The Impact of Structural Genomics: Expectations and Outcomes

Running head: Same

Authors: John-Marc Chandonia[1] and Steven E. Brenner[1,2]

Address for correspondence:
Steven E. Brenner
Department of Plant and Microbial Biology
461A Koshland Hall
University of California
Berkeley, CA 94720-3102
email: brenner@compbio.berkeley.edu
fax: (413) 280-7813

Affiliations:
1 - Berkeley Structural Genomics Center, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
2 - Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA

Keywords: cost efficiency, Pfam, SCOP

ABSTRACT

Structural Genomics (SG) projects aim to expand our structural knowledge of biological macromolecules, while lowering the average costs of structure determination. We quantitatively analyzed the novelty, cost, and impact of structures solved by SG centers, and contrast these results with traditional structural biology. The first structure from a protein family is particularly important to reveal the fold and ancient relationships to other proteins. In the last year, approximately half of such structures were solved at a SG center rather than in a traditional laboratory. Furthermore, the cost of solving a structure at the most efficient U.S. center has now dropped to one-quarter the estimated cost of solving a structure by traditional methods. However, top structural biology laboratories are much more efficient than the average, and comparable to SG centers despite working on very challenging structures. Moreover, traditional structural biology papers are cited significantly more often, suggesting greater current impact.

INTRODUCTION

Structural genomics (SG) is an international effort to determine the three-dimensional shapes of all important biological macromolecules, with a primary focus on proteins *(1 and references therein)*. A major secondary goal is to decrease the average cost of structure determination through high throughput methods for protein production and structure determination. In the United States, the National Institutes of Health initiated pilot SG projects at 9 centers through the Protein Structure Initiative (PSI), beginning in 2000. As the PSI project moves from its pilot phase to full production this year, the total funding at four large-scale centers and six specialized centers is expected to be approximately $60 million annually. Considerable resources have also been spent internationally, with SG projects in Japan, Canada, Israel, and Europe underway since the late 1990s. With over 5 years of data from SG projects worldwide, this is an opportune time to examine their impact, and to evaluate how much progress has been made towards the major goals.

As with other large-scale, goal-based projects, it is important to establish objective, quantitative measures of success. We aim to measure the biological importance and difficulty of solving macromolecular structures, and we rely on several proxies to estimate these. Although every new experimental structure adds to our repository of structural data, most structural biologists would agree that novel structures (e.g., the first high-resolution structures of ribosomal subunits [*2, 3*]) are especially valuable. For example, the first protein structure in a family may be used to understand function and mechanism, infer the fold of other family members, create detailed comparative models of the most similar proteins (*4*), or identify previously uncharacterized evolutionary relationships (*5*). Novelty is

not necessarily limited to new families: the structure of a previously solved protein in a different conformation or with a different binding partner could provide insight into its functional mechanisms. Consideration might also be given to the size, complexity, or quality of a structure, as estimates of its difficulty. Over time, a structure's impact on the field may be crudely evaluated by the number of subsequently published papers that cite the original reference.

In this review, we focus on quantifying the impact of SG on expanding structural coverage of protein families, as that is the primary goal of the PSI and several international projects (*6*). We examine several sequence- and structure-based definitions of a protein family, in order to reduce the potential for bias introduced by use of any single standard and to directly compare current results with expectations projected at the outset of the project (*7*). We contrast the number of new families solved and the costs of structure determination at SG centers with the same metrics compiled for structural biology laboratories that are not affiliated with a SG center. We also examine several of the most productive non-SG groups as measured by our standards. Finally, we performed a preliminary analysis of citations of structural publications from both SG and non-SG laboratories.

We expect that this analysis will be helpful for informing future strategy in both SG and structural biology projects, as well as serve as a model for quantitative analysis of the impact of a large-scale project. A complete description of our methodology, and additional detailed results, are provided as Supporting Online Material. Although we focus on PSI centers, we analyze the output of all SG centers that report their results to TargetDB (*8*); these

centers and the specific goals of each are listed in Table S-I of the Supporting Online Material.

Impact of Structural Genomics on Coverage of Protein Families

The Pfam database (*9*) is a manually curated database of protein families from sequenced genomes. As of 1 February 2005, 36% (2,736 of 7,677) of Pfam families (*9*) contain a member with known structure, which allows the folds of all other members of the family to be inferred. We mapped each Pfam family to SG targets and proteins of known structure from the Protein Data Bank (PDB, *10*), and identified the earliest structural representative from each family using the database deposition dates. The rate of first structural characterization of families rose steadily throughout the 1990s, but has leveled off at around 20 new families per month since 1999 (Figure 1b), even as the total number of structures solved continues to increase (Figure 1a). Surprisingly, the rate of solution of first structures in a Pfam family by non-SG structural biologists has decreased in recent years, while SG centers have made up the deficit. SG centers worldwide now account for approximately half of new structurally characterized families, even though they contribute only approximately 20% of the new structures. PSI centers account for approximately two-thirds of the worldwide SG contribution. Only 5% of non-SG structures reported since 2000 represent a new Pfam family, while the PSI average was almost four times greater.

We analyzed the individual contributions of each of the 9 U.S. pilot centers, and compared them to other SG and structural biology efforts (Table I). Results vary widely for the 9 PSI centers. The MCSG was the most productive based on both the total number of

structures solved and total new families, while the BSGC (with which we are affiliated) had the highest fraction of new families and the largest total number of proteins in new families. The bulk of non-PSI SG results were produced by the Japanese center RIKEN. Note that the output of non-PSI SG centers is not expected to be equivalent to PSI centers due to varying budgets and goals, and that two of the PSI centers (CESG and SGPP) started a year later than the others.

Quantifying Novel Structures by Direct Sequence Comparison

To alleviate bias introduced by Pfam, we also examined the number of structures that could not be matched to any prior solved structure using the local sequence comparison methods BLAST (*11*) and PSI-BLAST (*12*), at several different levels of sequence similarity. Results are shown in Figure 2a and Table I.

The overall fraction of structures that were classified as novel according to PSI-BLAST has decreased in the last 15 years, from approximately 20% in 1990 to 10% today. SG structures account for 44% of the total number of novel structures reported in the last year, according to the PSI-BLAST criteria. This result is slightly lower than the Pfam metric for several reasons: although Pfam families often contain more members than can be detected in a single PSI-BLAST search, Pfam does not include many species-specific proteins; moreover, the rate of curation of new families may be lagging behind the rate of discovery of new sequences.

A surprising result is the high proportion of solved SG targets that matched prior structures at 95% ID (sequence identity) or 30% ID thresholds of similarity. For four of the PSI centers (see Figure 2a), over 50% of the structures solved had 30% or more sequence identity to previously solved structures. The fraction of solved targets that were 95% identical to a previously known structure ranged from 4% (SGPP and MCSG) to 21% (CESG), with an average of 8% for PSI centers and 17% for all SG efforts. Some of the variation is due to differing policies between SG centers on what is reported as a target, as we discuss further in the Supporting Online Material.

Impact of Structural Genomics on Identifying New Folds, Superfamilies, and Families

To complement our sequence-based analyses, we evaluated the novelty of protein structures from all sources in the context of the SCOP database (*13*). SCOP provides a widely used, manually curated hierarchy indicating different levels of structural and evolutionary relationship between the protein domains. Domains classified together at the "family" level have a clear common evolutionary origin, and in many cases are sufficiently similar to allow reasonably accurate comparative models to be constructed for any family member using the structure of another as a template (*4*). Groups of families with common structural features or functions that imply a common evolutionary origin are grouped together in "superfamilies." Typically, superfamily relationships are very distant and can only be recognized using structural information. The structure of a single member of a superfamily may be used to confidently predict the overall fold of the other members. Superfamilies that share similar secondary structural features and topology, but for which there is little or no

evidence to suggest a common evolutionary origin, are classified together at the "fold" level. We evaluated each PDB structure to determine how many of its domains represented the first instance of a fold, superfamily, family, protein, or species in SCOP 1.67. For non-SG structures, over 70% of protein domains solved in the last 10 years represent a new experiment on a protein already structurally characterized, although possibly with mutations, bound ligands, or in a different complex. The percentage of domains that represent a new family in SCOP has fallen from 9.6% in 1995 to 4.4% in 2004. This number reflects structural biologists' intentions, as they choose whether to characterize a new family as part of their research design.

Comparison of Structural Genomics Results with Expectations

In 2000, Brenner and Levitt (*7*) predicted that by using standard sequence comparison techniques such as BLAST and PSI-BLAST to avoid targeting homologs of known structures (*1, 14*), SG centers might increase the percentage of new SCOP folds and superfamilies discovered to approximately 40%. Projections based on 2004 data (described in the Supporting Online Material) are remarkably similar.

How well have SG centers met these expectations? We analyzed all targets solved in time to be included in version 1.67 of SCOP (i.e., deposited and released by the PDB prior to 15 May 2004). Results are shown in Figure 2b and Table I. For PSI centers, the percentage of domains that represented a new SCOP fold or superfamily was 16.0%, higher than the non-SG average of 4.0%, but lower than the target of 40%. Results for individual centers varied widely, with much of the difference presumably due to differences in the specific

focus of each center, which resulted in differing strategies for target selection and deselection. The relatively early cutoff date for SCOP limits this analysis: for most centers, between half and three-quarters of the total output has occurred in the last year, too late for analysis by this method. For example, the analysis of SGPP data represents only 4 of 25 targets solved, although 2 of these 4 structures represent new folds. However, the centers with the highest novelty rates in sequence-based tests (BSGC, MCSG, and NESGC) also had the highest rates of discovery of new folds, superfamilies, and families.

Costs of Determining Novel Structures and Families

In cost and productivity data presented to an open session of the NIGMS Advisory Council in 2003, the average cost of solving a protein structure under an R01 grant was estimated as $250,000 - $300,000 (15, 16). Because the methodology behind the estimate is not published, we extrapolate an upper and lower estimate for direct comparison to PSI results. The upper estimate is $300,000 for each PDB entry and the lower estimate is $250,000 for each PDB entry with less than 95% sequence identity to any previously solved entry. We suspect that the lower estimate is closer to the actual figure (see the Supporting Online Material for details). Since the PSI project began in September 2000, the average cost per structure at the pilot centers (including direct and indirect costs) has been $211,000, or 70% to 92% of the estimated cost of solving a structure using traditional methods. In the last year of our study (1 February 2004 - 31 January 2005), the average cost at PSI centers was $138,000 per structure, 46% to 59% of the cost of traditional methods. The most productive center, MCSG, is more than twice as efficient as the average center, having achieved an average cost of only $67,000 per structure over the last year. However,

structures solved by SG centers are on average smaller and contain fewer non-identical polypeptide chains than those from traditional structural biology: when normalized to account for both those factors, per-residue costs for SG in the last year are 66% to 85% (rather than 46% to 59%) of those for non-SG structural biology. This normalization accounts for a presumably higher average degree of difficulty in solving larger structures.

When the costs per novel structure are compared, SG becomes even more efficient. Because the average structural biology laboratory directs most of its research effort towards structures sequence-similar to those already solved, often in order to test hypotheses concerning the function of a particular protein, novel structures are discovered relatively infrequently. Thus, the extrapolated ranges of costs per novel structure using traditional methods are relatively high: $532,000 to $1.9 million per novel structure at the 30% ID level, $1.5 to $5.5 million per new Pfam family, and $2.0 to $7.3 million per new SCOP superfamily or fold. Over the lifetime of the project, PSI centers have averaged costs of $364,000 per novel structure at the 30% ID level, $1.0 million per new Pfam family, and $2.2 million per new SCOP superfamily or fold, with costs in each category lowered by at least 20% in the most recent year of the project. The most efficient center, the MCSG, was 5 to 17 times more cost-efficient than traditional labs in each category in the most recent year of the project. When normalized for structure size, this advantage is still 4- to 14-fold.

These cost data should be interpreted with great caution, since there are many factors not explicitly considered. Besides the imprecision of the traditional structure cost estimate, many SG centers collaborate with non-SG biologists, a process that shifts some of the costs of protein production and structure determination to other groups not supported

by the centers' budgets—and this inflates the apparent productivity of SG. Most SG centers also included targets in their lists that were solved prior to the official start of PSI funding, and the costs of these structures were also not included. On the other hand, most SG centers have invested substantial funds in capital equipment and technology development during the PSI pilot phase. While some technology is already widely used throughout the field (*17*) recent investments may not have yet paid off in increased throughput. Equipment costs are presumably a major factor in structural biology laboratories as well, especially at startup. SG centers also bear additional costs of computation, data reporting, and analysis that are not required of non-SG structural biology labs. Costs of synchrotron time and NMR facilities may not be included in the total cost estimates for either SG centers or other structural biology laboratories. Finally, many structural biology projects benefit from potentially extensive prior work on the biochemical characterization of particular proteins, which is especially important for more challenging structures.

Comparison with Leading Structural Biologists

We include in Table I results for several individual structural biologists who have been among leaders in determining novel structures according to our metrics since 1 January 2000. Tom Steitz's laboratory is best known for solving the structures of protein-nucleic acid complexes, including the large ribosomal subunit (*2*). Robert Huber's group has solved the structures of many macromolecular complexes, including the proteasome (*18*), DNA primase (*19*), and light harvesting complexes (*20, 21*). So Iwata is a leader in membrane crystallography and recently solved the structure of the photosystem II complex (*22*). The total output of each of their laboratories is comparable to the average SG center, and the

output of novel structures surpasses the lowest-performing PSI centers, although both are lower than for the best-performing SG center. The area in which the three groups stood out is in solving large challenging complexes: the Steitz group solved much larger complexes (an average of 12.2 non-identical polypeptide chains per entry) than SG centers, while the Huber and Iwata groups solved somewhat larger complexes composed of larger individual subunits. We caution that our metrics may be biased towards heteromeric complexes.

