Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 11911–11916, October 1997
Biophysics

A structural census of the current population of protein sequences

(sequence analysis/genome comparison/fold family/databank statistics/protein evolution)

MARK GERSTEIN*† AND MICHAEL LEVITT‡

*Molecular Biophysics and Biochemistry Department, P.O. Box 208114, Yale University, New Haven, CT 06520-8114; and ‡Department of Structural Biology, Stanford University, Stanford, CA 94305

Communicated by Thomas A. Steitz, Yale University, New Haven, CT, August 13, 1997 (received for review May 14, 1997)

ABSTRACT **We examine the occurrence of the ≈300 known protein folds in different groups of organisms. To do this, we characterize a large fraction of the currently known protein sequences (≈140,000) in structural terms, by matching them to known structures via sequence comparison (or by secondary-structure class prediction for those without structural homologues). Overall, we find that an appreciable fraction of the known folds are present in each of the major groups of organisms (e.g., bacteria and eukaryotes share 156 of 275 folds), and most of the common folds are associated with many families of nonhomologous sequences (i.e., >10 sequence families for each common fold). However, different groups of organisms have characteristically distinct distributions of folds. So, for instance, some of the most common folds in vertebrates, such as globins or zinc fingers, are rare or absent in bacteria. Many of these differences in fold usage are biologically reasonable, such as the folds of metabolic enzymes being common in bacteria and those associated with extracellular transport and communication being common in animals. They also have important implications for database-based methods for fold recognition, suggesting that an unknown sequence from a plant is more likely to have a certain fold (e.g., a TIM barrel) than an unknown sequence from an animal.**

There is some evidence that there is a limited number of different protein folds (estimated to be ≈1,000) and that this "molecular parts list" is sufficient for all organisms to get on with life (1, 2). Given that this is true, one is led to ask to what degree the obvious morphological differences among organisms arise from their using different selections from this master parts list. In somewhat extreme terms, are people different from plants because they have distinctly different protein folds? On the opposite extreme, it may be that most folds occur in every organism in the same way that the genetic code and many basic biochemical pathways (such as glycolysis) are almost universally shared. Up to now, it has only been possible to address this question anecdotally in terms of individual examples. Herein we attempt a more comprehensive answer, by structurally characterizing all the known protein sequences in the databanks, i.e., by doing a structural census of the current protein universe (>140,000 sequences). Briefly, we find that the distribution of folds and structural features is different between different groups of organisms (e.g., prokaryote vs. eukaryote), a fact that has strong implications for fold recognition. However, we also find evidence that many folds are shared rather evenly among a wide variety of organisms.

Surveys of the representation of folds in the structure databank (Protein Data Bank, PDB) (3) have been reported and calculations have been done estimating the number and size of sequence families with known folds (1, 4–7). However, there have not been any studies comparing the occurrence of the known fold families among different groups of organisms. This type of comparative work has been done in studies that focused purely on sequences (8–10). For instance, it has been possible to identify sequences, called ancient conserved regions, that have been conserved over long evolutionary time scales between phylogenetically distant species (11, 12). Herein we have similar aims, but endeavor to do the work in a more structural fashion, expecting that the greater conservation of structure (as compared with sequence) and its closer relation to function will reveal more about distant evolutionary relationships (8, 13–15).

Results

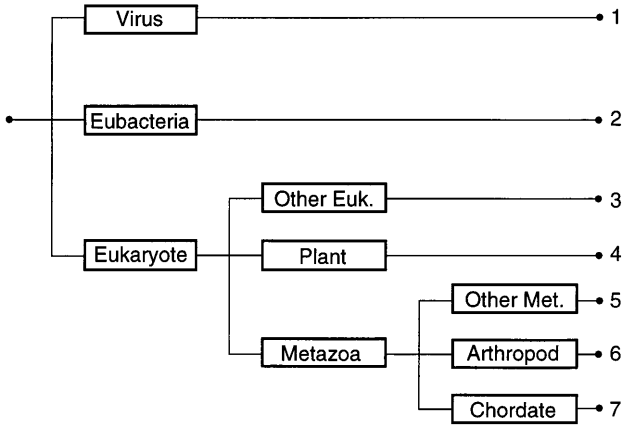
Overall Division of the Databank. Our approach is straightforward. As shown in Fig. 1, we began with all protein sequences in the publicly accessible databanks [the 142,737 sequences in the OWL composite database (17)]. We then partitioned them in two ways. First, we assigned each sequence to one of the seven groups of organisms shown in Fig. 1 and then divided the sequences into those with and without a homologue in the structure databank. All the taxonometric classification and sequence analysis was done with standard methodology [in particular, the FASTA program (22, 23) with conservative thresholds to find sequence homologues]. In the process of partitioning the sequences, we removed from our data set sequence fragments less than 40 residues, sequences that did not fit into our seven taxa (e.g., the few archaean, unclassified, and artificial sequences), and low-complexity sequences. This gave us 120,068 sequences to work with. Most of these (57%) were from eukaryotes with the remainder split between viruses and eubacteria. About 28% of these sequences had a homologue with known structure. Interestingly, eukaryotic sequences (especially chordate ones) were almost twice as likely to have a structural homologue as bacterial or viral sequences (e.g., 46% of chordate sequences had structural homologues versus 25% of bacterial sequences).

