Inversion Medians Outperform Breakpoint Medians in Phylogeny Reconstruction from Gene-Order Data

Bernard M.E. Moret[1], Adam C. Siepel[2], Jijun Tang[1], and Tao Liu[1]

[1] Department of Computer Science
University of New Mexico
Albuquerque, NM 87131, USA
{moret,jtang,tigerliu}@cs.unm.edu
[2] Department of Computer Science and Engineering
University of California at Santa Cruz
Santa Cruz, CA 95064
acs@cse.ucsc.edu

Abstract. Phylogeny reconstruction from gene-order data has attracted much attention over the last few years. The two software packages used for that purpose, BPAnalysis and GRAPPA, both use so-called breakpoint medians in their computations. Some of our past results indicate that using inversion scores rather than breakpoint scores in evaluating trees leads to the selection of better trees. On that basis, we conjectured that phylogeny reconstructions could be improved by using inversion medians, which minimize evolutionary distance under an inversions-only model of genome rearrangement. Recent algorithmic developments have made it possible to compute inversion medians for problems of realistic size.
Our experimental studies unequivocally show that inversion medians are strongly preferable to breakpoint medians in the context of phylogenetic reconstruction from gene-order data. Improvements are most pronounced in the reconstruction of ancestral genomes, but are also evident in the topological accuracy of the reconstruction as well as, surprisingly, in the overall running time. Improvements are strongest for small average distances along tree edges and for evolutionary scenarios with a preponderance of inversion events, but occur in all cases, including evolutionary scenarios with high proportions of transpositions.
All of our tests were run using our GRAPPA package, available (under GPL) at www.cs.unm.edu/∼moret/GRAPPA/; the next release will include the inversion median software we used in this study. The software used includes RevMed, developed by the authors and available at www.cs.unm.edu/∼acs/, and A. Caprara's inversion median code, generously made available for testing.

Keywords: breakpoint, genome rearrangement, genomic distance, inversion, reversal

1 Introduction

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 and transpositions and, because they evolve slowly, give biologists an important new source of data for phylogeny reconstruction (see, e.g., [7, 16, 17, 19]). 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 was devoted to this topic, as was a workshop at DIMACS.

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 under such models is NP-hard [4, 18]. All approaches to phylogeny reconstruction for such data must therefore find ways of handling significant computational difficulties. Moreover, because suboptimal solutions can yield very different evolutionary reconstructions, exact solutions are strongly preferred over approximate solutions in all phylogenetic work (see [25]).

For some datasets (e.g., chloroplast genomes of land plants), biologists conjecture that rearrangement events are predominantly *inversions*. In other datasets, transpositions and inverted transpositions are viewed as possible, but their relative preponderance with respect to inversions is unknown, so that it is difficult to define a suitable distance measure based on these three events. Sankoff proposed using the *breakpoint* distance (the number of pairwise gene adjacencies present in one genome but absent in the other), a measure of distance between genomes that is independent of any particular mechanism of rearrangement, to reconstruct phylogenies; the *breakpoint phylogeny*, introduced by Blanchette *et al.* [2], is the most parsimonious tree with respect to breakpoint distances.

The two software packages for reconstructing the breakpoint phylogeny, the original `BPAnalysis` of Sankoff and Blanchette [21] and the more recent and much faster `GRAPPA` [14], both use as their basic optimization tool an algorithm for computing the breakpoint median of three genomes. Recent work, however, based on the elegant theory of Hannenhalli and Pevzner [9], has shown that inversion distance can be computed in linear time [1], a development that has allowed us to base the reconstruction process on the inversion score of a tree rather than on its breakpoint score, with significant resulting improvements in the accuracy of reconstructions [13]. Other recent results have shown that the inversion median of three genomes can be obtained quickly for a reasonable range of instances [5, 23, 24] (in spite of the NP-hardness of the problem). These developments have enabled us to extend `GRAPPA` by replacing the breakpoint median routine with an inversion median routine—which we conjectured would yield better phylogenetic reconstructions (at least when inversions are the dominant mechanism of rearrangement), because inversion medians score significantly better in terms of inversion distance, and are much closer to being unique, than breakpoint medians [24]. Note that it would be even more desirable to use medians based on a general measure of distance that considers transpositions (and inverted transpositions) as well as inversions. Currently, however, efficient algorithms are not available either to find transposition distance or to compute medians that consider transpositions.

In this paper, we present the results of a series of experiments designed to compare the quality of reconstructions obtained by `GRAPPA` running with breakpoint medians

[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.

and with inversion medians. To our knowledge, no such comparison has previously been conducted, despite much speculation about the value of inversion medians in phylogeny reconstruction. In brief, we found that inversion medians are strongly preferable to breakpoint medians in all but the most extreme cases, in terms of quality of tree topologies and of ancestral genome reconstruction, and also (surprisingly) preferable in many cases in terms of speed of execution.

The rest of the paper is organized as follows. We begin by reviewing the pertinent prior work. We then introduce the required terminology, present our experimental design, and briefly highlight the main attributes of the median-finding routines described in [5, 23, 24]. Next we present a cross-section of our results and discuss their implications. Finally, we conclude with suggestions for further work.

2 Prior Results

BPAnalysis. Blanchette *et al.* [2] proposed the breakpoint phylogeny (finding the tree with the fewest breakpoints) and developed a reconstruction method, `BPAnalysis` [21], for that purpose. Their method examines every possible tree topology in turn and for each topology, it generates a set of ancestral genomes so as to minimize the total breakpoint distance in the tree. This method returns good results, but takes exponential time: the number of topologies is exponential and generating a set of ancestral genomes is achieved through an unbounded iterative process that must solve an instance of the Travelling Salesperson Problem (TSP) for each internal node at each iteration, so that the total running time is exponential in both the number of genes and the number of genomes.

