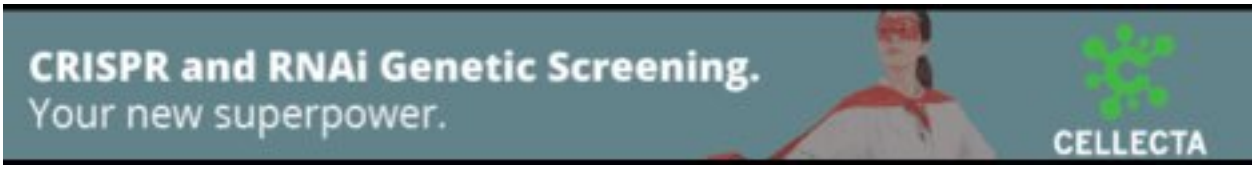
Comparative Genomics of the Archaea (Euryarchaeota): Evolution of Conserved Protein Families, the Stable Core, and the Variable Shell

Kira S. Makarova, L. Aravind, Michael Y. Galperin, et al.

Genome Res. 1999 9: 608-628
Access the most recent version at doi:10.1101/gr.9.7.608



Comparative Genomics of the Archaea (Euryarchaeota): Evolution of Conserved Protein Families, the Stable Core, and the Variable Shell

Kira S. Makarova,[1,2,4] L. Aravind,[1,3] Michael Y. Galperin,[1] Nick V. Grishin,[1] Roman L. Tatusov,[1] Yuri I. Wolf,[1,4] and Eugene V. Koonin[1,5]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894 USA; [2]Department of Pathology, F.E. Hebert School of Medicine, Uniformed Services University of the Health Sciences, Bethesda, Maryland 20814-4799 USA; [3]Department of Biology, Texas A&M University, College Station, Texas 70843 USA

Comparative analysis of the protein sequences encoded in the four euryarchaeal species whose genomes have been sequenced completely (*Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum*, *Archaeoglobus fulgidus*, and *Pyrococcus horikoshii*) revealed 1326 orthologous sets, of which 543 are represented in all four species. The proteins that belong to these conserved euryarchaeal families comprise 31%–35% of the gene complement and may be considered the evolutionarily stable core of the archaeal genomes. The core gene set includes the great majority of genes coding for proteins involved in genome replication and expression, but only a relatively small subset of metabolic functions. For many gene families that are conserved in all euryarchaea, previously undetected orthologs in bacteria and eukaryotes were identified. A number of euryarchaeal synapomorphies (unique shared characters) were identified; these are protein families that possess sequence signatures or domain architectures that are conserved in all euryarchaea but are not found in bacteria or eukaryotes. In addition, euryarchaea-specific expansions of several protein and domain families were detected. In terms of their apparent phylogenetic affinities, the archaeal protein families split into bacterial and eukaryotic families. The majority of the proteins that have only eukaryotic orthologs or show the greatest similarity to their eukaryotic counterparts belong to the core set. The families of euryarchaeal genes that are conserved in only two or three species constitute a relatively mobile component of the genomes whose evolution should have involved multiple events of lineage-specific gene loss and horizontal gene transfer. Frequently these proteins have detectable orthologs only in bacteria or show the greatest similarity to the bacterial homologs, which might suggest a significant role of horizontal gene transfer from bacteria in the evolution of the euryarchaeota.

Phylogenetic analysis of rRNA and a set of proteins involved in translation, transcription, and replication has led to the concept of archaea as a third division of life, distinct from either bacteria or eukaryotes (Woese et al. 1978, 1990; Woese and Gupta 1981; Pace et al. 1986; Zillig 1991). Furthermore, rooting of paralogous trees for translation elongation factors and proton ATPases suggested that archaea are a sister group of eukaryotes (Gogarten et al. 1989a,b; Iwabe et al. 1989; Gribaldo and Cammarano 1998). This concept appears to be gaining further support from the generally eukaryotic layout of the genome expression systems, particularly the system of DNA replication whose principal components are orthologous to the respective replication proteins of eukaryotes but apparently do not have counterparts in bacteria (Mushegian and Koonin 1996; Brown and Doolittle 1997; Edgell and Doolittle 1997). However, it has been aptly noted that archaea have a "eubacterial form and eukaryotic content" (Keeling et al. 1994). Indeed, beyond the common "negative" trait, namely the small cell size and the absence of a nucleus, archaea and bacteria share major aspects of genome organization and expression strategy. The most important of these common features include the (typically) single circular chromosome, the absence of introns in protein-coding genes, the operonic organization of many genes, and the absence of a 5′-terminal cap and the presence of a ribosomal-binding (Shine-Dalgarno) site in archaeal mRNAs (Brown and Doolittle 1997). Furthermore, several operons, particularly those encoding ribosomal proteins, are conserved in archaea and bacteria (Brown and Doolittle 1997; Koonin and Galperin 1997).

The analysis of the first two completely sequenced archaeal genomes, those of *Methanococcus jannaschii* (Bult et al. 1996) and *Methanobacterium thermoautotro-*

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

9:608–628 ©1999 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/99 $5.00; www.genome.org

phicum (Smith et al. 1997), showed, somewhat unexpectedly given the already established archaeal–eukaryotic clade, that the bacterial form of archaea is complemented by considerable bacterial content. It has become clear that the majority of archaeal proteins show the greatest similarity to their bacterial homologs, which is likely to indicate bacterial origin, and only a minority look "eukaryotic" (Koonin et al. 1997; Smith et al. 1997). In functional terms, there is a clear split between the bacterial and eukaryotic components of the archaeal genomes—the eukaryotic genes are primarily those coding for components of the translation, transcription, and replication machineries, whereas the bacterial ones typically encode metabolic enzymes and proteins involved in cell division and cell wall biogenesis (Koonin et al. 1997; Smith et al. 1997). These findings raised the issue of possible extensive gene exchange between bacteria and archaea (Feng et al. 1997; Koonin et al. 1997; Doolittle and Logsdon 1998).

Subsequently, the complete genome sequences of two additional archaeal species, namely *Archaeoglobus fulgidus* (Klenk et al. 1997) and *Pyrococcus horikoshii* (Kawarabayasi et al. 1998a,b), have been reported. All four available complete archaeal genomes represent only one of the two (or possibly three) main archaeal subdivisions—the Euryarchaeota (Olsen et al. 1994; Pace 1997). Nevertheless, they show sufficient diversity to allow us, for the first time, to embark on a systematic comparative analysis of archaeal genomes. We describe here the results of a detailed comparative analysis of the four complete euryarchaeal protein sets. Our principal approach included the delineation of sets of orthologous genes and examination of phylogenetic patterns in these families (Tatusov et al. 1997; Koonin et al. 1998).

RESULTS AND DISCUSSION

Orthologous Families Delineated by Comparison of Four Euryarchaeal Genomes and the Principal Types of Events in Archaeal Evolution

The proteins encoded in the genomes of the four euryarchaeal species comprise a very good set for the delineation of families of likely orthologs [designated clusters of orthologous groups (of proteins), COGs; Tatusov et al. 1997)]. In the original COG analysis, we emphasized that to use consistency between different genomes to support the derivation of COGs, the sequences of the compared proteins should be maximally independent; therefore, this criterion works best with phylogenetically distant genomes. At large phylogenetic distances, however, correct identification of COGs may be hampered by other problems, such as difficulty in distinguishing orthologs from paralogs, and in some cases, very low similarity between orthologs that precludes their detection altogether. As a result, the final step in the construction of the original collection of COGs involved considerable manual correction. The distances separating the four archaeal species are intermediate between those that are seen among close bacterial species such as *Escherichia coli* and *Haemophilus influenzae* (in the original COG analysis, these species were not considered independently) and those between phylogenetically remote species such as bacteria and eukaryotes. In quantitative terms, the mean percent identity of the best hits in all-against-all interspecies comparisons of protein sequences is in the range of 41%–46% for the archaea, 57% for *E. coli* versus *H. influenzae*, and between 30%–35% for most distant bacterial lineages and bacteria versus eukaryotes or archaea (N.V. Grishin, unpubl.; ftp://ncbi.nlm.nih.gov/pub/koonin/gen2gen). It appears that the intermediate level of sequence conservation seen among the archaea is high enough to prevent most, if not all, artificial lumping of COGs attributable to paralogous families, but low enough for the consistency criterion to be valid and useful. For these reasons, most of the archaeal COGs delineated by the automatic procedure were corroborated by subsequent case-by-case evaluation. Furthermore, given the typically highly significant similarity between archaeal orthologs, it is most unlikely that any significant number of them have been missed as a result of low sequence conservation.

Figure 1 shows the breakdown of the archaeal protein set in terms of their conservation in the four complete genomes. The majority of the proteins in each species—from 58% for *P. horikoshii* to 71% for *M. jannaschii*—belong to the archaeal families of likely orthologs (COGs), and another sizable fraction (from 7%

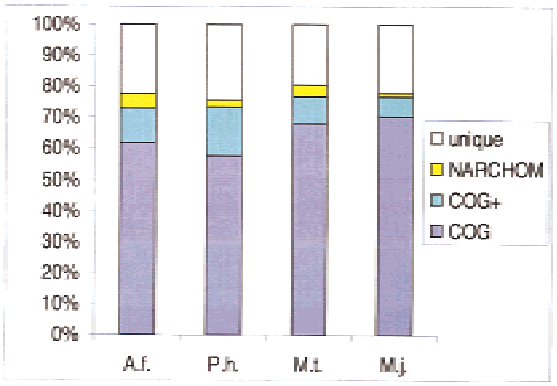


Figure 1 Conserved families and unique proteins encoded in the four complete archaeal genomes. (COGs+) Distant homologs of COGs; (NARCHOM) nonarchaeal homologs (only); (unique) proteins without detectable homologs in other species (for details see text); (Af) *Archaeoglobus fulgidus*; (Ph) *Pyrococcus horikoshii*; (Mt) *Methanobacterium thermoautrophicum*; (Mj) *Methanococcus jannaschii*.

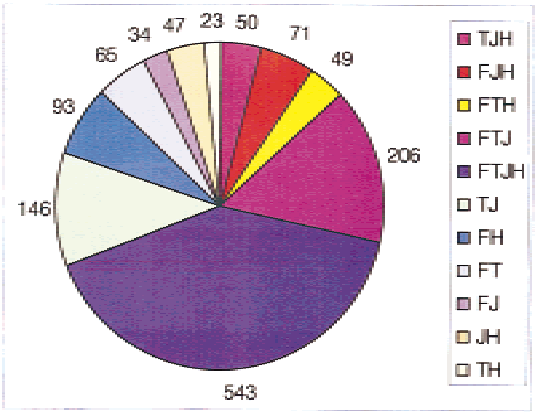


Figure 2 Representation of the four archaeal species in the COGs. (F) *Archaeoglobus fulgidus*; (T) *Methanobacterium thermoautrophicum*; (J) *Methanococcus jannaschii*; (H) *Pyrococcus horikoshii*.

for *M. jannaschii* to 11% for *A. fulgidus*) were identified as distant homologs of the COGs. Among the remaining proteins that had no archaeal homologs, for a relatively small fraction (from 1% in *M. jannaschii* to 4% in *A. fulgidus*), homologs were detected in other taxa (primarily bacteria), and the rest (~20%) had no detectable homologs. This distribution suggests that a conserved archaeal gene set does exist. This core gene set, however, includes a minority of the archaeal genes as indicated by the fact that only 543 of the 1326 identified COGs (40%) are represented in all four archaeal species; the remaining COGs are roughly equally divided between those that include three and two species (Fig. 2). The universal archaeal COGs encompass 31%–35% of the proteins encoded in each of the individual genomes. This number appears to be an important measure of the evolutionary stability of the genomes—the rest of the gene complement in each of the archaea must have been subject to evolutionary events other than vertical inheritance, such as duplication with subsequent rapid divergence, horizontal gene transfer, and lineage-specific gene loss.

These results provide at least a rough estimate of the likely amount of gene loss in each species, as well as the number of COGs represented in the ancestral euryarchaeon. A conservative estimate of the number of genes that might have been lost in each genome is provided by the number of COGs that include three archaeal species other than the given one. This number is in the range of 50 to 70 for *M. jannaschii*, *M. thermoautotrophicum*, and *A. fulgidus*, as opposed to 206 in *P. horikoshii* (Fig. 2). The greatest number of COGs that are not represented in *P. horikoshii* is not surprising as it is a heterotrophic organism that lacks a number of biosynthetic capabilities (Gonzalez et al. 1998). The majority of the archaea are autotrophs and it seems most likely that the ancestral form also had been autotrophic; thus, the absence of the representatives of many COGs in *P. horikoshii* is best explained by lineage-specific gene elimination. At least some of the archaeal COGs with two members are also likely to reflect gene loss. Thus, a higher estimate for the number of ancestral genes lost in each genome can be obtained by adding up all COGs with three or two members that do not include the given species. The result varies from a total of 220 genes for *M. jannaschii* to 451 genes for *P. horikoshii*.

Thus, the analysis of the conserved archaeal families reveals major genome plasticity, with only a minority of families represented in all genomes. These observations make all the more pertinent the question: which essential cellular functions are provided by the set of 543 universal archaeal COGs and which are not represented by it, and, accordingly, are performed by nonorthologous (unrelated or paralogous) proteins in different species—the phenomenon described as nonorthologous gene displacement (Koonin et al. 1996a; Mushegian and Koonin 1996).