We analyzed the sequences with structural homologues in detail by using the Structural Classification of Proteins (scop) (19). This classification attempts to comprehensively systematize all known structural resemblances, many of which were originally pointed out on the basis of case-by-case observations of crystal structures [e.g., Rossmann *et al.* (47) and Harrison (48)]. In total, scop divides the 4,432 structures in the PDB into 8,330 domains, which, in turn, are classified into 318 different fold families. Thus, we were able to associate each sequence matching a structure with a particular scop "fold identifier," essentially a molecular part number, and then to see how these identifiers were distributed among our seven taxonometric categories. Other classifications of protein structure also divide the structure databank among ≈300 fold families [e.g.,

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: PDB, Protein Data Bank; scop, Structural Classification of Proteins.

†To whom reprint requests should be addressed. e-mail: Mark.Gerstein@yale.edu.



	Total	Total			Eukaryote			Metazoa		
		Virus	Eubacteria	Eukaryote	Plant	Other Euk.	Metazoa	Arthropod	Other Met.	Chordate
Total sequences (OWL 27.1)	120068	23567	27731	68770	10540	12315	45915	4276	6595	35044
Have a PDB Homologue	37706	3139	7032	27535	4960	3256	19319	1716	1435	16168
... and well-defined class	29531	2394	5489	21648	3177	2331	16140	1467	1090	13583
Observed Class										
membrane (%)	1	8	3	<1	1	1	0	0	0	0
all-α (%)	22	21	17	24	17	15	26	29	41	25
all-β (%)	35	51	20	37	23	18	43	29	12	46
α/β (%)	31	10	59	26	53	61	16	23	34	14
small (%)	10	10	1	13	6	5	15	18	14	15
No PDB Homologue	82362	20428	20699	41235	5580	9059	26596	2560	5160	18876
... but can predict class	27453	5394	8170	13889	1727	2712	9450	912	1317	7221
Predicted Class										
membrane (%)	25	17	26	27	21	36	25	16	12	29
all-α (%)	21	13	28	20	20	25	19	18	30	17
all-β (%)	3	3	1	4	1	1	6	4	2	7
α/β (%)	21	21	24	20	26	23	18	26	35	14
small (%)	30	46	22	28	32	15	31	36	21	33

FIG. 1. How the total population of sequences is divided into seven taxa in three steps is shown at the top. Other Euk. includes mostly fungi and protists, and Other Met. includes mostly nematodes. Beneath each of the taxa is shown the total number of sequences, the total with and without a structural homologue, and the total with a well-defined structural class. Below this is shown the percentage of the sequences that have each of the five well-defined structural classes.

CATH, FSSP, Entrez-MMDB, LPFC (5, 13, 49–51)] and should give similar results.

Most of our calculations were done in the most straightforward way based on exhaustive enumeration—counting everything equally. We are fully aware that such an approach tends to give results biased to some degree by the current composition of the sequence databank, i.e., toward proteins that scientists have chosen to study. There are a variety of approaches toward counting sequences (involving differential weighting or polling of selected samples, such as whole genomes) that attempt to address biases in a systematic fashion, and we discuss the application of some of these in detail. However, on a basic level, we do not believe it is possible to remove all traces of "investigator-preference" bias from any sample drawn from the current databanks. Consequently, we believe an approach of straightforward enumeration (just as in governmental censuses) provides the clearest reflection of what we currently know.

Top-10 Folds in Various Taxa. The overall distribution of folds shows that most folds have about ≈130 homologues, but there are a few folds with many more, as shown in the "top-10 lists" in Fig. 2. In particular, the top-7 folds (which include the TIM barrel, the Ig fold, the Rossmann fold, the homeo domain-like three-helix bundle, and the ferrodoxin fold) each have more than 1,000 homologues, and the top-25 folds match almost two-thirds (61%) of the sequences with structural homologues. Some of these folds, such as the nucleotide-binding Rossmann fold, the zinc finger, or the DNA-binding three-helix bundle, perform a single function and tend to be recombined as modules in a variety of proteins. Other folds act as multipurpose parts that can perform a variety of diverse functions within the same structural framework. For instance, the ribonuclease H fold can act either as a structural protein or a nuclease; the ferrodoxin fold appears in ribosomal proteins, transcription factors, and enzymes; and the Ig fold provides a scaffolding for enzymes, transcription factors, and viral envelope proteins, in addition to its well-known role in the immune system.

