

Chapter 1

The Importance of Target Selection Strategies in Structural Biology

Enrique E. Abola and Raymond C. Stevens*

Introduction

The industrialization of biology — the large-scale acquisition of biological data — has been pioneered by the sequencing of entire genomes and is being applied to the characterization of other important biological molecules such as the proteome, the interactome (arguably a subset of the proteome), the glycome, and the metabolome. As of October 2007, the complete genomic sequences of 676 organisms have been published and more projects are underway. The DNA sequence data alone is insufficient to generate the level of understanding of biological systems that most biologists seek. Understanding how biological systems operate from the level of single proteins and enzymes, to the level of protein-protein interactions, and finally at the level of intact cellular physiological pathways, a goal of systems biology, will require detailed, quantitative characterization of cellular proteins and their interactions, which is facilitated by access to protein structural information. Thus, the number and types of questions that can now be addressed by structural biologists has increased dramatically. The scope of protein structure space is still too immense for a

*Corresponding author: stevens@scripps.edu. Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA.

completely unfocused approach to data acquisition. Therefore, target selection is still a critical step in establishing an industrial-scale protein structure project.

About 50 years after the publication of the first 3-dimensional structure of a protein, that of sperm whale myoglobin completed in John Kendrew's laboratory¹ (a structural proteomics (SP) project in itself as myoglobin from multiple species were pursued), structural biologists have started to explore the possibility of conducting high-throughput (HT) structural studies to permit the structural characterization of proteomes. HT approaches, such as parallel studies of multiple protein targets, are expected to revolutionize the way structures are determined by moving away from one-by-one structural studies. These new approaches are expected to produce important scientific results at reduced costs by using economy of scales and by generating standardized and more generalizable protocols and evaluation metrics for protein expression, purification and structural characterization. Over the past 10 years, the results of successful pilot studies have been reported by the various SP programs. This leads to the important question of what target selection strategies should then be used in the future by both SP and non-SP laboratories in the light of what has been learned from these pilot studies.

This chapter summarizes the strategies in target selection and prioritization used by various SP groups and provides a brief summary of their recent results. We explore the potential of small and medium sized laboratories as well as larger collaborative efforts to make use of the new technologies, protocols and approaches developed by these initial SP initiatives, with a special emphasis on studying biological systems through class-directed target selection approaches.

Global Structural Efforts and their Target Selection Strategies

By 2000, several groups, both academic and for-profit companies, were being setup to establish HT structure determination production lines. Initial efforts were focused on developing new technologies and protocols for each step in the process, from initial cloning to final

deposition of coordinate data sets to public databases. The immediate aim was to convert the one-by-one structure determination process, using both single-crystal X-ray diffraction and solution NMR techniques, to work in a HT mode. Although the initial mandate was technology development, there was also a requirement to solve a relatively large number of protein structures to serve both as proof of concept and as a justification of the approach adopted by each center. Major government-sponsored consortium efforts were formed: the Protein Structure Initiative (PSI-1) in the USA, the Structural Proteomics (SPINE) integrated project in Europe, and Project 3000 in Japan, and smaller efforts in other countries. Another major effort was a joint venture between government and industry. The Structural Genomics Consortium, an international project funded by Canada, Sweden, the Wellcome Trust in the UK and industry, with laboratories in Oxford, Stockholm and Toronto. For-profit companies, such as SYRRX, SGX, and ASTEX, were setup with the goal of improving the drug discovery and development process by reducing the risks and cost of getting at the structures of drug targets and their complexes.

Each major consortium had an overarching target selection and prioritization strategy. By and large, all the SP groups mentioned above pursued targets based on general principles which followed the class-directed target strategies outlined in papers published as part of the dialogue on definition and implementation of structural proteomics.^{2,3} This is exemplified by the paper by Terwilliger *et al.* (see Table 1; Ref. 3), which put forward a list of protein classes and a scientific rationale for selecting them, and also suggested a protocol for implementation of the target selection strategy. Four classes were suggested: 1) the construction of a database of structural motifs; 2) the study of proteins from microorganisms, including pathogens and thermophiles; 3) a large-scale target class including human targets of biomedical interest, protein assemblies, proteins from plants or animals; and 4) a small-scale target class that is the study of important protein families (e.g. protein kinases, transcription factors). Class 1 attempts to generate structural annotations, while the rest are motivated by the goal of generating functional annotation. These two goals are somewhat related (i.e. fold may

Table 1 Target Classes for a Protein Structure Initiative[†]

| Class of Proteins | Importance |
|--|---------------------------------|
| (1) Database of protein structure motifs | Prediction of protein structure |
| (2) Proteins from a microorganism | |
| Proteins from a pathogen | Potential drug targets |
| Proteins from a thermophile | Robust enzyme |
| (3) Large-scale targets | |
| Human proteins | Medical applications |
| Plant or animal proteins | Biotechnology |
| Protein assemblies | Protein interactions |
| (4) Small-scale targets | |
| Groups of structurally-similar proteins | Predicting protein evolution |
| Proteins from a metabolic pathway | Biocatalysis |

[†] From Terwilliger *et al.*³

provide clues to function), and one can argue that the target lists generated by some of the PSI-1 centers are more focused on an attempt to functionally annotate an organism's proteome (e.g. the Joint Center for Structural Genomics (JCSG)'s studies of *Thermotoga maritima*). Below, we summarize the activities of the various SP centers, their initial target selection strategies, and the outcomes. Although the Japanese effort, Protein 3000, has produced more than 2500 structures, accounting for half of the 5000 structures solved in all SP centers worldwide, we have not included an extensive discussion of their efforts. At this time, their target selection strategies remain unpublished, although early descriptions⁴ indicate that their main effort was geared towards the generation of structural annotations, viz. looking for new folds. Final statistics indicate that only 34% of their structures have novel sequences and the list of their solved structures appears to indicate that secondary target selection criteria were used.

Selecting and Prioritizing Targets

Initially, most target selection processes centered on choosing proteins based on primary and secondary objectives, which were later

supplemented by additional target prioritization schemes. For example, projects that aimed at exploring fold space focused on methodology and approaches to find clusters of related sequences for which 3-dimensional structures of any of its members were not available in the PDB. There was also an interest in developing and using tools for the identification of domains to facilitate the production of expression constructs. This was particularly important for proteins anticipated to be difficult to crystallize and thus more amenable to NMR structure determination. Targets were further prioritized based on the predicted novelty of sequences or on the prediction that structures of new folds would be obtained.

