BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–9

Estimating True Evolutionary Distances under the DCJ Model

Yu Lin and Bernard M.E. Moret *

Laboratory for Computational Biology and Bioinformatics, Swiss Federal Institute of Technology (EPFL), EPFL-IIS-LCBB, INJ 230, Station 14, CH-1015 Lausanne, Switzerland

ABSTRACT

Motivation: Modern techniques can yield the ordering and strandedness of genes on each chromosome of a genome; such data already exists for hundreds of organisms. The evolutionary mechanisms through which the set of the genes of an organism is altered and reordered are of great interest to systematists, evolutionary biologists, comparative genomicists, and biomedical researchers. Perhaps the most basic concept in this area is that of evolutionary distance between two genomes: under a given model of genomic evolution, how many events most likely took place to account for the difference between the two genomes?

Results: We present a method to estimate the true evolutionary distance between two genomes under the "double-cut-and-join" (DCJ) model of genome rearrangement, a model under which a single multichromosomal operation accounts for all genomic rearrangement events: inversion, transposition, translocation, block interchange, and chromosomal fusion and fission. Our method relies on a simple structural characterization of a genome pair and is both analytically and computationally tractable. We provide analytical results to describe the asymptotic behavior of genomes under the DCJ model, as well as experimental results on a wide variety of genome structures to exemplify the very high accuracy (and low variance) of our estimator. Our results provide a tool for accurate phylogenetic reconstruction from multichromosomal gene rearrangement data as well as a theoretical basis for refinements of the DCJ model to account for biological constraints.

Availability: All of our software is available in source form under GPL at `http://lcbb.epfl.ch`

Contact: `bernard.moret@epfl.ch`

1 INTRODUCTION

The ordering and strandedness of genes on each chromosome of many organisms have become available, with many more added every year. Using this information, one can represent a genome as a collection of chromosomes, each of which is a linear or circular sequence of gene identifiers. Variations in the placement of the same genes, as well as variations in gene content and multiplicity, among organisms can then be analyzed. This data is of great interest to evolutionary biologists, but also to comparative genomicists and to any researcher interested in understanding evolutionary changes in pathogens. In the past ten years, there has been a large increase in work done on analyzing such data (see, e.g., Moret *et al.* (2005)).

Perhaps the most basic requirement in the analysis of such data is the ability to estimate the amount of evolutionary change between two genomes—that is, to compute a pairwise *evolutionary distance*. Since the *true distance*, that is, the actual number of changes in the gene order and content that took place during the course of evolution, is not something we can compute, researchers have used a two-stage process, in which a well defined measure is first computed (such as an *edit distance*, that is, the smallest number of evolutionary changes—from a defined set—needed to transform one genome into the other), then a statistical model of evolution is used to infer an estimate of the true distance by deriving the effect of a given number of changes in the model on the computed measure and (algebraically or numerically) inverting the derivation to produce a maximum-likelihood estimate of the true distance under the model. This second step is often called a distance "correction" and has long been used for sequence (DNA) data (see, e.g., Swofford *et al.* (1996)) as well as, more recently, for gene-order data, for which see Moret *et al.* (2001, 2002); Sankoff and Blanchette (1999); Wang (2001); Wang and Warnow (2001).

The measures commonly used in the first step (edit distances, synteny measures, etc.) are bounded and typically reflect only the endstate of an evolutionary process, whereas the true evolutionary distance can be arbitrarily large. Thus these first-step measures typically underestimate the true distance, by an amount that grows quickly as the true distance grows large. This is an aspect of the problem of *saturation*, in which the evolutionary process may take a convoluted path to its endstate, possibly even undoing earlier changes along the way. For very small distances, the problem does not arise, while, for extremely large ones, the problem is essentially insurmountable, as the variance of any estimate will be huge. For most distance values, however, one can view the goal of distance correction as postponing the onset of saturation, that is, making it possible to deliver an accurate estimate of the true distance up to as large a value as possible.

Evolutionary events that affect the gene order of genomes include a number of rearrangements, which affect only the order, as well as gene duplication and loss, which affect both the gene content and, indirectly, the order. Handling both together has proved challenging—first steps were taken recently by Marron *et al.* (2004), Swenson *et al.* (2005), and Swenson *et al.* (2008). Rearrangements themselves include inversion, transposition, and block exchange, which act on a single chromosome, and translocation, fusion, and fission, which act on two chromosomes. Inversion, translocation, fusion, and fission were characterized in the seminal work of Hannenhalli and Pevzner (1995a,b), while Bader *et al.* (2001)

* to whom correspondence should be addressed

showed how to compute edit distances for these operations in linear time. In contrast, transpositions remain poorly understood. Efforts at unifying some of these operations in a statistical framework have had some success (see Durrett *et al*. (2004)). However, Yancopoulos *et al.* (2005) recently defined and studied a unifying operation that can produce each of these rearrangements in one or two steps: the so-called "double-cut-and-join", or DCJ, operation. Bergeron *et al.* (2006) subsequently generalized the DCJ operation and showed how to compute an edit distance for it (assuming that every operation has unit cost) in linear time with a simple formula.

In this paper, we address the problem of estimating a true evolutionary distance under the DCJ model of evolution, assuming no change in gene content and a uniform distribution of all possible DCJ events—the same simplifying assumptions used to date in all rearrangement analyses. Our estimate is in the style of the IEBP estimate for the true inversion distance for a single chromosome due to Wang (2001) (see also Wang and Warnow (2001)), in that it does not require computing an edit distance, but only a simple count of shared gene adjacencies (or, equivalently, breakpoints, as in the seminal work of Sankoff and Blanchette (1998, 1999)) and chromosome endpoints. We characterize the asymptotic behavior of genome structure under the uniform DCJ model and present experimental results showing that our estimates are very precise, and exhibit very little variance, under both realistic and extreme parameter settings.

2 BACKGROUND

2.1 Genomes as Gene-Order Data

A gene is a stranded sequence of DNA that starts with a tail and ends with a head. The tail of a gene a is denoted by a^t and its head by a^h. We write $+a$ ($a^t \rightarrow a^h$) if gene a is transcribed from $3'$ to $5'$ and write $-a$ ($a^h \rightarrow a^t$) otherwise. We are interested, not in the strand of one single gene, but in the connection of two consecutive genes in one chromosome. Due to different strandedness, two consecutive genes b and c can be connected by one *adjacency* of the following four types, $\{b^t, c^t\}$, $\{b^h, c^t\}$, $\{b^t, c^h\}$ and $\{b^h, c^h\}$. If gene d lies at one end of a linear chromosome, the we have a singleton set, $\{d^t\}$ or $\{d^h\}$, called *telomere*.

In the simplest case, we assume equal gene content and no duplicate gene. A *genome* is then represented as a set of adjacencies and telomeres such that the tail or the head of any gene appears in exactly one adjacency or telomere. For example, the genome G illustrated in Figure 1, composed of two linear and one circular chromosomes, $(a, -c, -f)$, (e), and (b, d, b), can be represented by the following set of adjacencies and telomeres: $\{\{a^t\}, \{a^h, c^h\}, \{c^t, f^h\}, \{f^t\}, \{b^h, d^t\}, \{d^h, b^t\}, \{e^t\}, \{e^h\}\}$.

