29 July 2017
Woods Hole MBL Workshop in Molecular Evolution
Scott Edwards

More methods for estimating species trees

Files can be found on wiki, Edwards page:

From your unix account prompt, type:

wget https://molevol.mbl.edu/images/7/75/Edwards_species_tree_lab.zip

Then: unzip Edwards_species_tree_lab.zip

Then navigate ('cd') to the appropriate folder for analysis (BEST, MP-EST, STAR, etc)

Outline:

1) BEST – Bayesian estimation of species trees (http://www.stat.osu.edu/~dkp/BEST/)
    i) Best format
    ii) Priors
    iii) Execution

2) MP-EST – maximum (pseudo)likelihood estimation of species trees (http://faculty.franklin.uga.edu/lliu/content/software)

Or download from webserver:

http://bioinformatics.publichealth.uga.edu

3) Depending on time…
    a) Using the Phybase R package (http://faculty.franklin.uga.edu/lliu/content/software) to simulate gene trees, make species trees, and conduct a multilocus bootstrap
        i) R environment of Phybase
        ii) Defining variables (sequences, trees, OTU names, species names)
        iii) Making a STAR tree
        iv) Executing multilocus bootstrap

******type all commands in blue from appropriate directory**********

******examine and edit all text files using nano filename**********

1. BEST - Liu (2008) Bioinformatics, 24:2542-2543; Liu & Pearl (2007) Syst Biol 56:504-514

a) Input file format – modified mrBayes file

BEST block:

```
partition Genes = 30:
        locus097,locus098,locus118,locus119,locus120,locus122,locus130,locus104,locus129,locus143,locus146,locus111,locus13
        5,locus148,locus182,locus200,locus209,locusB098,locus184,locus185,locus186,locus187,locus188,locus192,locus193,locu
        s195,locus198,locus199,locusB200,locus103;
set partition=Genes;
taxset species1=P_acuticauda;
taxset species2=P_hecki;
taxset species3=P_cincta;
taxset species4=T_guttata;
prset thetapr=invgamma(3,0.003) GeneMuPr=uniform(0.5,1.5) best=1;
unlink topology=(all) brlens=(all) statefreq=(all) genemu=(all);
mcmc ngen=3000000 samplefreq=100 nrun=1 nchain=1;
quit;
end;
```

Type: username@class-01 ~]$ best
Best> execute finch-best-star-steac.nex

When analysis is done you can type:
Best> execute `finch-best-star-steac.nex.sumt`

Examine output files (similar to mrBayes): Finch_BEST.nex.run1.p, Finch_BEST.nex.run2.p, Finch_BEST.nex.sptree.con, and *.t, *.parts, *.tprobs files

2. MP-EST - maximum (pseudo)likelihood estimation of species trees – Liu et al. (2010) BMC Evolutionary Biology, 10:302

A) You need rooted gene trees, with following flexibility:

1. Different gene trees may have different outgroups.

2. The outgroup may have multiple species

3. The program can optimize the branch lengths for a fixed tree

4. The program can calculate the log-likelihood score for a fixed tree with branch lengths

5. The program can optimize the placements of a subset of taxa, while keeping the placements of the remaining taxa fixed.

B) control file: ("Maluridae_control_v15.txt") contains information on where the gene trees are, how the gene tree OTUs map onto species, etc.

```
Maluridae_gene.trees #name of gene tree file
0 #1: calculate triple distances among trees; 0: do not calculate
73455 #random seed number
1 #number of independent runs
18 26 #number of genes and species
Kalkadoon_Grasswren 1 Kalkadoon_Grasswren # species name, number of alleles, allele name(s) in gene trees
```

Grey_Grasswren 1 Grey_Grasswren
Carpentarian_Grasswren 1 Carpentarian_Grasswren
…
White_shouldered_fairy_wren 1 White_shouldered_fairy_wren
White_winged_Fairy_Wren 1 White_winged_Fairy_Wren
White_throated_Gerygone 1 White_throated_Gerygone
0 #1 use user tree below; 0: do not use

(((((((Kalkadoon_Grasswren,Dusky_Grasswren),Black_Grasswren),Eyrean_Grasswren),Thick_billed_Grasswren),(Grey_Grasswren,(Carpentarian_Grasswren,Striated_Grasswren)),Short_tailed_Grasswren),((((Lovely_Fairy_wren,Red_winged_Fairy_wren,Blue_breasted_Fairy_wren,Variegated_Fairy_Wren),((((Superb_Fairy_wren,Splendid_Fairy_wren),(Red_backed_Fairy_wren,White_shouldered_fairy_wren)),Purple_crowned_Fairy_wren,White_winged_Fairy_Wren),Emperor_Fairy_Wren)),(Southern_Emu_wren,(Mallee_emu_wren,Red_crowned_Emu_Wren))),(Broad_billed_Fairy_Wren,Orange_crowned_Fairy_wren))),White_throated_Gerygone);

Execute:

username@class-01 ~]$ mpest Maluridae_control_v15.txt

Output tree search (Maluridae_gene.trees.tre) will be in Nexus format; examine last (mp-est) tree in file for results. Branch lengths are in coalescent units, unless length > 9 in which case length is 9 (an arbitrarily long number that can't be estimated with these data).

3) Testing the likelihood of an a priori species tree using MP-EST

The new version MP-EST (1.5) can estimate branch lengths for a fixed species tree, or can estimate branch lengths and the likelihood for a fixed topology. Suppose the null species tree is T.