Halfway through the first phase of PSI, it became possible to select and prioritize targets utilizing databases created from the results of the large number of SP experiments carried out in the various PSI laboratories that used different target selection strategies on a number of classes of proteins.⁵⁻⁷ Data on diverse data sets were, therefore, available to attempt to understand successes (e.g. produces soluble constructs or crystallizes) and failures. Two papers from the JCSG exemplify what can be done. Canaves *et al.*,⁶ analyzing the JCSG data for *T. maritima* proteins, and Slabinski, *et al.*⁷ analyzing all the available SP data, developed a number of sequence-based metrics, which provides a measure of the difficulty/ease of crystallizing protein (<http://ffas.burnham.org/XtalPred-cgi/xtal.pl>). Both studies use 12 parameters to arrive at an index. Once crystals are obtained, statistics indicate a 32–38% chance of completing the structure (see Table 2). Interestingly, the Barton laboratory⁸ used PDB entries to develop a normalized score, OB_SCORE, based on just two parameters, pI and the Gravy index. This z-score estimates the chances of producing diffraction-quality crystals.⁸ These metrics, along with bioinformatics resources established by the various SP centers, were then used to construct prioritized target lists; in the case of the PSI, prioritization was assigned based primarily on the novelty of the sequence. A more integrated approach to target selection is now provided by a web-based system, sgTarget (<http://www.ysbl.york.ac.uk/sgTarget/>), that produces homology information that measures the uniqueness of the sequence, as well as the calculated physiochemical properties that

Table 2 Success Rates for All Structural Proteomics Centers (October 2007)[†]

| Status | Total Number of Targets | (% Relative to “Cloned” Targets | | (% Relative to “Expressed” Targets | | (% Relative to “Purified” Targets | | (% Relative to “Crystallized” Targets | |
|------------------------------|-------------------------------|---------------------------------------|------|--|------|---|---|---|---|
| | | | | | | | | | |
| Cloned | 102306 | 100 | — | — | — | — | — | — | — |
| Expressed | 64773 | 63.3 | 100 | — | — | — | — | — | — |
| Soluble | 27886 | 27.3 | 43.1 | — | — | — | — | — | — |
| Purified | 25659 | 25.1 | 39.6 | 100 | — | — | — | — | — |
| Crystallized | 9357 | 9.1 | 14.4 | 14.4 | 36.5 | 100 | — | 100 | — |
| Diffraction-quality crystals | 4863 | 4.8 | 7.5 | 7.5 | 19 | 19 | — | 52 | — |
| Diffraction | 4055 | 4 | 6.3 | 6.3 | 15.8 | 15.8 | — | 43.3 | — |
| NMR assigned | 1727 | 1.7 | 2.7 | 2.7 | 6.7 | 6.7 | — | — | — |
| HSQC | 2950 | 2.9 | 4.6 | 4.6 | 11.5 | 11.5 | — | — | — |
| Crystal structure | 3746 | 3.7 | 5.8 | 5.8 | 14.6 | 14.6 | — | 40 | — |
| NMR structure | 1642 | 1.6 | 2.5 | 2.5 | 6.4 | 6.4 | — | — | — |
| In PDB | 5195 | 5.1 | 8 | 8 | 20.2 | 20.2 | — | 38 | — |
| Work stopped | 25278 | — | — | — | — | — | — | — | — |
| Test target | 3 | — | — | — | — | — | — | — | — |
| Other | 1 | — | — | — | — | — | — | — | — |

[†] Table downloaded from http://sg.pdb.org/target_centers.html on October 2007.

may affect expression, solubility, and the likelihood that crystals will be obtained.⁹ In addition to the target selection activities, statistics derived from data-mining activities on production databases (e.g. TargetDB, PepCDB), as well as process and technology evaluation that measured the performance of the various pipelines, are now able to provide a robust estimate and a better understanding of the risks involved in structural studies which had previously been estimated using anecdotal and *ad hoc* approaches.¹⁰

The Protein Structure Initiative (PSI)

The PSI project was established by the National Institutes of General Medical Sciences (NIGMS) at the U.S. National Institutes of Health. PSI phase I (PSI-1) studies, initiated in year 2000 and 2001, were conducted in nine centers (Table 3), and were completed in 2005, with over 1100 structures having been deposited in the PDB. Phase II (PSI-2) studies were immediately started, and involved four large-scale production centers, six specialized centers for development, two homology modeling centers, and a research grants program focusing on improving the accuracy of the comparative protein structure modeling. Production centers in PSI-2 were required to produce 4000 new structures within five years, while the specialized centers were given the mission of developing new tools and approaches to handle challenging targets, including eukaryotic proteins, integral membrane proteins, and large macromolecular complexes. Within two years of operation, the four PSI-2 production centers had deposited about 1200 structures in the PDB, thus exceeding the five-year combined production output of the PSI-1 centers.

The focus of the PSI-1 pilot centers was primarily the development of new tools, technologies, and methodology to increase the success rates and lower the costs of structure determination. Each center was responsible for automating protein sample production and the structure determination pipelines, and for meeting production goals. The final production numbers for the PSI-1 centers are presented in Table 4. As the initial goal of the consortium was to set up the pipelines and test their scalability, about 40% of the total number of

