Published in final edited form as:

Infect Disord Drug Targets. 2009 November ; 9(5): 459–474. doi:10.2174/187152609789105704.

Benefits of Structural Genomics for Drug Discovery Research

Marek Grabowski^{1,2,3}, Maksymilian Chruszcz^{1,2,3}, Matthew D. Zimmerman^{1,2,3}, Olga Kirillova^{1,2}, Wladek Minor^{1,2,3,*}

¹Department of Molecular Physiology and Biological Physics, University of Virginia, 1340 Jefferson Park Avenue, Charlottesville, VA 22908, USA

²Midwest Center for Structural Genomics

³Center for Structural Genomics of Infectious Diseases

Abstract

While three dimensional structures have long been used to search for new drug targets, only a fraction of new drugs coming to the market has been developed with the use of a structure-based drug discovery approach. However, the recent years have brought not only an avalanche of new macromolecular structures, but also significant advances in the protein structure determination methodology only now making their way into structure-based drug discovery. In this paper, we review recent developments resulting from the Structural Genomics (SG) programs, focusing on the methods and results most likely to improve our understanding of the molecular foundation of human diseases. SG programs have been around for almost a decade, and in that time, have contributed a significant part of the structural coverage of both the genomes of pathogens causing infectious diseases and structurally uncharacterized biological processes in general. Perhaps most importantly, SG programs have developed new methodology at all steps of the structure determination process, not only to determine new structures highly efficiently, but also to screen protein/ligand interactions. We describe the methodologies, experience and technologies developed by SG, which range from improvements to cloning protocols to improved procedures for crystallographic structure solution that may be applied in “traditional” structural biology laboratories particularly those performing drug discovery. We also discuss the conditions that must be met to convert the present high-throughput structure determination pipeline into a high-*output* structure-based drug discovery system.

Keywords

Structural Genomics; structure determination; protein-small molecule complexes; infectious disease; drug discovery

Introduction

Three-dimensional structures of both *bona fide* and putative drug targets have long been one of the many tools used by drug discovery researchers to identify and improve lead

---

*Corresponding authors: Wladek Minor (wladek@iwonka.med.virginia.edu).

compounds. While only a subset of drugs coming to market at the moment were developed in complex with structure-based methods, such methods may become increasingly important in the future. This is due to an exponential growth of the number of protein structures available and an explosion of new technologies for determining and analyzing those structures. The field of structural genomics (SG) has played a major role in both of these advances.

Structural genomics will soon celebrate the 10th anniversary of the Protein Structure Initiative (PSI), which was funded by the U.S. National Institutes of Health in September of 2000 [1]. The PSI is one of the largest of many SG programs (Table 1) that have been established on almost every continent. These worldwide programs have (or *had*) very different motivations but share one universal goal: to develop a high-throughput, or preferably high-*output*, structure determination pipeline for very fast progress from cloning to deposition of elucidated structures into the Protein Data Bank [2,3,4].

The output of some SG programs or groups within those programs was lower than initial estimates in terms of deposited structures. It was discovered that the solution of protein structures on a massive scale is much more complicated than initially expected, revealing that those early expectations were probably unrealistic. Accordingly, there have been some reductions of scale and scope of SG programs around the world. However, the most productive SG groups still deposit around 200 structures per year. In other words, a structure is deposited every other day including weekends and holidays. At first, SG groups put substantial effort into automation of the experimental stages of the structure determination process: cloning, expression, crystallization, data collection, processing, phasing, model building, and deposition. In contradiction to anecdotal experience, it was shown that there is no clear single bottleneck in the structure determination process [5]. Both X-ray crystallography and NMR spectroscopy are multi-step processes and a failure at any single step of the staircase leading up to structure solution results in a complete failure.

Improving the chances for success of each of the steps in the process is not always an easy task, either within SG projects or small traditional structural biology laboratories, including those engaged in structure-based drug discovery. For lack of a better term, in this review we will use the term “traditional” structural biology (tSB) for structural biology research conducted outside structural genomics initiative. In solving thousands of structures, SG groups have amassed a vast amount of experience. This allows not only the discovery and ultimately mastery of the most efficient experimental techniques, but most of all, a more precise evaluation of the probability of success for any particular step of the structure solution staircase. The benefits of that knowledge to traditional structural biology laboratories could be enormous, assuming that SG groups adequately share their expertise. Many opponents of SG argue that majority of the structures solved by SG would still be solved, *albeit* at a slower pace, by tSB researchers. Even if we disregard SG structures totally, we have no doubt that the knowledge acquired by solving hundreds of structures, encapsulated in the form of new experimental protocols and new software, greatly surpasses the costs to date of SG programs worldwide.

There is one bottleneck that should be addressed separately. The number of papers published by SG groups, although impressive, is an order of magnitude smaller than the number of structures. This shows that despite all the automation, processes that require significant “brain engagement” still remain a critical bottleneck. The high output of PSI production centers set the groundwork for SG programs that follow a model of research that pays more attention to the relationship between structure and function. This is the main aim of two recently formed US centers: the Center for Structural Genomics of Infectious Diseases (CSGID) and the Seattle Structural Genomics Center for Infectious Diseases (SSGCID), funded by the National Institute of Allergy and Infectious Diseases (NIAID), both of which work on proteins from viral, bacterial and eukaryotic pathogens. The previously formed Structural Genomics Consortium (SGC), a collaboration of researchers from three countries, focuses on eukaryotic proteins related to public health. The fast ramp-up of productivity for these centers has been made possible by application of technologies developed by the PSI. These centers strive for determination of the three-dimensional structure of known and potential drug targets, which might provide helpful insights into the mechanisms of molecular interactions. They are also studying complexes of proteins with compounds that are considered candidates for inhibitor design. This review will discuss developments in the first decade of the SG programs that have made the most significant impact on traditional structural biology laboratories that are working on projects related to human health.

Target Proteins and Structural Results of SG

At first, it may seem that the strategies for target selection in SG projects are completely irrelevant to a traditional biology or drug discovery laboratory, which usually works on a specific project for many years and may spend only a small fraction of that time on structure elucidation. On the other hand, SG projects, especially phase 2 PSI centers, have focused on maximizing the coverage of structural representation of protein families [6], which may provide structural insights into new projects. While it has been well established that SG deposits provide, on average, more structural coverage of the protein fold universe than traditional deposits [7-8], opponents of SG have claimed that the structures determined in these programs are mostly irrelevant to hypothesis-driven research [9].

However, in the process of structurally charting the uncharacterized regions of the protein fold universe, SG has significantly contributed to a number of structurally characterized Gene Ontology (GO) terms (Fig. (1)). Gene Ontologies are standardized vocabularies describing biological processes, molecular functions and cellular components [10]. By their very nature, these dictionaries describe areas of biology being investigated by the wider scientific community. As of 2007, SG programs provided the first structural representatives for 59 molecular functions and 46 biological processes assigned to Pfam [11] families, which constitute about 8% of all structurally characterized functions and processes. About half of all structurally characterized functions and processes (as classified by Pfam family) had structural representatives both among SG and tSB deposits. By maximizing structural coverage of molecular functions and biological processes, SG may provide structural representatives for potential new drug targets even before they are identified as such.

As SG has only existed for about a decade, and drug development is a long process, there are few, if any, drugs in the marketplace whose development could be directly attributable to SG. In fact, the authors are not aware of any. Indeed, the number of drugs whose discovery could be unequivocally credited to structural biology as a whole was recently estimated to be around 10 [12, 13]. It is difficult to predict which protein structures will lead to new drugs in the marketplace, but SG has solved many structures that are drug targets or potential drug targets. For example, the CSGID determined the 1.8 Å structure of a mutant of the pre-cleavage cysteine protease domain of the *V. cholerae* RTX toxin in complex with the activating factor inositol hexaphosphate (PDB id 3FZY). The RTX toxin is a large, multifunctional bacterial virulence factor which uses the protease domain for autoproteolysis and delivery of catalytic domains of the toxin in the eukaryotic cytosol [14].

