Estimating True Evolutionary Distances Between Genomes

Li-San Wang
Department of Computer Sciences,
University of Texas at Austin

Tandy Warnow [*]
Department of Computer Sciences,
University of Texas at Austin

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, provided that sufficiently accurate methods can be developed to reconstruct evolutionary trees in these models [3, 6, 7, 15, 17]. A necessary component of any such method is the ability to accurately estimate *true evolutionary distances* between two genomes, which is the number of rearrangement events that took place in the evolutionary history between them. We present a new technique called *IEBP*, for estimating the true evolutionary distance between two genomes, whether signed or unsigned, circular or linear, and for any relative probabilities of rearrangement event classes. The method is highly accurate, as our simulation study shows. This simulation study also shows that the distance estimation technique improves the accuracy of the phylogenetic trees reconstructed by the popular distance-based method, neighbor joining [1, 20].

1. INTRODUCTION

The genomes of some organisms have a single chromosome or contain single chromosome organelles (such as mitochondria [4, 18] or chloroplasts [7, 17, 18, 19]) whose evolution is largely independent of the evolution of the nuclear genome for these organisms. Many single-chromosome organisms and organelles have circular chromosomes. Given a single chromosome, whether linear or circular, gene maps and whole genome sequencing projects can provide us with information about the ordering and strandedness of the genes, thus representing each chromosome by an ordering (linear or circular) of signed genes (where the sign of the gene indicates which strand it is located on).[1] The evolutionary process that operates on the chromosome can thus be seen as a transformation of signed orderings of genes.[2]

We are particularly interested in the following rearrangements: inversion, transposition, and inverted transposition. Starting with a genome $G = (g_1, g_2, \ldots, g_n)$, an *inversion* between indices i and j, $i < j$, (for unsigned genomes we also require $j \geq i + 1$) produces the genome with linear ordering

$$(g_1, g_2, \ldots, g_{i-1}, -g_j, -g_{j-1}, \ldots, -g_i, g_{j+1}, \ldots, g_n)$$

If we have $j < i$, we can still apply an inversion to a circular (but not linear) genome by simply rotating the circular ordering until g_i precedes g_j in the representation — we consider all rotations of the complete circular ordering of a circular genome as equivalent.

A *transposition* on the (linear or circular) genome G acts on three indices, i, j, k, with $i < j$ and $k \notin [i, j]$, and operates by picking up the interval $g_i, g_{i+1}, \ldots, g_j$ and inserting it immediately after g_k. Thus the genome G above (with the additional assumption of $k > j$) is replaced by

$$(g_1, \ldots, g_{i-1}, g_{j+1}, \ldots, g_k, g_i, g_{i+1}, \ldots, g_j, g_{k+1}, \ldots, g_n)$$

An *inverted transposition* is the combination of a transposition and an inversion on the transposed subsequence, so that G is replaced by[3]

$$(g_1, \ldots, g_{i-1}, g_{j+1}, \ldots, g_k, -g_j, -g_{j-1}, \ldots, -g_i, g_{k+1}, \ldots, g_n)$$

In the *Generalized Nadeau-Taylor model* [16], inversions, transpositions, and inverted transpositions occur on each edge. Different inversions have equal probability, as do different transpositions and inverted transpositions. Each model tree has two parameters α and β, which are the probability a rearrangement event is a transposition or an inverted transposition, respectively. (The probability for a rearrangement to be an inversion is thus $1 - \alpha - \beta$.) The number of events

[*] This research is supported by the David and Lucile Packard Foundation. Mailing address: Dept. of Computer Sciences, Taylor Hall 2.124, UT Austin, Austin, TX 78712-1188 USA. E-mail: (lisan,tandy)@cs.utexas.edu

STOC'01, July 6-8, 2001, Hersonissos, Crete, Greece.

[1] For circular genomes, the "strandedness" of a gene, also the sign in the representation, is relative; it depends on which strand is chosen that corresponds to the "positive" strand in the identity genome. Also, the position in the representation depends on which gene is chosen in the first position. That is, the representation is equivalent under rotation, and reversal and changing the sign of the whole sequence at the same time. We define the *canonical representation* of genome G to be the representation such that gene 1 has positive sign and is at the first position.

[2] We assume no gene duplication or deletion events occur, so all genomes contain one copy of each gene.

[3] For unsigned genomes, the subsequence to be inverted has at least two genes.

on each edge e is Poisson distributed with mean λ_e. This process produces a set of signed gene orders at the leaves of the model tree.

Let T be the true tree on which a set of genomes $\{G_1, \dots, G_n\}$ has evolved. Every edge e in T is associated with a number k_e, the actual number of rearrangements along edge e. The *true evolutionary distance* (*t.e.d.*) between two leaves G_i and G_j in T is $k_{ij} = \sum_{e \in P_{ij}} k_e$, where P_{ij} is the simple path on T between G_i and G_j. Using good estimates of true evolutionary distances between genomes greatly improves the performance of distance-based methods (in particular, the neighbor joining method (NJ) [20] – see [1] for the proof for NJ's error tolerance). Obtaining *t.e.d.* estimates when analyzing DNA sequences (under stochastic models of DNA sequence evolution) is understood, and well-studied [22]. However, very little work has been done on obtaining *t.e.d.* estimates between whole genomes; only the special case of estimating the actual number of inversions between genomes has been solved [5, 21].

In this paper we present a general analytical technique for estimating the true evolutionary distances under the Generalized Nadeau-Taylor model. This technique applies to datasets that are either signed or unsigned, circular or linear, and for many other rearrangement classes.

The organization of this paper is as follows. We define our notation in Section 2. In Section 3 we show the derivation and error bounds for our estimation. The technique is then checked by computer simulations in Section 4; our study shows the performance of neighbor joining, the most popular distance-based method, is improved when using the IEBP distance correction. We summarize our results in Section 5.

2. DEFINITIONS

We first define the *breakpoint distance* between two genomes. Let genome $G_0 = (g_1, \dots, g_n)$, and let G be a genome obtained by rearranging G_0. As in [21], we say that two genes g_i and g_j are *adjacent* in genome G if g_i is immediately followed by g_j in G_0, or, equivalently, if $-g_j$ is immediately followed by $-g_i$. A breakpoint in G with respect to G_0 is defined as an ordered pair of genes (g_i, g_j) such that g_i and g_j are adjacent in G_0, but are not adjacent in G (neither (g_i, g_j) nor $(-g_j, -g_i)$ appear consecutively in that order in G). The breakpoint distance between two genomes G and G_0 is the number of breakpoints in G with respect to G_0 (or vice versa, since the breakpoint distance is symmetric). For example, in the two circular genomes $G_1 = (1, 2, 3, 4, 5)$ and $G_2 = (1, -4, -3, 2, 5)$, there are three pairs of adjacent genes in G_1 but not in G_2: $(1, 2)$, $(2, 3)$, and $(4, 5)$, so the breakpoint distance between them is 3.

