

# Large Scale Phylogenetic Reconstruction from Arbitrary Gene-Order Data

Jijun Tang

[jtang@cs.unm.edu](mailto:jtang@cs.unm.edu)

Department of Computer Science  
University of New Mexico

# Acknowledgement

- **Joint work with Bernard Moret**
- **Supported by National Science Foundation**

# Overview

- **GRAPPA**
- **Scaling up using DCM-GRAPPA**
- **Unequal gene content**
- **Conclusion**

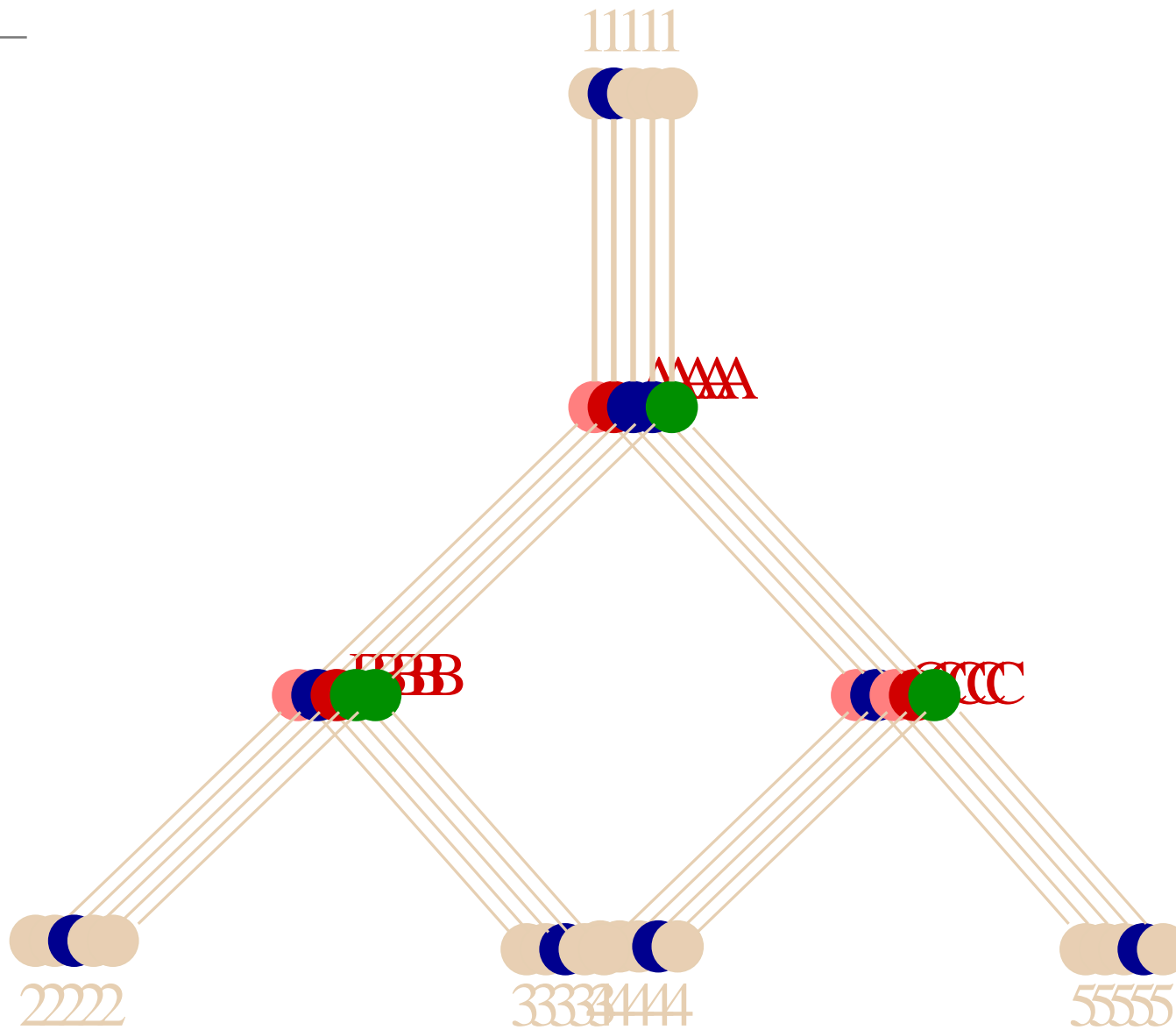
# GRAPPA

- **Genome Rearrangements Analysis under Parsimony and other Phylogenetic Algorithms**
- **Started as an effort to reimplement the BPA analysis of Sankoff and Blanchette**
- **Used algorithmic techniques to improve the speed**