Many folds (125 in all) can be associated with more than one sequence family. That is, the sequences corresponding to each of these folds can be clustered into groups of similar sequences that have no detectable homology between them (see techniques section for more precise definitions). Orengo *et al.* (5) suggested that protein folds associated with many sequence

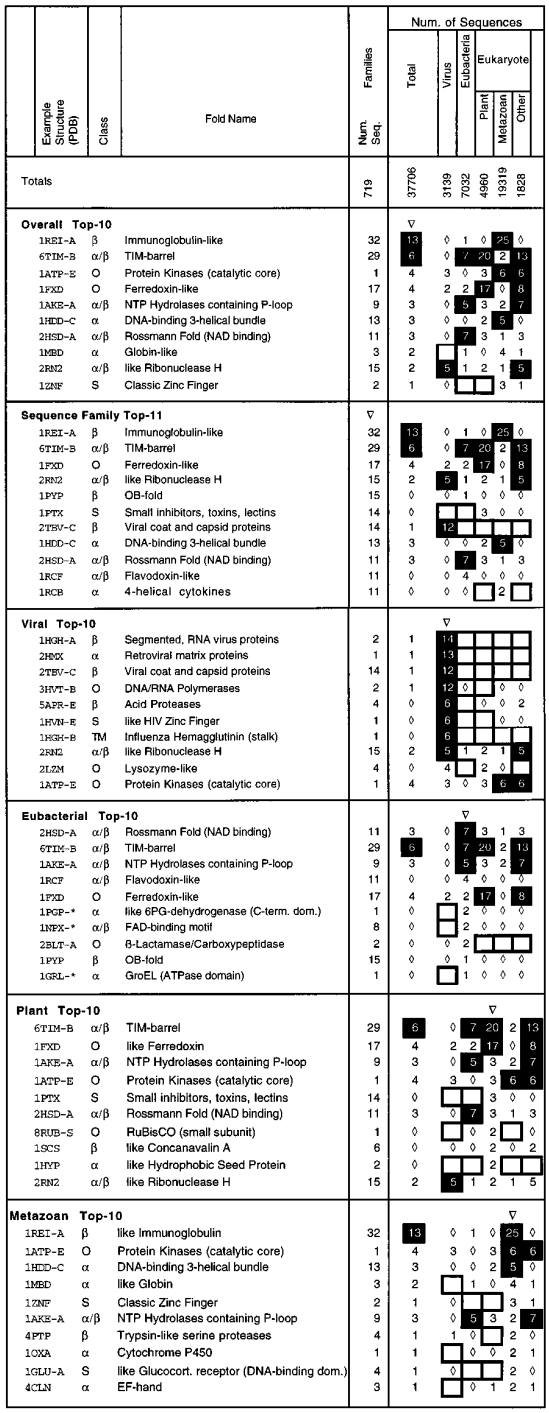


FIG. 2. Top-10 folds, overall, in terms of number of sequence families and in each of four taxa. In each of the top 10, the number of sequences with a particular fold is shown as a percentage of the total number of sequences in the corresponding taxa that have a structural homologue (this last value is shown as an absolute number at the top). Values: 0%, □; between 0% and 1%, ◇; greater than 5%, ■. Also shown in column 1 is a representative structure with that fold. (The syntax is PDB identifier followed by chain. For the three identifiers marked with an *, a particular residue selection is also necessary: for 1PGP, residues 177–473; for 1NPX, residues 120–242; for 1GRL, residues 6–136 and 410–523.) In column 2 is the structural class of the fold, derived from scop (S, is small; TM, transmembrane; O, not one of the five well-defined classes—in this case usually an $\alpha + \beta$ protein). The fold name is in column 3. In column 4, the number of sequence families is shown as an absolute number, not a percentage. This is derived from clustering the domains in the PDB with an e-value threshold of 0.001 (see techniques section). Counting only the number of clusters (i.e., families) effectively represents a particularly stringent form of sequence weighting.

families, which they dubbed "superfolds," may represent particularly favorable structural architectures, accommodating to a wide variety of sequences.

We created a new list of superfolds, by ranking the folds considered herein in terms of the number of sequence families they are associated with. [It is not possible to directly compare our most common folds to the superfolds in Orengo *et al.* (5), because they use slightly different fold definitions and because the databank has grown considerably since their work was published.] Comparing the most common folds in terms of sequence families with the most common overall indicates that most of the common folds are associated with many sequence families and that multifunction folds tend to have more associated families than single-function ones. Specifically, 7 of the top-10 folds have more than nine sequence families (with the exceptions being the globin, protein kinase, and zinc-finger folds, all of which have highly specific functions). However, the converse is not as true: 5 of the folds in the "sequence-family top 11" are not in the overall top 10. These, notably, include the OB fold and four-helical cytokine family, a four-helix bundle.

