J. Mol. Biol. (1997) **274**, 562–576

JMB

A Structural Census of Genomes: Comparing Bacterial, Eukaryotic, and Archaeal Genomes in Terms of Protein Structure

Mark Gerstein

Department of Molecular Biophysics & Biochemistry
266 Whitney Avenue
Yale University
PO Box 208114, New Haven
CT 06520, USA

Representative genomes from each of the three kingdoms of life are compared in terms of protein structure, in particular, those of *Haemophilus influenzae* (a bacteria), *Methanococcus jannaschii* (an archaeon), and yeast (a eukaryote). The comparison is in the form of a census (or comprehensive accounting) of the relative occurrence of secondary and tertiary structures in the genomes, which particular emphasis on patterns of supersecondary structure. Comparison of secondary structure shows that the three genomes have nearly the same overall secondary-structure content, although they differ markedly in amino acid composition. Comparison of super-secondary structure, using a novel ''frequent-words'' approach, shows that yeast has a preponderance of consecutive strands (e.g. beta–beta–beta patterns), *Haemophilus*, consecutive helices (alpha–alpha–alpha), and *Methanococcus*, alternating helix-strand structures (beta–alpha–beta). Yeast also has significantly more helical membrane proteins than the other two genomes, with most of the differences concentrated in proteins containing two transmembrane segments. Comparison of tertiary structure (by sequence matching and domain-level clustering) highlights the substantial duplication in each genome (≈30% to 50%), with the degree of duplication following similar patterns in all three. Many sequence families are shared among the genomes, with the degree of overlap between any two genomes being roughly similar. In total, the three genomes contain 148 of the ≈300 known protein folds. Forty-five of these 148 that are present in all three genomes are especially enriched in mixed super-secondary structures (alpha/beta). Moreover, the five most common of these 45 (the ''top-5'') have a remarkably similar super-secondary structure architecture, containing a central sheet of parallel strands with helices packed onto at least one face and beta–alpha–beta connections between adjacent strands. These most basic molecular parts, which, presumably, were present in the last common ancestor to the three kingdoms, include the TIM-barrel, Rossmann, flavodoxin, thiamin-binding, and P-loop-hydrolase folds.

Introduction

In the past two years the complete genome sequences of a number of free-living organisms have been announced, generating tremendous interest (Nowak, 1995; Wade, 1997). This provides a unique opportunity to perform comprehensive comparisons between different organisms on a molecular level. Here three genomes are compared in terms of the protein structures that they encode, those of *Haemophilus influenzae* (Fleischmann *et al*., 1995a,b), *Methanococcus jannaschii* (Bult *et al*., 1996), and the yeast *Saccharomyces cerevisiae* (Goffeau *et al*., 1996, 1997). These organisms are representatives of the three major kingdoms of life (bacteria, archaea and eukarya) and so provide a most diverse comparison.

Structurally, each of these organisms appears different on the micron scale as they have different

Abbreviations used: HI, *Haemophilus influenzae*; MJ, *Methanococcus jannaschii*; SC, *Saccharomyces cerevisiae* or yeast; Å, Angström; ORF, open reading frame; PDB, Protein Data Bank.

0022–2836/97/490562–15 $25.00/0/mb971412

internal cell structures, but on the scale of single Ångströms they appear nearly the same, containing similar proportions of C, H, O, N, P, and S atoms. The question addressed here is how they compare on the scale of protein structure (10 to 100 Å). For instance, since these organisms live in such different physical environments, from deep-sea vents at high temperature and pressure for *Methanococcus* (85°, 200 atm) to rotting figs and grapes at normal temperature and pressure (for yeast) it is possible that particular types of secondary or tertiary structure would be favored over others.

To address this and other related questions, the three genomes (or more properly proteomes) were analyzed by established techniques of sequence matching and secondary-structure prediction. They are found to differ in terms of overall amino acid composition, yet, surprisingly, have a similar composition in terms of secondary structures. The results of the sequence comparison and structure prediction were clustered and combined using new methods to address questions related the distribution of supersecondary structures. It is found that the genomes have different frequencies of supersecondary structures (e.g. αβαβ), with yeast having relatively more consecutive strands (e.g. βββ), *Haemophilus* having more consecutive helices (e.g. αα), and *Methanococcus* having more alternating helix-strand structures (e.g. αβαβ). Yeast also has more helical membrane proteins, especially those with only one or two transmembrane elements.

The genomes also have a different composition of tertiary structures (or folds). The number of distinct folds contained in the three genomes is considerably less than the aggegrate number of sequences (≈10,000) because of substantial duplication within each genome (involving between a third and half the sequences in each genome) and the many protein families shared between genomes. A small group of sequence families are common to all three genomes and presumably were present in the last common ancestor to bacteria, eukaryotes and archaea, hypothesized to exist two billion years ago (Doolittle *et al*., 1996). In total the three genomes contain 148 known protein folds, which mostly have an α/β architecture. A disproportionate number of the shared sequence families have a known structure (45), and five of these 45 are among the most common folds in each genome (i.e. in the ''top-10''). These five, which appear to be basic ''molecular parts,'' are the TIM-barrel fold, Rossmann fold, flavodoxin fold, thiamin-binding fold, and P-loop-hydrolase fold. They have a remarkably similar architecture, containing a central sheet of parallel strands with helices packed on at least one face and with, almost exclusively, βαβ connections between adjacent strands.

There has been much recent work analyzing genomes (or partial genomes) that this work is following upon. Automated methods have been developed for comparing a whole genome against a number of interlocking databases, and these have been used to characterize a number of recently sequenced genomes (Bork *et al*., 1992a,b; Scharf *et al*., 1994; Casari *et al*., 1995; Ouzounis *et al*., 1995a,b). Genes have been related to the metabolic pathways, enabling determination of whether or not a pathway is present in a given organism and making possible estimation of the minimal set of genes necessary for life (Karp *et al*., 1996a,b; Koonin *et al*., 1996a; Mushegian & Koonin, 1996; Tatusov *et al*., 1996). The number of membrane proteins in genomes has been surveyed (Arkin *et al*., 1997; Goffeau *et al*., 1993; Rost *et al*., 1995, 1996), and duplications (i.e. clusters of paralogous genes) in genomes have been identified (Brenner *et al*., 1995; Koonin *et al*., 1996b; Riley & Labedan, 1987; Wolfe & Shields, 1997). Genomes have been compared based on the frequencies of oligonucleotide and oligopeptide words, focusing especially on the relative abundance of dinucleotides as a unique genomic signature (Blaisdell *et al*., 1996; Karlin & Burge, 1995; Karlin *et al*., 1992, 1996). Finally, it has been possible to identify certain sequences, called ancient conserved regions, that have been conserved over long evolutionary time scales between phylogenetically distant organisms (Green *et al*., 1993; Koonin *et al*., 1995). Recently, comparisons have just been made focussing on archea, highlighting the sequence families unique to this kingdom (Ouzounis *et al*., 1995a,b; Clayton *et al*., 1997).

As the work here involves comparing the protein structures implied by the genome sequences it also rests upon the great amount of recent work systematizing protein structures and classifying them into fold families (Gibrat *et al*., 1996; Holm & Sander, 1996; Murzin *et al*., 1995; Orengo *et al*., 1994; Schmidt *et al*., 1996; Pascarella & Argos, 1992; Sander & Schneider, 1991). In particular, a census similar to this one has recently been done, comparing the occurrence of fold families in different species (Gerstein & Levitt, 1997a).

One expects analysis of structure to reveal more about distant evolutionary relationships than just sequence comparison since structure is usually more conserved than sequence (Chothia & Gerstein, 1997; Chothia & Lesk, 1986). In other words, it is at the level of protein structure, where one sees the greatest redundancy and reuse in biology. Specifically, it is believed that there is only a very limited number of protein motifs, and elucidation of this limited repertoire of molecular parts is seen as one of the principal future challenges for biology (Chothia, 1992; Lander, 1996).

Results and Discussion

Overall genome size and composition

As shown in Table 1, the *Haemophilus* (HI) and *Methanococcus* (MJ) genomes are approximately the same size, both in terms of the number and average length of the sequences (≈1700 and ≈295, respectively). However, the yeast (SC) genome is considerably larger, coding for six times as many

Table 1. Overall statistics for secondary structures

	HI	MJ	SC
All sequences in genome			
Total number	1680	1735	6218
Average length (residues)	301	287	466
Average strand propensity per residue (kcal/mole)	−0.33	−0.37	−0.35
Average helix propensity per residue (kcal/mole)	−1.02	−1.03	−0.98
Average TM-helix propensity per residue (kcal/mole)	1.35	1.74	1.64
Sequences corresponding to soluble proteins			
Total number	1376	1502	4810
(as fraction of number of sequences in genome)	82%	87%	77%
Average length (residues)	285	280	432
Fraction of residues that are predicted to be in a sheet	16%	18%	16%
Fraction of residues that are predicted to be in a helix	42%	41%	35%
Predicted number of secondary structure elements	20	21	29
Fraction of elements that are strands	51%	53%	54%
Fraction of elements that are helices	49%	47%	46%
Average number of residues per strand	4.4	4.5	4.4
Average number of residues per helix	11.8	11.9	11.2

The number of soluble protein is the total number of sequences less those predicted to contain at least two transmembrane elements (see Methods). For reference, the size of a domain in a protein of known structure is 174.1. This was determined by averaging the lengths of the 971 non-homologous chains in scop (see Methods). The length distribution for the structure domains and for the ORFs in the three genomes are unimodal. There is no periodicity observed (e.g. for multiples of 125, as suggested by Berman *et al*. (1994)). The average strand, α-helix, and TM-helix propensities are derived by computing a weighted average of the propensities in Table 2, using as weighting factors for each residue the fractional composition of it in the whole genome.

residues, spread over more than 6200 sequences with an average size of ≈470. For comparison the average size of a protein domain in known crystal structures is ≈175 residues (calculation described in the Table legend), about a third the size of a protein in yeast or 60% of one in *Haemophilus* or *Methanococcus*.