Another possible drug target structurally characterized by SG is dipeptidyl-peptidase III, an aminopeptidase that is involved in the mammalian pain-modulation pathway [15,16] and in the endogenous oxidative stress management system in humans [17]. The SGC determined the structure of the enzyme and found that while some elements are conserved, the peptidase fold is distinct from other families of dipeptidyl proteases (PDB id 3FVY).

Other drug targets include sortases, which anchor surface proteins to the cell wall of Gram-positive pathogens through specific interactions, and play a crucial role in virulence. The MCSG determined the structure of sortase B from both *Bacillus anthracis* (PDB id 1RZ2) and *Staphylococcus aureus* (PDB id 1NG5) [18, 19]. Another example is aminopeptidase N, which is a zinc-dependent exopeptidase that is involved in the process of tumor invasion and metastasis. The MCSG solved the structure of the aminopeptidase N from *Neisseria meningitidis*, which is a suitable target for antibacterial drugs and anti-cancer drug design (PDB id 2GTQ) [20],

There is also a non-trivial impact of SG on improvement of the structure determination process in tSB labs. First of all, SG has fundamentally changed the structure determination culture, as the typical time between a diffraction experiment and the deposit of a refined structure is now measured in days or weeks rather than months or years for both SG and tSB labs. For example, 249 structures collected after 1 Jan 2008 were refined, validated and deposited 15 days or less after data collection. In recent years, SG produced around 50% of unique PDB deposits. These are defined as those having no more than 30% sequence identity to any PDB structure deposited previously. These unique structures can often help in determination of the structure of homologous proteins. In particular, as of March 2009, 303 SG structures served as a starting model for molecular replacement methods used to solve 441 tSB structures. In addition, out of about 17,000 tSB structures deposited in the PDB since the start of the second phase of PSI (October 2005 to February 2009), around 6% (950) exhibited at the moment of their deposition more than 30% sequence identity to a structure deposited previously by SG (Fig. (2)). These data suggest an interesting synergy between SG and traditional hypothesis-driven research. In some sense, they could be regarded as follow-ups of SG structures done, in most cases unintentionally, by tSB researchers. Though it cannot be proved how much information from SG-solved homologs has been used by tSB researchers, having a homologous structure in the PDB with known experimental conditions

may be of great value not only for the determination of new protein structures but also for further functional studies.

It should be also noted that all large-scale PSI centers accept requests from the larger academic community for pursuing high-impact targets. The four large-scale PSI centers have so far solved around 100 structures (out of about 1850 submitted requests) of community-nominated targets. The two new structural genomics centers CSGID and SSGCID solve structures of proteins from infectious disease (ID) pathogens. These protein targets are chosen both from the list of NIAID Category A-C priority pathogens [12], which includes drug targets, drug target homologs, essential genes, and virulence factors, and by request from the larger scientific community. For example, the CSGID target list contains at present over 150 proteins requested by the infectious diseases research community.

Both the PSI and the new infectious disease SG centers have already made important contributions to the structural mapping of infectious pathogens. The Tuberculosis Structural Genomics Consortium (TBsgc) [21], which started in 2000 as part of the PSI, determined the structures of 215 proteins from *Mycobacterium tuberculosis*, out of 594 total structures for this organism that are currently deposited in the PDB. The center solved the structures of two-thirds of the unique *M. tuberculosis* structures (more than 100 as of 2006) in the PDB [22, 23]. The knowledge of three-dimensional structures, integrated with bioinformatics and biochemical studies, is gradually extending our understanding of tuberculosis and enabling formulation of testable functional hypotheses. Unfortunately, four years passed after the closure of the TB as a PSI center before its impact was finally realized [24]. The Structural Genomics of Pathogenic Protozoa (SGPP) [25] and its continuation Medical Structural Genomics of Pathogenic Protozoa (MSGPP) focused on the proteomes of *Plasmodium falciparum*, *Leishmania major*, *Trypanosoma brucei*, and *Trypanosoma cruzi*. These organisms are responsible for the major global health threats of malaria, leishmaniasis, and both sleeping and Chagas sickness. These centers worked out numerous problems in bacterial production of eukaryotic proteins and have so far contributed 70 structures of proteins from parasitic organisms into the PDB. The Structural Genomics Consortium (SGC) deposited over 700 protein structures, most from the human genome and many containing functional ligands [26,27]. For example, the enzymes of the polyamine biosynthesis pathway in *P. falciparum*, the causative agent of malaria, have been proposed as possible drug targets to treat the disease. The SGC determined the structure of spermidine synthase, both in its apo-form (PDB id 2PSS) and in complex with two strong inhibitors, 4MCHA (PDB id 2I7C) and AdoDATO (PDB id 2PT9) [28].

Among the structures of proteins from pathogens deposited since 2004, the PSI and the ID centers solved half of the total number from *Shigella flexneri* and *Vibrio cholerae*, about a third of the total from *Salmonella typhimurim*, and *Lysteria monocytogenes*, close to a fifth from *Bacillus anthracis*, and almost all from *Vibrio parahaemolyticus* (Table 2). Several structures solved by the infectious disease centers are of proteins found in pharmaceutical screens or identified in DrugBank [29]. For example, 9 out of the 73 structures determined so far by the CSGID fall into this category. These drug targets include holo-acyl carrier protein synthases, IspF orthologs, ACP S-malonyltransferases, peptide deformylase, glutamate racemase, and a menaquinone-specific isochorismate synthase.

Another example of important drug targets are human protein kinases, due to the role that the superfamily plays in many different signaling pathways. In the past 2 years, the number of unique human kinase domains solved by SG efforts has surpassed the number published by the pharmaceutical industry (Aled Edwards, personal communication).

Besides the direct effect due to structural determination of drug targets, SG benefits drug discovery indirectly, by improving structural biology tools and creating new resources. In the future, understanding of infectious diseases should benefit from the synergy between the work of SG centers, particularly the infectious disease centers, and the larger scientific community in this field.

Cloning, expression, purification and crystallization

Taken together, the four steps needed to produce diffraction-quality crystals of macromolecules present a major obstacle in the structure determination staircase. As a consequence, well-diffracting single crystals of macromolecules are much more expensive than any diamond on Earth, either by weight or by volume. The four steps are linked to one another and should be treated as a single process leading to high-quality crystals (Fig. (3)). For example, even if protein production is successful, there may be a need to modify it and repeat it multiple times in search of more favorable crystallization conditions.

The cloning step is relatively straightforward, as there is an abundance of commercially available kits and services, including *de novo* gene synthesis. Moreover, cloning can be relatively easily automated, or at least semi-automated, and it is not surprising that most SG centers are able to generate thousands of clones per year. The clones created by PSI centers are available through a material depository system [30] which delivers clones by request to any laboratory worldwide. Similarly, CSGID and SSGCID will deposit their clones into the Biodefenses and Emerging Infections Research Resource Repository [31]. Although each center has its own protocols for cloning and expression, the protocols often share many common features, such as the usage of ligation-independent cloning (LIC) methods, and T7 RNA polymerase-based expression by the use of derivatives of *E. coli* BL21(DE3) [32].

SG centers have developed many new cloning vectors (Fig. (4)) which have been tested over time and optimized, taking into account a large number of successful [33-35] and arguably more important *unsuccessful* experiments. Though there are many kinds of affinity tags, SG centers most often initially use polyhistidine tags, linked to the main protein sequence *via* a tobacco etch virus (TEV) protease cleavage site [36]. Fusion of the affinity tag to the protein of interest greatly facilitates purification of that protein and rarely affects biological activity. Purification in the case of polyhistidine-tagged proteins is done by immobilized metal affinity chromatography (IMAC), and the tag is typically cleaved prior to crystallization [32-37].

