eResearch Solutions for High Throughput Structural Biology

Noel Faux[1,2], Anthony Beitz[3], Mark Bate[1], Abdullah A. Amin[1,2], Ian Atkinson[4], Colin Enticott[3], Khalid Mahmood[1,2], Matthew Swift[3], Andrew Treloar[3], David Abramson[3], James C. Whisstock[1,2], Ashley M. Buckle[1]

[1]The Department of Biochemistry and Molecular Biology
[2]The ARC Centre of Excellence in Structural and Functional Microbial Genomics
Faculty of Medicine,
[3]CSIT,
Monash University, Clayton, Victoria 3800, Australia
[4]High Performance Computing & School of Information Technology
James Cook University, Townsville, QLD, 4814, Australia
Email: Ashley.Buckle@med.monash.edu.au

Abstract

Structural biology research places significant demands upon computing and informatics infrastructure. Protein production, crystallization and X-ray data collection require solutions to data management, annotation, target tracking and remote experiment monitoring. Structure elucidation is computationally demanding and requires user-friendly interfaces to high-performance computing resources. Here we discuss how these challenges are being met at the Protein Crystallography Unit at Monash University. Specifically, we have developed informatics solutions for each stage in the structural biology pipeline, from DNA cloning through to protein structure determination. This infrastructure will be pivotal for accelerating the process of structural discovery and will be of significant interest to other laboratories worldwide.

1. Introduction

Proteins perform the functions necessary for life in all organisms. Protein function is to a large extent dictated by the 3-dimensional structure, and thus knowledge of the atomic structure of a protein is a prerequisite to understanding its function. The understanding of protein structure now has a firm role in the molecular basis of all diseases, and as such is a vital underpinning for the future promise of de novo drug design. X-ray crystallography is the most common technique for the structure elucidation of proteins. Briefly, this method involves first the production of large amounts of (usually recombinant) pure protein, followed by crystallization and X-ray diffraction analysis. The atomic structure is then calculated from the diffraction pattern using one of several methods. Each stage of this process is technically challenging and a potential bottleneck. Over the last 5 years adoption of automation technologies has eased the bottlenecks at the cloning, protein production and crystallization stages. Availability of synchrotron radiation has increased the rate at which high-quality diffraction data can be collected. Although the development of computational methods of structure elucidation has also undergone significant improvement, the high-throughput nature of the pipeline places an increasing emphasis on informatics and data management requirements for all stages in the process.

2. Protein Crystallography at Monash

Protein crystallography at Monash has grown considerably over the past 5 years (Figure 1). To date the unit includes six independent groups and over 100 researchers. In order to cope with the demand for crystallography, the unit has, where possible, deployed robotics to enhance throughput. There are several factors that will increase further this growth over the next 5 years: (1) Recent establishment of *ProteinExpress* in 2005, a High-Throughput protein

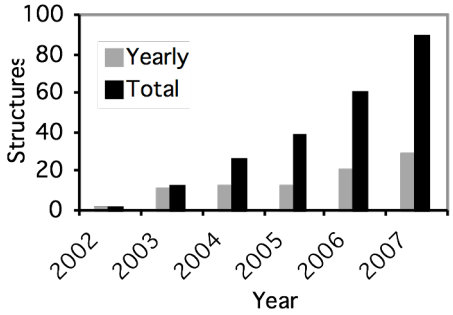

Figure 1. Protein structures produced at Monash

production facility at Monash, which automates and accelerates the production of quality-quality proteins; (2) the procurement of a more sophisticated and greater capacity (estimated 1 million experiments) high-throughput crystallization and imaging facility in 2007 will address the bottleneck of producing diffraction-quality crystals; (3) the Australian synchrotron, 500 meters from the protein crystallization unit, is now operational. Together, these three technologies will require radical advances in informatics and data management capabilities. In particular the current *ad-hoc* management of crystallographic software setup and execution will be inadequate to meet future research challenges. These technologies will converge in 2008 and therefore the necessary informatics infrastructure will need to be in place to take full advantage of the high-throughput approach.

