

Chapter 19

Structural Genomics and Structural Proteomics: A Global Perspective

Lucia Banci^{*}, Wolfgang Baumeister[†], Udo Heinemann[‡],
Gunter Schneider[§], Israel Silman[¶]
and Joel L. Sussman^{||}

The concept of Structural Genomics (SG) arose towards the mid-1990s as a consequence of the availability of whole-genome information and the success of high-throughput (HTP) methods in DNA sequencing. It was envisaged that similar HTP methods could be applied to determining the 3-D structures of “all” the proteins (the “proteome”) of an organism. As a part of a general research strategy for functional genomics, systematic, genome-driven and high-throughput crystal and NMR structure determination projects were planned. The rationale was that these data would significantly advance our understanding, at the

^{*} Centro Risonanze Magnetiche, University of Florence, Via Luigi Sacconi 6, Sesto Fiorentino, Florence 50019, Italy, Email: banci@cerm.unifi.it.

[†] Max Planck Institute of Biochemistry, Am Klopferspitz 18a, Martinsried D-82152, Germany.

[‡] Max-Delbrück-Center for Molecular Medicine, Robert-Roessle-Str 10, Berlin D-13125, Germany.

[§] Karolinska Institutet, Scheelevägen 2, Stockholm S-171 77, Sweden.

[¶] Department of Neurobiology, Weizmann Institute of Science, Rehovot 76100, Israel

^{||} Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel.

molecular, and eventually, at higher levels, of the functional processes underlying function and dysfunction of the cell and the organism. An interim objective was to provide an efficient way of filling existing gaps in “fold-space,” i.e. to try to determine at least one structure for every existing sequence family, so as to provide suitable templates for modeling the structures of all the proteins present in a given genome. Till now, other gene products such as regulatory RNAs and ribozymes have remained outside the focus of SG projects.

Until quite recently, many structural biologists and protein chemists would have questioned the value of the use of homology modeling, in accurately predicting novel protein structures or for their use in drug design. But there are now an increasing number of examples where predicted structures have proved invaluable in both contexts,¹⁻³ and indeed in engineering proteins capable of performing novel functions.^{4,5}

The SG “vision” led to the investment of very large sums of money in large scale projects, both in the USA (~\$300 million invested by the NIH/NIGMS Protein Structure Initiative (PSI) in nine large centers from 2000 to 2005,⁶ (<http://www.nigms.nih.gov/psi>), and in Japan (~US\$70 million per annum invested in the Protein 3000 national project from 2002 to 2007,⁷ with the bulk going to the RIKEN Research Institute, <http://www.rsgi.riken.go.jp>). Both these national programs were characterized by the concentration of large resources in a small number of big centers, by the concomitant development of novel, automated technologies for implementing a HTP pipeline approach to structure determination; a focus on novel folds as the major criterion for success; and for the US initiative, a policy requiring immediate public deposition of structural data.

In June 2005, the USA NIH/NIGMS activity moved into Phase 2, which involved the large-scale funding of four production units, and funding on a smaller scale of several other centers focussed on the development of complementary new technologies. Phase 2 will run through 2005–2010, again, with a total investment of ~\$300M.⁶

Japan was one of the first countries to embrace SG — Japanese-led projects oriented towards such an approach were conceptualized

as early as 1995. Officially, the SG program in Japan began with the Protein Folds Project, which was initiated at the RIKEN Institute in 1997, and in the following year was transferred to the newly established RIKEN Genomic Sciences Center GSC (<http://www.riken.go.jp>). Another project, known as the Structurome Project began in October, 1999 at the RIKEN Harima Institute of the SPring-8 synchrotron; this project focussed on proteins of the extremophile bacterium, *Thermus thermophilus*.⁸ Although the structurome project used mainly X-ray crystallography for structure determination, the Protein Folds project at the GSC was intimately linked from its inception to GSC's large new nuclear magnetic resonance facility. Research activities within the Protein Folds project focussed on structure determination of mouse and plant proteins, being synergistically aligned with work on DNA libraries developed by scientists at the GSC.

Europe proceeded more slowly than both the USA and Japan in implementing large-scale SG programs. The Protein Structure Factory in Berlin, Germany (<http://www.proteinstrukturfabrik.de>) led the way, followed by the OPPF at Oxford (<http://www.oppf.ox.ac.uk>), England, and the Genopoles in France (notably at Gif-sur-Yvette, Marseille and Strasbourg, <http://rng.cnrg.fr>). However, it was not until October 2002 that the first Europe-wide project began. This was a three-year Integrated Project, funded by the EU FP5 program, which bore the acronym SPINE, standing for Structural Proteomics in Europe (<http://www.spineurope.org>). SPINE was a second-generation project with respect to the evolution of the concept of SG projects, and was deliberately called a Structural Proteomics (SP) project. This name was chosen to make a distinction from the earlier SG initiatives, from which it radically departed, while obviously benefiting from their experience and from the technologies already developed by other projects. Additional SG-related integrated projects were subsequently established and funded by the EC, which had either specific methodological (e.g. BIOHXIT) or thematic (e.g. 3D Repertoire, VIZIER, Interaction PROTEOME) aims, as well as related smaller scale projects. It is worth noting that, even taken together, these projects were, in terms of financial

investment, on a much smaller scale than the corresponding Japanese and US initiatives.

The worldwide activities briefly surveyed above led the SG/SP community to establish an organization called the *International Structural Genomics Organization* (ISGO), so as to exchange and coordinate views and information. ISGO is now a well-established body, which, among other activities, publishes, as its official journal, the *Journal of Structural and Functional Genomics* (JSFG), and organizes a biannual international SG conference.

