Gene Transfers from Nanoarchaeota to an Ancestor of Diplomonads and Parabasalids

Jan O. Andersson,*† Stewart W. Sarchfield,* and Andrew J. Roger*†

*The Canadian Institute for Advanced Research, Program in Evolutionary Biology, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada; and †Institute of Cell and Molecular Biology, Uppsala University, Biomedical Center, Uppsala, Sweden

Rare evolutionary events, such as lateral gene transfers and gene fusions, may be useful to pinpoint, and correlate the timing of, key branches across the tree of life. For example, the shared possession of a transferred gene indicates a phylogenetic relationship among organismal lineages by virtue of their shared common ancestral recipient. Here, we present phylogenetic analyses of prolyl-tRNA and alanyl-tRNA synthetase genes that indicate lateral gene transfer events to an ancestor of the diplomonads and parabasalids from lineages more closely related to the newly discovered archaeal hyperthermophile Nanoarchaeum equitans (Nanoarchaeota) than to Crenarchaeota or Euryarchaeota. The support for this scenario is strong from all applied phylogenetic methods for the alanyl-tRNA sequences, whereas the phylogenetic analyses of the prolyl-tRNA sequences show some disagreements between methods, indicating that the donor lineage cannot be identified with a high degree of certainty. However, in both trees, the diplomonads and parabasalids branch together within the Archaea, strongly suggesting that these two groups of unicellular eukaryotes, often regarded as the two earliest independent offshoots of the eukaryotic lineage, share a common ancestor to the exclusion of the eukaryotic root. Unfortunately, the phylogenetic analyses of these two aminoacyl-tRNA synthetase genes are inconclusive regarding the position of the diplomonad/parabasalid group within the eukaryotes. Our results also show that the lineage leading to Nanoarchaeota branched off from Euryarchaeota and Crenarchaeota before the divergence of diplomonads and parabasalids, that this unexplored archaean diversity, currently only represented by the hyperthermophilic organism Nanoarchaeum equitans, may include members living in close proximity to mesophilic eukaryotes, and that the presence of split genes in the Nanoarchaeum genome is a derived feature.

Introduction

Nanoarchaeum equitans is an enigmatic symbiont that attaches to the surface of the crenarchaeon Ignicoccus in submarine hot vents and represents the only species of the newly discovered phylum Nanoarchaeota to be cultivated to date (Huber et al. 2002; Waters et al. 2003). However, a large unexplored archaean diversity may exist; phylogenetic analyses show N. equitans as an early branching archaean (Waters et al. 2003), and nanoarchaeota-like ribosomal RNA samples have been amplified from hot environments around the world (Hohn, Hedlund, and Huber 2002). It is unknown whether this potential diversity also includes representatives in cooler environments or whether hyperthermophily is an ancient and exclusive trait of this phylum.

Like the Archaea, an improved understanding of the deep phylogeny of Eukaryota has been also achieved by the identification of novel unicellular lineages in combination with improved phylogenetic methods. It now seems clear that the eukaryotic diversity falls into at least half a dozen eukaryotic “supergroups” (superkingdoms) (Cavalier-Smith 2002; Simpson and Roger 2002; Baldauf 2003). One of these, Excavata (Cavalier-Smith 2003; Simpson 2003), contains two important human pathogens: Giardia lamblia, a diplomonad commonly causing diarrhea in humans, and Trichomonas vaginalis, a parabasalid that causes trichomoniasis. For the past 2 decades, it has been widely thought that these organisms were the most “primitive” eukaryotes, branching as the first two lineages of eukaryotes in rDNA and other molecular phylogenies (Sogin 1991). However, the deep-branching emergence of diplomonads and parabasalids in phylogenetic trees of eukaryotes has recently been called into question; this phylogenetic position may result from artifacts of the methods employed, rather than reflect their true evolutionary positions (Embly and Hirt 1998; Philippe et al. 2000; Simpson et al. 2002; Cavalier-Smith 2003; Simpson 2003).

