PROTEINS: Structure, Function, and Genetics 33:518–534 (1998)

Patterns of Protein-Fold Usage in Eight Microbial Genomes: A Comprehensive Structural Census

Mark Gerstein*
Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut

ABSTRACT **Eight microbial genomes are compared in terms of protein structure. Specifically, yeast, *H. influenzae, M. genitalium, M. jannaschii, Synechocystis, M. pneumoniae, H. pylori,* and *E. coli* are compared in terms of patterns of fold usage—whether a given fold occurs in a particular organism. Of the ~340 soluble protein folds currently in the structure databank (PDB), 240 occur in at least one of the eight genomes, and 30 are shared amongst all eight. The shared folds are depleted in all-helical structure and enriched in mixed helix-sheet structure compared to the folds in the PDB. The top-10 most common of the shared 30 are enriched in superfolds, uniting many non-homologous sequence families, and are especially similar in overall architecture—eight having helices packed onto a central sheet. They are also very different from the common folds in the PBD, highlighting databank biases. Folds can be ranked in terms of expression as well as genome duplication. In yeast the top-10 most highly expressed folds are considerably different from the most highly duplicated folds. A tree can be constructed grouping genomes in terms of their shared folds. This has a remarkably similar topology to more conventional classifications, based on very different measures of relatedness. Finally, folds of membrane proteins can be analyzed through transmembrane-helix (TM) prediction. All the genomes appear to have similar usage patterns for these folds, with the occurrence of a particular fold falling off rapidly with increasing numbers of TM-elements, according to a "Zipf-like" law. This implies there are no marked preferences for proteins with particular numbers of TM-helices (e.g. 7-TM) in microbial genomes. Further information pertinent to this analysis is available at http://bioinfo.mbb.yale.edu/genome. Proteins 33:518–534, 1998.** © 1998 Wiley-Liss, Inc.

Key words: structure databank; superfold; protein structure

INTRODUCTION

In the last 3 years the genomes of a number of free-living organisms have been completely sequenced, generating tremendous interest, popular as well as scientific.[1–3] This event provides a unique opportunity to perform comprehensive comparisons between organisms on a molecular level. One of the most interesting questions that can be addressed through such comparisons is whether different organisms have distinctly different patterns of protein fold usage. That is, to what degree is there a common set of molecular parts (or shapes) that are shared universally among different organisms? Or, conversely, to what degree do certain protein folds occur only in one group of organisms and not in others (e.g., in eukaryotes but not in eubacteria)?

This type of "occurrence" analysis has been performed previously in terms of sequence motifs, families, functions, and biochemical pathways. Starting from the most basic units, genomes have been compared in terms of the relative frequencies of short oligonucleotide and oligopeptide "words."[4–7] The degree of gene duplication in a number of genomes has been ascertained.[8–13] Other analyses have looked at how many highly conserved sequence families in one organism are present in another.[14–19] Finally, if sequences can be related to specific functions and pathways, one can see whether homologous sequences in two organisms truly have the same role (ortholog vs. paralog) and whether particular pathways are present or absent in different organisms.[8,16,20–23] This work has yielded many interesting conclusions in terms of pathways that are modified or absent in certain organisms. For instance, the essential citric acid cycle is found to be highly modified in *Haemophilus influenzae*.[23,24] Furthermore, identifying pathways and proteins unique to certain microbes may prove useful for developing drugs (e.g., antibiotics against bacteria).[16,18,25–29]

The analysis of structure and fold families is expected to be particularly advantageous from the point of view of occurrence analyses for three reasons:

First, structures allow one to define more precisely the "module" or part that is shared. This is particularly true for groups of aligned structures, which

*Correspondence to: Mark Gerstein, Department of Molecular Biophysics and Biochemistry, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520. E-mail: Mark.Gerstein@yale.edu
Received 28 May 1998; Accepted 11 August 1998

TABLE I. Genomes and Abbreviations Used

Abbrev.	Kingdom (subgroup)	Genome	Size (Mb)	No. of ORFs	Reference	Website URL (http://. . .)
EC	Bacteria (gram negative)	*Escherichia coli*	4.60	4290	87	www.genetics.wisc.edu
HI	Bacteria (gram negative)	*Haemophilus influenzae*	1.83	1680	24	www.tigr.org/tdb/mdb/hidb/hidb.html
HP	Bacteria (gram negative)	*Helicobacter pylori*	1.66	1577	53	www.tigr.org/tdb/mdb/hpdb/hpdb.html
MG	Bacteria (gram positive)	*Mycoplasma genitalium*	0.58	468	120	www.tigr.org/tdb/mdb/mgdb/mgdb.html
MJ	Archaea (Euryarchaeota)	*Methanococcus jannaschii*	1.66	1735	121	www.tigr.org/tdb/mdb/mjdb/mjdb.html
MP	Bacteria (gram positive)	*Mycoplasma pneumoniae*	0.81	677	122	www.zmbh.uni-heidelberg.de/M_pneumoniae/MP_Home.html
SC	Eukarya (fungi)	*Saccharomyces cerevisiae*	13	6218	123	genome-www.stanford.edu/Saccharomyces
SS	Bacteria (Cyanobacteria)	*Synechocystis* sp.	3.57	3168	124	www.kazusa.or.jp/cyano/cyano.html

allow the definition of a structural core.[30,31] It is possible (and quite productive) to define modules purely in terms of conserved regions in sequence alignments.[32–38] However, functioning protein modules fundamentally consist of units of 3D structure, usually folding domains, and relating modules defined on the sequence level to structure enables them to be better characterized.

Second and more importantly, one expects analysis of structure to reveal more about distant evolutionary relationships than just sequence comparison, since structure is more conserved than sequence or function.[39,40] In other words, it is at the level of protein structure where one sees the greatest redundancy and reuse in biology. It is believed that the number of structural motifs is very limited, and elucidation of this limited repertoire of molecular parts is seen as one of the principal future challenges for biology.[41,42]

A final reason that structure is advantageous for genome comparisons is that the relationship between sequence similarity and structural similarity is much better defined than the corresponding relationship between sequence and function.

It is generally accepted that proteins with similar sequences usually have similar structures. A decade ago Chothia and Lesk[39,43] systematically investigated this relationship. They found that the extent of the structural changes is directly related to the extent of the sequence changes. The relationship between sequence similarity and functional similarity is much less clear.[44] In part, this is because it is much more difficult to specify precisely a function than a sequence or a structure. Moreover, even when the functional identification is well specified, there are several examples in which highly similar sequences have completely different functions, i.e., same fold but different function.

Here the eight genomes listed in Table I are compared in terms of their usage of protein folds. These eight, which are among the first to be completely sequenced, provide a most diverse comparison. They represent microbes from the three kingdoms of life (Eukarya, Eubacteria, Archaea), from different environments (room temperature and pressure to high temperature and pressure, and neutral pH to highly acidic), with a wide range of genome sizes (0.6–13 Mb), and with a variety of modes of life (from parasite to autotroph).

The comparisons here follow up on recent work comparing fold usage in representative collections of sequences from different species or in complete inventories of predicted structures in a genome.[12,45,46,127,128] There has also been much work focusing specifically on surveying the occurrence of membrane proteins in genomes.[12,47–55] As the work here implicitly involves comparison of protein structures, it also rests on a foundation provided by the emerging protein fold classifications.[56–63]

CATEGORIES OF FOLDS

The protein folds in the genomes can be divided into three categories:

1. *Those corresponding to known structures of soluble proteins.* Based on current technology, folds in this category represent 6–14% of the total residues in the genomes, 9% on average (involving 11–20% of the open reading frames [ORFs]). Similar fractions have been found in many previous analyses.[28,44] These folds are the part of the genome that can be best characterized in terms of protein structure and will be dealt with first, in the next section.

2. *Those made from transmembrane helices.* Folds in this category are analyzed by transmembrane helix prediction.
3. *Other*. Proteins in this category are either soluble proteins with (currently) unknown fold, membrane proteins composed of transmembrane β-strands (such as porins[64]), or proteins that do not assume a fixed conformation (such as repetitive, low-complexity regions[65]). Although some of these proteins can be surveyed structurally to a limited degree by prediction methods, here they are filtered out and excluded from the analysis. (For an example of the identification of β-membrane proteins in genomes, see Champion et al.[66].)

Division of the Protein Databank Into Families, Folds, and Superfolds

An important preliminary step in characterizing the known folds in the genomes is clustering the proteins in the structure databank (the PDB[67]) into sequence families, groups of homologous sequences for which there is no significant similarity between groups. Doing this, via a new clustering approach described below, gives 990 distinct sequence families in the (current) databank. Then, using the structural similarity relationships in the SCOP database, sequence families that share the same fold but that have no detectable homology can be combined into folds. There are currently 338 folds in the PDB, with an average of three sequence families per fold.[68] The fact that the number of folds is considerably less than the number of sequence families suggests that many of the evolutionary similarities between highly diverged organisms may only be apparent in terms of structure, all the sequence similarity having been eroded away.[69]

The known folds can be ranked by how many different families of nonhomologous sequences with which they are associated. Folds uniting many distinct sequence families have been dubbed superfolds.[59] These may represent intrinsically stable and favorable structural arrangements, as suggested by a variety of analyses.[59,70,71] Here the 25 known folds associated with the most sequence families are defined to be superfolds.

Thus the analysis begins by dividing the structure databank at three levels: into 990 sequence families, which are apportioned among 338 folds, which, in turn, contain 25 superfolds.

ANALYSIS OF SOLUBLE-PROTEIN FOLDS IN GENOMES

Fold Tables and Usage Patterns in Terms of Binary Numbers

Having been clustered, the known structures in the PDB were compared against the eight genomes. The raw results take the form of two "fold" tables, listing how many times each of the 990 sequence families and 338 folds in the PDB occur in each of the eight genomes. The complete tables are quite large (8×338 and 8×990), so showing them in full is not possible. Only the top quarter of one is reproduced here, listing the 54 folds that occur in at least seven of the eight genomes (Fig. 1). However, as described below, the complete tables (and other associated information) are available over the web in a variety of convenient formats (see, in particular, http://bioinfo.mbb.yale.edu/genome/browser/fold-report).

The raw fold tables are condensed and cross-tabulated into summaries (Tables II and III), indicating how often particular patterns of fold usage occur (i.e., how often a fold is in yeast and *Escherichia coli* but not in the other six genomes). One way of achieving this condensation is through the use of Venn diagrams.[45,72,73] However, this is awkward for eight genomes. A more convenient representation for these patterns is through an 8-digit binary number, where a digit is "1" if the fold occurs in the corresponding genome and "0" if it does not assuming the genomes are listed in distinct order as shown in Table II). (Also "*" matches both occurs and does not.) There are 255 possible patterns of fold usage (28–1). However, as indicated in Table II, only about a quarter (62, 24%) of these patterns are observed.

