The Deep Archaeal Roots of Eukaryotes

Natalya Yutin, Kira S. Makarova, Sergey L. Mekhedov, Yuri I. Wolf, and Eugene V. Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD

The set of conserved eukaryotic protein-coding genes includes distinct subsets one of which appears to be most closely related to and, by inference, derived from archaea, whereas another one appears to be of bacterial, possibly, endosymbiotic origin. The "archaeal" genes of eukaryotes, primarily, encode components of information-processing systems, whereas the "bacterial" genes are predominantly operational. The precise nature of the archaeo–eukaryotic relationship remains uncertain, and it has been variously argued that eukaryotic informational genes evolved from the homologous genes of Euryarchaeota or Crenarchaeota (the major branches of extant archaea) or that the origin of eukaryotes lies outside the known diversity of archaea. We describe a comprehensive set of 355 eukaryotic genes of apparent archaeal origin identified through ortholog detection and phylogenetic analysis. Phylogenetic hypothesis testing using constrained trees, combined with a systematic search for shared derived characters in the form of homologous inserts in conserved proteins, indicate that, for the majority of these genes, the preferred tree topology is one with the eukaryotic branch placed outside the extant diversity of archaea although small subsets of genes show crenarchaeal and euryarchaeal affinities. Thus, the archaeal genes in eukaryotes appear to descend from a distinct, ancient, and otherwise uncharacterized archaeal lineage that acquired some euryarchaeal and crenarchaeal genes via early horizontal gene transfer.

Introduction

The emergence of eukaryotes is one of the central, and hotly debated, themes in evolutionary biology. These debates have led to multiple, competing hypotheses that present drastically different scenarios for the origin of the complex eukaryotic cell (Martin and Muller 1998; Embley and Martin 2006; Kurland et al. 2006; Martin and Koonin 2006; Dagan and Martin 2007; Poole and Penny 2007b). Two key observations that must be taken into account by any concept of eukaryotic origin are currently not contested seriously.

1. All extant eukaryotes evolved from a common ancestor that already possessed an α-proteobacterial endosymbiont that gave rise to the mitochondria and their degraded relatives, hydrogenosomes and mitosomes (van der Giezen and Tovar 2005; Embley 2006).
2. Eukaryotes possess 2 distinct sets of genes, one of which shows apparent phylogenetic affinity with homologs from archaea and the other one is more closely related to bacterial homologs (not all eukaryotic genes belong to these 2 sets, of course; many are of uncertain origin, and many more appear to be unique to eukaryotes). There is a clear functional divide between the "archaeal" and "bacterial" genes of eukaryotes, with the former encoding, largely, proteins involved in information processing (translation, transcription, replication, and repair) and the latter encoding proteins with "operational" functions (metabolic enzymes, components of membranes, and other cellular structures, etc.) (Esser et al. 2004; Rivera and Lake 2004). In some of the informational and operational systems, the archaeal and bacterial affinities, respectively, of eukaryotic genes are manifest qualitatively: Thus, the key proteins involved in DNA replication in archaea and eukaryotes are not homologous to the functionally analogous proteins of bacteria (Leipe et al. 1999), and conversely, some of the principal enzymes of membrane biogenesis are homologous in eukaryotes and bacteria but not in archaea (Pereto et al. 2004).

Apparently, the most parsimonious scenario of eukaryogenesis combining these 2 key facts is that the first eukaryote was an archaeal–bacterial chimera that emerged as a result of an invasion of an archaeon by an α-proteobacterium, the well-established ancestor of the mitochondria (Martin and Muller 1998; Rivera and Lake 2004; Martin and Koonin 2006). However, this is by no means the only scenario of eukaryotic origins that is currently actively considered (Embley and Martin 2006; Poole and Penny 2007b). The main competitor is, probably, the archezoan hypothesis under which the host of the α-proteobacterial endosymbiont was not an archaeon but a primitive, obviously, amitochondrial, proto-eukaryote that already possessed the hallmarks of the eukaryotic cell, such as the endomembrane system, the nucleus, and the cytoskeleton (Kurland et al. 2006; Poole and Penny 2007a). The symbiotic scenarios substantially differ from the archezoan hypothesis with respect to the level of complexity that is attributed to the host of the mitochondrial endosymbiont. Under the symbiotic hypotheses, the host was a "garden variety" archaeon, with the dramatic complexification of the cellular organization being triggered by the symbiosis. In contrast, the archezoan hypothesis posits that, at least, some substantial aspects of the characteristic eukaryotic complexity (e.g., the endomembrane system) evolved prior to and independent of the symbiosis and were already in place in the organism that hosted the mitochondrion. Under the archezoan scenario, the presence of archaea-like genes in the ancestral eukaryotic gene set is, then, explained either by postulating that the proto-eukaryotic lineage was a sister group of archaea and/or by horizontal transfer of archaeal genes. The archezoan hypothesis was seriously undermined by the realization that all unicellular eukaryotes previously thought to be primitively amitochondrial actually possess degraded organelles of α-proteobacterial descent. Nevertheless, the archezoan scenario stays alive, with the proviso that the ancestral archezoan lineage had gone extinct (Poole and Penny 2007a). In addition, more complex scenarios have been considered, with an ancient, primary symbiosis

Key words: archaea, eukaryotes, Euryarchaeota, Crenarchaeota, phylogenetic analysis.

E-mail: koonin@ncbi.nlm.nih.gov.

Mol. Biol. Evol. 25(8):1619–1630. 2008
doi:10.1093/molbev/msn108
Advance Access publication May 6, 2008

Published by Oxford University Press 2008.

leading to the emergence of a nucleated, amitochondriate, proto-eukaryotic cell and antedating the acquisition of an α-proteobacterium that gave rise to the mitochondria. A γ-proteobacterium (Horiike et al. 2004), a δ-proteobacterium (Moreira and Lopez-Garcia 1998; Lopez-Garcia and Moreira 2006), a Clostridium-like gram-positive bacterium (Karlin et al. 1999), or a spirochaete (Margulis 1996) have been variously proposed as bacterial counterparts of this putative primary symbiosis. The possibility also has been considered that the nucleus itself is a derived endosymbiont, a descendant of a Crenarchaeon (Lake and Rivera 1994) or a Euryarchaeon, such as Pyrococcus (Horiike et al. 2004) that invaded a bacterial host.

The rapidly growing collection of sequenced genomes from different domains and lineages of life provides for empirical testing of these hypotheses by phylogenetic analyses on genome-wide data. The problem of eukaryogenesis is extremely hard and complex, given the depth of the divergences involved, and arguably, has to be tackled piecemeal, by deciphering the origins of particular subsets of eukaryotic genes and signature eukaryotic functional system through thorough phylogenetic analysis. Here we address the specific evolutionary origins of those eukaryotic genes that appear to show an affinity with archaeal homologs. In particular, we asked whether the archaea-related "parent" of eukaryotes comes from within the phylogenetic span of the extant archaea, that is, originates from either Euryarchaeota or Crenarchaeota, or outside that span, perhaps, representing a distinct archaeal branch, or even a distinct domain of life. Clearly, in the first case, eukaryotes are expected to be rooted within either Crenarchaeota or Euryarchaeota in phylogenetic trees whereas, in the second case, Eukarya should branch outside of the archaeal clade.

Phylogenetic analyses and other types of evolutionary reconstructions aimed at elucidating the evolutionary relationship between archaea and eukaryotes have yielded conflicting results. Some early comparisons of ribosomal structure and phylogenetic analyses have suggested a specific affinity between eukaryotic genes and their orthologs from Crenarchaeota (dubbed eocytes on the basis of this observation) (Lake et al. 1984; Lake 1988; 1998; Rivera and Lake 1992). Support for the eocyte hypothesis has been subsequently claimed from comparative analysis of ribosomal protein sequences (Vishwanath et al. 2004) and from a novel approach to whole-genome–based phylogenetic analysis (Rivera and Lake 2004).

