A HIGH-THROUGHPUT APPROACH TO PROTEIN STRUCTURE ANALYSIS

Babu A. Manjasetty[1,2], Wuxian Shi[1,2], Chenyang Zhan[3], András Fiser[3] and Mark R. Chance[1,2,3*]

[1]New York Structural GenomiX Research Consortium (NYSGXRC)
Center for Synchrotron Biosciences
National Synchrotron Light Source
Brookhaven National Laboratory
Upton NY 11973
[2]Case Center for Proteomics
Case Western Reserve University
10900, Euclid Avenue
Cleveland, OH 44106
[3]New York Structural GenomiX Research Consortium (NYSGXRC)
Department of Biochemistry
Albert Einstein College of Medicine
Bronx, NY 10461

*To whom correspondence should be addressed.

Genetic Engineering, Volume 28, Edited by J. K. Setlow

INTRODUCTION

Structural Genomics was defined at the 2nd International Structural Genomics conference in 2001 as, "A large-scale project to determine the three-dimensional shapes of all proteins and other important biological molecules encoded by the genomes of key organisms". The structural genomics projects aim at the discovery, analysis and dissemination of three-dimensional structures of all proteins and other biological macromolecules in the universe of protein folds (Figure 1). The major structural genomics initiatives around the world are listed in Table 1.

The Protein Structure Initiative (PSI), which comprise the major efforts in structural genomics in the United States, has established centers for the project that have achieved automation of all the steps involved in determining protein structures, including target selection, cloning, expression, purification, biophysical characterization, crystallization, data collection, structure solution, refinement, validation and functional annotation (Figure 2). In addition, international coordination was put in place among the different centers in the U.S.A. and worldwide to avoid duplication of efforts and waste of resources (http://targetdb.pdb.org/).

Among the many structural genomics research projects around the world, in the U.S.A., the National Institutes of General Medical Science (NIGMS) of the National Institutes of Health (NIH) sponsored nine pilot structural genomics centers through the first phase of PSI (1). During this pilot phase (PSI1), these centers have established infrastructure for high-throughput production of protein

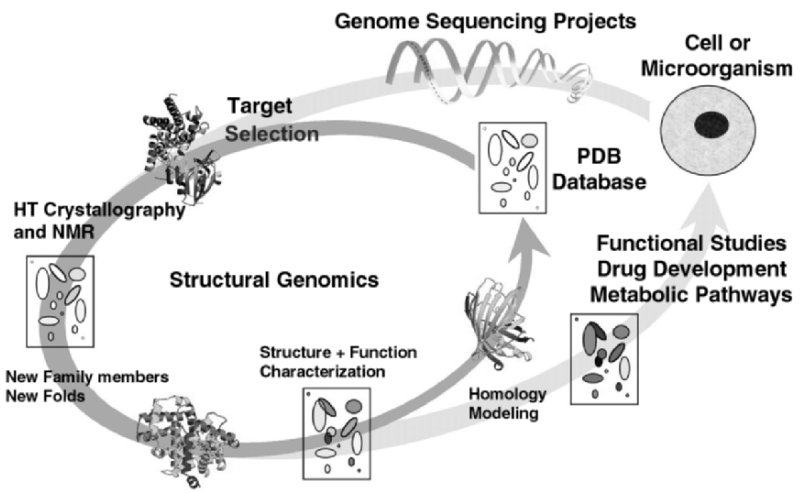

Figure 1. An example representing global structural genomics efforts for completing the protein family and fold landscape. The rectangular panels represent our current knowledge of the set of protein sequence families, showing whether they contain any 3D structural examples (black encircled regions) or not (white encircled regions). The amount of black increases as more structures are determined experimentally. Only a small fraction of the protein families may not contain a known 3D structure (small circles), but the majority of the fold landscape will be represented, permitting homology modeling of most of the remaining and new gene sequences. Diagram taken from Stevens, R.C., Yokoyama, S., Wilson, I.A. (2001) Global efforts in structural genomics, Science, 294, 89-92.

Table 1. Major structural Genomics Centers around the world.

Country	Initiative	Web Address
Japan	RIKEN Structural Genomics Initiative (RSGI)	http://www.riken.go.jp/engn/index.html
England	Structural Proteomics in Europe (SPINE)	http://www.spineurope.org/page.php?page=home
	Oxford Protein Production Facility (OPPF)	http://www.oppf.ox.ac.uk/index.php?module=ContentExpress&func=display&ceid=1&meid=-1
U.S.A.	NIGMS Protein Structure Initiative (PSI)	http://www.nigms.nih.gov/Initiatives/PSI
Canada	Montreal-Kingston Bacterial Genomics Initiative (BSGI)	http://euler.bri.nrc.ca/brimsg/bsgi.html
Germany	Protein Structure Factory (PSF)	http://www.proteinstrukturfabrik.de/
	Mycobacterium Tuberculosis Structural Proteomics Project (XMTB)	http://xmtb.org/start.html
Israel	The Israel Structural Proteomics Center	http://www.weizmann.ac.il/ISPC/
France	Yeast Structural Genomics (YSG)	http://genomics.eu.org/spip/index.php
	Bacterial Targets at IGS-CNRS (BIGS)	http://igs-server.cnrs-mrs.fr/Str_gen/

structures and tested the feasibility of a high-throughput structure production pipeline. PSI1 proved to be very productive, with more than 1100 structures solved over the five-year period, illustrating the immense potential for expediting protein structure solution through focused investments. The new technologies pioneered have already found their applications in conventional structural biology laboratories to facilitate the structural characterization of more difficult targets. The results and progress of major structural genomics initiatives in the U.S.A. and around the world have been recently summarized (2).

In July, 2005, the PSI advanced into a production phase. PSI2 consists of two major components: large-scale centers to increase the structural coverage of sequenced genomes by high-throughput production of structures and specialized centers to reduce technical barriers to high-throughput structure solution of challenging proteins (such as integral membrane proteins and multi-protein complexes). In addition to the production centers, centralized databases are being set up to coordinate the target selection from each center and to disseminate results to the public (Target Search for Structural Genomics (TARGETDB) at http://targetdb.pdb.org/ and Protein Expression Purification and Crystallization Database (PEPCDB) at http://pepcdb.pdb.org/). The objective of PSI2 is to solve 3000-4000 protein structures in a five-year period at a cost of ~$50,000-75,000 per structure and efficiently fill in the gaps in protein 'fold space' from all kingdoms of life. The large influx of the protein structures will benefit all structural biologists and other scientific communities, and ultimately be used to assist in drug discovery. The consortia selected for PSI2 are listed in Table 2.

