The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods

Peter G. Foster, Cymon J. Cox and T. Martin Embley


**Supplementary data**

"Data Supplement"
http://rstb.royalsocietypublishing.org/content/suppl/2010/06/11/364.1527.2197.DC1.html

**References**

This article cites 42 articles, 25 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/364/1527/2197.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/364/1527/2197.full.html#related-urls

**Subject collections**

Articles on similar topics can be found in the following collections

- evolution (680 articles)

**Email alerting service**

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here
The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods

Peter G. Foster 1, Cymon J. Cox 2 and T. Martin Embley 3,*

1 Department of Zoology, Natural History Museum, Cromwell Road, London SW7 5BD, UK

2 Centro de Ciências do Mar, Universidade do Algarve, Campus de Gambelas, 8005-139 Faro, Portugal

3 Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle NE2 4HH, UK

The three-domains tree, which depicts eukaryotes and archaeabacteria as monophyletic sister groups, is the dominant model for early eukaryotic evolution. By contrast, the ‘ecocyte hypothesis’, where eukaryotes are proposed to have originated from within the archaeabacteria as sister to the Crenarchaeota (also called the eocytes), has been largely neglected in the literature. We have investigated support for these two competing hypotheses from molecular sequence data using methods that attempt to accommodate the across-site compositional heterogeneity and across-tree compositional and rate matrix heterogeneity that are manifest features of these data. When ribosomal RNA genes were analysed using standard methods that do not adequately model these kinds of heterogeneity, the three-domains tree was supported. However, this support was eroded or lost when composition-heterogeneous models were used, with concomitant increase in support for the ecocyte tree for eukaryotic origins. Analysis of combined amino acid sequences from 41 protein-coding genes supported the ecocyte tree, whether or not composition-heterogeneous models were used. The possible effects of substitutional saturation of our data were examined using simulation; these results suggested that saturation is delayed by among-site rate variation in the sequences, and that phylogenetic signal for ancient relationships is plausibly present in these data.

Keywords: universal tree of life; eukaryote origins; archaeabacteria; ecocyte; heterogeneous phylogenetic models

1. INTRODUCTION

Phylogenetic reconstruction of the earliest diverging lineages in the tree of life is one of the most difficult, but important, problems in evolutionary biology. At present there are two main hypotheses concerning the primary divisions in the tree of life based upon different analyses of molecular sequence data (figure 1). The ‘three-domains hypothesis’ (Woese et al. 1990) posits that eubacteria (or Bacteria), archaeabacteria (Archaea) and eukaryotes (Eukarya) form separate monophyletic groups (domains). The three-domains tree has provided support for theories of eukaryotic origins that have eukaryotes as old as archaeabacteria and derived from a common ancestor, sometimes called a neomuran, shared with that group (Cavalier-Smith 2002; Pace 2006). By contrast, the ‘ecocyte hypothesis’ (Rivera & Lake 1992) posits that essential components of eukaryotes branch from within the archaeabacteria, sharing common ancestry with a specific group of archaeabacteria called the Crenarchaeota (Woese et al. 1990) or eocytes. Current versions of both hypotheses hold that the root of the tree of life is either on the branch separating the eubacteria from the archaeabacteria and eukaryotes, in line with rooting studies using ancient paralogous genes (e.g. Baldauf et al. 1996; Zhaxybayeva et al. 2005), or it lies within the eubacteria based on the polarization of clades characters or indels (Cavalier-Smith 2006; Skopfhammer et al. 2007). For the purpose of this paper we also follow the convention of a eubacterial root, while recognizing that there is still a healthy debate about its reliability (Philippe & Forterre 1999; Zhaxybayeva et al. 2005; Lake et al. 2008, and references therein).