By contrast, the origin of the archaeal genes of eukaryotes from euryarchaea, and specifically, from methanogens has been postulated on biological considerations, within the framework of the hydrogen hypothesis (Martin and Muller 1998) and one of the so-called syntrophic hypotheses (Lopez-Garcia and Moreira 2006) that are based on different forms of metabolic cooperation between the archaeal and bacterial partners of the primary endosymbiosis. The origin of these genes from methanogens also appears compatible with the results of some phylogenetic analyses (Moreira and Lopez-Garcia 1998; Horiike et al. 2004). Other researchers have argued that the archaeal parent of eukaryotes lies outside the currently known diversity of archaea, on the basis of biological considerations (Vellai et al. 1998) or phylogenetic analysis results (Hedges et al. 2001; Tekaia and Yeramian 2005; Ciccarelli et al. 2006; Fukami-Kobayashi et al. 2007).

Most of these analyses employed a relatively small number of concatenated, most highly conserved protein sequences (e.g., those of ribosomal proteins) for phylogenetic tree construction (Ciccarelli et al. 2006; Hartman et al. 2006) or built trees on the basis of gene composition comparisons (phyletic patterns) (Horiike et al. 2004; Tekaia and Yeramian 2005), or else employed other features, such as domain architectures of multidomain proteins, as phylogenetic characters (Fukami-Kobayashi et al. 2007). The most complete of the relevant phylogenetic studies that have appeared so far is the work of Pisani et al. (2007) who analyzed nearly 6,000 gene sets from 185 genomes using the supertree approach. These authors reported an apparent phylogenetic affinity between eukaryotes and *Thermoplasma* although the signal came with limited statistical support. *Thermoplasma* or a related archaeon also has been suggested as the likely archaeal parent of eukaryotes on the basis of biochemical and cytological considerations (Searcy et al. 1978; Margulis and Stolz 1984; Hixon and Searcy 1993; Margulis 1996; Margulis et al. 2000).

Given these conflicting conclusions on the nature of the archaeal–eukaryotic affinity that have been reached over the years using widely different methods along with a variety of biological considerations, we were compelled to attempt an exhaustive phylogenetic analysis of eukaryotic genes of apparent archaeal origin, with a minimal set of assumptions. We do not take it for granted that genes in a lineage share a common history (Gogarten et al. 2002; Bapteste et al. 2005; Doolittle and Bapteste 2007) and avoid concatenation of sequences of individual genes or a supertree-type analysis of individual trees. Instead, trees for orthologous gene sets were built separately, their topologies were assessed with several independent methods, and a post hoc census was taken.

We conclude that neither Crenarchaeota nor Euryarchaeota made the decisive contribution to the archaeal component of the ancestral eukaryotic gene set. The bulk of the eukaryotic genes with an apparent archaeal affinity seem to originate from a distinct archaeal lineage that branched off the trunk of the archaeal tree prior to the radiation of Crenarchaeota and Euryarchaeota. A limited amount of horizontal gene transfer (HGT) might have led to the acquisition of the few eukaryotic genes that do show Crenarchaeal and Euryarchaeal affinities.

Materials and Methods

Clusters of Orthologous Genes

The database of archaeal clusters of orthologous genes (arCOGs) (Makarova et al. 2007) includes 41 archaeal genomes (13 Crenarchaeota, 27 Euryarchaeota, and 1 Nanoarchaeon; see species names in supplementary table S2, Supplementary Material online) and contains 7,672 arCOGs of which 3,164 included proteins from at least 6 species and were used for the present analysis.

Eukaryotic clusters of orthologous domains (KODs) represent a further development of the eukaryotic KOG database (Tatusov et al. 2003) and were constructed from 67 eukaryotic proteomes (supplementary table S2, Supplementary Material online). Orthologs from new genomes were added to the clusters of orthologs from 7 eukaryotes (Tatusov et al. 2003) using a modification of the COGNITOR method (Tatusov et al. 2000). New KODs, not represented in the KOG database, were created by manual analyses of species-specific Blast hits. The current KOD database contains over 40,000 clusters of orthologous domains; this study was restricted to a subset of 5,377 "ancient" KODs, which were operationally (and liberally) defined as those that included at least one plant gene and at least one fungal or animal gene.

Prokaryotic clusters of orthologs, representing 110 archaeal and bacterial genomes, were an extension of the 2003 version of the COG database (Tatusov et al. 2003). COGs used in this study represent the original 2003 COGs with new members added using the modified COGNITOR method (Tatusov et al. 2000).

Relationships between Gene Clusters

Multiple alignments of the protein sequences comprising the archaeal and eukaryotic clusters were constructed using MUSCLE (Edgar 2004) and used to construct position-specific scoring matrices (PSSMs). The PSSMs were compared with custom databases consisting of individual sequences representing the target set of clusters using single-pass PSI-Blast (Altschul et al. 1997) (for details, see supplementary methods, Supplementary Material online). Target clusters were ranked according to the mean score of member sequences. Specifically, the PSSMs for each of the 5,377 eukaryotic KODs were run against the database of the archaeal sequences from the arCOGs, and reciprocally, PSSMs for each of the 7,672 arCOGs were run against the database of the eukaryotic sequences from the KODs. A cluster pair was registered when the 2 clusters were, reciprocally, the top-ranked hits for each other. The archaeo–eukaryotic (A–E) clusters resulting from this step were linked to COGs. Both arCOGs and KOGs have been previously assigned to COGs through a combination of automatic comparison and manual curation (Makarova et al. 2005, 2007). When both members of an A–E pair were assigned to the same COG, an arCOG–KOD–COG triplet was formed. The archaeal and eukaryotic PSSMs from the remaining A–E pairs were compared with the database of COG sequences using single-pass PSI-Blast. Results were examined manually and, when appropriate, a COG was linked to the A–E pair, adding to the list of arCOG–KOD–COG triplets.

Selection of Sequences and Preparation of Alignments for Phylogenetic Analysis

For the preliminary phylogenetic analysis, archaeal, eukaryotic, and bacterial sequences in each cluster of putative orthologs were clustered using Blastclust with a 80% identity threshold to avoid the appearance of nearly identical sequences in the tree, that is, to eliminate redundancy. The sequences were aligned using MUSCLE (Edgar 2004). Poorly conserved positions, positions including greater than one-third gaps, and sequences covering less than 80% of the alignment length were removed from the alignment (see details in the supplementary methods, Supplementary Material online). For the detailed phylogenetic analysis, for each of the groups, a branch with the best compromise between the number of members of the given group (the more the better) and the number of members of the other groups (the less the better) was identified. The representatives of Euryarchaeota, Crenarchaeota, Eukaryota, and Bacteria were selected from this branch with the aim to maximize diversity but avoid unusually long branches (see the details of the algorithm in supplementary methods, Supplementary Material online); up to 10 representatives from each branch were chosen. Representative sequences were aligned using MUSCLE; weakly conserved positions and positions with gaps in more than one-third of the aligned sequences were removed.

Phylogenetic Analysis

In the first round of phylogenetic analysis, maximum likelihood (ML) trees were constructed using the PhyML software with the following parameters: number of relative substitution rate categories was 4; the proportion of invariable sites and alpha (gamma distribution parameter) was adjustable (estimated) (Guindon and Gascuel 2003). Eight PhyML runs were performed for each alignment, for 8 substitution models (Blosum62, Dayhoff, JTT, DCMut, RtREV, CpREV, VT, and WAG). The best tree was chosen by maximum log likelihood of 8 trees (likelihood was used instead of AIC or BIC because all used models have the same number of parameters). For a majority of the clusters, WAG matrix was the best (69% from 672 alignments), the second matrix was Blosum62 (19%), and the others RtREV (6%) and CpREV(6%), see supplementary figure S4 (Supplementary Material online).

