Distribution of Protein Folds in the Three Superkingdoms of Life

Yuri I. Wolf,[1,4] Steven E. Brenner,[2] Paul A. Bash,[3] and Eugene V. Koonin[1,5]

[1]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894 USA;* [2]*Department of Structural Biology, Stanford University, Stanford, California 94305-5126 USA;* [3]*Department of Molecular Pharmacology and Biological Chemistry, Northwestern University, Chicago, Illinois 60611 USA*

A sensitive protein-fold recognition procedure was developed on the basis of iterative database search using the PSI-BLAST program. A collection of 1193 position-dependent weight matrices that can be used as fold identifiers was produced. In the completely sequenced genomes, folds could be automatically identified for 20%–30% of the proteins, with 3%–6% more detectable by additional analysis of conserved motifs. The distribution of the most common folds is very similar in bacteria and archaea but distinct in eukaryotes. Within the bacteria, this distribution differs between parasitic and free-living species. In all analyzed genomes, the P-loop NTPases are the most abundant fold. In bacteria and archaea, the next most common folds are ferredoxin-like domains, TIM-barrels, and methyltransferases, whereas in eukaryotes, the second to fourth places belong to protein kinases, β-propellers and TIM-barrels. The observed diversity of protein folds in different proteomes is approximately twice as high as it would be expected from a simple stochastic model describing a proteome as a finite sample from an infinite pool of proteins with an exponential distribution of the fold fractions. Distribution of the number of domains with different folds in one protein fits the geometric model, which is compatible with the evolution of multidomain proteins by random combination of domains.

[Fold predictions for proteins from 14 proteomes are available on the World Wide Web at ftp://ncbi.nlm.nih.gov/pub/koonin/FOLDS/index.html. The FIDs are available by anonymous ftp at the same location.]

Knowledge of the three-dimensional structures of proteins is indispensable for understanding biological processes. Ideally, determination of the structures of all proteins encoded in a genome should follow genome sequencing promptly. In reality, the recent substantial progress in experimental structural biology notwithstanding, structures are being determined for only a miniscule fraction of the gene products even for a bacterial genome containing a few thousand genes, not to mention the human genome with its estimated 100,000 genes (Holm and Sander 1996). Fortunately, however, considerable information on protein structure can be extracted by computer from sequence alone. This stems from two related principles of protein sequence-structure relationships: (a) there is only a limited number of distinct protein folds, perhaps no more than 1000 altogether, and ~400, presumably the most common ones, are already represented by experimentally determined structures (Dorit et al. 1990; Chothia 1992; Hubbard et al. 1997); (b) proteins with similar sequences tend to have similar structures; in homologous proteins, structure is generally more conserved than sequence, and therefore even subtle but reliable sequence similarity is likely to signify structure conservation (Doolittle 1981; Holm and Sander 1996, 1997).

The latter principle is essentially a recast of the Anfinsen's postulate: A protein's sequence determines its structure (Anfinsen and Scheraga 1975). Theoretically, structure should thus be predictable from sequence. Currently, however, such ab initio prediction is possible only for peptides and very small proteins (Abagyan 1997; Ortiz et al. 1998). Therefore for typical proteins, the only practical route to deriving structural information from sequence is through similarity to proteins with known structures, and success of structure prediction critically depends on the resolution and robustness of the methods used to detect such similarity. There are two basic categories of such methods: (1) sequence similarity analysis; and (2) sequence–structure threading (Godzik and Skolnick 1992; Bryant and Altschul 1995; Murzin and Bateman 1997). The threading approaches have been designed to address the problem of sequence-based structure prediction directly by assessing the compatibility of a given sequence with each known structure. These methods, however, generally lack statistical rigor and are computationally expensive (Lathrop 1994; T.F. Smith et al. 1997). Sequence similarity search is much faster, and at least the most popular method, BLAST, has a solid statistical foundation (Karlin and Altschul 1990; Karlin et al. 1991). The problem with these methods is that, as shown by a recent extensive evaluation, they detect only a small fraction of all homologous relationships

[4]*Permanent address: Institute of Cytology and Genetics, Russian Academy of Sciences, Novosibirsk 630090, Russia.*
[5]**Corresponding author.**
E-MAIL koonin@ncbi.nlm.nih.gov; FAX (301) 480-9241.

9:17–26 ©1999 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/99 $5.00; www.genome.org

that can be inferred from the comparison of the known protein structures (Brenner et al. 1998).

Accumulation of complete genome sequences from several bacteria, archaea, and eukaryotes creates new possibilities for assessing the phylogenetic distribution of protein folds in connection to organism phenotypes. Clearly, such a survey will be meaningful only if for each known fold, the majority of the representatives are recognized correctly. Recently several attempts have been made to analyze fold distribution in the complete protein sequence database (Gerstein and Levitt 1997) or in individual proteomes (Fischer and Eisenberg 1997; Gerstein 1997). These efforts relied primarily on standard methods for sequence comparison whose relatively low performance in fold recognition has been demonstrated (Brenner et al. 1998), with an additional contribution from secondary structure-based threading (Fischer and Eisenberg 1997). We sought to increase the sensitivity of fold recognition by using position-dependent weight matrices that are produced by the PSI-BLAST program concomitantly with database search (Altschul et al. 1997). In several studies on individual protein families, PSI-BLAST has demonstrated its ability to detect subtle sequence similarities that led to fold prediction, in part already confirmed by experiment (Mushegian et al. 1997; Aravind et al. 1998a,b; Aravind and Koonin 1998). We reasoned that matrices produced by PSI-BLAST could serve as sensitive identifiers for protein folds and proceeded to develop such identifiers for all folds present in the Structural Classification of Proteins (SCOP) database (Murzin et al. 1995; Hubbard et al. 1997) and to apply them for a comparative analysis of fold distribution in complete proteomes of bacteria, archaea, and eukaryotes.