Let $\mathcal{G}_n$ be the set of all genomes of n genes. As will always be specified by the context under discussion, all genomes have the same form: signed or unsigned, circular or linear. Each genome G can be represented as a permutation of $\{1, 2, \dots, n\}$ if it is unsigned, or as a permutation of $\{\pm 1, \pm 2, \dots, \pm n\}$ if signed. A rearrangement ρ is a signed permutation on the n genes; the domain of ρ is $\mathcal{G}_n$. For any $G \in \mathcal{G}_n$, let ρG be the genome obtained by applying ρ on G. Two rearrangements ρ_1 and ρ_2 are *e*quivalent if $\rho_1 G = \rho_2 G, \forall G \in \mathcal{G}_n$. A rearrangement class $\mathcal{E}$ acting on $\mathcal{G}_n$ is a pair $(A(\mathcal{E}), f_{\mathcal{E}})$, where $A(\mathcal{E})$ is a set of rearrangements with nonzero probability of taking place, and $f_{\mathcal{E}}(\rho|G)$ is the probability that rearrangement ρ takes place on genome G, for a given $\rho \in A(\mathcal{E})$ and $G \in \mathcal{G}_n$. We say the random variable of rearrangements ρ on genome G is of rearrangement class $\mathcal{E}$ if ρ is in $A(\mathcal{E})$ and has distribution $f_{\mathcal{E}}(\rho|G)$.

Let $G_0 = (g_1, g_2, \dots, g_n)$ be the signed genome of n genes at the beginning of the evolutionary process. (If G_0 is circular, we assume the representation is canonical; that is, $g_1 = 1$.) For any $k \geq 1$, let $\rho_1, \rho_2, \dots \rho_k$ be k random rearrangements of rearrangement class $\mathcal{E}$, and let $G_k = \rho_k \rho_{k-1} \dots \rho_1 G_0$ (i.e. G_k is the result of applying these k rearrangements to G_0). We say that G_k *evolves under* $\mathcal{E}$. Given any linear genome $G = (g'_0, g'_1, g'_2, \dots, g'_n, g'_{n+1}) \in \mathcal{G}_n$, where g'_0 and g'_{n+1} are sentinel genes, we define the function $B_i(G), 0 \leq i \leq n$ by setting $B_i(G) = 0$ if genes g_i and g_{i+1} are adjacent, and $B_i(G) = 1$ if not; in other words, $B_i(G) = 1$ if and only if G has a breakpoint between g_i and g_{i+1}. We have the same definition when G is circular, except $B_0(G) = 0$ for all $G \in \mathcal{G}_n$, and $g'_{n+1} = g'_1$ for $B_n(G)$, so compatibility in summation indices is preserved. We denote the breakpoint distance between two genomes G and G' by $BP(G, G')$. Let $P_{i|k} = \Pr(B_i(G_k) = 1)$; then $E[BP(G_0, G_k)] = \sum_{i=0}^{n} P_{i|k}$.

Assume the rearrangement to act on G is ρ. We make the following definitions:

- $s(i|G, \mathcal{E}) = \Pr(B_i(\rho G) = 1 \mid B_i(G) = 0)$,
- $u(i|G, \mathcal{E}) = \Pr(B_i(\rho G) = 0 \mid B_i(G) = 1)$,
- $\mathrm{Sep}(i|G, \mathcal{E}) = \{\rho \in A(\mathcal{E}) : B_i(\rho G) = 1\}$, and
- $\mathrm{Uni}(i|G, \mathcal{E}) = \{\rho \in A(\mathcal{E}) : B_i(\rho G) = 0\}$.

We focus on rearrangement classes $\mathcal{E}$ where $f_{\mathcal{E}}$ is independent of k and G, and $s(i|G, \mathcal{E})$ is independent of G. The three rearrangement classes in the Generalized Nadeau-Taylor model, namely the class of random inversions, the class of random transpositions, and the class of random inverted transpositions, satisfy these requirements.

3. THE IEBP T.E.D ESTIMATOR

3.1 Introduction

In this section we show the derivation and properties of our *t.e.d.* estimator. We start in Section 3.2 with the simple case of rearrangement event classes where the breakpoints satisfy the Markov property, and find the expected number of breakpoints after k random rearrangements. The result is extended in Section 3.3, where the requirement on the Markov property is relaxed; the result is an approximation to the expected number of breakpoints. The error bounds on the approximation are shown in Section 3.4. The main result is in Section 3.5, where we develop the technique for rearrangement classes that are mixtures of other rearrangement classes. The technique is then applied to the Generalized Nadeau-Taylor model of genome rearrangements in Section 3.6.

3.2 Single Rearrangement Class Models Where the Breakpoints Satisfy Markov Property

We start with a simpler case by considering any rearrangement class $\mathcal{E}$ that has the property the two quantities $s(i|G, \mathcal{E})$ and $u(i|G, \mathcal{E})$ are independent of the past history and the current genome G to be acted upon. Then $\{B_i(G_k) | k \geq 0\}$ is a Markov process, as is shown in the following theorem:

THEOREM 1. *Assume $\mathcal{E}$ is a class of rearrangements such that $s(i|G,\mathcal{E})$ and $u(i|G,\mathcal{E})$ do not depend upon genome G. Let their common values be $s(i|\mathcal{E})$ and $u(i|\mathcal{E})$, respectively. Then*

$$P_{i|k} = s(i|\mathcal{E})\left(\frac{1-(1-s(i|\mathcal{E})-u(i|\mathcal{E}))^k}{1-(1-s(i|\mathcal{E})-u(i|\mathcal{E}))}\right).$$

PROOF. We have the following recurrence:

$$\begin{aligned}
P_{i|k+1} &= \Pr(B_1(G_{k+1})=1) \\
&= (1-P_{i|k})s(i|\mathcal{E}) + P_{i|k}(1-u(i|\mathcal{E})) \\
&= P_{i|k}(1-s(i|\mathcal{E})-u(i|\mathcal{E})) + s(i|\mathcal{E}) \\
P_{i|0} &= 0
\end{aligned}$$

The proof follows by solving the recurrence. □

3.3 The Lower and Upper Bounds Technique for Single Rearrangement Class Models

For many other classes of rearrangements, the parameters regarding transitions of $B_i(G)$'s state depend not only on $B_i(G)$ but on other properties of G. For example, the number of inversions that make genes g_1 and g_2 adjacent on signed genomes depend on the number of genes between these two genes. However, for the rearrangement classes $\mathcal{E}$ where $s(i|G,\mathcal{E})$ does not depend on G, we can obtain upper and lower bounds on the expected number of breakpoints and thus *t.e.d.* estimators.

