J Mol Evol (2006) xx:1–12
DOI: 10.1007/s00239-005-0216-y

JOURNAL OF MOLECULAR EVOLUTION

Distance-Based Genome Rearrangement Phylogeny

Li-San Wang,[1] Tandy Warnow,[2] Bernard M. E. Moret,[3] Robert K. Jansen,[4] Linda A. Raubeson[5]

[1] Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA
[2] Department of Computer Sciences, University of Texas at Austin, Austin, TX 78712, USA
[3] Department of Computer Science, University of New Mexico, Albuquerque, NM 87131, USA
[4] Section of Integrative Biology and Institute of Cellular and Molecular Biology, University of Texas at Austin, Austin, TX 78712, USA
[5] Department of Biological Sciences, Central Washington University, Ellensburg, WA 98926, USA

Received: 13 September 2005 / Accepted: 26 June 2006 [*Reviewing Editor:* Dr. Martin Kreitman]

Abstract. Evolution operates on whole genomes through direct rearrangements of genes, such as inversions, transpositions, and inverted transpositions, as well as through operations, such as duplications, losses, and transfers, that also affect the gene content of the genomes. Because these events are rare relative to nucleotide substitutions, gene order data offer the possibility of resolving ancient branches in the tree of life; the combination of gene order data with sequence data also has the potential to provide more robust phylogenetic reconstructions, since each can elucidate evolution at different time scales. Distance corrections greatly improve the accuracy of phylogeny reconstructions from DNA sequences, enabling distance-based methods to approach the accuracy of the more elaborate methods based on parsimony or likelihood at a fraction of the computational cost. This paper focuses on developing distance correction methods for phylogeny reconstruction from whole genomes. The main question we investigate is how to estimate evolutionary histories from whole genomes with equal gene content, and we present a technique, the empirically derived estimator (EDE), that we have developed for this purpose. We study the use of EDE on whole genomes with identical gene content, and we explore the accuracy of phylogenies inferred using EDE with the neighbor joining and minimum evolution methods under a wide range of model conditions. Our study shows that tree reconstruction under these two methods is much more accurate when based on EDE distances than when based on other distances previously suggested for whole genomes.

Key words: Distance-based methods — Genome rearrangements — Neighbor joining — Fast ME — Nadeau-Taylor model — Breakpoint — Inversion

Correspondence to: Li-San Wang, 203 Goddard Laboratories, 415 South University Avenue, Philadelphia, PA 19104, USA; *email:* lswang@mail.med.upenn.edu

Introduction

Success in phylogeny reconstruction depends on the qualities of the underlying data and the accuracy of the methods of analysis. Gene order changes are attractive characters for phylogeny reconstruction because these events are rare, and thus they have the potential to provide information about ancient events in evolutionary history (Rokas and Holland 2000). Gene order changes in mitochondrial (Boore and Brown 1998; Boore 1999) or chloroplast (reviewed by Downie and Palmer 1992; Raubeson and Jansen 2005) genomes have been utilized as phylogenetic characters. In most cases, a small number of changes, perhaps only one, have been characterized and the phylogenetic implications of the changes determined. For example, a single 32-kb inversion in the chloroplast genome supported lycopsids as the basal lineage of vascular plants (Raubeson and Jansen 1992) and animal mitochondrial gene orders support the

monophyly of Arthropoda (Boore et al. 1995). However, as more genomic information becomes available, the need for computational methods to analyze gene order data will increase.

There is a variety of methods for reconstructing phylogenies, such as distance-based methods, maximum parsimony, and maximum likelihood. However, except for the distance-based approaches, all are computationally intensive. When analyzing gene order data phylogenetically these methods are even more computationally intensive than the corresponding problems in sequence-based phylogenetics. Indeed, the elementary problem of computing the minimum number of events needed to transform one gene order into another (a trivial problem for sequences) is solved only for inversions—the best solution of this problem (Bader et al. 2001) was the culmination of 10 years of research and several combinatorial breakthroughs. Several new methods have been developed for estimating phylogenies from gene order data; see Moret et al. (2005) and Moret and Warnow (2005) for detailed surveys. In this paper, we focus on distance-based methods for phylogeny reconstruction, largely because they are fast and have running times that grow only in polynomial time relative to the number of taxa and genes involved.

Distance-based phylogeny reconstruction involves two steps: a matrix of pairwise distances is computed, and then a tree is constructed based on the distance matrix. The two most widely used distance-based tree reconstruction methods are neighbor joining (NJ) and minimum evolution (ME). For these methods to produce highly accurate estimates of the evolutionary tree, the distance matrices must be close to the matrix of true evolutionary distances; this means that the estimated pairwise distance between any two given taxa should be as close as possible to the number of evolutionary events along the tree path that connects the two taxa. Since that distance cannot be computed directly, statistical techniques, based on the assumed model of evolution, are used. For example, in a phylogenetic analysis of DNA sequences under the Jukes-Cantor model, first the p-distance (i.e., the normalized Hamming distance) matrix is computed, and then this distance is "corrected" through the use of the Jukes-Cantor distance correction; similar, but more complicated, calculations exist for more complex models of evolution. Such corrections are routine in the computation of pairwise distances between DNA sequences; assuming that the model is well chosen, the corrections ensure statistical consistency of the distance method and clearly improve the accuracy of trees estimated using distance-based methods (Sourdis and Krimbas 1987; Swofford et al. 1996).

The first study that used a distance-based method to reconstruct phylogenies from gene orders was published by Blanchette et al. (1999). They conducted a phylogenetic analysis of six metazoan groups using NJ applied to a matrix of breakpoint distances (BP) defined on a set of mitochondrial genomes. The breakpoint distance (Blanchette et al. 1997) is the number of gene adjacencies that are present in one gene order but not in the other—a measure of the dissimilarity between two gene orders rather than a measure of the amount of evolution between the two genomes. An alternative measure is the inversion (INV) distance, which is the minimum number of inversions needed to transform one gene order into the other. However, as we will show, breakpoint and inversion distances are not very accurate, largely because they can seriously underestimate evolutionary distances, especially when there is a large number of evolutionary events. Therefore, the challenge is to develop a method for correcting the inversion distance to produce a more accurate estimate of the true evolutionary distance.

In this paper, we investigate a statistically based technique for estimating evolutionary distances between genomes called the "empirically derived estimator" (EDE) (Moret et al. 2001). This technique corrects the minimum inversion distance between two genomes, thus obtaining a more accurate estimate of the number of events in their evolutionary history. We study the performance of BP, INV, and EDE distances in simulations using both NJ and ME tree reconstruction methods under a variety of model settings. Our study establishes that phylogenies reconstructed using EDE distances are much more accurate than phylogenies reconstructed using either BP or INV distances, for both NJ and ME tree reconstruction methods. We also find that a new ME implementation (FastME) outperforms NJ on these data. As a result, we find that FastME(EDE) provides the most accurate reconstruction of gene order phylogenies within the conditions tested in our experiments.