The main evidence for the two competing hypotheses comes from analyses of molecular sequences, often the same ones, at different times and using different methods (Lake 1988; Yang & Roberts 1995; Baldauf et al. 1996; Barns et al. 1996; Touroasse & Gouy 1999; Brown et al. 2001; Katoh et al. 2001; Harris et al. 2003). It has been suggested that the strongest support for archaeabacterial monophyly, and hence the three-domains tree, comes from the simplest methods (Touroasse & Gouy 1999; Katoh et al. 2001); the inference being that archaeabacterial monophyly might be a phylogenetic artefact of model mis-specification. With this in mind, we recently re-investigated the support for the three-domains tree and the ecocyte tree from the small number of genes, typically
2. MATERIAL AND METHODS

Two combined datasets, one rRNA and one protein, were constructed based on the data presented in Cox et al. (2008). Seven additional Crenarchaeota/eocytes that had been sequenced since our original analyses (Cox et al. 2008) were added to the datasets, namely, *Caldivigia maquiltingensis*, *Cenarchaeum symbiosum*, *Hyperthermus butylicus*, *Ignicoccus hospitalis*, *Nitrosopumilus maritimus*, *Staphylothermus marinus* and *Thermofilum pendens*, and a total of 12 taxa from the eubacteria, euryarchaeota and eukaryotes were removed to reduce the computational complexity (table 1). In total, there were 35 taxa included in the combined protein dataset and 34 in the combined rRNA dataset (*Phytophora ramorum* rRNA sequences were unavailable). The protein dataset consisted of 41 proteins, those used by Cox et al. (2008), but excluding chaperonin TCPl subunits 1 (α), 3 (γ), 4 (δ), 7 (η), which are paralogues of chaperonin-containing TCPl subunit 5 (ε). The combined protein dataset was also recoded into Dayhoff groups (Hrdy et al. 2004). Dayhoff recoding defining the following six groups of aminoacids corresponding to the PAM matrix: 1, cysteine; 2, alanine, serine, threonine, proline, glycine; 3, asparagene, aspartic acid, glutamic acid, glutamine; 4, histidine, arginine, lysine; 5, methionine, isoleucine, leucine, valine; 6, phenylalanine, tyrosine, tryptophan. Constant sites and singletons were removed from the datasets as they do not contribute to topological resolution and their composition differs from that of the variable sites in a $\chi^2$-test of significance ($p = 0$).

Maximum-parsimony (MP) and maximum-likelihood (ML) analyses were conducted in PAUP* (v. 4.10; Swofford 2002) and RAxML (v. 2.2.3; Stamatakis 2006), respectively. Bayesian Markov Chain Monte Carlo (MCMC) analyses were conducted in MrBayes (v. 3.1.2; Ronquist & Huelsenbeck 2003), P4 (v. 0.86; Foster 2004) and PHYLORBAYES (v. 2.3; Lartillot & Philippe 2004). MP bootstrap analyses of the small subunit (SSU) and large subunit (LSU) combined rRNA dataset (1045 sites) were performed with the data in a single partition using 1000 heuristic search replicates with trees-bisection reconnection (TBR) branch-swapping. Neighbour-joining (NJ) analysis using log-determinant distances (LogDet; Lake 1994; Lockhart et al. 1994) were conducted in PAUP* with 500 bootstrap replicates. ML bootstrap analyses (500 replicates) were conducted with each rRNA partition modelled separately by the general-time reversible (GTR) plus gamma-distributed among-site rate variation ($\Gamma$) model (labelled GTRGAMMA in RAxML). Tree-homogeneous MCMC analysis of the rRNA data was performed in P4 for 2 000 000 generations with a separate GTR + $\Gamma$ model for each partition, and with the polytomous prior, and a free among-partition rate parameter. Covarion model analyses were performed in MrBayes, with a GTR + $\Gamma$ applied to each rRNA partition, and the MCMC run for 2 000 000 generations. Bayesian MCMC analysis using the NDCH and the NDRH models was performed using P4. The NDCH (node-discrete composition heterogeneity) model allows different compositions on different branches, and the NDRH model allows different rate matrices on different branches. Ten replicate MCMC runs were performed for each configuration of NDCH and NDRH, for four to six million generations. A prior
probability ratio on changing composition vectors or rate matrices associated with nodes was used, as described further in the supplemental materials. CAT model analyses were performed in PHYLOBAYES with a GTR rate matrix and $G$ distribution of rates among sites. Two independent CAT model runs were conducted, each 2,000,000 cycles, to check convergence to the same posterior probability distribution.

