Fast Phylogenetic Methods for the Analysis of Genome Rearrangement Data: An Empirical Study

Li-San Wang
Dept. of Computer Sciences
University of Texas, Austin, TX 78712

Robert K. Jansen
Section of Integrative Biology
University of Texas, Austin, TX 78712

Bernard M.E. Moret
Dept. of Computer Science
University of New Mexico
Albuquerque, NM 87131

Linda A. Raubeson
Dept. of Biological Sciences
Central Washington University
Ellensburg, WA 98926

Tandy Warnow
Dept. of Computer Sciences
University of Texas
Austin, TX 78712

Evolution operates on whole genomes through mutations that change the order and strandedness of genes within the genomes. Thus analyses of gene-order data present new opportunities for discoveries about deep evolutionary events, provided that sufficiently accurate methods can be developed to reconstruct evolutionary trees. In this paper we present two new methods of character coding for parsimony-based analysis of genomic rearrangements: one called MPBE-2, and a new parsimony-based method which we call MPME (based on an encoding of Bryant), both variants of the MPBE method. We then conduct computer simulations to compare this class of methods to distance-based methods (NJ under various distance measures). Our empirical results show that two of our new methods return highly accurate estimates of the true tree, outperforming the other methods significantly, especially when close to saturation.

1 Introduction

1.1 Gene Orders as a Source of Phylogenetic Data

While DNA sequences have greatly improved our understanding of evolutionary relationships, they have also left open many crucial phylogenetic questions. The research community has thus sought other sources of phylogenetic signal, looking for characters that evolve slowly or have a large number of states, since such characters generally have a higher signal-to-noise ratio than DNA sequences. One source of such characters is the category of "rare genomic changes" [1]. Rare genomic changes are defined as large-scale mutational events in genomes; among many possibilities are *genomic rearrangements*, which include both gene duplications [2] and changes in gene order [3]. The relative rarity of genomic rearrangements makes these characters very attractive as sources of phylogenetic data. Although it has been suggested that there are not enough genomic rearrangements to provide sufficient numbers of characters for resolving phylogenetic relationships in most groups (e.g., chloroplast genomes [4]), increased genome sequencing efforts are uncovering many new genome rearrangements for use in phylogeny reconstruction. For example, gene-order comparisons for two ascomycete fungal nuclear genomes (*Saccharomyces cervisiae* and *Candida albicans*) estimated that there have been approximately 1100 single-gene inversions since the divergence of these species [5].

1.2 Genome rearrangement evolution

Some organisms have a single chromosome or contain single-chromosome organelles (such as mitochondria[6,7] or chloroplasts[3,4]), whose evolution is largely independent of the evolution of the nuclear genome for these organisms. Whole-genome sequencing projects are providing us with information about the ordering and orientation of the genes, enabling us to represent the chromosome by an ordering (linear or circular) of signed genes (where the sign of the gene indicates its orientation). Evolutionary processes on the chromosome can thus be seen as transformations of signed orderings of genes. With a number assigned to the same gene in each genome, a genome can be represented by a *signed permutation* of $\{1, \ldots, n\}$—a permutation in which each number is given a sign; if the genome is circular, so is the permutation.

An *inversion* lifts a contiguous subpiece of the permutation, reverses its order and the orientation of every gene within it, then puts the resulting piece back in the same location; for it to happen requires two concurrent breaks in the DNA. A *transposition* lifts a contiguous subpiece of the permutation and puts it back unchanged between two contiguous permutation elements not in the subpiece; it requires three DNA breaks. An *inverted transposition* is a transposition that also reverses the order of the subpiece and the orientation of every gene within it.

The *Generalized Nadeau-Taylor (GNT)* model[8,9] describes the process responsible for the change in gene order along the edges of a given phylogeny. The model includes the three types of rearrangement events just described; within each type, all events have equal probability (e.g., any inversion is as likely as any other), but the model includes two parameters to indicate the probabilities of each type of event: α and β are the probabilities that an event is a transposition or an inverted transposition, respectively—and thus $(1–\alpha–\beta)$ is the probability that an event is an inversion. Each edge e of the tree has an associated parameter λ_e, which is the expected value of a Poisson distribution for the number of events taking place along this edge. The process that this model describes, when given a rooted binary tree with an ancestral gene order at the root, as well as the values of the various parameters, produces a set of signed gene orders at the leaves of the model tree.

