1

DISTANCE-BASED GENOME REARRANGEMENT PHYLOGENY

Li-San Wang
Department of Biology
University of Pennsylvania
Philadelphia, PA USA
lswang@mail.med.upenn.edu

Tandy Warnow
Department of Computer Sciences
University of Texas
Austin, TX USA
tandy@cs.utexas.edu

Abstract

Evolution operates on whole genomes through mutations, such as inversions, transpositions, and inverted transpositions, that rearrange genes within genomes. In this chapter we survey distance-based techniques for estimating evolutionary history under these events. We present the mathematical derivation of three statistically-based evolutionary distance estimators, and show that the use of these new distance estimators with methods such as neighbor joining and Weighbor can result in improved reconstructions of evolutionary history.

1.1 Introduction

The genomes of some organisms have a single chromosome or contain single chromosome organelles (such as mitochondria [5, 25] or chloroplasts [10, 24, 25, 27]) whose evolution is largely independent of the evolution of the nuclear genome for these organisms. Evolutionary events can alter these orderings through rearrangements such as inversions and transpositions, collectively called genome rearrangements. These events fall under the general category of "rare genomic changes", and are thought to have great potential for clarifying deep evolutionary histories [28]. In the last decade or so, a few researchers have used such data in their phylogenetic analyses [3, 5–7, 10, 24, 27, 31].

Of the various techniques for estimating phylogenies from gene order data, only distance-based methods are polynomial time. The first study that used distance-based methods to reconstruct phylogenies from gene orders was done

by Blanchette, Kunisawa, and Sankoff [5]. Their study gave a phylogenetic analysis using the neighbor joining [29] method applied to a matrix of "breakpoint distances" defined on a set of mitochondrial genomes for six metazoan groups. However, as this chapter will show, breakpoint distances do not provide particularly accurate estimations of evolutionary distances, and better estimations of trees can be obtained using other distance estimators.

The rest of the chapter is organized as follows. Section 1.2 provides the background on genome rearrangement evolution and describes the Generalized Nadeau-Taylor model. In Section 1.3 we discuss distance-based phylogeny reconstruction. We describe three new distance estimators for genome rearrangement evolution in Sections 1.4 and 1.5. We report on simulation studies evaluating the accuracy of these estimators, and of phylogenies estimated using these estimators on random tree topologies in Section 1.6. Finally, in Section 1.7 we discuss recent extensions to the Generalized Nadeau-Taylor model and discuss some relevant open problems in phylogeny reconstruction that arise.

1.2 Whole genomes and events that change gene orders

In this chapter we will study phylogeny reconstruction on whole genomes under the assumption that all genomes have exactly one copy of each gene; thus, all genomes have exactly the same gene content.

1.2.1 *Inversions and transpositions*

The events we consider do not change the number of copies of a gene, but only scramble the order of the genes within the genomes. Thus we will not consider events such as duplications, insertions, or deletions, but will restrict ourselves to inversions (also called "reversals") and transpositions.

Inversions operate by picking up a segment within the genome and reinserting the segment in the reverse direction; thus, the order and strandedness of the genes involved change. A transposition has the effect of moving a segment from between two genes to another location (between two other genes), without changing the order or strandedness of the genes within the segment. If the transposition is combined with an inversion, then the order and strandedness change as well - this is called an inverted transposition. Examples of these events are shown in Fig. 1.1.

1.2.2 *Representations of genomes*

In order to analyze gene order evolution mathematically, we represent each genome (whether linear or circular) as a signed permutation of $(1, 2, \ldots, n)$, where n is the number of genes and where the sign indicates the strand on which the gene occurs. Thus, a circular genome can be represented as a signed circular permutation, and a linear genome can be represented as a signed linear permutation. In the case of circular genomes, we use linear representations by beginning at any of its genes, in either orientation. We consider two such representations of a circular genome equivalent. As an example, the circular genome

(a) 1 2 3 4 5 6 7 8 9 10
(b) 1 2 3 -8 -7 -6 -5 -4 9 10
(c) 1 2 3 9 4 5 6 7 8 10
(d) 1 2 3 9 -8 -7 -6 -5 -4 10

FIG. 1.1. Examples of genome rearrangements. Genome (a) is the starting point for all the events we demonstrate. Genome (b) is obtained by applying an inversion to Genome (a). Genome (c) is obtained by applying a transposition to Genome (a). Genome (d) is obtained by applying an inverted transposition to Genome (a). In each of these events we have affected the same target segment of genes (genes 4 through 8, underlined in Genome (a)), and indicated its location (also by underlining) in the resultant genome.

given by the linear ordering $(1, 2, 3, 4, 5)$ is equivalently represented by the linear orderings $(2, 3, 4, 5, 1)$ and $(-2, -1, -5, -4, -3)$. As an example of how an inversion acts, if we apply an inversion on the segment $2, 3$ to $(1, 2, 3, 4, 5)$, we obtain $(1, -3, -2, 4, 5)$. For an example of a transposition, if we then apply a transposition moving the segment $-2, 4$ to between 1 and -3, we obtain $(1, -2, 4, -3, 5)$.

For the rest of the chapter we focus on circular genomes unless stated otherwise (our simulations show that all results can be directly applied to linear genomes without any significant difference in accuracy).

1.2.3 *Edit distances between genomes: inversion and breakpoint distances*

The kinds of distances we are most interested in estimating are evolutionary distances – the number of events that took place in the evolutionary history between two genomes. However, the two common ways of defining distances between genomes are breakpoint distances and inversion distances, neither of which provides a good estimate of evolutionary distances. We obtain our evolutionary distance estimators (described later in the chapter) by "correcting" these two distances.

Inversion distance The inversion distance between genomes G and G' is the minimum number of inversions needed to transform G into G' (or vice-versa, as it is symmetric); we denote this distance by $d_{\text{INV}}(G, G')$. The first polynomial time algorithm for computing this distance was obtained by Hannenhalli and Pevzner [15], and later improved by Kaplan, Shamir, and Tarjan [16] and Bader, Moret, and Yan [2] (the latter obtained an optimal linear-time algorithm). See Chapter 10 in this Volume for a review of these algorithms.

Breakpoint distance Another popular distance measure between genomes is the breakpoint distance [4]. A breakpoint occurs between genes g and g' in genome G' with respect to genome G if g is not followed immediately by g' in G. As an example, consider circular genomes $G = (1, 2, -3, 4, 5)$ and $G' = (1, 2, 3, -5, -4)$. There is a breakpoint between 2 and 3 in G', since 2 is not followed by 3 in G, but there is no breakpoint between -5 and -4 in G' (since G can be equivalently written as $(-1, -5, -4, 3, -2)$). The breakpoint distance between two genomes is the number of breakpoints in one genome with respect to the other, which is

clearly symmetric; we denote this distance by $d_{\mathrm{BP}}(G, G')$. In the example above the breakpoint distance is 3.

1.2.4 *The Nadeau-Taylor model and its generalization*

The *Nadeau-Taylor* model [22] assumes that only inversions occur (i.e., no transpositions or inverted transpositions occur), and all inversions have the same probability of occurring. This assumption that inversions are equiprobable was inspired by the observation made in [22] that the length of conserved segments between the human and mouse genomes (relative to each other) seems to be uniformly randomly distributed.

In [40] we proposed a generalized version of the Nadeau-Taylor model which allows for transpositions and inverted transpositions to occur. In the *Generalized Nadeau-Taylor* (GNT) model, all inversions have equal probability, as do different transpositions and 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 a rearrangement event is a transposition, and w_{IT} is the probability that a rearrangement event is an inverted transposition. Because we assume that all events are of these three types, $w_I + w_T + w_{IT} = 1$. Given a model tree, we will let $X(e)$ be the random variable for the number of evolutionary events that takes place on the edge e. We assume that $X(e)$ is a Poisson random variable with mean λ_e; hence, λ_e can be considered the length of the edge e. We also assume that events on one edge are independent of the events on other edges. Thus, the GNT model requires $O(m)$ parameters, where m is the number of genomes (i.e., leaves): the length λ_e of each edge e, and the triplet w_I, w_T, w_{IT}. We let $\mathrm{GNT}(w_I, w_T, w_{IT})$ denote the set of model trees with the triplet (w_I, w_T, w_{IT}). Thus, the Nadeau-Taylor model is simply the $\mathrm{GNT}(1, 0, 0)$ model.

1.3 Distance-based phylogeny reconstruction

There are many methods for reconstructing phylogenies, such as *maximum parsimony and maximum likelihood*, which are computationally intensive. In this chapter we focus on phylogeny reconstruction techniques that are polynomial time. For gene order phylogeny reconstruction, the fast methods are primarily *distance-based* methods. We briefly review the basic concepts here, and direct the interested reader to the chapter in this volume by Desper and Gascuel on distance-based methods for a more in-depth discussion.

1.3.1 *Additive and near-additive matrices*

Suppose we have a phylogenetic tree T on m leaves, and we assign a positive length $l(e)$ to each edge e in the tree. Consider the $m \times m$ matrix (D_{ij}) defined by $D_{ij} = \sum_{e \in P_{ij}} l(e)$, where P_{ij} is the path in T between leaves i and j. This matrix is said to be "additive". Interestingly, given the matrix (D_{ij}), it is possible to construct T and the edge lengths in polynomial time, up to the location of the root [41, 42], provided that we assume that T has no nodes of degree two.

The connection between this discussion and the inference of evolutionary histories is obtained by setting $l(e)$ to be the actual number of changes on the edge e. Then, $D_{ij} = \sum_{e \in P_{ij}} l(e)$ is the actual number of events (in our case, inversions, transpositions, and inverted transpositions) that took place in the evolutionary history relating genomes i and j.