We calculated the average cost per novel structure solved by Tom Steitz's laboratory, which operates on a total budget of approximately $1.5 million per year (personal communication), compared to approximately $5.7 million for the average PSI center. Since January 2000, the average cost per structure is approximately $166,000, but only $14,000 per non-identical chain (less than one-quarter that of the most recent year of MCSG output). The Steitz lab is also comparable in cost efficiency to PSI centers at solving novel structures. The large ribosomal subunit structure (PDB entry 1ffk, *2*) is especially remarkable in that it revealed 6 proteins with novel folds. Furthermore, our protein-based metrics underestimate the novelty of structures solved by the Steitz lab due to the large number of novel nucleic acid macromolecular structures that were solved.

Comparison of Citations

Several structural biologists have suggested that one measure of the level of interest in a scientific field is the number of published papers in the field, and the impact of a scientific report may be roughly estimated by the number of subsequent citations. We examined the number of citations to the primary reference in each PDB entry for the 104 SG

structures deposited between 1 September 2001 and 31 August 2002. As of November 2005, 34 of the 104 structures remain unpublished, and thus have no citations. The mean number of citations for the 104 structures was 11.0 and the median number was 4. Several factors bias this analysis: the two most-cited references (with 107 and 61 citations, respectively) describe the overall work of a center rather than individual structures, and each was the primary reference for two PDB entries. Also, there were several additional cases in which multiple structures shared the same primary reference, often a functional study, and these were cited more on average than other references. For comparison, we randomly selected 104 non-SG structures solved in the same time period, of which all but six had been published. Like the SG structures, several shared primary references. The 104 structures had a mean of 21.0 citations and a median of 11.5 citations. Thus, publications of SG structures have significantly fewer citations than publications of structures from non-SG laboratories ($p<0.0001$ in a 2-tailed Mann-Whitney test, *23*). For SG structures, novelty did not appear to correlate with the citation rate. Among non-SG structures, novel structures were cited more often than non-novel structures, as traditional structural biologists solved structures likely to have immediate impact on established biochemical research communities.

DISCUSSION

Structural genomics has been extremely successful at increasing the scope of our structural knowledge of protein families. SG efforts worldwide account for nearly half of the protein families for which the first representative was reported solved during the most recent year of our study (February 2004 - January 2005). Despite the pace of SG, the quality of SG

structures has been found to be similar to that of non-SG structures (*24*). The difference in output between the most efficient center and the average is striking.

Despite solving unprecedented numbers of novel protein families, the fraction of structures solved that are novel could be improved at all SG centers. The specific focus of a center may not be entirely compatible with the goal of producing novel structures; for example, a center focusing on medically relevant proteins may need to target multiple members of a family of therapeutic importance. Also, work on a target is not always abandoned when a detectably homologous structure is solved elsewhere, since finishing a near-complete structure may be a worthwhile use of resources. Finally, a structure may not be considered novel because the preceding structure was solved elsewhere but not reported immediately. Rapid reporting of the sequences of newly solved structures could reduce wasted effort at SG centers by at least 4-8% (the minimal level of redundancy observed across all SG centers), saving millions of dollars per year in the U.S. alone.

Compared to other structural biology laboratories, SG centers have published relatively few papers describing their structures, and these papers have a lower average number of citations. This suggests that publication is a bottleneck not easily adapted to high throughput environments. Currently, our estimated costs per citation are similar between SG and non-SG structural biology laboratories, in contrast to other areas in which SG has shown greatly improved efficiency. Although SG centers are reporting results through channels other than traditional publications (*25*), such as public websites and centralized databases (*8*), it is unclear whether structures reported in this manner will individually have the same scientific impact as those reported in traditional publications. Highly cited publications

often describe detailed studies of protein function, and such studies were not funded at the PSI centers in the pilot phase; however, PSI structures may be used as a starting point for such studies. Ultimately, the cumulative impact of SG, by providing comprehensive structural information covering the majority of proteins, is likely to be greater than sum of the impact of the individual structures (as was the case for genome sequencing projects).

Finally, the cost estimates suggest a strategy for direction of future structural biology resources. New families predicted to be tractable with high throughput methods could have basic structural characterization attempted by SG centers, due to the substantial cost savings. These families should be prioritized by significance, for example, family size or biological role (*26, 27*). Non-SG structural biology could focus on hypothesis-driven research on the function or mechanism of individual proteins, as well as characterization of particularly challenging proteins and complexes, and other research that is currently impractical to conduct using high throughput methods. Stephen Harrison points out that leading-edge structural biology studies often rely on integration of data from multiple length and time scales, for which most steps are not currently amenable to high throughput experiments (*28*). Considerable resources will be spent during PSI phase 2 on specialized centers aimed at development of technology for high throughput solution of more challenging structures, such as membrane proteins, eukaryotic proteins, and small protein complexes, which we hope will lead to further gains in efficiency. We view SG and traditional structural biology as playing complementary roles. Structural genomics offers an efficient means to comprehensively survey the protein structure landscape; by structurally characterizing proteins whose significance is not yet understood, it provides a foundation for the next generation of biomedical research. On the other hand, non-SG structural biology focuses on

proteins whose importance is already appreciated, delving deep into particularly rewarding areas to provide immediate scientific impact.

ACKNOWLEDGEMENTS

We thank Jasper Rine, Tom Alber, Tom Steitz, Al Edwards, and Guy Montelione for helpful comments. This work is supported by grants from the NIH (1-P50-GM62412 and 1-K22-HG00056), the Searle Scholars Program (01-L-116), and the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

1. S. E. Brenner, *Nat Rev Genet* **2**, 801-9 (Oct, 2001).
2. N. Ban, P. Nissen, J. Hansen, P. B. Moore, T. A. Steitz, *Science* **289**, 905-20 (Aug 11, 2000).
3. B. T. Wimberly *et al.*, *Nature* **407**, 327-39 (Sep 21, 2000).
4. D. Baker, A. Sali, *Science* **294**, 93-6 (Oct 5, 2001).
5. S. E. Brenner, C. Chothia, T. J. Hubbard, A. G. Murzin, *Methods Enzymol* **266**, 635-43 (1996).
6. P. Smaglik, *Nature* **403**, 691 (Feb 17, 2000).
7. S. E. Brenner, M. Levitt, *Protein Sci* **9**, 197-200 (Jan, 2000).
8. L. Chen, R. Oughtred, H. M. Berman, J. Westbrook, *Bioinformatics* (May 6, 2004).
9. A. Bateman *et al.*, *Nucleic Acids Res* **32 Database issue**, D138-41 (Jan 1, 2004).
10. H. M. Berman *et al.*, *Nucleic Acids Res* **28**, 235-42 (Jan 1, 2000).
11. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J Mol Biol* **215**, 403-10 (Oct 5, 1990).
12. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389-402 (Sep 1, 1997).
13. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, *J Mol Biol* **247**, 536-40 (Apr 7, 1995).
14. S. E. Brenner, *Nat Struct Biol* **7 Suppl**, 967-9 (Nov, 2000).
15. R. Service, *Science* **307**, 1554-8 (Mar 11, 2005).
16. E. Lattman, *Proteins* **54**, 611-5 (Mar 1, 2004).
17. R. C. Stevens, *Nat Struct Mol Biol* **11**, 293-5 (Apr, 2004).
18. J. Lowe *et al.*, *Science* **268**, 533-9 (Apr 28, 1995).
19. M. A. Augustin, R. Huber, J. T. Kaiser, *Nat Struct Biol* **8**, 57-61 (Jan, 2001).
20. J. Deisenhofer, O. Epp, K. Miki, R. Huber, H. Michel, *J Mol Biol* **180**, 385-98 (Dec 5, 1984).
21. R. Huber, *Embo J* **8**, 2125-47 (Aug, 1989).

22. K. N. Ferreira, T. M. Iverson, K. Maghlaoui, J. Barber, S. Iwata, *Science* **303**, 1831-8 (Mar 19, 2004).
23. B. L. v. d. Waerden, *Mathematical statistics*, Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen mit besonderer Berèucksichtigung der Anwendungsgebiete ; Bd. 156 (Springer-Verlag, Berlin, New York,, 1969).
24. A. E. Todd, R. L. Marsden, J. M. Thornton, C. A. Orengo, *J Mol Biol* **348**, 1235-60 (May 20, 2005).
25. A. Wlodawer, *Nat Struct Mol Biol* **12**, 634; discussion 634 (Aug, 2005).
26. J. M. Chandonia, S. E. Brenner, *Proteins* **58**, 166-79 (Jan 1, 2005).
27. J. M. Chandonia, S. E. Brenner, *Proceedings of the 27th International Conference of the IEEE Engineering In Medicine and Biology Society (EMBS)* (2005).
28. S. C. Harrison, *Nat Struct Mol Biol* **11**, 12-5 (Jan, 2004).

FIGURE LEGENDS

Figure 1: Structural Characterization of New Families

a) Black lines indicate the total number of new structures reported per month. Blue lines are contributions of non-SG structural biologists, red lines are from SG centers, and green lines from the PSI centers. The orange line indicates structures that were deposited into the PDB for which the sequence is not available; these structures, which presumably come mainly from structural biologists, were not included in our analysis. b) Total number of new Pfam families with a first representative solved per month, divided into the same categories as in panel a. Monthly totals and a 1-year moving average are shown.

Figure 2: Novelty Rates by Center

a) The fraction of structures from each SG center, and from non-SG structural biologists, that were classified as novel according to each similarity criterion examined. Each structure was classified at the most stringent novelty threshold attained. For example, structures classified as novel at the 95% ID level were between 30% and 95% identical in sequence to a previously reported structure. b) Novelty of domains from SG targets classified in SCOP, by center. StrBio includes all domains solved by non-SG structural biologists (1972 - present). Filtered StrBio includes only domains from non-SG structural biologists filtered but filtered to remove all proteins with sequence similarity to previously solved structures; this represents what structural biologists might produce if they used PSI-BLAST filtering to avoid targeting structures similar to those previously solved. Note that panel a includes data on all structures reported through the end of January 2005, while panel b only includes those structures released by the PDB prior to the cutoff date for inclusion in SCOP 1.67 (15 May 2004).

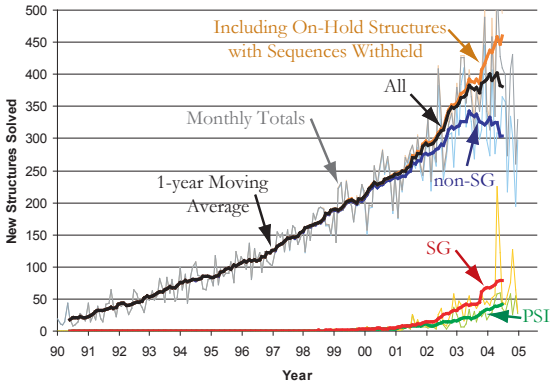
Table I. Novel Structures solved by Structural Genomics Centers and Leading Structural Biology Groups

This shows the total number of novel structures and non-identical polypeptide chains first structurally characterized by SG centers and several leading structural biology groups not affiliated with SG centers. Totals for non-SG structural biology groups were compiled from 1 January 2000. For non-SG centers, each PDB entry was counted as a separate target. The number of non-identical polypeptide chains is also given for each group; this was calculated as the total number of chains with a distinct sequence from other chains within each PDB entry. The number of Pfam families for which the first structure was solved by each group is shown, along with the total number of proteins in these families. The number of novel structures shown is the number of chains with less than 30% sequence identity to any chain from a previously solved structure. The number of new SCOP folds and superfamilies are the number of domains from each group that represented the earliest reported instance of a particular fold or superfamily in the SCOP 1.67 classification.