The most common folds are present in all taxa. However, the degree of their representation varies greatly. This is particularly true for the Ig fold, the most common one. It constitutes 25% of metazoan sequences with structural matches (and 40% of the human sequences), but only ≈1% of the plant, bacterial, and viral sequences. It is, consequently, of interest to look at the most common folds in each of the seven taxa, and this is shown in Fig. 2. Clearly, viruses have the most unique distribution of folds, reflecting their special functional requirements. In fact, four of their top-10 folds do not occur in any of the other six taxa. This is understandable as they are all associated with the viral envelope, which has a highly symmetrical structure unique to viruses. Viruses share with other organisms folds associated with essential viral functions (polymerases, acid proteases, and ribonucleases), but they have few of the folds associated with metabolic enzymes (e.g., TIM barrels, Rossmann folds, or NTP hydrolases). The bacterial top 10, in contrast, shows a great preponderance of folds for metabolic enzymes, in particular glycolytic enzymes. It also contains one fold unique to bacteria (and bacteriophages), that for β-lactamases and D-Ala carboxypeptidases. These enzymes perform functions associated with the unique structure of the bacterial cell wall (i.e., antibiotic resistance and cleavage of D-Ala peptides).

The top-10 folds for multicellular animals (metazoa) are very different from those for bacteria. They contain fewer folds for enzymes and more folds associated with intercellular communication, defense, and transport (e.g., EF hand, Igs, globins, protein kinases, and also within the top 15 are cysteine-knot and four-helical cytokines). There are also three folds of DNA-binding regulatory proteins and one for trypsin-like proteases, which are usually involved in extracellular digestion.

Like the metazoan top 10, the plant top 10 also contains the protein kinase fold, which is involved in signaling. However, it has many more metabolic enzymes, making it in some ways more like the bacterial top 10. It also contains a few folds unique to plants. In particular, the fold of the protein rubisco, which has a crucial role in fixing carbon in photosynthesis, is featured twice in the plant top 10—once for its small subunit, which has a fold unique to plants, and a second time for its large subunit, which contains a ferrodoxin fold.

Top-10 Folds in a Representative Genome. The list of top-10 folds in various taxa, although comprehensive, is to some degree skewed by investigator preference. We can get a sense of this bias by sampling the databanks selectively, specifically, with the sample corresponding to the entire genome of a particular organism. This is done in Fig. 3, which shows a top-10 list derived from the genome of a representative bacterium, *Haemophilus influenzae* (52). This list is clearly similar to the eubacterial top-10 list in Fig. 2, with 7 of the 10

Example Structure (PDB)	Fold Name	Percentage of known folds in genome	Rank in eubacterial Top-10
Top-10 in a bacterial genome (H. influenzae)			
2HSD-A	Rossmann Fold (NAD binding)	9.6	1
1AKE-A	NTP Hydrolases containing P-loop	5.7	3
1RCF	Flavodoxin-like	5.1	4
6TIM-B	TIM-barrel	4.5	2
1FXD	Ferredoxin-like	4.2	5
2RN2	like Ribonuclease H	3.0	16
1SBP	like Periplasmic binding protein (class II)	3.0	11
2DRI	like Periplasmic binding protein (class I)	3.0	19
1SRY-*	Class II aaRS and biotin synthetases	2.7	50
1PYP	OB-fold	2.7	9

FIG. 3. The figure shows the top-10 folds in a representative bacterial genome in a format similar to eubacterial top-10 in Fig. 2. For each of the top-10, column 1 shows a representative structure with that fold. (The syntax is PDB identifier followed by chain. For 1SRY, marked with a "*", a particular residue selection is also necessary, A:111–421.) Column 2 gives the fold name, derived from scop. Column 3 shows the number of sequences with a particular fold as a percentage of the 248 sequences in the *Haemophilus* genome (1,680 sequences in total) that have a structural homologue (using the relatively conservative thresholds described in section on sequence analysis techniques). Column 4 shows the rank of this particular fold in the eubacterial top-10 list in Fig. 2. Folds that appear in roughly the same position in both top-10 lists are shown with black boxes. The data in this table are adapted from the expanded analysis in ref. 53.

entries having similar positions. However, it is important to realize that the folds in a genome top 10 are still biased to a degree, by the selection of the representative genome itself and, more importantly, because they depend on the selection of known folds in the structure databank (i.e., in scop and the PDB). With many new genome sequences coming out, it will be possible to perform this common fold analysis, comparatively, on a number of microbial genomes. Some initial analyses show quite revealing differences (53).

