Chapter 1

FIVE YEARS OF INCREASING STRUCTURAL BIOLOGY THROUGHPUT- A RETROSPECTIVE ANALYSIS

Enrique Abola[1], Dennis D. Carlton[1], Peter Kuhn[2] and Raymond C. Stevens [1]
The Scripps Research Institute, Departments of Molecular Biology [1] and Cell Biology [2]. 10550 North Torrey Pines Road, La Jolla, CA 92037, USA

1. INTRODUCTION

1.1 Structural biology and genomics

The completed sequencing and initial characterization of the human genome in 2001 (Lander et al 2001; Venter et al 2001) and that of other organisms such as *Drosophila melanogaster* (Adams et al 2000) and the SARS Corona Virus (Marra et al 2003), have educated us on the vast complexity of the proteome. Full genome characterization efforts highlight how critical it is to understand at a molecular level all of the protein products from multiple organisms. An important issue for addressing the molecular characterization challenge is the need to quickly and economically characterize normal and diseased biological processes in order to understand the basic biology and chemistry of the systems and to facilitate the discovery and development of new therapeutic and diagnostic protocols. In order to fully characterize the proteins at the molecular level, three-dimensional protein structure determination has proven to be invaluable, complementing biological and biochemical information from other types of experiments. Structural information is also the ultimate rational drug design tool, with the potential to save an estimated 50% of the cost of drug discovery (Stevens 2004). However, the best means

H. Jhoti and A. Leach (eds.), Structure-based Drug Discovery, 1–26.

by which to attain structural knowledge is a topic of controversy. The traditional approach was a complex and labor-intensive process in which one protein or complex was studied at a time. The alternative is a high throughput (HT), discovery-oriented approach wherein entire families, pathways or genomes are characterized. Benefits include the economy of scale, the speed of mass production, and a dramatic increase in discovery rates through the systematic collection and analysis of data. Prior to the late 1990's, the technologies and approaches were too slow and unreliable to allow for such larger scale analyses.

In the past, we have reviewed some of the technology developments in miniaturizing and streamlining structure determination pipelines (Stevens 2004; Abola et al 2000). For this chapter, we summarize the input and output of several structural genomics efforts that have validated new technology efforts over the first 5 years of the HT structural biology era. These technologies have been used by various HT pipelines that have contributed to the determination of over 1600 new structures, a high percentage of which were novel folds, and 70% had less than 30% identity to any other protein in the Protein Data Bank (PDB) at the time of release. As an example of the implementation of the HT pipeline, we discuss in some detail the specific approach of the Joint Center for Structural Genomics (JCSG) that we have been involved in.

1.2 Protein structure initiative

In 1999, several initiatives in Japan, Europe, and the United States were created to investigate the feasibility of HT structural biology, structural genomics, and increased throughput structure based drug dicovery (Table 1-1). For this chapter, we will focus on the efforts in the United States since we are most familiar with them, and they have recently completed the first phase of their efforts (PSI-1). However, we would be remiss in not mentioning a number of critical players in the global effort. Japan has prehaps made the largest financial global investment in this area of structural genomics at the Riken Genomics Center through the creation of the NMR Farm, and development of such innovative technologies as cell-free protien expression (Kigawa et al 2004). In Europe, several efforts in the UK (e.g. SPINE) and Germany (e.g. Protein Structure Factory) were both early movers in this area and have contributed significantly to the field.

More recent efforts included MepNet and the Structural Genomics Consortium (Toronto, Oxford, and Sweden). In addition to academic and government-led efforts, a number of structural genomics companies were set up during this period and have also contributed to the rapid growth of the

Table 1-1. History of Structural Genomics.

Feb 1995	LBNL structural genomics expression/crystallization technology development initiated
1995	Proposal of structural genomics projects in Japan
Jan 1997	The workshop on Structural Genomics (Argonne, IL, USA)
Apr 1997	Start of structural genomics pilot project at RIKEN Institute
1997	Initiating study of structural genomics at DOE and NIGMS/NIH in USA
1998	Start of the initial pilot projects in Germany, Canada, and USA
Feb 1999	Formation of the Berlin Protein Structure Factory
Feb 1999	Formation of Syrrx (previously called Agencor)
Jun 1999	Call for grant applications for NIGMS/NIH pilot projects (PSI-1)
Dec 1999	Formation of Structural GenomiX (previously called Protarch)
Dec 1999	Formation of Astex Technology
Apr 2000	First International Structural Genomics Meeting (Hinxton, UK)
Aug 2000	Formation of Affinium Pharmaceuticals (previously called Integrative Proteomics)
Sep 2000	Structural Genomics: From Gene to Structure to Function (Cambridge, UK)
Sep 2000	Start of the NIGMS Protein Structure Initiatives in USA with seven Centers
Nov 2000	1st International Conference on Structural Genomics 2000 (ICSG 2000) (Yokohama, Japan)
Apr 2001	Second International Structural Genomics Meeting (Airlie House, USA) - Start of International Structural Genomics Organization (ISGO)
Jun 2001	Formation of Plexxikon
Sep 2001	Start of the new two centers for NIGMS Protein Structure Initiatives in USA (9 total)
Mar 2002	Start of the European drive for post-genome research, Structural Proteomics in Europe (SPINE)
Apr 2002	Start of the National Project on Protein Structural and Functional Analyses in Japan
Oct 2002	2nd ISGO International Conference on Structural Genomics (ICSG 2002) (Berlin, Germany)
Early 2003	RIKEN– 100th structure solved at Riken deposited in PDB
April 2003	Formation of The Structural Genomics Consortium (SGC)
April 2004	RFA for next generation structural genomics centers in USA
November 2004	3rd ISGO International Conference on Structural Genomics (ICSG 2004) (Washington, D.C. USA)
Feb 2005	PSI 1000th structure milestone achieved
July 2005	Start of PSI-2 with 4 Large Scale and 6 Specialized Centers

field (e.g. Syrrx, Structural GenomiX, Astex Therapeutics, Affinium Pharmaceuticals, Plexxikon).

As one of the initiators of the structural genomics movement in the late 1990's, the National Institute of General Medical Sciences (NIGMS) created the Protein Structure Initiative (PSI), a national program with the long-range goal of making three-dimensional, high-resolution protein structures obtainable from knowledge of their corresponding DNA sequences. Completed in the summer of 2005, the pilot phase (referred here as PSI-1) supported a 5-year effort with 9 pilot centers throughout the U.S. to evaluate "if" HT structural biology pipelines could be established and then incorporated into scaleable production pipelines capable of solving hundreds of protein structures per year. Early in 2005 the NIGMS PSI announced its first major milestone, that the combined output of the nine PSI centers had exceeded 1,000 structures.

In Phase II (referred here as PSI-2), NIGMS is providing additional funding for four large-scale centers that will scale-up their production lines to provide another 3,000 to 5,000 structures (NIH 2005). A critical component of the second phase will be the careful target selection procedures that will be managed by the NIGMS PSI-2 Network. A part of this coordinated target selection management is the focus on biomedically relevant protein structures. In addition to the four production centers, six technology development centers have been created to continue the development of innovative technologies for the more challenging problems including studies on membrane proteins, large protein assemblies, and the more difficult eukaryotic proteins.