Methods

We studied phylogeny reconstruction on whole genomes under the simplifying assumption that all genomes have exactly one copy of each gene and that they have exactly the same gene content. We focused on circular genomes, although our methods can also be applied to linear genomes. We represented each genome by an ordering of genes, and used the sign to indicate the strand on which the gene is located. For circular genomes, it does not matter which gene we begin with, nor which strand is positive or negative. Thus, the circular genome given by the linear ordering (1,2,3,4,5,6,7) is equivalent to the linear orderings (2,3,4,5,6,7,1) and (−2,−1, −7,−6,−5,−4,−3) as well as numerous other permutations. Each genome comprises one complex character in the phylogenetic analysis, which is represented by the signed gene order.

Because we studied the case where gene content remains constant, we did not consider events such as duplications, insertions, or deletions, but restricted ourselves to inversions (also called "reversals") and transpositions. Inversions operate by reversing a

segment at a location within the genome; thus, the order and strandedness of the genes involved change. For example, if we apply an inversion on the segment 2,3 in (1,2,3,4,5,6,7), we obtain (1,−3,−2,4,5,6,7). In contrast, a transposition has the effect of moving a segment from between two genes to another location (between two other genes). This can occur without changing the order or strandedness of the genes within the segment (transposition) or with the reversal of order and strandedness of the moved segment (inverted transposition). For example, a change of (1,2,3,4,5,6,7) to (1,4,2,3,5,6,7) could be explained by a transposition of the segment 2,3 from between 1 and 4 to between 4 and 5. An inverted transposition of the same segment to the same location would result in (1,4,−3,−2,5,6,7). The term *transposition* here is intended only to describe this resulting pattern of change; the change does not necessarily result from the movement of genomic segments by transposable elements. Inverted transpositions are considered distinct from transpositions for computational rather than biological reasons. Any transposition event could be explained alternatively by three inversions, whereas an inverted transposition can be explained by two inversions.

We used simulation studies to evaluate the accuracy of EDE relative to other distances (INV and BP) for estimating evolutionary distances between genomes affected by inversions, transpositions, and inverted transpositions. The details of our methodology are described below; here we give a brief outline of the study design. We generated model trees under either a uniform distribution or under the birth-death model. We simulated the evolution of genomes down the different trees using the GNT model (the generalized Nadeau-Taylor, defined below), thus producing synthetic data (genomes) at the tips (leaves) of the trees. We then computed distances between these genomes, using the various estimators (BP, INV, EDE). Each distance matrix was analyzed using NJ and FastME (Desper and Gascuel 2002), producing trees for each matrix. Accuracy of the resulting trees was measured relative to the model tree using false-negative and false-positive rates. We explored performance on datasets containing either 40 or 160 genomes and either 37 or 120 genes (typical values for mitochondrial and chloroplast genomes, respectively), for a variety of settings of the relative probabilities of the three types of events (inversions, transpositions, and inverted transpositions). Each of these aspects of the study is discussed in more detail below.

Model Trees

As the basis of our experiments we produced model trees; a model tree consists of a rooted tree topology and branch lengths, where the branch length indicates the expected number of evolutionary events (inversions, transpositions, and inverted transpositions) on the branch. The relative probabilities of these different events are given by other parameters in the GNT model and are defined below.

There are several ways to produce trees with branch lengths, and the choice of technique can influence the relative performance of phylogenetic reconstruction methods. We therefore studied two techniques for generating random trees: (1) birth-death trees, with rate variation across lineages to deviate the trees away from ultrametricity (i.e., away from clocklike behavior), and (2) uniform distribution on tree topologies, with branch lengths drawn from a distribution. Note that birth-death trees are more balanced (in the sense of Heard 1992) than trees drawn from a uniform distribution.

Birth-death trees. We generated birth-death model trees through the use of the r8s software (Sanderson 2003), which produces a rooted binary tree along with branch lengths; these branch lengths are ultrametric (i.e., they obey a molecular clock). We now describe how we modified the model tree so that it did not fit the molecular clock. First, we selected a parameter c. Then, for each branch, we picked a random number s (called the "stretch") where $\ln(s)$ was drawn uniformly from the interval $[-c,c]$; the length of the branch was then multiplied by s. Thus, each branch length is multiplied by a potentially different random number. This process yields a model tree which is not ultrametric; furthermore, by varying the parameter c we can vary the deviation from the molecular clock. For $c = 8.8$ and 26.1, the expected deviation $E[s]$ from the molecular clock is 2 and 4, respectively. Finally, we then rescale all branch lengths (by multiplying all lengths by the same fixed value) in order to achieve a target evolutionary diameter D for the tree, where the "evolutionary diameter" is the maximum pairwise path length between taxa in the resulting tree. The target diameters were drawn from $0.1n$, $0.2n$, $0.4n$, $0.8n$, $1.6n$, $2.4n$, and $3.2n$, where n is the number of genes; these resulted in datasets that have maximum normalized pairwise inversion distances ranging from approximately 0.1 up to almost 1, which is the maximum possible.

Uniform tree topologies. Under this approach, we selected tree topologies from the uniform distribution on binary, unrooted trees with leaves labeled by 1, 2, ... , m, for $m = 40$ or 160 (the two tree sizes we investigated). We assigned branch lengths to each tree using the following three steps: (1) we picked a target diameter D, drawn from $0.1n$, $0.2n$, $0.4n$, $0.8n$, $1.6n$, $2.4n$, and $3.2n$, where n is the number of genes; (2) we assigned an initial length for each branch by drawing integers randomly between 1 and 15; and (3) we then multiplied all branch lengths by the same constant in order to obtain the selected target diameter. The use of small target diameters defines model trees that produce simulated datasets with small maximum pairwise inversion distances, while the use of large target diameters defines model trees that produce simulated datasets with maximum pairwise inversion distances close to n, the maximum possible.

The Generalized Nadeau-Taylor Model

We simulated genome evolution on the trees using the GNT model. Under this model, any inversion is equally likely to occur, regardless of where the two endpoints are; the same assumption of uniform probability applies to the set of all transpositions and to the set of all inverted transpositions. Each model tree thus has parameters w_I , w_T , and w_{IT} , where w_I is the probability that a rearrangement event is an inversion, w_T is the probability that a rearrangement event is a transposition, and w_{IT} is the probability that a rearrangement event is an inverted transposition. Because we assumed that all events are of these three types, $w_I + w_T + w_{IT} = 1$, and so there are two free parameters. Given a model tree, we let $X(e)$ be the random variable for the number of evolutionary events that takes place on branch e. We assumed that $X(e)$ is a Poisson random variable with mean λ_e; hence, λ_e is the length of the branch e and indicates the expected number of events that will occur on branch e. We also assume that events on one branch are independent of the events on other branches. Thus, in the GNT model the number of parameters is proportional to the number of genomes (i.e., taxa): the length λ_e of each branch e and the triplet w_I, w_T , w_{IT}. We let GNT(w_I,w_T ,w_{IT}) denote the set of model trees with the triplet w_I, w_T, w_{IT}.

