Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Automation of the collection and processing of X-ray diffraction data – a generic approach

A. G. W. Leslie,[a]* H. R. Powell,[a] G. Winter,[a] O. Svensson,[b] D. Spruce,[b] S. McSweeney,[b] D. Love,[c] S. Kinder,[c] E. Duke[c] and C. Nave[c]

[a]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England, [b]ESRF, BP220, 38043 Grenoble, France, and [c]CLRC Daresbury Laboratory, Daresbury, Warrington WA4 4AD, England

Correspondence e-mail: andrew@mrc-lmb.cam.ac.uk

Received 3 May 2002
Accepted 19 September 2002

With modern detectors and synchrotron sources, it is now routine to collect complete data sets in 10–30 min. To make the most efficient use of these resources, it is desirable to automate the collection and processing of the diffraction data, ideally to a level at which multiple data sets can be acquired without any intervention. A scheme is described to allow fully automated data collection and processing. The design is modular, so that it can easily be interfaced with different beamline-control programs and different data-processing programs. An expert system provides a communication path between the data-processing software and the beamline-control software and takes decisions about the data collection based on project information provided by the user and experimental data provided by the data-processing program.

1. Introduction

The advent of very high brilliance beamlines at third-generation synchrotron sources and advances in detector technology have resulted in a dramatic increase in the speed of macromolecular diffraction data collection (Cassetta *et al.*, 1999; Mitchell *et al.*, 1999; Abola *et al.*, 2000; Hendrickson, 2000). Data sets can now be collected in tens of minutes rather than several hours (Walsh *et al.*, 1999). As a result, the time involved to set up the experiment (sample mounting, crystal centring, determination of data-collection parameters) has become a significant proportion of the total time. To maximize the efficiency of use of the beamline, it is therefore desirable to automate these steps as far as possible. Equally, automation of the processing of the data allows the user to monitor the progress of the experiment more readily and to collect additional data or abort the experiment according to the circumstances. Structural genomics programs involving high-throughput structure determination would obviously benefit from automation. Quite apart from the realms of high-throughput crystallography, many of the more challenging structural problems require screening of a large number of crystals in order to find one giving suitable diffraction. The ability to automatically screen and rank order a large number of crystals and then collect data from the best would greatly reduce the burden on the experimentalists. While the benefits of automation will be greatest at synchrotron beamlines, there will be many instances in which automation would be valuable when using laboratory sources, either for data collection or for pre-screening crystals prior to visiting a synchrotron.

The issue of automation is being addressed at many synchrotrons and script-based procedures for automatic data processing have already been developed (Ferrer, 2001; Holton, 2002; Roth *et al.*, 2002). However, these procedures have no control over how the data are collected. A fully

automated system requires communication with the software controlling the beamline (including the detector) as well as the data-processing programs. Because most synchrotrons have made a significant investment in developing their own software for beamline control and because there are several data-processing packages available, there is an advantage in developing a modular system that can readily be integrated with different existing software packages.

2. The expert system

When collecting data at a synchrotron, it is quite common to use two or even three different computers in order to control the beamline and process the diffraction images. The scientist will have to make decisions about the parameters of the experiment (*e.g.* exposure time, rotation range, oscillation angle, detector distance, beam size, wavelength) based on their experience, the visual appearance of the images and information provided by the data-processing programs. The goal of an automated system is to replace the scientist with intelligent software, which will be referred to as the 'expert system' (Fig. 1). The role of the expert system is to issue commands to the beamline-control software (and sample-loading software) to collect the initial image(s) necessary to characterize the sample. It will then instruct the data-processing software to process these images (autoindex and integrate). On the basis of the resulting information and project information provided by the user (stored in the database), the expert system will then take a decision on whether the sample is suitable for data collection. If data is to be collected, it will use information obtained from processing the initial images to determine the data-collection parameters and instruct the beamline-control software to collect the appropriate images. Finally, it will instruct the processing software to integrate the data and actively monitor the data quality, checking for excessive radiation damage or other experimental problems.

