Maximum pseudo-likelihood estimation of species trees (MPEST)

(manual_July_20_2013)

Liang Liu (lliu@uga.edu)

The MPEST method estimates species trees from a set of gene trees by maximizing a pseudo-likelihood function. The program can run independent searches (chains) in parallel using the parallel version of the program. Each chain starts with a different seed. The program will find the estimate of the species tree with the largest pseudo-likelihood score across chains. **MPEST can take gene trees w/o branch lengths, though branch lengths will not be used for inferring species trees**. To run MPEST, you need to follow the following steps.

(1) Estimate ML gene trees from phylogenetic programs (phyml, phylip, paup).
(2) Root ML gene trees (gene trees estimated from phyml are unrooted trees). Different gene trees can have different outgroups, but all gene trees must have a single species as an outgroup.
(3) Put the rooted ML gene trees (phylip format) into a single file (this is the gene tree file).
(4) Create a control file for running MPEST.
(5) Type "mpest control" to run MPEST on the terminal window (DOS command window on PC)

1. Installation

The program is written in C. The source code of the program MPEST is available at http://code.google.com/p/mp-est/. To compile the source code, type "make" (without quote) at the terminal window (Dos command window on PC). You need to set "architecture = ?" to the correct platform (mac, unix, or windows) in makefile. If you want to use the parallel version of the program, set "MPI=yes" in makefile. Otherwise, set "MPI=no".

If the number of taxa (or genes) exceeds the constant defined in mpest.h, you need to increase these constants and recompile the source code. These constants include NTAXA, NGENE, MAXROUND, NUM_NOCHANGE,

```
#define NTAXA              200          /* max # of species */
#define NGENE              50000        /* max # of loci */
#define MAXROUND       10000000             /* MAX # OF ROUNDS*/
#define NUM_NOCHANGE 20000       /* # OF ROUNDS THAT NO BIGGER LIKELIHOOD VALUES ARE FOUND*/
```

For example, if the number of taxa in your dataset is 285, you need to increase

NTAXA to 300. MAXROUND defines the maximum number of rounds the algorithm will run. The algorithm will be terminated when the MAXROUNDth round is reached. If you think the current setting MAXROUND=10000000 are not enough for the algorithm to find the maximum pseudo-likelihood estimate (MPE) 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. Run the program

type "mpest control". The control file can be created as follows. The example control and tree files are in the "data" folder.

When the data involve a large number of taxa, it is likely that MP-EST cannot find the global optimal tree. For these cases, it is highly recommended to conduct multiple independent runs to thoroughly search the tree space.

2.1 Create a control file

Explanation lines (in red in the example control file) are not allowed in the control file. Thus the red lines must be removed before using the example control file to run mpest. A control file starts with the name of the gene tree file (genetree.tree in the example file). The gene tree file and 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. If you want to find the MPE of the species tree, set it to 0. Since the taxa names in gene trees may not match species names, you have to specify the species name as well as the number and names of taxa that belong to this species. In the example file, the line "**species1 1 S1**" in the example file indicates that the species name is species1 and it contains only one taxon whose name is S1 in gene trees.

```
genetree.tree    # the name of the gene tree file
0                  # 1: calculate triple distance among trees. 0: donot calculate
-1          # seed; -1: random seed; or a large integer (at least 6 digits)
20000 9           # number of gene trees in the gene tree file, 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   # 1: use the user tree, 0: use a random tree, as the staring species tree
in the algorithm
(((species8:5.000000,((species5:5.000000,(species4:5.000000,(species3:5.0000
00,(species2:5.000000,species1:5.000000):0.547063):0.537160):0.604559):1.825
150,species7:5.000000):0.474750):0.368258,species6:5.000000):0.386622,specie
s9:5.000000);  # user tree
```

An example control file

3. Output

The output file, “xx.tre”, is in the nexus format. Species trees are saved every 1000 circles during the search for the maximum of the likelihood function in the tree block of the output file. For multiple chains (parallel version), the trees in the tree block are the ones with the largest pseudo-likelihood score across chains. The very last tree “mpest” is the MPE of the species tree. The pseudo-likelihood score for each saved 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 “9.00”. The value “9.00” 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 due to the lack of topological variance among gene tree triples.