As shown in Table 2, the three genomes have some significant differences in terms of their overall amino acid composition. The average difference for any amino acid is 45%. The greatest particular differences are between *Methanococcus* in comparison to *Haemophilus* and yeast, perhaps reflecting the radically different environment that *Methanococcus* lives in. For instance, Ile and Lys are more common (by about half) in the *Methanococcus* genome than in those of yeast and *Haemophilus*, and Ser and Gln are much less common (Gln by a factor of almost five). These differences are somewhat larger than the compositional differences usually found in comparing sets of sequences from different species (Doolittle, 1987).

Each amino acid has a different propensity to confer secondary structure, whether α-helices, transmembrane helices, or β-strands (also shown in Table 2). Consequently, the differences in composition might be expected to give rise to more of one type of secondary structure, e.g. more helices. This can be tested to some degree through prediction of secondary structure.

Overall secondary-structure and transmembrane-helix composition

Bulk prediction of secondary structure was done for every protein in the three genomes by a number of standard approaches (see Methods). First, the proteins that contain transmembrane helices were determined. There appears to be relatively more helical membrane proteins in yeast than in *Haemophilus* and more in *Haemophilus* than *Methanococcus* (as a fraction of the total genome, 23% *versus* 18% to 13%). The large fraction of membrane proteins in the yeast genome may reflect the greater number of membranes in a eukaryotic cell. (However, it may also result from the larger average size of a yeast protein if there is an assumed constant propensity for transmembrane segment formation per residue.)

Second, the membrane proteins were set aside and conventional helix–turn–strand secondary structure was predicted for the remaining proteins. Surprisingly, despite the differences in amino acid composition, the overall statistics for secondary structure composition (the number and size of helices and strands) are nearly identical in the three genomes: about 54% of secondary structural elements are predicted to be strands (with an average length of 4.5 residues) and the remaining fraction are helices (with an average length of 11.5 residues). (By residue, about 40% of the genomes are predicted to be helical, and about 17%, strand.)

How can the genomes have such similar secondary structure composition, while having such a markedly different amino acid composition? This is analogous to how genomes can have very different base compositions (AT or GC rich) while coding for proteins with similar amino acid composition. To some degree it has do with a ''degeneracy'' in the coding of secondary structure propensities and the ''trading-off'' of residues with

Table 2. Differences in amino acid composition

	Amino acid composition (%)			Max diff. between genomes (%)	Propensity (kcal/mol)		
	HI	MJ	SC		TM helix	α helix	β strand
Q	4.6	1.5	3.9	105	+4.1	−1.3	−0.4
S	5.6	4.5	9.0	67	−0.6	−1.1	−0.9
K	6.3	10.4	7.3	49	+8.8	−1.5	−0.4
I	7.1	10.5	6.6	46	−3.1	−1.2	−1.3
W	1.1	0.7	1.0	43	−1.9	−1.1	−1.0
H	2.1	1.4	2.2	40	+3.0	−1.1	−0.4
A	8.2	5.5	5.5	40	−1.6	−1.9	0.0
T	5.2	4.0	5.9	37	−1.2	−0.6	−1.4
Y	3.1	4.4	3.4	33	+0.7	−1.2	−1.6
E	6.5	8.7	6.5	29	+8.2	−1.2	−0.2
G	6.6	6.3	5.0	29	−1.0	0.0	+1.2
P	3.7	3.4	4.3	25	+0.2	+3.0	>3.0
C	1.0	1.3	1.3	24	−2.0	−1.1	−0.8
N	4.9	5.3	6.1	22	+4.8	−1.0	−0.5
V	6.7	6.9	5.6	20	−2.0	−0.8	−0.9
R	4.5	3.9	4.5	16	+12.3	−1.9	−0.4
D	5.0	5.5	5.8	15	+9.2	−1.0	+0.9
M	2.4	2.2	2.1	14	−3.4	−1.4	−0.9
L	10.5	9.5	9.6	10	−2.8	−1.6	−0.5
F	4.5	4.2	4.5	7	−3.7	−1.0	−1.1

The Table shows the amino acid composition of the three genomes. The average rms difference in composition for an individual amino acid is 0.023, and expressed as a fraction of 5%, this is 45%. (0.023 is computed from determining the rms difference between each of pair of composition vectors, *viz*:

$$\sqrt{\frac{(\mathbf{V}_{HI}-\mathbf{V}_{MJ})^2+(\mathbf{V}_{MJ}-\mathbf{V}_{SC})^2+(\mathbf{V}_{SC}-\mathbf{V}_{HI})^2}{60}}$$

where $\mathbf{V}_{HI}$ is the composition vector for HI and other vectors are named correspondingly.) The fifth column shows the differences in composition in more detail. For each amino acid the maximum difference is expressed as a percentage, viz: $2(X-y)/(X+y)$, where is X is the maximum composition amount of a particular amino acid in the three genomes and y is the minimum. The TM-helix scale gives the energy in kcal/mol for inserting this amino acid into a membrane (Engelman *et al.*, 1986). As described in the Methods it is used here for the identification of membrane proteins. The α-helix and β-stand propensity scales illustrate how different compositions of amino acids would be expected, to a first approximation, to give rise to different secondary structures. They are also expressed in kcal/mol. Both scales are derived from protein-unfolding experiments (Chakrabartty *et al.*, 1994; Smith *et al.*, 1994), but similar scales can be determined from doing statistics on solved crystal structures (King & Sternberg, 1996).

equivalent propensities between genomes. This is evident in the similar values calculated for each genome for average helix and strand propensity per residue (Table 1).

Frequent super-secondary structure words

Groups of linked secondary structures (e.g. αβαβ) form super secondary structures, and the frequency of these in each genome can also be analyzed. For this analysis the ''frequent words'' approach often used for characterizing oligopeptide and oligonucleotide sequences is applied to the predicted secondary structures. The results, shown in Table 3, are for super secondary structure ''words'' of length 2 to 7. Unlike secondary structures, there are great differences in the frequency that super secondary structures occur in the three genomes. In general the greatest differences occur for symmetrical or repeating patterns of super secondary structure (i.e. symmetrical patterns, such as ααααα and αβαβ, which are underlined in the Table, *versus* asymmetrical ones, such as αβββ or αββαα). These differences become more pronounced as one looks at longer and longer words (i.e. βββ *versus* βββββββ).

There is a much greater chance of finding all-β words (e.g. ββ, βββ, ββββ, and so forth) in yeast than in the other genomes. Specifically, the chance of finding ββββ in yeast is 54% greater than finding it in *Haemophilus* and a 27% greater than finding it in *Methanococcus*, for 41% greater on average. This is interesting in that all-β proteins are expected to be common in metazoa, which contain a great number of all-β immunoglobulin-like and fibronectin type III-like folds, but not necessarily in lower eukaryotes such as yeast (Doolittle, 1995; Gerstein & Levitt, 1997a).

Conversely, all-α proteins are more common in *Haemophilus* than in the other two organisms. In particular, four consecutive helices (as in a four-helix bundle) are more frequent in *Haemophilus* than *Methanococcus* or yeast by 25%. As is necessarily implied, alternating alpha–beta structures are more common in *Methanococcus* than the other two

Table 3. Frequencies of supersecondary structure words

A.