1.3 Approaches for Reconstructing Phylogenies from Gene Order Data

We now describe two basic classes of methods for reconstructing phylogenies from whole genomes. In the first class of *distance-based methods*, we use a method such as the Neighbor Joining Method[10] in conjunction with an estimator of evolutionary distances, to infer an edge-weighted tree whose matrix of leaf-to-leaf distances approximates the estimated distance matrix. Thus, the estimation of evolutionary distances is an important component of distance-based estimation of phylogenies.

The *true evolutionary distance* (*t.e.d.*) between two leaves in the true tree is simply the length (in terms of actual numbers of rearrangements) of the unique simple path between these two leaves in the tree. Theory has established that if we can estimate all t.e.d.s sufficiently accurately, we can reconstruct the tree T using even very simple methods [11,12]. Estimates of pairwise distances that are close to the t.e.d.s will in general be more useful for evolutionary tree reconstruction than edit distances, because edit distances usually *underestimate* t.e.d.s, by an amount that can be very significant as the number of rearrangements increases [13,14].

The other basic approach we examine are called "maximum parsimony" type approaches, since they are similar to the maximum parsimony problem for biomolecular sequences. Given a set R of allowed rearrangement events, the *length* of a tree T with all nodes labeled by genomes is defined as the sum of the edit distances with respect to R over all edges in T. The *parsimony score* of T with respect to R is the minimum length over all possible labelings of the internal nodes. The *Maximum Parsimony on Rearranged Genomes* problem asks for the tree topology T that has minimum parsimony score with respect to R. The problem is difficult even when R is very restricted: the time complexity is unknown (but believed to be NP-hard) when R is the set of all transpositions and is NP-hard when R is the set of all inversions.

Sankoff *et al.* [15] proposed a different optimization problem for phylogeny reconstruction from gene-order data: seek the tree with the minimum number of breakpoints rather than that with the minimum number of evolutionary events. The resulting tree is called the *breakpoint phylogeny*. When the breakpoint distance is linearly correlated with the t.e.d., minimizing the number of breakpoints also minimizes the total number of evolutionary events; Blanchette *et al.* [6] observed such a relationship in a group of metazoan mitochondrial genomes. Computing the breakpoint phylogeny is NP-hard for just three genomes [16], a special case known as the *Median Problem for Breakpoints (MPB)*. Blanchette *et al.* reduced the MPB to the travelling salesman problem and developed the software suite `BPAnalysis` to approximate the breakpoint phylogeny; this approach was subsequently refined and enormously accelerated by Moret *et al.* with the `GRAPPA` software suite [17]. However, these approaches fail on large datasets—a 16-taxon problem gives rise to several quadrillion trees!

Our experiments have shown that selecting trees with smaller total edge length (under any of these measures) leads to more accurate reconstructions [18,19].

1.4 Our Contribution

This paper provides the first thorough empirical study of fast phylogenetic reconstruction methods for gene-order data, using both distance-based and parsimony-based approaches. It also introduces two new analysis methods based on encodings of gene orders as sequences of state characters. In Section 2 we describe the various methods tested in our experiments; in Section 3 we discuss the experimental setup; and in Section 4 we present our results, in terms of efficiency and of topological accuracy. We find that the NJ method, used with our t.e.d. estimator *EDE*, and the MPME method, are both highly accurate, outperforming all of the other methods in this study.

2 Phylogenetic Methods Under Study

The methods used in our experiments can be grouped under *distance-based methods*, which use various distance estimators to recover *true evolutionary distances*, and *parsimony-based methods*, which convert the gene-order data into character codings and use conventional parsimony algorithms to reconstruct the phylogeny.

2.1 Distance-Based Methods

Our basic distances are the breakpoint (`BP`) and the inversion (`INV`) distances. The first measures the number of adjacencies that are disrupted in moving from one ordering into the other, while the second measures the minimum number of inversions required to transform one ordering to the other. Both are computable in linear time (the second through the method of Bader *et al.* [20]). Using these distances, we have three t.e.d. estimators, all of which can be computed in low-order polynomial time.

The `a-IEBP` (Approximately Inverting the Expected BreakPoint distance) method [9] approximates, with known error bound, the expected breakpoint distance obtained after k random events in the GNT model, for any setting of the parameters α and β. Consequently, given two genomes, we can estimate the true evolutionary distance (t.e.d.) between them by selecting the number of events most likely to have created the observed breakpoint distance. Simulations [9] show that the method is robust even under wrong assumptions about model parameters.