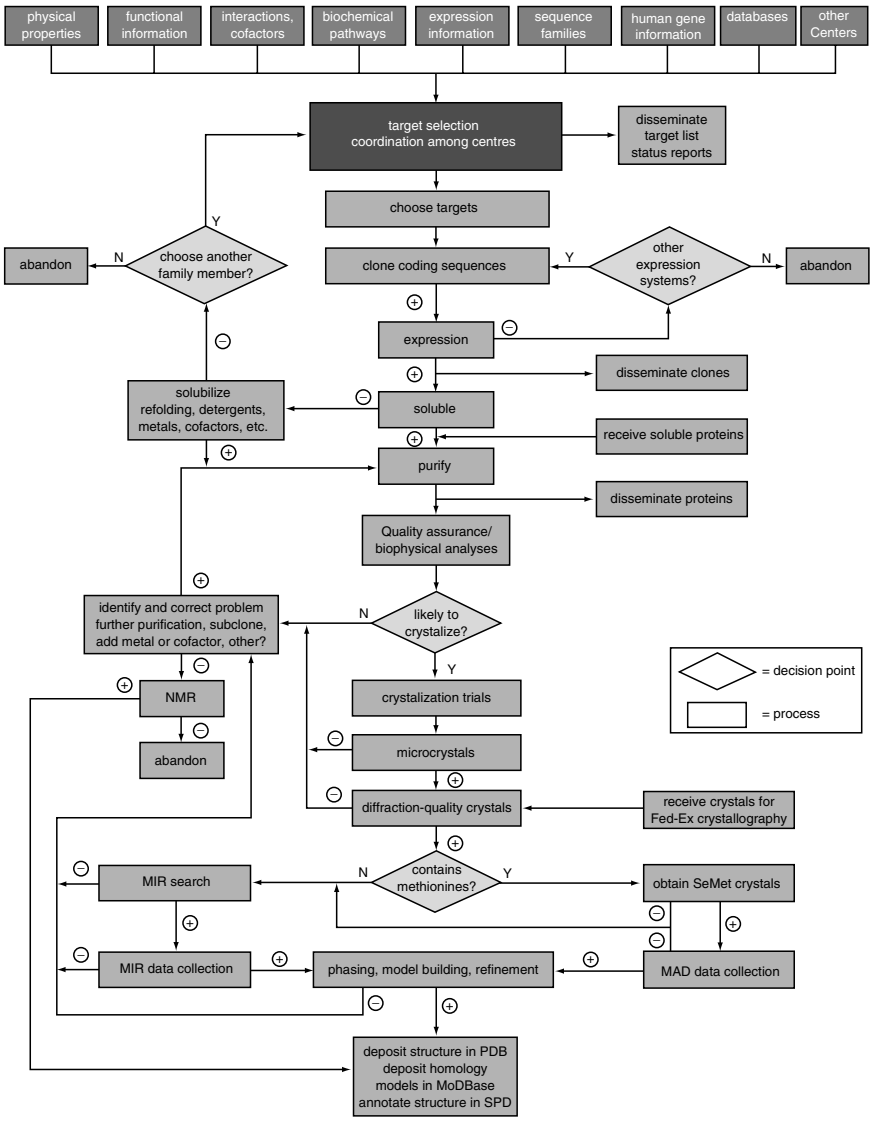

Figure 2. Schematic flow diagram of the NYSGXRC high-throughput strategy. NYSGXRC is a collaborative industrial / academic research consortium devoted to large production of protein structures and is one of the large scale centers selected for PSI2.

The New York Structural GenomiX Research Consortium (NYSGXRC), a collaborative industrial/academic research consortium devoted to large-scale production of protein structures, is one of the large-scale centers funded within PSI2. This review will focus on methodologies and technologies developed by NYSGXRC for high-throughput protein structure determination, as well as those contributed from other PSI Centers. Within NYSGXRC, proteins purified by a biotechnology/pharmacology company, SGX Pharmaceuticals Inc.

Table 2. Protein Structure Initiative (PSI)—2 in USA.

Consortium	Led By	Focus / Web Address
NIH Affiliated Large-Scale Protein Structure Production Centers		
Joint Centre for Structural Genomics JCSG	Ian Wilson Scripps Research Institute La Jolla, CA	Novel cell signaling proteins from *C. elegans*, *human*, *mouse* and *Drosophila*. http://www.jcsg.org/
Midwest Center for Structural Genomics MCSG	Andrzej Joachimiak Argonne National Laboratory Near Chicago, IL	Plan to solve quickly large number of "easy" targets through highly cost-effective methods http://www.mcsg.anl.gov/
New York Structural GenomiX Research Consortium NYSGXRC	Stephen Burley Structural GenomiX Pharmaceuticals San Diego, CA	Novel folds and biologically important proteins from three kingdoms of life. http://www.nysgxrc.org/
Northeast Structural Genomics Consortium NESGC	Gaetano Montelione Rutgers University New Brunswick, NJ	Eukaryotic model organisms which are subjects of extensive functional genomics research, including *S. cerevisiae*, *C. elegans* and *D. melanogaster*, as well as homologs from the human genome. http://www.nesg.org/
NIH Affiliated Specialized Centers		
Accelerated Technologies Center for Gene to 3D Structure	Lance Stewart deCODE biostructures Bainbridge Island, WA	Development, operation and deployment of novel approaches in miniaturization, integration and automation with an aim towards lowering the overall cost of gene to structure. http://www.atcg3d.org
Center for Eukaryotic Structural Genomics	John Markley University of Wisconsin Madison, WI	NMR spectroscopy and its biological applications; structure function relationships in proteins. http://www.uwstructural genomics.org/
Center for High-Throughput Structural Biology	George De Titta Hauptman-Woodward Medical Research Institute Buffalo, NY	Development of crystal growth methods and techniques. High-throughput structural biology. Website: forthcoming.
Center for Structures of Membrane Proteins	Robert Stroud University of California San Francisco, CA	Large effort to express eukaryotic membrane proteins with the end goal of determining their molecular structures. http://csmp.ucsf.edu/index.htm
Integrated Center for Structure and Function Innovation ISFI	Thomas Terwilliger Los Alamos National Laboratory Los Alamos, NM	Powerful methods for screening whether a molecule is properly folded and whether it will crystallize. Website : forthcoming.
New York Consortium on Membrane Protein Structure	Wayne Hendrickson New York Structural Biology Center New York, NY	Key class of proteins that serve as the portals through which cells and some components within cells communicate with the external environment. Membrane proteins lead to the development of disease and many are pharmaceutical targets of prime interest. http://www.nysbc.org/

(SGX Pharma), are distributed to four academic institutions: Albert Einstein College of Medicine (AECOM), Brookhaven National Laboratory (BNL), Case Western Reserve University (CWRU), and Columbia University (CU), for structural studies. Bioinformatics, target selection and data management tasks are performed by AECOM and University of California San Francisco (UCSF) (Table 3).

TARGET SELECTION

During PSI1, NYSGXRC and other structural genomics centers independently developed strategies for target selection. In NYSGXRC, targets were selected from microbes to human, with particular emphasis on proteins of biomedical relevance and 'hypothetical' proteins with unknown function. Because of their

Table 3. NYSGXRC Scientific Organization.

Organization Name	Scientific Team Leader	Tasks
SGX Pharmaceuticals Inc. SGX Pharma	Stephen K Burley *Principal Invstigator (PI)*	Protein Production
Albert Einstein College of Medicine AECOM	Steven Almo Institutional *Co-PI* *Department of Biochemistry* almo@aecom.yu.edu	Protein Structure Determination
	Andras Fiser *Co-PI* *Department of Biochemistry Center for Bioinformatics* fiser@fiserlab.org	Target selection, Data management and functional annotation
Case Western Reserve University CWRU	Mark R. Chance *Institutional Co-PI* *Case Center for Proteomics* mark.chance@case.edu	Metalloproteomics, Protein Structure Annotation and Publication
Columbia University CU	Lawrence Shapiro *Institutional Co-PI* *Department of Biophysics* shapiro@convex.hhmi.columbia.edu	Protein Structure Determination
University of California San Francisco UCSF	Andrej Sali *Institutional Co-PI* *California Institute for Quantitative Biomedical Research* sali@salilab.org	Comparative modeling
Brookhaven National Laboratory BNL	S. Swaminathan *Institutional Co-PI* *Biology Department*, swami@bnl.gov	Protein Structure Determination
	F. William Studier *Co-PI* Biology Department studier@bnl.gov	Protein expression strategies

biological importance and relative ease to work with, enzymes associated with small molecule metabolic pathways were also frequently selected. The target proteins were chosen based on their low sequence homology to the proteins with known structures. In addition, orthologues from several species were simultaneously cloned and purified to maximize the chance of solving the fold. Once the representative structure is solved, the efforts to solve the other orthologues are abandoned.

In PSI2, the majority of targets will be chosen using a centralized strategy as imposed by NIH target selection committee (http://grants2.nih.gov/grants/guide/rfa-files/RFA-GM-05-001.html) (3). Several strategies have been suggested and discussed in detail (4, 5) such as the "Pfam5000" strategy, which involves selecting the 5,000 largest families from the Pfam database as sources for targets. It is estimated that if at least one structure is solved from each of these 5,000 families, it will provide sequence coverage of 68% of prokaryotic proteins and 61% of eukaryotic proteins, and greatly increase our ability to assign folds for all sequenced genomes through modeling and threading methods. Pfam5000 strategy complements the other strategies such as random target selection strategies and single-genome strategies (5). In PSI2, NIH requests a target selection strategy that combines coarse-grained coverage of sequence space, proteins of known medical interest, and contributions from the scientific community (5). NYSGXRC will follow the method suggested by NIH target selection committee. About 70% of the targets will be selected from available genomes in coordination with the other 3 large scale structural genomics centers in PSI2.