For detailed phylogenetic analysis, ML trees were constructed using the TreeFinder program (Jobb et al. 2004), with the estimated site rates heterogeneity and with the substitution model found best for a given alignment in the first-round analysis. Tree topologies were compared using the TreeFinder program according to either their expected likelihood weights (ELWs [Strimmer and Rambaut 2002]) or by the approximately unbiased (AU) test *P* value (Shimodaira 2002). Unconstrained ML trees were compared with 3 constrained topologies, which represented the "deep," "crenarchaeal," and "euryarchaeal" models of the origin of the eukaryotic genes.

Availability of the Data and Results

The list of the 980 identified A–E orthologous clusters, alignments of the respective protein sequences, and the resulting PhyML and ML trees are available at ftp://ftp.ncbi.nih.gov/pub/koonin/yutin.

Results

Eukaryotic Genes of Archaeal and Bacterial Origins

The analysis of the evolutionary relationships between eukaryotic and archaeal genes involved a series of steps

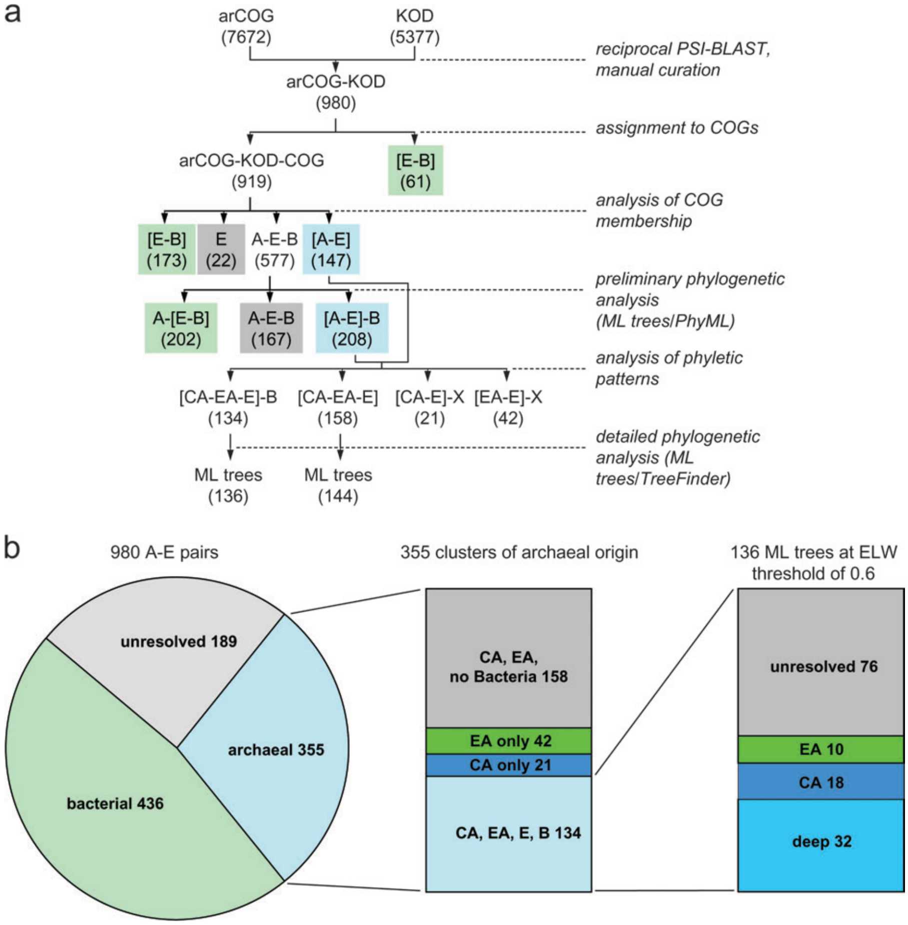


FIG. 1.—Phylogenetic classification of the archaeal–eukaryotic orthologs. (*a*) Flowchart of the procedure. (*b*) Breakdown of orthologous clusters by inferred origin A, archaea; B, bacteria; CA, Crenarchaeota; E, eukaryotes; and EA, Euryarchaeota. For further details, see text.

from identification of orthologs to phylogenetic hypothesis testing, organized in a pipeline (fig. 1*a*; for details, see Materials and Methods). In order to derive a comprehensive set of orthologous genes of eukaryotes and prokaryotes, we performed a comparison between updated versions of previously developed clusters of orthologous genes for eukaryotes (originally, KOGs but currently renamed KOD, euKaryotic clusters of orthologous domains) and prokaryotes, COGs (Tatusov et al. 2003), and a new collection of orthologous clusters for archaea, the arCOGs (Makarova et al. 2007). The identification of orthologs was based on the bidirectional best hit scheme (Tatusov et al. 1997), with the difference that the best hits were identified by running a PSSM (Altschul et al. 1997) derived from the alignment of orthologous protein sequences for each arCOG against the database of KOD sequences and, reciprocally, comparing KOD PSSMs with the arCOG sequence collection. This strategy was chosen to increase the detection sensitivity and also because, for the purpose of this study, it was desirable to collapse the lineage-specific paralogs, that is, those genes that evolved by duplication either in eukaryotes or in archaea. It can be easily demonstrated that, if (PSI-) Blast scores change monotonically with evolutionary distances, the bidirectional best hit scheme is guaranteed to choose one and only one reciprocal pair of genes (COGs) on both sides of a tree branch, reducing the many-to-many relationship between members of lineage-specific expansions of paralogous families to one-to-one relationship between the selected representatives. Deviations of the (PSI-) Blast scores from this relationship with the distances might lead to the appearance of additional pairs from the same multigene family; however, such deviations are seen only on rare occasions (Czabarka E, Wolf YI, unpublished observations). This step yielded 980 archaea–eukaryote (A–E) pairs of putative orthologous gene sets (fig. 1*a*). At the next step, whenever possible, bacterial orthologs from the COG database were added using either the same approach or the previous, manually curated assignments of KODs and arCOGs to COGs (Makarova et al. 2005, 2007). As the result of this procedure, 919 archaeal–eukaryotic–bacterial (A–E–B) triplets were obtained. For the remaining 61 A–E pairs, the eukaryotic proteins were found to be closely associated with bacterial only COGs, whereas the archaeal component had a different bacterial counterpart in the COG database. These 61 KODs were considered to be of bacterial origin; the affinity between the respective arCOGs and KODs, probably, reflects ancient paralogy (an example is given in supplementary fig. S1, Supplementary Material online).

The clusters of putative orthologs were then classified according to the representation of the 3 domains. Among the 919 A–E–B triplets, 147 were found to be predominantly archaeal (no more than 2 bacterial proteins); in the absence of ancient bacterial orthologs, the eukaryotic members of these clusters were tentatively assigned archaeal origin. Conversely, 173 clusters were found to contain an abundance of bacteria but only a few (<6) archaea; accordingly, the eukaryotic members of these triplets were tentatively assigned bacterial origin. These assignments rely on the assumption, akin to the parsimony principle, that the dominance of a particular gene in one domain (e.g., archaea) reflects its origin in that domain, with the few instances seen in the other domain (bacteria) attributable to horizontal transfer. A possible caveat is that the gene in question is of bacterial origin but spread horizontally among diverse archaea, perhaps, owing to its special utility and was concomitantly lost in the majority of bacteria. This scenario seems highly unlikely, especially, given the drastic lifestyle differences between the majority of the archaea (mostly, thermophiles) and bacteria in the current genome collections; nevertheless, there is no means to rule it out other than parsimony considerations. In 22 cases, neither archaeal nor bacterial members of the clusters made the cutoff (6 and 10 species, respectively), so these clusters were classified into an unresolved category. The remaining 577 A–E–B triplets were subject to a round of preliminary phylogenetic analysis. The rationale behind choosing this and only this set of orthologous clusters for phylogenetic analysis was that only clusters with appropriate taxon sampling from each of the 3 domains are suitable for meaningful analysis of the likely origins of the eukaryotic genes (Pollock et al. 2002; Zwickl and Hillis 2002).

