Published in final edited form as:
Structure. 2008 January ; 16(1): 5–11. doi:10.1016/j.str.2007.12.002.

Contributions to the NIH-NIGMS Protein Structure Initiative from the PSI Production Centers

Stephen K. Burley[1,*], **Andrzej Joachimiak**[2,*], **Gaetano T. Montelione**[3,*], and **Ian A. Wilson**[4,*]

[1]*New York SGX Research Center for Structural Genomics (NYSGXRC), SGX Pharmaceuticals, Inc., 10505 Roselle Street, San Diego, CA 92121, USA*

[2]*Midwest Center for Structural Genomics (MCSG), Biosciences Division, Argonne National Laboratory, Argonne, IL 60439, USA*

[3]*Northeast Center for Structural Genomics (NESG), Center for Advanced Biotechnology and Medicine, Rutgers University, Piscataway, NJ 08854, USA*

[4]*Joint Center for Structural Genomics (JCSG), The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA*

In biology, function follows form. Less succinctly put, the biological/biochemical function of a protein is dictated by its three-dimensional (3D) structure. To fully comprehend how proteins work in isolation and how proteins cooperate with one another and with other molecules large or small to support life, or through mutations and other perturbations cause disease, they must often be visualized at the atomic level. Our knowledge of protein structure is limited to a small fraction of proteins encoded by the human genome or those found elsewhere in nature. With the advent of high-throughput genome sequencing and the availability of genome sequences for >500 organisms, mechanistic studies in biology have been transformed by the need to examine gene products in parallel. This broadened scope of biomedical research stimulated development of numerous high-throughput methods. Among early adopters of this approach was a cadre of biologists, both academic and industrial, enabled by numerous technological advances, who undertook high-throughput structural characterization of proteins to improve our understanding of biology and human health/disease. Such efforts catalyzed initiation of structural genomics programs globally.

The Protein Structure Initiative

The goal of the United States National Institutes of Health (NIH) National Institute of General Medical Sciences (NIGMS) Protein Structure Initiative (PSI; http://www.nigms.nih.gov/Initiatives/PSI/) is to make 3D atomic-level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences. The PSI is primarily focused on structure determination of proteins for which there is no public structural information. Upon completion, PSI structures are made public via deposition to the Protein Data Bank (PDB; http://www.pdb.org). The PSI also aims to make structure determination more accurate/efficient, and seeks to disseminate such technological advances widely. When defining the purview of the PSI, the NIH-NIGMS Council sought to complement and support R01-funded research and explicitly disallowed PSI support for experimental functional annotation to focus all efforts on novel structure determination, as reflected in the PSI-2 Requests for Applications (RFAs:

*E-mail: sburley@sgxpharma.com (S.K.B.), E-mail: andrzejj@anl.gov (A.J.), E-mail: guy@cabm.rutgers.edu (G.T.M.), E-mail: wilson@scripps.edu (I.A.W.).

http://grants.nih.gov/grants/guide/rfa-files/RFA-GM-05-001.html; http://grants.nih.gov/grants/guide/rfa-files/RFA-GM-05-002.html).

The PSI, initiated in September 2000, is now in its second 5-year phase (PSI-2; July 1, 2005–June 30, 2010). Four Large-Scale Centers, six Specialized Centers, two Homology Modeling Centers, a Materials Repository, and the Knowledgebase, plus R01 and P01 grantees currently receive PSI funding. The Large-Scale Centers (or Production centers) are determining >600 experimental protein structures/year and are pursuing various technology development efforts. The Specialized Centers are exploring technological developments with which to further speed structure determination and reduce costs, particularly for challenging targets (i.e., integral membrane proteins, eukaryotic protein domains, and protein-protein complexes). The Homology Modeling Centers are charged with improving modeling of evolutionarily related sequences using experimental structures. Tens of thousands of protein expression clones will be distributed by the Materials Repository (http://www.plasmid.harvard.edu) to enable functional studies of PSI targets. Finally, the recently-launched Knowledge-base (http://kb-test.psi-structuralgenomics.org/KB/) is charged with enhancing access to PSI results to maximize their value to researchers/educators worldwide.

The PSI is a geographically distributed research network, involving large teams spread across two continents and more than sixty academic, government, and not-for-profit research institutions, and three companies. The program funds many molecular biology, microbiology, structural biology, and bioinformatics research laboratories, including those of successful young investigators. PSI researchers constitute a cohesive team that uses a range of resources to foster innovation and scientific excellence. They have also trained hundreds of graduate and undergraduate students, postdoctoral researchers, and faculty members, and synchrotron and NMR facility staff members. Finally, minority outreach/training represents an important component of the PSI.

