Structural Genomics at the National Synchrotron Light Source

K.R. Rajashankar[6], M.R. Chance[1,2,3], S.K. Burley[6,5], J. Jiang[7], S.C. Almo[1,3], A. Bresnick[3], T. Dodatko[1,3], R. Huang[1,2], G. He[6,5], H. Chen[6,5], M. Sullivan[1,2], J. Toomey[1,2], RA. Thirumuruhan[1,2], W.A. Franklin[4], A. Sali[6], U. Pieper[6], N. Eswar[6], V. Ilyin[6], and L. McMahan[6]

[1]Center for Synchrotron Biosciences, [2]Department of Physiology & Biophysics, [3]Department of Biochemistry, [4]Department of Radiology, Albert Einstein College of Medicine, Bronx, NY
[5]Howard Hughes Medical Institute, [6]The Rockefeller University, New York, NY
[7]Department of Biology, Brookhaven National Laboratory, Upton, NY

Progress in understanding the organization and sequences of genes in model organisms and humans is rapidly accelerating. Availability of genome sequences from several organisms (Green et al., 2001) has prompted a scientific inquiry to understand the structure and function of all genes, including the pathways leading to the organization and biochemical function of macromolecular assemblies, organelles, cells, organs, and whole life forms. At this juncture, structural biologists have embraced high-throughput biology by developing and implementing technologies that will enable the structures of hundreds of protein domains to be solved in a relatively short time. Although thousands of structures are deposited annually in the Protein Data Bank, the vast majority are identical or very similar in sequence to a structure previously existing in the databank (Brenner et al., 1997). Providing structural information for a broader range of sequences requires a focused effort on determining structure for sequences that are divergent from those already in the database. Although structure does not always elucidate function, it makes functional annotations possible that are not recognizable at the sequence level (Burley et al., 1999). The ultimate goal of the structural genomics projects recently funded by the National Institutes of Health is to determine the structure of at least one member of each family so that most proteins can be characterized structurally. Recent estimates indicate that with a 30% sequence identity cutoff a minimum of 16,000 targets have to be solved so as to cover 90% of all protein domain families, including those of membrane proteins (Vitkup et al., 2001). Structure of the members of the family that are not determined directly will be modeled with useful accuracy (Sali, 1998, Burley et al., 1999, Shi et al., 2002)

New York Structural Genomics Research Consortium

The New York Structural Genomics Research Consortium (nysgrc.org), one of the nine NIH funded centers (www.nigms.nih.gov/funding/psi.html), is a cooperative effort between the Albert Einstein College of Medicine, Rockefeller University, Brookhaven National Laboratories, Weill-Cornell Medical College, and Mount Sinai School of Medicine. The goal of the NYSGRC is to develop and implement high-throughput technology to identify, obtain and model protein structures. NSLS beamlines X9A, X9B, X12B, X12C and X25 are major experimental resources for the NYSGRC. At this stage of the program we are identifying bottlenecks in our structure discovery pipeline, and developing or implementing technologies to remove these bottlenecks so as to increase the rate of structure determination. The structure determination process includes target selection; cloning the coding sequence; protein expression & purification; biophysical characterization of the expressed protein; defining and refining crystallization conditions and incorporating suitable heavy atoms; collecting multiple wavelength anomalous dispersion (MAD) data at an X-ray synchrotron beamline; determining the phases of the reflections, building the model and refining the structure; making functional inferences from the structure; modeling unknown open reading frames based on the identification of new sequence-structure relationships and disseminating our findings (Burley et al., 1999, Chance et al., 2002).

In this article, we recount recent progress of the NYSGRC in some of the pipeline steps identified above and briefly report the consortium's 27 new structures solved during its first-year's effort.