Since estimations of evolutionary distances have some error, the matrices (d_{ij}) given as input to distance-based methods generally are not additive. Therefore, we may wish to understand the conditions under which a distance-based method will still correctly reconstruct the tree, even though the edge lengths may be incorrect. Research in the last few years has established that various methods, including neighbor joining [1], will still reconstruct the true tree as long as $L_\infty(D, d) = \max_{ij} |D_{ij} - d_{ij}|$ is small enough, where (d_{ij}) is the input matrix and (D_{ij}) is the matrix for the true tree (see [1,17] and Chapter 1 in this Volume).

Consequently, methods such as neighbor joining which have some error tolerance will yield correct estimates of the true tree, as long as each D_{ij} can be estimated with sufficient accuracy.

1.3.2 *The two steps of a distance-based method*

Using these observations, it is clear why distance-based methods have these two steps:

- **Step 1:** Estimate "evolutionary distances" (expected or actual number of changes) between every pair of taxa, producing matrix (d_{ij}).
- **Step 2:** Use a method (such as neighbor joining) to infer an edge-weighted tree from (d_{ij}).

The second step is fairly standard at this point, with neighbor joining [29] the most popular of the distance-based methods. However, the first step is very important as well. Extensive simulation studies under DNA models of site substitution have shown that phylogenies obtained using distance-based methods (such as neighbor joining) applied to statistically-based distance estimation techniques are closer to the true tree than when used with uncorrected distances. If, however, the evolutionary model obeys the molecular clock, so that the expected number of changes is proportional to time, then statistically based estimations of distance are unnecessary – correct trees can be reconstructed by applying simple reconstruction methods such as UPGMA [33] applied to Hamming distances. However, since the molecular clock assumption is not generally applicable, better distance estimation techniques are necessary for phylogeny reconstruction purposes.

The use of breakpoint distances and inversion distances in whole genome phylogeny reconstruction is problematic because these typically underestimate the actual number of events; therefore, they are not statistically consistent distance-estimators under the GNT model. This theoretical observation, coupled with

empirical results, motivates us to produce statistically-based distance estimators for the GNT model.

1.3.3 *Method of moments estimators*

The distance estimators we describe in this chapter are all method of moments estimators. Let X be a real-valued random variable whose distribution is parametrized by p; as a result $E[X]$ is a function f of p. The estimator $\widehat{p} = f^{-1}(x)$, where x is the observed value for the mean of X, is a method of moments estimator of the parameter p. In our case, since there is only one observation for X, the mean of X is simply the observed value for X. Method of moments estimators are common in many statistical applications, and generally have good accuracy; see any standard statistics textbook (such as Section 7.1 in [9]) for details.

In the context of gene order phylogeny, we have developed two functions which estimate the expected breakpoint distance produced by k random events under the GNT(w_I, w_T, w_{IT}) model, for each way of setting w_I, w_T and w_{IT}. One of these two functions is provably correct, and the other is approximate (with provable error bounds), but both have almost identical performance in simulation. We also have a function which estimates the expected inversion distance produced by k random inversions (i.e., random events in the GNT$(1, 0, 0)$ model).

Each of these functions is invertible, and thus can be used to estimate the number of events in the evolutionary history between two genomes in a simple way. For example, given the function $f(k)$ for the expected breakpoint distance produced by k random events in the GNT(w_I, w_T, w_{IT}) model on n genes (see Section 1.5), we can define a distance estimation technique, which we call `IEBP`, for "Inverting the Expected Breakpoint Distance" as follows:

- **Step 1:** Given genomes G and G', compute their breakpoint distance d.
- **Step 2:** Using the assumed values for w_I, w_T and w_{IT}, compute $f^{-1}(d)$. This is the estimate of the evolutionary distance between G and G'.

We demonstrate this technique in Fig. 1.2.

We have also developed a distance estimation technique called `EDE`, for the "Empirically Derived Estimator", which estimates the evolutionary distance between two genomes by inverting the expected inversion distance. (See Section 1.4 for the derivation of `EDE`.)

In the next sections we describe these three distance-estimators: `Exact-IEBP`, which is based upon an exact formula for the expected breakpoint distance, `Approx-IEBP`, which is based upon an approximate formula (with guaranteed error bounds) for the expected breakpoint distance, and `EDE`, which is based upon a heuristic for the expected inversion distance. All three estimators improve upon both breakpoint and inversion distances as evolutionary distance estimators, and produce better phylogenetic trees, especially when the datasets come from model trees with high evolutionary diameters (so that the datasets are close to saturation). Of the three, `Exact-IEBP` and `Approx-IEBP` have the best accuracy

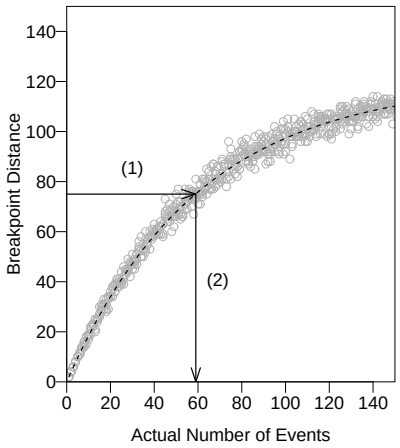


FIG. 1.2. Illustration of the IEBP technique, a method of moments estimator. The backdrop is the scatter plot of simulations with 120 genes, inversion-only evolution. The dashed line is the expected breakpoint distance (the function f in the paragraph describing IEBP), as a function of the number of inversions. In the first step we compute the breakpoint distance d (the y-axis coordinate); in the second step we find $f^{-1}(d)$ as the estimate of the actual number of inversions.

with respect to distance estimation, but surprisingly phylogeny reconstruction based upon EDE is somewhat more accurate than phylogeny reconstruction based upon the other estimators.

In the next sections we provide the derivations for these three evolutionary distance estimators. We begin with EDE because it is the simplest to explain, and the mathematics is the least complicated.

1.4 EDE: The "Empirically Derived Estimator"

Our first method of moments estimator is EDE, which is based upon inverting the expected inversion distance produced by random inversions. Because our technique in deriving EDE is empirical (i.e., we do not have theory to establish any performance guarantees for EDE's distance estimation), we call it the "Empirically Derived Estimator." However, despite the lack of provable theory, of our three evolutionary distance estimators, EDE produces the best results whether we use neighbor joining or Weighbor [8] (a variant of neighbor joining that uses the variance of the evolutionary distance estimators as well). EDE is quite robust, and performs well even when the model does not permit inversions. The results in this section are taken from [20, 39].

1.4.1 *The method of moments estimator: EDE*

The EDE estimator is based upon inverting the expectation of the inversion distance produced by a sequence of random inversions under the GNT$(1, 0, 0)$

model. Thus, to create EDE we have to find a function which will estimate the expected 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 $Q(\frac{k}{n})$ that approximates $E[d_{\text{INV}}(G_0, G_k)/n]$ 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. This function Q should have the following properties:

(1) $0 \leq Q(x) \leq x$, since the inversion distance is always less than or equal to the actual number of inversions.
(2) $\lim_{x\to\infty} Q(x) \simeq 1$, as simulation shows the normalized expected inversion distance is close to 1 when a large number of random inversions is applied.
(3) $Q'(0) = 1$, since a single random inversion always produces a genome that is inversion distance 1 away.
(4) $Q^{-1}(y)$ is defined for all $y \in [0, 1]$, so that we may invert the function.

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

$$Q(x) = \frac{ax^2 + bx}{x^2 + cx + b}$$

satisfies constraints (2)-(4).

The quantity $\lim_{x\to\infty} Q(x) = a$ in constraint (2) has the following interpretation. When a large number of random inversions are being applied to a genome G, the resultant genome should look random with respect to G. This quantity is very close to one as n, the number of genes in G, increases, but for finite n a does not equal 1. Nonetheless, by simply setting $a = 1$ the formula produces very accurate results in practice.

The estimation of b and c amounts to a least-squares nonlinear regression. We found that setting $b = 0.5956$ and $c = 0.4577$ produced a good fit to the empirical data. However, with this setting for a, b, and c, the formula does not satisfy the first constraint. Hence, we modify the formula to ensure that constraint (1) holds, and obtain:

$$Q^*(x) = \min\{x, Q(x)\} = \min\{x, \frac{ax^2 + bx}{x^2 + cx + b}\}.$$

Please refer to Fig. 1.3 for our simulation study evaluating the performance of this formula in fitting the expectation.

EDE's algorithm We can define a method of moments estimator EDE, using the function Q^*, as follows:

- **Step 1:** Given genomes G and G', compute the inversion distance d.

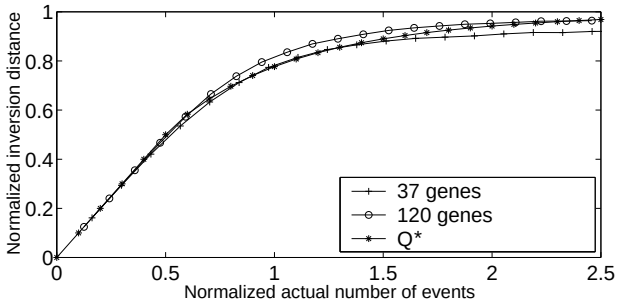


FIG. 1.3. Comparison of the regression formula Q^* for the expected inversion distance in EDE with simulated data. Both the x- and y-axis coordinates are normalized – both are divided by the number of genes.

- **Step 2:** Return $k = n(Q^*)^{-1}(\frac{d}{n})$, where n is the number of genes.

As the number of actual events must be an integer, another way to obtain an estimate of the evolutionary distance is to choose either $\lfloor k \rfloor$ and $\lceil k \rceil$. However, in practice there is almost no difference in the accuracy of the tree inferred whether we use the inverted function or the closest integer criterion to compute the EDE distance matrix.

We summarize the EDE distance estimator as follows.

Definition 1.1 *Let G and G' be two genomes with genes $\{1, 2, \ldots, n\}$. Define*

$$Q^*(x) = \min\{x, Q(x)\} = \min\{x, \frac{x^2 + 0.5956x}{x^2 + 0.4577x + 0.5956}\}.$$

The EDE distance between G and G' is

$$EDE(G, G') = n(Q^*)^{-1}(\frac{d}{n}),$$

where $d = d_{INV}(G, G')$ is the inversion distance between G and G'.

