Comparing function and structure between entire proteomes

JINFENG LIU[1,2] AND BURKHARD ROST[1]

[1]CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10032, USA
[2]Graduate Program in Pharmacology, Columbia University, New York, New York 10032, USA

(RECEIVED March 15, 2001; FINAL REVISION July 6, 2001; ACCEPTED July 12, 2001)

Abstract

More than 30 organisms have been sequenced entirely. Here, we applied a variety of simple bioinformatics tools to analyze 29 proteomes for representatives from all three kingdoms: eukaryotes, prokaryotes, and archaebacteria. We confirmed that eukaryotes have relatively more long proteins than prokaryotes and archaes, and that the overall amino acid composition is similar among the three. We predicted that ~15%–30% of all proteins contained transmembrane helices. We could not find a correlation between the content of membrane proteins and the complexity of the organism. In particular, we did not find significantly higher percentages of helical membrane proteins in eukaryotes than in prokaryotes or archae. However, we found more proteins with seven transmembrane helices in eukaryotes and more with six and 12 transmembrane helices in prokaryotes. We found twice as many coiled-coil proteins in eukaryotes (10%) as in prokaryotes and archaes (4%–5%), and we predicted ~15%–25% of all proteins to be secreted by most eukaryotes and prokaryotes. Every tenth protein had no known homolog in current databases, and 30%–40% of the proteins fell into structural families with >100 members. A classification by cellular function verified that eukaryotes have a higher proportion of proteins for communication with the environment. Finally, we found at least one homolog of experimentally known structure for ~20%–45% of all proteins; the regions with structural homology covered 20%–30% of all residues. These numbers may or may not suggest that there are 1200–2600 folds in the universe of protein structures. All predictions are available at http://cubic.bioc.columbia.edu/genomes.

Keywords: Protein sequence analysis; analyzing entire genomes; helical membrane proteins; coiled-coil proteins; signal peptides; comparative modeling

Supplemental material: See www.proteinscience.org.

Comparative genomics begins with collecting and describing. Sequencing the entire genome of the first free-living organism, *Haemophilus influenzae*, opened the new era of flooding data in molecular biology (Fleischmann et al. 1995). Since then, over 40 genomes have been sequenced, mostly for pathogens and model organisms. These include the first eukaryotic genome, *Saccharomyces cerevisiae* (1997), and the first animal genomes, *Caenorhabditis elegans* (1998), *Drosophila melanogaster* (Adams et al. 2000), and *Homo sapiens* (The Genome International Sequencing Consortium 2001; Venter et al. 2001). What can we learn from all the data? Like zoology and botany a century ago, we are just commencing to catalog the com-

Reprint requests to: Burkhard Rost, CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168 Street, BB217, New York, New York 10032, USA; e-mail: rost@columbia.edu; fax: (212) 305-7932.

Abbreviations: 3D structure, three-dimensional structure (i.e., coordinates of all residues/atoms in a protein); `COILS`, prediction of coiled-coil regions from sequence based on statistics and expert rules; ORF, open reading frame (protein predicted by genome-sequencing project); PDB, protein data bank of protein structures; `PHDhtm`, profile-based neural network prediction of transmembrane helices; `PSI-BLAST`, fast and reliable database search method; `SignalP`, neural network based prediction of signal peptides; SWISS-PROT, curated database with protein sequences and functional annotations; TM, transmembrane helices; TrEMBL, automatic translation of EMBL nucleotide database of protein sequences.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1101/ps.10101.

ponents of life while trying to find common features and systematic schemes. Thus, comparative genomics in its infancy is confined to describing similarities and differences. So far, synopses of the whole genome have focused on archiving the functional and the structural content of organisms (Koonin et al. 1996; Odgren et al. 1996; Tamames et al. 1996; Frishman and Mewes 1997; Gerstein and Levitt 1997; Wallin and von Heijne 1998; Zhang et al. 1998). Classifying proteins according to functional criteria is difficult because function is a complex phenomenon associated with many mutually overlapping levels: chemical, biochemical, cellular, physiological, organism mediated, and developmental. These levels are related in complex ways. For example, protein kinases can be related to different cellular functions (such as cell cycle) and to a chemical function (transferase), plus a complex control mechanism by interaction with other proteins.

Bioinformatics identifies most helical membrane proteins. Membrane proteins are crucial for survival; one reason is that they mediate communication across the cell membrane. Despite the great biological and medical importance, we still have very little experimental information about the 3D structures: <1% of the proteins of known structure are membrane proteins. In contrast, helical membrane proteins are relatively easy to identify by bioinformatics. How many helical membrane proteins are in a genome (Goffeau et al. 1993; Boyd et al. 1998; Wallin and von Heijne 1998)? Is the fraction of helical membrane proteins constant, or does the fraction of helical membrane proteins correlate with the complexity of the organism (Wallin and von Heijne 1998)? Are there preferences for particular numbers and topologies of transmembrane helices (TM) in some organisms (Jones 1998)? Some analyses concluded that there was no preference for proteins with a certain number of membrane helices (Arkin et al. 1997; Gerstein and Hegyi 1998), whereas others reported some preferences (Jones 1998; Wallin and von Heijne 1998).

Many extracellular proteins were identified through signal peptides. Signal peptides at the N-terminal end target many prokaryotic and eukaryotic proteins to the secretory pathway (Cleves 1997; Nielsen et al. 1997; Nakai 2000; Thanassi and Hutltgren 2000). Signal peptides are predicted accurately (Nielsen et al. 1997; Schneider 1999; Emanuelsson et al. 2000). However, although secreted proteins were studied in various bacteria (Schneider 1999), few groups analyzed entirely sequenced eukaryotes.

Coiled-coil proteins are the most heterogeneous identifiable class. Coiled-coils are typically formed as bundles of several right-handed alpha helices twisted around each other, forming a left-handed super helix (Crick 1953; Lupas 1996a). Coiled-coil structures are often used to mediate protein–protein interaction or to build filaments and other macroscopic structures. Most known coiled-coil proteins can be detected based on particular sequence signals (Lupas 1997). Thus, we can identify most coiled-coil proteins in a proteome. Six percent of all proteins in GenBank (Benson et al. 2000) appear to contain coiled-coil regions, and the percentages appear to vary between organisms (Odgren et al. 1996). Does this finding hold up for entire proteomes?

Here, we analyzed predictions for transmembrane helices, coiled-coil regions, and signal peptides for 28 entire proteomes and for 24,000 human proteins. Because of the comprehensive size of the data, we reanalyzed the distribution of protein length explored by others (Das et al. 1997; Gerstein and Hegyi 1998).

