Rearrangement Phylogeny of Genomes in Contig Form

Adriana Muñoz and David Sankoff

Abstract—There has been a trend in increasing the phylogenetic scope of genome sequencing while decreasing the quality of the published sequence for each genome. With reduced finishing effort, there is an increasing number of genomes being published in contig form. Rearrangement algorithms, including gene order-based phylogenetic tools, require whole genome data on gene order, segment order, or some other marker order. Items whose chromosomal location is unknown cannot be part of the input. The question we address here is, for gene order-based phylogenetic analysis, how can we use rearrangement algorithms to handle genomes available in contig form only? Our suggestion is to use the contigs directly in the rearrangement algorithms as if they were chromosomes, while making a number of corrections, e.g., we correct for the number of extra fusion/fission operations required to make contigs comparable to full assemblies. We model the relationship between contig number and genomic distance, and estimate the parameters of this model using insect genome data. With this model, we use distance matrix methods to reconstruct the phylogeny based on genomic distance and numbers of contigs. We compare this with methods to reconstruct ancestral gene orders using uncorrected contig data.

Index Terms—Contigs, assembly, genome rearrangements, genomic distance, Drosophila, phylogeny.

1 Introduction

While the increasing pace of genome sequencing is adding phylogenetic breadth to the inventory of species available for comparative genomics, the sequencing goal for many of these species is not to produce completely assembled genomes. Instead, the published and archived data remain in draft form, with all or many of the numerous contigs not assigned chromosomal locations, and there are often no resources allocated to further polishing. The price paid for increasing phylogenetic scope in genome sequencing is thus the decreasing sequencing quality for each genome.

While such data may be adequate for many types of comparative genomic studies, they are not directly usable as input to genome rearrangement algorithms. These algorithms require whole genome data, i.e., complete representations of each chromosome in terms of gene order, conserved segment order, or some other marker order, in order to calculate the rearrangement distance d between two genomes. Items, whose chromosomal location is unknown, cannot be part of the input.

The present paper deals with gene order-based phylogeny. The question we ask here: Is there any way to use genome rearrangement algorithms to compare genomes available in draft form only? One elegant answer was provided by Gaul and Blanchette [9] for the comparison of two genomes. Other approaches have also been investigated [4]. These methods construct a number of intermediate structures before or during the actual comparison of the genomes. Since we will be using distance matrix methods for phylogenetic analysis, these intermediate structures are largely irrelevant; we need distances and not the detailed reconstruction of the structures used in calculating the distance. For these purposes, involving more than two genomes, our suggestion is to use the contigs directly in the rearrangement algorithms as if they were chromosomes. This introduces a number of biases, such as increasing the distance to accommodate the count of extra fusion/fission operations necessary to compare genomes with different numbers of chromosomes. This bias and other problems with rearrangement distances in general, and with contig-based distances in particular, must be corrected during the construction of a distance matrix to input into a phylogenetic analysis.

We apply our methods to data originating mostly in the 12-genome *Drosophila* project [7]. We compare 10 *Drosophila* genomes with two other dipteran genomes and two outgroup insect genomes. We discuss these data in Section 2.

In Section 3, we model the behavior of the genomic distance as a function of evolutionary time, and discuss how to invert this function in order to infer elapsed time. In Section 4, we study the case where one of the two genomes being compared is fully assembled and the other is in contig form. Simulations are used to understand the consequences on evolutionary time inference of using incomplete assemblies. The ideas developed there are then extended to the more complex case where both genomes are fragmented into contigs, in Section 5. We can then construct a matrix of corrected evolutionary divergence times between all pairs of genomes in the database and carry out a phylogenetic analysis of the 14 genomes, in Section 6. For the *Drosophila* data only, we compare this phylogeny to one generated by gene order reconstruction algorithms applied to the genomes without any corrections for contigs or distance-time nonlinearities.

- *A. Muñoz is with the Department of Computer Science, School of Information Technology and Engineering, University of Ottawa, Ottawa K1N 6N5, Canada. E-mail: amuno010@uottawa.ca.*
- *D. Sankoff is with the Department of Mathematics and Statistics, University of Ottawa, Ottawa K1N 6N5, Canada. E-mail: sankoff@uottawa.ca.*

Manuscript received 18 July 2009; revised 27 Oct. 2009; accepted 28 Oct. 2009; published online 6 Aug. 2010.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBBSI-2009-07-0124.
Digital Object Identifier no. 10.1109/TCBB.2010.66.

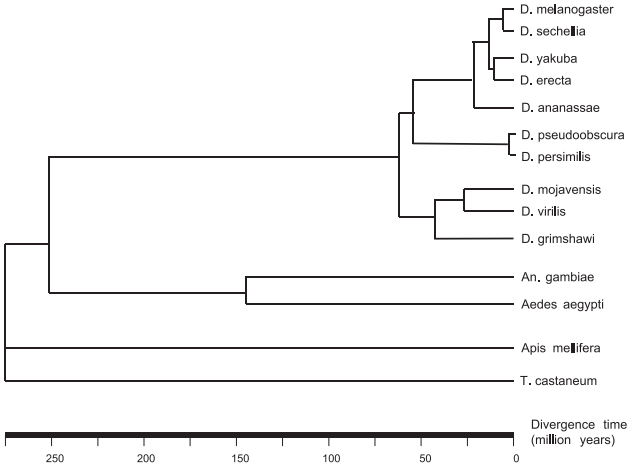


Fig. 1. Phylogeny of Drosophila and outgroups abstracted from the literature, with divergence times.

TABLE 1
Number of Contigs Constructed for Each Genome

species (abbreviation)	genes	contigs
D. melanogaster (Dmel)	8867	6
D. sechellia (Dsec)	8851	66
D. yakuba (Dyak)	8809	30
D. erecta (Dere)	8866	9
D. ananassae (Dana)	8844	40
D. pseudoobscura (Dpse)	8778	51
D. persimilis (Dper)	8779	87
D. virilis (Dvir)	8855	32
D. mojavensis (Dmoja)	8853	14
D. grimshawi (Dgri)	8801	35
Anopheles gambiae (Anoph)	6168	6
Aedes aegypti (Aedes)	6318	869
Apis mellifera (Apis)	4898	702
Tribolium castaneum (Trib)	5647	89

2 The Data

One of the difficulties in using gene order rearrangement algorithms is the lack of curated gene order databases for the higher eukaryotes with sequenced genomes. Because the gene identification and homology identification have already been done in [7], we use a carefully constructed inventory of neighboring gene pairs (NGPs) in 10 *Drosophila* species and four outgroup insects, rather than raw contig data. A.J. Bhuktar provided us with a file listing all NGPs and the genomes in which they appear. By the time of writing, the assembly of these genomes has progressed, but for our purposes, i.e., to show how to handle genomes in contig form, the original data set is preferred.

