Functional Coverage of the Human Genome by Existing Structures, Structural Genomics Targets, and Homology Models

Lei Xie, Philip E. Bourne*

San Diego Supercomputer Center and Department of Pharmacology, University of California, San Diego, California, United States of America

The bias in protein structure and function space resulting from experimental limitations and targeting of particular functional classes of proteins by structural biologists has long been recognized, but never continuously quantified. Using the Enzyme Commission and the Gene Ontology classifications as a reference frame, and integrating structure data from the Protein Data Bank (PDB), target sequences from the structural genomics projects, structure homology derived from the SUPERFAMILY database, and genome annotations from Ensembl and NCBI, we provide a quantified view, both at the domain and whole-protein levels, of the current and projected coverage of protein structure and function space relative to the human genome. Protein structures currently provide at least one domain that covers 37% of the functional classes identified in the genome; whole structure coverage exists for 25% of the genome. If all the structural genomics targets were solved (twice the current number of structures in the PDB), it is estimated that structures of one domain would cover 69% of the functional classes identified and complete structure coverage would be 44%. Homology models from existing experimental structures extend the 37% coverage to 56% of the genome as single domains and 25% to 31% for complete structures. Coverage from homology models is not evenly distributed by protein family, reflecting differing degrees of sequence and structure divergence within families. While these data provide coverage, conversely, they also systematically highlight functional classes of proteins for which structures should be determined. Current key functional families without structure representation are highlighted here; updated information on the "most wanted list" that should be solved is available on a weekly basis from http://function.rcsb.org:8080/pdb/function_distribution/index.html.

Citation: Xie L, Bourne PE (2005) Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. PLoS Comp Biol 1(3): e31.

Introduction

The three-dimensional structure of a protein is an essential component in elucidating the biological function(s) at the molecular level and in understanding the details of molecular recognition. Traditional structural biology supports a paradigm in which biochemical evidence of function is confirmed and further understood through the study of structure [1]. Structural genomics [2] has changed this paradigm, being motivated by a variety of criteria, including a desire to increase the coverage of known fold space [3]. Concomitantly, complete genome sequences are becoming available at an increasing rate and both putative functions and structures defined for coding regions. Even given the limitations of these assignments, it is an appropriate time to assess the current coverage of protein structure space from a functional perspective relative to the perceived functional coverage of complete genomes, notably human. Further, the registering of structural genomics targets (sequences subject to structure determination) by most projects worldwide [4] provides an excellent opportunity to assess what the perceived coverage of functional space by structure will likely be going forward. This paper makes this assessment, discusses where a change of strategy in selecting targets may be appropriate, and reports on functional classes that are well represented in the human genome but without the existence of structures—the so-called "most wanted list."

Many authors have noted the structural and functional bias in the Protein Data Bank (PDB), but few have attempted to quantify it [5–8]. Rather, general statements are made that refer to the limitations associated with structure determination methods, such as the propensity for small, globular, soluble proteins solved by X-ray crystallography and nuclear magnetic resonance. Beyond physical limitations, there is a bias toward proteins identified as potential drug targets and a historical bias toward structures that, without the benefit of modern techniques, were, from the point of view of protein isolation and structure determination, the most tractable. Where does that put us today, and how can we estimate this bias? A problem that has thwarted such studies is the lack of a common reference frame. This problem has been partially addressed by systems of consistent nomenclature; notwithstanding, depth of coverage is neither complete nor consistent across protein families. Recently, the Enzyme Commission (EC) classification has been used to study the

Received May 16, 2005; Accepted July 18, 2005; Published August 19, 2005
DOI: 10.1371/journal.pcbi.0010031

Abbreviations: EC, Enzyme Commission; GO, Gene Ontology; OMIM, Online Mendelian Inheritance in Man; PDB, Protein Data Bank

Editor: Janet Thornton, European Bioinformatics Institute, United Kingdom

*To whom correspondence should be addressed. E-mail: bourne@sdsc.edu

A previous version of this article appeared as an Early Online Release on July 24, 2005 (DOI: 10.1371/journal.pcbi.0010031.eor).

Synopsis

The sequencing of the human genome provides biologists with new opportunities to understand the molecular basis of physiological processes and disease states. To take full advantage of these opportunities, the three-dimensional structures of the gene products are needed to provide the appropriate level of detail. Since protein structure determination lags behind protein sequence determination, an important and ongoing question becomes: what degree of coverage of the human proteome do we have from experimental structures, and what can we infer by modeling? Or, turning the question around: what structures do we need to determine (the "most wanted list") to further our understanding of the human condition? This paper addresses these questions through integration of existing data resources correlated using comparative functional features, namely the Gene Ontology, which describes biochemical process, molecular function, and cellular location for all types of proteins, and the Enzyme Commission classification for enzymes. Genetic disease states are linked through the Online Mendelian Inheritance in Man resource. Readers can ask their own questions of the resource at http://function.rcsb.org:8080/pdb/function_distribution/index.html. The resource should prove particularly useful to the structural genomics community as it strives to undertake large-scale structure determination with a goal of improving the understanding of protein functional space.

relationships among sequence, structure, and functions [9–15]. Similarly, the Gene Ontology (GO) [16], while still evolving, provides a consistent view of molecular function, biological process, and cell component beyond enzymes. Further, with consistent sequence annotation as a common feature between resources describing structure and human genetic disease, structure–disease relationships can be inferred. Inference requires that care must be taken to assess the statistical significance of the outcome. Using EC and GO to define a common functional framework and highly significant sequence relationships to infer relationships between structure (either solved or under study) and disease, we can measure the biased nature of the PDB and the structures under consideration by structural genomics and suggest protein structures that should be determined to further our understanding of structure and function space.

