J Mol Evol (1999) 49:413–423

JOURNAL OF MOLECULAR EVOLUTION

Universal Protein Families and the Functional Content of the Last Universal Common Ancestor

Nikos Kyrpides,[1,2] Ross Overbeek,[2] Christos Ouzounis[3]

[1] Department of Microbiology, University of Illinois at Urbana–Champaign, 407 South Goodwin Avenue, Urbana IL 61801, USA
[2] Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Case Avenue, Argonne IL 60439, USA
[3] Computational Genomics Group, Research Programme, The European Bioinformatics Institute, EMBL Cambridge Outstation, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK

Abstract. The phylogenetic distribution of *Methanococcus jannaschii* proteins can provide, for the first time, an estimate of the genome content of the last common ancestor of the three domains of life. Relying on annotation and comparison with reference to the species distribution of sequence similarities results in 324 proteins forming the universal family set. This set is very well characterized and relatively small and nonredundant, containing 301 biochemical functions, of which 246 are unique. This universal function set contains mostly genes coding for energy metabolism or information processing. It appears that the Last Universal Common Ancestor was an organism with metabolic networks and genetic machinery similar to those of extant unicellular organisms.

Key words: Last Universal Common Ancestor — Cellular evolution — Metabolic reconstruction — Biochemical pathways — Genomics

Introduction

With the completion of the sequencing of the first archaeal genome, that of *Methanococcus jannaschii* (Bult et al. 1996), it has been possible to describe the similarities of a complete set of archaeal proteins with the other two domains, Bacteria and Eukarya. This universal set of protein families present in all domains can then be used as an estimate for the genome content of the Last Universal Common Ancestor (Woese 1982; Woese and Fox 1977). Evidently, factors such as gene loss, horizontal transfer across domains and species peculiarities make this task difficult, especially when only interspecies comparisons are used (Becerra et al. 1997). Previous such approaches have ignored the above problems (Mushegian and Koonin 1996), resulting in descriptions that may be relevant only from a functional but not an evolutionary viewpoint, despite such claims (Koonin and Mushegian 1996).

Until recently, there has been no description and only a single prediction for the set of universal protein families, based on the presence or absence of sequence pattern sets in the three domains of life (Ouzounis and Kyrpides 1996b). However, this approach was limited primarily by two factors: first, sequence patterns cannot adequately describe an exact molecular function, but only protein families; and second, the archaeal families were clearly underrepresented (Ouzounis et al. 1995a), before the availability of the *M. jannaschii* genome. Although only 77 protein families were found to be universal, a much larger number of protein families was predicted to be present in Archaea, based on functional relationships such as metabolic pathways (Ouzounis and Kyrpides 1996b).

Herein, we describe for the first time a list of universal functions based on sequence comparison and detailed

Correspondence to: C. Ouzounis; e-mail: ouzounis@embl-ebi.ac.uk functional annotation. The current approach takes into account, but is not restricted to, complete genomes, thus providing a basis for the identification of the broadest possible number of functions present in representative species from all three domains of life.

Methods

All function assignments were derived after detailed family analysis of every single *M. jannaschii* ORF with continuing updates (Andrade et al. 1997; Kyrpides et al. 1996a). In addition, intradomain similarities were considered through the complete collection of species in public databases, thus eliminating some of the problems with pairwise interspecies comparisons encountered in similar studies (Tatusov et al. 1997). It must be emphasized that if a protein is present in any species for a given domain, it is irrelevant whether this protein may be absent from complete genomes, as is sometimes thought (Mushegian and Koonin, 1996). For this particular problem, what is needed is abundant sequence information, and not complete genome sequences (Ouzounis and Kyrpides 1996b). Manually derived results were compared with an automatic analysis obtained by the WIT system (<http://wit.mcs.anl.gov/WIT/>) (Overbeek et al. 1997). Metabolic information was obtained through the EMP database (Selkov et al. 1997a).

Each *M. jannaschii* ORF was used as a query in Blast2 similarity searches against the nonredundant protein sequence database at the National Center of Biotechnology Information (NCBI). The complete genome sequence of *Methanobacterium thermoautotrophicum* (Smith et al. 1997) was obtained from the Genome Therapeutics Corporation (<http://www.cric.com/>). The complete genome sequence of *Archaeoglobus fulgidus* (Klenk et al. 1997) was retrieved from TIGR (<http://www.tigr.org/>). These genome sequences were only used for comparisons within the archaeal domain. Database and bibliographic searches and function assignments were performed as described previously (Andrade et al. 1997; Kyrpides et al. 1996a). All functional annotations are available at <http://geta.life.uiuc.edu/~nikos/MJannotations.html> and mirrored at <http://www.ebi.ac.uk/research/cgg/annotation/MJannotations.html>. Updates of the phylogenetic distribution for all *M. jannaschii* genes are available at <http://geta.life.uiuc.edu/~nikos/Domain.Comparisons.html>.