The most common single pattern of fold usage is for a fold to occur in all eight genomes, and this occurs 30 times, as shown in Figure 1.

The 30 shared folds presumably represent a most ancient and essential set of molecular parts, as they are present in all three kingdoms of life, in a wide variety of environments, and in genomes of very different size. They include a number of ribosomal protein folds (e.g., L14 and S5, domain 2), folds that

Fig. 1. Usage of each of the known folds in eight different genomes. The entire table is available over the web at http://bioinfo.mbb.yale.edu/genome/browser/fold-report. Here only the top part of the table is shown, corresponding to the folds that occur in at least seven genomes. In all columns inverted (white-on-black) squares are for numbers >10, gray squares are for 2–9, and white squares are for 1. Column 1 ("class") is the structural class that the fold belongs to, as determined by SCOP.[58] Column 2 ("Fold#") is the fold number in SCOP 1.35. Columns 3–10 ("EC" to "MG") give the total number of matches in one genome for a particular fold. This is on a domain level so there can potentially be more than one match per ORF (see text). For instance, the first row shows that there are 19 Rossmann fold domains in the HP genome. The columns are sorted in terms of the total number of matches in the genome, with EC having the most and MG the least. Column 11 ("Total") is the row total of columns 3–10, the total number of times the fold occurs in all eight genomes. Column 12 ("Fam.") gives the number of sequence families with a particular fold in the PDB. This column is used to determine whether or not a fold is a superfold (top 25 in terms of the number of sequence families), and inverted boxes highlight the superfolds. Column 13 ("PDB") gives the number of times a particular fold occurs in the PDB, i.e., how many structures have been solved with this fold. This column should be compared with column 11 ("Total") to highlight the biases in the PDB. Column 14 ("Rep. Struc.") gives a representative structure with this fold, including residue selection. (The residue selection for GroEL is A:2–136, A:410–525.) Abbreviations: dom, domain; Nt-dom, N-terminal domain; Ct-dom, C-terminal domain.

class	Fold#	EC	SC	HI	SS	HP	MJ	MP	MG	total	Fam.	PDB	Rep.	Struc.	Name
α/β	18	60	46	23	40	19	7	4	3	202	16	183	1xel	-	NAD(P)-binding Rossmann Fold
α/β	24	20	69	17	19	17	16	10	11	179	13	132	1gky	-	P-loop Containing NTP Hydrolases
α+β	31	37	28	18	16	12	40	3	3	157	23	160	1fxd	-	like Ferrodoxin
α/β	01	45	36	13	22	11	10	5	4	146	37	399	1byb	-	TIM-barrel
α/β	23	18	17	7	9	4	8	2	2	67	5	36	1pyd	a:2-181	Thiamin-binding
α/β	04	15	11	7	10	1	9	5	5	63	13	132	2tmd	a:490-645	FAD/NAD(P)-binding
α+β	55	8	9	7	8	9	3	6	6	56	4	23	1sry	a:111-421	Class-II-aaRS/Biotin Synthetases
β	27	7	10	8	8	4	4	3	3	47	5	19	1fnb	19-154	Reductase/Elongation Factor Domain
β	24	13	7	4	3	3	3	3	3	39	18	177	1snc	-	OB-fold
α+β	11	10	8	4	8	2	2	2	1	37	11	48	1igd	-	beta-Grasp
β	55	9	10	5	5	2	2	2	2	37	7	19	1bdo	-	Barrel-sandwich hybrid
α/β	15	5	5	4	4	5	6	3	3	35	3	22	2ts1	1-217	ATP pyrophoshatases
α/β	05	10	4	2	4	2	2	2	3	29	4	35	1zym	a:	The "swivelling" beta/beta/alpha domain
α/β	60	5	7	4	6	3	2	1	1	29	3	18	3pmg	a:1-190	Phosphoglucomutase, first 3 domains
α+β	68	4	2	3	6	4	2	4	3	28	2	3	1mat	-	Creatinase/methionine aminopeptidase
α+β	39	6	4	3	4	4	1	1	1	24	3	42	1gad	o:149-312	like G3P dehydrogenase, Ct-dom
α+β	18	5	4	4	1	2	2	1	2	21	3	23	1fkd	-	FKBP-like
α/β	41	3	3	3	3	1	3	1	1	18	3	16	1opr	-	Phosphoribosyltransferases (PRTases)
α	78	1	9	1	2	1	1	1	1	17	1	23	1oel	a:(*)	GroEL, the ATPase domain
α+β	10	2	2	2	4	2	1	2	2	17	2	5	1dar	477-599	Ribosomal protein S5 domain 2-like
α+β	43	4	3	2	2	1	1	2	2	17	4	50	3grs	364-478	FAD/NAD-linked reductases, dimer-dom.
α+β	09	3	4	3	1	2	1	1	1	16	3	12	1kpa	a:	HIT-like
α/β	47	4	2	3	1	2	1	1	1	15	2	10	1ulb	-	Purine and uridine phosphorylases
α+β	33	3	1	3	3	2	1	1	1	15	2	3	1tig	-	IF3-like
α+β	26	2	3	1	2	2	1	1	1	13	3	4	1stu	-	dsRBD & PDA domains
α+β	29	2	5	1	1	1	1	1	1	13	3	26	1one	a:1-141	like Enolase, Nt-dom.
M	11	2	1	2	1	2	2	1	1	12	1	1	1ecl	-	type I DNA topoisomerase
β	23	1	3	1	1	1	1	1	1	10	1	1	1whi	-	Ribosomal protein L14
α/β	31	2	2	1	1	1	1	1	1	10	1	10	1trk	a:535-680	Transketolase, Ct-dom.
α/β	61	1	1	1	1	1	1	1	1	8	1	4	3pgk	-	Phosphoglycerate kinase
α/β	13	49	8	14	57	12	5		1	146	15	100	3chy	-	Flavodoxin-like
α/β	38	24	54	15	11	4		4	5	117	19	112	2rn2	-	Ribonuclease H-like motif
α	02	7	18	6	9	4		5	5	54	4	33	1hdj	-	Long alpha-hairpin
β	21	14	13	3	3	2		2	1	38	2	44	1lep	a:	GroES-like
α/β	30	7	13	4	10	2		1	1	38	7	83	1srx	-	Thioredoxin-like
α/β	56	8	4	2	4	2	4	2		26	3	105	2at2	a:	Asp-carbamoyltransferase, Cat.-chain
α+β	70	3	6	3	3	3		3	3	24	3	24	1mxa	1-101	S-adenosylmethionine synthetase. MAT
α/β	44	2	1	3	5	6	4	2		23	5	16	1vid	-	SAM-dependent methyltransferases
M	12	4	1	4	3	2		4	4	22	1	1	1bgw	-	type II DNA topoisomerase
M	16	3	10	2	3	1		1	1	21	1	4	1dkz	a:	like HSP70, Ct-dom.
β	31	4	2	3	3	3		2	1	18	3	20	1bmf	a:24-94	like F1 ATP synthase, a & b sub., A-dom.
α	21	4	2	4	3	2		1	1	17	5	54	1fha	-	Ferritin-like
α/β	55	3	6	1	2	1	2		1	16	1	29	1xaa	-	Isocitrate/isopropylmalate dehydrogenases
α+β	71	3	2	3	3	2	2	1		16	5	10	2pol	a:1-122	DNA clamp
α	49	2	2	2	2	2		2	2	14	2	18	1bmf	a:380-510	Left-handed superhelix
α/β	50	4	4	1	2	1		1	1	14	3	27	2ctb	-	Zn-dependent exopeptidases
α/β	43	4	1	2	3	1		1	1	13	1	7	1cde	-	Glycinamide ribonucleotide transformylase
β	53	2	1	2	2	2	1		1	11	1	4	1lxa	-	Single-stranded left-handed beta-helix
β	38	2	2	1	2		1	1	1	10	1	7	1pkn	116-217	Pyruvate kinase beta-barrel domain
β	28	2	1	2	1	1		1	1	9	1	6	1efu	a:297-393	EF-Tu, Ct-dom.
α/β	03	2	2	1	1	1		1	1	9	1	1	1rlr	221-748	ribonucleotide reductase, R1 sub., Ct-dom.
α+β	85	1	3	1	1		1	1	1	9	3	43	1mld	a:145-313	like LDH/MDH, Ct-dom.
α	15	1	1	1	1	1		1	1	7	1	3	1bmf	g:	F1-ATPase, gamma subunit
α+β	24	1	1	1	1	1		1	1	7	1	1	1ctf	-	Ribosomal protein L7/12, Ct-dom.

Figure 1.

act as scaffolds for many different functions (TIM-barrel and OB-fold), folds for the binding of cofactors and certain other molecules (NAD-binding Rossmann fold and the thiamin-binding fold), and folds associated with specific metabolic functions (the phosphoglycerate kinase fold and the P-loop containing NTP hydrolase fold). There are fewer sequence families than folds present in all the genomes (26), dramatically illustrating how structure is conserved more than sequence.

TABLE II. The Observed Patterns of Fold Usage†

ESHSHMMM CCISPJPG	(##)	ESHSHMMM CCISPJPG	(##)	ESHSHMMM CCISPJPG	(##)	ESHSHMMM CCISPJPG	(##)	ESHSHMMM CCISPJPG	(##)
11111111	(30)	.1.....	(23)	1.....	(19)	11111.11	(16)	111111..	(16)
1111....	(09)	11111...	(08)	1.1.....	(08)	1.111.11	(06)	11.....	(06)
...1....	(06)	1.11....	(05)	.1.1....	(05)	1.111...	(04)	11.1....	(04)
.1...1..	(04)	..1.....	(04)	111111.1	(03)	1111111.	(03)	1111..11	(03)
1111.1..	(03)	.....1..	(03)	1111.111	(02)	111...11	(02)	111.11..	(02)
1.11.1..	(02)	..111...	(02)	.1.11...	(02)	1..1.1..	(02)	1.1..1..	(02)
111.....	(02)	.11.....	(02)	......1.	(02)	....1...	(02)	111..111	(01)
111.1.11	(01)	1.111..1	(01)	1.1111..	(01)	.1.1..11	(01)	.1.11.1.	(01)
.11.1..1	(01)	1....111	(01)	1..111..	(01)	1.1...11	(01)	1.1..11.	(01)
11....11	(01)	11.1.1..	(01)	11.11...	(01)	111..1..	(01)	111.1...	(01)
.11...1.	(01)	1.....11	(01)	1...11..	(01)	1.1.1...	(01)	11	(01)
....1..1	(01)	...1.1..	(01)	...11...	(01)	..1.1...	(01)	.1....1.	(01)
1....1..	(01)	1	(01)						

†Each of the 8-bit binary numbers represents a particular pattern of fold usage: "1" if a fold is present and "." if it is absent. From left to right, the bits correspond to EC, SC, HI, SS, HP, MJ, MP, and MG. The number in parentheses after the binary number is the number of folds that have this pattern of usage. For instance, "..1..... (04)" means that there are four folds present in HI and in no other of the genomes. There are 255 possible patterns of fold usage, but only the 62 that are observed are shown here.