Here, we address the phylogenetic relationships of diplomonads and parabasalids by using lateral-gene-transfer events as discrete evolutionary markers, by extending the eukaryotic taxonomic sampling of the prolyl-tRNA and alanyl-tRNA synthetase genes (proS and alaS), which were previously shown to have been transferred from the Archaea to diplomonads (Andersson et al. 2003). Our phylogenetic analyses of the updated data sets indicate that diplomonads and parabasalids shared a most recent common anaerobic protist ancestor that was the recipient of transfers of homologs of both of these genes from an archaean of the Nanoarchaeota lineage. These data indicate that diplomonads and parabasalids shared a common ancestor exclusive of the eukaryotic root, as well as provide insights into the evolution and molecular biology of the archaean phylum Nanoarchaeota. Thus, gene-transfer events may provide a range of biological information unrelated to the encoded functions of the transferred genes.

Material and Methods

PCR Amplifications

Using degenerate primers against conserved parts of the proS and alaS amino acid alignments, with genomic DNA in PCR reactions, we amplified proS and alaS gene sequences from diverse eukaryotes. The proS gene was amplified with PCR from Entamoeba invadens,
E. moshkovskii, E. terrapinae (cell lysates were gifts from C. G. Clark, London School of Hygiene and Tropical Medicine), and the heterolobosean Naegleria gruberi (genomic DNA and cDNA were gifts from Å. Sjögren, Dalhousie University) using the forward primer PROSeu2 (AAYTTYCGNGARTGGTA) and the reverse primers PROSeur2 (TGGCTTCACCCANGNGGC) and PRO-Seur1 (GCNTRTGNCCYTCYTGCCTA) for the Entamoeba species and Naegleria, respectively. The C-terminal part of N. gruberi proS were amplified with PCR from cDNA using a specific internal primer and a primer complimentary to the 3’ end of the cDNA. alaS gene sequences from E. invadens and E. moshkovskii were amplified using the forward primer alaSswsF1 (GGATCTAYTGYTYWCANC) and the reverse primer alaSswsR4 (TGGCTCCNCCRCA-NAYTCT), the N. gruberi alaS was amplified using the forward primers alaSswsBTr1 (CAYCAYAINTTTYTGARATG) and the reverse primer alaSswsBTr2 (TGGCTCCNCCRCA-NAYTCT), and the pelobiont Mastigamoeba balamuthi alaS was amplified using a specific reverse primer based on the partial cDNA clone released to GenBank (gi: 18055573) and the degenerate forward primer alaSBTf1 (genomic DNA and cDNA were gifts from A˚. Sjo¨gren, Medicine), and the heterolobosean C. G. Clark, London School of Hygiene and Tropical Medicine), and the heterolobosean N. gruberi proS and alaS gene sequences were obtained from cDNA clones (gifts from P. Tang, Chang Gung University, Taiwan). The PCR products and the cDNA clones were sequenced using standard methods.

Assembly of the Data Sets

Unpublished proS and alaS sequences were retrieved from various genome projects. Phytophthora sojae (oomycete) and Thalassiosira pseudonana (diatom) sequences were retrieved from the DOE-Joint Genome Institute (http://www.jgi.doe.gov/), sequences from Tetrahymena thermophila (ciliate) sequences were retrieved from Genoscope (http://www.genoscope.cns.fr/), a proS Dictyostelium discoideum (mycetozoa) sequence was retrieved from the Dictyostelium discoideum Genome Project (http://www.uni-koeln.de/dictyostelium), and sequences from the ciliate Paramecium tetraurelia, Entamoeba histolytica, the kinetoplastid Trypanosoma brucei, and the apicomplexan Cryptosporidium parvum, were retrieved from The Institute for Genomic Research (http://www.tigr.org/). Subsequently, 305 and 365 unambiguously aligned amino acid positions of prolyl-tRNA synthetase and alanyl-tRNA synthetase, respectively, were identified. The \( \chi^2 \) tests for deviation of amino acid frequencies implemented in Tree-Puzzle version 5.1 (Strimmer and von Haeseler 1999) were applied. Because none of the commonly used phylogenetic methods are able to deal with compositional heterogeneity in the data (Foster and Hickey 1999), sequences that failed the test were excluded from further analyses, except that the Nequitans and the G. lamblia proS sequences and the Nequitans alaS sequence were retained although they failed the test. In principle, the recently published method for calculating LogDet pairwise distances between amino acid sequences (Thollesson 2004) could overcome the problem with amino acid composition heterogeneity within the data sets, which would make the exclusion of sequences that fail the test unnecessary. However, the relative performance of the LogDet method compared with maximum-likelihood (ML) methods has not yet been rigorously tested for single-gene-size data sets.