EDE therefore is a method of moments estimator of the actual number of inversions that took place in transforming G into G' under the GNT$(1, 0, 0)$ model (i.e., inversion-only evolution).

Let m be the number of genomes and let n be the number of genes. Computing the inversion distance between each pair of genomes takes only $O(n)$ time, for a total of $O(nm^2)$ time. Once the inversion distance matrix is computed, as the formula Q^* used in EDE is directly invertible, computing the entire EDE distance matrix takes an additional $O(m^2)$ time.

Note that EDE, our first method of moments estimator, was derived on the basis of a simulation study involving 120 genes under an inversion-only evolutionary model. Therefore, the distance estimated by EDE is independent of the model condition: we will get the same estimated distance no matter what we

know about the model conditions. Despite this rigidity in EDE's structure and origin, we can apply EDE to any pair of genomes and use it to estimate evolutionary distances. Interestingly, we will see that EDE is quite robust to model violations, and can be used with methods such as neighbor joining to produce highly accurate estimations of phylogenies. See Section 1.6 for experimental results evaluating the accuracy of EDE and of distance-based tree reconstruction methods using EDE in simulation.

1.4.2 *The variance of the inversion and EDE distances*

In order to use EDE with methods such as Weighbor, we need also to have an estimate for the variance of the EDE distance. We therefore developed an estimator (presented in [39]) for the standard deviation of the normalized inversion distance produced by nx random inversions, where n is the number of genes. The approach we used to obtain this estimate is similar to the approach we used to derive EDE.

The variance of the inversion distance The first step is to obtain the variance of the inversion distance. After several experiments with simulated data, we decided to use the following regression formula:

$$\sigma_n(x) = n^q \frac{ux^2 + vx}{x^2 + wx + t}.$$

The constant term in the numerator is zero because we know $\sigma_n(0) = 0$. As we did in our derivation of the EDE technique, we make the assumption that the actual number of inversions is no more than $3n$.

Note that

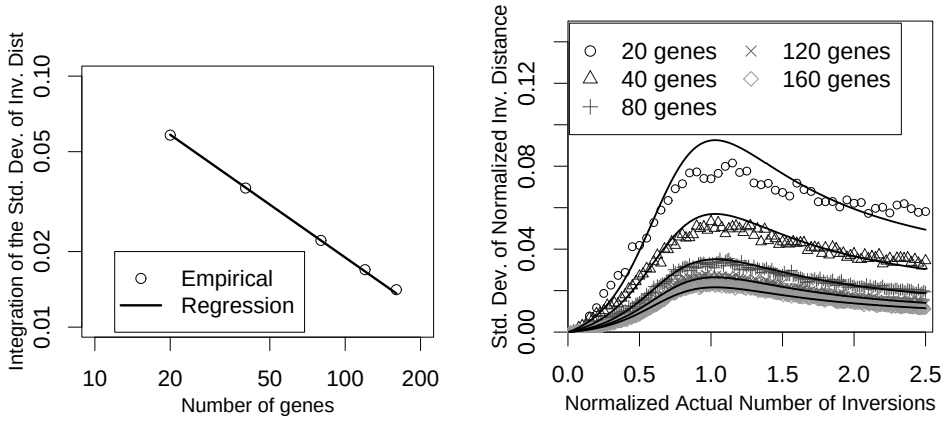
$$\begin{aligned}\ln\left(\frac{1}{3n}\sum_{i=0}^{3n}\sigma_n\left(\frac{i}{n}\right)\right) &= q\ln n + \ln\left(\frac{1}{3n}\sum_{i=0}^{3n}\frac{u\left(\frac{i}{n}\right)^2 + v\left(\frac{i}{n}\right)}{\left(\frac{i}{n}\right)^2 + w\left(\frac{i}{n}\right) + t}\right)\\ &\simeq q\ln n + \ln\left(\frac{1}{3}\int_0^3 \frac{ux^2 + vx}{x^2 + wx + t}dx\right)\end{aligned}$$

is linear in $\ln n$. Thus we can obtain q as the slope in the linear regression using $\ln n$ as the independent variable and $\ln(\frac{1}{3n}\sum_0^{3n}\sigma_n(i/n))$ as the dependent variable. Our simulation results, shown in Fig. 1.4(a), suggest that $\ln(\frac{1}{3n}\sum_{i=0}^{3n}\sigma_n(\frac{i}{n}))$ indeed is (almost) linear in $\ln n$.

After obtaining $q = -0.6998$, we applied nonlinear regression to obtain u, v, w, and t, using the simulated data for 40, 80, 120, and 160 genes, and obtained the values $q = -0.6998$, $u = 0.1684$, $v = 0.1573$, $w = -1.3893$, and $t = 0.8224$. The resultant functions are shown as the solid curves in Fig. 1.4(b).

Estimating the variance of EDE The variance of EDE can now be obtained using a common statistical technique called the *delta method* [23] as follows. Assume Y is a random variable with variance $\mathrm{Var}[Y]$, and let $X = f(Y)$. Then $\mathrm{Var}[X]$ can be approximated by $\left(\frac{dX}{dY}\right)^2 \mathrm{Var}[Y]$.

To apply the delta method to EDE, we set Y to be the normalized inversion distance between genomes G and G' (i.e., the inversion distance divided by



(a) (b)

FIG. 1.4. (a) regression of coefficient q (see Section 1.4); for every point corresponding to n genes, the y coordinate is the average of all data points in the simulation. (b) simulation (points) and regression (solid lines) of the standard deviation of the inversion distance.

the number of genes), and set $X = Q^{-1}(Y)$ (we do not use Q^* since it is not differentiable in its entire range).

Let G and G' be two genomes with genes $\{1, 2, \ldots, n\}$. Let $x = \texttt{EDE}(G, G')/n$. Since $\frac{d}{dY}Q^{-1}(Y) = \left(Q'(Q^{-1}(Y))\right)^{-1}$, the variance of the EDE distance can be approximated as

$$\mathrm{Var}[\texttt{EDE}(G, G')] \simeq n^2 \left(\frac{1}{Q'(x)}\right)^2 \left(n^{-0.6998} \frac{0.1684x^2 + 0.1573x}{x^2 - 1.3893x + 0.8224}\right).$$

Here $Q(x)$ is the function defined in Section 1.4, upon which Q^*, the expected inversion distance, is based.

1.5 IEBP: "Inverting the Expected Breakpoint distance"

Exact-IEBP and Approx-IEBP are two method of moments estimators based upon functions for estimating the expected breakpoint distance produced by k random events under the GNT(w_I, w_T, w_{IT}) model, where w_I, w_T, and w_{IT} are given. Thus, "IEBP" stands for "inverting the expected breakpoint distance". Exact-IEBP is based upon an exact calculation of the expected breakpoint distance, and Approx-IEBP is based upon an approximate estimation of the expected breakpoint distance which we can prove has very low error. In order to

use `IEBP` (Exact- or Approx-) with Weighbor, we also developed a technique for estimating the variance of the `IEBP` distance; this is presented in Section 1.5.3.

1.5.1 *The method of moments estimator,* `Exact-IEBP`

We begin with the derivation of the expected breakpoint distance produced by a sequence of random events under the GNT(w_I, w_T, w_{IT}) model. By linearity of expectation and symmetry of the model, it suffices to find the distribution of the presence/absence of a single breakpoint (a zero-one variable).

We consider how a circular genome evolves under the Generalized Nadeau-Taylor model (the analysis for linear genomes can be obtained easily using the same techniques). Let the number of genes in the genome be n. We start with genome $G_0 = (1, 2, \ldots, n)$, and we let G_k denote the genome obtained after k random rearrangement events are applied under the Generalized Nadeau-Taylor model.

We begin by defining a character L on circular genomes which will have states in $\{\pm 1, \pm 2, \ldots, \pm(n-1)\}$. The state of this character on a genome is defined as follows:

- In G', do genes 1 and 2 have the same sign, or different signs? If the same sign, then $L(G')$ is positive, and otherwise $L(G')$ is negative.
- We then count the number of genes between 1 and either 2 or -2 in G' (depending upon which one appears in G''s representation when we use gene 1 in its positive strand), and add 1 to that value; this is $|L(G')|$.

We present some examples of how L is defined on different genomes with 6 genes. If $G' = (1, 2, 4, 5, -3, 6)$ then $L(G') = 1$, while if $G' = (1, -2, 3, 4, 5, 6)$ then $L(G') = -1$. A somewhat harder example is $G' = (1, 5, 3, -2, 4, 6)$, for which $L(G') = -3$ (gene 2 is the third gene to follow gene 1, and it is located on the other strand).

The following lemma shows the number of rearrangement events transforming G into genome G' only depends on $L(G)$, $L(G')$ and the number n of genes. Thus, the distribution of a breakpoint is a $(2n-2)$-state Markov chain, and we use the character L defined above to assign states to genomes. We sketch the proof for the transposition-only situation.

To facilitate the proof, we formally characterize transpositions on circular genomes. A transposition on G has three indices, a, b, c, with $1 \leq a < b \leq n$ and $2 \leq c \leq n$, $c \notin [a, b]$, and operates on G by picking up the interval $g_a, g_{a+1}, \ldots, g_{b-1}$ and inserting it immediately after g_{c-1}. Thus the genome $G = (g_1, g_2, \ldots, g_n)$ (with the additional assumption of $c > b$) is replaced by

$$(g_1, \ldots, g_{a-1}, g_b, g_{b+1}, \ldots, g_{c-1}, g_a, g_{a+1}, \ldots, g_{b-1}, g_c, \ldots, g_n)$$

Lemma 1.2 *[38] Let n be the number of genes. Let $\iota_n(u, v)$, $\tau_n(u, v)$ and $\nu_n(u, v)$ be the number of inversions, transpositions, and inverted transpositions, respectively, that bring a genome in state u to state v. Assume the genome is circular. Then*

$$\iota_n(u,v) = \begin{cases} \min\{|u|,|v|,n-|u|,n-|v|\} & \text{if } uv<0 \\ 0 & \text{if } u\neq v, uv>0 \\ \binom{|u|}{2}+\binom{n-|u|}{2} & \text{if } u=v \end{cases}$$

$$\tau_n(u,v) = \begin{cases} 0 & \text{if } uv<0 \\ (\min\{|u|,|v|\})(n-\max\{|u|,|v|\}) & \text{if } u\neq v, uv>0 \\ \binom{|u|}{3}+\binom{n-|u|}{3} & \text{if } u=v \end{cases}$$

$$\nu_n(u,v) = \begin{cases} (n-2)\iota_n(u,v) & \text{if } uv<0 \\ \tau_n(u,v) & \text{if } u\neq v, uv>0 \\ 3\tau_n(u,v) & \text{if } u=v \end{cases}$$

Proof The formula for ι is first shown in [32]. Here we sketch the proof for τ.

