

Genome Rearrangement Phylogeny Using Weighbor

Li-San Wang

Department of Computer Sciences, University of Texas
Austin, TX 78712 USA
lisan@cs.utexas.edu

Abstract. Evolution operates on whole genomes by operations that change the order and strandedness of genes within the genomes. This type of data presents new opportunities for discoveries about deep evolutionary rearrangement events. Several distance-based phylogenetic reconstruction methods have been proposed [12, 21, 19] that use neighbor joining (NJ) [16] with the expected breakpoint or inversion distances after k rearrangement events. In this paper we study the variance of the breakpoint and inversion distances. The result is combined with Weighbor [5], an improved version of NJ using the variance of true evolutionary distance estimators, to yield two new methods, Weighbor-IEBP and Weighbor-EDE. Experiments show the new methods have better accuracy than all previous distance-based methods, and are robust against model parameter misspecifications.

1 Background

Distance-based phylogenetic reconstruction. A (phylogenetic) tree T on a set of taxa S is a tree representation of the evolutionary history of S : T is a tree leaf-labeled by S , such that the internal nodes reflect past speciation events. Given a tree T on a set S of genomes and given any two leaves i, j in T , we denote by $P(i, j)$ the path in T between i and j . We let λ_e denote the number of evolutionary events on the edge e during the evolution of the genomes in S within the tree T ; we call it the *true evolutionary distance* between the two endpoints. We can then define the matrix of true evolutionary distances, $d_{ij} = \sum_{e \in P(i, j)} \lambda_e$, which is additive (a distance D is additive if it satisfies the *four-point condition*: for every distinct four points $\{i, j, l, m\}$, $D_{ij} + D_{lm} \leq \max\{D_{il} + D_{jm}, D_{im} + D_{jl}\}$). Given an additive matrix, many distance-based methods are guaranteed to reconstruct the tree T and the edge weights.

Neighbor joining and its variants BioNJ and Weighbor. Neighbor joining (NJ) is the most popular distance-based tree inference method. The input to the method is a matrix of estimated leaf-to-leaf distances $\{D_{ij} | 1 \leq i, j \leq n\}$ on n leaves. Starting with these leaves as one-node subtrees, the algorithm creates new subtrees iteratively by picking two subtrees using a least-squares criterion of the distances to other roots of the subtrees (the

pairing step), and updates the distances of the root of the new subtree to other roots of the subtrees (the distance update step) according to some least-squares criterion [7].

In reality, we do not get exact distance estimates between leaves due to the random nature of evolution. Atteson showed NJ is guaranteed to reconstruct the true tree topology if the input distance matrix is sufficiently close to a distance matrix defining the same tree topology [1]. Consequently, techniques that yield a good estimate of the matrix $\{d_{ij}\}$ are of significant interest.

In this paper we will use two modified versions of neighbor joining called **BioNJ** [7] and **Weighbor** [5]. Both methods use the variance of the true distance estimators in the pairing (**Weighbor** only) and distance update steps to improve the accuracy of the tree reconstruction.

Genome rearrangement evolution. Modern laboratory techniques yield the ordering and strandedness of genes on a chromosome, allowing us to represent each chromosome by an ordering of signed genes (where the sign indicates the strand). Evolutionary events can alter these orderings through rearrangements such as inversions and transpositions, collectively called genome rearrangements. Because these events are rare, they give us information about ancient events in the evolutionary history of a group of organisms. In consequence, many biologists have embraced this new source of data in their phylogenetic work [6, 14, 15]. Appropriate tools for analyzing such data remain primitive when compared to those developed for DNA sequence data; thus developing such tools is becoming an important area of research.

The genomes of some organisms have a single chromosome or contain single chromosome organelles (such as mitochondria [4] or chloroplasts [14, 15]) whose evolution is largely independent of the evolution of the nuclear genome for these organisms. When each genome has the same set of genes and each gene appears exactly once, a genome can be described by an ordering (circular or linear) of these genes, each gene given with an orientation that is either positive (g_i) or negative ($-g_i$). If the genome is circular, we always represent the genome “linearly” so g_1 is positive and at the first position. A close inspection shows a circular genome with n genes, when represented linearly, is identical to a linear genome with $n - 1$ genomes.

Let G be the genome with signed ordering g_1, g_2, \dots, g_k . An *inversion* between indices a and b , for $a \leq b$, produces the genome with linear ordering $(g_1, g_2, \dots, g_{a-1}, -g_b, -g_{b-1}, \dots, -g_a, g_{b+1}, \dots, g_k)$. A *transposition* on the (linear or circular) ordering G acts on three indices, a, b, c , with $a \leq b$

and $c \notin [a, b]$, picking up the interval g_a, g_{a+1}, \dots, g_b and inserting it immediately after g_c . Thus the genome G above (with the assumption of $c > b$) is replaced by $(g_1, \dots, g_{a-1}, g_{b+1}, \dots, g_c, g_a, g_{a+1}, \dots, g_b, g_{c+1}, \dots, g_k)$. An *inverted transposition* is a transposition followed by an inversion of the transposed subsequence.

