Geobiology (2003), **1**, 15–26

Geobiological analysis using whole genome-based tree building applied to the Bacteria, Archaea, and Eukarya

CHRISTOPHER H. HOUSE,[1] BRUCE RUNNEGAR[2] AND SOREL T. FITZ-GIBBON[3]

[1]*Penn State Astrobiology Research Center and Department of Geosciences, Pennsylvania State University, 212 Deike Building, University Park, PA 16802, USA*
[2]*Institute of Geophysics and Planetary Physics and NASA Astrobiology Institute, University of California, Los Angeles, CA 90095–1567, USA*
[3]*3845 Slichter Hall, IGPP Center for Astrobiology, University of California, Los Angeles, CA 90095–1567, USA*

ABSTRACT

We constructed genomic trees based on the presence and absence of families of protein-encoding genes observed in 55 prokaryotic and five eukaryotic genomes. There are features of the genomic trees that are not congruent with typical rRNA phylogenetic trees. In the bacteria, for example, *Deinococcus radiodurans* associates with the Gram-positive bacteria, a result that is also seen in some other phylogenetic studies using whole genome data. In the Archaea, the methanogens plus Archaeoglobus form a united clade and the Euryarchaeota are divided with the two Thermoplasma genomes and *Halobacterium* sp. falling below the Crenarchaeota. While the former appears to be an accurate representation of methanogen-relatedness, the misplacement of *Halobacterium* may be an artefact of parsimony. These results imply the last common ancestor of the Archaea was not a methanogen, leaving sulphur reduction as the most geochemically plausible metabolism for the base of the archaeal crown group. It also suggests that methanogens were not a component of the Earth's earliest biosphere and that their origin occurred sometime during the Archean. In the Eukarya, the parsimony analysis of five Eukaryotes using the Crenarchaeota as an outgroup seems to counter the Ecdysozoa hypothesis, placing *Caenorhabditis elegans* (Nematoda) below the common ancestor of *Drosophila melanogaster* (Arthropoda) and *Homo sapiens* (Chordata) even when efforts are made to counter the possible effects of a faster rate of sequence evolution for the *C. elegans* genome. Further analysis, however, suggests that the gene loss of 'animal' genes is highest in *C. elegans* and is obscuring the relationships of these organisms.

Received 07 November 2002; accepted 18 March 2003

Corresponding author: Professor Christopher H. House. E-mail: chouse@geosc.psu.edu

INTRODUCTION

By the early 1980s, several studies had shown that ribosomal RNA (rRNA) held promise for phylogenetic reconstruction (Fox *et al.*, 1980) and by the end of the decade, analysis of universally conserved nucleic acid sequences (particularly those of the small subunit rRNA gene) had become a powerful tool for microbial taxonomy, allowing identification of specific taxa on the basis of only a single gene sequence (Woese *et al.*, 1990). In spite of this success, single gene taxonomy has failed to reveal clearly the evolutionary relationships between major groups of prokaryotes, chiefly because single gene sequences lack sufficient information to resolve much of the divergence pattern of the major microbial branches. Furthermore, misalignment of sequences and differences of evolutionary rates among the various lineages can result in well-supported phylogenetic trees with the wrong topology (Marshall, 1997; Gribaldo & Philippe, 2002). Moreover, additional complexity is introduced by the horizontal transfer of genes from one taxon to another, providing a means by which each gene may tell of an independent history. In principle, moving to multi-gene and whole-genome-based systematics might alleviate these problems inherent to single gene molecular systematics.

Multi-gene systematics can take one of two forms. In some studies, a large number of conserved genes are each individually analysed. These studies have yielded incongruent phylogenetic results (e.g. Feng *et al.*, 1997; Ribeiro & Golding, 1998; Rivera *et al.*, 1998). The extent of these incongruencies has led to the speculation that there may not be a single tree that can be used to represent the history of life on Earth

(Doolittle, 1999). Alternatively, multi-gene molecular systematics can entail a single phylogenetic analysis of a combined dataset containing a large number of conserved proteins (Hansmann & Martin, 2000; Brown *et al.*, 2001; Wolf *et al.*, 2001; Brochier *et al.*, 2002; Daubin *et al.*, 2002; Matte-Tailliez *et al.*, 2002).

Evolutionary relationships and significant evolutionary events can be studied using whole genome sequences by, for example, building genomic trees using methods based on the presence and absence of genes (gene content) in each genome. Several different methods for the generation of trees using gene content have been developed (Fitz-Gibbon & House, 1999; Snel *et al.*, 1999; Tekaia *et al.*, 1999; Lin & Gerstein, 2000; Montague & Hutchison, 2000; Wolf *et al.*, 2001; Bansal & Meyer, 2002; Clarke *et al.*, 2002; Li *et al.*, 2002). Clarke *et al.* have developed a method that takes into account not only the gene content, but also blast score-derived distance measures for the individual protein pairs that are shared between the genomes (Clarke *et al.*, 2002).

For the most part, these different processes for building genome trees can be divided into two broad categories: those based on the presence and absence of suspected ortholog pairs, the 'Ortholog method' (e.g. Snel *et al.*, 1999), and those based on the presence and absence of gene families or protein folds, the 'Homolog method' (e.g. Fitz-Gibbon & House, 1999; Lin & Gerstein, 2000).

Generally, published genomic trees with relatively few genomes have been similar to rRNA trees and trees based on the analysis of multi-gene datasets. However, as more genomes are published and included in genomic trees, incongruencies with the universal rRNA tree can be identified and studied to reveal the origin of the non-agreement.

Here, prokaryotic genome trees have been constructed using the presence and absence of protein families (homologs) within each of 55 genomes providing an opportunity to find incongruencies with typical rRNA phylogenetic trees and explore what such incongruencies mean for genome tree construction, for the topology of the tree of life and for the implied history of microbes on Earth.

In addition, analysis of five eukaryotic genomes has been used to explore the placement of animal phyla with respect to each other. The placement of the Nematoda with respect to other animal phyla is contentious (Blair *et al.*, 2002). Although nematodes, which have a pseudocoelom, have traditionally been placed in a phylogenetic position basal to animals with a true coelom (Coelomata), recent analyses of 18S rRNA genes have placed the nematodes in a clade (Ecdysozoa) comprised of moulting animals, including the Arthropoda (Aguinaldo *et al.*, 1997; Peterson & Eernisse, 2001; Mallatt & Winchell, 2002). This alternative phylogeny (the Ecdysozoa hypothesis) has gained wide acceptance within developmental biology, influencing many interpretations of early animal evolution (Valentine & Collins, 2000; Carroll *et al.*, 2001; Davidson, 2001). In order to help understand animal evolution, we analysed relationships within the non-protist eukaryotes by constructing a homolog-based tree of all eukaryotes with complete (or nearly complete) genomes at the time of analysis, including the three animal phyla Nematoda, Arthropoda and Chordata.

MATERIALS AND METHODS

For this analysis, we used all of the published complete genome sequences (55 genomes) available at the time that were larger than 1.5 Mb (Table 1). Trees were constructed based upon the presence and absence of informative gene families, with gene families defined to be groups of homologs. Thus, as previously described (Fitz-Gibbon & House, 1999; House & Fitz-Gibbon, 2002), gene families were determined by single linkage clustering of all genes similar to each other above a specified similarity score cutoff. FASTA3 (Pearson, 1998) software was used to identify sequence similarities by comparing each individual gene sequence to a series of databases of gene sequences for each organism. The Smith-Waterman statistic (Smith & Waterman, 1981) calculated by FASTA3 was used to define similarity score cutoffs (SW-cut) used for clustering genes. The presence or absence of each gene family was scored for each genome to construct the data matrices. Data matrices are available at http://www.geosc.psu.edu/~chouse/geobiology1/.