We abstracted best-judgment divergence times among the genomes from a number of somewhat contradictory publications [10], [12], [13] available when we carried out our research, as summarized in Fig. 1. For *Drosophila*, this turns out to be the same as the tree now widely accepted (see [5], [1]), except that the time scale is now thought to be 40 percent shorter than in Fig. 1. This is of little pertinence to us, since our calculations are all in terms of numbers of rearrangements, not absolute time, which, as a first approximation, is merely assumed to be proportional to the rearrangement numbers. The position of the mosquitos *Anopheles gambiae* and *Aedes aegypti*, relative to *Drosophila*, is uncontroversial within the Diptera, while the uncertainty of the branching order of the Hymenoptera represented by *Apis mellifera* and Coleoptera represented by *Tribolium castaneum* leads us to postulate a trichotomy giving rise to these three orders of holometabolous insects.

Bhutkar et al. [3], [5] have already used the NGP data for a phylogenetic analysis of *Drosophila*, inferring phylogenies, rearrangements and synteny blocks, but our use of the NGPs here is different. It is simply to reconstruct the gene orders in the contigs; we wish to create a data set for testing our method for gene order-based phylogenetics from genomes in contig form.

For each genome, we constructed contigs by amalgamating overlapping NGPs. Whenever we arrived at a gene in only one NGP in a genome, this terminated a contig. Our reconstruction then does not necessarily correspond completely to the original contigs in the 12-genome *Drosophila* sequencing project [7], but this has little importance for our work—how the genomes are fragmented into contigs, and into how many, is a methodological question that depends on laboratory resources and techniques and has nothing directly to do with how the genome has evolved. (Both contig ends and rearrangement breakpoints may be enriched for duplicated sequence, but this indirect connection has no consequence for the problem we are attacking.)

Table 1 gives the number of contigs reconstructed for each genome. Note that the reconstructions of *D. melanogaster*, *D. erecta*, and *An. gambiae* reflect the complete, or almost complete assembly of these genomes.

3 Genomic Distance and Evolutionary Time

3.1 The Operations

We assume familiarity with the classical genetics notions of inversion, transposition, and reciprocal translocation of chromosome segments, as well as chromosomal fission and fusion. These are formalized in such papers as those by Tesler [14], Yancopoulos et al. [16], and Bergeron et al. [2]. Briefly, representing a chromosome as a string of genes $h_1 \cdots h_l$, where a pair of successive genes $h_u h_{u+1}$ is termed an *adjacency*, we can illustrate:

- an inversion (implying change of sign, i.e., change of strand) of a chromosomal segment:

$$h_1 \cdots h_u \cdots h_v \cdots h_m \rightarrow h_1 \cdots -h_v \cdots -h_u \cdots h_m,$$

  disrupting the two adjacencies $h_{u-1}h_u$ and $h_v h_{v+1}$,
- a transposition of a chromosomal segment:

$$h_1 \cdots h_u \cdots h_v \cdots h_w \cdots h_m \rightarrow h_1 \cdots h_{u-1} h_v \cdots h_w h_u \cdots h_{v-1} h_{w+1} \cdots h_m,$$

  disrupting the three adjacencies $h_{u-1}h_u$, $h_{v-1}h_v$ and $h_w h_{w+1}$,
- a reciprocal translocation between two chromosomes:

$$h_1 \cdots h_u \cdots h_l, k_1 \cdots k_v \cdots k_m \rightarrow h_1 \cdots h_{u-1} k_v \cdots k_m, k_1 \cdots k_{v-1} h_u \cdots h_l,$$

  disrupting the two adjacencies $h_{u-1}h_u$ and $k_{v-1}k_v$,

- a chromosome fission:

$$h_1 \cdots h_v \cdots h_l \rightarrow h_1 \cdots h_v, h_{v+1} \cdots h_l,$$

  disrupting the adjacency $h_v h_{v+1}$, and
- the fusion of two chromosomes:

$$h_1 \cdots h_l, k_1 \cdots k_m \rightarrow h_1 \cdots h_l k_1 \cdots k_m.$$

The genomic distance is the minimum number of operations of these types (or some specified subset of types) required to transform one of the genomes being compared into the other. The authors mentioned above also provide rapid algorithms for deriving the distance, given genomes composed of ordered chromosomes represented by the same n genes, markers or segments in the two genomes, assuming the strandedness, or reading direction of each gene is known.

3.2 The Relation between True and Inferred Distance

Even assuming that rearrangements occur at a relatively constant rate over time and are randomly positioned in the genomes, we have no simple, exact probability relationship between the actual number τ of rearrangements after a certain time t has elapsed and the number of rearrangements d inferred by applying the genomic distance algorithms to compare the initial and the derived genomes [6], [8], [15]. We can, however, model the proportion of adjacencies that will be disrupted versus the proportion that will remain intact after τ random rearrangements. For each of the adjacencies in the original genome, the probability that it will remain intact after τ rearrangements is $(1-\lambda/n)^{\tau}$ or approximately $e^{-\lambda\tau/n}$, where λ depends on the proportions of the various kinds of rearrangements in the model. Thus the number of disrupted adjacencies will be approximately $n(1-e^{-\lambda\tau/n})$.