1. Use MP-EST to fit branch lengths of T and calculate the log-likelihood L0 (option 2 or 3 in control file; user tree must be rooted).
2. Use MP-EST to fit branch lengths of a fixed user tree and evaluate the likelihood (option 2) or simply evaluate the likelihood of a user tree with branch lengths given (option 3) and note the loglikelihood L1.
3. The test statistic is t = L1-L0
4. Use bootstrap across gene trees to find the null distribution of t.
5. Calculate p-value

We can use alternate control files for these excercises: Maluridae_control_v15_1st_tree.txt ; Maluridae_control_v15_mid_tree.txt. The "1st_tree" file contains as the user tree the first (random) tree used in the main species tree search just performed; the "mid_tree" file contains an estimate of the species tree from the middle of the likelihood run.

4) Phybase, an R module for estimating, analyzing and simulating species trees

Load phybase: library(phybase)

a) First: some simple coalescent simulations
Simulate a coalescent tree with n alleles from a single population with theta t:
sim.coaltree(n,t) (for example: sim.coaltree(12,0.01)

b) Making a STAR tree:

- Input file: rooted or unrooted gene trees in phylip or nexus format. But if trees are unrooted, you must first root them to make a STAR tree (can be done in R).

> wrentrees<-read.tree.string(file="Maluridae_gene.trees",format="phylip")

#variable genetrees has 3 values: vector of trees; species names; and TRUE or FALSE for rooted or note.

> wrengenetrees<-wrentrees$tree ##extracts trees from the file and assigns them to variable "genetreevector"

> wrentaxanames<-species.name(wrengenetrees[1]) ##gets gene tree names from the first gene tree; make sure this gene tree has all taxa in it.

> wrenspnames<-species.name(wrengenetrees[1]) ##assigns same names to species tree as in first gene tree

Now, link names in gene tree with names in species tree via a matrix called "species.structure"

> wrentreematrix<-matrix(0,26,26) ##a matrix for 26 species, filled with 0s

>diag(wrentreematrix)<-1 #1s on the diagonal indicate a 1-to-1 correspondence of gene and species names

##now, make a star tree:

>wrenstartree<-
star.sptree(wrengenetrees,speciesname=wrenspnames,taxaname=wrentaxanames,species.structure=wrentreematrix,outgroup="White_throated_Gerygone",method="nj")

Now write the STAR tree to a nexus of newick file:

>write.tree.string(wrenstartree, file="wrenstartree.nex")
>write.tree.string(wrenstartree, format="phylip",file="wrenstartree.phy")

Representing species trees as matrices and simulating gene trees will wait for another time. The Phybase manual has useful instructions for these two topics.

4) The multilocus bootstrap

Input file for DNA sequence data: same as for BEST (Nexus/mrbayes file with BEST block)
#read in a sequence file

>wrenfile<-"Maluridae_seqs.nex"

#assign DNA sequences in that file to a variable "wrenfile"
>wrendata<-read.dna.seq(wrenfile)

>wrensequence<-wrendata$seq #assigns sequences in wrensequence to the file "wrendata"

> wrengenes<-wrendata$gene #assign gene partitions to variable "wrengenes"

> wrennames<-wrendata$name #get taxa names – these are the OTUs in the gene trees

> write.dna(sequence=wrensequence, file= "wrenseqs.phy", format="phylip", name=wrennames)
#can export DNA sequence in nexus of phylip format

#bootstrap the data set
> bootstrap.mulgene(sequence=wrensequence, gene=wrengenes, name=wrennames, boot=100,

outfile="wrenboot_seqs_100.txt")

Can look at bootstrapped data set using nano or other text editor (nano wrenboot_seqs_100.txt)

Then you need to root your gene trees (using R or the STRAW web server) for use in STAR, MP-EST, or other methods.

Multilocus bootstrap replicates can be used for many species tree methods, such as STAR, MDC, MP-EST, SVDquartets and many other species tree methods.

References:

Castillo-Ramírez, S., L. Liu, D. Pearl, and Edwardsm S. V. 2010. Bayesian estimation of species trees: a practical guide to optimal sampling and analysis, Pages 15-33 in L. L. Knowles, and L. S. Kubatko, eds. Estimating Species Trees: Practical and Theoretical Aspects. New Jersey, Wiley-Blackwell.

Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenation. Proceedings of the National Academy of Sciences (USA) 104:5936-5941.

Edwards, S. V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1-19.

Kubatko, LS. 2009. Identifying Hybridization Events in the Presence of Coalescence via Model Selection, Systematic Biology 58(5): 478-488.

Liu, L. (2008). BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics (Oxford, England), 24(21), 2542–3. doi:10.1093/bioinformatics/btn484

Liu, L., L. Yu, and S. Edwards. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evolutionary Biology 10:302.

Liu, L., L. Yu, D. K. Pearl, and S. V. Edwards. 2009. Estimating species phylogenies using coalescence times among sequences. Syst Biol 58:468-477.

Liu, L., Yu L., & Pearl D. K. 2009. Maximum tree: a consistent estimator of the species tree. Journal of Mathematical Biology 60(1): 95-106

Liu, L., L. Yu, L. Kubatko, D. K. Pearl, and S. V. Edwards. 2009. Coalescent methods for estimating phylogenetic trees. Mol Phylogenet Evol 53:320-328.

Liu, L., D. K. Pearl, R. T. Brumfield, and S. V. Edwards. 2008. Estimating species trees using multiple-allele DNA sequence data. Evolution 62:2080-2091.

Liu, L., and D. K. Pearl. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst Biol 56:504-514.