GRAPPA. We reimplemented `BPAnalysis` in order to analyze our larger datasets and also to experiment with alternative approaches. Our program, called `GRAPPA` [14], includes all of the features of `BPAnalysis`, but is more flexible and runs up to six orders of magnitude faster [12]. It also allows the user to base the optimization on a tree's inversion score or on its breakpoint score (although medians are still computed with the TSP methods of Sankoff and Blanchette, which is optimal only for breakpoint medians).

Inversion distance and inversion medians. Inversion distance can be computed in linear time [1]; an efficient implementation of this algorithm is provided in `GRAPPA`. As mentioned earlier, computing the inversion (or breakpoint) median of three signed permutations is an NP-hard problem. Breakpoint medians can be computed very efficiently for reasonable problem sizes [14]. Two algorithms have recently been proposed for computing inversion medians (also called "reversal medians") and shown to run efficiently (although much more slowly than the algorithm for breakpoint medians) for a range of parameters that covers most organellar genomes; one by Caprara [5] and one by Siepel and Moret [24] (subsequently refined in [22, 23]). Both algorithms use branch-and-bound strategies, but the algorithm of Siepel and Moret directly searches the space of genome rearrangements for an optimal solution, using the metric property of inversion distance for bounding, while the algorithm of Caprara uses edge contraction on a "multibreakpoint graph" (a version of the breakpoint graph generalized to accommodate more than two signed permutations). The edge-contraction operation modifies a

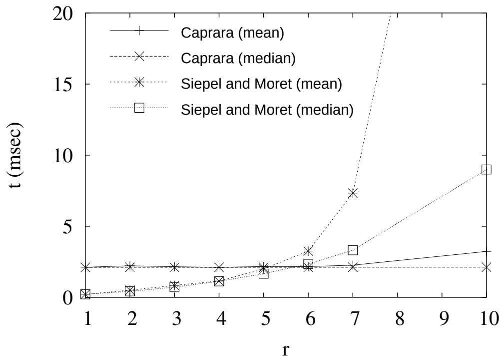


Fig. 1. Mean and median running times (in *ms*) of the two algorithms for three genomes of 50 genes as a function of the evolutionary rate r (expected number of events per edge) under an inversion-only scenario. Values shown are based on 100 experiments. When r is large, mean times significantly exceed median times, because the algorithms occasionally take much longer than usual to find a median.

multibreakpoint graph by removing an edge and its adjoining nodes, and making appropriate adjustments to *matchings* associated with the graph. Caprara's branch-and-bound algorithm uses the property that the best solution to an instance of the problem can be expressed in terms of the best solutions to subproblems in which edges have been contracted. (The algorithms of Caprara and of Siepel and Moret are both moderately complicated, and fuller descriptions are beyond the scope of this paper; we refer readers to [5] and [22] for details). The algorithm of Caprara generally runs faster than the other, although the "sorting median" refinement of Siepel is faster when edge lengths are small between permutations, as seen in Figure 1. Both algorithms are sensitive to the distances separating input permutations (although Caprara's much less so); hence their use in phylogeny reconstruction depends on relatively short edge lengths between nodes. Timings in this study use Caprara's algorithm; note, however, that a dynamic switch to the algorithm of Siepel and Moret for small distances (which are common within most phylogenies) would considerably improve execution time.

3 Background and Terminology

3.1 Evolutionary events

When each genome has the same set of genes and each gene appears exactly once, a genome can be described by an ordering (circular or linear) of these genes, each gene given with an orientation that is either positive (g_i) or negative ($-g_i$). Let G be the genome with signed ordering $g_1, g_2, \ldots, g_n$. An *inversion* between indices i and j, $i \leq j$, produces the genome with linear ordering

$$g_1, g_2, \ldots, g_{i-1}, -g_j, -g_{j-1}, \ldots, -g_i, g_{j+1}, \ldots, g_n$$

A *transposition* on the ordering G acts on three indices, i, j, k, with $i \leq j$ and $k \notin [i, j]$, picking up the interval $g_i, g_{i+1}, \ldots, g_j$ and inserting it immediately after g_k. Thus the genome G above (for $k > j$) is replaced by

$$g_1, \ldots, g_{i-1}, g_{j+1}, \ldots, g_k, g_i, g_{i+1}, \ldots, g_j, g_{k+1}, \ldots, g_n$$

An *inverted transposition* is a transposition followed by an inversion of the transposed piece. The *distance* between two gene orders is the minimum number of inversions, transpositions, and inverted transpositions needed to transform one gene order into the other. When only one type of event occurs in the model, we speak of *inversion distance* or *transposition distance*.

Given two genomes G and G' on the same set of genes, a *breakpoint* in G is defined as an ordered pair of genes (g_i, g_j) such that g_i and g_j appear consecutively in that order in G, but neither (g_i, g_j) nor $(-g_j, -g_i)$ appear consecutively in that order in G'. The number of breakpoints in G relative to G' is the *breakpoint distance* between G and G'.

The *Nadeau-Taylor* model [15] of genome evolution, as generalized by Wang and Warnow [27], uses only genome rearrangement events, so that all genomes have equal gene content. The model assumes that each of the three types of events obeys a Poisson distribution on each edge—with the three means for the three types of events in some fixed ratio.