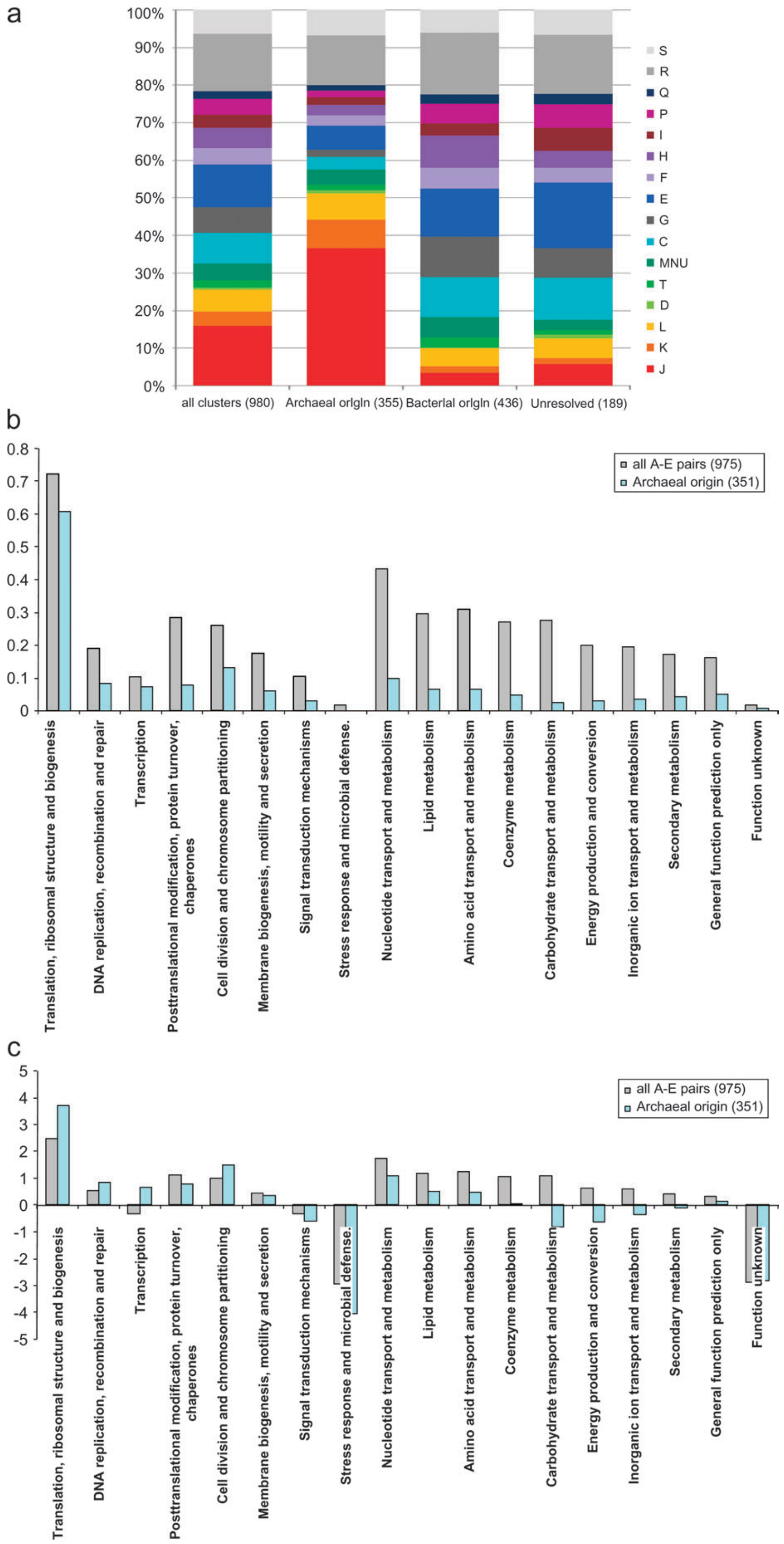
The assignment of archaeal or bacterial origin to eukaryotic genes is a nontrivial task except for the above cases where orthologs were detected in one of the domains, archaea or bacteria, (almost) exclusively. For the majority of the highly conserved eukaryotic genes that had numerous apparent orthologs in both archaea and bacteria, the root position in the phylogenetic tree is not known, so technically, the archaeal or bacterial origin of the eukaryotic genes cannot be determined. Therefore, to infer eukaryotic gene origin, we made an additional assumption. Considering the well-established endosymbiotic scenario of eukaryotic evolution, under which the endosymbiosis of an archaeon (or an ancestral form related to archaea) with an α-proteobacterium that led to the origin of mitochondria occurred well after the divergence of the major bacterial lineages (Gray 1992; Gray et al. 2001), we assumed that eukaryotic genes of bacterial origin would be rooted within the bacterial subtree rather than outside the bacteria. Conversely, eukaryotic genes that branched off between the bacterial and archaeal subtrees could be considered candidates for archaeal origin. A potential problem with this assumption is the well-known long-branch attraction (LBA) artifact because of which eukaryotic genes of bacterial origin that underwent a period of accelerated evolution during eukaryogenesis could fall outside the bacterial subtree. Owing to the LBA problem, the resulting set of candidates for archaeal origin is expected to include some false positives, that is, fast-evolving proteins of bacterial origin. Hence, we included an additional filtering step in which the trees for eukaryotic proteins that, on average, showed a greater similarity to bacterial than to archaeal orthologs were inspected on a case-by-case basis and removed from the archaeal set unless the high similarity to bacterial orthologs could be attributed to HGT from eukaryotes to bacteria.

In the first round of phylogenetic analysis, eukaryotic, archaeal, and bacterial proteins from each of the 577 A–E–B triplets were aligned (for details, see Materials and Methods). For each alignment, 8 phylogenetic trees with different substitution models were constructed using the PhyML software (Guindon and Gascuel 2003). The best tree for each A–E–B triplet was chosen for further analysis according to its log-likelihood value. The similarity filter was applied after which the trees were examined case-by-case and classified into the following 3 categories (fig. 1*a*).

- Unresolved: 167 trees. Trees, in which archaea, eukaryotes, or bacteria were not monophyletic, and starlike trees.
- Probable bacterial origin of eukaryotic genes: 202 trees. Trees in which eukaryotes grouped within the bacterial subtree.
- Probable archaeal origin of eukaryotic genes: 208 trees. Trees where eukaryotes formed a clade either within the archaeal subtree or between archaea and bacteria.

Altogether, after pooling those genes that had no or few bacterial orthologs with those that were marked as archaeal at the stage of preliminary phylogenetic analysis, we identified 355 ancestral eukaryotic genes of apparent archaeal descent (36% of the identified A–E orthologous pairs) (fig. 1*b*). The distribution of the functional assignments among the bacterial and archaeal genes seems to support the classification of conserved eukaryotic genes by their likely origin that was derived by using phyletic patterns and preliminary tree analysis. As expected from previous studies (Rivera et al. 1998; Esser et al. 2004), the proteins of apparent archaeal origin were mostly components of information-processing systems, whereas the proteins of bacterial origin were predominantly operational, that is, involved in various metabolic, cellular, and signaling processes (fig. 2*a*). Thus, we assumed that the classification of a gene as archaeal at this stage, typically, correctly reflected its origin.