Parsimony and distance analyses were performed using PAUP v.4.0b (Swofford, 2002) for a series of data matrices derived using a Smith–Waterman score cutoff of 160, and a range of cutoffs for the Eukaryotes. Also, compatibility and threshold parsimony analysis was applied using the Phylip software package (Felsenstein, 1993). Bootstrap scores and consistency indices were calculated using PAUP v.4.0b. The consistency index for all characters on a tree is the minimum possible tree length divided by the observed tree length (Farris, 1989). The decay index (also called Bremer support) is defined as the number of additional steps required to collapse the branch in question (Bremer, 1988) and was calculated using AUTODECAY v.4.0 (Eriksson & Wikstroem, 1995) and PAUP v.4.0b.

RESULTS AND DISCUSSION

Prokaryotic relationships

The genomic tree building results using maximum parsimony (MP), compatibility and threshold parsimony (TP) are shown in Fig. 1. While MP builds a tree minimizing the total number of events needed to build a tree from the data matrix, compatibility finds the tree representing the largest number of fully consistent characters, and TP samples trees that are intermediate between MP and compatibility by allowing character state changes for a particular character to each be counted until a threshold value is reached after which no more are counted for that character (Felsenstein, 1981). With high thresholds, the TP tree is identical to that of MP.

Table 1 Fifty-five prokaryotes used for genome tree building. Columns list: (1) code used in Figs 1 and 2, (2) organism name, (3) number of genes in the genome, (4) number of gene families within each genome (after single linkage clustering with a Smith–Waterman cutoff of 160), (5) number of gene families after single linkage clustering with all 55 genomes, (6) column 5/average of column 5.

Code	Organism	Genes	Within genome gene families	55 taxa group gene families	Proportion of average gene families
aa	*Aquifex aeolicus* VF5	1522	1120	434	0.4
af	*Archaeoglobus fulgidus* DSM4304	2407	1593	857	0.8
ap	*Aeropyrum pernix* K1	2694	2336	1692	1.6
at	*Agrobacterium tumetaciens* C58	5299	2779	1441	1.4
bh	*Bacillus halodurans* C-1 25	4066	2418	1332	1.3
bm	*Brucella melitensis*	3198	2141	956	0.9
bs	*Bacillus subtilis* 168	4021	2429	1324	1.3
ca	*Clostridium acetobutylicum* ATCC 824	3672	2243	1240	1.2
cc	*Caulobactercrescentus* CB15	3737	2381	1220	1.2
cg	*Corynebacterium glutamicum*	3040	2104	1045	1.0
cj	*Campylobacter jejuni* NCTC 11168	1731	1358	592	0.6
cpe	*Clostridium perfringens*	2723	1726	900	0.9
cte	*Chlorobium tepidum* TLS	2252	1752	951	0.9
dr	*Deinococcus radiodurans* R1	3117	2072	1143	1.1
ec	*Escherichia coli* K-12 Strain MG1655	4289	2547	1217	1.2
fn	*Fusobacterium nucleatum* ATCC 25586	2067	1408	687	0.7
hi	*Haemophilus influenzae* Road KW20	1717	1363	529	0.5
hp	*Helicobacter pylon* 26695	1565	1220	580	0.6
Insp	*Halobacterium* sp. NRC-1	2429	1652	947	0.9
ll	*Lactococcuslactis* IL1403	2266	1493	720	0.7
ma	*Methanosarcina acetivorans* C2A	4540	2369	1444	1.4
mj	*Methanococcus jannaschii* DSM 2661	1680	1211	652	0.6
rnk	*Methanopyrus kandleri* AV1 9	1687	1218	713	0.7
Ml	*Mesorhizobium loti* MAFF303099	7281	3569	2120	2.0
mle	*Mycobacterium leprae*	1605	1223	559	0.5
mma	*Methanosarcina mazei* Goel	3371	1898	993	0.9
mt	*Methanobacterium thermoautotrophicum*	1871	1344	715	0.7
nmm	*Neisseria meningitidis* MC58	2025	1593	783	0.7
nmz	*Neisseria meningitidis* Z2491	2065	1621	775	0.7
ns	*Nostoc* sp. PCC7120	6129	3372	2239	2.1
pab	*Pyrococcus abyssi*	1765	1174	564	0.5
pag	*Pyrobaculum aerophilurn* IM2	3060	2254	1562	1.5
pf	*Pyrococcus furiosus* DSM3638	2065	1395	692	0.7
ph	*Pyrococcus horikoshn* OT3	1975	1440	852	0.8
pm	*Pasteurella multocida* Pm70	2014	1477	577	0.5
psa	*Pseudomonas aeruginosa* PA01	5565	2840	1478	1.4
rs	*Ralstonia solanacearum*	5116	2836	1558	1.5
sam	*Staphylococcus aureus* Mu50	2748	1815	918	0.9
sco	*Streptomyces coelicolor* A3(2)	7897	3255	2170	2.1
sm	*Sinorhizobium meliloti* 1021	6205	2960	1594	1.5
sp	*Streptococcus pyogenes* M1	1696	1280	587	0.6
ss	*Sulfolobus soltataricus* P2	3249	1617	836	0.8
st	*Sulfolobus tokodaii* 7	2826	1815	1064	1.0
sty	*Salmonella typhimunum* LT2	4553	2722	1400	1.3
sy	*Synechocystis* sp. PCC 6803	3166	2027	1028	1.0
ta	*Thermoplasma acidophilurn*	1478	1110	467	0.4
tb	*Mycobacterium tuberculosis* H37Rv	3924	2115	1095	1.0
tm	*Thermotoga maritima* MS138	1849	1260	575	0.5
tt	*Thermoanaerobacter tengcongensis* MB4T	2588	1623	819	0.8
tv	*Thermoplasma volcanium* GSS1	1526	1143	520	0.5
vc	*Vibriocholerae* N16961	5565	2563	1418	1.4
xa	*Xanthomonas axonopodis* pv citri 306	4312	2682	1352	1.3
xc	*Xanthomonas campestris* ATCC 33913	4181	2517	1222	1.2
xf	*Xylella fastidiosa* 9a5c	2831	2186	1307	1.2
yp	*Yersinia pestis* CO-92 Biovar Orientalis	4083	2489	1280	1.2