A Venn Diagram for Shared Folds. The variation shown in Figs. 2 and 3 in the common folds among different taxa and selected genomes directly addresses the issue of whether the differences between organisms reflect their having fundamentally different folds. However, it only addresses this issue on a case-by-case basis, in terms of specific folds. In Fig. 4, we attempt a more systematic analysis by asking what fraction of all the known folds (in scop) are present in each of our seven taxa and what fraction of those that are present are shared among different taxa. We performed this analysis by dividing the database into subsets of progressively more related organisms (as shown in Figs. 1 and 4). The major division is between eubacterial and eukaryote sequences with the remainder of the sequences falling into a third miscellaneous division (viral sequences). Next, we divide the largest of these divisions (eukaryotes) into major and minor subsets (plants and metazoa) and a third category that includes the remaining eukaryotic sequences (mostly from protists and fungi). Finally, we again perform a three-way major-minor-miscellaneous division on the largest group of eukaryotic sequences (metazoan ones), partitioning them into chordate, arthropod, and other (mostly round worms).

We find that more related organisms have fewer folds in total but share a larger fraction of them. That is, the "top-level" division, which includes the least related organisms (bacteria, eukaryotes, and others), contains 282 folds in total, but only 18% of these (50 folds) are shared among all three subsets. The next division (plants, metazoa, and others) contains fewer folds (229), but a larger fraction of these are shared (42% or 96 folds). This trend continues in the division containing the most related taxa (chordates, arthropods, and others): these have only 191 folds in total but share 45% of them (87 folds).

If we look only at the two principal subsets at each level of division, we find that they share about half their folds (i.e., eukaryotes and eubacteria share 156 of 275, plants and metazoa share 104 of 214, and chordate and arthropod share 102 of 184). There are only 19 universal folds shared through all divisions. These include the Ig fold, the TIM barrel fold, the Rossmann fold, the ferrodoxin fold, and the ribonuclease H fold.

Characterizing Sequences Without Structural Homologues. Thus far, we have concentrated on the 37,706 sequences that have a homologue in the structure databank. What can we say about the remaining 82,362 sequences that have no homologue? By using standard methods of secondary structure prediction, we have attempted to place each of these sequences into one of five well-defined structural classes, expanded somewhat from the original class definitions in Levitt and Chothia (45): all-α, all-β, α/β, transmembrane, and small. Most of the sequences with structural homologues can be placed into these five classes by observation. However, because of our fairly strict class definitions and due to a variety of complications (in particular, the difficulty in determining domain definitions in sequences without structural homologues), we could confidently place only about a third of the sequences without a structural homologue into the five categories.

Our results, shown in Fig. 1, indicate that the proportion of proteins in the five well-defined classes varies considerably between taxonomic groups. Most importantly, the results we obtained by looking at sequences that do not have a structural homologue are consistent with the results obtained for sequences that do, suggesting that some of our firmer conclusions about the former can be reliably extrapolated to the later. In particular, we find that the percentage of small proteins is much larger in complex multicellular animals (e.g., arthropods and chordates) than in bacteria and simple eukaryotes (fungi and protists). Also, the percentage of all-β proteins is larger in viruses and eukaryotes than prokaryotes, and within eukaryotes the percentage increases as one moves from the simpler organisms to the more complex metazoans (e.g., arthropods and chordates). This may be understandable for the sequences that have a structural homologue in terms of the large number of all-β Igs in vertebrates. However, note that it is also true for the sequences that do not have structural homologues (and consequently probably do not have Ig folds).

Continuing our focus on structural class, we repeated the Venn-diagram analysis in Fig. 4, this time looking at the number of distinct folds in each taxa individually for each of the five structural classes. Except for small proteins, we found that each structural class gave essentially similar results to the overall results shown in Fig. 4. However, eukaryotes, especially chordates, had a

