Ideas in Ecology and Evolution 7: 3–7, 2014
doi:10.4033/iee.2014.7.2.n

Received 18 November 2013; Accepted 18 February 2014

New Idea

Fishing for significance in phylogenies: too many alternatives for the same outcome, or an appeal to Journal editors

Jorge O. Chiapella, Joseph C. Kuhl, Pablos H. Demaio, and Leonardo D. Amarilla

Jorge O. Chiapella (jchiapella@imbiv.unc.edu.ar), Sistemática y Filogeografía de Plantas, Instituto Multidisciplinario de Biología Vegetal (IMBIV-CONICET-Universidad Nacional de Córdoba), Velez Sarsfield 1611, X5016GCA Córdoba, Argentina

Joseph C. Kuhl (jkuhl@uidaho.edu),Department of Plant, Soil, and Entomological Sciences, University of Idaho, Moscow, ID 83844-2339

Pablos H. Demaio (phdemaio@gmail.com), Sistemática y Filogeografía de Plantas, Instituto Multidisciplinario de Biología Vegetal (IMBIV-CONICET-Universidad Nacional de Córdoba), Velez Sarsfield 1611, X5016GCA Córdoba, Argentina

Leonardo D. Amarilla (amarillaleonardo@conicet.gov.ar), Sistemática y Filogeografía de Plantas, Instituto Multidisciplinario de Biología Vegetal (IMBIV-CONICET-Universidad Nacional de Córdoba), Velez Sarsfield 1611, X5016GCA Córdoba, Argentina

Abstract

The ever increasing number of computer programs developed for phylogenetic research does not necessarily facilitate the construction of biologically relevant phylogenies. Regardless of the algorithm utilized by new software, the vast majority result in treelike graphs. We suggest that a new, more inclusive framework for phylogenetic studies needs to be developed, which includes trees as an alternative in the absence of conflicting signals in the sequence data set. Conflicts are caused by noisy phylogenetic signal deriving from hybridization, allopolyploidy and lateral gene transfer—biological processes that undermine the construction of simple dichotomic bifurcating graphs. A robust framework for determining biologically relevant phylogenetic relationships should include quality analysis of the phylogenetic signal, a thorough determination of homology, analyses for phylogenetic networks, and exploration of the data for character or tree conflicts.

Keywords: trees, phylogenetic networks, phylogenetic signal, noise, new framework

We live in a world that offers a huge number of choices in almost every single aspect of life. In a book that examines how we make choices, Iyengar (2011: 207) writes: "To begin with, we have to change our attitudes toward choice, recognizing that it is not an unconditional good. We must respect the constraints on our cognitive abilities and resources that prevent us from fully exploring complex choices and stop blaming ourselves for not finding the very best option every time."

In science, as in life, having too many alternatives can be as challenging as having a single one. Consider a standard plant phylogenetic study based on molecular data. Multiple sequence alignments attempt to identify homology in a set of three or more sequences. Nearly all the available programs (23 packages, Felsenstein 2011) work mainly based on hierarchical clustering algorithms that first obtain an alignment of the most similar sequences and add progressively less similar sequences in each iteration (Koonin and Galperin 2004, Notredame 2007). Therefore, although there are numerous programs to choose from, there are relatively few algorithms that are significantly different.

Once the alignment process is completed, you have an aligned matrix and can begin the search for trees. You may choose between parsimony (45 available programs, Felsenstein, 2011), distance (73 programs), likelihood (93 programs) and Bayesian (26 programs) methods. In Bayesian and likelihood approaches you must also choose the evolutionary model, for which other programs are available (14 programs, Felsenstein 2011). Considering the pressure to get results worth publishing, having so many alternatives might be an advantage. The existence of the "publication bias" suggests positive results are more likely to be published (Sterling 1959, Boulesteix 2010, Szapkowicz 2010), which could push researchers to explore multiple methodological alternatives until a positive result is found. To avoid negative results (i.e. an unresolved phylogeny), it may be more profitable to use some of the many options identified above, to tinker with data until a positive result (i.e. a resolved phylogeny) is finally identified, rather than analyzing the quality of the data and verifying if a tree model is adequate. The circumstance in which intensive optimization of a data set yields a positive result has been called "*fishing for significance*" (Boulesteix 2010); derived from bioinformatics research, it means that the researcher searches (or fishes) for results that are the product of intensive optimization or adaptation of a new algorithm to a given dataset. It is typically difficult to reproduce such results, which should therefore be considered a weak representation of biological reality and could be considered unreliable or even false (Ioannidis 2005).

