**Corrections**

**EARTH, ATMOSPHERIC, AND PLANETARY SCIENCES**

The authors note that Fig. 3 and its corresponding legend appeared incorrectly. The corrected figure and legend appear below.

**BIOPHYSICS AND COMPUTATIONAL BIOLOGY**

The authors note that the following statement should be added as a new Acknowledgments section: “This work was supported by grants from the National Institutes of Health (GM094665, GM094662, GM096041).”

**IMMUNOLOGY**

The authors note that the author name Natalia de Val Alda should instead appear as Natalia de Val. The corrected author line appears below. The online version has been corrected.


---

**Fig. 3.** Generalized Changhsingian to Anisian carbonate carbon isotopic composition from South China. Bed thickness and number, carbonate carbon isotopic composition, weighted mean 206Pb/238U dates, and extinction interval (gray) within Fig. 1, Inset from this study and Cao et al. (11). remainder of carbonate carbon isotopic composition from Payne et al. (52) and geochronology from Galfetti et al. (54). Permian and Triassic conodont zones from Ogg et al. (63). Stage/substage names are global standard chronostratigraphic units used by the International Commission on Stratigraphy.

www.pnas.org/cgi/doi/10.1073/pnas.1403228111
Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative

Kamil Khafizov\textsuperscript{a,b,c,d,1}, Carlos Madrid-Aliste\textsuperscript{a,b,c,d}, Steven C. Almo\textsuperscript{b,c,d,e}, and Andras Fiser\textsuperscript{a,b,c,d,2}

\textsuperscript{a}Department of Systems and Computational Biology, \textsuperscript{b}Department of Biochemistry, \textsuperscript{c}New York Structural Genomics Research Consortium, \textsuperscript{d}Immune Function Network, and \textsuperscript{e}Department of Physiology and Biophysics, Albert Einstein College of Medicine, Bronx, NY 10461

Edited by Barry Honig, Columbia University and Howard Hughes Medical Institute, New York, NY, and approved January 29, 2014 (received for review November 21, 2013)

The exponential growth of protein sequence data provides an ever-expanding body of unannotated and misannotated proteins. The National Institutes of Health-supported Protein Structure Initiative and related worldwide structural genomics efforts facilitate functional annotation of proteins through structural characterization. Recently there have been profound changes in the taxonomic composition of sequence databases, which are effectively redefining the scope and contribution of these large-scale structure-based efforts. The faster-growing bacterial genomic entries have overtaken the eukaryotic entries over the last 5 y, but also have become more redundant. Despite the enormous increase in the number of sequences, the overall structural coverage of proteins—including proteins for which reliable homology models can be generated—on the residue level has increased from 30% to 40% over the last 10 y. Structural genomics efforts contributed ~50% of this new structural coverage, despite determining only ~10% of all new structures. Based on current trends, it is expected that ~55% structural coverage (the level required for significant functional insight) will be achieved within 15 y, whereas without structural genomics efforts, realizing this goal will take approximately twice as long.

The revolution in DNA sequencing technologies has resulted in an enormous, and ever-growing, number of gene sequences over the last decade (1). At the same time, the number of experimentally determined protein structures has lagged increasingly behind, owing to the inherently slower, more resource-intensive, and less-predictable nature of these experiments (2). Three-dimensional structural information often provides the insight required to understand macromolecular function (3, 4), to design new drugs and probe reagents (5), and to progress toward a greater understanding of proteomes through analysis of macromolecular complexes and assemblies (6).

Comprehensive structural coverage and accurate functional annotation are universal goals that are being actively pursued by a wide range of individual investigators, as well as by multidisciplinary, multi-institutional consortia. For example, structural genomics (SG) centers have been established worldwide to contribute to the structural coverage of the protein universe. The technical goals of SG efforts are to reduce the costs associated with structure determination and to accelerate the rate of discovery through the implementation of high-throughput (HTP) methodologies. The infrastructure, resources, and expertise needed to support these efforts are well beyond the capabilities of any single laboratory and require the large multidisciplinary teams that have come to characterize SG.