The Generalized Nadeau-Taylor Model. The Generalized *Nadeau-Taylor* (GNT) model [21] of genome evolution uses only genome rearrangement events which do not change the gene content. The model assumes that the number of each of the three types of events obeys a Poisson distribution on each edge, that the relative probabilities of each type of event are fixed across the tree, and that events of a given type are equiprobable. Thus we can represent a GNT model tree as a triplet $(T, \{\lambda_e\}, (\alpha, \beta))$, where the (α, β) defines the relative probabilities of transpositions and inverted transpositions (hence an event is an inversion with probability $1 - \alpha - \beta$). For instance, $(\frac{1}{3}, \frac{1}{3})$ indicates that the three event classes are equiprobable, while the pair $(0, 0)$ indicates that only inversions happen.

Estimating true evolutionary distances using genome rearrangements. Let us be given a set of genome rearrangement events with a particular weighting scheme; the weight of a sequence of rearrangement events from this set is the sum of the weights of these events. The *edit distance* between two gene orders is the minimum of the weights of all sequences of events from the given set that transform one gene order into the other. For example, the *inversion distance* is the edit distance when only inversions are permitted and all inversions have weight 1. The inversion distance can be computed in linear time [2], and the transposition distance is of unknown computational complexity [3].

Another type of genomic distance that received much attention in the genomics community is the *breakpoint distance*. Given two genomes G and G' on the same set of genes, a *breakpoint* in G is an ordered pair of genes (g_a, g_b) such that g_a and g_b appear consecutively in that order in G , but neither (g_a, g_b) nor $(-g_b, -g_a)$ appear consecutively in that order in G' . The number of breakpoints in G relative to G' is the *breakpoint distance* between G and G' . The breakpoint distance is easily calculated by inspection in linear time.

Estimating the true evolutionary distance requires assumption about the model; in the case of gene-order evolution, the assumption is that the genomes have evolved from a common ancestor under the Nadeau-Taylor model of evolution. The technique in [4], applicable only to inversions, calculates this value exactly, while **Approx-IEBP** [21] and **EDE** [12], applicable to very general models of evolution, obtain approximations of these

values, and **Exact-IEBP** [19] calculates the value exactly for any combination of inversions, transpositions, and inverted transpositions. These estimates can all be computed in low polynomial time.

Variance of genomic distances. The following problem has been studied in [17, 19]: given any genome G with n genes, what is the expected breakpoint distance between G and G' when G' is the genome obtained from G by applying k rearrangements according to the GNT model? Both papers approach the problem by computing the probability of having a breakpoint between every pair of genes; by linearity of the expectation the expected breakpoint distance can be obtained by n times the aforementioned probability. Each breakpoint can be characterized as a Markov process with $2(n - 1)$ states. But the probability of a breakpoint is a sum of $O(n)$ terms that we do not know yet how to further simplify.

However the variance cannot be obtained this way since breakpoints are not independent (under any evolutionary model) by the following simple observation: the probability of having a breakpoint for each breakpoint position is nonzero, but the probability of the breakpoint distance being 1 is zero (the breakpoint distance is always 0 or at least 2). Thus, to compute the variance (or the second moment) of the breakpoint distance we need to look at two breakpoints at the same time. This implies we have to study a Markov process of $O(n^2)$ states and a sum of $O(n^2)$ terms that is hard to simplify. As for the inversion distance, even the expectation is still an open problem.

Estimating the variance of breakpoint and inversion distances is important for several reasons. Based on these estimates we can compute the variances of the **Approx-IEBP** and **Exact-IEBP** estimators (based on the breakpoint distance), and the **EDE** estimator (based on the inversion distance). It is also informative when we compare estimators based on breakpoint distances to other estimators, e.g. the inversion distance and the **EDE** distance. Finally, variance estimation can be used in distance-based methods to improve the topological accuracy of tree reconstruction.

Outline of this paper. We start in Section 2 by presenting a stochastic model approximating the breakpoint distance, and derive the analytical form of the variance of the approximation, as well as the variance of the **IEBP** estimators. In Section 3 the variance of the inversion and the **EDE** distances are obtained through simulation. Based on these variance estimates we propose four new methods, called **BioNJ-IEBP**, **Weighbor-IEBP**, **BioNJ-EDE**, and **Weighbor-EDE**. These methods are based on **BioNJ** and **Weighbor**, but the variances in these algorithms have been replaced with

the variances of IEBP and EDE. In Section 4 we present our simulation study to verify the accuracy of these new methods.

2 Variance of the Breakpoint Distance

The approximating model. We first define the following notation in this paper: $\binom{a}{b}$ is the number of choosing b objects from a (the binomial coefficient) when $a \geq b \geq 0$; $\binom{a}{b}$ is set to 0 otherwise.

