You should, however, take the twinning (the twin law in your case is probably 0 1 0 1 0 0 0 0 -1) into account when you refine the data.
You can indeed solve a MR search against twinned data - you will find two overlapping molecules of course.
SHELX and CNS can both refine your models against the twinned data, and also refine the twinning factor.You will have difficulties with generating maps for rebuilding though.. There are ways to detwin the Iobs using the calculated values as a guide but I have no experience of how much bias it introduces.
From my very limited experience, I got the impression that CNS can handle partial, but not perfect merohedral twinning.To my knowledge, all structures solved using perfect twins so far have been refined with SHELXL.
Yes, CNS does have a routine to detwin perfect hemihedral twins (detwin_perfect.inp). It detwins data based on a twinning operator, twinning fraction and model amplitudes. I don't quite understand how that works though but I think it works better when you have a starting model.
From my experience with replacement of perfectly twinned crystals (so far two proteins), it is possible to get a correct replacement-solution by following the stategy described in "THE twin-paper":
First do a rotation-search in the apparent space group, then do the translation search in the real space group (can be conveniently done with molrep). The advantage is that you won't get overlapping molecules as you already decide for only one of the twins in your rotation search. OK. So far it sounds promising. R-values decreased to slightly over 30 - and stuck (refinement done with CNS-twin-refine; that means you apply the twin-law to your F-calc and compare these to your F-obs. You should fix your twin ratio for this even with not perfectly twinned data, otherwise - for not really clear reasons - you can refine everything!). The problem starts now: even though additional density shows up omit-maps really look discouraging *e.g.* you omit conserved parts from your model, where it shows good density and it won't show up in the newly calculated map... I calculated maps with the CNS-script. To calculate maps for a perfect twin, you MUST have a model.

Best aproximation is:with TWOP being the twin-operator in the "hkl-form" and Icalc('hkl') is the calulated twinned intensity (the 0.5 are not really important...) as

```
Iobs('hkl') = 0.5 I(hkl) + 0.5 I(TWOPhkl)

Icalc('hkl') = 0.5 Icalc(hkl) + 0.5 Icalc(TWOP[hkl])
```

so

```
FRAC(hkl) = Icalc(hkl)/Icalc('hkl')
```

that means if you apply FRAC(hkl) to your Iobs('hkl') you should get the best approximation with regard to your Iobs(hkl)..... (BEWARE: FRAC has NOTHING to do with your twin-fraction! It is the fraction of the 'real' reflection hkl contributing to the 'twinned' reflection 'hkl'; it is a number different for EVERY reflection you have!). However as this fraction is to be computed for each reflection and is solely dependent on your actual model you end up with nearly no new information.
another try was - as I had several molecules asymmetric - a 'cyclic' detwinning. This should reduce model-bias. The idea was:

1. compute Fcalc and phicalc with model in the real space-group
2. determine NCS-ops
3. determine monomer masks
4. take Fcalc
5. square them
6. apply twin-law
7. determine fraction for each reflection

```
8. calculate your theoretical I(hkl) from I('hkl')
9. truncate (will give detwinned "Fobs")
10.      phase with phicalc from step 4. (for a "2fofc" you use fc from
   step 4 but in this case a "Fobs"-map should work better)
11.      average density with NCS-op and mask from steps 2 and 3.
12.      back transform density
13.      goto 4
```

Calulations can be done with SFTOOLS... averaging can be performed with MAIN, as you can conveniently create masks and operators and transform forwards and backwards.. and you see what is happening.
While the idea was, that the estimate of the fraction of I(hkl) and I(TWOP[hkl]) gradually improves and 'forgets' about the model, I ended up with somthing that could be displayed on the screen but had nothing to to with an electron density of a protein ... I didn't try on, however for the desperate it would be worth another try....
But all in all I think that efforts in recrystallizing and/or recloning (other constructs, a little shorter/longer, fusion-proteins etc.) are worth more energy than trying to refine a perfect twin.... (by the way: I was told this also and didn't believe it first.... ;-))
But probably the success is strongly dependent on the space-group, data-quality and 'amount' of data (*i.e.* resolution....)

Please teach about twinning

In connection with the twinning discussion, I am open to admit that many of us are in the situation:
Everything you wanted to know about twinning but afraid to ask... Can some experienced and kind soul (including the program authors) post a mini tutorial, with protocols, pointers and computer program steps, for:

1. *Symptoms of twinning (and details of the buzz words like twin factor, merohedral and hemihedral twinning)*
2. *Data collection, reduction and correction*
3. *Structure determination and refinement*

Obviously, there is a fear if you realize that a crystal is twinned and immediately we just throw the crystal away or pretend to do something, which leads to nowhere.
These tutorials are available already. See *e.g.* Twin-refinement with SHELXL.
There is something in CCP4 - it is in your local documentation: $CHTML, GENERAL --> twinning (or on the www: http://www.ccp4.ac.uk/dist/html/twinning.html). It gives information on symptoms of twinning, possible twin operators, twin factor (ratio of the volumes of the two overlapping crystals), merohedral and hemihedral twinning. There are definitions in the SHelx 97 manual. There is material on http://www.doe-mbi.ucla.edu/Services/Twinning/.

Twin problem

I am currently trying to solve a structure of a twinned crystal. I actually have 2 measured crystals, both of them appear to be hemihedrally twinned, one (A) very much (almost perfectly) and the other one (B) only a bit.

I first tried to solve the structure of crystal (A). It was possible to integrate this dataset under assumption of the space group C222 and I got an Rmerge of about 6%. However, when I tried to solve the structure by molecular replacement using AMORE and a search

model with 45% identity, I couldn't find a solution. I then used SCALEPACK to check the cumulative intensity distribution and found, that the crystal (A) was twinned. Measuring my second, less twinned crystal (B) yielded a dataset that I couldn't successfully integrate using the space group C222. Instead, this crystal appeared to belong to space group P2.

I concluded, that our crystals belong to space group P2 and that the additional axis is an artifact caused by the twinning. Since molecular replacement using the (admittedly not very good) search model again gave no result, and modifying and changing my search model didn't help either. I didn't get an AMORE solution that I could use to determine the twin-operator. Now, I have heard that it is theoretically possible to calculate the twin operator and the twin fraction from the dataset first and then to use the dataset and the twin information to solve the structure. Does anybody know, how this works?

Without knowing cell dimensions this sounds a lot like twinning that occurs when the P2 or P21 lattice approximates C222 or C2221 via a combination of a, c, beta lattice values. Typically there would be 2 (or more) molecules per asymmetric unit, with the non-crystallographic symmetry axis parallel to the twin-operation axis.
If this is correct, the pseudo-orthorhombic axis would be coincident with the monoclinic a, the pseudo-ortho c/c* axis would be along the monoclinic c* axis. I don't have the math for the desired a/c/beta relationship to hand, but if this is the case, the twin relationship is (h,k,l) -> (h,k,-h-l)
If this is the case I'd expect that your "P2" data would show pseudo-mmm symmetry that would be relatively obvious in an (*e.g.*) self-rotation function using POLARRFN (which would also confirm the direction of the twin-axes and therefore the twin-operator).
It may be possible to solve the structure using the data with the lower twin fraction using AMORE and the space group P21 or P2 even without twinning. Detwinning data with a relatively high twin fraction is prone to introducing error as the difference between the twin-related intensities approaches the noise level in the data, but there's no reason not to try it.
The article to read would be T.O. Yeates, Detecting and Overcoming Crystal Twinning (1997) Meth in Enzymology vol. 276, page. 344. There are a bunch of scripts in CNS to do what you want, and http://www.doe-mbi.ucla.edu/Services/Twinning/intro.html may come in useful.