The Core Set of Conserved Euryarchaeal Genes, Lineage-Specific Gene Loss, and Nonorthologous Gene Displacement

The COGs represented in all four euryarchaeal species are significantly enriched in proteins that are involved in genome expression, compared to the entire collec-

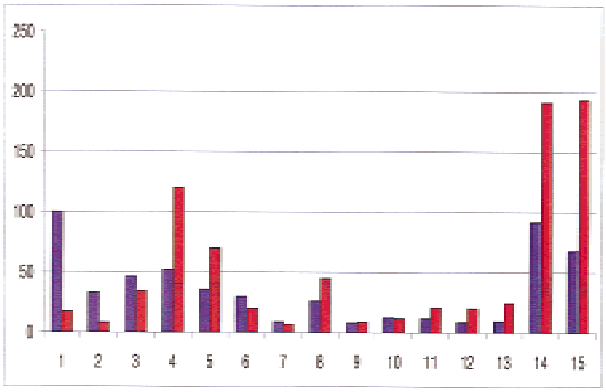


Figure 3 Distribution of predicted protein functions in the universal and nonuniversal subsets of the archaeal COGs. (Blue bars) The universal subset (543 COGs with four members each); (red bars) the nonuniversal subset (783 COGs with two or three members each). (Vertical axis) Number of COGs; (horizontal axis) functional categories: 1, translation, ribosome structure, and biogenesis; 2, transcription; 3, DNA replication, repair, recombination; 4, energy production and methanogenesis; 5, amino acid metabolism; 6, nucleotide metabolism; 7, carbohydrate metabolism; 8, coenzyme metabolism; 9, lipid metabolism; 10, molecular chaperones and related functions; 11, cell wall biogenesis and cell division; 12, secretion and motility; 13, inorganic ion transport; 14, general functional prediction only; 15, no functional prediction.

tion of the archaeal COGs. In particular, most of the basic components of the translation, transcription, and replication systems are conserved consistently in all four species; the same is true of a number of proteins implicated in repair and recombination (Fig. 3).

In other functional categories of genes, the genome plasticity revealed by COG analysis is more pronounced. Because of the apparent loss of a number of biosynthetic pathways in the heterotrophic *P. horikoshii*, there are relatively few metabolic enzymes among the all-archaeal COGs, and in fact, it does not seem possible to delineate even a single metabolic pathway that would be completely orthologous in all four archaea (Table 1). Among the three autotrophic species, most of the steps of the central pathways are represented by orthologs; nevertheless, almost each pathway has at least one step where nonorthologous displacement is likely (Table 1). The biosynthesis of branched chain aliphatic amino acids (leucine, isoleucine, valine) is an example of a complex pathway that is, in its entirety, represented by orthologs in the three autotrophic archaea as well as in most bacteria. This is, however, an exception rather than the rule among the archaeal metabolic pathways—few of them consist exclusively of orthologs of bacterial enzymes. In most pathways, at least one or two reactions are predicted to be catalyzed either by known archaea-specific enzymes or by yet uncharacterized ones (Table 2). In the readily detectable cases of nonorthologous gene displacement, one of the alternative solutions is frequently based on orthologs of the respective bacterial enzymes, whereas the other one seems to be unique for archaea and is not always identifiable. This is, for example, the situation with a critical reaction in glycolysis, namely the formation of pyruvate from phosphoenolpyruvate. *M. jannaschii* and *P. horikoshii* encode an ortholog of the bacterial pyruvate kinase that is predicted to catalyze this reaction. Pyruvate kinase, however, is not detectable in the other two archaea. Given that the other components of the trunk portion of the glycolytic pathway are present and that the reaction catalyzed by pyruvate kinase is indispensable for the completion of glycolysis, nonorthologous displacement must be invoked. The most likely displacing enzyme is phosphoenolpyruvate synthase, which is conserved in all archaea and might produce pyruvate by reversing its typical reaction.

Nonorthologous gene displacement is notable also in the archaeal amino acid metabolism. For example, different archaeal species apparently use radically different pathways to synthesize proline. In *M. thermoautotrophicum* and *A. fulgidus*, proline can be formed from ornithine in a single reaction catalyzed by ornithine cyclodeaminase (Sans et al. 1988). *M. jannaschii* and *P. horikoshii* lack this enzyme, and while the latter is expected to be a proline auxotroph, the only possible route for proline biosynthesis in *M. jannaschii* appears to be through the deacetylation of *N*-acetylglutamate γ-semialdehyde into γ-glutamic semialdehyde, followed by its conversion into pyrroline-5-carboxylate and then to proline as shown for bacteria and yeast (Adams and Frank 1980). *M. jannaschii* encodes an ortholog of the *N*-acetylornithine deacetylase (ArgE) that catalyzes the first step of this pathway. The second step of the pathway, conversion of γ-glutamic semialdehyde to pyrroline-5-carboxylate, occurs spontaneously. However, the ortholog of the bacterial enzyme for the last step of proline biosynthesis, namely pyrroline-5-carboxylate reductase (ProC), is not encoded in the *M. jannaschii* genome and should have been displaced by another dehydrogenase that remains to be identified experimentally. Remarkably, *A. fulgidus* encodes only the ArgE ortholog and *M. thermoautotrophicum* only the ProC ortholog. It appears that in this case, we observe nonorthologous displacement of an entire (albeit short) pathway whereby acquisition of the ornithine cyclodeaminase gene by *A. fulgidus* and *M. thermoautotrophicum* has made the enzymes of the original pathway of proline biosynthesis dispensable.

In addition to the cases of apparent nonorthologous displacement, there are several important gaps in our understanding of metabolic pathways in *all* euryarchaeota. The archaeal version of sugar metabolism is particularly puzzling. There is no doubt that autotrophic archaea possess the capabilities to synthesize ribose, deoxyribose, and the sugar components of the cell envelope. It is unclear, however, how they accomplish this in the absence of aldolase, fructose bisphosphatase, transaldolase, transketolase, and pentose-5-phosphate 3-epimerase (see Table 1). Genes for all these enzymes are missing in *M. thermoautotrophicum* and *A. fulgidus*, whereas *M. jannaschii* has genes coding for the three latter enzymes but not the former two. It appears that compared to bacteria, the archaeal sugar metabolism shows systematic nonorthologous displacement of enzymes. Interestingly, one of the archaeal COGs includes predicted aldolases that are highly conserved in all four archaea and are orthologous to the recently identified class I fructose-biphosphate aldolase from *E. coli* (Thomson et al. 1998). There are two paralogous representatives of this family of aldolases in *M. jannaschii* and *A. fulgidus* and only one member in *M. thermoautotrophicum* and *P. horikoshii* (Table 1). These enzymes are likely to catalyze key reactions both in pentose and in hexose biosynthesis; the exact pathways remain to be studied experimentally.

Archaeal COGs that contain four or three members account for the majority of known housekeeping functions, with several notable exceptions (e.g., those in the translation machinery discussed above), and in a sense, may be considered an idealized minimal ar-

Table 1. Orthologous and Nonorthologous Metabolic Pathways and Enzymes in Archaea

Pathway	Enzymes (genes) in the pathway[a]	Orthologs of bacterial genes found in all four archaeal genomes	Genes missing in all four archaeal genomes	Nonorthologous gene displacement: orthologs of bacterial genes found only in some of the archaeal genomes	Consequences for archaeal metabolism
Glycolysis	hexokinase (*glk*), phosphoglucoisomerase (*pgi*), phosphofructokinase (*pfkA*), aldolase (*fba/dhnA*), triosephosphate isomerase (*tpi*), glyceraldehyde 3-phosphate dehydrogenase (*gapA*), 3-phosphoglycerate kinase (*pgk*), phosphoglyceromutase (*pgm/yibO*), enolase (*eno*), pyruvate kinase (*pykA*)	*dhnA, tpi, gapA, pgk, pgm*[d]*, eno*	*glk pfkA*	*pgi* is present only in Mj; *pykA* is present in Mj and Ph, not found in Mt and Af	bacterial-type hexokinase and phosphofructokinase are apparently displaced by nonorthologous ADP-dependent enzymes; the lack of pyruvate kinase in Mt and Af is probably compensated by phosphoenolpyruvate synthase working in the reverse direction.
Gluconeogenesis	phosphoenolpyruvate synthase (*ppsA*), enolase (*eno*), phosphoglyceromutase (*pgm*), 3-phosphoglycerate kinase (*pgk*), glyceraldehyde 3-phosphate dehydrogenase (*gapA*), triosephosphate isomerase (*tpi*), aldolase (*fba/dhnA*), fructose bisphosphatase (*fbp*), phosphoglucoisomerase (*pgi*)	*ppsA, eno, pgm, pgk, gapA, tpi, dhnA*	*fbp*	*pgi* is present only in Mj, not found in Mt, Af, and Ph	other phosphohexomutases (e.g., phosphomannomutase) are probably used for polysaccharide biosynthesis in Mt, Af, and Ph.
Pentose phosphate shunt and pentose biosynthesis	glucose-6-phosphate dehydrogenase (*zwf*), 6-phosphogluconate dehydrogenase (*gnd*), transketolase (*tktA*); transaldolase (*talA*), pentose-5-phosphate-3-epimerase (*yhfD*), ribose 5-phosphate isomerase (*rpiA*), deoxyribose-phosphate aldolase (*deoC*)	*rpiA*	*zwf, gnd*	*tktA* gene is split in Mj, absent in Mt, Af, and Ph; *talA* and *yhfD* are present in Mj, not in Mt, Af, or Ph; *deoC* is present only in Mt	this pathway is not functional in any of these archaea. The mechanism of pentose phosphate biosynthesis is not clear. A predicted DhnA-type aldolase that is highly conserved in all four archaea (MJ0400, MJ1585; MTH579; AF0108, AF0230; PH0082) may catalyze the formation of ribose from glyceraldehyde-3-phosphate and acetaldehyde. Alternatively, in Mt, this reaction might be catalyzed by the DeoC ortholog (MTH818).
Entner–Doudoroff pathway[b,c]	glucose-6-phosphate dehydrogenase (*zwf*), 6-phosphogluconate dehydratase (*edd*), 2-keto-3-deoxy-6-phosphogluconate aldolase (*eda*)	*edd*	*zwf, eda*		archaea lack the classical Entner–Doudoroff path-way and instead appear to possess a modified, nonphosphorylated ver-sion. In Mj, Af, and Mt, members of the above new family of aldolases may function in this pathway as 2-keto-3-de-oxygluconate aldolase (a nonorthologous displace-ment of Eda). The archaeal gluconate dehydratase remains unknown, which precludes a complete reconstruction of this pathway.
TCA cycle[c]	citrate synthase (*gltA*), aconitase (*acnA*), isocitrate dehydrogenase (*icd*),	*acnA, icd, sucC, sucD, frdA, frdB,*	*sucA, sucB*	*gltA* is present in Mt and Af, but not in Mj or Ph; Ph has	Mt and Af can reduce α-ketoglutarate to citrate and further to succinate

Table 1. (*Continued*)

Pathway	Enzymes (genes) in the pathway[a]	Orthologs of bacterial genes found in all four archaeal genomes	Genes missing in all four archaeal genomes	Nonorthologous gene displacement: orthologs of bacterial genes found only in some of the archaeal genomes	Consequences for archaeal metabolism
	α-ketoglutarate dehydrogenase (*sucA, sucB*), succinyl-CoA synthase (*sucC, sucD*), fumarate reductase (*frdA, frdB*), fumarase (*fumA*), malate dehydrogenase (*mdh*)	*fumA, mdh*		*fumA* and, possibly, *acnA* genes	or succinyl-CoA; in Mj only the part from oxaloacetate to succinyl-CoA is operative.
Purine biosynthesis	phosphophosphoribosylpyrophosphate synthase (*prsA*), amidophosphoribosyltransferase (*purF*), GAR synthase (*purD*), GAR transformylase (*purN/purT*), FGAM synthase (*purL*), AIR synthase (*purM*), NCAIR synthase (*purK*), NCAIR mutase (*purE*), SAICAR synthase (*purC*), adenylosuccinate lyase (*purB*), AICAR transformylase (*purH2*), IMP cyclohydrolase (*purH1*), adenylosuccinate synthase (*purA*), IMP dehydrogenase (*guaB*), GMP synthase (*guaA*),	*prsA, purF, purD, purL, purM, purE, purC, purB, purA, guaA*	*purK, purH2*	*purT* is present in Mj and Ph but not in Mt and Af; *purH1* is present only in Af; *guaB* is present in Mj, Mt, and Ph, but not in AF	all four archaea are probably capable of purine synthesis de novo; carboxylation of AIR probably occurs spontaneously. The still unidentified enzymes that catalyze formylation of SAICAR and AICAR in all four archeaea and formylation of GAR in Mt and Af apparently use formate and ATP as substrates. IMP dehydrogenase is missing AF and is probably displaced by a nonorthologous dehydrogenase.
Pyrimidine biosynthesis	carbamoylphosphate synthase (*carA, carB*), aspartate carbamoyltransferase (*pyrB*), dihydroorotase (*pyrC/ygeZ*), dihydroorotate dehydrogenase (*pyrD*), orotate phosphoribosyl-transferase (*pyrE*), orotidine-5′-phosphate decarboxylase (*pyrF*), UMP kinase (*pyrH*), NDP kinase (*ndk*), CTP synthase (*pyrG*)	*pyrB, ygeZ, pyrD, pyrE, pyrF, pyrH, ndk, pyrG*	none	*carA* and *carB* are missing in Ph	Mj, Mt, Af and probably Ph are capable of pyrimidine synthesis de novo. The identity of carbamoylphosphate synthase in Ph remains unclear.
Histidine biosynthesis[c]	phosphophosphoribosylpyrophosphate synthase (*prsA*), ATP-phosphoribosyltransferase (*hisG*), phosphoribosyl-ATP pyrophosphatase (*hisI2*), phosphoribosyl-AMP cyclohydrolase (*hisI1*), 5′-ProFAR isomerase (*hisA*), imidazoleglycerol phosphate synthase (*hisH, hisF*), imidazoleglycerol phosphate dehydratase (*hisB2*), histidinol phosphate aminotransferase (*hisC*), histidinol phosphatase (*hisB1*), histidinol dehydrogenase (*hisD*)	*prsA, hisG, hisI1, hisA, hisH, hisF, hisB2, hisC, hisD*	none	*hisI2* is present in Mj and Mt, but not in Af; *hisB1* is found only in Mj; *prsA* and *hisC* genes are present in Ph	Mj, Mt, and Af are all capable of histidine biosynthesis; phosphoribosyl-ATP pyrophosphatase in Af is probably displaced by some other ATP/ADPase; all archaea encode distant homologs of yeast histidinol phosphatase (HD superfamily hydrolases; Aravind and Koonin 1998b), one of which might displace HisB1 in Af and Mt.
Branched chain amino acids biosynthesis[c]	threonine deaminase (*ilvA*), acetohydroxyacid synthase (*ilvB, ilvN*), acetohydroxyacid isomeroreductase (*ilvC*), dihydroxyacid dehydratase (*ilvD*), 2-isopropylmalate synthase (*leuA*), isopropylmalate isomerase (*leuC, leuD*), 3-isopropyl-malate	*ilvA, ilvB, ilvN, ilvC, ilvD, leuA, leuC, leuD, leuB, ilvE*	none	none	all enzymes of leucine, isoleucine, and valine biosynthesis in bacteria, archaea, and yeast are orthologous.

