Journal of Structural and Functional Genomics **6**: 225–232, 2005.
DOI 10.1007/s10969-005-6827-0

New York-Structural GenomiX Research Consortium (NYSGXRC): a large scale center for the protein structure initiative

Jeffrey B. Bonanno[1], Steven C. Almo[2], Anne Bresnick[2], Mark R. Chance[2], Andras Fiser[2], S. Swaminathan[3], J. Jiang[3], F. William Studier[3], Lawrence Shapiro[4], Christopher D. Lima[5], Theresa M. Gaasterland[6], Andrej Sali[7], Kevin Bain[1], Ingeborg Feil[1], Xia Gao[1], Don Lorimer[1], Aurora Ramos[1], J. Michael Sauder[1], Steven R. Wasserman[1], Spencer Emtage[1], Kevin L. D'Amico[1] & Stephen K. Burley[1,]*

*[1]Structural GenomiX, Inc., 10505 Roselle Street, San Diego, CA 92121, USA; [2]Albert Einstein College of Medicine, Bronx, NY 10461, USA; [3]Department of Biology, Brookhaven National Laboratory, Upton, NY 11973, USA; [4]Department Biochemistry, Columbia University, New York, NY 10032, USA; [5]Structural Biology Program, Sloan-Kettering Institute, New York, NY 10021, USA; [6]University of California, San Diego, CA 92093, USA; [7]University of California, San Francisco, CA 94143, USA; *Author for correspondence (e-mail: sburley@stromix.com; fax +1-858-558-6079)*

Received 25 October 2004; accepted in revised form 2 May 2005

Key words: crystallography, NYSGXRC, protein structure initiative, structural genomics

Abstract

Structural GenomiX, Inc. (SGX), four New York area institutions, and two University of California schools have formed the New York Structural GenomiX Research Consortium (NYSGXRC), an industrial/academic Research Consortium that exploits individual core competencies to support all aspects of the NIH-NIGMS funded Protein Structure Initiative (PSI), including protein family classification and target selection, generation of protein for biophysical analyses, sample preparation for structural studies, structure determination and analyses, and dissemination of results. At the end of the PSI Pilot Study Phase (PSI-1), the NYSGXRC will be capable of producing 100–200 experimentally determined protein structures annually. All Consortium activities can be scaled to increase production capacity significantly during the Production Phase of the PSI (PSI-2). The Consortium utilizes both centralized and de-centralized production teams with clearly defined deliverables and hand-off procedures that are supported by a web-based target/sample tracking system (SGX Laboratory Information Data Management System, LIMS, and NYSGXRC Internal Consortium Experimental Database, ICE-DB). Consortium management is provided by an Executive Committee, which is composed of the PI and all Co-PIs. Progress to date is tracked on a publicly available Consortium web site (http://www.nysgxrc.org) and all DNA/protein reagents and experimental protocols are distributed freely from the New York City Area institutions. In addition to meeting the requirements of the Pilot Study Phase and preparing for the Production Phase of the PSI, the NYSGXRC aims to develop modular technologies that are transferable to structural biology laboratories in both academe and industry. The NYSGXRC PI and Co-PIs intend the PSI to have a transforming effect on the disciplines of X-ray crystallography and NMR spectroscopy of biological macromolecules. Working with other PSI-funded Centers, the NYSGXRC seeks to create the structural biology laboratory of the future. Herein, we present an overview of the organization of the NYSGXRC and describe progress toward development of a high-throughput Gene → Structure platform. An analysis of current and projected consortium metrics reflects progress to date and delineates opportunities for further technology development.

NYSGXRC scientific organization

The NYSGXRC (S.K. Burley, PI; SGX) is comprised of members from Albert Einstein College of Medicine (AECOM; S. Almo, Institutional Co-PI; Additional Co-PIs: A. Bresnick, M.R. Chance, and A. Fiser), Brookhaven National Laboratory Department of Biology (BNL; S. Swaminathan, Institutional Co-PI; Additional Co-PIs: J. Jiang and F.W. Studier), Columbia University (CU; L. Shapiro, Institutional Co-PI), Sloan-Kettering Institute (SKI; C.D. Lima, Institutional Co-PI), the University of California at San Diego (UCSD; T. Gaasterland, Institutional Co-PI), the University of California at San Francisco (UCSF; A. Sali, Institutional Co-PI), and Structural GenomiX, Inc (SGX; J.B. Bonanno, Institutional Co-PI). The NYSGXRC has improved efficiency through centralization of early stage activities, which lend themselves to economies of scale, at SGX, while retaining a de-centralized business model for later stage activities that are more appropriately distributed. All of the constitutive activities of a PSI-1 center required to proceed from selection of targets to dissemination of annotated structures, are represented within the NYSGXRC and are described below.