If the target protein is soluble, purification is straightforward, and the purified protein may be used immediately in crystallization experiments. This is not the case when the purified protein is insoluble. Poor protein solubility greatly reduces the rate of success of the steps between clone and crystal (Fig. (3)), and much of the experimental attention of SG centers

is directed toward solving this problem [38-39]. A protein's solubility may be improved by expression with a fusion partner, such as maltose-binding protein (MBP) [34-40], or by denaturing and refolding the protein [41-42]. Other approaches, like a change of expression vectors, which provide alternate affinity tags, different variants of protein sequence, etc., or expression conditions are usually performed routinely. Less frequently, SG centers will change expression systems, co-express with interacting proteins [43-44], or supplement the expression with biologically relevant ligands [45]. However, the latter two methods are limited by the lack of knowledge about the exact function of a significant number of SG targets. Finally, in ortholog screening [46], a set of orthologs of the protein of interest are purified and crystallized in parallel, increasing the probability that at least one will be soluble and readily crystallize.

Low molecular weight compounds, such as ligands and so-called additives) greatly influence crystallization and their addition may be critical for success of this process [47-48]. Isothermal denaturation may be used to screen relatively large number of compounds and detect their ability to bind to proteins [49]. Other physico-chemical methods (such as SDS and native polyacrylamide gel electrophoresis (PAGE), dynamic light scattering (DLS), and nuclear magnetic resonance (NMR)) for characterizing protein solutions can identify possible problems such as lack of homogeneity, polydispersity, or aggregation of the sample, as well as provide additional information which is helpful for crystallization [50]. For example, NMR may be used for screening of protein samples prior to their crystallization. Even one-dimensional ^1H NMR spectroscopy, although limited by protein size, helps to identify the proteins that are likely to crystallize and those that are not, due to their degree of disorder [51].

SG centers have collected a great deal of information about failed experiments, and because of this, are able to correlate intrinsic protein properties with the outcome of particular steps leading to structure determination. Given the huge number of crystallization experiments prepared by SG centers, some of the properties of proteins or protein constructs that correlate with crystallization success have been identified [50]. The recently proposed method for increasing the chance of crystallization *via* protein oligomerization or symmetrization [52] may be validated by the results of SG projects. Every additional piece of information that identifies possible problems with the protein sample prior to crystallization may lead to significant cost reductions of the whole structure determination process. As was estimated by the Joint Center for Structural Genomics (JCSG), more than 60% of the cost of structure determination process may be attributed to failed experiments [53]. Discovering ways to overcome or prevent these failures is a large focus of the SG centers and is great benefit to the entire structural biology community.

Initial crystallization conditions usually come from either commercial or locally-developed screens (which are often later commercialized). The most successful labs use only a limited number of screens (96 or 192)—if a protein does not crystallize after 200-300 crystallization conditions are screened, it is not likely to crystallize at all [54-55]. For such cases, pursuit of alternate strategies, such as modification of the protein itself, has proven to be more successful. Two common and simple approaches for protein modification are reductive methylation [56-57] and limited proteolysis [58]. Their application does not

require additional cloning and may be performed using previously purified protein. Surface entropy reduction (SER) may also be successfully used in a salvage pathway [59], although SER requires more effort by comparison, namely site-directed mutagenesis of one or more surface Lys or Glu residues. In a number of cases SER has yielded crystals when the unmodified protein failed to crystallize.

In general, the experience of SG centers has shown that the production of soluble protein and successful crystallization appear to be the most difficult steps in the process of structure determination *via* X-ray diffraction.

Data Collection and Structure Solution

The fact that more than three fourths of SG targets are not homologous to previously solved protein structures (as determined by sequence identity lower than 30%) has pushed the development of technology and protocols that use Single-wavelength Anomalous Diffraction (SAD) and Multiple-wavelength Anomalous Diffraction (MAD) techniques. Indeed, today both techniques almost completely overshadow Multiple Isomorphous Replacement (MIR) experiments (Fig. (5)). Two other factors also hastened that transition, namely modern molecular biology techniques for easy production of selenomethionine-substituted protein, and the substantial growth, particularly in the past five years, in the number of synchrotron stations for macromolecular diffraction experiments.

Currently there are over 120 synchrotron stations in the world, most of them dedicated, that are suitable for X-ray macromolecular diffraction experiments. In recent years, almost 80% of tSB deposits in PDB report the use of a synchrotron source for diffraction experiments [60]. This percentage is even higher, about 90%, for SG deposits (Fig. (6)). Many of these stations have automatic crystal mounting robots and some of them have systems that allow for automatic or remote data collection, or both. Assuming that a complete data set can be collected within one hour—on 3rd generation synchrotrons like the Advanced Photon Source, it may take less than 10 minutes—and that an average synchrotron station works 4000 hours per year, the current synchrotron stations are capable of producing an astronomical 500,000 data sets per year. However, this number is almost two orders of magnitude higher than the real number of deposits to the PDB per year. A ratio of 100 data sets per 1 successful PDB deposit seems very high but it roughly agrees with an analysis of the output of an Advanced Light Source beamline, which on average reported that 57 collected data sets were required to produce one deposit [61]. Moreover, an analysis of PDB deposits shows that in many cases a single group uses many synchrotron stations to solve a deposited structure. Fig (7) presents the top synchrotron beamlines in terms of number of unique deposits in the PDB in recent years.

Further analysis shows that two SG centers, MCSG and CSGID, collect far fewer data sets per deposit, or in other words use fewer crystals per deposit, than most other SG centers. Neither center uses highly automatic or remote data collection systems but instead uses the new HKL-3000 system [62]. HKL-3000 integrates data collection, reduction and structure solution into one process and thus provides direct feedback from the structure solution process back to the data collection step. In this case, the result of a synchrotron experiment

is not reduced diffraction data or even an electron density map, but a partial model or sometimes a complete and partially refined model, obtained when the crystal is still in the X-ray beam. There are numerous advantages of that approach. First, it confirms that the structure is practically solved and functional studies may begin even before the structure is completely refined. Second, in cases where the map contains a metal atom which is difficult to identify, it is often possible to perform an anomalous diffraction experiment in order to accurately identify the metal atom while the crystal is still mounted on the goniometer head. In fact, applying this strategy is much more important for a tSB laboratory than for a SG center, since a crystallographer who has already solved over hundred structures can more readily identify why a structure may be ‘stubborn’. This highlights the need to encapsulate the experience gained by structural genomics in both training of crystallographers and commercialization of SG technologies.

There are two main sources of problems in data collection: the crystal and the experimental setup. The variety of crystals used in macromolecular crystallography is enormous. For example, unit cell dimensions may vary between 10 and 2000 Å, crystal mosaicity may vary between 0.05 and 3.5°, and every crystal may have additional problems like anisotropic mosaicity, twinning, etc. The correlations between crystal and experimental setup pathologies, sometimes called “features,” are very unintuitive. For example, goniostats using servomotors for driving the spindle axis produce poor quality data when almost perfect crystals with low mosaicity are measured. The same goniostat rotating a more typical macromolecular crystal, with a mosaicity of 1°, may yield perfect data, as the effectively larger measurement time averages the nonuniformity of the spindle axis movement. Structural experiments influenced by this issue may still be solved by molecular replacement, but effectively make use of SAD or MAD techniques impossible. Experimenters that are aware of this correlation may use one of two equivalent approaches: either use another synchrotron station where the spindle axis is driven by a stepper motor, or select a more mosaic crystal for SAD/MAD experiments.

The example described above illustrate that a good experimental protocol for data collection is critical for the final success of the experiment. Customized data collection protocols are a part of the strategy algorithms embedded into several diffraction data collection and processing programs. Most of these strategy algorithms minimize the amount of time required by experiment, by presenting a strategy, which unfortunately is sometimes very complicated, for collecting the minimum number of crystal oscillation frames to yield nearly 100% complete data. This was a very important calculation when synchrotron time was in very short supply and every minute of synchrotron time had to be used efficiently. It is somewhat surprising that despite good strategy algorithms, the overall data completeness and completeness in the last shell for PDB deposits are relatively low (Fig. (8)). Apparently the estimation of overlapping reflection profiles is not routinely performed. Incomplete high-resolution data result in an effective resolution lower than the nominal one. Another surprising result is the distribution of $I/\sigma I$ for the highest resolution shell (Fig. (9)). Clearly despite the tremendous investment in large area detectors (the price of the largest CCD detector on the market today exceeds a million dollars), a significant number of structural data are still collected to a resolution worse than the possible diffraction limit. Subsequently,

many structures are refined to a limited resolution, which may very strongly affect both the identification and placement of ligands, and the overall accuracy of structures. This is particularly troublesome for drug design projects, because the resolution of diffraction data and accuracy of a structure of a protein-ligand complex is of critical importance to drug discovery studies.