Super-secondary structure "word"	Maximum difference between 3 genomes (%)	Relative abundance (odds ratio) HI	MJ	SC	PDB
ββ	26	0.96	1.06	1.24	1.22
αα	15	0.97	0.85	0.83	0.85
αβ	10	1.09	1.09	0.99	0.95
βα	7	0.98	1.00	0.93	0.99
βββ	41	0.96	1.15	1.46	1.62
ααα	19	1.01	0.83	0.84	0.92
αβα	18	1.04	1.03	0.87	1.16
ααβ	15	1.03	0.97	0.89	0.70
βαβ	12	1.15	1.24	1.10	1.19
βαα	11	0.93	0.87	0.83	0.78
ββα	9	0.90	0.94	0.99	0.82
αββ	6	0.97	0.98	1.03	0.80
ββββ	54	1.03	1.35	1.78	2.28
αααα	29	1.10	0.82	0.89	1.18
βββα	25	0.85	0.94	1.10	0.98
βαβα	23	1.11	1.18	0.94	1.48
αβαβ	21	1.21	1.23	0.99	1.39
αβαα	21	1.00	0.95	0.81	1.00
αβββ	20	0.93	0.95	1.14	0.93
ααβα	20	0.97	0.88	0.80	0.91
αααβ	19	1.03	0.94	0.85	0.50
βααα	14	0.92	0.84	0.79	0.63
βααβ	13	1.06	1.01	0.93	0.92
βαββ	12	1.04	1.13	1.18	0.98
ββαβ	12	1.07	1.21	1.20	0.97
ββαα	8	0.84	0.78	0.85	0.59
αββα	8	0.95	0.94	0.87	0.68
ααββ	6	0.90	0.84	0.89	0.58
βββββ	67	1.14	1.71	2.27	3.43
ααααα	41	1.27	0.84	0.97	1.63
ββββα	38	0.85	0.98	1.25	1.15
αβαβα	32	1.21	1.24	0.90	2.06
αββββ	30	0.96	0.99	1.30	1.05
ββαββ	29	0.99	1.13	1.32	1.04
βαβαβ	28	1.33	1.48	1.11	1.93

25 more 5-letter words follow...

B.

BBBBB (67%),	aaaaa (41%),	BBBBa (38%),	aBaBa (32%),	aBBBB (30%),
BBaBB (29%),	BaBaB (28%),	aaaBa (27%),	BBBaB (27%),	aBaaa (26%),
BaBaa (21%),	aaBBB (24%),	aaBaB (23%),	aBaaB (23%),	BBBaa (22%),
aaBaa (21%),	BaBaa (21%),	aaaaB (20%),	BaaaB (18%),	aBBaa (17%),
Baaaa (14%),	aaBBa (14%),	aBBaB (13%),	BaaBa (13%),	BBaBa (12%),
BBaaB (11%),	aaaBB (09%),	aBBBa (09%),	aBaBB (07%),	BaBBa (04%),
BBaaa (04%),	BaaBB (03%)			

BBBBBBB	aaaaaaa	BBBBBBa	BBBaBBB	BaBBBBB	BBaBBBB	BBBBaBB	BBBBBaa
BaBaBaa	BBBBBaB	aBaBaBa	BBaaBBB	BaaaBaa	aBBBBBB	aaaaBBB	aaBaBaB
aBaBaaa	aBaaaaa	aBaaBBa	BaBaBaB	aBaBaaB	aaBBBBB	aBBBaBB	BaaBaBa
BBBBaaa	aBBBBBa	BaaBBBB	aaaBaaa	aBaBBaa	aaaaBaB	aaBBaBa	BBaBaBa
Baaaaaa	aaaaaaB	aaaBaaB	BBBaaBa	BaBaaaB	BaaBBaB	aaaBBBB	aBaaaaB
BaBBaBa	aaaaaBa	BBBBaaB	aBBaaBa	BaBBaaa	BBBaaBB	BBBaBBa	BaBaaaa
aaBBBaa	aaBBBBa	aBaBaBB	aBBaBaa	BBaaaBa	aBBaBaB	aaBaBaa	BBBaaaa
aaaBBaB	aaaaaBB	aaBaaaa	BBBaBaB	aBaaaBa	BaBBaBB	BBaBBBa	aBBaaaa
aBaaaBB	aaaBaBa	aaaaBaa	aBaaBaB	BaaaaaB	BaBaaBB	BaBaBBB	aaBaaaB
BBaBBaB	BBBBaBa	aaBBaaB	aBBaBBB	aBaaBaa	aaBaaBa	BaaaaBa	
aaBaBBa	BBaaBaB	aaBaBBB	aaaBBaa	BaaBaaB	aBaBBBa	BBaaaaa	aBBaaaB
BBaBaBB	aBBBaaB	BBaaaBB	BaBaBBa	aaaBaBB	aBaBBBB	aBaBBaB	BaaaBaB
BaBBBBa	BaBBaaB	BaBaaBa	aBBBBaB	aaaBBBa	aBBBBaa	aaBaaBB	aBBaBBa
BBaaBaa	BaaaBBa	BBaBaaa	aBBaaBB	BaBBBaa	BaaBaBB	BaaaBBB	BaaBBaa
BBBaBaa	aBaaBBB	aBBBaaa	aBBBaBa	BBaBaaB	BaaaaBB	BBaBBaa	BBaaaaB
aaBBBaB	BaaBBBa	BBBaaaB	aaBBaBB	BBaaBBa	aaaaBBa	aaBBaaa	BaaBaaa

A. The different frequencies of super secondary patterns in the predicted secondary structure for the three genomes. The frequencies are shown as odds ratios of the observed number to the expected number (so values greater than 1.0 denote frequent patterns). The second column of the Table provides a measure of the differences between the genomes *viz*: $D = 2(X - y)/(X + y)$, where is X is the maximum odds ratio for a particular word and y is the minimum. Symmetrical and repeating words are underlined. For five-letter words, 32 in total, only the seven that exhibit the greatest differences between genomes are shown. For comparison the frequencies of super-secondary words observed in a representative set of known protein structures is shown in the last column, labeled "PDB" (see Methods for discussion of the representative set). The structure-databank appears to have an even higher representation of some super-secondary structure patterns (particularly all-β) than any of the genomes, perhaps reflecting its biased composition toward such folds as the immunoglobulins. B. All the 5 and 7 character words ordered according to the difference D between genomes. Symmetrical and repeating words are underlined. Note how symmetrical words often appear to exhibit the greatest difference between genomes.

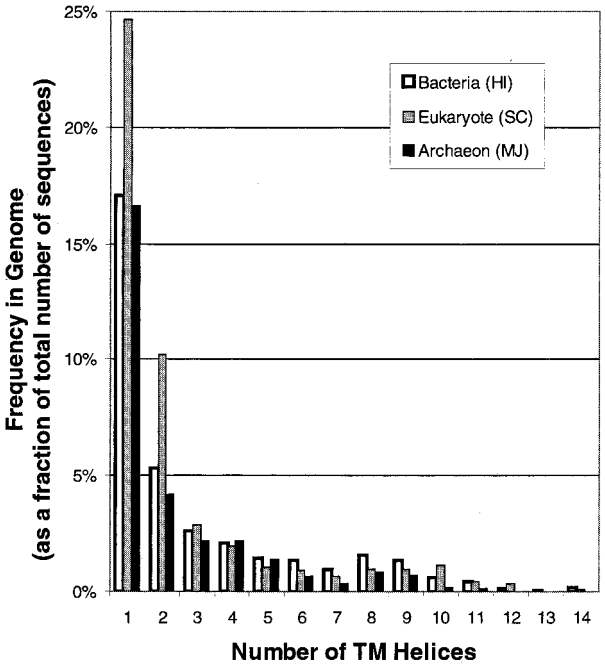

Figure 1. Transmembrane segments in the three genomes. The frequency of membrane proteins with a given number of transmembrane (TM) helices is shown for the three genomes.

genomes. However, this difference is not as great as with all-α and all-β words. Specifically, αβαβ and βαβα have a 13.5% greater chance of occurring in *Methanococcus* than in yeast or *Haemophilus*.

The preponderance of all-β super-secondary structures in yeast, all-α ones in *Haemophilus*, and alternating αβ structures in *Methanococcus* is obviously related to the a greater frequency of all-β domains in the yeast genome, all-α domains in the *Haemophilus* genome, and mixed (α/β and $\alpha+\beta$) domains in the *Methanococcus* genome (in the sense of the original classification of Levitt & Chothia (1976)). However, because of the difficulty in determining domain boundaries in many of these sequences it is difficult to make these conclusions more precise.

It is possible to look at the super secondary structure of membrane proteins in a simple fashion by classifying them in terms of the number of transmembrane helices they contain. Such an analysis is shown in Figure 1. Overall, the size of the family with a given number of transmembrane helices drops off sharply with increasing numbers of helices, as has been observed in previous studies (Arkin *et al.*, 1997; Goffeau *et al.*, 1993; Rost *et al.*, 1995, 1996). This behavior is observed fairly consistently for the three genomes. However, yeast has relatively more membrane proteins with only one or two transmembrane helices, and this difference principally accounts for the higher proportion of membrane proteins in yeast.

Duplications: families of paralogous genes

The number of sequence families within each of the genomes was determined by doing all-*versus*-all sequence comparisons and clustering the results. In performing these calculations, one has to take into account that sequence similarity occurs on the domain level, so two completely dissimilar sequences can both match a third, intermediate sequence at different places. This situation is illustrated in Figure 2, and as described in Methods, a multiple linkage approach was used here.

