The Impact of Structural Genomics: Expectations and Outcomes

John-Marc Chandonia, *et al.*

Science **311**, 347 (2006);
DOI: 10.1126/science.1121018

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

The Impact of Structural Genomics: Expectations and Outcomes

John-Marc Chandonia and Steven E. Brenner*

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 we contrast these results with traditional structural biology. The first structure identified in a protein family enables inference of the fold and of ancient relationships to other proteins; in the year ending 31 January 2005, about 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 SG center in the United States has dropped to one-quarter of the estimated cost of solving a structure by traditional methods. However, the efficiency of the top structural biology laboratories—even though they work on very challenging structures—is comparable to that of SG centers; moreover, traditional structural biology papers are cited significantly more often, suggesting greater current impact.

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 nine 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 about $60 million annually. Considerable resources have also been spent internationally, with SG projects in Japan, Canada, Israel, and Europe under way since the late 1990s. With more than 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 toward 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 a way to estimate its difficulty. Over time, a structure's impact might be crudely evaluated by the number of subsequently published papers that cite the original work.

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 examined several sequence- and structure-based definitions of a protein family so as to reduce the potential for bias introduced by use of any single standard and to directly compare current results with expectations at the outset of the project (*7*). We contrasted 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 examined 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, and that it will 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 in (*8*). Although we focus on PSI centers, we analyze the output of all SG centers that report their results to TargetDB (*9*); these centers and the specific goals of each are listed in table S1.

Berkeley Structural Genomics Center, Physical Biosciences Division, Lawrence Berkeley National Laboratory, and Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA.

*To whom correspondence should be addressed. E-mail: brenner@compbio.berkeley.edu

Impact of Structural Genomics on Coverage of Protein Families

The Pfam database (*10*) is a manually curated database of protein families from sequenced genomes. As of 1 February 2005, 36% of Pfam families (2736 of 7677) (*10*) 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 (*11*)], and we used the database deposition dates to identify the earliest structural representative from each family. 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 (Fig. 1B), even as the total number of structures solved continues to increase (Fig. 1A). Surprisingly, in recent years, the rate of solution of first structures in a Pfam family by non-SG structural biologists has decreased while SG centers have made up the deficit. SG centers worldwide now account for about half of new structurally characterized families, even though they contribute only about 20% of the new structures. PSI centers account for about two-thirds of the worldwide SG contribution. Only 5% of non-SG structures reported since 2000 represent a new Pfam family, whereas the PSI average was 20.4%.

We analyzed the individual contributions of each of the nine U.S. pilot centers and compared them to other SG and structural biology efforts (Table 1). Results vary widely for the nine PSI centers. The MCSG was the most productive, as measured by the total number of structures solved and the total number of new families; 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 because of 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 used the local sequence comparison methods BLAST (*12*) and PSI-BLAST (*13*), at several different levels of sequence similarity, to examine the number of structures that could not be matched to any prior solved structure. Results are shown in Fig. 2A and Table 1.

The overall fraction of structures that were classified as novel according to PSI-BLAST has decreased in the past 15 years, from about

