J Mol Evol (1999) 49:551–557

JOURNAL OF MOLECULAR EVOLUTION



# Functional Classes in the Three Domains of Life

**M.A. Andrade,[1] C. Ouzounis,[1] C. Sander,[1] J. Tamames,[2] A. Valencia[2]**

[1] EMBL-EBI, Wellcome Trust Genome Campus, Cambridge, UK
[2] Protein Design Group, CNB-CSIC, Madrid, Spain

Received: 21 July 1997 / Accepted: 5 May 1999

**Abstract.** The evolutionary divergence among the three major domains of life can now be addressed through the first set of complete genomes from representative species. These model species from the three domains of life, *Haemophilus influenzae* for Bacteria, *Saccharomyces cerevisiae* for Eukarya, and *Methanococcus jannaschii* for Archaea, provide the basis for a universal functional classification and analysis. We have chosen 13 functional classes and three superclasses (ENERGY, COMMUNICATION and INFORMATION) as global descriptors of protein function. Compositional comparison of the three complete genomes reveals that functional classes are ubiquitous yet diverse in the three domains of life. Proteins related with ENERGY processes are generally represented in all three domains, while those related with COMMUNICATION represent the most distinctive functional feature of each single domain. Finally, functions related with INFORMATION processing (translation, transcription, and replication) show a complex behaviour. In Archaea, proteins in this superclass are related with proteins in either Eukarya or Bacteria, as recognized previously. The distribution of functional classes in the three domains accurately reflects the principal characteristics of cellular life forms.

**Key words:** Genome comparison — Functional classes — *Haemophilus influenzae* — *Saccharomyces cerevisiae* — *Methanococcus jannaschii* — Archaea

*Correspondence to:* Alfonso Valencia, CNB-CSIC, Campus U. Autonoma, Cantoblanco, Madrid 28049, Spain; e-mail: valencia@cnb.uam.es

## Introduction

The three domains of cellular life—Bacteria, Archaea, and Eukarya—exhibit differences that can possibly be attributed to their genome structure and composition. The availability of the first bacterial (*Haemophilus influenzae*) (HI) (Fleischmann et al. 1995), eukaryotic (*Saccharomyces cerevisiae*) (SC) (Goffeau et al. 1996), and archaeal (*Methanococcus jannaschii*) (MJ) (Bult et al. 1996) genomes as well as a large number of additional eukaryotic and bacterial sequences provide a unique opportunity to approach this question. However, a comparative analysis will be meaningful only to the extent that these three species have representative genomes for the corresponding domains. For example, the genome of *Mycoplasma genitalium* (Fraser et al. 1995) does not adequately represent Bacteria, given the lack of many important functions derived from its parasitic lifestyle.

To minimize domain misrepresentation by the abovementioned genomes, we compared them with the full set of sequences in the database using phylogenetic criteria, instead of restricting the analysis to species intersections. The number of sequences in the archaeal domain is still insufficient to be used as a reference set, limiting the analysis to the distribution of sequences in the eukaryotic and bacterial domains.

The analysis is based on the presence or absence of key cellular functions in these model genomes. Cellular function was reduced to a comprehensive set of 13 classes of functions derived from a previously proposed scheme (Riley 1993), later applied to the HI (Fleischmann et al. 1995) and MJ (Bult et al. 1996) genomes. For the SC genome, a similar classification was provided

**Table 1.** Percentages of the SC and HI proteins of each functional class with at least one homologue in Bacteria and Eukarya, respectively

