Progress on Estimating Large Species Trees

Tandy Warnow

University of Texas at Austin

March 7, 2012

Phylogenomic Analyses

- Input: set of estimated gene alignments and/or trees
- Output: species tree

Supertree Methods

- If the true gene trees should all be topologically identical to the true species tree, then supertree methods make sense.
- There are many supertree methods (MRP, Robinson-Foulds Supertrees, Min Flip, etc.). Which ones work well?
- We have developed the SuperFine method (Swenson et al., Systematic Biology 2012).

Superfine Study

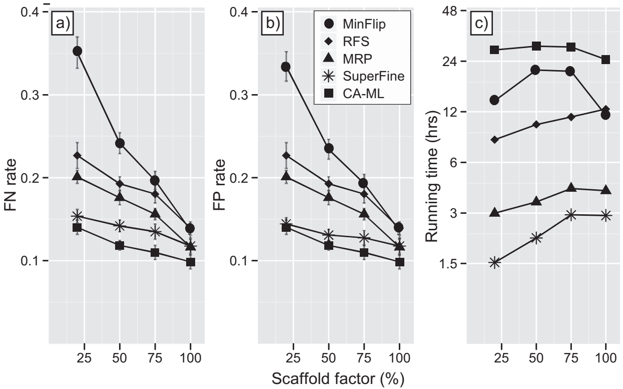


Figure: Comparison of MinFlip, Robinson-Foulds Supertree (RFS), MRP, SuperFine, and combined analysis using maximum likelihood (CA-ML) on simulated 1000-taxon datasets. Running time (c) is given in hours on a logarithmic scale; for the supertree methods, running time shown includes the time needed to calculate ML source trees using RAxML.

Extensions of SuperFine

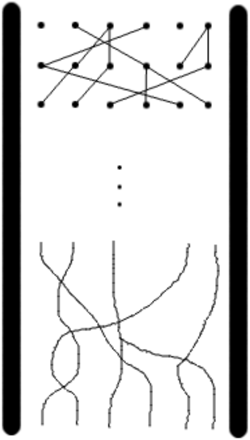
- “MRL and SuperFine+MRL: new supertree methods” by Nguyen, Mirarab, and Warnow, in Journal of Algorithms for Molecular Biology, 2012.
- “Parallelizing SuperFine” by Neves et al., in 27th Symp. on Applied Computing, Bioinformatics

Species Trees / Gene Trees Discordance

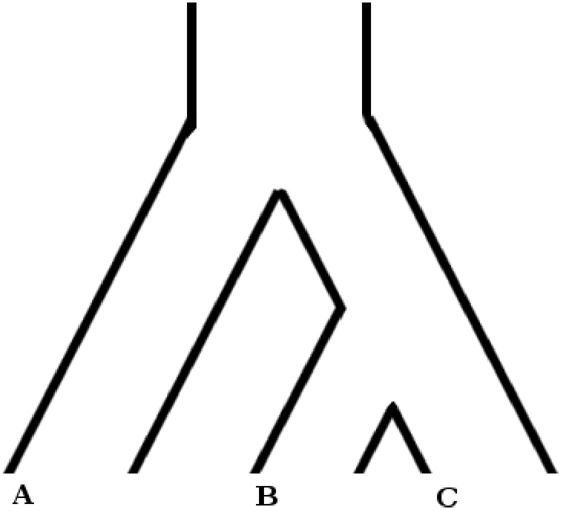
Causes:

- Gene duplication and loss
- Incomplete lineage sorting (ILS), commonly studied under the coalescent model
- Horizontal gene transfer (HGT), hybridization, recombination, etc.