The overall structure quality of X-ray structures coming from SG is a little better than the PDB average [5,63,64]. This is because in addition to the validation tools used by the PDB and other popular validation tools (Coot [65], Molprobity [66-67]), excellent new tools developed as part of the JCSG effort have now been made available as online server services (<http://smb.slac.stanford.edu/jcsg/QC>). Unfortunately, there are relatively few validation servers for non-proteinaceous moieties and their interactions with the macromolecule. These servers include PROSIT (<http://cactus.nci.nih.gov/prosit/>), which analyzes nucleotide ligands, VaLigURL (<http://eds.bmc.uu.se/eds/valligurl.php>), which helps validate non-proteinaceous ligand conformations [68], and LPCCCU (<http://bip.weizmann.ac.il/oca-bin/lpcccu/>), which quantifies overall ligand fit [69]. Tools for validation of carbohydrates are currently under development and are not yet fully implemented [70]. Although the primary goal of SG so far is the elucidation of new structures, many contain metals and/or various agents that were introduced during crystallization and purification. An analysis of metal environments in protein structures [71] showed some abnormally high or low values of bond lengths and B-factors in metal-binding sites, especially for structures utilizing data with resolutions between 2.0 – 2.5 Å.

Structure Analysis

The initial goals of the PSI did not include advanced analysis or further functional studies of target proteins. It was assumed that this task would be taken over by researchers outside of the SG centers. However, production of structures by SG programs has been growing much faster than scientists' ability to analyze them and absorb them into functional studies. Faced with the deluge of structures, SG centers have developed semi-automatic methods for functional annotation and structure analysis such as the ProFunc server [72] at MCSG, or the Protein Sequence Comparative Analysis (PSCA; <http://www1.jcsg.org/psat/>) affiliated with JCSG. In practice, functional annotation of the deposited structures at the PSI centers is performed by human curators making use of automatic annotation servers [73-74].

Up to the date of this writing (February 2009), PSI programs have produced more than 1300 publications in scientific journals, 605 of these structural and 696 methodological. So far, PSI structural publications have averaged 16 citations per paper, whereas methodological publications so far have had an average of 32 citations. These numbers are much higher than the average citation numbers for biochemical papers and suggest that the methodology developed by SG is being adopted by traditional biomedical laboratories.

In recent years a new interesting paradigm has started to emerge in which structure analysis is done by the community as a whole, *a la* Wikipedia. For example, Proteopedia [75-76] strives to “collect, organize and disseminate structural and functional knowledge about proteins” and permits any interested person to contribute his or her knowledge. The

Open Structure Annotation Network (TOPSAN), developed at JCSG, is a wiki designed to “collect, share and distribute information about protein three-dimensional structures, and to advance it towards knowledge about functions”. This wiki provides a forum in which any interested scientist may contribute his or her knowledge about a particular structure solved by PSI, hopefully leading to the elucidation of its functional role, start of a collaboration on a follow-up study, and a joint publication. SG centers are also starting to find new ways to disseminate three-dimensional information about proteins to the wider scientific public outside the structural biology community. For example, SGC has developed an “interactive structurally enhanced experience” (iSee) offering users a predefined ‘guided tours’ of specific structures, linking a 3-D visualization with narrative written by experts, containing biological and functional interpretation [77]. If nothing else, the wikis and other Internet resources may expose structural knowledge and related challenges to a new audience, outside the confines of academic structural biology: researchers and students from other fields, medical doctors and even amateur scientists. (In some fields, such as astronomy, the contributions of amateurs have been non-negligible). The combination of different backgrounds and points of view along with the exchange of ideas among this audience may yield fresh approaches to concrete, medically relevant structural problems. Ultimately, if this shift to collaborative, community-wide structure analysis succeeds, it may turn out to be one of the most important achievements of structural genomics.

Ligands

Perhaps the most important aspect of structural biology for drug design is the determination of protein-ligand complexes. The interpretation of structures in terms of biochemical and biomedical properties usually requires detailed study of these complexes. Analysis of the PDB shows that about 70% of structures contain small molecule agents. Some ligands are not identified and marked as unknown. It may seem counterintuitive that the fraction of structures with unknown ligands is higher for higher resolution structures (Fig. (10)) but as unambiguous identification and refinement of ligand is more reliable for higher resolution data, it is less likely for ligands to be placed in density where they may not belong. Lower resolution structures probably allow for more “wishful thinking” in ligand placement, resulting in ligand structures that might be incorrect. Even for metal ions, in many cases identification and refinement of the binding sites is clearly incorrect [71], when compared to very high resolution structures in the Cambridge Structural Database [78].

In most cases, the structure of a protein’s *apo*-form is not sufficient to understand the mechanism of action of that protein. Several additional structures of protein (or protein mutants) complexed with small molecule ligands may be required to elucidate the biochemistry of the process in which the molecule is involved. Often the whole process of cloning, expression, purification and crystallization has to be repeated without any guarantee that new structural information will be acquired. Preparation of protein and ligand complexes involves co-crystallization or soaking of crystals or both, and most often is not trivial. It may be necessary not only to find new crystallization conditions, but new crystal forms more suitable for ligand binding studies—for example, crystal forms where ligand-binding sites are not blocked by crystal-symmetry-related molecules.

One example of a putative drug target in complex both with a natural ligand and a putative lead compound is cystathionine- γ -lyase (CSE). CSE is one of the two enzymes mainly responsible for production of the gaseous signaling molecule H₂S, and inhibition of the enzyme has been shown in animal models to be therapeutic for treatment of disorders of sulfur metabolism such as cystathioninuria. The SGC determined the structure of human CSE in an *apo*-form (PDB id 2NMP), and in complex with the natural substrate pyrioxidal-5'-phosphate and the inhibitor DL-propargylglycine (PDB id 3COG) [17].

Although some of the ligands in protein crystal structures are artifacts of the purification and crystallization processes, they may still provide useful information for functional studies. One example is that of TM1030, a TetR homolog from *Thermatoga maritima*, which contains a PEG molecule mimicking the potential effector molecule for the regulator [79]. Unfortunately, despite an avalanche of ligand structural data present in both the PDB and the private databases of pharmaceutical companies, the intelligent structure-based drug discovery process is still extremely difficult. The structural data provided by SG has already been a tremendous benefit to the development of protein fold modeling technologies and software—for example, the majority of CASP targets [80-81] have come from SG. It is hoped that the abundance of ligand-bound structures that have come from SG will soon aid the development of *in silico* ligand-docking tools in a similar manner.

Currently available *in silico* ligand screening tools are still not mature. The reliability of these programs to separate useful solutions from background noise could be compared to the reliability of the results produced by the risk assessment programs recently used on Wall Street [82]. The many structures of ligand-protein complexes provided in part by SG should help to develop better programs for virtual screening. Better *in silico* analysis could convert high-throughput ligand screening studies into high-*output* drug discovery. Better use of existing information may greatly facilitate the process of finding treatment for some diseases by revisiting compounds which are already clinically approved. Sometimes compounds developed to treat one disease may be efficient in treatment of other diseases, and in such a situation, providing the therapeutic solution is much faster than the development of a new drug. For example, nitrogen-containing bisphosphonates (Fig. (11)), which are clinically approved for treatment of bone metabolism disorders, were found to be good candidates for treatment of cryptosporidiosis [75]. In that case, structural analysis of protein – small molecule interactions was the result of a fruitful collaboration between two SG centers (MCSG and SGC). In general, the main difficulty of SG (and modern science as a whole) lies in the conversion of many data into a useful synthesis of information.

