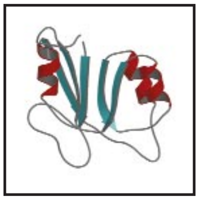
Current state of automated crystallographic data analysis

Victor S. Lamzin[1] and Anastassis Perrakis[2,3]

A goal of structural biology — and of structural genomics in particular — is to improve the underlying methodology for high-throughput determination of three-dimensional structures of biological macromolecules. Here we address issues related to the development, automation and streamlining of the process of macromolecular X-ray crystal structure solution.



Structural genomics will depend in large part on high-throughput determination of structures by X-ray crystallography[1]. There are currently several bottlenecks that must be overcome to achieve the goal of high-throughput. These include the expression, purification, and crystallization of proteins and the structure solving process. The first three are discussed elsewhere in this issue (by Edwards and colleagues, and Stevens and colleagues). Here, we focus on the last step, which includes data collection and analysis.

As soon as a well diffracting crystal of a macromolecule is available, several consecutive steps have to be taken before a reliable model can be deposited into a database (Fig. 1). Good diffraction data must be collected initially and subsequent data processing and reduction derives the intensities of the diffraction spots. To construct an electron density map, into which a macromolecular model can be built, both the amplitudes and the phases of the diffracted X-rays need to be known (see Box 1 in the article by Stevens and colleagues for a general overview of X-ray crystallography). While the amplitudes can be straightforwardly derived from the intensities, the phases have to be obtained indirectly. Solution of this so-called crystallographic phase problem can be carried out by the use of the purely computational *ab initio* methods if extremely high resolution X-ray data (1.2 Å or higher) are available[2], by molecular replacement[3] if a homologous structure is known, or by heavy atom substitution techniques.

Recent advances in molecular biology (such as expression of seleno-methionine substituted proteins) coupled with the availability of tuneable synchrotron radiation were the main driving forces for the success of the multi-wavelength anomalous dispersion (MAD) technique[4]. The ability to analyze metalloproteins by this method, or to analyze protein crystals that have been soaked in solutions containing heavy atoms or halides, have popularized the MAD and single-wavelength (SAD) measurements[5]. This technology provided an efficient route to solving the crystallographic phase problem but is crucially dependent on obtaining the highest quality experimental data. Considering the rapidly growing database of known macromolecular structures, additional developments are foreseen that would allow generalized molecular replacement, where various topological motifs and templates will be screened for potential structural homology.

When phases are available, the electron density map, which is the net result of a crystallographic experiment, can be computed (a variety of electron density modification techniques are available to improve the map quality[6]). Then, the map has to be interpreted in terms of a macromolecular model. The macromolecular model is then subjected to a refinement procedure in which the parameters of the model are optimized to best fit the experimental data and stereochemical expectations. During the refinement the electron density map improves, and this may require considerable adjustment and rebuilding of the model. The structure determination finishes with model validation and deposition of the model into the Protein Data Bank (PDB).

In an automated process all of these 'sequential' modules in structure determination should not only be tied together with a user-friendly interface, but they should also really be connected bi-directionally (Fig. 1). Information gathered from any step within the process should be fed back to the optimization of the preceding steps[7]. This 'feedback' procedure is often done manually by expert crystallographers but remains the major challenge for automation. The software for macromolecular crystal structure determination has always been explorative and was not originally developed with a goal for a high-throughput. The applications are not limited by the size of the macromolecule or the resolution of the diffraction data; they cover a wide range from small polypeptides to the large ribosomal subunit, for example. High-throughput plans however, can be hampered by the time that is required to complete a crystallographic structure determination, which can vary from hours to years.