2. NEW TECHNOLOGIES IN PSI-1 PIPELINES

2.1 High-throughput structural determination pipeline

Nine centers successfully completed PSI-1 operations in the summer of 2005 (Table 1-2 and Table 1-3) each of which developed HT pipelines using new technologies, most of which were created before the start of PSI-1 and were critically evaluated during PSI-1. The HT structure determination pipelines covered all activities from target selection to analysis and deposition of solved structures in the PDB. Both single crystal X-ray diffraction and solution NMR structural determination approaches were used.

Table 1-2. PSI Centers.

Center	***Home Institution and website***
PSI-1 Pilot Centers	
Joint Center for Structural Genomics	The Scripps Research Institute, http://www.jcsg.org
Midwest Center for Structural Genomics	Argonne National Laboratory, http://www.mcsg.anl.gov
New York Structural GenomiX Research Consortium	Structural GenomiX, Inc., http://www.nysgrc.org
Northeast Structural Genomics Consortium	Rutgers University, http://www.nesg.org
Southeast Collaboratory for Structural Genomics -	University of Georgia, Athens, http://www.secsg.org/
Berkeley Structural Genomics Center	University of California, Berkeley, http://www.strgen.org/
Tuberculosis (TB) Structural Genomics Consortium -	Los Alamos National Laboratory, http://www.doe-mbi.ucla.edu/TB/
Structural Genomics of Pathogenic Protozoa Consortium	University of Washington, http://www.sgpp.org/
Center for Eukaryotic Structural Genomics -	University of Wisconsin, Madison, http://www.uwstructuralgenomics.org/
PSI-2 Large Scale Production Centers	
Joint Center for Structural Genomics	The Scripps Research Institute, http://www.jcsg.org
Midwest Center for Structural Genomics	Argonne National Laboratory, http://www.mcsg.anl.gov
New York Structural GenomiX Research Consortium	Structural GenomiX, Inc. http://www.nysgrc.org
Northeast Structural Genomics Consortium	Rutgers University, http://www.nesg.org
PSI-2 Specialized Centers	
Accelerated Technologies Center for Gene to 3D Structure	deCODE Biostructures and The Scripps Research Institute, http://www.atcg3d.org
Center for Eukaryotic Structural Genomics	University of Wisconsin, Madison, http://www.uwstructuralgenomics.org
Center for High-Throughput Structural Biology	Hauptman-Woodward Medical Research Institute, http://www.chtsb.org
Center for Structures of Membrane Proteins	University of California, San Francisco http://www.csmp.ucsf.edu
Integrated Center for Structure and Function Innovation	Los Alamos National Laboratory http://www.techcenter.mbi.ucla.edu
New York Consortium on Membrane Protein Structure	New York Structural Biology Center http://www.nycomps.org

Table 1-3. Production summary for PSI-1 Structural Genomics Centers based on TargetDB XML distribution file (http://targetdb.pdb.org/target_files/). The table given below was downloaded from http://olenka.med.virginia.edu/mcsg/html/recent_results.html which was last updated on August 2, 2005. Only distinct target sequences are taken into account for each center and in the total count (hence numbers of "distinct" targets reported for centers where sequences are duplicated or missing in XML files may be lower than those reported by the centers; note also that the number of targets in the total count may be less than the sum of targets for the centers due to target overlaps).

Center	All Targets	Cloned	Targets With Crystals	Diffracting Targets	Total Solved (X-ray, NMR)	Median Length
MCSG	15359	5675	838	349	281	319
JCSG	6594	3650	1166	265	226	415
NESGC	12205	5309	162	115	206	193
NYSGRC	2145	1538	388	185	185	454
TB	1756	1547	208	118	107	574
SECSG	14786	14377	223	118	76	214
BSGC	911	812	94	65	60	374
CESG	6582	4476	104	40	52	222
SGPP	19503	10154	175	45	28	200
TOTAL	74899	45189	3257	1277	1206	361

HT pipelines employed a manufacturing style approach in that responsibilities were compartmentalized by function and processes were standardized through the use of quality assurance practices such as standard operating procedures (SOP). Whenever possible, common quality control practices were employed to monitor processes and materials from beginning to end. Data was uploaded to a common database to facilitate target management, process monitoring, and regular reporting. Laboratory information management systems (Zolnai et al 2003; Bertone et al 2001) were used to manage and track experiments. A good example is the java-based SESAME system developed by the CESG group in Wisconsin. As all the projects were run as multi-institutional collaborations, specific pipeline processes were implemented in separate institutions. For example, in the case of the JCSG, steps from protein production to crystallization as well as crystal mounting were carried out at TSRI and GNF, while diffraction screening and data collection were done at the Stanford Synchrotron Radiation Laboratory (SSRL). Most of the pipelines were established as learning platforms wherein experimental results and operational experiences were applied using a feedback loop to incrementally introduce improvements to the process.

Achievement of the 1000 structure milestone by these pipelines validates the hypothesis that structural genomics pipelines could be constructed and scaled-up. It also demonstrates the feasibility of using HT approaches for protein production, a notion that was not clear at the start of the PSI as it was generally thought that the variability in protein properties would not make them amenable to handling by simplified processes. Much remains to be done, for example expression and purification of eukaryotic proteins, some of which may require folding partners, remains to be developed and is the focus of a number of the PSI-2 specialized centers.

2.2 Pipeline technologies

A central theme in the technology development area has been automation, integration, and miniaturization of processes in the pipeline. These goals have reduced the cost per structure by decreasing time from gene to structure, material usage, and number of personnel needed to accomplish large numbers of tasks. As mentioned above, most centers developed database and software products to manage their pipeline, in addition a number of essential technologies were also developed, most of which are now in general use by the community. In this section we mention a number of notable technologies that have contributed significantly to the effort.

2.2.1 Protein Expression and Fermentation

Several innovative approaches have resulted in a marked increase in productivity and through put in expression particularly in *E. coli*. Studier from the NYSGRC has formulated growth media (Studier 2005) in which expression strains can grow uninduced to relatively high cell densities and then be induced automatically without any intervention by the experimenter. Cell densities attained in these auto-inducing cultures have produced 10-fold more target protein per volume of culture than with the standard IPTG induction protocol. Auto-induction also allows many cultures to be inoculated in parallel and induced simply by growing to saturation, making auto-induction a powerful tool for screening clones for expression and solubility in an automated setting.

Two JCSG-related innovations that greatly increased capacity were a high throughput, 96-tube *E. coli* expression system (Page et al 2004) and a scalable 96-well micro-expression device (Page et al 2004). The GNF fermentor is a production device capable of 96 simultaneous 65 mL fermentations in either native or selenomethionine (SeMet) media. Pelleted cell mass after 6 hours of growth varies from 1-3g/tube for SeMet to

3-5g/tube in native media. This device has already resulted in a more than ten-fold reduction in the culture volume required for protein production when compared to conventional expression in shaker flasks.