2.1 The Structural biology pipeline

In low to medium throughput scenarios, individual targets enter the structural biology pipeline in a serial fashion, as a DNA clone (Figure 2). This typically does not present any informatics challenges. However, it is now commonplace for many targets (10s - 100s) to enter the pipeline at the same time and proceed towards structure determination in a parallel fashion. This creates an overall need for target tracking functionality, such that the progress of any individual target can be tracked easily at any stage in the pipeline. The many Structural Genomics Consortiums worldwide provide the ultimate example of this approach, where all representative proteins expressed in an organism are targeted (typically >20,000 for a eukaryote). Indeed these efforts have largely driven the biological and informatics technological advances.

Once the DNA clone has been obtained and purified, protein production can begin. This is typically performed using robotics and 96-well protein expression plates. Miniaturization and the ability to screen for properly folded proteins allow the rapid and easy identification of a small-scale protein expression system for each target. Often only a few milligrams of pure material are sufficient for successful crystallization. Expressed protein is then purified using standard chromatographic techniques, often requiring multiple attempts in parallel to determine the optimum protocol. Quality control measures employed to check purity often produce a large amount of digital image data, such as gel electrophoresis images and Mass Spectrometry

Pure protein samples then enter crystallization trials. As in protein production, this often involves the screening of several hundred conditions before suitable diffraction-quality crystals are obtained. This also requires automated crystal imaging technology and a high degree of data annotation by multiple users.

Crystals can then be subjected to X-ray analysis in a dedicated facility. Diffraction data is transferred to a client workstation and data processing and analysis begins. Structure determination can then take days to months, and is frequently disjointed due to the heterogeneity of hardware, software and data formats encountered. Dedicated hardware and software configurations required use either the UNIX command-line or X-windows based interfaces for a wide-range of necessary programs. Despite efforts to improve the user-friendliness of software the learning curve for novice structural biologists can be steep, particularly for researchers with a biological sciences background, rather than physical or mathematical sciences, as was more typical at the end of the last century. Thus, the informatics requirements have changed drastically over the past 10 years, and there is an unmet need for more user-friendly, integrated and highly automated approaches to protein structure determination.

Figure 2. Structural biology pipeline

Furthermore, the power of high-performance distributed computing is yet to be harnessed in mainstream protein crystallography in a user-friendly way.

This paper describes continuing efforts to develop and implement a highly integrated informatics infrastructure for the structural biology research pipeline at Monash. Implementation at each stage of the pipeline is described below.

2.2 AutoBLAST – Pre-processing targets

In the initial pre-processing and feasibility stages of a structural biology pipeline, sequence analysis of each target is typically performed using PSI-blast [1]. PSI-Blast is a widely used bioinformatics program that searches for distant relatives of a protein, based upon sequence comparisons. This is typically performed by submitting the target protein sequence to the NCBI blast website (www.ncbi.nlm.nih.gov/BLAST). With the first search iteration a list of all closely related proteins is created. This list is used to create a profile reflecting patterns of sequence conservation, which is then used as input to the second search iteration. This process is repeated until convergence.

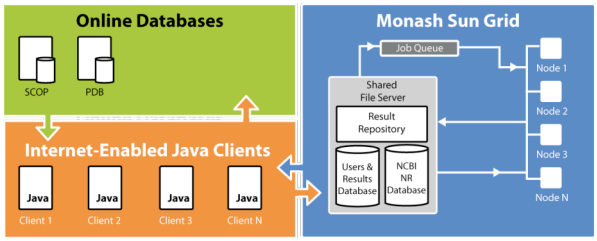

Figure 3. AutoBLAST architecture

Despite its power, the web-based blast has several shortcomings: (1) iterations are controlled in a manual fashion; (2) calculations can take a relatively long time (significant time savings can be achieved by distributing multiple PSI-Blast requests to multiple nodes of a computing cluster, since they are independent of each other); (3), it does not have sort, results filtering, or search-within functionality; (4) Results are not stored in a persistent database that ca be accessed at a later date.

