Minireview

Topological variation in single-gene phylogenetic trees

Jose Castresana

Address: Department of Physiology and Molecular Biodiversity, Institute of Molecular Biology of Barcelona, CSIC, 08034 Barcelona, Spain.
Email: jcvagr@ibmb.csic.es

Published: 13 June 2007

Genome ***Biology*** 2007, **8:**216 (doi:10.1186/gb-2007-8-6-216)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/6/216

Abstract

A recent large-scale phylogenomic study has shown the great degree of topological variation that can be found among eukaryotic phylogenetic trees constructed from single genes, highlighting the problems that can be associated with gene sampling in phylogenetic studies.

In 1982, Penny, Foulds and Hendy [1] made a test of the theory of evolution by comparing phylogenetic trees constructed from different protein-coding genes from the same set of species. Specifically, they tested whether a unique evolutionary tree relating these genes existed or not, and whether it could be recovered. The existence of such a tree was important, not only to confirm the theory of evolution, but also to show that this theory allowed quantitative and falsifiable predictions. At that time, five proteins from 11 mammalian species were available for the study, but each protein produced different trees. At first sight, this contradicted the existence of a unique tree. However, the authors, being aware of the methodological difficulties in phlyogenetic reconstruction, did not expect the five trees to be identical. Rather, they expected them to be similar. To measure topological dissimilarity between trees they made use of the symmetric difference distance (also know as the Robinson-Foulds distance), which had just been introduced [2], and found that the trees obtained from the five genes were indeed more similar than expected by chance, proving the existence of a unique tree relating these sequences. This was a simple but powerful study that opened the way to test evolutionary hypotheses by means of multi-gene studies or ‘phylogenomics’.

Now, 25 years later, a much larger multi-gene study has been published by Huerta-Cepas *et al.* [3], using the complete human proteome and the homologous genes in the other complete genome sequences now available. This time, 21,588 trees, each including a different human gene, were obtained from the genomes of 39 species of eukaryotes. What can we learn from such a large number of phylogenies? Nowadays, we take it as read that eukaryotes are related by a unique phylogenetic tree. Instead, the focus of recent phylogenetic work has shifted to studying whether we can determine the exact branching pattern of this tree. The results have been mixed. Many nodes, or branchpoints, on the eukaryotic tree were well known before the advent of molecular sequences, and molecular phylogenies have simply confirmed them. In other cases, unexpected nodes, such as the one that splits off a group of mammals of African origin into the Afrotheria [4,5], or the one that groups the hemichordates and echinoderms together in the Ambulacraria [6,7], have found strong and congruent support from different genes. On the other hand, there is contradictory support for a large number of nodes or phylogenetic relationships, and even the analysis of complete genomes has not helped to resolve some of these. With all the data accumulated in the past few years, philosophical concerns about being able to trace the existence of a unique tree have vanished, at least in eukaryotes (here, only some exceptional genes, such as those coming from endosymbiotic events, confuse attempts to trace lineages by the usual rules), but there are still many methodological problems that can blur the outlines of the tree.

The human phylome

Huerta-Cepas *et al.* [3] have based their analysis exclusively on genes present in humans, and thus produced a set of phylogenies that all include the lineage leading to humans. The authors call this collection of trees ‘the human phylome’. They then focused on three parts of the eukaryotic tree that have been the subject of some contention. There are not