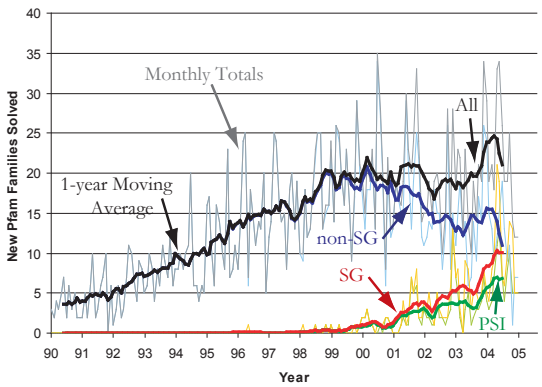
Group or SG Center	Targets and non-identical chains	New Pfam families (total family size)	Novel Structures (30% ID)	New SCOP folds	New SCOP fold or superfamily
SG Centers:					
Berkeley Structural Genomics Center (BSGC)	57 (57 chains)	22 (5,757)	41	4	6
Center for Eukaryotic Structural Genomics (CESG)	48 (48 chains)	7 (387)	28	0	0
Joint Center for Structural Genomics (JCSG)	186 (187 chains)	32 (4,875)	92	3	4
Midwest Center for Structural Genomics (MCSG)	224 (229 chains)	55 (5,512)	163	18	25
Northeast Structural Genomics Consortium (NESGC)	159 (159 chains)	52 (4,811)	108	15	26
New York Structural Genomics Research Consortium (NYSGRC)	166 (171 chains)	27 (3,982)	90	6	9
Southeast Collaboratory for Structural Genomics (SECSG)	67 (67 chains)	6 (1,079)	25	0	1
Structural Genomics of Pathogenic Protozoa Consortium (SGPP)	26 (26 chains)	1 (19)	8	2	2
TB Structural Genomics Consortium (TB)	99 (99 chains)	9 (3,938)	42	0	1
PSI Centers (total of 9 centers above)	1,032 (1,043 chains)	211 (30,360)	597	48	74
Japanese Center (RIKEN)	686 (718 chains)	50 (6,860)	289	10	20
Other International SG (total, excluding all centers above)	169 (183 chains)	33 (5,877)	69	6	9
Non-SG Groups (since 2000):					
Non-SG Structural Biology (total since 2000)	17,096 (23,747 chains)	928 (249,171)	2,521	269	478
Tom Steitz (since 2000)	46 (559 chains)	23 (4,190)	31	7	12
Robert Huber (since 2000)	185 (273 chains)	8 (679)	38	5	10
So Iwata (since 2000)	14 (54 chains)	14 (7,960)	20	2	3

Chandonia and Brenner, Figure 1.

a) New structures solved per month

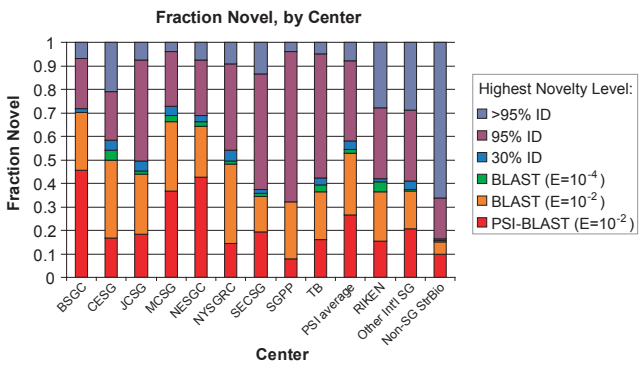


b) Pfam families with a first representative solved, per month

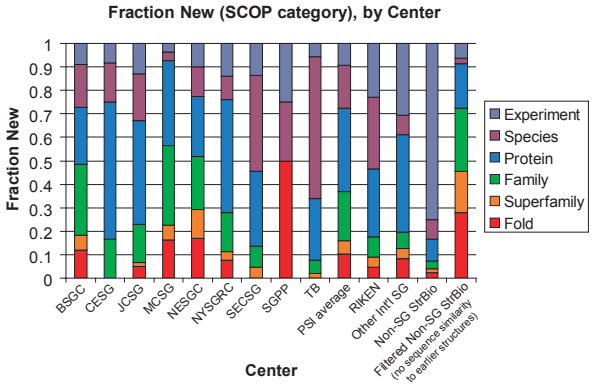


Chandonia and Brenner, Figure 2.

a) Novelty of Structural Genomics Targets, by direct sequence comparison with earlier structures



b) Novelty of Structural Genomics Targets in SCOP



The Impact of Structural Genomics: Expectations and Outcomes

Authors: John-Marc Chandonia[1] and Steven E. Brenner[1,2]

Supporting Online Material

Affiliations:
1 - Berkeley Structural Genomics Center, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
2 - Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA

INTRODUCTION

We present detailed results and descriptions of our methodology in this online supplement. This information is primarily of interest to specialists in the field, and it is required to reproduce our analysis.

RESULTS

SG Centers Included in our Analysis

We analyzed results from all SG centers that report their results to TargetDB (*1*), and which had reported at least one solved structure. These are listed in Table S-I.

Additional Results from Direct Sequence Comparison

To alleviate bias introduced by Pfam, we also examined the number of structures that could not be matched to any prior solved structure using the local sequence comparison methods BLAST (*2*) and PSI-BLAST (*3*), at several different levels of sequence similarity. In addition to the calculations on the number of novel structures solved by each Structural Genomics center and presented in the Results section of the primary manuscript, we present complete results here in Table S-II and in Figure S-1.

Overall, the results from the most sensitive of our direct sequence comparison tests (PSI-BLAST) were most similar to the results from the Pfam metric. However, unlike the number of newly solved Pfam families, the number of newly solved novel structures according to PSI-BLAST has continued to increase rather than leveling off in recent years (Figure S-1a). This result is mostly due to SG efforts: while the number of non-SG novel structures has been fairly level for the last five years, the number of novel SG structures has increased rapidly. SG structures currently account for approximately 44% of the total number of novel structures, according to the PSI-BLAST criteria. Note that SG structures currently account for only about 20% of the total structures being solved, as shown in Figure 1a in the primary manuscript.

Figure S-1b shows the overall fraction of structures that are considered novel according to each similarity criteria tested. The fraction of structures that were classified novel according to PSI-BLAST has decreased in the last 15 years, from approximately 20% in 1990 to approximately 10% today. For the last 15 years, approximately 80% of structures solved have had at least 30% sequence identity to an existing structure. Modeling tools developed in the 1990s have allowed comparative models of moderate accuracy to be constructed for such proteins (*4*). Almost 2/3 of structures solved in the last 15 years had at least 95% sequence identity to an existing structure. This fraction has decreased slightly in recent years, possibly due to the development of more accurate modeling tools (*5*).

As reported in the primary manuscript, we found the number of SG targets that matched previously solved structures at a 95% identity level varied between PSI centers from 4% to 21%. Some of the discrepancy is caused by differing policy on what is reported as a solved target. For example, the BSGC (with which we are affiliated) solved multiple structures for some proteins (e.g., with bound ligands), and reported each PDB entry to TargetDB as different structure of a single target. In this survey, this target would only be counted once, with novelty determined on the earliest date a structure for the target was reported solved (as further explained in the Methods section). Had the BSGC chosen to report each PDB entry as a separate target in TargetDB, this would have resulted in more solved targets and a lower novelty rate, as any subsequent targets would be at least 95% identical in sequence to the first target solved. At the CESG, six proteins were reported to TargetDB as solved twice, in each case using two different target identifiers. As these were reported under separate identifiers, each was counted as a solved structure; however, at least one target from each pair was not considered novel. Had the six additional targets been excluded from our data set, this would have resulted in the CESG having solved 42 targets, with 10% of the targets matching a previously solved target at 95% sequence identity. Since each center sets its own policy on what is reported to TargetDB, we did not attempt to manually curate such cases.

In addition to identifying novel structures at various similarity criteria, as reported above, we also identified "completely novel" structures. In the former set, local similarity to prior structures was allowed, provided at least one region of 50 or more consecutive residues (the size of a small domain) had no local similarity to a prior structure. In the latter set, no local similarity to prior structures was allowed. For example, a multi-domain structure in which only one domain was identified as similar to a prior structure would be characterized as "novel" but not "completely novel." Further details are given in the Methods section. Results on the

number of "completely novel" structures solved by each center are given in Table S-III and shown in Figure S-2.

Comparison of Pfam and PSI-BLAST results

As shown in Figure S-1c, over 90% of the structures solved prior to 1999 are classified in Pfam version 16.0. However, more recent structures are less likely to have been classified in Pfam; only about 60% of structures solved from 2000 to 2004 and classified as novel by PSI-BLAST were from Pfam families. This suggests that the manually curated Pfam-A database has fallen behind the exponentially increasing amounts of sequence data produced in recent years. Although the Pfam authors prioritize the curation of families containing a member with known structure, there is some time required for curation after a novel structure is reported. About 30% of structures classified as novel by PSI-BLAST were members of a previously structurally characterized Pfam family, indicating that many Pfam families contain more members than can be detected in a single PSI-BLAST search.

Complete Results from SCOP Analysis

In addition to the data presented in Table I and Figure 2b in the primary manuscript, we present additional data on the number of novel domains at each level in the SCOP hierarchy (fold, superfamily, family, protein, or species) in Table S-IV.

Figure S-3a shows that for non-SG structures solved in the last 10 years, over 70% of protein domains solved represented a new experiment on a protein already structurally characterized. In 2000, Brenner and Levitt (*6*) predicted that by using standard sequence comparison techniques such as BLAST and PSI-BLAST to avoid targeting homologs of known structures, SG centers might increase the percentage of new folds and superfamilies discovered to approximately 40%. Projections based on current data (shown in Figure S-3b) are remarkably similar.

Detailed Cost Estimates

The PSI centers' approximate total direct and indirect costs are available from the NIH and were calculated for each center as described in the Methods. We can thus calculate the average cost per structure at each PSI center, as well as the cost per novel structure, family, or fold. Detailed results are given in Table S-V, and summarized in the primary manuscript.

Comparison of Structure Size at SG and non-SG Laboratories

We compared the average size of structures produced by both SG and non-SG laboratories, as the size of structures is assumed to roughly correlate with the degree of difficulty. The average number of chains per structure and the average number of residues per chain for each group are given in Table S-VI. To avoid double-counting crystallographically related monomers, only a single chain from each group of 100% identical PDB chain sequences in a single PDB entry was included in our analysis. We also investigated the number of non-identical chains in PDB entries where at least one chain was classified as novel in the direct sequence comparison metric, at the BLAST fine (30% sequence identity) level. Finally, we calculated the average number of "novel residues" in each chain classified as novel; these were defined as all residues in regions not covered by a BLAST hit of at least 30% local sequence identity and 50 residues long to a previously solved structure.

Several results are apparent from Table S-VI. First, while the average number of non-identical chains in structural biology structures was 1.40, few heteromeric structures were solved by structural genomics centers. The average length of chains in non-SG structural biology structures was 10 residues longer than the average for PSI structures, although shorter than structures solved by international SG centers. Therefore, if we calculate cost per residue rather than cost per structure, the cost advantage of structural genomics over the 5-year pilot period is erased. Although the average cost per structure in PSI centers was approximately 70% to 92% of the cost in non-SG laboratories, the average cost per residue (including the effects of chain length and multiple chains) at PSI centers was 2% to 32% higher than for non-SG laboratories. However, in the most recent year, PSI centers are more cost-effective by either measure: while per-structure costs are approximately 46% to 59% of non-SG structural biology costs, per-residue costs are 66% to 85% of those for non-SG structural biology.

Interestingly, novel structures were rarely discovered in heteromeric complexes by either group. No novel structure (at the 30% identity level) was discovered in a heteromeric complex by a SG center. The average number of chains in novel structures solved by non-SG structural biology groups was 1.09, considerably less than the figure of 1.40 for all non-SG structures. In both SG and non-SG groups, the number of "novel residues" per chain in novel structures was somewhat lower than the average number of residues per chain for all structures. When normalized for differences in size (both in the number of novel residues and the number of chains), the cost ratio for novel structures from non-SG structural biology laboratories relative to SG centers is 83% of the original ratio calculated from the data in Table S-V (or 80% when compared to the most productive center, the MCSG). In other words, the cost advantages of structural genomics are reduced by 17 to 20% after normalizing based on the size of structures: the 5- to 18-fold cost advantage of the MCSG in the most recent year over non-SG laboratories at discovering new SCOP folds and superfamilies is reduced to a 4- to 14-fold advantage.

Details of Citation Analysis

In the primary manuscript, we compare the number of citations to structural publications from SG centers to similar publications from non-SG structural biology laboratories. Citations to each publication were obtained using the ISI Web of Science index (http://isiknowledge.com). We initially surveyed 20 randomly chosen structures from among three groups: PSI structures, novel (by either our Pfam or PSI-BLAST criteria) non-SG structures, and non-novel non-SG

structures deposited between 1 September 2001 and 31 August 2002. This time period was chosen to correspond to the second year of the PSI project, as we suspected that many of the PSI publications from the first year would describe the work of the center rather than individual structures, while publications of structures from later years would have had little time to garner citations. We also conducted a more extensive survey of structures from the same time period, as described in the section on "extended citation analysis," below. We caution that both surveys are preliminary.

Details of the PDB entries selected from among each of the three groups described above are given in Table S-VII, Table S-VIII, and Table S-IX respectively. As of 8 July 2005, 8 of the 20 SG structures remain unpublished, and thus have no citations. One SG structure (1kq3) had 86 citations for its paper (*7*), but this report describes the overall work of the center rather than any individual structure. Two other SG structures (1l7n and 1l7o) share a single reference (*8*) that was cited 43 times. The remaining 9 SG publications were cited a total of 48 times. Overall, the publications for the 20 SG structures were cited a total of 218 times, for a mean of 11.0 citations/structure and a median of 1 citation. As of 8 July 2005, the 20 publications of novel non-SG structural biology structures had a mean of 26.2 citations, and a median of 15 citations. All had been published, and each publication was cited a minimum of 7 times. Non-SG structures that were not considered novel had a lower number of citations than the novel non-SG structures: the mean for these 20 structures was 17.6 citations, and the median number was 13.5. Only one had not been published, and one other had not yet been cited.

We compared all three distributions to each other using a 2-tailed Mann-Whitney test (*9*). The calculated p-values were: $p=0.0003$ for SG vs. non-SG novel; $p=0.02$ for SG vs. non-SG non-novel; $p=0.06$ for non-SG novel vs. non-SG non-novel. Thus, publications of SG structures have significantly fewer citations than publications of structures from non-SG laboratories, and novel structures have more citations on average than non-novel structures. For SG structures, novelty did not seem to correlate with citation level: the structure with the most citations (1kq3) was more than 30% identical to a previously deposited structure and received a large number of citations due to referencing a paper describing the overall accomplishments of the center. The paper describing the only novel SCOP fold among the 20 SG structures sampled, 1lql (*10*), had not yet been cited.

