Properties of Markov Chain Monte Carlo Performance Across Many Empirical Alignments

Acquisition and processing of phylogenetic datasets

This document describes the methods and scripts we used to obtain datasets from the phylota database <http://ceiba.biosci.arizona.edu/pb/> (Sanderson et al. 2008) and conduct phylogenetic analyses. First, we downloaded the raw fasta files, cleaned, and renamed them. We then aligned the reads using MUSCLE, and trimmed ragged ends from the alignments. Next, we used jmodeltest to evaluate model fit and appended Bayes blocks to each alignment file. Finally, we uploaded these dataset files to a supercomputer cluster for phylogenetic analysis in MrBayes. Once complete, we downloaded output files for diagnostics and evaluation.

Preparing files for the pipeline

These scripts are located in the PhyLoTA\_dataset\_download directory.

We downloaded raw datasets using a wget script to pull all datasets matching predefined criteria for phylota version 184 get184datasets. Once the raw datasets were downloaded, we removed html headers and unreadable characters by downstream programs, replacing them with '\_' using the script 001\_cleanfiles.py. Finally, we renamed files to a simple common format encompassing the pertinent details (version number, root taxon ID, and cluster ID; 002\_renamefiles. In total, we downloaded 27,307 datasets. Of these, 35 were empty (Zero bytes), and are removed, leaving 27,272 datasets for analysis.

Initial steps - Writing directories and copying fasta files

The bash script 003\_directoryMaker writes necessary directories for the pipeline. The fasta input files can be copied to the 01\_FASTA directory to start the pipeline by running 004\_copyfiles.

General Pipeline Procedures

The following scripts are located within the PhyLoTA\_analysis\_pipeline directory. All General Pipeline Procedures can be run in sequence by evoking the 00\_masterScript file in the 00\_SCRIPTS directory.

Pipeline Step 1 - Removing duplicate taxa

Duplicate taxa are removed using the python script 01\_duplicateRemover.py which reads the original fasta file in the 01\_FASTA directory, and writes a new file using only the first unique taxon ID's encountered. Thus, the new file will contain only unique taxa. These single taxon fasta files carry the .output extension.

Intermediate step - Moving output files

The bash script 02\_outputMover moves the original fasta files (with duplicate taxa) to the /ORIGINAL\_FASTA\_FILES directory. The .output files (single-taxon files) are moved to the ../02\_1TAXON\_FASTA directory.

Intermediate step – Appending taxon size to dataset files

The bash script 03\_n.tax.sh will count the number of taxa in each .output file within the 02\_1TAXON\_FASTA directory and prepend this number to each filename. We retained datasets with 25 – 250 taxa in this analysis.

Pipeline Step 2 - Performing MUSCLE alignments

The bash script 04\_MUSCLE\_aligner performs MUSCLE alignments (Edgar 2004) on all files in the 02\_1TAXON\_FASTA directory with a .output extension. MUSCLE alignments are performed using default settings.

Intermediate step - Moving alignment files, and shortening taxa names

The bash script 05\_alignmentMover moves all files which carry a .fasta extension (alignment files) to ../03\_../MUSCLE\_ALIGNMENTS. Files carrying the .output file extension are moved to ../02\_1TAXON\_FASTA/SINGLE\_SPECIES\_FASTA\_FILES/.

The perl script 06\_fasta\_shorten.pl (Estill and Bennetzen 2009) changes the headers in the fasta files to give shorter names. In this case, all names are shortened to 15 characters by issuing the following command:

./06 \_fasta\_shorten.pl -l 15 -i ../03\_MUSCLE\_ALIGNMENTS/ -o ../03\_MUSCLE\_ALIGNMENTS/SHORT\_ID

Pipeline Step 3 – Removing ragged ends from alignment files

08\_trim\_ends\_fasta.R is an R script that removes ragged ends of alignment files to improve phylogenetic inference. The 07\_loop.sh script will loop over all .fasta files in the SHORT\_ID directory and write output to this folder with the .fas file extension. The 00\_masterScript will automatically copy these two files to the SHORT\_ID subdirectory in the 03\_MUSCLE\_ALIGNMENTS directory and execute the command on all fasta files there.

Pipeline Step 4 - Nexus file conversion

Since jmodeltest requires a nexus input file, it is necessary to convert the fasta files to nexus file format. This is accomplished by running the script 09\_nexusConveter. This script runs Seqmagick, a utility that converts between several phylogenetic file types. In this case, we are using it convert from fasta to nexus. Seqmagick can be installed on your system by running pip install seqmagick.

Intermediate step - Moving nexus files

The bash script 10\_nexusMover moves all nexus output files (.nex) to ../04\_NEXUS\_CONVERSION. It then moves all .fasta files (the alignment files) to ../03\_MUSCLE\_ALIGNMENTS/ALIGNMENTS. Finally, all trimmed output files TRIMMED\_OUTPUT.

Pipeline Step 5 - Model testing

jModelTest (Guindon and Gascuel, 2003) is a tool that is used to carry out statistical selection of best-fit models of nucleotide substitution. The python script 11\_jmodeltest.py implements this. Specifically, models are tested and evaluated using the following command line arguments:

-g 4 [include models with rate variation among sites, and set the number of categories to 4]

-i [include models with proportion invariable sites]

-f [include models with unequal base frequencies]

-t BIONJ [sets the base tree for likelihood calculations to Neighbor-Joining (NJ) topology for each model] (while NJ may be less accurate than a Maximum Likelihood [ML] tree topology search for each model, ML is computationally prohibitive for the amount of data considered here).

-s 3 [sets the number of substitution schemes to 3: JC/F81, K80/HKY, SYM/GTR, i.e., those implemented in MrBayes]

-AICc [select best-fit models using the corrected Akaike Information Criterion]

Pipeline Step 6 - Write Bayes block

The python script 12\_mbb.py appends a Bayes block to the end of the nexus file. This contains important information for running phylogenetic analysis in MrBayes (Huelsenbeck and Ronquist, 2001), such as: specifying the substitution model to be used; priors; number of runs and chains; generations to run; diagnostic, print, and sample frequencies; as well as number of runs to be discarded as burn-in. MrBayes uses Markov Chain Monte Carlo to sample from the posterior probability distribution, and metropolis-coupling to speed up convergence where three "heated" chains are in parallel to the "cold" chain. The "heated" chains are simply sampling the same probability distribution, but one that has been "heated" to "melt" or "flatten" the landscape of peaks (optima) defined by the (posterior) probability distribution. Sampling switches between the cold and heated chains so the cold chain can escape local optima when sampling along the flattened peaks of probability space and more reliably identify the globally optimum parameters.

Intermediate steps - Moving analysis files, writing SLURM files, writing directories

The bash script 13\_analysisFilesMover moves jModelTest output files to ../04\_NEXUS\_CONVERSION/jmodeltestOutputFiles and copies nexus (\*.nex) output files to ../05\_ANALYSIS\_FILES/NEXUS\_FILES.

Datasets are now ready for phylogenetic analysis.

References

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res 32:1792-97.

Estill, J. C. and J. L. Bennetzen 2009. The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes. Plant Methods. 5:8

Guindon S. and O. Gascuel. 2003. A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. Systematic Biology 52:696-704.

Huelsenbeck, J. P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17:754-755.

Sanderson, M. J., D. Boss, D. Chen, K. A. Cranston, and A. Wehe. 2008. The PhyLoTA Browser: processing GenBank for molecular phylogenetics research. Syst. Biol. 57:335-346.