The number of adjacencies and telomeres in one genome only captures the number of linear chromosomes: k adjacencies from circular chromosomes could come from a single circular chromosome of size k or from k circular chromosomes of one gene each, or any other combination. In particular, every genome on n genes made entirely of circular chromosomes has the same number of adjacencies and telomeres.

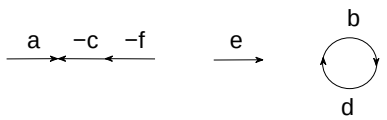

Fig. 1. A very small genome G

2.2 Preliminaries on the DCJ model

The double-cut-and-join operation, in the formulation of Bergeron *et al.* (2006), can model all classical rearrangements: inversion, translocation, fusion, fission, transposition and block interchange. In that formulation, a DCJ operation makes a pair of cuts in the chromosomes and reglues the cut ends on two adjacencies or telomeres (which can be in the same chromosome or in different chromosomes), giving rise to four cases:

1. A pair of adjacencies $\{i^u, j^v\}$ and $\{p^x, q^y\}$ can be replaced by the pair $\{i^u, p^x\}$ and $\{j^v, q^y\}$ or by the pair $\{i^u, q^y\}$ and $\{j^v, p^x\}$.
2. An adjacency $\{i^u, j^v\}$ and a telomere $\{p^x\}$ can be replaced by the adjacency $\{i^u, p^x\}$ and telomere $\{j^v\}$ or by the adjacency $\{j^v, p^x\}$ and telomere $\{i^u\}$.
3. A pair of telomeres $\{i^u\}$ and $\{j^v\}$ can be replaced by the adjacency $\{i^u, j^v\}$.
4. An adjacency $\{i^u, j^v\}$ can be replaced by the pair of telomeres $\{i^u\}$ and $\{j^v\}$.

THEOREM 1. *Let G be a genome with n genes, n_1 adjacencies, and n_2 telomeres. If m is the number of the different possible DCJ operations on G, we can write*

$$\begin{aligned} n &= n_1 + \frac{n_2}{2} \\ m &= n_1^2 + 2n_1n_2 + \frac{1}{2}n_2^2 - \frac{1}{2}n_2 \\ n^2 &\leq m \leq 2n^2 - n \end{aligned}$$

PROOF. G has n genes and thus $2n$ tails and heads of genes; as the tail or the head of any gene appears in exactly one adjacency or telomere, we have

$$2n = 2n_1 + n_2 \tag{1}$$

Now consider the four cases of DCJ operations:

1. There are $\binom{n_1}{2}$ ways to select two adjacencies and 2 possible DCJ operations for each such choice, for a total of $\binom{n_1}{2} \times 2$ operations.
2. There are $n_1 \times n_2$ ways to select one adjacency and one telomere and 2 possible DCJ operations for each combination, for a total of $n_1 \times n_2 \times 2$ operations.
3. There are $\binom{n_2}{2}$ ways to select two telomeres and 1 possible DCJ operation for each such choice, for a total of $\binom{n_2}{2}$ operations.
4. There are n_1 different ways to select one adjacency and 1 possible DCJ operation for each such choice, for a total of n_1 operations.

Thus the total number of possible DCJ operations is

$$m = n_1^2 + 2n_1n_2 + \frac{1}{2}n_2^2 - \frac{1}{2}n_2$$

Combining this result with (1), we get

$$m = -\frac{1}{4}n_2^2 + (n - \frac{1}{2})n_2 + n^2 \tag{2}$$

Now we also have $0 \leq n_2 \leq 2n$, and so we can write

$$n^2 \leq m \leq 2n^2 - n \qquad \square$$

3 METHODS

3.1 An overview of our technique for estimating the true evolutionary distance

The problem of estimating the true evolutionary distance under DCJ model is defined as follows:

Input: The original genome G and the final genome G^F, two genomes on the same n genes represented as adjacencies and telomeres.

Output: An estimate of the actual number of DCJ operations that took place in the evolutionary history to transform G into G^F.

Based on the original genome G, for any genome G^* (of same gene content as G), we can divide the adjacencies and telomeres of G^* into four sets S_A^*, S_T^*, D_A^* and D_T^*, where S_A^* is the set of adjacencies of G^* that also appear in G, S_T^* is the set of telomeres of G^* that also appear in G, D_A^* is the set of adjacencies of G^* that do not appear in G, and D_T^* is the set of telomeres of G^* that do not appear in G. Then we can calculate a vector $V_G(G^*) = (s_A^*, s_T^*, d_A^*, d_T^*)$ to represent the genome G^* based on G, where s_A^*, s_T^*, d_A^* and d_T^* are the cardinalities of the sets S_A^*, S_T^*, D_A^* and D_T^*, respectively. (V_G may be viewed as producing a fingerprint of G^*.) Obviously, we have

$$2n = 2s_A^* + s_T^* + 2d_A^* + d_T^* \tag{3}$$

Let G^k be the genome obtained from $G = G^0$ by applying k randomly selected DCJ operations—under our model, the $(i+1)$st DCJ operation is selected from a uniform distribution of all possible DCJ operations on the current genome G^i. We can compute the vector $V_G(G^k) = (s_A^k, s_T^k, d_A^k, d_T^k)$ to represent the genome G^k with respect to G. In the next section, we will show that, given G, we can also produce the estimate $\widetilde{E}(V_G(G^k)) = (\widetilde{s_A^k}, \widetilde{s_T^k}, \widetilde{d_A^k}, \widetilde{d_T^k})$ for the expected vector $E(V_G(G^k)) = (\overline{s_A^k}, \overline{s_T^k}, \overline{d_A^k}, \overline{d_T^k})$, for any integer $k > 0$. We use $\widetilde{s_A^k}$ to approximate the expected number of adjacencies present in both G and G^k. Our approach for estimating the true evolutionary distance is then to return the integer k that minimizes the difference $|s_A^k - \widetilde{s_A^k}|$.

3.2 Estimation of the expected vector after some number of random DCJ operations

We show how to estimate the expected vector $E(V_G(G^k)) = (\overline{s_A^k}, \overline{s_T^k}, \overline{d_A^k}, \overline{d_T^k})$ under our DCJ model for any integer $k > 0$. Let G and G^k be as defined above; the vector for $G^0 = G$ is clearly just $V_G(G^0) = (n_1, n_2, 0, 0)$. We first show how to compute $E(V_G(G^1))$.

THEOREM 2. *Let m be the number of possible DCJ operations applicable to G. We have $E(V_G(G^1)) = (\overline{s_A^1}, \overline{s_T^1}, \overline{d_A^1}, \overline{d_T^1})$, where*

$$\begin{aligned}
\overline{s_A^1} &= n_1 - \frac{2n_1^2 + 2n_1 n_2 - n_1}{m} \\
\overline{s_T^1} &= n_2 - \frac{2n_1 n_2 + n_2^2 - n_2}{m} \\
\overline{d_A^1} &= \frac{2n_1^2 - 2n_1 + 2n_1 n_2 + \frac{1}{2}n_2^2 - \frac{1}{2}n_2}{m} \\
\overline{d_T^1} &= \frac{2n_1 n_2 + 2n_1}{m}
\end{aligned}$$

PROOF. Write $V_G(G^0) = (s_A^0, s_T^0, 0, 0)$ and consider the four cases for DCJ operations.

