Published in final edited form as:

Curr Opin Struct Biol. 2007 June ; 17(3): 347–353. doi:10.1016/j.sbi.2007.06.003.

Structural genomics: keeping up with expanding knowledge of the protein universe

Marek Grabowski¹, Andrzej Joachimiak², Zbyszek Otwinowski³, and Wladek Minor¹

¹Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA 22903, USA

²Structural Biology Center, Biosciences Division, Argonne National Laboratory, Argonne, IL 60439, USA

³Department of Biochemistry, UT Southwestern Medical Center at Dallas, Dallas, TX 75235, USA

Abstract

Structural characterization of the protein universe is the main mission of Structural Genomics (SG) programs. However, progress in gene sequencing technology, set in motion in the 1990s, has resulted in rapid expansion of protein sequence space — a twelvefold increase in the past seven years. For the SG field, this creates new challenges and necessitates a reassessment of its strategies.

Nevertheless, despite the growth of sequence space, at present nearly half of the content of the Swiss-Prot database and over 40% of Pfam protein families can be structurally modeled based on structures determined so far, with SG projects making an increasingly significant contribution. The SG contribution of new Pfam structures nearly doubled from 27.2% in 2003 to 51.6% in 2006.

Introduction

The initial long-term goal of the Structural Genomics (SG) endeavor was to map all protein folds, so that the structures of virtually all proteins could be either found in the Protein Data Bank (PDB) or derived by computational methods. The National Institutes of Health (NIH)-funded Protein Structure Initiative (PSI) stated its mission as “making the three-dimensional atomic-level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences” (see mission statement at <http://www.nigms.nih.gov/Initiatives/PSI/>). Similar goals were declared by other worldwide SG programs, such as RIKEN in Japan and SPINE in Europe. In 2000, the known protein universe, defined as the space of all known and predicted protein sequences (open reading frames, ORFs), contained about 300 000 sequences, as measured by the number of entries in the Swiss-Prot and TrEMBL [1] protein sequence databases. It was estimated that, to cover this space with experimental 3D structures, it would be necessary to solve 16 000 carefully selected targets [2]. Through subsequent homology modeling, this set of experimental structures would be sufficient to provide structural models for 90% of sequence space. Based on the projected exponential growth of the PDB [3], it appeared feasible that near-complete structural mapping could be achieved in a decade [2,4••].

In this review, we confront these predictions with what has happened since the start of the SG programs. We examine the coverage of the content of the Swiss-Prot database, an important sample of the protein universe containing most of the proteins studied so far in biomedical

research. We also discuss SG strategies in dealing with the growth in the number of known proteins and protein families during the past seven years.

Rapid expansion of the known protein universe

Today, from the perspective of the past seven years of SG programs (in one of which we have been fortunate to participate), we can look back at the issue of protein structure coverage completeness with the benefit of hind-sight. Over the same time-frame, the size of the known protein sequence universe has been rapidly expanding, set in motion by the ‘Big Bang’ of genome sequencing projects. The content of Swiss-Prot plus TrEMBL grew 12-fold (to 3.8 million sequences in January 2007) and this growth will continue, nourished by the expansion of low-cost sequencing projects, such as Global Ocean Sampling (GOS) [5] and other metagenomics efforts, recovering genomes from environmental samples for organisms that can not be easily cultured in the laboratory. In each newly sequenced genome, the majority of the predicted proteins belong to already established protein families. However, there are no signs of saturation of sequence space: each new sequenced genome adds many novel sequences and new families, including singletons (single-member families; ORFs with no sequence homologue in known genomes) [6•]. It has been estimated that the total number of protein families in 1000 bacterial genomes (currently there are 759 in the NCBI database) will be 250 000 [4•], with the vast majority of families being small or singletons [7].

Growth of the Protein Data Bank

The number of known structures, as measured by the number of deposits in the PDB, has also been growing very fast (Figure 1). In January 2007, the PDB contained almost 35 000 protein structures determined by X-ray crystallography and over 5000 NMR deposits (including 2607 X-ray and 1109 NMR deposits from SG programs). On a linear scale, new deposits are added at an increasing rate. However, despite enormous progress in protein production, crystallization and structure determination methodologies and technologies [8,9•,10,11], the growth of the PDB has been moderate in recent years compared to the previous burst [12•]. For the period 2000–2006, the rate of growth of the number of unique structures has been in the range 18–22% per year (unique structures are defined as distinct at the level of 30% sequence identity). Additionally, increased knowledge of protein structures has shown that the range of potential protein folds is more like a continuum rather than a series of discrete, well-defined categories. For this reason, there is no agreement on the total number of protein folds existing in nature [6•]. Current estimates range from a few thousand [13•] to tens of thousands [14], depending on the methodology adopted.