RESULTS AND DISCUSSION

The Fold Recognition Procedure

The starting material for our fold recognition protocol (Fig. 1) was the set of protein sequences represented in SCOP 1.35, in which individual structural domains have been isolated. With the goal of increasing the resolution power of the resulting profiles, these domain sequences were enriched with those of obvious homologs from the nonredundant (NR) protein sequence database at the National Center for Biotechnology Information (NCBI) and then clustered by sequence similarity to select representative sequences for fold recognition (see Methods). The resulting 1193 domains belong to significantly different proteins with reliable fold assignment, classified by fold (fold representative sequences or FRS). Each FRS was used as a starting point for iterative PSI-BLAST search of the NR database, producing significant hits to 77,279 proteins, which comprises ~27% of the entire database. Compared to single-pass searches, the iterative searches retrieved a total of 26,275 extra hits corresponding to 228 out of the 284 folds. The current version of PSI-BLAST has the option of saving position-dependent weight matrices constructed during the iterative search. Such a matrix contains information on all the database sequences significantly similar to a FRS and can be used to search another database, greatly increasing the sensitivity and selectivity compared to a search with a single query sequence. Thus the 1193 matrices produced for each of the FRS by the PSI-BLAST searches were stored for subsequent use as fold identifiers (FIDs).

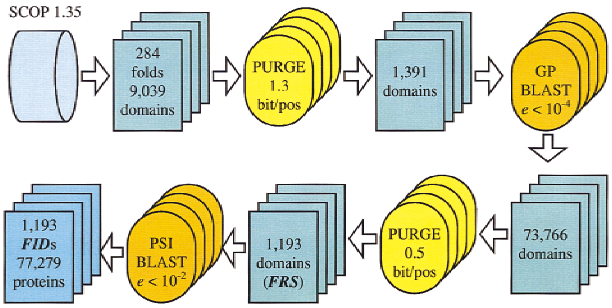

Figure 1 The fold recognition procedure.

Under this approach, a fold may be represented by one or multiple FIDs, depending on the number of FRS. Cross-recognition between FIDs (in other words, overlap between PSI-BLAST outputs) within one fold measures the ability of the method to detect subtle similarities that escape standard sequence comparison procedures. Of the 176 folds with more than one FRS, 74 (42%) showed perfect intrafold recognition (there was overlap within each pair of outputs), 58 (33%) showed partial intrafold recognition, and in the remaining 44 (25%), there was no recognition between different FIDs.

In contrast, recognition between different folds typically is false; thus overlaps between the database search results for FRS representing different folds should be considered false positives for one or both of the two folds involved. To estimate the error rates conservatively, both assignments were counted as false positives. (Parenthetically, it should be noticed that this may not be true in the rare cases in which certain folds in the SCOP classification seemed to have been split artificially. Thus it was noticed that two pairs of folds, namely seven- and eight-bladed β-propellers and two Rossmann-like nucleotide-binding folds, are in fact related closely at the sequence level. These folds were combined for the purpose of this analysis). Of the 284 folds included in our analysis, for 198 (70%), no false positives were detected. For the remaining folds, point estimates of the false-positive rates were obtained after clustering the complete sets of hits and the sets of overlaps, to account for nonindependent (homologous) sequences.

A point estimate of the error rate may give false confidence when the number of involved cases is small. To obtain an interval estimate, we assumed that clusters of database hits for each FRS were obtained independently (a realistic assumption given the low cut-off used for clustering; see Methods). A Bernoulli model was then applied to find the upper limit for a background error rate that may lead to the given number of independent false positives out of the given number of independent clusters. The upper limits of the 95% confidence interval of the false positive rate for the most common folds are shown in Table 1. With the exception of two cases, the maximum expected error rate is well below 10%. This is in a good agreement with the empirical results (see below).

Whereas the evaluation of the false-positive rate in fold recognition is more or less straightforward, the critical issue of false negatives (that is, how many proteins with a known fold are missed) is much harder to address. Some estimates, however, could be made concomitantly with a detailed analysis of the distribution of predicted folds in complete proteomes as discussed below.

Phylogenetic Distribution of Protein Folds

At the time of this analysis (April 1998), 13 complete genome sequences were publicly available: *Haemophilus influenzae* (Fleischmann et al. 1995), *Mycoplasma genitalium* (Fraser et al. 1995), *Mycoplasma pneumoniae* (Himmelreich et al. 1996), *Synechocystis* sp. (Kaneko and Tabata 1997), *Helicobacter pylori* (Tomb et al. 1997), *Escherichia coli* (Blattner et al. 1997), *Bacillus subtilis* (Kunst et al. 1997), *Borrelia burgdorferi* (Fraser et al. 1997), *Aquifex aeolicus* (Deckert et al. 1998), *Methanococcus jannaschii* (Bult et al. 1996), *Methanobacterium thermoautotrophicum* (D.R. Smith et al. 1997), *Archaeoglobus fulgidus* (Klenk et al. 1997), and *Saccharomyces cerevisiae* (Goffeau et al. 1996). In addition, the proteome of the nematode *Caenorhabditis elegans* that was ~85% complete (Kuwabara 1997) also was included in the analysis (Table 2).

Table 1. Upper Limit of 95% Confidence Interval for False-Positive Rate

Fold[a]	False-positive rate (%)
β-propeller	2.1
PLP-dependent transferases	2.5
α/β-hydrolase	2.8
SAM-methyltransferases	3.0
Zn-β-lactamases	3.3
Periplasmic-binding II	3.4
Ferredoxin-like	3.7
ATP-pyrophosphatases	4.3
TIM-barrel	5.3
P-loop NTPases	5.6
RNase H-like	5.8
NR-ligand binding	6.1
Rossmann-like	6.2
Protein kinases	6.3
Flavodoxin-like	7.6
Rossmann-fold	9.6
ATP-grasp	17.1
Class II aaRS and biotin synthetases	23.9

[a] Abbreviated SCOP 1.35-fold names.

