Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

From information management to protein annotation: preparing protein structures for drug discovery

Tom Peat, Eric de La Fortelle, Janice Culpepper* and Janet Newman

Structural GenomiX Inc., USA

Correspondence e-mail: janice@stromix.com

Received 1 May 2002
Accepted 19 September 2002

In contrast to academic pursuits of structural genomics, Structural GenomiX (SGX) solves protein structures at high throughput for the main purpose of enhancing drug-discovery projects, either internally or in partnership with pharmaceutical/biotechnology companies. This involves a radical redesign of the pipeline of methods that turn a gene sequence into a three-dimensional protein structure. The various processes all report electronically to a Laboratory Information Management System (LIMS) to make sure all the parameters of the experiment are recorded in an accessible and 'mineable' form, helping guarantee reproducibility of results. Quality control at several key points keeps the process from branching out on a wrong hypothesis. Protein annotation, in a broad sense, takes care of the interpretation of a protein crystal structure or the crystal structure of one or several protein–ligand complexes. This interpretation both gathers all necessary biological information (protein function, mechanism, specific features within a protein family *etc.*) and hands over this information in a form accessible to medicinal chemistry teams designing specific small-molecule agonists or antagonists.

1. Description of the technology platform

1.1. Introduction

Structural GenomiX (SGX) has developed a high-throughput gene-to-structure platform that is being used to decrease the time necessary for drug discovery. An important aspect of this platform is that it captures the data generated in a database for data mining and quality-control purposes. This is an Oracle database that was generated in-house at SGX specifically as a tool to continually improve the gene-to-structure process.

1.2. Protein production

The platform starts with the identification of a target sequence in the bioinformatics department. A genomics approach is taken: *i.e.* sequence-alignment tools align multiple orthologous sequences from various organisms and suggest a possible domain structure for the protein. Multiple orthologs and start/stop points for the domains of interest are generated and oligonucleotide primers are ordered in a 96-well plate for the PCR. This plate is received by the molecular-biology department and put into our robotics suite where the PCR, cloning and testing of expression occur (all in 96-well or higher density plates). Positive clones, those that express and where the protein is soluble, are then grown up on a large scale by the fermentation group and the cell pellets are passed on to the purification group. Using a combination of affinity, ion-exchange and gel-filtration columns, set up on Amersham Biosciences Äkta machines modified for high throughput, the proteins are purified to greater than 95% purity (preferably 99%). Homogeneity (in terms of isoforms and low-molecular-weight additives that are not separated on an SDS gel) is systematically tested by mass spectroscopy.

1.3. Crystallization

The proteins are then set up for crystallization and stored in a custom storage/retrieval system, which is integrated with a proprietary imaging system. Both systems were developed and manufactured for SGX by Robodesign Inc. (Carlsbad, CA, USA) according to specifications provided by SGX. This system provides a complete automated solution to the need to store and schedule regular inspections of 2 000 000 crystallization drops. The crystallization team is then notified when crystallization events occur. Semi-automatic scoring of the crystallization plates provides a template for crystal optimization: a majority of crystallization trays used for initial screening, even when they do not produce initial crystals, will provide sufficient statistical information to guide an optimization of crystallization conditions. The LIMS captures the results of each crystallization experiment, both as images and as an associated numeric score. These data are used to generate the experimental conditions for second and further rounds of crystallization refinement.

1.4. Structure solution

Once diffraction-quality crystals are obtained, if necessary after optimizing the crystallization conditions around the best screening hits, these crystals are looped out of the tray, frozen and sent for data collection to SGX's dedicated synchrotron beamline, SGX-CAT, at the Advanced Photon Source (APS). Each crystal mounted for X-ray analysis is entered into the LIMS with a unique crystal identifier and this becomes the basis for the automatic generation of filenames and directory structures for the crystallographic analysis. This allows any crystallographer to immediately see the extent and progress of any project. The raw X-ray diffraction images are integrated and converted into structure-factor files at the APS and these files are sent back to San Diego (*via* a high-bandwidth Internet connection). A multi-step semi-automated procedure is then used to turn diffraction data into partially interpreted electron-density maps, which are corrected, completed and refined by the crystallography group. Particular attention is directed to maintaining the highest quality standards for the structures we solve. For this purpose, we have established a certification procedure (Badger & Hendle, 2002), with a separate group in charge of 'accepting' or 'asking for more information' for structures submitted to the SGX database. The certification information is also used, for selected high-value structures, as a template for the submission of patent applications.