Results/Discussion

The relationship between protein structures and their function(s) is complex. A single structure superfamily often displays variations in function. Conversely, the same function can be achieved by proteins with different structures [10,11]. Domain recombination and shuffling leads to further functional diversity [17–19]; hence, this study is undertaken at the level of both single domain and whole protein to provide an in-depth view of the functional distribution of protein structures. Single-domain coverage is defined such that at least one domain in the protein has structural information available from the PDB, structural genomics, or homology models. Whole-protein coverage means that structure information for all domains, including their organization, can be directly or indirectly inferred. Similarities between functional distributions from the human genome and from experimental structures or theoretical models are measured with Kendall's tau correlation, which ranges from −1.0 to 1.0. A large positive value indicates that two measurements have similar ranks. The structure–function relationship analysis is based on non-redundant sequence clusters with less than 40% sequence identity and 90% overlap, since functional similarity usually breaks down below these thresholds [10,15].

The Functional Bias of PDB Structures

As stated, several studies have noted structural and functional bias in the PDB [5–8]. In general, protein domains such as transmembrane domains, low complexity regions, and disordered regions, which are not suited to current structure determination methods by X-ray crystallography and nuclear magnetic resonance, are highly underrepresented in the PDB. The columns labeled "PDB/Genome" in Tables 1–4 quantify this bias relative to the known functional classification within the human genome using EC (Table 1) and GO (Tables 2–4) classifications. This bias is examined from the perspective of both a single domain and the whole structure, since many proteins have intracellular and extracellular domains that have been solved without their domain spanning regions. For example, proteins associated with transporter activity (Table 2) have the lowest coverage at the domain level (21.0%), but are further underrepresented at the structure level (12.1%) because of the presence of transmembrane domains. Proteins with two or more contiguous domains, where each of the domains has structure information available, may result in different structures when those domains are swapped. This impacts the observed relationship between values of coverage and correlation computed with Kendall's tau (see Materials and Methods) for single domains versus whole structures, as will be described subsequently.

Beyond obvious experimental limitations, skewed functional distributions of PDB structures are observed for almost all types and levels of EC and GO classifications (Tables 1–4). For example, consider classification by EC number at all levels (only the top-level EC classification is given in Table 1, but current values for all levels of the EC hierarchy are available from the Web site). The correlation coefficients by Kendall's tau between the genome sequences and PDB structures with single-domain coverage for EC: *.*.*.* (all), 2.*.*.* (transferases), EC 2.7.*.* (transferring phosphorous-containing groups), and EC 2.7.1.* (phosphotransferase with an alcohol group as acceptor) are 0.867, 0.889, 0.806, and 0.383, with coverage of 29.9%, 25.4%, 28.7%, and 32.1%, respectively. Thus, even for one of the most structurally studied superfamilies, the protein kinase-like superfamily (all belong to EC 2.7.1), the structures of the majority of atypical kinases (proteins that phosphorylate a variety of substrates) have not been determined, and the protein kinase family itself is slightly underrepresented. The Kendall's tau correlation coefficient is only 0.383 and 0.192 for single-domain and whole-protein coverage, respectively, for the protein kinase-like superfamily.

The functional bias of PDB structures is also notable when using GO molecular function annotations that extend beyond enzyme activity. A total of 16,211 proteins within the human genome can be annotated at this time. As shown in Table 2, according to their Kendall's tau at the whole structure level, several subcategories are underrepresented, notably transporter activity (already noted) and structural molecule activity. Looking deeper (refer to http://function.rcsb.org:8080/pdb/function_distribution/index.html), there are 16 GO subcategories of molecular function associated with

Table 1. Coverage/Kendall's Tau Correlations for Major Categories of Enzyme for Both Single Domains and Whole Proteins

EC Classification	Number of Gene Clusters	PDB/Genome		SG/Genome		Model/Genome	
		Domain	Structure	Domain	Structure	Domain	Structure
1. Oxidoreductases	804	0.274/0.641	0.215/0.616	0.308/0.436	0.173/0.495	0.536/0.737	0.350/0.628
2. Transferases	3,995	0.254/0.889	0.176/0.766	0.310/0.915	0.213/0.915	0.489/0.889	0.304/0.944
3. Hydrolases	3,718	0.369/0.884	0.237/0.899	0.322/0.884	0.201/0.841	0.577/0.873	0.351/0.873
4. Lyases	307	0.278/0.596	0.222/0.552	0.264/0.414	0.208/0.276	0.528/0.733	0.375/0.467
5. Isomerases	347	0.268/0.931	0.244/0.931	0.293/0.966	0.195/0.966	0.610/1.000	0.488/0.733
6. Ligases	638	0.194/0.414	0.081/0.775	0.339/0.690	0.194/0.645	0.629/1.000	0.226/0.745

Current values for nodes of these major branches can be determined from http://function.rcsb.org:8080/pdb/function_distribution/index.html.
DOI: 10.1371/journal.pcbi.0010031.t001

Table 2. Coverage/Kendall's Tau Correlations for Major Categories of GO Molecular Function for Both Single Domains and Whole Proteins

GO Molecular Function[a]	Number of Gene Clusters	PDB/Genome		SG/Genome		Model/Genome	
		Domain	Structure	Domain	Structure	Domain	Structure
Binding (GO 5488)	5,997	0.487/0.729	0.329/0.672	0.360/0.727	0.220/0.654	0.575/0.816	0.303/0.826
Catalytic activity (GO 3824)	4,117	0.336/0.912	0.251/0.857	0.297/0.826	0.165/0.850	0.593/0.941	0.351/0.805
Enzyme regulator activity (GO 30234)	466	0.595/0.753	0.419/0.834	0.419/0.612	0.311/0.636	0.689/0.897	0.432/0.739
Signal transducer activity (GO 4871)	1,509	0.392/0.913	0.155/0.913	0.270/1.000	0.182/1.000	0.541/1.000	0.155/0.913
Structural molecule activity (GO 5198)	559	0.400/0.643	0.400/0.651	0.600/0.706	0.533/0.561	0.400/0.700	0.333/0.581
Transcription regulator activity (GO 30528)	970	0.542/0.636	0.333/0.799	0.750/0.925	0.625/0.828	0.750/0.833	0.417/0.854
Transporter activity (GO 5215)	1,123	0.210/0.671	0.121/0.608	0.304/0.623	0.112/0.453	0.327/0.717	0.164/0.626

Current values for nodes of these seven major branches and other eight minor categories can be determined from http://function.rcsb.org:8080/pdb/function_distribution/index.html.
[a] Eight minor categories are not listed in the table, and can be browsed from the Web site.
DOI: 10.1371/journal.pcbi.0010031.t002

Table 3. Coverage/Kendall's Tau Correlations for Major Categories of GO Biological Process for Both Single Domains and Whole Proteins

GO Biological Process[a]	Number of Gene Clusters	PDB/Genome		SG/Genome		Model/Genome	
		Domain	Structure	Domain	Structure	Domain	Structure
Behavior (GO 7610)	62	0.538/0.672	0.231/0.696	0.077/0.468	0.000/0.000	0.615/0.504	0.077/0.468
Cellular process (GO 9987)	7,133	0.484/1.000	0.355/1.000	0.426/1.000	0.280/1.000	0.667/1.000	0.377/1.000
Development (GO 7275)	1,171	0.376/0.798	0.257/0.638	0.468/0.882	0.231/0.705	0.688/0.803	0.321/0.825
Physiological process (GO 7582)	4,483	0.503/0.795	0.371/0.682	0.435/0.841	0.304/0.828	0.654/0.932	0.369/0.787
Regulation of biological process (GO 50789)	395	0.455/0.909	0.321/0.815	0.312/0.679	0.205/0.605	0.607/0.889	0.268/0.780

Current values for nodes of these five major branches and other two minor categories can be determined from http://function.rcsb.org:8080/pdb/function_distribution/index.html.
[a] Two minor categories—viral life of cycle and biological process unknown—are not listed in the table, and can be browsed from the Web site.
DOI: 10.1371/journal.pcbi.0010031.t003

Table 4. Coverage/Kendall's Tau Correlations for Major Categories of GO Cell Component for Both Single Domains and Whole Proteins

GO Cellular Location[a]	Number of Gene Clusters	PDB/Genome		SG/Genome		Model/Genome	
		Domain	Structure	Domain	Structure	Domain	Structure
Cell (GO 5623)	4,599	0.441/0.615	0.275/0.692	0.549/0.846	0.373/0.854	0.618/0.680	0.314/0.530
Extracellular matrix (GO 31012)	235	0.333/0.105	0.190/0.447	0.286/0.738	0.143/0.598	0.524/0.800	0.190/0.359
Extracellular region (GO 5576)	639	0.200/0.400	0.200/0.400	0.000/0.000	0.000/0.000	0.600/0.516	0.200/0.400
Organelle (GO 43226)	4,104	0.444/1.000	0.350/1.000	0.538/1.000	0.402/1.000	0.718/1.000	0.444/1.000
Protein complex (GO 43234)	1,203	0.306/0.357	0.214/0.352	0.410/0.550	0.249/0.468	0.520/0.532	0.295/0.431

Current values for nodes of these five major branches and other two minor categories can be determined from http://function.rcsb.org:8080/pdb/function_distribution/index.html.
[a] Two minor categories—virion and cell component unknown—are not listed in the table, and can be browsed from the Web site.
DOI: 10.1371/journal.pcbi.0010031.t004

structural molecule activity. Twelve of them have been mapped to the human genome. The most structurally underrepresented proteins include the structural constituent of ribosome (two structures but 180 annotations), of myelin sheath (zero structures and two annotations), of epidermis (zero structures and six annotations), of tooth enamel (zero structures and five annotations), of bone (zero structures and three annotations), of chorion (zero structures and one annotation), and of cell wall (zero structures and one annotation).

Table 3 provides the distribution of protein domains and whole structures according to the subcategories of GO biological process. A total of 14,876 proteins within the human genome can be annotated at this time. Thus, biological process is less well characterized than molecular function, presumably since molecular function cannot necessarily be related to a role in a complex biological process. Notwithstanding, both single-domain and whole-protein structures with an identified role in cellular process are underrepresented.