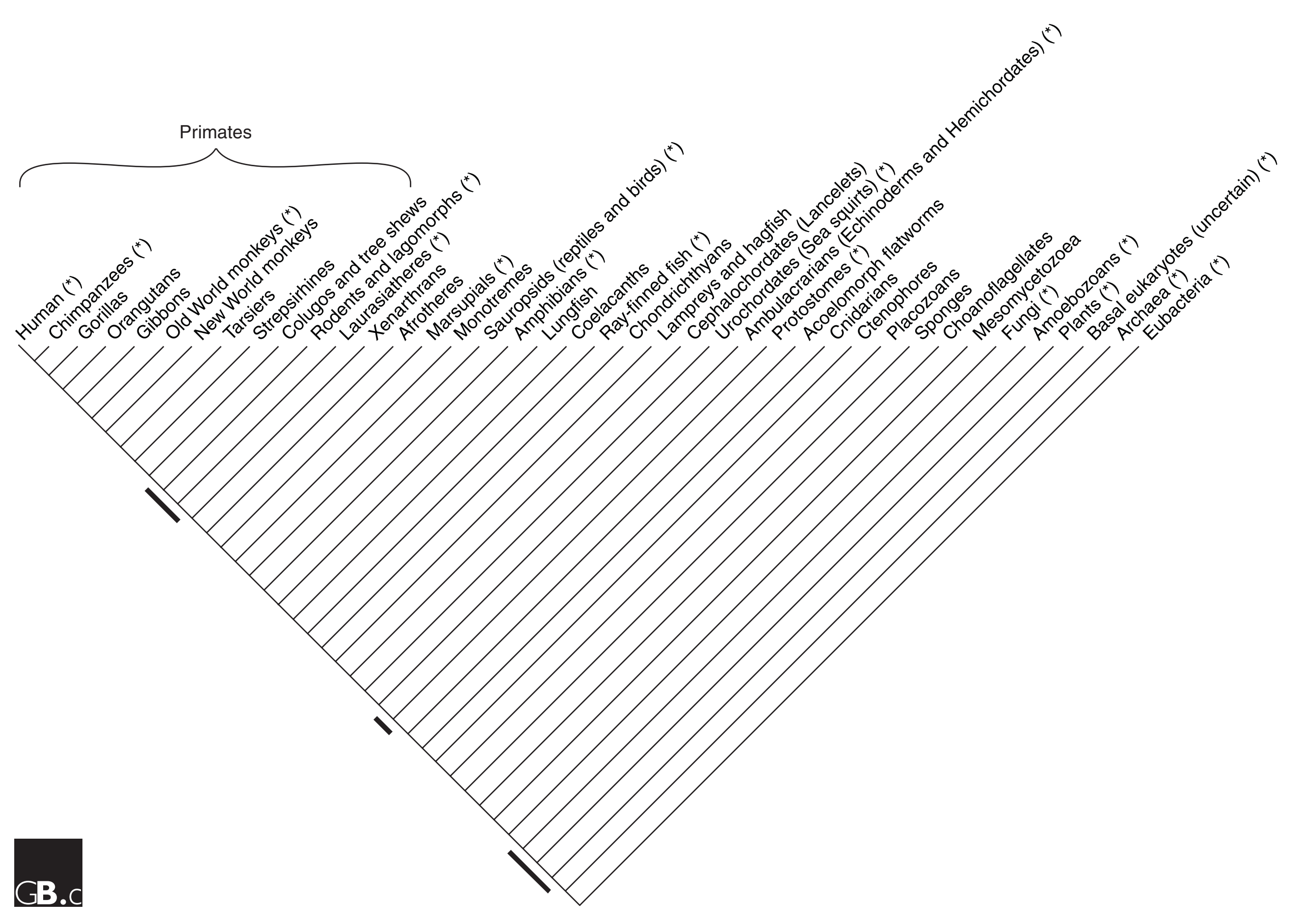

Figure 1

A phylogenetic tree depicting all nodes (branchpoints) on the evolutionary line leading to humans. The species groupings (clades) given here along the top are those used in [8]. The 'basal eukaryotes' are a diverse polyphyletic group (not a single clade) of mainly unicellular organisms such as excavates and chromalveolates, and this grouping is thus labelled 'uncertain'. The nodes studied by Huerta-Cepas *et al.* [3] are indicated by the black bars. The clades in which at least one complete genome sequence is available are marked with an asterisk. All eukaryotic clades with a genome sequence were included in the phylogenetic analysis of Huerta-Cepas *et al.*, except for the Ambulacraria, for which a genome sequence (of the echinoderm *Strongylocentrotus purpuratus*) has only recently become available.

many nodes that can be addressed if you wish to always include humans in the tree. In his recent book, *The Ancestor's Tale: A Pilgrimage to the Dawn of Life*, Richard Dawkins [8] noted around 39 'splits' of the tree of life that contain the human lineage as one of their branches (Figure 1). This number is provisional of course, and may get a bit bigger when the basal eukaryotic radiation is better resolved. It is pure coincidence that the number of fully sequenced genomes available to Huerta-Cepas *et al.* was also 39, and there is no close relationship between these 39 genomes and the 39 splits that contain the human lineage: some of the clades connecting to the human path have received much attention from sequencing projects (fungi, for example) whereas others have received none so far (gorillas or cnidarians, to name just a few). (A clade is a grouping of species descended from a particular common ancestor that is not the ancestor of any other species.) In addition, more than two genomes per node are necessary to study some relationships of interest (for example, the relationship between nematodes and arthropods to resolve the protostome node).