Assume the current genome is in state u. Let v be the new state of the genome after the transposition with indices (a,b,c), $1 \le a < b < c \le n$. Since transpositions do not change the sign, $\tau_n(u,v) = \tau_n(-u,-v)$, and $\tau_n(u,v) = 0$ if $uv < 0$. Therefore we only need to analyze the case where $u, v > 0$.

We first analyze the case when $u = v$. Suppose that either $a \le u < b$ or $b \le u < c$. In the first case, we immediately have $v = u + (c-b)$, therefore $v - u = c - b > 0$. In the second case, we have $v = u + (a-b)$, therefore $v - u = a - b < 0$. Both cases contradict the assumption that $u = v$, and the only remaining possibilities that makes $u = v$ are when $1 \le u = v < a$ or $c \le u = v \le n-1$. This leads to the third line in the $\tau_n(u,v)$ formula.

Next, the total number of solutions (a,b,c) for the following two problems is $\tau_n(u,v)$ when $u \neq v$ and $u, v > 0$:

$$(1)\ u < v : b = c - (v-u),\ \ 1 \le a \le u < b < c \le n, u < v \le c,$$
$$(2)\ u > v : b = a + (u-v),\ \ 1 \le a < b \le u < c \le n, a \le v < u.$$

In the first case $\tau_n(u,v) = u(n-v)$, and in the second case $\tau_n(u,v) = v(n-u)$. The second line in the $\tau_n(u,v)$ formula follows by combining the two results. □

We now derive the distribution of the Markov chain. To simplify the formulas, we index all vectors and matrices by the states $\{\pm1, \pm2, \ldots, \pm(n-1)\}$. Let G_k be the result of applying k random rearrangements to genome G_0 under $\mathrm{GNT}(w_I, w_T, w_{IT})$. We first obtain the transition matrix.

Lemma 1.3 *Let M_I, M_T, and M_{IT} be the transition matrices of the Markov chain when only inversions, transpositions, or inverted transpositions occur, respectively. We let w_I be the probability of an inversion, w_T be the probability of a transposition, and w_{IT} be the probability of an inverted transposition (with $w_I + w_T + w_{IT} = 1$). Then*

(a) $M_I[u,v] = \dfrac{\iota_n(u,v)}{\binom{n}{2}}, \quad M_T[u,v] = \dfrac{\tau_n(u,v)}{\binom{n}{3}}, \quad M_{IT}[u,v] = \dfrac{\nu_n(u,v)}{3\binom{n}{3}}.$

(b) The transition matrix M of the breakpoint Markov chain is

$$M = w_I M_I + w_T M_T + w_{IT} M_{IT}.$$

Proof Results in (a) follow from Lemma 1.2 together with the observation that there are $\binom{n}{2}$ distinct inversions, $\binom{n}{3}$ distinct transpositions, and $3\binom{n}{3}$ distinct inverted transpositions. □

Theorem 1.4 *Let M be the transition matrix of the breakpoint Markov chain as described above. Then*

$$E[d_{BP}(G_0, G_k)] = n(1 - M^k[1,1]).$$

Proof Let L be the character defined for the Markov chain (i.e., $L(G')$ is the state of genome G') and let x_k be the distribution vector of $L(G_k)$. Because $L(G_0) = 1$, we can set x_0 as follows:

$$\begin{aligned} x_0[1] &= 1, \\ x_0[u] &= 0, \quad u \in \{-1, \pm 2, \ldots, \pm(n-1)\}. \end{aligned}$$

Since $x_k = M^k x_0$,

$$\begin{aligned} \Pr(L(G_k) = 1) &= (M^k x_0)[1,1] = M^k[1,1] \\ \Rightarrow E[d_{\mathrm{BP}}(G_0, G_k)] &= n \Pr(L(G_k) \neq 1) = n(1 - M^k[1,1]). \end{aligned}$$

□

We summarize the `Exact-IEBP` distance as follows.

Definition 1.5 *Assume the evolutionary model is GNT(w_I, w_T, w_{IT}). Let G and G' be two genomes with genes $\{1, 2, \ldots, n\}$. Let*

$$Y(k) = n(1 - M^k[1,1]),$$

where M is defined in Lemma 1.3. The `Exact-IEBP` distance is the nonnegative integer k that minimizes $|Y(k) - d|$:

$$\textit{Exact-IEBP}(G, G') = \operatorname*{argmin}_{\text{integer } k \geq 0} |Y(k) - d|,$$

where $d = d_{BP}(G, G')$ is the breakpoint distance between G and G'.

Thus, `Exact-IEBP` is a method of moments estimator of the actual number of evolutionary events under the GNT model, which uses assumed values of w_I, w_T and w_{IT}.

Note the following. First, computing the expected breakpoint distance produced by k random events is done recursively, and the calculation takes $O(n^2 k)$ time. Second, because breakpoints are not independent, extending the approach in order to study higher order statistics such as the variance is difficult. To see why breakpoints are not independent, consider the following argument. If breakpoints were independent, then the probability of having breakpoint distance 1 would be positive, as it is a product of n positive values. Since no two genomes can differ by one breakpoint, this is impossible.

Let m be the number of genomes, and n be the number of genes in each genome. Computing the breakpoint distance matrix takes $O(m^2n)$ time total. To compute the `Exact-IEBP` distance matrix the first step of the algorithm is to compute $Y(k)$, the expected breakpoint distance produced by k random events, for each k between 1 and $3n$. This amounts to $3n$ (transition) matrix-(state probability) vector multiplications, and uses $O(n^3)$ time. To invert $Y(k)$ (as a method of moments estimator requires) we use binary search in $O(\log n)$ time (we assume the number of rearrangement events never exceed $3n$). Because there are at most n different breakpoint distance values, computing the `Exact-IEBP` distance matrix when the breakpoint distance is known takes $O(n^3 + m^2 + \min\{m^2, n\} \log n)$ time.

1.5.2 *The method of moments estimator, `Approx-IEBP`*

In this section we present an approximate version of `Exact-IEBP`, which we call `Approx-IEBP` (see [40] for the details). Rather than exactly computing the expected number of breakpoints produced by a sequence of random events in the GNT model, we compute an approximation of that value. Because we allow an approximation, we can obtain the estimation faster; thus, the main advantage over `Exact-IEBP` is the running time. Fortunately, we are able to provide very good error bounds on the estimation. Our simulation results, shown later in this chapter, also show that `Approx-IEBP` is almost as accurate as `Exact-IEBP`, and that trees inferred based upon either version of `IEBP` are almost indistinguishable. The technique we used to obtain `Approx-IEBP` is based upon an analysis using 2-state Markov Chains. We describe that approach here.

Without loss of generality, consider the 2-state stochastic process indicating the presence of a breakpoint between genes 1 and 2. We let 0 denote the absence of a breakpoint between 1 and 2 (i.e., that gene 2 immediately follows gene 1), and we let 1 indicate the presence of the breakpoint (i.e., that gene 1 is not immediately followed by gene 2). The 2-state stochastic process is shown in Fig. 1.5. While the transitional probability s of jumping from state 0 to 1 in one step is a constant, the transitional probability u of jumping from state 1 to 0 in one step depends on both the sign of gene 2 and the number of genes between the two genes. Thus, no Markov chain with only these two states (presence or absence of a breakpoint) can completely specify the stochastic process. However, we can always find tight bounds on u.

Lemma 1.6 *Let G_0 be a signed circular genome with n genes. Let the model of evolution be GNT(w_I, w_T, w_{IT}). The transitional probability s of jumping from state 0 to state 1 after a rearrangement event occurs is given by*

$$s = \frac{2 + w_T + w_{IT}}{n}$$

and the transitional probability u of jumping from state 1 to state 0 after a rearrangement event occurs is between 0 and u_H, where

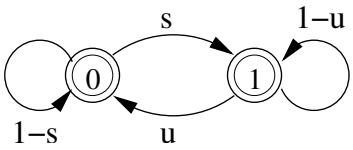


FIG. 1.5. The two-state stochastic process of the breakpoint between genes 1 and 2 under the Generalized Nadeau-Taylor model.

$$u_{_H} = \frac{2(n-2) + 4w_T(n-2) + 2w_{IT}n}{n(n-1)(n-2)}.$$

Based on these bounds we can devise two 2-state Markov chains with different values of u (s is always fixed) so that the probability of having a breakpoint can be bounded. A good approximation of the expected breakpoint distance can then be obtained by taking the product of n with the average of the two probabilities of having a breakpoint.

Theorem 1.7 *(From [40]) Assume the genome is signed and circular, and the evolutionary model is GNT(w_I, w_T, w_{IT}). Let B_k be the random variable for the presence of a breakpoint between genes 1 and 2 after k rearrangement events. Let*

$$L(k) = s\left(\frac{1-(1-s-u_{_H})^k}{s+u_{_H}}\right), \text{ and}$$
$$H(k) = s\left(\frac{1-(1-s)^k}{s}\right) = 1-(1-s)^k.$$

Then for any integer $k \geq 0$, $L(k) \leq \Pr(B_k = 1) \leq H(k)$. The function

$$F(k) = \frac{n}{2}\left(L(k) + H(k)\right)$$

provides an approximation of the expected breakpoint distance between G_0 and G_k with small absolute and relative error:

$$|F(k) - E[d_{BP}(G_0, G_k)]| = O(1), \text{ and}$$
$$\phi^{-1} \leq \frac{F(k)}{E[d_{BP}(G_0, G_k)]} \leq \phi,$$

where $\phi = 1 + O(\frac{1}{n})$.

