Phylogenetic Reconstruction from Arbitrary Gene-Order Data

Jijun Tang and Bernard M.E. Moret
University of New Mexico
Department of Computer Science
Albuquerque, NM 87131, USA
jtang,moret@cs.unm.edu

Liying Cui and Claude W. dePamphilis
Pennsylvania State University
Department of Biology and Institute of Molecular Evolutionary Genetics
208 Mueller Lab, University Park, PA 16802, USA
liying,cwd3@psu.edu

Abstract

Phylogenetic reconstruction from gene-order data has attracted 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 requires that all genomes have identical gene content, with each gene appearing exactly once in each genome. Some progress has been made in handling genomes with unequal gene content, both in terms of computing pairwise genomic distances and in terms of reconstruction. In this paper, we present a new approach for computing the median of three arbitrary genomes and apply it to the reconstruction of phylogenies from arbitrary gene-order data. We implemented these methods within* GRAPPA *and tested them on simulated datasets under various conditions as well as on a real dataset of chloroplast genomes; we report the results of our simulations and our analysis of the real dataset and compare them to reconstructions made by using neighbor-joining and using the original* GRAPPA *on the same genomes with equalized gene contents. Our new approach is remarkably accurate both in simulations and on the real dataset, in contrast to the distance-based approaches and to reconstructions using the original* GRAPPA *applied to equalized gene contents.*

1. Introduction

Reconstructing phylogenies from gene-order data has been studied intensely since the pioneering papers of Sankoff [3, 4, 19]. For smaller genomes, such as the single chromosome of organelles (mitochondria and chloroplasts), it has become possible to obtain the complete, ordered list of genes along the chromosome; animal mitochondria typically have around 40 genes, while chloroplast DNAs have around 120 genes. In such genomes, rearrangement of genes under inversion or other operations that change gene order (such as transposition) may form the principal evolutionary mechanism [7, 8]; other mechanisms may include duplication, insertion, and deletion of genes. Deletion is also a ubiquitous process in chloroplast genomes: an ongoing process of gene migration to the nucleus results in a low, but observable rate of gene loss from cp genomes in plants and algae ([13, 14]). In some cases, such as the cpDNA of a parasitic plant *Epifagus virginiana*, many deletion events, but only one inversion, separate its plastid genome from that of tobacco, a photosynthetic relative [23]. Thus, inversion and deletion represent two dominant processes in chloroplast genome evolution.

Because it uses the complete genome, gene-order data does not suffer from the notorious gene-tree vs. species-tree problem; and because rearrangements, insertions, and deletions of genes are rare events, gene-order data enables the reconstruction of events far back in the evolutionary history of organisms. Simulations studies conducted by our group [16, 22] indicate that gene-order data leads to very accurate reconstructions—far more accurate than those obtained from analyses of DNA sequence data.

However, gene-order data is much harder to work with than DNA sequence data. For instance, computing the *edit distance* between two genomes, (the smallest number of evolutionary events that can transform one genome into the other), an easy task with DNA sequence data, remains

Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04)
0-7695-2173-8/04 $ 20.00 © 2004 IEEE

IEEE COMPUTER SOCIETY

mostly unsolved for gene-order data—an exact solution is possible only when the permitted events are restricted to inversions and deletions. Almost all of the approaches to date have assumed that the genomes have equal gene content and do not contain any duplicate genes, both quite unrealistic assumptions (even if useful in first approximation). Recently, Marron *et al.* [12] gave a polynomial-time approximation algorithm to compute the distance between two genomes without any assumptions about their content. The most accurate reconstruction tool to date, our GRAPPA software, requires repeated (potentially tens of millions of times) computations of the *inversion median* of three genomes—a problem proved NP-hard some years ago [5]. While exact approaches to the median problem have been published [6, 20] and incorporated in the GRAPPA software [1], where they work well at the scale of organellar genomes [16, 22], these approaches only work for equal gene contents without duplication.

Recently, we proposed a simple approach to the median problem in the presence of limited duplications and deletions [21], but that approach, based on reducing the problem to one with equal gene contents by evaluating all possible resolutions of the duplications and deletions, does not scale to instances where the gene contents of the three genomes can differ significantly.