Phylogenetic Analyses

We chose to use two protein ML methods in our analyses because they have different advantages. The PROML is an established method that employs a fairly extensive heuristic tree-searching algorithm and is widely used (Felsenstein 1989). By contrast, the PHYML method is a much newer method with an ultrarapid (but less extensive) heuristic tree-searching algorithm that has not yet been applied to a wide array of data sets. It is much faster than PROML, thereby allowing a larger number of bootstrap replicates to be evaluated (Guindon and Gascuel 2003). ML trees were inferred using PROML within the PHYLIP version 3.6a3 program (Felsenstein 1989), using the Jones-Taylor-Thorton (JTT) substitution model, a mixed four-category discrete-gamma model of among-site rate variation plus invariable sites (JTT + Inv) and 10 jumbles with global rearrangements. The \( \Gamma \) shape parameter, \( \alpha \), and the fraction of invariable sites, \( P_{inv} \), were estimated by using Tree-Puzzle version 5.1 (Strimmer and von Haeseler 1996). Protein ML bootstrap values based on 500 resampled data sets generated using SEQBOOT within the PHYLIP package (Felsenstein 1989) were calculated using PHYML version 2.1b1 (Guindon and Gascuel 2003), with JTT + Inv models. Bootstrap support values using other methods and models were calculated for comparison. Protein ML bootstrap values were calculated with PROML with JTT + Inv models based on 100 resampled data sets. All other analyses were based on 500 resampled data sets. Protein ML bootstrap values assuming uniform site rates were calculated using PHYML (Guindon and Gascuel 2003), and protein ML distance bootstrap values with JTT + Inv models were calculated using PUZZLEBOOT (http://hades.biochem.dal.ca/Rogerlab/). LogDet amino acid distance bootstrap support values were calculated with the LDDist Perl module using the companion Perl script PLD.pl (Thollesson 2004) with a model of invariable sites and one category of variable site rates. The fractions of invariable sites were calculated using Tree-Puzzle version 5.1 (Strimmer and von Haeseler 1996). The distance matrices were analyzed using FITCH within the PHYLIP package with global rearrangements and one jumble. LogDet analyses with more site rate categories were not performed, because the sequence lengths in our data sets were judged too small to give valid results when divided into several rate classes (Thollesson 2004; M. Thollesson, personal communication). Finally, maximum-parsimony bootstrap support values were calculated using PROTPARS (Felsenstein 1989).