Centralized target selection

Sali (UCSF) leads NYSGXRC target selection efforts. Sali coordinates the activities of his own laboratory for identifying priority targets on the basis of homology modeling impact plus activities within the Gaasterland (UCSD) and Fiser (AECOM) laboratories for screening annotated genomes and protein databases for proteins that meet certain biological criteria for NYSGXRC target selection.

NYSGXRC target proteins are selected on a consensus basis with input from the member institutions. All NYSGXRC targets are publicized *via* the NYSGXRC web site (http://www.nysgxrc.org), and uploaded to the central target tracking system maintained by the Protein Data Bank (PDB, http://www.rcsb.org/pdb).

Centralized target tracking

Fiser (AECOM) leads NYSGXRC target tracking efforts and is responsible for maintenance of the internal consortium experimental database (ICE-DB) and interactions with TargetDB and PepcDB, which are maintained by the RCSB (http://www.rcsb.org/pdb).

Centralized protein expression and purification

SGX is responsible for all gene cloning and expression/purification of NYSGXRC target proteins. At SGX, the NYSGXRC Project Team includes Staff Scientists and Research Assistants drawn from the Departments of Molecular Biology, Protein Purification, Protein Characterization, Bioinformatics, and Data Management.

The highly automated, industrial scale platform at SGX is being used to prepare expression clones, and express and purify NYSGXRC target proteins. Soluble, purified proteins produced at SGX on the 10+ mg scale are shared equally among De-Centralized Crystallization Teams at AECOM, BNL, CU, and SKI (see below). Progress on production of each target protein is tracked by the Information Data Management system at SGX, and these data together with experimental protocols are automatically uploaded to http://www.nysgxrc.org on a weekly basis.

For each target protein yielding diffraction-quality crystals, SGX furnishes purified Se-Met protein on the 5+ mg scale (as appropriate for Met-containing proteins) to the relevant De-Centralized Crystallization Team for production of Se-Met crystals.

For all targets cloned and purified by SGX, materials derived from the experimental process (PCR products, plasmids, expression strains, and archival quantities of protein), are separately stored at −80 °C for distribution to qualified scientific investigators and/or transfer to a central archive operated by the NIH-NIGMS.

Within SGX, transfers of reagents and protocols are monitored electronically to ensure facile communication and orderly exchange. The same procedures are used to monitor hand-offs of purified proteins, plasmid DNAs, and protocols between SGX and the New York City Area institutions.

The results of cloning, expression/solubility testing, and purification efforts by SGX on NYSGXRC targets during PSI-1 Years 3 and 4 are provided in Tables 1 and 2.

Centralized protein characterization

For every soluble NYSGXRC target protein purified by SGX, the Protein Characterization

Table 1. Single pass SGX Gene → Protein pipeline success rates for the SGX_EC bacterial target list of enzyme-related sequences.

Process step	SGX_EC throughput PSI-1 (12/5/03–10/15/04)		
	Actual SGX_EC throughput	% Success of process step	% Success from 1511 targets
Targets attempted	1511		
PCR products	1292	85.5	85.5
Plasmids	1243	96.2	82.3
Transformed	1164	93.6	77.0
Expressed	858	73.7	56.8
Soluble	567	66.1	37.5
Purified	420	74.1	27.8
Passed QA/QC	384	91.4	25.4

Table 2. Current metrics for SGX-produced target proteins.