From Data to Information

The amount of experimental data generated in SG projects is enormous. In January 2009 the number of SG targets surpassed 200,000, of which almost 70% have been cloned. For the large-scale PSI centers the numbers of targets ranged from 15,000 to almost 50,000. As every week there are around 140 new structures deposited into the PDB, the list of targets has to be continually reevaluated, which is alone a significant (albeit simple) bioinformatics task. Typically, as each target progresses through the experimental pipeline there are several hundred data items that must be recorded, and these data must be consistent with one

another for effective data analysis. Efficient data management techniques are required to keep track of this data, minimize duplication of effort and maximize the chances of success at each step. As various steps in the experimental pipeline for an SG center are often performed in different locations, the system of tools and databases required for this task acts like a “nervous system”. Blocking unpleasant information by reporting only successful experiments is equivalent to numbing some pain receptors – pleasant, but dangerous for long-term health.

At the start of PSI, there were attempts to standardize the database architecture and create all-encompassing data dictionaries for SG labs [83-84], to meet the requirement to provide a weekly report of the progress of each target to the public. In practice, each PSI center adopted its own solution, customized to account for the specific experimental circumstances of each center. Nevertheless, it appears that SG projects have significantly contributed to the growing need for a very reliable Laboratory Information Management Systems (LIMS) not only in the context of a SG center, but also within traditional structural biology labs [85]. Most labs (even some participating in SG centers) have been using lab notebooks or, slightly more ambitiously, Excel spreadsheets as a substitute for a LIMS. This has worked fine when the number of projects in a lab was small, and the post-doc or student was able to write the paper before they move to another laboratory, but this approach becomes impossible when the number of parallel projects in a the lab exceeds a certain threshold.

Researchers used to working with laboratory notebooks only sometimes complain that the requirements of updating databases slow down the actual experimental work. That may be true, especially when data harvesting interfaces to the databases are not optimally designed; however in practice the costs of such slowdowns are usually minuscule in comparison with the time and effort involved in rectifying the errors which are almost impossible to avoid without an efficient LIMS system. Currently, data harvesting interfaces to databases are evolving towards better integration with laboratory equipment and minimization of intervention by human beings. A number of LIMS developed in SG centers are aspiring to achieve wider distribution outside the SG environment. Among those are Sesame [86] developed by Center for Eukaryotic Structural Genomics, HalX [87], developed by the European consortium SPINE, and Xtaldb [88], developed at the Midwest Center for Structural Genomics.

The PSI and infectious disease SG programs, under the requirements of the NIH, and most other SG centers, under the International Structural Genomics Organization (ISGO) agreement, have adopted a policy of sharing data about projects that they are working on. Initially, only basic information about the progress of SG projects was collected by TargetDB [89]. Despite the very limited number of data per project collected, TargetDB has provided a very valuable resource for analysis of properties of proteins, which has been used in numerous research studies. PepcDB is an enhanced version of TargetDB that contains additional data about protocols and experimental details. Both databases have now been integrated in the PSI Knowledgebase [2], providing a collection of SG resources concerned with target selection and tracking, models, annotations, and publications.

Conclusions

In terms of the drugs currently available on the market, the structure-based approach in drug discovery has so far been a rather modest contributor. On the other hand, it is generally agreed that the knowledge of three-dimensional protein structures often offers functional insights indirectly benefiting drug discovery [12]. Structural Genomics programs have significantly increased structural coverage of many important infectious pathogens. Technologies developed by Structural Genomics programs, such as new cloning vectors and crystal screens, methylation, limited proteolysis *etc.*, provide effective and cost-efficient tools for high-throughput structure determination by X-ray crystallography. There remain experimental challenges in obtaining soluble, expressing protein and in getting diffraction-quality crystals. There also remains a less obvious, but no less important, bottleneck related to processing of experimental data and transforming it into information. This last step requires 'brain engagement' and is impossible to automate. It is our opinion that the impact of the technologies developed by SG will enhance drug discovery in the near future. The technological expertise accumulated in SG, together with the growing structural coverage of bio-medically relevant proteins, and community-wide structure analysis are conditions in which the present high-throughput structure determination pipeline might be transformed into a future high-output drug research pipeline, one that might combine both public funding and corporate support.

Acknowledgments

The authors would like to thank Alex Wlodawer, Andrzej Joachimiak, Wayne Anderson, Tom Terwilliger, Al Edwards and Zbyszek Dauter for valuable discussions. The work described in the paper was supported by GM74942, GM53163 and with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200700058C.

References

1. NIH Press Release. <http://www.nigms.nih.gov/Initiatives/PSI/>
2. Berman HM, Westbrook JD, Gabanyi MJ, Tao W, Shah R, Kouranov A, Schwede T, Arnold K, Kiefer F, Bordoli L, Kopp J, Podvinec M, Adams PD, Carter LG, Minor W, Nair R, La Baer J. Nucleic Acids Res. 2009; 37: D365. [PubMed: 19010965]
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. Nucleic Acids Res. 2000; 28: 235. [PubMed: 10592235]
4. Berman H, Henrick K, Nakamura H, Markley JL. Nucleic Acids Res. 2007; 35: D301. [PubMed: 17142228]
5. O'Toole N, Grabowski M, Otwinowski Z, Minor W, Cygler M. Proteins. 2004; 56: 201. [PubMed: 15211505]
6. Chandonia JM, Brenner SE. Proteins. 2005; 58: 166. [PubMed: 15521074]
7. Grabowski M, Joachimiak A, Otwinowski Z, Minor W. Curr Opin Struct Biol. 2007; 17: 347. [PubMed: 17587562]
8. Liu J, Montelione GT, Rost B. Nat Biotechnol. 2007; 25: 849. [PubMed: 17687356]
9. Petsko GA. Genome Biol. 2007; 8: 107. [PubMed: 17608958]
10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matrese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Nat Genet. 2000; 25: 25. [PubMed: 10802651]