Results and Discussion

Interdomain Gene Transfers of proS and alaS

Previously we showed that the genes encoding prolyl-tRNA synthetase and alanyl-tRNA synthetase, respectively,
were most likely laterally transferred from Archaea to diplomonads (Andersson et al. 2003). To explore these interdomain transfers further, we have increased the eukaryotic taxonomic sampling for these genes to include sequences from three additional Entamoeba species, the heterolobosean Naegleria, the pelobiont Mastigamoeba, the parabasalid Trichomonas, the diatom Thalassiosira, the oomycete Phytophthora, and the ciliates Paramecium and Tetrahymena. On the whole, the phylogenetic analyses on the updated data sets of the two proteins agree fairly well with expected organismal phylogeny, with only a few easily identified exceptions (fig. 1). Only two of these unexpected branching patterns with high bootstrap support are observed among the prokaryotes, both in the proS tree; the Pirellula sequence shows close relationship with α-proteobacteria and the Halobacterium sequence is found in a distinct position from the other euryarchaeota sequences (fig. 1A). Surprisingly, the eukaryotes show up in a handful unexpected positions. The D. discoideum and P. sojae alas sequences are found outside the main eukaryotic clade (fig. 1B). Unfortunately, the statistical support for the separation is weak and, therefore, the origins of these sequences are
uncertain. Both proS and alaS plant sequences are found nested within the Eubacteria with strong support (fig. 1), most likely indicating two interdomain gene transfer events—the alaS sequence almost certainly via the chloroplast. Finally, a subset of the eukaryotes is found nested within the Archaea in both trees (fig. 1), strongly suggesting transfer of the genes between Archaea and eukaryotes. The seemingly similar frequency of observed gene transfer events affecting eukaryotes compared with prokaryotes for these two aminoacyl-tRNA genes may be somewhat surprising (fig. 1). However, we previously reported that gene transfer appears to have affected prokaryotes and microbial eukaryotes to a similar extent in the glutamate dehydrogenase gene families (Andersson and Archibald et al. 2003; Bergthorsson et al. 2003), indicating that lateral gene transfer may indeed be a widespread evolutionary mechanism in microbial eukaryotes (Andersson et al. 2003; Archibald et al. 2003; Gogarten 2003).

A Single Archaeal Origin of the Diplomonad and Parabasalid proS

The topology with a eukaryotic clade including parabasalid and diplomonad sequences nested with the Archaea to the exclusion of other eukaryotes is strongly supported in the proS tree by all phylogenetic methods, likely indicating an interdomain gene-transfer event from the Archaea to a common ancestor of these eukaryotic groups. The specific archaeal origin of the proS gene in diplomonads and parabasalids is more difficult to identify because of large disagreements between the results from the various phylogenetic methods used. A specific relationship between the N. equitans and the eukaryotic sequences is supported by a bootstrap value of 73% in the ML analysis, and the relationship is recovered by the other ML methods, as well as with parsimony (fig. 1A). However, the LogDet distance analysis disagrees. The Nanoarchaeum relationship is recovered in only 25% of the bootstrap replicates (fig 1A), whereas 52% of the replicates place the eukaryotic sequences basal to the archaea (data not shown)—a position that is only found in 4% of the replicates in the ML bootstrap analyses that incorporate the gamma model of rate heterogeneity (data not shown). Because both the N. equitans and the G. lambia sequences failed the test for amino acid compositional heterogeneity applied to the data sets, and the applied phylogenetic methods assume a uniform amino acid composition within the data set (with the exception of LogDet analysis), the specific relationship between the Nanoarchaeota, diplomonad, and parabasalid sequences could be the result of an artifactual attraction caused by the amino acid compositional heterogeneity. On the other hand, a model that incorporates rate heterogeneity within the data set could not be applied to the LogDet analysis, because that would require a larger data set (Thollesson 2004; M. Thollesson, personal communication), and the attraction between the diplomonad and parabasalid sequences—which represent the longest branches within the archaea/diplomonad/parabasalid subtree (fig. 1A)—and the long internal branch in this analysis could be an artifact caused by the absence of a rate heterogeneity model. Obviously, in the absence of an efficient method that simultaneously can incorporate models of rate and amino acid heterogeneities, the relationship within the archaea/diplomonad/parabasalid subtree cannot be resolved with a high degree of confidence for proS.

Gene-Transfer Events Have Distributed Archaeal alaS in Diverse Microbial Eukaryotes