As has been determined previously (Brenner *et al.*, 1995; Bult *et al.*, 1996; Tamames *et al.*, 1997), there is a substantial amount of duplication in each genome, giving rise to families of paralogous sequences. The number of individual sequences involved in duplications is about 30% for *Haemophilus*, 37% for *Methanococcus*, and 46% for yeast, with the larger genomes having a greater amount of duplication. Overall, *Haemophilus* and *Methanococcus* have ≈1350 sequence families, and yeast ≈4300 (exact numbers in legend to Figure 3). The number of sequences in each duplication ranges from 2 to more than 30 (for two families in yeast). As shown in Figure 3, the number of sequence families N of a given size S, expressed as a fraction of the total number of sequences in the genome G, falls off in a similar fashion for all three genomes, approximately according to the following formula:

$$\frac{N}{G} = \frac{1}{2S^3}.$$

Sequence similarity between genomes

In addition to the similarity between sequences within the same genome, there is similarity between sequences in different genomes (resulting, in part, in families of orthologous genes). The degree of overlap can be represented in terms of Venn diagrams, similar to those by Ouzounis & Kyrpides (1996), as shown in Figure 4. Clearly, an appreciable fraction of the sequences in any one genome are similar to those in other two, with smaller genomes having a proportionately greater fraction (39% for *Haemophilus*, 33% for *Methanococcus*, and 19% for yeast). It is somewhat difficult to make absolute comparisons regarding similarity since the much larger size of the yeast genome distorts the results. However, from the perspective of each individual genome roughly an equal proportion of its sequences are homologous to sequences in the other two genomes, a reasonable finding if the genomes have diverged equally from a common ancestor.

Sequence similarity to known structures

An appreciable fraction of the sequences in each of the genomes were homologous (or identical) to sequences corresponding to known structures (in the PDB). Looking just at these and using the fold definitions in the scop database, it is possible to see

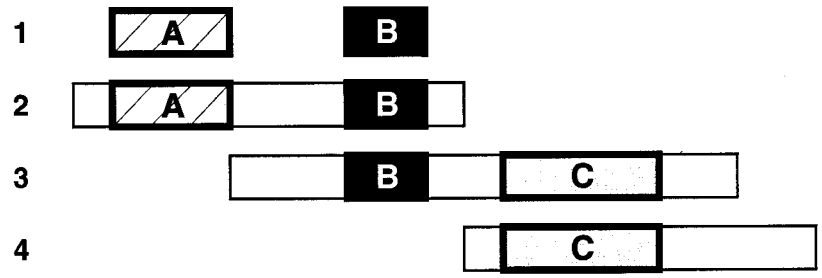

Figure 2. "Domain problem" in clustering sequence families. The Figure illustrates some of the complexities in clustering sequences. Domains are fundamentally defined at the structural level. Here the definitions in the scop database are used (Murzin *et al.*, 1995). It can be the case that a given ORF matches more than one domain-level fold (e.g. yeast ORF YAL026C). This is illustrated by the top two lines in the Figure, where sequence 2 matches both the A and B domain folds. Statistics based on this type of matching (where a single ORF can count more than once) and are used in the construction of Figure 5 and Table 4. However, for sequences that do not correspond to known structures, it is not possible to rigorously or consistently define domains, although some approaches exist (Sonnhammer & Kahn, 1994; Sonnhammer *et al.*, 1996). Consequently, the sequence-level clustering in Figure 3 is done only in terms of individual genes. If this approach is taken, single-linkage clustering can give potentially misleading results, as has been pointed before (Koonin *et al.*, 1996b; Riley & Labedan, 1997), it will group together two sequences (i.e. 2 and 4) that have similarity to different domains (B and C) in a third, intermediate sequence (3). One can get around this problem in two ways: one can split sequence 3 in half during the clustering or one can use a multiple-linkage algorithm and only create a cluster where all the members have similarity to each other. The first case is in a sense more accurate. However, it greatly biases the resulting statistics on the fraction of duplication since one knows how many genes there are in the genome but not how many domains there are. Here the later approach is taken. In a sense this is what happens in a governmental census. Even if one has multiple homes or jobs, the annual census forces one into a single category to keep the statistics fair.

how the known folds are distributed amongst the three genomes. This is shown in Figure 5 and Table 4.

The known domain structures presently correspond to 971 sequence families (at the 40% homology level, see Methods: S. Brenner, C. Chothia & T. Hubbard, unpublished results), and a total of 355 of these are present in at least one of the genomes. Obviously more are represented in yeast than *Haemophilus* and more in *Haemophilus* than *Methanococcus*, reflecting the greater size of the yeast genome as well as the biases of investigators. Comparison of Figures 4 and 5 shows that sequence families common to one or more genome have a greater chance of having a known structure. This finding could again reflect the biases of investigators but it could also indicate the omnipresent character of the ancient shared families.

Using the structural similarity relationships in the scop database, sequence families that share the same fold but which have no detectable homology can be combined into fold families. These are currently 299 folds in scop, and of these about half (148) are contained in at least one of the three genomes. The fact that these numbers are considerably less than the number of sequence families shows how many of the evolutionary similarities between these highly diverged organisms are only apparent in terms of structure, all the sequence similarity having been eroded away (Doolittle, 1995). Although the *Haemophilus* and *Methanococcus* genomes are approximately the same size about twice as many folds are known for the former in comparison to the latter. This undoubtedly reflects the biased nature of the structure database.

It is possible to classify each fold as all-α, all-β, α/β, $\alpha+\beta$, or other using the original definitions of Levitt & Chothia (1976) and then to see how the folds corresponding to each of the structural classes are distributed among the genomes (Figure 5). Overall, the genomes contain a disproportionate number of mixed folds (α/β and $\alpha+\beta$, 83/148). Yeast also has the most all-β folds (with 14 of the 18 all-β folds in the three genomes).

There are 45 domain folds shared between the three genomes. These presumably represent a most ancient set of molecular parts. They include such diverse folds as that of a common, metabolic enzyme, similar in structure to flavodoxin, and that of a domain involved in tRNA recognition (specifically, the substrate-binding domain of the dehydrogenase 2DLD and the C-terminal, anticodon-binding domain of the glutamyl-tRNA synthetase 1GLN). These 45 folds are especially enriched in α/β supersecondary structures, with 38 of 45 having one of the mixed architectures.

Top-10 folds in each genome and five basic folds

Finally, it is possible to look at the frequency with which the known folds occur in the genomes. This is shown in the form of top-10 lists (Table 4). As was the case for the folds overall, most of the common folds have an α/β architecture. This is especially true for the common, "shared" folds that are present in all three genomes. The nine most common of these folds all have an α/β architecture.

The five most common folds that are present in all three genomes are shown in Figure 6. Ordered in terms of the frequency of their occurrence (see legend to Figure 6), they are: the P-loop containing NTP hydrolase fold, the Rossmann fold, the TIM-barrel fold, the flavodoxin fold, and the thiamin-binding fold. Each of these overall "top-5" folds

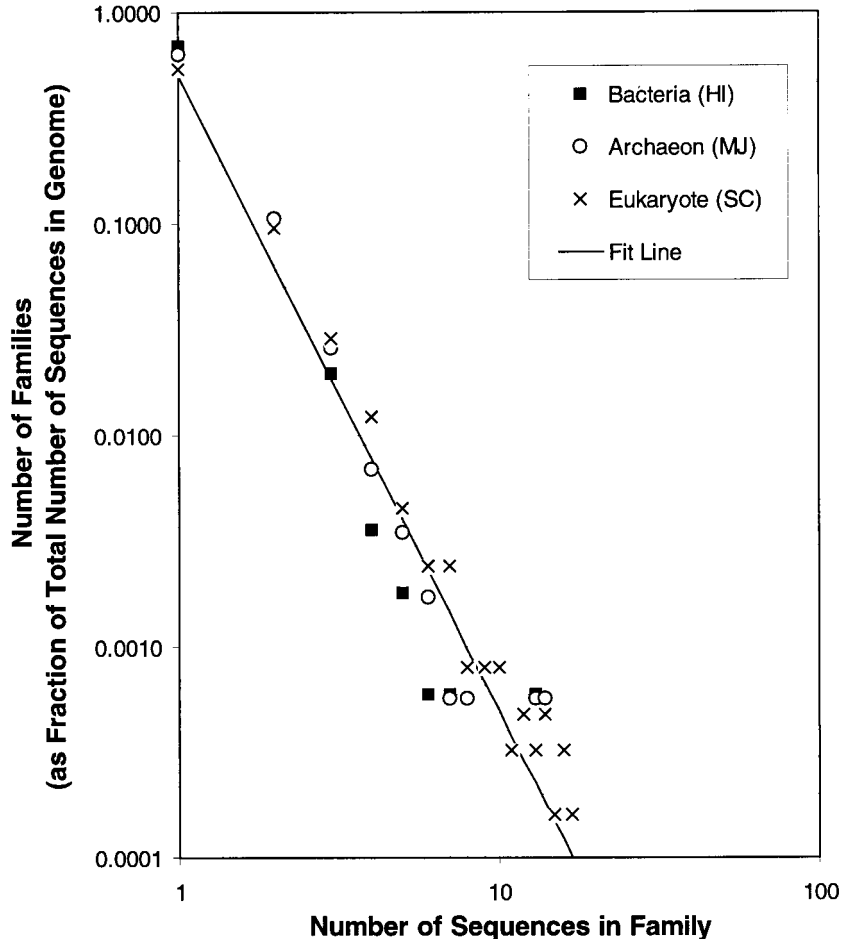

Figure 3. Number of duplications in the three genomes. Log–log graph showing how the number of sequence families (i.e. clusters of paralogs, shown on the vertical axis) of a given size (shown on the horizontal axis) drops off in nearly the same way in each of the three genomes. The exact numbers that this chart is based on are shown below.

