Rearrangement Phylogeny of Genomes in Contig Form

Adriana Muñoz[1] and David Sankoff[2]

[1] School of Information Technology and Engineering, University of Ottawa
[2] Department of Mathematics and Statistics, University of Ottawa

Abstract. There has been a trend in increasing phylogenetic coverage for genome sequencing while decreasing the sequencing coverage for each genome. With lower coverage, 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 can then reconstruct the phylogeny based on genomic distance and numbers of contigs.

1 Introduction

While the increasing pace of genome sequencing is adding phylogenetic breadth to the inventory of species available for comparative genomics, the sequencing coverage of many of these species is not sufficient to produce completely assembled genomes. Instead the published and archived data remain in contig form, not necessarily associated with chromosomal scaffolds, and there are often no resources allocated to further polishing. The price paid for increasing phylogenetic coverage in genome sequencing is thus the decreasing the sequencing coverage for each genome. With lower coverage, more genomes are being published in contig form.

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

I. Măndoiu, G. Narasimhan, and Y. Zhang (Eds.): ISBRA 2009, LNBI 5542, pp. 160–172, 2009.

genomes available in contig form only? One elegant answer was provided by Gaul and Blanchette [7] for the comparison of two genomes. Their method constructs a number of intermediate structures before actually comparing the genomes. Since we will be using distance matrix methods for phylogenetic analysis, the Gaul and Blanchette procedure is 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 [5]. We compare ten *Drosophila* genomes with two other dipteran genomes and two outlier insect genomes. We discuss this data in Section 2.

In Section 3, we model the behaviour 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 fourteen genomes, in Section 6. Finally, in the Conclusion, we suggest a simplifying hypothesis for further mathematical and empirical work on the contig problem.

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 has already been done Ref [5], we use a carefully constructed inventory of neighbouring gene pairs (NGP) in ten *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-judgement divergence times among the genomes from a number of somewhat contradictory recent publications [8,10,11], as summarized in Figure 1.

Bhutkar *et al.* [2,3] have already used the NGP data for a phylogenetic analysis of *Drosophila*, inferring phylogenies, rearrangements and synteny blocks, but our

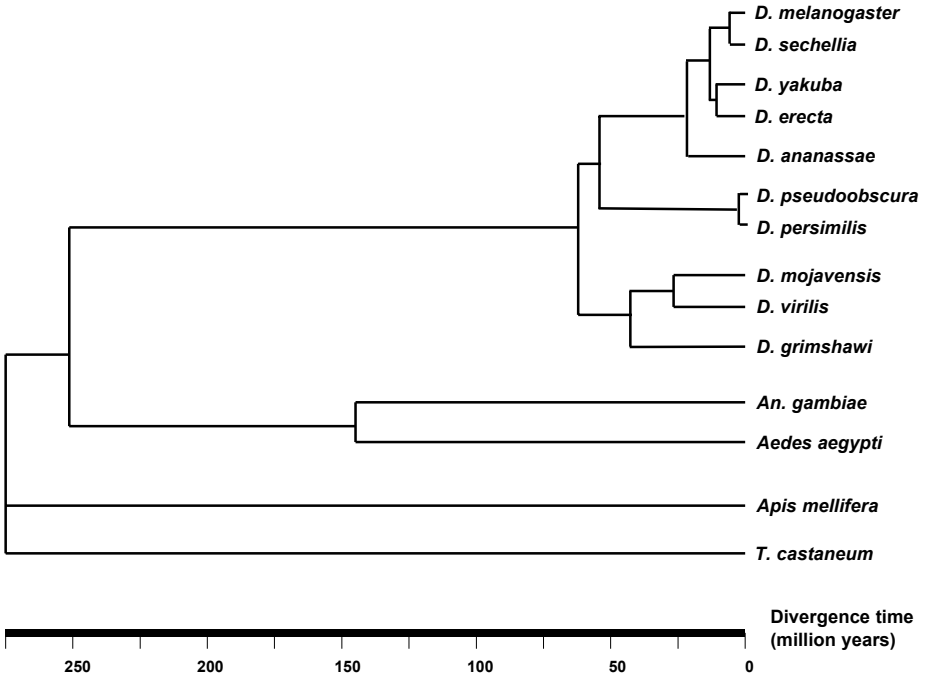

Fig. 1. Phylogeny of Drosophila and outgroups abstracted from the literature, with divergence times

