Phylogenetic Reconstruction from Gene-Rearrangement Data with Unequal Gene Content

Jijun Tang and Bernard M.E. Moret

Dept. of Computer Science, University of New Mexico, Albuquerque, NM 87131, USA
email jtang,moret@cs.unm.edu, URL http://compbio.unm.edu

Abstract. Phylogenetic reconstruction from gene-rearrangement data has seen increased attention over the last five years. Existing methods are limited computationally and by the assumption (highly unrealistic in practice) that all genomes have the same gene content. We have recently shown that we can scale our reconstruction tool, GRAPPA, to instances with up to a thousand genomes with no loss of accuracy and at minimal computational cost. Computing genomic distances between two genomes with unequal gene contents has seen much progress recently, but that progress has not yet been reflected in phylogenetic reconstruction methods. In this paper, we present extensions to our GRAPPA approach that can handle limited numbers of duplications (one of the main requirements for analyzing genomic data from organelles) and a few deletions. Although GRAPPA is based on exhaustive search, we show that, in practice, our bounding functions suffice to prune away almost all of the search space (our pruning rates never fall below 99.995%), resulting in high accuracy and fast running times. The range of values within which we have tested our approach encompasses mitochondria and chloroplast organellar genomes, whose phylogenetic analysis is providing new insights on evolution.

Keywords

computational biology, phylogenetic reconstruction, gene-order data, whole-genome data, signed permutations, lower bounds, Hannenhalli-Pevzner theory, inversion distance, reversal distance, edit distance, gene duplications, experimental assessment

1 Introduction

A phylogeny is the evolutionary history of a group of organisms; in most cases, it is represented (in obviously simplified form) by a tree where the leaves represent current organisms and the internal nodes represent ancestral organisms, and where the edges denote evolutionary relationships. Such phylogenies have long been reconstructed on the basis of morphological data and more recently on the basis of molecular data such as DNA sequence data.

Biologists can infer the ordering and strandedness of genes on a chromosome, and thus represent each chromosome by an ordering of signed genes (where the sign indicates the strand). These gene orders can be rearranged by evolutionary events such as inversions (also called reversals) and transpositions and, because they evolve slowly (much more slowly, for instance, than DNA sequences), give biologists an important new source of data for phylogeny reconstruction (see, e.g., [10, 20, 21, 23]). Appropriate tools for analyzing such data may help resolve some difficult phylogenetic reconstruction problems. Developing such tools is thus an important area of research—indeed, the recent DCAF symposium [27] was devoted to this topic.

A natural optimization problem for phylogeny reconstruction from gene-order data is to reconstruct an evolutionary scenario with a minimum number of the permitted evolutionary events on the tree. This problem is NP-hard for most criteria—even the very simple problem of computing the median[1] of *three* genomes with identical gene content under such models is NP-hard [7, 22]—although the algorithms of Caprara [8] and of Siepel and Moret [28] have done well in practice (see, e.g., [18]). Indeed, even the problem of computing the edit distance between two genomes is difficult: for instance, even with equal gene content and with only inversions allowed, the problem is NP-hard for unsigned permutations [6].

2 Background

2.1 Genomic distances

Hannenhalli and Pevzner [12] made a fundamental breakthrough by developing an elegant theory for signed permutations and providing a polynomial-time algorithm to compute the edit distance (and the corresponding shortest edit sequence) between two signed permutations under inversions; Bader et al. [2] later showed that this edit distance can be computed in linear time. El-Mabrouk [11] extended the results of Hannenhalli and Pevzner to the computation of edit distances for inversions and deletions and also for inversions and non-duplicating insertions; she also gave an approximation algorithm with bounded error for computing edit distances in the presence of all three operations (inversions, deletions, and non-duplicating insertion). Sankoff had proposed the so-called exemplar strategy [25] (itself an NP-hard problem [4]) to handle duplications: only one copy of each gene is retained, that which minimizes a breakpoint scoring function. Experiments we conducted suggested that too much information is lost in reducing the genomes to a single copy of each gene; working to use all duplicates in the computation, our group recently extended the work of El-Mabrouk by providing tight approximations for edit distances under arbitrary operations (including duplications) [17].