The Differing Approaches of SG/SP and Classical Structural Biology (SB)

The strategy and implementation of the first SG projects launched in the USA were at the center of a major and thorough debate,⁶ similar to that which preceded the funding and launching of the Human Genome Project a decade earlier. It was realized that implementation of SG programs required even more demanding technological developments than those required for the Human Genome Project. It was necessary to develop HTP procedures for a series of stages, from gene cloning through expression, protein purification, crystallization, data collection to structure analysis and refinement. It is a tribute to the efforts of the various SG projects taken together that automatized procedures have been developed for all these steps, although, hardly surprisingly, there is still much scope for improvement.

Although all the SG projects share the common objective of contributing to the “fold space,” which will permit structural modeling of any protein with a known sequence, the individual SG consortia differ in the criteria for selecting their protein targets. Thus, for example, even within the framework of the US PSI initiative, some consortia chose family-based criteria for target selection, whereas others focussed on the genome of a given organism.

Much discussion was devoted to the productivity that might be expected from the various SG projects in terms of the number of structures determined. The first round of the PSI set a goal of

determining about 10 000 structures. But it soon became clear that this initial goal was unrealistic; by the end of the first five years, ~1300 structures had been solved. The first round of the PSI was, however, successful in developing automatized technologies at a level that permitted the second round to enter into a “production” phase. It was also anticipated that, after an initial peak in generation of novel structures, a decline would occur after the easiest structures, the so-called ‘low-hanging fruit,’ had been determined. Moreover, to quote John Norvell, Director of the PSI at NIGMS/NIH, “... the fact remains that some proteins are not amenable to high-throughput approaches.” Nevertheless, SG projects have already made, and are continuing to make, significant contributions to the determination of new folds and new domains, thus providing the various databases, such as CATH and SCOP,^{9,10} with a substantially larger number of unique new domains than had been provided by the standard structural biology (SB) approach. SG centers have indeed contributed to about half of the new SCOP families, superfamilies and folds in the two and a half years since January 1, 2004. Moreover, the structures solved by SG projects are ~4-fold less sequence-redundant than typical PDB structures.¹¹

Extensive discussions were also directed towards the comparison between the approaches and impact of “SG/SP” versus “Structural Biology (SB)” endeavors. A significant proportion of the structures generated by SG/SP centers have lower citation levels than those generated by SB studies,¹² suggesting that the biological/functional characterization of a protein performed in the context of a classical SB study has a broader impact on the biochemical/biological community. Ultimately, however, the cumulative impact of SG/SP, by providing comprehensive structural data applicable to the majority of proteins, will most certainly exceed the sum of the impacts of the individual structures solved. SG/SP projects aim to achieve as broad a coverage of the proteome as possible. As a consequence, target selection has, in general, been directed towards unique proteins, defined as proteins whose sequence has <30% identity with structures already present in the PDB. In contrast, a SB approach is usually devoted to the detailed study of a limited

number of proteins, often already well characterized in terms of mechanism, specificity and biological role. This may result in the deliberate choice of a number of closely related proteins, or of complexes of a given protein with a number of ligands, in order to address in depth certain aspects of its mode of action and biological function.

It is now becoming apparent that the number of folds is quite limited, and that quite different sequences can assume a similar fold.¹³ An awareness is also emerging that the classical SB approach and the SG/SP approach are, in fact, complementary, as the structure of a given protein is essential for understanding its function; but such an isolated snapshot does not suffice to provide complete functional knowledge.

Another major issue that has attracted the attention of the scientific community, and has promoted an ongoing debate, concerns both the size of the proteins studied and the quality of the structures determined, within the various SG/SP projects, as compared with the individual SB projects. Some scientists and officers of funding agencies had indeed expressed concern that, due to the HTP approaches adopted, the structures determined in SG/SP projects would be of lower quality than of those determined in the individual SB projects. It is widely accepted that the quality of the structures determined in the framework of SP/SG projects is quite comparable to, or even better than that of structures determined in SB projects.¹⁴

If one compares the efforts of the PSI centers with those of traditional SB laboratories in terms of cost/structure, it has been calculated that novel structures solved by PSI centers are significantly less costly than those solved by traditional SB laboratories, whether the structures involved are individual novel structures per se, or new PFAM families or new SCOP superfamilies or folds. Furthermore, there is significant evidence that the cost for solution of structures at the PSI Centers is decreasing quite substantially. However, for large high-impact structures, like the ribosome, which contains a significant number of non-identical polypeptide chains, this is not; in fact, the cost per individual polypeptide chains is significantly lower.¹⁴

The Goals and Policies of International SG/SP Projects

A compilation of the worldwide SG/SP initiatives, which updates progresses on the basis of the Target Registration Database (TB) of the PDB,^{15,16} is presented in Table 1, which also lists the main focus of each project and its website. The first round of the PSI adopted an almost “pure” SG approach, which favored a high production rate for protein structures, and oriented target selection towards the principal goal of completing “fold space.” This focus has been revised in PSI-II, where the focus has also been put on function in target selection, through the funding of specialized centers devoted to specific classes of proteins.

The PSI also invested major efforts, especially in PSI-1, in methodological developments essential for implementation of HTP approaches, and major technological advances were made as a consequence. These advances resulted in a dramatic reduction in the cost per structure. It has been estimated that the average cost per structure at the PSI centers, during the period from 1 February 2004 to 31 January 2005, was US\$138 000.

PSI-2 is more oriented towards structure “production,” exploiting the technical advances obtained during PSI-1. The total number of structures solved during PSI-1 (September 2000 to June 2005) was just over 1100 (<http://www.nigms.nih.gov/Initiatives/PSI/Background/PilotFacts.htm>). During PSI-2, which is still ongoing, ~1200 structures have been solved so far, still far short of the 10 000 structures envisaged at the onset of PSI-1.