Results

Continued Annotation

At the time of the original publication, the *M. jannaschii* genome was thought to be unique in that it contained only a few functionally characterized homologues, 38% of the total genome (Bult et al. 1996). One explanation was that this was the first completely sequenced archaeal genome and its uniqueness represented the peculiarity of the yet unexplored archaeal domain. However, with continued updates, the level of functional annotation has now surpassed 50% (Fig. 1). Previous claims that have raised this number to as high as 70% (Koonin et al. 1997) should be discarded due to a large number of false-positive identifications and overpredictions (Kyrpides and Ouzounis 1999a). Three independent analyses (Andrade et al. 1997; Bult et al. 1996; Kyrpides et al. 1996a) have concluded that the level of functional assignment for *M. jannaschii* is much lower. Approximately another quarter of the genome contains proteins with homologues of unknown function in the database and the remaining quarter of the genome contains unique sequences, with no homologues (even after the completion of two additional archaeal genomes). This is an important observation, since these levels of function and similarity relationships are now comparable to those in other model species (Ouzounis et al. 1996).

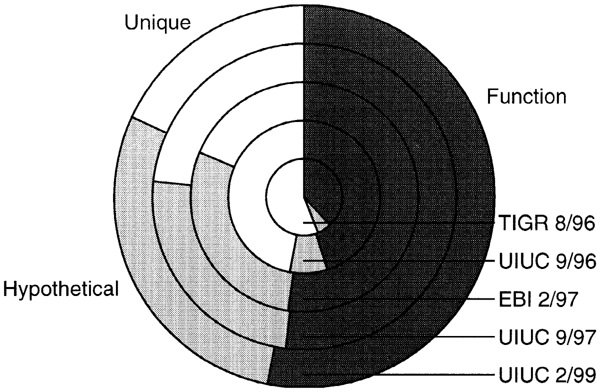

Fig. 1. Evolution of functional annotations for the *M. jannaschii* genome. “Function” signifies any functional assignment with a varying degree of accuracy, “hypothetical” denotes similarity to uncharacterized sequences, and “unique” represents unique sequences without any functional annotation. Within a year, the level of function assignment increased to 52%, while homologies to other proteins have covered another quarter of the total genome. Continuing annotation is available at <http://geta.life.uiuc.edu/~nikos/MJannotations.html>.

Phylogenetic Distribution

There are 324 proteins in *M. jannaschii,* with at least one homologue present in some species from both the other two domains, Bacteria and Eukarya, forming the universal protein family set (Table 1, Fig. 2). Of those, only 23 are hypothetical (families without functional annotation). The universal set of proteins contains metabolic enzymes, transporters, various ATP/GTP-binding proteins, protein tyrosine phosphatases (Stravopodis and Kyrpides 1999), ribosomal proteins, aminoacyl-tRNA synthetases, translation initiation factors (Kyrpides and Woese 1998), helicases, and RNA polymerase subunits. Most of these proteins were previously predicted using family patterns (Ouzounis and Kyrpides 1996b). About a quarter of these proteins are paralogues within *M. jannaschii,* therefore limiting the number of predicted universal functions to 246 (Table 2). Structural RNA (rRNA and tRNA) genes can also be considered to belong to this universal function set but are not further discussed in this context.

Another 522 proteins have at least one homologue in the bacterial domain only, of which 132 are hypothetical (Table 1, Fig. 2). This “uneukaryotic” set (Ouzounis and Kyrpides 1996b) of proteins contains electron-transport

Table 1. Domain and class distribution of the *M. jannaschii* proteins

	Domains/classes				
	Energy	Information	Communication	Hypothetical	Total (domain)
Universal (ABE)	178	103	20	23	324
Uneukaryotic (AB)	264	94	32	132	522
Unbacterial (AE)	10	79	7	27	123
Archaeal (A)	80	34	11	662	787
Total (class)	532	310	70	844	1756

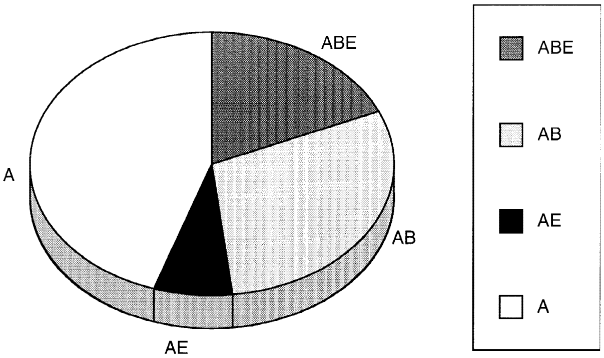

Fig. 2. Phylogenetic distribution of the *M. jannaschii* proteins. Universal proteins (ABE) are 324 (18%), uneukaryotic (AB) are 522 (30%), unbacterial (AE) are 123 (7%), and archaeal-only (A) are 787 (the remaining 45%). It should be emphasized that these percentages contain all proteins that have a homologue in the database, according to the species distribution, without any consideration of whether they have a predicted function: in other words, both "function" and "hypothetical" categories are included, with the exception of the archaeal-only section, which also includes the "unique" sequences. Continuing updates are available at <http://geta.life.uiuc.edu/~nikos/Domain.Comparisons.html>.

systems such as F420-reducing hydrogenase subunits and small electron-transport proteins, nitrogen fixation factors, cobalt- or tungsten-binding proteins, cofactor biosynthesis enzymes, bacterial-type transporter systems, bacteriochlorophyll synthases, cell wall components, bacterial-type transcriptional regulators (Kyrpides and Ouzounis, 1999b), modification methylases, and restriction enzymes.