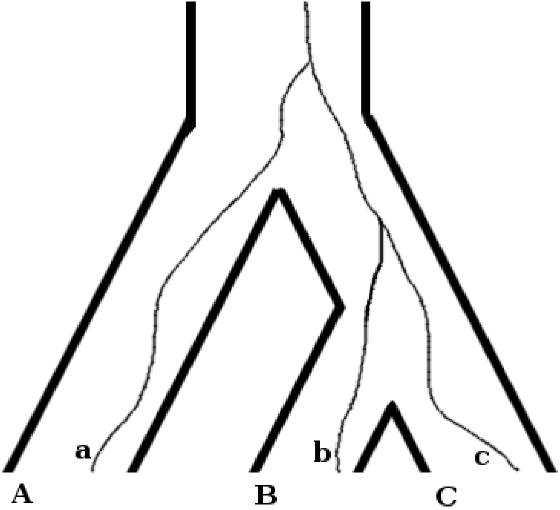
Coalescent Model



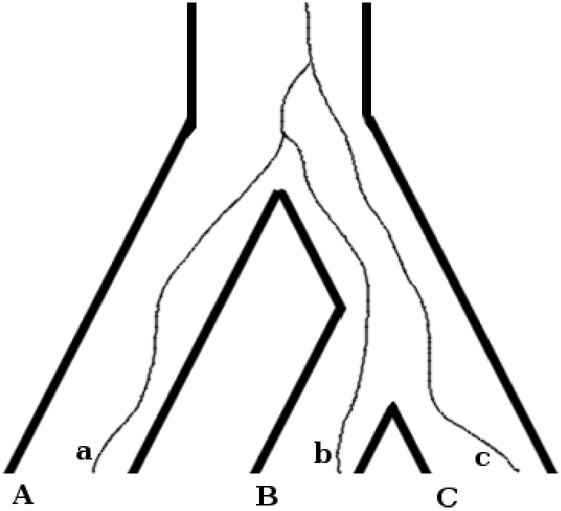
Multispecies Coalescent Model



Multispecies Coalescent Model



Multispecies Coalescent Model



Questions

- Which methods produce the most accurate species trees? How do these methods scale (in terms of computational requirements) with the number of taxa?
- Can we improve species tree estimations by considering gene tree estimation error? For example, as in Yu, Warnow, Nakhleh (RECOMB 2011), by contracting low support edges in estimated gene trees, or as in BUCKy (Ané et al.), by using gene tree distributions?
- Are there fast methods with accuracy competitive with the most promising statistical methods (e.g., BUCKy, *BEAST, BEST)?

Results for estimating species trees under duplication+loss

- Optimizing duplication and duplication+loss score: exact polynomial time algorithm for constrained species tree estimation
- Empirical study: new method several orders of magnitude faster than iGTP and Duptree, with same scores on most datasets (for complete gene tree case)
- Handling incomplete gene trees presents additional empirical challenges

Bayzid, Mirarab, and Warnow, in preparation

Results for estimating species trees under ILS

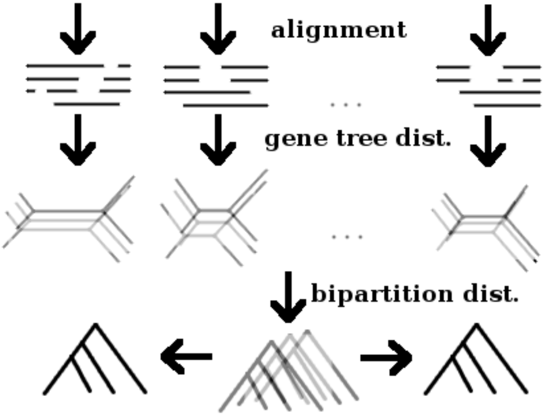
- Complete gene trees:
  - MDC optimization on unrooted, unresolved trees (Yu, Warnow, and Nakhleh, Recomb 2010) - polynomial time exact algorithm for constrained species tree estimation
  - Simulation study (Yang and Warnow, RECOMB-CG 2011) of BUCKy in comparison to fast methods
- Incomplete gene trees
  - Bayzid and Warnow, J Computational Biology:
    - MDC optimization: reconciling incomplete rooted binary gene trees with a species tree, and shows iGTP and Phylonet solve different problems
    - Experimental study shows *BEAST much more accurate than tested fast methods, but computationally too expensive to run on larger datasets.