Table 3 List of Major Structural Genomics Organizatons

| Center/Consortium | Target Selection Criteria and Target Organism(s) |
|--|--|
| 1. Berkeley Structural Genomics Center (BSGC), USA http://www.strgen.org | Novel sequences Minimal organisms — <i>M. genitalium</i> , <i>M. pneumoniae</i> |
| 2. Center for Eukaryotic Structural Genomics (CESG), USA http://www.uwstructuralgenomics.org | Novel sequences <i>Arabidopsis thaliana</i> |
| 3. Joint Center for Structural Genomics (JCSG), USA http://www.jcsg.org | Novel sequences <i>Thermatoga Maritima</i> , mouse |
| 4. Midwest Center for Structural Genomics (MCSG), USA http://www.mcsg.anl.gov | Novel sequences Proteins from all three kingdoms of life |
| 5. Mycobacterium Tuberculosis Structural Genomics Consortium (TBSGC), USA http://www.doe-mpi.ucla.edu/TB/ | Novel sequences, <i>Mycobacterium tuberculosis</i> |
| 6. New York Structural Genomics Consortium (NYSGC), USA http://www.nysgrc.org | Novel sequences Disease-related proteins from eukaryotes and bacteria |
| 7. Northeast Structural Genomics Consortium (NESG), USA http://www.nesg.org | Novel sequences Eukaryotic domain families from <i>D. melanogaster</i> , <i>S. cerevisiae</i> , <i>C. elegans</i> , mouse, human |
| 8. Southeast Collaboratory for Structural Genomics (SECSG), USA http://www.scsg.org | Novel sequences <i>P. furiosus</i> , <i>C. elegans</i> , human |
| 9. Structural Genomics of Pathogenic Protozoa (SGPP), USA http://www.sgpp.org Medical Structural Genomics of Pathogenic Protozoa (MSGPP) | Novel sequences, Pathogenic protozoans — <i>Leishmania major</i> , <i>Trypanosoma brucei</i> , <i>Trypanosoma cruzi</i> , <i>Plasmodium falciparum</i> , <i>Entamoeba histolytica</i> , <i>Gardia</i> , <i>Lambliia</i> , <i>Toxomplasma gondii</i> , <i>Cryptosporidium parvum</i> |

(Continued)

Table 3 (Continued)

| | Center/Consortium | Target Selection Criteria and Target Organism(s) |
|-----|--|--|
| 10 | Structural Proteomics in Europe (SPINE), UK http://www.spineurope.org | Bacterial and viral pathogens — <i>B. anthracis</i> , <i>M. tuberculosis</i> , <i>SARS-CoV</i> , <i>Herpes virus</i> Cancer-related proteins Immune defense, neuronal development and neurodegenerative diseases |
| 11. | Structural Genomics Centers, Canada, Sweden, UK http://www.thesgc.com | Human proteins related to diseases and human pathogens |
| 12 | Project 3000, Japan http://www.rsgi.riken.jp | Novel sequences and biologically important or human health related |

structures were determined in the last year of operation. Overall, these studies now show that, based on the results of PSI-1 activities, there is a 5–10% probability of success for a given target in the class of targets included in the PSI-1 list.

Summary of Target Selection and Results from PSI Centers

The overall scientific goal of the PSI effort was to determine enough structures to completely populate a database which could then be used to construct homology-based models covering most of protein space. Thus, the PSI's primary targets were Class 1 proteins of Table 1. However, target selection efforts were not centralized in PSI-1. All that the centers were required to do was to ensure that a significant number of their targets were unique, i.e. have <30% sequence identity with structures already deposited in the PDB. Target selection was also driven by the informal goal of being able to define the complete fold space of proteins; hence, additional selection criteria were applied by the centers themselves, with higher priority given to proteins for which there was a higher expectation of discovering a new fold.

Table 4 Final Production Numbers for PSI-1 Centers†

| Center | All targets | Structures (novel) | | | Crystals | Diffr | NMR | X-Ray | In PDB (novel; unique) | Deposits after Oct 1, 2000 (novel; unique) | % Unique | Annual Rate (last 2 months) | Median Length |
|-----------|-------------|--------------------|----------|-------|----------|------------|-----------------|-----------------|------------------------|--|----------|-----------------------------|---------------|
| | | Cloned | Crystals | Diffr | | | | | | | | | |
| MCSG | 15 565 | 5730 | 888 | 363 | 0 | 296 (274) | 296 (274; 235) | 291 (269; 230) | 79 | 120 (108) | 319 | | |
| NESGC | 12 213 | 5484 | 163 | 116 | 93 | 116 (97) | 198 (169; 138) | 186 (157; 128) | 68.8 | 54 (42) | 191 | | |
| NYSGRC | 2145 | 1538 | 397 | 196 | 0 | 195 (157) | 178 (146; 106) | 171 (139; 100) | 58.5 | 48 (30) | 454 | | |
| JCSG | 6594 | 3650 | 1167 | 268 | 8 | 221 (180) | 198 (160; 104) | 198 (160; 104) | 52.5 | 78 (54) | 415 | | |
| BSGC | 911 | 812 | 94 | 65 | 3 | 58 (50) | 52 (45; 37) | 43 (37; 30) | 69.8 | 0 (0) | 374 | | |
| SECSG | 14 786 | 14 378 | 223 | 118 | 2 | 74 (52) | 71 (51; 29) | 71 (51; 29) | 40.8 | 0 (0) | 214 | | |
| TB | 1758 | 1547 | 209 | 120 | 2 | 107 (70) | 67 (44; 25) | 62 (40; 23) | 37.1 | 12 (12) | 611 | | |
| CESG | 6582 | 4476 | 104 | 40 | 18 | 34 (22) | 47 (33; 27) | 47 (33; 27) | 57.4 | 0 (0) | 166 | | |
| SGPP | 19 503 | 10 154 | 175 | 45 | 0 | 28 (17) | 22 (15; 10) | 22 (15; 10) | 45.5 | 0 (0) | 200 | | |
| Total PSI | 75 104 | 45 391 | 3311 | 1307 | 125 | 1114 (919) | 1111 (937; 711) | 1074 (901; 681) | 63.4 | 312 (246) | 358 | | |

† This table was downloaded from <http://www.mcsig.anl.gov/index.html>

The centers elected to build their programs around certain organisms and/or classes of organisms, prioritizing their target lists by giving higher preferences to novel sequences within these organisms. Thus, a “spread” of protein classes studied was achieved. Table 5 summarizes the target organisms tackled by the PSI-1 centers. Targets from both Class 2 and Class 3 were chosen by these centers. Each center complemented their lists of proteins from their target organisms with orthologs, thereby increasing the chances of obtaining the desired structures.