11. Finn RD, Tate J, Mistry J, Coghill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. *Nucleic Acids Res.* 2008; 36: D281. [PubMed: 18039703]
12. Congreve M, Murray CW, Blundell TL. *Drug Discov Today.* 2005; 10: 895. [PubMed: 15993809]
13. Scapin G. *Curr Pharm Des.* 2006; 12: 2087. [PubMed: 16796557]
14. Sheahan KL, Cordero CL, Satchell KJ. *EMBO J.* 2007; 26: 2552. [PubMed: 17464284]
15. Chiba T, Li YH, Yamane T, Ogikubo O, Fukuoka M, Arai R, Takahashi S, Ohtsuka T, Ohkubo I, Matsui N. *Peptides.* 2003; 24: 773. [PubMed: 12895665]
16. Barsun M, Jajcanin N, Vukelic B, Spoljaric J, Abramic M. *Biol Chem.* 2007; 388: 343. [PubMed: 17338643]
17. Liu Y, Kern JT, Walker JR, Johnson JA, Schultz PG, Luesch H. *Proc.*
18. Maresso AW, Wu R, Kern JW, Zhang R, Janik D, Missiakas DM. *Natl Acad Sci U S A.* 2007; 104: 5205. Duban ME, Joachimiak A, Schneewind O. *J Biol Chem.* 2007; 282: 23129. [PubMed: 17545669]
19. Zhang R, Wu R, Joachimiak G, Mazmanian SK, Missiakas DM, Gornicki P, Schneewind O, Joachimiak A. *Structure.* 2004; 12: 1147. [PubMed: 15242591]
20. Nocek B, Mulligan R, Bargassa M, Collart F, Joachimiak A. *Proteins.* 2008; 70: 273. [PubMed: 17876832]
21. Terwilliger TC, Park MS, Waldo GS, Berendzen J, Hung LW, Kim CY, Smith CV, Sacchettini JC, Bellinzoni M, Bossi R, De Rossi E, Mattevi A, Milano A, Riccardi G, Rizzi M, Roberts MM, Coker AR, Fossati G, Mascagni P, Coates AR, Wood SP, Goulding CW, Apostol MI, Anderson DH, Gill HS, Eisenberg DS, Taneja B, Mande S, Pohl E, Lamzin V, Tucker P, Wilmanns M, Colovos C, Meyer-Klaucke W, Munro AW, McLean KJ, Marshall KR, Leys D, Yang JK, Yoon HJ, Lee BI, Lee MG, Kwak JE, Han BW, Lee JY, Baek SH, Suh SW, Komen MM, Arcus VL, Baker EN, Lott JS, Jacobs W Jr, Alber T, Rupp B. *Tuberculosis (Edinb).* 2003; 83: 223. [PubMed: 12906835]
22. Arcus VL, Lott JS, Johnston JM, Baker EN. *Drug Discov Today.* 2006; 11: 28. [PubMed: 16478688]
23. Murillo AC, Li HY, Alber T, Baker EN, Berger JM, Cherney LT, Cherney MM, Cho YS, Eisenberg D, Garen CR, Goulding CW, Hung LW, Ioerger TR, Jacobs WR, James MN, Kim C, Krieger I, Lott JS, Sankaranarayanan R, Segelke BW, Terwilliger TC, Wang F, Wang S, Sacchettini JC. *Infect Disord Drug Targets.* 2007; 7: 127. [PubMed: 17970224]
24. Baker EN. *J Struct Funct Genomics.* 2007; 8: 57. [PubMed: 17668294]
25. Fan E, Baker D, Fields S, Gelb MH, Buckner FS, Van Voorhis WC, Phizicky E, Dumont M, Mehlin C, Grayhack E, Sullivan M, Verlinde C, Detitta G, Meldrum DR, Merritt EA, Earnest T, Soltis M, Zucker F, Myler PJ, Schoenfeld L, Kim D, Worthey L, Lacount D, Vignali M, Li J, Mondal S, Massey A, Carroll B, Gulde S, Luft J, Desoto L, Holl M, Caruthers J, Bosch J, Robien M, Arakaki T, Holmes M, Le Trong I, Hol WG. *Methods Mol Biol.* 2008; 426: 497. [PubMed: 18542886]
26. Gileadi O, Knapp S, Lee WH, Marsden BD, Muller S, Niesen FH, Kavanagh KL, Ball LJ, von Delft F, Doyle DA, Oppermann UC, Sundstrom M. *J Struct Funct Genomics.* 2007; 8: 107. [PubMed: 17932789]
27. Edwards A. *Drug Discov Today.* 2008; 13: 731. [PubMed: 18790412]
28. Dufe VT, Qiu W, Muller IB, Hui R, Walter RD, Al-Karadaghi S. *J Mol Biol.* 2007; 373: 167. [PubMed: 17822713]
29. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. *Nucleic Acids Res.* 2006; 34: D668. [PubMed: 16381955]
30. PSI Material Repository. <http://www.hip.harvard.edu/PSIMR/>
31. BEI Resources. <http://www.beiresources.org>
32. Graslund S, Nordlund P, Weigelt J, Hallberg BM, Bray J, Gileadi O, Knapp S, Oppermann U, Arrowsmith C, Hui R, Ming J, dhe-Paganon S, Park HW, Savchenko A, Yee A, Edwards A, Vincentelli R, Cambillau C, Kim R, Kim SH, Rao Z, Shi Y, Terwilliger TC, Kim CY, Hung LW, Waldo GS, Peleg Y, Albeck S, Unger T, Dym O, Prilusky J, Sussman JL, Stevens RC, Lesley SA, Wilson IA, Joachimiak A, Collart F, Dementieva I, Donnelly MI, Eschenfeldt WH, Kim Y, Stols L, Wu R, Zhou M, Burley SK, Emtage JS, Sauder JM, Thompson D, Bain K, Luz J, Gheyi T, Zhang

- F, Atwell S, Almo SC, Bonanno JB, Fiser A, Swaminathan S, Studier FW, Chance MR, Sali A, Acton TB, Xiao R, Zhao L, Ma LC, Hunt JF, Tong L, Cunningham K, Inouye M, Anderson S, Janjua H, Shastry R, Ho CK, Wang D, Wang H, Jiang M, Montelione GT, Stuart DI, Owens RJ, Daenke S, Schutz A, Heinemann U, Yokoyama S, Bussow K, Gunsalus KC. *Nat Methods*. 2008; 5: 135. [PubMed: 18235434]
33. Stols L, Gu M, Dieckman L, Raffin R, Collart FR, Donnelly MI. *Protein Expr Purif*. 2002; 25: 8. [PubMed: 12071693]
34. Donnelly MI, Zhou M, Millard CS, Clancy S, Stols L, Eschenfeldt WH, Collart FR, Joachimiak A. *Protein Expr Purif*. 2006; 47: 446. [PubMed: 16497515]
35. Klock HE, Koesema EJ, Knuth MW, Lesley SA. *Proteins*. 2008; 71: 982. [PubMed: 18004753]
36. Kapust RB, Tozser J, Fox JD, Anderson DE, Cherry S, Copeland TD, Waugh DS. *Protein Eng*. 2001; 14: 993. [PubMed: 11809930]
37. Kim Y, Dementieva I, Zhou M, Wu R, Lezondra L, Quartey P, Joachimiak G, Korolev O, Li H, Joachimiak A. *J Struct Funct Genomics*. 2004; 5: 111. [PubMed: 15263850]
38. Pedelacq JD, Piltch E, Liong EC, Berendzen J, Kim CY, Rho BS, Park MS, Terwilliger TC, Waldo GS. *Nat Biotechnol*. 2002; 20: 927. [PubMed: 12205510]
39. Bursey EH, Kim CY, Yu M, Terwilliger TC, Hung LW. *J Struct Funct Genomics*. 2006; 7: 139. [PubMed: 17541730]
40. Esposito D, Chatterjee DK. *Curr Opin Biotechnol*. 2006; 17: 353. [PubMed: 16781139]
41. Maxwell KL, Bona D, Liu C, Arrowsmith CH, Edwards AM. *Protein Sci*. 2003; 12: 2073. [PubMed: 12931005]
42. Cowieson NP, Wensley B, Listwan P, Hume DA, Kobe B, Martin JL. *Proteomics*. 2006; 6: 1750. [PubMed: 16475229]
43. Stols L, Zhou M, Eschenfeldt WH, Millard CS, Abdullah J, Collart FR, Kim Y, Donnelly MI. *Protein Expr Purif*. 2007; 53: 396. [PubMed: 17363272]
44. Strong M, Sawaya MR, Wang S, Phillips M, Cascio D, Eisenberg D. *Proc Natl Acad Sci U S A*. 2006; 103: 8060. [PubMed: 16690741]
45. Hozjan V, Guo K, Wu X, Oppermann U. *Expert Rev Proteomics*. 2008; 5: 137. [PubMed: 18282129]
46. Page R. *Methods Mol Biol*. 2008; 426: 345. [PubMed: 18542875]
47. McPherson A, Cudney B. *J Struct Biol*. 2006; 156: 387. [PubMed: 17101277]
48. Chruszcz M, Zimmerman MD, Wang SR, Koclega KD, Zheng H, Evdokimova E, Kudritska M, Cymborowski M, Savchenko A, Edwards A, Minor W. *Crystal Growth & Design*. 2008; 8: 4054. [PubMed: 19898606]
49. Senisterra GA, Soo Hong B, Park HW, Vedadi M. *J Biomol Screen*. 2008; 13: 337. [PubMed: 18448703]
50. Price WN 2nd, Chen Y, Handelman SK, Neely H, Manor P, Karlin R, Nair R, Liu J, Baran M, Everett J, Tong SN, Forouhar F, Swaminathan SS, Acton T, Xiao R, Luft JR, Lauricella A, DeTitta GT, Rost B, Montelione GT, Hunt JF. *Nat Biotechnol*. 2009; 27: 51. [PubMed: 19079241]
51. Page R, Peti W, Wilson IA, Stevens RC, Wuthrich K. *Proc Natl Acad Sci U S A*. 2005; 102: 1901. [PubMed: 15677718]
52. Banatao DR, Cascio D, Crowley CS, Fleissner MR, Tienson HL, Yeates TO. *Proc Natl Acad Sci U S A*. 2006; 103: 16230. [PubMed: 17050682]
53. Slabinski L, Jaroszewski L, Rodrigues AP, Rychlewski L, Wilson IA, Lesley SA, Godzik A. *Protein Sci*. 2007; 16: 2472. [PubMed: 17962404]
54. Page R, Grzechnik SK, Canaves JM, Spraggon G, Kreusch A, Kuhn P, Stevens RC, Lesley SA. *Acta Crystallogr D Biol Crystallogr*. 2003; 59: 1028. [PubMed: 12777766]
55. Kimber MS, Vallee F, Houston S, Necakov A, Skarina T, Evdokimova E, Beasley S, Christendat D, Savchenko A, Arrowsmith CH, Vedadi M, Gerstein M, Edwards AM. *Proteins*. 2003; 51: 562. [PubMed: 12784215]
56. Walter TS, Meier C, Assenberg R, Au KF, Ren J, Verma A, Nettleship JE, Owens RJ, Stuart DI, Grimes JM. *Structure*. 2006; 14: 1617. [PubMed: 17098187]