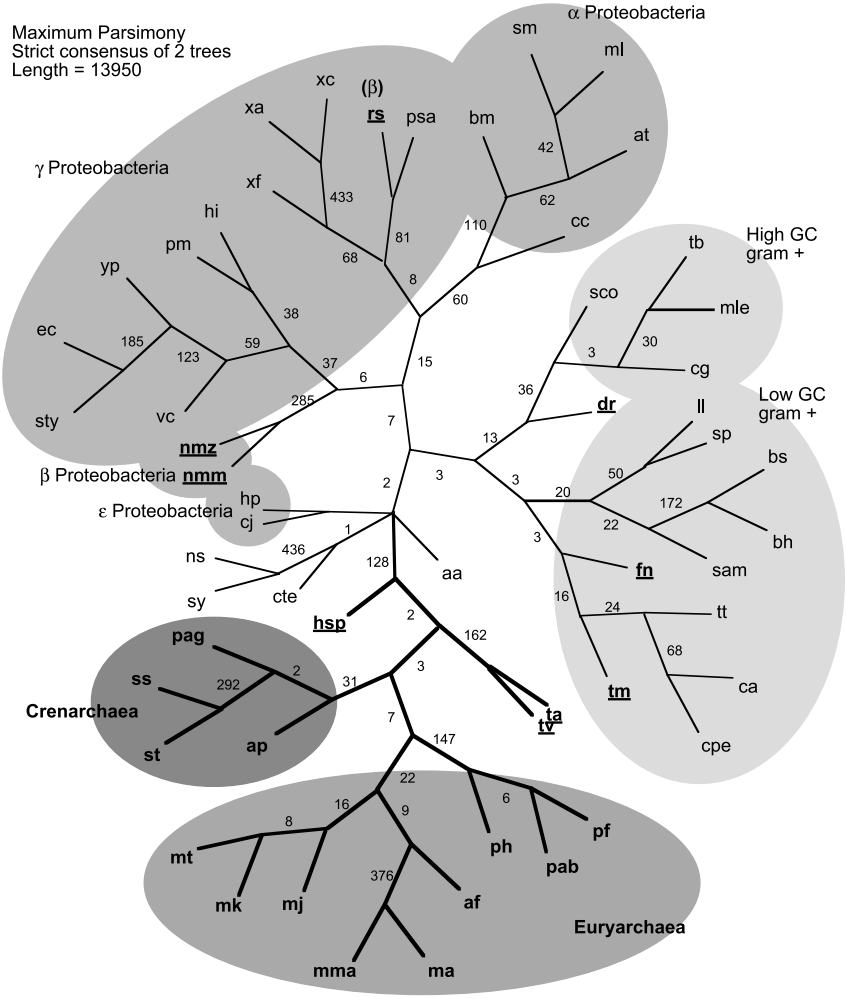

Fig. 1 Fifty-five taxa gene content trees of prokaryotes built using (a) maximum parsimony, (b) maximum compatibility and (c) threshold parsimony. Decay indices are shown for each branch of the maximum parsimony tree. See Table 1 for taxa codes.

As with other phylogenetic methods, gene content trees topologies often vary depending on the set of included taxa. In order to apply a tree-building algorithm that was not as sensitive to taxon sampling, we constructed the tree shown in Fig. 2 using a novel 'triplets' method. We first found the apparent root for each of all possible three taxon groupings, using the number of apparent synapomorphies between each pair of taxa to the exclusion of the third. By assuming that most of the signal in homolog-based genomic trees is gene family origins rather than losses, the apparent root for each group of three taxa can be determined. Our past research has suggested that most of the character state changes for a homolog-based genome tree are gene family origins (House & Fitz-Gibbon, 2002), while more complex models suggest gene loss is more significant at the ortholog level (Snel *et al.*, 2002). After the apparent root of each of the three taxon groupings was determined, a genomic tree was built by searching for the full tree that had the fewest conflicts with the 'rooted' three taxon groupings. Because of the high number of taxa used, not enough different tree topologies were searched to identify confidently the best tree with respect to bacterial relations. In contrast, the optimal archaeal topology for this analysis was robust and easily determined (Fig. 2).

The results across these various methods are not identical, but often have similar features, many of which are also seen in genomic trees built using other methods. All of our trees separate the three domains, Bacteria, Archaea and Eukarya (Fitz-Gibbon & House, 1999), and most major phylogenetic groups are reasonably well clustered (Fig. 1). For example, the proteobacteria are often in a monophyletic clade with the subdivisions (α, βγ & ε) well separated. The low GC and high GC Gram-positive clades are often neighbours, but usually have a small number of non-Gram-positive organisms among them, perhaps indicating polyphyly for the Gram-positive clade. On the whole, the relationships between major bacterial lineages remain unresolved, with no consensus emerging from the variety of whole-genome-based phylogenetic methods. There is some support for an association of *Deinococcus radiodurans* with the high GC Gram-positive clade. This is supported by several types of whole genome studies: gene content (Wolf *et al.*, 2001), concatenation of orthologous proteins (Brown *et al.*, 2001; Brochier *et al.*, 2002), combining of multiple

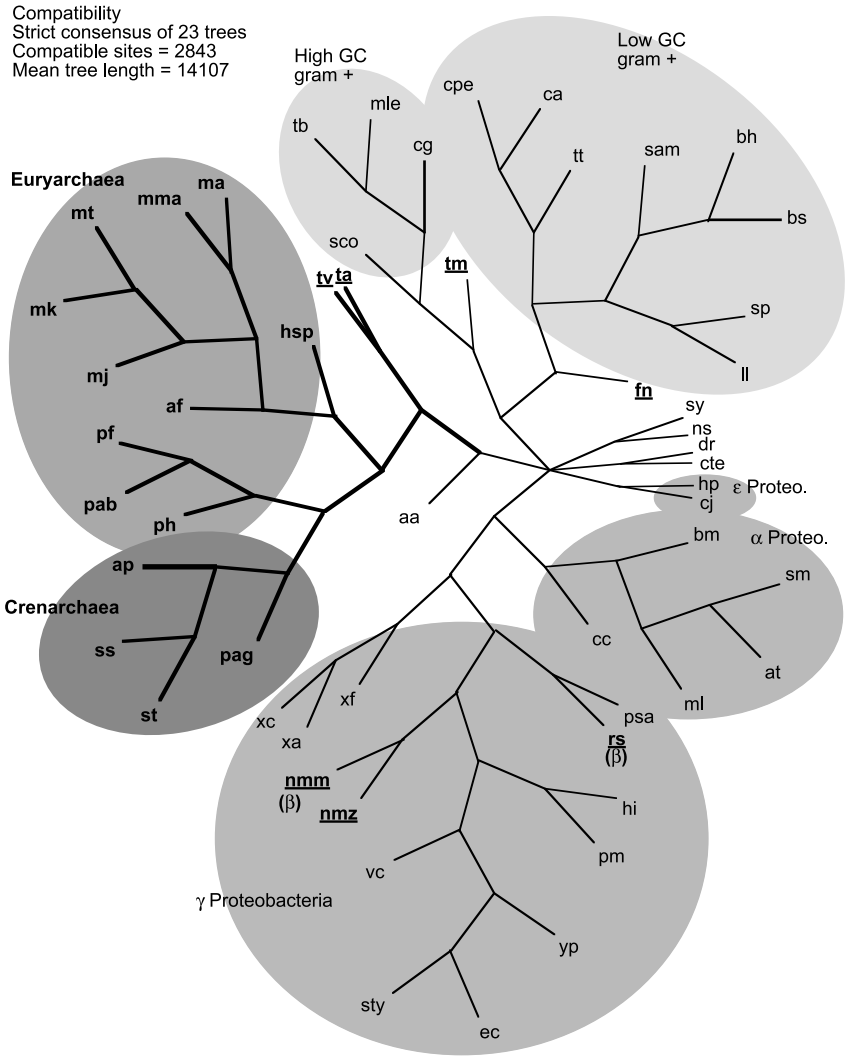

Fig. 1 *Continued*

single gene trees (supertrees) (Daubin *et al.*, 2002) and concatenation of rRNA genes (Brochier *et al.*, 2002). Repeating the analysis on the bacterial taxa only, without the archaeal outgroup, resulted in a maximum parsimony tree with exactly the same branching pattern with the minor exception of *Aquifex aeolicus* and the two epsilon Proteobacteria moving to the base of the cyanobacteria/Chlorobium clade. However, this change is very weakly supported as are most of the basal branches between major taxonomic groups in all trees.