Another common and simple to understand fold-usage pattern is when a fold is present in only a single genome, i.e., for folds unique to specific organisms. As shown in Table III, each of the eight genomes has at least one unique fold. Yeast has the most unique folds, followed by *E. coli* (EC; 23, then 19). At the other extreme, *Methanococcus jannaschii* (MJ) has three unique folds and *Mycoplasma genitalium* (MG) only one.

The converse of a fold being present in a single genome is for it to be absent from only one of the eight genomes. Twenty-four folds are present in exactly seven of the eight of the genomes. They are shown in Figure 1.

It never occurs that a fold is missing from EC, *Saccharomyces cerevisiae* (SC), *Haemophilus influenzae* (HI), or *Synechocystis* sp. (SS) and is present in all the other genomes. When a fold is missing from one genome, it is usually missing from MJ (16 times out of 24). In a similar fashion, most of the time when a fold is present in six of the eight genomes it is absent from *Mycoplasma pneumoniae* (MP) and MG.

Disregarding whether a fold occurs in another genome, one finds that 240 of the 338 known folds in the PDB (and 547 of the 990 sequence families) occur in at least one of the eight genomes. Thus, ~10% of the residues in these primitive organisms match ~70% of the known folds.

Overall, EC has the most distinct (not unique) folds, followed by SC, and predictably MG, MP, and MJ have the fewest. As a fraction of the number of its ORFs, MJ has the fewest known folds.

Fold-Usage Tree

One can also take the observed patterns of fold usage and use this to cluster the genomes. A "distance" between two genomes can be reasonably defined as the number of common folds shared between two genomes as a fraction of the total folds in the genomes. This is similar to the definitions of distance used in traditional phylogenetic analysis, in which the number of shared taxonometric characters or features is used as the basis for classification.[74] Other definitions are, of course, possible. A tree built with this distance metric is shown in Figure 3.

For comparison the fold-usage tree is shown next to a number of other trees constructed by different distance measures:

1. The overall difference in amino acid over the whole genome
2. The number of shared sequence families (very similar to the number of shared folds)
3. The sequence divergence of related proteins that share the same fold and are present in all eight genomes, a measure that is most similar to the customary measure based on individual proteins

Remarkably, even though they are derived from such different properties of the organisms, these trees are all very similar to each other in topology and also similar to the conventional classification of microbes, based on 16S ribosomal RNA sequences (summarized in Table 1).[8,75] That is, they group together the gram-positive bacteria (MP and MG)

TABLE III. Folds Present in or Absent From Only a Single Genome†

	Overall totals		No. of times a "fold" is absent only from this genome				No. of times a "fold" is present only in this genome			
	Matches to the PDB	Distinct folds	Fold	Fam.	All-α	All-β	Fold	Fam.	All-α	All-β
EC	848	174	.	.	.	.	19	51	1	6
SC	1,073	157	.	3	.	.	23	84	5	3
HI	394	146	.	.	.	.	4	11	.	1
SS	534	140	.	1	.	.	6	24	.	3
HP	254	107	2	4	.	1	2	11	.	1
MJ	233	82	16	29	4	3	3	6	1	1
MP	140	76	3	3	.	1	2	3	1	.
MG	134	74	3	1	.	.	1	2	1	.
Total	3,610	338	24	41	4	5	60	192	9	15

†Column 2 ("Matches") shows how many total homologues there are for a particular genome to structures in the PDB (i.e., total PDB hits in the genome). Column 3 ("Distinct folds") shows how many different folds are contained in this genome, regardless of whether these folds occur in other genomes. Columns 4–7 (under "absent") show how many times one of the 338 folds, 990 sequence families, 48 all-α folds, or 39 all-β folds are absent only from this genome (and not from the other seven). Columns 8–11 ("present"), conversely, show how many times a fold or family is *only* present in this genome, i.e., how many unique folds the genome has.

and the gram-negative bacteria (HI and EC) and position these two bacterial lineages with the cyanobacteria SS a distance from the eukaryote SC and the archeon MJ. The major difference between the trees is the treatment of HP, which is closer to the mycoplasms in the composition tree and closer to EC and HI in the other trees. HP is a gram-negative proteobacterium and should be grouped with EC and HI. However, it has been found to be rather problematical in terms of evolutionary classification.[53,76]

Distribution of Fold Classes and Superfolds

To gain more structural insight into what types of folds are shared between genomes, it is possible to classify each fold as all-α, all-β, α/β, α + β, or other (using the definitions of Levitt and Chothia[77]) and then to see how the folds corresponding to each of the structural classes are distributed among the genomes (Fig. 1 and Tables 3 and 4). Compared with the PDB, the most common folds in the genomes (i.e., those that occur in many genomes) are enriched in mixed domains containing both helices and strands and depleted in all-α ones. Specifically, of the 30 folds that occur in all eight genomes, one is all-α, four are all-β, and the rest are mixed (3% all-α and 13% all-β). In contrast, of the 338 folds in the PDB, 75 are all-α and 55 are all-β (22% and 16%). EC has most of the unique all-β folds and yeast the most unique all-α ones.

The superfolds are much more highly represented in the folds present in all (or many) genomes than they are in the PDB. In particular, 7 of the 25 superfolds are present in all eight genomes, and only 2 are not present in at least one genome. Figure 2 shows how superfolds are shared to a greater degree between genomes than are folds and likewise, how folds are shared more than sequence families.

Number of Structure Matches and the Top 10 Folds

In addition to being ordered by how many genomes they occur in, folds can also be arranged by how often they occur in total in all eight genomes. That is, the number of "matches" that PDB structures with a given fold have in all eight genomes can be used to rank the folds. The 240 folds present in at least one genome have 3,610 total matches in all the genomes, about 15 per fold and 2 per fold per genome. The yeast genome contributes the most structure matches of the eight genomes (1,073), reflecting its large size and highly duplicated character.[11]

The 10 folds of the 30 that occur most often in all eight genomes, i.e., those with the most matches, are drawn in Figure 4.

These are the top 10 folds. They include the seven superfolds that are present in all eight genomes. Furthermore, they have a remarkably similar architecture, containing interleaved helices and sheets. They can be divided into barrel folds (reductase/elongation-factor, OB-fold, and TIM-barrel), classic α/β folds with helices packed on either side of a central sheet (P-loop hydrolase, Rossmann fold, and thiamin-binding), folds with helices packed onto a single face of a sheet (ferrodoxin, FAD-binding, and β-grasp), and a fold with a more complex structure (class II synthetase). Overall, 8 of the top 10 contain a clear central sheet with helices packed onto at least one face. The two exceptions are the OB-fold and the reductase/elongation-factor fold, which are mostly structured by strands.

Ranking Folds by Expression Level

The top 10 list in Figure 4 ranks folds by how often they occur in the genome, tending to emphasize highly duplicated genes. Folds can also be ordered by a number of other criteria. In particular, they can be

TABLE IV. Yeast Folds Ranked by Duplication and Expression†

Common yeast folds (SCOP)	Rep. structure	Genome duplication	Expression (aerobic)	Expression (anaerobic)
Protein kinases (cat. core)	1hcl	1	3	4
NTP hydrolases with P-loop	1gky	2	1	2
Classic Zn finger	1ard	3	9	5
Ribonuclease H-like motif	2rn2	4	2	1
Rossmann fold	1xe1	5	4	3
Zn2/Cys6 DNA-binding domain	125d	6	6	7
7-Bladed β-propeller	2bbk-H	7	8	16
TIM-barrel	1byb	8	5	6
Like ferrodoxin	1fxd	9	7	10
DNA-binding 3-helix bundle	1enh	10	30	36
—	—			
GroES-like	11ep-A	17	10	9
—	—			
Like HSP70, Ct domain	1dkz-A	22	11	8

†This table shows the most common folds in the yeast genome ranked according to duplication and expression. Columns 1 and 2 give the name for the fold, as determined by SCOP[58] and a representative structure with this fold. Ct domain stands for C-terminal domain. Column 3 gives an ordering of folds in terms of the number of times they are found in the yeast genome. For instance, the top fold (kinase) is found 110 times, and the second fold (NTP hydrolase) is found 69 times (from data in the Fig. 1 fold table). Columns 4 and 5 give an ordering of folds in terms of their degree of expression. Using the data from DeRisi et al.,[80] the total expression E of a fold F is calculated as a sum of the expression levels of all the ORFs that contain this fold. The expression level of a given ORF (i.e., ORF i) is the degree of its "Red" color on a cDNA microarray $R(i)$, less the background $R_{back}(i)$, viz.: $E(F) = \Sigma_{\forall i \text{ containing } F} R(i) - R_{back}(i)$. Column 4 gives the expression in aerobic conditions (high sugar, second time-series data point in DeRisi et al.), and Column 5, in anaerobic conditions (low sugar, high ethanol, last time-series data point in DeRisi et al.). Note how some folds that are in the top 10 in terms of duplication are not in this select list in terms of expression (e.g., "DNA-binding 3-helical bundle").

ranked in terms of expression level, essentially by mRNA occurrence in the cell. This has already been done in nonstructural terms for all the genes in yeast.[78–80] Table IV shows how this expression level ranking maps onto folds. Using data from DeRisi et al.,[80] Table IV shows the most highly expressed folds in yeast grown in two different conditions (high sugar and low sugar, aerobic vs. anaerobic conditions). This ranking of folds is clearly different from that purely based on duplication. In particular, note how two DNA-binding folds (the Zn finger and the DNA-binding 3-helix bundle) are ranked high in terms of duplication but low in terms of expression. This is quite reasonable given that one usually expects DNA-binding proteins to be common in the genome but expressed at very low levels. Note also how the most common fold in terms of expression changes in the different conditions.

ANALYSIS OF MEMBRANE-PROTEIN FOLDS IN GENOMES

Overall Numbers

The usage of membrane-protein folds was surveyed by first performing simple transmembrane-helix (TM) predictions and then seeing what structural class the prediction was in, i.e., 3-TM, 5-TM, and so on.

