Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The high-throughput protein-to-structure pipeline at SECSG

Zhi-Jie Liu,[a] Wolfram Tempel,[a] Joseph D. Ng,[b] Dawei Lin,[a] Ashit K. Shah,[a] Lirong Chen,[a] Peter S. Horanyi,[a] Jeff E. Habel,[a] Irina A. Kataeva,[a] Hao Xu,[a] Hua Yang,[a] Jessie C. Chang,[a] Lei Huang,[a] Shu-Huey Chang,[a] Weihong Zhou,[a] Doowon Lee,[a] Jeremy L. Praissman,[a] Hua Zhang,[a] M. Gary Newton,[a] John P. Rose,[a] Jane S. Richardson,[c] David C. Richardson[c] and Bi-Cheng Wang[a]*

[a]Southeast Collaboratory for Structural Genomics, Department of Biochechemistry and Molecular Biology, University of Georgia, Athens, GA 30602, USA, [b]Laboratory for Structural Biology and Department of Biological Sciences, University of Alabama in Huntsville, Huntsville, AL 35899, USA, and [c]Department of Biochemistry, Duke University, Durham, NC 27710, USA

Correspondence e-mail:
wang@bcl1.bmb.uga.edu

Received 14 November 2004
Accepted 25 April 2005

Using a high degree of automation, the crystallography core at the Southeast Collaboratory for Structural Genomics (SECSG) has developed a high-throughput protein-to-structure pipeline. Various robots and automation procedures have been adopted and integrated into a pipeline that is capable of screening 40 proteins for crystallization and solving four protein structures per week. This pipeline is composed of three major units: crystallization, structure determination/validation and crystallomics. Coupled with the protein-production cores at SECSG, the protein-to-structure pipeline provides a two-tiered approach for protein production at SECSG. In tier 1, all protein samples supplied by the protein-production cores pass through the pipeline using standard crystallization screening and optimization procedures. The protein targets that failed to yield diffraction-quality crystals (resolution better than 3.0 Å) become tier 2 or salvaging targets. The goal of tier 2 target salvaging, carried out by the crystallomics core, is to produce the target proteins with increased purity and homogeneity, which would render them more likely to yield well diffracting crystals. This is performed by alternative purification procedures and/or the introduction of chemical modifications to the proteins (such as tag removal, methylation, surface mutagenesis, selenomethionine labelling *etc.*). Details of the various procedures in the pipeline for protein crystallization, target salvaging, data collection/processing and high-throughput structure determination/validation, as well as some examples, are described.

1. Introduction

The Southeast Collaboratory for Structural Genomics (SECSG; http://www.secsg.org; Adams *et al.*, 2003), one of the nine structural genomics pilot centers funded under the National Institutes of Health (NIH) Protein Structure Initiative (PSI; http://www.nigms.nih.gov/psi), is a networked center consisting of five partner institutions located in the southeast USA. SECSG structure-determination efforts are currently targeting members of large Pfam families (Bateman *et al.*, 2004) that lack a representative structure in the Protein Data Bank (PDB; Berman *et al.*, 2000). The genomes of *Pyrococcus furiosus* (Robb *et al.*, 2001) and *Caenorhabditis elegans* (*C. elegans* Sequencing Consortium, 1998), as well as selected human proteins from the Mammalian Gene Collection (MGC) maintained by the American Type Culture Collection (http://mgc.nci.nih.gov) are the focus of these efforts (Table 1). SECSG is composed of four research cores: protein production, X-ray crystallography, NMR and bioinformatics.

A high throughput protein-to-structure pipeline has been developed by the crystallography core. It integrates robotics and other automation technologies into three modules that interact closely: crystallization, crystallomics (target salvaging) and structure determination/validation. Relational databases provide the backend for communication between these relatively independent modules. While nearly four years of experience with the pipeline confirm the significance of automation, we also became acutely aware of the importance of data management. The multiplicity of samples and the large amount of data associated with a high-throughput operation extend the challenge beyond a simple scale-up of traditional laboratory practices. More planning, testing and fine-tuning and a more complex

doi:10.1107/S0907444905013132

approach to project management are an absolute requirement. Success in this environment relies heavily on the integration not only of hardware and software, but also of the various pipeline stages.