The situation in phylogenetic research is comparable to bioinformatics in that researchers may fish for phylogenetic values of a particular group using different approaches (parsimony, likelihood, or Bayesian). The result in the end is a positive result; this is a statistically significant supported topology. But this situation could be interpreted as "Corollary 4" in Ioannidis (2005: 698): "The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true." Furthermore, statistical significance is not always coupled with biological relevance, as shown in several examples analyzed by Wägele and Mayer (2007); topologies can show good support values but with little biological meaning, because of conflicts in the raw data (Wägele and Mayer 2007). The conflicts arise mainly, but not only, from long-branch effects, which are caused by selection of taxa and noise, the latter defined as the opposite of phylogenetic signal. Noise stems from random variations in the base composition of sequences, whereas the phylogenetic signal is defined as identifiable, heritable, homologous character states (Wägele and Mayer 2007). Phylogenetic signal is a desirable feature of a dataset, while noise is not. The quality of the information contained in the multiple alignment dataset is thus a crucial fact that is not usually evaluated before tree construction (Wägele and Mayer 2007).

Regarding tree construction, even the most theoretically reliable approach (Bayesian inference) is not free from criticism, due to the tendency to overestimate support values (Rokas et al. 2003, Simmons et al. 2004, Randle et al. 2005). Developed at nearly the same time as Bayesian methods, phylogenetic networks are based on concepts by Bandelt and Dress (1992) and Bandelt (1994), and show phylogenetic relationships in non-treelike graphs when the phylogenetic signal is affected by issues of hybridization, recombination, or horizontal gene transfer. However, the utility of phylogenetic networks has been reduced to not much more than a tool to detect conflicts in the data set (Vriesendorp and Bakker 2005), despite the fact that they can better reflect phylogenetic relationships in situations where conflicting data sets would result in weakly supported trees (Bapteste et al. 2013).

At this point is worthwhile to recall, "simple dichotomous branching diagrams cannot do justice to the real world of higher plants phylogeny" (Stuessy 1997: 115). This expression renders the thought, "[o]nce an alignment method process is completed, you have an aligned matrix and you can start the search for trees," somehow misleading (Stuessy 1997: 115). We don't have to search for trees, although the "tree-thinking" paradigm (de Queiroz 1988, O'Hara 1998) has considered species evolution only in a phylogenetic context as part of a tree. The search for phylogenetic relationships has to be independent of the outcome. But the many issues affecting tree building pose a profound confounding effect with the methods resulting in treelike graphs, in plants it is hybridization and allopolyploidy, while in bacteria or fungi it is lateral gene transfer. Contradictory trees telling two different evolutionary histories (with 100% bootstrap support) result not only from different programs applied to the same data set, but from the same data set and the same program but different settings (Philipps et al. 2003). However, this is not a call for abandoning the use of trees as a metaphor in phylogenies (Morrison *unpublished*). Trees should be seen as one of the alternatives of phylogenetic analysis, not the mandatory result. The search for a well-supported tree has become a goal in itself, instead of the search for a biologically-sound and plausible evolutionary history, encompassing all aspects of plant biology: molecular, cytological, morphology, ecology, and geographic. This probably will not happen unless Journal editors start accepting that it is no longer possible to ignore the abundant evidence on issues like hybridization, allopolyploidy, and lateral gene transfer that undermine the simple dichotomist tree concept.

New tree-based software increases the already numerous alternatives of the dominant paradigm in phylo-

genetic research, and at this point more alter-natives does not necessarily mean progress. We believe that the numerous software alternatives provide researchers only with variations to the same end, i.e. to build a tree. But even well-supported trees can be misleading; substant-ive alternatives for the study of evolutionary relation-ships might perhaps be sought in methods that do not result in a tree. Such a conceptual framework where a tree is a possible outcome (equally possible as a net-work) and not an obligatory result is still lacking in plant phylogenetic studies. The new framework should also include quality analysis of the phylogenetic signal of the sequence alignment prior to analysis, such as SAMS (Wägele and Mayer 2007), which provides a thorough assessment of homology as proposed by Ochoterena (2009), and an "exploration" of the data as described by Morrison (2010), that allows the detection of character or tree conflicts in a data set.

Acknowledgments

We thank the CONICET (Consejo Nacional de Investi-gaciones Científcas y Tecnológicas) of Argentina for continued financial support and David A. Morrison (SWEPAR, Swedish University of Agricultural Sci-ences, Uppsala) for discussion and for sharing his unpublished manuscript.