Now, we can expect at the τth step that the increase in d will also be closely connected to the proportion of the adjacencies between genes that have not been created, i.e., have never been disrupted, by the previous $\tau-1$ rearrangements—if the τth rearrangement only disrupts adjacencies created in previous steps, it is quite likely that the inference algorithm will suggest an optimal evolutionary history requiring no more rearrangements than were required after the $(\tau-1)$th step. Then, though we do not know the precise probability law of d, we can hypothesize as a first approximation

$$\mathrm{E}(d) \approx n(1-e^{-\lambda\tau/n}), \tag{1}$$

where n is the number of ordered genes or markers in both genomes, and λ in this case is a constant close to 1, since we know that $d \approx \tau$ for small τ and that $d/n \rightarrow 1$, as $\tau \rightarrow \infty$. Then if we knew λ, we could estimate τ using

$$\hat{\tau} = -\frac{n}{\lambda}\log\left(1-\frac{d}{n}\right). \tag{2}$$

In fact, the relationship between the actual and inferred numbers of rearrangements (not shown) deviates considerably from the one-parameter model in (1), both for small and large τ. Combinatorial effects result in $\mathrm{E}(d) < \tau$ even for rather small values of τ. And the approach to the asymptote $\frac{\mathrm{E}(d)}{n} \nearrow 1$ is faster than (1) would suggest. We thus have recourse to a two-parameter model by adding a quadratic correction to the linear term in the exponent, so that the model becomes

$$\mathrm{E}(d) \approx n(1-e^{-\lambda_1\tau/n-\lambda_2(\tau/n)^2}), \tag{3}$$

in which case the estimate of τ becomes

$$\hat{\tau} = \frac{n}{2\lambda_2}\left(-\lambda_1 + \sqrt{\lambda_1^2 - 4\lambda_2\log\left(1-\frac{d}{n}\right)}\right). \tag{4}$$

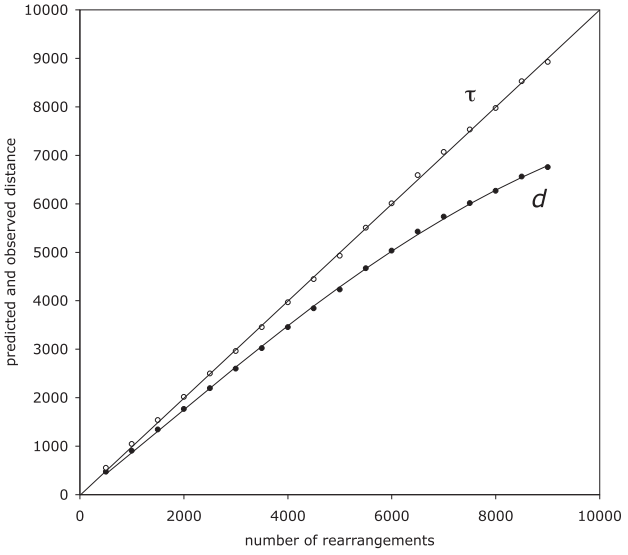


Fig. 2. Comparison of model and simulations for genomic distance d and of true value and estimated τ. Predicted (curve) and observed (dots) values of d, and inferred (open dots) values of $\hat{\tau}$ versus true (diagonal line) values. (See Sections 3.2 and 3.3.)

3.3 Simulation-Based Estimates

To estimate the parameters λ_1 and λ_2, we simulate pairs of genomes with $n = 8{,}867$, the maximum number of genes used in our *Drosophila melanogaster* comparisons, and τ up to 9,000 random rearrangements to derive one genome from the other. It is well known (e.g., [5]) that rearrangements in *Drosophila* are almost exclusively inversions within Muller elements (chromosome arms), although this does not pertain to other insects. We carry out our simulations accordingly, with about 99.8 percent inversions, and only a few reciprocal translocations. We use a DCJ (double-cut-and-join) algorithm [16], [2] to calculate d from the genomes. This is repeated 100 times, and d averaged, to estimate $\mathrm{E}(d)$.

Fig. 2 shows the relationship between τ and both $\mathrm{E}(d)$ and $\hat{\tau}$, using the values $\lambda_1 = 0.846$ and $\lambda_2 = 0.576$, found by a least sum of squares criterion applied to the set of τ and $\hat{\tau}$ values. The way τ and d are normalized means that the parameters should not be very sensitive to n, though we do not study this here, since all the experimental genomes are of comparable size.

Note that if τ is very small, our model predicts $\mathrm{E}(d) \approx \lambda_1\tau$ instead of $\mathrm{E}(d) \approx \tau$, though this bias occurs at a scale not visible in Fig. 2.

3.4 Comparison with Previous Work

This analysis resembles the "empirical" approach in [15] to the relationship between d and τ, which also makes use of two parameters, except that our starting point is the intuitive development leading to (1) at the beginning of this section, whereas [15] takes a purely curve-fitting approach from the outset. These authors propose that

$$\mathrm{E}(d) \approx n \min\left(\frac{\tau}{n}, \frac{\frac{\tau}{n}^2 + b\frac{\tau}{n}}{\frac{\tau}{n}^2 + c\frac{\tau}{n} + b}\right). \quad (5)$$

They find the parameter values $b = 0.6$ and $c = 0.46$ fit simulated data well when n is in the region of a few dozen, justifying the use of the normalized variable $\frac{\tau}{n}$. When n is hundred times greater, however, neither these nor any other parameter values allow the model in (5) to fit the simulated values of d, without the introduction of additional parameters.

4 The Effect of Genome Fragmentation

4.1 Contigs as Chromosomes

Consider one completely assembled genome B, and another A, in contig form only. The basic idea is that if we treat each contig as a chromosome, a rearrangement algorithm will automatically carry out a number of "fusions" to assemble the χ_A contigs in A into a small number of inferred chromosomes equal to the number χ_B in B, in calculating d. At the same time, it will find other rearrangements, but we know that the number of fusions required will be at least the difference between the number of contigs in A and the number of chromosomes in B. Indeed, in almost all optimal rearrangement scenarios there will not be both fusions *and* chromosome fissions; thus, when we use a rearrangement algorithm to compare a genome A in contig form with an assembled genome B, obtaining a preliminary distance d', it may seem appropriate to correct this to

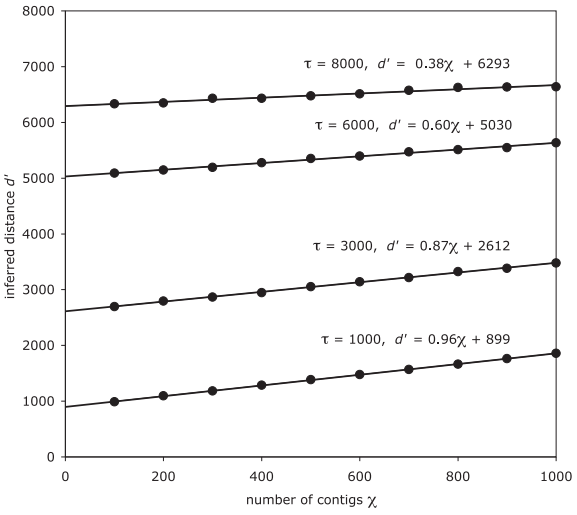
$$d = d' - (\chi_A - \chi_B), \quad (6)$$

where we assume $\chi_A > \chi_B$.