Let $u_{min}(i|\mathcal{E})$ and $u_{max}(i|\mathcal{E})$ be the lower and upper bounds of $u(i|G,\mathcal{E})$ over all genomes G. Observe that a larger value of $u(i|G,\mathcal{E})$ means that genes g_i and g_{i+1} are more likely to be made adjacent, given that they are currently not adjacent. This means $P_{i|k}$, the probability of having a breakpoint between gene g_i and g_{i+1} after k rearrangements, is monotone decreasing on $u(i|G,\mathcal{E})$.

THEOREM 2. *Assume $\mathcal{E}$ is a class of rearrangements such that $s(i|\mathcal{E})$ is independent of the genome G currently acted upon. Let $u_{min}(i|\mathcal{E})$ and $u_{max}(i|\mathcal{E})$ be defined as in the previous paragraph. We have $P^L_{i|k} \leq P_{i|k} \leq P^H_{i|k}$ for all k, where*

$$\begin{aligned}
P^L_{i|k} &= s(i|\mathcal{E})\left(\frac{1-(1-s(i|\mathcal{E})-u_{max}(i|\mathcal{E}))^k}{1-(1-s(i|\mathcal{E})-u_{max}(i|\mathcal{E}))}\right) \\
P^H_{i|k} &= s(i|\mathcal{E})\left(\frac{1-(1-s(i|\mathcal{E})-u_{min}(i|\mathcal{E}))^k}{1-(1-s(i|\mathcal{E})-u_{min}(i|\mathcal{E}))}\right)
\end{aligned}$$

PROOF. The two recursions determined by $u_{min}(i|\mathcal{E})$ and $u_{max}(i|\mathcal{E})$ can be solved using Theorem 1. The last step is to prove the inequality bounding $P_{i|k}$ by $P^L_{i|k}$ and $P^H_{i|k}$ for all k using induction. When $k=0$, all three quantities are 0. The induction step is as follows:

$$\begin{aligned}
P^L_{i|k+1} &= P^L_{i|k}(1-s(i|\mathcal{E})-u_{max}(i|\mathcal{E})) + s(i|\mathcal{E}) \\
&\leq P_{i|k}(1-s(i|\mathcal{E})-u(i|G_k,\mathcal{E})) + s(i|\mathcal{E}) = P_{i|k+1} \\
&\leq P^H_{i|k}(1-s(i|\mathcal{E})-u_{min}(i|\mathcal{E})) + s(i|\mathcal{E}) = P^H_{i|k+1}
\end{aligned}$$

□

DEFINITION 1. *Given any class of rearrangements $\mathcal{E}$ that satisfies the assumption in Theorem 2, we set*

$$\mathcal{F}_k(\mathcal{E}) = \sum_{i=0}^{n} \frac{P^L_{i|k}+P^H_{i|k}}{2}.$$

The function $\mathcal{F}_k(\mathcal{E})$ is an approximation to the expected number of breakpoints after k random rearrangements drawn from $\mathcal{E}$.

3.4 Error Bounds on the Technique Using Upper and Lower Bounds

In this section we bound the absolute and relative errors of the estimator $\mathcal{F}_k(\mathcal{E})$ with respect to $E[BP(G_k,G_0)]$. Let $R^L_i = 1-s(i|\mathcal{E})-u_{max}(i|\mathcal{E})$, and $R^H_i = 1-s(i|\mathcal{E})-u_{min}(i|\mathcal{E})$. Note $(R^L_i)^k \leq (R^H_i)^k, \forall k \geq 0$. We now bound the error of the estimator $\mathcal{F}_k(\mathcal{E})$.

LEMMA 1.

$$\frac{1}{2}(P^H_{i|k}-P^L_{i|k}) \leq \frac{u_{max}(i|\mathcal{E})-u_{min}(i|\mathcal{E})}{2\ s(i|\mathcal{E})}$$

PROOF.

$$\begin{aligned}
\frac{1}{2}(P^H_{i|k}-P^L_{i|k}) &= \frac{1}{2}s(i|\mathcal{E})\left(\frac{1-(R^H_i)^k}{1-R^H_i} - \frac{1-(R^L_i)^k}{1-R^L_i}\right) \\
&= \frac{1}{2}s(i|\mathcal{E})\sum_{j=0}^{k-1}((R^H_i)^j-(R^L_i)^j) \\
&\leq \frac{1}{2}s(i|\mathcal{E})\sum_{j=0}^{\infty}((R^H_i)^j-(R^L_i)^j) \\
&= \frac{1}{2}s(i|\mathcal{E})\left(\frac{1}{1-R^H_i}-\frac{1}{1-R^L_i}\right) \\
&= \frac{s(i|\mathcal{E})(u_{max}(i|\mathcal{E})-u_{min}(i|\mathcal{E}))}{2(s(i|\mathcal{E})+u_{min}(i|\mathcal{E}))(s(i|\mathcal{E})+u_{max}(i|\mathcal{E}))} \\
&\leq \frac{u_{max}(i|\mathcal{E})-u_{min}(i|\mathcal{E})}{2\ s(i|\mathcal{E})}
\end{aligned}$$

□

THEOREM 3. *For all $k \geq 0$,*

$$|\mathcal{F}_k(\mathcal{E}) - E[BP(G_k,G_0)]| \leq \sum_{i=0}^{n}\frac{u_{max}(i|\mathcal{E})-u_{min}(i|\mathcal{E})}{2s(i|\mathcal{E})}.$$

In addition, for all $i : 0 \leq i \leq n$, if $u_{max}(i|\mathcal{E})$ (and thus $u_{min}(i|\mathcal{E})$) is $O(s(i|\mathcal{E})/n)$, (the case for random inversions, transpositions, and inverted transpositions), then

$$|\mathcal{F}_k(\mathcal{E}) - E[BP(G_k,G_0)]| = O(1).$$

PROOF. The error is at most one half of the maximum difference between $\sum_{i=0}^{n} P^H_{i|k}(\mathcal{E})$ and $\sum_{i=0}^{n} P^L_{i|k}(\mathcal{E})$; the result follows from Lemma 1.