Given these limitations and the need to perform multiple psi-blast calculations for multiple targets concurrently we have developed *AutoBLAST* (Al-Amin, Faux, Buckle and Whisstock, unpublished data) –a multi-layered distributed application that allows multiple users to submit psi-blast requests, distributes the load onto multiple compute nodes interconnected by a grid layer, and structures the resulting data into a MySQL relational database, which can then be analyzed and manipulated using an interactive Java Swing desktop application (Figure 3).

The *AutoBLAST* engine, a Java backend application residing on the head node, obtains PSI-Blast requests from the database and prioritizes the list based on urgency, then distributes the load onto multiple cluster nodes in a round-robin manner. Upon completion each node writes the result to an XML file on the networked file server. The presence of new XML files is checked periodically, which are then parsed and stored in the central database. Searches are organized using the *Blast Wish List Manager*, which allows targets to be auto-blasted as well as performing some preliminary sequence analyses. Search results are presented in tabular format that offers much post analysis, such as links to PDB files that can be also viewed using the implemented Jmol viewer. We have successfully deployed AutoBLAST over 32 nodes of the Monash Sun Grid cluster, offering significant improvements in both the throughput and quality of target sequence analysis.

2.3 High-throughput protein production pipeline

As a target progresses from DNA cloning through protein production, crystallization and structure elucidation and analysis, data is generated at each stage – this requires facile retrieval, analysis, annotation and archival. Whereas a standard laboratory notebook performs this role adequately for low-throughput scenarios, the pace and volume of a high-throughput laboratory demand simple yet effective means of organizing and managing large amounts of data. One of the most effective means of managing and organizing scientific research data is the use of a Laboratory Information Management System, or LIMS. These consist of graphical and text-based interfaces to databases that store laboratory data in a highly structured manner. There are currently over 100 commercially available LIMS packages for protein crystallization (http://www.limsource.com). However, there are far fewer freely available, or open source alternatives. Although these range from generic solutions to highly specific, customized databases and interfaces, each laboratory often requires a tailored system.

We have therefore developed a user-friendly graphical user interface to a relational database that allows users to manage the cloning, expression and purification of targets in the pipeline (Figure 4). Design Highlights include:

1. Centralized, Decoupled, Secure Database: to store and manage data generated in each step of the pipeline

2. User Access Control: to enable user to manipulate data based on hierarchical role.
3. Platform independent & Extensible Electronic Data Manager (EDM), to provide a streamlined platform to host plug-in applications to work on the central dataset.
4. Standard set of modular plug-ins: to decouple module interdependency
5. Plug-in Development Protocol (PDP): to enable third parties to develop plug-ins
and run in the pipeline environment.
6. Report generation and Data mining: to provide view only data to various report engines, and data mining agents in standard formats (such as XML, CSV, Text).
7. Interactive standalone and web Graphical User Interface, to access the pipeline and analyze data through graphical visualization aids.

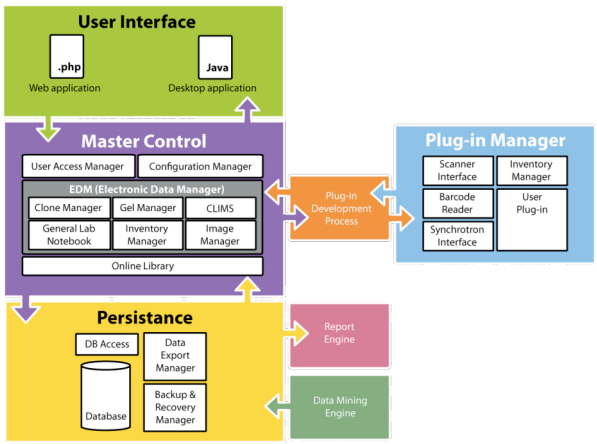

Figure 4. Architecture of target tracking laboratory information management system (LIMS)