Table 1. (*Continued*)

Pathway	Enzymes (genes) in the pathway[a]	Orthologs of bacterial genes found in all four archaeal genomes	Genes missing in all four archaeal genomes	Nonorthologous gene displacement: orthologs of bacterial genes found only in some of the archaeal genomes	Consequences for archaeal metabolism
	dehydrogenase (*leuB*), glutamate transaminase (*ilvE*)				
Aromatic amino acids biosynthesis[c]	3-deoxyheptulosonate 7-phosphate synthase (*aroG/kdsA*), 3-dehydroquinate synthase (*aroB*), 3-dehydroquinate dehydratase (*aroD*), shikimate dehydrogenase (*aroE*), shikimate kinase (*aroK*), 5-enolpyruvoylshikimate 3-phosphate synthase (*aroA*), chorismate synthase (*aroC*), chorismate mutase (*pheA1*), prephenate dehydratase (*pheA2*), prephenate dehydrogenase (*tyrA2*), tyrosine aminotransferase (*tyrB*), antranilate synthase (*trpD1, trpE*), antranilate phosphoribosyl-transferase (*trpD2*), phosphoribosylanthranilate isomerase (*trpC2*), indole-glycerol phosphate synthase (*trpC1*), tryptophan synthase (*trpA, trpB*)	*aroD, aroE, aroA, aroC, pheA1, pheA2, tyrA2, tyrB, trpD1, trpE, trpD2, trpC2, trpC1, trpA, trpB*	*aroG, aroB, aroK*	none	the mechanism of 3-dehydroquinate synthesis remains unclear; shikimate phosphorylation in all autotrophic archaea is probably performed by an archaea-specific kinase.
Threonine biosynthesis	aspartokinase (*thrA1*), aspartate semialdehyde dehydrogenase (*asd*), homoserine dehydrogenase (*thrA2*), homoserine kinase (*thrB*), threonine synthase (*thrC*)	*thrA1, asd, thrA2, thrC*	none	*thrB* is present in Mj and Ph, but not in Mt and Af	in Mt and Af, homoserine kinase is probably displaced by a different kinase.
Methionine biosynthesis	aspartokinase (*metL1*), aspartate semialdehyde dehydrogenase (*asd*), homoserine dehydrogenase (*metL2*), homoserine transsuccinylase (*metA*), cystathionine γ-synthase (*metB*), β-cystathionase (*metC*), methionine synthase (*metE/metH*)	*metL1, asd, metL2, metE*	*metA,*	*metB/metC* is found only in Ph, not in Mj, Mt, or Af	in all four archaea, the three steps leading from homoserine to homocysteine are probably displaced by a single reaction catalyzed by a sulfur transferase.
Arginine biosynthesis	acetylglutamate synthase (*argA2*), acetylglutamate kinase (*argB*), acetylglutamate phosphate reductase (*argC*), acetylornithine aminotransferase (*argD*), acetylornithinase (*argE*), ornithine carbamoyltransferase (*argF*), argininosuccinate synthase (*argG*), argininosuccinate lyase (*argH*)	*argA2, argB, argC, argD, argF, argH*	none	*argE* is present in Mj, Af; and Ph, but not in Mt; *argG* is present in Mj, Mt and Af but not in Ph	in Mt, acetylornithinase is probably displaced by a different acetyltransferase.
NAD biosynthesis	aspartate oxidase (*nadB*), quinolinate synthase (*nadA*), quinolinate phosphoribosyltransferase (*nadC*), nicotinic acid mononucleotide adenylyltransferase (*nadD*), deamido-NAD ammonia ligase (*nadE*)	*nadB, nadC, nadE*	none	*nadA* is present in Mj, Mt, and Ph, but not in AF	*nadA* is probably displaced in AF by a different enzyme; *nadD* gene (predicted *E. coli yneB*) remains unidentified in bacteria and archaea.
Riboflavin biosynthesis[c]	GTP cyclohydrolase II (*ribA*), pyrimidine deaminase (*ribD1*), pyrimidine reductase (*ribD2*),	*ribD2, ribB, ribE*	*ribC*	*ribA* and *ribD1* are present in Af but not in Mj or Mt	*ribC* is displaced by an archaea-specific riboflavin synthase (Eberhardt et al.,

Table 1. (*Continued*)

Pathway	Enzymes (genes) in the pathway[a]	Orthologs of bacterial genes found in all four archaeal genomes	Genes missing in all four archaeal genomes	Nonorthologous gene displacement: orthologs of bacterial genes found only in some of the archaeal genomes	Consequences for archaeal metabolism
	3,4-dihydroxybutanone-4-phosphate synthase (*ribB*), 6,7-dimethyl-8-ribityllumazine synthase (*ribE*), riboflavin synthase (*ribC*)				1997); the mechanism of 2,5-diamino-6-ribosyl-amino-4-pyrimidone 5′-phosphate formation Mj and Mt remains unclear.
Siroheme biosynthesis[c]	Glutamyl-tRNA reductase (*hemA*), glutamate 1-semialdehyde aminotrans-ferase (*hemL*), probilinogen III synthase (*hemB*), hydroxy-methylbilane synthase (*hemC*), uroporphyrinogen III synthase (*hemD*), uroporphirinogen methyltransferase (*cysG2*), dimethyluroporphirinogen III dehydrogenase (*cysG1*)	*hemA, hemL, hemB, hemC, hemD, cysG2, cysG1*	none	none	all enzymes of siroheme biosynthesis in archaea are orthologous to bacterial ones.
Cobalamin biosynthesis[c]	uroporphyrinogen III methylase (*cysG2*), precorrin-2 methylase (*cbiL*), precorrin-3B methylase (*cbiH*), precorrin-4 methylase (*cbiF*), precorrin-6A reductase (*cbiJ*), precorrin 6B methylase (*cbiE*), precorrin 6B decarbox-ylase (*cbiT*), precorrin-8x isomerase (*cbiC*), cobyrinic acid *a,c*-diamide synthase (*cbiA*), cobalt insertion protein (*cobN*), cob(I)alamin adenosyl-transferase (*cobA*), cobyric acid synthase (*cbiP*), cobyric acid aminotransferase (*cobD*), cobinamide synthase (*cbiB*), nicotinate-nucleotide:dimethyl-benzimidazole phosphoribosyl-transferase (*cobT*), cobalamin synthase (*cobS*)	*cysG2, cbiH, cbiF, cbiE, cbiC, cbiA, cbiP, cbiB, cobS*	*cobA, cobD*	*cbiL, cbiJ, cbiT,* and *cobN* are present in Mj and Mt but not in Af; *cobT* is found only in Af	in Mj, Mt, and Af, *cobA* and *cobD* gene products are probably displaced by archaea-specific adenosyl- and aminotransferases, respectively; it is not clear whether this pathway is functional in Af.
Biotin biosynthesis	pimeloyl-CoA synthetase (*bioW*)[e], 7-keto-8-aminopel-argonate synthetase (*bioF*), 7,8-diaminopelargonate aminotransferase (*bioA*), dethiobiotin synthetase (*bioD*), biotin synthetase (*bioB*), biotin-[acetyl-CoA carboxylase] holoenzyme synthetase (*birA*)	none	none	all the enzymes of the pathway are present in Mj but only one or two enzymes can be found in Mt, Af, and Ph	probably only Mj is capable of biotin biosynthesis.

[a]The genes and pathways follow the biochemical data and nomenclature described for *E. coli* and *S. typhimurium* (Neidhardt et al. 1996). Genes coding for multidomain proteins with more than one enzymatic activity are divided into separate domains, starting from the amino-terminal domain. Known cases of nonorthologous gene displacements are indicated with a slash; genes encoding different subunits of the same enzyme are separated by commas.
[b]6-Phosphogluconate dehydratase gene (*edd*) is apparently present in Mj, Mt, and Af, but these ORFs probably function as dihydroxyacid dehydratases (*ilvD*), which is a paralog of *edd*.
[c]Most enzymes of these pathways are not encoded in the *Pyrococcus horikoshii* genome; exceptions are indicated.
[d]Orthologs of *B. subtilis pgm* gene, corresponding to the *E. coli yibO.*
[e]Orthologs of *B. subtilis* gene *bioW,* not found in *E. coli.*

Table 2. Synapomorphies in Euryarchaeota (Examples)

COG description	Representatives				Unique shared characters nonarchaeal homologs (see also Fig. 8)
	Mj	Mt	Af	Ph	
DNA polymerase II large subunit	MJ1630	MTH1536	AF1722	PHBN021	a highly conserved archaeal enzyme without similarity to any other proteins, with the exception of a C4 Zn finger resembling those in eukaryotic DNA polymerase δ.
DNA polymerase II small subunit, predicted phosphohydrolase of the calcineurin-like superfamily	MJ0702	MTH1405	AF1790	PHBN023 PHAZ021	Predicted active phosphohydrolase (phosphatase) in archaea and an inactivated form in eukaryotes (Aravind and Koonin 1998a).
Predicted ATP-dependent DNA ligase	MJ0414	MTH1221	AF0849	PHBG013	very limited similarity to other ATP-dependent DNA ligases except for one from *Aquifex aeolicus,* probably due to horizontal transfer (Altschul and Koonin 1998).
DNA excision repair enzyme	MJ1505	MTH1415	AF0358	PHAI012	consists of a typical helicase domain and a nuclease domain as opposed to the apparent eukaryotic orthologs (ERCC4/RAD1) in which the helicase domain appears to be inactivated (Aravind et al. 1999).
DnaG-type primase-like proteins	MJ1206	MTH891	AF1899	PHAN003	unique domain organization with a N-terminal helicase motif combined with the DnaG-type (Toprim) domain (Aravind et al. 1998).
DNA-directed RNA polymerase subunit (E′/E″)	MJ0396/ MJ0397	MTH264/ MTH265	AF1116/ AF1117	PHBT008/ inPH744	two single domain (S1 and C4 Zn finger domains) proteins in all Euryarchaeota; a fusion in *Sulfolobus;* only the S1 domain protein is a RNA polymerase subunit in eukaryotes (Fig. 8).
DNA-dependent RNA polymerase A′/A″ subunits	MJ1042/ MJ1043	MTH1051/ MTH1052	AF1888/ AF1889	PHCB020/ PHCB021	the split of the largest RNA polymerase subunit gene into two adjacent genes is unique to archaea. Both eukaryotes and bacteria encode highly conserved orthologs of the archaeal A′ and A″ subunits as a single polypeptide (the β′-subunit in bacteria).
Predicted HTH transcriptional regulators	MJ0188	MTH1282	AF1259	PHCN020	unique domain organization: two CBS domains fused with an HTH domain.
Archaeosine synthetase (archaea-specific tRNA modification)	MJ1022	MTH1665	AF0587	PHBN035	two-domain architecture, with an additional, predicted RNA-binding PUA domain, as opposed to bacterial homologs (queuine synthetases) that consist of the enzymatic domain alone (Fig 8; Aravind and Koonin 1999a).
Translation elongation factor EF-1β	MJ0459	MTH1699	AF0574	AP000001	the archaeal EF-1β is a small protein of ~120 amino acids whereas all eukaryotic homologs (orthologs?) contain an additional domain homologous to GSTs (Koonin et al. 1994).
ATP-dependent protease Lon	MJ1417	MTH785	AF0364	PHBH031	only the carboxy-terminal, protease domain is highly conserved in archaea and bacteria; the amino-terminal ATPase domain in the archaeal proteins is distinct from the ATPase domain of Lon.
PilT family ATPase	MJ1533	MTH246	AF1951	PHBP012	unique domain organization—ATPase + amino-terminal PIN domain.
GMP synthetase subunits—PP-family ATPases and glutamine amidotransferase	MJ1131	MTH710	AF0253	PHAU017	ATPase (top row) and glutamine amidotransferase (bottom row) moieties of the GMP synthetase are separate polypeptides. Orthologs of each subunit in bacteria and eukaryotes are domains of a single protein.
	MJ1575	MTH709	AF1320	PHAU016	
Predicted enzyme with an ATP-grasp domain and a redox active center	MJ0202	MTH1744	AF1104	PHBQ042	ortholog with the same domain architecture only in *Aquifex;* all other homologs are distantly related and lack the redox center.

chaeal gene complement. The COGs with two members appear to account for more specific functions linked to the organism's particular life style, for example, a number of COGs that include enzymes involved in methanogenesis in *M. jannaschii* and *M. thermoautotrophicum*.