We considered three models:

- GNT(1,0,0) (inversion only),
- GNT(0.5,0.25,0.25) (half inversions, half transpositions), and
- GNT(0,0.5,0.5) (transposition only).

It would seem reasonable that EDE would perform well under the inversion-only model because it is a distance correction based on an inversion-only simulation; similarly, it is reasonable to presume that INV should perform well under inversion-only scenarios, though perhaps not as well as EDE. However, it still remains to be seen whether EDE performs well for phylogeny estimation in scenarios other than inversion-only. The inclusion of these two other

models (one with half inversions and one with no inversions) was meant to explicitly test the robustness of EDE.

Distances Between Genomes

We compared three genomic distances: BP, INV, and EDE. For each distance, we tested its accuracy in estimating true evolutionary distance and the accuracy of tree estimation based on the distance.

Breakpoint distance. The first measure proposed for the estimation of evolutionary rearrangement distance between genomes was the breakpoint distance (Blanchette et al. 1997). A breakpoint occurs between gene g and gene g′ in genome G′ with respect to genome G if g is not followed immediately by g′ in G. As an example, consider the comparison of circular genomes G = (1,2,−3,4,5,6,7) and G′ = (1,2,3,−7,−6,−5,−4). There is a breakpoint between 3 and 5 in G′, since 3 is not followed by 5 in G, but there is no breakpoint between 5 and 4 in G′ since G can be equivalently written as (1,−7,−6,−5,−4,3,−2). The breakpoint distance between two genomes is the number of breakpoints in one genome with respect to the other, which is clearly symmetric. In this example, the breakpoint distance between G and G′ is 3.

Inversion distance. The inversion distance between genome G and genome G′ is the minimum number of inversions needed to transform G into G′ (or vice versa, as it is symmetric). For example, if G = (1,2,3,4,5,6,7) and G′ = (1,−4,−3,−2,5, −7,−6), then the inversion distance between G and G′ is 2, since we can transform G into G′ in two inversions, but not in one. The first polynomial-time algorithm for computing this distance was obtained by Hannenhalli and Pevzner (1995) and later improved by Bader et al. (2001) (the latter obtained an optimal linear time algorithm).

Our statistically based distance estimator, EDE. We have developed a statistical technique, called EDE, for correcting inversion distances. EDE (Moret et al. 2001) is our "empirically derived estimator" because we developed this technique based on data obtained in simulation. The basic structure of the EDE distance is described here (a more detailed derivation is given in the Appendix). Suppose we have a function $f(x)$ that is the expected normalized inversion distance produced by nx random inversions, where n is the number of genes. Then, given two genomes, to estimate the actual number of inversions that took place between them we do the following: compute their inversion distance, then use the values computed for the function $f(x)$ to look up the number of inversions that would have produced that inversion distance (refer to the Appendix for more details). EDE was derived on the basis of a simulation study under an inversion-only evolutionary model, in which the number of genes ranged from 20 to 160 per genome. Thus, we can expect the estimated evolutionary distance to be accurate if the evolutionary process is inversion only and if the genomes have between 20 and 160 genes (typical values are 37 for mitochondria and 120 for chloroplasts). But what if the evolutionary process is not inversion only, and may in fact consist only of transpositions? Will distances estimated by EDE still be highly accurate? And will phylogenies obtained from EDE distance matrices be highly accurate, or will it be better to use some other distance estimation technique, such as the breakpoint distance? The study we present in this paper explores these questions using simulations under a wide range of model conditions.

Phylogeny Reconstruction Techniques

In our study, we used two different tree reconstruction methods, NJ (Saitou and Nei 1987) and a fast implementation of the ME method, FastME (Desper and Gascuel 2002). Each of these two methods is applied to distance matrices obtained using the BP, INV, and EDE distance estimation techniques. We used PAUP* (Swofford 2001) to compute the NJ trees. We downloaded the source code of FastME from the authors' web site and compiled it using GCC on Debian Linux. Since running time was not the criterion by which we compared methods, we were not concerned with obtaining the most efficient implementations. However, both methods were very fast even on large datasets (160 genomes on 120 genes each).

Performance Criteria

These trees were then compared to the true tree (the model tree minus the zero-event branches) for topological accuracy. A reconstructed tree can have two types of errors: false positives, which are non-zero-length reconstructed branches that are not present in the true tree, and false negatives, which are non-zero-length branches in the true tree that fail to appear in the reconstructed tree. When both the true tree and the inferred tree are binary, then the number of false positives and the number of false negatives are equal; however, in our case, since the model tree may have branches without any changes, the true tree may not be fully resolved. Hence, we will report both types of errors.

The false-negative and false-positive rates were obtained by dividing the number of false negatives and false positives, respectively, by $m - 3$ (the number of internal branches in a binary tree on m taxa). Our experiments examined performance under a range of evolutionary rates, and the performance under higher rates of evolution allowed us to evaluate whether or not tree reconstruction can be done accurately when every branch is expected to have changes on it.

Experiments

For each experimental setting we ran the simulation 50 times. We then computed the average diameter (maximum pairwise inversion distance between any two genomes) of the 50 replicates and the false-negative and false-positive rates of trees produced by the different distance-based methods. We report the topological accuracy of the trees we obtained, and we explore how that accuracy is impacted by the distance estimator used, but also by the different parameters of the model tree, specifically focusing on the number of genes, number of genomes, maximum pairwise inversion distance, and relative probability of the different events (inversion, transposition, and inverted transposition).

Results

Accuracy of the True Evolutionary Distance Estimators

Compared to EDE, BP and INV are highly biased when the number of rearrangement events is large (Fig. 1 and Supplemental Figs. 1 and 2). EDE maintains a more or less linear relationship with true evolutionary distance at even very large numbers of events, whereas BP and INV lose their ability to estimate the number of events reasonably well when this number approaches or exceeds the number of genes in the dataset. There is never a linear relationship between BP and the actual number of events. However, under an inversion-only scenario, both INV and EDE scale almost linearly when the actual