The `e-IEBP` (Exactly Inverting...) method [21] improves the accuracy of `a-IEBP` by providing an exact calculation of the expected breakpoint distance, at the cost of increased running time. In our simulations [21], `e-IEBP` produces more accurate trees than `a-IEBP` when used with NJ.

The `EDE` (Empirically Derived Estimator) method [18,19] estimates the t.e.d. by inverting the expected inversion distance. We derived the estimator through a nonlinear regression on simulation data. The evolutionary model in the simulations uses only inversions, but NJ using `EDE` distances shows high accuracy in simulations [21,18] even when transpositions and inverted transpositions are present.

2.2 Parsimony-Based Methods

All methods discussed in this section are based on character-encodings generated from the signed permutation. These character matrices are then subjected to parsimony searches—for which good implementations have long been available.

The *Maximum Parsimony on Binary Encodings (MPBE)* [22,23] has running time exponential in the number of genomes, but runs very fast in practice. In MPBE, each gene ordering is translated into a binary sequence, where each site from the binary sequence corresponds to a pair of genes. (The ordering of the sites is immaterial in this encoding.) For the pair (g_i, g_j), the sequence has a 1 at the corresponding site if g_i is immediately followed by g_j in the gene ordering and a 0 otherwise (note that g_i and g_j can be negative and that, since (g_i, g_j) and ($-g_j$,$-g_i$) denote the same adjacency, w need only one site for both). There are $\binom{n}{2}$ pairs, where n is the number of genes in each genome, but we drop the sites where every sequence has the same value.

Our first new encoding, *MPBE-2*, is a subset of an MPBE encoding designed to eliminate any character denoting the ancestral condition (the identity permutation in our simulations). For example, if the adjacency 1–2 is scored as one character for MPBE, but we can safely assume it is also the state of the common ancestor of the taxa, then we will not include this character in the MPBE-2 encoding. That is, we attempt in MBE-2 to be true to the cladistic goal of using only shared derived mutations to support sister-group relationships. This also has the consequence of reducing dependencies among characters, although it cannot fully eliminate these dependencies.

Our second new encoding builds on this observation by developing an ordered multistate encoding that avoids multiple encodings for each position. This new encoding, which we call *Maximum Parsimony on Multistate Encodings* (MPME). is inspired by a proposal of Bryant's [24], itself based on an earlier characterization approach of Sankoff and Blanchette. Let n be the number of genes in each genome; then each gene order is translated into a sequence of length $2n$. For every i, $1 \leq i \leq n$, site i takes the value of the gene immediately following gene i and site $n+i$ takes the value of the gene immediately following gene -i. For example, the circular gene ordering $(1,\text{-}4,\text{-}3,\text{-}2)$ corresponds to the MPME sequence of $(\text{-}4, 3, 4, \text{-}1, 2, 1, \text{-}2, \text{-}3)$. Each site can take up to $2(n-1)$ different values; the unbounded number of states per characters is a drawback in practical implementations, which usually assume that this number is bounded by a small constant. For example, the bound is 32 in PAUP* 4.0 [25]; even after remapping the set of successor values into a consecutive set of symbols, the number of symbols often exceeds the PAUP bound for larger problems.

3 Design of the Experiments

The goal of our experiments is to compare the tradeoffs (time vs. accuracy) offered by NJ with those offered by the parsimony-based methods; thus we present results for both running time and accuracy.

3.1 Quantifying Accuracy

Given an inferred tree, we assess its *topological accuracy* by computing the *normalized false negative (FN) rate* with respect to the *true tree*. The true tree may not be the model tree itself: the evolutionary process may cause no changes on some edges of the model tree, in which case we define the true tree to be the result of *contracting* those edges in the model tree.

For every tree there is a natural association between every edge e and the bipartition on the leaf set induced by deleting e from the tree. Let T be the true tree and let T' be the inferred tree. An edge of T is *missing* in T' if T' does not contain an edge defining the same bipartition; such an edge is then called a *false negative* (FN). We normalize these values by dividing the number of false negatives by the number of internal edges in the true tree, thus producing a value between 0 and 1 (which we express in terms of percentages).

3.2 The Experiments

For each setting of the parameters (number of leaves, number of genes, probability of each type of rearrangement, and edge lengths), we generate 60 *runs*. In each run, we generate a model tree, and a set of genomes at the leaves as follows. First, we generate a random leaf-labeled tree (from the uniform distribution on topologies); the leaf-labeled tree and the parameter settings thus define a model tree in the GNT model. We run the GNT simulator on the model tree and produce a set of genomes at the leaves. The numbers of genes in each genome are 37 (typical of genes in animal mitochondrial genomes [6]) and 120 (typical of genes in plant chloroplast genomes [23]).