Relationships Between Euryarchaeal Protein Families and Their Bacterial and Eukaryotic Homologs

The majority of the archaeal COGs have homologs in other taxa. In the present analysis, we attempted to distinguish carefully between true orthology (see Methods) and other homologous relationships that typically include weak sequence conservation or differences in domain architectures. There are notable differences in the distribution of the apparent phylogenetic affinities for the COGs represented in all archaea (universal) and those that include only three or two archaeal species. For >50% of the universal archaeal COGs, orthologs were identified in both bacteria and eukaryotes, in a sharp contrast to the nonuniversal COGs for which this fraction comprised of only 28% (Fig. 4A,B). A significant majority of the COGs that have *only* bacterial orthologs are not conserved in all archaea, whereas most of the COGs that have *only* eukaryotic orthologs belong to the universal subset (Fig. 4A,B). Furthermore, those COGs that do not have any homologs outside the archaea are poorly represented in the universal subset.

A complementary, quantitative analysis of the distribution of sequence similarities supports these observations. Archaeal proteins from the COGs that include only two or three species typically show the greatest similarity to bacterial homologs, in contrast to the universal COGs that are significantly enriched in proteins most similar to the eukaryotic homologs (Fig. 5). This

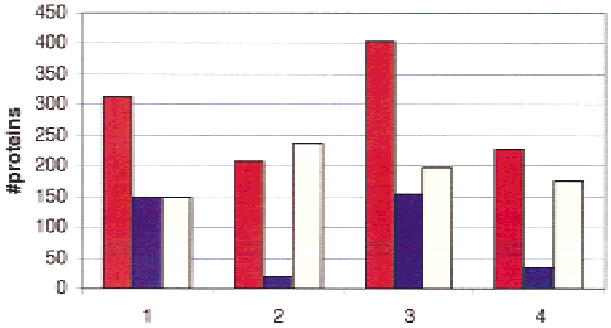


Figure 5 Relationship between members of the universal and nonuniversal subsets of the euryarchaeal COGs from *M. jannaschii* and *A. fulgidus* and their bacterial and eukaryotic homologs (1) *M. jannaschii*, the universal subset; (2) *M. jannaschii*, the nonuniversal subset; (3) *A. fulgidus*, the universal subset; (4) *A. fulgidus*, the nonuniversal subset. (Bacterial) Reliable best hits to bacterial proteins; (eukaryotic) reliable best hits to eukaryotic proteins. A reliable best hit was defined as one with an e-value at least 10000 times lower than that for the other divisions (eukaryotic or bacteria, respectively). Only the hits with e-values <0.001 were analyzed. (Red bars) Bacterial; (blue bars) eukaryotic; (yellow bars) uncertain.

difference might reflect true phylogenetic affinities, difference in evolutionary rates in different functional categories of proteins, or both. However, the finding that COGs consisting of two to three euryarchaeal members typically show a greater similarity to bacterial homologs, might suggest a significant contribution of horizontal transfer of bacterial genes into archaea.

The functional distinction between bacterial and eukaryotic COGs in archaea is clear-cut and is related to the functional difference between the universal and specialized subsets discussed above (see Fig. 3). The bacterial COGs within the universal subset comprise primarily proteins involved in energy production (e.g., ferredoxins and numerous components of hydrogenase complexes), certain metabolic functions, such as coenzyme biosynthesis, and transport system components. Interestingly, this bacterial set also includes enzymes involved in protein degradation and potentially in chaperone-like functions, such as three families of previously undetected predicted zinc-dependent proteases (K.S. Makarova, L. Aravind, and E.V. Koonin, unpubl.). Furthermore, the bacterial component of the universal COG subset includes several repair enzymes, proteins involved in cell division, for example, chromosome partitioning ATPases and stress response proteins, such as the homologs of the bacterial universal stress protein UspA.

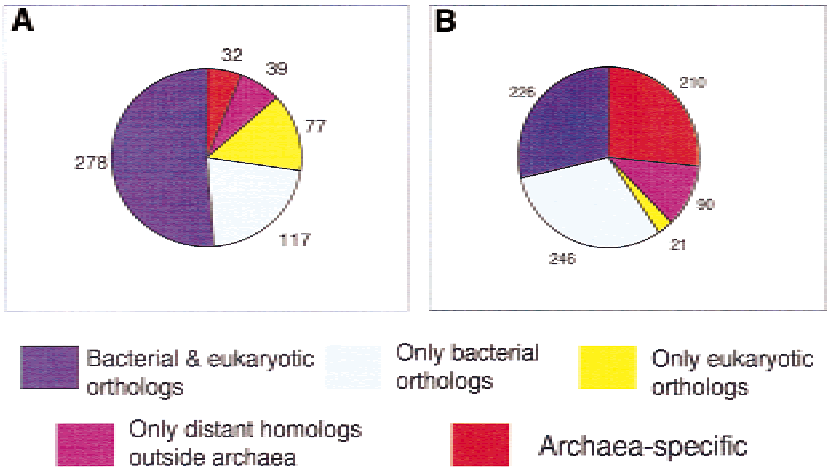


Figure 4 Taxonomic distribution of nonarchael homologs for universal and nonuniversal subsets of the archael COGs. (*A*) The universal subset (543 COGs with four members each); (*B*) the nonuniversal subset (783 COGs with two or three members each).

The UspA homologs are an example of a protein superfamily that originally has not been recognized in archaeal genome analyses but, in fact, is conserved in all archaea, most bacteria, plants, and fungi; all archaea and many bacteria encode multiple, paralogous members of this superfamily (Fig. 6). Most of the proteins in the superfamily consist of one or more copies of the UspA domain, but in the *A. fulgidus* protein AF1612

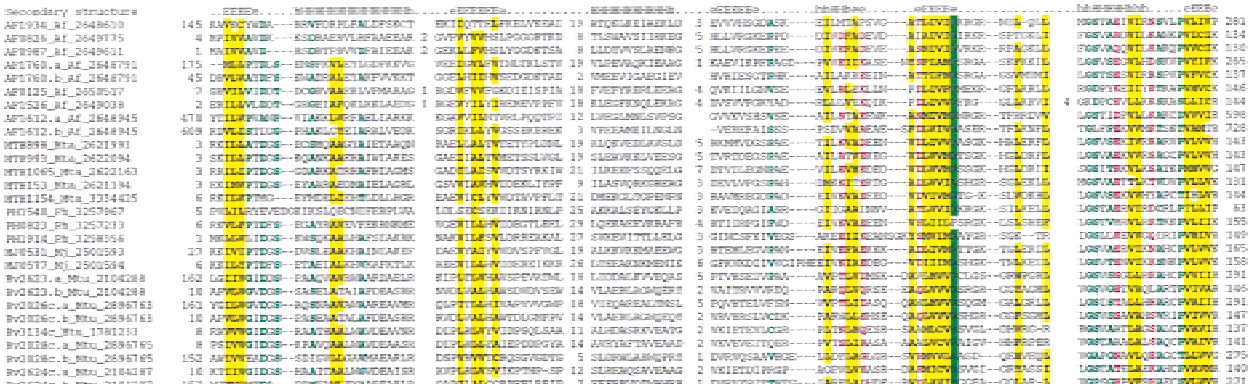


Figure 6 Previously undetected protein family conservation in archaea, bacteria, and eukaryotes—the UspA superfamily of predicted nucleotide-binding, regulatory proteins. The alignment was constructed on the basis of the PSI-BLAST results using the Clustal W program. The inclusion of each sequence in the superfamily was statistically supported by the PSI-BLAST analysis with an e-value of at least 0.01. The left column includes the protein (gene) names, and the gene identification (GI) numbers (after the underscore). A consensus derived using the 80% cutoff is shown underneath the alignment and the respective alignment columns are highlighted; (b) a "big" residue (E,K,R,I,L,M,F,Y,W); (h) hydrophobic residues (A,C,F,I,L,M,V,W,Y); (s) small residues (A,C,S,T,D,N,V,G,P); (u) "tiny" residues (G,A,S); (p) polar residues (D,E,H,K,N,Q,R,S,T); (c) charged residues (K,R,D,E,H); (_) negatively charged residues (D,E). The distances from the aligned regions to the protein termini and the distances between the conserved blocks, where more variable regions were omitted, are indicated by numbers. The secondary structure elements predicted using the PHD program, with the multiple alignment as the input (Rost and Sander 1994), is shown above the alignment; (E) extended conformation (β-strand); (H) α-helix. Species name abbreviations: (Aae *Aquifex aeolicus*; (Ab) *Azospirillum brasilense*; (Ac) *Acanthamoeba castellanii*; (Af) *Archaeoglobus fulgidus*; (At) *Arabidopsis thaliana*; (Bj) *Bradyrhizobium japonicum*; (Bs) Bacillus subtilis; (Clab) *Clostridium acetobutylicum*; (Cxb) *Coxiella burnetti*; (Ec) *Escherichia coli*: (Hi) *Haemophilus influenzae*; (Mj) *Methanococcus jannaschii*; (Mta) *Methanobacterium thermoautotrophicum*; (Mtu) *Mycobacterium tuberculosis*; (Ph) *Pyrococcus horikoshii*; (Pd) *Paracoccus denitrificans*; (Rc) *Rhodobacter capsulatus*; (Sp) Schizosaccharomyces pombe; (Ssp) *Synechocystis* sp.

and a *Synechocystis* protein, the UspA domain is fused to a cation transporter. In addition, fusions of the UspA domain to bacterial sensor proteins (e.g., KdpD) and to plant protein kinases were detected. The *E. coli* UspA protein has been reported to possess autophosphorylation activity (Freestone et al. 1997). Very recently, the x-ray structure of the *M. jannaschii* protein MJ0577 that we identified as a UspA homolog has been determined and the protein has been shown to tightly bind ATP (Zarembinski et al. 1998). It appears likely that the UspA superfamily proteins and domains are nucleotide-binding signal transducers that play a central regulatory role in both archaeal and bacterial cells.

Within the bacterial component of the euryarchaeal core gene set, 8 COGs with 4 members and 13 COGs with 3 members include archaeal proteins that contain the helix–turn–helix (HTH) domain and are predicted to function as transcription regulators. The conservation of these families in all or all but one of the archaea whose genomes have been sequenced, along with the existence of a number of more specific HTH protein families, emphasizes the combination of bacterial and eukaryotic features in the archaeal transcription machinery. Indeed, all archaeal RNA polymerase subunits and several basal transcription factors are most closely related to their eukaryotic counterparts, and some of them have no detectable orthologs in bacteria (Leffers et al. 1989; Puhler et al. 1989; Zillig et al. 1989; Langer et al. 1995; Bell and Jackson 1998; Bell et al. 1998). This is in a stark contrast with the bacterial affinities of the predicted transcriptional regulators; a detailed analysis of the archaeal transcription machinery and its evolutionary implications will be presented elsewhere (L. Aravind and E.V. Koonin, unpubl.).

Nearly all of the eukaryotic COGs in archaea, with only a few exceptions, consist of proteins involved in translation, modification of translation machinery components, transcription, replication, and repair. The present analysis resulted in the identification of previously undetected archaeal orthologs for several characteristically eukaryotic proteins that function in transcription and replication. Three such findings include the orthologs of the large subunit of DNA primase, the P30 subunit of RNAse P, and the nascent polypeptide-associated complex (NAC) α-subunit. The detection of the second eukaryotic-type primase subunit further supports the concept of a eukaryotic-type replication machinery in archaea but, in addition, is of particular interest given the existence of archaeal homologs of bacterial DNA G-type primases (Aravind et al. 1998).

The NACα family seems to be of special interest and we present this case in some detail. NACα is a multifunctional eukaryotic protein that is involved in translation and subcellular targeting of nascent polypeptides (Wang et al. 1995; Wickner 1995; Powers and Walter 1996) but it has been shown to function also as a transcription coactivator (Yotov et al. 1998). All archaea encode an apparent ortholog of NACα with a conserved domain organization; a further detailed sequence analysis showed that the amino-terminal domain of these proteins is distantly related to the general transcription factor BTF3 (Fig. 7A,B). Unexpectedly, we found that the small, carboxy-terminal domain of NACα and its archaeal counterparts, which is missing in BTF3, showed significant similarity to the distinct amino-terminal domain of the bacterial translation elongation factor EF-Ts and is likely to adopt the same structure (Fig. 7A,C,D,). The amino-terminal domain of EF-Ts has been implicated in its interaction with EF-Tu (Zhang et al. 1997); a similar interaction with the archaeal and eukaryotic elongation factors might be involved in the translational function of NACα. It appears likely that the ancestral form of NACα already performed a dual role in transcription and translation; as the result of our present analysis, each of these functions was mapped tentatively to a distinct domain.