Size of family	Frequency in genome HI	MJ	SC
1	1172	1095	3347
2	161	183	599
3	33	45	180
4	6	12	76
5	3	6	28
6	1	3	15
7	1	1	15
8	0	1	5
9	0	0	5
10	0	0	5
11	0	0	2
12	0	0	3
13	1	1	2
14	0	1	3
15	0	0	1
16	0	0	2
17	0	0	1
18	0	0	3
19	0	0	0
20	0	0	0
>20	1	0	4

occurs in the top-10 list for each of the individual genomes (with the exception of the flavodoxin fold, which does not occur in the yeast top-10). They are all associated with basic metabolism (as opposed to other functions such as transcription or regulation). They have a remarkably similar super-secondary structure architecture. They are all classic α/β proteins, containing a central sheet of parallel strands with helices packed onto at least one face of this sheet. (For this discussion, it is convenient to imagine the barrel in an unrolled fashion as shown in the Figure.) As emphasized in the schematic part of the Figure, the topology of the central sheet is very similar in the proteins. Almost all of the connections are right-handed links between adjacent parallel strands through an intervening helix packed onto the central sheet. (Specifically, 18 of the total 24 connections fall into this pattern, with five other ones being very similar, but involving connections between strands two apart in the sheet.)

Conclusion

Three genomes have been compared in terms of the protein structure, particularly supersecondary structure, they encode. This has demonstrated that even using as crude a measure as secondary-structure prediction, one can find marked, statistical differences between the genomes in terms of overall protein-structure features. It is found that yeast has more all-β supersecondary structure and *Haemophilus*, more all-α. This particular result is further borne out from looking at the distribution of known folds.

Secondly, using straightforward sequence comparison, it is possible to find that these diverse organisms share many common folds. In particular, there are 45 known folds common to the three kingdoms. Presumably, these are ancient folds that were present in the last common ancestor that predated the divergence of the major kingdoms about two billion years ago (Doolittle *et al.*, 1996). Five of these are amongst the most common folds in each organism and share a remarkably similar α/β architecture.

The two different general types of calculations performed here, structure prediction and sequence matching, give conclusions with complementary strengths and weaknesses. Structure prediction can be applied to the whole genome in a uniform fashion, providing comprehensive, statistical conclusions. However, these conclusions suffer from the inherent inaccuracy of the prediction methods, especially given that the application of structure prediction methods to genomes is such a great extrapolation from the data that the methods were

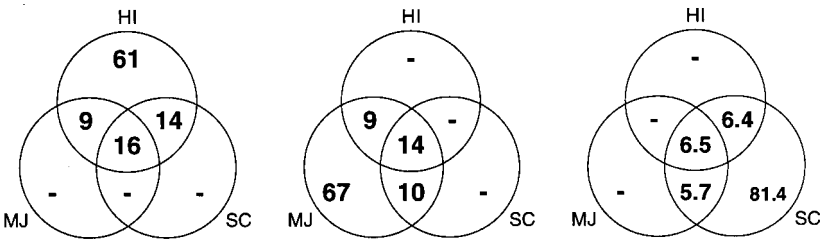

Figure 4. Sequence similarity between genomes. The Figure shows for each genome what fraction of its sequences (expressed as a percentage of the total, 100%) have homologs in the other two genomes. For instance, yeast has roughly equal number of sequences with homologs to *Methanococcus* (12.2% = 6.5% + 5.7%) and *Haemophilus* (12.9% = 6.5 + 6.4%).

trained on (i.e. the PDB). In contrast, sequence comparison to known structure gives almost completely accurate ''predictions'' (about the fold). However, only a small sample of each genome can be surveyed (7% to 15% of the ORFs), giving conclusions that are necessarily anecdotal and biased to a degree.

More refined methods of sequence comparison and secondary structure prediction (if developed) would allow a greater percentage of the genome to be matched with known folding patterns, further developing the conclusions of this paper. In any case, the general idea of comparing genomes in terms of protein structures is expected to be a very fertile topic in the future. There are currently seven microbial genomes completed and at least 36 more being worked on (Kerlavage, 1997), so there will be many possibilities for comparison soon.

Methods

A relational database of genome sequences and structure assignments

Translated genome sequences were taken from the web sites (www.tigr.org and genome-www.stanford.edu). Note that there is some uncertainty regarding whether all of the translated open reading frames (ORFs) are really genes. For instance, in yeast 5888 of the 6218 ORFs are definitely believed to be genes, but there is some uncertainty about the remaining 330 (Goffeau *et al.*, 1996). Non-genome sequences were taken from the non-redundant OWL databank (version 27.1) (Bleasby *et al.*, 1994), structures from the PDB (Bernstein *et al.*, 1977), and domain fold definitions from scop (version 1.32, May 1996) (Brenner *et al.*, 1996; Murzin *et al.*, 195). Core structures for each

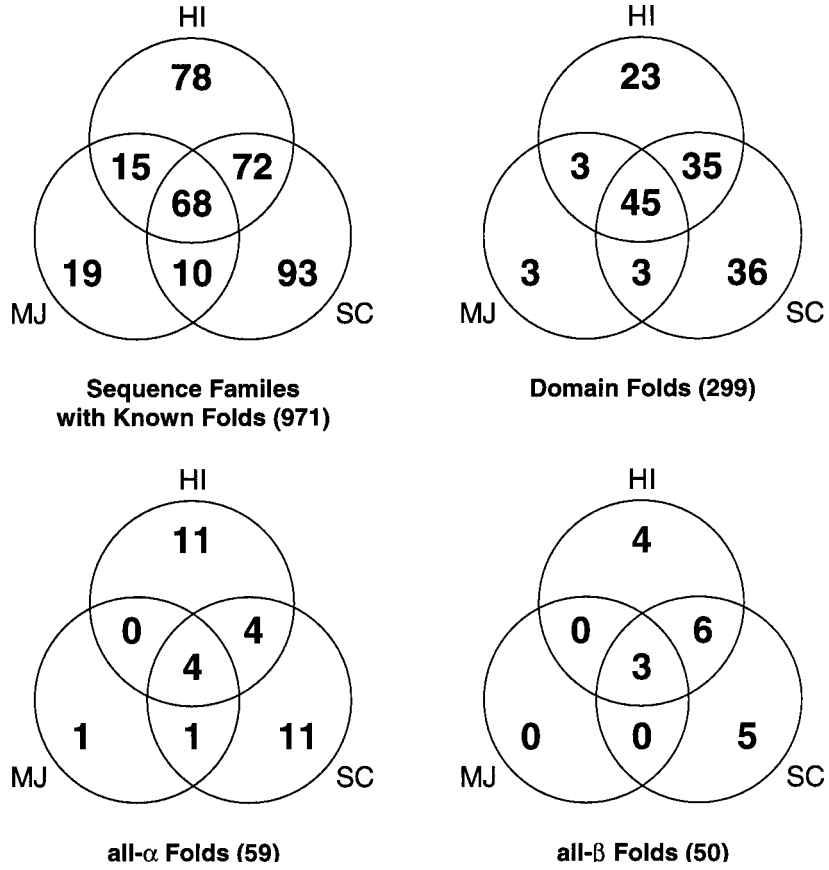

Figure 5. Folds shared between genomes. This Figure shows how the known folds are distributed amongst the three genomes in terms of a Venn diagram. All the domains in the PDB can be clustered into 971 sequence families at 40% homology (Brenner *et al.*, unpublished results). Of these 355 appear in at least one of the three genomes and how they are distributed is indicated in the top-left panel. As shown in the top-right panel, the structural similarities in scop (Murzin *et al.*, 1995) collapse many of the 971 sequence families into a smaller number (299) of fold families. Of these 148 appear in the three genomes, with 45 shared between all three. The bottom-left and bottom-right panels show selections of fold families shown at top-right, corresponding, respectively, to all-α or all-β structures. Note that the total number of folds consists of essentially five parts: all-α folds, all-β folds, α/β folds, α + β folds, and miscellaneous folds (148 = 32 + 18 + 47 + 36 + 15). Thus, from subtracting the two bottom panels from the top-right one, one can see how the overall prevalence of mixed (α/β and α + β) folds is distributed amongst the genomes.

