Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Industrial perspective on X-ray data collection and analysis

Tadeusz Skarzynski* and James Thorpe

GlaxoSmithKline, UK

Correspondence e-mail: tadeusz.j.skarzynski@gsk.com

Received 5 August 2005
Accepted 21 October 2005

Protein crystallography methods applied by research teams in the pharmaceutical industry to support the process of discovery of new medicines are not greatly different from those used by academic structural biology groups. However, owing to the specific aims of the pharmaceutical industry, the approaches and working practices are often quite distinct. This applies to both the determination of novel structures of drug targets and complexes of these targets with potential drugs. To make any significant impact on ongoing medicinal chemistry projects, crystal structures have to be delivered on time and must provide answers to specific questions. Owing to the high number of crystal structures typically solved by industrial research groups, development of technology and computational methods which speed up the process and increase throughput is of vital importance. This paper presents an overview of current approaches to X-ray data collection and processing within the industrial environment, with examples of how technology is used to address the challenges structural biology faces in this 'high-throughput-everything' period.

1. Role of protein crystallography in the discovery of new medicines

When structural biology techniques started to make their inroads into the laboratories of pharmaceutical companies over a decade ago, the initial expectation was that we would be able to rationally design potent chemical molecules using the knowledge of the atomic structure of the active site. This turned out to be a much more difficult task for a variety of reasons; most importantly, our inability to create an adequate computational model of the complex array of physical forces involved in molecular interactions. Therefore, the main and ever-increasing role of macromolecular crystallography in the drug-design process has been in the optimization of original 'hits' from high-throughput screening (HTS) of large compound collections. However, although the compound collections of large pharmaceutical companies are growing at ever-increasing speed owing to combinatorial chemistry and may include in excess of a million compounds, the number of possible chemical entities is almost infinitely large. Therefore, HTS quite often does not result in any tractable 'hits'. Then, the detailed knowledge from the crystal structure about interactions between the active site of the target and a natural substrate or known agonist or antagonist can help to generate chemical ideas about possible classes of molecules to synthesize for subsequent screening. Also, the use of crystals of protein targets for direct screening of low-complexity compounds as starting points for the development of potent molecules is a strategy that has become an integral part of modern drug discovery (Hann & Oprea, 2004). The initial

'structure-based drug design' concept seems to have evolved into a more realistic 'structure-guided drug discovery' over the last 10–15 y.

The typical range of activities of a structural biology team in a pharmaceutical company setting range from crystal structure determination of a difficult novel mammalian protein target to obtain exclusive insights into the architecture of its active site(s) and better understand its biology to solving large numbers of routine structures of complexes with newly synthesized compounds for established targets of interest. For example, structural biology teams within GlaxoSmithKline determine about 20–30 structures of new protein targets and in excess of 400 structures of fully refined protein–ligand complexes per year. The actual number of crystal structures solved is significantly higher because often there is no evidence of compound binding (usually owing to poor compound solubility in aqueous solutions) or the occupancy of the ligand is not high enough to give a detailed picture of protein–ligand interactions.