In Japan, the SG initiative at the RIKEN Institute focussed on the “fold” approach, i.e. aiming at the determination of the structures of a large number of distinct protein domains. To select the target proteins, mouse and plant genomes were clustered into families on the basis of amino acid sequences, and families for which no experimental structure was yet available were selected. Then, families of particular biological interest were prioritized. For protein production, the cell-free protein production method pioneered at RIKEN¹⁷ has been used

Table 1 Structural Genomics and Proteomics Project List: — Worldwide Initiatives — 2002–2007

Acronym.	Coordinator	Short Description	URL
BSGC PSI-1	Sung-Hou Kim, Lawrence Berkeley Natl. Lab.	The main focus of this initiative is an integrated SG effort on minimal organisms, <i>Mycoplasma genitalium</i> and <i>Mycoplasma pneumoniae</i> , to study proteins essential for life. The goals include classification of fold families, obtaining representative proteins from each family, inferring molecular functions of proteins of unknown function, and optimizing key steps for structure determination. Structures are determined by X-ray crystallography.	http://www.strgen.org
CESG PSI-2	John Markley Univ. Wisconsin Madison	The center is elucidating 3D structures of proteins encoded by the genome of <i>Arabidopsis thaliana</i> , an important model plant. The initial focus of the center is to develop HTP methods for protein production, characterization and structure determination, using X-ray crystallography and NMR spectroscopy.	http://www.uwstructuralgenomics.org

(Continued)

Table 1 (Continued)

Acronym.	Coordinator	Short Description	URL
CHTSB PSI-2	Michael Malkowski	The broad goal of this center is to overcome the most significant obstacles to structure determination by focusing on technology development in areas related to sample preparation for X-ray diffraction studies.	http://www.chtsb.org
CSMP PSI-2	Suzan Bethel	Atomic structure determination of both bacterial and human membrane proteins. Human membrane proteins encode the targets for ~40% of all therapeutic drugs currently used, but understanding of their mechanisms of action at the atomic level is still lacking. Many of the human protein structures sought have therapeutic importance, and their solution will provide atomic-level templates for drug design/discovery.	http://csmmp.ucsf.edu/index.htm
JCSG PSI-2	Ian Wilson	This center is developing HTP methodologies for target selection, protein production, crystallization, and structure determination by	http://www.jcsg.org

(Continued)

Table 1 (Continued)

Acronym.	Coordinator	Short Description	URL
ISFI PSI-2	Thomas C. Terwilliger	X-ray crystallography. Initial focus is on novel structures from <i>Thermotoga maritima</i> , <i>C. elegans</i> and on human proteins thought to be involved in cell signaling. It will also cover the structures of similar proteins from other organisms to ensure the inclusion of the greatest number of different protein folds. The 5-year goal is to generate 3D structures of approximately two thousand proteins.	http://techcenter.mbi.ucla.edu
MCSG PSI-2	Andrzej Joachimiak	This is an NIH PSI Specialized Center focussed on developing and applying a set of synergistic technologies designed to overcome recognized bottlenecks in structure determination at the key steps of production of soluble protein and protein crystallization. The group will select protein targets from Eukarya, Archaea, and Bacteria, with an emphasis on previously unknown folds and on proteins from	http://www.mcs.g.anl.gov/index.html

(Continued)

Table 1 (Continued)

Acronym.	Coordinator	Short Description	URL
NESG PSI-2	Gaetano Montelione	<p>disease-causing organisms. Another focus of this group is to establish methodologies for highly cost-effective protein production, crystallization, structure determination by X-ray crystallography, and refinement, with to the objective of reducing the average cost per structure from \$100 000 to \$20 000.</p> <p>This consortium is targeting proteins from eukaryotic model organisms, which are subjects of extensive functional genomics research, including <i>S. cerevisiae</i>, <i>C. elegans</i>, and <i>D. melanogaster</i>, as well as homologues from the human genome. Its aim is to develop integrated key technologies such as protein expression and structure determination by both X-ray crystallography and NMR spectroscopy. By developing HTP and cost-effective platforms, it plans to solve >180 protein structures per year at a cost, excluding capital equipment, of \$10 000–\$20 000 per structure.</p>	http://bioinfo5.mbb.yale.edu/nescg

(Continued)

Table 1 (Continued)

Acronym.	Coordinator	Short Description	URL
NYSGXRC PSI-2	Stephen Burley	The consortium expects to solve several hundred protein structures from organisms ranging from bacteria to humans, with an emphasis on developing leads for drug discovery. The consortium is also focusing on development of key HTP technologies such as computational methods for protein family classification and target selection, protein production, purification, and structure determination by X-ray crystallography. Its long-term goal is to determine >10000 3D structures.	http://www.nysgrc.org
SECSG PSI-1	Bi-Cheng Wang	This consortium will analyze part of the human genome and the entire genomes of two representative organisms, the eukaryotic microorganism, <i>Caenorhabditis elegans</i> , and its more primitive prokaryotic ancestor, <i>Pyrococcus furiosus</i> . There is an emphasis on technology developments, especially automation of various X-ray crystallography and NMR spectroscopy data collection techniques.	http://www.secsg.org

(Continued)

Table 1 (Continued)

Acronym.	Coordinator	Short Description	URL
SGPP PSI-1	Wim Hol	The focus of this consortium is the development of methods and technologies for determining structures of proteins from pathogenic protozoans, many of which cause deadly diseases such as sleeping sickness (<i>Trypanosoma brucei</i>), Chagas' disease (<i>Trypanosoma cruzi</i>), leishmaniasis (<i>Leishmania</i>) and malaria (<i>Plasmodium falciparum</i> and <i>Plasmodium vivax</i>). Using X-ray crystallography, the consortium plans to discover novel folds and templates for drug design.	http://www.sgpp.org
TBSGC PSI-1	Thomas Terwilliger	The consortium plans to determine and analyze the structures of over 400 proteins from <i>Mycobacterium tuberculosis</i> , including ~40 novel folds and 200 representatives of new protein families, and to analyze these structures in the context of functional information. This will be strongly directed to the design of new and improved drugs and vaccines for tuberculosis. HTP methodology developments have also been carried out as a pilot project using a hyperthermophile. The consortium uses X-ray crystallography for structure determination.	http://www.doe-mbi.ucla.edu/TB