1. When we select two adjacencies out of S_A^0 , the number of possible DCJ operations is $\binom{s_A^0}{2} \times 2$. Neither of the resulting adjacencies will be in G, so that every such operation reduces s_A^0 by 2 and increase d_A^0 by 2.
2. When we select one adjacency out of S_A^0 and one telomere out of S_T^0, the number of possible DCJ operations is $s_A^0 \times s_T^0 \times 2$. Neither of the resulting adjacency nor telomere will be in G, so that every such operation reduces both s_A^0 and s_T^0 by 1 and increases both d_A^0 and d_T^0 by 1.
3. When we select two telomeres out of S_T^0, the number of possible DCJ operations is $\binom{s_T^0}{2}$. The resulting adjacency will not be in G, so that every such operation will reduce s_T^0 by 2 and increase d_A^0 by 1.
4. When we select one adjacency out of S_A^0, the number of possible DCJ operations is s_A^0. Neither of the resulting telomeres will be in G, so that every such operation reduces s_A^0 by 1 and increases d_T^0 by 2.

Adding up the 4 cases and normalizing by the total m, we get

$$\begin{aligned}
\overline{s_A^1} &= s_A^0 + \frac{2\binom{s_A^0}{2}}{m} \cdot (-2) + \frac{2s_A^0 s_T^0}{m} \cdot (-1) + \frac{s_A^0}{m} \cdot (-1) \\
&= s_A^0 - \frac{2{s_A^0}^2 + 2s_A^0 s_T^0 - s_A^0}{m} \\
\overline{s_T^1} &= s_T^0 + \frac{s_A^0 \cdot s_T^0 \cdot 2}{m} \cdot (-1) + \frac{\binom{s_T^0}{2}}{m} \cdot (-2) \\
&= s_T^0 - \frac{2s_A^0 s_T^0 + {s_T^0}^2 - s_T^0}{m} \\
\overline{d_A^1} &= 0 + \frac{\binom{s_A^0}{2} \cdot 2}{m} \cdot 2 + \frac{s_A^0 \cdot s_T^0 \cdot 2}{m} \cdot 1 + \frac{\binom{s_T^0}{2}}{m} \cdot 1 \\
&= \frac{2{s_A^0}^2 - 2s_A^0 + 2s_A^0 s_T^0 + \frac{1}{2}{s_T^0}^2 - \frac{1}{2}s_T^0}{m} \\
\overline{d_T^1} &= 0 + \frac{s_A^0 \cdot s_T^0 \cdot 2}{m} \cdot 1 + \frac{s_A^0}{m} \cdot 2 \\
&= \frac{2s_A^0 s_T^0 + 2s_A^0}{m} \qquad \square
\end{aligned}$$

Let G^k be a genome obtained from G by applying k randomly selected DCJ operations and let $\dot{G}^{k+1}$ be the genome obtained from the genome G^k by applying one more randomly selected DCJ operation. We show how to calculate the expected value of $V_G(\dot{G}^{k+1})$ given G^k and G.

THEOREM 3. *Let $V_G(G^k) = (s_A^k, s_T^k, d_A^k, d_T^k)$ and let m_k be the number of possible DCJ operations on G^k. For conciseness,*

write $A^k = s_A^k + d_A^k$ (the number of adjacencies in G^k) and $T^k = s_T^k + d_T^k$ (the number of telomeres in G^k). Then we can write

$$\begin{aligned} m_k &= (A^k)^2 + 2(A^k)(T^k) + \frac{1}{2}(T^k)^2 - \frac{1}{2}(T^k) \\ E(V_G(\dot{G}^{k+1})) &= (\dot{s}_A^{k+1}, \dot{s}_T^{k+1}, \dot{d}_A^{k+1}, \dot{d}_T^{k+1}) \end{aligned}$$

where we have

$$\dot{s}_A^{k+1} = s_A^k + \tfrac{1}{m_k}[\quad n_1 - 2s_A^k(A^k + T^k)] \tag{4}$$

$$\dot{s}_T^{k+1} = s_T^k + \tfrac{1}{m_k}[\quad n_2(T^k + 1) - 2s_T^k(A^k + T^k)] \tag{5}$$

$$\begin{aligned} \dot{d}_A^{k+1} &= d_A^k + \tfrac{1}{m_k}[\quad 2s_A^k(A^k + T^k) + \binom{T^k}{2} \\ &\qquad -(A^k) - n_1] \\ \dot{d}_T^{k+1} &= d_T^k + \tfrac{1}{m_k}[\quad 2s_T^k(A^k + T^k) - n_2(T^k + 1) \\ &\qquad -2\binom{T^k}{2} + 2(A^k)] \end{aligned} \tag{6}$$

PROOF. From Theorem 1, we have

$$m_k = (A^k)^2 + 2(A^k)(T^k) + \frac{1}{2}(T^k)^2 - \frac{1}{2}(T^k)$$

There are $n_1 - s_A^k$ adjacencies in G that do not appear in G^k and they must fall into one the following 3 cases:

1. n_{AA} pairs with members in two different adjacencies in D_A^k.
2. n_{TT} pairs with members in two telomeres of D_T^k.
3. n_{AT} pairs with one member in D_A^k and the other in D_T^k.

There also are $n_2 - s_T^k$ telomeres in G that do not appear in G^k and so must be members of D_A^k.

Now we complete the proof by running through the possible cases. From the proof for Theorem 2, we have already covered 4 cases where adjacencies and telomeres were selected only from S_A^k and S_T^k. The remaining 8 cases cover selections from D_A^k and D_T^k as well. In the last 5 of these 8 cases, the outcome of a particular operation in terms of adjacency and telomere counts is not fixed, but the total count over all possible operations can still be computed; we use the expression "recover" (an adjacency or a telomere) to indicate a case in which the count increases.