Structural coverage of the Swiss-Prot database

Has the growth in the number of determined structures kept pace with the expansion of known sequence space? We specifically analyzed the structural coverage of the Swiss-Prot database, which contains only curated and manually annotated protein sequences. It includes information about most of the proteins being studied in detail by biological and biomedical laboratories around the world. In January 2000, it contained 77 109 sequences. The families of these sequences were largely covered by the set of SG targets selected during the pilot phase of PSI [15,16]. At that time, the PDB contained 7707 non-redundant protein sequences (11 568 protein deposits), enabling the prediction of 3D structures for 28.0% of Swiss-Prot entries by homology modeling. This number is based on the assumption that a good-quality 3D model can be obtained for protein sequences that can be aligned with more than 30% sequence identity to at least one PDB deposit.

In January 2007, Swiss-Prot contained 250 259 protein sequences from nearly 11 000 species. Close to half of these (48.4%) had more than 30% sequence identity to proteins in the PDB.

Thus, measured by the percentage of sequences that could be modeled from known protein structures, the structural coverage of Swiss-Prot has almost doubled in the past seven years (Figure 2). If no new structures were deposited in the PDB after January 2000, this rate of structural coverage would drop to 23.9%. Thus, the additional modeling gain produced by structures deposited after January 2000 has been about 60 000 sequences. A significant fraction of this additional modeling coverage resulted from structures determined by SG projects worldwide. More than 39 000 sequences (15.6% of all Swiss-Prot proteins) had a match above 30% sequence identity to an SG deposit. From these, for over 15 000 sequences (6.0% of Swiss-Prot), the SG deposits either were exclusive matches or were deposited before any other matching PDB structures determined using 'traditional structural biology', outside SG centers. About 8200 sequences (3.3% of Swiss-Prot) were exclusively within the modeling leverage of SG deposits and could not be modeled from 'traditional' structure deposits. This translates to a ratio of 10.7 models of Swiss-Prot proteins per average SG deposit versus 3.1 models per average 'traditional' deposit. Moreover, the contribution of SG structures in extending modeling coverage is increasing with time: in 2002, SG deposits provided 17.6% of added modeling leverage and in 2005 it nearly doubled to 32.6%.

Looking at the set of 77 109 proteins that constituted the Swiss-Prot database in January 2000, we find that seven years later 43.6% of those protein sequences are within current modeling leverage of the PDB (10.4% are within modeling leverage of SG deposits). Surprisingly, coverage for this set is not as good as for subsequent additions to Swiss-Prot, despite the fact that these proteins have been available for study for a much longer time. These 'overlooked' proteins are probably those for which protein production and/or crystallization and structure determination has presented considerable difficulties.

Re-assessment of structural genomics strategies

Our analysis has been performed for Swiss-Prot only, which at present constitutes about 6% of all sequences in the combined Swiss-Prot plus TrEMBL database. If the current average of 10.7 models per SG deposit is extrapolated to that larger set, it becomes apparent that achieving near-complete coverage is not realistic with current technology. A comprehensive analysis of a set of more than 630 000 sequences from 203 completely sequenced genomes [17••] showed that fine-grained (at the 30% sequence identity level) coverage of 70% of that space would require more than 90 000 structures. On the other hand, for coarse-grained (at any detectable sequence homology level) coverage of the same percentage of sequences, it would be sufficient to determine structures for the largest 2000 protein families constructed from this set (one structure per family). Faced with the growth of sequence space and its fragmentation into a large number of small families, target selection in the second phase of PSI focuses mainly on structural coverage of large protein families (>10 members). Because of the change in strategy, SG modeling leverage is increasing (e.g. at the beginning of 2007, each chain of each structure deposited by the Midwest Center for Structural Genomics consortium generated, on average, the ability to model 135 new sequences). Moreover, there are ongoing efforts to improve homology modeling, which are also likely to increase modeling leverage by lowering the sequence identity level at which reasonable models can be built.

It may be argued that the number of unrelated protein sequence families is a more appropriate measure of the size of the known protein universe. The constant increase in the number of new protein families with the addition of newly sequenced genomes indicates that, so far, we have only explored a very tiny portion of the protein universe (with millions of organisms still waiting to be sequenced) [17••]. At present, the only way to structurally classify new families is through the experimental determination of 3D structures of one or more representatives of that family. The number of new structural families is also increasing with the addition of new genomes, but at a significantly lower rate. As the number of possible protein folds in nature

appears to be much smaller than the number of protein families classified by sequence similarity, the coarse-grained experimental sampling of the biggest families should maximize the combined experimental and modeling coverage of the protein universe.