Target selection for structural genomics

Structural genomics aims to structurally characterize most protein sequences by an efficient combination of experiment and modeling (Burley et al., 1999, Sali, 1998, Vitkup et al., 2001). Central to the success of these efforts is effective target selection. There are a variety of target selection schemes, ranging from focusing on only novel folds to selecting all proteins in a model genome (Brenner, 2000). NYSGRC has multiple target selection strategies one of which involves targeting enzymes. We have prepared a conservative list of protein families that contain human enzymes of unknown structure. First, all sequences with an annotated enzyme classification (EC) number were extracted from the TrEMBL database (9/1/01), resulting in 19,382 presumptive enzymes from a wide variety of organisms. This list included human enzymes in 204 classes with unique EC numbers. For each of the 204 representative human enzymes, homologs from 10 other organisms with more than 30% sequence identity were identified from the PSI-BLAST profiles in ModBase (guitar.rockefeller.edu/modbase). The resulting 204 families contain 903 enzymes, all of which were deposited into a web-based target tracking system, which is our on-line database and lab notebook. The final target list was further refined to (i) limit sequence length to 500 residues, (ii) avoid those sequences containing any predicted trans-membrane spanning regions, (iii) eliminate outliers in a sequence alignment of the families and (iv) avoid enzymes that can be related to a protein of known structure with a PSI-BLAST E-value greater than 10^{-4} over any segment in their sequences. Based on these criteria, the current list of selected targets contains approximately 300 enzymes from approximately 100 EC classes, with each class containing a single human sequence and multiple homologs from the selected organisms.

Production and Testing of Expression Clones

A key feature of the Structural Genomics pipeline is the automated production of expression vectors from identified targets. Within the NYSGRC, we have established a Centralized Cloning Facility that is responsible for operating an automated platform for all of the molecular biological steps required to subclone open reading frames from genomic DNAs and/or cDNA libraries, insert these coding sequences into expression vectors, transform *E. coli*, and test the resulting expression strains for production of soluble protein. A Beckman Biomek FX Robotic Platform has been programmed to perform all of the steps with bar code tracking of sample and reagent plates. Small-scale purification of recombinant proteins is then performed with Millipore Metal Chelating ZipTips, loaded with Ni^{2+} ions. The resulting purified recombinant proteins can be spotted onto a matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) sample plate.

Our initial experiences with this technology platform are extremely encouraging. The first cloning candidate (NYSGRC ID T136) subjected to the entire process yielded purified recombinant protein with a measured molecular mass within 12 mass units of the predicted mass of 22,845.8 (observed mass=22,834.4). After cleavage with polio viral protease, we obtained a measured molecular mass within 73 mass units of the predicted mass of 17,990.5 (observed mass=17,917.6). Thus, we can rapidly confirm soluble expression of the desired protein with the correct molecular mass and removal of the His_6 purification tag without DNA sequencing of the expression plasmid. Expression strains meeting these criteria are transferred to one of the five decentralized protein production/crystallization teams located at each of the five participating institutions. At present, one molecular biologist working with the robotics technician can work through the entire process of cloning-transformation-insert testing-retransformation-expression testing in less than two weeks. Preliminary results showed a 90% success rate in producing BL21(DE3) Star cells transformed with the desired expression vector using the above enzyme targets as input to the cloning strategy. Solubility tests demonstrated that 75% of these robotically-engineered expression clones yielded soluble proteins with correct apparent masses (as judged by gel electrophoresis). In the long term, small-scale, automated biophysical characterization will include domain mapping by limited proteolysis and MALDI-MS, and detection of non-specific aggregates using dynamic light scattering. The results of these studies will be used to guide target redesign within the NYSGRC.

High throughput identification of metals in structural genomics targets (metallomics)

Metal containing proteins are common and many of these metals have X-ray absorption K-edges that can be accessed using synchrotron sources. Knowledge about the presence of such metals may speed up the process of structure determination, as it may bypass the bottleneck of producing heavy atom derivatives or preparing seleno-methionyl substituted protein. From the lessons learned in the process of determining the structure of a uracil-DNA glycosylase from *T.Maritima* (TMUDG, Sandigursky et al. 2001), one of the targets of NYSGRC, we have developed a high throughput procedure to examine the presence of metals in proteins using x-ray absorption. This requires microgram quantities of protein and hundreds of samples can be examined in a high throughput manner. In this method a dried powder of protein sample is exposed to high energy synchrotron radiation (eg.