Table 1. Number of contigs constructed for each genome

species (abbreviation)	genes	contigs	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

use of the NGP 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 [5], 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

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 [12], Yancopoulos *et al.* [14], and Bergeron *et al.* [1] Briefly, representing a chromosome a a string of genes $h_1 \cdots h_l$, where a pair of successive genes $h_u h_{u+1}$ are 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:
  $h_1 \cdots h_u \cdots h_l, k_1 \cdots k_v \cdots k_m \rightarrow h_1 \cdots k_v \cdots k_m, k_1 \cdots h_u \cdots h_l$, 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.

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 [4,6,13]. 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 undisrupted 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 τ-th 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 τ-th 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 $\tau-1$-st 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 $d/n \to 1$, as $\tau \to \infty$. 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 Eq 1 both for small and large τ. Combinatorial effects result in $\mathrm{E}(d) < \tau$ even for very small values of τ. And the approach to the asymptote $\frac{\mathrm{E}(d)}{n} \nearrow 1$ is faster than Eq 1 would suggest. We thus have recourse to a 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

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

This analysis resembles the "empirical" approach in Ref [13] 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 Eq 1 at the beginning of this section, whereas Ref [13] takes a purely curve-fitting approach from the outset.

To estimate the parameters λ_1 and λ_2, we simulate pairs of genomes with $n = 8867$, the maximum number of genes used in our *Drosophila melanogaster* comparisons, and τ up to 9000 random rearrangements to derive one genome from the other. We assume the rearrangements are almost exclusively inversions (around 99.8%), reflecting the evolutionary history of *Drosophila*. We use a DCJ

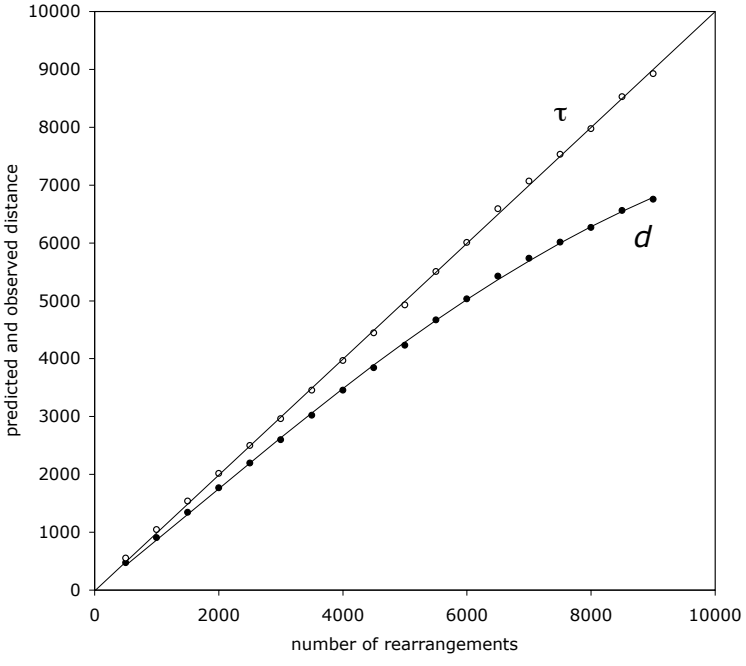

Fig. 2. Predicted (curve) and observed (dots) values of genomic distance d, and inferred (open dots) values of $\hat{\tau}$ versus true (diagonal line) values

algorithm [14,1] to calculate d from the genomes. This is repeated 100 times, and d averaged, to estimate $\mathrm{E}(d)$.

Figure 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 the experimental genomes are of comparable sizes.

4 The Effect of Genome Fragmentation

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 fusions can be separated out as an initial step without changing the total number of rearrangements required. Furthermore, we know exactly how many fusions are required, namely the difference between the number of contigs in A and the number of chromosomes in B. (The optimal scenario will never require both fusions and 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

$$d = d' - |\chi_A - \chi_B|. \tag{5}$$

The absolute value signs accommodate the rare case where $\chi_A < \chi_B$. Since the rearrangement distance can be achieved by doing all the translocations and fusions first, before all the inversions, the correction $|\chi_A - \chi_B|$ is a fixed value and is not dependent on the details of the rearrangement scenario, for which there may be many for a particular data set.