57. Kim Y, Quartey P, Li H, Volkart L, Hatzos C, Chang C, Nocek B, Cuff M, Osipiuk J, Tan K, Fan Y, Bigelow L, Maltseva N, Wu R, Borovilos M, Duggan E, Zhou M, Binkowski TA, Zhang RG, Joachimiak A. *Nat Methods*. 2008; 5: 853. [PubMed: 18825126]
58. Dong A, Xu X, Edwards AM, Chang C, Chruszcz M, Cuff M, Cymborowski M, Di Leo R, Egorova O, Evdokimova E, Filippova E, Gu J, Guthrie J, Ignatchenko A, Joachimiak A, Klostermann N, Kim Y, Korniyenko Y, Minor W, Que Q, Savchenko A, Skarina T, Tan K, Yakunin A, Yee A, Yim V, Zhang R, Zheng H, Akutsu M, Arrowsmith C, Avvakumov GV, Bochkarev A, Dahlgren LG, Dhe-Paganon S, Dimov S, Dombrowski L, Finerty P Jr, Flodin S, Flores A, Graslund S, Hammerstrom M, Herman MD, Hong BS, Hui R, Johansson I, Liu Y, Nilsson M, Nedyalkova L, Nordlund P, Nyman T, Min J, Ouyang H, Park HW, Qi C, Rabeh W, Shen L, Shen Y, Sukumard D, Tempel W, Tong Y, Tresagues L, Vedadi M, Walker JR, Weigelt J, Welin M, Wu H, Xiao T, Zeng H, Zhu H. *Nat Methods*. 2007; 4: 1019. [PubMed: 17982461]
59. Derewenda ZS, Vekilov PG. *Acta Crystallogr D Biol Crystallogr*. 2006; 62: 116. [PubMed: 16369101]
60. Chruszcz M, Wlodawer A, Minor W. *Biophys J*. 2008; 95: 1. [PubMed: 18441029]
61. Holton, J. Teaching Elves to Collect Data: An Analysis of the Last Million Diffraction Images from ALS 8.3.1; Annual Meeting of the American Crystallographic Association; Orlando,FL, USA. 2005;
62. Minor W, Cymborowski M, Otwinowski Z, Chruszcz M. *Acta Crystallogr D Biol Crystallogr*. 2006; 62: 859. [PubMed: 16855301]
63. Todd AE, Marsden RL, Thornton JM, Orengo CA. *J Mol Biol*. 2005; 348: 1235. [PubMed: 15854658]
64. Brown EN, Ramaswamy S. *Acta Crystallogr D Biol Crystallogr*. 2007; 63: 941. [PubMed: 17704562]
65. Emsley P, Cowtan K. *Acta Crystallogr D Biol Crystallogr*. 2004; 60: 2126. [PubMed: 15572765]
66. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB 3rd, Snoeyink J, Richardson JS, Richardson DC. *Nucleic Acids Res*. 2007; 35: W375. [PubMed: 17452350]
67. Davis IW, Murray LW, Richardson JS, Richardson DC. *Nucleic Acids Res*. 2004; 32: W615. [PubMed: 15215462]
68. Kleywegt GJ, Harris MR. *Acta Crystallogr D Biol Crystallogr*. 2007; 63: 935. [PubMed: 17642521]
69. Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M. *Bioinformatics*. 1999; 15: 327. [PubMed: 10320401]
70. Lutteke T. *Acta Crystallogr D Biol Crystallogr*. 2009; 65: 156. [PubMed: 19171971]
71. Zheng H, Chruszcz M, Lasota P, Lebiada L, Minor W. *J Inorg Biochem*. 2008; 102: 1765. [PubMed: 18614239]
72. Laskowski RA, Chistyakov VV, Thornton JM. *Nucleic Acids Res*. 2005; 33: D266. [PubMed: 15608193]
73. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A. *Genome Biol*. 2009; 10: 207. [PubMed: 19226439]
74. Laskowski RA, Watson JD, Thornton JM. *Nucleic Acids Res*. 2005; 33: W89. [PubMed: 15980588]
75. Hodis E, Prilusky J, Martz E, Silman I, Moulton J, Sussman JL. *Genome Biol*. 2008; 9: R121. [PubMed: 18673581]
76. Martz E. *Biopolymers*. 2009; 92: 76. [PubMed: 19117028]
77. Abagyan R, Lee WH, Raush E, Budagyan L, Totrov M, Sundstrom M, Marsden BD. *Trends Biochem Sci*. 2006; 31: 76. [PubMed: 16406633]
78. Allen FH. *Acta Crystallogr B*. 2002; 58: 380. [PubMed: 12037359]
79. Koclega KD, Chruszcz M, Zimmerman MD, Cymborowski M, Evdokimova E, Minor W. *J Struct Biol*. 2007; 159: 424. [PubMed: 17588774]
80. Moulton J. *Curr Opin Struct Biol*. 2005; 15: 285. [PubMed: 15939584]

81. Moulton J, Fidelis K, Rost B, Hubbard T, Tramontano A. *Proteins*. 2005; 61 (Suppl 7) 3. [PubMed: 16187341]
82. Dennis B, O'Harrow JA. *Washington Post*. Dec 30. 2008. A01.
83. Pajon A, Ionides J, Diprose J, Fillon J, Fogh R, Ashton AW, Berman H, Boucher W, Cygler M, Deleury E, Esnouf R, Janin J, Kim R, Krimm I, Lawson CL, Oeuillet E, Poupon A, Raymond S, Stevens T, van Tilbeurgh H, Westbrook J, Wood P, Ulrich E, Vranken W, Xueli L, Laue E, Stuart DI, Henrick K. *Proteins*. 2005; 58: 278. [PubMed: 15562521]
84. Fogh RH, Boucher W, Vranken WF, Pajon A, Stevens TJ, Bhat TN, Westbrook J, Ionides JM, Laue ED. *Bioinformatics*. 2005; 21: 1678. [PubMed: 15613391]
85. Haquin S, Oeuillet E, Pajon A, Harris M, Jones AT, van Tilbeurgh H, Markley JL, Zolnai Z, Poupon A. *Methods Mol Biol*. 2008; 426: 49. [PubMed: 18542857]
86. Zolnai Z, Lee PT, Li J, Chapman MR, Newman CS, Phillips GN Jr, Rayment I, Ulrich EL, Volkman BF, Markley JL. *J Struct Funct Genomics*. 2003; 4: 11. [PubMed: 12943363]
87. Prilusky J, Oeuillet E, Ulryck N, Pajon A, Bernauer J, Krimm I, Quevillon-Cheruel S, Leulliot N, Graille M, Liger D, Tresaugues L, Sussman JL, Janin J, van Tilbeurgh H, Poupon A. *Acta Crystallogr D Biol Crystallogr*. 2005; 61: 671. [PubMed: 15930618]
88. Zimmerman MD, Chruszcz M, Koclega K, Otwinowski Z, Minor W. *Acta Cryst. A*. 2005; 61: C178.
89. Chen L, Oughtred R, Berman HM, Westbrook J. *Bioinformatics*. 2004; 20: 2860. [PubMed: 15130928]

Fig. (1).

