Annotated Bibliography: Species Trees and the Multispecies Coalescent

Carrie Tribble & Gabriel Damasco

Coalescent methods for estimating phylogenetic trees

Liang Liu, Lili Yu, Laura Kubatko, Dennis K. Pearl, Scott V. Edwards
Molecular Phylogenetics and Evolution 53.1 (2009): 320-328.

Here, the authors review different methods for resolving species trees from genetic data, including the full implementation of the multispecies coalescent, concatenation, and summary statistic methods. It summarizes the two likelihood functions involved in estimating a species tree using the full coalescent and the general principles behind the various summary methods. The paper summarizes the computational and statistical differences between the models and ends with a discussion of future directions for species tree estimation.

The BPP program for species tree estimation and species delimitation

Ziheng Yang
Current Zoology 61 (5): 854–865, 2015

This paper introduces the BPP implementation of the full multispecies coalescent. It reviews four scenarios under which the program can be used:

- A00: estimation of divergence times and effective population sizes on a fixed species phylogeny
- A10: estimation of a species tree given species assignment and species delimitation
- A01: species delimitation given a guide species tree
- A11: joint estimation of species delimitation and species tree

The paper goes through a brief tutorial on how to use the BPP program and provides examples for all four scenarios using the example of East Asian brown frogs. The paper ends with a discussion of the limitations of the basic multispecies coalescent, including the assumptions of no recombination within genes and no migration between species, and some practical suggestions for implementing the models. It's an interesting approach to publish a tutorial along with a more theoretical explanation of the model, but I think working through the tutorial is a useful way to truly understand the various uses of the multispecies coalescent.

ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes

Siavash Mirarab and Tandy Warnow
Bioinformatics, Volume 31, Issue 12, 15 June 2015, Pages i44–i52,
https://doi.org/10.1093/bioinformatics/btv234

This paper introduces the program ASTRAL-II, an upgrade of ASTRAL, a program developed to enable coalescent-based analyses of larger datasets. Despite the availability of coalescent-based methods, many biological datasets are too large for the available methods. For example, *BEAST (Heled and Drummond, 2010), a method that co-estimates gene trees and the species tree, cannot be used with more than about 25 species. ASTRAL solves this problem by constraining the allowed search space to those species trees that derive their bipartitions from a set of input gene trees provided by the user. However, ASTRAL's running time increased quickly for large datasets and that large sets of gene trees reduced the accuracy for species trees estimated by ASTRAL under certain model conditions. ASTRAL-II improved the running time asymptotically by a factor of the number of species, as well as it explores a larger search space. It also can handle polytomies in the input trees now. The authors compared ASTRAL to others coalescent-based species tree estimation methods and to concatenation using maximum likelihood. They showed that ASTRAL outperforms the coalescent-based methods handling 1000 species and 1000 genes in about a day. The comparison between ASTRAL and concatenation shows that ASTRAL is more accurate whenever the ILS level is sufficiently high and comes close to concatenation under very low ILS levels. Moe interestingly, their simulations show how the choice of the best method to use can often depend on the amount of gene tree error, number of genes and the level of discordance.

Estimating phylogenetic trees from genome-scale data

Liang Liu, Zhenxiang Xi, Shaoyuan Wu, Charles C. Davis and Scott V. Edwards
Annals of the New York Academy of Science, Volume 1360 Pages 36–53, 14 April 2015
http://onlinelibrary.wiley.com/doi/10.1111/nyas.12747/full

In this review paper, the authors showed theoretical and empirical examples that clarifies conflicts between species tree and concatenation methods, as well as misconceptions in the literature about the accuracy of species tree methods. Different methods of species tree inference incorporate different amounts of detail of the MSC process, and there is a trade-off between model accuracy and computational burden. Most species coalescent models (MSC) in phylogenomics assume simple models of instantaneous speciation with no gene flow after species divergence. Moreover, MSC assumes complete neutrality, no recombination within loci, and free recombination between loci, such that loci can be treated as independent neutral replicates conditional on the phylogenetic history of the lineages under study. In contrast, the problem with concatenation is the simplification on the number of parameters, because all gene trees in the concatenation model are treated as the same variable. Thus, the estimates of parameters in the concatenation model tend to have a smaller variance. Since small variance corresponds to high bootstrap support or posterior probability, overestimation of bootstrap support by concatenation methods is problematic. Although concatenation and species tree approaches often yield similar estimates of phylogeny, an increasing number of examples of strong conflict between concatenation and coalescent analyses shows that the conditions for conflict among methods are consistent with empirical data. The justification for species tree methods relies not in the ubiquity of gene tree heterogeneity in empirical data sets, but in their acknowledgement of fundamental genetic processes inherent in all organisms, including recombination along the chromosome, which renders gene histories independent of one another and conditional on the phylogeny; and genetic drift, which generates stochasticity in gene tree topologies and branch lengths. Thus, even when all gene trees are topologically similar, species tree methods can yield results differing from concatenation methods, if not in phylogenetic topology then often in phylogenetic support, because species tree methods better model the accumulation of signal that is accrued with increasingly large data sets.