As reported previously, bacterial homologs of some of the protein families that appeared to be confined to archaea and eukaryotes could be identified by structural comparison or through sequence searches using sensitive methods. An example of a structural comparison that has convincingly demonstrated the existence of a bacterial homolog (probably a highly diverged ortholog) of a archaeal–eukaryotic protein family is the relationship between the clamp subunits of DNA polymerases, that is, the eukaryotic proliferating cell nuclear antigen (PCNA), its highly conserved archaeal orthologs, and bacterial DNA polymerase β subunit (Krishna et al. 1994). More recently, bacterial homologs were detected by detailed sequence analyses for several translation factors that appeared to be exclusively archaeal–eukaryotic, such as eIF-5A whose highly diverged ortholog in bacteria is the elongation factor P (Tatusov et al. 1997; Kyrpides and Woese 1998). In the same vein, we observed that eukaryotic–archaeal initiation factor eIF6 contains a diverged ribosomal protein S1-type RNA-binding domain and thus, has homologs, although apparently not true orthologs, among bacterial proteins (data not shown). Other examples of eukaryotic–archaeal families, for which distant bacterial homologs become detectable as a result of detailed sequence analysis, are the transcription factors TFIIE and MBF1 (multiprotein bridging factor 1), in which we identified HTH domains (L. Aravind and E.V. Koonin, unpubl.). A number of other families, however, remained refractory to the detection of bacterial homologs despite extensive searches [e.g., several families of ribosomal proteins, translation initiation

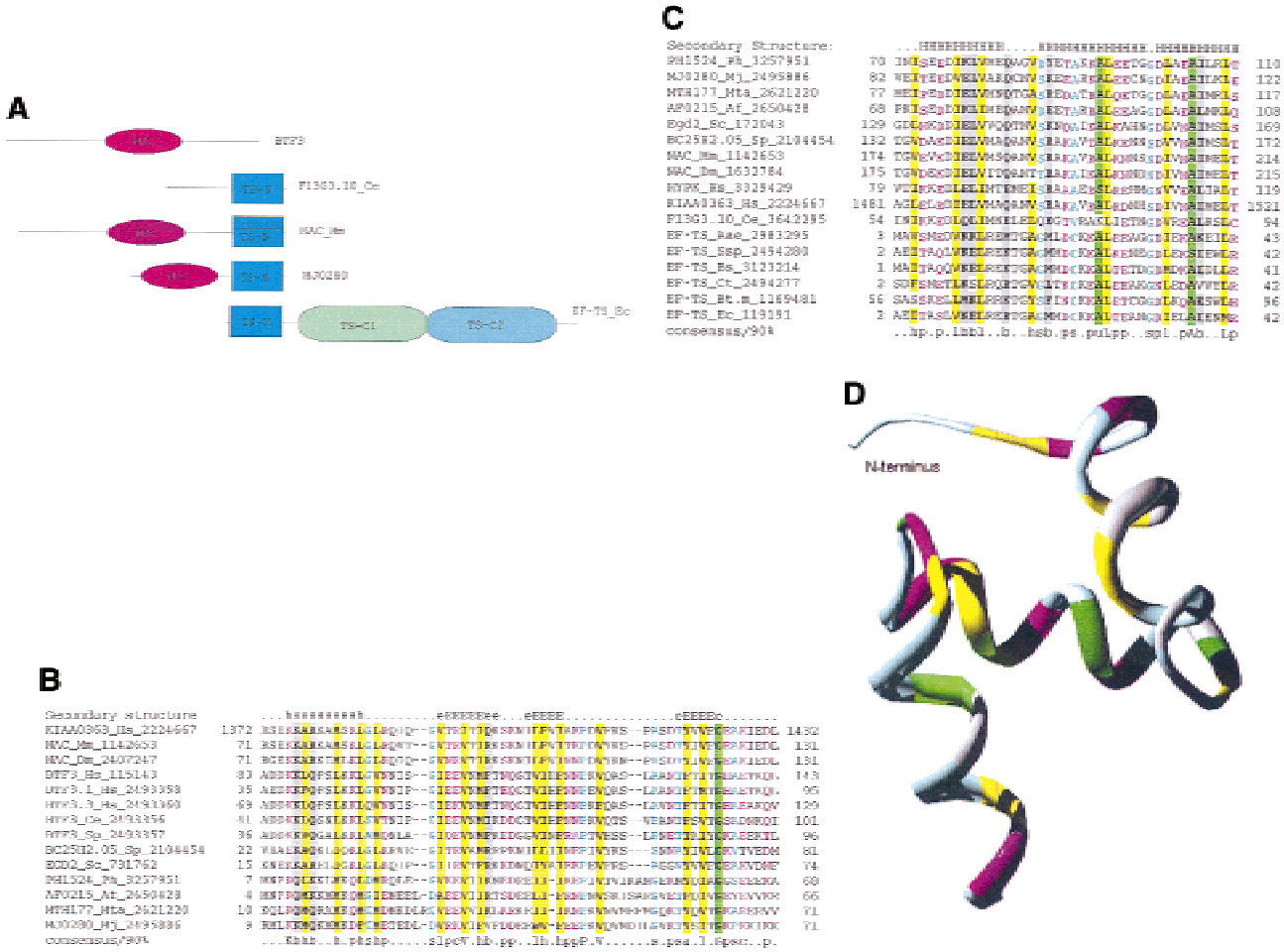


Figure 7 The NACα–BTF3 protein family—bifunctional proteins involved in both transcription and translation. (*A*) Domain architecture. (NAC) Amino-terminal domain of NACα that is conserved in BTF3 and is involved in transcription activation; (TS-N) amino-terminal domain of the bacterial translation factor Ts that is conserved in NACa and its orthologs; (TS-C1,2) two carboxy-terminal domains of Ts. Species name abbreviations: (Ce) *Caenorhabditis elegans*; (Ec) *Escherichia coli*; (Mm) *Mus musculus*. (*B*) Multiple alignment of the amino-terminal, BTF3-related domain. For details of the designation, see legend to Fig. 6. Species name abbreviations: (Af) *Achaeoglubus fulgidus*; (Bm) *Bombyx mori*; (Ce) *Caenorhabditis elegans*; (Dm) *Drosophila melanogaster*; (Hs) *Homo sapiens*; (Mj) *Methanocccocus jannaschii*; (Mm) *Mus musculus*; (Mta) *Methanobacterium thermoautotrophicum*; (Ph) *Pyrococcus horikoshii*; (Sc) *Saccharomyces cerevisiae*; (Sp) *Schizosaccharomyces pombe*. (*C*) Multiple alignment of the carboxy-terminal, EF-Ts-related domain. For details of the designations, see legend to Fig. 8. Species name abbreviations: (Aae) *Aquefex aeolicus*; (Af) *Archaeoglobus fulgidus*; (Bs) *Bacillus subtilis*; (Bt.m) bovine mitochondria; (Ce) *Caenorhabditis elegans*; (Ct) *Chlamydia trachomatis*; (Dm) *Drosophila melanogaster*; (Ec) *Escherichia coli*; (Hs) *Homo sapiens*; (Mj) *Methanococcus jannaschii*; (Mta) *Methanobacterium thermoautotrophicum*; (Ph) *Pyrococcus horikoshii*; (Sc) *Saccharomyces cerevisiae*; (Sp) *Schizosaccharomyces pombe*; (Ssp) *Synechocystis* sp. (*D*) Structure of the carboxy-terminal domain modeled using the amino-terminal domain of EF-Ts (Kawashima et al. 1996; PDB code 1efu) as a template. The conserved amino acid residues are colored as in *C*.

factor eIF-1β, three subunits (K, L, and N) of DNA-dependent RNA polymerase, and two DNA primase subunits].

Synapomorphies (shared-derived characters) Among Archaeal Protein Families and Archaea-Specific Family Expansions

Shared-derived characters present in the members of the given lineage to the exclusion of all other taxa under comparison (synapomorphies) are perhaps the most reliable indicators of monophyly that are free of the uncertainties that plague conventional methods of tree analysis, particularly when ancient evolutionary events are involved. At the level of conserved proteins, it is natural to define a synapomorphy as a family (COG) that does not have orthologs in other taxa. Typically, this conclusion can be reached either when there are no detectable homologs for a given family outside a particular clade, or when it has a unique domain architecture, with homologs found only for individual domains. According to these criteria, the 71 COGs that are represented in all four archaeal genomes but do not have detectable orthologs outside archaea (see Fig. 4B) should be considered archaeal synapomorphies (Table 2). The most obvious of these are the 32 universal archaeal COGs that do not have any detect-

able nonarchaeal homologs. Unfortunately, the information on the functions of these proteins is scant. A striking exception is the recently discovered archaeal DNA polymerase II (Uemori et al. 1997; Cann et al. 1998; Ishino et al. 1998) that is one of the most highly conserved proteins among the four archaea, but does not show any detectable similarity to other known polymerases (or any other proteins) except for a zinc finger domain.

In fact, however, the 71 COGs that have no obvious nonarchaeal orthologs mark only the lower bound of the number of synapomorphies. There is a considerable number of COGs that show readily definable unique features, although a traceable line of vertical descent seems to exist, suggesting orthologous relationships with bacterial or eukaryotic genes. Three examples in this category are translation elongation factor EF-1β, the small subunit of archaeal DNA polymerase II, and the archaeal ortholog of the eukaryotic repair protein ERCC4. The eukaryotic EF-1β all contain an additional domain that is homologous to glutathione *S*-transferases (Koonin et al. 1994) and is fused to the main domain that is conserved in the archaeal counterparts (Table 2; Fig. 8). In the case of the polymerase subunit and the ERCC4 protein, the archaeal counterparts contain the conserved sequence motifs that strongly suggest, respectively, a phosphohydrolase and a helicase activity; in eukaryotes, these motifs are disrupted, indicating that the respective enzymatic activities are abolished (Aravind and Koonin 1998; Aravind et al. 1999).

The most interesting synapomorphies are those

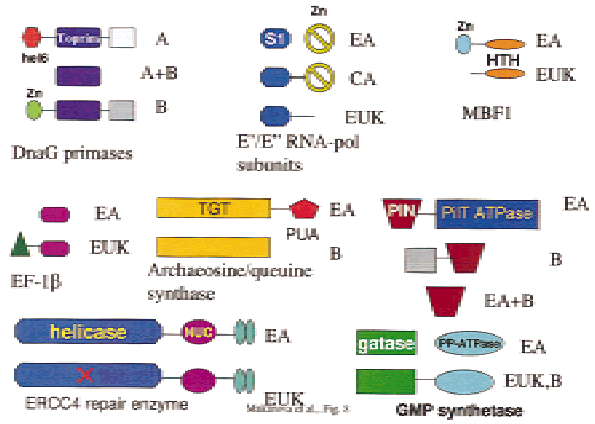


Figure 8 Examples of euryarchaeal synapomorphies—unique domain architectures in conserved euryarchaeal proteins. (Troprim) The catalytic domain conserved in primases and topoisomerases; (HTH) helix-turn-helix domain; (TGT) tRNA-guanine transglycosylase; (PIN) PilT-amino-terminal domain; (NUC) nuclease; (HEL6) helicase superfamily II motif 6; (gatase) glutamine aminotransferase; (PP-ATPase) PP-loop superfamily ATPase; (A) archaea (for domain architectures found in both Euryarchaeota and Crenarchaeota); (EA) Euryarchaeota; (CA) Crenarchaeota; (B) bacteria; (EUK) eukaryotes.

COGs that consist of proteins whose individual domains are conserved in other taxa but the domain architecture is unique (Table 2; Fig. 8). The recently described archaeal homologs of bacterial DnaG-type primases represent one such example where the primase domain is highly conserved in archaea and bacteria but the domains implicated in DNA binding are unrelated (Aravind et al. 1998). Table 2 and Figure 8 show additional instances of unique domain architectures in archaea. These include both archaea-specific domain fusions, as in the archaeal counterpart of the eukaryotic multiprotein bridging factor MBF1 (a transcriptional coactivator), and splitting of multidomain proteins into subunits encoded by distinct genes, as in the cases of the largest subunit of DNA-directed RNA polymerase and GMP synthetase. Interestingly, in *M. thermoautotrophicum* and *P. horikoshii* (but not in the other two archaeal species) the genes for the two GMP synthetase subunits are adjacent (Table 2), which strongly suggests that an ancestral gene that encoded the two-domain enzyme had been split early in archaeal evolution.

In addition to the protein families that are genuine synapomorphies, the uniqueness of a clade is defined by significant expansions of gene families that are less abundantly represented in other lineages. Several archaea-specific gene family expansions were detected as well as gene expansions confined to one or two archaeal species (Fig. 9). In only one case, that of ferredoxins, a correlation between a protein superfamily expansion and distinct features of archaeal physiology, such as iron-dependent respiration (Schafer et al. 1996a,b) and methanogenesis, seems obvious. Some of the other expanded families, [e.g., metal-dependent β-lactamase-like hydrolases (Aravind 1998)] include enzymes with versatile functions whose connection with the specifics of the archaeal lifestyle (if any) remains unclear.

Three expanded archaeal families include P-loop-containing ATPases, namely the RecA/RadA superfamily and two archaea-specific groups that have undergone species-specific amplification in *M. jannaschii* and *P. horikoshii*, respectively (Mj-type and Ph-type predicted ATPases). In the present analysis, the RecA/RadA ATPases formed two distinct COGs. One of these is represented by a single member in each of the four archaea and is orthologous to eukaryotic RadA-type ATPases. The second COG consists of different numbers of paralogs from each of the archaeal species and includes, in addition to typical RecA-like ATPases, forms with a duplicated ATPase domain, inactivated forms and fusions with other domains (e.g., GTPases; Aravind et al. 1999, L. Aravind, unpubl.). Interestingly, the members of this COG that contain the duplication of the ATPase domain are highly similar and apparently orthologous to a family of cyanobacterial RecA-

like ATPases at least one of which is involved in circadian clock regulation (Ishiura et al. 1998) (Fig. 10). Taken together with the observed inactivation and fusion with other domains, this functional connection may suggest that this second type of archaeal RecA-like ATPases is involved in signal transduction rather than repair. It appears likely that the duplication of the ATPase domain, which is unique within the RecA/RadA family of ATPases, occurred in one of the two lineages—euryarchaeota or cyanobacteria—with a subsequent horizontal gene transfer; the direction of transfer in this case is uncertain.