Structural coverage of the Pfam classification of proteins

Targets for the second phase of PSI have been primarily selected from Pfam families [18] that have not been previously characterized structurally. Pfam, a manually curated database of families of protein motifs and domains, is skewed towards big families and contains annotations providing an approximate functional classification. Pfam grew from 2008 families in release 5 in January 2000 to 8957 families in release 21 in November 2006. Currently, about 2300 families belong to the 'Domain of unknown function' or 'Uncharacterized protein' category. About 74% of all proteins in Swiss-Prot and TrEMBL could be assigned to one or more Pfam families (<http://www.sanger.ac.uk/Software/Pfam/>).

We have examined the structural coverage of Pfam using family assignment based on hidden Markov models [19]. In January 2000, there were 898 families with PDB representatives (out of 1903 families that remained in Pfam release 21), corresponding to 46.8% structural coverage. Between 2000 and 2007, the number of Pfam families increased to 8957 and those with structural representatives quadrupled (to 3815 families in release 21) (Figure 3a). The growth of classified protein families happened at roughly the same rate that their novel structural representatives were found, with overall structural Pfam coverage now at the level of 42.6%. SG contributed the first structural representatives for 650 Pfam classifications (about one-fourth of families for which structural representatives were solved after January 2000). The rate of determination of new Pfam representatives has been fluctuating (Figure 3b), but overall the SG contribution has grown: in 2003 it was 27.2% and in 2006 it nearly doubled to 51.6%, following the trend observed by Chandonia and Brenner [20•,21•]. Looking back at the initial Pfam data as of January 2000, we find that the subsequent growth of the PDB increased the coverage of this set by January 2007 from 898 to 1429 families (i.e. from 46.8% to 75.1%).

Will it be possible to predict a structure for any protein with known sequence using structure databases and modeling tools? This question is not easy to answer, as inherent disorder seems to be a characteristic feature of a large fraction of eukaryotic proteins and this issue has so far been barely studied. Thus, the issue of predicting structures is not precisely defined for uncharacterized proteins. However, when proteins are evolutionarily conserved, it is reasonable to expect that they also have conserved structures, making protein families the natural target of SG. Even when targets are carefully chosen to maximize structural coverage, there is no guarantee that it would be possible to determine the structure of all of these targets. Structure determination by X-ray crystallography and NMR is a multistage process (cloning, expression, purification, crystallization, diffraction and structure determination), with many experimental difficulties for various classes of proteins. Even assuming an extremely low attrition of 10% at each stage, the resulting theoretical overall success ratio would be only about 53%. In practice, the most efficient SG centers have an average success rate on the order of 10 or less deposits per 100 clones [22•]. As we have revealed, 56% of the original Swiss-Prot entries from January 2000 remain structurally uncharacterized even after seven years.

Applications of structural genomics

Aside from satisfying the fundamental quest for knowledge, why would we really want the complete structural characterization of the entire protein universe? Perhaps this issue can be considered in analogy to the study of our physical universe. There are some fundamental physical constraints that prevent us from exploring our entire universe, but through careful studies we have a good understanding of its properties. Similarly, there are some fundamental limitations that prevent us from determining, experimentally or computationally, all protein

structures. However, it seems that careful structural studies can provide sufficient data to understand the properties of the protein universe. Already, some attempts have been made to generalize protein folds and their relationships [23,24••].

Several fields of science have already directly benefited from SG results. These include bioinformatics and computational biology. In addition, high-throughput methods and technologies in molecular biology and structural biology are currently being applied in many academic and commercial laboratories outside of SG programs. 325 SG structures have already been used as starting models in molecular replacement. Finally, in the biomedical sciences, structure-function relationships and structural explanations of biological function in numerous cases have either been a direct result of SG programs or come from projects in which structural studies followed an SG result.

Moreover, macromolecular crystallography has become a standard technique used by many pharmaceutical and biotechnology companies for structure-based drug design [25]. This process has accelerated recently because of progress in structural bioinformatics and computational biochemistry resulting from SG efforts.

The intent of high-throughput approaches in SG is to provide a starting point for structure-function studies, rather than to replace them. SG results are most important when they are followed by detailed studies in a traditional structural biology environment, in both academic and industrial settings. Therefore, SG efforts may need to ensure as high as possible completeness of structural coverage for targets of biomedical importance, not only in the sense of covering sequence space, but also in terms of adequate structural quality and the level of detail. This could be achieved, for example, by a fine granularity strategy for biomedical targets and by imposing very high quality requirements for SG deposits.