In this paper, we extend the approach of Siepel and Moret [20] to the median problem. We use a two-phase method, in which we begin by computing the best gene contents for the median, then use a branch-and-bound approach, with new lower bounds, to determine the best ordering of these gene contents. (New lower bounds are necessary, as the fixed gene contents of the median may make the lower bounds used in [20] inapplicable.) We then present the results of experimental tests on simulated datasets as well as on a biological dataset of green plant chloroplasts, The simulations show that our method produces very accurate results (no false positives and typically one false negative on datasets of 10 and 11 genomes) at reasonable costs—a typical dataset of 10 or 11 genomes takes from a few minutes to a few hours of computation. The biological dataset, which contains genomes with both high and low rearrangement rates, illustrates the power of our approach: our method reconstructs the tree posited by biologists and also reproduces the uncertainty among them about the position on the tree of one of the species. In contrast, neighbor-joining (using a linear measure of inversion and deletion distance) returns a tree with false positives and GRAPPA run on datasets with equalized gene contents returns a tree with only one resolved edge when using the breakpoint distance and with errors (and huge computation times) when using the inversion distance. In summary, our new method is both faster and more accurate than previous approaches.

We recently demonstrated [22] that GRAPPA can be scaled up from 10–14 genomes to a thousand genomes (with equal gene content) by combining it with the Disk-Covering methods (DCM) of Warnow and her group; the same approach will work with our new methodology for unequal gene content, enabling us finally to realize the promise of gene-order data in phylogenetic reconstruction.

2. Definitions and Notation

Suppose a dataset has N genomes and a fixed set of n genes $S = g_1, g_2, \ldots, g_n$. Each genome G_i contains a subset $S_i = \{g_{i_1}, \ldots, g_{i_k}\}$ of these genes (with $k \leq n$); we call S_i the *gene content* of G_i. Then the genome G_i can be represented as a *signed permutation* $\pi_i = (\pi_{i_1}, \ldots, \pi_{i_k})$ defined on subset S_i.

The *median problem* for a set of permutations $\pi_i = \pi_1, \pi_2, \ldots, \pi_m$ is to find a permutation π_M that minimizes the *median score*, $\sum_{i=1}^{m} d_{i,M}$, where $d_{i.M}$ is the pairwise edit distance between π_i and π_M. We denote the optimal (minimal) median score by $D(M)$. In phylogenetic practice, we examine binary trees, so that the median problem has $m = 3$.

For any genome, we can define an undirected graph $G = (V, E)$, in which each vertex in V corresponds to a signed permutation and an edge connects two vertices v_i and v_j if and only if v_i can be transformed into v_j by a single inversion, deletion, or insertion. We call this graph the *evolution graph*; if we only allow inversions, then the graph is reduced to an *inversion graph*. In such a graph, the *neighbors* of a permutation π are those permutations that can be obtained from π by subjecting it to one evolutionary event; if only inversions are allowed, then π has $\binom{n}{2}$ neighbors. A *shortest path* between two permutations π_i and π_j is the simple path of shortest length in G between the vertices corresponding to π_i and π_j.

3. An Algorithm for the Median Problem

GRAPPA has two methods to solve the problem of inversion medians. One was developed by Caprara and is based on an extension of the breakpoint graph; its foundation relies on equal gene contents and thus makes it very difficult to extend to events such as insertions and deletions. The other one is a branch-and-bound method developed by Siepel and Moret [20]; it is slower than Caprara's method because of a rather loose bound (we will show a better bound in the next subsection), but it can use any definition of distance and thus forms a good starting point for the development of our new median solver. We now briefly review that algorithm.

IEEE COMPUTER SOCIETY

3.1. The Algorithm of Siepel and Moret

This algorithm uses a simple branch-and-bound approach:

- Given the three permutations π_1, π_2 and π_3, compute the lower bound on the median score, $D(M) = \frac{d_{1,2}+d_{2,3}+d_{3,1}}{2}$.
- Pick one permutation from π_1, π_2, π_3 as start (a so-called *trivial median*) and push it into a queue; its median score is the initial best-so-far.
- Iteratively remove a permutation π from the queue until the queue is empty:
  - If the median score of π meets the lower bound, $d_{\pi,1} + d_{\pi,2} + d_{\pi,3} = D(M)$, then stop.
  - If the median score of π is less than the current best-so-far, update the latter, create all $\binom{n}{2}$ neighboring permutations (one inversion away from π), discard those with lower bounds exceeding the best-so-far, and queue up the surviving ones.

Clearly, the success of this algorithm relies on good lower bounds. In [20], the authors proved two bounds:

Lemma 1 *For three permutations π_1, π_2 and π_3, the optimal median score, $D(M)$, obeys:*

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

Lemma 2 *If ϕ is a permutation on the shortest path from π_1 to the median, then it obeys:*

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

The left side of Lemma 2 is the lower bound used in GRAPPA to check if a new permutation can be discarded; in practice, this bound is loose and the search space remains too large even for datasets with small edge lengths (with an average of 4 evolutionary events per tree edge).