The scientific goal of SG is to identify protein structures in areas meriting extensive and focused effort, including the identification of new folds or highly diverged folds within partially characterized superfamilies (often with unknown function) and high-priority biomedical areas. This strategy is of particular practical importance, given that individual laboratories tend to work within well-established research areas, where the existence of substantial in vitro and in vivo data justifies the efforts associated with structure determination of specific high-value targets. Experience clearly demonstrates the challenges in securing resources to work in unexplored areas owing to a lack of compelling hypotheses. Without discovery-driven efforts, large amounts of biology will continue to remain obscure, and it will not be possible to systemically examine the structural features of full proteomes, which undoubtedly would lead to unexpected functional and biological insights (7–9).

The Protein Structure Initiative (PSI), supported by the US National Institutes of General Medical Sciences, was established in 2000 and is the largest ongoing coordinated effort in the field of structural biology. The PSI has evolved through three phases (10). The first phase (PSI-1; 2000–2005) demonstrated the feasibility of HTP cloning, protein expression, purification, and structure determination. The implementation of this infrastructure was realized and applied during the production phase (PSI-2; 2005–2010) to significantly expand our knowledge of sequence–structure relationships and to complement efforts in computational biology, such as homology modeling (11), as well as to address specific bottlenecks, such as those associated with membrane protein structure determination (12).

The current phase, termed PSI-Biology, established in 2010, consists of a centralized network including four HTP production centers that support a broad range of biological problems in collaboration with more than a dozen specialized High-Throughput Enabled (HTE) Biological Partnership Centers. These HTE Partnership Centers provide the initial target lists for protein production and structure determination, and perform functional analysis to enable the systematic exploration of specific biological themes, such as mammalian immunity, cell adhesion processes, nucleocytoplasmic transport, and chemokine-related processes. Along with supporting the HTE Partnership Centers, the production centers continue to focus on sequence–structure coverage.

Significance

The Protein Structure Initiative and related worldwide efforts are engaged in the large-scale structural annotation of proteins. In this work, we investigated the dynamic changes that have occurred in this complex race, where sequence databases double every 1.5 y but are becoming increasingly redundant and have exhibited profound changes in taxonomic composition over the last 5 y. Meanwhile, the number of known protein structures is approximately 200 times smaller, and the pace of discovery of new folds is slowing. Nevertheless, the overall structural coverage of proteins has increased from 30% to 40% over the last 10 y. Assuming current trends, ~55% coverage will be achieved within 15 y, a level considered sufficient to fully characterize the metabolic network of an organism.

Author contributions: K.K. and A.F. designed research; K.K. performed research; C.M.-A., S.C.A., and A.F. analyzed data; and K.K., S.C.A., and A.F. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

\textsuperscript{1}Present address: Moscow Institute of Physics and Technology, Moscow 141700, Russia.

\textsuperscript{2}To whom correspondence should be addressed. E-mail: andras.fiser@einstein.yu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1321614111/-/DCSupplemental.
These efforts are complemented by nine centers focused on membrane protein structure determination, as well as additional resources for data tracking and dissemination (www.sbkb.org) (13) and a materials repository (14).

**The Challenge of Structural Coverage**

The feasibility of comprehensive structural coverage is complicated by the exponential growth of the sequence databases. The practical impact of this reality is the need for continual reevaluation of strategies for target selection, given that the source databases of sequences doubles every 18 mo on average (15). For instance, the number of protein sequences in the nonredundant (NR) database, which collects all of the sequences from GenBank translations together with sequences from other databases, reached almost ~17 million at the beginning of 2012, whereas in 2005, when strategies for PSI-2 were being developed, there were only 2 million.

