A generic bioinformatics pipeline to integrate large-scale trait data with large phylogenies

Conference Paper · November 2017

DOI: 10.1109/BIBM.2017.8218008

5 authors, including:

Pasan C Fernando
University of South Dakota
4 PUBLICATIONS **8** CITATIONS
SEE PROFILE

Laura M. Jackson
University of South Dakota
7 PUBLICATIONS **221** CITATIONS
SEE PROFILE

Erliang Zeng
University of Iowa
84 PUBLICATIONS **923** CITATIONS
SEE PROFILE

Paula Mabee
Battelle Memorial Institute
113 PUBLICATIONS **3,091** CITATIONS
SEE PROFILE

A generic bioinformatics pipeline to integrate large-scale trait data with large phylogenies

Pasan C. Fernando, Laura M. Jackson, Erliang Zeng, Paula M. Mabee
Biology Department
University of South Dakota
Vermillion, USA
pasan.fernando@coyotes.usd.edu,
laura.jackson@coyotes.usd.edu, erliang.zeng@usd.edu,
paula.mabee@usd.edu

James P. Balhoff
Renaissance Computing Institute
University of North Carolina
Chapel Hill, USA
balhoff@renci.org

Abstract— **Ancestral state reconstructions are used to infer the evolutionary history of phenotypic traits to understand their evolution. Current ancestral state reconstructions require integration of larger trait matrices with larger phylogenies, which introduces several challenges. We identified these challenges and developed a generic pipeline that uses the Phenoscape Knowledgebase (Phenoscape KB) to retrieve raw trait matrices and convert them to an efficient version that can be easily integrated with large phylogenies downloaded from the Open Tree of Life (Open Tree). We demonstrate the performance of this pipeline using the evolution of the pectoral and pelvic fins as the use case, which involves integration of a large-scale phylogeny containing over 38,000 taxa.**

Keywords- Ancestral state reconstruction; pipeline; evolution; phylogenetic trees; phenotypes; trait matrix.

I. INTRODUCTION

To understand evolutionary mechanisms, it is important not only to observe extant species, but also understand the characteristics of ancestral taxa, which are extinct or available only as fossils. The process of inferring unknown ancestral states based on observed states of existing species is identified as the ancestral state reconstruction [1], [2]. This usually requires a trait matrix and a phylogenetic tree. The current rate of biological data acquisition mandates performing ancestral state reconstructions in a large scale, which introduces several challenges, such as reconciling taxon names between matrices and trees and minimizing data loss. Individual steps of our pipeline address these challenges and maintain an efficient data transfer between raw data matrices retrieved from the Phenoscape Knowledgebase (Phenoscape KB) [3] and large phylogenetic trees from the Open Tree of Life (Open Tree) [4]. The output matrices of the pipeline can be easily mapped to large phylogenies, allowing the user to perform ancestral state reconstructions on larger taxonomic groups, even containing over 30,000 taxa.

II. METHODS

A. Data sources

Phenoscape KB is an efficient data resource to retrieve large trait matrices of vertebrates that are primarily based on published character matrices and some monographic treatments [5]. It uses an ontology-based system to generate supermatrices from individual trait matrices. The user can access these matrices via the OntoTrace query tool [6] by searching for the desired trait name. These trait names are equivalent to term names from the Uberon anatomy ontology [7].

Open Tree is a popular data source to retrieve large phylogenetic trees. It uses the 'propinquity' supertree pipeline [8] to synthesize large phylogenetic trees using multiple taxonomic data sources including published phylogenetic trees. The user can easily download a phylogenetic tree for a desired taxonomic group from the Open Tree web interface [4].

B. The generic pipeline for the conversion of Phenoscape KB trait matrices

Despite the availability of resources for downloading large trait matrices and large phylogenetic trees, there are limitations when integrating them due to differences in original data sources. The developed pipeline has six major steps that sequentially convert the trait matrices downloaded from the Phenoscape KB to a version that can be efficiently mapped to Open Tree phylogenies. The pipeline is implemented in Python and a command line version of the pipeline is openly available on GitHub [9]. The inputs to the pipeline are a presence/absence trait matrix retrieved from the Phenoscape KB where presence is usually represented by '1' and absence by '0' and an Open Tree phylogenetic tree file. The steps of the pipeline are explained below.