PROTEIN PRODUCTION AND BIOPHYSICAL ANALYSIS

During PSI1, SGX Pharmaceuticals Inc, (http://www.sgxpharma.com/) established a modular industrial platform for the recombinant protein production. cDNAs of interest were cloned by PCR amplification and inserted into a suitable expression vector. The procedure can be operated in a parallel fashion for high-throughput (6). The protocol developed by NYSGXRC laboratories, using the T7 RNA polymerase-dependent *E.coli* expression vector system (pET-vectors), is a universal system to generate recombinant protein for structural analysis (7, 8). pET vectors are usually combined with *E. coli* B strain BL21 or the derivatives that are engineered to carry the T7 RNA polymerase gene. These strains, however, have limitations in cloning and stable propagation of the expression constructs. The approach based on the concept of topoisomerase mediation, which involves directional flap ligation of a blunt-ended PCR product into pET100/D-TOPO Vector (Invitrogen) was adopted by NYSGXRC. It creates a fusion protein bearing an N-terminal His_6-tag followed by a polio viral protease cleavage site followed by the protein sequence of interest. Recently, NYSGXRC implemented an additional vector for recombinant protein expression based on N-terminal fusions with a yeast form of SUMO, a small ubiquitin-like modifier that frequently enhances the solubility to the recombinant fusion protein (9, 10). The pSUMO system utilizes an N-terminal His_6-tag SUMO fusion with the respective target sequence. The protein is expressed in bacteria, purified by metal affinity chromatography, and liberated from the His_6-SUMO fusion by cleavage with a modified version of the desumoylating enzyme Ulp1.

To facilitate the high-throughput production of proteins, a Beckman Biomek FX robotic platform has been adopted to perform many of the steps required from PCR to transformation in 96-well format with bar code tracking of sample and reagent plates (6). Some steps are conducted off-line with multichannel pipetting. Small-scale (1 μg) purification of recombinant proteins followed by spotting onto a matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) sample plate allows rapid identification of constructs expressing the appropriate product (11).

All soluble proteins purified were subjected to biophysical analyses, including mass spectrometry for construct verification and protein purification, analytical gel filtration for homogeneity, domain mapping by limited proteolysis combined with mass spectrometry (LPMS) to analyze for "floppy-ends", peptide mapping of posttranslational modifications *via* mass spectrometry, and UV/vis. absorbance spectroscopy to identify possible bound co-factors (6).

Proteins produced at SGX Pharmaceuticals (10+ mg) were shipped to AECOM, BNL and CU for protein structure determination (Table 3) and smaller amounts of samples (0.1 mg) were also shipped to Case Center for Proteomics, CWRU for intrinsic metal detection through automated X-ray absorption spectroscopy (XAS) at the NSLS beamline X9B (12, 13).

The results of these measurements are made available to the crystallographers to facilitate *de novo* phasing and structure solution using the intrinsic metals (14). SGX Pharmaceuticals provides selenomethionine (SeMet)-labeled proteins for X-ray single/multiple anomalous dispersion (SAD/MAD) studies at synchrotron sites after adequate crystallization conditions have been established for native crystals. The use of synchrotron radiation is crucial to the NYSGXRC pipeline. High brilliance and energy tuneability (mainly Se Edge) are prerequisites for fast data collection from small protein crystals (15).

PROTEIN STRUCTURE ANALYSIS

X-ray crystallography is the primary technique used for protein structure determination in most of structural genomics centers. The efforts to solve the protein structures have seen great improvements over the past decade and resulted in dramatic accumulation of protein structures in the PDB (http://www.rcsb.org/pdb).

Crystallization

It is usually regarded as a major bottleneck for structure determination by X-ray crystallography with low success rates evident in all structural genomics centers. As of October 2005, for all structural genomics centers worldwide, only 4,692 targets (8% of the proteins cloned) yielded crystals and 2,034 targets (3.5%) resulted in crystal structures. In PSI centers, 3,629 targets (7% of the proteins cloned) were crystallized yielding 1,210 (2%) crystal structures (as of October 2005). For structural genomics centers focusing on medically related human proteins, the statistics are even lower. For example, within the Protein Structure Factory (Germany), only 63 targets (7% of the proteins cloned) were crystallized and 11 (1.2%) protein structures were solved. Frequently, the failure of producing diffracting quality crystals is attributed to disordered regions, particularly at

N- and C-termini. These full length proteins have high tendency to aggregate with low yield (2 mg/L) and are difficult to concentrate (>1 mg/ml). Several techniques have been developed to identify the stable domains and to remove the structural micro-heterogeneity of the proteins. The technique, termed limited proteolysis by mass spectroscopy (LPMS) (16, 17), has been implemented in the NYSGXRC pipeline (18). The constructs obtained by LPMS possess enhanced qualities such as high yield (30 mg/L), stability overtime and greater tendency to crystallize (19). It has been demonstrated that the targets resistant to proteolysis are good candidates for crystallographic studies with 27% of these targets yielding 3-dimensional structures thus far, whereas only 9% of the targets showing partial proteolysis yielded 3-dimensional structures to date. Large-scale sub-cloning and subsequent testing of expression, solubility, and crystallization are currently underway (18). Another technique, the hydrogen/deuterium exchange mass spectroscopy (DXMS) that allows rapid identification of unstructured regions in proteins, was developed at JCSG (20). These targets, TM0160, TM1171, TM0613 and TM0021, have been successfully crystallized after DXMS analysis (21). Bioinformatics strategies can also be applied to identify disorder region using computational tools such as *PONDR* (22), *GlobPlot* (23) and *DisEMBL* (24), and to help redesign constructs utilizing different expression systems and genes from different species. For example, crystals suitable for X-ray crystallographic studies for Protein Structure Factory (Germany) target PSF200001226 (PDB ID 1U2H) were obtained following the truncation of the first 14 N-terminal amino acids that were predicted to be structurally disordered and these crystals diffracted to 0.96 Å resolution (25).

The field of protein crystallization is revolutionized with the development of robotic technology. All steps in crystallization have been automated including crystallization plate setup and bar coding, movement of crystallization plates into and out of the storage vault, and crystallization plate imaging, image processing, storage and display. These traditionally tedious manual procedures have been addressed to save proteins, time and cost of the crystallization experiments. The robotic imaging systems (26) and macromolecular crystallization using free-interface diffusion method at the nanoliter scale (27) have been described recently. The system is capable of performing multidimensional screening (mixing 5-10 solutions) to explore more crystallization space, maximizing the chance of obtaining crystals. Furthermore, capillary-containing protein crystals can be directly mounted on the goniometer, eliminating the need of crystal manipulation and mounting. Currently, NYSGXRC has implemented parallel robotic stations for high-throughput crystallization screening at each crystallography site utilizing 96-well “sitting drop” vapor diffusion method. The optimization screens are still performed manually.

Synchrotron Data Collection

New third generation synchrotrons with beamlines equipped with insertion devices provide more intense, tunable and stable X-ray beams, allowing crystallographers to collect higher quality data much more rapidly. As of August 2005, beamline 19ID-APS (183 deposits) has the highest number of deposited PDBs among the beamlines utilized by PSI centers. Several bending magnet

beamlines including X4A (90 deposits), 19BM-APS (88 deposits) and X9A-NSLS (53 deposits) are also making significant contributions. Other factors such as flash-freezing techniques, faster and larger CCD X-ray detectors have led to dramatic increases in the rate of structure determination. Novel methods for automatic crystal mounting, optical crystal centering, data collection and indexing of the crystals have been developed at many synchrotron sites (28). During PSI1, NYSGXRC built highly collimated and extremely intense beamline, X29-NSLS, a novel mini-gap undulator beamline, for efficient high-resolution data collection from very small crystals to facilitate rapid structure determination (29). The X29 optical system comprises a double crystal monochromator with a sagittally bent second crystal providing horizontal focusing, followed by a cylindrically bent mirror providing vertical focusing and harmonics rejection. The photon energy range of the monochromator is 4-18 keV which covers the absorption edges of all commonly used heavy atoms (30). The method of MAD phasing requires X-ray diffraction measurements at two to four X-ray energies near an atomic absorption edge of the heavy atom, chosen to maximize the real and imaginary components of anomalous scattering. MAD phasing on data collected from crystals containing variety of anomalous scatterers including Se, Fe, Cu, Br, Tb, Pt, Hg, W, Au and Zn, is the method of choice for determining new crystal structures. In addition, X29 is equipped with state-of-the-art ADSC Q315 detector system (near 100 μm resolution with near 2 second readout time) in order to take advantage of the short exposure time (~5 sec per frame) and to collect data on the crystals with large unit cells (>600 Å). Furthermore, the X29 station is equipped with gaseous liquid-nitrogen cooling, highly automated beamline control, efficient software packages to facilitate high-throughput data collection. Installation of sample changing and crystal alignment robotics which automate the initial crystal screening step are underway at X29.

