

MP-EST (1.6) Manual

Liang Liu

August 26, 2017

MP-EST estimates species trees from a set of rooted binary gene trees by maximizing a pseudo-likelihood function. Users may choose to run multiple independent searches for the maximum likelihood tree. Each search starts with a random tree. The program outputs the best trees and their log-likelihood scores generated from the multiple runs. The tree with the largest log-likelihood score is the estimate of the species tree. MP-EST can take gene trees w/o branch lengths, though branch lengths will not be used for inferring species trees.

1. Installation

MP-EST is written in C. To compile the source code, type *make* at the terminal window on Mac (or the Dos command window on PC). You need to set *architecture* to the correct platform (Mac, Unix, or Windows) in the file *makefile*. Currently, the parallel version is not available. Thus, please set *MPI = no*.

If the number of taxa in your data set exceeds the constant (NTAXA=400) defined in the file *mpest.h*, you need to increase the constant and recompile the source code. Other important constants include NGENE, MAXROUND, and NUM_NOCHANGE. MAXROUND defines the maximum number of rounds that the algorithm can run. The algorithm is terminated when the MAXROUND round is reached. If you think the current setting MAXROUND = 10000000 is too small for the algorithm to find the maximum pseudo-likelihood estimate of the species tree, you have to increase MAXROUND. The algorithm will be terminated if no higher pseudo-likelihood scores are found for a consecutive NUM_NOCHANGE of rounds. Increasing NUM_NOCHANGE can make it more likely to find the maximum pseudo-likelihood score.

2. Input data

The input data of MP-EST are rooted binary gene trees. The input gene trees may be produced by the ML phylogenetic programs RAxML, PhymI, Phylip, or Paup from sequence data.

As most ML programs output unrooted trees, the ML gene trees must be rooted by the outgroup species. Different gene trees can have different outgroups, and the outgroup may have multiple species.

Species may have multiple alleles and MP-EST calculates the total frequencies of triples across multiple alleles.

3. The control file

Comments (in red in the example control file below) are not allowed in the control file. Thus, the red lines must be removed before using the example control file to run MP-EST. The control file starts with the name of the gene tree file. The gene tree file and the control file must be in the same folder. Otherwise, you have to specify the full path of the gene tree file. If you just want to calculate triple distances among gene

trees, set the second line to 1. Otherwise, set it to 0. Species may have any arbitrary names, but following each species name you have to specify the number of alleles for that species and the names of the alleles.

```
genetree.tree # the name of the gene tree file
0 # 1: calculate triple distance among trees. 0: donot calculate
6950387 # seed
5 # number of independent runs
20000 9 # number of genes and number of species
species1 1 S1 # species, number of alleles, allele names in gene trees
species2 1 S2
species3 1 S3
species4 1 S4
species5 1 S5
species6 1 S6
species7 1 S7
species8 1 S8
species9 1 S9
1 # usertree (see below for detail)
(((species8:9.0,((species5:9.0,(species4:9.0,(species3:9.0,(species2:9.0,species1
:9.0):0.547063):0.537160):0.604559):1.825150,species7:9.0):0.474750):0.368258,spe
cies6:9.0):0.386622,species9:9.0);
2 # number of taxa being optimized; this is used only for usertree = 4
species1 species2 #the name of the taxa being optimized; used only for usertree=4
```

usertree = 0: the program will generate a random tree as the starting tree

usertree = 1: the program will use the user tree as the starting tree

usertree = 2: optimize the branch lengths for a fixed tree

usertree = 3: the program will calculate the log-likelihood score for the user tree provided with branch lengths

usertree = 4: the program will optimize the placements of a subset of taxa, while keeping the placements of the remaining taxa fixed.

4. output

The output file *besttree.tre* contains the best tree in the nexus format. For multiple runs, the tree block consists of multiple best trees produced from different runs. The pseudo-likelihood score for each best tree is saved within []. The branch lengths of the trees in the output file are in coalescent units. Users should be cautious about the branches of length “1.0”. The value “1.0” is not the actual length of the branch. It indicates that all gene tree triples support the same topology (strong support for the topology), but the corresponding branch in the species tree is not estimable. All trees found in the search for the best tree are saved in the output file *output.tre*.