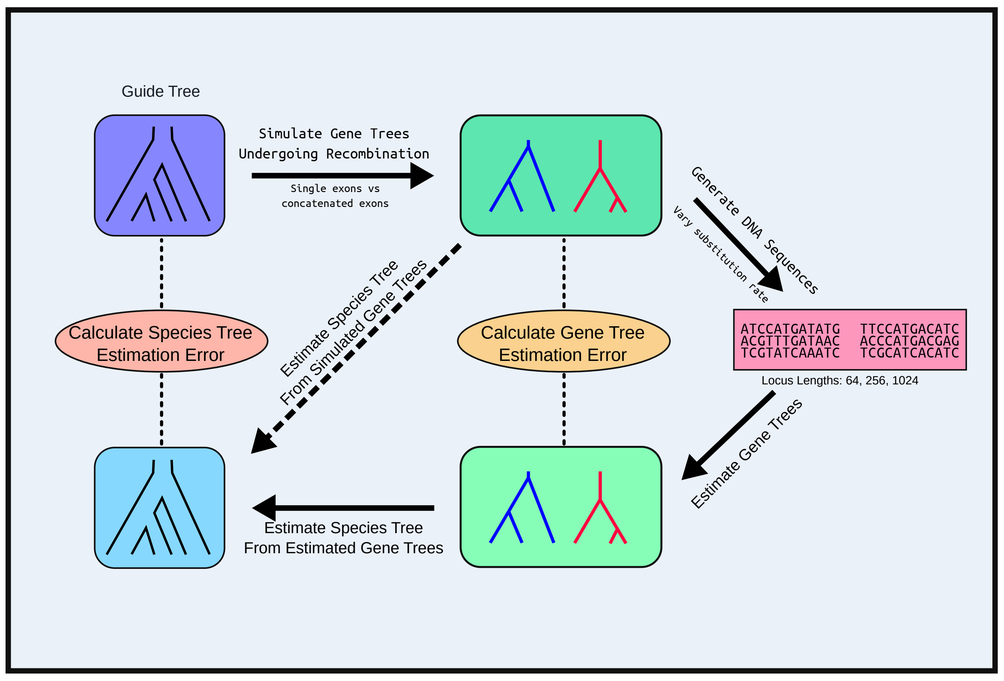
Gene Tree Estimation Accuracy Under Varying Locus Lengths, Divergence Times, and Substitution Rates

Michael Conry Alan R Lemmon

Abstract

One of the central goals of evolutionary biology is to understand the evolutionary relationships among organisms by constructing phylogenetic estimates, commonly known as evolutionary trees. The accuracy of phylogenetic estimates can be strongly affected by the particular evolutionary processes that are taken into account during an analysis. One important process, genetic recombination, has been shown to lead to inaccurate phylogenetic estimates when ignored. In this novel simulation study we explore a subset of important parameters (locus length, species divergence time, and substitution rate) that require careful consideration during the preparation of DNA alignments used to estimate gene trees. Gene trees resulting from this simulation are compared to the true trees using the Robinson-Foulds (RF) metric to determine the extent of the discordance caused by short loci that do not contain sufficient information about the evolutionary history of a set of taxa, shallow divergence times that lead to incomplete lineage sorting (ILS), and extreme DNA substitution rates that exemplify species relationships that are difficult to resolve. This research is part of a larger study that examines the extent to which neglecting to accommodate for recombination in data partitioning approaches can lead to inaccurate species trees estimates. We intend to provide recommendations to empirical researchers as to when it is most beneficial to treat exons as independent evolutionary units in phylogenetic analyses.