The current breakdown of archaeal–eukaryotic orthologs suggests that the archaeal parent contributed at least 355 genes to the emerging eukaryotic cell (the number of present-day eukaryotic genes that are traceable to these archaeal genes is, ca., 2-fold greater as a result of ancient duplications during eukaryogenesis [Makarova et al. 2005]). This hardly could be the complete list because, inevitably, some archaeo–eukaryotic orthologs escape detection owing to extreme divergence, whereas others could be missed due to the complex history of duplications and accompanying changes in the evolution rate obscured the relationships with the prokaryotic homologs such that some of the eukaryotic genes of archaeal origin might end up in the unresolved set (fig. 1*a* and *b*). Nevertheless, the set derived here seems to be a reasonable, if conservative, approximation of the archaeal contribution to eukaryogenesis. Indeed, these genes include the majority of the protein components of the archaeal translation system, cover most of the core functions involved in archaeal translation,


a
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
all clusters (980)
Archaeal origin (355)
Bacterial origin (436)
Unresolved (189)
S
R
Q
P
I
H
F
E
G
C
MNU
T
D
L
K
J
b
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
all A-E pairs (975)
Archaeal origin (351)
Translation, ribosomal structure and biogenesis
DNA replication, recombination and repair
Transcription
Posttranslational modification, protein turnover, chaperones
Cell division and chromosome partitioning
Membrane biogenesis, motility and secretion
Signal transduction mechanisms
Stress response and microbial defense.
Nucleotide transport and metabolism
Lipid metabolism
Amino acid transport and metabolism
Coenzyme metabolism
Carbohydrate transport and metabolism
Energy production and conversion
Inorganic ion transport and metabolism
Secondary metabolism
General function prediction only
Function unknown
c
5
4
3
2
1
0
-1
-2
-3
-4
-5
all A-E pairs (975)
Archaeal origin (351)
Translation, ribosomal structure and biogenesis
DNA replication, recombination and repair
Transcription
Posttranslational modification, protein turnover, chaperones
Cell division and chromosome partitioning
Membrane biogenesis, motility and secretion
Signal transduction mechanisms
Stress response and microbial defense.
Nucleotide transport and metabolism
Lipid metabolism
Amino acid transport and metabolism
Coenzyme metabolism
Carbohydrate transport and metabolism
Energy production and conversion
Inorganic ion transport and metabolism
Secondary metabolism
General function prediction only
Function unknown

transcription, and replication, and potentially could approximate a minimal information-processing apparatus at the outset of eukaryogenesis (fig. 2*b* and *c*).

Of the 208 A–E–B clusters for which archaeal origin of the eukaryotic genes was inferred, 134 contained enough representatives of Crenarchaeota, Euryarchaeota, Eukaryota, and Bacteria to detect possible euryarchaeal, crenarchaeal, or deep archaeal ancestry of the eukaryotic gene family (see Detailed Phylogenetic Analysis below). The remaining 74 clusters were combined with the 147 clusters of archaeal origin that lacked representation of bacterial orthologs and further classified on the basis of the representation of the group members (fig. 1*a* and *b* and see below).

Detailed Phylogenetic Analysis of 134 A–E–B Clusters of Orthologs

For the purpose of detailed phylogenetic analysis, 134 clusters of orthologs that included eukaryotic genes of apparent archaeal origin (A–E–B clusters) were selected as described above. Among these, 2 clusters each encompassed 2 distinct eukaryotic clades (nonmonophyly of eukaryotes was validated using the AU test; see Materials and Methods). These clusters were analyzed twice, once with each of the eukaryotic clades, bringing the total number of analyzed orthologous sets to 136. To select representatives from each of the 4 major clades (Crenarchaeota, CA; Euryarchaeota, EA; Eukaryota, E; and Bacteria, B), we implemented an algorithm that analyzes the midpoint-rooted PhyML tree (for details, see Materials and Methods and supplementary methods [Supplementary Material online]). For each of the groups, the most representative branch was identified using the criteria of the maximum diversity of the members of the given major clade combined with the minimum diversity of the members of the other clades. The required number of representatives was selected from this branch with the aim to maximize diversity while excluding unusually long branches. For each cluster, an unconstrained ML tree and 3 constrained ML trees were constructed. Each of the constrained trees enforced the monophyly of each of the 4 major clades, and, additionally, the tree was forced into one of the following 3 topologies:

- deep ([B],[E],[{CA},{EA}]): Crenarchaeota group with Euryarchaeota to the exclusion of Eukaryota and Bacteria;
- crenarchaeal ([B],[EA],[{CA},{E}]): Eukaryota group with Crenarchaeota to the exclusion of Euryarchaeota and Bacteria;
- euryarchaeal ([B],[CA],[{EA},{E}]): Eukaryota group with Euryarchaeota to the exclusion of Crenarchaeota and Bacteria.

First, all clusters were tested for compatibility with the hypothesis of monophyly of the representatives of all 4 major groups. The AU test *P* values for the constrained trees were compared with that of the unconstrained tree; if at least 1 of the 3 *P* values was greater than the cutoff (0.05), the cluster was deemed to consist of monophyletic major groups. All 136 sets passed this test.

Second, log-likelihood values, ELW values, and AU test *P* values were recorded for the comparison of the 3 constrained topologies that represent 3 distinct evolutionary scenarios for the respective eukaryotic genes (supplementary table S1, Supplementary Material online). These data were subject to further statistical analysis. Most of the analyzed clusters did not contain enough phylogenetic information to unambiguously resolve the position of the eukaryotic branch. In 9 clusters, only 1 of the 3 competing topologies passed the AU test (5 deep and 4 crenarchaeal). In another 8 clusters, one of the topologies was rejected (1 deep, 4 crenarchaeal, and 3 euryarchaeal). For the rest of the clusters, all 3 topologies passed the *P* value threshold of 0.05, that is, none of the topologies could be rejected.

The lack of statistical power in the analysis of most of the individual clusters does not preclude a meaningful examination of trends in the data set as a whole. When plotted on a simplex, ELW values indicate the tendency of the entire data set to resolve into each of the 3 competing topologies (fig. 3). Within a wide range of threshold ELW values (0.33–0.85), the ratio of clusters for which the tree topology was best compatible with the deep, crenarchaeal, and euryarchaeal origins of the respective eukaryotic genes was fairly stable and close to 0.54:0.30:0.16 (33:18:10 clusters at the threshold ELW value of 0.6; fig. 4). Log-likelihood values for each of the 3 topologies, summed over the 136 clusters, indicate the overall likelihood support for the corresponding evolutionary scenarios (−1668665.05, −1668864.95, and −1669217.11 for the deep, crenarchaeal, and euryarchaeal topologies, respectively). Thus, the deep topology had, marginally, the greatest total likelihood. We performed 10,000 bootstrap resamplings of the set of 136 clusters of orthologous genes to assess the robustness of this result and found that 73% of the bootstrap samples supported the overall dominance of the deep topology, 27% supported the crenarchaeal scenario, and none ($P < 1 \times 10^{-4}$) supported the euryarchaeal affinity of the eukaryotic branch. Thus, even in those trees where the euryarchaeal topology was formally preferred, the signal was weak.