2.4 pxPortal – Remote monitoring of x-ray data collection

X-ray data collection experiments require constant monitoring to ensure the experiment proceeds without interruption and adverse changes in conditions. The *Dataset Acquisition, Accessibility, and Annotation e-Research Technologies (DART: http://dart.edu.au) Project* has developed a software toolkit that facilitates the construction of collaborative e-research infrastructure. Using the DART toolkit we have developed *pxPortal*, a web portal for protein crystallography that allows remote monitoring of the X-ray data collection experiment as well as management of resulting diffraction data.

Monitoring the room temperature and humidity allow the user to monitor important physical conditions in the experiment enclosure that may have detrimental effects on the experiment if left unchecked. Both parameters are particularly important when collecting data at cryo-temperatures. High humidity may cause accumulation of ice crystals on the loop housing the crystal that may lead to the observation of ice diffraction rings in the diffraction data that can ruin a structure determination. In such cases, the ability to monitor the ice build-up via both live video feed of the loop and/or direct inspection of the current diffraction image, both which are important features of the portal, can prove useful. Data from the X-ray diffractometer is captured on the local instrument control computer, and is available for processing over the local network.

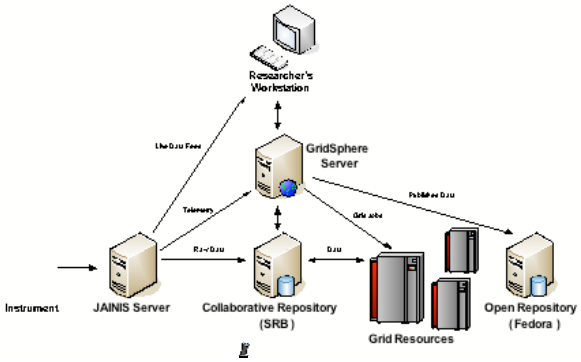

Figure 5. DART *pxPortal* system architecture

The ability to access diffraction images is critical because of the need to check the data processing statistics (especially data quality) as the experiment proceeds. Time-allocated use of the X-ray equipment puts pressures on users to perform the experiment as quickly as possible whilst ensuring the dataset is near 100% complete and of acceptable quality. Pseudo-symmetry or ambiguities in spacegroup sometimes cannot be resolved until the later stages of structure elucidation, and thus the data processing outcome has a direct effect on the precise time the experiment can be safely terminated. Further, protein crystals are particularly prone to radiation damage, and although the time needed to collect a near-100% dataset can be calculated easily from the symmetry apparent in the first image, real-time inspection of the images may prompt the user to terminate the experiment and thus prevent significant wasted time collecting useless data. Data collected during the experiments are recognized as very valuable, both academically and commercially. Even the hint from a filename, as to the nature of the work, may provide an IP risk. Therefore, the collected data and its associated metadata need to be transferred and stored in a secure manner.

The web portal was built in Java utilizing GridSphere's portlet containers (JSR-168 compliant), and was supported by a number of services, as shown in Figure. 5. Salient features of the architecture can be summarized as follows. The JAINIS server uses the Common Instrument Middleware Architecture (CIMA) [2] to create a proxy access to an instrument's collected data and its associated telemetry, transports this data to a workflow engine, where data can be reduced, meta data added and then moves the results into a Collaborative Repository.

The Collaborative Repository stores the data collected from the instrument as well downstream results processed from the analysis tools. This was implemented using Storage Resource Broker (SRB), as this technology is capable of managing very large data sets and data federations. File information and metadata (associated with the experiment) is stored in SRB's metadata catalog (MCAT). Objects in the Collaborative Repository can be made more accessible to the research community by publishing them to *Fedora*, open source digital repository software, featuring rich support for metadata.