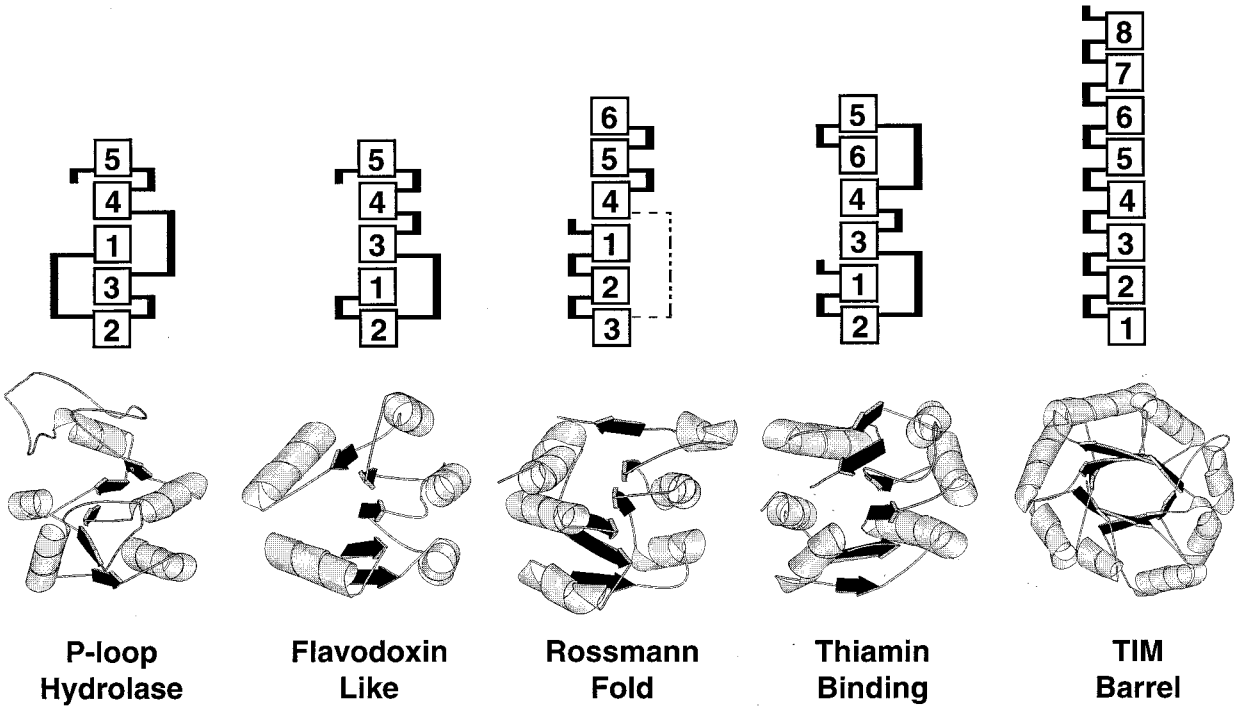

Figure 6. Five basic molecular parts common in all three genones. The Figure shows five basic molecular parts, five folds that are most common in all three genomes. Here ''commonness'' is determined by the average frequency rank $R_{overall}$ of the fold over each of the three genomes. Specifically, $R_{overall} = (R_{HI} + R_{SC} + R_{MJ})/3$, where R_{HI} is rank in the *Haemophilus* ''top-10'' list in Table 4 and the other quantities are defined analogously. Using this definition, the five most common folds are: P-loop hydrolase (average rank of 2), Rossmann (2.7), TIM-barrel (3.7), flavodoxin (7), and thiamin-binding (7.7). All folds are drawn with MOLSCRIPT (Kraulis, 1991) using residue selections from Table 4. They are somewhat simplified so that coil geometry is smoothed out and insertions not packing against the central sheet are de-emphasized. Also shown are highly schematic views of the sheet topology. Boxes indicate parallel strands in a beta-sheet with their order noted. (Strands are coming out of the page.) Solid arcs joining the boxes indicate right-handed connections between the parallel strands. All of these involve skipping no more than two strands and are through a parallel helix packed onto the sheet, from above or below. Half of an arc indicates that there is a parallel helix connected to either the first or last strand of the sheet. There is one exceptional connection, indicated with a dotted line: in the Rossmann fold there is a connection across three strands through a parallel helix.

scop domain were based on refinement of structural alignments (Altman & Gerstein, 1994; Gerstein & Altman, 1995; Gerstein & Levitt, 1996, 1997b). In total the domain definitions in scop correspond to a set of 971 non-homologous sequences (Brenner *et al.*, unpublished results), after they have been clustered to level of 40% identity by a similar procedure to that of Hobohm & Sander (1994).

Analysis and processing of the data was greatly expedited by the use of a simple relational database formalism, implemented in DBM, Perl5 (Wall *et al.*, 1996) and mini-SQL (http://Hughes.com.au), with tables linking sequence identifiers, structure matches, fold identifiers, and so forth. A number of these tables will be made available over the internet from the following URL: http://bioinfo.mbb.yale.edu/census.

Sequence comparisons

All sequence matching was done with the FASTA program (version 2.0) (Lipman & Pearson, 1985; Pearson & Lipman, 1988) with k-tup 1 and a conservative ''e-value'' cutoff of 0.001. The e-value describes the number of false positives expected in a single database scan, so a value of 0.001 means that no more than one out of every thousand cluster linkages will be in error (Brenner *et al.*, 1995, and unpublished results; Pearson, 1996). This error rate has been verified by empirical tests on a database of known protein relationships (Brenner *et al.*, 1995, and unpublished results). An e-value cutoff should give similar results to a more conventional threshold in terms of percent identity, but it has been shown to be better calibrated and more sensitive for marginal similarities, taking into account compositional biases of the databank and the query sequence (Altschul *et al.*, 1994; Karlin & Altschul, 1993). Low complexity sequences were filtered out using the SEG program (Wooton & Federhen, 1993). Many of the sequence comparisons took a very long time, particularly running the yeast genome against all known sequences. (This took about a month and four days on a DEC workstation with a 266 MHz alpha CPU.) However, the calculations are intrinsically quite parallel and could be greatly speeded up by distributing them over a number of processors.

Table 4. Top-10 folds in the three genomes

No. in genome	Class	Fold name	Representative structure (PDB selection)
Top-10 in a eukaryotic genome (SC)			
84	α + β	Protein kinases (catalytic core)	1irk
49	α/β	P-loop containing NTP hydrolasses	1gky
35	α/β	Rossmann fold	2ohx A:175–324
31	A/B	TIM barrel	1tim A:
25	α/β	Ribonuclease H-like	2rn2
18	S	Classic zinc finger	1zaa C:
14	α + β	Ubiquitin conjugating enzyme	1aak
12	β	GroES-like	1acy L: 109–211
10	α/β	Thioredoxin-like	1trx
9	α/β	Thiamin-binding fold	1pvd A:2–181
5 × 8	...	...	...
7	α/β	Flavodoxin-like	3chy
Top-11 in a eubacterial genome (HI)			
18	α/β	Rossmann fold	2ohx A:175–324
13	α/β	P-loop containing NTP hydrolases	1gky
12	α/β	Flavodoxin-like	3chy
10	α/β	TIM barrel	1tim A:
10	α + β	Ferredoxin-like	1fxd
10	α/β	Ribonuclease H-like	2rn2
6	α/β	Periplasmic binding protein-like II	1sbp
5	α/β	Periplasmic binding protein-like I	2dri
5	α + β	Like class II aaRS synthetases	1sry A:111–421
4	β	OB-fold	1pyp
4	α/β	Thiamin-binding fold	1pvd A:2–181
Top-11 in an archaeal genome (MJ)			
19	α + β	Ferredoxin-like	1fxd
10	α/β	P-loop containing NTP hydrolases	1gky
7	α/β	TIM barrel	1tim A:
6	α/β	Rossmann fold	2ohx A:175–324
5	α	Histone-fold	1ntx
4	α/β	Thiamin-binding fold	1pvd A:2–181
4	α/β	Flavodoxin-like	3chy
4	β	Reductase/elongation factor common	1efg A:283–403
3	α + β	ATP-grasp	1bnc A:115–330
3	α/β	PLD-dependent transferases	1dka
3	α/β	ATP pyrophoshatases	1gpm A:208–404

The most common of the known folds in the various genomes. Folds common to all three genomes are underlined. The flavodoxin-like fold occurs seven times in the yeast genome, making it the 16th most common fold in this genome.

There are more sensitive methods of comparing sequences to structures than the FASTA program, e.g. profiles, Hidden-Markov models, motif analysis, and threading (Bowie & Eisenberg, 1993; Jones & Thornton, 1996; Eddy, 1996). These methods would be expected to find more homologues for certain folds. (For instance, using careful sequence comparison and motif analysis Hunter & Plowman (1997) recently reported that yeast had 113 protein kinases, considerably more than the number reported here.) However, the sensitivity improvement would *not be uniform* over all folds. This is not advantageous for a large-scale census since uniform sampling and treatment of the data is more important than sensitivity (as one is more concerned with relative rather than absolute numbers).

