

(such as TEV or PreScission) could also be encoded to remove large flexible regions from the N-termini of the proteins which would potentially interfere with crystallisation. In addition, site-directed mutagenesis could be employed to remove undesirable surface residues or loops. Using this approach, it could be anticipated that many constructs of a particular protein in a form suitable for crystallisation may be generated.

What would be the implications of this increase in the number of proteins for crystallisation? It is estimated that the production of 1000 crystal structures per year would require approximately 10,000 proteins to be screened for crystallisation (about 800 crystallisation screens per month). If each crystallisation screen comprised 200-500 conditions, then each screen would require 2-5mg of protein at a concentration of 10mg/ml and 95% purity. In order to collect sufficient X-ray data to solve this quantity of crystal structures a dedicated MAD (Multiple-wavelength Anomalous Diffraction) beamline would be essential. Using current structure solution methods, automated electron density map interpretation and refinement packages each structure would take approximately 2-4 weeks to complete. A

team of 50-100 protein crystallographers would be required to undertake a task of this magnitude.

The key to completion of this number of structures is miniaturisation and automation at every step of the process. Cloning, expression and purification is relatively straightforward to automate; robots exist to perform liquid handling, but expression analysis would be time consuming by electrophoresis (SDS-PAGE), so an alternative must be developed. Crystallisation can also be automated (possibly at the nanolitre scale), but crystal mounting and freezing is currently a time-consuming and manual procedure, and therefore a serious bottleneck in this process. Finally, new programs need to be developed to automate structure solving and perform docking in a high-throughput mode.

It is not impossible to do high-throughput structural determination. Indeed, there are several biotech companies (*e.g.* Syrrx, Structural Genomics) with business plans based on structural genomics. Large pharmaceutical companies have a slightly different focus, with needs closer to 'functional' genomics. Although crystal structures of pure

proteins can be used for compound docking and SAR (Structure-Activity Relationship studies), more information for drug design and optimisation can be obtained from structures of protein/ligand or protein/ inhibitor complexes. This is particularly true if inhibitors of many structural classes are available for co-crystallisation. However, this is just a difference in emphasis in the needs of structural genomics projects versus drug discovery, and the methods discussed in this article can be applied and adapted to achieve both sets of aims. ■

#### REFERENCES

- [1] S. K. Burley, *Nature Struc. Biol.* 7, 932-934 (2000).
- [2] J. C. Norvell and A. Z. Machalek, *Nature Struc. Biol.* 7, 931 (2000).

#### ACKNOWLEDGEMENTS

We would like to thank all of the members of the Structural Biology, Gene Expression Sciences, Protein Biochemistry and Computational Chemistry departments at GSK in Harlow that took part in the discussions described in this article.

## AUTOMATED DATA COLLECTION AND PROCESSING FOR MACROMOLECULAR CRYSTALLOGRAPHY

A.G.W. LESLIE

MRC LABORATORY OF MOLECULAR BIOLOGY, CAMBRIDGE (UK)

**Presentation given at the ESRF Workshop  
"High Throughput Structural Biology", 20-21 February 2001.**

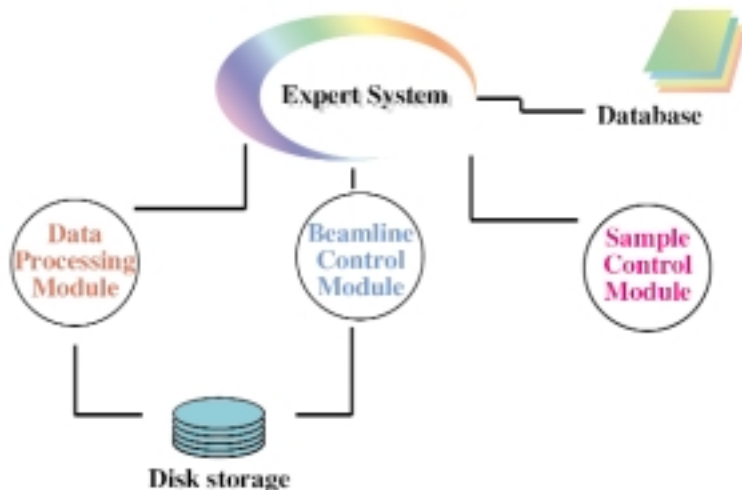
*Fully automated data collection and processing is an achievable and highly desirable objective for modern synchrotron protein crystallography beamlines. A possible scheme for reaching a high level of automation with modest programming resources is outlined.*