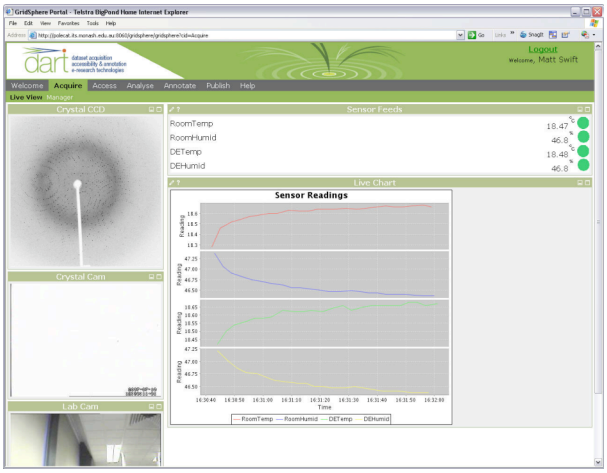

Figure 6. pxPortal web interface

The web portal consists of several well-defined areas that contain useful functionality for the diffraction experiment. Currently, these consist of a portlet that provides a low resolution version of the diffraction pattern from the diffractometer, a live-video feed of the crystal, a live-video feed of the lab, and various temperature and humidity readings (Figure 6). In addition, browsing, image previewing and metadata search capabilities for the Collaborative Repository are provided by James Cook University's Portlet Grid Library (https://plone.jcu.edu.au/hpc/staff/projects/hpc-software/personal-grid-library). Finally, initial proof-of-concept data processing functionality is provided by allowing data reduction using the program XIA2 [3].

In summary, the DART toolkit has been used to build a protein crystallography portal that allows the remote management of a X-ray data collection experiments, as well as file transfer and visualization. In order to realize the full potential of a simple portal-based approach and to hide much of the computational complexity, we have begun to integrate many of the common crystallographic tasks into the portal. As such the user will be able to perform the majority of the data processing, structure solution and analysis tasks via the portal itself. We envisage that the user-friendliness and simplicity of its interface will be of significant interest to many structural biology laboratories worldwide.

2.5 pxGrid - High-throughput structure determination

The most common method of protein structure determination is Molecular Replacement (MR). This involves using the structure of a protein that shares significant sequence similarity with the protein of unknown structure as a starting point in the structure determination (otherwise known as solving the *phase problem*). The process generally involves three steps: (1) Using sequence-searching methods such as PSI-BLAST to identify suitable structures that can be used for molecular replacement; (2) modification of structures (e.g., removal of flexible loop regions and non-identical side chains), to yield *search models*; (3) Finding the orientation and position of the search model in the unit cell of the target crystal; (4) Refinement of the model using iterative model-building and maximum likelihood atomic refinement. Although there are other methods of structure determination, molecular replacement is predicted to become an increasingly common technique, for two reasons. Firstly, the probability that the unknown target structure belongs to a known fold is steadily increasing, due to the exponential growth of the Protein Database (PDB) [4]. Secondly, the emergence of more sophisticated sequence searching algorithms, such as profile-profile matching [5], improve the probability of finding a suitable search model, even in cases of very low similarity (<20% identity).

In this final part of the structural biology pipeline we are developing intelligently guided brute force approaches to identify candidate models for structure determination in the event that no obvious search model (based on sequence similarity) is available. Below we summarize the background, aims and preliminary data for this project.

Conventional protein crystallography dictates that when sequence similarity with known structures cannot be detected, other methods of structure solution must be employed, for example multiple isomorphous replacement (MIR) or multiwavelength anomalous dispersion (MAD). Both techniques are powerful but bring their own requirements, costs, and difficulties. For example MIR requires trial-and-error heavy atom soaking experiments, and MAD requires a selenomethionine-containing protein (which can be challenging to generate, for example, if recombinant material is produced using eukaryotic expression systems), and synchrotron radiation.