Table 4 provides the distribution of protein domains according to the subcategories of GO cell component. Overall, the distribution between PDB structures and the human genome is comparable, with a Kendall's tau of 0.714; however, proteins identified within the cell (GO 5623) are underrepresented at both the structure and domain levels. Coverage is not as favorable as distribution: only 38.3% of the subcategories of cell location have at least one structure domain representative. Of those, the vast majority are annotated as cell (4,599 out of 8,936 gene clusters). Under cell (see http://function.rcsb.org:8080/pdb/function_distribution/index.html), there are 12 subcategories that have been assigned to the human genome, five have no structure representation and seven have at least one structure domain representative (membrane [3,354 gene and 236 structure clusters], intracellular [1,237 gene and 169 structure clusters], cell surface [23 gene and four structure clusters], cell projection [26 gene and one structure cluster], cell fraction [621 gene and 75 structure clusters], apical part of cell [six gene and one structure cluster], and basal part of cell (two gene and one structure cluster]). As expected, membrane is structurally underrepresented, and intracellular and cell fraction is structurally overrepresented. There is no structural information available for five small subcategories of cell: site of polarized growth (three gene clusters), periplasmic space (three gene clusters), midbody (one gene cluster), external encapsulation (two gene clusters), and cell soma (one gene cluster).

Structural Coverage of Human Genetic Diseases

Three-dimensional protein structures are important in understanding the mechanisms of human genetic diseases [20], predicting the effect of non-synonymous single nucleotide polymorphisms [20,21], and developing new personalized medicines [22]. For example, a recent study highlights the application of PDB structure and homology models in understanding a predisposition to breast and ovarian cancer [23]. However, the current identification and coverage of human genetic disease space, as identified by the Online Mendelian Inheritance in Man (OMIM), is limited: 218 non-redundant human genome sequence clusters, 46 structure clusters, and 34 structural genomics target clusters. The PDB currently covers 69.9% of the disease categories described by OMIM, but the distribution based on class of disease is uneven. For example, diseases of the central nervous system have the largest single representation (20 structure clusters) with a disproportionately large number of structural genomics targets (23 target clusters). Blood and lymph based diseases have a disproportionate number of ten solved structures, but an underrepresentation of targets (three clusters). Conversely, diseases of the ear, nose, and throat are underrepresented (six structure and seven target clusters). Overall, cancers have an appropriate level of structures and targets; however, digging deeper reveals a different situation. For example, there are no structures available for the five human proteins that have been associated with prostate cancer, although homology models can be inferred for domains such as prostate specific kallikrein of serine proteases [24]. Data showing measurable differences in protein and gene regulatory networks between the early- and the late-stage prostate cancer [25] only highlight the need to further understand the structural basis of this disease. Human genetic disease distributions, while limited, are undoubtedly influenced by historical precedent, preventative and treatable conditions, and social and, hence, funding pressures.

The Contributions of Homology Modeling

While the number of three-dimensional structures of proteins has increased close to the near-exponential rate predicted by Dickerson in 1978 as number of structures = $\exp(0.19 \times \text{year})$ [26], there are still a vast number of protein sequences without structure information available. Homology modeling can potentially provide putative structure information for these sequences to facilitate our understanding of their function and evolution [27,28]. Reliable homology modeling usually requires that the query sequence share at least 30% sequence identity with the template structure for each domain [27]. Domain rearrangements and lack of domain structures reduce the effectiveness of homology modeling for whole structures, as shown in the columns labeled "Model/Genome" in Tables 1–4. In almost all EC and GO classifications, coverage and distribution falls for whole structures versus domains. From a biological perspective, modeling of only a subset of domains within a structure limits the value of modeling.

As expected, the distribution of homology models is highly correlated with the availability of PDB structures. Single-domain coverage across the whole human genome indicates that our ability to provide homology models for domains in the different GO molecular function categories varies from 32% to 75%. For the modeling of whole proteins, coverage drops, varying from 16% to 41%. Transporter activity and signal transducer activity is the most difficult to model at the whole-protein level. GO functional coverage for signal transducer activity drops from 0.541 to 0.155. Thus, while catalytic domains involved in signal transduction are well represented and can be modeled in 54% of cases, these data quantitatively show that the associated non-transmembrane domains of the whole protein are significantly underrepresented, thereby limiting our ability to model whole proteins in 38% of cases.

Considering enzymes alone, our ability to homology model single domains is fairly evenly distributed across all major

classes (Table 1). At the whole-protein level, this picture changes. Retaining a high Kendall's tau even as coverage drops significantly could imply that functional diversity comes primarily from domain recombination rather than from new domains that cannot be modeled. Indeed, it has recently been reported that contemporary ligases evolved by domain fusions [29], a fact supported by a relatively small drop in Kendall's tau from 1.000 to 0.745 for single domains versus the whole protein.

Low correlation within a functional class implies that homology models can be inferred from structures in different functional subclasses and other species. For example, in the oxidoreductases (EC 1.x.x.x), five classifications (1.7, acting on other nitrogenous compounds as donors; 1.9, acting on a heme group of donors; 1.10, acting on diphenols and related substances as donors; 1.17, acting on $-CH_2$ groups; and 1.97, other oxidoreductase) are not structurally covered at all. However, with the exception of 1.97, other oxidoreductase, the proteins in the four remaining subclasses can be modeled from structural templates present in the other functional subcategories, implying a close evolutionary relationship within this functional class.

Conversely, experimental structural coverage is more critical for functional classes with more distinct evolutional origins, such as the protein kinase-like superfamily, which is in the transferase category. It has been suggested [30] that atypical kinases diverged early in evolution from protein kinases; therefore, homology models of atypical kinases derived from protein kinases are likely insufficient to infer their functional and evolutional roles. In 13% of cases, while the homology model is identified to belong to the protein kinase-like superfamily, the specific family cannot be determined.