Within the Archaea, there are several features of note in our results. First, for all of the algorithms used (MP, Compatibility, TP, and the novel 'triplets' method (Figs 1 and 2)), there is a clade containing all of the methanogens, plus *Archaeoglobus fulgidus*, an archaeal sulphate-reducer that has similar biochemistry to methanogens (Klenk *et al.*, 1997) and is micromethanogenic (Stetter *et al.*, 1987; Stetter, 1988). Archaeoglobales are probably derived from a methanogen that acquired sulphate reduction genes via lateral gene transfer (Klein *et al.*, 2001; Stahl *et al.*, 2002). Monophyly of the Methanoarchaea (including the Archaeoglobales) is not usually seen in rRNA trees. The most striking difference is the position of *Methanopyrus kandleri*, which is positioned around the base of the Archaea in rRNA trees (Burggraf *et al.*, 1991). All kinds of gene content trees consistently place *Methanopyrus* well within the cluster of other methanogens (Figs 1 and 2; Slesarev *et al.*, 2002). The robustness of this phylogenetic placement across different studies using different genome tree building methods suggests that *Methanopyrus*'s placement on the rRNA tree of life is incorrect and that in this case genomic trees are revealing a more correct tree of life topology. We cannot rule out the possibility that the Methanoarchaea cluster together in gene content trees only due to their shared (and perhaps laterally transferred) genes involved in the methanogenic life style. However, the derived position of *Methanopyrus kandleri* within the methanoarchaea argues against this. Furthermore, long branch artefacts may be effecting the placement of *Methanopyrus* in rRNA trees as these effects are expected to move taxa to the base of major clades such

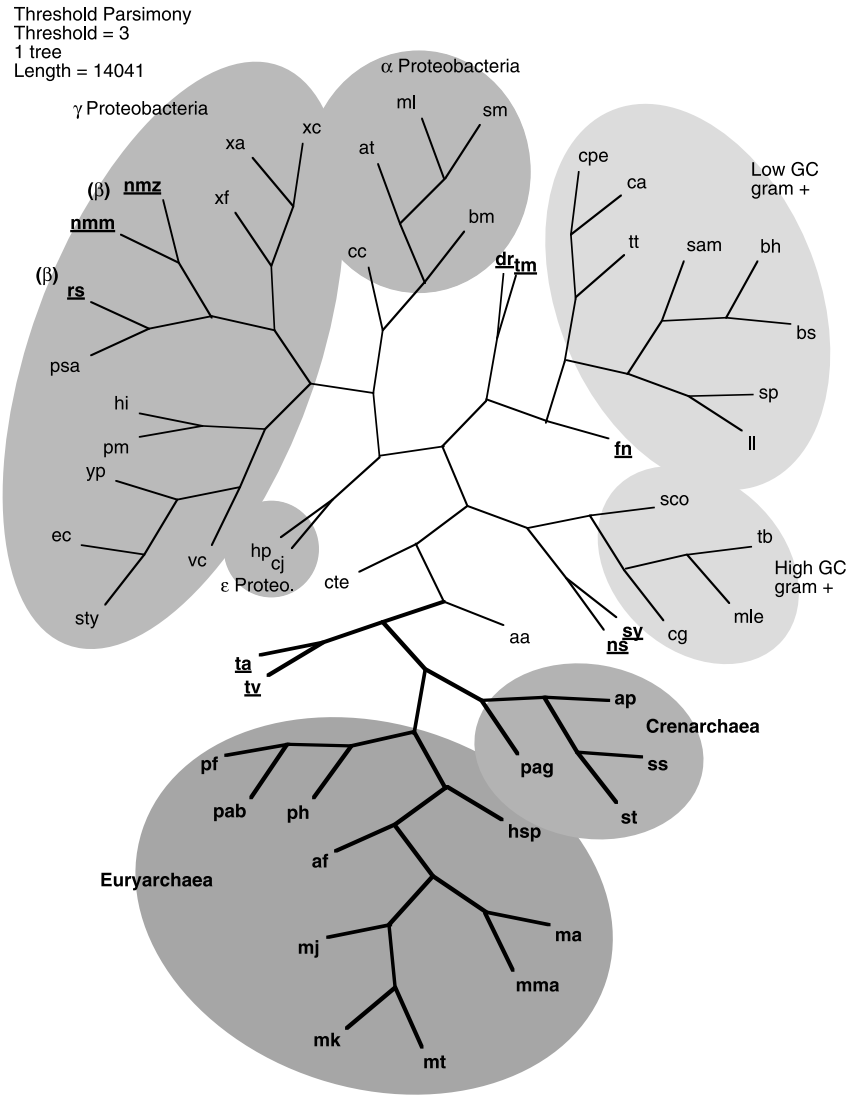

Fig. 1 *Continued*

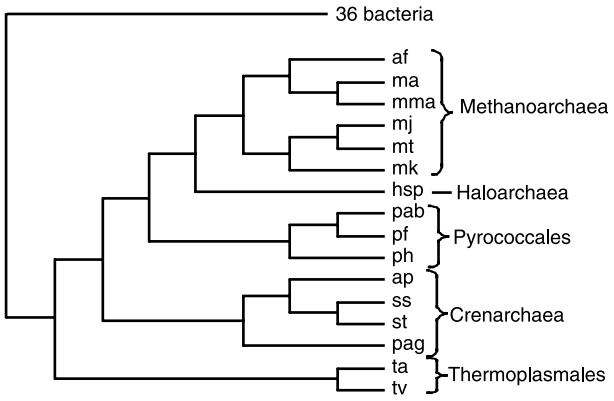

Fig. 2 Tree of Archaea (rooted using Bacteria) and built by minimizing the number of inconsistencies with all possible three taxa groupings when the number of apparent synapomorphies between each pair of two taxa to the exclusion of the third is used to define topology.

as the base of the Euryarchaeota where *Methanopyrus* is often placed. In contrast, our genomic tree places *Methanopyrus* in a derived clade, a result unlikely to be caused by long-branch artefacts.

Another notable archaeal result is the division of the Euryarchaeota, with the two *Thermoplasma* genomes and *Halobacterium* sp. falling below the Crenarchaeota. These basal positions for Halobacteria and Thermoplasma are another common feature of gene content trees (Wolf *et al.*, 2001, 2002; Clarke *et al.*, 2002) and are sometimes seen in concatenated ortholog trees (Brown *et al.*, 2001; Wolf *et al.*, 2001). However, in the case of *Halobacterium* sp., the basal position is not robust to alternate tree-building algorithms. Threshold parsimony (Fig. 1c) and the 'triplets' method (Fig. 2) both place *Halobacterium* sp. within the Euryarchaeota in a position analogous to its rRNA tree position. Compatibility (Fig. 1b) also places *Halobacterium* within the Euryarchaeota; however, this tree also moves the Pyrococcales to an unprecedented

position within the Crenarchaeota, for unknown reasons. Given the fragility of *Halobacterium*'s basal position, we suspect that it is artefactual and may be due to increased lateral gene transfer with mesophilic Bacteria (Zhaxybayeva & Gogarten, 2002).

The basal position of the Thermoplasma clade is resilient to our alternate tree-building methods and is more frequently found in concatenated ortholog trees than is the basal position for *Halobacterium* (Brown *et al.*, 2001; Wolf *et al.*, 2001). Because of the small size of the Thermoplasma genomes (<1.6 Mb), we cannot confidently confirm or refute our placement of it at the base of Archaea, as very small genomes have a tendency to be attracted to the root (House & Fitz-Gibbon, 2002).

The results shown in Fig. 1 represent unrooted topologies. In general, it is difficult to form a Tree of Life that is rooted. The most common root used is between the Archaea and Bacteria with stem Eukaryotes as a sister group to the Archaea based on paralogous gene duplications prior to the last common ancestor (Gogarten *et al.*, 1989; Iwabe *et al.*, 1989). This rooting is, however, controversial because much of the signal may be long-branch artefacts and because it is not found for all paralogous gene duplications (Gribaldo & Philippe, 2002). Given the results shown in Fig. 2, if one assumes an ancient origin of Archaea, and that the root of the tree of life is between the Archaea and the Bacteria, important geobiological implications are clear. Because of the scarcity of oxidized inorganic substrates prior to the evolution of oxygenic photosynthesis, the two most plausible microbial metabolisms that could have been present in the last common ancestor of the Archaea are methanogenesis based on CO_2 and H_2 and sulphur reduction using H_2.