Various

Multi-channel pipettors and 96 well trays for crystallization

(January 2002)

We have been trying to convert to 96-well plates and multichannel pipettors for crystallization. We have tried a few pipettors (both manual and automatic; Eppendorf, Genex, Biorad) with very limited success. Among other, we encounter the following problems:

- *The channels are not well aligned with the wells (particularly Eppendorf; the tips are too long).*
- *The pipettors do not accurately pipette small volumes (0.5-1 µl), despite trying many different types of tips.*
- *The blowout feature introduces bubbles in drops.*

Could people share their experience about using multi-channel pipettors and 96-well plates, in particular:

- *What drop size do people use?*
- *What brands of pipettors do people recommend?*

Any other information regarding this topic will also be very welcome.

(March 2002) - A related question:

I am looking into trying out the (supposedly) high-throughput 96 well vapor diffusion setups and was looking for some opinions. In particular I am looking at the Greiner or Corning plates sold by Hampton. I was just wondering what comments people had on their ease of use, visibility for seeing crystals, use with mulitchannel pipets or robotics setups, or what alternative sorts of 96-well setups people might be using that are readily available?

The first reaction involves 'manual slogging':

We have observed the same issues with multi-channel pipets. Our "solution" has been to use a 12 channel pipettor for dispensing the well solutions, a repeating pipet for putting protein in the mini-wells, and then manually slogging through dispensing appropriate volumes of ppt soln into each protein drop. Ultimately, we will use a robot for these functions and no longer have this concern.....

The next reply caused a little controversy:

It might not be exactly what you need but TECAN with their Genesis workstation (8 needles in // that can pipette .5 µl without problem) combined with Greiner 96 sitting drop well plate (with 3 micro wells/ well) can do a pretty good job at crystallisation.

The controversy being:

I don't agree completely. We have a Tecan robot since this summer and tested it extensively. It is true, it does a good job, but I don't think you can really get down to .5 + 0.5. The problem is mainly that the ejection is not strong enough, plastic plates are electrostatic and small drops aren't deposited on the plate but travel up the outside of the needle. 1.5 + 1.5 works fairly well. Another option in Cartesian, with selenoid valve technology: can do nano-drops, but is very expensive. Still to verify if nano-technology really works for crystallisation (kinetics might be to fast).

Then a very promising one, but not tried-and-tested by the person who posted it:

You should check out FastDrops, which is being developed with Corning. Here's Armando Villaseñor's abstract from the ACA: Fast Drops: A Speedy Approach to Setting Up Protein Crystallization Trials. Armando Villaseñor[1], Ma Sha[2], Michelle Browner[3]. [1,3]Roche Bioscience, 3401 Hillview Ave, Palo Alto, CA 94304, [2]Corning Inc. Life Sciences, 45 Nagog Prk, Acton, MA 01720. Imagine if you could set up Hampton Screens I and II against four protein complexes in 1 hour without using a robot. That's a total of 392 conditions in one hour! It is possible using the procedure and materials described in this poster. The procedure is simple, cost effective and minimized physical strain due to repetitive manipulations. This poster shows the details of the speedy manual procedure using a prototype plate described below.

If your needs demand faster crystallization set-ups, a high throughput (HT) solution is right around the corner. We have developed, in collaboration with Corning Life Sciences, a prototype 96 well crystallization plate that meets the stringent footprint standard for SBS (Society For Biomolecular Screening) microplates. Our plate will be the first HT crystallization plate on the market that is compatible with HT Automation Robotics. Crystallography will soon enjoy screening rates that are comparable to those currently available for High Throughput Drug Screening!

Pksearch

(January 2002)

Question for the code-savvy: Given a 3D-map filled with e-density or Patterson values. What is the smart way (I discovered some of the others..) to code a peak search? I would need this to update my web tutorial - any code fragments, suggestions, etc. are welcome.

The consensus is that a maximum search in a nearest neighbor 3x3x3 cube is the standard algorithm for peak search. Code examples suggested peakmax.f (CCP4), xdlmapman.f and xdlmapman_subs.f (Kleywegt cornucopia), pksrch.f (Bill Fury's PHASES). I have a feeling that with f90 array masking and maxval this could be coded quite simply - 'let you know if I get somewhere.

Archeal protein expression in *E. coli*

(January 2002)

I am trying to express an archeal protein in Ecoli. Can someone name some references that list commonly used expression and purification protocols in such cases?

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00111.html

Unusually high solvent content

(February 2002)

According to Matthews, (Matthews, B. W. (1968). J. Mol. Biol. 33, 491-497.), the Matthew's coefficient of protein crystals is usually comprehended between 1.7-3.5 Å^3/Da. Are there many proteins with a much higher solvent content than the Matthews limits? Do you have experience with protein crystals with high solvent contents? Could you point me some references to published works, or reviews about protein crystal structures with unusual VM (above 70%)?

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00205.html

Poly A to Poly S

(March 2002)

I have a polyalanine model, and would like to convert it to a poly serine model. MOLEMAN doesn't seem to be able to do this. Anyone know of an easy way to do this?

The following programmes have been suggested:

- SOD, or write a little jiffy that generates the appropriate mutate_replace commands for O for all your residues
- SEAMAN
- CNS (Generate script)

Anisotropic B-factors

(April 2002)

I am working with some data at very high resolution (1.5Å). The protein has a substrate and a cofactor (NADH). I have refined the structure both isotropically and anisotropically (in the presence and absence of the cofactor). We are interested in the puckering of the NADH ring. I would like to plot the anisotropic B-factors of the cofactor as ellipsoids. So far, I have only tried Rastep in RASTER3D but I do not really understand the way that it draws the plot. Does anybody have any suggestions on how to do so?

The following programs were suggested to me:

- ORTEP
- MapView
- XtalView
- XFIT
- Rastep in RASTER3D - using the script:
-
```
grep NAD file.pdb | rastep -auto -Bcol 5. 35. > ellipsoids.r3d
render -jpeg < ellipsoids.r3d > ellipsoids.jpeg
```

I tried all the above programs and in my point of view Rastep and Ortep give nice graphic output for the B-factors.
Then a late entry:

- povscript+

MIR-test case

(April 2002)

Before trying to reinvent the wheel, I thought I should take advantage of the vast experience that is available on this bulletin board. We wish to have a simple test case for MIR technique, for the benefit of new students in our lab. As for any good test-case scenario, we are looking for:

i. a protein that is commercially available,
ii. easily crystallizable and
iii. well established protocol for heavy atom derivatives should be available

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00428.html

A simple question of resolution

(April 2002)

My question is a simple one: How do you determine what resolution to report your structure at?