We motivate the approximating model by the case of inversion-only evolution on signed circular genomes. Let n be the number of genes, and b be the number of breakpoints of the current genome G . When we apply a random inversion (out of $\binom{n}{2}$ possible choices) to G , we have the following cases according to the two endpoints of the inversion [10]:

1. None of the two endpoints of the inversion is a breakpoint. The number of breakpoints is increased by 2. There are $\binom{n-b}{2}$ such inversions.
2. Exactly one of the two endpoints of the inversion is a breakpoint. The number of breakpoints is increased by 1. There are $b(n-b)$ such inversions.
3. The two endpoints of the inversion are two breakpoints. There are $\binom{b}{2}$ such inversions. Let g_i and g_{i+1} be the left and right genes at the left breakpoint, and let g_j and g_{j+1} be the left and right genes at the right breakpoint. There are three subcases:
 - (a) None of $(g_i, -g_j)$ and $(-g_{i+1}, g_{j+1})$ is an adjacency in G_0 . The number of breakpoints is unchanged.
 - (b) Exactly one of $(g_i, -g_j)$ and $(-g_{i+1}, g_{j+1})$ is an adjacency in G_0 . The number of breakpoints is decreased by 1.
 - (c) $(g_i, -g_j)$ and $(-g_{i+1}, g_{j+1})$ are adjacencies in G_0 . The number of breakpoints is decreased by 2.

When $b \geq 3$, out of the $\binom{b}{2}$ inversions from case 3, case 3(b) and 3(c) count for at most b inversions; this means given that an inversion belongs to case 3, with probability at least $1 - b/\binom{b}{2} = \frac{b-3}{b-1}$ it does not change the breakpoint distance; this probability is close to 1 when b is large. Furthermore, when $b \ll n$ almost all the inversions belong to case 1 and 2. Therefore, when n is large, we can drop cases 3(b) and 3(c) without affecting the distribution of breakpoint distance drastically.

The approximating model we use is as follows. Assume first the evolutionary model is such that each rearrangement creates r breakpoints on an unrearranged genome (for example, $r = 2$ for inversions and $r = 3$ for transpositions and inverted transpositions). Let us be given n boxes, initially empty. At each iteration r boxes will be chosen randomly (without replacement); we then place a ball into each of these r boxes if it is empty. The number of nonempty boxes after k iterations, b_k , can be used

to estimate the number of breakpoints after k rearrangement events are applied to an unrearranged genome. This model can also be extended to approximate the GNT model: at each iteration, with probability $1 - \alpha - \beta$ we choose 2 boxes, and with probability $\alpha + \beta$ we choose 3 boxes.

Mean and variance of the approximating model. Fix n (the number of boxes) and k (the number of times we choose r boxes). Consider the expansion of the following expression

$$S = ((x_1x_2 + x_1x_3 + \cdots + x_{n-1}x_n)/\binom{n}{2})^k.$$

Each term corresponds to the number of ways of choosing $r = 2$ boxes for k times where the total number of times box i is chosen is the power of x_i , and 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 (when $k = 6$) is the probability of choosing box 1 three times, box 2 once, and box 3 twice. Let u_i be the coefficient of the terms with i distinct symbols; $\binom{n}{i}u_i$ is the probability i boxes are nonempty after k iterations. The identity of u_i for all terms of the same set of power indices holds as long as the probability of each box being chosen is identical; in other words, S is not changed by permuting $\{x_1, x_2, \dots, x_n\}$ arbitrarily.

To solve for u_i exactly for all k is difficult and unnecessary. Instead we can find the expectation and variance of b_k directly. Actually the following results give all moments of b_k . Let $S(a_1, a_2, \dots, a_n)$ be the value of S when we substitute x_i by a_i , $1 \leq i \leq n$, and let S_j be the value of S when $a_1 = a_2 = \cdots = a_j = 1$ and $a_{j+1} = a_{j+2} = \cdots = a_n = 0$. For integers j , $0 \leq j \leq n$, we have

$$\sum_{i=0}^j \binom{j}{i} u_i = S(\underbrace{1, 1, 1, \dots, 1}_{j \text{ 1's}}, 0, \dots, 0) = S_j.$$

Let

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

for all a , $1 \leq a \leq n$. We want to express Z_a by some linear combination of S_i , $0 \leq i \leq n$. The following lemma, which is a special case of equation (5.24) in [9], finds the coefficients of the linear combination.

Lemma 1. *Let a be some given integer such that $1 \leq a \leq n$. Let us be given $\{u_i : 0 \leq i \leq n\}$ that satisfy $\sum_{j=0}^i \binom{i}{j} u_j = \sum_{j=0}^n \binom{i}{j} u_j = S_i$, where $0 \leq i \leq n$. We have $\sum_{i=n-a}^n (-1)^{n-i} \binom{a}{n-i} S_i = \sum_{j=0}^n \binom{n-a}{j-a} u_j$.*

The following results follow from Lemma 1; we state them without proof due to space limitations.

Theorem 1. For all a , $1 \leq a \leq n$,

$$Z_a = n(n-1) \cdots (n-a+1) \sum_{i=n-a}^n (-1)^{n-i} \binom{a}{n-i} S_i.$$

Corollary 1. (a) $Eb_k = Z_1 = n(1 - S_{n-1})$.