| | SC | | | HI | | |
|---|---|---|---|---|---|---|
| Functional class | Number | In Bacteria | % | Number | In Eukarya | % |
| Amino acid biosynthesis | 99 | 84 | 85 | 67 | 53 | 79 |
| Biosynthesis of cofactors | 76 | 44 | 58 | 54 | 30 | 56 |
| Central & int. metabolism | 186 | 103 | 55 | 30 | 19 | 63 |
| Energy metabolism | 288 | 199 | 69 | 105 | 54 | 51 |
| Fatty acids & phospholipids | 86 | 56 | 65 | 26 | 17 | 65 |
| Nucleotide biosynthesis | 220 | 134 | 61 | 53 | 33 | 62 |
| Transport | 416 | 165 | 40 | 120 | 38 | 32 |
| **Energy** | **1371** | **785** | **57** | **455** | **244** | **54** |
| Replication | 273 | 127 | 47 | 83 | 23 | 28 |
| Transcription | 305 | 126 | 41 | 26 | 11 | 42 |
| Translation | 283 | 227 | 80 | 142 | 86 | 61 |
| **Information** | **861** | **480** | **56** | **251** | **120** | **48** |
| Cell envelope/cell wall | 156 | 25 | 16 | 76 | 5 | 7 |
| Cellular processes | 424 | 202 | 48 | 53 | 20 | 38 |
| Regulatory functions | 355 | 75 | 21 | 66 | 19 | 29 |
| **Communication** | **935** | **302** | **32** | **195** | **44** | **23** |
| Other | 37 | | | 94 | | |
| Unassigned | 3074 | | | 685 | | |
| **Total** | **6278** | | | **1680** | | |

by an automatic functional classification system (Tamames et al. 1996, 1998).

## Methods

### *Similarity Searches and Family Definition*

The complete set of open reading frames (ORFs) available for HI, MJ, and SC was used for the analysis (1680 HI, 1735 MJ, and 6278 SC ORFs). Similarity searches have been carried out with the GeneQuiz system (Casari et al. 1996a, b; Andrade et al. 1999), which includes a careful selection of updated databases, a composition bias masking procedure (Ouzounis and Casari, unpublished), and an integrated selection of searching methods (Altschul et al. 1990; Sander and Schneider 1991; Henikoff and Henikoff 1996; Pearson 1996; Bairoch et al. 1997). A combination of scores from different sequence similarity methods is used to describe sequence similarities as "clear." At this level, we have estimated the reliability of the system at 95% correct similarity assignments (Ouzounis et al. 1996). The limitations of the system include inaccurate (or wrong) database annotations that are used for assignment and false-positive or -negative cases. This clear level provided by GeneQuiz is equivalent to the more familiar $P(n)$ value at 10e-10 in BLAST (Altschul et al. 1990) or a FASTA score (Pearson 1996) of 135. This single cutoff for the definition of family boundaries was used, to make results comparable. A similar procedure has been adopted by other authors in large-scale sequence comparisons (Sander and Schneider 1991; Riley 1993). We have not observed significant differences using stricter cutoffs (data not shown). Results on the searches and functional classifications for the three presented genomes are available at http://www.sander.ebi.ac.uk/overlap_2/.

### *Assignment of Sequences to Functional Classes*

Different authors have classified sequences into functional classes (Fleischmann et al. 1995; Bult et al. 1996; Tamames et al. 1996) according to the scheme of Riley (1993). The scheme used here is a slight modification of the original one, using a simple hierarchy (Table 1). Classes have been defined as sets of proteins that have a general overall function, and are usually not related by evolution. Generally, proteins are assigned to a single class, although this may not always reflect biological reality (some proteins could in principle be assigned to multiple classes).

Even if there are a number of caveats around this classification, in practice, besides being the only one available, it is not clear if the field is mature enough for reaching a consensus classification of protein function. Therefore, we have adopted the original classifications for the MJ (Bult et al. 1996) and HI (Fleischmann et al. 1995) proteins. A similar classification is unavailable for SC, and the one deposited at MIPS (http://speedy.mips.biochem.mpg.de/mips/SC/) is difficult to equate with the ones of HI and MJ. Such an equivalent classification is essential for the present work. Therefore, we have used a method (Tamames et al. 1996, 1998) that performs automatic class assignment based on the SwissProt (Bairoch and Apweiler 1997) key word usage, from a large corpus of sequences coherently classified in any given scheme.

We were able to classify automatically approximately half of the SC ORFs (3167 sequences) into the 13 classes. There exist a number of sequences impossible to classify automatically because their function (or the function of similar sequences) has not been annotated in SwissProt. Given the large number of sequences in SC, the limited coverage of the assignment to functional classes did not constitute a major problem.

The accuracy of automatic function assignment is approximately 80% for all functional classes and 92% for the three superclasses, with MJ genome as a test case (Andrade et al. 1997; Tamames and Valencia, unpublished). The main sources of error are ill-defined functional classes, such as CENTRAL INTERMEDIARY METABOLISM and REGULATORY FUNCTIONS. Since the SC has been better characterized biochemically and has more similar sequences in current databases than MJ, levels of accuracy are expected to be higher.

### *Species Names Assignment*

For each sequence and its homologues, the species (and domain) distribution is recorded. This task is a nontrivial one, given the lack of systematic species annotations in different databases. Species names

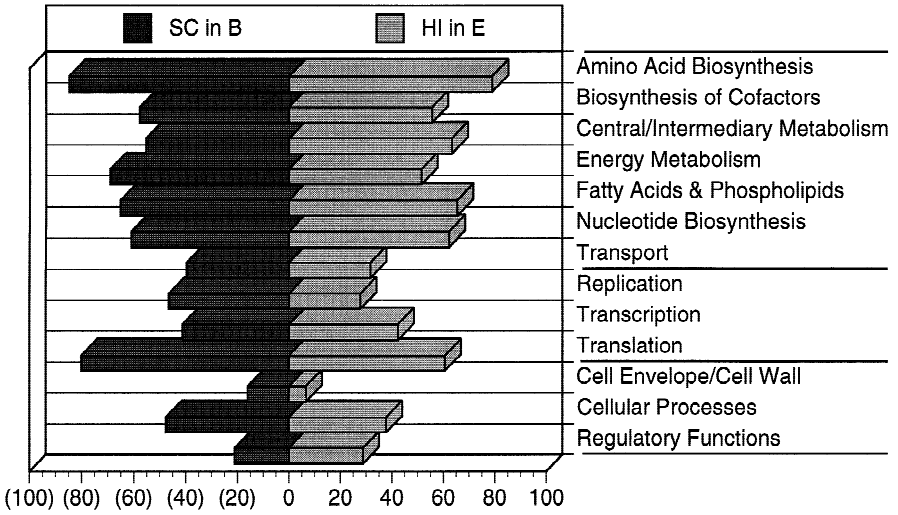


**Fig. 1.** Comparison of SC with Bacteria and HI with Eukarya by functional class. The percentage of SC sequences with at least one homologue in bacterial species (**left**) and HI sequences with at least one homologue in eukaryotic species (**right**). The results are represented for 13 classes of cellular function (see also Table 1).

for all sequences in the EMBL (Stoesser et al. 1997), PIR (George et al. 1996), and Genbank (Benson et al. 1997) databases were translated to the corresponding species and higher taxa names used in SwissProt (Bairoch and Apweiler 1997). Corrections for typographical errors were taken into account. Around 45,000 sequences from EMBL and Genbank databases remain unclassified (December 1997), corresponding mainly to incomplete species annotation such as Bacteria or to species not yet present in SwissProt, e.g., *Pristionchus pacificus* (704 sequences in the EMBL database). For most of these cases, we were able to extract enough taxonomic information for the purposes of the present comparison. Finally, only 200 sequences were not classified into taxa, including mostly artificial sequences.

### *Additional Technical Details*

A total of 9693 sequence searches was performed using GeneQuiz. The EMBL nonredundant protein database was used, with more than 200,000 entries. All computations were done on a SGI IRIX64 using eight processors. The computation rate was about 60 sequences per day per processor. The storage of results was about 2000 kbytes/sequence. Information on GeneQuiz, the present genomes and ongoing genome analysis efforts can be accessed at http://www.sander.ebi.ac.uk/genequiz/. The present results and additional material can be accessed at: http://www.sander.ebi.ac.uk/overlap_2/.

## Results

The analysis presented here was done at the level of functional classes corresponding to major cellular functions, instead of focusing on particular protein families. Two different levels of description are introduced, classifying proteins in three superclasses or 13 classes of cellular function (Table 1). Each sequence of the three model genomes (HI, MJ, and SC) was classified in one of these functional classes, if possible, and similar sequences in the corresponding domains were recorded.

