Phylogenetic Experimental Design via Signal-Noise Framework

J. Nicholas Fisk, Alexander Dornburg, Jeffrey Townsend
Interdepartmental Program in Computational Biology and Bioinformatics, Yale University
Email: jeffrey.fisk@yale.edu Twitter: @inSiliConjurer

Goal

Development of a R package for comprehensive phylogenetic experimental design.

Motivation

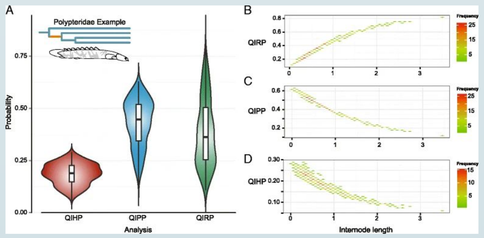
Phylogenetic trees represent a hodgepodge of interconnected hypotheses, only some of which are of interest to particular research programs. Determining the optimal gene-taxa sampling schema prospectively allows for hypothesis-driven data collection and retrospective filtering to maximize the probability of achieving sufficient power to resolve specific hypotheses.

Abstract

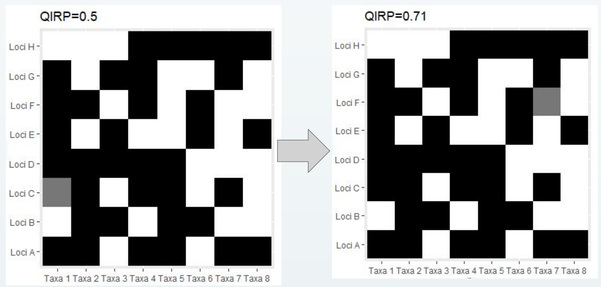
In the emergent big-data world of phylogenomics, it is clear that big data results bulwarked by the traditional hallmarks of strong support are sometimes in conflict with one another, and that the resolution of this conflict requires rigorous thought about the sources of conflict and consequently the relative power of data to address phylogenetic hypotheses. Theoretical tools have been derived to address long-standing controversies in experimental design that have occasionally engendered contentious academic debate, such as i) the power of different genes and phylogenetic characters, ii) the relative utility of increased taxonomic versus character sampling iii.) the potential to design taxonomically dense phylogenetic studies optimized by taxonomically sparse genome-scale data. Here, we present an implementation of these theoretical tools to guide phylogenetic experimental design using advances to the phylogenetic signal framework to iteratively rank-order gene-taxa sampling schema and to ensure proposed sampling schema reach the desired power to answer specific phylogenetic hypotheses.

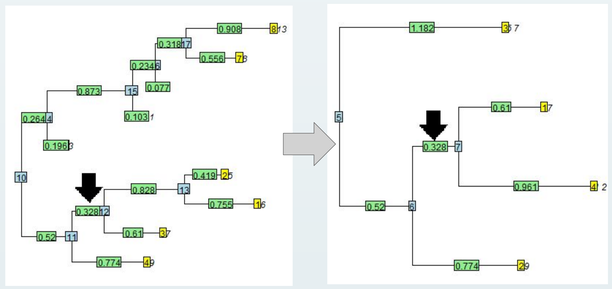
Introduction

- Previously, Dornburg et al [1], developed the PhyInformR R package implementing phylogenetic informativeness theory described in Townsend et al[2]
- Calculations in PhyInformR quantify not only the probability of correct resolution (QIRP), but also quartet internode homoplasy probabilities (QIHP) as well as quartet internode polytomy probabilities (QIPP).
- While PhyInformR is a useful tool for data exploration, it requires many user-decisions and interpretation of data to use in experimental design
- Additionally, advancements have been made in the theory of quartet-based phylogenetic informativeness that further empower tools for phylogenetic experimental design.