It is not immediately clear how this tremendous expansion has affected the composition of existing databases; i.e., whether the new sequences differ appreciably from previously existing ones or share a high degree of similarity. If a significant fraction of the new sequences arise from organisms (or strains) related to those already sequenced, then little additional structural diversity would be expected. However, it is also plausible that as sequencing projects extend to previously uncharacterized areas (e.g., metagenomes), the composition of the sequence databases could be drastically altered, requiring not only redefinition of PSI target selection priorities, but also a reevaluation of whether existing infrastructure can meet the demands presented by the increased number of potentially important targets.

One of the first assessments of SG contributions in 2006 concluded that SG centers contributed approximately one-half of all newly characterized families, while determining only ~20% of the total number of new structures through 2006 (2). PSI centers were responsible for approximately two-thirds of the worldwide SG contribution. A subsequent report introduced the concept of the “novel modeling leverage,” defined as the approximate number of structurally uncharacterized proteins that can be computationally modeled on the basis of a new structure (16). This type of contribution from SG structures on the residue level increased from 10% in 2001 to 31% in 2005.

Despite many years of continuous effort, 60% of known protein families in the Pfam database (17) still lack structural characterization; i.e., no homologous structure exists for any of the member of the family. It is important to note that Pfam itself expanded from ~6,000 to ~13,000 family definitions over the same 5-y period, further complicating this analysis. Nair et al. (18) showed that PSI-2 contributed approximately 8% of all structures deposited into the Protein Data Bank (PDB), but >20% of all novel structures. The per-structure leverage of PSI is reported to be fivefold to eightfold greater than that of non-SG structures. Levitt (19) introduced the “weighted count method” to measure structural novelty. According to this metric, during 2000–2006, the level of structural novelty increased by a factor of 3.8, but without the SG contributions, it would have increased by only a factor of 2.9.

In this work, we explore (i) how the composition of the sequence databases has changed owing to their ~30-fold growth since 2000; (ii) the likelihood that a protein of a particular type (classified by physicochemical properties, cellular localization, and organismal and phylogenetic occurrence) has a known structure or high-quality model, and the role of SG in past and future efforts; and (iii) the estimated effort required to achieve certain coverage levels, and the associated contributions of SG centers.

**Results**

**Rapid Evolution of Protein Sequence Databases.** We analyzed the taxonomic composition of the NR protein sequence database and how it changed over time (Fig. 1A). In 2000, almost 60% of all proteins were derived from eukaryotic sources (with more than 50% of the sequences deriving from genomes of *Homo sapiens, Mus musculus, Caenorhabditis elegans, Saccharomyces cerevisiae, Drosophila melanogaster,* and *Arabidopsis thaliana*), whereas bacterial and viral sources contributed 22% and 17%, respectively. However, by 2005, the fraction of bacterial sequences in the NR database had exceeded those from eukaryotes, and it has kept growing ever since, reaching ~58% in 2012. At the same time, the relative fractions of both eukaryotic and viral sequences have dropped, accounting for 33% and 5%, respectively, in 2012. Archaeal sequences constituted ~2% of the entire database, whereas synthetic proteins and proteins originating from unknown organisms contributed less than 1%. With regard to the absolute numbers of sequences, the bacterial and eukaryotic proteins were both growing at close to exponential rates, with bacterial sequences outpacing eukaryote sequences.

It is well established that evolution has resulted in protein families with wide-ranging distributions of representation. When
proteins are clustered by structural similarity, the 10 most frequently observed protein folds (i.e., superfolds) account for more than one-third of all genes in a typical genome (20); however, the number of distinct folds is estimated as several thousand (21). A critical question is whether the large increase in bacterial sequences is the result of many similar entries from related species (strains and homologs) or whether unique proteins are being discovered. A related issue is how these factors are varying over time. The absolute number of sequences provides a distorted view of composition, considering that many hundreds of strains or variants have been sequenced for some species.