To investigate the extent to which older structures accumulate more citations, and whether novel structures had accumulated more citations than non-novel structures over a longer time period, we randomly selected 20 novel and 20 non-novel PDB entries from among all PDB entries solved prior to 1 September 2002 in traditional structural biology laboratories. These entries are shown in Table S-X and Table S-XI, respectively. All of the novel structures had accumulated citations as of 8 July 2005: the median number was 50.5, the mean was 78.0, and the standard deviation was 89.3. Among the non-novel entries, one had not yet been published. The median number of citations was 23.5, the mean was 41.4, and the standard deviation was 65.2. The data indicate that novel structures result in approximately twice as many citations as non-novel ones over time; a 2-tailed Mann-Whitney test indicates the distributions are significantly different from each other with a p-value of 0.02. The same test revealed that the more recent novel non-SG structures (Table S-VIII) had accumulated significantly fewer citations ($p=0.01$) than the sampling of all novel non-SG structures (Table S-X), but that the differences between more recent non-novel non-SG structures (Table S-IX) and the sample of all non-novel non-SG structures (Table S-XI) was less significant ($p=0.15$).

Extended Citation Analysis

As suggested by reviewers, we expanded the citation analysis above to include all 104 PSI structures deposited to the PDB between 1 September 2001 and 31 August 2002, and an equivalent number of non-SG PDB entries from the same time period. The latter structures were randomly selected without regard to novelty. A table showing the number of citations for each SG and non-SG structure as of 22 November 2005 are given in Table S-XII and Table S-XIII, respectively. These results are summarized in the primary manuscript.

Costs per Citation

If we extrapolate the citation rates observed in these samples to all structures, we can estimate the average cost per citation (measured at a time point approximately 3 years after publication of each individual structure). Over the entire five-year period, the cost of SG structures has been approximately $211,000, so with 11.0 citations per structure (in both the limited and extended surveys), the average cost per citation is approximately $19,000. As the cost per structure in SG centers has decreased to $138,000 in the last year, the average cost per citation is expected to be approximately $13,000. For non-SG centers, the average number of citations per structure is approximately 21.0 (based on the more recent sample of 104 structures), so the average cost per citation (based on an estimated cost of $250,000-$300,000 per structure) is approximately $12,000-$14,000. These results should be interpreted with great caution, as a comprehensive study of citations was not possible to perform due to our inability to automatically extract data from the ISI Web of Science product. Because of this limitation, we were not able to account for multiple PDB entries that share a single primary citation, as is often the case for a group of sequence-similar structures involved in a functional study. Furthermore, older structures were observed to have many more citations on average than more recent structures, so it is premature to use the citation metric to estimate the impact of structures solved by structural genomics at this time.

Time Course of PSI Results

For the PSI centers, we plotted a time course for each column of data in Table I in the primary manuscript. These plots are shown in Figure S-4. Note that many centers first reported results prior to their official start date, and that the

relative order of centers when ranked by each metric varied throughout the pilot period.

Final PSI Pilot Phase Results

The pilot phase of the PSI ended on 31 August 2005. Although complete analysis of data deposited after February 2005 is beyond the scope of this study, we show the total number of solved targets reported by each PSI center to TargetDB in Table S-XIV. We caution that this data was not curated as was the data in Table I in the primary manuscript. However, it shows that several hundred additional structures were solved by pilot centers in the last seven months of the PSI pilot phase.

METHODS

Databases

Our database of known protein structures, or "knownstr" was created on 1 Feb 2005. This database contained sequences of every protein chain released by the PDB (*11*), including those of obsolete entries, sequences of proteins deposited in the PDB and made available while the structures were still "on hold," and sequences from TargetDB (*1*), for which a structure had been solved by a participating structural genomics center. These centers are listed in Table S-I. Each protein in knownstr was annotated with a "report date," the date the structure was first reported to the public as solved in one of the above databases. Released PDB entries were annotated as having been reported solved on the deposit date indicated in the entry. Chains from PDB entries on hold were annotated as having been reported solved on the first day the chain was made available by the PDB; we have downloaded all sequences of structures on hold weekly since October 2001, and thus have accurate dates for most if not all of the structures currently on hold. Structural genomics targets were annotated as reported solved on the first date that their status was reported to TargetDB as "Crystal Structure" or "NMR Structure."

The family classification of known structures was evaluated using Pfam version 16.0 (*12*). The HMMER tool (version 2.3.2) (*13*) was used to compare the Pfam_ls library of hidden Markov models to the knownstr database, using the family-specific "trusted cutoff" score as a threshold for assigning significance.

The SCOP (*14, 15*) classification of known structures was evaluated using SCOP version 1.67. Sequences for each ASTRAL domain, and SCOP sccs identifiers (*16*), were obtained from version 1.67 of the ASTRAL database (*17*). The sccs identifiers contain a compact representation of the classification of each domain in SCOP, and were used to look up the degree of similarity in the classification of pairs of domains within the SCOP hierarchy. Obsolete PDB entries were classified in the same way as the entries that superseded them.

The "snr" database of known sequences included all sequences in the `swissprot` and `trembl` files (downloaded 9 November 2004) from Swiss-Prot (*18*), which had been filtered with the SEG (*19*) and PFILT (*20*) programs using default options.

Mapping Equivalent Structures

Because the knownstr database is made from three different sources, it contains some redundancy. For example, a single protein could be present in the database as a structural genomics target from TargetDB, a chain from the PDB on-hold structures, and later as a chain from a released PDB entry. In order to count each protein only once, we created a map of equivalent entries. On-hold PDB structures were mapped to released PDB structures using the PDB identifiers. Structural genomics targets from TargetDB were mapped to PDB entries according to TargetDB annotations. However, because these annotations contained some errors, the target sequences reported in TargetDB were required to have at least 95% sequence identity (calculated using BLAST, as below) to at least one chain in the PDB entry in order to map the entry. In addition, some targets in TargetDB were manually mapped to PDB entries based on examination of the PDB entry headers and sequence alignments. In cases where several knownstr entries were mapped as representing the same protein, but were annotated with different report dates, the earliest report date was used. In cases where reported sequences differed between equivalent entries in the PDB and TargetDB, the sequence from the PDB was considered authoritative and used for all calculations.

Evaluations of Sequence Similarity

To identify sequence similarity among sequences in the knownstr databases, BLAST (version 2.2.4) was used to compare each sequence in the database to all other sequences, using a fixed effective database length of 10^8 residues. Regions of local similarity less than 50 residues long were not considered. Four different similarity criteria were examined. "Coarse" matches required a BLAST E-value of at least 10^{-2}. "Medium" matches required a BLAST E-value of at least 10^{-4}. "Fine" matches required a BLAST E-value at least as significant as 10^{-4} and sequence identity of at least 30% over the region of local similarity. "Ultrafine" matches required a BLAST E-value at least as significant as 10^{-4} and sequence identity of at least 95% over the region of local similarity. Regions of local similarity between two sequences were considered regardless of which sequence was used as the query.

We also evaluated sequence similarity among knownstr sequences using PSI-BLAST version 2.2.4. Position-specific scoring matrices (PSSMs) were constructed for each knownstr sequence using 10 rounds of searching our "snr" database with the default matrix inclusion threshold E-value of 5×10^{-3}. These PSSMs were used to search the database of knownstr sequences, using a fixed effective database length of 10^8 residues. As with the BLAST matches, regions of local similarity less than 50 residues long were eliminated. We examined the remaining regions with PSI-BLAST E-values at least as significant as 10^{-2}.

To evaluate sequence similarity among ASTRAL domains, BLAST (version 2.2.4) was used to compare each sequence in the database to all other sequences, using an effective database length of 10^8 residues. PSI-BLAST position-specific scoring matrices (PSSMs) were constructed for each ASTRAL sequence using 10 rounds of searching our "snr" database with the default matrix inclusion threshold E-value of $5x10^{-3}$. These PSSMs were used to search the database of ASTRAL sequences, using an effective database length of 10^8 residues.

Evaluating Novelty of Structures using Pfam

Each of the 7,677 Pfam-A families from Pfam version 16.0 was mapped to structures in the knownstr database, using HMMER as described above. At least one structural representative was identified for 2,736 families. The structure with the earliest report date (described above) was identified. If the structure was identified as a structural genomics target (either from TargetDB, or a PDB entry mapped as equivalent to a target from TargetDB), the corresponding center was credited with having first solved the family. Otherwise, the family was credited as having been solved by non-SG structural biologists. In cases where the authors of the entry could be identified (using the AUTHOR field in PDB headers), each author was also credited as having first solved the family. Family size for each Pfam family was calculated as the total number of proteins in Pfamseq 16.0 annotated by the Pfam authors as belonging to that family.

Evaluating Novelty of Structures using Direct Sequence Comparison

Each sequence in knownstr was compared to every other sequence using BLAST and PSI-BLAST, as described above. All sequences in knownstr were ordered according to the report date, with ties resolved arbitrarily. Each sequence was tested for novelty (as described below) and then used to mask out regions of sequences with subsequent report dates. All residues in regions of local similarity to an earlier sequence were masked in all subsequently reported sequences. As each sequence was tested for novelty, it was classified as "completely novel" if it was at least 50 residues long and no part of the sequence had been masked by an earlier sequence. Structures were classified as "novel" if there was at least one region of 50 consecutive residues that had not been masked by an earlier sequence. This process was repeated for each of the 4 BLAST similarity criteria we examined, and for PSI-BLAST at an E-value cutoff of 10^{-2}. To mitigate potential problems with incorrectly converged PSI-BLAST PSSMs, regions identified by BLAST with E-values at least as significant as 10^{-2} were included when examining the PSI-BLAST matches. Each novel and completely novel structure at each level of similarity criteria was credited to its authors and/or a structural genomics center, as was done in the Pfam evaluation method described above.

Evaluating Novelty of Structures using SCOP

Domains from all structures released by the PDB and classified in SCOP version 1.67 (cutoff date 15 May 2004) were evaluated for novelty in the context of the SCOP 1.67 hierarchy. To avoid classifying homomers or crystallographically related molecules as redundant, only a single representative of each domain type in a PDB entry with identical SCOP classifications was included in our analysis. The first reported structural representative of every class, fold, superfamily, family, protein, and species in the main classes (1–7) of the SCOP 1.67 classification were determined. Every domain was classified according to the highest level of new information it contained; e.g., an entry that included the first structural representative of a superfamily within a fold that had an earlier structural representative was labeled a "new superfamily." Those entries that did not contain a new domain at any level of the SCOP hierarchy were labeled "new experiments," since they represented a new structure of a previously characterized protein, possibly with different ligands, mutations, or in a different complex than previously deposited structures. We calculated the number of novel domains at every level of the SCOP hierarchy solved by each structural genomics center, and by every author listed in the AUTHORS field of PDB entries. Obsolete PDB entries were assumed to contain the same repertoire of domains as the entries that superseded them. The number of residues in each domain was calculated as the length of the domain sequence in version 1.67 of the ASTRAL database (*17*).

Projecting Expectations of Structural Genomics using SCOP

We used methods described previously (*6*) to filter the full set of genetic domain sequences (*21*) from ASTRAL 1.67 (*17*). We identified a subset of domains that did not have a BLAST or PSI-BLAST E-value score at least as significant as 10^{-2} to any other ASTRAL sequence from a PDB structure deposited at an earlier date, regardless of which of the matching domains was used as a search query. Obsolete entries were not considered in this analysis. Thus, every sequence in this subset represented a "novel" sequence according to criteria similar to the direct sequence criteria described above, although sequence and local alignment length restrictions were not considered, since ASTRAL sequences may be as short as 20 residues. This procedure was designed to directly compare results derived from SCOP version 1.67 to results previously described (*6*) based on SCOP 1.40s. The filtering criteria mimic a target selection strategy that eliminates all potential targets for which a match to a known structure can be found using BLAST and PSI-BLAST searches at a high level of sensitivity (*22, 23*).

Costs per Structure at PSI centers

We based our calculation of the average cost per structure at PSI centers on total direct and indirect costs of $30 million in Y1 (1 Sep 2000 - 31 Aug 2001), $39 million in Y2 (1 Sep 2001 - 31 Aug 2002), $ 52 million in Y3 (1 Sep 2002 - 31 Aug 2003), $68 million in Y4 (1 Sep 2003 - 31 Aug 2004), and $68 million in Y5 (1 Sep 2004 - 31 Aug 2005). For purposes

of calculating an approximate cost per structure, funds were assumed to have distributed evenly among the centers active in a given year. Differing overhead rates at different centers, which affect indirect costs, were also ignored. Costs per month were assumed to be 1/12 of the total annual budget. Seven of the nine centers started in September 2000, and thus have been active for 4 years and 5 months as of the time of this study's data set (through the end of January 2005). The total funding at each of these seven centers was approximately $25.1 million. Two of the centers, CESG and SGPP, have been active since September 2001, or 3 years and 5 months total. The total funding at each of these centers is approximately $20.8 million. The total funding for all 9 centers to date is approximately $217.3 million. SCOP 1.67 includes all structures released by the PDB prior to 15 May 2004, so we calculated costs for SCOP-based metrics beginning with the start of each center through 1 May 2004, assuming a minimum 2-week processing time at the PDB.