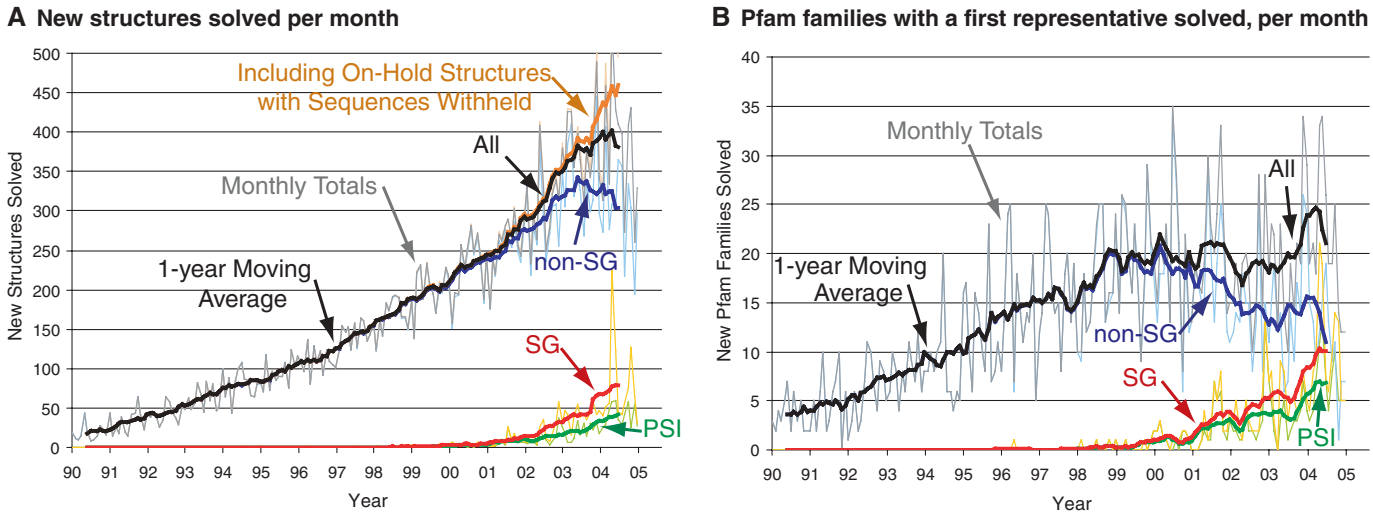

Fig. 1. Structural characterization of new families. (**A**) Black lines indicate the total number of new structures reported per month. Blue lines are contributions from non-SG structural biologists, red lines from SG centers, and green lines from the PSI centers. The orange line indicates structures 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 (A). Monthly totals and a 1-year moving average are shown.

Table 1. Novel structures solved by structural genomics centers and leading structural biology groups (see also fig. S4 and table S14). Shown are the total numbers of novel structures and nonidentical 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 nonidentical 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. Numbers of new SCOP folds and superfamilies are the numbers 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 nonidentical 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 (5757)	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 (4875)	92	3	4
Midwest Center for Structural Genomics (MCSG)	224 (229 chains)	55 (5512)	163	18	25
Northeast Structural Genomics Consortium (NESGC)	159 (159 chains)	52 (4811)	108	15	26
New York Structural Genomics Research Consortium (NYSGRC)	166 (171 chains)	27 (3982)	90	6	9
Southeast Collaboratory for Structural Genomics (SECSG)	67 (67 chains)	6 (1079)	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 (3938)	42	0	1
PSI centers (total of 9 centers above)	1032 (1043 chains)	211 (30,360)	597	48	74
Japanese center (RIKEN)	686 (718 chains)	50 (6860)	289	10	20
Other international SG (total, excluding all centers above)	169 (183 chains)	33 (5877)	69	6	9
Non-SG groups (since 2000)					
Non-SG structural biology (total)	17,096 (23,747 chains)	928 (249,171)	2,521	269	478
Steitz group	46 (559 chains)	23 (4190)	31	7	12
Huber group	185 (273 chains)	8 (679)	38	5	10
Iwata group	14 (54 chains)	14 (7960)	20	2	3

20% in 1990 to 10% today (fig. S1). SG structures account for 44% of the total number of novel structures reported in the year ending 31 January 2005, 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 Fig. 2A), more than 50% of the structures solved had 30% or more