A mere 123 proteins have at least one homologue in the eukaryotic domain only, of which 27 are hypothetical (Table 1, Fig. 2). Members of this "unbacterial" set are proteins such as eukaryotic-type ribosomal proteins, cell division control proteins, some oxidoreductases, translation initiation factors, core histone fold-containing proteins (Ouzounis and Kyrpides 1996a), general transcription initiation factors, fibrillarin-like pre-rRNA processing proteins and RNA maturases, signal recognition particle proteins, and proteasome components.

Finally, there are 787 proteins that remain uniquely archaeal at present, with only 125 characterized (Table 1, Fig. 2). Some of these proteins include various subunits of the methanogenesis systems, hydrogenases, desulfoferredoxins, ATPases, flagellins, methyltransferases, and endonucleases.

There are two unexpected observations from this analysis: first, the universal set of proteins is relatively small, but very well characterized (Tables 1 and 2); second, Archaea, as represented by *M. jannaschii,* seem to contain four times more bacterial-type than eukaryotic-type proteins. This final point comes in sharp contrast to previous beliefs (Keeling et al. 1994). It remains to be seen whether the sequencing of primitive eukaryotic genomes will change this pattern.

Functional Classification

To obtain a more complete picture of the nature of the proteins shared with the other two domains, we have classified all characterized proteins from *M. jannaschii* into three functional superclasses (Table 1, Fig. 3) (Ouzounis et al. 1996; Tamames et al. 1996). These superclasses of Energy-, Information-, and Communication-related proteins reflect the involvement of the corresponding proteins in small-molecule, nucleic acid, and protein–protein interactions, respectively (Ouzounis et al. 1995b).

The Energy superclass is composed of universal proteins and a dominant fraction of bacterial proteins, as predicted (Ouzounis and Kyrpides 1996b). Only 10 proteins in this class are shared between Archaea and Eukarya exclusively (Table 1, Fig. 3). The results are in agreement with the metabolic reconstruction of *M. jannaschii* (Selkov et al. 1997b), despite the presumed absence of certain enzymes from this species.

In the Information superclass, the *M. jannaschii* proteins are almost equally distributed as universal, uneukaryotic, and unbacterial (Ouzounis and Kyrpides 1996b), underlining the sharing of components with the eukaryotic information-processing machinery (Olsen and Woese 1997; Ouzounis and Kyrpides 1996c). Yet the extent of common elements of archaeal and bacterial information-processing systems has also been extensively documented (Kyrpides and Ouzounis, 1995, 1997; Kyrpides et al. 1996b). Finally, Communication-related proteins do not display any discernible patterns (Table 1, Fig. 3).

The above numbers are expected to vary when the complete repertoire of protein functions for *M. jannaschii* becomes available. The assumption in all similar

Table 2. The 246 functions with a universal distribution, derived from the 301 universal proteins of known function: this set can be used as an estimate for the functional content of the last common ancestor

Function	EC No.
Amino acid biosynthesis	
Aromatic amino acid family	
3-Dehydroquinate dehydratase	EC 4.2.1.10
5-Enolpyruvylshikimate 3-phosphate synthase	EC 2.5.1.19
Anthranilate synthase I	EC 4.1.3.27
Anthranilate synthase II″	EC 4.1.3.27
Anthranilate synthase II′	EC 2.4.2.18
Chorismate synthase	EC 4.6.1.4
Chorismate mutase	EC 5.4.99.5
Indole-3-glycerol phosphate synthase	EC 4.1.1.48
N-Phosphoribosyl anthranilate isomerase	EC 5.3.1.24
Prephenate dehydratase	EC 4.2.1.51
Shikimate 5-dehydrogenase	EC 1.1.1.25
Tryptophan synthase, subunit α	EC 4.2.1.20
Tryptophan synthase, subunit β	EC 4.2.1.20
Aspartate family	
Asparagine synthetase	EC 6.3.5.4
Aspartate-semialdehyde dehydrogenase	EC 1.2.1.11
Aspartokinase I	EC 2.7.2.4
5-Methyltetrahydrofolate—homodysteine *S*-methyltransferase	EC 2.1.1.13
3-Isopropylmalate dehydratase	EC 4.2.1.33
Dihydrodipicolinate synthase	EC 4.2.1.52
Homoserine dehydrogenase (HDH)	EC 1.1.1.3
Homoserine kinase (HK)	EC 2.7.1.39
L-Asparaginase	EC 3.5.1.1
Threonine synthase	EC 4.2.99.2
Glutamate family	
Acetylglutamate kinase	EC 2.7.2.8
Argininosuccinate lyase	EC 4.3.2.1
Argininosuccinate synthase	EC 6.3.4.5
Glutamate synthase (NADPH), subunit α	EC 1.4.1.13
Glutamine synthetase	EC 6.3.1.2
Glutamate decarboxylase	EC 4.1.1.15
N-Acetyl-γ-glutamyl-phosphate reductase	EC 1.2.1.38
N-Acetylornithine aminotransferase	EC 2.6.1.69
Ornithine carbamoyltransferase subunit F	EC 2.1.3.3
Pyruvate family	
3-Isopropylmalate dehydratase	EC 4.2.1.33
Acetolactate synthase, large subunit	EC 4.1.3.18
Acetolactate synthase, small subunit	EC 4.1.3.18
Branched-chain amino acid aminotransferase	EC 2.6.1.42
Dihydroxy-acid dehydratase	EC 4.2.1.9
2-Isopropylmalate synthase	EC 4.1.3.12
Ketol-acid reductoisomerase	EC 1.1.1.86
Serine family	
Glycine hydroxymethyltransferase	EC 2.1.2.1
Phosphoglycerate dehydrogenase	EC 1.1.1.95
Phosphoserine phosphatase	EC 3.1.3.3
Aspartate transaminase	EC 2.6.1.1
Histidine family	
ATP phosphoribosyltransferase	EC 2.4.2.17
Histidinol dehydrogenase	EC 1.1.1.23
Histidinol-phosphate aminotransferase (hisH)	EC 2.6.1.9
Imidazoleglycerol-phosphate dehydrogenase	EC 4.2.1.19
Amidotransferase	EC 2.4.2.—
Imidazoleglycerol-phosphate synthase	
Phosphoribosyl-AMP cyclohydrolase	EC 3.5.4.19
Biosynthesis of cofactors, prosthetic groups, and carriers	
Glutamate-1-semialdehyde-2,1-aminomutase	EC 5.4.3.8
Porphobilinogen deaminase	EC 4.3.1.8
Quinolinate phosphoribosyltransferase	EC 2.4.2.19
S-Adenosylhomocysteine hydrolase	EC 3.3.1.1

