IIB 290- Species Tree Inference- Annotated Bibliography
24 October 2019

Chou, Jed, et al. "A comparative study of SVDquartets and other coalescent-based species tree estimation methods." BMC genomics 16.10 (2015): S2.

- (Methods)
- SVDquartets, ASTRAL-2, NJst are three coalescent-based summary statistic methods. ASTRAL-2 (and similar methods) are known to be vulnerable to short gene sequences. SVDquartets is designed specifically to deal with short gene sequences. Using a series of simulated datasets, this paper employs maximum likelihood to compare the above three methods with each other and with the concatenation method. Different ILS levels, numbers of taxa, and number of sites per locus are explored, including gene sequence alignments as short as 10 sites per locus. The results show that ASTRAL-2 provides the most accurate results under higher ILS conditions. SVDquartets is competitive under lower ILS and small numbers of sites per locus, and concatenation has the best accuracy under the lowest ILS conditions.

Degnan, James H., and Noah A. Rosenberg. "Gene tree discordance, phylogenetic inference and the multispecies coalescent." ***Trends in ecology & evolution*** **24.6 (2009): 332-340.**

- (Foundation/ Concepts)
- This review paper highlights the complexities of genealogical discordance in phylogenies inferred using multilocus genomic data. It describes the conceptual basis for gene tree discordance, discusses its implications and how to deal with them when inferring species trees. The paper also reviews the issues that the tree inference methods must address to successfully account for the variability in the genome. It explains why applying more data (as in the consensus and concatenation methods) can be more likely to result in an incorrect species tree. The article includes a glossary and a proposed list of questions for framing a multilocus phylogenetics investigation.

Liu, Liang, et al. "Modern Phylogenomics: Building Phylogenetic Trees Using the Multispecies Coalescent Model." Evolutionary Genomics. Humana, New York, NY, 2019. 211-239.

- (Methods)
- This book chapter introduces the multispecies coalescent (MSC) model as a framework for building phylogenetic trees from multilocus DNA sequence data. It starts with the initial MSC methods, which combine information from multiple gene trees into a single “supergene”, but as more gene trees and sequence data are added, the potential for converging on the incorrect species tree increases. Then comes the second-generation implementations of the MSC, employing Bayesian or likelihood models. These remain consistent in all regions of gene tree space, but Bayesian methods in particular are incapable of handling the large phylogenomic data sets. So the Two-step methods are introduced, where gene trees are first estimated and then combined to estimate an overarching species tree. These methods (e.g. MP-EST, ASTRAL) can handle large phylogenomic data sets but sometimes provide inappropriate measures of tree support which is verified with a likelihood ratio test. The LRT is regarded as a useful alternative to the multilocus bootstrap, which only indirectly tests the appropriateness of competing species trees.

Liu, Yang, et al. "Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear genomes." Nature communications 10.1 (2019): 1485.

- (Application)
- This paper is an application of species tree inference as well as concatenation methods -- they use both! In fact, reading the methods of this paper makes it seem like the authors did absolutely every method out there. They didn’t, but it is an impressive amount of work. The part that is most interesting for this discussion is the comparison of the results from the different methods and data sources, as well as the discordance in the gene trees. I find this to be a good example of how diving deeper into discordance, rather than treating it like a problem to deal with, can provide extra insight into the process that led to your study group. For example, the authors noted that, when looking at concatenation trees of the three genomes (nuclear, chloroplast, and mitochondria), the nuclear and mitochondrial genomes agree on one topology of a

particular splitting order and the plastome another. Looking at these histories separately allows the authors to hypothesize that the discordance is likely not from hybridization as the organellar genomes would more likely have the same history, in that case. Similarly, when looking at the gene tree discordance using ASTRAl-II, most gene trees agree with the nuclear-mitochondrial topology, but there is a runner up and that is the same as the plastome history. The point I want to get across here is not the intricacies of moss evolutionary history, but that gene tree discordance is not (necessarily) noise in the data that has to be overcome: it may well represent real evolutionary history and only by looking into it with species tree methods can you begin to see what that history might have been.

