title

author

April 29, 2013

Abstract

We surveyed 101 producers and consumers of phylogenetic data to determine their perceptions of the ease and importance of providing phylogenetic metadata when archiving results in a data repository. We found that producers labeled a majority of metadata types as easy to produce, suggesting that the production of metadata is not as much of a barrier to data sharing and metadata quality in this field as was previously thought. As more sciences use computational methods and reuse data, understanding scientists' attitudes towards computational metadata will be critical to the effective functioning of data repositories and development of data management policies.

I. Introduction

Phylogenetics is the science of inferring probable evolutionary relationships from characteristics of organisms. 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 published trees. 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, distributed 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 experiments 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 analyses of

computational analysis. In this way, examinations of this discipline are likely to yield 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 (Leebens-Mack et al., 2006)
- Data formats (Vos et al., 2012)
- Data reuse (Stoltzfus et al., 2012)

In this study we are 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

To address these questions, we surveyed researchers in the field about the importance of phylogenetic metadata. The survey questions built upon previous work compiling a checklist of metadata elements that would comprise a minimum information standard for phylogenetics (MIAPA - minimum information about a phylogenetic analysis) (Leebens-Mack et al., 2006) and `http://wiki.tdwg.org/twiki/bin/view/Phylogenetics/MIAPAWorkshop2011`. The metadata elements describe both the data that were the inputs to an analysis as well as the computational parameters used to infer the tree and summarize the results. The full list of metadata elements can be found in the survey (supplemental material / appendix).

The survey contained the following questions (paraphrased):

1. Are you a producer of trees or a user of trees, or both?
2. What is your career stage
3. What are your fields of study
4. Users of trees: categorize metadata elements as critical / important / useful for interpreting, evaluating, and using trees produced by others.
5. Producers of trees: categorize metadata elements as easy / medium / difficult to provide when archiving or databasing your trees.

Question 4 used the following definitions: CRITICAL = without this information, you would not (or could not) use a tree; IMPORTANT = it would make use much easier, or allow use with more confidence; USEFUL = could think of uses for this data, but I would use a tree without it

Question 5 used the following definitions: EASY = included in the standard input or output files for your analysis; MEDIUM = data readily available, but you would need to enter this separately from the input / output files; DIFFICULT = you would have to search for this information and / or would not generally be available for you data

The complete survey is included in the supplementary materials. We received and IRB exemption from Duke University before proceeding with the survey.

We sent the link to various mailing lists in evolutionary biology and advertised it via social media channels.

IV. Results

There were 101 respondents. Results in csv and XML form are provided in the supplemental material.

Figure 1 shows a stack of one dimensional plots of the

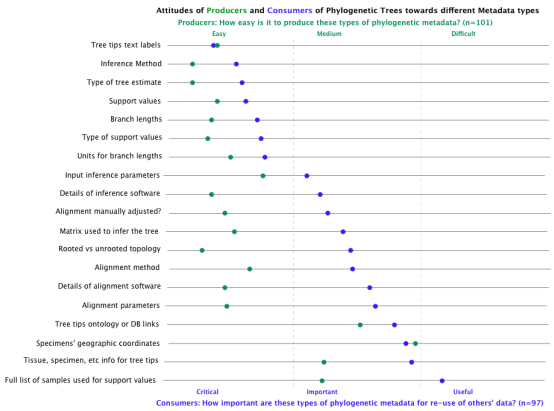

Figure 1: *Caption*

V. Discussion

We found evidence that both of our hypotheses were false.

H1: producers reported that much of the metadata we asked about was produced automatically by the programs that they use.

H2: Consumers seemed confident that they could effectively evaluate phylogenetic data with relatively few metadata types

V.1 Preliminary Verification

Our results suggest that the next step in confirming these results will be to investigate the actual data and metadata in common repositories. To inform our discussion, we undertook a small, uncontrolled examination of XXstudies in data repository Dryad Of these, we found...

V.2 Limitations

There are several limitations of our study that suggest caution in interpreting the results as we have here. First, the low n (=101)

VI. Conclusion

We surveyed 101 producers and consumers of phylogenetic data to determine their perceptions of the ease and importance of providing phylogenetic metadata when archiving results in a data repository. We found that producers labeled a majority of metadata types as easy to produce, suggesting that the production of metadata is not as much of a barrier to data sharing and metadata quality in this field as was previously thought. These results suggest that computational sciences such as phylogenetics may have different barriers to data sharing, reuse, and quality than other sciences. As more sciences use computational methods and reuse data, understanding these new needs will be critical to the effective functioning of data repositories and development of data management policies.

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.

Leebens-Mack, J., Vision, T., Brenner, E., Bowers, J. E., Cannon, S., Clement, M. J., Cunningham, C. W., DePamphilis, C. W., DeSalle, R., Doyle, J. J., Eisen, J. A., Gu, X. U. N., Harshman, J., Jansen, R. K., Kellogg, E. A., Koonin, E., Mishler, B. D., Philippe, H., Pires, J. C., Qiu, Y.-L., Rhee, S. Y., Sjölander, K., Soltis, D. E., Soltis, P. S., Stevenson, D. W., Wall, K., Warnow, T., and Zmasek, C. (2006). Taking the First Steps towards a Standard for Reporting on Phylogenies: Minimal Information about a Phylogenetic Analysis (MIAPA). *OMICS: A Journal of Integrative Biology*, 10(2):231–237.

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*, 61(4).