New Approaches for Reconstructing Phylogenies from Gene Order Data

Bernard M.E. Moret* Li-San Wang† Tandy Warnow† Stacia K. Wyman†

Abstract

We report on new techniques we have developed for reconstructing phylogenies on whole genomes. Our mathematical techniques include new polynomial-time methods for bounding the inversion length of a candidate tree and new polynomial-time methods for estimating genomic distances which greatly improve the accuracy of neighbor-joining analyses. We demonstrate the power of these techniques through an extensive performance study based upon simulating genome evolution under a wide range of model conditions. Combining these new tools with standard approaches (fast reconstruction with neighbor-joining, exploration of all possible refinements of strict consensus trees, etc.) has allowed us to analyze datasets that were previously considered computationally impractical. In particular, we have conducted a complete phylogenetic analysis of a subset of the Campanulaceae *family, confirming various conjectures about the relationships among members of the subset and about the principal mechanism of evolution for their chloroplast genome. We give representative results of the extensive experimentation we conducted on both real and simulated datasets in order to validate and characterize our approaches. We find that our techniques provide very accurate reconstructions of the true tree topology even when the data are generated by processes that include a significant fraction of transpositions and when the data are close to saturation.*

Keywords

gene order, genome rearrangement, inversion, transposition, breakpoint, distance estimation, distance correction, neighbor-joining, algorithm engineering, high-performance computing

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 us an important new source of data for phylogeny reconstruction. Many biologists have already embraced this new source of data in their phylogenetic work [13, 22, 23, 25]. 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 [11] was devoted to this topic, as is an upcoming workshop at DIMACS [12].

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 of *three* genomes under such models is NP-hard [7, 24]. All approaches

*Contact author, Dept. of Computer Science, U. of New Mexico, Albuquerque, NM 87131, moret@cs.unm.edu, phone 505-277-5699, FAX 505-277-6927, supported by NSF grant ITR 00-81404

†Dept. of Computer Sciences, U. of Texas, Austin, TX 78712, lisan,tandy,stacia@cs.utexas.edu; supported by the David and Lucile Packard Foundation.

to phylogeny reconstruction for such data must therefore find ways of handling the significant computational difficulties. Moreover, because suboptimal solutions can yield very different evolutionary reconstructions, exact solutions are strongly preferred over approximate solutions (see [28]).

For some datasets (e.g., chloroplast genomes of land plants), biologists conjecture that the only rearrangement events that occur are *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. Researchers have used breakpoint distance as an independent measure of distance between genomes and the *breakpoint phylogeny*, proposed by Blanchette *et al.* [6], is the most parsimonious tree with respect to breakpoint distances.

2 Prior Results

We build upon several major prior results.

BPAnalysis. Blanchette *et al.* [6] proposed the breakpoint phylogeny (finding the tree with the fewest breakpoints) and developed a reconstruction method, `BPAnalysis` [26], 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. And hence, the total running time is exponential in *both* the number of genes and the number of genomes.

MPBE. We developed an alternate method, based on a binary encoding of breakpoints, to take advantage of existing parsimony software [9, 10]. This method, called *Maximum Parsimony on Binary Encodings* (MPBE), is exponential only in the number of genomes (because the parsimony problem is NP-hard), runs very fast in practice, but returns only candidate tree topologies and so must make use of the labeling phase of `BPAnalysis` in order to return ancestral genomes. (Similar approaches based on neighbor-joining suffer from the same problem.)

GRAPPA. We reimplemented `BPAnalysis` in order to analyze our larger datasets and also to experiment with alternate approaches. Our program, called `GRAPPA` [20], includes all of the features of `BPAnalysis`, but runs about three orders of magnitude faster. As part of the development of `GRAPPA`, we designed a new and very fast linear-time algorithm for computing inversion distances [5], which has enabled us to extend our work on breakpoint phylogeny to the inversion phylogeny.

IEBP. We developed a mathematical technique for estimating the maximum likelihood evolutionary distance between two genomes [29]. This technique, called IEBP for "Inverting Expected Breakpoint Distances," has provable error bounds and performs well empirically. Furthermore, using IEBP distances for neighbor-joining analyses results in improved estimations of the true phylogenetic tree.

3 New Results

We present several new results in this paper:

- EDE, a new polynomial-time technique for estimating evolutionary distances between genomes. EDE is not as good an estimator of evolutionary distances as IEBP, but neighbor-joining trees based upon EDE estimations of distances are more accurate than neighbor-joining trees based upon any other distance, including IEBP distances.

- A simulation study examining the relationship between topological accuracy and two definitions of tree length: the number of breakpoints on the tree and the number of inversions on the tree. We find that both definitions for tree length are correlated with topological accuracy, but that the correlation is weakest for genomes on 37 genes (the mitochondrial genome case), especially when the dataset is close to saturation.
- A detailed study of the efficacy of using a simple lower bound on the inversion length of a candidate phylogeny. We observe that this simple bound can quickly eliminate close to 100% of the candidate trees when the evolutionary rates are sufficiently low.
- A successful analysis of the *Campanulaceae* dataset using a combination of these techniques, resulting in a *million-fold* speedup over previous approaches.