When both $u_{min}(i|\mathcal{E})$ and $u_{max}(i|\mathcal{E})$ are $O(\frac{s(i|\mathcal{E})}{n})$, the error is at most

$$\sum_{i=0}^{n}\frac{u_{max}(i|\mathcal{E})-u_{min}(i|\mathcal{E})}{2\ s(i|\mathcal{E})} = \sum_{i=0}^{n} O(\frac{1}{n}) = O(1)$$

□

LEMMA 2. *Let* $s_l = \min_{0\le i\le n}\{s(i|\mathcal{E})\}$, $s_h = \max_{0\le i\le n}\{s(i|\mathcal{E})\}$, $r_l = \min_{0\le i\le n}\{s(i|\mathcal{E}) + u_{min}(i|\mathcal{E})\}$, *and* $r_h = \max_{0\le i\le n}\{s(i|\mathcal{E}) + u_{max}(i|\mathcal{E})\}$. *For all* $k \ge 1$,

$$\frac{s_l r_l}{s_h r_h} \le \frac{\mathcal{F}_k(\mathcal{E})}{E[BP(G_k, G_0)]} \le \frac{s_h r_h}{s_l r_l}$$

PROOF. We only prove the upper bound, as the lower bound is the reciprocal of the upper bound and can be proven similarly. Let $w = 1 - r_l$ and $v = 1 - r_h$; we have $v \le w$, $1 - w^k \le 1 - v^k$, and

$$\begin{aligned}
\frac{\mathcal{F}_k(\mathcal{E})}{E[BP(G_k, G_0)]} &\le \frac{\sum_{i=0}^{n} P^H_{i|k}}{\sum_{i=0}^{n} P^L_{i|k}} \le \frac{\max\limits_{0\le i\le n} P^H_{i|k}}{\min\limits_{0\le i\le n} P^L_{i|k}} \\
&\le \frac{s_h \dfrac{1-w^k}{1-w}}{s_l \dfrac{1-v^k}{1-v}} = \left(\frac{s_h(1-v)}{s_l(1-w)}\right)\left(\frac{1-w^k}{1-v^k}\right) \\
&\le \frac{s_h(1-v)}{s_l(1-w)} = \frac{s_h r_h}{s_l r_l}
\end{aligned}$$

□

THEOREM 4. *If* $s_h/s_l = 1 + \Theta(\frac{1}{n})$ *and* $\forall i : 0 \le i \le n$, $u_{max}(i|\mathcal{E})$ *(and thus* $u_{min}(i|\mathcal{E})$*) is* $O(s(i|\mathcal{E})/n)$, *then*

$$\frac{\mathcal{F}_k(\mathcal{E})}{E[BP(G_k, G_0)]} = 1 + O(\frac{1}{n})$$

PROOF. Follows from Lemma 2. □

3.5 Upper and Lower Bounds Estimation with Multiple Rearrangement Classes

We can easily extend the results to a mixture of different rearrangement classes. Consider m classes of rearrangements, $\mathcal{E}_1, \dots, \mathcal{E}_m$, where $\mathcal{E}_i = (A(\mathcal{E}_i), f_{\mathcal{E}_i}), 1 \le i \le m$. For any rearrangement ρ, let $\gamma_j = \Pr(\rho \in \mathcal{E}_j), 1 \le j \le m$. Assume γ_j does not depend on genome G, the genome currently acted upon. Let $s(i|\mathcal{E}_j)$, $u(i|G, \mathcal{E}_j)$, $u_{min}(i|\mathcal{E}_j)$, and $u_{max}(i|\mathcal{E}_j)$ be the parameters corresponding to $\mathcal{E}_j$ as defined in Theorem 2. Let $\mathcal{E} = (A(\mathcal{E}), f_{\mathcal{E}})$ be the rearrangement class such that $A(\mathcal{E}) = \cup_{j=1}^{m} A(\mathcal{E}_j)$, and $f_{\mathcal{E}}(r|G) = \sum_{j=1}^{m} \gamma_j \; f_{\mathcal{E}_j}(r|G)$. Then $\text{Sep}(i|G, \mathcal{E}) = \cup_{j=1}^{m}\text{Sep}(i|G, \mathcal{E}_j)$, and $\text{Uni}(i|G, \mathcal{E}) = \cup_{j=1}^{m}\text{Uni}(i|G, \mathcal{E}_j)$.

The hierarchical way of choosing rearrangements (first choose the rearrangement class, then choose one rearrangement among others in the class chosen) during evolution allows two rearrangements in different rearrangement classes to produce the same results, while retaining the additivity of probability:

$$\begin{aligned}
& \Pr(\rho = \rho_0 | G = G_0) \\
= & \sum_{j=1}^{m} \Pr(\rho = \rho_0 | G = G_0, \mathcal{E}_j \text{ chosen}) \Pr(\mathcal{E}_j \text{ chosen} | G = G_0) \\
= & \sum_{j=1}^{m} \gamma_j f_{\mathcal{E}_j}(\rho_0 | G_0)
\end{aligned}$$

The new recurrence is

$$\begin{aligned}
s(i|\mathcal{E}) &= \Pr(B_i(G_{k+1}) = 1 | B_i(G_k) = 0) \\
&= \Pr(\rho_k \in \text{Sep}(i|G_k, \mathcal{E}) \;\; | \; B_i(G_k) = 0) \\
&= \sum_{j=1}^{m} \Pr(\rho_k \in \text{Sep}(i|G_k, \mathcal{E}_j) \;\; | \; B_i(G_k) = 0) \\
&= \sum_{j=1}^{m} \gamma_j s(i|\mathcal{E}_j)
\end{aligned}$$

Similarly,

$$\begin{aligned}
u(i|G_k, \mathcal{E}) &= \Pr(B_i(G_{k+1}) = 0 | B_i(G_k) = 1) \\
&= \sum_{j=1}^{m} \gamma_j u_j(i|G_k, \mathcal{E}_j), \forall k \ge 0 \\
u_{min}(i|\mathcal{E}) &= \sum_{j=1}^{m} \gamma_j u_{min}(i|\mathcal{E}_j) \\
u_{max}(i|\mathcal{E}) &= \sum_{j=1}^{m} \gamma_j u_{max}(i|\mathcal{E}_j) \\
P_{i|k+1} &= (1 - P_{i|k}) s(i|\mathcal{E}) + P_{i|k}(1 - u(i|G_k, \mathcal{E})) \\
&= P_{i|k}(1 - s(i|\mathcal{E}) - u(i|G_k, \mathcal{E})) + s(i|\mathcal{E}) \\
P_{i|0} &= 0
\end{aligned}$$

Results similar to Theorems 3 and 4 on error bounds can be obtained for multiple classes:

THEOREM 5. *Consider the estimator* $\mathcal{F}_k(\mathcal{E})$ *defined in Definition 1 with the parameters* $s(i|\mathcal{E})$, $u_{min}(i|\mathcal{E})$, *and* $u_{max}(i|\mathcal{E})$ *in the previous paragraphs. If the assumptions in Theorems 3 and 4 regarding these parameters are satisfied, then*

$$\begin{aligned}
& |\mathcal{F}_k(\mathcal{E}) - E[BP(G_k, G_0)]| = O(1), \;\; \textit{and} \\
& \phi^{-1} \le \frac{\mathcal{F}_k(\mathcal{E})}{E[BP(G_k, G_0)]} \le \phi
\end{aligned}$$

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

PROOF. Follows from Theorems 3 and 4. □

3.6 A *t.e.d.* Estimator for the Generalized Nadeau-Taylor Model

Recall that in the Generalized Nadeau-Taylor model, all three types of rearrangements can occur: inversions, transpositions, and inverted transpositions. Given as part of the model are two values α and β such that the probability a rearrangement is an inversion, a transposition, or an inverted transpositions is $1-\alpha-\beta$, α, and β, respectively. Let $\mathcal{H}$ denote this model. In this section we use the techniques given above in order to derive an estimator of the true evolutionary distance (t.e.d.) between genomes, when the permitted rearrangements include inversions, transpositions, and inverted transpositions, and given arbitrarily defined probabilities on the three classes of rearrangements.