(Continued)

Table 1 (Continued)

Acronym.	Coordinator	Short Description	URL
BSGI	Mirek Cygler	The aim of this project is to allow researches to investigate the function and structure of genes and proteins that can be used in developing new drugs. The facility will emphasize protein mapping, identification and characterization. The project will bring together investigators who use biochemical assays, cell biology methodologies, genomics, protein engineering, DNA chip technology, protein sequence analysis, and X-ray crystallography, among other tools.	http://culer.bri.nrc.ca/brimg/bsgi.html
S2F	John Moulton and Osnat Herzberg	The project determines structures of hypothetical proteins, i.e. those whose structures cannot be related to any previously characterized proteins and whose functions are thus, as yet, unknown. The initial targets have been selected from <i>Hemophilus influenzae</i> . Structure determination utilizes both X-ray crystallography and NMR spectroscopy.	http://s2f.umbi.umd.edu/families.php

(Continued)

Table 1 (Continued)

Acronym.	Coordinator	Short Description	URL
RSGI	Shigeyuki Yokoyama	It focusses on the “fold” approach, i.e. aiming to determine the structures of a large number of distinct protein domains. It has established a high-throughput pipeline for protein sample preparation for structural genomics and proteomics by using cell-free protein synthesis. This center has had a very high success rate, i.e. as of 15 Jan 2008, determining 1343 crystal structures, and 1373 NMR structures.	http://www.rsgi.riken.jp/rsgi_e
KSPRO	Se Won Suh	Its major focus is on proteins from organisms such as <i>Mycobacterium tuberculosis</i> and <i>Helicobacter pylori</i> that may result in novel targets for drug discovery. X-ray crystallography and NMR are being used for structure determination.	http://kspro.org

extensively, being particularly advantageous for producing isotope-labeled proteins for NMR structure solution. Mission-oriented infrastructures were established which exploited an impressive park of NMR spectrometers in Yokohama, as well as the Spring-8 synchrotron at Harima. As of Sep 2007 *ca.* 1914 structures had been released in the PDB, of which 1040 had been solved by NMR.

In China, the Structural Genomics Consortium of the Chinese Academy of Sciences was established in the spring of 2001. Five universities and institutions have joined together to form this consortium, *viz.* the University of Science and Technology of China; the Institute of Biophysics, CAS; the Shanghai Institute for Biological Sciences, and the Shanghai Second Medical University. Five X-ray crystallography groups, three NMR groups, one bioinformatics group and four molecular biology/biochemistry groups are involved in these SG activities. The consortium is focusing on proteins expressed in human hematopoietic stem/progenitor cells, and on proteins related to blood diseases.^{18,19}

In Taiwan, the new synchrotron-based Protein Crystallography Facility at the NSRRC was inaugurated in November 2005 (<http://www.nsrcc.org.tw>). With the NSRRC's protein crystallography beamlines having become operational, Taiwan is a new player in the fields of proteomics and structural genomics.

The Korean Structural Proteomics Research Organization was established in February 2002 to promote and coordinate proteomics research activities in Korea (<http://xtalg.gist.ac.kr>). Its major focus is on proteins of bacteria such as *Mycobacterium tuberculosis* and *Helicobacter pylori*, which may lead to discovery of new drugs for treatment of tuberculosis and ulcers, respectively. Both X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are being used for structure determination.

The Israel Structural Proteomics Center (ISPC) (<http://www.weizmann.ac.il/ISPC>) was established by scientists from the Weizmann Institute of Science, Rehovot, ISRAEL to increase the efficiency of all stages of 3D protein structure determination.²⁰ Targets submitted to the ISPC are primarily related to human health and disease. The center has a unique combination of scientific expertise and state-of-the-art

instrumentation for high-throughput production and crystallization of proteins. Each target is cloned into multiple vectors, using ligation independent cloning. Expression is extensively screened in several bacterial strains with different fusion proteins. Proteins which are not soluble are expressed either in bacterial cell free extracts or in yeast (*Pichia pastoris*). Parallel purification of up to six proteins can be performed using an AKTA3D. Purified proteins are screened for crystallization using a Douglas Instruments ORYX6 robot, which employs the batch method under oil, and a TTP-Labtech MOSQUITO robot for sitting and hanging drops crystallization. This has yielded a high-percentage of high quality diffracting crystals. All the different stages are manipulated by a laboratory information management system (LIMS) in which several bioinformatics tools have been incorporated to facilitate the analysis of our targets. The ISPC now receives targets from scientists both in academia and industry. The ISPC believes that making structural information accessible to the entire scientific community will stimulate novel studies and developments related to health and disease.

The Taiwanese, Korean and Israeli projects show that even relatively small countries are capable of developing domestic SG/SP projects, evidence for the worldwide relevance and impact of the SG/SP endeavor.

In Australia, three SG projects are at the planning stage. They will focus on microbial virulence factors, macrophage proteins, and cold-adapted organisms (<http://www.isgo.org/list/index.php#Australia>).