The availability of powerful computers contributes to high speed data collection by automation in selection of optimum data acquisition parameters and processing protocols (31). As of August 2005, the program HKL was used on fly for integration (716 PSI deposits) and for scaling (743 PSI deposits). Another other popular program for data integration is MOSFLM (154 PSI deposits) and for data scaling is SCALA (174 PSI deposits). However, since protein crystals differ enormously in their diffraction properties, it is difficult to develop a complete automated system using a single data collection strategy that can satisfy all possible scenarios (32).

Phasing

One of the primary problems in macromolecular X-ray crystallography is the phase problem. Single/multiple anomalous dispersion (SAD/MAD), single/multiple isomorphous replacement (SIR/MIR) and molecular replacement (MR) methods are commonly used to solve the phase problem. Recently, phasing using SAD/MAD with SeMet-substituted proteins has become a routine process in protein crystallography (33). In 2004, the percentage of newly-deposited structures, which share less than 30% of sequence identity to any known structures at the time of deposition, are 61% (555 of 915), 63% (326 of 521) and 77%

(62 of 82), respectively, for all SG (Structural Genomics) centers, PSI centers and NYSGXRC. These statistics indicate that the majority of protein structures from Structural Genomics centers are determined using SAD or MAD methods. Novel methods, such as heavy atom derivatization with halides, SAD with sulfur atoms, phasing using Hg radiation damage and brute force molecular replacement, were developed by NYSGXRC to facilitate high-throughput structure determination (34-36).

Automated Structure Solution

Development of integration and extension of existing crystallographic software provides user-friendly tools for rapid automated structure determination. Several integrated program packages are now in general use and listed in Table 4. Two popular automated protocols are commonly used by the NYSGXRC. First, the program *HKL2MAP* connects *SHELX* suite (37). The processed data from HKL-2000 collected at different wavelengths are scaled for data analysis, phase calculation and the electron density map are displayed using XFIT (38). The electron density map can be interpreted and fitted through automatic model building program such as *ARP/wARP* (39). Second, *SOLVE/RESOLVE* suit is fully automated and can function with data resolution as low as 3 Å (40). With these approaches, initial models can be built and displayed while the data are still being collected. With fast computers and the automated crystallographic software, structure solution is straightforward in many cases. The initial model built with automated programs is further completed with manual model fitting using programs O (41) or COOT (42), and subjected to refinement with programs REFMAC and/or CNS (43). As of August 2005, 469 PSI deposits are refined using CNS and 459 deposits are refined with REFMAC programs. Web-based tools such as AutoDep are available to deposit the coordinates and structure factors to the PDB with immediate release. Attempts are underway around the world including NYSGXRC to build user-friendly automated tools for protein structure determination (44-50).

Project Management Systems and Progress Report

An important feature of structural genomics is the on-line documentation of progress that allows data mining for evaluating the enterprise. Status information for all the steps of the high-throughput pipeline are archived through a centralized NYSGXRC database (http://www.nysgxrc.org) and linked to the Protein Data Bank (PDB) through the target database (http://targetdb.rcsb.org). The progress of all NYSGXRC targets is shown in Table 5. As of August, 2005, 190 protein structures have been determined in NYSGXRC. So far 11.2% of the cloned targets have yielded deposited structures (1,685 cloned:190 structures in PDB). However, this success rate is well above the 4.4% success rate indicated in the target database for all structural genomics centers worldwide (56,146 cloned: 2475 structures in PDB as of October, 2005) and the 2.3% success rate for all PSI centers (51,131 cloned: 1180 structures in PDB). Interestingly, 64% of cloned NYSGXRC targets were purified and 18% of purified proteins yielded crystal structures.

Table 4. Selected Computational Resources for Protein Structure Analysis.

Programs	Features	Web Address
DENZO/HKL-2000	Analysis and process X-ray data collected from single crystals	http://www.hkl-xray.com/
MOSFLM	Analysis and process X-ray data collected on the image plate and CCD	http://www.mrc-lmb.cam.ac.uk/harry/mosflm/
HKL2MAP/SHELX	Direct methods, phasing refinement and heavy atoms	http://shelx.uni-ac.gwdg.de/SHELX/
SnB	Direct methods, phasing refinement and heavy atoms	http://www.hwi.buffalo.edu/SnB/
BnP	Substructure determination and refinement	http://www.hwi.buffalo.edu/BnP/
SOLVE/RESOLVE	Heavy atom refinement, phasing and chain tracing	http://www.solve.lanl.gov/
AUTOSHARP	Statistical heavy atom refinement and phasing	http://www.globalphasing.com/sharp/
PHENIX	Python-based Hierarchical Environment for Integrated Xtallography	http://www.phenix-online.org/
CCP4	Collaborative Computing Project 4	http://www.ccp4.ac.uk/main.html
ARP/wARP	Automatic model refinement and chain tracing	http://www.embl-hamburg.de/ARP/
CNS	Refinement, phasing, heavy atoms, molecular replacement	http://asdp.bio.bnl.gov/cns_solve_1.1/doc/html/index.html
O	Graphics—model and map visualization and building	
XtalView	Solving and building crystal structures	http://www.sdsc.edu/CCMS/Packages/ XTALVIEW/xtalview.html
COOT	Crystallographic Object-Oriented Toolkit. Model building and validation.	http://www.ysbl.york.ac.uk/~emsley/coot/
PyMol	A molecular graphics system with an embedded Python interpreter	http://pymol.sourceforge.net/
USF	Uppsala Software Factory—supportive programs	http://alpha2.bmc.uu.se/usf/
PROCHECK	Check stereochemistry quality of protein models	http://www.biochem.ucl.ac.uk/~roman/ procheck/procheck.html
SFCHECK	Assessment between Structure Factors and models	http://www.ysbl.york.ac.uk/~alexei/sfcheck.html
AutoDep	Model validation and deposition.	http://deposit.rcsb.org/adit/ http://www.ebi.ac.uk/msd-srv/autodep4/index.jsp

Table 5. Progress of NYSGXRC as of August 16, 2005*.

Different Stages of the Pipeline	
Targets Selected	2306
Cloned	1685
Expressed	1375
Soluble	1164
Purified	1068
Crystallized	391
Diffraction Quality Crystals	240
Protein Structures in PDB	190
Protein Structures Released by year	
<2000	6
2001	12
2002	14
2003	35
2004	82
2005	41

*Updates available at http://www.nysgxrc.org and mirrored http://targetdb.pdb.org/statistics/sites/NYSGRC.html.

In contrast, only 19% of the cloned targets from all PSI centers were purified and 13% of purified proteins resulted in crystal structures. The statistics are similar for all SG centers where only 21% of cloned targets were purified and 17% of purified proteins resulted in crystal structures. The growth of protein structures released by NYSGXRC by year is shown in Table 5. The number of protein structures released in 2004 by all SG and PSI centers (including NYSGXRC) are 914 and 523, respectively. The contribution from NYSGXRC alone (82 structures for 2004) is about 9.7% of structures by all SG centers and 18% by PSI centers.

The quantity of structures solved may not be the best measure and it is important to analyze the quality of the structures solved within the various projects. The productivity depends upon the nature of targets and the availability of high-throughput methodologies and the technical infrastructure to tackle them. Several beamlines at National Synchrotron Light Source (NSLS) and Advanced Photon Source (APS) have been utilized for X-ray diffraction data collection. The majority of structures were determined from the data collected at NSLS X9A (52 structures) and APS 31ID (33 structures). In addition, 19 structures were from NSLS X29 (29). The average resolution of all target structures by the consortium is 2.26Å. The average R_{work} and R_{free} for these structures are 0.211 and 0.25, respectively, indicating the high quality of the structures determined by NYSGXRC. The average sequence length for NYSGXRC deposited structures is 296 residues (as of August 2005), significantly higher than the average by all PSI centers (358 residues). For the other three large PSI centers selected for the production phase, the average lengths are 211, 177 and 255, respectively, for MCSG, NESGC and JCSG. The range of organisms from which the NYSGXRC targets were selected from reflects the broad focus of the NYSGXRC (Table 6). As of August, 2005, 149 structures are from prokaryotes (archaea and bacteria) and 32

Table 6. NYSGXRC Structures by Organism (as of August 16, 2005*).