We now provide an improved lower bound.

Lemma 3 *If ϕ is a permutation on the shortest path from π_1 to the median and γ is derived from ϕ by applying one inversion, then, if γ is also on the shortest path from π_1 to M, it obeys $d_{1,\gamma} + d_{2,\gamma} + d_{3,\gamma} \leq d_{1,\phi} + d_{2,\phi} + d_{3,\phi} + 1$*

Figure 1 illustrates the situation.

Proof From the triangle inequality, we have $d_{2,\gamma} + d_{3,\gamma} \leq d_{2,\phi} + d_{3,\phi}$; since we have $d_{\phi,\gamma} = 1$, we can write $d_{1,\phi} + 1 \leq d_{1,\gamma}$ and we immediately get

$$d_{1,\gamma} + d_{2,\gamma} + d_{3,\gamma} \leq d_{1,\phi} + d_{2,\phi} + d_{3,\phi} + 1$$

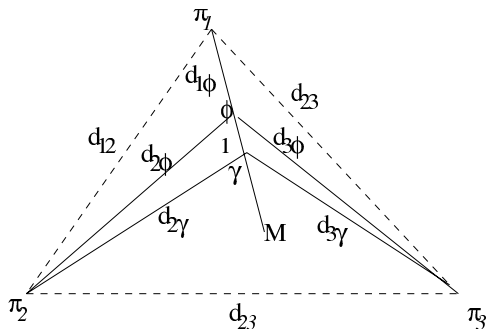

Figure 1. An illustration of Lemma 3

If the bound is not met, then either ϕ or γ is not on the shortest path; in other words, γ (and perhaps ϕ as well) should be discarded. However, it is hard to verify that the failure is due to ϕ; hence, in our implementation, we discard only γ. This bound is much tighter than the original bound of Siepel and Moret and no more expensive to compute; using it in the median solver achieves very significant speed gains.

3.2. New Median Bounds

Lemma 3 depends on the existence of a trivial median. When we want to handle unequal gene contents, it is not always possible to find a trivial median (as we explain below), hence we must find a new way to compute a good lower bound.

The following lemma is the immediate extension of Lemma 1:

Lemma 4 *Given three permutations π_1, π_2, and π_3, and another permutation π_0, then the optimal median score, $D(M)$, obeys:*

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

If the branch-and-bound process starts with a permutation that is not one of π_1, π_2 or π_3, then we can use the following bound:

Lemma 5 *Given three permutations π_1, π_2, and π_3, and another permutation π_0 close to π_1, if a permutation ϕ is on the shortest path from π_0 to the optimal median permutation, then the optimal median score $D(M)$ obeys:*

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

Figure 2 illustrates the situation.

Proof Since ϕ is on the shortest path from π_0 to M, then, as proved in [20], it obeys

$$d_{0,\phi} + \frac{d_{2,\phi} + d_{3,\phi} + d_{2,3}}{2} \leq d_{0,M} + d_{2,M} + d_{3,M}$$

IEEE COMPUTER SOCIETY

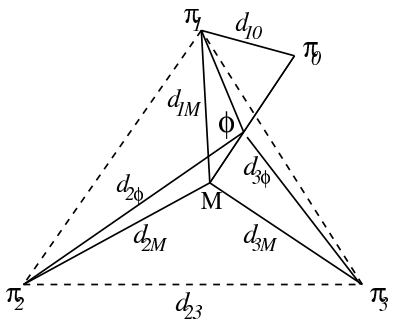

Figure 2. An illustration of Lemma 5

Now, because we have $d_{0.M} \leq d_{0,1} + d_{1,M}$, we can write

$$d_{0,\phi} + \frac{d_{2,\phi} + d_{3,\phi} + d_{2,3}}{2} \leq d_{1,M} + d_{2,M} + d_{3,M} + d_{0,1}$$

which directly implies

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

If π_0 coincides with one of the three given permutations, then this lower bound is the same as defined in Lemma 2. In the same manner, we can extend Lemma 3.