The PSI Production Centers

Composition/Management

The four Large-Scale Centers—Joint Center for Structural Genomics (JCSG; http://www.jcsg.org), Midwest Center for Structural Genomics (MCSG; http://www.mcsg.anl.gov), New York SGX Research Center for Structural Genomics (NYSGXRC; http://www.nysgxrc.org), and the Northeast Center for Structural Genomics (NESG; http://www.nesg.org)—began as technology development centers during the pilot phase of the PSI (PSI-1; September 1, 2000–June 30, 2005). Using high-throughput X-ray crystallography or solution state nuclear magnetic resonance (NMR) spectroscopy, these multi-disciplinary, multi-institution research collaboratories are largely responsible for PSI-2 protein production/structure determination efforts. Their activities are coordinated by an Operations Management Group consisting of center principal investigators (Burley, NYSGXRC; Joachimiak, MCSG; Montelione, NESG; Wilson; JCSG), the NIH-NIGMS PSI Program Director (John Norvell), and the PSI Steering Committee Chair (Wayne Hendrickson).

Target Selection

Three classes of PSI-2 targets were approved by the NIH-NIGMS Council for structure determination by the Production centers, as reflected in the PSI-2 RFAs. Network Targets are chosen from large protein sequence families for which there is no structural information and from very large phylogenetically diverse protein families, which are inadequately characterized at the level of structure. Network Targets are identified jointly by bioinformatics experts drawn from the four Production centers (PSI Bioinformatics Group: Fiser, NYSGXRC; Godzik, JCSG; Orengo, MCSG; Rost, NESG) with advice from the scientific community (http://www.nigms.nih.gov/News/Reports/psi_targetselect.htm). A draft-style selection process then permits each Production center to select nonredundant Network Target protein families in turn. To date, this quarterly process has allocated >2300 target protein families across the four centers. Once a target protein family is “drafted,” individual structure determination targets are chosen at the center. Factors influencing this final stage of Network Target selection include availability of genomic DNA or cDNA, polypeptide chain length, the presence of low complexity regions or predicted membrane-spanning segments, amino acid composition, predicted hydrophobicity, isoelectric point, and extent of disordered or coiled-coil segments. Seventy percent effort is mandated for Network Targets.

As required by the NIH-NIGMS Council, the second and third classes of PSI-2 targets are chosen independently by each Production center on the basis of biological/biomedical interest (~15% effort, Biomedical Targets) or adopted via nomination from the scientific community (~15%, Community-Nominated Targets). No formal mechanisms have been required to eliminate overlap among the four centers for these two minor target classes. Instead, the four centers coordinate their activities to avoid redundancy.

Experimental Structure Determination Metrics

Between initiation of PSI-1 (September 1, 2000) and October 31, 2007, the entire PSI deposited 2734 experimental protein structures into the PDB, of which 2130 came from the four Production centers. For PSI-2 to date (July 1, 2005–October 31, 2007), the four Production centers were together responsible for 1261 PDB depositions. Their annual rate of PDB deposition for Year Two of PSI-2 (July 1, 2006–June 30, 2007) was 619. Similar annual productivity metrics are targeted for Year Three and beyond. In aggregate, the PSI expects to determine >4000 structures by mid-2010. The average total cost per structure among the Production centers during Year Two of PSI-2 was ~$66K, which is substantially lower than the corresponding amount from Year Five of PSI-1 (~$138K). Average total cost/structure is expected to decline further during the balance of PSI-2.

Contributions to Structural Novelty

Assessments of structure novelty (defined as <30% amino acid sequence identical to any public domain structure upon PDB deposition/release) have been performed by the PSI Knowledgebase. Between July 1, 2006 and June 30, 2007, the PDB received 1324 novel depositions from all sources, of which ~39% (519) came from the PSI (>95% of which came from the Production centers). These 519 depositions account for 73% of all novel PDB depositions from U.S. sources. This remarkable statistic shows that the scope of the PSI is indeed complementary to that of R01-funded structural biology research in the US, which appears to be largely focused on structural/functional studies of proteins for which initial experimental structures already exist in the PDB. Independent assessments of the growth of novel protein structure data and contributions thereto from structural genomics efforts have been made by Chandonia and Brenner (2006) and Levitt (2007).

Thus far in PSI-2, most effort has been devoted to Network Targets from Pfam (http://www.sanger.ac.uk/Software/Pfam/). To date, 1369 Pfam families (with ≥10 members) were identified by the Production centers and 1269 selected for prosecution. Experimental structures covering 198 distinct Pfam target families have been determined, giving an overall success rate of ~14% (12%–18% across the Production centers). These statistics are, however, misleading, because they represent a snapshot in time; many of the targets are still making their way through center structure determination pipelines. Recent Knowledgebase analyses of PSI success rates across all target classes revealed an overall success rate in going from purified protein to PDB deposition of ~12%. Again, this statistic represents a snapshot in time and is not the final word. Many new technologies are currently under development within the PSI, which should further increase structure determination success rates.