Yang and Warnow, RECOMB-CG 2011

Compared BUCKy to fast methods on estimated gene trees that could differ from the true species tree due to ILS and/or estimation error.
BUCKy (Ané et al., MBE 2007, and Larget et al., Bioinformatics 2010)

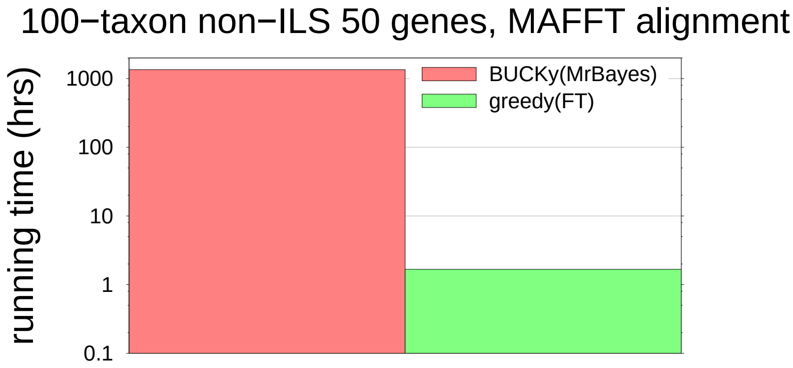
- BUCKy-pop/con, takes gene tree distributions as input, uses *concordance factors* on quartets to compute the population tree and concordance factors on clades to compute the concordance tree.
- BUCKy-pop is statistically consistent under ILS.
- BUCKy-con is not statistically consistent under ILS.

BUCKy(MrBayes) Analysis



1. MAFFT
2. MrBayes
3. BUCKy (concordance and population tree)

BUCKy(MrBayes) vs. Greedy



Memory usage:

- BUCKy: 34-234 GB
- greedy: $<$ 9 MB

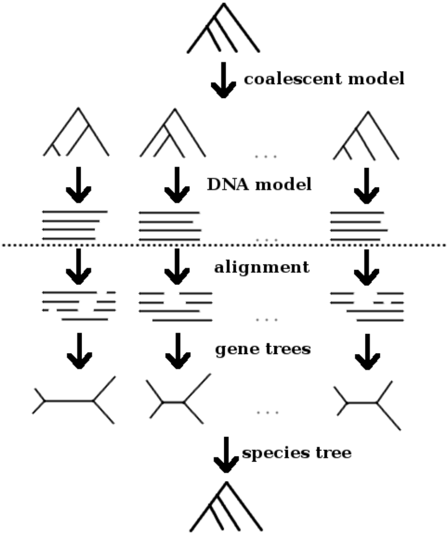
Using MrBayes to estimate gene tree distributions

- Computational issues:
    - Long running times
    - Convergence to stationarity
    - Large numbers of sampled gene trees makes BUCKy slow and memory-intensive
- Alternatives to “proper” MrBayes analysis
    - non-converged distributions
    - sparse MrBayes samples
    - replacing MrBayes with other methods (e.g., bootstrap trees using RAxML)

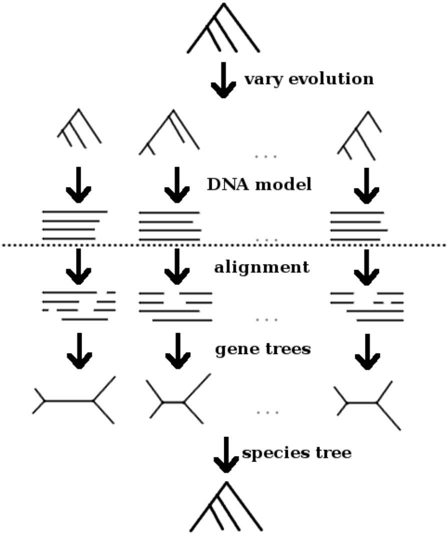
Other methods

- GLASS, distance-based (statistically consistent)
- Phylonet and iGTP for MDC
- iGTP for duplication and duplication/loss
- Greedy consensus