Combined protein analyses were all performed with the entire dataset (5222 sites standard amino acid coding and 4088 sites Dayhoff-recoded data) treated as a single partition. NJ bootstrap analyses of protein LogDet distances were performed with 1000 replicates in P4. Equally weighted MP analyses were performed in PAUP with 500 bootstrap heuristic search replicates with TBR branch-swapping. ML bootstrap analyses (300 replicates) under a WAG (Whelan & Goldman 2001) rate matrix with gamma-distributed among-site rate variation (PROTGAMMAWAG) were performed in RAxML. Homogeneous MCMC analyses of the combined protein dataset were conducted in MRBAYES under a WAG $+$ $G$, with two independent runs each of 1,000,000 generations. Covarion MCMC analyses in MRBAYES were performed with two independent runs, each with two chains, under a WAG $+$ $G$ substitution model, and run for 800,000 generations. MCMC analyses in PHYLOBAYES were conducted with the CAT-Poisson model with a $G$ distribution of rates among sites. Two independent CAT-poison MCMC runs were performed, each of 3,000,000 cycles, to check for convergence to the

Table 1. Taxa and data provenance.

<table>
<thead>
<tr>
<th>taxonomy</th>
<th>taxon</th>
<th>provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>eubacteria</td>
<td>Campylobacter jejuni</td>
<td>GB: NC_003912</td>
</tr>
<tr>
<td></td>
<td>Chlamydia trachomatis</td>
<td>GB: NC_000117</td>
</tr>
<tr>
<td></td>
<td>Clostridium acetobutylicum</td>
<td>GB: NC_003030</td>
</tr>
<tr>
<td></td>
<td>Escherichia coli</td>
<td>GB: NC_00913</td>
</tr>
<tr>
<td></td>
<td>Rhodopirellula baltica</td>
<td>GB: NC_05027</td>
</tr>
<tr>
<td>archaeabacteria</td>
<td>Campylobacterales</td>
<td>GB: NC_000854</td>
</tr>
<tr>
<td></td>
<td>Chlamydiae</td>
<td>GB: NC_000117</td>
</tr>
<tr>
<td></td>
<td>Firmicutes</td>
<td>GB: NC_000854</td>
</tr>
<tr>
<td></td>
<td>Gammaproteobacteria</td>
<td>GB: NC_000117</td>
</tr>
<tr>
<td></td>
<td>Planctomycetes</td>
<td>GB: NC_000117</td>
</tr>
<tr>
<td></td>
<td>Alphaproteobacteria</td>
<td>GB: NC_000117</td>
</tr>
<tr>
<td></td>
<td>Cyanobacteria</td>
<td>GB: NC_000117</td>
</tr>
<tr>
<td></td>
<td>Spirochaetes</td>
<td>GB: NC_000117</td>
</tr>
<tr>
<td>eukaryotes</td>
<td>Viridiplantae</td>
<td>TIGR$^a$</td>
</tr>
<tr>
<td></td>
<td>Mycetozoa</td>
<td>dictyBase$^b$</td>
</tr>
<tr>
<td></td>
<td>Fungi</td>
<td>GB: nr$^d$</td>
</tr>
<tr>
<td></td>
<td>Diplomonadida</td>
<td>GiardiaDB$^c$</td>
</tr>
<tr>
<td></td>
<td>Metazoa</td>
<td>EMBL-EBI$^e$</td>
</tr>
<tr>
<td></td>
<td>Euglenozoa</td>
<td>WTSH$^f$</td>
</tr>
<tr>
<td></td>
<td>Oomycetes</td>
<td>JGI$^g$</td>
</tr>
<tr>
<td></td>
<td>Fungi</td>
<td>GB: nr$^e$</td>
</tr>
<tr>
<td></td>
<td>Bacillariophyta</td>
<td>JGI$^g$</td>
</tr>
<tr>
<td></td>
<td>Parabasalidea</td>
<td>TIGR$^g$</td>
</tr>
</tbody>
</table>

$^a$The Institute for Genomic Research, Center for the Study of Biological Complexity.