Summary We summarize the `Approx-IEBP` distance as follows.

Definition 1.8 *Assume the evolutionary model is GNT(w_I, w_T, w_{IT}). Let G and G' be two genomes with genes $\{1, 2, \ldots, n\}$. Let $d = d_{BP}(G, G')$ be the breakpoint distance between G and G'. Let F be the function defined in Theorem 1.7. The `Approx-IEBP` distance is the nonnegative integer k minimizing $|F(k) - d|$:*

$$\textit{Approx-IEBP}(G, G') = \operatorname*{argmin}_{\text{integer } k \geq 0} |F(k) - d|.$$

Thus, `Approx-IEBP` is a method of moments estimator which estimates the actual number of rearrangement events between two genomes in the GNT model. Like `Exact-IEBP`, it requires values for w_I, w_T and w_{IT}.

Let m be the number of genomes, and n be the number of genes in each genome. Computing the breakpoint distance matrix takes $O(m^2 n)$ time total. To compute the `Approx-IEBP` distance matrix, we invert $F(k)$, the estimate of the expected breakpoint distance in `Approx-IEBP`, for each pairwise breakpoint distance between two genomes. Computing $F(k)$ takes constant time for each k. To invert $F(k)$ for each pairwise breakpoint distance (as a method of moments estimator requires) we use binary search, which takes $O(\log n)$ time (we assume the number of rearrangement events never exceed $3n$). Because there are at most n different breakpoint distance values, computing the `Approx-IEBP` distance matrix when the breakpoint distance is known takes $O(m^2 + \min\{n, m^2\} \log n)$ time.

1.5.3 *The variance of the breakpoint and `IEBP` distances*

In this section, we show how to calculate the variance of the breakpoint distance, so that we can use `IEBP` with methods such as Weighbor.

The variance of the breakpoint distance To estimate the variance of the breakpoint distance, we have to examine at least two breakpoints at the same time. To use a straightforward approach like `Exact-IEBP` we have to analyze a Markov chain with $O(n^3)$ states, where n is the number of genes in each genome. However, if we are willing to relax the model a bit, we can get a good approximation of the variance, and in fact of all the moments of the breakpoint distance under the Generalized Nadeau-Taylor model, through the use of a "box model." We present this box model here (see [39] for the full details).

Assume all genomes are circular, and that the genome before random rearrangements occur is $(1, \ldots, n)$. Note that if the number of genes is sufficiently large, once the breakpoint between genes i and $i+1$ is created, it is unlikely that a later rearrangement event will bring the two genes back together.

We let $G' = G_k$ denote the genome obtained by k rearrangement events. As k increases, G' changes, and so new breakpoints appear in G with respect to G'. We will let each box represent the presence of a breakpoint in G relative to G'. Thus, for $i = 1, 2, \ldots, n-1$, box i will be empty if there is no breakpoint in G between genes i and $i+1$, and non-empty otherwise. We let box n indicate the presence or absence of a breakpoint between n and 1.

The box model for the inversion-only scenario To illustrate the box model, we begin with the $\mathrm{GNT}(1, 0, 0)$ model in which only inversions occur. We start with n empty boxes, and repeat the following procedure k times. In each iteration we choose two distinct boxes (since an inversion creates two breakpoints). For each box chosen, if the box is empty, we put a ball in it, and otherwise we do not change anything. We let b_k denote the number of nonempty boxes obtained after k iterations. Under our assumption that breakpoints do not disappear, this is an estimate of the number of breakpoints produced by k random inversions.

Let

$$S(x_1, x_2, \ldots, x_n) = \frac{x_1x_2 + x_1x_3 + \cdots + x_{n-1}x_n}{\binom{n}{2}}.$$

Consider $S^k(x_1, x_2, \ldots, x_n)$, the expansion of S to the k^{th} power. Each term in the expansion corresponds to a particular combination of choosing two boxes k times so the total number of times box i is chosen is the power of x_i, for each $1 \leq i \leq n$. The coefficient of that term is the total probability of these ways. For example, the coefficient of $x_1^3x_2x_3^2$ in S^k (when $k = 3$) is the probability of choosing box 1 three times, box 2 once, and box 3 twice. Let $u_{i,k}$ be the sum of the coefficients of all terms taking the form $x_1^{a_1}x_2^{a_2}\cdots x_i^{a_i}$ ($a_j > 0$, $1 \leq j \leq i$), in the expansion of S^k. Then $\binom{n}{i}u_{i,k}$ is the probability i boxes are nonempty after k iterations. This is due to the symmetry in S, in the sense that S is not changed by permuting $\{x_1, x_2, \ldots, x_n\}$. Let A_j be the value of S when we make the following substitutions: $x_1 = x_2 = \cdots = x_j = 1$ and $x_{j+1} = x_{j+2} = \cdots = x_n = 0$. For integers j, $0 \leq j \leq n$, we have

$$\sum_{i=0}^{j} \binom{j}{i} u_{i,k} = S^k(\underbrace{1, 1, 1, \ldots, 1}_{j \ 1's}, 0, \ldots, 0) = A_j^k.$$

Let

$$Z_a = \sum_{i=0}^{n} i(i-1)\cdots(i-a+1)\binom{n}{i}u_{i,k} = \sum_{i=a}^{n} n(n-1)\cdots(n-a+1)\binom{n-a}{i-a}u_{i,k}$$

for all a, $1 \leq a \leq n$. However, each Z_a can be represented as a linear combination of A_i, for $0 \leq i \leq n$. To obtain the variance of b_k we only need Z_1 and Z_2.

Lemma 1.9

(a) $Z_1 = nu_{1,k} = n(A_n^k - A_{n-1}^k)$.
(b) $Z_2 = n(n-1)u_{2,k} = n(n-1)(1 - 2A_{n-1}^k + A_{n-2}^k)$.

We then have the following theorem.

Theorem 1.10 *[39] Let b_k be the number of nonempty boxes in the box model after k iterations. The expectation and variance of b_k are*

$$E[b_k] = n(1 - A_{n-1}^k),$$
$$Var[b_k] = nA_{n-1}^k - n^2A_{n-1}^{2k} + n(n-1)A_{n-2}^k,$$

where

$$A_{n-1} = 1 - \frac{2}{n}, \text{and}$$
$$A_{n-2} = \frac{(n-3)(n-2)}{n(n-1)}.$$

Proof The first identity follows immediately from the fact that $E[b_k] = Z_1$ and that $A_n = 1$. To prove (b), note

$$\begin{aligned}
E[b_k(b_k-1)] &= Z_2 = n(n-1)(1-2A_{n-1}^k + A_{n-2}^k) \\
\Rightarrow E[b_k^2] &= E[b_k(b_k-1)] + E[b_k] = Z_2 + Z_1 \\
&= n(n-1)(1-2A_{n-1}^k + A_{n-2}^k) + n(1-A_{n-1}^k) \\
&= n^2 - n(2n-1)A_{n-1}^k + n(n-1)A_{n-2}^k \\
\Rightarrow \mathrm{Var}[b_k] &= E[b_k^2] - (E[b_k])^2 \\
&= n^2 - n(2n-1)A_{n-1}^k + n(n-1)A_{n-2}^k - n^2(1-A_{n-1}^k)^2 \\
&= nA_{n-1}^k - n^2A_{n-1}^{2k} + n(n-1)A_{n-2}^k
\end{aligned}$$

□

A natural idea is to use A_{n-1} as an estimate of the expected breakpoint distance in computing `IEBP`. The estimate is quite accurate when n is large, though unlike `Approx-IEBP` the formula does not have provable error bounds.

The box model for the general case. Though we assumed only inversions occur in the derivation of Theorem 1.10, it is only reflected in our definition of S. The derivation of Theorem 1.10 only requires S is symmetric, i.e. that S is not changed when we permute $x_1, \ldots, x_n$. Therefore, it is easy to extend the result to the general case, i.e., to $\mathrm{GNT}(w_I, w_T, w_{IT})$: at each iteration, with probability w_I we choose two boxes, and with probability $w_T + w_{IT}$ we choose three boxes (since each transposition and inverted transposition creates at most three breakpoints). Therefore, we can prove the following generalization:

Corollary 1.11 *[39] Let b_k be the number of nonempty boxes in the box model after k iterations. Assume in each iteration, with probability w_I two boxes are picked at random, and with probability $w_T + w_{IT} = 1 - w_I$ three boxes are picked at random. The expectation and variance of b_k are*

$$\begin{aligned}
E[b_k] &= n(1-A_{n-1}^k), \\
Var[b_k] &= nA_{n-1}^k - n^2A_{n-1}^{2k} + n(n-1)A_{n-2}^k,
\end{aligned}$$

where

$$\begin{aligned}
A_{n-1} &= 1 - \frac{3-w_I}{n}, \text{and} \\
A_{n-2} &= \frac{(n-3)(n-4+2w_I)}{n(n-1)}.
\end{aligned}$$

Proof We set S as follows:

$$S = \frac{w_I}{\binom{n}{2}}\Big(\sum_{1\le i_1<i_2\le n} x_{i_1}x_{i_2}\Big) + \frac{w_T + w_{IT}}{\binom{n}{3}}\Big(\sum_{1\le i_1<i_2<i_3\le n} x_{i_1}x_{i_2}x_{i_3}\Big).$$

The values A_{n-1}, A_{n-2} in Theorem 1.10 are changed according to S.

$$A_{n-1} = \frac{w_I\binom{n-1}{2}}{\binom{n}{2}} + \frac{(w_T + w_{IT})\binom{n-1}{3}}{\binom{n}{3}} = 1 - \frac{3 - w_I}{n},$$
$$A_{n-2} = \frac{w_I\binom{n-2}{2}}{\binom{n}{2}} + \frac{(w_T + w_{IT})\binom{n-2}{3}}{\binom{n}{3}} = \frac{(n-3)(n-4+2w_I)}{n(n-1)}.$$

□

The variance of the IEBP distance We begin by observing that `Exact-IEBP` and `Approx-IEBP` have almost identical performance, and so we will refer to them collectively as `IEBP`.