The results are presented in the following order.

(I) We analyze the distribution of the SC proteins in Bacteria.

(II) We perform the same analysis of the HI proteins in Eukarya.

These two comparisons represent symmetrical views of the same functional overlap between Bacteria and Eukarya. None of these genomes is compared to Archaea, because of lack of sequence information (see next).

(III) The archaeal genome (MJ) is compared with both Bacteria and Eukarya. This comparison provides an overview of the functional differences between Archaea and the other two domains.

### *Analysis of* S. cerevisiae *as the First Eukaryotic Genome*

The 6278 yeast (SC) ORFs (Goffeau et al. 1996) have been compared with current database entries. For each one of them, the set of sequences in the database with a "clear" sequence similarity were considered equivalent, at the level of functional class, and selected for the analysis. The number of sequences (and percentage of the SC genome) with respect to sequence similarity and/or function assignment into the 13 classes are as follows: 3167 ORFs (50%) have homologues, function, and class assignment; 1228 (20%) have homologues and predicted function but no class; 1097 (17%) have homologues but no function or class assignment; and finally, another 749 (12%) do not have any homologues. There remain only 37 ORFs that are classified in another category, not considered further here (Table 1).

For all SC ORFs, the presence of a bacterial homologue was recorded and summarized by functional class (Table 1 and Fig. 1). Some functional classes, for example, NUCLEOTIDE BIOSYNTHESIS, are well represented in Bacteria, while other classes such as CELL ENVELOPE/CELL WALL, are obviously very different (Fig. 1). The same comparison can be performed at the superclass level, where it is evident that COMMUNICATION is the least well-represented SC superclass in Bacteria (Table 1).

The most common class in the ENERGY superclass is AMINO ACID BIOSYNTHESIS, containing many key

enzymes, usually with highly conserved role in cellular physiology (Table 1). Overall, the ENERGY super-class is shared widely with Bacteria.

The same seems to hold for the INFORMATION superclass, while at the same time, a clear difference between both REPLICATION and TRANSCRIPTION is observed. TRANSLATION-related proteins are common between SC and Bacteria (Fig. 1), with aminoacyl-tRNA synthases and elongation factors being most prominent.

Finally, COMMUNICATION-related proteins are most specific of Eukarya. This is especially clear for the CELL ENVELOPE/CELL WALL class, where less than 20% of the 156 assigned SC sequences have homologues in Bacteria. On the contrary, CELLULAR PROCESSES are more common in Bacteria. An interesting case is represented by heat-shock proteins with different distributions of their components. For example, GrpE is found in Bacteria, but not so far in Eukarya, while the hsc70-DnaK family is present in both domains.

From the analysis of the SC genome, two major conclusions can be drawn: (i) COMMUNICATION classes are rather distinct in Eukarya, while (ii) ENERGY and INFORMATION, with some variation, are shared between Bacteria and Eukarya.

### *Analysis of the HI Bacterial Genome*

The same analysis was carried out for the HI genome (Fleischmann et al. 1995) (Table 1 and Fig. 1). Both comparisons, SC proteins in Bacteria and HI in Eukarya provide two complementary views of the same problem, highlighting the differences and similarities between the two domains.

As in the parallel analysis of SC in Bacteria, classes related with ENERGY are well represented in both Bacteria and Eukarya, with more than 50% of the HI sequences having a eukaryotic homologue (Table 1). In particular, the AMINO ACID BIOSYNTHESIS was the best represented class between Bacteria and Eukarya.

The symmetry of the analysis of HI and SC is also evident in the INFORMATION classes (Table 1). In particular, TRANSLATION contains the largest number of HI sequences present in Eukarya (Fig. 1). It must be stressed that these are relative values, and the actual number of proteins in each one of the classes is usually much larger for SC (see Table 1).

In the COMMUNICATION superclass, HI shares very little with Eukarya, as found previously. In particular, CELL ENVELOPE/CELL WALL is the most extreme case, including bacterial-only proteins such as penicillin-binding proteins, glycosyl transferases, opacity proteins, and UDP-processing related proteins.