(b) $\text{Var } b_k = nS_{n-1} - n^2 S_{n-1}^2 + n(n-1)S_{n-2}$.

These results work for all integers r , $1 \leq r \leq n$. When there are more than one type of rearrangement events with different r 's we can change S accordingly. For example, let $\gamma = \alpha + \beta$; for the GNT model we can set

$$S = \left(\frac{1-\gamma}{\binom{n}{2}} \left(\sum_{1 \leq i_1 < i_2 \leq n} x_{i_1} x_{i_2} \right) + \frac{\gamma}{\binom{n}{3}} \left(\sum_{1 \leq i_1 < i_2 < i_3 \leq n} x_{i_1} x_{i_2} x_{i_3} \right) \right)^k. \quad (1)$$

Mean and variance of the breakpoint distance under the GNT model. We begin this section by finding the mean and variance for b_k with respect to the GNT model. By substituting into equation (1):

$$S_{n-1} = \left(1 - \frac{2+\gamma}{n}\right)^k, S_{n-2} = \left(\frac{(n-3)(n-2-2\gamma)}{n(n-1)}\right)^k.$$

For the GNT model, we have the following results:

$$\frac{d}{dk} Eb_k = -nS_{n-1} \left(\frac{1}{k} \ln S_{n-1}\right) = -nS_{n-1} \ln\left(1 - \frac{2+\gamma}{n}\right) \quad (2)$$

$$\begin{aligned} \text{Var } b_k &= nS_{n-1} - n^2 S_{n-1}^2 + n(n-1)S_{n-2} \\ &= (nS_{n-1} - n^2 S_{n-1}^2 + n(n-1)S_{n-2}) \end{aligned} \quad (3)$$

Using BioNJ and Weighbor. Both **BioNJ** and **Weighbor** are designed for DNA sequence phylogeny using the variance of the true evolutionary distance estimator. In **BioNJ**, the distance update step of NJ is modified so the variances of the new distances are minimized. In **Weighbor**, the pairing step is also modified to utilize the variance information. We use the variance for the GNT model in this section and the expected breakpoint distance in [19] in the two methods. The new methods are called **BioNJ-IEBP** and **Weighbor-IEBP**. To estimate the true evolutionary distance, we use **Exact-IEBP**, though we can also use Equation (2) which is less accurate. Let $\hat{k}(b)$ denote the **Exact-IEBP** distance given the breakpoint distance is b ; $\hat{k}(b)$ behaves as the inverse of Eb_k , the expected breakpoint distance

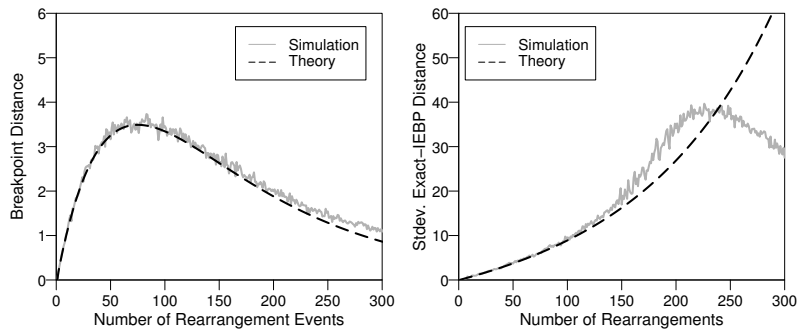


Fig. 1. Accuracy of the estimator for the variance. Each figure consists of two sets of curves, corresponding to the values of simulation and theoretical estimation. The number of genes is 120. The number of rearrangement events, k , range from 1 to 220. The evolutionary model is inversion-only GNT; see Section 1 for details. For each k we generate 500 runs. We then compute the standard deviation of b_k for each k , and those of $\hat{k}(b_k)$ for each k , and compare them with the values of the theoretical estimation.

after k rearrangement events. The variance of $\hat{k}(b)$ can be approximated using a common statistical technique, the delta method [13], as follows:

$$\text{Var } \hat{k}(b) \simeq \left(\frac{d}{dk} E b_k\right)^{-2} \text{Var } b_k = \frac{\left(1 - nS_{n-1} + (n-1)\left(\frac{S_{n-2}}{S_{n-1}}\right)\right)}{nS_{n-1}\left(\ln\left(1 - \frac{2+\gamma}{n}\right)\right)^2}.$$

When the number of rearrangements are below the number of genes (120 in the simulation), these results are accurate approximations to the mean and variance of the breakpoint distance under the GNT model, as the simulation in Figure 1 shows. As the number of rearrangements k is so high the breakpoint distance is close to the maximum (the resulting genome is random with respect to the genome before evolution), the simulation shows the variance is much lower than the theoretical formula. This is due to the application of the delta method: while the method assumes the **Exact-IEBP** distance is continuous, in reality it is a discrete function. The effect gets more obvious as k is large: different values of k all give breakpoint distances close to the maximum, yet the **Exact-IEBP** can only return one estimate for k , hence the very low variance. This problem is less serious as n increases.