Let G and G' be two genomes with genes $\{1, 2, \ldots, n\}$. Let $D_b = \texttt{IEBP}(G, G')$ and $J(k) = E[b_k]$ be the expected number of nonempty boxes in the box model. The variance of the `IEBP` distance can be approximated using the delta method (see Section 1.4.2) toegether with the expectation and variance of the box model:

$$\mathrm{Var}[\texttt{IEBP}(G, G')] \simeq (\frac{1}{J'(D_b)})^2 \mathrm{Var}[J(D_b)].$$

1.6 Simulation studies

In this section we report on the accuracy of the various techniques for defining distances between genomes (both the original inversion and breakpoint distances, and also `EDE`, `Exact-IEBP`, and `Approx-IEBP`). All these studies are based upon simulation under the Generalized Nadeau-Taylor model, for various settings of the model parameters. All model trees are drawn from the uniform distribution.

We also report on the accuracy of trees reconstructed using either neighbor joining or Weighbor under these various distances. We test these distance estimators under optimal conditions – where the true model parameters are known – as well as under conditions where the true model parameters are incorrectly specified. We explore performance on datasets containing 40 or 160 genomes (i.e., moderate and large size), and examine performance for both 37 and 120 genes (typical values for mitochondria and chloroplast genomes, respectively).

1.6.1 *Accuracy of the evolutionary distance estimators*

In this section we report on our simulation studies evaluating the performance of the evolutionary distance estimators, by comparison to breakpoint and inversion distances.

In our simulations we see that distances estimated by `Exact-IEBP` and `Approx-IEBP` have almost identical error (there is a slight advantage of `Exact-IEBP` over `Approx-IEBP`, but it is fairly negligible); therefore, we refer to them collectively as `IEBP`.

The results of our simulations show how using either breakpoint and inversion distances is problematic: compared to `IEBP` and `EDE`, breakpoint and inversion distances are highly biased when the number of rearrangement events is large. The inversion distance is a good evolutionary distance estimator when the underlying evolutionary model is inversion-only and the rates of evolution are low

(see Fig. 1.6), but is in general not as accurate as either EDE or IEBP under an inversion-only model.

We also explored the robustness of our estimators by simulating evolution under models other than inversion-only, or by giving incorrect parameter values to IEBP. In these cases we see that all five estimators (BP, INV, EDE, Exact-IEBP, and Approx-IEPB) become less accurate; thus, none of these estimators, including our new ones, is robust to model violations (data not shown).

On the other hand, 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 evolutionary distances. This is because any such matrix is still an additive matrix for the same underlying tree, but with different edge lengths. Therefore, the estimated distances can be evaluated according to whether they scale linearly with the number of events. Our simulations (data not shown) reveal that all the distance estimators initially scale linearly, implying that all are able to reconstruct good trees when the evolutionary rate is low enough (as indicated by the evolutionary diameter in the dataset). Interestingly, each of the three evolutionary distance estimators seem to scale linearly for a long initial range (IEBP more so than EDE), even when their assumptions about the model are violated. The worst with respect to linear scaling is clearly BP, as seen in Fig. 1.6. These observations may suggest that trees reconstructed from breakpoint distances will have the worst accuracy, especially close to saturation, than trees reconstructed from other methods, and that trees reconstructed from IEBP or EDE should have the greatest accuracy.

1.6.2 *Performance of NJ and Weighbor Using IEBP and EDE*

As we saw in the previous section, the best estimator of evolutionary distances is IEBP (whether Approx- or Exact-), but EDE is also quite accurate, and each is more accurate (except under unusual circumstances) than INV and BP. The question we investigate in this section is whether the improvement in accuracy of the distance estimators corresponds to an improvement in the accuracy of the resultant phylogenies, as predicted.

We see that the accuracy of trees computed by neighbor joining using either Exact-IEBP or Approx-IEBP is essentially unchanged, and we similarly see unchanged behavior for Weighbor. Therefore, we will collectively call both distances IEBP. In particular, the results shown in Fig. 1.7 for Exact-IEBP apply to Approx-IEBP as well.

Model tree generation In our simulations we produce model trees under the GNT model with 40 or 160 leaves. These model trees have topologies drawn from the uniform distribution on trees leaf-labeled by $1, 2, \ldots, m$, where $m = 40$ or 160.

For each model tree we must define branch lengths λ_e, where λ_e is the expected number of changes on the edge. We define these branch lengths in two steps: we assign an initial length, and then we scale all edge lengths to obtain a fixed target maximum path length D for the tree. This maximum path length is defined by $\Delta = \max_{ij} D_{ij}$, where $D_{ij} = \sum_{e \in P_{ij}} \lambda_e$ and P_{ij} is the path in T

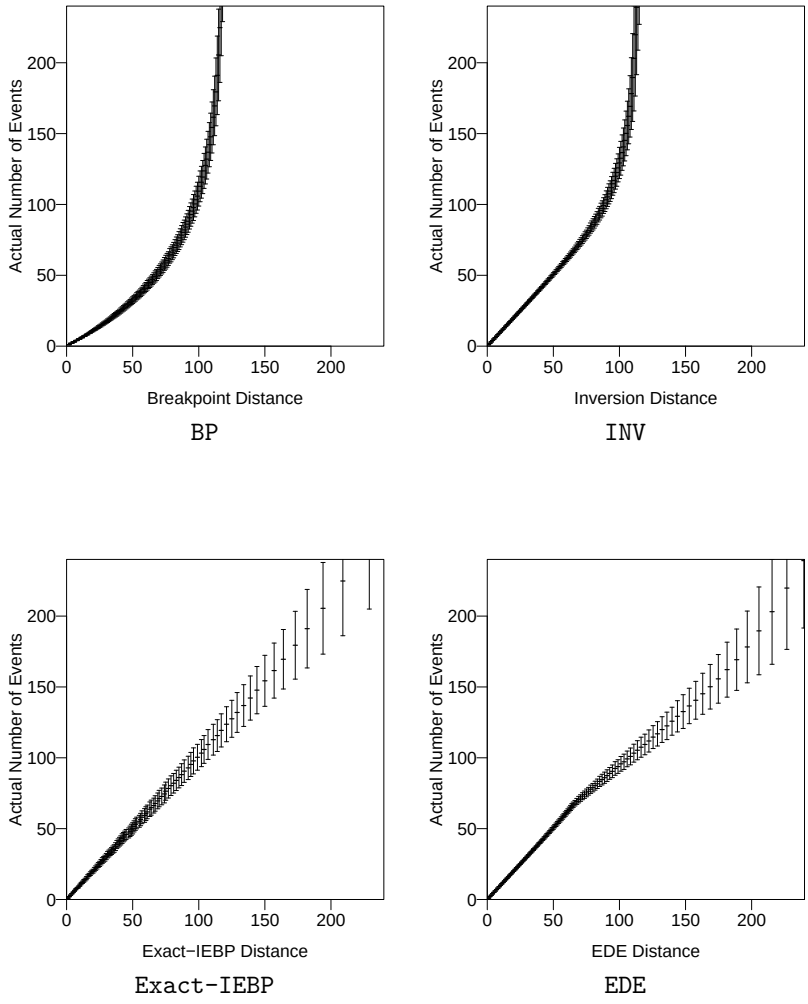


FIG. 1.6. The distribution of genomic distances under the Nadeau-Taylor model (i.e., GNT(1,0,0), or inversion-only evolution). The number of genes is 120, the x-axis is the measured distance, and the y-axis is the actual number of rearrangement events (inversions in this case). For each vertical line, the middle point is the mean, and the top and bottom tips of the line represent one standard deviation away from the mean. In computing `Exact-IEBP` we use correct values of w_I, w_T and w_{IT}.

between leaves i and j. This value Δ is called the "evolutionary diameter" of T. Our initial assignment of lengths is obtained by choosing random positive integers between 1 and 18 for each edge independently. Then, for each target value of Δ, we scale the edge lengths to obtain the desired evolutionary diameter. The target diameters are drawn from $0.1n, 0.2n, 0.4n, 0.8n, 1.6n$, and $3.2n$, where n is the number of genes; these settings result in datasets which have maximum normalized inversion lengths ranging from approximately 0.1 up to almost 1, the maximum possible.

Performance criteria We study the performance of trees reconstructed using these five distances (`BP, INV, EDE, Approx-IEBP,` and `Exact-IEBP`). We used neighbor joining [29], the most frequently used distance-based method, and Weighbor [8], for comparative purposes. We evolved genomes down different GNT model trees, using different values for w_I, w_T and w_{IT}, thus producing synthetic data (genomes) at the leaves of the trees. During each run, we noted which edges of the model tree have had no events on them (these are the "zero-event" edges); these edges are not included in the comparison to the reconstructed trees. We then computed distances between the genomes, using the five different distance estimators. (Since `IEBP` requires values for w_I, w_T, and w_{IT}, in order to test robustness we included incorrect as well as correct values for these parameters.) Each distance matrix was then given to neighbor joining and Weighbor, thus producing trees for each matrix. These trees were then compared to the *true tree* (the model tree minus the zero-event edges) for topological accuracy.