Fold assignment was performed by searching the sequences from each of these proteomes using the PSI-BLAST program (a single pass), with each of the 1193 FIDs as the query. All hits with an *e*-value $\leq 10^{-2}$ after an adjustment to the NR database size were considered automatic fold assignments. For the 30 most common folds, additional, case-by-case analysis was performed by searching for the conservation of motifs typical of known protein families in the outputs of the FID-initiated searches (regardless of the statistical significance). Additionally, all the sequences from complete proteomes were searched against the NR database using PSI-BLAST and the outputs were examined in the same fashion.

The results of this analysis (Table 2) show that the fraction of false positives (erroneous fold assignments) among automatic predictions typically was ~1%–2% (maximum 3.2%); the detected fraction of false negatives (additional assignments made by the case-by-case screening) was 9%–13% (maximum 13.3%). These findings suggest that FIDs predict protein folds at a genome scale with a reasonable reliability.

The overall fraction of proteins with fold assignments in the proteomes typically varied in the range of 24%–35% with a few exceptions: the highly compositionally biased proteome of *B. burgdorferi* and the incomplete proteome of *C. elegans*, which was analyzed only automatically, have the lowest fraction of proteins with assigned folds (19% and 21%, respectively), whereas the smallest known proteome of *M. genitalium* has the highest (39%). This prediction rate is considerably higher than those reported in the previous studies (Fischer and Eisenberg 1997; Gerstein 1997; Gerstein and Levitt 1997). Furthermore, the information is now available for a greater number of genomes, at least for bacteria and archaea. Thus, though the prediction evidently is still incomplete, it was of interest to explore some patterns in the fold distribution.

Figure 2 shows the distribution of predicted folds in the three superkingdoms of life, Bacteria, Archaea, and Eukarya. Almost one-half of the folds are universal. It is remarkable that nearly all folds found in archaea belong to this ubiquitous set, whereas a very small number is shared by archaea with bacteria or eukaryotes, to the exclusion of the third superkingdom. By contrast, over 20% of the recognized folds are shared by bacteria and eukaryotes, but not by archaea,

Table 2. Fold Assignment in Complete Proteomes

Species	No. of proteins	Automatically predicted	Total predicted[a]		False negatives[b]		False positives[c]		No. of recognized folds
M. genitalium	467	159	182	39.0%	24	13.2%	1	0.6%	81
M. pneumoniae	677	173	197	29.1%	26	13.2%	2	1.2%	84
B. burgdorferi	1,256	216	241	19.2%	32	13.3%	7	3.2%	94
A. aeolicus	1,521	481	530	34.8%	53	10.0%	4	0.8%	108
H. pylori	1,565	343	392	25.0%	50	12.8%	1	0.3%	112
H. influenzae	1,717	506	555	32.3%	53	9.5%	4	0.8%	144
Synechocystis sp.	3,169	826	894	28.2%	72	8.1%	4	0.5%	151
B. subtilis	4,099	1021	1124	27.4%	115	10.2%	12	1.2%	165
E. coli	4,289	1104	1212	28.3%	136	11.2%	28	2.5%	175
M. jannaschii	1,770	389	439	24.8%	56	12.8%	6	1.5%	98
M. thermoautotrophicum	1,893	459	509	26.9%	54	10.6%	4	0.9%	103
A. fulgidus	2,407	608	679	28.2%	75	11.0%	4	0.7%	112
S. cerevisiae	6,529	1462	1575	24.1%	148	9.4%	35	2.4%	165
C. elegans	13,743	2831	2831	20.6%	N/A	N/A	N/A	N/A	172

(N/A) Not applicable.
[a]Percentage of total prediction is indicated relative to the proteome size.
[b]Percentage of false negatives is indicated relative to total prediction figures
[c]Percentage of false positives is indicated relative to automatic prediction figures.

most likely caused by the transfer of bacterial genes from organellar genomes to the nuclear genomes of eukaryotes, and perhaps to additional horizontal transfer events. Whereas major gene exchange has most likely occurred also between bacteria and archaea (Koonin et al. 1997), it appears that these events involved primarily genes encoding proteins with ubiquitous folds, for example., central metabolic enzymes. The near absence of archaea-specific folds, which contrasts the considerable and almost equal number of specifically bacterial and eukaryotic folds, probably reflects the currently insufficient structural characterization of archaeal proteins.

In all three superkingdoms, the most common fold is the P-loop NTPase. Four folds, namely P-loop NTPases, TIM barrels, ferredoxin-like domains, and Rossmann fold domains, are present in all three top 10 lists (Table 3). The abundance of each of the common folds, but particularly P-loop NTPases and SAM-dependent methyltransferases (the third and fourth-ranking fold in bacteria and archaea, respectively), seems to have been underestimated in the previous studies (Gerstein 1997; Gerstein and Levitt 1997). The P-loops have not been detected as the most common fold in any genome or taxonomic division, whereas the methyltransferases never made the top 10 list at all, apparently because of the relative difficulty of their recognition. In agreement with the previous findings reported for a small set of complete genomes (Gerstein 1997), all top folds in bacteria and archaea, and 8 out of the top 10 folds in eukaryotes belong to two structural classes: α/β and mixed α + β proteins.

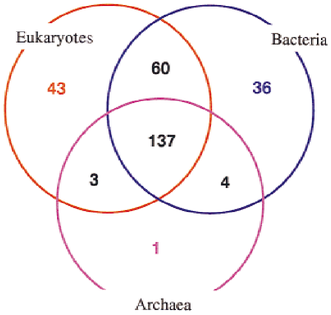

Figure 2 Distribution of the recognized folds in bacteria, archaea, and eukaryotes. The number of recognized folds is indicated for each part of the diagram.