13500 eV, just above Br K-edge) to excite the energy levels of frequent metals in proteins (eg. Fe, Ni, Co, Zn etc.). Fluorescence emission from these metals is detected using a solid-state detector and the metal signals can be individually discriminated. This procedure provided valuable insights for determining the structure of TMUDG. This protein was found to aggregate within an hour of expression/purification, requiring setting up of crystal trays soon after purification. A 2.8Å native data collected at 9881 eV indicated the presence of weak anomalous signal. At this stage the 'metallomics' experiment was implemented, taking TMUDG as a first test sample. The fluorescence spectrum unambiguously showed the presence of Fe in the protein. On examining the electron density maps, it was recognized that the anomalous scatterer is an Iron-Sulfur cluster covalently bound to four cysteine residues (Figure 1). If this knowledge had been available at early stages of the structure analysis, it would have aided the structure solution significantly. Now, each of crystallized NYSGRC targets is screened through this 'metallomics' pipeline. This is in accordance with the dictum of the NYSGRC – developing and/or implementing technologies to remove bottlenecks so as to increase the rate of structure determination.

Automated Structure Determination Platform

An Automated Structure Determination Platform (asdp.bnl.gov) has been established by the NYSGRC for high throughput structure determination. Various publicly accessible software packages are organized as a production pipeline that provides a highly efficient computational environment with a web-based interface, spanning all steps of structure determination from data collection to PDB submission. The platform is implemented on a substantial and expandable computing server. The registered users are provided a web-based home data directory with disk space. X-ray data collected at synchrotron beam lines can be linked directly to a user's web-home within ASDP. After setting up necessary experimental data files, ASDP automatically performs appropriate data conversions required to use the various crystallographic programs, and writes script files in the user's directory for running programs. Users are presented with a list of steps/programs, each with suggested input/script files for submitting jobs. The jobs run on the large computing cluster (Linux or SGI) available to the ASDP user community and the progress of each structure determination is archived in an internal database.

ADSP was tested in the structure determination of on of the NYSGRC target P097, a hypothetical yeast

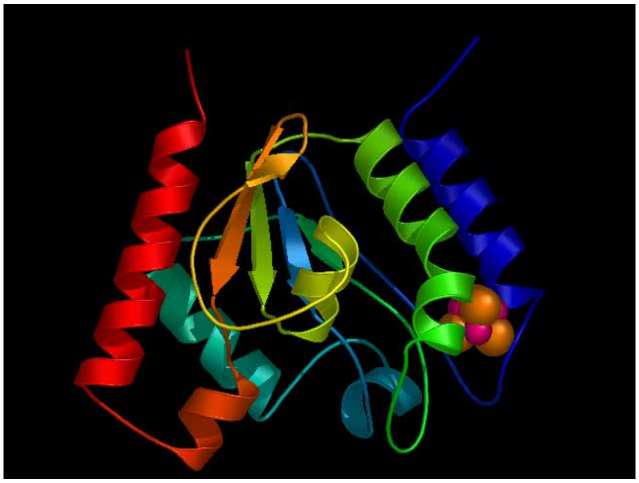

Figure 1. *Structure of Uracil-DNA glycosylase from* T. maritima*, blue to red ramp colored from N to C terminus. In CPK representation is the Iron-Sulfur cluster covalently bound to four cysteine residues.*