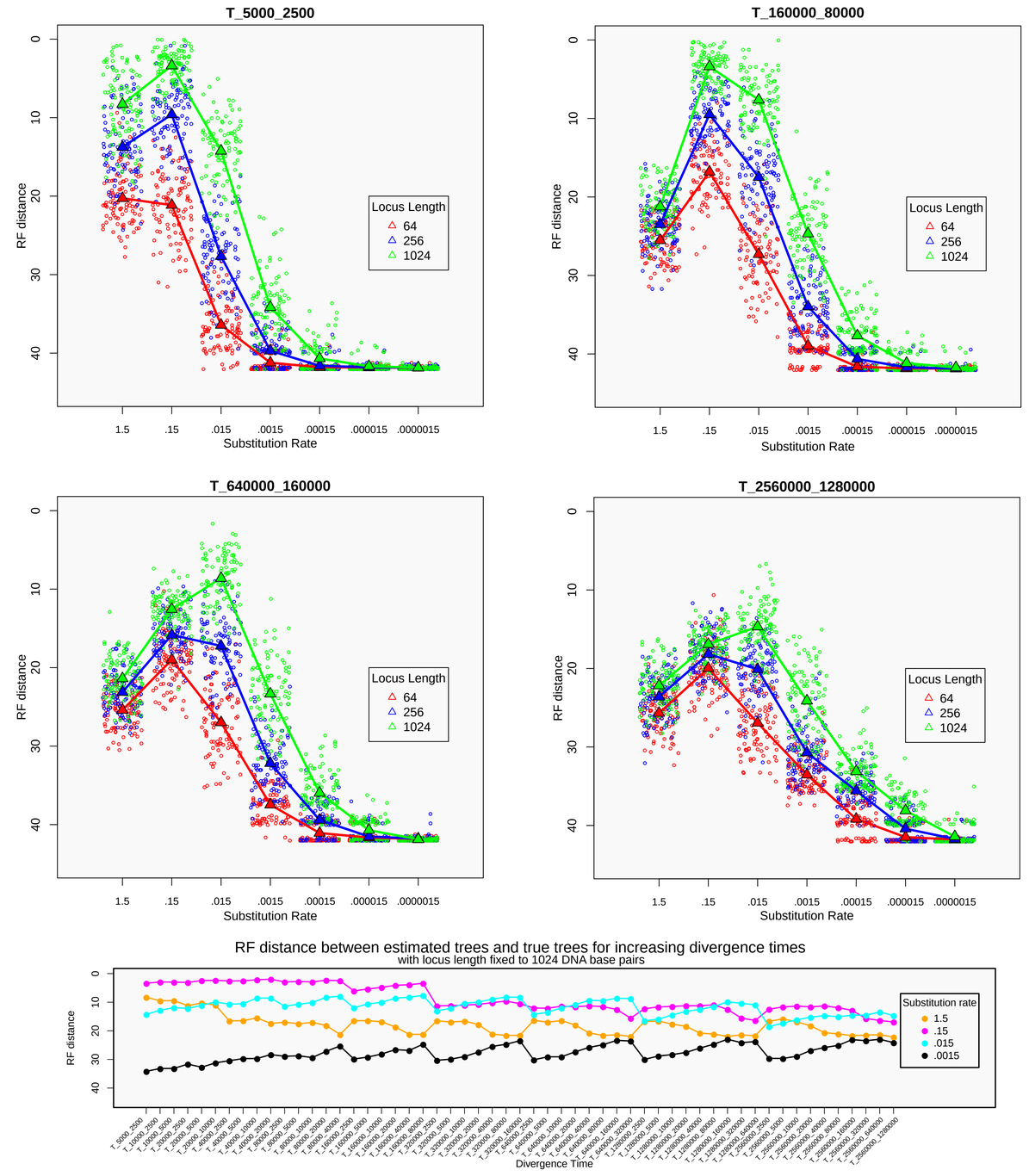
Workflow



The simulation study begins with a 3 species guide tree that is used to control divergence times in MS, a backwards-in-time coalescent simulator. The simulator generates 24-taxa gene trees which are separated into 12 exons and, depending on the recombination parameter, the model may permit recombination to occur between each of them. We convert the simulated gene trees into simulated DNA alignments using SeqGen, a DNA simulator with a choice of substitution models. Using the simulated alignments we estimate gene trees via the maximum likelihood gene tree estimation software, RAxML. Species trees are then inferred from the estimated gene trees using Astrid, a multi-species coalescent summary method.