This accuracy was measured as follows. Each edge e in a tree T defines a bipartition $\pi_e = A|B$ on the leaves of T in the obvious way (deleting e separates S into two sets A and B); we let $C(T) = \{\pi_e : e \in E(T)\}$. However, we do not include zero-event edges in the character encoding. Similarly we can define the set $C(T')$, where T' is the inferred tree. The set of *false positives* is $C(T') - C(T)$, and the set of *false negatives* is $C(T) - C(T')$. The false negative and false positive *rates* are obtained by dividing the number of false negatives and false positives, respectively, by $n-3$ (the number of internal edges in a binary tree on n leaves). The false negative rate is informative of the true tree edges that are found in the inferred tree (i.e., the true positive rate). A low false negative rate does not indicate that the inferred tree obtained is highly resolved and close to the true tree, but only that it does not miss many edges in the true tree. Therefore, when the true tree has very low resolution, a low false positive rate is not indicative of a highly resolved accurate inferred tree. The false negative rate will be most significant when the true tree is close to fully resolved, i.e., when the datasets are close to saturation. Our experiments examine performance under all rates of evolution, but the performance under higher rates of evolution allows us to observe whether tree reconstruction can be done accurately when every edge is expected to have changes on it.

Results In Figs. 1.7 and 1.8 we present a sample of the simulation study, showing the accuracy of neighbor joining and Weighbor trees constructed using the different distance estimators.

Our model trees have 160 leaves, and we evolve genomes with 120 genes down the model trees. The model conditions include both an inversion-only scenario (GNT$(1, 0, 0)$) and a scenario with half inversions and half transpositions/inverted transpositions (GNT$(.5, .25, .25)$).

We gave `IEBP` correct parameter values for w_I, w_T, w_{IT} in this experiment. The model trees have rates of evolution that range from low to almost saturated, as indicated by the x-axis which measures the normalized maximum inversion distance in the dataset. For each experimental setting, we bin the datasets according to their diameters (maximum pairwise inversion distance between any two genomes). The $x-$ and y-axis coordinates of each point in the figure are the average diameter and average false negative rates of the corresponding bin, respectively.

False positive rates Trees returned by neighbor joining or Weighbor are always binary. However, since true trees may not be binary (due to the presence of zero-event edges), some false positive edges may be artifacts. In fact, in our experiments, except when quite close to saturation, the true tree will in general be quite unresolved. As a result, any reconstruction method that always returns binary trees will necessarily have a high false positive rate, since the false positive rate must be at least as high as the percentage of edges missing in the true tree. However, in our experiments we see that the false positive rates we obtain generally are not much higher than the percentage of missing edges, indicating quite good performance (see Fig. 1.7).

False negative rates We see clearly from Fig. 1.8 that for extremely low evolutionary diameters, all methods can reconstruct a good estimate of the true tree, but as the diameter increases, the false negative rates increase for all methods. We also see that overall NJ(`BP`) has the worst performance, and that Weighbor(`IEBP`) is generally inferior to the other methods (even when it is given the correct parameter values, for a reason we do not understand). On the other hand, Weighbor(`EDE`) is extremely accurate, even when the model condition is not inversion-only. Second best is NJ(`EDE`), which is also quite accurate even when the model condition is not inversion-only. Thus, although we saw that `EDE` is not robust to model violations with respect to estimating distances correctly, its apparent linear scaling with the actual distance makes it a good technique for phylogeny reconstruction.

Some of the other trends are also worth noting:

1. As the number of genes increases, the inferred trees become more accurate, at all evolutionary diameters (data not shown). Thus, inferring phylogenies from chloroplast genomes (which contain on average 120 genes) is more reliable than inferring phylogenies from mitochondrial genomes (which contain on average 37 genes).

2. As the number of taxa increases, the inferred trees become less accurate, at all evolutionary diameters (data not shown).

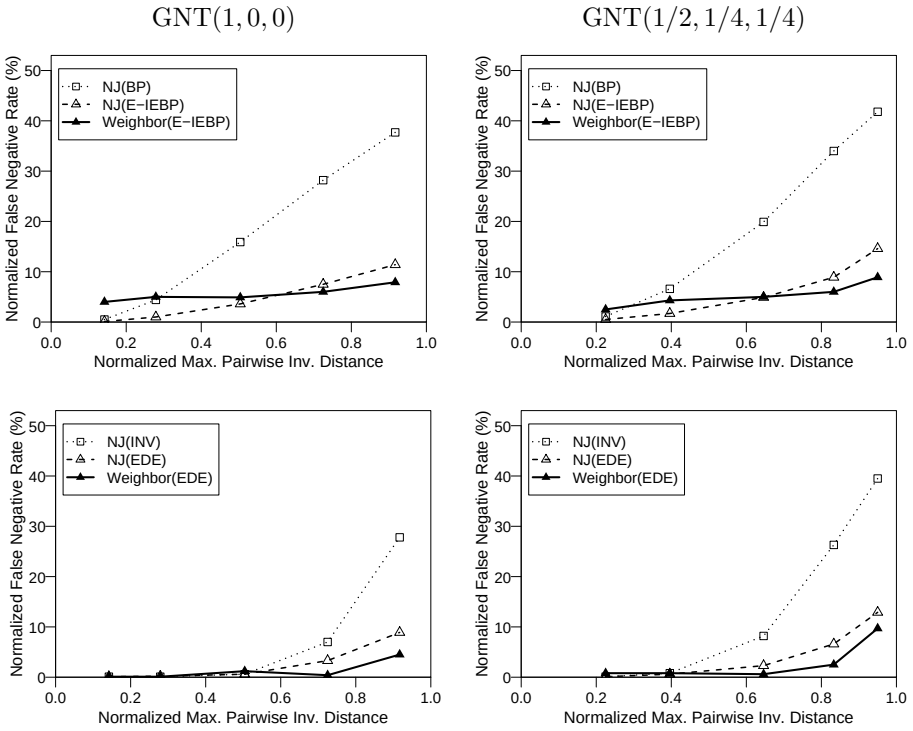


FIG. 1.7. Simulation study of false negative rates of distance-based tree reconstruction methods on 160 circular genomes with 120 genes: (Top) Breakpoint distance based methods, (Bottom) Inversion distance based methods. The x-axis is the normalized diameter (maximum inversion distance between all pairs of genomes) of the dataset, and the y-axis is the false negative rate. The model of evolution is (left) the Nadeau-Taylor model (i.e., GNT$(1, 0, 0)$), or (right) the GNT model with half inversions, one-fourth transpositions and one-fourth inverted transpositions (i.e., GNT$(1/2, 1/4, 1/4)$). In computing `Exact-IEBP` we use correct values of w_I, w_T and w_{IT}.

3. Neighbor joining trees are more accurate when based upon corrected distances (`IEBP` or `EDE`) than uncorrected distances (breakpoint or inversion distance). The distinction is the greatest when the dataset has a high evolutionary diameter (i.e., when the dataset contains some pair of genomes that look almost random with respect to each other).

4. NJ(`IEBP`) and Weighbor(`IEBP`) perform comparably with incorrect values for the parameters as with correct values; however, Weighbor(`IEBP`) is not particularly accurate, and neither is as good as

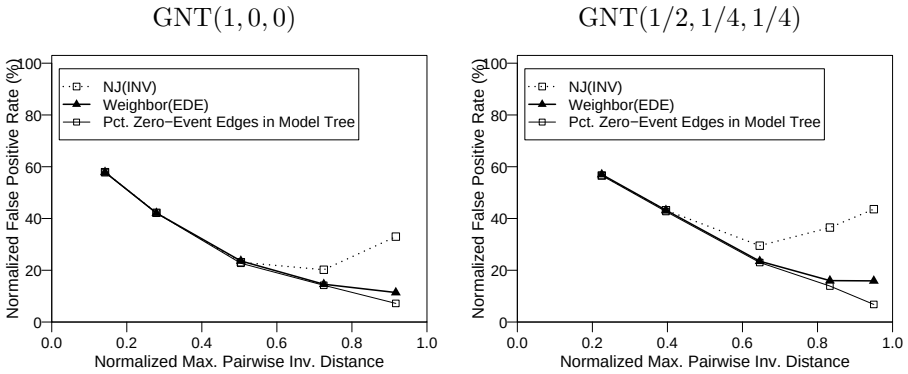


FIG. 1.8. Simulation study of the false positive rates of NJ(`INV`) and Weighbor(`EDE`) on 160 circular genomes with 120 genes. We do not include the false positive rates of NJ(`EDE`) because the curve is very close to that of Weighbor(`EDE`). The x-axis is the normalized diameter (maximum inversion distance between all pairs of genomes) of the dataset, and the y-axis is the false positive rate. The model of evolution is (left) the Nadeau-Taylor model (i.e., GNT$(1, 0, 0)$), or (right) the GNT model with half inversions, one-fourth transpositions and one-fourth inverted transpositions (i.e., GNT$(1/2, 1/4, 1/4)$). Refer to Section 1.6.1 for how these figures are generated.

NJ(`EDE`) or Weighbor(`EDE`).

5. In general, Weighbor(`EDE`) seems to provide better estimates of evolutionary history than all other methods we examined, especially when the number of genomes and genes are large, and the evolutionary rate is high, but NJ(`EDE`) is a close second. Both give highly accurate estimations of phylogenies even when the model is not inversion-only.

These observations are specifically for the uniform tree topology case, but most of them hold for other models, including birth-death trees generated by the r8s program [30]. In particular, Weighbor(`EDE`) is still the most accurate of these methods.

We conclude this section with the following observation. Perhaps the most significant indicator of the difficulty of a dataset is its evolutionary diameter: if the diameter is low, all methods will get a good estimate of the tree, even if the distance estimation is based upon incorrect assumptions, but for the largest diameters (approaching saturation), only Weighbor and NJ on `EDE` distances are reliably accurate.

1.7 Summary

We have shown that statistically-based estimations of evolutionary distances can be quite robust to some model violations, and can help make phylogeny

reconstructions much more accurate – especially when the dataset is close to saturation. However, one of the interesting observations to come out of our experiments is that the accuracy of a phylogeny reconstruction is *usually, but not always* improved by having a better estimate of the evolutionary distance. For example, NJ(`EDE`) gives better estimates of trees than NJ(`IEBP`), although `IEBP` gives more accurate estimates of distances than `EDE`. Clearly, the interplay between phylogeny reconstruction methods themselves, and the distance estimates, cannot be simply summarized and explained.