The eukaryotic group showing the largest diversity within the alaS tree is found within eubacteria, and a eukaryotic cluster including diplomonads, parabasalid, Entamoeba, and ciliate sequences is found as a sister group to the N. equitans sequence with high bootstrap support (≥98% for all methods [fig. 1B]). Given current accounts of eukaryotic phylogeny (Baldauf et al. 2000; Baptiste et al. 2002; Cavalier-Smith 2002; Simpson and Roger 2002; Baldauf 2003), this topology is most easily explained by a gene transfer to a common ancestor of diplomonads and parabasalids, followed by two eukaryote-to-eukaryote gene transfers that replaced the ancient eukaryotic version in the ciliate and Entamoeba lineages. Specifically, the well-supported relationship in ML and parsimony analyses between the Trichomonas sequence and the Entamoeba and ciliate sequences to the exclusion of diplomonads in the alaS tree suggests that a parabasalid was the donor eukaryotic lineage for the first of the two eukaryote-to-eukaryote gene transfers. Interestingly, the LogDet analysis places the Entamoeba and ciliate sequences as a sister clade to the diplomonads with a high bootstrap support (92%). Because all sequences in this cluster passed the test for amino acid compositional heterogeneity, the inconsistency between ML and LogDet analyses probably is explained by the absence of a model incorporating rate heterogeneity in the latter—the Trichomonas sequence represents the longest branch in the cluster and, indeed, is attracted to the root of the cluster in the LogDet analysis. Although only a few lateral gene transfers between eukaryotes have been described (Andersson et al. 2003; Archibald et al. 2003; Berghthorson et al. 2003), the inferred intradomain transfers should not be surprising, because both Entamoeba and ciliates are phagotrophic lineages that may ingest both microbial eukaryotes and prokaryotes. The alternative hypothesis—that both alaS versions were present in the last common eukaryotic ancestor and subsequently differentially lost—is much less likely for several reasons. Ciliates, Entamoeba, and parabasalids/diplomonads are specifically related to the apicomplexa, pelobiont, and heterolobosea, respectively (Baldauf et al. 2000; Baptiste et al. 2002; Cavalier-Smith 2002; Simpson and Roger 2002; Baldauf 2003), which all have the “bacterial” version (fig. 1B). Plasmodium falciparum (apicomplexa) and M. balamuthi (pelobiont) sequences were excluded from the phylogenetic analyses because of strong amino acid heterogeneity and short length, respectively (data not shown). Thus, many parallel independent losses have to be posited if the “archaeal” version were ancestral to all eukaryotes. Also, all extant eukaryotes have either the “bacterial” or the “archaeal” version of alaS—none has been found to encode both—which argues against retention of both versions in a single
genome over a long evolutionary timescale, as required by an “ancient paralogy and differential loss” scenario.

Transfer of Two Unsplit Nanoarchaeota Genes in a Single Event?

A gene-transfer ratchet mechanism could explain the presence of the two archaeal genes in the eukaryotic genomes (Doolittle 1998); the food ingested by the common ancestor of diplomonads and parabasalids may have been rich in members of the Nanoarchaeota or close relatives to the phylum, or a symbiotic relationship between such an organism and the eukaryote may have existed. Noticeably, the only described species from Nanoarchaeota lives as a symbiont with a crenarchaeon (Huber et al. 2002; Waters et al. 2003). However, the organization of the genes in Nanoarchaeum hints that the transfer of the two tRNA synthetase genes may have occurred in a single rather than in multiple events. The N. equitans alaS gene is one of several in that genome shown to be “split” into two noncontiguous pieces (Waters et al. 2003). Curiously, the gene corresponding to the C-terminus of alaS shows a close genetic linkage to the proS gene; they are separated by only a single gene encoding a hypothetical protein in the Nanoarchaeum genome. If the ancestral unsplit alaS gene in the nanoarchaibial lineage was located in the current position of the C-terminal gene, the transfer of a single DNA fragment would be sufficient to transfer of both alaS and proS in a single event to a common ancestor of diplomonads and parabasalids. The fact that Nanoarchaeum is the only archaeal genome (among 18 full-genome sequences available from this domain) that shows such a close linkage of these two genes circumstantially supports this scenario.