$^b$http://www.dictybase.org/.


$^e$European Molecular Biology Laboratory—European Bioinformatics Institute.

$^f$Wellcome Trust Sanger Institute.

$^g$Joint Genome Institute.
same posterior probability distribution. Maximum parsimony analyses of the Dayhoff-recoded protein dataset were performed with 1000 bootstrap replicates and equally weighted characters. The Dayhoff-recoded dataset was analysed using a homogeneous GTR + $\Gamma$ model (table 2), that is, a model with a GTR + $\Gamma$ rate matrix and 14 composition vectors, with the polytomy prior (Lewis et al. 2005). The MCMC was run for 2,000,000 generations. CAT + $\Gamma$ analyses were also conducted in PhyloBayes with the ‘dayhoff’ option, and two independent MCMC’s run for $>8$ 500 000 cycles.

### 3. RESULTS AND DISCUSSION

#### (a) Support for alternative trees of life based on rRNA genes

Historically, it has mainly been conflicting phylogenetic analyses of large and small subunit ribosomal RNA sequences (Lake 1988; Yang & Roberts 1995; Barns et al. 1996; Tourasse & Gouy 1999) that have fuelled the debate over which tree, three-domains or eocyte, is better supported. We therefore analysed concatenated SSU rRNA and LSU rRNA gene sequences using a variety of methods to investigate support for the competing hypotheses. Most analyses showed high support for monophyletic Euryarchaeota and monophyletic Crenarchaeota/eocytes (table 2); these groups are generally, but not universally, considered to be monophyletic. In accord with previous studies, MP recovered high bootstrap support for the three-domains tree (electronic supplementary material, figure S2), GTR + $\Gamma$: 82% Bayesian posterior probability (BPP) in figure 2a, 76% BPP in electronic supplementary material, figure S3).

We used the GTR + $\Gamma$ analyses (figure 2a, electronic supplementary material, figure S3) as a baseline and improved the models by accommodating covarion, tree-heterogeneous or data-heterogeneous processes. The improved fit of the models to the data was indicated by an improvement in the log marginal likelihood. Bayes factors are the ratio of the marginal likelihood of two models, and can be used to compare models: a log(BF) of 5 or more is considered to be a very strong support in favour of the better fitting model (Kass & Raftery 1995). In all cases described here the log(BF) greatly exceeded 5, suggesting that we were adding important aspects to the evolutionary model (table 2).

The dataset we analysed was heterogeneous in composition; a $\chi^2$-test for compositional homogeneity over the tree failed for both data partitions ($p \approx 0.0$ for both SSU and LSU). Compositional heterogeneity over the tree was accommodated with the NDCH model, and the model fit was assessed by posterior predictive simulation (Bollback 2002). The $X^2$ values for the original data were 434 for the SSU and 839 for the LSU (figure 3, arrows). (Here we distinguish the statistic $X^2$ sensu Sokal & Rohlf (1981) from the $\chi^2$ curve that is often used to assess its significance). $X^2$ values from simulations of posterior samples using the GTR + $\Gamma$ model (table 2, row D) were mostly less than 100 for both partitions (figure 3, black bars), showing that the tree-homogeneous GTR + $\Gamma$ model did not adequately fit the data. However, a tree-heterogeneous NDCH model using two
composition vectors for each data partition (table 2, row G; electronic supplementary material, figure S5) adequately modelled the data by this test (figure 3, white bars). Furthermore, use of this model increased the likelihood by 334 log units compared with the GTR + G model; this was done by the addition of two extra composition vectors over the GTR + G model, for a total of six extra parameters. This model eroded support for the three-domains tree somewhat (60% BPP) and increased support for the eocyte tree (39% BPP). Although the NDCH(2,2) model adequately modelled the composition as shown by posterior predictive simulation, addition of two more composition parameters (table 2, row H, NDCH(4,4) model; electronic supplementary material, figure S6) improved the likelihood by a further 82 log units. This latter analysis recovered a marked increase in support for the eocyte hypothesis (87% BPP) over a monophyletic archaea (12% BPP). By accommodating a covarion process in the model (table 2, row F; electronic supplementary material, figure S4), the likelihood improved greatly by 224 log units over the GTR + G model. With this model, support for the three-domains tree also increased greatly (to 99% BPP); however, the improvement to the likelihood imparted by the covarion was not as great as that for compositional heterogeneity in these data, with contrary support.