If this whole line of argument were universally valid, we could simply substitute correction Eq 5 into Eq 2 or 4 to estimate τ. In reality, this correction is only appropriate for small values of τ (e.g. $\tau < 0.1n$). For larger values, the apparent rearrangement distance d' based on contigs is inflated less than $|\chi_A - \chi_B|$ over one based on the correctly assembled genomes. The fragmentation of the genome into contigs allows the algorithm, in effect, to compare more similar, albeit incorrect, assemblies. This effect was previously noted in Ref [9]. To circumvent it, we should only remove a proportion α of $|\chi_A - \chi_B|$ from d'. How large a proportion?

To answer this, we undertook a series of simulations, starting from an initial genome B containing 8867 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 on the left of Figure 3. First the linearity of the response to increasing χ_A is clear, at least in the range studied $\chi_A < 1000$, indicating that Eq 5 should be replaced by

$$d = d' - \alpha(\tau)|\chi_A - \chi_B|, \tag{6}$$

where $\alpha(\tau)$ is a decreasing function of the number of rearrangements τ. This decrease is not linear; for practical purposes, we can fit $\alpha(\tau)$ with a quadratic function. Also, as we already know from Eq 3 and Figure 2, d/τ is a decreasing function of τ. This dependence of α and d/τ on τ, as derived from the simulations, are shown on the right of Figure 3.

Given d', then, we can solve Eqs 3 and 6 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 Eqs 4 and 6, 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 on the left of Figure 4. 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 fruitfly and mosquito families within the dipteran order and among the three orders in the class Insecta represented in these data.

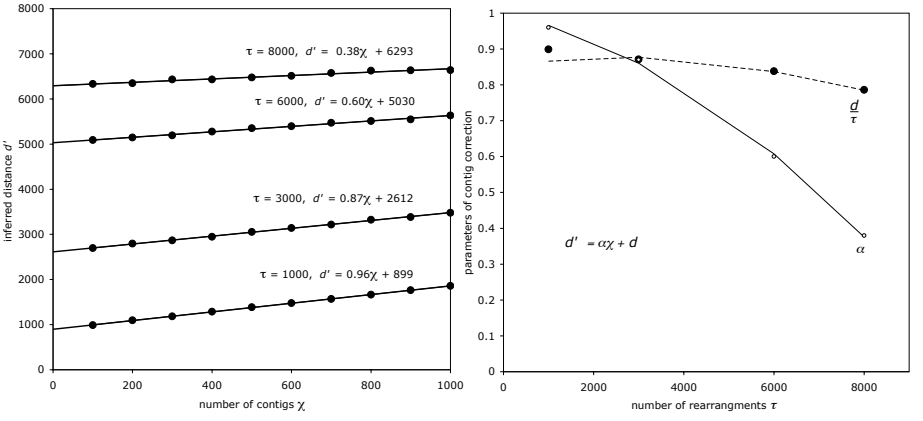

Fig. 3. For genomes generated by $\tau = 1000, 3000, 6000$ or 8000 rearrangements, broken into $\chi = 100, 200, \ldots, 1000$ contigs: (left) the relationship between uncorrected genomic distance d' and χ and equations of trend lines. (right) the parameters α and d of the linear dependence of d' on χ, as a function of τ. Dotted line represents predicted behaviour based on Eq 3. Solid line represents quadratic fit $\alpha(\tau) = 1 - 0.0276(\tau/1000) - 0.0063(\tau/1000)^2$.