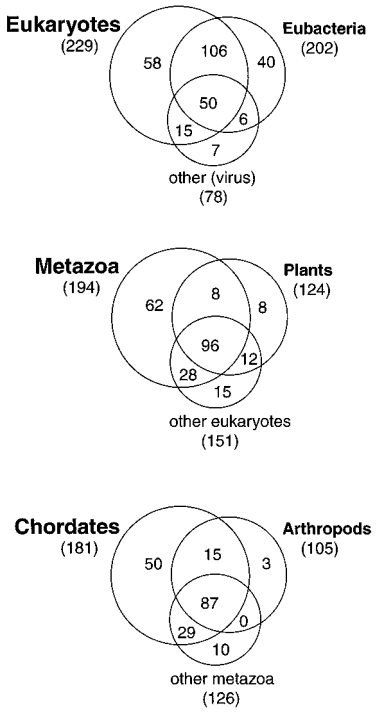


FIG. 4. Venn diagrams showing the number of folds in each group of organisms and how many of these folds are shared between different groups of organisms. Note that in total there are 318 folds in scop. However, we excluded folds associated only with membrane proteins, designed proteins, and model proteins, as well as folds only from archaea and folds not currently in the PDB, giving 282 folds. In the top-most division, these 282 folds are distributed among a major group of sequences (eukaryotes), a minor group (eubacteria), and a miscellaneous group (other and viruses). In the middle division, the major group from the top level (eukaryotic sequences) is subdivided into a major group (metazoa), a minor group (plants), and a miscellaneous group (mostly fungi and protist sequences). This pattern of major, minor, and miscellaneous division is repeated at bottom, where metazoa is subdivided into chordates, arthropods, and other (metazoa).

far greater proportion of the small folds. In particular, of the 35 small protein folds, 30 occur in eukaryotes but only 8 occur in bacteria, and of the 30 in eukaryotes, 27 occur in metazoa and 23 occur in chordates. Thus, there is prevalence of small proteins in eukaryotes, both in terms of relative numbers of sequences (with and without structural homologues) and in terms of number of folds. In some sense, this is counterintuitive as one might expect simpler smaller organisms, such as bacteria, to have more small folds. However, it is explained to some degree by the number of small protein folds involved in intercellular communication and regulation in vertebrates (e.g., insulin, kringle domains, fibronectin, or zinc fingers).

Implications, Especially for Fold Recognition

We have conducted a census of the current population of proteins. It is in a sense skewed and incomplete because we do not have all possible proteins for a given taxa. However, because the number of protein sequences is growing at a tremendous rate [more than doubling every 2 years (17)], we would expect this situation to improve to some degree in the future and we would hope that our conclusions give one a clear taste of what is to come. Furthermore, having a clear grasp of the current state of the databanks provides an important yardstick for measuring the results of future sequencing projects and for assessing the hidden biases in database-based methods for structure prediction and fold recognition.

Specifically, we have found that although there are large numbers of folds shared between organisms, different organisms have a markedly different distribution of folds. In addition to its obvious evolutionary implications, this finding is very important for approaches to fold recognition, the matching of a query sequence to a target structure, where there is no detectable homology between the sequence and the structure (29, 30). That is, our census indicates that knowing the species of an unknown sequence gives one clear clues as to its fold. For instance, there are 282 folds in the current protein universe (Fig. 4). However, *a priori* it is reasonable to rule out 80 of them (282–202) for an unknown bacterial sequence. In particular, we would not expect this unknown sequence to have many of the common eukaryotic folds associated with transcription or signaling, such as zinc fingers or EF hands. Likewise, knowing that an unknown sequence is from a plant means that it would be much more likely to have a TIM-barrel fold than if it were from an animal, in which case it would be much more likely to have an Ig fold.

Sequence Analysis Techniques Employed

A Relational Database of Folds, Sequences, and Taxa. Our census was greatly expedited by use of simple relational database implemented by using DBM and "object-oriented" PERL (version 5) (16). Relational tables linked the 142,737 sequence identifiers, the 37,706 structure matches, the 282 fold identifiers, and the 7 taxonometric ranks. We will make available over the Internet a number of these tables at the following URL: http://bioinfo.mbb.yale.edu/census.

Sequences were taken from the OWL composite database (April 1996) (17, 18) and the *Haemophilus* genome project website (www.tigr.org), structures were from the PDB, and fold defintions were from scop 1.32 (May 1996) (19, 20). We assigned specific taxonometric ranks to sequences based on the classification scheme associated with GenBank (21). All archaean, artificial, and unclassified sequences were excluded.

Sequence Comparison and Clustering into Families. All sequence matching was done with the FASTA program (version 2.0) (22–24) with k-tup 1 and an e-value cutoff of 0.001. This is a very conservative threshold, and empirical tests have shown that it should give one error every 1,000 comparisons (25, 26). Low complexity sequences were filtered out first by using the SEG program (27, 28).

There are more sensitive methods of comparing sequences to structures than the FASTA program, e.g., profiles, Hidden–Markov models, and threading (29–32). These methods would be expected to find more homologues for certain folds. However, the sensitivity improvement would not be uniform over all folds. The more sensitive methods tend to do better for large fold families (with many associated sequences) or for fold families with clearly defined sequence motifs. Thus, using these methods would bias the results even more toward highly populated and well-characterized folds. This is not advantageous because for a large-scale census, where uniform sampling and treatment of the data are more important than sensitivity (as one is more concerned with relative rather than absolute numbers).

The number of sequence families for each fold is derived from clustering all the domains in the PDB (using scop domain definitions). FASTA is used for the sequence comparisons; a pair of domains matching with an e-value of 0.001 or less is taken as connected (significantly related). Each cluster consists of the domains connected to one another by at least one linkage. This is a similar approach to that taken in Hobohm *et al.* (33) but with a somewhat different method of sequence comparison.