Conclusions

Despite the rapid growth in the number of known protein sequences in the period 2000–2007, structural coverage of Swiss-Prot is actually increasing. It appears that, because of the latest endeavors, structural biology with a strong contribution from SG programs has been able to keep pace with the rapid expansion in the number of Pfam classifications. Therefore, SG plays an important role in determining structures of novel proteins and extending the structural coverage of the protein universe. In recent years, SG has been responsible for about a third of the new structural coverage of Swiss-Prot and nearly half of first structural representatives in Pfam. Moreover, it appears that SG is on track to further improve the coverage of protein families. Progress for some regions of sequence space that are significantly more challenging and difficult to characterize structurally will probably require new methodologies and approaches. Several SG programs have already invested considerable effort to develop and implement new methodologies and technologies.

World-wide SG efforts have determined 3716 protein structures; 70% were solved using X-ray crystallography. It is clear that macromolecular X-ray crystallography will continue to be the most important experimental technique to achieve the goals of SG and expand the coverage of the most challenging protein families.

Taking into account the importance of precise structural information for applications in biomedicine and the pharmaceutical industry, the quality of protein structures has to be as high as possible. In the pilot stage, the quality of high-throughput structures was, on average, similar to that produced by traditional approaches. Currently, SG structures exceed the average quality of PDB structures, which, in the future, will better satisfy the requirements of the consumers of structural information.

Acknowledgments

We would like to thank Ian Wilson, Alex Wlodawer, Adam Godzik and Matt Zimmerman for helpful discussions. This work has been supported, in part, by NIH grant GM074942. The authors are members of the Midwest Center for Structural Genomics consortium.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- • of outstanding interest

1. O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief Bioinform* 2002;3:275–284. [PubMed: 12230036]
2. Vitkup D, Melamid E, Moulton J, Sander C. Completeness in structural genomics. *Nat Struct Biol* 2001;8:559–566. [PubMed: 11373627]
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–242. [PubMed: 10592235]
4. Yan Y, Moulton J. Protein family clustering for structural genomics. *J Mol Biol* 2005;353:744–759. [PubMed: 16185712] The authors re-examine target selection strategies for SG in light of the growth in the number of known proteins and protein families, revising earlier estimates from [2]
5. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, et al. The Sorcerer II Global Ocean Sampling Expedition: expanding the universe of protein families. *PLoS Biol* 2007;5:e16. [PubMed: 17355171]
6. Orengo CA, Thornton JM. Protein families and their evolution—a structural perspective. *Annu Rev Biochem* 2005;74:867–900. [PubMed: 15954844] A comprehensive review
7. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. *Nature* 2002;420:218–223. [PubMed: 12432406]
8. Stevens RC. Long live structural biology. *Nat Struct Mol Biol* 2004;11:293–295. [PubMed: 15048098]
9. Minor W, Cymborowski M, Otwinowski Z, Chruszcz M. HKL-3000: the integration of data reduction and structure solution—from diffraction images to an initial model in minutes. *Acta Crystallogr D Biol Crystallogr* 2006;62:859–866. [PubMed: 16855301] This paper presents a new approach that significantly accelerates the process of structure determination
10. Dauter Z. Current state and prospects of macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 2006;62:1–11. [PubMed: 16369088]
11. Rosenbaum G, Alkire RW, Evans G, Rotella FJ, Lazarski K, Zhang RG, Ginell SL, Duke N, Naday I, Lazarz J, et al. The structural biology center 19ID undulator beamline: facility specifications and protein crystallographic results. *J Synchrotron Radiat* 2006;13:30–45. [PubMed: 16371706]
12. Levitt M. Growth of novel protein structural data. *Proc Natl Acad Sci USA* 2007;104:3183–3188. [PubMed: 17360626] A new method to determine novel structures is introduced and used to analyze the historical growth of the PDB. The main conclusion of the analysis is that the rate of growth of structural data in the PDB has slowed in recent years, in comparison with earlier periods of exponential expansion
13. Grant A, Lee D, Orengo C. Progress towards mapping the universe of protein folds. *Genome Biol* 2004;5:107. [PubMed: 15128436] This study discusses various ways of estimating the total number of folds in nature and prospects for the determination of representatives of major protein families by SG
14. Coulson AF, Moulton J. A unfold, mesofold, and superfold model of protein fold use. *Proteins* 2002;46:61–71. [PubMed: 11746703]
15. Chen L, Oughtred R, Berman HM, Westbrook J. TargetDB: a target registration database for structural genomics projects. *Bioinformatics* 2004;20:2860–2862. [PubMed: 15130928]

