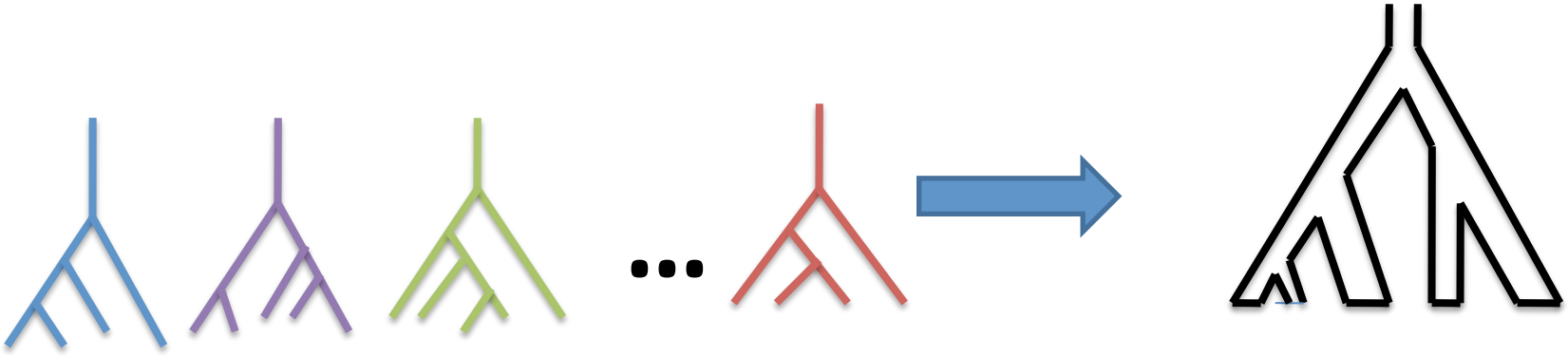
CS 598 : Mid-Term Report

ASTRID with Ninja

Problem Statement



Input
Estimated Complete
Gene Trees

Output
Species Tree

Proposed methdology

Step 1: Construct $n \times n$ matrix M_i for all $i = 1, \ldots, k$

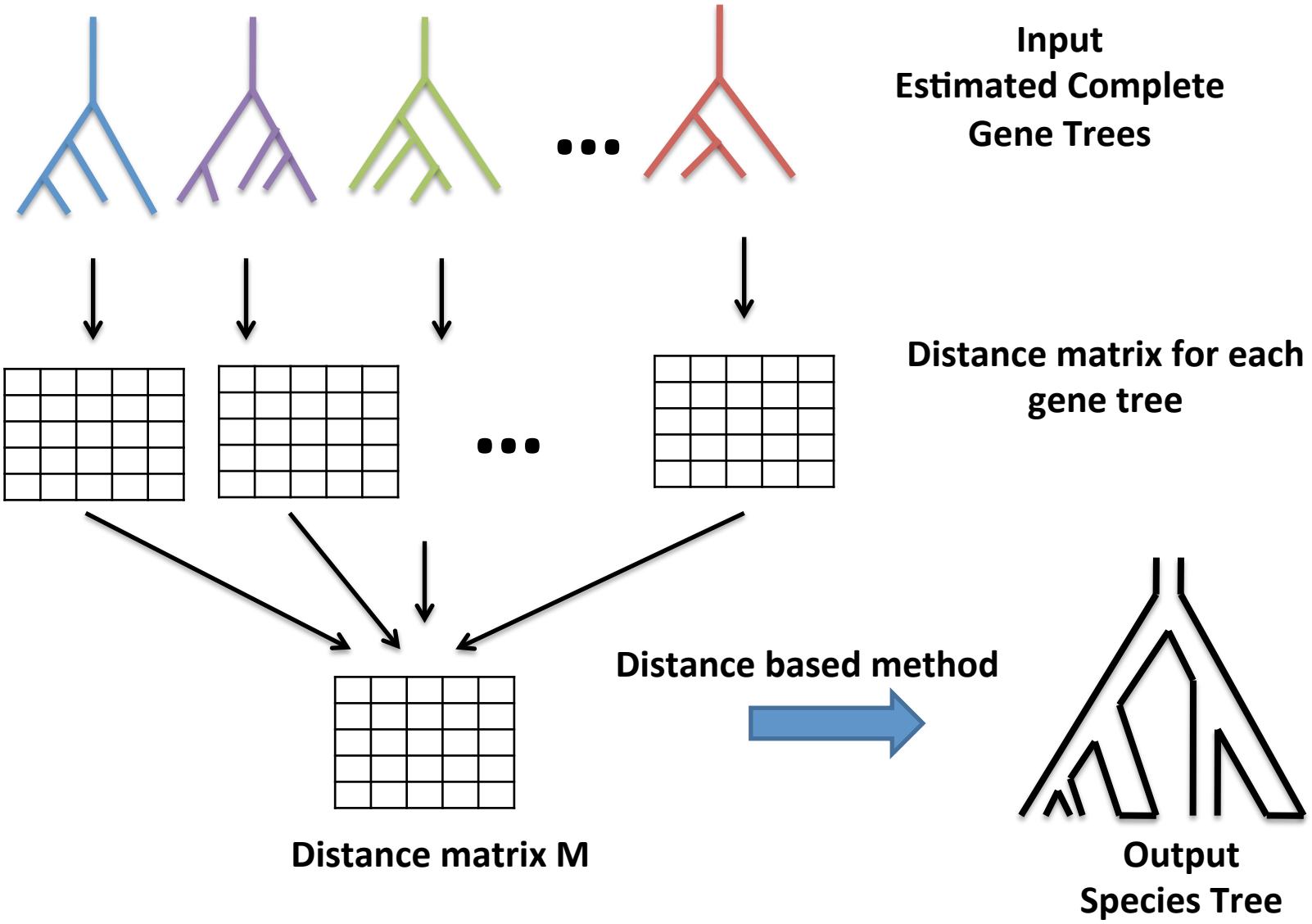
$\forall p, q \in S$, set $M_i(p, q) \leftarrow \#$ of edges in the path between p and q in $\mathcal{T}_i$