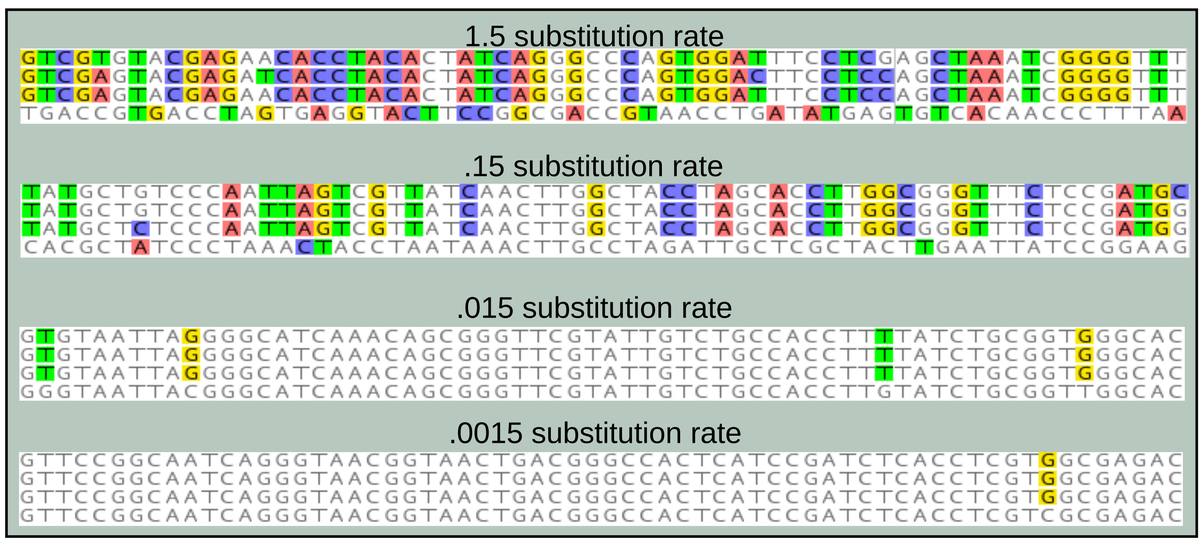
The goal is to determine if the final species tree matches the initial species tree that was used as a guide tree for our simulations. If the species trees are discordant we should be able to determine what parameter conditions caused the discordance. Although recombination in the presence of ILS may lead to inaccuracy in the species tree estimation, it is also possible that the inaccuracy is caused by poor substitution model parameter choice or insufficient locus length when simulating DNA sequences. To determine if model parameters have caused species tree discordance we will also estimate a species tree directly from the simulated gene trees, bypassing the DNA sequence generation and the possibly incorrect substitution model assumptions.