Our research combines the development of mathematical techniques with extensive experimental performance studies. We present a cross-section of the results of the experimental study we conducted to characterize and validate our approaches. We used a large variety of simulated datasets as well as several real datasets (chloroplast and mitochondrial genomes) and tested speed (in both sequential and parallel implementations), robustness (in particular against mismatched models), efficacy (for our new bounding technique), and accuracy (for reconstruction and distance estimation).

4 Basic material

4.1 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, for $i \leq j$, 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 (linear or circular) 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 (with the additional assumption of $k > j$) 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 an inversion composed with a transposition. 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 is used in the model, we speak of inversion distance or transposition distance.

A dataset of genomes is said to be *saturated* if it contains a pair of genomes whose inversion distance is as large as the expected distance between two completely unrelated genomes. This saturation value depends upon the number of genes and is bounded from above by n, the maximum distance between any two genomes on n genes [18]. Reconstructing trees from saturated datasets is difficult because of the seeming randomness in the data—this is well understood for gene sequences [15, 16, 28], so it is no surprise that it applies to gene rearrangements as well.

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' (and is symmetric).

4.2 The Nadeau-Taylor model

The *Nadeau-Taylor* model [21] of genome evolution uses only genome rearrangement events.

This results in all genomes having 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.

4.3 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, which defines the probability that a rearrangement event is an inversion, transposition, or inverted transposition. We used weights equal to 1:0:0 (inversions only), 0:1:0 (transpositions only), and 1:1:1 (all three events are equally probable).

In our experimental studies, we used random leaf-labelled trees as the topologies and assigned uniform edge lengths to these trees. The simulator generates signed circular orderings of the genes as follows. The root is assigned the identity gene ordering $g_1, g_2, \ldots, g_k$. When traversing an edge e with expected number of events λ_e, three random numbers are generated according to the model weight: the first determines the actual number of inversions on that edge, the second that of transpositions, and the third that of inverted transpositions. Once the number of events of each type is determined, the order of these events is randomly selected, as are the indices on which these events operate. This process produces a set of circular, signed gene orders for each genome at the leaves of the model tree.

4.4 Labelling internal nodes

The approach proposed by Sankoff and Blanchette to derive ancestral genomes for the internal nodes of a tree is iterative, using a local optimization strategy. After initial labels have been assigned in some way, their procedure repeatedly traverses the tree, computing the breakpoint median of its three neighbors for each node, and using it as the new label if this change improves the overall breakpoint score. 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 guarantees, but does very 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 get a potential label for the 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 on the inversion score of the tree. In addition, `GRAPPA` can substitute approximate TSP solvers (greedy and variations of Lin-Kernighan [17]) for the exact one whenever the exact solver gets bogged down by a TSP instance.

4.5 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 G 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 bipartitions that appear in T that do not appear in T'. This number is normalized by dividing by the number of non-trivial bipartitions of T. Note that, if this rate is 0, then T' equals T or refines it.

5 Estimating Distances

5.1 Introduction

Because the distance between two genomes is defined to be the minimum number of events needed to transform one genome into another, it may underestimate the number of events that actually took place during evolution. While the evolutionary processes that lead to changes in gene sequences are different from those that lead to genome rearrangements, the extensive literature on the improvement obtained with NJ by giving it "corrected" distances (so that they are good estimates of the actual number of events) [28] suggests strongly that comparable improvements might be obtained by correcting distances for genome rearrangements. Furthermore, there is both theoretical and empirical evidence that the trees reconstructed by most distance methods, including NJ, degrade significantly (in terms of topological accuracy) under high rates of evolution [1, 15]. For these reasons, we have given a lot of attention to providing improved estimates of inversion distances.

5.2 IEBP—a theoretical correction

We have developed a general analytical technique for estimating the expected number of breakpoints produced by a sequence of random events in the Nadeau-Taylor model with arbitrary weights γ_I, γ_T, and γ_{IT}, for inversion, transposition, and inverted transposition, respectively [29]. The technique applies to datasets that are either signed or unsigned, circular or linear, and for many other event classes. The technique has two steps:

1. Given a pair of genomes G and G', compute the breakpoint distance.
2. Estimate k, the number of events along the simple path P in the evolutionary tree between G and G'.

Computing the number of breakpoints can be done in linear time; finding k can be done by numerical methods such as bisection.