With the genome sequences currently available (marked by asterisks in Figure 1), less than half of the 39 nodes containing the human lineage can be tackled. Although this is not a large number compared to the millions of ramifications in the tree of life, these are precisely the nodes that have received most scientific scrutiny by phylogenetic analysis of either genomes or single genes. The most recent split between the human lineage and an extant lineage has been thoroughly studied during the past few years, and a consensus has gradually arisen in which chimpanzees, and not gorillas, are our closest relatives [9]. The most ancient split in Figure 1 separates the Eukaryote-Archaea clade and the Eubacteria, but this phylogeny is still highly debatable and other possibilities exist [10]. The nodes selected by Huerta-Cepas *et al.* [3] for further study lie between the two ends of the path that goes from humans to the last common ancestor of all living species (or "the pilgrimage to the dawn of life" [8]).

Topological variation

One of the phylogenetic problems analyzed by Huerta-Cepas *et al.* [3] is the relationship between primates, rodents, and laurasiatherians (the latter comprising the Cetartiodactyla, which include whales and artiodactyles, as well as the Carnivora, and certain other mammalian orders). By means of an algorithm that scans topologies in the trees of the human phylome, the authors quantified the number of trees supporting different relationships. They found, after eliminating unstable trees, 4,806 phylogenetic trees supporting the grouping of primates and laurasiatherians into a clade with the exclusion of rodents, 3,459 trees supporting a primates and rodents grouping (a clade known as Euarchontoglires or Supraprimates, and supported by recent molecular phylogenies [5]; this is the arrangement depicted in Figure 1), and 2,258 trees grouping rodents and laurasiatherians in a single clade. Thus, the topological variation found was extreme, not far from the maximum possible, and represents a serious methodological challenge, especially as all these trees are statistically well supported, with a Bayesian posterior probability higher than 0.9 in the node of interest. Given the large numbers of genes supporting each of the three possible arrangements of these mammalian lineages, it is not surprising that recent phylogenomic studies have produced different trees relating human, mouse and dog [11,12]. Huerta-Cepas *et al.* [3] did not calculate a consensus tree (this was not the purpose of this study), and thus it is not straightforward to determine the 'true' tree topology relating these mammals. Just getting the best-supported topology is not enough, and even using all genes in a genome may not help you come to an unambiguous solution. This is because different genes produce different biases, and rigorous criteria for selecting the genes to be used to build a species tree are necessary to get less ambiguous results, as has been done in other work (see [13] for a review). The important message from this part of the study is that, whatever the true tree may be, trees derived from single genes are more likely than not to point to a wrong topology.

Huerta-Cepas *et al.* [3] also looked at the relationship among chordates, arthropods and nematodes, a tree that has been the subject of much recent work (see references in [3]). In this case, 2,431 trees support a grouping of chordates and arthropods (Coelomata), 1,759 trees support a nematode-arthropod clade (the Ecdysozoa; in Figure 1, this group is included in the protostomes) and 1,040 trees support a grouping of chordates and nematodes. A great diversity of topologies was also found and we can see again that, even without knowing the true tree, most trees must be wrong. A third problem studied by Huerta-Cepas *et al.* [3] regarding the position of several basal eukaryotic lineages is more difficult to interpret, as there are more than three possible topologies, but the results also point to a high variability among topologies.

It is true that the three examples discussed above are inherently difficult phylogenies, but the authors indicate that they found considerable levels of topological diversity in trees of other, undisputed, phylogenies. These very instructive results should make us realize that not all single-gene trees, even those with high support, must necessarily be coincident with the real species tree. Thus, the methodological approach of the pioneering work of Penny *et al.* [1], which implied a certain degree of topological variation among different genes without denying the existence of a unique tree, is largely supported from this much larger analysis using the most up-to-date methods of statistical analysis.

What causes this variation?

There are many factors that can cause different genes to give different topologies, and there are excellent reviews on this topic [14,15]. Briefly, there are three basic sources of variation. First, there is an important natural source of variation between genes due to the stochastic nature of mutation; short genes are most affected by this randomness, so that the mutations found in such a gene in different species may not be enough to truly reflect their phylogeny. Lineage sorting, which implies the random retention of ancestral polymorphisms in diverging lineages, is also an important natural source of variation in the phylogenies of closely related species. There are also phylogenetic reconstruction artifacts, such as those due to base-compositional bias, saturation of substitutions, or the artificial grouping of the most rapidly evolving lineages (long-branch attraction). The use of time-consuming Bayesian phylogenetic methods by Huerta-Cepas *et al.* [3] (despite the huge number of trees involved) certainly helped in reducing these problems. Finally, there are methodological problems related to the assessment of homology: this includes determining which genes are true orthologs and building the multiple alignments on which phylogenetic reconstructions are based. Orthologous genes are homologous genes that have been separated by speciation and not by gene duplication, and only orthologs should be used for building a species tree. If undetected, genes related by duplication events (paralogs) can lead to serious misinterpretations of species trees. These considerations are particularly problematic for the most ancient phylogenies, where large numbers of gene duplications and gene losses have to be recognized and resolved.