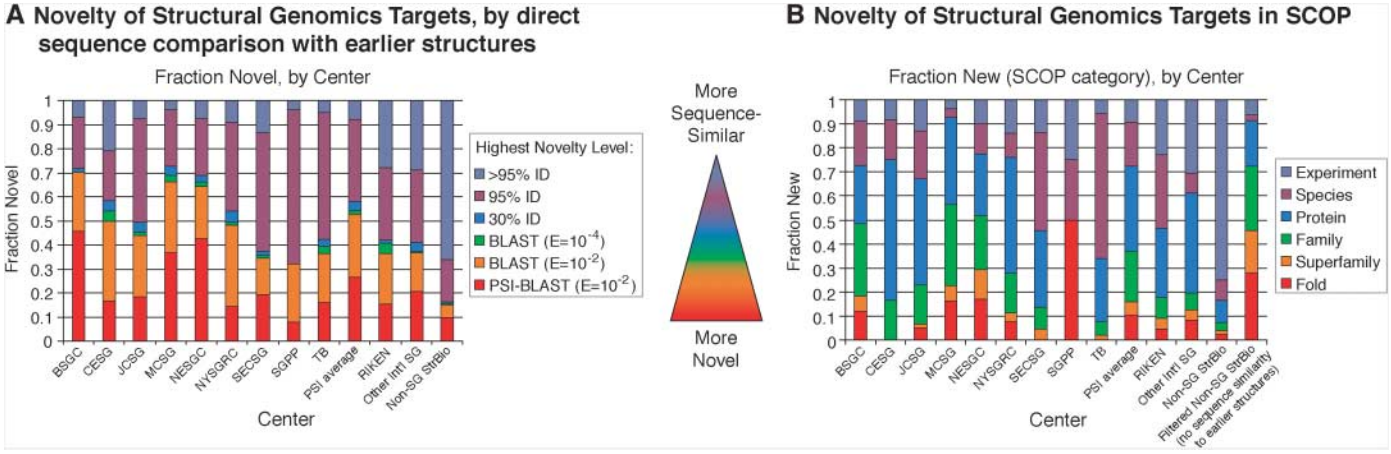

Fig. 2. Novelty rates by center. (**A**) Fractions of structures from each SG center and from non-SG structural biologists that were classified as novel according to each similarity criterion examined. Each structure was classified at the most stringent novelty threshold attained. For example, structures 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. Non-SG StrBio includes all domains solved by non-SG structural biologists (1972 to 2005). Filtered non-SG StrBio includes only domains from non-SG structural biologists 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 (A) includes data on all structures reported through the end of January 2005, whereas (B) only includes those structures released by the PDB before the cutoff date for inclusion in SCOP 1.67 (15 May 2004).

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 (*8*).

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 Structural Classification of Proteins (SCOP) database (*14*). SCOP provides a widely used, manually curated hierarchy indicating different levels of structural and evolutionary relationship between protein domains. Domains classified together at the "family" level have a clear common evolutionary origin, and in many cases they are sufficiently similar to allow reasonably accurate comparative models to be constructed for any family member by 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 with the use of 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 (table S4). For non-SG structures, more than 70% of protein domains solved in the past 10 years represent a new experiment on a protein already structurally characterized, although possibly with mutations, with 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 (fig. S3). 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*, *15*), SG centers might increase the percentage of new SCOP folds and superfamilies discovered to about 40%. Projections based on 2004 data (fig. S3) 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 before 15 May 2004). Results are shown in Fig. 2B and Table 1. 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 year ending 31 January 2005, too late for analysis by this method. For example, the analysis of SGPP data represents only 4 of 25 targets solved, although two of these four 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 National Institute of General Medical Sciences Advisory Council in 2003, the average cost of solving a protein structure under an R01 grant was estimated as $250,000 to $300,000 (*16*, *17*). Because the methodology behind the estimate was not published, we extrapolated 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 (*8*). 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 with

traditional methods. In the last year of our study (1 February 2004 to 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 of our study. However, structures solved by SG centers are on average smaller and contain fewer nonidentical polypeptide chains than those from traditional structural biology (table S6). When normalized to account for both of these factors, per-residue costs for SG in the year ending 31 January 2005 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 toward structures with sequences 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 with 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 (table S5). The most efficient center, MCSG, was more cost-efficient than traditional labs in each category in the most recent year of the project by a factor of 5 to 17 (or, when normalized for structure size, a factor of 4 to 14).

These cost data should be interpreted with great caution because many factors are 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 before 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. Although some technology is already widely used throughout the field (*18*), 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 nuclear magnetic resonance 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 1 results for several individual structural biologists who have been among the 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 (*19*), DNA primase (*20*), and light-harvesting complexes (*21*, *22*). So Iwata is a leader in membrane crystallography and recently solved the structure of the photosystem II complex (*23*). The total output of each of their laboratories is comparable to that of 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 nonidentical polypeptide chains per entry) than did SG centers, whereas the Huber and Iwata groups solved somewhat larger complexes composed of larger individual subunits. We caution that our metrics may be biased toward heteromeric complexes.

We calculated the average cost per novel structure solved by Steitz's laboratory, which operates on a total budget of about $1.5 million per year (*24*), versus about $5.7 million for the average PSI center. Since January 2000, the average cost per structure is about $166,000, but only $14,000 per nonidentical 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 six proteins with novel folds. Furthermore, our protein-based metrics underestimate the novelty of structures solved by the Steitz lab because of 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 on average 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 (table S12). 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 (table S13). 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 two-tailed Mann-Whitney test (*25*)]. For SG structures, novelty did not appear to correlate with the citation rate (*8*). 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 to January 2005). Despite the pace of SG, the quality of SG structures has been found to be similar to that of non-SG structures (*26*). The difference in output between the most efficient center and the average is striking.