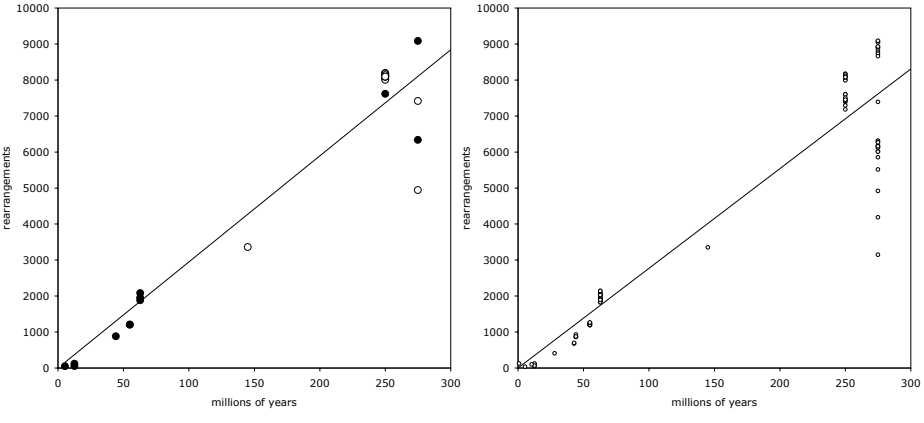

Fig. 4. (left) Comparison of *D. melanogaster* with 13 other genomes (solid dots). Comparison of *Anopheles gambiae* with 13 other genomes (open dots). Divergence, in total number of genome rearrangements, estimated from genomic distances through Eqs 3 and 6, compared to divergence times abstracted from the literature. Line represents least squares fit to all points. (right) Pairwise comparison of all pairs of 14 genomes, as discussed in Section 5. Line represents least squares fit.

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 $n = 8867$ separated by $\tau = 1000, 3000, 6000$ and 8000 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 Figure 5.

We observe on the left of Figure 5 that for fixed τ and χ_A, the response of d' to increasing χ_B is systematically linear. This is clear up to $\tau = 6000$ and only starts to break down for $\tau = 8000$ and $\chi_A \geq 600$, where examination of the data on an expanded scale shows that d' actually decreases somewhat initially, then increases, as χ_B increases (not discernible in Figure 5). The linear rate of increase of d', plotted as $\beta(\tau, \chi_A)$ on the right of the figure, is the same as the $\alpha(\tau)$ in Figure 3 for low 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 Figure 3, depending on τ, after which the rate of increase drops off somewhat.

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 Eq 4 simultaneously with

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

where $\beta(\tau, \chi_A) = \alpha(\tau) - (.00027 - .00003\tau)\chi_A$, and where the coefficient of χ_A is estimated by a least squares fit to the slopes of the four trend lines in Figure 6 (right).

Plotting the inferred values of τ against values extracted from the literature produced the results on the right of Figure 4.