Several problems for the Generalized Nadeau-Taylor model are still open. First, the distribution of the inversion distance is still unknown, as are its expectation and variance. Results along these lines will help us understand why neighbor joining based upon the inversion distance gives better results in phylogeny reconstruction than neighbor joining based upon the breakpoint distance. Also several studies suggest minimum evolution methods also produce highly accurate trees (see Chapter 1 in this Volume) for DNA sequence evolution. It will be interesting to see whether minimum evolution methods produce accurate trees for gene order data.

A maximum likelihood approach for genome rearrangement phylogeny estimation is another approach that will be interesting to explore. MCMC methods are also interesting, but have not been able to scale to reasonable dataset sizes [18]. Maximum likelihood distance estimation is another interesting area to investigate, and it is unknown if the method of moments estimators used for correcting breakpoint and inversion distances are maximum likelihood distance estimators. Another challenging problem is to estimate w_I, w_T, and w_{IT} from the data.

The models we have studied have all presumed that evolutionary events occur with probabilities that only depend upon the type of event. Therefore, a main research question is to explore the estimation of evolutionary distances under newer models of genome evolution. Such models might assume that the probability of the rearrangement events may depend upon the lengths of the affected segments (see [26] for one such model), or may make other assumptions that incorporate hotspots or break the chromosome into distinct regions and require events to stay within these regions [37]. Also of interest are models which allow for deletions, duplications, and other events which change the gene content and not just the gene order. Calculations of distances in these models are much more complicated; initial results along these lines have been obtained by El-Mabrouk, Moret, and others (see [11–14, 19, 34] and Chapters 11 and 12 in this Volume). Similarly, models which handle multiple chromosomes, and which allow for translocations, need to be considered, and there is much less that has been established for this multi-chromosomal case than for the unequal gene content case [35, 36].

Finally, as we have noted, the reconstructions of trees we obtain can have a high false positive rate, due to the high incidence of zero-event edges in the model tree (and hence low resolution in the true tree). Determining which edges

in the reconstructed tree are valid, and which are not, is a general problem facing phylogenetic analysis. In DNA systematics, bootstrapping and other techniques can be used to assess the confidence in a given edge, and so potentially identify the false positive edges. However, in gene order phylogeny it is not possible to perform bootstrapping, since there is only one character. Consequently, other techniques would need to be used to identify false positives.

One potential approach would be to use `GRAPPA` (see [21], and also Chapter 12 in this Volume) to try to identify the false positives, as follows. First we could assign genomes (i.e., signed circular permutations of $(1, 2, \ldots, n)$) to internal nodes in order to minimize the total number of events on the tree, and then we could contract all edges that are assigned the same genomes at the endpoints. Such a technique might be able to identify edges on the tree that have no events on them, but is most likely to succeed when the reconstructed tree is a refinement of the true tree. In our experiments, since the false negative rate is either 0 or close to 0, this would be the case. Future research will investigate this as a potential second phase in the phylogenetic analysis.

1.8 Acknowledgments

The authors would like to thank the two anonymous reviewers for their very helpful criticism. This research was supported by National Science Foundation grants EIA-0121680, EF-0331453, DEB-0120709, and IIS-0113654. The first author was supported in part by a NIH Training Grant in Cancer and Immunopathobiology (1 T32 CA101968). The second author 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

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

[2] Bader, D.A., Moret, B.M.E., and 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.

[3] Bailey, J.A., Baertsch, R., Kent, W.J., Haussler, D., and Eichler, E.E. (2004). Hotspots of mammalian chromosomal evolution. *Genome Biol. 5*(4), R23. Epub 2004 Mar 08.

[4] Blanchette, M., Bourque, G., and Sankoff, D. (1997). Breakpoint phylogenies. In *Genome Informatics* (ed. S. Miyano and T. Takagi), pp. 25–34. Univ. Acad. Press.

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

[6] Boore, J. L., Collins, T. M., Stanton, D., Daehler, L. L., and Brown, W. M. (1995). Deducing arthropod phylogeny from mitochondrial DNA rearrangements. *Nature*, **376**, 163–165.

[7] Bourque, G., Pevzner, P.A., and Tesler, G. (2004). Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res. 14*(4), 507–16.

[8] Bruno, W.J., Socci, N.D., and Halpern, A.L. (2000). Weighted Neighbor Joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.*, **17**, 189–197.

[9] Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Duxbury Press.

[10] Downie, S.R. and Palmer, J.D. (1992). Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In *Molecular Systematics of Plants* (ed. P. Soltis, D. Soltis, and J. Doyle), Volume 49, pp. 14–35. Chapman & Hall.

[11] El-Mabrouk, N. (2001). Sorting signed permutations by reversals and insertions/deletions of contiguous segments. *Journal of Discrete Algorithms 1*(1), 105–122.

[12] El-Mabrouk, N. (2002). Reconstructing an ancestral genome using minimum segments duplications and reversals. *Journal of Computer and System Sciences*, **65**, 442–464.

[13] El-Mabrouk, N. and 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*, Volume 1, pp. 537–550. Kluwer Academic Publishers.

[14] El-Mabrouk, N. and Sankoff, D. (2003). The reconstruction of doubled

genomes. *SIAM Journal on Computing 32*(1), 754–792.

[15] Hannenhalli, S. and Pevzner, P. (1995). Transforming cabbage into turnip (polynomial algorithm for genomic distance problems). In *Proc. 27th Annual ACM Symp. on Theory of Comp. (STOC'95)*, pp. 178–189. ACM Press, NY.

[16] Kaplan, H., Shamir, R., and Tarjan, R.E. (1997). Faster and simpler algorithm for sorting signed permutations by reversals. In *Proc. 8th Annual Symp. on Discrete Alg. (SODA'97)*, pp. 344–351. ACM Press, NY.

[17] Kim, J. and Warnow, T. (1999). Tutorial on phylogenetic tree estimation. http://kim.bio.upenn.edu/∼jkim/media/ISMBtutorial.pdf.

[18] Larget, B., Simon, D. L., and Kadane, J.B. (2002). On a Bayesian approach to phylogenetic inference from animal mitochondrial genome arrangements. *J. Royal Stat. Society B 64*(4), 681–693.

[19] Marron, M., Swenson, K.M., and Moret, B.M.E. Genomic distances under deletions and insertions. *Theoretical Computer Science*. To appear (special issue on the best papers from COCOON'03).

[20] Moret, B.M.E., Wang, L.-S., Warnow, T., and Wyman, S. (2001*a*). New approaches for reconstructing phylogenies based on gene order. In *Proc. 9th Intl. Conf. on Intel. Sys. for Mol. Bio. (ISMB'01)*, pp. 165–173.

[21] Moret, B.M.E, Wyman, S.K., Bader, D.A., Warnow, T., and Yan, M. (2001*b*). A new implementation and detailed study of breakpoint analysis. In *Proc. 6th Pacific Symp. Biocomputing (PSB'01)*, pp. 583–594.

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

[23] Oehlert, G. W. (1992). A note on the delta method. *Amer. Statist.*, **46**, 27–29.

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

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

[26] Pinter, R.Y. and Skiena, S. (2002). Genomic sorting with length-weighted reversals. *Genome Informatics*, **13**, 103–111.

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

[28] Rokas, A. and Holland, P. W. H. (2000). Rare genomic changes as a tool for phylogenetics. *Trends in Ecology and Evolution*, **15**, 454–459.

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

[30] Sanderson, M. J. Analysis of rates (r8s) of evolution. v 1.6.

[31] Sankoff, D. (2003). Rearrangements and chromosomal evolution. *Curr. Opin. Genet. Dev. 13*(6), 583–7.

[32] Sankoff, D. and Blanchette, M. (1999). Probability models for genome rearrangements and linear invariants for phylogenetic inference. *Proc. 3rd Int'l*

Conf. on Comput. Mol. Bio. (RECOMB'99), 302–309.

[33] Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy.* W.H. Freeman & Co.

[34] Tang, J. and Moret, B.M.E. (2003). Phylogenetic reconstruction from gene rearrangement data with unequal gene contents. In *Lecture Notes in Computer Science No. 2748: Proc. 8th Workshop on Algs. and Data Structs. (WADS'03)*, pp. 37–46.

[35] Tesler, G. (2002*a*). Efficient algorithms for multichromosomal genome rearrangements. *Journal of Computer and System Sciences 65*(3), 587–609.

[36] Tesler, G. (2002*b*). GRIMM: genome rearrangements web server. *Bioinformatics 18*(3), 492–493.

[37] Tesler, G. and Pevzner, P. (2003). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Nat'l Acad. Sci. USA 100*(13), 7672–7677.

[38] Wang, L.-S. (2001). Improving the accuracy of evolutionary distances between genomes. In *Lec. Notes in Comp. Sci.: Proc. 1st Workshop for Alg. & Bio. Inform. (WABI'01)*, pp. 175–188. Springer Verlag.

[39] Wang, L.-S. (2002). Genome rearrangement phylogeny using Weighbor. In *Lec. Notes in Comp. Sci.: Proc. 2nd Workshop for Alg. & Bio. Inform. (WABI'02)*, pp. 112–125. Springer-Verlag.

[40] Wang, L.-S. and Warnow, T. (2001). Estimating true evolutionary distances between genomes. In *Proc. 33th Annual ACM Symp. on Theory of Comp. (STOC'01)*, pp. 637–646. ACM Press.

[41] Waterman, M. S., Smith, T. F., Singh, M., and Bayer, W. A. (1977). Additive evolutionary trees. *J. Theor. Biol.*, **64**.

[42] Zaretskii, K. (1965). Constructing a tree on the basis of a set of distance between the hanging vertices. *Uspekhi Mat. Nauk.*, **20**, 90–92. (in Russian).

INDEX