In the second phase of the PSI (PSI-2), representatives of the four production centers served as members of a centralized target selection committee responsible for generating and maintaining a target list. Prioritization of targets is being carried out within each center, and is usually based on individual scientific interests and technical capabilities of the center. Overall, the goal of PSI-2 remains the same as that of PSI-1, viz. to attempt to characterize protein space completely. This goal is being approached by coarsely sampling pfam and other large protein families for clusters of sequence-related proteins which lack structural representatives in the PDB. Unlike PSI-1, which provided more latitude to the centers in selecting their targets, a set of clearly defined objectives have been set in order to attain this overall goal (see Table 6). Each center is expected to prioritize the members of a sequence family assigned to it by applying a number of criteria, including: 1) families containing representatives from selected model organisms or groups of organisms; 2) families containing representatives with known or postulated disease associations; 3) families containing representatives with predicted or known biological/biochemical functions; and 4) families containing representatives from all the three kingdoms of life.

PSI Targets from Minimal Organisms

The Berkeley Structural Genomics Center (BSGC) has developed its pipeline to work on minimal organisms (i.e. microbes with the smallest genomes), studying proteins from *M. genitalium*, with 486 ORFs, and *M. pneumoniae*, with 687 ORFs, in their respective genomes. The idea was that by studying the proteomes of these minimal organisms, one might gain insight into the minimal requirements for a viable

Table 5 Statistics on PSI Production Levels Classified by Organism[†]

| Organism | Total Number ¹ | Work Stopped | Cloned | Expressed | Purified | Crystallized | Crystal Structure | NMR Structure | In PDB ² |
|--------------------|---------------------------|--------------|--------------|--------------|--------------|--------------|-------------------|---------------|---------------------|
| Total Viruses | 368 | 96 | 204 | 151 | 87 | 15 | 12 | 5 | 17 |
| Archaea | 8901 | 1479 | 6715 | 4174 | 1991 | 526 | 211 | 39 | 245 |
| Bacteria | 66101 | 9933 | 50239 | 34797 | 13557 | 4860 | 1731 | 130 | 1768 |
| Total | 75002 | 11412 | 56954 | 38971 | 15548 | 5386 | 1942 | 169 | 2013 |
| Prokaryotes | | | | | | | | | |
| Yeast | 1983 | 624 | 1411 | 733 | 578 | 84 | 33 | 7 | 35 |
| Plasmodium | 5197 | 268 | 2974 | 1260 | 196 | 65 | 16 | 0 | 16 |
| Trypanosoma | 6403 | 62 | 3953 | 1909 | 299 | 58 | 10 | 0 | 8 |
| Leishmania | 9581 | 288 | 4557 | 2202 | 403 | 146 | 21 | 0 | 17 |
| Arabidopsis | 7525 | 4118 | 4122 | 1663 | 439 | 229 | 38 | 19 | 54 |
| Rice | 130 | 94 | 128 | 62 | 12 | 7 | 1 | 0 | 1 |
| Nematode | 15057 | 3459 | 12634 | 5504 | 417 | 97 | 29 | 3 | 32 |
| Fly | 651 | 270 | 142 | 69 | 20 | 1 | 1 | 0 | 1 |

(Continued)

Table 5 (Continued)

| Organism | Total Number ¹ | Work | | Cloned | Expressed | Purified | Crystallized | Crystal Structure | NMR Structure | In PDB ² |
|--------------|---------------------------|--------------|--------------|--------------|--------------|-------------|--------------|-------------------|---------------|---------------------|
| | | Stopped | Worked | | | | | | | |
| Mouse | 1376 | 660 | 1179 | 866 | 241 | 152 | 28 | 4 | 32 | |
| Human | 8714 | 3466 | 3751 | 2171 | 714 | 188 | 47 | 17 | 64 | |
| Other | 1294 | 126 | 1153 | 962 | 327 | 116 | 36 | 4 | 38 | |
| Eukaryotes | | | | | | | | | | |
| Total | 57911 | 13435 | 36004 | 17401 | 3646 | 1143 | 260 | 54 | 298 | |
| Eukaryotes | | | | | | | | | | |
| Synthetic | 3 | 0 | 3 | 2 | 3 | 1 | 1 | 2 | 3 | |
| Unknown | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | |
| Total | 133285 | 24943 | 93166 | 56526 | 19284 | 6545 | 2215 | 230 | 2331 | |

[†] This table was downloaded from http://sg.pdb.org/target_centers.html

Table 6 PSI-2 Objective of Improving Coverage of Protein Structural Space

-
1. To determine at least one structure for each large, hitherto uncharacterized, protein sequence family using coarse sampling (BIG families)
 2. To determine representative structures for each branch of very large, diverse protein sequence families (MEGA families) to span the structural and functional diversity within that family (moderate sampling to increase structural coverage and to provide structural coverage of selected families with high biomedical relevance)
 3. To determine representative structures for families that are over-represented in the microbiome and metagenome sequence data (moderate sampling of META-family)
 4. To determine the structures of biomedical targets and community-proposed targets.
-

organisms.^{11,12} They have been working on 1036 protein targets from these organisms, determining 87 structures of 61 proteins, 52 of which turned out to be novel proteins. Upon completion of their PSI-1 effort, they announced that their efforts had contributed significantly to the almost complete characterization of the *M. genitalium* proteome, which now has fold assignment for 87% of its proteins. The 486 ORFs in *M. genitalium* include 82 integral membrane proteins, 44 soluble globular proteins, and 10 soluble non-globular proteins. Most importantly, they report that their recent efforts in structural biology and SP have succeeded in enabling fold assignments for over ~90% of the soluble globular proteins in five minimal organisms, *Buchnera aphidicola*, *Blochmannia floridanus*, *Wigglesworthia glossinidia*, *Mycoplasma genitalium*, and *Tropheryma whipplei*, thus, providing a rich data set to further drive attempts to understand the minimal requirements for sustaining life.¹³

PSI Targets from Extremophiles

Two centers opted to study extremophile targets. The JCSG worked on *Thermotoga maritima* (*T. maritima*) and the Southeast Collaboratory for Structural Genomics (SECSG) worked on *Pyrococcus furiosus*. Extremophiles were thought to be ideal targets for early SG studies, since their proteins were expected to be more stable and, therefore, amenable to simplified SG sample production protocols.¹⁴