The archaea-specific family of Ph-type ATPases contains, in addition to the ATPase domain proper, a predicted HTH domain, whereas the distinct, although distantly related Mj-type family, contains a putative metal-binding motif (Koonin 1997; data not shown.). Given the presence of an HTH, the Ph-type family is most likely involved in ATP-dependent transcription regulation; by analogy, a similar role may be proposed for the Mj-type ATPases, the conserved metal-binding site being involved in DNA binding.

Other proteins and domains that are unusually abundant in archaea probably perform regulatory and signaling functions, such as the CBS domain (Bateman 1997; Ponting 1997) and the newly identified PIN domain (Figs. 9 and 11), although their functions are not understood in detail. The PIN (PilT amino terminus) domain is of particular interest. It is a compact domain that consists of ~100 amino acids, with the sequence conservation centered at two nearly invariant aspartates that cap predicted β-strands and two additional acidic residues found in the majority of PIN domains (Fig. 11). Each of the archaeal species encodes multiple stand-alone versions of the PIN domain as well as fusions with other domains; two of these fusions, namely those with the PilT-type ATPase domain and a C4 zinc finger, are archaeal synapomorphies (Figs. 9 and 11). PIN domains are sporadic and much less common in bacteria and eukaryotes except for the major expansion in *Mycobacteria* that appears to be independent of the archaeal expansion (Figs. 9, 11; L. Aravind, unpubl.). The function of the PIN domain is not known but a role in signaling appears likely given the presence of this domain in the plasmid-encoded transcriptional repressor StbB (Tabuchi et al. 1992) and the DIS3 family of eukaryotic proteins that are involved in mitosis regulation (Kinoshita et al. 1991; Noguchi et al. 1996; Shiomi et al. 1998). The yeast Dis3P is a 3′-5′ exonuclease, which is a subunit of the exosome (Mitchell et al. 1997), and consists of the PIN domain fused to a RNase II domain and a dsRNA-binding domain. The DIS3 proteins appear to perform a regulatory function mediated by their binding to the GTP–Ran and RCC1 proteins (Noguchi et al. 1996). Given the conservation of the PIN domain in DIS3 proteins from yeast to mammals (Fig. 11), it is likely to perform an important signaling function in all eukaryotes and, by implication, in archaea and bacteria.

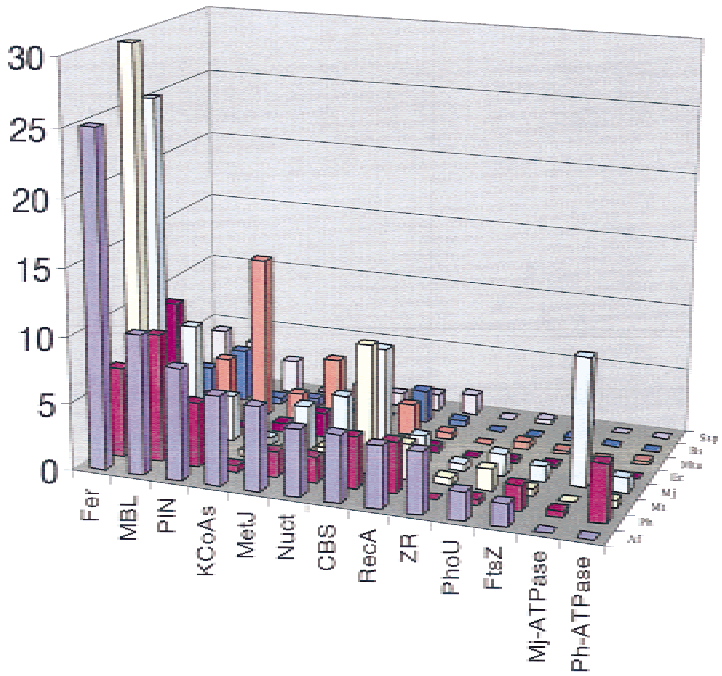


Figure 9 Specidic expansion of protein families in Euryarchaeota. The members of the families were identified using family-specific PSSMs are described in Methods. (Vertical axis) Number of proteins (domains) per 1000 genes. (Fer) Ferredoxins; (MBL) metallo-β-lactamase; (Nuct) "minimal" nucleotidyltransferase (Koonin et al. 1997; Aravind and Koonin 1999b); (PIN) PilT-amino-terminal domain (see text); (CBS) cystathionine-β-synthase domain; (FtsZ) GTPases involved in cell division, orthologs of the bacterial FtsZ protein; (RecA) superfamily ATPases; (MetJ) Arc/Met-repressor class of transcription regulators; (PhoU) regulators of phosphate uptake, orthologs of the bacterial PhoU protein; (KCoAS) ketoacyl-coenzyme A synthetases; (ZR) a distinct, archaea-specific family of predicted nucleic acid-binding protein containing the zinc ribbon domain (L. Aravind, unpubl.). Archaea; (Af) *A. fulgidus*; (Mj) *M. jannaschii*; (Mta) *M. thermoautotrophicum*; (Ph) *P. horikoshii*. Bacteria: (Bs) *Bacillus subtilis*; (Ec) *Escherichia coli*; (Mtu) *Mycobacterium tuberculosis*; (Ssp) *Synechocystis* sp.

Concluding Remarks

The analysis of the orthologous gene families (COGs) among the four completely sequenced archaeal genomes resulted in the delineation of the core gene set that is conserved in euryarchaeota. This core set includes only 31%–35% of the genes from each of the genomes but seems to account for most of the principal functions in genome replication, expression, and repair, as well as the majority of the reactions in several central metabolic pathways. This core gene set appears

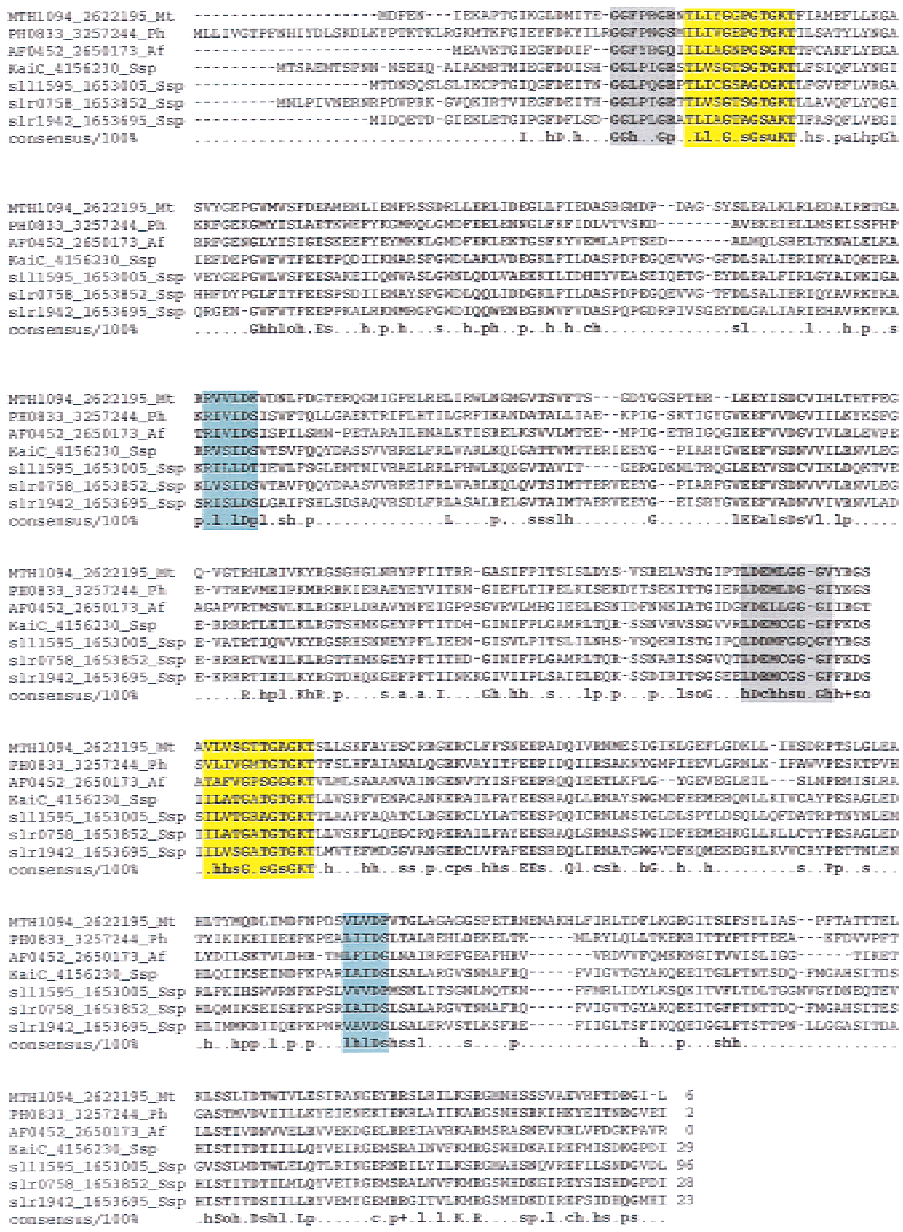


Figure 10 The unusual family of RecA-type ATPases with duplicated ATPase domains conserved in archaea and cyanobacteria. The designations are the same in Figs. 8 and 9 except that consensus residues are not highlighted. Instead, the highlighting shows the P-loop involved in the binding of the phosphates of ATP (yellow), the Mg-binding motif (cyan), and the glycine-rich amino-terminal motif that is typical of the RecA family of ATPases. Note the duplication of the predicted ATPase domain that encompasses all these motifs. KaiC is the cyanobacterial protein that is a component of the circadian rhythm system (Ishiura et al. 1998); the remaining proteins have not been experimentally characterized.

to have been relatively stable throughout the evolution of euryarchaeota. It defines the euryarchaeal clade through a number of synapomorphies–unique features, such as specific domain architectures of proteins that are conserved among the members of archaeal COGs but are not found outside the euryarchaea.

The evolution of the variable "shell" of the euryarchaeal genomes should have included multiple events other than vertical inheritance, namely horizontal gene exchange and lineage-specific gene loss, in archaeal evolution. Likely horizontal gene transfer may be manifest as nonorthologous gene displacement—apparent substitution of an unrelated or distantly related but functionally equivalent gene for the ancestral archaeal gene.

Generally, the comparison of the 4 archaeal genomes confirms the observations first made for *M. jannaschii* and *M. thermoautotrophicum*: the majority of archaeal proteins, particularly the metabolic enzymes and proteins involved in cell division and cell wall biogenesis, are most similar to their bacterial counterparts, and a minority, primarily proteins involved in genome replication and expression, most closely resemble their eukaryotic orthologs. The comparative analysis made it clear that the eukaryotic component belongs almost entirely to the families that are conserved in all four genomes, whereas much of the bacterial component comprises more variable families and species-specific genes. This might suggest a significant role of horizontal gene transfer from bacteria in the evolution of the euryarchaeota.

Comparative analysis of the four available genomes of euryarchaeota, aided by the availability of a number of complete bacterial genome sequences and one complete eukaryotic genome, provides some glimpses of archaeal evolution and the relationships between the three divisions of life. Once complete genomes of at least one crenarchaeon and some early-branching eukaryotes arrive, it will become possible to strive for a more coherent picture.

METHODS

Databases

The databases used in this study were the nonredundant (NR)