For any genome X, define the function $B_i(X)$ as follows: $B_i(X) = 1$ if there is a breakpoint between genes i and $i+1$ in genome X, and $B_i(X) = 0$ otherwise. Let X_k be the genome obtained by applying a random sequence of k events to X_0. (All X_k's have the same gene content.) The following theorem shows that $\mathcal{F}_k$, our estimate for $E[BP(X_k, X_0)]$, has low error:

Theorem 1 *(From [29].) Let $\mathcal{E}$ be a class of rearrangement events such that for any genome X, $\rho \in \mathcal{E}$, and breakpoint position i,* $\Pr(B_i(\rho X) = 1 \mid B_i(X) = 0)$ *is independent of X. Then* $|\mathcal{F}_k(\mathcal{E}) - E[BP(X_k, X_0)]| = O(1)$.

Denote by $\mathcal{H}$ the class of rearrangement events in the Nadeau-Taylor model of evolution associated with any setting of γ_I, γ_T, and γ_{IT}.

Theorem 2 *(From [29].)* $\forall k \geq 0$,

$$|\mathcal{F}_k(\mathcal{H}) - E[BP(X_k, X_0)]| \leq 1 + \frac{1}{n-1}.$$

5.3 EDE—an empirical correction

Although NJ using our IEBP estimator shows marked improvement over NJ using breakpoint or inversion distances, it too degrades in accuracy when given data close to saturation. This degradation motivated us to design a correction function to apply to input distance matrices so as to improve the behavior of NJ on nearly saturated data. We used extensive simulations to obtain large amounts of information on the relationship between actual and minimal distances, then designed a correction function, EDE, using various fitting tools and numerical techniques.

To develop an estimator for the actual number of inversions under which NJ performs well, we simulated evolution for a large range of numbers n of genes and numbers k of (random) inversions. We then normalized observed values by the number of genes, plotted the (normalized) actual number of inversions against the (normalized) minimum inversion distance, computed the means of the sets of values for which x is fixed, and graphed this mean. The curve we obtained suggested a function F mapping normalized numbers of actual inversions to normalized inversion distances. This function F must have the following properties:

1. $0 \leq F(x) \leq x$ (obviously)
2. $\lim_{x\to\infty} F(x) = a_n$, where a_n is the expected inversion distance between two random genomes on n genes, divided by n

3. $F'(0) = 1$, because initially every inversion increases the inversion distance by 1
4. $F^{-1}(y)$ is defined for all $y \in [0,1]$. We also assume that F is monotone increasing (additional inversions generally, if not always, increase the inversion distance) to allow us to infer F^{-1}

A ratio of second-degree polynomials satisfies constraints (2)–(4), so we used $F_0(x) = \frac{ax^2+bx}{x^2+cx+b}$. Experiments showed that setting $a = 1$ for all values of n produces the best results. To estimate b and c, we minimized the least-square error between F_0 and the empirical data—that is, we minimized $\sum_{(x,y)} |F_0(x) - y|^2$. Using gradient descent methods, we obtained $b = 0.5956$ and $c = 0.4577$. Because this definition of F_0 does not always satisfy constraint (1), we set $F(x) = \min\{F_0(x), x\}$; this is the "fixed-point modification."

We can now define EDE to be the nonnegative inverse of F. EDE overestimates the actual number of inversions for large inversion distances. However, this overestimation appears not to affect the performance of NJ (we explored several ways of modifying the latter values, but did not obtain an improvement).

5.4 Comparison of different distances

We simulated the Nadeau-Taylor model under different weight settings to study the behavior of different distance methods. All datasets have 120 genes. For each dataset in the experiment, we chose a number between 1 and 300 (2.5 times the number of genes) as the number of rearrangement events. We then computed the BP (breakpoint) and INV (inversion) distances and corrected them using IEBP and EDE distances. Figure 1 shows our findings. The figure suggests that BP and INV distances underestimate the number of events, although they are highly accurate and have small variance when the number of events is low. The linear region—the range of the x-coordinate values where the curve is a straight line—is larger for INV than for BP, so that INV produces unbiased estimates in a larger range than does BP. IEBP produces good estimates over all ranges. EDE has more erratic behavior—the curve initially has the same shape as INV due to the fixed-point modification, but it constantly underestimates thereafter. Both EDE and IEBP have larger variances than BP and INV.

5.5 Neighbor-joining performance

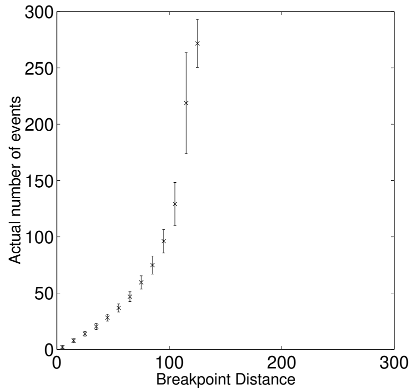
We conducted a simulation study to compare the performance of NJ using four different distances: BP, INV, IEBP, and EDE. Figure 2 shows our findings. This figure shows false negative rates for a best case—inversion-only scenarios—and for a nearly worst-case—scenarios with both transpositions and inverted transpositions. Note that NJ with EDE is remarkably robust: even though it was engineered for inversions only, it handles datasets with a large number of transpositions and inverted transpositions almost as well. NJ with EDE can recover 90% of the edges even for the close-to-saturation datasets where the maximum pairwise inversion distance is close to 90% of the maximum value. NJ with EDE is even competitive with the more computationally intensive MPBE (Maximum Parsimony on Binary Encoding) method (data not shown).