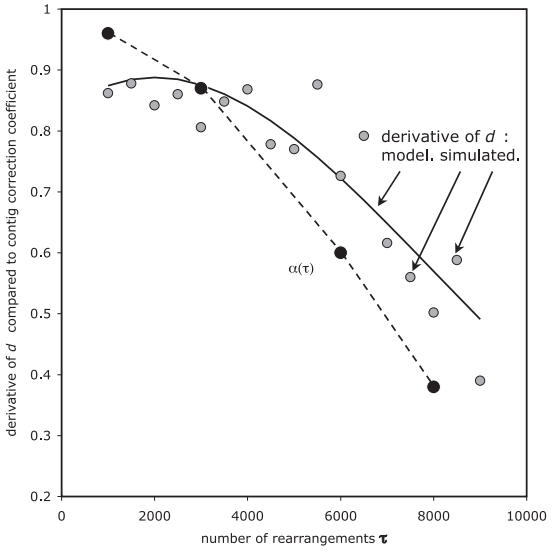
This argument holds independent of the other details of the optimal rearrangement scenario, for which there may be many for a particular data set.

4.2 Contig Fusion and $\frac{d\mathrm{E}(d)}{dt}$

We cannot simply substitute correction (6) into (2) or (4) to estimate τ. Even if the number of contig fusions is exactly $\chi_A - \chi_B$, we know that these fusions are done by the inference algorithm in such a way as to minimize the total d', including inversions and other operations. To take account of this, we should only remove a proportion α of $C = \chi_A - \chi_B$ from d'. How large a proportion? **The natural hypothesis is that the effect of the C fragmentation operations, which have exactly the same properties as chromosome fission operations, and which create C extra contigs, should have the same effect as the addition of C of any other types of rearrangement to the genome.** In any model, such as (3), where we relate d and τ as if they were real variables (i.e., not just integer variables), the expected rate of increase of d with τ is the derivative $\frac{d\mathrm{E}(d)}{d\tau}$. Thus, increasing τ by C should increase d by approximately $\frac{d\mathrm{E}(d)}{d\tau}C$.



(a)



(b)

Fig. 3. Effect of genome fragmentation on genomic distance. For genomes generated by $\tau = 1{,}000, 3{,}000, 6{,}000$, or $8{,}000$ rearrangements, broken into $\chi = 100, 200, \ldots, 1{,}000$ contigs: (a) The relationship between uncorrected genomic distance d' and χ, with equations of trend lines. (b) The parameter α as a function of τ (large dots and dotted line), with values taken from the coefficient of χ in the trend lines in the top part of the figure. Solid line represents the derivative of $\mathrm{E}(d)$ in (3) and Fig. 2, while the shaded dots represent the simulated values of the derivative, calculated from the data presented in Fig. 2. Calculated and simulated derivatives for small values of τ biased downwards, as discussed in Section 3.3.

To verify this hypothesis, we undertook a series of simulations, starting from an initial genome B containing 8,867 genes in $\chi_B = 6$ chromosomes, generating 100 rearranged genomes, each through τ random rearrangements applied to B to produce a new genome, and each then fragmented into χ_A contigs. This was repeated for a range of values of τ and χ_A.

The average results for d' are summarized in Fig. 3a. First the linearity of the response to increasing χ_A is clear, at least

in the range studied $\chi_A < 1{,}000$, indicating that (6) should be replaced by

$$d = d' - \alpha(\tau)(\chi_A - \chi_B), \tag{7}$$

where $\alpha(\tau)$ is a decreasing function of the number of rearrangements τ. As can be seen in Fig. 3b, this decrease approximately parallels the theoretical derivative of d, but is somewhat steeper. For purposes of interpolation, we fit $\alpha(\tau)$ with a quadratic function $1 - 0.0276(\tau/1{,}000) - 0.0063(\tau/1{,}000)^2$ by minimizing the sum of squares over the four values of τ, where we have calculated α.

Given d', then, we can solve (3) and (7) simultaneously to find τ and d, since $n, \lambda_1, \lambda_2, \chi_A$, and χ_B are known, as is the dependence of α on τ. In practice, this can be done by successive iteration of (4) and (7), which converges rapidly, initializing with, for example, $\tau_0 = d'$.

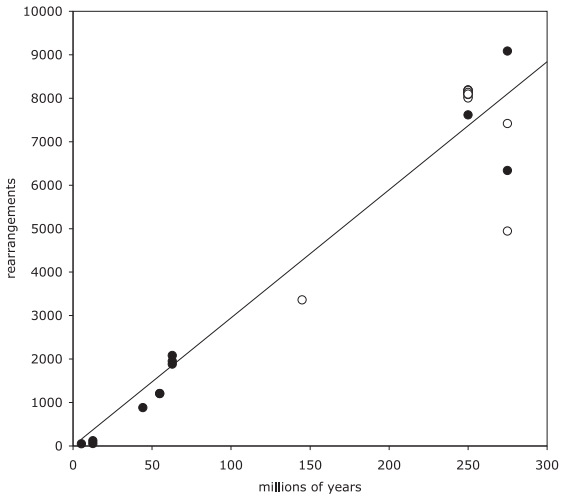
Applying this to the comparison of the completely assembled *D. melanogaster* genome with each of the other 13 genomes, and to the comparison of the completely assembled *Anopheles gambiae* genome with each of the other 13 genomes, gives the results in Fig. 4a. The high degree of scatter at higher divergence times reflects both the uncertainty of the divergence dates and the inhomogeneity of rearrangement rates, both between the fruit-fly and mosquito families within the dipteran order and among the three orders in the class *Insecta* represented in these data.

5 The Case of Both Genomes in Contig Form

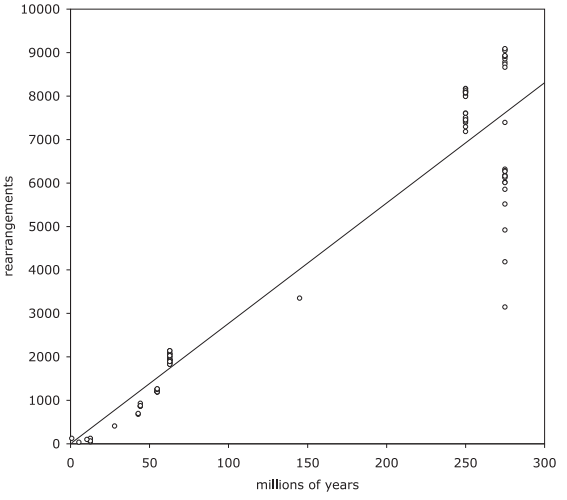
When we compare two incompletely assembled genomes A and B, we may still wish to remove some quantity depending on χ_A and χ_B from d' to account for the fusions (and/or fissions), but this is not as easy to analyze, for two reasons. One is that we are not comparing a fragmented genome to a complete genome, so we can no longer consider this correction as a way of using the assembled genome as a guide for reconstructing the fragmented genome, simultaneous with the distance calculation. The second problem is that there is no obvious way, within the formula, of combining (adding, multiplying, ...) the number of contigs in one genome with the number in the other. This reflects the lack of intuition on how the contigs increase the distance (because of artificial fusions and fissions) on one hand, and how they decrease it (by multiplying the number of economical but false rearrangements) on the other hand. These reasons lessen the intuitive appeal of the kind of correction we used in the previous section. Nevertheless, we can try to find an appropriate correction using the same simulation approach as in the previous sections.

We simulated 50 runs each of two genomes of size $n = 8{,}867$ separated by $\tau = 1{,}000, 3{,}000, 6{,}000$, and $8{,}000$ random rearrangements as before, but with both genomes independently and randomly fragmented into $\chi = 100, 200, 400, 600$, or 800 contigs, i.e., $5 + \binom{5}{2} = 15$ pairs of contig configurations for each degree of rearrangement. We applied the DCJ algorithm and calculated the mean d' for each configuration. The results are summarized in Fig. 5.

We observe in Fig. 5a that for fixed τ and χ_A, the response of d' to increasing χ_B is systematically linear. This is clear up to $\tau = 6{,}000$ and only starts to break down for $\tau = 8{,}000$ and $\chi_A \geq 600$, where examination of the data on an expanded scale shows that d' actually



(a)



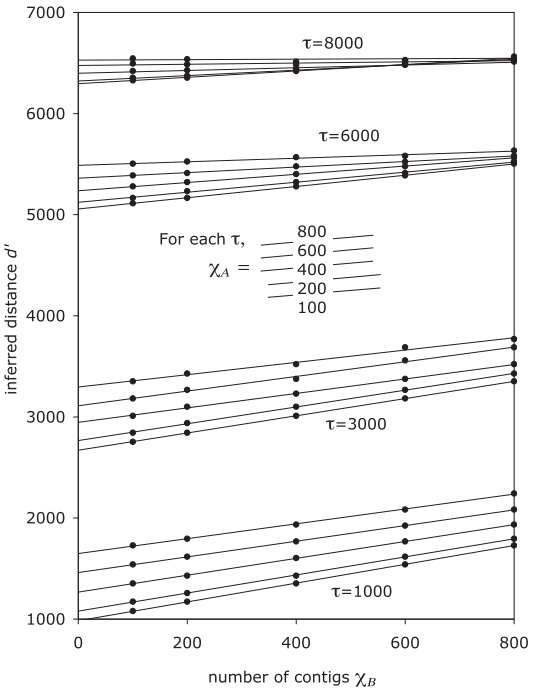
(b)

Fig. 4. (a) Divergence, in total number of genome rearrangements, estimated from genomic distances through (3) and (7), compared to divergence times abstracted from the literature. Estimates of τ for *D. melanogaster* compared with 13 other genomes (solid dots). Estimates for *Anopheles gambiae* with 13 other genomes (open dots). Line represents least squares fit to all points. (b) Pairwise comparison of all pairs of 14 genomes, as discussed in Section 5. Line represents least squares fit.

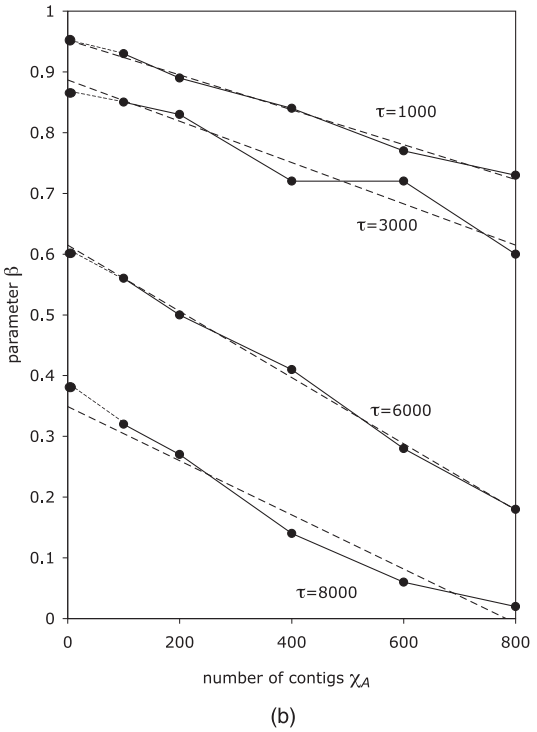
decreases somewhat initially, then increases, as χ_B increases (not discernible in Fig. 5). The linear rate of increase of d', plotted as $\beta(\tau, \chi_A)$ in Fig. 5b, is the same as the $\alpha(\tau)$ in Fig. 3 for small values of χ_A. In fact, d' shows the same linear increase as a function of $\chi_A + \chi_B$ up to moderate values of this sum, as in Fig. 6, depending on τ, after which the rate of increase drops off markedly. As both genomes are increasingly fragmented, the mutational process becomes saturated so that inference of distance can only be very approximate.

Nevertheless, as with the case of only one genome fragmented into contigs studied in Section 4, we can infer d and τ from observed values of d' by solving (4) simultaneously with

$$d = d' - \alpha(\tau)\chi_A - \beta(\tau, \chi_A)\chi_B, \tag{8}$$



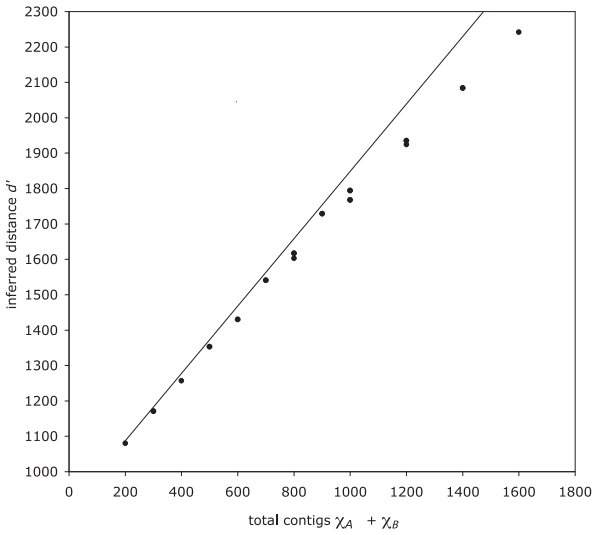
(a)



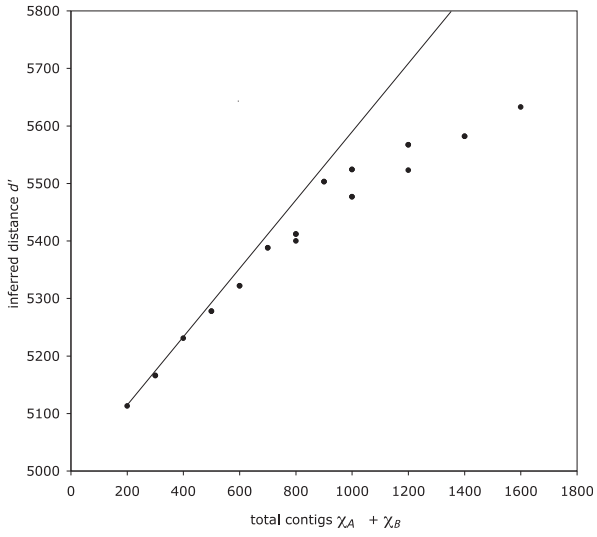
(b)

Fig. 5. Effect of fragmentation of both genomes on genomic distance. For genomes A and B separated by $\tau = 1{,}000, 3{,}000, 6{,}000,$ or $8{,}000$ random rearrangements, broken into $\chi = 100, 200, 400,$ or 800 contigs: (a) The relationship between uncorrected genomic distance d', χ_A, and χ_B, with trend lines for each χ_A connecting the values of d' for a range of χ_B. (b) The coefficient β of the linear dependence of d' on χ_B in the top diagram, as a function of χ_A. Dotted segments connect $\beta(\tau, 100)$ to $\beta(\tau, 0) = \alpha(\tau)$ from Fig. 3. Dashed line is the linear trend line, not taking account of $\beta(\tau, 0)$.

where $\beta(\tau, \chi_A) = \alpha(\tau) - (0.00027 - 0.00003\tau)\chi_A$, and where the coefficient of χ_A is estimated by a least squares fit to the slopes of the four trend lines in Fig. 5b.



(a)



(b)

Fig. 6. Dependence of d' on the total number of contigs in the two genomes, for (a) $\tau = 1{,}000$ and (b) $\tau = 6{,}000$. Straight lines represent $d' = d + \alpha(\tau)(\chi_A + \chi_B)$ where $\alpha(\tau)$ is as in Fig. 3.

Plotting the values of τ inferred from (8) against values extracted from the literature produced the results in Fig. 4b. Note that (8) is asymmetric with respect to A and B, which could have consequences for the analysis in the next section. However, this is avoided if A always denotes the more fragmented of the two genomes.

6 Phylogeny

6.1 Using a Distance Matrix

If we input the inferred pairwise values of τ into a neighbor-joining routine, we produce the phylogeny in Fig. 7. When this is compared to Fig. 1, the only structural difference is at one node where we see *D. sechellia* branching off just before *D. melanogaster* rather than branching off together as sister groups. More striking is the long branch leading to the *Drosophila* group, suggesting a rapid rate of evolution at the moment of divergence from

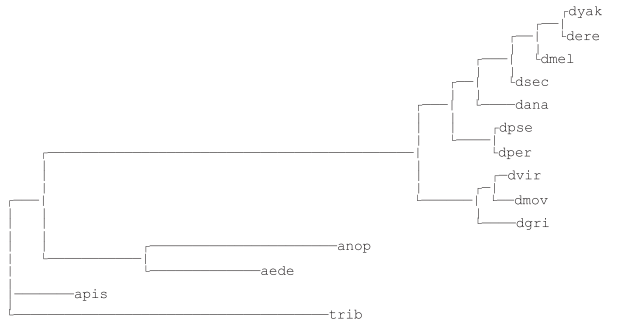


Fig. 7. Neighbor-joining phylogeny based on a matrix of inferred number of rearrangements τ.

other *Diptera*. Note that using the uncorrected matrix of d' as input to neighbor joining does not show this rate effect as clearly as τ, and also introduces other structural errors into the phylogeny.

6.2 Phylogeny with Reconstruction of Ancestral Genomes

While the use of neighbor-joining on a distance matrix is convenient and rapid, it does not infer anything about the ancestral genomes in the resulting phylogeny. Tools are available, however, for inferring these ancestral genomes, given the topology of the tree.

As schematized in Fig. 8, we solved the gene-order median problem at each node of the ancestral tree, using the path groups approach [17], iterating over all ancestral nodes of the tree until convergence. We did this separately for the trees in Figs. 1 and 7, using genomes containing the

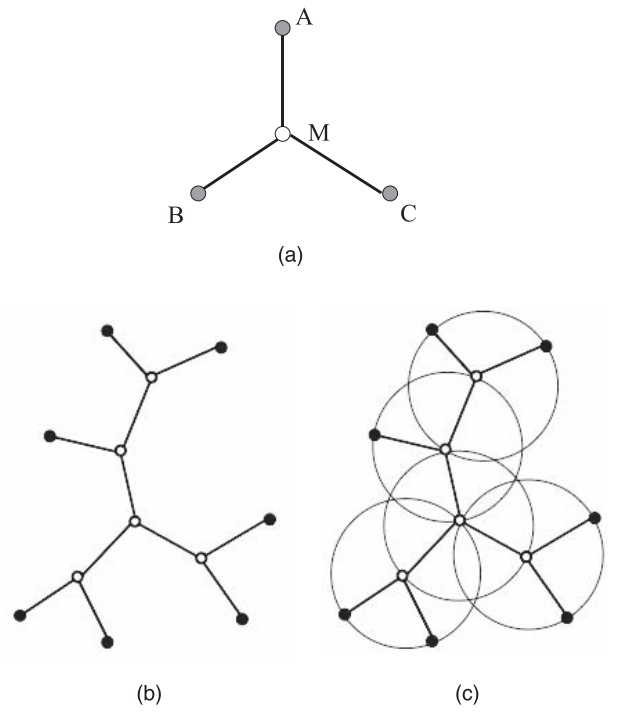


Fig. 8. (a) Median problem: Given genomes A, B, C, find M such that $d(A,M) + d(B,M) + d(C,M)$ is minimized. (b) Example of unrooted phylogeny with given present-day genomes at terminal nodes (dark dots) and genomes to be inferred at the ancestral nodes (white dots). (c) Inference of genomes at ancestral nodes found by iterating through the ancestral vertices, solving a median problem at each step.

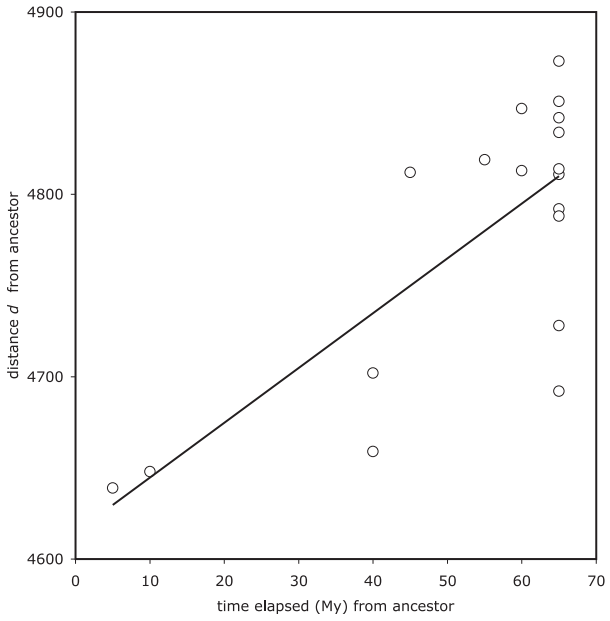


Fig. 9. Rearrangement distance between Bhutkar et al. *Drosophila* ancestor [3] and 10 data (vertically aligned on right edge of graph) and eight reconstructed ancestral genomes. Least squares line shown.

8,500 genes in common among all the *Drosophila* species, ignoring the outgroups for this analysis. This method attempts to minimize the sum of d over all branches. The path groups median is constructed by greedily building three "breakpoint graphs" simultaneously, relating the genome under construction M to each of A, B, and C in Fig. 8a , and attempting to maximize the total number of cycles over the three separate graphs. In the last iterations, the construction employs a look-ahead routine to increase accuracy, at the cost of increasing individual median calculations from a minute or so with 8,560 genes, to an hour or two depending on how different the three genomes are.

The *Drosophila* part of the tree in Fig. 7 proved to be substantially less costly in terms of rearrangement operations (3,862 rearrangements versus 3,872 for the tree in Fig. 1), showing that the incorrect early branching of *D. sechellia* in Fig. 7 is inherent in the gene-order data, and is not an artifact of the distance correction.

The rearrangement analysis was carried out on the raw data, namely the genomes assembled or partially assembled from the NGP data. The only preprocessing was to discard the small number (200-300 per genome) genes not in common to all genomes. The median program contains a bias towards fusions in the direction of the median, though this does not bias the total cost of the median analysis. Thus the ancestral genomes contain relatively few contigs.

How can we validate the rearrangement analysis? If we knew the gene order of the ancestral *Drosophila*, we could compare the reconstructed ones with it. Data provided in Additional file 7 in [3] include the 1,000+ syntenic blocks that Bhutkar et al. were able to reconstruct using conservative criteria based on 12 *Drosophila* genomes. Treating these blocks as contigs of the ancestral genome "A", we calculated the rearrangement distance between A and each of our 10 data and eight ancestral genomes (reconstructed according to Fig. 1), after removing from our genomes the several thousand genes absent in A. The distance in Fig. 9 is

plotted against elapsed time from A as in Fig. 1. Despite the large amount of noise in the analysis, largely due to large number of contigs in A, it is clear that our reconstructed ancestors tend to be closer to A than are the data genomes. We may conclude that the rearrangement phylogeny is reconstructing aspects of ancestral gene order that are not apparent in all the individual data genomes.

7 Conclusion

We have developed a principled approach to correcting genome rearrangement distance when comparing genomes in contig form. Features of this include:

- A model for the τ—d relationship motivated by an intuitive negative exponential connection between asymptotic genomic distance ($\mathrm{E}(d) \nearrow n$) and adjacency retention, including an empirically motivated quadratic correction term to improve the fit to simulated values.
- A reasoned procedure for subtracting artificial fusions and fissions due to the fragmentation of one or both of the genomes into contigs.
- The discovery and quantitative characterization of the linear relation between the uncorrected distance and the number of contigs, when only one or both of the genomes are fragmented into contigs. These linearities hold for a wide range of τ, up to 6,000 for genomes of size around $n = 9{,}000$, and up to $\chi = 1{,}000$ contigs, though saturation is more of a problem when both genomes are highly fragmented.
- Improved phylogenetic reconstruction for a data set on 14 insect genomes. We recovered a tree that accurately reflects almost all the phylogenetic information extracted from the literature, and pinpointed a period of evolutionary acceleration on one lineage.
- A validation approach based on the reconstruction of uncorrected ancestral genomes using gene-order medians biased towards contig fusion operations.