To achieve automation, it is essential to utilize common data formats. At least two major software packages are used commonly for macromolecular crystallography, the Collaborative Computational Project Number 4, CCP4 (ref. 8), and the Crystallography and NMR system, CNS (ref. 9). Most other software can directly use or convert the data formats of these. The definitions of mmCIF (macromolecular crystallography information file), a project undertaken by the International Union of Crystallography (IUCr), should eliminate remaining format incompatibilities and allow all required experimental details to be submitted automatically upon deposition of the

[1]European Molecular Biology Laboratory, Hamburg Outstation, c/o DESY, Notkestrasse 85, 22603 Hamburg, Germany. [2]European Molecular Biology Laboratory, Grenoble Outstation, c/o ILL, B.P. 156, 6 Rue Jules Horowitz, 38043 Grenoble, France. [3]Present address: Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. Correspondence should be addressed to V. S. L. *email: Victor @embl–hamburg.de*

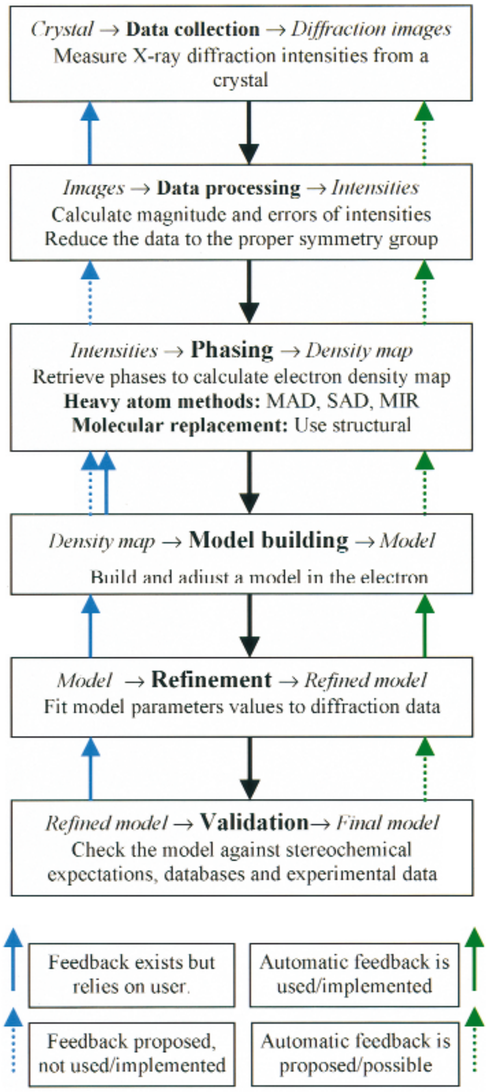

Crystal → Data collection → Diffraction images
Measure X-ray diffraction intensities from a crystal
Images → Data processing → Intensities
Calculate magnitude and errors of intensities
Reduce the data to the proper symmetry group
Intensities → Phasing → Density map
Retrieve phases to calculate electron density map
Heavy atom methods: MAD, SAD, MIR
Molecular replacement: Use structural
Density map → Model building → Model
Build and adjust a model in the electron
Model → Refinement → Refined model
Fit model parameters values to diffraction data
Refined model → Validation→ Final model
Check the model against stereochemical expectations, databases and experimental data
Feedback exists but relies on user
Automatic feedback is used/implemented
Feedback proposed, not used/implemented
Automatic feedback is proposed/possible

Fig. 1 Procedure for structure determination by X-ray crystallography.

model. Most software is free of charge for academics, a tactic that is proving priceless for development.

Data collection

Data collection is the last experimental step in the structure determination. Although the analysis of diffraction data starts after the data has been recorded, the experiment and the analysis are tightly linked together. A compromised data collection experiment will undoubtedly make all the subsequent steps of computational analysis difficult or even impossible. Data collection at record speed is now feasible on third generation synchrotron sources; with the use of charged coupled device (CCD) detectors a data set can be collected within minutes. Fears have been expressed about a likely loss of data quality (due to imperfections in instrument synchronization and increased human error in decision making under time constraints) and about the possibility that high-throughput data collection may not necessarily result in high-throughput determination of good quality macromolecular models.