# Algorithm Outline

- **Consider each tree topology in turn**
- **For each tree**
  - Test the lower bound, if it exceeds the best so far, continue to the next tree
  - Initialize the internal nodes by some means
  - Compute medians of three iteratively until no change occurs
- **Return the lowest score tree**

# Scoring a Tree



# Problem of Exhaustive Search

- **The search space is too large**
  - $(2N - 5) \times (2N - 7) \times \dots \times 1$  unrooted trees for  $N$  genomes
  - It will take 200,000 years to finish a 20-genome dataset
  - Exhaustive search cannot be used for more than 17 genomes
- **We need a new method to deal with large scale datasets**

# DCM (Disk Covering Method)

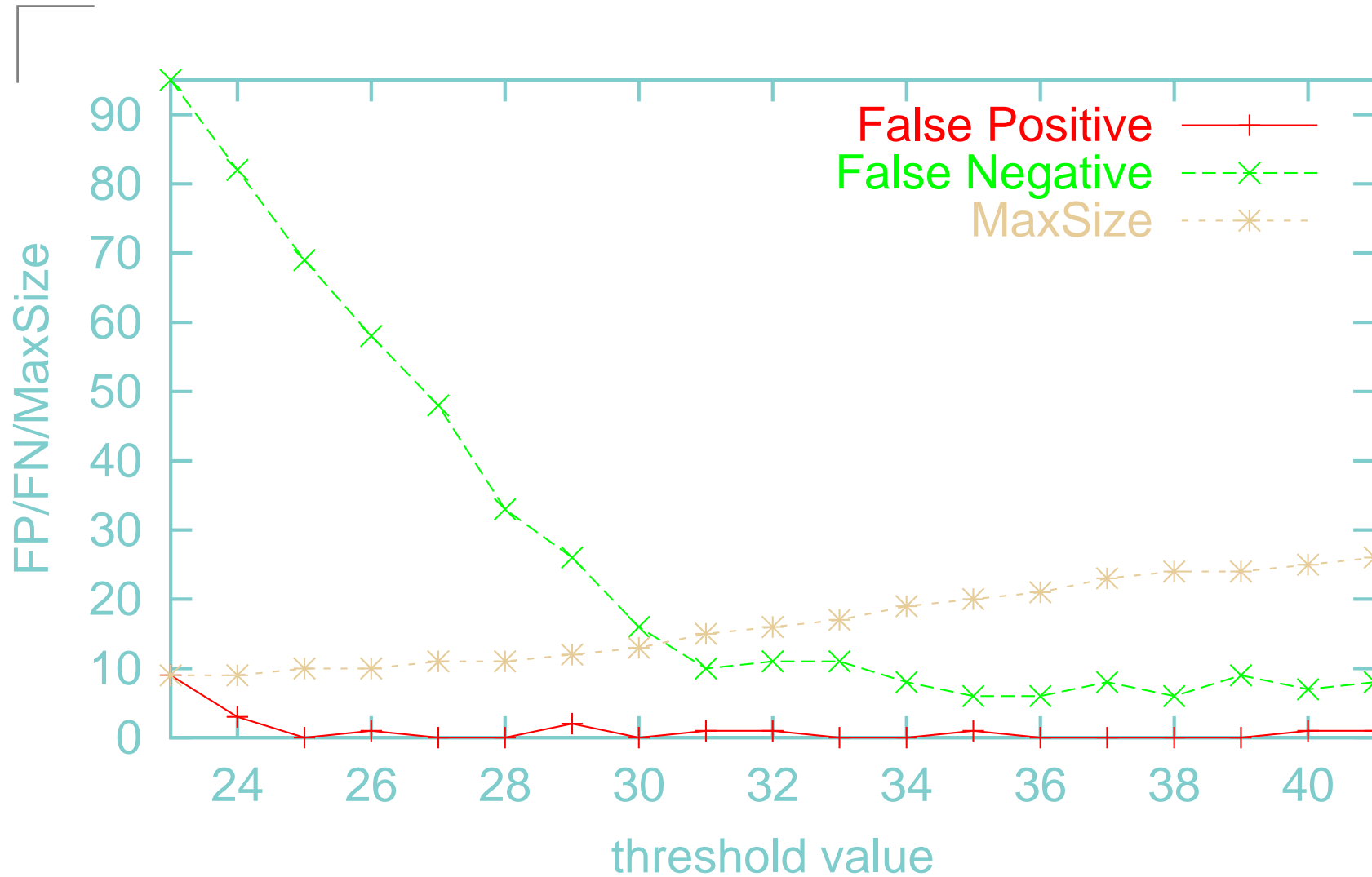
- **Developed by Tandy Warnow and her group**
- **Decomposes the input dataset into smaller overlapping sets of closely related genomes**
- **Reconstructs phylogenies for the smaller subproblems by some base methods**
  - Neighbor-joining (DCM-NJ)
  - GRAPPA (DCM-GRAPPA)
- **Combines the subtrees into one tree on the entire dataset**