NYSGXRC target proteins produced at SGX	
Total # protein deliveries to NY	1044
Unique protein deliveries	786
Native proteins resupplied	32
Se-Met labeled proteins deliveries	226
Proteins crystallized	203
Structures of delivered proteins	83

N.B.: A large number of experiments are ongoing in New York area laboratories (242 unique proteins were delivered between July 15th 2004 and October 25th 2004). Current crystallization and structure determination metrics represent snapshots subject to change during the remainder of PSI-1 Year 5.

Group at SGX conducts comprehensive biophysical characterization studies, including mass spectrometry for construct verification and protein purification QA/QC, analytical gel filtration, domain mapping by limited proteolysis combined with mass spectrometry (LP/MS), peptide mapping of post-translational modifications via mass spectrometry (where necessary), and UV/vis. absorbance spectroscopy to identify possible bound co-factors.

Experimental and computational aspects of this part of the research program (directed by Burley) are aimed at producing a comprehensive public domain database of biophysical parameters, both measurable and calculable, for correlation with outcomes in crystallization and structure determination efforts. Results are accumulated and data analyzed by the Data Management group at SGX. Statistically significant correlations will be disseminated *via* publications and http://www.nysgxrc.org.

Also reported in this issue of the *Journal of Structural and Functional Genomics*, is an initial assessment of our large scale LP/MS analysis of NYSGXRC target proteins produced for crystallization trials by New York area laboratories [1]. Members of the NYSGXRC have very considerable experience in using limited proteolysis combined with mass spectrometry and other biophysical tools to identify protein samples that represent promising candidates for successful crystallization trials.

The Burley and Chait laboratories at The Rockefeller University together pioneered application of limited proteolysis combined with mass spectrometry (LP/MS) to map domain boundaries in globular proteins [2]. At SGX, this information is being used prospectively to design new expression constructs in a salvage pathway for protein targets resisting crystallization attempts. Experience has shown that the presence of flexible regions within N- and/or C-terminal segments encoded by expression constructs often fail to yield diffraction quality crystals. This is often the result of construct design based on imperfect bioinformatic analyses of globular domain boundaries for targets of unknown structure.

As a critical first step in establishing a salvage pathway for NYSGXRC targets, large scale LP/MS efforts were carried out at SGX for NYSGXRC targets during PSI-1 years 3 and 4 [1]. LP/MS results for 164 target proteins documented that ~50% were totally resistant to proteolysis (with the exception of occasional affinity tag removal), ~45% were partially degraded to a stable domain, and ~5% were rapidly degraded to oligopeptide fragments. Not surprisingly, the proteolytically resistant NYSGXRC targets proved to be good candidates for crystallization, with 24 of 86 (27%) such targets yielding three-dimensional (3-D) structures to date. The second group of NYSGXRC targets, which demonstrated minimal proteolysis yielding a stable domain, often produced crystals with suboptimal diffraction limits (i.e., worse that 3.5 Å resolution). Thus far, only 7 of 75 (9%) such targets have yielded 3-D structures. Finally, none of the NYSGXRC targets exhibiting massive proteolysis have crystallized.

These systematic observations combined with earlier anecdotal evidence based on small scale applications of LP/MS domain mapping form an important element of the NYSGXRC system for salvaging structure determination targets. Large scale subcloning of NYSGXRC targets exhibiting minimal proteolysis with new N- and/or C-termini and expression of the resulting truncations in *E. coli* are nearly complete. Truncated forms of >200 salvaged targets will be passed to the De-Centralized Crystallization teams in the New York area laboratories (see below) in late 2004.

Centralized metalloprotein detection

For every soluble NYSGXRC target protein, SGX furnishes ~0.1 mg to Chance et al. (AECOM) for automated X-ray absorption spectroscopy (XAS) detection of bound metal ions using NSLS Beamline X9B. The results of these measurements are uploaded to http://www.nysgxrc.org. As appropriate, phasing of diffraction measurements from S-Met crystals using intrinsic metal ions is attempted in parallel with, or *in lieu* of, Se-Met structure determinations. Approximately 1/3 of NYSGXRC target proteins contain an intrinsic metal ion, which suggests that full reliance on this opportunistic phasing strategy should further increase the efficiency of the NYSGXRC Gene → Structure platform.

De-centralized protein crystallization

Table 2 summarizes the results of crystallization efforts with proteins furnished by SGX to AECOM, BNL, CU, and SKI during PSI-1 Years 3 and 4.

