Connecting Algebraic Geometry to Phylogenies via Singular Value Decomposition

Undergraduate Researchers: Wendi Chen, Sijia (Scarlett) Huo, Pengzheng Zhang, Ruizhe (Rex) Zhou
Graduate Mentor: Erin Molloy
Faculty Mentor: Dr. Ruth Davidson

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Introduction

GENERAL PROBLEM

Estimate a species tree from DNA data.

Concepts

- *Phylogenies are mathematical models describing the evolutionary history of a set of species. A phylogeny on n species is a binary tree, i.e., a tree for which each node has at most two children, with n leaves.*
- *Species trees describe the evolutionary pathway of species.*
- *Gene trees describe the substitution process of DNA sequences.*
- *One common method for species tree estimation is to estimate gene trees from DNA sequences and then estimate the species tree from the set of gene trees.*
- *As estimated gene trees may have considerable error, SVDQuartets, which bypasses gene tree estimation, is considered an exciting new method for species tree estimation.*

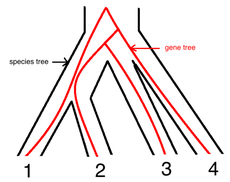


Figure: Gene Trees VS. Species Trees

Goal

Simulate various types of data to test the robustness of the SVDquartets method and to figure out under which circumstance the method has the best performance.

Background

Three possible unrooted trees exist on 4 species $\{1, 2, 3, 4\}$: 12|34, 13|24, and 14|23. We wish to identify which of the three quartets exists in the species tree.

Method Overview

SVDQuartets [1, 2] is a method for finding species quartet trees from DNA sequences. For each possible quartet,

- *Compute the flattening matrix, $\hat{\mathbf{P}}$, which is a 16×16 matrix that approximates the probability distribution of site patterns in DNA.*

$$\hat{\mathbf{P}} = \begin{pmatrix} \hat{p}_{AA|AA} & \hat{p}_{AA|AC} & \cdots & \hat{p}_{AA|TT} \\ \hat{p}_{AC|AA} & \hat{p}_{AC|AC} & \cdots & \hat{p}_{AC|TT} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{p}_{TT|AA} & \hat{p}_{TT|AC} & \cdots & \hat{p}_{TT|TT} \end{pmatrix}$$

Example: Consider the DNA data on 4 taxa $\{1, 2, 3, 4\}$ below. For all three possible quartet trees, $\hat{p}_{AA|AA}$ = 1/5. For the quartet tree defined by 12|34, $\hat{p}_{CC|GG}$ = 1/5; however, $\hat{p}_{CC|GG}$ = 0 for the other two quartet trees defined by 13|24 and 14|23.

1	A C A A C
2	A C A G C
3	A G A C T
4	A G G C T

- *Compute the SVD score, which measures the distance to the nearest rank 10 matrix in the Frobenius norm by Eckart-Young theorem:*
  SVDscore$(\hat{\mathbf{P}}) = \sqrt{\sum_{i=11}^{16} \sigma_i^2}$
- *Select the quartet tree with minimum SVDscore.*

Methodology and Results

Overview of the Simulation Study

1. Simulate species trees.
2. Simulate gene trees from species trees using **SimPhy** [3].
3. Simulate DNA sequence data from gene trees using **INDELible** [4].
4. Estimate species trees using SVDquartets.
5. Compare estimated species trees to true species trees.

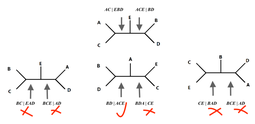
Features

- ***Simphy*** *is a software that can generate gene trees and species trees based on the parameters you give it. It can handle various aspects of the evolutionary process.*
  Example 1*: ./SIMPHY -rs 10 -rl F:50 -rg 1 -si F:1 -sl F:6 -sb F:0.000001 -sp F:2000000 -v 6*
  The letters after the slash specifies the properties controlled by the arguments, for instance, "-rs" tell the number of replicates, -rl and -rg control the number of species and the number of gene trees generated from the species tree.
  Specific definitions of the each command can be find in https://github.com/adamallo/SimPhy
  Example 2: *./SIMPHY -rs 10 -rl F:1000 -rg 1 -st F:2000000 -si F:1 -sl F:5 -sb F:0.000001 -sp F:200000 -hs LN:1.5,1 -hl LN:1.2,1 -hg LN:1.4,1 -su E:10000000 -so F:1 -od 1*
  The above command is based on Dr. Davidson's script, which adds commands to control more factors, for example "-so" to simulate outer group who's distance is relatively far from other groups. It also adds commands to control the branch lengths to help us analyze the accuracy of SVDQuartets method.
- ***INDELible*** *is a tool to simulate biological sequence evolution. We use this to generate the DNA sequences for each gene tree generated by Simphy.*
  Example:
  [*TYPE*] *NUCLEOTIDE 1 ← the name (choose from NUCLEOTIDE, AMINOACID and CODON) controls the type of simulation and the number (1 or 2) decides algorithm to use.*
  [*TYPE*] *command.*
  [*MODEL*] *modelname*
  [*submodel*] *JC ← we used JC69 model and don't use any model involves insertion and deletion for we want to test simple cases first.*
  [*TREE*] *treename*
  ((5_{00} : 15330942.24951079, (2_{00} : 10089899.29155450, ((4_{00} : 6020540.19650047, 3_{00} : 6020540.19650047) : 3255117.93412819, 6_{00} : 9275658.13062866) : 814241.16092584) : 5241042.95795629) : 4284184.56241053, 1_{00} : 19615126.81192132);
  ← that's the gene tree file we got before, here we used a outcome from the first example.
  [*PARTITIONS*] *partitionname*
  treename modelname 1000 ← controls length of sequence generated and model used on each tree
  [*EVOLVE*] *partitionname 1 outputname ← controls the output name and number of replicates.*