Our group at TSRI adapted a low-cost, high-velocity incubating Glas-Col (Glas-Col, LLC, Terre Haute, IN, USA) Vertiga shaker to develop an efficient, HT *E. coli* microliter-scale expression screening protocol which accurately predicts parameters that can be used for scale-up studies (milliliter and liter fermentation) (Page et al 2004). The apparatus shakes cultures in three-dimensions at speeds of up to 1000 rpm, allowing small-scale (~750 μL) cultures grown in 2 mL deep-well 96-well blocks to achieve optical densities (OD_{600}) as high as 10-20. This generates sufficient material for analysis of expression, solubility, binding to affinity purification matrices, and initial crystallization/NMR analysis. Moreover, this screening strategy has also been used to identify clones which express and are soluble under SeMet or $^{15}N/^{13}C$-labeled expression conditions that are necessary for the production of labeled recombinant proteins for direct structural analysis. It also provided an early quality control step in that one 96-well micro-purification step produced enough of each protein for characterization by MALDI, electrophoresis, or size exclusion chromatography.

Protein purification from *E. coli* at the JCSG has also been largely automated using GNFuge, a robot developed at GNF. The fully automated GNFuge harvests, sonicates, centrifuges, and aspirates 96 bacterial cultures in parallel. In addition, it facilitates fully automated affinity purification of tagged proteins from the resulting lysates and for insoluble proteins. An on-column refolding strategy compatible with this automation was recently implemented.

2.2.2 Crystallization

The last five years has seen rapid development and deployment of technologies and systems designed to carry out large-scale crystallization experiments. These include the use of nanoliter volumes (Santarsiero et al 2002), use of microfluidics (Hansen et al 2002), the development of rapid and large-scale crystallization imaging and storage systems (Hosfield et al 2003), and finally integration of these technologies into a complete system (e.g., CrystalMotion, available from MSC, http://www.rigakumsc.com/). The system that we developed along with a team of engineers and scientists at GNF and Syrrx is capable of performing 100,000 sitting drop experiments per-day, imaging one 96-well plate in one minute and storing and managing up to 40,000 plates in a cold room (Hosfield et al 2003). This system has been in operation for almost five years and continues to process reliable and productive experiments.

The majority of the key developments were created just prior to the start of PSI-1, but PSI-1 was critical in the validation of these advancements. Results from JCSG and other groups implementing the nanovolume crystallization technology clearly demonstrated the power of this new approach. Smaller volumes of protein allowed for the exploration of a broader universe of crystallization conditions, leading to significant costs savings, shorter crystallization times, improved crystal quality, and the successful crystallization of targets, previously difficult to achieve with larger volumes (Santarsiero et al 2002; Carter et al 2005). Capitalizing upon increases in intensity and focus of X-rays beams at modern synchrotron facilities, early JCSG-related studies showed that crystals for X-ray data collection could be reproducibly generated in volumes as low as 50nL. Although 100-200nL proved to be more practical in a production setting, an order of magnitude decrease in protein consumption was realized in the TSRI pipeline. Of critical importance is that all of these technologies are now available at "reasonable" cost to the scientific community, with young start-up labs now able to afford crystallization and imaging robotic systems.

2.2.3 X-ray Diffraction Screening and Data Collection

The process of mounting flash-cooled crystals, aligning them with the X-ray beam and evaluating and collecting their diffraction was clearly a major bottleneck for any HT structure determination pipeline. Efforts to automate these processes were in the planning and prototyping stage in 2000 when PSI-1 started (Abola et al 2000). By 2005, the start of PSI-2, the majority of beamlines used in structural genomics efforts have been automated with new robotic and software systems (McPhillips et al 2002; Cohen et al 2002; Snell et al 2004). In addition, new products became available that have been installed for use in-house. A good example is the ACTOR system from Rigaku/MSC which is the first commercially available off-the-shelf system for automatically changing samples for screening or data collection (Muchmore et al 2000).

Beamlines equipped with a crystal mounting robot can now handle hundreds of samples mounted in 96-format cassettes that can be screened in a few hours (e.g. at SRRL it takes about 5 hours to process 3 cassettes) without any human intervention. Automated crystal mounting at the beamline permits a more thorough and systematic approach to the screening process, which in turn translates into a higher structure determination success rate, as the crystal quality cannot be judged solely from their physical appearance. All diffraction data are processed in real-time to evaluate both quality and completeness. Real-time data reduction and analysis allow accurate determination of the amount of data required to solve any given structure. Data collection is terminated once sufficient data are collected and the sample restored in liquid nitrogen. At the SSRL, all protein crystallography beamlines now have automation systems and are integrated with the Blu-Ice/DCS data collection environment (McPhillips et al 2002).

The PSI-1 also funded the development of the Compact Light Source (CLS, Lyncean Technologies Inc.) through a Small Business Innovation Research (SBIR) program. The CLS is a breakthrough technology that offers the possibility of a “synchrotron beamline” for home laboratory applications. This tunable, tabletop X-ray source can be used in much the same way as a typical X-ray beamline at a large facility; but it is small enough to bring state-of-the-art methods of macromolecular crystallography directly into an experimenter's local laboratory.

2.2.4 NMR

NMR spectroscopy is a well-established technique for protein structure determination, as well as to screen for the folded state of globular proteins (Muchmore et al 2000; Markley et al 2003; Wüthrich 2003). Since NMR spectroscopy has intrinsically low sensitivity, milligram amounts of protein are required for screening and structure determination with conventional equipment. At the beginning of PSI-1 in 2000, about 6 weeks of NMR instrument time per protein structure was considered to be a realistic estimate for ~1 mM protein samples with molecular weights up to 15 kDa. By 2005, the start of PSI-2, NMR had been successfully transformed. It is now being used for HT structural determination efforts as well as for screening protein samples to determine suitability for crystal structural studies. A total of 123 structures were determined by NMR in PSI-1 of which 91 were done at the NESG. Microcoil NMR probes had been developed for use in biomolecular NMR spectroscopy (Olson et al 1995; Peti et al 2004). Specifically, small diameter coils enable up to ten-fold (mass-based) sensitivity gain so that microgram amounts of protein are now sufficient for screening by NMR spectroscopy. At the JCSG, by the 4th year of operations, most samples were being screened for the folded state with the microcoil probe before undergoing crystallization studies and assigned a grade of A, B, C, or D (Table 1-4; Page et al 2005). Using the microcoil probe, such information could be collected with 5 µL of protein and in 5 minutes. At this time, miniaturization is primarily aimed at identifying promising targets for structure determination. This methodology effectively guides efforts to focus on targets with a high probability of success, and either eliminates poor targets or replaces them by improved constructs. Overall, this process increases the efficiency of the entire pipeline and results in a reduction of the cost per structure. Further developments including optimized miniaturization may, at least for some proteins, lead the way directly to structure determination (Page et al 2005).

Table 1-4. Results of crystallographic studies with 79 mouse homologue proteins that were graded 'A' to 'D' based on 1D ^{1}H NMR screening.

Grade[a]	Proteins[b]	Crystal Hits[c]	> 5.0 Å Diffraction; *No Structure*[d]	< 5.0 Å Diffraction; *No Structure*[e]	Structures Solved[f]
A	24	16 (67%)	0 (0%)	4 (17%)	4 (17%)
B	26	22 (85%)	0 (0%)	1 (4%)	9 (35%)
C	22	18 (82%)	4 (18%)	5 (23%)	2 (9%)
D	7	6 (86%)	2 (29%)	2 (29%)	0 (0%)
Total	79	62	6	12	15

[a] The classification into four grades, 'A' to 'D', by 1D ^{1}H NMR screening is described in the published manuscript by (Page et al., 2005). 'A' and 'B' are proteins that are now routinely forwarded for extensive coarse and fine-screen crystallization trials, while 'C' and 'D' proteins are only subjected to coarse-screen crystallization trials.