I'm not lucky enough enough to be in a situation with multiple complete MAD datasets that were solved used an ingenious program while I was drinking coffee. Instead I don't like coffee and my data is incomplete and gets quite sparse at high resolution. I'm interested in what is a proper method for reporting resolution in the worst/poor case scenarios, any references to papers on the topic would also be greatly appreciated.

This sparked a deluge of reactions - check the CCP4BB archive, starting from the original question. Also check out a discussion on the CCP4BB in 1998, threads "weak reflections" and "Reflections + Geometry". The long and short of it:

See Validation of protein crystal structures. Acta D56 (2000), 249-265. Briefly, you have a choice of reporting *nominal resolution*, *effective resolution* a la bart hazes (can be a sobering experience), and/or the *optical resolution*. Nominal resolution is something you decide subjectively. Effective resolution can be calculated using DATAMAN. Optical resolution is calculated by SFCHECK.

Docking programs

(April 2002)

Are there some free programs (free for academics) to perform docking of ligands to a protein?

- http://www.scripps.edu/pub/olson-web/doc/autodock/
- http://www.cmpharm.ucsf.edu/kuntz/dock.html
- http://www.biochem.abdn.ac.uk/hex/
- http://reco3.ams.sunysb.edu/gramm/
- http://www.ks.uiuc.edu/Development/biosoftdb/ - this site has other useful programs
- http://www.bmm.icnet.uk/docking/

Dry shipper

(May 2002)

We are in the process of acquiring a dry shipper. The 'Taylor-Wharton' CX100 seems to be widely used (although I have only seen a CP100 mentioned, but they should be similar). The 'Air Liquide' Voyageur 5 looks also good. Similar price, too. And there would be a slightly more expensive model from Messer-Griesheim. Would anybody have a comment on the alternatives?

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00895.html

Cross-platform NIS

(May 2002)

Has anyone managed to setup NIS across Linux and SG workstations? I currently have my NIS master as a RedHat Linux 7.2 machine and another as a client, which works fine. What I would like to do is include an SG machine running Irix 6.5 as an NIS client in this domain as well. I have had a go at this: by copying SG format .login and .cshrc files to a user home directory on the linux box I can login to this account as if it were on an SG

machine. I can start programs by clicking on desktop icons eg. showcase and jot, and I have priviledges to write files. However I can't open a shell ("xwsh: No such file or diectory: can't start command"). Incidentally this SG machine was formerly part of an entirely SG domain with NIS functioning normally. Any help much appreciated. If it's just not possible I would also like to know!

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00706.html

Mapstretch; also: fitting atomic models into cryoEM maps

(February and June 2002)

We have a 13-15Å resolution cryoEM map of a mutliprotein assembly. For some of the proteins from this assembly we have their atomic structures obtained by X-ray crystallography. I am looking for a program to help us in fitting of these proteins into the cryoEM map. There are two programs I have tried already: EMFIT and ESSENS. Are there any other programs that can be used for fitting?

A closely related topic:

Is there a way to "stretch" maps equally in all three dimensions using ccp4, mapman or another tool? I have an averaged map from EM (brix, ccp4 and xplor format) which is about 1.5 times too small to fit my crystallographic model into. I guess something went wrong in the calculation of the EM-magnification, but I don't know with which program that map was made. Before people attack me - I just want to make a nice picture with bobscript showing which part of the protein my partial model fits into.

Although the map stretch idea seems like 'cheating', it is, according to some, a necessary step in combining EM and crystallographic work: "We find this is essential before using EM images for molecular replacement - it is worth stretching or shrinking them by small fractions (1%) then seeing which image gives the best signal."

A list of programs, scripts and other reactions:

Fitting:

- http://www.bmsc.washington.edu/people/diller/blob/
- Some places to look:
  - September 2001 issue of Structure had a ways&means about this
  - Acta Cryst D56 October 2000 (CCP4 proceedings) had several papers about this
  - More papers about this can certainly be found if you browse through some old volumes of J.Struct.Biol.
  - Entrez-PubMed
- Try the SITUS package by Willy Wrigger at SCRIPPS. It has a bit of a learning curve (there are tutorials), but it is quite powerful.

Mapstretch:

- http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00764.html
- Write the map in an ASCII format (*e.g.*, EZD format with MAPMAN); edit this file and multiply all three unit cell lengths by 1.5; convert the map back into

CCP4 or whatever format with MAPMAN - I think this should work
Come to think of it, it might even work if you change the cell lengths with the CEll command inside MAPMAN and then mappage the map

- You can stretch a map equally in all dimensions by increasing all three unit cell parameters by the same amount. I'm not sure if this is easy to do in the map file header records but you can always backtransform to structure factors, make the changes to the MTZ header (SFTOOLS) and transform back.
- I do it by hand - if you use MAPTONA4 you get an ASCII dump of the map, and by altering the cell dimension by 1% you effectively shrink or expand the image. Then you can run MAPTONA4 again to convert the file back to give a modified map.
  Here is script of sorts:

```
maptona4 mapin hao-thelot_mlp1.map mapout hao-thelot_mlp1.ascii
```

(You need to halve the cell dimensions: equivalent to scaling by 2)

```
TITLE
Hand 1 mlp map
AXIS            Z       X       Y
GRID           50      52      56
XYZLIM          0      49       0      51       0      55
SPACEGROUP              1
MODE            2
CELL         20.125    21.260    23.240    91.700   113.300   107.700
RHOLIM      -1.33909         1.00938       -0.476994E-09    0.158896
INDXR4     0  22
END HEADER

SECTION       0
maptona4 mapout hao-thelot_mlp1_by_2.map mapin hao-thelot_mlp1.ascii
```

Side Chain Assignment of more-or-less unknown protein

(June 2002)

I am refining a structure at 1.9Å resolution, spacegroup P212121, one molecule per assymetric unit. Unfortunatelly the primary structure for this protein is unkonw (purified protein), but its tertiary structure was solved (main chain and pseudo-side chains). This "pseudo-model" is refined and the R-factor and R-free are around 19%. Initial MIRAS phases are reliable and initial map is practically continuous in all its extension. My question is: Is there any program that uses this pseudo-model and the initial MIRAS map (or an omit-map) to validate the side chain and/or to guess the most probably side-chain? I am not sure if at 1.9Å I will be able to distinguish between Leu, Asn and Asp (as an example) solely by the electron density... Cysteine, Methione residues can be identified by the anomalous signal and the peak in the 2Fo-Fc map. Other residues are easily identified (Proline, Arginine, Tyrosine ...) but what can I do distinguish between Leu/Asn/Asp, Gln/Glu, Val/Thr...? Amino acid sequence analysis is under way, but it may take a while since this is a big protein.