The Contribution of Structural Genomics

One approach to the selection of structural genomics targets has been to focus on increasing the coverage of fold space [31–33]. A recent review suggests that the first phase of structural genomics has been successful in this regard [34]. It is anticipated that functional roles will be given greater precedence in future phases of the project [35–37]. If so, a question to address is: does the current complement of structural genomics targets and the structures solved by these projects reduce the functional bias present in the PDB? The short answer is yes, but only significantly in some functional categories (Tables 1–4, columns labeled "SG/Genome") and assuming two to three times the number of structures than we have now (based on the relative numbers of clusters between structure genomics targets and PDB structures, given a 40% sequence identity cutoff).

Within the enzymes (Table 1), ligases will benefit the most and lyases the least. Based on GO molecular function (Table 2), structural molecule activity and transcription regulator activity (single domain) will be impacted the most; binding, catalytic activity, signal transducer activity, and transporter activity the least. In terms of GO biological processes (Table 3), structural genomics will contribute almost nothing to our understanding of behavior and about equally to cellular processes, development, physiological processes, and regulation of biological processes. Finally, current structural genomics targets will not contribute to our understanding of extracellular region of cell component (Table 4). The most notable impact of structural genomics overall is in our potential understanding of transcription regulator activity, which shows an improvement in coverage from 0.542 to 0.750 and an improved Kendall's tau of 0.636 to 0.925 for a single domain.

Drilling down into one of these categories, the previously described structurally underrepresented GO class for molecular function—namely, structural molecule—becomes better populated such that targets will increase the coverage of the structural constituent of tooth enamel (one structure but five annotations), of myelin sheath (one structure and two annotations), and of ribosome (48 structures and 180 annotations). There remains no anticipated experimental structure information for the structural constituent of epidermis, bone, chorion, and cell wall (total 11 annotations).

Given these findings, it is timely to consider the choice of structural genomics targets. It has been suggested that solving the structures of proteins from the 5,000 Pfam families will cover more of fold space than focusing on a single genome [38]. Here, we look at target selection from a functional perspective and provide a tool for comparing the functional coverage by the existing PDB and what the existing complement of structural genomics targets do to that functional coverage. The remainder of the paper considers one application of the tool in providing a strategy for selecting structures that could be used to maximize our understanding of structure–function relationships with respect to the human genome.

Defining Structures That Should Be Determined

To date, approximately 50% of human genes (16,211 terms for GO molecular function, 14,876 terms for GO biological process, and 13,322 terms for GO cell component) have at least one GO annotation. However, approximately 70% of these GO molecular function categories are yet to be covered by experimental structures with even one identifiable domain. The structural coverage of the human genome is even lower with respect to sequence space: approximately 10% coverage by structure at 40% sequence identity. Stated another way, 5% of the human genome, which covers 30% of functional space, has structure representation for at least one domain in a protein. If all current structural targets were determined, it is estimated that coverage of the human genome and functional space would increase to 20% and 50%, respectively. Homology modeling would increase genome and functional coverage to 40% and 60%, but what these putative high-throughput models add to our understanding of molecular function remains questionable. When taking domain recombination into account, the functional coverage of the human genome by existing experimental structures and anticipated structures being determined by structure genomics decreases to approximately 25% of the functional space.

This lack of coverage perhaps calls for a new strategy to select targets for structure determination. Here, one such strategy is outlined for choosing targets to increase the coverage of functional space. It is based on the following criteria: (1) functional categories are preferred where proteins with experimental or theoretical three-dimensional models are underrepresented; (2) from (1), proteins without SCOP superfamily assignments are preferred; (3) if the protein is identified as being associated with a disease or is

identified in multiple functional categories, it has a higher priority; and (4) less experimentally tractable proteins—for example, those with transmembrane segments—can be filtered out.

From our initial analysis, approximately 2,000 non-redundant human genes with GO annotation have no experimental structure in the PDB, nor are they identified structural genomics targets or amenable to homology modeling. Of this 2,000, approximately 50% include transmembrane domains. After removing transmembrane and low-complex regions, about 1,800 include at least one domain that is potentially solvable. The most understudied proteins of this 1,800 are various types of transporters and receptors. It should be noted that it requires fewer targets to cover this functional space than the equivalent sequence space.

Ranked by the size of the cluster of proteins, examples of the most pressing biological molecule functions for which structural representation is needed and soluble structure domains are probably present are listed here (Tables 5 and 6) and at http://function.rcsb.org:8080/pdb/function_distribution/index.html, which is updated regularly. For catalytic activity, most of them are involved in protein syntheses and gene regulation. For binding, most of them are involved in signal transduction and have additional benefit as potential drug targets.

Several genes without experimental structures and not found in the structural genomics target list are annotated by both GO and disease terms (see http://function.rcsb.org:8080/pdb/function_distribution/index.html). For example, congenital adrenal hyperplasia is associated with three gene clusters. Two of them are annotated with oxygen binding (GO ID: 19825), and one with steroid 11-beta-monooxygenase activity (GO ID: 4507).