Table 2. (*Continued*)

Function	EC No.
Biotin	
Adenosylmethionine-8-amino-7-oxononanoate aminotransferase	EC 2.6.1.62
Biotin—acetyl-CoA carboxylase synthetase	EC 6.3.4.15
Biotin synthetase	EC 2.8.1.—
Heme and porphyrin	
Uroporphyrin-III C-methyltransferase	EC 2.1.1.107
Porphobilinogen synthase	EC 4.2.1.24
Molybdopterin	
Molybdenum cofactor biosynthesis prtein (moaB)	
Molybdenum cofactor biosynthesis protein (moaC)	
Molybdenum cofactor biosynthesis protein (moeA)	
Thioredoxin, glutaredoxin, and glutathione	
Thioredoxin reductase	EC 1.6.4.5
Thiamine	
Thiamine biosynthesis protein NMT2	
Cell envelope	
Membranes, lipoproteins, and porins	
Dolichyl-phosphate mannose synthase	EC 2.4.1.83
Murein sacculus and peptidoglycan	
Amidophosphoribosyltransferase precursor	EC 2.4.2.14
Surface polysaccharides, lipopolysaccharides, and antigens	
UDP-glucose 4-epimerase	EC 5.1.3.2
UDP-glucose dehydrogenase	EC 1.1.1.22
UDP-*N*-acetylglucosamine-dolichyl-phosphate	EC 2.7.8.15
Cellular processes	
Cell division	
Cell division control protein CDC48	
Cell division protein (ftsJ)	
Chromosome segregation protein, SMC	
Chaperones	
Thermosome, chaperonin	
Detoxification	
Alkyl hydroperoxide reductase	
Protein and peptide secretion	
Signal recognition particle, 54 kDa	
Preprotein translocase SECY	
SecE/Sec61, γ subunit	
Central intermediary metabolism	
Amino sugars	
Glutamine-fructose-6-phosphate transaminase	EC 2.6.1.16
Degradation of polysaccharides	
Glucan 1,4-α-glucosidase	EC 3.2.1.3
Other	
2-Hydroxyhepta-2,4-diene-1,7-dioate isomerase	EC 4.1.1.68
Agmatine ureohydrolase	EC 3.5.3.11
Phosphorus compounds	
N-Methylhydantoinase	
Polyamine biosynthesis	
Acetylpolyamine aminohydolase	EC 3.5.1.—
Spermidin synthase	EC 2.5.1.16
Nitrogen metabolism	
ADP-ribosylglycohydrolase	EC 3.2.—.—
Eneregy metabolism	
Adenylate kinase	EC 2.7.4.3
ATP-proton motive force interconversion	
ATP synthase, subunit A	EC 3.6.1.34
ATP synthase, subunit B	EC 3.6.1.34
ATP synthase, subunit D	EC 3.6.1.34
ATP synthase, subunit I	EC 3.6.1.34
Glycogen metabolism	
Glycogen phosphorylase	EC 2.4.1.1
Gluconeogenesis	
Alanine aminotransferase 2	EC 2.6.1.2

Table 2. (*Continued*)