Some of the key points, which have been raised at the synchrotron sites at European Molecular Biology Laboratory (EMBL) in Hamburg and within the Joint Structural Biology Group in Grenoble (JSBG) and are most relevant in analysis, are outlined here. (i) Easy access to synchrotrons and availability of sufficient beam time is vital. An increasing number of high quality beam lines dedicated to macromolecular crystallography may be able to satisfy the demand, but only if the developments continue at the present rate. (ii) Most synchrotron beamlines have fast read-out CCD detectors. Pixel array or amorphous silicon detectors are expected to speed up data collection even further in the midterm future. Software must keep pace. (iii) Cryo-crystallography[10] is now a standard technique for measurements at both in-house and synchrotron sources. Radiation damage, which occurs even for flash-frozen crystals[11], remains a concern and affects decisions for the strategy of an experiment. (iv) Data collection strategies must depend on the overall plan of data analysis. For phasing only, modest resolution data collected with minimal exposure are often sufficient. For the refinement of the model, the data should be collected to the maximal possible resolution.

Data processing

In data processing, the goal is to derive the intensities of the diffraction spots and their standard deviations from the X-ray images and reduce the data to the appropriate crystallographic space group (see Box 1 of the article by Stevens and colleagues for an overview). There are excellent, user friendly, fast and reasonably automated software suites to do this. The HKL2000 (DENZO) package[12] may outperform others in popularity, but MOSFLM[13], XDS[14], D*TREK[15], DPS[16], POW[17] and others do an excellent job and collectively have a large user community. These programs recognize a variety of detector formats. An important step in data processing will certainly be the use of the image-CIF format — the next release of MOSFLM (H. Powell and A. Leslie, pers. comm.) will take this pioneering step. The crystallographic community looks forward to detector developers and other packages adopting that standard.

Feedback from data processing to data collection

The so-called 'strategy' software is used to suggest to the user the optimal way to obtain a good quality data set for a particular purpose. Most of the data processing software has strategy options, and stand-alone programs are also popular. Normally the software makes predictions on a purely geometrical basis and using the first diffraction image only. However, it cannot foresee possible deterioration of crystal diffraction due to radiation damage, crystal anisotropy, and so forth. Internal data quality indicators, which are provided by the processing software, are typically inspected after the measurements have been completed. To proceed towards true automated feedback, the tools for 'at-the-time' data evaluation should be implemented, notably to take account of crystal deterioration.

Phasing

The SOLVE[18] program revolutionized the field by providing fully automated phasing from the diffraction intensities of a heavy atom substitution experiment. Automated versions of other software, such as auto-SHARP[19] (C. Vonrhein, pers. comm.) and CHART[20] which invoke programs from the CCP4 suite, will soon become available. Direct methods approaches, such as SnB[21] and SHELXD (G. Sheldrick, pers. comm.) are very efficient, but only provide the positions of the heavy atom sites. The CCP4 graphical user interface and CNS offer easy means for proceeding from the heavy atom sites or a molecular replacement solution to a density map, but require user intervention to a varying extent. There are excellent molecular replacement routines in all major software packages.

Feedback from phasing to data processing

No tool exists that can feed information about the quality of an electron density map back into better processing of the X-ray data. Several groups are working on theoretical aspects of this, but it is likely to be some time before practical implementations emerge.

Model building

Subsequent to the phasing step, efficient, accurate and objective modeling presents a challenge. The initial density map or a molecular replacement solution (that is, the phase data) is often of medium quality. It is not straightforward to build an accurate model, and highly tedious and time-consuming iterations with refinement and model reconstruction steps are required. QUANTA© (ref. 22) offers highly automated tools to facilitate model building but requires user intervention — although minimal in high-resolution maps. XtalView[23], Main[24], Turbo Frodo[25] and the pioneer of these, O[26], also offer good quality tools. Approaches for recognition of patterns in density maps (such as with ESSENS[27] and FFFEAR[28]) are promising but cannot deliver a model automatically. Various tools exist to build models from the Cα coordinates but it is the accurate recognition of Cα positions of the protein backbone that remains a challenge. ARP/wARP[29] offers complete automation for building protein structures at resolution around 2.3 Å or higher and will be described below.