Table 5. Most Wanted Structures According to EC Numbers[a]

EC Class	Number of Gene Clusters	Function
2.3.1.51	12	1-acylglycerol-3-phosphate O-acyltransferase
3.6.3.1	10	Phospholipids-translocating ATPase
2.4.99.6	10	N-acetyllactosaminide alpha-2,3-sialyltransferase
3.1.13.4	6	Poly(A)-specific ribonuclease
1.3.99.5	5	3-oxo-5-alpha-steroid 4-dehydrogenase
3.6.1.6	5	Nucleoside-diphosphatase
1.5.99.2	5	Dimethylglycine dehydrogenase
3.1.3.56	4	Inositol-polyphosphate 5-phosphatase
1.13.11.40	4	Arachidonate 8-lipoxygenase
2.7.7.13	4	Mannose-1-phosphate guanylyltransferase
3.4.11.2	4	Membrane alanyl aminopeptidase
3.6.3.28	3	Phosphonate-transporting ATPase
2.4.1.68	3	Glycoprotein 6-alpha-L-fucosyltransferase
3.4.24.61	3	Nardilysin
2.4.1.69	3	Galactoside 2-alpha-L-fucosyltransferase
3.6.1.5	3	Apyrase
3.5.4.6	3	AMP deaminase
3.6.3.19	3	Maltose-transporting ATPase
1.3.3.2	3	Lathosterol oxidase
1.5.99.1	3	Sarcosine dehydrogenase
3.1.3.4	3	Phosphatidate phosphatase
2.4.99.7	3	(Alpha-N-acetylneuraminyl-2,3-beta-galactosyl-1,3)-N-acetyl- galactosaminide 6-alpha-sialyltransferase
2.4.99.1	3	Beta-galactoside alpha-2,6-sialyltransferase
2.1.1.62	3	mRNA (2′-O-methyladenosine-N(6)-)-methyltransferase
3.6.1.15	3	Nucleoside-triphosphatase
2.8.2.23	3	(Heparan sulfate)-glucosamine 3-sulfotransferase 1
2.4.1.109	3	Dolichyl-phosphate-mannose—protein mannosyltransferase
3.6.3.30	3	O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase
1.3.1.70	3	Delta(14)-sterol reductase
2.4.1.102	3	Beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase
1.5.1.2	3	Pyrroline-5-carboxylate reductase
3.1.21.7	3	Deoxyribonuclease V

The proteins are clustered with 40% sequence identity.

[a] Data for clusters of fewer than three can be obtained from http://function.rcsb.org:8080/pdb/function_distribution/index.html.

DOI: 10.1371/journal.pcbi.0010031.t005

Table 6. Most Wanted Structures According to GO Classification[a]

GO	GO ID	Number of Gene Clusters	Function
	15171	15	Amino acid transporter activity
Molecular function	5328	12	Neurotransmitter:sodium symporter activity
	45028	8	Purinergic nucleotide receptor activity, G-protein coupled
	5338	7	Nucleotide-sugar transporter activity
	16526	7	G-protein coupled receptor activity, unknown ligand
	30165	7	PDZ domain binding
	30280	6	Structural constituent of epidermis
	8717	6	D-alanyl-D-alanine endopeptidase activity
	4985	6	Opioid receptor activity
	4500	5	Dopamine beta-monooxygenase activity
	4012	5	Phospholipid-translocating ATPase activity
	8508	5	Bile acid:sodium symporter activity
	8518	5	Reduced folate carrier activity
	4994	5	Somatostatin receptor activity
	8503	5	Benzodiazepine receptor activity
	15321	5	Sodium-dependent phosphate transporter activity
	8158	5	Hedgehog receptor activity
	4993	5	Serotonin receptor activity
	4579	5	Dolicyl-diphospholigosaccharide-protein glycotransferase activity
	4709	5	MAP kinase kinase kinase activity
	17022	5	Myosin binding
	15520	5	Tetracycline:hydrogen antiporter activity
Biological process	15893	10	Drug transport
	7026	10	Negative regulation of microtubule depolymerization
	46777	9	Autophosphorylation
	15711	8	Organic anion transport
	7214	7	Gamma-aminobutyric acid signaling pathway
	42832	6	Defense response to pathogenic protozoa
	46803	6	Reduction of virulence
	15780	5	Nucleotide-sugar transport
	6829	5	Zinc ion transporter
	15904	5	Tetracycline transport
Cell component	5852	7	Eukaryotic translation initiation factor 3 complex
	5747	6	Respiratory chain complex I
	5883	5	Neurofilament
	5678	5	Chromatin assembly complex
	922	5	Spindle pole

The proteins are clustered with 40% sequence identity.

[a] Data for clusters of fewer than five can be obtained from http://function.rcsb.org:8080/pdb/function_distribution/index.html.

DOI: 10.1371/journal.pcbi.0010031.t006

In summary, by using common annotation as found in the GO and the EC classification scheme, we have been able to correlate the biological functions of proteins and their constituent domains for both experimentally derived structures and those under determination by structural genomics projects worldwide. Further, by using empirical sequence limitations known from homology modeling experiments and by clustering human genome sequences according to sequence identity, we can estimate the impact that current structure determination strategies will have on our understanding of structure–function relationships from homology modeling. Finally, by introducing relationships between gene products and known disease states, we have provided pointers for choosing structures to be determined to have the maximum impact on our understanding of human genetic disease. To facilitate these choices, a Web resource has been established at http://function.rcsb.org:8080/pdb/function_distribution/index.html to allow readers to make their own assessments of the progress of structural biology. The resource will be updated on a weekly basis to provide a current view. The resource itself will be the subject of a separate publication.

Materials and Methods

Date sources and annotation mapping. The human genome sequences (version 26.35.1) were downloaded from Ensembl database [39]. Wild-type sequences associated with PDB structures were generated by associating the structural sequence with that from UniProt [40] using database cross references records. Subsequently, all wild-type PDB sequences of the human proteins were mapped to the genes in the human genome through sequence alignment using Blast [41]. A gene was considered to have a structure representation if it had 100% sequence identity with the wild-type sequence of the PDB structure. Structural genomics targets were taken from targetdb [42], the worldwide repository of all sequences representing structures being attempted. Among more than 5,000 registered human target sequences, there were 3,141 and 4,784 targets mapped to the 3,200 and 4,218 Ensembl human genes with sequence identity 100% and greater than 90%, respectively. The 4,784 targets with sequence identity above 90% were used in our analysis, with 2,180 of them having GO or EC terms assigned.