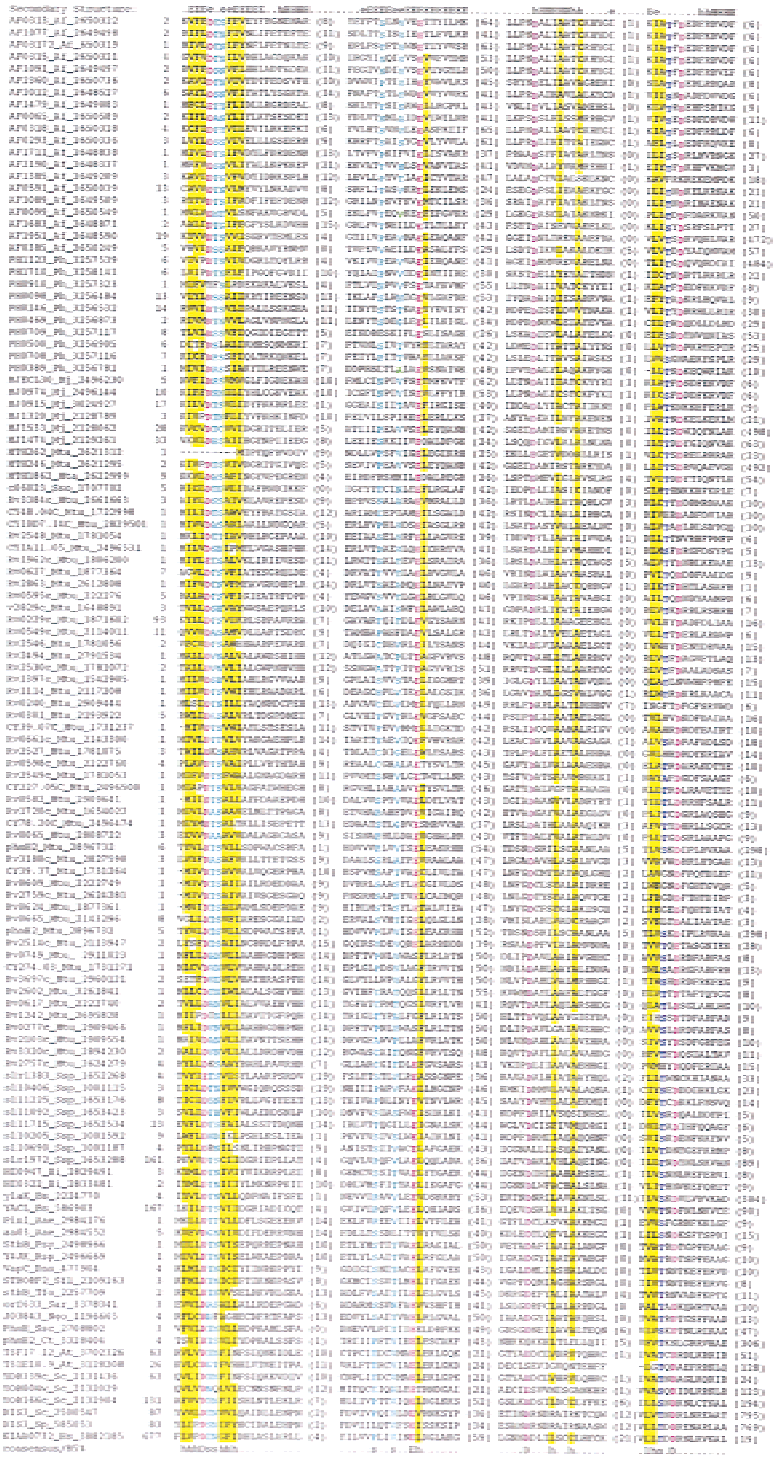


Figure 11 PIN—a novel domain superfamily with possible signaling function. For details for alignment construction and designations, see legend to Fig. 8. Species name abbreviations; (Aae) *Aquifex aeolicus*; (Af) *Archaeoglobus fulgidus*; (At) *Arabidopsis thaliana*; (Bs) *Bacillus subtilis*; (Ct) *Chlamydia trachomatis*; (Dno) *Dichelococcus nodosus*; (Hi) *Haemophilus influenzae; (Hs) Homo sapiens*; (Mj) *Methanococcus jannaschii*; (Mta) *Methanobacterium thermoautotrophicum*; (Mtu) *Mycobacterium tuberculosis*; (Ngo) Neisseria gonorrhoeae; (Ph) *Pyrococcus horikoshii*; (Psy) *Pseudomonas syringaea*; (Rsp) *Rhizobium* sp. NGR234; (Sar) *Sphingomonas aromaticivorans*; (Sfl) *Shigella flexner*; (Sc) *Saccaromyces cerevisiae*; (Sp) *Schizosaccharomyces pombe*; (Ssc) *Synechococcus* PCC7002; (Sso) *Sulfolobus solfataricus*; (Ssp) *Synechocystis* sp; (Tfo) *Thiobacillus ferrooxidans*.

database and a separate database containing the protein sequences encoded in the complete genomes of four archaea, namely *M. jannaschii* (Bult et al. 1996), *M. thermoautotrophicum* (Smith et al. 1997), *A. fulgidus* (Klenk et al. 1997), and *P. horikoshii* (Kawarabayasi et al. 1998a,b). The archaeal protein complements and the complete nucleotide sequences of the archaeal genomes were extracted from the Genomes division of Entrez.

Database Searches

The protein sequence database searches were performed using the gapped BLAST program and the PSI-BLAST program (Altschul et al. 1997). The PSI-BLAST program constructs a position-specific matrix (PSSM) from a multiple alignment generated from the BLAST hits above a certain expectation value (e-value) and carries out iterative database searches using the PSSM as the query (Altschul et al. 1997; Altschul and Koonin 1998). PSI-BLAST also has the capability to save the PSSM after a user-defined number of iterations or at convergence and to reuse for searching another database (Wolf et al. 1999). The estimates of statistical significance of the PSI-BLAST results are based on the extreme value distribution statistics originally developed by Karlin and Altschul for local alignments without gaps (Karlin and Altschul 1990; Karlin et al. 1991) and subsequently shown to apply to gapped alignments as well (Altschul and Gish 1996; Altschul et al. 1997). There is no analytical proof of the applicability of the Karlin–Altschul statistics to searches that use PSSM as queries, but extensive computer simulations showed a nearly perfect fit of the score distribution produced searches to the extreme value distribution (Altschul et al. 1997). Therefore, e-values reported for each retrieved sequence at the point when its alignment with the query exceeds the cutoff for the first time should be considered reliable estimates of the statistical significance of the observed similarity. Clearly, after a sequence is included in the model, e-values reported for it (and its closely related homologs) in subsequent iterations become inflated and do not represent accurately the statistical significance (Altschul and Koonin 1998). All reported e-values are for the first appearance of the given sequence above the cutoff.

The main source of artifacts that arise in database searches and are inevitably amplified in PSI-BLAST iterations are regions of low compositional complexity in protein sequences that typically correspond to nonglobular domains (Wootton 1994). To avoid such artifacts, database searches were routinely run after masking the low complexity regions in the query sequences using the SEG program with default parameters (Wootton

and Federhen 1996). However, because masking may also prevent the detection of subtle but functionally and evolutionarily important sequence similarities, filtering for low complexity was omitted in case-by-case analyses aimed at the detection of distant homologs.

The current default e-value cutoff for PSI-BLAST to include a sequence in the PSSM for use in the next iteration is 0.001. However, the original evaluation of the accuracy of PSI-BLAST and a number of subsequent analyses, including both large-scale benchmarking experiments and detailed case studies, have shown that an e-value of 0.01 (and in some cases, even higher e-values) is an appropriate cutoff for PSI-BLAST provided that (1) regions of low complexity in the query are masked before the search, and (2) the search results are subsequently examined for the conservation of sequence motifs that are typical of the particular protein superfamily. Accordingly, the cutoff of 0.01 was used as the default for PSI-BLAST searches in this work. The outcome of the analysis performed using PSI-BLAST critically depends on the optimal choice of the queries used to seed the iterative search (Aravind and Koonin 1999a). Therefore, all protein families that were analyzed in detail were investigated using multiple starting points. All PSI-BLAST outputs were manually examined for the conservation of characteristic sequence motifs to corroborate the relevance of the results and facilitate the prediction of protein functions.

Construction and Analysis of COGs of Proteins

After comparing the archaeal protein set to itself using the gapped BLAST program, conserved archaeal families that consist of likely orthologs, termed COGs, were delineated using the previously described approach (Tatusov et al. 1997; Koonin et al. 1998). Briefly, this procedure first identifies and clusters obvious paralogs within each proteome; that is, those proteins that show a greater similarity to each other than to any protein from the other proteomes. At the next step, for each protein or group of paralogs, the most similar protein in each of the other proteomes is found, consistent triangles of such intergenomic best hits are identified, and triangles with a common side are merged to form COGs.

Multiple alignments were constructed for each potential COG using the ClustalW program (Thompson et al. 1994); the default parameters for ClustaW, namely the BLOSUM62 matrix for amino acid residue comparison, gap opening penalty 10, and gap extension penalty 0.1 were used. The resulting multiple alignments were examined, in conjunction with the BLAST search outputs, to identify proteins that contain two or more distinct, independently evolving regions. The distinguishing feature of such independently evolving units in proteins is that they are fused in some species to form a single protein, but in other species are encoded by two distinct genes, resulting in independent proteins (Doolittle and Bork 1993; Doolittle 1995; Riley and Labedan 1997).

Typically, when the respective three-dimensional structures are available, the independently evolving regions are recognized as sequence cognates of compact structural units, and therefore, these regions are frequently called domains, whereas proteins containing more than one such region are called multidomain proteins. However, a one-to-one correspondence between independently evolving regions of proteins and domains defined as fundamental units of three-dimensional structure (Branden and Tooze 1991) may or may not exist, as a single independently evolving region may contain more than one domain. In our analysis, independently evolving regions of proteins were recognized on the basis of statistically significant sequence similarity (typically, e-value below 0.01) detected using the BLAST or PSI-BLAST programs; the recognition of such regions is facilitated by use of the graphical output of the database search implemented in WWW-BLAST (http://www.ncbi.nlm.nih.gov/BLAST). Multidomain proteins may artificially connect unrelated single-domain proteins into a cluster (Watanabe and Otsuka 1995; Koonin et al. 1996b; Riley and Labedan 1997). Clusters that appeared to contain two COGs artificially merged, because of the presence of multidomain proteins, were manually split into single-domain COGs.

These procedures resulted in the identification of COGs that included at least three archaeal species. In addition, all symmetrical intergenomic best hits (Tatusov et al. 1997) between proteins not included in this set of COGs were analyzed to identify COGs that contained only two species. The protein sequences from each COG were compared to the rest of the archaeal proteins using the PSI-BLAST program, which was run for four iterations, to detect possible distant, nonorthologous homologs of the COGs encoded in the archaeal genomes. In addition, the protein sequences from the COGs including three or two archaeal species were compared to the complete sequences of the remaining archaeal genomes translated in all six reading frames using the gapped version of the TBLASTN program (Altschul et al. 1990, 1997), to detect possible orthologs that might have been missed in the original translation of the genome sequences.

The archaeal protein sequences included in the COGs were compared to the NR database using the PSI-BLAST program (four iterations), to detect orthologs and nonorthologous homologs in other taxa, even in cases of low sequence conservation. The search outputs were analyzed using the Tax_Break and Tax_Collector programs of the SEALS package (Walker and Koonin 1997), to evaluate the phylogenetic distribution of homologs for each COG. The Tax_Break program outputs the complete taxonomic breakdown of database hits above the chosen cutoff (e-value of 0.01 in this work) and the Tax_Collector program outputs the lineage-specific best hits using the taxonomy tree structure embedded in the Entrez system. The alignments of archaeal proteins with most similar proteins from different taxa were examined manually to assess the orthologous relationships (or lack thereof). The assignment of likely orthologs was based on a combination of statistical significance of the best lineage-specific hits and the conservation of domain architecture (Tatusov et al. 1996, 1997).

The PSI-BLAST searches with the same settings were performed for the archaeal proteins not included in the COGs. To enumerate the members of large protein or domain families encoded in the archaeal genomes, a profile for each family was developed using the PSI-BLAST program and run as a query against the archaeal protein sequence database using the e-value of 0.01 (adjusted to the size of the NR database) as the cutoff (Aravind et al. 1998; Chervitz et al. 1998; Wolf et al. 1999).

Other Methods for Protein Sequence and Structure Analysis

Protein secondary structure prediction on the basis of a multiple sequence alignment was carried out using the PHD program (Rost and Sander 1994). Homology modeling of protein structures was performed using the ProMod program (Peitsch

1996). Protein databank (PDB) files were visualized using SWISS-PDB viewer version 2.6 (Peitsch 1996).

Availability of the Complete Results

The complete, annotated list of archaeal COGs is available at http://www.ncbi.nlm.nih.gov/CBBresearch/Koonin/COGS/Archaea. This list is also available, together with multiple alignments for each of the COGs at ftp://www.ncbi.nlm.nih.gov/pub/koonin/Archaea.

ACKNOWLEDGMENTS

K.M. is supported by U.S. Department of Energy OBER grant DE-FG02-98ER62583.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, E. and L. Frank. 1980. Metabolism of proline and the hydroxyprolines. *Annu. Rev. Biochem.* **49:** 1005–1061.

Altschul, S.F. and W. Gish. 1996. Local alignment statistics. *Methods Enzymol.* **266:** 460–480.

Altschul, S.F. and E.V. Koonin. 1998. PSI-BLAST—A tool for making discoveries in sequence databases. *Trends Biochem. Sci.* **23:** 444–447.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

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

Aravind, L. 1998. An evolutionary classification of the metallo-beta lactamase fold proteins. *In Silico Biol.* **1:** 8.

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

———. 1998b. The HD domain defines a new superfamily of metal-dependent phosphohydrolases. *Trends Biochem. Sci.* **23:** 469–472.

———. 1999a. Gleaning non-trivial structural, functional, and evolutionary information about proteins by iterative database searches. *J. Mol. Evol.* **287:** 1023–1040.

———. 1999b. DNA polymerase beta-like nucleotidyltransferase superfamily: Identification of three new families, classification and evolutionary history. *Nucleic Acids Res.* **27:** 1609–1618.

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

Aravind, L., D.R. Walker, and E.V. Koonin. 1999. Conserved domains in DNA repair proteins and evolution of repair systems. *Nucleic Acids Res.* **27:** 1223–1242.

Bateman, A. 1997. The structure of a domain common to archaebacteria and the homocystinuria disease protein. *Trends Biochem. Sci.* **22:** 12–13.

Bell, S.D. and S.P. Jackson. 1998. Transcription and translation in Archaea: A mosaic of eukaryal and bacterial features. *Trends Microbiol.* **6:** 222–228.

Bell, S.D., C. Jaxel, M. Nadal, P.F. Kosa, and S.P. Jackson. 1998. Temperature, template topology, and factor requirements of archaeal transcription. *Proc. Natl. Acad. Sci.* **95:** 15218–15222.

Branden, C. and J. Tooze. 1991. *Introduction to protein structure*. Garland Publishing, Inc., New York-London.

Brown, J.R. and W.F. Doolittle. 1997. Archaea and the prokaryote-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.* **61:** 456–502.

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

Cann, I.K.O., K. Komori, H. Toh, S. Kanai, and Y. Ishino. 1998. A heterodimeric DNA polymerase: Evidence that members of euryarchaeota possess a distinct DNA polymerase. *Proc. Natl. Acad. Sci.* **95:** 14250–14255.

Chervitz, S.A., L. Aravind, G. Sherlock, C.A. Ball, E.V. Koonin, S.S. Dwight, M.A. Harris, K. Dolinski, S. Mohr, T. Smith et al. 1998. Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* **282:** 2022–2028.