Our GNT simulator[9,18] takes as input a rooted leaf-labeled tree and the associated parameters (edge lengths and the relative probabilities of inversions, transpositions, and inverted transpositions). On each edge, it applies random rearrangement events to the genome at the ancestral node according to the model with given parameters α and β. We use `tgen` (from D. Huson) to generate random trees. These trees have topologies drawn from the uniform distribution, and edge lengths drawn from the discrete uniform distribution on intervals $[a, b]$, where we specify a and b. Table 1 summarizes the settings. We then compute NJ trees on each of the five distance matrices (`BP`, `INV`, `a-IEBP`, `e-IEBP`, and `EDE`) and the most parsimonious trees from the heuristic search using the three encodings (MPBE, MPBE-2, and MPME). When the parsimony search returns more than one tree, we use the majority-rule consensus (generally not a fully resolved tree) for comparison to the true tree. We use PAUP* 4.0b8 [25] for NJ, to compute the false negative rate between two trees, and for the parsimony search using the three encodings. The upper bound for the running time is 240 mins., the heuristic search uses Tree-Bisection-Reconnection (TBR) operations to generate new trees, at any time we hold the 5 trees having the lowest parsimony score, and we use the NJ trees with our five distances as the starting trees. All experiments were conducted on Pentium-class machines.

Table 1: Settings For The Empirical Study.

Parameter	Value ($*$ for 120 genes only)
# genes	37, 120
# leaves	40, 80*, and 160*
expected # events/edge	uniform within $[1,3]$, $[1,5]$, $[1,10]$, $[3,5]^*$, $[3,10]^*$, and $[5,10]^*$
probability settings: (α, β)	$(0,0)$, $(1,0)$, $(0,1)$, $(\frac{1}{2}, \frac{1}{2})$, $(0, \frac{1}{2})$, $(\frac{1}{2}, 0)$, $(\frac{1}{3}, \frac{1}{3})$
datasets per setting	60

4 Results of the Experiments

As mentioned, MPME will exceed 32 states per character for large problems. The problem worsens with increasing rate of evolution; for runs with 120 genes, 160 taxa, and edge lengths in $[5, 10]$, PAUP *always* rejects the MPME data matrix. We ignore all MPME datasets rejected by PAUP—thereby introducing an unknown, but undoubtedly favorable bias in our accuracy results for MPME on large problems.

Figure 1 shows histograms of the running times of the parsimony-based methods for two sizes of problems; on smaller problems (40 taxa), the parsimony search ran quickly (20 mins.), but larger numbers of taxa caused sharp increases in running times—to the point where MPME generally reached the time limit. In comparison, the NJ-based methods ran faster—typically in 8 minutes or less (this time includes calculation of pairwise distances, which can be computationally expensive for the `IEBP` methods), with no variation among runs using a particular estimator.

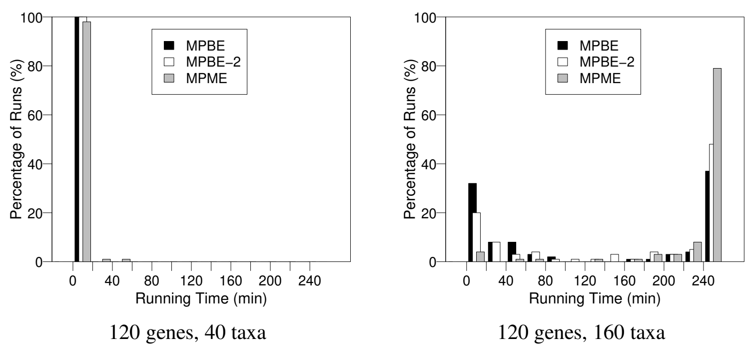


Figure 1: PAUP running times for the three parsimony-based methods. The vertical bars right of 240 mins. represent the runs that exceeded the parsimony search limit and were cut off.