Full structural and functional characterization of a proteome is now within reach for *Tm*. Although other SP centers have studied proteins from *Tm*, the efforts at the JCSG led to this microbe becoming one of the larger organisms for which a high percentage of its proteome has been structurally characterized. Of its 1877 ORFs, 273 protein structures have been deposited in the PDB, 180 of which were determined by the JCSG. Taking into consideration all of the proteins for which structural information is available, i.e. including those for which structures of homologs are known, it is estimated that 62% of the *T. maritima* proteome now has protein-fold coverage. It should be noted that about 28% of the members of the *T. maritima* proteome have been identified as transmembrane proteins and/or predicted to be of low complexity or to be inherently disordered. Thus, the overall structural coverage from which functional annotation can potentially be generated is quite significant. In addition to the structures in the PDB, over 1000 *Tm* proteins have been purified, about 800 at the JCSG, providing a valuable resource for the community interested in doing further functional and/or biological studies on these proteins. Another important consequence of this work is that questions relating to the putative inherent higher stability of proteins from extremophiles may now be addressed systematically.^{15,16}

PSI Targets from Pathogens

The Structural Genomics of Pathogenic Protozoa (SGPP) center selected proteins exclusively from disease-causing protozoans, while the Mycobacterium Tuberculosis Structural Genomics Consortium (TBSGC) studied proteins from *Mycobacterium tuberculosis* (*Mtb*). Their main interests were in developing a better understanding of the biology of these pathogens, as well as in the development of new therapeutics. Upon completion of the PSI-I program, these two centers continued their studies with funding from the National Institute for Allergy and Infectious Diseases (NIAID), which had decided to fund SP centers focused on protein targets of infectious agents. In 2007, two additional HT centers were funded by the NIAID to determine at least 100 new structures per year, focusing on protein targets from

Table 7 Summary of Progress of the Structure Genomics Consortium (SGC) in Various Target Focus Areas[†]

| Groups | No. of Targets | Targets with | | Targets Purified | Targets in | | Targets with | | Approved Structures |
|---------|----------------|-------------------|--------|------------------|----------------|----------------|--------------|----------|---------------------|
| | | Constructs Cloned | Cloned | | Crystal Trials | Crystal Trials | Crystals | Crystals | |
| TO1 | 308 | 254 | 158 | 95 | 74 | 48 | 53 | | |
| TO2 | 442 | 401 | 177 | 119 | 56 | 46 | 49 | | |
| TO3 | 299 | 219 | 146 | 100 | 55 | 51 | 53 | | |
| Malaria | 333 | 323 | 246 | 171 | 110 | 52 | 51 | | |
| OX1 | 279 | 263 | 137 | 129 | 101 | 66 | 67 | | |
| OX2 | 308 | 198 | 108 | 101 | 80 | 55 | 65 | | |
| OX3 | 383 | 297 | 144 | 139 | 92 | 63 | 64 | | |
| KI | 426 | 376 | 239 | 204 | 108 | 57 | 49 | | |
| Totals | 2778 | 2331 | 1355 | 1058 | 676 | 438 | 451 | | |

Target focus of research groups:

TO1: Proteases, ubiquitylation pathway and cyclophilins

TO2: Chromatin biology and epigenetics

TO3: ATPases and GTPases

Malaria: Malaria and related diseases caused by apicomplexan parasites

OX1: Oxidoreductases and metabolic enzymes

OX2: Membrane receptor signaling

OX3: Phosphorylation dependent signaling

KI: Nucleotide and amino acid metabolism, signalling domains in apoptosis and inflammation, phosphoinositol and lipid signaling and RNA helicases

[†] From http://www.thesgc.com/structures/target_progress.php accessed on October 2007.

organisms implicated in infectious diseases, with the goal of using these new structures for developing new therapeutic protocols.

TBSGC operates as a global structural proteomics effort, with 400 members in 80 institutions.^{17,18} The primary goal of the consortium is to study the function of proteins in the pathogen, targeting those which are potential drug targets or are believed to play key role in *Mtb* biology. By now ~200 unique *Mtb* protein structures and ~250 ligand complexes have been deposited in the PDB. Two thirds of these were produced by the TBSGC, the rest by SPINE, the European structural proteomics consortium, and by other laboratories worldwide. Prior to the SP efforts, there were only 8 *Mtb* protein structures in the PDB. Only ~29% of the protein structures solved by the TBSGC have novel sequences, reflecting less emphasis on finding new folds, and more on working on targets of high relevance to the disease. A notable contribution from the TBSGC is the development of a protocol that can be scaled-up on the genome level to identify (using computational approaches), characterize, and determine the crystal structures of protein-protein complexes.

The SGPP center conducted structural studies on proteins from major pathogenic protozoans. These challenging targets included *Plasmodium falciparum*, the causative agent of the most deadly form of malaria, which is responsible for over one million deaths a year, mostly of children. Using the tools of HT SP, they initiated the task of structurally and functionally characterizing the proteome by attempting expression of about 1000 ORFs, leading to the high-level expression of 63 proteins and to solution of 16 structures. Further, protein engineering studies are underway to improve these success rates. The initial pilot study that led to the expression of these targets is now being analyzed so as to lay the ground work for prioritizing and ultimately eliminating barriers to producing protein samples for structural and functional studies.

PSI Targets from Eukaryotic Organisms

Although proteins from eukaryotes remain a difficult class of targets for SP studies, they include important biomedical targets, and a large number of them have been subjected to analysis. As of October 2007,

the PSI centers have studied 57 911 eukaryotic proteins, but only about 300 structures of such proteins have been completed. Focused development work is currently underway to produce technologies and processes capable of handling these difficult targets.

In most cases, eukaryotic targets were selected by the PSI centers primarily to solve the structures of proteins with novel sequences; thus, most of the centers had several such targets on their list. However, the plant model organism, *Arabidopsis thaliana*, was the primary target proteome for the Center for Eukaryotic Structural Genomics (CESG). They studied more than 4000 targets from this organism and solved 52. They also worked on 4000 other eukaryotic targets, solving 43. All of the other centers, except for the BCSG, worked on a variety of eukaryotic organisms, including mouse and human. The SECSG extensively studied *C. elegans*, working on almost 12 000 protein targets, but solving only 12 of them. These results again highlight the difficulties associated with working with these eukaryotic protein targets in HT studies. Three centers, the Northeast Structural Genomics Consortium (NESG), the Midwest Center for Structural Genomics (MCSG), and the New York Structural Genomics Consortium (NYSGC), did not focus on particular organisms, but rather worked on novel prokaryotic and eukaryotic targets.