Homology Modeling Impact/Leverage

Once an experimental structure of a novel protein is determined by the PSI, homology or comparative protein structure modeling is used to calculate structural models of evolutionarily-related proteins or protein domains. Today, “accurate” homology modeling is thought to be limited to computing structural models for proteins or protein domains that are ≥ 40% identical to the amino acid sequence of the experimental structure, or modeling template (Vitkup et al., 2001). This limitation arises because of challenges inherent in constructing valid sequence-sequence alignments between modeling candidate and modeling template and from the inherent divergence of structures below this threshold. Contrary to assertions from some quarters, homology models do in fact represent more accurate estimates of the true structure of the modeling candidate than the structure of the modeling template itself (Chakravarty and Sanchez, 2004). For reference, typical errors in Cα atomic positions of homology models range from ~0.5–1.0 Å (Cα atomic pair root-mean-square deviation, rmsd) for closely related templates (>80% identity with an error-free alignment) to ~1.0–2.0 Å for templates of intermediate sequence identity (40%–80%) to ~2.0–4.0 Å for distantly related templates (<40%). Clearly, the closer in sequence identity the modeling candidate is to the template, the more accurate the model. Therefore, the Production centers attempt to produce multiple experimental structures for each large protein sequence family to better support the homology modeling process. For reference, Cα atomic coordinate accuracy of a 2 Å resolution experimental X-ray structure is ~0.2 Å. Solution state NMR structures are generally thought to be less accurate than those coming from X-ray crystallography.

To further the PSI goal of making the “structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences,” multiple structures are often determined across large protein families. Broad structural coverage of a protein sequence family or superfamily involves complementary experimental and computational approaches to direct target selection within the family with the goal of ensuring coverage at the level of ~40% identity. Although such homology models tend to be less accurate than either experimental X-ray or NMR structures, they have proven useful for both understanding protein function from structure and for planning functional studies with which to confirm or refute hypotheses. To cite but one of many examples, Dalhus et al. (2007) recently used a homology model based on one of our structures (~40% sequence identity) to guide highly informative functional studies of a new superfamily of DNA glycosylases bearing HEAT-like repeats. We acknowledge that homology models are not generally thought to be useful in drug discovery and other applications requiring the highest possible accuracy.

To rigorously assess the impact of PSI structures on our structural knowledge of other proteins or protein domains, the Production centers report “modeling leverage” statistics annually, corresponding to the number of new “accurate” homology models yielded by each experimental structure. In the absence of standardization, leverage estimates vary with software used to calculate models and model quality assessment, and which protein sequence database is used to select modeling candidates. The PSI Bioinformatics Group has defined standard procedures for assessing modeling leverage across the PSI.

Early in PSI-2, the Milestones and Goals Subcommittee to the PSI Steering Committee (http://www.nigms.nih.gov/Initiatives/PSI/Background/SteeringCommittee.htm) established consensus definitions of program metrics. Among these metrics is “novel modeling leverage” (Liu et al., 2007) estimating the number of proteins (and residues) that can be modeled using each experimental protein structure, and which could not be accurately modeled at the time of its PDB deposition. Acceptable candidates for homology modeling are defined to be those proteins (or protein domains) that satisfy a PSI-Blast-based sequence similarity measure (~40% identity or better) between modeling candidate and template. As described above, this criterion qualifies modeling candidates likely to give “accurate” homology models (Cα atomic coordinate rmsd < ~1.0–2.0 Å). This definition also permits meaningful and reliable comparisons across the PSI and the PDB. Between PSI inception and August 15, 2007, the Production centers alone enabled calculation of new homology models for ~110,000 proteins or protein domains (> 22,000,000 residues). In aggregate, structural genomics centers worldwide generated an impressive ~27% of the novel modeling leverage derived from all PDB depositions during same period (~179,000 versus ~653,000).

PSI Production Centers Contributions to Biology

The overarching goal of the PSI is to explore the universe of protein sequences/structures. It has become evident during the past decade that discovery of polypeptide chain “folds” has not kept pace with discovery of new protein sequences, which are increasing exponentially (i.e., many proteins seemingly unrelated at the level of amino acid sequence are similar in overall polypeptide chain fold, but have diverged substantially in sequence and to a lesser extent structure). In the context of the burgeoning PDB (now archiving more than 48,000 entries), we appreciate that the concept of discrete protein folds may no longer be useful. Protein structure space is probably best thought of as a continuum (Honig, 2007), albeit one with densely populated volume elements, such as those encompassing the $(\alpha\beta)_8$ triose phosphate isomerase (TIM) barrel superfamily.