To take this into account, we repeated the foregoing analysis on a set of nonredundant sequences defined on the basis of clustering at 50% sequence identity (Fig. 1B). The trends in compositional changes in the NR database are quite different from those in the whole (nonclustered) set. The bacterial contribution has dominated the database, with a slightly positive trend over the last decade (increasing from 44% to 58%). Sequences of eukaryotic proteins contribute ~40%, a proportion that has remained rather steady over time. The increased share of bacterial sequences comes at the expense of viral and archaeal entries, which decreased from 8% to 3% and from 7% to 2%, respectively.

To better explore these trends, we plotted the ratio of the clustered number of sequences with respect to the total number of sequences as a function of time (Fig. S1). Our data demonstrate that recent bacterial additions were becoming more redundant, whereas the eukaryotic additions remained rather steady over time. It could be speculated that the large number of bacterial genomes sequenced (2,568, plus metagenomes) has provided a rather complete picture of bacterial gene repertoire with ample opportunities for redundancy, whereas the more limited number of sequenced eukaryotic genomes (311) leaves considerable space for uncovering new genes (22). Interestingly, the low ratio and steady decline of the relative contribution of viral sequences shows that these are highly similar to one another; their overall fraction without filtering was as high as 17% around 2000, but only 5% by 2012 (Fig. S1).

A much more balanced picture emerges if the compositional changes of the sequence database are examined with respect to various physicochemical features and cellular localization, such as proteins with membrane segments, signal peptides, or disordered regions (Fig. S2). Here the difference between the trends of whole and clustered sequence datasets is negligible and not statistically significant. Overall, this finding suggests that the emergence of genomic information from a vast number of new species can drastically transform the composition of databases, but from the standpoint of physicochemical composition, the trends have essentially leveled out.

The median lengths of the protein sequences from different domains of life have generally increased from ~240 to ~270 residues over the last 10 y (Fig. S3), but those for the sequences filtered at 50% have remained quite steady, with a slight decrease. Interestingly, in 2011 the median length of the nonredundant eukaryotic sequences was ~30 residues longer than that of the redundant sequences, suggesting that the more populated protein families are shorter. This effect is reversed in the bacterial entries, where nonredundant sequences are ~20 residues shorter. The shorter bacterial proteins apparently are less redundant.

**Structural Coverage of Proteins: Strong Progress, but Far from Completion.** Assessing progress in the structural coverage of proteins is a nontrivial task, as various parameters vary over time. On one hand, the size of the sequence databases has increased considerably over the last several years, the composition of these sequences is changing, the level of redundancy is shifting, and the pace of discovery of large protein families is slowing rapidly. On the other hand, the rate of protein structure determination is much slower, with only ~80,000 total entries in the PDB as of 2012 (Fig. S4). Conservation of protein structure is much higher than that of sequence structure, resulting in a much lower number of distinct structural families (23). The size distribution of protein fold families is very uneven, and the most frequently occurring ones (e.g., Ig, TIM barrel, Rossmann fold) likely have already been identified.

To assess the structural coverage of the protein universe, we compared all of the sequences in the NR database against all of the sequences in the PDB for each year between 2000 and 2012. We defined residues as “structurally covered” if BLAST aligns them to a sequence of known structure at a given e-value cutoff (SI Materials and Methods), and we assessed coverage on the residue level, by counting the number of “covered” residues divided by their total number. Over the past decade, all structural biology efforts, including SG, have led to an overall increase in the structural coverage of existing proteins from ~30% to ~40% at the residue level (Fig. S5A, black line), despite the tremendous growth of the underlying sequence database, from ~0.5 million to ~16 million (Fig. I4). A breakdown of structural coverage on a residue basis for different domains of life reveals some interesting insights. Viral proteins show the highest coverage, reaching almost 60%, whereas eukaryotic proteins show the lowest, remaining nearly stationary over the last decade (~30%), demonstrating an insufficient focus on these proteins and highlighting inherent challenges in expression and structure determination. Archaeal and bacterial proteins exhibited sharp increases in structural coverage during this period, from 23% to 40% and from 28% to 45%, respectively.