Sequences were assigned GO terms from the EBI GOA resource (http://www.ebi.ac.uk/GOA). The query sequence was aligned with the UniProt GO annotated sequence with Blast [41]. If the Blast sequence identity was above 40%, and the overlap was above 90%, the annotated GO terms were mapped to the query gene (16,211 for GO molecular function, 14,876 for GO biological process, and 13,322 for GO cell component). The threshold is based on the observation that below 40% sequence identity with global alignment, the functional similarity relationship breaks down [10,15]. Sequences were also mapped to enzyme classification numbers with the annotations and sequences in the UniProt database as the reference. The 40% sequence identity and 90% overlap threshold was also applied to EC mapping.

Genome sequences were masked for low-complexity regions, coiled-coils, and transmembrane helical domains, using SEG [43], Coils [44], and TMHMM [45], respectively. SCOP superfamily domains [46] of unmasked regions of human genome sequences were assigned with HMMER [47]. A set of hidden Markov Models of SCOP domains was taken from SUPERFAMILY 1.65 [48]. Given the current stage of homology modeling, the model was usually reliable when the sequence identity was above 30% between the query sequence and the template structure [25]. Thus, only those assigned domains with sequence identity above 30% in the alignment were considered as homology models. The sequence regions that were not assigned by SCOP domains were further parsed with Pfam 16.0 [49]. The remaining unmasked sequence segments that were not annotated by either SCOP or Pfam but longer than 30 residues were considered as novel domains. Moreover, for contiguous domains, their orders were recorded in the database. The two domains were considered as contiguous with each other if they were not separated by the filtered sequence segments.

All genome sequences were clustered with 40% sequence identity and 90% overlap using CD-HIT [50].

For PDB structures and structural genomics targets, SCOP domains and their arrangements were computed with the same procedure as for genome sequences.

The original mapping of structures to OMIM numbers was taken from SWISS-PROT [51]. The mapping of genome sequences to OMIM numbers was from NCBI [52]. These mappings were recorded and used from the PDB beta site [26].

Data analysis. For each functional or structural category, the number of sequence or structure clusters in the subcategory was normalized with that of sequence clusters from the genome. The overall similarity between two distributions—for example, the PDB structure and the human genome—was measured with Kendall's tau correlation coefficient τ [53]. For N pairs of measurements (xi, yi), each of them has N(N–1)/2 pairs of data points. τ is computed as:

$$\tau = \frac{con - dis}{\sqrt{(con + dis + ey)(con + dis + ex)}}$$

con is defined as the number of pairs where (xi, xj) ranks the same as (yi, yj). *dis* is the number of pairs where (xi, xj) ranks the opposite to (yi, yj). *ey* is the number of pairs where yi = yj, and *ex* is the number of pairs where xi = xj.

Kendall's tau correlation coefficient ranges from −1.0 to 1.0. If two measurements have the similar ordering, it will be close to 1.0. The opposite ordering will give values close to −1.0. The coverage was also computed and defined as the ratio between the number of functional categories that have at least one structure representative and all functional categories.

Data access. Data were warehoused in a single relational database where relations represent the mappings between the individual data sources. From a user's perspective, data appear in a multi-dimensional space. Each of the functional or structural categories is considered one dimension in the multi-dimensional space. A PDB structure or a genome sequence occupies a cube in this space. Any combination of two dimensions can be selected, and the distributions corresponding to the selected dimensions are calculated and displayed. The dimensions are organized in a hierarchal fashion according to their functional or structural taxonomies. Thus, data mining tasks such as drill-down or roll-up are supported. The database is accessible from http://function.rcsb.org:8080/pdb/function_distribution/index.html.

Supporting Information

Accession Numbers

The UniProt (http://www.pir.uniprot.org/) accession number for prostate specific kallikrein of serine proteases is P07288.

Acknowledgments

This work is supported by the Protein Data Bank through a multi-agency grant (NSF DBI 9814284), and PEB is supported in part by a National Institutes of Health grant (GM63208).

Competing interests. The authors have declared that no competing interests exist.

Author contributions. LX and PEB conceived and designed the experiments, analyzed the data, and wrote the paper. LX performed the experiments and contributed reagents/materials/analysis tools. ■

References

1. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA (2000) From structure to function: Approaches and limitations. Nat Struct Biol 7: 991–994.
2. Brenner SE, Levitt M (2000) Expectations from structural genomics. Protein Sci 9: 197.
3. Portugaly E, Linial M (2000) Estimating the probability for a protein to have a new fold: A statistical computational model. Proc Natl Acad Sci U S A 97: 5161.
4. Westbrook J, Feng Z, Chen L, Yang H, Berman HM (2003) The Protein Data Bank and structural genomics. Nucleic Acids Res 31: 489–491.
5. Peng K, Obradovic Z, Vucetic S (2004) Exploring bias in the Protein Data Bank using contrast classifiers. Pac Symp Biocomput 2004: 435–446.