16. O'Toole N, Raymond S, Cygler M. Coverage of protein sequence space by current structural genomics targets. *J Struct Funct Genomics* 2003;4:47–55. [PubMed: 14649288]
17. Marsden RL, Lee D, Maibaum M, Yeats C, Orengo CA. Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space. *Nucleic Acids Res* 2006;34:1066–1080. [PubMed: 16481312] The analysis presented in this paper shows that the number of protein families is expanding and that singletons are an intrinsic part of genomes. A coarse-grained coverage strategy for SG is discussed
18. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, et al. Pfam: clans, web tools and services. *Nucleic Acids Res* 2006;34:D247–D251. [PubMed: 16381856]
19. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;14:755–763. [PubMed: 9918945]
20. Chandonia JM, Brenner SE. The impact of structural genomics: expectations and outcomes. *Science* 2006;311:347–351. [PubMed: 16424331] A quantitative analysis of the novelty, cost and impact of SG structures, compared with structures solved by traditional structural biology
21. Chandonia JM, Brenner SE. Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. *Proteins* 2005;58:166–179. [PubMed: 15521074] An assessment of the various strategies proposed for phase 2 of the PSI
22. O'Toole N, Grabowski M, Otwinowski Z, Minor W, Cygler M. The structural genomics experimental pipeline: insights from global target lists. *Proteins* 2004;56:201–210. [PubMed: 15211505] The overall throughput of the SG pipeline and its bottlenecks are analyzed using survival analysis techniques
23. Hou J, Sims GE, Zhang C, Kim SH. A global representation of the protein fold space. *Proc Natl Acad Sci USA* 2003;100:2386–2390. [PubMed: 12606708]
24. Hou J, Jun SR, Zhang C, Kim SH. Global mapping of the protein structure space and application in structure-based inference of protein function. *Proc Natl Acad Sci USA* 2005;102:3651–3656. [PubMed: 15705717] Protein homology and function can be predicted using distances in 3D 'protein structure space'
25. Scapin G. Structural biology and drug discovery. *Curr Pharm Des* 2006;12:2087–2097. [PubMed: 16796557]
26. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 1988;85:2444–2448. [PubMed: 3162770]
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402. [PubMed: 9254694]
28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410. [PubMed: 2231712]

Figure 1.

Growth of the number of unique structures in the PDB. **(a)** Cumulative growth. **(b)** Number of unique structures released quarterly, showing that the current rate of increase in unique structures (compare red versus black curves) is almost exclusively due to SG deposits. Unique structures are defined as having less than 30% sequence identity to prior deposits in the PDB. XRAY: SG denotes unique X-ray deposits produced by SG programs; XRAY: tSB denotes unique X-ray deposits produced by 'traditional' structural biology. NMR: SG and NMR: tSB denote unique SG and traditional NMR deposits, respectively. The solid lines are exponential trend lines that represent the unique structures obtained from traditional structural biology (X-ray plus NMR; black line) and structures deposited by traditional structural biology and SG

(red line). The time axis corresponds to the time of release of deposits in three-month increments. Sequence identities were calculated using FASTA [26], with a cutoff expectation value $E = 0.01$, where percentage identity was defined as the number of identical residues divided by the length of the target protein. Calculations were performed on the PDB as of February 1, 2007.

Figure 2.

Structural coverage of the Swiss-Prot database. Coverage was calculated as the percentage of sequences with a homologous structure in the PDB (defined as greater than 30% sequence identity as calculated using BLAST [27,28], with the same cutoff and rescaling of identity as for Figure 1) obtained by SG or traditional structural biology methods. tSB + SG represents sequences for which both traditional and SG structures were solved and the traditional structure was solved first, whereas SG + tSB represents the reverse. Calculations were performed on the PDB as of February 1, 2007 and Swiss-Prot as of January 1, 2007.

Figure 3.

Structural coverage of Pfam families. **(a)** Cumulative growth in the number of new families determined by SG or traditional methods. Each time point represents the number of families in the then-current release of Pfam with at least one structural representative in the PDB at that time. The calculations were performed using hidden Markov models (hmmpfam) [19], with an expectation value cutoff of 0.01. Only families surviving to Pfam-A release 21 (November 2006) were taken into account. Sharp increases in the number of families correspond to Pfam releases. **(b)** Number of families in Pfam 21 with a first structural representative deposited in the PDB, grouped by the quarter of the deposition date to the PDB. Calculations were performed

on the PDB as of February 1, 2007. In this representation, the major recent impact of SG is apparent.