The picture is very different for the 50% clustered sets of sequences (Fig. S5A, dotted lines). Although the overall trend was toward increased structural coverage, this increase was significantly lower for all categories, with overall coverage reaching only ~18% by 2012 (from 13.5% in 2001). The most significant difference was observed for viruses, which had only ~11% coverage by 2012 on a nonredundant set, compared with nearly 60% when no clustering was used. This effect appears to be related to structure determinations of similar viral proteins and to the extensive sequencing of highly related viral strains of the same viruses. For instance, influenza virus, hepatitis C norovirus, rhinovirus...
(A and B), and epizootic hemorrhagic disease virus have 98, 82, 228, 97, and 130 strains sequenced, respectively, among the ∼3,300 currently known full viral genomes (www.ebi.ac.uk/genomes). Bacterial, eukaryotic, and archaeal proteins reached 23%, 15%, and 23% coverage, respectively.

We next analyzed the structural coverage of all of the proteins when classified according to physicochemical properties irrespective of the domain of life from which they are derived (Fig. S5). As expected, globular proteins were covered better than any other category, reaching ∼45% by 2012. Membrane proteins had significantly lower coverage, although there was a noticeable increase in coverage after 2009, reaching almost 28% by the end of 2011. However, when we calculated structural coverage of transmembrane (TM) segments only (i.e., excluding globular parts), coverage was even lower, reaching only ∼25% in 2011.

When redundancy was removed at 50% level (Fig. S6, dashed lines), the coverage of the so-called “singletons” and proteins in small families again started to play a more significant role in the overall assessment. The total coverage dropped significantly for all categories of proteins; globular proteins had slightly higher coverage at ∼21%, whereas TM domains had only 7% coverage.

“Bacterialization” of Eukaryotic Sequences. From the foregoing results, it is clear that eukaryotic proteins show the lowest structural coverage (only the viral proteins clustered at 50% rank slightly lower). It is possible that even this modest coverage might have been achieved by determining structures of homologous bacterial proteins. Interestingly, this effect of “bacterialization” of eukaryotic genomes is clearly detectable in practice, with twofold to fourfold higher structural coverage for eukaryotic proteins that are similar to bacterial proteins compared with those that are not similar. This is likely a consequence of the conscious selection of bacterial targets and/or the easier experimental access to bacterial targets.

To define the limits of this phenomenon, we evaluated the overlap between bacterial and eukaryotic sequences, aiming to identify the common “core” between these two domains of life. For this purpose, we BLASTed four complete eukaryotic proteomes (A. thaliana, D. rerio, H. sapiens, and S. cerevisiae) against all of the bacterial sequences found in the NR database in each of the last 11 y. Although the number of bacterial sequences increased considerably over this period (Fig. 1), the common core between them and the eukaryotic proteomes increased at a much slower rate (Fig. S6). Only ∼29% of all residues in A. thaliana could be aligned with any of the bacterial sequences in the NR database by the end of 2011, indicating a mere 7% increase since 2001. These numbers were comparatively modest for S. cerevisiae (27%, a 7% increase since 2001), D. rerio (22%, up 9%) and, H. sapiens (20%, up 8%). Thus, it is unlikely that determining structures of bacterial proteins (i.e., homologs) will significantly bolster the number of computationally modelable structures in the proteomes of eukaryotic organisms. Adequate coverage of eukaryotic genomes will require direct targeting of these proteins, which will necessitate enhanced investment in eukaryotic expression systems (24).

