title

author

March 13, 2013

Abstract

Text of the Abstract

I. Introduction

Phylogenetics is the science of deriving probable evolutionary relaitonships from characteristics of clades (i.e. groups of organisms of varying size). The Open Tree of Life project is a multi-institutional, NSF-funded phylogenetic synthesis project. It is motivated by the fact that, despite many years of study, there still doesn't exist a comprehensive account of our knowledge of the evolutionary relationships of all known species. The project seeks to collect and synthesize these trees, beginning with two example taxa. This project is of interest to those of us studying the archiving, sharing, and reuse of data for several reasons. First, as a large synthesis project, it is exemplary of the kind of large scale, distrubuted scientific data projects that are increasingly shaping the availability of research data in several fields... Second, phylogenetics is a prime example of a computational science, where experements are performed by computation (*in silico*) as opposed to in physical laboratories (*in vivo*). Phylogenetics is also notable in that much of the data used to produce trees is itself the result of computations (e.g. genetic sequence alignments). This adds an added layer of complexity to the field's data reuse that presages a coming trend in science: the expansion of secondary computational anayses of computational analysis. In this way, examinations of this discipline are likely to yeild insights robust to coming changes in science.

Phylogenetics as a field also has a history of direct involvement in its own informatics (i.e. phyloinformatics). Cite these: Metadata quality Minimum information standards (**?**) Data formats (Vos et al., 2012) Data reuse

In this study we were most interested in examining scientists' perceptions of the ease and importance of metadata, to understand how these might shed light on data sharing and reuse in the field User-centered evaluation

I.1 Research Questions

There are two primary research questions in this study.

Q1 What perceptions do producers of phylogenetic trees have regarding the ease of providing metadata?

Q2 What perceptions do consumers of phylogenetic trees have regarding the importance of metadata for evaluating and reusing trees from repositories?

Regarding Q1, previous studies such as Stoltzfus et al. (2012) indicating that phylogenetic data sharing was as low as 4%, and Drew et al. (2013) estimate that only 64% of these have metadata sufficient for reuse. This led us to believe that producers' attitudes would be biased against metadata production and could be an explanation for low data submission rates and poor metadata quality. Consquently, the specific hypothesis we developed to test Q1 can be stated thus:

H1 Producers perceive most phylogenetic metadata types as difficult to produce

If true, this hypothesis would suggest that increasing the ease of metadata production and data submission to repositories is the path forward for improving the quantity and quality of phylogenetic data shared and its associated metadata. Regarding Q2, previous studies...In general, we assumed that consumers of trees would view a majority of metadata categories as critical to reuse. Stoltzfus et al. (2012) found 21 of 40 articles in a sample reused genetic sequences. While the rate of reuse of phylogenies was much lower, it's worth noting that fully 5 of the 40 studies reused trees from the same, high quality data source, the phylogeny of plant taxa maintained by the Angiosperm Phylogeny Group. This suggests that the availability of high quality, well annotated phylogenetic data will increase reuse. Our hypothesis regarding Q2 is thus:

H2 Consumers will perceive a majority of metadata types as critical to reuse

If true, this hypothesis would explain the observed low rate of phylogenetic data reuse.

II. Related Literature

Phylogenetics
  Science more broadly (Angela)
  Edge: Computational Science in silico vs in vivo automated provenance
  Methods etc User-centered evaluation
  - asdf

III. Methods

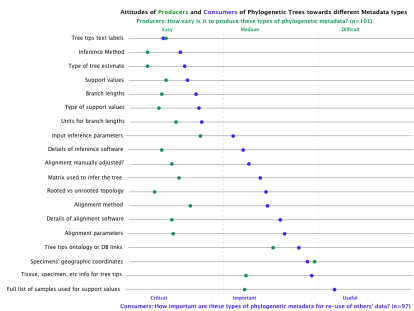## IV. Results

Figure 1

graph - wide.png

Figure 1: *Caption*

V.   Discussion

VI.   Conclusion

References

Drew, B. T., Gazis, R., Padilla, P. C., Swithers, K. S., Soltis, D. E., Hibbett, D. S., Crandal, K. A., and Katz, L. A. (2013). Data Deposition: Missing data mean holes in the Tree of Life. *Nature*, 493(7432):305.

Stoltzfus, A., O'Meara, B., Whitacre, J., Mounce, R., Gillespie, E. L., Kumar, S., Rosauer, D. F., and Vos, R. a. (2012). Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis. *BMC research notes*, 5:574.

Vos, R., Balhoff, J. P., Caravas, J., Holder, M. T., Lapp, H., Maddison, W. P., Midford, P. E., Priyam, A., Sukumaran, J., Xia, X., and Stoltzfus, A. (2012). NeXML: rich, extensible, and verifiable representation of comparative data and metadata. *Systematic biology*, (Bisby 2000).