At this point in time, structural genomics still suffers from the relatively high cost and low success rate in going from purified protein samples to structures. In high-throughput mode, each protein target receives an equal amount of attention and consequently easier targets ('low-hanging fruit') will be solved while the more difficult ones will generally be abandoned before they yield a structure. Based on the current statistics of the nine NIH PSI centers, on average only 13% of purified protein samples resulted in a successful structure determination. Therefore, 87% of the purified samples are abandoned before a structure is obtained. In order to balance the throughput and the success rate of the structure-determination process, an alternate path (crystallomics) for rescuing these failed targets was implemented at SECSG in the third quarter of 2003 using NIH supplemental funding providing a two-tiered approach for protein production. SECSG tier 1 protein-production activities are focused on producing all proteins (both low-hanging and high-hanging fruit) from the *P. furiosus* and *C. elegans* genomes using high-throughput methods and selected MGC human proteins using more traditional methods. The SECSG tier 2 protein production efforts support tier 1 production activities. In that role, the crystallomics group provides scaled-up amounts of tier 1 protein repeats for further crystallization trials, where necessary, and prepares labeled and otherwise modified proteins for crystal optimization and structure-determination purposes.

Here, we describe briefly the protein to structure pipeline developed by the SECSG including procedures of crystallization, crystal diffraction characterization, data collection/processing, structure refinement/validation and data management (see Fig. 1).

2. Materials and methods

2.1. Crystallization

When protein samples become available for crystallization trials, the protein-production group responsible for the protein notifies the crystallization group *via* the ExpressSG database. A product sheet listing all relevant information about the protein including sequence, predicted isoelectric point, concentration, buffer, metal content and the scanned image of an SDS–PAGE gel accompanies each sample. The product information is also available on the web (http://www.secsg.org/cgi-bin/report.pl). Prior to screening, protein samples are assigned a barcode ID and again checked for purity by both PAGE and dynamic light scattering (Kadima *et al.*, 1990). Each sample is screened against 384 reagent mixtures made up from seven commercial sparse-matrix screens: Crystal Screen, MemFac, PEG Ion, Crystal Screen Cryo, Crystal Screen II (Hampton Research), Wizard I and II (Dedode Genetics) and the locally developed SECSG MP1 screen containing 48 conditions (Shah *et al.*, 2005). Condition Nos. 25 and 27 were removed from Crystal Screen I and Crystal Screen Cryo because they have had the lowest success rate in crystallizing proteins in the past and thus were omitted from the screens.

Table 1
Accumulated summary of protein structure determination at SECSG as of 30 June 2004.

Protein	Purified	Crystallized	X-ray data	Structure
P. furiosus	201	102	53	22
C. elegans	196	73	20	14
Human	41	5	4	3
Total	438	180	77	39