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

Doolittle, R.F. and P. Bork. 1993. Evolutionarily mobile modules in proteins. *Sci. Am.* **269:** 50–56.

Doolittle, W.F. and J.M. Logsdon Jr. 1998. Archaeal genomics: Do archaea have a mixed heritage? *Curr. Biol.* **8:** R209–211.

Eberhardt, S., S. Korn, P. Lottspelch, and A. Bacher. 1997. Biosynthesis of riboflavin synthase of methenobacterium thermoantotrophium. J. Bacteriol. **179:** 2938–2940.

Edgell, D.R. and W.F. Doolittle. 1997. Archaea and the origin(s) of DNA replication proteins. *Cell* **89:** 995–998.

Feng, D.F., G. Cho, and R.F. Doolittle. 1997. Determining divergence times with a protein clock: Update and reevaluation. *Proc. Natl. Acad. Sci.* **94:** 13028–13033.

Freestone, P., T. Nystrom, M. Trinei, and V. Norris. 1997. The universal stress protein, UspA, of Escherichia coli is phosphorylated in response to stasis. *J. Mol. Biol.* **274:** 318–324.

Gogarten, J.P., H. Kibak, P. Dittrich, L. Taiz, E.J. Bowman, B.J. Bowman, M.F. Manolson, R.J. Poole, T. Date, T. Oshima et al. 1989a. Evolution of the vacuolar H+–ATPase: Implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci.* **86:** 6661–6665.

Gogarten, J.P., T. Rausch, P. Bernasconi, H. Kibak, and L. Taiz. 1989b. Molecular evolution of H+–ATPases. I. Methanococcus and Sulfolobus are monophyletic with respect to eukaryotes and Eubacteria. *Z. Naturforsch. Teil C* **44:** 641–650.

Gonzalez, J.M., Y. Masuchi, F.T. Robb, J.W. Ammerman, D.L. Maeder, M. Yanagibayashi, J. Tamaoka, and C. Kato. 1998. Pyrococcus horikoshii sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal vent at the Okinawa Trough. *Extremophiles* **2:** 123–130.

Gribaldo, S. and P. Cammarano. 1998. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery [In Process Citation]. *J. Mol. Evol.* **47:** 508–516.

Ishino, Y., K. Komori, I.K. Cann, and Y. Koga. 1998. A novel DNA polymerase family found in Archaea. *J. Bacteriol.* **180:** 2232–2236.

Ishiura, M., S. Kutsuna, S. Aoki, H. Iwasaki, C.R. Andersson, A. Tanabe, S.S. Golden, C.H. Johnson, and T. Kondo. 1998. Expression of a gene cluster kaiABC as a circadian feedback process in cyanobacteria. *Science* **281:** 1519–1523.

Iwabe, N., K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci.* **86:** 9355–9359.

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

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

Kawarabayasi, Y., M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S. Baba, H. Kosugi, A. Hosoyama et al. 1998a. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res.* **5:** 55–76.

———. 1998b. Complete sequence and gene organization of the

genome of a hyper-thermophilic archaebacterium, Pyrococcus horikoshii OT3 (suppl.). *DNA Res.* **5:** 147–155.
Kawashima, T., C. Berthet-Colominas, M. Wulff, S. Cusack, and R. Leberman. 1996. The structure of the Escherichia coli EF-Tu.EF-Ts complex at 2.5 A resolution. *Nature* **379:** 511–518.
Keeling, P.J., R.L. Charlebois, and W.F. Doolittle. 1994. Archaebacterial genomes: Eubacterial form and eukaryotic content. *Curr. Opin. Genet. Dev.* **4:** 816–822.
Kinoshita, N., M. Goebl, and M. Yanagida. 1991. The fission yeast dis3+ gene encodes a 110-kDa essential protein implicated in mitotic control. *Mol. Cell. Biol.* **11:** 5839–5847.
Klenk, H.P., R.A. Clayton, J.F. Tomb, O. White, K.E. Nelson, K.A. Ketchum, R.J. Dodson, M. Gwinn, E.K. Hickey, J.D. Peterson et al. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390:** 364–370.
Koonin, E.V. 1997. Evidence for a family of archaeal ATPases. *Science* **275:** 1489–1490.
Koonin, E.V. and M.Y. Galperin. 1997. Prokaryotic genomes: The emerging paradigm of genome-based microbiology. *Curr. Opin. Genet. Dev.* **7:** 757–763.
Koonin, E.V., A.R. Mushegian, R.L. Tatusov, S.F. Altschul, S.H. Bryant, P. Bork, and A. Valencia. 1994. Eukaryotic translation elongation factor 1 gamma contains a glutathione transferase domain—Study of a diverse, ancient protein superfamily using motif search and structural modeling. *Protein Sci.* **3:** 2045–2054.
Koonin, E.V., A.R. Mushegian, and P. Bork. 1996a. Non-orthologous gene displacement *Trends Genet.* **12:** 334–336.
Koonin, E.V., R.L. Tatusov, and K.E. Rudd. 1996b. Protein sequence comparison at genome scale. *Methods Enzymol.* **266:** 295–322.
Koonin, E.V., A.R. Mushegian, M.Y. Galperin, and D.R. Walker. 1997. Comparison of archaeal and bacterial genomes: Computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.* **25:** 619–637.
Koonin, E.V., R.L. Tatusov, and M.Y. Galperin. 1998. Beyond complete genomes: From sequence to structure and function. *Curr. Opin. Struct. Biol.* **8:** 355–363.
Krishna, T.S., X.P. Kong, S. Gary, P.M. Burgers, and J. Kuriyan. 1994. Crystal structure of the eukaryotic DNA polymerase processivity factor PCNA. *Cell* **79:** 1233–1243.
Kyrpides, N.C. and C.R. Woese. 1998. Universally conserved translation initiation factors. *Proc. Natl. Acad. Sci.* **95:** 224–228.
Langer, D., J. Hain, P. Thuriaux, and W. Zillig. 1995. Transcription in archaea: Similarity to that in eucarya. *Proc. Natl. Acad. Sci.* **92:** 5768–5772.
Leffers, H., F. Gropp, F. Lottspeich, W. Zillig, and R.A. Garrett. 1989. Sequence, organization, transcription and evolution of RNA polymerase subunit genes from the archaebacterial extreme halophiles *Halobacterium halobium* and *Halococcus morrhuae*. *J. Mol. Biol.* **206:** 1–17.
Mitchell, P., E. Petfalski, A. Shevchenko, M. Mann, and D. Tollervey. 1997. The exosome: A conserved eukaryotic RNA processing complex containing multiple 3′ → 5′ exoribonucleases. *Cell* **91:** 457–466.
Mushegian, A.R. and E.V. Koonin. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci.* **93:** 10268–10273.
Neidhardt, F.C., R. Curtiss III, J.L. Ingraham, E.C.C. Lin, K.B. Low, B. Magasanik, W.S. Reznekoff, M. Riley, M. Schaechter, and H.E. Umbarger, eds. 1996. Escherichia coli *and* Salmonella. *Cell. molecular biology*. 2nd ed. ASM Press, Washington, D.C.
Noguchi, E., N. Hayashi, Y. Azuma, T. Seki, M. Nakamura, N. Nakashima, M. Yanagida, X. He, U. Mueller, S. Sazer et al.. 1996. Dis3, implicated in mitotic control, binds directly to Ran and enhances the GEF activity of RCC1. *EMBO J.* **15:** 5595–5605.
Olsen, G.J., C.R. Woese, and R. Overbeek. 1994. The winds of (evolutionary) change: Breathing new life into microbiology. *J. Bacteriol.* **176:** 1–6.
Pace, N.R. 1997. A molecular view of microbial diversity and the biosphere. *Science* **276:** 734–740.
Pace, N.R., G.J. Olsen, and C.R. Woese. 1986. Ribosomal RNA phylogeny and the primary lines of evolutionary descent. *Cell* **45:** 325–326.
Peitsch, M.C. 1996. ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem. Soc. Trans.* **24:** 274–279.
Ponting, C.P. 1997. CBS domains in CIC chloride channels implicated in myotonia and nephrolithiasis (kidney stones). *J. Mol. Med.* **75:** 160–163.
Powers, T. and P. Walter. 1996. The nascent polypeptide-associated complex modulates interactions between the signal recognition particle and the ribosome. *Curr. Biol.* **6:** 331–338.
Puhler, G., H. Leffers, F. Gropp, P. Palm, H.P. Klenk, F. Lottspeich, R.A. Garrett, and W. Zillig. 1989. Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. *Proc. Natl. Acad. Sci.* **86:** 4569–4573.
Riley, M. and B. Labedan. 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. and C. Sander. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19:** 55–72.
Sans, N., U. Schindler, and J. Schroder. 1988. Ornithine cyclodeaminase from Ti plasmid C58: DNA sequence, enzyme properties and regulation of activity by arginine. *Eur. J. Biochem.* **173:** 123–130.
Schafer, G., W. Purschke, and C.L. Schmidt. 1996a. On the origin of respiration: Electron transport proteins from archaea to man. *FEMS Microbiol. Rev.* **18:** 173–188.
Schafer, G., W.G. Purschke, M. Gleissner, and C.L. Schmidt. 1996b. Respiratory chains of archaea and extremophiles. *Biochim. Biophys. Acta* **1275:** 16–20.
Shiomi, T., K. Fukushima, N. Suzuki, N. Nakashima, E. Noguchi, and T. Nishimoto. 1998. Human dis3p, which binds to either GTP- or GDP-Ran, complements Saccharomyces cerevisiae dis3. *J. Biochem. (Tokyo)* **123:** 883–890.
Smith, D.R., L.A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert et al. 1997. Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: Functional analysis and comparative genomics. *J. Bacteriol.* **179:** 7135–7155.
Tabuchi, A., Y.N. Min, D.D. Womble, and R.H. Rownd. 1992. Autoregulation of the stability operon of IncFII plasmid NR1. *J. Bacteriol.* **174:** 7629–7634.
Tatusov, R.L., A.R. Mushegian, P. Bork, N.P. Brown, W.S. Hayes, M. Borodovsky, K.E. Rudd, and E.V. Koonin. 1996. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.* **6:** 279–291.
Tatusov, R.L., E.V. Koonin, and D.J. Lipman. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.
Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Thomson, G.J., G.J. Howlett, A.E. Ashcroft, and A. Berry. 1998. The dhnA gene of Escherichia coli encodes a class I fructose bisphosphate aldolase. *Biochem. J.* **331:** 437–445.
Uemori, T., Y. Sato, I. Kato, H. Doi, and Y. Ishino. 1997. A novel DNA polymerase in the hyperthermophilic archaeon, Pyrococcus furiosus: Gene cloning, expression, and characterization. *Genes Cells* **2:** 499–512.
Walker, D.R. and E.V. Koonin. 1997. SEALS: A system for easy analysis of lots of sequences. *ISMB* **5:** 333–339.
Wang, S., H. Sakai, and M. Wiedmann. 1995. NAC covers ribosome-associated nascent chains thereby forming a protective environment for regions of nascent chains just emerging from the peptidyl transferase center. *J. Cell. Biol.* **130:** 519–528.
Watanabe, H. and J. Otsuka. 1995. A comprehensive representation

of extensive similarity linkage between large numbers of proteins. *Comput. Appl. Biosci.* **11:** 159–166.
Wickner, W. 1995. The nascent-polypeptide-associated complex: Having a "NAC" for fidelity in translocation. *Proc. Natl. Acad. Sci.* **92:** 9433–9434.
Woese, C.R. 1994. There must be a prokaryote somewhere: Microbiology's search for itself. *Microbiol. Rev.* **58:** 1–9.
Woese, C.R. and R. Gupta. 1981. Are archaebacteria merely derived prokaryotes? *Nature* **289:** 95–96.
Woese, C.R., L.J. Magrum, and G.E. Fox. 1978. Archaebacteria. *J. Mol. Evol.* **11:** 245–251.
Woese, C.R., O. Kandler, and M.L. Wheelis. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci.* **87:** 4576–4579.
Wolf, Y.I., S.E. Brenner, P.A. Bash, and E.V. Koonin. 1999. Distribution of protein folds in the three superkingdoms of life. *Genome Res.* **9:** 17–26.
Wootton, J.C. 1994. Non-globular domains in protein sequences: Automated segmentation using complexity measures. *Comput. Chem.* **18:** 269–285.
Wootton, J.C. and S. Federhen. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266:** 554–571.
Yotov, W.V., A. Moreau, and R. St-Arnaud. 1998. The alpha chain of the nascent polypeptide-associated complex functions as a transcriptional coactivator. *Mol. Cell. Biol.* **18:** 1303–1311.
Zarembinski, T.I., L.W. Hung, H.J. Mueller-Dieckmann, K.K. Kim, H. Yokota, R. Kim, and S.H. Kim. 1998. Structure-based assignment of the biochemical function of a hypothetical protein: A test case of structural genomics. *Proc. Natl. Acad. Sci.* **95:** 15189–15193.
Zhang, Y., V. Sun, and L.L. Spremulli. 1997. Role of domains in Escherichia coli and mammalian mitochondrial elongation factor Ts in the interaction with elongation factor Tu. *J. Biol. Chem.* **272:** 21956–21963.
Zillig, W. 1991. Comparative biochemistry of Archaea and Bacteria. *Curr. Opin. Genet. Dev.* **1:** 544–551.
Zillig, W., H.P. Klenk, P. Palm, G. Puhler, F. Gropp, R.A. Garrett, and H. Leffers. 1989. The phylogenetic relations of DNA-dependent RNA polymerases of archaebacteria, eukaryotes, and eubacteria. *Can. J. Microbiol.* **35:** 73–80.

Received January 7, 1999; accepted in revised form May 27, 1999.