1.1. Structures of novel targets

Crystal determination of the new macromolecular target to guide discovery of active compounds (potential medicines) in the pharmaceutical industry environment is a highly focused and concerted effort involving a large number of people. Its principal aim is to create a crystal system suitable for studying protein–ligand complexes at resolution of at least $\sim$2.5 Å where details of intermolecular interactions can be unambiguously 'seen'. A stable crystal form which allows compound-soaking experiments without disruption of the crystal lattice and produces high-resolution diffraction of X-rays is ideal for long-term ligand-optimization studies. However, in some cases a novel structure on its own, especially with a natural substrate or known active compound bound to its active site, can give a pharmaceutical company a huge competitive advantage. Therefore, a significant effort is put into molecular biology and expression of the chosen target, with tens or even hundreds of different constructs expressed and purified for initial characterization, including small-scale crystallization experiments. Since initially only small amounts of protein samples are available, crystallization trials have to be performed using sub-microlitre volumes using crystallization robotics and other small-volume systems such as Fluidigm Topaz (Fluidigm Corporation, San Francisco). Also, automated crystallization systems are needed to cope with the large number of protein samples to screen and analyse results. When crystals are produced in the initial screen, they are often small. A powerful, well collimated X-ray source is then needed to properly evaluate their diffracting properties before the crystals are further optimized. Ideally, the initial diffraction tests should be performed as soon as possible, so a bright in-house source with high-quality X-ray optics becomes a necessity. However, for new crystal structures, data-collection methods and practices in industrial and academic laboratories are very similar, with the overall goal of collecting data of the highest possible quality (high resolution, completeness, high redundancy, low noise, good spot definition *etc*.).

1.2. Protein–ligand complexes

The structural details of the ligand-binding site and its interactions with a small molecule of interest can be only obtained from the structure of a relevant complex, which can be formed at all stages of protein sample preparation. A small molecule can be picked up by the protein from an expression system and the complex carried on through subsequent purification, concentration and crystallization steps. Sometimes it is the only way to create protein–ligand complexes when the target is structurally unstable. For example, autoproteolysis of some proteases or folding problems of nuclear receptors can be resolved by expressing and purifying proteins in the presence of an antagonist or agonist. Alternatively, the target protein can be complexed with a ligand and incubated prior to being concentrated for crystallization trials. This often helps to create complexes with poorly soluble compounds or when the target aggregates at higher concentrations in the absence of a ligand. However, the most often used methods of complex formation are performed either by cocrystallization of concentrated protein with the compound added before the crystallization experiment is set up or by soaking of fully grown crystals of the protein is a solution containing the ligand. Sometimes, a combination of above methods is used; for example, cocrystals with an inhibitor or natural substrates are grown and then the original ligand is substituted by another compound by soaking ('replacement soaking'). While well formed cocrystals do not normally present any specific data-collection issues, crystals of complexes obtained by compound soaking may become damaged, change their diffraction properties or even change the space group during the soaking experiment! Specific issues related to X-ray data collection from soaked crystals and their treatment are presented further below.

2. High-throughput data collection for protein–ligand complexes

A large number of crystallographic experiments is often required to support the process of iterative optimization of chemical 'leads' found at an early stage of drug-discovery projects. For some targets, structures of tens and even hundreds of complexes with different compounds are solved over the lifetime of the project. This is especially true for fragment-based approaches to finding an active molecule. The high turnover of structures is only possible with readily available access to bright synchrotron sources and good in-house facilities. A typical 24 h data-collection run on an ESRF or APS beamline may result in about 30–50 complete data sets. This number can be even higher where sample-changing equipment is available. Also, for in-house data collection the X-ray equipment can be used much more effectively where a reliable sample changer is installed, with automatic loop centring or manual pre-centring of individual samples.

2.1. Formation of protein–ligand complexes by soaking

Soaking of crystals in appropriate solutions containing active compounds is the fastest way of creating protein–ligand complexes for suitable crystal systems, but there are a number of issues related to this method.

(i) A stock of fresh crystals has to be maintained and periodically replenished. In many cases, protein crystals deteriorate over time or become cross-linked and lose the ability to diffract X-rays. Although we have seen cases where such cross-linked crystals were still usable, it is an exception rather than a rule.

(ii) Most protein crystals are sensitive to DMSO and other organic solvents used routinely to dissolve compounds. For example, our experiments with crystals of HCV protease showed that the crystals can be soaked in a solution containing up to 30% DMSO, while crystals of other proteins used in our recent projects can typically withstand only up to 5–10% DMSO for relatively short periods of time.