In addition to the SP/SG projects ongoing in the USA, Europe and Japan, transnational consortia are also being established. The most prominent, to date, is the Structural Genomics Consortium (SGC), headed by Aled Edwards, which was established in 2004, and maintains research centers in Toronto, Oxford and Stockholm. It focuses on human proteins of medical relevance, and is the first consortium to have solved the structures of a large number of human proteins, which are far harder to produce than prokaryotic proteins.²¹

In order to coordinate the efforts of the multiple SG/SP projects currently functioning, the SG/SP community, in particular the publicly funded projects, have agreed on a series of actions directed

towards making all the targets public, ensuring the prompt release of all structures analyzed, and facilitating the open exchange of new technologies as they come “on-line.” A direct outcome of this policy has been the establishment of web sites and repository databases, which are providing the scientific community as a whole with open access to a wealth of data. Particularly relevant are the databases of selected targets, which allow researchers to avoid duplication in target selection (<http://targetdb.pdb.org>). This approach has indeed proved successful, as a recent survey²² reported that only 14% of the structures determined by the various consortia have close homology (>30% sequence identity) with structures analyzed by other consortia. Databases containing information on methodological issues, such as cloning, expression, and purification, are also available. For example, the Protein Expression Cloning and Purification Database, PepcDB (<http://pepcdb.pdb.org>), was established to collect detailed status information and experimental details of each step in the protein production pipeline.¹⁶

Achievements of SG/SP Projects

The few years during which the various SG/SP projects have produced data and results can be used to measure their effectiveness and their impact. A simple way to measure their effectiveness is to count the number of experimentally determined structures that they have generated in terms of their absolute number, the fraction of the total structures deposited in the PDB, and, perhaps more importantly, the fraction of unique structures (defined as such on the basis of the sequence identity being <30% of that of any other structure deposited in the PDB).

In a recent paper, Chandonia and Brenner¹² reviewed the results and impact of SG efforts worldwide, and presented extensive statistics, with particular emphasis on structural novelty. Their analysis showed that the numbers of new structures or, more importantly, of the first structure reported for a PFAM family, came far more often from an SG/SP project than from a classical SB project. SG centers worldwide now account for about half of all new structurally characterized

families. For PSI centers, for example, the percentage of domains representing a new SCOP fold or superfamily was 16%, significantly higher than for the non-SG average, which was 4%. For non-SG/SP structures, >70% of those solved in the past 10 years were related to proteins which had already been structurally characterized in a different state, i.e. with mutations, with bound ligands, or in a different complex.¹²

The analysis of the achievements of SG projects and of the advancements in structural knowledge only in terms of the number of structures, of novel structures and reduced cost per structure is quite reductive. An additional major outcome has been the development of pioneering HTP technologies in the fields of protein production, purification and crystallization, as well as structure determination, using both X-rays and NMR. These achievements have fall-out well beyond the SG projects themselves, also contributing significantly to SB and to life science studies in general.

The structural knowledge provided by the SG/SP projects can suggest functional properties or a biological role for proteins of unknown function. Indeed, in a few cases, newly analyzed structures have been used to infer the functional properties and the mechanism of action of a given protein.

Finally, SG/SP projects, as a spin-off of their HTP approach, which involves screening for expression of several constructs of a number of orthologous genes for each of tens of thousands of targets worldwide, have produced huge archives of cloned and expressed proteins. The vast majority have not resulted in crystals suitable for X-ray data collection, or in samples suitable for NMR spectroscopy. Nevertheless, these archives contain a wealth of precious information for other biochemical and biological studies.

The European Structural Genomics Project SPINE

Europe, which tackled the SP/SG scientific challenge later than both the USA and Japan, has developed an approach combining features of both SP/SG and SB, and exploiting the positive aspects of the two

disciplines. In particular, this has been the approach of SPINE, which was the first Structural Proteomics project to be funded at the European level.

SPINE developed an approach that combines technical and methodological development with the generation of protein structures of high medical relevance, selected from pathogens (as was done in the TB Structural Genomics Consortium (<http://www.doe-mbi.ucla.edu/TB>)) or from human proteins involved in diseases.

A principal contribution of SPINE has been to serve as a catalyst for the development of a pan-European network of laboratories with HTP SG/SP capabilities. SPINE has contributed to the spread of novel technologies (e.g. affordable nano-crystallization and expression screening robotics), rather than establishing large central facilities. It has taken advantage of the diversity of European laboratories so as to generate novel ideas or to benchmark alternative strategies, the best of which have then been more widely adopted.

SPINE has pushed the development of European standards in several areas of HTP technology, notably the development of LIMS systems and automatization of the handling of frozen crystals at synchrotrons (http://www.spineurope.org/page.php?page=protocol_vials), which is already progressing towards courier mail transfer of crystals from the users to synchrotrons, and thus to monitoring of data collection by scientists from their home laboratories. Furthermore, it has been driven by the notion of selecting “high-value targets for human health” rather than by “filling fold space” by solving many of the structures of an entire small proteome, or even by selecting “low-hanging fruit” in the context of development of techniques and methodologies. By so doing, it has provided a pragmatic working definition of the term “structural proteomics.” Surprisingly, despite the fact that many of the targets selected by SPINE were difficult ones, the success rate that it has achieved in the structure determination of human proteins compares favorably with the success rates of other major SG programs focussed on bacterial proteins. Thus, the Joint Center for Structural Genomics (JCSG; <http://www.jcsg.org>), which started in 2000, is one of the most effective large US projects, and has focussed mainly on the proteome of the bacterial thermophile,

Thermotoga maritima (with annual funding substantially greater than that allocated to SPINE). The scoreboard for this project, after seven years of operation, was (on 7/9/07): targets selected: 19 749; cloned: 16 213; expressed: 14 819; crystallized: 1082; solved: 465 (X-ray), 15 (NMR); deposited in PDB: 468. The corresponding output of SPINE after ~5 years operation is highly encouraging and on a par with the US projects: targets selected: 2395; cloned: 1534; expressed: 1177; soluble: 687; solved: 252 (X-ray), 56 (NMR); deposited in PDB: 122 (Jul, 2006). These figures also conceal considerable parallel work on many targets, with the total number of expression trials being ~14 000. The SPINE statistics, showing a total of 308 structures solved, reflect novel structures only; the number including ligand- and metal ion-bound isoforms is >370, with more than 200 being human proteins. To put this in perspective, the total number of new human structures (with <95% identity to prior structures) deposited into the PDB during the first 11 months of 2005 was 337. It should be stressed, however, that the funding for structures that have been “counted” as SPINE targets, has not always been exclusively funded by SPINE only, as was the case for the PSI.