Lemma 6 *Given permutations π_1, π_2, and π_3, and permutation π_0 close to π_1, if a permutation ϕ is on the shortest path from π_0 to the optimal median permutation and its neighbor γ is also on the shortest path, then γ obeys*

$$d_{1,\gamma} + d_{2,\gamma} + d_{3,\gamma} \leq d_{1,\phi} + d_{2,\phi} + d_{3,\phi} + 1$$

3.3. The New Median Solver

Using the new set of bounds, we can easily extend the algorithm of Siepel and Moret to permutations with unequal gene content. Before solving the median problems, we determine the gene content for each internal node, as discussed in the next section. Once these contents have been determined, we proceed to determine the gene orders. Thus we need to pick starting permutations that respect the predetermined contents at each internal node, after which we can start the branch-and-bound search as in the algorithm of Siepel and Moret, except that we now use the distance computation for unequal gene content and the new bounds. Note that the number of neighbors of a permutation remains $\binom{n}{2}$: that is, we only take into account inversions in the search, since insertions and deletions were accounted for in the process of assigning gene contents.

From Lemmata 5 and 6, when $d_{0,1}$ is small and π_0 is very close to π_1, the lower bound will be tight. Our experiments confirm that the choice of π_0 has a huge impact on the search procedure. For instance, choosing a random permutation of the assigned gene content results in very poor performance. The best method we found is to choose a permutation π_0 which is closest to one of the three given permutations and which minimizes the initial upper bound $d_{0,1}+d_{0,2}+d_{0,3}$. Since finding such a permutation is itself a complex optimization problem, we use a simpler heuristic:

- From π_1, create π_0 by first deleting all genes not in the median content, then inserting all genes needed to complete the median content in a single block to the end; now compute $d_0 = d_{0,1}$ and $l = d_{0,2} + d_{0,3}$.
- Repeat the above procedure for π_2 and π_3.
- Pick that π_0, if any among the three constructed, which minimizes both d_0 and l; otherwise, pick that which minimizes d_0 (because the lower bound is more important in pruning than the initial upper bound).

4. Determining the Gene Content of Internal Nodes

The gene content is determined based on the same assumptions that we used in earlier work [21], namely that deletions and insertions are rarer events than inversions and that concurrent change in two neighbors is much less likely than a single complementary change from the third neighbor. (That is, whenever we face the choice of assigning the gene loss or gain to the parent or to both of its children, we always assign it to the parent.) Thus, at each internal node, when the contents of its two children are known, we face three possibilities in deciding the presence or absence of a gene g:

1. If both children have g, then g should be in the median; otherwise, g would be inserted into both children—with vanishingly small probability.
2. If neither child has g, then, for the same reason, g should not be in the median.
3. If one child has g and the other does not, then, because the tree is unrooted, we face a deletion or insertion and cannot assign a higher probability to one than to the other. If we also know the gene content of the parent, we can break the tie in the obvious way; otherwise, we are left with an undetermined outcome for g.

Cases (1) and (2) have been used by biologists to construct phylogenies (e.g., [15]). When the tree is rooted, then the problem is greatly simplified, because gene loss is much more likely than gene gain [13, 14, 15].

If a gene is undetermined at some internal node, it may become determined through a propagation of content decisions from the leaves (of known prior content) to the root. However, `GRAPPA` only deals with unrooted trees. (It does use a temporary root in its computation, but this root is picked arbitrarily and is thus unlikely to be the biologically correct choice.) Thus, in order to resolve undetermined gene contents at internal nodes, we decided to use

IEEE COMPUTER SOCIETY

an iterative improvement algorithm similar to the core algorithm in GRAPPA itself:

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 the gene appears in one leaf, but not the other, we mark its status as undetermined in the parent.
2. Starting from the (arbitrary) root, we carry out a depth-first search of the tree to propagate resolutions according to our standard rule (if two neighbors have the gene, so will the node; if two neighbors are lacking that gene, so will the node) and thus to resolve undetermined states through look-ahead and cost propagation.

This model extends naturally to gene duplications: if g is duplicated, with one child having s_1 copies and the other s_2 copies, then we have:

- With $s = s_1 = s_2$, then the node has s copies of g.
- With $s_1 > s_2$, then the node has at least s_2 copies of g, with another $s_1 - s_2$ undetermined copies.

Undetermined copies are again resolved in the iterative improvement phase through propagation.

5. Putting the Pieces Together

Our new method can be summarized as follows:

- Compute the NJ tree and use its score as the initial upper bound.
- For each possible tree:
  - Test the lower bound based on the distance matrix.
  - If the lower bound exceeds the upper bound, discard the tree and move to the next.
  - Determine the gene content of each internal node, initialize the gene order of internal nodes, and iteratively solve the median problem until no change occurs.
  - Update the upper bound if this tree's score improves it.
- Return the tree(s) with the lowest score.