2.1. Traditional data collection

To assess the feasibility of automation, it is useful to consider the steps taken by an experienced user in setting up a data-collection experiment. Firstly, the detector distance is set to provide the desired resolution limit and two diffraction images are then collected with a conservative exposure time and an oscillation angle determined by the unit-cell size (if known) or a small rotation (*e.g.* 0.25–0.5°). The crystal is rotated by 90° between the two images to provide two orthogonal samples of the reciprocal lattice. These initial images are then examined carefully to determine the effective resolution limit, to check for the presence of disorder, twinning or multiple crystals and to obtain an approximate estimate of the mosaic spread. If the quality of diffraction is acceptable, the images are autoindexed to determine the cell dimensions and possible space groups. The indexing is checked by comparing the predicted diffraction patterns with the images. This provides another opportunity for detecting the presence of a second (weaker) lattice. Assuming successful indexing, the mosaic spread can be estimated more accurately. Finally, the data-collection strategy [total rotation angle and oscillation angle(s)] is worked out and an exposure time is chosen. If necessary, the detector distance is then reset to provide the required resolution limit and data collection is started. Ideally, the data is processed during collection, although this can be challenging on very intense beamlines with typical exposure times of a few seconds.

2.2. The use of autoindexing to assess crystal quality

The most difficult step to automate in the procedure outlined in the previous section is also the most critical one, namely the visual assessment of the quality of the diffraction pattern. While a trained crystallographer can instantly recognize the presence of disorder, multiple crystals or excessive mosaicity, to try to emulate this with image-recognition software would not be trivial. A less sophisticated but more practical approach is to use the success or failure of the autoindexing as the initial test of crystal quality. Autoindexing algorithms have become very robust, particularly those based on Fourier methods (Steller *et al.*, 1997; Otwinowski & Minor, 1997). The most common causes of failure are now the following.

(i) Incorrect physical parameters (direct-beam position, detector distance, wavelength).

(ii) Insufficient reflections.

(iii) Multiple lattices or split spots.

(iv) Excessive mosaic spread, leading to lune overlap.

(v) Algorithm failure (rare).

As the physical parameters will be obtained directly from the beamline software, they should not be in error and therefore all the common sources of failure can be attributed to poor crystal quality (assuming that insufficient reflections is the result of very weak diffraction). With the sole exception of algorithm failure, any failure in the autoindexing can therefore be interpreted as an indication that the crystal is not suitable for data collection. The two images can be indexed separately and in combination, with the requirement that indexing is successful in each case. This should help to identify those cases

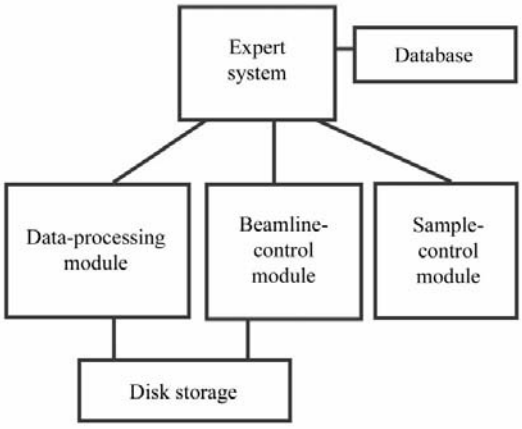

Figure 1
A schematic outline of the proposed automation scheme.

where disorder or multiple lattices are only apparent in some sections of reciprocal space.

The success of the autoindexing can be judged according the following criteria.

(i) A solution must be found.

(ii) The r.m.s. difference between observed and predicted spot positions must be less than a preset value.

(iii) The number of spots rejected from either the indexing or the refinement must be less than a preset value.

Autoindexing results in a number of possible solutions for the crystal lattice type, each of which will have a penalty associated with it which is a measure of how well the experimentally determined unit cell conforms to the restrictions imposed by the symmetry of the solution. In many cases, there will be a cluster of solutions with low penalties, followed by other solutions with much higher penalties. Typically, the solution with the highest symmetry from the cluster of low-penalty solutions will be chosen. In the absence of measured intensities, it is not possible at this stage to distinguish between Laue groups 3, 3/*m*, 6, 6/*mmm* or between 4 and 4/*mmm*. A conservative approach would be to adopt the lower symmetry solution when working out the data-collection strategy. Alternatively, the higher symmetry could be assumed and the symmetry tested as soon as a sufficient number of images had been collected and processed. If the symmetry was lower than anticipated, the data-collection strategy would be re-evaluated using the correct symmetry and taking account of the data that had already been collected. The same procedure would be followed for cases of pseudo-symmetry; for example, a monoclinic cell with a β angle close to 90°, which could have been identified as orthorhombic.