Results and Discussion

Simple biophysical criteria: Protein length and amino acid composition

Protein lengths differ among kingdoms. Generally, prokaryotes and archaes appeared to have an asymmetric bell-shape distribution of ORF lengths with a peak of ~100–300 residues, whereas the ORFs of eukaryotes were distributed much more evenly within the range of 100–600 residues (Fig. 1A). This was also reflected in the cumulative distributions (Fig. 1A, inset): The distributions for prokaryotes and archaes were steepest at ~300 residues, whereas those for eukaryotes were relatively constant between 100 and 600 residues. Furthermore, every tenth eukaryotic protein had >1000 residues, whereas <5% of the ORFs in prokaryotes and archaes were as long (Fig. 1A, inset). The length of ORFs is reported to follow an extreme value distribution (Gerstein 1998). Although our data agreed with this (Fig. 1A, bold lines), a detailed statistical analysis did not support the hypothesis for all three kingdoms. In particular, eukaryotes deviated significantly from an extreme value distribution.

Amino acid compositions did not differ between proteomes. The codon usage differs among the genomes we analyzed. In contrast, amino acid compositions were rather similar (Fig. 1B). Leucine (L), valine (V), serine (S), and alanine (A) were the most abundant amino acids, and cysteine (C), tryptophan (W), histidine (H), and methionine (M) the most underrepresented. The only differences were that eukaryotes tended to have fewer alanines (A) and more asparagines (N), and archaes had fewer glutamines (Q).

Prediction-based classifications of proteomes

Errors of prediction methods affected averages marginally. The following results are based on predictions that may be wrong. In particular, PHDhtm (Rost et al. 1996) may have missed ~5% of the helical membrane proteins, and may have falsely predicted membrane helices in ~3% of all globular proteins. We adjusted the number of predicted membrane proteins accordingly (see equation 1 below). The

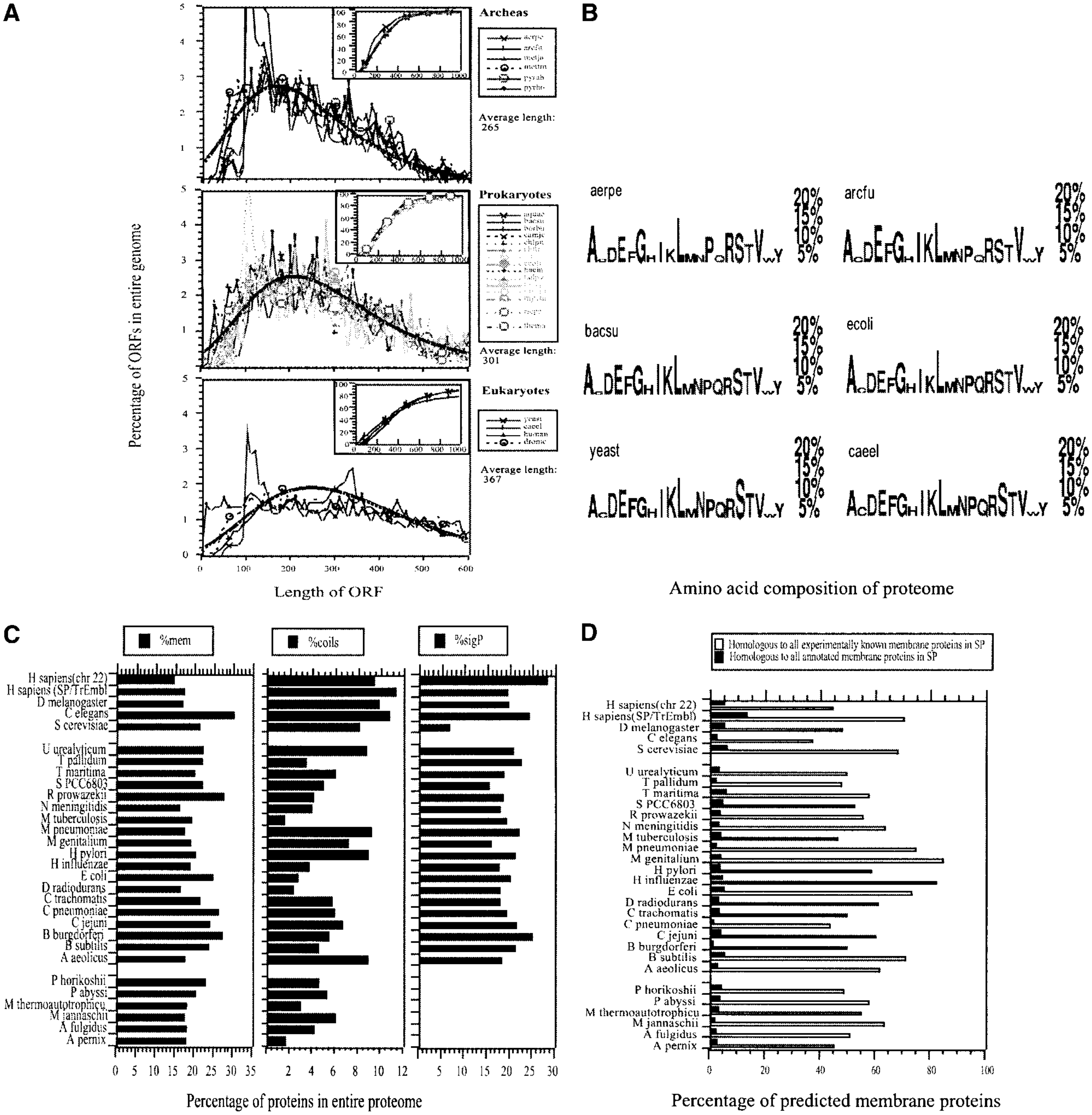

Fig. 1. (*A*) The distribution of length of ORFs (in bins of 10 residues) in 29 genomes (cumulative values in the insets). The extreme value distribution fit is shown in bold. The abbreviations for the organisms are given in Table 1. (*B*) Amino acid composition for six representative genomes: The letter height is proportional to the observed composition of the respective amino acid (one-letter code). (*C*) Percentage of membrane proteins, coiled-coil proteins, and proteins with signal peptides in 29 genomes. (*D*) Less than half of the predicted membrane proteins could have been identified through homology with known membrane proteins.

accuracy of `SignalP` was estimated to be ~90% (Nielsen et al. 1997; Emanuelsson et al. 2000), implying that the predicted compositions underestimated the actual percentage of secreted proteins. For `COILS` (Lupas 1996b), we used a threshold assuring high accuracy (probability > 0.9). At this level, the prediction would have 95% coverage for parallel coiled coils and 80% for antiparallel coiled coils (A. Lupas, pers. comm.). When transferring the functional assignment of SWISS-PROT, we used a threshold estimated to yield ~70% accuracy (Devos and Valencia 2000). Be-