2.2 Gene-order reconstruction

Extending the computation of genomic distances to genomes with unequal gene contents is but the first step in a reconstruction effort. While it is possible to reconstruct the tree's topology on the basis of pairwise distances only (using standard methods such as neighbor-joining [24]), reconstructing ancestral genomes requires additional steps. Sankoff had proposed an iterative strategy which he called breakpoint analysis [26],

[1] The median of k genomes is a genome that minimizes the sum of the pairwise distances between itself and each of the k given genomes.

which we subsequently improved by combining it with our fast inversion distance computation and various speedup heuristics to produce the software suite GRAPPA [1, 19]. Other approaches include classical parsimony analysis based on binary encodings of the genome data [9], fast heuristic uses of the reversal distance for ancestral genome reconstruction [3], and a recent endeavor based on likelihood maximization [16].

2.3 GRAPPA

GRAPPA is based on Sankoff's breakpoint analysis. It works by enumerating every possible tree topology for the given collection of organisms and, for each tree, by reconstructing the ancestral genomes associated with the internal nodes of the tree, thereby making it possible to score the tree. The trees of lowest score are then returned. Ancestral genomes are reconstructed through iterative refinement: on successive traversals of the tree, each ancestral genome is compared with the median of its three neighbors and replaced by that median if the tree score is thereby improved. Since computing the median is itself an NP-hard problem, scoring each tree is computationally intensive and should be avoided if at all possible. We devised and built into GRAPPA an effective bounding scheme, which runs in linear time, to prune most candidate trees without having to score them, using nothing more than the triangle inequality. The resulting speed-up (of up to a billion-fold on many datasets) enabled us to solve datasets of up to 15 genomes. Most recently, we combined the disk-covering approach of Warnow and her colleagues [13–15] with GRAPPA, thereby scaling up the approach to up to one thousand genomes with no loss of accuracy [29].

3 Our Approach to Duplication

We assume a fixed set of genes $\{g_1, g_2, ..., g_k\}$. Let $d_i \geq 1$ be the number of copies of g_i, which we assume to be equal for all genomes—unequal numbers of duplicates introduce the possibility of deletions, which we address in a later section. Since the number of copies for a gene is identical for all genomes, we can define the multiset

$$\{\underbrace{g_1, \cdots, g_1}_{d1}, \underbrace{g_2, \cdots, g_2}_{d2}, \cdots, \underbrace{g_k, \cdots, g_k}_{dk}\}$$

and each genome is then an ordering (circular or linear) of this superset, with each gene copy given an orientation (sign).

We assume that copies of a gene are rearranged as if they were distinct genes; by renaming the copies, we then obtain a signed permutation of a set of $\sum_{i=1}^{k} d_i$ distinct genes. For example, given the ordering $(1, 2, -3, 4, 3, 5)$, in which gene 3 appears twice, we can relabel one of the copies as gene 6, yielding two possible new orderings: $(1, 2, -3, 4, 6, 5)$ and $(1, 2, -6, 4, 3, 5)$, both signed permutations of the set $\{1, 2, 3, 4, 5, 6\}$. We call the collection of $\Pi_{i=1}^{k} d_i$ possible new orderings obtained through the relabeling of copies as new genes a *differentiated genome family*. Each genome in the input data has a family of that size. We can then define the inversion distance between two genomes with identical duplications as the minimum pairwise inversion distance between a member of the differentiated family of one genome and a member of the family of the other genome. Because inversion distances between genomes with equal gene content can be computed efficiently in linear time, this definition can be computed quickly for modest numbers of duplications by checking all $\Pi_{i=1}^{k} d_i^2$ pairs.