3.2 Model trees: simulating evolution

A model tree is a rooted binary tree in which each edge e has an associated non-negative real number, λ_e, denoting the expected number of events on e. The model tree also has a weight parameter, a triple of values which defines the probability that a rearrangement event is an inversion, transposition, or inverted transposition. We denote this triple by $(1 - a - b, a, b)$.

Given a model tree with N leaves, a set of N "contemporary" gene orderings can be generated as follows. First, the root is labeled with the identity gene ordering $g_1, g_2, \ldots, g_n$; then the tree is traversed recursively, and each node is assigned a label that is derived by applying random rearrangements to the label of its parent. Suppose a node u is separated from its parent p by an edge having expected number of events λ_e, and suppose the parent is labeled with gene ordering π_p. The label π_u for u is determined by drawing numbers of inversions, transpositions, and inverted transpositions from the appropriate Poisson distributions (having expected values $\lambda_e(1 - a - b)$, $\lambda_e a$, and $\lambda_e b$, respectively) and applying those events to π_p in random order, at randomly-selected indices. If an effective reconstruction algorithm is applied to the labels at the leaves of such a tree, it should infer an evolutionary history resembling the model tree.

3.3 Labelling internal nodes

One of the major advantages of median-based phylogeny reconstruction over alternative methods (such as the encoding methods of [13, 26]) is that it estimates the configurations of ancestral genomes as well as the topology of the tree. The approach proposed by Sankoff and Blanchette to estimate ancestral genomes is iterative, using a local optimization strategy. It is applied in turn to each plausible tree topology; in the end, the topologies are selected that require the fewest total breakpoints to explain. After initial labels have been assigned in some way to internal (ancestral) nodes in a given topology, the procedure repeatedly traverses the tree, computing for each node the breakpoint median of its three neighbors and using the median as the new label if this change improves the overall breakpoint score—the entire process is also known as "Steinerization." The median-of-three subproblems are transformed into instances of the Travelling Salesperson Problem (TSP) and solved optimally. The overall procedure is a heuristic without any approximation guarantee, but does well in practice on datasets with a small number of genomes.

`GRAPPA` uses the same overall iterative strategy and also solves the median-of-three problem in its TSP formulation to obtain potential labels for internal nodes. `GRAPPA`, however, has the option of accepting a relabelling of an internal node based on either the breakpoint score (as in `BPAnalysis`) or the inversion score of the tree.

3.4 Performance criteria

Let T be a tree leaf-labelled by the set S. Deleting some edge e from T produces a bipartition π_e of S into two sets. Let T be the true tree and let T' be an estimate of T. Then the *false negatives* of T' with respect to T are those non-trivial bipartitions[2] that appear in T, but not in T'; conversely, the *false positives* of T' with respect to T are those non-trivial bipartitions that appear in T', but not in T. The numbers of false positives and false negatives are normalized by dividing by the number of non-trivial bipartitions of T, to obtain the *fraction of false positives* and the *fraction of false negatives*, respectively. If the fraction of false negatives is 0, then T' equals T or refines it; if the fraction of false positives is also 0, then T' equals T (assuming differences in zero-length edges are not considered important).

4 Our Experimental Design

We designed a set of experiments to assess the impact of replacing breakpoint medians with inversion medians on three critical aspects of phylogenetic reconstruction: speed of execution, topological accuracy of reconstructed trees, and accuracy of estimated ancestral labels. As mentioned, we sought to test the conjecture that exact inversion medians—which score better than breakpoint medians in terms of inversion distance and are more unique [24]—would lead to more accurate reconstructions (both in terms of topology and ancestral labels); at the same time, we sought to characterize the running-time penalty for using the considerably slower inversion median computations. We mostly used simulated datasets (where we know the true tree and the true ancestral labels and thus can directly assess accuracy), but also used two datasets of chloroplast genomes from a number of land plants and algae (where we can compare tree scores and running times). A simulated dataset is determined by four parameters: (i)

[2] An edge of zero length is said to produce a *trivial* bipartition. If an edge e' in T' has zero length but a corresponding edge e in T (one producing the same bipartition) does not have zero length, e will count as a false negative; but if e' and e both have zero length, no penalty will occur. The converse holds for false positives.

the number of genes n; (ii) the number of genomes N; (iii) the triple $(1-a-b, a, b)$ describing the proportions of inversions, transpositions, and inverted transpositions used in the simulated evolution; and (iv) the amount of evolution taking place, represented by r, the expected number of evolutionary events occurring along a tree edge. A simulated dataset is obtained by selecting a tree topology uniformly at random, labeling its edges uniformly with an expected number r of rearrangement events (that is, for simplicity, we use the same λ_e value everywhere), and simulating evolution down the tree, as explained earlier. Because we know the true tree and the true ancestral labels, we can compare the reconstructions to the true tree in terms of topology as well as in terms of ancestral labels. In order to assess variability, we repeated all simulation experiments for at least 10 datasets generated using the same parameters.

Because computing the exact inversion median of three genomes is much more expensive than computing their exact breakpoint median (while both tasks are NP-hard, the breakpoint version has a much simpler structure), replacing breakpoint medians by inversion medians was expected to cause considerable slowdown. However, if inversion medians lead quickly to better tree scores, some performance gain can accrue from more effective pruning of tree topologies. The net effect on running time will thus be the result of a trade-off between more expensive median computations and better pruning. We evaluated this net effect by running the fastest version of GRAPPA, with the improved bounding and layered search described in [12]; we used both breakpoint medians (evaluated with inversion scores, as previously supported) and true inversion medians (supported by new extensions). We compared both total running times and pruning rates, the latter being of interest as an implementation-independent measure.