By its policy of maintaining an open decentralized network, together with a focus on high-value targets, SPINE has overcome the potentially divisive dichotomy between the “traditional” way of doing SB (“one post-doc/one project” with in-depth complementary functional investigations) and “factory-style” SG (multiple parallel projects, abandoning of failures, target proteins of often unknown function). The SPINE mode of work, whereby HTP techniques are exploited for high-value targets, is likely to become the norm for SB. SPINE has put in place strong links with a number of companies that have stimulated technology transfer to SMEs, and encouraged beta-testing of new products in SPINE laboratories. Furthermore, the output of SPINE in terms of published papers is outstanding with, to date, 219 publications citing SPINE support.

The current and earlier SG/SP projects have revolutionized the way in which structural biology is now being done worldwide, through the introduction of novel automated, systematic and methodological strategies at each step of the structure determination

pipeline (although their cost-effectiveness, particularly as a method for discovering new folds, is open to discussion). Of perhaps greater importance than the numbers of structures delivered, SPINE has had a remarkable impact in Europe, acting as the springboard for the second generation of FP6 Integrated Projects, such as VIZIER and BIOXHIT, which apply and further hone the appropriate technologies for specific target areas, as well as SPINE2-Complexes, initiated in the summer of 2006 (<http://www.spine2.eu/index.php>), which represents a step forward with respect to the classical, as one might say “old style” structural biology approach, as these new projects exploit a HTP “factory style” approach to address functional processes at the cellular level in their complexity and in their entirety. SPINE2-Complexes has moved on from the goals of SPINE, which were to advance technologies and solve structures of single proteins, to developing approaches for solving structures of protein complexes, with the eventual challenging objective of integrating such complexes into higher-order cellular structures. The measure of the success of the project will not be the number of structures solved but rather their biological impact.

The Scientific Advisory Board of SPINE, in their final review of its achievements during the three-year term for which it was funded, wrote to the European Commission as follows: “*The SPINE impact on the European Community has been very significant and there is no other funding mechanism to accomplish what they have done. SPINE has been a tremendous success as the catalyst for structural biology throughout Europe. This model programme should be duplicated for other EU projects.*”

Highlights of SPINE’S Achievements

The following provides a snapshot of some of the major achievements of the SPINE project that have laid significant foundations on which future SG/SP research can build.

1. Efficient small-scale automated HTP pipelines for protein cloning, expression and purification in prokaryotes, now utilized by many European laboratories both within and outside SPINE.

2. New mammalian expression technologies and refinement of procedures for optimization of expression in eukaryotic systems.
3. Incorporation of quality assurance (QA) into the HTP protein production pipeline, including technologies such as mass spectrometry, ThermoFluor analysis and small-angle X-ray scattering.
4. Methods for achieving soluble expression of protein domains and subdomains, suitable for structural analysis, from previously intractable proteins.
5. Dissemination of nanoliter crystallization technologies.
6. Progress in crystal imaging and image recognition testing.
7. Development of ^{13}C protonless NMR spectroscopy methodology that provides a significant breakthrough in structure solution, particularly for larger proteins.
8. Establishment and testing of an expert system for crystal diffraction data collection from user laboratory to synchrotron; this involves utilization of automated procedures from sample loading through crystal alignment, to data collection and reporting.
9. Development of a SPINE sample holder standard has been adopted across Europe and, more recently, also in China (http://www.spineurope.org/page.php?page=protocol_vials).
10. An integrated protein information server for SG/SP, providing a comprehensive resource for protein selection, annotation and data collection: including PipeAlign, OPAL, OPTIC, FoldIndex, SeqAlert, SeqFacts, RONN, BestPrimers, OPINE, eHTPX hub, ISPyB, DNA automated data collection, ProFunc server, SURFNET, and many others.
11. Solution of the structures of 30 *Bacillus anthracis* structures out of 361 target proteins selected.
12. Analysis of more than 50 high-impact structures, including pathogen and human proteins (see <http://www.spineurope.org>).
13. Contributed to benchmarking definition in SG via a series of multi-lab comparisons of the various stages of expression and protein production.
14. SPINE played a major role in providing credibility for the consideration of structural biology as a research area whose requirements

for infrastructure were eventually incorporated into the ESFRI Roadmap. This resulted in the funding of the preparatory phase of the new infrastructure INSTRUMENT at the beginning of 2008.

The Legacy of SPINE

In large part due to large-scale EU support, SPINE has given visibility and identity to European scientists engaged in SP/SG, and has achieved an international stature comparable to that attained by equivalent large-scale projects in the USA and Japan. This provided an effective mechanism by which worldwide opportunities for scientific exchange in the field could be funnelled through SPINE to individual European laboratories. In a similar way, SPINE, and now SPINE2-COMPLEXES, due to the extensive network developed, can serve as a natural contact point for companies wishing to beta-test new technologies relevant to SG/SP, as positive results can be rapidly disseminated.

SPINE has been exemplary in combining the expertise of the consortium members with that of related consortia, both inside and outside the EC, to pioneer benchmark procedures (e.g. for constructs, expression vectors, folding protocols, crystallization screens and their visual analysis, data collection and rapid structure determination), all of which may result in the adoption of pan-European standards. The establishment of such standards will be greatly facilitated through maintenance of careful quantitative records of both successes and failures, at all stages of the HTP pipeline, by means of the LIMS being built around the PIMS initiative, which arose largely out of preparatory work within SPINE. PIMS is destined to become a *de facto* standard in the area of SP/SG, for which such a standard is sorely lacking.