1. When we select one adjacency out of S_A^k and another out of D_A^k, the number of possible DCJ operations is $s_A^k \times d_A^k \times 2$. Neither resulting adjacency will be in G, so that every such operation reduces s_A^k by 1 and increases d_A^k by 1.
2. When we select one adjacency out of S_A^k and one telomere out of D_T^k, the number of possible DCJ operations is $s_A^k \times d_T^k \times 2$. Neither the resulting adjacency nor telomere will be in G, so that every such operation reduces s_A^k by 1 and increases d_A^k by 1.
3. When we select one telomere out of S_T^k and one telomere out of D_T^k, the number of possible DCJ operations is $s_T^k \times d_T^k$. Neither the resulting adjacency nor telomere will be in G, so that every such operation reduces s_T^k and d_T^k by 1 and increases d_A^k by 1.
4. When we select one telomere out of S_T^k and one adjacency out of D_A^k, the number of possible DCJ operations is $s_T^k \times d_A^k \times 2$. The resulting adjacency will not be in G, while the resulting telomere can be in G or not. There are $s_T^k \times (n_2 - s_T^k)$ ways to recover one telomere out of $n_2 - s_T^k$ telomeres in G that do not appear in G^k.
5. When we select two adjacencies out of D_A, the number of possible DCJ operations is $\binom{d_A}{2} \times 2$. The two resulting adjacencies can be in G or not. There are n_{AA} ways to recover one adjacency out of $n_1 - s_A^k$ adjacencies in G that do not appear in G^k.
6. When we select one one adjacency out of D_A^k and one telomere out of D_T^k, the number of possible DCJ operations is $d_A^k \times d_T^k \times 2$. The resulting adjacency and telomere can be in G or not. There are $d_T^k \times (n_2 - s_T^k)$ ways to recover one telomere out of $n_2 - s_T^k$ telomeres in G that do not appear in G^k and n_{AT} ways to recover one adjacency out of $n_1 - s_A^k$ adjacencies in G that do not appear in G^k.
7. When we select one adjacency out of D_A^k, the number of possible DCJ operations is d_A^k. The two resulting telomeres can be in G or not and there are $n_2 - s_T^k$ ways to recover one telomere out of $n_2 - s_T^k$ telomeres in G that do not appear in G^k.
8. When we select two telomeres out of D_T^k, the number of possible DCJ operations is $\binom{d_T}{2}$. The resulting adjacency can be in G or not and there are n_{TT} ways to recover one adjacency out of $n_1 - s_A^k$ adjacencies in G that do not appear in G^k.

Adding up the 12 cases and normalizing by the total m_k, we get

$$\begin{aligned} \dot{s}_A^{k+1} &= s_A^k + \tfrac{1}{m_k}[\quad n_1 - 2s_A^k(A^k + T^k)] \\ \dot{s}_T^{k+1} &= s_T^k + \tfrac{1}{m_k}[\quad n_2(T^k + 1) - 2s_T^k(A^k + T^k)] \\ \dot{d}_A^{k+1} &= d_A^k + \tfrac{1}{m_k}[\quad 2s_A^k(A^k + T^k) + \binom{T^k}{2} \\ &\qquad -(A^k) - n_1] \\ \dot{d}_T^{k+1} &= d_T^k + \tfrac{1}{m_k}[\quad 2s_T^k(A^k + T^k) - n_2(T^k + 1) \\ &\qquad -2\binom{T^k}{2} + 2(A^k)] \quad \square \end{aligned}$$

Given G^0, we estimate $E(V_G(G^k))$ for $k > 0$ by iterating k times the matching formula in Theorem 3, and every time we identify $E(V_G(G^{k-1}))$ with the actual vector $V_G(G^{k-1})$.

COROLLARY 1. *Let G be one genome on n genes, the estimated vector $\widetilde{E}(V_G(G^i)) = (\widetilde{s_A^i}, \widetilde{s_T^i}, \widetilde{d_A^i}, \widetilde{d_T^i})$ for all integers i $(0 \le i \le k)$ can be computed in $O(k)$ time.*

3.3 Asymptotic behavior of our estimation

We can use our results to derive the "stable" structure of a genome under the random DCJ model—the structure reached after sufficiently many events.

COROLLARY 2. *Let G have n $(n \ge 2)$ genes; then the estimated vector $\widetilde{E}(V_G(G^k)) = (\widetilde{s_A^k}, \widetilde{s_T^k}, \widetilde{d_A^k}, \widetilde{d_T^k})$ and the estimated number of possible DCJ operation $\widetilde{m}_k$ for genome G^k satisfy*

$$\lim_{k \to +\infty} (\widetilde{s_T^k} + \widetilde{d_T^k}) = \sqrt{2n} \tag{7}$$

$$\lim_{k \to +\infty} \widetilde{m}_k = n^2 + n\sqrt{2n} - \frac{n}{2} - \frac{\sqrt{2n}}{2} \tag{8}$$

The fairly technical proof is attached in appendix; the approach is to define $\widetilde{s_T^0} + \widetilde{d_T^0} = \sqrt{2n} + \varepsilon_0$, with $-\sqrt{2n} \leq \varepsilon_0 \leq 2n - \sqrt{2n}$, and to consider separately the cases where ε_0 is positive and negative, showing in each case that ε_k keeps the sign of ε_0 and that the limit of ε_k as k grows is zero.

COROLLARY 3. *If the estimated vector is* $\widetilde{E}(V_G(G^k)) = (\widetilde{s_A^k}, \widetilde{s_T^k}, \widetilde{d_A^k}, \widetilde{d_T^k})$ *and if we have* $n \geq 2$*, then we can write*

$$\begin{aligned}
\lim_{k\to+\infty} \widetilde{s_A^k} &= \frac{n_1}{2n + \sqrt{2n}} \\
\lim_{k\to+\infty} \widetilde{s_T^k} &= \frac{n_2(\sqrt{2n}+1)}{2n + \sqrt{2n}} \\
\lim_{k\to+\infty} \widetilde{d_A^k} &= n - \frac{\sqrt{2n}}{2} - \frac{n_1}{2n + \sqrt{2n}} \\
\lim_{k\to+\infty} \widetilde{d_T^k} &= \sqrt{2n} - \frac{n_2(\sqrt{2n}+1)}{2n + \sqrt{2n}}
\end{aligned}$$

PROOF. We first calculate $\lim_{k\to+\infty} \widetilde{s_A^k}$. From formula (4) in Theorem 3 and formula (3), we have

$$\widetilde{s_A^{i+1}} = \widetilde{s_A^i} + \frac{1}{m_i}[n_1 - \widetilde{s_A^i}(2n + (\widetilde{s_T^i} + \widetilde{d_T^i}))] \qquad (9)$$

Combining formulae (7), (8), and (9), together with $0 \leq s_A^0 (= n_1) \leq 2n$, we get

$$\lim_{k\to+\infty} \widetilde{s_A^k} = \frac{n_1}{2n + \sqrt{2n}}$$

Similarly, we can calculate the limits for $\widetilde{s_T^k}$, $\widetilde{d_A^k}$ and $\widetilde{d_T^k}$. □

COROLLARY 4. *If we have* $n_1 \geq 1$*, then our estimated value* $\widetilde{s_A^k}$ *decreases monotonically with k until* $\widetilde{s_A^k} \leq \frac{1}{2}$.