SPINE – Structural Proteomics in Europe Project, Function-based Target Selection

The SPINE project was a three-year project that commenced in 2002, and was funded through the EU FP5 program. In 2006, a special issue of *Acta Crystallographica*, Volume 62, was devoted to a description of the work done by this consortium, and provides a comprehensive discussion of its vision for SP, as well as a description of the development of technologies needed to carry out SP projects. Finally, several of the papers in this volume communicate quite succinctly target selection strategies and their correlation with success rates. SPINE produced 375 structures, of which 305 were unique proteins, and the rest of protein–ligand complexes. It had an overall success rate comparable to that of the PSI-1 centers, determining the structures of 12%

of selected targets. Like the PSI effort, SPINE used HT approaches, and was deliberately named as a Structural Proteomics program, so as to differentiate it from structural genomics efforts. SPINE was driven by the notion of “human health targets,” rather than by a bioinformatics-based “fold space” approach.¹⁹ Thus, the criteria of working with novel sequences were not necessarily the primary determinant in selecting targets. Indeed, one early criterion used was to select proteins that could be solved by molecular replacement, which was done primarily to help refine protein production pipelines. The project was subdivided into workpackages, three of which focused on a particular target space of interest. Workpackage 9 focused on proteins from bacterial and viral pathogens, with *B. anthracis* and *Mtb* as the primary bacterial target organisms, and the SARS-CoV and Herpes virus as the viral ones. Workpackage 10 worked on cancer-related proteins (i.e. kinesins, kinases, proteins from the ubiquitin pathway), while Workpackage 11 studied proteins involved in immune defense mechanisms, neuronal development, and on proteins implicated in neurodegenerative diseases. Workpackages 1–8 focused on technology and process development.

New and more challenging targets are being tackled by SPINE2-COMPLEXES, a continuation of the SPINE integrated project funded by the EC within FP6. The project is titled “From Receptor to Gene: Structures of Complexes from Signaling Pathways Linking Immunology, Neurobiology, and Cancer,” and reflects the challenging nature of new efforts as well as the establishment of HT centers throughout Europe. The new targets will be protein–protein and protein–nucleic acid complexes that are related to the areas of investigation as shown in Table 8. This new initiative will require the development of new technologies for the HT study of these complexes, ushering in a new era in structural biology.

Target Selection in SPINE

Discussion of target selection strategies used in the SPINE studies are presented in several papers of the *Acta Crystallographica* special issues.^{20–23} The overall target selection activity was carried out in two distinct phases, the first involving the identification of targets with

Table 8 SPINE2 Target Selection – Complexes in the Following Areas will be Studied

| |
|---|
| WP1.1: Complexes in the ubiquitin signaling pathway |
| <ul style="list-style-type: none"> • Ubiquitination • De-ubiquitination • Proteasome regulation |
| WP1.2: Cell life, death and integrity |
| <ul style="list-style-type: none"> • Cell cycle (including ankyrin repeat protein complexes) • Apoptotic pathways (p53-dependent and p53-independent) • Control of cell integrity (e.g. Lon type protease) |
| WP1.3: Complexes in development and synaptic signaling |
| <ul style="list-style-type: none"> • Development and synaptic signaling assemblies • Protein complexes involving neuronal proteins dependent on copper |
| WP1.4: Protein kinases |
| <ul style="list-style-type: none"> • Signaling and regulatory complexes involving protein kinases |
| WP1.5: Protein phosphatases |
| <ul style="list-style-type: none"> • Protein phosphatases in regulatory cell pathways |
| WP1.6: Receptors/activators associated with transcription complexes |
| <ul style="list-style-type: none"> • Nuclear receptor and transcription factor assemblies • Chromatin remodelling motors |
| WP1.7: Innate immune system |
| <ul style="list-style-type: none"> • Complexes involved in pathogen recognition and subsequent signaling (including TLRs, NOD/NALP proteins, TIR domains, dectin) |
| WP1.8: Adaptive immune system |
| <ul style="list-style-type: none"> • Cell surface receptors and recognition complexes (including MHC, TCR, NK receptor families) |
| WP1.9: Viral subversion of cellular signaling and immune modulation |
| <ul style="list-style-type: none"> • Proteins modulating host responses (including those from EBV and poxviruses) • Proteins involved in host interactions such as receptor binding and fusion |

possible important biomedical roles, and the second being an assessment of whether the target was amenable to structural studies. The criteria used for the second activity were similar to those described by Canaves *et al.*,⁶ using sequence-based predictions of the probability of successfully completing a structural study on a given target. Target selection strategies used by both the SPINE project and the SGC (see below), as well as their success rates, are perhaps more reflective of

what could be achieved by those interested in generating functional annotations and/or interested in generating high-resolution structures, in contrast to the bioinformatics-driven, fold-space coverage approach taken by the PSI efforts described above. For example, for drug design applications, targets that can be solved using molecular replacement approaches are not necessarily excluded.

SPINE Studies on Bacillus anthracis

A total of 359 proteins from *Bacillus anthracis* were targeted for study resulting in the determination of 46 structures.²³ Two rounds of target selection were carried out. The first was used to identify targets with a high probability of successful completion, primarily to help in developing and establishing the HT pipelines and to help fine tune the target selection process. The criteria used to select these “easy” proteins were (1) sizes of <50 kDa; (2) were possible candidates for molecular replacement; (3) were not part of a complex; (4) did not contain signal peptides or transmembrane regions; and (5) were predicted to be soluble, based on the near absence of disordered regions. The second round of selection relied more on biomedical criteria, selecting the proteins that were predicted to be involved in pathogenesis, and of biomedical interest. Finally, additional challenging targets were included, such as those annotated as hypothetical proteins, or those for which putative molecular replacement models were not available.