The general conclusion for the HI genome comparison is that metabolic functions are conserved in Bacteria and Eukarya, whereas there are many differences in the functions related with COMMUNICATION.

### *Analysis of MJ: Closer to Bacteria?*

The genome of MJ, as the first complete archaeal genome available (Bult et al. 1996), was compared with Eukarya and Bacteria. The amount of archaeal sequences is still too small to allow a systematic comparison against this domain (this is exactly the reason why HI and SC were not compared against Archaea).

The total number of ORFs in MJ is 1735, of which a relatively small percentage (606 sequences; 35%) was classified into 1 of the 13 classes (plus 20 classified in another category, not further considered here). Of these 606 sequences, 518 had detectable homologues in either the bacterial or the eukaryotic domains. In these 518 ORFs, the overall proportion of shared sequences was 160 (31%) with only Bacteria, 71 (14%) with only Eukarya, and 287 (55%) with both domains. Surprisingly, the amount of bacterial-only sequences is more than twice as many as the eukaryotic-only ones. This is evident from the functional class distribution with the other two domains (Fig. 2).

Classes related with ENERGY are common to the three domains. In particular, homologues of proteins involved in AMINO ACID BIOSYNTHESIS and NUCLEOTIDE BIOSYNTHESIS can be found in both Bacteria and Eukarya (Fig. 2). More specialized processes, such as BIOSYNTHESIS OF COFACTORS or TRANSPORT, are largely shared with Bacteria, while ENERGY METABOLISM appears to be rather specific to Archaea, with more than a third of the proteins exclusive in MJ, mostly associated with methane-based metabolic enzymes (Fig. 2).

The results on INFORMATION classes are more complex: first, both TRANSCRIPTION and TRANSLATION contain a large number of sequences shared with Eukarya only. These are mainly transcription and translation initiation factors. The TRANSCRIPTION class contains a large proportion of proteins common to MJ and Eukarya, for example, DNA-dependent RNA polymerases. Similarly, the TRANSLATION class shows the same pattern, with cases such as the GTP1/OBG family (Hudson and Young 1993), various ribosomal proteins and the MJ0829 family—similar to eukaryotic peptide chain release factors (Frolova et al. 1994), all characteristic of MJ and Eukarya. Second, in the REPLICATION class, an outstanding 30% of the proteins were related with Bacteria only, like MJ0124 (similar to type I restriction enzyme). This may have to do with the similarity of genome structure of Archaea and Bacteria, as proposed previously (Ouzounis and Kyrpides 1996).

In the COMMUNICATION superclass, only two classes (CELL ENVELOPE/CELL WALL and CELLULAR PROCESSES) contain a significant number of observations. In both of them, around 40% of the proteins have homologues in both Bacteria and Eukarya. The second largest fraction of sequences was found only in Bacteria. These include (a) several proteins related to the

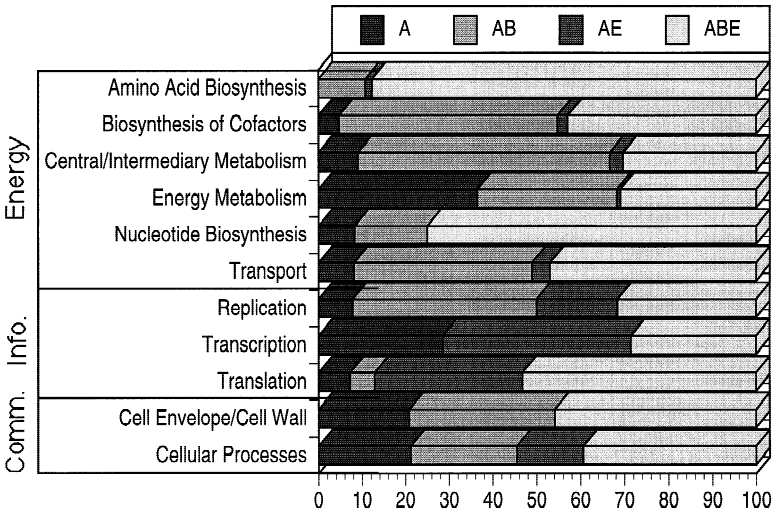


**Fig. 2.** Comparison of MJ proteins with Bacteria and Eukarya by functional class. The percentage of MJ sequences with at least one homologue in both Bacteria and Eukarya (ABE), Eukarya only (AE), Bacteria only (AB) and Archaea only (A). The results are represented for 11 classes, with 2 classes (Fatty Acid & Phospholipid Biosynthesis and Regulatory Functions) omitted, due to the low count of observations (Bult et al. 1996). Note that the MJ genome shares more components with Bacteria, with the exception of two classes, Transcription and Translation.