Number of Gene Ontology terms for molecular functions **(a)** and biological processes **(b)** for structurally characterized Pfam families. “SG first” is the number of families for which SG determined the first structural representative, “tSB only” represents families for which structural representatives were determined only by traditional structural biology (outside SG centers), and “tSB+SG” denotes families for which both SG and tSB structural representatives exist, and the tSB structures were solved first. The data used were taken from Pfam-A release 21. For these and all other statistics, the list of “SG centers” were defined as those making depositions to TargetDB [89].

Fig. (2). Traditional structural biology deposits in PDB with significant sequence identity to previous SG deposits. Number of tSB structures which had at the moment of their deposition sequence identity greater than 30% or 90% to a previous SG deposit and no matches with identity greater than 30% or 90%, respectively, to a previous tSB deposit.

Fig. (3).
Schematic representation of the experimental pipeline of structure determination by X-ray crystallography.

Fig. (4).

Cloning vectors used by MCSG. The relative frequency of usage of all clones is shown as a pie chart (a), and the current (as of February 2009) distribution of the number of targets by experimental step of MCSG targets for which pMCSG7 and pMCSG19 vectors were used are shown as histograms (b).

Fig. (5).

The number of structures solved by anomalous diffraction (MAD/SAD) and isomorphous replacement (MIR/SIR) methods as a fraction of all PDB deposits with experimentally determined phases. (This includes all structures released as of March 2009.)

Fig. (6).

The fraction of X-ray PDB deposits reporting the use of synchrotron sources for data collection for the PDB as a whole, and for SG deposits only. (This includes all structures released as of March 2009.)

Fig. (7).

The synchrotron beamlines that were most frequently reported in PDB deposits of unique protein X-ray crystallography structures. (A unique deposit is defined as having no earlier PDB deposit with sequence identity greater than 30%). Numbers of deposits denotes number of all X-ray structures deposited to the PDB between January 2006 and March 2009 sorted by the beamline on which data were collected and subdivided into unique and non-unique SG and tSB contributions. The synchrotrons with beamlines shown on the chart are the Advanced Photo Source (APS; Argonne, Illinois: beamlines 19ID, 19BM, 31ID, 22ID, 23ID, 23IDD, 24IDC, 23IDB), the Stanford Synchrotron Radiation Lightsource (SSRL; Stanford, California: beamlines BL11-1, BL9-A), the National Synchrotron Light Source (NSLS; Brookhaven, New York: beamlines X29A, X4A, X12C), the Advanced Light Source (ALS; Berkeley, California: beamline 8.2.2), the Swiss Light Source (SLS; Switzerland: beamlines X10SA, X06SA), the European Synchrotron Radiation Facility (ESRF; Grenoble, France: beamlines ID2, ID29, ID14-4, ID14-1), Spring-8 (Nishi Harima, Japan: beamline BL26B2), and the Photon Factory (Tsukuba, Japan: beamline NW12).

Fig. (8).

The distribution of X-ray deposits in the PDB since 2001, grouped by overall completeness (a) or completeness in the highest-resolution shell (b), are shown as histograms. 46.8% of the SG deposits in this time frame had overall completeness above 99%, in comparison with 32.7% of tSB deposits. In the highest resolution shell, 35.7% of SG deposits had completeness above 99%, versus 27.5% of tSB structures.

Fig. (9).

The distribution of the average I/σ for the highest resolution shell for all X-ray structures in the PDB is plotted as a histogram.

Fig. (10).

The distributions of all X-ray PDB deposits and the deposits with ligands, both grouped by resolution, are plotted as a histogram (a). Overall about 70% of PDB deposits contain one or more ligands. The number of deposits with unidentified ligands, grouped by resolution and measured as a percentage of all X-ray deposits is plotted as a histogram (b). The data reflect deposits released as of July 2008.

Fig. (11). Binding of risedronate (shown in blue) by a nonspecific polyprenyl pyrophosphate synthase from *Cryptosporidium parvum* (PDB code 2O1O). Magnesium ions and water molecules, shown in sphere representation, are colored in green and red respectively. The side chains of residues forming the risedronate binding site are drawn in stick representation.

Table 1

Selected worldwide SG centers and resources

Acronym	Name	Description	URL
PSI	Protein Structure Initiative	NIH-funded SG program in the US. In its current (second) phase it has 4 large-scale centers and 6 specialized centers	http://www.nigms.nih.gov
JCSG	Joint Center for Structural Genomics	One of the four large-scale PSI centers	http://www.jcsg.org
MCSG	Midwest Center for Structural Genomics	One of the four large-scale PSI centers	http://mcsg.anl.gov
NESG	Northeast Center for Structural Genomics	One of the four large-scale PSI centers	http://www.nesg.org
NYSGXRC	New York SGX Research Center for Structural Genomics	One of the four large-scale PSI centers	http://www.nysgxrc.org
SGC	Structural Genomics Consortium	Non-profit international organization focused on proteins of medical relevance	http://www.sgc.utoronto.ca
CSGID	Center for Structural Genomics of Infectious Disease	NIAID-funded center focused on infectious diseases	http://csgid.org
SSGCID	Seattle Structural Genomics Center for Infectious Disease	NIAID-funded center focused on infectious diseases	http://ssgcid.org
CESG	Center for Eukaryotic Structural Genomics	PSI specialized center focused on eukaryotic proteins	http://www.uwstructuralgenomics.org/
SGPP	Structural Genomics of Pathogenic Protozoa	PSI-phase 1 center focused on pathogenic protozoa	http://www.sgpp.org
MSGPP	Medical Structural Genomics of Pathogenic Protozoa	Focuses on major global parasitic diseases	http://www.msgpp.org/
TBsgc	Tuberculosis Structural Genomics Consortium	International consortium of over 70 labs devoted to structural biology of <i>M. tuberculosis</i>	http://www.doe-mbi.ucla.edu/TB
SPINE	Structural Proteomics in Europe	European collaborative program aimed at proteins relevant to human health	http://spineurope.orgm
RSGI	RIKEN Structural Genomics/Proteomics Initiative	Large-scale Japanese program focused on eukaryotic and biomedical targets	http://protein.gsc.riken.jp
ISGO	International Structural Genomics Organization	Devoted to international cooperation of structural genomics centers	http://www.isgo.org
SGKB	PSI Nature Structural Genomics Knowledgebase	Web portal for PSI-related resources	http://kb.psi.structural-genomics.org
TargetDB	Structural Genomics target registration database	Repository of data on progress of targets from worldwide SG projects	http://targetdb.pdb.org

Acronym	Name	Description	URL
PepcDB	Protein expression purification and crystallization Data Base	Repository of data from SG projects, extending the content of TargetDB with experimental details	http://pepcdb.pdb.org

Table 2

Contributions of the PSI and the infectious disease (ID) centers (CSGID and SSGCID) to the structural coverage of proteins in species targeted by the ID centers. “Total” represents the number of all structures of proteins from a given species deposited to the PDB between January 2004 and February 2009. PSI, CSGID, and SSGCID represent the depositions from those centers in the same time frame

Species	Total	PSI	CSGID	SSGCID	PSI+ID centers
<i>Escherichia coli</i>	2116	130	0	0	6.1%
<i>Salmonella typhimurium</i>	218	53	17	0	32.1%
<i>Vibrio cholerae</i>	104	40	10	0	48.0%
<i>Vibrio parahaemolyticus</i>	37	36	0	0	97.3%
<i>Staphylococcus aureus</i>	337	35	10	0	13.3%
<i>Shigella flexneri</i>	53	27	2	0	54.7%
<i>Listeria monocytogenes</i>	50	18	0	0	36.0%
<i>Bacillus anthracis</i>	97	11	6	0	17.5%
<i>Campylobacter jejuni</i>	39	9	0	0	23.1%
<i>Clostridium difficile</i>	15	9	0	0	60.0%
<i>Burkholderia pseudomallei</i>	62	3	0	40	64.5%
<i>Bartonella henselae</i>	6	3	0	2	83.3%
<i>Brucella melitensis</i>	23	2	0	12	60.9%
<i>Yersinia pestis</i>	22	0	3	0	13.6%
<i>Dengue virus</i>	38	0	1	0	2.6%
<i>Rickettsia prowazekii</i>	3	0	0	3	100.0%