Step 2 : Construct $n \times n$ matrix M

Set $M(p, q) \leftarrow \frac{\sum_{i=1}^{k} M_i(p,q)}{k}$ where k is the total number of gene trees present in the dataset.

Step 3: Now we have a distance matrix on the entire set of taxa S. Apply distance-based method lik
Ninja to construct a tree on S

Methodology



Motivation

- Distance based method
- FastME, BioNJ, Weighbor, NINJA, FastTree etc. - statistically consistent
- Ninja is sometimes faster than FastMe

Dataset

Dataset Name	# taxa	Gene Trees	ILS (AD%)	Sequence Length
MC11 - 1000 taxa	1001	400	35 AD %	300-1500bp
MC11 - 1000 taxa	1001	600	35 AD %	300-1500bp
MC11 - 1000 taxa	1001	800	35 AD %	300-1500bp
MC11 - 1000 taxa	1001	1000	35 AD %	300-1500 bp
MC12 - 1000 taxa	1001	400	52 AD %	100bp
MC12 - 1000 taxa	1001	600	52 AD %	100bp
MC12 - 1000 taxa	1001	800	52 AD %	100bp
MC12 - 1000 taxa	1001	1000	52 AD %	100bp

TABLE I: Dataset

Results

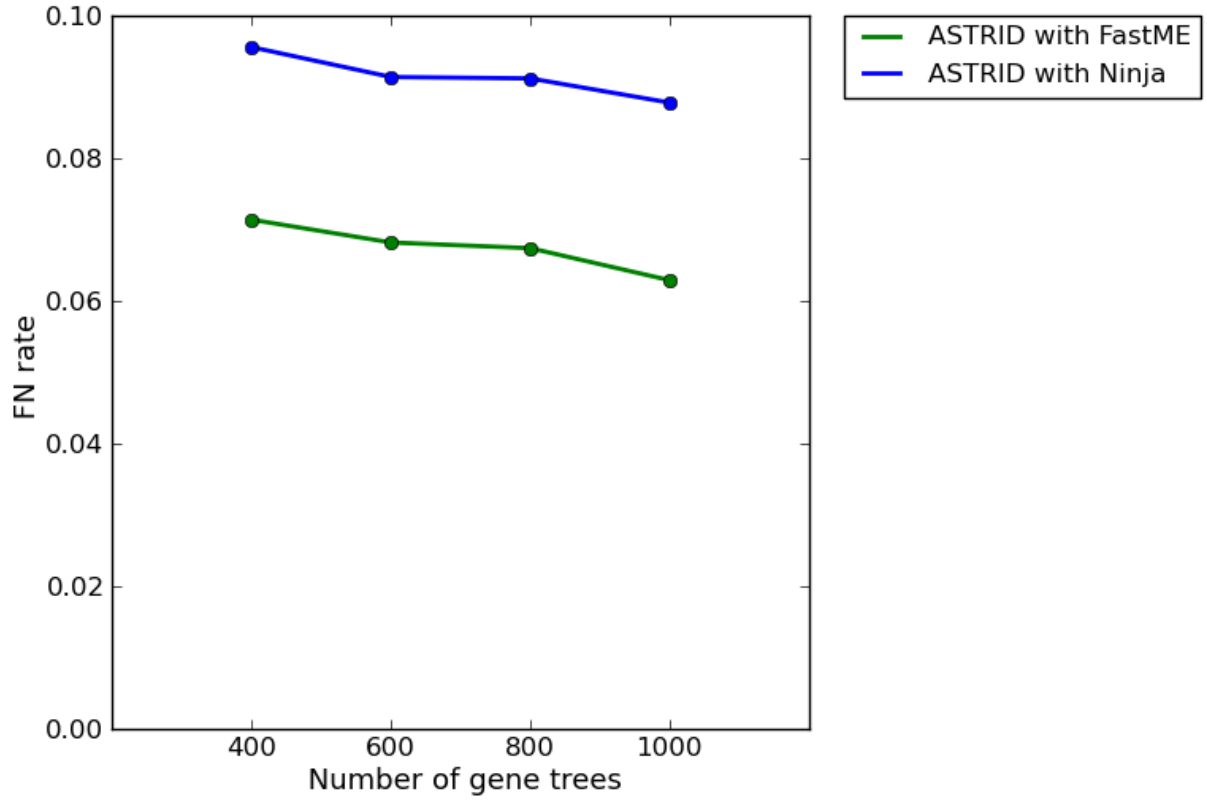


Figure showing the topological accuracy of ASTRID with the two distance based methods namely Ninja and FastME on the Astral-2 MC11 dataset where the level of ILS is 35 AD % and sequence length varies from 300-1500bp. The experiments are carried out on 10 replicates.

