Update TRENDS in Ecology and Evolution Vol.20 No.3 March 2005

Research Focus

New methods ring changes for the tree of life

James O. McInerney[1] and Mark Wilkinson[2]

[1]Bioinformatics Laboratory, Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland
[2]Department of Zoology, The Natural History Museum, London, UK, SW7 5BD

Relationships among prokaryotes and the origin of eukaryotes have both proven controversial, with results depending upon the gene sequences and methods used. Extensive horizontal gene transfer is one possible reason why inferring such deep phylogenetic relationships is difficult. In two recent papers, Lake and Rivera introduce new methods that can be used to reconstruct the genomic tree in the presence of horizontal gene transfers, but which suggest that a ring rather than a tree is a better representation of some parts of the history of life on Earth.

The development of eukaryotic cellular organization from an anucleate prokaryotic ancestor is one of the major transitions in the history of life, and theories of eukaryotic origins and their phylogenetic relationships with prokaryotes abound [1–5]. One reason for the seeming lack of consensus is that inferring deep divergences in the history of life is difficult. There has been enough time for signal to be overwritten as noise and for systematic biases to accumulate in the sequence data that are usually used to reconstruct phylogeny. There has been enough time for hidden paralogy and horizontal gene transfer (HGT), both of which can yield incorrect species or genome trees even when the gene trees are correctly inferred. It has been suggested that HGT is sufficiently extensive to call into question the existence of a genomic phylogeny [6], and some theories postulate the occurrence of genomic fusion events [7], which are not accommodated in phylogenetic trees. Now, new methods that appear to be insensitive to HGT and that have the potential to reconstruct genomic fusions have been developed [8,9] and applied to an exhaustive collection of putative orthologs from completed genomic sequences to provide the strongest test to date of eukaryotic origins.

Eukaryotic origins

The theory of eukaryotic origins that is most common in textbooks is based on phylogenetic trees that are inferred from the small subunit ribosomal RNA sequence [4,5] with a root provided by analysis of paralogous H1-ATPase genes [10]. In this scheme, the most recent common ancestor of all life was prokaryotic, the prokaryotes are divided into monophyletic Bacteria and Archaea, and the eukaryote lineage separated from the archaeal lineage before the diversification of the extant Archaea. The eukaryotic cell type then developed, gradually or otherwise, after this diversification. Analysis of archaeal translation indicates that it is more similar to eukaryotic translation than to bacterial translation and this seemed to justify the rRNA tree [11].

In addition to the presence of nuclei, most eukaryotes also differ from prokaryotes in having mitochondria, organelles that are essential for aerobic metabolism and that are generally accepted to be descended from an endosymbiotic proteobacterium [12]. Early phylogenetic trees, such as the rRNA tree, placed the few amitochondriate eukaryotes (those lacking mitochondria) as early-branching lineages, which resulted in the Archaezoa hypothesis [4]: that is, that these eukaryotes are primitively amitochondriate (i.e. have never had mitochondria) and that the mitochondrion endosymbiosis was a relatively late event in the history of eukaryotes [2].

One of the first indications that this scenario was too simplistic was the finding that one group of amitochondriate eukaryotes (the microsporidia) were highly modified fungi and, therefore, secondarily amitochondriate [13]. It has since been suggested that all eukaryotes have other organelles, such as hydrogenosomes, that are highly derived mitochondria [14].

The notion of a fusion event that would have created the eukaryote dates back to 1980 [3]. Genomic fusion events, distinct from the mitochondrial endosymbiosis, have been speculated many times but without strong empirical evidence (rather than simply isolated incidences of HGT problems with phylogeny reconstruction or absence of sufficient data) [7]. Six years ago, Martin and Muller [15] proposed a detailed and thoughtful scenario where equal contributions from a methanogen and a bacterium could have given rise to the eukaryote. In 2004, it was shown that genes of archaeal and/or bacterial origins contribute significantly to the yeast genome although two fusion partners were not identified [1]. However, until now, no phylogenetic reconstruction method has produced a scenario where completed genome information has been unequivocal about the origin of the eukaryote.

