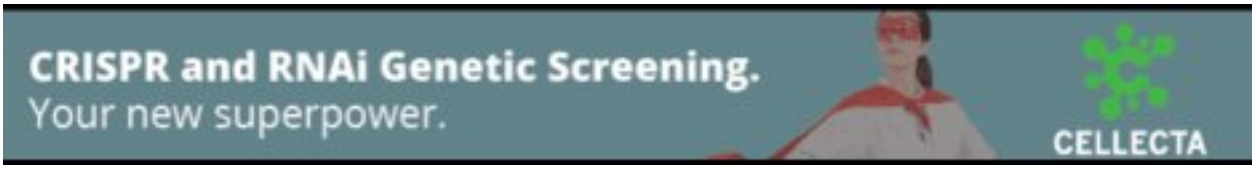
The Genetic Core of the Universal Ancestor

J. Kirk Harris, Scott T. Kelley, George B. Spiegelman, et al.

Genome Res. 2003 13: 407-412
Access the most recent version at doi:10.1101/gr.652803



Letter

The Genetic Core of the Universal Ancestor

J. Kirk Harris,[1,2,4] Scott T. Kelley,[1,4] George B. Spiegelman,[3] and Norman R. Pace[1,5]

[1]Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado 80309-0347, USA; [2]Graduate Group in Microbiology, University of California, Berkeley, Berkeley, California 94720, USA; [3]Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z3

Molecular analysis of conserved sequences in the ribosomal RNAs of modern organisms reveals a three-domain phylogeny that converges in a universal ancestor for all life. We used the Clusters of Orthologous Groups database and information from published genomes to search for other universally conserved genes that have the same phylogenetic pattern as ribosomal RNA, and therefore constitute the ancestral genetic core of cells. Our analyses identified a small set of genes that can be traced back to the universal ancestor and have coevolved since that time. As indicated by earlier studies, almost all of these genes are involved with the transfer of genetic information, and most of them directly interact with the ribosome. Other universal genes have either undergone lateral transfer in the past, or have diverged so much in sequence that their distant past could not be resolved. The nature of the conserved genes suggests innovations that may have been essential to the divergence of the three domains of life. The analysis also identified several genes of unknown function with phylogenies that track with the ribosomal RNA genes. The products of these genes are likely to play fundamental roles in cellular processes.

Phylogenetic studies of ribosomal RNA (rRNA) revolutionized our understanding of biological diversity by revealing that modern organisms fall into three phylogenetic domains: Archaea, Bacteria, and Eucarya (Woese and Fox 1977; Woese et al. 1990). rRNA sequence information in principle is well suited for determining deep phylogenetic relationships for several reasons. The rRNA sequences occur in all organisms, they have evolved at a sufficiently slow rate to retain phylogenetic information between distantly related organisms, and the rRNA genes have undergone limited or no horizontal transfer (i.e., transfer between distantly related organisms; Asai et al. 1999). Since the original description of the three-domain phylogeny, correlations of biochemical properties between organisms and data from genomic sequences have lent support to this classification of life (Woese et al. 1990; Wettach et al. 1995; Brown et al. 2001b).

At the same time, it also has become evident that many genes do not exhibit the same phylogenetic pattern as rRNA genes. Data from complete genomic sequences and phylogenetic studies of particular genes have revealed that genomes contain many genes that have undergone horizontal as well as vertical evolutionary change (Brown and Doolittle 1997). Moreover, a large number of genes appear to have been lost from, or never acquired by, various lineages over evolutionary time (Snel et al. 2002). Although gene loss or gain and horizontal transfer are common themes in evolution, phylogenetic analyses nonetheless have identified a number of genes in the nucleic acid-based information-processing pathway that have phylogenetic histories congruent with that of rRNA. For instance, the phylogenetic relationships among the core subunits of the DNA-dependent RNA polymerases, or most ribosomal protein genes, are the same as those seen in phylogenetic analyses of rRNA sequences (Iwabe et al. 1991; Klenk et al. 1993; Liao and Dennis 1994). Additionally, recent studies of concatenated datasets recovered the three-domain topology even when component members analyzed separately clearly demonstrated lateral transfers between organisms (Brown et al. 2001a). Collectively, the results indicate that the phylogenetic pattern of rRNA is representative of the evolutionary history of some portion of cellular components, which we term the 'genetic core.'