In parallel to work on structural analysis of the component proteins of the proteome, major efforts are now underway to map the interactions of human proteins (the so-called human “interactome”). This requires that the definition of human complexes be placed on a more systematic and complete basis, and European laboratories are playing an important role in this effort, building on the strong platform of

achievement established by SPINE, and with the FP6 Integrated Project SPINE2-COMPLEXES positioned to play a leading role in this endeavor.

Other European Structural Genomics Projects 2002–2006

Following the success of SPINE, the EC funded a series of other programs in the area of SG/SP, focussing on specific SG/SP technologies, targets, standardization of methods, and plans for the future. The EC has several different instruments to fund projects, all of which require partners from at least three member countries (or associated countries, such as Switzerland and Israel). These programs, together with a brief summary of their activities, are listed in Table 2.

Perspectives

The functional perspective is becoming increasingly relevant both to target selection and prioritization. Analysis of the entries in the PDB has shown that approximately 70% of the human genes with a Gene Ontology annotation (molecular function, biological process or cell component) are not yet structurally characterized by even one identifiable domain. The structural coverage of the human genome is even lower with respect to sequence space; there is approximately 10% coverage by structures with >40% sequence identity.²³

Indeed PDB content, not surprisingly, is significantly enriched in terms of functional coverage in “low-hanging fruit” and validated drug targets. Accordingly, SG projects are beginning to turn their attention from coverage of fold space to that of functional space. This includes individual proteins that are often hard to study, such as membrane proteins; however, attention is increasingly being turned to higher order structures, starting with functional complexes and the long term objective of obtaining structures of organelles and cellular structures.

Table 2 EC Funded SG/SP Projects 2002–2007

Acronym.	Coordinator	Short Description	URL
BIOXHIT	Victor Lamzin	Coordinates scientists at all European synchrotrons, together with leading software developers, in an unprecedented joint effort to develop, assemble and provide a highly effective technology platform for SG.	http://www.bioxhit.org
Vizier	Bruno Canard	Aims to have a groundbreaking impact on the identification of potential new drug targets in RNA viruses through comprehensive structural characterization of the replicative machinery of a carefully selected and diverse set of viruses.	http://www.vizier-europe.org
IMPS	Jean-Luc Popot	Aims to develop broad-range tools for SP of membrane proteins.	http://cordis.europa.eu/fetch?CALLER=FP6_PROJ&ACTION=D&DOC=117&CAT=PROJ&QUERY=1179147785074&RCN=78727
Opticryst	Roslyn Bill	Development, implementation and exploitation of new technologies to overcome bottlenecks in optimization of protein crystallization.	http://www.opticryst.info

(Continued)

Table 2 (Continued)

Acronym.	Coordinator	Short Description	URL
thera-cAMP	Enno Klufmann	Identification of lead compounds that specifically modulate protein-protein interactions in cAMP signaling networks.	http://www.thera-camp.eu
SPINE	David Stuart	Development of new methodologies and technologies for HTP structural biology.	http://www.spineurope.org
SPINE2-Complexes	David Stuart	Structure determination of protein complexes associated with signaling pathways involved in human health and disease, and concomitant development of cutting edge technologies for the production and structure determination of such complexes.	www.spinc2.eu
E-MeP	Roslyn Bill	Development and implementation of new technologies to overcome bottlenecks that preclude the HTP determination of high-resolution structures of membrane proteins and membrane protein complexes.	http://www.e-mep.org
3D Repertoire	Luis Serrano	This project brings together the top European structural biology institutions in a collaboration aimed at solving the structures of a large number of functional protein complexes in yeast.	www.3drepertoire.org

(Continued)

Table 2 (Continued)

AcroAcronym.	Coordinator	Short Description	URL
NMR-Life	Ivano Bertini	Development of cutting-edge NMR technologies for studying functional protein complexes <i>in vitro</i> and <i>in situ</i> .	http://www.postgenomicnmr.net
EXTEND-NMR	Ernest D. Laue	Development of novel computational tools that extend the scope of NMR spectroscopy and make possible functional and structural studies on large proteins and biomolecular complexes.	http://www.biocompetence.eu/index.php/kb_5/io_3577/io.html
UPMAN	Harald Schwalbe	Use of NMR to understand protein misfolding and aggregation.	http://schwalbe.org.chemie.uni-frankfurt.de/upman
FSB-V-RNA	Syben Wijmenga	The structural, functional and virological analysis of RNA and RNA-protein complexes from viruses.	http://www.fsgyrna.nmr.ru.nl
NDDP	Rolf Boelens	Use of cutting-edge NMR techniques to develop a fast, integrated approach for support of structure-based drug design.	http://projects.bijvoet-center.nl/nddp
3D-EM	Andreas Engel	3D-EM aims to establish a standardized platform of advanced technology and methodology that.	http://www.3dem-noc.org

(Continued)

Table 2 (Continued)

Acronym.	Coordinator	Short Description	URL
HT-3DEM	Andreas Engel	will allow Europe to maintain the lead in structural research. It will allow the coordination of research, training activities, research-industry collaboration, and the transfer of knowledge, via publications and focused scientific meetings, in the field of electron microscopy. To enhance European leadership in 3D EM, this project proposes the development of an automated platform permitting HTP screening and analysis of native protein complexes and protein crystals using EM.	http://www.ht3dem.org
MSGP	Christian Cambillau	The project is conducted by a joint CNRS and industrial consortium aimed at the discovery of new anti-bacterial and antiviral targets. The targets include proteins from <i>Escherichia coli</i> and <i>Mycobacterium tuberculosis</i> as well as viral proteins.	http://www.afmb.univ-mrs.fr/rubrique93.html
BIGS	Chantal Abergel	Focuses on the discovery of new antibacterial gene targets among evolutionary conserved genes of uncharacterized function.	http://igs-server.cnrs-mrs.fr