# Challenge

- **Different threshold values will result in different decompositions and we cannot predict which threshold value will produce the best tree**
- **GRAPPA is limited to handle up to 14 genomes**

Will the disks be too large?

# Threshold Value: Example



# Integrating DCM and GRAPPA

- **DCM and GRAPPA communicate through files**
- **Recursively call DCM if disk has more than 13 genomes**
- **Results on each disk are cached**
  - Normally, for each threshold value, at least 40% of the disks have been examined during the computation of previous threshold values
  - The running time is reduced substantially

# Experimental Results: Speed

- **GRAPPA will not finish a 640-genome dataset within the life of the universe**
- **DCM-GRAPPA can finish a 640-genome dataset within a day**

**Effectively infinite speed-up**

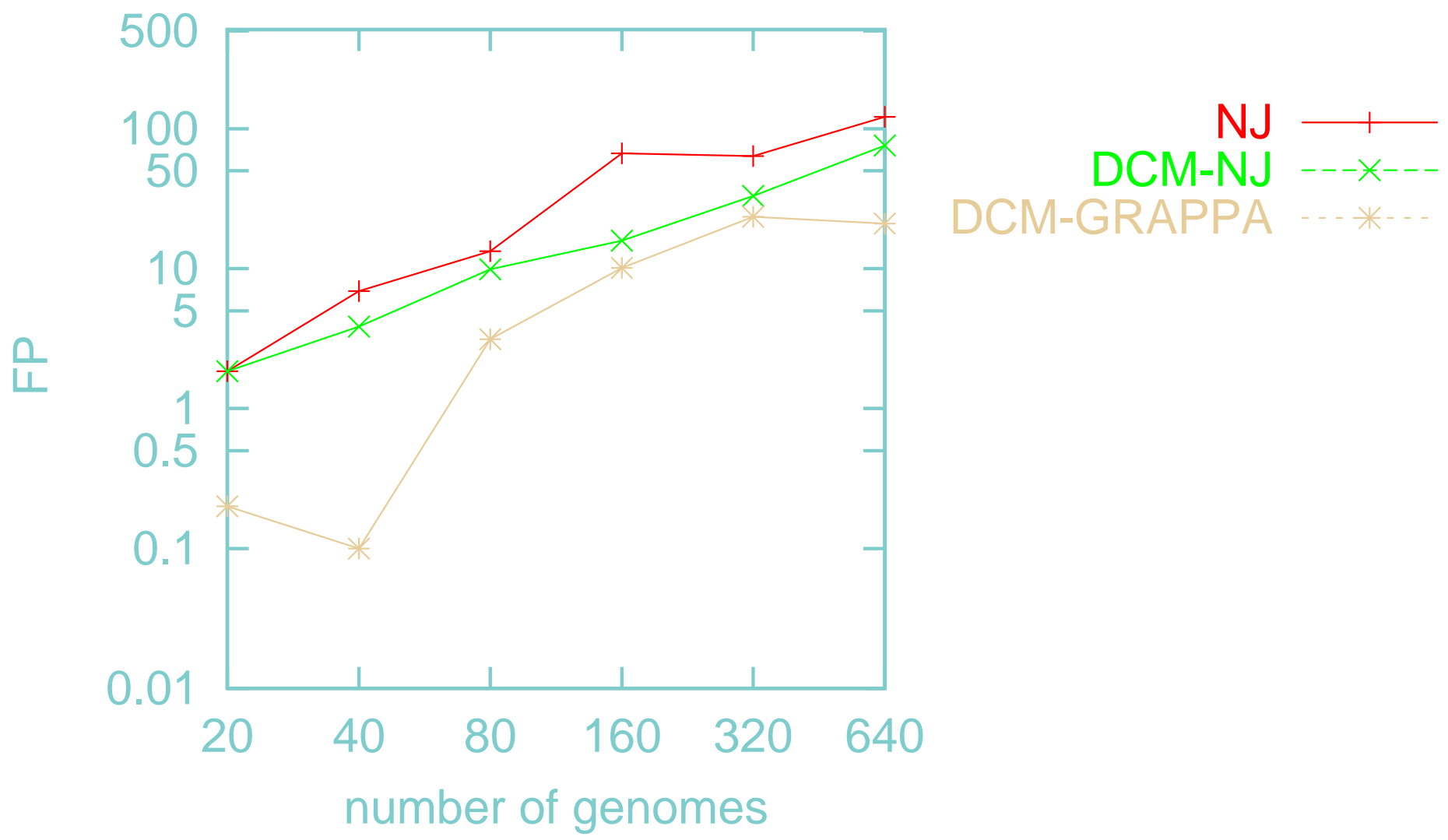
# Experimental Results: Accuracy

- **DCM-GRAPPA is accurate**
- **Error rates stayed below 2%**

For 640-genome datasets, fewer than 10 edges are wrong out of 637 internal edges

# Accuracy: Log/Log Scale

100 genes, expected edge length is 2



# Unequal Gene Content

- **GRAPPA requires equal gene content**
  - Each genome should have the same number of genes