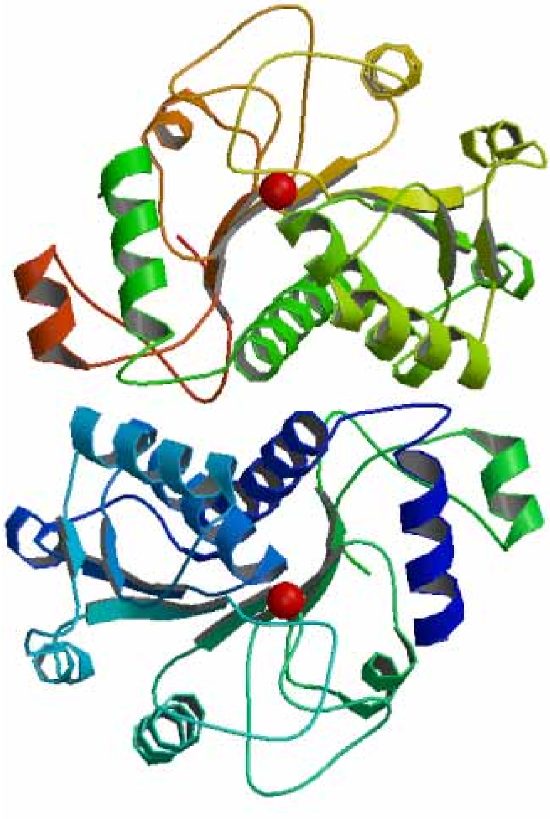

Figure 2. *P097 forms a tightly packed dimer a three-layer α-ß-α sandwich. ß strands are painted as cyan, α helices are painted as red and loops are painted as gray. Large red spheres represent putative metal ions.*

protein (YNL200C) and a member of a large protein family (23 sequences currently represented in ProDom - domain PD005835) with unknown biological function. The initial structure was solved by ASDP within two hours after completing a three wavelength seleno-methionine MAD experiment. The P097 structure revealed a three layer α-β-α sandwich, with two molecules forming a tightly packed dimer (Figure 2). A search using the DALI (Holm & Sander, 1993) server showed that P097 is similar to the non-catalytic domain of D-glycerate dehydrogenase (1GDH), with a Z-score of 8.5, a sequence identity of 10% and RMSD of 4.0 Å for C_a atoms.

A Summary of structures determined during the first year

As of August 31, 2001, the NYSGRC completed 27 X-ray structures that resulted from the examination of more than 500 independent constructs expressed in *E. coli*. Comparative protein structure modeling with these 27 experimentally determined structures produced additional structural information for thousands of protein sequences. These models are publicly available via ModBase (nysgrc.org). Of our 27 structures (Table 1) over half were distantly or entirely unrelated to known structures or folds. A striking feature is that our solved structures are derived from working with recombinant proteins from all three phyla (*Eukarya*, *Archaea*, and *Eubacteria)* and not narrowly distributed evolutionarily. Also, the targets were of significant size, with an average length of 280 residues (Table 1). With respect to crystal types and diffraction, large asymmetric units and long unit cell dimensions (including a very challenging case at circa 510Å) and lower symmetry crystal systems (16/27 in monoclinic or orthorhombic) were overcome and the average resolution limit was 2.3 Å. Although two-thirds of the structures were solved using selenium-MAD, other phasing methods were also critical to productivity.

Conclusion

The progress of the NYSGRC after its first year of funding from the NIH indicates that cautious optimism about the overall progress of the Structural Genomics initiatives is warranted. Although significant progress is evident in this report, bottlenecks remain in the structural genomics pipeline, particularly in generating sufficient homogeneous protein from a wide array of targets for crystallization trials. Also, as greater numbers of structures are produced, providing adequate annotation for these structures will become a significant bottleneck. However, this activity cannot be neglected, as it provides the core of the structural genomics effort. In addition, an increasing need for direct functional analysis by biochemical methodologies, not just annotation, is expected to arise, especially for proteins whose function is inferred from structure and where direct experimental tests can be easily imagined.

Lastly, some comments are appropriate regarding the impact of structural genomics initiatives on the practice of structural biology in individual laboratories. It is important that structural genomics efforts do not become the only source of structure determination for small, single domain proteins. In many cases our efforts will not acquire the high resolution data required for understanding chemical mechanism. Moreover, examination of mutant proteins and substrate or inhibitor complexes is critical in the evaluation of biological and chemical function and such studies are not part of the mandate of the NIH-funded centers. However, it is expected that there will be some impact, necessitating a focus shift for some laboratories to “hard” problems, e.g. proteins that are difficult to express or structures of multi-component complexes. The goal of structural genomics is to provide structural models for the biologist thus permitting improved functional annotation of proteins involved in a wide array of biochemical and cellular processes. This should bring more biologists into the structural “fold” and promote interest in in-depth structural studies for molecules of biological interest.

Table 1: Summary information for the first 27 proteins crystallized and solved by the NYSGRC.

