Research article
Open Access
CrossMark

Assessing clade wise concordance between phylogenetic trees and corresponding taxonomic trees

Abstract

Taxonomic trees are based on a large number of characters while phylogenetic trees consider single or multiple traits of a fixed set of species. We compute the clade-by-clade similarity between two trees as Taxonomic fidelity Index (F). In contrast to monogenic traits, the topology of phylogenetic trees increasingly resembles the taxonomic tree.

Keywords: taxonomic tree, phylogenetict, clade, tree topology, clustering algorithm, taxonomic, fidelity index f

Volume 4 Issue 1 - 2016

Milner Kumar,[1] Sohan P Modak[2]

[1]Department of Zoology, Karnataka University, India
[2]Department of Biotechnology, Institute of Bioinformatics and Biotechnology University of Pune, India

Correspondence: Sohan P Modak, 2Department of Biotechnology, Institute of Bioinformatics and Biotechnology University of Pune, Open vision, 759,75, Deccan Gymkhana, Pune 411004, India, Email spmodak@gmail.com

Received: September 07, 2016 | **Published:** September 27, 2016

Introduction

Darwin,[1] used morphological characters or polygenic traits to describe the hierarchy in complexity among related and distant species although he was unaware of the source of variation. Only later, after the discovery of Mendel's laws explained that the genetic variations are generated by mutations. Speciation is based on the analysis of the extent of similarity among a wide variety of morphological, physiological, biochemical, genetic, behavioral traits[2] allows establishment of evolutionary relationship among organisms that are expressed in form of phylogenetic trees that mimic taxonomic trees. However, a phylogenetic tree based on comparison of a monogenic trait such as specific gene/polypeptide sequences differs from that based on morphological and functional phenotypes/traits that are necessarily polygenic and represent consensus topology.

With the availability of nucleotide- and amino acid sequences the field of molecular systematics has emerged that complemented phylogenetic systematics that provoked the controversy between classical taxonomy and Phylogenetic cladistics.[3–11] Indeed, evolution manifests in form of changes in the whole organism and not a single gene, which is subject to random mutations at variable rates and cannot alone affect the principal phenotype of the organism. Instead, one would expect that a number of gene cohorts operating in concerted manner lead to changes in the polygenic phenotype. In contrast to multicellular organism in which all cells develop from the same original founder cell with identical genetic makeup, closely related, but not identical, organisms designated as a different species, will be expected to possess a very similar genetic makeup except for those structures/functions that differ in the DNA or polypeptide script. By aligning such sequence strings, alphabet by alphabet, for the same gene or polypeptide from different organisms allows quantifying the extent of their closeness or differences. To compare multiple species, one carries out multiple sequence alignment wherein the comparison is still carried out between all possible species pairs to obtain a matrix of all-pairs distances that serves as the basis for building a dendrogram / tree.[12,13]

One, of course, needs benchmarks against which the topology of a given phylogenetic tree is assessed. For example, one such popular benchmark involves 16s ribosomal RNA that is a relatively conserved housekeeping molecule in cells.[7] rDNA sequences available from a wide variety of organisms are compared to construct DNA phylogenetic trees and used as to supplement or even supplant classical taxonomic trees.[14,15] However, multiple sequence alignment of different biophysical traits, namely, isoelectric points and immuno-cross-reactivity or monogenic traits based on nucleotide sequences of a gene, the coding region in mRNA and amino acid sequences from the same set of species reveal considerable differences in phylogenetic tree topologies leading to controversial interpretations even on the relative phylogenetic position of taxa that are considered as evolutionary links.[16–22] One would think that a comparison of entire genome sequences would yield meaningful insight in the evolutionary relationships. While this has yet to happen,[23] in depth analysis and visualization of genomic signatures based on the fractal structure of nucleotide sequences do reveal considerable phylogenetic differences.[24,25] In any case, these need further analysis to elucidate the positional differences in the frequencies of occurrence and localization of discrete nucleotide sequence clusters in entire genomes. The issue is complex as major portion of genomes of eukaryotes with increasing complexity contain variable amounts of coding as well as noncoding sequences; the latter involve a variety of repetitive sequences that act as structural signals as well as positional and functional signals within and flanking the coding regions in order to render these retrievable. Finally, there exists the extreme case of the C-value paradox illustrated by dramatic differences in the size of haploid genome of *Triturus cristatus* with 7 times more DNA than *Xenopus laevis*, although both contain the nearly identical amount of coding sequences Rosbash et al.[26]