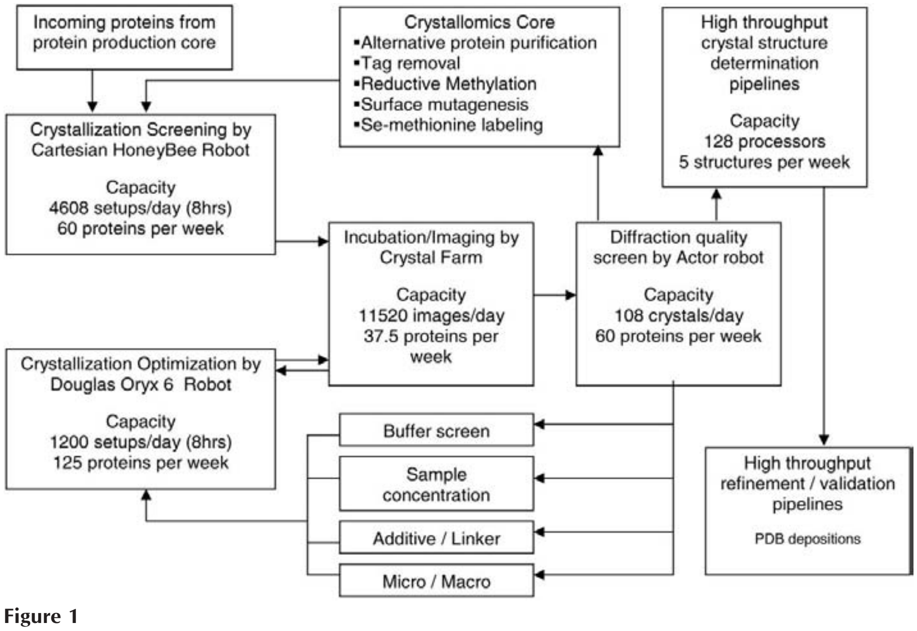

Figure 1
Diagram outlining the SECSG high-throughput protein-to-structure pipeline.

Initial screening is carried out by the sitting-drop vapor-diffusion process using Greiner Crystalquick plates set up with a Cartesian Honeybee crystallization robot (Genomic Solutions). Drop compositions consist of 200 nl protein solution plus 200 nl reservoir solution, giving a total drop volume of 400 nl. Once a plate has been barcoded and set up, it is moved to the CrystalFarm incubator (Discovery Partners International) for storage, imaging and scoring. Crystals are scored as follows: (1) clear, (2) precipitate, (3) crystal and (4) harvestable. Images and scores are currently recorded in the CrystalFarm database and transferred to the Crystal Monitor (Decode Genetics) database application. Crystal hits are optimized by the modified microbatch method (Baldock *et al.*, 1996) using a single or double grid-screen approach around the selected condition. The optimization is performed using locally modified Douglas Instruments ORYX robots using 1 µl drops containing equal volumes of protein and precipitant solutions (Shah *et al.*, 2004) on a 72-well Nunc plate (Nalge Nunc International). A Genesis RSP robot (Tecan) is used to reformat the commercial screens for the initial trials and to prepare grid screens for optimization.

2.2. Crystallomics and target salvaging

Several techniques are applied, some in combination, to salvage protein targets that fail to produce well diffracting crystals. In the case of overexpression of recombinant hyperthermophile proteins, the cell lysate may undergo heat treatment at 343 K for 1 h to precipitate contaminants. In a modification of the standard purification procedure for oligohistidine-tagged proteins, the batch elution step is replaced by gradient elution with an increasing imidazole concentration (0 to 1 *M*). Reductive methylation of the amino groups of surface lysine residues is performed according to a published procedure (Rayment, 1997).

2.3. Robotic crystal diffraction-quality screening

Harvestable crystals (dimensions greater than 50 μm) are mounted, flash-frozen and screened for diffraction quality in-house using a Rigaku/MSC ACTOR robot (http://www.rigakumsc.com). Diffraction-quality crystals are then recovered and stored at cryogenic temperatures awaiting data collection. All information from the diffraction screening, including crystal size, unit cell, space group, resolution, mosaicity and storage location is recorded in the XtalDB database (a web-enabled system developed at SECSG) for process control and future reference.