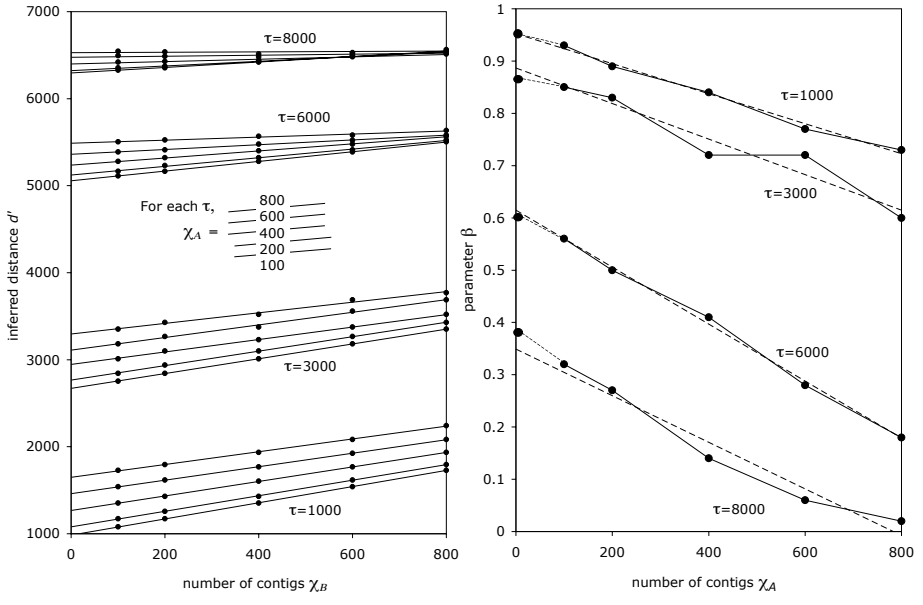

Fig. 5. For genomes A and B separated by $\tau = 1000, 3000, 6000$ or 8000 random rearrangements, broken into $\chi = 100, 200, 400$ or 800 contigs: (left) the relationship between uncorrected genomic distance d', χ_A and χ_B, with trend lines for each χ_A connecting the values of d' for a range of χ_B. (right) the coefficient β of the linear dependence of d' on χ_B in the left hand diagram, as a function of χ_A. Dotted segments connect $\beta(\tau, 100)$ to $\beta(\tau, 0) = \alpha(\tau)$ from Figure 3. Dashed line is the linear trend line, not taking account of $\beta(\tau, 0)$.

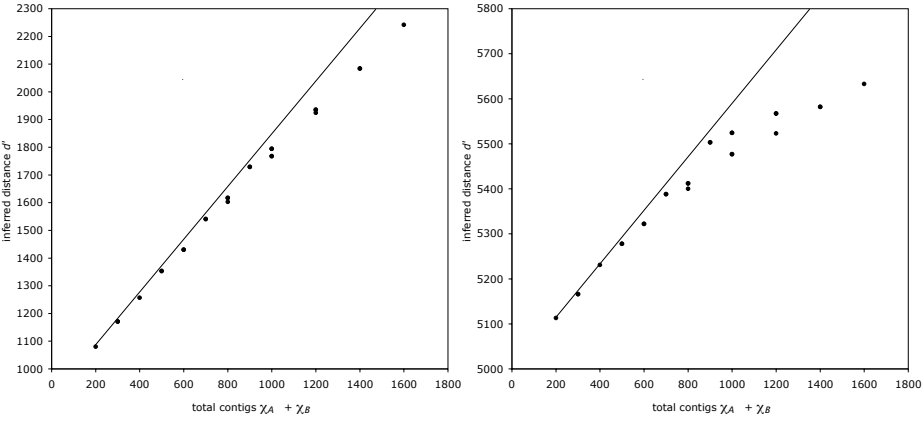

Fig. 6. Dependence of d' on the total number of contigs in the two genomes, for $\tau = 1000$ (left) and $\tau = 6000$ (right). Straight lines represent $d' = d + \alpha(\tau)(\chi_A + \chi_B)$ where $\alpha(\tau)$ is as in Figure 3.

6 Phylogeny

If we input the inferred pairwise values of τ into a neighbour-joining routine, we produce the phylogeny in Figure 7. When this is compared to Figure 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 other *Diptera*. Note that using the uncorrected matrix of d' as input to neighbour joining does not show this rate effect as clearly as τ and also introduces other structural errors into the phylogeny.

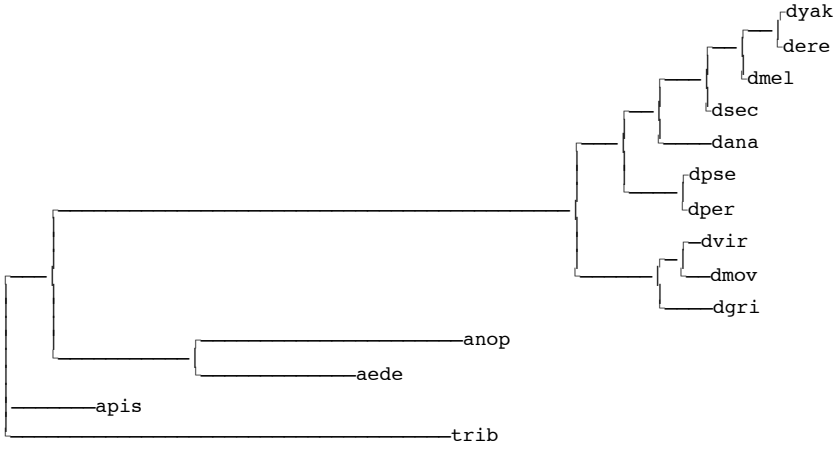

Fig. 7. Neighbour-joining phylogeny based on matrix of inferred number of rearrangements τ

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 intuitive connections between genomic distance and adjacency disruption.
- 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 6000 for genomes of size around $n = 9000$, and up to $\chi = 1000$ contigs.