Organisms from Three Kingdoms of Life	No. of organisms × structures	Total
Escherichia coli	1 × 37	37
Bacillus subtilis	1 × 16	16
Saccharomyces cerevisiae	1 × 13	13
Enterococcus faecalis; Pseudomonas aeruginosa;	2 × 8	16
Agrobacterium tumefaciens; Haemophilus influenzae; Methanococcus jannaschii; Mus Musculus; Vibrio cholerae;	5 × 5	25
Archaeoglobus fulgidus; Deinococcus radiodurans; Homo sapiens; Thermotoga maritima	4 × 4	16
Bacillus halodurans; Mycobacterium tuberculosis; Neisseria meningitides; Streptococcus pneumoniae; Streptococcus pyogenes;	5 × 3	15
Campylobacter jejuni; Clostridium acetobutylicum ATCC 824; Listeria monocytogenes; Phleum pretense; Salmonella typhimurium; Schizosaccharomyces pombe; Staphylococcus aureus.	7 × 2	14
Aquifex aeolicus; Arabidopsis thaliana; Bacteroides thetaiotaomicron; Borrelia burgdorferi; Bradyrhizobium japonicum; Caenorhabditis elegans; Caulobacter crescentus; Chlorobium tepidum TLS; Encephalitozoon cuniculi; Helicobacter pylori J99; Klebsiella pneumoniae; Listeria innocua Clip11262; Salmonella enterica; Shigella flexneri; Streptococcus mutans UA159; Thermoplasma acidophilum; Xanthomonas campestris	17 × 1	17

*Updates available at http://www.nysgxrc.org and mirrored http://targetdb.pdb.org/statistics/sites/NYSGRC.html.

structures are from eukaryotes (yeast, plasmodium, arabidopsis, nematode, fly, mouse, humans). The statistics indicate that NYSGXRC has been very productive. The number of high-quality crystal structures by NYSGXRC through the high-throughput pipeline is promising, however, it could also be argued that many of these structures are "easy" targets (low hanging fruit), so that both quantity and quality are expected to be high. More challenging targets, such as human and other eukaryotic proteins, and large macromolecule assemblies, pose a greater challenge. The functional coverage of NYSGXRC structures based on enzyme classification, biological process, cell component, molecular function and disease is shown in Table 7. About 37% of the solved structures are hypothetical proteins with unknown function.

In order to annotate proteins with unknown function, high-throughput tools are needed at each step of the experimental pipeline, including the timely release of protein structures to biologists and other scientists. For instance, out of 190 protein structures by NYSGXRC to date, 140 (74%) of them are "to be published" and only about 50 (26%) have peer-reviewed publications. Similar statistics can also be found for other structural genomics centers around the world.

Table 7. NYSGXRC Structures by Functional Classification (as of August 16, 2005).

Classification	Functional coverage × structures	Total
Unknown Function	1 × 62	62
Transferase	1 × 23	23
Hydrolase	1 × 16	16
Oxidoreductase	1 × 14	14
Lyase, Isomerase	2 × 8	16
Transcription	1 × 7	7
DNA Binding	2 × 4	8
Structural Protein	1 × 3	3
Signaling; Protein Binding; Lipid Binding; Ligase; Harmone/Growth Factor; Biosynthetic; Allergen	7 × 2	14
Penicillin Binding; Immune System	2 × 1	2

HOMOLOGY MODELING OF REPRESENTATIVE PROTEIN FAMILY MEMBERS

Homology modeling or comparative modeling takes advantage of structural similarities within protein families. This technique is based on the assumption that all the homologous members of the protein family are related by divergent evolution from a common ancestor and must share a common basic fold. Solving the structure of any single member of a protein family clustered at 30% or more identity allows comparative modeling of the entire family in most cases. Basic approaches to homology modeling were initiated by Greer in 1981 (51) and Sali and Blundell in 1993 (52), and the methods were recently reviewed (53). Automated homology modeling with MODWEB has been fully implemented by NYSGXRC and is now being used routinely by NYSGXRC members, other PSI centers and researchers around the world (54). About 146,236 protein structure models including 12,651 accurate models have been generated using 181 NYSGXRC structures with an average number of models per structure of 807. The quality and usefulness of homology models depend critically on the level of sequence identity. The accuracy of a model based on a template with >50% sequence identity is equal to a medium-resolution crystal structure (3.0 Å resolution). Models based on >30% but less than 50% sequence identity are suitable for many applications including fold assignment and molecular replacement for phasing. Models based on <30% identity have the possibility of significant alignment errors. As a general rule, 30% sequence identity is the arbitrary cutoff for effective homology modeling. Despite the possible errors, less accurate models are useful in many applications in structural biology. For examples, they can be used to identify putative active site residues, redesign expression constructs, and as templates for structure determination by molecular replacement. Recently, combining homology modeling results with low resolution electron microscopic maps have been shown to help model more difficult targets, such as macromolecular complexes and eukaryotic proteins, and this approach is becoming well accepted (55-57).

STRUCTURE TO FUNCTION

Several bioinformatics servers, such as ProFunc (58) and ProTarget (59) servers, have been developed and are available for public access for protein structure and sequence analysis which includes prediction of the function of proteins from the solved structures. The information available from the three-dimensional structure of a protein, relating to its function, is summarized in Figure 3a. The theory and practice of how to predict function from sequence and structure have been thoroughly discussed (60-63). Currently, there are two sequence-based approaches for protein annotation. Enzyme classification of enzymes (EC numbers) has been used to study the sequence, structure and function relationships (64-66). Second, the Gene Ontology (GO) provides a consistent view of molecular function, biological process and cell component beyond enzymes (67).

The summary of biological function information extracted from the protein structures and structure-based function discovery has been described (57, 68). One of the most powerful methods for functional inference is identification of homologous proteins and protein structures through structural comparisons (69, 70). The proteins can diverge beyond significant sequence similarity but still retain the 3D fold of their ancestor and even similar functions. Commonly used web-based servers to scan the novel protein structure against the known protein structure database (PDB) and retrieve closest matches are DALI (71) and VAST (72). For example, the crystal structure of *E.coli* L-Arabinose isomerase (NYSGXRC target T2031, PDB ID 2AJT) shows significant similarity to that of *E.coli* fucose isomerase (PDB ID 1FUI) despite the very low sequence identity (9.7%) shared by the two enzymes (Figure 3b). Both structures retain hexameric subunit assembly for enzyme activity based on the results from electron microscopy studies (73). In addition, the two enzymes show similar substrate specificities (74).

However, DALI and VAST servers usually fail to produce any match if the new protein structure possesses a novel fold. In this case, identification of functional sites across different folds is required (75). Using databases of active site templates, programs such as PINTS (76), PROCAT (77) and Rigor (78) identify conservation of functional patterns within the structures of different folds. If both structure comparison and functional site comparison for proteins with unknown function fail to yield any match, analysis of the sequence conservation through evolution may reveal their functions. Conservation score can be calculated for each residue in the sequence by comparing the residue variability at each position in a multiple sequence alignment of homologous proteins, and mapped onto the protein surface. The web-based server ConSurf (79, 80) identifies most likely functional or protein-protein interaction patches on the surface of the protein structure. Further, analysis of clefts and cavities on the protein surface can be useful to locate its putative active site and sometimes provide clues to its function. The program SURFNET (81) performs the analysis of clefts and their surface properties automatically and points to the regions that are most likely to be functionally important. Many of the bioinformatics approaches described above are routinely in use to annotate protein functions.

Presence of metal atoms in proteins often marks active centers and provides a guide to functional annotation. For example, the crystal structure of ybeY protein from *E.coli* (NYSGXRC target T842, PDB ID 1XM5) reveals that the