The continuum nature of protein structure/fold space probably lies at the heart of our inability to accurately predict protein structure/fold family membership for many newly discovered protein sequences. High-throughput structure determination efforts in general, and the PSI in particular, are providing new insights into the tremendous diversity of structures that have evolved to support diverse biochemical/biological functions. Among targets with highest potential impact in this regard are so-called “singletons” (i.e., individual proteins or small groups of proteins with no discernible sequence-sequence relationship to a protein of known experimental structure). Almost invariably, experimentally determined structures of such singletons or “fewtons” reveal similarity to a protein of known structure, sometimes connecting disparate protein sequence families for the first time and providing unexpected insights into evolutionary relationships.

When the PSI was launched we could not have anticipated the richness and diversity of the genomes and corresponding protein sequences from which we can now choose structure determination targets. Recent forays into various so-called “metagenomes” from various sources (The New Science of Metagenomics; http://www.nap.edu/catalog/11902.html) suggest that there will soon be an even more dramatic increase in the number and diversity of protein sequences that lack structural information. During PSI-2, the Production centers have responded to this challenge by expanding Network Target selection and Community-Nominated Target adoption to encompass large protein superfamilies (many of which have > 10,000 members) from both genomic and metagenomic sources with the goal of providing structural coverage across as many branches of protein superfamilies as possible.

Illustrative examples from the efforts of each of the Production centers follow: The JCSG has analyzed the FMN-binding split barrel family, wherein the protein core has remained unchanged despite acquisition of various appendages that support binding of different cofactors and distinct enzymatic properties. The MCSG is studying bacterial transcription factors, which control the cell’s genetic program and are often linked to human disease. The NESG has provided extensive structural coverage of proteins involved in FeS cluster assembly, ubiquitin-like proteins, and START domains that are important in intracellular metabolite transport. The

NYSGXRC is studying the enolase/amido-hydrolase superfamily with the goal of helping to elucidate how addition of small domains that cap the active site lying at the mouth of the TIM barrel support substrate recognition (Hermann et al., 2007). The enolases/amidohydrolases were nominated by an NIH-NIGMS funded Program Project team led by John Gerlt and Frank Raushel.

Independent of PSI Network and Community-Nominated Targets, each Production center is pursuing its own biomedical targets. The JCSG is studying the "Central Machinery of Life" by focusing on proteins that are conserved across all domains of life. The MCSG is studying proteins from human pathogens, and recently discovered an important protein secretion apparatus within a *P. aeruginosa* virulence locus (Mougous et al., 2006). NESG is studying and modeling a large number of human proteins involved in cancer and developmental biology, with emphasis on protein-protein interaction networks (http://nmr.cabm.rutgers.edu:9090/HCPIN/). The NYSGXRC has undertaken a systematic study of human and pathogen protein phosphatases of unknown structure, thereby making significant contributions to our understanding of the phosphatome (Almo et al., 2007).

Many PSI structures represent important drug targets, such as essential pathogen enzymes, protein kinases and phosphatases, proteases, DNA-and RNA-binding proteins, and molecular chaperones. During PSI-1, the TB Structural Genomics Consortium contributed approximately two-thirds of the structures of *Mycobacterium tuberculosis* (TB) proteins now available in the PDB. Today, at least nine of these TB protein structures represent industry drug discovery targets.

During PSI-1, the JCSG began exploring complete structural coverage of *Thermotoga maritima*, a thermophilic bacterium that lives in geothermal marine sediments. Their efforts fostered a community-wide collaboration that today claims experimental structures for 273 proteins (with contributions from other PSI centers and the structural biology community) and homology models for 886 proteins, which together cover ~62% of the *T. maritima* proteome and 92% of predicted soluble, nonorphan proteins (L. Jaroszewski and A. Godzik, personal communication). Of particular importance coming from this collective effort is the opportunity to compare and contrast diversity of the metabolic pathways in different organisms, wherein new enzymes have evolved as alternates to those found in standard textbook representations. Some of these proteins may represent drug discovery targets as they occur in certain microorganisms and not in humans. For example, the JCSG structure of a *T. maritima* thymidylate synthase complementing protein (TM0449; PDB code: 1024; Kuhn et al., 2002) revealed differences in protein fold, cofactor requirement, and enzymatic mechanism as compared to eukaryotic thymidylate synthase. Further evidence of PSI impact on the *Thermotoga* community came recently in the guise of a broadly attended *Thermotoga* community workshop sponsored by the JCSG (http://research.calit2.net/metagenomics/thermotoga/index.php).

Finally, PSI centers are involved in collaborations with other structural genomics consortia, The NIH Roadmap, and many individual laboratories. To date, PSI centers have published more than 990 scholarly articles describing both experimental protein structures and technology developments. We submit that the PSI represents a valuable hypothesis-generating enterprise, producing reagents, 3D structures, and insights that together can serve as cornerstones for investigator-initiated, hypothesis-driven research programs.