[b] Number of proteins in each category.

[c] The number of proteins that crystallized in at least one coarse screen crystallization condition. Two 'A' proteins had been removed from the pipeline for structure determination by NMR.

[d] The number of proteins for which the best crystals diffracted to no higher than 5.0 Å.

[e] The number of proteins for which the best crystals diffracted to better than 5.0 Å, but no structure is as yet available.

[f] The number of proteins for which high resolution crystal structures have been determined.

In addition to the miniaturization efforts for NMR, cryogenic probes were evolving and becoming more robust and useful, offering approximately a 3-fold increase in sensitivity in routine applications in biological NMR spectroscopy (Monleon et al 2002), and potentially an order-of-magnitude reduction in measurement times. The use of this probe has led to the development of G-matrix Fourier Transform (GFT) NMR by the NESG Center which enables researchers to optimally adjust NMR measurement times to sensitivity requirements and allows them to take full advantage of highly sensitive cryogenic probes for HT NMR structure determination.

3. THE JCSG PROTEIN AND CRYSTAL PRODUCTION PIPELINES

3.1 Protein targets

Although target species changed over the 5 years of the project, the goal of proving the feasibility of attacking an entire genome remained the principal focus of the JCSG. *C. elegans* was chosen as the initial target set. Shortly after start of the project, it became clear that the pipeline was not ready to tackle a complete eukaryotic system and hence *T. maritima* became the principal prokaryotic genome of the JCSG. However, within the following year, the mouse genome was providing a eukaryotic source of additional protein targets. By the end of year four, approximately 70% of the total structures solved by the JCSG were proteins from *T. maritima*.

3.2 Production strategies

The JCSG adopted a three-tiered shotgun strategy for the crystallization of the *T. maritima* proteome in order to identify and focus the majority of crystallization efforts on those proteins with a demonstrated propensity to crystallize (Lesley et al 2002). This strategy is founded on the hypothesis that proteins which crystallize readily, even under suboptimal conditions, will do so again during focused crystallization attempts. In tier 1, the goal is to identify those targets which have a propensity to crystallize under the conditions tested; the quality of the crystals produced is not significantly important. To maximize throughput, the protein samples are purified with only one round of affinity purification and screened for crystal formation against a limited number of crystallization conditions; it is expected that some of the proteins will not be sufficiently pure or in the optimal state to crystallize. In tier 2, the objective is to obtain diffraction-quality crystals suitable for structure determination. In this stage, the targets that crystallized in tier 1 are reprocessed to contain SeMet, purified extensively and screened against an expanded set of crystallization conditions. Selected difficult targets that did not produce high quality crystals in tier 2 were subjected to further batch processing in tier 3 which used a loosely defined *ad-hoc* batch process referred to as a "*salvage pathway*".

3.3 Cloning and expression

Target constructs were generally produced in multiples of 96-well plates. Upon generation of selected target sequences and primers, the TSRI pipeline utilized PCR to generate target DNA from appropriate American Type Culture Centre (ATCC) available genomic DNA. Typically, the insert was ligated into a modified Invitrogen pBAD backbone to create a plasmid that specified ampicillin resistance, arabinose inducibility, and that would place a 6 His N-terminal tag on the protein for use in expression quality control testing and purification. Restriction sites for ligation were engineered using Pm1I (N-terminal) and FseI (C-terminal). Variations of the TSRI protocol included a TEV protease cleavage site and T7 promoter. Cell transformation was by heat shock, with competent cell storage as glycerol stocks at -80^{o}C. A Qiagen BioRobot 3000 provided the necessary automation.

3.4 Purification

Purification starts with cell harvest, sonication, and clarification of the *E. coli* extracts. At JCSG this was accomplished in a single step for up to 96 samples using the GNFuge. Proteolysis and denaturation were minimized by cooling, inclusion of a protease inhibitor cocktail, and the addition of a mild reducing agent. Viscosity reduction for subsequent steps was provided by adding a DNAse. Verification of expression in the clarified extract was provided by SDS PAGE and anti-His western blotting.

Up to 96 gravity fed, immobilized metal chelate columns (IMAC) were run in parallel to provide one-step purification for native proteins entering initial crystallization screening. Elution was via a single step gradient. IMAC was also applied to all targets entering tiers 2 and 3 of the pipeline for structure determination or salvage. At TSRI, an agarose-based cobalt resin provides low non-specific binding and allowed a low salt elution that facilitated a subsequent ion exchange step. For some *E. coli* studies and for targets expressed in insect cells, a second IMAC step was performed after TEV cleavage of the His tag. Post-IMAC quality control included SDS PAGE for purity, a Bradford assay for yield, and MALDI to verify identity by molecular weight.

Anion or cation exchange chromatography (IEX) were used to both purify and concentrate all samples entering tiers 2 or 3 of the pipeline. Sample loading, separation, and peak cutting are automated through the use of various Pharmacia-Amersham automated FPLC systems. The JCSG capacity was as high as 60 targets per day from a single production shift. TSRI found Waters AP-1 columns packed with Poros HQ resins to be most amenable to the 10mL/min gradient conditions required for maximum throughput. Quality control post-IEX typically included SDS PAGE, MALDI of tryptic-digests, and analytical size exclusion chromatography (SEC).

Preparative SEC was optionally employed for samples showing less than 95% purity by analytical SEC or SDS PAGE. Columns and conditions varied between the two pipelines, but TSRI generally employed Superdex 200. Although relatively slow, the separations were automated and thus ran unattended. If multiple species of a single target were readily resolved, each was screened separately in crystallization trials.

3.5 Suitability testing

In year four of the JCSG, NMR screening represented a major improvement to the pipeline that was the result of a collaboration with Professor Kurt Wüthrich (Page et al 2005). Averaging only 5 minutes per measurement, a 1D ^{1}H NMR spectrum was recorded for subsequent evaluation of band broadening. Proteins were then categorized into one of four groupings that reflected their folded state and the likelihood of structure determination by NMR or crystallography. A study of 79 mouse homolog targets showed that despite a nearly equal ability to crystallize, only proteins graded A-B or flagged as potential multimers produced high resolution structures (Peti et al 2005). Most recently the NMR screening process was further refined by implementing use of a 1mm probe to reduce sample requirements to only 5µL of a 0.5-2mM protein solution, as well as by automation of the sample loading and measurement steps.

A second key component of the pipeline was implemented in year 4, stemming from a collaboration with Dr. Virgil Woods at UCSD (Pantazatos et al 2004). Deuterium exchange mass spectrometry (DXMS) measures the solvent exchange rates of amide hydrogen atoms to identify unstructured regions within a protein. Deletion mutants were then generated to reduce the level of disorder, a process which proved effective in the generation of crystals yielding high resolution structures. Further refinement of the technology to increase throughput continues.