To solve the median problem—the central computational problem for GRAPPA—we can extend this simple idea and consider all triples of elements from the three differentiated families, for a total of $\Pi_{i=1}^{k} d_i^3$ possibilities. This number can quickly grow uncomfortably large since each median computation is potentially very expensive, so we need to avoid as many of these computations as possible. We can use the same bounding strategy at this stage as is used by GRAPPA in bounding the cost of individual trees: by the triangle inequality, the sum of the distances from the median to its three neighbors is at least as large as half of the sum of the three pairwise distances between the three neighbors. (Bryant [5] developed a slightly tighter bound, but it is limited to breakpoint distances and our previous experimental work [19] showed that it is too slow and gains too little to be useful in pruning the search space.) We can compute the pairwise distances in linear time and avoid computing the median of a particular triple of family members whenever their lower bound exceeds the current best median score. (We still need to examine all $\Pi_{i=1}^{k} d_i^3$ triples of family members, but our intended application, to organellar genomes with only a hundred or so genes and typically fewer than 10 duplicates in all, yields reasonable values for this product.) Clearly, we will get better bounding if we can start with some reasonable choice of median; in particular, the choice of initial family members for the genomes at the leaves has a huge impact on the pruning rate for median computations. We found the following initialization method to be very effective: for each leaf genome, we pick that member of the differentiated family which minimizes the sum of the minimum pairwise distances to the other leaf genomes.

4 Experimental Results for Duplications

We ran simulation tests on trees of 10, 11, and 12 genomes (sizes easily handled by the basic GRAPPA), under two different models of topologies (uniform random trees and birth-death trees) and three different rates of evolution (with r, the expected number of inversions per edge, set to 2, 4, and 8). For a given rate of evolution r, we generate an actual number of evolutionary events for each edge by using a random integer in the set $\{0, 1, \ldots, 2r\}$. We then start with the identity permutation on the genes at the root of the tree topology and evolve permutations down the tree by applying to the parent permutation the number of events prescribed by the edge; each event is an inversion, with its two endpoints chosen independently and uniformly at random. All genomes have a total of 100 genes; a duplication is generated by selecting two of the genes at random and calling them duplicates of each other. We used three scenarios with limited duplications: (i) one gene is duplicated once; (ii) two genes are duplicated, one once and the other twice; and (iii) three genes are duplicated, two once and one twice. Overall, then, we used 54 combinations of parameters; we generated 20 datasets for each combination and report in the tables below the average value of the 20 runs. We ran all of our experiments on a 2.4GHz desktop Pentium-4 machine with 1GB of memory running Linux; running times naturally increased with the number of genomes, but, even at 12

Table 1. Average numbers of edges in error for one duplication: (a) uniform trees, (b) birth-death trees.

(a)

	$r=2$		$r=4$		$r=8$	
n	FP	FN	FP	FN	FP	FN
10	0	0	0	0.05	0	0
11	0.15	0.20	0	0	0	0.10
12	0.10	0.10	0	0.10	0	0

(b)

	$r=2$		$r=4$		$r=8$	
n	FP	FN	FP	FN	FP	FN
10	0	0	0	0	0	0
11	0	0	0.10	0.10	0	0
12	0.10	0.15	0	0	0	0

Table 2. Average numbers of edges in error for two duplications $(2,3)$: (a) uniform trees, (b) birth-death trees.

(a)

	$r=2$		$r=4$		$r=8$	
n	FP	FN	FP	FN	FP	FN
10	0.20	0.20	0	0	0	0
11	0.10	0.10	0	0	0	0.10
12	0	0	0.10	0.10	0	0

(b)

	$r=2$		$r=4$		$r=8$	
n	FP	FN	FP	FN	FP	FN
10	0.05	0.15	0	0	0	0
11	0	0	0	0	0	0
12	0	0	0	0	0	0.20

Table 3. Average numbers of edges in error for three duplications $(2,2,3)$: (a) uniform trees, (b) birth-death trees.

(a)

	$r=2$		$r=4$		$r=8$	
n	FP	FN	FP	FN	FP	FN
10	0.10	0.10	0	0	0	0
11	0	0	0	0	0	0.05
12	0.05	0.20	0	0	0	0.10

(b)

	$r=2$		$r=4$		$r=8$	
n	FP	FN	FP	FN	FP	FN
10	0.10	0.10	0	0	0	0
11	0	0	0	0.10	0	0
12	0	0.05	0	0	0.10	0.10

genomes (cases in which GRAPPA must examine half a billion trees), the running time never exceeded five minutes.