Although it is known that some cellular genes show the same phylogenetic patterns as rRNA, the purpose of the present study was to determine the entire set of universal genes with this property; this set constitutes a 'genetic core' of the known cellular lines of descent. Abundant new sequence information from a rapidly expanding database of genome sequences allows a more complete assessment of the genes that comprise such a genetic core that traces its ancestry back to the last common ancestor (LCA) of life. We used the Clusters of Orthologous Groups of proteins (COG) database (Tatusov et al. 2001) to search for constituents of the genetic core by identifying the universally conserved set of related genes that have the same phylogenetic history as rRNA. If a gene that is universally present in cells shares the same phylogenetic history as rRNA, two important properties of the gene can be inferred: (1) The gene occurred in the LCA and is not present in all organisms, as a result of subsequent horizontal transfer between lineages; and (2) the gene has resisted both nonorthologous displacement and extensive amino acid substitution since that time of the LCA. We note that this analysis will not yield a minimal genome for the LCA, because it should focus primarily on the mechanisms of the universal function of transfer of genetic information.

The analyses presented here were based exclusively on fully sequenced genomes and have two primary advantages over single-gene surveys. First, the complete set of genes from the organisms being examined is known, which allowed for a comprehensive analysis of gene coevolution. Second, the absence of a gene in an analysis of complete genome sequences is not a negative result; rather, it is a finding that the gene is truly not present in the organism. This contrasts with PCR- or

[4]These authors contributed equally to this work.
[5]Corresponding author.
E-MAIL nrpace@colorado.edu; FAX (303) 492-7744.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.652803. Article published online before print in February 2003.

homology-based analyses of particular genes, where a negative result is ambiguous.

RESULTS

Of the roughly 3100 COGs analyzed, only 80 were found to occur in all organisms. Fifty of these universally present genes showed the same phylogenetic relationships as rRNA (Fig. 1A presents examples). For brevity, we refer to universally conserved genes that share the rRNA topology as 'three-domain' genes. The majority of universally conserved three-domain COG genes (37 of 50) are physically associated with the ribosome in modern cells. For the 30 COGs that were not three-domain, there was no single evolutionary pattern (e.g., Fig. 1B). In some cases the relationships were simply unresolved, whereas for others one domain clearly separated and the other two domains remained intermixed (Table 1). The 80 COGs were classified into six groups based on phylogeny and known, or presumed, function (Table 1). The six groups are described below.

Group 1: Ribosomal Proteins and Translation Initiation Factors

Group 1 contains genes that recapitulate the three-domain phylogeny and whose products are directly linked to the function of the ribosome. This group includes genes for 29 universally conserved ribosomal proteins (rproteins) and the four universally conserved initiation and elongation factors (RNAs were not considered in this compilation.) In the case of the 30 small subunit rprotein COGs, 15 were universal, six were found only in Bacteria, and nine were found only in Eucarya and Archaea. The majority of universal COGs for small subunit rproteins showed strong support for a three-domain phylogeny (14 of 15; Table 1).

The genes for large subunit ribosomal proteins were a more complex group, and a smaller fraction of these were universally conserved (17 of 51 COGs, with 15 being three-domain). COGs encoding 11 large subunit proteins appeared to be three-domain by either maximum parsimony or neighbor-joining analysis, but bootstrap support for the three-domain topology was not strong (< 50%). We presume that the lack of statistical support in the calculations resulted from random evolutionary convergence in the relatively small data sets (ranging between 125 and 200 parsimony-informative characters) for these COGs. Because the best and most resolved topology was three-domain, we classified them as such.

Group 2: Proteins Associated With the Ribosome or Protein Modification

Group 2 includes universal three-domain genes that encode a diverse set of nonribosomal proteins with known functions that potentially link the genes to ribosome function or to modification of proteins. COG0024, methionine aminopeptidase (*map*, in *E. coli*), cleaves the initiator methionine during the process of translation (Lowther and Matthews 2000). COG0006, XaaPro amino peptidase (*pepP* in *E. coli*), also encodes a protease, initially identified by enzymatic activity against dipeptides with proline as the penultimate residue (Yaron and Mlynar 1968). COG0112 encodes the GlyA protein in *E. coli*. GlyA is required for amino acid catabolism and for donation of methyl groups to S-adenosyl-methionine-dependent methyltransferases and other methylating enzymes. In modern organisms, proteins encoded by three members of Group 2 (COG0201 [*secY*], COG0552 [*ffh*, SRP54], and COG0541 [*ftsY* SRP54 receptor]) are involved in protein export or insertion into membranes, guiding leader peptides to the membrane during translation (Walter and Johnson 1994).