Referees

L.J.J. van Iersel – l.j.j.v.iersel@gmail.com
Centrum Wiskunde & Informatica

Root Gorelick – root_gorelick@carleton.ca
Carleton University

References

Bandelt, H-J. and A.W.M. Dress. 1992. Split Decom-position: A new and useful approach to phylo-genetic analysis of distance data. Molecular Phylo-genetics and Evolution 1:242–252. CrossRef

Bandelt, H-J. 1994. Phylogenetic networks. Verhand-lungen Naturwissenchaft Vereins Hamburg (NF) 34:51–71.

Bapteste, E., van Iersel, L., Janke, A., Kelchner, S., Kelk, S., McInerney, J.O., Morrison, D.A., Nakhleh, L., Steel, M., Stougie, L. and J. Whitfield. 2013. Net-works: expanding evolutionary thinking. Trends in Genetics 29:439–441. CrossRef

Boulesteix, A.-L. 2010. Over-optimism in bioinfor-matics research. Bioinformatics 26:437–439. CrossRef

De Queiroz, K. 1988. Systematics and the Darwinian revolution. Philosophy of Science 55: 238–259. CrossRef

Felsenstein, J. 2011. http://evolution.genetics.washing ton.edu/phylip/software.html (accessed September 13, 2011).

Ioannidis, J.P.A. 2005. Why most published research findings are false. PLoS Medicine 2: e124. CrossRef

Iyengar, S. 2011. The art of choosing. Twelve, Hachette Book Group, New York.

Koonin, E.V. and M.Y. Galperin. 2004. Principles and Methods of Sequence Analysis. Pages 111–192, *in* Koonin, E.V. and M.Y. Galperin, editors. Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Kluwer Academic, Boston.

Morrison, D.A. 2010. Using data-display networks for exploratory data analysis in phylogenetic studies. Molecular Biology and Evolution 27:1044–1057. CrossRef

Morrison, D.A. (unpublished manuscript). Is the tree of life the best metaphor, model or heuristics for phylogenetics?

Notredame, C. 2007. Recent evolutions of multiple sequence alignment algorithms. PLoS computational biology 3:e123. CrossRef

Ochoterena, H. 2009. Homology in coding and non-coding DNA sequences: a parsimony perspective. Plant Systematics and Evolution 282:151–168. CrossRef

O'Hara, R.J. 1998. Population thinking and tree think-ing in systematics. Zoologica Scripta 26:323–329. CrossRef

Phillips, M.J., Delsuc, F. and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Molecular Biology and Evolution 21:1455–1458. CrossRef

Randle, C.P., Mort, M.E. and D.J. Crawford. 2005. Bayesian inference of phylogenetics revisited: devel-opments and concerns. Taxon 54:9–15. CrossRef

Rokas, A., Williams, B.L., King, N. and S.B. Carroll. 2003. Genome-scale approaches to resolving incon-gruence in molecular phylogenies. Nature 425:798–804. CrossRef

Simmons, M.P., Pickett, K.M. and M. Miya. 2004. How meaningful are Bayesian support values? Molecular Biology and Evolution 21:188–199. CrossRef

Sterling, T.D. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. Journal American Stat-istics Association 54:30–34.

Stuessy, T.F. 1997. Classification: more than just branching patterns of evolution. Aliso 15:113–124.

Szpakowicz, S. 2010. Failure is an orphan (let's adopt). Computational Linguistics 36:157–158. CrossRef

Vriesendorp, B. and F.T. Bakker. 2005. Reconstructing patterns of reticulate evolution in angiosperms: what can we do? Taxon 54:593–604. CrossRef

Wägele, J.W. and C. Mayer. 2007. Visualizing differ-ences in phylogenetic information content of align-

ments and distinction of three classes of long-branch effects. BMC Evolutionary Biology 7:147. CrossRef

Response to referee

We agree with Gorelick (2014) in that the dependence of displaying phylogenetic relationships only as trees is a zoocentric relict of Mayr's biological species concept, but it is also a consequence of the success of another zoologist's (Willi Hennig) methodological development. Cladistics, probably the most complete framework to study evolutionary relationships ever developed, provides concrete ways to handle characters and describe relationships, as well as how to define taxonomic boundaries. A sort of 'all-in-one' methodology to produce clean and reproducible results, cladistics has firmly engrained in the collective consciousness of evolutionary biologists and taxonomists (as well as scientists working in related fields such as genetics or biogeography) the concept that only a dichotomic tree can depict evolutionary relationships. If something goes wrong in the process of tree building, the burden must be a consequence of problems in the data, the organisms being studied, or even the capacity of the researcher to adequately build trees (as suggested recently by Anisimova 2013 and Anisimova et al. 2013), but never in the concept of a tree itself. Trees are still regarded by many (including many top-ranked journal editors) as the only way to depict evolutionary relationships. This is one point we would like to now emphasize—the attachment of journal editors to the tree paradigm. The application of networks to provide an alternative depiction of evolutionary relationships has been treated in detail (Than et al. 2008, Huson and Scornavacca 2012, Bapteste et al. 2013). Furthermore, these authors, and numerous others concur that common biological issues such as hybridization and reticulation cannot be adequately depicted by trees. Therefore, the attachment of journal editors to a paradigm which can produce anomalous results (in the form of conflicting or unresolved topologies) is counterproductive. These same journal editors, when presented with alternative and new analytical developments, resist including such advances in their author guidelines. Kuhn (1996: 151–152), in his seminal work, refers to such resistance "[t]he transfer of allegiance from paradigm to paradigm is a conversion experience that cannot be forced," adding, "[t]he source of resistance is the assurance that the older paradigm will ultimately solve all problems, that nature can be shoved into the box the paradigm provides."