It is reasonable to assume that, unlike the phylogenetic trees based on single traits, the topology of trees based on comparison of multiple traits would offer a consensus representation approaching relationships in classical taxonomic trees. Recently, this has been attempted by concatenation, or end-to-end ligation, of aligned nucleotide sequences of multiple genes aligned amino acid sequences to generate large polyphenic strings for comparison to construct phylogenetic trees. However, this method requires a selection of known representative sequences that are aligned in a specific order before concatenation in order to avoid low computational efficiencies in comparing long strings.[18,20,22] During past 13years, we have been constructing phylogenetic trees using a novel method that compares multiple sets of polygenic traits or parameters (e.g., MW, pI, Immuno-cross reactivity) as well as monogenic traits such as nucleotide- and amino acid sequences.[13,16–18,27] In this method, using Euclidean geometry we determine all pairs distances for a consortium of at least three traits/parameters, such as3 mitochondrial polypeptides for a set of 74 eukaryotes with emphasis on the phylogeny of mammals and protochordates,[18] to construct a phylogenetic tree that can be visualized in either 2- or 3-dimensional space. More recently, we have constructed phylogenetic trees by comparing and 15 aminoacyl tRNA synthetase sequences from 119 prokaryotes to achieve a polygenic 'consensus' topological representation of phylogeny that near-parallels the classical polygenic taxonomic trees.[27] Indeed, comparing trees for individual tRNA synthetases, rDNA alone, a *consensus* tree for 15 aminoacyl-tRNA synthetase sequences and the classical taxonomic tree, we found that the consensus tree for 15 synthetases is the closest to the classical taxonomic tree while trees for individual tRNA synthetase or 16s rDNA exhibited substantial differences in the tree topologies at the level of clades of families and even genera.[27] It is in this context, that we have developed a method that carries out clade-by-clade comparison of uniparamettric or multiparametric phylogenetic trees with classical taxonomic trees as benchmarks. Here, we describe the method that allows assessment of the relative closeness between a phylogenetic tree and taxonomic tree for the same species, based on a clustering algorithm for Taxonomic Fidelity (F).

Methods results and discussion

Estimating taxonomic fidelity of phylogenetic trees

Phylogenetic trees constructed using different parameters differ substantially in their topologies. Therefore, it is necessary to validate the fidelity of clades (a group consisting of an organism/ancestor and all its descendants) in a phylogenetic tree against a known classification scheme such as taxonomy. Here, a Taxonomic Clade is the group of species consisting of an ancestor and its descendant/s from an established taxonomy, while a Phylogenetic Clade is the group of species from a phylogenetic tree. Taxonomic fidelity of a phylogenetic tree should reflect the extent of topological similarity with the corresponding taxonomic tree. The taxonomic fidelity is estimated using the equation

$$F=z/(x+y-z)$$

where, F-the fidelity, **z**-number of species common to the Taxonomic clade/s and the corresponding Phylogenetic Clade/s, **x**-number of species in "Taxonomic Clade" and y-number of species in "Phylogenetic Clade". There are three possible cases one can expect when these two trees are compared.