Our results suggest that the last common ancestor of Archaea was not a methanogen and that methanogenesis arose later during subsequent microbial evolution. This leaves sulphur reduction as the most geochemically plausible metabolism for the base of the archaeal crown group (Fig. 3). Sulphur-reduction is common in the Archaea where it is a widespread chemolithotrophic metabolism in lineages of the Crenarchaeota, as well as being present in the heterotrophic euryarchaeal Pyrococcales. Furthermore, the phylogenetically uncertain *Thermoplasma* are also capable of sulphur-reduction. While the small genome-size makes it hard for us to confirm or refute the placement of the *Thermoplasma* at the base of the Archaea, their capacity to perform anaerobic sulphur-reduction (Segerer *et al.*, 1988) and their noted lack of a cell wall and thus similarity to Eukaryotic cells (Searcy & Hixon, 1991; Margulis, 1993) are consistent with such a phylogenetic position. The widespread and basal positions of sulphur reducers support the early origin of sulphur reduction; however, it is also possible that sulphur reduction genes were spread among these taxa at a later date by horizontal gene transfer (Gogarten *et al.*, 2002).

In any case, sulphur-reduction remains the most plausible metabolism for the base of the archaeal crown group. Therefore, we suggest, based on this study, that attempts to understand the microbial biosphere during the Archean consider the possibility that methanogens were not present from the beginning, but rather have a distinct origin sometime during that geological eon. It is even possible that their origin coincides with the advent of extremely ^{13}C-depleted kerogen at around 2.7 Ga. The fact that methanogens are not present during the entire Archean eon is most relevant to models that require methanogenesis as a mechanism for hydrogen-escape from the atmosphere (Catling *et al.*, 2001). If methanogens were not present during some portion of the Archean and the mantle was buffered at the present fayalite, magnetite, quartz (FMQ) redox state, then hydrogen loss rates would be lower than has been suggested by Catling *et al.* (2001).

In contrast, models that invoke a change in mantle redox early in Earth history (Kasting *et al.*, 1993; Kump *et al.*, 2001) are less affected by a lack of biogenic methanogenesis in the early Archean because a more reduced mantle will result in higher hydrogen escape with or without global biogenic methanogenesis. Other implications for the early Archean include the possible elimination of biogenic methane as an agent of greenhouse warming, as invoked by Pavlov *et al.* (2000), and inefficient early Archean carbon remineralization due to the lack of methanogenesis coupled with the lack of sulphate reduction (Habicht *et al.*, 2002).

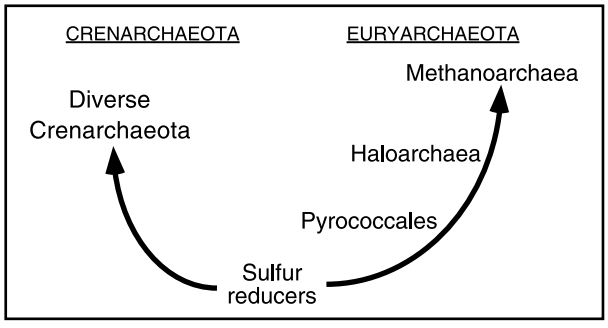

Fig. 3 Geobiological interpretations of the history of the Archaea and methanogenesis based on this genomic tree building study.

Metazoan relationships

An important problem remaining for geobiology is understanding the Cambrian explosion of multicellular life, especially the radiation of the bilaterian animals (Bilateria). As a small first step towards a full understanding of the order of appearance of the various animal phyla, we have focused on trying to help resolve the branching order of the two or three major groups of bilaterian animals.

Bilaterian animals had traditionally been separated to major groups, formalized as 'Coelomata' and 'Acoelomata' based on the assumed presence of or absence of a true cavity (coelom) within the body wall. This concept was overturned recently

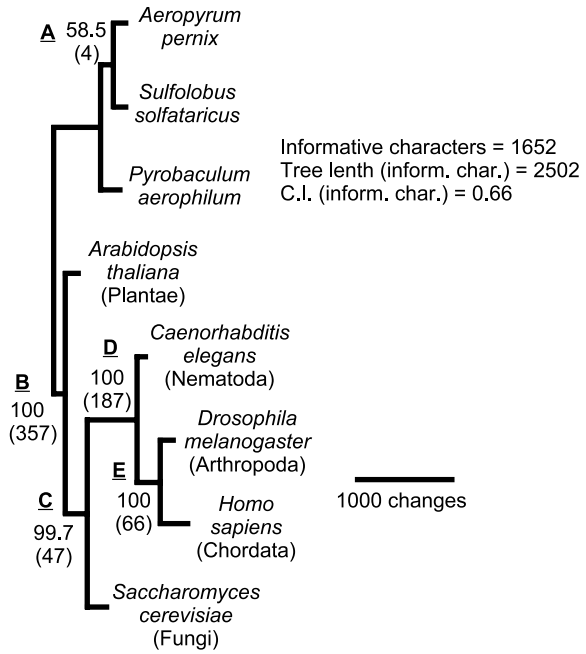

Fig. 4 Results of maximum parsimony analysis of five eukaryotes, with three crenarchaeotes as an outgroup, using the presence or absence of informative families of protein-encoding genes in each genome as characters. Bootstrap values obtained from 1000 replicates are listed at each node, followed by Decay Indices in parentheses. Each node has been assigned a letter for cross-reference to Fig. 5.

(Aguinaldo *et al.*, 1997) with the proposal, based on a carefully researched 18S rRNA tree plus morphological arguments, that the living Bilateria belong to one of two major clades which they named the Ecdysozoa (molting animals) and the Lophotrochozoa (animals having either a lophophore for feeding or a trochophore larva). This proposal has been enthusiastically accepted by many workers, but some molecular datasets also support the alternative (Coelomata) hypothesis (Blair *et al.*, 2002). The fossil record is not able at present to resolve this issue and knowing the true tree is important for any attempt to use molecular sequence data to date the Precambrian divergences of the principal animal phyla.

Maximum Parsimony analysis was used to construct the tree shown in Fig. 4 from the data matrix. The tree shown uses the 1652 informative characters (out of 16 454 total characters, 75 of which are universal to all taxa). The remaining 14 727 uninformative characters represent gene families found in only one genome. In general, there appears to be more gene content signal in eukaryotic genomes than in Prokaryotes due to larger genome size and probably a smaller fraction of transferred genes.

As a first test of the reliability of our whole genome method to deal with the much more complex and disparate eukaryotic genomes, we note that the tree (Fig. 4) has the animal phyla clustered as a sister group to *Saccharomyces cerevisiae* (Fungi) with strong statistical support for excluding *Arabidopsis thaliana* (Plantae). This degree of support for the expected pairing of animals with the fungi is notable, given that the yeast genome is greatly reduced in size relative to more normal fungal genomes. The tree also unites the Chordata with the Arthropoda to the exclusion of the Nematoda. This arrangement of animal phyla is consistent with the classical view of animal evolution where the Coelomata, which includes both Chordata and Arthropoda, are united to the exclusion of the Nematoda, members of which lack a true coelom. Although our tree has high statistical support, incorrect trees can have high levels of consistency and phylogenetic signal due to persistent biases such as 'long branch' artefacts. It has been suggested that *C. elegans* has a fast rate of gene sequence evolution compared with other animals (Aguinaldo *et al.*, 1997). This could lead to an overall decrease in the number of linkages formed by *C. elegans* genes causing *C. elegans* to fall towards the root of the tree. This could change the topology from one supporting the Ecdysozoa hypothesis to the one observed here. In order to test for this possibility, we reran our analysis accommodating the possible higher rate of sequence evolution for *C. elegans* genes. Pairwise comparisons involving a *C. elegans* gene were allowed to match at a lower similarity score cutoff than comparisons not involving a *C. elegans* gene. Figure 5 shows the support for each node of the tree with increasing differences (ΔSW) between the general score cutoff (SW-cut) and the lowered *C. elegans* score cutoff. We routinely use a low (i.e. inclusive) cutoff (SW-cut = 160) for gene family clustering, as was done to produce the tree in Fig. 4, in order to minimize the effect of variable rates of sequence divergence. The left-hand side of Fig. 5 shows the results for keeping SW-cut = 160, and allowing even lower cutoffs for *C. elegans* genes. Only a few steps are possible before the number of spurious matches allowed by the lowered cutoff overwhelm the analysis by causing high numbers of unrelated gene families (characters) to collapse together. At ΔSW = 50, only 628 of the original 1652 informative characters remain. The only topology change (Node support < 0) is within the Crenarchaeota for a node that was initially only weakly supported (see Node A, Fig. 5). At the last step (ΔSW = 50) the support for *A. thaliana* basal to the small *S. cerevisiae* genome is near zero. This is consistent with the expectation that increased numbers of spurious matches will have the largest effect on the position of the smallest genome. The support for the Coelomata topology remains high even when *C. elegans* genes are allowed to cluster at scores that are 50 below the normal cut off of 160 (ΔSW = 50).