1.5. Annotation of crystal structures for lead discovery

Newly solved structures are analyzed and annotated in collaboration between the crystallography group and the structural bioinformatics group. The purpose is to determine the function of the protein if it is unknown and the molecular basis of the mechanism whenever possible. SGX has developed proprietary software to find the most likely active site (if any) and compare it with other active sites in the PDB. Local structural homology is able to generate a hypothesis about function for the protein in cases where there is neither sequence homology nor overall structural homology. In cases where the function of the protein is well known (for example, within the kinase family), the same bioinformatics tools can be applied to compare active sites across the family and find differences to be exploited for the design and optimization of selective small-molecule ligands.

1.6. Co-crystallization platform and medicinal chemistry

The 'first structure' of a novel protein target, be it an unliganded protein or already in complex with a ligand that facilitates crystallization, opens the possibility of in-depth structural studies of co-crystal structures between this target and a wide variety of small molecules selected through functional screening, affinity-based screening, virtual screening or similarity searches. The high throughput of the APS beamline (capacity of 50–100 co-crystal diffraction experiments per day) makes it possible to visually validate large numbers of compounds and binding modes in a very short time. The information is then used either by SGX's medicinal chemistry group or by pharmaceutical/biotechnology partner companies to guide lead discovery and optimization efforts.

2. Laboratory Information Management System (LIMS)

2.1. Introduction

Each step of the process, from gene selection and cloning to co-crystallization and chemistry, has its own LIMS interface. These interfaces allow specialized data input by anyone participating in the process and allow supervisors to monitor where each target is in the system. The LIMS captures much of the data automatically as downloads from individual robots, chromatograms from purification runs *etc.* All data relating to a given clone, from initial sequence to structure and annotation, can be seen in the database by simply selecting that clone name. Quality-control measures, such as mass-spectrometry data and activity assays, are also linked to the clone name in the database. As processes change or new technology develops, new tables are created in the database and these data are also captured for future data-mining efforts. This iterative cycle is extremely useful in improving the efficiency of the process.

2.2. Example 1: optimization of crystal screens

Crystallization provides an example of how LIMS data are used. As described in §1.3 above, the crystallization procedure

Acta Cryst. (2002). D**58**, 1968–1970 Peat *et al.* • Protein structures for drug discovery **1969**

for a novel protein consists of a first 'general' screen followed by one or several optimization screens that use the hits (defined as the appearance of crystalline material in crystallization drops) in the first screen as a guide to the optimal crystallization area in 'reagent space'. Whereas it is rare to find a diffraction-quality crystal in the first screen, the number of crystalline hits providing information for follow-on screens is directly correlated to the probability of successful optimization.

The first two 96-well solution 'general screens' were developed early at SGX by pooling together data from public sources (Hampton screens, macromolecular crystallization database; Jancarik & Kim, 1991) with additional insights from crystallographers' experience. The large number of proteins coming through the platform, most of them purified in the same way and handed over in a standard buffer, gives us a relatively unbiased sample to detect crystallization trends. Data-mining the first 600 000 scored crystallization experiments recorded in the LIMS enabled the SGX statistics group to identify imbalanced areas of crystallization space and significantly upgrade the crystallization screen. After two such upgrades, we currently use a fourth-generation 96-well screen that is significantly more effective (in terms of the number of 'crystal hits' obtained) than the original screens we started with.

An experiment in support of this claim was designed where the test protein, hen egg-white lysozyme (50 mg ml^{-1} in water; protein from Sigma Chemicals), had not been used in the statistical analysis that led to the improvement of the screens. The results of an overnight crystallization experiment at 294 K are described in Table 1.

Table 1
Improvements in crystallization score.

	Hampton screen	SGX 'Third 96'	SGX 'Fourth 96'
No. of conditions that yield crystals	9	12	18

2.3. Example 2: protein domain definition

Using computational and experimental data together to improve the process is a theme that runs throughout the platform. The genomics approach of using multiple start/end points to delimit a given domain, as well as different orthologs of the same gene, maximizes the chances of obtaining soluble protein for crystallization. However, there are instances when this protein does not crystallize or the protein proves to be unstable. Under these circumstances, limited proteolysis followed by mass spectrometry is used to experimentally define domain boundaries of specific targets. These data are stored in the database and as the target class is expanded, more data are collected. These data are then used to improve the algorithms for domain definition in the programs used for both initial target selection and for modelling.

3. Conclusion

The SGX platform and people have shown that they were able to solve and process large numbers of protein structures by X-ray crystallography. However, the value created in that platform can only find its full expression in a lead discovery process, extensively nourished by structural information, that is able to reproducibly deliver high-quality lead compounds for pharmaceutically important protein targets.

The authors with to acknowledge the entire SGX staff, in particular Jon Christopher, John Badger and the LIMS group for their dedication to proteins, structures and the constantly improving process.

References

Badger, J. & Hendle, J. (2002). *Acta Cryst.* D**58**, 284–291.
Jancarik, J. & Kim, S.-H. (1991). *J. Appl. Cryst.* **24**, 409–411.