5.1. Distance Computations

We implemented a linear-time algorithm to compute the distance for two genomes with deletion and insertion, which is based on El-Mabrouk's algorithm [9]. Because there is no polynomial-time algorithm available for genomic distances under a combination of insertions, deletions, and duplications, we adopted the renaming strategy presented in [21] to handle cases when genes are duplicated, which is appropriate for our intended genomes of cpDNAs—as opposed to the more general, but more complex and, for smaller genomes, less accurate method of Marron *et al.* [12].

True distance estimators have proved very useful both in distance-based reconstruction and with GRAPPA [16]—with equal gene content in both cases. We have no true distance estimator as yet for unequal gene content; however, since we separately compute the number of inversions and that of insertions/deletions, we use our EDE correction [17] for the number of inversions, thereby obtaining a partial distance correction that both decreased the running time and improved the quality of reconstruction.

5.2. Initialization

The median problem requires the gene orders of the three neighboring nodes to be known. Initially, however, none of the internal nodes of a tree has a known gene order, so each must be initialized in some fashion. In previous work [18, 22], we showed that this initialization is crucial to both speed and quality of reconstruction with GRAPPA. Among the various initialization methods used in GRAPPA, the *nearest-neighbor* method is the best: it picks the three closest (in terms of number of edges) leaves to the internal node, then solves the median problem of these three leaves and assigns the resulting gene order to that internal node. We use the same strategy in our new system, with some enhancements. The gene content determined from the scope of the whole tree may differ from that obtained by considering only the three nearest leaves; in that case, we can use the gene content determined by the three nearest leaves and later add to the resulting gene order the additional genes needed to complete the gene content determined in the first phase. (Our experiments show that simply deleting the unwanted, or inserting the missing, genes introduces large errors.) Because the three nodes are further away from the internal node than the three neighbors, the median problems in the initialization procedure are typically harder than those encountered in the iterative improvement stage. Since we just want reasonable initial gene orders for a start, but do not really require optimal solutions to the leaf-median problems, we increase the lower bound by a fixed factor (e.g., 10%), which may eliminate some good solutions, but allows the initialization to run in time comparable to the scoring phase.

6. Experimental Results

6.1. Simulations

For the simulation study, we chose datasets of 10 and 11 genomes; these sizes, while appearing small, in fact formed the bulk of the subproblems solved by our DCM-GRAPPA when working on datasets of one thousand taxa [22].

IEEE COMPUTER SOCIETY

We chose genome sizes of 50 and 100 (roughly matching typical animal mitochondria and land plant chloroplast genomes, respectively). Finally, we used three evolutionary rates, of 2, 4 and 6 expected events per tree edge—the last one representing a high rate of evolution even on small datasets. In our simulations, each tree node (internal and external) has a 5% chance to lose one segment of genes; the length of this segment is at most 10% of the number of genes in its parent. For each combination of parameter settings, we ran 10 datasets; all our tests were run on 1.8GHz Pentium 4 desktop with 512MB of memory.

Tables 1 and 2 show the average false positives and false negatives for each dataset. The new method achieves high accuracy: for each dataset, the expected number of edges in error is considerably less than one. Running times (not shown) are about $10 \sim 15$ times slower than what we have typically seen with equal gene content (using Caprara's median solver). Given the significant additional complexity of the task, this is quite acceptable: a typical 12-genome dataset takes about 15–20 minutes to complete.

	$r=2$		$r=4$		$r=6$	
10	0.2	0.6	0.1	0.7	0.2	0.7
11	0.0	0.8	0.1	0.5	0.0	0.9

	$r=2$		$r=4$		$r=6$	
10	0.1	0.6	0.1	0.3	0.1	0.4
11	0.2	0.5	0.3	0.3	0.2	0.6

Table 2. Average number of edges in error for 50 genes

6.2. A cpDNA Dataset

Molecular phylogenies using concatenated plastid genes and different methods provide clear evidence that the chloroplast genomes are derived from a cyanobacteria-like ancestral genome with many subsequent gene losses and gene transfers to the nucleus [13]. Typical chloroplast genomes or cpDNAs are circular, encoding 50-200 genes involved in transcription, translation, ATP synthesis, electron transport, photosynthesis and other functions. Most chloroplast DNAs (cpDNA) include two almost identical regions in opposite orientation, called the *inverted repeat (IR)*. Genes located in the inverted repeats are therefore duplicated.