Impact of PSI on Structural Coverage. After obtaining the overall picture of structural coverage, we assessed the impact of the PSI contribution when the target NR database was clustered at 50% sequence identity. The relationship between increased coverage and structural novelty can be assessed more directly on this set. The overall structural coverage increased from 13.5% to 18.7% from 2001 to 2011 (Fig. S7). (Year 2012 data were incomplete at the time of this study and thus were excluded.) In the breakdown, traditional structural biology (TSB) laboratories contributed 2.3%, PSI centers contributed 2.1%, and other SG centers contributed 1.0%. Of note, this apparently modest growth occurred while the underlying sequence database itself increased by more than 10-fold.

To assess the efficacy of these contributions from different sources, they should be compared with the number of PDB structures deposited by the three types of contributors: 54,286 by TSB laboratories, 5,609 by PSI centers, and 5,204 by other SG centers together. These data indicate that PSI centers contributed almost the same new structural coverage as TSB laboratories, but did so by determining only ∼10% as many structures. This finding confirms earlier studies (16, 18) demonstrating the effectiveness of PSI, which was invaluable for increasing the structural coverage of the protein universe. Combining the efforts of both PSI and non-PSI SG centers, the contribution to total structural coverage during the 11 y was 57%, with only ∼17% of all of the PDB entries deposited during the same period arising from these sources.

As an alternative method of characterizing efficacy, we also calculated the percentage of the average number of residues structurally covered per PDB entry from each of the three contributors. We set the year 2000 as the reference point, and for each subsequent year, divided the observed new structural coverage by the number of structures released in the same year. It is clear that achieving new structural coverage per structure is becoming increasingly difficult (the effect of gradually disappearing “low-hanging fruit”); nevertheless, we can conclude that an “average” structure from PSI and other SG centers provided 8.4-fold and 4.5-fold more new coverage, respectively, than an average structure from TSB laboratories during 2001–2011.

Significant Structural Coverage of the Largest 5,000 Pfam Families by PSI Centers. We revisited an earlier report on the structural coverage of the largest 5,000 Pfam families and how it has changed over the last decade (2). This set can be considered a manually curated set of protein families that removes redundancy. We consider Pfam family structurally covered if there exists at least one PDB entry that shares a significant degree of similarity to its hidden Markov model (HMM) profile. The overall structural coverage increase on this dataset was impressive: from 1,682 to 3,593 families between 2001 and 2012 (Fig. S8). We also determined which PDB deposition was the “first” to structurally define these families. TSB laboratories contributed 1,224 “novel” structures, whereas PSI and non-PSI SG centers deposited 540 and 147 structures, respectively, to the PDB. Approximately 10% of the structures from PSI released in the years 2001–2011 provided novel coverage for Pfam families from this list. The corresponding proportions for TSB laboratories and non-PSI SG centers were between 2% and 3%. These results highlight the power of concerted consortium-wide efforts and, in particular, their utility for defining structural families.

We also explored those 1,407 targets that are still not structurally covered in the Pfam5000 list. We downloaded data from the TargetTrack database (www.sbb.org), which monitors all structural genomics efforts from all centers on each target, and used Hmmer to scan all documented PSI targets against HMM profiles of 1,407 protein families. Nearly all of the remaining targets (1,234 of 1,407) can be found in the database, which means that they have received some level of attention. More specifically, 1,134 of these reached at least “cloning” stage, whereas the rest were at “selected” stage or work officially stopped. Thus, only ∼3% of the Pfam5000 list was not visited by an experimental laboratory. Thus, further progress in coverage would require revisiting targets with alternative experimental strategies or selecting additional targets, because only a few sequences (often only one) were examined in each of these highly populated and evolutionary diverged families.