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00767.html

Indexing problem

(June 2002)

I have collected some data for a new structure and I have problems with my indexing. In Mosflm I get the following possible space groups:

```
  17  59     cP   137.26   137.69   153.06    90.1  90.1  90.0
P23,P213,P432,P4232,P4332,P4132
  16  58     hR   194.36   205.36   247.82    95.2  89.9 118.1   R3,R32
  15  58     hR   194.36   205.82   247.13    95.1  89.8 118.1   R3,R32
  14  57     tP   137.69   153.06   137.26    90.1  90.0  90.1
P4,P41,P42,P43,P422,P4212,P4122,P41212,P4222,P42212 P4322,P43212
  13  57     oC   205.69   206.06   137.26    90.1  90.1  83.9   C222,C2221
  12  57     mC   205.69   206.06   137.26    90.1  90.1  83.9   C2
  11  56     mC   206.06   205.69   137.26    89.9  90.1  96.1   C2
  10   3     oC   194.47   194.36   153.06    90.0  90.2  89.8   C222,C2221
   9   3     tP   137.26   137.69   153.06    90.1  90.1  90.0
P4,P41,P42,P43,P422,P4212,P4122,P41212,P4222,P42212 P4322,P43212
   8   2     mC   194.47   194.36   153.06    90.0  90.2  90.2   C2
   7   2     mC   194.47   194.36   153.06    90.0  90.2  89.8   C2
   6   1     oP   137.26   137.69   153.06    90.1  90.1  90.0
P222,P2221,P21212,P212121
   5   1     mP   137.26   137.69   153.06    90.1  90.1  90.0   P2,P21
   4   1     mP   137.26   153.06   137.69    90.1  90.0  90.1   P2,P21
   3   1     mP   137.69   137.26   153.06    90.1  90.1  90.0   P2,P21
   2   0     aP   137.26   137.69   153.06    89.9  89.9  90.0   P1
   1   0     aP   137.26   137.69   153.06    90.1  90.1  90.0   P1
```

So in Mosflm from 10 to 1 I have good fitting.
When I put those data in SCALA in all the possible groups I get the following:

	I/sigma(overall)	multiplicity(overall)
for p1 Rmerge of 5%	5.3	1.7
for p21 Rmerge of 11.7%	2.3	2.2
for c2 Rmerge of 11.7%	1.6	2.2
for p222, p2221..... 17%	1.6	4.2
for c222, ... 17%	1.6	4.2
for p4, ... 36%	0.8	4.4

The I/sigma and redundancy varies according to the space group. As you can see my a=b, which I think it causes problems with the right indexing.. In my Scala files it seems that I have p4 pseudosymmetry. In the p212121 space group I have good definition for the h. Due to that (misindexing) I have problems with my Molecular Replacement. Is there something that I might be doing wrong or something that I have not noticed? Any help or suggestions are welcomed!
Suggestions:

- Process data in P1 and then run a self rotation and self translation on the data to see what symmetry is present. Calculate Matthew's volumes.
- HKLVIEW plots of the P1 data to see symmetry.
- There are a couple of extra R-factors calculated by SCALA which include some sort of multiplicity weighting - these might be a better indicator between the choices you have. For me the clincher is the significance (I/sigI), if you merge reflections which are truly equivalent then sigI ought to be reduced.
- Rmerge in the lowest bin is a useful indicator for the spacegroup, because I tend to be liberal in what I include as "data", said he, donning a bulletproof vest. But if possible try to get more redundant data in P1, so that the scaling comparison makes more sense.

The inquirer's answer to these was: "Even though, I tried all the above, nothing really clear came out. It seems that my space group is P1 at the moment. The problem that I will have to phase now is to locate all the dimmers in the a.u. (expect 12-16 from Matthews coef.!)." This raised one final remark: It sounds scary but they will probably obey pseudo crystallographic packing, and you can analyse the interesting differences!

Molecular Replacement woes!

(June 2002)

I have synchrotron data from a nucleosome crystal which was indexed and scaled without any major problems in the space group p222(1). I get an overall completion of 96.7%. I run into problems at the molecular replacement stage (using CNS). The spacegroup of the model is p2(1)2(1)2(1). When I do a cross rotational search, going by the rotation function peak height values it seems to have worked (highest values are .1591 and .1496 and the rest of the values are in the .06 range). This is on using "fastdirect". I haven't tried "direct". When I do a translational search I do get a solution with a correlation coefficient (E2E2) which is significantly higher than the other solutions (monitor no. is .262) but on minimization refinement I get unreasonable R and free_R values. My final R is 0.5163 and the final free_R is 0.5231. The model and my molecule supposedly have a high degree of homology.
My questions are

1. *Should I expect problems in molecular relacement on account of the model and my molecule being in different space groups (I've been told that I shouldnt)? If so, what would be the nature of the problem?*
2. *What could be the possible reasons for the high R and free_R values?*

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00886.html

One addendum to this: If you use AMoRe from CCP4i it is rather easy to test whether it is P 2i 2i 2i for all 8 posibilities: P222, P21 2 2, P21 21 2, P21 21 21, P 2 21 2, etc etc.. The rotation solutions are the same for each, and then you run the translation search in all possibilities and see if one gives a significantly better result. The difficulty is that all spacegroups will be "correct" for the reflections where h & k & l are even, and others correct for subsets of data with h or k or l even.

Structure-based sequence alignment

(July 2002)

Recently, I want to prepare a structure-based-sequence-alignment figure, but I don't know how to do it. In addition, how should I deal with those residues that are of similar physicochemical properties but are not of equal positions.

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00914.html
N.B.: One of the links in the summary (http://www.prosci.uci.edu/Articles/Vol9/issue11/0235/0235.html) is out of date. Volume 9 issue 11 can now be found at: http://www.proteinscience.org/content/vol9/issue11/

Examples of pH affecting ligand conformation

(July 2002)

Does anybody know of any published example of a case where pH affects the observed conformation of a cocrystallised ligand? I.e., the same ligand in complex with the same protein assumes different conformations at different pH levels.

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00887.html

Selenomethionine prep

(July 2002)

Would anybody have a reference to production of seleno-methionine variants IN A FERMENTER?
We work rather succesfully in shaker flasks with the recipe (i.e. a variation) of Van Duyne et al., but in the fermenter, we get ODs of 40 or more, and then, I anticipated (and found), things would look a bit different. Any hint would greately contribute to minimizing the amount of selenium contaminated broth etc. I have to dispose of while trying to optimize the procedure. Any ideas about periplasmic expression (where I will likely end up oxidizing all my seleno methionine?) would also be appreciated. It does work with the regular procedure, but there's always room for improvement, I guess.

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00894.html

PC crystallography - notably 'portable crystallography'

I noticed that PC crystallography has been discussed quite a lot on this mailing list from processors to compilers. But I have not seen much practical experiences on doing so (using CCP4, CNS, O etc.) on a today's top-performance laptop. If you have done so, I would greatly appreciate your input on the practicality of crystallography on-the-go (against lab servers).

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg00899.html

CNS composite SA omit

(July 2002)

I just made a composite simulated annealling omit map in CNS, and it looks almost completely indistinguishable from the sigma-a weighted 2Fo-Fc. Unfortunately, it is clearly not because we've built a perfect model. So ... is this a common result? Is it that sigma-a does such a wondrous job of weighting the normal 2Fo-Fc, or is something running amok with this composite SA omit map script?

Rfree is just under 30, the data extend to 2.2Å in the best direction (2.5 or so in the worst), I used cartesian dynamics and a starting T of 1000, and v1.1 of cns_solve.