PSI Production Centers Contributions to Technology Development

To establish high-throughput structure determination pipelines, PSI-1 centers were challenged not only to revamp and streamline existing methods, but also to develop and invent new tools and technologies. In doing so, we frequently "stood on the shoulders of giants," exploiting many earlier discoveries and innovations. With PSI infrastructure (specifically TargetDB and PepcDB), however, tool/technology development took on a very different hue from that seen previously in structural biology. Given the enormity of the challenge of determining thousands of novel and diverse protein structures, we were forced to test rigorously competing, parallel approaches instead of relying on anecdotal reports of individual successes. Technologies coming from the PSI and other structural genomics efforts worldwide represent important new means of producing both protein samples and structures that can be applied to much of biology and can be implemented in many laboratories, big and small.

During the past seven plus years, we have addressed three major imperatives facing structural biologists: 1) increasing access to biological samples for structural studies and reducing the amount of material required for structure determination; 2) increasing production of single crystals and reducing the size and number of crystals needed for structure determination; and 3) improving phasing methods to accelerate structure determination. We now have faster, more robust, and more cost effective approaches to gene cloning and protein production, isotope labeling for NMR, crystallization, structure determination using X-ray crystallography and solution state NMR, structure refinement and validation, and comparative protein structure modeling. We also understand many of the limitations of these strategies and what opportunities remain to improve further protein production and structure determination.

Given the breadth of contributions, a comprehensive review of the technologies developed with PSI support is well beyond the scope of this commentary; instead, we provide below selected examples of technological advances from each of the Production centers.

Ligation Independent Cloning

PSI laboratories have established Ligation Independent Cloning (LIC) as an effective source of open reading frames for target genes of interest (Dieckman et al., 2002). Unlike traditional methods reliant on restriction enzymes and DNA ligases, LIC provides unique cloning sites, is directional and highly efficient, and can be implemented readily in parallel formats with minimal optimization. The LIC approach can also be used to generate multiple constructs from a single template, is compatible with multiple expression vectors, and is well suited to both robotic and manual cloning/expression testing.

Expression Vectors

Multiple expression vectors optimized for structural biology applications have been developed within the PSI. Design and use of these new vectors has been rigorously tested with thousands of genes, many of which encode challenging targets such as large proteins (>60 kD) and domains of eukaryotic proteins (Donnelly et al., 2006). The LIC vector pMCSG19 produces highly soluble fusion proteins in which maltose binding protein (MBP), a His$_6$-tag, and the desired target protein are separated by two distinct protease cleavage sites: MBP-site1-His$_6$-site2-target protein. In vivo cleavage in *E. coli* at the first site by a coexpressed protease separates untagged MBP from the His $_6$-site2-target protein fusion construct. Following Ni^{2+}-ion affinity chromatography, the target protein can be released from the His$_6$-tag by the second protease and purified to homogeneity. The presence of MBP during expression improves target protein solubility. PSI expression vectors are available from the Materials Repository.

Protein Expression via Autoinduction

Most PSI target proteins are expressed in *E. coli* using inducible T7 RNA polymerase expression systems. F. William Studier (NYSGXRC) has developed autoinducing culture media to support high density growth (Studier, 2005). Autoinduction allows efficient screening of many clones in parallel for expression and solubility, as cultures only have to be inoculated and grown to saturation. Yields of target protein are typically several-fold higher than with conventional IPTG induction. Autoinducing media have also been developed for labeling proteins with selenomethionine and $^{15}N/^{13}C$. Studier's Overnight Express autoinducing media are available from Novagen (http://www.emdbiosciences.com/html/NVG/home.html).

Protein Domain Elucidation via Limited Proteolysis/Mass Spectrometry

Anecdotal reports have suggested that limit digests of proteins (i.e., compact globular domains) represent good candidates for crystallization/structure determination. The NYSGXRC and SGX Pharmaceuticals, Inc. (formerly Structural GenomiX, Inc.) together developed parallel/automated instrumentation and procedures for limited proteolytic digestion of target proteins followed by mass spectrometry (MS) data acquisition/analysis. Their results demonstrated that integration of MS with limited proteolysis provides accurate definition of domain boundaries within larger polypeptide chains. A study of 400 PSI-1 proteins documented that targets yielding diffraction-quality crystals are indeed typically resistant to proteolysis, whereas those failing crystallization trials are not (Gao et al., 2005). Moreover, targets failing "first pass" attempts with the NYSGXRC or SGX high-throughput pipelines can be re-engineered following MS/proteolysis domain mapping (i.e., flexible N or C termini are truncated or mobile internal loops are deleted), thereby increasing likelihood of successful structure determination. NMR chemical shift data on full-length and truncated proteins have repeatedly documented that truncation does not significantly disturb the well-ordered regions of the protein structure, while also providing samples with improved spectral properties. Other structural genomics centers and structural biology laboratories have adopted similar MS/proteolysis strategies for their recalcitrant targets.