Table 1. *Genomes analyzed*

Latin name	Abbreviation[a]	No ORF[b]
Archael bacteria		
Aeropyrum pernix K1	aerpe	2694
Archaeoglobus fulgidus	arcfu	2383
Methanococcus jannaschii	metja	1735
Methanobacterium thermoautotrophicum	metth	1871
Pyrococcus abyssi	pyrab	1765
Pyrococcus horikoshii	pyrho	2064
Prokaryotes		
Aquifex aeolicus	aquae	1522
Bacillus subtilis	bacsu	4099
Borrelia burgdorferi	borbu	850
Campylobacter jejuni	camje	1731
Chlamydia pneumoniae	chlpn	1052
Chlamydia trachomatis	chltr	894
Deinococcus radiodurans	deira	3103
Escherichia coli	ecoli	4285
Haemophilus influenzae	haein	1716
Helicobacter pylori	helpy	1788
Mycoplasma genitalium	mycge	470
Mycoplasma pneumoniae	mycpn	677
Mycobacterium tuberculosis	myctu	3918
Neisseria meningitidis	neime	2081
Rickettsia prowazekii	ricpr	834
Synechocystis PCC6803	syny3	3169
Thermotoga maritima	thema	1846
Treponema pallidum	trepa	1031
Ureaplasma urealyticum	ureur	613
Eukaryotes		
Caenorhabditis elegans	caeel	18944
Drosophila melanogaster	drome	14218
Subset of Homo sapiens[c]	human	24235
Homo sapiens (chromosome 22)	hw22	887
Saccharomyces cerevisiae	yeast	6307

[a] Abbreviations: Taken from SWISS-PROT.

[b] No ORFs: The number of open reading frames (predicted proteins) is taken from the respective original publication.

[c] Human sequences: The only noncomplete data were the human sequences, taken from SWISS-PROT release 39 (Bairoch and Apweiler 2000) and from TrEMBL release 15 (Bairoch and Apweiler 2000).

Note: All sequences used are available on our Web site (Liu and Rost 2000).

cause this error seems not to be class-specific, we expect the relative proportions to be relatively accurate.

Multicellular organisms appeared to have membrane content similar to unicellular organisms. For most organisms, <25% of all ORFs encoded helical membrane proteins (Fig. 1C). Whereas we predicted the highest content for *C. elegans* (30%), *Drosophila* appeared to have <18%; three prokaryotes had >25%: *Rickettsia prowazeki*, *Borrelia burgdorferi*, and *Chlamydia trachomatis*. Two archaes had >20%: *Pyrococcus horikoshii* and *Pyrococcus abyssi*. Of all the membrane proteins identified by PHDhtm, 40%–80% could have been detected by homology with known membrane proteins (Bairoch and Apweiler 2000). Excluding proteins annotated as Putative, Possible, Probable, or By Similarity in SWISS-PROT, the number of previously known membrane proteins dropped to 1%–7%; the exception: human with 14% (Fig. 1D). Thus, we could not verify that more complex organisms need larger fractions of membrane proteins (Goffeau et al. 1993; Boyd et al. 1998; Wallin and von Heijne 1998). The discrepancy probably resulted from the insufficient amount of human and fly data available earlier.

Number of transmembrane helices differed among kingdoms. We confirmed that most membrane proteins have <4 helices (Fig. S-1A S-1 in electronic supplementary material). In contrast to previous results (Arkin et al. 1997), we found that 7-TM proteins were significantly over represented in *C. elegans* and *H. sapiens*, as were proteins with six and 12 TM in most prokaryotes (Fig. 2A). These three classes were also the only ones with an imbalance in the distribution between the two possible orientations of the transmembrane helices: Eukaryotic 7-TM proteins were dominated by topology "in" (G protein-coupled receptors) and 6- and 12-TM proteins in prokaryotes by the topology "out" (transporters). Surprisingly, we found relatively few 7-TM proteins in fly (Vosshall et al. 1999, 2000) and many in worm (Fig. 2A). The worm appears to contain 1000 smell receptors, whereas the fly has <100 (Vosshall et al. 1999, 2000). Thus, the difference in this single class of proteins might explain the observed differences.

Many membrane proteins had almost no globular regions. Relating protein length to the number of transmembrane helices, we observed two clusters (Figs. 2B, S-1B). One contained proteins of varying length with one helix; the other was populated by proteins with ~35 residues per helix. Most transmembrane helices spanned over 17–33 residues. Thus, the second cluster contained almost no globular regions. These data confirmed earlier findings (Wallin and von Heijne 1998). Long nontransmembrane regions in membrane proteins are likely to form structurally compact globular domains; most of these were longer than 100 residues and were as often inside as they were outside of the membrane (Fig. S-2 S-2).

Almost every tenth eukaryotic protein contained coiled-coil regions. We found coiled-coil regions for ~8%–11% of all eukaryotic and 2%–9% of all prokaryotic and archae proteins (Fig. 1C). Most eukaryotes had more coiled-coil proteins than prokaryotes, and most prokaryotes more than archaes. Exceptions were *Mycoplasma pneumoniae*, *Helicobacter pylori*, and *Aquifex aeolicus*, which all had higher percentages of coiled-coil proteins than *C. elegans*. Surprisingly low contents of coiled-coil proteins appeared in *Deinococcus radiodurans*, *Mycobacterium tuberculosis*, and *Aeropyrum pernix K1*. A previous survey of GenBank suggested that 4.96% of bacterial, 9.47% of invertebrate, and 6.80% of vertebrate proteins have coiled-coil regions (Odgren et al. 1996). We could not find significant differ-