The distributions of the most common folds in bacterial and archaeal proteomes are very similar (8 of the top 10 folds are the same; Table 3), though the much higher abundance of ferredoxin-like proteins and metallo-β-lactamase-like proteins and the underrepresentation of the Rossmann fold in archaea are notable. Eukaryotes show a different ranking of folds—five of the folds among the eukaryotic top 10 hits are not in the bacterial or archaeal top 10 lists, and one, namely the ligand-binding domains of nuclear receptors, is unique for eukaryotes (Table 3). In bacteria and archaea, the most common folds correspond to enzymes involved in genome replication and expression (e.g., ATPases and GTPases) and metabolic enzymes. Particularly notable is the abundance of methyltransferases (Table 3; see above), most of which are involved in modification of nucleic acids and proteins. By contrast, among the eukaryotic top 10 folds, proteins involved in regulation and signal transduction, such as protein kinases and β-propellers, are prominent; it is of further note that in the multicellular eukaryote *C. elegans*, protein kinases are the most common fold (Table 3). Perhaps unexpectedly, in spite of the great importance of methylation in the regulation of eukary-

Table 3. Top 30 Folds in Complete Proteomes

Fold	*Mg*	*Mp*	*Bb*	*Aa*	*Hp*	*Hi*	*Ss*	*Bs*	*Ec*	*Mj*	*Mt*	*Af*	*Sc*	*Ce*	B[a]		A[a]		E[a]		All[b]
P-loop NTPases	53	54	83	102	92	108	163	177	186	125	105	125	239	216	22.5%	*1*	22.5%	*1*	11.4%	*1*	18.8%
TIM-barrel	5	8	13	35	26	36	51	89	105	37	40	47	93	98	6.0%	*2*	7.7%	*3*	4.7%	*4*	6.1%
Ferredoxin-like	3	3	6	22	18	20	26	20	55	50	61	67	68	85	3.0%	*9*	11.1%	*2*	3.7%	*5*	5.9%
SAM methyltransferases	7	11	11	27	39	32	40	37	39	40	23	30	45	50	5.1%	*3*	6.0%	*4*	2.3%	*12*	4.5%
Protein kinases, catalytic core	1	1	0	2	1	1	11	3	3	3	2	3	128	305	0.4%	*27*	0.5%	*26*	9.5%	*2*	3.5%
Rossmann-fold domains	2	2	6	24	16	19	48	68	53	15	19	22	51	89	3.6%	*6*	3.5%	*8*	3.2%	*9*	3.4%
ATP pyrophosphatases	13	13	13	18	17	15	16	22	23	27	20	22	25	16	3.9%	*4*	4.4%	*5*	1.1%	*19*	3.1%
Rossmann-like fold domain	6	6	5	21	9	16	49	43	63	10	15	42	37	32	3.6%	*7*	3.8%	*7*	1.7%	*14*	3.0%
Flavodoxin-like	3	3	7	16	15	16	70	57	66	11	26	35	22	11	3.8%	*5*	4.3%	*6*	0.9%	*20*	3.0%
7- and 8-bladed β-propeller	0	0	1	1	0	2	10	5	6	2	2	3	132	131	0.3%	*28*	0.4%	*27*	6.5%	*3*	2.4%
Ribonuclease H-like motif	3	3	6	7	8	21	23	18	36	4	5	13	85	50	2.2%	*12*	1.3%	*17*	3.6%	*6*	2.4%
PLP-dependent transferases	2	2	2	17	10	16	23	44	39	15	15	18	32	30	2.4%	*11*	3.0%	*10*	1.5%	*15*	2.3%
α/β-Hydrolases	4	4	1	4	1	5	27	36	24	0	4	13	44	114	1.6%	*14*	0.9%	*23*	3.4%	*8*	2.0%
Class II aaRS and biotin synth.	9	9	8	10	8	11	10	11	12	10	11	10	20	13	2.4%	*10*	2.0%	*13*	0.9%	*21*	1.8%
L-2-Haloacid dehalogenase	4	4	4	7	8	10	17	21	25	6	12	6	30	25	1.9%	*13*	1.5%	*15*	1.4%	*16*	1.6%
DNA-binding 3-helical bundle	0	0	0	3	1	8	29	23	34	4	5	7	30	89	1.2%	*22*	1.0%	*22*	2.5%	*10*	1.5%
Periplasmic binding protein-like II	6	12	9	4	8	20	27	28	39	3	5	11	5	13	3.1%	*8*	1.1%	*20*	0.4%	*29*	1.5%
Zn metallo-beta-lactamase	2	2	2	5	2	3	14	16	8	15	11	25	8	7	1.0%	*23*	3.1%	*9*	0.4%	*30*	1.5%
ATP-grasp	2	2	1	11	5	6	11	15	12	13	10	16	11	14	1.2%	*20*	2.4%	*11*	0.6%	*25*	1.4%
Thiamin-binding	3	3	0	7	6	7	10	14	13	8	14	13	14	11	1.2%	*19*	2.2%	*12*	0.6%	*24*	1.3%
Thioredoxin-like	1	1	1	7	2	13	16	15	20	1	3	9	24	66	1.2%	*21*	0.7%	*25*	1.9%	*13*	1.3%
Reductase/elongation factor domain	4	4	5	7	5	8	9	8	15	7	6	6	20	11	1.5%	*15*	1.2%	*18*	0.8%	*22*	1.2%
Metallo-dependent phosphatases	1	1	1	4	3	4	6	13	8	7	8	8	19	44	0.7%	*26*	1.4%	*16*	1.4%	*17*	1.2%
Ligand-binding domain of NR	0	0	0	0	0	0	0	0	0	0	0	0	0	194	-	-	-	-	3.4%	*7*	1.1%
OB-fold	5	5	2	6	5	8	5	10	17	7	7	2	11	10	1.4%	*16*	1.1%	*21*	0.5%	*26*	1.0%
Phosphoribosyltransferases	4	4	4	7	4	10	7	10	8	5	6	7	13	5	1.4%	*17*	1.1%	*19*	0.5%	*27*	1.0%
Class I glutamine amidotransferases	0	0	2	6	4	6	7	11	9	7	8	10	12	2	0.7%	*25*	1.5%	*14*	0.4%	*28*	0.9%
Long alpha-hairpin	4	4	5	3	5	6	8	4	10	0	2	0	23	25	1.3%	*18*	0.1%	*28*	1.2%	*18*	0.9%
C-type lectin-like	0	0	0	0	0	0	0	0	0	0	0	0	0	138	-	-	-	-	2.4%	*11*	0.8%
Zn-dependent exopeptidases	1	1	2	3	2	6	4	19	16	3	4	5	11	20	0.8%	*24*	0.7%	*24*	0.7%	*23*	0.8%