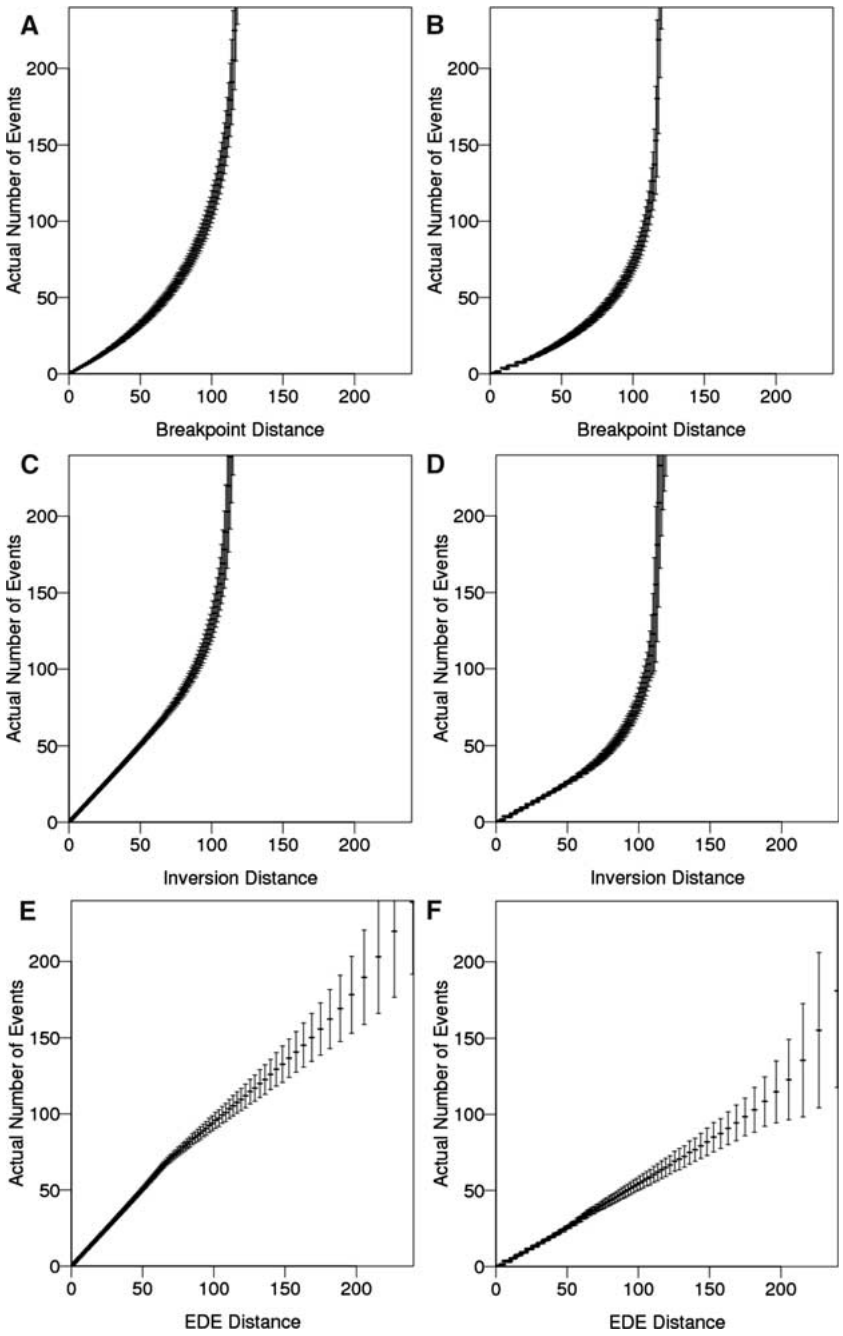

Fig. 1. The distribution of genomic distances on 120-gene genomes under the generalized Nadeau-Taylor model. The *x*-axis is the measured distance, and the *y*-axis is the actual number of rearrangement events. For each vertical line, the middle point is the mean, and the top and bottom tips of the line represent one standard deviation away from the mean.

number of inversions is below some threshold, with EDE having a higher threshold to which it scales linearly.

Under transposition-only evolution, none of the estimators accurately measures true evolutionary distance (Fig. 1 and Supplemental Figs. 1 and 2). However, EDE maintains a scalar linear relationship with the actual number of events, although it consistently overestimates the true evolutionary distance. This overestimation results from the fact that each inverted transposition would require two inversions to explain it and each transposition, three inversions. However, inaccuracy in distances may not lead to inaccuracy in the trees that are constructed using those distances, provided that the estimated distances are just scalar multiples of the true evolutionary distances.

Performance of the Estimators in Tree Reconstruction

At lower amounts of evolution, all estimators perform similarly under all conditions. However, distinctions in performance become noticeable as the rate of evolution (and hence the evolutionary diameter) increases. In general, both NJ and FastME perform better on EDE than on INV, and better on INV than on BP. Within each method (NJ and FastME; see Fig. 2 and Supplemental Figs. 3 and 4),

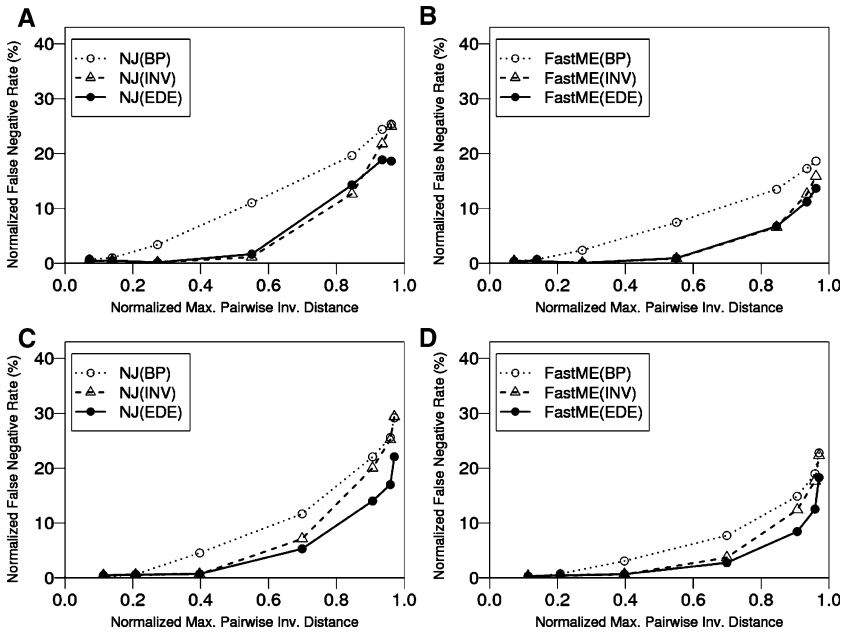

Fig. 2. Simulation study of the false-negative rates of neighbor-joining (NJ) and FastME using the three genomic distances on birth-death trees with 160 genomes containing 120 genes in each genome. The model trees are birth-death trees generated derived from the r8s software and have a moderate deviation from the molecular clock (expected stretch 2). The *x*-axis is the normalized diameter (maximum inversion distance between all pairs of genomes) of the dataset, and the *y*-axis is the false-negative rate of the inferred tree.