Feedback from model building to phasing

The introduction of the statistical σ_A approach[30] is a commonly used technique that combines information from the model and experimental phases. It is our strong belief that much information from the model building step can be fed back no only to the phasing step but also to data processing step. Indeed, if the model building is difficult or even unsuccessful, the problem has most likely occurred at earlier stages.

Refinement

Refinement involves the adjustment of the parameters of the model, through the minimization of residuals between diffraction amplitudes that are calculated from the current model and those obtained from the X-ray experiment. The most commonly used programs for this process are CNS[9] and REFMAC[31] (CCP4), TNT[32] have a faithful user community, while SHELXL[33] remains favored for high-resolution refinement. The generalized maximum likelihood program, BUSTER is under development and will be particularly valuable for incomplete models (G. Bricogne, pers. comm.). Explicit definition of the supplementary stereochemical restraints is required, although REFMAC 5 (G. Murshudov, pers. comm.) is able to automatically restrain a variety of ligands and prosthetic groups from a pre-built library.

Table 1 Software

General packages	
CCP4	http://www.dl.ac.uk/CCP/CCP4/
CNS	http://cns.csb.yale.edu/v1.0/
Data processing	
D*TREK	http://www.msc.com/brochures/dTREK/
DPS	http://ultdev.chess.cornell.edu/MacCHESS/DPS/
HKL2000/DENZO	http://www.hkl-xray.com/
MOSFLM	http://www.mrc-lmb.cam.ac.uk/harry/mosflm/
XDS	not available
STRATEGY	http://www.crystal.chem.uu.nl/distr/strategy.html
PREDICT	http://biop.ox.ac.uk/www/distrib/predict.html
Phasing	
Molecular replacement	
AMORE (and in CCP4)	ftp://b3sgi3.cep.u-psud.fr/pub/
CNS	http://cns.csb.yale.edu/v1.0/
MOLREP (CCP4)	http://www.dl.ac.uk/CCP/CCP4/dist/html/molrep.html
Heavy atom sites identification	
SHELXD	http://shelx.uni-ac.gwdg.de/SHELX/
SNB	http://www.hwi.buffalo.edu/SnB/
RANTAN, RSPS (CCP4)	http://www.dl.ac.uk/CCP/CCP4/dist/html/rantan.html; http://www.dl.ac.uk/CCP/CCP4/dist/html/rsps.html
Heavy atom phasing	
CHART	http://crick.chem.gla.ac.uk/~paule/chart/
MLPHARE (CCP4)	http://www.dl.ac.uk/CCP/CCP4/dist/html/mlphare.html
PHASES	not available
SHARP	http://lagrange.mrc-lmb.cam.ac.uk/
SOLVE	http://www.solve.lanl.gov/
Model Building	
Pattern searching	
ESSENS	http://alpha2.bmc.uu.se/~gerard/manuals/gerard_manuals.html
FFFEAR	http://www.ysbl.york.ac.uk/~cowtan/fffear/fffear.html
Interactive graphics for model building	
MAIN	http://www-bmb.ijs.si/
O	http://origo.imsb.au.dk/~mok/o/
QUANTA	http://www.msi.com/life/products/quanta/index.html
TURBO-FRODO	http://afmb.cnrs-mrs.fr/TURBO_FRODO/
XTALVIEW	http://www.scripps.edu/pub/dem-web/
Automated model building	
ARP/wARP	http://www.embl-hamburg.de/ARP/
Refinement	
BUSTER	http://lagrange.mrc-lmb.cam.ac.uk/
CNS	http://cns.csb.yale.edu/v1.0/
REFMAC (CCP4)	http://www.dl.ac.uk/CCP/CCP4/dist/html/refmac.html
SHELXL	http://shelx.uni-ac.gwdg.de/SHELX/
TNT	http://www.uoxray.uoregon.edu/tnt/welcome.html
Validation	
PROCHECK (and in CCP4)	http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html
SFCHECK (CCP4)	http://www.dl.ac.uk/CCP/CCP4/dist/html/sfcheck.html
WHATCHECK	http://swift.embl-heidelberg.de/whatcheck/