To assess improvements in accuracy, we ran experiments with fixed evolutionary rate and increasing numbers of genomes. We repeated experiments for several different numbers of genes—a parameter that affects computational cost more than it affects accuracy, but that does have an impact on accuracy through discretization (with a small number of genes, the range of possible values is limited and any error is therefore magnified). To assess topological accuracy, we ran GRAPPA with breakpoint medians and with inversion medians and compared the best reconstructed topologies with the true tree. While the trees returned by GRAPPA are ostensibly binary, they often contain edges of zero length and thus are not fully resolved. (Such lack of resolution is much more pronounced with breakpoint medians than with inversion medians.) We thus measured both false positive and false negative rates (a reconstruction can avoid introducing false positives by leaving relationships unresolved, but only at the cost of false negatives).

To assess the accuracy of reconstructed ancestral genomes, we ran GRAPPA on the leaf genomes as before, but this time with a single topology—that of the true tree. In effect, we used GRAPPA to score a single tree by computing ancestral genomes for the internal nodes of the tree. The reconstructed tree is thus always isomorphic to the true tree, and we can compare internal nodes directly, without having to adjust for inconsistent topologies or lack of branch resolution. We assessed overall accuracy by summing, over all internal nodes of the tree, the pairwise distances (in both inversion and breakpoint versions, with no significant changes in outcomes) between the true and the reconstructed ancestral genomes and normalizing this sum by the number of internal nodes in the tree (for independence from tree size). This simple measure describes how far are the ancestral nodes of the reconstructed tree (in the aggregate) from being precisely correct. (Notice that this measure has the potential quickly to become saturated when reconstructions are poor, because errors accumulate as one moves from leaves to root; we found, however, that reconstructed trees were good enough to avoid this problem).

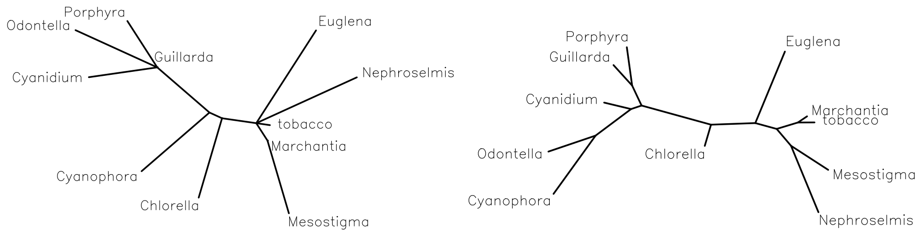
To assess robustness of results, we ran experiments with many different evolutionary scenarios, including inversions only $(1, 0, 0)$, equally weighted inversions, transpositions, and inverted transpositions $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, a weighting of $(\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$ that is believed to be more realistic (recommended by Sankoff), and the completely mismatched transposition-only scenario, $(0, 1, 0)$. We can expect inversion medians to be beneficial in inversion-only evolutionary scenarios, but need to evaluate their usefulness in more adverse scenarios, including the completely mismatched scenarios where we have $a + b = 1$. (One argument for breakpoints is that they provide a model-independent measure: our goal then must be to test whether using a model-dependent measure such as inversion distance provides sufficient benefits when the model is well matched and avoids excessive problems when the model is poorly matched.)

5 Results and Discussion

5.1 Speed

We used two real datasets of chloroplast genomes that we had previously analyzed using a variety of methods. Our first set contains 12 *Campanulaceae* plus a Tobacco outgroup, for a total of 13 taxa, each with 105 genes. Evolutionary rates on this set appear low and most evolutionary changes are believed to have occurred through inversions [6]. Earlier analyses reduced the running times from an original estimate of several centuries with `BPAnalysis` to 9 hours on a fast workstation [12], with the best-scoring tree having an inversion score of 67. Running the same code with the inversion median routine reduced the running time to just one hour on the same machine, while yielding many trees with an inversion score of 64—but with topologies identical to the previous best trees. (Bourque and Pevzner [3] had found one tree with a score of 65 using a somewhat different approach, while Larget *et al.* [11] just reported finding a number of trees that, like ours, required 64 inversions—again these trees have the same topologies as our previous best trees.) While the inversion median routine itself was much slower than the breakpoint median routine, the pruning rate more than made up for it: of the 13.7 billion trees to be examined, all but $100,000$ were pruned away when using the inversion medians (a pruning rate of over 99.999%), while 8.7 million remained unpruned when using breakpoint medians (a pruning rate of 99.94%).

Our second set has 11 taxa (including land plants as well as red and green algae), each with 33 genes. This set was also analyzed with a variety of methods [20]; unlike the *Campanulaceae* set, it has high divergence, with intertaxa inversion distances averaging over 10. The analyses of [20] used a variety of techniques, including `GRAPPA` with breakpoint medians, which yielded, in about 14 hours of computation (with a pruning rate of 25%), trees with inversion scores of 95; using inversion medians brought the pruning rate up to 99.9%, the running time down to $\frac{1}{4}$ hour, and the inversion score down to a very low 82, with complete binary resolution, as illustrated in Figure 2. The



using breakpoint medians *using inversion medians*

Fig. 2. Phylogenies of 11 plants based on chloroplast gene orders.

new tree (on the right-hand side) has monophyletic greens, correctly groups together tobacco and Marchantia, and offers a resolution of red algae that, while not the accepted one (the placement of Cyanophora, in particular, is curious), is not entirely impossible; in contrast, the old tree has poor resolution and does not group tobacco and Marchantia. Of interest in this reconstruction is the fact that a method predicated on an inversion model did much better (both in biological terms and in terms of scoring) than one using the neutral model of breakpoints, which could be viewed as additional support for the conjecture that most of the rearrangements in the chloroplast genome have occurred through inversions.