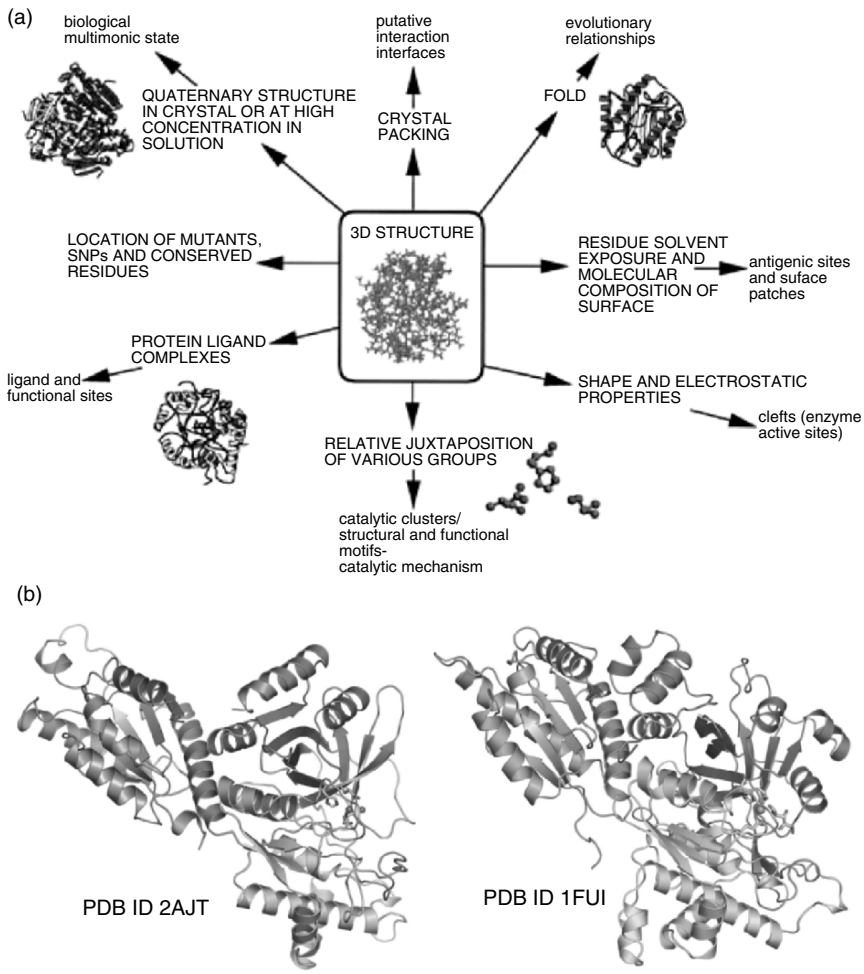

Figure 3. Structure to function and examples. (a) Summary of information derived from protein structure, with biological function related. Taken from Thornton, J.M., Todd, A.E., Milburn, D., Borkakoti, N., Orengo, C.A. (2000) From Structure to function: approaches and limitations. Nat. Struct. Biol. 7(Suppl.), 991-994. (b) Structural conservation in distant evolutionary relatives, *E.coli* L-Arabinose isomerase (PDB ID 2AJT, left) and *E.coli* fucose isomerase (PDB ID 1FUI, right), in the absence of significant sequence identity; and

protein binds to a metal ion in a tetrahedral geometry with three histidine residues (Figure 3c). The fourth coordination site might be a water molecule which was not seen in the structure. The structure of ybeY and its sequence similarity to a number of predicted metal-dependent hydrolases suggests a potential functional assignment for this protein (82). A high-throughput technology to identify proteins containing metals has been developed based on X-ray fluorescence analysis to analyze for transition metal content at beamline X9B of National Synchrotron Light Source. The initial results and potential application towards protein annotation have been discussed recently (13).

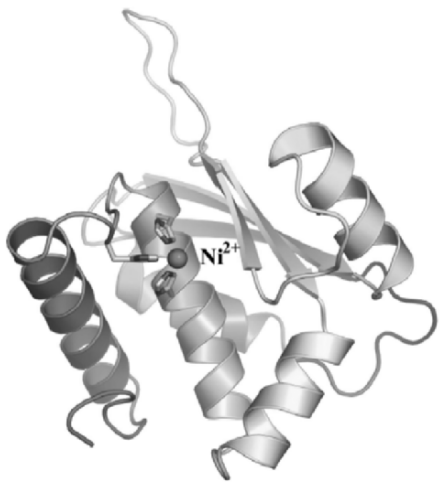

Figure 3. (*Continued*) (c) Presence of a metal ion, Ni^{2+}, guides protein annotation for NYSGXRC target T842 (PDB ID 1XM5).

The observation of an unexpected bound ligand sometimes gives clues to protein function annotation (83-85). For example, the crystal structure of the *E.coli* Ycei periplasmic protein (NYSGXRC target T792, PDB ID 1Y0G) revealed a dimer of β-barrels (similar to lipocalin superfamily folds) with a continuous electron density feature running along the entire length of the central axis of the β-barrels. The electron density was interpreted as 2-octaprenylphenol (OPP) and mass spectroscopic studies are under way to confirm the identity. The OPP bound to Ycei helps to identify the active site. In principle, experimental approaches such as functional assays and site-directed mutagenesis should follow to confirm the annotation (86, 87). High-throughput methods to automate enzymatic analysis, termed as enzyme genomics, by screening for protein–ligand complex libraries using mass spectrometry and matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) have been initiated (88-90). The development of proteomics strategies for genome annotation are in progress by utilizing the structure-based functional discovery (91-93).

Auto Publish Web Tool

To speed up protein structure publications and annotation, NYSGXRC is developing an automated server to prepare structure reports automatically in short structure report format of journal Acta Crystallographica F. The web-tool aims at facilitating publication of newly-solved structures by automating major steps in data analysis and manuscript preparation, such as producing tables, figures, and performing standard-functional analysis based on structure and sequence. The server generates five outputs as shown in the flowchart (Figure 4). Users will be able to start with the desired PDB code and obtain a raw manuscript that consist all the standard requirements of a regular report on a crystallographic protein structure. First, a WORD format file useful for 'Materials and

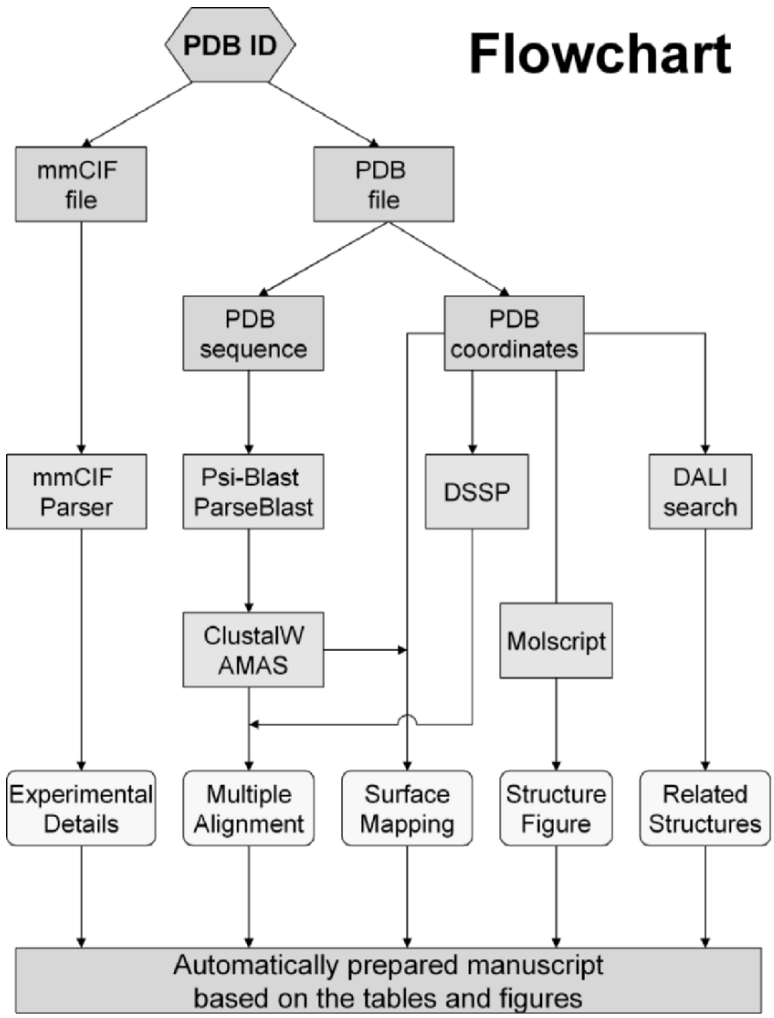

Figure 4. Flowchart of Auto Publish web server which generates five outputs including experimental details, structure images and standard functional and structural analysis to facilitate rapid publication.

Methods' section is generated along with a table containing statistics for data collection, structure solution and refinement by extracting parameters from the mmCIF PDB file. Second, the amino acid sequence of the protein structure is compared using PSI-Blast (94) with the sequences of homologs from databases and sequence conservation analysis is performed with AMAS (95) and displayed with ALSCRIPT (96) programs. The conservation scores for each residue are placed in the temperature factor column in PDB file and the modified PDB file can be automatically uploaded into graphic programs such as PyMol (97) to generate a protein surface plot with the conserved regions appropriately color coded. In addition, a standard sequence alignment figure with selected homologs is also generated and conserved residues are identified and highlighted to assist in

protein family classification and functional annotation. Third, protein structure is automatically uploaded into a structure alignment program (DALI (71)). The comparative analysis results can be used for the protein fold assignment and active site identification. The web-tool has been tested on one of the NYSGXRC target structure (1XM5) (82). Currently, in-depth testing of the server is underway to prepare the structure reports more automatically by using NYSGXRC protein structures.