2.3. Automatic mosaicity estimation

An estimate of the crystal mosaicity is important for ranking different crystals and when determining the optimum oscillation angle per image for data collection. Once the crystal has been indexed, an estimate of the mosaicity can be obtained by integrating the diffraction image using a series of increasing values for the mosaic spread (typically from 0 to 1° in steps of 0.1°). The total integrated intensity for all predicted reflections is calculated in each case and the mosaicity is set to the value at which further increases in mosaic spread do not result in a significant increase in the total integrated intensity. The presence of diffuse scatter limits the accuracy of this approach, but current experience suggests that this procedure, which is implemented in *MOSFLM* (Leslie, 1992), works well in the majority of cases.

2.4. Data-collection strategy

Several software packages are available to calculate a data-collection strategy based on the crystal orientation and assumed Laue group (see Dauter, 1999, for an overview). In the case of *MOSFLM* (Leslie, 1992), allowance can also be made for any data that have already been collected. It is also possible to calculate which segments of data should be collected to provide maximum data completeness for a specified total rotation. For example, a total rotation of 60° in two 30° segments will often give 95% completeness for an orthorhombic crystal. This procedure is useful if the crystal lifetime (owing to radiation damage) is unknown, as it allows an almost complete data set to be collected with minimum exposure time and additional data can then be collected to improve multiplicity if the radiation damage is acceptable. In an automated system, more sophisticated data-collection strategies can be envisaged. For example, the first image could be recollected at regular intervals to provide a rapid assessment of the level of radiation damage.

2.5. Determination of the resolution limit and exposure time

Following estimation of the crystal mosaicity, the two images used to autoindex the crystal are integrated, giving an estimate of the mean $I/\sigma(I)$ as a function of resolution. These experimental values can be used in conjunction with an error model to estimate the mean $I/\sigma(I)$ as a function of both resolution and exposure time, as has already been achieved in the program *BEST*, which has been implemented on the beamlines at the EMBL Outstation at DESY (Popov & Bourenkov, 2001). Based on a user-supplied value for the required resolution [defined for example as the resolution at which the mean $I/\sigma(I)$ drops to less than 2 or 3], the expert system calculates the exposure time required. If this exceeds the maximum time allowed (specified by the user) then data collection is abandoned. More sophisticated procedures could take account of any observed anisotropy in the diffraction.

2.6. Integration of the data

Providing that the relevant experimental parameters (direct-beam position, wavelength, dispersion, polarization, beam divergence, detector type, detector distance *etc*) are known, then automated integration of the images is quite straightforward for most data sets (see Rossmann & Arnold, 2001, for an overview of data-processing procedures). Accurate cell parameters are first determined using post-refinement techniques and one or two small wedges of data (two are essential for orthorhombic or lower crystal symmetries). The required additional data would automatically be collected as part of the data-collection strategy. For example, if the crystal symmetry is orthorhombic, data collection would start with 1–2° of data at a φ value 90° away from the starting φ value determined by the strategy software. These images, together with the first few images collected at the true starting φ, would be used to refine the cell. The post-refinement will also provide a more accurate estimate of crystal mosaicity. Automatic determination of the peak/background mask is already incorporated into *MOSFLM* and typically no manual intervention is required.

In situations where data collection is faster than data integration on a single processor, segments of data (for example 10–20 images) can be farmed out to different processors and the results merged at the scaling step. The expert system would be responsible for overall control of the integration and statistics on the data quality and completeness would be made

available to the user in order to monitor the progress of the experiment.

3. Implementation

The expert system uses a set of high-level commands to issue requests to the beamline-control and data-processing software (Fig. 1) and expects a defined response to each command. The format of both the commands and the responses has to be rigorously defined, particularly if the system is to be truly generic. A decision was taken to use the eXtensible Markup Language (XML) as the format for both, as this is a widely used and flexible standard (see Fig. 2 for an example). The commands or responses, formatted as XML documents, are transmitted using HTTP as the data-transfer protocol, which facilitates checking for transmission errors. A 'translator' is required to convert these high-level commands to a set of (keyworded) commands which are recognized by the programs and also to pass back the required information to the expert system. However, assuming that the other programs already have the necessary functionality, which is likely to be the case, then this translator is the only new software required in addition to the expert system itself. High-level commands issued to the data-processing program currently include the following.