The number of proteins in which the given fold was recognized is indicated for each proteome. Average percentage relative to the number of proteins with predicted folds is indicated for each superkingdom; cells for top 10 folds are shaded. The folds are sorted by overall rank.

[a]Average fraction and rank in the given superkingdom

[b]Average fraction among the three superkingdoms.

otic gene expression, the methyltransferases are relatively much less abundant in eukaryotes than in prokaryotes (rank 12; Table 3).

The fraction of proteins with the P-loop fold strongly depends on the proteome size—the smaller the proteome, the larger the share of P-loop-containing proteins (Fig. 3). This reflects the fact that many ATPases and GTPases are involved in housekeeping processes (e.g., translation and replication), and their loss is incompatible with life. The other common folds do not show a similar distribution, and their contribution to a given proteome seems to depend more on the respective organism's lifestyle than on the total number of proteins. Thus the fraction of TIM barrels is the greatest in heterotrophic bacteria with diverse metabolism, for example, *E. coli*, whereas ferredoxins are most prominent in autotrophs with long electron transfer chains such as archaea and *Synechocystis* sp. Even more specifically, in the free-living bacterium *A. aeolicus*, whose proteome size is close to those of the parasites *B. burgdorferi* and *H. pylori*, the folds involved in metabolic functions, namely TIM barrels and Rossmann fold domains, are clearly more abundant (Fig. 3). Some observations, however, for example the obvious overrepresentation of methyltransferases in *H. pylori* (Fig. 3), are not so easily explained and may hint at completely unknown aspects of the organism's physiology.

Clustering of Organisms on the Basis of Fold Composition

Even a superficial inspection of the distributions of the top 30 folds reveals certain similarities between different organisms (Table 3). To address the issue in a systematic manner, a matrix of correlation coefficients between the fold distributions was constructed and used to produce a similarity dendrogram (Fig. 4). The dendrogram emphasizes the already mentioned dramatic difference in the fold composition between eukaryotes and prokaryotes (bacteria and archaea). Archaea form a distinct branch, whereas bacteria fall into two clusters—free-living and parasitic ones. The hyperthermophilic bacterium *A. aeolicus* is close to the common branching point on the dendrogram, which may reflect massive horizontal gene transfer from archaea, resulting in a chimeric composition of its genome (Aravind et al. 1998c).

It should be emphasized that the observed clustering is clearly different from that observed in phylogenetic reconstructions; for example, such phylogenetically close bacteria as *E. coli* and *H. influenzae* (Fleischmann et al. 1995) are in different branches of the fold composition dendrogram. It appears that the observed clustering of parasitic bacteria and their separation from the free-living ones reflects the elimination of a similar subset of folds in the course of a genome-scale adaptation to parasitism that has occurred independently in different bacterial lineages.

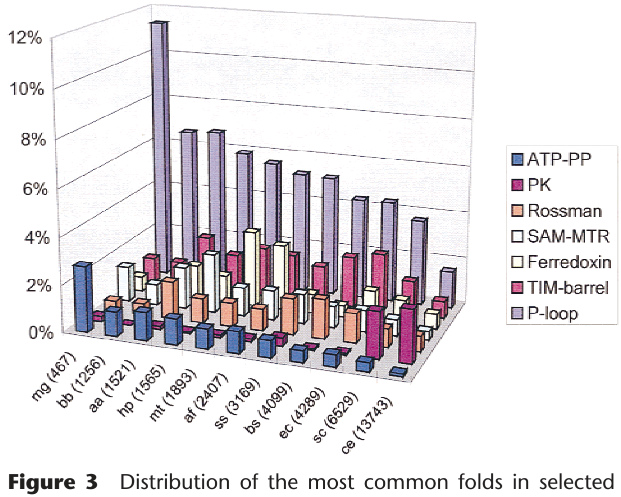

Figure 3 Distribution of the most common folds in selected bacterial, archaeal, and eukaryotic proteomes. (ATP-PP) ATP pyrophosphatases; (PK) serine–threonine protein kinases; (SAM-MTR) *S*-adenosyl methionine-dependent methyltransferases. (mg) *Mycoplasma genitalium*; (bb) *Borrelia burgdorferi*; (aa) *Aquifex aeolicus*; (hp) *Helicobacter pylori*; (mt) *Methanobacterium thermoautotrophicum*; (af) *Archaeoglobus fulgidus*; (ss) *Synechocystis* sp.; (bs) *Bacillus subtilis*; (ec) *Escherichia coli*; (sc) *Saccharomyces cerevisiae*; (ce) *Caenorhabditis elegans*. For each genome, the total number of encoded proteins is indicated in parenthesis.

