Naive binning improves phylogenomic analyses

Md Shamsuzzoha Bayzid and Tandy Warnow

Bioinformatics
2013

Motivation

- Poor accuracy when the individual gene sequence alignments have low phylogenetic signal.
- discordance between the gene trees and the true species tree
- Scalability : *BEAST is computationally very intensive and challenging to run it on a dataset with 100 or more gene trees and 200 taxa.
- Running time

Proposed Mythology

- Step 1: The input genes sequences are partitioned into bins such that each bin has approximately the same number of genes.
- Step 2: A supergene alignment is computed for each bin from the concatenation of the gene sequences in the bin.
- Step 3: Inside each bin, compute a *supergene tree* using maximum likelihood based on the supergene alignment
- Step 4: Estimated species tree is computed from the supergene trees using either a summary method or from the supergene alignments.

Datasets

	11-taxon	17-taxon
Varying ILS	Yes	Yes
	Jukes cantor model	Jukes cantor model
molecular clock	No	Yes
Sequence Length	Short	Long

Experiment 1

- Evalutation of fast methods :

CA-ML is compared with MP-EST, MRP, Phylonet, and GC.

Observation :- CA-ML had the best accuracy with very large improvements over other methods on the 11-taxon datasets and small improvements on the 17-taxon datasets.

Experiment 2

- Comparison between BUCKy-pop, MP-EST,*BEAST, CA-ML, and BUCKy-con. This experiment was limited to 20 replicates as *BEAST has higher running time.

- Observation : Does not substantially increase topological accuracy but increases scalability and decreases running time.
-

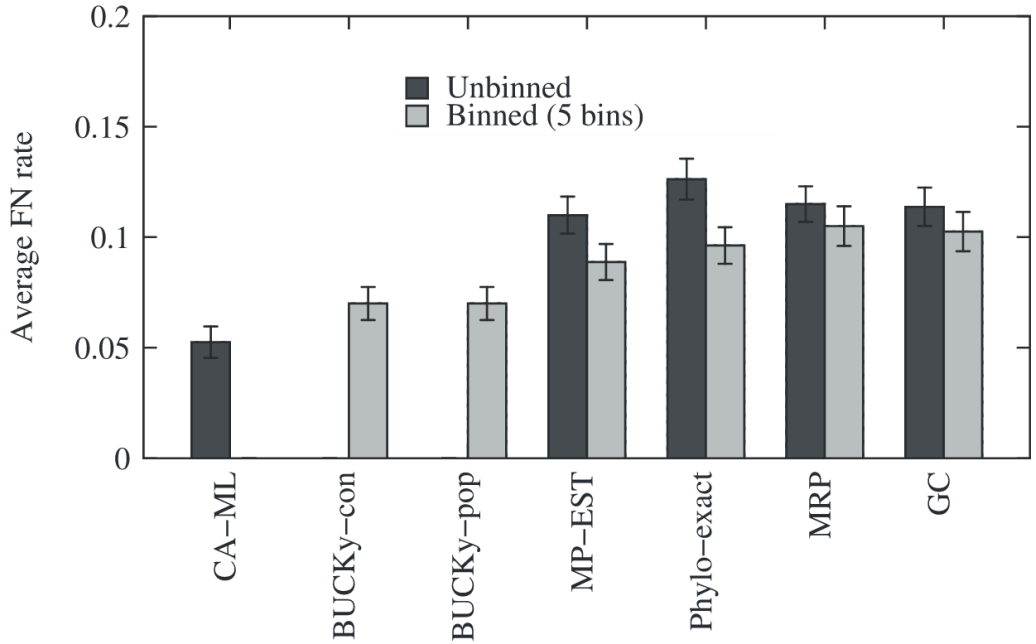

Fig. 8. Results of binning experiments of the fast methods on 100 replicates of the 11-taxon 25-gene strongILS datasets. Each bin contains five genes. We omit BUCKy on unbinned genes and *BEAST (binned or unbinned) because these are too slow to run on all 100 replicates within our time limits. CA-ML is not impacted by binning because it uses an unpartitioned analysis