Inferring the ring of life

Lake and Rivera's new method, termed 'conditioned reconstruction' [8], is based on a Markov model of genomic evolution, analogous to models of nucleotide substitution that are routinely used to infer trees from aligned sequence data [16]. In genomic history, individual genes are stochastically lost or gained to produce the patterns of presence (P) and absence (A) of orthologous genes

Corresponding author: McInerney, J.O. (James.O.Mcinerney@may.ie).

observed in sampled genomes. The multiple possible causes of gene loss or gain, such as the causes of mutations in models of sequence evolution, are not specified. Thus HGT, which can give rise to gene trees that are not species trees, will not similarly mislead the inference of genomic history, provided that it, and the other causes of change in gene content, are adequately described by the model of gene loss or gain. Inasmuch as patterns of gene presence and absence produced by HGT must sometimes conflict with the genomic phylogeny, these conflicts are analogous to those produced by homoplasy (multiple hits) in sequence evolution. In both contexts, homoplasy and HGT are unproblematic to the extent that they can be adequately modelled and do not occur at too high a rate. The authors also consider gene presence and absence characters to be evolving more slowly than are nucleotide characters and, thus, as more useful for discerning deep phylogenetic signals.

Genomes contain information about the rates of gene loss and gain in the frequencies of patterns of joint presence or absence of genes. Given two genomes, for any gene there are four possible patterns of joint presence or absence (PP, PA, AP, and AA). The proportions of the first three patterns across any two genomes can be readily determined, but a problem for any Markov method using absence–presence data is how to estimate the numbers of shared absences of genes (AA). The authors' solution is to use one or more conditioning genomes. Genes that are present in the conditioning genome, but absent in the two genomes under consideration, provide an estimate of shared absence. Choice of conditioning genome is arbitrary, but use of a range of conditioning genomes enables the potential impact of conditioning to be assessed.

An equally important innovation of Lake and Rivera's method is that it can be used to determine whether the conflict among alternative trees yielded through bootstrap analyses can be reconciled in terms of well supported cyclic, rather than tree-like, interrelationships. This means that, unlike previous methods, conditioned reconstruction can, in principle, detect genomic fusion events, representing them as cycle graphs (Figure 1). Usefully, only the removal of genomes that are the result of fusion can break the circle, enabling identification of those genomes that are the product of a merger. The subsequent analysis of Lake and Rivera's method [9] used LogDet/Paralinear distances [17] to counter potential big genome artefacts, with pattern filtering [18] to counter rate heterogeneity and bootstrappers gambit [19] to construct and assign probabilities to multitaxon trees, methods also developed by Lake to better resolve deep phylogenetic relationships. The analysis provides strong bootstrap support for the hypothesis that the eukaryotic genome is the product of the fusion of the genomes of an archeum and a bacterium. Consistency across a range of conditioning genomes further boosts this support. Rivera and Lake's conclusion is that, at this major point in history, there is a ring rather than a tree of life (Figure 1).

Rivera and Lake's findings explain the apparent chimaeric genome of eukaryotes [1] and, although their analysis did not enable the precise identification of the source of the bacterial contribution to the eukaryotic