Genome rearrangement is frequently observed in algal chloroplast genomes, including the expansion and deletion of inverted repeats, a unique feature of chloroplast genomes. Cosner *et al.* [7] first tested several phylogenetic methods on the cpDNA gene-order data of the flower plant family Campanulaceae and discovered an unusual variety of rearrangements; later, this dataset was extensively analyzed by our group, with the best reconstructions provided by an inversion-only estimation.

We extracted 77 genes from each of 7 fully sequenced chloroplast genomes, with *Chlorella vulgaris* missing 3 genes. The data set includes land plants (*Nicotiana tabacum*, *Marchantia polymorpha*), a coleochaetales green alga (*Chaetosphaeridium globosum*), where most cpDNAs have very few rearrangements, as well as chorophytic green algae (*Nephroselmis olivacea*, *Chlorella vulgaris*, and *Chlamydomonas reinhardtii*) with extensive rearrangements in cpDNA gene order. This implies very heterogeneous rates of evolution along the branches, which is a unique challenge to phylogenetic reconstruction—one that we did not include in the simulation study. Chloroplast genomes from most photosynthetic land plants and the coleochaetales algae, including *Chaetosphaeridium globosum*, share nearly identical gene content and gene order. *Chlamydomonas reinhartii* is closely related to the green alga *Chlorella*, while the latter contains only one copy of the IR; however, the cpDNA of *Chlamydomonas* is highly rearranged with respect to related green algal cpDNA sequences [14]. The phylogenetic position of the dinoflagellate *Mesostigma viride* is not fully resolved: it is either an early branch off all the green plants or more closely related to the Charophycean green algae [2, 10, 11]. We ran maximum parsimony and neighbor-joining analyses using concatenated protein sequences of 50 chloroplast genes, with *Cyanophora paradoxa* as the outgroup, and used the bootstrap consensus phylogeny (using PAUP 4.0b with 1,000 bootstrap replicates) as a reference tree; each node had bootstrap support 60 or higher in both analyses.

We then ran four different reconstruction methods on this gene-order data: (i) our new method; (ii) neighbor-joining on a distance matrix computed by using our exact linear-time algorithm for inversion/deletion distance; (iii) the regular `GRAPPA` code using breakpoint medians on a reduced dataset with the largest common gene content; and (iv) the same as in (iii), but using inversion medians.

Our method returned two phylogenies with equal score. We presented the one with higher congruence to the reference phylogeny. The other was only different in the relationship of *Chaetosphaeridium* and land plants. The reference phylogeny and phylogenies returned by methods (i) through (iii) are shown in Figure 3. In the reference phylogeny, the position of *Mesostigma* is unresolved due to low bootstrap support. The phylogeny returned by the new method is congruent with the reference phylogeny. (The neighbor-joining phylogeny, in contrast, introduces false positive edges among the three algae *Mesostigma*, *Chlamydomonas*