A key problem in bioinformatics is that structural similarity can be retained long after detectable sequence similarity is lost [6] (the so called midnight zone). Thus it is common for similarity between a protein of unknown structure and a "known fold" to become apparent only after structure determination. For such proteins, an MR-based approach may have been achievable, however, the inability to detect the fold or family by sequence matching methods restricts its application.

MR calculations are embarrassingly parallel: An approach to address the scenario of no detectable sequence similarity is to attempt brute force molecular replacement experiments using every single structure in the PDB (>3000 families). Up until recently, the computational resources required for such an approach would be prohibitive. However, the exponential growth of computing power and recent advances in harnessing this power in a massively parallel fashion, using grid computing, means this approach is now feasible. This report describes the methods we are developing in order to apply Grid computing to these previously intractable cases.

To date, the structures of approximately 1000 different 3D folds have been described [4], from ~3000 families. Further, structural genomics programs have launched targeted attempts in order to provide the biological community with representatives of all folds [7; 8], estimated at ~1700 [9]. We are developing a "brute force" molecular replacement approach using all known folds, which does not rely on sequence similarity. Using the SCOP database [10], we have developed a library consisting of ~3000 MR search models derived from the representative highest resolution structure of each SCOP family. In proof-of-concept experiments, we have developed a resource where each family representative is used as a search model in a PHASER [11] MR calculation. The PHASER calculation is the main bottleneck in the MR process, typically requiring 1-100 hours on a high-end workstation, per calculation. In order to perform >3000 PHASER calculations in a timeframe of days rather than years we have developed a highly parallel approach using computational facilities at VPAC (*Brecca* – 97 dual Xeon 2.8 GHz CPUs, 160 GB (2 GB per node) total memory; *Edda* – 185 Power5 CPUs, 552 GB (8-16 GB per node) total memory) and Monash University ITS Sun Grid (54 dual 2.3 GHz CPUs, 208.7 GB (3.8 GB per node) total memory). The PostgreSQL database (www.postgresql.org) system is used to store and manage the MR jobs and results. PERL scripts are used to farm out MR jobs to free CPU's, launch the MR programs and collect the results. The web front end is written using PHP software (www.php.net) and served using Apache server software (www.apache.org). This is represented schematically in Figure 7.

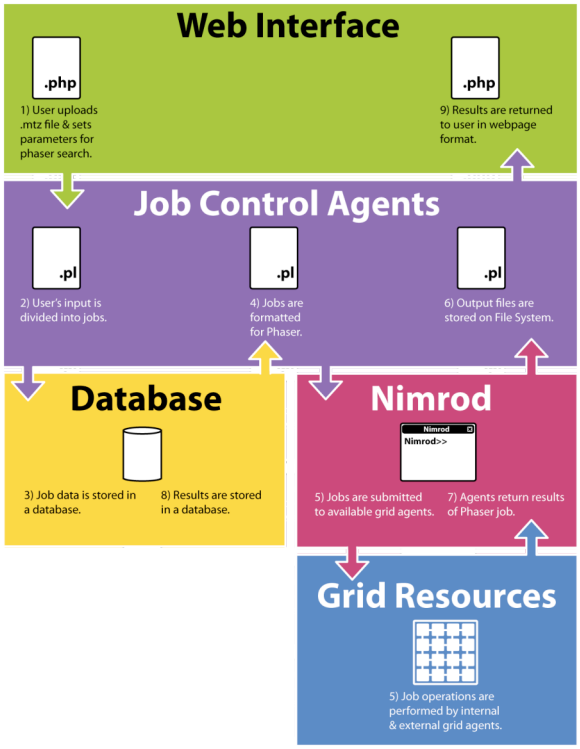

Figure 7. pxGrid architecture

In order to expand this approach to all protein domains in the PDB (~80, 000) we require an order of magnitude increase in computing nodes. This is being achieved using the software tool *Nimrod/G [12],* which distributes individual jobs over the PRAGMA testbed. As such, the availability of ~1000 nodes makes the scale of this task practical in a timeframe of days and at most, weeks. Specifically, each individual PHASER job consists of the csh script (describing the PHASER job), reflection file, search PDB, and any PDB file that will be fixed during the run. These files are copied

over to the resource and the csh script is then run / submitted on the allocated resource. Upon completion of the job, the results files are copied back to the submission machine and the initial copied files are removed.