The composition and covarion are not the only aspects of heterogeneity over the tree. The NDRH model developed here allows the rate matrix to differ over the tree in the same way that the composition can differ over the tree in the NDCH model. A tree-homogeneous model such as the GTR + G model for two data partitions can be considered a NDRH(1,1) model; if we have two rate matrices in each of two data partitions, then we have a NDRH(2,2) model. The NDCH and NDRH models can be used together, and are independent of each other.

The primary divisions of life P. G. Foster et al.

Figure 2. Bayesian analysis of combined SSU and LSU rRNA genes. In panel (a) the model is GTR + G, separate in the two data partitions, with free partition rate. This is the analysis in table 2, row D. Panel (b) shows an analysis with the NDCH(4,4) NDRH(2,2) tree heterogeneous model from table 2, row J.

Phil. Trans. R. Soc. B (2009)
Combining extra composition vectors with heterogeneous rate matrices in the NDCH(4,4) + NDRH(2,2) model (figure 2b) improves the likelihood by 70 log units over the NDCH(4,4) model alone. All of these models (table 2, rows H–J) show 86–88 per cent BPP support for the eocyte hypothesis. Interestingly, the LogDet analysis (table 2, row B; electronic supplementary material, figure S8) has 68 per cent BS support for the eocyte hypothesis; LogDet distances are relatively immune to compositional differences over the tree (but see Hirt et al. 1999).

The CAT model accommodates heterogeneity over the data, and using this model increased the fit to the data greatly, as shown by an increase in the likelihood of 1769 log units over the GTR + Γ model, making it by far the best-fitting model used (electronic supplementary material, figure S9). This model had 100 per cent BPP support for the eocyte hypothesis. A notable feature of this analysis is that there was low support for the monophyly of both the euryarcheotes and crenarcheotes.

(b) Support for alternative trees of life based on protein-coding genes

It has been demonstrated previously (Brown et al. 2001; Cox et al. 2008) that equally weighted MP analyses of concatenated proteins across the tree of life support the traditional three-domains hypothesis. With our expanded eocyte sampling, this remained the case with both standard amino acid coding (86% BS for a monophyletic archaeabacteria—table 3, row A; electronic supplementary material, figure S10) and Dayhoff-recoded data (88% BS—table 3, row G; electronic supplementary material, figure S11). However, other methods and models we employed supported the alternative tree of life; the eocyte hypothesis. One possible explanation of these contradictory results is that they are due to the increased tendency of the MP method to suffer from the distorting effects of long-branch attraction (LBA—Felsenstein (1978)), whereby taxa are grouped by an excess of unrecognized homoplasy rather than homologous changes. That we obtain similar results (i.e. the three-domains tree) to previous studies under the MP criterion suggests that our selection of genes and site inclusion was not responsible for our obtaining a different result (i.e. the eocyte tree) from other methods. Further highlighting the importance of composition heterogeneity in the tree of life are the results of LogDet distance analysis, which clearly identifies the eocyte hypothesis in preference to a monophyletic archaeabacteria (88% versus 8% BS, respectively; table 3, row B; electronic supplementary material, figure S12). ML analyses (table 3, row C; electronic supplementary material, figure S13) also resolve the eocyte tree, but with little support (68% BS). By contrast, topologies identifying a monophyletic archaeabacteria are not recovered from samples of the posterior probability distributions of any Bayesian analysis; in fact, all of these analyses support the eocyte hypothesis at, or near, maximum support values (figure 4a,b, table 3, rows D–F; H–J; electronic supplementary material, figures S14–S17). It is notable that including a covarion parameter had little effect on the fit of the model, increasing the log marginal likelihood by only 2 units (table 3, row D versus E). This result suggests that covarion-like rates of lineage evolution are not an important factor in these data. Nevertheless, caution should be urged with regard to this conclusion as the result of combining loci into a single partition may have the effect of homogenizing covarion structures of lineage substitution rates particular to individual loci. Despite this caution, we note that Cox et al. (2008) were only able to identify three genes included in the current dataset where a significant covarion structure was present.