Feedback from refinement to model building

The most traditional feedback (manual examination of model iterated with the refinement) was the first one to be automated. The program ARP/wARP offers complete automation in this iteration. The basic concept underlying the ARP/wARP approach is the unified view of the refinement and the iterative building of model fragments in the interpretable regions of the density map. The 'hybrid' macromolecular model (a mixture of protein fragments and free atoms) has been introduced and not only the parameters of the model but the model itself are allowed to change on the fly after every refinement cycle. Thus, the software effectively mimics human intervention.

Accounting for the limitations in resolution of the X-ray data required (2.3 Å or higher), which spans about two thirds of the Protein Data Bank content, and allowing a reasonable failure rate, we estimate that when initial phases are available, ~50% of structure solutions can proceed with automatic model building using ARP/wARP. Although this is a step towards full automation, further developments are needed to address the remaining 50% (often the most difficult ones) to deliver a final and well-validated model. This is an example of the 'bottleneck within the bottleneck' in the procedure of macromolecular structure determination.

The introduction of a changeable 'hybrid' model is a step towards more elaborate description of crystallographic structures. Whether we like it or not, the crystal of a macromolecule cannot be adequately represented as a set of connected atoms in the way that we understand a macromolecular model in the conventional sense. The nature of the crystals is much more complex: atomic motion and multiple conformations are two examples of complexity. Much effort is now put into the development of improved parameterization of macromolecules. It is hoped that within the next few years, we will witness essentially automatic techniques for model building and refinement at a resolution approaching 3 Å.

Validation

A variety of software (WHATCHECK[34], PROCHECK[35], SFCHECK[36]) and easy to use web servers are available for assessment of the model quality. The Protein Data Bank (http://www.rcsb.org/pdb/) provides submission tools for model validation and is an outstanding example of a common format for model coordinates. A number of issues remain to be resolved, especially regarding the validation targets[37,38].

Feedback from validation to refinement

This is common practice for an expert crystallographer. Going back to further refine and adjust the model may be required if suggested by the validation software. Feedback of this process is not yet automatic but should be explored as soon as possible.

Feedback across many steps

Although full feedback-based automation may not be feasible in the near future, theoretical considerations are being undertaken. Some emerging ideas (for example, that towards the end of a structure determination a more objective re-examination and better statistical treatment of initial X-ray data is needed; G. Bricogne and R. Read, pers. comm.) will hopefully lead to better and more accurate models resulting from the structure determination process.

Web resources

The html links to the current web sites of the software mentioned in this review are available in Table 1.

Associations with structural genomics

V.S.L. is Deputy Head of the EMBL Outstation in Hamburg and Group Leader in protein crystallography. A.P. was a Team Leader at the EMBL Outstation in Grenoble and from October 1st, 2000 is a Group Leader at the Netherlands Cancer Institute in Amsterdam. Both authors are collaborators in several NIH applications for structural genomic centers and members of the EU-funded AUTOSTRUCT network aiming at coordinated software development for macromolecular structure determination.