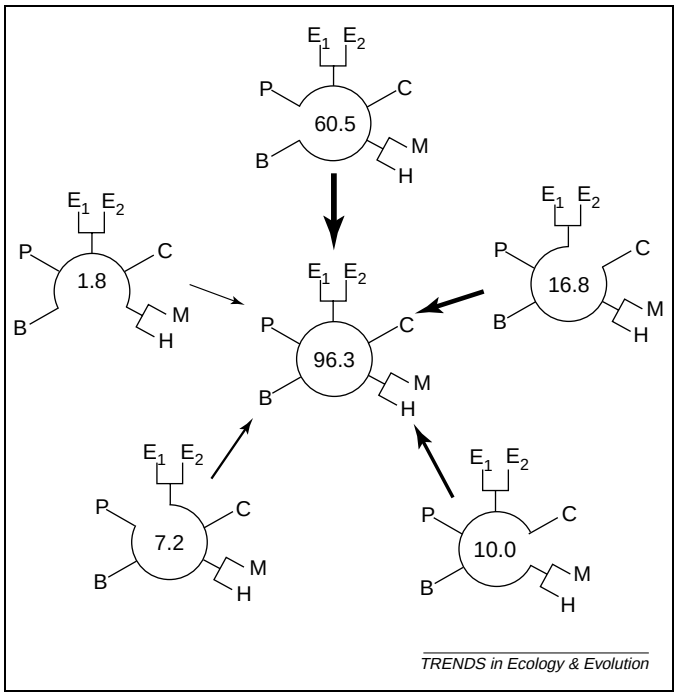

Figure 1. An example of conditioned reconstruction of the ring of life. The figure shows five best-supported unrooted trees arranged clockwise in decreasing order of their frequency of occurrence in bootstrap analyses, and, in the centre of the figure, the cyclic phylogeny that reconciles the conflict among the unrooted trees with high cumulative bootstrap support. Reconstructions are based on gene presence and absence data for two eukaryotes (E_1 and E_2), a crenarchaeote or eocyte (C), a halobacterium (H), a methanococcus (M), a bacillus (B) and a proteobacterium (P), conditioned on the genome of an archaeoglobium (not shown). Numbers are bootstrap proportions. Modified, with permission, from [9].

genome, it is consistent with it being one and the same as the ancestor of the mitochondrion. It also implies that eukaryotes arose from within both the Bacteria and the Archaea, and that neither group is monophyletic. This scenario could not be more different to those based on the rRNA tree [5], and suggest that the position of eukaryotes in such rRNA trees is a long-branch attraction artefact. The longest branches on the rRNA tree are usually the branch leading to the eukaryotes and the branch separating the Bacteria from the Archaea.

Prospects

The power and limitations of the new methods require further theoretical and empirical scrutiny. Model-based methods generally perform well when the model is adequate, but all are known to fail in cases where the model is not. Conditioned reconstruction is well founded, but the impact of violations of the assumptions on which it is based, and the adequacy of those assumptions in practice, are insufficiently known. The method depends crucially upon the identification of sets of orthologous genes and its sensitivity to how this is achieved is unknown. The extent to which the methods will recover a meaningful genomic phylogeny when rates of HGT, or other mechanisms of gene loss or gain, are high is unclear, as are the effects of rate variation across and non-independence of genes. It is also unclear to what extent the method generalizes to cases where there are multiple rings in the cycle graphs. Using only presence–absence data, the method makes no use of the information

available from comparisons of sequences and it might be that the method of Rivera and Lake of reconciling incongruence in terms of rings rather than trees can also be extended to incongruent gene trees.

The new method has been applied just twice, with careful analyses providing promising and challenging results that should be taken seriously and scrutinized further. The ring of life also raises questions concerning the nature and timescale of the eukaryotic genome fusion and to what extent genomic fusions might have occurred elsewhere in the early history of life.