Accuracy

Number of Correct Quartets is used to assess the accuracy of SVDQuartets in this study. This is the number of quartets from the estimated tree that exist in the true tree.

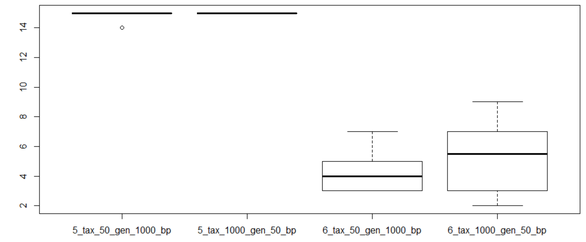
Robinson-Foulds Distance quantifies the distance between a simulated and estimated unrooted tree by summing splits in the true tree that are not in the estimated tree (false negatives) and splits in the estimated tree that are not in the true tree (false positives).



For example, The top center tree is the tree used in all comparisons. Since this unrooted tree contains two non-trivial branches, we have two possible splits. By comparing each split in the top tree to each split in the three simulated trees, the error is 4, 2, and 4, respectively.

Result

The box plots below shows the number of correct of quartets across 10 replicates for each model condition. This first two box plots correspond SimPhy Example 2, and the second two boxplots correspond to SimPhy Example 1. The means between these model conditions significantly different, 14.9 to 4.2 (out of 15 possible quartets). For each example, there are also boxes that are corresponding to '1000 gene tree and 50 base pairs each', which give the same total amount of information as '50 gene tree and 1000 base pairs each'. The former models do perform better, which is expected due the properties of the SVDquartet method.



Hamming Distance In information theory, the Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different.

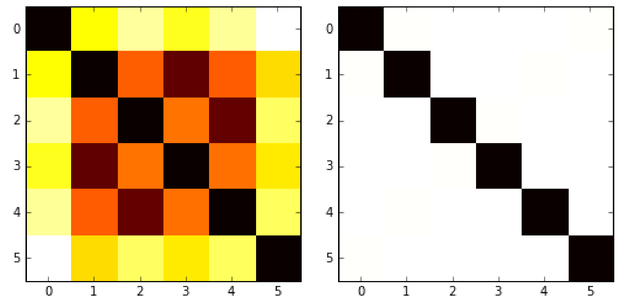


Figure: SimPhy Example 2 and SimPhy Example 1

The heat plots above show the Hamming distance between each species for two trees. The first box plot is corresponding to the left heat plot above and the the third box plot is corresponding to the right heat plot above. It is easy to see that the differences between Hamming distances between species in the Example 1 are much larger than that of Example 2. Therefore will not be surprising to see that the Example 1 will be more consistent with the true species tree than the species tree from Example 2.

References

[1] J. Chifman and L. Kubatko. Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics*, vol. 30, pp. 3317–3324, 2014.
[2] J. Chifman and L. Kubatko. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *Journal of Theoretical Biology*, vol. 374, pp. 35–47, 2015.
[3] Mallo D, de Oliveira Martins L, Posada D. SimPhy: Phylogenomic Simulation of Gene, Locus and Species Trees. *Systematic Biology*, vol. 65 pp. 334–344, 2015.
[4] W. Fletcher and Y. Ziheng. INDELible: A Flexible Simulator of Biological Sequence Evolution, *Molecular Biology and Evolution*, vol. 26, pp. 1879–1888, 2009.

External Funding

This project is based upon work supported by the National Science Foundation under Grant Number DMS-1449269. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Support for this project was provided by the Illinois Geometry Lab and the Department of Mathematics at the University of Illinois at Urbana-Champaign.