(iii) Binding of potent compounds often causes conformational changes of the protein molecules and either complete or partial disruption of the crystal lattice. This effect can be very dependent on time and compound concentration. A number of initial trial soaking experiments are usually needed to establish conditions where compound binding is likely to occur without significantly deteriorating the diffraction properties of the crystal.

(iv) Sometimes, poorly soluble compounds crystallize under soaking conditions and diffraction patterns from crystals of small molecules may severely interfere with the diffraction pattern of the protein crystal sample.

(v) Very often, no binding is observed for active compounds, despite their potency under biochemical or biological assay conditions. This may be a consequence of poor water solubility of the compound, different pH of the soaking experiment compared with the assay, the presence of other chemicals in the solution or inaccessibility of the binding site in the crystal.

The successful outcome of a data-collection experiment is often hampered by one or more of these factors and a carefully chosen strategy may be required to collect the best possible data set using available crystal samples.

2.2. Data quality and meaningful structural information

To make impact on the ligand-optimization process within a drug-discovery project, structures of complexes have to be solved quickly and results fed back to project teams on time. Often, the rule ‘the structure today is better than a better structure tomorrow’ is true. While we always aim at obtaining the best possible experimental data, we often face a dilemma of collecting data from less-than-perfect crystals of complexes or delaying the diffraction experiment until a better quality crystal becomes available. In many cases, even the most carefully optimized soaking experiment results in a degradation of the diffraction pattern. Yet, if these imperfect diffraction images can be processed by one of the standard data-processing programs available today then, after processing and computation of electron-density map, they may give us the information about protein–compound interaction which may be vital to the project at this point in time. Data-processing and scaling statistics may be far from ideal, but the electron density may be sufficiently clear to allow fitting of the model of the compound molecule unambiguously. This is especially true during synchrotron data-collection trips, where the number of crystal samples for each complex is limited and there is a considerable pressure to collect as many different data sets as possible in the time available.

For example, Fig. 1(*a*) shows a diffraction pattern from the best crystal sample out of three samples for this particular complex available during a data-collection trip. The crystals were previously soaked overnight in 1 m*M* inhibitor solution also containing 2% DMSO and 10% glycerol. They were then cryocooled in liquid nitrogen and shipped to the synchrotron site in a standard dry shipper. Although the diffraction pattern was not ideal, a complete data set was collected and data processing was attempted. The autoindexing step proved difficult, but eventually after several attempts using images