Organisms	Methods of Structure Determination	Crystal Systems	Overall average
S. cerevisiae: 12 Eubacteria: 8 Archaea: 3 Human or Mouse: 4	MAD/Se: 18 MAD/MIR Other Elements: 1 SIRAS: 4 (Pt or Hg) MIR: 1 MR from NYSGRC structure: 2 Part of Larger Complex: 1	Monoclinic: 5 Orthorhombic: 11 Tetragonal: 5 Trigonal or Hexagonal: 6	Protein size: 280 residues Resolution: 2.3Å Unit Cell Dimension 86Å Number of Residues per Asymmetric Unit: 700 Number of Protomers per Asymmetric Unit: 2.5

Acknowledgements

The P097 X-ray diffraction data was collected using X12C beamline at NSLS and support of its staff is acknowledged. Please contact J. Jiang for further information on P097. The TUMDG structure was solved on beamline X9A, please contact K. Rajashankar for further information on this structure. This work was supported in part by grants from NIH-GM54762 (AS), the Mathers Foundation (AS), a Merck Genome Research Award (AS), and the NIH structural genomics pilot center grant GM62529. Also, support from the Biomedical Technology Resource programs located at the NSLS and funding by the National Center for Research Resources (P41-RR01633 and P41-RR12408) are acknowledged. The National Synchrotron Light Source, Brookhaven National Laboratory, is supported by the U.S. Department of Energy, Division of Materials Sciences and Division of Chemical Sciences, under Contract No. DE-AC02-98CH10886. A. Sali and S. Almo are Irma T. Hirschl Trust Career Scientists.

References

S.E. Brenner, “Target Selection for Structural Genomics,” Nat. Struct. Biol., **7**, Suppl: 967, 2000.

S.E. Brenner, C. Chothia and T. Hubbard, “Population statistics of protein structures:lesson from structural classifications”, Curr. Opin. Stru. Biol. **7**, *369*, 1997.

S.K. Burley, S.C. Almo, J.B. Bonanno, M. Capel, M.R. Chance, T. Gaasterland, D. Lin, A. Sali, F.W. Studier and S. Swaminathan, “Structural genomics: beyond the Human Genome Project”, Nat. Genet., **23**, *151,* 1999.

M.R. Chance, A.R. Bresnick, S.K. Burley, J. Jiang, C.D. Lima, A. Sali, S.C. Almo, J.B. Bonanno, J.A. Buglino, S. Boulton, H. Chen, N. Eswar, G. He, R. Huang, V. Ilyin, L. McMahan, U. Pieper, S. Ray, M. Vidal, L. K. Wang, “High throughput structural biology: A pipeline for providing structures for the biologist” Protein Science, **in press**, 2002.

E.D. Green, “Strategies for the systematic sequencing of complex genomes”, Nat Rev Genet, **2**, *573*, 2001.

L. Holm and C. Sander, “Protein Structure Comparison by Alignment of Distant Matrices,” J. Mol. Biol., **233**, 123, 1993.

M.A. Marti-Renom, A.C. Stuart, A. Fiser, R. Sanchez, F. Melo, A. Sali, “Comparative protein structure modeling of genes and genomes”, Annu Rev Biophys Biomol Struct., **29**, *291*, 2000.

A. Sali, “100,000 protein structures for the biologist”, Nat Struct Biol., **5**, *1029*, 1998.

M. Sandigursky, A. Faje and W.A. Franklin, “Characterization of the full length uracil-DNA glycosylase in the extreme thermophile thermotoga maritima,” Mutation Res. **485**, *187*, 2001.

W. Shi, D. Ostrov, S. Gerchman, H. Kycia, W. Studier, W. Edstrom, A. Bresnick, J. Ehrlich, J. Blanchard, S.C. Almo and M.R. Chance, “High-Throughput Structural Biology and Proteomics” in Proteomics: The Next Phase of Genomics Discovery, Marcel Dekker, Pubs. **in press**, 2002.

D. Vitkup, E. Melamud, J. Moult and C. Sander, “Completeness in structural genomics”, Nat Struct Biol., **8**, *559*, 2001.