Function	EC No.
Glycolysis	
Phosphoglycerate kinase	EC 2.7.2.3
Enolase	EC 4.2.1.11
Glucose-6-phosphate isomerase B (GPI B)	EC 5.3.1.9
Glyceraldehyde 3-phosphate dehydrogenase	EC 1.2.1.12
L-Lactate dehydrogenase	EC 1.1.1.27
Pyruvate kinase	EC 2.7.1.40
Triosephosphate isomerase	EC 5.3.1.1
Pentose phosphate pathway	
Pentose-5-phosphate-3-epimerase	
Ribose 5-phosphate isomerase A	EC 5.3.1.6
Transaldolase	EC 2.2.1.2
Transketolase, subunit A	EC 2.2.1.1
Transketolase, subunit B	EC 2.2.1.1
TCA cycle	
Aconitase	EC 4.2.1.3
Isocitrate dehydrogenase	EC 1.1.1.42
Succinate dehydrogenase, flavoprotein subunit	EC 1.3.99.1
Succinyl-CoA synthetase, α subunit	EC 6.2.1.5
Succinyl-CoA synthetase, β subunit	EC 6.2.1.5
Fatty acid and phospholipid metabolism	
Bifunctional short-chain isoprenyl diphosphate synthase	EC 2.5.1.1
Biotin carboxylase	EC 6.3.4.14
CDP-diacylglycerol-serine *O*-phosphatidyltransferase	EC 2.7.8.8
Acetyl-CoA synthase	EC 6.2.1.1
Purines, pyrimidines, nucleosides, and nucleotides	
2′-Deoxyribonucleotide metabolism	
Glycinamide ribonucleotide synthetase	EC 6.3.4.13
Purine ribonucleotide biosynthesis	
Adenylosuccinate lyase	EC 4.3.2.2
Adenylosuccinate synthetase	EC 6.3.4.4
GMP synthetase	EC 6.3.5.2
Inosine-5′-monophosphate dehydrogenase (IMP)	EC 1.1.1.205
Nucleoside diphosphate kinase	EC 2.7.4.6
Phosphoribosylaminoimidazole carboxylase	EC 4.1.1.21
Phosphoribosylaminoimidazolesuccinocarboxamide synthase	EC 6.3.2.6
Phosphoribosylformylglycinamidine cycloligase	EC 6.3.3.1
Phosphoribosylformylglycinamidine synthase subunit I	EC 6.3.5.3
Phosphoribosylformylglycinamidine synthase II	EC 6.3.5.3
Phosphoribosylglycinamide formyltransferase 2	EC 2.1.2.2
Ribose-phosphate pyrophosphokinase	EC 2.7.6.1
Exopolyphosphatase	EC 3.6.1.11
Pyrimidine ribonucleotide biosynthesis	
Aspartate carbamoyltransferase catalytic subunit	EC 2.1.3.2
Carbamoyl-phosphate synthase, large subunit	EC 6.3.5.5
Carbamoyl-phosphate synthase, small subunit	EC 6.3.5.5
CTP synthase	EC 6.3.4.2
Dihydroorotase	EC 3.5.2.3
Dihydroorotase dehydrogenase	EC 1.3.3.1
Thymidylate kinase	EC 2.7.4.9
Orotidine 5′-monophosphate decarboxylase	EC 4.1.1.23
Uridine 5′-monophosphate synthase	EC 2.4.2.10
Salvage of nucleosides and nucleotides	
Adenine phosphoribosyltransferase	EC 2.4.2.7
Methylthioadenosine phosphorylase	EC 2.4.2.28
Thymidine phosphorylase	EC 2.4.2.4
Sugar-nucleotide biosynthesis and conversions	
Glucose-1-phosphate thymidylyltransferase	EC 2.7.7.24
Replication	
Degradation of DNA	
Endonuclease III	EC 4.2.99.18
DNA replication, restriction, modification, recombination, and repair	
Dimethyladenosine transferase	EC 2.1.1.—

Table 2. (*Continued*)

Function	EC No.
DNA repair protein RAD$_{51}$	
DNA topoisomerase I	EC 5.99.1.2
Methylated DNA protein cysteine methyltransferase	EC 2.1.1.63
Proliferating-cell nucleolar antigen	
Similar to ribonuclease HII (rnhB)	
Replication factor C	
Cell division inhibitor (minD)	
Transcription	
DNA-dependent RNA polymerases	
DNA-dependent RNA polymerase, subunit A′	EC 2.7.7.6
DNA-dependent RNA polymerase, subunit A″	EC 2.7.7.6
DNA-dependent RNA polymerase, subunit B′	EC 2.7.7.6
DNA-directed RNA polymerase, subunit B″	EC 2.7.7.6
Translation	
PET$_{112}$ protein	
Amino acyl tRNA synthetases	
Tyrosyl-tRNA synthetase	EC 6.1.1.1
Tryptophanyl-tRNA synthetase	EC 6.1.1.2
Threonyl-tRNA synthetase	EC 6.1.1.3
Leucyl-tRNA synthetase	EC 6.1.1.4
Isoleucyl-tRNA synthetase	EC 6.1.1.5
Alanyl-tRNA synthetase	EC 6.1.1.7
Valyl-tRNA synthetase	EC 6.1.1.9
Methionyl-tRNA synthetase	EC 6.1.1.10
Seryl-tRNA synthetase	EC 6.1.1.11
Aspartyl-tRNA synthetase	EC 6.1.1.12
Glycyl-tRNA synthetase	EC 6.1.1.14
Prolyl-tRNA synthetase	EC 6.1.1.15
Glutamyl-tRNA synthetase	EC 6.1.1.17
Arginyl-tRNA synthetase	EC 6.1.1.19
Phenylalanyl-tRNA synthetase, subunit alpha	EC 6.1.1.20
Phenylalanyl-tRNA synthetase, subunit beta	EC 6.1.1.20
Histidyl-tRNA synthetase	EC 6.1.1.21
Degradation of proteins, peptides, and glycopeptides	
ATP-dependent protease La	
O-sialoglycoprotein endopeptidase	EC 3.4.24.5
xaa-pro dipeptidase	
Protease I	
ATP-dependent 26S protease regulatory subunit 4	
Protein modification	
L-Isoaspartyl protein carboxyl methyltransferase	EC 2.1.1.77
Methionine aminopeptidase	EC 3.4.11.18
Acetyltransferase complex, subunit ARD1	EC 2.3.1.—
Selenium donor protein	EC 2.7.9.3
Ribosomal proteins: synthesis and modification	
SSU ribosomal protein S2P	
SSU ribosomal protein S3P	
SSU ribosomal protein S4E	
SSU ribosomal protein S4P	
SSU ribosomal protein S5P	
SSU ribosomal protein S7P	
SSU ribosomal protein S8P	
SSU ribosomal protein S10P	
SSU ribosomal protein S11P	
SSU ribosomal protein S12P	
SSU ribosomal protein S13P	
SSU ribosomal protein S14P	
SSU ribosomal protein S15P	
SSU ribosomal protein S17P	
SSU ribosomal protein S18P	
SSU ribosomal protein S19P	
LSU ribosomal protein L1P	
LSU ribosomal protein L2P	