spore coat polysaccharide biosynthesis in the CELL ENVELOPE/CELL WALL class, not present in HI, yet found in *Bacillus subtilis* (Zhang et al. 1994), and (b) bacterial protein–export membrane proteins or cell division-related proteins as FtsZ (a distant member of eukaryotic tubulin) in the CELLULAR PROCESSES class. Eukaryotic-only proteins in the latter class can still be found, example cases being a CDC45-related protein (MJECL31) or alkyl hydroperoxide reductase (MJ0736).

The MJ genome is singular in that it shares many components with either of the other two domains in a very distinctive way, pointing to the peculiar nature of Archaea.

## Discussion

### *Are Genomes Good Representatives of the Phylogenetic Domains?*

Any completely sequenced genome can be considered as a possible model for the corresponding phylogenetic domain, but is this true for these first forerunners? It can be argued that SC has many particularities derived from its unique cell cycle, HI has a relatively small genome and is able to survive only within a host, and MJ has a unique methane-based metabolism. Despite their uniqueness, these genomes can be considered to contain a sample of proteins from their phylogenetic domains.

To minimize the possible problem of the lack of domain representation, we chose to perform the comparison of each one of the genomes (SC, HI, and MJ) with the full set of available bacterial and eukaryotic proteins. A meaningful comparison against the archaeal domain is still not possible. There are proteins in other archaeal species without homologues in MJ, e.g., gas vesicle proteins from several *Halobacterium* species, DnaK, rhodopsins, citrate synthase, superoxide dismutase, and others, but the extent of their distribution remains unknown.

### *What Are the Functional Differences Among the Three Domains?*

The answer to this question requires the analysis of the phylogenetic distribution of protein families and cellular functions. The sequencing of the first full genomes in each one of the three domains provides an opportunity for carrying out such an analysis with the assumption of having a near-complete range of cellular functions.

The complexity of the data, with thousands of proteins with different characteristics, and the intrinsic limitation of the biological knowledge available (protein function is in many cases inferred by sequence comparison and not by direct experimentation) require an appropriate abstraction level for global comparative analyses. We believe that the analysis of proteins grouped in few classes provides such a level of generalization.

### *Comparing Organisms by Functional Composition*

Early attempts have been reported on the comparison of taxa at the level of functional composition (Tamames et al. 1996). This comparative analysis gives a clear indication of an evolutionary trend toward the increase in the proportion of functions related with COMMUNICATION processes across the evolutionary scale, accompanied by a parallel decrease in the proportion of resources associated with ENERGY.

### *Comparing Organisms at the Level of Protein Families*

The alternative to the general comparison by functional composition is the detailed analysis of each protein based on sequence comparison. Each one of the recently completed genomes has been examined using different combinations of human expertise and automatic analysis (Riley 1993; Koonin et al. 1994; Bult et al. 1996; Koonin and Mushegian 1996; Mushegian and Koonin 1996; Tatusov et al. 1996; Bairoch and Apweiler 1997; Himmel-

reich et al. 1997). We have participated in this effort by analyzing systematically each genome with GeneQuiz (Casari et al. 1995, 1996a, b; Ouzounis et al. 1996; Andrade et al. 1997).

In particular, Koonin and collaborators have studied the definition of the minimal set of proteins common to the bacterial domain comparing HI and *Mycoplasma genitalium* (MG) (Mushegian and Koonin 1996) and later including the minimal core proteins found in SC (Koonin and Mushegian 1996). They have reported interesting examples of protein functions represented in the set of proteins common to MG, SC, and HI, even though the analysis is influenced by the small size and parasitic style of life of MG. This approach, however, has the strong limitation that with the inclusion of more genomes, the intersection between them is monotonically reduced (Kyrpides et al. 1999). A general picture requires the complete and systematic analysis of more representative genomes of the three domains.