– 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.

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. 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 Ref [13], 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 Ref [6] than by further characterization of empirical models such Eq 3.

For example, if we knew the probability law of d under random rearrangements, or even its expectation, we could most easily investigate the following hypothesis, suggested by our results on the linearity of the dependence of d' on χ: *The imposition of a contig structure has the same effect on* d *as adding further rearrangements.* In a continuous approximation of the τ—d relationship,

$$\frac{d\mathrm{E}(d)}{d\chi} = \frac{d\mathrm{E}(d)}{d\tau} = \alpha(\tau). \tag{8}$$

If this could be verified, analytically or, failing that, empirically, it would make for an elegant framework for our results.

Acknowledgements

We thank Arjun Bhutkar for providing the NGP files with pair occurrence tabulated by species. We also thank Chunfang Zheng for guidance in using her programs for DCJ distance and rearrangement simulations. Research funded in part by NSERC. DS holds the Canada Research Chair in Mathematical Genomics.

References

1. Bergeron, A., Mixtacki, J., Stoye, J.: A unifying view of genome rearrangements. In: Bücher, P., Moret, B.M.E. (eds.) WABI 2006. LNCS (LNBI), vol. 4175, pp. 163–173. Springer, Heidelberg (2006)
2. Bhutkar, A., Gelbart, W.M., Smith, T.: Inferring genome-scale rearrangement phylogeny and ancestral gene order: a *Drosophila* case study. Genome Biology 8, R236 (2007)
3. Bhutkar, A., Schaeffer, S.W., Russo, S.M., Xu, M., Smith, T.F., Gelbart, W.: Chromosomal rearrangement inferred from comparisons of 12 *Drosophila* genomes. Genetics 179, 1657–1680 (2008)

4. Dalevi, D., Eriksen, N.: Expected gene-order distances and model selection in bacteria. Bioinformatics 24, 1332–1338 (2008)
5. *Drosophila* 12 Genomes Consortium. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450, 203–218 (2007)
6. Eriksen, N., Hultman, A.: Estimating the expected reversal distance after a fixed number of reversals. Advances in Applied Mathematics 32, 439–453 (2004)
7. Gaul, E., Blanchette, M.: Ordering partially assembled genomes using gene arrangements. In: Bourque, G., El-Mabrouk, N. (eds.) RECOMB-CG 2006. LNCS (LNBI), vol. 4205, pp. 113–128. Springer, Heidelberg (2006)
8. Krzywinski, J., Grushko, O.G., Besansky, N.: 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 39, 417–423 (2006)
9. Sankoff, D., Zheng, Wall, C.P.K., de Pamphilis, C.W., Leebens-Mack, J., Albert, V.A.: Internal validation of ancestral gene order reconstruction in angiosperm phylogeny. In: Vialette, S., Nelson, C. (eds.) RECOMB-CG 2008. LNCS, vol. 5267, pp. 252–264. Springer, Heidelberg (2008)
10. Savard, J., Tautz, D., Richards, S., Weinstock, G.M., Gibbs, R.A., Werren, J.H., Tettelin, H., Lercher, M.J.: Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. Genome Research 16, 1334–1338 (2006)
11. Severson, D.W., DeBruyn, B., Lovin, D.D., Brown, S.E., Knudson, D.L., Morlais, I.: Comparative genome analysis of the yellow fever mosquito *Aedes aegypti* with *Drosophila melanogaster* and the malaria vector mosquito *Anopheles gambiae*. Journal of Heredity 95, 103–113 (2004)
12. Tesler, G.: Efficient algorithms for multichromosomal genome rearrangements. Journal of Computer and System Sciences 65, 587–609 (2002)
13. Wang, L.-S., Warnow, T.: Distance-based genome rearrangement phylogeny. In: Gascuel, O. (ed.) Mathematics of Evolution and Phylogeny, ch. 13, Oxford, pp. 353–383 (2005)
14. Yancopoulos, S., Attie, O., Friedberg, R.: Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21, 3340–3346 (2005)