⟵

FIG. 2.—Functional classification of ancestral eukaryotic genes of different probable origins. (*a*) Distribution of ancestral eukaryotic genes of different inferred origins by functional categories. The functional classes are as in the COG database: C, energy production and conversion; D, cell division; E, amino acid metabolism and transport; F, nucleotide metabolism and transport; G, carbohydrate metabolism and transport; H, coenzyme metabolism; I, lipid metabolism; J, translation; K, transcription; L, replication and repair; O, posttranslational modification, protein turnover, and chaperone functions; Q, secondary metabolism; T, signal transduction; U, intracellular trafficking and secretion; V, defense and resistance; R, general functional prediction only (typically, prediction of biochemical activity); and S, function unknown. (*b*) Fractions of arCOGs belonging to different functional classes in the set of 975 A–E pairs and in the set of 351 eukaryotic genes of inferred archaeal origin. (*c*) Log-odds ratio of the fraction of arCOGs belonging to different functional classes relative to the fraction of arCOGs that belong to the A–E set and the set of eukaryotic genes of inferred archaeal origin (1,008/7,538 and 286/7,538, respectively). The log base is 2.

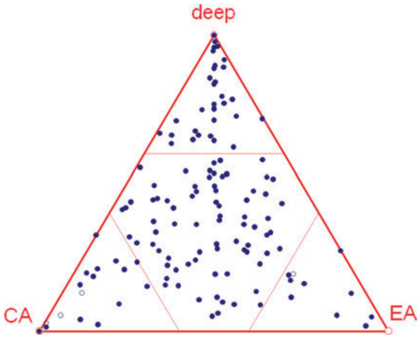


FIG. 3.—The evolutionary relationship between archaea and eukaryotes assessed by phylogenetic analysis of 136 A–E–B clusters. The ELW values are plotted on a simplex surface.

The principal problem with the interpretation of these results is the possibility that the deep archaeal origin of eukaryotes is an LBA artifact (Felsenstein 2004), perhaps, caused by an acceleration of evolution of genes derived from the archaeal parent during the emergence of eukaryotes, such that genes of euryarchaeal or crenarchaeal origin are placed below the base of the archaeal subtree. The ML tree methods have been shown to be less sensitive to LBA within a wide range of evolutionary rate variation than other phylogenetic methods although major differences in branch length still lead to inaccurate phylogenies (Anderson and Swofford 2004; Philippe et al. 2005). However, the fact that we found roughly the same number of crenarchaeal and euryarchaeal trees, and no obvious functional differences between genes that yielded the 3 subsets of trees seem to argue against a major contribution of LBA to the observed breakdown of trees. Indeed, should, for example, the crenarchaeal topology be correct, one would expect to find 2 groups of trees: accurate ones, with a eukaryotic–crenarchaeal clade, and those distorted by LBA, with the eukaryotic branch positioned outside the divergence of Creanrchaeota and Euryarchaeota; the euryarchaeal trees are not predicted by this scenario. Furthermore, we found no significant differences in the long-term evolutionary rates (relative to the respective bacterial orthologs) of the genes that showed the euryarchaeal, crenarchaeal, and deep topologies (supplementary fig. S3, Supplementary Material online). This seems to be poorly compatible with acceleration of evolution being the major underlying cause behind the prevalence of the deep topology: indeed, should that be the case, one would expect significantly higher rates to be detected for the genes with the deep topology compared with those with, at least, one of the other 2 topologies.

Shared Derived Characters in Archaea and Eukaryotes

In an attempt to obtain additional evidence of eukaryotic origins, we searched the 136 A–E–B alignments for possible shared derived characters (synapomorphies) that could help resolve the phylogenetic affinities among the 4 groups. Operationally, we looked for insertions–deletions (indels) where a gap was shared between bacteria and one of the other 3 groups (ensuring the derived state of the insert), whereas the remaining 2 groups shared an apparently homologous insert. The approach is similar to that previously employed by others (Rivera and Lake 1992; Griffiths and Gupta 2001) except that special attention was given to the sequence conservation in the insert itself. The position of the gap was validated by the presence of highly conserved "anchoring" alignment positions in the vicinity of both ends of the indel (for details, see Materials and Methods). Four potential synapomorphies supported the monophyly of archaea relative to bacteria and eukaryotes (inserts shared between EA and CA), which correspond to the deep archaeal origin of the respective eukaryotic genes, whereas one synapomorphy favored grouping of eukaryotes with crenarchaea and one synapomorphy grouped eukaryotes with euryarchaea (supplementary fig. S2, Supplementary Material online). In all 6 cases, the synapomorphies agreed with the most likely tree topology that was inferred on the basis of the highest ELW value for the respective cluster.

Other Eukaryotic Genes of Archaeal Origin

For 221 (147 + 74) clusters of orthologs that included eukaryotic genes of apparent archaeal origin, the detailed

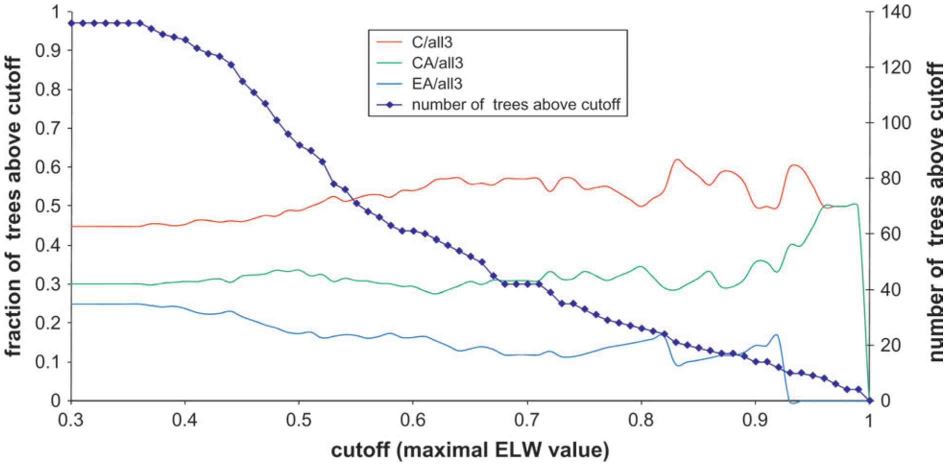


FIG. 4.—Distribution of the deep, crenarchaeal, and euryarchaeal topologies depending on the ELW value cutoff.

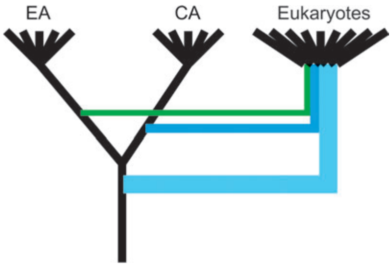


FIG. 5.—A cartoon representation of the deep, crenarchaeal, and euryarchaeal inferred origins of the archaeal genes in eukaryotes.

phylogenetic analysis described above was not feasible due to the low number of representatives of one or more of the major groups (CA, EA, or B) (fig. 1*a* and *b*). Specifically, CA or EA were considered to be "absent" when a cluster included fewer than 4 crenarchaeal or 10 euryarchaeal genes (in which case no PhyML tree was built) or when CA or EA did not represent a clade group in the PhyML tree. Similarly, clusters that included no or only one bacterial representative were classified as "no Bacteria" without phylogenetic analysis; clusters with 2 or more bacterial members were classified as "no Bacteria" when the respective PhyML trees included no credible bacterial outgroup. Among these, 42 clusters did not include Crenarchaeota, 21 clusters did not include Euryarchaeota, whereas in 158 clusters, both Crenarchaeota and Euryarchaeota were well represented but there was no apparent ancestral bacterial clade. For the former 2 categories, the crenarchaeal and euryarchaeal origin, respectively, of the eukaryotic genes could be inferred although, formally, the loss of the corresponding gene in the other archaeal branch could not be ruled out. For the remaining 158 clusters, determination of the specific archaeo–eukaryotic relationship was, generally, not feasible as there is no information on the root position in a tree of 3 clades. Therefore, trees where crenarchaea, euryarchaea, and eukaryotes are monophyletic are bound to remain unresolved. However, there might be special cases where the eukaryotic clade would fall within one of the archaeal clades (Crenarchaeota or Euryarchaeota). Thus, the 144 CA–EA–E clusters (excluding 4 short alignments) were analyzed to identify such situations. For each of the data sets, 2 ML trees were constructed, namely, an unconstrained tree and a tree constrained for the monophyly of the 3 major groups (CA, EA, and E). The 2 trees were compared using the AU test. If the constrained topology was rejected, this would indicate that one of the groups was paraphyletic with respect to another, allowing for the elucidation of the origin of the respective eukaryotic genes. All 144 trees passed the test for monophyly of each of the 3 groups, so the origin of the respective eukaryotic genes remains uncertain. Clearly, this result is compatible with the deep version of the archaeo–eukaryotic relationship.

