Ideas in Ecology and Evolution 7: 8–10, 2014
doi:10.4033/iee.2014.7.3.c

Received 19 December 2013; Accepted 18 February 2014

Commentary

Fishing for philosophical phylogenetic foibles

Root Gorelick

Root Gorelick (Root.Gorelick@carleton.ca), Department of Biology, School of Mathematics & Statistics, and Institute of Interdisciplinary Studies, Carleton University, 1125 Colonel By, Ottawa, Ontario, CANADA K1S 5B6

Jorge Chiapella and colleagues (2014) do a stellar job of applying John Ioannidis's (2005) famous paper to the problem of phylogenetic inference. They highlight how reticulation in the form of either inter-specific hybridization or lateral gene transfer makes a mockery of analyses that seek to find the most appropriate tree topology. They also highlight how often these problems arise from editorial decisions, rather than necessarily from researchers' beliefs. But, if anything, Chiapella et al. were too generous regarding the status quo. Therefore, instead of criticizing their paper, I want to extend and generalize their critique.

Our dependence on tree topologies is probably an anachronistic, zoocentric relict of Ernst Mayr's (1942) horrendously named biological species concept. (Are other species concepts really not biological?) The biological species concept posits that, as soon as populations become sufficiently reproductively isolated, they never again exchange genomes. For fine-scales, albeit much broader scales than Mayr had envisioned, tree topologies make no sense because of reticulation due to rampant hybridization. For example, see the remarkable number of inter-generic orchid hybrids in the "Quarterly Supplement to the International Register of Orchid Hybrids (Sander's List)" (http://www.rhs.org.uk/Plants/RHS-Publications/Orchid-hybrid-lists). At very course scales, such as all eukaryotes, tree topologies make no sense because of reticulation involved in endosymbiosis (Margulis and Sagan 1986, 1988). By counting the number of membranes around organelles in eukaryotes, it is obvious that there have been many independent secondary and tertiary origins of chloroplasts in chromalveolates (including both the SAR group and Cryptophyta) and excavates such as euglenids (McManus and Qiu 2008). Secondary origins of chloroplasts mean that these lineages incorporated eukaryotic autotrophs into their cells, in lieu of eubacterial autotrophs. Thus, the only possible hope for tree topologies being sensible is at intermediate taxonomic levels, but even here there seems to be too much genomic chimaerism due to lateral gene transfers (Katz 2002). Tree topologies are convenient and may be all that we can computationally manage at this juncture (although see Huson et al. 2005, Baroni et al. 2006, Baroni and Steel 2006, etc.), which is why tree topologies are so pervasive, but that does not mean they reflect reality.

For many green plants, phylogenic trees or more general phylogenetic graphs suffer from another unfortunate convenience, namely exclusive use of chloroplast markers in their construction. Chloroplast markers cannot tell us much about species-level or even family-level phylogenies in many green plants because of flexible chloroplast inheritance (Corriveau and Coleman 1988). Many studies assume strict maternal inheritance of chloroplasts when constructing phylogenic trees, even when the authors know this assumption is incorrect. For example, in the cactus family (Cactaceae), Corriveau and Coleman (1988) showed that chloroplasts are sometimes biparentally inherited. Yet a quarter century later, phylogenetic trees are being published based on the assumption of strict maternal inheritance of chloroplast DNA. Thus it seems unsurprising that Schlumpberger and Renner (2012) found the congeneric *Espostoa lanata* and *E. guentheri* to be only distantly related, even though there is no way to tell whether this reflects reality (presaging discussion of ontology) or merely a confounding assumption, made strictly for convenience, that chloroplasts are maternally inherited.

Fishing is the right metaphor. Say you have an unknown fish species that you are trying to identify by comparing it with various fish for which you know species names. If the known species are only the ones in your home fish tank, then the closest possible match is unlikely to provide much biologically useful

information. This is akin to searching for the best phylogenetic tree when more general phylogenetic graphs better describe evolutionary patterns.

Fishing for phylogenetic significance is a classic example of making the best of a bad situation because hypothesized models may poorly reflect reality. Fishing is a much bigger problem than just with phylogenetic analysis. This problem plagues all Fisherian statistics and *ad hoc* model selection, such as with the Akaike Information Criterion (AIC). The curious thing is that many contemporary scientists are devoutly Fisherian statisticians, even though practitioners in many fields (especially operations research) have abandoned this approach for a more Bayesian perspective. Note that Bayesian methods in phylogenetics, such as Markov chain Monte Carlo (MCMC) methods contain some Bayesian mechanics, but are not Bayesian at heart (see the final sentence of the next paragraph for my broad definition of 'Bayesian').

Chiapella et al. (2014) state that, "[p]hylogenetic signal is a desirable feature of a dataset, while noise is not." The problem is discerning what is signal and what is noise. Usually this is model-dependent, i.e. noise is anything not explained by the model, aka the residuals. By considering alternative models, such as reticulation, noise can be artificially converted into signal. With over-modeling (think of adjusted R-squared values), this might not necessarily be a benefit in ascertaining biological reality. As Nate Silver (2012) emphasized, the trick is to appropriately apportion data between signal and noise, which is an utterly non-trivial task, chocked full of both type I and type II errors. Evolutionary biology in particular and science in general should be Bayesian in the broad sense: new data should be used to update hypotheses, such as best-fit phylogenetic trees and graphs, as well as update our notions of what constitutes signal and noise.