As argued in Section 3, the values of the parameters λ_1 and λ_2 are not likely to be very sensitive to n, especially for n in the thousands, since the model relates the normalized variables τ/n and d/n. Nor should they depend on details of the rearrangement model such as the number of chromosomes or the proportions of different types of rearrangement, assuming the latter are naturally weighted as in the double-cut-and-join framework. Thus, the values we have estimated through simulation for λ_1, λ_2, and α (considered as a function of τ/n), should hold for a range of data sets. This stability reassures us that our methods should be widely applicable beyond the *Drosophila* data we have used, but only partly mitigates the main shortcoming of this and other models such as in [15], namely that they are not analytically derived. Thus, the mathematical foundation of probability models and statistical analyses of genomic problems like the one addressed here, would benefit more from advances like those in [8] than by further characterization of empirical models such (3).

Acknowledgments

The authors thank Arjun Bhutkar for providing the NGP files with pair occurrence tabulated by species. The authors also thank Chunfang Zheng for guidance in using her distance and median programs and rearrangement simulations. This research project was funded in part by the Natural Sciences and Engineering Research Council of Canada.

References

[1] "Assembly/Alignment/Annotation of 12 Related Drosophila Species," http://rana.lbl.gov/drosophila/, 2010.
[2] A. Bergeron, J. Mixtacki, and J. Stoye, "A Unifying View of Genome Rearrangements," *Algorithms in Bioinformatics,* P. BüCher and B.M.E. Moret, eds., pp. 163-173, Springer, 2006.
[3] A. Bhutkar, W.M. Gelbart, and T.F. Smith, "Inferring Genome-Scale Rearrangement Phylogeny and Ancestral Gene Order: A *Drosophila* Case Study," *Genome Biology,* vol. 8, p. R236.1-15, 2007.
[4] A. Bhutkar, S. Russo, T.F. Smith, and W.M. Gelbart, "Techniques for Multi-Genome Synteny Analysis to Overcome Assembly Limitations," *Genome Informatics,* vol. 17, pp. 152-161, 2006.
[5] A. Bhutkar, S.W. Schaeffer, S.M. Russo, M. Xu, T.F. Smith, and W.M. Gelbart, "Chromosomal Rearrangement Inferred from Comparisons of 12 *Drosophila* Genomes," *Genetics,* vol. 179, pp. 1657-1680, 2008.
[6] D. Dalevi and N. Eriksen, "Expected Gene-Order Distances and Model Selection in Bacteria," *Bioinformatics,* vol. 24, pp. 1332-1338, 2008.
[7] Drosophila 12 Genomes Consortium, A.G. Clark et al., "Evolution of Genes and Genomes on the *Drosophila* Phylogeny," *Nature,* vol. 450, pp. 203-218, 2007.
[8] N. Eriksen and A. Hultman, "Estimating the Expected Reversal Distance after a Fixed Number of Reversals," *Advances in Applied Math.,* vol. 32, pp. 439-453, 2004.
[9] E. Gaul and M. Blanchette, "Ordering Partially Assembled Genomes Using Gene Arrangements," *Proc. RECOMB Comparative Genomics Satellite 2006, Lecture Notes in Computer Science 4205,* G. Bourque and N. El-Mabrouk, eds., pp. 113-128, Springer, 2006.
[10] J. Krzywinski, O.G. Grushko, and N.J. Besansky, "Analysis of the Complete Mitochondrial DNA from *Anopheles funestus*: An Improved Dipteran Mitochondrial Genome Annotation and a Temporal Dimension of Mosquito Evolution," *Molecular Phylogenetics and Evolution,* vol. 39, no. 2, pp. 417-423, 2006.
[11] D. Sankoff, C. Zheng, P.K. Wall, C.W. dePamphilis, J. Leebens-Mack, and V.A. Albert, "Internal Validation of Ancestral Gene Order Reconstruction in Angiosperm Phylogeny," *Proc. RECOMB Comparative Genomics Satellite 2008, Lecture Notes in Computer Science 5267,* S. Vialette and C. Nelson, eds., pp. 252-264, Springer, 2008.
[12] J. Savard, D. Tautz, S. Richards, G.M. Weinstock, R.A. Gibbs, J.H. Werren, H. Tettelin, and M.J. Lercher, "Phylogenomic Analysis Reveals Bees and Wasps (Hymenoptera) at the Base of the Radiation of Holometabolous Insects," *Genome Research,* vol. 16, pp. 1334-1338, 2006.
[13] D.W. Severson, B. DeBruyn, D.D. Lovin, S.E. Brown, D.L. Knudson, and I. Morlais, "Comparative Genome Analysis of the Yellow Fever Mosquito *Aedes Aegypti* with *Drosophila melanogaster* and the Malaria Vector Mosquito *Anopheles gambiae*," *J. Heredity,* vol. 95, pp. 103-113, 2004.
[14] G. Tesler, "Efficient Algorithms for Multichromosomal Genome Rearrangements," *J. Computer and System Sciences,* vol. 65, pp. 587-609, 2002.
[15] L.-S. Wang and T. Warnow, "Distance-Based Genome Rearrangement Phylogeny," *Mathematics of Evolution and Phylogen,* O. Gascuel, ed., pp. 353-383, Oxford Univ. Press, 2005.
[16] S. Yancopoulos, O. Attie, and R. Friedberg, "Efficient Sorting of Genomic Permutations by Translocation, Inversion, and Block Interchange," *Bioinformatics,* vol. 21, pp. 3340-3346, 2005.
[17] C. Zheng, "Path Groups: A Common Data Structure for Rapid Heuristic Solutions to Ancestral Gene Order Reconstruction Problems," *Bioinformatics,* vol. 26, pp. 1587-1594, 2010.

Adriana Muñoz is working toward the PhD degree in the Department of Computer Science at the School of Information Technology and Engineering at the University of Ottawa. She received the master's degree in computer science from the University of Alberta. She was a member of the Scientific Research Division and project manager for Nortel for a number of years.

David Sankoff holds the Canada research chair in mathematical genomics at the University of Ottawa.