6 Better Analyses

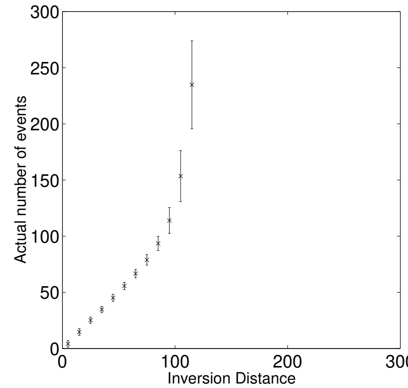
6.1 Parsimony improves accuracy

Since the main goal of phylogeny reconstruction is producing the correct tree topology, the parsimony approach we have taken needs to be evaluated in terms of the topological accuracy of the trees produced. We ran a large series of tests on model trees to test the hypothesis that reducing the total breakpoint or inversion length of trees would yield more topologically accurate trees.

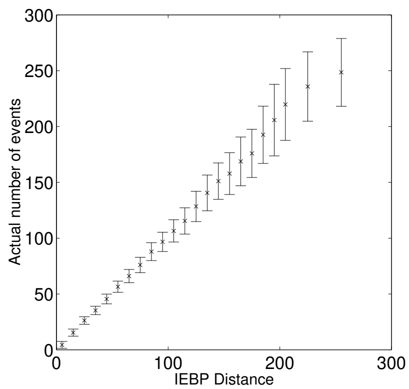
We ran a total of 209 tests of NJ with each of inversion and breakpoint distances, each test with at least 12 data points, on sets of up to



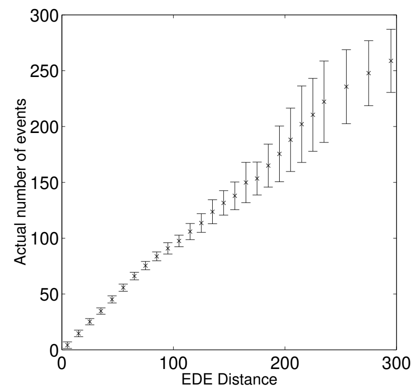
(a) BP distance



(b) INV distance



(c) IEBP distance



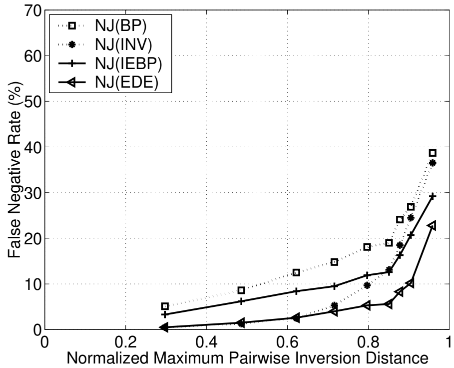
(d) EDE distance

Figure 1: Mean and standard deviation plots for different distance estimators, for 120 genes under inversion-only scenarios. The datasets are divided into bins according to their *x*-coordinate values (the BP or INV distance).

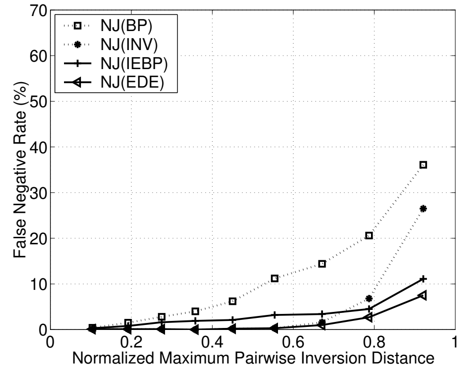
40 genomes. We used two genome sizes (37 and 120 genes, representative of mitochondrial and chloroplast genomes, respectively) and various ratios of inversions to transpositions and inverted transpositions, as well as various evolutionary rates. For each dataset, we computed the total inversion and breakpoint distances and compared their values with the percentage of errors (measured as false negatives). We used the nonparametric Cox-Stuart test [8] for detecting trends—i.e., for testing whether reducing breakpoint or inversion distance consistently reduced topological errors. Using a 95% confidence level, we found that over 97% of the datasets with inversion distance and over 96% of those with breakpoint distance exhibited such a trend. Indeed, even at the 99.9% confidence level, over 82% of the datasets still exhibited such a trend.

Figure 3 shows the results of scoring the different NJ trees under the two optimization criteria: breakpoint score and inversion length. Only the inversion-only scenario is shown here, since other evolutionary settings produce similar behavior. In general, the relative ordering and trend of the curves agree with the curves of Figure 2, suggesting that decreasing the number of inversions or breakpoints should lead to an improvement in topological accuracy. This correlation is strongest for the 120-gene case and somewhat weaker for the 37-gene case. Finally, this trend still holds under the other evolutionary models (such as when only transpositions occur).