Tables 1 through 3 show the average numbers of false positive (FP) and false negative (FN) edges in the reconstructed trees when compared to the model trees generated by the simulations. A false negative arises when the reconstructed tree does not include an edge present in the model tree; conversely, a false positive arises when the reconstructed tree includes an edge not present in the model tree. The reconstructed trees are not always binary (fully resolved), because some of their edges may have length zero; edges of zero length are removed and thus may give rise to false negatives—unless the true tree itself had edges of zero length, something that does occur at lower evolutionary rates. Observe that both FN and FP remain extremely low, often even zero, indicating high accuracy in the reconstruction, a consequence, we conjecture, of matching all duplicates in the reconstruction process.

Tables 4 and 5 show the pruning rates for the median computations and for the tree enumeration. The latter shows very high pruning rates: in most cases fewer than one tree in $100,000$ remains to be scored. Pruning rates for medians are also excellent: we only rarely have to compute more than a couple of medians per node.

Table 4. Pruning rates (percentage of eliminated problems) for uniform trees.

		$r=2$		$r=4$		$r=8$	
	n	*Medians*	*Overall*	*Medians*	*Overall*	*Medians*	*Overall*
(a) one duplication	10	85.4	100	85.0	100	85.5	99.999
	11	85.5	100	85.1	100	85.6	100
	12	82.1	100	83.6	100	82.9	100
	n	*Medians*	*Overall*	*Medians*	*Overall*	*Medians*	*Overall*
(b) two duplications $(2,3)$	10	99.4	99.999	99.4	100	99.5	99.999
	11	99.0	99.999	99.1	100	98.7	100
	12	98.8	100	98.8	100	99.2	100
	n	*Medians*	*Overall*	*Medians*	*Overall*	*Medians*	*Overall*
(v) three duplications $(2,2,3)$	10	99.8	99.999	99.7	100	99.8	100
	11	99.8	99.999	99.8	100	99.6	100
	12	99.8	100	99.8	100	99.5	100

Table 5. Pruning rates (percentage of eliminated problems) for birth-death trees.

		$r=2$		$r=4$		$r=8$	
	n	*Medians*	*Overall*	*Medians*	*Overall*	*Medians*	*Overall*
(a) one duplication	10	85.1	99.999	85.6	99.999	79.8	99.999
	11	85.5	99.999	80.2	99.999	74.5	100
	12	85.6	100	85.5	99.999	84.0	100
	n	*Medians*	*Overall*	*Medians*	*Overall*	*Medians*	*Overall*
(b) two duplications $(2,3)$	10	98.9	99.995	99.0	99.998	99.3	99.999
	11	99.3	100	98.9	100	99.0	99.999
	12	99.3	100	99.2	100	98.7	100
	n	*Medians*	*Overall*	*Medians*	*Overall*	*Medians*	*Overall*
(c) three duplications $(2,2,3)$	10	99.6	100	99.8	100	99.8	99.999
	11	99.8	100	99.6	100	99.5	100
	12	99.8	100	99.8	100	99.7	100

5 Our Approach to Deletions

For simplicity, we now consider genomes without duplications to present our approach to deletions. (The two strategies can easily be combined to handle both duplications and deletions, albeit at the usual multiplicative cost of cases.) We need to devise strategies for computing pairwise distances between two genomes, for computing the median of three genomes, and for initializing ancestral labels at internal nodes. In developing these strategies, we will ignore “silent” changes (such as a gene loss followed by an insertion that restores the same gene). We will also assume that the probability of a gene loss is small enough that, when faced with the choice of assigning the loss to a parent or assigning it to both children, we always choose to assign it to the parent, since the probability of that one loss is some small p, but the probability of its being lost within the same time frame by both children is an infinitesimally small p^2.