References

1 Esser, C. *et al*. (2004) A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol. Biol. Evol.* 21, 1643–1660
2 Sogin, M.L. and Silberman, J.D. (1998) Evolution of the protists and protistan parasites from the perspective of molecular systematics. *Int. J. Parasitol.* 28, 11–20
3 Van Valen, L.M. and Maiorana, V.C. (1980) The archaebacteria and eukaryotic origins. *Nature* 287, 248–250
4 Woese, C.R. and Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5088–5090
5 Woese, C.R. (1987) Bacterial evolution. *Microbiol. Rev.* 51, 221–271
6 Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science* 284, 2124–2129
7 Golding, G.B. and Gupta, R.S. (1995) Protein-based phylogenies support a chimeric origin for the eukaryotic genome. *Mol. Biol. Evol.* 12, 1–6
8 Lake, J.A. and Rivera, M.C. (2004) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Mol. Biol. Evol.* 21, 681–690
9 Rivera, M.C. and Lake, J.A. (2004) The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* 431, 152–155
10 Gogarten, J.P. *et al*. (1989) Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 86, 6661–6665
11 Marsh, T.L. *et al*. (1994) Transcription factor IID in the Archaea: sequences in the *Thermococcus celer* genome would encode a product closely related to the TATA-binding protein of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 91, 4180–4184
12 Andersson, S.G. *et al*. (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396, 133–140
13 Hirt, R.P. *et al*. (1999) Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proc. Natl. Acad. Sci. U. S. A.* 96, 580–585
14 Embley, T.M. *et al*. (2003) Hydrogenosomes, mitochondria and early eukaryotic evolution. *IUBMB Life* 55, 387–395
15 Martin, W. and Muller, M. (1998) he hydrogen hypothesis for the first eukaryote. *Nature* 392, 37–41
16 Yang, Z. (1995) A space-time process model for the evolution of DNA sequences. *Genetics* 139, 993–1005
17 Lake, J.A. (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proc. Natl. Acad. Sci. U. S. A.* 91, 1455–1459
18 Lake, J.A. (1998) Optimally recovering rate variation information from genomes and sequences: pattern filtering. *Mol. Biol. Evol.* 15, 1224–1231
19 Lake, J.A. (1995) Calculating the probability of multitaxon evolutionary trees: bootstrappers gambit. *Proc. Natl. Acad. Sci. U. S. A.* 92, 9662–9666

0169-5347/$ - see front matter
doi:10.1016/j.tree.2005.01.007

Letters

Providing baselines for biodiversity measurement

Katherine J. Willis, Lindsey Gillson, Terry M. Brncic and Blanca L. Figueroa-Rangel

Oxford Long-term Ecology Laboratory, School of Geography and the Environment, University of Oxford, Mansfield Road, Oxford, UK, OX2 7LE

One of the major, yet largely unacknowledged, weaknesses in biodiversity measurement is that many of the data sets used in biodiversity assessments span less than one full generation of the organisms under study. At least seven international biodiversity assessments have been published in the past five years, but rarely do they use temporal records that are longer than 50 years (Table 1). As a result, policy documents such as the EC Biodiversity Action Plan for the Conservation of Natural Resources (http://www.epbrs.org/epbrs_library.html) neglect the historical dimension altogether. Whereas 50 years might be an acceptable timeframe for some herbaceous plants and animals, the average generation time of many organisms, such as trees, is much greater than this. There are many instances where the use of longer-term data could add much to biodiversity assessments and provide the very type of information, (e.g. time-series data, data on abundance and how it varies, and long-term distribution data) highlighted in many of these reports as a 'critical gap in our knowledge'.

Why is longer-term data not being included? Outside of Quaternary science, there is a general lack of awareness of palaeoecological techniques and what they can tell us. Over the past 20 years, there have been significant advances in palaeoecological research and, contrary to popular belief, it is possible to obtain high-resolution temporal and taxonomic analyses that reveal annual variations in communities over hundreds to thousands of years. However, this information is not filtering through to the biodiversity community, one of the underlying reasons being the problem of its dissemination. Much of this longer-term data is published in journals that are not read by the conservation community and the data are presented in ways that are difficult for non-specialists to interpret. Pollen diagrams are a case in point. Although these are the traditional display tools of palaeoecologists, to those outside the discipline they represent confusing diagrams where multiple axes are plotted the wrong way

Corresponding author: Willis, K.J. (kathy.willis@ouce.ox.ac.uk).
Available online 6 January 2005