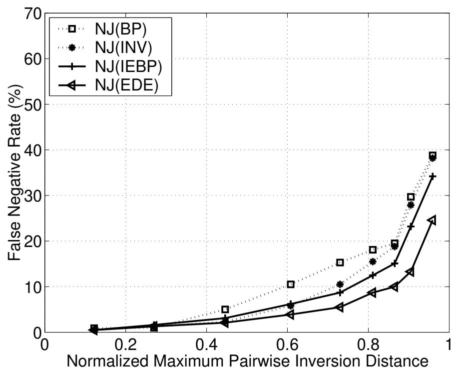
These experiments support the conjecture that improving the inversion length or breakpoint length should lead to improved topological ac-



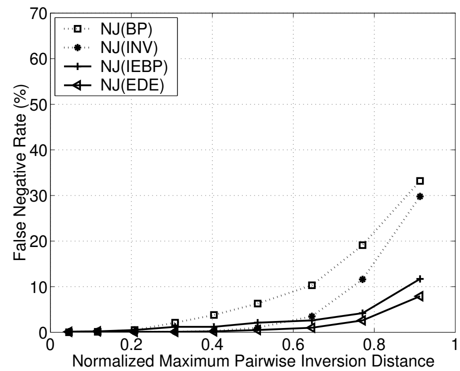
(a) inversion only, 37 genes



(b) inversion only, 120 genes



(c) three classes equally likely, 37 genes



(d) three classes equally likely, 120 genes

Figure 2: False negative rates of NJ methods under various distance estimators as a function of the maximum pairwise inversion distance, for 10, 20, 40, 80, and 160 genomes. Model weight settings are 1:0:0 (inversion only) and 1:1:1 (equally likely events).

curacy, at least for the case of chloroplast genomes (which have many genes) and of mitochondrial genomes where the rates of evolution are sufficiently low to keep the dataset below saturation.

6.2 A lower bound using circular orderings

The following theorem is well known:

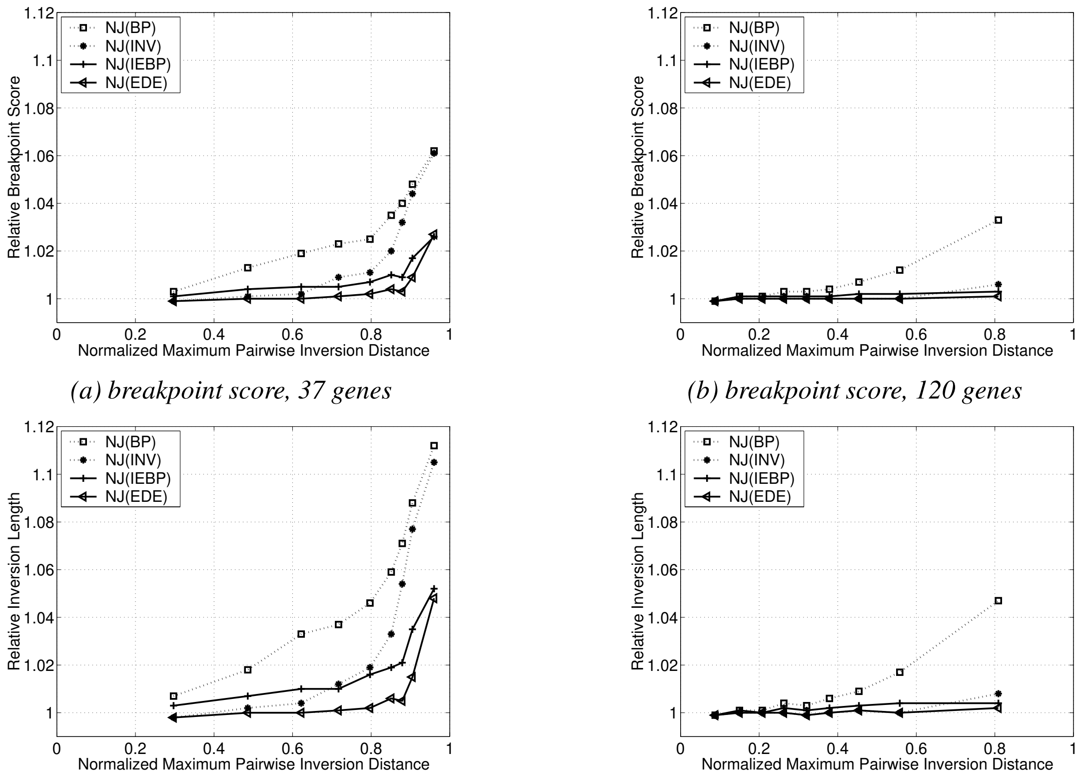
Theorem 3 *Let d be a $n \times n$ matrix of pairwise distances between the taxa in a set S; let T be a tree leaf-labelled by the taxa in S; and let w be an edge-weighting on T, so that we have $w_{ij} = \sum_{e \in P_{ij}} w(e) \geq d_{ij}$. Set $w(T) = \sum_{e \in E(T)} w(e)$. If $1, 2, \ldots, n$ is a circular ordering of the leaves of T, under some planar embedding of T, then we have $2w(T) \geq d_{1,2} + d_{2,3} + \ldots + d_{n,1}$.*