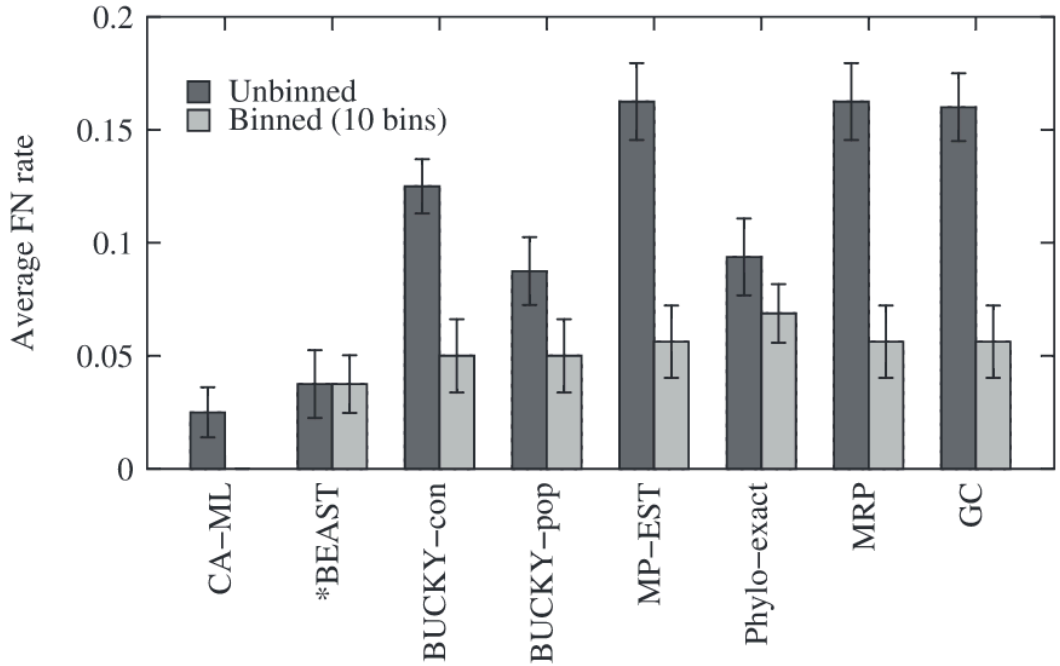

Fig. 11. Results of the binning experiment for all methods on 20 replicate 11-taxon 50-gene strongILS datasets. CA-ML is not impacted by binning because it is an unpartitioned analysis. Each bin contains five genes. Average and standard error bars shown

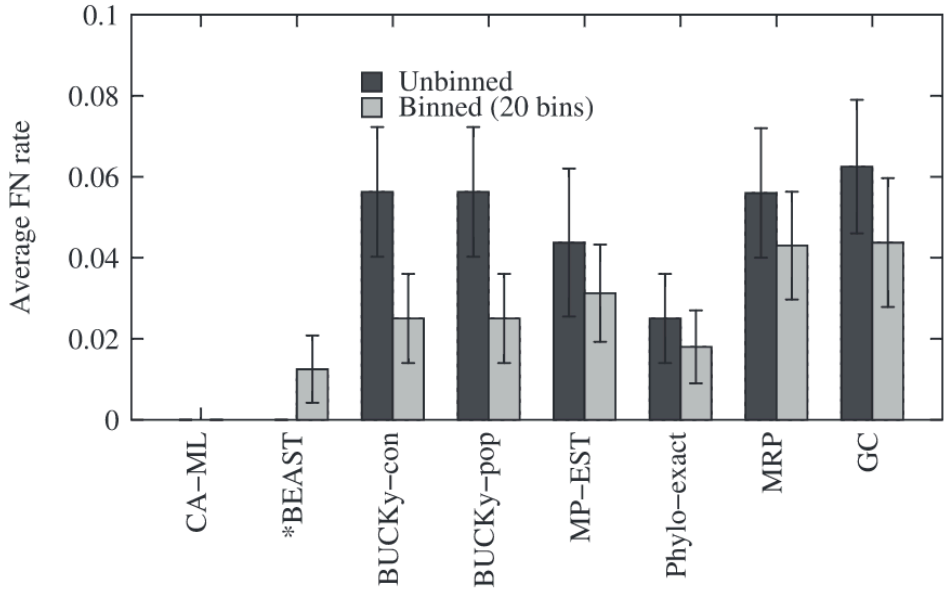

Fig. 12. Results of the binning experiment for all methods (except *BEAST) on 20 replicate 11-taxon 100-gene strongILS datasets. Each bin contains five genes. Average and standard error bars shown. We omit *BEAST on unbinned genes because it could not run to convergence on this dataset within the time limit; however, we show results for *BEAST on the binned datasets. CA-ML returns the true tree on these data

Comments

1) This method has been tested for the 17- axon and 15-taxon dataset. It would be interesting to it for a larger dataset.

2) Instead of Naive binning the decomposition can be done on the basis of the CT-5 algorithm. (CT-5 algorithm is used in Sate I). Moreover many intelligent methods for binning have been proposed in the recent past like weighted statistical binning etc which can helpful for dealing with large datasets.