Structural Coverage of Important Organisms. Evaluating structural coverage of sequence databases is hindered by the difficulty of attempting to compare entries between sequence and structure databases that expand nonlinearly and at different rates. These factors potentially could result in a situation where despite a large and increasing number of structures, the overall structural
database of genes
Menten-like saturation curves were fit to the coverage in terms of covering specific vol. 111 PNAS 2011 using the Levenberg–Marquardt algorithm (human) demonstrated the best coverage among E. coli (red) Comparisons of novelty and effectiveness of structures from SG no. 10 (in terms of novel coverage) were calculated, and PSI/TSB and non-PSI SG/TSB ratios are shown. (Fig. 2). The best coverage was observed for D. rerio (~27% by 2012), and the worst coverage was seen for S. cerevisiae (~11%).
We also analyzed the contributions of PSI, non-PSI SG, and TSB laboratories to the structural coverage of these organisms, as well as the novelty and "usefulness" (in terms of novel coverage) of structures from these contributors during 2001–2011. We first calculated fractions of novel structures/total number of structures for PSI, non-PSI SG, and TSB contributors. Fig. 3A shows the PSITSB and non-PSI SG/TSB ratios of such fractions for the eight organisms studied. This measure also can indicate how many times a structure determined by a PSI or non-PSI SG center is more or less "novel" in terms of covering specific proteomes compared with determination by a TSB laboratory. To account for the different number of structures determined by these contributors, the corresponding numbers of PDB structures were normalized by the number of PDB depositions.

The y-axis in Fig. 3A shows log2 values of these ratios, with 0 indicating equal contributions from any of these sources compared with TSB laboratories. For instance, for E. coli, one PSI structure was on average ~5.4 (2^5.4) times more likely to provide novelty compared with an average TSB structure. For non-PSI SG, this ratio was 3.7 (2^3.7). The strong results of PSI centers on bacterial genomes are consistent with the fact that in the past, SG laboratories focused on producing structural coverage rather than on functional questions, such as effects of single point mutations or new bound ligands/cofactors. Non-PSI SG centers also made a strong, albeit more modest, contribution. For eukaryotic species, the contributions from SG centers were significantly lower. For A. thaliana, PSI and non-PSI structures provided 1.7-fold and 2-fold greater novelty, respectively. For human and zebrafish, only non-PSI SG centers made significantly greater contributions (2.0-fold and 2.4-fold, respectively), likely related to the explicit focus on these species by some of the European centers, such as Structural Proteomics IN Europe (25). Interestingly, no specific contributions from SG centers were observed for yeast proteins (only ~1.3-fold greater for both PSI and non-PSI SG centers).

It is relevant to consider not only the number of novel structures generated, but also the extent of the novelty that these structures provide; i.e., how many previously uncharacterized residues the new structure can cover. Fig. 3B shows PSI/TSB and non-PSI SG/TSB ratios for a number of residues per novel structure. Both PSI and non-PSI SG structures on average cover a few selected proteomes that were obtained before or shortly after 2000. This provided a naturally “frozen” database of genomes that would not change over time, allowing us to monitor the effect of the expanding PDB independently. We considered four bacterial species from different phyla, as well as plant, yeast, fish, and human genomes.

Bacterial species exhibited the best coverage, which was particularly high (58%) for the model organism E. coli (Fig. 2). The lowest coverage among the bacterial species was observed for S. meliloti (~38%). Overall, the eukaryotic organisms had lower structural coverage than the bacterial species. D. rerio (zebrafish) and H. sapiens (human) demonstrated the best coverage among the eukaryotes examined, both of which increased from ~17% to ~35% from 2000 to 2011. S. cerevisiae (baker’s yeast) and A. thaliana (plant) had 14–15% coverage in 2001 and ~29% coverage in 2011.

When only the globular proteins of these species were considered, the structural coverage was higher (Fig. S9A), reaching 68% for E. coli, 37% for D. rerio, and 36% for H. sapiens. For transmembrane proteins (Fig. S9B), the best coverage was observed for D. rerio (~27% by 2012), and the worst coverage was seen for S. cerevisiae (~11%).
twofold to fourfold more residues than TSB PDBs for bacterial proteins. Noticeably, for eukaryotic species, either not much difference was observed (i.e., in human and zebrafish) or TSB covered twice as many residues than PSI per structure. This reflects the fact that PSI centers did not focus on eukaryotic species and that most coverage comes from orthologous bacterial protein structures.