Case I-identical

When the clades and tree topology in both phylogenetic tree and the corresponding benchmark taxonomic tree are identical, obtain the representation as In Figure 1. Here, the members of mammalian clade/s, Human, Monkey, Rat and Mouse are identical in both phylogenetic tree and benchmark taxonomic tree. Therefore applying the number of species in equation 1 we get the following results for mammalian clade

- Number of mammalian species in taxonomic tree x=4
- Number of mammalian species in phylogenetic tree y=4
- Number of common species z=4
- Fidelity, F= z=(x +y-z)=4=(4 + 4 -4)=1

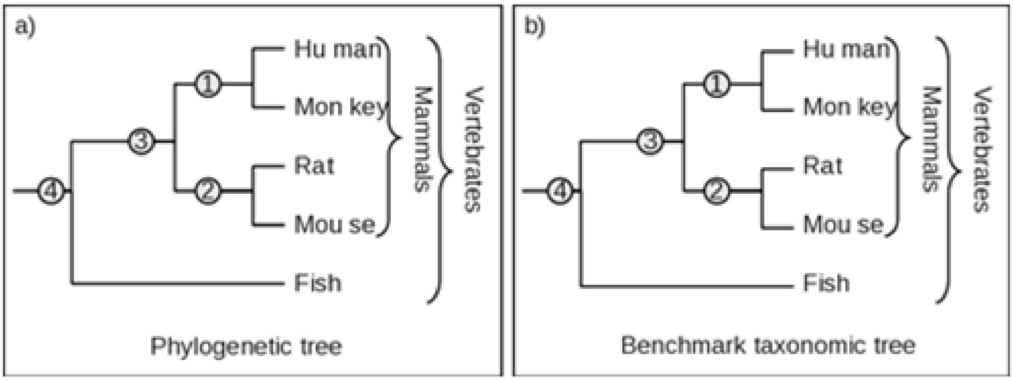

Figure 1 A phylogenetic tree of vertebrates that is identical to benchmark taxonomic tree.

Case II–missing

One or more species from the phylogenetic tree are missing. For example, as seen in Figure 2, the rat is missing from the mammalian clade in the phylogenetic tree and has been displaced elsewhere or associated with altogether different clade. We therefore estimate the Fidelity F as follows

- Number of mammalian species in taxonomic tree x=4
- Number of mammalian species in phylogenetic tree y=3
- Number of common species z=3
- Fidelity, F=z=(x + y -z)=3=(4 + 3 -4)=0.75

Case III–additional

One or more species have been added to a clade in the phylogenetic tree but absent in the corresponding clade in the taxonomic tree. As shown in Figure 3, the taxonomic clade number 4 contains four species, while the corresponding clade in the phylogenetic tree has the pig in addition to other four mammals. Therefore for the estimation of Fidelity F, is based on

- Number of mammalian species in taxonomic tree x=4
- Number of mammalian species in phylogenetic tree y=5
- Number of common species z=4
- Therefore, fidelity, F=z=(x + y - z)=3=(4 + 5 - 4)=0.80

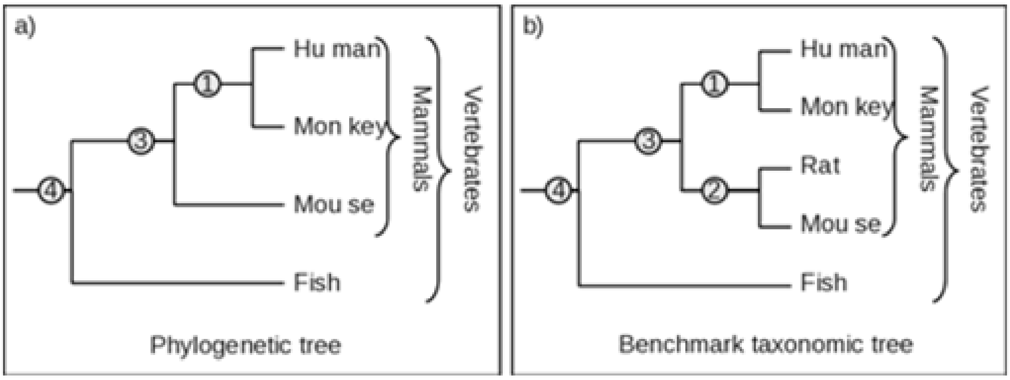

Figure 2 Phylogenetic tree of vertebrates and benchmark taxonomic tree. Here the rat is missing from mammalian clade in phylogenetic tree.