Almo (AECOM), Swaminathan (BNL), Shapiro (CU), and Lima (SKI) are responsible for coordinating crystallization trials with NYSGXRC target proteins at their respective New York area institutions. This activity is distributed evenly among the four experimental crystallography laboratories, each one utilizing identical robotic platforms. SGX has already implemented a cost-effective, automated approach to 96-well format sitting drop crystallization trials using TECAN robotic devices. This strategy has been replicated within each of the NYSGXRC crystallization laboratories.

With installation of identical robotic crystallization platforms at each member institution, protocols for both initial screening and optimization can be used uniformly for every target protein across multiple geographically distributed laboratories. SGX and Lima (SKI) have independently developed systematic approaches to these screening and optimization methods, which are currently being evaluated for amalgamation and adoption throughout the NYSGXRC.

Partially centralized synchrotron data measurement

New York area consortium members at AECOM, BNL, CU, and SKI oversee synchrotron data measurements using beamlines X4A, X9A, X9B, X12B, X12C, X25, and X29 at the National Synchrotron Light Source (NSLS). Both MAD/SAD (with Se-Met or intrinsic metals) and MIRAS/SIRAS methods are used for phasing.

During PSI-1 Years 3 and 4, members of the NYSGXRC also made significant use of the SGX-CAT undulator beamline at the Argonne National Laboratory Advanced Photon Source (APS). SGX-CAT staff members coordinate NYSGXRC access, which amounts to multiple days of beamtime/month. Crystals are frozen by each of the De-Centralized Crystallization Teams and shipped by courier for data collection at SGX-CAT.

De-centralized structure determination/PDB coordinate deposition

A summary of 161 structures determined by the NYSGXRC (as of October 25th 2004) can be found at http://www.nysgxrc.org.

Almo (AECOM), Swaminathan (BNL), Shapiro (CU), and Lima (SKI) are responsible for structure determinations with the results of synchrotron measurements on their respective targets, and timely PDB deposition as required by the NIGMS.

Jiang (BNL) provides the NYSGXRC crystallography laboratories with access to the Automated Structure Determination Package (ASDP), which is currently under development at BNL. This web-based ensemble of virtually all public domain crystallographic software can be used for automated structure determination, including electron density map interpretation, manual structure determination, refinement, and PDB deposition.

Almo *et al.* (AECOM) provide the NYSGXRC crystallography laboratories with access to a six-dimensional rotation/phased translation server, BRUTEPTF. This tool is used to overcome limitations of poor phasing (typically with heavy atom derivatives in the absence of diffraction quality Se-Met crystals) when a low-accuracy homology model of the target protein is available. This approach is particularly useful in cases with high non-crystallographic symmetry. Burley et al. at The Rockefeller University first described use of the BRUTEPTF strategy to determine a structure containing 16 RNA recognition motifs for which homology models were available. 8-fold non-crystallographic symmetry averaging rendered an uninterpretable electron density map fully traceable, revealing clear signal for all portions of 8 polypeptide chains and 8 bound RNA oligonucleotides [3].

Centralized comparative protein structure modeling

The homology modeling impact of 112 selected NYSGXRC structures is summarized in Table 3.

Sali (UCSF) leads the NYSGXRC homology modeling effort. Automated homology modeling with MODWEB has been fully implemented, and is now being used routinely by NYSGXRC members, other PSI Centers, and researchers throughout the world. The MODWEB software tools have been reviewed in detail [4].

Centralized structure annotation

Fiser (AECOM) is responsible for leading the NYSGXRC annotation effort. A system for automated structure annotation is currently under development by Fiser et al. NYSGXRC members are routinely using an initial prototype of this system, which permits automated preparation of summary structure reports.

Table 3. Comparative protein structure results for 112 representative NYSGXRC structures as modeling templates including breakdown by sequence identity of the template to the model.

Total number of modeled sequences	84,491
Average number of models per structure	754
Fraction of modeled sequences <30% ID	~90.7%
Fraction of modeled sequences 30–50% ID	~7.7%
Fraction of modeled sequences >50% ID	~1.6%

De-centralized structure publication and collaborative functional studies

More than 70 publications have come either directly or indirectly from NYSGXRC PSI-related activities (as of June 30th 2004).