Expression Construct Definition via Enhanced H/D Exchange MS

The JCSG has used high-throughput deuterium exchange MS (DXMS) for rapid identification of unstructured regions in proteins (Pantazatos et al., 2004). Initial tests with 24 *T. maritima* proteins accurately mapped unstructured/disordered regions for 21 of these cases. DXMS results were then correlated with the propensity of such targets to crystallize. Truncations defined via DXMS have demonstrated improved crystallization properties and increased likelihood of successful structure determination. This approach, like MS/proteolysis, represents a rapid, general method suitable for any target of interest in any laboratory. Recently, DXMS has also been adopted by the NESG to optimize protein expression constructs for NMR.

Crystallization Feasibility Analyses

As discussed earlier, TargetDB and PepcDB have enabled retrospective process analyses with which to establish optimal protocols for every step between gene cloning and structure determination. These tools have also been used to analyze the likelihood of success for a particular target by seeking to correlate out-comes with target protein properties. The JCSG identified sequence features and biophysical properties that correlate with successful protein production and crystallization, which were combined into crystallization and X-ray structure determination feasibility scores (Slabinski et al., 2007). Their method was tested with a jackknife procedure and validated with independent benchmark data. Application of "crystallization feasibility" scoring to target selection is helping increase success rates, lower costs, and increase speed. This a priori strategy permits recognition of target protein sequences that are more likely to yield soluble protein and crystals. In situations where one has the choice of cloning a particular target protein gene from multiple organisms, such estimates permit prioritization and increase success rates. Alternatively, if the choice of target protein gene is fixed, such an a priori approach can be used for prioritizing expression attempts with distinct truncated forms of the target protein. Analyses of non-PSI PDB depositions suggest that these same features and properties influence success rates in non-high-throughput structure determination, thereby making the approach one of general interest.

Semiautomated X-Ray Structure Determination

The MCSG has integrated synchrotron data collection, data reduction, phasing, and model building into an automated procedure that accelerates structure determination (Minor et al., 2006). Their software attempts to determine the structure using different algorithms and approaches, rapidly converting diffraction data into an interpretable electron density map and, for smaller structures, into an initial refinement model. The typical out-come is an interpretable electron density map with a partially built protein structure (usually 70%–80% of the polypeptide chain). In favorable cases, a near final refined structure can be produced without manual intervention. Current developments are aimed at very fast structure determination to provide feedback during the synchrotron experiment. Efforts are also devoted to improving the phasing estimation and model building. Similar developments are also ongoing in other PSI centers. The software has been successfully tested on over 370 PDB deposits and is in use in non-PSI laboratories. Similarly successful strategies have been developed in other PSI centers. Considerable Production center efforts/resources have also been devoted to synchrotron X-ray beamline automation at SBC-CAT and SGX-CAT (Advanced Photon Source), and at the National Synchrotron Light Source and the Stanford Synchrotron Radiation Laboratory, all of which contribute to PSI structure determinations of novel proteins.

G-Matrix FT NMR Data Collection and Automated NMR Data Analysis

The NESG has developed automated methods of protein NMR data analysis, NMR structure quality assessment, and reduced-dimensionality G-matrix FT-NMR (GFT-NMR) data collection (Liu et al., 2005). GFT-NMR provides 4D/5D NMR spectral information in 3D or 2D spectra, thereby reducing data collection times. This approach is particularly advantageous when using high-sensitivity cryogenic probes, as the times required for carrying out 4D and 5D spectral data collection are much longer than the minimum time needed to achieve adequate sensitivity with 3D data collection. The peak patterns that appear in GFT NMR spectra are also useful in developing automated peak recognition software. GFT NMR data collection has been integrated with methodology for semiautomated data analysis and is routinely used by the NESG, along with conventional tripleresonance NMR, to determine smaller protein structures (6–25 kD). Reduced dimensionality and GFT-like data collection methods have also been adopted by other structural genomics consortia.

Microgram Sample NMR Structure Determination and Sample Optimization

The NESG has demonstrated near complete resonance assignment and 3D structure determination for a 68-residue *M. mazei* protein, TRAM, using only 72 μg of protein (6 μl, 1.4 mM). This success demonstrates the utility of 1 mm microcoil-probe NMR for proteins that can only be produced in limited quantities (Aramini et al., 2007). The 1 mm micro-probe, utilizing sample volumes of <10 μl, has been incorporated into an NESG solution state NMR screening pipeline for expression construct/sample buffer optimization. Recently developed cryogenic microcoil NMR probes have the potential to make such microscale analyses of proteins routine.