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg01030.html

A few late entries:

Another option that you might wish to try is:
http://www.edencrystallography.com/index.html
You can run this in "correction" mode to look for model errors and there is also an option for randomly perturbing the map (and then re-refining it) as a further mechanism for minimizing model bias. It also works quite well when large chunks of the model are missing, when run in "completion" mode. It does not do structural refinement per se, but rather real-space improvement of the maps. As a consequence, there are no difference maps -- you just have to contour lower to see density for the missing (or wrong) parts of the model.
Tom Allen with Garib's help used a different trick which works slightly better. This uses REFMAC5, and does a reciporal space omit calculation. 20 different free R-sets, each time starting from random perturbation of the structure. Works well if you dont have NCS and to resolutions worse than about 2.2A. If you have good high resolution data, these look like the refmac maps themselves (which to me is not surprising). it takes very little time to run compared to the composite omit map of CNS too.
Ask twallen@blue.weeg.uiowa.edu and he will e-mail the script for you - whatever result you get we would like to know what happened.

A SAD case

(July 2002)

Perhaps some of you have come across a situation similar to this: I have a fairly complete selenium MAD dataset (10 Se atoms in total), reduced and scaled with excellent statistics in P212121, with a decent anomalous signal. Shake-and-bake finds all 10 peaks, with I/sigma(I) between 15.0-5.0, using the peak wavelength as a SAS dataset. When I refine those peaks with MLPHARE against the entire MAD dataset, the statistics look very good: phasing power of 1.1 and R-Cullis of 0.85 across all resolution bins (for the refinement I am using the strategy outlined by Ian Tickle's tutorial). Only the real occupancies tend to a low value, 0.1, while the anomalous occupancies remain at 0.85. Density modification (solvent content of 50%) seems to improve the matter bringing the figure of merit from 0.36 to 0.55 (both RESOLVE and DM). Even better, the map is fairly interpretable, and RESOLVE can trace almost 500f the protein as a polyalanine model. But, when I inspect the fom-weighted maps, all of the Se's are in the solvent!

So, the question is: How is it possible this lack of density around the peaks if the figure of merit is 0.55, the phasing power 1.0-1.1 and the R-Cullis 0.85? Have I got to discard the Se positions calculated by Shake-and-Bake? Where could the error be? (As for the scripts I use to produce maps, they are pretty simple and had always worked, so it cannot be just that.)

I leave a brief summary of all replies with a few comments, it seems that I have not yet worked out why my selenium atoms are in the middle of the solvent. I have tried almost all suggestions but regrettably with no success, and I would like to try changing the origin of the output pdb file, could perhaps someone explain how to do it?

1. It was suggested to try both hands, or to check that RESOLVE did use the same hand that I represented, since an interpretable map is the best indicator of a successful phasing.
2. Another suggstion is doing an anomalous difference Fourier phased with the DM phases (assuming the DM phases were reasonable, as they looked so), and perhaps also peak-picking the RESOLVE map to obtain the right peaks.

Oddly enough, peak-picking produced peaks mostly on top of the main chain's density, while the anomalous difference Fourier produced peaks onto the SnB calculated peaks, which I find most confusing.

3. It was mentioned that SnB versions prior to 2.2 use an orthogonalization code different from the standard CCP4 codes, but that fractional coordinates should be OK. I am using the latest version, so this should not be a problem, I hope. I checked it nonetheless with COORCONV, and it seems SnB 2.2 produced the right pdb file.
4. A few people hinted at the possibility that the origin of the substructure could be shifted with respect to the map's origin. Could someone explain how to produce all the origin-shifted pdb's? The space group is P212121.

A reply to this last query:

In this space group, you have 8 possible origins:

```
0.  0.  0.
0.  0.  0.5
0.  0.5 0.
0.5 0.  0.
0.  0.5 0.5
0.5 0.  0.5
0.5 0.5 0.
0.5 0.5 0.5
```

After shifting your coordinates according to these values (PDBSET, keyword SYMGEN), you might need to bring your atoms back in the unit cell corresponding to your electron density map. Your graphic program should be able to do display neighbouring unit cells, or you can translate the coordinates by +1 or -1 in the appropriate direction (PDBSET again).

Then a few remarks about point 2:

Not so surprising - the RESOLVE map wants to show the protein, and will have flattened the SE sites..
It is good that the DANO map reproduces your Se sites.. Are you displaying the map with all symmetry generated? The simplest explanation is that your Se sites have symmetry equivalent positions within the RESOLVE density and you are not displaying all the equivalent sets..

PEG 550 MME as cryoprotectant

(August 2002)

Has someone used PEG 550 MME successfully as cryo-protectant? If so, at what concentrations?

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg01137.html

Riding hydrogens

(September 2002)

I am in the process of refining a 2.1Å structure using the latest version of Refmac. The default for restrained refinement seems to be to generate all hydrogens. This doesn't seem very resonable at 2Å resolution. I usually change the default to not include the hydrogens,

but I was a bit sleepy one afternoon and forgot to change the default. Upon noticing that I had set it wrong, I decided to re-run refinement again. Surprisingly, I found that adding the hydrogens in lowered both R and R free by about 1%. Still skeptical, I tried a 2.3Å data set, in and early stage of model building, here are the results:

without hydrogens:

```
REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   CROSS-VALIDATION METHOD          : THROUGHOUT
REMARK   3   FREE R VALUE TEST SET SELECTION  : RANDOM
REMARK   3   R VALUE     (WORKING + TEST SET) : 0.23753
REMARK   3   R VALUE           (WORKING SET) :  0.23305
REMARK   3   FREE R VALUE                    :  0.32008
REMARK   3   FREE R VALUE TEST SET SIZE   (%) :  5.2
REMARK   3   FREE R VALUE TEST SET COUNT      :  1196
```

with hydrogens:

```
REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   CROSS-VALIDATION METHOD          : THROUGHOUT
REMARK   3   FREE R VALUE TEST SET SELECTION  : RANDOM
REMARK   3   R VALUE     (WORKING + TEST SET) : 0.23091
REMARK   3   R VALUE           (WORKING SET) :  0.22655
REMARK   3   FREE R VALUE                    :  0.31141
REMARK   3   FREE R VALUE TEST SET SIZE   (%) :  5.2
REMARK   3   FREE R VALUE TEST SET COUNT      :  1196
```

So I am unclear on on the requirements for adding the hydrogens on in refinement. At what resolutions do they become justified? Why are they lowering my R and R-free?
http://www.ysbl.york.ac.uk/ccp4bb/2002/msg01215.html

PDB2CIF/CIF2PDB

I have kind of a funny problem: I retrieved a structure file from the Cambridge Database in CIF format - but I cannot get it into the PDB format. I tried:

- *CIFTr gives a run time error.*
- *CIF2PDB doesn't compile properly - at least in my hands.*

Are there any other conversion programs? I hardly can belive that we have such a big compatibility problem between small - and large molecule crystallography. Actually, aren't we supposed to switch to (mm)CIF anyways sooner or later ??
http://www.ysbl.york.ac.uk/ccp4bb/2002/msg01249.html
then more reactions, and Roberto's 'Babel' link got lost somewhere.