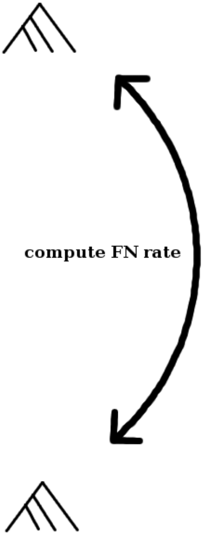
Simulation Study



Simulation Study

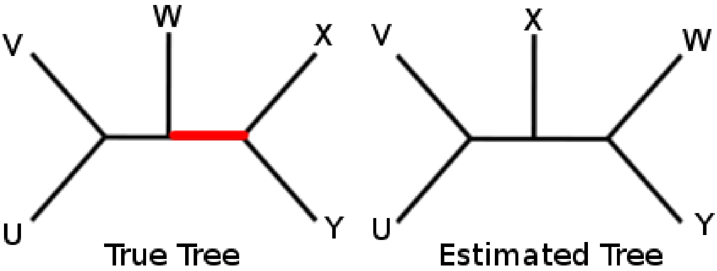


Simulation Study



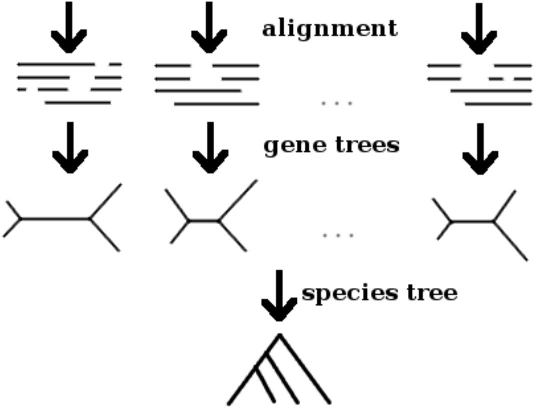
compute FN rate

Comparing Trees



- False Negative: edge in the true tree missing from the estimated tree
- FN rate (missing branch rate): 50%

