

# Instructions for installation and implementation of AUGIST (Accommodating Uncertainty in Genealogies while Inferring Species Trees) for the Mesquite software system.

by Jeffrey C. Oliver

Latest revision: January 1, 2009.

## Installation Instructions

Before you begin, the AUGIST package must be installed in an already installed version of Mesquite. If you do not have Mesquite installed yet, download the latest version and follow the instructions on Mesquite installation (<http://mesquiteproject.org/mesquite/download/download.html>).

Download the file “augistForMesquite.zip”, available from <http://www.lycaenid.org/augist/>.

Extract the directory within this archive, called “augist”.

Move the augist directory into the Mesquite\_Folder/mesquite directory.

## Mesquite Instructions

These instructions detail how to infer species trees using deep coalescences while accommodating gene tree uncertainty. Users are strongly recommended to read Maddison (1997) and Maddison and Knowles (2006), and to be familiar with the Mesquite software system (Maddison & Maddison 2008). The process has three steps:

1. Generating gene tree distributions
2. Inferring species trees accommodating uncertainty in gene genealogies
3. Processing species trees to compute consensus with clade frequencies

These instructions will guide users on the entire process; however, steps 1 and 3 are not novel and will be very general.

### Before you begin

These instructions assume a basic understanding of the software necessary to generate gene tree distributions and standard Mesquite conventions. For the former, the instructions are generally written for use with MrBayes, although any software which produces a NEXUS TREE block corresponding to a distribution of gene trees can be used. For the latter, it is recommended the user understand how to set up associations between taxa (see

[http://mesquiteproject.org/Mesquite\\_Folder/docs/mesquite/popGen/popGen.html#establishing](http://mesquiteproject.org/Mesquite_Folder/docs/mesquite/popGen/popGen.html#establishing)).

### **1. Generating gene tree distributions**

In this step, gene tree distributions for each locus are generated, such that the frequency at which a clade occurs in the distribution reflects the likelihood or probability of that clade based on the data (Holder and Lewis 2003; if Bayesian analyses are being used to generate gene tree distributions for loci, users are recommended to also consult Alfaro and Holder [2006] for information concerning the influence of topological priors on tree inference). To use these gene tree distributions, a single gene tree block for each locus must be created to correspond to those trees sampled from all runs. Omitting burnin trees at this step will make the process easier in later steps, but if a single tree file is used for a locus (e.g. from a single MCMC run) the burnin trees can be ignored at a later step (2.j, below). This can be done in a text editor, or in a tree managing program such as PAUP\*. Once there is a NEXUS file of trees for each locus (i.e. if you are using four loci, you need four tree files), you may proceed to step 2. For a discussion on recommended sampling design, regarding the number of individuals and the number of loci to sample, users should see Maddison and Knowles (2006).

### **2. Inferring species trees using deep coalescences**

In this step, gene trees are sampled from the distributions generated in step 1 and used to infer a species tree, minimizing the number of deep coalescences of the multiple loci. To begin, you will need a Mesquite file with the following:

- Species Taxa Block: contains the species for which the species tree is being inferred
- Genes Taxa Block: contains individual samples used in gene tree distribution generation in step 1. Note: the names in this Mesquite file *must* exactly match names in the tree-containing

NEXUS files produced in step 1. This taxa block will include all sampled alleles for all loci. The example below shows a gene taxon block for three species (A, B, and C) and two loci (locus1 and locus2); taxa are named using a <Species>\_<Locus>\_<Allele> convention:

```
A_Locus1_1
A_Locus1_2
A_Locus2_1
A_Locus2_2
B_Locus1_1
B_Locus1_2
B_Locus2_1
B_Locus2_2
C_Locus1_1
C_Locus2_1
```

- Association: a single association between Genes Taxa Block and Species Taxa Block

The file `augistExample.nex` is an example of a file containing these three parts. The Species Taxa Block contains 12 taxa (Species A through K); the Genes Taxa Block contains 24 alleles (two per species) of three loci, for a total of 72 taxa. The tree files corresponding to gene tree distributions for each of the three loci are called `augistExampleLocus1.nex`, `augistExampleLocus2.nex`, `augistExampleLocus3.nex`.

To create a distribution of species trees:

- 2.1. Select Taxa & Trees > Save Copies of Tree Blocks > Tree Block Combiner.
- 2.2. Select the Species Taxa Block.
- 2.3. Choose Tree Search as the Tree Block Source.
- 2.4. If prompted, select Heuristic (Add & rearrange) as the Tree Searcher.
- 2.5. Select Deep Coalescences Multiple Loci as the Criterion for tree search.
- 2.6. For the Contained (gene) tree interpretation, uncheck “Contained polytomies auto-resolve” and “Use Branch lengths of Contained tree”; unless the gene trees can be reliably rooted, check the “Treat contained as unrooted” option.
- 2.7. For the source of contained trees, select Sample Trees from Multiple Sources (a secondary choice).
- 2.8. You will then be prompted to enter the number of tree sources; this is the number of loci you are using to infer the species tree and should be the same as the number of tree files you created in step 1 (e.g. if you are using 4 loci, enter “4”).
- 2.9. Enter “1” as the Number of trees per source when asked “How many random trees to sample per source”.
- 2.10. You will then be asked a series of queries regarding each of the gene genealogies. For each locus you will first enter the Tree Block Source as Randomly Sample Trees from Separate NEXUS File (do not choose Use Trees from Separate NEXUS File nor Sample Trees from Separate File), then choose the file corresponding to the gene tree distribution you created in step 1. Finally, you will be asked to enter the “Number of Trees to Ignore”; if you are using a tree file that contains burnin trees you do not want to include in your analysis, enter the length of the burnin at this step (e.g. if you sampled 10,000 trees, but the first 1,000 are part of the burnin, enter “1,000” as

the number of trees to ignore). However, if the tree file does not contain any burnin trees, enter "0".

These three queries will be repeated as many times as the number of tree sources entered in step h above. Be sure to keep track of which loci you have selected (numbering the loci 1, 2, 3, etc. beforehand may help in bookkeeping).

- 2.11. Select a tree rearranger. The SPR rearranger will take longer than the NNI rearranger, but will search more treespace. For data with few species (less than 14), the NNI rearranger may be sufficient, but SPR should be used whenever possible.
- 2.12. Enter the maximum number of trees to store at each step during branch swapping. The default is 100. If search replicates are consistently recovering the maximum number of trees, it is very likely that there are additional trees of equal score that are not recovered, and it increases the probability that the search does not recover the true optimal tree(s).
- 2.13. For the Tree Block source options, enter the number of tree sources from which to save trees. This is the number of species tree searches that will be performed. So, for 100 species tree search replicates, enter "100." If you want tree weights to be stored, check "Store Tree Weights." Trees are weighted as the inverse of the number of most optimal trees encountered for that tree search replicate. For example, if a search replicate recovered 15 equally optimal trees, each tree will have a weight of 1/15.

The optimal number of search replicates to run will depend on the specific conditions of each search, but you can check for accuracy of clade frequencies by comparing a consensus species tree based on the first half of the replicates to a consensus species tree based on the second half of the replicates. The average standard deviation of clade frequencies can be calculated and used as a sampling statistic. If the average standard deviation is below 0.02, the clade frequencies based on the consensus of the entire sample of species trees is likely an accurate representation of the true clade frequencies implied by the underlying gene tree distributions. The minimum number of recommended species tree search replicates is 100, but authors are encouraged to perform at least 500-1000 replicates to ensure accurate clade frequency estimates.

- 2.14. Enter a name for the file to which the species trees will be saved.
- 2.15. For the number of tree blocks to save copies, enter "1". If you enter a number larger than one, multiple iterations of the entire process will be carried out and saved in that many files. The file will be named <filename>0.nex; remember this filename - it will be used in step 3 to compute the consensus species tree.

### 3. Processing species trees to compute consensus with clade frequencies

Using a Mesquite file with a Species Taxa Block of the taxa included in the species tree inferred in step 2 (you can use the same file as in step 2), Select Taxa & Trees > Make New Trees Block from > Consensus Tree and select (if prompted) the Species Taxa Block. Select Use Trees from Separate NEXUS File as the source of trees for consensus and choose the file created in step 2.15, above. Choose Majority Rules Consensus as the consensus calculator and enter options for the consensus calculation. If you checked "Store Tree Weights" in step 2.13 above, you can select "consider tree weights" in the calculation. If you want a list of frequencies for all those partitions encountered in the species tree inference procedure (not just those that were encountered at a frequency higher than the consensus threshold value), also check the "write group frequency list" box; the frequency list will be written to the Mesquite Log. To see support values for partitions in the species tree, open a Tree Window displaying the consensus tree. Select Tree > Node-Associated Values > Choose Value to Show... and check the box that says "consensusFrequency". The values displayed on nodes represent the proportion of inferred species trees containing that clade. The values are also available from the "Text" view of the tree.

The resulting consensus tree can also be exported as a NEXUS (File > Export... > Export NEXUS Tree File) tree for use with other programs.

**How to cite:** If you use the AUGIST approach, please cite the original description (Oliver 2008), as well as Mesquite (Maddison & Maddison 2008).

### References

- Alfaro, M.E. & M.T. Holder. 2006. The posterior and the prior in Bayesian phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* 37:19-42.
- Holder, M. & P.O. Lewis. 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics* 4:275-284.
- Maddison, W.P. 1997. Gene trees in species trees. *Systematic Biology* 46:523-536.
- Maddison, W.P. & L.L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* 55:21-30.
- Maddison, W.P. & D.R. Maddison. 2008. Mesquite: a modular system for evolutionary analysis. Version 2.5, build j55. <http://mesquiteproject.org>.
- Oliver, J.C. 2008. AUGIST: inferring species trees while accommodating gene tree uncertainty. *Bioinformatics* 24:2932-2933.