Almo (AECOM), Swaminathan (BNL), Shapiro (CU), and Lima (SKI) are separately responsible for publication of the results of NYSGXRC structure determinations at their respective institutions. The NYSGXRC has cleared much of its publication backlog for structures determined during PSI-1 Years 1 and 2. Development of an automated publication preparation system, utilizing mmCIF files for efficient transfer of experimental/calculable results for each newly structure, is underway in Fiser's laboratory. NYSGXRC members have contributed to numerous functional studies arising from PSI-1 activities. This work has been funded various non-PSI sources, including a supplemental PSI grant from NIH-NIGMS to AECOM scientists.

Technology development

The NYSGXRC has actively pursued a number of technology developments aimed at overcoming bottlenecks identified in our Gene → Structure process. Three selected programs are highlighted below.

Auto-induction media for E. coli *expression*

Studier *et al.* (BNL) have formulated bacterial growth media in which expression strains can grow uninduced to relatively high cell densities and are then induced automatically without human intervention. Cell densities attained in these auto-inducing cultures have produced 10-fold more target protein per unit volume of culture than with standard IPTG induction protocols. Auto-induction also allows many cultures to be inoculated in parallel and induced simply by growth to saturation, making auto-induction a powerful tool for screening clones for expression and solubility in an automated setting. The initial test of a similar medium formulated for labeling proteins with Se-Met by auto-induction produced 5 mg of purified, fully labeled protein from 500 ml of culture, which sufficed for structure determination. We continue to evaluate the effectiveness of these media and protocols by conducting auto-induction expression/solubility testing in parallel with conventional IPTG induction in LB medium.

Auto-induction protocols and recipes for producing proteins were distributed to more than 150 laboratories and described at several national and international scientific meetings before publication [5]. A patent application has been submitted, products based on this work have been commercialized by Novagen as Overnight Express Autoinduction Systems, and this work has been recognized by an R&D 100 Award from R&D Magazine as one of the 100 most significant innovations commercialized in 2004. Auto-induction provides substantial improvements in convenience and efficiency of high-throughput protein production and is used routinely by NYSGXRC and several other PSI-1 structural genomics centers for parallel screening, production of 10–100 mg amounts in shake flasks, and labeling with Se-Met.

A novel, facile system for protein expression

Lima et al. (SKI) developed a novel SUMO-based protein expression system (US Patent Number 6, 872, 551 B2) for high-throughput cloning and protein expression which has been utilized within the NYSGXRC to rapidly clone, affinity purify, and proteolytically liberate NYSGXRC targets from the His_6-SMT3 fusion partner for further purification. This proprietary vector has been Topo-adapted for rapid (5 min) topoisomerase mediated directional flap-ligation to facilitate efficient, high-throughput restriction enzyme independent cloning for DNA generated from the coding regions containing our protein targets. This measure alleviates the need for engineering longer than necessary primers to contain convenient restriction enzymes, the need to enzymatically digest and purify the DNA inserts, and the need for long, low temperature ligations traditionally utilized to clone DNA fragments into expression vectors. A description of the commercially available system can be found by searching the Invitrogen website for the keyword “SUMO”, or by visiting: https://catalog.invitrogen.com/index.cfm?fuseaction = viewCatalog.viewProductDetails &sku = &productDescription = 1045&CMP = LEC-GCMSSEARCH&HQS = sumo&

A mammalian cell expression system for disulfide rich proteins

Production of properly folded mammalian proteins in heterologous systems is frequently challenging. Proteins can be mis-folded due to lack of cognate chaperones or absence of the proper cellular machinery for post-translational modification. Shapiro *et al.* (CU) have developed an efficient system for selection of highly expressing stable mammalian cells, based on fluorescence detection of a co-expressed marker, the green fluorescent protein (GFP).