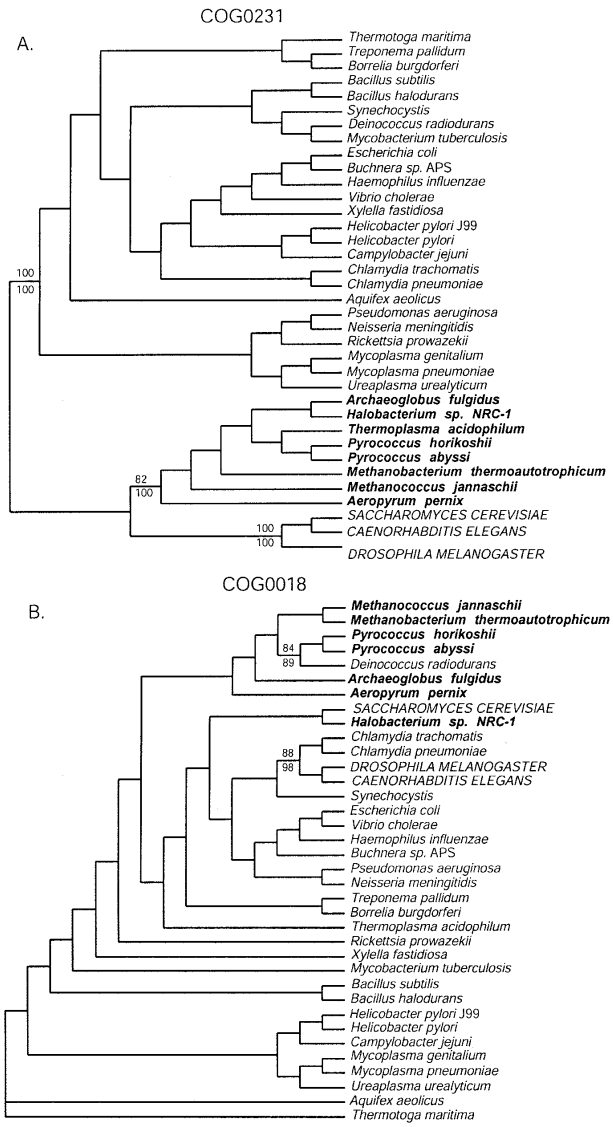


Figure 1 Examples of three-domain and non-three-domain phylogenetic trees from analyses of the COG database protein alignments. The trees are the single shortest trees found by a maximum parsimony (MP) analysis of the amino acid alignments (neighbor-joining [NJ] analysis gave the same topology). Names of organisms belonging to the Bacteria are in italics; names of Archaea are in bold italics, and names of Eucarya are in all capital letters. (*A*) The phylogeny of COG0231 (*efp* in *E. coli*) recapitulates the basic three-domain topology given by ribosomal RNA, and the numbers indicate bootstrap support for the monophyly of the archaeal, bacterial, and eucaryal sequences in this COG. Results from MP bootstrap analysis are given above the branches, and results from NJ bootstrap analysis are given below. (*B*) The phylogeny of COG0018 (*argS* in *E. coli*) violates the three-domain paradigm, as none of the three domains are monophyletic. Indications of horizontal gene transfer events are presented in enlarged font along with corresponding bootstrap support for the nodes demarking lateral transfer.

Table 1. List of COGs and *E. coli* Gene Designations by Groups

Group[a]	COG number[b]	*E. coli* gene designation[b]
1	0048, 0049, 0051, 0052, 0096, 0098, 0099, 0100[c], 0103, 0184[c], 0185[c], 0186, 0199[c], 0522, 0080[c], 0081, 0087, 0088, 0089[c], 0090[c], 0091, 0093, 0094, 0097, 0102, 0197, 0198[c], 0244, 0256[c], [0050], [0231], 0361, [0480], 0532	rpsL, rpsG, rpsJ, rpsB, rpsH, rpsE, rpsM, rpsK, rpsI, rpsO, rpsS, rpsQ, rpsN, rpsD, rplK, rplA, rplC, rplD, rplW, rplB, rplV, rplN, rplE, rplF, rplM, rplP, rplX, rplJ, rplR, [tufB, cysN, tufA, selB], [yeiP, efp], infA, [fusA, prfC], infB
2	0006[d], 0024, 0112[d], 0201, 0541, 0552	[pepP, pepQ, ec1788728], map, glyA, secY, ffh, ftsY
3	0085, 0086, 0180, 0202, 0250, 0258, 0468[d], 0592	rpoB, rpoC, trpS, rpoA, [nusG, rfaH], [exo, polA_1], recA, dnaN
4	0012, [0037]	ychF, [mesJ, ydaO]
5	[0008[f]], 0013[g], 0016[f], 0018[f], 0030[g], 0060[f], 0072[f], 0092[e,g], 0101[e,g], 0124[g], 0125[g], 0143[f], 0162[f], 0172[f], 0200[e,f] 0255[e,g], 0441[f], 0442[f], 0459[d,f], 0470[f], [0492[e,f]], 0495[f], 0525[g], 0533[g], [0550[g]], [0575[d,g]], 0636[e,g], [1109[f]]	[glnS, yadB, gltX], alaS, pheS, argS, ksgA, ileS, pheT_2, rpsC, truA, hisS, tmk, metG_1, tyrS, serS, rplO, rpmC, thrS, proS, groL, holB, [trxB, ahpF], leuS, valS, ygjD, [topA_1, topB], [cdsA, ec1787677], atpE, [cpsG, mrsA]
6	[0073], [0526]	[pheT_1, ygiH, metG_2], [trxA, yfiG, dsbD, yejO, ybbN, dsbA]