Multiple linkage clustering

The sequences were grouped into families by applying single and multiple linkage clustering (Kaufman & Rousseeuw, 1990) to an all-against-all comparison of the three genomes, taken individually and jointly, plus a set of non-homologous sequences corresponding to the structures in scop. The basic clustering procedure was very similar to algorithm 1 (''select until done'') by Hobohm *et al.* (1992) with the inclusion of a test for multiple linkage. Initially one starts with an empty list of sequences in the first cluster and a list of N unassigned sequences (which can be ordered according to a useful criteria such as length). One then takes the first unassigned sequence (s_0) and compares it sequentially to each of the other unassigned sequences (s_i). At each step, if the sequence being compared (s_i) matches the first sequence (s_0) *and* every other sequence in the current first cluster list, it is added to the (growing) first cluster. The test here for multi-linkage addresses the ''domain problem'' described in Figure 2. Then one repeats the whole procedure, creating new clusters at each iteration, until all the unassigned sequences are exhausted. If the mean cluster size is C, it is possible to show that this procedure requires

approximately:

$$\left(\frac{N-1-(C+1)}{C+1}\right)\left(\frac{N-1}{2}\right)$$

comparisons, with approaches $N^2/2(C+1)$ for large N and $N \gg C$. This is significantly less than the number of operations when one considers all pairs: $N(N-1/2)$, which approaches $N^2/2$ for large N.

Transmembrane helix prediction

Transmembrane segments were identified by using the GES hydrophobicity scale, shown in Table 2 (Engelman *et al*., 1986). 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 seven, 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 (Arkin *et al*., 1997; Tomb *et al*., 1997). Here detection of at least two transmembrane helices was necessary for the protein to be classified as a ''membrane protein.'' This rather conservative threshold was used for two reasons: (i) sequences with only a single transmembrane element are not really integral membrane proteins, and (ii) the false positive rate for misclassifying proteins as membrane proteins was much higher on the basis of one transmembrane helix than multiple ones. (Specifically, the error rate on a test set of 395 non-homologous chains was ≈4% (21/395) for classifications based on a single transmembrane helix *versus* ≈1% (4/395) for those based on two or more.)

Transmembrane identifications based on the GES scale were compared against those based on the Kyte–Doolittle scale using both a strict and more lax threshold (Jähnig, 1990; Kyte & Doolittle, 1982). They were also compared against a neural network approach (Rost *et al*., 1995, 1996). There are differences in the predictions for a number of the sequences. However, the overall results of the bulk prediction are fairly insensitive to this. For instance, the number of membrane proteins predicted using the GES scale in *Haemophilus, Methanococcus*, and yeast were 303, 232, and 1407, respectively. The corresponding numbers for the Kyte–Doolittle scale (with the strict threshold) were 386, 358, and 1845. The GES scale is a little more conservative, but the relative amounts of membrane proteins in the three genomes are similar. The differences between the scales reflect the extremely small size of the membrane protein database that the predictions extrapolate from.

Secondary structure prediction

Secondary structure prediction was done using the GOR program (Garnier *et al*., 1978, 1996; Gibrat *et al*., 1987). This is a well-established and commonly used method. It is statistically based so that the prediction for a particular residue (say Ala) to be in a given state (i.e. helix) is directly based on the frequency that this residue (and taking into account neighbors at ±1, ±2, and so forth) occurs in this state in a database of solved structures. Specifically, version 4 of the GOR program is used here (Garnier *et al*., 1996). This bases the prediction for residue i on a window from $i-8$ to $i+8$ around i, and within this window, the 17 individual residue frequencies (singlets) are combined with the frequencies of all 136 possible di-residue pairs (doublets). The GOR method only uses single sequence information and because of this achieves lower accuracy (65% *versus* 71%) than the current ''state-of-art'' methods that incorporate multiple sequence information (King & Sternberg, 1996; Rost, 1996; Rost & Sander, 1993; Salamov & Solovyev, 1995). However, it is not possible to obtain multiple sequence alignments for most of the proteins in each of the three genomes. Consequently, bulk predictions of all the proteins in a genome based on multiple-alignment approaches are in a sense skewed. One gets two distinctly different types of prediction, depending on how many homologues a given protein has. Consequently, for the bulk prediction approaches used here the simpler single sequence approach was deemed more consistent. However, for one genome (*Haemophilus*) tests were also done using the PHD server (Rost, 1996). This does predictions using a neural-network, multiple sequence alignment approach. While obviously the predictions differed at individual positions from that of the GOR approach, most of the aggregate properties (e.g. total number and size of helices and strands) were fairly consistent. (In particular, on a random sample of 125 proteins that did not contain transmembrane elements or have PDB homologues, PHD predicted 38% of the residues to be helical, 18% to be strand, and the rest coil: the corresponding numbers for GOR are 36% and 18%.)

Note also that the analysis here is not at all focused on the particular secondary structure prediction for any individual residue. What is of concern is aggregate secondary structure content of whole proteins (and genomes) and the prediction of secondary structure elements (i.e. whether or not a helix is present, regardless of its length). Prediction of aggregate quantities is expected to be more accurate than the prediction of individual residues (Rost & Sander, 1993).

Frequent words for super-secondary structures

To simplify the secondary structure predictions, they were ''condensed'' into simple ''αβ-code'' where ''α'' stands for the position of a helix and

''β'' for a strand. Specifically, for the condensation, two or more adjacent residues of strand or helix were merged into single ''α'' or ''β'' characters. (The few isolated residues predicted to be in different conformation than both their neighbors were sometimes merged with their neighbors or directly promoted to ''α'' or ''β'' characters.)

The frequency of various two to ten letter words in each sequence's αβ-code was assessed using methods developed for analyzing frequent words in nucleotide sequences (Karlin *et al.*, 1992, 1996; Karlin & Cardon, 1994). The observed frequency of words of a given type was divided by the expected frequency for this type, based on the frequencies of individual α and β elements, to form an odds ratio *R*. For instance, for the pattern αβββ the odds ratio is:

$$R = \left(\frac{N(\alpha\beta\beta\beta)}{N(????)}\right) \Big/ \left(\frac{N(\alpha)}{N(?)}\frac{N(\beta)}{N(?)}\frac{N(\beta)}{N(?)}\frac{N(\beta)}{N(?)}\right)$$

where N (*xy*) is the number of words matching pattern ''*xy*'' with ''?'' as the wildcard, so $N(????)$ is the total number of four-character words and $N(\alpha)$ is the number of α's (helices). This analysis assumes that the expected frequency of words only depends on the frequency of individual characters (e.g. $N(\alpha)$). However, it is also possible to do the analysis taking into account higher-order frequencies and conditional probabilities so the odds ratio becomes:

$$R = \left(\frac{N(\alpha\beta\beta\beta)}{N(????)}\right) \Big/ \left(\frac{N(\alpha)}{N(?)}\frac{N(\alpha\beta)}{N(??)}\frac{N(\beta\beta)}{N(??)}\frac{N(\beta\beta)}{N(??)}\right)$$

Both analyses were used here, but there was little difference in the results.

Acknowledgments

D. Engelman, L. Regan, F. Richards, A. Brünger, W. Krebs, and T. Johnson are acknowledged for comments on the manuscript. Helpful correspondence with M. Cherry and A. Murzin is appreciated. Support for this work was provided by an ONR Young Investigator Grant (N00014-97-1-0725).

References

Altman, R. & Gerstein, M. (1994). Finding an average core structure: application to the globins. In *Proceedings of the Second International Conference on Intelligent Systems in Molecular Biology,* pp. 19–27, AAAI Press, Menlo Park, CA.

Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994). Issues in searching molecular sequence databases. (Review). *Nature Genet.* **6**, 119–129.

Arkin, I., Brunger, A. & Engelman, D. (1997). Are there dominant membrane protein families with a given number of helices? *Proteins: Struct. Funct. Genet.* In the press.

Berman, A. L., Kolker, E. & Trifonov, E. N. (1994). Underlying order in protein sequence organization. *Proc. Natl Acad. Sci. USA,* **91**, 4044–4047.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **122**, 535–542.

Blaisdell, B. E., Campbell, A. M. & Karlin, S. (1996). Similarities and dissimilarities of phage genomes. *Proc. Natl Acad. Sci. USA,* **93**, 5854–5859.

Bleasby, A. J., Akrigg, D. & Attwood, T. K. (1994). OWL – a non-redundant composite protein sequence database. *Nucl. Acids Res.* **22**, 3574–3577.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. & Sonnhammer, E. (1992a). Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome iii. *Protein Sci.* **1**, 1677–1690.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. & Sonnhammer, E. (1992b). What's in a genome? *Nature,* **358**, 287.

Bowie, J. U. & Eisenberg, D. (1993). Inverted protein structure prediction. *Curr. Opin. Struct. Biol.* **3**, 437–444.

Brenner, S., Hubbard, T., Murzin, A. & Chothia, C. (1995). Gene duplication in *H. Influenzae*. *Nature,* **378**, 140.

Brenner, S., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996). Understanding protein structure: using scop for fold interpretation. *Methods Enzymol.* **266**, 635–642.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., Fitzgerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J.-F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geohagen, N. S. M., Weidman, J. F., Fuhrmann, J. L., Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H.-P., Fraser, C. M., Smith, H. O., Woese, C. R. & Venter, J. C. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science,* **273**, 1058–1073.