Sequence Weighting and Databank Sampling. We did not attempt to use explicit sequence weighting in our census [see Altschul *et al.* (34), Sander and Schneider (6), Vingron and Sibbald (35), Gerstein *et al.* (36), and Miyazawa and Jernigan (37)]. Thus, our conclusions to some degree directly reflect the biases inherent in the databanks. We feel that completely removing these biases is impossible and that assessing them is to a large degree a subjective issue [e.g., see Altschul *et al.* (34)]. Furthermore, insofar as our conclusions about the current state of the databanks reveal

bias, we feel they are useful for assessing hidden biases in database-based structure-prediction methods.

Note, however, that aspects of our census did involve four forms of implicit (and reasonable) sequence weighting. (*i*) In compiling the OWL composite databank, all mutant and identical sequences were removed. This is, in effect, a very simple type of sequence weighting that removes one of the major problems in doing calculations on the PDB, the problem of compensating for the many structures (e.g., T4 lysozyme) solved in different liganded states or as mutants. (*ii*) The enumeration of sequence families shown in Fig. 2 is a specific and particularly stringent form of sequence weighting (see above for the method). It greatly down weights highly homologous sequences, giving all the sequences in a family, even a large one, an aggregate weight of 1.0. (*iii*) Many of the conclusions in Fig. 2 and all the conclusions in Fig. 4 are completely independent of sequence weighting because they are only concerned with membership, whether or not a given fold is present in a particular taxa. (*iv*) Finally, the genome top-10 list in Fig. 3 is constructed from a complete genome sequence that is not biased by the preferences of investigators to sequence proteins of functional importance. However, it is still skewed by the selection of the known structures in the PDB matched against the genome.

For the numbers reported herein that are the result of exhaustive enumeration—counting everything—we found that we could achieve essentially the same results through randomly sampling (i.e., polling) small subsets, the same approach that has been argued to be effective in the American governmental census (38).

Class Prediction. For the class predictions, we used a standard "off-the-shelf" approach. We first divided all sequences on the basis of length. Those with less than 40 residues were excluded, and those with between 40 and 80 residues were classed as small. For sequence with more than 80 residues, we applied the following protocol: Based on the annotations [from Swiss-Prot (39)], we decided whether or not a sequence was transmembrane. By doing this, we found that only about 10% of the sequences without structural homologues are transmembrane. This is probably an underestimate, and we tried to assess its magnitude by testing each sequence with the Kyte–Doolittle and GES hydropathy scales (40–42). By using a strict threshold, we found 5% of the sequences to be transmembrane but by using the lax one, we found that 30% were (but with many documented false positives).

After removing small and transmembrane proteins, we ran the GOR program for secondary-structure prediction (43). We used commonly accepted thresholds (44) for placing a protein in the various classes: all-α has $\alpha > 40\%$ and $\beta < 5\%$, all-β has $\beta > 40\%$ and $\alpha < 5\%$, and α/β has $\alpha > 30\%$ and $\beta > 20\%$. Sequences that did not fit in any of the previous classes were considered not to have a "well-defined" class. Note this includes (*i*) sequences with a structural homologue where the structure has the $\alpha + \beta$ class (45), (*ii*) sequences without a structural homologue that code for single-domain proteins naturally falling into the $\alpha + \beta$ class, and (*iii*) sequences without a structural homologue that code for multidomain proteins where each domain has a well-defined class (e.g., all-α and all-β) but where the protein is considered as whole for class prediction. We tested our class predictions against the observed classes in scop and found about 80% agreement. We also tested our structural class predictions by comparing a sample of them with the results of running the PHD server (46) and found substantial agreement.

We thank L. Stryer, A. Murzin, S. Brenner, C. Chothia, T. Johnson, and E. Brodkin for helpful conversations or reading the manuscript. M.G. acknowledges the Office of Naval Research for support (Young Investigator Grant N00014–97-1–0725) and M.L. acknowledges the Department of Energy (DOE DE-FG03-95ER62135).