[a]Group 1: rproteins and translation factors; 2: ribosome associated proteins; 3: transcription and replication proteins; 4: proteins of unknown function; 5: proteins that do not exhibit 3 domain phylogeny; and 6: protein families.
[b]Square brackets are used to show COGs that contain more than one *E. coli* ORF.
[c]COGs for ribosomal proteins that show the Archaea to be polyphyletic, but both the Bacteria and Eucarya are strongly supported monophyletic groups.
[d]These COGS are missing an ORF for a single bacterium that contains a highly reduced genome and therefore are included in this analysis.
[e]Additional non-three-domain COGs that are missing from a single genome analyzed.
[f]Non-three-domain COGs with statistically supported lateral gene transfers.
[g]Non-three-domain COGs with no statistical support for lateral gene transfers.

Group 3: Proteins Associated With Transcription and Replication of DNA

Four of the universally conserved three-domain COGs in Group 3 encode proteins involved in transcription, including three subunits of DNA-dependent RNA polymerase (COG0085, COG0086, and COG0202 [*RpoB, RpoC,* and *RpoA,* respectively in *E. coli*]), and the gene for a transcription antiterminator (COG0250, *NusG* in *E. coli*).

The number of universal genes involved in DNA replication and repair was surprisingly small, only four. Of these universal genes, only three were found to be three-domain: COG0592 (*DnaN*, in *E. coli*) that encodes the sliding clamp subunit of DNA polymerase III, COG0258 (*Pol1-A* in *E. coli*) that encodes the 5′-3′ exonuclease function of the DNA polymerase I, and COG0468 that encodes the recombination enzyme RecA.

Group 4: Uncharacterized Proteins

Two universally conserved genes that displayed three-domain phylogeny (> 95% bootstrap for all domains) but have no known functions were also found. In both cases, at least one property of the COG proteins could be predicted from its sequences. COG0037 (*mesJ* and *ydaO* in *E. coli*) encodes a predicted ATPase, and COG0012 (*ychF* in *E. coli*) encodes a predicted GTPase.

Group 5: Universal, Non-Three-Domain Proteins

Twenty-eight universally conserved COGs did not show three-domain phylogeny. Presumably, therefore, these genes encode essential functions and have been subjected to lateral gene transfer at some point in evolution (Doolittle 1999; Glansdorff 2000). For example, 14 of the eucaryal amino acyl tRNA synthetase genes did not form a monophyletic group, and rather were always nested within either the bacterial or the archaeal groups (Woese et al. 2000). The non-three-domain universal COGs also include COG0125 (thymidine kinase), COG0550 (topoisomerase 1A), and COG1109 (phosphomanomutase; *mrsA* in *E. coli*). COG0533 has been predicted to encode a metal-dependent protease (*ygjD* in *E. coli*), but the precise function of the gene product has not been identified in any organism. The remaining COG representing a subunit of DNA polymerase III that was found to be universal was COG0470 (*holB* in *E. coli*). In modern organisms this subunit is required to load the modern sliding clamp, but, like the rest of the essential DNA polymerase genes, COG0470 has been transferred between domains. Groups 5 also includes a number of COGs that are missing from only one of the 36 genomes included in the survey and do not show three-domain phylogeny (Table 1).

Group 6: Protein Families and Domain Families

The COG database contains two gene families that occur universally, COG0073 (EMAP domain) and COG0526 (thiodisulfide isomerases). These were not analyzed in this study due to the large number of paralogs in these COGs.