Casari, G., Andrade, M., Bork, P., Boyle, J., Daruvar, A., Ouzounis, C., Schneider, R., Tamames, J., Valencia, A. & Sander, C. (1995). Challenging times for bioinformatics. *Nature,* **376**, 647–648.

Chakrabartty, A., Kortemme, T. & Baldwin, R. L. (1994). Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Sci.* **3**, 843–852.

Chothia, C. (1992). Proteins – 1000 families for the molecular biologist. *Nature,* **357**, 543–544.

Chothia, C. & Gerstein, M. (1997). Protein evolution. How far can sequences diverge? *Nature,* **385**, 579–581.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.

Clayton, R. A., White, O., Ketchum, K. A. & Venter, J. C. (1997). The first genome from the third domain of life (news). *Nature,* **387**, 459–462.

Doolittle, R. F. (1987). *Of Urfs and Orfs.* University Science Books, Mill Valley, CA.

Doolittle, R. F. (1995). The multiplicity of domains in proteins. (Review). *Annu. Rev. Biochem.* **64**, 287–314.

Doolittle, R. F., Feng, D. F., Tsang, S., Cho, G. & Little, E. (1996). Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science,* **271**, 470–477.

Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361–365.

Engelman, D. M., Steitz, T. A. & Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. (Review). *Annu. Rev. Biophys. Biophys. Chem.* **15**, 321–353.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. & Venter, J. C. (1995a). Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd. *Science (Washington DC),* **269**, 496–498.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. & Venter, J. C. (1995b). Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd. *Science (Washington DC),* **269**, 507–512.

Garnier, J., Osguthorpe, D. & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97–120.

Garnier, J., Gibrat, J. F. & Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* **266**, 540–553.

Gerstein, M. & Altman, R. (1995). Average core structures and variability measures for protein families: Application to the immunoglobulins. *J. Mol. Biol.* **251**, 161–175.

Gerstein, M. & Levitt, M. (1996). Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology,* pp. 59–67, AAAI Press, Menlo Park, CA.

Gerstein, M. & Levitt, M. (1997a). A structural census of the current population of protein sequences. *Proc. Natl Acad. Sci. USA*, In the press.

Gerstein, M. & Levitt, M. (1997b). Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Sci.* In the press.

Gibrat, J., Garnier, J. & Robson, B. (1987). Further developments of protein secondary structure prediction using information theory. *J. Mol. Biol.* **198**, 425–443.

Gibrat, J. F., Madej, T. & Bryant, S. H. (1996). Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6**, 377–385.

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

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S. G. (1996). Life with 6000 genes. *Science,* **274**, 546–567.

Goffeau, A., *et al.* (1997). The yeast genome directory. *Nature,* **387** (Suppl.), 5.

Green, P., Lipman, D., Hillier, L., Waterston, R., States, D. & Claverie, J. M. (1993). Ancient conserved regions in new gene sequences and the protein databases. *Science,* **259**, 1711–1716.

Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522.

Hobohm, W., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.

Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science,* **273**, 595–602.

Hunter, T. & Plowman, G. D. (1997). The protein kinases of budding yeast: six score and more. *Trends Biochem. Sci.* **22**, 18–22.

Jähnig, F. (1990). Structure predictions of membrane proteins are not that bad. *Trends Biochem. Sci.* **15**, 93–95.

Jones, D. T. & Thornton, J. M. (1996). Potential energy functions for threading. *Curr. Opin. Struct. Biol.* **6**, 210–216.

Karlin, S. & Altschul, S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl Acad. Sci. USA,* **90**, 5873–5877.

Karlin, S. & Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. (Review). *Trends Genet.* **11**, 283–290.

Karlin, S. & Cardon, L. R. (1994). Computational DNA sequence analysis. (Review). *Annu. Rev. Microbiol.* **48**, 619–654.

Karlin, S., Burge, C. & Campbell, A. M. (1992). Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucl. Acids Res.* **20**, 1363–1370.

Karlin, S., Mrazek, J. & Campbell, A. M. (1996). Frequent oligonucleotides and peptides of the haemophilus influenzae genome. *Nucl. Acids Res.* **24**, 4263–4272.

Karp, P., Riley, M., Paley, S. & Pellegrini-Toole, A. (1996a). EcoCyc: electronic encyclopaedia of *E. coli* genes and metabolism. *Nucl. Acids Res.* **24**, 32–40.

Karp, P. D., Ouzounis, C. & Paley, S. M. (1996b). HinCyc: a knowledge base of the complete genome and metabolic pathways of H. influenze. *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology,* pp. 116–124, AAAI Press, Menlo Park, CA.

Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New York.

Kerlavage, A. R. (1997). TIGR Microbial Genome Database. http://www.tigr.org/mdb (as of 2/97).

King, R. D. & Sternberg, M. J. E. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.* **5**, 2298–2310.

Koonin, E. V., Tatusov, R. L. & Rudd, K. E. (1995). Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications. *Proc. Natl Acad. Sci. USA,* **92**, 11921–11925.

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

Koonin, E. V., Tatusov, R. L. & Rudd, K. E. (1996b). Protein sequence comparison at a genome scale. *Methods Enzymol.* **266**, 295–322.
Kraulis, P. J. (1991). MOLSCRIPT – a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.
Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* **157**, 105–132.
Lander, E. S. (1996). The new genomics: global views of biology. *Science,* **274**, 536–539.
Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature,* **261**, 552–558.
Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science,* **227**, 1435–1441.
Murzin, A., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
Mushegian, A. R. & Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl Acad. Sci. USA,* **93**, 10268–10273.
Nowak, R. (1995). Bacterial genome sequence bagged. *Science,* **269**, 468–470.
Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature,* **372**, 631–634.
Ouzounis, C. & Kyrpides, N. (1996). The emergence of major cellular processes in evolution. *FEBS Letters,* **390**, 119–123.
Ouzounis, C., Bork, P., Casari, G. & Sander, C. (1995a). New protein functions in yeast chromosome VIII. *Protein Sci.* **4**, 2424–2428.
Ouzounis, C., Kyrpides, N. & Sander, C. (1995b). Novel protein families in archaean genomes. *Nucl. Acids Res.* **23**, 565–570.
Pascarella, S. & Argos, P. (1992). A databank merging related protein structures and sequences. *Protein Eng.* **5**, 121–137.
Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* **266**, 227–259.
Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence analysis. *Proc. Natl Acad. Sci. USA,* **85**, 2444–2448.
Riley, M. & Labedan, B. (1997). 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.
Rost, B. (1996). PHD: Predicting one-dimensional protein secondary structure by profile-based neural networks. *Methods Enzymol.* **266**, 525–539.
Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599.
Rost, B., Fariselli, P., Casadio, R. & Sander, C. (1995). Prediction of helical transmembrane segments at 95% accuracy. *Protein Sci.* **4**, 521–533.
Rost, B., Fariselli, P. & Casadio, R. (1996). Topology prediction for helical transmembrane segments at 95% accuracy. *Protein Sci.* **7**, 1704–1718.
Salamov, A. & Solovyev, V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.* **247**, 11–15.
Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.
Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C. & Sander, C. (1994). GeneQuiz: a workbench for sequence analysis. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology,* pp. 348–353, AAAI Press, Menlo Park, CA.
Schmidt, R., Gerstein, M. & Altman, R. (1996). LPFC: an internet library of protein family core structures. *Protein Sci.* **6**, 246–248.
Smith, C. K., Withka, J. M. & Regan, L. (1994). A thermodynamic scale for the beta-sheet forming tendencies of the amino acids. *Biochemistry,* **33**, 5510–5517.
Sonnhammer, E. L. L. & Kahn, D. (1994). Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.* **3**, 482–492.
Sonnhammer, E., Eddy, S. & Durbin, R. (1996). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Struct. Funct. Genet.* **28**, 405–420.
Tamames, J., Casari, G., Ouzounis, C. & Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**, 66–73.
Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P., Hayes, W. S., Borodovsky, M., Rudd, K. E. & Koonin, E. V. (1996). Metabolism and evolution of *Haemophilus influenzae* deduced from a whole- genome comparison with *Escherichia coli*. *Curr. Biol.* **6**, 279–291.
Tomb, J.-F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., Fitzegerald, L. M., Lee, N., Adams, M. D., Hickey, E. K., Berg, D. E., Gocayne, J. D., Utterback, T. R., Peterson, J. D., Kelley, J. M., Cotton, M. D., Weidman, J. M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W. S., Borodovsky, M., Karpk, P. D., Smith, H. O., Fraser, C. M. & Venter, J. C. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature,* **388**, 539–547.
Wade, N. (1997). Thinking small paying off big in gene quest. In *New York Times* 3 February 1997, p. A1 (front page).
Wall, L., Christiansen, D. & Schwartz, R. (1996). *Programming PERL.* O'Reilly and Associates, Sebastapol, CA.
Wolfe, K. H. & Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature,* **387**, 708–713.
Wooton, J. C. & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17**, 149–163.

Edited by F. E. Cohen

(Received 25 February 1997; received in revised form 4 September 1997; accepted 4 September 1997)