3.6 Crystallization

Targets were concentrated and placed into a suitable delivery buffer for crystallization by ultrafiltration over regenerated cellulose membranes having a molecular weight cut-off of ~10,000Da. At TSRI, coarse and fine screening was typically initiated at protein concentrations of 0.5 and 1mM, respectively, with further optimization guided by the solubility results obtained from crystallization trials. Increasing attention was paid to repeated over-concentration during buffer exchange, which could facilitate aggregation or irreversible precipitation. TSRI utilized a buffer composition of 10mM Tris pH 7.8, 100mM NaCl, and 0.25mM TCEP for proteins entering crystallization trials. Delivery buffer optimization for soluble proteins received attention only as a salvage pathway, but simply reviewing crystallization results for buffers that yielded clear drops after an overnight incubation provided dramatic improvements in tier 3 buffer selection and crystal quality. Further development of this option continues, based upon previously reported successes at the Berkley Center for Structural Genomics (Jancarik et al 2004).

Crystallization screening at JCSG has been ripe with innovation, owing in part to the systematic capture and analysis of production data (Stevens 2000; Page and Stevens 2004). Roughly 480 crystallization conditions were evaluated in over 320,000 individual experiments on 28% of the *T. maritima* proteome (Page et al 2003). Approximately 86% of purified proteins produced crystals. Prioritization of a subset of these proteins for SeMet labeling and more extensive purification resulted in 68 of 69 proteins yielding crystals in tier 2 and the percentage of total crystals that were harvestable nearly doubled (41%). Further review of pipeline processes revealed that over 75% of the commonly used crystallization conditions found in tier 1 were redundant, with a subset of only 108 of the best conditions yielding crystals for all previously successful 465 tier 1 purified proteins. This led to the establishment of a set of core screening conditions that provided an estimate of a protein's compatibility with tier 2 of the crystallomics core pipeline. It should be noted that the more extensively purified SeMet proteins tended to crystallize under different conditions than their less pure tier 1 counterparts and hence, tier 2 screening maintained its more aggressive wide-sampling of crystallization space. In the last year of PSI-1, TSRI was using a 96-well, 200nL sitting drop format that limited consumption to only 20μL of protein for a complete tier 1 screen at two temperatures.

Review of the results from crystallization studies of *T. maritima* proteins has led to a proposed target filtering strategy (Canaves et al 2004). To identify useful criteria for future protein target selection and to determine ways to improve current pipeline protocols to increase crystallization success of active targets, the distribution of various parameters in the proteome and in the subset of crystallized proteins was analyzed for trends in crystallization success. The parameters analyzed were: (a) biophysical properties, including sequence length, isoelectric point, protein hydropathy, and percentage of charged residues, (b) predicted transmembrane helices and signal peptide sequences, (c) predicted bacterial lipoprotein lipid-binding sites (hydrophobicity pockets), (d) predicted coiled-coils, and (e) predicted low-complexity regions that might lead to disorder.

Seven sequence-derived parameters shown to have a direct effect on protein crystallization were selected for these filtering strategies, including protein length, calculated isoelectric point, percent charged residues, gravy index to indicate hydrophobicity, the number of SEG residues to identify low complexity, the number of predicted trans-membrane helices and the number of predicted signal peptides (Table 1-5). The first strategy proposed is based on the absolute maxima and minima at which crystallization has been observed for each parameter, i.e. none of the observed crystals would be lost, but would still result in an increase in the ratio of successfully crystallized proteins and selected targets (37.7%, Table 1-5). The second strategy is based on more stringent cut-offs that tolerate the loss of up to 5% of the crystals per parameter. The goal is to further reduce the pool of potential targets with respect to the first strategy, while further increasing the ratio between successfully crystallized proteins and selected targets (39.5%, Table 1-5). The loss of a small number of outlier crystallized proteins is tolerated because it allows for a higher success rate for new targets, resulting in an overall increase of successfully crystallized targets. Finally, we opted for an even more stringent filtering strategy that uses as limits the area where most of crystallized proteins cluster in the distribution defined by each protein attribute (Maximum Clustering Strategy, MCS). Whereas the number of lost crystallized proteins and solved structures is higher than in the second strategy, the ratio of crystallized and solved proteins to selected targets is even greater (45.1%, Table 1-5), indicating that this is a superior target selection or design strategy.

4. DISCUSSION

4.1 Production results, JCSG and other centers

A comparison of production statistics for all nine of the PSI-1 large-scale centers is given in Table 1-3. Several factors make it difficult to draw inferences from these numbers. For example, it clearly is hard to gauge the relative difficulties of working with the protein sets that each center was targeting, although prokaryotic proteins made up the majority of targets for PSI-1 (e.g. the JCSG focused almost exclusively on protein from *Thermotoga maritima*). These numbers however represent a valid documentation of the success rates of the overall process and provide an estimate of the approximate cost of doing structural studies. A more detailed breakdown of yield by process step for JCSG is shown in Table 1-6. It is interesting to note that the projected success rates for a number of

Table 1-5. Three different target filtering strategies and statistics calculated from primary sequence.

	No Filtering	Target Filtering Schema		
	Proteome Limits	Absolute Limits	95% Crystal Conservation Per Attribute	Maximum Clustering Strategy
Filtering Parameters: Targets not fulfilling these filtering requirements would be discarded as potential targets.	30<=*Length*<=1690 3.7<=*pI*<=12.3 7.4<=*%Charged*<=52.5 -1.56<=*GRAVY*<=1.62 *0<=SEG*<=177 *0<=TMHMM*<=16 *0<=SignalP*<=1	41<*Length*<813 4.3<*pI*<11.2 17<*%Charged*<47 -0.96<*GRAVY*<0.61 *SEG*<62 *TMHMM*<1 *SignalP*<1	65<*Length*<500 4.6<*pI*<10.2 25<*%Charged*<41 -0.70<*GRAVY*<0.10 *SEG*<32 *TMHMM*<1 *SignalP*<1	90<*Length*<480 4.6<*pI*<7.4 25<*%Charged*<40 -0.50<*GRAVY*<0.10 *SEG*<35 *TMHMM*<1 *SignalP*<1
Final number of targets after filtering (Initial target pool = 1877, i.e. *T. maritima* proteome)	1877	1232	875	606
Number of proteins which crystallized eliminated (total = 465)	0	0	-118	-191
Number of protein structures eliminated (total = 86)	0	0	-14	-18
Chances of Crystallization per target	24.7%	37.7%	39.5%	45.1%
Chances of Structure solution per target	4.6%	7.0%	8.2%	11.2%
Number of proteins crystallized with similar JCSG experimental effort	464 (actual)	708 (theoretical)	741 (theoretical)	847 (theoretical)
Number of protein structures produced with similar JCSG experimental effort	86 (actual)	131 (theoretical)	154 (theoretical)	210 (theoretical)
Theoretical gain in number of structures	n/a	+45	+68	+124
Theoretical increase in JCSG pipeline throughput	n/a	52%	79%	145%

The initial pool of potential targets, the *Thermotoga maritima* proteome, contains 1877 ORFs. The analysis shows the predicted effect of different target selection strategies on total pipeline throughput, assuming that the same experimental effort that was devoted to the full shotgun analysis of the *T. maritime* proteome had been used on focused efforts against selected (filtered) sets of targets.

[@] Number of crystals or structures lost with respect to tier1, if the proposed filtering schemas had been applied to the *T. maritima* proteome.