Future Prospects. While considerable progress has been realized, there remain many challenges for reaching widespread structural coverage of model organisms and the protein universe. Here we provide some simple projections for future efforts aimed at enhancing structural coverage.

We first focused on the structural coverage of the entire NR database. From 2001 to 2011, total structural coverage increased from ~30% to ~40% (Fig S5A), with the increase approximately linear. Studies on Thermotoga maritima have shown that 55% coverage is sufficient for a complete metabolic mapping of that organism (7). On a universal level, this 55% coverage can be achieved from current levels within ~15 y. Undoubtedly, the closer we get to these asymptotic values, the more difficult it will be to gain additional coverage.

We were also interested in estimating the contributions of SG efforts to structural coverage in the future. TSB laboratories alone contributed ~0.5% of the structural coverage per year, whereas PSI and non-PSI SG centers contribute 0.39% and 0.13%, respectively. Thus, without the SG contribution, the growth in coverage would be twice as slow, and achieving 55% coverage would take 30 y instead of 15 y.

We also performed this analysis on the NR database clustered at 50% sequence identity. Although the total structural coverage growth was 0.53% per year, TSB laboratories alone contributed 0.20% of this growth, PSI centers provided 0.22%, and other SG centers provided 0.10%. In this case, the SG contribution was even more significant, ~60% overall, despite the fact that the number of deposited structures was lower.

Finally, we focused on the proteomes of two organisms, *E. coli* and *H. sapiens*, to predict when near-complete coverage might be anticipated. For this purpose, we fitted a Michaelis–Menten-like saturation curve and extrapolated the results (Fig 4). By the year 2030, the coverage of *E. coli* could reach ~80%, which is approximately the current theoretical limit given the fraction of disordered regions. For *H. sapiens*, coverage could be ~50% by 2030. This prediction reflects issues discussed earlier, i.e., the realization of higher coverage for eukaryotic organisms will require more substantial targeted efforts.

Discussion

In this work, we have analyzed the compositional changes of sequence databanks and the contribution of structural coverage from various sources (TSB laboratories and SG centers, including PSI laboratories) over time. Despite the enormous increase in sequence databanks, from 2 million in 2005 to 17 million by 2012, newly determined structures have not only kept pace, but have steadily increased the structural coverage of proteins from ~30% to ~40% in the whole NR database and from 13% to 18% when monitored on a nonredundant NR set. With existing technologies and strategies, we project that it would take 15 y to reach an ~55% coverage level, which provides considerable utility for defining large-scale functional characterization of organism-specific properties (e.g., metabolism). These efforts would take twice as long in the absence of the SG contributions, given that SG centers contributed 50–60% of novel coverage despite accounting for <10% of all structure depositions.

The structural coverage of eukaryotic genomes is distinctly lower (29–35%) than that of bacterial genomes (35–58%), and the growth trends are nearly flat, reflecting less focus on novel eukaryotic targets. The contribution of SG centers (especially PSI centers) to the structure determination of eukaryotic targets is barely distinguishable from, or lower than, that of TSB laboratories. To increase the structural coverage of eukaryotic gene products, SG centers need to directly focus on these targets, using technically more expensive and challenging expression systems. This situation is exacerbated by the fact that the vast majority of eukaryotic genes do not have suitable bacterial homologs.

When the PSI centers focused on specific target lists, such as P0m5000, they were very effective. Coverage of these families increased from 34% to 72%, with 36% of the 38% increase coming from SG centers alone, and nearly every family was subjected to at least some degree of experimental effort.

Materials and Methods

The following publicly available sequence alignment, clustering, and feature prediction programs were used in this study: BLAST, Phobius, IUpred, and CDHIT. For sequence information, various National Center for Biotechnology Information resources were utilized. Detailed information on databases and procedures is provided in SI Materials and Methods.