2.4. Data collection and data processing

Data is currently collected at beamlines administered by the Southeast Regional Collaborative Access Team (SER-CAT, Sector 22, Advanced Photon Source, Argonne National Laboratory) using both beam time allocated to the University of Georgia and the mail-in data-collection program. Additionally, data can be collected in-house on CCD (Bruker SMART 6000 and Rigaku Saturn92) detectors using a copper rotating-anode source (Rigaku FR-D X-ray generator) or on an Rigaku R-AXIS IV image-plate detectors using a chromium rotating anode and associated confocal optics. Standard phasing protocols include sulfur/metal SAS phasing from native crystals (Liu *et al.*, 2000; Wang, 1985), Xe SAS phasing, iodine SAS phasing using quick soaks (Dauter *et al.*, 2000) and Se-Met SAS/MAD phasing. A promising phasing technique using Cr $K\alpha$ radiation and native crystals is also being developed as part of SECSG's Direct Crystallography approach (Chen *et al.*, 2004; Yang *et al.*, 2003). Data reduction is carried out using either the *d*TREK* (Rigaku/MSC) or *HKL*2000 (Otwinowski & Minor, 1997) suites. Development and testing of an SECSG-developed data processing package is ongoing and emphasizes the extraction of weak anomalous signal.

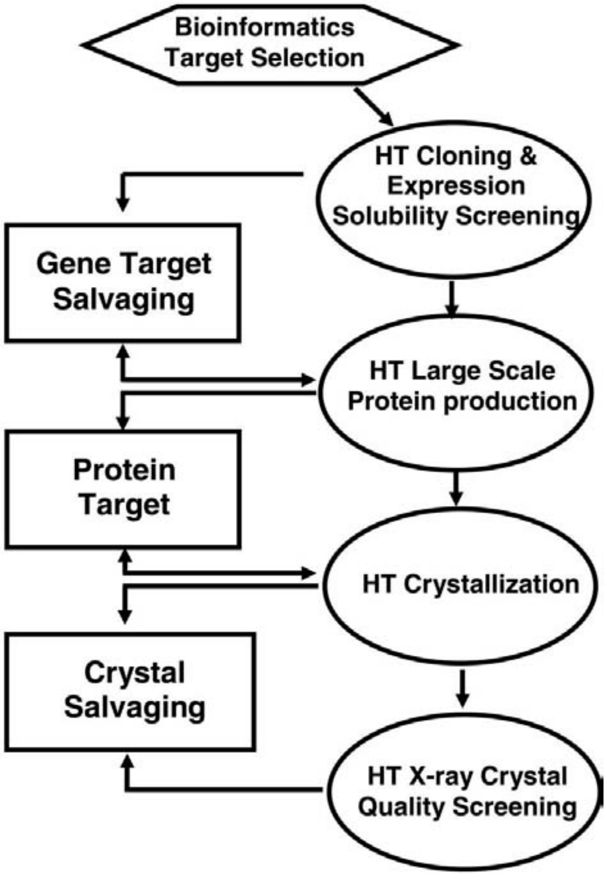
Bioinformatics Target Selection
HT Cloning & Expression Solubility Screening
Gene Target Salvaging
HT Large Scale Protein production
Protein Target
HT Crystallization
Crystal Salvaging
HT X-ray Crystal Quality Screening

Figure 2
Operation of the high-throughput (HT) pipeline from protein expression to crystallization. Tier 1 is the mainstream production unit and the major tasks are diagramed in circles. Tier 2 shows the salvaging pathways and the salvaging components are outlined in squares. In this report, a total of 50 targeted proteins were run through the salvaging pathways and consequently structures were obtained where they would not have been acquired otherwise.

2.5. High-throughput crystal structure determination pipelines

Characterization of the heavy-atom or anomalous scattering substructure for SAS, MAD, SIRAS and MIRAS experiments is carried out using *SOLVE* (Terwilliger & Berendzen, 1999) or *SHELXD* (Schneider & Sheldrick, 2002). Initial phasing is carried out using either the *SOLVE* or *ISAS* (Wang, 1985) packages. Phase improvement utilizes the programs *RESOLVE* (Terwilliger & Berendzen, 1999) and *DM* (Cowtan & Main, 1998). For molecular-replacement calculations, the programs *AMoRe* (Navaza, 2001), *EPMR* (Kissinger *et al.*, 1999), *PHASER* (Storoni *et al.*, 2004) or scripts from *CNS* (Brünger *et al.*, 1998) suite are used. The programs *SOLVE*, *RESOLVE*, *ISAS* and *AMoRe* have been integrated into HT pipelines running on a multi-node computer cluster (manuscript in preparation).