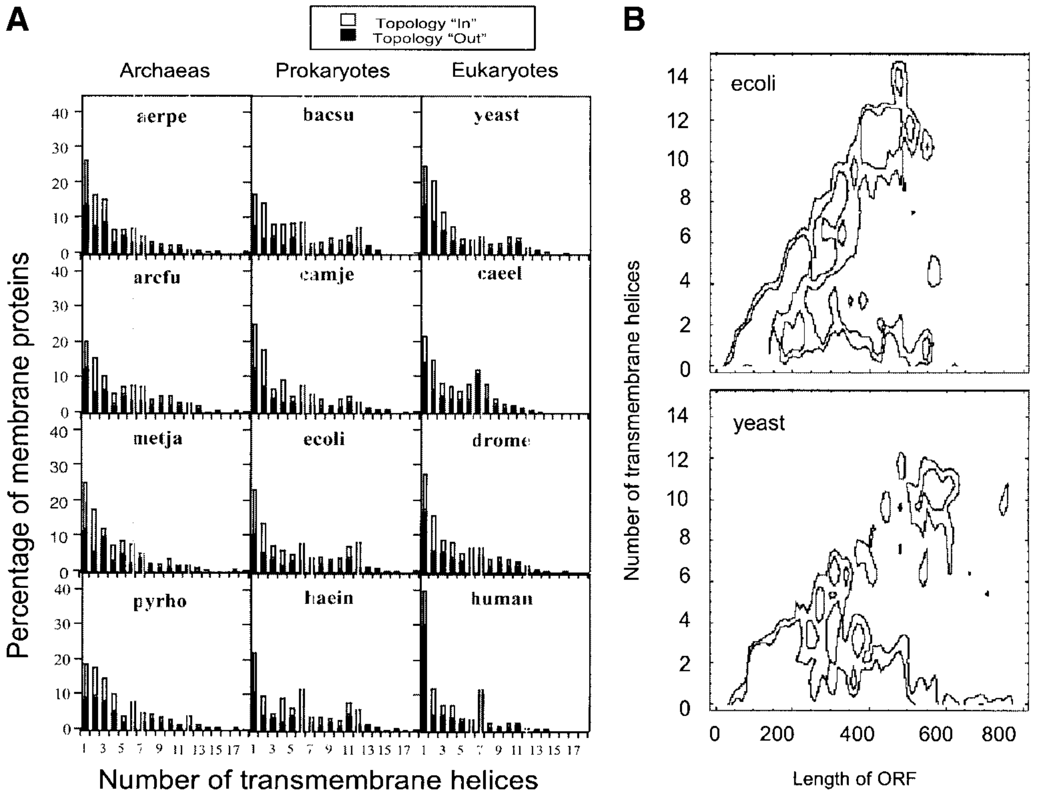

Fig. 2. (*A*) Fraction of membrane proteins with different numbers of predicted transmembrane segments. White bars: proteins with topology "in." Black bars: proteins with topology "out." (*B*) Contour plot showing the relation between ORF length (in bins of 10 residues) and the number of predicted membrane helices for two representative organisms.

ences between vertebrate (human) and invertebrates (*C. elegans* and *Drosophila*). The vast majority of coiled-coil proteins contained only one long helix and most coiled-coil regions extended over 28 residues (Fig. S-3 S-3). Protein length did not correlate with the number of coiled-coil regions (Fig. S-3C). As expected, the amino acid composition of coiled-coil regions (Fig. S-3D) differed from that of all other regions (Fig. 1B).

Prokaryotes and eukaryotes had similar fractions of secreted proteins. We predicted 15%–25% of the ORFs in prokaryotes and eukaryotes to have signal peptides (Fig. 1C). The exception was yeast for which we found only 7% secreted proteins. Supposedly our estimates constitute lower bounds because of the fact that the prediction program, `SignalP`, misses secreted proteins.

Too many functionally unclassified proteins hampered comparing function. Using `EUCLID` (Tamames et al. 1996, 1998), we could classify ~45%–65% of all proteins into one of 13 functional classes at a level reported to yield 70% correct classifications with >30% pairwise sequence identity (Devos and Valencia 2000). When grouping the 13 classes into three super-classes, energy, information and communication (Tamames et al. 1996), we found similar compositions within the archaen and eukaryotic kingdoms (Fig. 3A). In contrast, the composition varied significantly between prokaryotes: For *Escherichia coli* and *Synechocystis PCC6803*, the composition resembled that of eukaryotes; for *Aquifex aeolicus* and *Thermotoga maritima*, that of archeas (Fig. 3A). Finally, we found the following differences in the 13 classes. Amino acid biosynthesis, Biosynthesis of cofactors, prosthetic groups, and carriers, and Energy metabolism were abundant in prokaryotes; human seemed to have a larger portion of the classes Transport and binding and Regulatory functions (Fig. 3B). Previously, bacteria were reported to have smaller fractions of proteins responsible for communication (5%) than plants (20%) and animals (45%), possibly because the complex organization of multicellular organisms required more proteins communicating with the environment (Tamames et al. 1996). Our data differed in that we found bacteria to contain more proteins associated with communication (15%–32%) than reported previously (Fig. 3A). The inherent bias of SWISS-PROT may have contributed most to these differences. The significant variation of the class composition between the various prokaryotic genomes may reflect the very different environments in which these organisms dwell. However, the most important result was that, although accepting classification errors of 30% or more, we still could classify only