EDE-based trees and INV-based trees have lower false-negative rates than BP-based trees. NJ(EDE) and FastME(EDE) as well as NJ(INV) and FastME(INV) have similar false-negative rates until the evolutionary rate is high, when EDE-based trees are more accurate than INV-based trees. The gap is larger when the trees are drawn from a uniform distribution (data not shown), where the error rate of NJ(BP) can be three times that of FastME(EDE) when the evolutionary rate is very high, presumably due to the better balance of birth-death trees. This pattern—NJ(EDE) > NJ(INV) > NJ(BP) and FastME(EDE) > FastME(INV) > FastME(BP), although data are not always shown for BP—is maintained under all experimental settings (Figs. 2–4 and Supplemental Figs. 3–9): number of genes, number of genomes, GNT model, and method of model tree generation (uniform or birth-death). The one exception occurs in the case of the FastME analysis of 160 genomes under the birth-death model for 37 genes (Supplemental Fig. 5), where FastME(EDE) and FastME(INV) perform about equally well over the entire range of distances. The experiments showed that analyses based on BP distances consistently produce trees inferior to analyses based on INV or EDE distances. We therefore focus our attention on comparisons involving either inversion or EDE distances, and do not discuss performance under breakpoint distances for the remainder of this paper.

Performance Under Different GNT Models

The parameter values for the GNT model, whether all inversions (1,0,0), all transpositions (0,0.5,0.5), or a mixture (0.5, 0.25, 0.25), do not affect the relative ranking of distance estimator performance described in the previous section (Fig. 3 and Supplemental Figs. 5 and 6). When transpositions are included, the results of the different methods are more similar and EDE no longer dominates the other methods quite as significantly (especially not at the higher rates of evolution). However, even here we maintain the relative performance EDE > INV. It is interesting to note that EDE operates as well under the mixed model as under the transposition-only model; this is surprising since EDE is based on an inversion-only assumption, and the transposition-only model deviates the most from this assumption.

Performance of the Phylogeny Reconstruction Algorithms—NJ and FastME

FastME is superior to NJ within each distance measure and setting (Figs. 2 and 3 and Supplemental Figs. 3–9); that is, FastME(EDE) performs better than NJ(EDE) within each experiment and FastME(INV) performs better than NJ(INV). Overall, the best method is FastME(EDE).

Interestingly, in many (although not all) cases, FastME(INV) outperforms NJ(EDE), suggesting that the choice of phylogeny reconstruction technique has a bigger impact than the distance estimation technique under some circumstances. However, NJ (EDE) outperforms FastME(INV) on small datasets with high evolutionary rates (data not shown). When model trees are based on the uniform distribution, all 40 genome cases show NJ(EDE) > FastME(INV); however, with birth-death model trees, NJ(EDE) improves on FastME(INV) for special cases: 120 genes, 40 genomes, under models that include transpositions (Supplemental Fig. 6).

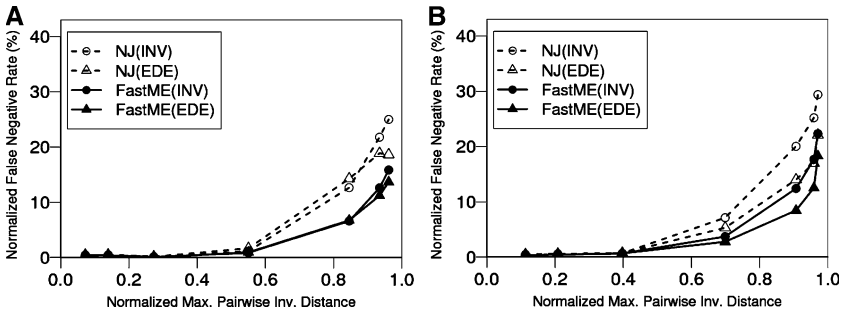

Fig. 3. Simulation study of the false-negative rate of distance-based tree reconstruction methods on birth-death trees with 120 genes in each genome. The model trees are birth-death trees derived from the r8s software and have a moderate deviation from the molecular clock (expected stretch 2). The *x*-axis is the normalized diameter (maximum inversion distance between all pairs of genomes) of the dataset, and the *y*-axis is the false-negative rate of the inferred tree.

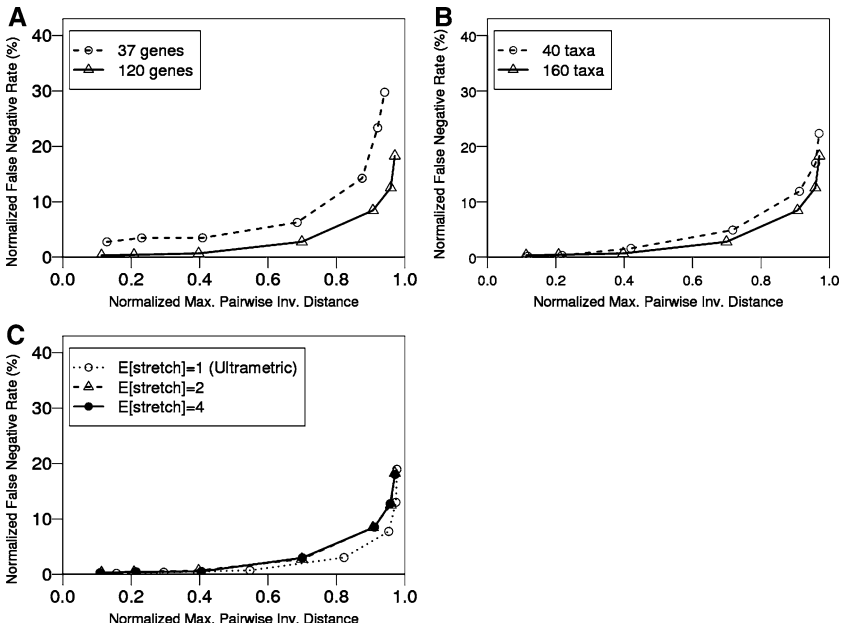

Fig. 4. Simulation study exploring how the false-negative rate of distance-based tree reconstruction methods is affected by (**a**) the number of genes in each genome, (**b**) the number of genomes, and (**c**) the deviation from ultrametricity. The model trees have 160 genomes and are birth-death trees generated using the r8s software with a moderate deviation from the molecular clock (expected stretch 2). The evolutionary model is GNT(0.5, 0.25, 0.25) (i.e, the "mixed" model, with equal probability of inversions and non-inversions). The *x*-axis is the normalized diameter (maximum inversion distance between all pairs of genomes) of the dataset, and the *y*-axis is the false-negative rate of the inferred tree.

Number of Genes

For every method, at every evolutionary diameter, phylogenies reconstructed on genomes containing 37 genes are less accurate than phylogenies reconstructed on genomes containing 120 genes (Fig. 4a and Supplemental Fig. 7). This difference in performance is consistent with the greater accuracy of the three distance estimators on genomes containing 120 genes compared to genomes containing 37 genes (data not shown). That is, distance estimation is more accurate when there are more genes, so phylogeny estimations on genomes with 120 genes are more accurate than phylogeny estimations on genomes with 37 genes.