1. Report on the First International Structural Genomics Meeting. http://www.nigms.nih.gov/news/meetings/hinxton.html (2000).
2. Uson, I. & Sheldrick, G.M. *Curr. Opin. Struct. Biol.* **9**, 643–648 (1999).
3. Turkenburg, J.P. & Dodson, E.J. *Curr. Opin. Struct. Biol.* **6**, 604–610 (1996).
4. Ogata, C.M. *Nature Struct. Biol.* **5**, 638–640 (1998).
5. Dauter, Z., Dauter, M. & Rajashankar, K.R. *Acta. Crystallogr. D* **56**, 232–237 (2000).
6. Abrahams, J.P. & De Graaff, R.A. *Curr. Opin. St.ruct. Biol.* **8**, 601–605 (1998).
7. Lamzin, V.S. *et al. Acta Crystallogr. D* **in the press** (2000).
8. Collaborative Computational Project Number 4. *Acta Crystallogr. D* **50**, 760–763 (1994).
9. Brunger, A.T. *et al. Acta Crystallogr. D* **54**, 905–921 (1998).
10. Garman, E.F. & Schneider, T.R. *J. Appl. Crystallogr.* **30**, 211–237 (1997).
11. Ravelli, R.B. & McSweeney, S.M. *Structure* **8**, 315–328 (2000).
12. Otwinowski, Z. & Minor, W. *Methods Enzymol.* **276**, 307–326 (1997).
13. Leslie, A.G.W. *CCP4 Newsletter* (1992).
14. Kabsch, W. *J Appl Crystallogr.* **26**, 795–800 (1993).
15. Pflugrath, J.W. *Acta Crystallogr. D* **55**, 1718–1725 (1999).
16. Rossmann, M.G. & van Beek, C.G. *Acta Crystalogr. D* **55**, 1631–1653 (1999).
17. Bourgeois, D. *Acta Crystallogr. D* **55**, 1733–1741 (1999).
18. Terwilliger, T.C. & Berendzen, J. *Acta Crystallogr. D* **55**, 849–861 (1999).
19. Fortelle de La, E. & Bricogne, G. *Methods Enzymol.* **276**, 590–620 (1997).
20. Emsley, P. *CCP4 Newsletter* (1999).
21. Weeks, C.M. & Miller, R. *Acta Crystallogr. D* **55**, 492–500 (1999).
22. Oldfield, T.J. *Meeting of the IUCr macromolecular computing school*, http://www.iucr.org/iucr-top/comm/ccom/School96/iucr.html (1996).
23. McRee, D.E. *J. Struct. Biol.* **125**, 156–165 (1999).
24. Turk, D. PhD thesis, Technische Universitaet Muenchen (1992).
25. Roussel, A. & Cambillau, C. *Silicon Graphics geometry partners directory* (Silicon Graphics Corp.;1991).
26. Jones, T.A., Zou, J.-Y., Cowan, S.W. & Kjeldgaard, M. *Acta Crystallogr. A* **47**, 110–119 (1991).
27. Kleywegt, G.J. & Jones, T.A. *Acta Crystallogr. D* **53**, 179–185 (1997).
28. Cowtan, K. *Acta Crystallogr. D* **54**, 750–756 (1998).
29. Perrakis, A., Morris, R. & Lamzin, V.S. *Nature Struct. Biol.* **6**, 458–463 (1999).
30. Read, R.J. *Acta Crystallogr. A* **42**, 140–149 (1986).
31. Murhudov, G.N., Vagin, A.A. & Dodson, E.J. *Acta Crystallogr. D* **53**, 240–255 (1997).
32. Tronrud, D.E. *Methods Enzymol.* **277**, 306–319 (1997).
33. Sheldrick, G.M. & Schneider, T.R. *Methods Enzymol.* 277, 319–343 (, 1997).
34. Vriend, G. *J. Mol. Graph.* **8**, 52–56 (1990).
35. Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
36. Vaguine, A.A., Richelle, J. & Wodak, S.J. *Acta Crystallogr. D* **55**, 191–205 (1999).
37. Dodson, E.J., Davies, G.J., Lamzin, V.S., Murshudov, G.N. & Wilson, K.S. *Structure* **6**, 685–690 (1998).
38. Wilson, K.S. *et al. J. Mol. Biol.* **276**, 417–436 (1998).