DISCUSSION

Systematic phylogenetic analyses of the universally conserved COG proteins revealed a genetic core of organisms containing a small number of genes that coevolved with the ribosomal RNAs since their divergence from a common ancestor. As expected, most of the three-domain genes belong to the nucleic acid-based central information pathway (ribosomal proteins, DNA/RNA polymerase subunits, elongation factors). However, we also discovered a number of three-domain COG proteins with little apparent connection to genetic transmission or gene expression (e.g., membrane insertion factors and proteases). Perhaps the most surprising finding of this analysis was the relatively small number of the COG gene sets that were three-domain in this analysis. Of the nearly 3100 COG gene sets in the database, only 80 were universal and, of these, only 50 were three-domain.

Comparison of the gene sets used in the analysis suggested four main reasons for the paucity of three-domain COG proteins. First, many of the proteins in the COG database are unique to subsets of organisms, a reflection of the enormous phenotypic diversity of modern cell types. For ex-

ample, genes required for synthesis of cell membranes, a required function for all modern organisms, are not universally conserved among the phylogenetic domains. This is because the biochemistry of archaeal ether-linked lipids is fundamentally different from that used in the other two domains, which produce ester-linked lipids (Koga et al. 1993). Second, the amino acid sequences of some proteins have diverged so radically since the LCA that the sequences are no longer recognizably homologous in different organisms (e.g., F_1F_0 ATP synthetase; Gruber et al. 2001). Third, gene loss without replacement is a common phenomenon in many genomes and appears to play an important role in shaping genome content (Snel et al. 2002). Finally, the low number of three-domain COG proteins reflects the importance of gene replacement by genes of independent origins through nonorthologous displacement by lateral gene transfer. As examples of the latter, DNA primase, DNA polymerization activity, and ribonuclease H activity all appear to have multiple independent origins (Leipe et al. 1999). This may also be true for other ribonucleases and reverse transcriptase. As pointed out earlier, it is certain that the LCA contained many genes other than the 80 three-domain COGs and that some of the COGs were added after the three domains diverged. However, by and large we found that late additions to the COGs and lateral transfers were obvious from their phylogenetic patterns.

Three-Domain Ribosome-Associated Proteins

Most of the 50 three-domain COGs identified were ribosomal proteins (29 of 50). This finding supports previous conclusions that the divergence of the three types of ribosomes (bacterial, archaeal, and eukaryal) occurred after a relatively efficient ribosome structure was in place (Ouzounis and Kyrpides 1996; Olsen and Woese 1997). The abundance of three-domain ribosomal proteins may be attributable to the specific physical association of these proteins with the rRNA. The crystal structure of the *Thermus thermophilis* 30S subunit suggests that many of the three-domain ribosomal proteins in the small subunit (SSU) are found at junctions between helices, such as S4, S7, and the cluster of proteins S8, S15, and S17. Other three-domain SSU proteins penetrate the RNA structural core, providing functional stability (Wimberly et al. 2000). The interactions of the SSU proteins with the 16S ribosomal RNA, as well as with each other, suggest a strong mutual dependency and perhaps a powerful selective constraint inhibiting radical sequence evolution or nonorthologous displacement.

In contrast to the three-domain SSU proteins, considerably fewer of the large ribosome subunit proteins were three-domain. In general, the large subunit (LSU) proteins tend to be less physically clustered in the ribosome than are those of the small subunit. The crystal structure of the *Haloarcula marismortui* 50S ribosomal subunit shows that only a few proteins, such as L3, L13, and L14, are sufficiently close to one another to interact physically. The primary interaction of the LSU proteins is with RNA rather than other proteins (Ban et al. 2000). This provides some rationale for the lower frequency of the large ribosome subunit proteins in the three-domain set. We note that the collection of three-domain rproteins emphasizes the deep divergence of the three domains of life, arguing against evolution models in which the Eucarya are derived from a fusion of other cell types.

In addition to many ribosomal proteins, a number of other proteins associated with the ribosome also are three-domain. As might be expected considering the relatively sophisticated protein synthesis machine of the LCA, the basic initiation and elongation factors are three-domain (Kyrpides and Woese 1998). More surprisingly, several proteins used for proteolytic modification of nascent peptides and for methylation events are three-domain. Methionine aminopeptidase (COG0024, *map*) is responsible for the proteolytic processing of nascent peptides during translation to remove the initiator methionine. In three genomes, the *pepP* (COG0006) gene (proteolytic modification) has been found directly adjacent to the gene for one of the universal three-domain elongation factors, implying a link to maturation of proteins during translation (Matos et al. 1998). Methylated nucleotides are a universal property of ribosomal RNA, and the presence of a methyl donor (COG0112, glyA) among the three-domain COGs suggests that methylation was required for the early function of the ribosome.