Due to space limitations, we present in Figures 2, 3, and 4 only a sample of our results. We show three different problem sizes, which we can think of as small, medium, and large. For 37 genes, both distance- and parsimony-based methods (except MPME) yield FN of at least 10%—the low number of genes reduces the amount of phylogenetic information. For 120 genes, trees produced by parsimony-based methods and NJ using `a−IEBP`, `e−IEBP`, and `EDE` have FN at most 20% (10% for higher rate and 40 taxa), and outperform NJ(INV) and NJ(BP) by a large margin when the amount of evolution is high. While MPME usually produces the most accurate trees among the parsimony-based methods, it is considerably slower than MPBE; indeed, we expect its performance on larger datasets is time-limited—had we given it more time to run, it would have surpassed the other MP-based methods easily. With 37 genes, increasing the rate of evolution improves the accuracy of MPME, but worsens that of MPBE and MPBE-2, whereas all three methods improve in accuracy for larger evolutionary rates with 120 genes.

NJ(EDE) is clearly the most accurate distance-based method, and this difference is most noticeable as the dataset gets close to saturation. Furthermore, it is also one of the fastest (calculating both `a-IEBP` and `e-IEBP` distances is more expensive). Also, NJ(EDE) is competitive with both MPBE and MPBE2, although mostly not as accurate as MPME (except for the inversion only scenario, and even then only for datasets which are far from saturated). Of all the methods we studied, MPME is the most accurate: it behaves well at all rates and is much better at high rates. Our results suggest that using an encoding that attempts to capture more details about the gene order (like MPME) preserves useful phylogenetic information that a parsimony-based search (with sufficient time) can put to good use. The choice between the best two methods (NJ(EDE) and MPME) may thus be dictated by running time rather than accuracy concerns: while NJ(EDE) is very fast and thus always usable, MPME will be too computationally expensive to use for some datasets.

5 Conclusion

We have introduced two new encoding methods for gene-order data and compared them to a previous encoding method (MPBE) and to NJ analyses based on various estimates of the true evolutionary distance (`EDE`, `a-IEBP`, and `e-IEBP`). MPME and NJ(EDE) are clearly the best two choices in our study, returning much more accurate trees than the other methods. Furthermore, MPME almost always outperforms NJ(EDE). However, while the advantage gained by MPME is significant, MPME is also the slowest of the methods we studied. An important direction for future research is thus to develop new heuristics that are as accurate as MPME, yet easy to implement for practical use.

Acknowledgement

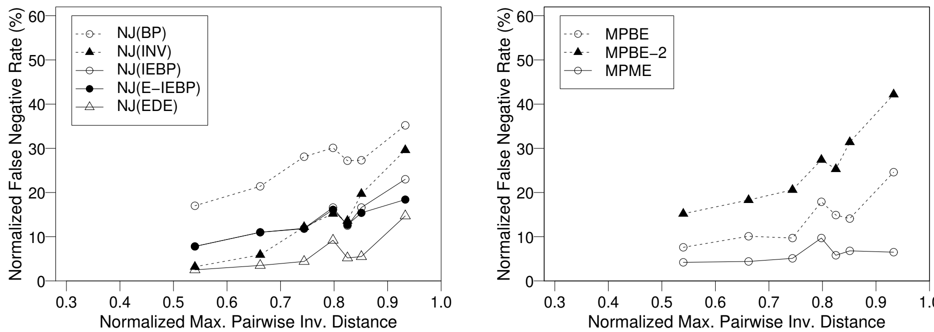
This work was supported in part by the National Science Foundation under grants to R.K. Jansen (DEB 99-82092), to B.M.E. Moret (ACI 00-81404), and to L.A. Raubeson (DEB 00-75700), and by the David and Lucile Packard Foundation to T. Warnow.

References

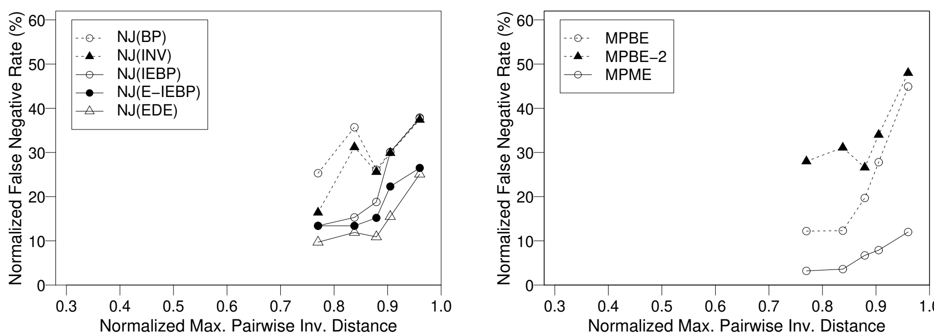
1. A. Rokas and P. W. H. Holland. Rare genomic changes as a tool for phylogenetics. *Trends in Ecology and Evolution*, 15:454–459, 2000.
2. S. Mathews and M. J. Donoghue. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science*, 286:947–950, 1999.
3. L.A. Raubeson and R.K. Jansen. Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. *Science*, 255:1697–1699, 1992.
4. R.G. Olmstead and J.D. Palmer. Chloroplast DNA systematics: a review of methods and data analysis. *Amer. J. Bot.*, 81:1205–1224, 1994.
5. C. Seoighe et al. Prevalence of small inversions in yeast gene order evolution. *Proc. Natl.*