More reactions after that:

- "I am sad to say that I am not surprised by your story. However, you might be able to help yourself if you enjoy programming at least a little bit." http://www.ysbl.york.ac.uk/ccp4bb/2002/msg01254.html
- "The CIF file has a lot of information that you don't wan't. Why the dickens would you need scattering factors to write a PDB file?" http://www.ysbl.york.ac.uk/ccp4bb/2002/msg01256.html

Strange NCS/refinement problem

(September 2002)

Good data set to 2Å, indexes very well by P2 (67.947 76.813 98.441, beta=101.21) or P1 (alpha and gamma within 0.5 close to 90). Two molecules per cell. Scaled as P2, systematic absences indicate P21 very clearly.Molrep finds what appears to be a good solution readily. Problems start after that: When I restrain or constrain NCS during refinement, R free goes way up (R~30, Rfree > 40%). If I refine without NCS, R factors slip right away to 27/29 but this strange thing happens: one copy of the protein refines very well - low B factors, very good looking map, two ligands totalling >100 non-H atoms show up perfectly well on 1fofc map. A completely different story with another copy of the molecule: B factors are sky high and the map looks crappy with most of it being probably model bias. Matthews coefficient is 6.1 with just one molecule...
At this point I have tried a lot of things hoping to find an error in previous steps and nothing shows up. Can this be twinned crystal? Yates server does not appear to think so. Short of trying to find a new well-diffracting crystal form, is there a reasonale solution?

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg01251.html

AMoRe Rotation Function Scoring

(September 2002)

the Amore cross-rotation function basically calculates a correlation coefficient between the observed and calculated Patterson function (CC_P). However, the output of the cross-rotation search is for some dubious reason sorted on the correlation coefficient between calculated and observed F (CC_F). This doesn't make much sense to me for the following reasons:

1. *The search function is the CC_P, thus, from a methodological point of view, the output should be sorted on this value and not on something else.*
2. *Both the calculated F and I of the model only make sense after it has been correctly positioned, which is not the case in the cross-rotation search.*
3. *Accordingly, the signal-to-noise must be much better for CC_P than for either CC_F or CC_I.*

To illustrate this, I have run a cross-rotation search with the refined protein-only model of the A. niger phytase (Kostrewa et al., NSB, 4, 185ff, 1995) against its observed data. The top 10 of the amore cross-rotation output looks like this (I've removed the TX,TY,TZ columns for better readability):

	ITAB	ALPHA	BETA	GAMMA	CC_F	RF_F	CC_I	CC_P	Icp
SOLUTIONRC	1	3.28	85.77	237.92	27.9	55.3	42.8	26.8	1
SOLUTIONRC	1	117.85	90.00	58.64	22.2	57.2	34.3	16.3	2
SOLUTIONRC	1	90.57	80.41	235.17	18.2	58.3	26.3	5.6	3
SOLUTIONRC	1	60.20	85.07	240.47	17.9	58.5	25.9	4.9	4
SOLUTIONRC	1	22.57	57.12	223.13	17.9	58.5	26.6	4.1	5
SOLUTIONRC	1	47.85	86.10	237.71	17.8	58.5	25.8	5.2	6
SOLUTIONRC	1	87.65	60.22	71.67	17.8	58.4	25.9	4.4	7
SOLUTIONRC	1	80.37	85.82	235.99	17.7	58.4	25.1	4.5	8
SOLUTIONRC	1	44.86	24.72	48.00	17.7	58.5	26.0	5.6	9
SOLUTIONRC	1	41.18	58.25	87.29	17.7	58.4	25.7	6.4	10

Interestingly, the correct top peak appears to be also the top peak in CC_F and CC_I. However, as you can clearly see, the signal-to-noise ratio is MUCH better for CC_P. Now,

imagine that you do not have a perfect search model. In this case, I think, the chances to find the correct peak are much poorer if the output is sorted on CC_F rather than on CC_P. I don't know what you other users of CCP4 think about this, but I would strongly prefer a sorting on the real search function values rather than on something else in order to get the best chances to find the correct molecular replacement solution. Unfortunately, CCP4 Amore apparently does not give the user the choice on which values he/she wants to sort the output. Thus, the request from my side to the CCP4 developers is to give the user the choice on which values the output should be sorted, and to set the sorting on CC_P as the default, and not the sorting on CC_F.

The thread summary (http://www.ysbl.york.ac.uk/ccp4bb/2002/msg01304.html) seems to provide overriding arguments to honour the above request, but:

Ad 1) I don't really agree with the logic of this. The fast rotation function takes the form it does because this can be computed quickly. The philosophy of AMoRe is to use a fast scoring scheme to generate plausible solutions quickly, but then to re-evaluate them based on a better score. There's no overriding reason for that score to be based on Patterson overlap like the fast function.

Ad 2) This is also not true. As we showed in the paper on BRUTE, a correlation coefficient on intensities is equivalent to the correlation coefficient between the corresponding origin-removed Patterson maps. Because the origin-removed Patterson map includes the self vectors from which the orientation can be judged, the calculated I (of a single oriented model in a P1 cell) does make sense before it is correctly positioned.

Ad 3) Perhaps Jorge is reading this and will comment, but my recollection of what he's said in talks is that he has chosen the criterion on which he sorts by running tests on a wide variety of cases. In an individual problem it may not give the best results.

Now a bit of a plug. If a molecular replacement problem is difficult enough that the different criteria in AMoRe give different choices of solution, then it's probably worth running Beast, because the likelihood-based score really does seem to discriminate better. You can even just rescore the top solutions output from AMoRe or Molrep to get them resorted by likelihood score. (Note that, if you're using AMoRe you have to be a bit careful, and use the reoriented/repositioned model to which the AMoRe results refer.)

And a few other late entries:

Actually the two views are not that far apart. Though the fast rotation function is formulated on Patterson overlap, its output in Amore is in the form of correlation function (CC_P), which is equivalent to correlation of intensities theoretically. But they surely do not look equivalent in current version of Amore. The old score (CC_P) does appear to have better discrimination. Is the difference due to specific definitions used in these scores? It will be easier to judge the scores if the means and standard deviations are provided as done in XPLOR/CNS with PC search or refinement. For Amore, these parameters are listed for CC_P but not for other scores. Finally no argument about the use of Beast.

First point: the CC_F is actually the CC(F-<F>) - equivalent to CC on E**2 -1 for normalised amplitudes.

Jorge Navaza did tests, and found that all indicators detect strong signals, but that this one was more likely to detect weak signals for low homology models than the CC_P. It also seemed less vulnerable to missing and unreliable data.

In fact the most statistically valid test is that used by BEAST but it is much slower, and since we only want a set of 50 or so trial orientations from the rotation function for further analysis in the translation, the CC_F seemed the most sensitive. It really doesn't matter whether the correct orientation is 1st or 17th - providing it is present...

Diffraction images to gif/jpg

(October 2002)

Does anyone have code to convert diffraction images to graphics format (gif/jpg etc)? Maybe an addon to MOSFLM?

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg01331.html

CCP4 - Pentium 4

(October 2002)

I recall that recently there was a brief discussion of compiling/running CCP4 on Linux-based Pentium 4 systems. I apologize in advance that I don't recall what the conclusions were. Have people been encountering any problems with compilation and/or execution on P4 machines?