Finally, components of two systems for insertion of proteins into membranes were found among the three-domain COG proteins (COG0201, SecY; COG0541 and COG0552, Ffh and FtsY, respectively). The three-domain nature of these membrane insertion factors suggests that functions linked to membranes were an ancient, required activity prior to the establishment of the three domains of life.

Three-Domain Proteins Not Directly Associated With the Ribosome

In contrast to the coordinated structure of the ribosome, relatively few genes encoding proteins involved in DNA replication or transcription from DNA to RNA proved to be three-domain. The majority of RNA polymerases found in modern organisms are not three-domain, which illustrates the diversity of proteins that can carry out this catalytic activity. Indeed, a number of studies have pointed to multiple origins for RNA polymerases (McAllister and Raskin 1993; Zhang et al. 1999; Cramer et al. 2000). We identified in this study only three subunits of the core DNA-dependent RNA polymerase as three-domain, as seen previously (Iwabe et al. 1991; Klenk et al. 1993). The three-domain nature of the core RNA polymerase subunits indicates that the LCA used DNA for genetic continuity. This supposition is supported by the occurrence in the three-domain set of two enzymes of DNA metabolism, RecA and Pol1A (Eisen and Hanawalt 1999). The only component of the replicative DNA polymerase in modern cells that was found to be three-domain is *DnaN* (COG0592), the gene for the "sliding clamp." Considerable evidence supports the idea that this protein is necessary for the high degree of processivity of DNA polymerase during replication (Kuriyan and O'Donnell 1993; Hingorani and O'Donnell 2000). Others have noted the sequence divergence of the subunits of the replicative DNA polymerase, where it has been suggested that the capacity for DNA polymerization arose several times (Leipe et al. 1999).

This collection of three-domain DNA metabolism and transcription enzymes suggests that the ability to synthesize and transcribe long DNA molecules was an important property of the LCA. This innovation would have increased genetic linkage, which in turn would have increased the ability to transmit genetic information through vertical inheritance. In particular, the sliding clamp function would have been required to allow accurate replication of linked genes. Additionally, both RecA and Pol1A would contribute to genetic continuity by gene conversion after recombination, and

would become increasingly useful in the maintenance of genetic information as the lengths of DNA strands increased (Eisen and Hanawalt 1999). A final protein that may have contributed to this general innovation is the three-domain COG protein NusG. As a transcription anti-terminator, NusG improves transcriptional efficiency. Moreover, NusG (along with other proteins, such as the ribosomal protein S4) has been proposed as a link between the ribosome and the process of transcription (Squires and Zaporojets 2000).

Two universal COGs consist of genes with functions predicted solely on the basis of sequence similarity to other functionally described protein motifs. These genes likely encode proteins involved in the central information processing of the cell. One of these COGs is a predicted GTPase (COG0012, *YchF* in *E. coli*) that could be disrupted in a *Mycoplasma genitalium* mutagenesis study, and so is not an essential gene for laboratory growth (Hutchison et al. 1999). The other COG is a predicted ATPase (COG0037, *YdaO* and *MesJ* in *E. coli*). The universal three-domain conservation of these genes suggests that they encode ancient and fundamental properties of all organisms, and identifies them as potentially fruitful targets for further experimentation.

The universal COGs that are not three-domain primarily contain genes that encode proteins that are not integrated into specific large macromolecular complexes, for instance, the aminoacyl tRNA synthetases (14 of 28 COGs, reviewed by Woese et al. 2000). One of the non-three-domain COGs represents a metal-dependent protease that is universally conserved, but of unknown specific function. The universality of this protein indicates that it is an important cellular component that is not highly integrated into a specific macromolecular complex. The function of this protein could be a useful subject for further investigation. Although lateral gene transfer is evident in this group of universally conserved non-three-domain genes, the numbers of transfers are still relatively low, indicating that lateral gene transfer was not extensive among these genes (Table 1; Snel et al. 2002).

METHODS

Phylogenetic Analyses

More than 3100 COGs from 34 sequenced bacterial, archaeal, and eucaryal genomes available in the COG database were surveyed. Although additional genome sequences continue to be determined, the generality of these results is unlikely to be affected substantially by additional genomic sequences. Eighty of the COGs surveyed were found to occur in all organisms (Table 1). The phylogenetic relationships of these COGs were then examined using PAUP version 4.0b8 (Swofford 1998). Alignments were obtained from the COG database, and orthologs from *Drosophila melanogaster* and *Caenorhabditis elegans* genomes (identified in the COG database) were added to the alignment using the CLUSTAL W program (Aiyar 2000).