Costs per Structure for non-SG Structures

In cost and productivity data presented to an open session of the NIGMS Advisory Council in 2003, the average cost of solving a protein structure under an R01 grant was estimated as $250,000 - $300,000, including direct and indirect costs (*24, 25*). We caution that the methodology behind the NIH estimate is not well documented, and may not represent the cost per PDB entry, but rather the cost per set of nearly sequence-identical entries. We therefore extrapolated both upper and lower bounds on the cost per structure based on the original estimate. As an upper estimate, we assumed that a "structure" was defined as a single PDB entry, and the average cost was $300,000. As a lower estimate, we assumed that a "structure" was defined as a PDB entry that was less than 95% identical in sequence to previously solved entries, and that the average cost was $250,000. As a check on these estimates, we note that traditional structural biology labs worldwide have deposited 17,096 PDB entries between 1 Jan 2000 and 1 Feb 2005 (Table I in the primary manuscript), and that 5,362 were considered novel by our metric at the 95% identity level (Table S-II). The total cost of solving the structures is therefore estimated to be between $1.34 billion (5,362 * $250,000) and $5.13 billion (17,096 * $300,000), or between $264 million and $1.0 billion annually. Although a precise estimate of the total worldwide public and private funds available for structural biology research is impossible to obtain, we suspect the lower estimate is closer to the actual figure.

To estimate the average cost per novel family or structure at non-SG structural biology projects, we extrapolated the upper and lower estimates of the average cost per structure, above, based on the relative numbers of novel structures discovered. For example, because 928 PDB files deposited by traditional labs since 2000 revealed the first structure for a Pfam family (Table I of the primary manuscript), the estimated cost per new Pfam family would range from a lower estimate of $1.5 million ($250,000 * 5,362 / 928) to an upper estimate of $5.5 million ($300,000 * 17,096 / 928).

Selection of non-SG Groups for Comparison to SG Centers

The three individual structural biology laboratories chosen as case studies (Huber, Iwata, and Steitz) were selected for having performed well in all three of our major metrics (Table I in the primary manuscript) despite not having been listed as authors of any PDB entry that was mapped to a SG target in our study. Any individual who appeared in the AUTHOR line of any PDB entry that was mapped to a SG target was excluded from consideration. The remaining individuals were ranked according to our metrics, informally clustered by laboratory, and three laboratories were selected that span a range of specializations. We caution that our metrics may be biased towards large complexes, as a single structure of a large complex may contain several novel chains and representatives of Pfam families.

REFERENCES CITED IN THE ONLINE SUPPLEMENT

1. L. Chen, R. Oughtred, H. M. Berman, J. Westbrook, *Bioinformatics* (May 6, 2004).
2. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J Mol Biol* **215**, 403-10 (Oct 5, 1990).
3. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389-402 (Sep 1, 1997).
4. D. Baker, A. Sali, *Science* **294**, 93-6 (Oct 5, 2001).
5. J. Moult, *Curr Opin Struct Biol* **15**, 285-9 (Jun, 2005).
6. S. E. Brenner, M. Levitt, *Protein Sci* **9**, 197-200 (Jan, 2000).
7. S. A. Lesley *et al.*, *Proc Natl Acad Sci U S A* **99**, 11664-9 (Sep 3, 2002).
8. W. Wang *et al.*, *J Mol Biol* **319**, 421-31 (May 31, 2002).
9. B. L. v. d. Waerden, *Mathematical statistics*, Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen mit besonderer Berèucksichtigung der Anwendungsgebiete ; Bd. 156 (Springer-Verlag, Berlin, New York,, 1969).
10. I. G. Choi *et al.*, *J Struct Funct Genomics* **4**, 31-4 (2003).
11. H. M. Berman *et al.*, *Nucleic Acids Res* **28**, 235-42 (Jan 1, 2000).
12. A. Bateman *et al.*, *Nucleic Acids Res* **32 Database issue**, D138-41 (Jan 1, 2004).
13. S. R. Eddy, *Bioinformatics* **14**, 755-63 (1998).
14. A. Andreeva *et al.*, *Nucleic Acids Res* **32 Database issue**, D226-9 (Jan 1, 2004).
15. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, *J Mol Biol* **247**, 536-40 (Apr 7, 1995).
16. L. Lo Conte, S. E. Brenner, T. J. Hubbard, C. Chothia, A. G. Murzin, *Nucleic Acids Res* **30**, 264-7 (Jan 1, 2002).
17. J. M. Chandonia *et al.*, *Nucleic Acids Res* **32 Database issue**, D189-92 (Jan 1, 2004).
18. B. Boeckmann *et al.*, *Nucleic Acids Res* **31**, 365-70 (Jan 1, 2003).
19. J. C. Wootton, *Comput Chem* **18**, 269-85 (Sep, 1994).
20. D. T. Jones, M. B. Swindells, *Trends Biochem Sci* **27**, 161-4 (Mar, 2002).

21. J. M. Chandonia *et al.*, *Nucleic Acids Res* **30**, 260-3 (Jan 1, 2002).
22. S. E. Brenner, *Nat Rev Genet* **2**, 801-9 (Oct, 2001).
23. S. E. Brenner, *Nat Struct Biol* **7 Suppl**, 967-9 (Nov, 2000).
24. E. Lattman, *Proteins* **54**, 611-5 (Mar 1, 2004).
25. R. Service, *Science* **307**, 1554-8 (Mar 11, 2005).

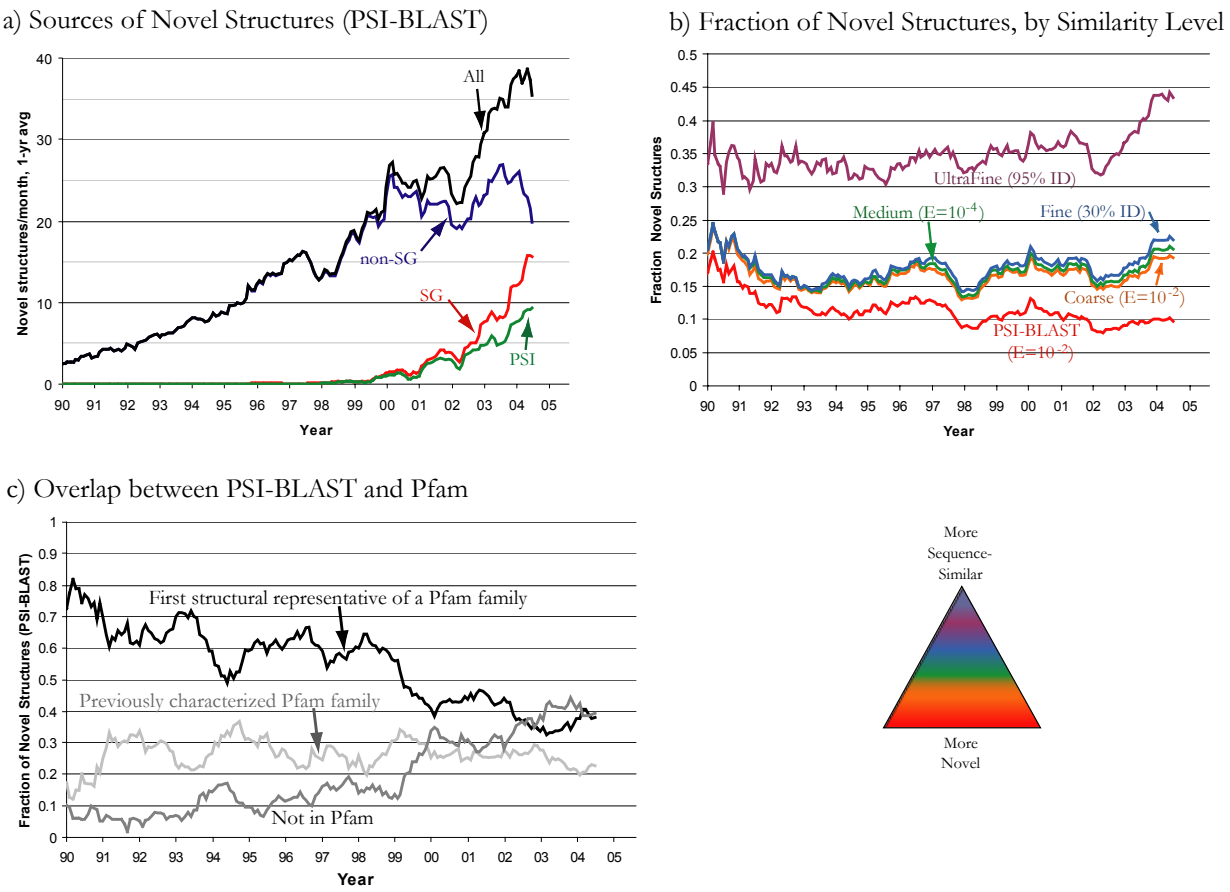
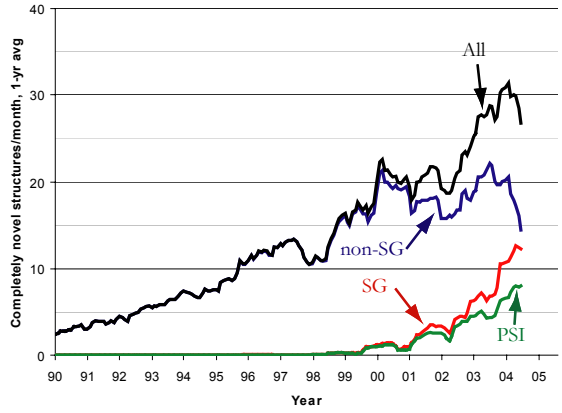


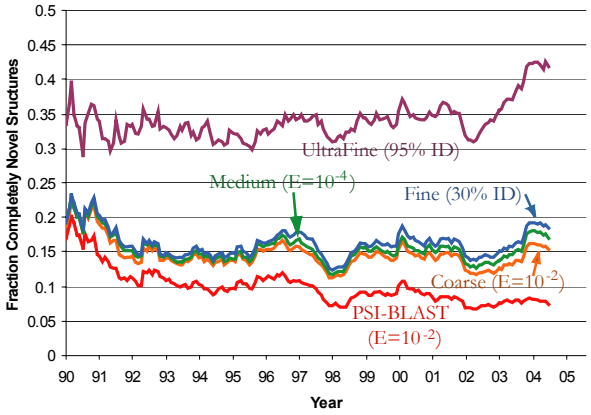
Figure S-1: Novel Structures as Determined by Sequence Comparison Methods

a) The black lines indicate the total number of novel structures solved per month, as determined by PSI-BLAST. The blue lines are contributions of non-SG structural biologists, the red lines are from all SG centers, and the green lines from the PSI centers. b) Fraction of all deposited structures that were novel at each similarity criterion examined. This was calculated as the number of novel chains divided by the number of structures (i.e., PDB entries). In homomers, only the first chain might be considered novel, so this method avoids counting the other chains as redundant. As described in the Methods, "Coarse" matches required a BLAST E-value of at least 10^{-2}. "Medium" matches required a BLAST E-value of at least 10^{-4}. "Fine" matches required a BLAST E-value at least as significant as 10^{-4} and sequence identity of at least 30% over the region of local similarity. "Ultrafine" matches required a BLAST E-value at least as significant as 10^{-4} and sequence identity of at least 95% over the region of local similarity. c) Overlap between structures considered novel according to PSI-BLAST and Pfam. Structures that were novel according to PSI-BLAST were divided into three categories: those that were the first structural representative of a Pfam family, those that belonged to a Pfam family with a prior structural representative, and those that were not classified in Pfam. The fraction in each category is displayed. A 1-year moving average of monthly totals is shown for data in all panels.

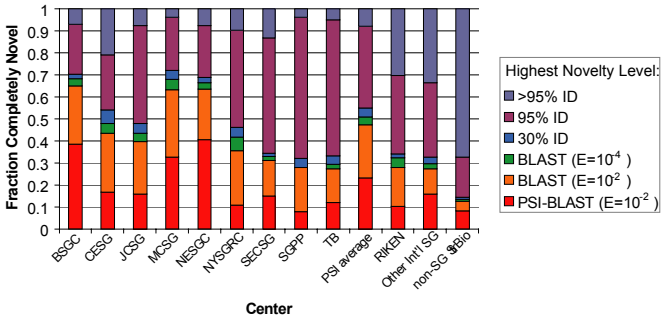
a) Sources of Completely Novel Structures (PSI-BLAST)



b) Fraction of Completely Novel Structures, by Similarity Level



c) Fraction of Completely Novel Structures, by Center and Level



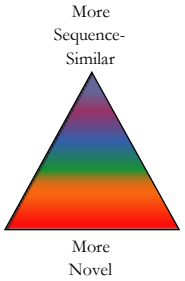
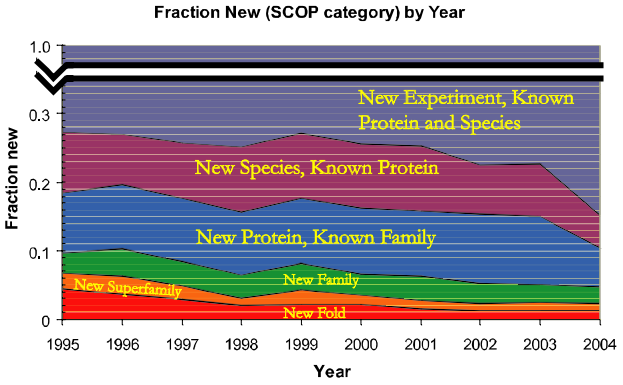


Figure S-2: Completely Novel Structures as Determined by Sequence Comparison Methods