RF Distances



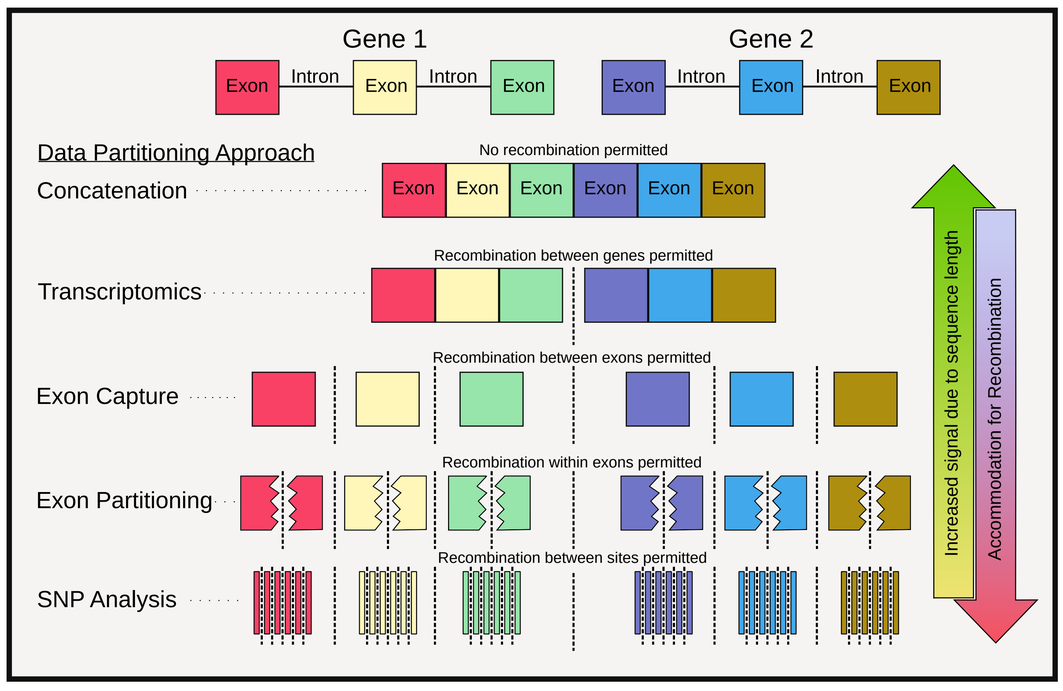
The top four graphs demonstrate the changes locus length and substitution rate transmit to gene tree estimation accuracy. The RF distances between the estimated trees and the true trees are minimized when the simulated loci have a substitution rate of .15 for shallow trees and .015 for deep coalescent times. This variance is intuitive because a short coalescence time will restrict the chance for the necessary ancestral information to arise in the form of substitution events. Consequently, the longer locus lengths shown in the top graphs both permit a lower substitution rate and help to increase gene tree accuracy by providing more ancestral information content. The question then becomes: What analysis approach can systematists employ to acquire longer sequences while still accommodating empirical assumptions like the presence of recombination?

Simulated Alignments



Gene tree inference models rely on DNA alignments for input that are required to contain enough substitution events among their sequences to allow for proper inference of relationships between taxa, but not so much substitution that all similarities between taxa are lost. This raises the question: How much DNA substitution is appropriate to achieve an accurate result from gene tree estimation for a given locus length? Increasing the length of a locus allows for more information content, and consequently can significantly offset the problems introduced by low substitution rates. Above are 4 sequence alignments of length 64 that were generated by a DNA alignment simulator, Seq-Gen, using various substitution rates. The extreme substitution rates often lead to inaccurate gene tree estimates. We can quantify the inaccuracy by first simulating alignments of varied lengths and substitution rates based on evolutionary histories with varying divergence times, then inferring gene trees from these alignments, before finally comparing the gene trees to the true trees using the RF distance metric.

Data Partitioning Approaches



Although recombination within gene segments occurs and may negatively affect accuracy, most researchers choose to ignore intra-segment recombination due to the computational challenges encountered when accounting for recombination within these segments. The question then arises: At what point does the advantage of having a longer gene segment outweigh the negative effects of unaccounted for recombination within the segment? This question becomes especially important in phylogenetic studies relying on transcriptome data, in which the exons contained within each RNA transcript may be separated by long genomic distances across which recombination is more likely to have occurred. In these studies, researchers are faced with the question: when should exons be concatenated or treated as separate evolutionary units?

The figure above shows multiple ways to prepare exon data for phylogenetic analysis. Before transcription, exons are separated by introns that are typically much longer in comparison which leads to the assumption that recombination is more likely within introns. However, that assumption can be incorrect, and can lead to inaccurate species trees. The dashed vertical lines show locations where recombination is permitted depending on the method of analysis. While possible locations for recombination events increase from concatenation to site-based methods, sequence length is reduced. As we have demonstrated, longer sequences are causally related to greater accuracy in gene tree estimation.

References

Hudson, R.R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-338.

Lanier, H.C. and L.L. Knowles. 2012. Is Recombination a Problem for Species-Tree Analyses? Sys. Bio. 61:691-701.

Rambaut, A. and N. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235-238.

A. Stamatakis: "RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies". In Bioinformatics, 2014