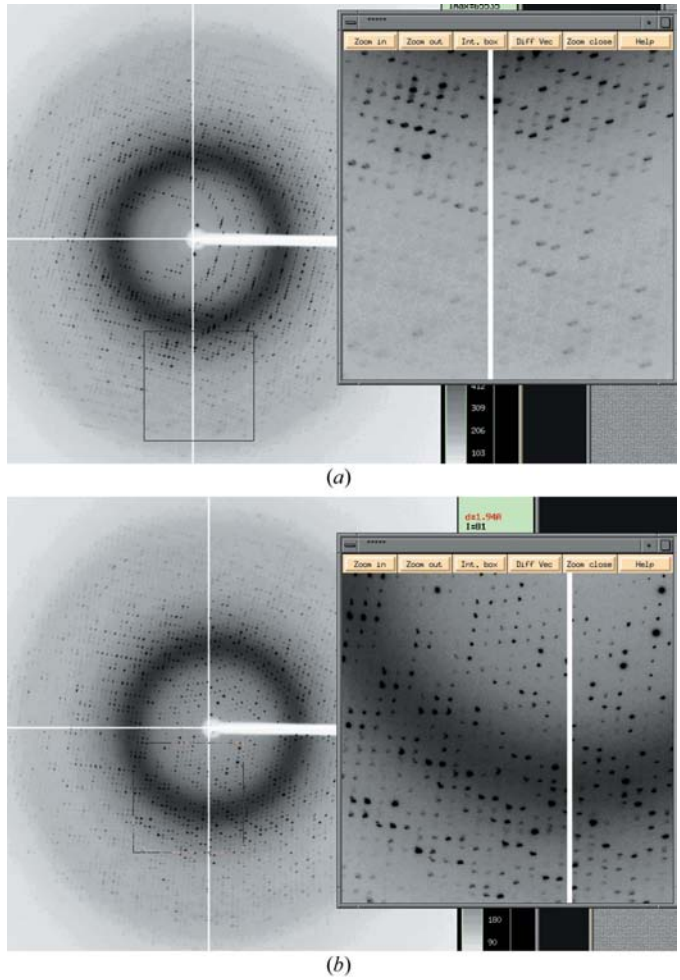

Figure 1
An example of a diffraction pattern from a soaked crystal with visible deterioration of diffraction spots. This was the best crystal of three samples available for data collection on a synchrotron trip. Image at (*a*) $\varphi = 0\,^{\circ}$, (*b*) $\varphi = 60^{\circ}$.

from different crystal orientations, it was possible to index the data using one of the images (Fig. 1*b*), about 60° away from the starting position, using *DENZO* (Otwinowski & Minor, 1997). As can be seen in Fig. 1, the difference in diffraction quality between individual images is not immediately obvious by visual inspection. The resulting orientation matrix allowed processing of the whole data set. However, after processing and scaling using *DENZO* and *SCALEPACK* (Otwinowski & Minor, 1997), with typical parameters used routinely for crystals of this protein, a significantly incomplete data set was obtained (51.3% complete to 1.80 Å resolution, $R_{\rm merge}$ = 0.081), owing to a large number of rejections during data scaling. Several attempts were made to process the data with modified parameters, which did not improve the completeness significantly. The only way to increase the completeness, was to remove the POSTREFINE command in the scaling program *SCALEPACK*, which made the data more complete (97.1%, $R_{\rm merge}$ = 0.122). Processing the same data with *MOSFLM* (Leslie, 1992) and *SCALA* (Evans, 1997), incorporated into our automatic protocol *DIY* (see below), resulted in a complete data set (99.3%, $R_{\rm merge}$ = 0.143), with default parameters. Despite relatively high values of $R_{\rm merge}$ for the