We first show how to compute s, u_{min}, and u_{max} by the following example. Assuming the inversion only ($\alpha = \beta = 0$) model on circular genomes, we can assume that g_1 has positive sign and is at the first position (i.e. we use the canonical representation), and we only need to consider the set of inversions that do not involve g_1. By symmetry of circular genomes and the model we only need to look at $P_{1|k}$, i.e. the distribution of the first breakpoint. Then s is the probability of breaking g_1 and g_2 apart assuming that g_1 is followed by g_2. Out of all $\binom{n}{2}$ inversions there are $n-1$ inversions that have one endpoint acting on g_2, so $s = (n-1)/\binom{n}{2} = 2/n$. Now consider u_{min} and u_{max}, the minimum and maximum probability of bringing g_1 and g_2 adjacent given they were not before the inversion is applied. The worst case is when the sign of g_2 is positive (by the assumption, g_2 cannot have the positive sign and follow g_1 before the inversion is applied) so that no inversion removes the breakpoint between g_1 and g_2, hence $u_{min} = 0$. If the sign of g_2 is negative, for every possible position for g_2 there is exactly one inversion that brings g_2 to the position right after g_1 and make it positive. So $u_{max} = 1/\binom{n}{2}$.

Recall the estimator given in Definition 1. We can tighten the error bounds obtained in Theorem 3, as follows:

Theorem 6. *For all $k > 0$,*

$$|\mathcal{F}_k(\mathcal{H}) - E[BP(G_k, G_0)]| \leq 1 + \frac{1}{n-1}, \text{ and}$$

$$\phi^{-1} \leq \frac{\mathcal{F}_k(\mathcal{H})}{E[BP(G_k, G_0)]} \leq \phi$$

where $\phi = 1 + \frac{2+4\alpha+2\beta}{2+\alpha+\beta}n^{-1} + O(n^{-2})$.

Proof. The proof involves in computing the quantities s, u_{min}, and u_{max}. The absolute error bound follows from the proof of Theorem 3. For the relative error bound, we look at $\frac{\sum_{i=0}^{n} P^H_{i|k}}{\sum_{i=1}^{n} P^L_{i|k}}$ directly to improve the result. □

We now define the *t.e.d.* estimator:

1. For every $i = 1, \ldots, n$ and $k = 1, \ldots, r$ (where r is some large integer enough to bring a genome to randomness) compute $s(i|\mathcal{H})$, $u_{min}(i|\mathcal{H})$ and $u_{max}(i|\mathcal{H})$ to obtain $P^L_{i|k}$, $P^H_{i|k}$, and thus $\mathcal{F}_k(\mathcal{H}) = \frac{1}{2}\sum_{i=0}^{n}(P^L_{i|k} + P^H_{i|k})$. These parameters only depend on n, k, α and β.

2. For each pair of genomes G and G', we estimate the true evolutionary distance $\widehat{k}(G, G')$ as follows:

   (a) Compute the breakpoint distance b between them.

   (b) Return integer $\widehat{k}(G, G') = k$ that minimizes

$$\left| \sum_{i=0}^{n} \mathcal{F}_k(\mathcal{H}) - b \right|.$$

We call the estimator IEBP, where IEBP stands for "the inverse of the expected number of breakpoints."

The parameters $s(i|G_k, \mathcal{H})$, $u_{min}(i|\mathcal{H})$ and $u_{max}(i|\mathcal{H})$ do not change as i takes different value in $\{1, \ldots, n\}$ for circular genomes. For linear genomes these parameters do not change for $i = 1, \ldots, n-1$, so we only need compute parameters for $i = 0, 1$, and n. This means $\mathcal{F}_k(\mathcal{H})$ can be computed in constant time for each value of k.

For the purpose of computing all $\binom{n}{2}$ pairwise distances, let N be the number of genomes and the dimension of the distance matrix. We need to compute the distance for at least $O(\min\{N^2, n\})$ distinct breakpoint distance values. Consider the value r, the number of inversions needed to produce a genome that is close to random; we can use this as an upper bound of k in computing the recursion. Both our simulation (see Section 4.2 and Figure 1) and the IEBP formula show that it is reasonable to set $r = un$ for some constant u that is sufficiently larger than 1 (in our experiment $u = 2.5$ is enough).

We can improve the running time by the following implementation. For each k, the value k that minimizes $|\mathcal{F}_k - D_{BP}(G)|$ can be found in $O(\log r)$ time using the bisection method. Since there are at most $n+1$ distinct nonzero breakpoint distance values, we create an array that stores the IEBP distances corresponding to each possible breakpoint distance value, and use the corresponding breakpoint distance values as indices. When a new breakpoint distance value is encountered we compute the IEBP distance and store it in the array. Therefore computing the IEBP distances for N genomes takes $O(N^2 + \min\{n, N^2\}\log r)$ time. We summarize the discussion as follows:

Theorem 7. *Let n be the number of genes in each genome, and let N be the number of genomes. We can compute the IEBP distances of all $\binom{N}{2}$ pairs of genomes in $O(N^2 + \min\{n, N^2\}\log n)$ time.*

4. EXPERIMENTS

4.1 Introduction

In this section we present the results of our experimental performance study in which we compare various techniques for estimating genomic distances, using simulations of evolution under the Generalized Nadeau-Taylor model. We first look at the behavior of the inversion and breakpoint distance as well as the IEBP distance correction on different amounts of evolution (i.e. number of rearrangements) under the Generalized Nadeau-Taylor model. We then apply neighbor joining to the three distances and compare the performance. Finally we show neighbor joining on IEBP is robust to errors in the model parameters, namely (α, β).

4.2 Fitness of the Estimators

In this study, we simulated evolution under the Generalized Nadeau-Taylor model of genome evolution under three probability settings (inversion only, equiprobable transposition and inverted transpositions, and equiprobable inversions, transpositions, and inverted transpositions). The three distances we look at are:

- BP, the breakpoint distance,
- INV, the inversion distance, meaning the minimum number of inversions needed to transform one genome into another (computed in linear time using the algorithm given in [2]; see also [9, 12]), and
- IEBP, the estimator described in Section 3.6, which estimates the actual number of rearrangements in the Generalized Nadeau-Taylor model.