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg01365.html

Then a few more reactions:

http://www.ysbl.york.ac.uk/ccp4bb/2002/msg01368.html
http://www.ysbl.york.ac.uk/ccp4bb/2002/msg01380.html
http://www.ysbl.york.ac.uk/ccp4bb/2002/msg01385.html

Radiation damage

(October 2002)

I am looking for a database with radiation damages (or adducts), may be shown together with the corresponding (difference) electron density. Does there exist such a base or home page? Nothing found with GOOGLE!

The resume is that there does not exist any homepage or picture collection on the WWW concerning the radiation damages in biological macromolecules, but there are some interesting papers:

1. Burmeister, W. P. (2000). Structural changes in a cryo-cooled protein crystal owing to radiation damage. Acta Crystallogr D Biol Crystallogr 56(Pt 3), 328-41.
2. Helliwell, J. R. (1988). Protein crystal perfection and the nature of radiation damage. J. Cryst. Growth 90, 259-272.
3. Schroder Leiros, H. K., McSweeney, S. M. & Smalas, A. O. (2001). Atomic resolution structures of trypsin provide insight into structural radiation damage. Acta Crystallogr D Biol Crystallogr 57(Pt 4), 488-497.
4. Ravelli, R. B. & McSweeney, S. M. (2000). The 'fingerprint' that X-rays can leave on structures. Structure Fold Des 8(3), 315-28.
5. Weik, M., Ravelli, R. B., Kryger, G., McSweeney, S., Raves, M. L., Harel, M., Gros, P., Silman, I., Kroon, J. & Sussman, J. L. (2000). Specific chemical and structural damage to proteins produced by synchrotron radiation. Proc Natl Acad Sci U S A 97(2), 623-628.

6. Weik, M., Ravelli, R. B., Silman, I., Sussman, J. L., Gros, P. & Kroon, J.(2001). Specific protein dynamics near the solvent glass transition assayed by radiation-induced structural changes. Protein Sci 10(10), 1953-1961.
7. T.Y. Teng and K. Moffat (2000) Primary radiation damage of protein crystals by an intense synchrotron X-ray beam J. Synchrotron Rad., Vol 7, 313-317.
8. T. M. Kuzay, M. Kazmierczak and B. J. Hsieh (2001) X-ray beam/biomaterial thermal interactions in third-generation synchrotron sources Acta Cryst. D, Vol D57, 69-81.

This summary sparked one final reaction:

I saw your request for info on this subject but didn't realize you were looking for pictures. You can find movie of radiation damage I did for for these folks: Weik, M., Ravelli, R. B., Kryger, G., McSweeney, S., Raves, M. L., Harel, M., Gros, P., Silman, I., Kroon, J. & Sussman, J. L. (2000). on the http://staff.chess.cornell.edu/~richard/. It is large (nearly 1MB), lasts several minutes and has sound. You'll need Quicktime to view it. The link is called "qtest2.mov (949KB) time-resolved electron density changes from x-ray snapshots"

Structure question - how long is a 9-residue peptide?

(October 2002)

If I have a 9 residue peptide, what is the farthest distance between the C-alpha atoms of the 1st and 9th residue. I know that ca-ca is 3.8Å. But I am not sure if it is 3.8x9 or is it different - can it be longer?

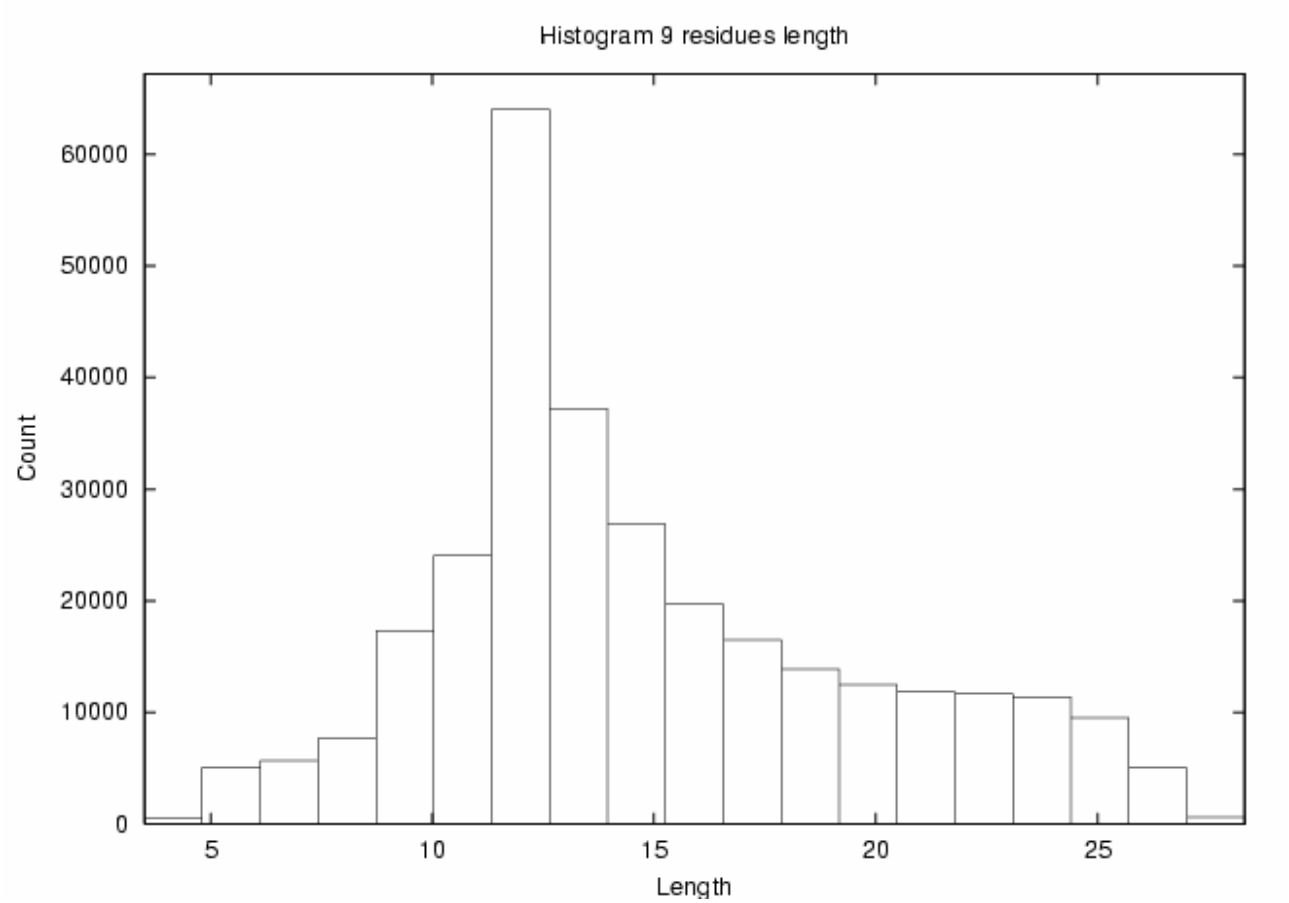
An interesting flow of reactions:

1. This sounds like putting your peptide on the rack.. no torture submissions to this web site please..
2. Here is a upper/lower/sloppy case estimate: 8 times only: *-1-*-2-*-3-*-4-*-5-*-6-*-7-*-8-*9 stars... Then: extended maximally considering no gyration around phi/psi ignoring steric restriction gives about 28Å (pulling hard, so to speak) ~ (n/2)*(3.8*cos(30)+3.8) (would look bad in Ramachandran plot with one cluster of overlying verbotens..., the so called torture case mentioned above). For helix you know 3.6 residues/turn and rise of 5.4 -> low estimate 12Å. Free gyration is around 18-22Å I wager.
3.
    - 9 residue alpha helix, approx 11.5Å from N to C-term CA
    - 9 residue beta strand, approx 26.5Å from N to C-term CA (in practice, this will about the maximum)
    - 9 residue loop - anything from 2.9Å up to the maximum (2.9 A occurs for a circular peptide with a cis-peptide bond linking nr 9 back to nr 1 ;-)

    Carugo and Pongor use ca(i)-ca(i+N) distances for fold comparison (J.Mol.Biol 315, 887-898 (2002), and J.Appl. Cryst.35, 648-649 (2002)), so they may have database-wide statistics of the distribution of ca(i)-ca(i+8).

4. I had some evil code lying around that cuts up structures in fragments so I decided to generate a little histogram for you (click on thumbnail for larger picture). The histogram was calculated using 1816 structures (one from each SCOP/ASTRAL

family, NMR structures were ignored). In total 300767 nine residue fragments were extracted.



Announcements, software releases and special places on the www

Clipper

(January 2002)

Clipper - a set of object-oriented libraries for the organisation of crystallographic data and the performance of crystallographic computation.

SHARP/autoSHARP

(January 2002)

SHARP/autoSHARP: autoSHARP: a fully automated structure solution system - from merged data to automatic model building (uses SHARP as phasing engine).

PARVATI

I am at the end of refinement of my structure. Things went great. Then now that my R-factors are 15.5 and Rfree is 17.0 (data to 1.1Å resolution) refined with anisotropic b factors, I am suddenly getting:

Problem in MAKE_U_POSITIVE -0.1387620
How do I find out the offending atom(s) - or did something else go crazy. The geometry are all well behaved. I have 8 Zn+2 in my structure and their equivalent isotropic B's are positive.....
Run the output PDB file through the PARVATI web server for analysis of the anisotropic refinement, a list of problematic atoms, and mug shots of offending residues: PARVATI - Protein Anisotropic Refinement Validation and Analysis Tool

povscript+

(February, March, April 2002)

povscript+ - a modified version of molscript and its complementary povray patch povscript.

MAPMAN server

(February 2002)

MAPMAN server - will run the program MAPMAN on your ASCII electron-density map or mask to generate an O-style map.

Gerard Kleywegt Reprint Mailer

(March 2002)

http://xray.bmc.uu.se/cgi-bin/gerard/reprint_mailer.pl

New version of PDB-mode for Xemacs/Emacs

(April 2002)

pdb-mode is a 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.

New features include the abilities to ...

- Insert new sequence
- Mutate residues
- Insert HETGROUPS directly from HICUP
- Directly submit coordinates to PRODRG server
- Interconvert fractional and orthogonal coordinates

... all within the comfort of your editor, (X)Emacs. See http://stein.bioch.dundee.ac.uk/~charlie/scripts/ for more info and downloads.

MOSFLM

(March, May 2002)

http://www.mrc-lmb.cam.ac.uk/harry/mosflm/

Uppsala Electron Density Server

(March, 2002)

http://portray.bmc.uu.se/eds/

CNS parameters

(March, 2002)

http://davapc1.bioch.dundee.ac.uk/programs/prodrg/prodrg.html

Then a recipe for 'how to obtain the topology and parameter files for acarbose for the CNS suite', combining PRODRG and HIC-Up:

1. go to http://xray.bmc.uu.se/hicup/
2. click on "Search HIC-Up"
3. enter "acarbose" in the Google search box and click the search button
4. this gives four hits (ACR, ABD, ABC, GAC); select the one that you want and click on the corresponding link
5. this gives you the HIC-Up page for your compound with loads of information and links
6. to run the PRODRG server on your acarbose compound, scroll down and hit "Run PRODRG"
7. in many cases, the PDB file you get out of PRODRG will have more sensible geometry than the one found in the PDB; save it in a file
8. select the "HIC-Up server" and upload the new coordinate file to get CNS, O, etc dictionaries calculated using the new coordinates
9. you need dictionaries for a hetero compound

Raster3D

Raster3D - a set of tools for generating high quality raster images of proteins or other molecules.

'ccp4get' CCP4 auto-installer

(June 2002)

http://www.yorvic.york.ac.uk/~cowtan/ccp4/ccp4.html

New services at the EBI-MSD

(June 2002)

** Announcement of New Services to the PDB **

- Secondary Structure Matching a tool for protein structure comparison URL : http://www.ebi.ac.uk/msd-srv/ssm Author: Eugene Krissinel <keb@ebi.ac.uk>. This is an interactive service for comparing protein

structures in 3D based on a new algorithm for common subgraph isomorphism.

- Hetgroup interface for accessing the ligands and small molecule dictionary of compounds found in the PDB URL: http://www.ebi.ac.uk/msd-srv/chempdb Author: Dimitris Dimitropoulos <dimitris@ebi.ac.uk>. An interface to an MSD reference data warehouse containing a consistent and enriched library of all the small molecules and monomers that are refered in any macromolecular structure.

** As of Wednesday Jun 19 2002 the following EBI/MSD services were moved from an SGI server to a SUN server:

http://oca.ebi.ac.uk	Integrates data query form for the PDB
http://pqs.ebi.ac.uk	Protein Quaternary Structure Query Form
http://autodep.ebi.ac.uk	PDB deposition form
http://capri.ebi.ac.uk	Home page for protein-protein docking for structure prediction
http://iims.ebi.ac.uk	Some aspects of Electron Microscopy data model (Home Page: http://www.ebi.ac.uk/msd/MSDProjects/IIMShome.html)

We would welcome feedback to msd@ebi.ac.uk on any problems user may encounter during this change over period. This weeks PDB update is the first under the new operating system.

Also note that

- the service http://relibase.ebi.ac.uk/ from 1-July-2002 will be temporarily directed to http://relibase.ccdc.cam.ac.uk/ until a linux version is re-instated at the EBI.
- the service http://www.ebi.ac.uk:80/dali/ is in the process of a port from SGI to Linux and will be maintained after 1-July-2002 by the MSD group.

ARP/wARP: version 6

(July 2002)

http://www.arp-warp.org/

HIC-Update: Release 6.1

(July 2002)

http://xray.bmc.uu.se/hicup/

Reminder X-ray generators can bite

(July 2002)

Reminder X-ray generators can bite, especially if you qualify for the http://www.improbable.com/projects/hair/hair-club-top.html.

PyMOL

(September 2002)

http://pymol.sourceforge.net/

AMoRe webpage

(Here and now)

http://www.ccp4.ac.uk/autostruct/amore/