Handling discordance in phylogenetic inferences

Emily Jane McTavish

Life and Environmental Sciences
University of California, Merced
ejmctavish@ucmerced.edu, twitter:snacktavish

What are reasons that a 'gene tree' may not show the same relationships as the species tree?

- incomplete lineage sorting
- hybridization
- horizontal gene transfer
- gene tree error

How do you estimate relationships in the face of these processes?

How do you estimate relationships in the face of these processes?
It depends what question you are trying to answer!

Concatenation

Combine all genes or regions for each sample or taxon.

Advantages:

- Relatively fast
- Provides single answer (interpretable!)
- May result in similar or same inferences are more complex methods, especially when gene tree incongruence is rare. Tonini et al. (2015)

Disadvantages:

- Final tree often doesn't match any of the individual gene trees, and can be the wrong spp tree Kubatko and Degnan (2007)
- Incorporating coalescent models can improve accuracy Edwards et al. (2016)
- Potentially interesting conflicting signals are lost Hahn and Nakhleh (2016)

Gene tree methods

Infer individual gene trees, and combine.

Advantages:

- Captures gene tree variation
- Can focus on loci of interest (e.g. Hahn and Nakhleh (2016))
- Can model variation across gene trees in multiple ways - ILS, HGT, hybridization, and assess model fit.

Disadvantages:

- Shorter sequences result in higher error
- Loci from many approaches for generating genomic data are too short to estimate individual trees (SNPs or short loci)

Gene tree methods

Infer individual gene trees, and combine.

Coalescent analyses

- Astral: `https://github.com/smirarab/ASTRAL/blob/master/astral-tutorial.md`
- MP-EST: `https://github.com/lliu1871/mp-est`

Network/hybridization inference

- PhyloNet: `https://wiki.rice.edu/confluence/pages/viewpage.action?pageId=39500205`
- SNaQ: `https://github.com/crsl4/PhyloNetworks.jl/wiki`

Full data methods

Joint inference of gene trees, model and species tree.

Advantages:

- Model describes the processes generating the data
- Full joint likelihood calculation

Disadvantages:

- Complex models, often very slow to infer for large numbers of taxa (months!)

Full data methods

Gene sequences

- *BEAST, starBEAST2: `https://taming-the-beast.org/tutorials/StarBeast-Tutorial/`
- BPP: *Can jointly estimate coalescence and introgression* https://hal.archives-ouvertes.fr/hal-02536475/document

SNPs or short loci from across the genome

- SVDQuartets: *fast, quartet based, so handles missing data well* `http://www.phylosolutions.com/tutorials/ssb2018/svdquartets-tutorial.html`
- SNAPP: `http://evomicsorg.wpengine.netdna-cdn.com/wp-content/uploads/2018/01/BFD-tutorial-1.pdf`

Conclusions:

- The importance of gene tree discordance, as well as how to address it, depends on both your data and your question!

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leaché, A. D., Liu, L., and Davis, C. C. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, 94:447–462.

Hahn, M. W. and Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution*, 70(1):7–17.

Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. *Systematic Biology*, 56(1):17–24. Publisher: Oxford Academic.

Tonini, J., Moore, A., Stern, D., Shcheglovitova, M., and Ortí, G. (2015). Concatenation and Species Tree Methods Exhibit Statistically Indistinguishable Accuracy under a Range of Simulated Conditions. *PLOS Currents Tree of Life*. Publisher: Public Library of Science.