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, because finishing a near-complete structure may be a worthwhile use of resources. Finally, a structure may not be considered novel because the preceding struc-

ture 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 to 8% (the minimal level of redundancy observed across all SG centers), saving millions of dollars per year in the United States alone.

Relative 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 finding 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 (*27*), such as public websites and centralized databases (*9*), 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 the 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 because of the substantial cost savings. These families should be prioritized according to significance, for example, family size or biological role (*28*,*29*). Non-SG structural biology could focus on hypothesis-driven research into the function or mechanism of individual proteins, the characterization of particularly challenging proteins and complexes, and other research that is currently impractical to conduct using high-throughput methods. 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 (*30*). During PSI phase 2, considerable resources will be spent on specialized centers aimed at developing 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 protein families; by structurally characterizing proteins whose importance 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 significance is already appreciated, delving deep into particularly rewarding areas to provide immediate scientific impact.

References and Notes

1. S. E. Brenner, *Nat. Rev. Genet.* **2**, 801 (2001).
2. N. Ban, P. Nissen, J. Hansen, P. B. Moore, T. A. Steitz, *Science* **289**, 905 (2000).
3. B. T. Wimberly *et al.*, *Nature* **407**, 327 (2000).
4. D. Baker, A. Sali, *Science* **294**, 93 (2001).
5. S. E. Brenner, C. Chothia, T. J. Hubbard, A. G. Murzin, *Methods Enzymol.* **266**, 635 (1996).
6. P. Smaglik, *Nature* **403**, 691 (2000).
7. S. E. Brenner, M. Levitt, *Protein Sci.* **9**, 197 (2000).
8. See supporting material on *Science* Online.
9. L. Chen, R. Oughtred, H. M. Berman, J. Westbrook, *Bioinformatics* **20**, 2860 (2004).
10. A. Bateman *et al.*, *Nucleic Acids Res.* **32**, D138 (2004).
11. H. M. Berman *et al.*, *Nucleic Acids Res.* **28**, 235 (2000).
12. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
13. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
14. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, *J. Mol. Biol.* **247**, 536 (1995).
15. S. E. Brenner, *Nat. Struct. Biol.* **7** (suppl.), 967 (2000).
16. R. Service, *Science* **307**, 1554 (2005).
17. E. Lattman, *Proteins* **54**, 611 (2004).
18. R. C. Stevens, *Nat. Struct. Mol. Biol.* **11**, 293 (2004).
19. J. Lowe *et al.*, *Science* **268**, 533 (1995).
20. M. A. Augustin, R. Huber, J. T. Kaiser, *Nat. Struct. Biol.* **8**, 57 (2001).
21. J. Deisenhofer, O. Epp, K. Miki, R. Huber, H. Michel, *J. Mol. Biol.* **180**, 385 (1984).
22. R. Huber, *EMBO J.* **8**, 2125 (1989).
23. K. N. Ferreira, T. M. Iverson, K. Maghlaoui, J. Barber, S. Iwata, *Science* **303**, 1831 (2004); published online 5 February 2004 (10.1126/science.1093087).
24. T. Steitz, personal communication.
25. B. L. van der Waerden, *Mathematical Statistics*, vol. 156 of *Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen mit besonderer Berücksichtigung der Anwendungsgebiete* (Springer-Verlag, Berlin, 1969).
26. A. E. Todd, R. L. Marsden, J. M. Thornton, C. A. Orengo, *J. Mol. Biol.* **348**, 1235 (2005).
27. A. Wlodawer, *Nat. Struct. Mol. Biol.* **12**, 634 (2005).
28. J. M. Chandonia, S. E. Brenner, *Proteins* **58**, 166 (2005).
29. J. M. Chandonia, S. E. Brenner, in Proceedings of the 27th International Conference of the IEEE Engineering in Medicine and Biology Society, Shanghai, 1 to 4 September 2005.
30. S. C. Harrison, *Nat. Struct. Mol. Biol.* **11**, 12 (2004).

31. We thank J. Rine, T. Alber, T. Steitz, A. Edwards, and G. Montelione for helpful comments. Supported by NIH grants 1-P50-GM62412, 1-K22-HG00056, and 1-R01-GM073109; the Searle Scholars Program (01-L-116); a Sloan Research Fellowship; the IBM Shared University Research Program; and the U.S. Department of Energy under contract DE-AC02-05CH11231.

Supporting Online Material

www.sciencemag.org/cgi/content/full/311/5759/347/DC1
Materials and Methods
References
Figs. S1 to S4
Tables S1 to S14

10.1126/science.1121018