A comparative study of SVDquartets and other coalescent-based species tree estimation methods

Jed Chou, Ashu Gupta, Shashank Yaduvanshi, Ruth Davidson, Mike Nute, Siavash Mirarab, and Tandy Warnow
BMC Genomics 2015 16 (Suppl 10):S2
https://doi.org/10.1186/1471-2164-16-S10-S2

This study compares the accuracy of some coalescent based summary statistic methods with concatenation under differing conditions of incomplete lineage sorting (ILS), loci size, and number of taxa. Summary methods are often used instead of the full implementation of the multispecies coalescent due to the computational cost of the full implementation (see Ogilvie et al. 2016). One common problem with summary methods when compared to concatenation is that longer gene regions commonly used in phylogenetics violate a key assumption of the multispecies coalescent model and thus the summary methods: that there is no recombination within the genes. Non recombining regions are often small, fewer than 100 base pairs, and thus gene tree uncertainty can be very high. They surprisingly found that SVDquartets, which was designed explicitly to deal with short gene sequences by dealing with single-site patterns, was rarely preferred. Instead, ASTRAL-2 performed best under high ILS conditions and concatenation performed best under low ILS. Overall, this is a useful paper for evaluating the various summary methods and considering which to use under which circumstances.

Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics

Scott V. Edwards, Zhenxiang Xi, Axel Janke, Brant C. Faircloth, John E. McCormack, Travis C. Glenn, Bojian Zhong, Shaoyuan Wu, Emily Moriarty, Lemmonh Alan, R. Lemmon, Adam D. Leaché, Liang Liu, Charles C. Davis
Molecular Phylogenetics and Evolution Volume 94, Part A, January 2016, Pages 447-462, https://doi.org/10.1016/j.ympev.2015.10.027

This paper is exciting because it gives an elegant and incisive response to a recent article published in Molecular Phylogenetics and Evolution by Mark Springer and John Gatesy (S&G) called “The gene tree delusion”. S&G critiqued the data sets analyzed by

several recent papers, including McCormack et al. (2012), Song et al. (2012), and Xi et al. (2014) - second author in this annotated paper, and stated that those studies violate and misapply aspects of the MSC, resulting in unreliable phylogenetic trees (all criticisms listed in the introduction). Edwards et al. replied saying that several concepts and attempts to discredit MSC models were invalid and reveal numerous misunderstandings of the MSC. Although the authors critiqued by S&G acknowledged the errors in their data sets, many of these errors had already been pointed out in the literature and do not imply fundamental changes in their conclusions. In this long review, Edwards et al. showed that ILS indeed accounts for a substantial fraction of gene tree heterogeneity observed in the critiqued data sets, and emphasized that MSC models are not predicated on the existence of ILS, whose presence or absence, ultimately, is irrelevant to the application of the MSC model. Rather, it is the conditional independence of loci used to make gene trees, not the presence of ILS, that is the fundamental assumption of MSC methods. The paper then reviews the conditional independence based on first principles of genetics, such as recombination and the chromosomal structure of genomes, and discuss the role of different data type and model violations in simulations studies. Finally, the authors admitted that S&G have performed an admirable service by identifying errors in phylogenomic data sets, but their recommendations for phylogenomics and the provocative language used to proffer them represent fundamental throwbacks that do not advance the field. BANG!

Challenges in Species Tree Estimation Under the Multispecies Coalescent Model

Bo Xu and Ziheng Yang
GENETICS December 1, 2016 vol. 204 no. 4 1353-1368,
https://doi.org/10.1534/genetics.116.190173

This paper reviews the methods that estimates species tree under the MSC and explore the conceptual and statistical challenges. They use simulations to explore the differences between the two classes of species trees methods: the full MSC methods (which includes both maximum likelihood and Bayesian inference) and the summary methods (ones that use gene trees vs sequenced loci as input data). Unfortunately, simulations were used on simple cases involving three or four species, with the justification that most summary methods are based on small species trees (e.g., triplet and quartet trees). They conclude that summary methods are inefficient due to the information loss when estimated gene tree topologies are used a priori, ignoring

information in the branch lengths. A number of challenges exist with current implementations of the full MSC model under the Bayesian inference methods. A major problem is the intensive computation involved and the inefficient mixing of the MCMC algorithms. They finally suggest that the utility of summary vs. full methods will depend on the nature of the species tree estimation problem. For easy problems with long internal branches in the species tree and little incomplete lineage sorting, different methods are likely to produce the same results, and simple methods such as concatenation may have even higher statistical efficiency than coalescent-based full likelihood methods. For shallow species phylogenies, characterized by recent divergences and short internal branches (as occurs in radiative speciation), full likelihood methods may have a big advantage over summary methods or simple methods such as concatenation.