PROOF. From the assumption $n_1 \geq 1$, we have $\widetilde{s_A^0} = s_A^0 = n_1 \geq 1$. Now it is enough to show that, for any integer k, if we have $\widetilde{s_A^k} > \frac{1}{2}$, then we get $\widetilde{s_A^{k+1}} < \widetilde{s_A^k}$. If we have $\widetilde{s_A^k} > \frac{1}{2}$, then, from formula (4) in Theorem 3, we have,

$$\begin{aligned}
\widetilde{s_A^{k+1}} - \widetilde{s_A^k} &= \frac{1}{m_k}[n_1 - 2\widetilde{s_A^k}(\widetilde{s_A^k} + \widetilde{s_T^k} + \widetilde{d_A^k} + \widetilde{d_T^k})] \\
&< \frac{1}{m_k}[n_1 - (\widetilde{s_A^k} + \widetilde{s_T^k} + \widetilde{d_A^k} + \widetilde{d_T^k})] \\
&\leq -\frac{1}{2m_k}(n_2 + \widetilde{s_T^k} + \widetilde{d_T^k}) \leq 0 \qquad \square
\end{aligned}$$

These three corollaries paint a picture of the long-term consequences on genomic structure of random DCJ events; in particular, they predict that the number of linear chromosomes (half of the number of telomeres) converges to $\frac{\sqrt{2n}}{2}$ and also, intuitively enough, that the number of shared adjacencies, $\widetilde{s_A^k}$, goes down to (effectively) zero. The prediction of the asymptotic number of linear chromosomes illustrates the limitations of the method: for humans, for instance, using a figure of $25,000$ genes, we get an asymptotic number of 112 chromosomes—and it is to be doubted whether, even with a billion more years of evolution, the human genome would ever feature these many chromosomes. Another example is that of bacteria, which usually have a single circular chromosome, not the 22–50 linear chromosomes that would go with $1,000$–$5,000$ genes. Clearly, the uniform model is too simple and constraints exist in organismal genomes that strongly dampen chromosomal fission. At present, however, there are too many ways in which to impose such constraints within the DCJ model and not enough data to decide which way is best.

4 EXPERIMENTS

We now present experimental results on the accuracy of our estimation of the expected vector after a given number of random DCJ operations and on the quality of our estimator for the true evolutionary distance (in terms of the actual number of DCJ operations). Our experiments all start with an original genome, G, with some chosen number of linear and circular chromosomes of various sizes; this genome is subjected to a prescribed number k of DCJ operations chosen uniformly at random to obtain a final genome G^k. We vary k from one to six times the number of genes—very large values in evolutionary terms. For each choice of parameters, we generate $10,000$ runs to obtain a tight estimate of variance. We compute the vector representations for all intermediate genomes and then use our method to estimate the evolutionary distance. We ran tests on a large variety of initial genomes, some designed to resemble actual organismal genomes, some entirely unrealistic to test extreme cases. Due to space limitations, we present results on just three initial genomes, all meant to resemble real organismal genomes: (a) $25,000$ genes and 25 linear chromosomes; (b) $10,000$ genes and 5 linear chromosomes; and (c) $1,000$ genes and 1 circular chromosome—the first two examples match metazoan genomes, the last matches a small bacterial genome.

4.1 Accuracy of the expected vector after k random DCJ operations

We study the behavior of our estimation $\widetilde{E}(V_G(G^k))$ by comparing its prediction to the sample mean for $E(V_G(G^k))$, as computed from our $10,000$ trials. We include an additional, extreme, genome with $5,000$ genes and $2,500$ linear chromosomes to show that our technique works across a very broad range of parameters. In all of our experiments, we find that $\widetilde{E}(V_G(G^k))$ is extremely close to the sample mean for $E(V_G(G^k))$: the maximum absolute error of corresponding values between our estimation and the sample mean is always less than 0.8. Figure 2 shows the four values in the vector as a function of the actual number of DCJ operations for genome (a) (the curves for genomes (a), (b), and (c) are similar) and for the "extreme" genome (where the curves are better differentiated). The figure does not distinguish our estimation $\widetilde{E}(V_G(G^k))$ and the sample mean for $E(V_G(G^k))$, because the difference is too small with respect to the actual value. We also compute the mean absolute difference for s_A, s_T, d_A, and d_T between our estimation $\widetilde{E}(V_G(G^k))$ and each experimental vector $V_G(G^k)$ in every single run for genomes (a), (b), and (c) and show the results in Figure 3. The sum of absolute difference of entries in the vector on the larger genomes never exceeds 0.5% (as a percentage of the sum of entries in the vector) and is typically well below 0.25%; even on the smaller genome, the difference does not exceed 2% and is typically below 1%.

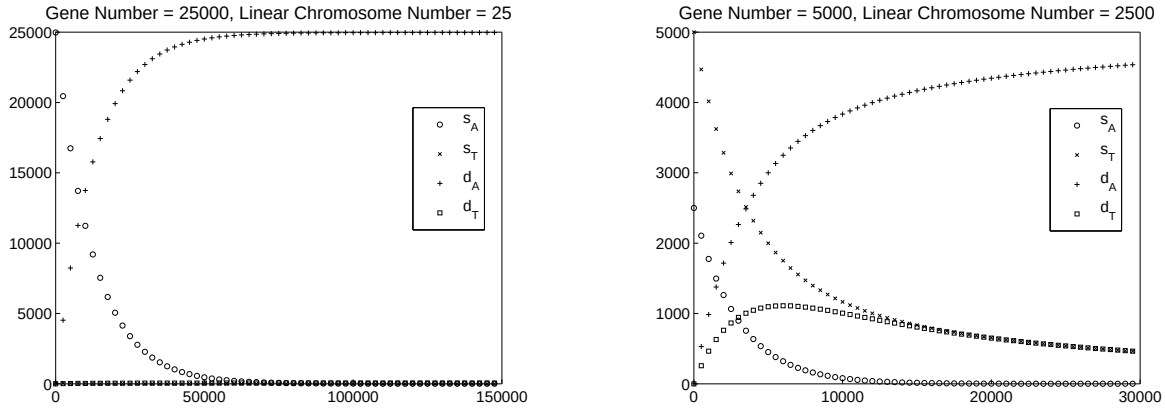

Fig. 2. The four vector values as a function of the actual number of DCJ operations.