**S**tructural genomics initiatives such as that recently funded by the National Institute of General Medical Sciences in

the USA will lead to a dramatic increase in the rate at which macromolecular structures are determined. This increase

in throughput will be achieved by applying automation to almost all steps in the structure determination pathway,

*Fig. 1: A schematic representation of the control software for an automated protein crystallography beamline.*



from protein expression and purification through to model building and refinement. Inevitably, the steps involving collection and processing of the diffraction data will also need to become far more automated than at present, if these steps are not to become rate limiting. Based on conservative estimates of exposure time and detector readout time (10 s each per 1° image), a typical third-generation beamline equipped with automated sample loading is capable of producing 24 complete (180° rotation) datasets per day. This could represent 12 *ab initio* structures (based on 2 wavelength MAD phasing) which translates into over 3000 structures per annum. In practice, the requirement to screen multiple crystals prior to commencing data collection and the fact that not all MAD datasets will lead to a structure, will significantly reduce this number. Nevertheless, this example helps to illustrate that data collection and processing must become more automated if the potential throughput of such a beamline is to be achieved.

At present, the scientists working on the beamline are responsible for mounting samples, for operating the beamline control software and for running the data processing software. They decide whether or not to collect data from a particular crystal on the basis of visual inspection of images and information obtained from the data processing programs. They set the optimal data collection parameters (resolution, exposure time, oscillation angle, phi range) and ideally try to process at least some of the images as they are collected. In an automated system, crystal mounting and centring would be

performed robotically, and "intelligent" software (an expert system) would take the decisions currently made by the scientists, based on information provided by the processing software and "project parameter information" stored in a database. One possible scheme is outlined in Figure 1. Its modular nature is deliberate because a more monolithic or integrated structure would take longer to develop and could not easily be installed on different beamlines, which typically have their own version of the beamline control software. Communication between the modules would be at a high level such that the expert system could instruct the processing module to characterise a crystal using images from a certain directory. In response, the processing module would return information such as crystal cell parameters, possible space group, mosaicity, and resolution limit. On the basis of this information, the expert system could then choose the acquisition parameters and instruct the beamline control module to proceed with the data acquisition.

To assess the feasibility of this type of automation, it is necessary to examine the steps involved in manually characterising a crystal for data collection. The first step is to collect two images, separated by 90° in rotation angle, and to verify whether the crystal diffracts to a useful resolution, whether it is a single lattice, and whether the mosaicity and spot shapes are acceptable. If the crystal passes this "quality check", the images are autoindexed to derive a unit cell and possible spacegroup(s). The autoindexing is also checked by predicting the spots for

the two images. A more quantitative estimate of the mosaicity is determined, and a data collection strategy is worked out (total phi range(s), oscillation angle required to avoid spatial overlaps). The images are integrated to obtain an estimate of the true resolution limit, and the exposure time and resolution for data collection are chosen. Finally, the data are collected and, ideally, processed simultaneously.

The most challenging step to automate in this procedure is an assessment of the "quality" of the two initial diffraction images. This is something that is quite straightforward for an experienced crystallographer, but it would require rather sophisticated image processing techniques to extract the same information automatically. A possible solution to this problem is to extract this information indirectly, based on the success or failure of the autoindexing of the two images. Algorithms for autoindexing are now generally very robust and further improvements should help to reduce the occurrence of algorithm failure. An examination of the situations in which the autoindexing fails reveals the following causes: incorrect direct beam co-ordinates, crystal-to-detector distance or wavelength; insufficient spots found (implying weak diffraction); multiple lattices or split spots; excessive mosaic spread or too large a rotation angle. On an automated beamline, the physical parameters would be passed from the beamline control software and therefore should not be erroneous. Many of the other reasons for the failure of autoindexing would suggest that the sample is actually unsuitable for

data collection. The two images should be autoindexed separately (to detect lattice imperfections which are not visible in some regions of reciprocal space) and also together, to give more accurate cell parameters. Rejection criteria based on the rms positional residual of indexed spots and the number of spots rejected from both the indexing and the cell parameter refinement will be applied to select successful solutions. Samples which fail this test are flagged for visual inspection (possible at a later time) by the beamline operator who may (e.g. by manual spot editing) be able to find a satisfactory solution. Autoindexing will produce a list of possible solutions with different symmetries, each with its own penalty score which evaluates how well the cell parameters comply with the restrictions appropriate for that spacegroup. Generally (unless the true symmetry is triclinic) there will be a clear separation between a number of solutions with low penalties and other solutions with much higher penalties. The solution with the highest lattice symmetry from the group with low penalties will normally be selected as an initial hypothesis for the true symmetry.