CONCLUSION

The NYSGXRC has implemented pipeline and potential for experimentally determining 100-200 protein structures annually. All consortium activities can be scaled up to increase capacity for protein structure production anticipated in PSI2. NYSGXRC is dedicated to unravel the shapes and to derive the functions of many hundreds of proteins in the next few years. The structural information generated in structural genomics will have a profound impact in many related fields including drug discovery by providing scientists a large structure database for structure-based drug design.

REFERENCES

1 Norvell, J.C. and Machalek, A.Z. (2000) Nat. Struct. Biol. 7 Suppl; 931.
2 Todd, A.E., Marsden, R.L., Thornton, J.M. and Orengo, C.A. (2005) J. Mol. Biol. 348(5), 1235-1260.
3 PSI-phase 1 and beyond. (2004) Nat. Struct. Mol. Biol. 11(3), 201.
4 Chandonia, J.M., Earnest, T.N. and Brenner, S.E. (2004) Genome Biol. 5, 343.
5 Chandonia, J.M. and Brenner, S.E. (2005) Proteins 58(1), 166-179.
6 Bonanno, J.B., Almo, S.C., Bresnick, A., Chance, M.R., Fiser, A,. Swaminathan, S., Jiang, J., Studier, F.W., Shapiro, L., Lima, C.D., Gaasterland, T.M., Sali, A., Bain, K., Feil, I., Gao, X., Lorimer, D., Ramos, A., Sauder, J.M., Wasserman, S.R., Emtage, S., D'Amico, K.L. and Burley, S.K. (2005) J. Struct. Funct. Genomics 6, 225-232.
7 Studier, F.W. (2005) Protein Expr. Purif. 41(1), 207-234.
8 Studier, F.W., Rosenberg, A.H., Dunn, J.J. and Dubendorff, J.W. (1990) Methods Enzymol. 185, 60-89.
9 Mossessova, E. and Lima, C.D. (2000) Mol. Cell 5(5), 865-876.
10 Reverter, D. and Lima C.D. (2005) Nature 435(7042) 687-692.
11 Huang, R.Y., Boulton, S.J., Vidal, M., Almo, S.C., Bresnick, A.R. and Chance, M.R. (2003) Biochem. Biophys. Res. Commun. 307(4), 928-934.
12 Chance, M.R., Fiser, A., Sali, A., Pieper, U., Eswar, N., Xu, G., Fajardo, J.E., Radhakannan, T. and Marinkovic, N. (2004) Genome Res. 14(10B), 2145-2154.
13 Shi, W., Zhan, C., Ignatov, A., Manjasetty, B.A., Marinkovic, N., Sullivan, M., Huang, R. and Chance, M.R. (2005) Structure (Cambr.) 13(10), 1473-1486.

14 Rajashankar, K.R., Chance, M.R., Burley, S.K., Jiang, J., Almo, S.C., Bresnick, A., Dodatko, T., Huang, R., He, G., Chen, H., Sullivan, M., Toomey, J., Thirumuruhan, R.A., Franklin, W.A., Sali, A., Pieper, U., Eswar, N., Ilyin, V. and McMahan, L. (2002) NSLS Activity Report 2, 28-32.
15 Kim, S.H. (1998) Nat. Struct. Biol. 5 Suppl; 643-645.
16 Nieves-Alicea, R., Focia, P.J., Craig, S.P., 3rd and Eakin, A.E. (1998) Biochim. Biophys. Acta 1388(2), 500-505.
17 Sharma, S., Singh, T.P. and Bhatia, K.L. (1997) Acta Crystallogr. D. Biol. Crystallogr. 53(Pt1), 116-118.
18 Gao, X., Bain, K., Bonanno, J.B., Buchanan, M., Henderson, D., Lorimer, D., Marsh, C., Reynes, J.A., Sauder, J.M., Schwinn, K., Thai, C. and Burley, S.K. (2005) J. Struct. Funct. Genomics 6(2-3), 129-134.
19 Matsui, T., Hogetsu, K., Akao, Y., Tanaka, M., Sato, T., Kumasaka, T. and Tanaka, N. (2004) Acta Crystallogr. D. Biol. Crystallogr. 60(Pt 1), 156-159.
20 Pantazatos, D., Kim, J.S., Klock, H.E., Stevens, R.C., Wilson, I.A., Lesley, S.A. and Woods, V.L. Jr., (2004) Proc. Nat. Acad. Sci. U.S.A. 101(3), 751-756.
21 Spraggon, G., Pantazatos, D., Klock, H.E., Wilson, I.A., Woods, V.L., Jr. and Lesley, S.A. (2004) Protein Sci. 13(12), 3187-3199.
22 Romero, P., Obradovic, Z. and Dunker, A.K. (2004) Appl. Bioinformatics 3(2-3), 105-113.
23 Linding, R., Russell, R.B., Neduva, V. and Gibson, T.J. (2003) Nucl. Acids Res. 31(13), 3701-3708.
24 Iakoucheva, L.M. and Dunker, A.K. (2003) Structure (Cambr.) 11(11), 1316-1317.
25 Manjasetty, B.A., Niesen, F.H., Scheich, C., Yuette, R., Gotz, F., Behlke, J., Sievert, V., Heinemann, U. and Bussow, K. (2005) BMC Struc. Biol., 5(1), 21.
26 Mayo, C.J., Diprose, J.M., Walter, T.S., Berry, I.M., Wilson, J., Owens, R.J., Jones, E.Y., Harlos, K., Stuart, D.I. and Esnouf, R.M. (2005) Structure (Cambr.) 13(2), 175-182.
27 Segelke, B. (2005) Expert Rev. Proteomics 2(2), 165-172.
28 Muchmore, S.W., Olson, J., Jones, R., Pan, J., Blum, M., Greer, J., Merrick, S.M., Magdalinos, P. and Nienaber, V.L. (2000) Structure Fold Des. 8(12), R243-246.
29 Robinson, H.H., Shi, W., Sullivan, M., Nolan, W., Schneider, D.K., Berman, L., Lynch, D., Rock, L., Rosenbaum, G., Johnson, E., Chance, M.R. and Sweet, R.M. (2005) Synchrotron Radiation News 18(5), 27-31.
30 Ogata, C.M. (1998) Nat. Struct. Biol. 5 Suppl; 638-640.
31 Leslie, A.G., Powell, H.R., Winter, G., Svensson, O., Spruce, D., McSweeney, S., Love, D., Kinder, S., Duke, E. and Nave, C. (2002) Acta Crystallogr. D. Biol. Crystallogr. 58(Pt 11), 1924-1928.
32 Dauter, Z. (2005) Prog. Biophys. Mol. Biol. 89(2), 153-172.
33 Hendrickson, W.A. and Ogata, C.M. (1997) Methods Enzymol. 276, 494-523.