- The matrix conversion: this step converts the input Phenoscape KB matrix from NeXML format to tab-delimited format, which is efficient for the downstream analysis.
- The pre-processing step: this removes the taxa with missing data from the converted matrix.
- The removal of apparent polymorphisms and conflicts: this step removes higher-level taxa (families, genera, etc.) that contains '0&1' as the state. These can be due to conflicting statements of presence and absence of a certain trait for the same taxon by authors (conflicts), or authors not giving details regarding which species lack or contain a

certain trait within a higher-level taxon, such as a family (apparent polymorphisms) [10].

- Distinguishing inferred versus asserted data: The use of the Uberon anatomy ontology to annotate traits allows unknown character states to be inferred through ontology-enabled reasoning [10]. For instance, pectoral fin is composed of pectoral fin rays, which can be represented as "every 'pectoral fin ray' is part of a 'pectoral fin'" within the anatomy ontology. If an author declares that pectoral fin rays are present in a fish, ontology-based reasoning automatically infers that pectoral fin is also present, because pectoral fin rays do not exist without a pectoral fin. Matrices from the Phenoscape KB contain both inferred data as explained above and also character states that are based on direct author statements (asserted data). However, if an author directly declares that the pectoral fin is present for a particular fish, the resulting character state of presence for the pectoral fin is more reliable than presence based on inference. This pipeline step distinguishes between asserted versus inferred states to differentially visualize them after an ancestral state reconstruction.

- Propagation: Matrices from the Phenoscape KB contain data for higher-level taxa, such as families and genera, notably when an author wants to state that a certain character like pectoral fin is present for an entire group, such as a family. However, it is impossible to map these higher-level data to phylogenetic trees, especially at a large scale. To avoid losing a large proportion of data at higher-levels, we implemented an algorithm to propagate trait data from higher-level taxa to corresponding species. This step propagates data from genera during the first iteration and families in the second (Fig. 1). During the propagation, existing species-level character states are not replaced and will be considered as propagation conflicts.

- Taxonomic name reconciliation: The taxon names from the Phenoscape KB matrices are annotated using ontology terms coming from the Vertebrate Taxonomic Ontology (VTO) [11], which is based mainly on the NCBI taxonomy system [12]. However, Open Tree taxon names are based on the NCBI taxonomy system and some additional sources not used by the VTO. This causes several mismatches when combining the Phenoscape KB trait matrices with Open Tree phylogenies. Furthermore, misspellings, abbreviations of taxon names, and synonyms further complicate the name reconciliation [10]. The final step of the pipeline attempts to overcome these challenges using an algorithm that first matches taxa using NCBI taxonomy identifiers and then using taxon names in the second iteration.

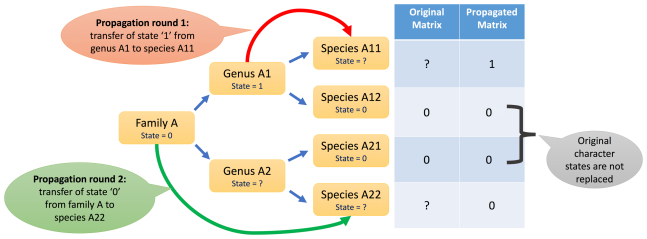

Figure 1. A schematic representation of the propagation algorithm. During the first iteration (red arrow), data is propagated from genera to corresponding species, and during the second iteration (green arrow), data is propagated from families to the remaining species with missing data. Species with existing data are not modified by the propagation.