2.6. Structure-refinement and validation pipeline

Automated model (re-)building is carried out with the programs *RESOLVE* (Terwilliger, 2003) and *ARP/wARP* (Perrakis *et al.*, 1999) in an iterative manner with maximum-likelihood positional and thermal parameter refinement by *REFMAC* (Murshudov *et al.*, 1997). Models are examined for errors based on their correlation with experimental data (*SFCHECK*; Vaguine *et al.*, 1999), main-chain and side-chain torsion angles and atom clashes after addition of H atoms (*MOLPROBITY*; Davis *et al.*, 2004). Manual rebuilding, when necessary, with *XFIT* (McRee, 1999) is iterated with *REFMAC* and further validation. For submission to the PDB, the program *PDB_EXTRACT* is used.

3. Results and discussion

3.1. Crystallization

The current SECSG crystallization facility at the University of Georgia was developed to handle ten new protein samples/constructs per 8 h day, aimed at analyzing 2000 or more proteins/constructs annually consistent with PSI-2 goals.

Incoming samples from the protein production and crystallomics cores are assigned an independent data matrix two-dimensional barcode and the physical characteristics of the sample (buffer, concentration, modifications *etc.*) and storage location are recorded in a sample-tracking database, SampleDB. All barcodes are printed using a high-resolution barcode printer. Any alteration of the initial sample (*e.g.* protein concentration change due to dilution) results in the assignment of a new barcode and database entry. Aliquots of the same protein sample can therefore be tracked independently in the crystallization records.

Incoming proteins are screened against seven commercially available screens and an SECSG developed MP1 screen, giving a total of 384 crystallization screening experiments using a total of less than 80 µl protein solution.

Crystal trays are barcoded (hexadecimal six character linear barcode) and logged into the crystallization database together with details of the protein sample and the screening matrix used. Currently, all crystallization experiments, once set up, are loaded into the CrystalFarm integrated imaging/incubator unit. The plate barcodes are read, imaging is scheduled and images and scoring results stored in the CrystalFarm database. Currently, drops are photographed once a week for four weeks. When crystals of sufficient size are found, they are harvested (see below). If further optimization is needed, the solutions of optimization screens are mixed and dispensed into a 96-well 1.5 ml deep block tray or 96-well tray using a Tecan liquid-handling robot. Optimization screens consist of a single (36-well) or double (72-well) grid screen based on variations of buffer pH, precipitant and salt concentrations. Optimization trays are stored and imaged using the CrystalFarm as described above. Further optimization using the Hampton Research Additive Screens 1 and 2 is carried out in cases where initial optimization fails to produce diffraction-quality crystals. Finally, if the additive screen still fails to produce diffraction-quality crystals, the protein target is diverted to the target-salvaging pathway (see §3.2). The accumulated summary for the crystallization effort is shown in Fig. 2.

3.2. Crystallomics and target salvaging

Using the purification and sample-modification techniques described previously, seven structures were produced from a set of 50 *P. furiosus* proteins that either failed to crystallize or gave poor diffracting crystals. The following examples illustrate the potential of the SECSG target-salvaging pathway. Conserved hypothetical protein PF0863 (Pfu-838710) gave only marginally diffracting crystals from the initial sample. Polyacrylamide gel electrophoresis (PAGE, overloaded gels) of the initial sample showed that it contained a number of minor but significant bands. Upon re-purification, crystals diffracting to 2.3 Å resolution were obtained. The structure was solved and a model is currently undergoing refinement. Conserved hypothetical protein PF0380 (Pfu-392566) failed to give crystals from the initial sample. Upon re-purification, the sample produced crystals that were too small for X-ray analysis. Reductive methylation of the re-purified protein however gave crystals that diffracted to 1.2 Å resolution and the structure was solved. A refined model was deposited in the PDB (code 1vk1).