IEEE COMPUTER SOCIETY

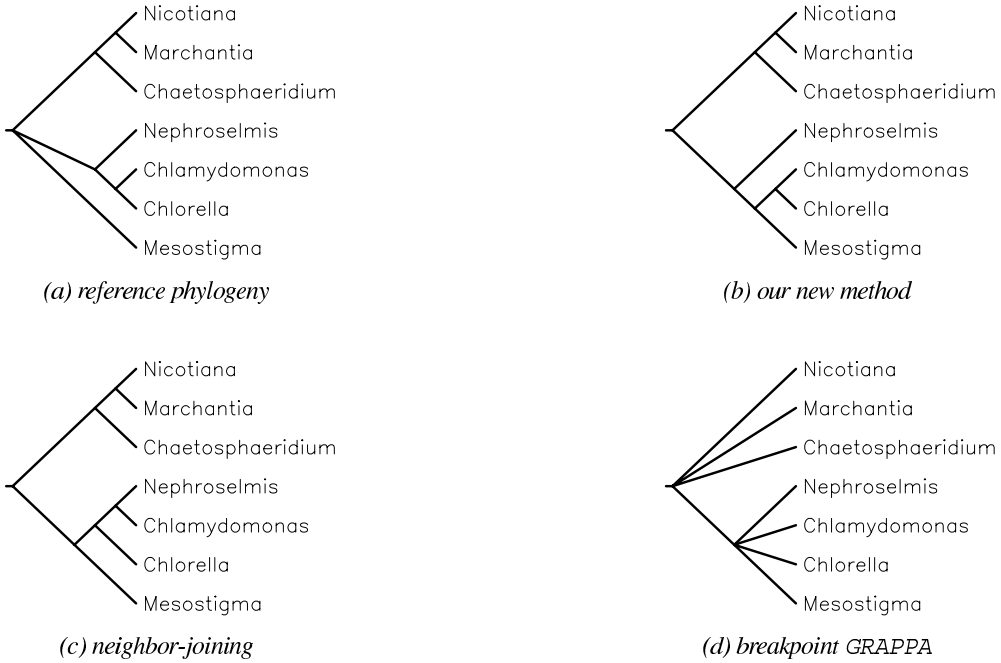

Figure 3. Phylogenies on the 7-taxa cpDNA dataset

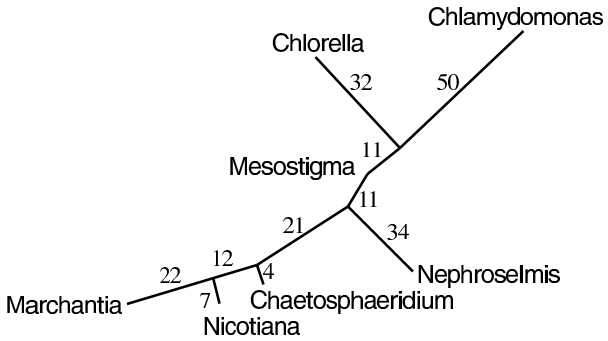

Figure 4. The phylogeny returned by our methods, showing estimated branch lengths

and *Chlorella*.) Our new method yields one polytomy due to a branch of zero length, but it involves *Mesostigma*, whose exact position in the tree is subject to debate [2, 10, 11]. As seen in Figure 4, the branch lengths, representing the number of inferred rearrangements, differ significantly among the land plant cluster (generally short branches) and the green algae cluster (generally long branches), a finding in agreement with previous observations [14].

7. Conclusion

We presented the first computational approach for the reconstruction of phylogenies from arbitrary gene-order data. Results from our simulations indicate that our approach is remarkably accurate; results from the real dataset confirm this finding. The real dataset posed phylogenetic challenges that we did not encounter in our simulations. The placement of Mesostigma is not well resolved by the data, which echoes the conflicts of Mesostigma position in DNA and protein based analyses. The major concordance of our results and the sequence-based results suggest the potential of our method for solving difficult deep phylogeny questions. These results are in contrast to the distance-based approaches and to our same method (GRAPPA) applied to equalized gene contents. The difference between the results is in fact striking, a stark reminder of how much informa-

IEEE COMPUTER SOCIETY

tion is lost when the gene contents are equalized. While we have only tested the method on small datasets, the approach can easily be extended to large datasets by using the Disk-Covering Method as we did recently in developing DCM-GRAPPA [22]; thus our approach will extend to datasets with hundreds of taxa and moderate gene losses and duplications. Extending our approach to much larger nuclear genomes remains a major computational challenge.

8. Acknowledgments

Our work is supported by the National Science Foundation under grants DEB 01-20709 (dePamphilis, Moret), EF 03-31654 (Moret), EIA 02-03584 (Moret), EIA 01-13095 (Moret), EIA 01-21377 (Moret), DBI-0115684 (Cui, dePamphilis), and by a grant from IBM Corporation (Moret).

References

[1] D. Bader, B. Moret, J. Tang, and T. Warnow. *GRAPPA (Genome Rearrangements Analysis under Parsimony and other Phylogenetic Algorithms)*. `www.cs.unm.edu/~moret/GRAPPA/`.

[2] D. Bhattacharya, K. Weber, S. An, and W. Berning-Koch. Actin phylogeny identifies Mesostigma viride as a flagellate ancestor of the land plants. *J. Mol. Evol.*, 47:544–550, 1998.

[3] M. Blanchette, G. Bourque, and D. Sankoff. Breakpoint phylogenies. In S. Miyano and T. Takagi, editors, *Genome Informatics*, pages 25–34. Univ. Academy Press, Tokyo, 1997.

[4] M. Blanchette, T. Kunisawa, and D. Sankoff. Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.*, 49:193–203, 1999.

[5] A. Caprara. Formulations and hardness of multiple sorting by reversals. In *Proceedings of the 3rd Int'l Conference on Comput. Mol. Biol. RECOMB99*, pages 84–93. ACM Press, 1999.