3 Variance of the Inversion and EDE Distances

The EDE distance. Given two genomes having the same set of n genes and the inversion distance between them is d , we define the **EDE** distance as $nf^{-1}\left(\frac{d}{n}\right)$: here n is the number of genes, and f , an approximation to the expected inversion distance normalized by the number of genes, is defined

as (see [12]):

$$f(x) = \min\left\{x, \frac{ax^2 + bx}{x^2 + cx + b}\right\}$$

We simulate the inversion-only GNT model to evaluate the relationship between the inversion distance and the actual number of inversion applied. Regression on simulation results suggests $a = 1$, $b = 0.5956$, and $c = 0.4577$. As the rational function is inverted, we take the larger (and only positive) root:

$$x = \frac{-(b - cy) \pm \sqrt{(b - cy)^2 + 4(a - y)by}}{2(a - y)}.$$

Let $y = \frac{d}{n}$. Thus

$$f^{-1}(y) = \max\left\{y, \frac{-(b - cy) \pm \sqrt{(b - cy)^2 + 4(a - y)by}}{2(a - y)}\right\}.$$

Here the coefficients do not depend on n , since for different values of n the curves of the normalized expected inversion distance are similar.

Regression for the Variance. Due to the success of nonlinear regression in the derivation of EDE, we use the same technique again for the variance of the inversion distance (and that of EDE). However for different numbers of genes, the curves of the variance are very different (see Figure 2). From the simulation it is obvious the magnitudes of the curves are inversely proportional to the number of genes (or some kind of function of it).

We use the following regression formula for the standard deviation of the inversion distance normalized by the number of genes after nx inversions are applied:

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

The constant term in the numerator is zero because we know $g(0) = 0$. Let r be the value such that rn is the largest number of inversions applied; we use $r = 2.5$. Note that

$$\ln\left(\frac{1}{rn} \sum_0^{rn} g_n(x)\right) \simeq \ln\left(\frac{1}{r} \int_0^r g_n(x) dx\right) = q \ln n + \ln\left(\frac{1}{r} \int_0^r \frac{ux^2 + vx}{x^2 + wx + t} dx\right)$$

is a linear function of $\ln n$. Thus we can obtain q as the slope in the linear regression using n as the independent variable and $\ln\left(\frac{1}{rn} \sum_0^{rn} g_n(x)\right)$ as the independent variable (see Figure 2(b); simulation results suggest the average of the curve indeed is inversely proportional to $\ln n$). When q is obtained we apply nonlinear regression to obtain u , v , w , and t using the simulation data for 40, 80, 120, and 160 genes. The resultant functions are shown as the solid curves in Figure 2, with coefficients $q = -0.6998$, $u = 0.1684$, $v = 0.1573$, $w = -1.3893$, and $t = 0.8224$.

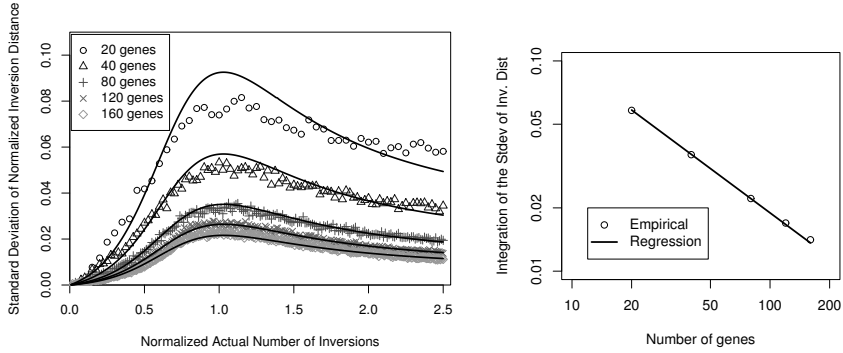


Fig. 2. Left: simulation (points) and regression (solid lines) of the standard deviation of the inversion distance. Right: regression of coefficient q (see Section 3); for every point corresponding to n genes, the y coordinate is the average of all data points in the simulation.

Variance of the EDE Distance. Let X_k and Y_k be the inversion and EDE distances after k inversions are applied to a genome of n genes, respectively. We again use the delta method. Let $x = \frac{k}{n}$. Since $X_k = nf(\frac{Y_k}{n})$, we have

$$\left| \frac{dY_k}{dX_k} \right|^{-1} = \left| \frac{dX_k}{dY_k} \right| = \frac{1}{n} \left| \frac{dX_k}{d(Y_k/n)} \right| = f'(x) = \frac{d}{dx} \left(\min \left\{ x, \frac{x^2 + bx}{x^2 + cx + b} \right\} \right).$$

The point where $x = \frac{x^2 + bx}{x^2 + cx + b}$ is when $x = 0.5423$. Therefore

$$f'(x) = \begin{cases} 1 & \text{if } 0 \leq x < 0.5423 \\ \frac{d}{dx} \left(\frac{x^2 + bx}{x^2 + cx + b} \right) = \frac{x^2(c - b) + 2bx + b^2}{(x^2 + cx + b)^2} & \text{if } x \geq 0.5423 \end{cases}$$

and $Var(Y_k) \simeq \left| \frac{dY_k}{dX_k} \right|^2 Var(X_k) = (f'(x))^{-2} (ng_n(x))^2 = (ng_n(\frac{k}{n})/f'(\frac{k}{n}))^2$.