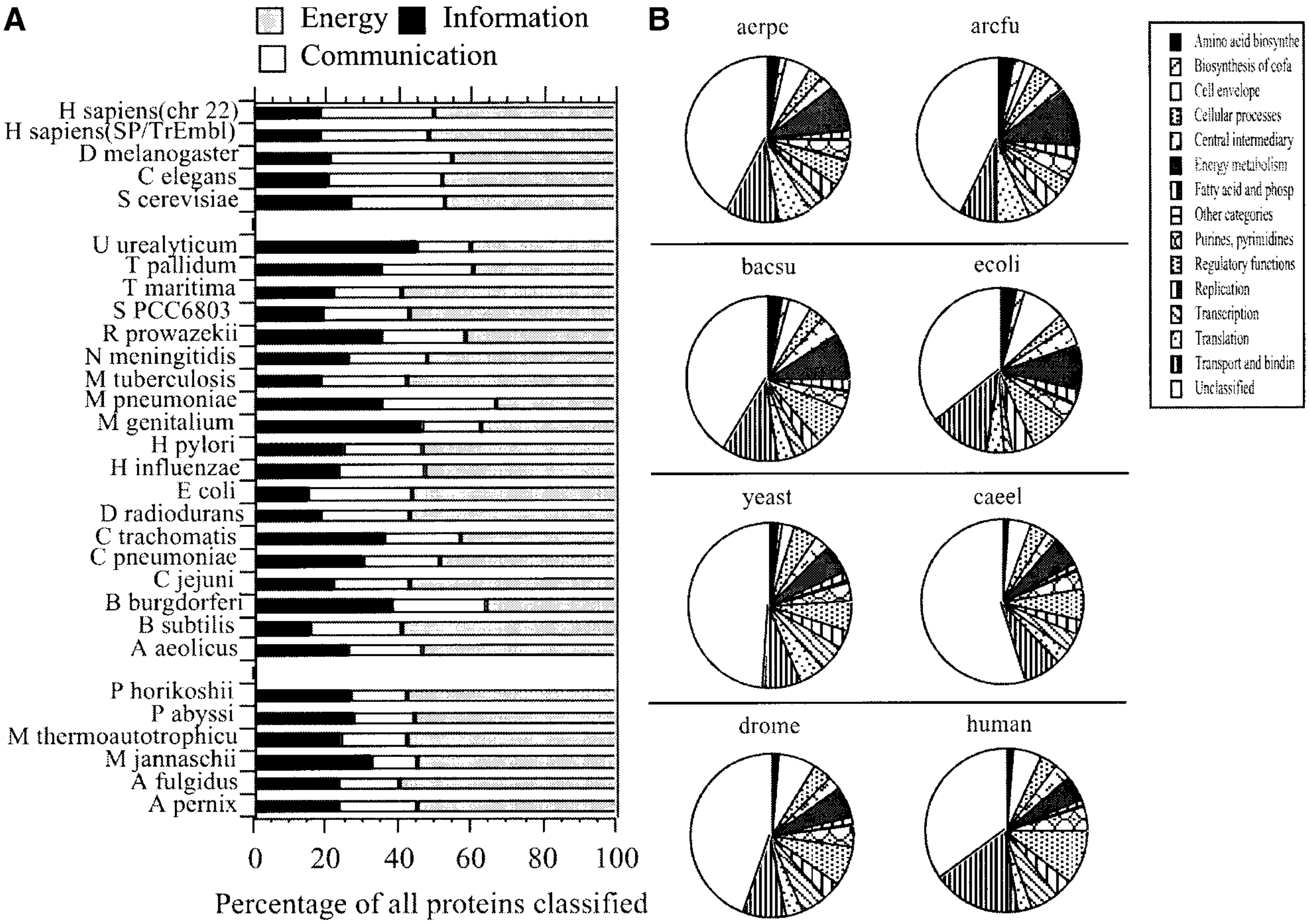

Fig. 3. Functional classification of genomes. (*A*) Super-class distribution for the genomes: We grouped the 14 EUCLID classes into Energy, Communication, and Information super-classes. (*B*) Distribution of 13-category classification for selected genomes, including those without functional classification.

about half of all proteins. Thus, conclusions about the meaning of the relative proportions remain highly speculative.

Of the proteins, ~10% were orphans. We examined the size of all families with proteins of similar structure (structural families) found in SWISS-PROT / TrEMBL (Bairoch and Apweiler 2000). We found that most proteomes contain 5%–10% orphan families; a few organisms deviated significantly from this level (Fig. 4A). Another 10% of the proteins had only one homolog (Fig. 4A). Overall, 30% of the proteins were in families with <10 proteins and 30%–40% in families with >100 members (Fig. 4B).

For 60% of all residues, we could not predict protein structure through comparative modeling. Comparative modeling could predict 3D structure at very low resolution for ~25%–40% of all proteins (Fig. 5A). Exceptions were the biased subset of human sequences found in SWISS-PROT and TrEMBL (>50%) and the proteome of *Aeropyrum pernix K1* (<19%). These numbers comprised the most optimistic values, in that we applied an iterated PSI-BLAST search (see Materials and Methods), which may yield a considerable number of false positives (Lindahl and Elofsson 2000). Using a more conservative cutoff for pairwise comparisons, we still found structural similarities in ~20%–35% of all proteins. The similarity to regions of experimentally known structure covered ~17%–36% of the entire residue mass of all proteomes (Fig. 5B). Large-scale initiatives in structural genomics aim at experimentally determining all protein structures (Rost 1998; Sali 1998; Burley et al. 1999; Shapiro and Harris 2000). Obviously membrane proteins, as well as other nonglobular proteins, will be left out in this search to cover protein structure space. Today we have experimental information about structure for <1% of the proteomes we analyzed. However, through comparative modeling, we could obtain good structure predictions, i.e., <3 Å rmsd for main chain (Eyrich et al. 2001; Marti-Renom et al. 2001), for ~6% of all proteins (Liu and Rost 2001). When we relaxed model accuracy to a level at which most models will be better than 6 Å rmsd (Marti-Renom et al. 2001), we found that ~25%–40% of all ORFs were similar to a protein