We focus on the performance for the case with 120 genes (the typical number of genes in a chloroplast genome [11]). For each setting for the relative probabilities of rearrangements, we generate 5000 datasets. In each dataset we choose

k, the number of actual rearrangements, uniformly between 0 and 300, then apply k rearrangements with the setting of relative probabilities on an unrearranged genome. Note the maximum number of events, 300, is 2.5 times the number of genes. The experiments show this amount of rearrangements is more than enough to transform a genome close to a random one, hence making it difficult for any method to reconstruct the evolutionary history. We then generate error bar plots by comparing the three distances with the actual number of rearrangements: INV distance, BP distance, and IEBP distance. The result of this experiment is shown in Figure 1.

As the experiment shows, IEBP produces linear estimates of the actual number of rearrangements, and produces much better fits than, for example, the breakpoint distance, or the minimum inversion distance. Consider the initial segment of the x-axis in which each estimator is close to linear. This segment is largest for the IEBP estimator; for the INV distance, this region is at least as large as that of the BP distance in all three probability settings. The linear regions of INV and BP distances have similar width under different probability settings except for the inversion-only scenario: the first region of INV distance curve is considerably larger, up to almost 100 in the x-axis coordinate, while the breakpoint distance curve is clearly bended when it is close to 100 in the x-axis coordinate. This is not surprising, since INV distances are expected to work best when only inversions are present, while the breakpoint distances are less dependent on the type of rearrangements. When the number of rearrangements between two genomes is high, both INV and BP distances tend to underestimate the actual number of events. In all three probability settings IEBP distance has a larger variance than the other two distances, and INV has the smallest variance.

4.3 Performance of Neighbor Joining under Various Estimators

In this section we explore the performance of neighbor joining (NJ) under different ways of calculating genomic distances. The settings are in Table 1.

Given an inferred tree by some methods to be tested, we compare its "topological accuracy" by computing "false negatives" with respect to the "true tree" [13, 8]. We begin by defining the true tree. During the evolutionary process, some edges of the model tree may have no changes 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 a reconstructed tree, by comparison to the true tree. Given the true tree T, we define the set $C(T)$ to be the set of bipartitions on the leaf set induced by edge-deletions from T. Similarly, given a tree T', we define $C(T')$. An edge of the true tree is "missing" in T' if T' does not contain an edge defining the same bipartition; such an edge is called a *false negative*. The *false negative rate* is the number of missing edges divided by the number of bipartitions in T.

The input to our simulator is a rooted leaf-labeled tree and the associated parameters (i.e. branch lengths, and the relative probabilities of inversions, transpositions, and inverted transpositions). For each setting of the parameters (number of leaves, probabilities of rearrangements, and branch lengths), we generate 100 datasets of genomes as follows. First, we generate a random leaf-labeled tree (from the uniform distribution on topologies). The leaf-labeled tree and the parameter settings thus define a model tree in the Generalized Nadeau-Taylor model. We run the simulator on the model tree, and produce a set of genomes at the leaves.

Given each set of genomes, we compute the breakpoint distance, the minimum inversion distance, and the IEBP distance. We then compute neighbor joining trees on each of these distance matrices, and compare the resultant trees to the true tree. The results of this experiment are in Figure 2(a)-(c). The x-axis is the maximum normalized edit distance (as computed by the linear time algorithm for minimum inversion distances given in [2]) between any two genomes in the input, and hence ranges between 0 and 1. Distance matrices with some normalized edit distances close to 1 are said to be "saturated", and the recovery of accurate trees from such datasets is considered to be very difficult [10]. The y-axis is the false negative rate (i.e. the proportion of missing edges), and hence also ranges between 0 and 1. False negative rates of less than 5% are excellent, but false negative rates of up to 10% can be tolerated.

Note that neighbor joining obtains highly accurate trees using IEBP distances, even when the maximum normalized edit distance is close to saturation. By comparison, neighbor joining obtains much poorer estimates of the true tree using BP distances, and demonstrates a sharp decline in accuracy as the input approaches saturation. The performance of neighbor joining under INV distances is interesting: it is better, in some cases, than our theoretically obtained estimates of the actual number of rearrangements in the inversion-only scenario; this is explained by the fitness experiment that the INV distance has a smaller variance than the breakpoint and IEBP distances when the uncorrected distance is small, and the linear region is wide enough only in this setting. In the other two settings, the INV distance curve has smaller linear regions, so the accuracy of the neighbor joining trees under INV distances declines.

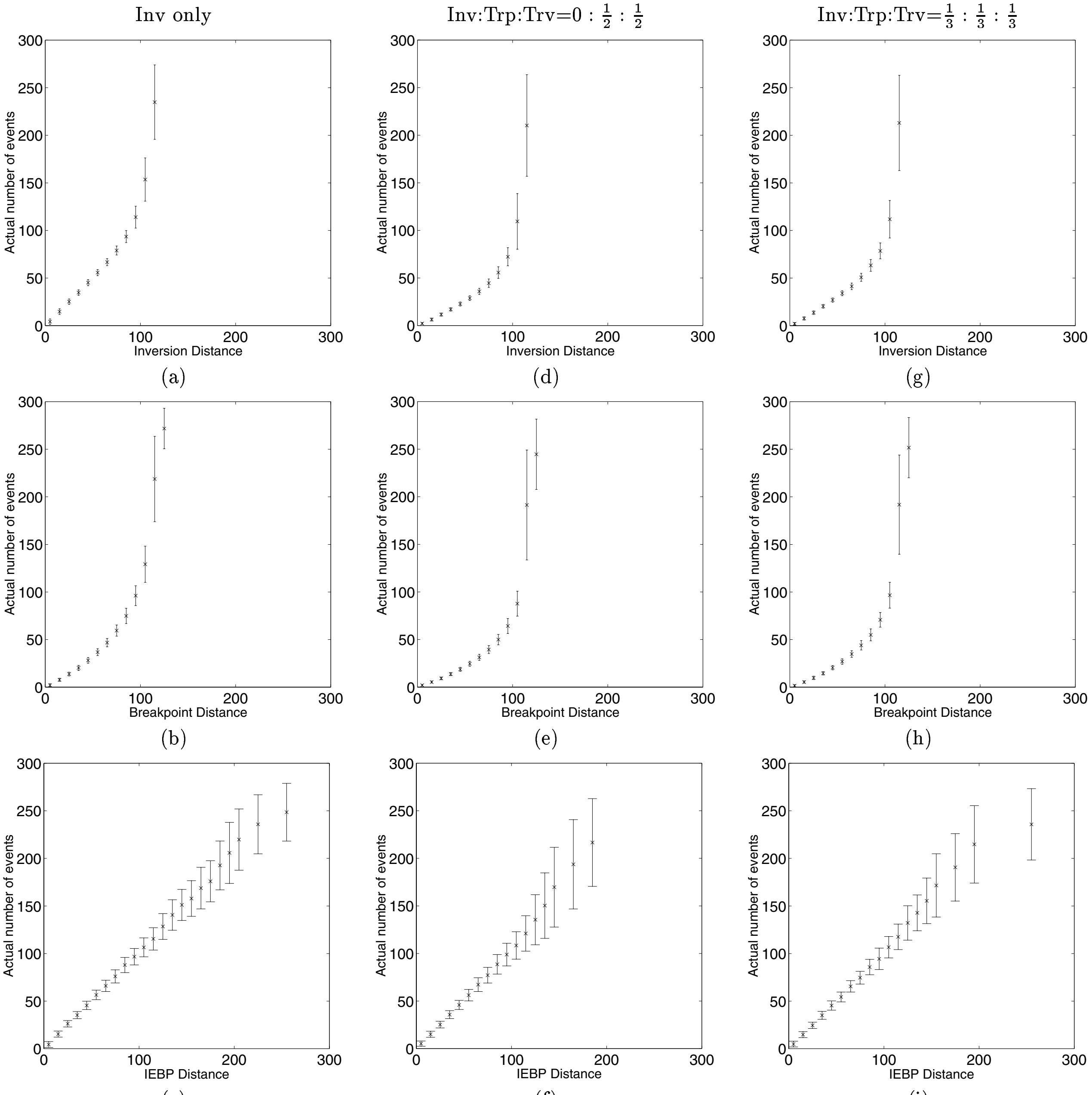
4.4 Robustness to Unknown Model Parameters