Another similar approach, the first report recording the presence or absence of protein families in the three domains, used the PROSITE functional and phylogenetic classification (Ouzounis and Kyrpides 1996). In a subsequent report, the MJ genome as a representative of the Archaea has been classified into the three domains (Kyrpides et al. 1999).

### *Phylogenetic Distribution of Protein Families and Functional Classes*

Here, instead of focusing on the definition of a "minimal gene set" we describe the main differences among the three domains of life at a higher level of functional annotation.

Unique features of the present analysis are as follows:

(i) *Automation.* The automatic analysis is reproducible. More relaxed cutoffs in the similarity searches can lead to higher numbers of functional assignments and GeneQuiz (Casari et al. 1996b, Andrede et al, 1999) generates these automatically. However, problems remain about the validity of this type of twilight assignments.

(ii) *Domain analysis.* A full genome was used as a starting point for each of the comparisons (HI for Bacteria, MJ for Archaea, and SC for Eukarya) but the distribution of each functional class is assessed scanning all eukaryotic or bacterial sequences. This broadened reference set is necessary, to expand domain coverage.

(iii) *Functional classes.* An adequate degree of generalization by classifying proteins in a small number of functional classes allows comparative analyses of whole taxa. It is an attempt to describe global similarities and differences of various organisms.

This approach can be extended to any taxonomic level and number of genomes with additional complexity, where automation may be the only solution for increasingly refined genome comparisons and functional classifications.

### *Horizontal Gene Transfer*

A caveat to the present approach is the existence of horizontal gene transfer between genomes (Kidwell 1993). Our method cannot distinguish these cases from vertically occurring genes, present in the taxa compared since the time of their divergence. The effect is that the differences observed between two organisms are an underestimation of the real differences. Since the major conclusions are based on these differences, they are by definition conservative.

Another consideration is how large is the extent of horizontal gene transfer. This is, obviously, a difficult question to answer, and in any case, the answer strongly depends on the analyzed taxa. A recent analysis of *E. coli* (Lawrence and Ochman 1998), using a model of DNA composition amelioration (Lawrence and Ochman 1997), estimates that about 18% of the *E. coli* genome has been taken from other Bacteria since its original divergence from the *Salmonella* lineage 100 million years ago. It seems reasonable that, given the different ecological niches of Archaea and Bacteria, one would expect the transfer between these two taxa to be much lower.

### *Classes and the Three-Domain Split*

It is interesting that well-known differences in the three phylogenetic domains, in particular genome organization, gene expression, metabolism, and cell membrane (wall, envelope) are reflected in the distribution of the functional classes as defined herein. The most eminent conclusions follow.

(1) ENERGY-related proteins can be considered to be generally common in the three domains. Bacterial and archaeal proteins in this class are well represented in Eukarya.

(2) INFORMATION-related proteins are more domain specific, especially in HI and SC. MJ shares proteins with both domains, as previously noted (Ouzounis and Kyrpides 1996).

(3) COMMUNICATION-related functions are the defining factor for the three-domain split, with a large number of proteins unique to eukaryotic cellular structure.

The descriptive level of protein functional classes allows the analysis of the functional divergence among Bacteria, Eukarya, and Archaea, and the major differences signify the principal source of function diversification during cellular evolution.

## Note Added at Proof

Since the preparation of this paper a number of new genomes have been published, including a second Eukarya (see Science vol. 282, 1998) and a new generation of comparisons of complete genomes have generated an interesting controversy about the structure of the common ancestor and the relations between the different kingdoms [for a review see Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science* 284:2124–2129]. This work can be seen as a contribution to this discussion previous to the current flourishing of genomic information.

*Acknowledgments.* This work is part of the activities of the GeneQuiz consortium supported by Grant ERB 4061 PL 95-0315 under the European Union TMR program.