Number of Genomes

For 120 genes (Fig. 4b and Supplemental Fig. 7), for all but the smallest evolutionary diameters, both NJ and FastME trees reconstructed on 160 taxa are more accurate than trees reconstructed on 40 taxa, and this relationship holds under all the models we examined. This is probably due to the fact that for a fixed evolutionary diameter, the average branch length will be smaller on 160-taxon model trees than on 40-taxon model trees (data not shown); therefore, trees with large diameters will be more easily estimated if they have more taxa.

Effect of the Model Tree

The accuracy of phylogeny reconstruction drops when the model tree deviates from ultrametricity (Fig. 4c and Supplemental Fig. 9), although EDE-based analyses are less affected than INV-based analyses. The best results were obtained by FastME(EDE) under uniform model tree generation. All but NJ(INV) perform better on uniform distribution model trees than on birth-death trees; we conjecture that the reason is simply a better spread of the pairwise distances within the predetermined range. Furthermore, the different methods vary less in their accuracy on birth-death trees

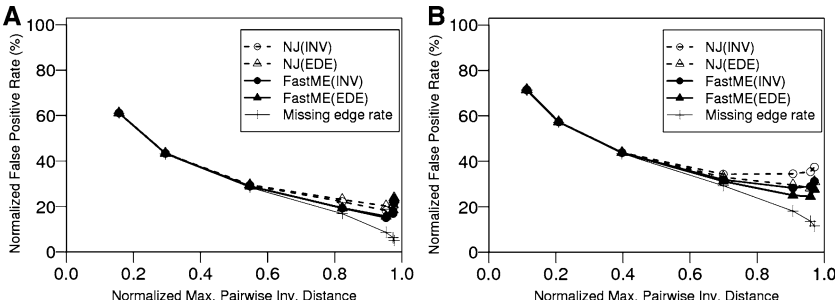

Fig. 5. Simulation study of the false-positive rate of distance-based tree reconstruction methods with 120 genes in each genome. The model trees are birth-death trees generated using the r8s software with (**a**) no deviation (ultrametric) and (**b**) a moderate deviation from the molecular clock (expected stretch 2). The evolutionary model is GNT(0.5, 0.25, 0.25) (i.e., the "mixed" model, with equal probability of inversions and noninversions). The *x*-axis is the normalized diameter (maximum inversion distance between all pairs of genomes) of the dataset, and the *y*-axis is the false-positive rate of the inferred tree. We also include the curve of "missing branch rates" for comparison. For this curve the *y*-axis is the percentage of internal branches in the model tree that are zero-event branches.

than on uniform trees. For all methods, the difference in performance between ultrametric trees (E[*s*] = 1) and those with a moderate deviation (E[*s*] = 2) from a molecular clock is large (Fig. 4c), although the difference between moderate deviation (E[*s*] = 2) and large deviation (E[*s*] = 4) from a molecular clock is insignificant (data not shown). A possible explanation is that for a given fixed diameter, the more the model tree deviates from the molecular clock, the larger the variance in branch lengths, which could make it harder to estimate. These experiments are also consistent with results of the study by Nakhleh et al. (2002), which examined the impact of deviation from a molecular clock on phylogeny estimation from DNA sequences.

False-Positive Rates

The false-positive rate indicates the percentage of the branches in the inferred tree that are not present in the true tree. When the true tree is not fully resolved (due to branches in the model tree in which no event occurs), if the estimation technique forces the inferred tree to be binary, the false-positive rate will necessarily be at least as large as the missing branch rate. Therefore, in our studies, we explored the false-positive rate and compared it against the theoretical minimum, which is the missing branch rate (Fig. 5).

For most of the range, all four methods (NJ and FastME on both INV and EDE distances) have an optimal or near-optimal false-positive rate, as reflected in the almost-exact match between their false-positive rates and the missing branch rate. However, NJ's false-positive rate is slightly worse than FastME's in our experiments.

Conclusions/Discussion

The simulation studies indicate that our new distance correction (EDE) produces more accurate estimations of evolutionary distances between gene orders than INV or BP, and that phylogenies estimated using EDE are more accurate than those estimated either INV or BP, whether one uses NJ or FastME. The improvement in accuracy is especially evident when the model has a high evolutionary rate, so that the dataset has a large maximum inversion distance. Our results mirror those reported previously for distance corrections for DNA sequence data (Saitou and Imanishi 1989; Rzhetsky and Sitnikova 1996).

As we have previously noted, the biggest improvement in using EDE comes when the dataset has a large diameter (maximum pairwise inversion distance). This parameter can vary significantly across real datasets, varying with the type of genomes and the divergence within the set of taxa in the analysis. For example, in the Campanulaceae dataset analyzed by Cosner et al. (2000), the diameter is on the lower end, mainly because of the close phylogenetic relationships the among the taxa. In another dataset with more than 40 animal mitochondrial genomes (Boore 1999), many of the pairwise distances have reached or are very close to the maximum possible for that dataset.

We have also shown that statistically based estimations of evolutionary distances can be quite robust to some model violations, making phylogenetic reconstruction much more accurate, especially when the amount of evolutionary change is high. Perhaps the most significant indicator of the difficulty of phylogenetic reconstruction for a dataset is its evolutionary diameter: if the amount of change (diameter) is low, all methods will give a good estimate of the tree even if the distance estimation is based on incorrect assumptions, but for the largest diameters (highest amounts of change), only FastME(EDE) is reliably accurate.

We have already noted that reconstructions of trees we obtain can have a high false-positive rate due to the high incidence of zero-event branches in the model tree (and hence low resolution in the true tree).

Determining which branches in the reconstructed tree are valid, and which are not, is a general problem facing distance-based phylogenetic analysis. In DNA sequences, bootstrapping and other techniques can be used to assess the confidence in a given branch and, so, potentially identify the false-positive branches. In gene order phylogeny it is not possible to perform bootstrapping, since there is only one character. For distance-based methods, the interior branch length test of Rzhetsky and Nei (1992) may be used to test if the length of a branch is significantly different from zero, though more research is warranted on the power of the test for gene order data.

Several factors can affect the accuracy of our method (and most phylogenetic methods in general) when studying genome rearrangements. One positive factor is the number of genes: our studies clearly show that trees based on 120 genes are more accurate than trees based on 37 genes. However, another interesting question is the impact of the number of taxa on the resultant analysis. In our opinion, this issue (which is related to the taxon sampling question) is still a subject for debate. For example, in some studies, estimated phylogenies are more accurate when taxonomic sampling is increased (Zwickl and Hillis 2002), but in others the accuracy can decrease (Nakhleh et al. 2002a). Theoretical investigations into this question have also suggested that the problem is more complex than it might seem (Kim 1998). Furthermore, the impact of taxon sampling interacts with the technique used to obtain a model tree, and so simulation studies can differ based on factors that we do not yet understand. As intriguing as these questions are, they are unfortunately beyond the scope of the current paper; we leave them for future research.