Overall, about 5% of the residues in the genomes are in transmembrane helices, ranging from a high of 7% in EC to a low of 3% in MJ. The number of ORFs with at least one transmembrane element ranges from ~35% in EC, SC, and SS to ~20% in MJ, for an average of 28%, indicating that EC, SC, and SS have more membrane proteins (as a fraction of the total) than the other genomes. This agrees with previous work by others, who found that ~20–30% of the proteins in microbial genomes are membrane proteins, the specific value depending somewhat on prediction method and threshold used.[12,47–55]

Zipf's Law Fit and 7-TM proteins

The number of TM-helices per protein follows a similar decreasing pattern in each genome, with fewer proteins having large numbers of TM-helices. As shown in Figure 5, the fraction F of proteins in the genome with a given number n of TM-helices can be fit with the expression $F(n) = 0.18\ n^{-1.8}$, where n ranges from 0 to 15. (Without great degradation of the fit, the even simpler expression $1/[5n^2]$ can be used as well.) This expression has a form like that of the Zipf's law that often occurs in the analysis of word frequency in documents.[81] Similar Zipf law-like expressions have been found to apply in a variety of other situations relating to the occurrence of proteins (e.g., in relation to the occurrence of oligopeptide words[82–84]). Moreover, this particular functional form for the occurrence of proteins with a given number of TM-helices falls off smoothly with

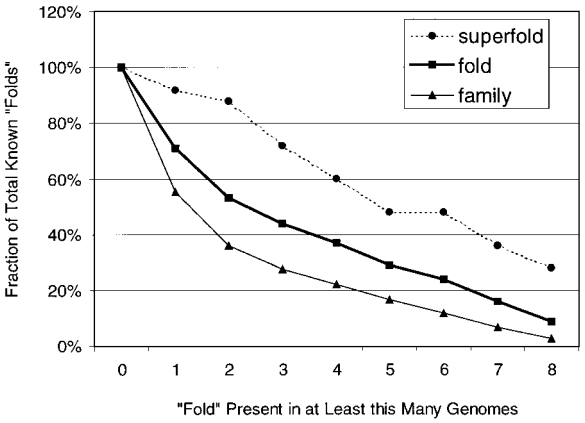

Fig. 2. Fraction of known families, folds, and superfolds present in a given number of genomes. Note how superfolds are shared to a greater degree between genomes than are folds, and likewise, how folds are shared more than sequence families. The data in this figure are derived from the table below, which gives the absolute number of folds present in a given number of genomes. (The graph shows the absolute number in the table divided by the total.) For instance, the third line indicates that 23 of 338 known folds (7%) and 54 of the 990 sequence families (with known fold) (5%) are present in exactly three genomes.

Present in this many genomes	fold	fam.	fold-α	fold-β	SF
0	98	443	27	16	2
1	60	192	9	15	1
2	32	82	15	4	4
3	23	54	6	3	3
4	27	53	4	6	3
5	17	50	3	2	0
6	27	49	6	0	3
7	24	41	4	5	2
8	30	26	1	4	7
Total	338	990	75	55	25
Total (1-8)	240	547	48	39	23

an increasing number of helices (*n*), implying that there is no particular preference (i.e., local maximum) for proteins with seven TM-helices. This suggests that this heavily studied group of proteins is not exceptionally important in the context of microbial genomes.

There are two additional points to note about Figure 5.

First, the logarithmic scale of the figure tends to emphasize proteins with many TM-helices. However, this can be a bit deceptive as the bulk of the membrane proteins in each genome has only a few TM-spans (e.g., 2-TM). Second, although the frequency of TM-helices fits the Zipf's law fairly well in an overall sense, on closer examination there are some notable differences between the genomes. In particular, MP appears to have more 7-TM folds than average (3% vs. 1%) and EC more 12-TM folds. Yeast has fewer 9-TM helix folds and MG fewer 10- and 11-TM folds.

Most of the membrane-protein surveys agree on this absence of 7-TM proteins in microbial genomes; some also claim to find more 6- and 12-TM proteins in bacterial genomes corresponding to well-known families of transporter proteins.[12,50,52,55] In contrast, surveys of the incomplete (and highly biased) set of human sequences and the unfinished worm genome find a relative abundance of 7-TM proteins in these multicellular organisms.[50,55]

Conclusions

Eight microbial genomes have been compared in terms of their usage of protein folds. To this end, a "binary-number" representation was developed for counting and comparing patterns of fold usage. It was found that the eight genomes contain 240 of the 338 known soluble protein folds, and 30 of these are shared among all eight. Compared with the PDB, the shared folds are enriched in mixed structure (both helices and sheets) and depleted in all-α domains. The 10 most common of the 30 shared folds (the top 10) are especially similar in structure, with 8 of the 10 having a classic architecture of helices packed against a central sheet. They are also particularly enriched in superfolds (7 of 10).

Each of the eight genomes, including the very minimal MG, has at least one unique protein fold not present in any of the others. Conversely, when a fold is absent from only one genome, it is usually absent from the archeon MJ. Overall, a tree clustering the genomes in terms of their number of shared folds has a remarkably similar structure to more conventional classifications that are based on amino acid composition or ribosomal sequences.

Finally, the folds of membrane proteins were analyzed through TM-helix prediction. All the genomes appear to have similar patterns of usage for these folds. The occurrence of a particular membrane-protein fold falls off rapidly with more TM-helices, according to a Zipf's law, and there are no marked preferences for folds with particular numbers of TM-helices (such as 7-TM proteins).

Limitations of the approach

The small, incomplete number of known folds. There are a number of limitations to the fold-usage analysis presented here. First, only a relatively small number of folds can be surveyed, involving no more than a fifth of the ORFs in a genome. (This number would be even smaller if one were to restrict attention to just the ORFs in a genome that have been studied directly by crystallography or nuclear magnetic resonance. For example, it is currently 52 out of 6,218 for yeast, as reported by Sacch3D.[85])

The situation is expected to improve in the future as new structures are determined, but it will be a while before all the folds in a genome are known—especially considering that the increase in new folds is much slower than the increase in new structures.[68] An important corollary of this is that the absolute counts found in a given genome survey are

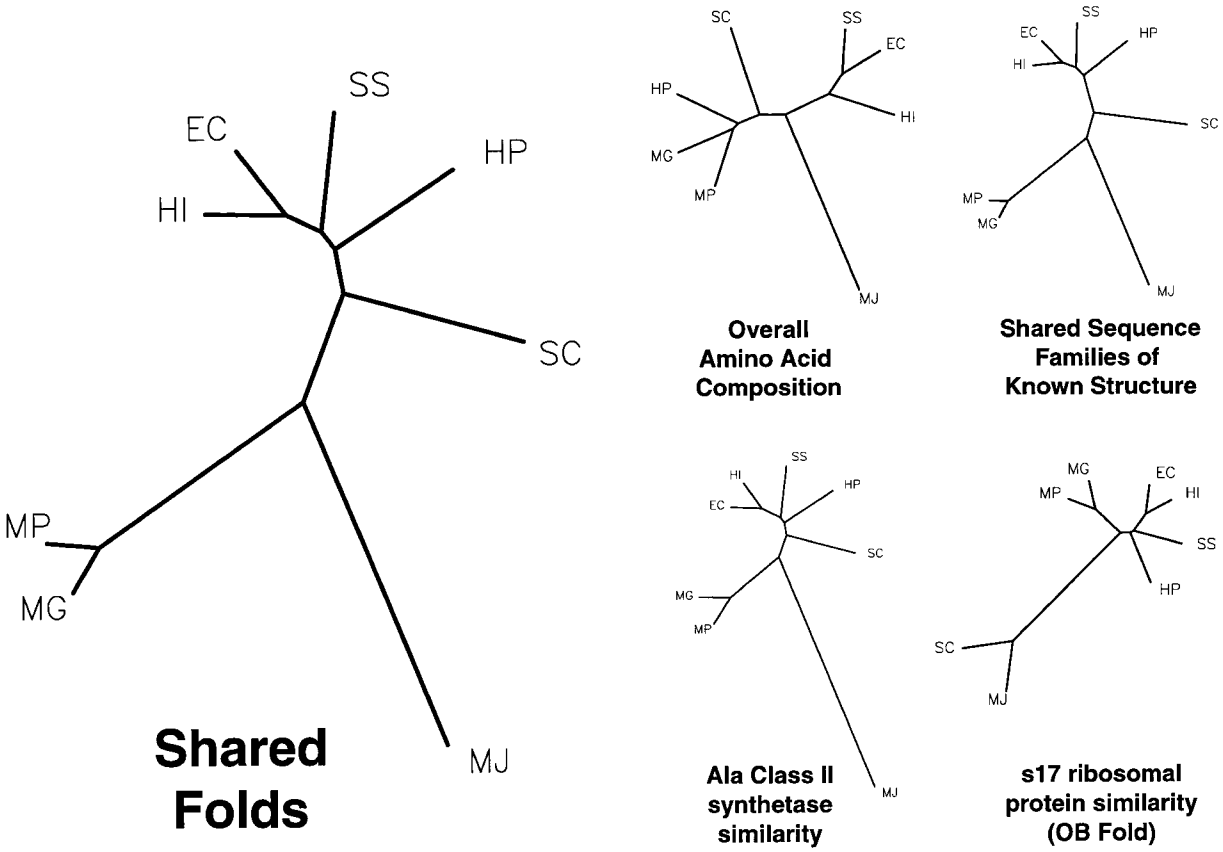

Fig. 3. Cluster trees based on fold usage and other criteria. The unrooted trees in this figure show the result of clustering the genome based on a variety of measures for distance between genomes. (The two-letter abbreviations for genomes are defined in Table I.) Far left, genomes arranged according to patterns of shared folds. Here the definition of distance between two genomes is in terms of fold usage:

$$D = N(11)/(N(10) + N(11) + N(01)),$$

where N(11) is the number of folds in both genomes A and B, N(10) is the number just in the first genome, and N(01) is the number just in the second. Top middle, a tree based on global differences in amino acid composition. The distance between two genomes A and B is defined through the following formula for euclidean distance:

$$D(AB) = \sqrt{\Sigma_{i=1}^{20}(C(i, A) - C(i, B))^2}$$

where C(g,i) is the composition of the ith amino acid in genome g. Other measures of distance were also tried, in particular, the Hellinger distance,[125] which is the same as D (AB) except for the replacement $C(i, \cdot) \rightarrow \sqrt{C(i, \cdot)}$. This treats small differences differently. However, it is found that the resulting tree topology is insensitive to the choice of distance metric—providing a test of the robustness of the results. Top right, just like the fold-usage tree but now based on the number of the 990 sequence families that are shared between genomes. Bottom middle and right, trees based on sequence similarities from pairwise comparison of selected families of orthologous sequences in the genomes, for which a fold is known. Consider the bottom middle tree first. This is for alanyl-tRNA synthetase, which has the class II synthetase fold. Its sequences were selected from the COGS database (specifically all sequences from COG0013 except YNL040w).[16] The distance between a pair of sequences was defined as 1/(S + C), where S is the Smith-Waterman score after a global alignment and C is the mean Smith-Waterman in doing global alignments of all proteins of this length (from the all-vs.-all[102]). The bottom right tree is similarly constructed. It corresponds to ribosomal protein S17, which has an OB-fold. Its sequences were selected from COG0186 (all sequences except YDR025w and YMR188c).