The PSI Large-Scale Centers have addressed significant technical and scientific challenges and substantially improved methods for structure determination of novel proteins. With these new tools and strategies, we have established robust structure determination pipelines with the aggregate capacity to determine more than 600 protein structures annually at a current average total cost/structure of less than $70,000. Such advances have not come at the expense of quality. Independent evaluations have documented that PSI structures are typically of higher quality than that of structures deposited into the PDB by nonstructural genomics laboratories (Brown and Ramaswamy, 2007). PSI experimental protocols, both general and specific, are freely available from TargetDB/PepcDB and PSI expression vectors/expression constructs can be obtained from the Materials Repository. Information regarding PSI robotics/instrumentation solutions, both custom and commercial, can be accessed via the Technology Development module of the Knowledgebase.

PSI Philosophy

The PSI aims to explore the protein structure universe by characterizing as many novel targets as possible. This holistic approach makes relatively few assumptions regarding the function of a given gene product. Instead, it emphasizes the benefits of understanding, to the fullest extent possible, the diversity of protein shapes represented in the natural world and the importance of providing reference structures for as many protein sequence families as possible. Adoption of such an approach by the architects of the Human Genome Project, while somewhat controversial at the time, created enormous value for the biomedical research community. The PSI has embarked on a similar “voyage of discovery,” which is being both vindicated and rewarded by an impressive array of publications and the same sense of excitement and wonder every time unexpected structural/mechanistic connections are made between or among seemingly disparate groups of proteins. As a discovery and knowledge-building endeavor, the PSI promises to improve our understanding of evolution at the molecular level by extending analyses from one to three dimensions with all of the richness and import that structure brings to biology.

Thoughts on the Future of the Protein Structure Initiative

The goals and achievements of the PSI are tremendously exciting. The success of the PSI in furnishing large numbers of experimental structures to the public, which have yielded even larger numbers of calculated homology models, has vindicated the central goal of the PSI to make 3D atomic-level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences. In thinking about future directions for the PSI, we consider the following issues: 1) recent growth in the number of newly discovered protein sequence families that is outpacing the number of novel structures being deposited into the PDB; 2) limitations in our ability to recognize structurally similar protein families and evolutionary relationships solely on the basis of sequence-sequence relationships; 3) limitations in the accuracy of high-throughput homology modeling of sequences that are distantly related to experimental template structures; 4) the “functional annotation” gap that exists for many PSI targets/structures; and 5) the promise of harnessing the machinery of the PSI to advance high-priority areas in biomedical research. The first three of these issues are part and parcel of the same fundamental challenge. Our understanding of evolutionary relationships at the level of protein sequence and structure significantly lags our knowledge of genome sequences, and this gap is widening every day. To fully exploit the wealth of available genomic and protein structural information and to realize the central goal of the PSI, we urge that current investments aimed at improving homology modeling be augmented and complemented with significant new investment in studies of protein evolution in 3D. Moving on to the issue of the “functional annotation” gap, we acknowledge some frustration with current PSI policies. A wealth of research opportunities pass before our eyes on a daily basis, making us sympathetic to calls from members of the scientific community urging more functional study of PSI targets/ structures. The challenge beyond PSI-2 will be to establish “rules of engagement” that are acceptable to the U.S. scientific community, particularly in the face of recent year-on-year reductions of NIH funding in real terms. We do recognize the logic of the NIGMS in discouraging use of PSI funds for experimental functional characterization that might limit output of structures and place the PSI mission in jeopardy. Such distractions, no matter how intellectually rewarding, would almost certainly have adversely impacted development of high-through-put methodologies, which benefit the entire scientific community. We look forward to contributing to this lively debate, both in the scientific literature and in policy making discussions. Finally, the infrastructure of the PSI represents an important national resource that should, in our view, be harnessed as part of one or more coordinated efforts to address high priority challenges in biomedical research. We might choose to apply the PSI structural genomics platform to systems biology, human microbiomes, human drug discovery targets, essential proteins found in microbial pathogens, or proteins from bacteria that could help address the global scourge of environmental pollution.

Postscript

In closing, it is remarkable that many of the criticisms currently leveled at the PSI were made by erstwhile detractors of the Human Genome Project.

ACKNOWLEDGMENTS

We wish to thank the members of the JCSG (supported by NIH-NIGMS Cooperative Agreement U54 GM074898), MCSG (U54 GM074942), NESG (U54 GM074958), and NYSGXRC (U54 GM074945), both past and present, for all of their efforts on behalf of the PSI. We also gratefully acknowledge the many contributions of the other PSI centers, collaborators outside the PSI, the PDB, and the NIH-NIGMS that have enhanced our own activities. Finally, we recognize the valuable contributions made by structural genomics researchers throughout the rest of the world. Drs. Burley, Joachimiak, Montelione, and Wilson and their Production centers receive support from the NIH-NIGMS PSI.