1. Chothia, C. (1992) *Nature (London)* **357,** 543–544.
2. Dorit, R. L., Schoenbach, L. & Gilbert, W. (1990) *Science* **250,** 1377–1382.
3. Bernsteinm, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542.
4. Holm, L. & Sander, C. (1996) *Science* **273,** 595–602.
5. Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994) *Nature (London)* **372,** 631–634.
6. Sander, C. & Schneider, R. (1991) *Proteins Struct. Funct. Genet.* **9,** 56–68.
7. Pascarella, S. & Argos, P. (1992) *Protein Eng.* **5,** 121–137.
8. Doolittle, R. F. (1995) *Annu. Rev. Biochem.* **64,** 287–314.
9. Koonin, E. V., Tatusov, R. L. & Rudd, K. E. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 11921–11925.
10. Ouzounis, C. & Kyrpides, N. (1996) *FEBS Lett.* **390,** 119–123.
11. Green, P., Lipman, D., Hillier, L., Waterston, R., States, D. & Claverie, J. M. (1993) *Science* **259,** 1711–1716.
12. Green, P. (1994) *Curr. Opin. Struct. Biol.* **4,** 404–412.
13. Gibrat, J. F., Madej, T. & Bryant, S. H. (1996) *Curr. Opin. Struct. Biol.* **6,** 377–385.
14. Chothia, C. & Lesk, A. M. (1986) *EMBO J.* **5,** 823–826.
15. Chothia, C. & Gerstein, M. (1997) *Nature (London)* **385,** 579–581.
16. Wall, L., Christiansen, D. & Schwartz, R. (1996) *Programming Perl* (O'Reilly and Associates, Sebastapol, CA).
17. Bleasby, A. J., Akrigg, D. & Attwood, T. K. (1994) *Nucleic Acids Res.* **22,** 3574–3577.
18. Bleasby, A. J. & Wootton, J. C. (1990) *Protein Eng.* **3,** 153–159.
19. Murzin, A., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536–540.
20. Brenner, S., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* **266,** 635–642.
21. Benson, D. A., Boguski, M., Lipman, D. J. & Ostell, J. (1996) *Nucleic Acids Res.* **24,** 1–5.
22. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227,** 1435–1441.
23. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.
24. Pearson,W. R. (1996) *Methods Enzymol.* **266,** 227–259.
25. Brenner, S., Hubbard, T., Murzin, A. & Chothia, C. (1995) *Nature (London)* **378,** 140.
26. Brenner, S., Chothia, C. & Hubbard, T. (1997) *Proc. Natl. Acad. Sci. USA*, in press.
27. Wootton, J. C. & Federhen, S. (1993) *Computers and Chemistry* **17,** 149–163.
28. Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266,** 554–571.
29. Jones, D. T. & Thornton, J. M. (1996) *Curr. Opin. Struct. Biol.* **6,** 210–216.
30. Bowie, J. U. & Eisenberg, D. (1993) *Curr. Opin. Struct. Biol.* **3,** 437–444.
31. Eddy, S. R. (1996) *Curr. Opin. Struct. Biol.* **6,** 361–365.
32. Gribskov, M., Lüthy, R. & Eisenberg, D. (1990) *Methods Enzymol.* **183,** 146–159.
33. Hobohm, W., Scharf, M., Schneider, R. & Sander, C. (1992) *Protein Sci.* **1,** 409–417.
34. Altschul, S. F., Carroll, R. J. & Lipman, D. J. (1989) *J. Mol. Biol.* **207,** 647–653.
35. Vingron, M. & Sibbald, P. R. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 8777–8781.
36. Gerstein, M., Sonnhammer, E. & Chothia, C. (1994) *J. Mol. Biol.* **236,** 1067–1078.
37. Miyazawa, S. & Jernigan, R. L. (1996) *J. Mol. Biol.* **256,** 623–644.
38. Ladd, E. C. Tempest in a Census. *Wall Street Journal,* 30 July 1997, A14.
39. Bairoch, A. & Boeckmann, B. (1992) *Nucleic Acids Res.* **20,** 2019–2022.
40. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157,** 105–132.
41. Engelman, D. M., Steitz, T. A. & Goldman, A. (1986) *Annu. Rev. Biophys. Biophys. Chem.* **15,** 321–353.
42. Jähnig, F. (1990) *Trends Biochem. Sci.* **15,** 93–95.
43. Garnier, J., Gibrat, J. F. & Robson, B. (1996) *Methods Enzymol.* **266,** 540–553.
44. Rost, B. (1996) *Methods Enzymol.* **266,** 525–539.
45. Levitt, M. & Chothia, C. (1976) *Nature (London)* **261,** 552–558.
46. Rost, B. & Sander, C. (1993) *J. Mol. Biol.* **232,** 584–599.
47. Rossmann, M. G., Liljas, A., Branden, C. I. & Banaszak, L. J. (1975) *Enzymes* **11,** 61–102.
48. Harrison, S. C. (1991) *Nature (London)* **353,** 715–719.
49. Holm, L. & Sander, C. (1994) *Nucleic Acids Res.* **22,** 3600–3609.
50. Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993) *Protein Eng.* **6,** 485–500.
51. Schmidt, R., Gerstein, M. & Altman, R. (1997) *Protein Sci.* **6,** 246–248.
52. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al*. (1995) *Science* **269,** 496–512.
53. Gerstein, M. (1997) *J. Mol. Biol.*, in press.