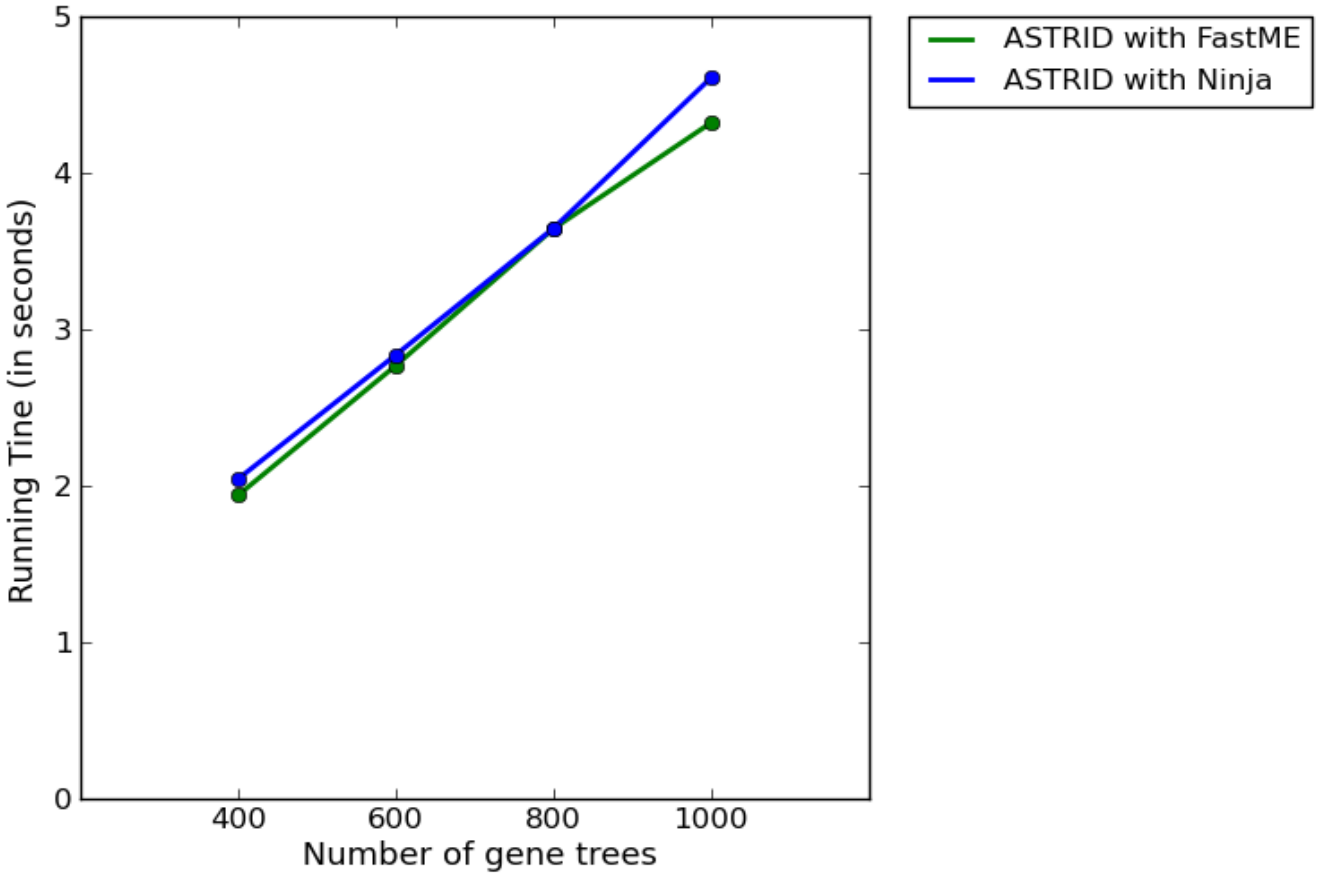


Figure showing the running time of ASTRID with the two distance based methods namely Ninja and FastME on the Astral-2 MC11 dataset where the level of ILS is 35 AD % and sequence length varies from 300-1500bp. The experiments are carried out on 10 replicates. The simulations are done on a node with 10 processors per node with 