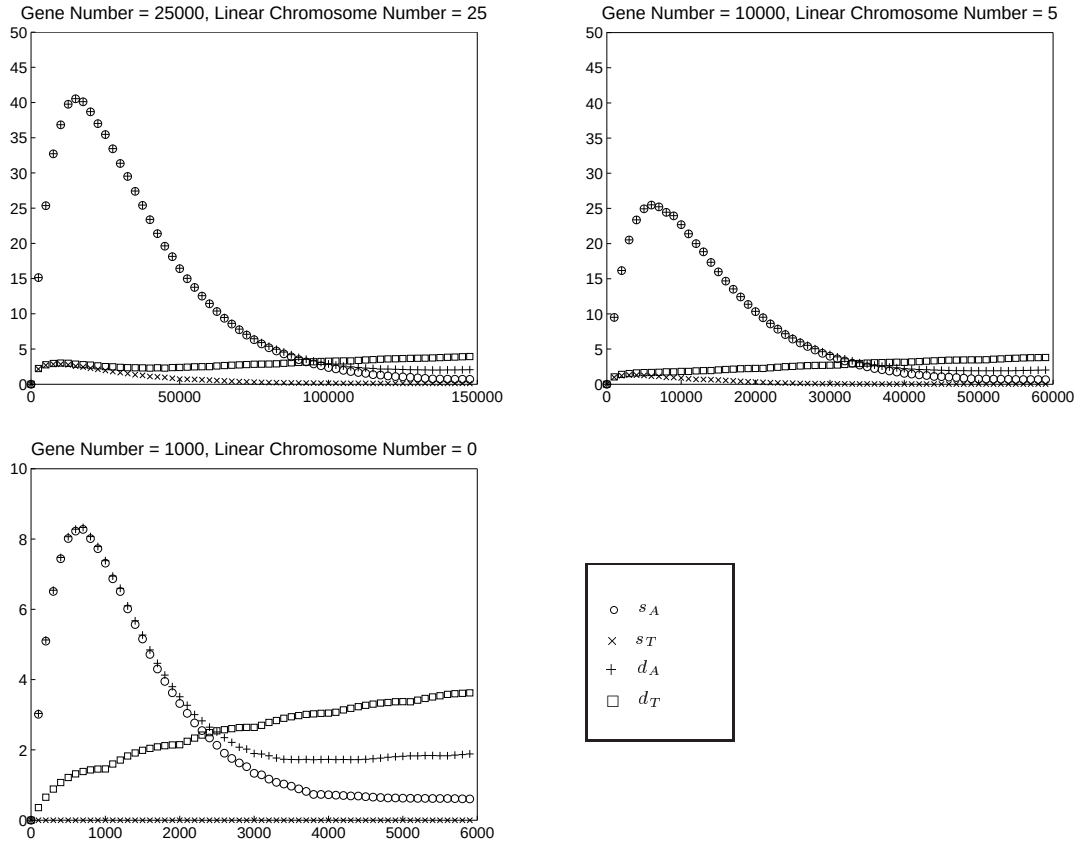

Fig. 3. The mean absolute difference for s_A, s_T, d_A and d_T between our estimation $\widetilde{E}(V_G(G^k))$ and each experimental vector $V_G(G^k)$ as a function of the actual number of DCJ operations.

4.2 Accuracy of the estimation of the actual number of DCJ operations

We want to study the threshold of saturation of our estimator in addition to its accuracy; in order to do that, we create simulations with controlled numbers of DCJ operations and set up a threshold for correction in the estimation procedure. Specifically, we choose a number between 1 and some upper bound B as the actual number of DCJ operations; B is chosen to be the smallest integer k that makes the expected value $\widetilde{s_A^k}$ smaller than 2, a point at which there are almost no shared adjacencies left (from Corollary 4). For genomes

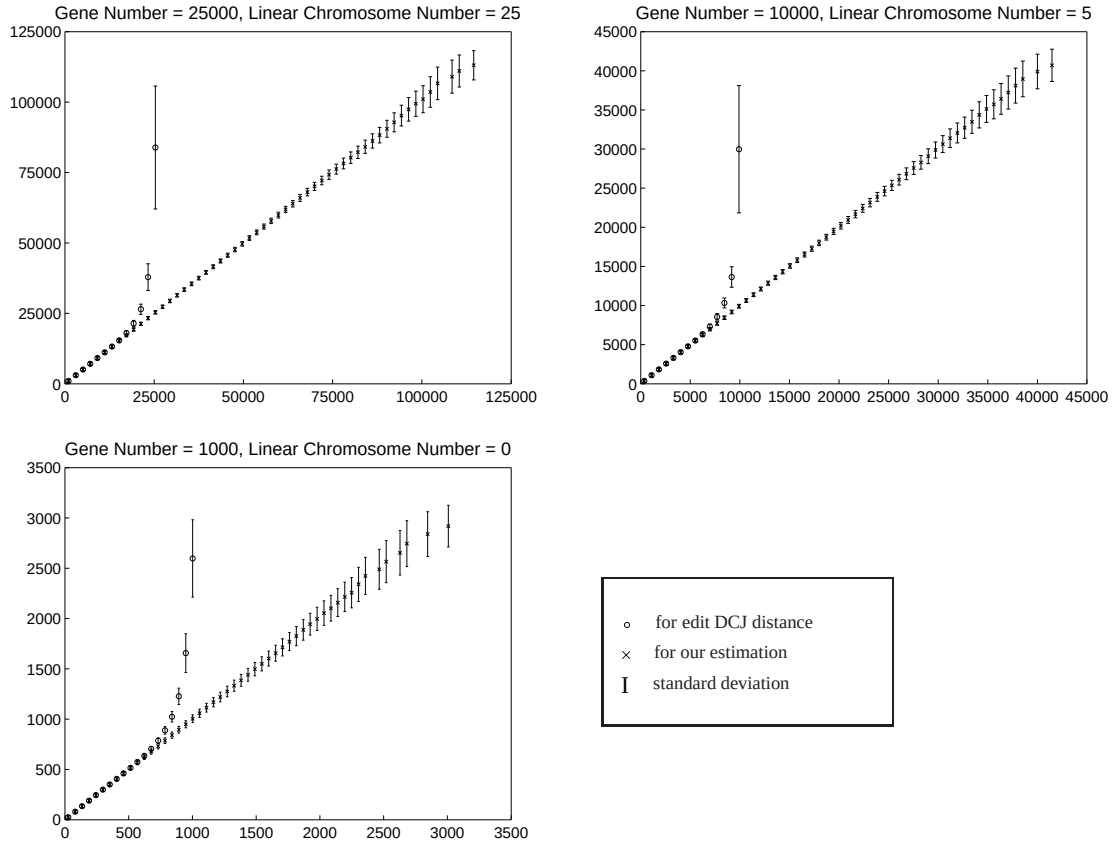

Fig. 4. Mean and standard deviation plots for the actual number of DCJ operations (y axis) vs. the edit DCJ distance and our estimator (x axis). The datasets are divided into 60 bins according to their x-coordinate values.

(a), (b) and (c), the corresponding upper bounds are $121,621$, $44,047$, and $3,253$, respectively. From Corollaries 3 and 4, and the fact $n_1 \leq n$, we have $0 \leq \lim_{k\to+\infty} \widetilde{s_A^k} < \frac{1}{2}$. Thus we use the smallest integer r that causes the expected value $\widetilde{s_A^r}$ to become smaller than $\frac{1}{2}$ as an upper limit on the maximum number of DCJ operations in the evolutionary history. Finally, if we have $s_A^F = 0$, we set k (the value normally chosen to minimize $|s_A^F - \widetilde{s_A^k}|$) to this upper limit r. For genomes (a), (b) and (c), r has values $211,442$, $81,329$, and $6,398$, respectively.

Figure 4 shows the mean and standard deviation for the actual number of DCJ operations estimated by the edit DCJ distance and by our approach. These figures indicate that, as expected, the edit DCJ distance underestimates, often severely, the actual number of events. In contrast, our approach provides highly accurate estimates, with very small variance.

We also study the mean absolute difference between the actual number of DCJ operations and our estimator for genomes (a), (b) and (c). The results are shown in Table 1. The estimates are highly accurate (even for small genomes) up to surprisingly large numbers of events. Rearrangements events fall under the category of "rare genomic events" (in the terminology of Rokas and Holland (2000)), yet our estimator works well even for what would be considered common events.