We ran a large number of tests on simulated data as well. Table 1 shows the ratio of the running times with breakpoint medians to the running times with inversion medians. Note that the ratios are very close to 1, indicating that the cost of the inversion median computation is neatly compensated by the reduction in the number of passes necessary to score a tree and by the improved pruning rate. Whenever the new code runs faster, it is a consequence of significantly better pruning due to tighter bounds. (The bounds are computed with exactly the same formula, relying only on the triangle inequality, but the distance estimates based on inversion distances are evidently much better than those based on breakpoint medians.) Figure 3 shows the number of calls made to each of the median procedures. The use of inversion medians, by deriving better initial solutions and by tightening upper bounds quickly, considerably reduces the number of trees that must be examined and, for these trees, the number of passes required to score them.

Table 1. Ratios of the running time of `GRAPPA` with breakpoint medians to that with inversion medians for two genome sizes, under three different evolutionary scenarios.

For 10 taxa:

	$(1,0,0)$		$(\frac{2}{3},\frac{1}{6},\frac{1}{6})$		$(\frac{1}{3},\frac{1}{3},\frac{1}{3})$	
	$r=2$	$r=4$	$r=2$	$r=4$	$r=2$	$r=4$
$n=50$	1.01	0.90	1.07	1.12	1.36	1.25
$n=100$	1.00	1.20	1.43	1.08	2.90	1.93

For 11 taxa:

	$(1,0,0)$		$(\frac{2}{3},\frac{1}{6},\frac{1}{6})$		$(\frac{1}{3},\frac{1}{3},\frac{1}{3})$	
	$r=2$	$r=4$	$r=2$	$r=4$	$r=2$	$r=4$
$n=50$	1.02	0.93	0.99	0.99	1.22	1.13
$n=100$	1.01	1.00	1.01	1.02	1.72	1.43

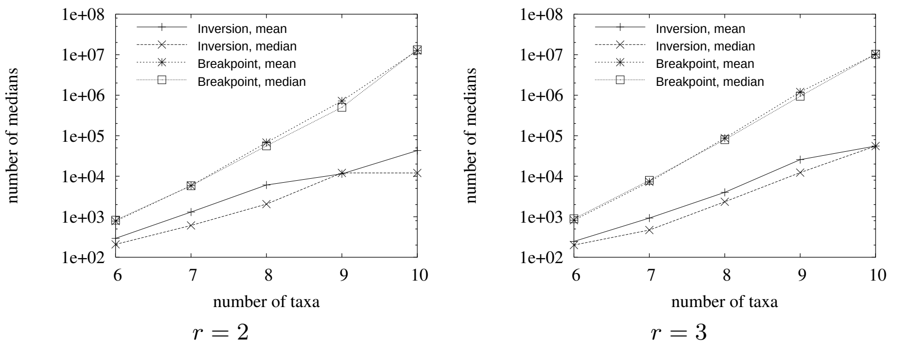


Fig. 3. The number of calls made to each median routine as a function of the number of taxa, for 30 genes and two values of r, plotted on a semilogarithmic scale.

5.2 Topological Accuracy

Figure 4 shows the fractions of false negative edges in the reconstructions for $r = 2$, $n = 30$, and three different proportions of rearrangement events. False positive fractions are negligible for both breakpoint and inversion medians—although even here, the values for inversion medians improve on those for breakpoint medians. These results are typical of what we observed with other parameter settings. The rate of false negatives is quite high for breakpoint medians, in good part because breakpoint medians are much more frequently "trivial" than inversion medians. That is, a valid breakpoint median for three gene orderings is frequently one of the orderings itself; a node in the tree and its parent are then assigned the same ordering and a zero-length edge results, which counts as a false negative as long as the corresponding edge in the true tree is nontrivial. Zero-length edges are undesirable for other reasons as well (see below), so that their avoidance constitutes an important advantage of inversion medians

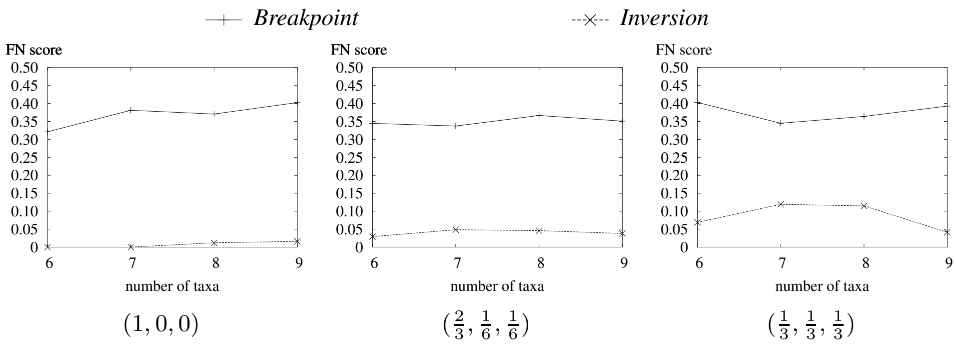


Fig. 4. False negative (FN) scores as a function of the number of taxa for 30 genes at evolutionary rate of $r = 2$.