Analyses using the CAT model with both standard amino acid coding and Dayhoff-recoded data fit the data much better than homogeneous models or the tree-heterogeneous NDCH model. That is, the CAT model fits the standard amino acid coded data 26 048 log marginal likelihood units better than the homogeneous WAG + Γ model (table 3, row D versus F), and the same model showed an improvement of 7312 log marginal likelihood units over the homogenous GTR + Γ model when the data were recoded into Dayhoff groups (table 3, row H versus J). Such remarkable improvements in model fit indicate the utility of the CAT model and the importance of modelling across-data compositional heterogeneity. Nevertheless, PHYLOBAYES analyses of the amino acid data did suffer from a lack of convergence between independent runs. In both cases, the standard and Dayhoff coding of the amino acid data, the two runs differed in topology with respect to the placement of Encephalitozoon cuniculi, a taxon whose placement is
known to be problematic (Embley & Martin 2006), but importantly, not with respect to the status of the eocytes or archaeabacterial monophyly. In the Dayhoff-recoded amino acid data analyses, the differences between runs were not well supported (i.e., <95% posterior probability). In the CAT analysis of the standard amino acid coding data, however, there was strong (>95%) support for two alternative placements of E. cuniculi; either at the base of the eukaryotic tree, or with the fungi. In both cases we chose to present the results from the run with the best log marginal likelihood score: Dayhoff-recoded: run 1 = 98 785 versus run 2 = 98 754, and standard coding: run 1 = 220 776 versus run 2 = 220 644.

In our previous analyses (Cox et al. 2008), we found only two proteins, the largest subunits of RNA polymerase I (RP1) and III (RPC1), that resolved archaeabacterial monophyly under the NDCH model. Further analyses of RNA polymerases (RP1, RPB1, RPC1, RPA2, RPB2 and RPC2) with additional eocyte taxa, under both the NDCH and CAT models, failed to find additional support for archaeabacterial monophyly. Indeed, support for a monophyletic archaeabacteria from RP1 under the NDCH(2) model was eroded from 99 per cent to 57 per cent, and NDCH(2) analyses of RPC1 failed to resolve a monophyletic archaeabacteria. All analyses of RNA polymerases under the CAT model failed to resolve a monophyletic archaeabacteria or strongly identify any group as most closely related to the eukaryotes.

(c) Decay of phylogenetic information is delayed by among-site rate variation

A potential criticism of any phylogenetic study based on anciently diverged molecules is that the sequences may be saturated with superimposed mutations, masking the historical signal (Philippe & Forterre 1999; Penny et al. 2001; Ho & Jermiin 2004). One way to visualize saturation is by using saturation plots (Philippe & Forterre 1999). These plots can be made for either simulated data, where the observed pairwise p-distances are plotted against the simulation branch lengths, or they can be made from empirical data where the p-distances are plotted against inferred branch lengths. In simulations using the simple Jukes-Cantor model for DNA (figure 5a), as branch lengths increased, the observed pairwise distances increased but plateaued at 0.75 as the sequences became randomized by superimposed mutations. This appeared to happen at branch lengths above about three mutations per site; if evolution behaved in this way it would be difficult or impossible to make reliable phylogenetic inferences based on such diverged sequences. A similar effect is shown in simulated protein sequences in figure 5b, which used the WAG model. Here saturation appears to have occurred at branch lengths above about 6.