Ranking and Diversity of Protein Folds in Proteomes

To explore the general features of protein-fold distribution in all organisms, the unweighted average fraction of each fold was calculated first within the superkingdoms and then between them (Table 3). This procedure gives equal weights to each proteome within a superkingdom and to each superkingdom in the total count, regardless of the sample size. A plot of the average fraction of the given fold representatives in a proteome versus fold rank (Fig. 5) shows that at least 29 of the top 30 folds fit an exponent with a strong statistical support [$P(\chi^2) >> 0.1$] (extending this plot to the rest of the 239 folds detected in 14 proteomes is statistically unfeasible since most of them are represented by only a few proteins). The first point that does not fit the curve in Figure 5 corresponds to the top-ranking P-loop ATPase fold, which is clearly overrepresented, given the exponential distribution. Computer simulations based on very simple models of pro-

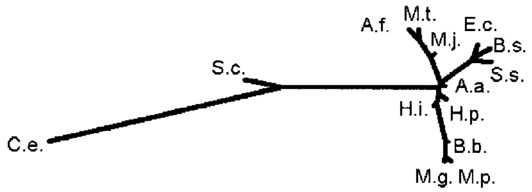

Figure 4 Clustering of proteomes by correlation between protein fold compositions.

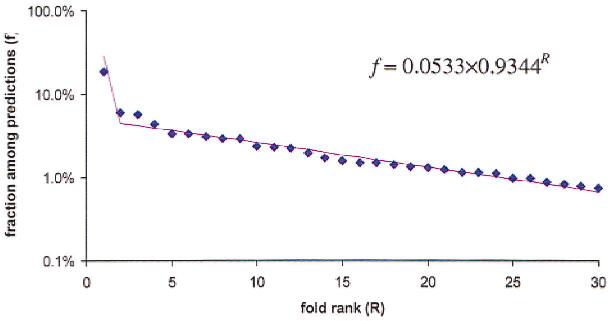

Figure 5 Rank distribution of the unweighted average fraction of the top 30 protein folds in proteomes. (Blue diamonds) The observed unweighted average fractions of folds; (magenta line) The best-fitting exponent approximation.

tein fold evolution (assuming a constant rate of protein duplication within a fold and in time, but different rates for different folds) show that the fraction versus rank plots fit exponent when the background probability of protein duplication (i.e., the growth rate of the number of fold representatives) is uniformly distributed among the folds (not shown).

The larger the proteome, the more different folds it contains (Table 2; Fig. 6). This reflects the intuitively obvious fact that proteomes of more complex organisms show a greater structural diversity. On the other hand, the increase of diversity follows from a purely stochastic model that describes a proteome as a finite sample from an infinite pool of proteins with a particular distribution of fold fractions (a bag of proteins). A series of numerical experiments was performed, simulating random sampling from a protein pool. The pool contained an infinite number of proteins, with fold fractions distributed exponentially except for one special point (the top-ranking fold); the parameters of the simulated distribution were optimized to fit the exponential part of the distribution of the top 30 folds from the 14 proteomes (Fig. 5). A comparison of the simulated and observed data (Fig. 6) shows that, whereas both real and simulated diversity seem to follow the logarithm law, the stochastic model underestimates the number of different folds approximately twofold. From the statistical viewpoint, these observations suggest that the distribution of lower-ranking folds (that can not be assessed directly because of the lack of statistically representative data) does not fit the exponential distribution observed for the higher-ranking folds (Fig. 5). In other words, the fold composition of the real proteomes does not seem to follow the protein bag model; their higher than expected diversity is likely to be a product of natural selection.

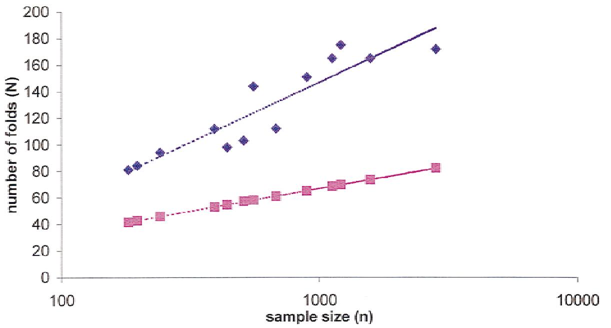

Figure 6 Observed and simulated fold diversity in complete proteomes. (Blue diamonds) Observed number of different folds in the proteomes, plotted against the number of proteins with predicted folds (sample size). (Magenta squares) The results of computer simulations under the stochastic model (see text). (Dotted lines) The respective best-fitting logarithm approximations.

Multidomain Proteins

Whereas most proteins contain only one recognizable domain, complex, multidomain proteins are not uncommon (Doolittle 1995). Aggregation of different domains within a single polypeptide chain obviously serves the purpose of bringing several different activities into spatial proximity to ensure proper coordination and regulation. One could speculate that evolution favors the formation of such multidomain proteins, or that their abundance should increase along with increasing complexity of the cellular machinery. To address these questions quantitatively, we examined the distribution of the number of domains in proteins from the three superkingdoms. The number of different folds predicted in each protein in the complete proteomes was counted, and the unweighted average fraction of the proteins with each given number of domains was calculated for each superkingdom (Fig. 7). Somewhat surprisingly, the three distributions do not significantly differ from each other [$P(\chi^2) > 0.1$ in all comparisons between superkingdoms], indicating that neither the proteome size nor the average protein length [both of which are considerably greater in eukaryotes (Das et al. 1997)] affect the statistics of do-

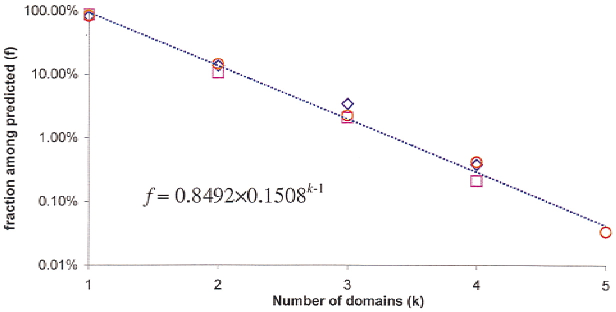

Figure 7 Distribution of multidomain proteins in complete proteomes. (Blue diamonds) Bacteria; (magenta squares) Archaea; (red circles) Eukarya. (Dotted line) Best-fitting exponent approximation (geometric distribution) for the data, averaged across the three superkingdoms.

main composition. All distributions show a very good fit [$P(\chi^2) >> 0.1$] to an exponential model (Fig. 7), where each next class contains approximately seven times less entries then the previous one. Such geometric distribution is typical of series of random independent events with the same background probability. This observation further supports the notion that the selective forces that affect the formation of multidomain proteins, if they exist, are well balanced by the forces that favor splitting of such proteins.