The next step is to obtain an estimate of the mosaic spread. The algorithm implemented in MOSFLM [1] relies on integrating an image with different values of mosaic spread and evaluating the total intensity of all predicted reflections. This total intensity will reach a plateau when the true mosaic spread is reached.

A data collection strategy can now be determined based on the assumed spacegroup, the known spot size and the mosaic spread. Initially, a minimum total rotation range, typically divided into two or three wedges of data that should result in a high (e.g. 90-95%) completeness, will be calculated. For this rotation range(s) the maximum oscillation angle that will avoid spatial overlap will also be calculated as a function of the phi angle. This data will be collected first, in order to minimise radiation damage. A second data collection range, designed to bring the completeness to 100% and increase the multiplicity of the observations, will also be calculated. For appropriate cases (MAD data collection), the completeness of the anomalous data will be maximised. All the functionality for these calculations already exists in MOSFLM. Finally, the exposure time and resolution of the dataset need to be

defined. To do this, the two initial images are integrated and the mean  $I/\sigma(I)$  is calculated as a function of resolution. The effective resolution is defined as that at which the mean  $I/\sigma(I)$  drops below a cutoff value (e.g. 2). Using Poisson statistics, it is possible to estimate the exposure time required to achieve the desired  $I/\sigma(I)$  value at any particular resolution. This would be compared with the "maximum allowable" exposure time to arrive at a final choice of resolution and actual exposure time. Data collection can then be started. For lower symmetry spacegroups (orthorhombic or below), a few images should be collected  $90^\circ$  away in  $\Phi$  from the first data wedge. These images, together with images from the first wedge, are used to determine accurate cell parameters by post-refinement. Using these cell parameters the images are processed as they are collected. During the data collection the first two images are re-collected at regular intervals to monitor radiation damage. The images should also be scaled and merged at the earliest opportunity, to check the initial assumption of the spacegroup symmetry. If it turns out to be incorrect, the data collection strategy will need to be recalculated, taking into account the data that has already been collected.

The modularity of the system should allow maximum flexibility in implementation while minimising the programming effort. The expert system should be beamline and synchrotron independent and could be a focal point for collaboration between different synchrotron sites. Only a command "translator" module would be required to interface a standard expert system with different beamline control software. Information specific to a given sample or project will need to be supplied to the expert system in order to assist in decision making. Some of the possible project parameters are listed below:

- (a) Minimum acceptable resolution (defined as  $I/\sigma(I) > X$ )
- (b) Highest resolution required
- (c) Maximum exposure time per dataset
- (d) Maximum acceptable mosaic spread
- (e) Maximum acceptable anisotropy
- (f) Maximum acceptable radiation damage (expressed as a B factor or  $\Delta I/\sigma(I)$  at the highest resolution)
- (g) Number of wavelengths required
- (h) Anomalous scatterer

- (i) Minimum exposure time per image
- (j) Maximum number of overloaded reflections
- (k) Minimum acceptable completeness (anomalous or overall)

The number of project parameters and their default values would have to be determined on the basis of experience.

Beamline performance could be monitored by evaluating a standard sample at regular intervals. Any deterioration in quality would result in a message being passed to the beamline operator, possibly via a mobile phone.

The different processes outlined in Figure 1 would probably run on different computers. In particular, a high performance machine with rapid disk access would be required for the data processing if this is to keep pace with data collection. Given the continual improvement in performance, it is not unrealistic to expect data collection and processing to proceed in step. A collaboration is currently being established with personnel at the ESRF and SRS (UK) to work towards a practical implementation of the scheme outlined above. ■

#### REFERENCES

- [1] A.G.W. Leslie, *Joint CCP4 + ESF EAMCB Newsletter on Protein Crystallography*, **26**, (1992); [http://www.mrc-lmb.cam.ac.uk/harry/mosflm/mosflm\\_user\\_guide.html](http://www.mrc-lmb.cam.ac.uk/harry/mosflm/mosflm_user_guide.html).

#### ACKNOWLEDGEMENTS

I would like to thank P. Evans and H. Powell for many useful discussions.