C. *Ancestral state reconstruction*

The final output of the pipeline is a matrix that can be easily mapped to an Open Tree phylogeny. To perform the ancestral state reconstruction, Mesquite [13], a popular software for evolutionary analysis can be used. To demonstrate the performance of the pipeline, we downloaded matrices for pectoral and pelvic fins from the Phenoscape KB and performed ancestral state reconstructions using a Teleostei phylogenetic tree retrieved from the Open Tree, which contained 38,419 species. The merging of the tree file and trait matrices, ancestral state reconstructions, and visualizations were performed using the Mesquite software.

III. Results and Discussion

One major challenge of large-scale data integration is reducing the missing data proportion. Usage of ontology-enabled reasoning in the Phenoscape KB and propagation in the pipeline significantly reduced the missing data percentage of the original matrices. When both matrices for pectoral and pelvic fins were taken together, ontology-enabled reasoning reduced the missing data percentage from 98% to 85.9%, and propagation further reduced it to 34.8% compared to the total amount of data mapped to the Teleostei phylogenetic tree. The extension of the number of species with data for pectoral and pelvic fin by inference and propagation is shown in Fig. 2.

The visualization of pectoral and pelvic fin evolution after the ancestral state reconstructions is shown in Fig. 3. For a better visualization, we merged the pipeline outputs of pectoral and pelvic fins to a single matrix and visualized both traits simultaneously. According to ancestral state reconstruction results, there is a possibility for a minimum of 19 losses and a minimum of 3 regains for the pectoral fin, and a minimum of 48 losses and a minimum of 14 regains for the pelvic fin [10].

Without using the pipeline, it is difficult to directly map the data from the Phenoscape KB to a phylogenetic tree, especially at a large scale. The pipeline not only facilitates the integration of trait data with phylogenies, but also ensures that original data is extended by propagation to preserve the data for higher-level taxa, which otherwise would be lost during the mapping. The minimization of missing data proportion achieved by the pipeline is essential for large-scale ancestral state reconstructions.

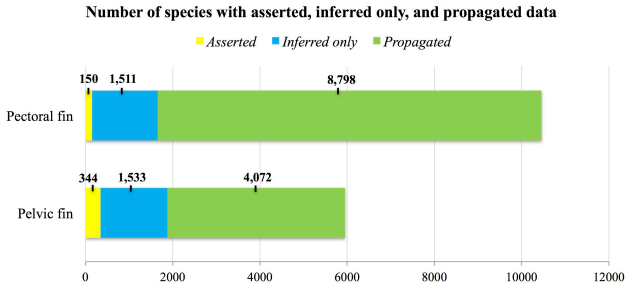

Figure 2. Combined usage of inference and propagation extends original data. The bar chart shows the number of species with asserted (yellow), inferred only (blue), and propagated (green) data for the pectoral fin and pelvic fin. The increase of the number of species with data after inference and then propagation demonstrate the importance of these steps in reducing missing data [10].

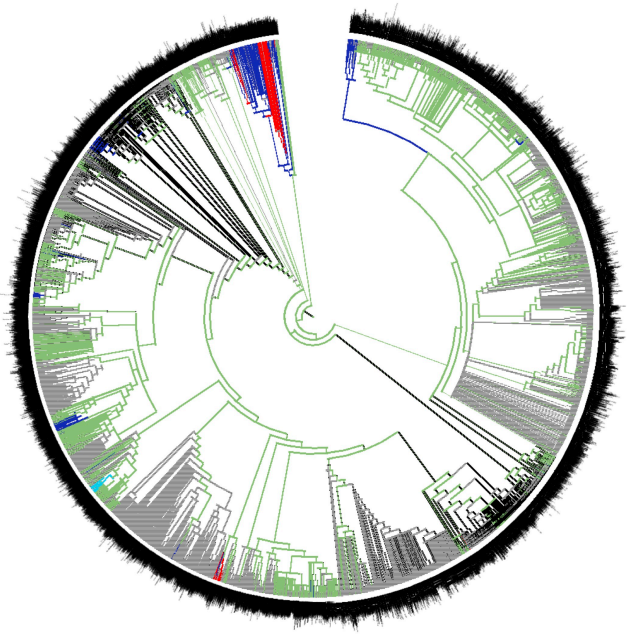

Figure 3. Evolution of the presence and absence of paired fins in a phylogeny of 38,419 teleost fishes. Light blue: absence of only the pectoral fin, dark blue: absence of only the pelvic fin, red: absence of both fins, green: presence of at least one of the fins, and grey/black: no data for either fin.