Table 1. *The mean absolute difference between actual number of DCJ operations and our estimation.*

	actual number of DCJ operations		
# genes	# genes $\times 1$	# genes $\times 2$	# genes $\times 3$
25,000	131.0 (0.5%)	447.5 (0.9%)	1280.2 (1.7%)
10,000	83.9 (0.8%)	282.0 (1.4%)	819.4 (2.7%)
1,000	27.2 (2.7%)	93.6 (4.7%)	441.8 (14.7%)

5 DISCUSSION AND CONCLUSIONS

From Sections 4.1 and 4.2, our estimation achieves very high accuracy, especially for larger (metazoan) genomes. From Figure 4, our approach postpones the threshold of saturation (viewed as a number of DCJ operations) from well under the number of genes to at least three times the number of genes for all three example genomes. This large gain in accuracy should translate into much better phylogenetic reconstructions as well as more accurate genomic alignments.

Moreover, Corollaries 2 and 3 make specific predictions about the structure of evolved genomes on n genes after many steps:

namely, that all should have approximately $\sqrt{2n}$ telomeres, that is $\frac{\sqrt{2n}}{2}$ linear chromosomes, and that shared adjacencies with other highly evolved genomes should be nearly absent. While the second prediction is intuitively reasonable, it is in fact rarely satisfied in actual organisms, especially for small genomes (such as prokaryotic genomes), where conservation pressures are very high and specific structures such as operons survive across broad ranges of evolutionary divergence. The first prediction is, as noted earlier, nearly always overstated: clearly, biological constraints prevent chromosomal fission to be as commonplace as the uniform DCJ mechanism would appear to suggest.

These predictions rely on the two main assumptions made in this work: no gene duplication or loss; and uniform distribution of DCJ operations. Both are clearly unrealistic, so our ability to gauge their effect on model predictions is crucial to future model refinements.

For instance, in their original paper, Yancopoulos *et al.* (2005) required that a chromosomal fission that creates a new small circular chromosome be immediately followed by a chromosomal fusion that re-absorbs this small circular chromosome, thereby causing a block exchange within the original chromosome and treating the extra circular chromosome as a transient artifact. Since circular chromosomes do not arise in organisms with a number of linear chromosomes, a similar constraint would strongly reduce the incidence of fission. A similar type of constraint could be used for prokaryotic genomes, which normally consist of a single circular chromosome. Using just such a constraint, Kothari and Moret (2007) found that the DCJ edit distance closely reflected both inversion and transposition operations. Evidence that paracentric rearrangements are more common than pericentric ones, at least in two *Drosophila* species (see York *et al.* (2007)), and that short inversions are more common than long ones, in some prokaryotes and in the aforementioned *Drosophila* (see Lefebvre *et al.* (2003); York *et al.* (2007)), can also be reflected into additional constraints on the DCJ model. Any additional constraint naturally creates complications, but we expect that at least a few natural constraints can be handled within the framework described here.

Some significant advances have been made by our group for handling duplications and losses in an inversion context (see, e.g., Tang *et al.* (2004); Marron *et al.* (2004); Swenson *et al.* (2005)); since duplications and losses are handled in that work mostly through the greedy approach of using rearrangements to place together genes that can then be gained or lost in a single operation, moving this work to a DCJ context appears uncomplicated. This combination could then yield the first reliable estimate of genomic pairwise distances for complex metazoan genomes based on rearrangements and duplication/loss mechanisms.

Finally, since the DCJ operation regroups all rearrangements studied to date, and since our results point to one way in which the behavior of this model can be studied for various constraints (such as where the cuts can be made), our results may shed light on the vexing issue of what constitutes a significant syntenic block in comparative genomics—an issue that has seen a lot of discussion over the last few years. (Sinha and Meller (2008) give a review of these discussions and some proposals, while Chaisson *et al.* (2006) give evidence that rearrangements occur at multiple scales.)

REFERENCES

Bader, D., Moret, B., and Yan, M. (2001). A fast linear-time algorithm for inversion distance with an experimental comparison. *J. Comput. Biol.*, **8**(5), 483–491.

Bergeron, A., Mixtacki, J., and Stoye, J. (2006). A unifying view of genome rearrangements. In *Proc. 6th Int'l Workshop Algs. in Bioinformatics (WABI'06)*, number 4175 in Lecture Notes in Computer Science, pages 163–173. Springer Verlag.

Chaisson, M., Raphael, B., and Pevzner, P. (2006). Microinversions in mammalian evolution. *Proc. Nat'l Acad. Sci., USA*, **103**(52), 19,824–19,829.

Durrett, R., Nielsen, R., and York, T. (2004). Bayesian estimation of the genomic distance. *Genetics*, **166**(1), 621–629.

Hannenhalli, S. and Pevzner, P. (1995a). Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *Proc. 27th Ann. ACM Symp. Theory of Comput. (STOC'95)*, pages 178–189. ACM Press.

Hannenhalli, S. and Pevzner, P. (1995b). Transforming mice into men (polynomial algorithm for genomic distance problems). In *Proc. 36th Ann. IEEE Symp. Foundations of Comput. Sci. (FOCS'95)*, pages 581–592. IEEE Press.

Kothari, M. and Moret, B. (2007). An experimental evaluation of inversion- and transposition-based genomic distances. In *Proc. 3rd IEEE Symp. on Comput. Intelligence in Bioinformatics and Comput. Biol. (CIBCB'07)*, pages 151–158. IEEE Press.

Lefebvre, J.-F., El-Mabrouk, N., Tillier, E., and Sankoff, D. (2003). Detection and validation of single gene inversions. In *Proc. 11th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'03)*, volume 19 of *Bioinformatics*, pages i190–i196.

Marron, M., Swenson, K., and Moret, B. (2004). Genomic distances under deletions and insertions. *Theor. Computer Science*, **325**(3), 347–360.

Moret, B., Wang, L.-S., Warnow, T., and Wyman, S. (2001). New approaches for reconstructing phylogenies from gene-order data. In *Proc. 9th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'01)*, volume 17 of *Bioinformatics*, pages S165–S173.

Moret, B., Tang, J., Wang, L.-S., and Warnow, T. (2002). Steps toward accurate reconstructions of phylogenies from gene-order data. *J. Comput. Syst. Sci.*, **65**(3), 508–525.

Moret, B., Tang, J., and Warnow, T. (2005). Reconstructing phylogenies from gene-content and gene-order data. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*, pages 321–352. Oxford University Press.

Rokas, A. and Holland, P. (2000). Rare genomic changes as a tool for phylogenetics. *Trends in Ecol. and Evol.*, **15**, 454–459.

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

Sankoff, D. and Blanchette, M. (1999). Probability models for genome rearrangement and linear invariants for phylogenetic inference. In *Proc. 3rd Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB'99)*, pages 302–309. ACM Press.

Sinha, A. and Meller, J. (2008). Sensitivity analysis for reversal distance and breakpoint reuse in genome rearrangements. In *Proc. 13th Pacific Symp. on Biocomputing (PSB'08)*, pages 37–48. World Scientific.

Swenson, K., Marron, M., Earnest-DeYoung, J., and Moret, B. (2005). Approximating the true evolutionary distance between two genomes. In *Proc. 7th SIAM Workshop on Algorithm Engineering & Experiments (ALENEX'05)*, pages 121–129. SIAM Press.