Bioinformatics researchers study ontologies, but seem largely uninterested in deciding which entities are real, even though this is the classic purview of ontology. By contrast and refreshingly, the editor of the *Biological Journal of the Linnean Society*, in their instructions to authors, seems acutely aware of ontology, stating that there is only one real phylogeny, therefore instructing authors to use phrases like 'phylogenetic hypothesis' in lieu of 'phylogeny'. Practicing biologists seem too wedded to the outdated Popperian philosophy of naïve ('dogmatic') falsificationism, abandoning reasonable hypotheses with even a scintilla of falsification (Faith 2004), but retaining seemingly unreasonable hypotheses (e.g. theoretically unjustified or logically inconsistent hypotheses) when there is insufficient data to reject them. Bayesian perspectives are much more like Imre Lakatos' (1970) iterative framework for making progress in science. Possibly a little more radically, Chiapella et al.'s (2014) critique resonates with Paul Feyerabend's (1975) anarchist philosophical framework in "Against Method" or maybe the more pragmatic approach of John Law's (2004) "After Method".

Chiapella et al. (2014) state that, "topologies can show good support values but with little biological meaning, because of conflicts in the raw data." This actually constitutes two problems. First, 'big data' obfuscates the fact that not all data are created equally. Some data points have more phylogenetic signal than others. Outliers and inconsistencies contain more information than long recurring patterns (Gorelick 2013). Second, not all theories are created equally. Some are more powerful, parsimonious, and aesthetic than others.

I applaud Chiapella et al. (2014) for cajoling us to think outside of the box with regards to phylogenetic analyses, encouraging us to do more than what is merely convenient, but instead focus on what is biologically appropriate, even if we cannot yet implement the mathematics. Furthermore, Chiapella and colleagues provide a fascinating perspective that the problems are not entirely with researchers, but also with editorial juggernauts, where the default position is to publish the best phylogenetic tree even if a tree topology does not reflect biological reality or does not reflect currently accepted biological theory in the face of lateral gene transfer and endosymbiosis. I do not wish to denigrate the entire field of phylogenetics, which has been fantastic at letting us conceptualize evolutionary relationships, making mathematically precise the notions that Charles Darwin sketched (Darwin 1859, Barrett et al. 1987). But it never hurts to reiterate that phylogenetic trees and more general phylogenetic graphs are nothing more than hypotheses (Dominguez and Wheeler 1997, Harlin 1998), hypotheses that are often constrained by our mathematics, computing power, theories, and imagination.

Acknowledgments

This work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Baroni, M., Semple, C., and M. Steel. 2006. Hybrids in real time. Systematic Biology 55:46–56. CrossRef

Baroni, M. and M. Steel. 2006. Accumulation phylogenies. Annals of Combinatorics 10: 19–30. CrossRef

Barrett, P.H., Gautrey, P.J., Herbert, S., Kohn, D., and S. Smith. 1987. Charles Darwin's notebooks, 1836–1844. Cornell University Press, Ithaca.

Chiapella, J.O., Kuhl, J.C., Demaioa, P.H., and L.D. Amarillaa. 2014. Fishing for significance in

phylogenies: too many alternatives for the same outcome, or an appeal to journal editors. Ideas in Ecology and Evolution 7:3–7. CrossRef

Corriveau, J.L. and A.W. Coleman. 1988. Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species. American Journal of Botany 75: 1443–1458. CrossRef

Darwin, C.R. 1859. On the origin of species by natural selection or the preservation of favoured races in the struggle for life. John Murray, London.

Dominguez, E. and Q.D. Wheeler. 1997. Taxonomic stability is ignorance. Cladistics 13: 367–372.

Faith, D.P. 2004. From species to supertrees: Popperian corroboration and some current controversies in systematics. Australian Systematic Botany 17: 1–16. CrossRef

Feyerabend, P. 1975. Against method: outline of an anarchistic theory of knowledge. New Left Books, London.

Gorelick, R. 2013. What is natural history? CSEE Bulletin 14: 17–19.

Harlin, M. 1998. Taxonomic names and phylogenetic trees. Zoologica Scripta 27: 381–390. CrossRef

Huson, D.H., Klöpper, T., Lockhart, P.J., and M.A. Steel. 2005. Reconstruction of reticulate networks from gene trees. Pages 233–249 *in* Miyano, S., Mesirov, J., Kasif, S., Istrail, S., Pevzner, P.A., and M. Waterman, editors. Research in Computational Molecular Biology. Springer-Verlag, Berlin.

Ioannidis, J.P.A. 2005. Why most published research findings are false. PLoS Medicine 2: 696–701. CrossRef

Katz, L.A. 2002. Lateral gene transfers and the evolution of eukaryotes: theories and data. International Journal of Systematic and Evolutionary Microbiology 52: 1893–1900. CrossRef

Lakatos, I. 1970. Falsification and the methodology of scientific research programmes. Pages 91–316 *in* I. Lakatos and A. Musgrave, editors. Criticism and the growth of knowledge. Cambridge University Press, Cambridge.

Law, J. 2004. After method: mess in social science research. Routledge, London.

Margulis, L. and D. Sagan. 1986. Origins of sex: three billion years of genetic recombination. Yale University Press, New Haven.

Margulis, L. and D. Sagan. 1988. Sex: the cannibalistic legacy of primordial androgynes. Pages 23–38 *in* R. Bellig and G. Stevens, editors. The evolution of sex. Harper & Row, San Francisco.

Mayr, E. 1942. Systematics and the origin of species. Columbia University Press, New York.

McManus, H.A. and Y-L Qiu. 2008. Life cycles in major lineages of photosynthetic eukaryotes, with a special reference to the origin of land plants. Fieldiana Botany 41: 17–33. CrossRef

Schlumpberger, B.O. and S.S. Renner. 2012. Molecualr phylogenetics of *Echinopsis* (Cactaceae): Polyphyly at all levels and convergent evolution of pollination modes and growth forms. American Journal of Botany 99: 1335–1349. CrossRef

Silver, N. 2012. The signal and the noise: Why so many predictions fail – but some don't. Penguin, New York.