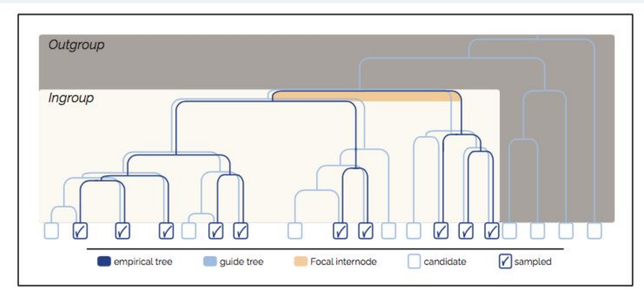


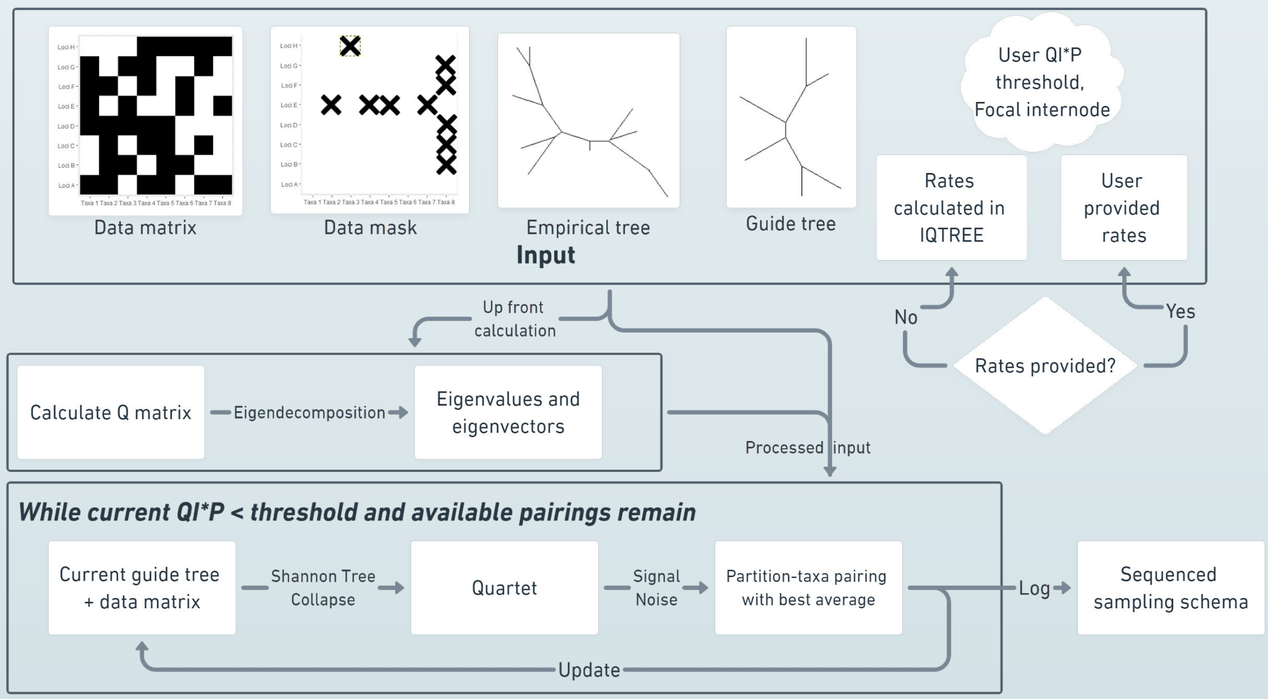
Visual Methodology





Left: One iteration of the design search schema. Grey boxes represent candidates for sampling. Shannon tree collapse is performed with respect to the internode designated by the black arrow. Data masks can be applied for inaccessible gene-taxa combination.
Bottom: Emblematic hypothesis and phylogenies
Far bottom: Diagrammatic workflow of UltimateSignalNoise





Methodology/Features

Shannon Information Collapse: Rather than rely on quartet decomposition, mutual information is used to collapse trees iteratively into quartets, minimizing information loss.

Uneven quartet branch length: PhyInformR previously limited the user to 2 distinct lengths (in accordance with previous theory). Here, implementation is expanded to allow uneven branch lengths in collapsed quartets.

Automated iterative design schema: Given the input and internode of interest, the program will generate a gene-taxon sampling schema until QIRP reaches desired power.

Future Work

- UltimateSignalNoise currently lacks the visualization functionality from PhyInformR. Such visualizations will be adapted and carried forward.
- PhyInformR was made available on CRAN; due to dependencies of UltimateSignalNoise on IQTree, it will have to be distributed via BioConductor.
- Parallelisation and GUI interface

Acknowledgements

This publication was made possible by an NIH/NLM-funded predoctoral fellowship to J. Nick Fisk (5T15LM007056-32). We also thank the Yale Center for Research Computing for their continued support.

Citations

1. Dornburg, A., Fisk, J.N., Tamagnan, J. et al. PhyInformR: phylogenetic experimental design and phylogenomic data exploration in R. BMC Evol Biol 16, 262 (2016). https://doi.org/10.1186/s12862-016-0837-3
2. Townsend JP, Su Z, Tekle YI. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. Syst Biol. 2012;61:835–49.

github.com/jnickfisk/UltimateSignalNoise