And this corollary immediately follows:

Corollary 1 *Let d be the matrix of minimum inversion distances between every pair of genomes in a set S, let T be a fixed tree on S, and let $1, 2, \ldots, n$ be the circular ordering of leaves in T. Then the inversion length of T is at least $\frac{1}{2}(d_{1,2} + d_{2,3} + \ldots + d_{n,1})$.*

(This corollary forms the basis of the old "twice around the tree" heuristic for the TSP based on minimum spanning trees [14].)

We use these bounds to help search tree space in the obvious way. First, we obtain a good upper bound on the minimum achievable inversion length by using our polynomial-time technique (NJ with EDE distances); we update this upper bound every time the search finds a better tree. For each tree, we quickly compute the circular lower bound of Corollary 1. If that lower bound



(a) breakpoint score, 37 genes *(b) breakpoint score, 120 genes*

(c) inversion length, 37 genes *(d) inversion length, 120 genes*

Figure 3: Scoring NJ methods under various distance estimators as a function of the maximum pairwise inversion distance for 10, 20, and 40 genomes. Plotted is the ratio of the NJ tree score to the model tree score (breakpoint or inversion) on an inversion-only model tree.

exceeds the upper bound, the tree can be discarded without being scored. Since scoring a tree involves solving numerous TSP instances, such bounding can dramatically reduce the running time.

6.3 The lower bound in practice

We ran three different experiments to quantify various aspects of our bounding techniques. Our first experiment measures the percentage of trees that are pruned through bounding (and thus not scored) as a function of the three model parameters: number of genomes, number of genes, and number of inversions per edge. We used an inversion-only scenario, as well as one with approximately half inversions and half transpositions or inverted transpositions. Our data consisted of two collections of 10 datasets each for a combination of parameters. The number of genomes varied over $\{10,20,40,80,160\}$, the number of genes varied over $\{10,20,40,80,160,320\}$, and the rate of evolution varied from 2 to 8 events per tree edge, for a total of 90 parameter combinations and thus 1,800 datasets. For each dataset, we generated 1,000 random circular orderings, scored their pairwise circular order inversion distance, and compared these scores to the upper bound. Table 1 shows the percentage of trees pruned away by the circular lower bound. The lower bound is surprisingly effective for certain datasets, but for many others would not eliminate any trees. For low rates of evolution and datasets with up to 160 genomes, we found that most random circular orderings were eliminated—a very encouraging result for chloroplast genomes (which can contain several

Table 1: Percentage of trees eliminated through bounding.

	$r = 2$					$r = 4$				$r = 8$				r value
	10	20	40	80	160	10	20	40	80	10	20	40	80	# genomes
10	0	0	0	1	1	0	0	0	1	0	0	0	1	
20	0	80	91	1	1	0	0	0	0	0	0	0	0	
40	91	100	100	100	100	0	0	0	0	0	0	0	0	
80	99	100	100	100	100	65	72	100	0	0	0	0	0	
160	100	100	100	100	100	98	100	100	100	0	0	0	0	
320	100	100	100	100	100	99	100	100	100	71	90	100	0	
# genes														

hundred genes) and quite sufficient for the analysis of mitochondrial datasets (where the number of genes is 37), since many have low evolutionary rates. However, note that at higher rates of evolution, the bound is not effective until the number of genes gets into the range of 80 to 160, while the bound is simply ineffective at truly high rates of evolution.

Our second experiment used our real dataset of 13 chloroplast genomes, 12 from the family *Campanulaceae* and with Tobacco as an outgroup. Each of the 13 genomes has 105 gene segments and, though highly rearranged, has what we consider to be a low rate of evolution. We ran `GRAPPA` on this dataset both with and without the lower bound and computed the percentage of trees eliminated using the bound. After running for 12 hours (on a 300MHz Pentium II workstation) and processing well over 50 million trees, the code using bounding had eliminated 85% of the trees from further computation.

Because computing the circular bound entails its own cost (linear in the number of genomes), we were interested in what kind of running time speedup `GRAPPA` would gain through this bounding technique. Our third experiment ran our code on the *Campanulaceae* dataset for 12 hours each with and without bounding: the version with bounding processed nearly 10 times as many trees. Thus the speedup over `BPAnalysis` reported in [20] is now increased by another factor of 5–10, to a value of over 5,000.