Maddison, Wayne P., and L. Lacey Knowles. "Inferring phylogeny despite incomplete lineage sorting." Systematic biology 55.1 (2006): 21-30.

- (Foundation/ Concepts)
- This paper acknowledges the phylogenetic difficulties due to lineage sorting and explores some approaches to overcome them. The authors examine how the reconstructability of a species phylogeny is affected by (a) the number of loci used to estimate the phylogeny and (b) the number of individuals sampled per species. Even in difficult cases with considerable incomplete lineage sorting, the authors found the reconstructed species trees to match the "true" species trees in at least three out of five partitions, as long as a reasonable number of individuals per species were sampled. They also examined the tradeoff between sampling more loci versus more individuals and concluded that increasing the number of loci gives more accurate trees for a given sampling effort with deeper species trees, while sampling more individuals often gives better results than sampling more loci with shallower species trees. Taken together, the authors conclude that gene sequences retain enough signal to achieve an accurate estimate of phylogeny despite widespread incomplete lineage sorting.

Mallo, Diego, and David Posada. "Multilocus inference of species trees and DNA barcoding." *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.1702 (2016): 20150335.

- (Methods)
- A methods review paper that compares models that account for ILS and other challenges in multilocus data. The methods comparison figure from the slides is from this paper.

Mirarab, Siavash, and Tandy Warnow. "ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes." Bioinformatics 31.12 (2015): i44-i52.

- (Methods)
- Given that different loci can have different trees (due to incomplete lineage sorting), inferring species phylogenies requires a dataset of multiple loci modeled by the multi-species coalescent model. Astral is a statistically consistent coalescent based method available in open source that can handle large datasets. Astral II is the newest version: faster, can analyze much larger datasets (up to 1000 species and 1000 genes) and has substantially better accuracy under some conditions.

Nakhleh, Luay. "Computational approaches to species phylogeny inference and gene tree reconciliation." Trends in ecology & evolution 28.12 (2013): 719-728.

- (Review)
- Another good review paper on how gene trees reflect the evolutionary processes that act on them within and across populations and how this can be used to elucidate information about those processes as well as to infer accurate phylogenies. It builds on Maddison's parsimony and likelihood approaches to reconciling gene trees and species trees (discussed by Maddison 16 years prior to the publication of this article) and highlights the contributions of computational biology and bioinformatics since then in devising methods for detecting and resolving incongruences.

Ogilvie, Huw A., et al. "Computational performance and statistical accuracy of* BEAST and comparisons with other methods." ***Systematic biology*** **65.3 (2016): 381-396.**

- (Methods)
- *BEAST, a fully Bayesian implementation of the multispecies coalescent model, is the subject of this paper. Using simulation studies, the authors examine the impact of the number of loci on both computational performance and statistical accuracy and compare them with summary methods. They show that the statistical performance of *BEAST relative to concatenation improves both as branch length is reduced and as the number of loci is increased. In addition, a performance comparison of the species tree and concatenation methods shows that using *BEAST with tens of loci can outperform using concatenation with thousands of loci.

Pamilo, Pekka, and Masatoshi Nei. "Relationships between gene trees and species trees." ***Molecular biology and evolution*** **5.5 (1988): 568-583.**

- (Foundation/ Concepts)
- This paper studies the role of ancestral genetic polymorphism in the difference seen between a gene tree and its corresponding species tree. the number of nucleotides examined on the topology of an estimated tree.

Rothfels, Carl J., et al. "The evolutionary history of ferns inferred from 25 low-copy nuclear genes." American Journal of Botany 102.7 (2015): 1089-1107

- (Application)
- An application of a species tree MSC summary method (ASTRAL) with the explicit goal of testing whether differences in topology from previous analyses (nuclear vs. plastome) were due to ILS. The authors found evidence to support the hypothesis that ILS may have caused reduced support on some nodes of the tree from concatenated supermatrix data but the plastome is less sensitive to ILS and, thus, had higher support.

Wang, Kun, et al. "Incomplete lineage sorting rather than hybridization explains the inconsistent phylogeny of the wisent." Communications biology 1.1 (2018): 169.

- (Application)
- Another application paper testing explicitly for ILS when past phylogenies produced inconsistent topologies.