We believe that evolutionary relationships are determined by the data, and that some relationships are best represented as trees, and others as networks. The deterministic attitude of subscribing automatically to a tree model presumes all data are perfectly segregating and binary (i.e. adequate) and therefore the resulting tree in a phylogenetic study will be biologically meaningful. However, if the organisms have experienced normal biological processes such as hybrid speciation, polyploidization, horizontal gene transfer or similar, the data will not be perfectly binary (i.e. not adequate), and the resulting trees will be poorly resolved or not resolved at all and the phylogenetic study will produce inconclusive results, in the form of several equally well-supported trees but with conflicting topologies.

Gorelick (2014) points out that "tree topologies are convenient" and that "practicing biologists seem too wedded to the outdated Popperian philosophy of naïve ('dogmatic') falsificationism." We also believe that choosing to be naïve provides biologists flexibility to reject and accept hypotheses (topologies) to meet their own agendas, i.e. fishing (Chiapella et al. 2014). Paraphrasing Groucho Marx, 'those are my principles (trees), and if you don't like them… well, I have others.'

Regarding phylogenetic signal, the subject has been discussed at length by Wägele and Mayer (2007), who analyzed several phylogenies with large datasets and robust supporting values for topologies, finding conflicting results in relation to earlier analysis (including morphology). Wägele and Mayer (2007) proposed a novel algorithm to analyze the quality of the phylogenetic signal contained in the data set *prior* to the building of trees. Signal was defined as identifiable homologous character states, while noise is made up of randomly distributed substitutions, including paralogus sequences (Wägele and Mayer 2007). The software SAMS (developed by C. Mayer) yields a graph similar to a spectral plot (Lento et al. 1995) providing unambiguous differentiation between phylogenetic signal and noise. In cases where reticulate evolutionary events may have occurred, network analysis (Huson and Scornavacca 2008, Than et al. 2012) will provide a better description of evolutionary relationships, and offer valuable tools contributing to the formulation of new hypotheses to explain discordance (Bapteste et al. 2013).

Anisimova, M. 2013. Tree-building: Maria Anisimova on pitfalls and solutions in phylogenetic analysis http://www.biomedcentral.com/biome/tree-building-maria-anisimova-on-pitfalls-and-solutions-in-phylogenetic-analysis/ Posted by Biome on 25th November 2013.

Anisimova, M., Liberles, D. A., Philippe, H., Provan, J., Pupko, T., and A. von Haeseler. 2013. State-of the art methodologies dictate new standards for phylogenetic analysis. BMC evolutionary biology 13:1–8. CrossRef

Bapteste, E., van Iersel, L., Janke, A., Kelchner, S., Kelk, S., McInerney, J.O., Morrison, D.A., Nakhleh, L., Steel, M., Stougie, L. and J. Whitfield. 2013. Networks: expanding evolutionary thinking. Trends in Genetics 29: 439–441. CrossRef

Gorelick, R. 2014. Fishing for philosophical phylogenyetic foibles. Ideas in Ecology and Evolution 7:8–10. CrossRef

Huson, D.H. and C. Scornavacca. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Systematic Biology 61:1061–1067. CrossRef

Kuhn, T.S. 1996. The structure of scientific revolutions. 3rd edition. The University of Chicago Press, Chicago and London. CrossRef

Lento, G.M., Hickson, R.E., Chambers, G.K. and D. Penny. 1995. Use of spectral analysis to test hypotheses on the origin of pinnipeds. Molecular Biology and Evolution 12: 28–52. CrossRef

Than, C., Ruths, D. and L. Nakhleh. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics 9: 322. CrossRef

Wägele, J.W. and C. Mayer. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. BMC Evolutionary Biology 7: 147. CrossRef