However, the situation changed greatly in our simulations when we allowed the process of evolution to have among-site rate variation. The simulations shown in figure 5c were performed using the WAG + \Gamma model, that is, with gamma-distributed among-site rate variation. Here, we can see that complete saturation was never reached even after an average of 50 mutations per site. When there is among-site rate variation there are both slow sites and fast sites; the fast sites will become saturated at small simulation branch lengths (even sooner than the average sites in simulations without among-site rate variation), but the slow sites will be relatively immune from saturation even at high simulation distances. The biological sequences that we used do show among-site rate variation, and together with invariant sites, these slow sites allow us to recognize that anciently diverged sequences are homologous, and allow us to align sequences from some conserved genes over the entire tree of life with confidence in positional homology.
Using simulations we asked whether we would expect saturation to cause problems for our methods of analysis and our data. In figure 5, points were from simulations based on samples from the posterior distribution of the tree-heterogeneous model analysis shown in row I of table 2. Lack of a plateau showed that in evolutionary scenarios such as this we would not expect complete saturation.

Turning now to the data that we used in this study, we asked whether plots of these data show saturation. Figure 6 shows that neither the rRNA genes, nor the protein sequences, nor the grouped amino acid sequences showed complete saturation. The rRNA genes are perhaps close to saturation, and we can speculate that this contributes to the generally ambiguous results that are shown in table 2 and in the published literature (e.g. Cox et al. 2008). It appears that the protein and grouped amino acid data are not saturated. The achaebacteria–eukaryote pairs are the most relevant to our problem, and those distances, isolated in figure 6, are clearly not saturated, as there are larger pairwise distances with larger p-distances evident in figure 6.

While the saturation curves in figure 6 show that the sequences are not saturated, we also approached the question from a different angle and asked whether we would expect there to be enough phylogenetic signal remaining in protein sequences that have evolved under a WAG + G model at the level of divergence seen in figure 6. Simulations were made using a WAG + G model on a four-taxon tree with terminal branch lengths of 1.5 and an internal branch length of 0.1. The simulation sequences were 5222 characters long, the same length as the concatenated protein...
Figure 5. Saturation plots from simulated data. Simulation distances are simulation branch lengths, measured in average mutations per site. Panel (a) is from DNA simulated under the Jukes-Cantor model. Panel (b) is from protein simulated under the WAG model. The simulations in panels (a) and (b) were performed without among-site rate variation. Panel (c) shows protein simulations under the WAG + Γ model, i.e. including among-site rate variation. Panel (d) shows DNA simulations based on samples from the posterior distribution of the analysis shown in row I of table 2. Lines were fit from the second half of the points.

dataset, and the alpha value for the gamma-distributed rate variation was set to 1.94, the posterior average. For each replicate simulation, the ML of the three possible four-taxon trees was calculated. Of 500 replicates, 483 ML trees (96.6%) were the simulation topology. If we had saturation, we would have expected that the three possible topologies would be the ML tree about one-third of the time each. This shows that even when the sequences had been hit by an average of three mutations per site between taxa pairs (more than is seen between taxa pairs in the Jukes-Cantor model. Panel (a) shows DNA simulated under the WAG model, i.e. including among-site rate variation. Panel (b) shows DNA simulations based on samples from the posterior distribution of the analysis shown in row I of table 2. Lines were fit from the second half of the points.