In summary, we are developing a major new tool to solve the three-dimensional structures of proteins in a significantly shorter timeframe than is currently possible. The ability to perform MR calculations using an exhaustive set of search models will offer a timesaving of weeks to months in a typical successful structure determination. Challenging structure determinations by MR currently can take more than 6 months, therefore it is extremely useful to know as quickly as possible when the MR approach might fail, and thus when to pursue alternative methods.

3. Conclusions

High-throughput structural biology requires significant informatics input at many stages of the pipeline. Many tasks are potential bottlenecks and present challenges for data management and target tracking. We are developing software solutions to many of these stages, which we are using in our daily research. Ultimately we aim to integrate these approaches into a single web-based resource providing a consistent, simple informatics solution that will have a positive impact on the long-term structural biology research programs at Monash.

4. Acknowledgements

We thank the NHMRC, ARC, Victorian Partnership for Advanced Computing, and the Victorian Bioinformatics Consortium for funding and support. JCW is an National Health and Medical Research Council of Australia Principle Research Fellow and Monash University Senior Logan Fellow. SPB and AMB are NHMRC Senior Research Fellows.

5. References

[1] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25 (1997) 3389-402.

[2] D.d.B. I.M. Atkinson, C. Chee, K. Chiu, T. King, D.F. McMullen, R. Quilici, N.G.D. Sim, P. Turner, M. Wyatt, CIMA Based Remote Instrument and Data Access: An Extension into the Australian e-Science Environment. , Proc. IEEE 2nd Int. Conf. on e-Science and Grid Computing Amsterdam, The Netherlands. , 2006.

[3] CCP4, The CCP4 suite: programs for protein crystallography. Acta Crystallogr D50 (1994) 760-763.

[4] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, The Protein Data Bank. Nucleic Acids Res 28 (2000) 235-42.

[5] L. Jaroszewski, L. Rychlewski, Z. Li, W. Li, and A. Godzik, FFAS03: a server for profile--profile sequence alignments. Nucleic Acids Res 33 (2005) W284-8.

[6] J.C. Whisstock, and A.M. Lesk, Prediction of protein function from protein sequence and structure. Q Rev Biophys 36 (2003) 307-40.

[7] M. Gerstein, A. Edwards, C.H. Arrowsmith, and G.T. Montelione, Structural genomics: current progress. Science 299 (2003) 1663.

[8] A. Yee, K. Pardee, D. Christendat, A. Savchenko, A.M. Edwards, and C.H. Arrowsmith, Structural proteomics: toward high-throughput structural biology as a tool in functional genomics. Acc Chem Res 36 (2003) 183-9.

[9] R.I. Sadreyev, and N.V. Grishin, Exploring dynamics of protein structure determination and homology-based prediction to estimate the number of superfamilies and folds. BMC Struct Biol 6 (2006) 6.

[10] A. Andreeva, D. Howorth, S.E. Brenner, T.J. Hubbard, C. Chothia, and A.G. Murzin, SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32 Database issue (2004) D226-9.

[11] A.J. McCoy, R.W. Grosse-Kunstleve, L.C. Storoni, and R.J. Read, Likelihood-enhanced fast translation functions. Acta Crystallogr D Biol Crystallogr 61 (2005) 458-64.

[12] D. Abramson, Giddy, J. and Kotler, L., High Performance Parametric Modeling with Nimrod/G: Killer Application for the Global Grid?, International Parallel and Distributed Processing Symposium (IPDPS), Cancun, Mexico, 2000.