The speedup obtained by bounding depends upon two factors: the percentage of trees that can be eliminated by the bounding and the difficulty of the TSP instances avoided by using the bounds. As Table 1 shows, when the rate of evolution is not too high, close to 100% of the trees can be eliminated by using the bounds. However, the TSP instances solved in `GRAPPA` can be quite small when the evolutionary rate is low, due to how we compress data (as described in [20]). Consequently, the speedup will also depend upon the rate of evolution, with lower rates of evolution producing easier TSP instances and thus smaller speedups. The *Campanulaceae* dataset is a good example of a dataset that is quite easy for `GRAPPA`, in the sense that it produces easy TSP instances—but even in this case, a significant speedup results. More generally, the speedup increases with larger numbers of genomes and, to a point, with higher rates of evolution. When one is forced to exhaustively search tree space, these speedups represent substantial savings in time.

Our last experiment used a combined heuristic. We analyzed the *Campanulaceae* dataset using NJ and MPBE and then took the strict consensus tree (the maximally resolved tree that is a common contraction) of the 8 trees returned by these procedures. We gave this tree as a constraint tree to `GRAPPA`; this makes `GRAPPA` search the space of all refinements of the constraint tree for the minimum inversion tree. The search space contained only 10,395 trees, which we can run to completion in much less than a minute (though not because of the bounding technique, since it did not eliminate any tree, an expected occurrence when all trees examined have a good topology). The search returned 216 optimal trees, with an inversion score of 67 and a breakpoint score of 84. Since earlier attempts to analyze this dataset found only four trees with an inversion score of 67 [9], this represents a significant advance.

7 High-Performance Computing

Even the best algorithms for phylogeny reconstruction are likely to take exponential time in many cases, so that we should take advantage of high-performance tools whenever possible. We used the best precepts of algorithm engineering [19] to improve the running time of our `GRAPPA` software, eventually achieving a 2,000-fold speedup, as reported in [20]. More recently, we parallelized our software (an easy task, since it offers "embarrassing parallelism") and used the 512-processor Los Lobos supercluster at the U. of New Mexico to run a complete analysis of the *Campanulaceae* dataset discussed in [9]. This analysis took only 1.5 hours instead of the several centuries estimated in [9], for a million-fold speedup [3]. We expect to effect similar speedups (by several orders of magnitude) in a future reimplementation of parsimony searches (both local, using TBR techniques, and global, using branch-and-bound searches), based on the same principles of high-performance algorithm engineering and parallel algorithm development. Although even a million-fold speedup will allow us to increase the number of taxa by only a few when using an exponential-time algorithm, the same speedup applied to a polynomial-time algorithm will represent the difference between solving a problem today or waiting a few generations. We have also produced the first ever linear parallel speedups for complex combinatorial problems [2], using shared-memory machines (SMPs). Branch-and-bound falls in this category of problems, so that we can now expect to see respectable parallel speedups in parsimony searches and other related optimization problems when using our newly developed parallel techniques [4].

8 Conclusions and Future Work

We have described new theoretical and experimental results that have enabled us to analyze significant datasets in terms of inversion events and that also extend to models incorporating transpositions. The work described here is part of an ongoing project to develop fast and robust techniques for reconstructing phylogenies from gene order data. Our current software suffers from limitations that we need to address; most limiting is the fact that it explicitly searches all of (constrained) tree space. However, the bounds we have described can be used in conjunction with branch-and-bound (based upon either inserting leaves into subtrees or extending circular orderings), as well as in heuristic techniques for searching through tree space. Our immediate work will implement these extensions to the software. In the long term, we plan to extend the techniques to solving the IT (inversion plus transposition) phylogeny problem, enable analysis of genomes with unequal gene sets, and handle multiple chromosomes.

Acknowledgments

We thank David Sankoff and Joseph Nadeau for inviting us to the DCAF meeting, during which some of the ideas in this paper came to fruition.

References

[1] Atteson, K., "The performance of the neighbor-joining methods of phylogenetic reconstruction," *Algorithmica* **25**, 2/3 (1999), 251–278.

[2] Bader, D.A., Illendula, A.K., & Moret, B.M.E., "Using PRAM algorithms on a uniform-memory-access shared-memory architecture," preprint.

[3] Bader, D.A., & Moret, B.M.E., "GRAPPA runs in record time," *HPC Wire*, **9**, 47 (Nov. 23), 2000.

[4] Bader, D.A., Moret, B.M.E., Warnow, T., Wyman, S.K., & Yan, M., "High-performance algorithm engineering for gene-order phylogenies," DIMACS Workshop on Whole Genome Comparison, Rutgers U., 2001.

[5] Bader, D.A., Moret, B.M.E., & Yan, M., "A fast linear-time algorithm for inversion distance with an experimental comparison," preprint.

[6] Blanchette, M., Bourque, G., & Sankoff, D., "Breakpoint phylogenies," in *Genome Informatics 1997*, Miyano, S., & Takagi, T., eds., Univ. Academy Press, Tokyo, 25–34.