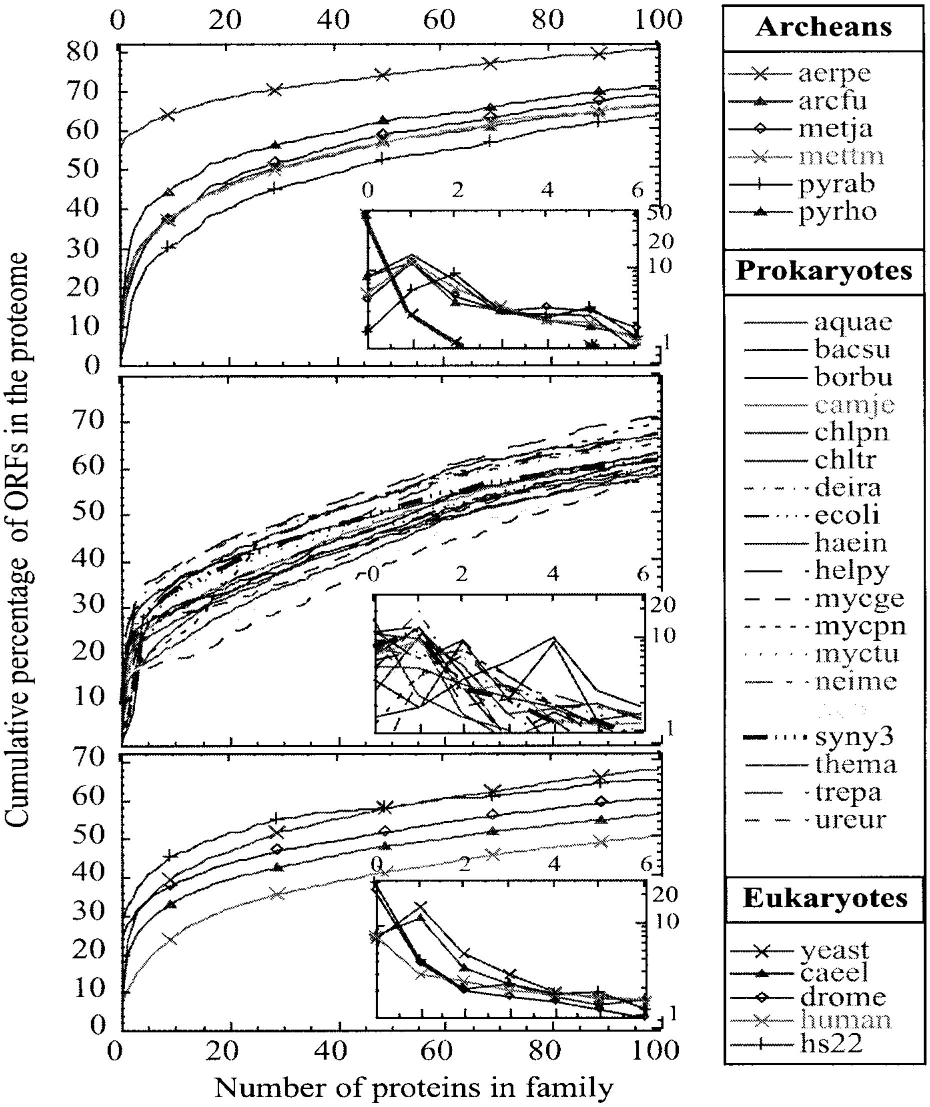

Fig. 4. For each protein in all proteomes, we counted the number of proteins found in the respective family at a PSI-BLAST E value < 10^{-3}. The graphs show the cumulative percentages of proteins found in families of particular sizes. For example, ~5%–10% of all ORFs were orphans, i.e., had no homolog in current databases; 30%–40% were in families with > 100 members.

of known structure (Fig. 5A). At this low level of accuracy, ~26% of all eukaryotic residues could be modeled. We found ~4% of the residues in transmembrane regions, ~2% in coiled-coil regions, and 11% in long regions that lack regular secondary structure (J. Liu and B. Rost, unpubl.). Thus, we estimated that structural genomics would have to cover ~60% of all the proteomes. Obviously, many of these proteins/fragments belong to the same structural families (Fig. 4). In fact, we found the 40,000 proteins from fly, worm, and yeast to cluster into ~17,000 structural families (data not shown). Assuming a similar reduction, structural genomics will have to experimentally determine structures for about one-fourth of all the proteomes we analyzed.

How many folds are there? Chothia (1992) estimated that the universe of proteins contains ~1000 different folds. Depending on what we consider different folds, we currently know of ~600–800 folds (Holm and Sander 1999; Orengo et al. 1999; Lo Conte et al. 2000). We found that these folds span ~25%–45% of the proteomes we analyzed. A more detailed analysis of regions with missing structural information (Liu and Rost 2001) suggested that ~15%–20% of the entire residue mass will not correspond to folds. Thus, we need folds for 80%–85% of all proteins. Of these, we currently cover 30%–50% with 600–800 folds. Hence, if we assume that today's folds and proteomes are representative of the universe of life, we estimate ~1200–2600 folds in the universe. However, we have many good reasons to assume that today's databases are biased and that our perspective of the world is too narrow to thoroughly support such conclusions.

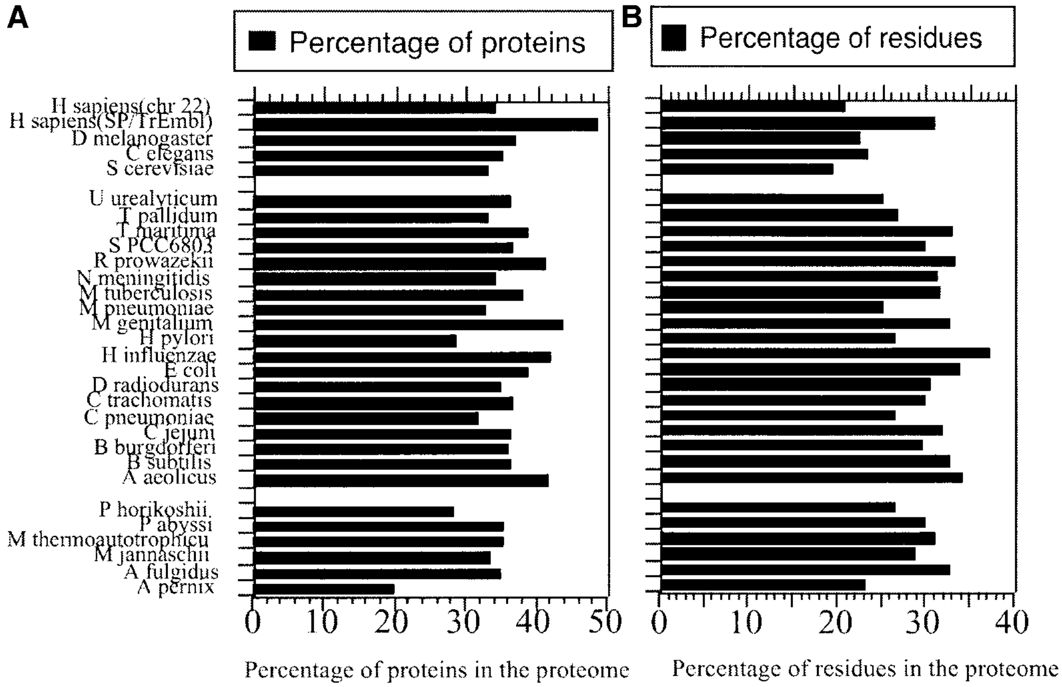

Fig. 5. Structural annotation of genomes. (*A*) 25%–40% of all ORFs were sequence similar to at least one PDB protein. (*B*) The total percentage of residues that could thus be homology modeled amounted to ~20%–30% of all residues.

Materials and methods

Source of sequences

We obtained the sequences for all 28 organisms we analyzed from the public domain (Fleischmann et al. 1995; Fraser et al. 1995; Bult et al. 1996; Himmelreich et al. 1996; Kaneko et al. 1996; Blattner et al. 1997; Fraser et al. 1997; Klenk et al. 1997; Kunst et al. 1997; Smith et al. 1997; Tomb et al. 1997; Andersson et al. 1998; Cole et al. 1998; Deckert et al. 1998; Fraser et al. 1998; Kawarabayasi et al. 1998; Dunham et al. 1999; Kalman et al. 1999; Kawarabayasi et al. 1999; White et al. 1999; Adams et al. 2000; Parkhill et al. 2000; Tettelin et al. 2000). We downloaded most ORFs from ftp://ncbi.nlm.nih.gov/genbank/genomes/. The exceptions were *H. sapiens* (from SWISS-PROT, release 39 and TrEMBL database, release 15) and *D. melanogaster* (from http://www.fruitfly.org/).

Prediction methods

Membrane proteins. We obtained multiple sequence alignments through a search with `MaxHom` (Sander and Schneider 1991) against SWISS-PROT. We filtered the resulting alignments (Rost 1999) and used them as input for `PHDhtm` (Rost et al. 1996), using the default threshold of 0.8. The total number of membrane proteins was adjusted according to the false-positive rate and false-negative rate published in the original paper:

$$n = (1 - FP)/(1 - FN - FP) \times n_{pred} \\ - FP/(1 - FN - FP) \times n_{total} \qquad (1)$$

where n was the final number of membrane proteins we reported, FP and FN were the false positive and false negative rates, respectively, n_{pred} was the number of predicted membrane proteins in the genome, and n_{total} was the total number of proteins in the genome.

