A Mechanism for Storing and Retrieving Statistical Properties of Phylogenetic Trees

Haleh Ashki, James C. Wilgenbusch, and Paul van der Mark

Department of Scientific Computing, Florida State University, Tallahassee, FL

INTRODUCTION

Most Biological inference is improved by knowing something about the evolutionary relationship among the biological objects that one is studying. A phylogenetic tree is a schematic used to show evolutionary relationships and plays an important role in both micro and macro evolution. Unfortunately, estimating evolutionary relationship is not an easy task. In general the practice can be broken down into two parts; scoring the phylogenetic tree and searching for an optimal tree based on a scoring criteria. The former is generally considered to be the small problem, while the later is consider the large problem because of how large the tree space gets when more sequences are added to an analysis.

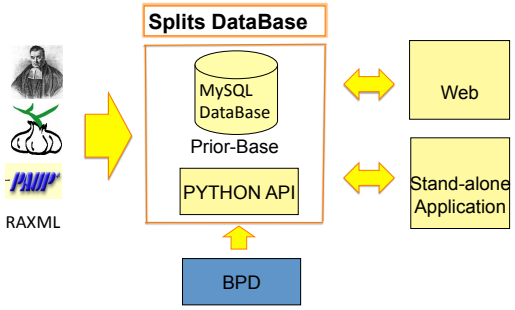

FIGURE 1: Structure of application

This work is motivated by the fact that there are numerous methods to evaluate phylogenetic trees and a growing number of applications that attempt to implement these methods. The splitDB project uses database tools, web-based applications, and graphical output to help users make sense of the different answers they typically obtained when analyzing phylogenetic data with different methods and software. At the core of this project is the database used to store the support values for each split.

WHY A DATABASE?

- **Data Retention:** store results for subsequent comparisons; use stored split frequencies for Bayesian priors
- **Speed**: queries can be construction to outperform flat file searches
- **Interoperability**: can parse, store, and access data from other packages
- **Fault Tolerance**: robust to network failures and computer failures

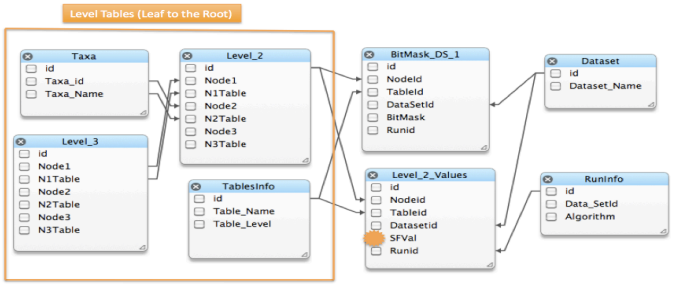

FIGURE 2 : Database Design

What is SplitDB

SplitDB is a database used in conjunction with python scripts that communicate with other web-based resources or the program can also be used as stand-alone application. SplitDB allows users to simultaneously view results generated by different methods and software packages, which allows users to quickly identify incongruent splits and to further investigate the cause of the discrepancy.

The SplitDB's input is the phylogenetic trees generated as output by other programs, e.g., MrBayes(4), Paup(3), RaxML(5). Some use Markov chain Monte Carlo (MCMC) methods(1) to estimate the posterior probability of splits. This estimating of posterior distribution is based on free parameters (e.g., tree topology and substitution model). The nonparametric bootstrap (Felsenstein, 1988)(2) can also be used to estimate the support for each split. SplitDB gives users a user-friendly and graphical view to compare split support values within or among separate MCMC runs or Bootstrap tests.

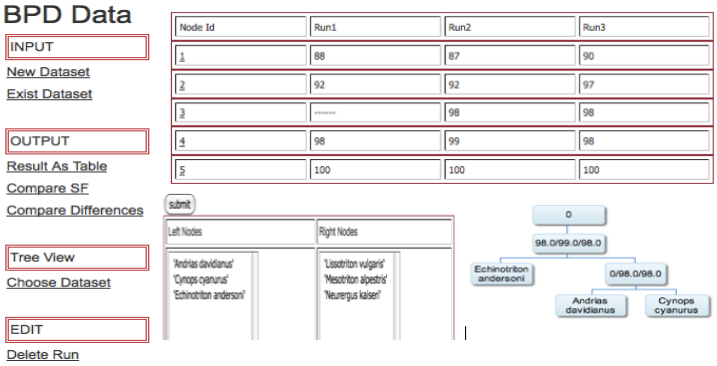

FIGURE 3:Left: SplitDB Menu; Top: Table Containing SF of all runs of each node(Link). Down: list of Left/Right child nodes and graphical representation of left child nodes.

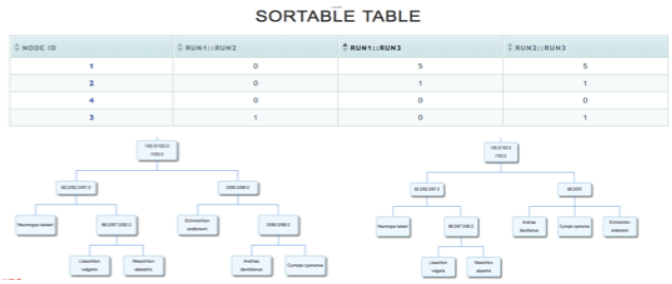

FIGURE 4. Top: Sortable Table comparing SF values by getting SF threshold and differences as input parameters. Down: Graphical View of a single dataset for different runs.

REFERENCES

1.Ronquist, F., B. Larget, J. Huelsenbeck, J. Kadane, D. Simon, and P. van der Mark. 2006. Comment on "Phylogenetic MCMC algorithms are misleading on mixtures of trees" Science 312:367a.
2. Joseph Felsenstein, "Confidence Limits on Phylogenies: An Approach Using the Bootstrap," Evolution 39, no. 4 (July 1985): 783-791.
3. D. L Swofford, PAUP*: phylogenetic analysis using parsimony (* and other methods) (version, 2002)
4.Huelsenbeck, J. P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17:754-755.
5.A. Stamatakis, T. Ludwig, and H. Meier. Raxml-iii: A fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics, 21(4):456-463, 2005
6.Zwickl, D. J., 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin