

# A Statistically Fair Comparison of Ancestral Genome Reconstructions, Based on Breakpoint and Rearrangement Distances

ZAKY ADAM<sup>1</sup> and DAVID SANKOFF<sup>2</sup>

## ABSTRACT

**We introduce a way of evaluating two mathematically different optimization approaches to the same problem, namely how good or bad each is with respect to the other's criterion. We illustrate this in a comparison of breakpoint and rearrangement distances between the endpoints of a branch, where total branch-length is minimized in reconstructing ancestral genomes at the nodes of a given phylogeny. We apply this to mammalian genome evolution and simulations under various hypotheses about breakpoint re-use. Reconstructions based on rearrangement distance are superior in terms of branch length and dispersion of the multiple optimal reconstructions, but simulations show that both sets of reconstructions are equally close to the simulated ancestors.**

**Key words:** genomic rearrangements, molecular evolution, sequence analysis, statistics.

## 1. INTRODUCTION

**B**REAKPOINT DISTANCE AND REARRANGEMENT DISTANCES provide alternative ways of evaluating phylogenetic trees and reconstructing ancestral genomes based on whole genome data. When applied to a set of sufficiently diverse genomes, these approaches will generally lead to different results. Ancestral genomes optimal under the breakpoint criterion will not minimize the total rearrangement distance on the edges of a given tree, and optimality according to rearrangement distance will not minimize the total number of breakpoints.

Can we say that one of these methods is superior to the other? Lacking any widely accepted probability model for genomes evolving through rearrangements, we can have no analytic framework for the statistical properties of reconstructions, in particular the accuracy and reliability of these reconstructions. Even assessment through simulations, though informative (Moret et al., 2002), is highly dependent on the assumptions necessary for generating the data, assumptions that are either highly simplified such as uniform weights on a small repertoire of rearrangement events or highly parametrized models pertinent to limited phylogenetic domains.

Is there any sense, then, in which we could affirm that one objective function on reconstructions is better than the other? In this paper we introduce a way of evaluating two different optimization approaches

---

<sup>1</sup>School of Information Technology and Engineering, University of Ottawa, Ottawa, Canada, and Department of Biology, The Pennsylvania State University, State College, Pennsylvania.

<sup>2</sup>Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada.

relative to each other, namely how good or bad each is with respect to the other's criterion. The idea is that *the approach that comes the closest to satisfying the other's criterion as well as its own, is more desirable.*

We will illustrate this method on two data sets on mammalian evolution, as well as three different simulations modelling each of these data sets in a different way, eight data sets in all. We use a given, well-accepted phylogeny for each data set and each simulation. We reconstruct all the ancestral genomes, once minimizing the total breakpoints over all tree branches (using the polynomial-time median method in Tannier et al., 2009), and once with the minimum rearrangement distance (Adam and Sankoff, 2008). Because optimal reconstructions are not unique, we sample five different reconstructions in each case.

We then apply our “fair” method to four aspects of these reconstructions. First we assess all the branch lengths in each reconstruction and then compare, in two different ways, the dispersion (or compactness), for each tree node, of the five different reconstructions. Finally, for the simulated data sets, we measure the distances of the reconstructed ancestors from the simulated ancestors.

For the analysis of the branch lengths and the dispersion of reconstructed ancestors, the results show a clear and systematic advantage of rearrangement distance over breakpoint distance. Despite this, the two methods prove to be equally good at reconstructing the known simulated ancestor genomes.

In Section 2, we first discuss the relationship between breakpoint distances and rearrangement distances. In Section 3, we formalize our proposal for a fair comparison of metrics, illustrating with four aspects of the “small” phylogeny problem, branch lengths, node dispersion (looked at two ways) among optimal solutions, and distance between reconstructed and true ancestral genomes in simulations. In Section 4, we formalize the breakpoint distance and the rearrangement distance, and sketch how they are used in solving the small phylogeny problem. In Section 5, we briefly describe the two real data sets on mammalian evolution as well as the simulations carried out under various “breakpoint re-use” conditions. The results are presented in Section 6 and discussed in Section 7.

## 2. BREAKPOINT DISTANCE AND REARRANGEMENT DISTANCE

We represent genomes here as signed permutations on  $(1, 2, \dots, n)$ , fragmented into a number of chromosomes. The integers represent genes and their sign indicates their strandedness, or reading direction.

The breakpoint distance between two genomes may be defined in a number of ways. The main difference among these definitions pertain to telomeres in one multichromosomal genome that are internal, i.e., non-telomeres in the other. Here we will define the breakpoint distance (Tannier et al., 2009) between two genomes as

$$d_{BP} = n - \text{the number of common gene adjacencies in the genomes} \\ + \frac{1}{2} \text{ the number of common chromosomal endpoints (telomeres)} \quad (1)$$

where gene adjacency requires conservation of their relative orientation (strandedness).

Among the various definitions of rearrangement distance, all closely related to each other, we focus on the DCJ metric (Yancopoulos et al., 2005; Bergeron et al., 2006)  $d_{DCJ}$ , which counts the minimal number of operations necessary to transform one genome into another, where the repertoire of operations includes inversions, reciprocal translocations, chromosome fissions and fusions, as well as block interchanges, which count as two operations. Transposition of chromosomal segments from one site to another are a special case of block interchange. All of these operations, except the general case of block interchange, are chromosomal mutations familiar in classical genetics.

Since each of these operations can create or remove at most two breakpoints, the breakpoint distance is no greater than twice the rearrangement distance. And since it is always possible to convert one genome into another by rearrangements that remove one breakpoint at a time, the rearrangement distance is no greater than the breakpoint distance. For any two genomes  $X$  and  $Y$ , then, we can write (Watterson et al., 1982)

$$\frac{1}{2} d_{BP}(X, Y) \leq d_{DCJ}(X, Y) \leq d_{BP}(X, Y). \quad (2)$$

Aside from the bounds in (2), there is no mathematical constraint between  $d_{DCJ}$  and  $d_{BP}$ . The two distances may be highly correlated for pairs of random genomes, but for a given set of present-day genomes,

phylogenies and ancestral genomes constructed to optimize one of these distances will not in general optimize the other.

It is often said that rearrangement distance is preferable to breakpoint distance in phylogenetic inference because only the former is connected to a model of genome evolution. Because of the minimality criterion inherent in rearrangement distance, however, this distance is virtually always substantially less than the actual number of rearrangements responsible for the divergent gene orders of two genomes (unless they are very closely related), and the actual rearrangements inferred in calculating the distance may bear little resemblance to the actual rearrangements. This is as true for simulations as it is for real genomes. Thus phylogenetic results from rearrangement analysis do not have an a priori privileged status over results from breakpoint analysis based on evolutionary models.

### 3. A FAIR COMPARISON

Let  $\mathcal{P}$  be a phylogeny where each of the  $M$  terminal nodes is labelled by a known genome, and let  $d_A$  and  $d_B$  be two metrics on the set of genomes. Each branch of  $\mathcal{P}$  may be incident to at most one terminal node and at least one of the  $N$  ancestral nodes. For any distance  $d_D$  on the set of genomes, and edge  $XY \in \mathcal{P}$ , by  $d_D(XY)$  we mean the distance between the genomes labelling  $X$  and  $Y$ . Suppose we label the ancestral nodes of  $\mathcal{P}$  using some set of genomes  $R = (G_1, \dots, G_N)$ . We define

$$L_D(R) = \sum_{\text{branch } XY \in \mathcal{P}} d_D(XY). \quad (3)$$

Consider reconstructions  $R_A = (G_1^A, \dots, G_N^A)$  and  $R_B = (G_1^B, \dots, G_N^B)$  of the set of genomes to label the ancestral nodes such that

$$L_A(R_A) = \sum_{\text{branch } XY \in \mathcal{P}} d_A(X^A Y^A) \quad (4)$$

is minimized, and

$$L_B(R_B) = \sum_{\text{branch } XY \in \mathcal{P}} d_B(X^B Y^B) \quad (5)$$

is also minimized, where we use superscripts on the vertices  $X$  and  $Y$  to distinguish between the labels (i.e., genomes) on the vertices determined by the two different distances. (Note that for a terminal node  $X$ , the genomes labelling  $X^A$  and  $X^B$  are the same, given, genome.) Without taking into account any additional, external criteria, there is no justification for saying one of  $d_A$  or  $d_B$  is better as a criterion for reconstructing the ancestral genomes. In general, we can expect that

$$L_A(R_A) < L_A(R_B), \quad (6)$$

i.e., strict inequality holds and

$$L_B(R_B) < L_B(R_A). \quad (7)$$

#### 3.1. Branch lengths

Inequalities (6) and (7) are not necessarily inherited by the individual terms in the sums, e.g.,  $d_A(X^A Y^A)$  may not always be less than  $d_A(X^B Y^B)$ . We call  $e_A(X^B Y^B) = d_A(X^B, Y^B) - d_A(X^A, Y^A)$  the *excess length* of branch  $X^B Y^B$  with respect to distance  $d_A$ , though it may sometimes be less than zero.

As schematized in Figure 1, we will calculate the least squares fit to  $d_A(X^B Y^B) = \alpha_{B/A} d_A(X^A, Y^A)$  and to  $d_B(X^A Y^A) = \alpha_{A/B} d_B(X^B, Y^B)$  to estimate the coefficients  $\alpha_{B/A}$  and  $\alpha_{A/B}$ . These coefficients measure how much larger the  $B$  constructs are compared to the  $A$  ones, in terms of the  $A$  criterion, and how much larger the  $A$  constructs are compared to the  $B$  ones, in terms of the  $B$  criterion, respectively. Then  $\alpha_{B/A} - 1$  is the excess rate of  $B$  with respect to  $d_A$  and  $\alpha_{A/B} - 1$  is the excess rate of  $A$  with respect to  $d_B$ . A metric that induces a reconstruction with a lower excess rate with respect to the other metric may be considered superior. I.e., if the excess rate of  $A$  with respect to  $d_B$  is less than the excess rate of  $B$  with respect to  $d_A$ , then  $d_A$  is better in the sense of being more universal or less “parochial”: the branches in the  $A$

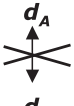
<i>incommensurable</i>	<i>compare</i>	<i>regression</i>	<i>commensurable</i>
	$d_B(X^A Y^A) \leftrightarrow d_B(X^B Y^B)$	$d_B(X^A Y^A) = \alpha_{A/B} d_B(X^B Y^B)$	$\alpha_{A/B} - 1$
	$d_A(X^B Y^B) \leftrightarrow d_A(X^A Y^A)$	$d_A(X^B Y^B) = \alpha_{B/A} d_A(X^A Y^A)$	$\alpha_{B/A} - 1$

FIG. 1. Strategy for comparing different metrics.

reconstruction are closer to optimal length according to  $d_B$  than the branches in the  $B$  reconstruction are according to  $d_A$ .

### 3.2. Node dispersion

Suppose we sample  $N_s$  optimal reconstructions under  $d_A$ . Let  $G_i^A = \{G_{i1}^A, \dots, G_{iN_s}^A\}$  be the set of reconstructions of ancestral genome  $G_i$ . Then

$$V^A(G_i^A) = \max_{0 < j < k \leq N_s} d_A(G_{ij}^A, G_{ik}^A) \quad (8)$$

is the dispersion of the  $N_s$  reconstructions. Insofar as these reconstructions achieve the goal of ancestral inference, they may be expected to cluster around or near the unknown “true” value of  $G_i$  as measured by  $d_A$ . We may also expect that

$$V^A(G_i^A) \leq V^A(G_i^B), \quad (9)$$

where

$$V^A(G_i^B) = \max_{0 < j < k \leq N_s} d_A(G_{ij}^B, G_{ik}^B) \quad (10)$$

since the genomes in  $G_i^B$  are not necessarily close to the true (but unknown)  $G_i$  as measured by  $d_A$ , and similarly

$$V^B(G_i^B) \leq V^B(G_i^A). \quad (11)$$

In both (9) and (11), the inequality would likely be strict.

As in Section 3.1, we will calculate the least squares fit to  $V^A(G_i^B) = \alpha_{B/A} V^A(G_i^A)$  and to  $V^B(G_i^A) = \alpha_{A/B} V^B(G_i^B)$  over all ancestral nodes. Then  $\alpha_{B/A} - 1$  is the excess rate of  $B$  with respect to  $d_A$  and  $\alpha_{A/B} - 1$  is the excess rate of  $A$  with respect to  $d_B$ . As with the branch lengths, we can see whether one of the two metrics has a systematically lower excess rate.

### 3.3. Distance to true genome

In contrast to the real data sets, with the simulated data sets, we actually know the ancestral genomes  $G_i$ . We can expect

---

**Algorithm 1.** Outline of phylogenetic reconstruction based on medians

---

**Algorithm for Small Phylogeny Using Metric  $d$**

**input**  $\mathcal{P}, g_1, \dots, g_M$

**set**  $L = \infty$

**initialize**  $G_1, \dots, G_N$

calculate  $L' = L(G_1, \dots, G_N)$

**while**  $L' < L$

set  $L = L'$

for each  $G_1, \dots, G_N$ , with neighbours  $H, J, K$

$G = d\text{-median}(H, J, K)$

$L' = L(G_1, \dots, G_N)$

**end while**

**output**

---

$$d_A(G_i, G_i^A) \leq d_A(G_i, G_i^B), \quad (12)$$

$$d_B(G_i, G_i^B) \leq d_B(G_i, G_i^A). \quad (13)$$

As in Sections 3.1 and 3.2, we will calculate the least squares fit to  $d_A(G_i, G_i^B) = \alpha_{B/A} d_A(G_i, G_i^A)$  and to  $d_B(G_i, G_i^A) = \alpha_{A/B} d_B(G_i, G_i^B)$  over all ancestral nodes. Then  $\alpha_{B/A} - 1$  is the excess rate of  $B$  with respect to  $d_A$  and  $\alpha_{A/B} - 1$  is the excess rate of  $A$  with respect to  $d_B$ . As with the branch lengths and node dispersion, we can see whether one of the two metrics has a systematically lower excess rate.

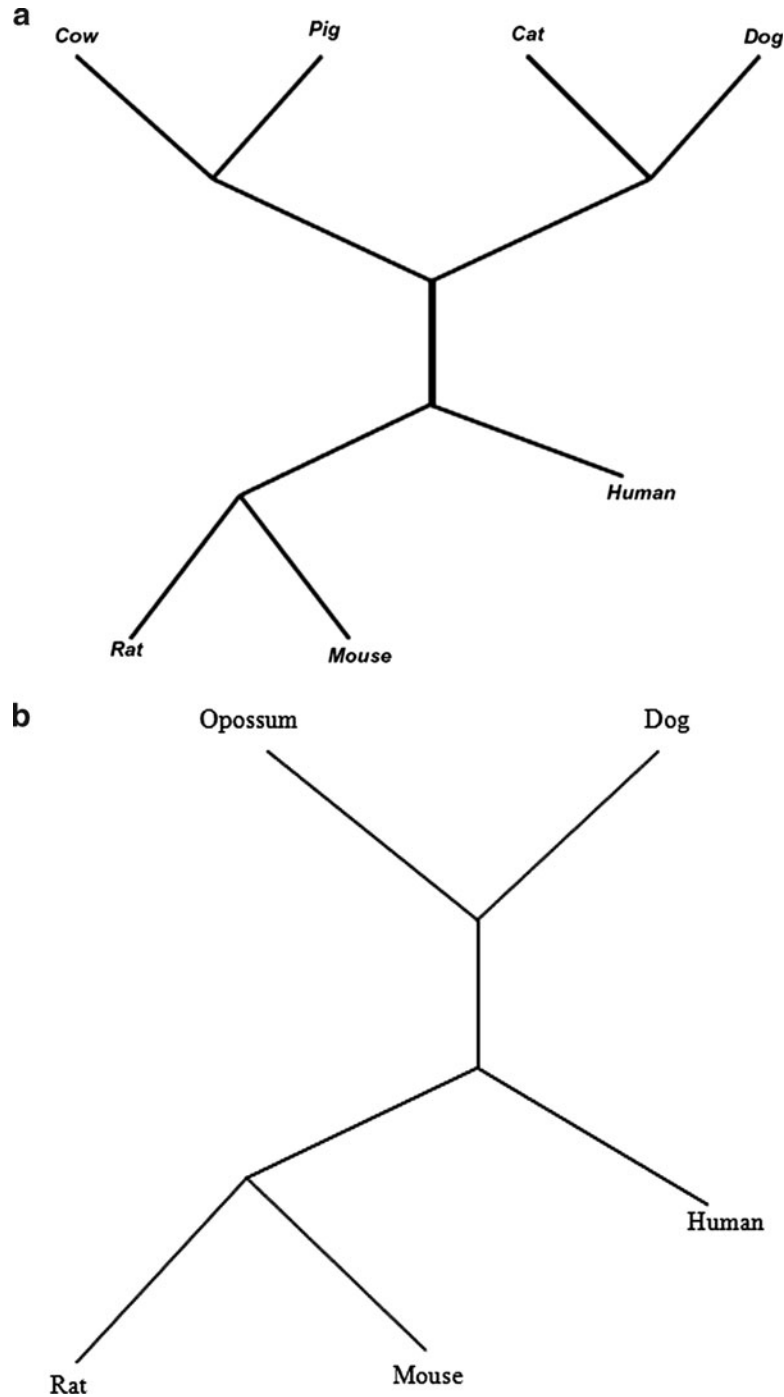


FIG. 2. (a, b) Phylogenies for two mammalian data sets.

#### 4. PHYLOGENY AND THE MEDIAN PROBLEM

Given genomes  $g_1, \dots, g_M$  associated with the terminal nodes of  $\mathcal{P}$ , the small phylogeny problem is to construct a set of genomes  $G_1, \dots, G_N$  to associate with the non-terminal nodes of  $\mathcal{P}$ , such that the phylogenetic tree length  $L$  is minimal under some metric  $d$ , as in Section 3. We consider the simplest structure for  $\mathcal{P}$ , namely an unrooted, binary-branching tree. All nodes are of degree one (terminal) or three (non-terminal). Our algorithms for searching for a minimum  $L$  depend on recurrent use of algorithms for computing the median of three genomes  $H, J$  and  $K$ , which we represent as  $d$ -**median** ( $H, J, K$ ), as shown in the pseudo-code presented as Algorithm 1 here.

Simply stated, the median problem is: considering three genomes  $H, J, K$  as points in some metric space  $(E, d)$ , find another genome  $C \in E$  such that  $d(C, H) + d(C, J) + d(C, K)$  is minimal. There is a large literature on median problems in comparative genomics (Tannier et al., 2009), with all studied versions except one proving to be NP-hard. For our purposes we take  $E$  to be the set of oriented multichromosomal or unichromosomal genomes on the same set of  $n$  elements or genes. The genomes in this set may have circular as well as linear chromosomes, a property which has little consequence for the numerical results of the median problem, but has computational advantages for both the two metrics we study: the double cut and join (DCJ) distance (Yancopoulos et al., 2005; Bergeron et al., 2006), where we have previously implemented and tested code (Adam and Sankoff, 2008) and the breakpoint distance, which gives rise to the only known version of the median problem that is of polynomial complexity, due to Tannier et al. (2009).

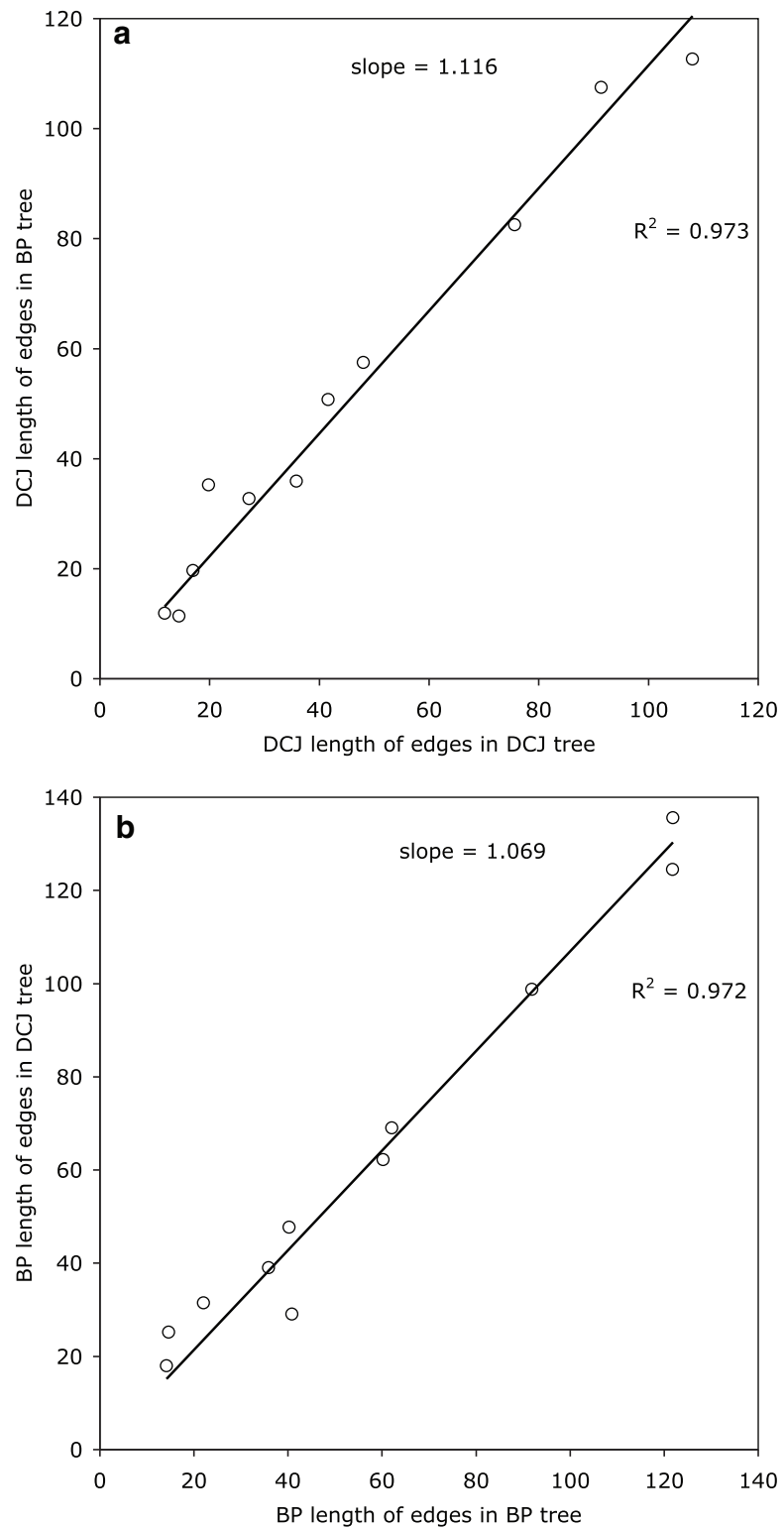
Recall from Section 2 that the breakpoint distance between two genomes is defined to be  $d_{BP} = n - \text{the number of common gene adjacencies} + \frac{1}{2} \text{the number of common chromosomal endpoints}$ . We have carried out the first implementation of the Tannier algorithm. The median problem turns out to be directly transformable to a version of the maximum weight perfect matching problem. We made use of the code in Lau (2006) for this purpose. Although the maximum weight perfect matching algorithm is polynomial, the execution time is not negligible, being at least  $O(n^3)$ .

The DCJ metric  $d_{DCJ}$  counts the minimal number of operations necessary to transform one genome into another, where the repertoire of operations includes inversions, reciprocal translocations, chromosome fissions and fusions, as well as block interchanges, which count as two operations. Transposition of chromosomal segments from one site to another are a special case of block interchange. Our version of the median solver in Adam and Sankoff (2008) is based on the MGR algorithm (Bourque and Pevzner, 2002), with some differences due to the different rearrangement distances used, but also because of extensive use of additional routines to escape from local minima.

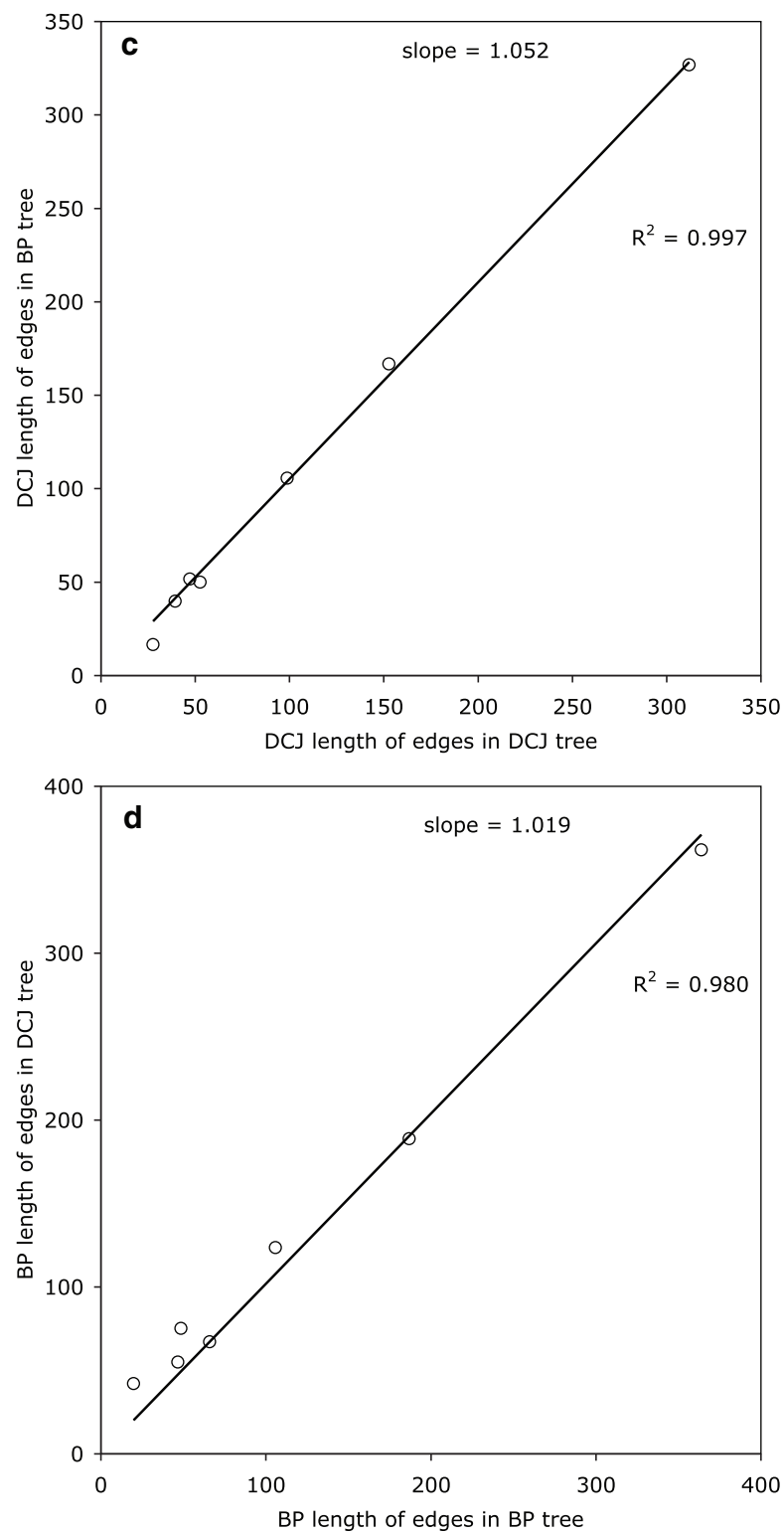
In the breakpoint method, the ancestral gene orders converge after 4-6 iterations within Algorithm 1, and each iteration takes 3-5 minutes for the data we consider in the next section. Our implementation of the DCJ method requires extensive computing time, taking more than 20 iterations to converge on the same data, where each iteration takes many hours to finish.

TABLE 1. EXCESS RATE (IN %) FOR EACH RECONSTRUCTION, MEASURED BY COMPETING CRITERION: AVERAGE BRANCH LENGTH RESULTS

Reconstruction	Dataset			
	Seven placentals		Marsupial, placentals	
	BP	DCJ	BP	DCJ
Real data	5.23	<b>1.94</b>	11.55	<b>6.92</b>
Simulated				
No re-use	6.31	<b>3.05</b>	7.22	<b>5.56</b>
Re-use 2	8.55	<b>2.02</b>	8.65	<b>0.44</b>
Actual re-use	6.06	<b>3.85</b>	8.69	<b>3.30</b>



**FIG. 3.** (a–d) Competing criterion average branch lengths versus reconstructing criterion. Top: seven placentals data. Bottom: placentals plus opossum data.

**FIG. 3.** (Continued)

## 5. THE EMPIRICAL STUDIES

### 5.1. The data

The first data set, drawn from Murphy et al. (2005), consists of the placental mammalian genomes from human, rat, mouse, cat, dog, pig, and cow. Each genome consists of 307 HSB (homologous syntenic blocks). The second data set includes the marsupial opossum along with the placental mammalian genomes human, rat, mouse and dog. These data are from the supplementary information for Mikkelsen et al. (2007). Each genome consists of 603 HSB. The given phylogenies are shown in Figure 2. Although these data sets are obviously not independent, the differences in the species involved, and the large number of extra HSB induced by the presence of the opossum genome, assures that these problems are rather different from the computational point of view.

### 5.2. The sample of optimal reconstructions and breakpoint re-use

For each of the two data sets, we ran the breakpoint phylogeny algorithm 40 times with different random initializations of  $G_1, \dots, G_N$ . We retained the runs that gave the five minimum total tree lengths, usually the same value. For each branch of the tree and for each of the five results, we computed the corresponding DCJ distance between the two breakpoint-inferred genomes determining that branch, and then computed the breakpoint re-use (Bourque and Pevzner, 2002; Sankoff, 2006) quotient  $r = 2d_{DCJ}/d_{BP}$ .

We ran the DCJ phylogeny algorithm five times only with different random initializations of  $G_1, \dots, G_N$ . For each branch of the tree and for each of the five results, we computed the corresponding breakpoint distance between the two DCJ-inferred genomes determining that branch, and then computed the quotient  $r = 2d_{DCJ}/d_{BP}$ .

In the simulations, we make use of the average  $\bar{r}(X, Y)$  of the ten values of the re-use statistic calculated for each branch  $XY$  in these two ways, as well as the average of the ten branch lengths  $\bar{d}_{DCJ}(X, Y)$ .

### 5.3. Simulated data experiments

Corresponding to each of the two data sets, we simulated data sets as follows:

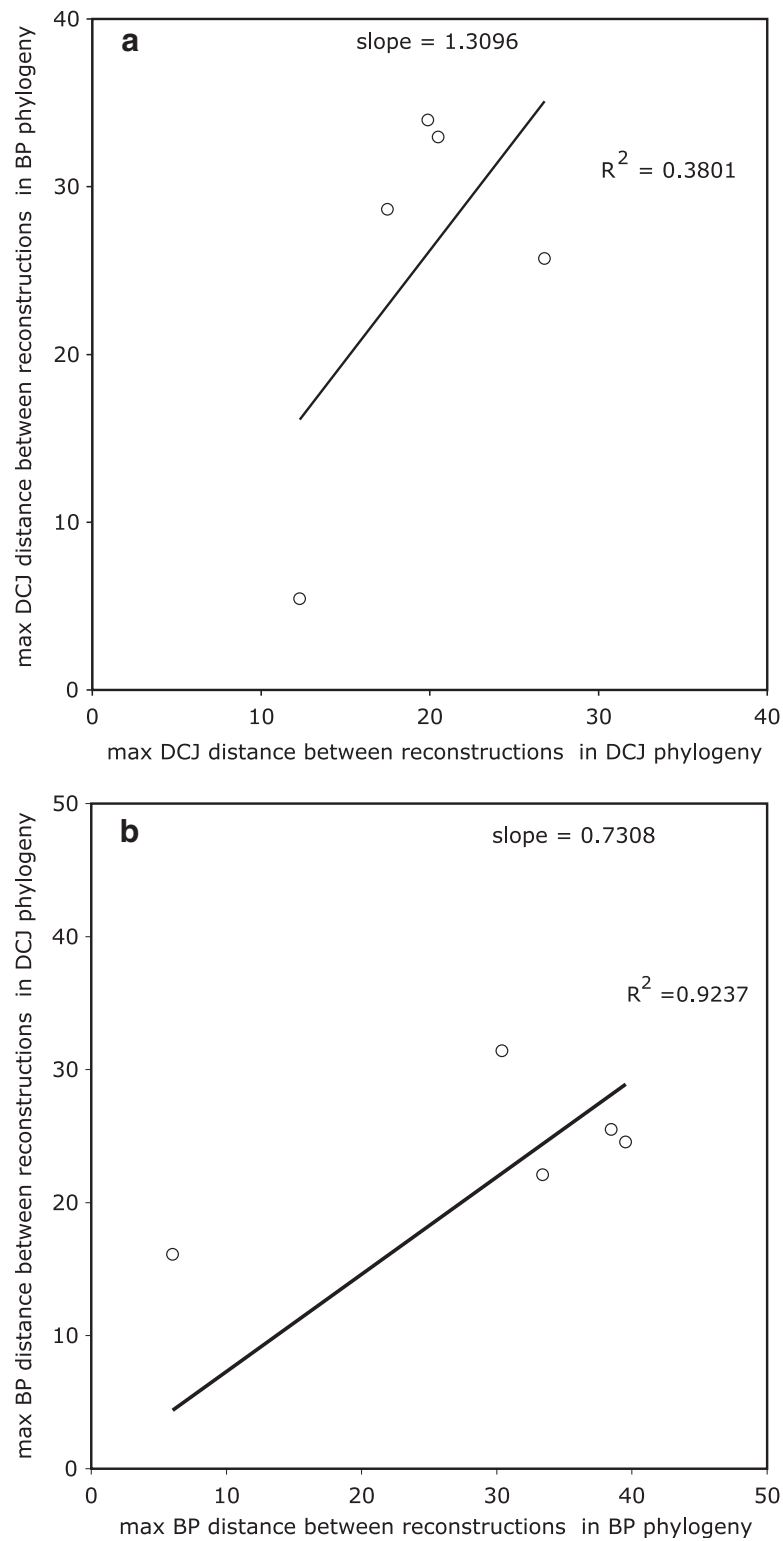
- We arbitrarily chose one of the ancestral nodes as “root.” Because of the reversibility of operations like inversion and reciprocal translocation, the choice of root has little consequence for the simulation and its analysis.
- We generated a random genome at the “root,” distributing the HSB over  $\chi$  chromosomes, where  $\chi = \text{mean}_i^M \chi(g_i)$ , and  $\chi(g_i)$  is the number of chromosomes in given genome  $g_i$ .
- We then generated the children and other descendants of the root along the tree branches using 90% inversions and 10% reciprocal translocations for the  $\bar{d}_{DCJ}(X, Y)$  random rearrangements on branch  $XY$ . The breakpoints for each rearrangement were chosen at random uniformly across the chromosome.
- We ran three separate simulations, constraining the rearrangements on a branch  $XY$  in three different ways: once assuring each operation on  $XY$  used two new breakpoints so that  $r = 1$ , once assuring each successive rearrangement on a branch used one new breakpoint and re-used one existing breakpoint so that  $r = 2 - 1/n$ , and once assuring  $r = \bar{r}(X, Y)$ .
- We then implemented the breakpoint and DCJ phylogeny algorithms exactly as for the real data, obtaining five reconstructions for each criterion.

TABLE 2. EXCESS RATE (IN %) FOR EACH RECONSTRUCTION, MEASURED BY COMPETING CRITERION: MAXIMUM INTRANODE DISTANCE DATA

Reconstruction	Dataset			
	Intranode		Intranode, intercriteria	
	BP	DCJ	BP	DCJ
Real data	34.1	−27.8	11.55	<b>6.92</b>
Simulated				
No re-use	40.0	−35.9	59.2	<b>14.3</b>
Re-use 2	8.09	<b>5.46</b>	<b>13.75</b>	26.4
Actual re-use	5.96	−3.26	62.5	<b>16.6</b>

#### 5.4. The measurements

For each of the  $8 = 2 \text{ criteria} \times (1 \text{ real} + 3 \text{ simulated})$  data sets, we calculated the following aggregate statistics over the reconstructions, in each case measured by *both* criteria, the one used to infer the reconstructions and the opposing one:



**FIG. 4.** (a, b) Competing criterion maximum intranode distance versus reconstructing criterion.

- average branch length  $d_A(X^A, Y^A)$  and  $d_B(X^A, Y^A)$ , for each branch  $XY$ ,
- maximum intranode distance,  $\max_{jk} d_A(G_{ij}^A, G_{ik}^A)$  and  $\max_{jk} d_B(G_{ij}^A, G_{ik}^A)$ , for each  $G_i$ ,
- maximum intranode, intercriteria distance,  $\max_{jk} d_A(G_{ij}^A, G_{ik}^B)$ , for each  $G_i$ ,
- (simulations) average distance between reconstruction and true ancestor,  $d_A(G_i, G_i^A)$  and  $d_B(G_i, G_i^A)$ , for each  $G_i$ .

## 6. RESULTS

Before examining the results, we reiterate that the key to this methodology is that it is fair, i.e., not inherently biased either towards BP or DCJ. Each comparison is made according to a single criterion; we do not compare BP scores with DCJ scores. The BP measurements are slightly worse on the DCJ reconstructions and the DCJ measurements are slightly worse on the BP reconstructions. How bad they are is measured in normalized terms—slope of a least squares line anchored at (0,0). The excess of this slope over 1.0 we call the excess explanatory rate (of using the ancestral genomes reconstructed under criterion A instead of those reconstructed under criterion B, when B is used to make the measurements). The smaller this cost, the closer the A ancestors are to being solutions, not only under criterion A but also B. Thus we see that the excess rate of the DCJ reconstructions is systematically lower than that of the BP reconstructions, for both real data sets and for all the simulations.

### 6.1. Average branch length

In Table 1, we see that the excess rate for DCJ (in boldface), which measures how far the DCJ reconstructions are from being BP-optimal, is only of the order of half the excess rate of the BP reconstructions. This is true in both real data sets and in all the simulations, regardless of re-use rate.

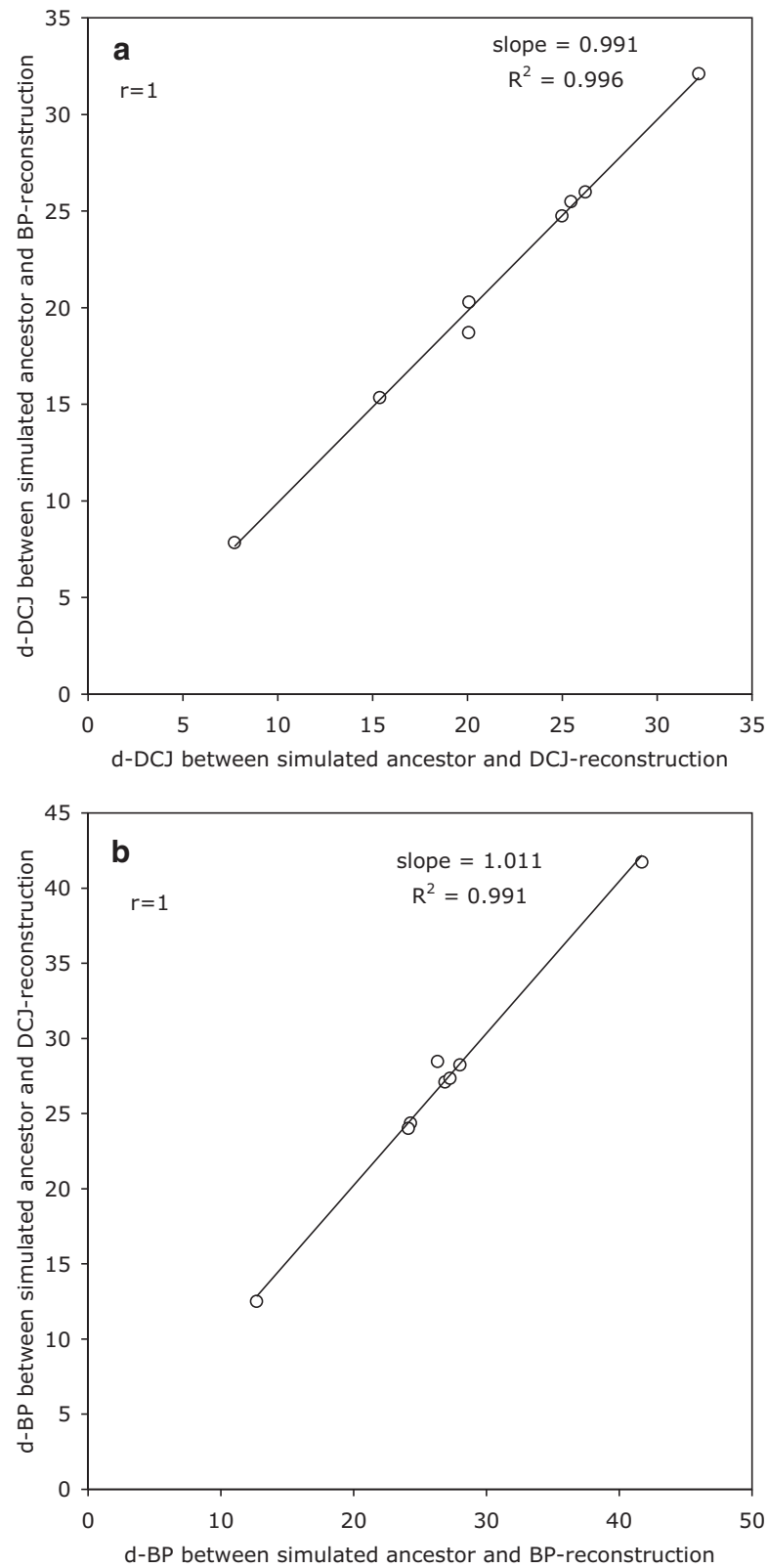
To illustrate the derivation of the excess rates as detailed in Section 3.1 and Figure 1, we present scattergrams of competing criterion versus reconstructing criterion branch lengths in Figure 3 for the real data sets, corresponding to the top row in Table 1.

### 6.2. Maximum intranode distance

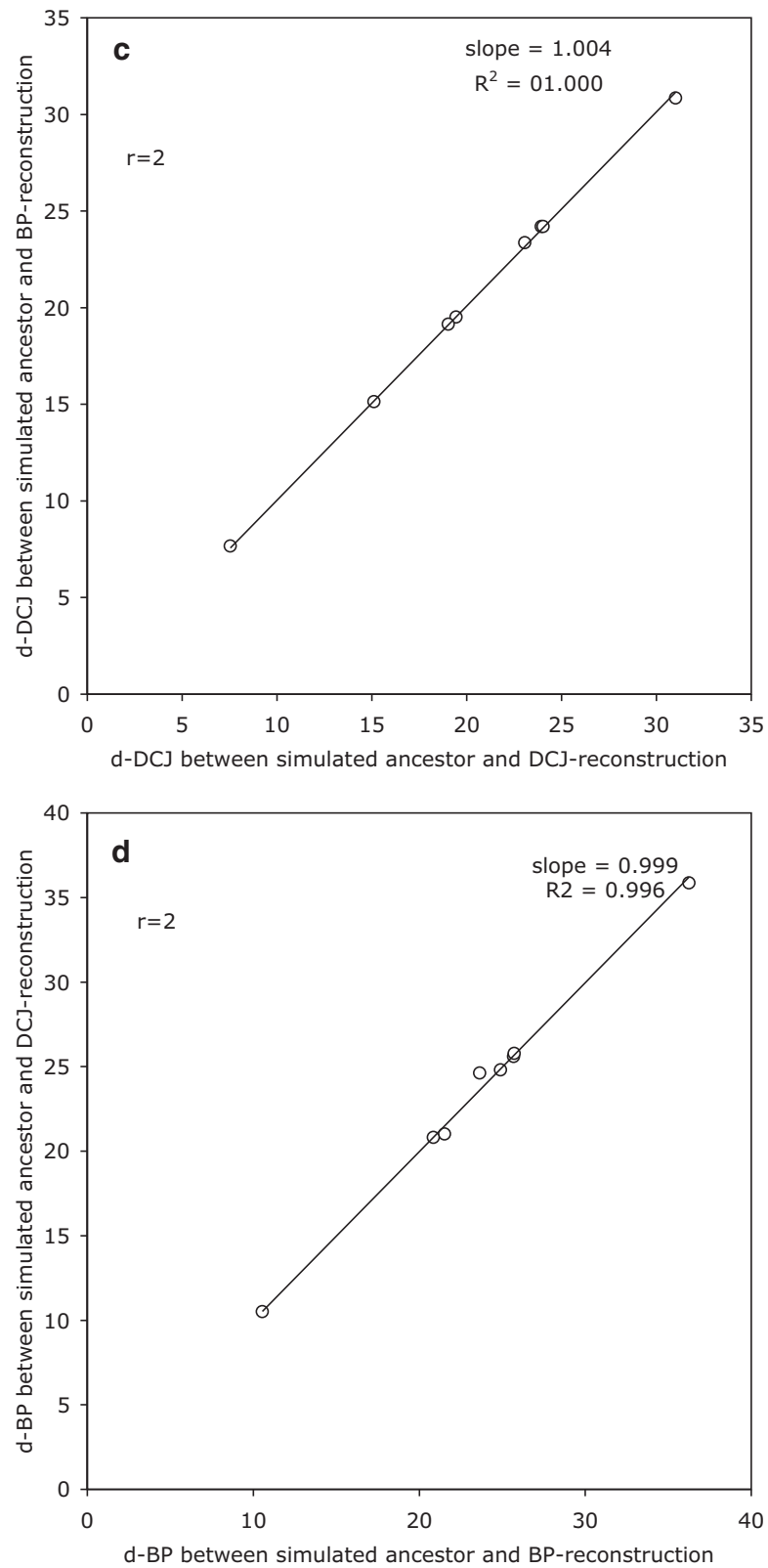
In calculating the intranode distances, no patterns could be discerned for the excess rates within the second data set, the one containing the opossum genome. In particular, there were only three points on each scattergram, making inference very sensitive to statistical fluctuation in any of them. The correlations were very poor, and sometimes even negative. We tracked this latter problem down to two very different DCJ reconstructions of the node closest to opossum, occurring in at least two of the sets of simulations. Generally local minima for these problems tend to be relatively close together and this is what justifies our using summary statistics for them. Exploring the large-scale structure of the set of optimal solutions is beyond the scope of this paper, however, so we confined the study of intranode distances to the seven placentals data set. Here, Table 2 shows that in seven out of eight comparisons, the smaller excess rate (in

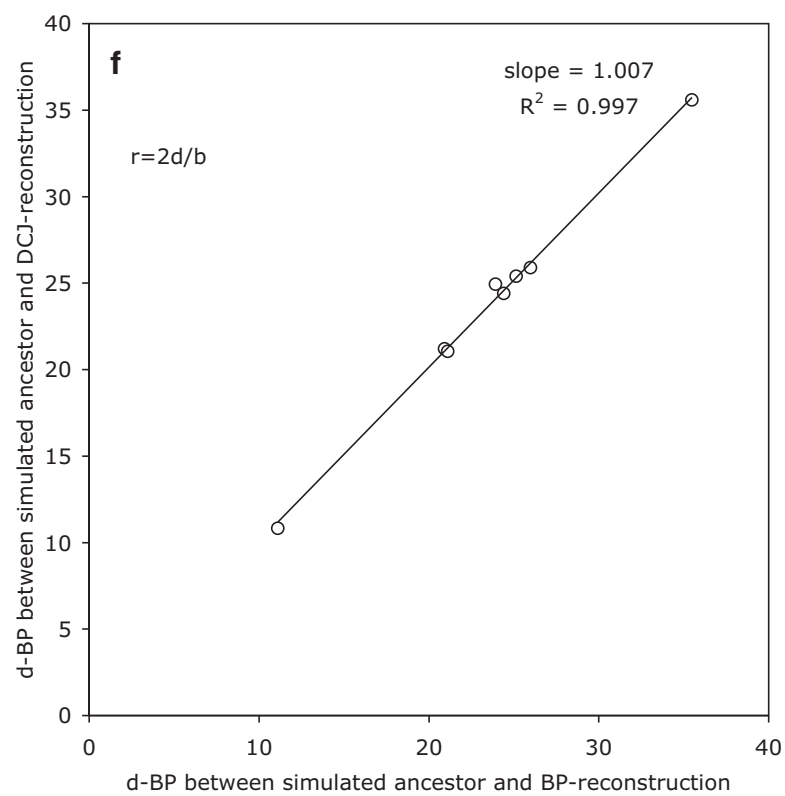
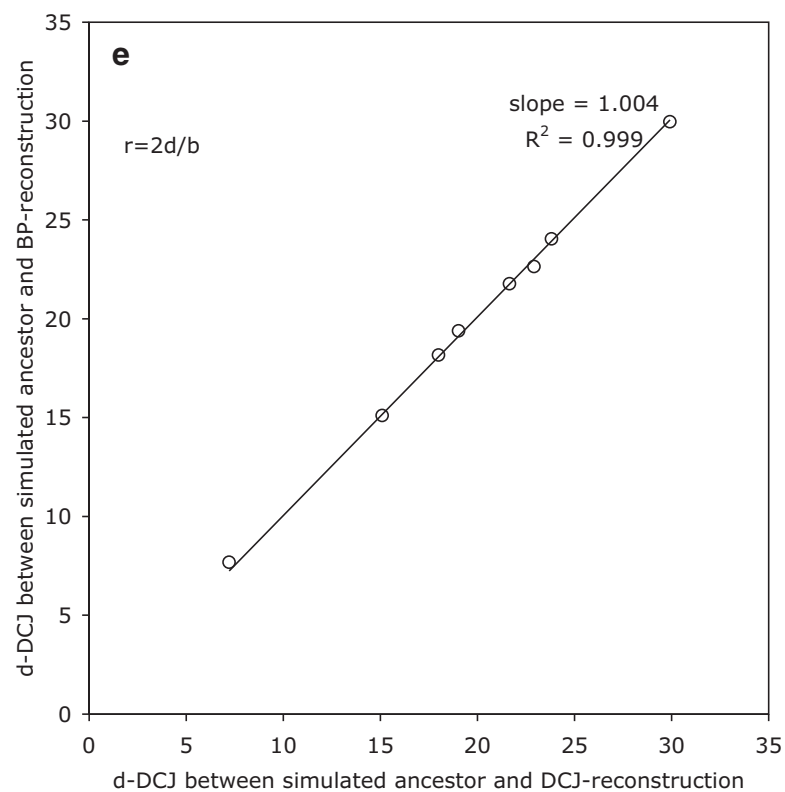
TABLE 3. EXCESS RATE (IN %) FOR EACH RECONSTRUCTION, MEASURED BY COMPETING CRITERION: AVERAGE DISTANCE BETWEEN RECONSTRUCTED AND SIMULATED ANCESTOR RESULTS