Secreted proteins and coiled-coil regions. We predicted signal peptides with the program `SignalP` (Nielsen et al. 1997), considering a protein to contain a signal peptide if the "mean S" value was above the default threshold. We predicted coiled-coil regions with `COILS` (Lupas 1996b), using a window size of 28 and a probability threshold of 0.9.

Structural homologs. We ran three iterations of `PSI-BLAST` (Altschul et al. 1997) searching against our local filtered databases (Przybylski and Rost 2001) to detect homologs of experimentally known structure in PDB (Berman et al. 2000). We included hits with E-values $<10^{-3}$. We obtained more conservative estimates, searching with a pairwise `BLAST` against PDB, reporting hits with E-values $<10^{-3}$.

Functional classification. We classified cellular function using `EUCLID` (Tamames et al. 1998). As input we used the SWISS-PROT homologs identified by `MaxHom`. `EUCLID` assigned the following 13+1 categories of cellular function (Fraser et al. 1995): Amino acid biosynthesis, Biosynthesis of cofactors, prosthetic groups, and carriers, Cell envelope, Cellular processes, Central intermediary metabolism, Energy metabolism, Fatty acid and phospholipid metabolism, Purines, pyrimidines, nucleosides, and nucleotides, Regulatory functions, Replication, Transcription, Translation, Transport and binding proteins, and Other categories. Proteins described as Unclassified either had no SWISS-PROT homolog or could not be classified by `EUCLID`.

Statistical analysis

To examine if the ORF lengths followed an extreme value distribution:

$$f(x) = \frac{1}{\beta} e^{-\frac{x-\mu}{\beta}} e^{-e^{\frac{x-\mu}{\beta}}}$$

we calculated the mean $\bar{x}$ and standard deviation s of the ORF length from archaes, prokaryotes, and eukaryotes, and estimated the parameter β and μ using the method of moments:

$$\beta = \frac{s\sqrt{6}}{\pi}, \mu = \bar{x} - 0.5772\beta$$

We then performed a goodness-of-fit test to determine whether the observed distribution was compatible with the extreme value distribution.

Electronic supplemental material

The supplemental material includes three figures showing the following: (1) cumulative percentage of membrane proteins with different numbers of predicted transmembrane segments relation between the ORF length and the frequency of proteins with given number of predicted transmembrane segments, (2) distribution of length of globular regions in membrane proteins, and (3) distribution of coiled-coil segments in the proteins and amino acid composition of the coiled-coil segments. All of the figures and figure legends are included in the file, supplement.pdf.

Acknowledgments

We thank Henrik Nielsen (CBS, Denmark) for the source code of `SignalP` and for his generous help in using this program, Andrei Lupas (SmithKline Beecham Pharmaceuticals) for helpful suggestions about running the `COILS` program, and Florencio Pazos, Damien Devos, and Alfonso Valencia (all CNB Madrid) for supplying and helping with the `EUCLID` program. The work of J.L. and B.R. was supported by grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institutes of Health. We thank all those who enabled this analysis by depositing experimental information into public databases and to those who maintain these databases.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

1997. The Yeast Genome Directory. *Nature* **387:** 5.

1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. The *C. elegans* Sequencing Consortium (publ. errata appear in *Science* 1999, **283:** 35, **283:** 2103, **285:** 1493). Science **282:** 2012–2018.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Altschul, S., Madden, T., Shaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. 1997. Gapped Blast and PSI-Blast: A new generation of protein database search programs. *Nucl. Acids Res.* **25:** 3389–3402.

Andersson, S.G., Zomorodipour, A., Andersson, J.O., Sicheritz-Ponten, T., Alsmark, U.C., Podowski, R.M., Naslund, A.K., Eriksson, A.S., Winkler, H.H., and Kurland, C.G. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396:** 133–140.

Arkin, I.T., Brunger, A.T., and Engelman, D.M. 1997. Are there dominant membrane protein families with a given number of helices? *Proteins* **28:** 465–466.

Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28:** 45–48.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.L. 2000. GenBank. *Nucl. Acids Res.* **28:** 15–18.

Berman, H.M., Westbrook, J., Feng, Z., Gillliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucl. Acids Res.* **28:** 235–242.

Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1474.

Boyd, D., Schierle, C., and Beckwith, J. 1998. How many membrane proteins are there? *Protein Sci.* **7:** 201–205.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273:** 1058–1073.

Burley, S.K., Almo, S.C., Bonanno, J.B., Capel, M., Chance, M.R., Gaasterland, T., Lin, D., Sali, A., Studier, F.W., and Swaminathan, S. 1999. Structural genomics: Beyond the human genome project. *Nature Gen.* **23:** 151–157.

Chothia, C. 1992. One thousand protein families for the molecular biologist. *Nature* **357:** 543–544.

Cleves, A.E. 1997. Protein transports: The nonclassical ins and outs. *Curr. Biol.* **7:** 318–320.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry 3rd, C.E., et al. 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393:** 537–544.

Crick, F.H.C. 1953. The packing of a-helices: Simple coiled-coils. *Acta Crystallogr. Sect. A* **6:** 689–697.

Das, S., Yu, L., Gaitatzes, C., Rogers, R., Freeman, J., Bienkowska, J., Adams, R.M., Smith, T.F., and Lindelien, J. 1997. Biology's new Rosetta stone. *Nature* **385:** 29–30.

Deckert, G., Warren, P.V., Gaasterland, T., Young, W.G., Lenox, A.L., Graham, D.E., Overbeek, R., Snead, M.A., Keller, M., Aujay, M., et al. 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392:** 353–358.

Devos, D. and Valencia, A. 2000. Practical limits of function prediction. *Proteins* **41:** 98–107.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300:** 1005–1016.

Eyrich, V., Martí-Renom, M.A., Przybylski, D., Fiser, A., Pazos, F., Valencia, A., Sali, A., and Rost, B. 2001. EVA: Continuous automatic evaluation of protein structure prediction servers. *Bioinformatics* (in prep.).

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496–512.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M. et al. 1995. The minimal gene complement of Mycoplasma genitalium. *Science* **270:** 397–403.

Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R., Lathigra, R., White, O., Ketchum, K.A., Dodson, R., Hickey, E.K., et al. 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* 390: 580–586.

Fraser, C.M., Norris, S.J., Weinstock, G.M., White, O., Sutton, G.G., Dodson, R., Gwinn, M., Hickey, E.K., Clayton, R., Ketchum, K.A., et al. 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281:** 375–388.