Completely novel structures are those with no local sequence similarity (at a given criterion) to chains from previously solved structures. a) The black lines indicate the total number of completely novel structures solved per month, as determined by PSI-BLAST. The blue lines are contributions of non-SG structural biologists, the red lines are from SG centers, and the green lines from the PSI centers. b) Fraction of all deposited structures that were completely novel at each similarity criterion examined. This was calculated as the number of completely novel chains divided by the number of structures (i.e., PDB entries). In homomers, only the first chain might be considered novel, so this method avoids counting the other chains as redundant. As described in the Methods, "Coarse" matches required a BLAST E-value of at least 10^{-2}. "Medium" matches required a BLAST E-value of at least 10^{-4}. "Fine" matches required a BLAST E-value at least as significant as 10^{-4} and sequence identity of at least 30% over the region of local similarity. "Ultrafine" matches required a BLAST E-value at least as significant as 10^{-4} and sequence identity of at least 95% over the region of local similarity. c) The fraction of structures from each SG center, and from non-SG structural biologists (Non-SG StrBio) that were classified as completely novel according to each criterion. A 1-year moving average of monthly totals is shown for data in panels a-b.

a) Novelty of non-SG StrBio PDB entries in SCOP



b) Novelty of non-SG StrBio PDB entries without Sequence Similarity to Previously Solved Structures

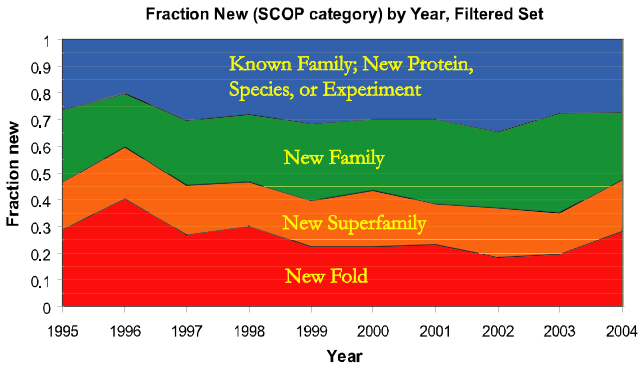
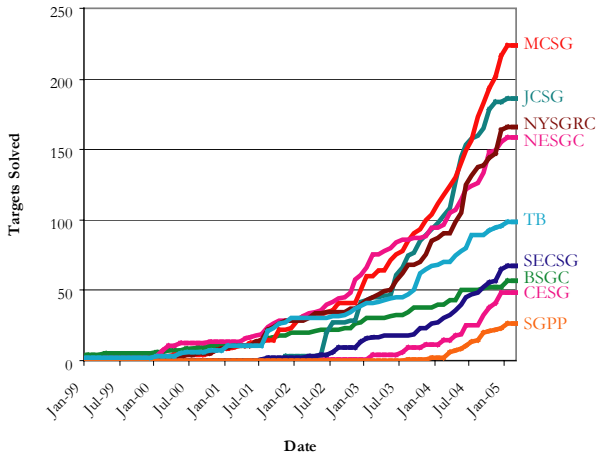


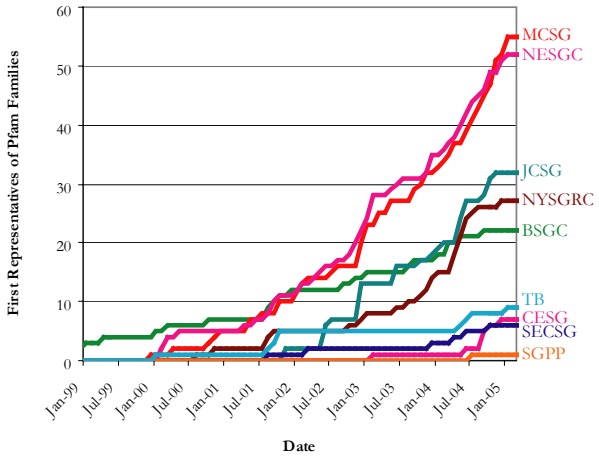
Figure S-3: Projections Based on SCOP

a) Non-SG structural biologists' selection of targets for structure determination. Domains from all PDB entries from 1995-2004 are evaluated as to their level of novelty in SCOP 1.67. PDB entries solved at SG centers were excluded, and only partial data (through 15 May) is available for 2004. The fraction of domains that were the first representatives of their SCOP category at several levels in SCOP (fold, superfamily, family, protein, species) is shown. Domains with identical SCOP classification to previously deposited domains were considered "new experiments." b) Novelty of domains from proteins without sequence similarity to previously solved structures. The same data are shown as in panel a, but filtered to remove all proteins with sequence similarity (by BLAST and PSI-BLAST, as described in the text) to previously solved structures. A summary of data in these panels, including statistics on individual SG centers, is provided in Figure 2b in the primary manuscript.

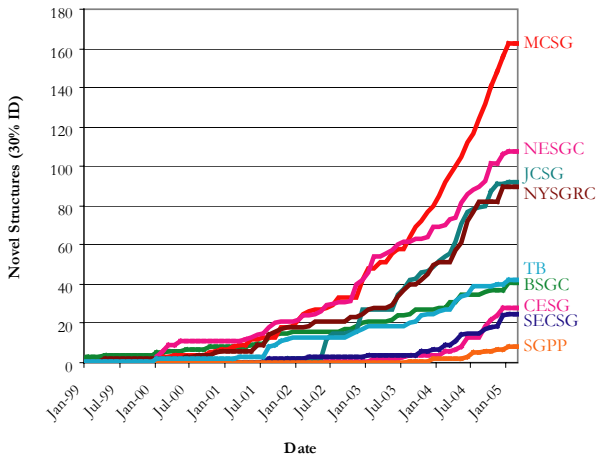
a) Number of Targets Solved at PSI Pilot Centers



b) Number of First Representatives of Pfam families



c) Number of Novel Structures (30% ID)



d) Number of New SCOP Folds or Superfamilies

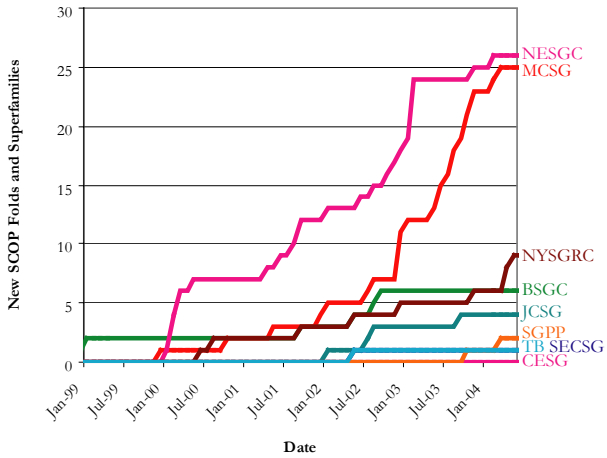


Figure S-4: Time Course of Results for PSI Centers

These plots show the time course for data in Table I in the primary manuscript, for the 9 PSI pilot centers. a) Total number of targets solved. b) Number of first structural representatives of a Pfam family. c) Novel structures at 30% identity. d) New SCOP folds or superfamilies. Note that two centers (CESG and SGPP) officially started a year later than the others.

Table S-I: Structural Genomics Centers Included in this Study.

The list includes all 9 pilot centers funded by the Protein Structure Initiative, as well as the 10 international centers that report results to TargetDB and had solved at least one structure by 1 Feb 2005.

Center	Stated Objective
Protein Structure Initiative (PSI) Centers:	
Berkeley Structural Genomics Center, http://www.strgen.org/ (BSGC)	Structural complement of minimal organisms *Mycoplasma genitalium* and *Mycoplasma pneumoniae.*
Center for Eukaryotic Structural Genomics, http://www.uwstructuralgenomics.org/ (CESG)	Novel eukaryotic proteins, with *A. thaliana* as a model genome.
Joint Center for Structural Genomics, http://www.jcsg.org/ (JCSG)	Structural genomics of *T. maritima* and *C. elegans.*
Midwest Center for Structural Genomics, http://www.mcsg.anl.gov/ (MCSG)	Novel protein folds from all kingdoms. Current targets are chosen from large sequence families of unknown structure.
Northeast Structural Genomics Consortium, http://www.nesg.org/ (NESG)	Novel folds of eukaryotic proteins including *S. cerevisiae, C. elegans, D. melanogaster, Homo sapiens,* or tractable prokaryotic homologs.
New York Structural Genomics Research Consortium, http://www.nysgrc.org/ (NYSGRC)	Novel structural data from all kingdoms of life with emphasis on medically relevant proteins. Current focus on enzymes.
Southeast Collaboratory for Structural Genomics, http://www.secsg.org/ (SECSG)	Structural proteomes of *P. furiosis, H. sapiens, and C. elegans.*
Structural Genomics of Pathogenic Protozoa Consortium, http://www.sgpp.org/ (SGPP)	Structural genomics of protozoan pathogens.
TB Structural Genomics Consortium, http://www.doe-mbi.ucla.edu/TB/ (TB)	Structures of *M. tuberculosis* proteome, with emphasis on functionally important proteins.
Non-PSI International Centers:	
Bacterial Targets at IGS-CNRS, France, http://igs-server.cnrs-mrs.fr/Str_gen/ (BIGS)	Proteins from the bacteria *Rickettsia*, as well as proteins with unique species-specific sequences (ORFans) from *Escherichia coli.*
Montreal-Kingston Bacterial Structural Genomics Initiative, Canada, http://euler.bri.nrc.ca/brimsg/bsgi.html (BSGI)	Novel representatives for protein families.
Israel Structural Proteomics Center, Israel, http://www.weizmann.ac.il/ISPC/ (ISPC)	Proteins related to human health and disease.
Marseilles Structural Genomics Program, France, http://afmb.cnrs-mrs.fr/stgen/ (MSGP)	Structural genomics of bacterial, viral, and human ORFs of known and unknown function.
Oxford Protein Production Facility , U.K., http://www.oppf.ox.ac.uk/ (OPPF)	Targets of biomedical interest: Human proteins, cancer and immune cell proteomes, and Herpes viruses.
Protein Structure Factory, Germany, http://www.proteinstrukturfabrik.de/ (PSF)	Structure of human proteins.
RIKEN Structural Genomics / Proteomics Initiative, Japan, http://www.riken.jp/engn/ (RIKEN)	Structural genomics of *Thermus thermophilus* HB8 and an archaeal hyperthermophile, *Pyrococcus horikoshii* OT3.
Structure 2 Function Project , U.S., http://s2f.carb.nist.gov/ (S2F)	Functional characterization of *Haemophilus influenzae* proteins.
Structural Proteomics in Europe, E.U., http://www.spineurope.org/ (SPINE)	Structures of a set of human proteins implicated in disease states.
Yeast Structural Genomics, France, http://genomics.eu.org/spip/ (YSG)	Structures of *Saccharomyces cerevisiae* proteins.

Table S-II. Novel Structures as Evaluated by Sequence Comparison Methods

This shows the total number of novel structures first structurally characterized by the nine PSI pilot centers, by international Structural Genomics efforts, and by other (non-SG) structural biologists in the last 5 years. Because targets shorter than 50 residues long were not counted here, the NESGC has two fewer targets in this table than in Table I in the primary manuscript, and the SGPP has one fewer.

Center	Targets Solved	Novel Structures at Similarity Criteria				
		PSI-BLAST	Coarse BLAST	Medium BLAST	Fine (30% ID)	Ultrafine (95% ID)
PSI Centers:						
BSGC	57	26	40	40	41	53
CESG	48	8	24	26	28	38
JCSG	186	34	82	84	92	172
MCSG	224	82	148	154	163	215
NESGC	157	67	101	104	108	145
NYSGC	166	24	80	82	90	151
SECSG	67	13	23	24	25	58
SGPP	25	2	8	8	8	24
TB	99	16	36	39	42	94
All PSI Centers (total)	1,029	272	542	561	597	950
Japanese Center (RIKEN)	686	105	250	280	289	494
Other International Centers:						
BIGS	12	3	4	4	7	10
BSGI	40	12	19	19	20	27
ISPC	2	0	0	0	0	0
MSGP	8	1	5	5	6	8
OPPF	3	1	1	1	1	3
PSF	19	4	6	6	6	12
S2F	2	1	1	1	1	1
SPINE	72	10	19	20	21	50
YSG	11	3	7	7	7	9
Total International SG, excluding PSI, RIKEN	169	35	62	63	69	120
Non-SG Structural Biology, since 2000	16,126	1,363	2,269	2,375	2,521	5,362

Table S-III. Completely Novel Structures as Evaluated by Sequence Comparison Methods

This shows the total number of completely novel structures first structurally characterized by the nine PSI pilot centers, by international Structural Genomics efforts, and by other (non- SG) structural biologists in the last 5 years. Like Table S-II, it excludes structures with less than 50 residues.

Center	Targets Solved	Completely Novel Structures at Similarity Criteria				
		PSI-BLAST	Coarse BLAST	Medium BLAST	Fine (30% ID)	Ultrafine (95% ID)
PSI Centers:						
BSGC	57	22	37	39	40	53
CESG	48	8	21	23	26	38
JCSG	186	30	74	81	89	172
MCSG	224	73	142	152	161	215
NESGC	157	64	100	104	108	145
NYSGC	166	18	59	69	77	150
SECSG	67	10	21	22	23	58
SGPP	25	2	7	7	8	24
TB	99	12	27	29	33	94
All PSI Centers (total)	1,029	239	488	526	565	949
Japanese Center (RIKEN)	686	72	192	223	234	478
Other International Centers:						
BIGS	12	2	4	4	5	10
BSGI	40	11	17	18	18	26
ISPC	2	0	0	0	0	0
MSGP	8	1	2	2	4	8
OPPF	3	1	1	1	1	3
PSF	19	3	5	5	5	11
S2F	2	1	1	1	1	1
SPINE	72	5	11	13	14	44
YSG	11	3	5	6	7	9
Total International SG, excluding PSI, RIKEN	169	27	46	50	55	112
Non-SG Structural Biology, since 2000	16,126	1,101	1,824	1,977	2,144	5,164

Table S-IV. Novel Structures Evaluated Using SCOP 1.67

This shows the total number of structures and domains characterized by the nine PSI pilot centers, by international Structural Genomics efforts, and by other (non- SG) structural biologists in the last 5 years. Targets analyzed were those that were released by the PDB prior to the SCOP 1.67 freeze date (15 May 2004). The number of domains in parentheses is the total number of non-redundant domains in these targets.