Table 2. (*Continued*)

Function	EC No.
LSU ribosomal protein L3P	
LSU ribosomal protein L4P	
LSU ribosomal protein L5P	
LSU ribosomal protein L6P	
LSU ribosomal protein L11P	
LSU ribosomal protein L13P	
LSU ribosomal protein L14P	
LSU ribosomal protein L18P	
LSU ribosomal protein L22P	
LSU ribosomal protein L23P	
LSU ribosomal protein L24P	
LSU ribosomal protein L29P	
LSU ribosomal protein L44P	
tRNA modification	
tRNA-pseudouridine synthase I	EC 5.4.99.12
Glutamyl-tRNA (Gln) amidotransferase subunit A	EC 6.3.5.—
Translation factors	
SUI1 family of translation factors	
Translation initiation factor, eIF-2B α subunit	
Translation initiation factor IF-2	
ATP-dependent RNA helicase, eIF-4A family	
Translation elongation factor, EF-2	
Transport and binding proteins	
ABC transporter ATP-binding protein	
Na$^+$/Ca$^+$ exchanger protein	
Amino acids, peptides, and amines	
Ammonium transporter	
Cationic amino acid transporter MCAT-2	
Carbohydrates, organic alcohols, and acids	
Malic acid transport protein	
Sodium-dependent noradrenaline transporter	
Cations	
Oxaloacetate decarboxylase, subunit α	EC 4.1.1.3
Potassium channel protein	
Other	
Arsenical pump-driving ATPase	
H$^+$-transporting ATPase	EC 3.6.1.35
Chloride channel protein	
Miscellaneous/unclassified	
Sodium-dependent phosphate transporter	
Galactoside acetyltransferase	EC 2.3.1.18
HIT protein, member of the HIT family	
Large helicase-related protein, LHR	
Acylphosphatase	EC 3.6.1.7
HAM$_1$ protein	
Polynucleotide kinase	EC 2.7.1.78
Inositol monophosphatase family	EC 3.1.3.25
Glycosulfatase	
Atrazine chlorohydrolase	
DNA/RNA helicase (Hfm1p)	
RNase L inhibitor	
SUA5 protein family	
Cleavage and polyadenylation specificity factor (CPSF)	
Phosphoadensoine phosphosulfate reductase	EC 1.8.99.4
Dolichyl-phosphate β-glucosyltransferase	EC 2.4.1.117
RNA 3′-terminal phosphate cyclase	EC 6.5.1.4

analyses is that the present set of more than 50% function assignments forms a representative sample of the total range of functions for a given genome, thus providing us with insights about the distribution of functional classes across the three domains of life (Overbeek et al. 1997). The hypothetical ORFs, proteins with no functional annotation, are largely archaeal only, with the next sizable component being uneukaryotic (shared with bacteria

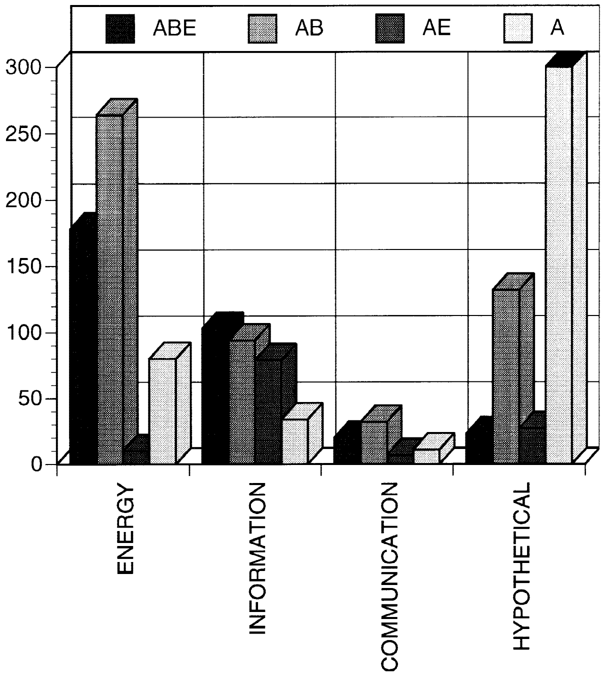

Fig. 3. Phylogenetic distribution of the *M. jannaschii* proteins into three functional superclasses (Ouzounis et al. 1995b). Exact numbers are given in Table 1. The two most outstanding features of this classification are that, first, the Energy class is dominated by uneukaryotic proteins (AB) followed by universal ones (ABE) and, second, the Information class is well represented in the other two domains, reflecting the presence of common elements in archaeal and eukaryotic transcription/translation machineries. For clarity, only 300 (of 662) hypothetical archaeal proteins are shown (data are clipped, marked by a *black top face*).

only) (Table 1, Fig. 3). This may reflect the abundance of complete sequence data generated by the many ongoing small-genome projects.