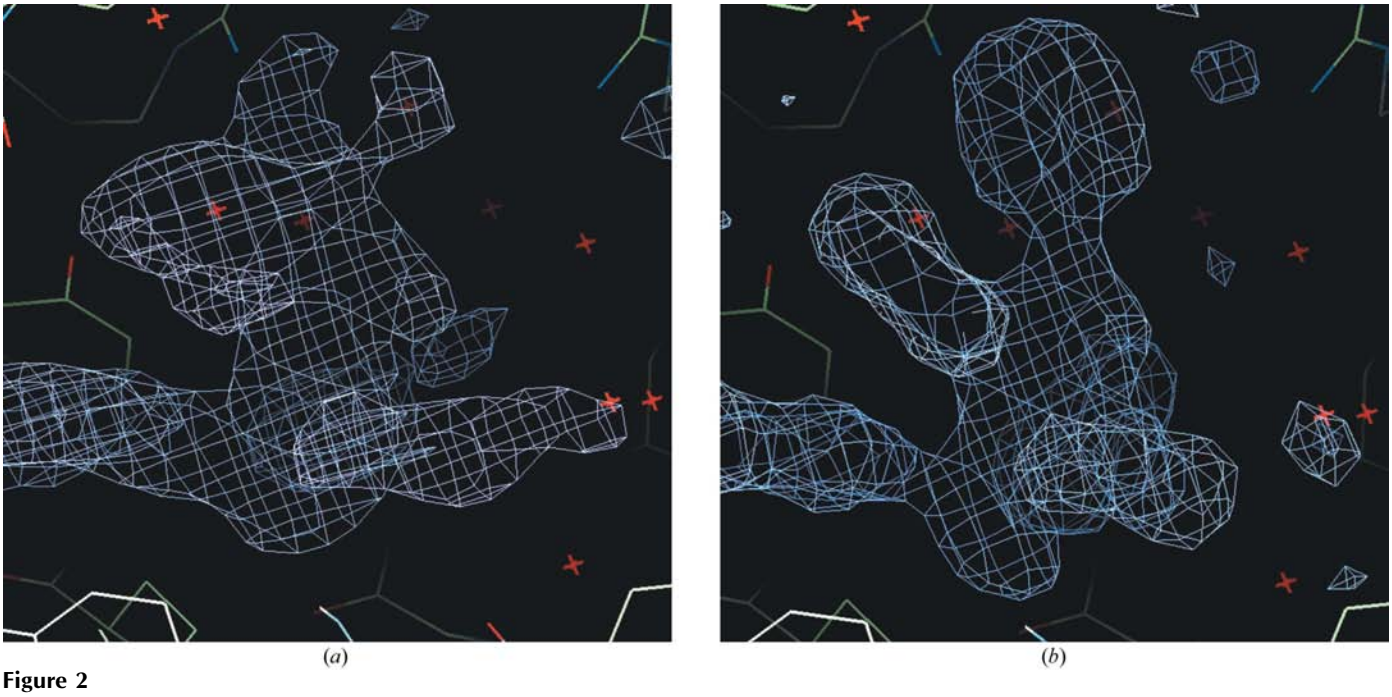

Figure 2
Electron density resulting from (*a*) 51.3% complete data set, $R_{\rm merge}$ = 0.081; (*b*) 99.3% complete data set, $R_{\rm merge}$ = 0.143

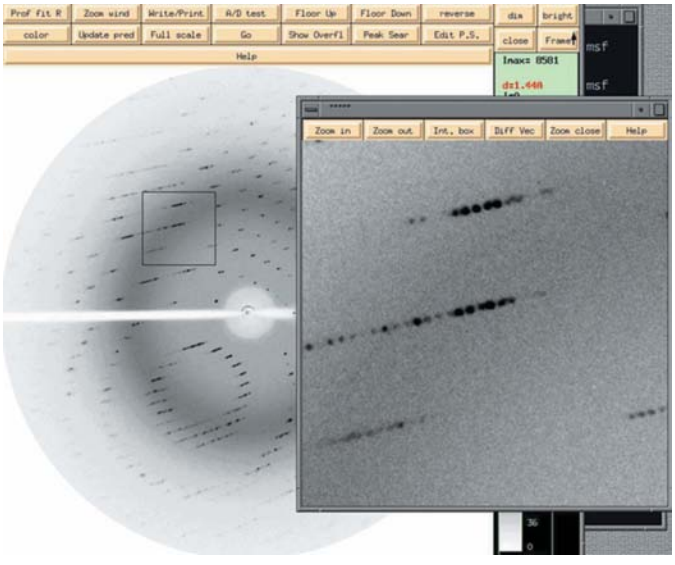

Figure 3
Diffraction pattern of a crystal diffracting to high resolution with a long unit cell (see text for details)

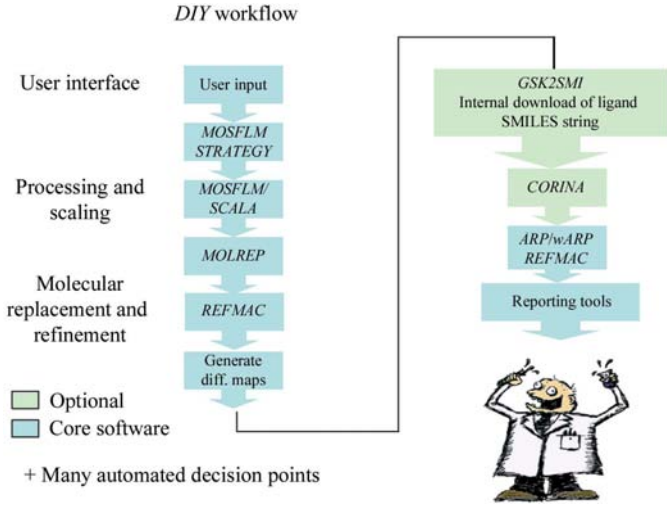

Figure 4
The main steps of *DIY*, the automatic data-processing and analysis procedure.