In order to allow for larger ΔSW values, we repeated the analysis with a higher general cutoff (SW-cut) of 300. As expected, consistently high support is seen for the nodes separating Eukarya from the crenarchaeal outgroup (B) and uniting the three animals (D). The support for the Coelomata clade (E) is seen to increase with the higher SW-cut and, as expected, to decrease with increasing ΔSW. Again, the only change in topology, with respect to Fig. 4, is within the Crenarchaeota (A) and in this case is directly due to the extremely

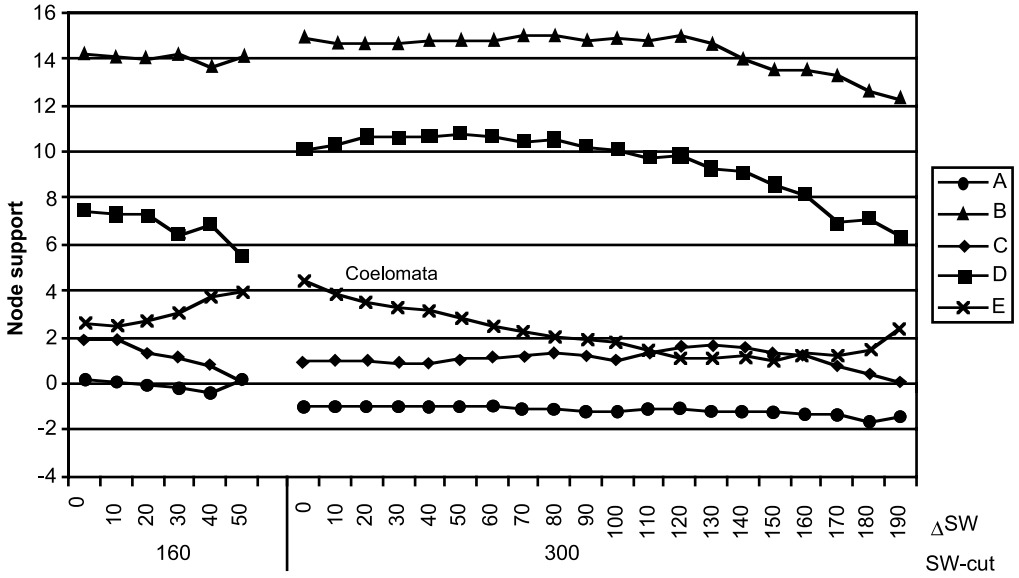

Fig. 5 Node support for Smith–Waterman score cutoffs (SW-cuts) of 160 and 300, and variable lower SW-cuts for *C. elegans* genes (SW-cut lowered by ΔSW). Node support here (in percentage) is the decay index divided by the tree length for informative characters multiplied by 100. Node support is given for each tree node (A, B, C, . . .) as labelled in Fig. 1. Negative node support here indicates that the topology has changed from that seen in Fig. 4, and is based upon the negative decay index being equal to the number of extra steps required to return to the Fig. 4 topology.

high general cutoff of 300, since the change is even observed when ΔSW = 0. The most parsimonious tree continues to show the Coelomata topology even when the *C. elegans* cutoff is 190 lower than the general cutoff, with levels of support similar to the support for uniting *S. cerevisiae* with the animals (C). We attempted to force the Ecdysozoa topology more directly by repeating the analyses, this time allowing lower cutoffs only for pairwise matches between *C. elegans* and *D. melanogaster*. Even with this direct biasing of the analysis in favour of uniting the Ecdysozoans, the Ecdysozoa topology (data not shown) was only seen at the extremes. With an SW-cut of 160, the Ecdysozoa topology was not seen even at the lowest *C. elegans* cutoff (ΔSW = 60). At an SW-cut of 300, the Ecdysozoa topology appeared at ΔSW = 120. By ΔSW = 130, however, the support for *A. thaliana* basal to *S. cerevisiae* was diminished to zero. Considering that this analysis directly generates support for the Ecdysozoa topology with every new match allowed (even spurious ones), the high level of manipulation required to alter the topology argues that the topology of the parsimony tree is not the result of differential rates of sequence evolution.

The support for a particular node on a tree can also be estimated with the decay index (also called Bremer support). Although the decay index is derived from a wider range of characters, most of its power comes from the number of characters shared exclusively by *D. melanogaster* and *H. sapiens* (160) and *D. melanogaster* and *C. elegans* (97). For reference, 34 characters are shared exclusively by *H. sapiens* and *C. elegans*. The overall tree lengths for the three different possible topologies of animal phyla given these taxa are: Coelomata, 2502; Ecdysozoa, 2568; and the third topology, 2645. Clearly, the highest support for the topology of these animals supports the Coelomata and conflicts with the Ecdysozoa, but it should be noted that the Coelomata and Ecdysozoa topologies get far more support than does the third hypothesis. This is somewhat unexpected because if the Coelomata hypothesis is correct, then one would expect the support for each of the 'false' hypotheses to be about equal. This suggests that the support for either the Coelomata or the Ecdysozoa hypothesis is inflated by homoplasy. Homoplasy can arise in homolog gene content trees through either gene family loss or lateral gene family transfer. While lateral gene transfer is more pervasive between microbial genomes than for animals, gene loss may be particularly important in this case because the Nematoda (including *C. elegans*) may have lost gene families during an evolutionary path from a complex animal form toward a seemingly more 'primative' morphology. For this reason, we investigated the loss of 'animal' genes from these three taxa in order to elucidate which hypothesis (Coelomata or Ecdysozoa) is getting inflated support.

For this new analysis, we identified gene families present in either the plant or the fungi genomes and at least one of the three animal genomes. From this list, gene families suspected to have been lost in each of the three animal genomes were identified. Table 2 shows the number of gene families identified as lost in each of the animal lineages based on its presence in other Eukaryotes. Also shown is the calculated relative gene family loss for each of these lineages based on the observed

Table 2 Gene families lost in each animal, but present in fungi (*S. cerevisiae*) and/or plant (*A. thaliana*) and at least one animal. Also shown, the relative gene family loss calculated from this data normalized to loss from *C. elegans* (Ce). On the right, the numbers of gene families that uniquely support each hypothesis are shown. Adjusted values of the number of gene families that uniquely support each hypothesis are shown in bold corrected using the calculated relative gene family loss for each animal. These values assume that the 34 gene families that support neither the Coelomata or Ecdysozoa hypotheses are the result of genes having been lost from *D. melanogaster* (Dm). *H. sapiens* has been abbreviated Hsa.