  - Each gene should appear exactly once in a genome
- **Deletion, insertion and duplication are common in reality**

# Progress in Distance Computation

- **Breakpoint** [Sankoff et al., 1998]:  
Distance counting the number of altered adjacencies for identical gene content; linear time.
- **INV** [Hannenhalli, Pevzner, 1995]:  
Edit distance (inversions) for identical gene content; [Bader, Moret, Yan, 2001] linear time.
- **INV-DEL** [El-Mabrouk, 2000]:  
Edit distance (inversions and insertions/deletions, but no duplications); doable in linear time [Liu, Moret, 2003].
- **ALL** [Marron, Swenson, Moret, 2003]:  
Estimated edit distance (inversions, insertions/deletions, duplications).
- **INV-DUP** [Tang, Liu, in progress]:  
Exact solution for inversions and duplications.

# Algorithm for Unequal Gene Content

- **Using the exhaustive approach of GRAPPA**
- **For each tree**
  - Test the lower bound
  - Determine the *gene content* of each internal node
  - Iteratively solving the median problem using new median solver
- **Return tree(s) with the lowest score**

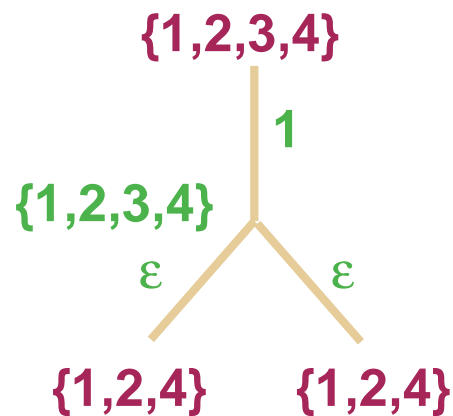
# Gene Content of Internal Node

## Assumptions:

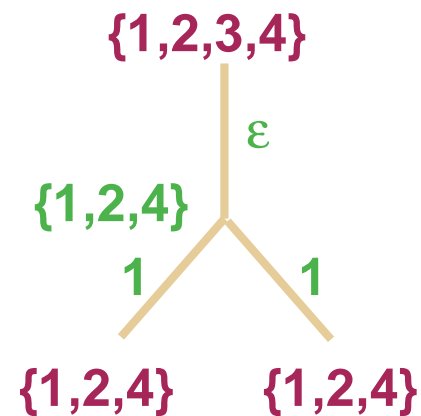
Probability of a deletion/insertion is  $\varepsilon$  (very small).