4. CONCLUSIONS AND IMPLICATIONS FOR THEORIES OF EUKARYOTIC ORIGINS

The alignments that we used were conservative to ensure that our hypotheses of positional homology between domains were as robust as we could make them. In doing so, we removed many positions that could be reliably aligned within domains but not between them. This inevitably removed some signal for relationships within domains and may have contributed to the recovery of some of the controversial or unconventional relationships that we observed in our trees. For example, in most analyses the long-branched microsporidian Encephalitozoon was recovered near the base of the eukaryote cluster, rather than with fungi where most data would place it (Embley & Martin 2006). The CAT method is reported to be more robust to long-branch artefacts than other methods (Lartillot et al. 2007); so it was interesting to see that CAT analyses of the concatenated proteins did indeed unite Encephalitozoon with Saccharomyces (figure 4b; electronic supplementary material, figure S17). Most analyses recovered the Crenarchaeota/eocytes and the euryarchaeotes as monophyletic groups, as classically depicted in both the three-domains and eocyte trees. By contrast, the CAT analyses recovered euryarchaeotes as a paraphyletic group, with Pyrococcus as the sister to the Crenarchaeota/eocytes plus eukaryotes. It will be interesting to test how robust these relationships are to increased taxon sampling and to an expanded sequence alignment. Taken at face value, they raise the possibility that Crenarchaeota/eocytes plus eukaryotes may have originated from within the Euryarchaeote radiation.

There is currently a debate about how far back molecular phylogenetics might be able to take us (e.g. Philippe & Forterre 1999; Penny et al. 2001; Ho & Jermiin 2004). Phylogenetic methods will generally construct trees irrespective of whether any historical signals for relationships remain in the data. The message from computer simulations is that success in recovering any ancient signal is related to the properties of the data, for example whether it contains a mixture of site rates, and how well the model fits the data (Penny et al. 2001; Ho & Jermiin 2004; Lartillot et al. 2007). Our simple simulations illustrate some of these issues and are consistent with the possibility of there being signal for historical relationships in the data that we analysed. There are unicellular microfossils (acritarchs) argued to be eukaryotic in strata of about 1.45 Gyr of age that provide one estimate for a minimal age for eukaryotes (Javaux et al. 2001). This figure is consistent with an age of between 950 and 1259 Myr for the diversification of major eukaryotic groups that has been estimated from concatenated molecular sequence data using a relaxed molecular clock (Douzery et al. 2004).

The three-domains hypothesis (Woese et al. 1990) explains the similarities in the eukaryotic and archaeobacterial transcription and translation machinery (Zillig et al. 1985; Olsen & Woese 1997), as originating in a common ancestor that was not shared with eubacteria. This putative common ancestor was subsequently called a neomuran by Cavalier-Smith (2002). By contrast, the eocyte hypothesis posits that the observed similarities reflect the origin of eukaryotes from within the archaeabacteria as the sister group of a specific group of archaeabacteria called the Crenarchaeota or eocytes. In the current work, we have investigated the support for these competing hypotheses from genes and proteins that largely, but not exclusively, comprise components of the salient genetic machinery (Cox et al. 2008). Our results, based upon an increased sampling of eocytes, are in agreement with those that we published earlier (Cox et al. 2008), namely, when methods are used that are designed to overcome across-tree (Foster 2004), or across-data compositional heterogeneity (Lartillot & Philippe 2004), features that are manifestly evident for these data, it is the eocyte tree that is favoured and not the three-domains tree.
Figure 6. Saturation plots of empirical data. Inferred distances are patristic distances between taxa pairs following the tree path. Panels (a)–(d) show all points; panels (e)–(h) isolate pairs where one member of the pair is a eukaryote sequence and the other is an archaeabacterial sequence. Panels (a) and (c): rRNA data with the tree-heterogeneous model shown in row I in table 2. Panels (b) and (f): protein sequences analysed with the WAG + Γ model. Panels (c) and (g): protein sequences analysed with the CAT model. Panels (d) and (h): protein sequences recoded into the six Dayhoff groups and analysed with a GTR + Γ-like model.

We thank Dan Swan for administration of the Newcastle University computing facilities. This work was supported in part by Biotechnology and Biological Sciences Research Council (UK) grant BB/C508777/1 to P.G.F and T.M.E and by a Wolfson Research Merit Award from the Royal Society to T.M.E.

REFERENCES


Pace, N. R. 2006 Time for a change. Nature 441, 289. (doi:10.1038/441289a)