Taxa	#G.F. losses	Relative G.F. loss	Hypothesis	#G.F. synapomorphies	Adjust. G.F. synap.
Ce	169	1	Celomata	160	**92**
Hsa	121	0.72	Ecdysozoa	97	**49**
Dm	85	0.50	Neither	34	**0**

Table 3 Expected relative rates of paired gene family loss due to two independent losses calculated by multiplication of the pair of rates from Table 1 followed by normalizing the result to the calculated rate of paired loss from Ce-Hsa. On the right are gene families absent from a pair of animals, but present in fungi (*S. cerevisiae*) and/or plant (*A. thaliana*) and the other animal, followed by the same data corrected to remove cases in which paired gene family loss is caused by two independent losses rather than synapomorphy, using the relative rates in Table 3 and assuming that the gene families absent from Ce-Hsa are the result of paired gene family loss due to two independent losses.

Taxa pair	Expected relative G.F. loss	Hypothesis	#Double losses	Adjust. double losses
Dm-Hsa	0.50	Celomata	15	**–1**
Ce-Dm	0.70	Ecclysozoa	42	**20**
Ce-Hsa	1	Neither	31	**0**

losses normalized to loss from *C. elegans*. On the right-hand side of Table 2, the apparent synapomorphies for each hypothesis is shown along with the same data adjusted by the relative gene family loss for each lineage assuming that all 34 gene families that support the third hypothesis (neither Coelomata or Ecdysozoa) are the result of gene loss from *D. melanogaster*. Because the observed gene loss is highest in *C. elegans*, the number of apparent synapomorphies for the Coelomata hypothesis falls from 160 to 92 with support for the Ecdysozoa failing from 97 to 49. After this adjustment for gene loss from these lineages, the results still favour Coelomata over Ecdysozoa, but by a narrower margin (92–49). The results are still troubling, however, because the analysis was not able fully to eliminate support for the second hypothesis leaving residual support for both Coelomata and Ecdysozoa.

Because of this residual support, we decided to investigate which hypothesis is favoured by the gene losses themselves as characters for phylogenetic analysis. In this case, the rooted topology of any three taxa dictates that the paired loss of gene families that were present in the ancestor of all three taxa and are now missing from two of the three taxa will be more common for the two taxa more closely related. Therefore, we first calculated the relative paired gene family loss that would be expected given the relative single gene family loss found in Table 2. These expected relative paired gene family losses (shown in Table 3) are highest in *C. elegans* – *H. sapiens* because each of these taxa has a higher level of single gene loss than does *D. melanogaster*. Next, we counted the observed number of double gene family losses for each taxa-pair, arranged in Table 3 by the hypothesis that they support, for gene families expected to be present in the last common ancestor of animals based on their presence in either fungi or plant. Finally, we adjusted the observed results using the expected relative double gene family loss and assuming that all of the paired gene family loss supporting the uniting of *C. elegans* and *H. sapiens* is due to two independent events of single gene family loss. These adjusted results are shown in bold in Table 3, and, in this case, they indicate fair support (20 gene families) for the Ecdysozoa hypothesis and no support for the Coelomata hypothesis.

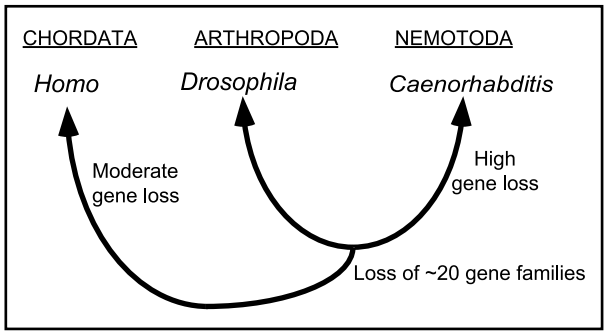

Fig. 6 Interpretation of the history of animal phyla based on this genomic tree building study.

Our results from the Eukarya demonstrate that care must be taken when interpreting gene content results in order to test particular phylogenetic hypotheses. We found that: (1) a robust parsimony tree could be constructed from the presence and absence of gene families seemingly contradicting the Ecdysozoa hypothesis, (2) the topology of the parsimony tree was not the result of differential rates of sequence evolution as our tree is not very sensitive to pairwise alignment scores because extremely large artificial variations are required to force a change in the topology, and (3) the topology of the parsimony tree does seem to be influenced by a high number of gene family losses in *C. elegans* since the divergence of animal phyla. Taken together, we believe that these results support, albeit weakly, the Ecdysozoa hypothesis over the Coelomata hypothesis with events of gene family loss proceeding as shown in Fig. 6. How this weak support for the Ecdysozoa relates to geobiological debates regarding the Vendian–Cambrian explosion is unclear, but it does demonstrate that fairly extensive genetic changes are occurring in animal genomes during this diversification, and it suggests that the seemingly 'primitive' Nematoda body plan has evolved from a more complex animal in a process including the loss of gene families.

CONCLUSION

The full extent to which gene family content reflects true organismal lineages has yet to be determined and will require careful analysis of many more genomes as they become available. Initial analyses are promising (Fitz-Gibbon & House, 1999; Snel *et al.*, 1999; Wolf *et al.*, 2001; House & Fitz-Gibbon, 2002), but they suggest that care must be taken when analysing organisms with the potential for substantial genome changes such as the massive gene loss associated with adaptation to a non-free-living life style (Douglas *et al.*, 2001). Several important conclusions can be drawn from this study using the presence and absence of gene families to investigate the Bacteria, Archaea and Eukarya. Principally, that incongruencies between genomic trees and those of rRNA can have diverse causes leading to situations in which the rRNA topology is correct and the genomic tree is false, as well as situations in which the genomic tree is correct and the rRNA tree is false. In particular, evidence suggests that the Methanoarchaea are correctly united in genomic trees while the placement of *Halobacterium* on genomic trees is problematic. This result is important as it indicates that models of the Earth's early biosphere must consider that methanogens may not have been present until sometime during the mid to late Archean. Our investigation of the relations of animal phyla found that parsimony is misleading for these taxa because gene loss of Eukaryotic genes is highest in *Caenorhabditis elegans* and appears to be obscuring the relationships of these organisms.

ACKNOWLEDGMENTS

This research was supported by NASA Astrobiology Institute (NAI) grants to the Penn State Astrobiology Research Center and to the UCLA Center for Astrobiology. IGPP Publication no. 5780

REFERENCES

Aguinaldo AMA, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387**, 489–493.

Bansal AK, Meyer TE (2002) Evolutionary analysis by whole-genome comparisons. *Journal of Bacteriology* **184**, 2260–2272.

Blair JE, Ikeo K, Gojobori T, Hedges B (2002) The Evolutionary Position of Nematodes. *BMC Evolutionary Biology* **2**, 7.

Bremer K (1988) The limits of amino acid sequence data in angiosperm phlyogenetic reconstruction. *Evolution* **42**, 795–803.

Brochier C, Bapteste E, Moreira D, Philippe H (2002) Eubacterial phylogeny based on translational apparatus proteins. *Trends in Genetics* **18**, 1–5.

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ (2001) Universal trees based on large combined protein sequence data sets. *Nature Genetics* **28**, 281–285.

Burggraf S, Stetter KO, Rouviere P, Woese CR (1991) Methanopyrus kandleri: an archaeal methanogen unrelated to all other known methanogens. *Systematic Applied Microbiology* **14**, 346–351.

Carroll SB, Grenier JK, Weatherbee SD (2001) *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design*. Blackwell Science, Oxford.

Catling DC, Zahnle KJ, McKay C (2001) Biogenic methane, hydrogen escape, and the irreversible oxidation of early Earth. *Science* **293**, 839–843.