A Common Ancestor of Diplomonads and Parabasalids to the Exclusion of the Root

These findings have several additional important implications. Both diplomonads (Chihade et al. 2000) and parabasalids (Keeling and Palmer 2000) have each been suggested, individually, to represent the deepest eukaryotic branch, and rDNA phylogenies have long depicted them as the two earliest emerging groups (Sogin 1991). Our data indicate that none of these proposals is correct—the presence of two aminoacyl-tRNA synthetase genes of archaeal ancestry are shared derived features that distinguish them from the other eukaryotes included in the study, indicating that they share a common ancestor. Thus, the root of eukaryotes can neither lie on the branch leading to diplomonads nor lie on the branch leading to parabasalids. Although this has been proposed previously (Embley and Hirt 1998; Cavalier-Smith 2002; Simpson and Roger 2002; Baldauf 2003; Cavalier-Smith 2003; Simpson 2003), the support from phylogenetic analyses has been relatively weak (Henze et al. 2001; Simpson et al. 2002; Cavalier-Smith 2003; Simpson 2003). Our data do not directly bear on the phylogenetic position of the diplomonad/parabasalid group within the eukaryotes. Nevertheless, the confirmation of the specific relationship between diplomonads and parabasalids will deepen the understanding of the evolution of two important human pathogens, G. lamblia and T. vaginalis, the genomes of which will both be completely sequenced shortly. For instance, this sister group relationship suggests that hydrogenosomes, mitochondrial-derived, hydrogen-evolving energy-generating organelles in T. vaginalis and the recently discovered mitochondrial remnant organelles (mitosomes) in G. lamblia (Tovar et al. 2003) may have common anaerobic ancestry. Indeed, because most extant archaeal lineages, including N. equitans, exist in oxygen-poor environments like those inhabited by free-living diplomonads and parabasalids, the transfers probably occurred in an anaerobic ancestor of these two protist lineages that could have already begun to lose canonical aerobic mitochondrial functions.

Ancient Gene Transfers Provide Insights into Nanoarchaeota

The phylogenies of the two transferred genes also indicate that the lineage leading to Nanoarchaeota diverged from Crenarchaeota and Euryarchaeota before the divergence between diplomonads and parabasalids. Moreover, the transfer of a continuous nanoarchaeon alaS gene to a eukaryote indicates that the presence of split noncontiguous genes—one of which is alaS—on the genome of N. equitans likely is a derived feature, rather than a reflection of the ancestral state of genes in early microbial evolution. Thus, the split N. equitans genes are probably not indicators that the lineage represents “a living microbial fossil” (Thomson et al. 2004). Unfortunately, it remains unclear whether the divergent nature of the N. equitans sequences is a consequence of the symbiotic lifestyle of the lineage (Boucher and Doolittle 2002) or indicates a truly ancient origin within the Archaea; further phylogenetic studies are needed to confirm the phylogenetic position of Nanoarchaeota within Archaea. In any case, our results suggest that there have been mesophilic archaea that are closer relatives to Nanoarchaeota than to Crenarchaeota or Euryarchaeota because the common ancestor of diplomonads and parabasalids most likely was a mesophile (Cavalier-Smith 2002) and physical proximity of the organisms is likely an important factor that vastly increases the probability of successful gene-transfer events. However, a transfer from a hyperthermophile to a mesophile living close to a hyperthermophilic environment cannot be excluded, and further studies are needed to clarify whether mesophilic organisms related to Nanoarchaeota still exist and whether they have symbionts with eukaryotes. Hopefully, further examples of interdomain lateral gene transfer discovered from genomic sequences will continue to resolve the phylogeny within prokaryotes and eukaryotes and allow us to determine the relative timing of major evolutionary events in disparate regions of the tree of life.

Supplementary Material

The sequences obtained in this study were deposited under GenBank accession numbers AY568067 to AY568076.
Acknowledgments

We thank P. Tang, E. Gill, and C. G. Clark for generous gifts of cDNA clones, genomic DNA, and cell lysates, respectively. We also thank Å. Sjögren for experimental assistance and critical reading of the manuscript. A.J.R. is supported by the Canadian Institute for Advanced Research, Program in Evolutionary Biology. This work was supported by a Canadian Institutes of Health Research (CIHR) Grant (MOP-62809) awarded to A.J.R and a Swedish Research Council (VR) Grant awarded to J.O.A.

Literature Cited


Martin Embley, Associate Editor

Accepted September 1, 2004