The coding sequence for the gene of interest is placed under the control of a strong constitutive promoter (such as a promoter element derived from cytomegalovirus, CMV). Downstream, after the termination codon for the gene of interest, an internal ribosome entry site (IRES) is followed by the coding sequence for GFP. Transcription from this construct produces a single bicistronic messenger RNA encoding both genes. The IRES permits binding of the ribosome at the initiation site for production of GFP. Thus, two separate proteins – the target and GFP – are translated from the same messenger RNA, and expression levels of these two proteins are thereby coupled. This system enables efficient selection of highly expressed targets by monitoring the fluorescence intensity of mammalian cells expressing variable amounts of GFP. Use of fluorescence-activated cell sorting (FACS) technology permits rapid selection of either clonal or non-clonal populations of highly expressing cells. Shapiro *et al.* have used this method to prepare non-clonal HEK-293 cell lines for production of the highly-disulfide-bonded proteins of the resistin hormone family [6]. Protein expressed in mammalian cells using this system permitted determination of high-resolution structures of resistin and RELM-β [7].

Current and projected consortium metrics

During PSI-1 Years 1 and 2, the NYSGXRC relied on a distributed network for molecular biology/protein production. During PSI-1 Years 3 and 4, NYSGXRC molecular biology/protein production and protein characterization was centralized at SGX, where an industry leading high-throughput platform was made available to the PSI.

Centralization of protein production at SGX has been, and will continue to be, an enormous strength for the consortium. To date, the NYSGXRC has placed 943 unique purified, soluble proteins in crystallization trials. Evidence of the quality of these preparations is the fact that 327 have crystallized and 161 have, thus far, produced X-ray crystal structures (as of October 25th 2004).

Gene → structure pipeline analysis

Since protein production was moved to SGX, 1044 purified soluble protein preparations (unique and resupply; Table 2) were shipped to the NY area laboratories. Most of these targets were selected from lists of bacterial enzymes, hypothetical bacterial proteins, and proteins of medical relevance. Despite the success of these activities and earlier NYSGXRC structure determination programs, many of NYSGXRC target lists emerged as an amalgamation of target lists from one or more member institutions, which do not provide a consistent data set with which to effectively evaluate the NYSGXRC Gene → Structure platform as it is now implemented.

To evaluate current NYSGXRC processes, we have examined our progress with one large NYSGXRC target list, which has recently completed a first pass through SGX molecular biology, fermentation, and protein purification modules. This SGX_EC target list was developed by Lima and Sali from sequences identified in the Enzyme Classification database that represent proteins of unknown structure. After applying appropriate bioinformatic filters to prioritize target genes from a diverse set of organisms, ~1500 target genes were selected from ~40 bacterial genomes available at SGX. Results of the first pass through the SGX Gene → Protein pipeline for SGX_EC targets are presented in Table 2. Efforts in crystallization and structure determination for proteins from this target list continue. Most of these proteins were in crystallization trials at the time of submission (October 25th 2004).

From all NYSGXRC target lists, 1044 purified soluble proteins have been shipped from SGX to the New York area laboratories (9/1/02–10/15/04). The average amount of each protein delivered was ~37 mg at concentrations of > 10 mg/ml. 786 of the 1044 samples were unique targets, 32 were resupplies of native protein (typically indented for continuation of crystallization optimization efforts), and 226 were Se-Met labeled samples produced for targets already yielding diffraction quality crystals. To date, 83 structures have been determined for proteins produced at SGX. Table 2 summarizes these metrics. SGX protein production levels during PSI-1 Year 4 were scaled up significantly as compared to PSI-1 Year 3 (224 unique proteins produced in PSI-1 Year 3; 450 produced in PSI-1 Year 4) due largely to ramp-up in molecular biology efforts and more efficient protocols developed for purification methods. A corresponding increase in structure determinations was realized. The total number of structures determined by the NYSGXRC for PSI-1 Years 1–3 was 68; an additional 74 structure determinations were conducted in PSI-1 Year 4 totaling 142. An additional 19 structures have been determined thus far in PSI-1 Year 5 (for a summary of NYSGXRC structures, see http://www.nysgxrc.org).

Comparative protein structure modeling impact

On average, each NYSGXRC structure determination permits high accuracy homology modeling of ~70 protein sequences, which are related to the experimental structure (modeling template) at 30% or greater sequence identity (Table 3).