Assume G_1 has N genes and G_2 has $N-m$ genes, i.e., G_2 lost m genes. There are $N \times N-1 \times \cdots \times N-m$, or roughly N^m different ways to equalize the two gene contents; using the same approach as for duplications, we define the distance between G_1 and G_2 to be the smallest pairwise inversion distance between G_1 and the various "completions" of G_2. In fact, we could directly use the method of El-Mabrouk [11] to solve this problem exactly in polynomial time, but we use the brute-force paradigm because it extends easily to the computation of medians, something that El-Mabrouk's approach does not. Consider now the computation of the median. For simplicity, assume we have $m = 1$—our reasoning easily extends to arbitrary values of m. Given three genomes G_1, G_2, and G_3, each of which could have lost a given gene, we face three cases:

- All three genomes lost that gene or none did. Then the median is in the same situation and the computation proceeds as currently implemented in GRAPPA.
- One genome, say G_1, lost that gene, but the other two still have it. Then the median retained the gene and thus a single loss event took place between the median and G_1. This gives us N choices of completion for G_1—we compute the median for each choice (if needed, since we can prune some choices through the same bounding strategy used for duplications).
- Two genomes lost that gene—say that only G_1 retains it. Then the median also lost that gene, so that we again have a single loss event, between the median and G_1. We remove that gene from G_1 and compute the median in the usual manner.

Thus we can compute the median in the case of a one-gene loss with at most N regular median computations; in the case of an m-gene loss, the number of regular median computations is on the order of N^m.

Before we can apply the median computations, however, we need to initialize the internal nodes of a tree with ancestral genomes. The first step in this initialization is to determine the gene content at each node. We accomplish this task using the same principle of always preferring a single loss event to a pair of concurrent loss events. Specifically, we run the following iterative algorithm:

- Identify all sibling pairs of leaves.
- For each sibling pair of leaves, assign to their parent the larger of the two gene contents—corresponding to a single loss event from the parent to the smaller child.
- Remove all processed leaves (thus turning their parents into leaves) and repeat.

6 Experimental Results for Deletions

Our test simulations are structured in the same manner as those we used for duplications. We tested for minimal gene loss (only two of the genomes lost a gene) and widespread gene loss (half of the genomes lost that gene). Once again, we ran 20 datasets for each combination of parameters and we report the average of the runs. Tables 6 and 7 show the average numbers of false positive and false negative edges in our reconstructions. As in the case of duplications, the reconstructions are remarkably accurate. Tables 8 and 9 give the corresponding pruning rates. The rates remain extremely high for tree pruning, but the simple triangle inequality proves fairly weak for pruning median computations.

Table 6. Average numbers of edges in error for one missing gene in two genomes: (a) uniform trees, (b) birth-death trees.

(a)

	$r=2$		$r=4$		$r=8$	
n	FP	FN	FP	FN	FP	FN
10	0.10	0.10	0	0	0	0
11	0.10	0.15	0	0	0.05	0.10
12	0.20	0.20	0	0	0	0

(b)

	$r=2$		$r=4$		$r=8$	
n	FP	FN	FP	FN	FP	FN
10	0	0.10	0	0	0	0.05
11	0	0	0	0.10	0	0
12	0	0.05	0	0	0	0

Table 7. Average numbers of edges in error for one missing gene in half the genomes: (a) uniform trees, (b) birth-death trees.

(a)

	$r=2$		$r=4$		$r=8$	
n	FP	FN	FP	FN	FP	FN
10	0	0	0	0	0	0
11	0.10	0.10	0	0	0.10	0.10
12	0	0	0	0.10	0	0

(b)

	$r=2$		$r=4$		$r=8$	
n	FP	FN	FP	FN	FP	FN
10	0.10	0.10	0	0.05	0	0.10
11	0	0	0	0	0	0
12	0.10	0.15	0	0	0.10	0.10

Table 8. Pruning rates (percentage of eliminated problems) for uniform trees.