(usually) an underrepresentation of the true numbers. Furthermore, they are contingent on the evolving contents of the databank. Thus, over time as more structures are added to the databank, one should expect such statistics as the most common folds and number of shared folds to change somewhat.

Comprehensive application of ab initio structure prediction and advanced sequence-comparison and fold recognition methods to complete genomes can overcome somewhat the limitations of only knowing a small number of folds,[12,46] allowing one to survey the complete inventory of proteins in an organism. However, in its present form, structure prediction is not a substitute for structure determination, especially in situations when the fold is completely new. Moreover (as discussed below), using state-of-the-art sequence comparison methods introduces a measure of variability and uncertainty into the results, as different methods will give different results at the margin.

The uncertainties in the analysis resulting from the small number of known folds are aptly illus-

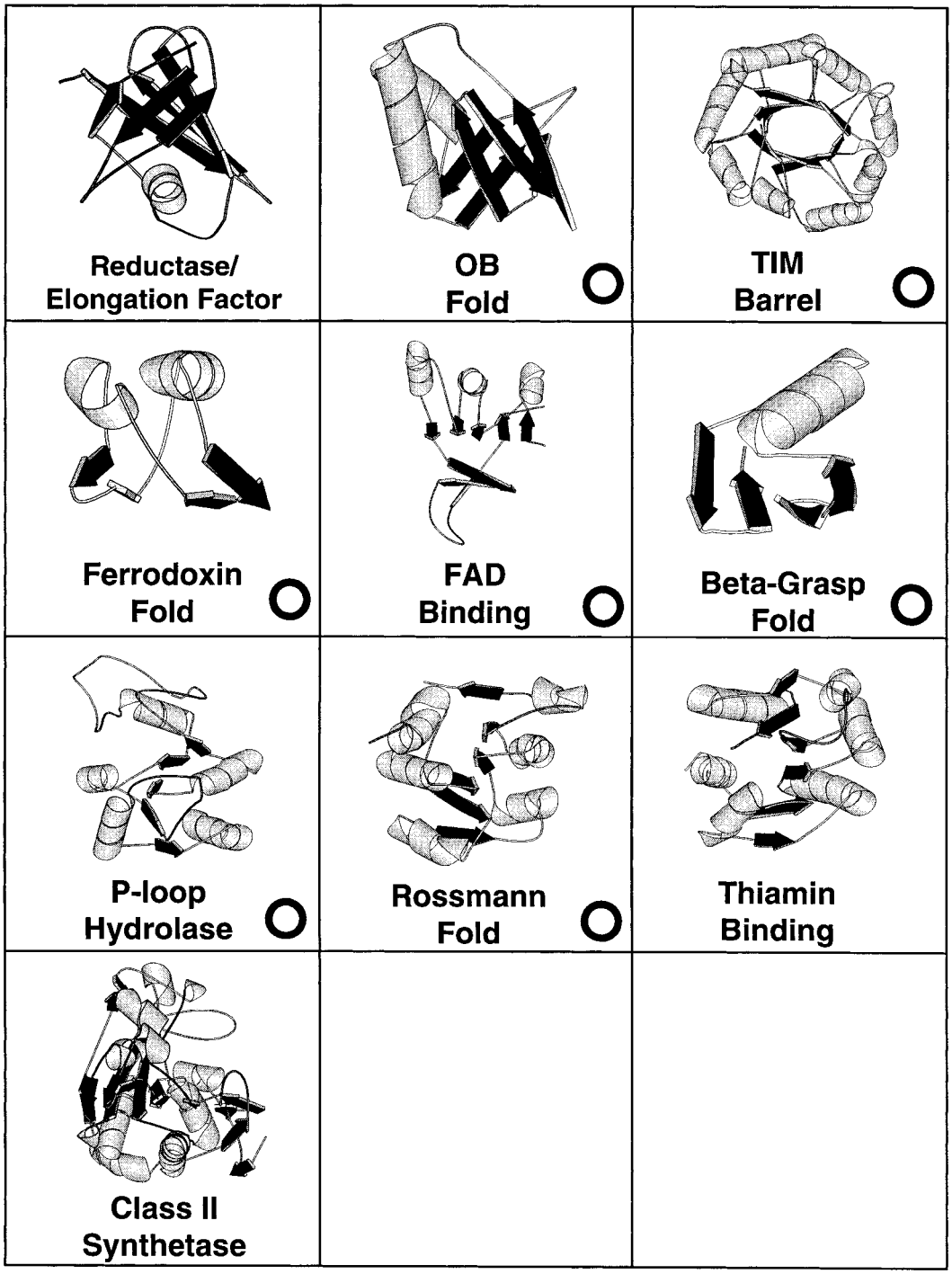

Fig. 4. Pictures of the 10 most common folds that are shared among all eight genomes. The figure is arranged from top row to bottom row: three barrel folds, three classic α/β folds with helices packed on either side of a central sheet, three folds with helices packed onto a single face of a sheet, and one fold with a more complex structure (class II synthetase). All folds are drawn with MOLSCRIPT[126] using residue selections from Figure 1. They are somewhat simplified so that coil geometry is smoothed out and insertions not packing against the central sheet are deemphasized. Folds that are superfolds are indicated by a black circle ("○") in the lower right-hand corner.

trated in Table V. This shows the top 10 folds in yeast calculated two ways, as done here versus as done for the Sacch3D database.[85] The differences between the two lists, which are described in detail in the table footnote, are easily understood and illustrate well how the exact ordering of the top 10 list depends on two factors: the sequence-structure comparison methods used and the contents of the current database of

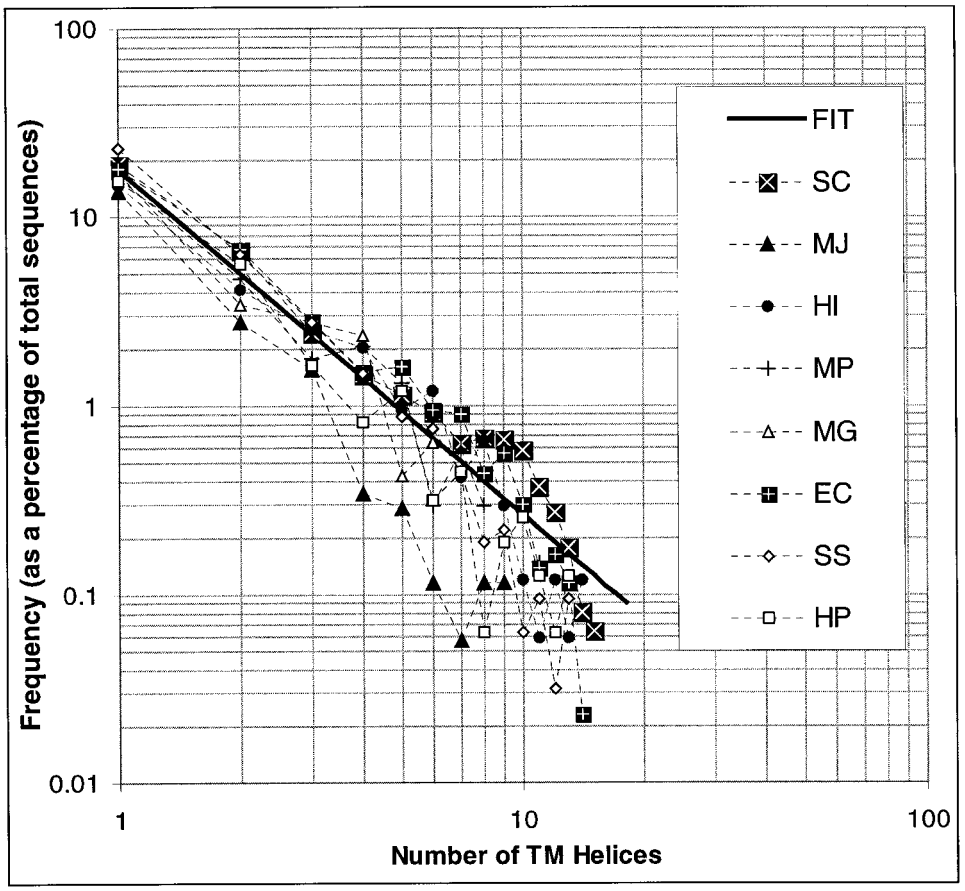

Fig. 5. Log-log graph showing the occurrence of membrane proteins with a given number of transmembrane (TM) helices in each of the eight genomes. The occurrence drops off sharply in a similar fashion in all eight genomes, according to a Zipf-like law. A fit to all eight is shown in the graph. The exact numbers that this chart is based on are listed in the table below, in which the number of proteins with a given number of TM-helices is expressed as a percentage of the total number of sequences in the genomes. For instance, the table shows that 6.6% of the 6,218 yeast ORFs contain two TM-helices. The derived fit values ("FIT" column) are determined by minimizing the chi-squared statistic between a linear model and the observed number of TM-helices in all the genomes:

$$\chi^2 = \sum_{n,g} \frac{(O(n, g) - E(n))^2}{E(n)}$$

where O(n,g) is the observed fraction of n-TM proteins in genome g and E(n) is the expected fraction. Obviously, some genomes fit the model better than other ones. This can be quantified by calculating a chi-squared statistic for the fit to each individual genome (i.e., the same sum as above, but now not summing over g, just n). This value is shown in the last row of the table ("chi-sq"). It shows that MJ followed by MG are the two genomes that fit the model worst—as might be expected given that these two organisms also differ most from the others in terms of the usage of soluble folds.