Clarke GD, Beiko RG, Ragan MA, Charlebois RL (2002) Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *Journal of Bacteriology* **184**, 2072–2080.

Daubin V, Gouy M, Perriere G (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Research* **12**, 1080–1090.

Davidson EH (2001) *Genomic Regulatory Systems: Development and Evolution*. Academic Press, San Diego.

Doolittle WF (1999) Phylogenetic classification and the universal tree. *Science* **284**, 2124–2129.

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu XN, Reith M, Cavalier-Smith T, Maier UG (2001) The highly reduced genome of an enslaved algal nucleus. *Nature* **410**, 1091–1096.

Eriksson T, Wikstroem N (1995) *Autodecay*, Version 3.0. Computer program available from. *http://www.zoo.toronto.edu/~mes/pub/Autodecay_3.0.3.sea.hqx*. v.3.0.

Farris JS (1989) The retention index and the rescaled consistency index. *Cladistics* **5**, 417–419.

Felsenstein J (1981) A likelihood approach to character weighting and what it tells us about parsimony and compatiblity. *Biological Journal of the Linnean Society of London* **16**, 183–106.

Felsenstein J (1993) *Phylip (Phylogeny Inference Package)*, v.3.6a3, Distributed by the Author. Department of Genetics, University of Washington, Seattle.

Feng DF, Cho G, Doolittle RF (1997) Determining divergence times with a protein clock: update and reevaluation. *Proceedings of the National Academy of Sciences of the USA* **94**, 13028–13033.

Fitz-Gibbon ST, House CH (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Research* **27**, 4218–4222.

Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LJ, Zablen LB, Blakemore R, Gupta R, Bonen L, Lewis BJ, Stahl DA, Luehrsen KR, Chen KN, Woese CR (1980) The phylogeny of prokaryotes. *Science* **209**, 457–463.

Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution* **19**, 2226–2238.

Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, *et al.* (1989) Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proceedings of the National Academy of Sciences of the USA* **86**, 6661–6665.

Gribaldo S, Philippe H (2002) Ancient phylogenetic relationships. *Theoretical Population Biology* **61**, 391–408.

Habicht KS, Gade M, Thamdrup B, Berg P, Canfield DE (2002) Calibration of sulfate levels in the archean ocean. *Science* **298**, 2372–2374.

Hansmann S, Martin W (2000) Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. *International Journal of Systematic and Evolutionary Microbiology* **50**, 1655–1663.

House CH, Fitz-Gibbon ST (2002) Using homolog groups to create a whole-genomic tree of free-living organisms: An update. *Journal of Molecular Evolution* **54**, 539–547.

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proceedings of the National Academy of Sciences of the USA* **86**, 9355–9359.

Kasting JF, Eggler DH, Raeburn SP (1993) Mantle redox evolution and the oxidation state of the Archean atmosphere. *Journal of Geology* **101**, 245–257.

Klein M, Friedrich M, Roger AJ, Hugenholtz P, Fishbain S, Abicht H, Blackall LL, Stahl DA, Wagner M (2001) Multiple lateral transfers of dissimilatory sulfite reductase genes between major lineages of sulfate-reducing prokaryotes. *Journal of Bacteriology* **183**, 6028–6035.

Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA, McKenney K, Adams MD, Loftus B, Venter JC *et al.* (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. *Nature* **390**, 364–370.

Kump LR, Kasting JF, Barley ME (2001) Rise of atmospheric oxygen and the 'upside-down' Archean mantle. *Geochemistry, Geophysics, and Geosystems* **2**, art. no. 2000GC000114.

Li W, Fang W, Ling L, Wang J, Xuan Z, Chen R (2002) Phylogeny based on whole genome as inferred from complete information set analysis. *Journal of Biology Physics* **28**, 439–447.

Lin J, Gerstein M (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Research* **10**, 808–818.

Mallatt J, Winchell CJ (2002) Testing the new animal phylogeny: First use of combined large-subunit and small-subunit rRNA gene sequences to classify the protostomes. *Molecular Biology and Evolution* **19**, 289–301.

Margulis L (1993) *Symbiosis in Cell Evolution: Microbial Communities in the Archean and Proterozoic Eons*. Freeman, New York.

Marshall CR (1997) Statistical and computational problems in reconstructing evolutionary histories from DNA data. *Computing Science and Statistics* **29**, 218–226.

Matte-Tailliez O, Brochier C, Forterre P, Philippe H (2002) Archaeal phylogeny based on ribosomal proteins. *Molecular Biology and Evolution* **19**, 631–639.

Montague MG, Hutchison CA (2000) Gene content phylogeny of herpesviruses. *Proceedings of the National Academy of Sciences of the USA*. **97**, 5334–5339.

Pavlov AA, Kasting JF, Brown LL, Rages KA, Freedman R (2000) Greenhouse warming by CH4 in the atmosphere of early Earth. *Journal of Geophysics Research* **105**, 11981–11990.

Pearson WR (1998) Empirical statistical estimates for sequence similarity searches. *Journal of Molecular Biology* **276**, 71–84.

Peterson KJ, Eernisse DJ (2001) Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. *Evolution and Development* **3**, 170–205.

Ribeiro S, Golding GB (1998) The mosaic nature of the eukaryotic nucleus. *Molecular Biology and Evolution* **15**, 779–788.

Rivera MC, Jain R, Moore JE, Lake JA (1998) Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences of the USA* **95**, 6239–6244.

Searcy DG, Hixon WG (1991) Cytoskeletal origins in sulfur-metabolizing archaebacteria. *Biosystems* **25**, 1–11.

Segerer A, Langworthy TA, Stetter KO (1988) Thermoplasma acidophilum and Thermoplasma volcanium sp. nov. from solfatara fields. *Systematic and Applied Microbiology* **10**, 161–171.

Slesarev AI, Mezhevaya KV, Makarova KS, Polushin NN, Shcherbinina OV, Shakhova VV, Belova GI, Aravind L, Natale DA, Rogozin IB, Tatusov RL, Wolf YI, Stetter KO, Malykh AG, Koonin EV, Kozyavkin SA (2002) The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens. *Proceedings of the National Academy of Sciences of the USA* **99**, 4644–4649.

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197.

Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. *Nature Genetics* **21**, 108–110.

Snel B, Bork P, Huynen MA (2002) Genomes in flux: The evolution of archaeal and proteobacterial gene content. *Genome Research* **12**, 17–25.

Stahl DA, Fishbain S, Klein MBJB, Wagner M (2002) Origins and diversification of sulfate-respiring microorganisms. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology* **81**, 189–195.

Stetter KO (1988) *Archaeoglobus fulgidus* Gen-Nov. Sp-Nov – A new taxon of extremely thermophilic Archaeobacteria. *Systematic and Applied Microbiology* **10**, 172–173.

Stetter KO, Lauerer G, Thomm M, Neuner A (1987) Isolation of extremely thermophilic sulfate reducers: evidence for a novel branch of Archaebacteria. *Science* **236**, 822–824.

Swofford DL (2002) *PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods)*, v.4.0b. Sinauer Associates, Sunderland, MA.

Tekaia F, Lazcano A, Dujon B (1999) The genomic tree as revealed from whole proteome comparisons. *Genome Research* **9**, 550–557.

Valentine JW, Collins AG (2000) The significance of moulting in Ecdysozoan evolution. *Evolution and Development* **2**, 152–156.

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the USA* **87**, 4576–4579.

Wolf YI, Rogozin IB, Grishin NV, Koonin EV (2002) Genome trees and the Tree of Life. *Trends in Genetics* **18**, 472–479.

Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evolutionary Biology* **1**, 8.

Zhaxybayeva O, Gogarten JP (2002) Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses. *BMC Genomics* **3**, 4.