4 Simulation Study

We present the results of two simulation studies in this section. The first experiment compares the accuracy of our new methods with other distance-based methods for genome rearrangement phylogeny. In the second experiment we show the robustness of `Weighbor-IEBP` against errors in the parameters of the GNT model, (α, β) .

4.1 Experiment 1: accuracy of the new methods

In this experiment we compare the trees reconstructed using `BioNJ-IEBP`, `Weighbor-IEBP`, `BioNJ-EDE`, and `Weighbor-EDE` to neighbor joining trees

Table 1. Settings for Experiments 1 and 2.

Parameter	Value
1. Number of genes	120 (plant chloroplast genome)
2. Model tree generation	Uniformly Random Topology (See the Model Tree paragraph in Section 4.1 for details.)
4. GNT Model parameters $(\alpha, \beta)^\dagger$	$(0, 0)$, $(\frac{1}{4}, \frac{1}{4})$
5. Datasets for each setting	30

\dagger The probabilities that a rearrangement is an inversion, a transposition, or an inverted transposition are $1 - \alpha - \beta$, α , and β , respectively.

using different distance estimators. The following four distance estimators are used with neighbor joining: (1) BP, the breakpoint distance, (2) INV [2], the inversion distance, and (3) **Exact-IEBP** [19] and (4) **EDE** [12], true evolutionary distance estimators based on BP and INV, respectively. The procedure of neighbor joining combined with distance X will be denoted by $\text{NJ}(X)$. According to past simulation studies [21, 12, 19], $\text{NJ}(\text{EDE})$ has the best accuracy, followed closely by $\text{NJ}(\text{Exact-IEBP})$. See Table 1 for the settings for the experiment.

Quantifying error. Given an inferred tree, we compare its “topological accuracy” by computing “false negatives” with respect to the “true tree” [11, 8]. During the evolutionary process, some edges of the model tree may have no changes (*i.e. evolutionary events*) on them. Since reconstructing such edges is at best guesswork, we are not interested in these edges. Hence, we define the true tree to be the result of contracting those edges in the model tree on which there are no changes.

We now define how we score an inferred tree, by comparison to the true tree. For every tree there is a natural association between every edge and the bipartition on the leaf set induced by deleting the edge from the tree. Let T be the true tree and let T' be the inferred tree. An edge e in T is “missing” in T' if T' does not contain an edge defining the same bipartition; such an edge is called a false negative. Note that the external edges (*i.e.* edges incident to a leaf) are trivial in the sense that they are present in every tree with the same set of leaves. The *false negative rate* is the number of false negative edges in T' with respect to T divided by the number of internal edges in T .

Software. We use PAUP* 4.0 [18] to compute the neighbor joining method and the false negative rates between two trees. We have implemented a simulator [12, 21] for the GNT model. The input is a rooted leaf-labeled model tree $(T, \{\lambda_e\})$, and parameters (α, β) . On each edge, the simulator

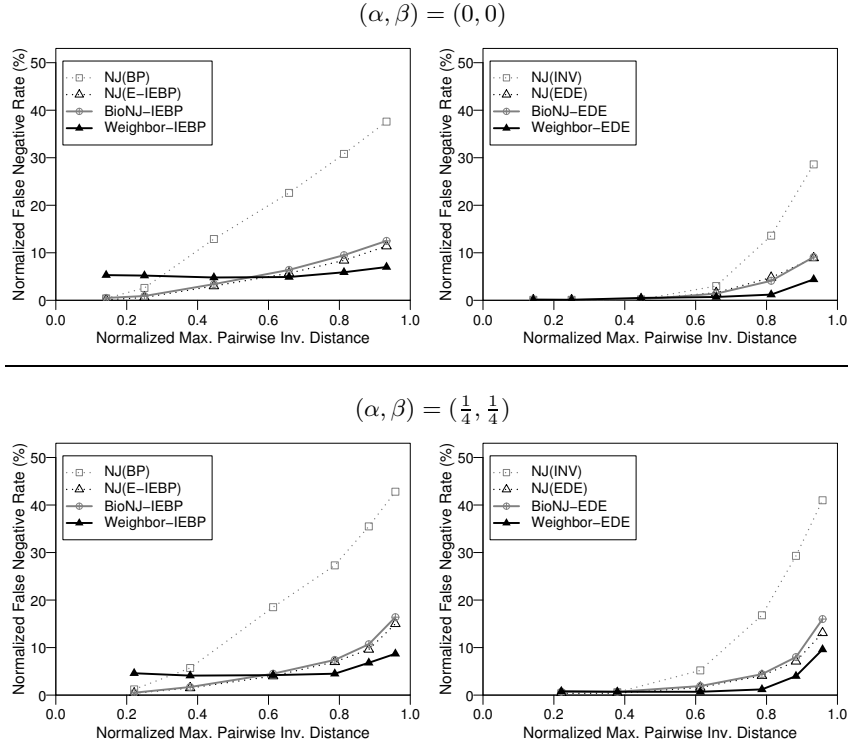


Fig. 3. The topological accuracy of various distance-based tree reconstruction methods. The number of genomes is 160. See Table 1 for the settings in the experiment.