Swenson, K., Arndt, W., Tang, J., and Moret, B. (2008). Phylogenetic reconstruction from complete gene orders of whole genomes. In *Proc. 6th Asia Pacific Bioinformatics Conf. (APBC'08)*, number 6 in Advances in Bioinformatics and Computational Biology, pages 241–250. Imperial Press.

Swofford, D., Olsen, G., Waddell, P., and Hillis, D. (1996). Phylogenetic inference. In D. Hillis, B. Mable, and C. Moritz, editors, *Molecular Systematics*, pages 407–514. Sinauer Assoc., Sunderland, MA.

Tang, J., Moret, B., Cui, L., and dePamphilis, C. (2004). Phylogenetic reconstruction from arbitrary gene-order data. In *Proc. 4th IEEE Symp. on Bioinformatics and Bioengineering BIBE'04*, pages 592–599. IEEE Press.

Wang, L.-S. (2001). Exact-IEBP: A new technique for estimating evolutionary distances between whole genomes. In *Proc. 33rd Ann. ACM Symp. Theory of Comput. (STOC'01)*, pages 637–646. ACM Press.

Wang, L.-S. and Warnow, T. (2001). Estimating true evolutionary distances between genomes. In *Proc. 1st Int'l Workshop Algs. in Bioinformatics (WABI'01)*, number 2149 in Lecture Notes in Computer Science, pages 176–190. Springer Verlag.

Yancopoulos, S., Attie, O., and Friedberg, R. (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, **21**(16), 3340–3346.

York, T., Durrett, R., and Nielsen, R. (2007). Dependence of paracentric inversion rate on tract length. *BMC Bioinformatics*, **8**(115).

APPENDIX

Proof of Corollary 2:

PROOF. We have $0 \leq s_T^0 + d_T^0 (= n_2) \leq 2n$, $\widetilde{s_T^0} = s_T^0$, and $\widetilde{d_T^0} = d_T^0$, and so can write

$$\widetilde{s_T^0} + \widetilde{d_T^0} = \sqrt{2n} + \varepsilon_0$$

with

$$-\sqrt{2n} \leq \varepsilon_0 \leq 2n - \sqrt{2n}$$

We now consider two cases: (i) $-\sqrt{2n} \leq \varepsilon_0 \leq 0$ and (ii) $0 \leq \varepsilon_0 \leq 2n - \sqrt{2n}$. In each case, we prove by induction on k the following result for $\varepsilon_k (= (\widetilde{s_T^k} + \widetilde{d_T^k}) - \sqrt{2n})$:

$$\lim_{k \to +\infty} \varepsilon_k = 0 \tag{10}$$

Case (i) We have $-\sqrt{2n} \leq \varepsilon_0 \leq 0$ and, by inductive hypothesis, $-\sqrt{2n} \leq \varepsilon_i (= \widetilde{s_T^i} + \widetilde{d_T^i} - \sqrt{2n}) \leq 0$. We now bound the range for ε_{i+1} $(= \widetilde{s_T^{i+1}} + \widetilde{d_T^{i+1}} - \sqrt{2n})$.

From formulae (5) and (6) in Theorem 3 as well as formula (3), we have

$$(\widetilde{s_T^{i+1}} + \widetilde{d_T^{i+1}}) = (\widetilde{s_T^i} + \widetilde{d_T^i}) + \frac{1}{m_i}[2n - (\widetilde{s_T^i} + \widetilde{d_T^i})^2]$$

Replacing $(\widetilde{s_T^{i+1}} + \widetilde{d_T^{i+1}})$ and $(\widetilde{s_T^i} + \widetilde{d_T^i})$ by $(\varepsilon_{i+1} + \sqrt{2n})$ and $(\varepsilon_i + \sqrt{2n})$, we get

$$\varepsilon_{i+1} \quad = \quad \varepsilon_i (1 - \frac{1}{m_i}(\varepsilon_i + 2\sqrt{2n})) \tag{11}$$

From formula (2), we have

$$m_i = -\frac{1}{4}(\sqrt{2n} + \varepsilon_i)^2 + (n - \frac{1}{2})(\sqrt{2n} + \varepsilon_i) + n^2 \tag{12}$$

From formula (12) and our inductive hypothesis, and using $n \geq 2$, we get

$$0 \leq 1 - \frac{1}{m_i}(\varepsilon_i + 2\sqrt{2n}) \leq 1 - \frac{\sqrt{2n}}{2n^2 - n} \tag{13}$$

Then from formulae (11) and (13), we can write

$$\varepsilon_i (1 - \frac{\sqrt{2n}}{2n^2 - n}) \leq \varepsilon_{i+1} \leq 0$$

and from the inductive assumption and by using $n \geq 2$, we can verify that ε_{i+1} satisfies

$$-\sqrt{2n} \leq \varepsilon_{i+1} \leq 0$$

Since we have $-\sqrt{2n} \leq \varepsilon_0 \leq 0$, then, by induction, we have, for any integer k,

$$\varepsilon_0 (1 - \frac{\sqrt{2n}}{2n^2 - n})^k \leq \varepsilon_k \leq 0$$

and thus, with $n \geq 2$,

$$\lim_{k \to +\infty} \varepsilon_k = 0$$

Case (ii) We have $0 \leq \varepsilon_0 \leq 2n - \sqrt{2n}$ and, by inductive hypothesis, $0 \leq \varepsilon_i \leq 2n - \sqrt{2n}$. We now bound the range for ε_{i+1} $(= \widetilde{s_T^{i+1}} + \widetilde{d_T^{i+1}} - \sqrt{2n})$.

From formula (12) and the inductive hypothesis, and using $n \geq 2$, we can write

$$0 \leq 1 - \frac{1}{m_i}(\varepsilon_i + 2\sqrt{2n}) \leq 1 - \frac{2\sqrt{2n}}{2n^2 - n} \tag{14}$$

Now using formulae (11) and (14), we get

$$0 \leq \varepsilon_{i+1} \leq \varepsilon_i (1 - \frac{2\sqrt{2n}}{2n^2 - n})$$

and from the inductive hypothesis and using $n \geq 2$, we can verify that ε_{i+1} satisfies

$$0 \leq \varepsilon_{i+1} \leq 2n - \sqrt{2n}$$

Since we have $0 \leq \varepsilon_0 \leq 2n - \sqrt{2n}$, then, by induction, we have, for any integer k,

$$0 \leq \varepsilon_k \leq \varepsilon_0 (1 - \frac{2\sqrt{2n}}{2n^2 - n})^k$$

and thus, with $n \geq 2$,

$$\lim_{k \to +\infty} \varepsilon_k = 0$$

Putting it all together, we have

$$\lim_{k \to +\infty} (\widetilde{s_T^i} + \widetilde{d_T^i}) = \sqrt{2n}$$

Moreover, from formulae (10) and (12), we can write

$$\lim_{k \to +\infty} \widetilde{m}_k \quad = \quad n^2 + n\sqrt{2n} - \frac{n}{2} - \frac{\sqrt{2n}}{2} \quad \square$$