General Notes and Conclusions

We developed a computer system for protein fold recognition that is based on position-dependent weight matrices constructed using the iterative PSI-BLAST methods, with structurally characterized domains from the SCOP database as starting points. A collection of 1193 position-dependent weight matrices that can serve as fold identifiers was constructed and is available for use. Folds were predicted for 20%–30% of the proteins in each of the 13 analyzed complete proteomes, with a greater prediction rate (39%) for the minimal proteome of *M. genitalium*. After this analysis was completed, two independent studies have been published that give a very close number of predicted folds for *M. genitalium* using PSI-BLAST with proteins from this bacterium as starting points (Huynen et al. 1998; Rychlewski et al. 1998). The congruence between the two approaches suggests that PSI-BLAST is a reasonably robust tool for fold prediction.

Given that another 20%–30% of each proteome seem to be comprised by integral membrane proteins and soluble nonglobular proteins (e.g., Koonin et al. 1997), 30%–50% of all predictable globular domains may be covered by the present analysis. Although incomplete, this coverage suggests that conclusions drawn from the comparative analysis of fold distributions among different phylogenetic lineages may be meaningful. These distributions show major differences between eukaryotes and prokaryotes (bacteria and archaea) in terms of the predominant folds. The most common folds in prokaryotes are those involved in housekeeping functions, such as P-loop-containing NTPases and TIM barrels, whereas the eukaryotic distribution is marked by the prominence of domains with primarily regulatory functions, such as protein kinases and β-propellers. Within the bacteria, there is a remarkable correlation in the fold distributions between phylogenetically distant parasitic species as opposed to their free-living relatives.

Computer simulation of the rank distribution of folds, when compared to the actual observations, indicates that the diversity of folds in each of the analyzed proteomes is about twice as great as that predicted on the basis of the exponential distribution seen among the top 30 folds. It is speculated that structural diversity may be selected for in the course of evolution. The observed distribution of the number of multidomain proteins fits the model of their origin by random domain combination. Further improvements in domain recognition, together with the experimental identification of new folds, will show how general these trends are.

It should be kept in mind that our conclusions may be to some extent affected by the existing biases both in the database of protein structure and in the available collection of complete genome sequences. In particular, folds that are specific to archaea and to multicellular eukaryotes are likely under-represented. Nevertheless, given that the most common folds are already clearly present in SCOP and that at least two genomes from each of the three superkingdoms are available, we do not expect that the ranking of the most abundant folds changes significantly.

METHODS

Databases

The sequences of individual domains from the SCOP 1.35 database were used as the learning set for the fold recognition procedure. All database searches were performed with the NCBI NR database (288,947 protein sequences) in which the regions with low compositional complexity have been masked using the SEG program [window length 60, trigger complexity threshold 3.4, extension complexity threshold 3.8 (Wootton 1994; Wootton and Federhen 1996)]. This version of the NR database is available on request. Proteome sequence data were extracted from the NCBI database of genomes (http://ncbi.nlm.nih.gov/Entrez/Genomes/org.html).

Sequence Analysis

All searches were performed using the PSI-BLAST program, version 2.0.4 (Altschul et al. 1997). The fold recognition protocol (Fig. 1) was developed using the domain sequences representing 284 folds from SCOP 1.35. The nonprotein, oligopeptide, and coiled–coil folds were excluded for obvious reasons; folds so far found only in viral proteins were irrelevant for the analysis of proteomes of cellular life forms and were not analyzed either; the immunoglobulin fold was excluded because of the over-representation of immunoglobulin-like sequences in the NR database that made the analysis of this fold very computationally intensive. To purge redundant entries, sequences belonging to each fold were clustered by a single-linkage algorithm. The pairwise BLAST alignment score divided by the length of the shorter sequence was used as the linkage criterion) with the linkage threshold of 1.3 bit/position; the longest sequence from each cluster was selected for further analysis, and the remaining sequences were discarded. With the goal of increasing the resolution power of the procedure, the resulting sequence set was used to search the NR database using the gapped BLASTP program. Database entries with highly significant ($e \leqslant 10^{-4}$) similarity to the query sequences were considered to be indisputable homologs with the same structural fold. The portions of the respective proteins that aligned with the query were extracted from the database and clustered again using the linkage threshold of 0.5 bit/position. The longest sequence was again retained and used as a query to initiate a PSI-BLAST search of the NR database that was run to convergence or to a maximum of 10 iterations, whichever comes first; the cutoff for inclusion of a sequence in matrix construction was set at $e \leqslant 10^{-2}$. In addition to the search results themselves, a position-specific weight matrix was saved for each search and stored for subsequent use.

For the construction of fold composition dendrogram, the matrix of correlation coefficients (r) between the species-specific fold composition vectors was converted into a distance matrix using the $1 - r^2$ transformation. The dendrogram was constructed from this distance matrix using the FITCH program from the PHYLIP package (Felsenstein 1996), which is based on the least-square algorithm of Fitch and Margoliash (Fitch and Margoliash 1967).