128GB RAM and Intel E5-2670 (Sandy Bridge) 2.60GHz, processor.

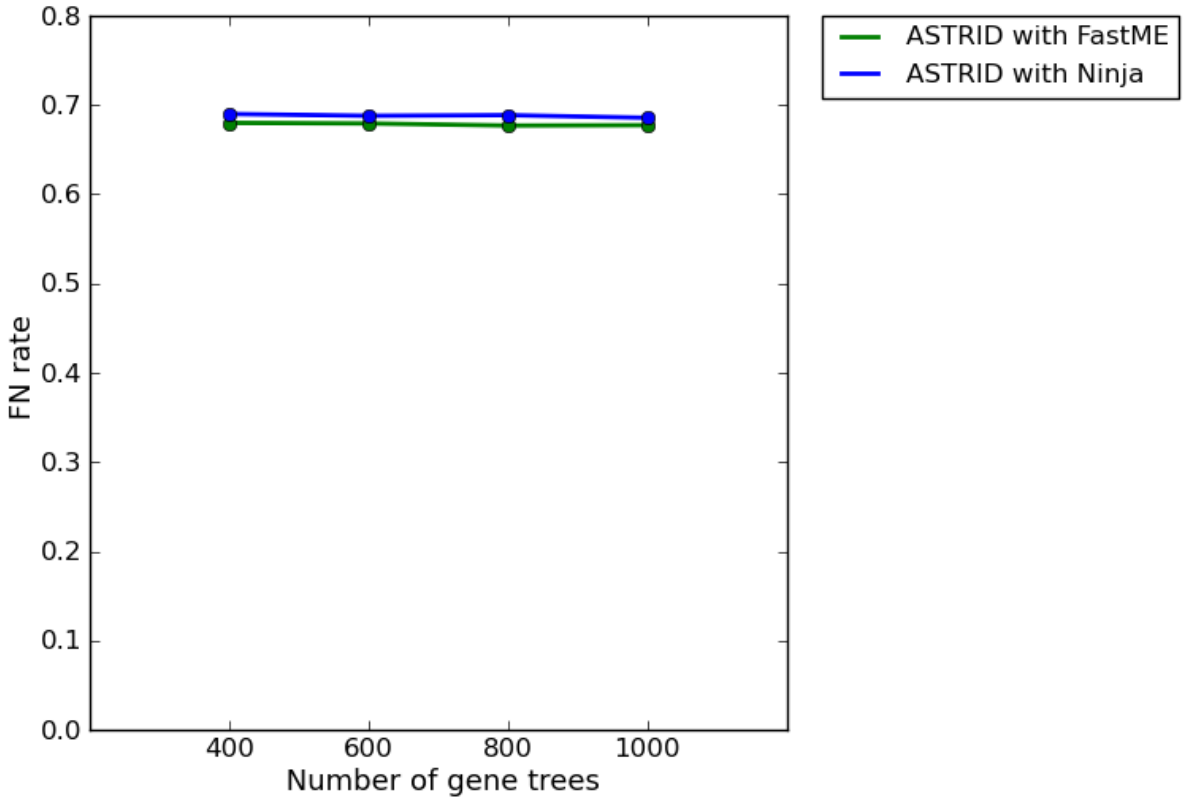


Figure showing the topological accuracy of ASTRID with the two distance based methods namely Ninja and FastME on the Astral-2 MC12 dataset where the level of ILS is 52 AD % and sequence length is fixed to 100b. The experiments are carried out on 10 replicates.