The development of improved methods for using gene order data for phylogeny reconstruction on the GNT model is still a very active area of research. In addition to the approach we have taken in this paper, researchers have developed methods based on minimizing tree length (Sankoff and Blanchette 1998; Cosner et al. 2000; Bourque and Pevzner 2002; Moret et al. 2002; Wang et al. 2002), as well as Bayesian methods (Larget et al. 2002; Larget et al. 2004). Furthermore, except for the distance-based methods we describe, these other approaches are limited to small datasets because of the computational difficulties involved in these analyses. Thus, our distance-based methods, which are very fast and can handle large numbers of genomes, will continue to be valuable for reconstructing phylogenies as the number of completely sequences genomes increases rapidly.

However, the GNT model applies to only a single chromosome, and only allows events that maintain the number of genes. Thus, these analyses only apply to datasets based on a single chromosome, in which all genomes have equal gene content. Several researchers are beginning to expand their methods for more complex models, which allow events that change the number of copies of each gene and which move genes between genomes. Calculations of distances in these models are much more complicated; initial results along these lines have been obtained by El-Mabrouk, Moret, and others (Marron et al. 2004; Swenson et al. 2005; Tang and Moret 2003; El-Mabrouk 2001, 2002; El-Mabrouk and Sankoff 2000; Belda et al. 2005; Moret et al. 2005; Moret and Warnow 2005), as well as models that handle multiple chromosomes (Tesler 2002). These more advanced models will be essential to expanding methodologies to the consideration of eukaryotic nuclear genome comparisons. In addition, some researchers are considering models in which the probability of the rearrangement events is not uniform within a class. For example, some newer models define the probability of the event so that it depends on the lengths of the affected segments (for one such model see Pinter and Skiena 2002), or make assumptions that incorporate hotspots, or break the chromosome into distinct regions and require events to stay within these regions (Tesler and Pevzner 2003). Future research will explore the estimation of evolutionary distances under more sophisticated models of genome evolution.

Appendix

EDE: The Empirically Derived Estimator

EDE produces the best results under all model conditions, even when the evolutionary model is exclusively transpositions. For details about the mathematical derivation of this technique, see Moret et al. (2001).

EDE is based on inverting a function for the expected minimum inversion distance produced by a sequence of random inversions. Theoretical approaches (i.e., actually trying to analytically solve the expected inversion distance produced by k random inversions) proved to be quite difficult, and so we studied this under simulation. Our initial studies showed little difference in the behavior under 120 genes (typical for chloroplasts) and 37 genes (typical of mitochondria) and, in particular, suggested that it should be possible to express the normalized expected inversion distance as a function of the normalized number of random inversions. Therefore, we attempted to define a simple function $f(k/n)$ that approximates $E[\text{dINV}(G_0, G_k)/n]$ well, for k the number of random inversions, n the number of genes, G_0 the initial genome, and G_k the result of applying k random inversions to G_0.

The function f should have the following properties.

1. $0 \leq f(x) \leq x$, since the inversion distance is always less than or equal to the actual number of inversions.
2. $\lim_{x \to \infty} f(x) \approx 1$, as our simulations show the normalized expected inversion distance is close to 1, when a large number of random inversions is applied.
3. $f'(0) = 1$, since a single random inversion always produces a genome that is inversion distance 1 away.
4. $f^{-1}(y)$ is defined for all y [0, 1].

We used $nf(x)$ to estimate $E[d_{INV}(G_{nx}, G_0)]$, the expected inversion distance after nx inversions are applied. The nonlinear formula

$$f(x) = (ax^2 + bx)/(x^2 + cx + b) \qquad (1)$$

satisfies constraints (2) and (4).

We tried several different values for the constant a, and observed in our experiments that setting $a = 1$ produced the best results in subsequent phylogeny reconstructions using neighbor joining, for all values of n (the number of genes). The estimation of the constants b and c then amounts to a least-squares nonlinear regression; using simulated data we obtained $b = 0.5956$ and $c = 0.4577$. However, with this setting for a, b, and c, the formula does not satisfy the first constraint. Hence, we modified the formula to ensure that constraint (1) holds, and obtained

$$f(x) = \min\{x, (ax^2 + bx)/(x^2 + cx + b)\} \qquad (2)$$

The inverse of f is given by the formula

$$f^{-1}(d) = \max\{d, (-(b - cd) + ((b - cd)^2 + 4bd(1 - d))^{1/2})/(2(1 - d))\} \qquad (3)$$

Using the function f given above, we can thus define EDE, a method of moments estimator, as follows.

- Step 1: Given genomes G and G′, compute the inversion distance d.
- Step 2: Return $n f^{-1}(d/n)$, where n is the number of genes, as the estimate of the actual number of rearrangement events.

Since the function f is directly invertible, this allows us to estimate distances efficiently. **Theorem 1 (Moret et al. 2001)**. Let m be the number of genomes and let n be the number of genes. We can compute the pairwise EDE distance between every pair of genomes in $O(nm^2)$ time. If the inversion distance matrix is already computed, then we can compute the EDE distance matrix in $O(m^2)$ time.

Acknowledgments. We thank the two anonymous reviewers for their comments and for the suggestion of the Rzhetsky-Nei interior branch length test from one of the reviewers. This research was supported by National Science Foundation Grants EIA0121680, EF0331453, DEB0120709, DEB0075700, IIS0113654, EF0331654, IIS0121377, IIS0113095, and ANI020203584. Bernard Moret would like to acknowledge support from the IBM Corporation under a DARPA grant for the HPCS initiative and from the NIH under Grant 2R01GM056120-05A1 through a subcontract to the University of Arizona. Li-San Wang was supported in part by a NIH Training Grant in Bioinformatics. Tandy Warnow would like to acknowledge the support of the David and Lucile Packard Foundation, the Radcliffe Institute for Advanced Study, the Program in Evolutionary Dynamics at Harvard, and the Institute for Cellular and Molecular Biology at the University of Texas at Austin.