ACKNOWLEDGMENTS

We thank L. Aravind and Roland Walker for valuable help with case-by-case fold prediction and automated database searches, respectively, and L. Aravind, Steven Bryant, Michael Galperin, David Landsman, David Lipman, Kira Makarova, and Roland Walker for critical reading of the manuscript.

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

Abagyan, R.A. 1997. Protein structure prediction by global energy optimization. In *Computer simulations of biomolecular systems: Theoretical and experimental applications* (ed. W.F. Van Gunsteren), Vol. 3, pp. 363–394. ESCOM Science Publishers BV, Leiden, The Netherlands.

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

Anfinsen, C.B. and H.A. Scheraga. 1975. Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.* **29:** 205–300.

Aravind, L. and E.V. Koonin. 1998. Phosphoesterase domains associated with DNA polymerases of diverse origins. *Nucleic Acids Res.* **26:** 3746–3752.

Aravind, L., M.Y. Galperin, and E.V. Koonin. 1998a. The catalytic domain of the P-type ATPase has the haloacid dehalogenase fold. *Trends Biochem. Sci.* **23:** 127–129.

Aravind, L., D.D. Leipe, and E.V. Koonin. 1998b. Toprim-a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. *Nucleic Acids Res.* **26:** 4205–4213.

Aravind, L., R.L. Tatusov, Y.I. Wolf, D.R. Walker, and E.V. Koonin. 1998c. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.* **14:** 442–444.

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

Brenner, S.E., C. Chothia, and T.J. Hubbard. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci.* **95:** 6073–6078.

Bryant, S.H. and S.F. Altschul. 1995. Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* **5:** 236–244.

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

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

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

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

Doolittle, R.F. 1981. Similar amino acid sequences: chance or common ancestry? *Science* **214:** 149–159.

———. 1995. The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64:** 287–314.

Dorit, R.L., L. Schoenbach, and W. Gilbert. 1990. How big is the universe of exons? *Science* **250:** 1377–1382.

Felsenstein, J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* **266:** 418–427.

Fischer, D. and D. Eisenberg. 1997. Assigning folds to the proteins encoded by the genome of Mycoplasma genitalium. *Proc. Natl. Acad. Sci.* **94:** 11929–11934.

Fitch, W.M. and E. Margoliash. 1967. Construction of phylogenetic trees. *Science* **155:** 279–284.

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

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

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

Gerstein, M. 1997. A structural census of genomes: Comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274:** 562–576.

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

Godzik, A. and J. Skolnick. 1992. Sequence-structure matching in globular proteins: Application to supersecondary and tertiary structure determination. *Proc. Natl. Acad. Sci.* **89:** 12098–12102.

Goffeau, A., B.G. Barrell, H. Bussey, R.W. Davis, B. Dujon, H. Feldmann, F. Galibert, J.D. Hoheisel, C. Jacq, M. Johnston et al. 1996. Life with 6000 genes. *Science* **274:** 546, 563–567.

Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B.C. Li, and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. *Nucleic Acids Res.* **24:** 4420–4449.

Holm, L. and C. Sander. 1996. Mapping the protein universe. *Science* **273:** 595–603.

———. 1997. New structure—Novel fold? *Structure* **5:** 165–171.

Hubbard, T.J.P., A.G. Murzin, S.E. Brenner, and C. Chothia. 1997. SCOP: A structural classification of proteins database. *Nucleic Acids Res.* **25:** 236–239.

Huynen, M., T. Doerks, F. Eisenhaber, C. Orengo, S. Sunyaev, Y. Yuan, and P. Bork. 1998. Homology-based fold predictions for mycoplasma genitalium proteins. *J. Mol. Biol.* **280:** 323–326.

Kaneko, T. and S. Tabata. 1997. Complete genome structure of the unicellular cyanobacterium Synechocystis sp. PCC6803. *Plant Cell Physiol.* **38:** 1171–1176.

Karlin, S. and S.F. Altschul. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci.* **87:** 2264–2268.

Karlin, S., P. Bucher, V. Brendel, and S.F. Altschul. 1991. Statistical methods and insights for protein and DNA sequences. *Annu. Rev. Biophys. Biophys. Chem.* **20:** 175–203.

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

Koonin, E.V., A.R. Mushegian, M.Y. Galperin, and D.R. Walker. 1997. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.* **25:** 619–637.

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

Kuwabara, P.E. 1997. Worming your way through the genome. *Trends Genet.* **13:** 455–460.

Lathrop, R.H. 1994. The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.* **7:** 1059–1068.

Murzin, A.G. and A. Bateman. 1997. Distant homology recognition using structural classification of proteins. *Proteins* (Suppl.) **1:** 105–112.

Murzin, A.G., S.E. Brenner, T. Hubbard, and C. Chothia. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536–540.

Mushegian, A.R., D.E. Bassett, Jr., M.S. Boguski, P. Bork, and E.V. Koonin. 1997. Positionally cloned human disease genes: Patterns of evolutionary conservation and functional motifs (see comments). *Proc. Natl. Acad. Sci.* **94:** 5831–5836.

Ortiz, A.R., A. Kolinski, and J. Skolnick. 1998. Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. *Proc. Natl. Acad. Sci.* **95:** 1020–1025.

Rychlewski, L., B. Zhang, and A. Godzik. 1998. Fold and function preditions for Mycoplasma genitalium proteins. *Folding Design* **3:** 229–238.

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

Smith, T.F., L. Lo Conte, J. Bienkowska, C. Gaitatzes, R.G. Rogers, Jr., and R. Lathrop. 1997. Current limitations to protein threading approaches. *J. Comput. Biol.* **4:** 217–225.

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

Wootton, J.C. 1994. Non-globular domains in protein sequences: Automated segmentation using complexity measures. *Comput. Chem.* **18:** 269–285.

Wootton, J.C. and S. Federhen. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266:** 554–571.

Received August 19, 1998; accepted in revised form November 24, 1998.