[*] Chances of crystallization and structural solution are calculated as 100*(Number of tier1 crystals or structures remaining after filtering / final number of targets after filtering).

[#] Theoretical number of structures gained with similar JCSG resources was calculated extrapolating the chances of structure solution per target for a certain filtering schema to a set of initial targets equivalent in size to the *T. maritime* proteome.

process steps which were presented in the original JCSG research proposal came close to what was achieved. Thus, the 3% overall success rate was anticipated in 2000. However, the original plan called for working with about 45,000 targets, clearly this was not achieved.

Table 1-6. Average yields at various stages of the JCSG pipeline as of October 2004 (number of structures as of August 2, 2005 was 226 as listed in Table 1-3). Numbers in parenthesis were the projected success rates as presented in the proposed JCSG plans in 2000.

Step	Total	% Overall	% Stage	Step	Total	% Overall	% Stage
Target Selection	6537	- -	100	Crystallized	985	17 (9.4)	85
Target Activation	5689	100	87	Screened (X-Ray)	384	7	39
Cloned	3131	55	55	Data Sets	205	4	53
Expressed	2811	49	90	Structure	171	3 (4.3)	83
Soluble Protein	1165	20 (24.4)	41				

4.2 Results from TSRI core

In years 3-5, a TSRI based group worked on a smaller subset of targets with a strong reliance on bioinformatics tools to expand the number of initial constructs per target. Of the 1,452 proteins processed by the TSRI pipeline in the final years of PSI-1, 42% represented homologs or orthologs of a parent target that was also being processed by the rest of JCSG. Secondly, a heavy emphasis was placed on extensive crystallization screening (coarse screening), which out of necessity evolved in year 4 to a more focused optimization of crystallization conditions (fine screening) and cryoprotectants. At the end of PSI-1, the TSRI target pipeline recorded a 44% success rate for turning crystallizable proteins into solved structures. A breakdown of the more than 309,000 screening experiments carried out by the TSRI pipeline alone showed: ~95% coarse screens, ~5% fine screens, and ~1% seeding, additive and chemical modification experiments. In contrast, the distribution of protein crystals that ultimately yielded a high resolution structure was 75% fine screen, 18% coarse screen, and 7%

additive screen. The distribution of structure generating crystals between 277 and 293K was nearly 50:50. Clearly opportunities still exist for improving the ability of sparse coarse screening to generate crystals suitable for high resolution X-ray crystallography. A sampling of TSRI production statistics during this period is shown in Table 1-7.

Table 1-7. Production statistics generated from the TSRI technology development crystallomics core using novel target selection filters. Note – this is a subset of the overall JCSG Crystallomics Core Production. The majority of targets were processed as part of the production crystallomics core at GNF.

		Total	Rate	Rate
Fermentation	Total constructs processed	576	100%	
	Total constructs expressing soluble	179	31%	
	Average ferments per construct that expressed soluble at least once	5.1		
	Average ferments per solved structure	5.4		
Purification	Total constructs processed	179	100%	100%
	Total native proteins processed	72	40%	
	Total SeMet proteins processed	154	86%	
	Total constructs purified successfully	164		92%
	Total purification runs	632		
	Average yield per purification (mg)	~6.4		
	Total purified protein generated (g)	>3.4		
	Average purifications per construct entering purification	3.5		
	Average purifications per solved structure	4.1		
Crystallization	Total constructs entering coarse screening	163	100%	
	Total constructs entering fine screening	67		
	Total proteins solved by X-ray diffraction	28	15.6%	
	Total plates	3,699	100.0%	
	Coarse screen plates	3,041	82.2%	
	Fine screen plates	476	12.9%	
	Seeding plates	4	0.1%	
	Additive plates	53	1.4%	
	Reductive methylation plates	12	0.3%	
	Uncharacterized plates	113	3.1%	
	Total experiments (wells)	309,303	100.0%	100.0%
	Native protein (wells)	95,091	30.7%	

Table 1-7. (Continued)

		Total	Rate	Rate
	SeMet protein (wells)	209,510	67.7%	
	Uncharacterized protein (wells)	4,702	1.5%	
	Coarse screen (wells)	289,825		95.2% 93.7%
	Fine screen (wells)	15,490		5.1%
	Seeding (wells)	132		0.0%
	Additive (wells)	2,704		0.9%
	Reductive methylation (wells)	1,152		0.4%
Structure Determination	Total construct structures solved	29		
	Total unique target structures solved	27		

The TSRI technology development crystallomics core target selection list has also been segmented into protein family or technology method approach (Table 1-8). Not surprisingly, those approaches that were particularly successful included the selection of bacterial homologs of eukaryotic targets, utilization of C-terminal truncations and domain isolation to generate multiple target constructs, and the DXMS-guided generation of deletion mutants for targets that had previously crystallized but diffracted poorly. Techniques that performed poorly included *in-silico* bioinformatics-guided rational target design, exploration of yeast homologs of eukaryotic targets, and utilization of physical measurements of disorder (DXMS) to optimize previously non-crystallizable proteins. Excluding the unsuccessful target selection protocols nearly doubles the average yield of structures per protocol from 2.9% to 5.4%.

4.3 Future directions

In PSI-2 the groups that had operated the JCSG have been funded to run a large-scale production center, JCSG-2 (www.jcsg.org), a specialized center, and a separate Road Map Initiative center. The Road Map Initiative center is called the Joint Center for Innovative Membrane Protein Technologies, and is located at TSRI with a focus on developing novel expression and stability systems for integral membrane proteins (JCIMPT; www.jcimpt.org). The PSI-2 specialized center is called the Accelerated Technologies Center for Gene to 3D Structure (ATCG3D; www.atcg3d.org), which is a collaboration centered at deCODE Biostructures and TSRI, with key collaborations at Lyncean Technologies and the University of Chicago. The ATCG3D is now assembling a new integrated pipeline using technologies currently being developed within the collaboration. Its overall

Table 1-8. A breakdown of the strategy of the TSRI technology development crystallomics core pipeline in years 3-5 (2002-2004) showing a heavy emphasis on homologs, orthologs, and rational target design. In general, the utilization of orthologs and generation of multiple constructs through C-terminal truncations proved more reliable than the utilization of more sophisticated bioinformatics techniques. Interestingly, DXMS produced an exceptionally high yield when used to guide the modification of constructs that had previously generated only poorly diffracting crystals.