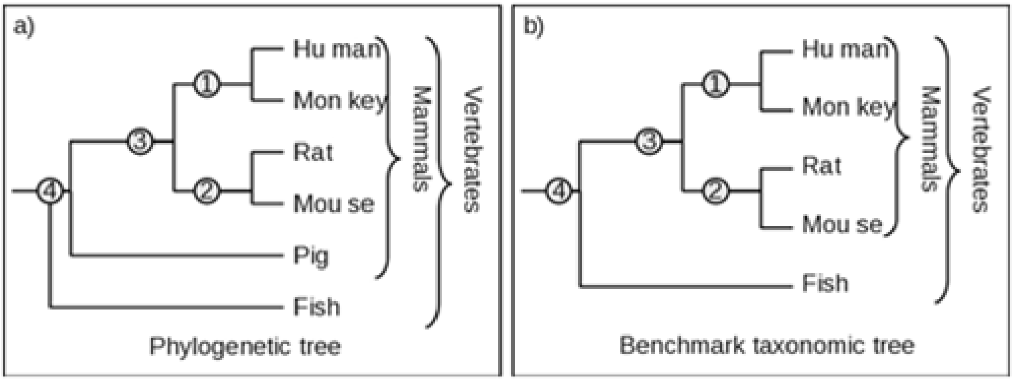

Figure 3 Phylogenetic tree of vertebrates and benchmark taxonomic tree. Note that pig is additional in mammalian clade of phylogenetic tree.

Implementation

1. All possible taxa (species associated with an internal node consist of all Operational Taxonomic Units/terminal nodes) are listed from a phylogenetic tree to test the taxonomic fidelity. For example in the phylogenetic trees in Figure 4 the possible taxa for the internal nodes are (a, b), (c, e), (a, b, c, d) and (a, b, c, d, e).
2. Select taxa from the benchmark taxonomic tree. For example we want to test the fidelity of the taxa a and b
3. With each listed taxa from phylogenetic tree estimate fidelity scores against selected taxa from the benchmark tree (Figure 4B). Here, the taxa designated in bold letters are from taxonomic tree and the remaining are from phylogenetic tree

   i. (a, b), (a, b); F=1

   ii. (a, b), (c, e); F=0

   iii. (a, b), (a, b, c, d); F=0.5

   iv. (a, b), (a, b, c, d, e); F=0.4

Thus, the maximum expected fidelity F=1 when all taxa in a given phylogenetic clade are at similar or identical position in the corresponding taxonomic clade.

4. Assign this fidelity score to the selected taxa in step 2 from benchmark taxonomic tree
5. Select the next taxa from benchmark taxonomic tree, e.g. (c, d)
6. Perform steps 3 and 4
7. Repeat step 6 till all taxa are accounted for in the benchmark taxonomic tree
8. Sum the fidelity scores obtained for all taxa
9. Assign this sum as total fidelity score of the entire phylogenetic tree.

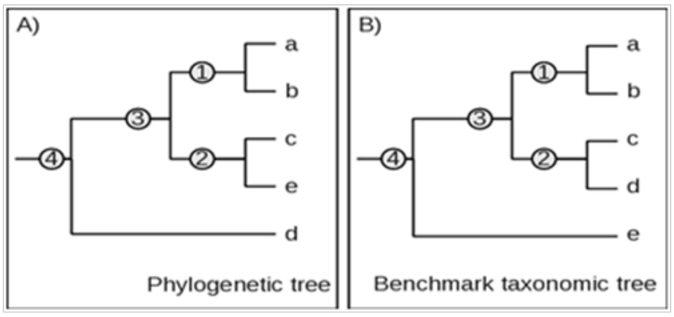

Figure 4 Phylogenetic tree & benchmark Taxonomic tree for five Operational Taxonomic Units a, b, c, d, e.

The maximum score any phylogenetic tree can obtain is the same as the number of possible taxa in benchmark taxonomic tree. Therefore, greater the total taxonomic fidelity score of a phylogenetic tree, closer it is in the topology to the taxonomic tree.

Acknowledgements

None.

Conflict of interest

The author declares no conflict of interest.