Multilocus inference of species trees and DNA barcoding

Diego Mallo and David Posada
Philosophical Transactions of Royal Society, Volume 371, issue 1702, 5 September 2016
http://rstb.royalsocietypublishing.org/content/371/1702/20150335?utm_source=TrendMD&utm_medium=cpc&utm_campaign=Philosophical_Transactions_B_TrendMD_0

This paper is part of the special issue called “From DNA barcodes to biomes”. This is very nice review for those who just started embracing phylogenomics and using next generation sequencing. In this review, they discussed the most important challenges to reconstruct phylogenetic histories and the pros and cons of using multilocus data in phylogenetic analysis. More interestingly, they described the evolutionary events that can result in discordance of species tree and gene tree and compared the most useful methods for species tree reconstruction. They also state that multilocus data could improve DNA barcoding analysis based on species-tree approaches. Species tree reconstruction methods rely either directly or indirectly on estimated gene trees and every condition able to mislead gene tree estimation (reviewed in this paper) will also affect final species tree accuracy. Regarding the use of high-throughput sequences, the authors brought the attention to data sets with non-randomly distributed missing data in combination with substitution-rate heterogeneity as the main cause of incongruence in species trees methods. There is also a short review on the most used programs to estimate species trees under the MSC model.

Computational Performance and Statistical Accuracy of *BEAST and Comparisons with Other Methods

Huw A. Ogilvie, Joseph Heled, Dong Xie, and Alexei J. Drummond
Systematic Biology, Volume 65, Issue 3, 1 May 2016, Pages 381–396.
https://doi.org/10.1093/sysbio/syv118

In this paper, the authors compare the performance and accuracy of various methods for species tree estimation, including a commonly implemented version of the full multispecies coalescent, *BEAST. They perform three tests using simulated data:

1) *Evaluate the performance of *BEAST as as the number of loci are increased*. They estimated the computational cost to analyze a set dataset with differing number of loci. For example, a tree with 5 species and 2 tips per species would take 369 CPU hours to run with 356 genes, but 1064 CPU **days** [emphasis added] with 1024 loci. The found that the relationship between increasing number of loci and decreasing error in species tree estimation was explained by a power law, such that adding more loci initially decreases error in species tree estimation but eventually the decrease in error is very small.

2) *Compare the accuracy of *BEAST to Bayesian supermatrix approaches*. The authors compared the accuracy of *BEAST to supermatrix approaches while varying the branch lengths of the trees in the simulations (because the importance of accounting for deep coalescence increases as branch lengths become smaller). They found that regardless of number of species, *BEAST outperformed Bayesian supermatrix estimation when branch lengths were small, though the specific value of branch lengths depended on the number of loci included.

3) *Compare parameters of two empirical datasets*. Here, the authors estimate parameters from two empirical datasets - one shallow tree (*Cyathophora*) and one deep tree (primates). They then run simulations based off of the estimated parameters and compare the accuracy of various estimation methods as functions of the number of loci. They show that *BEAST in the shallow simulations, *BEAST was the best-performing method, and the topological accuracy of both *BEAST and MP-EST was improved given two individuals per

species. However in the deep simulations, all methods other than *BEAST and MP-EST converged at near-zero topological error given 512 loci.

Ultimately, the authors conclude that the multispecies coalescent is most useful when branch lengths are small/ in estimating shallower evolutionary relationships. They also say that using *BEAST with many loci at the scale of phylogenomic data is feasible as long as you have access to supercomputers and know how to parallelize properly, but as soon as you start talking in units of CPU DAYS I get worried...

Multispecies coalescent delimits structure, not species.

Jeet Sukumaran and L. Lacey Knowles.
Proc Natl Acad Sci U S A. 2017 Feb 14;114(7):1607-1612.
doi: 10.1073/pnas

This paper gets into some fun controversy about the commonly implemented multispecies coalescent model BPP (Bayesian Phylogenetics and Phylogeography - see Yang 2015), and other possible implementations of the multispecies coalescent for species delimitation. Since the use of the multispecies coalescent for species delimitation is a literal application of the biological species concept, where a ‘species’ is defined by a group of interbreeding individuals along the branches of the tree, it has increasingly been used to identify cryptic species and has been used as the sole evidence to support newly described species. This paper points out that it is possible to identify population structure within species using BPP, and that relying solely on BPP and lack of ongoing gene flow may inflate our estimates of species counts. Furthermore, the authors point out that speciation is a gradual process, unlike how it is modeled in the popular birth-death process. To address these concerns, they simulate data under the protracted speciation model, which considers speciation to be gradual rather than immediate. They test the ability of BPP to detect ‘actual’ species rather than structured population on these simulated data. They find that BPP overestimates the number of species. Honestly, this all just seems like one long extended argument about what you consider a species rather than a critique of a statistical method. It also seems unfair to test the ability of a model to detect ‘species’ on data simulated under another model. In conclusion - it is definitely useful for phylogeneticists and systematists to remember that speciation is a long and complex process, but the specific critique of BPP doesn’t seem as relevant.