(Continued)

Table 2 (Continued)

Acronym.	Coordinator	Short Description	URL
OPPF	David Stuart, Ernest Laue	Explores the biomedical relevance of human pathogens, in particular of herpes viruses.	http://www.oppf.ox.ac.uk/OPPF/
XMTB	Matthias Wilmanns	Is focussed on the identification of lead compounds against <i>Mycobacterium tuberculosis</i> (TB), using a structure-based approach.	http://xmtb.org
YSG	Herman van Tilbeurgh	A lab-scale platform for the systematic production and structure determination of proteins is being tested on 250 yeast non-membrane proteins of unknown structure. Strategies and final statistics are evaluated.	http://genomics.eu.org/spip
PSF	Udo Heinemann	Target proteins are human proteins relevant to health and disease.	http://www.proteinstrukturfabrik.de
ISPC	Joel Sussman	Aims to increase the efficiency of protein structure determination. Targets submitted to the ISPC are primarily related to human health and disease.	http://www.weizmann.ac.il/ISPC
SGC	Aled Edwards	The SGC operates out of the Universities of Oxford and Toronto and Karolinska Institutet, Stockholm.	http://www.sgc.ox.ac.uk

(Continued)

Table 2 (Continued)

Acronym.	Coordinator	Short Description	URL
FESP	Joel L. Sussman	<p>The primary focus of the Oxford laboratory is the study of human proteins involved in phosphorylation and integral membrane proteins, as well as of enzymes associated with metabolic pathways.</p> <p>The Toronto group seeks to determine the 3D structures of human proteins of therapeutic relevance to diseases such as cancer, diabetes, and metabolic disorders.</p>	http://www.cc-fesp.org

Acknowledgements

This work was supported, in part, by European Commission Grant for the Forum for European Structural Proteomics (FESP), contract number: LSSG-CT-2005-018750.

References

1. Okumoto S, Looger LL, Micheva KD, *et al.* (2005) Detection of glutamate release from neurons by genetically encoded surface-displayed FRET nanosensors. *Proc Natl Acad Sci USA* **102**(24): 8740–45.
2. Becker OM, Dhanoa DS, Marantz Y, *et al.* (2006) An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT_{1A} agonist (PRX-00023) for the treatment of anxiety and depression. *J Med Chem* **49**(11): 3116–35.
3. Dooley AJ, Shindo N, Taggart B, *et al.* (2006) From genome to drug lead: identification of a small-molecule inhibitor of the SARS virus. *Bioorg Med Chem Lett* **16**(4): 830–33.
4. Looger LL, Dwyer MA, Smith JJ, Hellinga HW. (2003) Computational design of receptor and sensor proteins with novel functions. *Nature* **423**(6936): 185–90.
5. Deuschle K, Okumoto S, Fehr M, *et al.* (2005) Construction and optimization of a family of genetically encoded metabolite sensors by semirational protein engineering. *Protein Sci* **14**(9): 2304–14.
6. Norvell J, Berg JM. (2008) Policies in structural genomics/structural proteomics — PSI. In: Sussman JL, Silman I (eds). *Structural Proteomics and its Impact on the Life Sciences* (in press): World Scientific Publishing.
7. Cyranoski D. (2006) “Big science” protein project under fire. *Nature* **443**(7110): 382–.
8. Yokoyama S, Hirota H, Kigawa T, *et al.* (2000) Structural genomics projects in Japan. *Nat Struct Biol* **7** (Suppl): 943–45.
9. Lo Conte L, Ailey B, Hubbard TJ, *et al.* (2000) SCOP: a structural classification of proteins database. *Nucl Acids Res* **28**(1): 257–79.
10. Greene LH, Lewis TE, Addou S, *et al.* (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucl Acids Res* **35**(Database issue): D291–97.
11. Levitt M. (2007) Growth of novel protein structural data. *Proc Natl Acad Sci USA* **104**(9): 3183–88.
12. Chandonia J-M, Brenner SE. (2006) The impact of structural genomics: expectations and outcomes. *Science* **311**(5759): 347–51.

13. 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 Cryst* **62**(Pt 10): 1196–207.
14. Brown EN, Ramaswamy S. (2007) Quality of protein crystal structures. *Acta Crystallogr D Biol Crystallogr* **63**(Pt 9): 941–50.
15. Westbrook J, Feng Z, Chen L, *et al.* (2003) The Protein Data Bank and structural genomics. *Nucl Acids Res* **31**(1): 489–91.
16. Kouranov A, Xie L, de la Cruz J, *et al.* (2006) The RCSB PDB information portal for structural genomics. *Nucl Acids Res* **34**(Database issue): D302–D5.
17. Kigawa T, Yabuki T, Yoshida Y, *et al.* (1999) Cell-free production and stable-isotope labeling of milligram quantities of proteins. *FEBS Lett* **442**(1): 15–19.
18. Shi Y, Wu J. (2007) Structural basis of protein-protein interaction studied by NMR. *J Struct Funct Genomics*: (in press).
19. Gong WM, Liu HY, Niu LW, *et al.* (2003) Structural genomics efforts at the Chinese Academy of Sciences and Peking University. *J Struct Funct Genomics* **4**(2–3): 137–39.
20. Albeck S, Burstein Y, Dym O, *et al.* (2005) 3D structure determination of proteins related to human health in their functional context at the Israel Structural Proteomics Center (ISPC). *Acta Cryst D* **61**: 1364–72.
21. 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* **8**: 107–19.
22. Todd AE, Marsden RL, Thornton JM, Orengo CA. (2005) Progress of structural genomics initiatives: an analysis of solved target structures. *J Mol Biol* **348**(5): 1235–60.
23. Xie L, Bourne PE. (2005) Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Comput Biol* **1**(3): e31.