

# Large-scale phylogeny reconstruction from arbitrary gene-order data<sup>1</sup>

Jijun Tang and Bernard M.E. Moret

Department of Computer Science, University of New Mexico, Albuquerque, NM 87131

jttang,moret@cs.unm.edu

Supported by the National Science Foundation under ACI 00-81404, DEB 01-20709, EIA 01-13095, and EIA 01-21377.

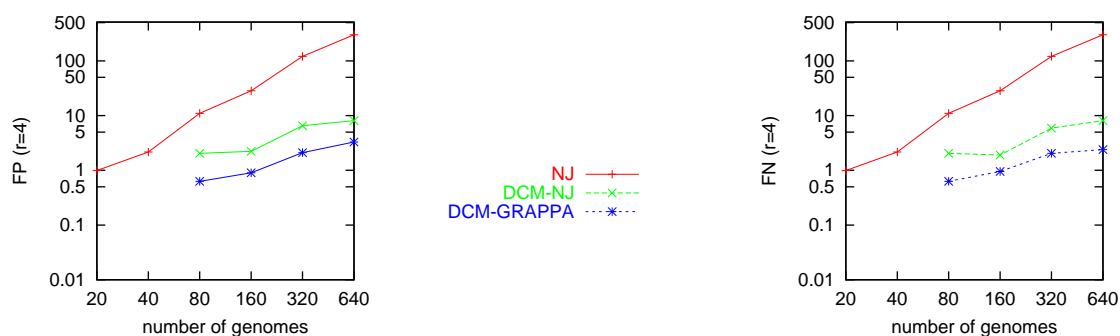
Phylogenetic reconstruction from gene-order data has attracted increasing attention from both biologists and computer scientists over the last few years. So far, our software suite GRAPPA is the most accurate approach, but it cannot practically compute cases when there are more than 15 genomes and it requires that all genomes have identical gene content, with each gene appearing exactly once in each genome. We report here on our successful efforts to scale up GRAPPA to hundreds of genomes and to handle unequal gene content and duplicate genes.

**Scaling up the current method:** GRAPPA exhaustively checks all  $(2n - 5)!!$  trees for  $n$  genomes. A lower bound is computed for each tree; trees whose lower bound exceeds the best solution found so far are discarded, while surviving trees must be scored. GRAPPA scores a tree using the approach pioneered by Sankoff: first initialize the gene order for each internal node, then repeatedly traverse the tree, attempting at each node to improve its gene order by replacing it with the median of its three neighbors. The tree(s) with the minimum score are then returned as the best. With our tightened lower bound and improved search order, GRAPPA is fast enough for up to 15 genomes, but unusable for 18 or more genomes.

To scale up GRAPPA, we combined it with the Disk-Covering Method (DCM) developed by Warnow *et al.* The new DCM-GRAPPA uses a two-step approach:

1. Decompose the dataset into smaller overlapping subsets and runs GRAPPA on each piece to reconstruct the subtrees.
2. Use strict consensus to produce a single tree from the subtrees returned by GRAPPA.

DCM-GRAPPA removes the computational limit of GRAPPA and can handle dataset with a thousand genomes, yet still retains GRAPPA's high accuracy. In our simulations, DCM-GRAPPA handles 640-genome datasets within a day to produce trees with less than 2% error.



False positives (left) and false negatives (right) for trees reconstructed with NJ, DCM-NJ, and DCM-GRAPPA.

**Handling unequal gene content:** Progress has been made in the last few years in computing the distance between two genomes with different gene content. El-Mabrouk proposed an exact method to compute edit sequences for inversions and deletions (extended to an approximate method for general operations by Marron, Swenson, and Moret). We have implemented a linear-time algorithm to compute El-Mabrouk’s edit distance and used this implementation to solve the median problem for inversions, deletions, insertions, and duplications. With this solver, we can score phylogenetic trees for, and thus also reconstruct phylogenies from, genomes with unequal gene content, as long as the content changes are small. Our approach proceeds in two phases: given a tree, we first determine the gene content at each internal node, then we reconstruct internal gene orders to minimize the sum of the edge lengths.

*Determining the gene content of internal nodes:* From the biological model, all events are rare, say with a probability bounded by a very small  $\epsilon$ . Exactly reversing an event then has infinitesimal probability at most  $\epsilon^2$ ; the same bound holds for observing two concurrent events along the two edges leading to the children of a node. If the gene contents of the three neighbors of an internal node are known, we can determine the contents of the internal node from these assumptions.

When using GRAPPA to score a tree, however, the gene content of at least one neighbor of each internal node is unknown. We then use an iterative improvement algorithm (much in the style of the core algorithm in GRAPPA itself) to determine gene contents for internal nodes.

1. For each sibling pair of *leaves*, if a gene appears in both, we place it in the parent (an internal node); if it is absent from both, we do not place it in the parent. If, on the other hand, the gene appears in one leaf, but not the other, we mark its status as undetermined in the parent.
2. Starting from the root, we carry out a depth-first search of the tree to propagate resolutions (if two neighbors have the gene present, the node will have it too; if two neighbors are lacking that gene, so will the node) and to resolve undetermined states through look-ahead and cost propagation.

*Solving the median problem with unequal gene content:* Once we have determined the gene content in each node, we use an optimization procedure similar to the branch-and-bound currently used in GRAPPA to determine an optimal ordering for these genes. To make the greedy search more efficient, we developed a method to optimize the starting point of the branch-and-bound procedure. Our simulation results indicate that GRAPPA, equipped with our new median solver for unequal gene content, can reconstruct phylogenies with very high accuracy (no error on datasets of up to a dozen genomes). This new version of GRAPPA thus can form the basis for a new version of DCM-GRAPPA and extend our results to large datasets.