Discussion

Ancient phylogenetic relationships are notoriously hard to decipher, and it is often argued, in some cases, with compelling demonstrations, that sequences of very distant orthologous genes, such as those from archaea and bacteria, do not carry sufficient phylogenetic signal for a meaningful tree reconstruction (Penny et al. 2001; Rokas et al. 2005; Rokas and Carroll 2006; Doolittle and Bapteste 2007). Genome trees that strive to utilize combined information from multiple genes on the genome scale have the potential to overcome these limitations but they run into their own major difficulties, primarily, because the evolutionary histories of different genes differ, often, substantially, and not only can it be technically challenging to derive a consensus but also the very legitimacy of such an exercise is dubious (Wolf et al. 2002; Bapteste et al. 2005; Snel et al. 2005; McInerney et al. 2008). Furthermore, it has been argued that early, transitional stages in the evolution of major divisions of life might not be amenable to a description through the tree metaphor in principle (Doolittle and Bapteste 2007; Koonin 2007; McInerney et al. 2008). However, all these formidable obstacles notwithstanding, the problems that are at stake when the deepest evolutionary relationships are examined are too fundamental to abandon attempts to extract the most of the modern phylogenetic approaches combined with comprehensive comparative genomic data.

The nature of the relationship between eukaryotes and archaea, arguably, is one of the most important and hardest problems in the reconstruction of the evolutionary history of life. Here we attempted to approach this problem by delineating the set of archaeo–eukaryotic orthologous genes as completely as possible and identifying the subset that is amenable to phylogenetic analysis aimed at resolution of the archaeo–eukaryotic relationships. At the first step, we identified ~1,000 archaeo–eukaryotic orthologous gene sets of which ~350 were classified as being of probable archaeal origin. As expected from previous studies (Koonin et al. 1997; Rivera et al. 1998; Esser et al. 2004), this archaeal gene set was strongly enriched in genes for information-processing system components and encompassed substantial parts of the core machineries of translation, transcription, and replication, suggesting that this is a reasonable, although conservative approximation of the archaeal contribution to eukaryogenesis. Due to either poor representation of bacterial orthologs or poor tree resolution, we ended up with only 136 archaeal genes that carried enough phylogenetic information to weigh in on the specifics of the relationships of Eukaryota with Euryarchaeota and Crenarchaeota. Although this is a relatively small number, it seems unlikely that many more genes can be rescued for future phylogenetic analysis to approach this problem. Thus, detailed phylogenetic analysis of this gene set is likely to approximate the best shot on resolving the archaeo–eukaryotic relationship.

We addressed this problem by using, primarily, the constrained tree method, which is a procedure for directly comparing the likelihoods of competing phylogenetic scenarios, and buttressed this approach by analysis of putative shared derived characters. None of these approaches supported the origin of eukaryotic genes, en masse, from either Euryarchaeota or Crenarchaeota although evidence of origin from each of these major archaeal divisions was obtained for relatively small minorities of the genes, with some excess of the crenarchaeal affinity. The favored topology is one where eukaryotic genes derive from a distinct, ancient archaeal branch that lies outside the extant diversity of archaea (fig. 5). The possibility of an LBA artifact could be a concern in the interpretation of these results. However, we found roughly the same number of crenarchaeal and euryarchaeal trees, and there was no systematic functional difference between genes that yielded the 3 subsets of trees, suggesting that LBA was not a major factor in the observed breakdown of trees. Furthermore, there was no significant difference in the evolutionary rates of the genes that showed the euryarchaeal, crenarchaeal, and deep topologies, an observation that, again, is poorly compatible with widespread LBA. In addition, shared derived characters that unite Euryarchaeota and Crenarchaeota to the exclusion of eukaryotes comprise independent evidence in support of the notion that the archaeal genes of eukaryotes branch off the trunk of the archaeal tree prior to the radiation of Euryarchaeota and Crenarchaeota.

Taken together, the results of the present analysis suggest that the archaeal genes in eukaryotes derive from an ancient, probably, extinct, and in any case, unknown archaeal lineage. There are relatively small subsets of archaeal genes for which either euryarchaeal or crenarchaeal origin appeared most likely (figs. 2*b* and 5). Under the hypothesis of deep archaeal roots of eukaryotic genes, these observations can be explained by HGT from Crenarchaeota and Euryarchaeota to the hypothetical eukaryotic ancestor, most likely, prior to the eukaryogenesis. However, these results are incompatible with those hypotheses that specifically derive the archaeal component of eukaryotes from one of the known branches of archaea, namely, the eocyte (crenarchaeal origin) hypothesis (Lake et al. 1984; Lake 1988, 1998), those versions of the hydrogen and syntrophic hypotheses that link eukaryotes to the known euryarchaeal methanogens (Moreira and Lopez-Garcia 1998; Lopez-Garcia and Moreira 1999), or the syntrophic hypotheses that derive the archaeal host of the mitochondrion from *Thermoplasmatales* (Searcy et al. 1978; Margulis and Stolz 1984; Hixon and Searcy 1993; Margulis 1996; Margulis et al. 2000; Pisani et al. 2007).

The nature of the archaeo–eukaryotic relationship has implications for the scenario of eukaryotic origins. Had it been found that the archaeal genes in eukaryotes originated from within the extant archaeal diversity, the archezoan hypothesis would have been, effectively, falsified. The present finding that the majority of archaeal genes in eukaryotes appear to derive from a distinct, ancient archaeal (or archaea-related) lineage is compatible with both archaeal and archezoan origins of eukaryotes. These results do not, in themselves, lend support to the archezoan hypothesis because all genes analyzed here encode typical archaeal proteins, so there is no implication that the archaea-related parent of eukaryotes was not a bona fide archaeon. Furthermore, recent genomic and phylogenetic analyses of "mesophilic Crenarchaeota" and Korarchaeota indicate that, although these groups encompass typical archaea, they represent deep branches that join the archaeal tree either close to or above the point of radiation of Euryarchaeota and Crenarchaeota (Brochier-Armanet et al. 2008; Elkins et al. 2008). The choice between the "deep archaeal" and archezoan scenarios requires further exploration of the archaeal diversity and, perhaps, even more importantly, a thorough analysis of the origins of the proteins that comprise the signature eukaryotic cellular structures.

Supplementary Material

Supplementary methods, tables S1 and S2, and figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Literature Cited

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Anderson FE, Swofford DL. 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. Mol Phylogenet Evol. 33:440–451.

Bapteste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF. 2005. Do orthologous gene phylogenies really support tree-thinking? BMC Evol Biol. 5:33.

Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P. 2008. Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat Rev Microbiol. 6:245–252.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science. 311:1283–1287.