The *B. anthracis* studies were carried out in two laboratories at the University of York and at Oxford University. Results from the first round of studies from both laboratories, using 48 targets, were quite encouraging. Oxford, after repeated efforts and refining of protocols, attained a structure solution success rate of 31%. In comparison, York attained a success rate of 21%, but did not carry out as many extensive retrials. These success rates were much reduced when the more challenging targets were studied. It is nevertheless quite gratifying to have learned that interesting questions could be answered using HT approaches through a careful target selection process, and that such a

procedure could lead to the solution of a relatively large number of structures at reduced costs.

SPINE Studies on Viral Pathogens

Attempts were also made to study viral pathogens on the SPINE pipelines. The structures of four SARS corona virus (SARS-CoV), four Epstein-Barr virus (EBV), and two vaccinia virus proteins were completed. These studies highlight the difficulties of working with viral proteins in a HT setting. Nevertheless, the results obtained represent significant achievements. The target selection strategy for EBV is discussed in Tarbouriech *et al.*²⁴ EBV selection was oriented towards enzymes and other proteins that had been predicted to have significant secondary structure, were small in size, and had been calculated to have a high stability index. Proteins were deselected if they were predicted to be parts of a multi-protein assembly and/or had trans-membrane domains.

SPINE Studies on Human Proteins of High Biomedical Value

One of the more challenging aspects of the SPINE project was the decision to focus a large portion of the work on human and other eukaryotic proteins that are potentially of high biomedical value. By the start of the SPINE project in 2002, it was clear from early results of the PSI that working with eukaryotic targets required careful attention to target selection in order for some level of success to be achieved. A total of 800 eukaryotic protein targets were selected for study by SPINE, the structures of 170 of which were determined. Biological importance was the primary selection criterion, with calculated physicochemical properties as the secondary criteria. As for the biological targets, preference was given based on the availability of models that could be used in molecular replacement studies.

SGC – Structural Genomics Consortium

The Structural Genomics Consortium (SGC) was organized in 2003 to address industrial and academic pharmaceutical research, and has

thus focused on proteins and protein families from human and apicomplexan (e.g. *Plasmodium falciparum*) that are either potential drug targets or have been implicated in human disease processes. The SGC operates with laboratories at the University of Oxford and University of Toronto, and the Karolinska Institute. Unlike other SP efforts, it uses HT pipelines to study protein-ligand (inhibitors, cofactors, substrates and substrate analogs) interactions and protein families. By 2007, the consortium had solved the structures of over 451 out of the 2778 targets selected for study, corresponding to a success rate of 19%, the highest rate for all the SP centers. Table 7 provides a summary of the status of the structural studies, as well as a list of the target focuses of the various member research groups.

Target Selection in the SGC

SGC target selection protocols, including their list of candidates have not been published, as there is a possibility that this information may be used for commercial advantage. What is available are the areas of interests from which these targets are being chosen (Table 7). The main criteria on is relevance to human health and disease. As an example, one of SGC's areas of interest is signaling pathways. The family of protein kinases, the kinome, are an important class of drug targets, and are a prominent category of protein targets in the list. The consortium has solved 21 novel human kinase structures, and along with other SP efforts, was responsible for raising the number of kinase structures from 38 to 93 by the end of 2006.²⁵ Rather than just solving a unique member of a family, the SGC has elected to attempt total coverage. One important advantage of this approach is that methods and procedures developed for one member of the family could be used for the other members for all the steps in the process, from expression all the way through to crystallization and structure solution. To date, as many as 95% of the targets studied by the SGC have a homologous structure available, simplifying structure solution. Another family that has been extensively studied is that of the human cytosolic sulfur transferases (hSULT).²⁶

Proteins in this family are involved in the metabolism of drugs and hormones, the bioactivation of carcinogens, and the detoxification of

xenobiotics. Knowledge of the structural and mechanistic basis of substrate specificity and activity is crucial for understanding steroid and hormone metabolism, drug sensitivity, pharmacogenomics, and response to environmental toxins. Thus, they form a class of important targets for the pharmaceutical industry. The SGC has solved the structures of five of the 12 hSULT's; these structures, along with those of six others previously characterized by other groups, have permitted the exploration of local and global structural features of members of this enzyme family. In addition to the structural studies, the enzymes were screened for binding and activity towards a panel of potential substrates and inhibitors, revealing unique “chemical fingerprints” for each protein.²⁶

The family-based approach also allows for the exploration of variations among the structures in attempts to develop selective therapeutics. It allows extensive study of small-molecule complexes, thus providing a powerful platform for developing a better understanding of the binding properties of members of the family. For example, in the case of kinases, this has led to the development of inhibitors with picomole potency for PIM kinases and glycogen synthetase kinase 3 (GSK-3) (see Ref. 27).

The SGC is also targeting proteins from *P. falciparum*, the causative organism of malaria, and related apicomplexan organisms. A total of 1008 genes from *P. falciparum* and related organisms have been studied, leading to the determination of 36 structures. This study provides yet another example of a SP survey of a complete organism paralleling those discussed above. The results of the studies are being applied to attempts to develop new vaccines and small molecule therapeutics against the organism.

Expanding the Target List

The ~5000 novel structures that have been produced by the SP efforts within the relatively short time of five years attest to the power of the new HT approaches, since it took almost 40 years to accumulate this number of structures in the PDB using traditional methods. As can be

seen from the various target selection approaches discussed above, the range of questions that now can be realistically addressed has increased dramatically. Although SG was initially considered primarily an enterprise whose only goal was to explore protein fold space, activities within the various centers operating within the PSI, as well as in other SP efforts, clearly demonstrate that other important biological questions can be addressed with these new tools of HT structural biology.

The PSI efforts have demonstrated the success of the bioinformatics-based approach by getting close to completing protein-fold coverage for minimal organisms and for a thermophilic organism. These efforts have also shown that this level of coverage of protein-fold space in eukaryotic organisms will require further advances, primarily in protein expression. One of the major lessons of the PSI efforts is that focusing on prokaryotic and thermophilic organisms to cover fold-space and then building homology models for eukaryotic proteins is an efficient, cost-effective route in cases where homologs exist.

The SPINE efforts have demonstrated the success of function-based approaches to target selection, demonstrating that using biological importance for target prioritization can also lead to high success rates. Furthermore, both the PSI and SPINE have demonstrated the value of using sequence-based metrics that provide estimates for success as a further aid to target selection. The SGC effort provides a powerful approach to target selection in which biological questions, which in their case focus on the needs of the biopharmaceutical industry, are addressed through the study of large families of proteins. This effort has generated high resolution structures that can be used for functional and drug design studies. We feel that this approach may be one that makes the best use of HT technologies to address important biomedical questions.