Universal Functions

The 301 universal proteins with some functional annotation can be clustered into 246 biochemical functions (Table 2). These include metabolic enzymes (amino acid interconversion and biosynthesis, nucleotide biosynthesis, electron transfer, energy transformations including carbohydrate catabolism), small-molecule transporters, transcription-related proteins (including RNA polymerase subunits), translation-related proteins (including ribosomal proteins and aminoacyl-tRNA synthetases), and proteins involved in protein modification and degradation. There is only a limited number of functions related to cell division, cell cycle, and intracellular signaling.

It is interesting to note that although most of these functions were predicted (Ouzounis and Kyrpides 1996b), it is now becoming clear that the universal set of functions is composed mostly of metabolic enzymes, transporters, and information processing elements. For instance, the lack of universal transcription factors or intra- and intercellular communication proteins is striking. In more general terms, "structural" components (metabolic enzymes, transporters, parts of translation) of the cellular biochemistry are present, but "regulatory" components (replication components, transcription factors, cell division factors) present in all domains are sparse or absent.

Discussion

> As a distinct primary kingdom, archae[bacteri]a are important in their own right. Yet I feel they are also important in a broader context. They serve to give us a badly needed perspective on early events in evolution of cells. Because of archae[bacteri]a, we will come to understand better the universal ancestor and will develop a new and better concept of eukaryotic origins. (Woese 1982)

It is evident that only now, in the genome era, can we better appreciate the above ideas and reach a deeper understanding of the problem of the nature of the Last Universal Common Ancestor. Having at least one complete genome from each domain of life, a totally new picture is emerging regarding the problem of the universal families and the genomic content of the common ancestor of all life forms (Forterre et al. 1992).

The current "intersection" [or backtrack (Becerra et al. 1997)] approach suffers from some well-understood limitations but nevertheless provides a lower estimate of the universal set of functions. An underlying assumption in this approach is that the branching order of the domains still remains unknown. If, for example, as is now becoming evident (Olsen and Woese 1997), Archaea and Eukarya are considered sister groups (Becerra et al. 1997), then the universal set should also include "uneukaryotic" proteins, shared exclusively between the two most primitive domains, Archaea and Bacteria. Yet, even with this limitation in mind, the universal set can be a basis for further characterizations of the Last Universal Common Ancestor.

Given that the functional intersection of the three domains, as presented herein, is based only on one species which certainly does not embrace the full diversity of Archaea, the universal function set is expected to grow as more complete genomes become available (Klenk et al. 1997; Smith et al. 1997). There exist, for instance, genes in *A. fulgidus* and *M. thermoautotrophicum* that are present in the other two domains but are absent from *M. jannaschii* (data not shown).

There are three major conclusions drawn from the analysis of the universal functional set. First, Archaea do not manifest a chimeric nature, a term suggestive of a derived instead of an ancestral form, as previously proposed (Koonin et al. 1997); they seem rather to be an ancient life form that gave rise to Eukarya (Olsen and Woese, 1997). As we have pointed out previously, it is the Eukarya that contain an archaeal-like basic transcription machinery, rather than the other way around (Ouzounis and Kyrpides 1996c). Second, Archaea seem to share a larger fraction of their genome with Bacteria, rather than Eukarya (Fig. 2), not necessarily suggesting a closer phylogenetic relationship of these two domains, at present. Third, the nature of the Last Universal Common. Ancestor is now revealed to be even more advanced and complex than previously believed. From the present analysis, it seems that it contained metabolic enzymes and genetic systems similar to those of extant unicellular organisms.

Was the Last Universal Common Ancestor a cellular entity? It appears that it was possibly a complex organism, with most "structural" components of metabolic pathways and some genetic information processing in place, but without "regulatory" elements such as replication, cell division, and intracellular regulation. It has been previously proposed that the Last Universal Common Ancestor may have been a rudimentary cell, called a "progenote," possibly without a full genetic information processing machinery (Woese, 1970). This argument is based on the fact that molecules participating in the contemporary translation machinery are so fundamentally different between Bacteria and Eukarya (Kyrpides and Woese 1998) that no refined translation process should have existed at the time of the domain split.

Although Archaea have bridged much of the differences between the other two domains, parts of information processing systems, including translation, transcription, replication, and regulation, are not universally conserved. There are two alternative, not mutually exclusive, explanations: there may have been massive replacements of these systems during early evolution, or information processing was never present in the Last Universal Common Ancestor, according to the "progenote" hypothesis (Woese 1970; Woese and Fox 1977). We have previously proposed a plausible scenario (Ouzounis and Kyrpides 1996b), which partially explains the differences in genome organization and transcription during cellular evolution. According to that hypothesis, the Last Common Ancestor most probably had molecular components of basic metabolism very similar to the contemporary ones, while having an archaeal-like transcription (Ouzounis and Kyrpides, 1996b). Overall, Archaea seem to be the most ancient forms of life on earth, and closer to the Last Universal Common Ancestor, while at the same time being the predecessors of Eukarya, providing us with invaluable perspectives on the nature of cellular life.