Num. TM-helices	FIT	SC	MJ	HI	MP	MG	EC	SS	HP
1	17.3	19.1	13.7	14.8	16.4	17.3	18.1	23.2	15.6
2	4.9	6.6	2.8	4.1	4.7	3.4	6.7	6.3	5.6
3	2.4	2.4	1.6	2.5	1.8	2.8	2.8	2.8	1.6
4	1.4	1.5	0.3	2.0	2.1	2.4	1.5	1.5	0.8
5	0.9	1.1	0.3	1.0	1.3	0.4	1.6	0.9	1.2
6	0.7	0.9	0.1	1.2	0.3	0.6	1.0	0.8	0.3
7	0.5	0.6	0.1	0.4	0.6		0.9	0.4	0.4
8	0.4	0.7	0.1	0.7	0.3		0.4	0.2	0.1
9	0.3	0.7	0.1	0.3			0.6	0.2	0.1
10	0.3	0.6		0.1			0.3	0.1	0.3
11	0.2	0.4		0.1	0.1		0.1	0.1	0.1
12	0.2	0.3		0.1			0.2	0.0	0.1
13	0.2	0.2		0.1			0.1	0.1	0.1
14	0.1	0.1		0.1			0.0		
15	0.1	0.1							
num ORFs		6218	1735	1680	677	468	4290	3168	1577
chi-sq		1.8	5.7	1.8	2.1	3.7	1.9	2.7.	1.7

folds. Different comparison programs and fold databases will give different numbers.

Biases in PDB and in the genomes. In addition to rendering the results here, in a sense, incomplete, the small number of known folds also means that the results may be influenced to some degree by the biases in the PDB. These biases are manifest in a number of ways.

First and most simply, there is a considerable disparity between how often a fold occurs in the genomes (i.e., how many total matches it has) and how often it occurs in the PDB (i.e., how many known

TABLE V. The Yeast Top 10, Determined Two Ways†

		GeneCensus		Sacch3D, SGD		
Common yeast folds	Rep. struct.	Count	Rank	Count	Rank	Diff.
Protein kinases (cat. core)	1hcl	110	1	109	1	
NTP hydrolases with P-loop	1gky	69	2	52	2	
Classic Zn finger	1ard	55	3	34	7	●
Ribonuclease H-like motif	2rn2	54	4	30	8	●
Rossmann fold	1xel	46	5	41	5	
Zn2/Cys6 DNA-binding domain	125d	46	6	30	9	●
7-Bladed β-propeller	2bbk-H	46	7	0	—	<
TIM-barrel	1byb	36	8	39	6	●
Ferrodoxin-like	1fxd	28	9	43	4	●
DNA-binding 3-helix bundle	1enh	22	10	22	10	
Long helix oligomers (coils)	1zta	1	—	47	3	<

†This table shows the top 10 folds in yeast calculated two ways, as done here (GeneCensus) versus as done for the Sacch3D database, which is part of SGD.[85] Columns 1 and 2 give the name for the fold, as determined by SCOP,[58] and a representative structure with this fold. The columns labeled "GeneCensus" show the top 10 folds determined with the methods used here: FASTA with e-value cutoff of 0.01 and strict overlap criteria, using a clustered version of the SCOP 1.35 database. The columns labeled "Sacch3D" show the top 10 folds calculated by different methods: WU-BLAST with a *P* value cutoff of 10e-4, using a differently clustered version of an earlier SCOP database, 1.32. For both GeneCensus and Sacch3D both the number of folds found ("count") and the rank in the top 10 list are shown. There is broad agreement between the two top 10 lists. However, there are some differences. These are flagged in the final column. The minor differences (indicated by "●") are the five folds that have slightly changed rank within the top 10. In comparison with BLAST, FASTA appears to be a bit better in finding homologues for the classic Zn finger, ribonuclease H-like motif, and the Rossmann fold, and a bit worse for the TIM-barrel and ferrodoxin-like folds. The two major differences (indicated by "<") are more substantial and warrant explanation: 1 *propeller fold*. The Sacch3D list does not have any "7-bladed β-propeller" folds, because of new structures that were added to SCOP between release 1.32 and 1.35, in particular the structure of transducin; this difference thus illustrates how the top 10 lists evolve with the growth of the PDB; and 2 *helix oligomers*. The Sacch3D list has many more long helix oligomer folds. These are small "folds" such as coiled-coils and leucine zippers. These undoubtedly occur in yeast in great frequency. However, they are very small in size—e.g. the representative coiled-coil in 1zta is only 30 residues. This makes them particularly problematical to define as a fold and to find with sequence searching programs. The difference between the BLAST and FASTA results illustrates dramatically how certain programs may differ in finding these marginal folds.

structures have this fold). This is indicated in Figure 1 (and in the web presentation). One can immediately see how different the common folds are in the PDB versus in the genomes. This illustrates in a direct sense the biases in the PDB—although these sorts of biases are not expected to affect the results (which are principally concerned with "membership" rather than absolute counts).

Second and more subtly, the composition of the PDB is biased toward folds that occur in more heavily studied organisms such as EC and SC. These biases are probably reflected in some of the results, specifically, in the finding that there are many more known folds and unique folds in the bacterium HI than in the archeon MJ, even though both of these organisms have genomes of approximately the same size.

Another subtle bias in the results here is in the selection of genomes. The eight organisms picked were the first with complete genomes to be sequenced, as has by necessity been done in all the other multigenome comparisons to date (e.g., Tatusov et al.[16]). A more balanced comparison would perhaps have a more comparable amount of eukaryotes and archaea to bacteria.

Prospects

Improvements in the results presented here will have to wait for more data, more genome sequences, and more determinations of structures. Despite these limitations, comparisons of genomes in terms of protein structure are certain to yield results about the fundamental differences between organisms on a molecular level. Currently more than 10 microbial genomes have been completed, and at least 35 more are being worked on,[86] so there will be many possibilities for comparison soon.

SEQUENCE AND STRUCTURE ANALYSIS TECHNIQUES

A Relational Database of Genome Sequences and Structure Assignments

Translated genome sequences were taken from the web sites (Table I). The genome data are constantly changing and are contingent on the current state of the art in gene finding. The data used in this paper reflect a particular snapshot of this ongoing process. For instance, the *E. coli* data file used was version M52, containing 4,290 ORFs. This is a more recent version and contains a different number of ORFs than one referred to in the official publication (M49, containing 4,288 ORFs[87]). For yeast there is some uncertainty regarding whether all of the ORFs in the web site file are really genes. In particular, 5,888 of the 6,218 ORFs are definitely believed to be genes, but there is some question about the remaining 330.[88] Furthermore, quite a number of yeast sequences (initially) annotated as ORFs are, in fact,

transposons, which should properly be segregated from the rest of the proteome.[89]

Structures were taken from the PDB via the PDB browser.[67, 90] Domain fold and class definitions were taken from SCOP (version 1.35, May 1996).[9,58,91] Specific values quoted about the composition of the PDB, e.g., that it has 5,493 total structures and 222 T4 lysozyme structures, refer to the state of the databank when SCOP 1.35 was built. Core structures for each domain were based on refinement of structural alignments.[12,30,92,93]

Analysis and processing of the data were greatly expedited by the use of a simple relational database, implemented in DBM, Perl5[94] and mini-SQL (http://Hughes.com.au). This was described previously.[12] It has tables cross-referencing sequence identifiers, structure matches, TM-helix positions, and so forth, as well as cross-tabulation reports giving the occurrence of various patterns. Most of these tables and reports will be made available over the Internet (as text tables and via a simple query interface) from the following URL: http://bioinfo.mbb.yale.edu/genome. The tables are structured in such a way that all the genome features (e.g., location of a TM-helix or PDB match) are annotated in a consistent fashion, with thresholds and scoring schemes applied consistently over multiple tools. This attempt at consistency is similar to what has been achieved in other genome annotation systems that aim to integrate multiple tools.[29,95]

Matching to Known Structures

All sequence comparison was done with the FASTA program (version 2.0) with k-tup 1 and an "e-value" threshold of 0.01.[96,97] The e-value describes the number of false positives expected in a single database scan, so a value of 0.01 means that about 1 out of 100 cluster linkages will be in error.[9,98–102] This error rate has been verified by empirical tests on a database of known protein relationships and is similar to the thresholds used in other multigenome comparisons.[9,16,98] Probabilistic scores, such as the e-value, should give similar results to more conventional scores, such as percent identity, but they have been shown to be better calibrated and more sensitive for marginal similarities, taking into account compositional biases of the databank and the query sequence.[99,102–104]

There are other, potentially more sensitive, methods of comparing sequences to structures than FASTA, e.g., profiles, Hidden-Markov models, motif analysis, secondary structure matching, and threading.[105–110] A number of these were tested, and as expected, they find more homologues for certain folds. However, the sensitivity improvement is not uniform over all folds. This is not advantageous for a large-scale census in which uniform sampling and treatment of the data are more important than sensitivity. In this instance one is more concerned with accurate relative numbers than with absolute values. Cobbling together a census through the use of a disparate collection of tools and patterns creates the problem of devising consistent scores and thresholds. This is particularly acute in the case of manually derived sequence patterns and motifs, since an expert on a particular fold or motif would expect his or her pattern to find relatively more homologues than a pattern not constructed by an expert. The approach here, applying the same objective procedure to each fold, circumvents these problems to some degree. Furthermore, it has the added advantage that it can be performed automatically without manual intervention and, consequently, can easily be scaled up to deal with much larger data sets.

Another issue to consider with regard to matching sequences to structures has to do with the fact that protein structure is fundamentally arranged around the level of folding domains whereas statistics for genomes are often calculated and best understood in terms of the number of genes. For instance, when one talks about how prevalent the kinase and Rossmann folds are in the yeast and *E. coli* genomes, one is implicitly comparing the number of matches that known kinase and Rossmann fold structures have in the ~6,200 yeast ORFs relative to the ~4,300 *E. coli* ORFs. However, it is possible for a single gene to contain a number of kinase fold domains or to contain simultaneously both a kinase and Rossmann fold. Thus, the total number of domains in a genome is probably a better standard for these comparisons. Unfortunately, this number is not known. However, it is known that the number of domains is not related simply to the number of genes. For instance, on average a protein is about 50% larger in yeast than in *E. coli* (317 vs. 466), meaning that there are probably twice as many possible domains in yeast as in *E. coli*. Here an intermediate approach is taken. The statistics are reported in terms of the number of domains matched but reference is always made to the number of ORFs in the genome.

Clustering and Trees

The structures in the PDB were clustered into 990 representative domains. The few membrane protein structures in the PDB were excluded from this clustering so that all the membrane proteins would be identified, in a uniform fashion, by prediction. (This is not expected to be a major factor as, for instance, the yeast genome contains only a single homologue to a known membrane protein structure). The clustering was similar in spirit to the many previous divisions of the PDB into representative chains.[68,98,111–113] However, a slightly different multiple-linkage algorithm was used.[114] It was designed to be internally consistent with the search method used to identify homologues in the genomes, using the same similarity criteria (a FASTA e-value threshold). The clustering algorithm takes the results of an

all-vs.-all comparison of the PDB and creates a graph that has one vertex for each sequence and one edge for each similarity score. Each vertex starts out as a cluster of size one. Since sequence similarity scores (i.e., e-values) are not commutative, this directed graph is converted to an undirected graph by removing the better scoring edges between pairs. Then, each edge is considered in turn, and the two clusters associated by this edge are merged into a single cluster if every member of the first cluster has a good scoring edge between it and every member of the second cluster, and vice versa. The edges are considered in order of decreasing similarity. This has the advantage that close relationships are considered before more distant ones, ensuring that distant relationships are not erroneously used to add a member to a cluster when there exists (for that member) a much closer relationship that would lead to an alternate clustering. Furthermore, this algorithm will produce the same result on the same data set every time, i.e., it is not affected by the order in which the data is traversed.