Frishman, D. and Mewes, H. 1997. PEDANTic genome analysis. *Trends Genetics* **13:** 415–416.

Gerstein, M. 1998. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold. Des.* **3:** 497–512.

Gerstein, M. and Levitt, M. 1997. A structural census of the current population of protein sequences. *Proc. Natl. Acad. Sci.* **94:** 11911–11916.

Gerstein, M. and Hegyi, H. 1998. Comparing genomes in terms of protein structure: Surveys of a finite parts list. *FEMS Microbiol. Rev.* **22:** 277–304.

Goffeau, A., Slonimski, P., Nakai, K., and Risler, J.L. 1993. How many yeast genes code for membrane-spanning proteins? *Yeast* **9:** 691–702.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.C., and Herrmann, R.

1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucl. Acids Res.* **24:** 4420–4449.

Holm, L. and Sander, C. 1999. Protein folds and families: Sequence and structure alignments. *Nucl. Acids Res.* **27:** 244–247.

Jones, D.T. 1998. Do transmembrane protein superfolds exist? *FEBS Lett.* **423:** 281–285.

Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Fan, J., Hyman, R.W., Olinger, L., Grimwood, J., Davis, R.W., and Stephens, R.S. 1999. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nature Gen.* **21:** 385–389.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., et al. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3:** 109–136.

Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., et al. 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res.* **5:** 55–76.

Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., et al. 1999. Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res.* **6:** 83–101, 145–152.

Klenk, H.P., Clayton, R.A., Tomb, J.F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K., Peterson, J.D., et al. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390:** 364–370.

Koonin, E.V., Mushegian, A.R., and Rudd, K.E. 1996. Sequencing and analysis of bacterial genomes. *Curr. Biol.* **6:** 404–416.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S., et al. 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390:** 249–256.

Lindahl, E. and Elofsson, A. 2000. Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* **295:** 613–625.

Liu, J. and Rost, B. 2000. Analyzing all proteins in entire genomes. CUBIC, Columbia University, Dept. of Biochemistry and Molecular Biophysics. http://cubic.bioc.columbia.edu/genomes/

———. 2001. Target space for structural genomics revisited. *Bioinformatics* (in prep.).

Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G., and Chothia, C. 2000. SCOP: A structural classification of proteins database. *Nucl. Acids Res.* **28:** 257–259.

Lupas, A. 1996a. Coiled coils: New structures and new functions. *Nucl. Acids Res.* **21:** 375–382.

———. 1996b. Prediction and analyis of coiled-coil structures. *Methods Enzymol.* **266:** 513–525.

———. 1997. Predicting coiled-coil regions in proteins. *Curr. Opin. Struct. Biol.* **7:** 388–393.

Marti-Renom, M.A., Madhusudhan, M.S., Fiser, A., and Sali, A. 2001. Accuracy of comparative modeling. http://pipe.rockefeller.edu/eva/cm/res/accuracy.html

Nakai, K. 2000. Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem.* **54:** 277–344.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engin.* **10:** 1–6.

Odgren, P.R., Harvie Jr., L.W., and Fey, E.G. 1996. Phylogenetic occurrence of coiled coil proteins: Implications for tissue structure in metazoa via a coiled coil tissue matrix. *Proteins* **24:** 467–484.

Orengo, C.A., Todd, A.E., and Thornton, J.M. 1999. From protein structure to function. *Curr. Opin. Struct. Biol.* **9:** 374–382.

Parkhill, J., Wren, B.W., Mungall, K., Ketley, J.M., Churcher, C., Basham, D., Chillingworth, T., Davies, R.M., Feltwell, T., Holroyd, S., et al. 2000. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403:** 665–668.

Przybylski, D. and Rost, B. 2001. Alignments grow, secondary structure prediction improves. Columbia University.

Rost, B. 1998. Marrying structure and genomics. *Structure* **6:** 259–263.

———. 1999. Twilight zone of protein sequence alignments. *Protein Engin.* **12:** 85–94.

Rost, B., Casadio, R., and Fariselli, P. 1996. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **5:** 1704–1718.

Sali, A. 1998. 100,000 protein structures for the biologist. *Nature Struct. Biol.* **5:** 1029–1032.

Sander, C. and Schneider, R. 1991. Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* **9:** 56–68.

Schneider, G. 1999. How many potentially secreted proteins are contained in a bacterial genome? *Gene* **237:** 113–121.

Shapiro, L. and Harris, T. 2000. Finding function through structural genomics. *Curr. Opin. Biotech.* **11:** 31–35.

Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., et al. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: Functional analysis and comparative genomics. *J. Bacteriol.* **179:** 7135–7155.

Tamames, J., Ouzounis, C., Sander, C., and Valencia, A. 1996. Genomes with distinct function composition. *FEBS Lett.* **389:** 96–101.

Tamames, J., Ouzounis, C., Casari, G., Sander, C., and Valencia, A. 1998. EUCLID: Automatic classification of proteins in functional classes by their database annotations. *Bioinformatics* **14:** 542–543.

Tettelin, H., Saunders, N.J., Heidelberg, J., Jeffries, A.C., Nelson, K.E., Eisen, J.A., Ketchum, K.A., Hood, D.W., Peden, J.F., Dodson, R.J., et al. 2000. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287:** 1809–1815.

Thanassi, D.G. and Hutltgren, S.J. 2000. Multiple pathways allow protein secretion across the bacterial outer membrane. *Curr. Opin. Cell Biol.* **12:** 420–430.

The Genome International Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Tomb, J.F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S., Dougherty, B.A., et al 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388:** 539–547.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The human genome. *Science* **291:** 1304–1351.

Vosshall, L.B., Amrein, H., Morozov, P.S., Rzhetsky, A., and Axel, R. 1999. A spatial map of olfactory receptor expression in the *Drosophila* antenna. *Cell* **96:** 725–736.

Vosshall, L.B., Wong, A.M., and Axel, R. 2000. An olfactory sensory map in the fly brain. *Cell* **102:** 147–159.

Wallin, E. and von Heijne, G. 1998. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **7:** 1029–1038.

White, O., Eisen, J.A., Heidelberg, J.F., Hickey, E.K., Peterson, J.D., Dodson, R.J., Haft, D.H., Gwinn, M.L., Nelson, W.C., Richardson, D.L., et al. 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286:** 1571–1577.

Zhang, L., Godzik, A., Skolnick, J., and Fetrow, J.S. 1998. Functional analysis of the *Escherichia coli* genome for members of the alpha/beta hydrolase family. *Fold. Des.* **3:** 535–548.