gene lost in two genomes

	$r=2$		$r=4$		$r=8$	
n	*Medians*	*Overall*	*Medians*	*Overall*	*Medians*	*Overall*
10	71.7	99.999	62.5	100	64.1	99.999
11	72.2	99.999	57.4	100	66.8	100
12	68.6	100	65.2	100	60.5	100

gene lost in half the genomes

n	*Medians*	*Overall*	*Medians*	*Overall*	*Medians*	*Overall*
10	57.6	99.999	73.4	99.999	63.3	100
11	77.9	100	52.9	100	55.8	100
12	78.5	100	62.1	100	65.6	100

Table 9. Pruning rates (percentage of eliminated problems) for birth-death trees.

gene lost in two genomes

	$r=2$		$r=4$		$r=8$	
n	*Medians*	*Overall*	*Medians*	*Overall*	*Medians*	*Overall*
10	70.8	99.999	53.6	100	53.8	99.999
11	58.2	100	64.8	100	62.6	100
12	72.5	100	54.5	100	55.4	100

gene lost in half the genomes

n	*Medians*	*Overall*	*Medians*	*Overall*	*Medians*	*Overall*
10	68.9	100	65.7	100	51.2	99.999
11	68.8	99.999	51.2	100	74.6	100
12	78.5	99.999	62.6	100	67.8	100

7 Conclusions

We have presented a simple approach to the handling of a limited number of gene duplications and gene losses in the reconstruction of phylogenies from gene-order data. While the exhaustive nature of our approach limits its applicability to a fairly modest number of duplication and deletion events, it does allow us to analyze many datasets of organellar genomes, particularly chloroplast and mitochondria genomes, which are of special interest to evolutionary biologists. The success of our approach on such datasets also opens up the possibility that it could be scaled up through some type of divide-and-conquer paradigm, much in the manner in which we successfully scaled up GRAPPA, usually limited to 13–14 genomes, to one thousand genomes through the application of the sophisticated divide-and-conquer approach known as disk-covering.

8 Acknowledgments

This research is supported by the National Science Foundation under grants ACI 00-81404, DEB 01-20709, EIA 01-13095, and EIA 01-21377.

References

1. D.A. Bader, B.M.E. Moret, T. Warnow, S.K. Wyman, and M. Yan. *GRAPPA (Genome Rearrangements Analysis under Parsimony and other Phylogenetic Algorithms)*. `www.cs.unm.edu/~moret/GRAPPA/`.
2. D.A. Bader, B.M.E. Moret, and M. Yan. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J. Comput. Biol.*, 8(5):483–491, 2001. A preliminary version appeared in WADS'01, pp. 365–376.
3. G. Bourque and P. Pevzner. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Research*, 12:26–36, 2002.
4. D. Bryant. The complexity of calculating exemplar distances. 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 207–212. Kluwer Academic Pubs., Dordrecht, Netherlands, 2000.
5. D. Bryant. A lower bound for the breakpoint phylogeny problem. In *Proc. 11th Ann. Symp. Combin. Pattern Matching (CPM'00)*, volume 1848 of *Lecture Notes in Computer Science*, pages 235–247. Springer-Verlag, 2000.
6. A. Caprara. Sorting by reversals is difficult. In *Proc. 1st Int'l Conf. on Comput. Mol. Biol. RECOMB97*, pages 75–83. ACM Press, 1997.
7. A. Caprara. Formulations and hardness of multiple sorting by reversals. In *Proc. 3rd Int'l Conf. on Comput. Mol. Biol. RECOMB99*, pages 84–93. ACM Press, 1999.
8. A. Caprara. On the practical solution of the reversal median problem. In *Proc. 1st Workshop on Algs. in Bioinformatics WABI 2001*, volume 2149 of *Lecture Notes in Computer Science*, pages 238–251. Springer-Verlag, 2001.
9. M.E. Cosner, R. K. Jansen, B.M.E. Moret, L. A. Raubeson, L.-S. Wang, T. Warnow, and S. K. 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.
10. S.R. Downie and J.D. Palmer. Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In P. Soltis, D. Soltis, and J.J. Doyle, editors, *Plant Molecular Systematics*, pages 14–35. Chapman and Hall, 1992.
11. N. El-Mabrouk. Genome rearrangement by reversals and insertions/deletions of contiguous segments. In *Proc. 11th Ann. Symp. Combin. Pattern Matching (CPM'00)*, volume 1848 of *Lecture Notes in Computer Science*, pages 222–234. Springer-Verlag, 2000.