Phylogenetic analyses were performed on all of the final alignments of the amino acid sequences. A maximum parsimony (MP) heuristic search with 10 random addition sequence searches was performed to find the most parsimonious tree or sets of trees (summarized by strict consensus). A distance analysis of the sequence was also performed using the neighbor-joining (NJ) method. To determine the confidence levels for each tree, an MP bootstrap analysis with 100 replicates (10 random addition sequence searches per replicate) and an NJ bootstrap with 500 replicates were conducted. Although the sequence alignments used in the phylogenetic analyses contained clear regions of homology between all of the sequences, they sometimes also contained poorly aligned sections due to insertions or deletions that had accumulated over evolutionary time. To test the effect of these poorly aligned regions on the phylogenetic analyses, we repeated the NJ and MP phylogenetic analyses on a selection of 10 different COG data sets after excluding the poorly aligned regions in these alignments (three-domain: COG0048, CO0080, COG0180, COG0198, COG0201; non-three-domain: COG0013, COG0092, COG0143, COG0495, COG0550). These particular alignments were chosen because they included a broad array of alignment sizes, varying from 248 positions in the smallest multiple sequence alignment to 1488 positions in the largest. Excluding poorly aligned regions of these alignments did not significantly alter the resulting phylogenetic topologies and had no effect on the interpretation of whether any of these particular COG protein groups were three-domain. Based on these results, we concluded that the CLUSTAL W alignments were appropriate for answering the question of whether particular COGs were three-domain and that the poorly aligned sections had a negligible effect on the phylogenetic analyses.

Because MP and uncorrected NJ analyses underestimate the rates of change in amino acid sequences, we tested whether rate-corrected distance and maximum likelihood (ML) analyses affected the interpretation of phylogenetic relationships within COG protein groups. NJ analyses using PAM amino acid distance corrections, available with the Phylip phylogeny package (Felsenstein 1993), were preformed with the various COG alignments. In addition, we used the ML approach for protein sequence data sets, available with the Molphy phylogenetic analysis package (http://www.ism.ac.jp/software/ismlib/softother.e.html#molphy), to determine whether there was support for alternative topologies. The protein ML analyses utilized the JTT (Jones, Taylor, and Thornton) model of protein sequence evolution (Jones et al. 1992). Because ML analyses tend to be computationally intensive, we used the NJ trees with the PAM distance corrections as starting trees and assessed the likelihood of local topology rearrangements. None of the rate-corrected analyses found tree topologies significantly different from the uncorrected analyses.

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

Aiyar, A. 2000. The use of CLUSTAL W and CLUSTAL X for multiple sequence alignment. *Methods Mol. Biol.* **132:** 221–241.

Asai, T., Zaporojets, D., Squires, C., and Squires, C.L. 1999. An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: Complete exchange of rRNA genes between bacteria. *Proc. Natl. Acad. Sci.* **96:** 1971–1976.

Ban, N., Nissen, P., Hansen, J., Moore, P.B., and Steitz, T.A. 2000. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* **289:** 905–920.

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

Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E., and Stanhope, M.J. 2001b. Universal trees based on large combined protein sequence data sets. *Nat. Genet.* **28:** 281–285.

Cramer, P., Bushnell, D.A., Fu, J., Gnatt, A.L., Maier-Davis, B., Thompson, N.E., Burgess, R.R., Edwards, A.M., David, P.R., and Kornberg, R.D. 2000. Architecture of RNA polymerase II and implications for the transcription mechanism. *Science* **288:** 640–649.

Doolittle, W.F. 1999. Lateral genomics. *Trends Cell Biol.* **9:** 5–8.

Eisen, J. and Hanawalt, P.C. 1999. A phylogenomic study of DNA repair genes, proteins, and processes. *Mutat. Res.* **435:** 171–213.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package). Distributed by the author, Department of Genetics, University of