Methods

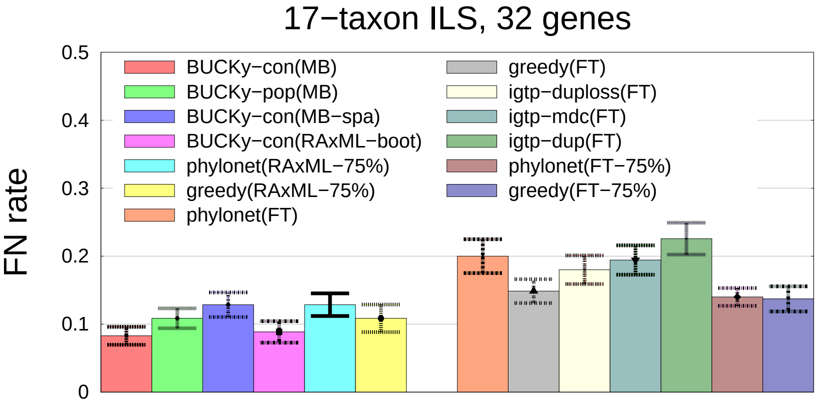


1. MAFFT
2. RAxML, FastTree, MrBayes
3. BUCKy, PhyloNet, iGTP, greedy consensus, GLASS

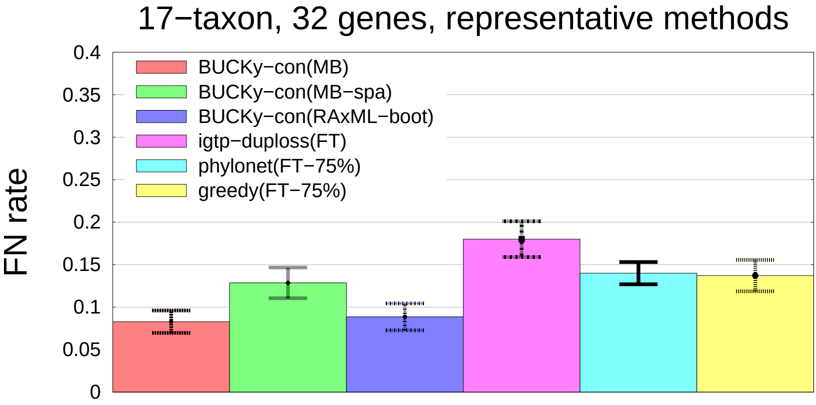
Simulation Parameters

	previous studies	this study
number of taxa	4-20	17-500
number of genes	≤ 100	25-50
evolution model	JC, HKY	GTR + Γ + Indels
cause of discord	ILS, HGT	none, ILS

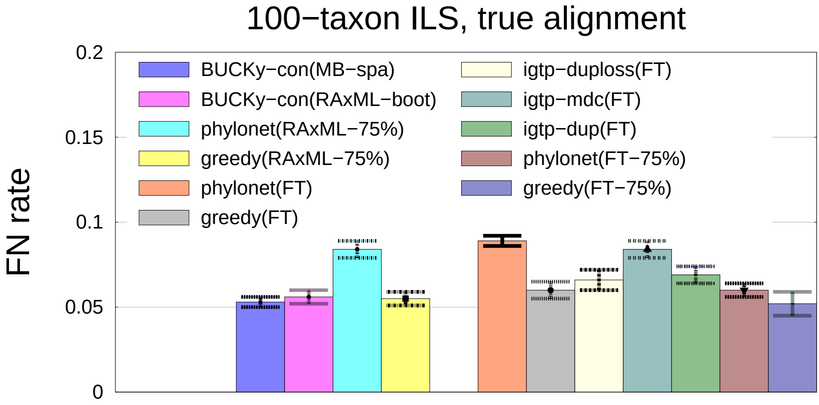
Results on 17-taxon datasets, all methods



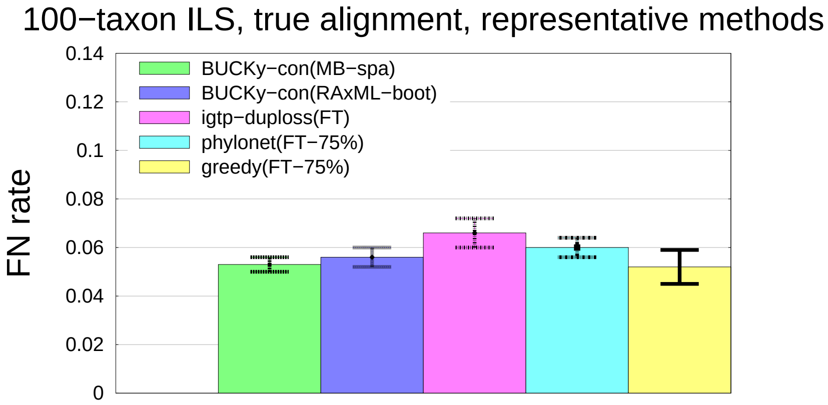
Results on 17-taxon datasets, representative methods



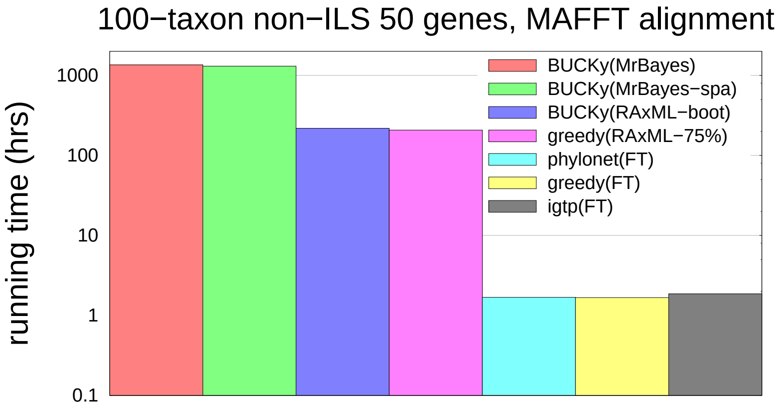
Results on 100-taxon datasets, all methods