Dagan T, Martin W. 2007. Testing hypotheses without considering predictions. Bioessays. 29:500–503.

Doolittle WF, Bapteste E. 2007. Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci USA. 104:2043–2049.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Elkins JG, Podar M, Graham DE, et al. (20 co-authors). 2008. A korarchaeal genome reveals new insights into the evolution of the Archaea. Proc Natl Acad Sci USA. doi: 10.1073/pnas.0801980105

Embley TM. 2006. Multiple secondary origins of the anaerobic lifestyle in eukaryotes. Philos Trans R Soc Lond B Biol Sci. 361:1055–1067.

Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. Nature. 440:623–630.

Esser C, Ahmadinejad N, Wiegand C, et al. (15 co-authors). 2004. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol Biol Evol. 21:1643–1660.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Fukami-Kobayashi K, Minezaki Y, Tateno Y, Nishikawa K. 2007. A tree of life based on protein domain organizations. Mol Biol Evol. 24:1181–1189.

Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. Mol Biol Evol. 19:2226–2238.

Gray MW. 1992. The endosymbiont hypothesis revisited. Int Rev Cytol. 141:233–357.

Gray MW, Burger G, Lang BF. 2001. The origin and early evolution of mitochondria. Genome Biol. 2: reviews1018.1–1018.5.

Griffiths E, Gupta RS. 2001. The use of signature sequences in different proteins to determine the relative branching order of bacterial divisions: evidence that Fibrobacter diverged at a similar time to Chlamydia and the Cytophaga-Flavobacterium-Bacteroides division. Microbiology. 147: 2611–2622.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hartman H, Favaretto P, Smith TF. 2006. The archaeal origins of the eukaryotic translational system. Archaea. 2:1–9.

Hedges SB, Chen H, Kumar S, Wang DY, Thompson AS, Watanabe H. 2001. A genomic timescale for the origin of eukaryotes. BMC Evol Biol. 1:4.

Hixon WG, Searcy DG. 1993. Cytoskeleton in the archaebacterium Thermoplasma acidophilum? Viscosity increase in soluble extracts. Biosystems. 29:151–160.

Horiike T, Hamada K, Miyata D, Shinozawa T. 2004. The origin of eukaryotes is suggested as the symbiosis of pyrococcus into gamma-proteobacteria by phylogenetic tree based on gene content. J Mol Evol. 59:606–619.

Jobb G, von Haeseler A, Strimmer K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol. 4:18.

Karlin S, Brocchieri L, Mrazek J, Campbell AM, Spormann AM. 1999. A chimeric prokaryotic ancestry of mitochondria and primitive eukaryotes. Proc Natl Acad Sci USA. 96:9190–9195.

Koonin EV. 2007. The Biological Big Bang model for the major transitions in evolution. Biol Direct. 2:21.

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

Kurland CG, Collins LJ, Penny D. 2006. Genomics and the irreducible nature of eukaryote cells. Science. 312:1011–1014.

Lake JA. 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. Nature. 331: 184–186.

Lake JA. 1998. Optimally recovering rate variation information from genomes and sequences: pattern filtering. Mol Biol Evol. 15:1224–1231.

Lake JA, Henderson E, Oakes M, Clark MW. 1984. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc Natl Acad Sci USA. 81:3786–3790.

Lake JA, Rivera MC. 1994. Was the nucleus the first endosymbiont? Proc Natl Acad Sci USA. 91:2880–2881.

Leipe DD, Aravind L, Koonin EV. 1999. Did DNA replication evolve twice independently? Nucleic Acids Res. 27:3389–3401.

Lopez-Garcia P, Moreira D. 1999. Metabolic symbiosis at the origin of eukaryotes. Trends Biochem Sci. 24:88–93.

Lopez-Garcia P, Moreira D. 2006. Selective forces for the origin of the eukaryotic nucleus. Bioessays. 28:525–533.

Makarova KS, Sorokin AV, Novichkov PS, Wolf YI, Koonin EV. 2007. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. Biol Direct. 2:33.

Makarova KS, Wolf YI, Mekhedov SL, Mirkin BG, Koonin EV. 2005. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. Nucleic Acids Res. 33:4626–4638.

Margulis L. 1996. Archaeal-eubacterial mergers in the origin of Eukarya: phylogenetic classification of life. Proc Natl Acad Sci USA. 93:1071–1076.

Margulis L, Dolan MF, Guerrero R. 2000. The chimeric eukaryote: origin of the nucleus from the karyomastigont in amitochondriate protists. Proc Natl Acad Sci USA. 97:6954–6959.

Margulis L, Stolz JF. 1984. Cell symbiosis theory: status and implications for the fossil record. Adv Space Res. 4:195–201.

Martin W, Koonin EV. 2006. Introns and the origin of nucleus-cytosol compartmentation. Nature. 440:41–45.

Martin W, Muller M. 1998. The hydrogen hypothesis for the first eukaryote. Nature. 392:37–41.

McInerney JO, Cotton JA, Pisani D. 2008. The prokaryotic tree of life: past, present.and future? Trends Ecol Evol. 23:276–281.

Moreira D, Lopez-Garcia P. 1998. Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. J Mol Evol. 47: 517–530.

Penny D, McComish BJ, Charleston MA, Hendy MD. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J Mol Evol. 53:711–723.

Pereto J, Lopez-Garcia P, Moreira D. 2004. Ancestral lipid biosynthesis and early membrane evolution. Trends Biochem Sci. 29:469–477.

Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol. 5:50.

Pisani D, Cotton JA, McInerney JO. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol. 24:1752–1760.

Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst Biol. 51:664–671.

Poole A, Penny D. 2007a. Eukaryote evolution: engulfed by speculation. Nature. 447:913.

Poole AM, Penny D. 2007b. Evaluating hypotheses for the origin of eukaryotes. Bioessays. 29:74–84.

Rivera MC, Jain R, Moore JE, Lake JA. 1998. Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci USA. 95:6239–6244.

Rivera MC, Lake JA. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science. 257:74–76.

Rivera MC, Lake JA. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature. 431:152–155.

Rokas A, Carroll SB. 2006. Bushes in the tree of life. PLoS Biol. 4:e352.

Rokas A, Kruger D, Carroll SB. 2005. Animal evolution and the molecular signature of radiations compressed in time. Science. 310:1933–1938.

Searcy DG, Stein DB, Green GR. 1978. Phylogenetic affinities between eukaryotic cells and a thermophilic mycoplasma. Biosystems. 10:19–28.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 51:492–508.

Snel B, Huynen MA, Dutilh BE. 2005. Genome trees and the nature of genome evolution. Annu Rev Microbiol. 59:191–209.

Strimmer K, Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. Proc Biol Sci. 269:137–142.

Tatusov RL, Fedorova ND, Jackson JD, et al. (17 co-authors). 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 4:41.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28:33–36.

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

Tekaia F, Yeramian E. 2005. Genome trees from conservation profiles. PLoS Comput Biol. 1:e75.

van der Giezen M, Tovar J. 2005. Degenerate mitochondria. EMBO Rep. 6:525–530.

Vellai T, Takacs K, Vida G. 1998. A new aspect to the origin and evolution of eukaryotes. J Mol Evol. 46:499–507.

Vishwanath P, Favaretto P, Hartman H, Mohr SC, Smith TF. 2004. Ribosomal protein-sequence block structure suggests complex prokaryotic evolution with implications for the origin of eukaryotes. Mol Phylogenet Evol. 33:615–625.

Wolf YI, Rogozin IB, Grishin NV, Koonin EV. 2002. Genome trees and the tree of life. Trends Genet. 18:472–479.

Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol. 51:588–598.

Aoife McLysaght, Associate Editor

Accepted May 1, 2008