12. S. Hannenhalli and P.A. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *Proc. 27th Ann. Symp. Theory of Computing STOC 95*, pages 178–189. ACM Press, 1995.
13. D. Huson, S. Nettles, K. Rice, T. Warnow, and S. Yooseph. The hybrid tree reconstruction method. *ACM J. Experimental Algorithmics*, 4(5), 1999. http://www.jea.acm.org/1999/HusonHybrid/.
14. D. Huson, S. Nettles, and T. Warnow. Disk-covering, a fast converging method for phylogenetic tree reconstruction. *J. Comput. Biol.*, 6(3):369–386, 1999.
15. D. Huson, L. Vawter, and T. Warnow. Solving large-scale phylogenetic problems using DCM-2. In *Proc. 7th Int'l Conf. on Intelligent Systems for Molecular Biology (ISMB99)*, pages 118–129. AAAI Press, 1999.
16. B. Larget, J.B. Kadane, and D. Simon. A Markov chain Monte Carlo approach to reconstructing ancestral genome rearrangements. Technical report, Carnegie Mellon University, Pittsburgh, PA, 2002. Available at `www.stat.cmu.edu/tr/tr765/`.
17. M. Marron, K.M. Swenson, and B.M.E. Moret. Genomic distances under deletions and insertions. In *Proc. 9th Ann. Int'l Conf. Computing and Combinatorics (COCOON'03)*, Lecture Notes in Computer Science. Springer-Verlag, 2003. Accepted, to appear.
18. B.M.E. Moret, A.C. Siepel, J. Tang, and T. Liu. Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In R. Guigo and D. Gusfield, editors, *Proc. 2nd Int'l Workshop Algorithms in Bioinformatics (WABI'02)*, volume 2452 of *Lecture Notes in Computer Science*, pages 521–536. Springer-Verlag, 2002.
19. B.M.E. Moret, J. Tang, L.-S. Wang, and T. Warnow. Steps toward accurate reconstructions of phylogenies from gene-order data. *J. Comput. Syst. Sci.*, 65(3):508–525, 2002.
20. R.G. Olmstead and J.D. Palmer. Chloroplast DNA systematics: a review of methods and data analysis. *Amer. J. Bot.*, 81:1205–1224, 1994.
21. J.D. Palmer. Chloroplast and mitochondrial genome evolution in land plants. In R. Herrmann, editor, *Cell Organelles*, pages 99–133. Springer Verlag, 1992.
22. I. Pe'er and R. Shamir. The median problems for breakpoints are NP-complete. *Elec. Colloq. on Comput. Complexity*, 71, 1998.
23. L.A. Raubeson and R.K. Jansen. Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. *Science*, 255:1697–1699, 1992.
24. N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.
25. D. Sankoff. Genome rearrangement with gene families. *Bioinformatics*, 15(11):909–917, 1999.
26. D. Sankoff and M. Blanchette. Multiple genome rearrangement and breakpoint phylogeny. *J. Comput. Biol.*, 5:555–570, 1998.
27. D. Sankoff and J. Nadeau, editors. *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment, and the Evolution of Gene Families*. Kluwer Academic Pubs., Dordrecht, Netherlands, 2000.
28. A.C. Siepel and B.M.E. Moret. Finding an optimal inversion median: Experimental results. In O. Gascuel and B.M.E. Moret, editors, *Proc. 1st Int'l Workshop Algorithms in Bioinformatics (WABI'01)*, volume 2149 of *Lecture Notes in Computer Science*, pages 189–203. Springer-Verlag, 2001.
29. J. Tang and B.M.E. 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, Suppl. 1 of *Bioinformatics*, pages i305–i312. Oxford U. Press, 2003.