CONCATENATION AND SPECIES TREE METHODS EXHIBIT STATISTIALLY INDISTINGUISHABLE ACCURACY UNDER A RANGE OF SIMULATED CONDITIONS

Joao Tonini, Andrew Moore, David Stern, Maryia Scheglovitova, Guillermo Orti
TREE CONCATENATION

Differences between gene tree and species tree
Average of gene trees = true species tree
Incomplete Lineage Sorting
Gene Duplication
Lateral Gene Transfer
Traditional Supermatrix approaches
SPECIES TREE METHODS

BATWING
BEAST
BEST
BUCK
BUCKy
GLASS
iGLASS
MCMCcoal
MDC
MP-EST
NJst
SNAPP
FINDINGS OF PREVIOUS STUDIES

There exists species trees for which discordant gene trees are more likely than genealogies that agree with the species tree.

- Species trees containing anomaly zones, polytomous gene trees are more probable than anomalous gene trees

- Short deep branches, gene trees are more likely to be just uninformative.

May still be adequate for densely sampled data matrices under non extreme rates of change

Where is the evidence?

- Studies proving inconsistencies usually do not prove superiority.

- Not reflective of typical empirical studies

Mutational Variance

Coalescent Variance
Discord is attributed to *mutational and coalescent variance*

Paper concluded that the accuracy of species tree estimation differs systematically depending on:
- the timing of divergence
- the sampling design
- the method used for species—tree estimation

Using more information contained in gene trees aside from topology such as branch lengths does not translate to gains in accuracy.

Accurate species-tree estimation should be dependent on the relative impact of mutational and coalescent variance.

Difficulty in estimating impact of mutational variance in the context of species tree estimation.
COMPARISON BETWEEN MDC, STEM, AND CONCATENATION

1) Generating a species tree under a **uniform speciation** model
2) Simulating coalescent gene trees for each species tree
3) Simulating DNA sequences under a specified model of nucleotide evolution along the branches of each gene tree
4) Estimating gene trees from the simulated DNA matrix
5) Estimating species trees from the estimated gene trees using **MDC and STEM**
6) Calculating **Discord** between both trees

Simulations do not estimate gene trees separately but estimate directly from concatenated matrixes. Comparisons from original Huang paper with identical parameterized simulations under concatenation. MDC and STEM were originally chosen for their inputs being gene trees.
SIMULATION

50 species tree of eight taxa from original paper

540 coalescent gene trees for each of the 50 trees under a neutral coalescent model, constant population size, and no migration using the same script.

1N and 10N generations for gene tree depth

Increasing the number of loci in the data matrix to obtain the true tree with matrices of 3, 9, and 27 genes.

540 genes were concatenated into respective matrices producing 180:3 matrices: loci, 60:9, 20:27

Typical single clock-like rate of sequence evolution across the tree.

\[
HKY = \begin{pmatrix}
- & \pi_C \beta & \pi_G \alpha & \pi_T \beta \\
\pi_A \beta & - & \pi_G \beta & \pi_T \alpha \\
\pi_A \alpha & \pi_C \beta & - & \pi_T \beta \\
\pi_A \beta & \pi_C \alpha & \pi_G \beta & -
\end{pmatrix}
\]

Each gene tree, SEQ-Gen produces 1000 base pairs

HKY model of nucleotide substitution to model evolution

Transition and transversion rate ratio of 3.0

Gamma distributed rate heterogeneity shape parameter of 0.8

Dirichlet distributed nucleotide frequencies in accordance with Huang.
MRBAYES AND R PHANGOM

Ran for each tree estimation until std. deviation < .01
Discarding the first 25% of the posterior results as burn in, 100 trees were sampled from posterior distribution to create the **majority rule consensus tree**

50% majority rule consensus with compatible clades
- Potential problem in clades with **polytomous** clades
- Manual inspection of subset of estimated species tree found no polytomous clades

Effectiveness of majority rule consensus tree?
**RF statistic** – measuring distance between two unrooted trees for 180, 60, and 20 concatenated matrixes
8-taxon matrices significantly smaller than matrices of most molecular phylogenetic analyses

Table 2. Results of ANOVA tests for differences in accuracy among methods (concatenation, STEM, MDC) with different numbers of loci and between methods with the same number of loci

<table>
<thead>
<tr>
<th>3 loci</th>
<th>9 loci</th>
<th>27 loci</th>
<th>Concatenation</th>
<th>MDC</th>
<th>STEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1N</td>
<td>&lt;0.001*</td>
<td>0.0255*</td>
<td>&lt;0.0001*</td>
<td>&lt;0.0001*</td>
<td>&lt;0.0001*</td>
</tr>
<tr>
<td>10N</td>
<td>0.3601</td>
<td>0.8233</td>
<td>0.8505</td>
<td>&lt;0.0002</td>
<td>&lt;0.0001*</td>
</tr>
</tbody>
</table>
```python
# Define variables from program arguments
infolder = sys.argv[1]  # Argument 1 = input folder
try:
    # Test for the presence of an optional output folder argument
    assert len(sys.argv) == 3
    outfolder = sys.argv[2]  # Argument 2 = output folder
except AssertionError:
    print("You did not provide an output directory. Please rerun with an output option.")
    exit(1)

# Perform action for files with common SETs

# This is an optional block of code to run Mr. Bayes for the concatenated Nexus files. Uncomment if you want.
NEXUSlist = [x for x in os.listdir(infolder) if ".nex" in x]
with open(infolder, "r") as infolder:
    for NexusFile in NEXUSlist:
        os.system(f"{Mb x} {NexusFile}")  # Run Mr. Bayes on the Nexus file
        print("Mr. Bayes ran successfully from file %s" % NexusFile)

# Another set of operations for processing Nexus files
```
PHYLOGENETICIST’S TOOLBOX: “NULL” METHODOLOGY

Concatenation can outperform well or better than methods that attempt to account for sources of error:

- ILS is low
- Few loci are used
- Gene trees have low phylogenetic signal

Regions of disparity between methods could result from unknown biological factors or critically violated assumptions in either method. Difficult for researchers to know to what extent discordance among gene trees may be due to methodological or sampling error. Criticism for lack of accountability in error due to ILS is reflected by shortcut coalescence methods assuming all error is ILS. Concatenation has practical merit for avoiding making assumptions to which sources of uncertainty influence the evolutionary history of the studied taxa.

Concatenation exhibits greater power to overcome sampling error and discrepant patterns of homoplasy and “concatalescence” methods can make use of concatenation in improving quality of tree inputs prior to modern coalescence based estimation of species tree.
“...EXHIBITS STATISTICALLY COMPARABLE ACCURACY UNDER A RANGE OF SAMPLING AND TREE DEPTH CONDITIONS VIS-A-VIS SOME EXISTING SPECIES TREE METHODS, AND URGE MOLECULAR PHYLOGENETICISTS TO THOROUGHLY EVALUATE THE PERFORMANCE OF METHODS THAT MODEL GENE TREE-SPECIES TREE DISCORD AGAINST CONCATENATION”
QUESTIONS?