applies random rearrangement events to the circular genome at the ancestral node according to the model with given parameters α and β . We use the original *Weighbor* and *BioNJ* implementations [5, 7] (downloadable from the internet) and make modifications so they use the new variance formulas.

Model Trees. The model trees have topologies drawn from the uniform distribution¹, and edge lengths drawn from the discrete uniform distribution on intervals $[1, b]$, where b is one of the following: 3, 6, 12, 18 (higher values of b makes the variance of the edge lengths higher). Then the length of each edge is scaled by the same factor so the diameter of the tree (the maximum pairwise leaf-to-leaf distance on the tree) is 36, 72, 120, 180, 360 (so low- to high-evolutionary rates are covered).

¹ This is easily done and well known by the community by add one leaf at a time to produce the whole tree. At each iteration, we choose an edge from the current tree (each edge has the same probability to be chosen) and attach the new leaf to it.

Discussion. The results of the simulation are in Figure 3. Due to space limitations, and because the relative order of accuracy remains the same, we describe our results only for a subset of our experiments. For each setting for simulation, we group methods based on the genomic distances they are based on: breakpoint or inversion distance. In either group, the relative order of accuracy is roughly the same. Let Y be the true distance estimator based on a genomic distance X ; e.g. $Y=IEBP$ if $X=BP$, and $Y=EDE$ if $X=INV$. The order of the methods, starting from the worst, is (1) $NJ(X)$, (2) $BioNJ-Y$, (3) $NJ(Y)$, and (4) $Weighbor-Y$ (except for very low evolutionary rates when $Weighbor-IEBP$ is worst, but only by a few percents). The differences between $NJ(Y)$ and $BioNJ-Y$ are extremely small. $Weighbor-EDE$ has the best accuracy over all methods.

When we compare methods based on breakpoint distance and methods based on inversion distance, the latter are always better (or no worse) than the former if we compare methods of the same complexity: $NJ(INV)$ is better than $NJ(BP)$, $NJ(EDE)$ is better than $NJ(Exact-IEBP)$, $BioNJ-EDE$ is better than $BioNJ-IEBP$, and $Weighbor-EDE$ is better than $Weighbor-IEBP$. This suggests INV is a better statistic than BP for the true evolutionary distance under the GNT model, even when transpositions and inverted transpositions are present. This is not surprising as INV , just like BP , increases by a small constant when a rearrangement event from the GNT model is applied. Also, though their maximum allowed values are the same (the number of genes for circular signed genomes), the fact the average increase in INV is smaller² than the average increase in BP gives INV a wider effective range.

Note $Weighbor-IEBP$ outperforms $NJ(Exact-IEBP)$ when the normalized maximum pairwise inversion distance, or the diameter of the dataset, exceeds 0.6; $Weighbor-IEBP$ (based on BP) is even better than $NJ(EDE)$ (based on the better INV) when the diameter of the dataset exceeds 0.9. This suggests the $Weighbor$ approach really shines under high amounts of evolution.

² An inversion creates two breakpoints; a transposition and an inverted transposition can be realized by three and two inversions, respectively, and they all create three breakpoints each. Thus under the GNT model with model parameters (α, β) and assumption that genome G has only a small breakpoint ($BP(G, G_0)$) and inversion ($INV(G, G_0)$) distance from the reference (ancestral) genome G_0 , the average increase in $BP(G, G_0)$ after a random rearrangement is applied to G is $2(1 - \alpha - \beta) + 3\alpha + 3\beta = 2 + \alpha + \beta$ and the average increase in $INV(G, G_0)$ is $(1 - \alpha - \beta) + 3\alpha + 2\beta = 1 + 2\alpha + \beta$. The latter is always smaller, and the two quantities are equal only when $\alpha = 1$, i.e. only transpositions occur.

Running time. NJ, BioNJ-IEBP, and BioNJ-EDE all finish within 1 second for all settings on our Pentium workstations running Linux. However, Weighbor-IEBP and Weighbor-EDE take considerably more time; both Weighbor-IEBP and Weighbor-EDE take about 10 minutes to finish for 160 genomes. As a side note, fast maximum parsimony methods for genome rearrangement phylogeny, MPME/MPBE, almost always exceed the four-hour running time limit in the experiment in [20]; Weighbor-EDE is almost as good as these two methods except for datasets with very high amount of evolution.