In addition, each newly determined NYSGXRC structure provides structural information for another ~685 more distantly related protein sequences (Table 3). These homology models contain inaccurate regions, because of errors in the alignment of the sequence of the experimental structure (modeling template) with the sequence of the protein to be modeled.

Despite such errors, lower accuracy homology models permit many sequences to be assigned to previously characterized fold families, which can be used to guide additional structure determination efforts required to provide one experimental structure for each 30% identity sequence family. BRUTEPTF, described above, has already shown considerable promise in this context.

Lower accuracy homology models also provide structural information that can be used to infer biological or biochemical function, identify putative active site residues, and help biologists design expression constructs, with which to obtain soluble truncated forms of proteins that cannot be expressed as full-length open reading frames.

These modeling results have been calculated using MODWEB, and the resulting datasets have been deposited in MODBASE. A more detailed and automatically updated version of the table including links to the resulting MODBASE datasets can be found at http://salilab.org/modbase/models_nysgxrc_latest.html. All protein sequences in SwissProt/TrEMBL that are detectably related to the NYSGXRC structures can be viewed in MODBASE. This feature facilitates detection of remote relationships and functional annotation of proteins previously annotated as hypothetical. The statistics in the various sequence identity ranges signify the number of modeled sequences with at least 75 amino acid residues of the template sequence aligned to the target sequence.

NYSGXRC PSI-1 year 5 projections

A total of 74 structures were determined by the NYSGXRC during PSI-1 Year 4. Given our consortium budget (9/1/2003–8/31/2004) of $8,113,500 (including administrative supplemental monies awarded in 2003), the average cost per structure was $109,641. This estimate includes both our Gene → Structure activities and ongoing NYSGXRC technology development efforts.

The NYSGXRC is on track to meet its goal of determining >100 structures during PSI-1 Year 5. Given the NYSGXRC PSI-1 Year 5 budget of $8,142,625 (September 1st 2004–August 31st 2005) we anticipate that the cost per structure will be further reduced to ~$80,000 or less.

Table 4. NYSGXRC structure determination cost analysis and projections.

Years	# Structures determined	Period budget	Average cost per structure
PSI-1 1–3	68[a]	~$15,000,000	~$220K[a]
PSI-1 four	74[a]	~$8,113,500	~$110K[a]
PSI-1 five	~100[b]	~$8,142,625	~$81K[b]
PSI-2 one	170–200[b]	12,000,000[b]	~$70K[b]

[a]Indicates actual value.
[b]Indicates projected value.

NYSGXRC projections regarding the production phase of the Protein Structure Initiative (PSI-2) suggest that further cost reductions will bring the average cost per structure to $60,000–$70,000 in PSI-2 Year 1. Efficiency and throughput improvements will come from full implementation of NYSGXRC salvage pathways (Table 4).

References

1. Gao, X., Adams, J., Bain, K., Bonanno, J.B., Buchanan, M., Emtage, S., Lorimer, D., Marsh, C., Reynes, J.A., Sauder, M., Schwinn, K., Thai, C. and Burley, S.K. (2004) *J. Struct. Funct. Genet.* (full citation to be provided at the proof stage, same issue of JSFG).
2. Cohen, S.L., Ferré-D'Amaré, A.R., Burley, S.K. and Chait, B.T. (1995) *Protein Sci.* **4**, 1088–1099.
3. Deo, R.C., Bonanno, J.B., Sonenberg, N. and Burley, S.K. (1999) *Cell* **98**, 835–845.
4. Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V.A., Pieper, U., Stuart, A.C., Marti-Renom, M.A., Madhusudhan, M.S., Yerkovich, B. and Sali, A. (2003) *Nucleic Acids Res.* **31**, 3375–3380.
5. Studier, F.W. (2005) *Protein Expr. Purif.* **41**, 207–234.
6. Mancia, F., Patel, S.D., Rajala, M.W., Scherer, P.E., Nemes, A., Ira Schieren, I., Hendrickson, W.A. and Shapiro, L. (2004) *Structure* **12**, 1355–1360.
7. Patel, S.D., Rajala, M.W., Rossetti, L., Scherer, P.E. and Shapiro, L. (2004) *Science* **304**, 1154–1158.