References

Bader D, Moret B, Yan M (2001) A linear time algorithm for computing inversion distances between signed permutations with an experimental study. J Comput Biol 8(5):483–491

Belda E, Moya A, Silva F (2005) Genome rearrangement distances and gene order phylogeny in γ-proteobacteria. Mol Biol Evol 22:1456–1467

Blanchette M, Bourque G, Sankoff D (1997) Breakpoint phylogenies. In: Miyano S, Takagi T (eds) Genome informatics. Univ. Acad. Press, pp 25–34 1

Blanchette M, Kunisawa M, Sankoff D (1999) Gene order breakpoint evidence in animal mitochondrial phylogeny. J Mol Evol 49:193–203

Boore J (1999) Animal mitochondrial genomes. Nucleic Acids Res 27:1767–1780

Boore J, Brown W (1998) Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. Curr Opin Genet Dev 8(6):668–674

Boore JL, Collins TM, Stanton D, Daehler LL, Brown WM (1995) Deducing arthropod phylogeny from mitochondrial DNA rearrangements. Nature 376:163–165

Bourque G, Pevzner P (2002) Genome scale evolution: reconstructing gene orders in the ancestral species. Genome Res 12(1):26–36

Cosner M, Jansen R, Moret B, Raubeson L, Wang LS, Warnow T, Wyman S (2000) A new fast heuristic for computing the breakpoint phylogeny and a phylogenetic analysis of a group of highly rearranged chloroplast genomes. In: Proceedings, 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'00). AAAI Press, pp 104–115 2

Desper R, Gascuel O (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle. J Comput Biol 19(5):687–705

Downie S, Palmer J (1992) Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In: Soltis P, Soltis D, Doyle J (eds) Molecular systematics of plants, Vol 49. Chapman & Hall, London, pp 14–35

El-Mabrouk N (2001) Sorting signed permutations by reversals and insertions/deletions of contiguous segments. J Discrete Algorithms 1(1):105–122

El-Mabrouk N (2002) Reconstructing an ancestral genome using minimum segments duplications and reversals. J Comput Syst Sci 65:442–464

El-Mabrouk N, Sankoff D (2000) Duplication, rearrangement and reconciliation. In: Comparative genomics: empirical and analytical approaches to gene order dynamics, map alignment and the evolution of gene families, Vol 1. Kluwer Academic, New York, pp 537–550

Hannenhalli S, Pevzner P (1995) Transforming cabbage into turnip (polynomial algorithm for genomic distance problems). In: Proceedings, 27th Annual ACM Symposium on the Theory of Computing (STOC'95). ACM Press, New York, pp 178–189

Heard SB (1992) Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. Evolution 46:1818–1826

Kim J (1998) Large scale phylogenies and measuring the performance of phylogenetic estimators. Syst Biol 47(1):43–60

Larget B, Simon D, Kadane J (2002) Bayesian phylogenetic inference from animal mitochondrial genome arrangements (with discussion). J Roy Stat Soc Ser B 64:681–693

Larget B, Simon D, Kadane J, Sweet D (2004) A Bayesian analysis of metazoan mitochondrial genome arrangements. Mol Biol Evol 22(3):486–495

Marron M, Swenson K, Moret B (2004) Genomic distances under deletions and insertions. Theor Comput Sci 325(3):347–360 (Special issue: papers from COCOON'03)

Moret B, Warnow T (2005) Advances in phylogeny reconstruction from gene order and content data. In: Zimmer E, Roalson E (eds) Molecular evolution, producing the biochemical data, Part B, 395. Elsevier, Amsterdam, pp 673–700

Moret B, Wang LS, Warnow T, Wyman S (2001) New approaches for reconstructing phylogenies based on gene order. Bioinformatics 17(Suppl):165–173

Moret B, Tang J, Wang LS, Warnow T (2002) Steps toward accurate reconstructions of phylogenies from gene order data. J Comput Syst Sci 65:508–525

Moret B, Tang J, Warnow T (2005) Reconstructing phylogenies from gene content and geneorder data. In: Gascuel O (ed) Mathematics of evolution and phylogeny. Oxford University Press, New York, pp 321–352

Nakhleh L, Moret B, Roshan U, John KS, Sun J, Warnow T (2002a) The accuracy of fast phylogenetic methods for large datasets. In: Proceedings, 7th Pacific Symposium on Biocomputing (PSB'02), pp 211–222

Nakhleh L, Roshan U, Vawter L, Warnow T (2002b) Estimating the deviation from a molecular clock. In: Lecture Notes in Computer Science: Proceedings of the 2nd Workshop for Algorithms and Bioinformatics (WABI'02), Vol 2452. Springer Verlag, New York, pp 287–299

Pinter R, Skiena S (2002) Genomic sorting with length weighted reversals. Genome Inform 13:103–111

Raubeson L, Jansen R (1992) Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. Science 255:1697–1699

Raubeson L, Jansen R (2005) Chloroplast genomes of plants. In: Henry R (ed) Diversity and evolution of plants genotypic and phenotypic variation in higher plants. CABI, London, pp 45–68

Rokas A, Holland PWH (2000) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15:454–459

Rzhetsky A, Nei M (1992) A simple method for estimating and testing minimumevolution trees. Mol Biol Evol 35:367–375

Rzhetsky A, Sitnikova T (1996) When is it safe to use an oversimplified substitution model in treemaking? Mol Biol Evol 13(9):1255–1265

Saitou N, Imanishi T (1989) Relative efficiencies of the Fitch-Margoliash, maximumparsimony, maximum likekihood, minimum evolution, and neighbor joining methods of phylogenetic tree construction in obtaining the correct tree. Mol Biol Evol 6(5):514–525

Saitou N, Nei M (1987) The neighbor joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sanderson MJ (2003) Analysis of rates (r8s) of evolution, v1.6; available at: Http://ginger.ucdavis.edu/r8s/

Sankoff D, Blanchette M (1998) Multiple genome rearrangement and breakpoint phylogeny. J Comput Biol 5:555–570

Sourdis J, Krimbas C (1987) Accuracy of phylogenetic trees estimated from DNA sequence data. Mol Biol Evol 4:159–166

Swenson K, Marron M, Earnest DeYoung J, Moret B (2005) Approximating the true evolutionary distance between two genomes. In: Proceedings, 7th Workshop on Algorithm Engineering and Experiments (ALENEX'05). SIAM Press, pp 37–46

Swofford D (2001) PAUP* 4.0. Sinauer Associates, Sunderland, MA

Swofford D, Olson G, Waddell P, Hillis D (1996) Phylogenetic inference. In: Hillis D, Moritz C, Mable B (eds) Molecular systematics, 2nd ed. Sinauer Associates, Sunderland, MA, chap 11

Tang J, Moret B (2003) Phylogenetic reconstruction from gene rearrangement data with unequal gene contents. In: Lecture Notes in Computer Science: Proceedings, 8th Workshop on Algorithms and Data Structures (WADS'03), Vol 3069, pp 37–46

Tannier E, Sagot M (2004) Sorting by reversals in subquadratic time. In: Lecture Notes in Computer Science: Proceedings, 15th Symposium on Combinatorial Pattern Matching (CPM'04), Vol 3109. Springer Verlag, New York, pp 1–13

Tesler G, Pevzner P (2003) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proc Natl Acad Sci USA 100(13):7672–7677

Tessler G (2002) Efficient algorithms for multichromosomal genome rearrangements. J Comput Syst Scie 65:587–609

Wang LS, Jansen R, Moret B, Raubeson L, Warnow T (2002) Fast phylogenetic methods for the analysis of genome rearrangement data: an empirical study. In: Proceedings of the Fifth Pacific Symposium of Biocomputing (PSB'02), pp 524–535

Zwickl DJ, Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. Syst Biol 51(4):588–598