Center	Targets Solved (Domains)	Novel Domains at SCOP Level					
		Fold	SF	Family	Species	Protein	Exper.
PSI Centers:							
BSGC	29 (33)	4	2	10	8	6	3
CESG	12 (12)	0	0	2	7	2	1
JCSG	51 (61)	3	1	10	27	12	8
MCSG	99 (110)	18	7	37	40	4	4
NESGC	84 (89)	15	11	20	23	11	9
NYSGC	61 (79)	6	3	13	38	8	11
SECSG	21 (22)	0	1	2	7	9	3
SGPP	4 (4)	2	0	0	0	1	1
TB	41 (53)	0	1	3	14	32	3
All PSI Centers (total)	402 (463)	48	26	97	164	85	43
Japanese Center (RIKEN)	172 (222)	10	10	19	64	68	51
Other International SG (total)	60 (72)	6	3	5	30	6	22
Non-SG Structural Biology, since 2000	11,638 (17,654)	269	209	521	1,703	1,458	13,494

Table S-V. Average Cost per Novel Structure

This shows the average cost per structure, novel structure, and novel family by the nine PSI pilot centers, and by other (non-SG) structural biologists. "Any Structure" is the average cost for all structures, including those highly similar to ones already known. The other 3 columns (Novel Structure, 30% ID; New Pfam family; and New SCOP fold or superfamily) are several measures of the average cost per novel structure. Average cost per novel Structural Biology structure is extrapolated from the cost per structure, as described in the Methods section. For PSI centers, the average cost over the lifetime of the center, and the average cost in the most recent 12-month period analyzed are shown. The latter calculation includes structures solved 1 Feb 2004 through 31 Jan 2005 for the first 3 columns, and structures released 16 May 2003 through 15 May 2004 for the SCOP column. "n/a" indicates no structures in a given category were solved.

Center	**Cost (1000s of $) per**				
	Any Structure	**Novel Structure (95% ID)**	**Novel Structure (30% ID)**	**New Pfam family**	**New SCOP fold or SF**
PSI Centers:					
BSGC	440	474	612	1,141	3,239
most recent year	444	472	581	1,889	3,481
CESG	434	548	743	2,974	n/a
most recent year	210	236	315	1,259	n/a
JCSG	135	146	273	784	4,858
most recent year	86	92	189	581	6,963
MCSG	112	117	154	456	777
most recent year	67	68	97	343	410
NESGC	158	173	232	483	747
most recent year	118	128	194	444	870
NYSGC	151	166	279	930	2,159
most recent year	96	99	194	630	1,393
SECSG	375	433	1,004	4,183	19,434
most recent year	189	204	420	2,519	n/a
SGPP	801	867	2,602	20,815	7,574
most recent year	315	343	1,259	7,556	3,481
TB	254	267	598	2,789	19,434
most recent year	244	270	472	1,889	n/a
All PSI Centers (average)	211	229	364	1,030	2,248
most recent year	138	147	249	829	1,790
Non-SG Structural Biology (lower estimate since 2000)	83	**250**	532	1,531	2,024
Non-SG Structural Biology (upper estimate since 2000)	**300**	902	1,919	5,526	7,304

Table S-VI. Size of Structural Genomics Structures

This table shows the average number of non-identical chains, and residues per chain, in structures solved by the nine PSI pilot centers, by international Structural Genomics efforts, and by other (non-SG) structural biologists in the last 5 years. Like Table I in the primary manuscript, this table includes data on structures with fewer than 50 residues.

Center	Average # of Non-identical Chains per Structure	Average # of Residues per Non-identical Chain	Average # of Non-identical Chains per Novel Structure	Average # of Novel Residues per Chain in Novel Structures
PSI Centers:				
BSGC	1.0	219.9	1.0	209.7
CESG	1.0	206.4	1.0	184.3
JCSG	1.01	267.0	1.0	247.6
MCSG	1.02	209.4	1.0	200.4
NESGC	1.0	167.1	1.0	155.0
NYSGC	1.03	271.0	1.0	250.4
SECSG	1.0	208.8	1.0	177.1
SGPP	1.0	220.0	1.0	213.3
TB	1.0	270.7	1.0	216.3
All PSI Centers (average)	1.01	229.9	1.0	207.2
Japanese Center (RIKEN)	1.05	252.9	1.0	190.7
Other International SG	1.08	241.8	1.0	225.4
Non-SG Structural Biology, since 2000	1.40	239.6	1.09	229.3

Table S-VII. Citations for Publications of 20 Randomly Selected Y2 PSI Structures

20 PDB entries were randomly selected from among 104 PDB entries with deposition dates between 1 September 2001 and 31 August 2002 that were mapped to PSI targets. The deposition date and center (abbreviated as per Table I) are given. "Novelty" indicates the level of novelty using the three categories of criteria: Pfam, BLAST/PSI-BLAST, and SCOP. Key: PF = novel Pfam, PB = novel PSI-BLAST, CB = novel by coarse BLAST, MB = novel by medium BLAST, FB = novel by fine BLAST, UFB = novel by ultra-fine BLAST, SFO = new SCOP fold, SSF = new SCOP superfamily, SFA = new SCOP family, SPR = new SCOP protein, SSP = new SCOP species. The year of publication of the primary reference and the number of citations reported for the primary reference in ISI Web of Science on 8 July 2005 are given. (1) - summarizes the accomplishments of the center, not the individual structure. (2) - two structures described in the same paper.

PDB Entry	Deposition Date	Center	Novelty	Year, # of Citations
1J5T	3 Jul 2002	JCSG	-	unpublished
1J5W	5 Jul 2002	JCSG	PF, PB, SPR	unpublished
1K0R	20 Sep 2001	TB	UFB, SSP	2001, 9
1K7K	19 Oct 2001	MCSG	UFB, SPR	unpublished
1KCX	11 Nov 2001	NYSGRC	CB, SPR	2004, 5
1KQ3	3 Jan 2002	JCSG	UFB, SSP	2002, 86(1)
1KUT	22 Jan 2002	MCSG	MB, SPR	unpublished
1KYH	4 Feb 2002	MCSG	PF, CB, SFA	2002, 4
1L7A	14 Mar 2002	MCSG	PF, CB, SFA	unpublished
1L7N	16 Mar 2002	BSGC	CB, SFA	2002, 43(2)
1L7O	16 Mar 2002	BSGC	CB, SFA	2002, 43(2)
1LA2	27 Mar 2002	NYSGRC	-	2002, 8
1LQL	10 May 2002	BSGC	PF, PB, SFO	2003, 0
1LVW	29 May 2002	NESGC	UFB, SSP	unpublished
1LW4	30 May 2002	NYSGRC	CB, SPR	2002, 6
1M1M	19 Jun 2002	TB	UFB, SSP	unpublished
1M1S	20 Jun 2002	NESGC	CB, SPR	unpublished
1M6Y	17 Jul 2002	MCSG	PF, CB, SSF	2003, 2
1M94	26 Jul 2002	NESGC	CB, SPR	2003, 0
1MKM	29 Aug 2002	MCSG	-	2002, 14
Mean number of Citations				11.0
Standard Deviation in Number of Citations				21.3
Median number of Citations				1

Table S-VIII. Citations for Publications of 20 Randomly Selected Novel non-SG structures from the PSI Y2 period

20 PDB entries were randomly selected from among 240 PDB entries with deposition dates between 1 September 2001 and 31 August 2002 that were not mapped to structural genomics targets and were considered novel according to the PSI-BLAST or Pfam criteria. The year of publication of the primary reference and the number of citations reported for the primary reference in ISI Web of Science on 8 July 2005 are given.

PDB Entry	Deposition Date	Year, # of Citations
1GMJ	14 Sep 2001	2001, 24
1H0X	1 Jul 2002	2002, 25
1H2S	15 Aug 2002	2002, 51
1IR6	11 Sep 2001	2002, 8
1JYA	11 Sep 2001	2001, 36
1K30	1 Oct 2001	2001, 8
1K6I	16 Oct 2001	2001, 13
1KHC	29 Nov 2001	2002, 41
1KMI	16 Dec 2001	2002, 31
1KMO	17 Dec 2001	2002, 93
1KWI	29 Jan 2002	2002, 12
1KY9	4 Feb 2002	2002, 71
1L6H	11 Mar 2002	2002, 11
1L6L	11 Mar 2002	2002, 15
1LMZ	2 May 2002	2002, 14
1LN0	2 May 2002	2002, 15
1LPV	8 May 2002	2000, 26
1LSH	17 May 2002	2002, 13
1LVA	28 May 2002	2002, 9
1M98	8 July 2002	2003, 7
Mean number of Citations		26.2
Standard Deviation in Number of Citations		22.3
Median number of Citations		15

Table S-IX. Citations for Publications of 20 Randomly Selected Non-Novel Non-SG Structures from the PSI Y2 period

20 PDB entries were randomly selected from among 2,724 PDB entries with deposition dates between 1 September 2001 and 31 August 2002 that were not mapped to structural genomics targets or considered novel according to the PSI-BLAST or Pfam criteria. The year of publication of the primary reference and the number of citations reported for the primary reference in ISI Web of Science on 8 July 2005 are given.

PDB Entry	Deposition Date	Year, # of Citations
1GSX	9 Jan 2002	2002, 4
1H09	12 Jun 2002	2003, 7
1H0A	12 Jun 2002	2002, 125
1H2I	9 Aug 2002	2002, 27
1ITT	3 Feb 2002	2001, 6
1JXO	7 Sep 2001	2001, 31
1K3D	2 Oct 2001	2001, 14
1K8Y	26 Oct 2001	2002, 8
1KA1	31 Oct 2001	2002, 4
1KEC	15 Nov 2001	2004, 0
1KFP	22 Nov 2001	2002, 13
1KG4	26 Nov 2001	unpublished
1KTG	16 Jan 2002	2002, 15
1KVM	27 Jan 2002	2002, 17
1KZ4	6 Feb 2002	2002, 14
1L2K	21 Feb 2002	2002, 24
1LC2	4 Apr 2002	2003, 1
1LE1	9 Apr 2002	2001, 5
1LMH	1 May 2002	2002, 17
1LNW	3 May 2002	2002, 19
Mean number of Citations		17.6
Standard Deviation in Number of Citations		26.1
Median number of Citations		13.5

Table S-X. Citations for Publications of 20 Randomly Selected Novel Non-SG Structures

20 PDB entries were randomly selected from among 2,131 PDB entries with deposition dates prior to 1 September 2002 that were not mapped to structural genomics targets and were considered novel according to the PSI-BLAST or Pfam criteria. The year of publication of the primary reference and the number of citations reported for the primary reference in ISI Web of Science on 8 July 2005 are given.

PDB Entry	Deposition Date	Year, # of Citations
1AOL	8 Jul 1997	1997, 96
1ATB	20 Mar 1994	1994, 31
1B34	17 Dec 1998	1999, 141
1DML	14 Dec 1999	2000, 44
1EL6	13 Mar 2000	2000, 20
1EMW	20 Mar 2000	2000, 7
1FZR	4 Oct 2000	2001, 38
1GSO	24 May 2002	2002, 68
1H4L	11 May 2001	2001, 44
1HCC	28 Nov 1990	1991, 111
1ID1	2 Apr 2001	2001, 57
1IJA	25 Apr 2001	2001, 31
1JFA	20 Jun 2001	2001, 22
1K0H	19 Sep 2001	2002, 2
1KU3	21 Jan 2002	2002, 99
1KWI	29 Jan 2002	2002, 12
1LGH	20 Mar 1996	1996, 424
1NKL	17 Apr 1997	1997, 102
1RYT	26 Apr 1996	1996, 81
1WJA	13 May 1997	1997, 130
Mean number of Citations		78
Standard Deviation in Number of Citations		89.3
Median number of Citations		50.5

Table S-XI. Citations for Publications of 20 Randomly Selected Non-Novel Non-SG Structures

20 PDB entries were randomly selected from among 17,840 PDB entries with deposition dates prior to 1 September 2002 that were not mapped to structural genomics targets or considered novel according to the PSI-BLAST or Pfam criteria. The year of publication of the primary reference and the number of citations reported for the primary reference in ISI Web of Science on 8 July 2005 are given.