Acknowledgments. We thank Prof. C.R. Woese for stimulating discussions and ideas. We are also grateful to Prof. Patrick Forterre, Prof. Antonio Lazcano, and "La Fondation des Treilles" for their hospitality at the meeting on the "Last Universal Common Ancestor" that took place in Les Treilles, South France, 18–24 June 1996. This work was partly supported by NASA (NAG 54500), U.S. Department of Energy grants to C.R. Woese (DEFG C02-95ER61963) and R.O. (Contract W-31-109-Eng-38), and the European Molecular Biology Laboratory.

References

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

Becerra A, Islas S, Leguina JI, Silva E, Lazcano A (1997) Polyphyletic gene losses can bias backtrack characterizations of the Cenancestor. J Mol Evol 45:115–117

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

Forterre P, Benachenhou-Lahfa N, Confalonieri F, Duguet M, Elie C, Labedan B (1992) The nature of the last universal ancestor and the root of the tree of life, still open questions. BioSystems 28:15–32

Keeling PJ, Charlebois RL, Doolittle WF (1994) Archaebacterial genomes: Eubacterial form and eukaryotic content. Curr Opin Genet Dev 4:816–822

Klenk H-P, Clayton RA, Tomb J-F, et al. (1997) The complete genome sequence of the hyperthermophilic, sulfate-reducing archaeon *Archaeoglobus fulgidus.* Nature 390:364–370

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, Mushegian AR, Galperin MY, Walker DR (1997) Comparison of archaeal and bacterial genomes: Computer analysis of protein sequences predicts novel functions and suggests a chimeric origin of the archaea. Mol Microbiol 25:619–637

Kyrpides NC, Ouzounis CA (1995) The eubacterial transcriptional activator *Lrp* is present in the Archaeon *Pyrococcus furiosus.* Trends Biochem Sci 20:140–141

Kyrpides NC, Ouzounis CA (1997) Bacterial σ^{70} transcription factor DNA-binding domains in the Archaeon *Methanococcus jannaschii.* J Mol Evol 45:706–707

Kyrpides NC, Ouzounis CA (1999a) Whole-genome sequence annotation: "Going wrong with confidence." Mol Microbiol 32:886–887

Kyrpides NC, Ouzounis CA (1999b) Transcription in Archaea. Proc Natl Acad Sci USA (in press)

Kyrpides NC, Woese CR (1998) Universally conserved translation initiation factors. Proc Natl Acad Sci USA 95:224–228

Kyrpides NC, Olsen GJ, Klenk H-P, White O, Woese CR (1996a) *Methanococcus jannaschii* genome: Revisited. Microbial Comp Genom 1:329–338

Kyrpides NC, Woese CR, Ouzounis CA (1996b) KOW: A novel motif linking a bacterial transcription factor with ribosomal proteins. Trends Biochem Sci 21:425–426

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

Olsen GJ, Woese CR (1997). Archaeal genomics: An overview. Cell 89:991–994

Ouzounis CA, Kyrpides NC (1996a) The core histone fold: Limits to functional versatility. J Mol Evol 43:541–542

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

Ouzounis CA, Kyrpides NC (1996c) Parallel origins of the nucleosome core and eukaryotic transcription from Archaea. J Mol Evol 42: 234–239

Ouzounis C, Kyrpides N, Sander C (1995a) Novel protein families in Archaean genomes. Nucleic Acids Res 23:565–570

Ouzounis C, Valencia A, Tamames J, Bork P, Sander C (1995b) The functional composition of living machines as a design principle for artificial organisms. In: Morán F, Moreno A, Merelo JJ, Chacón P (eds) 3rd European Conference on Artificial Life (ECAL95). Springer, Granada, Spain, pp 843–851

Ouzounis C, Casari G, Sander C, Tamames J, Valencia A (1996) Computational comparisons of model genomes. Trends Biotechnol 14: 280:285

Overbeek R, Larsen N, Smith W, Maltsev N, Selkov E (1997) Representation of function: The next step. Gene 191:GC1–GC9

Selkov E, Galimova M, Goryanin I, et al. (1997a) The metabolic pathway collection: an update. Nucleic Acids Res 25:37–38

Selkov E, Maltsev N, Olsen GJ, Overbeek R, Whitman WB (1997b) A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data. Gene 197:GC11–GC26

Smith DR, Doucette-Stamm LA, Deloughery C, et al. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: Functional analysis and comparative genomics. J Bacteriol 179: 7135–7155

Stravopodis DJ, Kyrpides NC (1999) Identification of protein tyrosine phosphatases in Archaea. J Mol Evol 48:625–627

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

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

Woese CR (1970) The genetic code in prokaryotes and eukaryotes. In: Charles HP, Knight BCJG (eds) Organization and control in prokaryotic and eukaryotic cells. Cambridge University Press, Cambridge, pp 39–54

Woese CR (1982) Archaebacteria and cellular origins: An overview. Zentrabl Bakt Hyg I Abt Orig C 3:1–17

Woese CR, Fox GE (1977) The concept of cellular evolution. J Mol Evol 10:1–6