Advanced methods in DNA sequence and multilocus data analyses

Lesson 3 – Species tree estimation from phylogenomic datasets

(Tomáš Fér – Faculty of Science, Charles University, Prague)

25th November 2016

Phylogenomic datasets

When using whole genomes or at least their substantial portions to estimate the evolutionary tree we speak about phylogenomics. There are many approaches allowing us to get sequences from hundreds to thousands of genes or sequences of the whole chloroplast/mitochondria, e.g., whole genome sequencing, transcriptome sequencing, multiplex amplicon sequencing or targeted hybrid enrichment (Hyb-Seq).

Hyb-Seq

This method combines hybridization-based target enrichment of NGS sequencing libraries with genome skimming and currently becomes a standard method of phylogenomics. It results with sequences for hundreds to thousands of loci and also nearly-complete chloroplast genomes can be obtained from off-target reads. Several pipelines for the data analysis have been proposed, e.g., HybPhyloMaker (https://github.com/tomas-fer/HybPhyloMaker) or HybPiper.

Species tree estimation from phylogenomic datasets

During phylogenomic data analysis individual trees are constructed for each gene (gene trees). However, we are mostly interested to construct a tree describing the phylogeny of species (species tree). Gene trees often differ from species trees, creating challenges to species tree estimation. One of the reasons for different topologies between gene trees and species trees is incomplete lineage sorting (ILS), which can be modelled by the multi-species coalescent.

Species tree can be reconstructed using different methods. The most accurate but computationally most intense is *BEAST which co-estimates gene trees and species trees. However, for larger datasets (including more than 50 samples and/or 25 genes, i.e., typical phylogenomic datasets) this approach is not feasible as the datasets will not converge in a reasonable time.

Recently, several methods that are feasible with large phylogenomic datasets including hundreds to thousands of loci appeared. The mostly used so-called **summary methods** are: (1) coalescent summary methods ASTRAL (Accurate Species Tree Reconstruction ALgorithm) and ASTRID (Accurate Species TRees from Internode Distances), (2) MP-EST, (3) supertree method using matrix representation with likelihood (MRL).

The input to **ASTRAL** is a set of unrooted gene trees and ASTRAL finds the species tree that agrees with the largest number of quartet trees induced by the set of gene trees. It gives weights to all three alternative quartet topologies according to their relative frequencies within gene trees. It means that if the most frequent quartet topology is much more frequent, trees that do not include this topology are penalized. If the three alternative quartet topologies all have similar frequencies (i.e., close to 0.33), that quartet will only little

contribute to the optimization. ASTRAL also provides (1) a quartet-based support for each branch as a local posterior probability that the branch is in the species tree, (2) the length of internal branches in coalescent units.

Maximum representation with likelihood (**MRL**) is a supertree methods which estimate species tree on full taxon sets from sets of smaller trees (i.e., with missing species). First it encodes a set of gene trees by a large randomized matrix (the "MRL matrix") over {0, 1, ?} (e.g., using `mrp.jar`; https://github.com/smirarab/mrpmatrix): for a given edge (branch) in a given gene tree, the column in the matrix has entries over {0,1, ?}, with '0' given for the taxa that are on one side of the edge, '1' for the taxa on the other side, and '?' for all the remaining taxa (i.e., the ones that do not appear in the tree). This is done for all edges in all gene trees. As a next step the MRL matrix is analyzed using heuristics for a symmetric 2-state Maximum Likelihood (implemented , e.g., in RAxML as 'BINCAT' model).

Test dataset

85 alignments ('alignments') and corresponding gene trees ('trees') constructed using RAxML from the family Zingiberaceae (6 species). Gene trees are also provided in a single file 'geneTrees.tre'. Concatenated alignments are in 'concatenated.phylip'.

Task 1: reconstruction of species tree using ASTRAL

Astral is a java-based (https://github.com/smirarab/ASTRAL/raw/master/Astral.4.10.0.zip). Extract the zip file and put the jar file and 'lib' directory to the folder with the file 'geneTrees.tre'. Go to command-line and write the following command:

```
java -jar astral.4.10.0.jar -i geneTrees.tre -o speciesTree.tre
```

Open the 'speciesTree.tre' in a tree viewer (e.g., FigTree) and look at branch support values (local posterior probabilities).

Task 2: reconstruction of species tree using MRL

Download java-based program MRP (https://github.com/smirarab/mrpmatrix/blob/master/mrp.jar). Put the jar file to the folder with the file 'geneTrees.tre'. Go to command-line and write the following command:

```
java -jar mrp.jar geneTrees.tre MRLmatrix.phylip PHYLIP -randomize
```

A matrix called 'MRLmatrix.phylip' is created.

Run RAxML with BINCAT model and 100 fast bootstrap replicates:

```
raxmlHPC -f a -s MRLmatrix.phylip -n MRL.tre -m BINCAT -p 1234 -x 1234 -N 100
```

Compare the resulting tree (topology and support values) with ASTRAL tree.

Task 3: species tree from concatenated dataset using FastTree

Compute a maximum likelihood tree from concatenated dataset using FastTree by typing:

```
fastree -nt concatenated.phylip > concatenated.tre
```

Compare the topology and branch support values with ASTRAL and MRL trees.

Task 4: calculating alignment and tree properties using R

It is important to select only 'good genes' for final phylogenomic species tree reconstruction. These 'good genes' can be selected based on various criteria. Using R script (`tree_props.R`) we can calculate for each locus: average bootstrap support, average branch length and saturation potential. The saturation potential is calculated as a simple linear regression on uncorrected p-distances against inferred distances, i.e., tree distance (slope and R^2; higher values of R^2 mean lower saturation potential, slope higher than 1 indicates saturation).

Select loci that are not saturated and have average support values over 70%. How many 'good loci' you got? Recalculate species trees (ASTRAL, MRL, concatenated) and compare topology and support values.