Target Selection Protocol	Total Constructs	Parent Targets	Constructs Yielding Crystals	Constructs Yielding Diffraction	Unique Parent Structures Per Row	% Parent Structures Solved
Metabolic pathway targets – Enzymes	65	57	7	6	5	8.80%
Optimized bacterial homologs of metabolic pathway targets	79	33	3	2	0	0.00%
Heart mitochondrial proteome bacterialized	167	73	6	6	3	4.10%
Bacterial homologs of Mouse targets	190	175	17	12	7	4.00%
Optimized, bacterial homologs of Mouse targets	190	161	5	5	3	1.90%
Yeast homologs of Mouse targets	190	184	13	11	3	1.60%
Optimized, yeast homologs of Mouse targets	190	184	3	3	0	0.00%
Multiple constructs of viral targets	190	24	5	5	2	8.30%
Multiple bacterial orthologs of poorly diffracting targets	95	40	9	9	3	7.50%
Sequence optimization by DXMS for non-crystallizable targets	65	14	0	0	0	0.00%
Optimization by DXMS for targets that crystallized but diffracted poorly	31	7	2	2	2	28.60%

goal is to significantly reduce the cost of doing eukaryotic protein structures by approximately 10-fold while maintaining the high quality of work carried out by the structural genomics efforts. Four main areas of technology development are now underway:

4.3.1 Integral membrane proteins

Perhaps the proteins most under-represented in terms of three-dimensional protein structure are membrane proteins, particularly eukaryotic membrane proteins. Research at TSRI for the next several years will focus on developing key technologies to improve the success rate for this family of proteins. These research efforts include cell free protein expression, and improved tools for eukaryotic cell expression (insect cells and other mammalian cell lines). Detergent and lipid chemistry efforts are also a key area where combinatorial chemistry methods will be applied, using nanovolumes of proteins to screen for improved stability reagents.

4.3.2 Cloning by whole gene synthesis

Two of the primary problems with eukaryotic protein structure determination are the acquisition of reliable cDNA clones and the expression of protein constructs that will express well and be amenable to the production of diffraction quality crystals or NMR spectra. ATCG3D proposes to eliminate these bottlenecks by synthesizing all genes (Stewart and Burgin 2005) directly from synthetic oligonucleotides and designing the constructs with protein modeling. This approach has been demonstrated successfully in the past and proof of concept experiments have already shown that large genes (>7kb) and even small viral genomes can be produced by Whole Gene Synthesis. Due to the dropping prices for synthetic oligonucleotides and sequencing reactions, the economics of gene synthesis has reached a point where it is easier, more reliable, and often less expensive to synthesize the gene than it is to source, purchase, and validate a cDNA clone. Most importantly, the entire process can be automated to significantly reduce effort and cost.

It has been shown previously by many different structural biology efforts that an increased success rate of structure solutions can be accomplished by processing an extended number of protein constructs that includes homologs, domain boundary variants and mutants of the desired target protein (Derewenda 2004; Cohi et al 2004; Longenecker et al 2001). It has also been previously shown that when a particular construct does not express well, codon optimization is a very powerful alternative strategy. ATCG3D will use molecular modeling and codon based expression optimization to design multiple constructs, which will be built from synthetic oligonucleotides. This project is expected to lead to significant cost savings and will greatly improve overall success rates. The full system will be a single instrument that runs the computational modeling, process control database and gene synthesis on a compact footprint.

4.3.3 Crystallization using micro-capillary and *in situ* x-ray screening and data collection

Capillary-based microfluidics technology development has radically changed almost every liquid-based instrument from laser printers to DNA/protein/small molecule analysis. During the late 1990's, companies such as Fluidigm demonstrated the feasibility of microfluidics-based protein crystallization. The current cost of microfluidic chips, however, remains prohibitive for most structural genomics efforts, especially in academia. While the microfluidic technology is established, the breakthrough cost reduction and full implementation into a structural proteomics pipeline has yet to be realized. ATCG3D will focus on developing novel microfluidic technology that is inexpensive and can be directly integrated into both upstream and downstream processing steps including purification, imaging, X-ray screening, and data collection. Of particular importance will be the integration of crystallization with direct X-ray screening of protein crystals.

4.3.4 Compact Light Source

The implementation of an in–house, MAD-capable synchrotron light source might at first appear out of reach; however, the prototype development is already in place, and is based on integrating well-established technologies. The CLS is a miniature synchrotron founded on the marriage of two mature technologies—particle accelerator technology and solid-state laser technology. Accelerators and related hardware have been developed over the past 40 years by the Department of Energy for high-energy physics and synchrotron light sources. Over the past 30 years this progress has led to a large number of high-energy synchrotron light sources worldwide with continuing and dramatic improvements in the performance and quality of X-ray beamlines. Lyncean Technologies has miniaturized this technology by reducing the electron beam energy and by replacing conventional undulator magnets with a laser (http://www. lynceantech.com). The miniature synchrotron has an average flux comparable to the most productive beamlines at the large synchrotrons.

ACKNOWLEDGEMENTS

The majority of work described in this chapter resulted from the highly collaborative efforts between researchers at The Scripps Research Institute, Genomics Institute of the Novartis Research Foundation, and Syrrx. Funding for the JCSG was provided through a NIGMS PSI-1 grant GM062411; JCIMPT was funded by a NIH RoadMap Initiative grant GM073197; and the ATCG3D was funded by NIH PSI-2 grant GM074961.

REFERENCES

Abola E., Kuhn P., Earnest T. and Stevens R. C. 2000. Automation of X-ray crystallography. Nat. Struct. Biol. **7**, 973–977.

Adams M. D., Celniker S. E., Holt R. A., Evans C. A., Gocayne J. D., Amanatides P. G., Scherer S. E., Li P. W., Hoskins R. A., Galle R. F. et al. 2000. The genome sequence of *Drosophila melanogaster*. Science **287**, 2185–2195.

Bertone P., Kluger Y., Zheng D., Edwards A. M., Arrowsmith C. H., Montelione G. T., and Gerstein M. 2001. SPINE: An integrated tracking database and data mining approach for prioritizing feasible targets in high-throughput structural proteomics. Nucleic Acids Res. **29**, 2884–2898.

Canaves J. M., Page R., Wilson I. A. and Stevens R. C. 2004. Protein biophysical properties that correlate with crystallization success in *Thermotoga maritima*: Maximum clustering strategy for structural genomics. J. Mol. Biol. **344**, 977–991.

Carter D. C., Rhodes P., McRee D. E., Tari L. W., Dougan D. R., Snell G., Abola E. and Stevens R. C. 2005. Reduction in diffuse-convective disturbances in nanovolume protein crystallization experiments. J. Appl. Cryst. **38**, 87–90.

Choi K. H., Groarke J. M., Young D. C., Rossmann M. G., Pevear D. C., Kuhn R. J. and Smith J. L. 2004. Design, expression, and purification of a Flaviviridae polymerase using a high-throughput approach to facilitate crystal structure determination. Protein Sci. **13**, 2685–92.

Cohen A. E., Ellis P. J., Miller M. D., Deacon A. M. and Phizackerley R. P. 2002. An automated system to mount cryocooled protein crystals on a synchrotron beam line, using compact sample cassettes and a small-scale robot. J. Appl. Crystallogr. **35,** 720–726.

Derewenda Z. 2004. Rational protein crystallization by mutational surface engineering structure. **12**, 529–35.

Hansen C. L., Skordalakes E., Berger J. M. and Quake S. R. 2002. A robust and scalable microfluidic metering method that allows protein crystal growth by free interface diffusion. PNAS **99**, 16531–16536.

Hosfield D., Palan J., Hilgers M., Scheibe D., McRee D. E. and Stevens R. C. 2003. A fully integrated protein crystallization platform for small-molecule drug discovery. J. Struct. Biol. **142**, 207–217.

Jancarik J., Pufan R., Hong C., Kim S. H. and Kim R. 2004. Optimum solubility (OS) screening: An efficient method to optimize buffer conditions for homogeneity and crystallization of proteins. Acta Crystallogr. **D60**, 1670–1673.