Acad. Sci. USA, 97:14433–14437, 2000.

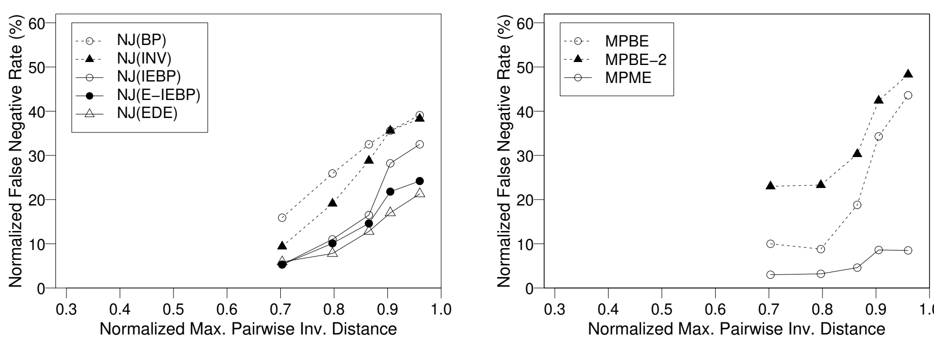
6. M. Blanchette, M. Kunisawa, and D. Sankoff. Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.*, 49:193–203, 1999.
7. J.D. Palmer. Chloroplast and mitochondrial genome evolution in land plants. In R. Herrmann, editor, *Cell Organelles*, pages 99–133. Wein, 1992.
8. J.H. Nadeau and B.A. Taylor. Lengths of chromosome segments conserved since divergence of man and mouse. *Proc. Nat'l Acad. Sci. USA*, 81:814–818, 1984.
9. L.-S. Wang and T. Warnow. Estimating true evolutionary distances between genomes. In *Proc. 33th Annual ACM Symp. on Theory of Comp. (STOC 2001).* ACM Press, 2001.
10. N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. & Evol.*, 4:406–425, 1987.
11. T. Warnow. Some combinatorial problems in phylogenetics. In *Proc. Int'l Colloq. Combinatorics & Graph Theory*, Balatonlelle, Hungary, 1996.
12. K. Atteson. The performance of the neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25(2/3):251–278, 1999.
13. D. Huson, S. Nettles, K. Rice, T. Warnow, and S. Yooseph. The hybrid tree reconstruction method. *J. Experimental Algorithmics*, 4:178–189, 1999. http://www.jea.acm.org/.
14. D. Swofford, G. Olson, P. Waddell, and D. Hillis. Phylogenetic inference. In D. Hillis, C. Moritz, and B. Mable, editors, *Molecular Systematics*. Sinauer Assoc. Inc., 1996.
15. D. Sankoff and M. Blanchette. Multiple genome rearrangement and breakpoint phylogeny. *J. Comp. Biol.*, 5:555–570, 1998.
16. I. Pe'er and R. Shamir. The median problems for breakpoints are NP-complete. *Elec. Colloq. on Comput. Complexity*, 71, 1998.
17. B.M.E. Moret, S.K. Wyman, D.A. Bader, T. Warnow, and M. Yan. A new implementation and detailed study of breakpoint analysis. In *Proc. 6th Pacific Symp. Biocomputing PSB 2001*, pages 583–594. World Scientific Pub., 2001.
18. B.M.E. Moret, L.-S. Wang, T. Warnow, and S. Wyman. New approaches for reconstructing phylogenies based on gene order. In *Proc. 9th Intl. Conf. on Intel. Sys. for Mol. Bio. (ISMB 2001)*, pages S165–S173, 2001. In *Bioinformatics* 17.
19. B.M.E. Moret, J. Tang, L.-S. Wang, T. Warnow, and S. Wyman. New approaches for reconstructing phylogenies based on gene order. *J. Comput. Syst. Sci.*, 2001. to appear.
20. D.A. Bader, B.M.E. Moret, and M. Yan. A fast linear-time algorithm for inversion distance with an experimental comparison. *J. Comput. Biol.*, 8(5):483–491, 2001.
21. L.-S. Wang. Improving the accuracy of evolutionary distances between genomes. In *Proc. 1st Workshop Algs. in Bioinformatics WABI'01*, pages 176–190. Springer Verlag. LNCS 2149.
22. D. Sankoff and J.H. Nadeau, editors. *Comparative Genomics : Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*. Kluwer Academic Pubs., 2000.
23. M.E. Cosner, R.K. Jansen, B.M.E. Moret, L.A. Raubeson, L. Wang, T. Warnow, and S.K. 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.
24. D. Bryant. A lower bound for the breakpoint phylogeny problem. In *Proc. 11th Ann. Symp. Combinatorial Pattern Matching CPM'00*, pages 235–247. Springer Verlag, 2000.
25. D. Swofford. *PAUP* 4.0*. Sinauer Associates Inc, 2001.