[7] Caprara, A., "Formulations and hardness of multiple sorting by reversals," *Proc. 3rd Int'l Conf. on Comput. Molecular Biology RECOMB99*, ACM Press, NY (1999), 84–93.

[8] Conover, W. J. *Practical Nonparametric Statistics*. 3rd ed., John Wiley & Sons (1999), 170–176.

[9] Cosner, M.E., Jansen, R.K., Moret, B.M.E., Raubeson, L.A., Wang, L.-S., Warnow, T., & Wyman, S., "A new fast heuristic for computing the breakpoint phylogeny and experimental phylogenetic analyses of real and synthetic data," *Proc. 8th Int'l Conf. on Intelligent Systems for Molecular Biology ISMB-2000*, San Diego (2000), 104–115.

[10] Cosner, M.E., Jansen, R.K., Moret, B.M.E., Raubeson, L.A., Wang, L.S., Warnow, T., & Wyman, S., "An empirical comparison of phylogenetic methods on chloroplast gene order data in Campanulaceae," in *Comparative Genomics*, D. Sankoff & J.H. Nadeau, eds., Kluwer Acad. Pubs. (2000), 99–122.

[11] *DCAF: Gene Order Dynamics, Comparative Maps, and Multigene Families*, proceedings available as *Comparative Genomics*, D. Sankoff & J.H. Nadeau, eds., Kluwer Acad. Pubs. (2000).

[12] *DIMACS Whole Genome Comparison Workshop*, to be held at DIMACS Center, Piscataway, NJ, Feb. 28 – Mar. 2, 2001.

[13] Downie, S.R., & Palmer, J.D., "Use of chloroplast DNA rearrangements in reconstructing plant phylogeny," in *Plant Molecular Systematics*, Soltis, P., Soltis, D., & Doyle, J.J., eds., Chapman & Hall, NY (1992), 14–35.

[14] Held, M., & Karp, R.M., "The travelling salesman problem and minimum spanning trees," *Ops. Res.* **18** (1970), 1138.

[15] Huson, D., Nettles, S., Rice, K., Warnow, T., & Yooseph, S., "Hybrid tree reconstruction methods," *ACM J. Experimental Algorithmics* **4**, 5 (1999), `www.jea.acm.org/1999/HusonHybrid/`.

[16] Huson, D., Smith, K.A., & Warnow, T., "Correcting large distances for phylogenetic reconstruction," *Proc. 3rd Workshop on Algorithm Engineering WAE99*, London, LNCS **1668**, Springer Verlag (1999), 273–286.

[17] Johnson, D.S., & McGeoch, L.A., "The traveling salesman problem: a case study," in *Local Search in Combinatorial Optimization*, E. Aarts & J.K. Lenstra, eds., John Wiley, NY (1997), 215–310.

[18] Meidanis, J., Walter, M.E.M.T., & Dias, Z., "Reversal distance of signed circular chromosomes," unpublished manuscript, 2000.

[19] Moret, B.M.E., "Towards a discipline of experimental algorithmics," to appear in *DIMACS Monographs*, 2001.

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

[21] Nadeau, J.H., & Taylor, B.A., "Lengths of chromosome segments conserved since divergence of man and mouse," *Proc. Nat'l Acad. Sci. USA* **81** (1984), 814–818.

[22] Olmstead, R.G., & Palmer, J.D., "Chloroplast DNA systematics: a review of methods and data analysis," *Amer. J. Bot.* **81** (1994), 1205–1224.

[23] Palmer, J.D., "Chloroplast and mitochondrial genome evolution in land plants," in *Cell Organelles*, Herrmann, R., ed., Springer Verlag (1992), 99–133.

[24] Pe'er, I., & Shamir, R., "The median problems for breakpoints are NP-complete," *Elec. Colloq. on Comput. Complexity*, ECCC-71, 1998.

[25] Raubeson, L.A., & Jansen, R.K., "Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants," *Science* **255** (1992), 1697–1699.

[26] Sankoff, D., & Blanchette, M., "Multiple genome rearrangement and breakpoint phylogeny," *J. Computational Biology* **5** (1998), 555–570.

[27] Saitou, N., & Nei, M., "The neighbor-joining method: A new method for reconstructing phylogenetic trees," *Mol. Biol. & Evol.* **4** (1987), 406–425.

[28] Swofford, D., Olson, G., Waddell, P., & Hillis, D., "Phylogenetic inference," in *Molecular Systematics* (2nd ed.), D. Hillis, C. Moritz, & B. Mable, eds., Sinauer Associates Inc., Sunderland, Mass. (1996), chap. 11.

[29] Wang, L.-S, & Warnow, T., "New polynomial-time methods for whole-genome phylogeny reconstruction," to appear in *Proc. 33rd Symp. on Theory of Comp. (STOC'01)*.