(i) INDEX. Autoindex one (or more) defined images and return the results to the expert system. The results will include unit cell, orientation, possible space groups, number of reflections (used and rejected from both indexing and refinement), r.m.s. positional error in predicted spot positions.

(ii) FIND-MOSAIC. Estimate the mosaicity of one (or more) diffraction images, given the unit cell and orientation. Results: mosaic spread.

(iii) STRATEGY. Work out a data-collection strategy, given the crystal cell, orientation, Laue group, spot size, mosaicity and whether anomalous data are required. Results: rotation range(s) and oscillation angle(s).

(iv) REFINE-CELL. Obtain accurate cell parameters using images in one or more segments, widely separated in φ, using post-refinement. Results: refined unit cell.

(v) INTEGRATE. Integrate a single image. Results: $\langle I/\sigma(I)\rangle$ as a function of resolution.

(vi) PROCESS. Integrate a series of images. Results: integrated reflection lists, positional residuals, $\langle I/\sigma(I)\rangle$ values.

Corresponding commands to the beamline-control software will include the following.

(i) COLLECT. Collect one or a series of images with a defined detector distance, exposure time, φ-value oscillation angle and wavelength.

(ii) ALIGN. Automatically align the beamline.

Each command will have a corresponding XML document defining all the parameters associated with that command and a second document which defines any results from that operation which need to be passed back to the expert system. The XML for the INDEX command is shown in Fig. 2 as an example.

```
<?xml version='1.0'?>
<!DOCTYPE index_request>
<index_request>
 <fileinfo>
  <template>test_1_###.img</template>
  <directory>/data/images/adsc</directory>
 </fileinfo>
 <detector>
  <type>adsc</type>
 </detector>
 <beam>
  <x>95</x><y>95</y>
 </beam>
 <image>1</image>
</index_request>

<?xml version='1.0'?>
<!DOCTYPE index_response>
<index_response>
 <status>
  <code>ok</code>
  <message></message>
 </status>
 <spot_search_response>
  <found>679</found>
  <used>646</used>
  <rejected>33</rejected>
 </spot_search_response>
 <solution>
  <symmetry>P2</symmetry>
  <number>3</number>
   <orientation>
    <a_matrix>
     <e11>0.00089776</e11>
          ...
     <e33>-0.00328923</e33>
    </a_matrix>
    <cell>
     <a>83.40</a><b>126.28</b><c>127.77</c>
     <alpha>90.00</alpha><beta>102.96</beta><gamma>90.00</gamma>
    </cell>
   </orientation>
   <spot_deviation>0.113</spot_deviation>
 </solution>
</index_response>
```

Figure 2
An abbreviated example of the XML for the INDEX command.

3.1. Role of the expert system

The functionality of the expert system is planned to increase as the project develops. In the initial stages, it will act primarily as a communication pathway between the beamline-control software and the data-processing software, while ultimately it will be in complete control of the data-collection experiment, including sample mounting, crystal characterization, crystal selection, data collection and data processing. The expert system will therefore have to take decisions about whether and how the data should be collected. These decisions will be based on information about the project supplied by the user (project parameters) and experimentally determined parameters provided by the data-processing software (sample

parameters). Project parameters will relate to the intended use of the data set and could include the following.

(i) Desired and minimum acceptable resolution limits.

(ii) Maximum acceptable mosaicity.

(iii) Maximum acceptable anisotropy of diffraction.

(iv) Maximum acceptable radiation damage.

(v) Maximum data-collection time.

(vi) Anomalous scatterers.

The experimentally determined parameters will include the following.

(i) Autoindexing solution.

(ii) Mosaic spread estimate.

(iii) $\langle I/\sigma(I)\rangle$ as a function of resolution.

(iv) Recommended data-collection strategy.

(v) Scaling results, including an estimate of radiation damage.

On the basis of the project parameters and the experimental data, the expert system will be able to take decisions about the following.

(i) Choice of the best sample from a number of crystals.

(ii) Whether a sample is suitable for data collection or should be rejected.

(iii) Choice of experimental parameters for the data collection.