$(\alpha, \beta) = (0, 0)$

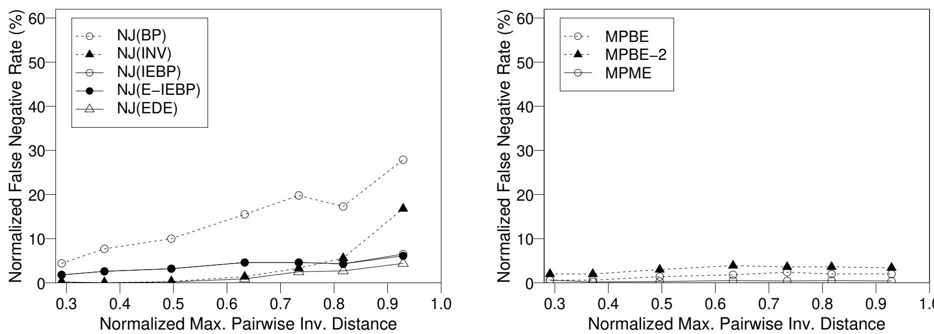


$(\alpha, \beta) = (\frac{1}{2}, \frac{1}{2})$

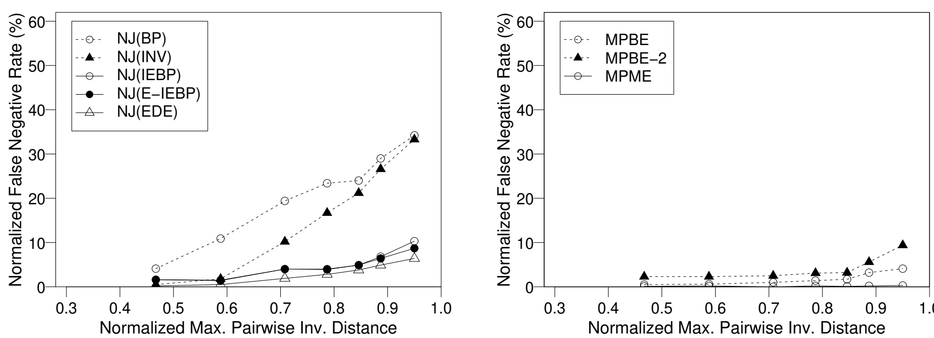


$(\alpha, \beta) = (\frac{1}{3}, \frac{1}{3})$

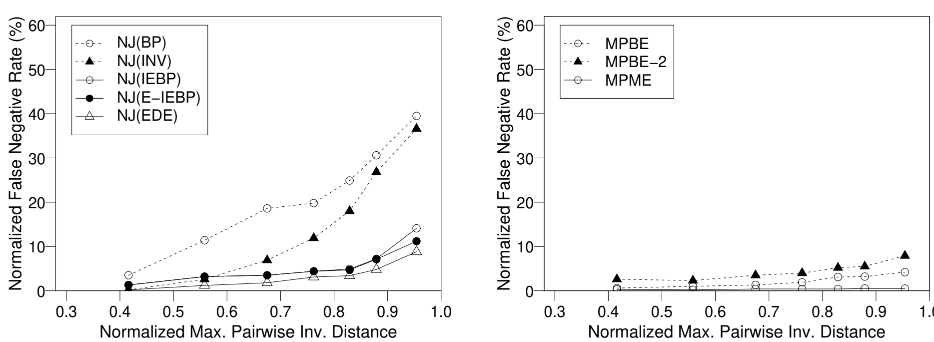
Figure 2: Topological accuracy of phylogenetic methods on problems with 37 genes and 40 taxa. The x-axis is normalized by the number of genes, the highest inversion distance two gene orders can have. Our plots result from binning the values into range of evolutionary distances (maximum pairwise inversion distance in the dataset) and plotting the average value for each bin. See Section 1.2 for the definition of the model weights (α, β).