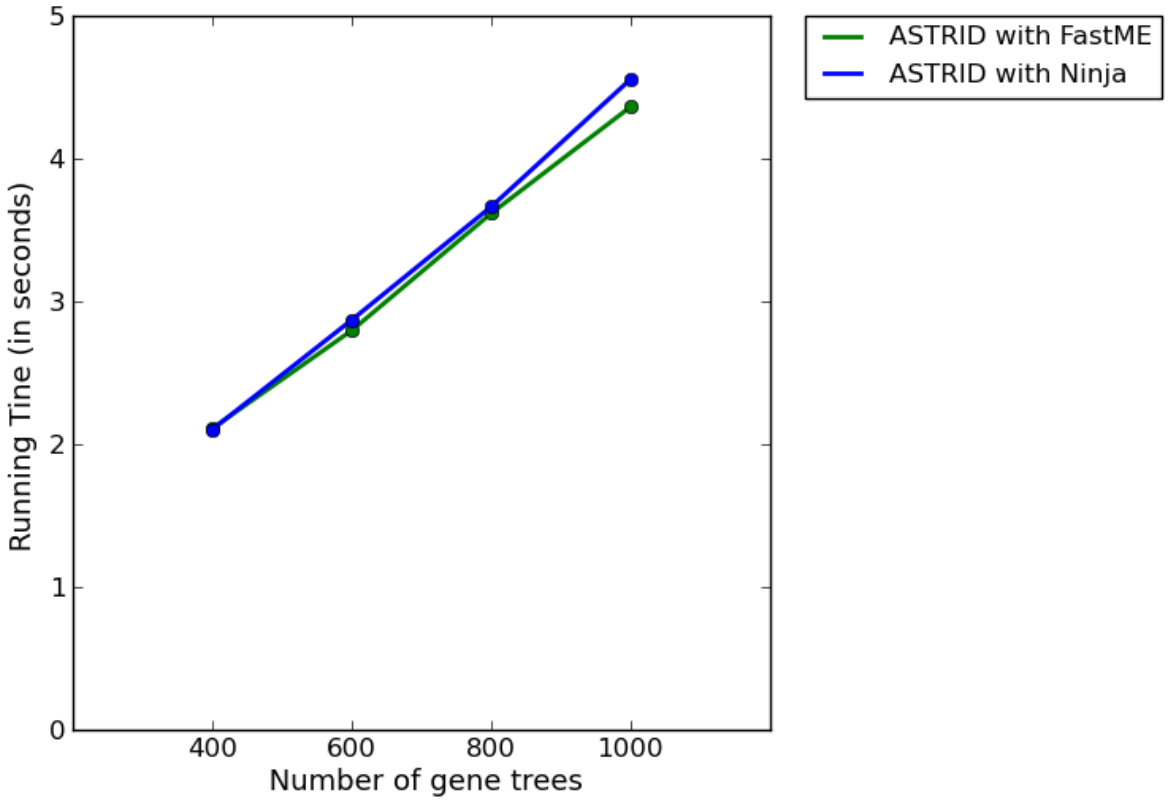


Figure showing the running time of ASTRID with the two distance based methods namely Ninja and FastME on the Astral-2 MC12 dataset where the level of ILS is 52 AD \% and sequence length is fixed to 100bp. The experiments are carried out on 10 replicates. The simulations are done on a node with 10 processors per node with 128GB RAM and Intel E5-2670 (Sandy Bridge) 2.60GHz, processor

Future work

Step 1: Construct $n \times n$ matrix M_i for all $i = 1, \ldots, k$

$\forall p, q \in S$, set $M_i(p, q) \leftarrow$: **Sum of edge weights in the path between p and q**

Step 2 : Construct $n \times n$ matrix M

Set $M(p, q) \leftarrow \frac{\sum_{i=1}^{k} M_i(p,q)}{k}$ where k is the total number of gene trees present in the dataset.

Step 3: Now we have a distance matrix on the entire set of taxa S.

Apply FastME as the distance based method