In this section we demonstrate the robustness of the IEBP estimator when the model parameters are unknown. The settings are the same in Table 1. The experiment is similar to that of Experiment 2, except we use both the correct and the incorrect values of (α, β) for the IEBP distance. The results are in Figure 3. These results suggest that neighbor joining on IEBP is robust against errors in (α, β).

This robustness is partially predicted by our earlier experiment shown in Figure 1. From Figure 1 we know the different breakpoint distance curves, despite the different model parameters (α, β), have very similar shapes. Since IEBP is based on the breakpoint distance, and neighbor joining only needs the input to be close to a constant multiple of the true evolutionary distance, the neighbor joining performances on IEBP should not be hampered even the model parameters are erroneous.

5. CONCLUSION

This paper has presented a new polynomial time algorithm for estimating the actual number of rearrangements that have taken place in the evolutionary history between two signed genomes, in which inversions, transpositions, and



Inv only

Inv:Trp:Trv=0 : $\frac{1}{2}$: $\frac{1}{2}$

Inv:Trp:Trv=$\frac{1}{3}$: $\frac{1}{3}$: $\frac{1}{3}$

(a) (d) (g)

(b) (e) (h)

(c) (f) (i)

Figure 1: Fitness of the estimators. The number of genes in the experiment is 120. Each plot is a comparison between some distance measure and the actual number of rearrangements: (a), (d), and (g) are on INV distances, (b), (e), and (h) are on BP distances, and (c), (f), and (i) are on IEBP distances. The x-axis is divided into 30 bins; the vertical bars indicate the amount of the standard deviation in each bin.

inverted transpositions are permitted. Our simulation study shows that neighbor joining produces highly accurate results using the corrected distance. We have successfully applied the method in getting accurate initial estimates for breakpoint and inversion phylogeny analysis [14]. They enable more accurate reconstructions of evolutionary trees on whole genomes, an increasingly important problem in comparative genomics.

6. ACKNOWLEDGMENT

We would like to thank Robert Jansen at the University of Texas at Austin for introducing us to the genome rearrangement phylogeny problem. We would like to thank David Sankoff and Joseph Nadeau for inviting us to the DCAF Conference last year in Montreal, Canada, during which some of the ideas in the paper came to fruition. We would like to thank Matthew Spencer at the University of Cambridge and Mike Steel at the University of Canterbury, New Zealand for their careful reading and helpful comments.

7. REFERENCES

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

[2] D. Bader, B. Moret, and M. Yan. A fast linear-time algorithm for inversion distance with an experimental comparison. In *Proc. 7th Workshop on Algs. and Data Structures. (WADS 2001)*, 2001. To appear.

[3] M. Blanchette, G. Bourque, and D. Sankoff. Breakpoint phylogenies. In S. Miyano and T. Takagi, editors, *Genome Informatics*, pages 25–34. Univ. Academy Press, 1997.

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

[5] A. Caprara and G. Lancia. Experimental and statistical analysis of sorting by reversals. In D. Sankoff and J. Nadeau, editors, *Comparative Genomics*, pages 171–184. Kluwer Academic Publishers, 2000.

[6] Workshop on Gene Order Dynamics, Comparative Maps and Multigene Families (DCAF). Montreal, Canada, August 2000.

[7] S. Downie and J. Palmer. Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In P. Soltis, D. Soltis, and J. Doyle, editors, *Molecular Systematics of Plants*, volume 49, pages 14–35. Chapman & Hall, 1992.

[8] O. Gascuel. Personal communication, April 2001.

[9] 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.

[10] D. Huson, S. Nettles, K. Rice, T. Warnow, and S. Yooseph. The hybrid tree reconstruction method. *J. Experimental Algorithmics*, 4:178–189, 1999. http://www.jea.acm.org/.

[11] R. Jansen. personal communication, October 3 2000.

[12] H. Kaplan, R. Shamir, and R. Tarjan. Faster and simpler algorithm for sorting signed permutations by reversals. In *Proc. 8th Annual Symp. on Discrete Alg. (SODA97)*, pages 344–351. ACM Press, NY, 1997.

[13] S. Kumar. Minimum evolution trees. *Mol. Biol. Evol.*, 15:584–593, 1996.

[14] B. 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)*. AAAI Press, 2001. To appear.

[15] B. Moret, S. Wyman, D. Bader, T. Warnow, and M. Yan. A new implementation and detailed study of breakpoint analysis. In *Proc. 6th Pacific Symp. Biocomputing (PSB 2001)*, pages 583–594, 2001.

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

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

[18] J. Palmer. Chloroplast and mitochondrial genome evolution in land plants. In R. Herrmann, editor, *Cell Organelles*, pages 99–133. Wein, 1992.

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

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

[21] 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.