Probability of no change is  $\sim 1$ .

## Example:



$$p = \varepsilon^2$$



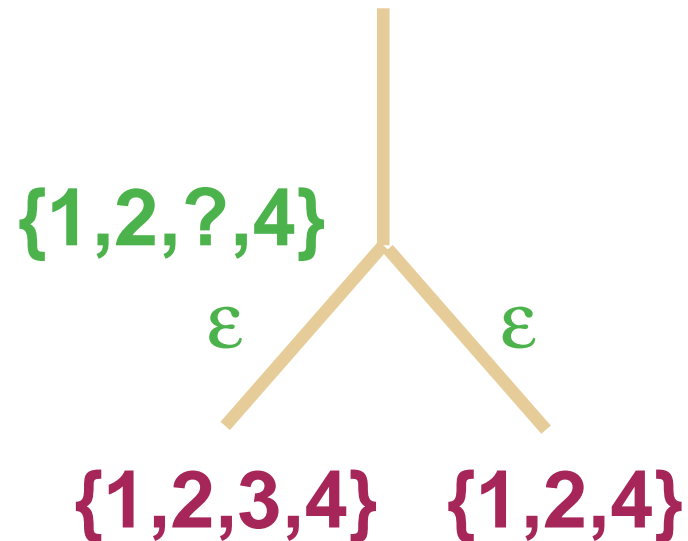
$$p = \varepsilon$$

# Ambiguity of the Algorithm

## Assumptions:

For unrooted tree, deletion and insertion have the same probability

## Example:



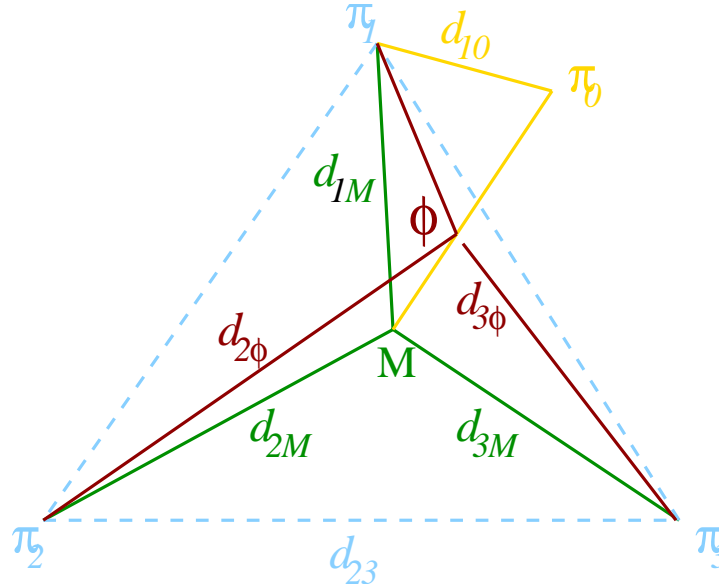
# Determine Gene Content

- **For each sibling pair of *leaves***
  - if a gene appears in both, place it in the parent (an internal node)
  - if it is absent from both, do not place it in the parent
  - if the gene appears in one leaf, but not the other, mark its status as undetermined in the parent
  - remove all processed leaves and repeat
- **Starting from an arbitrary leaf, do a depth-first search to propagate resolutions**

# Solving the Median Problem

- **Pick a start permutation, push it into a queue**
  - It has the same gene content as the median node
  - Be as close as possible to one of the three nodes
- **Pop permutations from the queue until it's empty**
  - if the score of this permutation meets the lower bound, STOP
  - otherwise create the  $\binom{n}{2}$  neighbors that are one inversion away, discard nodes for which the lower bound exceeds the best so far, and push the remaining nodes into the queue