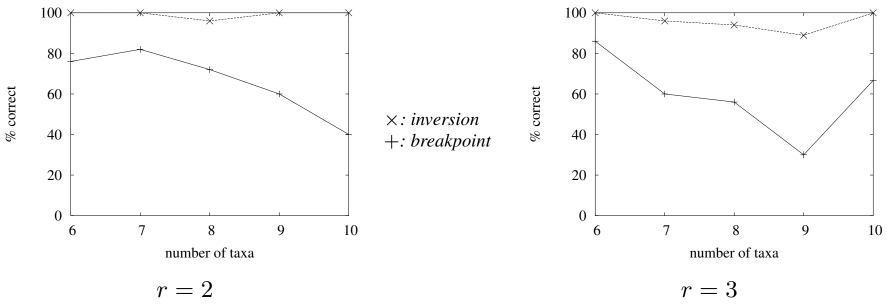


Fig. 5. Percentage of datasets, as a function of the number of taxa, for which `GRAPPA` returns at least one topology identical to the true tree; all datasets have 30 genes and were generated under an inversion-only scenario. Each point is based on 50 datasets.

over breakpoint medians. Remarkably, the percentage of false negatives in reconstructions using the inversion median hovers at or below 10, for all scenarios and all tree sizes tested—demonstrating extraordinary accuracy in reconstruction even when only one third of the events are in fact inversions. In contrast, the reconstructions based on breakpoint medians fare poorly under all scenarios, although their performance is, as expected, insensitive to the proportions of the three types of events.

As suggested by these small fractions of false positive and negative edges, `GRAPPA` was highly effective at accurately reconstructing model trees. Indeed, in most cases, the program found a tree that was identical in topology to the true tree (it generally reports multiple trees of equal score, each of which has a different topology). Figure 5 shows the percentage of cases in which at least one exact reconstruction was obtained, under various conditions. With less favorable evolutionary scenarios (having a substantial percentage of transpositions), we see a slow decrease in performance, but the gap between breakpoint medians and inversions medians remains pronounced. Surprisingly, even in unfavorable cases, the percentage of perfect topologies obtained using inversion medians remains high (Figure 6). In the course of performing these experiments, we found that many more trees of equal score tend to be returned when using breakpoint medians, presumably because breakpoint medians are highly non-unique [24] and because trivial medians are common. We also found in many cases that `GRAPPA` fails to find an exact reconstruction with the breakpoint median because of false negatives: that is, it finds trees consistent with the true tree, except that they have many unresolved nodes. The main advantages of the inversion median with respect to topology, then, seem to be that it produces more fully-resolved trees and fewer equally scoring trees.

Also of interest, in regard to topological accuracy, is the effect on overall parsimony scores of using the inversion median. While parsimony scores (the sum of the minimum inversion distances along the edges of the tree) do not have direct topological significance, the goal of the `GRAPPA` optimization is to obtain a most parsimonious tree, so that any reduction in the score is an indication that the algorithm is behaving more effectively. Table 2 shows percent reductions for a variety of settings; note that reductions of 3% or more generally translate into significant changes in tree topology.

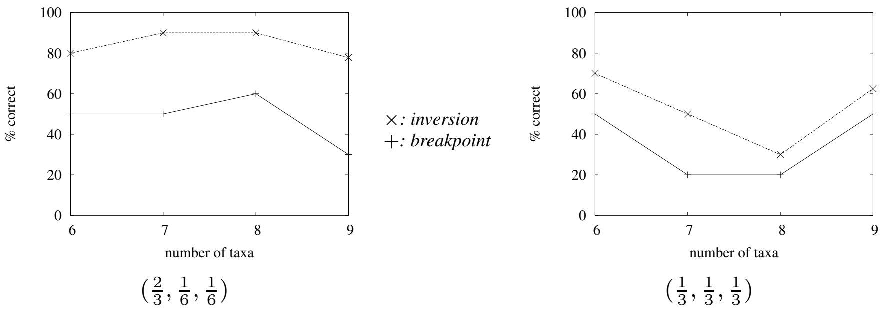


Fig. 6. Percentage of datasets, as a function of the number of taxa, for which GRAPPA returns at least one topology identical to the true tree; all datasets have 30 genes and were generated with an evolutionary rate of $r = 2$. Each point is based on 10 datasets.

Table 2. The reduction (in percent) in the parsimony score of trees when using inversion medians, for two genome sizes, under three different evolutionary scenarios.

For 10 taxa:

	$(1,0,0)$		$(\frac{2}{3},\frac{1}{6},\frac{1}{6})$		$(\frac{1}{3},\frac{1}{3},\frac{1}{3})$	
	$r=2$	$r=4$	$r=2$	$r=4$	$r=2$	$r=4$
$n = 50$	3.2%	5.8%	2.7%	5.1%	1.6%	6.2%
$n = 100$	2.7%	2.4%	1.5%	2.1%	1.9%	2.0%

For 11 taxa:

	$(1,0,0)$		$(\frac{2}{3},\frac{1}{6},\frac{1}{6})$		$(\frac{1}{3},\frac{1}{3},\frac{1}{3})$	
	$r=2$	$r=4$	$r=2$	$r=4$	$r=2$	$r=4$
$n = 50$	4.8%	9.9%	2.1%	3.9%	2.9%	5.4%
$n = 100$	0.0%	2.8%	0.5%	1.6%	3.5%	2.7%