34 Boggon, T.J. and Shapiro, L. (2000) Structure Fold Des. 8(7), R143-149.
35 Strokopytov, B.V., Fedorov, A., Mahoney, N.M., Kessels, M., Drubin, D.G. and Almo, S.C. (2005) Acta Crystallogr. D. Biol. Crystallogr. 61 (Pt 3), 285-293.
36 Ramagopal, U.A., Dauter, Z., Thirumuruhan, R., Fedorov, E. and Almo, S.C. (2005) Acta Crystallogr. D. Biol. Crystallogr. 61(Pt 9), 1289-1298.
37 Schneider, T.R. and Sheldrick, G.M. (2002) Acta Crystallogr. D. Biol. Crystallogr. 58(Pt 10 Pt 2), 1772-1779.
38 McRee, D.E. (1999) J. Struct. Biol. 125(2-3), 156-165.
39 Cohen, S.X., Morris, R.J., Fernandez, F.J., Ben Jelloul, M., Kakaris, M., Parthasarathy, V., Lamzin, V.S., Kleywegt, G.J. and Perrakis, A. (2004) Acta Crystallogr. D. Biol. Crystallogr. 60(Pt 12 Pt 1), 2222-2229.
40 Terwilliger, T. (2004) J. Synchrotron Radiat. 11(Pt 1), 49-52.
41 Jones, T.A., Zou, J.Y. and Cowan, S.W. (1991) Acta Crystallogr. A. 47(Pt 2), 110-119.
42 Emsley, P. and Cowtan, K. (2004) Acta Crystallogr. D. Biol. Crystallogr. 60(Pt 12 Pt 1), 2126-2132.
43 Brunger, 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., Rick, L.M., Simonson, T. and Warren, G.L. 1998 Acta Crystallogr. D. Biol. Crystallogr. 54(Pt 5), 905-921.
44 Ness, S.R., de Graaff, R.A., Abrahams, J.P. and Pannu, N.S. (2004) Structure (Cambr.) 12(10), 1753-1761.
45 Brunzelle, J.S., Shafaee, P., Yang, X., Weigand, S., Ren, Z. and Anderson, W.F. (2003) Acta Crystallogr. D. Biol. Crystallogr. 59(Pt 7), 1138-1144.
46 Holton, J. and Alber, T. (2004) Proc. Nat. Acad. Sci. U.S.A. 101(6), 1537-1542.
47 Liu, Z.J., Lin, D., Tempel, W., Praissman, J.L., Rose, J.P. and Wang, B.C. (2005) Acta Crystallogr. D. Biol. Crystallogr. 61(Pt 5), 520-527.
48 Panjikar, S., Parthasarathy, V., Lamzin, V.S., Weiss, M.S. and Tucker, P.A. (2005) Acta Crystallogr. D. Biol. Crystallogr. 61(Pt 4), 449-457.
49 Jiang, J.S. and Lin, Z. (2005) http://asdp.bnl.gov/. Private communications.
50 Fu, Z.Q., Rose, J.P. and Wang, B.C. (2005) Acta Crystallogr. D. Biol. Crystallogr. 61(Pt 7), 951-959.
51 Greer, J. (1981) J. Mol. Biol. 153(4), 1027-1042.
52 Sali, A. and Blundell, T.L. (1993) J. Mol. Biol. 234(3), 779-815.
53 Contreras-Moreira, B., Fitzjohn, P.W. and Bates, P.A. (2002) Appl. Bioinformatics 1(4), 177-190.
54 Pieper, U., Eswar, N., Braberg, H., Madhusudhan, M.S., Davis, F.P., Stuart, A.C., Mirkovic, N., Rossi, A., Marti-Renom, M.A., Fiser, A., Webb, B., Greenblatt, D., Huang, C.C., Ferrin, T.E. and Sali, A. (2004) Nucl. Acids Res. 32 (Database issue), D217-222.
55 Topf, M. and Sali, A. (2005) Curr. Opin. Struct. Biol. 15(5), 578-585.

56 Topf, M., Baker, M.L., John, B., Chiu, W. and Sali, A. (2005) J. Struct. Biol. 149(2), 191-203.
57 Jung, J.W. and Lee, W. (2004) J. Biochem. Mol. Biol. 37(1) 28-34.
58 Laskowski, R.A., Watson, J.D. and Thornton, J.M. (2005) Nucl. Acids Res. 33 (Web Server issue), W89-93.
59 Sasson, O. and Linial, M. (2005) Nucl. Acids Res. 33 (Web Server issue), W81-84.
60 Devos, D. and Valencia, A. (2000) Proteins 41(1), 98-107.
61 Iliopoulos, I., Tsoka, S., Andrade, M.A., Enright, A.J., Carroll, M., Poullet, P., Promponas, V., Liakopoulos, T., Palaios, G., Pasquier, C., Hamodrakas, S., Tamames, J., Yagnik, A.T., Tramontano, A., Devos, D., Blaschke, C., Valencia, A., Brett, D., Martin, D., Leroy, C., Rigoutsos, I., Sander, C. and Ouzounis, C.A. (2003) Bioinformatics 19(6), 717-726.
62 Hegyi, H. and Gerstein, M. (1999) J. Mol. Biol. 288(1), 147-164.
63 Wilson, C.A., Kreychman, J. and Gerstein, M. (2000) J. Mol. Biol. 297(1), 233-249.
64 Rost, B. (2002) J. Mol. Biol. 318(2), 595-608.
65 Tian, W. and Skolnick, J. (2003) J. Mol. Biol. 333(4), 863-882.
66 Tian, W., Arakaki, A.K. and Skolnick, J. (2004) Nucl. Acids Res. 32(21), 6226-6239.
67 Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Nat. Genet. 25(1), 25-29.
68 Thornton, J.M., Todd, A.E., Milburn, D., Borkakoti, N. and Orengo, C.A. (2000) Nat. Struct. Biol. 7 Suppl; 991-994.
69 Orengo, C.A., Todd, A.E. and Thornton, J.M. (1999) Curr. Opin. Struct. Biol. 9(3), 374-382.
70 Moult, J. and Melamud, E. (2000) Curr. Opin. Struct. Biol. 10(3), 384-389.
71 Holm, L. and Sander, C. (1993) J. Mol. Biol. 233(1), 123-138.
72 Gibrat, J.F., Madej, T. and Bryant, S.H. (1996) Curr. Opin. Struct. Biol. 6(3), 377-385.
73 Wallace, L.J., Eiserling, F.A. and Wilcox, G. (1978) J. Biol. Chem. 253(10), 3717-3720.
74 Manjasetty, B.A. and Chance, M.R. (2006) J. Mol. Biol., 360(2): 297–309.
75 Stark, A., Shkumatov, A. and Russell, R.B. (2004) Structure (Cambr.) 12(8), 1405-1412.
76 Stark, A. and Russell, R.B. (2003) Nucl. Acids Res. 31(13), 3341-3344.
77 Wallace, A.C., Laskowski, R.A. and Thornton, J.M. (1996) Protein Sci. 5(6), 1001-1013.
78 Kleywegt, G.J. (1999) J. Mol. Biol. 285(4), 1887-1897.
79 Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T. and Ben-Tal, N. (2005) Nucl. Acids Res. 33 (Web Server issue), W299-302.

80 Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E. and Ben-Tal, N. (2003) Bioinformatics 19(1), 163-164.
81 Laskowski, R.A. (1995) J. Mol. Graph. 13(5), 323-330, 307-328.
82 Zhan, C., Fedorov, E.V., Shi, W., Ramagopal, U.A., Thirumuruhan, R., Manjasetty, B.A., Almo, S.C., Fiser, A., Chance, M.R. and Fedorov, E. (2005) Acta Crystallogr. F F61, 959-963.
83 Manjasetty, B.A., Delbruck, H., Pham, D.T., Mueller, U., Fieber-Erdmann, M., Scheich, C., Sievert, V., Bussow, K., Niesen, F.H., Weihofen, W., Loll, B., Saenger, W. and Heinemann, U. (2004) Proteins 54(4), 797-800.
84 Manjasetty, B.A., Powlowski, J. and Vrielink, A. (2003) Proc. Nat. Acad. Sci. U.S.A. 100(12), 6992-6997.
85 Turnbull, A.P., Kummel, D., Prinz, B., Holz, C., Schultchen, J., Lang, C., Niesen, F.H., Hofmann, K.P., Delbruck, H., Behlke, J., Müller, E.-C., Jarosch, E., Sommer, T. and Heinemann, U. (2005) Embo J. 24(5), 875-884.
86 Rajashankar, K.R., Bryk, R., Kniewel, R., Buglino, J.A., Nathan, C.F. and Lima, C.D. (2005) J. Biol. Chem. 280(40), 33977-33983.
87 Yang, Z., Savchenko, A., Yakunin, A., Zhang, R., Edwards, A., Arrowsmith, C. and Tong, L. (2003) J. Biol. Chem. 278(10), 8804-8808.
88 Kuznetsova, E., Proudfoot, M., Sanders, S.A., Reinking, J., Savchenko, A., Arrowsmith, C.H., Edwards, A.M., Yakunin, A.F. (2005) FEMS Microbiol. Rev. 29(2), 263-279.
89 Powell, K.D. and Fitzgerald, M.C. (2004) J. Comb. Chem. 6(2), 262-269.
90 Mathur, S., Hassel, M., Steiner, F., Hollemeyer, K. and Hartmann, R.W. (2003) J. Biomol. Screen 8(2), 136-148.
91 Yakunin, A.F., Yee, A.A., Savchenko, A., Edwards, A.M. and Arrowsmith, C.H. (2004) Curr. Opin. Chem. Biol. 8(1), 42-48.
92 Yee, A., Pardee, K., Christendat, D., Savchenko, A., Edwards, A.M. and Arrowsmith, C.H. (2003) Acc. Chem. Res. 36(3), 183-189.
93 Xie, L. and Bourne, P.E. (2005) PLoS Comput. Biol. 1(3), e31.
94 Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucl. Acids Res. 25(17), 3389-3402.
95 Livingstone, C.D. and Barton, G.J. (1993) Comput. Appl. Biosci. 9(6), 745-756.
96 Barton, G.J. (1993) Protein Eng. 6(1), 37-40.
97 DeLano, W.L. (2002) http://www.pymol.org.