[6] A. Caprara. On the practical solution of the reversal median problem. In *Proc. 1st Int'l Workshop Algorithms in Bioinformatics (WABI01)*, volume 2149 of *Lecture Notes in Computer Science*, pages 238–251. Springer-Verlag, 2001.

[7] M. Cosner, R. Jansen, B. Moret, L. Raubeson, L.-S. Wang, T. Warnow, and S. Wyman. An empirical comparison of phylogenetic methods on chloroplast gene order data in Campanulaceae. In D. Sankoff and J. Nadeau, editors, *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment, and the Evolution of Gene Families*, pages 99–121. Kluwer Academic Pubs., Dordrecht, Netherlands, 2000.

[8] S. Downie and J. Palmer. Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In D. Soltis, P. Soltis, and J. Doyle, editors, *Molecular Systematics of Plants*, pages 14–35. Chapman and Hall, New York, 1992.

[9] N. El-Mabrouk. Genome rearrangement by reversals and insertions/deletions of contiguous sements. In *Proc. 11th Ann. Symp. Combinatorial Pattern Matching (CPM'00)*, volume 1848 of *Lecture Notes in Computer Science*, pages 222–234, 2000.

[10] K. Karol, R. McCourt, M. Cimino, and C. Delwiche. The closest living relatives of land plants. *Science*, 294:2351–2353, 2001.

[11] C. Lemieux, C. Otis, and M. Turmel. Ancestral chloroplast genome in Mesostigma viride reveals an early branch of green plant evolution. *Nature*, 403:649–652, 2000.

[12] M. Marron, K. Swenson, and B. Moret. Genomic distances under deletions and insertions. In *Proc. 9th Int'l Conf. Computing and Combinatorics (COCOON03)*, volume 2697 of *Lecture Notes in Computer Science*, pages 537–547. Springer-Verlag, 2003.

[13] W. Martin, B. Stoebe, V. Goremykin, S. Hansmann, and M. Hasegawa. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature*, 393:162–165, 1998.

[14] J. Maul, J. Lilly, L. Cui, C. dePamphilis, W. Miller, E. Harris, and D. Stern. Chlamydomonas chloroplast chromosome: islands of genes in a sea of repeats. *Plant Cell*, 14:2659–2679, 2002.

[15] A. McLysaght, P. Baldi, and B. Gaut. Extensive gene gain associated with adaptive evolution of poxviruses. *Proc. Nat'l Acad. Sci., USA*, 100:15655–15660, 2003.

[16] B. Moret, J. Tang, L.-S. Wang, and T. Warnow. Steps toward accurate reconstruction of phylogenies from gene-order data. *J. Comput. Syst. Sci.*, 65(3):508–525, 2003.

[17] B. Moret, L.-S. Wang, T. Warnow, and S. K. Wyman. New approaches for reconstructing phylogenies from gene-order data. In *Proc. 9th Int'l Conf. on Intelligent Systems for Molecular Biology (ISMB01)*, volume 17 of *Bioinformatics*, pages S165–S173. Oxford U. Press, 2001.

[18] B. Moret, S. K. Wyman, D. A. Bader, T. Warnow, and M. Yan. A new implementation and detailed study of breakpoint analysis. In *Proc. 6th Pacific Symp. on Biocomputing (PSB01)*, pages 583–594. World Scientific Pub., 2001.

[19] D. Sankoff and M. Blanchette. Multiple genome rearrangement and breakpoint phylogeny. *J. Comput. Biol.*, 5:555–570, 1998.

[20] A. Siepel and B. Moret. Finding an optimal inversion median: experimental results. In *Proc. 1st Int'l Workshop Algorithms in Bioinformatics (WABI01)*, volume 2149 of *Lecture Notes in Computer Science*, pages 189–203. Springer-Verlag, 2001.

[21] J. Tang and B. Moret. Phylogenetic reconstruction from gene rearrangement data with unequal gene contents. In *Proc. 8th Workshop on Algorithms and Data Structures (WADS03)*, volume 2748 of *Lecture Notes in Computer Science*, pages 37–46. Springer-Verlag, 2003.

[22] J. Tang and B. Moret. Scaling up accurate phylogenetic reconstruction from gene-order data. In *Proc. 11th Int'l Conf. on Intelligent Systems for Molecular Biology (ISMB03)*, volume 19 of *Bioinformatics*, pages i305–i312. Oxford U. Press, 2003.

[23] K. Wolfe, C. Morden, and J. Palmer. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc. Nat'l Acad. Sci., USA*, 89:10648–10652, 1992.

IEEE COMPUTER SOCIETY