6. Brenner SE, Chothia C, Hubbard TJ (1997) Population statistics of protein structures: Lessons from structural classifications. Curr Opin Struct Biol 7: 369.
7. Gerstein M (1998) How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. Fold Des 3: 497.
8. Liu J, Rost B (2002) Target space for structural genomics revisited. Bioinformatics 18: 922.
9. Devos D, Valencia A (2000) Practical limits of function prediction. Proteins: Struct Funct Genet 41: 98–107.
10. Hegyi H, Gerstein M (1999) The relationship between protein structure and function: A comprehensive survey with application to the yeast genome. J Mol Biol 288: 147–164.
11. Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307: 1113–1143.
12. Wilson CA, Kreychman J, Gerstein M (2000) Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 297: 233–249.
13. Martin ACR, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, et al. (1998) Protein folds and functions. Structure 6: 875–884.
14. Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318: 595–608.
15. Tian WD, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333: 863–882.
16. Consortium TGO (2000) Gene Ontology: Tool for the unification of biology. Nat Genet 25: 25–29.
17. Hegyi H, Gerstein M (2001) Annotation transfer for genomics: Measuring functional divergence in multi-domain proteins. Genome Res 11: 1632–1640.
18. Apic G, Huber W, Teichmann SA (2003) Multi-domain protein families and domain pairs: Comparison with known structures and a random model of domain recombination. J Struct Funct Genomics 4: 67–78.
19. Littler SJ, Hubbard SJ (2005) Conservation of orientation and sequence in protein domain-domain interactions. J Mol Biol 345: 1265–1279.
20. Wang Z, Moult J (2001) SNPs, protein structure, and disease. Hum Mutat 17: 263–270.
21. Sunyaev S, Ramensky V, Bork P (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet 16: 198–200.
22. Sander C (2000) Genomic medicine and the future of health care. Science 287: 1977–1978.
23. Mirkovic N, Marti-Renom MA, Weber BL, Sali A, Monteiro AN (2004) Structure-based assessment of missense mutations in human BRCA1: Implications for breast and ovarian cancer predisposition. Cancer Res 64: 3790–3797.
24. Carvalho AL, Sanz L, Barettino D, Romero A, Calvete JJ, et al. (2002) Crystal structure of a prostate kallikrein isolated from stallion seminal plasma: A homologue of human PSA. J Mol Biol 322: 325–337.
25. Lin B, White JT, Lu W, Xie T, Utleg AG, et al. (2005) Evidence for the presence of disease-perturbed networks in prostate cancer cells by genomics and proteomic analyses: a systems approach to disease. Cancer Res 65: 3081–3091.
26. Deshpande N, Addess KJ, Bluhm WF, Merino-Ott JC, Townsend-Merino W, et al. (2005) The RCSB Protein Data Bank: A redesigned query system and relational database based on the mmCIF schema. Nucleic Acids Res 33: D233–D237.
27. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, et al. (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29: 291–325.
28. Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294: 93–96.
29. Shuman S, Lima CD (2004) The polynucleotide ligase and RNA capping enzyme superfamily of covalent nucleotidyltransferases. Curr Opin Struct Biol 14: 757–764.
30. Scheeff ED, Bourne PE (2005) Structural evolution of the protein kinase-like superfamily. Submitted.
31. Sanchez R, Pieper U, Melo F, Eswar N, Marti-Renom MA, et al. (2000) Protein structure modeling for structural genomics. Nat Struct Biol 7: 986–990.
32. Brenner SE (2000) Target selection for structural genomics. Nat Struct Biol 7: 967–969.
33. Vitkup D, Melamud E, Moult J, Sander C (2001) Completeness in structural genomics. Nat Struct Biol 8: 559–566.
34. Todd AE, Marsden RL, Thornton JM, Orengo CA (2005) Progress of structural genomics initiatives: An analysis of solved target structures. J Mol Biol. In press.
35. Yee A, Pardee K, Christendat D, Savchenko A, Edwards AM, et al. (2003) Structural proteomics: Toward high-throughput structural biology as a tool in functional genomics. Acc Chem Res 36: 183–189.
36. Kyogoku Y, Fujiyoshi Y, Shimada I, Nakamura H, Tsukihara T, et al. (2003) Structural genomics of membrane proteins. Acc Chem Res 36: 199–206.
37. Lundstrom K (2004) Structural genomics on membrane proteins: The MePNet approach. Curr Opin Drug Discov Devel 7: 342–346.
38. Chandonia JM, Brenner SE (2005) Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. Proteins 58: 166–179.
39. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, et al. (2002) The Ensembl genome database project. Nucleic Acids Res 30: 38–41.
40. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, et al. (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res 33: D154–D159.
41. F. AS, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
42. Chen L, Oughtred R, Berman HM, Westbrook J (2004) A target registration database for structural genomics projects. Bioinformatics 20: 2860–2862.
43. Wootton JC (1994) Non-globular domains in protein sequences: Automated segmentation using complexity measures. Comput Chem 18: 269–285.
44. Lupas AN, Van Dyck M, Stock J (1991) Predicting coiled coils from protein sequences. Science 252: 1162–1164.
45. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J Mol Biol 305: 567–580.
46. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.
47. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755–763.
48. Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313: 903–919.
49. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam Protein Families Database. Nucleic Acids Res 32: D138–D141.
50. Li W, Jaroszewski L, Godzik A (2001) Clustering of highly homologous sequences to reduce the size of large protein database. Bioinformatics 17: 282–283.
51. Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28: 45–48.
52. Wheeler DL, Barret T, Benson DA, Bryant SH, Canese K, et al. (2005) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 33: D39–D45.
53. Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1992) Numerical recipes in C: The art of scientific computing. Cambridge: Cambridge University Press. 994 p.