REFERENCES

Almo SC, Bonanno JB, Sauder JM, Emtage JS, Dilorenzo TP, Malashkevich V, Wasserman SR, Swaminathan S, Eswaramoorthy S, Agarwal R, et al. Structural genomics of protein phosphatases. J. Struct. Funct. Genomics 2007;8:121–140. [PubMed: 18058037]Published online December 5, 2007

Aramini JM, Rossi P, Anklin C, Xiao R, Montelione GT. Microgram-scale protein structure determination by NMR. Nat. Methods 2007;4:491–493. [PubMed: 17496898]

Brown EN, Ramaswamy S. Quality of protein crystal structures. Acta Crystallogr. D Biol. Crystallogr 2007;63:941–950. [PubMed: 17704562]

Chakravarty S, Sanchez R. Systematic analysis of added-value in simple comparative models of protein structure. Structure 2004;12:1461–1470. [PubMed: 15296739]

Chandonia J-M, Brenner SE. The impact of structural genomics: expectations and outcomes. Science 2006;311:347–351. [PubMed: 16424331]

Dalhus B, Helle IH, Backe PH, Alseth I, Rognes T, Bjøråsm M, Laerdahl JK. Structural insight into repair of alkylated DNA by a new superfamily of DNA glycosylases comprising HEAT-like repeats. Nucleic Acids Res 2007;35:2451–2459. [PubMed: 17395642]

Dieckman L, Gu M, Stols L, Donnelly MI, Collart FR. High throughput methods for gene cloning and expression. Protein Expr. Purif 2002;25:1–7. [PubMed: 12071692]

Donnelly MI, Zhou M, Millard CS, Clancy S, Stols L, Eschenfeldt WH, Collart FR, Joachimiak A. An expression vector tailored for large-scale, high-throughput purification of recombinant proteins. Protein Expr. Purif 2006;47:446–454. [PubMed: 16497515]

Gao X, Bain K, Bonanno JB, Buchanan M, Henderson D, Lorimer D, Marsh C, Reynes JA, Sauder JM, Schwinn K, et al. High-throughput limited proteolysis/mass spectrometry for protein domain elucidation. J. Struct. Funct. Genomics 2005;6:129–134. [PubMed: 16211509]

Hermann JC, Marti-Aborna R, Federov AA, Federov E, Almo SC, Shoichet BK, Raushel FM. Structure-based activity prediction for an enzyme of unknown function. Nature 2007;448:775–779. [PubMed: 17603473]

Honig B. Protein structure space is much more than the sum of its folds. Nat. Struct. Mol. Biol 2007;14:458. [PubMed: 17549077]

Kuhn P, Lesley SA, Mathews II, Canaves JM, Brinen LS, Dai X, Deacon AM, Elsliger MA, Eshaghi S, Floyd R, et al. Crystal structure of thy1, a thymidylate synthase complementing protein from *Thermotoga maritima* at 2.25 Å resolution. Proteins 2002;49:142–145. [PubMed: 12211025]

Levitt M. Growth of novel protein structural data. Proc. Natl. Acad. Sci. USA 2007;104:3183–3188. [PubMed: 17360626]

Liu G, Shen Y, Atreya HS, Parish D, Shao Y, Sukumaran DK, Xiao R, Yee A, Acton TB, Arrowsmith CH, et al. NMR data collection and analysis protocol for high-throughput protein structure determination. Proc. Natl. Acad. Sci. USA 2005;102:10487–10492. [PubMed: 16027363]

Liu J, Montelione GT, Rost B. Novel leverage of structural genomics. Nat. Biotechnol 2007;25:849–851. [PubMed: 17687356]

Minor W, Cymborowski M, Otwinowski Z, Chruszcz M. HKL-3000: the integration of data reduction and structure solution–from diffraction images to an initial model in minutes. Acta Crystallogr. D Biol. Crystallogr 2006;62:859–866. [PubMed: 16855301]

Mougous JD, Cuff ME, Raunser S, Shen A, Zhou M, Gifford CA, Goodman AL, Joachimiak G, Ordonez CL, Lory S, et al. A virulence locus of *Pseudomonas aeruginosa* encodes a protein secretion apparatus. Science 2006;312:1526–1530. [PubMed: 16763151]

Pantazatos D, Kim JS, Klock HE, Stevens RC, Wilson IA, Lesley SA, Woods VL Jr. Rapid refinement of crystallographic protein construct definition employing enhanced hydrogen/deuterium exchange MS. Proc. Natl. Acad. Sci. USA 2004;101:751–756. [PubMed: 14715906]

Slabinski L, Jaroszewski L, Rodrigues AP, Rychlewski L, Wilson IA, Lesley SA, Godzik A. The challenge of protein structure determination lessons from structural genomics. Protein Sci 2007;16:2472–2482. [PubMed: 17962404]

Studier FW. Protein production by auto-induction in high density shaking cultures. Protein Expr. Purif 2005;41:207–234. [PubMed: 15915565]

Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat. Struct. Biol 2001;8:559–566. [PubMed: 11373627]