data sets with high completeness produced by *DENZO*/*SCALEPACK* and *MOSFLM*/*SCALA*, the electron density resulting from these data sets was very similar, informative and of high quality, while the electron density obtained from the incomplete data set was not interpretable (Fig. 2).

One of our protein targets crystallizes with a long *c* axis of 270 Å (Fig. 3) and the crystals diffract to a resolution of at least 1.9 Å in-house using a Rigaku Micromax 007 generator and beyond 1.7 Å at a synchrotron. To collect data for a series of complexes with this protein in a quick succession, dictated by medicinal project demands, we had to find a compromise between the highest possible resolution, allowing us to study protein–ligand interactions in detail, and data quality, which deteriorated significantly when diffraction spots were too close to each other. After some trials, we established a protocol that produced 1.95 Å resolution data with good processing and scaling statistics using the program *d*TREK* (Pflugrath, 1999). The protocol involved tilting the MAR 345 detector up by 15°, with a crystal-to-detector distance of 280 mm, an oscillation range of 0.2° and slits on the MAR DTB set to 0.2 × 0.2 mm to reduce the X-ray beam cross-section. Processing these data with *DENZO* and *MOSFLM* was problematic owing to the close spot separation and resulting overlaps.

Our experience with large number of 'challenging' data sets indicates that valuable information can be extracted from such data as long as the data sets are highly complete. We have also found that while all the major data-processing programs give excellent results with high-quality diffraction data, their treatment of imperfect data differs owing to different approaches to indexing, spot integration and the treatment of errors. Therefore, we now use a number of data-processing packages [*DENZO*/*SCALEPACK*, *HKL*2000 (Otwinowski & Minor, 1997), *MOSFLM*/*SCALA* and *d*TREK*], matching them with specific data-collection needs. With default or typical input parameters, *DENZO* and *SCALEPACK* seem to give best merging statistics for good crystals, with well resolved spots. The strength of *d*TREK* is in its ability of resolving very tightly populated diffraction patterns, with very close spacing between spots, while *MOSFLM*/*SCALA* seem to cope well with a wide range of typical and 'difficult' crystals, although its autoindexing is still somewhat less powerful than that provided by *DENZO*, presumably owing to *MOSFLM*'s greater dependence on the precise information about the direct-beam position. In addition, both *MOSFLM* and *d*TREK* can be run from a script, which makes them more suited for automation.

3. Automation of data processing and analysis

With the fast turnover of crystals at synchrotron beamlines and the large number of images generated in the process, efficient data processing becomes an issue. There are a number of current activities aiming at integrating data collection and processing in order to reduce the time required to successfully collect X-ray data. For example, software generated by the DNA project funded by BioXhit (the collaborative structural genomics project funded by the European Commission), currently implemented at ESRF beamlines in Grenoble, combines the control of data-collection hardware with widely used data-processing packages in a single automated system which minimizes user intervention and streamlines the whole process. Similar systems are being implemented at other synchrotron sources. However, no publicly available system exists that would combine data processing with other routine steps of protein–ligand structure determination, which is the main day-to-day activity of structural biology teams in the pharmaceutical industry. To address this problem, a prototype software solution has been created in-house, which in addition to data processing and scaling, performs subsequent steps automatically, including ligand fitting and structure refinement. The program is called *DIY* and was originally written as a C-shell script. It uses standard crystallographic programs to perform each of the steps required in the process and makes decisions based on outcomes of each of the steps. Fig. 4 illustrates the overall work flow within the program and shows the main programs used for each of the steps.