3.3. Robotic crystal diffraction-quality screening

Crystal classification according to diffraction quality is handled by the Rigaku ACTOR/Director system that mounts flash-frozen crystals (Hope, 1988; Teng, 1990) from an LN2 dewar onto the goniometer and automatically centers the crystal, collects images, indexes the crystal and if desired collects data unattended. Within 18 h, divided into two daily shifts, the unit cell, Laue group and diffraction limits of more than 100 crystals can be determined. Automation of this tedious procedure relieves personnel and reduces the probability of missing well diffracting specimens among a large number of poorly diffracting crystals. Based on the diffraction information, available beam time on the home or synchrotron sources can be prioritized.

3.4. Data collection and data processing

Several X-ray sources with different characteristics are available for a range of data-collection problems. Several beamlines providing tunable, intense and brilliant X-rays are accessible to SECSG at the Advanced Photon Source (Argonne National Laboratory) for phasing and high-resolution experiments. Several SECSG structures were solved during the commissioning stages of both SER-CAT beamlines. The University of Georgia share in SER-CAT (Sector 22) provides approximately 24 h beam time per month of operation on its both 22ID insertion-device beamline and 22BM bending-magnet beamline once it is fully commissioned. The SER-CAT beamlines are equipped with state-of-the-art MAR Research (http://www.mar-usa.com) 300 mm (22ID) and 225 mm (22BM) CCD detectors. A dual-port chromium rotating anode at the University of Georgia is equipped with large R-AXIS IV image-plate detectors, chromium confocal optics and helium beam paths. Its soft X-rays (λ = 2.29 Å) are suitable for the exploitation of weak anomalous scatterers such as sulfur in the phasing of protein structures. This practical implementation of direct crystallography, has resulted in the solution of eight new structures since its installation in 2003. The structure determination of the hyperthermophile protein Sso10a (Chen *et al.*, 2004) is a recent example of the application of this. A dual port ultrahigh-intensity copper rotating anode equipped with confocal optics and CCD detectors at the University of Georgia is also available for data collection and crystal characterization (see §3.3).

3.5. High-throughput crystal structure-determination pipelines

Pipelines for *de novo* and molecular-replacement structure solution have been implemented on a Linux-based multi-node computer cluster (provided by an IBM SUR Award). The combination of software pipelines and a high-throughput computing environment permits the easy setup of hundreds of jobs. Thus, for a given set of data, multiple SAS, MAD or molecular-replacement computations can be set up each using a slightly different set of program input parameters. For example, computations can be carried out using different data resolution ranges simply by defining the minimum and maximum resolution to be screened and the size of the resolution step used to cover the desired range. The pipeline workflow manager then uses this information to generate the appropriate program inputs for N jobs that are required to satisfy the request. This approach has been applied with success in more than 30 cases. For example, *P. furiosus* DNA-directed RNA polymerase subunit ε'' crystallizes in a trigonal space group, which could only be determined from the successful structure determination. The pipeline was used to carry out computations in all candidate space groups and the correct solution identified. A model of this protein refined at 1.38 Å resolution has been deposited in the PDB with code 1ryq. The *AMoRe* pipeline allows easy screening of resolution limits for the rotation and translation searches as well as radius used for intramolecular Patterson peaks and has produced solutions in cases where the search model exhibited less than 30% sequence identity (for example *P. furiosus* NADH oxidase/nitrite reductase). A model for this protein is currently undergoing refinement.