IV. Conclusion

The era of 'big data' mandates conventional biological analyses to be conducted on a large scale. Evolutionary analyses performed using ancestral state reconstructions were limited due to challenges that occurred when integrating large trait matrices with large phylogenetic trees, but our pipeline can significantly extend the scale of the analyses using public databases to retrieve large data sets. We conducted an evolutionary analysis using a phylogenetic tree that contained over 38,000 taxa, because of the efficiency of the pipeline. A user can retrieve a trait matrix for any character that is available in the Phenoscape KB and use this pipeline to efficiently convert it to a version that can be integrated with a large phylogenetic tree from the Open Tree. The automation of tedious tasks, such as reconciling taxon names between the two systems and the extension of data achieved by computational techniques, such as ontology-enabled inference and propagation makes this pipeline a valuable tool for evolutionary biologists.

Acknowledgment

We thank the Phenoscape team for their help with the Phenoscape KB and Karen Cranston for her assistance with the Open Tree. This work is supported by a grant from the National Science Foundation (DGE-1633213). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation

References

[1] G. Didier, "Time-Dependent-Asymmetric-Linear-Parsimonious Ancestral State Reconstruction," Bull. Math. Biol., vol. 79, pp. 2334-2355, 2017.

[2] C. W. Cunningham, "Some limitations of ancestral character-state reconstruction when testing evolutionary hypotheses," Syst. Biol., vol. 48, pp. 665-674, 1999.

[3] Phenoscape Knowledgebase.[Online]. Available: kb.phenoscape.org

[4] Open Tree of Life.[Online]. Available: http://opentreeoflife.org

[5] R. C. Edmunds, B. Su, J. P. Balhoff, B. F. Eames, W. M. Dahdul, H. Lapp, et al., "Phenoscape: Identifying Candidate Genes for Evolutionary Phenotypes," Mol. Biol. Evol., vol. 33, pp. 13-24, 2016.

[6] T. A. Dececchi, J. P. Balhoff, H. Lapp, and P. M. Mabee, "Toward synthesizing our knowledge of morphology: Using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies," Syst. Biol., vol. 64, pp. 936-952, 2015.

[7] M. A. Haendel, J. P. Balhoff, F. B. Bastian, D. C. Blackburn, J. A. Blake, Y. Bradford, *et al.*, "Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon," J. Biomed. Semant., vol. 5, p. 21, 2014.

[8] B. D. Redelings and M. T. Holder, "A supertree pipeline for summarizing phylogenetic and taxonomic information for millions of species," PeerJ, vol. 5, p. e3058, 2017.

[9] pasanfernando/generic_pipeline_for_trait_integration.[Online]. Available: https://github.com/pasanfernando/generic_pipeline_for_trait_integration

[10] L. M. Jackson, P. C. Fernando, J. S. Hanscom, J. P. Balhoff, and P. M. Mabee, "Automated integration of trees and traits: a case study using paired fin loss across teleost fishes," Syst. Biol., in review.

[11] P. E. Midford, T. A. Dececchi, J. P. Balhoff, W. M. Dahdul, N. Ibrahim, H. Lapp, *et al.*, "The vertebrate taxonomy ontology: a framework for reasoning across model organism and species phenotypes," J. Biomed. Semant., vol. 4, p. 34, 2013.

[12] S. Federhen, "The NCBI Taxonomy database," Nucleic Acids Res., vol. 40, pp. D136-D143, 2012.

[13] W. Maddison and D. Maddison, "Mesquite: a modular system for evolutionary analysis," Version 3.2, http://mesquiteproject.org, 2017.