Trees based on distance matrices were built with simple UPGMA clustering using the Kitsch program, which is part of the Phylip package.[115,116] Trees were built on the basis of the difference in amino acid composition vectors, as described in the legend to Figure 3. Di-amino acid composition was also used and gave a similar tree.

After the clustering was completed, it was found that the PDB consisted of 990 nonhomologous domains, each of which represents a single sequence family. These 990 domains were grouped into 338 fold families by the structural relationships in SCOP.[58] Each of the 338 folds can be ranked in terms of how many of the 990 sequence families it contains. It was decided to define a superfold as one of the top 25 folds in terms of the number of associated sequence families. Each of these contains at least 10 sequence families. This threshold is arbitrary and is similar but not identical to past usage.[59]

Transmembrane Helix and Low-Complexity Region Identification

Transmembrane segments were identified using the GES hydrophobicity scale.[117] The values from the scale for amino acids in a window of size 20 (the typical size of a transmembrane helix) were averaged and then compared against a cutoff of -1 kcal/mole. A value under this cutoff was taken to indicate the existence of a transmembrane helix. Initial hydrophobic stretches corresponding to signal sequences for membrane insertion were excluded. (These have the pattern of a charged residue within the first 7, followed by a stretch of 14 with an average hydrophobicity under the cutoff.) These parameters have been used, tested, and refined on surveys of membrane protein in genomes.[50,53,55]

Low-complexity sequences were identified with the SEG program[65,118,119] using the standard parameters K(1) = 3.4 and K(2) = 3.75, and a window of length 45. These parameters are the ones used to find "long" domain-size low-complexity regions. The average size of a low-complexity region found here is ~110 residues. Many of these transmembrane regions are also low-complexity regions (almost half). Taking a conservative approach, it was decided to annotate these doubly identified regions as low-complexity, not as transmembrane. This will tend to reduce the total amount of identified TM-helices. This is especially true for MJ, which has the largest amount of low-complexity regions.

ACKNOWLEDGMENTS

Thanks to Guy Plunket and Mike Cherry, for providing information about the genome data; Steve Chervitz, for information about Sacch3D; George Weinstock and Steve Norris, for providing information on transmembrane folds; Hedi Hegyi, for carefully reading the manuscript; Ted Johnson, for help in structure clustering; and Janice Murphy, for manuscript preparation.

REFERENCES

1. Nowak, R. Bacterial genome sequence bagged. Science 269:468–470, 1995.
2. Langreth, R. Scientists unlock sequence of ulcer bacterium's genes. Wall Street Journal B1, August 7, 1997.
3. Wade, N. Thinking small paying off big in gene quest. New York Times, 3 February 1997, A1.
4. Blaisdell, B.E., Campbell, A.M., Karlin, S. Similarities and dissimilarities of phage genomes. Proc. Natl. Acad. Sci. USA 93:5854–5859, 1996.
5. Karlin, S., Burge, C. Dinucleotide relative abundance extremes: a genomic signature [Review]. Trends Genet. 11:283–290, 1995.
6. Karlin, S., Burge, C., Campbell, A.M. Statistical analyses of counts and distributions of restriction sites in DNA sequences. Nucleic Acids Res. 20:1363–1370, 1992.
7. Karlin, S., Mrazek, J., Campbell, A.M. Frequent oligonucleotides and peptides of the *Haemophilus influenzae* genome. Nucleic Acids Res. 24:4263–4272, 1996.
8. Koonin, E.V., Mushegian, A.R., Rudd, K.E. Sequencing and analysis of bacterial genomes. Curr Biol. 6:404–416, 1996.
9. Brenner, S., Hubbard, T., Murzin, A., Chothia, C. Gene duplication in *H. influenzae*. Nature 378:140, 1995.
10. Riley, M. Genes and proteins of *Escherichia coli* K-12 (GenProtEC). Nucleic Acids Res 25:51–52, 1997.
11. Wolfe, K.H., Shields, D.C. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713, 1997.
12. Gerstein M. A structural census of genomes: comparing eukaryotic, bacterial and archaeal genomes in terms of protein structure. J. Mol. Biol. 274:562–576, 1997.
13. Tamames, J., Casari, G., Ouzounis, C., Valencia, A. Conserved clusters of functionally related genes in two bacterial genomes. J. Mol. Evol. 44:66–73, 1997.
14. Green, P., Lipman, D., Hillier, L., Waterston, R., States, D., Claverie, J.M. Ancient conserved regions in new gene sequences and the protein databases. Science 259:1711–1716, 1993.
15. Koonin, E.V., Tatusov, R.L., Rudd, K.E. Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications. Proc. Natl. Acad. Sci. USA 92:11921–11925, 1995.

16. Tatusov, R.L., Koonin, E.V., Lipman, D.J. A genomic perspective on protein families. Science 278:631–637, 1997.
17. Ouzounis, C., Kyrpides, N., Sander, C. Novel protein families in Archaean genomes. Nucleic Acids Res. 23:565–570, 1995.
18. Ouzounis, C., Bork, P., Casari, G., Sander, C. New protein functions in yeast chromosome VIII. Protein Sci. 4:2424–2428, 1995.
19. Clayton, R.A., White, O., Ketchum, K.A., Venter, J.C. The first genome from the third domain of life [news]. Nature 387:459–462, 1997.
20. Karp, P., Riley, M., Paley, S., Pellegrini-Toole, A. EcoCyc: electronic encyclopedia of *E. coli* genes and metabolism. Nucleic Acids Res. 24:32–40, 1996.
21. Karp, P.D., Ouzounis, C., Paley, S.M. HinCyc: a knowledge base of the complete genome and metabolic pathways of *H. influenzae*. In: "Proceedings of the Fourth International Conference on Intelligence Systems in Molecular Biology." Menlo Park, CA: AAAI Press, 1996:116–124.
22. Mushegian, A.R., Koonin, E.V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes [see comments]. Proc. Natl. Acad. Sci. USA 93:10268–10273, 1996.
23. Tatusov, R.L., Mushegian, A.R., Bork, P., et al. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. Curr. Biol. 6:279–291, 1996.
24. Fleischmann, R.D., Adams, M.D., White, O., et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd. Science 269:496–512, 1995.
25. Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., Sonnhammer, E. What's in a genome? Nature 358:287, 1992.
26. Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., Sonnhammer, E. Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome iii. Protein Sci. 1:1677–1690, 1992.
27. Scharf, M., Schneider, R., Casari, G., et al. GeneQuiz: a workbench for sequence analysis. In: "Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology." Menlo Park, CA: AAAI Press, 1994:348–353.
28. Casari, G., Andrade, M., Bork, P., et al. Challenging times for bioinformatics. Nature 376:647–648, 1995.
29. Gaasterland, T., Sensen, C.W. Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. Biochimie 78:302–310, 1996.
30. Gerstein, M., Altman, R. Average core structures and variability measures for protein families: application to the immunoglobulins. J. Mol. Biol. 251:161–175, 1995.
31. Gerstein, M., Altman, R. A structurally invariant core for the globins. CABIOS 11:633–644, 1995.
32. Henikoff, S., Henikoff, J.G. Automated assembly of protein blocks for database searching. Proc. Natl. Acad. Sci. USA 19:6565–6572, 1993.
33. Henikoff, S., Henikoff, J.G. Protein family classification based on searching a database of blocks. Genomics 19:97–107, 1994.
34. Henikoff, S., Greene, E.A., Pietrokovski, S., Bork, P., Attwood, T.K., Hood, L. Gene families: the taxonomy of protein paralogs and chimeras. Science 278:609–614, 1997.
35. Sonnhammer, E.L.L., Kahn, D. Modular arrangement of proteins as inferred from analysis of homology. Protein Sci. 3:482–492, 1994.
36. Sonnhammer, E., Eddy, S., Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28:405–420, 1997.
37. Riley, M., Labedan, B. Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module. J. Mol. Biol. 268:857–868, 1997.
38. Fabian, P., Murvai, J., Hatsagi, Z., Vlahovicek, K., Hegyi, H., Pongor, S. The SBASE protein domain library, release 5.0: a collection of annotated protein sequence segments. Nucleic Acids Res. 25:240–243, 1997.
39. Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826, 1986.
40. Chothia, C., Gerstein, M. Protein evolution. How far can sequences diverge? Nature 385:579–581, 1997.
41. Chothia, C. Proteins—1000 families for the molecular biologist. Nature 357:543–544, 1992.
42. Lander, E.S. The genomics: global views of biology. Science 274:536–539, 1996.
43. Lesk, A.M., Levitt, M., Chothia, C. Alignment of amino acid sequences of distantly related proteins using variable gap penalties. Protein Eng. 1:77–78, 1986.
44. Bork, P., Ouzounis, C., Sander, C. From genome sequences to protein function. Curr. Opin. Struct. Biol. 4:393–403, 1994.
45. Gerstein, M., Levitt, M. A structural census of the current population of protein sequences. Proc. Natl. Acad. Sci. USA 94:11911–11916, 1997.
46. Fischer, D., Eisenberg, D. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium* [In Process Citation]. Proc. Natl. Acad. Sci. USA 94:11929–11934, 1997.
47. Goffeau, A., Slonimski, P., Nakai, K., Risler, J.L. How many yeast genes code for membrane-spanning proteins? Yeast 9:691–702, 1993.
48. Rost, B., Fariselli, P., Casadio, R., Sander, C. Prediction of helical transmembrane segments at 95% accuracy. Protein Sci. 4:521–533, 1995.
49. Rost, B. PHD: predicting one-dimensional protein secondary structure by profile-based neural networks. Methods Enzymol. 266:525–539, 1996.
50. Arkin, I., Brunger, A., Engelman, D. Are there dominant membrane protein families with a given number of helices? Proteins 28:465–466, 1997.
51. Boyd, D., Schierle, C.,Beckwith, J. How many membrane proteins are there? Protein Sci. 7:201–205, 1998.
52. Jones, D.T. Do transmembrane protein superfolds exist? FEBS Lett. 423:281–285, 1998.
53. Tomb, J.-F., White, O., Kerlavage, A.R., et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388:539–547, 1997.
54. Fraser, C.M., Casjens, S., Huang, W.M., et al. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi* [see comments]. Nature 390:580–586, 1997.
55. Wallin, E., von Heijne, G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms [In Process Citation]. Protein Sci 7:1029–1038, 1998.
56. Gibrat, J.F., Madej, T., Bryant, S.H. Surprising similarities in structure comparison. Curr. Opin. Struct. Biol. 6:377–385, 1996.
57. Holm, L., Sander, C. Mapping the protein universe. Science 273:595–602, 1996.
58. Murzin, A., Brenner, S.E., Hubbard, T., Chothia, C. SCOP: a structural classification of proteins for the investigation of sequences and structures. J. Mol. Biol. 247:536–540, 1995.
59. Orengo, C.A., Jones, D.T., Thornton, J.M. Protein superfamilies and domain superfolds. Nature 372:631–634, 1994.
60. Schmidt, R. Gerstein, M.,Altman, R. LPFC: an Internet library of protein family core structures. Protein Sci. 6:246–248, 1997.
61. Pascarella, S., Argos, P. A databank merging related protein structures and sequences. Protein Eng. 5:121–137, 1992.
62. Sander, C., Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68, 1991.
63. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M. CATH—a hierarchic classification of protein domain structures. Structure 5:1093–1108, 1997.