3.6. Structure-refinement and validation pipeline

While always striving to increase the throughput of the protein-to-structure pipeline, SECSG also aims at producing structural models of the highest quality. Procedures for the combination of newly developed structure-validation tools with refinement programs in use at SECSG for all stages of refinement have been evaluated and have become part of standard procedures. Our current approach uses (i) updated versions of the standard Ramachandran side-chain rotamer database and bond-angle criteria (Lovell *et al.*, 2000, 2003), (ii) crystallographic R, $R_{\rm free}$ (Brünger, 1992) and difference map peaks,

(iii) hydrogen-bonding and analysis of side-chain amide and imidazole orientation (Word, Lovell, Richardson *et al.*, 1999) and (iv) H-atom addition and all-atom steric clashes (Richardson *et al.*, 2003; Word, Lovell, LaBean *et al.*, 1999). All recent structure submissions from the University of Georgia's SECSG crystallography core have undergone the automatic correction of Asn/Gln/His flips available in *REDUCE* or online at the *MOLPROBITY* site (http://kinemage.biochem.duke.edu) and *MOLPROBITY*'s rotamer, Ramachandran and clash information has been incorporated early on in the refinement process. As the procedures became more integrated, the final structures improved in all criteria (Arendall *et al.*, 2005).

Structural models are deposited into the Protein Data Bank utilizing the *PDB_EXTRACT* tool (Yang *et al.*, 2004). For this purpose, relevant program output for diffraction data reduction, phasing and refinement is stored in a centralized location, ensuring that at the time of deposition, all information pertaining to a given model is easily accessible for the extraction of relevant data items.

4. Conclusions

The crystallography core at SECSG has developed a high-throughput protein-to-structure pipeline using a variety of robots and automation procedures. This pipeline is composed of three relatively independent yet closely connected activities: crystallization, crystallomics and structure determination/validation. Communications between the modules are supported by several relational databases. After four years of implementation and testing of the pipeline, we have found that the use of the robots and automation is essential to high-throughput operations. Additionally, since a large amount of experimental data must be archived, analyzed and shared among the various pipeline components the appropriate implementation of a data-management system is another key to the success of the pipeline.

The execution of the salvage pathways can begin at three different levels during high-throughput operations. The earliest can take place at the gene level where new recombinant constructs are made to increase expression or solubility. The second is at the protein-purification and preparation level. Finally, the third is to optimize conditions for proteins that crystallize but do not diffract. These applications constitute the tier 2 pathways designed to couple those of tier 1. The overall approach is schematized in Fig. 2. This report focuses on the latter two salvaging pathways. The 'target salvaging' effort is an important component of the pipeline. With comparatively little additional effort, many valuable 'failed' target proteins can eventually yield structures.

Target salvaging increases the cost-effectiveness of the structural genomics operations by reducing the number of targets that are abandoned after considerable initial efforts in the upsteam stages. At SECSG, the target-salvaging effort is carried out by the crystallomics core. In tier 1, the protein production core is focused on producing target proteins in rapid pace. The purified proteins go through the crystallization unit and the proteins that fail to produce useful crystals for structure determination enter tier 2 or 'target salvaging'. In addition to the preparation of samples with increased purity and homogeneity, the crystallomics core introduces chemical modifications to the proteins (such as tag removal, methylation, surface mutagenesis, selenomethionine labeling) as required.

The high-throughput operation in structural genomics is not a simple scale-up of a traditional crystallographic laboratory; it requires additional resources for planning, testing, fine-tuning and project management. Success in high throughput and automation relies on the harmonious integration of hardware and software at each stage and a seamless transition and effective bi-directional communication mechanisms between the various stages.

Work was supported in part with funds from the National Institute of General Medical Sciences (GM62407), The Georgia Research Alliance and The University of Georgia Research Foundation. Data were collected at Southeast Regional Collaborative Access Team (SER-CAT) 22-ID beamline at the Advanced Photon Source, Argonne National Laboratory. Supporting institutions may be found at http://www.ser-cat.org. Use of the IMCA-CAT beamline 17-ID at the Advanced Photon Source was supported by the companies of the Industrial Macromolecular Crystallography Association through a contract with Illinois Institute of Technology. Data were also collected at the Structural Biology Center, Advanced Photon Source. Use of the Advanced Photon Source was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences under contract No. W-31–109-ENG-38. International Business Machines is acknowledged for an equipment grant.