5.3 Accuracy of Ancestral Genomes

Figure 7 shows the average inversion distance between corresponding ancestral genomes in true and reconstructed trees, as a function of the number of taxa for fixed evolutionary rates under different evolutionary scenarios. This measure was computed by fixing tree topologies and letting GRAPPA simply label ancestral nodes. Again, the high degree of accuracy of reconstructions based on inversion medians is striking and in sharp contrast with the poor reconstructions obtained with breakpoint medians. Even in the optimal case of inversion-only scenarios, the results exceeded our expectations, as almost all internal genomes are correctly reconstructed; also of note is that, in the case of transposition-only scenarios. the inversion-based reconstructions remain superior to the breakpoint-based ones, in spite of the complete mismatch of evolutionary models. We have shown elsewhere that inversion medians provide highly accurate estimates of actual intermediate genomes when $N = 3$ and evolution occurs by inversions only [24]. Our results here indicate that they remain accurate—and clearly superior to breakpoint medians—as N is increased (with multiple intermediate nodes determined by Sankoff and Blanchette's Steinerization method) and as other types of rearrangement events are introduced.

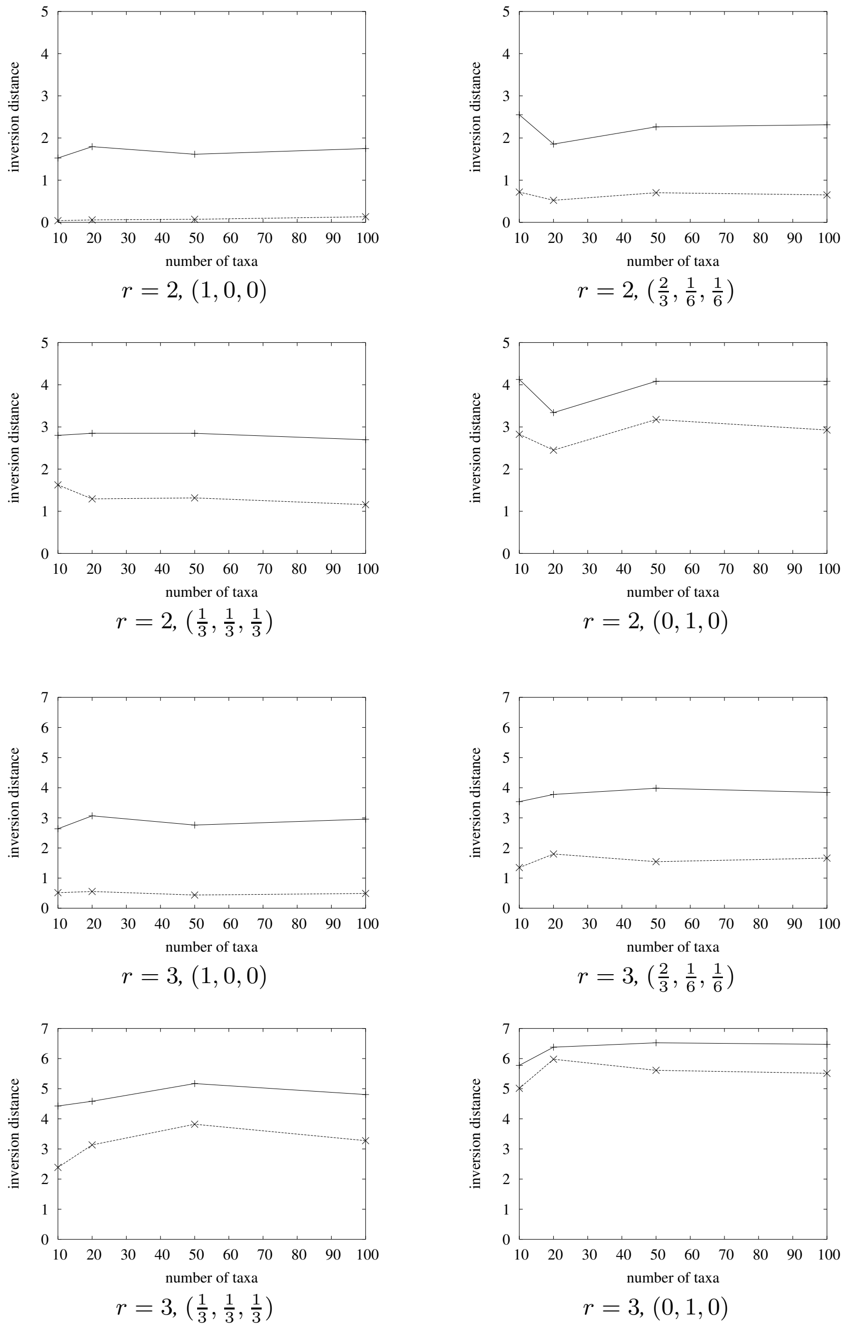


Fig. 7. Average inversion distance between corresponding ancestral genomes in the true tree and the reconstructed tree as a function of the number of taxa, under various evolutionary scenarios, for 30 genes. Upper curve corresponds to reconstructions based on breakpoint medians, lower curve to reconstructions based on inversion medians.

6 Conclusions and Future Work

Finding the inversion median of three genomes is a complex optimization task; finding the median of three genomes under a mix of evolutionary events (inversions, transpositions, and inverted transpositions) has not yet been addressed and seems considerably harder. Yet our experiments indicate that further research on these problems should prove very useful: the reconstructions we have obtained by using inversion medians are clearly better, in every respect, than those obtained with breakpoint medians. In particular, we obtained very accurate reconstructions of the ancestral genomes, a crucial step in the scoring of a particular tree topology as well as an important biological datum.