Reconstruction	Dataset	
	Combined data	
	BP	DCJ
Simulated		
No re-use	<b>-0.9</b>	1.1
Re-use 2	0.4	<b>-0.1</b>
Actual re-use	<b>0.4</b>	0.7



**FIG. 5.** (a–f) Competing criterion distance between simulated ancestor and reconstruction.

**FIG. 5.** (Continued)

**FIG. 5.** (Continued)

boldface) is that for the DCJ reconstructions. These rates are often negative, showing that the DCJ reconstructions are often more closely clustered in terms of  $d_{BP}$  than the BP reconstructions are.

Figure 4 illustrates the derivation of the excess rates for the real data. Note the smaller correlations compared with the average branch length comparisons.

### 6.3. Average distance between reconstruction and true ancestor

In the case of the simulated data sets, we actually know the ancestral genomes. Thus, as in Section 3.3, we can assess the relative performance of  $d_{BP}$  and  $d_{DCJ}$  with respect to how close the reconstructed genomes are to the true genomes. Table 3 shows the results of combining the results of both data sets to estimate the  $\alpha_{DCJ/BP}$  and the  $\alpha_{BP/DCJ}$ . Figure 5 illustrates the calculations of these values. We note the extremely small excess rates, surely within the noise level when we compare with Tables 1 and 2, so that it appears that the reconstructed genomes are equally close to the true simulated ancestors no matter which method was used to infer them, or used to measure the distance.

## 7. DISCUSSION

It is often taken for granted that rearrangement distance is “better” than breakpoint distance in phylogenetic inference because only the former is connected to a model of genome evolution. This reasoning is specious since (i) the rearrangements inferred in reconstructing the phylogeny are virtually always different and far fewer than those that actually generated the tree, (ii) the set of rearrangement operations in the evolutionary model cannot include all possible mechanisms, (iii) the uniform costs accorded to all operations is a weakness of the rearrangements approach, (iv) breakpoint distance is in fact closely related to evolutionary models, especially those with relatively unconstrained rearrangement operations. In fact, there is no *a priori* biological or statistical reason to prefer one approach over the other. This is what motivates our search for the criterion that does less poorly as judged by the competing criterion.

The general picture that emerges from our analysis is that the ancestral genomes reconstructed according to the breakpoint criterion are more dispersed in the space of genomes than those reconstructed under the rearrangements criterion. This explains the dispersion analyses and by extension the branch-length results, all of which give the impression that the alternative optimal reconstructions under the DCJ criterion are relatively more compact in the set of genomes.

The larger dispersion does not entail that the breakpoint reconstructions are on the average farther from the true simulated ancestor. The explanation most likely lies in the biases of the methods, so that DCJ reconstructions are less dispersed than BP reconstructions, but the biases are of comparable size.

## ACKNOWLEDGMENTS

We would like to thank Richard Durbin for suggesting the role of bias in accounting for the dispersion results. Research was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). D.S. holds the Canada Research Chair in Mathematical Genomics.

## DISCLOSURE STATEMENT

No competing financial interests exists.

## REFERENCES

- Adam, Z., and Sankoff, D. 2008. The ABCs of MGR with DCJ. *Evol. Bioinform. J.* 4, 69–74.
- Bergeron, A., Mixtacki, J., and Stoye, J. 2006. A unifying view of genome rearrangements. *Lect. Notes Comput. Sci.* 4175, 163–173.
- Bourque, G., and Pevzner, P. 2002. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.* 12, 26–36.

- Cosner, M., Jansen, R., Moret, B.M.E., et al. 2001. An empirical comparison of phylogenetic methods on chloroplast gene order data in *Campanulaceae*, 99–121. In Sankoff, D., Nadeau, J., eds. *Comparative Genomics*. Kluwer, Dordrecht.
- Lau, H. 2006. *A Java Library of Graph Algorithms and Optimization*. Chapman, New York.
- Mikkelsen, T.S., Wakefield, M.J., Aken, B., et al. 2007. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* 447, 167–177.
- Moret, B.M.E., Siepel, A.C., Tang, J., et al. 2002. Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. *Lect. Notes Comput. Sci.* 2452, 521–536.
- Moret, B.M.E., Wang, L., Warnow, T., et al. 2001. New approaches for reconstructing phylogenies from gene order data. *Bioinformatics* 17 (Suppl.), 165–173.
- Murphy, W.J., Larkin, D.M., Wind, A.E., et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* 309, 613–617.
- Sankoff, D. 2006. The signal in the genomes. *PLoS Comput. Biol.* 2, e35.
- Swenson, K.M., Marron, M., Earnest-DeYoung, J.V., et al. 2005. Approximating the true evolutionary distance between two genomes. *Proc. 7th Workshop Algorithm Eng. Exp. (ALENEX 2005) SIAM* 121–129.
- Tannier, E., Zheng, C., and Sankoff, D. 2009. Multichromosomal median and halving problems under different genomic distances. *BMC Bioinform.* 10, 120.
- Watterson, G., Ewens, W., Hall, T., et al. 1982. The chromosome inversion problem. *J. Theoret. Biol.* 99, 1–7.
- Yancopoulos, S., Attie, O., and Friedberg, R. 2005. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics* 21, 3340–3346.

Address correspondence to:

Dr. David Sankoff

Department of Mathematics and Statistics

University of Ottawa

585 King Edward Avenue

Ottawa, ON, K1N 6N5, Canada

E-mail: sankoff@uottawa.ca