Clearly, there remain important challenges that must be addressed in order to allow for the expansion of target selection strategies to include other classes of important proteins. For example, statistics for the PSI SP efforts reveal very high failure rates for eukaryotic proteins, with only 298 structures being obtained from ~58 000 targets. Membrane proteins and components of large multimeric/multiprotein

complexes, remain outside the scope of HT efforts. New initiatives, in the frameworks of PSI-2, SPINE2-COMPLEXES, and the NIH sponsored Roadmap Initiative (<http://nihroadmap.nih.gov/structuralbiology/index.asp>), are now trying to address these target areas. As mentioned above, the SPINE2 project is focused on working with protein-protein and protein-nucleic acid complexes, and is now developing new tools and methodology to work with these targets using the HT approaches.

References

1. Kendrew JC, Bodo G, Dintzls HM, *et al.* (1958) "A three-dimensional model of the myoglobin molecule obtained by X-ray analysis." *Nature* **181**: 662–66.
2. Brenner S. (2000) "Target selection for structural genomics." *Nature Struct Biol* **7 Suppl**: 967–69.
3. Terwilliger TC, Waldo G, Peat TS, *et al.* (1998) "Class-directed structure determination: foundation for a protein structure initiative." *Protein Sci* **7**: 1851–56.
4. Yokoyama S, Hirota H, Kigawa T, *et al.* (2000) "Structural genomics in Japan." *Nature Struct Biol* **7 Suppl**: 943–45.
5. Smialowski P, Schmidt T, Cox J, *et al.* (2006) "Will my protein crystallize? A sequence-based predictor." *Proteins* **62**: 343–55.
6. Canaves JM, Page R, Wilson IA, Stevens RC. (2004) "Protein biophysical properties that correlate with crystallization success in the *Thermatoga maritima*: maximum clustering strategy for structural genomics." *J Mol Biol* **344**: 977–91.
7. Slabinski L, Jaroszewski L, Rodrigues AP, *et al.* (2007) "The challenge of protein structure determination lessons from structural genomics." *Protein Sci* **16**: 2472–82.
8. Overton IM, Barton GJ. (2006) "A normalised scale for structural genomics target ranking: the OB-Score." *FEBS Lett* **580**: 4005–09.
9. Rodrigues APC, Grant BJ, Hubbard RE. (2006) "SgTarget: a target selection resource for structural genomics." *Nucl Acid Res* **34**: W225–30.
10. Abola E, Carlton DC, Kuhn P, Stevens RC. (2007) "Five years of increasing structural biology throughput — a retrospective analysis." In H Jhoti and A Leach (eds), *Structure-Based Drug Discovery*, Springer, The Netherlands.
11. Kim SH. (2000) "Structural genomics of microbes: an objective." *Curr Opin Struct Biol* **10**: 380–83.
12. Chandonia J-M, Kim S-H, Brenner SE. (2006) "Target selection and deselection at the Berkeley Structural Genomics Center." *Proteins* **62**: 356–70.
13. Chandonia J-M, Kim S-H. (2006) "Structural proteomics of minimal organisms: conservation of protein fold usage and evolutionary implications." *BMC Struct Biol* **6**: 7.

14. Christendat D, Yee A, Dharamsi A, *et al.* (2000) "Structural proteomics of an archaeon." *Nature Struct Biol* **10**: 903–09.
15. Robinson-Rechavi M, Godzik A. (2005) "Structural genomics of *Thermotoga maritima* proteins shows that contact order is a major determinant of protein thermostability." *Structure (Camb)* **13**: 857–60.
16. Robinson-Rechavi M, Alibes A, Godzik A. (2006) "Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of *Thermotoga maritima*." *J Mol Biol* **356**: 547–57.
17. Baker EN. (2007) "Structural Genomics as an approach towards understanding the biology of tuberculosis." *J Struct Funct Genomics* 2007 Aug 1; (Epub ahead of print).
18. Strong M, Sawaya MR, Wang S, *et al.* (2006) "Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*." *Proc Natl Acad Sci USA* **103**: 8060–65.
19. Stuart DI, Jones EY, Wilson KS, Daenke S. (2006) "SPINE: Structural Proteomics IN Europe: the best of both worlds." *Acta Crystallogr D* **62**: preface.
20. Banci L, Bertini I, Cusack S, *et al.* (2006) "First steps towards effective methods in exploiting high-throughput technologies for the determination of human protein structures of high biomedical value." *Acta Crystallogr D* **62**: 1208–17.
21. Fogg MJ, Alzari P, Bahar M, *et al.* (2006) "Application of the use of high-throughput technologies to the determination of protein structures of bacterial and viral pathogens." *Acta Crystallogr D* **62**: 1196–207.
22. Albeck S, Alzari P, Andreini C, *et al.* (2006) "SPINE bioinformatics and data-management aspects of high-throughput structural biology." *Acta Crystallogr D* **62**: 1184–95.
23. Au K, Berrow NS, Blagova E, *et al.* (2006) "Application of high-throughput technologies to a structural proteomics-type analysis of *Bacillus anthracis*." *Acta Crystallogr D* **62**: 1267–75.
24. Tarbouriech N, Buisson M, Géoui T, *et al.* (2006) "Structural genomics of the Epstein-Barr virus." *Acta Crystallogr D*. **62**: 1276–85.
25. Gileadi O, Knapp S, Lee WH, *et al.* (2007) "The scientific impact of the structural genomics consortium: a protein family and ligand-centered approach to medically-relevant human proteins." *J Struct Funct Genomics* 2007 Oct 12; (Epub ahead of print).
26. Allai-Hassani A, Pan W, Dombrowski L, *et al.* (2007) "Structural and Chemical Profiling of the Human Cytosolic Sulfotransferases." *PLoS Biology* **5**: 1063–78.
27. Fedorov O, Sundström M, Marsden B, Knapp S. (2007) "Insights for the development of specific kinase inhibitors by targeted structural genomics." *Drug Discovery Today* **12**: 365–72.