Results on 100-taxon datasets, representative methods



Computational Requirements



Memory usage:

- BUCKy: 34-234 GB
- PhyloNet, GLASS, iGTP, greedy: $<$ 9 MB

Incomplete Gene Datasets

- In the incomplete gene dataset case, not all taxa are present in all the gene trees.
- Not all methods can be used on incomplete gene datasets (e.g., BUCKy and greedy cannot be used).
- We performed a simulation study to evaluate *BEAST and fast methods that can be run on incomplete gene datasets. We explored performance on 11, 17, and 100-taxon datasets.
- *BEAST was far from converging on the 100-taxon datasets, but could run on the others. We show results for 11-taxon datasets.

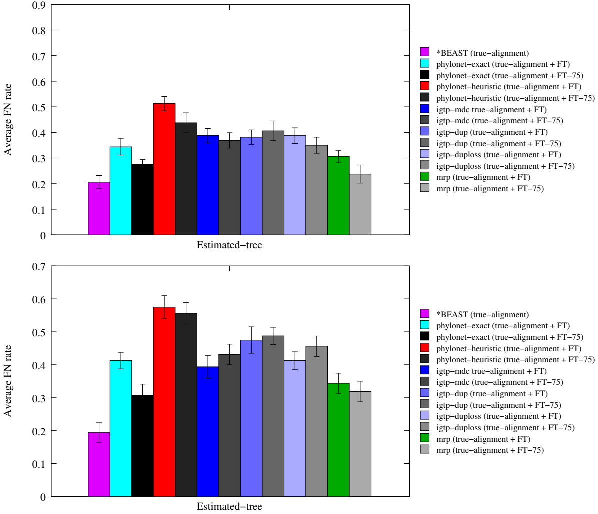


Figure: Average missing branch rates of methods on twenty (20) 11-taxon 10-gene datasets on true alignments. From top to bottom, number of missing taxa = 2 and 3.

Findings

- Accounting for gene tree estimation error improves methods
- MrBayes is expensive to run correctly - even on 17-taxon inputs. Using other methods to estimate the gene tree distribution does not reduce accuracy for BUCKy very much.
- Some fast methods (e.g., Greedy(FT)) have accuracy close to that of BUCKy-con(MrBayes).
- BUCKy-con more accurate than BUCKy-pop
- iGTP-duploss more accurate than iGTP-MDC
- GLASS fast but not competitive with other methods
- *BEAST is very accurate when it can be run well (on the 100-taxon datasets, it produced poor trees, but also failed to converge).

Observations:

- Statistical guarantees are often not predictive of performance on finite data
- Performance on large datasets can be different than on small datasets
- Estimating highly accurate species trees from incomplete gene trees is difficult
- Discrete optimization problems (duplication scores, duploss scores, MDC scores, etc.) popular, but may not produce good trees. Statistical methods are often more accurate, but typically do not run on large datasets.

Open Questions:

- Why are Greedy and MRP so accurate?
- How do methods perform when gene tree incongruence is due to other factors than ILS?
- Can we develop methods with accuracy close to the accuracy of the best statistical methods, with much lower computational costs?

Acknowledgements

- Jimmy Yang and Md Shamsuzzoha Bayzid for simulation study results
- Luay Nakhleh and Yun Yu for 17-taxon datasets
- Cecile Ané and Yujin Chung for 11-taxon datasets
- Steve Evans for statistical testing advice
- NSF
- Guggenheim Foundation