The results we have presented here, along with other results on handling transversions and gene duplication (see [8, 10]) justify a cautious optimism: over the next few years, the reconstruction of phylogenies from gene-order data should become applicable to much larger datasets, extending the approach from organellar genomes to some nuclear genomes. We also expect that some of the lessons learned in the process will yield distinct improvement to the reconstruction of phylogenies from simpler molecular data (such as RNA, DNA, or amino-acid sequences).

Acknowledgments

We thank Alberto Caprara from the Università di Bologna for letting us use his code for inversion medians. Bernard Moret's work is supported by the National Science Foundation under grants ACI 00-81404, EIA 01-13095, EIA 01-23177, and DEB 01-20709.

References

1. D. Bader, B. Moret, and M. Yan. A fast linear-time algorithm for inversion distance with an experimental comparison. *J. Comput. Biol.*, 8(5):483–491, 2001.
2. M. Blanchette, G. Bourque, and D. Sankoff. Breakpoint phylogenies. In S. Miyano and T. Takagi, editors, *Genome Informatics 1997*, pages 25–34. Univ. Academy Press, 1997.
3. G. Bourque and P. Pevzner. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Research*, 12:26–36, 2002.
4. 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.
5. 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.
6. M. Cosner, R. Jansen, B. Moret, L. Raubeson, L. Wang, T. Warnow, and S. Wyman. A new fast heuristic for computing the breakpoint phylogeny and experimental phylogenetic analyses of real and synthetic data. In *Proc. 8th Int'l Conf. on Intelligent Systems for Mol. Biol. ISMB-2000*, pages 104–115, 2000.
7. S. Downie and J. Palmer. Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In P. Soltis, D. Soltis, and J. Doyle, editors, *Plant Molecular Systematics*, pages 14–35. Chapman and Hall, 1992.
8. 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.

9. S. Hannenhalli and P. 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.
10. S. Hannenhalli and P. Pevzner. Transforming mice into men (polynomial algorithm for genomic distance problems). In *Proc. 36th Ann. IEEE Symp. Foundations of Comput. Sci. FOCS 95*, pages 581–592. IEEE Press, 1995.
11. B. Larget, J. 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/`.
12. B. Moret, J. Tang, L.-S. Wang, and T. Warnow. Steps toward accurate reconstructions of phylogenies from gene-order data. *J. Comput. Syst. Sci.*, 2002. in press.
13. B. Moret, L.-S. Wang, T. Warnow, and S. Wyman. New approaches for reconstructing phylogenies from gene-order data. In *Proc. 9th Int'l Conf. on Intelligent Systems for Mol. Biol. ISMB 2001*, volume 17 of *Bioinformatics*, pages S165–S173, 2001.
14. B. Moret, S. Wyman, D. Bader, T. Warnow, and M. Yan. A new implementation and detailed study of breakpoint analysis. In *Proc. 6th Pacific Symp. Biocomputing PSB 2001*, pages 583–594. World Scientific Pub., 2001.
15. J. Nadeau and B. Taylor. Lengths of chromosome segments conserved since divergence of man and mouse. *Proc. Nat'l Acad. Sci. USA*, 81:814–818, 1984.
16. R. Olmstead and J. Palmer. Chloroplast DNA systematics: a review of methods and data analysis. *Amer. J. Bot.*, 81:1205–1224, 1994.
17. J. Palmer. Chloroplast and mitochondrial genome evolution in land plants. In R. Herrmann, editor, *Cell Organelles*, pages 99–133. Springer Verlag, 1992.
18. I. Pe'er and R. Shamir. The median problems for breakpoints are NP-complete. *Elec. Colloq. on Comput. Complexity*, 71, 1998.
19. L. Raubeson and R. Jansen. Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. *Science*, 255:1697–1699, 1992.
20. L. A. Raubeson, B. M. Moret, J. Tang, S. K. Wyman, and T. Warnow. Inferring phylogenetic relationships using whole genome data: A case study of photosynthetic organelles and chloroplast genomes. Technical Report TR-CS-2001-19, U. of New Mexico, Albuquerque, New Mexico, 2001.
21. D. Sankoff and M. Blanchette. Multiple genome rearrangement and breakpoint phylogeny. *J. Comp. Biol.*, 5:555–570, 1998.
22. A. Siepel. Exact algorithms for the reversal median problem. Master's thesis, U. New Mexico, Albuquerque, NM, 2001. Available at `www.cs.unm.edu/∼acs/thesis.html`.
23. A. Siepel. An algorithm to find all sorting reversals. In *Proc. 6th Int'l Conf. on Comput. Mol. Biol. RECOMB02*. ACM Press, 2002. to appear.
24. A. Siepel and B. Moret. Finding an optimal inversion median: experimental results. In *Proc. 1st Workshop on Algs. in Bioinformatics WABI 2001*, volume 2149 of *Lecture Notes in Computer Science*, pages 189–203. Springer-Verlag, 2001.
25. D. Swofford, G. Olson, P. Waddell, and D. Hillis. Phylogenetic inference. In D. Hillis, C. Moritz, and B. Mable, editors, *Molecular Systematics, 2nd ed.*, chapter 11. Sinauer Associates, 1996.
26. L.-S. Wang, R. Jansen, B. Moret, L. Raubeson, and T. Warnow. Fast phylogenetic methods for the analysis of genome rearrangement data: an empirical study. In *Proc. 7th Pacific Symp. Biocomputing PSB 2002*, pages 524–535. World Scientific Pub., 2002.
27. L.-S. Wang and T. Warnow. Estimating true evolutionary distances between genomes. In *Proc. 33rd Symp. on Theory of Comp. STOC01*, pages 637–646. ACM Press, 2001.