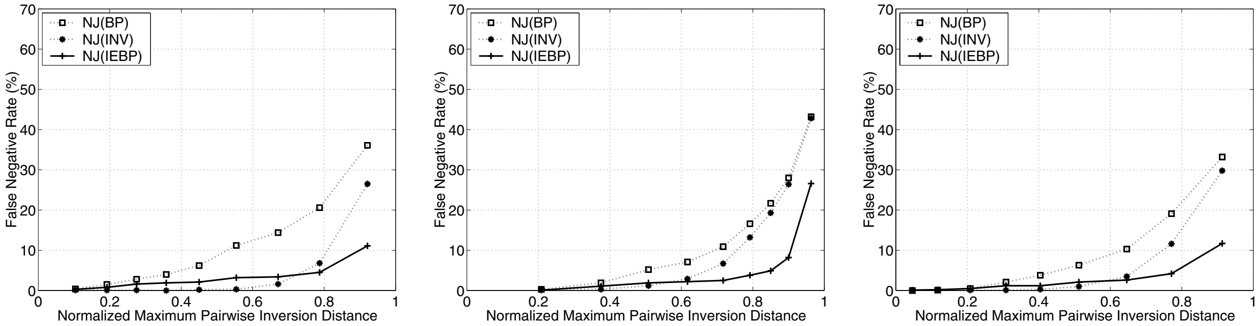
[22] D. Swofford, G. Olson, P. Waddell, and D. Hillis. Phylogenetic inference. In D. Hillis, C. Moritz, and B. Mable, editors, *Molecular Systematics*, Chapter 11. Sinauer Associates Inc, Second edition, 1996.

Parameter	Value
1. Number of genes	120
2. Number of leaves	10, 20, 40, 80, and 160
3. Expected number of rearrangements in each edge	Discrete Uniform within the following intervals: [1,3], [1,5], [1,10], [3,5], [3,10], and [5,10]
4. Probability settings: $(\alpha, \beta)^\dagger$	(0,0) (Inversion only) $(1, 0)$ (Transposition only) $(\frac{1}{3}, \frac{1}{3})$ (All three rearrangement classes equiprobable)
5. Datasets for each setting	100

† The probability that a rearrangement is an inversion, a transposition, or an inverted transposition is $1 - \alpha$, α, and β, respectively.

Table 1: Settings for the neighbor joining performance simulation study.

Experiment 2

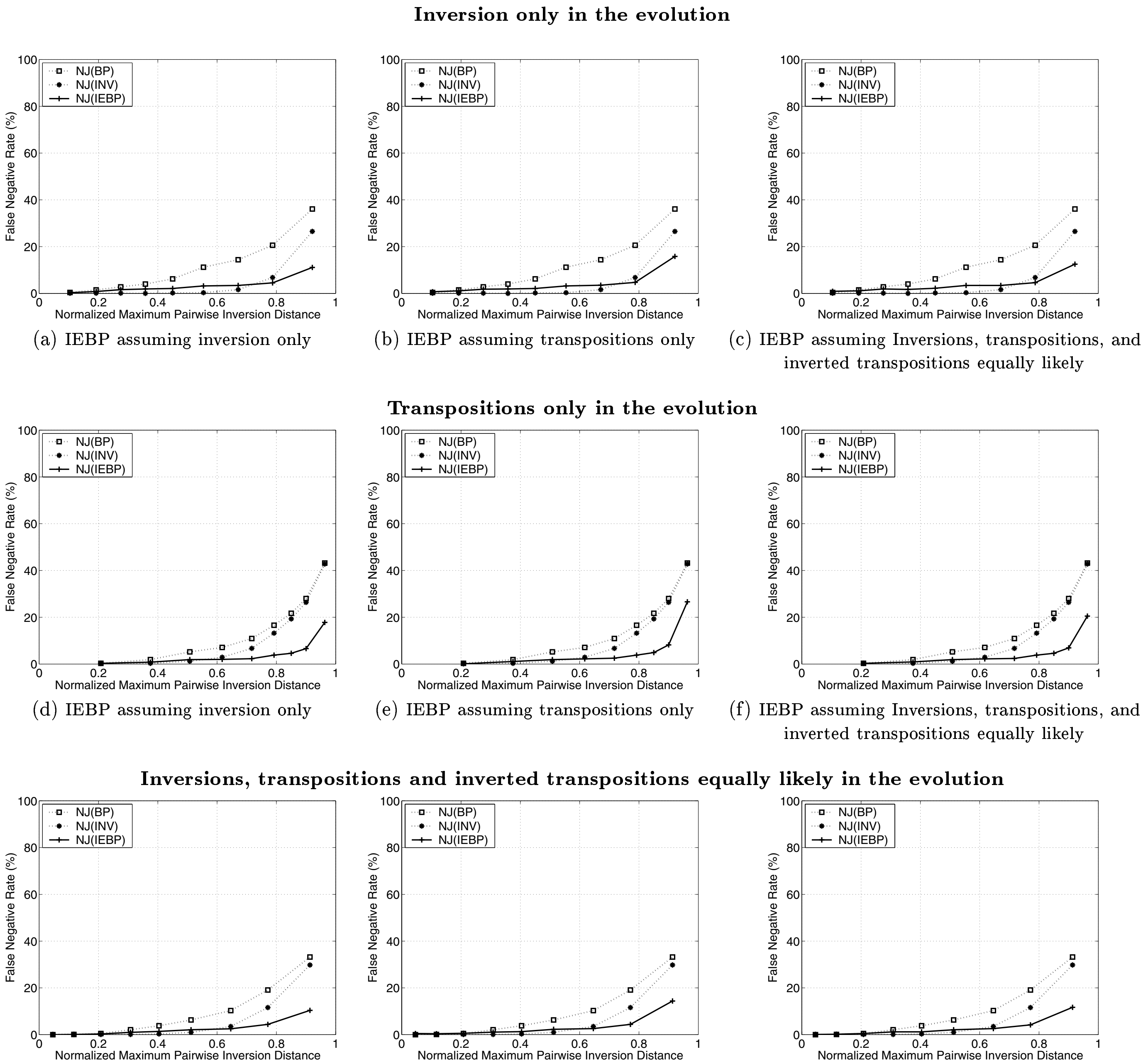


(a) Inversions only

(b) Transpositions only

(c) Inversions, transpositions, and inverted transpositions equally likely

Figure 2: Neighbor joining performance under several distance corrections. See Table 1 for the settings.



(a) IEBP assuming inversion only

(b) IEBP assuming transpositions only

(c) IEBP assuming Inversions, transpositions, and inverted transpositions equally likely

(d) IEBP assuming inversion only

(e) IEBP assuming transpositions only

(f) IEBP assuming Inversions, transpositions, and inverted transpositions equally likely

(g) IEBP assuming inversion only

(h) IEBP assuming transpositions only

(i) IEBP assuming Inversions, transpositions, and inverted transpositions equally likely

Figure 3: Neighbor joining performance for various distance estimators using correct and incorrect evolutionary model parameters. See Table 1 for the settings. Note that BP and INV distances are uniquely defined by the pairs of genomes and do not require the knowledge of the model. NJ(BP) and NJ(INV) are included for comparative purposes.