3.2. Progress of the project

The project is currently at an early stage of development. A new server has been written (in C) for *MOSFLM* which will perform the role of the 'translator' and *MOSFLM* has been modified to allow communication by sockets with the server. The functionality of *MOSFLM* has also been modified to allow a series of operations (autoindexing, mosaicity estimation, cell refinement, integration) to be carried out without any manual intervention, with the appropriate results being passed back (in an XML document) to the server. A preliminary version of the expert system has been written (in Python) and the ProDC beamline-control software used at the ESRF has also been modified to allow communication with the expert system. As an initial proof of principle, a new button labelled 'Characterize Crystal' was added to the ProDC graphical user interface (in later stages of the project, it is anticipated that the expert system will have its own graphical user interface). Selecting this button resulted in the expert system (in this case running on a different computer but in communication with ProDC) issuing a command to ProDC to collect two oscillations images at $\varphi = 0^\circ$ and $\varphi = 90^\circ$. When the images had been collected, the expert system sent an 'INDEX' command to *MOSFLM*, which autoindexed the images, selected the best solution and sent back the results. The expert system passed these results to ProDC, where they were displayed in a message window. The Characterize Crystal button will be made available to users at the ESRF in the near future, but including additional information on the strength of diffraction [$\langle I/\sigma(I)\rangle$ as a function of resolution] and a data-collection strategy. The same functionality has recently been added to the beamline-control software PXGEN++ under development at the SRS, Daresbury, demonstrating the generic nature of the approach described here.

4. Conclusions

Tests carried out at the ESRF and the SRS have demonstrated the feasibility of providing a close coupling between data-collection and data-processing software. This is achieved with a minimal programming effort by using existing software packages and adding the capability to interpret a limited set of 'high-level' commands and to return a limited amount of information. This allows the major effort to be invested in developing the expert system, which would form the core of a fully automated system. The modular nature of the solution outlined here has several advantages. It allows different beamline-control or data-processing software to be implemented with relative ease and provides the option of having different parts of the software running on different computers. For example, the data processing may be performed on a central multiple-processor computer server located remote to the beamline. While some issues, such as a robust procedure for dealing with a variety of error conditions, still need to be addressed in detail, the results of the initial trials have been promising.

We wish to acknowledge the help of staff at SSRL in developing the server for *MOSFLM*. Financial support for this initiative has been provided by the European Community under the Quality of Life and Management of Living Resources Programme (AUTOSTRUCT) and the Access to Research Infrastructure Action of the Improving Human Potential Programme (MAX-INF project) and HRP is supported by CCP4. The initiative is described on the website http://www.dna.ac.uk.

References

Abola, E., Kuhn, P., Earnest, T. & Stevens, R. C. (2000). *Nature Struct. Biol.* **7**, 973–977.

Cassetta, A., Deacon, A. M., Ealick, S. E., Helliwell, J. R. & Thompson, A. W. (1999). *J. Synchrotron Rad.* **6**, 822–833.

Dauter, Z. (1999). *Acta Cryst.* D**55**, 1703–1717.

Ferrer, J.-L. (2001). *Acta Cryst.* D**57**, 1752–1753.

Hendrickson, W. A. (2000). *Trends Biochem. Sci.* **25**, 637–643.

Holton, J. (2002). In preparation.

Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EACMB Newsl. Protein Crystallogr.* **26**.

Mitchell, E., Kuhn, P. & Garman, E. (1999). *Structure*, **7**, R111–R121.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.

Popov, A. N. & Bourenkov, G. P. (2001). Personal communication.

Rossmann, M. G. & Arnold, E. (2001). Editors. *International Tables for Crystallography*, Vol. F. Dordrecht: Kluwer Academic Publishers.

Roth, M., Carpentier, P., Kaikati, O., Joly, J., Charrault, P., Pirocchi, M., Kahn, R., Fanchon, E., Jacquamet, L., Borel, F., Bertoni, A., Israel-Gouy, P. & Ferrer, J.-L. (2002). *Acta Cryst.* D**58**, 805–814.

Steller, I., Bolotovsky, R. & Rossmann, M. G. (1997). *J. Appl. Cryst.* **30**, 1036–1040.

Walsh, M. A., Dementieva, I., Evans, G., Sanishvili, R. & Joachimiak, A. (1999). *Acta Cryst.* D**55**, 1168–1173.