64. Weiss, M.S., Abele, U., Weckesser, J., Welte, W., Schiltz, E., Schulz, G.E. Molecular architecture and electrostatic properties of a bacterial porin. Science 254:1627–1630, 1991.
65. Wootton, J.C., Federhen, S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol. 266:554–571, 1996.
66. Champion, C.I., Blanco, D.R., Exner, M.M., et al. Sequence analysis and recombinant expression of a 28-kilodalton *Treponema pallidum* subsp. *pallidum* rare outer membrane protein (Tromp2). J. Bacteriol. 179:1230–1238, 1997.
67. Abola, E., Sussman, J., Prilusky, J., Manning, N. Protein Data Bank archives of three-dimensional macromolecular structures. Methods Enzymol. 277:556–571, 1997.
68. Brenner, S.E., Chothia, C., Hubbard, T.J. Population statistics of protein structures: lessons from structural classifications [In Process Citation]. Curr. Opin. Struct. Biol. 7:369–376, 1997.
69. Doolittle, R.F. The multiplicity of domains in proteins. [Review]. Annu. Rev. Biochem. 64:287–314, 1995.
70. Govindarajan, S., Goldstein, R.A. Why are some proteins structures so common? Proc. Natl. Acad. Sci. USA 93:3341–3345, 1996.
71. Li, H., Helling, R., Tang, C., Wingreen, N. Emergence of preferred structures in a simple model of protein folding [see comments]. Science 273:666–669, 1996.
72. Sonnhammer, E.L., Durbin, R. Analysis of protein domain families in *Caenorhabditis elegans*. Genomics 46:200–216, 1997.
73. Ouzounis, C., Kyrpides, N. The emergence of major cellular processes in evolution. FEBS Lett. 390:119–123, 1996.
74. Sneath, P.H.A., Sokal, R.R. "Numerical Taxonomy." San Francisco, W.H. Freeman, 1973.
75. Olsen, G.J., Woese, C.R., Overbeek, R. The winds of (evolutionary) change: breathing new life into microbiology. J. Bacteriol. 176:1–6, 1994.
76. Doolittle, R.F. A bug with excess gastric avidity [news; comment]. Nature 388:515–516, 1997.
77. Levitt, M., Chothia, C. Structural patterns in globular proteins. Nature 261:552–558, 1976.
78. Velculescu, V.E., Zhang, L., Zhou, W., et al. Characterization of the yeast transcriptome. Cell 88:243–251, 1997.
79. Lashkari, D.A., DeRisi, J.L., McCusker, J.H., et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc. Natl. Acad. Sci. USA 94:13057–13062, 1997.
80. DeRisi, J.L., Iyer, V.R., Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680–686, 1997.
81. Knuth, D. "The Art of Computer Programming: Vol. 3, Sorting and Searching." Readin, MA: Addison-Wesley, 1973.
82. Konopka, A.K., Martindale, C. Noncoding DNA, Zipf's law, and language [letter]. Science 268:789, 1995.
83. Flam, F. Hints of a language in junk DNA [news] [see comments]. Science 266:1320, 1994.
84. Bornberg-Bauer, E. How are model protein structures distributed in sequence space? [In Process Citation]. Biophys. J. 73:2393–2403, 1997.
85. Cherry, J.M., Adler, C., Ball, C., et al. SGD: Saccharomyces Genome Database. Nucleic Acids Res. 26:73–79, 1998.
86. Kerlavage, A.R. TIGR Microbial Genome Database. http://www.tigr.org/mdb (as of 2/97), 1997.
87. Blattner, F.R., Plunkett, G.P., III, Bloch, C.A., et al. The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1462, 1997.
88. Goffeau, A., Barrell, B.G., Bussey, H., et al. Life with 6000 genes. Science 274:546–567, 1996.
89. Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A., Voytas., D.F. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res. 8:464–478, 1998.
90. Stampf, D.R., Felder, C.E., Sussman, J.L. PDBbrowse—a graphics interface to the Brookhaven Protein Data Bank. Nature 374:572–574, 1995.
91. Hubbard, T.J.P., Murzin, A.G., Brenner, S.E., Chothia, C. SCOP: a structural classification of proteins database. Nucleic Acids Res. 25:236–239, 1997.
92. Altman, R., Gerstein, M. Finding an average core structure: application to the globins. In: "Proceedings of the Second International Conference on Intelligent Systems in Molecular Biology." Menlo Park, CA: AAAI Press, 1994:19–27.
93. Gerstein, M., Levitt, M. Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. In: "Proceedings of the Fourth International Conference on Intelligent Systems in Molecular Biology." Menlo Park, CA: AAAI Press, 1996:59–67.
94. Wall, L., Christiansen, D., Schwartz, R. "Programming Perl." Sebastapol, CA: O'Reilly and Associates, 1996.
95. Medigue., C., Moszer, I., Viari, A., Danchin, A. Analysis of a *Bacillus subtilis* genome fragment using a co-operative computer system prototype. Gene 165:GC37–51, 1995.
96. Lipman, D.J., Pearson, W.R. Rapid and sensitive protein similarity searches. Science 227:1435–1441, 1985.
97. Pearson, W.R., Lipman, D.J. Improved tools for biological sequence analysis. Proc. Natl. Acad. Sci. USA 85:2444–2448, 1988.
98. Brenner, S., Chothia, C., Hubbard, T. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc. Natl. Acad. Sci. USA 95:6073–6078, 1998.
99. Pearson, W.R. Empirical statistical estimates for sequence similarity searches. J. Mol. Biol. 276:71–84, 1998.
100. Pearson, W.R. Effective protein sequence comparison. Methods Enzymol. 266:227–259, 1996.
101. Pearson, W.R. Identifying distantly related protein sequences. Comput. Appl. Biosci. 13:325–332, 1997.
102. Levitt, M., Gerstein, M. A unified statistical framework for sequence comparison and structure comparison. Proc. Natl. Acad. Sci. USA 95:5913–5920, 1998.
103. Altschul, S.F., Boguski, M.S., Gish, W., Wootton, J.C. Issues in searching molecular sequence databases. [Review]. Nature Genet. 6:119–129, 1994.
104. Karlin, S., Altschul, S.F. Applications and statistics for multiple high-scoring segments in molecular sequences. Proc. Natl. Aad. Sci. USA 90:5873–5877, 1993.
105. Bowie, J.U., Eisenberg, D. Inverted protein structure prediction. Curr. Opin. Struct. Biol. 3:437–444, 1993.
106. Jones, S., Thornton, J. Principles of protein-protein interactions. Proc. Natl. Acad. Sci. USA 93:13–20, 1996.
107. Eddy, S.R. Hidden Markov models. Curr. Opin. Struct,. Biol. 6:361–365, 1996.
108. Tatusov, R.L., Altschul, S.F., Koonin, E.V. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc. Natl. Acad. Sci. USA 91:12091–12095, 1994.
109. Dubchak, I., Muchnik, I., Holbrook, S.R., Kim, S.H. Prediction of protein folding class using global description of amino acid sequence. Proc. Natl. Acad. Sci. USA 92:8700–8704, 1995.
110. Aurora, R., Rose, G.D. Seeking an ancient enzyme in *Methanococcus jannaschii* using ORF, a program based on predicted secondary structure comparisons. Proc. Natl. Acad. Sci. USA 95:2818–2823, 1998.
111. Hobohm, U., Scharf, M., Schneider, R., Sander, C. Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Sci. 1:409–417, 1992.
112. Hobohm, U., Sander, C. Enlarged representative set of protein structures. Protein Sci. 3:522, 1994.
113. Boberg, J., Salakoski, T., Vihinen, M. Selection of a representative set of structures from Brookhaven Protein Data Bank. Proteins 14:265–276, 1992.
114. Kaufman, L., Rousseeuw, P.J. "Finding Groups in Data: An Introduction to Cluster Analysis." New York: John Wiley & Sons, 1990.

115. Felsenstein, J. PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 5:164–166, 1989.
116. Felsenstein, J. "PHYLIP (Phylogeny Inference Package) version 3.5c." Seattle: Department of Genetics, University of Washington, 1993.
117. Engelman, D.M., Steitz, T.A., Goldman, A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. [Review]. Annu. Rev. Biophys. Biophys. Chem, 15:321–353, 1986.
118. Wootton, J.C., Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. Comput. Chem. 17:149–163, 1993.
119. Wootton, J.C. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput. Chem. 18:269–285, 1994.
120. Fraser, C.M., Gocayne, J.D., White, O., et al. The minimal gene complement of *Mycoplasma genitalium* [see comments]. Science 270:397–403, 1995.
121. Bult, C.J., White, O., Olsen, G.J., et al. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273:1058–1073, 1996.
122. Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.C., Herrmann, R. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. 24:4420–4449, 1996.
123. Goffeau, A., Aert, R., Agostini-Carbone, M.L., et al. The Yeast Genome Directory. Nature 387(Suppl.):5–105, 1997.
124. Kaneko, T., Sato, S., Kotani, H., et al. 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, 1996.
125. Amari, S. Ann. Stat. 10:357–387, 1982.
126. Kraulis, P.J. MOLSCRIPT—a program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. 24:946–950, 1991.
127. Gerstein, M., Hegyi, H. Comparing microbial genomes in terms of protein structure: Surveys of a finite parts list. FEMS Microbiol. Rev. 1998, in press.
128. Gerstein, M. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. Folding & Design 1998, in press.