$(\alpha, \beta) = (0, 0)$

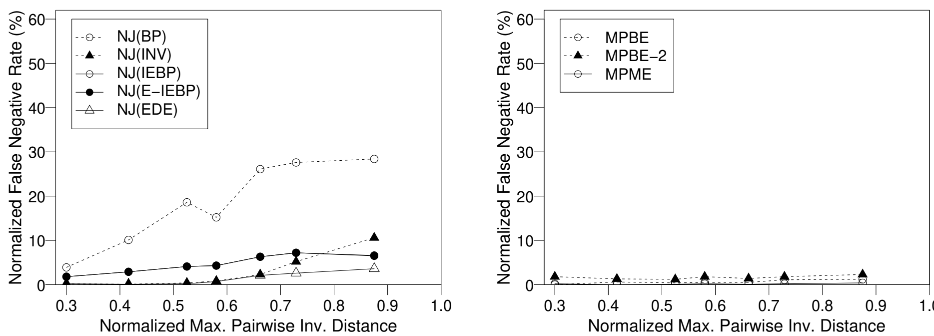


$(\alpha, \beta) = (\frac{1}{2}, \frac{1}{2})$

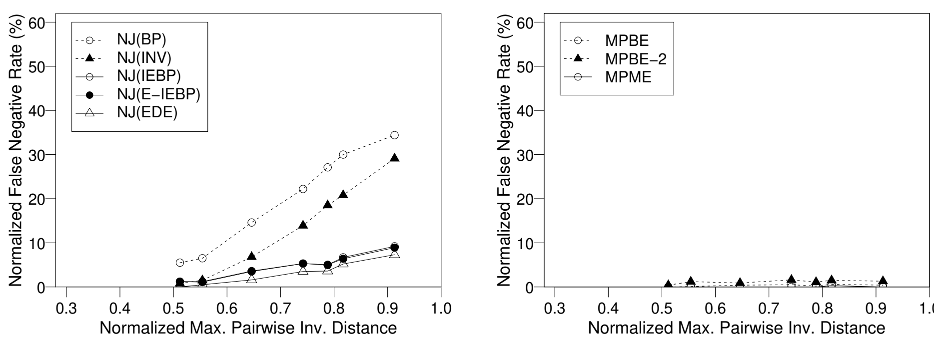


$(\alpha, \beta) = (\frac{1}{3}, \frac{1}{3})$

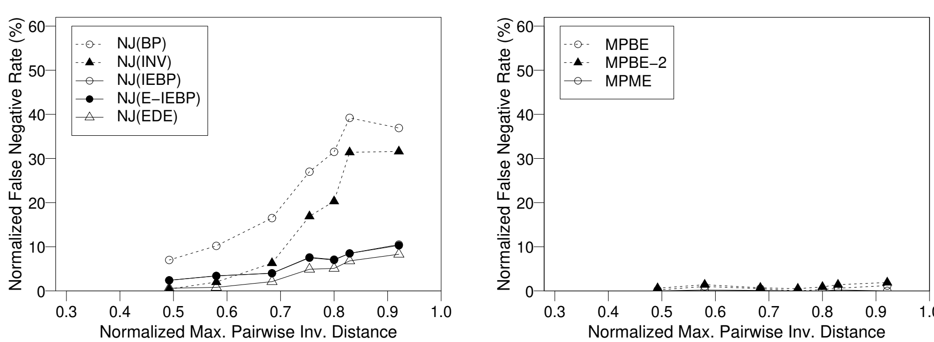
Figure 3: Topological accuracy of phylogenetic methods on problems with 120 genes and 40 taxa. See Section 1.2 for the definition of the model weights (α, β).



$(\alpha, \beta) = (0, 0)$



$(\alpha, \beta) = (\frac{1}{2}, \frac{1}{2})$



$(\alpha, \beta) = (\frac{1}{3}, \frac{1}{3})$

Figure 4: Topological accuracy of phylogenetic methods on problems with 120 genes and 160 taxa. See Section 1.2 for the definition of the model weights (α, β).