The study of Huerta-Cepas *et al.* [3] has uncovered a degree of topological variation among single-gene phylogenies much greater than previously thought. Their conclusions, although based on eukaryotes, may be applicable to the whole tree of life, and may be important to prokaryote phylogeny. In prokaryotes, besides attempts to determine the phylogenetic position of species or lineages by means of an accurate selection of genes [16,17], there are many studies where the main purpose is to deduce the phylogenetic history of individual genes. Such lineages that do not coincide with their expected position in a species tree are often assigned to lateral gene transfer [18,19], but in many cases this ignores the fact that similar, rather than identical, trees should be expected from different genes [1,3]. In addition, when paralogy problems occur, very dissimilar trees, even with high support, are also to be expected [3]. Thus, this large study by Huerta-Cepas *et al.* [3] reinforces the idea that the details of complex phylogenies (and most of the interesting nodes are complex) can only be solved by means of multi-gene studies after a careful selection of genes. However, in many circumstances a single-gene phylogeny may be interesting in itself. In such cases, not only should we be aware of the problems of orthology assignment and tree reconstruction artifacts, we should try hard to identify them to avoid erroneous speculations from such trees.

Acknowledgements

JC is supported by grant number CGL2005-01341/BOS from the Plan Nacional I+D+I of the MEC (Spain), co-financed with FEDER funds.

References

1. Penny D, Foulds LR, Hendy MD: **Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences.** *Nature* 1982, **297:**197-200.
2. Robinson DF, Foulds LR: **Comparison of phylogenetic trees.** *Math Biosci* 1981, **53:**131-147.
3. Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T: **The human phylome.** *Genome Biol* 2007, **8:**R109.
4. Springer MS, Cleven GC, Madsen O, de Jong WW, Waddell VG, Amrine HM, Stanhope MJ: **Endemic African mammals shake the phylogenetic tree.** *Nature* 1997, **388:**61-64.
5. Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ: **Molecular phylogenetics and the origins of placental mammals.** *Nature* 2001, **409:**614-618.
6. Turbeville JM, Schulz JR, Raff RA: **Deuterostome phylogeny and the sister group of the chordates: evidence from molecules and morphology.** *Mol Biol Evol* 1994, **11:**648-655.
7. Castresana J, Feldmaier-Fuchs G, Yokobori S, Satoh N, Pääbo S: **The mitochondrial genome of the hemichordate *Balanoglossus carnosus* and the evolution of deuterostome mitochondria.** *Genetics* 1998, **150:**1115-1123.
8. Dawkins R: *The Ancestor's Tale: A Pilgrimage to the Dawn of Life.* London: Phoenix; 2005.
9. Ruvolo M: **Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets.** *Mol Biol Evol* 1997, **14:**248-265.
10. Gribaldo S, Philippe H: **Ancient phylogenetic relationships.** *Theor Popul Biol* 2002, **61:**391-408.
11. Cannarozzi G, Schneider A, Gonnet G: **A phylogenomic study of human, dog, and mouse.** *PLoS Comput Biol* 2007, **3:**e2.
12. Nikolaev S, Montoya-Burgos JI, Margulies EH, NISC Comparative Sequencing Program, Rougemont J, Nyffeler B, Antonarakis SE: **Early history of mammals is elucidated with the ENCODE multiple species sequencing data.** *PLoS Genet* 2007, **3:**e2.
13. Delsuc F, Brinkmann H, Philippe H: **Phylogenomics and the reconstruction of the tree of life.** *Nat Rev Genet* 2005, **6:**361-375.
14. Jeffroy O, Brinkmann H, Delsuc F, Philippe H: **Phylogenomics: the beginning of incongruence?** *Trends Genet* 2006, **22:**225-231.
15. Rokas A, Carroll SB: **Bushes in the tree of life.** *PLoS Biol* 2006, **4:**e352.
16. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life.** *Science* 2006, **311:**1283-1287.
17. Soria-Carrasco V, Valens-Vadell M, Peña A, Antón J, Amann R, Castresana J, Rosselló-Mora R: **Phylogenetic position of *Salinibacter ruber* based on concatenated protein alignments.** *Syst Appl Microbiol* 2007, **30:**171-179.
18. Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau ME, Nesbo CL, Case RJ, Doolittle WF: **Lateral gene transfer and the origins of prokaryotic groups.** *Annu Rev Genet* 2003, **37:**283-328.
19. Brown JR: **Ancient horizontal gene transfer.** *Nat Rev Genet* 2003, **4:**121-132.