Kigawa T., Yabuki T., Matsuda N., Matsuda T., Nakajima R., Tanaka A. and Yokoyama S. 2004. Preparation of Escherichia coli cell extract for highly productive cell-free protein expression. J. Struct. Funct. Genomics **5**, 63–68.

Klock H. E., White A., Koesema E. and Lesley S. A. 2005. Methods and results for semi-automated cloning using integrated robotics. 2005. J. Struct. Funct. Genomics., **6**, 89-94.

Lander E. S., Linton L. M., Birren B., Nusbaum C., Zody M. C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W. et al. 2001. Initial sequencing and analysis of the human genome. Nature **409**, 860–921.

Lesley S. A., Kuhn P., Godzik A., Deacon A. M., Mathews I., Kreusch A., Spraggon G., Klock H. E., McMullan D., Shin T. 2002. Structural genomics of the Thermotoga maritima proteome implemented in a high-throughput structure determination pipeline. Proc. Natl. Acad. Sci. **99**, 11664–11669.

Longenecker K. L., Garrard S. M., Sheffield P. J. and Derewenda Z. S. 2001. Protein crystallization by rational mutagenesis of surface residues: Lys to Ala mutations promote crystallization of RhoGDI. Acta. Crystallogr. D. Biol. Crystallogr. **57**, 679–88.

Markley J. L., Ulrich E. L., Westler W. M. and Volkman B. F. 2003. Macromolecular structure determination by NMR spectroscopy. Methods Biochem. Anal. **44**, 89–113.

Marra M. A., Jones S. J. M., Astell C. R., Holt R. A., Brooks-Wilson A., Butterfield Y. S. N., Khattra J., Asano J. K., Barber S. A., Chan S. Y. 2003. The genome sequence of the SARS-Associated Coronavirus. Science **300**, 1399–1404.

McPhillips T. M., McPhillips S. E., Chiu H.-J., Cohen A. E., Deacon A. M., Ellis P. J., Garman E., Gonzalez A., Sauter N. K., Phizackerley R. P. et al. 2002. Blu-Ice and the distributed control system: Software for data acquisition and instrument control at macromolecular crystallography beamlines. J. Synchrotron Rad. **9,** 401–406.

Monleon D., Colson K., Moseley H. N. B., Anklin C., Oswald R., Szyperski T. and Montelione G. T. J. 2002. Rapid analysis of protein backbone resonance assignments using cryogenic probes, a distributed Linux-based computing architecture and a set of semi-automated spectral analysis tools. J. Struct. Funct. Genomics **2**, 93–101.

Muchmore S. W., Olson J., Jones R., Pan J., Blum M., Greer J., Merrick S. M., Magdalinos P. and Nienaber V. 2000. Automated crystal mounting and data collection for protein crystallography. Structure **8**, R243-R246.

NIH press release 2005. http://www.nih.gov/news/pr/jul2005/nigms-01.htm.

Olson D. L., Peck T. L., Webb A., Magin R. L. and Sweedler J. V. 1995. High-resolution microcoil ^{1}H-NMR for mass-limited, nanoliter-volume samples. Science **270**, 1967–1970.

Page R., Moy K., Sims E. C., Velasquez J., McManus B., Grittini C., Clayton T. L. and Stevens R. C. 2004. Scalable high-throughput micro-expression device for recombinant proteins. Biotechniques **37**, 364–370.

Page R., Peti W., Wilson I. A., Stevens R. C. and Wuthrich K. 2005. NMR screening and crystal quality of bacterially expressed prokaryotic and eukaryotic proteins in a structural genomics pipeline. Proc. Natl. Acad. Sci. U. S. A. **102**, 1901–1905.

Page R. and Stevens R. C. 2004. Crystallization data mining in structural genomics: using positive and negative results to optimize protein crystallization screens. Methods **34**, 373.

Page R., Grzechnik S. K., Canaves J. M., Spraggon G., Kreusch A., Kuhn P., Stevens R. C. and Lesley S. A. 2003. Shotgun crystallization strategy for structural genomics: An optimized two-tiered crystallization screen against the thermotoga maritima proteome. Acta Crystallogr. **D59**, 1028–1037.

Pantazatos D., Kim J. S., Klock H. E., Stevens R. C., Wilson I. A., Lesley S. A. and Woods V. L., Jr. 2004. Raid refinement of crystallographic protein construct definition employing enhanced hydrogen/deuterium exchange MS. Proc. Natl. Acad. Sci. U. S. A. **101**, 751–756.

Peti W., Norcross J., Eldridge G., O'Neil-Johnson M. 2004. Biomolecular NMR using a microcoil NMR probe-new technique for the chemical shift assignment of aromatic side chains in proteins. J. Am. Chem. Soc. **126**, 5873–5878.

Peti W., Page R., Moy K., O'Neil-Johnson M., Wilson I. A., Stevens R. C. and Wuthrich K. 2005. Towards miniaturization of a structural genomics pipeline using micro-expression and microcoil NMR. J. Struct. Funct. Genomics **6**, 259-67.

Rota P. A., Oberste M. S., Monroe S. S., Nix W. A., Campagnoli R., Icenogle J. P., Peñaranda S., Bankamp B., Maher K., Chen M. 2003. Characterization of a Novel Coronavirus associated with severe acute respiratory syndrome. Science **300**, 1394–1399.

Santarsiero B. D., Yegian D. T., Lee C. C., Spraggon G., Gu J., Scheibe D., Uber D.C., Cornell E. W., Nordmeyer R. A., Kolbe W. F. et al. 2002. J. Appl. Cryst. **35**, 278.

Snell G., Cork C., Nordmeyer R., Cornell E., Meigs G., Yegian D., Jaklevic J., Jin J., Stevens R. C. and Earnest T. 2004. Automated sample mounting and alignment system for biological crystallography at a synchrotron source. Structure **12,** 537–545.

Stevens R. C. 2000. High-throughput protein crystallization. Curr. Opin. Struct. Biol. **10**, 558–563.

Stevens R. C. 2004. Long live structural biology. Nat. Struct. Mol. Biol. **11**, 293–295.

Stewart L. and Burgin A. B. 2005. Whole gene synthesis: A Gene-O-Matic future. In *Frontiers in Drug Design and Discovery*. Bentham Science Publishers, Ltd, Co-Editors Atta-ur-Rahman, B. A. Springer, G. W. Caldwell.

Studier F. W. 2005. Protein production by auto-induction in high density shaking cultures. Protein Expr. Purif. **41**, 207–34.

Venter J. C., Adams M. D., Myers E. W., Li P. W., Mural R. J., Sutton G. G., Smith H. O., Yandell M., Evans C. A., Holt R. A. et al. 2001. The sequence of the human genome. Science. **291**, 1304–1351.

Wüthrich K. 2003. NMR studies of structure and function of biological macromolecules. J. Biomol. NMR **27**, 13–39.

Wüthrich K. 1986. NMR of proteins and nucleic acids. New York, Wiley.

Zolnai Z., Lee P. T., Li J., Chapman M. R., Newman C. S., Phillips G. N., Jr., Rayment I., Ulrich E. L., Volkman B. F. and Markley J. L. 2003. Project management system for structural and functional proteomics: Sesame. J. Struct. Funct. Genomics **4,** 11–23.