# Bounds



**Note:**  $\pi_0$  has the same gene content as  $M$

If  $\phi$  is on the shortest path from  $\pi_0$  to the desired median permutation, then the median score  $D(M)$  obeys:

$$\frac{d_{2,\phi} + d_{3,\phi} + d_{2,3}}{2} + d_{0,\phi} - d_{0,1} \leq D(M) \leq d_{1,\phi} + d_{2,\phi} + d_{3,\phi}$$

# Experimental Results

- **For simulated datasets:**

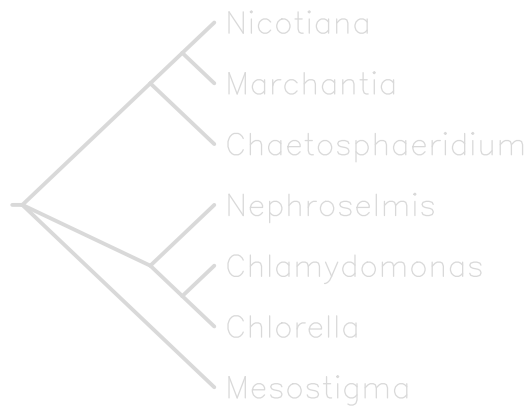
- The FP rate is very small ( $< 1\%$ )
- The FN rate is a bit higher than typically seen in equal gene content ( $< 10\%$ )

- **Tested on a biological dataset**

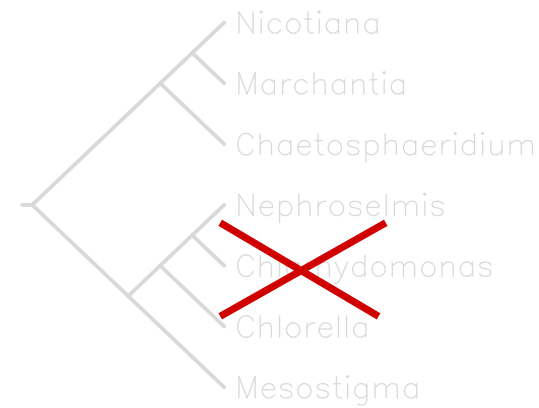
- 7 chloroplast genomes from green plants
- Three land plants, three green algae, one flagellate
- All genomes have 77 genes, except *Chlorella vulgaris* lost 3 genes

# Results from Different Methods

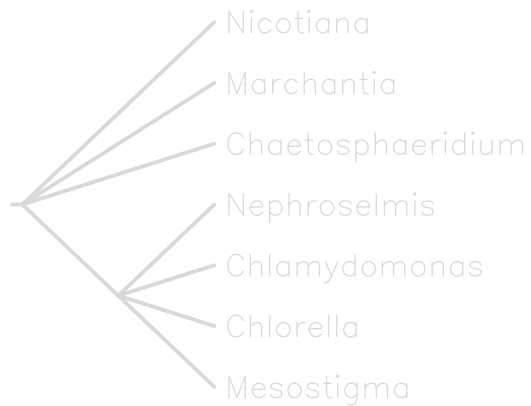
Proposed tree



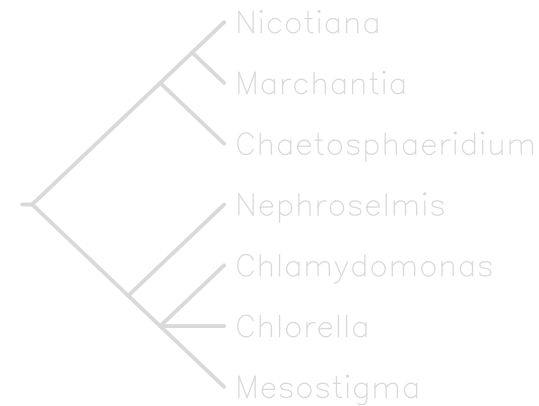
NJ tree



BP median



New median



# Conclusions

- Current algorithmic approaches scale to significant sizes (1,000 for DCM-GRAPPA)—comparable to the best achievable with sequence data and with better accuracy.
- We have a general solution for unequal gene content
- With the advance in distance computation, we will be able to handle arbitrary gene content in the near future