4.2 Experiment 2: robustness of Weighbor-IEBP

In this section we demonstrate the robustness of the Weighbor-IEBP method when the model parameters are unknown. The settings are the same in Table 1. The experiment is similar to the previous experiment, except here we use both the correct and the incorrect values of (α, β) for the Exact-IEBP distance and the variance estimation. Due to space limitations the results are not shown here. The false negative curves are similar even when different values of α and β are used. These results suggest that Weighbor-IEBP is robust against errors in (α, β) .

5 Conclusion and Future Work

In this paper we study the variance of the breakpoint and inversion distances under the Generalized Nadeau-Taylor model. We then used these results to obtain four new methods: BioNJ-IEBP, Weighbor-IEBP, BioNJ-EDE, and Weighbor-EDE. Of these Weighbor-IEBP and Weighbor-EDE yield very accurate phylogenetic trees, and are robust against errors in the model parameters. Future research includes analytical estimates of the expectation and variance of the inversion distance, and the difference between the approximating model and the breakpoint distance under the true GNT model.

6 Acknowledgements

The author thanks Tandy Warnow for suggesting the direction of research which leads to this paper, and the two anonymous referees for their helpful comments.

References

- [1] K. Atteson. The performance of the neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25(2/3):251–278, 1999.
- [2] D. A. Bader, B. M. E. Moret, and M. Yan. A linear-time algorithm for computing inversion distances between signed permutations with an experimental study. *J. Comput. Biol.*, 8(5):483–491, 2001.

- [3] V. Bafna and P. Pevzner. Sorting permutations by transpositions. In *Proc. 6th Annual ACM-SIAM Symp. on Disc. Alg. SODA95*, pages 614–623. ACM Press, 1995.
- [4] M. Blanchette, M. Kunisawa, and D. Sankoff. Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.*, 49:193–203, 1999.
- [5] W. J. Bruno, N. D. Socci, and A. L. Halpern. Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.*, 17:189–197, 2000. <http://www.t10.lanl.gov/billb/neighbor/>.
- [6] S. R. Downie and J. D. Palmer. Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In P. Soltis, D. Soltis, and J.J. Doyle, editors, *Molecular Systematics of Plants*, volume 49, pages 14–35. Chapman & Hall, 1992.
- [7] O. Gascuel. BIONJ: an improved version of the nj algorithm based on a simple model of sequence data. *Mol. Biol. Evol.*, 14:685–695, 1997. <http://www.crt.umontreal.ca/~olivierg/bionj.html>.
- [8] O. Gascuel. Personal communication, April 2001.
- [9] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, 1994. 2nd ed.
- [10] S. Hannenhalli and P. Pevzner. Transforming cabbage into turnip (polynomial algorithm for genomic distance problems). In *Proc. 27th Annual ACM Symp. on Theory of Comp. STOC95*, pages 178–189. ACM Press, NY, 1995.
- [11] S. Kumar. Minimum evolution trees. *Mol. Biol. Evol.*, 15:584–593, 1996.
- [12] B. M. E. Moret, L.-S. Wang, T. Warnow, and S. Wyman. New approaches for reconstructing phylogenies based on gene order. In *Proc. 9th Intl. Conf. on Intel. Sys. for Mol. Bio. (ISMB 2001)*, pages 165–173. AAAI Press, 2001.
- [13] G. W. Oehlert. A note on the delta method. *Amer. Statist.*, 46:27–29, 1992.
- [14] R. G. Olmstead and J. D. Palmer. Chloroplast DNA systematics: a review of methods and data analysis. *Amer. J. Bot.*, 81:1205–1224, 1994.
- [15] L. A. Raubeson and R. K. Jansen. Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. *Science*, 255:1697–1699, 1992.
- [16] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. & Evol.*, 4:406–425, 1987.
- [17] D. Sankoff and M. Blanchette. Probability models for genome rearrangements and linear invariants for phylogenetic inference. *Proc. 3rd Int'l Conf. on Comput. Mol. Bio. (RECOMB99)*, pages 302–309, 1999.
- [18] D. Swofford. *PAUP* 4.0*. Sinauer Associates Inc, 2001.
- [19] L.-S. Wang. Improving the accuracy of evolutionary distances between genomes. In *Lec. Notes in Comp. Sci. No. 2149: Proc. 1st Workshop for Alg. & Bio. Inform. WABI 2001*, pages 175–188. Springer Verlag, 2001.
- [20] L.-S. Wang, R. K. Jansen, B. M. E. Moret, L. A. Raubeson, and T. Warnow. Fast phylogenetic methods for the analysis of genome rearrangement data: An empirical study. In *Proc. 7th Pacific Symp. Biocomputing (PSB 2002)*, pages 524–535, 2002.
- [21] L.-S. Wang and T. Warnow. Estimating true evolutionary distances between genomes. In *Proc. 33th Annual ACM Symp. on Theory of Comp. (STOC 2001)*, pages 637–646. ACM Press, 2001.