References

Adams, M. W., Dailey, H. A., DeLucas, L. J., Luo, M., Prestegard, J. H., Rose, J. P. & Wang, B.-C. (2003). *Acc. Chem. Res.* **36**, 191–198.

Arendall, W. B. III, Tempel, W., Richardson, J. S., Zhou, W., Wang, S., Davis, I. W., Liu, Z.-J., Rose, J. P., Carson, W. M., Luo, M., Richardson, D. C. & Wang, B.-C. (2005). In the press.

Baldock, P., Mills, V. & Stewart, P. S. (1996). *J. Cryst. Growth*, **168**, 170–174.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. & Eddy, S. R. (2004). *Nucleic Acids Res.* **32**, D138–D141.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

C. elegans Sequencing Consortium (1998). *Science*, **282**, 2012–2018.

Chen, L., Chen, L. R., Zhou, X. E., Wang, Y., Kahsai, M. A., Clark, A. T., Edmondson, S. P., Liu, Z. J., Rose, J. P., Wang, B.-C., Meehan, E. J. & Shriver, J. W. (2004). *J. Mol. Biol.* **341**, 73–91.

Cowtan, K. & Main, P. (1998). *Acta Cryst.* D**54**, 487–493.

Dauter, Z., Dauter, M. & Rajashankar, K. R. (2000). *Acta Cryst.* D**56**, 232–237.

Davis, I. W., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2004). *Nucleic Acids Res.* **32**, W615–W619.

Hope, H. (1988). *Acta Cryst.* B**44**, 22–26.

Kadima, W., McPherson, A., Dunn, M. F. & Jurnak, F. A. (1990). *Biophys. J.* **57**, 125–132.

Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.

Liu, Z. J., Vysotski, E. S., Chen, C. J., Rose, J. P., Lee, J. & Wang, B.-C. (2000). *Protein Sci.* **9**, 2085–2093.

Lovell, S. C., Davis, I. W., Arendall, W. B. III, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). *Proteins*, **50**, 437–450.

Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). *Proteins*, **40**, 389–408.

McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.

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

Navaza, J. (2001). *Acta Cryst.* D**57**, 1367–1372.

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

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

Rayment, I. (1997). *Methods Enzymol.* **276**, 171–179.

Richardson, J. S., Arendall, W. B. III & Richardson, D. C. (2003). *Methods Enzymol.* **374**, 385–412.

Robb, F. T., Maeder, D. L., Brown, J. R., DiRuggiero, J., Stump, M. D., Yeh, R. K., Weiss, R. B. & Dunn, D. M. (2001). *Methods Enzymol.* **330**, 134–157.

Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.

Shah, A. K., Liu, Z. J., Stewart, S., Schubot, P. D., Rose, J. P., Newton, M. G. & Wang, B.-C. (2005). *Acta Cryst.* D**61**, 123–129.

Storoni, L. C., McCoy, A. J. & Read, R. J. (2004). *Acta Cryst.* D**60**, 432–438.

Teng, T.-Y. (1990). *J. Appl. Cryst.* **23**, 387–391.

Terwilliger, T. C. (2003). *Methods Enzymol.* **374**, 22–37.

Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.

Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.

Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K., Richardson, J. S. & Richardson, D. C. (1999). *J. Mol. Biol.* **285**, 1711–1733.

Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999). *J. Mol. Biol.* **285**, 1735–1747.

Yang, H., Guranovic, V., Dutta, S., Feng, Z., Berman, H. M. & Westbrook, J. D. (2004). *Acta Cryst.* D**60**, 1833–1839.

Yang, C., Pflugrath, J. W., Courville, D. A., Stence, C. N. & Ferrara, J. D. (2003). *Acta Cryst.* D**59**, 1943–1957.