References

1. Darwin C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray; 1859. p. 1–446.
2. Mayr E. *Populations, species, and evolution: an abridgment of Animal species and evolution*. USA: Belknap Press of Harvard University Press; 1970. p. 1–453.
3. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, editors. *Evolving Genes and Proteins*. USA: Academic Press; 1965. p. 97–166.
4. Fitch WM, Margoliash E. Construction of phylogenetic trees. *Science*. 1967;155(3760):279–284.
5. Goodman M, Moore, GW. Darwinian evolution in the genealogy of haemoglobin. *Nature*. 1975;53(5493):603–608.
6. King MC and Wilson AC. Evolution at two levels in humans and chimpanzees. *Science*. 1975;188:107–116.
7. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA*. 1977;74(11):5088–5090.
8. Woese C. The universal ancestor. *Proc Natl Acad Sci USA*. 1998;95(12):6854–6859.
9. Mayr E. *The species problem*. USA: Arno Press; 1974.
10. Hennig W. Phylogenetic Systematics. *Annu Rev Entomol*. 1965;10:97–116.
11. Hennig W. Cladistic Analysis or Cladistic Classification?" A Reply to Ernst Mayr. *Syst Zool*. 1975;24(2):244–256.
12. Mount DW. *Bioinformatics: sequence and genome analysis*. USA: CSHL Press; 2004.
13. Modak SP, Milner Kumar, Bargaje R. Molecular Phylogenetic Trees: Topology of multiparametric poly–Genic/Phenic tree exhibits higher taxonomic fidelity than uniparametric trees for mono–Genic/Phenic traits. *Evolutionary Biology: Mechanisms and Trends*. 2012:79–101.
14. Federhen S. The NCBI Taxonomy database. *Nucleic Acids Res*. 2012;40:D136–D143.
15. http://www.ncbi.nlm.nih.gov/nuccore/rDNA
16. Milner M, Patwardhan V, Bansode A, et al. Constructing 3Dphylogenetic trees. *Curr Sci*. 2003;85:1471–1478.
17. Milner M Bansode AG, Lawrence AL, Nevagi SA, et al. Molecular Phylogeny in 3–D. *Curr Issues Mol Biol*. 2004;6:189–200.
18. Milner Kumar M. *Multiparametric molecular phylogenetic trees in 3D*. Ph.D. thesis. India: Department of Zoology, Karnatak University; 2009.
19. Gadagkar SR, Rosenberg MS, Kumar S. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J Exp Zool B Mol Dev Evol B*. 2005;304(1):64–74.
20. Blair JE, Hedges SB. Molecular phylogeny and divergence times of deuterostome animals. *Mol Biol Evol*. 2005;22(11):2275–2284.
21. Brocchieri L. Phylogenetic inferences from molecular sequences: review and critique. *Theor Popul Biol*. 2001;59(1):27–40.
22. Delsuc F, Brinkmann H, Chourrout D, et al. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*. 2006;439(7079):965–968.
23. Boake CR, Arnold SJ, Breden F, et al. Genetic tools for studying adaptation and the evolution of behavior. *Am Nat*. 2002;160 Suppl 6:S143–S159.
24. Deschavanne PJ, Giron A, Vilain J, et al. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol*. 1999;16(10):1391–1399.
25. Fertil B, Massin M, Lespinats S, et al. GENSTYLE: exploration and analysis of DNA sequences with genomic signature. *Nucleic Acids Res*. 2005;33(Web Server issue):W512–W515.
26. Rosbash M, Ford PJ, Bishop JO. Analysis of the C–value paradox by molecular hybridization. *Proc Natl Acad Sci USA*. 1974;71(9):3746–3750.
27. Bargaje R, Milner Kumar M, Modak SP. Consensus Phylogenetic trees of Fifteen Prokaryotic Aminoacyl– tRNA synthetase polypeptides based on Euclidean Geometry of All–Pairs Distances and Concatenation. *Bio Rxiv*. 2016:2–44.

Citation: Kumar M, Modak SP. Assessing clade wise concordance between phylogenetic trees and corresponding taxonomic trees. *MOJ Proteomics Bioinform*. 2016;4(1):174–177. DOI: 10.15406/mojpb.2016.04.00111