Washington, Seattle. http://evolution.genetics.washington.edu/phylip.html
Glansdorff, N. 2000. About the last common ancestor, the universal life-tree and lateral transfer: A reappraisal. *Mol. Microbiol.* **38:** 177–185.
Gruber, G., Wieczorek, H., Harvey, W.R., and Muller, V. 2001. Structure-function relationships of A-, F- and V-ATPases. *J. Exp. Biol.* **204:** 2597–2605.
Hingorani, M. and O'Donnell, M. 2000. A tale of toroids in DNA metabolism. *Nat. Rev. Mol. Cell Biol.* **1:** 22–30.
Hutchison, C.A., Peterson, S.N., Gill, S.R., Cline, R.T., White, O., Fraser, C.M., Smith, H.O., and Venter, J.C. 1999. Global transposon mutagenesis and a minimal Mycoplasma genome. *Science* **286:** 2165–2169.
Iwabe, N., Kuma, K.-I., Kishino, H., Hasegawa, M., Osawa, S., and Miyata, T. 1991. Evolution of RNA polymerases and branching patterns of the three major groups of archaebacteria. *J. Mol. Evol.* **32:** 70–78.
Jones D.T., Taylor, W., and Thornton, J.M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8:** 275–282.
Klenk, H.-P., Palm, P., and Zillig, W. 1993. DNA-dependent RNA polymerases as phylogenetic marker molecules. *Syst. Appl. Microbiol.* **16:** 138–147.
Koga, Y., Nishihara, M., Morii, H., and Akagawa-Matsushita, M. 1993. Ether polar lipids of methanogenic bacteria: Structures, comparative aspects, and biosynthesis. *Microbiol. Rev.* **57:** 164–182.
Kuriyan, J. and O'Donnell, M. 1993. Sliding clamps of DNA polymerases. *J. Mol. Biol.* **234:** 915–925.
Kyrpides, N. and Woese, C.R. 1998. Universally conserved translation initiation factors. *Proc. Natl. Acad. Sci.* **95:** 224–228.
Leipe, D., Aravind, L., and Koonin, E.V. 1999. Did DNA replication evolve twice independently? *Nucleic Acids Res.* **27:** 3389–3401.
Liao, D. and Dennis, P.P. 1994. Molecular phylogenies based on ribosomal protein L11, L1, L10, and L12 sequences. *J. Mol. Evol.* **38:** 405–419.
Lowther, W.T. and Matthews, B.W. 2000. Structure and function of the methionine aminopeptidases. *Biochim. Biophys. Acta* **1477:** 157–167.
Matos, J., Nardi, M., Kumura, H., and Monnet, V. 1998. Genetic characterization of pepP, which encodes an aminopeptidase P whose deficiency does not affect Lactococcus lactis growth in milk, unlike deficiency of the X-prolyl dipeptidyl aminopeptidase. *Appl. Environ. Microbiol.* **64:** 4591–4595.
McAllister, W. and Raskin, C.A. 1993. The phage RNA polymerases are related to DNA polymerases and reverse transcriptases. *Mol. Microbiol.* **10:** 1–6.
Olsen, G. and Woese, C.R. 1997. Archaeal genomics: An overview. *Cell* **89:** 991–994.
Ouzounis, C. and Kyrpides, N. 1996. The emergence of major cellular processes in evolution. *FEBS Lett.* **390:** 119–123.
Snel, B., Bork, P., and Huynen, M.A. 2002. Genomes in flux: The evolution of archaeal and proteobacterial gene content. *Genome Res.* **12:** 17–25.
Squires, C. and Zaporojets, D. 2000. Proteins shared by the transcription and translation machines. *Annu. Rev. Microbiol.* **54:** 775–798.
Swofford, D. 1998. PAUP: *Phylogenetic analysis using parsimony (and other methods)*. Sinauer Associates, Sunderland, MA.
Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. 2001. The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29:** 22–28.
Walter, P. and Johnson, A.E. 1994. Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane. *Annu. Rev. Cell Biol.* **10:** 87–119.
Wettach, J., Gohl, H., Tschochner, H., and Thomm, M. 1995. Functional interaction of yeast and human TATA-binding proteins with an archaeal RNA polymerase and promoter. *Proc. Natl. Acad. Sci.* **92:** 472–476.
Wimberly, B., Brodersen, D.E., Clemons Jr., W.M., Morgan-Warren, R.J., Carter, A.P., Vonrhein, C., Hartsch, T., and Ramakrishnan, V. 2000. Structure of the 30S ribosomal subunit. *Nature* **407:** 327–339.
Woese, C.R. and Fox, G.E. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci.* **74:** 5088–5090.
Woese, C.R., Kandler, O., and Wheelis, M.L. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci.* **87:** 4576–4579.
Woese, C.R., Olsen, G.J., Ibba, M., and Soll, D. 2000. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.* **64:** 202–236.
Yaron, A. and Mlynar, D. 1968. Aminopeptidase-P. *Biochem. Biophys. Res. Commun.* **32:** 658–663.
Zhang, G., Campbell, E.A., Minakhin, L., Richter, C., Severinov, K., and Darst, S.A. 1999. Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 Å resolution. *Cell* **98:** 687–690.

WEB SITE REFERENCES

http://www.ism.ac.jp/software/ismlib/softother.e.html#molphy; MOLPHY computer package that allows the user to run either the ProtML or NucML programs on their sequence data.

Received July 24, 2002; accepted in revised form December 11, 2002.