PDB Entry	Deposition Date	Year, # of Citations
193L	1 Sep 1995	1996, 82
1AOG	3 Jul 1997	1996, 25
1CF9	24 Mar 1999	1999, 17
1ELZ	10 Feb 1998	1998, 23
1EQS	6 Apr 2000	1999, 24
1ET1	12 Apr 2000	2000, 37
1EYH	6 May 2000	unpublished
1F2U	29 May 2000	2000, 299
1F4H	7 Jun 2000	2000, 35
1FE7	21 Jul 2000	2000, 7
1FPM	31 Aug 2000	2000, 6
1GW4	4 Jun 1997	1997, 24
1H1H	15 Jul 2002	2002, 3
1J9E	25 May 2001	2002, 2
1KHD	29 Nov 2001	2002, 1
1QO9	7 Nov 1999	2000, 53
1QRJ	14 Jun 1999	1999, 2
2EBO	24 Dec 1998	1999, 79
8ICO	15 Dec 1995	1996, 92
9NSE	13 Jan 1999	2000, 16
Mean number of Citations		41.4
Standard Deviation in Number of Citations		65.2
Median number of Citations		23.5

Table S-XII. Citations for Publications of All Y2 PSI Structures

This table contains the 104 PDB entries that were deposited between 1 September 2001 and 31 August 2002, and mapped to PSI targets. The deposition date and center (abbreviated as per Table I) are given. The year of publication of the primary reference and the number of citations reported for the primary reference in ISI Web of Science on 22 November 2005 are shown in the rightmost column.

PDB Entry	Deposition Date	Center	Year, # of Citations
1GR0	10 Dec 2001	TB	2002, 13
1GTD	14 Jan 2002	NESGC	2002, 4
1H2H	8 Aug 2002	NESGC	2003, 7
1IY9	26 Jul 2002	NESGC	unpublished
1J5P	27 Jun 2002	JCSG	unpublished
1J5R	3 Jul 2002	JCSG	unpublished
1J5S	2 Jul 2002	JCSG	2003, 5
1J5T	3 Jul 2002	JCSG	unpublished
1J5U	3 Jul 2002	JCSG	unpublished
1J5V	5 Jul 2002	JCSG	unpublished
1J5W	5 Jul 2002	JCSG	unpublished
1J5X	5 Jul 2002	JCSG	unpublished
1J5Y	5 Jul 2002	JCSG	unpublished
1J6O	9 Jul 2002	JCSG	unpublished
1J6P	9 Jul 2002	JCSG	unpublished
1J6R	10 Jul 2002	JCSG	unpublished
1J6U	29 Aug 2002	JCSG	2004, 0
1JW2	2 Sep 2001	NESGC	2002, 61
1JW3	2 Sep 2001	NESGC	2002, 61
1JX7	5 Sep 2001	BSGC	2002, 8
1JXC	6 Sep 2001	CESG	2002, 10
1JYH	12 Sep 2001	NYSGRC	2002, 5
1JZT	17 Sep 2001	NYSGRC	unpublished
1K0R	20 Sep 2001	TB	2001, 10
1K3R	3 Oct 2001	MCSG	2003, 11
1K47	5 Oct 2001	NYSGRC	2002, 19
1K4N	8 Oct 2001	MCSG	2003, 0
1K6D	15 Oct 2001	MCSG	2002, 2
1K77	18 Oct 2001	MCSG	2002, 0
1K7J	19 Oct 2001	MCSG	unpublished
1K7K	19 Oct 2001	MCSG	unpublished
1K8F	24 Oct 2001	NYSGRC	unpublished
1KAG	1 Nov 2001	NYSGRC	2002, 8
1KCX	11 Nov 2001	NYSGRC	2004, 9
1KJN	4 Dec 2001	MCSG	unpublished
1KKG	7 Dec 2001	NESGC	2003, 20
1KMJ	16 Dec 2001	NYSGRC	2002, 20
1KMK	16 Dec 2001	NYSGRC	2002, 20
1KP9	30 Dec 2001	TB	2002, 26
1KPG	30 Dec 2001	TB	2002, 26
1KPH	30 Dec 2001	TB	2002, 26
1KPI	30 Dec 2001	TB	2002, 26
1KQ3	3 Jan 2002	JCSG	2002, 107

1KQ4	3 Jan 2002	JCSG	2002, 107
1KR4	8 Jan 2002	MCSG	2004, 1
1KS2	10 Jan 2002	MCSG	2003, 16
1KTN	16 Jan 2002	MCSG	unpublished
1KU9	21 Jan 2002	NYSGRC	2003, 5
1KUT	22 Jan 2002	MCSG	unpublished
1KXJ	31 Jan 2002	MCSG	2002, 4
1KYH	4 Feb 2002	MCSG	2002, 4
1KYT	5 Feb 2002	MCSG	unpublished
1L1E	15 Feb 2002	TB	2002, 26
1L1S	19 Feb 2002	MCSG	2002, 6
1L2F	20 Feb 2002	BSGC	2003, 3
1L6R	13 Mar 2002	MCSG	2004, 9
1L7A	14 Mar 2002	MCSG	unpublished
1L7B	14 Mar 2002	NESGC	unpublished
1L7L	15 Mar 2002	SECSG	2002, 0
1L7M	15 Mar 2002	BSGC	2002, 46
1L7N	16 Mar 2002	BSGC	2002, 46
1L7O	16 Mar 2002	BSGC	2002, 46
1L7P	16 Mar 2002	BSGC	2002, 46
1L7Y	18 Mar 2002	NESGC	2002, 3
1L9G	22 Mar 2002	NYSGRC	unpublished
1LA2	27 Mar 2002	NYSGRC	2002, 8
1LFP	11 Apr 2002	BSGC	2002, 8
1LJ9	19 Apr 2002	MCSG	2003, 15
1LKN	25 Apr 2002	NESGC	unpublished
1LME	1 May 2002	JCSG	2003, 10
1LMI	1 May 2002	TB	2002, 10
1LNZ	4 May 2002	NYSGRC	2002, 15
1LPL	8 May 2002	SECSG	2002, 27
1LQL	10 May 2002	BSGC	2003, 0
1LQT	13 May 2002	TB	2002, 17
1LQU	13 May 2002	TB	2002, 17
1LU4	21 May 2002	TB	2004, 10
1LUR	23 May 2002	NESGC	unpublished
1LV3	24 May 2002	NESGC	2002, 3
1LVW	29 May 2002	NESGC	unpublished
1LW4	30 May 2002	NYSGRC	2002, 7
1LW5	30 May 2002	NYSGRC	2002, 7
1LX7	4 Jun 2002	NYSGRC	2003, 10
1LXJ	5 Jun 2002	NESGC	2003, 4
1LXN	5 Jun 2002	NESGC	2003, 4
1M0S	14 Jun 2002	NESGC	unpublished
1M0T	14 Jun 2002	NYSGRC	2002, 4
1M0W	14 Jun 2002	NYSGRC	2002, 4
1M1M	19 Jun 2002	TB	unpublished
1M1S	20 Jun 2002	NESGC	unpublished
1M33	26 Jun 2002	MCSG	2003, 18
1M3S	28 Jun 2002	MCSG	2004, 0
1M6Y	17 Jul 2002	MCSG	2003, 3
1M94	26 Jul 2002	NESGC	2003, 0

1MGP	15 Aug 2002	BSGC	2003, 8
1MI1	21 Aug 2002	NESGC	2002, 18
1MJF	27 Aug 2002	SECSG	unpublished
1MK4	28 Aug 2002	MCSG	unpublished
1MKF	29 Aug 2002	MCSG	2002, 24
1MKI	29 Aug 2002	MCSG	unpublished
1MKM	29 Aug 2002	MCSG	2002, 16
1MKZ	29 Aug 2002	MCSG	2004, 0
1ML8	30 Aug 2002	MCSG	unpublished
1O0U	30 Aug 2002	JCSG	unpublished
Mean number of Citations			11.0
Standard Deviation in Number of Citations			18.7
Median number of Citations			4

Table S-XIII. Citations for Publications for 104 Non-SG Structures

104 PDB entries were randomly selected (without regard to novelty) from among 2,964 PDB entries with deposition dates between 1 September 2001 and 31 August 2002 that were not mapped to structural genomics targets. The year of publication of the primary reference and the number of citations reported for the primary reference in ISI Web of Science on 22 November 2005 are given.

PDB Entry	Deposition Date	Year, # of Citations
1GNV	10 Oct 2001	unpublished
1GP7	30 Oct 2001	2002, 2
1GQ7	20 Nov 2001	2002, 8
1GR9	15 Dec 2001	unpublished
1GSJ	7 Jan 2002	2002, 17
1GT4	10 Jan 2002	2004, 1
1GTH	15 Jan 2002	2002, 6
1GUI	27 Jan 2002	2002, 26
1GVG	12 Feb 2002	2002, 30
1GWC	14 Mar 2002	2002, 18
1GX6	27 Mar 2002	2002, 76
1GXM	8 Apr 2002	2002, 20
1GZ4	15 May 2002	2002, 10
1H0U	27 Jun 2002	2002, 22
1H2F	8 Aug 2002	2003, 6
1H3D	27 Aug 2002	2004, 2
1IU5	27 Feb 2002	2004, 5
1IV5	14 Mar 2002	2002, 7
1IW4	19 Apr 2002	2002, 5
1IXL	27 Jun 2002	2004, 1
1IXY	9 Jul 2002	2002, 7
1IYA	30 Jul 2002	unpublished
1IYB	5 Aug 2002	2002, 2
1J4B	7 Sep 2001	2001, 7
1JWP	4 Sep 2001	2002, 30
1JWX	5 Sep 2001	2002, 27
1JWZ	5 Sep 2001	2002, 30
1JY6	11 Sep 2001	2002, 13
1JZ3	13 Sep 2001	2001, 24
1JZN	16 Sep 2001	2004, 7
1K07	18 Sep 2001	2003, 24
1K2F	26 Sep 2001	2002, 27
1K3I	3 Oct 2001	2001, 23
1K3N	3 Oct 2001	2001, 11
1K41	5 Oct 2001	2001, 3
1K4K	8 Oct 2001	2002, 9
1K52	9 Oct 2001	2001, 16
1K56	10 Oct 2001	2001, 29
1K5O	11 Oct 2001	2001, 13
1K63	15 Oct 2001	2003, 6
1K8K	24 Oct 2001	2001, 120
1K9D	29 Oct 2001	2004, 4
1K9M	29 Oct 2001	2002, 116

1KDH	13 Nov 2001	2002, 31
1KE9	14 Nov 2001	2001, 50
1KEO	16 Nov 2001	2002, 13
1KEX	18 Nov 2001	2003, 10
1KFR	22 Nov 2001	2002, 0
1KFT	23 Nov 2001	2002, 8
1KGD	26 Nov 2001	2002, 6
1KH2	29 Nov 2001	2002, 3
1KH8	29 Nov 2001	2005, 0
1KH9	29 Nov 2001	2002, 6
1KHF	29 Nov 2001	2002, 24
1KJ4	4 Dec 2001	2002, 20
1KK7	6 Dec 2001	2002, 17
1KK8	6 Dec 2001	2002, 17
1KKO	10 Dec 2001	2002, 13
1KMI	16 Dec 2001	2001, 31
1KMT	17 Dec 2001	2002, 23
1KN4	18 Dec 2001	2002, 2
1KPJ	31 Dec 2001	2001, 242
1KS8	11 Jan 2002	2002, 7
1KSG	13 Jan 2002	2002, 32
1KTC	15 Jan 2002	2002, 21
1KTL	16 Jan 2002	2003, 21
1KTO	17 Jan 2002	unpublished
1KX3	31 Jan 2002	2002, 87
1KY3	2 Feb 2002	2002, 13
1L1L	18 Feb 2002	2002, 51
1L3J	27 Feb 2002	2002, 34
1L3S	1 Mar 2002	2003, 43
1L4G	6 Mar 2002	2002, 3
1L4T	5 Mar 2002	2002, 12
1L5H	6 Mar 2002	2002, 29
1L5O	7 Mar 2002	2002, 3
1L8J	20 Mar 2002	2002, 40
1L9C	22 Mar 2002	2002, 11
1L9F	22 Mar 2002	1999, 53
1L9P	26 Mar 2002	2003, 2
1LBF	3 Apr 2002	2002, 6
1LEV	10 Apr 2002	2003, 5
1LGL	16 Apr 2002	2002, 23
1LQB	9 May 2002	2002, 118
1LQF	10 May 2002	2002, 24
1LR4	14 May 2002	2005, 0
1LTK	20 May 2002	unpublished
1LUD	22 May 2002	2002, 3
1LWF	31 May 2002	2002, 16
1LXM	5 Jun 2002	2002, 17
1LYC	7 Jun 2002	2003, 2
1M0N	13 Jun 2002	2002, 7
1M1P	20 Jun 2002	2002, 16
1M27	21 Jun 2002	2003, 51

1M53	8 Jul 2002	2002, 9
1M6T	17 Jul 2002	2002, 11
1M7S	22 Jul 2002	2003, 8
1M8B	24 Jul 2002	2003, 3
1M8W	26 Jul 2002	2002, 33
1MBY	4 Aug 2002	2002, 21
1MBZ	4 Aug 2002	2002, 10
1MDM	7 Aug 2002	2002, 10
1MEX	8 Aug 2002	unpublished
1MIE	23 Aug 2002	2003, 0
Mean number of Citations		21.0
Standard Deviation in Number of Citations		31.8
Median number of Citations		11.5

Table S-XIV. Final PSI Pilot Phase Report

This shows the total number of targets reported to TargetDB as solved (either Crystal Structure or NMR Structure) by the nine PSI pilot centers at the end of the PSI pilot phase (31 August 2005). Note that two centers (CESG and SGPP) started a year later than the others.

PSI Center	Targets Reported Solved by X-ray Crystallography	Targets Reported Solved by NMR	Total Targets Reported Solved
BSGC	58	3	**61**
CESG	43	19	**62**
JCSG	221	8	**229**
MCSG	291	0	**291**
NESGC	116	93	**209**
NYSGC	195	0	**195**
SECSG	75	2	**77**
SGPP	39	0	**39**
TB	104	2	**106**
All PSI Centers (total)	**1142**	**127**	**1269**