## References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Andrade MA, Casari G, Daruvar A, et al. (1997) Sequence analysis of the *Methanococcus jannaschii* genome and the prediction of protein function. Comput Appl Biosci 13:481–483

Andrade MA, Brown NP, Leroy C, et al. (1999) Automated genome analysis and annotation. Bioinformatics 15:391–412

Bairoch A, Apweiler R (1997) The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. Nucleic Acids Res 25:31–36

Bairoch A, Bucher P, Hofmann K (1997) The PROSITE database, its status in 1997. Nucleic Acids Res 25:217–221

Benson DA, Boguski MS, Lipman DJ, Ostell J (1997) GenBank. Nucleic Acids Res 25:1–6

Bult CJ, White OW, Olsen GJ, et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii.* Science 273:1058–1073

Casari G, Andrade MA, Bork P, et al. (1995) Challenging times for bioinformatics. Nature 376:647–648

Casari G, De Daruvar A, Sander C, Schneider R (1996a) Bioinformatics and the discovery of gene function. Trends Genet 12:244–245

Casari G, Ouzounis C, Valencia A, Sander C (1996b) GeneQuiz II: Automatic function assignment for genome sequence analysis. In: Proceedings of the First Annual Pacific Symposium on Biocomputing. World Scientific, Honolulu, pp 707–709

Fleischmann RD, Adams MD, White O, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fraser CM, Gocayne JD, White O, et al. (1995) The minimal gene complement of *Mycoplasma genitalium.* Science 270:397–403

Frolova L, Le Goff X, Rasmussen HH, et al. (1994) A highly conserved eukaryotic protein family possessing properties of polypeptide chain release factor. Nature 372:701–703

George DG, Hunt LT, Barker WC (1996) PIR-international protein sequence database. Methods Enzymol 266:41–59

Goffeau A, Barrell BG, Bussey H, et al. (1996) Life with 6000 genes. Science 274:546–567

Henikoff JG, Henikoff S (1996) Blocks database and its applications. Methods Enzymol 266:88–104

Himmelreich R, Plagens H, Hilbert H, Reiner B, Herrmann R (1997) Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium.* Nucleic Acids Res 25: 701–712

Hudson JD, Young PG (1993) Sequence of the *Schizosaccharomyces pombe* gtp1 gene and the identification of a novel family of putative GTP-binding proteins. Gene 125:191–193

Kidwell M (1993) Lateral transfer in natural populations of eukaryotes. Annu Rev Genet 27:235–256

Koonin EV, Mushegian AR (1996) Complete genome sequences of cellular life forms: Glimpses of theoretical evolutionary genomics. Curr Opin Genet Dev 6:757–762

Koonin EV, Bork P, Sander C (1994) Yeast chromosome III: New gene functions. EMBO J 13:493–503

Kyrpides N, Overbeek R, Ouzounis C (1999) Universal protein families and the functional content of the last universal common ancestor. J Mol Evol (in press)

Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: Rates of change and exchange. J Mol Evol 44:383–397

Lawrence JG, Ochman H (1998) Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci USA 95:9413–9417

Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 93:10268–10273

Ouzounis C, Kyrpides N (1996) The emergence of major cellular processes in evolution. FEBS Lett 390:119–123

Ouzounis C, Casari G, Valencia A, Sander C (1996) Novelties from the complete genome of *Mycoplasma genitalium.* Mol Microbiol 20: 897–899

Pearson WR (1996) Effective protein sequence comparison. Methods Enzymol 266:227–258

Riley M (1993) Function of the gene products in *Escherichia coli.* Microbiol Rev 57:862–952

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

Stoesser G, Sterk P, Tuli MA, Stoehr PJ, Cameron GN (1997) The EMBL nucleotide sequence database. Nucleic Acids Res 25:7–14

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

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

Tatusov RL, Mushegian AR, Bork P, et al. (1996) Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli.* Curr Biol 6:279–291

Zhang J, Ichikawa H, Halberg R, Kroos L, Aronson AI (1994) Regulation of the transcription of a cluster of *Bacillus subtilis* spore coat genes. J Mol Biol 240:405–415