The complete script for *DIY* is generated automatically from a short C-shell script containing user input parameters such as a selection of steps to be performed, criteria for data resolution, input file names *etc.* In a typical scenario, the program is started in the directory where the X-ray image files are stored even before the data collection is completed and it starts running as soon as the number of images in the directory reaches the number specified by the user. In the first step, *MOSFLM* is used to perform automatic autoindexing, refinement and data integration, followed by *SCALA*, where in addition to data scaling, data resolution is adjusted from the values of pre-selected $R_{\rm merge}$ and $I/\sigma(I)$ parameters. A number of diagnostic parameters and warnings are generated at this and all subsequent steps. After the successful completion of data processing and scaling, an initial round of structure refinement is performed using *REFMAC* (Murshudov *et al.*, 1997) in a rigid-body mode, followed by a complete refinement of atomic parameters utilizing atomic coordinates from a file specified in the input script. If the initial refinement results in high *R* factors, the user can choose to run the molecular-replacement program *MOLREP* (Vagin & Teplyakov, 1997) or reindex the data. *MOLREP* can be also run before the refinement step on the user's request. In the next step, an attempt is made to automatically fit the ligand into the difference electron-density map generated by *REFMAC*. A SMILES string of the ligand (Weininger, 1988) is generated from its GSK registry number by an in-house program called *GSK2SMI*, followed by the generation of atomic coordinates by *CORINA* (Molecular Networks GmbH; Sadowski *et al.*, 1994). From these, a dictionary file for *REFMAC* is created and a round of *ARP*/*wARP* (Perrakis *et al.*, 1999) automated ligand-fitting protocol is run followed by the final round of refinement. After completion of the whole job, a summary report is produced by *DIY*, in addition to the log files generated by each of the programs used. So far, the success rate of the automated ligand fitting with *ARP*/*wARP* is limited (about

40–50%) and it strongly depends on the class of compounds and the level of conformational changes in the protein part of the structure. However, the modular design of *DIY* allows incorporation of other automated ligand-fitting protocols.

The *DIY* script can be used to perform any part of the whole process; for example, it can be used to calculate the data-collection strategy at the beginning of an experiment, to only run *SCALA* or to perform the ligand-fitting task using previously generated input files. The script is now routinely used by the GSK structural biology team both during synchrotron trips and in-house.

4. Conclusions

The role protein crystallography plays in supporting medicinal chemistry in the process of discovering new medicines often puts a significant pressure on the researchers to deliver crystal structures, especially structures of target–ligand complexes, as soon as practically possible. A large number of informative data sets are expected from every synchrotron trip which, with a limited number of available crystal samples, forces the scientists to collect data from imperfect crystals. Despite experimental problems resulting from compound soaking and crystal deterioration, as long as the X-ray data are highly complete, the answer to the original question can be obtained and the details of the target protein–compound interactions can significantly impact directions of synthetic medicinal chemistry effort. In addition, automatic processing and analysis of more routine complex data sets can dramatically increase the throughput and allow structural biologists to spend more time on difficult cases.

References

Evans, P. R. (1997). *Proceedings of the CCP4 Study Weekend. Recent Advances in Phasing*, edited by K. S. Wilson, G. Davies, A. W. Ashton & S. Bailey, pp. 97–102. Warrington: Daresbury Laboratory.

Hann, M. M. & Oprea, T. I. (2004). *Curr. Opin. Chem. Biol.* **8**, 255–263.

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

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

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

Perrakis, A., Morris, R. J. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.

Pflugrath, J. W. (1999). *Acta Cryst.* D**55**, 1718–1725.

Sadowski, J., Gasteiger, J. & Klebe, G. (1994). *Chem. Inf. Comput. Sci.* **34**, 1000–1008.

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

Weininger, D. (1988). *J. Chem. Inf. Comput. Sci.* **28**, 31–36.